# Efficient Analysis of Complex, Multi-modal Genomic Data

by

Chaitanya Ramanuj Acharya

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

_____
Andrew S. Allen, Supervisor

_____
Sayan Mukherjee, Chair

_____
Sandeep S. Dave

_____
Kouros Owzar

Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate Program in Computational Biology and Bioinformatics
in the Graduate School of Duke University
2016

# Abstract

# Efficient Analysis of Complex, Multi-modal Genomic Data

by

Chaitanya Ramanuj Acharya

Graduate Program in Computational Biology and Bioinformatics
Duke University

Date: _____

Approved:

_____
Andrew S. Allen, Supervisor

_____
Sayan Mukherjee, Chair

_____
Sandeep S. Dave

_____
Kouros Owzar

An abstract of a dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate Program in Computational Biology and Bioinformatics
in the Graduate School of Duke University
2016

# Abstract

Our primary goal is to better understand complex diseases using statistically disciplined approaches. As multi-modal data is streaming out of consortium projects like Genotype-Tissue Expression (GTEx) project, which aims at collecting samples from various tissue sites in order to understand tissue-specific gene regulation, new approaches are needed that can efficiently model groups of data with minimal loss of power. For example, GTEx project delivers RNA-Seq, Microarray gene expression and genotype data (SNP Arrays) from a vast number of tissues in a given individual subject. In order to analyze this type of multi-level (hierarchical) multi-modal data, we proposed a series of efficient-score based tests or score tests and leveraged groups of tissues or gene isoforms in order map genomic biomarkers. We model group-specific variability as a random effect within a mixed effects model framework. In one instance, we proposed a score-test based approach to map expression quantitative trait loci (eQTL) across multiple-tissues. In order to do that we jointly model all the tissues and make use of all the information available to maximize the power of eQTL mapping and investigate an overall shift in the gene expression combined with tissue-specific effects due to genetic variants. In the second instance, we showed the flexibility of our model framework by expanding it to include tissue-specific epigenetic data (DNA methylation) and map eQTL by leveraging both tissues and methylation. Finally, we also showed that our methods are applicable on different data type such as whole transcriptome expression data, which is designed to analyze genomic events such alternative gene splicing. In order to accomplish this, we proposed two different mod-

els that exploit gene expression data of all available gene-isoforms within a gene to map biomarkers of interest (either genes or gene-sets) in paired early-stage breast tumor samples before and after treatment with external beam radiation. Our efficient score-based approaches have very distinct advantages. They have a computational edge over existing methods because they do not need parameter estimation under the alternative hypothesis. As a result, model parameters only have to be estimated once per genome, significantly decreasing computation time. Also, the efficient score is the locally most powerful test and is guaranteed a theoretical optimality over all other approaches in a neighborhood of the null hypothesis. This theoretical performance is born out in extensive simulation studies which show that our approaches consistently outperform existing methods both in statistical power and computational speed. We applied our methods to publicly available datasets. It is important to note that all of our methods also accommodate the analysis of next-generation sequencing data.

To my beloved parents, Sridevi and Srinivas, for their love and sacrifices, and Kelly, who is the love of my life, and my son, Sameer, a new source of strength and inspiration.

# Contents

# List of Tables

# List of Figures

# List of Abbreviations and Symbols

## Abbreviations

| | |
|---|---|
| SNP | Single Nucleotide Polymorphism |
| QTL | Quantitative Trait Loci |
| eQTL | expression Quantitative Trait Loci |
| GWAS | Genome-Wide Association Study |
| CpG | CG-rich oligodeoxynucleotide island |
| TBT | Tissue-by-Tissue or Transcript-by-Transcript |
| GTEx | Genotype-Tissue Expression Project |
| BMA | Bayesian Model Averaging |
| FE | Fixed-effects |
| RE | Random-effects |
| MT | MetaTissue model |
| PVE | Proportion of Variation Explained |
| CRBLM | Cerebellum |
| FCTX | Frontal Cortex |
| TCTX | Temporal Cortex |
| HIPP | Hippocampus |
| MEDU | Medulla |
| OCTX | Occipital Cortex |
| PUTM | Putamen |

| | |
|---|---|
| SNIG | Substantia Nigra |
| THAL | Thalamus |
| WHMT | Intra-lobular White Matter |
| dbGAP | Database of Genotypes and Phenotypes |
| GEO | Gene Expression Omnibus |
| UKBEC | UK Brain Expression Consortium |
| BH | Benjamini-Hochberg method |
| FDR | False Discovery Rate |
| FWER | Family-wise Error Rate |
| TME | Tumor Microenvironment |
| RMA | Robust Multi-array Average method |
| GSVA | Gene Set Variational Analysis |
| PLAGE | Pathway Level Analysis of Gene Expression |
| GSEA | Gene Set Enrichment Analysis |
| ssGSEA | single sample GSEA |
| ZSCORE | Combined z-score method |
| GSA | Gene Set Analysis |
| MSigDB | Molecular Signatures Database |
| DE | Differentially Expressed |
| ROC | Receiver-Operator Characterisitic Curve |
| AUC | Area Under the ROC Curve |

# Acknowledgements

Immense thanks to Dr. Andrew S. Allen, my primary advisor, for giving me an opportunity to work with him, when all seemed lost for me at the beginning of my PhD program. It took courage and conviction from him to bring me under his wings, train me and give me direction during my years as a graduate student, and I am forever indebted to him. I would like to acknowledge the pivotal role played by Dr. Kouros Owzar, who is my *de-facto* co-advisor, for his guidance and support during the early years of my graduate school. I would poach his invaluable time, mostly without any prior appointments, either towards the end of the day or during a bathroom break. My sincere thanks to my other committee members, Dr. Sayan Mukherjee and Dr. Sandeep S. Dave, for their willingness to help me with any issue I had during my graduate school. I would also like to thank Dr. Elizabeth DeLong, who lent an invisible hand in starting my tenure as a graduate student within the auspices of the Department of Biostatistics and Bioinformatics at Duke University. It would be a shame if I did not recall the role played by my former mentor and now retired, Dr. Joseph Nevins, who was instrumental in shaping my career as a computational biologist, with a focus in genomic medicine. He impressed upon me the significance of curiosity and industry in science, and I would always remember the lessons I have learned from him for the rest of my life. I would also like to thank Drs. Janice McCarthy and Yu Jiang, for giving me some time for useful discussions that helped me achieve my research goals. Working with high-throughput data requires immense computational resources, and I would like to thank Thomas Milledge at the Duke Scalable Computing Support Center

for his patience and IT support throughout my graduate program. I would also like to thank Dr. Mark DeLong, a friend and a favorite colleague, for his guidance and enriching friendship that helped me wade through the tricky waters of graduate school.

Duke CBB program had been a blessing in disguise for me. I would first like to thank each of the CBB directors, John Harer, Alex Hartemink, and Paul Magwene, and DGSs, Jeanette McCarthy, Steve Haase, and Scott Schmidler, for their service and dedication to the program. I would remiss if I did not thank Drs. Alexander Hartemink, Scott Schmidler, Steve Haase and Jeanette McCarthy for their advice in time of need. Liz Labriola is the backbone of the program and this is no secret. I would like to thank her for her guidance and most importantly, her patience with me. Duke CBB student community is very diverse and I would like to single out my class-mates including Jason Belsky, Andrea Moffit, Jianling Zhong, Kaixuan Luo, Yu Jiang, Samuel Ramirez, Zach Scholl and David Winski for great memories and lasting friendships.

Finally, I'd like to thank my parents, Sridevi and Srinivas Maddali for their eternal love and unwavering support. They have instilled the values in me in order to make me the person that I am today. For a kid, who grew up in a lower middle-class family in India, dreaming big was not without its perils. They showed me the way against insurmountable odds, not without making innumerable sacrifices along the way. In the year 2008, I met my wife, Kelly Acharya, who gave me a reason to succeed in life. An accomplished physician herself, she inspires me to excel in life and make me a better person every day. I would like to thank her for being my bedrock. My life changed for good when my son, Sameer, was born. I want to thank him for giving me a renewed sense of purpose in life, and for his magical ability to turn my bad days into good with his incandescent smile.

# 1

# Introduction

Modern molecular high-throughput assays include diverse technologies such as microarray gene expression, single nucleotide polymorphism (SNP) chip array, microRNA expression, proteomics, metabolomics, DNA methylation, and next-generation sequencing among many others. These technologies have matured and have become increasingly accessible, and have enabled us to transition from the days of studying one gene at a time to analyzing thousands of them, simultaneously. The next frontier is to collect multi-modal data from each individual subject with a goal to understand complex diseases, and "quantify the biology" by conducting an integrative analysis that befits the hierarchical nature of the data. In that process, developing studies with statistically disciplined approaches must be emphasized in order to integrate multi-subject, multi-modal data. Our ultimate goal - dissect and understand the underlying biology of complex disease.

Many serious research efforts in the form of consortia studies have been undertaken in order to close our knowledge gap, such as, Genotype-Tissue Expression (GTeX) project (Lonsdale and et al, 2013), 1000 Genomes project (1000 Genomes Project Consortium et al., 2015), HapMap project (Consortium., 2003), Human Epigenome project (Bradbury, 2003) and Encyclopedia of DNA Elements (ENCODE) (Consortium, 2004) just to name a

few. Each of these projects focus on different data modalities within the same individual. For example, GTEx project focuses on creating a repository of human tissue bank, which aims at studying human gene expression and its relationship to genetic variation. This involves two data modalities within the same subject i.e. gene expression from different tissue samples and germline variation in the form of SNPs. The objective of GTEx project is to map regions of the genome containing DNA sequence variants that influence the expression of one or more genes called expression quantitative trait loci (eQTL) (see figure 1.1). Genetic variation can also influence gene expression through alterations in splicing (alternative splicing), non-coding RNAs (regulatory regions), and RNA stability. As more and more types of data are collected, the relationships between genetic variation and gene expression can be expanded to include correlations with other data modalities such as epigenetics or proteomics.

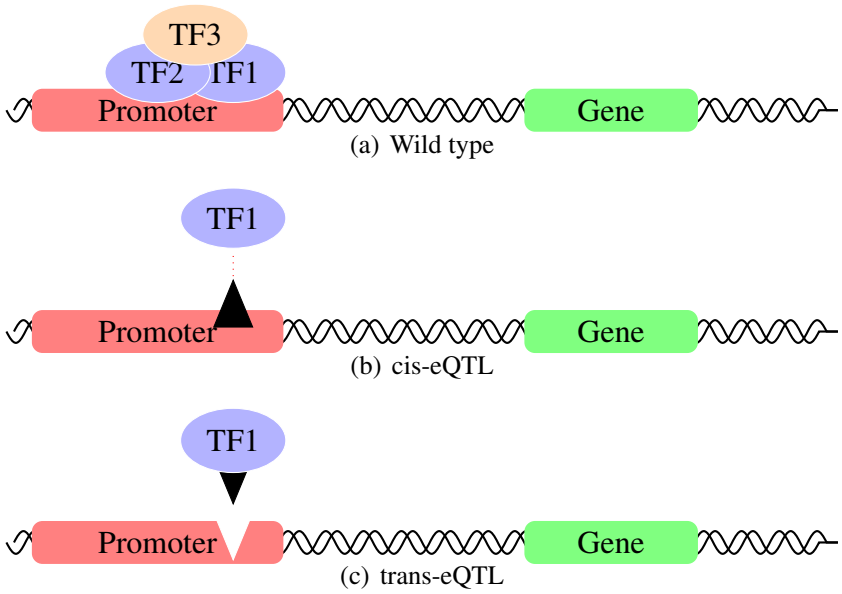



FIGURE 1.1: **Expression quantitative trait loci (eQTL):** A) In a wildtype, both transcription factor (TF) and all genes are transcribed to their full potential. B) A variant in the promoter region of a gene hinders the TF binding, causing in the reduction in the rate at which the gene is transcribed. C) A variant occurring in the region coding for a transcription factor hinders its binding to the promoter region of many genes causing in the reduction of the rate at which genes are transcribed.

As next-generation sequencing (NGS) technologies have made it possible to generate high-resolution genomic data much more efficiently, researchers in many fields have turned to next-generation sequencing to identify particular features of the genome, such as the aforementioned alternative splicing, that contribute to specific phenotypes. Whole genome, exome and transcriptomes are more accessible than ever for interpretation using the high resolution NGS assays. However, as the complexity of the genomic data increases, the statistical and analytical challenges mount. Statistical approaches to analyze individual data modalities are well developed. However, new statistical approaches that accommodate the analysis of multi-modal genomic data are increasingly in demand.

The objective of this dissertation thesis is to propose statistically disciplined framework for complex data analysis with an aim to dissect and understand the molecular phenotypes. There is a need to build a coherent theoretical framework to accommodate correlations for various types of outcomes in relation to many sources of variations. We provide three topics of discussion in the form of three chapters that focus of three specific problems involving multiple different data modalities.

### *1.0.1 TOPIC 1: Leveraging multiple tissues to map eQTLs.*

Our first chapter focuses on the first case study (Acharya et al., 2016), which involves mapping eQTL in multiple tissues. Regulatory regions in higher eukaryotes activate gene transcription in a tissue-specific manner, and genetic variants found within these regulatory regions may have variable effects on gene expression across different tissues or cell-types. The approach described in our manuscript exploits this fact, by identifying expression quantitative trait loci (eQTL) by explicitly looking for germ-line variation that results in either overall shift in gene expression or tissue-specific differences over the most popular approach, a tissue-by-tissue approach (Shabalin, 2012; Gatti et al., 2009; Lippert et al., 2011; Scott-Boyer et al., 2012; Sun, 2009), which is by far the most inefficient method of eQTL discovery in multi-tissue studies.

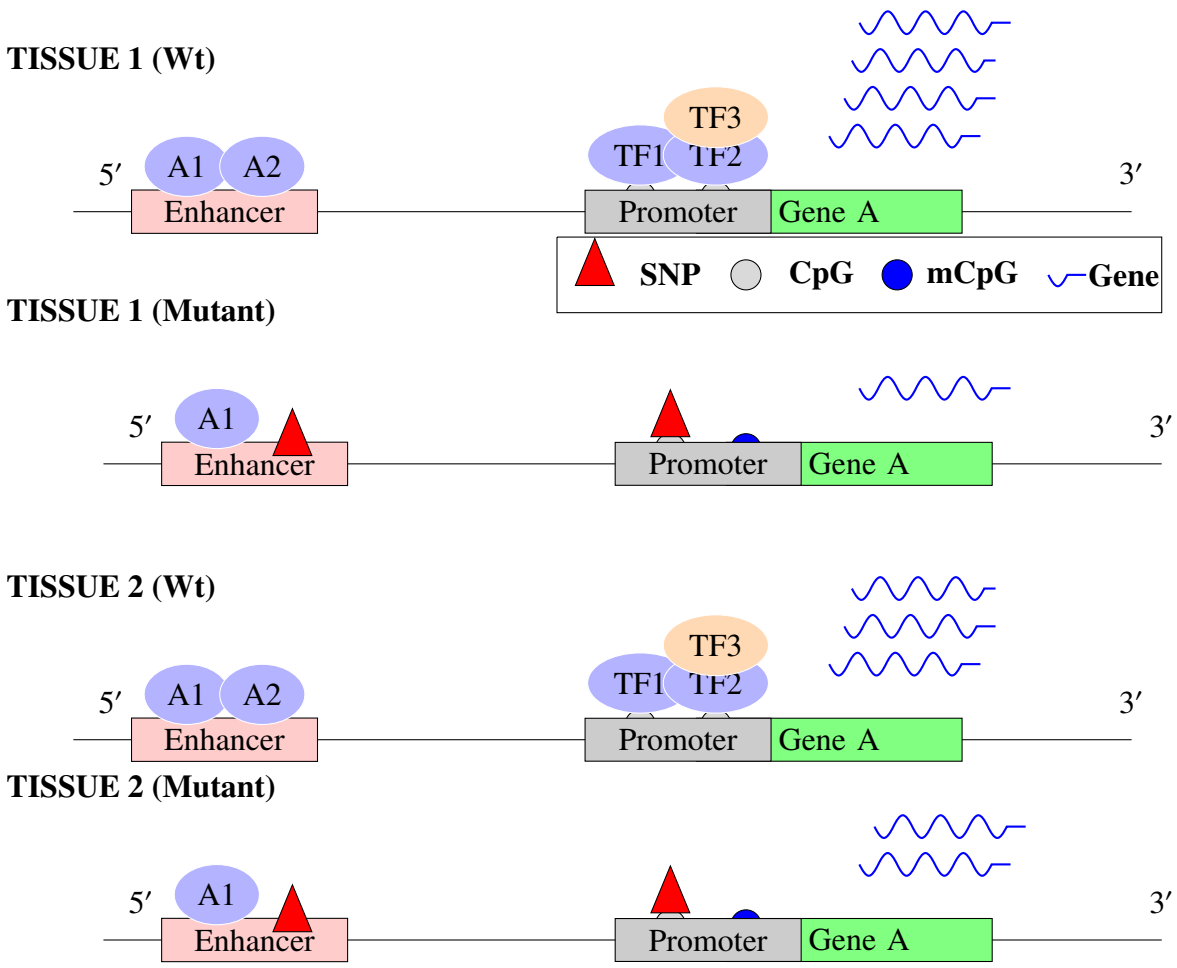



FIGURE 1.2: **Tissue-specific gene expression.**An illustration of tissue-specific gene expression of Gene A (quantified by blue squiggly lines) and the effect of the same genetic variant (denoted by red triangle labeled SNP) and the methylation status of the proximal CpG island (denoted by shaded semi-circle) in its expression in tissues 1 and 2

Though other approaches have been proposed that jointly model gene expression across multiple tissues, our approach provides a number of notable improvements (Flutre et al., 2013; Sul et al., 2013). Specifically, for a given gene-SNP pair, we model gene expression across tissues using a linear mixed model (Jiang, 2007; Verbeke and Molenberghs, 2000) in which both fixed and random effects are used to capture the effect of a variant on gene expression. Our approach is based on efficient score statistics, and as such model parameters only need to be estimated under the null hypothesis. This simple difference results

in a tremendous benefit computationally, as, in our approach, model parameters only need to be estimated once per genome. This increased efficiency is important, as it allows more accurate, permutation or Monte Carlo based, assessments of statistical significance and the ability to address denser marker or sequencing-based studies. Finally, the efficient score is the locally most powerful test and is guaranteed a theoretical optimality over all other approaches in a neighborhood of the null hypothesis. This theoretical performance is born out in extensive simulation studies which show that our approach consistently outperforms existing methods both in statistical power and computational speed.

We illustrate our approach by analyzing two publicly available expression datasets from normal human brains. The first is comprised of four brain regions from 150 neuropathologically normal samples (Gibbs et al., 2010). The second, from the UK Brain Expression Consortium (UKBEC), is comprised of ten brain regions from 134 neuropathologically normal samples (Ramaswamy et al., 2014). We show that our approach can identify eQTL within more genes than existing methods.

We implement our approach within the R package "JAGUAR", which is now available at the Comprehensive R Archive Network (CRAN) repository (Acharya and Allen, 2016).

### *1.0.2 TOPIC 2: Leveraging DNA methylation and multiple tissues to map eQTLs.*

The second chapter would focus on the second case study (**?**), which expands our model proposed in the first case study by including epigenetic data in tissue-specific DNA methylation. DNA methylation is an important tissue-specific epigenetic event that influences transcriptional regulation of gene expression (Deaton and Bird, 2011; Wrzodek et al., 2012; Shoemaker et al., 2010). Differentially methylated CpG sites may act as mediators between genetic variation and gene expression, and this relationship can be exploited while mapping multi-tissue expression quantitative trait loci (eQTL). Current multi-tissue eQTL mapping techniques are limited to only exploiting gene expression patterns across multiple tissues either in a joint tissue or tissue-by-tissue frameworks.

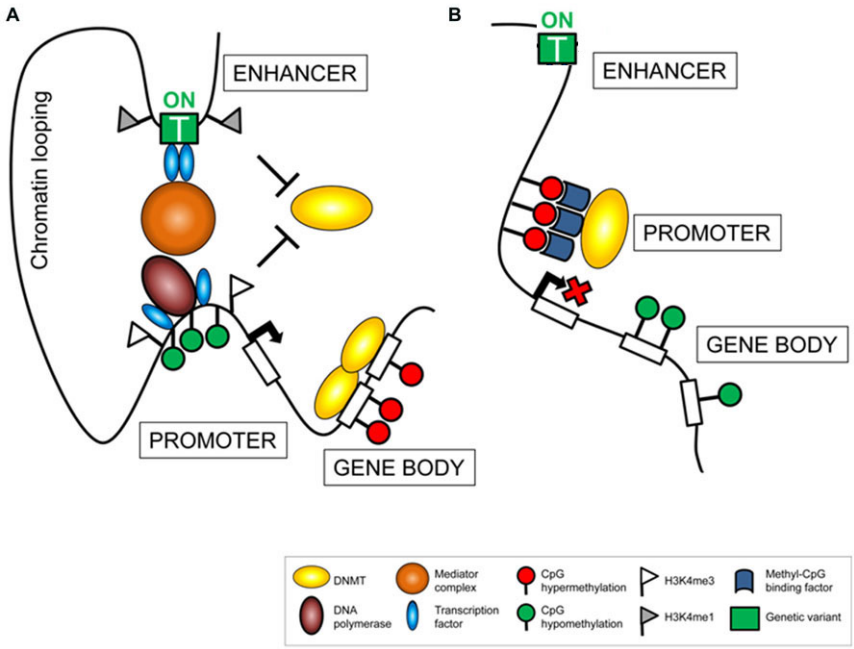



FIGURE 1.3: **Epigenetics and genetics.** The interaction between DNA methylation and genetic variant in the form of a single nucleotide polymorphism regulates gene expression. Source: A symbiotic liaison between the genetic and epigenetic code. Heyn, H. Frontiers in Genetics, 01 May 2014. http://dx.doi.org/10.3389/fgene.2014.00113

Consider a hypothetical DNA strand with a gene body and its associated regulatory region consisting of an upstream promoter and enhancer regions (Figure 1.3). The common element in both figures 1.3a and 1.3b is the presence of a genetic variant in the enhancer region. However, in figure 1.3a, it is clear that the presence of undermethylation or hypomethylated CpG islands in the promoter region is facilitating the formation of the transcription factor complex, and thereby promoting the transcription process. Hypermethylation of the same promoter CpG islands is disrupting the formation of the transcription factor complex and thus repressing the gene transcription process. This latent interaction effect between DNA methylation and genetic variant across multiple tissues is not explored by any current methods. We present a new statistical approach that enables us to model the effect of germ-line variation on tissue-specific gene expression in the presence of effects due to DNA methylation. Our method efficiently models genetic and epigenetic variation to identify genomic regions of interest containing combinations of mRNA transcripts,

CpG sites, and SNPs by jointly testing for genotypic effect and higher order interaction effects between genotype, methylation and tissues. We demonstrate using Monte Carlo simulations that our approach, in the presence of both genetic and DNA methylation effects, gives an improved performance (in terms of statistical power) to detect eQTLs over the current eQTL mapping approaches. When applied to an array-based dataset from 150 neuropathologically normal adult human brains, our method identifies eQTLs that were undetected using standard tissue-by-tissue or joint tissue eQTL mapping techniques. As an example, our method identifies eQTLs in a BAX inhibiting gene (TMBIM1), which may have a role in the pathogenesis of Alzheimer disease (Kudo et al., 2012; Lisak et al., 2015).

As previously stated, our score test-based approach does not need parameter estimation under the alternative hypothesis. As a result, our model parameters are estimated only once for each mRNA - CpG pair. Our model specifically studies the effects of non-coding regions of DNA (in this case, CpG sites) on mapping eQTLs. However, we can easily model micro-RNAs instead of CpG sites to study the effects of post-transcriptional events in mapping eQTL. Finally, our model's flexible framework also allows us to investigate other genomic events such as alternative gene splicing by extending our model to include gene isoform-specific data.

### *1.0.3 TOPIC 3: Leveraging multiple gene isoforms to map differentially expressed genes and differentially enriched gene-sets.*

Chapter 3 shifts focus from mapping eQTLs to mapping biomarkers including genes and gene-sets post treatment. In an effort to understand the underlying biology of radiation response along with whole transcriptome effects of preoperative radiotherapy in early-stage breast tumors, we propose two new statistical methods that exploit gene expression patterns across all available gene transcript isoforms and identify potential biomarkers, which are representative of tumor microenvironment (Carter et al., 2006; Chi et al., 2006;

Liu et al., 2007; Hsu et al., 2010; Widschwendter et al., 2007; Chang et al., 2004; Viemann et al., 2006) and tumor biology of radiation response (Liberzon et al., 2015).

Standard microarray designs demonstrated that they can detect alternative splicing, but many types of alternative splicing have probably been missed because of technical limitations such as 3′ labeling bias and the absence of probes designed to detect splice junctions. In order to address this issue and perform whole transcriptome analysis, Affymetrix Human Transcriptome Array 2.0 GeneChip arrays (Affymetrix, 2016) were introduced as a cheaper alternative to RNA-Seq assays. In order to ensure uniform coverage of the transcriptome, Human Transcriptome Array 2.0 was designed with approximately ten probes per exon and four probes per exon-exon splice junction. At the top level, each transcript cluster roughly corresponds to a gene. Each transcript cluster is comprised of exon clusters that a) shared splice sites, b) or were derived from overlapping exonic sequences, c) or were single-exon clusters bounded on the genome by spliced content. Each exon cluster is further fragmented into probe selection regions (PSRs), which are non-overlapping contiguous sequences. This hierarchy of probe design resembles that of Affymetrix Human Exon ST Array however, it is more comprehensive than any of the previous microarray platforms.

We demonstrate the effectiveness of our two methods using extensive simulation studies that show that both of our methods give an improved performance, in terms of statistical power, over the most commonly used methods. By exploiting radiation-induced changes in all available gene transcript isoforms i.e. human transcriptome, we identified several statistically significant differentially genes related to PI3K-AKT and JAK-STAT signaling pathways along with radiation-induced oncogenic signaling pathways and tumor microenvironment gene signatures that could be potential targets to improve response to radiotherapy in breast tumors.

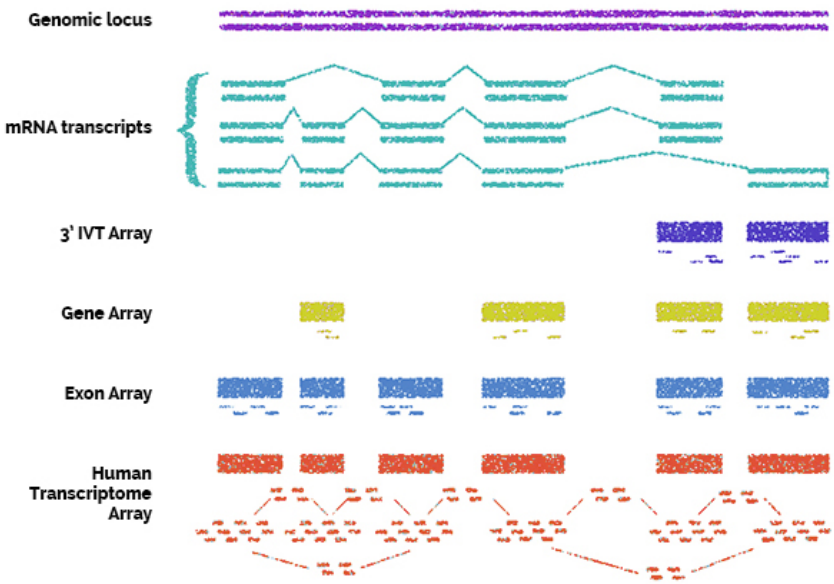



FIGURE 1.4: **A schematic overview of the Affymetrix Transcriptome Array 2.0.** The top row is a gene complete with exons and introns, the second row indicates the mRNA transcripts that are formed following splicing events. Splicing leads to alternative isoforms of the same gene. Transcriptome array probe set includes the short dashes shown in the last row underneath the exon bars. These probe-sets also include the exon splice junctions. Image Source: http://mbcf.dfci.harvard.edu/

# 2

# Exploiting expression patterns across multiple tissues to map expression quantitative trait loci

## 2.1 Introduction

Combining genetic and gene expression data has emerged as a powerful strategy for systematically unraveling the effects of genetic variation on disease (Brem et al., 2002). A common approach is to identify genetic variants that are correlated with gene expression in one or more genes (Cookson et al., 2009). Such variants are referred to as expression quantitative trait loci (eQTL). Since regulatory regions in higher eukaryotes activate gene transcription in a tissue-specific manner, genetic variants found within these regulatory regions may have variable effects on gene expression across different tissues or cell-types (See Figure 1.1).

For example, a genetic variant found near the promoter region of the catechol-O-methyl transferase (*COMT*) gene, which has been implicated in schizophrenia, is associated with differential *COMT* expression across regions of the brain during the course of the illness(Harrison and Weinberger, 2005). This spatio-temporal gene expression pattern has been shown to be strongly associated with structural abnormalities such as the loss

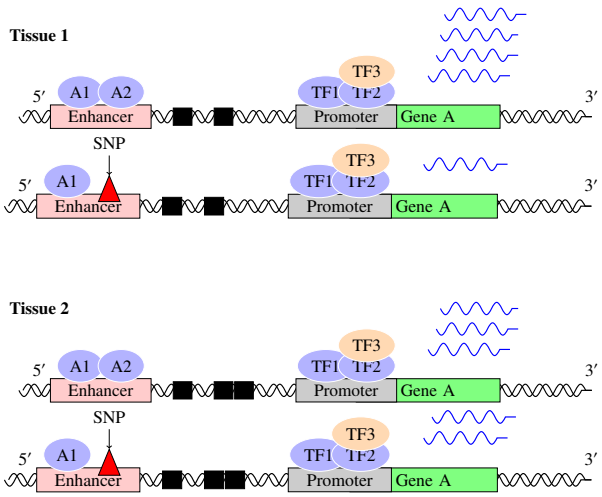



FIGURE 2.1: **An illustration of the regulation of tissue-specific gene regulation.** In this example, we illustrate the concept of tissue-specific gene expression using Gene A (quantified by blue squiggly lines) and its genetic variant (denoted by red triangle labeled SNP) in two tissues, tissue 1 and tissue 2. In both tissues 1 and 2, left panel indicates the wild-type gene expression of Gene A and the right panel indicates a reduced gene expression in the presence of a genetic variant, shown here by the reduced number of blue squiggly lines. It is clear from the figure that there is a difference in baseline gene expression levels of Gene A in tissues 1 and 2 and there is a difference in the degree to which the gene expression is repressed by the genetic variant.

of brain volume in the frontal cortex and hippocampus that is part of the natural progression of schizophrenia(Harrison, 1999). Studying the underlying biology of this tissue- or region-specific gene expression variation is essential in understanding various complex diseases (Cookson et al., 2009).

Many approaches to identifying eQTL utilize a marginal analysis of a single variant's effect on gene expression in a single tissue. Such analyses are then repeated on each tissue leading to a tissue-by-tissue (TBT) approach(Shabalin, 2012; Broman et al., 2003; Sun, 2009; Pletcher et al., 2004; Gatti et al., 2009; Scott-Boyer et al., 2012). However, such an approach has at least three significant limitations: First, a TBT analysis fails to fully exploit expression patterns across the tissues either by pooling information when a variant has a similar effect across multiple tissues or by explicitly identifying effects that differ

across tissues. Second, marginal analyses of individual tissues lead to a proliferation of hypotheses tested, which can negatively impact the power of eQTL discovery. Third, even when one identifies a variant that affects expression of a given gene in a given tissue via a tissue-by-tissue approach, it is not clear whether the effect is tissue-specific or is shared across multiple tissues since such a hypothesis is not explicitly tested. Hence, multi-tissue eQTL studies, such as the Genotype-Tissue Expression (GTEx) project(Lonsdale and et al, 2013), in which expression is measured in up to 30 tissue sites in each individual, require new analytic approaches to fully exploit the information in these samples. Recently, two methods have been proposed that attempt to take better advantage of the information across multiple tissues. Sul *et al.* proposed the MetaTissue (MT) approach(Sul et al., 2013), which combines tissue-specific effects across multiple tissues in a meta-analytic framework. MT uses a mixed effects meta analytic framework that not only accounts for the correlation of gene expression between tissues but also heterogeneity of the effects across tissues. Flutre *et al.* proposed a Bayesian hierarchical model (eQTL-BMA) that models the joint distribution of gene expression across tissues and "combines information across genes to estimate the relative frequency of patterns of eQTL sharing among tissues"(Flutre et al., 2013).

Both MT and eQTL-BMA require optimization under the alternative hypothesis (the given SNP is an eQTL in at least 1 tissue), and thus require the estimation of all model parameters for each gene by variant combination. As a result, the computational demands of both approaches scale very poorly with increasing numbers of variants or genes. To address this issue, we propose a score test-based approach which does not require parameter estimation under the alternative hypothesis. As a result, model parameters only have to be estimated once per genome, significantly decreasing computation time. Further, our score-based approach only requires estimation of the first two moments of the random effects, thus it is robust to misspecification of the random effect distribution(Lin, 1997). We

evaluate our method using extensive simulation studies that show a significant increase in power to detect eQTL when compared to a TBT approach. Furthermore, we show that our method surpasses currently existing joint modeling approaches such as MetaTissue eQTL and eQTL-BMA in terms of computational speed and yet provides a comparable performance with respect to statistical power. Finally, we demonstrate its effectiveness by applying it to two publicly available expression datasets from normal brains and show that by jointly analyzing multiple brain regions (tissues), we identify eQTL within more genes relative to a TBT analysis.

## 2.2 Results and Discussion

### *2.2.1 Methods overview*

For a given gene-SNP pair, our approach models gene expression across tissues using a linear mixed model in which both fixed and random effects are used to capture the effect of a variant on gene expression. Briefly, for each tissue $t$ and individual $i$ we model the potential genetic association between a target SNP and the expression levels of a target gene $j$ at a single locus by using the following vectorized form of the linear-mixed model (the $t$-variate normal law with mean $\mu \in \mathbb{R}^t$ and variance $\Sigma \in \mathbb{R}^{t\times t}$ will be denoted as $N_t(\mu, \Sigma)$) –

$$y_{ij} = \alpha_j + \mathbf{1}\beta_j g_i + \mathbf{1}u_i + g_i v_j + \xi_{ij} \qquad \xi_{ij} \overset{i.i.d.}{\sim} N_t(0, \epsilon\mathbb{I}) \tag{2.1}$$

where $y_{ij}$ is a $t \times 1$ vector of gene expression data, $\mathbb{I}$ denotes the corresponding $t \times t$ diagonal matrix, $\alpha \in \mathbb{R}^t$ is the fixed effect for the mRNA level for $t$ tissues, $\beta_j$ is the fixed effect for the SNP ($\beta_j \in \mathbb{R}^1$), $g_i$ is the value of a bi-allelic genotype such that $g_i \in (0, 1, 2)$, which represents the number of copies of the minor allele. $\mathbf{1}$ denotes a column vector of $t$ ones. The random effect $v_j \in \mathbb{R}^t$ represents tissue-specific interaction with the genotype and $u_i \in \mathbb{R}^1$ is a subject-specific random intercept. We assume that the random effects are

independent and that $v_j \sim N_t(0, \gamma\mathbb{I})$ and $u_i \sim N_1(0, \tau)$.

Since tissue-specific effects are modeled as random effects, a test of whether there are tissue-specific effects is equivalent to testing whether the variance of the random effect ($\gamma$) is zero. Thus our approach involves testing only two scalar parameters ($\beta$ and $\gamma$), regardless of the number of tissues being considered. We develop a score test of the null hypothesis that both of these parameters are zero, i.e., that the variant does not affect gene expression across any of the tissues. We present this model and the resulting score test in detail in the methods section.

### *2.2.2 Simulations*

We evaluate our approach through extensive simulation studies. We begin with a single locus and a single gene, of which the expression is measured across either 5 or 10 tissues. Genotypes are first generated assuming Hardy-Weinberg equilibrium and common minor allele frequency (> 5%). Given this genotype, gene expression is generated according to equation 1. Type I error is evaluated using 10,000 data replicates; 1,000 replicates are used for power calculations. Simulations under the null hypothesis confirm that our method has the correct type I error.

Table 2.1: Table comparing the type I error of the joint score test statistic, $U_\psi$ with tissue-by-tissue (TBT) analysis, MetaTissue (MT) model (FE = Fixed Effects model; RE = Random Effects model) and multivariate Bayesian Model Averaging . Note that all the results are based on 5,000 simulations on 100 observations at a nominal level of $\alpha = 0.05$.

| | **Number of tissues = 5** | | | | | **Number of tissues = 10** | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| MAF | TBT | $MT(FE)$ | MT(RE) | BMA | $U_\psi$ | TBT | $MT(FE)$ | MT(RE) | BMA | $U_\psi$ |
| 0.05 | 0.0422 | 0.0410 | 0.0488 | 0.0488 | 0.0456 | 0.0434 | 0.0362 | 0.0404 | 0.0392 | 0.0416 |
| 0.10 | 0.0476 | 0.0442 | 0.0510 | 0.0488 | 0.0494 | 0.0512 | 0.0368 | 0.0472 | 0.0448 | 0.0480 |

We compare the performance of our method with TBT, MT, and eQTL-BMA. Figure

2.2 shows that our method outperforms other methods in the presence of an additive genetic effect and tissue-specific interaction with the genotype ($PVE_\gamma$) when the number of tissues is 5 at a moderately rare allele frequency of 0.05. When the number of tissues is increased to 10, MT seems to outperform all other methods, including ours as seen in figure 2.3.

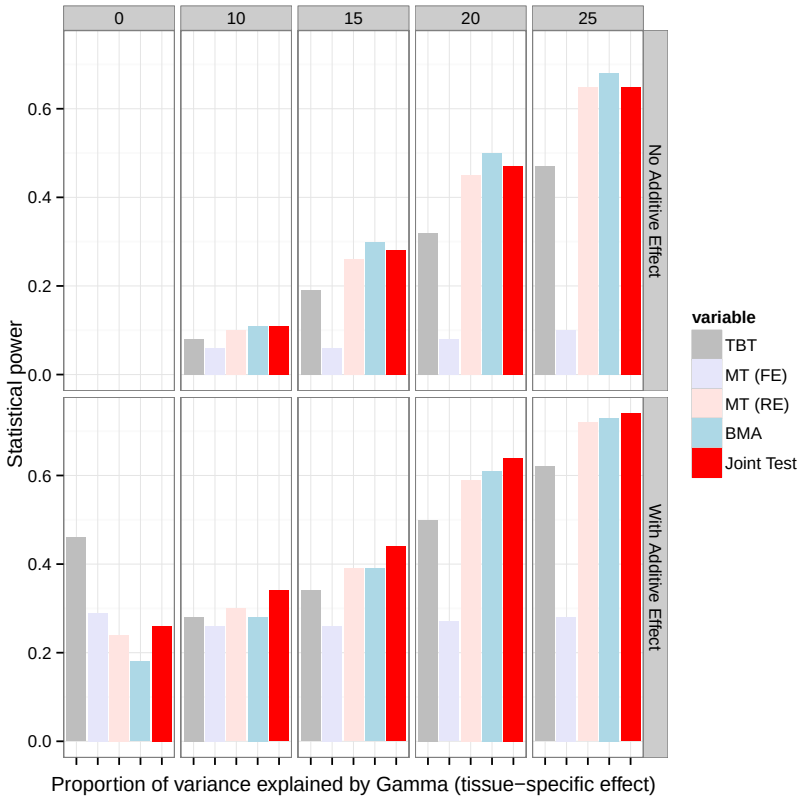



FIGURE 2.2: **Statistical power comparison at a minor allele frequency of 0.05 (moderately rare variant minor allele frequency) when the number of tissues is 5.** Barplot depicting the statistical power comparison between the joint score test method and other methods such as MetaTissue model (Fixed Effects, FE; Random Effects, RE) and eQTL-BMA when the number of tissues is 5 and 10 at a minor allele frequency of 0.05. We varied the proportion of variance explained by $\gamma$ ($PVE_\gamma$) between 0 - 25% and $\beta$ fixed effect for the additive effect of the SNP. Each vertical grid labeled 0 through 25 represents varying levels of $PVE_\gamma$ whereas each horizontal grid represents the presence or absence of the additive effect due to SNP.

On the other hand, at a more common variant frequency of 0.10 and when the number of tissues if 5 (See figure 2.4), our method outperforms eQTL-BMA and MT in the

presence of both additive genetic effect and $PVE_\gamma$. In the absence of any additive genetic effect, eQTL-BMA seems to work the best.

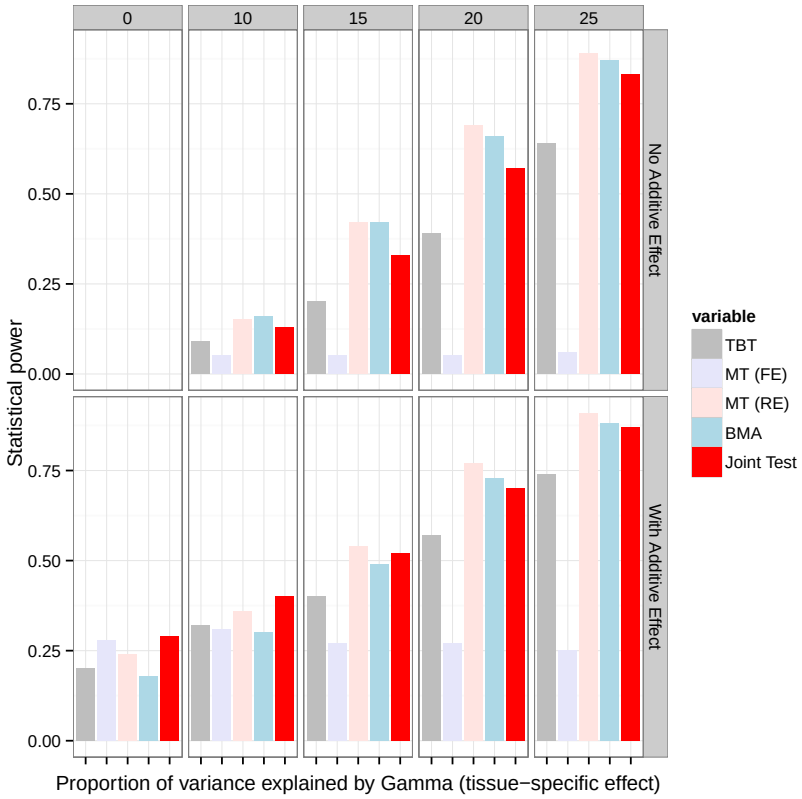



FIGURE 2.3: **Statistical power comparison at a minor allele frequency of 0.05 (moderately rare variant minor allele frequency) when the number of tissues is 10.** Barplot depicting the statistical power comparison between the joint score test method and other methods such as MetaTissue model (Fixed Effects, FE; Random Effects, RE) and eQTL-BMA when the number of tissues is 5 and 10 at a minor allele frequency of 0.05. We varied the proportion of variance explained by $\gamma$ $(PVE_\gamma)$ between 0 - 25% and $\beta$ fixed effect for the additive effect of the SNP. Each vertical grid labeled 0 through 25 represents varying levels of $PVE_\gamma$ whereas each horizontal grid represents the presence or absence of the additive effect due to SNP.

However, when the number of tissues is increased to 10 (See figure 2.5) in the presence of any additive genetic effect, our method is comparable to MT and eQTL-BMA and better than TBT.

The CPU time for the analyses performed on a simulated dataset are summarized in the table below. It is important to note that these times are reflective of the algorithm and

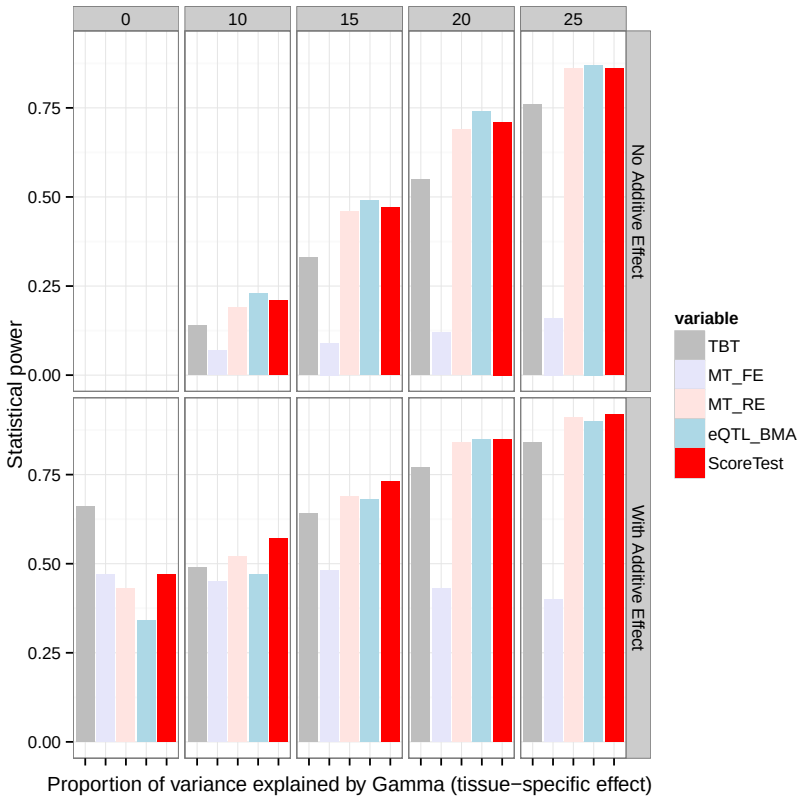



FIGURE 2.4: **Statistical power comparison at a minor allele frequency of 0.1 (common variant minor allele frequency) when the number of tissues is 5.** Barplot depicting the statistical power comparison between the joint score test method and other methods such as MetaTissue model (Fixed Effects, FE; Random Effects, RE) and eQTL-BMA when the number of tissues is 5 and 10 at a minor allele frequency of 0.10. We varied the proportion of variance explained by $\gamma$ $(PVE_\gamma)$ between 0 - 25% and $\beta$ fixed effect for the additive effect of the SNP. Each vertical grid labeled 0 through 25 represents varying levels of $PVE_\gamma$ whereas each horizontal grid represents the presence or absence of the additive effect due to SNP.

do not account for data pre-processing. It is clear from the table that our method is computationally faster than currently used multi-tissue eQTL methods, MT and eQTL-BMA. This computational efficiency is attributed to the existence of a closed-form solution to the distribution of our joint score test statistic, which can be written as a function of the number of genes and variants. It is important to note that the computational efficiency of MT and eQTL-BMA methods is estimated using publicly available software versions.

Table 2.2: Performance of different methods on a simulated dataset. All the computations were performed on a single core of an Intel Xeon E5-2650 2.60GHz CPU. These times do not include any data preparation time and are reflective of the core algorithm alone.

| **Method** | **Number of tissues = 5** | **Number of tissues = 10** | **Core algorithm implementation** |
|---|---|---|---|
| Joint score test | 0.48 s (with no permutations)<br>45 s (with permutations) | 0.7s (with no permutations)<br>72 s (with permutations) | RcppArmadillo |
| eQTL-BMA | 176 s | 244 s | C++ |
| MetaTissue | 157 s | 822 s | Java |

We demonstrate the effectiveness of our approach by applying it to two datasets in which gene expression data, measured across various regions in normal brains, is paired with genome-wide single nucleotide polymorphism data. These datasets have previously been analyzed using a region-by-region (i.e., tissue-by-tissue) approach. We consider two different types of analyses: One that focuses on SNPs that lie within 100 kilobase up- and down-stream of the transcription start site of a gene (referred to here as the *cis* candidate region), and another that focuses on a genome-wide analysis. Due to the computational burden of genome-wide analyses using eQTL-BMA and MT methods, we only apply our joint score test and the TBT approach and assessed their performance by comparing the total number of genome-wide gene-SNP pairs deemed statistically significant at a Bonferroni threshold.

### *2.2.3 Region-specific analysis of normal adult human brains*

Adult human brains have distinct expression patterns across each brain region (Ramaswamy et al., 2014), and understanding the genetic control of gene expression across the brain regions may further our understanding of brain diseases by identifying possible disease susceptibility regions. We hypothesize that our approach will identify more genes with eQTL than were previously identified using a TBT approach. The first brain dataset, originally analyzed by Gibbs *et al.*(Gibbs et al., 2010), consists of four brain regions (cerebellum, CRBLM; frontal cortex, FCTX; pons, PONS; temporal cortex, TCTX) from 150

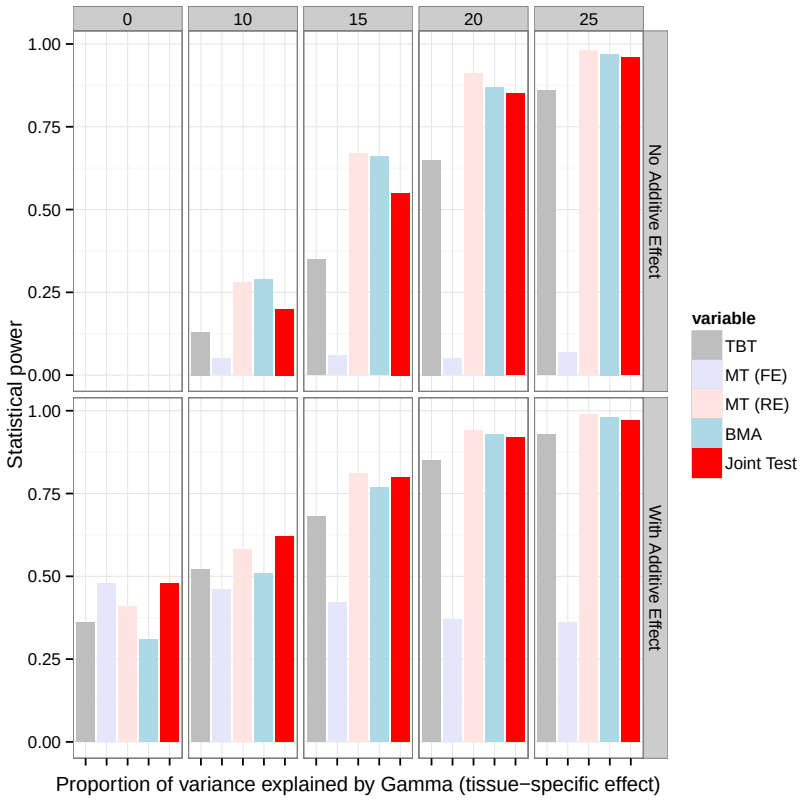



FIGURE 2.5: **Statistical power comparison at a minor allele frequency of 0.1 (common variant minor allele frequency) when the number of tissues is 10.** Barplot depicting the statistical power comparison between the joint score test method and other methods such as MetaTissue model (Fixed Effects, FE; Random Effects, RE) and eQTL-BMA when the number of tissues is 5 and 10 at a minor allele frequency of 0.10. We varied the proportion of variance explained by $\gamma$ ($PVE_\gamma$) between 0 - 25% and $\beta$ fixed effect for the additive effect of the SNP. Each vertical grid labeled 0 through 25 represents varying levels of $PVE_\gamma$ whereas each horizontal grid represents the presence or absence of the additive effect due to SNP.

neuropathologically normal patients. Gene expression and genotype data were assayed on Illumina HumanRef-8 Expression BeadChips and Infinium HumanHap550 Beadchips, respectively. Genotype data was preprocessed to remove rare variants (minor allele frequency less than 0.05) and the population being analyzed is homogeneous with respect to patient ethnicity. After standard genotype and gene-expression preprocessing and quality control procedures (see methods), the resulting dataset consisted of 400,973 SNPs and 18,983 genes. We considered two different analyses - a *cis* analysis, which effectively re-

stricts our analysis to only *cis*-SNPs that are 100 kilobase pairs up- and down-stream of the transcription start site, and a genome-wide analysis. A *cis* analysis reduces the search space of potential gene-SNP pairs to 511,458, thus reducing the total number of hypotheses being tested. A region-by-region *cis* analysis of Gibbs *et al* neuropathologically normal human brain data yielded the following statistically significant results (i.e., passed the Bonferroni multiple testing threshold of $2.44 \times 10^{-8} = \frac{0.05}{\text{\# tissues} \times 511{,}458}$): 1,547 gene-*cis*SNP pairs in CRBLM, 1,609 gene-*cis*SNP pairs in FCTX, 1,148 gene-*cis*SNP pairs in PONS and 1,341 gene-*cis*SNP pairs in TCTX. After 10,000 permutations, a region-by-region analysis yielded 2,367 genes with at least one *cis*-eQTL while our approach identifies 3,913 genes with at least one *cis*-eQTL or approximately 65% more genes. Of note, approximately 98% of these genes are present in the list of genes identified by the TBT approach. For comparison, while eQTL-BMA approach identifies 2,919 genes (23% more than the TBT approach with a 73% gene-overlap), MT method identifies 3,743 genes (58% more than the TBT approach with a 74% gene-overlap) using its fixed-effects model and 3,843 genes (62% more than the TBT approach with a 79% gene-overlap)) using its random effects model. A genome-wide TBT analysis of the same data yielded the following statistically significant results (i.e., passed the Bonferroni multiple testing threshold of $1.64 \times 10^{-12} = \frac{0.05}{\text{\# tissues} \times \text{\# genes} \times \text{\# SNPs}}$): 716 gene-SNP pairs in CRBLM, 779 gene-SNP pairs in FCTX, 473 gene-SNP pairs in PONS and 630 gene-SNP pairs in TCTX; 245 gene-SNP pairs are shared among all the brain regions with a total of 1,277 gene-SNP pairs being unique among all the regions of the brain. The smaller numbers are attributed to the more conservative Bonferroni threshold due to increased number of hypotheses tested. In contrast, our score test approach significantly (Bonferroni threshold = $6.57 \times 10^{-12} = \frac{0.05}{\text{\# genes} \times \text{\# SNPs}}$) implicates 2,602 unique gene-SNP pairs more than twice the number identified by the TBT approach.

The second brain dataset, originally analyzed by Ramaswamy *et al*(Ramaswamy et al.,

2014), consists of ten brain regions (cerebellum, CRBLM; frontal cortex, FCTX; hippocampus, HIPP; medulla, MEDU; occipital cortex, OCTX; putamen, PUTM; substantia nigra, SNIG; temporal cortex, TCTX; thalamus, THAL; intralobular white matter, WHMT) from 134 neuropathologically normal patients. The gene expression data was assayed on Affymetrix Human Exon 1.0 ST Array, which has both exon-level and gene-level gene expression data. In order to simplify our analysis, we made use of gene-level expression data where the expression levels of each exon for each gene/transcript were aggregated using the Winsorized mean (Wilcox and Keselman, 2003a). Genotype data were assayed on Illumina Infinium HumanHap550 v3 and were subjected to standard quality control preprocessing (see methods). This dataset contains many samples with missing gene expression values however, the missing data does not affect parameter estimation or inference under our method. In fact, the likelihood of the observed data has the same form as that of the missing data (see Appendix 1). After standard genotype and gene-expression preprocessing and quality control procedures (see methods), there were 627,126 SNPs and 25,501 genes to be analyzed. *cis*-eQTL analysis of this data yielded 2,714 genes with at least one eQTL in a TBT analysis while our approach yielded 5,413 genes with at least one eQTL with a 94% gene-overlap. For comparison, while eQTL-BMA approach identifies 5,316 genes, MT method identifies 4,984 and 5,176 genes with eQTL using its fixed-effects model and its random effects model, respectively. A genome-wide TBT analysis of the same data yielded 6,698 unique gene-SNP pairs after adjusting for multiple hypotheses (Bonferroni threshold = $3.13 \times 10^{-13} = \frac{0.05}{\#\text{ tissues} \times \#\text{ genes} \times \#\text{ SNPs}}$). As was observed with Gibbs *et al.* data, our approach again identifies substantially more eQTLs, significantly implicating 10,392 unique gene-SNP pairs.

## 2.3 Materials and Methods

### *2.3.1 Efficient score functions for $\beta$ and $\gamma$*

We begin with a linear mixed effects model that models expression patterns across tissues as a function of genotype. In a matrix notation, for a given gene-SNP pair

$$Y = J\alpha + G\beta + Zu + Xv + \xi \tag{2.2}$$

where $Y$ is a $nt$-dimensional matrix of expression levels in $t$ tissues and $n$ individuals, $\alpha$ is a fixed effect representing the tissue-specific intercepts, $G$ is a $nt$-dimensional matrix of genotypes, $\beta$ is a fixed effect of genotype across tissue, $u \sim N(0, \tau ZZ^T)$ is a $nt$-dimensional matrix of subject-specific random effect, $v \sim N(0, \gamma XX^T)$ is a $nt$-dimensional matrix of tissue-specific random effects, and $\xi \sim N(0, \epsilon I_{nt})$ and $I$ is the identity matrix . The matrices $J$, $Z$ and $X$ are design matrices with $X$ being a function of genotype. $J$ is $nt \times t$ dimensional matrix denoting the design matrix for the tissue-specific intercepts. $Z$ is $nt \times nt$ design matrix for the subject-specific intercepts. $X$ is a $nt \times t$ design matrix of stacked genotypes. The parameters of interest are $\beta$ and $\gamma$; $\alpha$, $\tau$ and $\epsilon$ are nuisance parameters.

We test the null hypothesis that $H_0 : \beta = \gamma = 0$, i.e. the variant does not affect gene expression across any of the tissues. To do so, we compute the efficient scores for $\beta$ and $\gamma$ by projecting off components correlated with the nuisance parameters. From equation 1, the log-likelihood function of Y conditional on the genotype is –

$$\ell(\beta, \theta) = c - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(Y - J\alpha - G\beta)^T \Sigma^{-1} (Y - J\alpha - G\beta)$$

where $\theta$ represents the vector of all the variance components involved in $\Sigma$ and c is a constant. Alternatively, under equation 1 and normality, we have

$$Y \sim N(J\alpha + G\beta, \Sigma) \quad with \quad \Sigma = \epsilon I + \tau ZZ^T + \gamma XX^T$$

The efficient scores evaluated under the null are given by –

$$U_\beta = (G - \bar{G})^T \hat{\Sigma}_n^{-1} (Y - J\hat{\alpha}) \tag{2.3}$$

and

$$U_\gamma = \frac{1}{2} (Y - J\hat{\alpha})^T \hat{\Sigma}_n^{-1} XX^T \hat{\Sigma}_n^{-1} (Y - J\hat{\alpha}) \tag{2.4}$$

where $\hat{\Sigma} = \hat{\tau} ZZ^T + \hat{\epsilon} I$ and $\hat{\tau}$ along with $\hat{\epsilon}$ are the maximum likelihood estimators of $\tau$ and $\epsilon$ under the null.

Following Huang et al (Huang et al., 2014), we propose a weighted sum of $U_\beta$ and $U_\gamma$ to arrive at our joint score test statistic, $U_\psi$. Since $U_\beta$ is linear in $Y$ while $U_\gamma$ is quadratic, we propose the following rule to combine them –

$$\begin{aligned} U_\psi &\equiv a_\beta U_\beta^2 + a_\gamma U_\gamma \\ &= (Y - J\hat{\alpha})^T \hat{\Sigma}_n^{-1} \left[ a_\beta (G - \bar{G}) (G - \bar{G})^T + a_\gamma \left( \frac{1}{2} XX^T \right) \right] \hat{\Sigma}_n^{-1} (Y - J\hat{\alpha}), \end{aligned} \tag{2.5}$$

where $a_\beta$ and $a_\gamma$ are scalar constants chosen to minimize the variance of $U_\psi$ (see Appendix 1 for details). Under the null, $U_\psi$ is distributed as a mixture of chi-square random variables. Several approximation and exact methods were proposed to obtain the distribution of $U_\psi$ (Duchesne and Lafaye De Micheaux, 2010). Here, we use the Satterthwaite method (Satterthwaite, 1946) to approximate the $p$ values from a scaled $\chi^2$ distribution by matching the first two moments as $U_\psi \sim \kappa \chi_\nu^2$ where $\kappa = \frac{Var(U_\psi)}{2E[U_\psi]}$ and $\nu = \frac{2E[U_\psi]^2}{Var(U_\psi)}$.

### *2.3.2 Simulation studies*

Each simulated dataset was comprised of data from a single locus and a single gene, whose expression is measured across either 5 or 10 tissues. The data are generated prospectively, i.e., first genotypes are generated (assuming Hardy-Weinberg equilibrium and > 5% minor allele frequency), then, given genotype, gene expression is generated according to equation 1 of methods. We use 10,000 data replicates when evaluating type I error and 1,000 for

power calculations. Simulations were performed by varying two parameters, $\beta$ (additive genetic effect) and the proportion of variation explained by the random effect of genotype ($PVE_\gamma \equiv \frac{\gamma}{\tau+\gamma+\epsilon}$). Additive genetic effect in the simulations was controlled by varying $\beta$ between 0 (indicates the absence of additive genetic effect) and 0.5 (indicates the presence of additive genetic effect), and the tissue-specific interaction effect was controlled by varying $\gamma$ between 0 (indicating no tissue-specific effects) and 25%. A linear mixed effects model was fit using the package *lme4*(Bates et al., 2014a,b) in the statistical environment R (R Core Team). MT and eQTL-BMA methods were run on the simulated datasets as per their respective software instructions. We picked the default option for calculating the Bayes Factors and performed joint analysis with permutations while using eQTL-BMA method. The statistic computed by the eQTL-BMA approach is given a frequentist interpretation by translating the test statistic (computed by the eQTL-BMA model) into a $p$ value for each gene by comparing the observed values with simulated values obtained under the null after permuting the sample labels. The significance of an association between a gene-SNP pair in a TBT analysis is assessed by the $p$ value obtained using *lm* function in R. The test statistic is the minimum $p$ value over the total number of tissues from linear regressions performed separately in each tissue for each gene-SNP pair. Statistical significance was determined at a nominal $p$ value of 0.05 for all power simulations (in case of TBT analysis, it is $\frac{0.05}{k}$ where $k$ is the number of tissues). In order to assess the computing times of various algorithms, we performed a series of simulations as noted in Flutre *et al* (Flutre et al., 2013), with five or ten tissues measured in 100 unrelated individuals. Each simulation consisted of 3,705 gene-SNP pairs, at least half of which were "null" (i.e. SNP was not an eQTL in any tissue) and the other half following an alternative hypothesis that the SNP was an eQTL in at least $k$ tissues with $k$ varying from 1 to 5 or 10. The genotypes at each SNP in each individual were simulated with minor allele frequency 30% and assuming Hardy-Weinberg equilibrium. Gene expression data was generated for 100 genes and 1,036 SNPs (in *cis* with at least one gene) as was explained in Flutre *et al*.

### 2.3.3 *Preprocessing Gibbs J.R.* et al *normal brain data*

Gene expression on four brain regions including cerebellum (CRBLM), frontal cortex (FCTX), caudal pons (PONS) and temporal cortex (TCTX) and SNP datasets are publicly available (Gene Expression Omnibus (GEO) Accession Number: **GSE15745**; dbGAP Study Accession: **phs000249.v1.p1**). Genotyping was done on Infinium HumanHap550 Beadchips to assay genotypes for 561,466 SNPs, from the cerebellum tissue samples while gene expression profiling of 22,184 mRNA transcripts was performed using Illumina HumanRef-8 Expression BeadChips. The genotype data was recoded into a SNP matrix of values 0, 1 and 2 representing minor allele counts under the additive model. Samples with African (GSM394931 in CRBLM, GSM395081 in FCTX, GSM395226 in PONS and GSM395374 in TCTX) and Asian (GSM394121 in CRBLM, GSM394263 in FCTX, GSM394405 in PONS and GSM394566 in TCTX) ancestry were removed from the analysis. These SNPs were filtered on the missing*ness* of the individual data (excluded samples with more than 10% missing genotypes) and the SNP data (excluded SNPs with missing values), followed by MAF (included SNPs with MAF $\geq 0.05$) and Hardy-Weinberg equilibrium (HWE; p-values $< 0.001$) in the same order using PLINK (Purcell et al., 2007) software. Top principal components on the filtered and pruned genotype data based on linkage disequilibrium measurements (window size of 1500, sliding window 150 SNPs at a time and an $r^2$ threshold of 0.04) were generated using EIGENSTRAT (Price et al., 2006) method for later use to correct for population stratification. Each gene expression probe was adjusted for the biological and methodological covariates such as tissue bank, gender, hybridization batch and numeric covariates such as post-mortem interval (PMI) and age in order to remove any associated confounding effects.

### 2.3.4 *Preprocessing Ramaswamy* et al *normal brain data (BrainEAC consortium study)*

Gene expression data from 10 brain regions including cerebellar cortex (CRBLM), frontal cortex (FCTX), hippocampus (HIPP), inferior olivary nucleus/medulla (MEDU), occipital cortex (OCTX), putamen (PUTM), substantia nigra (SNIG), temporal cortex (TCTX), thalamus (THAL) and intralobular white matter (WHMT) was obtained from GEO under the accession id **GSE46706**. The authors have kindly provided us with the SNP data. This data is part of the UK Brain Expression Consortium (UKBEC) and the brain samples were collected by the Medical Research Council Sudden Death Brain and Tissue Bank, Edinburgh, UK, and the Sun Health Research Institute an affiliate of Sun Health Corporation, USA. Exon-specific RNA expression was quantified using Affymetrix Human Exon 1.0 ST arrays and the genotyping was done on Illumina Omni1-quad and Immunochip arrays. We followed the same steps to preprocess the SNP data. Preprocessed gene-level expression profile was obtained as a 'Series Matrix File' from GEO where the gene-level expression data for every gene is aggregated over all the probes representing it. Gene-level summary signals were then generated by calculating the Winsorized mean of expression values of all probe sets annotated to a transcript. There are a total of 25,501 genes represented on this microarray platform.

It is important to note that our data preprocessing methods for both the aforementioned datasets are different from the original methods used by the authors in their respective publications.

### 2.3.5 *Data analysis*

We performed two different types of data analyses - 1) one that focuses on *cis* candidate regions that are defined by the proximity of an eQTL to the transcription start site of a gene not exceeding 100 kilobase up- and down-stream of the transcription start site of a gene,

and 2) a genome-wide analysis, which tests all gene-SNP pairs in a given eQTL dataset.

The performance of all the methods for the *cis* analysis is assessed by comparing the number of genes identified as having at least one eQTL in any given tissue at a 5% false discovery rate (FDR). eQTL-BMA computes a test statistic for all genes as an average of all Bayes Factors for the given gene and its *cis*-SNP. This test statistic is then converted to a $p$ value for each gene using an adaptive permutation-resampling technique performed on each gene separately, which compares the observed test statistic with the value of the test statistic obtained from repeated permutations (10,000 in this case). The MT model was run on the same set of *cis* SNPs for all the genes and the resulting $p$ values using both fixed and random effects model were adjusted using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). Since the MT model outputs $p$ values for each gene-SNP pair being tested, we identified the genes with at least one eQTL by finding the minimum adjusted $p$ value for a given gene. In the case of a TBT analysis, which tests for the presence of an eQTL in a single tissue, the test statistic is the minimum $p$ value of the linear regressions between a gene and a candidate *cis*-SNP in that tissue. While applying our joint score test approach for every gene, we computed the minimum $p$ value across all the candidate *cis*-SNPs and the adjusted $p$ value for that gene was computed as the average number of times the permuted $p$ values are smaller than the observed minimum $p$ value.

Genome-wide analysis of the brain datasets was performed only using our joint score test and TBT approaches due to the computational burden such analyses create on MT and eQTL-BMA methods. The performance of both TBT and our method for the genome-wide analysis of the datasets was evaluated at Bonferroni thresholds by comparing the number of significant gene-SNP pairs. A generic TBT approach on the normal human brain datasets was implemented as a linear regression model in the R-package MatrixEQTL (Shabalin, 2012).

## 2.4 Conclusion

Our method investigates the presence of two types of genetic effects: 1) an overall shift in gene expression due to genotype across all tissues, and 2) tissue-specific effects of genotypes on gene expression. Our approach models tissue-specific effects as random effects, resulting in a test that involves far fewer parameters. Further, our simulations demonstrate improved power over previously proposed methods for eQTL analysis.

Both of the datasets examined here used genome-wide association study (GWAS) SNP array platforms to interrogate germ-line variation. By design, the SNPs on these platforms are overwhelmingly common, and as a result, individual SNP based analyses will have reasonable power to detect association with gene expression. However, as eQTL studies transition from GWAS platforms to whole genome sequencing as the primary approach to assaying genetic variation, rare variation will also need to be considered. In this setting, an individual-variant analysis is no longer viable. Thus the dominant paradigm for rare variant analyses takes a gene-based or regional approach, accumulating information across a gene or other genetic unit. Score statistics are often used in this context and have been leveraged to form both burden- (Liu et al., 2014) and kernel-based(Wu et al., 2011) tests. The score test based framework presented here could be used in a similar way to develop tests that accumulates rare variant contributions across genomic regions that are annotated to have regulatory potential.

Another important issue in the context of eQTL studies is the vast number of hypotheses being tested. Because of this, it is important that multiple testing corrections are used to appropriately control for either family-wise error rate (FWER) or false discovery rate (FDR). Bonferroni (FWER) and Bejamini and Hochberg (FDR) adjustments are simple approaches to maintaining such control but they can be overly conservative

when there is substantial correlation among the tests, which is likely in eQTL studies. Permutation based adjustment can address this limitation but may be computationally prohibitive. Resampling-based (Huang et al., 2014) and Monte Carlo (Lin, 2005) approaches have been proposed that allow the characterization of the permutation distribution of the statistics without the computational demands required of permutation. Investigating how these approaches can be adapted to the joint score test presented here is a topic for future research.

We are currently considering a couple of ways our model can be extended to incorporate additional data types. First, our approach could accommodate the analysis of RNA-Seq by modeling gene transcripts in an analogous fashion to tissues in our current formulation. Thus, one would be able to test for both a variant's overall effect across all isoforms of a gene as well as transcript specific effects. Second, since DNA methylation can impact gene expression patterns in a tissue specific manner, we are considering an extension of the model that explicitly incorporates methylation in modeling the effect germ-line variation has on gene expression patterns.

# 3

# Mapping eQTL by leveraging multiple tissues and DNA methylation

## 3.1 Background

It has been long established that regulatory regions in higher eukaryotes activate gene transcription in a tissue-specific manner (Ong and Corces, 2011; Geyer et al., 1990). These regulatory regions, which affect the binding affinities of transcription factors, are susceptible to both genetic variation and epigenetic modifications that play a coordinated role in regulating tissue-specific gene expression (Bell et al., 2011; Gibbs et al., 2010; Shoemaker et al., 2010; Wrzodek et al., 2012; Lemire et al., 2014). One form of epigenetic variation is DNA methylation that targets nonmethylated and noncoding GC-rich and CpG-rich regions of the DNA sequence, which constitute approximately 70% of all annotated promoters (Deaton and Bird, 2011). DNA methylation is linked to transcriptional silencing, and many CpG island promoters are active in a tissue-specific manner. Previous studies have shown that inter-individual variation in DNA methylation at distinct CpG sites has been consistently linked to genetic variation such as single nucleotide polymorphisms (SNPs), known as methylation eQTLs (mQTLs) (Wagner et al., 2014; Hellman and Chess,

2010; Gutierrez-Arcelus et al., 2015). Since an increased DNA methylation at any of the distinct CpG sites located in the promoter regions necessitate chromatin remodeling and subsequent decrease in gene expression, any DNA sequence variation within the CpG-rich regions that disrupts the methylation process may have an opposite effect on gene expression.

Even though, mechanisms which regulate DNA methylation are unclear, it is clear that there is some association between genetic variation and quantitative changes in methylation levels (Banovich et al., 2014). For example, Catechol-O-methyltransferase (COMT) gene, which is implicated in schizophrenia has a SNP, $Val^{158}Met$ (rs4680) that is associated with differential COMT expression across regions of the brain during the course of the illness (Swift-Scanlan et al., 2014). More specifically, the substitution of a methionine (Met) for a valine (Val) at position 158 results in reduced activity of the COMT enzyme due to reduced protein stability. Methylation of CpG islands associated with the aforementioned variant affect the region-specific expression of COMT (Swift-Scanlan et al., 2014). Identifying and studying the mechanisms through which genetic variation, DNA methylation and gene expression interact may provide us yet another clue to understanding regions within the genome that are associated with complex disease phenotypes (Figure 3.1).

Current approaches to delineate the role played by both genetic and epigenetic variation in gene expression are limited to identifying statistically significant pairs of mRNA - SNPs and CpG - SNPs by performing independent eQTL and mQTL analyses, respectively, within a tissue-by-tissue (TBT) framework (Gibbs et al., 2010; Gutierrez-Arcelus et al., 2015, 2013). These pairs are then expanded to combinations of mRNA transcript, CpG site and a SNP wherever the SNP was significantly correlated with either mRNA or CpG site of the mRNA - CpG pair. First, any such TBT analyses have been shown to fall short in fully exploiting patterns across the tissues thus impacting eQTL or mQTL discovery (Flutre et al., 2013; Sul et al., 2013; Acharya et al., 2016). Second, independent eQTL and mQTL analyses do not reveal any underlying effects of genetic variation

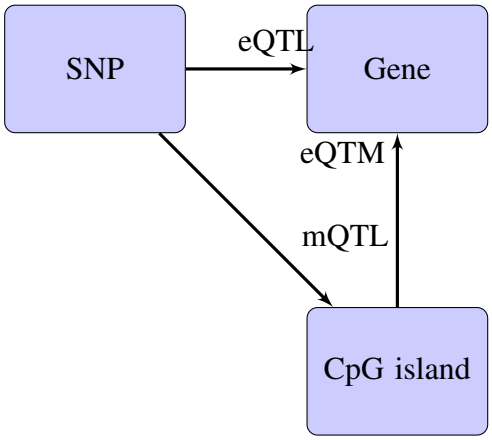



FIGURE 3.1: Tissue-specific gene expression is controlled by genetic, epigenetic and transcriptional regulatory mechanisms. Genetic control of gene expression can be defined in terms of SNPs and their associations with gene expression. Expression quantitative trait loci (eQTL) are correlations between SNPs and gene expression. Similarly, epigenetic control of gene expression can be defined in terms of CpG island methylation and their interactions with SNPs. Expression quantitative trait methylations (eQTM) are correlations between gene expression and methylation while methylation quantitative trait loci (mQTL) are correlations between SNPs and methylation. Gene expression can be modeled as a function of SNP and CpG islands.

on tissue-specific gene expression due to DNA methylation. Consequently, we propose to map eQTLs by leveraging DNA methylation and testing for any higher order interactions among methylation, genotype and tissues. We have previously proposed a score test-based approach to map multi-tissue eQTLs where we model tissue-specificity as a random effect and investigated an overall shift in the gene expression combined with tissue-specific effects due to genetic variants (Acharya et al., 2016). We extend this framework to include methylation-specific effects and model the combined effect of genetic and epigenetic variation on gene expression.

We show using Monte Carlo simulations that our joint score test is more powerful in teasing out eQTLs by controlling for methylation than any TBT approach that uses methylation as a covariate (TBTm-eQTL). We also show that the new joint score test is better at identifying eQTLs in the presence of DNA methylation than our previously proposed multi-tissue eQTL and TBT methods. Finally, we show that in cases where the interaction effects of DNA methylation are absent, our approach remains competetive.

We demonstrate the effectiveness of our method by applying it to a publicly available expression, methylation and SNP array datasets from adult normal brains (Gibbs et al., 2010) and show that by jointly analyzing multiple brain regions (tissues), we identify eQTL that may otherwise be not identified by multi-tissue eQTL methods.

## 3.2 Methods

### *3.2.1 Our model*

For a given mRNA transcript, tissue-specific gene expression is modeled as a function of genotype and methylation –

$$Y = J\alpha + G\beta + M\lambda + MG\phi + Au + Bv + Cw + Dx + \xi \tag{3.1}$$

where $Y$ is $nt$-dimensional vector of expression levels in $t$ tissues and $n$ individuals, $\alpha$ is a vector of tissue-specific intercepts, $G$ is $nt$-dimensional vector of genotypes, $\beta$ is a fixed effect of genotype across tissue, $M$ is $nt$-dimensional vector of methylation levels, $\lambda$ is an overall methylation-specific fixed effect, $MG$ is $nt$-dimensional vector of the product of methylation and genotype, $\phi$ is the regression coefficient for genotype and methylation interaction (fixed effect), $u \sim N(0, \tau AA^T)$ is a vector of subject-specific random effect, $v \sim N(0, \gamma BB^T)$ is a vector of tissue-specific random effects, $w \sim N(0, \delta CC^T)$ is a vector of tissue-specific random effects that describe the interaction effect between genotype and methylation is a vector of random effects describing the interaction between genotype, methylation and tissue, $x \sim N(0, \theta DD^T)$ is a vector of tissue-specific random effects describing methylation effects and $\xi \sim N(0, \epsilon I_{nt})$. The matrices $J$, $A$, $B$, $C$, and $D$ are design matrices with $B$ being a function of genotype, $C$ is a function of both genotype and methylation data and finally, $D$ is a function of just the methylation data. $J$ is $nt \times t$ dimensional matrix denoting the design matrix for the tissue-specific intercepts. $A$ is $nt \times nt$ design matrix for the subject-specific intercepts. $B$ is a $nt \times t$ design matrix of stacked genotypes. $C$ is a $nt \times t$ design matrix of stacked (product of) tissue-specific methylation and genotype

data. $D$ is $nt \times t$ design matrices of stacked tissue-specific methylation data. The parameters of interest are $\gamma$, $\delta$, $\beta$ and $\phi$; $\alpha$, $\lambda$, $\tau$, $\theta$ and $\epsilon$ are nuisance parameters. Alternatively, we can represent the distribution of $Y$ conditional on methylation and genotype as –

$$(Y|M = m, G = g) \sim N\left(J\alpha + G\beta + M\lambda + MG\phi, \Sigma\right)$$

From our model, the log-likelihood function of the parameters conditional on the genotype and methylation data is given by–

$$\begin{aligned} \ell\left(\Theta; Y|M = m, G = g\right) &= -c - \frac{1}{2}\log \mid \Sigma \mid \\ &-\frac{1}{2}\left(Y - J\alpha - G\beta - M\lambda - MG\phi\right)^T \Sigma^{-1}\left(Y - J\alpha - G\beta - M\lambda - MG\phi\right) \end{aligned} \tag{3.2}$$

where $\Theta$ represents the vector of all the variance components involved in $\Sigma$ and $c$ is a constant. We test the null hypothesis that $H_0 : \beta = \phi = \gamma = \delta = 0$, i.e. the variant does not affect gene expression across any of the tissues. To do so, we compute the efficient scores for $\gamma$, $\delta$, $\beta$ and $\phi$ by projecting off components correlated with the nuisance parameters. The reduced model under the null is –

$$Y_{H_0} = J\alpha + M\lambda + Au + Dx + \xi$$

The efficient scores evaluated under the null are given by –

$$\text{Additive Genetic Effect} \coloneqq U_{\beta|H_0} = \hat{Y}^T\hat{\Sigma}_n^{-1}\left(G - \bar{G}\right)$$

$$G \times M \text{ Effect} \coloneqq U_{\phi|H_0} = \hat{Y}^T\hat{\Sigma}_n^{-1}\left(MG - \overline{MG}\right)$$

$$G \times T \text{ Effect} \coloneqq U_{\gamma|H_0} = \frac{1}{2}\hat{Y}^T\hat{\Sigma}_n^{-1}BB^T\hat{\Sigma}_n^{-1}\hat{Y}$$

$$G \times M \times T \text{ Effect} \coloneqq U_{\delta|H_0} = \frac{1}{2}\hat{Y}^T\hat{\Sigma}_n^{-1}CC^T\hat{\Sigma}_n^{-1}\hat{Y}$$

where $\hat{Y}$ are the residuals from the model, $\bar{G}$ is an $nt$-dimensional vector of mean-centered genotypes, $\overline{MG}$ is an $nt$-dimensional vector of mean-centered product of genotypes and methylation, and $\hat{\Sigma} = \hat{\epsilon}I + \hat{\tau}ZZ^T + \hat{\theta}DD^T$. Our joint score test will test for the effect of genotype on 1) an overall shift in the gene expression, 2) tissue-specific interaction ($G \times T$), 3) overall methylation ($G \times M$), and 4) tissue-specific methylation ($G \times M \times T$). More on the individual components of our score test can be found in Appendix 2.

We propose a weighted sum of the above components (under the null) to arrive at our joint score test statistic, $U_\zeta$. Since $U_\beta$ and $U_\phi$ are linear in $Y$ while $U_\gamma$ and $U_\delta$ are quadratic, we propose the following rule to combine them –

$$\begin{aligned} U_\zeta &\equiv \left(\boldsymbol{a_\beta} U_\beta^2 + \boldsymbol{a_\phi} U_\phi^2 + \boldsymbol{a_\gamma} U_\gamma + \boldsymbol{a_\delta} U_\delta\right) \\ &\equiv \hat{Y}^T \hat{\Sigma}_n^{-1} \left[ a_\beta \left(G - \bar{G}\right)\left(G - \bar{G}\right)^T + a_\phi \left(MG - \overline{MG}\right)\left(MG - \overline{MG}\right)^T + a_\gamma \frac{1}{2} BB^T + a_\delta \frac{1}{2} CC^T \right] \hat{\Sigma}_n^{-1} \hat{Y} \end{aligned} \tag{3.3}$$

where $a_\beta$, $a_\phi$, $a_\gamma$ and $a_\delta$ are scalar constants chosen to minimize the variance of $U_\zeta$. Under the null, $U_\zeta$ is distributed as a mixture of chi-square random variables. We use Satterthwaite method (Satterthwaite, 1946) to approximate the $p$ values from a scaled $\chi^2$ distribution by matching the first two moments as $U_\zeta \sim \kappa \chi_\nu^2$ where $\kappa = \frac{2Var(U_\zeta)}{E[U_\zeta]}$ and $\nu = \frac{2E[U_\zeta]^2}{Var(U_\zeta)}$.

### *3.2.2 Simulations*

For a positive integer $t$ that represents number of tissues, if $\mathbf{1}$ denotes a column vector of $t$ ones and $\mathbb{I}$ denotes the corresponding $t \times t$ diagonal matrix, following the $t$-variate normal law denoted by $N_t[\mu, \Sigma]$ with mean $\mu \in \mathbb{R}^t$ and variance $\Sigma \in \mathbb{R}^{t \times t}$, expression levels of a target gene $j$ at a single locus by using the following vectorized form of the linear mixed model –

$$y_{ij} = \alpha_j + \mathbf{1}\beta_j g_i + \mathbf{1}\lambda_j m_{ij} + \mathbf{1}\phi_j m_{ij} g_i + \mathbf{1}a_i + b_j g_i + c_j m_{ij} g_i + d_j m_{ij} + \xi_{ij} \qquad \xi_{ij} \overset{i.i.d.}{\sim} N(0, \epsilon\mathbb{I}) \tag{3.4}$$

where $y_{ij}$ is a $t \times 1$ vector of gene expression data, $\alpha_t$ is the tissue-specific intercept ($\alpha_t \in \mathbb{R}^t$), $\beta_j$ describes the main additive genotypic effect ($\beta_j \in \mathbb{R}^1$), $\lambda_j$ describes the overall effect due to methylation ($\lambda_j \in \mathbb{R}^1$), $\phi$ describes the interaction effect between the overall methylation and genotype ($\phi_j \in \mathbb{R}^1$), $g_i$ is the value of a bi-allelic genotype such that $g \in (0, 1, 2)$ represents the number of copies of the minor allele. The random effect $b_j \in \mathbb{R}^t$ represents tissue-specific effect of the genotype, $c_j \in \mathbb{R}^t$ represents tissue-specific interaction effect between methylation and genotype, $d_j \in \mathbb{R}^t$ represents tissue-specific methylation effect, and $a_i \in \mathbb{R}^1$ is a subject-specific random intercept. We assume that all the random effects are independent and that $a_i \sim N_1(0, \tau)$, $b_j \sim N_t(0, \gamma\mathbb{I})$, $c_j \sim N_t(0, \delta\mathbb{I})$ and $d_j \sim N_t(0, \theta\mathbb{I})$. Methylation data for 5 tissues was generated independently from a multivariate normal distribution with mean zero and positive definite variance-covariance matrix.

We use 1,000 data replicates to evaluate the type I error and for power calculations. Simulations were performed by varying the following parameters- $\beta$ (additive genetic effect), $\phi$ ($G{\times}M$ effect), the proportion of variation explained by the $G{\times}T$ effect $\left(PVE_\gamma \equiv \left(\frac{\gamma}{\theta+\tau+\epsilon+\gamma+\delta}\right)\right)$ and the proportion of variation explained by the $G \times M \times T$ effect $\left(PVE_\delta \equiv \left(\frac{\delta}{\theta+\tau+\epsilon+\gamma+\delta}\right)\right)$. A linear mixed effects model was fit using the package *lme4*(Bates et al., 2014a,b) in the statistical environment R (R Core Team). The significance of an association between a mRNA - SNP pair in a tissue-by-tissue (TBT-eQTL) analysis is assessed by the $p$ value obtained using *lm* function in R by fitting the following linear regression model.

For each mRNA - *cis*SNP pair, TBT-eQTL analysis was performed using the following linear regression model –

$$Y = \beta_0 + \beta_1 G + \epsilon$$

where $Y$ is either gene expression data and $G$ represents genotypes encoded as the number of copies of minor allele. The test statistic is the minimum $p$ value over the total number of tissues from linear regressions performed separately in each tissue for each mRNA - SNP

pair. Statistical significance was determined at a nominal $p$ value of 0.05 for all power simulations (in case of TBT-eQTL analysis, it is $\frac{0.05}{k}$ where $k$ is the number of tissues).

### *3.2.3 Preprocessing Gibbs et al datasets*

*Data description*

Fresh frozen tissue samples of the cerebellum (CRBLM), frontal cortex (FCTX), caudal pons (PONS) and temporal cortex (TCTX) were obtained from 150 neuropathologically normal samples (Gibbs et al., 2010). Genotyping was performed using Infinium HumanHap550 beadchips (Illumina) to assay genotypes for 561,466 SNPs, from the cerebellum tissue samples. CpG methylation status was determined using HumanMethylation27 BeadChips (Illumina), which measure methylation at 27,578 CpG dinucleotides at 14,495 genes. Profiling of 22,184 mRNA transcripts was performed using HumanRef-8 Expression BeadChips (Illumina) The datasets are publicly available (GEO Accession Number: **GSE15745**; dbGAP Study Accession: **phs000249.v1.p1**).

*Gene expression data*

Gene expression on four brain regions are publicly available as rank-invariant (Schmid et al., 2010) normalized gene expression data ("series matrix file"). All the negative values in the gene expression dataset are changed to a 1 and the entire dataset was then log2 transformed. Before generating the PCA plots, samples with African and Asian ancestry were removed from the analysis. All the gene expression probes on sex chromosomes X and Y were removed from the analysis. In order to identify outliers in the PCA analysis, a simple yet standard approach or rule has been adopted. All the samples that did not follow the inter-quartile range (IQR) rule ($median + 1.5 * IQR$) were excluded from further analysis (one CRBLM, one FCTX and two PONS). These samples were also eliminated by Gibbs *et al* in their original analysis.

Each gene expression probe was then adjusted for the biological and methodological

covariates such as tissue bank, gender, hybridization batch and numeric covariates such as post-mortem interval (PMI) and age in order to remove any associated confounding effects using the following linear model –

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + PC_1 + \ldots + PC_{50} + \epsilon$$

where Y is the gene expression data, $X_1 \ldots X_n$ represent the aforementioned biological covariates and systematic hybridization batch effects while $PC_1 \ldots PC_{50}$ are the top 50 principal components obtained from the original gene expression data. In order to target the difference in the genetic variation of expression among tissues, global variation in expression among tissues was removed by using the residual expression for each probe in each tissue after removing 50 PCs for further downstream analyses. It was shown in the past that the number of *cis*-eQTL detected significantly improved when 50 PCs were removed from the expression data (Fu et al., 2012).

*Methylation data*

Methylation data, obtained as a "series matrix file" consisted of Beta-values, which represent the ratio of methylated probe intensity and the overall intensity (sum of methylated and unmethylated probe intensities) (Du et al., 2010). The methylation data was later adjusted for the biological and methodological covariates such as tissue bank, gender, hybridization batch and numeric covariates such as post-mortem interval (PMI) and age in order to remove any associated confounding effects using the following linear model –

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

where Y is the methylation expression data and $X_1 \ldots X_n$ represent the aforementioned biological covariates and hybridization batch effects. The residual methylation expression was later used in the subsequent downstream analyses.

*Genotype data*

The genotype data was obtained from dbGAP database (**phs000249.v1.p1**) following requisite author permissions. The genotype data was recoded into a SNP matrix of values 0, 1 and 2 representing minor allele counts. Samples with African and Asian ancestry were removed from the analysis. These SNPs were filtered on the missing-ness of the individual data and the SNP data (excluded SNPs with missing values), followed by MAF (included SNPs with MAF $\geq 0.05$)and Hardy-Weinberg equilibrium (HWE; p-values $\leq 0.001$) in the same order using PLINK (Purcell et al., 2007) software. The resulting dataset has 400,097 SNPs after preprocessing.

## 3.3 Results and Discussion

### *3.3.1 Evaluating our new score test using Monte Carlo simulations*

We evaluate our approach through extensive simulation studies. Briefly, each Monte Carlo simulated dataset was comprised of data from a single locus and a single gene, whose expression is measured across 5 tissues in 100 observations. For a given mRNA - SNP pair, the genotypes at each SNP in all the individuals were simulated as Binomial(2,0.3), i.e. a minor allele frequency 0.30 and assuming Hardy-Weinberg equilibrium. Methylation data for 5 tissues was generated independently from a multivariate normal distribution with a positive definite variance-covariance matrix. Since all the tissue-specific effects are modeled as random effects, a test of whether there are any tissue-specific effects is equivalent to testing whether the variances of the random effects ($\gamma$ and $\delta$) are zero. Thus, our model involves testing four scalar parameters ($\beta$, $\phi$, $\gamma$ and $\delta$). Simulations under the null hypothesis confirm that our method has the correct type 1 error (see Appendix 2). Since we model the effects of both epigenetic and genetic variation, we evaluated any power loss in identifying mRNA - SNP associations in the absence of any epigenetic effect. This was accomplished by comparing our method's performance with TBT-eQTL approach by keeping all the pa-

rameters associated with methylation in equation 1 at zero (i.e. $\lambda = \phi = \delta = \theta = 0$). We also compared our method with a previously proposed multi-tissue eQTL method, implemented in our software JAGUAR (Acharya and Allen, 2016), which is made available at Comprehensive R Archive Network (CRAN) repository. Briefly, JAGUAR implements an approach that jointly models the overall shift in the gene expression due to genotype together with tissue-specific interaction with genotype in order to efficiently identify multi-tissue eQTL. From figure 3.2, we see that JAGUAR outperforms both TBT-eQTL and our new joint score test.

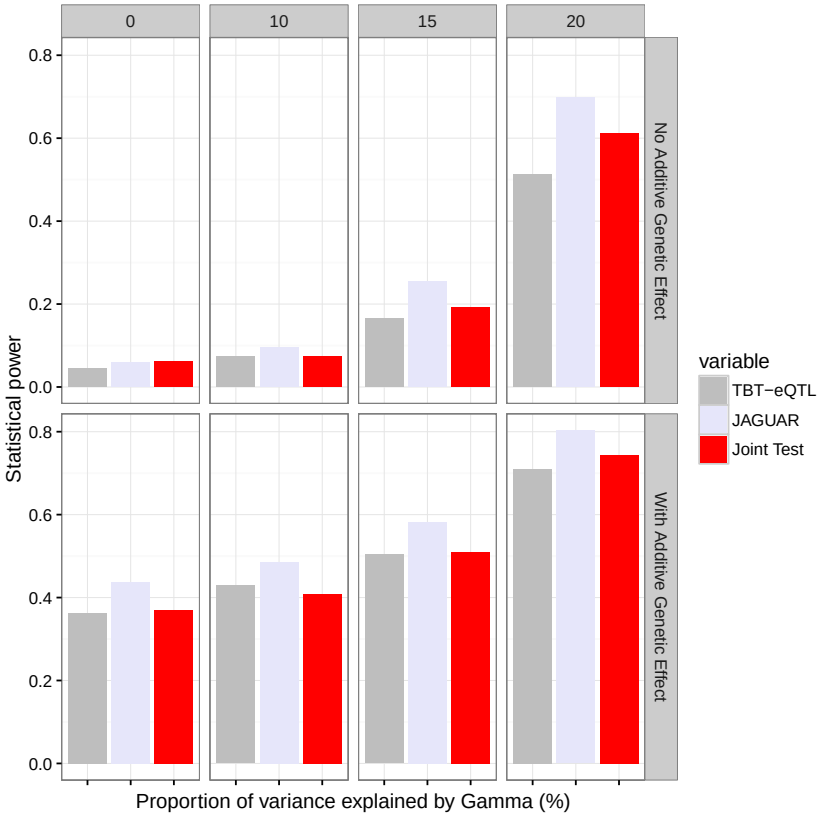



FIGURE 3.2: **Mapping eQTLs in multiple tissues using TBT-eQTL, JAGUAR and Joint Score Test methods.** This panel illustrates eQTL mapping in the absence of higher order methylation effects.

This loss of power, though not substantial, may be attributed to testing for an inexistent methylation effect. However, in the presence of a methylation effect our method

outperforms both TBT-eQTL and JAGUAR as evidenced by figure 3.3.

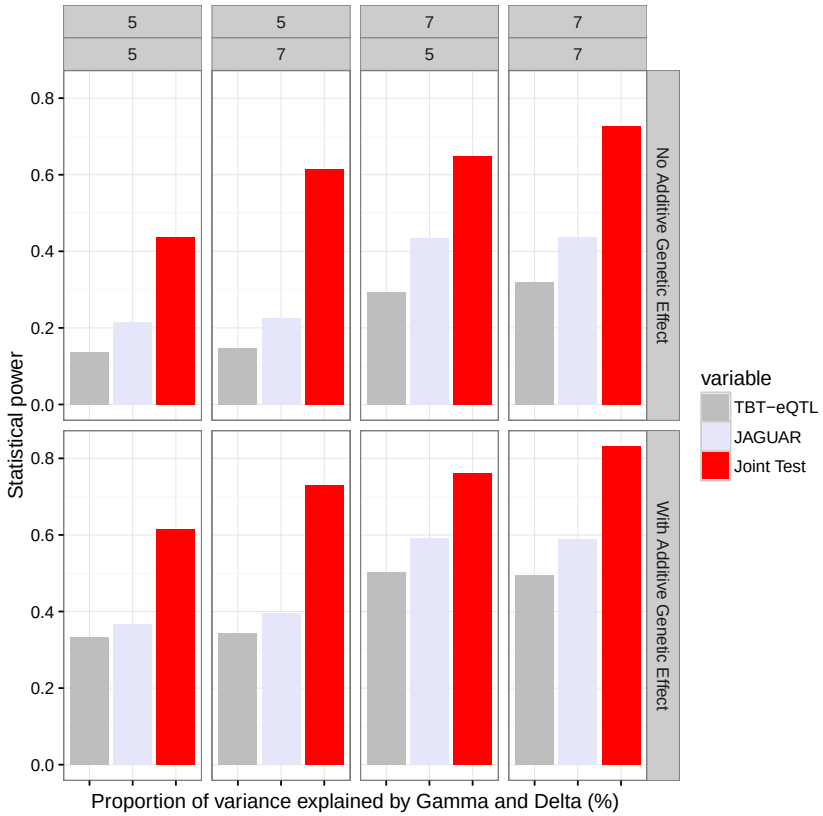



FIGURE 3.3: **Mapping eQTLs in multiple tissues using TBT-eQTL, JAGUAR and Joint Score Test methods.** Right panel illustrates eQTL mapping in the presence of higher order methylation effects and the numbers in the top two rows indicate the proportions of variance explained by both $\gamma$ and $\delta$, respectively.

We also compared our joint score test to a TBT-eQTL approach that included methylation as a baseline covariate (Gutierrez-Arcelus et al., 2013), henceforth referred to as TBTm-eQTL analysis, using the following linear regression model –

$$Y = M\alpha + G\beta + GM\phi + \xi \tag{3.5}$$

where $Y$ is a $nt$-dimensional matrix of expression levels in $t$ tissues and $n$ individuals, $\alpha$ is a fixed effect representing the tissue-specific intercepts, $G$ is a $nt$-dimensional matrix of genotypes, $\beta$ is a fixed effect of genotype across all tissues, $M$ is a $nt$-dimensional matrix of methylation information and $\phi$ is genotype $\times$ methylation interaction effect (fixed effect).

Minimum $p$ value from the TBTm-eQTL analysis across all the tissues is computed for power calculations. Table 1 shows that out method significantly outperforms TBTm-eQTL approach showing a clear statistical advantage in using our joint score test over the TBTm-eQTL approach.

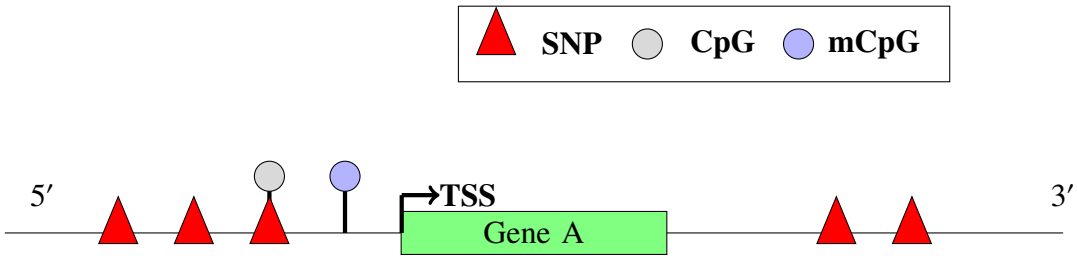



FIGURE 3.4: Here is an illustration of the analysis design using a hypothetical Gene A. The red triangles indicate SNPs and the circles, CpG sites (gray = unmethylated; blue = methylated). CpG sites that are at least 1.5 Kb from the transcription start site (TSS) of a gene were picked for the analysis. All the SNPs that were picked did not exceed 100 kilobase up- and down-stream of the transcription start site of a gene (cis-SNPs).

### *3.3.2 Region-specific DNA methylation impacts eQTL mapping in adult human brains*

In order to demonstrate the effectiveness of our method, we applied it to Gibbs *et al* (Gibbs et al., 2010) dataset comprising of 150 individual data obtained from four regions of human brain. We performed data analyses that focused on only *cis* candidate regions. The proximity of an eQTL to the transcription start site of a gene did not exceed 100 kilobase up- and down-stream of the transcription start site of a gene/mRNA transcript (*cis*-SNP). CpG islands that were less than 1.5 kilobase up- and down-stream of the transcription start site of the same mRNA transcript were paired with the mRNA transcripts. Figure 3.4 illustrates the analysis design in one tissue. Each mRNA transcript was tested for an association with every *cis*-SNP in the presence of a (methylated or unmethylated) CpG site located in the promoter region.

Our joint score test method performed a total of 491,496 tests (totaling 11,095 mRNA transcripts, 14,663 CpG sites and 144,571 *cis*SNPs). Each mRNA transcript may have

Table 3.1: Table comparing the statistical power of our method and TBTm-eQTL approach. This data were generated from 1,000 simulations run on 100 individuals and five tissues with genotypes generated at a common variant allele frequency (MAF = 0.3).

| Additive Genetic Effect | $PVE_{G\times M}$ | $PVE_{G\times M\times T}$ | $PVE_{G\times T}$ | TBTm-eQTL | Joint Test |
|---|---|---|---|---|---|
| NO | NO | 0 | 0 | 0.053 | 0.056 |
| NO | NO | 0 | 7 | 0.097 | 0.171 |
| NO | NO | 0 | 10 | 0.222 | 0.425 |
| NO | NO | 7 | 0 | 0.18 | 0.195 |
| NO | NO | 7 | 7 | 0.202 | 0.303 |
| NO | NO | 7 | 10 | 0.368 | 0.55 |
| NO | NO | 10 | 0 | 0.426 | 0.429 |
| NO | NO | 10 | 7 | 0.453 | 0.523 |
| NO | NO | 10 | 10 | 0.554 | 0.719 |
| NO | YES | 0 | 0 | 0.325 | 0.179 |
| NO | YES | 0 | 7 | 0.396 | 0.361 |
| NO | YES | 0 | 10 | 0.522 | 0.618 |
| NO | YES | 7 | 0 | 0.519 | 0.355 |
| NO | YES | 7 | 7 | 0.584 | 0.519 |
| NO | YES | 7 | 10 | 0.66 | 0.706 |
| NO | YES | 10 | 0 | 0.669 | 0.588 |
| NO | YES | 10 | 7 | 0.706 | 0.697 |
| NO | YES | 10 | 10 | 0.779 | 0.805 |
| YES | NO | 0 | 0 | 0.143 | 0.235 |
| YES | NO | 0 | 7 | 0.248 | 0.365 |
| YES | NO | 0 | 10 | 0.392 | 0.586 |
| YES | NO | 7 | 0 | 0.288 | 0.395 |
| YES | NO | 7 | 7 | 0.393 | 0.526 |
| YES | NO | 7 | 10 | 0.512 | 0.698 |
| YES | NO | 10 | 0 | 0.514 | 0.566 |
| YES | NO | 10 | 7 | 0.589 | 0.696 |
| YES | NO | 10 | 10 | 0.683 | 0.812 |
| YES | YES | 0 | 0 | 0.48 | 0.4 |
| YES | YES | 0 | 7 | 0.549 | 0.541 |
| YES | YES | 0 | 10 | 0.678 | 0.747 |
| YES | YES | 7 | 0 | 0.641 | 0.588 |
| YES | YES | 7 | 7 | 0.686 | 0.683 |
| YES | YES | 7 | 10 | 0.754 | 0.83 |
| YES | YES | 10 | 0 | 0.772 | 0.721 |
| YES | YES | 10 | 7 | 0.765 | 0.789 |
| YES | YES | 10 | 10 | 0.847 | 0.898 |

multiple CpG sites in its promoter region. Thus, each such mRNA - CpG pair is tested for an association with a *cis*SNP. It is important to note that our method does not test any direct association between an mRNA transcript and its corresponding CpG site. Any resulting combinations of mRNA transcript, CpG site and a SNP would describe the relationship between the mRNA and SNP in the presence of the corresponding promoter CpG site. Our method identified a total of 13,212 such combinations corresponding to 9,065 eQTLs that are statistically significant at 5% false discovery rate (FDR). In order to account for the number of traits being tested, the $p$ values obtained from applying our joint score test were adjusted for multiple testing using an optimized FDR approach to obtain $q$ values (FDR adjusted $p$ values) (Storey and Tibshirani, 2003). We observed that majority of these significant results are driven by a combination of additive genetic effect (86%) and $G \times T$ effect (43%) while the G $\times$ M and G $\times$ M $\times T$ effects were barely observed. This may be due to a lack of any distinct tissue-specificity in the methylation data, which we observed while preprocessing Gibbs *et al* data (see Methods section). However, we expect that the aforementioned effects may be well pronounced across diverse cell-types such as the ones made available by the Genotype-Tissue Expression Project (GTEx) (Lonsdale and et al, 2013).

We performed two region-by-region or TBT approaches on the same set of mRNA transcripts, CpG sites and SNPs as above, one with DNA methylation as a covariate (TBTm-eQTL) and the other with no methylation (TBT-eQTL) and compared the results with our approach. We estimated $q$ values from each set of $p$ values (originated from each region-by-region analysis) and minimum $q$ value for a given mRNA - SNP pair across all the brain regions was computed, which indicates the presence of a statistically significant pair in at least one brain region. The number of significant associations in at least one brain region were then assessed at 5% FDR ($q$ value $\leq \frac{0.05}{4}$ where 4 is the number of brain regions). TBT-eQTL approach identified a total of 11,014 mRNA-*cis*SNP pairs significant in at least one region of the brain at 5% FDR. Roughly 79% of these TBT-eQTLs over-

lap with eQTLs identified using our method. On the other hand, TBTm-eQTL approach identified 10,926 combinations of RNA transcripts, CpG sites and SNPs corresponding to 7,496 eQTLs with a 73% overlap with eQTLs identified using our method.

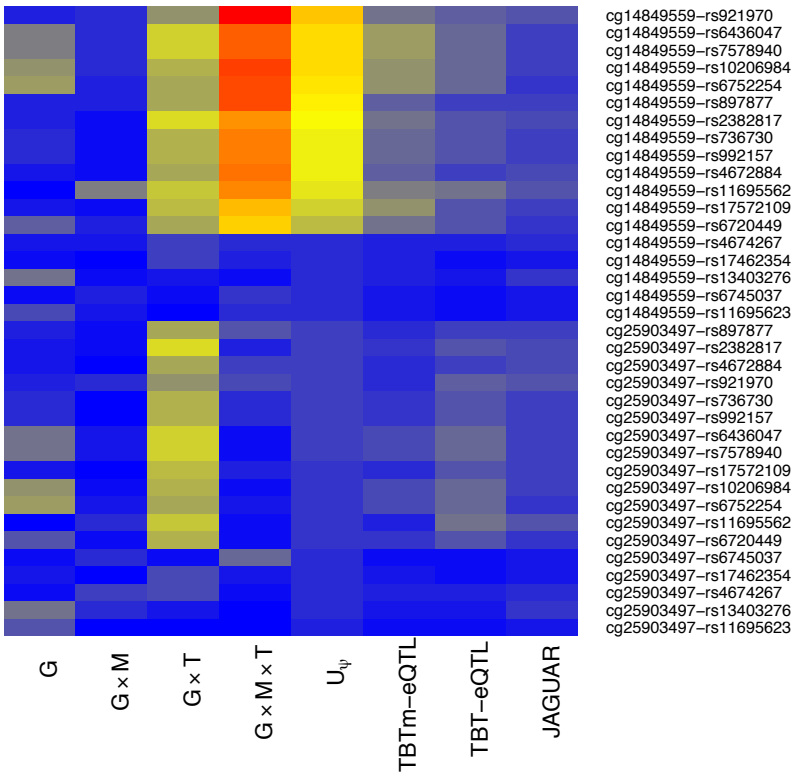



FIGURE 3.5: **Exploring the output from all different analyses.** Heatmap of $-log_{10}\, q$ values of all 36 possible combinations of TMBIM1 gene, two CpG sites and 18 SNPs. Red in the heatmap indicates a lower $q$ value and thus higher statistical significance where as blue indicates otherwise.

In order to assess the role of brain region-specificity on gene expression and the advantages in jointly modeling all the brain regions on mapping eQTLs, we compared our joint score test approach with a previously proposed multi-tissue eQTL mapping method (Acharya et al., 2016) implemented by our software JAGUAR. JAGUAR identifies 16,919 eQTLs (96% of them overlap with the TBT-eQTLs, 80% of them overlap with TBTm-eQTLs, and 92% of them overlap with the joint tests's eQTLs) at 5% FDR. All the eQTLs that overlap between JAGUAR and our new joint score test are mostly driven by the addi-

tive genetic effect and $G \times T$ effect and not higher order methylation interaction effects such as $G \times M$ and $G \times M \times T$. This absence of any pronounced region-specific DNA methylation effect explains the lower number of eQTLs identified by our method. However, as we have shown using simulation data, in the presence of any region-specific interaction effects involving methylation, our joint score test is far more informative than the results from JAGUAR. There are a total of 744 eQTLs identified by our method that weren't found to be statistically significant by JAGUAR. This is because 90% of these eQTLs were driven by $G \times M \times T$ interaction effect, which is not tested by JAGUAR. For example, let us consider a protein coding gene Transmembrane BAX inhibitor (TMBIM1; Ensemble ID - ENSG00000135926), located on chromosome 2, which has 18 annotated SNPs (possibly in LD with each other) and two promoter CpG sites in our preprocessed datasets. Out of these 36 (number of mRNA - CpG pairs $\times$ the number of SNPs) combinations of mRNA transcript, CpG sites and SNPs and a possible 18 eQTLs, our method identified 13 to be statistically significant. None of the 36 eQTLs were found to be statistically significant by any TBT-based or the multi-tissue eQTL approaches (Figure 3.5). Upon close inspection, we observed that there is an absence of any additive genetic effect or a $G \times T$ interaction effect however, the presence of $G \times M \times T$ interaction effect is driving the statistical significance. This is a good example of mapping eQTLs by leveraging effects due to DNA methylation. Of note, TMBIM1 is ubiquitously expressed in brain (Lisak et al., 2015) and BAX-inhibiting peptides have been known to prevent neuronal cell-death induced by oligomeric $\beta$-amyloid, which plays an important role in the pathogenesis of Alzheimer disease (Kudo et al., 2012). Out of the 13 significant eQTLs for TMBIM1 that our method identified, figure 3.6 illustrates the association between the gene, top CpG site (Methylation Probe ID: cg14849559) and SNP (SNP ID: rs921970) with the lowest $q$ value ($q$ val = 0.00386).

On the other hand, we also see many instances of eQTLs that were observed to be statistically significant using JAGUAR but not our joint score test method due to the lack of

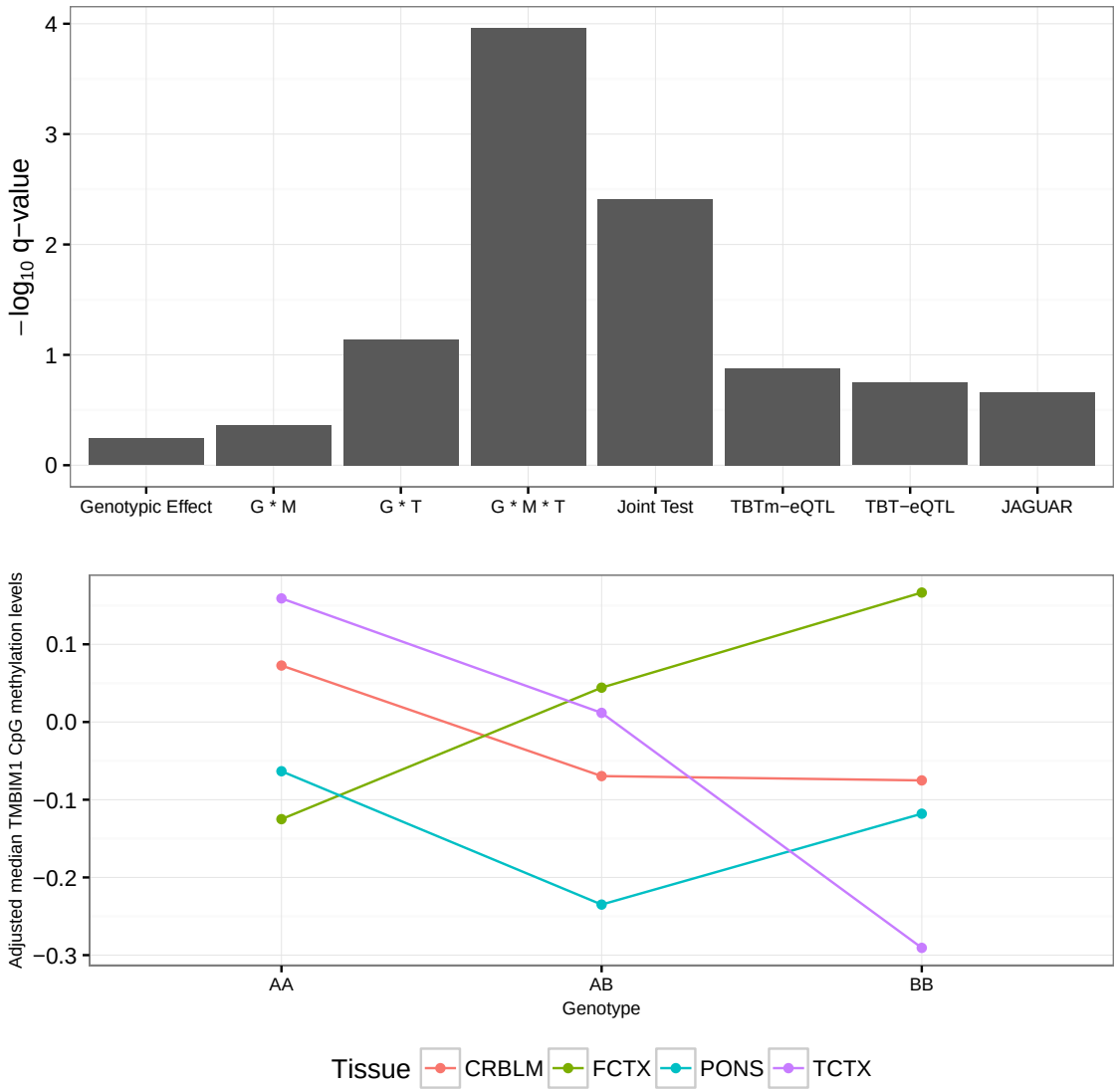



FIGURE 3.6: **TMBIM1 gene, CpG probe cg14849559 and SNP rs929170.** Top panel displays all the statistics computed for TMBIM1 gene, CpG probe cg14849559 and SNP rs929170. The first four bars indicate the four different effects tested by our joint score test and $U_\psi$ represents the omnibus $q$ value of our joint test. Bottom panel illustrate the $G \times M \times T$ interaction plot.

any DNA methylation effects. For example, JAGUAR identified gene B-Cell CLL/Lymphoma 2 (BCL2; Ensembl ID - ENSG00000171791), a gene that promotes neuronal cell death or apoptosis (Akhtar et al., 2004), to have a statistically significant association with a promoter eQTL (SNP ID: rs17676919), as illustrated by figure 3.7. As seen in this figure

(from $-log_{10}\ q$ values), the lack of any higher order methylation effects may have resulted in not being identified as a potential eQTL by our joint score test method.

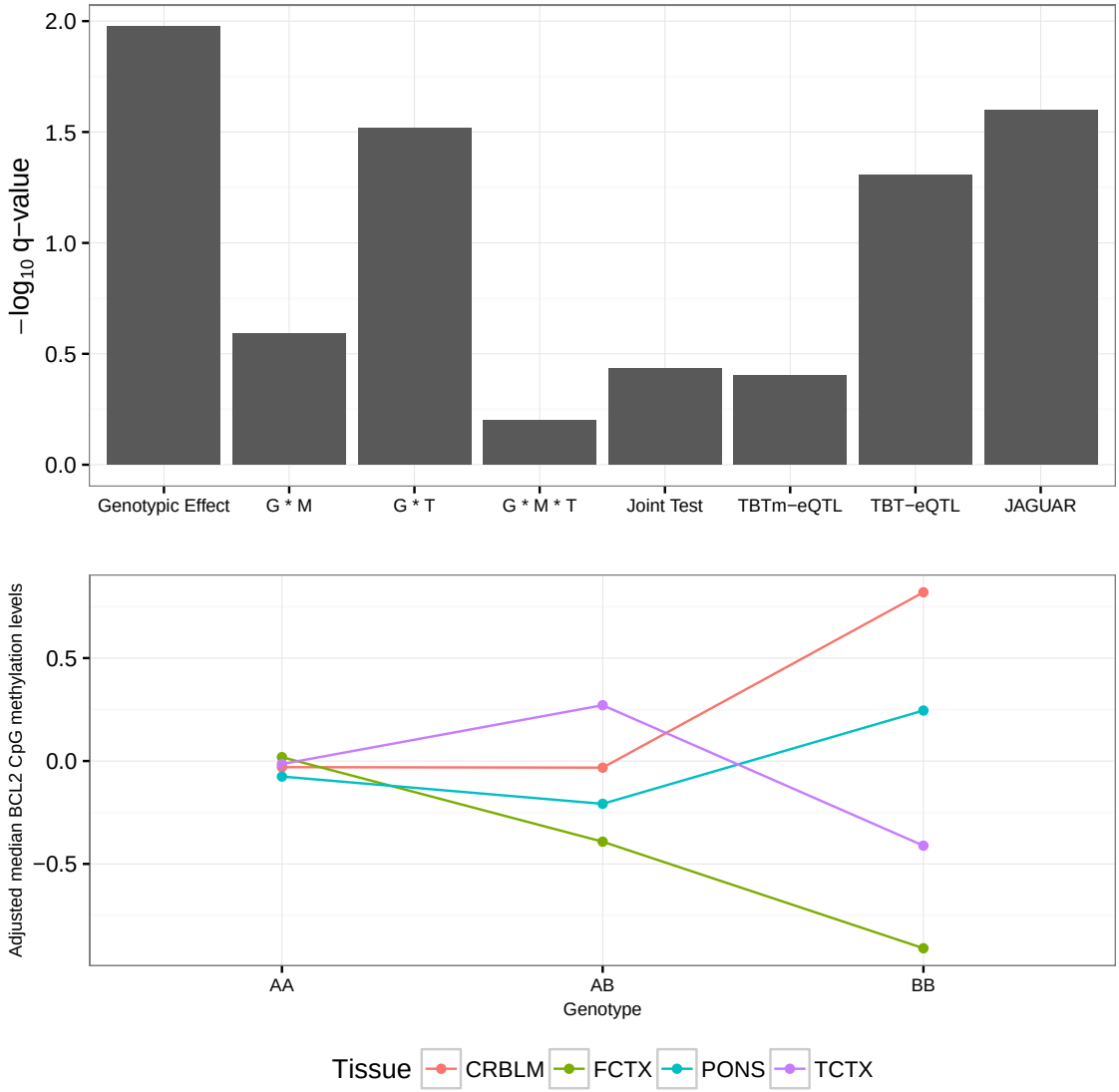



FIGURE 3.7: **An example of an eQTL for gene BCL2 not identified as statistically significant by our joint score test method.** Top panel displays all the statistics computed for BCL2 gene, CpG probe cg14455307 and SNP rs17676919. The first four bars indicate the four different effects tested by our joint score test and $U_\psi$ represents the omnibus $q$ value of our joint score test. Bottom panel illustrate the $G \times M \times T$ interaction plot.

To assess the biological relevance of the genes with eQTLs identified by TBT or multi-tissue methods including our new joint score test, we performed a KEGG pathway term enrichment analysis (Yu et al., 2014) for each set of results separately (see Appendix 2). KEGG pathways were considered overrepresented if a set of at least three genes from different linked regions is observed to be overrepresented with an adjusted significance level of $q$ value $< 0.05$, calculated from a hypergeometric test. Our method identified 5 overrepresented pathways (Metabolic pathways, Ribosome, Fatty acid degradation, Purine and Pyramidine metabolism), JAGUAR identified 2 pathways while TBT-eQTL identified 1 overrepresented pathway. The overrepresented pathway, "Metabolic Pathways" (KEGGID: hsa01100) is the only common pathway between TBT-eQTL, JAGUAR and our method methods. On the basis of prior knowledge of function, the overrepresented pathways "Purine metabolism" (KEGGID: hsa00230) and "Pyramidine metabolism" (KEGGID: hsa00240) are plausible functional candidate pathways for schizophrenia (Micheli et al., 2011). These information can be used to guide genetic analyses by selecting these relevant pathways and genes associated with the pathways for schizophrenia.

## 3.4 Conclusion

Overall, our efforts are primarily directed to understanding two very specific aspects – 1) the overall effect of a genetic variant on gene expression regulation by accounting for any changes in tissue DNA methylation levels, and 2) map eQTLs by leveraging tissue-specific methylation effects. Currently, there are no methods that jointly model the epigenetic and genetic control of tissue-specific gene expression. Many eQTL studies fail to account for the masking effect on a genetic variant due to DNA methylation, which may regulate gene expression across multiple tissues. Our method provides an efficient framework to integrate SNPs, DNA methylation and gene expression, and investigate how the different forms of variation inter-relate.

The dataset examined here used genome-wide association (GWA) study SNP array platform to interrogate germline variation that includes an overwhelming number of common variants. Although GWA studies have been able to explain a small fraction of the genetic components of common human diseases, it is hypothesized that some of the missing heritability may be due to rare variation. Since standard common disease common variant approaches are severely underpowered to tease out any underlying variants that are moderate to extremely rare, there is an emphasis on large sample sizes and gene-based association tests in order to securely identify genetic risk factors that may otherwise be outside the range detectable by GWA studies (McCarthy et al., 2008). One solution to the aforementioned issue would be to prioritize genetic variants in a non *ad-hoc* framework that preferentially weights genetic variants. Our method can provide a statistically disciplined weighting framework within which genetic variants can be either up- or down-weighted for any subsequent downstream analyses. Our method may also be useful in generating weights to any methods that use a reference data set in which both genome variation and gene expression levels have been measured to develop prediction models for gene expression (Gamazon et al., 2015).

Since we are modeling the effects of non-coding regions (via CpG sites) on gene expression using our model, we can easily use micro-RNA (miRNA) data instead of CpG site methylation data and model post-transcriptional regulation of tissue-specific gene expression. miRNA expression, also a tissue-specific phenomenon, have been known to post-transcriptionally silence expression of mRNA transcripts. The presence of genetic variants such as SNPs may have an effect on the biogenesis and function of miRNA molecules leading to a downstream effect on gene expression (Sun et al., 2009). This tissue-specific interaction between miRNA and SNP can be modeled in a similar fashion, analogous to modeling the interaction effects of tissue-specific DNA methylation and SNPs. The flexibility of our model also enables us to incorporate new information such as gene isoform data and accommodate the analysis of next-generation sequencing data (such as RNA-seq)

by modeling gene transcripts in an analogous fashion to tissues in our current model formulation. This type of analysis would aggregate expression over all the splice variants of a gene across multiple tissues and inform us of tissue-specific alternative splice variant of a gene. These results become relevant to studying genetic effects on alternative splicing and its key role in important cellular networks.

# 4

# Exploiting expression patterns across multiple gene isoforms to identify radiation response biomarkers in early-stage breast cancer patients.

## 4.1 Background

Radiation therapy or radiotherapy is utilized as a curative therapy in many solid tumors including gynecologic, head and neck, gastrointestinal, breast, prostate, lung, central nervous system and pediatric malignancies. Approximately 60% of cancer patients receive radiotherapy as part of their treatment either as a stand alone pre-operative therapy or combined with other modalities such as chemotherapy following surgery in an adjuvant setting (Brady et al., 2007). Radiotherapy has played a significant role in treating both invasive and non-invasive breast tumors over the years. However, response to radiation in breast cancer patients has not been uniform across all breast tumor subtypes (for example, basal, luminal, etc.) leading to a significant percentage of patients being either over- or under-treated (Horton et al., 2015; Langlands et al., 2013). This can be attributed to variable transcriptional response (through variable activation of transcription factors) to radiation, which is very similar to response to chemotherapy except that the mechanisms

underlying radiation response have not been well understood and studied (Macaeva et al., 2016; Wushou et al., 2015). Many genes have multiple transcript isoforms that result from alternative-splicing events. We measure the overall gene expression of a given gene by measuring the relative abundances of these isoforms, which may provide new insights into disease and biology.

Constantly evolving high-throughput gene expression profiling technologies, such as RNA-Seq or ultra high-resolution microarrays, have enabled us to interrogate all transcript isoforms in the human transcriptome by targeting coding transcripts, exon-exon splice junctions, and non-coding transcripts. The end goal of using these technologies is to exploit the gene expression patterns across multiple isoforms or gene transcripts in order to map biomarkers such as genes and gene-sets that help illuminate the molecular pathology of complex diseases at the RNA level. Existing analytic tools or methods for biomarker analysis such as differential expression analysis involves combining gene expression over all gene isoforms prior to data analysis, resulting in a gene-level interrogation of biological conditions (Miller et al., 2011; Irizarry et al., 2003; Efron and Tibshirani, 2006; Subramanian et al., 2005; Hänzelmann et al., 2013; Wang and Cairns, 2014). For example, the overall expression level of a gene can be represented by a single number and is measured by averaging the signals of many transcripts for the gene. Individual transcripts that have high variability compared to the average expression of a gene will be removed from the analysis (outliers). Such an approach has at least two significant limitations. First, it fails to fully exploit expression patterns across gene isoforms either by combining information across multiple transcripts or by not explicitly identifying effects that differ across transcripts. Second, and more importantly, it fails to account for alternative splicing or alternative $3'$ poly-adenylation events by removing gene isoforms that seems to be significantly differentially expressed. We propose two distinct approaches, one to identify radiation-induced gene expression biomarkers in an isoform-specific differential expression (DE) analysis and another to perform isoform-specific gene-set or pathway enrichment analysis. We test

these methods extensively using simulation studies and then evaluate the effectiveness of these two methods on a microarray-based gene expression dataset containing 26 paired early-stage breast cancer patient samples. Briefly, these tumor samples originated from a unique preoperative radiotherapy Phase I trial (Horton et al., 2015), and were assayed on the new Affymetrix Human Transcriptome 2.0 array (Affymetrix, 2016). The transcriptome response to radiation exposure was derived by comparing gene expression in samples before and after irradiation. While demonstrating the effectiveness of our method to identify differentially enriched gene-sets, we investigated the effects of radiation on 7 tumor microenvironment and 24 hallmark oncogenic signaling gene-sets that are associated with radiation response.

We hypothesize that our methods are effective in identifying biomarkers (in this case, differentially expressed genes and differentially enriched gene-sets) when compared to most commonly used approaches. Investigating the tumor microenvironment and the oncogenic signaling pathways before and after radiation will help us understand any radiation-induced changes in individual patients, which may serve as a surrogate to understand patient response to radiation and can make for potential therapeutic targets.

## 4.2 Materials and Methods

### *4.2.1 Microarray analysis of the breast cancer dataset*

Raw microarray data for twenty six early-stage breast cancer patients were obtained from NCBI's gene expression omnibus (GEO ID: GSE65505) repository (Edgar et al., 2002). All the patients are at least 55 years old, clinically node negative, ER-positive and/or PR-positive, HER2-negative (biologically favorable tumors) with T1 invasive carcinomas or low-intermediate grade *in situ* disease $\leq 2cm$. These patients received pre-operative radiotherapy (radiation dose prior to surgical resection of tumor). All the samples were arrayed on Affymetrix Human Transcriptome Array 2.0 (Affymetrix, 2016), which was

designed with approximately ten probes per exon and four probes per exon-exon splice junction. At the top level, each transcript cluster roughly corresponds to a gene. Each transcript cluster is comprised of exon clusters that a) shared splice sites, b) or were derived from overlapping exonic sequences, c) or were single-exon clusters bounded on the genome by spliced content. Each exon cluster is further fragmented into probe selection regions (PSRs), which are non-overlapping contiguous sequences. Gene-level and gene isoform/transcript-level expression data were obtained using R/Bioconductor packages *oligo* (Carvalho and Irizarry, 2010), *affyio* (Bolstad, 2016) and *pd.hta.2.0* (MacDonald, 2016), and pre-processed by robust multi-array average (RMA) method (Irizarry et al., 2003; Carvalho and Irizarry, 2010; Bolstad, 2016), which summarizes the probe level expression data into a probe set level expression value. Principal component analysis was conducted to check for batch effects in both gene-level and transcript-level data, and any batch effects that were identified were corrected using a popular Empirical Bayes approach (ComBat) (Johnson et al., 2007). DE analysis was performed on genes with at least two transcript isoforms. This resulted in a dataset with more than 800,000 transcripts.

### *4.2.2 Strategy to identify gene expression biomarkers of radiation: Differential Expression (DE) analysis*

Given two distinct biological groups (before and after radiation treatment), gene expression for each gene transcript, $Y$, can be modeled in the following way

$$Y = T\alpha + R\beta + Au + Bv + \xi \tag{4.1}$$

where $Y$ is a $ntg \times 1$ matrix of expression values, $T$ is a $ntg \times t$ dimensional matrix of gene expresion levels in $t$ isoforms of a gene in $g$ groups and $n$ individuals, $\alpha$ is a fixed effect representing the isoform-specific intercepts, $R$ is a $ngt \times 1$-dimensional matrix of radiation dose identifiers such that $R \in \{0, 1\}$, 0 indicates no radiation and 1 indicates radiation, $\beta$ is a fixed effect indicating the average effect of radiation on gene expression. $u \sim N\left(0, \tau AA^T\right)$

indicates subject-specific random intercept, $v \sim N\left(0, \gamma BB^T\right)$ is random effect that denotes the interaction between gene-isoform and radiation (isoform-specific radiation effect), and $\xi \sim N\left(0, \epsilon I\right)$. $I$ is $ntg \times ntg$ dimensional identity matrix . The matrices $J$, $A$ and $B$ are design matrices with $B$ being a function of radiation dose. $J$ is $ntg \times t$ dimensional matrix denoting the design matrix for the tissue-specific intercepts. $A$ is $ntg \times n$ design matrix for the subject-specific intercepts. $B$ is a $ntg \times t$ design matrix of stacked radiation dose identifiers.

We test the null hypothesis that $H_0 : \beta = 0; \gamma = 0$ i.e radiation does not affect gene expression. From our model above, we derive our score test statistic, $U_\psi$ as

$$U_\psi \equiv \hat{Y}^T \hat{\Sigma}^{-1} \left[ a_\beta \left(R - \bar{R}\right) \left(R - \bar{R}\right)^T + a_\gamma \left( \frac{1}{2} BB^T \right) \right] \hat{\Sigma}^{-1} \hat{Y}, \tag{4.2}$$

where $a_\beta$ and $a_\gamma$ are scalar constants chosen to minimize the variance of $U_\psi$ (see Supplementary methods). The $p$ values are approximated using Satterthwaite method (Satterthwaite, 1946). The maximum likelihood estimates, obtained from fitting a standard linear mixed model using lme4 (Bates et al., 2014a), are computed only once per gene since under the null, there is no effect due to radiation on the gene expression. The $p$ values obtained from applying our method were adjusted for multiple hypothesis within the false discovery rate (FDR) framework. Genes with FDR adjusted $p$ values ($q$ values) less than 0.05 were selected to be differentially expressed. More information on our method is available in the supplementary methods. As a side note, this model is very similar to a previous one we proposed (Acharya et al., 2016) with the exception that this is a paired data.

### *4.2.3 Strategy to perform radiation-induced isoform-specific gene-set enrichment analysis*

Gene expression data for each pathway, $Y$, is modeled in the following way

$$Y = T\alpha + G\lambda + R\beta + Au + Bv + Cw + \xi \tag{4.3}$$

where $Y$ is $ntjg \times 1$ dimensional matrix of expression values, $T$ is a $ntjg \times t$-dimensional matrix of expression levels in $t$ isoforms of a gene, $j$ genes, $g$ groups and $n$ individuals, $\alpha$ is a fixed effect representing $t$ isoform-specific intercepts, $\lambda$ is a fixed effect representing $g$ gene-specific intercepts, $R$ is a $ntjg \times 1$ dimensional matrix of radiation dose identifiers such that $R \in \{0, 1\}$, 0 indicates no radiation and 1 indicates radiation, $\beta$ is a fixed effect indicating the average effect of radiation on a pathway or gene-set. $u \sim N(0, \tau AA^T)$ indicates subject-specific random intercept, $v \sim N(0, \gamma BB^T)$ is a random effect that denotes the interaction between gene-isoform and radiation (isoform-specific radiation effect), $w \sim N(0, \phi CC^T)$ is a random effect that denotes the interaction between gene and radiation (gene-specific radiation effect), and $\xi \sim N(0, \epsilon I)$. $I$ is $ntjg \times ntjg$-dimensional identity matrix . The matrices $J$, $A$ and $B$ are design matrices with $B$ being a function of radiation dose. $J$ is $ntjg \times t$ dimensional matrix denoting the design matrix for the tissue-specific intercepts. $A$ is $ntjg \times n$ design matrix for the subject-specific intercepts. $B$ is a $ntjg \times t$ design matrix of stacked radiation dose identifiers and $C$ is a $ntjg \times g$ dimensional design matrix of the $R \times G$ effect.

We test the null hypothesis that $H_0 : \beta = 0; \gamma = 0; \phi = 0$ i.e radiation does not affect gene expression. From our model above, we derive our score test statistic, $U_\zeta$ as

$$U_\zeta \equiv \hat{Y}^T \hat{\Sigma}^{-1} \left[ a_\beta (R - \bar{R}) (R - \bar{R})^T + a_\gamma \left( \frac{1}{2} BB^T \right) + a_\phi \left( \frac{1}{2} CC^T \right) \right] \hat{\Sigma}^{-1} \hat{Y}, \tag{4.4}$$

where $a_\beta$, $a_\gamma$ and $a_\phi$ are scalar constants chosen to minimize the variance of $U_\zeta$. The $p$ values are approximated using Satterthwaite method (Satterthwaite, 1946). Similar to our earlier method, the maximum likelihood estimates, obtained from fitting a standard linear mixed model using lme4 (Bates et al., 2014a), are computed only once per gene-set since under the null, there is no effect due to radiation on the gene expression. The $p$ values obtained from applying our method were adjusted for multiple hypothesis within the false discovery rate (FDR) framework. Genes with FDR adjusted $p$ values ($q$ values)

less than 0.05 were selected to be differentially expressed. More details on our method are available in supplementary methods.

### *4.2.4 Simulations*

*Testing our method for DE analysis*

We have performed the following two simulation studies in order to verify our approach. In our first study, we simulated one gene at a time from the following linear model and varied the following parameters- $\beta$ (additive effect due to radiation), the proportion of variation explained by $\gamma$ or $R \times T$ effect $\left(PVE_\gamma \equiv \left(\frac{\gamma}{\tau+\epsilon}\right)\right)$ and the number of transcripts. For a positive integer $tg$ that represents the combined number of transcripts ($t$) and groups ($g$), if $\mathbf{1}$ denotes a column vector of $tg$ ones and $\mathbb{I}$ denotes the corresponding $tg \times tg$ diagonal matrix, following the $tg$-variate normal law denoted by $N_{tg}\left[\mu, \Sigma\right]$ with mean $\mu \in \mathbb{R}^{tg}$ and variance $\Sigma \in \mathbb{R}^{tg \times tg}$, expression levels of a target gene $j$ by using the following vectorized form of the linear mixed model –

$$y_{ijg} = \alpha_j + \beta_j r_g + \mathbf{1}a_i + b_j r_g + \xi_{ijg} \qquad \xi_{ijg} \overset{i.i.d.}{\sim} N\left(0, \epsilon\mathbb{I}\right) \tag{4.5}$$

where $y_{ijg}$ is a $tg \times 1$ vector of gene expression data, $\alpha_t$ is the transcript-specific intercept ($\alpha_t \in \mathbb{R}^t$), $\beta_j$ describes the main additive effect ($\beta_j \in \mathbb{R}^1$), $r_g$ is a vector of length $tg$ such that $r \in (0_t, 1_t)$. The random effect $b_j \in \mathbb{R}^{tg}$ represents transcript-specific interaction effect of radiation, and $a_i \in \mathbb{R}^1$ is a subject-specific random intercept. We assume that all the random effects are independent and that $a_i \sim N_1\left(0, \tau\right)$, $b_j \sim N_{tg}\left(0, \gamma\mathbb{I}\right)$. A linear mixed effects model was fit using the package *lme4*(Bates et al., 2014a) in the statistical environment R (R Core Team).

We then compared our method with a standard paired t-test and a non-parametric alternative in Wilcoxon's test (Wilcoxon, 1945). The test statistic in case of transcript-by-transcript (TBT) analysis is the minimum $p$ value over the total number of transcripts from either t-test or Wilcoxon's test performed separately in each transcript for each paired

sample. A gene-level test was constructed over all the transcripts by taking the median expression value across the transcripts followed by a standard paired t-test. Statistical significance was determined at a nominal $p$ value of 0.05 for all power simulations (in case of TBT analysis, it is $\frac{0.05}{k}$, where $k$ is the number of transcripts). We used 10,000 data replicates to evaluate the type I error and 1,000 data replicates for power calculations.

We have also tested our method on a synthetic dataset simulated from a multivariate normal distribution containing two classes of data. Each gene was simulated to have variable number of transcripts. We used this dataset with increasing number of genes (by also keeping a small proportion differentially expressed) and tested our approach at both transcript-level (paired t-test and Wilcoxon's test) and gene-level. The most commonly used method to combine $p$ values of all the transcripts of a gene is Fisher's method however, under the assumption that all the $p$ values are independent (Fisher, 1934). This assumption may be frequently violated since different isoforms of a gene may be correlated and the resulting $p$ values are dependent on each other. At the gene-level, paired t-tests were run on gene expression values of a gene that were aggregated over its transcripts by either their median expression values or Winsorized mean (Wilcox and Keselman, 2003b) expression values.

*Testing our method for gene-set enrichment analysis*

Similar to the above analyses, we have performed two simulations studies in order to verify our approach. In our first study, we simulated one gene-set at a time from the following linear model and varied the following parameters- $\beta$ (additive effect due to radiation), the proportion of variation explained by $\gamma$ or $R\times T$ effect $\left(PVE_\gamma \equiv \left(\frac{\gamma}{\tau+\phi+\epsilon}\right)\right)$, the proportion of variation explained by $\phi$ or $R\times G$ effect $\left(PVE_\phi \equiv \left(\frac{\phi}{\tau+\phi+\epsilon}\right)\right)$ and the number of transcripts. For a positive integer $tjg$ that represents the combined number of transcripts ($t$), genes ($j$) and groups ($g$), if $\mathbf{1}$ denotes a column vector of $tjg$ ones and $\mathbb{I}$ denotes the corresponding

$tjg \times tjg$ diagonal matrix, following the $tjg$-variate normal law denoted by $N_{tjg}\left[\mu, \Sigma\right]$ with mean $\mu \in \mathbb{R}^{tjg}$ and variance $\Sigma \in \mathbb{R}^{tjg \times tjg}$, expression levels of a target geneset $k$ by using the following vectorized form of the linear mixed model –

$$y_{ijkg} = \alpha_j + \beta_k r_g + \lambda_k + \mathbf{1}a_i + b_j r_g + c_k r_g + \xi_{ijkg} \qquad \xi_{ijkg} \overset{i.i.d.}{\sim} N\left(0, \epsilon\mathbb{I}\right) \tag{4.6}$$

where $y_{ijkg}$ is a $tjg \times 1$ vector of gene expression data, $\alpha_t$ is the transcript-specific intercept ($\alpha_t \in \mathbb{R}^t$), $\beta_k$ describes the main additive effect ($\beta_k \in \mathbb{R}^1$), $r_g$ is a vector of length $tjg$ such that $r \in \left(0_{tg}, 1_{tg}\right)$. The random effect $b_t \in \mathbb{R}^{tjg}$ represents transcript-specific interaction effect of radiation, the random effect $c_j \in \mathbb{R}^{tjg}$ represents transcript-specific interaction effect of radiation, and $a_i \in \mathbb{R}^1$ is a subject-specific random intercept. We assume that all the random effects are independent and that $a_i \sim N_1\left(0, \tau\right)$, $b_t \sim N_{tg}\left(0, \gamma\mathbb{I}\right)$ and $c_j \sim N_{jg}\left(0, \gamma\mathbb{I}\right)$. A linear mixed effects model was fit using the package *lme4*(Bates et al., 2014a) in the statistical environment R (R Core Team).

We then compared our method with a standard paired t-test and a non-parametric alternative in Wilcoxon's test (Wilcoxon, 1945). The test statistic in case of transcript-by-transcript (TBT) within a gene analysis is the minimum $p$ value over the total number of transcripts and genes from either t-test or Wilcoxon's test performed separately in each transcript for each paired sample. A gene-level test was constructed over all the transcripts by taking the median expression value across the transcripts followed by a standard paired t-test. Statistical significance was determined at a nominal $p$ value of 0.05 for all power simulations (in case of TBT analysis, it is $\frac{0.05}{k}$, where $k$ is the product of the number of transcripts and genes). We used 10,000 data replicates to evaluate the type I error and 1,000 data replicates for power calculations.

A second set of simulations involved generating a synthetic gene expression data from a multivariate normal distribution containing two classes of data. Each gene was simulated to have variable number of transcripts. We defined two types of gene-sets, one with overlapping genes and the other with non-overlapping genes, and randomly assigned some

gene-sets to to contain differentially expressed genes. Since most, if not all of the current methods involve gene-set analysis at the gene level, we compared our method with Gene Set Variational Analysis (GSVA) (Hänzelmann et al., 2013), Pathway Level Analysis of Gene Expression (PLAGE) (Tomfohr et al., 2005), single sample GSEA (ssGSEA) (Barbie et al., 2009) and the combined z-score (ZSCORE) (Lee et al., 2008) methods. Both, PLAGE and the ZSCORE are parametric and assume that gene expression profiles are jointly normally distributed. More about these methods in the supplementary material.

### *4.2.5 Defining the gene-sets and gene-set analysis*

All the hallmark oncogenic signaling pathways used in our primary data analysis were obtained from the Molecular Signature Database version 3 (MSigDB) collection (Liberzon et al., 2015). We focussed our attention on 24 specific oncogenic signaling pathways that were most likely associated with radiation response. We defined tumor microenvironment as a collection of proteins produced by cells present in and around the tumor that support the growth of the cancer cells. We included gene-sets representing hypoxia (Chi et al., 2006), invasiveness/metastases gene signature (Liu et al., 2007), epigenetic stem cell signature in cancer (Widschwendter et al., 2007), inflammatory pathway involving tumor necrosis factors (Viemann et al., 2006), angiogenesis (Chang et al., 2004), immune signatures (Hsu et al., 2010) and a form of genomic instability called chromosomal instability (Carter et al., 2006), which determines the tumor cell's ability to respond to its microenvironment. In order to visualize sample set enrichment of these gene-sets (enrichment level of a gene-set in a sample), we employed Gene Set Analysis (GSA) software (Efron and Tibshirani, 2010), which implements a supervised method (class labels are known before the analysis) that computes a "maxmean" summary statistic for each gene-set. Briefly, GSA computes the average of both positive and negative aspects of gene-scores (for example, fold changes) over each gene in a gene-set, and choose the one that is larger in absolute value (Efron and Tibshirani, 2006).

### *4.2.6 Multiple hypothesis correction*

Wherever applicable, we use multiple hypothesis correction based on the Benjamini-Hochberg (BH) approach (Benjamini and Hochberg, 1995) to obtain corrected $p$ values. In case of gene-set analysis, BH approach may result in a conservative estimate of the false discovery rate (FDR) because of overlapping gene-sets that have highly correlated genes. We used the BH method only as a demonstration of statistical power.

## 4.3 Results

Whole transcriptome expression profile analysis usually focuses on a gene-level analysis by combining gene expression data over all transcripts of a gene. This approach has a significant limitation in that it fails to exploit expression patterns across the transcripts by not explicitly identifying effects that differ among the gene transcripts. Marginal analyses of individual gene transcripts may also lead to a proliferation of hypotheses tested, which can negatively impact the power of biomarker discovery. Popular method used to combine $p$ values such as Fisher's approach assume independence among all the transcripts of a gene, which may not be entirely true in this case. We address the aforementioned issues by proposing two score-test based approaches, one to discover differentially expressed genes and another to identify differentially enriched gene-sets. Score test-based approaches do not require parameter estimation under the alternative hypothesis. As a result, model parameters only have to be estimated once per genome, significantly decreasing computation time. Further, our score-based approaches only require estimation of the first two moments of the random effects, and therefore are robust to misspecification of the random effect distribution (Lin, 1997).

### 4.3.1 *Evaluating our method to identify differentially expressed (DE) genes using simulated data*

We evaluated our method to detect DE genes using two simulation studies. Briefly, each Monte Carlo simulated dataset from the first simulation study was comprised of data for a single gene, whose expression is measured across 5 or 10 transcripts in 50 paired individuals. Each individual pair's radiation status is either a zero or a one indicating before and after radiotherapy, respectively. Since the transcript-specific effect is modeled as a random effect, a test of whether there is any transcript-specific effect due to radiation is equivalent to testing whether the variance of the random effect ($\gamma$) is zero. Thus, our model to detect DE genes involves testing two scalar parameters in $\beta$ and $\gamma$. Simulations under the null hypothesis (no effect of radiation on overall gene expression) confirm that our method has the right type I error. More details in the supplementary section.

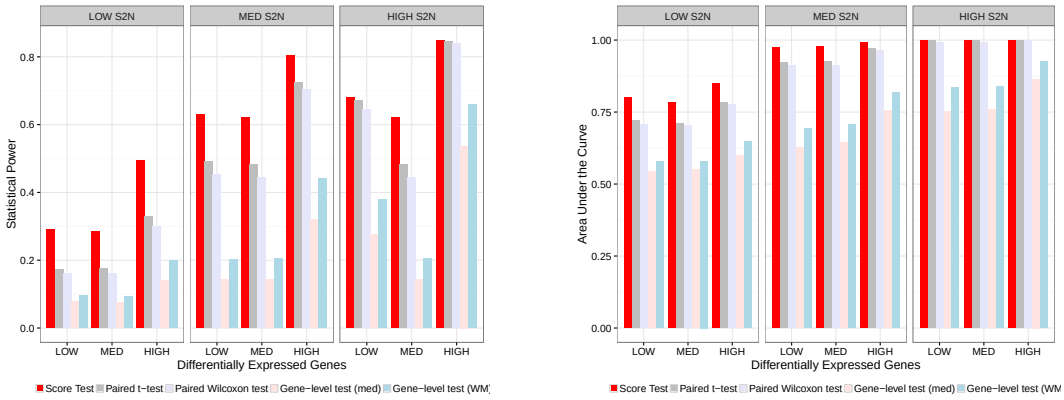
| Additive Effect | $PVE_\gamma(\%)$ | DE Score Test | TBT Paired t-test | TBT Wilcoxon's test | Gene-level paired t-test |
|---|---|---|---|---|---|
| | | **Number of transcripts per gene = 5** | | | |
| NO | 0 | 0.051 [0.038-0.067] | 0.052 [0.036-0.073] | 0.054 [0.037-0.075] | 0.044 [0.032-0.059] |
| NO | 9 | 0.36 [0.33-0.391] | 0.291 [0.255-0.329] | 0.263 [0.228-0.3] | 0.114 [0.095-0.135] |
| NO | 13 | 0.629 [0.598-0.659] | 0.536 [0.495-0.577] | 0.504 [0.463-0.545] | 0.205 [0.18-0.231] |
| YES | 0 | 0.373 [0.343-0.404] | 0.259 [0.224-0.296] | 0.239 [0.205-0.275] | 0.385 [0.355-0.416] |
| YES | 9 | 0.634 [0.603-0.664] | 0.515 [0.474-0.556] | 0.493 [0.452-0.534] | 0.418 [0.387-0.449] |
| YES | 13 | 0.759 [0.731-0.785] | 0.66 [0.62-0.698] | 0.627 [0.587-0.666] | 0.447 [0.416-0.478] |
| | | **Number of transcripts per gene = 10** | | | |
| NO | 0 | 0.053 [0.04-0.069] | 0.043 [0.027-0.064] | 0.039 [0.024-0.059] | 0.059 [0.045-0.075] |
| NO | 9 | 0.534 [0.503-0.565] | 0.352 [0.31-0.396] | 0.318 [0.277-0.361] | 0.135 [0.114-0.158] |
| NO | 13 | 0.861 [0.838-0.882] | 0.682 [0.639-0.723] | 0.642 [0.598-0.684] | 0.21 [0.185-0.237] |
| YES | 0 | 0.539 [0.508-0.57] | 0.302 [0.262-0.344] | 0.264 [0.226-0.305] | 0.646 [0.615-0.676] |
| YES | 9 | 0.831 [0.806-0.854] | 0.633 [0.589-0.675] | 0.588 [0.543-0.632] | 0.63 [0.599-0.66] |
| YES | 13 | 0.92 [0.901-0.936] | 0.832 [0.796-0.864] | 0.803 [0.766-0.837] | 0.604 [0.573-0.634] |

Table 4.1: DE of genes - Simulation results at 5% FDR with 95% confidence interval. We varied additive effect i.e. average effect of radiation on the whole transcriptome and proportion of variation explained by $\gamma$ i.e. radiation $\times$ transcripts interaction effect. Our score test is referred to as "DE Score Test".

Power simulations were performed by varying the following parameters- 1) additive effect of radiation ($\beta$), 2) the proportion of variation explained by the interaction effect between radiation and transcript ($PVE_\gamma$) and 3) the number of transcripts. The results in table 1 shows that our method significantly outperforms transcript-by-transcript paired t-

test and Wilcoxon test (a non-parametric alternative to t-test) in all simulated situations. However, the gene-level paired t-test seems to work the best when there is an overall shift in gene expression due to radiation but absence of any transcript-specific effects.

In the second simulation study, each Monte Carlo dataset, comprised of gene expression data for 50 genes over 50 observations, each gene with unequal number of isoforms, was simulated from a multivariate normal distribution with a known variance-covariance matrix. We varied the mean difference in differential gene expression between the two phenotypes (signal-to-noise ratio), and the proportion of differentially expressed gene-isoforms. At the transcript level, we applied paired t-test and a non-parametric alternative in Wilcoxon's paired t-test and combined the $p$ values over all the transcripts of a gene using Fisher's method. At the gene-level, we combined the gene expression values by computing either the median or Winsorized mean of all the transcripts within a given gene. Paired t-test was run on this gene-level data. We varied the proportion of genes that are differentially expressed and the signal-to-noise ratio. Statistical power and empirical type I error rates were estimated based on a nominal FDR of 5%. Figure 4.1 displays the performance of all the methods, measured both in terms of statistical power and area under the curve (AUC). AUC for all the methods was estimated using R package ROCR (Sing et al., 2005). We see that our method does well compared to the rest of the methods based on AUC plot. Given how the gene expression data were generated, every gene may have a fraction of transcripts differentially expressed. Consequently, any method for identifying DE genes must account for this transcript-specific variability. By combining gene expression values over all the transcripts of a gene (as evidenced by any gene-level methods), we are not able to fully exploit transcript-specific gene expression patterns. This is evident in Figures 4.1a and 4.1b, where the gene-level tests perform poorly compared to the transcript-level tests, including our approach.




(a) Power simulations (b) Area under the ROC curve

FIGURE 4.1: The performance of all the methods in detecting DE genes. A) Bar plot depicting the statistical power of each method under changing number of differentially expressed genes and the mean difference in gene expression (signal-to-noise ratio; S2N) between the two phenotypes (before and after radiation). We compared our method with two transcript-level tests in paired t-test and paired wilcoxon test ($p$ values combined at gene-level by Fisher's method), and with two gene-level tests, where the gene expression values are combined by median and Winsorized mean values followed by a paired t-test. B) Bar plot depicting the area under the curve (AUC) of all the methods under the aforementioned conditions.

### *4.3.2 Evaluating our method to identify DE gene-sets using simulated data*

We evaluated our method to detect DE gene-sets or pathways using two simulation studies. Briefly, each Monte Carlo simulated dataset from the first simulation study was comprised of data for a single gene-set comprising of 5 genes, whose expression is measured across 3 transcripts in 50 paired individuals. Each individual pair's radiation status is either a zero or a one indicating before and after radiotherapy, respectively. Since the transcript-specific effect is modeled as a random effect, a test of whether there is any transcript-specific effect on the gene-sets due to radiation is equivalent to testing whether the variances of the random effects ($\gamma$ and $\phi$) are zero. Thus, our model to detect enriched gene-sets involves testing three scalar parameters in $\beta$, $\gamma$ and $\phi$. Simulations under the null hypothesis (no effect of radiation on overall gene expression) confirm that our method has the right type I error (see supplementary material).

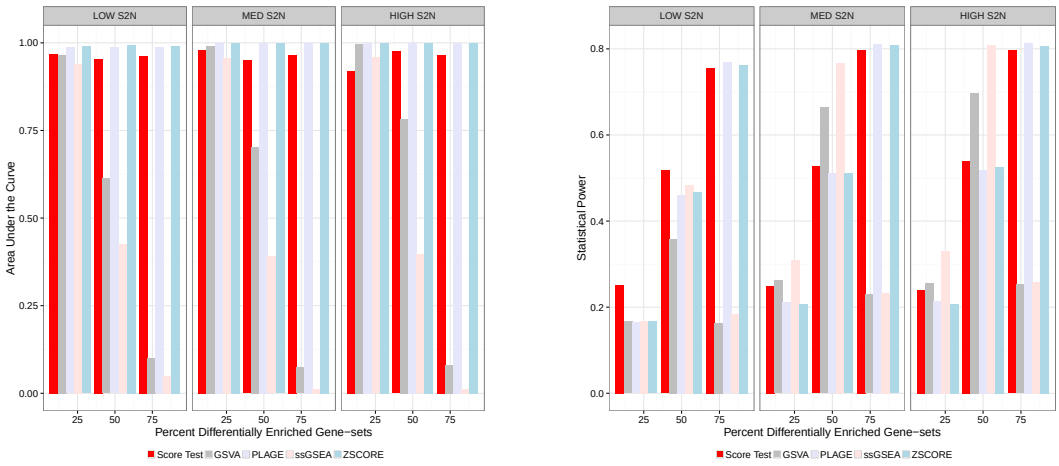
Power simulations were performed by varying the following parameters- 1) additive

| Additive Effect | $PVE_\gamma(\%)$ | $PVE_\phi(\%)$ | Gene-set Score Test | TBT Paired t-test | TBT Wilcoxon's test | Gene-level paired t-test |
|---|---|---|---|---|---|---|
| NO | 0 | 0 | 0.048 [ 0.036-0.063] | 0.047 [ 0.03-0.07 ] | 0.044 [ 0.027-0.066 ] | 0.042 [ 0.027-0.061 ] |
| NO | 0 | 7 | 0.546 [ 0.515-0.577 ] | 0.234 [ 0.196-0.275 ] | 0.198 [ 0.162-0.237 ] | 0.316 [ 0.279-0.355 ] |
| NO | 0 | 9 | 0.753 [ 0.725-0.779 ] | 0.384 [ 0.339-0.43 ] | 0.313 [ 0.271-0.358 ] | 0.465 [ 0.424-0.506 ] |
| NO | 7 | 0 | 0.408 [ 0.377-0.439 ] | 0.202 [ 0.166-0.242 ] | 0.17 [ 0.137-0.207 ] | 0.12 [ 0.095-0.149 ] |
| NO | 7 | 7 | 0.756 [ 0.728-0.782 ] | 0.413 [ 0.367-0.46 ] | 0.386 [ 0.341-0.432 ] | 0.376 [ 0.337-0.416 ] |
| NO | 6 | 9 | 0.859 [ 0.836-0.88 ] | 0.558 [ 0.511-0.604 ] | 0.515 [ 0.468-0.562 ] | 0.526 [ 0.485-0.567 ] |
| NO | 9 | 0 | 0.584 [ 0.553-0.615 ] | 0.353 [ 0.309-0.399 ] | 0.294 [ 0.253-0.338 ] | 0.178 [ 0.148-0.211 ] |
| NO | 9 | 6 | 0.806 [ 0.78-0.83 ] | 0.546 [ 0.499-0.592 ] | 0.481 [ 0.434-0.528 ] | 0.415 [ 0.375-0.456 ] |
| NO | 8 | 8 | 0.897 [ 0.876-0.915 ] | 0.655 [ 0.61-0.699 ] | 0.601 [ 0.555-0.646 ] | 0.606 [ 0.565-0.646 ] |
| YES | 0 | 0 | 0.716 [ 0.687-0.744 ] | 0.178 [ 0.144-0.216 ] | 0.167 [ 0.134-0.204 ] | 0.289 [ 0.253-0.327 ] |
| YES | 0 | 7 | 0.801 [ 0.775-0.825 ] | 0.386 [ 0.341-0.432 ] | 0.334 [ 0.291-0.379 ] | 0.483 [ 0.442-0.524 ] |
| YES | 0 | 9 | 0.878 [ 0.856-0.898 ] | 0.542 [ 0.495-0.588 ] | 0.483 [ 0.436-0.53 ] | 0.651 [ 0.611-0.69 ] |
| YES | 7 | 0 | 0.738 [ 0.71-0.765 ] | 0.414 [ 0.368-0.461 ] | 0.365 [ 0.321-0.411 ] | 0.334 [ 0.296-0.374 ] |
| YES | 7 | 7 | 0.876 [ 0.854-0.896 ] | 0.588 [ 0.541-0.634 ] | 0.538 [ 0.491-0.584 ] | 0.549 [ 0.508-0.59 ] |
| YES | 6 | 9 | 0.924 [ 0.906-0.94 ] | 0.654 [ 0.609-0.698 ] | 0.607 [ 0.561-0.652 ] | 0.66 [ 0.62-0.698 ] |
| YES | 9 | 0 | 0.763 [ 0.735-0.789 ] | 0.478 [ 0.431-0.525 ] | 0.438 [ 0.392-0.485 ] | 0.349 [ 0.31-0.389 ] |
| YES | 9 | 6 | 0.88 [ 0.858-0.899 ] | 0.654 [ 0.609-0.698 ] | 0.598 [ 0.551-0.643 ] | 0.57 [ 0.529-0.61 ] |
| YES | 8 | 8 | 0.944 [ 0.928-0.957 ] | 0.727 [ 0.684-0.767 ] | 0.682 [ 0.637-0.725 ] | 0.65 [ 0.61-0.689 ] |

Table 4.2: DE of gene-sets - Gene-set simulation results at 5% FDR with 95% confidence interval. We varied additive effect i.e. average effect of radiation on the whole transcriptome, proportion of variation explained by $\gamma$ i.e. radiation $\times$ transcripts interaction effect, and the proportion of variation explained by $\phi$ i.e. radiation $\times$ genes interaction effect. Our score test is referred to as "DE Score Test".

effect of radiation ($\beta$), 2) the proportion of variation explained by the interaction effect between radiation and transcript ($PVE_\gamma$) and 3) the proportion of variation explained by the interaction effect between radiation and gene ($PVE_\phi$) . We kept the number of transcripts and genes constant for all these simulations. The results in table 2 show that our method significantly outperforms both transcript-level and gene-level methods. More specifically, our method captures the transcript-specific variability due to radiation within each gene more efficiently than the other tests.

In our second simulation study, each Monte Carlo simulation consisted of 100 genes over 5 observations across the two phenotypes. We generated gene expression data using the same approach as described in the previous section. We simulated 10 gene-sets under both scenarios (with non-overlapping and overlapping genes) and compared the performance of our method with the other gene-set enrichment methods at the gene-level. We varied the sizes of gene-sets between 2 and 10 genes. Gene-level analysis is performed by computing the median gene expression values across all the transcripts within a gene




(a) Power simulations

(b) Area under the ROC curve

FIGURE 4.2: The performance of all the methods in detecting differentially enriched gene-sets when each gene-set is comprised of unique set of genes. A) Bar plot depicting the statistical power of each method under changing number of differentially enriched gene-sets and the mean difference in gene expression (signal-to-noise ratio) between the two phenotypes (before and after radiation). We compared our method with several gene-level tests,by computing the median gene expression values across all the transcripts within a gene. B) Bar plot depicting the area under the curve (AUC) of all the methods under the aforementioned conditions

followed by an implementation of gene set variational analysis (GSVA), Pathway Level analysis of Gene Expression (PLAGE), single sample GSEA (ssGSEA) and the combined z-score (ZSCORE). We estimated the empirical type I error rate at 5% FDR both in the presence and absence of any gene overlap among the simulated gene-sets. See supplementary methods for more details. In case on no gene overlap, we simulated 10 gene-sets with varying degrees of gene overlap (20%, 50% and 80%), and varying the signal-to-noise ratio between low, medium and high. We compared the performance of all the methods by measuring statistical power and area under the curve in case of gene-sets with no overlapping genes. In the case where gene-sets shared genes, we measured only statistical power.

Figues 4.2a and 4.2b show the performance of all the methods when the gene-sets do not share any genes. Even though, this is not a general scenario, our method is competitive with the rest of the methods. In situations where the power of our method is low (relative

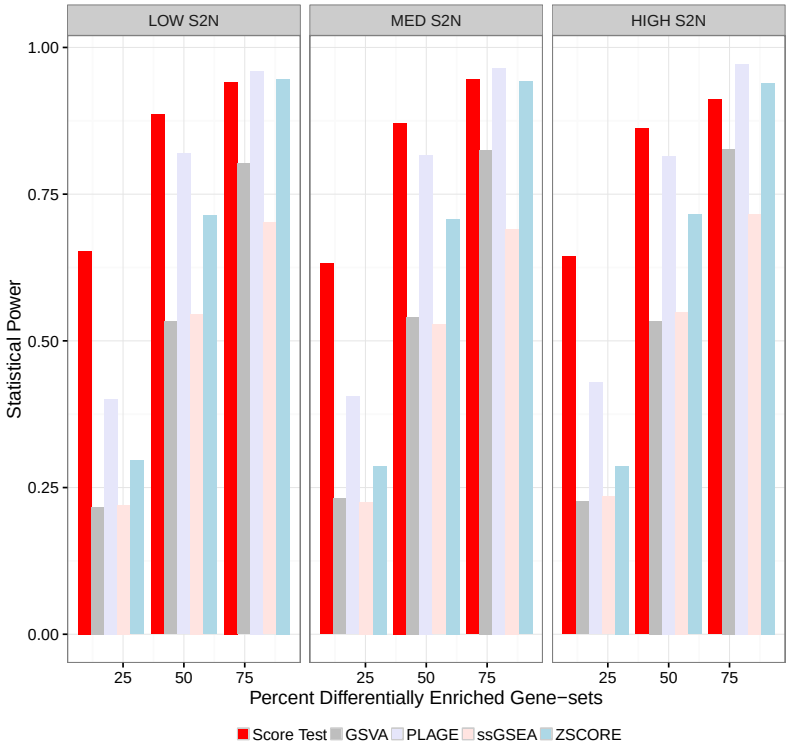



FIGURE 4.3: The performance of all the methods in detecting differentially enriched gene-sets when each gene-set is comprised of shared genes. Bar plot depicting the statistical power of each method under changing number of differentially enriched gene-sets and the mean difference in gene expression (signal-to-noise ratio; S2N) between the two phenotypes (before and after radiation). We compared our method with several gene-level tests,by computing the median gene expression values across all the transcripts within a gene.

to the other methods), the accuracy of our method is high given the AUC values. Figure 4.3 displays the performance of all the methods when the gene-sets have over-lapping genes or shared genes. This is the most common scenario and our method performs well, in terms of statistical power, in almost all cases.

### *4.3.3 Transcriptome-wide response to radiotherapy in breast tumors*

*Isoform-specific DE analysis*

Transcriptome expression profiling of the early-stage breast cancer patients before and after preoperative radiotherapy using our method has revealed many DE genes. Current

methods perform DE analysis at the gene-level and not at the transcript-level. One method performs a standard paired t-test at the transcript-level and combines the resulting $p$ values using Fisher's method (Fisher, 1934; Birnbaum, 1959). Fisher's method tests a global null hypothesis that the combined $p$ values are jointly significant. However, Fisher's method assumes that the transcript-level $p$ values for each gene are independent. Standard paired t-test followed by Fisher's method identified 11,944 genes at 5% FDR. Another most commonly used approach is to combine the gene expression values of all transcripts of a gene *a priori* by computing either the median expression values or Winsorized mean expression values (which is robust to any outliers). Paired t-tests were then run on the combined data. These two ways of combining the data identified 4,729 and 3,353 genes, respectively at 5% FDR. Our method identified a total of 12,414 DE genes at 5% FDR, which is more than the ones identified by the aforementioned methods. To assess the biological relevance of the DE genes, we performed a KEGG pathway term enrichment analysis (Kanehisa et al., 2016) for each set of results separately. KEGG pathways were considered overrepresented if a set of at least three genes from different linked regions is observed to be overrepresented with an adjusted significance level of an adjusted $p$ value $< 0.05$, calculated from a hypergeometric test (Yu et al., 2012).

| KEGG ID | Description | $p$ values | Adjusted $p$ values | $q$ value |
|---|---|---|---|---|
| hsa05200 | Pathways in cancer | 4.56E-09 | 1.34E-06 | 7.68E-07 |
| hsa04151 | PI3K-AKT signaling pathway | 3.69E-08 | 5.45E-06 | 3.11E-06 |
| hsa01100 | Metabolic pathways | 1.64E-07 | 1.61E-05 | 9.20E-06 |
| hsa04060 | Cytokine-cytokine receptor interaction | 1.21E-06 | 8.91E-05 | 5.09E-05 |
| hsa04510 | Focal adhesion | 1.59E-06 | 9.38E-05 | 5.36E-05 |
| hsa04630 | JAK-STAT signaling pathway | 3.35E-06 | 0.000164626 | 9.40E-05 |
| hsa04144 | Endocytosis | 9.88E-06 | 0.000386927 | 0.000220904 |
| hsa05166 | HTLV-I infection | 1.13E-05 | 0.000386927 | 0.000220904 |
| hsa04360 | Axon guidance | 1.24E-05 | 0.000386927 | 0.000220904 |
| hsa04210 | Apoptosis | 1.43E-05 | 0.000386927 | 0.000220904 |

Table 4.3: A list of top 10 over-represented KEGG pathways based on the functional enrichment of our DE gene list.

The results in table 3 show a list of top 10 signaling pathways that were shown be overrepresented in the dataset without any specifics on the directionality (up- or down-regulation) of the pathway deregulation. For example, PI3K-AKT signaling pathway shown in the table, a potential target for radiosensitizing cancer cells, is one the many pro-survival signaling pathways that get activated by radiation that may lead to suppression of apoptosis, initiation of DNA repair mechanisms and induction of cell-cycle arrest (Hein et al., 2014). Together with mTOR signaling pathway, PI3K-AKT are activated in many different cancers. Drugs like rapamycin, CCI-779 and RAD-001 target mTOR signaling pathway while perofisine, PX-866 target AKT pathway (LoPiccolo et al., 2008).

*Isoform-specific gene-set analysis*

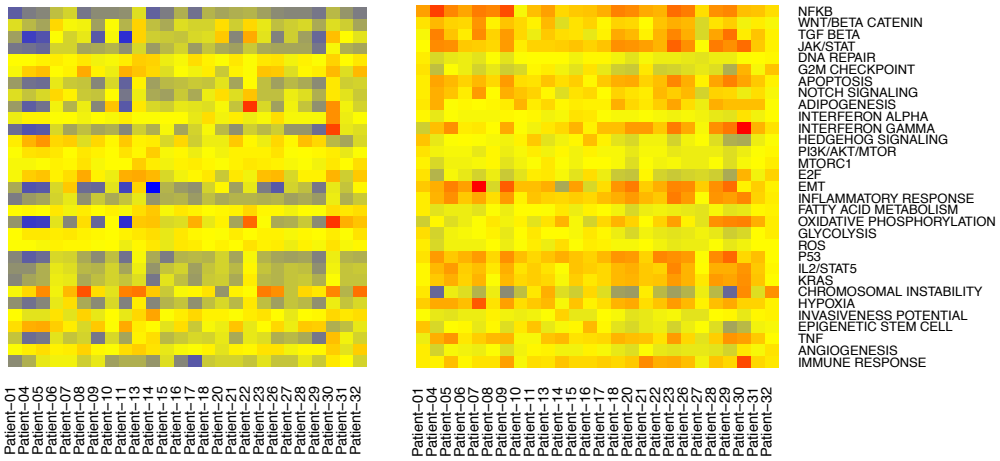
Instead of focusing on individual genes, we turned our focus on functionally related genes referred to as gene-sets or pathways and assess their behavior before and after treatment with radiation. Gene-set enrichment analysis (GSEA) and other similar methods such as Gene Set Analysis (GSA) make use of the entire gene expression profile in order to assess changes of small magnitude in functionally related genes. The aforementioned methods are supervised, which require an *a priori* knowledge of the phenotypes. In contrast, methods such as single sample GSEA (Barbie et al., 2009), GSVA (Hänzelmann et al., 2013), PLAGE (Tomfohr et al., 2005), and ZSCORE (Lee et al., 2008) are unsupervised and focus on the relative enrichment of pathways across all the samples rather than the absolute enrichment with respect to a phenotype. All of these methods work at a gene-level and require us to combine gene expression values at the transcript level before any analysis. Our method identified differentially expressed gene-sets by leveraging transcript-specific effects without having to aggregate gene expression over all the probes of a gene. On this basis, we interrogated critical radiation-associated oncogenic signaling pathways and tumor microenvironment signatures and compared the performance of our method with the rest of the methods. Many of the radiation-associated oncogenic signaling pathways

were obtained from the hallmark gene-set collection of the Molecular Signatures Database (MSigDB), which were generated by a hybrid approach that combines an automated computational procedure with manual expert curation (Liberzon et al., 2015). All of the investigated 24 oncogenic signaling pathways and 7 tumor microenvironment gene signatures were found to be statistically significant at 5% FDR by our method.

All other methods were applied at the gene-level i.e. aggregated gene expression values over all isoforms using median expression values. GSVA identified 22 gene-sets (70.9%), PLAGE identified 26 gene-sets (83.8%), ssGSEA identified 25 gene-sets (80.6%), and ZSCORE identified 27 gene-sets (87%) at 5% FDR. In order to visualize the patterns of pathway regulation, we obtained a matrix containing sample set enrichment scores of all the 31 gene-sets over all the samples using the popular GSA method. From the heat plots in figure 4.4, radiation induces a hypoxic state, enhances tumor necrosis factors and suppresses angiogenesis. Radiation-induced inflammatory pathways and immune response signatures can be targeted by therapeutics that improve the clinical outcome of radiotherapy by enhancing the radiosensitivity and decreasing any putative metabolic effects.

## 4.4 Discussion

Tailoring a patient's treatment to exploit an individual's tumor biology remains an elusive goal in cancer therapy. Similar to cytotoxic therapy, response to radiation in a given population of 'eligible' patients is markedly heterogeneous. While chemotherapy serves to address systemic disease, radiation acts as effective local therapy. In many instances, patients resistant to radiation have limited to no options to control local disease (Torres-Roca, 2012; Torres-Roca et al., 2005; Baumann et al., 2016); thus, prospectively determining tumor radiosensitivity is important to identify cohorts of patients most likely to respond and to minimize the incidence of radiation-related adverse events in patients who might not otherwise respond. Also, if the molecular underpinnings of radiation response can be elu-




(a) Pre-irradiation (b) Post-irradiation

FIGURE 4.4: Heat plot showing differentially enriched oncogenic signaling pathways and signatures of tumor microenvironment between patients before and after receiving radiotherapy. The matrix containing sample set enrichment score as computed by the GSA software were used to generate this heat plot. Red indicates a higher collective expression and blue indicates a lower collective expression of genes in that gene-set.

cidated and exploited, the radioresistance of tumors could potentially be abrogated with novel therapeutics. While many mechanisms of radiation resistance, including alterations in DNA repair mechanisms (Willers et al., 2004), upregulation of pathways regulating angiogenesis (Li et al., 2005), apoptosis (Balcer-Kubiczek, 2012) and cell cycle (Bernhard et al., 1995), have been previously described, a comprehensive evaluation of biological events to identify key oncogenic signaling events regulating radiation response, at a genomic and transcriptomic level, is largely unknown. Recent technological advances in quantifying gene expression (i.e. high-throughput sequencing assays) will allow us to interrogate whole exomes or transcriptomes with a higher precision than mRNA expression microarrays thus, overcoming the limitations in detecting and quantifying coding transcript isoforms. However, current statistical methods allow us to interrogate genes and gene-sets at the gene-level by aggregating gene expression across all possible gene isoforms thus, not taking advantage of alternative splicing mechanisms that result in multiple isoforms of the same gene.

Overall, our efforts are primarily directed to understanding two very specific aspects - 1) the effect of radiation-induced gene isoform-level variability on gene expression, oncogenic signaling pathways involved in radiation response and tumor microenvironment, and 2) the overall effect of radiation on gene expression. Currently, there are no established methods that leverage gene isoform-specific effects in order to quantify gene expression and investigate tumor biology at a higher resolution. Our methods provide an efficient framework to model transcript-specific and gene-specific effects to map biomarkers association with radiation response. The dataset used here used a high-resolution array-based platform that includes an overwhelming number of gene transcripts in the human transcriptome with >6 million probes targeting coding transcripts, exon-exon splice junctions, and non-coding transcripts. We predict that our methods will also be applicable to gene expression data quantified using RNA-Seq analysis since we make distributional assumptions that preclude their direct application to RNA-Seq count data.

Finally, our methods and analyses are only helpful in generating biological hypotheses, which require substantial verification using *in vitro* and *in vivo* model systems. Eventually, by correctly interpreting these data, we enhance our ability to accurately identify individuals most likely to be resistant to radiotherapy based on the patterns of pathway activation, which further emphasizes the need to identify novel compounds/drugs that could modulate radiation response and function as radiosensitizers.

# 5

# Conclusion

Genomic data is complex and heterogeneous. The heterogeneity exists between multiple levels and within the levels. Each level is described by a specific data type or modality. For example, the genome level is described by SNPs, copy number variants (CNVs), loss of heterozygosity (LOH), and genomic rearrangements in the form of genomic translocations, while the epigenome level is described by DNA methylation and histone modifications, which alter the accessibility to the chromatin network. Other such levels deal with mRNA expression (transcriptome level) and protein expression (proteome level) profiling. All of these levels are eventually contributing to our understanding of the phenome. Statistical methods (both frequentist and Bayesian) that analyze high throughput data from each level or modality are well developed. However, when some big consortium projects such as GTEx are delivering high resolution, multi-modal genomic data within a single individual, there is a demand for newer and more computationally efficient models. We decided to make use of linear mixed effects regression framework, given the attractiveness of and advantages associated with them. This regression framework enables us to make used to efficient score-based statistics or score test statistics. Some of the advantages of efficient score statistics are 1) they are locally most powerful tests and are guaranteed a

theoretical optimality over all other approaches in a neighborhood of the null hypothesis, 2) the model parameters are estimated only once and as such reduce the computational burden, and 3) they are robust to misspecification of random effect distribution. This increased efficiency is important, as it allows more accurate, permutation or Monte Carlo based, assessments of statistical significance and the ability to address denser marker or sequencing-based studies. The score test based framework presented here could also be used in a similar way to develop tests that accumulates rare variant contributions across genomic regions annotated to have regulatory potential. Finally, the methods we proposed are used to generate biological hypothesis, which must be followed by substantial verification using *in vitro* and *in vivo* model systems. In the end, in our eternal pursuit to understand the underlying complexity of molecular phenotypes, I am hopeful that these models play a small yet significant role.

# 6

# APPENDIX 1: Supplementary information for "Exploiting expression patterns across multiple tissues to map expression quantitative trait loci"

## 6.1 Our model

For a given gene-SNP pair, we begin with a linear mixed effects model that models expression patterns across tissues as a function of genotype, i.e.,

$$\mathbf{Y} = J\alpha + G\beta + Zu + Xv + \xi \tag{6.1}$$

where $Y$ is a $nt$-dimensional vector of expression levels in $t$ tissues and $n$ individuals, $\alpha$ is a vector of tissue-specific intercepts, $G$ is a $nt$-dimensional vector of genotypes, $\beta$ is a fixed effect of genotype across tissue, $u \sim N(0, \tau ZZ^T)$ is a vector of subject-specific random effect, $v \sim N(0, \gamma XX^T)$ is a vector of tissue-specific random effects, and $\xi \sim N(0, \epsilon I_{nt})$. The matrices $J$, $Z$ and $X$ are design matrices with $X$ being a function of genotype. $J$ is $nt \times t$ dimensional matrix denoting the design matrix for the tissue-specific intercepts. $Z$ is $nt \times nt$ design matrix for the subject-specific intercepts. $X$ is a $nt \times t$ design matrix of stacked genotypes. The parameters of interest are $\beta$ and $\gamma$; $\alpha$, $\tau$ and $\epsilon$ are

nuisance parameters.

We test the null hypothesis that $H_0 : \beta = \gamma = 0$, i.e. the variant does not affect gene expression across any of the tissues.

## 6.2 Derivation

From equation 1, the log-likelihood function of Y conditioned on the genotype is –

$$\ell(\beta, \theta) = c - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(Y - J\alpha - G\beta)^T \Sigma^{-1}(Y - J\alpha - G\beta) \tag{6.2}$$

where $\theta$ represents the vector of all the variance components involved in $\Sigma$ and c is a constant. Alternatively, under equation 1 and normality, we have

$$Y \sim N(J\alpha + G\beta, \Sigma) \quad with \quad \Sigma = \epsilon I + \tau ZZ^T + \gamma XX^T$$

From Jiming Jiang's *Linear and Generalized Linear Mixed Models and Their Applications* (Jiang, 2007) –

$$\frac{\partial \ell}{\partial \beta} = G^T\Sigma^{-1}Y - G^T\Sigma^{-1}G\beta$$

$$\frac{\partial \ell}{\partial \theta_r} = \frac{1}{2}\left\{(Y - G\beta - J\alpha)^T \Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_r}\Sigma^{-1}(Y - G\beta - J\alpha) - \mathrm{Tr}\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_r}\right)\right\}$$

where $\theta_r$ is the $r^{th}$ component of $\theta$ such that $\theta \in (\tau, \gamma, \epsilon)$.

$$E\left[\frac{\partial^2 \ell_i}{\partial \beta \partial \beta^T}\right] = -G^T\Sigma^{-1}G$$

$$E\left[\frac{\partial^2 \ell_i}{\partial \beta \partial \theta_r}\right] = 0$$

$$E\left[\frac{\partial^2 \ell_i}{\partial \theta_r \partial \theta_s}\right] = -\frac{1}{2}\mathrm{Tr}\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_r}\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_s}\right)$$

### 6.2.1 Score test

Let the parameters of interest be $\psi = (\beta, \gamma)^T$ and the nuisance parameters be $\eta = (\alpha, \tau, \epsilon)^T$. The following is constructed under the null ($H_0$)

$$U_\psi = \begin{pmatrix} \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \gamma} \end{pmatrix} - \begin{bmatrix} I_{\beta\alpha} & I_{\beta\tau} & I_{\beta\epsilon} \\ I_{\gamma\alpha} & I_{\gamma\tau} & I_{\gamma\epsilon} \end{bmatrix} \begin{bmatrix} I_{\alpha\alpha} & I_{\alpha\tau} & I_{\alpha\epsilon} \\ I_{\tau\alpha} & I_{\tau\tau} & I_{\tau\epsilon} \\ I_{\epsilon\alpha} & I_{\epsilon\tau} & I_{\epsilon\epsilon} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial l}{\partial \alpha} \\ \frac{\partial l}{\partial \tau} \\ \frac{\partial l}{\partial \epsilon} \end{bmatrix}$$

Some algebra will result in the following –

$$U_\beta = (G - \bar{G})^T \hat{\Sigma}_n^{-1} (Y - J\hat{\alpha}) \tag{6.3}$$

and

$$U_\gamma = \frac{1}{2} (Y - J\hat{\alpha})^T \hat{\Sigma}_n^{-1} XX^T \hat{\Sigma}_n^{-1} (Y - J\hat{\alpha}) \tag{6.4}$$

$$\begin{aligned} U_\psi &= \left( \boldsymbol{a_\beta} U_\beta^2 + \boldsymbol{a_\gamma} U_\gamma \right) \\ &= (Y - J\hat{\alpha})^T \hat{\Sigma}_n^{-1} \left[ a_\beta (G - \bar{G}) (G - \bar{G})^T + a_\gamma \frac{1}{2} XX^T \right] \hat{\Sigma}_n^{-1} (Y - J\hat{\alpha}) \end{aligned} \tag{6.5}$$

### 6.2.2 Missing response data

Let $Y_i = \left\{ Y_i^o, Y_i^m \right\}$ with $Y_i^o$ the observed part and $Y_i^m$ the missing part. Also, let $R_{i,j} = 1$ if $Y_{i,j}$ is observed and $R_{i,j} = 0$ otherwise. Assume that all the explanatory variables are completely observed. $\theta$ and $\psi$ describe the measurement and missingness, respectively.

$$f(Y^o, R|\theta, \psi) = \int f(Y^o, Y^m|\theta) f(R|Y^o, Y^m, \psi) dY^m$$

Assuming that the data are missing at random (MAR),

$$f\left(Y^o, R|\theta,\psi\right) = \int f\left(Y^o, Y^m|\theta\right) f\left(R|Y^o,\psi\right) dY^m$$

$$= f\left(R|Y^o, Y^m,\psi\right) \int f\left(Y^o, Y^m|\theta\right) dY^m$$

$$= f\left(R|Y^o,\psi\right) f\left(Y^o|\theta\right)$$

If the parameter space of $\left(\theta',\psi'\right)'$ is the product of the parameter spaces of $\theta$ and $\psi$ (separability condition), then the inference is based on the observable data only (ignorability) (Verbeke and Molenberghs, 2000; Little and Rubin, 2002).

If $x = \left[x_1, x_2\right]$ and $x \sim N\left(x,\mu,\Sigma\right)$ and $x_1$ constitute gene expression data available for samples with all the tissues/groups while $x_2$ constitutes gene expression data for samples with depleted tissues/groups. The multivariate gaussian theorem states that the marginal distribution of $x_1$ and $x_2$ are also normal with mean vector $\mu_i$ and covariance matrix $\Sigma_{ii}\left(i = 1, 2\right)$, respectively. The conditional distribution of $x_i$ given $x_j$ is also normal with mean vector such that $\mu_{i|j} = \mu_i + \Sigma_{ij}\Sigma_{ij}^{-1}\left(x_j - \mu_j\right)$ and $\Sigma_{i|j} = \Sigma_{jj} - \Sigma_{ij}^T\Sigma_{ij}^{-1}\Sigma_{ij}$.

The joint density of $x$ is given by

$$L_n\left(x_1, x_2\right) = \prod_{i=1}^{n} \left(2\pi\right)^{-\frac{n}{2}} \mid \Sigma_i \mid^{-\frac{1}{2}} exp\left[-\frac{1}{2} Q\left(x_1, x_2\right)\right]$$

where

$$Q\left(x_1, x_2\right) = \left(x-\mu\right)^T \Sigma^{-1}\left(x-\mu\right)$$

After some algebra, we can show that the marginal distribution of $x_1$ can be written as

$$f_1\left(x_1\right) = \int f\left(x_1, x_2\right) dx_2 = \frac{1}{\left(2\pi\right)^{\frac{p}{2}} \mid \Sigma_{11} \mid^{\frac{1}{2}}} exp\left[-\frac{1}{2}\left(x_1-\mu_1\right)^T \Sigma_{11}^{-1}\left(x_1-\mu_1\right)\right]$$

...and the conditional distribution of $x_2$ given $x_1$ is given by

$$f_{2|1}(x_1, x_2) = \frac{f(x_1, x_2)}{f(x_1)}$$
$$= \frac{1}{(2\pi)^{\frac{q}{2}} \mid A \mid^{\frac{1}{2}}} exp\left[-\frac{1}{2}(x_2 - b)^T A^{-1}(x_2 - b)\right]$$

where $b = \mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1}(x_1 - \mu_1)$ and $A = \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}$.

In this way, we can show that the observed data likelihood has the exact same model form as the full data likelihood.

## 6.3 Variance-covariance of $U_\beta^2$ and $U_\gamma$

We have

$$\mathbf{Y} = J\alpha + G\beta + Zu + Xv + \xi \qquad Y \sim N(J\alpha + G\beta, \Sigma) \tag{6.6}$$

From section 2.1, at global null i.e. $H_0 : \beta = 0; \gamma = 0$, we have

$$U_\gamma = \frac{1}{2} Y^T \Sigma_n^{-1} X X^T \Sigma_n^{-1} Y \tag{6.7}$$

where $\Sigma = \hat{\epsilon} I + \hat{\tau} Z Z^T$, $\hat{\tau}$ and $\hat{\epsilon}$ are the maximum likelihood estimates of the individual-specific and tissue-specific random effects. Using the theory of quadratic forms (Jiang, 2007), estimated variance of $U_\gamma$ under the null is given by

$$Var_{H_0}(U_\gamma) = 2\,\mathrm{Tr}\left[\left(\Sigma^{-1}\frac{1}{2} X X^T \Sigma^{-1}\right)\Sigma\left(\Sigma^{-1}\frac{1}{2} X X^T \Sigma^{-1}\right)\Sigma\right] \tag{6.8}$$

Similarly, from section 2.1,

$$U_\beta^2 = Y^T \Sigma^{-1}(G - \bar{G})(G - \bar{G})^T \Sigma^{-1} Y \tag{6.9}$$

From the theory of quadratic forms (Jiang, 2007), estimated variance of $U_\beta^2$ under the null is

$$Var_{H_0}\left(U_\beta^2\right) = 2\,\mathrm{Tr}\left[\left(\Sigma^{-1}\left(G-\bar{G}\right)\left(G-\bar{G}\right)^T\Sigma^{-1}\right)\Sigma\left(\Sigma^{-1}\left(G-\bar{G}\right)\left(G-\bar{G}\right)^T\Sigma^{-1}\right)\Sigma\right] \tag{6.10}$$

$U_\beta^2$ and $U_\gamma$ share the same $\epsilon$. Again, from the theory of the quadratic forms, the covariance between $U_\beta^2$ and $U_\gamma$ is

$$Cov_{H_0}\left(U_\beta^2, U_\gamma\right) = 2\,\mathrm{Tr}\left[\left(\Sigma^{-1}\left(G-\bar{G}\right)\left(G-\bar{G}\right)^T\Sigma^{-1}\right)\Sigma\left(\Sigma^{-1}\frac{1}{2}XX^T\Sigma^{-1}\right)\Sigma\right] \tag{6.11}$$

### *6.3.1 Optimal weights for minimum variance linear combination*

Let $a = (a_\beta, a_\gamma)^T$, $U_\psi = \left(U_\beta^2, U_\gamma\right)$, and $V_\psi = Var\left(U_\psi\right)$. We want to find the minimum variance linear combination $a^T V_\psi$, subject to the constraint that $a_\beta + a_\gamma = 1$ or $a^T\mathbf{1} = 1$. Specifically, we wish to minimize $a^T V_\psi a$.

Using Lagrangian multipliers to perform constrained optimization, we see that

$$\mathcal{L}\left(a|\lambda\right) = a^T V_\psi a - \lambda\left(a^T\mathbf{1} - 1\right)$$

where $\mathbf{1} = \begin{bmatrix}1 & 1\end{bmatrix}^T$ and $\lambda > 0$.

$$\frac{\partial}{\partial\left(a^T, \lambda\right)} = \left(a^T V_\psi a - \lambda\left(a^T\mathbf{1} - 1\right)\right) = 0$$

From the above equations, we have the following system of equations–

$$2V_\psi a - \lambda\mathbf{1} = 0 \qquad\qquad a^T\mathbf{1} = \mathbf{1}^T a = 1$$

$$a = \frac{\lambda}{2}V_\psi^{-1}\mathbf{1}$$

and

$$1 = a\mathbf{1}^T = \frac{\lambda}{2}\mathbf{1}^T V_\psi^{-1}\mathbf{1}$$

so that,

$$\lambda = \frac{2}{\mathbf{1}^T V_\psi^{-1}\mathbf{1}}$$

This gives our optimal weights –

$$a = \frac{V_\psi^{-1}\mathbf{1}}{\mathbf{1}^T V_\psi^{-1}\mathbf{1}}$$

$$a_\gamma = \frac{var\left(U_\beta^2\right) - cov\left(U_\beta^2, U_\gamma\right)}{var\left(U_\beta^2\right) + var\left(U_\gamma\right) - 2cov\left(U_\beta^2, U_\gamma\right)} \tag{6.12}$$

and

$$a_\beta = \frac{var\left(U_\gamma\right) - cov\left(U_\beta^2, U_\gamma\right)}{var\left(U_\beta^2\right) + var\left(U_\gamma\right) - 2cov\left(U_\beta^2, U_\gamma\right)} \tag{6.13}$$

## 6.4 MetaTissue method

MetaTissue (MT) method was proposed by Sul *et al* that jointly models all tissues by utilizing a meta-analysis by extending it to a mixed-model framework. A mixed model is used to account for the correlation of expression between tissues, and perform meta-analysis to combine results from multiple tissues. The following model description is from the original paper.

Consider the following mixed-model –

$$Y = 1\alpha + X_j\beta + u + e$$

where $u \sim N\left(0, \sigma_\mu^2 D\right)$ and $e \sim (0, \sigma_e^2 I)$ and $X_j$ is the matrix denoting SNP $j$ for $T$ tissues. The variances are estimated using *EMMA* and $\hat{\beta}$s are jointly estimated using the following equation –

$$\hat{\beta} = \left(X_j' \Sigma^{-1} X_j\right)^{-1} X_j' \Sigma^{-1} Y$$

Given the $\hat{\beta} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_T\right)$, information from multiple tissues is combined by applying meta-analysis to $\hat{\beta}$.

### *6.4.1 Fixed-effects model*

A statistic of FE and its distribution under the null hypothesis are –

$$S_{FE} = \frac{\sum_{i=1}^{T} V_i^{-1} B_i}{\sqrt{\sum_{i=1}^{T} V_i^{-1}}} \sim N(0, 1)$$

where $B_1 \ldots B_T$ and $V_1, \ldots, V_T$ are the estimates of effect-size and the standard error of $B_i$ in T tissues. Let $\mu$ be the unknown true effect size and so the null hypothesis of FE is $\mu = 0$ or in other words the effect size in all tissues is zero. A p-value of $S_{FE}$ is obtained from the standard normal distribution.

Under the null hypothesis, $\frac{\hat{\beta}}{\sqrt{var\left(\hat{\beta}\right)}}$ will approximately follow *t-distribution* with $k$ degrees of freedom.

$$p_t = 2\left(1 - \phi_{t(k)}\left(\frac{|\hat{\beta}|}{\sqrt{var\left(\hat{\beta}\right)}}\right)\right)$$

### 6.4.2 *Random-effects model*

The general assumption behind the random-effects model is that the effect size of a variant is different among datasets and follows a probability distribution with mean $\mu$ and variance $\tau^2$. The $H_0$ is the same as that of the fixed-effects model – $H_0 : \mu = 0$. The statistic for the random effects model is defined as –

$$S_{RE} = \sum log\left(\frac{V_i}{V_i + \hat{\tau^2}}\right) + \sum \frac{B_i^2}{V_i} - \sum \frac{(B_i - \hat{\mu})^2}{V_i + \hat{\tau^2}}$$

where $\hat{\mu}$ and $\hat{\tau^2}$ are estimated mean and variance of the effect size, and the *maximum likelihood estimates* of the two parameters that are iteratively calculated using *Hardy and Thompson approach* or some other iterative approach. The statistic follows a half and half mixture of $\chi_0^2$ and $\chi_1^2$ under the null.

## 6.5 eQTL-BMA method

eQTL-BMA, proposed by Flutre *et al*, investigates whether the SNP is an eQTL in any tissue, and, if so, in which tissues. The primary model is

$$y_{si} = \mu_s + \beta_s g_i + \epsilon_{si} \qquad \epsilon_{si} \sim N\left(0, \sigma_s^2\right)$$

where $y_{si}$ denotes gene expression vector in tissue $s$, for $i^{th}$ individual, $\mu_s$ is the mean expression level of the gene in tissue $s$, $\beta_s$ is the effect of the gene on the genotype in tissue $s$ and $g_i$ is the genotype of individual $i$ coded as 0,1 or 2 copies of the reference allele. Statistical inference is made on $\gamma$, a binary variable (called configuration) whose status indicates the presence or absence of an eQTL. The length of $\gamma$ depends on the number of tissues. Null hypothesis is indicated by $\gamma = \{0, \ldots, 0\}$ and any other combination is considered an alternative hypothesis. The statistical inference on $\gamma$ is done using Bayes Factors such that –

$$BF_{\gamma} = \frac{P\left(data|\gamma\right)}{P\left(data|H_0\right)}$$

In order to account for many possible alternatives, the overall strength of evidence against at the candidate SNP is obtained by "Bayesian Model Averaging" (BMA), which involves averaging over the possible alternative configurations, weighting each by its prior probability, $\eta_{\gamma} = P\left(\gamma|H_0 = FALSE\right)$:

$$BF_{BMA} = \frac{P\left(data|H_0 = false\right)}{P\left(data|H_0 = true\right)} = \sum_{\gamma \neq 0} \eta_{\gamma} BF_{\gamma}$$

Large values of $BF_{BMA}$ indicate a strong evidence against the $H_0$. Another flavor $BF_{BMA}^{HM}$ indicates a hierarchical model where the hyperparameters are estimated from the data (data-driven approach). $BF_{BMAlite}$ is a more computationally scalable version of the above flavors as it averages the test statistic over $S + 1$ configurations. In general, eQTL-BMA method does not scale well with increasing number of tissues because the number of terms in the sum of above equation is $2^S - 1$.

In the presence of a strong evidence against the $H_0$, posterior probability on each configuration indicating that the SNP is an eQTL in tissue $s$ is computed by –

$$P\left(\gamma = TRUE|data, H_0 = false\right) = \frac{\eta_{\gamma} BF_{\gamma}}{\sum\limits_{\gamma=0} \eta_{\gamma} BF_{\gamma}}$$

A frequentist interpretation to Bayes Factors computed by eQTL-BMA is given by performing adaptive permutations at the gene-level at a given FDR.

## 6.6 A note on statistical software

Our simulations were run to compare the statistical power (and type I error rate) between our method, eQTL-BMA, MetaTissue and Tissue-by-Tissue methods.

eQTL-BMA software is available for download at `https://github.com/timflutre/eqtlbma`. In order to expedite the analysis, we ran $BF_{BMAlite}$ version of the software, 1,000 adaptive permutations (using trick 1) to obtain the gene-level $p$ value. These $p$ values were then extracted from *output.log-jointPermPvals.txt.gz* file for further analysis. We used eQTL-BMA software version 1.2 to perform all the analyses. In case of the real data analyses, we increased the number of permutations to the author recommended 10,000.

MetaTissue model software is made available at `http://genetics.cs.ucla.edu/metatissue/`. We used default setting for each step described by the author on the website. We used MetaTissue software version 0.3 for our analyses.

# 7

# APPENDIX 2: Supplementary information for "Mapping eQTL by leveraging multiple tissues and DNA methylation"

## 7.1 Our model

For a given gene-SNP pair, gene expression is modeled as a function of genotype and methylation -

$$Y = J\alpha + G\beta + M\lambda + MG\phi + Au + Bv + Cw + Dx + \xi \tag{7.1}$$

where $Y$ is a $nt$-dimensional vector of expression levels in $t$ tissues and $n$ individuals, $\alpha$ is a vector of tissue-specific intercepts, $G$ is a $nt$-dimensional vector of genotypes, $\beta$ is a fixed effect of genotype across tissue, $\lambda$ is an overall methylation-specific fixed effect, $\phi$ is genotype $\times$ methylation interaction effect (fixed effect), $u \sim N(0, \tau AA^T)$ is a vector of subject-specific random effect, $v \sim N(0, \gamma BB^T)$ is a vector of tissue-specific random effects, $w \sim N(0, \delta CC^T)$ is a vector of tissue-specific random effects that describe the interaction effect between genotype and methylation is a vector of random effects describing the interaction between genotype, methylation and tissue, $x \sim N(0, \theta DD^T)$ is a vector

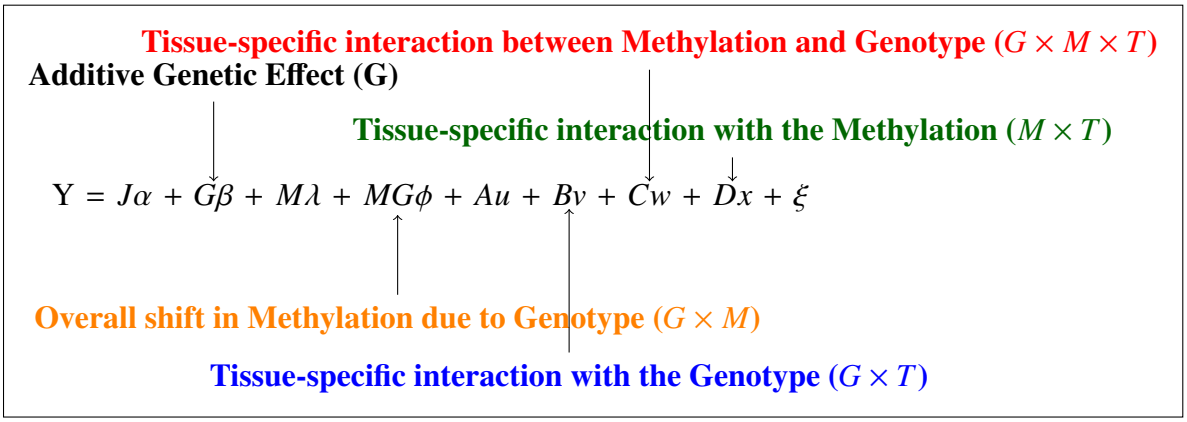



FIGURE 7.1: Description of all the terms in our model

of tissue-specific random effects describing methylation effects and $\xi \sim N(0, \epsilon I_{nt})$. The matrices $J$, $A$, $B$, $C$, and $D$ are design matrices with $B$ being a function of genotype, $C$ is a function of both genotype and methylation data and finally, $D$ is a function of just the methylation data. $J$ is $nt \times t$ dimensional matrix denoting the design matrix for the tissue-specific intercepts. $A$ is $nt \times nt$ design matrix for the subject-specific intercepts. $B$ is a $nt \times t$ design matrix of stacked genotypes. $C$ is a $nt \times t$ design matrix of stacked (product of) tissue-specific methylation and genotype data. $D$ is $nt \times t$ design matrices of stacked tissue-specific methylation data.

Parameters of interest are $\gamma$, $\delta$, $\beta$ and $\phi$; $\alpha$, $\lambda$, $\tau$, $\theta$ and $\epsilon$ are nuisance parameters. We test the null hypothesis that $H_0 : \beta = \phi = \gamma = \delta = 0$, i.e. the variant does not affect gene expression across any of the tissues. Our joint score test will test for the effect of genotype on 1) an overall shift in the gene expression, 2) tissue-specific interaction ($G \times T$), 3) overall methylation ($G \times M$), and 4) tissue-specific methylation ($G \times M \times T$)

## 7.2 Individual components of our joint score test statistic

### *7.2.1 Additive genetic effect on the gene expression under the global null*

$$\mathrm{Y} = J\alpha + G\beta + M\lambda + MG\phi + Au + Bv + Cw + Dx + \xi$$

**Additive genetic effect**

The score test for the fixed effect $\beta$ takes the following form under the global null –

$$U_\beta = (G - \bar{G})^T \Sigma_n^{-1} \hat{Y} \tag{7.2}$$

where $(G - \bar{G})$ is a vector of mean-centered genotypes for all individuals and $\hat{Y} = \left(Y - J\hat{\alpha} - M\hat{\lambda}\right)$. $U_\beta$ is a scalar quantity in a linear form and follows a $\chi^2_1$ distribution.

Squaring $U_\beta$ gives us the following quadratic form, which will be useful while aggregating all the score test statistics.

$$U_\beta^2 = \hat{Y}^T \Sigma_n^{-1} (G - \bar{G}) (G - \bar{G})^T \Sigma_n^{-1} \hat{Y} \tag{7.3}$$

### *7.2.2 The effect of SNP on gene expression via differential methylation patterns under the global null ($G \times M$ effect)*

$$\mathrm{Y} = J\alpha + G\beta + M\lambda + MG\phi + Au + Bv + Cw + Dx + \xi$$

**Methylation and genotype interaction** ($G \times M$)

The score test for the interaction effect $\phi$ takes the following form under the global null –

$$U_\phi = \left(MG - \overline{MG}\right)^T \Sigma_n^{-1} \hat{Y} \tag{7.4}$$

where $\Sigma_n = diag\,(\Sigma, \cdots, \Sigma)$ is an $nt \times nt$ block diagonal matrix and $\hat{Y} = \left(Y - J\hat{\alpha} - M\hat{\lambda}\right)$. $U_\phi$ is a scalar quantity in a linear form and follows a $\chi^2_1$ distribution. Squaring $U_\phi$ gives us the following quadratic form, which will be useful while aggregating all the score test statistics.

$$U^2_\phi = \hat{Y}^T \Sigma_n^{-1} \left(MG - \overline{MG}\right) \left(MG - \overline{MG}\right)^T \Sigma_n^{-1} \hat{Y} \tag{7.5}$$

### 7.2.3 *The tissue-specific effect due to genotype on the gene expression under the global null ($G \times T$ effect)*

$$\mathrm{Y} = J\alpha + G\beta + M\lambda + MG\phi + Au + \underset{\uparrow}{Bv} + Cw + Dx + \xi$$

**Genotype and tissue interaction** ($G \times T$)

The score for the variance component $\gamma$ under the global null is –

$$\frac{1}{2}\left\{\hat{Y}^T \Sigma_n^{-1} BB^T \Sigma_n^{-1} \hat{Y} - Tr\left(\Sigma_n^{-1} BB^T\right)\right\} \tag{7.6}$$

where $\Sigma_n = diag\,(\Sigma, \cdots, \Sigma)$ is an $nt \times nt$ block diagonal matrix and $\hat{Y} = \left(Y - J\hat{\alpha} - M\hat{\lambda}\right)$. As the *trace* term does not depend on the data, we use the first term to construct the test statistic.

$$U_\gamma = \frac{1}{2} \hat{Y}^T \Sigma_n^{-1} BB^T \Sigma_n^{-1} \hat{Y} \tag{7.7}$$

$U_\gamma$ follows a mixture of chi-square distribution and the $p$ value is approximated using a scaled $\chi^2$ distribution (the Satterthwaite method) by matching the first two moments as $U_\gamma \sim \kappa \chi^2_\nu$ where $\kappa = \frac{Var(U_\gamma)}{2E[U_\gamma]}$ and $\nu = \frac{2E[U_\gamma]^2}{Var(U_\gamma)}$.

### 7.2.4 *Latent effect (masking effect) of SNP on gene expression via tissue-specific methylation patterns ($G \times M \times T$ effect)*

$$\mathrm{Y} = J\alpha + G\beta + M\lambda + MG\phi + Au + Bv + \underset{\uparrow}{Cw} + Dx + \xi$$

**Methylation, genotype and tissue interaction** ($G \times M \times T$)

The score for the variance component $\boldsymbol{\delta}$ under the global null is –

$$\frac{1}{2}\left\{\hat{Y}^T\Sigma_n^{-1}CC^T\Sigma_n^{-1}\hat{Y} - Tr\left(\Sigma_n^{-1}CC^T\right)\right\} \tag{7.8}$$

where $\Sigma_n = diag\left(\Sigma, \cdots, \Sigma\right)$ is an $nt \times nt$ block diagonal matrix and $\hat{Y} = \left(Y - J\hat{\alpha} - M\hat{\lambda}\right)$. As the *trace* term does not depend on the data, we use the first term to construct the test statistic.

$$U_\delta = \frac{1}{2}\hat{Y}^T\Sigma_n^{-1}CC^T\Sigma_n^{-1}\hat{Y} \tag{7.9}$$

$U_\delta$ follows a mixture of chi-square distribution, the $p$ value can be approximated using a scaled $\chi^2$ distribution by matching the first two moments as $U_\delta \sim \kappa\chi^2_\nu$ where $\kappa = \frac{Var(U_\delta)}{2E[U_\delta]}$ and $\nu = \frac{2E[U_\delta]^2}{Var(U_\delta)}$.

### 7.2.5 *Joint score test statistic*

We propose a weighted sum of the above components to arrive at our joint score test statistic, $U_\zeta$. Since $U_\beta$ and $U_\phi$ are linear in $Y$ while $U_\gamma$ and $U_\delta$ are quadratic, we propose the following rule to combine them –

$$\begin{aligned} U_\zeta &\equiv \left(\boldsymbol{a_\beta}U_\beta^2 + \boldsymbol{a_\phi}U_\phi^2 + \boldsymbol{a_\gamma}U_\gamma + \boldsymbol{a_\delta}U_\delta\right) \\ &\equiv \hat{Y}^T\hat{\Sigma}_n^{-1}\left[a_\beta\left(G - \bar{G}\right)\left(G - \bar{G}\right)^T + a_\phi\left(MG - \overline{MG}\right)\left(MG - \overline{MG}\right)^T + a_\gamma\frac{1}{2}BB^T + a_\delta\frac{1}{2}CC^T\right]\hat{\Sigma}_n^{-1}\hat{Y} \end{aligned} \tag{7.10}$$

where $a_\beta$, $a_\phi$, $a_\gamma$ and $a_\delta$ are scalar constants chosen to minimize the variance of $U_\psi$. Under the null, $U_\psi$ is distributed as a mixture of chi-square random variables. We use Satterthwaite method (Satterthwaite, 1946) to approximate the $p$ values from a scaled $\chi^2$ distribution by matching the first two moments as $U_\psi \sim \kappa\chi^2_\nu$ where $\kappa = \frac{Var(U_\psi)}{2E[U_\psi]}$ and $\nu = \frac{2E[U_\psi]^2}{Var(U_\psi)}$.

Our joint score test will test for the effect of genotype on 1) an overall shift in the gene expression, 2) tissue-specific interaction ($G \times T$), 3) overall methylation ($G \times M$), and 4) tissue-specific methylation ($G \times M \times T$)

## 7.3 Evaluating our joint score test statistic

For a positive integer $t$ that represents number of tissues, if $\mathbf{1}$ denotes a column vector of $t$ ones and $\mathbb{I}$ denotes the corresponding $t \times t$ diagonal matrix, following the $t$-variate normal law denoted by $N_t[\mu, \Sigma]$ with mean $\mu \in \mathbb{R}^t$ and variance $\Sigma \in \mathbb{R}^{t\times t}$, expression levels of a target gene $j$ at a single locus by using the following vectorized form of the linear mixed model –

$$y_{ij} = \alpha_j + \mathbf{1}\beta_j g_i + \mathbf{1}\lambda_j m_{ij} + \mathbf{1}\phi_j m_{ij} g_i + \mathbf{1}a_i + b_j g_i + c_j m_{ij} g_i + d_j m_{ij} + \xi_{ij} \qquad \xi_{ij} \overset{i.i.d.}{\sim} N(0, \epsilon\mathbb{I}) \tag{7.11}$$

where $y_{ij}$ is a $t\times 1$ vector of gene expression data, $\alpha_t$ is the tissue-specific intercept ($\alpha_t \in \mathbb{R}^t$), $\beta_j$ describes the main additive genotypic effect ($\beta_j \in \mathbb{R}^1$), $\lambda_j$ describes the overall effect due to methylation ($\lambda_j \in \mathbb{R}^1$), $\phi$ describes the interaction effect between the overall methylation and genotype ($\phi_j \in \mathbb{R}^1$), $g_i$ is the value of a bi-allelic genotype such that $g \in (0, 1, 2)$ represents the number of copies of the minor allele. The random effect $b_j \in \mathbb{R}^t$ represents tissue-specific effect of the genotype, $c_j \in \mathbb{R}^t$ represents tissue-specific interaction effect between methylation and genotype, $d_j \in \mathbb{R}^t$ represents tissue-specific methylation effect, and $a_i \in \mathbb{R}^1$ is a subject-specific random intercept. We assume that all the random effects are independent and that $a_i \sim N_1(0, \tau)$, $b_j \sim N_t(0, \gamma\mathbb{I})$, $c_j \sim N_t(0, \delta\mathbb{I})$ and $d_j \sim N_t(0, \theta\mathbb{I})$.

Methylation data for 5 tissues was generated independently from a multivariate normal distribution with mean zero and positive definite variance-covariance matrix.

Under the global null, the reduced model is –

$$y_{ij} = \alpha_j + \mathbf{1}\lambda_j m_{ij} + \mathbf{1}a_i + d_j m_{ij} + \xi_{ij} \qquad \xi_{ij} \overset{i.i.d.}{\sim} N(0, \epsilon\mathbb{I}) \tag{7.12}$$

We use 1,000 data replicates to evaluate type I error and power calculations. Simulations were performed by varying $\beta$, the proportion of variance explained by the random effect describing the interaction between genotype and tissue, $G \times T = PVE_\gamma \equiv \left(\frac{\gamma}{\theta+\tau+\epsilon+\gamma+\delta}\right)$, and the proportion of variance explained by the random effect describing the interaction between genotype, methylation and tissue, $G \times M \times T = PVE_\delta \equiv \left(\frac{\delta}{\theta+\tau+\epsilon+\gamma+\delta}\right)$. A linear mixed effects model was fit using the package *lme4* (Bates et al., 2014a) in the statistical environment R (R Core Team) (Ihaka and Gentleman, 1996).

## 7.4 Gibbs et al Data Preprocessing

The following preprocessing is done on Gibbs *et al* data (Gibbs et al., 2010) comprising of 150 samples from four regions of normal brain (cerebellum, frontal cortex, pons and temporal cortex).

### *7.4.1 Genotype data*

The genotype data is recoded into a SNP matrix of values 0, 1 and 2 representing minor allele counts. Samples with African and Asian ancestry were removed from the analysis. These SNPs were filtered on the missing-ness of the individual data and the SNP data (excluded SNPs with missing data), followed by MAF (included SNPs with MAF $\geq 0.05$)and Hardy-Weinberg equilibrium (HWE; p-values $\leq 0.001$) in the same order using PLINK (Purcell et al., 2007) software. SNPs with missing values were removed from the analysis. We ended with 400,097 SNPs after preprocessing.

Table 7.1: Table comparing the statistical power of the joint score test statistic, $U_\psi$ and the contributions from its main components, $U^2_\beta$, $U^2_\phi$, $U_\gamma$ and $U_\delta$, all under the global null. These data were generated from 1,000 simulations run on 100 individuals and five tissues with genotypes generated at a common variant allele frequency (MAF = 0.3).

| G Effect | $G \times M$ Effect | $PVE_{G\times M\times T}$ | $PVE_{G\times T}$ | $U^2_{\beta\|H_0}$ | $U^2_{\phi\|H_0}$ | $U_{\gamma\|H_0}$ | $U_{\delta\|H_0}$ | $U_{\psi\|H_0}$ |
|---|---|---|---|---|---|---|---|---|
| NO | NO | 0 | 0 | 0.058 | 0.045 | 0.061 | 0.053 | 0.06 |
| NO | NO | 0 | 5 | 0.071 | 0.055 | 0.278 | 0.052 | 0.161 |
| NO | NO | 0 | 8 | 0.092 | 0.064 | 0.602 | 0.062 | 0.427 |
| NO | NO | 5 | 0 | 0.053 | 0.151 | 0.047 | 0.28 | 0.173 |
| NO | NO | 5 | 5 | 0.079 | 0.153 | 0.294 | 0.288 | 0.325 |
| NO | NO | 5 | 8 | 0.107 | 0.143 | 0.641 | 0.274 | 0.571 |
| NO | NO | 8 | 0 | 0.055 | 0.251 | 0.072 | 0.549 | 0.383 |
| NO | NO | 8 | 5 | 0.081 | 0.255 | 0.312 | 0.622 | 0.585 |
| NO | NO | 8 | 8 | 0.107 | 0.263 | 0.645 | 0.604 | 0.734 |
| NO | YES | 0 | 0 | 0.058 | 0.883 | 0.039 | 0.674 | 0.171 |
| NO | YES | 0 | 5 | 0.08 | 0.884 | 0.28 | 0.696 | 0.385 |
| NO | YES | 0 | 8 | 0.101 | 0.888 | 0.629 | 0.674 | 0.616 |
| NO | YES | 5 | 0 | 0.047 | 0.825 | 0.061 | 0.772 | 0.381 |
| NO | YES | 5 | 5 | 0.065 | 0.844 | 0.314 | 0.751 | 0.525 |
| NO | YES | 5 | 8 | 0.102 | 0.834 | 0.611 | 0.747 | 0.725 |
| NO | YES | 8 | 0 | 0.071 | 0.762 | 0.072 | 0.83 | 0.573 |
| NO | YES | 8 | 5 | 0.084 | 0.76 | 0.309 | 0.837 | 0.677 |
| NO | YES | 8 | 8 | 0.099 | 0.719 | 0.579 | 0.848 | 0.826 |
| YES | NO | 0 | 0 | 0.287 | 0.05 | 0.054 | 0.045 | 0.208 |
| YES | NO | 0 | 5 | 0.308 | 0.058 | 0.295 | 0.055 | 0.357 |
| YES | NO | 0 | 8 | 0.322 | 0.059 | 0.648 | 0.056 | 0.588 |
| YES | NO | 5 | 0 | 0.303 | 0.127 | 0.053 | 0.275 | 0.355 |
| YES | NO | 5 | 5 | 0.301 | 0.147 | 0.291 | 0.253 | 0.484 |
| YES | NO | 5 | 8 | 0.356 | 0.116 | 0.642 | 0.249 | 0.689 |
| YES | NO | 8 | 0 | 0.325 | 0.279 | 0.064 | 0.566 | 0.585 |
| YES | NO | 8 | 5 | 0.323 | 0.268 | 0.306 | 0.584 | 0.68 |
| YES | NO | 8 | 8 | 0.329 | 0.284 | 0.631 | 0.606 | 0.823 |
| YES | YES | 0 | 0 | 0.322 | 0.916 | 0.062 | 0.693 | 0.421 |
| YES | YES | 0 | 5 | 0.327 | 0.88 | 0.308 | 0.669 | 0.552 |
| YES | YES | 0 | 8 | 0.341 | 0.89 | 0.637 | 0.691 | 0.74 |
| YES | YES | 5 | 0 | 0.318 | 0.81 | 0.084 | 0.742 | 0.57 |
| YES | YES | 5 | 5 | 0.305 | 0.809 | 0.294 | 0.734 | 0.666 |
| YES | YES | 5 | 8 | 0.349 | 0.802 | 0.589 | 0.757 | 0.809 |
| YES | YES | 8 | 0 | 0.288 | 0.761 | 0.076 | 0.832 | 0.705 |
| YES | YES | 8 | 5 | 0.32 | 0.767 | 0.333 | 0.84 | 0.816 |
| YES | YES | 8 | 8 | 0.351 | 0.737 | 0.623 | 0.817 | 0.869 |

Table 7.2: *A description of brain data*

| Accession ID | Repository | Data type | Platform | Number of probes |
|---|---|---|---|---|
| GSE15745 | GEO | Gene expression data | Illumina humanRef-8 v2.0 expression beadchip | 22,184 |
| GSE15745 | GEO | Methylation data | Illumina Human-Methylation27 BeadChip | 27,578 |
| phs000249.v1.p1 | dbGaP | Genotype data | HumanHap550v3.0 7 | 561,466 |

### *7.4.2 Gene Expression data*

Gene expression on four brain regions are publicly available (Gene Expression Omnibus (GEO) Accession Number: GSE15745) as rank-invariant (Schmid et al., 2010) normalized gene expression data. All the negative values in the gene expression dataset are changed to a 1 and the entire dataset was then log2 transformed. Variance due to tissue source, tissue bank and the hybridization batch were visualized using principal component analysis (PCA). Before generating the PCA plots, samples with African and Asian ancestry were removed from the analysis. All the gene expression probes on sex chromosomes X and Y were removed from the analysis. As evident from the PCA plot, there are four distinct clusters of samples based on tissue source. Tissue bank and hybridization batch do not influence the cluster formation. In order to identify outliers in the PCA analysis, a simple yet standard approach or rule has been adopted. All the samples that did not follow the *'IQR rule'* ($median + 1.5 * IQR$) were excluded from further analysis (one CRBLM, one FCTX and two PONS). These samples were also eliminated by Gibbs *et al* in their original analysis. More information on data preprocessing and results from it are found in the supplementary section.

Each gene expression probe was then adjusted for the biological and methodological covariates such as tissue bank, gender, hybridization batch and numeric covariates such as

post-mortem interval (PMI) and age in order to remove any associated confounding effects using the following linear model –

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + PC_1 + \ldots + PC_{50} + \epsilon$$

where Y is the gene expression data, $X_1 \ldots X_n$ represent the aforementioned biological and methodological covariates while $PC_1 \ldots PC_{50}$ are the top 50 principal components obtained from the original gene expression data. In order to target the difference in the genetic variation of expression among tissues, global variation in expression among tissues was removed by using the residual expression for each probe in each tissue after removing 50 PCs for further downstream analyses. It was shown in the past that the number of *cis*-eQTL detected significantly improved when 50 PCs were removed from the expression data (Fu et al., 2012).

### *7.4.3 Methylation data*

Methylation data, obtained as a "series matrix file" from GEO consisted of Beta-values, which represent the ratio of methylated probe intensity and the overall intensity (sum of methylated and unmethylated probe intensities) (Du et al., 2010). Therefore, Beta value for an $i^{th}$ interrogated CpG site is –

$$Beta_i = \frac{max(y_{i,methyl}, 0)}{max(y_{i,methyl}, 0) + max(y_{i,unmethyl}, 0) + const}$$

where $y_{i,methyl}$ and $y_{i,unmethyl}$ are the intensities measured by the $i^{th}$ methylation and unmethylated probes, respectively. Beta values range between 0 and 1. Samples with African and Asian ancestry were removed from the analysis. The methylation data was later adjusted for the biological and methodological covariates such as tissue bank, gender, hybridiza-

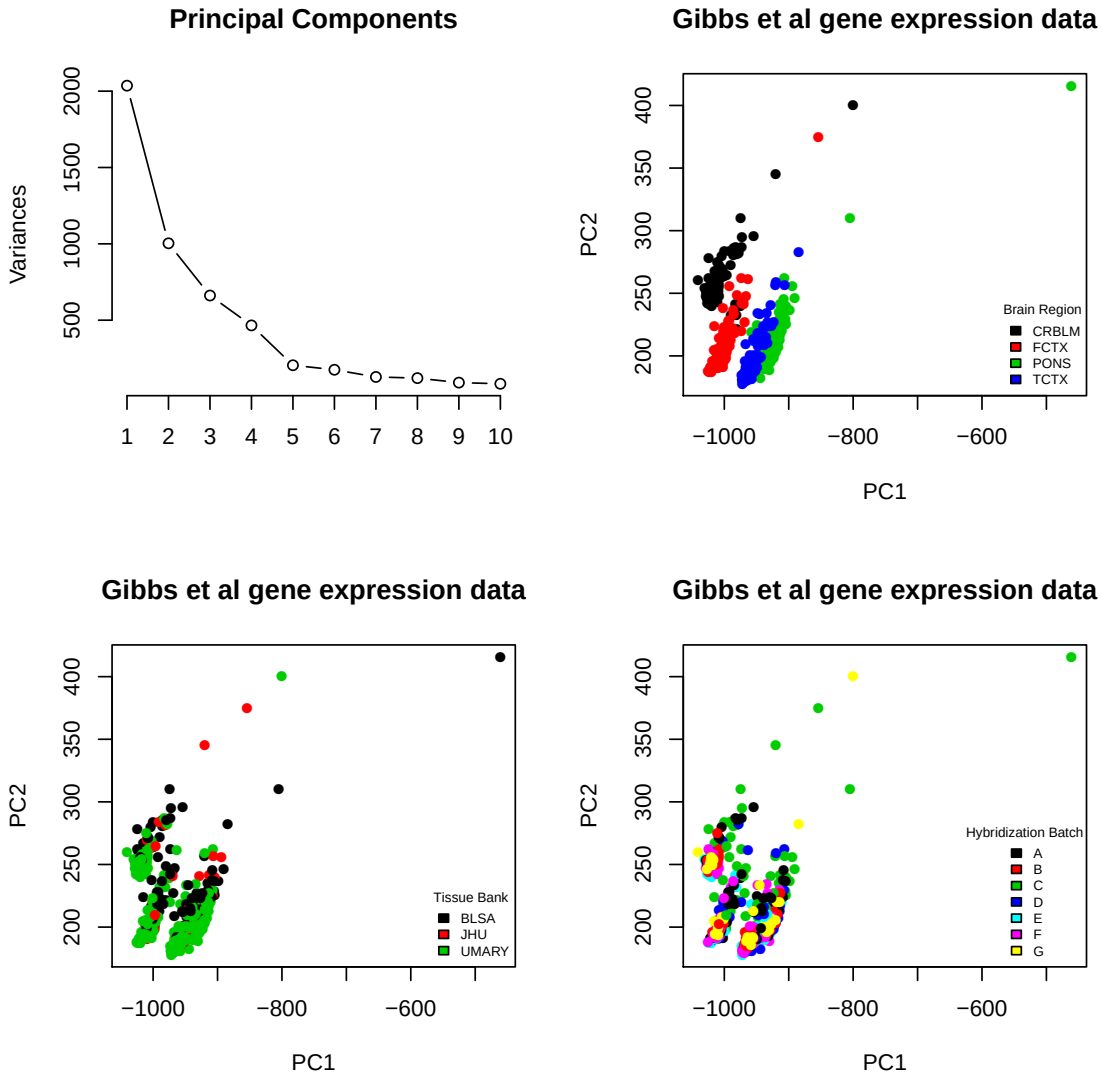



FIGURE 7.2: PCA plots exploring the presence of any biological or methodological variation using the first two principal components of the unadjusted rank-invariant normalized gene expression data.

tion batch and numeric covariates such as post-mortem interval (PMI) and age in order to remove any associated confounding effects using the following linear model –

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \epsilon$$

where Y is the methylation expression data and $X_1 \ldots X_n$ represent the aforementioned biological and methodological covariates. The residual methylation expression was later used in the subsequent downstream analyses.

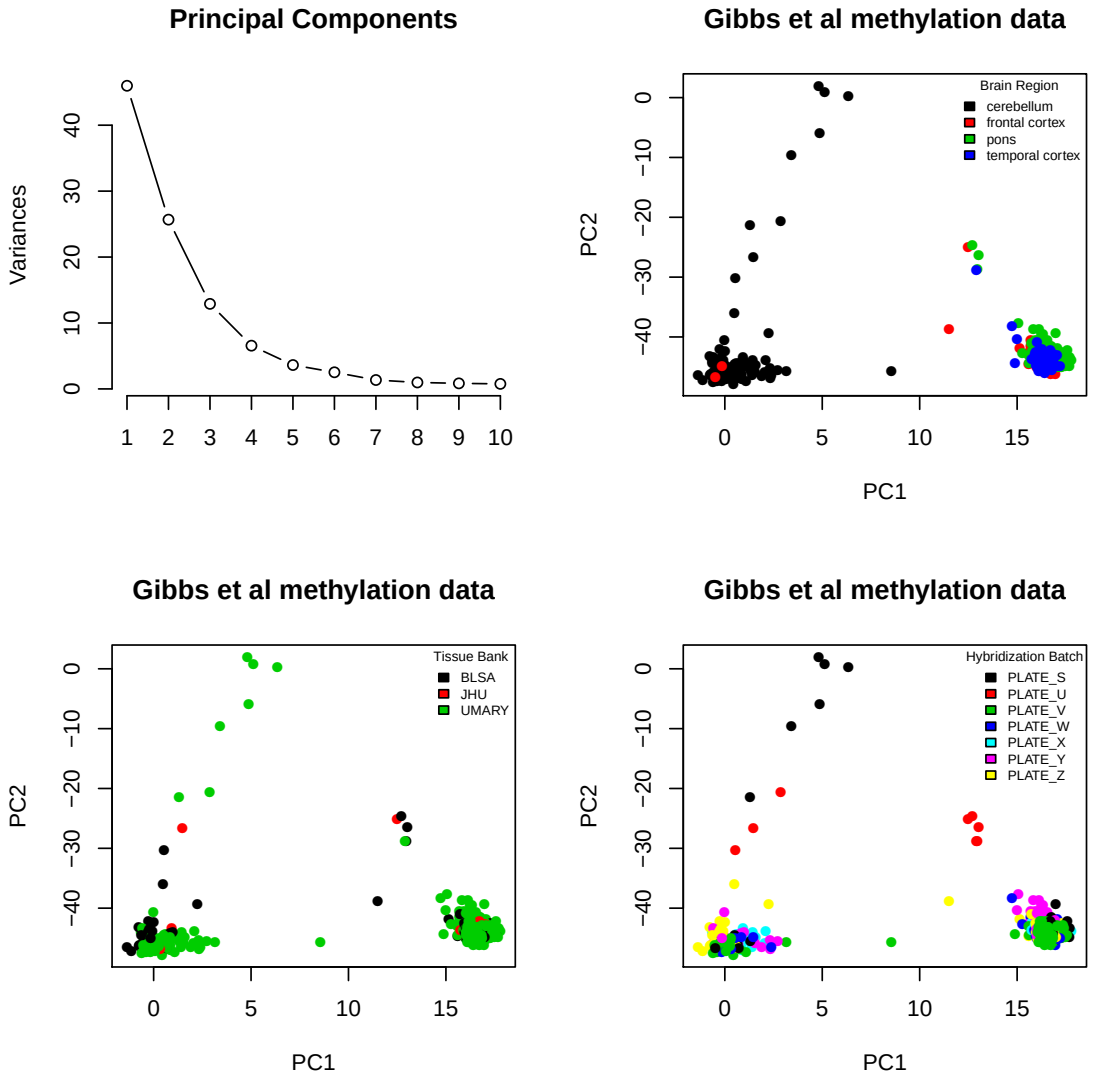



FIGURE 7.3: PCA plots exploring the presence of any biological or methodological variation using the first two principal components of the unadjusted methylation data.

## 7.5 Results from applying KEGG pathway analysis on results from Gibbs et al data

KEGG pathway analysis shows the biological relevance of this discovery.

Table 7.3: Enriched KEGG pathways in TBT and our Joint Test model

| Cluster | KEGG ID | Pathway Name | qvalue |
|---|---|---|---|
| TBT | hsa01100 | Metabolic pathways | 0.019519816 |
| JAGUAR | hsa00480 | Glutathione metabolism | 0.008110498 |
| JAGUAR | hsa01100 | Metabolic pathways | 0.018006598 |
| Joint score test | hsa01100 | Metabolic pathways | 0.00014368 |
| Joint score test | hsa03010 | Ribosome | 0.00014368 |
| Joint score test | hsa00240 | Pyrimidine metabolism | 0.00014368 |
| Joint score test | hsa00230 | Purine metabolism | 0.000490871 |
| Joint score test | hsa00071 | Fatty acid degradation | 0.003921085 |

An advantage to using our approach is the knowledge of individual contributions of each different effect (see table below), which gives us a hint on the extent to which tissue-specific effect is driving the association between every mRNA - CpG pair and SNP.

Table 7.4: The distribution of the different types of effects as measured by our joint score test statistic. $U_\beta$ is a measure of the main additive genetic effect. $U_\phi$ is a measure of $G \times M$ effect. $U_\gamma$ measures $G \times T$ effect while $U_\delta$ is a measure of $G \times M \times T$.

| $q\ value \leq 0.05$ | Additive Genotypic Effect | $G \times M$ Effect | $G \times T$ Effect | $G \times M \times T$ Effect |
|---|---|---|---|---|
| TRUE | 11,994 (90.8%) | 1,140 (8.6%) | 11,036 (83.5%) | 1,985 (15%) |
| FALSE | 1,218 (9.2%) | 12,072 (91.4%) | 2,176 (16.5%) | 11,227 (85%) |

## 7.6 JAGUAR

For a given gene-SNP pair, JAGUAR (Acharya et al., 2016) models gene expression across tissues using a linear mixed model in which both fixed and random effects are used to

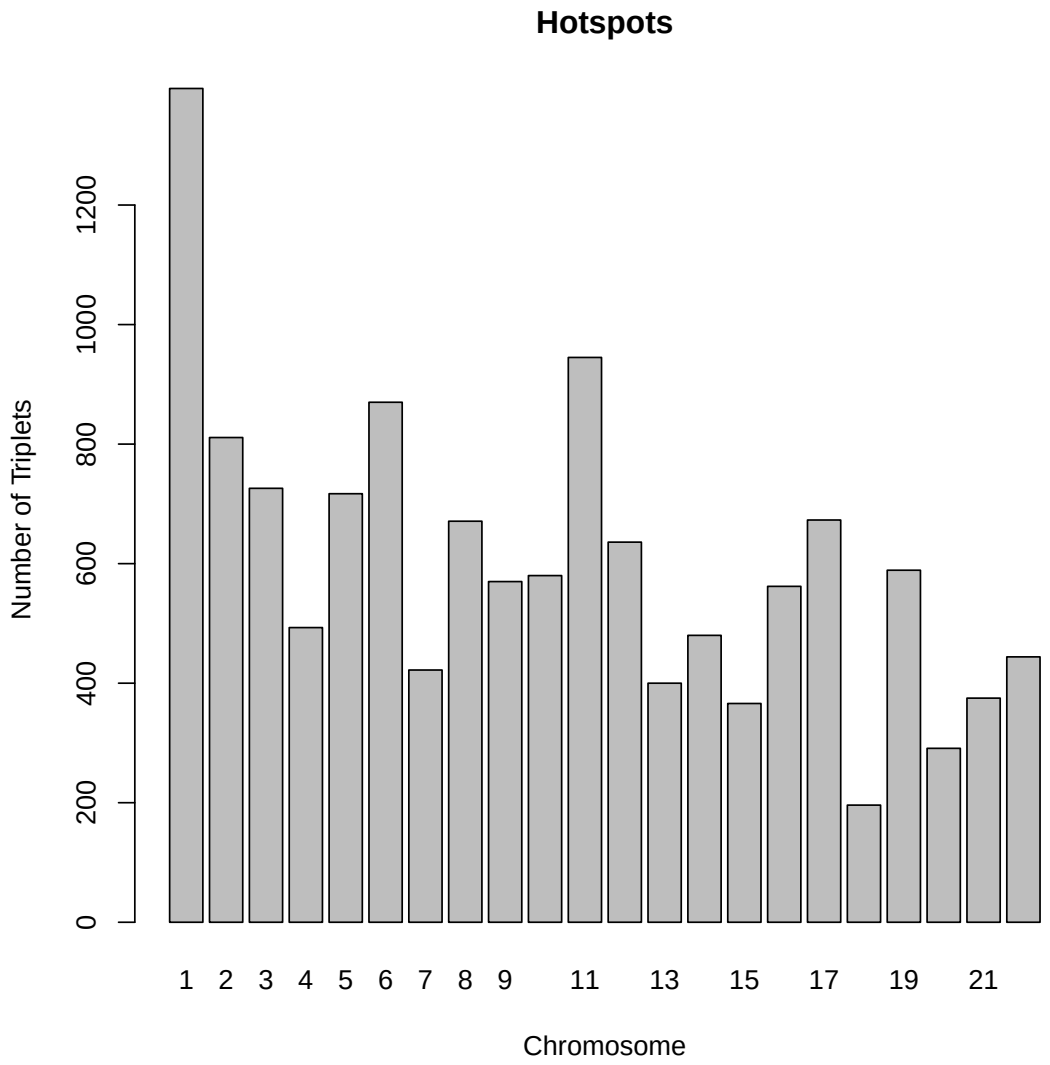



FIGURE 7.4: Number of mRNA - SNP pairs in each chromosome as identified by our method

capture the effect of a variant on gene expression. We begin with a linear mixed effects model that models expression patterns across tissues as a function of genotype, i.e.,

$$Y = J\alpha + G\beta + Au + Bv + \xi \tag{7.13}$$

where $Y$ is a $nt$-dimensional vector of expression levels in $t$ tissues and $n$ individuals, $\alpha$ is a vector of tissue-specific intercepts, $G$ is a $nt$-dimensional vector of genotypes, $\beta$ is a fixed effect of genotype across tissue, $u \sim N(0, \tau AA^T)$ is a vector of subject-specific random

effect, $v \sim N(0, \gamma BB^T)$ is a vector of tissue-specific random effects, and $\xi \sim N(0, \epsilon I_{nt})$. The matrices $J$, $A$ and $B$ are design matrices with $B$ being a function of genotype. $J$ is $nt \times t$ dimensional matrix denoting the design matrix for the tissue-specific intercepts. $A$ is $nt \times nt$ design matrix for the subject-specific intercepts. $B$ is a $nt \times t$ design matrix of stacked genotypes. The parameters of interest are $\beta$ and $\gamma$; $\alpha$, $\tau$ and $\epsilon$ are nuisance parameters.

We test the null hypothesis that $H_0 : \beta = \gamma = 0$, i.e. the variant does not affect gene expression across any of the tissues. To do so, we compute the efficient scores for $\beta$ and $\gamma$ by projecting off components correlated with the nuisance parameters.

The efficient scores evaluated under the null are given by –

$$U_\beta = (G - \bar{G})^T \hat{\Sigma}_n^{-1} (Y - J\hat{\alpha}) \tag{7.14}$$

and

$$U_\gamma = \frac{1}{2} (Y - J\hat{\alpha})^T \hat{\Sigma}_n^{-1} BB^T \hat{\Sigma}_n^{-1} (Y - J\hat{\alpha}) \tag{7.15}$$

where $\hat{\Sigma} = \hat{\tau} AA^T + \hat{\epsilon} I$ and $\hat{\tau}$ along with $\hat{\epsilon}$ are the maximum likelihood estimators of $\tau$ and $\epsilon$ under the null.

Following Huang et al (Huang et al., 2014), we propose a weighted sum of $U_\beta$ and $U_\gamma$ to arrive at our joint score test statistic as described by $U_\psi$. Since $U_\beta$ is linear in $Y$ while $U_\gamma$ is quadratic, we propose the following rule to combine them –

$$\begin{aligned} U_\psi &\equiv a_\beta U_\beta^2 + a_\gamma U_\gamma \\ &= (Y - J\hat{\alpha})^T \hat{\Sigma}_n^{-1} \left[ a_\beta (G - \bar{G})(G - \bar{G})^T + a_\gamma \left( \frac{1}{2} BB^T \right) \right] \hat{\Sigma}_n^{-1} (Y - J\hat{\alpha}), \end{aligned} \tag{7.16}$$

where $a_\beta$ and $a_\gamma$ are scalar constants chosen to minimize the variance of $U_\psi$. Under the null, $U_\psi$ is distributed as a mixture of chi-square random variables. We use Satterthwaite method (Satterthwaite, 1946) to approximate the $p$ values from a scaled $\chi^2$ distribution by matching the first two moments as $U_\psi \sim \kappa \chi_\nu^2$ where $\kappa = \frac{Var(U_\psi)}{2E[U_\psi]}$ and $\nu = \frac{2E[U_\psi]^2}{Var(U_\psi)}$.

## 7.7 Reproducibility of the analysis

All the scripts and the accompanied documentation for reproducing our analyses are located at `https://github.com/cramanuj/Epigen_Rcodes`.

# 8

# APPENDIX 3: Supplementary information for "Exploiting expression patterns across multiple gene isoforms to identify radiation response biomarkers in early-stage breast cancer patients"

## 8.1 Our models

### *8.1.1 Model for differential expression (DE) analysis*

$$Y = T\alpha + R\beta + Au + Bv + \xi \tag{8.1}$$

where $T$ is a $ntg \times t$ dimensional matrix of gene expresion levels in $t$ isoforms of a gene in $g$ groups and $n$ individuals, $\alpha$ is a fixed effect representing the isoform-specific intercepts, $R$ is a $ngt \times 1$-dimensional matrix of radiation dose identifiers such that $R \in \{0, 1\}$, 0 indicates no radiation and 1 indicates radiation, $\beta$ is a fixed effect indicating the average effect of radiation on gene expression. $u \sim N(0, \tau AA^T)$ indicates subject-specific random intercept, $v \sim N(0, \gamma BB^T)$ is random effect that denotes the interaction between gene-isoform and radiation (isoform-specific radiation effect), and $\xi \sim N(0, \epsilon I)$. $I$ is $ntg \times ntg$ dimensional identity matrix . The matrices $J$, $A$ and $B$ are design matrices with $B$ being a function of radiation dose. $J$ is $ntg \times t$ dimensional matrix denoting the design matrix for the tissue-

specific intercepts. $A$ is $ntg \times n$ design matrix for the subject-specific intercepts. $B$ is a $ntg \times t$ design matrix of stacked radiation dose identifiers.

Parameters of interest are $\gamma$, $\delta$, $\beta$ and $\phi$; $\alpha$, $\lambda$, $\tau$, $\theta$ and $\epsilon$ are nuisance parameters. We test the null hypothesis that $H_0 : \beta = \phi = \gamma = \delta = 0$, i.e. the variant does not affect gene expression across any of the tissues. Our joint score test will test for the effect of genotype on 1) an overall shift in the gene expression, 2) tissue-specific interaction ($G \times T$), 3) overall methylation ($G \times M$), and 4) tissue-specific methylation ($G \times M \times T$). All the random effects are independent to each other.

From the above model, the log-likelihood function conditioned on radiation is –

$$\ell(\beta, \theta) = c - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(Y - T\alpha - R\beta)^T \Sigma^{-1}(Y - T\alpha - R\beta) \tag{8.2}$$

where $\theta$ represents the vector of all the variance components involved in $\Sigma$, and $\beta$ while c is a constant.

From Jiming Jiang's *Linear and Generalized Linear Mixed Models and Their Applications* (Jiang, 2007) –

$$\frac{\partial \ell}{\partial \beta} = R^T\Sigma^{-1}Y - R^T\Sigma^{-1}R\beta$$

$$\frac{\partial \ell}{\partial \theta_r} = \frac{1}{2}\left\{(Y - T\alpha - R\beta)^T \Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_r}\Sigma^{-1}(Y - T\alpha - R\beta) - \mathrm{Tr}\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_r}\right)\right\}$$

where $\theta_r$ is the $r^{th}$ component of $\theta$ such that $\theta \in (\tau, \gamma, \epsilon)$.

$$E\left[\frac{\partial^2 \ell_i}{\partial \beta \partial \beta^T}\right] = -R^T\Sigma^{-1}R$$

$$E\left[\frac{\partial^2 \ell_i}{\partial \beta \partial \theta_r}\right] = 0$$

$$E\left[\frac{\partial^2 \ell_i}{\partial \theta_r \partial \theta_s}\right] = -\frac{1}{2}\mathrm{Tr}\left(\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_r}\Sigma^{-1}\frac{\partial \Sigma}{\partial \theta_s}\right)$$

Let the parameters of interest be $\psi = (\beta, \gamma)^T$ and the nuisance parameters be $\eta = (\alpha, \tau, \epsilon)^T$. The following is constructed under the null ($H_0$)

$$U_\psi = \begin{pmatrix} \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \gamma} \end{pmatrix} - \begin{bmatrix} I_{\beta\alpha} & I_{\beta\tau} & I_{\beta\epsilon} \\ I_{\gamma\alpha} & I_{\gamma\tau} & I_{\gamma\epsilon} \end{bmatrix} \begin{bmatrix} I_{\alpha\alpha} & I_{\alpha\tau} & I_{\alpha\epsilon} \\ I_{\tau\alpha} & I_{\tau\tau} & I_{\tau\epsilon} \\ I_{\epsilon\alpha} & I_{\epsilon\tau} & I_{\epsilon\epsilon} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial l}{\partial \alpha} \\ \frac{\partial l}{\partial \tau} \\ \frac{\partial l}{\partial \epsilon} \end{bmatrix}$$

Some algebra will result in the following –

$$U_\beta = (R - \bar{R})^T \hat{\Sigma}_n^{-1} (Y - T\hat{\alpha}) \tag{8.3}$$

and

$$U_\gamma = \frac{1}{2} (Y - T\hat{\alpha})^T \hat{\Sigma}_n^{-1} BB^T \hat{\Sigma}_n^{-1} (Y - T\hat{\alpha}) \tag{8.4}$$

$$\begin{aligned} U_\psi &= \left( \boldsymbol{a_\beta} U_\beta^2 + \boldsymbol{a_\gamma} U_\gamma \right) \\ &= (Y - T\hat{\alpha})^T \hat{\Sigma}_n^{-1} \left[ a_\beta (R - \bar{R}) (R - \bar{R})^T + a_\gamma \frac{1}{2} BB^T \right] \hat{\Sigma}_n^{-1} (Y - T\hat{\alpha}) \end{aligned} \tag{8.5}$$

### *8.1.2 Model for differentially enriched gene-set analysis*

$$Y = T\alpha + G\lambda + R\beta + Au + Bv + Cw + \xi \tag{8.6}$$

where $T$ is a $ntjg \times t$-dimensional matrix of expression levels in $t$ isoforms of a gene, $j$ genes, $g$ groups and $n$ individuals, $\alpha$ is a fixed effect representing $t$ isoform-specific intercepts, $\lambda$ is a fixed effect representing $g$ gene-specific intercepts, $R$ is a $ntjg \times 1$ dimensional matrix of radiation dose identifiers such that $R \in \{0, 1\}$, 0 indicates no radiation and 1 indicates radiation, $\beta$ is a fixed effect indicating the average effect of radiation on a pathway or gene-set. $u \sim N(0, \tau AA^T)$ indicates subject-specific random intercept, $v \sim N(0, \gamma BB^T)$ is

a random effect that denotes the interaction between gene-isoform and radiation (isoform-specific radiation effect), $w \sim N(0, \phi CC^T)$ is a random effect that denotes the interaction between gene and radiation (gene-specific radiation effect), and $\xi \sim N(0, \epsilon I)$. $I$ is $ntjg \times ntjg$-dimensional identity matrix . The matrices $J$, $A$ and $B$ are design matrices with $B$ being a function of radiation dose. $J$ is $ntjg \times t$ dimensional matrix denoting the design matrix for the tissue-specific intercepts. $A$ is $ntjg \times n$ design matrix for the subject-specific intercepts. $B$ is a $ntjg \times t$ design matrix of stacked radiation dose identifiers and $C$ is a $ntjg \times g$ dimensional design matrix of the $R \times G$ effect. All the random effects are independent to each other.

The log-likelihood function conditioned on the radiation is given by –

$$\ell(\beta, \theta) = c - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(Y - T\alpha - G\lambda - R\beta)^T \Sigma^{-1}(Y - T\alpha - G\lambda - R\beta) \tag{8.7}$$

where $\theta$ represents the vector of all the variance components involved in $\Sigma$, and $\beta$ while c is a constant.

Let the parameters of interest be $\psi = (\beta, \gamma, \phi)^T$ and the nuisance parameters be $\eta = (\alpha, \lambda, \tau, \epsilon)^T$. The following can be constructed under the null $(H_0)$ –

$$U_\zeta = \begin{bmatrix} \frac{\partial l}{\partial \beta} \\ \frac{\partial l}{\partial \gamma} \\ \frac{\partial l}{\partial \phi} \end{bmatrix} - \begin{bmatrix} I_{\beta\alpha} & I_{\beta\lambda} & I_{\beta\tau} & I_{\beta\epsilon} \\ I_{\gamma\alpha} & I_{\gamma\lambda} & I_{\gamma\tau} & I_{\gamma\epsilon} \\ I_{\phi\alpha} & I_{\phi\lambda} & I_{\phi\tau} & I_{\phi\epsilon} \end{bmatrix} \begin{bmatrix} I_{\alpha\alpha} & I_{\alpha\lambda} & I_{\alpha\tau} & I_{\alpha\epsilon} \\ I_{\lambda\alpha} & I_{\lambda\lambda} & I_{\lambda\tau} & I_{\lambda\epsilon} \\ I_{\tau\alpha} & I_{\tau\lambda} & I_{\tau\tau} & I_{\tau\epsilon} \\ I_{\epsilon\alpha} & I_{\epsilon\lambda} & I_{\epsilon\tau} & I_{\epsilon\epsilon} \end{bmatrix}^{-1} \begin{bmatrix} \frac{\partial l}{\partial \alpha} \\ \frac{\partial l}{\partial \lambda} \\ \frac{\partial l}{\partial \tau} \\ \frac{\partial l}{\partial \epsilon} \end{bmatrix}$$

Some algebra will result in the following –

$$U_\beta = (R - \bar{R})^T \hat{\Sigma}_n^{-1} \left(Y - T\hat{\alpha} - G\hat{\lambda}\right) \tag{8.8}$$

$$U_\gamma = \frac{1}{2}\left(Y - T\hat{\alpha} - G\hat{\lambda}\right)^T \hat{\Sigma}_n^{-1} BB^T \hat{\Sigma}_n^{-1} \left(Y - T\hat{\alpha} - G\hat{\lambda}\right) \tag{8.9}$$

$$U_\phi = \frac{1}{2}\left(Y - T\hat{\alpha} - G\hat{\lambda}\right)^T \hat{\Sigma}_n^{-1} CC^T \hat{\Sigma}_n^{-1}\left(Y - T\hat{\alpha} - G\hat{\lambda}\right) \tag{8.10}$$

$$\begin{aligned} U_\zeta &= \left(\boldsymbol{a_\beta} U_\beta^2 + \boldsymbol{a_\gamma} U_\gamma + \boldsymbol{a_\phi} U_\phi\right) \\ &= \left(Y - T\hat{\alpha} - G\hat{\lambda}\right)^T \hat{\Sigma}_n^{-1}\left[a_\beta\left(R - \bar{R}\right)\left(R - \bar{R}\right)^T + a_\gamma \frac{1}{2} BB^T + a_\phi \frac{1}{2} CC^T\right] \hat{\Sigma}_n^{-1}\left(Y - T\hat{\alpha} - G\hat{\lambda}\right) \end{aligned} \tag{8.11}$$

Optimal weights $a_\beta$, $a_\gamma$ and $a_\phi$ were derived using Lagrangians.

## 8.2 Null simulations

We carried out two simulation studies for each method. While the results from the power simulations were made available in the main manuscript, here are the tables showing the results from 'null' simulations in order to estimate the type I error at $\alpha = 0.05$.

### *8.2.1 Null simulations for the DE score test statistic*

We evaluated our method to detect DE genes using two simulation studies. Here we present results from null simulations from both simulation studies. Briefly, each Monte Carlo simulated dataset from the first simulation study was comprised of data for a single gene, whose expression is measured across 5 or 10 transcripts in 50 paired individuals. Each individual pair's radiation status is either a zero or a one indicating before and after radiotherapy, respectively. Since the transcript-specific effect is modeled as a random effect, a test of whether there is any transcript-specific effect due to radiation is equivalent to testing whether the variance of the random effect ($\gamma$) is zero.

In the second simulation study, each Monte Carlo dataset, comprised of gene expression data for 50 genes over 50 observations, each gene with unequal number of isoforms, was simulated from a multivariate normal distribution with a known variance-covariance

| Test | Type I error | Lower CI | Upper CI |
|---|---|---|---|
| DE Score Test | 0.0481 | 0.04398826 | 0.0524774 |
| TBT paired t-test | 0.0469 | 0.04162092 | 0.05261547 |
| TBT Wilcoxon test | 0.0438 | 0.03869686 | 0.04934269 |
| Gene-leve paired t-test | 0.0449 | 0.04092429 | 0.04914339 |

Table 8.1: DE of genes - Null Simulation results from our first simulation study at 5% FDR with 95% confidence interval. Our score test is referred to as "DE Score Test".

matrix. We varied the mean difference in differential gene expression between the two phenotypes, and the proportion of differentially expressed gene-isoforms. At the transcript level, we applied paired t-test and a non-parametric alternative in Wilcoxon's paired t-test and combined the $p$ values over all the transcripts of a gene using Fisher's method. At the gene-level, we combined the gene expression values by computing either the median or Winsorized mean of all the transcripts within a given gene. Paired t-test was run on this gene-level data.

| | DE Score Test | TBT Paired t-test | TBT Wilcoxon's test | Gene-level paired t-test |
|---|---|---|---|---|
| Type I error | 0.0578 | 0.0528 | 0.0456 | 0.0528 |

Table 8.2: DE of genes - Null Simulation results at 5% FDR from our second simulation study. Our score test is referred to as "DE Score Test".

### *8.2.2 Null simulations for the gene-set enrichment score test statistic*

We evaluated our method to detect DE gene-sets or pathways using two simulation studies. Here we present results from two simulations studies. Briefly, each Monte Carlo simulated dataset from the first simulation study was comprised of data for a single gene-set comprising of 5 genes, whose expression is measured across 3 transcripts in 50 paired individuals. Each individual pair's radiation status is either a zero or a one indicating before and after radiotherapy, respectively. Since the transcript-specific effect is modeled as a random effect, a test of whether there is any transcript-specific effect on the gene-sets

due to radiation is equivalent to testing whether the variances of the random effects ($\gamma$ and $\phi$) are zero.

| Test | Type I error | Lower CI | Upper CI |
|---|---|---|---|
| Gene-set Score Test | 0.0436 | 0.03968114 | 0.04778736 |
| TBT paired t-test | 0.0505 | 0.04489782 | 0.05655453 |

Table 8.3: Gene-set enrichment analysis - Null Simulation results from our first simulation study at 5% FDR with 95% confidence interval. Our score test is referred to as "Gene-set Score Test".

In our second simulation study, each Monte Carlo simulation consisted of 100 genes over 5 observations across the two phenotypes. We generated gene expression data using the same approach as described in the previous section. We simulated 10 gene-sets under both scenarios (with non-overlapping and overlapping genes) and compared the performance of our method with the other gene-set enrichment methods at the gene-level. We varied the sizes of gene-sets between 2 and 10 genes. Gene-level analysis is performed by computing the median gene expression values across all the transcripts within a gene followed by an implementation of gene set variational analysis (GSVA), Pathway Level analysis of Gene Expression (PLAGE), single sample GSEA (ssGSEA) and the combined z-score (ZSCORE). We estimated the empirical type I error rate at 5% FDR both in the presence and absence of any gene overlap among the simulated gene-sets.

| | Gene-set overlap | Gene-set Score Test | GSVA | PLAGE | ssGSEA | ZSCORE |
|---|---|---|---|---|---|---|
| Type I error | Yes | 0.053 | 0.057 | 0.054 | 0.052 | 0.044 |
| Type I error | No | 0.045 | 0.057 | 0.050 | 0.056 | 0.049 |

Table 8.4: Gene-set enrichment analysis - Null Simulation results from our second simulation study at 5% FDR. We present type I error rates for two cases - gene-sets share genes (overlap) and gene-sets with unique genes (no overlap). Our score test is referred to as "Gene-set Score Test".

## 8.3 Gene-set analysis methods

We compared the performance of our score test method for gene-set analysis with gene-set variational analysis (GSVA) (Hänzelmann et al., 2013), pathway level analysis of gene expression (PLAGE) (Tomfohr et al., 2005), the combined z-score method (ZSCORE) (Lee et al., 2008) and single sample GSEA (ssGSEA) (Barbie et al., 2009). Briefly, PLAGE standardizes expression profiles into z-scores over the samples and then calculates the singular value decomposition $Z = UDV'$ on the z-scores of the genes in the gene-set. The coefficients of the first right-singular vector (first column of V ) are taken as the gene-set summaries of expression over the samples. ZSCORE method also standardizes expression profiles into z-scores over the samples, but combines them together for each gene-set at each individual sample in the following way. Given a gene-set $\gamma = \{1, \cdots, k\}$ with z-scores $Z_1, \cdots, Z_k$ for each gene, the combined z-score $Z_\gamma$ for the gene-set $\gamma$ is $\frac{\sum_{i=1}^{k} Z_i}{\sqrt{k}}$. ssGSEA, on the other hand, calculates a gene-set enrichment score per sample as the normalized difference in empirical cumulative distribution functions of gene expression ranks inside and outside the gene-set.

All of these methods do not perform analysis at the transcript-level. In order to apply the aforementioned methods on both simulated and real data, we combined gene expression values over all the transcripts for a given gene by computing the median expression values thus, performing the analysis at the gene-level to obtain a matrix single sample enrichment scores. Paired t-test was then performed on this matrix.

## 8.4 Reproducibility of the analysis

All the scripts and the accompanied documentation for reproducing our analyses are located at `https://github.com/cramanuj/Radiation_Rcodes`.

# Bibliography

1000 Genomes Project Consortium, Auton, A., Brooks, L., Durbin, R., Garrison, E., Kang, H., Korbel, J., Marchini, J., McCarthy, S., McVean, G., and Abecasis, G. (2015), "A global reference for human genetic variation," *Nature*, 526, 68–74.

Acharya, C., McCarthy, J., Owzar, K., and Allen, A. (2016), "Exploiting expression patterns across multiple tissues to map expression quantitative trait loci," *BMC Bioinformatics*, 17, DOI: 10.1186/s12859–016–1123–5.

Acharya, C. R. and Allen, A. S. (2016), *JAGUAR: Joint Analysis of Genotype and Group-Specific Variability Using a Novel Score Test Approach to Map Expression Quantitative Trait Loci (eQTL)*, R package version 3.0.1.

Affymetrix (2016), "Affymetrix, GeneChip Human Transcriptome Array 2.0 Data Sheet." .

Akhtar, R. S., Ness, J. M., and Roth, K. A. (2004), "Bcl-2 family regulation of neuronal development and neurodegeneration," *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1644, 189 – 203, Bcl-2 family members: integrators of survival and death signals in physiology and pathology.

Balcer-Kubiczek, E. (2012), "Apoptosis in radiation therapy: a double-edged sword," *Experimental Oncology*, 34, 277–285.

Banovich, N., Lan, X., McVicker, G., van de Geijn, B., Degner, J., Blischak, J., Roux, J., Pritchard, J., and Gilad, Y. (2014), "Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels," *PLoS Genetics*, 10.

Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., Schinzel, A. C., Sandy, P., Meylan, E., Scholl, C., Frohling, S., Chan, E. M., Sos, M. L., Michel, K., Mermel, C., Silver, S. J., Weir, B. A., Reiling, J. H., Sheng, Q., Gupta, P. B., Wadlow, R. C., Le, H., Hoersch, S., Wittner, B. S., Ramaswamy, S., Livingston, D. M., Sabatini, D. M., Meyerson, M., Thomas, R. K., Lander, E. S., Mesirov, J. P., Root, D. E., Gilliland, D. G., Jacks, T., and Hahn, W. C. (2009), "Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1," *Nature*, 462, 108–112.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2014a), *lme4: Linear mixed-effects models using Eigen and S4*, R package version 1.1-7.

Bates, D., Maechler, M., Bolker, B. M., and Walker, S. (2014b), "lme4: Linear mixed-effects models using Eigen and S4," ArXiv e-print; submitted to *Journal of Statistical Software*.

Baumann, M., Krause, M., Overgaard, J., Debus, J., Bentzen, S., Daartz, J., Richter, C., and Zips, D. andBortfeld, T. (2016), "Radiation oncology in the era of precision medicine," *Nature Reviews Cancer*, 16, 234–49.

Bell, J., Pai, A., Pickrell, J., Gaffney, D., Pique-Regi, R., Degner, J., Gilad, Y., and Pritchard, J. (2011), "DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines," *Genome Biology*, 12.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, 57, 289–300.

Bernhard, E., Maity, A., Muschel, R., and McKenna, W. (1995), "Effects of ionizing radiation on cell cycle progression. A review." *Radiation and environmental biophysics*, 34, 79–83.

Birnbaum, A. (1959), "Combining Independent Tests of Significance," *Journal of American Statistical Association*, 59, 559–574.

Bolstad, B. (2016), *affyio: Tools for parsing Affymetrix data files*, R package version 1.40.0.

Bradbury, P. (2003), "Human Epigenome Project—Up and Running." *PLoS Biology*, 1, e82. doi:10.1371/journal.pbio.0000082.

Brady, L., Perez, C., and Halperin, E. (2007), *Principles and Practice of Radiation Oncology*, Lippincott Williams and Wilkins, 5th edition edn.

Brem, R., Yvert, G., Clinton, R., and Kruglyak, L. (2002), "Genetic dissection of transcriptional regulation in budding yeast," *Science*, 296, 752–5.

Broman, K., Wu, H., Sen, S., and Churchill, G. (2003), "R/qtl: QTL mapping in experimental crosses," *Bioinformatics*, 19, 889–890.

Carter, S., Eklund, A., Kohane, I., Harris, L., and Z, S. (2006), "A signature of chromosomal instability inferred from gene expression profiles predicts clinical outcome in multiple human cancers." *Nature Genetics*, 38, 1043–1048.

Carvalho, B. and Irizarry, R. (2010), "A framework for oligonucleotide microarray preprocessing," *Bioinformatics*, 26, 2363–2367.

Chang, H., Sneddon, J., Alizadeh, A., Sood, R., West, R., Montgomery, K., Chi, J.-T., van de Rijn, M., Botstein, D., and Brown, P. (2004), "Gene Expression Signature of Fibroblast Serum Response Predicts Human Cancer Progression: Similarities between Tumors and Wounds." *PLoS Biology*, 2, e7. doi:10.1371/journal.pbio.0020007.

Chi, J., Wang, Z., Nuyten, D., Rodriguez, E., Schaner, M., Salim, A., Wang, Y., Kristensen, G., Helland, A., Børresen-Dale, A., Giaccia, A., Longaker, M., Hastie, T., Yang, Y., van de Vijver, M., and Brown, P. (2006), "Gene expression programs in response to hypoxia; cell type specificity and prognostic significance in human cancers," *PLoS Medicine*, 3, e47. doi:10.1371/journal.pmed.0030047.

Consortium, T. E. P. (2004), "The Encode (ENCyclopedia of DNA Elements) Project," *Science*, 306, 636–640.

Consortium., T. I. H. (2003), "The International HapMap Project." *Nature*, 426, 789–796.

Cookson, W., Liang, L., Abecasis, G., Moffat, M., and Lathrop, M. (2009), "Mapping complex disease traits with global gene expression," *Nature Reviews Genetics*, 10, 184–194.

Deaton, A. and Bird, A. (2011), "CpG islands and the regulation of transcription," *Genes and Development*.

Du, P., Zhang, X., Huang, C., Jafari, N., Kibbe, W., Hou, L., and Lin, S. (2010), "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis," *BMC Bioinformatics*, 11.

Duchesne, P. and Lafaye De Micheaux, P. (2010), "In order to expedite the analysis, the gene expression data is split into many partitions such that there exist at least 100 genes in each such partitions and all such partitions are analyzed simultaneously (parallel processing). Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods." *Computational Statistical Data Analysis*, 54.

Edgar, R., Domrachev, M., and Lash, A. (2002), "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nulciec Acids Res.*, 30, 207–10.

Efron, B. and Tibshirani, R. (2006), "On testing the significance of sets of genes," Tech. rep., Stanford University.

Efron, B. and Tibshirani, R. (2010), *GSA: Gene set analysis*, R package version 1.03.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998), "Cluster analysis and display of genome-wide expression patterns," *PNAS*, 95, 14863–8.

Fisher, R. (1934), *Statistical Methods for Research Workers*, Oliver and Boyd, fifth edn.

Flutre, T., Wen, X., Pritchard, J., and Stephens, M. (2013), "A Statistical Framework for Joint eQTL Analysis in Multiple Tissues," *PLoS Genetics*, 9.

Fu, J., Wolfs, M., Deelen, P., Westra, H., and et al (2012), "Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression," *PLoS Genetics*, 8.

Gamazon, E. R., Wheeler, H. E., Shah, K. P., Mozaffari, S. V., Aquino-Michaels, K., Carroll, R. J., Eyler, A. E., Denny, J. C., Consortium, G., Nicolae, D. L., Cox, N. J., and Im, H. K. (2015), "A gene-based association method for mapping traits using reference transcriptome data," *Nat Genet*, 47, 1091–1098.

Gatti, D., Shabalin, A., Lam, T., Wright, F., Rusyn, I., and Nobel, A. (2009), "FastMap: fast eQTL mapping in homozygous populations," *Bioinformatics*, 25.

Geyer, P. K., Green, M. M., and Corces, V. G. (1990), "Tissue-specific transcriptional enhancers may act in trans on the gene located in the homologous chromosome: the molecular basis of transvection in Drosophila." *EMBO J.*, 9.

Gibbs, J., van der Brug, M., Hernandez, D., Traynor, B., Nalls, M., Lai, S.-L., Arepally, S., Dillman, A., Rafferty, I., Troncoso, J., Johnson, R., Zielke, H., Ferrucci, L., Longo, D., Cookson, M., and Singleton, A. (2010), "Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain," *Plos Genet*, 6.

Gutierrez-Arcelus, M., Lappalainen, T., Montgomery, S., Buil, A., Ongen, H., Yurovsky, A., Bryois, J., Giger, T., Romano, L., Planchon, A., Falconnet, E., Biesler, D., Gagnebin, M., Padioleau, I., Borel, C., Letourneau, A., Makrythanasis, P., Guipponi, M., Gehrig, C., Antonarakis, S., and Dermitzakis, E. (2013), "Passive and active DNA methylation and the interplay with genetic variation in gene regulation," *eLife*, 2.

Gutierrez-Arcelus, M., Ongen, H., Lappalainen, T., Montgomery, S., Buil, A., Yurovsky, A., Bryois, J., Padioleau, I., Romano, L., Planchon, A., Falconnet, E., Biesler, D., Gagnebin, M., Giger, T., Borel, C., Letourneau, A., Makrythanasis, P., Guipponi, M., Gehrig, C., Antonarakis, S., and Dermitzakis, E. (2015), "Tissue-Specific Effects of Genetic and Epigenetic Variation on Gene Regulation and Splicing," *PLoS Genetics*.

Hänzelmann, S., Castelo, R., and Guinney, J. (2013), "GSVA: gene set variation analysis for microarray and RNA-Seq data," *BMC Bioinformatics*, 14, DOI: 10.1186/1471–2105–14–7.

Harrison, P. (1999), "The neuropathology of schizophreniaA critical review of the data and their interpretation," *Brain*, 122, 593–624.

Harrison, P. and Weinberger, D. (2005), "Schizophrenia genes, gene expression, and neuropathology: on the matter of their convergence," *Molecular Psychiatry*, 10.

Hein, A., Ouellette, M., and Yan, Y. (2014), "Radiation-induced signaling pathways that promote cancer cell survival (Review)," *International Journal of Oncology*, 45, 1813–1819.

Hellman, A. and Chess, A. (2010), "Extensive sequence-influenced DNA methylation polymorphism in the human genome." *Epigenetics Chromatin*, 24, 1.

Horton, J., Siamakpour-Reihani, S., Lee, C., Zhou, Y., Chen, W., Geradts, J., Fels, D., Hoang, P., Ashcraft, K., Groth, J., Kung, H., Dewhirst, M., and Chi, J.-T. (2015), "FAS Death Receptor: A Breast Cancer Subtype-Specific Radiation Response Biomarker and Potential Therapeutic Target," *Radiation Research*, 184, 456–69.

Hsu, D., Kim, M.K.and Balakumaran, B., Acharya, C., Anders, C., Clay, T., Lyerly, H., Drake, C., Morse, M., and Febbo, P. (2010), "Immune Signatures Predict Prognosis in Localized Cancer," *Cancer Investigation*, 28, 765–773.

Huang, Y., VanderWeele, T., and Lin, X. (2014), "Joint analysis of snp and gene expression data in genetic association studies of complex diseases," *Annals of Applied Statistics*, 8, 352–376.

Ihaka, R. and Gentleman, R. (1996), "A language for data analysis and graphics," *Journal of Computational and Graphical Statistics*, 5, 299–314.

Irizarry, R., Hobbs, B., Collin, F., Beazer-Barclay, Y., Antonellis, K., Scherf, U., and Speed, T. (2003), "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." *Biostatistics*, 4, 249–264.

Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, Springer Series in Statistics, Springer-Verlag New York, 233 Springer Street, New York, NY 10013, USA, 1 edn.

Johnson, W., Li, C., and Rabinovic, A. (2007), "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, 8.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016), "KEGG as a reference resource for gene and protein annotation," *Nucleic Acids Research*, 44, D457–62.

Kudo, W., Lee, H.-P., Smith, M. A., Zhu, X., Matsuyama, S., and Lee, H.-g. (2012), "Inhibition of Bax protects neuronal cells from oligomeric A[beta] neurotoxicity," *Cell Death Dis*, 3, e309–.

Langlands, F. E., Horgan, K., Dodwell, D. D., and Smith, L. (2013), "Breast cancer subtypes: response to radiotherapy and potential radiosensitisation." *British Journal of Radiology*, 86, 20120601.

Lee, E., Chuang, H.-Y., Kim, J.-W., Ideker, T., and Lee, D. (2008), "Inferring Pathway Activity toward Precise Disease Classification," *PLoS Computational Biology*, 4, e1000217. doi:10.1371/journal.pcbi.1000217.

Lemire, M., Zaidi, S., Ban, M., and Ge, B. e. a. (2014), "Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci," *Nature Communications*, 6.

Li, J., Huang, S., Armstrong, E., Fowler, J., and Harari, P. (2005), "Angiogenesis and radiation response modulation after vascular endothelial growth factor receptor-2 (VEGFR2) blockade," *International Journal of Radiation Oncology*Biology*Physics*, 62, 1477–1485.

Liberzon, A., Birger, C., Thorvaldsdottir, H., Ghandi, M., Mesirov, J., and Tamayo, P. (2015), "The Molecular Signatures Database Hallmark Gene Set Collection," *Cell Systems*, 1, 417–425.

Lin, D. (2005), "An efficient Monte Carlo approach to assessing statistical significance in genomic studies," *Bioinformatics*, 21.

Lin, X. (1997), "Variance component testing in generalized linear models with random effects," *Biometrika*, 84, 309–326.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C., Davidson, R., and Heckerman, D. (2011), "FaST linear mixed models for genome-wide association studies." *Nature Methods*, 8.

Lisak, D. A., Schacht, T., Enders, V., Habicht, J., Kiviluoto, S., Schneider, J., Henke, N., Bultynck, G., and Methner, A. (2015), "The transmembrane Bax inhibitor motif (TMBIM) containing protein family: Tissue expression, intracellular localization and effects on the {ER} CA2+-filling state," *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, 1853, 2104 – 2114, 13th European Symposium on Calcium.

Little, R. and Rubin, D. (2002), *Statistical Analysis with Missing Data*, John Wiley and Sons, 111 River Street, Hoboken, NJ 07030, USA, 2 edn.

Liu, D., Peloso, G., Zhan, Z., Holmen, O., and et al (2014), "Meta-analysis of gene-level tests for rare variant association," *Nature Genetics*, 46, 200–204.

Liu, R., Wang, X., Chen, G., Dalerba, P., Gurney, A., Hoey, T., Sherlock, G., Lewicki, J., Shedden, K., and Clarke, M. (2007), "The prognostic role of a gene signature from tumorigenic breast-cancer cells. The prognostic role of a gene signature from tumorigenic breast-cancer cells," *N Engl J Med*, 356, 217–226.

Lonsdale, J. and et al (2013), "The Genotype-Tissue Expression (GTEx) project," *Nature Genetics*, 45, 580–585.

LoPiccolo, J., Blumenthal, G., Bernstein, W., and Dennis, P. (2008), "Targeting the PI3K/Akt/mTOR pathway: effective combinations and clinical considerations," *Drug Resistance Update*, 11, 32–50.

Macaeva, E., Saeys, Y., Tabury, K., Janssen, A., Michaux, A., Benotmane, M. A., De Vos, W. H., Baatout, S., and Quintens, R. (2016), "Radiation-induced alternative transcription and splicing events and their applicability to practical biodosimetry," *Scientific Reports*, 6, 19251 EP –.

MacDonald, J. (2016), *pd.hta.2.0: Platform Design Info for Affymetrix HTA-2˙0*.

McCarthy, M., Abecasis, G., Cardon, L., Goldstein, D., Little, J., Ioannidis, J., and Hirschhorn, J. (2008), "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." *Nature Reviews Genetics*, 9, 356–69.

Micheli, V., Camici, M., Tozzi, M., Ipata, P., Sestini, S., Bertelli, M., and Pompucci, G. (2011), "Neurological disorders of purine and pyrimidine metabolism," *Current Topics in Medicinal Chemistry*, 11.

Miller, J., Cai, C., Langfelder, P., Geschwind, D., Kurian, S., Salomon, D., and Horvath, S. (2011), "Strategies for aggregating gene expression data: The collapseRows R function," *BMC Bioinformatics*, 12, DOI: 10.1186/1471–2105–12–322.

Ong, C.-T. and Corces, V. (2011), "Enhancer function: new insights into the regulation of tissue-specific gene expression," *Nature Reviews Genetics*, 12.

Pletcher, M., McClurg, P., Batalov, S., and et al (2004), "Use of a dense single nucleotide polymorphism map for in silico mapping in the mouse," *PLoS Biology*, 2.

Price, A., Patterson, N., Plenge, R., Weinblatt, M., Shadick, N., and Reich, D. (2006), "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, 38, 904–909.

Purcell, S., Neale, B., Todd-Brown, K., and et al (2007), "PLINK: a tool-set for whole-genome association and population-based linkage analyses," *American Journal of Human Genetics*, 81.

Ramaswamy, A., Trabzuni, D., Guelfi, S., Varghese, V., Smith, C., Walker, R., De, T., and et al (2014), "Genetic variability in the regulation of gene expression in ten regions of the human brain," *Nature Neuroscience*, 17.

Satterthwaite, F. (1946), "An approximate distribution of estimates of variance components," *Biometrics Bulletin*, 2, 110–114.

Schmid, R., Baum, P., Ittrich, C., Fundel-Clemens, K., Huber, W., Brors, B., Eils, R., Weith, A., Mennerich, D., and Quast, K. (2010), "Comparison of normalization methods of Illumina BeadChip HumanHT-12 v3," *BMC Genomics*, 11.

Scott-Boyer, M., Imholte, G., Tayeb, A., Labbe, A., Deschepper, C., and Gottardo, R. (2012), "An integrated hierarchical Bayesian model for multivariate eQTL mapping," *Statistical Applications in Genetics and Molecular Biology*, 11.

Shabalin, A. (2012), "Matrix eQTL: ultra-fast eQTL analysis via large matrix operations," *Bioinformatics*, 28.

Shoemaker, R., Deng, J., Wang, W., and Zhang, K. (2010), "Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome," *Genome Res.*, 20, 883–889.

Sing, T., Sander, O., Beerenwinkel, N., and T, L. (2005), "ROCR: visualizing classifier performance in R." *Bioinformatics*, 21, 3940–1.

Storey, J. and Tibshirani, R. (2003), "Statistical significance for genome-wide experiments." *PNAS*.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005), "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proceedings of the National Academy of Sciences*, 102, 15545–15550.

Sul, J., Han, B., Ye, C., Choi, T., and Eskin, E. (2013), "Effectively Identifying eQTLs from Multiple Tissues by Combining Mixed Model and Meta-analytic Approaches," *PLoS Genetics*, 9.

Sun, G., Yan, J., Noltner, K., Feng, J., Li, H., Sarkis, D., Sommer, S., and Rossi, J. (2009), "SNPs in human miRNA genes affect biogenesis and function," *RNA*, 15, 1640–1651.

Sun, W. (2009), *eMAP: eQTL analysis by Linear Model*, University of North Carolina, Chapel Hill.

Swift-Scanlan, T., Smith, C., Bardowell, S., and Boettiger, C. (2014), "Comprehensive interrogation of CpG island methylation in the gene encoding COMT, a key estrogen and catecholamine regulator," *BMC Medical Genomics*.

Tomfohr, J., Lu, J., and Kepler, T. (2005), "Pathway level analysis of gene expression using singular value decomposition," *BMC Bioinformatics*, 6, DOI: 10.1186/1471–2105–6–225.

Torres-Roca, J. (2012), "A molecular assay of tumor radiosensitivity: a roadmap towards biology-based personalized radiation therapy," *Per Med.*, 9, 547–557.

Torres-Roca, J., Eschrich, S., Zhao, H., Bloom, G., Sung, J., McCarthy, S., Cantor, A., Scuto, A., Li, C., Zhang, S., Jove, R., and Yeatman, T. (2005), "Prediction of radiation sensitivity using a gene expression classifier," *Cancer Research*, 65, 7169–76.

Verbeke, G. and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, Springer-Verlag New York, 175 Fifth Avenue, New York NY 10010, USA, 1 edn.

Viemann, D., Goebeler, M., Schmid, S., Nordhues, U., Klimmek, K., Sorg, C., and Roth, J. (2006), "TNF induces distinct gene expression programs in microvascular and macrovascular human endothelial cells." *Journal of Leukocyte Biology*, 80, 174–185.

Wagner, J., Busche, S., Ge, B., Kwan, T., Pastinen, T., and Blanchette, M. (2014), "The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts," *Genome Biology*, 15.

Wang, X. and Cairns, M. (2014), "SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing." *Bioinformatics*, 30, 1777–9.

Widschwendter, M., Fiegl, H., Egle, D., Mueller-Holzner, E., Spizzo, G., Marth, C., Weisenberger, D. J., Campan, M., Young, J., Jacobs, I., and Laird, P. W. (2007), "Epigenetic stem cell signature in cancer," *Nat Genet*, 39, 157–158.

Wilcox, R. and Keselman, H. (2003a), "Modern Robust Data Analysis Methods: Measures of Central Tendency," *Psychological Methods*, 8, 254–274.

Wilcox, R. and Keselman, H. (2003b), "Modern robust data analysis methods: Measures of central tendency," *Psychological Methods*, 8, 254–274.

Wilcoxon, F. (1945), "Individual Comparisons by Ranking Methods," *Biometrics Bulletin*, 1, 80–83.

Willers, H., Dahm-Daphi, J., and Powell, S. (2004), "Repair of radiation damage to DNA," *British Journal of Cancer*, 90, 1297–1301.

Wrzodek, C., Büchel, F., Hinselmann, G., Eichner, J., Mittag, F., and Zell, A. (2012), "Linking the Epigenome to the Genome: Correlation of Different Features to DNA Methylation of CpG Islands," *Plos ONE*, 7.

Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011), "Rare-Variant association testing for sequence data with sequence kernel association test," *American Journal of Human Genetics*, 89, 82–93.

Wushou, A., Jiang, Y.-Z., Hou, J., Liu, Y.-R., Guo, X.-M., and Shao, Z.-M. (2015), "Development of triple-negative breast cancer radiosensitive gene signature and validation based on transcriptome analysis," *Breast Cancer Research and Treatment*, 154, 57–62.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012), "ClusterProfiler: An r package for comparing biological themes among gene clusters." *OMICS: A Journal of Integrative Biology*, 16, 284–287.

Yu, G., Wang, L., Yan, G., and He, Q. (2014), "DOSE: an R/Bioconductor package for Disease Ontology Semantic and Enrichment analysis," *Bioinformatics*, 31, 608–9.

# Biography

Chaitanya R. Acharya was born on August 30th, 1981, in the southern state of Andhra Pradesh in India. He grew up in India for the first 18 years of his life and moved to the United States in pursuit of higher education. He earned his undergraduate degree in Biological Sciences from Michigan Technological University, MI, in 2003. He attended Arizona State University to earn a graduate Professional Science Master's degree in Computational Biosciences and Bioinformatics in 2006. After working as a Research Technician in a genomics lab at Duke University for nearly four years, he matriculated into the Computational Biology and Bioinformatics Program at Duke University in 2010. While being a Research Technician at Duke University, he earned Duke Young Investigator Award in 2007 and Citizen's Advisory Council Young Investigator Award in 2009. He plans on working as a Post-doctoral Research Fellow in Dr. Herbert Kim Lyerly's lab, who is George Barth Geller Professor in the Department of Surgery at Duke University Medical Center, and a Principal Investigator in the Center for Applied Therapeutics.

## Relevant Publications


1. **Acharya, C.R.**, McCarthy, J.M., Owzar, K. and Allen, A.S. Exploiting expression patterns across multiple tissues to map expression quantitative trait loci. BMC Bioinformatics. 17(257):DOI: 10.1186/s12859-016-1123-5

2. **Acharya, C.R.** and Allen, A.S. (2016), JAGUAR: Joint Analysis of Genotype and

Group- Specific Variability Using a Novel Score Test Approach to Map Expression Quantitative Trait Loci (eQTL), R package version 3.0.1.

3. **Acharya, C.R.**, Owzar, K., Allen, A.S. Mapping eQTL by leveraging multiple tissues and DNA methylation. bioRxiv 069534; doi: http://dx.doi.org/10.1101/069534. *In review.*

4. **Acharya, C.R.**, Owzar, K., Horton, J.K. and Allen, A.S. Exploiting expression patterns across multiple gene isoforms to identify radiation response biomarkers in early-stage breast cancer patients. *Manuscript In Preperation.*