Bioinformatic analysis of complex, high-throughput genomic and epigenomic data in the context of $CD4^+$ T-cell differentiation and diagnosis and treatment of transplant rejection

A thesis presented

by

Ryan C. Thompson

to

The Scripps Research Institute Graduate Program

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in the subject of Biology

for

The Scripps Research Institute

La Jolla, California

October 2019

For Dan, who helped me through the hard times again and again.

He is, and will always be, fondly remembered and sorely missed.

Acknowledgements

My path through graduate school has been a long and winding one, and I am grateful to all the mentors I have had through the years – Drs. Terry Gaasterland, Daniel Salomon, and Andrew Su – all of whose encouragement and support have been vital to my development into the scientist I am today. I am also thankful for my collaborators in the Salomon lab: Drs. Sarah Lamere, Sunil Kurian, Thomas Whisenant, Padmaja Natarajan, Katie Podshivalova, and Heather Kiyomi Komori; as well as the many other lab members I have worked with in small ways over the years. In addition, Steven Head, Dr. Phillip Ordoukhanian, and Terri Gelbart from the Scripps Genomics core have also been instrumental in supporting my work. And of course, I am thankful for the guidance and expertise provided by my committee, Drs. Nicholas Schork, Ali Torkamani, Michael Petrascheck, and Luc Teyton.

Finally, I wish to thank my parents, for instilling in me a love of science and learning from an early age and encouraging me to pursue that love as a career as I grew up. I am truly lucky to have such a loving and supportive family.

Contents

Copyright notice **i**

Dedication **ii**

Acknowledgements **iii**

Contents **iv**

List of Tables **x**

List of Figures **xi**

List of Abbreviations **xiv**

Abstract **xviii**

1 Introduction **1**

1.1 Biological motivation 1

1.1.1 Rejection is the major long-term threat to organ and tissue allografts 1

1.1.2 Diagnosis and treatment of allograft rejection is a major challenge 2

1.1.3 Memory cells are resistant to immune suppression 4

1.1.4 Infusion of allogenic mesenchymal stem cells modulates the alloimmune response 6

1.2 Overview of bioinformatic analysis methods 7
1.2.1 `Limma`: The standard linear modeling framework for genomics 7
1.2.2 `edgeR` provides `limma`-like analysis features for read count data 10
1.2.3 Calling consensus peaks from ChIP-seq data 12
1.2.4 Normalization of high-throughput data is non-trivial and application-dependent 14
1.2.5 ComBat and SVA for correction of known and unknown batch effects 17
1.2.6 Interpreting p-value distributions and estimating false discovery rates 20
1.3 Structure of the thesis 22

2 Reproducible genome-wide epigenetic analysis of H3K4 and H3K27 methylation in naïve and memory CD4$^+$ T-cell activation 24

2.1 Introduction 24
2.2 Approach 25
2.3 Methods 26
2.3.1 RNA-seq differential expression analysis 27
2.3.2 ChIP-seq analyses 28
2.3.3 MOFA analysis of cross-dataset variation patterns 34
2.4 Results 38
2.4.1 Interpretation of RNA-seq analysis is limited by a major confounding factor 38
2.4.2 H3K4 and H3K27 methylation occur in broad regions and are enriched near promoters 39
2.4.3 Correlations between gene expression and promoter methylation follow expected genome-wide trends 41

2.4.4 Gene expression and promoter histone methylation patterns show convergence between naïve and memory cells at day 14 . 43
2.4.5 Location of H3K4me2 and H3K4me3 promoter coverage associates with gene expression 45
2.4.6 Patterns of H3K27me3 promoter coverage associate with gene expression 51
2.5 Discussion 53
2.5.1 Each histone mark's "effective promoter extent" must be determined empirically 53
2.5.2 Day 14 convergence is consistent with naïve-to-memory differentiation 54
2.5.3 The location of histone modifications within the promoter is important 56
2.5.4 A reproducible workflow aids in analysis 57
2.6 Future Directions 59
2.6.1 Previous negative results 59
2.6.2 Improve on the idea of an effective promoter radius 59
2.6.3 Design experiments to focus on post-activation convergence of naïve & memory cells 61
2.6.4 Follow up on hints of interesting patterns in promoter relative coverage profiles 62
2.6.5 Investigate causes of high correlation between mutually exclusive histone marks 64

3 Improving array-based diagnostics for transplant rejection by optimizing data preprocessing 66

3.1 Introduction 66
3.1.1 Proper pre-processing is essential for array data 66

3.2 Approach 67
3.2.1 Clinical diagnostic applications for microarrays require single-channel normalization 67
3.2.2 Heteroskedasticity must be accounted for in methylation array data 69
3.3 Methods 71
3.3.1 Evaluation of classifier performance with different normalization methods 71
3.3.2 Generating custom fRMA vectors for hthgu133pluspm array platform 73
3.3.3 Modeling methylation array M-value heteroskedasticity with a modified voom implementation 74
3.4 Results 76
3.4.1 Separate normalization with RMA introduces unwanted biases in classification 76
3.4.2 fRMA and SCAN maintain classification performance while eliminating dependence on normalization strategy 76
3.4.3 fRMA with custom-generated vectors enables single-channel normalization on hthgu133pluspm platform 80
3.4.4 SVA, voom, and array weights improve model fit for methylation array data 86
3.5 Discussion 90
3.5.1 fRMA achieves clinically applicable normalization without sacrificing classification performance 90
3.5.2 Robust fRMA vectors can be generated for new array platforms 93

3.5.3 Methylation array data can be successfully analyzed using existing techniques, but machine learning poses additional challenges . . . 94
3.6 Future Directions . . . 96
3.6.1 Improving fRMA to allow training from batches of unequal size 96
3.6.2 Developing methylation arrays as a diagnostic tool for kidney transplant rejection . . . 97

4 Globin-blocking for more effective blood RNA-seq analysis in primate animal model 99

4.1 Introduction . . . 100
4.2 Approach . . . 101
4.3 Methods . . . 102
4.3.1 Sample collection . . . 102
4.3.2 Globin blocking oligonucleotide design . . . 102
4.3.3 RNA-seq library preparation . . . 103
4.3.4 Read alignment and counting . . . 104
4.3.5 Normalization and exploratory data analysis . . . 105
4.3.6 Differential expression analysis . . . 106
4.4 Results . . . 107
4.4.1 Globin blocking yields a larger and more consistent fraction of useful reads . . . 107
4.4.2 Globin blocking lowers the noise floor and allows detection of about 2000 more low-expression genes . . . 109
4.4.3 Globin blocking does not add significant additional noise or decrease sample quality . . . 112
4.4.4 More differentially expressed genes are detected with globin blocking . . . 116

4.5 Discussion 117
4.6 Future Directions 119

5 Conclusions **120**

5.1 Every high-throughput analysis presents unique analysis challenges . 121
5.2 Successful data analysis requires a toolbox, not a pipeline 122

Bibliography **125**

List of Tables

2.1 Estimated and detected differentially expressed genes. 35

2.2 Summary of peak-calling statistics. 39

2.3 Effective promoter radius for each histone mark. 41

2.4 Number of differentially modified promoters between naïve and memory cells at each time point after activation. 47

3.1 Summary of analysis variants for methylation array data. 75

3.2 ROC curve AUC values for internal and external validation with 6 different normalization strategies. 78

3.3 Association of sample weights with clinical covariates in methylation array data. 89

3.4 Estimates of degree of differential methylation in for each contrast in each analysis. 92

4.1 Fractions of reads mapping to genomic features in GB and non-GB samples. 108

4.2 Comparison of significantly differentially expressed genes with and without globin blocking. 116

List of Figures

1.1 Example of empirical Bayes squeezing of per-gene variances. 9

1.2 Example IDR consistency plot. 15

1.3 Example MA plot of ChIP-seq read counts in 10kb bins for two arbitrary samples. 18

1.4 Example p-value histogram. 21

2.1 **Overview of the experimental design.** 27

2.2 PCoA plots of RNA-seq data showing effect of batch correction. . . . 29

2.3 RNA-seq sample weights, grouped by experimental and technical covariates. 30

2.4 Strand cross-correlation plots for ChIP-seq data, before and after blacklisting. 31

2.5 PCoA plots of ChIP-seq sliding window data, before and after subtracting surrogate variables. 33

2.6 Promoter abundance filtering for relative coverage profiles. 36

2.7 MOFA latent factors identify shared patterns of variation. 37

2.8 PCoA plot of RNA-seq samples after ComBat batch correction. . . . 40

2.9 Enrichment of peaks in promoter neighborhoods. 42

2.10 Expression distributions of genes with and without promoter peaks. . 44

2.11 PCoA plots for promoter ChIP-seq and expression RNA-seq data . . 46

2.12 K-means clustering of promoter H3K4me2 relative coverage depth in naïve day 0 samples. 49
2.13 K-means clustering of promoter H3K4me3 relative coverage depth in naïve day 0 samples. 50
2.14 K-means clustering of promoter H3K27me3 relative coverage depth in naïve day 0 samples. 52
2.15 Lamere 2016 Figure 8 "Model for the role of H3K4 methylation during $CD4^+$ T-cell activation." 55
2.16 Dependency graph of steps in reproducible workflow. 58

3.1 Sigmoid shape of the mapping between β and M values. 70
3.2 Classifier probabilities on validation samples when normalized with RMA together vs. separately. 77
3.3 ROC curves for PAM using different normalization strategies. 79
3.4 Effect of batch size selection on number of batches and number of samples included in fRMA probe weight learning. 81
3.5 Violin plot of log ratios between normalizations for 20 biopsy samples. 83
3.6 Violin plot of log ratios between normalizations for 20 blood samples. 84
3.7 Representative MA plots comparing RMA and custom fRMA normalizations. 85
3.8 Mean-variance trend modeling in methylation array data. 87
3.9 Box-and-whiskers plot of sample quality weights grouped by diabetes diagnosis. 89
3.10 Probe p-value histograms for each contrast in each analysis. 91

4.1 Fraction of genic reads in each sample aligned to non-globin genes, with and without GB. 110

4.2 Distributions of average group gene abundances when normalized separately or together. 111
4.3 Gene detections as a function of abundance thresholds in GB and non-GB samples. 112
4.4 MA plot showing effects of GB on each gene's abundance. 114
4.5 Comparison of inter-sample gene abundance correlations with and without GB. 115

List of Abbreviations

ADNR acute dysfunction with no rejection

APC antigen-presenting cell

AR acute rejection

AUC area under ROC curve

BCV biological coefficient of variation

BH Benjamini-Hochberg

CAN chronic allograft nephropathy

ChIP chromatin immunoprecipitation

ChIP-seq chromatin immunoprecipitation followed by high-throughput DNA sequencing

CpGi CpG island

CPM counts per million

ENCODE Encyclopedia Of DNA Elements

FDR false discovery rate

FPKM fragments per kilobase per million fragments

fRMA frozen Robust Multichip Average

GB globin blocking

GEO Gene Expression Omnibus

GLM generalized linear model

GRCh38 Genome Reference Consortium Human Build 38

GRSN Global Rank-invariant Set Normalization

HTS high-throughput sequencing

ID identifier

IDR irreproducible discovery rate

IFNγ interferon gamma

LF latent factor

logCPM $\log_2$ counts per million

logFC $\log_2$ fold change

M-value $\log_2$ ratio

MACS Model-based Analysis of ChIP-seq

MHC major histocompatability complex

MOFA Multi-Omics Factor Analysis

mRNA messenger RNA

MSC mesenchymal stem cell

NB negative binomial

ncRNA non-coding RNA

oligo oligonucleotide

PAM Prediction Analysis for Microarrays

PC principal component

PCA principal component analysis

PCoA principal coordinate analysis

PCR polymerase chain reaction

RMA Robust Multichip Average

RNA-seq high-throughput RNA sequencing

ROC receiver operating characteristic

SCAN Single-Channel Array Normalization

SICER Spatial Clustering for Identification of ChIP-Enriched Regions

SRA Sequence Read Archive

SVA surrogate variable analysis

SVD singular value decomposition

SWAN subset-quantile within array normalization

T1D Type 1 diabetes

T2D Type 2 diabetes

TCR T-cell receptor

TMM trimmed mean of M-values

TSS transcription start site

TX healthy transplant

Abstract

Transplant rejection mediated by adaptive immune response is the major challenge to long-term graft survival. Rejection is treated with immune suppressive drugs, but early diagnosis is essential for effective treatment. Memory lymphocytes are known to resist immune suppression, but the precise regulatory mechanisms underlying immune memory are still poorly understood. High-throughput genomic assays such as microarrays, RNA-seq, and ChIP-seq are heavily used in the study of immunology and transplant rejection. Here we present 3 analyses of such assays in this context. First, we re-analyze a large data set consisting of H3K4me2, H3K4me3, and H3K27me3 ChIP-seq data and RNA-seq data in naïve and memory $CD4^+$ T-cells using modern bioinformatics methods designed to address deficiencies in the data and extend the analysis in several new directions. All 3 histone marks are found to occur in broad regions and are enriched near promoters, but the radius of promoter enrichment is found to be larger for H3K27me3. We observe that both gene expression and promoter histone methylation in naïve and memory cells converges on a common signature 14 days after activation, consistent with differentiation of naïve cells into memory cells. The location of histone modifications within the promoter is also found to be important, with asymmetric associations with gene expression for peaks located the same distance up- or downstream of the TSS. Second, we demonstrate the effectiveness of fRMA as a single-channel normalization for using expression arrays to diagnose transplant rejection in a clinical diagnostic setting, and we develop a custom fRMA normaliza-

tion for a previously unsupported array platform. For methylation arrays, we adapt methods designed for RNA-seq to improve the sensitivity of differential methylation analysis by modeling the heteroskedasticity inherent in the data. Finally, we present and validate a novel method for RNA-seq of cynomolgus monkey blood samples using complementary oligonucleotides to prevent wasteful over-sequencing of globin genes. These results all demonstrate the usefulness of a toolbox full of flexible and modular analysis methods in analyzing complex high-throughput assays in contexts ranging from basic science to translational medicine.

Chapter 1

Introduction

1.1 Biological motivation

1.1.1 Rejection is the major long-term threat to organ and tissue allografts

Organ and tissue transplants are a life-saving treatment for people who have lost the function of an important organ. In some cases, it is possible to transplant a patient's own tissue from one area of their body to another, referred to as an autograft. This is common for tissues that are distributed throughout many areas of the body, such as skin and bone. However, in cases of organ failure, there is no functional self tissue remaining, and a transplant from another person – a donor – is required. This is referred to as an allograft [1].

Because an allograft comes from a donor of the same species who is genetically distinct from the recipient (with rare exceptions), genetic variants in protein-coding regions affect the polypeptide sequences encoded by the affected genes, resulting in protein products in the allograft that differ from the equivalent proteins produced by the graft recipient's own tissue. As a result, without intervention, the recipient's immune system will eventually identify the graft as foreign tissue and begin attacking

it. This is called an alloimmune response, and if left unchecked, it eventually results in failure and death of the graft, a process referred to as transplant rejection [2]. Rejection is the primary obstacle to long-term health and survival of an allograft [1]. Like any adaptive immune response, an alloimmune response generally occurs via two broad mechanisms: cellular immunity, in which $CD8^+$ T-cells recognizing graft-specific antigens induce apoptosis in the graft cells; and humoral immunity, in which B-cells produce antibodies that bind to graft proteins and direct an immune response against the graft [2]. In either case, alloimmunity and rejection show most of the typical hallmarks of an adaptive immune response, in particular mediation by $CD4^+$ T-cells and formation of immune memory.

1.1.2 Diagnosis and treatment of allograft rejection is a major challenge

To prevent rejection, allograft recipients are treated with immune suppressive drugs [3, 2]. The goal is to achieve sufficient suppression of the immune system to prevent rejection of the graft without compromising the ability of the immune system to raise a normal response against infection. As such, a delicate balance must be struck: insufficient immune suppression may lead to rejection and ultimately loss of the graft; excessive suppression leaves the patient vulnerable to life-threatening opportunistic infections [2]. Because every patient's matabolism is different, achieving this delicate balance requires drug dosage to be tailored for each patient. Furthermore, dosage must be tuned over time, as the immune system's activity varies over time and in response to external stimuli with no fixed pattern. In order to properly adjust the dosage of immune suppression drugs, it is necessary to monitor the health of the transplant and increase the dosage if evidence of rejection or alloimmune activity is observed.

However, diagnosis of rejection is a significant challenge. Early diagnosis is essen-

tial in order to step up immune suppression before the immune system damages the graft beyond recovery [4]. The current gold standard test for graft rejection is a tissue biopsy, examined for visible signs of rejection by a trained histologist [5]. When a patient shows symptoms of possible rejection, a "for cause" biopsy is performed to confirm the diagnosis, and immune suppression is adjusted as necessary. However, in many cases, the early stages of rejection are asymptomatic, known as "sub-clinical" rejection. In light of this, is is now common to perform "protocol biopsies" at specific times after transplantation of a graft, even if no symptoms of rejection are apparent, in addition to "for cause" biopsies [6, 7, 8, 9].

However, biopsies have a number of downsides that limit their effectiveness as a diagnostic tool. First, the need for manual inspection by a histologist means that diagnosis is subject to the biases of the particular histologist examining the biopsy [5]. In marginal cases, two different histologists may give two different diagnoses to the same biopsy. Second, a biopsy can only evaluate if rejection is occurring in the section of the graft from which the tissue was extracted. If rejection is localized to one section of the graft and the tissue is extracted from a different section, a false negative diagnosis may result. Most importantly, extraction of tissue from a graft is invasive and is treated as an injury by the body, which results in inflammation that in turn promotes increased immune system activity. Hence, the invasiveness of biopsies severely limits the frequency with which they can safely be performed [8]. Typically, protocol biopsies are not scheduled more than about once per month [7]. A less invasive diagnostic test for rejection would bring manifold benefits. Such a test would enable more frequent testing and therefore earlier detection of rejection events. In addition, having a larger pool of historical data for a given patient would make it easier to evaluate when a given test is outside the normal parameters for that specific patient, rather than relying on normal ranges for the population as a whole. Lastly, the accumulated data from more frequent tests would be a boon to

the transplant research community. Beyond simply providing more data overall, the better time granularity of the tests will enable studying the progression of a rejection event on the scale of days to weeks, rather than months.

1.1.3 Memory cells are resistant to immune suppression

One of the defining features of the adaptive immune system is immune memory: the ability of the immune system to recognize a previously encountered foreign antigen and respond more quickly and more strongly to that antigen in subsequent encounters [2]. When the immune system first encounters a new antigen, the T-cells that respond are known as naïve cells – T-cells that have never detected their target antigens before. Once activated by their specific antigen presented by an antigen-presenting cell in the proper co-stimulatory context, naïve cells differentiate into effector cells that carry out their respective functions in targeting and destroying the source of the foreign antigen. The T-cell receptor (TCR) is cell-surface protein complex produced by T-cells that is responsible for recognizing the T-cell's specific antigen, presented on a major histocompatability complex (MHC), the cell-surface protein complex used by an antigen-presenting cell (APC) to present antigens to the T-cell. However, a naïve T-cell that recognizes its antigen also requires a co-stimulatory signal, delivered through other interactions between APC surface proteins and T-cell surface proteins such as CD28. Without proper co-stimulation, a T-cell that recognizes its antigen either dies or enters an unresponsive state known as anergy, in which the T-cell becomes much more resistant to subsequent activation even with proper co-stimulation. The dependency of activation on co-stimulation is an important feature of naïve lymphocytes that limits "false positive" immune responses against self antigens, because APCs usually only express the proper co-stimulation after the innate immune system detects signs of an active infection, such as the presence of common bacterial cell components or inflamed tissue.

After the foreign antigen is cleared, most effector cells die since they are no longer needed, but some differentiate into memory cells and remain alive indefinitely. Like naïve cells, memory cells respond to detection of their specific antigen by differentiating into effector cells, ready to fight an infection [2]. However, the memory response to antigen is qualitatively different: memory cells are more sensitive to detection of their antigen, and a lower concentration of antigen is suffiicient to activate them [10, 11, 12]. In addition, memory cells are much less dependent on co-stimulation for activation: they can activate without certain co-stimulatory signals that are required by naïve cells, and the signals they do require are only required at lower levels in order to cause activation [11]. Furthermore, mechanisms that induce tolerance (non-response to antigen) in naïve cells are much less effective on memory cells [11]. Lastly, once activated, memory cells proliferate and differentiate into effector cells more quickly than naïve cells do [12]. In combination, these changes in lymphocyte behavior upon differentiation into memory cells account for the much quicker and stronger response of the immune system to subsequent exposure to a previously-encountered antigen.

In the context of a pathogenic infection, immune memory is a major advantage, allowing an organism to rapidly fight off a previously encountered pathogen much more quickly and effectively than the first time it was encountered [2]. However, if effector cells that recognize an antigen from an allograft are allowed to differentiate into memory cells, preventing rejection of the graft becomes much more difficult. Many immune suppression drugs work by interfering with the co-stimulation that naïve cells require in order to mount an immune response. Since memory cells do not require the same degree of co-stimulation, these drugs are not effective at suppressing an immune response that is mediated by memory cells. Secondly, because memory cells are able to mount a stronger and faster response to an antigen, all else being equal stronger immune suppression is required to prevent an immune response mediated by memory cells.

However, immune suppression affects the entire immune system, not just cells recognizing a specific antigen, so increasing the dosage of immune suppression drugs also increases the risk of complications from a compromised immune system, such as opportunistic infections [2]. While the differences in cell surface markers between naïve and memory cells have been fairly well characterized, the internal regulatory mechanisms that allow memory cells to respond more quickly and without co-stimulation are still poorly understood. In order to develop methods of immune suppression that either prevent the formation of memory cells or work more effectively against memory cells, a more complete understanding of the mechanisms of immune memory formation and regulation is required.

1.1.4 Infusion of allogenic mesenchymal stem cells modulates the alloimmune response

One promising experimental treatment for transplant rejection involves the infusion of allogenic mesenchymal stem cells (MSCs). MSCs have been shown to have immune modulatory effects, both in general and specifically in the case of immune responses against allografts [13, 14, 15, 16]. Furthermore, allogenic MSCs themselves are immune-evasive and are rejected by the recipient's immune system more slowly than most allogenic tissues [17, 18]. In addition, treating MSCs in culture with interferon gamma (IFNγ) is shown to enhance their immunosuppressive properties and homogenize their cellulat phenotype, making them more amenable to development into a well-controlled treatment [19, 20]. The mechanisms by which MSCs modulate the immune system are still poorly understood. Despite this, there is signifcant interest in using IFNγ-activated MSC infusion as a supplementary immune suppressive treatment for allograft transplantation.

Note that despite the name, none of the above properties of MSCs are believed to involve their ability as stem cells to differentiate into multiple different mature cell

types, but rather the intercellular signals they produce [17].

1.2 Overview of bioinformatic analysis methods

The studies presented in this work all involve the analysis of high-throughput genomic and epigenomic assay data. Assays like microarrays and high-throughput sequencing (HTS) are powerful methods for interrogating gene expression and epigenetic state across the entire genome. However, these data present many unique analysis challenges, and proper analysis requires identifying and exploiting genome-wide trends in the data to make up for the small sample sizes. A wide array of software tools is available to analyze these data. This section presents an overview of the most important methods and tools used throughout the following analyses, including what problems they solve, what assumptions they make, and a basic description of how they work.

1.2.1 `Limma`: The standard linear modeling framework for genomics

Linear models are a generalization of the t-test and ANOVA to arbitrarily complex experimental designs [21]. In a typical linear model, there is one dependent variable observation per sample and a large number of samples. For example, in a linear model of height as a function of age and sex, there is one height measurement per person. However, when analyzing genomic data, each sample consists of observations of thousands of dependent variables. For example, in a high-throughput RNA sequencing (RNA-seq) experiment, the dependent variables may be the count of RNA-seq reads for each annotated gene, and there are tens of thousands of genes in the human genome. Since many assays measure other things than gene expression, the abstract term "feature" is used to refer to each dependent variable being measured, which may

include any genomic element, such as genes, promoters, peaks, enhancers, exons, etc.

The simplest approach to analyzing such data would be to fit the same model independently to each feature. However, this is undesirable for most genomics data sets. Genomics assays like HTS are expensive, and often the process of generating the samples is also quite expensive and time-consuming. This expense limits the sample sizes typically employed in genomics experiments, so a typical genomic data set has far more features being measured than observations (samples) per feature. As a result, the statistical power of the linear model for each individual feature is likewise limited by the small number of samples. However, because thousands of features from the same set of samples are analyzed together, there is an opportunity to improve the statistical power of the analysis by exploiting shared patterns of variation across features. This is the core feature of `limma`, a linear modeling framework designed for genomic data. `Limma` is typically used to analyze expression microarray data, and more recently RNA-seq data, but it can also be used to analyze any other data for which linear modeling is appropriate.

The central challenge when fitting a linear model is to estimate the variance of the data accurately. Out of all parameters required to evaluate statistical significance of an effect, the variance is the most difficult to estimate when sample sizes are small. A single shared variance could be estimated for all of the features together, and this estimate would be very stable, in contrast to the individual feature variance estimates. However, this would require the assumption that all features have equal variance, which is known to be false for most genomic data sets (for example, some genes' expression is known to be more variable than others'). `Limma` offers a compromise between these two extremes by using a method called empirical Bayes moderation to "squeeze" the distribution of estimated variances toward a single common value that represents the variance of an average feature in the data (Figure 1.1) [22]. While the individual feature variance estimates are not stable, the common variance estimate for

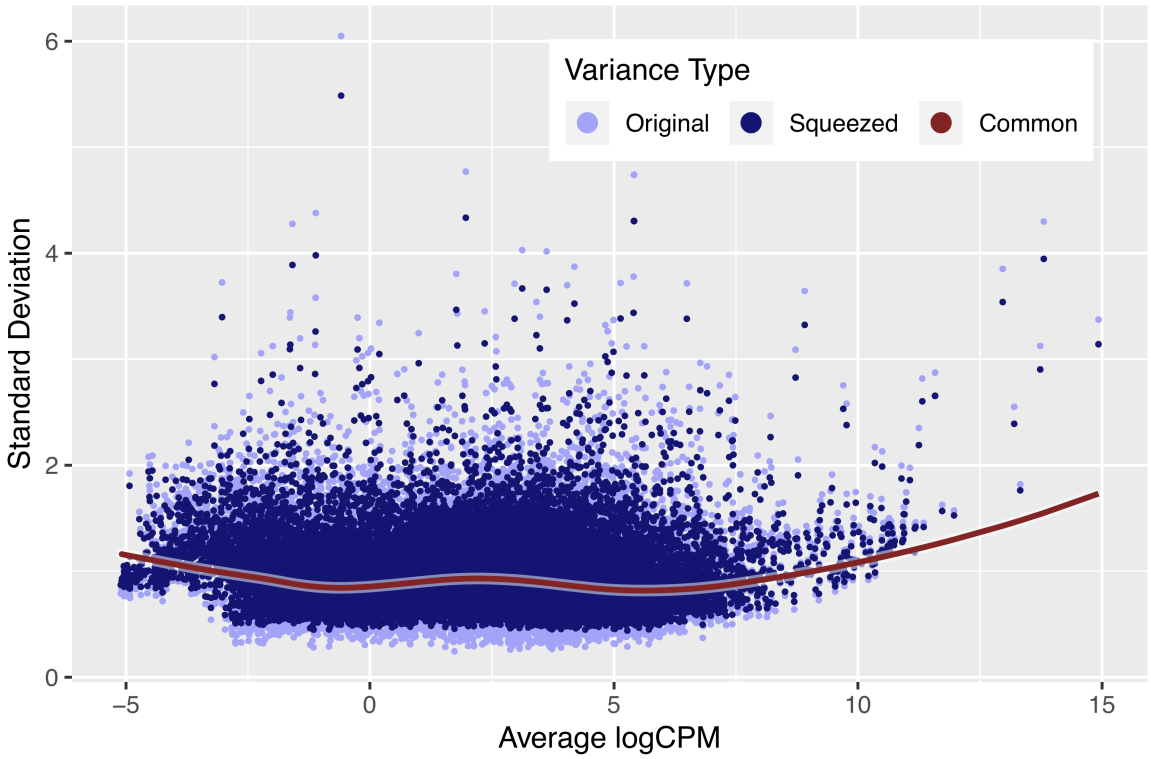


Figure 1.1: **Example of empirical Bayes squeezing of per-gene variances.** A smooth trend line (red) is fitted to the individual gene variances (light blue) as a function of average gene abundance (logCPM). Then the individual gene variances are "squeezed" toward the trend (dark blue).

the entire data set is quite stable, so using a combination of the two yields a variance estimate for each feature with greater precision than the individual feature variances. The trade-off for this improvement is that squeezing each estimated variance toward the common value introduces some bias – the variance will be underestimated for features with high variance and overestimated for features with low variance. Essentially, `limma` assumes that extreme variances are less common than variances close to the common value. The squeezed variance estimates from this empirical Bayes procedure are shown empirically to yield greater statistical power than either the individual feature variances or the single common value.

On top of this core framework, `limma` also implements many other enhancements that, further relax the assumptions of the model and extend the scope of what kinds of data it can analyze. Instead of squeezing toward a single common variance value,

`limma` can model the common variance as a function of a covariate, such as average expression [23]. This is essential for RNA-seq data, where higher gene counts yield more precise expression measurements and therefore smaller variances than low-count genes. While linear models typically assume that all samples have equal variance, `limma` is able to relax this assumption by identifying and down-weighting samples that diverge more strongly from the linear model across many features [24, 25]. In addition, `limma` is also able to fit simple mixed models incorporating one random effect in addition to the fixed effects represented by an ordinary linear model [26]. Once again, `limma` shares information between features to obtain a robust estimate for the random effect correlation.

1.2.2 `edgeR` provides `limma`-like analysis features for read count data

Although `limma` can be applied to read counts from RNA-seq data, it is less suitable for counts from chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq) and other sources, which tend to be much smaller and therefore violate the assumption of a normal distribution more severely. For all count-based data, the `edgeR` package works similarly to `limma`, but uses a generalized linear model (GLM) instead of a linear model. Relative to a linear model, a GLM gains flexibility by relaxing several assumptions, the most important of which is the assumption of normally distributed errors. This allows the GLM in `edgeR` to model the counts directly using a negative binomial (NB) distribution rather than modeling the normalized log counts using a normal distribution as `limma` does [27, 28, 29].

The NB distribution is a good fit for count data because it can be derived as a gamma-distributed mixture of Poisson distributions. The reads in an RNA-seq sample are assumed to be sampled from a much larger population, such that the sampling process does not significantly affect the proportions. Under this assumption, a gene's

read count in an RNA-seq sample is distributed as Binomial(n, p), where n is the total number of reads sequenced from the sample and p is the proportion of total fragments in the sample derived from that gene. When n is large and p is small, a Binomial(n, p) distribution is well-approximated by Poisson(np). Hence, if multiple sequencing runs are performed on the same RNA-seq sample (with the same gene mixing proportions each time), each gene's read count is expected to follow a Poisson distribution. If the abundance of a gene, p, varies across biological replicates according to a gamma distribution, and n is held constant, then the result is a gamma-distributed mixture of Poisson distributions, which is equivalent to the NB distribution. The assumption of a gamma distribution for the mixing weights is arbitrary, motivated by the convenience of the numerically tractable NB distribution and the need to select *some* distribution, since the true shape of the distribution of biological variance is unknown.

Thus, `edgeR`'s use of the NB is equivalent to an *a priori* assumption that the variation in gene abundances between replicates follows a gamma distribution. The gamma shape parameter in the context of the NB is called the dispersion, and the square root of this dispersion is referred to as the biological coefficient of variation (BCV), since it represents the variability in abundance that was present in the biological samples prior to the Poisson "noise" that was generated by the random sampling of reads in proportion to feature abundances. Like `limma`, `edgeR` estimates the BCV for each feature using an empirical Bayes procedure that represents a compromise between per-feature dispersions and a single pooled dispersion estimate shared across all features. For differential abundance testing, `edgeR` offers a likelihood ratio test based on the NB GLM. However, this test assumes the dispersion parameter is known exactly rather than estimated from the data, which can result in overstating the significance of differential abundance results. More recently, a quasi-likelihood test has been introduced that properly factors the uncertainty in dispersion estimation into the estimates of statistical significance, and this test is recommended over the likelihood

ratio test in most cases [30].

1.2.3 Calling consensus peaks from ChIP-seq data

Unlike RNA-seq data, in which gene annotations provide a well-defined set of discrete genomic regions in which to count reads, ChIP-seq reads can potentially occur anywhere in the genome. However, most genome regions will not contain significant ChIP-seq read coverage, and analyzing every position in the entire genome is statistically and computationally infeasible, so it is necessary to identify regions of interest inside which ChIP-seq reads will be counted and analyzed. One option is to define a set of interesting regions *a priori*, for example by defining a promoter region for each annotated gene. However, it is also possible to use the ChIP-seq data itself to identify regions with ChIP-seq read coverage significantly above the background level, known as peaks.

The challenge in peak calling is that the immunoprecipitation step is not 100% selective, so some fraction of reads are *not* derived from DNA fragments that were bound by the immunoprecipitated protein. These are referred to as background reads. Biases in amplification and sequencing, as well as the aforementioned Poisson randomness of the sequencing itself, can cause fluctuations in the background level of reads that resemble peaks, and the true peaks must be distinguished from these. It is common to sequence the input DNA to the ChIP-seq reaction alongside the immunoprecipitated product in order to aid in estimating the fluctuations in background level across the genome.

There are generally two kinds of peaks that can be identified: narrow peaks and broadly enriched regions. Proteins that bind specific sites in the genome (such as many transcription factors) typically show most of their ChIP-seq read coverage at these specific sites and very little coverage anywhere else. Because the footprint of the protein is consistent wherever it binds, each peak has a consistent width, typically tens

to hundreds of base pairs, representing the length of DNA that it binds to. Algorithms like Model-based Analysis of ChIP-seq (MACS) exploit this pattern to identify specific loci at which such "narrow peaks" occur by looking for the characteristic peak shape in the ChIP-seq coverage rising above the surrounding background coverage [31]. In contrast, some proteins, chief among them histones, do not bind only at a small number of specific sites, but rather bind potentially almost everywhere in the entire genome. When looking at histone marks, adjacent histones tend to be similarly marked, and a given mark may be present on an arbitrary number of consecutive histones along the genome. Hence, there is no consistent "footprint size" for ChIP-seq peaks based on histone marks, and peaks typically span many histones. Hence, typical peaks span many hundreds or even thousands of base pairs. Instead of identifying specific loci of strong enrichment, algorithms like Spatial Clustering for Identification of ChIP-Enriched Regions (SICER) assume that peaks are represented in the ChIP-seq data by modest enrichment above background occurring across broad regions, and they attempt to identify the extent of those regions [32].

Regardless of the type of peak identified, it is important to identify peaks that occur consistently across biological replicates. The Encyclopedia Of DNA Elements (ENCODE) project has developed a method called irreproducible discovery rate (IDR) for this purpose [33]. The IDR is defined as the probability that a peak identified in one biological replicate will *not* also be identified in a second replicate. Where the more familiar false discovery rate measures the degree of correspondence between a data-derived ranked list and the (unknown) true list of significant features, IDR instead measures the degree of correspondence between two ranked lists derived from different data. IDR assumes that the highest-ranked features are "signal" peaks that tend to be listed in the same order in both lists, while the lowest-ranked features are essentially noise peaks, listed in random order with no correspondence between the lists. IDR attempts to locate the "crossover point" between the signal and the

noise by determining how far down the list the rank consistency breaks down into randomness (Figure 1.2).

In addition to other considerations, if called peaks are to be used as regions of interest for differential abundance analysis, then care must be taken to call peaks in a way that is blind to differential abundance between experimental conditions, or else the statistical significance calculations for differential abundance will overstate their confidence in the results. The `csaw` package provides guidelines for calling peaks in this way: peaks are called based on a combination of all ChIP-seq reads from all experimental conditions, so that the identified peaks are based on the average abundance across all conditions, which is independent of any differential abundance between conditions [34].

1.2.4 Normalization of high-throughput data is non-trivial and application-dependent

High-throughput data sets invariably require some kind of normalization before further analysis can be conducted. In general, the goal of normalization is to remove effects in the data that are caused by technical factors that have nothing to do with the biology being studied.

For Affymetrix expression arrays, the standard normalization algorithm used in most analyses is Robust Multichip Average (RMA) [35]. RMA is designed with the assumption that some fraction of probes on each array will be artifactual and takes advantage of the fact that each gene is represented by multiple probes by implementing normalization and summarization steps that are robust against outlier probes. However, RMA uses the probe intensities of all arrays in the data set in the normalization of each individual array, meaning that the normalized expression values in each array depend on every array in the data set, and will necessarily change each time an array is added or removed from the data set. If this is undesirable, frozen

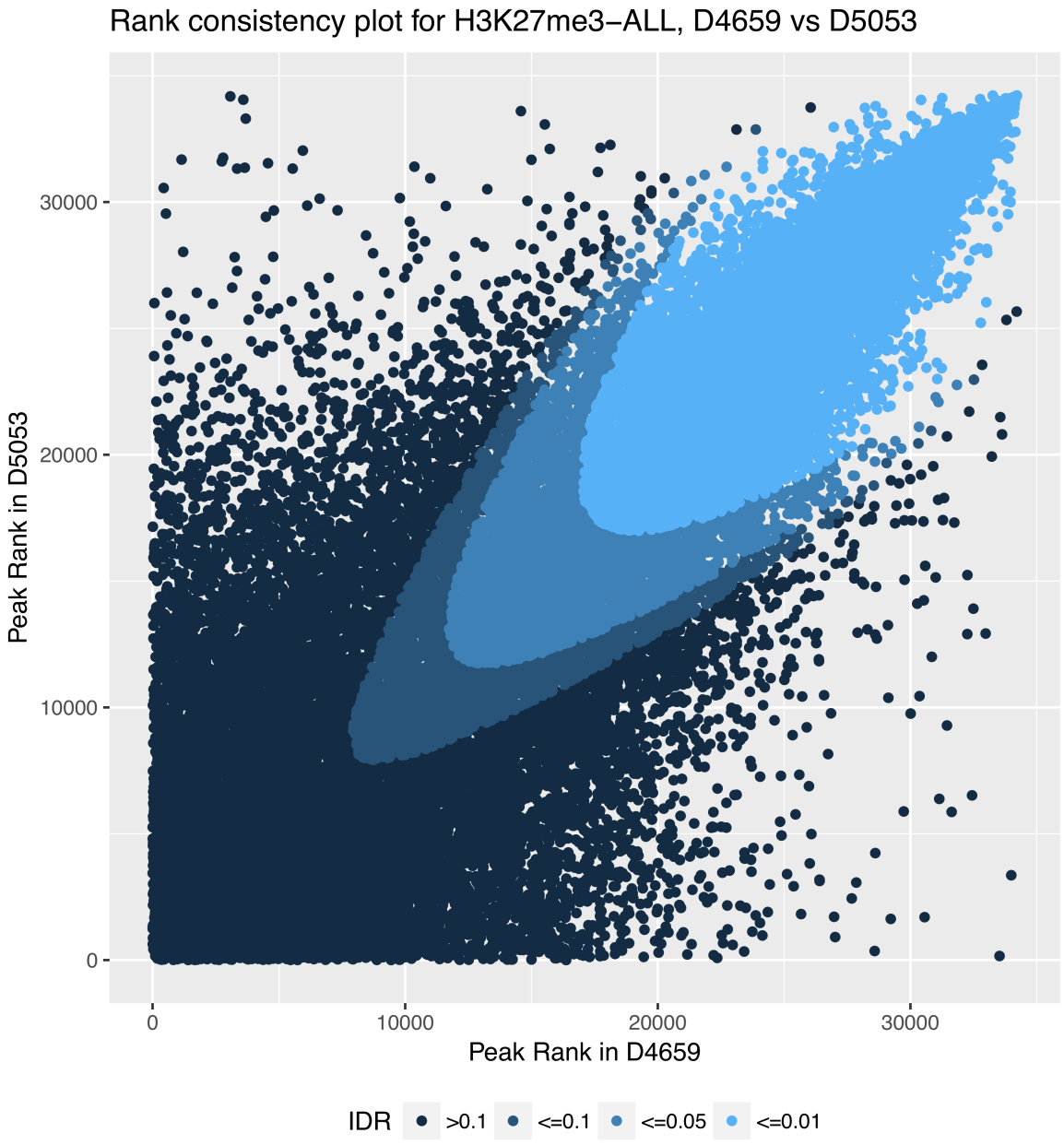


Figure 1.2: **Example IDR consistency plot.** Peak calls in two replicates are ranked from highest score (top and right) to lowest score (bottom and left). IDR identifies reproducible peaks, which rank highly in both replicates (light blue), separating them from "noise" peak calls whose ranking is not reproducible between replicates (dark blue).

Robust Multichip Average (fRMA) implements a variant of RMA where the relevant distributional parameters are learned from a large reference set of diverse public array data sets and then "frozen", so that each array is effectively normalized against this frozen reference set rather than the other arrays in the data set under study [36]. Other available array normalization methods considered include dChip, Global Rank-invariant Set Normalization (GRSN), and Single-Channel Array Normalization (SCAN) [37, 38, 39].

In contrast, HTS data present very different normalization challenges. The simplest case is RNA-seq in which read counts are obtained for a set of gene annotations, yielding a matrix of counts with rows representing genes and columns representing samples. Because RNA-seq approximates a process of sampling from a population with replacement, each gene's count is only interpretable as a fraction of the total reads for that sample. For that reason, RNA-seq abundances are often reported as counts per million (CPM). Furthermore, if the abundance of a single gene increases, then in order for its fraction of the total reads to increase, all other genes' fractions must decrease to accommodate it. This effect is known as composition bias, and it is an artifact of the read sampling process that has nothing to do with the biology of the samples and must therefore be normalized out. The most commonly used methods to normalize for composition bias in RNA-seq data seek to equalize the average gene abundance across samples, under the assumption that the average gene is likely not changing [40, 41]. The effect of such normalizations is to center the distribution of $\log_2$ fold changes (logFCs) at zero. Note that if a true global difference in gene expression is present in the data, this difference will be normalized out as well, since it is indistinguishable from composition bias. In other words, RNA-seq cannot measure absolute gene expression, only gene expression as a fraction of total reads.

In ChIP-seq data, normalization is not as straightforward. The `csaw` package implements several different normalization strategies and provides guidance on when

to use each one [34]. Briefly, a typical ChIP-seq sample has a bimodal distribution of read counts: a low-abundance mode representing background regions and a high-abundance mode representing signal regions. This offers two mutually incompatible normalization strategies: equalizing background coverage or equalizing signal coverage (Figure 1.3). If the experiment is well controlled and chromatin immunoprecipitation (ChIP) efficiency is known to be consistent across all samples, then normalizing the background coverage to be equal across all samples is a reasonable strategy. If this is not a safe assumption, then the preferred strategy is to normalize the signal regions in a way similar to RNA-seq data by assuming that the average signal region is not changing abundance between samples. Beyond this, if a ChIP-seq experiment has a more complicated structure that doesn't show the typical bimodal count distribution, it may be necessary to implement a normalization as a smooth function of abundance. However, this strategy makes a much stronger assumption about the data: that the average logFC is zero across all abundance levels. Hence, the simpler scaling normalization based on background or signal regions are generally preferred whenever possible.

1.2.5 ComBat and SVA for correction of known and unknown batch effects

In addition to well-understood effects that can be easily normalized out, a data set often contains confounding biological effects that must be accounted for in the modeling step. For instance, in an experiment with pre-treatment and post-treatment samples of cells from several different donors, donor variability represents a known batch effect. The most straightforward correction for known batches is to estimate the mean for each batch independently and subtract out the differences, so that all batches have identical means for each feature. However, as with variance estimation, estimating the differences in batch means is not necessarily robust at the feature

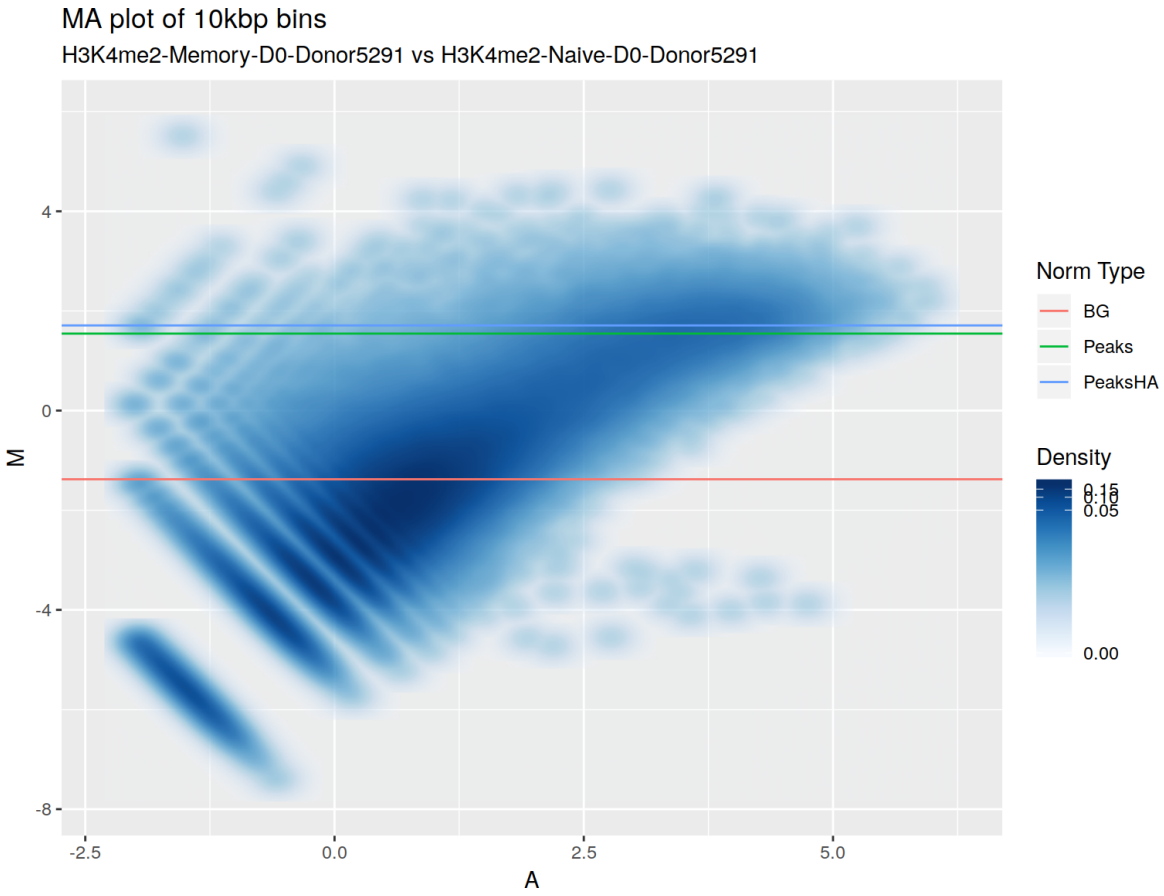


Figure 1.3: **Example MA plot of ChIP-seq read counts in 10kb bins for two arbitrary samples.** The distribution of bins is bimodal along the x axis (average abundance), with the left mode representing "background" regions with no protein binding and the right mode representing bound regions. The modes are also separated on the y axis (logFC), motivating two conflicting normalization strategies: background normalization (red) and signal normalization (blue and green, two similar signal normalizations).

level, so the ComBat method adds empirical Bayes squeezing of the batch mean differences toward a common value, analogous to `limma`'s empirical Bayes squeezing of feature variance estimates [42]. Effectively, ComBat assumes that modest differences between batch means are real batch effects, but extreme differences between batch means are more likely to be the result of outlier observations that happen to line up with the batches rather than a genuine batch effect. The result is a batch correction that is more robust against outliers than simple subtraction of mean differences.

In some data sets, unknown batch effects may be present due to inherent variability in the data, either caused by technical or biological effects. Examples of unknown batch effects include variations in enrichment efficiency between ChIP-seq samples, variations in populations of different cell types, and the effects of uncontrolled environmental factors on gene expression in humans or live animals. In an ordinary linear model context, unknown batch effects cannot be inferred and must be treated as random noise. However, in high-throughput experiments, once again information can be shared across features to identify patterns of un-modeled variation that are repeated in many features. One attractive strategy would be to perform singular value decomposition (SVD) on the matrix of linear model residuals (which contain all the un-modeled variation in the data) and take the first few singular vectors as batch effects. While this can be effective, it makes the unreasonable assumption that all batch effects are completely uncorrelated with any of the effects being modeled. surrogate variable analysis (SVA) starts with this approach, but takes some additional steps to identify batch effects in the full data that are both highly correlated with the singular vectors in the residuals and least correlated with the effects of interest [43]. Since the final batch effects are estimated from the full data, moderate correlations between the batch effects and effects of interest are allowed, which gives SVA much more freedom to estimate the true extent of the batch effects compared to simple residual SVD. Once the surrogate variables are estimated, they can be included as

coefficients in the linear model in a similar fashion to known batch effects in order to subtract out their effects on each feature's abundance.

1.2.6 Interpreting p-value distributions and estimating false discovery rates

When testing thousands of genes for differential expression or performing thousands of statistical tests for other kinds of genomic data, the result is thousands of p-values. By construction, p-values have a Uniform$(0, 1)$ distribution under the null hypothesis. This means that if all null hypotheses are true in a large number N of tests, then for any significance threshold T, approximately $N * T$ p-values would be called "significant" at that threshold even though the null hypotheses are all true. These are called false discoveries.

When only a fraction of null hypotheses are true, the p-value distribution will be a mixture of a uniform component representing the null hypotheses that are true and a non-uniform component representing the null hypotheses that are not true (Figure 1.4). The fraction belonging to the uniform component is referred to as π_0, which ranges from 1 (all null hypotheses true) to 0 (all null hypotheses false). Furthermore, the non-uniform component must be biased toward zero, since any evidence against the null hypothesis pushes the p-value for a test toward zero. We can exploit this fact to estimate the false discovery rate (FDR) for any significance threshold by estimating the degree to which the density of p-values left of that threshold exceeds what would be expected for a uniform distribution. In genomics, the most commonly used FDR estimation method, and the one used in this work, is that of Benjamini and Hochberg [44]. This is a conservative method that effectively assumes $\pi_0 = 1$. Hence it gives an estimated upper bound for the FDR at any significance threshold, rather than a point estimate.

We can also estimate π_0 for the entire distribution of p-values, which can give an

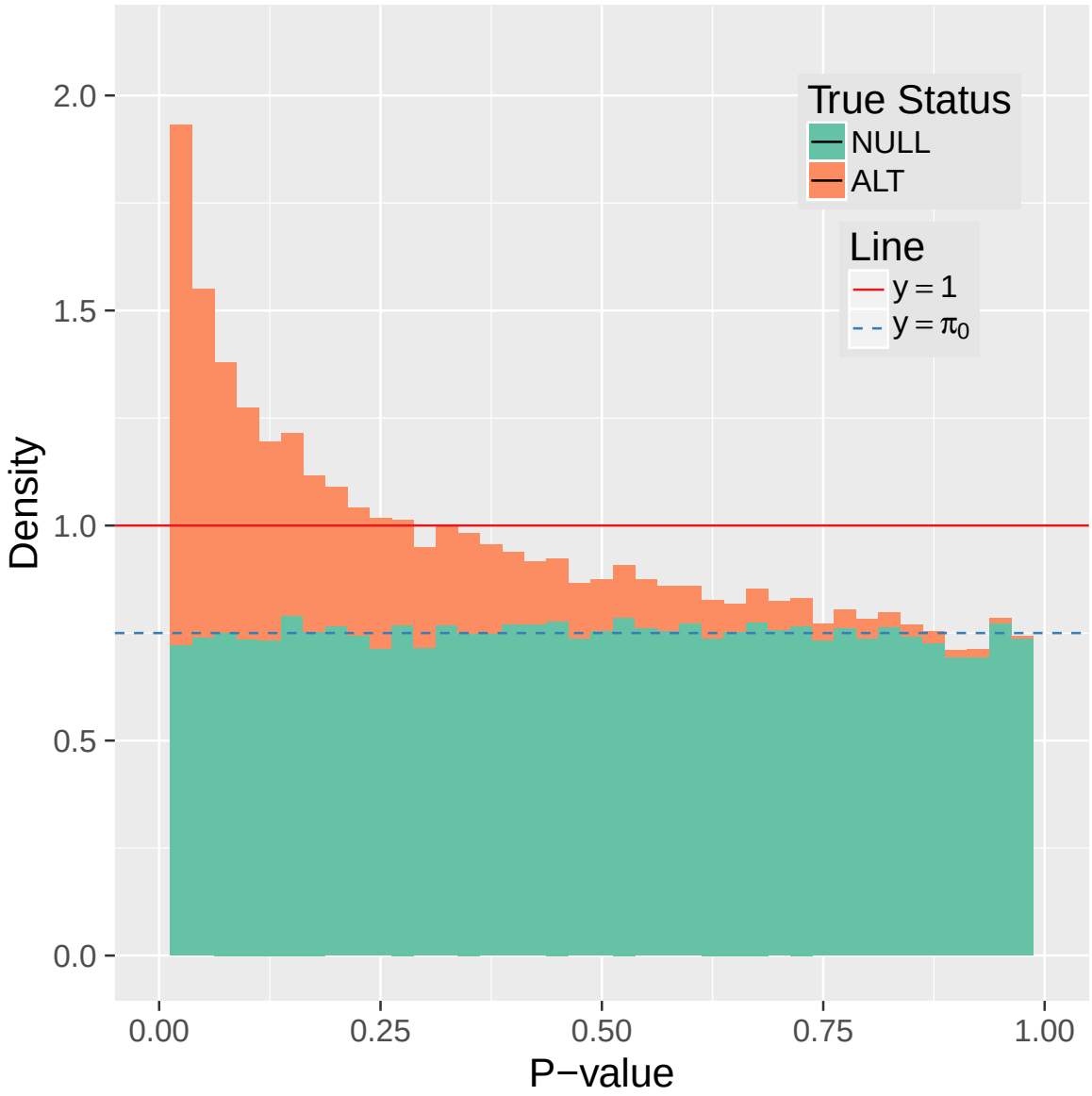


Figure 1.4: **Example p-value histogram.** The distribution of p-values from a large number of independent tests (such as differential expression tests for each gene in the genome) is a mixture of a uniform component representing the null hypotheses that are true (blue shading) and a zero-biased component representing the null hypotheses that are false (red shading). The FDR for any column in the histogram is the fraction of that column that is blue. The line $y = \pi_0$ represents the theoretical uniform component of this p-value distribution, while the line $y = 1$ represents the uniform component when all null hypotheses are true. Note that in real data, the true status of each hypothesis is unknown, so only the overall shape of the distribution is known.

idea of the overall signal size in the data without setting any significance threshold or making any decisions about which specific null hypotheses to reject. As FDR estimation, there are many methods proposed for estimating π_0. The one used in this work is the Phipson method of averaging local FDR values [45]. Once π_0 is estimated, the number of null hypotheses that are false can be estimated as $(1 - \pi_0) * N$.

Conversely, a p-value distribution that is neither uniform nor zero-biased is evidence of a modeling failure. Such a distribution would imply that there is less than zero evidence against the null hypothesis, which is not possible (in a frequentist setting). Attempting to estimate π_0 from such a distribution would yield an estimate greater than 1, a nonsensical result. The usual cause of a poorly-behaving p-value distribution is a model assumption that is violated by the data, such as assuming equal variance between groups (homoskedasticity) when the variance of each group is not equal (heteroskedasticity) or failing to model a strong confounding batch effect. In particular, such a p-value distribution is *not* consistent with a simple lack of signal in the data, as this should result in a uniform distribution. Hence, observing such a p-value distribution should prompt a search for violated model assumptions.

1.3 Structure of the thesis

This thesis presents 3 instances of using high-throughput genomic and epigenomic assays to investigate hypotheses or solve problems relating to the study of transplant rejection. In Chapter 2, ChIP-seq and RNA-seq are used to investigate the dynamics of promoter histone methylation as it relates to gene expression in T-cell activation and memory. Chapter 3 looks at several array-based assays with the potential to diagnose transplant rejection and shows that analyses of this array data are greatly improved by paying careful attention to normalization and preprocessing. Chapter 4 presents a custom method for improving RNA-seq of non-human primate blood

samples by preventing reverse transcription of unwanted globin transcripts. Finally, Chapter 5 summarizes the overarching lessons and strategies learned through these analyses that can be applied to all future analyses of high-throughput genomic assays.

Chapter 2

Reproducible genome-wide epigenetic analysis of H3K4 and H3K27 methylation in naïve and memory $CD4^+$ T-cell activation

Ryan C. Thompson, Sarah A. Lamere, Daniel R. Salomon

2.1 Introduction

$CD4^+$ T-cells are central to all adaptive immune responses, as well as immune memory [2]. After an infection is cleared, a subset of the naïve $CD4^+$ T-cells that responded to that infection differentiate into memory $CD4^+$ T-cells, which are responsible for responding to the same pathogen in the future. Memory $CD4^+$ T-cells are functionally distinct, able to respond to an infection more quickly and without the co-stimulation required by naïve $CD4^+$ T-cells. However, the molecular mechanisms underlying this functional distinction are not well-understood. Epigenetic regulation via histone

modification is thought to play an important role, but while many studies have looked at static snapshots of histone methylation in T-cells, few studies have looked at the dynamics of histone regulation after T-cell activation, nor the differences in histone methylation between naïve and memory T-cells. H3K4me2, H3K4me3 and H3K27me3 are three histone marks thought to be major epigenetic regulators of gene expression. The goal of the present study is to investigate the role of these histone marks in $CD4^+$ T-cell activation kinetics and memory differentiation. In static snapshots, H3K4me2 and H3K4me3 are often observed in the promoters of highly transcribed genes, while H3K27me3 is more often observed in promoters of inactive genes with little to no transcription occurring. As a result, the two H3K4 marks have been characterized as "activating" marks, while H3K27me3 has been characterized as "deactivating". Despite these characterizations, the actual causal relationship between these histone modifications and gene transcription is complex and likely involves positive and negative feedback loops between the two.

2.2 Approach

In order to investigate the relationship between gene expression and these histone modifications in the context of naïve and memory $CD4^+$ T-cell activation, a previously published data set of high-throughput RNA sequencing (RNA-seq) data and chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq) data was re-analyzed using up-to-date methods designed to address the specific analysis challenges posed by this data set. The data set contains naïve and memory $CD4^+$ T-cell samples in a time course before and after activation. Like the original analysis, this analysis looks at the dynamics of these histone marks and compares them to gene expression dynamics at the same time points during activation, as well as compares them between naïve and memory cells, in hope of discovering evidence

of new mechanistic details in the interplay between them. The original analysis of this data treated each gene promoter as a monolithic unit and mostly assumed that ChIP-seq reads or peaks occurring anywhere within a promoter were equivalent, regardless of where they occurred relative to the gene structure. For an initial analysis of the data, this was a necessary simplifying assumption. The current analysis aims to relax this assumption, first by directly analyzing ChIP-seq peaks for differential modification, and second by taking a more granular look at the ChIP-seq read coverage within promoter regions to ask whether the location of histone modifications relative to the gene's transcription start site (TSS) is an important factor, as opposed to simple proximity.

2.3 Methods

A reproducible workflow was written to analyze the raw ChIP-seq and RNA-seq data from previous studies (Gene Expression Omnibus (GEO) accession number GSE73214) [46, 47, 48, 49]. Briefly, this data consists of RNA-seq and ChIP-seq from $CD4^+$ T-cells from 4 donors. From each donor, naïve and memory $CD4^+$ T-cells were isolated separately. Then cultures of both cells were activated with CD3/CD28 beads, and samples were taken at 4 time points: Day 0 (pre-activation), Day 1 (early activation), Day 5 (peak activation), and Day 14 (post-activation). For each combination of cell type and time point, RNA was isolated and sequenced, and ChIP-seq was performed for each of 3 histone marks: H3K4me2, H3K4me3, and H3K27me3. The ChIP-seq input DNA was also sequenced for each sample. The result was 32 samples for each assay (see Figure 2.1).

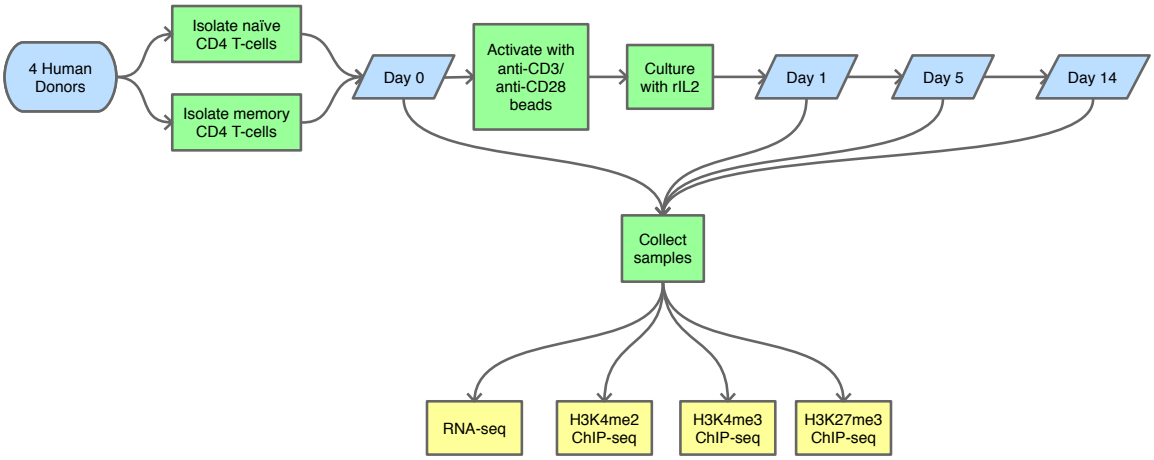


Figure 2.1: **Overview of the experimental design.**

2.3.1 RNA-seq differential expression analysis

Sequence reads were retrieved from the Sequence Read Archive (SRA) [50]. Five different alignment and quantification methods were tested for the RNA-seq data [51, 52, 53, 54, 55, 56, 57]. Each quantification was tested with both Ensembl transcripts and GENCODE known gene annotations [58, 59]. Comparisons of downstream results from each combination of quantification method and reference revealed that all quantifications gave broadly similar results for most genes, with non being obviously superior. Salmon quantification with regularization by shoal with the Ensembl annotation was chosen as the method theoretically most likely to partially mitigate some of the batch effect in the data [55, 56].

Due to an error in sample preparation, the RNA from the samples for days 0 and 5 were sequenced using a different kit than those for days 1 and 14. This induced a substantial batch effect in the data due to differences in sequencing biases between the two kits, and this batch effect is unfortunately confounded with the time point variable (Figure 2.2a). To do the best possible analysis with this data, this batch effect was subtracted out from the data using ComBat [42], ignoring the time point variable due to the confounding with the batch variable. The result is a marked improvement, but the unavoidable confounding with time point means that certain

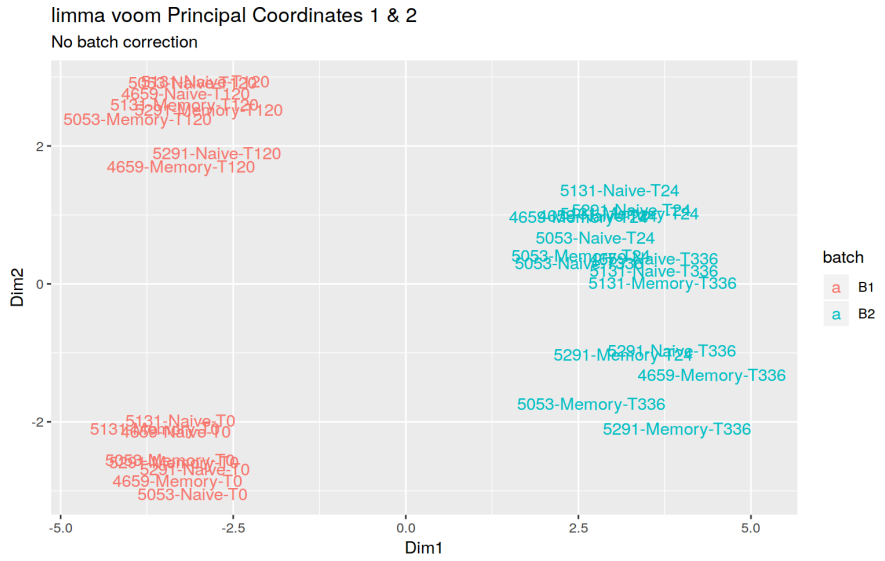
real patterns of gene expression will be indistinguishable from the batch effect and subtracted out as a result. Specifically, any "zig-zag" pattern, such as a gene whose expression goes up on day 1, down on day 5, and back up again on day 14, will be attenuated or eliminated entirely. In the context of a T-cell activation time course, it is unlikely that many genes of interest will follow such an expression pattern, so this loss was deemed an acceptable cost for correcting the batch effect.

However, removing the systematic component of the batch effect still leaves the noise component. The gene quantifications from the first batch are substantially noisier than those in the second batch. This analysis corrected for this by using `limma`'s sample weighting method to assign lower weights to the noisy samples of batch 1 (Figure 2.3) [24, 25]. The resulting analysis gives an accurate assessment of statistical significance for all comparisons, which unfortunately means a loss of statistical power for comparisons involving samples in batch 1.

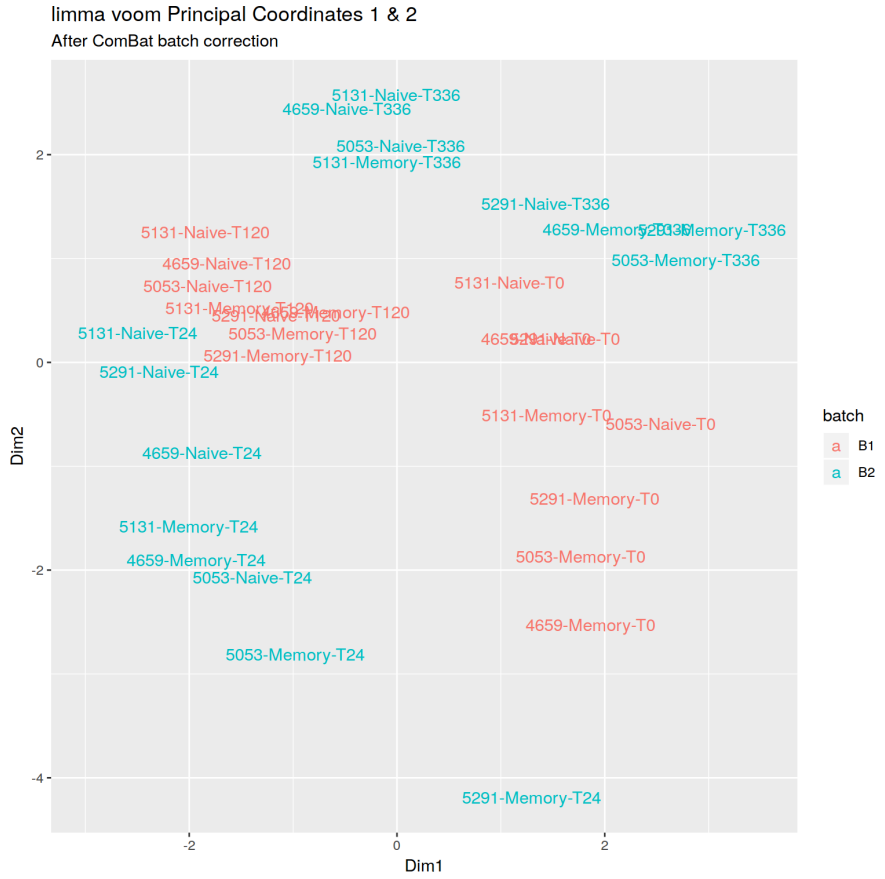
In any case, the RNA-seq counts were first normalized using trimmed mean of M-values (TMM) [40], converted to normalized $\log_2$ counts per million (logCPM) with quality weights using `voomWithQualityWeights` [23, 25], and batch-corrected at this point using ComBat. A linear model was fit to the batch-corrected, quality-weighted data for each gene using `limma`, and each gene was tested for differential expression using `limma`'s empirical Bayes moderated t-test [60, 23, 61]. P-values were corrected for multiple testing using the Benjamini-Hochberg (BH) procedure for false discovery rate (FDR) control [44].

2.3.2 ChIP-seq analyses

Sequence reads were retrieved from SRA [50]. ChIP-seq (and input) reads were aligned to the Genome Reference Consortium Human Build 38 (GRCh38) genome assembly using Bowtie 2 [62, 63, 57]. Artifact regions were annotated using a custom implementation of the `GreyListChIP` algorithm, and these "greylists" were merged with



(a) Before batch correction



(b) After batch correction with ComBat

Figure 2.2: **PCoA plots of RNA-seq data showing effect of batch correction.** The uncorrected data (a) shows a clear separation between samples from the two batches (red and blue) dominating the first principal coordinate. After correction with ComBat (b), the two batches now have approximately the same center, and the first two principal coordinates both show separation between experimental conditions rather than batches. (Note that time points are shown in hours rather than days in these plots.)

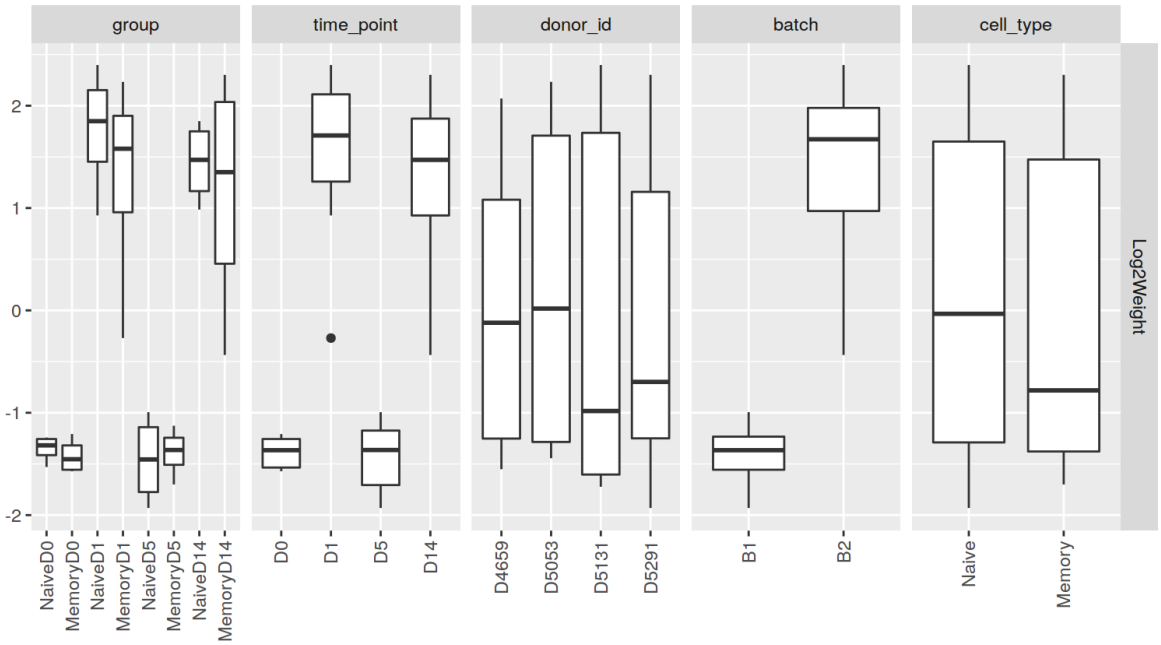
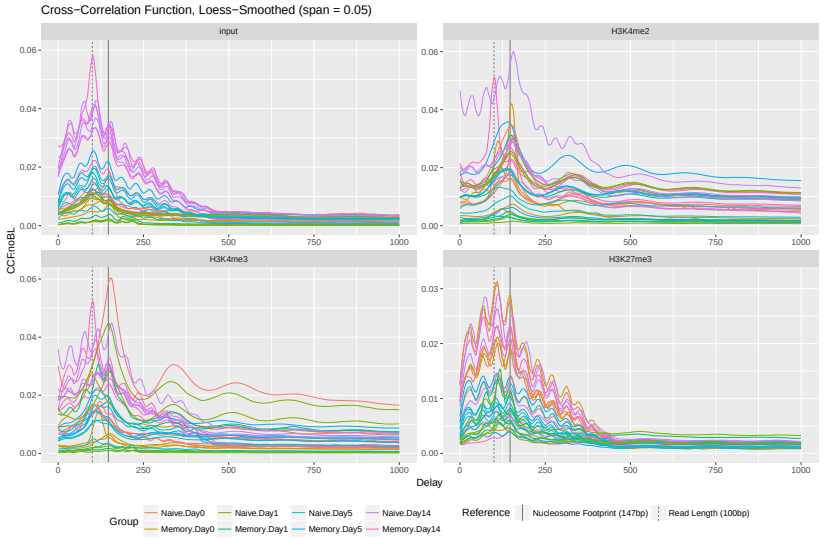


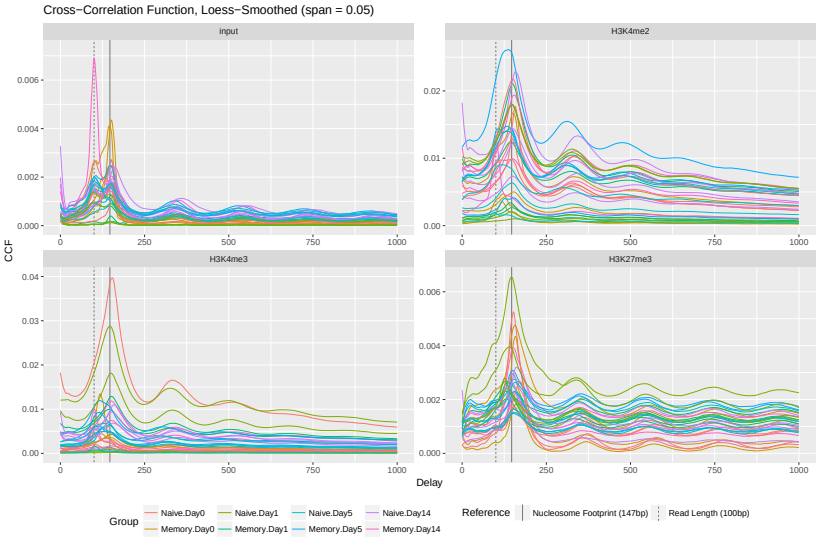
Figure 2.3: **RNA-seq sample weights, grouped by experimental and technical covariates.** Inverse variance weights were estimated for each sample using `limma`'s `arrayWeights` function (part of `voomWithQualityWeights`). The samples were grouped by each known covariate and the distribution of weights was plotted for each group.

the published Encyclopedia Of DNA Elements (ENCODE) blacklists [**greylistchip**, 64, 65, 46]. Any read or called peak overlapping one of these regions was regarded as artifactual and excluded from downstream analyses. Figure 2.4 shows the improvement after blacklisting in the strand cross-correlation plots, a common quality control plot for ChIP-seq data [66, 34]. Peaks were called using `epic`, an implementation of the Spatial Clustering for Identification of ChIP-Enriched Regions (SICER) algorithm [32, 67]. Peaks were also called separately using Model-based Analysis of ChIP-seq (MACS), but MACS was determined to be a poor fit for the data, and these peak calls are not used in any further analyses [31]. Consensus peaks were determined by applying the irreproducible discovery rate (IDR) framework [33, 68] to find peaks consistently called in the same locations across all 4 donors.

Promoters were defined by computing the distance from each annotated TSS to the nearest called peak and examining the distribution of distances, observing that



(a) **Cross-correlation plots without removing blacklisted reads.** Without blacklisting, many artifactual peaks are visible in the cross-correlations of the ChIP-seq samples, and the peak at the true fragment size (147 bp) is frequently overshadowed by the artifactual peak at the read length (100 bp).



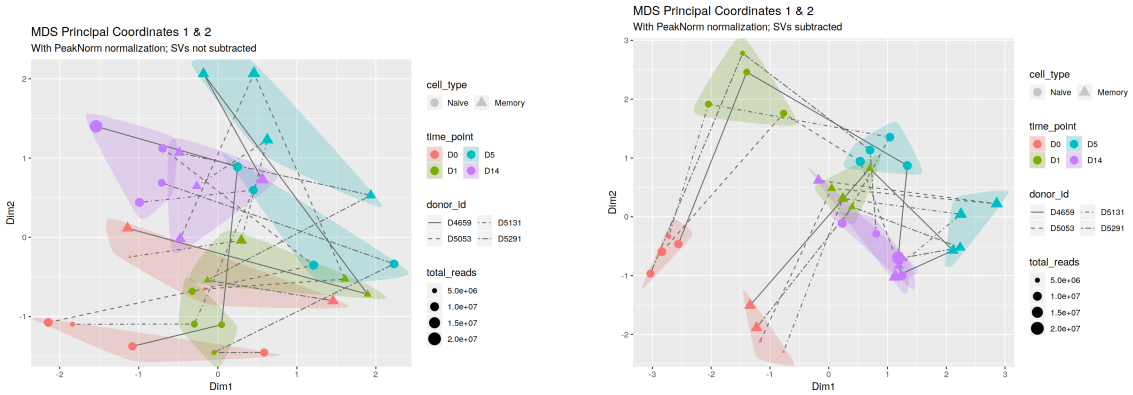
(b) **Cross-correlation plots with blacklisted reads removed.** After blacklisting, most ChIP-seq samples have clean-looking periodic cross-correlation plots, with the largest peak around 147 bp, the expected size for a fragment of DNA from a single nucleosome, and little to no peak at the read length, 100 bp.

Figure 2.4: **Strand cross-correlation plots for ChIP-seq data, before and after blacklisting.** The number of reads starting at each position in the genome was counted separately for the plus and minus strands, and then the correlation coefficient between the read start counts for both strands (cross-correlation) was computed after shifting the plus strand counts forward by a specified interval (the delay). This was repeated for every delay value from 0 to 1000, and the cross-correlation values were plotted as a function of the delay. In good quality samples, cross-correlation is maximized when the delay equals the fragment size; in poor quality samples, cross-correlation is often maximized when the delay equals the read length, an artifactual peak whose cause is not fully understood.

peaks for each histone mark were enriched within a certain distance of the TSS. (Note: this analysis was performed using the original peak calls and expression values from GEO [48].) For H3K4me2 and H3K4me3, this distance was about 1 kbp, while for H3K27me3 it was 2.5 kbp. These distances were used as an "effective promoter radius" for each mark. The promoter region for each gene was defined as the region of the genome within this distance upstream or downstream of the gene's annotated TSS. For genes with multiple annotated TSSs, a promoter region was defined for each TSS individually, and any promoters that overlapped (due to multiple TSSs being closer than 2 times the radius) were merged into one large promoter. Thus, some genes had multiple promoters defined, which were each analyzed separately for differential modification.

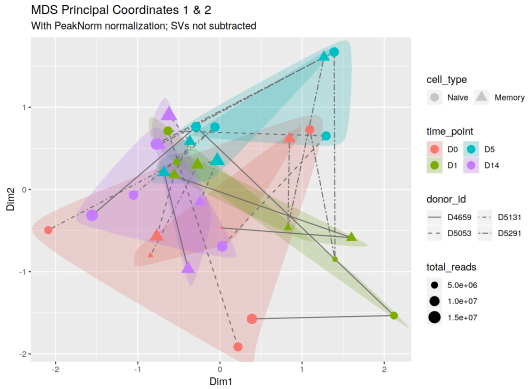
Reads in promoters, peaks, and sliding windows across the genome were counted and normalized using `csaw` and analyzed for differential modification using `edgeR` [69, 34, 30, 61]. Unobserved confounding factors in the ChIP-seq data were corrected using surrogate variable analysis (SVA) [43, 70]. Principal coordinate plots of the promoter count data for each histone mark before and after subtracting surrogate variable effects are shown in Figure 2.5.

To investigate whether the location of a peak within the promoter region was important, "relative coverage profiles" were generated. First, 500-bp sliding windows were tiled around each annotated TSS: one window centered on the TSS itself, and 10 windows each upstream and downstream, thus covering a 10.5-kb region centered on the TSS with a total of 21 windows. Reads in each window for each TSS were counted in each sample, and the counts were normalized and converted to logCPM as in the differential modification analysis. An abundance threshold was chosen such that 99% of peak-containing promoters have an average logCPM above this threshold (Figure 2.6). Then *all* promoters with an average logCPM above this threshold were included, and all below that thereshold were filtered out, regardless of whether

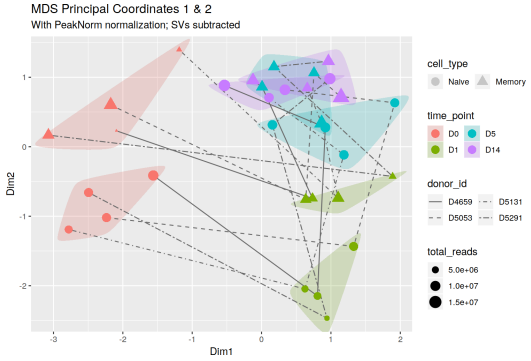


(a) **H3K4me2, no correction**

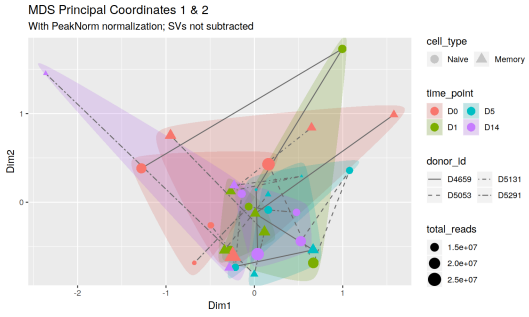
(b) **H3K4me2, SVs subtracted**



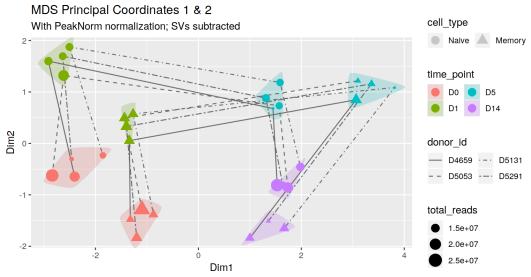
(c) **H3K4me3, no correction**



(d) **H3K4me3, SVs subtracted**



(e) **H3K27me3, no correction**



(f) **H3K27me3, SVs subtracted**

Figure 2.5: **PCoA plots of ChIP-seq sliding window data, before and after subtracting surrogate variables (SVs).** For each histone mark, a PCoA plot of the first 2 principal coordinates was created before and after subtraction of SV effects. Time points are shown by color and cell type by shape, and samples from the same time point and cell type are enclosed in a shaded area to aid in visial recognition (this shaded area has no meaning on the plot). Samples of the same cell type from the same donor are connected with a line in time point order, showing the "trajectory" of each donor's samples over time.

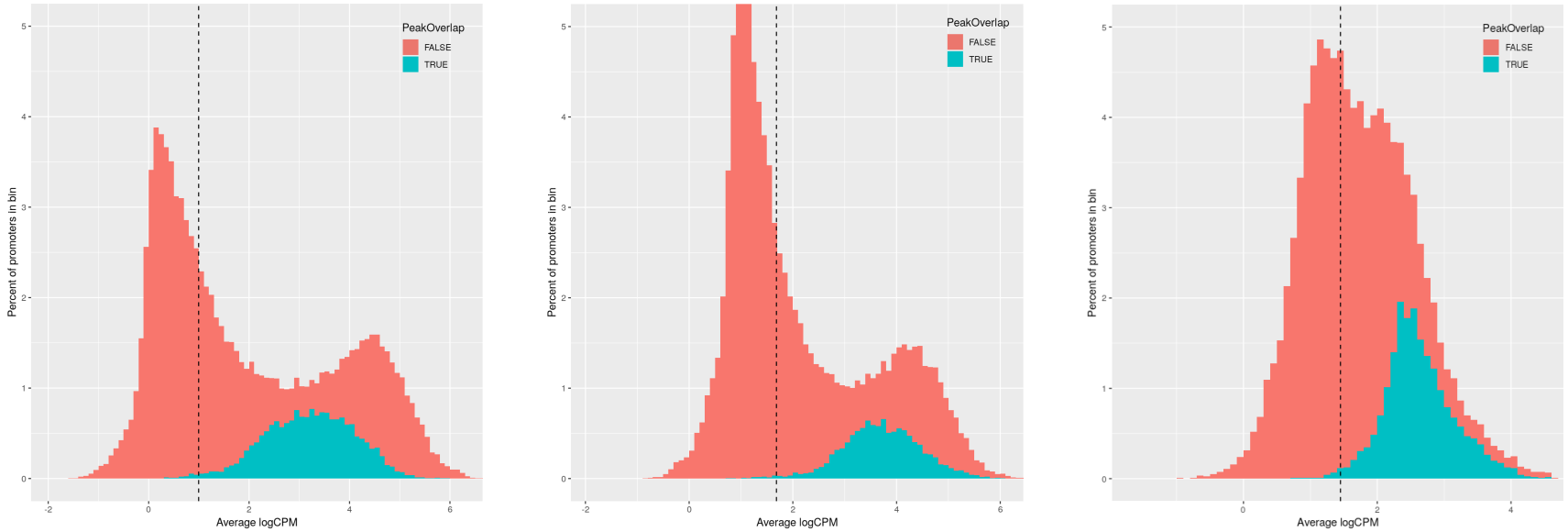
they actually contained a called peak. This ensures that even promoters containing undetected peaks will be included, at the cost of likely including many promoters that do not contain any true peak. Then, the logCPM values of the bins within each promoter were normalized to an average of zero, such that each window's normalized abundance now represents the relative read depth of that window compared to all other windows in the same promoter. The normalized abundance values for each window in a promoter are collectively referred to as that promoter's "relative coverage profile".

2.3.3 MOFA analysis of cross-dataset variation patterns

Multi-Omics Factor Analysis (MOFA) was run on all the ChIP-seq windows overlapping consensus peaks for each histone mark, as well as the RNA-seq data, in order to identify patterns of coordinated variation across all data sets [71]. The results are summarized in Figure 2.7. Latent factors (LFs) 1, 4, and 5 were determined to explain the most variation consistently across all data sets (Figure 2.7a), and scatter plots of these factors show that they also correlate best with the experimental factors (Figure 2.7b). LF2 captures the batch effect in the RNA-seq data. Removing the effect of LF2 using MOFA theoretically yields a batch correction that does not depend on knowing the experimental factors. When this was attempted, the resulting batch correction was comparable to ComBat (see Figure 2.2b), indicating that the ComBat-based batch correction has little room for improvement given the problems with the data set.

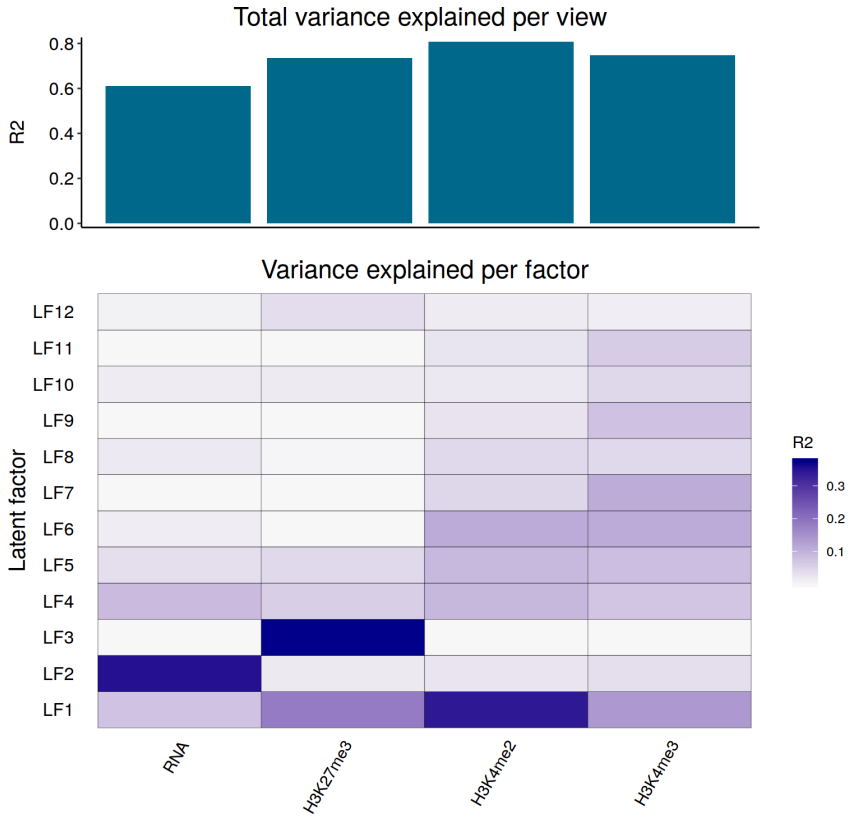
Test	Est. non-null	FDR $\leq 10\%$
Naïve Day 0 vs Day 1	5992	1613
Naïve Day 0 vs Day 5	3038	32
Naïve Day 0 vs Day 14	1870	190
Memory Day 0 vs Day 1	3195	411
Memory Day 0 vs Day 5	2688	18
Memory Day 0 vs Day 14	1911	227
Day 0 Naïve vs Memory	0	2
Day 1 Naïve vs Memory	9167	5532
Day 5 Naïve vs Memory	0	0
Day 14 Naïve vs Memory	6446	2319

Table 2.1: **Estimated and detected differentially expressed genes.** "Test": Which sample groups were compared; "Est non-null": Estimated number of differentially expressed genes, using the method of averaging local FDR values [45]; "FDR $\leq$ 10%": Number of significantly differentially expressed genes at an FDR threshold of 10%. The total number of genes tested was 16707.

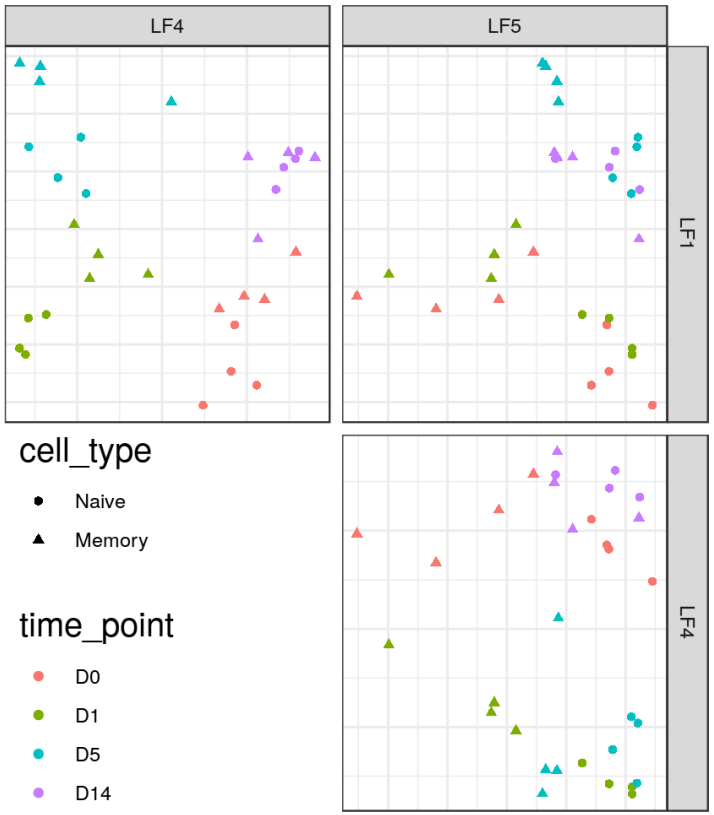


(a) H3K4me2 (b) H3K4me3 (c) H3K27me3

Figure 2.6: **Promoter abundance filtering for relative coverage profiles.** For each histone mark, a histogram of promoter logCPM values was plotted, colored by whether each promoter contains a called peak. The abundance filter for each histone mark (dotted vertical line) was set such that 99% of peak-containing promoters (blue) are above the threshold, and then all promoters above this threshold were included in downstream analyses.



(a) **Variance explained in each data set by each latent factor estimated by MOFA.** For each LF learned by MOFA, the variance explained by that factor in each data set ("view") is shown by the shading of the cells in the lower section. The upper section shows the total fraction of each data set's variance that is explained by all LFs combined.



(b) **Scatter plots of specific pairs of MOFA latent factors.** LFs 1, 4, and 5 explain substantial variation in all data sets, so they were plotted against each other in order to reveal patterns of variation that are shared across all data sets. These plots can be interpreted similarly to PCA and PCoA plots.

Figure 2.7: **MOFA latent factors identify shared patterns of variation.** MOFA was used to estimate latent factors (LFs) that explain substantial variation in the RNA-seq data and the ChIP-seq data (a). Then specific LFs of interest were selected and plotted (b).

2.4 Results

2.4.1 Interpretation of RNA-seq analysis is limited by a major confounding factor

Genes called as present in the RNA-seq data were tested for differential expression between all time points and cell types. The counts of differentially expressed genes are shown in Table 2.1. Notably, all the results for Day 0 and Day 5 have substantially fewer genes called differentially expressed than any of the results for other time points. This is an unfortunate result of the difference in sample quality between the two batches of RNA-seq data. All the samples in Batch 1, which includes all the samples from Days 0 and 5, have substantially more variability than the samples in Batch 2, which includes the other time points. This is reflected in the substantially higher weights assigned to Batch 2 (Figure 2.3). The batch effect has both a systematic component and a random noise component. While the systematic component was subtracted out using ComBat (Figure 2.2), no such correction is possible for the noise component: Batch 1 simply has substantially more random noise in it, which reduces the statistical power for any differential expression tests involving samples in that batch.

Despite the difficulty in detecting specific differentially expressed genes, there is still evidence that differential expression is present for these time points. In Figure 2.8, there is a clear separation between naïve and memory samples at Day 0, despite the fact that only 2 genes were significantly differentially expressed for this comparison. Similarly, the small numbers of genes detected for the Day 0 vs Day 5 comparisons do not reflect the large separation between these time points in Figure 2.8. In addition, the MOFA LF plots in Figure 2.7b. This suggests that there is indeed a differential expression signal present in the data for these comparisons, but the large variability in the Batch 1 samples obfuscates this signal at the individual gene level. As a result, it is

Histone Mark	# Peaks	Mean peak width	genome coverage	FRiP
H3K4me2	14,965	3,970	1.92%	14.2%
H3K4me3	6,163	2,946	0.588%	6.57%
H3K27me3	18,139	18,967	11.1%	22.5%

Table 2.2: **Summary of peak-calling statistics.** For each histone mark, the number of peaks called using SICER at an IDR threshold of 0.05, the mean width of those peaks, the fraction of the genome covered by peaks, and the fraction of reads in peaks (FRiP).

impossible to make any meaningful statements about the "size" of the gene signature for any time point, since the number of significant genes as well as the estimated number of differentially expressed genes depends so strongly on the variations in sample quality in addition to the size of the differential expression signal in the data. Gene-set enrichment analyses are similarly impractical. However, analyses looking at genome-wide patterns of expression are still practical.

2.4.2 H3K4 and H3K27 methylation occur in broad regions and are enriched near promoters

Table 2.2 gives a summary of the peak calling statistics for each histone mark. Consistent with previous observations, all 3 histone marks occur in broad regions spanning many consecutive nucleosomes, rather than in sharp peaks as would be expected for a transcription factor or other molecule that binds to specific sites. This conclusion is further supported by Figure 2.4b, in which a clear nucleosome-sized periodicity is visible in the cross-correlation value for each sample, indicating that each time a given mark is present on one histone, it is also likely to be found on adjacent histones as well. H3K27me3 enrichment in particular is substantially more broad than either H3K4 mark, with a mean peak width of almost 19,000 bp. This is also reflected in the periodicity observed in Figure 2.4b, which remains strong much farther out for H3K27me3 than the other marks, showing H3K27me3 especially tends to be found

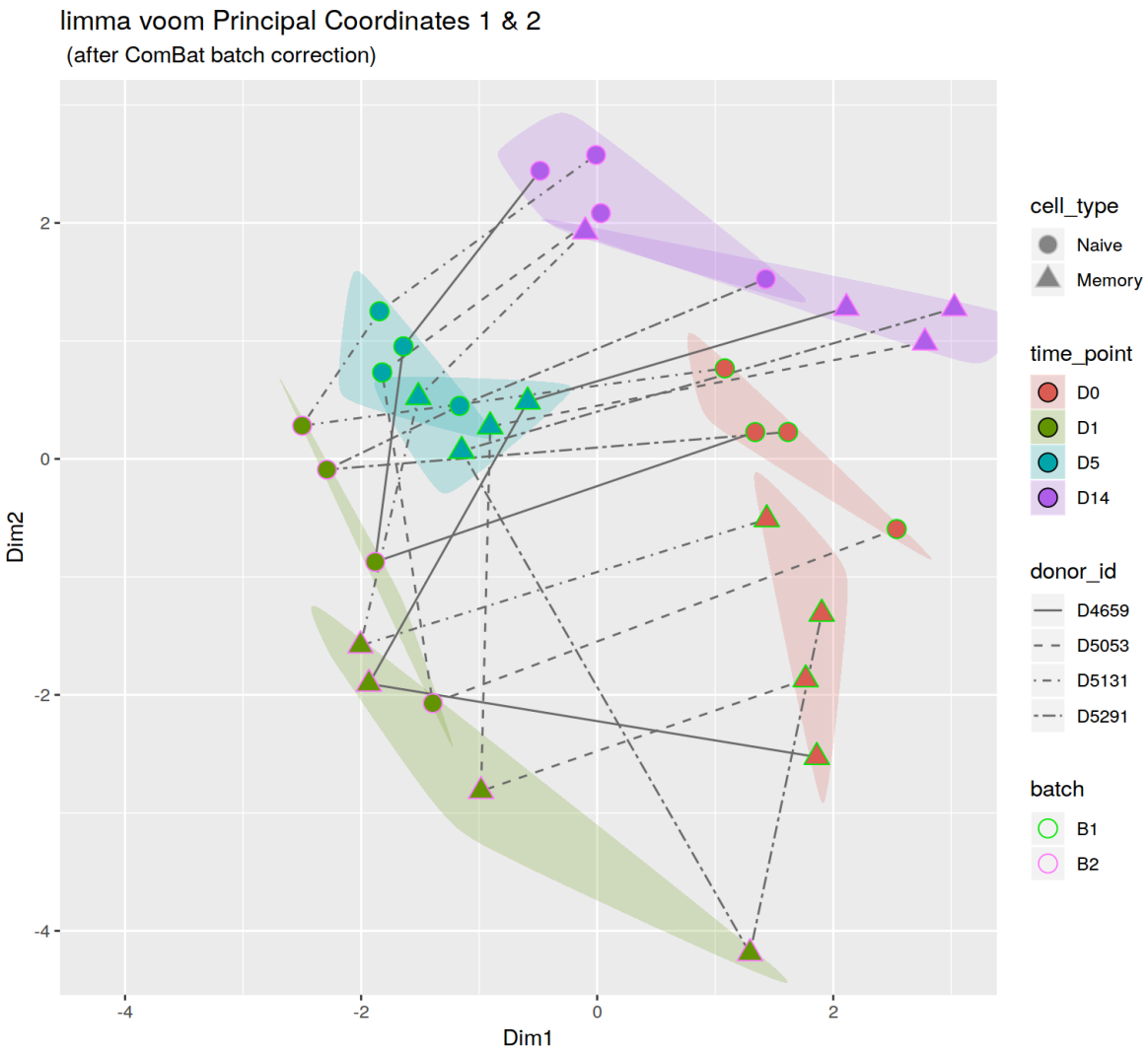


Figure 2.8: **PCoA plot of RNA-seq samples after ComBat batch correction.** Each point represents an individual sample. Samples with the same combination of cell type and time point are encircled with a shaded region to aid in visual identification of the sample groups. Samples of the same cell type from the same donor are connected by lines to indicate the "trajectory" of each donor's cells over time in PCoA space.

Histone mark	Effective promoter radius
H3K4me2	1 kbp
H3K4me3	1 kbp
H3K27me3	2.5 kbp

Table 2.3: **Effective promoter radius for each histone mark.** These values represent the approximate distance from transcription start site positions within which an excess of peaks are found, as shown in Figure 2.9.

on long runs of consecutive histones.

All 3 histone marks tend to occur more often near promoter regions, as shown in Figure 2.9. The majority of each density distribution is flat, representing the background density of peaks genome-wide. Each distribution has a peak near zero, representing an enrichment of peaks close to TSS positions relative to the remainder of the genome. Interestingly, the "radius" within which this enrichment occurs is not the same for every histone mark (Table 2.3). For H3K4me2 and H3K4me3, peaks are most enriched within 1 kbp of TSS positions, while for H3K27me3, enrichment is broader, extending to 2.5 kbp. These "effective promoter radii" remain approximately the same across all combinations of experimental condition (cell type, time point, and donor), so they appear to be a property of the histone mark itself. Hence, these radii were used to define the promoter regions for each histone mark in all further analyses.

2.4.3 Correlations between gene expression and promoter methylation follow expected genome-wide trends

H3K4me2 and H3K4me2 have previously been reported as activating marks whose presence in a gene's promoter is associated with higher gene expression, while H3K27me3 has been reported as inactivating [48, 49]. The data are consistent with this characterization: genes whose promoters (as defined by the radii for each histone mark listed in 2.3) overlap with a H3K4me2 or H3K4me3 peak tend to have higher expression than those that don't, while H3K27me3 is likewise associated with lower gene

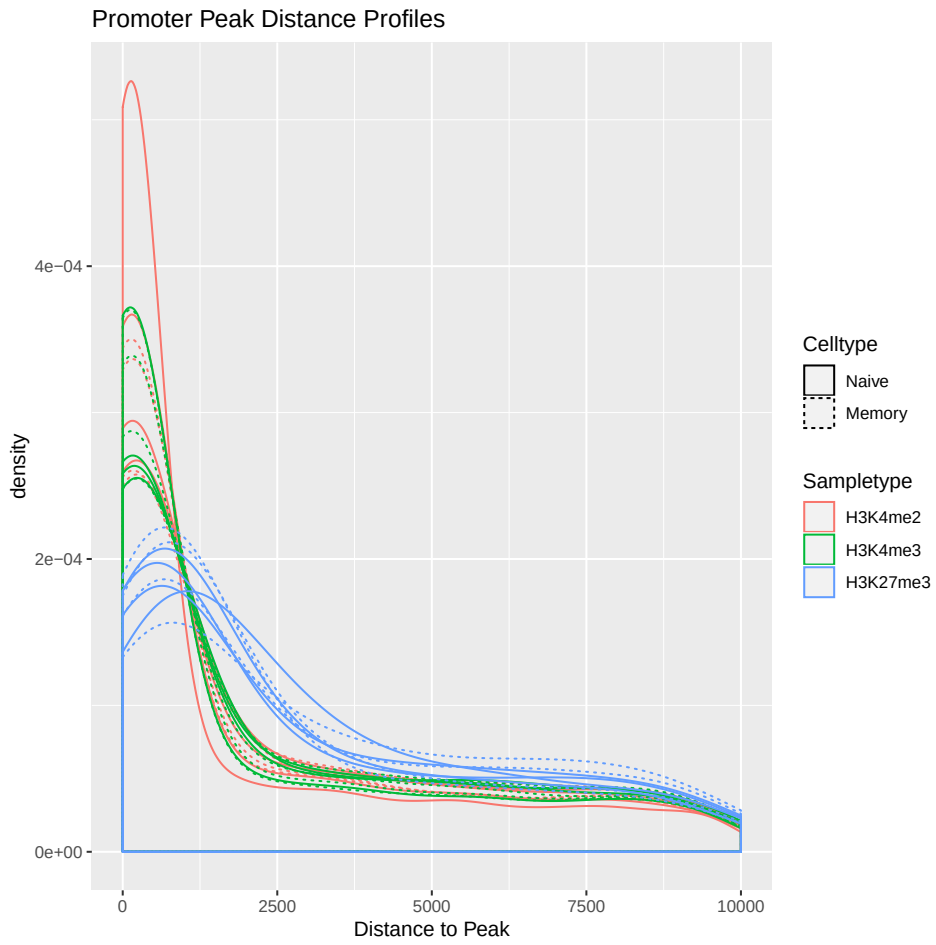


Figure 2.9: **Enrichment of peaks in promoter neighborhoods.** This plot shows the distribution of distances from each annotated transcription start site in the genome to the nearest called peak. Each line represents one combination of histone mark, cell type, and time point. Distributions are smoothed using kernel density estimation. TSSs that occur *within* peaks were excluded from this plot to avoid a large spike at zero that would overshadow the rest of the distribution. (Note: this figure was generated using the original peak calls and expression values from GEO [48].)

expression, as shown in 2.10. This pattern holds across all combinations of cell type and time point (Welch's t-test, all p-values $\ll 2.2 \times 10^{-16}$). The difference in average $\log_2$ fragments per kilobase per million fragments (FPKM) values when a peak overlaps the promoter is about +5.67 for H3K4me2, +5.76 for H3K4me2, and −4.00 for H3K27me3.

2.4.4 Gene expression and promoter histone methylation patterns show convergence between naïve and memory cells at day 14

We hypothesized that if naïve cells had differentiated into memory cells by Day 14, then their patterns of expression and histone modification should converge with those of memory cells at Day 14. Figure 2.11 shows the patterns of variation in all 3 histone marks in the promoter regions of the genome using principal coordinate analysis (PCoA). All 3 marks show a noticeable convergence between the naïve and memory samples at day 14, visible as an overlapping of the day 14 groups on each plot. This is consistent with the counts of significantly differentially modified promoters and estimates of the total numbers of differentially modified promoters shown in Table 2.4. For all histone marks, evidence of differential modification between naïve and memory samples was detected at every time point except day 14. The day 14 convergence pattern is also present in the RNA-seq data (Figure 2.11d), albeit in the 2nd and 3rd principal coordinates, indicating that it is not the most dominant pattern driving gene expression. Taken together, the data show that promoter histone methylation for these 3 histone marks and RNA expression for naïve and memory cells are most similar at day 14, the furthest time point after activation. MOFA was also able to capture this day 14 convergence pattern in LF5 (Figure 2.7b), which accounts for shared variation across all 3 histone marks and the RNA-seq data, confirming that

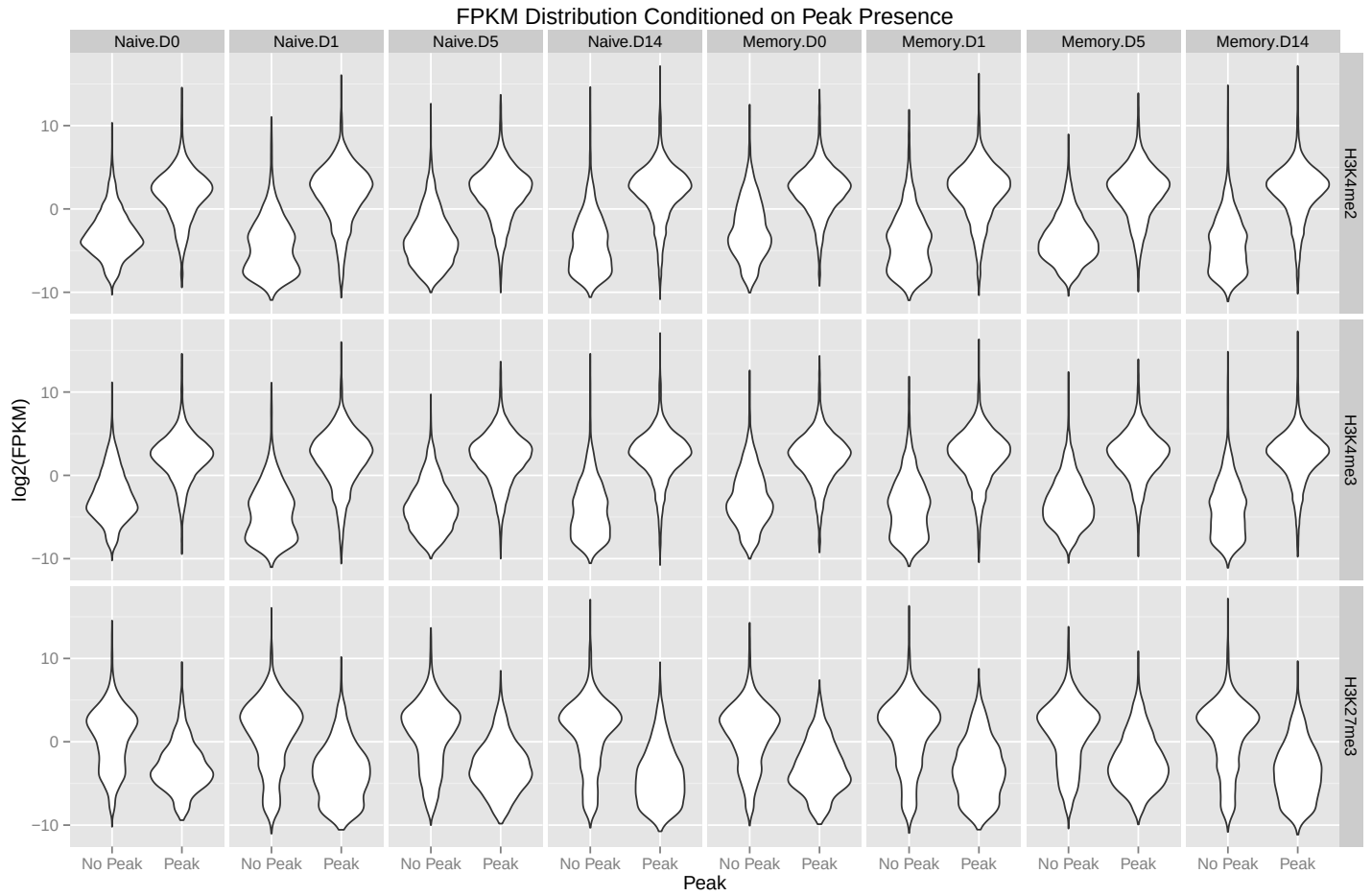


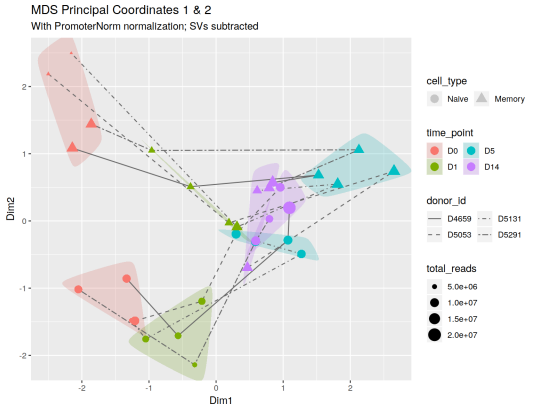
Figure 2.10: **Expression distributions of genes with and without promoter peaks.** For each histone mark in each experimental condition, the average RNA-seq abundance ($\log_2$ FPKM) of each gene across all 4 donors was calculated. Genes were grouped based on whether or not a peak was called in their promoters in that condition, and the distribution of abundance values was plotted for the no-peak and peak groups. (Note: this figure was generated using the original peak calls and expression values from GEO [48].)

this convergence is a coordinated pattern across all 4 data sets. While this observation does not prove that the naïve cells have differentiated into memory cells at Day 14, it is consistent with that hypothesis.

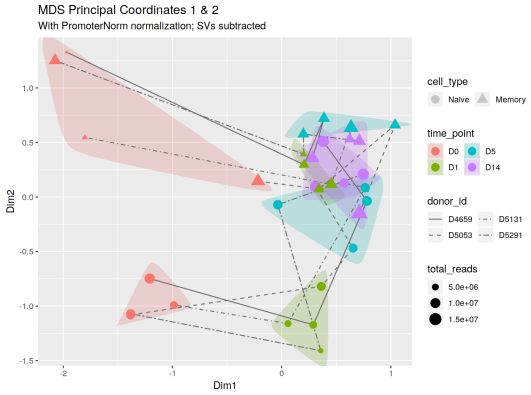
2.4.5 Location of H3K4me2 and H3K4me3 promoter coverage associates with gene expression

To test whether the position of a histone mark relative to a gene's TSS was important, we looked at the "landscape" of ChIP-seq read coverage in naïve Day 0 samples within 5 kbp of each gene's TSS by binning reads into 500-bp windows tiled across each promoter logCPM values were calculated for the bins in each promoter and then the average logCPM for each promoter's bins was normalized to zero, such that the values represent coverage relative to other regions of the same promoter rather than being proportional to absolute read count. The promoters were then clustered based on the normalized bin abundances using k-means clustering with $K = 6$. Different values of K were also tested, but did not substantially change the interpretation of the data.

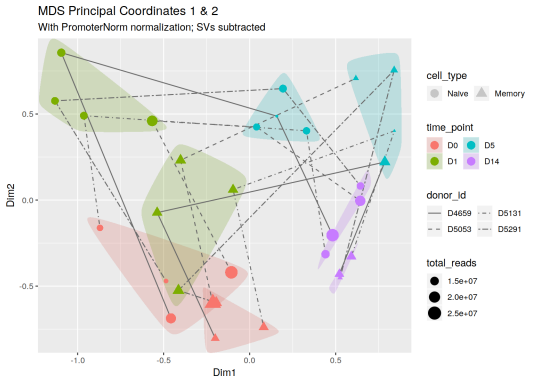
For H3K4me2, plotting the average bin abundances for each cluster reveals a simple pattern (Figure 2.12a): Cluster 5 represents a completely flat promoter coverage profile, likely consisting of genes with no H3K4me2 methylation in the promoter. All the other clusters represent a continuum of peak positions relative to the TSS. In order from most upstream to most downstream, they are Clusters 6, 4, 3, 1, and 2. There do not appear to be any clusters representing coverage patterns other than lone peaks, such as coverage troughs or double peaks. Next, all promoters were plotted in a principal component analysis (PCA) plot based on the same relative bin abundance data, and colored based on cluster membership (Figure 2.12b). The PCA plot shows Cluster 5 (the "no peak" cluster) at the center, with the other clusters arranged in a counter-clockwise arc around it in the order noted above, from most upstream peak to most downstream. Notably, the "clusters" form a single large "cloud" with no ap-



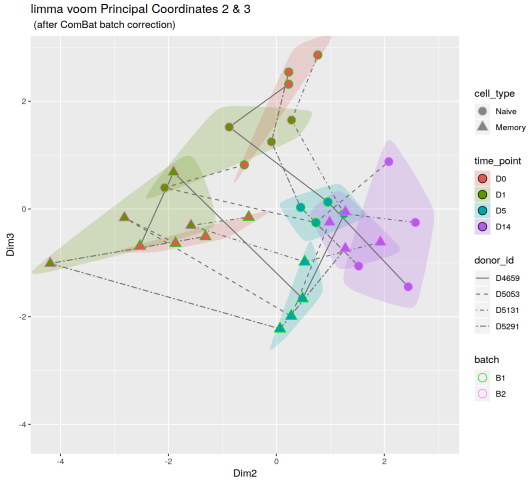
(a) PCoA plot of H3K4me2 promoters, after subtracting surrogate variables.



(b) PCoA plot of H3K4me3 promoters, after subtracting surrogate variables.



(c) PCoA plot of H3K27me3 promoters, after subtracting surrogate variables.



(d) RNA-seq PCoA, after ComBat batch correction, showing principal coordinates 2 and 3.

Figure 2.11: **PCoA plots for promoter ChIP-seq and expression RNA-seq data.** Each point represents an individual sample. Samples with the same combination of cell type and time point are encircled with a shaded region to aid in visual identification of the sample groups. Samples of the same cell type from the same donor are connected by lines to indicate the "trajectory" of each donor's cells over time in PCoA space.

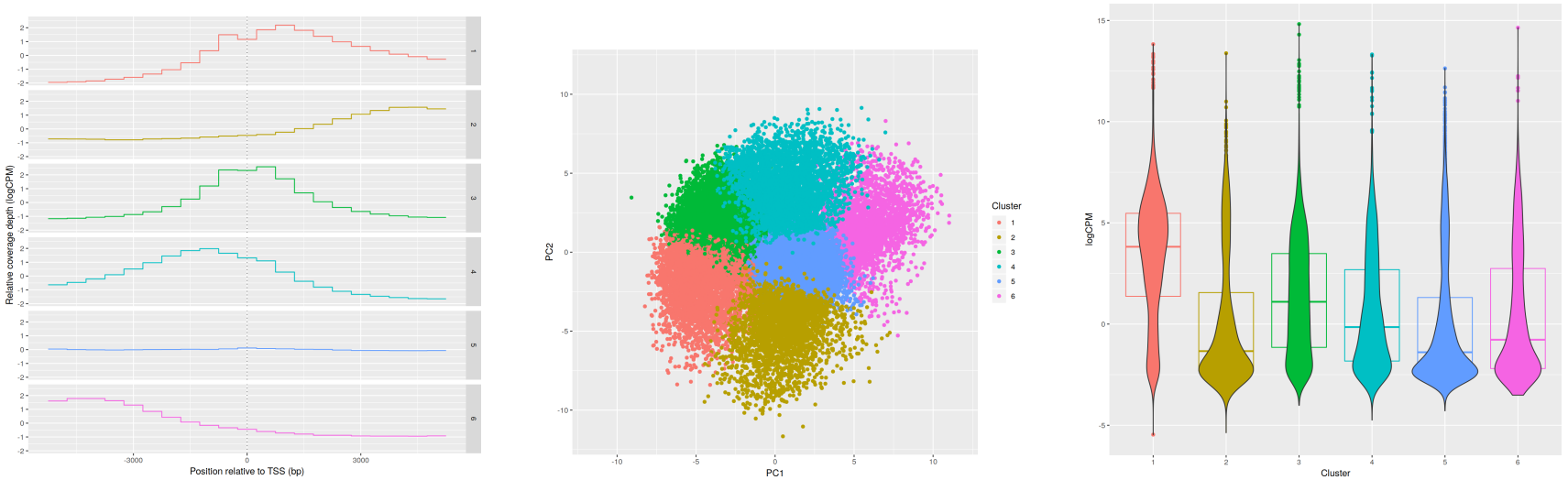
	Number of significant promoters			Est. differentially modified promoters		
Time Point	H3K4me2	H3K4me3	H3K27me3	H3K4me2	H3K4me3	H3K27me3
Day 0	4553	927	6	9967	4149	2404
Day 1	567	278	1570	4370	2145	6598
Day 5	2313	139	490	9450	1148	4141
Day 14	0	0	0	0	0	0

Table 2.4: **Number of differentially modified promoters between naïve and memory cells at each time point after activation.** This table shows both the number of differentially modified promoters detected at a 10% FDR threshold (left half), and the total number of differentially modified promoters estimated using the method of averaging local FDR estimates [61] (right half).

parent separation between them, further supporting the conclusion that these clusters represent an arbitrary partitioning of a continuous distribution of promoter coverage landscapes. While the clusters are a useful abstraction that aids in visualization, they are ultimately not an accurate representation of the data. The continuous nature of the distribution also explains why different values of K led to similar conclusions.

To investigate the association between relative peak position and gene expression, we plotted the Naïve Day 0 expression for the genes in each cluster (Figure 2.12c). Most genes in Cluster 5, the "no peak" cluster, have low expression values. Taking this as the "baseline" distribution when no H3K4me2 methylation is present, we can compare the other clusters' distributions to determine which peak positions are associated with elevated expression. As might be expected, the 3 clusters representing peaks closest to the TSS, Clusters 1, 3, and 4, show the highest average expression distributions. Specifically, these clusters all have their highest ChIP-seq abundance within 1kb of the TSS, consistent with the previously determined promoter radius. In contrast, cluster 6, which represents peaks several kbp upstream of the TSS, shows a slightly higher average expression than baseline, while Cluster 2, which represents peaks several kbp downstream, doesn't appear to show any appreciable difference. Interestingly, the cluster with the highest average expression is Cluster 1, which represents peaks about 1 kbp downstream of the TSS, rather than Cluster 3, which represents peaks centered directly at the TSS. This suggests that conceptualizing the promoter as a region centered on the TSS with a certain "radius" may be an oversimplification – a peak that is a specific distance from the TSS may have a different degree of influence depending on whether it is upstream or downstream of the TSS.

All observations described above for H3K4me2 ChIP-seq also appear to hold for H3K4me3 as well (Figure 2.13). This is expected, since there is a high correlation between the positions where both histone marks occur.

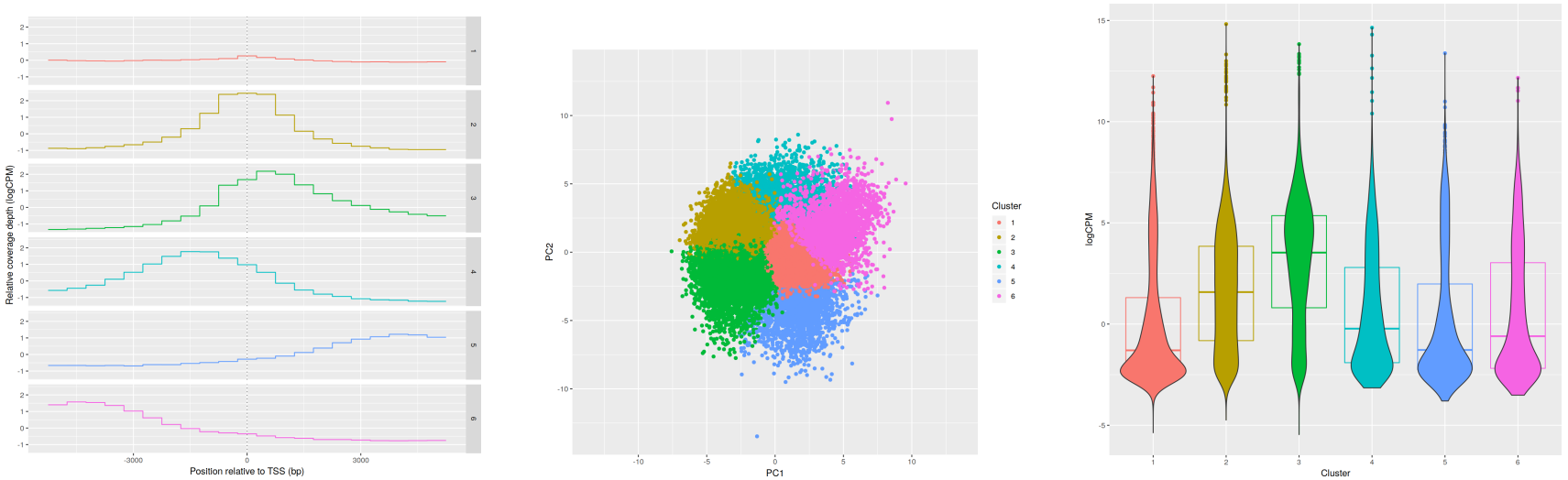


(a) **Average relative coverage for each bin in each cluster.**

(b) PCA of relative coverage depth, colored by K-means cluster membership.

(c) Gene expression grouped by promoter coverage clusters.

Figure 2.12: **K-means clustering of promoter H3K4me2 relative coverage depth in naïve day 0 samples.** H3K4me2 ChIP-seq reads were binned into 500-bp windows tiled across each promoter from 5 kbp upstream to 5 kbp downstream, and the logCPM values were normalized within each promoter to an average of 0, yielding relative coverage depths. These were then grouped using K-means clustering with $K = 6$, and the average bin values were plotted for each cluster (a). The x-axis is the genomic coordinate of each bin relative to the the transcription start site, and the y-axis is the mean relative coverage depth of that bin across all promoters in the cluster. Each line represents the average "shape" of the promoter coverage for promoters in that cluster. PCA was performed on the same data, and the first two PCs were plotted, coloring each point by its K-means cluster identity (b). For each cluster, the distribution of gene expression values was plotted (c).



(a) Average relative coverage for each bin in each cluster.

(b) PCA of relative coverage depth, colored by K-means cluster membership.

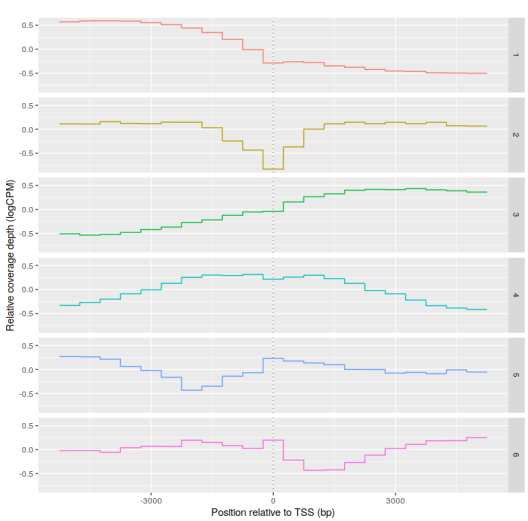
(c) Gene expression grouped by promoter coverage clusters.

Figure 2.13: **K-means clustering of promoter H3K4me3 relative coverage depth in naïve day 0 samples.** H3K4me3 ChIP-seq reads were binned into 500-bp windows tiled across each promoter from 5 kbp upstream to 5 kbp downstream, and the logCPM values were normalized within each promoter to an average of 0, yielding relative coverage depths. These were then grouped using K-means clustering with $K = 6$, and the average bin values were plotted for each cluster (a). The x-axis is the genomic coordinate of each bin relative to the the transcription start site, and the y-axis is the mean relative coverage depth of that bin across all promoters in the cluster. Each line represents the average "shape" of the promoter coverage for promoters in that cluster. PCA was performed on the same data, and the first two PCs were plotted, coloring each point by its K-means cluster identity (b). For each cluster, the distribution of gene expression values was plotted (c).

2.4.6 Patterns of H3K27me3 promoter coverage associate with gene expression

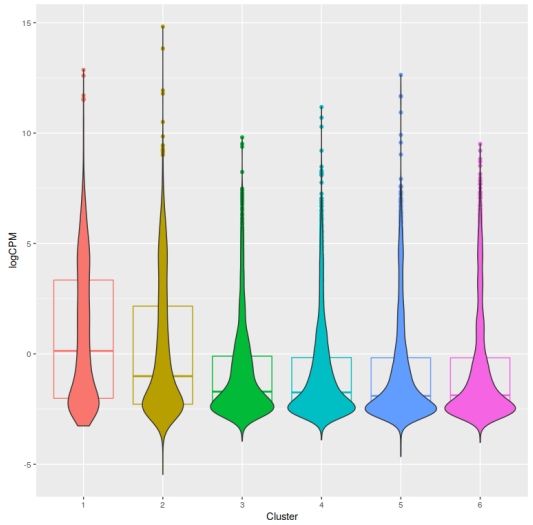
Unlike both H3K4 marks, whose main patterns of variation appear directly related to the size and position of a single peak within the promoter, the patterns of H3K27me3 methylation in promoters are more complex (Figure 2.14). Once again looking at the relative coverage in a 500-bp wide bins in a 5kb radius around each TSS, promoters were clustered based on the normalized relative coverage values in each bin using k-means clustering with $K = 6$ (Figure 2.14a). This time, 3 "axes" of variation can be observed, each represented by 2 clusters with opposing patterns. The first axis is greater upstream coverage (Cluster 1) vs. greater downstream coverage (Cluster 3); the second axis is the coverage at the TSS itself: peak (Cluster 4) or trough (Cluster 2); lastly, the third axis represents a trough upstream of the TSS (Cluster 5) vs. downstream of the TSS (Cluster 6). Referring to these opposing pairs of clusters as axes of variation is justified, because they correspond precisely to the first 3 principal components (PCs) in the PCA plot of the relative coverage values (Figure 2.14b). The PCA plot reveals that as in the case of H3K4me2, all the "clusters" are really just sections of a single connected cloud rather than discrete clusters. The cloud is approximately ellipsoid-shaped, with each PC being an axis of the ellipse, and each cluster consisting of a pyramidal section of the ellipsoid.

In Figure 2.14c, we can see that Clusters 1 and 2 are the only clusters with higher gene expression than the others. For Cluster 2, this is expected, since this cluster represents genes with depletion of H3K27me3 near the promoter. Hence, elevated expression in cluster 2 is consistent with the conventional view of H3K27me3 as a deactivating mark. However, Cluster 1, the cluster with the most elevated gene expression, represents genes with elevated coverage upstream of the TSS, or equivalently, decreased coverage downstream, inside the gene body. The opposite pattern, in which H3K27me3 is more abundant within the gene body and less abundance in



(a) Average relative coverage for each bin in each cluster.

(b) PCA of relative coverage depth, colored by K-means cluster membership.



(c) Gene expression grouped by promoter coverage clusters.

Figure 2.14: **K-means clustering of promoter H3K27me3 relative coverage depth in naïve day 0 samples.** H3K27me3 ChIP-seq reads were binned into 500-bp windows tiled across each promoter from 5 kbp upstream to 5 kbp downstream, and the logCPM values were normalized within each promoter to an average of 0, yielding relative coverage depths. These were then grouped using k-means clustering with $K = 6$, and the average bin values were plotted for each cluster (a). The x-axis is the genomic coordinate of each bin relative to the the transcription start site, and the y-axis is the mean relative coverage depth of that bin across all promoters in the cluster. Each line represents the average "shape" of the promoter coverage for promoters in that cluster. PCA was performed on the same data, and the first two PCs were plotted, coloring each point by its K-means cluster identity (b). (Note: In (b), Cluster 6 is hidden behind all the other clusters.) For each cluster, the distribution of gene expression values was plotted (c).

the upstream promoter region, does not show any elevation in gene expression. As with H3K4me2, this shows that the location of H3K27 trimethylation relative to the TSS is potentially an important factor beyond simple proximity.

2.5 Discussion

2.5.1 Each histone mark's "effective promoter extent" must be determined empirically

Figure 2.9 shows that H3K4me2, H3K4me3, and H3K27me3 are all enriched near promoters, relative to the rest of the genome, consistent with their conventionally understood role in regulating gene transcription. Interestingly, the radius within this enrichment occurs is not the same for each histone mark. H3K4me2 and H3K4me3 are enriched within a 1 kbp radius, while H3K27me3 is enriched within 2.5 kbp. Notably, the determined promoter radius was consistent across all experimental conditions, varying only between different histone marks. This suggests that the conventional "one size fits all" approach of defining a single promoter region for each gene (or each TSS) and using that same promoter region for analyzing all types of genomic data within an experiment may not be appropriate, and a better approach may be to use a separate promoter radius for each kind of data, with each radius being derived from the data itself. Furthermore, the apparent asymmetry of upstream and downstream promoter histone modification with respect to gene expression, seen in Figures 2.12, 2.13, and 2.14, shows that even the concept of a promoter "radius" is likely an oversimplification. At a minimum, nearby enrichment of peaks should be evaluated separately for both upstream and downstream peaks, and an appropriate "radius" should be selected for each direction.

Figures 2.12 and 2.13 show that the determined promoter radius of 1 kbp is approximately consistent with the distance from the TSS at which enrichment of

H3K4 methylation correlates with increased expression, showing that this radius, which was determined by a simple analysis of measuring the distance from each TSS to the nearest peak, also has functional significance. For H3K27me3, the correlation between histone modification near the promoter and gene expression is more complex, involving non-peak variations such as troughs in coverage at the TSS and asymmetric coverage upstream and downstream, so it is difficult in this case to evaluate whether the 2.5 kbp radius determined from TSS-to-peak distances is functionally significant. However, the two patterns of coverage associated with elevated expression levels both have interesting features within this radius.

2.5.2 Day 14 convergence is consistent with naïve-to-memory differentiation

We observed that all 3 histone marks and the gene expression data all exhibit evidence of convergence in abundance between naïve and memory cells by day 14 after activation (Figure 2.11, Table 2.4). The MOFA LF scatter plots (Figure 2.7b) show that this pattern of convergence is captured in LF5. Like all the LFs in this plot, this factor explains a substantial portion of the variance in all 4 data sets, indicating a coordinated pattern of variation shared across all histone marks and gene expression. This is consistent with the expectation that any naïve $CD4^+$ T-cells remaining at day 14 should have differentiated into memory cells by that time, and should therefore have a genomic and epigenomic state similar to memory cells. This convergence is evidence that these histone marks all play an important role in the naïve-to-memory differentiation process. A histone mark that was not involved in naïve-to-memory differentiation would not be expected to converge in this way after activation.

In H3K4me2, H3K4me3, and RNA-seq, this convergence appears to be in progress already by Day 5, shown by the smaller distance between naïve and memory cells at day 5 along the y-axes in Figures 2.11a, 2.11b, and 2.11d. This agrees with the model

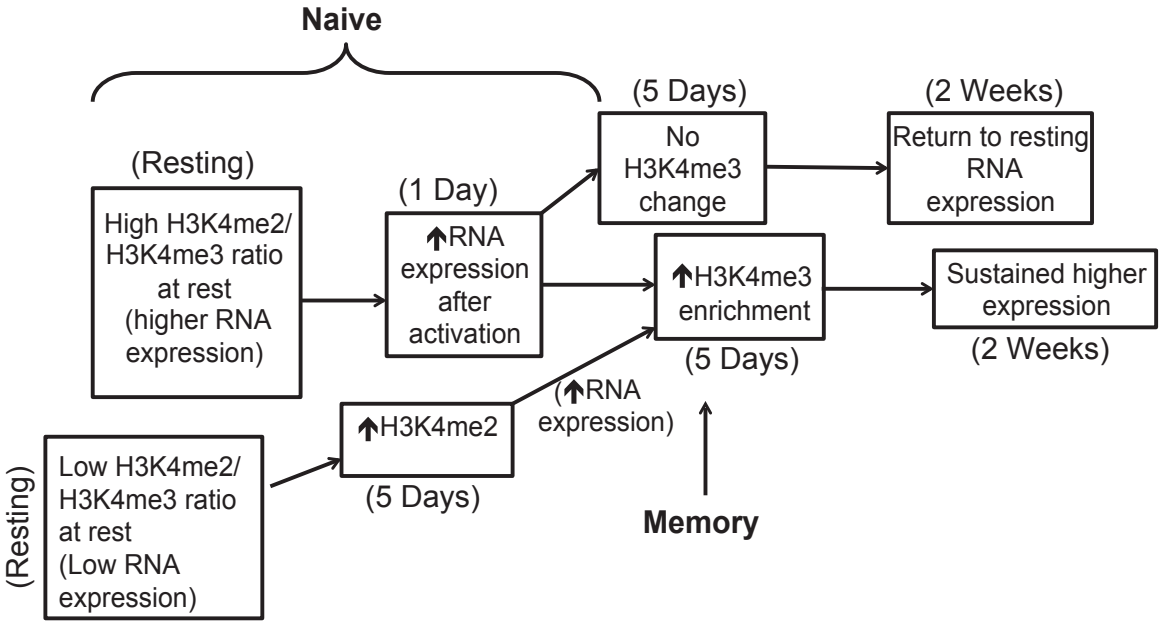


Figure 2.15: **Lamere 2016 Figure 8 [48], "Model for the role of H3K4 methylation during CD4$^+$ T-cell activation."** (Reproduced with permission.)

proposed by Sarah Lamere based on an prior analysis of the same data, shown in Figure 2.15, which shows the pattern of H3K4 methylation and expression for naïve cells and memory cells converging at day 5. This model was developed without the benefit of the PCoA plots in Figure 2.11, which have been corrected for confounding factors by ComBat and SVA. This shows that proper batch correction assists in extracting meaningful patterns in the data while eliminating systematic sources of irrelevant variation in the data, allowing simple automated procedures like PCoA to reveal interesting behaviors in the data that were previously only detectable by a detailed manual analysis. While the ideal comparison to demonstrate this convergence would be naïve cells at day 14 to memory cells at day 0, this is not feasible in this experimental system, since neither naïve nor memory cells are able to fully return to their pre-activation state, as shown by the lack of overlap between days 0 and 14 for either naïve or memory cells in Figure 2.11.

2.5.3 The location of histone modifications within the promoter is important

When looking at patterns in the relative coverage of each histone mark near the TSS of each gene, several interesting patterns were apparent. For H3K4me2 and H3K4me3, the pattern was straightforward: the consistent pattern across all promoters was a single peak a few kbp wide, with the main axis of variation being the position of this peak relative to the TSS (Figures 2.12 & 2.13). There were no obvious "preferred" positions, but rather a continuous distribution of relative positions ranging all across the promoter region. The association with gene expression was also straightforward: peaks closer to the TSS were more strongly associated with elevated gene expression. Coverage downstream of the TSS appears to be more strongly associated with elevated expression than coverage at the same distance upstream, indicating that the "effective promoter region" for H3K4me2 and H3K4me3 may be centered downstream of the TSS.

The relative promoter coverage for H3K27me3 had a more complex pattern, with two specific patterns of promoter coverage associated with elevated expression: a sharp depletion of H3K27me3 around the TSS relative to the surrounding area, and a depletion of H3K27me3 downstream of the TSS relative to upstream (Figure 2.14). A previous study found that H3K27me3 depletion within the gene body was associated with elevated gene expression in 4 different cell types in mice [72]. This is consistent with the second pattern described here. This study also reported that a spike in coverage at the TSS was associated with *lower* expression, which is indirectly consistent with the first pattern described here, in the sense that it associates lower H3K27me3 levels near the TSS with higher expression.

2.5.4 A reproducible workflow aids in analysis

The analyses described in this chapter were organized into a reproducible workflow using the Snakemake workflow management system [73]. As shown in Figure 2.16, the workflow includes many steps with complex dependencies between them. For example, the step that counts the number of ChIP-seq reads in 500 bp windows in each promoter (the starting point for Figures 2.12, 2.13, and 2.14), named `chipseq_count_tss_neighborhoods`, depends on the RNA-seq abundance estimates in order to select the most-used TSS for each gene, the aligned ChIP-seq reads, the index for those reads, and the blacklist of regions to be excluded from ChIP-seq analysis. Each step declares its inputs and outputs, and Snakemake uses these to determine the dependencies between steps. Each step is marked as depending on all the steps whose outputs match its inputs, generating the workflow graph in Figure 2.16, which Snakemake uses to determine order in which to execute each step so that each step is executed only after all of the steps it depends on have completed, thereby automating the entire workflow from start to finish.

In addition to simply making it easier to organize the steps in the analysis, structuring the analysis as a workflow allowed for some analysis strategies that would not have been practical otherwise. For example, 5 different RNA-seq quantification methods were tested against two different reference transcriptome annotations for a total of 10 different quantifications of the same RNA-seq data. These were then compared against each other in the exploratory data analysis step, to determine that the results were not very sensitive to either the choice of quantification method or the choice of annotation. This was possible with a single script for the exploratory data analysis, because Snakemake was able to automate running this script for every combination of method and reference. In a similar manner, two different peak calling methods were tested against each other, and in this case it was determined that SICER was unambiguously superior to MACS for all histone marks studied. By enabling these types of

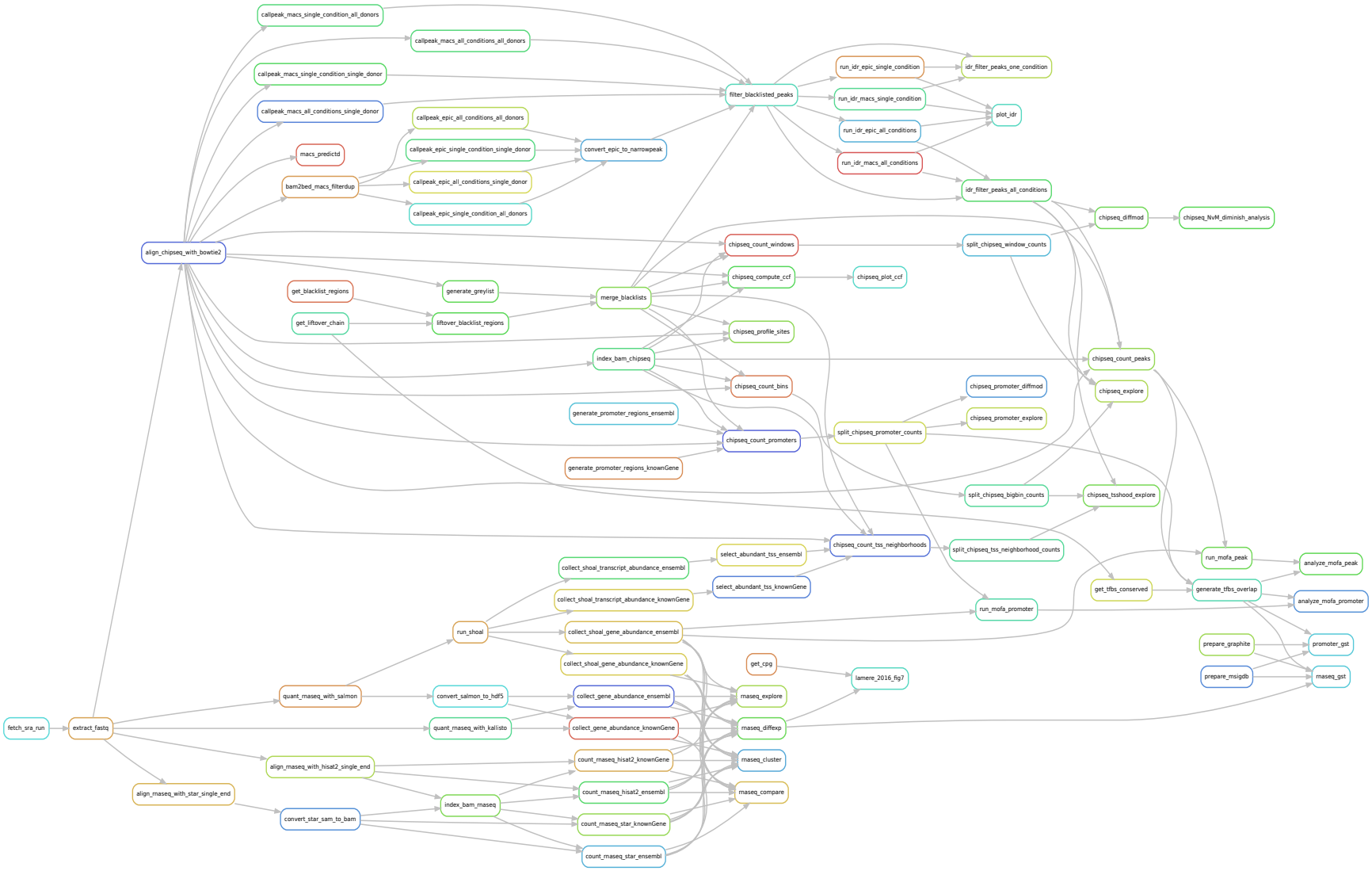


Figure 2.16: **Dependency graph of steps in reproducible workflow.** The analysis flows from left to right. Arrows indicate which analysis steps depend on the output of other steps.

comparisons, structuring the analysis as an automated workflow allowed important analysis decisions to be made in a data-driven way, by running every reasonable option through the downstream steps, seeing the consequences of choosing each option, and deciding accordingly.

2.6 Future Directions

The analysis of RNA-seq and ChIP-seq in $CD4^+$ T-cells in Chapter 2 is in many ways a preliminary study that suggests a multitude of new avenues of investigation. Here we consider a selection of such avenues.

2.6.1 Previous negative results

Two additional analyses were conducted beyond those reported in the results. First, we searched for evidence that the presence or absence of a CpG island (CpGi) in the promoter was correlated with increases or decreases in gene expression or any histone mark in any of the tested contrasts. Second, we searched for evidence that the relative ChIP-seq coverage profiles prior to activations could predict the change in expression of a gene after activation. Neither analysis turned up any clear positive results.

2.6.2 Improve on the idea of an effective promoter radius

This study introduced the concept of an "effective promoter radius" specific to each histone mark based on distance from the TSS within which an excess of peaks was called for that mark. This concept was then used to guide further analyses throughout the study. However, while the effective promoter radius was useful in those analyses, it is both limited in theory and shown in practice to be a possible oversimplification. First, the effective promoter radii used in this study were chosen based on manual inspection of the TSS-to-peak distance distributions in Figure 2.9, selecting round

numbers of analyst convenience (Table 2.3). It would be better to define an algorithm that selects a more precise radius based on the features of the graph. One possible way to do this would be to randomly rearrange the called peaks throughout the genome many (while preserving the distribution of peak widths) and re-generate the same plot as in Figure 2.9. This would yield a better "background" distribution that demonstrates the degree of near-TSS enrichment that would be expected by random chance. The effective promoter radius could be defined as the point where the true distribution diverges from the randomized background distribution.

Furthermore, the above definition of effective promoter radius has the significant limitation of being based on the peak calling method. It is thus very sensitive to the choice of peak caller and significance threshold for calling peaks, as well as the degree of saturation in the sequencing. Calling peaks from ChIP-seq samples with insufficient coverage depth, with the wrong peak caller, or with a different significance threshold could give a drastically different number of called peaks, and hence a drastically different distribution of peak-to-TSS distances. To address this, it is desirable to develop a better method of determining the effective promoter radius that relies only on the distribution of read coverage around the TSS, independent of the peak calling. Furthermore, as demonstrated by the upstream-downstream asymmetries observed in Figures 2.12, 2.13, and 2.14, this definition should determine a different radius for the upstream and downstream directions. At this point, it may be better to rename this concept "effective promoter extent" and avoid the word "radius", since a radius implies a symmetry about the TSS that is not supported by the data.

Beyond improving the definition of effective promoter extent, functional validation is necessary to show that this measure of near-TSS enrichment has biological meaning. Figures 2.12 and 2.13 already provide a very limited functional validation of the chosen promoter extents for H3K4me2 and H3K4me3 by showing that spikes in coverage within this region are most strongly correlated with elevated gene expression.

However, there are other ways to show functional relevance of the promoter extent. For example, correlations could be computed between read counts in peaks nearby gene promoters and the expression level of those genes, and these correlations could be plotted against the distance of the peak upstream or downstream of the gene's TSS. If the promoter extent truly defines a "sphere of influence" within which a histone mark is involved with the regulation of a gene, then the correlations for peaks within this extent should be significantly higher than those further upstream or downstream. Peaks within these extents may also be more likely to show differential modification than those outside genic regions of the genome.

2.6.3 Design experiments to focus on post-activation convergence of naïve & memory cells

In this study, a convergence between naïve and memory cells was observed in both the pattern of gene expression and in epigenetic state of the 3 histone marks studied, consistent with the hypothesis that any naïve cells remaining 14 days after activation have differentiated into memory cells, and that both gene expression and these histone marks are involved in this differentiation. However, the current study was not designed with this specific hypothesis in mind, and it therefore has some deficiencies with regard to testing it. The memory $CD4^+$ samples at day 14 do not resemble the memory samples at day 0, indicating that in the specific model of activation used for this experiment, the cells are not guaranteed to return to their original pre-activation state, or perhaps this process takes substantially longer than 14 days. This difference is expected, as the cell cultures in this experiment were treated with IL2 from day 5 onward [48], so the signalling environments in which the cells are cultured are different at day 0 and day 14. This is a challenge for testing the convergence hypothesis because the ideal comparison to prove that naïve cells are converging to a resting memory state would be to compare the final naïve time point to the Day 0 memory

samples, but this comparison is only meaningful if memory cells generally return to the same "resting" state that they started at.

Because pre-culture and post-culture cells will probably never behave identically even if they both nominally have a "resting" phenotype, a different experiment should be designed in which post-activation naive cells are compared to memory cells that were cultured for the same amount of time but never activated, in addition to post-activation memory cells. If the convergence hypothesis is correct, both post-activation cultures should converge on the culture of never-activated memory cells.

In addition, if naïve-to-memory convergence is a general pattern, it should also be detectable in other epigenetic marks, including other histone marks and DNA methylation. An experiment should be designed studying a large number of epigenetic marks known or suspected to be involved in regulation of gene expression, assaying all of these at the same pre- and post-activation time points. Multi-dataset factor analysis methods like MOFA can then be used to identify coordinated patterns of regulation shared across many epigenetic marks. Of course, $CD4^+$ T-cells are not the only adaptive immune cells that exhibit memory formation. A similar study could be designed for $CD8^+$ T-cells, B-cells, and even specific subsets of $CD4^+$ T-cells, such as Th1, Th2, Treg, and Th17 cells, to determine whether these also show convergence.

2.6.4 Follow up on hints of interesting patterns in promoter relative coverage profiles

The analysis of promoter coverage landscapes in resting naive $CD4^+$ T-cells and their correlations with gene expression raises many interesting questions. The chosen analysis strategy used a clustering approach, but this approach was subsequently shown to be a poor fit for the data. In light of this, a better means of dimension reduction for promoter landscape data is required. In the case of H3K4me2 and H3K4me3, one option is to define the first 3 principal componets as orthogonal promoter "state

variables": upstream vs downstream coverage, TSS-centered peak vs trough, and proximal upstream trough vs proximal downstream trough. Gene expression could then be modeled as a function of these three variables, or possibly as a function of the first N principal components for N larger than 3. For H3K4me2 and H3K4me3, a better representation might be obtained by transforming the first 2 principal coordinates into a polar coordinate system (r, θ) with the origin at the center of the "no peak" cluster, where the radius r represents the peak height above the background and the angle θ represents the peak's position upstream or downstream of the TSS.

Another weakness in the current analysis is the normalization of the average abundance of each promoter to an average of zero. This allows the abundance value in each window to represent the relative abundance of that window compared to all the other windows in the interrogated area. However, while using the remainder of the windows to set the "background" level against which each window is normalized is convenient, it is far from optimal. As shown in Table 2.2, many enriched regions are larger than the 5 kbp radius., which means there may not be any "background" regions within 5 kbp of the TSS to normalize against. For example, this normalization strategy fails to distinguish between a trough in coverage at the TSS and a pair of wide peaks upstream and downstream of the TSS. Both cases would present as lower coverage in the windows immediately adjacent to the TSS and higher coverage in windows further away, but the functional implications of these two cases might be completely different. To improve the normalization, the background estimation method used by SICER, which is specifically designed for finding broad regions of enrichment, should be adapted to estimate the background sequencing depth in each window from the ChIP-seq input samples, and each window's read count should be normalized against the background and reported as a $\log_2$ fold change (logFC) relative to that background.

Lastly, the analysis of promoter coverage landscapes presented in this work only

looked at promoter coverage of resting naive $CD4^+$ T-cells, with the goal of determining whether this initial promoter state was predictive of post-activation changes in gene expression. Changes in the promoter coverage landscape over time have not yet been considered. This represents a significant analysis challenge, by adding yet another dimension (genomic coordinate) in to the data.

2.6.5 Investigate causes of high correlation between mutually exclusive histone marks

The high correlation between coverage depth observed between H3K4me2 and H3K4me3 is both expected and unexpected. Since both marks are associated with elevated gene transcription, a positive correlation between them is not surprising. However, these two marks represent different post-translational modifications of the *same* lysine residue on the histone H3 polypeptide, which means that they cannot both be present on the same H3 subunit. Thus, the high correlation between them has several potential explanations. One possible reason is cell population heterogeneity: perhaps some genomic loci are frequently marked with H3K4me2 in some cells, while in other cells the same loci are marked with H3K4me3. Another possibility is allele-specific modifications: the loci are marked in each diploid cell with H3K4me2 on one allele and H3K4me3 on the other allele. Lastly, since each histone octamer contains 2 H3 subunits, it is possible that having one H3K4me2 mark and one H3K4me3 mark on a given histone octamer represents a distinct epigenetic state with a different function than either double H3K4me2 or double H3K4me3.

The hypothesis of allele-specific histone modification can easily be tested with existing data by locating all heterozygous loci occurring within both H3K4me3 and H3K4me2 peaks and checking for opposite allelic imbalance between H3K4me3 and H3K4me2 read at each locus. If the allele fractions in the reads from the two histone marks for each locus are plotted against each other, there should be a negative

correlation. If no such negative correlation is found, then allele-specific histone modification is unlikely to be the reason for the high correlation between these histone marks.

To test the hypothesis that H3K4me2 and H3K4me3 marks are occurring on the same histones. A double chromatin immunoprecipitation (ChIP) experiment can be performed [74]. In this assay, the input DNA goes through two sequential immunoprecipitations with different antibodies: first the anti-H3K4me2 antibody, then the anti-H3K4me3 antibody. Only bearing both histone marks, and the DNA associated with them, should be isolated. This can be followed by high-throughput sequencing (HTS) to form a "double ChIP-seq" assay that can be used to identify DNA regions bound by the isolated histones [75]. If peaks called from this double ChIP-seq assay are highly correlated with both H3K4me2 and H3K4me3 peaks, then this is strong evidence that the correlation between the two marks is actually caused by physical co-location on the same histone.

Chapter 3

Improving array-based diagnostics for transplant rejection by optimizing data preprocessing

Ryan C. Thompson, Sunil M. Kurian, Thomas Whisnant, Padmaja Natarajan, Daniel R. Salomon

3.1 Introduction

3.1.1 Proper pre-processing is essential for array data

Microarrays, bead arrays, and similar assays produce raw data in the form of fluorescence intensity measurements, with each intensity measurement proportional to the abundance of some fluorescently labelled target DNA or RNA sequence that base pairs to a specific probe sequence. However, the fluorescence measurements for each probe are also affected my many technical confounding factors, such as the concentration of target material, strength of off-target binding, the sensitivity of the imaging

sensor, and visual artifacts in the image. Some array designs also use multiple probe sequences for each target. Hence, extensive pre-processing of array data is necessary to normalize out the effects of these technical factors and summarize the information from multiple probes to arrive at a single usable estimate of abundance or other relevant quantity, such as a ratio of two abundances, for each target [76].

The choice of pre-processing algorithms used in the analysis of an array data set can have a large effect on the results of that analysis. However, despite their importance, these steps are often neglected or rushed in order to get to the more scientifically interesting analysis steps involving the actual biology of the system under study. Hence, it is often possible to achieve substantial gains in statistical power, model goodness-of-fit, or other relevant performance measures, by checking the assumptions made by each preprocessing step and choosing specific normalization methods tailored to the specific goals of the current analysis.

3.2 Approach

3.2.1 Clinical diagnostic applications for microarrays require single-channel normalization

As the cost of performing microarray assays falls, there is increasing interest in using genomic assays for diagnostic purposes, such as distinguishing healthy transplants (TX) from transplants undergoing acute rejection (AR) or acute dysfunction with no rejection (ADNR). However, the the standard normalization algorithm used for microarray data, Robust Multichip Average (RMA) [35], is not applicable in a clinical setting. Two of the steps in RMA, quantile normalization and probe summarization by median polish, depend on every array in the data set being normalized. This means that adding or removing any arrays from a data set changes the normalized values for all arrays, and data sets that have been normalized separately cannot be

compared to each other. Hence, when using RMA, any arrays to be analyzed together must also be normalized together, and the set of arrays included in the data set must be held constant throughout an analysis.

These limitations present serious impediments to the use of arrays as a diagnostic tool. When training a classifier, the samples to be classified must not be involved in any step of the training process, lest their inclusion bias the training process. Once a classifier is deployed in a clinical setting, the samples to be classified will not even *exist* at the time of training, so including them would be impossible even if it were statistically justifiable. Therefore, any machine learning application for microarrays demands that the normalized expression values computed for an array must depend only on information contained within that array. This would ensure that each array's normalization is independent of every other array, and that arrays normalized separately can still be compared to each other without bias. Such a normalization is commonly referred to as "single-channel normalization".

Frozen Robust Multichip Average (fRMA) addresses these concerns by replacing the quantile normalization and median polish with alternatives that do not introduce inter-array dependence, allowing each array to be normalized independently of all others [36]. Quantile normalization is performed against a pre-generated set of quantiles learned from a collection of 850 publicly available arrays sampled from a wide variety of tissues in the Gene Expression Omnibus (GEO). Each array's probe intensity distribution is normalized against these pre-generated quantiles. The median polish step is replaced with a robust weighted average of probe intensities, using inverse variance weights learned from the same public GEO data. The result is a normalization that satisfies the requirements mentioned above: each array is normalized independently of all others, and any two normalized arrays can be compared directly to each other.

One important limitation of fRMA is that it requires a separate reference data

set from which to learn the parameters (reference quantiles and probe weights) that will be used to normalize each array. These parameters are specific to a given array platform, and pre-generated parameters are only provided for the most common platforms, such as Affymetrix hgu133plus2. For a less common platform, such as hthgu133pluspm, is is necessary to learn custom parameters from in-house data before fRMA can be used to normalize samples on that platform [77].

One other option is the aptly-named Single Channel Array Normalization (SCAN), which adapts a normalization method originally designed for tiling arrays [39]. SCAN is truly single-channel in that it does not require a set of normalization parameters estimated from an external set of reference samples like fRMA does.

3.2.2 Heteroskedasticity must be accounted for in methylation array data

DNA methylation arrays are a relatively new kind of assay that uses microarrays to measure the degree of methylation on cytosines in specific regions arrayed across the genome. First, bisulfite treatment converts all unmethylated cytosines to uracil (which are read as thymine during amplification and sequencing) while leaving methylated cytosines unaffected. Then, each target region is interrogated with two probes: one binds to the original genomic sequence and interrogates the level of methylated DNA, and the other binds to the same sequence with all cytosines replaced by thymidines and interrogates the level of unmethylated DNA.

After normalization, these two probe intensities are summarized in one of two ways, each with advantages and disadvantages. β values, interpreted as fraction of DNA copies methylated, range from 0 to 1. β values are conceptually easy to interpret, but the constrained range makes them unsuitable for linear modeling, and their error distributions are highly non-normal, which also frustrates linear modeling. M-values, interpreted as the log ratios of methylated to unmethylated copies for each

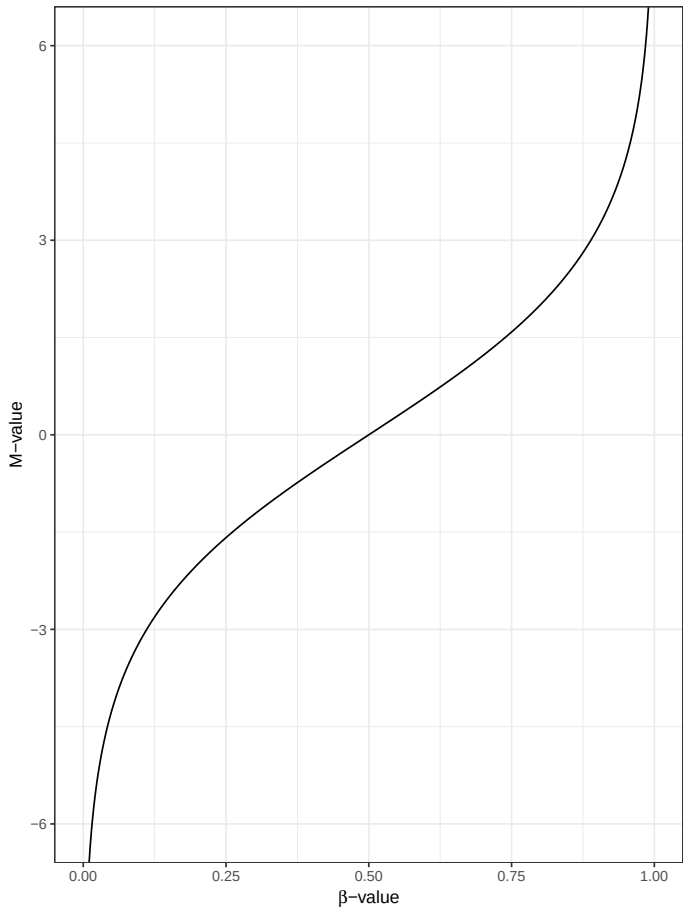


Figure 3.1: **Sigmoid shape of the mapping between β and M values.** This mapping is monotonic and non-linear, but it is approximately linear in the neighborhood of $(\beta = 0.5, M = 0)$.

probe region, are computed by mapping the beta values from $[0, 1]$ onto $(-\infty, +\infty)$ using a sigmoid curve (Figure 3.1). This transformation results in values with better statistical properties: the unconstrained range is suitable for linear modeling, and the error distributions are more normal. Hence, most linear modeling and other statistical testing on methylation arrays is performed using M-values.

However, the steep slope of the sigmoid transformation near 0 and 1 tends to over-exaggerate small differences in β values near those extremes, which in turn amplifies the error in those values, leading to a U-shaped trend in the mean-variance curve: extreme values have higher variances than values near the middle. This mean-variance dependency must be accounted for when fitting the linear model for differen-

tial methylation, or else the variance will be systematically overestimated for probes with moderate M-values and underestimated for probes with extreme M-values. This is particularly undesirable for methylation data because the intermediate M-values are the ones of most interest, since they are more likely to represent areas of varying methylation, whereas extreme M-values typically represent complete methylation or complete lack of methylation.

High-throughput RNA sequencing (RNA-seq) read count data are also known to show heteroskedasticity, and the voom method was introduced for modeling this heteroskedasticity by estimating the mean-variance trend in the data and using this trend to assign precision weights to each observation [23]. While methylation array data are not derived from counts and have a very different mean-variance relationship from that of typical RNA-seq data, the voom method makes no specific assumptions on the shape of the mean-variance relationship – it only assumes that the relationship can be modeled as a smooth curve. Hence, the method is sufficiently general to model the mean-variance relationship in methylation array data. However, while the method does not require count data as input, the standard implementation of voom assumes that the input is given in raw read counts, and it must be adapted to run on methylation M-values.

3.3 Methods

3.3.1 Evaluation of classifier performance with different normalization methods

For testing different expression microarray normalizations, a data set of 157 hgu133plus2 arrays was used, consisting of blood samples from kidney transplant patients whose grafts had been graded as TX, AR, or ADNR via biopsy and histology (46 TX, 69 AR, 42 ADNR) [5]. Additionally, an external validation set of 75 samples was gathered

from public GEO data (37 TX, 38 AR, no ADNR).

To evaluate the effect of each normalization on classifier performance, the same classifier training and validation procedure was used after each normalization method. The Prediction Analysis for Microarrays (PAM) algorithm was used to train a nearest shrunken centroid classifier on the training set and select the appropriate threshold for centroid shrinking [78]. Then the trained classifier was used to predict the class probabilities of each validation sample. From these class probabilities, receiver operating characteristic (ROC) curves and area under ROC curve (AUC) values were generated [79]. Each normalization was tested on two different sets of training and validation samples. For internal validation, the 115 TX and AR arrays in the internal set were split at random into two equal sized sets, one for training and one for validation, each containing the same numbers of TX and AR samples as the other set. For external validation, the full set of 115 TX and AR samples were used as a training set, and the 75 external TX and AR samples were used as the validation set. Thus, 2 ROC curves and AUC values were generated for each normalization method: one internal and one external. Because the external validation set contains no ADNR samples, only classification of TX and AR samples was considered. The ADNR samples were included during normalization but excluded from all classifier training and validation. This ensures that the performance on internal and external validation sets is directly comparable, since both are performing the same task: distinguishing TX from AR.

Six different normalization strategies were evaluated. First, 2 well-known non-single-channel normalization methods were considered: RMA and dChip [37, 35]. Since RMA produces expression values on a $\log_2$ scale and dChip does not, the values from dChip were $\log_2$ transformed after normalization. Next, RMA and dChip followed by Global Rank-invariant Set Normalization (GRSN) were tested [38]. Post-processing with GRSN does not turn RMA or dChip into single-channel methods, but it may help mitigate batch effects and is therefore useful as a benchmark. Lastly,

the two single-channel normalization methods, fRMA and SCAN, were tested [36, 39]. When evaluating internal validation performance, only the 157 internal samples were normalized; when evaluating external validation performance, all 157 internal samples and 75 external samples were normalized together.

For demonstrating the problem with separate normalization of training and validation data, one additional normalization was performed: the internal and external sets were each normalized separately using RMA, and the normalized data for each set were combined into a single set with no further attempts at normalizing between the two sets. This represents approximately how RMA would have to be used in a clinical setting, where the samples to be classified are not available at the time the classifier is trained.

3.3.2 Generating custom fRMA vectors for hthgu133pluspm array platform

In order to enable fRMA normalization for the hthgu133pluspm array platform, custom fRMA normalization vectors were trained using the `frmaTools` package [77]. Separate vectors were created for two types of samples: kidney graft biopsy samples and blood samples from graft recipients. For training, 341 kidney biopsy samples from 2 data sets and 965 blood samples from 5 data sets were used as the reference set. Arrays were groups into batches based on unique combinations of sample type (blood or biopsy), diagnosis (TX, AR, etc.), data set, and scan date. Thus, each batch represents arrays of the same kind that were run together on the same day. For estimating the probe inverse variance weights, frmaTools requires equal-sized batches, which means a batch size must be chosen, and then batches smaller than that size must be ignored, while batches larger than the chosen size must be downsampled. This downsampling is performed randomly, so the sampling process is repeated 5 times and the resulting normalizations are compared to each other.

To evaluate the consistency of the generated normalization vectors, the 5 fRMA vector sets generated from 5 random batch samplings were each used to normalize the same 20 randomly selected samples from each tissue. Then the normalized expression values for each probe on each array were compared across all normalizations. Each fRMA normalization was also compared against the normalized expression values obtained by normalizing the same 20 samples with ordinary RMA.

3.3.3 Modeling methylation array M-value heteroskedasticity with a modified voom implementation

To investigate the whether DNA methylation could be used to distinguish between healthy and dysfunctional transplants, a data set of 78 Illumina 450k methylation arrays from human kidney graft biopsies was analyzed for differential methylation between 4 transplant statuses: TX, transplants undergoing AR, ADNR, and chronic allograft nephropathy (CAN). The data consisted of 33 TX, 9 AR, 8 ADNR, and 28 CAN samples. The uneven group sizes are a result of taking the biopsy samples before the eventual fate of the transplant was known. Each sample was additionally annotated with a donor identifier (ID) (anonymized), sex, age, ethnicity, creatinine level, and diabetes diagnosis (all samples in this data set came from patients with either Type 1 diabetes (T1D) or Type 2 diabetes (T2D)).

The intensity data were first normalized using subset-quantile within array normalization (SWAN) [80], then converted to intensity ratios (beta values) [81]. Any probes binding to loci that overlapped annotated SNPs were dropped, and the annotated sex of each sample was verified against the sex inferred from the ratio of median probe intensities for the X and Y chromosomes. Then, the ratios were transformed to M-values.

From the M-values, a series of parallel analyses was performed, each adding additional steps into the model fit to accommodate a feature of the data (see Table 3.1).

Analysis	random effect	eBayes	SVA	weights	voom
A	Yes	Yes	No	No	No
B	Yes	Yes	Yes	Yes	No
C	Yes	Yes	Yes	Yes	Yes

Table 3.1: **Summary of analysis variants for methylation array data.** Each analysis included a different set of steps to adjust or account for various systematic features of the data. Random effect: The model included a random effect accounting for correlation between samples from the same patient [26]; eBayes: Empirical bayes squeezing of per-probe variances toward the mean-variance trend [82]; SVA: Surrogate variable analysis to account for unobserved confounders [43]; Weights: Estimate sample weights to account for differences in sample quality [25, 24]; voom: Use mean-variance trend to assign individual sample weights [23]. See the text for a more detailed explanation of each step.

For analysis A, a "basic" linear modeling analysis was performed, compensating for known confounders by including terms for the factor of interest (transplant status) as well as the known biological confounders: sex, age, ethnicity, and diabetes. Since some samples came from the same patients at different times, the intra-patient correlation was modeled as a random effect, estimating a shared correlation value across all probes [26]. Then the linear model was fit, and the variance was modeled using empirical Bayes squeezing toward the mean-variance trend [82]. Finally, t-tests or F-tests were performed as appropriate for each test: t-tests for single contrasts, and F-tests for multiple contrasts. P-values were corrected for multiple testing using the Benjamini-Hochberg (BH) procedure for false discovery rate (FDR) control [44].

For the analysis B, surrogate variable analysis (SVA) was used to infer additional unobserved sources of heterogeneity in the data [43]. These surrogate variables were added to the design matrix before fitting the linear model. In addition, sample quality weights were estimated from the data and used during linear modeling to down-weight the contribution of highly variable arrays while increasing the weight to arrays with lower variability[24]. The remainder of the analysis proceeded as in analysis A. For analysis C, the voom method was adapted to run on methylation array data and used to model and correct for the mean-variance trend using individual observation weights

[23], which were combined with the sample weights [25, 24]. Each time weights were used, they were estimated once before estimating the random effect correlation value, and then the weights were re-estimated taking the random effect into account. The remainder of the analysis proceeded as in analysis B.

3.4 Results

3.4.1 Separate normalization with RMA introduces unwanted biases in classification

To demonstrate the problem with non-single-channel normalization methods, we considered the problem of training a classifier to distinguish TX from AR using the samples from the internal set as training data, evaluating performance on the external set. First, training and evaluation were performed after normalizing all array samples together as a single set using RMA, and second, the internal samples were normalized separately from the external samples and the training and evaluation were repeated. For each sample in the validation set, the classifier probabilities from both classifiers were plotted against each other (Fig. 3.2). As expected, separate normalization biases the classifier probabilities, resulting in several misclassifications. In this case, the bias from separate normalization causes the classifier to assign a lower probability of AR to every sample.

3.4.2 fRMA and SCAN maintain classification performance while eliminating dependence on normalization strategy

For internal validation, the 6 methods' AUC values ranged from 0.816 to 0.891, as shown in Table 3.2. Among the non-single-channel normalizations, dChip outper-

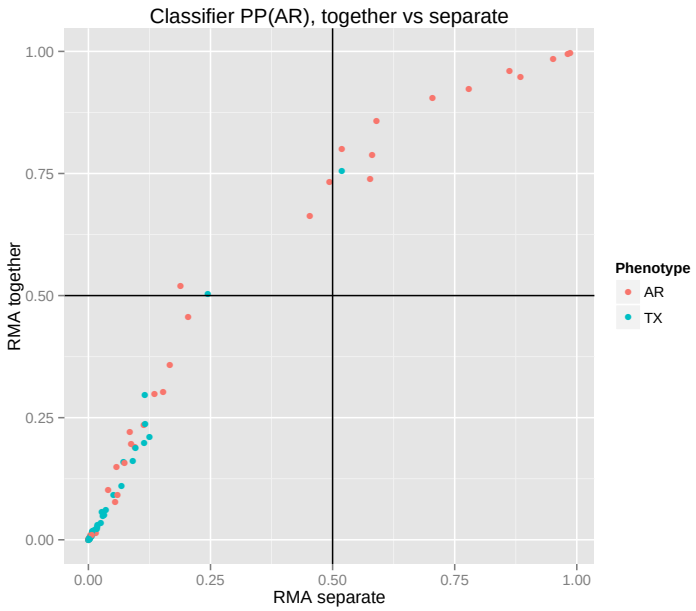


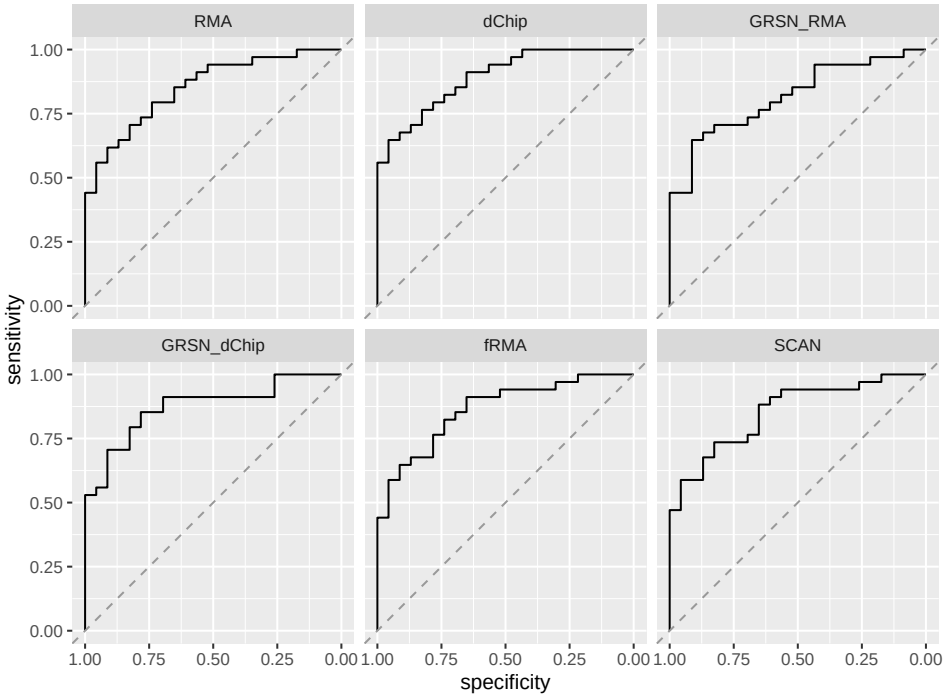
Figure 3.2: **Classifier probabilities on validation samples when normalized with RMA together vs. separately.** The PAM classifier algorithm was trained on the training set of arrays to distinguish AR from TX and then used to assign class probabilities to the validation set. The process was performed after normalizing all samples together and after normalizing the training and test sets separately, and the class probabilities assigned to each sample in the validation set were plotted against each other. Each axis indicates the posterior probability of AR assigned to a sample by the classifier in the specified analysis. The color of each point indicates the true classification of that sample.

Normalization	Single-channel?	Internal Val. AUC	External Val. AUC
RMA	No	0.852	0.713
dChip	No	0.891	0.657
RMA + GRSN	No	0.816	0.750
dChip + GRSN	No	0.875	0.642
fRMA	Yes	0.863	0.718
SCAN	Yes	0.853	0.689

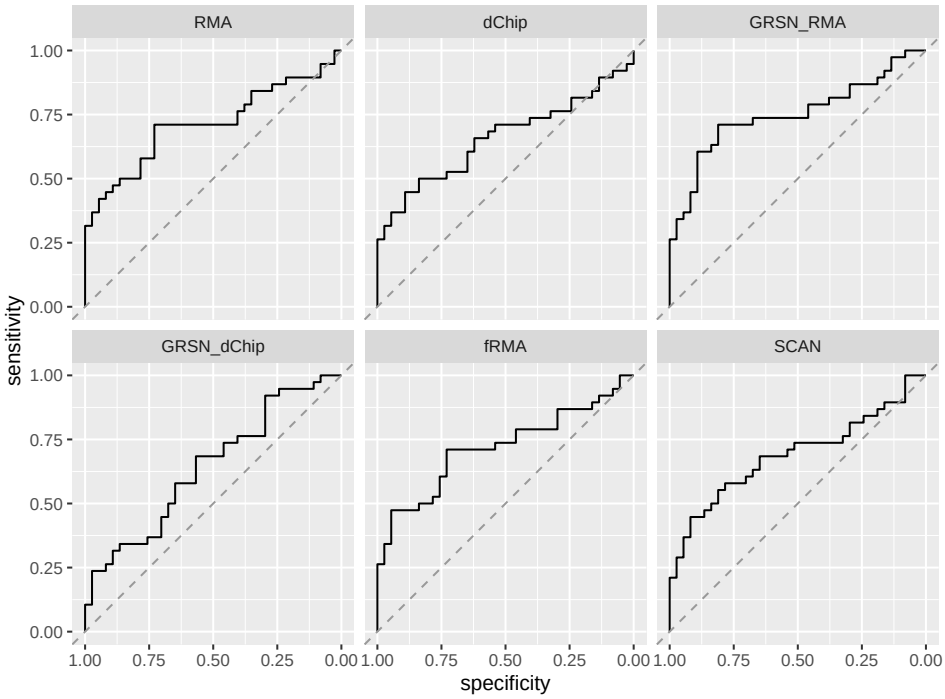
Table 3.2: **ROC curve AUC values for internal and external validation with 6 different normalization strategies.** These AUC values correspond to the ROC curves in Figure 3.3.

formed RMA, while GRSN reduced the AUC values for both dChip and RMA. Both single-channel methods, fRMA and SCAN, slightly outperformed RMA, with fRMA ahead of SCAN. However, the difference between RMA and fRMA is still quite small. Figure 3.3a shows that the ROC curves for RMA, dChip, and fRMA look very similar and relatively smooth, while both GRSN curves and the curve for SCAN have a more jagged appearance.

For external validation, as expected, all the AUC values are lower than the internal validations, ranging from 0.642 to 0.750 (Table 3.2). With or without GRSN, RMA shows its dominance over dChip in this more challenging test. Unlike in the internal validation, GRSN actually improves the classifier performance for RMA, although it does not for dChip. Once again, both single-channel methods perform about on par with RMA, with fRMA performing slightly better and SCAN performing a bit worse. Figure 3.3b shows the ROC curves for the external validation test. As expected, none of them are as clean-looking as the internal validation ROC curves. The curves for RMA, RMA+GRSN, and fRMA all look similar, while the other curves look more divergent.



(a) ROC curves for PAM on internal validation data



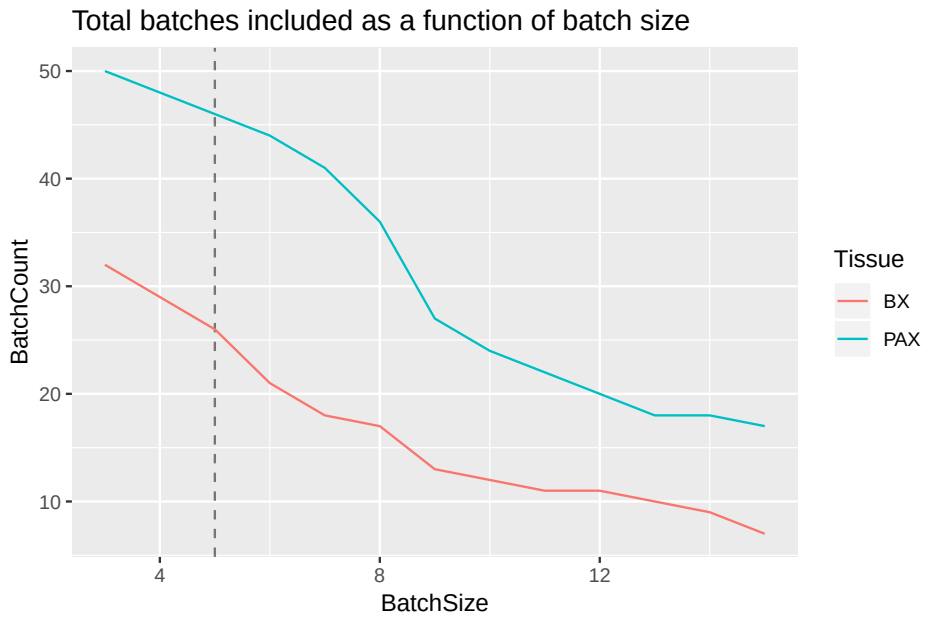
(b) ROC curves for PAM on external validation data

Figure 3.3: **ROC curves for PAM using different normalization strategies.** ROC curves were generated for PAM classification of AR vs TX after 6 different normalization strategies applied to the same data sets. Only fRMA and SCAN are single-channel normalizations. The other normalizations are for comparison.

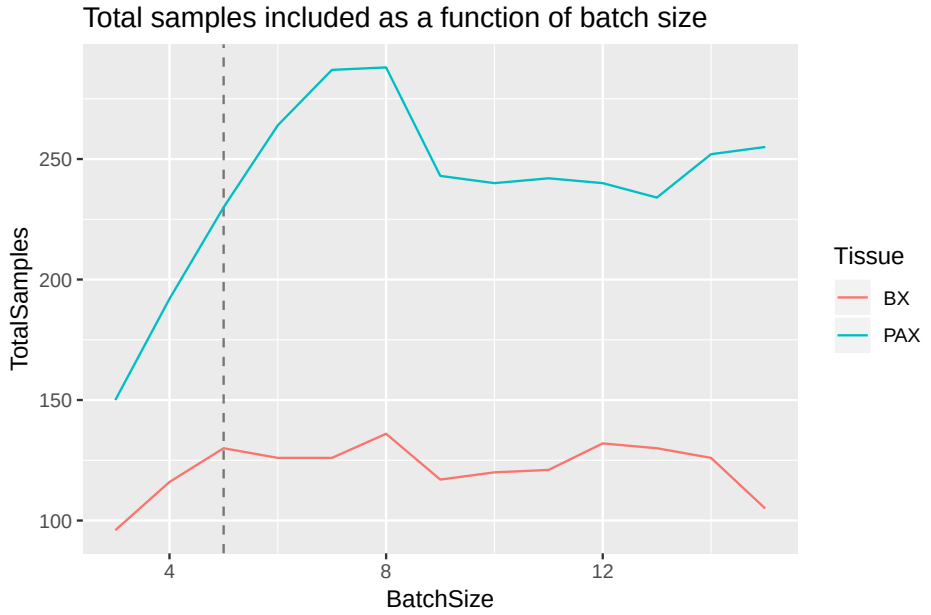
3.4.3 fRMA with custom-generated vectors enables single-channel normalization on hthgu133pluspm platform

In order to enable use of fRMA to normalize hthgu133pluspm, a custom set of fRMA vectors was created. First, an appropriate batch size was chosen by looking at the number of batches and number of samples included as a function of batch size (Figure 3.4). For a given batch size, all batches with fewer samples that the chosen size must be ignored during training, while larger batches must be randomly downsampled to the chosen size. Hence, the number of samples included for a given batch size equals the batch size times the number of batches with at least that many samples. From Figure 3.4b, it is apparent that a batch size of 8 maximizes the number of samples included in training. Increasing the batch size beyond this causes too many smaller batches to be excluded, reducing the total number of samples for both tissue types. However, a batch size of 8 is not necessarily optimal. The article introducing frmaTools concluded that it was highly advantageous to use a smaller batch size in order to include more batches, even at the cost of including fewer total samples in training [77]. To strike an appropriate balance between more batches and more samples, a batch size of 5 was chosen. For both blood and biopsy samples, this increased the number of batches included by 10, with only a modest reduction in the number of samples compared to a batch size of 8. With a batch size of 5, 26 batches of biopsy samples and 46 batches of blood samples were available.

Since fRMA training requires equal-size batches, larger batches are downsampled randomly. This introduces a nondeterministic step in the generation of normalization vectors. To show that this randomness does not substantially change the outcome, the random downsampling and subsequent vector learning was repeated 5 times, with a different random seed each time. 20 samples were selected at random as a test set and normalized with each of the 5 sets of fRMA normalization vectors as well as ordinary RMA, and the normalized expression values were compared across normalizations.



(a) **Number of batches usable in fRMA probe weight learning as a function of batch size.**



(b) **Number of samples usable in fRMA probe weight learning as a function of batch size.**

Figure 3.4: **Effect of batch size selection on number of batches and number of samples included in fRMA probe weight learning.** For batch sizes ranging from 3 to 15, the number of batches (a) and samples (b) included in probe weight training were plotted for biopsy (BX) and blood (PAX) samples. The selected batch size, 5, is marked with a dotted vertical line.

Figure 3.5 shows a summary of these comparisons for biopsy samples. Comparing RMA to each of the 5 fRMA normalizations, the distribution of log ratios is somewhat wide, indicating that the normalizations disagree on the expression values of a fair number of probe sets. In contrast, comparisons of fRMA against fRMA, the vast majority of probe sets have very small log ratios, indicating a very high agreement between the normalized values generated by the two normalizations. This shows that the fRMA normalization's behavior is not very sensitive to the random downsampling of larger batches during training.

Figure 3.7a shows an MA plot of the RMA-normalized values against the fRMA-normalized values for the same probe sets and arrays, corresponding to the first row of Figure 3.5. This MA plot shows that not only is there a wide distribution of M-values, but the trend of M-values is dependent on the average normalized intensity. This is expected, since the overall trend represents the differences in the quantile normalization step. When running RMA, only the quantiles for these specific 20 arrays are used, while for fRMA the quantile distribution is taking from all arrays used in training. Figure 3.7b shows a similar MA plot comparing 2 different fRMA normalizations, corresponding to the 6th row of Figure 3.5. The MA plot is very tightly centered around zero with no visible trend. Figures 3.6, 3.7c, and 3.7b show exactly the same information for the blood samples, once again comparing the normalized expression values between normalizations for all probe sets across 20 randomly selected test arrays. Once again, there is a wider distribution of log ratios between RMA-normalized values and fRMA-normalized, and a much tighter distribution when comparing different fRMA normalizations to each other, indicating that the fRMA training process is robust to random batch sub-sampling for the blood samples as well.

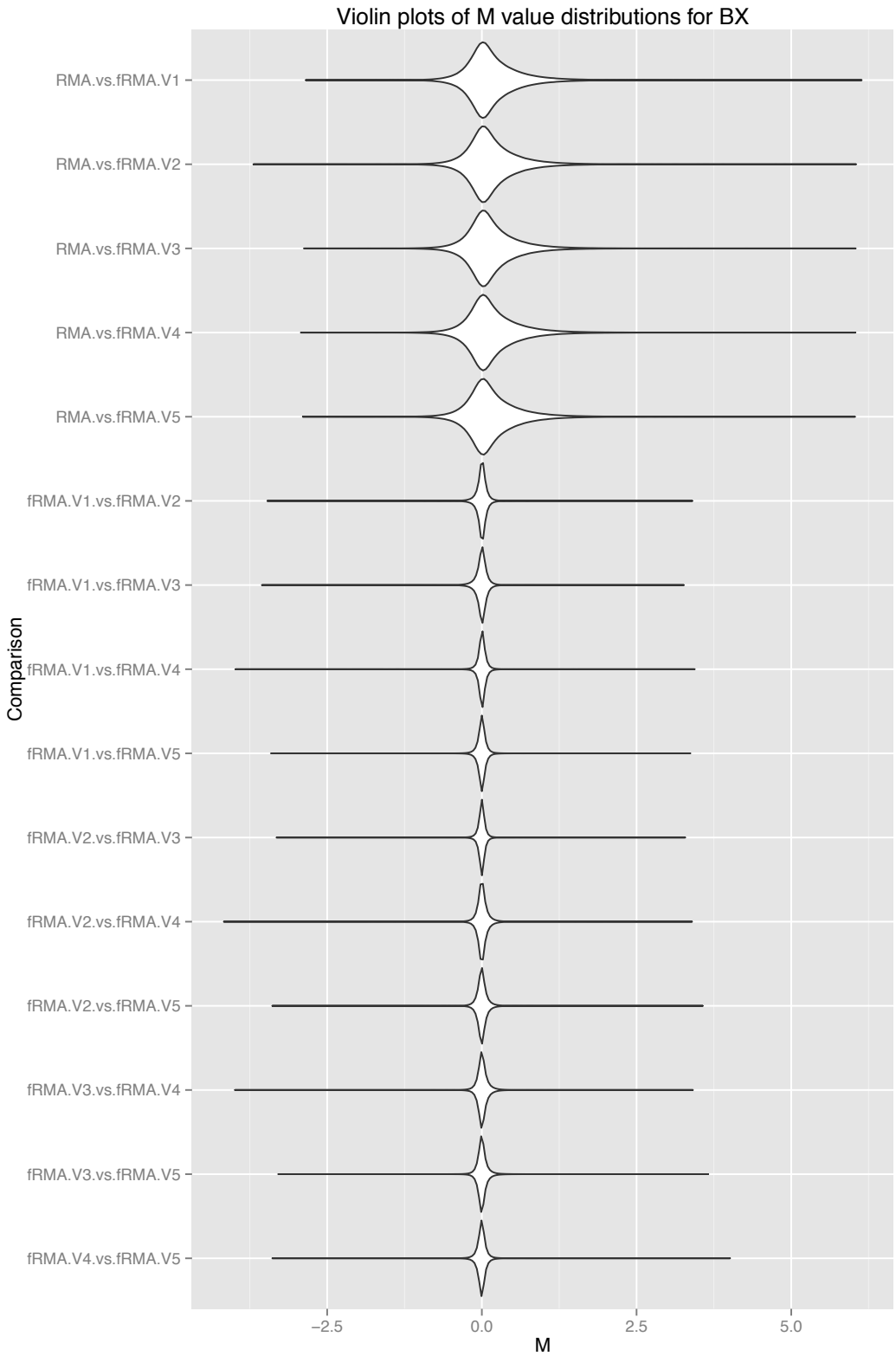


Figure 3.5: **Violin plot of log ratios between normalizations for 20 biopsy samples.** Each of 20 randomly selected samples was normalized with RMA and with 5 different sets of fRMA vectors. The distribution of log ratios between normalized expression values, aggregated across all 20 arrays, was plotted for each pair of normalizations.

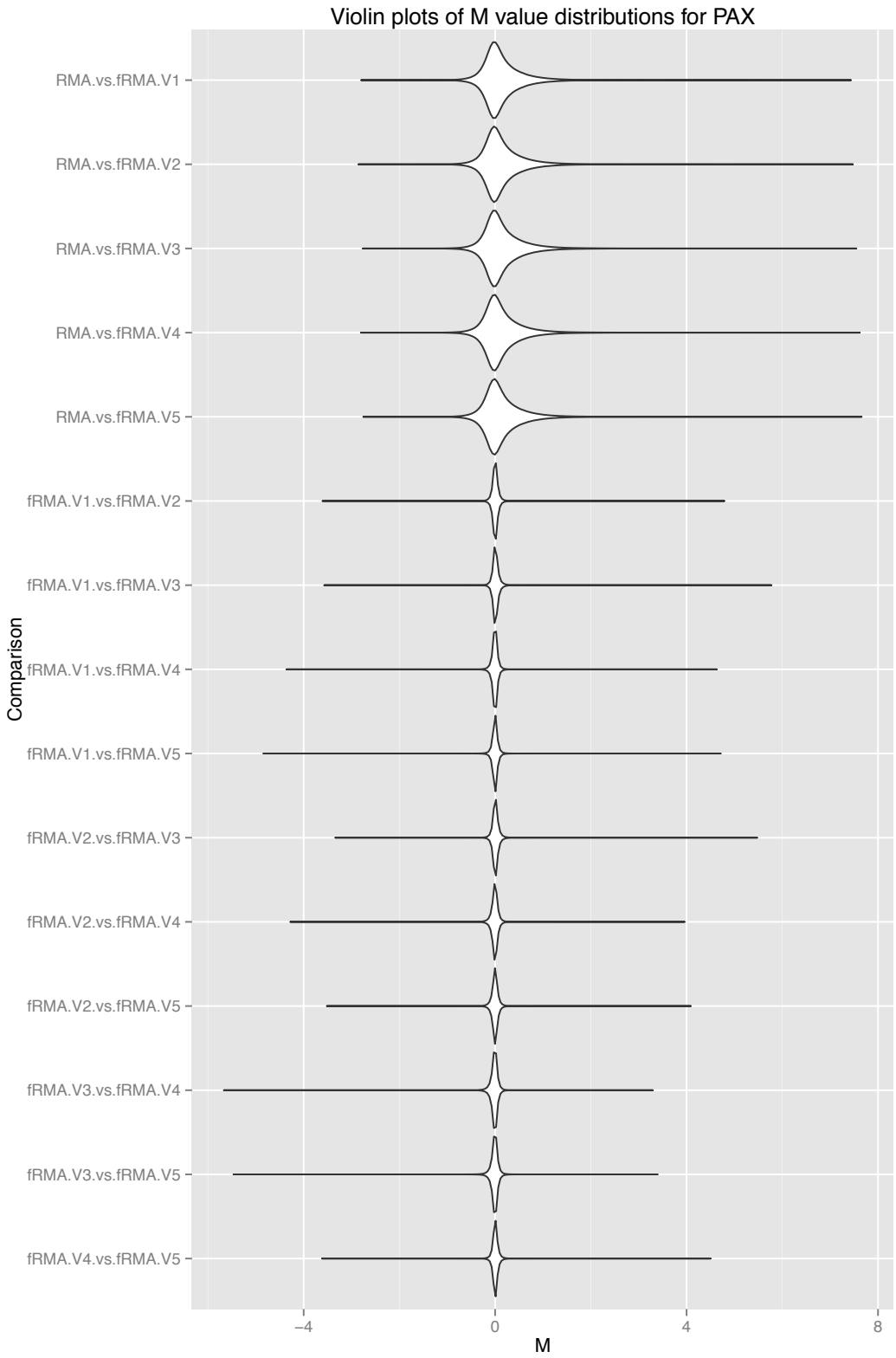
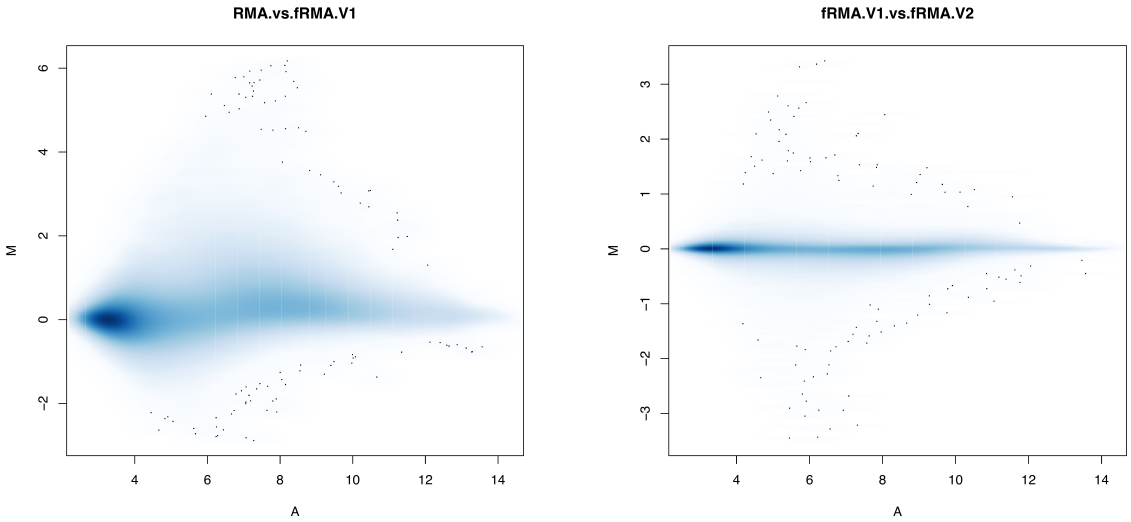
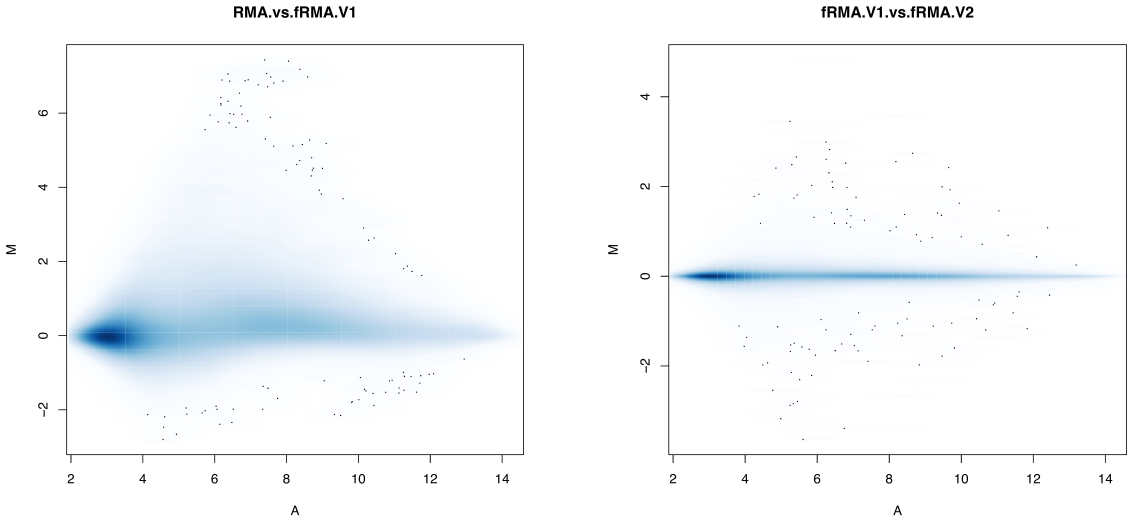


Figure 3.6: **Violin plot of log ratios between normalizations for 20 blood samples.** Each of 20 randomly selected samples was normalized with RMA and with 5 different sets of fRMA vectors. The distribution of log ratios between normalized expression values, aggregated across all 20 arrays, was plotted for each pair of normalizations.



(a) RMA vs. fRMA for biopsy samples.

(b) fRMA vs fRMA for biopsy samples.



(c) RMA vs. fRMA for blood samples.

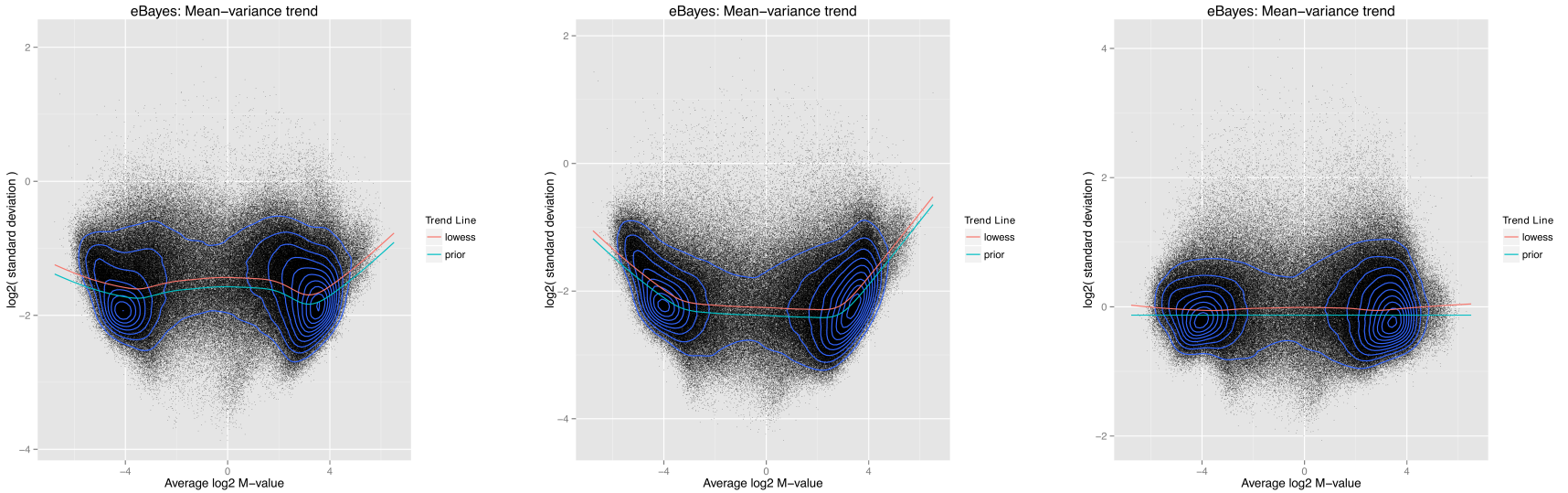
(d) fRMA vs fRMA for blood samples.

Figure 3.7: **Representative MA plots comparing RMA and custom fRMA normalizations.** For each plot, 20 samples were normalized using 2 different normalizations, and then averages (A) and log ratios (M) were plotted between the two different normalizations for every probe. For the "fRMA vs fRMA" plots (b & d), two different fRMA normalizations using vectors from two independent batch samplings were compared. Density of points is represented by blue shading, and individual outlier points are plotted.

3.4.4 SVA, voom, and array weights improve model fit for methylation array data

Figure 3.8a shows the relationship between the mean M-value and the standard deviation calculated for each probe in the methylation array data set. A few features of the data are apparent. First, the data are very strongly bimodal, with peaks in the density around M-values of +4 and -4. These modes correspond to methylation sites that are nearly 100% methylated and nearly 100% unmethylated, respectively. The strong bimodality indicates that a majority of probes interrogate sites that fall into one of these two categories. The points in between these modes represent sites that are either partially methylated in many samples, or are fully methylated in some samples and fully unmethylated in other samples, or some combination. The next visible feature of the data is the W-shaped variance trend. The upticks in the variance trend on either side are expected, based on the sigmoid transformation exaggerating small differences at extreme M-values (Figure 3.1). However, the uptick in the center is interesting: it indicates that sites that are not constitutively methylated or unmethylated have a higher variance. This could be a genuine biological effect, or it could be spurious noise that is only observable at sites with varying methylation.

In Figure 3.8b, we see the mean-variance trend for the same methylation array data, this time with surrogate variables and sample quality weights estimated from the data and included in the model. As expected, the overall average variance is smaller, since the surrogate variables account for some of the variance. In addition, the uptick in variance in the middle of the M-value range has disappeared, turning the W shape into a wide U shape. This indicates that the excess variance in the probes with intermediate M-values was explained by systematic variations not correlated with known covariates, and these variations were modeled by the surrogate variables. The result is a nearly flat variance trend for the entire intermediate M-value range from about -3 to +3. Note that this corresponds closely to the range within which the



(a) Mean-variance trend for analysis A.

(b) Mean-variance trend for analysis B.

(c) Mean-variance trend after voom modeling in analysis C.

Figure 3.8: **Mean-variance trend modeling in methylation array data.** The estimated $\log_2$(standard deviation) for each probe is plotted against the probe's average M-value across all samples as a black point, with some transparency to make over-plotting more visible, since there are about 450,000 points. Density of points is also indicated by the dark blue contour lines. The prior variance trend estimated by eBayes is shown in light blue, while the lowess trend of the points is shown in red.

M-value transformation shown in Figure 3.1 is nearly linear. In contrast, the excess variance at the extremes (greater than +3 and less than -3) was not "absorbed" by the surrogate variables and remains in the plot, indicating that this variation has no systematic component: probes with extreme M-values are uniformly more variable across all samples, as expected.

Figure 3.8c shows the mean-variance trend after fitting the model with the observation weights assigned by voom based on the mean-variance trend shown in Figure 3.8b. As expected, the weights exactly counteract the trend in the data, resulting in a nearly flat trend centered vertically at 1 (i.e. 0 on the log scale). This shows that the observations with extreme M-values have been appropriately down-weighted to account for the fact that the noise in those observations has been amplified by the non-linear M-value transformation. In turn, this gives relatively more weight to observations in the middle region, which are more likely to correspond to probes measuring interesting biology (not constitutively methylated or unmethylated).

To determine whether any of the known experimental factors had an impact on data quality, the sample quality weights estimated from the data were tested for association with each of the experimental factors (Table 3.3). Diabetes diagnosis was found to have a potentially significant association with the sample weights, with a t-test p-value of 1.06×10^{-3}. Figure 3.9 shows the distribution of sample weights grouped by diabetes diagnosis. The samples from patients with T2D were assigned significantly lower weights than those from patients with T1D. This indicates that the T2D samples had an overall higher variance on average across all probes.

Table 3.4a shows the number of significantly differentially methylated probes reported by each analysis for each comparison of interest at an FDR of 10%. As expected, the more elaborate analyses, B and C, report more significant probes than the more basic analysis A, consistent with the conclusions above that the data contain hidden systematic variations that must be modeled. Table 3.4b shows the estimated

Covariate	Test used	p-value
Transplant Status	F-test	0.404
Diabetes Diagnosis	t-test	0.00106
Sex	t-test	0.148
Age	linear regression	0.212

Table 3.3: **Association of sample weights with clinical covariates in methylation array data.** Computed sample quality log weights were tested for significant association with each of the variables in the model (1st column). An appropriate test was selected for each variable based on whether the variable had 2 categories (t-test), had more than 2 categories (F-test), or was numeric (linear regression). The test selected is shown in the 2nd column. P-values for association with the log weights are shown in the 3rd column. No multiple testing adjustment was performed for these p-values.

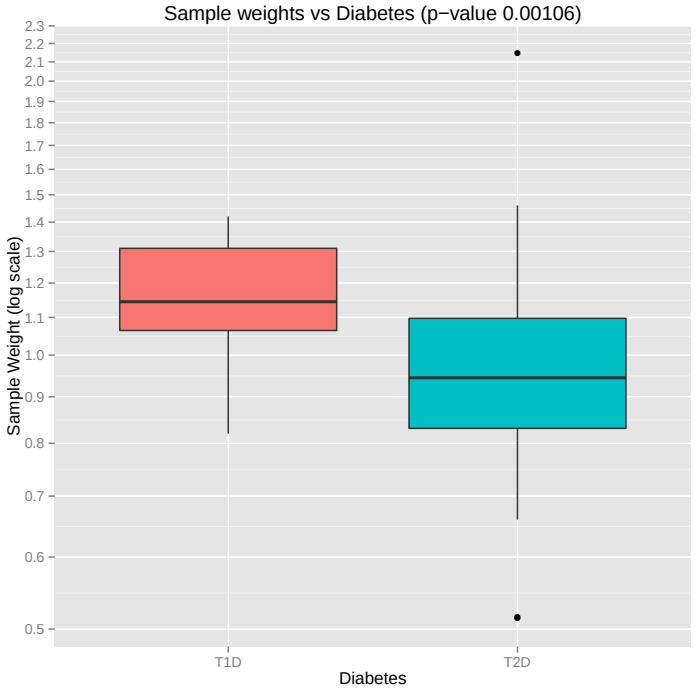


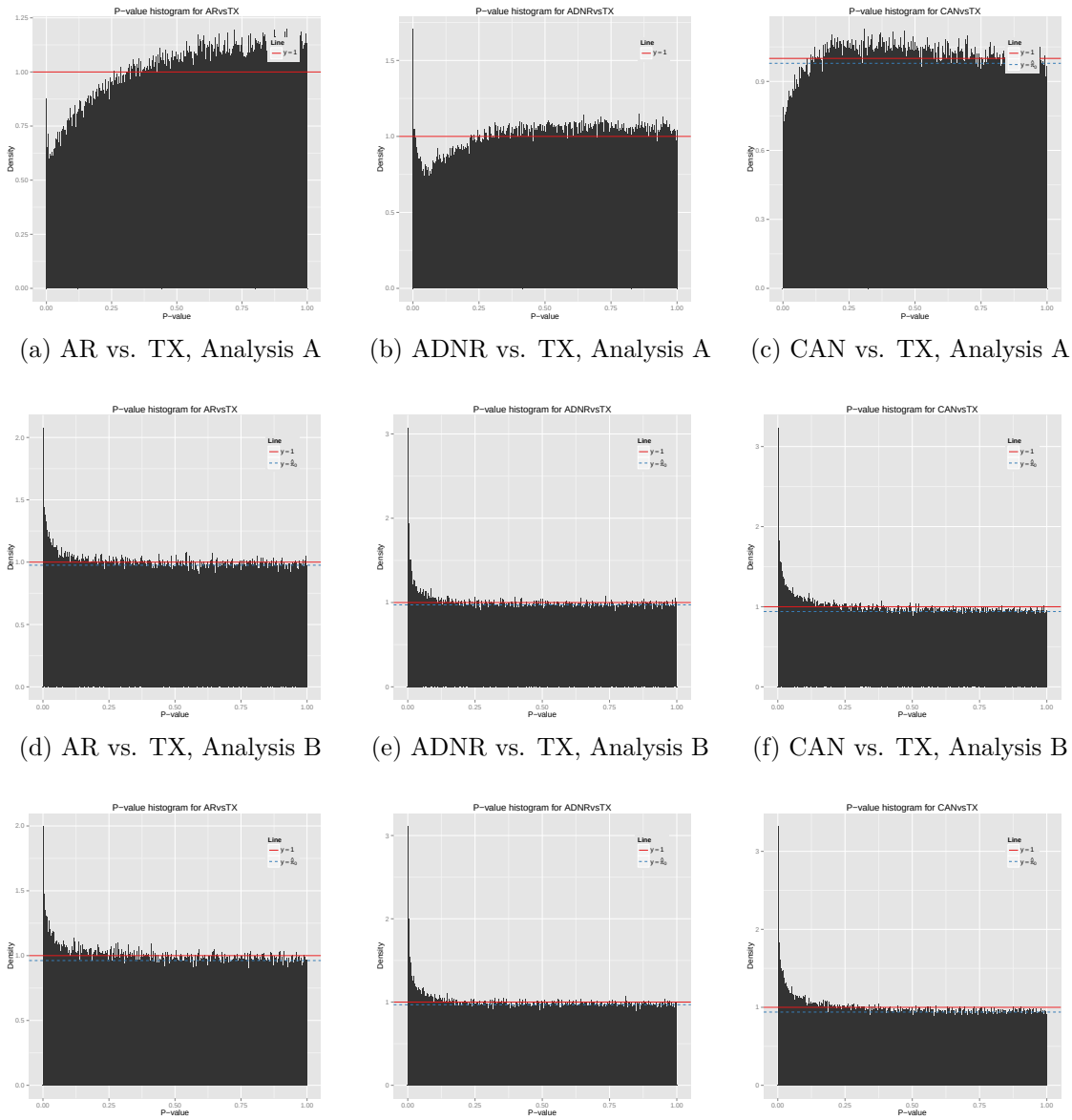
Figure 3.9: **Box-and-whiskers plot of sample quality weights grouped by diabetes diagnosis.** Samples were grouped based on diabetes diagnosis, and the distribution of sample quality weights for each diagnosis was plotted as a box-and-whiskers plot [83].

number differentially methylated probes for each test from each analysis. This was computed by estimating the proportion of null hypotheses that were true using the method of [45] and subtracting that fraction from the total number of probes, yielding an estimate of the number of null hypotheses that are false based on the distribution of p-values across the entire dataset. Note that this does not identify which null hypotheses should be rejected (i.e. which probes are significant); it only estimates the true number of such probes. Once again, analyses B and C result it much larger estimates for the number of differentially methylated probes. In this case, analysis C, the only analysis that includes voom, estimates the largest number of differentially methylated probes for all 3 contrasts. If the assumptions of all the methods employed hold, then this represents a gain in statistical power over the simpler analysis A. Figure 3.10 shows the p-value distributions for each test, from which the numbers in Table 3.4b were generated. The distributions for analysis A all have a dip in density near zero, which is a strong sign of a poor model fit. The histograms for analyses B and C are more well-behaved, with a uniform component stretching all the way from 0 to 1 representing the probes for which the null hypotheses is true (no differential methylation), and a zero-biased component representing the probes for which the null hypothesis is false (differentially methylated). These histograms do not indicate any major issues with the model fit.

3.5 Discussion

3.5.1 fRMA achieves clinically applicable normalization without sacrificing classification performance

As shown in Figure 3.2, improper normalization, particularly separate normalization of training and test samples, leads to unwanted biases in classification. In a controlled experimental context, it is always possible to correct this issue by normalizing all



(a) AR vs. TX, Analysis A (b) ADNR vs. TX, Analysis A (c) CAN vs. TX, Analysis A

(d) AR vs. TX, Analysis B (e) ADNR vs. TX, Analysis B (f) CAN vs. TX, Analysis B

(g) AR vs. TX, Analysis C (h) ADNR vs. TX, Analysis C (i) CAN vs. TX, Analysis C

Figure 3.10: **Probe p-value histograms for each contrast in each analysis.** For each differential methylation test of interest, the distribution of p-values across all probes is plotted as a histogram. The red solid line indicates the density that would be expected under the null hypothesis for all probes (a Uniform$(0, 1)$ distribution), while the blue dotted line indicates the fraction of p-values that actually follow the null hypothesis ($\hat{\pi}_0$) estimated using the method of averaging local FDR values [45]. A blue line is only shown in each plot if the estimate of $\hat{\pi}_0$ for that p-value distribution is smaller than 1.

	Analysis		
Contrast	A	B	C
TX vs AR	0	25	22
TX vs ADNR	7	338	369
TX vs CAN	0	231	278

(a) Number of probes significant at 10% FDR.

	Analysis		
Contrast	A	B	C
TX vs AR	0	10,063	11,225
TX vs ADNR	27	12,674	13,086
TX vs CAN	966	20,039	20,955

(b) Estimated number of non-null tests, using the method of averaging local FDR values [45].

Table 3.4: **Estimates of degree of differential methylation in for each contrast in each analysis.** For each of the analyses in Table 3.1, these tables show the number of probes called significantly differentially methylated at a threshold of 10% FDR for each comparison between TX and the other 3 transplant statuses (a) and the estimated total number of probes that are differentially methylated (b).

experimental samples together. However, because it is not feasible to normalize all samples together in a clinical context, a single-channel normalization is required.

The major concern in using a single-channel normalization is that non-single-channel methods can share information between arrays to improve the normalization, and single-channel methods risk sacrificing the gains in normalization accuracy that come from this information sharing. In the case of RMA, this information sharing is accomplished through quantile normalization and median polish steps. The need for information sharing in quantile normalization can easily be removed by learning a fixed set of quantiles from external data and normalizing each array to these fixed quantiles, instead of the quantiles of the data itself. As long as the fixed quantiles are reasonable, the result will be similar to standard RMA. However, there is no analogous way to eliminate cross-array information sharing in the median polish step, so fRMA replaces this with a weighted average of probes on each array, with the weights learned from external data. This step of fRMA has the greatest potential to diverge from RMA in undesirable ways.

However, when run on real data, fRMA performed at least as well as RMA in both the internal validation and external validation tests. This shows that fRMA can be used to normalize individual clinical samples in a class prediction context without

sacrificing the classifier performance that would be obtained by using the more well-established RMA for normalization. The other single-channel normalization method considered, SCAN, showed some loss of AUC in the external validation test. Based on these results, fRMA is the preferred normalization for clinical samples in a class prediction context.

3.5.2 Robust fRMA vectors can be generated for new array platforms

The published fRMA normalization vectors for the hgu133plus2 platform were generated from a set of 850 samples chosen from a wide range of tissues, which the authors determined was sufficient to generate a robust set of normalization vectors that could be applied across all tissues [36]. Since we only had hthgu133pluspm for 2 tissues of interest, our needs were more modest. Even using only 130 samples in 26 batches of 5 samples each for kidney biopsies, we were able to train a robust set of fRMA normalization vectors that were not meaningfully affected by the random selection of 5 samples from each batch. As expected, the training process was just as robust for the blood samples with 230 samples in 46 batches of 5 samples each. Because these vectors were each generated using training samples from a single tissue, they are not suitable for general use, unlike the vectors provided with fRMA itself. They are purpose-built for normalizing a specific type of sample on a specific platform. This is a mostly acceptable limitation in the context of developing a machine learning classifier for diagnosing a disease from samples of a specific tissue.

3.5.3 Methylation array data can be successfully analyzed using existing techniques, but machine learning poses additional challenges

Both analysis strategies B and C both yield a reasonable analysis, with a mean-variance trend that matches the expected behavior for the non-linear M-value transformation (Figure 3.8b) and well-behaved p-value distributions (Figure 3.10). These two analyses also yield similar numbers of significant probes (Table 3.4a) and similar estimates of the number of differentially methylated probes (Table 3.4b). The main difference between these two analyses is the method used to account for the mean-variance trend. In analysis B, the trend is estimated and applied at the probe level: each probe's estimated variance is squeezed toward the trend using an empirical Bayes procedure (Figure 3.8b). In analysis C, the trend is still estimated at the probe level, but instead of estimating a single variance value shared across all observations for a given probe, the voom method computes an initial estimate of the variance for each observation individually based on where its model-fitted M-value falls on the trend line and then assigns inverse-variance weights to model the difference in variance between observations. An overall variance is still estimated for each probe using the same empirical Bayes method, but now the residual trend is flat (Figure 3.8c), indicating that the mean-variance trend is adequately modeled by scaling the estimated variance for each observation using the weights computed by voom.

The difference between the standard empirical Bayes trended variance modeling (analysis B) and voom (analysis C) is analogous to the difference between a t-test with equal variance and a t-test with unequal variance, except that the unequal group variances used in the latter test are estimated based on the mean-variance trend from all the probes rather than the data for the specific probe being tested, thus stabilizing the group variance estimates by sharing information between probes. Allowing voom

to model the variance using observation weights in this manner allows the linear model fit to concentrate statistical power where it will do the most good. For example, if a particular probe's M-values are always at the extreme of the M-value range (e.g. less than -4) for ADNR samples, but the M-values for that probe in TX and CAN samples are within the flat region of the mean-variance trend (between -3 and $+3$), voom is able to down-weight the contribution of the high-variance M-values from the ADNR samples in order to gain more statistical power while testing for differential methylation between TX and CAN. In contrast, modeling the mean-variance trend only at the probe level would combine the high-variance ADNR samples and lower-variance samples from other conditions and estimate an intermediate variance for this probe. In practice, analysis B shows that this approach is adequate, but the voom approach in analysis C performs at least as well on all model fit criteria and yields a larger estimate for the number of differentially methylated genes, *and* it matches up slightly better with the theoretical properties of the data.

The significant association of diabetes diagnosis with sample quality is interesting. The samples with T2D tended to have more variation, averaged across all probes, than those with T1D. This is consistent with the consensus that T2D and the associated metabolic syndrome represent a broad dysregulation of the body's endocrine signaling related to metabolism [84, 85, 86]. This dysregulation could easily manifest as a greater degree of variation in the DNA methylation patterns of affected tissues. In contrast, T1D has a more specific cause and effect, so a less variable methylation signature is expected.

This preliminary analysis suggests that some degree of differential methylation exists between TX and each of the three types of transplant disfunction studied. Hence, it may be feasible to train a classifier to diagnose transplant disfunction from DNA methylation array data. However, the major importance of both SVA and sample quality weighting for proper modeling of this data poses significant challenges

for any attempt at a machine learning on data of similar quality. While these are easily used in a modeling context with full sample information, neither of these methods is directly applicable in a machine learning context, where the diagnosis is not known ahead of time. If a machine learning approach for methylation-based diagnosis is to be pursued, it will either require machine-learning-friendly methods to address the same systematic trends in the data that SVA and sample quality weighting address, or it will require higher quality data with substantially less systematic perturbation of the data.

3.6 Future Directions

3.6.1 Improving fRMA to allow training from batches of unequal size

Because the tools for building fRMA normalization vectors require equal-size batches, many samples must be discarded from the training data. This is undesirable for a few reasons. First, more data is simply better, all other things being equal. In this case, "better" means a more precise estimate of normalization parameters. In addition, the samples to be discarded must be chosen arbitrarily, which introduces an unnecessary element of randomness into the estimation process. While the randomness can be made deterministic by setting a consistent random seed, the need for equal size batches also introduces a need for the analyst to decide on the appropriate trade-off between batch size and the number of batches. This introduces an unnecessary and undesirable "researcher degree of freedom" into the analysis, since the generated normalization vectors now depend on the choice of batch size based on vague selection criteria and instinct, which can unintentionally introduce bias if the researcher chooses a batch size based on what seems to yield the most favorable downstream results [87].

Fortunately, the requirement for equal-size batches is not inherent to the fRMA

algorithm but rather a limitation of the implementation in the `frmaTools` package. In personal communication, the package's author, Matthew McCall, has indicated that with some work, it should be possible to improve the implementation to work with batches of unequal sizes. The current implementation ignores the batch size when calculating with-batch and between-batch residual variances, since the batch size constant cancels out later in the calculations as long as all batches are of equal size. Hence, the calculations of these parameters would need to be modified to remove this optimization and properly calculate the variances using the full formula. Once this modification is made, a new strategy would need to be developed for assessing the stability of parameter estimates, since the random sub-sampling step is eliminated, meaning that different sub-samplings can no longer be compared as in Figures **??** and 3.7. Bootstrap resampling is likely a good candidate here: sample many training sets of equal size from the existing training set with replacement, estimate parameters from each resampled training set, and compare the estimated parameters between bootstraps in order to quantify the variability in each parameter's estimation.

3.6.2 Developing methylation arrays as a diagnostic tool for kidney transplant rejection

The current study has showed that DNA methylation, as assayed by Illumina 450k methylation arrays, has some potential for diagnosing transplant dysfunctions, including rejection. However, very few probes could be confidently identified as differentially methylated between healthy and dysfunctional transplants. One likely explanation for this is the predominant influence of unobserved confounding factors. SVA can model and correct for such factors, but the correction can never be perfect, so some degree of unwanted systematic variation will always remain after SVA correction. If the effect size of the confounding factors was similar to that of the factor of interest (in this case, transplant status), this would be an acceptable limitation, since re-

moving most of the confounding factors' effects would allow the main effect to stand out. However, in this data set, the confounding factors have a much larger effect size than transplant status, which means that the small degree of remaining variation not removed by SVA can still swamp the effect of interest, making it difficult to detect. This is, of course, a major issue when the end goal is to develop a classifier to diagnose transplant rejection from methylation data, since batch-correction methods like SVA that work in a linear modeling context cannot be applied in a machine learning context.

Currently, the source of these unwanted systematic variations in the data is unknown. The best solution would be to determine the cause of the variation and eliminate it, thereby eliminating the need to model and remove that variation. However, if this proves impractical, another option is to use SVA to identify probes that are highly associated with the surrogate variables that describe the unwanted variation in the data. These probes could be discarded prior to classifier training, in order to maximize the chance that the training algorithm will be able to identify highly predictive probes from those remaining. Lastly, it is possible that some of this unwanted variation is a result of the array-based assay being used and would be eliminated by switching to assaying DNA methylation using bisulphite sequencing. However, this carries the risk that the sequencing assay will have its own set of biases that must be corrected for in a different way.

Chapter 4

Globin-blocking for more effective blood RNA-seq analysis in primate animal model

Ryan C. Thompson, Terri Gelbart, Steven R. Head, Phillip Ordoukhanian, Courtney Mullen, Dongmei Han, Dora Berman, Amelia Bartholomew, Norma Kenyon, Daniel R. Salomon

Abstract

Background Primate blood contains high concentrations of globin messenger RNA (mRNA). Globin reduction is a standard technique used to improve the expression results obtained by DNA microarrays on RNA from blood samples. However, with high-throughput RNA sequencing (RNA-seq) quickly replacing microarrays for many applications, the impact of globin reduction for RNA-seq is less well-studied. Moreover, no off-the-shelf kits are available for globin reduction in nonhuman primates.

Results Here we report a protocol for RNA-seq in primate blood samples that uses complimentary oligonucleotides (oligos) to block reverse transcription of the alpha and beta globin genes. In test samples from cynomolgus monkeys (*Macaca fascicularis*), this globin blocking (GB) protocol approximately doubles the yield of informative (non-globin) reads by greatly reducing the fraction of globin reads, while also improving the consistency in sequencing depth between samples. The increased yield enables detection of about 2000 more genes, significantly increases the correlation in measured gene expression levels between samples, and increases the sensitivity of differential gene expression tests.

Conclusions These results show that GB significantly improves the cost-effectiveness of RNA-seq in primate blood samples by doubling the yield of useful reads, allowing detection of more genes, and improving the precision of gene expression measurements. Based on these results, a globin reducing or blocking protocol is recommended for all RNA-seq studies of primate blood samples.

4.1 Introduction

As part of a multi-lab PO1 grant to study mesenchymal stem cell (MSC) infusion as a treatment for graft rejection in cynomolgus monkeys (*Macaca fascicularis*), a large number of serial blood draws from cynomolgus monkeys were planned in order to monitor the progress of graft healing and eventual rejection after transplantation. In order to streamline the process of performing high-throughput RNA sequencing (RNA-seq) on these blood samples, we developed a custom sequencing protocol. In the developement of this protocol, we required a solution for the problem of excess globin reads. High fractions of globin messenger RNA (mRNA) are naturally present in mammalian peripheral blood samples (up to 70% of total mRNA) and these are known to interfere with the results of array-based expression profiling [88]. Globin re-

duction is also necessary for RNA-seq of blood samples, though for unrelated reasons: without globin reduction, many RNA-seq reads will be derived from the globin genes, leaving fewer for the remainder of the genes in the transcriptome. However, existing strategies for globin reduction require an additional step during sample preparation to deplete the population of globin transcripts from the sample prior to reverse transcription [89, 90, 91]. Furthermore, off-the-shelf globin reduction kits are generally targeted at human or mouse globin, not cynomolgus monkey, and sequence identity between human and cyno globin genes cannot be automatically assumed. Hence, we sought to incorporate a custom globin reduction method into our RNA-seq protocol purely by adding additional reagents to an existing step in the sample preparation.

4.2 Approach

We evaluated globin reduction for RNA-seq by blocking reverse transcription of globin transcripts using custom blocking oligonucleotides (oligos). We demonstrate that globin blocking (GB) significantly improves the cost-effectiveness of RNA-seq in blood samples. Thus, our protocol offers a significant advantage to any investigator planning to use RNA-seq for gene expression profiling of nonhuman primate blood samples. Our method can be generally applied to any species by designing complementary oligo blocking probes to the globin gene sequences of that species. Indeed, any highly expressed but biologically uninformative transcripts can also be blocked to further increase sequencing efficiency and value [92].

4.3 Methods

4.3.1 Sample collection

All research reported here was done under IACUC-approved protocols at the University of Miami and complied with all applicable federal and state regulations and ethical principles for nonhuman primate research. Blood draws occurred between 16 April 2012 and 18 June 2015. The experimental system involved intrahepatic pancreatic islet transplantation into Cynomolgus monkeys with induced diabetes mellitus with or without concomitant infusion of mesenchymal stem cells. Blood was collected at serial time points before and after transplantation into PAXgene Blood RNA tubes (PreAnalytiX/Qiagen, Valencia, CA) at the precise volume:volume ratio of 2.5 ml whole blood into 6.9 ml of PAX gene additive.

4.3.2 Globin blocking oligonucleotide design

Four oligos were designed to hybridize to the 3′ end of the transcripts for the Cynomolgus alpha and beta globin, with two hybridization sites for each gene. All oligos were purchased from Sigma and were entirely composed of 2′O-Me bases with a C3 spacer positioned at the 3′ ends to prevent any polymerase mediated primer extension.

HBA1/2 site 1: `GCCCACUCAGACUUUAUUCAAAG-C3spacer`

HBA1/2 site 2: `GGUGCAAGGAGGGGAGGAG-C3spacer`

HBB site 1: `AAUGAAAAUAAAUGUUUUUUAUUAG-C3spacer`

HBB site 2: `CUCAAGGCCCUUCAUAAUAUCCC-C3spacer`

4.3.3 RNA-seq library preparation

Sequencing libraries were prepared with 200 ng total RNA from each sample. Polyadenylated mRNA was selected from 200 ng aliquots of cynomolgus blood-derived total RNA using Ambion Dynabeads Oligo(dT)25 beads (Invitrogen) following the manufacturer's recommended protocol. PolyA selected RNA was then combined with 8 pmol of HBA1/2 (site 1), 8 pmol of HBA1/2 (site 2), 12 pmol of HBB (site 1) and 12 pmol of HBB (site 2) oligos. In addition, 20 pmol of RT primer containing a portion of the Illumina adapter sequence (B-oligo-dTV: GAGTTCCTTGGCACCCGAGAATTCCATTTTTTTTTTTTTTTTTV) and 4 μL of 5X First Strand buffer (250 mM Tris-HCl pH 8.3, 375 mM KCl, 15 mM $MgCl_2$) were added in a total volume of 15 µL. The RNA was fragmented by heating this cocktail for 3 minutes at 95°C and then placed on ice. This was followed by the addition of 2 µL 0.1 M DTT, 1 µL RNaseOUT, 1 µL 10 mM dNTPs 10% biotin-16 aminoallyl-2′- dUTP and 10% biotin-16 aminoallyl-2′-dCTP (TriLink Biotech, San Diego, CA), 1 µL Superscript II (200 U/µL, Thermo-Fisher). A second "unblocked" library was prepared in the same way for each sample but replacing the blocking oligos with an equivalent volume of water. The reaction was carried out at 25°C for 15 minutes and 42°C for 40 minutes, followed by incubation at 75°C for 10 minutes to inactivate the reverse transcriptase.

The cDNA/RNA hybrid molecules were purified using 1.8X Ampure XP beads (Agencourt) following supplier's recommended protocol. The cDNA/RNA hybrid was eluted in 25 µL of 10 mM Tris-HCl pH 8.0, and then bound to 25 µL of M280 Magnetic Streptavidin beads washed per recommended protocol (Thermo-Fisher). After 30 minutes of binding, beads were washed one time in 100 µL 0.1 N NaOH to denature and remove the bound RNA, followed by two 100 µL washes with 1X TE buffer.

Subsequent attachment of the 5′ Illumina A adapter was performed by on-bead random primer extension of the following sequence (A-N8 primer: `TTCAGAGTTCTACAGTCCGACGATCNNN`

Briefly, beads were resuspended in a 20 µL reaction containing 5 µM A-N8 primer, 40 mM Tris-HCl pH 7.5, 20 mM $MgCl_2$, 50 mM NaCl, 0.325 U/µL Sequenase 2.0 (Affymetrix, Santa Clara, CA), 0.0025 U/µL inorganic pyrophosphatase (Affymetrix) and 300 µM each dNTP. Reaction was incubated at 22°C for 30 minutes, then beads were washed 2 times with 1X TE buffer (200 µL).

The magnetic streptavidin beads were resuspended in 34 µL nuclease-free water and added directly to a polymerase chain reaction (PCR) tube. The two Illumina protocol-specified PCR primers were added at 0.53 µM (Illumina TruSeq Universal Primer 1 and Illumina TruSeq barcoded PCR primer 2), along with 40 µL 2X KAPA HiFi Hotstart ReadyMix (KAPA, Willmington MA) and thermocycled as follows: starting with 98°C (2 min-hold); 15 cycles of 98°C, 20sec; 60°C, 30sec; 72°C, 30sec; and finished with a 72°C (2 min-hold).

PCR products were purified with 1X Ampure Beads following manufacturer's recommended protocol. Libraries were then analyzed using the Agilent TapeStation and quantitation of desired size range was performed by "smear analysis". Samples were pooled in equimolar batches of 16 samples. Pooled libraries were size selected on 2% agarose gels (E-Gel EX Agarose Gels; Thermo-Fisher). Products were cut between 250 and 350 bp (corresponding to insert sizes of 130 to 230 bp). Finished library pools were then sequenced on the Illumina NextSeq500 instrument with 75 bp read lengths.

4.3.4 Read alignment and counting

Reads were aligned to the cynomolgus genome using STAR [**Dobin2012**, 93]. Counts of uniquely mapped reads were obtained for every gene in each sample with the `featureCounts` function from the `Rsubread` package, using each of the three possibilities for the `strandSpecific` option: sense, antisense, and unstranded [53]. A few artifacts in the cynomolgus genome annotation complicated read counting. First, no

ortholog is annotated for alpha globin in the cynomolgus genome, presumably because the human genome has two alpha globin genes with nearly identical sequences, making the orthology relationship ambiguous. However, two loci in the cynomolgus genome are annotated as "hemoglobin subunit alpha-like" (LOC102136192 and LOC102136846). LOC102136192 is annotated as a pseudogene while LOC102136846 is annotated as protein-coding. Our globin reduction protocol was designed to include blocking of these two genes. Indeed, these two genes together have almost the same read counts in each library as the properly-annotated HBB gene and much larger counts than any other gene in the unblocked libraries, giving confidence that reads derived from the real alpha globin are mapping to both genes. Thus, reads from both of these loci were counted as alpha globin reads in all further analyses. The second artifact is a small, uncharacterized non-coding RNA gene (LOC102136591), which overlaps the HBA-like gene (LOC102136192) on the opposite strand. If counting is not performed in stranded mode (or if a non-strand-specific sequencing protocol is used), many reads mapping to the globin gene will be discarded as ambiguous due to their overlap with this non-coding RNA (ncRNA) gene, resulting in significant undercounting of globin reads. Therefore, stranded sense counts were used for all further analysis in the present study to insure that we accurately accounted for globin transcript reduction. However, we note that stranded reads are not necessary for RNA-seq using our protocol in standard practice.

4.3.5 Normalization and exploratory data analysis

Libraries were normalized by computing scaling factors using the `edgeR` package's trimmed mean of M-values (TMM) method [40]. Log_2 counts per million (logCPM) values were calculated using the `cpm` function in `edgeR` for individual samples and `aveLogCPM` function for averages across groups of samples, using those functions' default prior count values to avoid taking the logarithm of 0. Genes were considered

"present" if their average normalized logCPM values across all libraries were at least -1. Normalizing for gene length was unnecessary because the sequencing protocol is $3'$-biased and hence the expected read count for each gene is related to the transcript's copy number but not its length.

In order to assess the effect of GB on reproducibility, Pearson and Spearman correlation coefficients were computed between the logCPM values for every pair of libraries within the GB non-GB groups, and `edgeR`'s `estimateDisp` function was used to compute negative binomial (NB) dispersions separately for the two groups [27].

4.3.6 Differential expression analysis

All tests for differential gene expression were performed using `edgeR`, by first fitting a NB generalized linear model (GLM) to the counts and normalization factors and then performing a quasi-likelihood F-test with robust estimation of outlier gene dispersions [30, 61]. To investigate the effects of GB on each gene, an additive model was fit to the full data with coefficients for GB and Sample identifier (ID). To test the effect of GB on detection of differentially expressed genes, the GB samples and non-GB samples were each analyzed independently as follows: for each animal with both a pre-transplant and a post-transplant time point in the data set, the pre-transplant sample and the earliest post-transplant sample were selected, and all others were excluded, yielding a pre-/post-transplant pair of samples for each animal ($N = 7$ animals with paired samples). These samples were analyzed for pre-transplant vs. post-transplant differential gene expression while controlling for inter-animal variation using an additive model with coefficients for transplant and animal ID. In all analyses, p-values were adjusted using the Benjamini-Hochberg (BH) procedure for false discovery rate (FDR) control [44].

4.4 Results

4.4.1 Globin blocking yields a larger and more consistent fraction of useful reads

The objective of the present study was to validate a new protocol for deep RNA-seq of whole blood drawn into PaxGene tubes from cynomolgus monkeys undergoing islet transplantation, with particular focus on minimizing the loss of useful sequencing space to uninformative globin reads. The details of the analysis with respect to transplant outcomes and the impact of mesenchymal stem cell treatment will be reported in a separate manuscript (in preparation). To focus on the efficacy of our GB protocol, 37 blood samples, 16 from pre-transplant and 21 from post-transplant time points, were each prepped once with and once without GB oligos, and were then sequenced on an Illumina NextSeq500 instrument. The number of reads aligning to each gene in the cynomolgus genome was counted. Table 4.1 summarizes the distribution of read fractions among the GB and non-GB libraries. In the libraries with no GB, globin reads made up an average of 44.6% of total input reads, while reads assigned to all other genes made up an average of 26.3%. The remaining reads either aligned to intergenic regions (that include long non-coding RNAs) or did not align with any annotated transcripts in the current build of the cynomolgus genome. In the GB libraries, globin reads made up only 3.48% and reads assigned to all other genes increased to 50.4%. Thus, GB resulted in a 92.2% reduction in globin reads and a 91.6% increase in yield of useful non-globin reads.

This reduction is not quite as efficient as the previous analysis showed for human samples by DeepSAGE (<0.4% globin reads after globin reduction) [89]. Nonetheless, this degree of globin reduction is sufficient to nearly double the yield of useful reads. Thus, GB cuts the required sequencing effort (and costs) to achieve a target coverage depth by almost 50%. Consistent with this near doubling of yield, the av-

GB	Percent of Total Reads				Percent of Genic Reads	
	Non-globin Reads	Globin Reads	All Genic Reads	All Aligned Reads	Non-globin Reads	Globin Reads
Yes	50.4% ± 6.82	3.48% ± 2.94	53.9% ± 6.81	89.7% ± 2.40	93.5% ± 5.25	6.49% ± 5.25
No	26.3% ± 8.95	44.6% ± 16.6	70.1% ± 9.38	90.7% ± 5.16	38.8% ± 17.1	61.2% ± 17.1

Table 4.1: **Fractions of reads mapping to genomic features in GB and non-GB samples.** All values are given as mean ± standard deviation.

erage difference in un-normalized logCPM across all genes between the GB libraries and non-GB libraries is approximately 1 (mean = 1.01, median = 1.08), an overall 2-fold increase. Un-normalized values are used here because the TMM normalization correctly identifies this 2-fold difference as biologically irrelevant and removes it.

Another important aspect is that the standard deviations in Table 4.1 are uniformly smaller in the GB samples than the non-GB ones, indicating much greater consistency of yield. This is best seen in the percentage of non-globin reads as a fraction of total reads aligned to annotated genes (genic reads). For the non-GB samples, this measure ranges from 10.9% to 80.9%, while for the GB samples it ranges from 81.9% to 99.9% (Figure 4.1). This means that for applications where it is critical that each sample achieve a specified minimum coverage in order to provide useful information, it would be necessary to budget up to 10 times the sequencing depth per sample without GB, even though the average yield improvement for GB is only 2-fold, because every sample has a chance of being 90% globin and 10% useful reads. Hence, the more consistent behavior of GB samples makes planning an experiment easier and more efficient because it eliminates the need to over-sequence every sample in order to guard against the worst case of a high-globin fraction.

4.4.2 Globin blocking lowers the noise floor and allows detection of about 2000 more low-expression genes

Since GB yields more usable sequencing depth, it should also allow detection of more genes at any given threshold. When we looked at the distribution of average normalized logCPM values across all libraries for genes with at least one read assigned to them, we observed the expected bimodal distribution, with a high-abundance "signal" peak representing detected genes and a low-abundance "noise" peak representing genes whose read count did not rise above the noise floor (Figure 4.2). Consistent with the 2-fold increase in raw counts assigned to non-globin genes, the signal peak

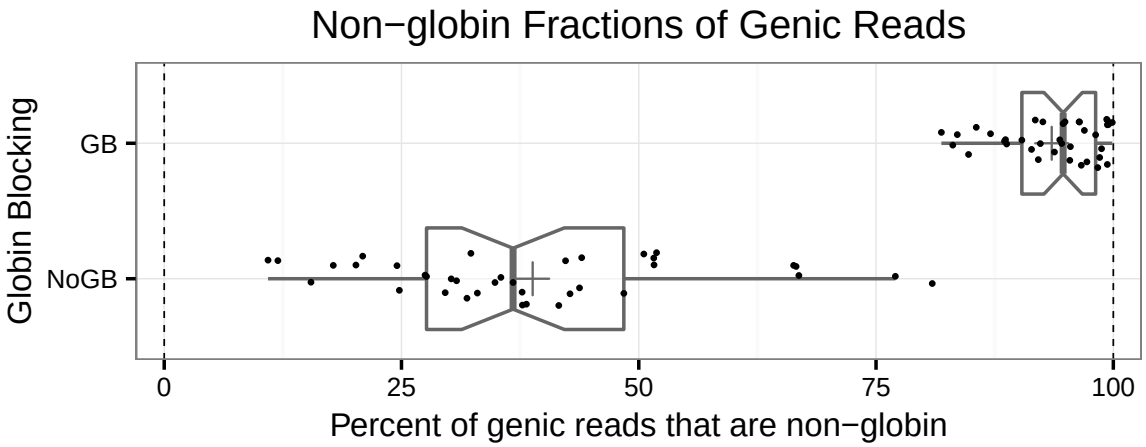


Figure 4.1: **Fraction of genic reads in each sample aligned to non-globin genes, with and without GB.** All reads in each sequencing library were aligned to the cyno genome, and the number of reads uniquely aligning to each gene was counted. For each sample, counts were summed separately for all globin genes and for the remainder of the genes (non-globin genes), and the fraction of genic reads aligned to non-globin genes was computed. Each point represents an individual sample. Gray + signs indicate the means for globin-blocked libraries and unblocked libraries. The overall distribution for each group is represented as a notched box plot. Points are randomly spread vertically to avoid excessive overlapping.

for GB samples is shifted to the right relative to the non-GB signal peak. When all the samples are normalized together, this difference is normalized out, lining up the signal peaks, and this reveals that, as expected, the noise floor for the GB samples is about 2-fold lower. This greater separation between signal and noise peaks in the GB samples means that low-expression genes should be more easily detected and more precisely quantified than in the non-GB samples.

Based on these distributions, we selected a detection threshold of -1, which is approximately the leftmost edge of the trough between the signal and noise peaks. This represents the most liberal possible detection threshold that doesn't call substantial numbers of noise genes as detected. Among the full dataset, 13429 genes were detected at this threshold, and 22276 were not. When considering the GB libraries and non-GB libraries separately and re-computing normalization factors independently within each group, 14535 genes were detected in the GB libraries while only 12460 were detected in the non-GB libraries. Thus, GB allowed the detection of 2000 extra

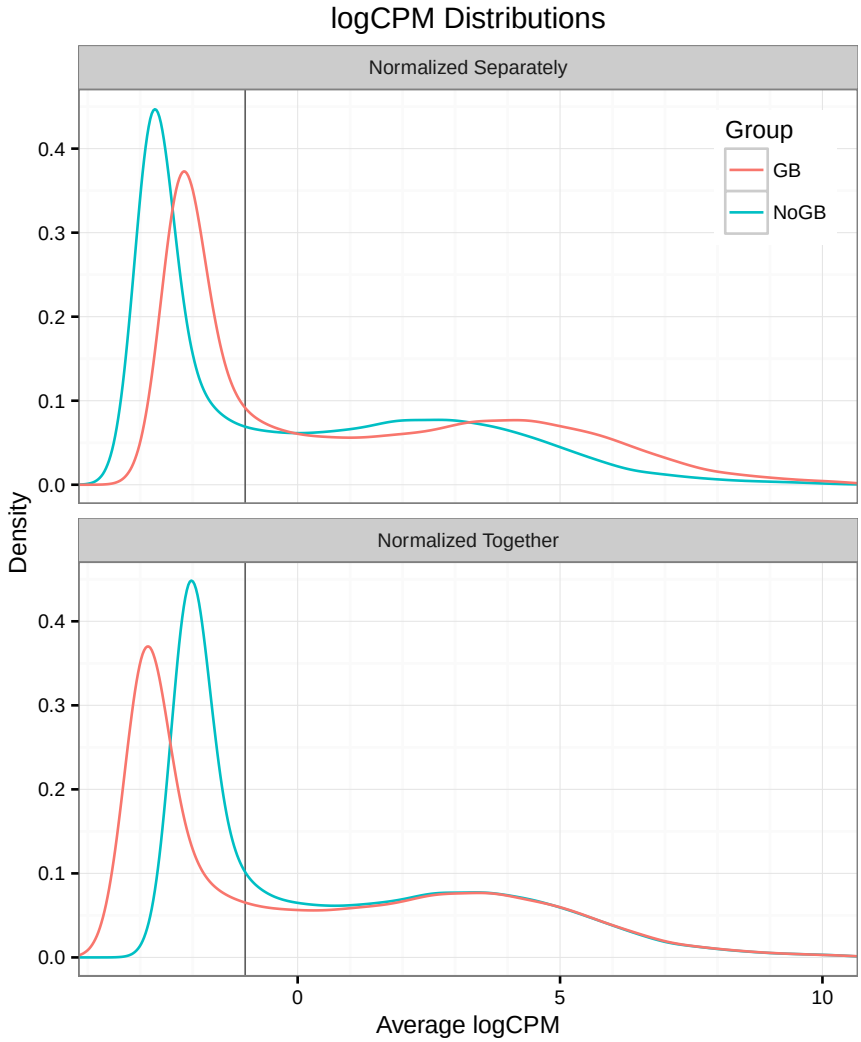


Figure 4.2: **Distributions of average group gene abundances when normalized separately or together.** All reads in each sequencing library were aligned to the cyno genome, and the number of reads uniquely aligning to each gene was counted. Genes with zero counts in all libraries were discarded. Libraries were normalized using the TMM method. Libraries were split into GB and non-GB groups and the average logCPM was computed. The distribution of average gene logCPM values was plotted for both groups using a kernel density plot to approximate a continuous distribution. The GB logCPM distributions are marked in red, non-GB in blue. The black vertical line denotes the chosen detection threshold of -1. Top panel: Libraries were split into GB and non-GB groups first and normalized separately. Bottom panel: Libraries were all normalized together first and then split into groups.

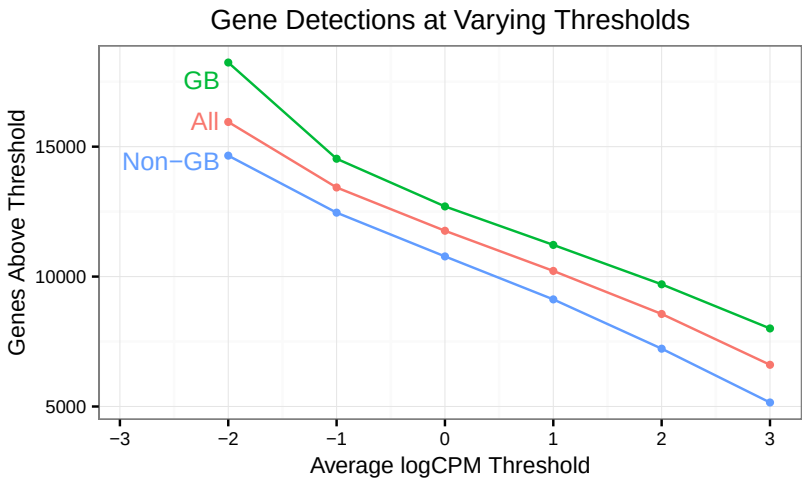


Figure 4.3: **Gene detections as a function of abundance thresholds in GB and non-GB samples.** Average logCPM was computed by separate group normalization as described in Figure 4.2 for both the GB and non-GB groups, as well as for all samples considered as one large group. For each every integer threshold from -2 to 3, the number of genes detected at or above that logCPM threshold was plotted for each group.

genes that were buried under the noise floor without GB. This pattern of at least 2000 additional genes detected with GB was also consistent across a wide range of possible detection thresholds, from -2 to 3 (see Figure 4.3).

4.4.3 Globin blocking does not add significant additional noise or decrease sample quality

One potential worry is that the GB protocol could perturb the levels of non-globin genes. There are two kinds of possible perturbations: systematic and random. The former is not a major concern for detection of differential expression, since a 2-fold change in every sample has no effect on the relative fold change between samples. In contrast, random perturbations would increase the noise and obscure the signal in the dataset, reducing the capacity to detect differential expression.

The data do indeed show small systematic perturbations in gene levels (Figure 4.4). Other than the 3 designated alpha and beta globin genes, two other genes

stand out as having especially large negative $\log_2$ fold changes (logFCs): HBD and LOC1021365. HBD, delta globin, is most likely targeted by the blocking oligos due to high sequence homology with the other globin genes. LOC1021365 is the aforementioned ncRNA that is reverse-complementary to one of the alpha-like genes and that would be expected to be removed during the GB step. All other genes appear in a cluster centered vertically at 0, and the vast majority of genes in this cluster show an absolute logFC of 0.5 or less. Nevertheless, many of these small perturbations are still statistically significant, indicating that the GB oligos likely cause very small but non-zero systematic perturbations in measured gene expression levels.

To evaluate the possibility of GB causing random perturbations and reducing sample quality, we computed the Pearson correlation between logCPM values for every pair of samples with and without GB and plotted them against each other (Figure 4.5). The plot indicated that the GB libraries have higher sample-to-sample correlations than the non-GB libraries. Parametric and nonparametric tests for differences between the correlations with and without GB both confirmed that this difference was highly significant (2-sided paired t-test: $t = 37.2$, $d.f. = 665$, $P \ll 2.2 \times 10^{-16}$; 2-sided Wilcoxon sign-rank test: $V = 2195$, $P \ll 2.2 \times 10^{-16}$). Performing the same tests on the Spearman correlations gave the same conclusion (t-test: $t = 26.8$, $d.f. = 665$, $P \ll 2.2 \times 10^{-16}$; sign-rank test: $V = 8781$, $P \ll 2.2 \times 10^{-16}$). The `edgeR` package was used to compute the overall biological coefficient of variation (BCV) for GB and non-GB libraries, and found that GB resulted in a negligible increase in the BCV (0.417 with GB vs. 0.400 without). The near equality of the BCV for both sets indicates that the higher correlations in the GB libraries are most likely a result of the increased yield of useful reads, which reduces the contribution of Poisson counting uncertainty to the overall variance of the logCPM values [28]. This improves the precision of expression measurements and more than offsets the negligible increase in BCV.

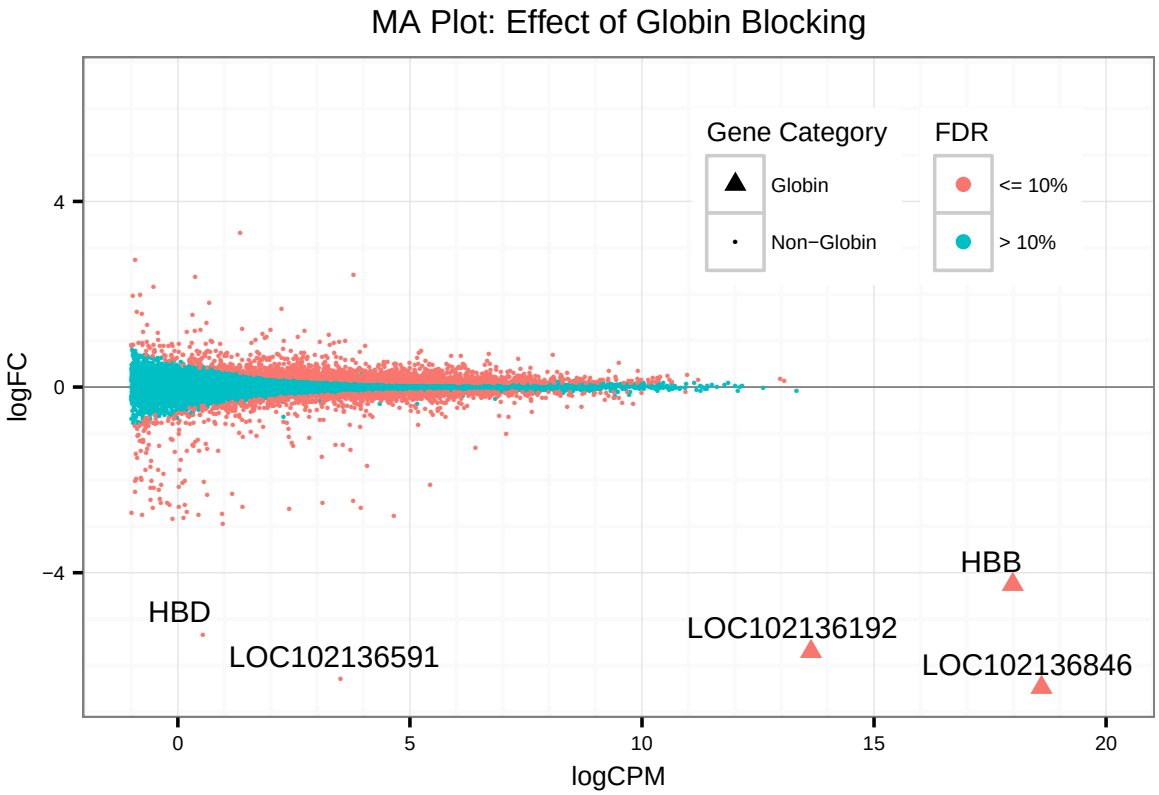


Figure 4.4: **MA plot showing effects of GB on each gene's abundance.** All libraries were normalized together as described in Figure 4.2, and genes with an average logCPM below -1 were filtered out. Each remaining gene was tested for differential abundance with respect to GB using `edgeR`'s quasi-likelihood F-test, fitting a NB GLM to table of read counts in each library. For each gene, `edgeR` reported average logCPM, logFC, p-value, and BH-adjusted FDR. Each gene's logFC was plotted against its logCPM, colored by FDR. Red points are significant at $\leq 10\%$ FDR, and blue are not significant at that threshold. The alpha and beta globin genes targeted for blocking are marked with large triangles, while all other genes are represented as small points.

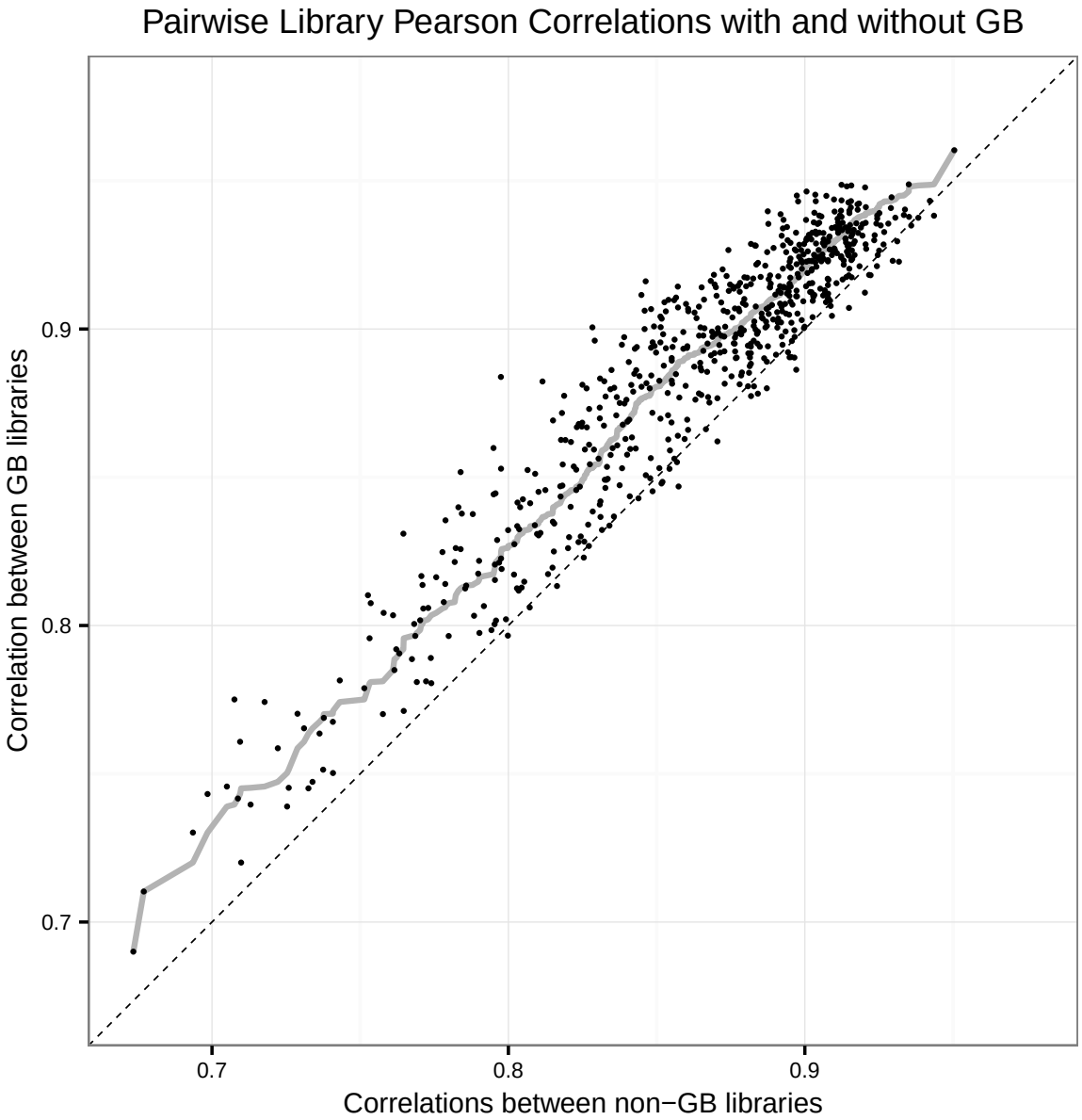


Figure 4.5: **Comparison of inter-sample gene abundance correlations with and without GB.** All libraries were normalized together as described in Figure 4.2, and genes with an average logCPM less than -1 were filtered out. Each gene's logCPM was computed in each library using `edgeR`'s `cpm` function. For each pair of biological samples, the Pearson correlation between those samples' GB libraries was plotted against the correlation between the same samples' non-GB libraries. Each point represents an unique pair of samples. The solid gray line shows a quantile-quantile plot of the distribution of inter-sample correlations with GB vs. without GB. The thin dashed line is the identity line, provided for reference.

		No Globin Blocking		
		Up	**NS**	**Down**
Globin-Blocking	**Up**	231	515	2
	NS	160	11235	136
	Down	0	548	127

Table 4.2: **Comparison of significantly differentially expressed genes with and without globin blocking.** Up, Down: Genes significantly up/down-regulated in post-transplant samples relative to pre-transplant samples, with a false discovery rate of 10% or less. NS: Non-significant genes (false discovery rate greater than 10%).

4.4.4 More differentially expressed genes are detected with globin blocking

To compare performance on differential gene expression tests, we took subsets of both the GB and non-GB libraries with exactly one pre-transplant and one post-transplant sample for each animal that had paired samples available for analysis ($N = 7$ animals, $N = 14$ samples in each subset). The same test for pre- vs. post-transplant differential gene expression was performed on the same 7 pairs of samples from GB libraries and non-GB libraries, in each case using an FDR of 10% as the threshold of significance. Out of 12,954 genes that passed the detection threshold in both subsets, 358 were called significantly differentially expressed in the same direction in both sets; 1063 were differentially expressed in the GB set only; 296 were differentially expressed in the non-GB set only; 2 genes were called significantly up in the GB set but significantly down in the non-GB set; and the remaining 11,235 were not called differentially expressed in either set. These data are summarized in Table 4.2. The differences in BCV calculated by `edgeR` for these subsets of samples were negligible (BCV = 0.302 for GB and 0.297 for non-GB).

The key point is that the GB data results in substantially more differentially expressed calls than the non-GB data. Since there is no gold standard for this dataset, it is impossible to be certain whether this is due to under-calling of differential expression in the non-GB samples or over-calling in the GB samples. However, given

that both datasets are derived from the same biological samples and have nearly equal BCVs, it is more likely that the larger number of differential expression calls in the GB samples are genuine detections that were enabled by the higher sequencing depth and measurement precision of the GB samples. Note that the same set of genes was considered in both subsets, so the larger number of differentially expressed gene calls in the GB data set reflects a greater sensitivity to detect significant differential gene expression and not simply the larger total number of detected genes in GB samples described earlier.

4.5 Discussion

The original experience with whole blood gene expression profiling on DNA microarrays demonstrated that the high concentration of globin transcripts reduced the sensitivity to detect genes with relatively low expression levels, in effect, significantly reducing the sensitivity. To address this limitation, commercial protocols for globin reduction were developed based on strategies to block globin transcript amplification during labeling or physically removing globin transcripts by affinity bead methods [88]. More recently, using the latest generation of labeling protocols and arrays, it was determined that globin reduction was no longer necessary to obtain sufficient sensitivity to detect differential transcript expression [94]. However, we are not aware of any publications using these currently available protocols with the latest generation of microarrays that actually compare the detection sensitivity with and without globin reduction. However, in practice this has now been adopted generally primarily driven by concerns for cost control. The main objective of our work was to directly test the impact of globin gene transcripts and a new GB protocol for application to the newest generation of differential gene expression profiling determined using next generation sequencing.

The challenge of doing global gene expression profiling in cynomolgus monkeys is that the current available arrays were never designed to comprehensively cover this genome and have not been updated since the first assemblies of the cynomolgus genome were published. Therefore, we determined that the best strategy for peripheral blood profiling was to perform deep RNA-seq and inform the workflow using the latest available genome assembly and annotation [93]. However, it was not immediately clear whether globin reduction was necessary for RNA-seq or how much improvement in efficiency or sensitivity to detect differential gene expression would be achieved for the added cost and effort.

Existing strategies for globin reduction involve degradation or physical removal of globin transcripts in a separate step prior to reverse transcription [89, 90, 91]. This additional step adds significant time, complexity, and cost to sample preparation. Faced with the need to perform RNA-seq on large numbers of blood samples we sought a solution to globin reduction that could be achieved purely by adding additional reagents during the reverse transcription reaction. Furthermore, we needed a globin reduction method specific to cynomolgus globin sequences that would work an organism for which no kit is available off the shelf.

As mentioned above, the addition of GB oligos has a very small impact on measured expression levels of gene expression. However, this is a non-issue for the purposes of differential expression testing, since a systematic change in a gene in all samples does not affect relative expression levels between samples. However, we must acknowledge that simple comparisons of gene expression data obtained by GB and non-GB protocols are not possible without additional normalization.

More importantly, GB not only nearly doubles the yield of usable reads, it also increases inter-sample correlation and sensitivity to detect differential gene expression relative to the same set of samples profiled without GB. In addition, GB does not add a significant amount of random noise to the data. GB thus represents a cost-

effective and low-effort way to squeeze more data and statistical power out of the same blood samples and the same amount of sequencing. In conclusion, GB greatly increases the yield of useful RNA-seq reads mapping to the rest of the genome, with minimal perturbations in the relative levels of non-globin genes. Based on these results, globin transcript reduction using sequence-specific, complementary blocking oligos is recommended for all deep RNA-seq of cynomolgus and other nonhuman primate blood samples.

4.6 Future Directions

One drawback of the GB method presented in this analysis is a poor yield of genic reads, only around 50%. In a separate experiment, the reagent mixture was modified so as to address this drawback, resulting in a method that produces an even better reduction in globin reads without reducing the overall fraction of genic reads. However, the data showing this improvement consists of only a few test samples, so the larger data set analyzed above was chosen in order to demonstrate the effectiveness of the method in reducing globin reads while preserving the biological signal.

The motivation for developing a fast practical way to enrich for non-globin reads in cyno blood samples was to enable a large-scale RNA-seq experiment investigating the effects of mesenchymal stem cell infusion on blood gene expression in cynomologus transplant recipients in a time course after transplantation. With the GB method in place, the way is now clear for this experiment to proceed.

Chapter 5

Conclusions

In this work, I have presented a wide range of applications for high-thoughput genomic and epigenomic assays based on sequencing and arrays in the context of immunology and transplant rejection. Chapter 2 described the use of high-throughput RNA sequencing (RNA-seq) and chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq) to investigate the interplay between promoter histone marks and gene expression during activation of naïve and memory $CD4^+$ T-cells. Chapter 3 explored the use of expression microarrays and methylation arrays for diagnosing transplant rejection. Chapter 4 introduced a new RNA-seq protocol for sequencing blood samples from cynomolgus monkeys designed to expedite gene expression profiling in serial blood samples from monkeys who received an experimental treatment for transplant rejection based on mesenchymal stem cells (MSCs). These applications range from basic science to translational medicine, but in all cases, high-thoughput genomic assays were central to the results.

5.1 Every high-throughput analysis presents unique analysis challenges

In addition, each of these applications of high-throughput genomic assays presented unique analysis challenges that could not be solved simply by stringing together standard off-the-shelf methods into a straightforward analysis pipeline. In every case, a bespoke analysis workflow tailored to the data was required, and in no case was it possible to determine every step in the workflow fully prior to seeing the data. For example, exploratory data analysis of the $CD4^+$ T-cell RNA-seq data uncovered the batch effect, and the analysis was adjusted to compensate for it. Similarly, analysis of the ChIP-seq data required choosing an "effective promoter radius" based on the data itself, and several different peak callers were tested before the correct choice became clear. In the development of custom frozen Robust Multichip Average (fRMA) vectors, an appropriate batch size had to be chosen based on the properties of the training data. In the analysis of methylation array data, the appropriate analysis strategy was not obvious and was determined by trying several plausible strategies and inspecting the model paramters afterward to determine which strategy appeared to best capture the observed properties of the data and which strategies appeared to have systematic errors as a result of failing to capture those properties. The globin blocking (GB) protocol went through several rounds of testing before satisfactory performance was achieved, and as mentioned, optimization of the protocol has continued past the version described here. These are only a few examples out of many instances of analysis decisions motivated by the properties of the data.

5.2 Successful data analysis requires a toolbox, not a pipeline

Multiple times throughout this work, I have attempted to construct standard, reusable, pipelines for analysis of specific kinds of data, such as RNA-seq or ChIP-seq. Each time, the very next data set containing this data broke one or more of the assumptions I had built into the pipeline, such as an RNA-seq dataset where some samples aligned to the sense strand while others aligned to the antisense strand, or the discovery that the effective promoter radius varies by histone mark. Each violation of an assumption required a significant rewrite of the pipeline's code in order to accommodate the new aspect of the data. The prospect of reusability turned out to be a pipe(line) dream. After several attempts to extend my pipelines to be general enough to handle an ever-increasing variety of data idiosyncrasies, I realized that it was actually *less* work to reimplement an analysis workflow from scratch each time rather than try to adapt an existing workflow that was originally designed for a different data set.

Once I embraced the idea of writing a bespoke analysis workflow for every data set instead of a one-size-fits-all pipeline, I stopped thinking of the pipeline as the atomic unit of analysis. Instead, I focused on developing an understanding of the component parts of each pipeline, which problems each part solves, and what assumptions it makes, so that when I was presented with a new data set, I could quickly select the appropriate analysis methods for that data set and compose them into a new workflow to answer the demands of a new data set. In cases where no off-the-shelf method existed to address a specific aspect of the data, knowing about a wide range of analysis methods allowed me to select the one that was closest to what I needed and adapt it accordingly, even if it was not originally designed to handle the kind of data I was analyzing. For example, when analyzing heteroskedastic methylation array data, I adapted the `voom` method from `limma`, which was originally designed to model

heteroskedasticity in RNA-seqdata [23]. While `voom` was designed to accept read counts, I determined that this was not a fundamental assumption of the method but rather a limitation of the specific implementation, and I was able to craft a modified implementation that accepted $\log_2$ ratios (M-values) from methylation arrays. In contrast, adapting another method such as `edgeR` for methylation arrays would not be possible, since many steps of the `edgeR` workflow, from normalization to dispersion estimation to model fitting, assume that the input is given on the scale of raw counts and take full advantage of this assumption [40, 29, 28, 27]. In short, I collected a "toolbox" full of useful modular analysis methods and developed the knowledge of when and where each could be applied, as well as how to compose them on demand into pipelines for specific data sets. This prepared me to handle the idiosyncrasies of any new data set, even when the new data has problems that I have not previously encountered in any other data set.

Reusable pipelines have their place, but that place is in automating established processes, not researching new science. For example, the custom fRMA vectors developed in Chapter 3, are being incorporated into an automated pipeline for diagnosing transplant rejection using biopsy and blood samples from transplant recipients. Once ready, this diagnostic method will consist of normalization using the pre-trained fRMA vectors, followed by classification of the sample by a pre-trained classifier, which outputs a posterior probability of acute rejection. This is a perfect use case for a proper pipeline: repeating the exact same sequence of analysis steps many times. The input to the pipeline is sufficiently well-controlled that we can guarantee it will satisfy the assumptions of the pipeline. But research data is not so well-controlled, so when analyzing data in a research context, the analysis must conform to the data, rather than trying to force the data to conform to a preferred analysis strategy. That means having a toolbox full of composable methods ready to respond to the observed properties of the data.

Bibliography

[1] Nicole M. Valenzuela and Elaine F. Reed. "Antibody-Mediated Rejection across Solid Organ Transplants: Manifestations, Mechanisms, and Therapies". In: *Journal of Clinical Investigation* 127.7 (June 12, 2017), pp. 2492–2504. ISSN: 0021-9738. DOI: 10/gbmrzf.

[2] Kenneth Murphy et al. *Janeway's Immunobiology*. 8th. OCLC: 733935898. New York: Garland Science, 2012. ISBN: 978-0-8153-4243-4 978-0-8153-4530-5.

[3] Richard Kowalski et al. "Immune Cell Function Testing: An Adjunct to Therapeutic Drug Monitoring in Transplant Patient Management". In: *Clinical Transplantation* 17.2 (2003), pp. 77–88. ISSN: 09020063. DOI: 10/btbdwp.

[4] Moshe Israeli et al. "Preceeding the Rejection: In Search for a Comprehensive Post-Transplant Immune Monitoring Platform". In: *Transplant Immunology* 18.1 (July 1, 2007), pp. 7–12. ISSN: 0966-3274. DOI: 10/dtdsbs.

[5] S M Kurian et al. "Molecular Classifiers for Acute Kidney Transplant Rejection in Peripheral Blood by Whole Genome Gene Expression Profiling". In: *American Journal of Transplantation* 14.5 (May 2014), pp. 1164–1172. ISSN: 16006135. DOI: 10/f5xswg.

[6] Daniel R Salomon. "Protocol Biopsies Should Be Part of the Routine Management of Kidney Transplant Recipients". In: *American Journal of Kidney Diseases* 40.4 (Oct. 1, 2002), pp. 674–677. ISSN: 0272-6386. DOI: 10/bjm7nv.

[7] Alan Wilkinson. "Protocol Transplant Biopsies: Are They Really Needed?" In: *Clinical journal of the American Society of Nephrology : CJASN* 1.1 (2006), pp. 130–137. ISSN: 1555905X. DOI: 10/cq7cjq.

[8] J Patel et al. "Determining the Utility of Protocol Biopsies in Kidney Transplant Recipients [Abstract]." In: *Poster Session D: Kidney: Acute Cellular Rejection*. 2018 American Transplant Congress. Seattle, WA, June 5, 2018. URL: https://atcmeetingabstracts.com/abstract/determining-the-utility-of-protocol-biopsies-in-kidney-transplant-recipients/.

[9] Mareena S. Zachariah et al. "Utility of Serial Protocol Biopsies Performed after 1 Year in Predicting Long-Term Kidney Allograft Function According to Histologic Phenotype". In: *Experimental and Clinical Transplantation* 16.4 (2018), pp. 391–400. ISSN: 13040855. DOI: 10/ggcxm6.

[10] Paul R. Rogers, Caroline Dubey, and Susan L. Swain. "Qualitative Changes Accompany Memory T Cell Generation: Faster, More Effective Responses at Lower Doses of Antigen". In: *The Journal of Immunology* 164.5 (2000), pp. 2338–2346. ISSN: 0022-1767. DOI: 10/gf4d35.

[11] Cheryl A. London, Michael P. Lodge, and Abul K. Abbas. "Functional Responses and Costimulator Dependence of Memory CD4 + T Cells". In: *The Journal of Immunology* 164.1 (2000), pp. 265–272. ISSN: 0022-1767. DOI: 10/ggcxkw.

[12] Marion Berard and David F. Tough. "Qualitative Differences between Naïve and Memory T Cells". In: *Immunology* 106.2 (2002), pp. 127–138. ISSN: 00192805. DOI: 10/fc8shc.

[13] K Le Blanc. "Immunomodulatory Effects of Fetal and Adult Mesenchymal Stem Cells". In: *Cytotherapy* 5.6 (Dec. 1, 2003), pp. 485–489. ISSN: 1465-3249. DOI: 10/dmxgm2.

[14] Sudeepta Aggarwal and Mark F. Pittenger. "Human Mesenchymal Stem Cells Modulate Allogeneic Immune Cell Responses". In: *Blood* 105.4 (Feb. 15, 2005), pp. 1815–1822. ISSN: 00064971. DOI: 10/fnb37s.

[15] Amelia Bartholomew et al. "Mesenchymal Stem Cells in the Induction of Transplantation Tolerance." In: *Transplantation* 87 (9 Suppl May 2009), S55–S57. ISSN: 15346080. DOI: 10/cf45q7.

[16] Dora M. Berman et al. "Mesenchymal Stem Cells Enhance Allogeneic Islet Engraftment in Nonhuman Primates". In: *Diabetes* 59.10 (2010), pp. 2558–2568. ISSN: 00121797. DOI: 10/c9r6nn.

[17] James A. Ankrum, Joon Faii Ong, and Jeffrey M. Karp. "Mesenchymal Stem Cells: Immune Evasive, Not Immune Privileged". In: *Nature Biotechnology* 32.3 (2014), pp. 252–260. ISSN: 15461696. DOI: 10/f5vjkk.

[18] Alix K. Berglund et al. "Immunoprivileged No More: Measuring the Immunogenicity of Allogeneic Adult Mesenchymal Stem Cells". In: *Stem Cell Research and Therapy* 8.1 (Dec. 22, 2017), p. 288. ISSN: 17576512. DOI: 10/ggcxjz.

[19] Manas K. Majumdar et al. "Characterization and Functionality of Cell Surface Molecules on Human Mesenchymal Stem Cells". In: *Journal of Biomedical Science* 10.2 (2003), pp. 228–241. ISSN: 10217770. DOI: 10/b7sw4z.

[20] J. M. Ryan et al. "Interferon-γ Does Not Break, but Promotes the Immunosuppressive Capacity of Adult Human Mesenchymal Stem Cells". In: *Clinical and Experimental Immunology* 149.2 (Aug. 2007), pp. 353–363. ISSN: 00099104. DOI: 10/b34t6x.

[21] John M. Chambers and Trevor J. Hastie, eds. *Statistical Models in S*. 1st ed. Routledge, 1992. ISBN: 978-0-203-73853-5. DOI: 10.1201/9780203738535.

[22] Gordon K Smyth. "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments". In: *Statistical Applications in Genetics and Molecular Biology* 3.1 (Jan. 12, 2004), pp. 1–25. ISSN: 1544-6115. DOI: 10/ddqzg9.

[23] Charity W. Law et al. "Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts". In: *Genome Biology* 15.2 (2014), R29. ISSN: 1465-6906. DOI: 10/gfghcz.

[24] Matthew E. Ritchie et al. "Empirical Array Quality Weights in the Analysis of Microarray Data". In: *BMC Bioinformatics* 7 (2006). ISSN: 14712105. DOI: 10/dxrnmz.

[25] Ruijie Liu et al. "Why Weight? Modelling Sample and Observational Level Variability Improves Power in RNA-Seq Analyses". In: *Nucleic Acids Research* 43.15 (Sept. 3, 2015), e97–e97. ISSN: 13624962. DOI: 10/f7rq7c.

[26] Gordon K Smyth, Joëlle Michaud, and Hamish S Scott. "Use of Within-Array Replicate Spots for Assessing Differential Expression in Microarray Experiments." In: *Bioinformatics (Oxford, England)* 21.9 (May 1, 2005), pp. 2067–75. ISSN: 1367-4803. DOI: 10/ffmjnv.

[27] Yunshun Chen, Aaron T. L. Lun, and Gordon K. Smyth. "Differential Expression Analysis of Complex RNA-Seq Experiments Using edgeR". In: *Statistical Analysis of Next Generation Sequencing Data*. Ed. by Somnath Datta and Dan Nettleton. Cham: Springer International Publishing, 2014, pp. 51–74. ISBN: 978-3-319-07212-8. DOI: 10.1007/978-3-319-07212-8_3.

[28] Davis J McCarthy, Yunshun Chen, and Gordon K Smyth. "Differential Expression Analysis of Multifactor RNA-Seq Experiments with Respect to Biological Variation." In: *Nucleic acids research* 40.10 (May 2012), pp. 4288–97. ISSN: 1362-4962. DOI: 10/fxwbrf.

[29] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. "edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." In: *Bioinformatics (Oxford, England)* 26.1 (Jan. 1, 2010), pp. 139–40. ISSN: 1367-4811. DOI: 10/drxgw2.

[30] Steven P Lund et al. "Detecting Differential Expression in RNA-Sequence Data Using Quasi-Likelihood with Shrunken Dispersion Estimates." In: *Statistical applications in genetics and molecular biology* 11.5 (Jan. 2012). ISSN: 1544-6115. DOI: 10/f95zdf.

[31] Yong Zhang et al. "Model-Based Analysis of ChIP-Seq (MACS)". In: *Genome Biology* 9.9 (Jan. 2008), R137. ISSN: 1465-6906. DOI: 10/dfst4f.

[32] Chongzhi Zang et al. "A Clustering Approach for Identification of Enriched Domains from Histone Modification ChIP-Seq Data". In: *Bioinformatics* 25.15 (2009), pp. 1952–1958. ISSN: 13674803. DOI: 10/fd9qhm.

[33] Qunhua Li et al. "Measuring Reproducibility of High-Throughput Experiments". In: *The Annals of Applied Statistics* 5.3 (Sept. 2011), pp. 1752–1779. ISSN: 1932-6157. DOI: 10/bwxxjt.

[34] Aaron T.L. Lun and Gordon K. Smyth. "Csaw: A Bioconductor Package for Differential Binding Analysis of ChIP-Seq Data Using Sliding Windows". In: *Nucleic Acids Research* 44.5 (Mar. 18, 2015), e45. ISSN: 13624962. DOI: 10/f8g6nw.

[35] Rafael a Irizarry et al. "Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data." In: *Biostatistics (Oxford, England)* 4.2 (Apr. 2003), pp. 249–64. ISSN: 1465-4644. DOI: 10/b9zc5m.

[36] Matthew N. Mccall, Benjamin M. Bolstad, and Rafael a. Irizarry. "Frozen Robust Multiarray Analysis (fRMA)". In: *Biostatistics (Oxford, England)* 11.2 (Apr. 1, 2010), pp. 242–53. ISSN: 1465-4644. DOI: 10/bxxhbc.

[37] C Li and W Hung Wong. "Model-Based Analysis of Oligonucleotide Arrays: Model Validation, Design Issues and Standard Error Application." In: *Genome biology* 2.8 (Jan. 2001), RESEARCH0032. ISSN: 1474-760X. DOI: 10/dtsm65.

[38] Carl R Pelz et al. "Global Rank-Invariant Set Normalization (GRSN) to Reduce Systematic Distortions in Microarray Data." In: *BMC bioinformatics* 9 (Jan. 2008), p. 520. ISSN: 1471-2105. DOI: 10/dmcdfj.

[39] Stephen R Piccolo et al. "A Single-Sample Microarray Normalization Method to Facilitate Personalized-Medicine Workflows." In: *Genomics* 100.6 (Dec. 2012), pp. 337–44. ISSN: 1089-8646. DOI: 10/f4gwz9.

[40] Mark D Robinson and Alicia Oshlack. "A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data". In: *Genome Biology* 11.3 (Jan. 2010), R25. ISSN: 1465-6906. DOI: 10/cq6f8b.

[41] Simon Anders and Wolfgang Huber. "Differential Expression Analysis for Sequence Count Data". In: *Genome Biology* 11.10 (Oct. 27, 2010), R106. ISSN: 1474-760X. DOI: 10/btmbk5.

[42] W Evan Johnson, Cheng Li, and Ariel Rabinovic. "Adjusting Batch Effects in Microarray Expression Data Using Empirical Bayes Methods". In: *Biostatistics* 8.1 (Jan. 1, 2007), pp. 118–127. ISSN: 1468-4357. DOI: 10/dsf386.

[43] Jeffrey T. Leek and John D. Storey. "Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis". In: *PLoS Genetics* 3.9 (Sept. 2007), pp. 1724–1735. ISSN: 15537390. DOI: 10/c9pc69.

[44] Y Benjamini and Y Hochberg. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society. Series B …* (1995). JSTOR: 10.2307/2346101.

[45] Belinda Phipson. "Empirical Bayes Modelling of Expression Profiles and Their Associations". The Walter, Eliza Hall Institute of Medical Research & The University of Melbourne, 2013. URL: http://hdl.handle.net/11343/38162.

[46] Ryan C. Thompson. *Reproducible Reanalysis of a Combined ChIP-Seq & RNA-Seq Data Set*. La Jolla, CA: The Scripps Research Institute, Aug. 9, 2019. URL: https://github.com/DarwinAwardWinner/CD4-csaw (visited on 11/14/2019).

[47] Sarah Adrianne Hutchison LaMere. "Dynamic Epigenetic Regulation of CD4 T Cell Activation and Memory Formation". The Scripps Research Institute, 2015. 371 pp.

[48] S. A. LaMere et al. "Promoter H3K4 Methylation Dynamically Reinforces Activation-Induced Pathways in Human CD4 T Cells". In: *Genes & Immunity* 17.5 (July 12, 2016), pp. 283–297. ISSN: 1466-4879. DOI: 10/f97x85.

[49] Sarah A. LaMere et al. "H3K27 Methylation Dynamics during CD4 T Cell Activation: Regulation of JAK/STAT and IL12RB2 Expression by JMJD3". In: *The Journal of Immunology* 199.9 (Nov. 1, 2017), pp. 3158–3175. ISSN: 0022-1767. DOI: 10/gchc9x.

[50] Rasko Leinonen, Hideaki Sugawara, and Martin Shumway. "The Sequence Read Archive". In: *Nucleic Acids Research* 39 (SUPPL. 1 2011), pp. 2010–2012. ISSN: 03051048. DOI: 10/c652z5.

[51] Alexander Dobin et al. "STAR: Ultrafast Universal RNA-Seq Aligner". In: *Bioinformatics* 29.1 (Jan. 2013), pp. 15–21. ISSN: 1460-2059, 1367-4803. DOI: 10/f4h523.

[52] Daehwan Kim et al. "Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-Genotype". In: *Nature Biotechnology* 37.8 (2019), pp. 907–915. ISSN: 1087-0156. DOI: 10/gf5395.

[53] Yang Liao, Gordon K. Smyth, and Wei Shi. "FeatureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features". In: *Bioinformatics* 30.7 (2014), pp. 923–930. ISSN: 14602059. DOI: 10/f5w7rp.

[54] Harold J Pimentel et al. "Differential Analysis of RNA-Seq Incorporating Quantification Uncertainty". In: *bioRxiv* (2016), p. 058164. DOI: 10/gfn5bn.

[55] Rob Patro et al. "Salmon Provides Fast and Bias-Aware Quantification of Transcript Expression". In: *Nature Methods* 14.4 (Apr. 6, 2017), pp. 417–419. ISSN: 1548-7091. DOI: 10/gcw9f5.

[56] *Shoal: Improved Multi-Sample Transcript Abundance Estimates Using Adaptive Priors*. COMBINE-lab, Nov. 28, 2018. URL: https://github.com/COMBINE-lab/shoal (visited on 11/14/2019).

[57] Ryan C. Thompson. *Workflow to Download/Generate Various Mapping Indices for the Human Hg38 Genome*. La Jolla, CA: The Scripps Research Institute, Jan. 20, 2018. URL: https://github.com/DarwinAwardWinner/hg38-ref (visited on 11/14/2019).

[58] Daniel R. Zerbino et al. "Ensembl 2018". In: *Nucleic Acids Research* 46.D1 (Jan. 4, 2018), pp. D754–D761. ISSN: 13624962. DOI: 10/gcwg6r.

[59] Jennifer Harrow et al. "GENCODE: The Reference Human Genome Annotation for the ENCODE Project". In: *Genome Research* 22.9 (2012), pp. 1760–1774. ISSN: 10889051. DOI: 10/f4w5m5.

[60] Gordon K Smyth. "Limma : Linear Models for Microarray Data". In: *Bioinformatics* pages (2005), pp. 397–420. ISSN: 00199567. DOI: 10/dv8chk.

[61] Belinda Phipson et al. "Robust Hyperparameter Estimation Protects against Hypervariable Genes and Improves Power to Detect Differential Expression". In: *Annals of Applied Statistics* 10.2 (June 2016), pp. 946–963. ISSN: 19417330. DOI: 10/gfgp3f. arXiv: 1602.08678.

[62] Ben Langmead and Steven L Salzberg. "Fast Gapped-Read Alignment with Bowtie 2." In: *Nature methods* 9.4 (Apr. 2012), pp. 357–9. ISSN: 1548-7105. DOI: 10/gd2xzn.

[63] Valerie A. Schneider et al. "Evaluation of GRCh38 and de Novo Haploid Genome Assemblies Demonstrates the Enduring Quality of the Reference Assembly". In: *Genome Research* 27.5 (2017), pp. 849–864. ISSN: 15495469. DOI: 10/f92cmg.

[64] Ian Dunham et al. "An Integrated Encyclopedia of DNA Elements in the Human Genome". In: *Nature* 489.7414 (2012), pp. 57–74. ISSN: 14764687. DOI: 10/bg9d.

[65] Haley M. Amemiya, Anshul Kundaje, and Alan P. Boyle. "The ENCODE Blacklist: Identification of Problematic Regions of the Genome". In: *Scientific Reports* 9.1 (2019), pp. 1–5. ISSN: 20452322. DOI: 10/gf4jsb.

[66] Peter V. Kharchenko, Michael Y. Tolstorukov, and Peter J. Park. "Design and Analysis of ChIP-Seq Experiments for DNA-Binding Proteins". In: *Nature Biotechnology* 26.12 (Dec. 2008), pp. 1351–1359. ISSN: 1546-1696. DOI: 10/d2rbh7.

[67] Endre Bakken Stovner. *Epic: Diffuse Domain ChIP-Seq Caller Based on SICER*. BioCore, NTNU, July 20, 2019. URL: https://github.com/biocore-ntnu/epic (visited on 11/14/2019).

[68] Nathan Boley. *Irreproducible Discovery Rate (IDR)*. Oct. 15, 2019. URL: https://github.com/nboley/idr (visited on 11/14/2019).

[69] Aaron T.L. Lun and Gordon K Smyth. "De Novo Detection of Differentially Bound Regions for ChIP-Seq Data Using Peaks and Windows: Controlling Error Rates Correctly". In: *Nucleic Acids Research* 42.11 (June 17, 2014), e95–e95. ISSN: 0305-1048. DOI: 10/f5874p.

[70] Jeffrey T Leek. "Svaseq: Removing Batch Effects and Other Unwanted Noise from Sequencing Data." In: *Nucleic acids research* 42.21 (Dec. 1, 2014), pp. 0-11. ISSN: 1362-4962. DOI: 10/f8k8kf.

[71] Ricard Argelaguet et al. "Multi-Omics Factor Analysis—a Framework for Unsupervised Integration of Multi-omics Data Sets". In: *Molecular Systems Biology* 14.6 (June 20, 2018), pp. 1–13. ISSN: 1744-4292. DOI: 10/gdqq3f.

[72] Matthew D. Young et al. "ChIP-Seq Analysis Reveals Distinct H3K27me3 Profiles That Correlate with Transcriptional Activity". In: *Nucleic Acids Research* 39.17 (Sept. 2011), pp. 7415–7427. ISSN: 1362-4962. DOI: 10/cp6h62.

[73] Johannes Köster and Sven Rahmann. "Snakemake-a Scalable Bioinformatics Workflow Engine". In: *Bioinformatics* 28.19 (2012), pp. 2520–2522. ISSN: 13674803. DOI: 10/gd2xzq.

[74] Chunyuan Jin and Gary Felsenfeld. "Nucleosome Stability Mediated by Histone Variants H3.3 and H2A.Z". In: *Genes and Development* 21.12 (2007), pp. 1519–1529. ISSN: 08909369. DOI: 10/bwwqb3.

[75] Chunyuan Jin et al. "H3.3/H2A.Z Double Variant-Containing Nucleosomes Mark 'nucleosome-Free Regions' of Active Promoters and Other Regulatory Regions". In: *Nature Genetics* 41.8 (Aug. 26, 2009), pp. 941–945. ISSN: 10614036. DOI: 10/bqfzgv.

[76] Robert Gentleman et al., eds. *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Red. by Wing Wong et al. Statistics for Biology and Health. New York, NY: Springer New York, 2005. ISBN: 978-0-387-25146-2 978-0-387-29362-2. DOI: 10.1007/0-387-29362-0. (Visited on 11/15/2019).

[77] Matthew N McCall and Rafael A Irizarry. "Thawing Frozen Robust Multi-Array Analysis (fRMA)". In: *BMC Bioinformatics* 12.1 (Dec. 16, 2011), p. 369. ISSN: 1471-2105. DOI: 10/fv34wn.

[78] Robert Tibshirani et al. “Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression.” In: *Proceedings of the National Academy of Sciences of the United States of America* 99.10 (May 14, 2002), pp. 6567–72. ISSN: 0027-8424. DOI: 10/d2h5n3.

[79] Natacha Turck et al. “A Multiparameter Panel Method for Outcome Prediction Following Aneurysmal Subarachnoid Hemorrhage”. In: *Intensive Care Medicine* 36.1 (2010), pp. 107–115. ISSN: 03424642. DOI: 10/d8hzhf.

[80] Jovana Maksimovic, Lavinia Gordon, and Alicia Oshlack. “SWAN: Subset-Quantile within Array Normalization for Illumina Infinium HumanMethylation450 BeadChips.” In: *Genome biology* 13.6 (2012), R44. ISSN: 14656914. DOI: 10/ggcxk2.

[81] Martin J. Aryee et al. “Minfi: A Flexible and Comprehensive Bioconductor Package for the Analysis of Infinium DNA Methylation Microarrays”. In: *Bioinformatics* 30.10 (May 15, 2014), pp. 1363–1369. ISSN: 14602059. DOI: 10/f3m42q.

[82] Matthew E. Ritchie et al. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies”. In: *Nucleic Acids Research* 43.7 (2015), pp. 1–13. ISSN: 13624962. DOI: 10/f7c4n5.

[83] Robert McGill, John W Tukey, and Wayne A Larsen. “Variations of Box Plots”. In: *The American Statistician* 32.1 (1978), pp. 12–16. ISSN: 00031305. DOI: 10/dsvtxr. JSTOR: 2683468.

[84] Michael Volkmar et al. “DNA Methylation Profiling Identifies Epigenetic Dysregulation in Pancreatic Islets from Type 2 Diabetic Patients”. In: *EMBO Journal* 31.6 (2012), pp. 1405–1426. ISSN: 02614189. DOI: 10/fzd945.

[85] Heather Hall et al. “Glucotypes Reveal New Patterns of Glucose Dysregulation”. In: *PLoS Biology* 16.7 (2018), pp. 1–23. ISSN: 15457885. DOI: 10/gdwxrm.

[86] Norihide Yokoi. "Epigenetic Dysregulation in Pancreatic Islets and Pathogenesis of Type 2 Diabetes". In: *Journal of Diabetes Investigation* 9.3 (2018), pp. 475–477. ISSN: 20401124. DOI: 10/ggcxm5.

[87] Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn. "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant". In: *Psychological Science* 22.11 (Nov. 17, 2011), pp. 1359–1366. ISSN: 14679280. DOI: 10/bxbw3c.

[88] ME Winn, MA Zapala, and Iiris Hovatta. "The Effects of Globin on Microarray-Based Gene Expression Analysis of Mouse Blood". In: *Mammalian …* 21.5-6 (June 2010), pp. 268–75. DOI: 10/bjjwht.

[89] Anastasios Mastrokolias et al. "Increased Sensitivity of next Generation Sequencing-Based Expression Profiling after Globin Reduction in Human Blood RNA." In: *BMC genomics* 13.1 (Jan. 2012), p. 28. ISSN: 1471-2164. DOI: 10/fx3s8g.

[90] Igseo Choi et al. "Increasing Gene Discovery and Coverage Using RNA-Seq of Globin RNA Reduced Porcine Blood Samples". In: *BMC Genomics* 15.1 (2014), pp. 1–10. ISSN: 14712164. DOI: 10/gb3g9j.

[91] Heesun Shin et al. "Variation in RNA-Seq Transcriptome Profiles of Peripheral Whole Blood from Healthy Individuals with and without Globin Depletion". In: *PLoS ONE* 9.3 (2014), pp. 1–11. ISSN: 19326203. DOI: 10/ggcxmp.

[92] Ophélie Arnaud et al. "Targeted Reduction of Highly Abundant Transcripts Using Pseudo-Random Primers". In: *BioTechniques* 60.4 (Apr. 1, 2016), pp. 169–74. ISSN: 1940-9818. DOI: 10/ggcxjv.

[93] Richard K Wilson and Wesley Warren. *Macaca Fascicularis (Cynomolgus Macaque) Sequence Assembly*. 2013. URL: http://www.ncbi.nlm.nih.gov/assembly/GCF_000364345.1.

[94] NuGEN. *Performance Verification of the Automated NuGEN Ovation Whole Blood Solution*. 2010. URL: http://www.nugeninc.com/nugen/?LinkServID=89366653-85CF-44AC-80672BBD775B0170.