STATISTICAL METHODS FOR EPIGENETIC DATA AND STRUCTURAL MAGNETIC RESONANCE IMAGING

by

Jean-Philippe Fortin

A dissertation submitted to The Johns Hopkins University in conformity with the requirements for the degree of Doctor of Philosophy.

Baltimore, Maryland

February, 2016

Abstract

This dissertation introduces novel statistical methods for the analysis of high-dimensional biomedical data. First, we present a normalization method for DNA methylation microarray data, called Functional normalization (Funnorm). The methods extends quantile normalization to remove unwanted variation using control probes. Using several cancer datasets from The Cancer Genome Atlas (TCGA), we show that Funnorm improves the replication of the biological findings in cancer studies. Second, we present a between-scan normalization method for structural magnetic resonance imaging (MRI) data. We use voxels that are not associated with the outcome of interest, for instance the cerebrospinal fluid (CSF) voxels in the ventricles, to model the unwanted variation across scans. We show that our method, called Removal of Artificial Voxel Effect by Linear regression (RAVEL), improves the replicability of the voxels associated with Alzheimer's diseases estimated from T1-weighted images. Third, we present a computational method that predicts A/B compartments as revealed by Hi-C data. using long-range correlations in epigenetic data. Analysis of Hi-C data has shown that the genome can be divided into two compartments called

ABSTRACT

A/B compartments. These compartments are cell-type specific and are associated with open and closed chromatin. We show that A/B compartments can reliably be estimated using data from the Illumina 450k DNA methylation microarray, DNase hypersensitivity sequencing, single-cell ATAC sequencing and single-cell whole-genome bisulfite sequencing. Finally, we present shinyMethyl, a Bioconductor package for interactive quality control of DNA methylation data from the Illumina 450k array. shinyMethyl makes it easy to perform quality assessment of large-scale methylation datasets, such as epigenome-wide association studies or the datasets available through TCGA.

Advisor: Kasper D. Hansen, Ph.D.

Thesis Readers: Jeffrey T. Leek, Brion Maher, Russell T. Shinohara and Steven L. Salzberg

Acknowledgments

To Kasper Hansen: Thank you for your good mentoring, generosity and infinite patience. Thank you for teaching me thoroughness, pragmatic skepticism and how to be a finisher, but also that deepskyblue3 and deeppink4 are deep and beautiful colors in R. You gave me the chance to really learn from you by spending hours on reviewing my code and spending days to improve my manuscripts. Thank you for sharing with me your passion for biology, and for teaching me how to never be satisfied with simulations only. Finally, thank you for understanding and accepting my variability in productivity and time constraints during the past four years at Johns Hopkins. This really made a difference.

To Taki Shinohara: Thank you for letting me work with you on those beautiful brain images, for introducing me to the exciting challenges of neuroimaging and for mentoring me in the past year and a half. I really enjoyed being mentored by you and I really hope that the RAVEL work is only the first of the stones of our collaboration!

To Masoud Asgharian, Celia Greenwood and Aurélie Labbe: Thank you for introducing me to the world of statistics and epigenetics, and for mentoring me at McGill

ACKNOWLEDGMENTS

University. I am here because of you.

To my committee members – Jeff Leek, Steven Salzberg and Brion Maher, and to my preliminary oral exam committee members– Christine Ladd-Acosta and Hans Bjornosson: Thank you for your mentoring and your valuable feedback along the road.

To Ciprian Crainiceanu, Brian Caffo, Elana Fertig, Marie-Diener West, Karen Bandeen-Roche, John McGready: Thank you for your mentorship, guidance and knowledge that made me a better researcher and teacher.

To the faculty, staff, postdocs and students in the Department of Biostatistics: a thousand times thank you. There is no other department on earth that would have been better suited for me than Hopkins Biostats. I came to Hopkins Biostats because I had a feel it would be a nice family to be in. And I was right.

To the Hansen lab folks – Leslie Myint, Kipper Fletez-Brant (Kippy) and Pete Hickey: Please keep shaving your heads (except Leslie - you don't have to). I will always miss you.

To Mandy Mejia, Aaron Fisher, Leonardo Collado-Torres, Prasad Patil, Emily Huang, Therri Usher, Vivek Charu: Thank you for everything my friends, Hopkins Biostats would not have been Hopkins Biostats without you.

To Kate Shearer, Shan Andrews and Kelly Bakulski: Thank you for being the best housemates!

To my friends Frazer Matthews and Cyrus Franklin: thank you for always being

there for me when it was not always easy to be there for me! To Sarah Gauvreau-Jean and Lisa Belley: There is not enough ink in the world to thank you. To Anthony Sylvain: You know that you are required to visit me wherever I go next right ?

To Roger Messick: Thank you for everything.

To John Muschelli, Francis Abreu and David Lenis: I should capitalize thanks for you, or exponentiate it. You were essentially my family in Baltimore. Together we have cried, laughed, worked and baked. I will miss you so much.

To Elizabeth Sweeney: Thank you for teaching me "proper English", for literally feeding me with the finest of the cuisines during my comprehensive exams, for making me discover the Indiana State Fair and its little treasures, for introducing me to the world of brain imaging and for helping me feel like home in Baltimore. You were the best partner in crime I could have dreamed of. I will miss you.

À ma famille: Merci pour tout l'amour inconditionnel, que je ressens même a distance. Merci de m'encourager à poursuivre mes rêves et merci de m'accompagner dans mes décisions. Je vous aime.

Contents

Abstract **ii**

Acknowledgments **iv**

List of Tables **viii**

List of Figures **ix**

1 Introduction **1**

2 Normalization of DNA methylation data **7**

2.1 Introduction 7

2.2 Methods 12

2.2.1 Infinium HumanMethylation450 BeadChip 12

2.2.2 Control probe summaries 13

2.2.3 Functional normalization: the general framework 13

2.2.4 Functional normalization for 450k arrays 17

2.2.5 Data 18
2.2.5.1 Data description 18
2.2.5.2 Data processing 21
2.2.6 Comparison to normalization methods 22
2.2.7 Comparison to SVA, RUV and ComBat 24
2.2.8 Identification of differentially methylated positions 27
2.2.9 Discovery-Validation comparisons 27
2.2.10 Sex validation analysis 29
2.2.11 Sample size simulation 29
2.3 Results 30
2.3.1 Control probes may act as surrogates for batch effects 30
2.3.2 Functional normalization 32
2.3.3 Funnorm improves the replication between experiments, even when a batch effect is present 33
2.3.4 Replication between experiments in cancer study (TCGA-KIRC datasets) 37
2.3.5 Funnorm preserves subtype heterogeneity in tumor samples (TCGA-AML datasets) 39
2.3.6 Replication between experiments with small changes 42
2.3.7 Funnorm improves X and Y chromosomes probes prediction in blood samples 42

2.3.8 Funnorm reduces technical variability 44
2.3.9 Number of principal components 45
2.3.10 Comparison to batch effect removal tools 47
2.3.11 The effect of normalization strategy on effect size estimates . . 49
2.3.12 The performance of Funnorm for smaller sample sizes 51
2.4 Discussion 53
2.4.1 Reproducibility 57
2.5 Supplementary Material 57
2.6 Supplementary Figures 60

3 Normalization of structural MRI images 67

3.1 Introduction 67
3.2 Materials and methods 71
3.2.1 Study population 71
3.2.2 Imaging sequences and preprocessing 72
3.2.3 RAVEL methodology 74
3.2.4 Estimation of the number of unwanted factors 76
3.2.5 Comparison to intensity normalization methods 78
3.2.6 Identification of voxels associated with clinical covariates . . . 79
3.2.7 Evaluating the replicability of the top voxels associated with AD 80
3.2.8 Pseudo-ROC curves and enrichment curves 81
3.3 Results 83

3.3.1 RAVEL reduces inter-subject variability 84
3.3.2 RAVEL improves replicability of large MRI studies 88
3.3.3 RAVEL uncovers known regions associated with AD 89
3.3.4 RAVEL-corrected intensities improve prediction of AD and MCI 93
3.4 Discussion 95

4 Reconstruction of A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data 104

4.1 Introduction 105
4.2 Results and discussion 108
4.2.1 A/B compartments are highly reproducible and are cell-type specific 108
4.2.2 Predicting A/B compartments from DNA methylation data . 113
4.2.3 Long-range correlations in DNA methylation data predict A/B compartment changes between cell types 117
4.2.4 The structure of long-range correlations in DNA methylation data 121
4.2.5 The impact of GC content on long-range correlations in DNA methylation data 130
4.2.6 Sometimes compartment prediction fails using DNA methylation data 132
4.2.7 Notes on processing of the DNA methylation data 135

4.2.8 An application to prostate cancer 136
4.2.9 Compartments across human cancers 140
4.2.10 Compartment prediction using DNase hypersensitivity data . 141
4.2.11 Compartment prediction using single-cell epigenetic data . . . 148
4.3 Conclusion 151
4.4 Methods 157
4.4.1 Infinium HumanMethylation450 BeadChip 157
4.4.2 Methylation Data 157
4.4.3 Processing of the methylation data 159
4.4.4 Construction of 450k correlation matrices 161
4.4.5 Hi-C Data 161
4.4.6 Processing of the Hi-C data 161
4.4.7 Construction of Hi-C matrices 163
4.4.8 DNase-Seq data 164
4.4.9 GC content correction of the DNase data 166
4.4.10 Single-cell ATAC-seq data 166
4.4.11 Single-cell WGBS data 167
4.4.12 Eigenvector analysis 168
4.4.13 Somatic mutations in PRAD 169
4.4.14 Data 170
4.4.15 Software 170

5 Interactive visualization of DNA methylation data **171**

5.1 Introduction 171

5.2 Methods 173

5.2.0.1 Raw data summarization 173

5.2.0.2 Pre-processed data summarization (optional) 174

5.2.1 Quality control assessment 175

5.2.2 Sex prediction 178

5.2.3 Example data 178

5.3 Discussion 180

5.3.1 Software Availability 180

Bibliography **181**

List of Tables

3.1 Summary statistics of the ADNI sample. 72
3.2 Scanning parameters for the ADNI data subset. 73
3.3 Brain regions previously reported to undergo a structural change in the progression of AD. 82

4.1 Correlation and agreement between Hi-C and 450k-based eigenvector estimates of genome compartments. 115
4.2 Number of somatic mutations per 100kb in PRAD stratified by compartment. 140
4.3 Methylation data sources. 159
4.4 Hi-C data sources. 162
4.5 DNase-Seq data sources. 164
4.6 Single-cell epigenetic data sources. 167

List of Figures

2.1 Global changes in DNA methylation for the EBV dataset. 9
2.2 RUV tuning plots. 26
2.3 Control probes acts as surrogates for batch effects. 31
2.4 Improvements in replication for the EBV dataset. 35
2.5 Improvements in replication for the TCGA KIRC dataset. 38
2.6 Improvements in replication of tumor subtype heterogeneity. 40
2.7 Performance Improvements on blood samples dataset. 43
2.8 Variance across technical triplicates. 45
2.9 Spatial location affects overall methylation. 46
2.10 The impact of the number of principal components. 48
2.11 Comparison to batch effect removal tools SVA, RUV and ComBat . . 50
2.12 Effect size of the top replicated loci. 51
2.13 Sample size simulation for the Ontario-EBV dataset. 52
2.14 Illustration of in silico batch effects. 61
2.15 Improvements in replication for the EBV dataset, all methods. 62
2.16 Improvements in replication for the TCGA KIRC dataset, all methods. 63
2.17 Plate effects and dye bias for the AML dataset. 64
2.18 Improvements in replication of tumor subtype heterogeneity. 65
2.19 Improvements in blood samples, all methods. 66

3.1 Schematic showing the RAVEL pipeline. 77
3.2 CAT plots with additional methods. 83
3.3 Estimation of technical variability using CSF control voxels. 85
3.4 Effect of RAVEL on the histograms of intensities. 87
3.5 RAVEL improves replicability of voxels associated with AD 90
3.6 The top voxels associated with AD are enriched for the hippocampus and parahippocampal regions 92
3.7 Voxel-level p-value maps from AD vs. healthy patient differential analysis 94
3.8 RAVEL improves the prediction of AD and MCI. 96

4.1 A/B compartments are reproducible and cell-type specific. 109
4.2 A/B compartments revealed by Hi-C data do not change at resolutions higher than 100kb. 110
4.3 Association between the Hi-C eigenvectors and GC content is reproducible and cell-type specific. 112
4.4 Association between the Hi-C eigenvectors and GC content is reproducible and cell-type specific. 113
4.5 The methylation correlation signal is a better predictor of A/B compartments than the average methylation signal. 115
4.6 Cell-type specific A/B compartments using Hi-C data are predicted using DNA methylation data. 118
4.7 Compartment predictions based on 450k data are cell-type specific. . 119
4.8 Densities of the correlations of the 450k methylation probes. 123
4.9 The relationship between a Hi-C contact matrix and a binned DNA methylation correlation matrix. 124
4.10 Between-chromosome correlations of DNA methylation. 125
4.11 The relationship between a Hi-C contact matrix and a binned DNA methylation correlation matrix. 127
4.12 Sample ranking based on methylation levels in the closed compartments replicate across experiments. 128
4.13 Mean methylation levels in the 450k-Fibroblast dataset are not associated with technical control probes. 129
4.14 Relationship between long-range correlations, GC content and methylation levels for the 450k-EBV dataset. 131
4.15 Between-sample variability in marginal methylation. 133
4.16 The methylation correlation signal of the 450k-Blood dataset does not correlate well with other datasets. 135
4.17 Comparison of the methylation levels and the Hi-C compartments signal for the 450k-PRAD datasets. 137
4.18 Relationship between A/B compartments and somatic mutation rate in prostate cancer. 139
4.19 Estimated A/B compartments across several human cancers. 142
4.20 DNase data can predict A/B compartments revealed by Hi-C. 144
4.21 Relationship between DNase scores and GC content. 146
4.22 Densities of the correlations of DNase data. 147
4.23 Single-cell ATAC-seq data. 150
4.24 Single-cell WGBS data. 152

5.1 The workflow of shinyMethyl 174
5.2 The shinyMethyl user interface for quality control. 176
5.3 Visualization of cancer/normal differences in the TCGA dataset, before and after normalization 177

5.4 Sex prediction interface. 179

Chapter 1

Introduction

In the past decade, there is been an undoubtful and exponential increase of biomedical data: thousands of individuals are now being sequenced for their DNA, the Human Connectome Project is collecting thousand of medical images to study the living brain, and it is common to have a friend who wears a portable accelerometer to keep track of their daily activity. For a statistician, it has never been more exciting than now to be part of biomedical research.

The main goal of conducting large biomedical studies is to increase our knowledge of biological phenomena at the population level. It is tempting to believe that biological studies with larger sample size systematically lead to a better statistical inference. In reality, the noise in those large biological studies is often not at random; there often exists a particular unknown structure in the noise that does not cancel out as the sample size increases. Such structured noise can be a consequence of the

specific study design constraints, but also can be due to systematic unknown technical artifacts, often hard to identify and measure. While randomization can help prevent confounding of unknown unwanted factors with the outcome of interest, the strength of the association between the outcome and the biological measurements can be diminished by such unwanted variation. In addition, in the situation where some of the unwanted factors are confounded with the outcome of interest, larger sample sizes will not necessarily improve the validity of the biological findings, but could rather exaggerate the effect of the confounding biases.

A good example of biomedical studies with structured noise are studies of gene expression. Gene expression microarray is an inexpensive technology to assess the level of gene expression for a large number of genes simultaneously, across hundreds of people. For large studies, it is not possible to process all of the microarrays synchronously on the same machine. The different laboratory conditions, machine protocols and sample preparations between the different processed batches often cause technical differences in the gene expression measurements. Those differences have been coined before under the term "batch effects", and if not corrected for, can lead to incorrect conclusions.[22] We note that high-throughput sequencing data are equally subject to those technical artifacts.

For the first part of the dissertation, we present statistical methods for the removal of unwanted variation in large biomedical studies: first in DNA methylation microarray studies, and second in structural MRI studies. For both data types, because of

the high-dimensional nature of the data (each observation contains hundreds of thousands of features) we show that it is possible to estimate the unwanted variation by selecting a subset of features that are not associated with the outcome of interest: control probes in the case of the DNA methylation data, and control voxels in the case of MRI images.

In particular, in Chapter 2, we present our work on the normalization of DNA methylation data. DNA methylation, an addition of methyl groups to the DNA, is an important epigenetic mark that occurs mostly at CpG dinucleotides in humans. Among others, DNA methylation is implicated in gene silencing. In 2011, Illumina released the HumanMethylation450 bead array,[2] also known as the 450k array. This array has enabled population-level studies of DNA methylation by providing a cheap, high-throughput and comprehensive assay for DNA methylation. Applications of this array to population-level data include epigenome-wide association studies (EWAS)[3,4] and large-scale cancer studies, such as the ones available through The Cancer Genome Atlas (TCGA). Today, around 10,000 samples are available from the Gene Expression Omnibus of the National Center for Biotechnology Information, and around 10,000 samples from TCGA have been profiled on the 450k array.

Studies of DNA methylation in cancer pose a challenging problem for array normalization. It is widely accepted that most cancers show massive changes their methylome compared to normal samples from the same tissue of origin, making the marginal distribution of methylation across the genome different between cancer and normal

samples.[5–9] Therefore, quantile normalization, the most widely used normalization for microarray

As a solution, we present a statistical normalization procedure, called Functional normalization (Funnorm). Funnorm is an extension to quantile normalization that removes unwanted technical variation using control probes. We adapt our algorithm to the Illumina 450k methylation array and address the open problem of normalization methylation data with global epigenetic changes. Using datasets from TCGA, we show that our algorithm outperforms all existing normalization methods with respect to replication of results between experiments, and yields robust results even in the present of batch effects.

In Chapter 3, we present our work on MRI images studies. In recent years, there has been an increase in the number of multi-site neuroimaging studies, including the Human Connectome Project (HCP), the Alzheimer's Disease NeuroImaging Initiative (ADNI) and the Australian Imaging, Biomarkers and Lifestyle (AIBL). Because MRI intensities are acquired in arbitrary units, it has often been found that the differences in MRI intensities between scanning parameters and studies are larger than the biological differences observed in these images. In order to study disease at the population level, it is necessary to obtain scans from different scanning sites, and therefore controlling for the different scanning parameters becomes a near-impossible task. It is inconceivable to scan al patients on the same machine, at the same time. Most of the time, patients are followed longitudinally and need to come at different

visits, with many years in between, during which the technology and scanner protocol may have changed considerably, thereby increasing the technical variability between scans. As an example, despite the trial of harmonizing the ADNI protocol across centers and visits, we show in this dissertation that a huge technical variation between scans remain present, and by far exceeding the biological differences between individuals.

As a solution, we present a novel methodology for the removal of unwanted variation in large MRI studies. We call the method Removal of Artificial Voxel Effect by Linear regression (RAVEL). As proposed by SVA and RUV,[26,27,29] two batch effect correction tools largely used in genomics, we decompose the voxel intensities of images registered to a template into a biological component and an unwanted variation component. The unwanted variation component is estimated from a control region obtained from the cerebrospinal fluid (CSF), where intensities are known to be unassociated with disease status and other clinical covariates. We assess the performance of RAVEL using T1-weighted (T1-w) images from more than 900 subjects with Alzheimers disease (AD) and mild cognitive impairment (MCI), as well as healthy controls from ADNI database. We show that the RAVEL-corrected intensities have the best performance in distinguishing between MCI subjects and healthy subjects using the mean hippocampal intensity (AUC=67%), a marked improvement compared to results from existing intensity normalizations.

For the second part of the dissertation, we present two pieces of work related to

the analysis of DNA methylation data. In particular, in Chapter 4, we present a computational method that predicts A/B compartments as revealed by Hi-C data using long-range correlations in epigenetic data, focusing on the DNA methylation microarray data discussed above. Analysis of Hi-C data has shown that the genome can be divided into two compartments called A/B compartments. These compartments are cell-type specific and are associated with open and closed chromatin. We show that A/B compartments can reliably be estimated using data from the Illumina 450k DNA methylation microarray, DNase hypersensitivity sequencing, single-cell ATAC sequencing and single-cell whole-genome bisulfite sequencing. We do this by exploiting that the structure of long-range correlations differs between open and closed compartments. This work makes A/B compartment assignment readily available in a wide variety of cell types, including many human cancers.

In Chapter 5, we present shinyMethyl, a Bioconductor package for interactive quality control of DNA methylation data from the Illumina 450k array. The package summarizes 450k experiments into small exportable R objects from which an interactive interface is launched. Reactive plots allow fast and intuitive quality control assessment of the samples. In addition, exploration of the phenotypic associations is possible through coloring and principal component analysis. Altogether, the package makes it easy to perform quality assessment of large-scale methylation datasets, such as epigenome-wide association studies or the datasets available through TCGA. The shinyMethyl package is implemented in R and available via Bioconductor.

Chapter 2

Normalization of DNA methylation data

This chapter describes work published in a separate form in the journal *Genome Biology*, with contributions from co-authors Aurélie Labbe, Mathieu Lemire, Brent W. Zanke, Thomas J. Hudson, Elana J. Fertig, Celia M.T. Greenwood and Kasper D. Hansen.[1]

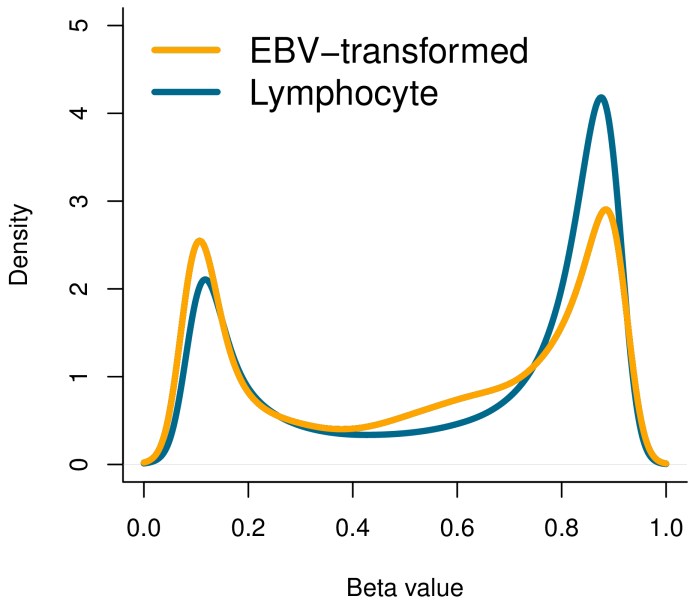
2.1 Introduction

In humans, DNA methylation is an important epigenetic mark occurring at CpG dinucleotides, implicated in gene silencing. In 2011, Illumina released the IlluminaHumanMethylation450 bead array,[2] also known as the "450k array". This array

has enabled population-level studies of DNA methylation by providing a cheap, high-throughput and comprehensive assay for DNA methylation. Applications of this array to population level data includes epigenome-wide association studies (EWAS)[3,4] and large-scale cancer studies such as the ones available through The Cancer Genome Atlas (TCGA). To date, around 9000 samples are available from NCBI GEO, and around 8000 samples from TCGA have been profiled on either the 450k array, the 27k array or both.

Studies of DNA methylation in cancer pose a challenging problem for array normalization. It is widely accepted that most cancers show massive changes in their methylome compared to normal samples from the same tissue of origin, making the marginal distribution of methylation across the genome different between cancer and normal samples;[5–9] see Figure 2.1 for an example of such a global shift. We refer to this as global hypomethylation. The global hypomethylation commonly observed in human cancers was recently shown to be organized into large, well-defined domains.[10,11] It is worth noting that there are other situations where global methylation differences can be expected, such as between cell types and tissues.

Several methods have been proposed for normalization of the 450k array, including quantile normalization,[12,13] SWAN,[14] BMIQ,[15] dasen,[16] and noob.[17] A recent review examined the performance of many normalization methods in a setting with global methylation differences and concluded "there is to date no between-array normalization method suited to 450K data that can bring enough benefit to counterbalance



(a)

Figure 2.1: Global changes in DNA methylation for the EBV dataset. (a) The average density of unnormalized beta values across both EBV transformed lymphocytes and normal lymphocytes, showing global hypomethylation caused by EBV transformation.

the strong impairment of data quality they can cause on some data sets".[18] The authors note that absence of normalization outperforms the methods they evaluate, highlighting the importance of benchmarking any method against raw data.

The difficulties in normalizing DNA methylation data across cancer and normal samples simultanously have been recognized for a while. In earlier work on the CHARM platform,[19] Aryee *et al.*[20] proposed a variant of subset quantile normalization[21] as a solution. For CHARM, input DNA is compared to DNA processed by a methylation dependent restriction enzyme. Aryee *et al.*[20] used subset quantile normalization to normalize the input channels from different arrays to each other. The 450k assay does not involve an input channel; it is based on bisulfite conversion. While not directly applicable to the 450k array design, the work on the CHARM platform is an example of an approach to normalizing DNA methylation data across cancer and normal samples.

Any high-throughout assay suffers from unwanted variation.[22] This is best addressed by experimental design.[22] In the gene expression literature, correction for this unwanted variation was first addressed by the development of unsupervised normalization methods such as RMA[23] and VSN.[24] As Mecham *et al.*,[25] we use the term "unsupervised" to indicate that the methods are unaware of the experimental design; all samples are treated equally. These methods lead to a substantial increase in signal-to-noise. As experiments with larger sample sizes were performed, it was discovered that substantial unwanted variation remained in many experiments despite

the application of an unsupervised normalization method. This unwanted variation is often – but not exclusively – found to be associated with processing date or "batch" and is therefore referred to as a batch effect. This led to the development of a series of supervised normalization tools such as SVA,[26,27] ComBat,[28] SNM[25] or RUV[29] also known as batch effect removal tools. The supervised nature of these tools allows them to aggressively remove unwanted variation while keeping variation associated with the covariate of interest (such as case/control status). Unsurprisingly, batch effects have been observed in studies using the 450K array[30]

As an example of unwanted variation which is biological in origin, we draw attention to the issue of cell type heterogeneity which has seen a lot of attention in the literature on DNA methylation.[31–35] This issue arises when primary samples are profiled; primary samples are usually a complicated mixture of cell types. This mixture can substantially increase the unwanted variation in the data and can even confound the analysis if the cell type distribution depends on a phenotype of interest. It has been shown that SVA can help mitigate the effect of cell type heterogeneity,[34] but other approaches are also useful.[31–33]

In this work, we propose an unsupervised method that we call functional normalization, which uses control probes to act as surrogates for unwanted variation. We apply this method to the analysis of 450k array data, and show that functional normalization outperforms all existing normalization methods for analysis of datasets with global methylation differences, including studies of human cancer. We also show

that functional normalization outperforms the batch removal tools SVA,[26,27] ComBat[28] and RUV[29] in this setting. Our evaluation metrics are focusing on assessing the degree of replication between large-scale studies, arguably the most important, biologically relevant end-point for such studies. Our method is available as the "preprocessFunnorm" function in the minfi package[13] through the Bioconductor project.[36]

2.2 Methods

2.2.1 Infinium HumanMethylation450 BeadChip

We use the following terminology, consistent with the minfi package:[13] the 450k array is available as slides consisting of 12 arrays. These arrays are arranged in a 6 rows by 2 columns layout. The scanner can process up to 8 slides in a single plate.

We use the standard formula $\beta = M/(M + U + 100)$ for estimating percent methylation given (un)methylation intensities U and M.

Functional normalization uses information from the 848 control probes on the 450k array, as well as the "out-of-band" probes discussed in Triche *et al.*[17] These control probes are not part of the standard output from GenomeStudio, the default Illumina software. Instead we use the IDAT files from the scanner together with the open source illuminaio[50] package to access the full data from the IDAT files. This step is implemented in minfi.[13] While not part of the standard output from GenomeStudio,

it is possible to access the control probe measures within this software by accessing the "Control Probe Profile".

2.2.2 Control probe summaries

We transform the 848 control probes, as well as the out-of-band probes[17] into 42 summary measures. The control probes contributes 38 of these 42 measures and the out-of-band contributes 4. An example of a control probe summary is the mapping of 61 "C" normalization probes to a single summary values, their mean. The out-of-band probes are the intensities of the Type I probes measured in the opposite color channel from the probe design. For the 450k platform, this means 92,596 green intensities, and 178,406 red intensities that can be used to estimate background intensity, and we summaries these values into 4 summary measures. A full description of how the control probes and the out-of-band probes are transformed into the summary control measures are listed in the Supplementary Material.

2.2.3 Functional normalization: the general framework

Functional normalization extends the idea of quantile normalization, by adjusting for known covariates measuring unwanted variation. In this section we present a general model that is not specific to methylation data. The adaptation of this general model

to the 450k data is discussed in the next section. The general model is as follows. Consider $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ high dimensional vectors each associated with a set of scalar covariates $Z_{i,j}$ with $i = 1, \ldots, n$ indexing samples and $j = 1, \ldots, m$ indexing covariates. Ideally these known covariates are associated with unwanted variation and unassociated with biological variation; functional normalization attempts to remove their influence. For each high-dimensional observation $\mathbf{Y}_i$, we form the empirical quantile function for its marginal distribution, and denote it by q_i^{emp}. Quantile functions are defined on the unit interval and we use the variable $r \in [0, 1]$ to evaluate them pointwise, like $q_i^{\text{emp}}(r)$. We assume the following model in pointwise form

$$q_i^{\text{emp}}(r) = \alpha(r) + \sum_{j=1}^{m} Z_{i,j}\beta_j(r) + \epsilon_i(r) \tag{2.1}$$

which has the functional form

$$q_i^{\text{emp}} = \alpha + \sum_{j=1}^{m} Z_{i,j}\beta_j + \epsilon_i \tag{2.2}$$

The parameter function α is the mean of the quantile functions across all samples, β_j are the coefficient functions associated with the covariates and ϵ_i are the error functions which are assumed to be independent and centered around 0.

In this model, the term

$$\sum_{j=1}^{m} Z_{i,j}\beta_j \tag{2.3}$$

represents variation in the quantile functions explained by the covariates. By specifying known covariates that measure unwanted variation and that are not associated with biological signal, functional normalization removes unwanted variation by regressing out the latter term. An example of a known covariate could be processing batch. In a good experimental design, processing batch will not be associated with biological signal.

In particular, assuming we have obtained estimates $\hat{\beta}_j$ for $j = 1, \ldots, m$, we form the functional normalized quantiles by

$$q_i^{\text{Funnorm}}(r) = q_i^{\text{emp}}(r) - \sum_{j=1}^{m} Z_{i,j}\hat{\beta}_j(r) \tag{2.4}$$

We then transform $\mathbf{Y}_i$ into the functional normalized quantity $\tilde{\mathbf{Y}}_i$ using the formula

$$\tilde{\mathbf{Y}}_i = q_i^{\text{Funnorm}}\big((q_i^{\text{emp}})^{-1}(\mathbf{Y}_i)\big) \tag{2.5}$$

This ensures that the marginal distribution of $\tilde{\mathbf{Y}}_i$ has q_i^{Funnorm} as its quantile function.

We now describe how to obtain estimates $\hat{\beta}_j$ for $j = 1, \ldots, m$. Our model 2.1 is an example of function-on-scalar regression, described in.[51] The literature on function-on-scalar regression makes assumptions about the smoothness of the coefficient functions and uses a penalized framework because the observations appear noisy and non-smooth. In contrast, because our observations $\mathbf{Y}_i$ are high-dimensional and continuous, the jumps of the empirical quantile functions are very small. This allows

us to circumvent the smoothing approach used in traditional function-on-scalar regression. We use a dense grid of H equidistant points between 0 and 1, and we assume that H is much smaller than the dimension of $\mathbf{Y}_i$. On this grid, model 2.1 reduces pointwise to a standard linear model. Because the empirical quantile functions $q^{\text{emp}}(r)$ have very small jumps, the parameter estimates of these linear models vary little between two neighbouring grid points. This allows us to use H standard linear model fits to compute estimates $\hat{\alpha}(h), \hat{\beta}_j(h), j = 1, \ldots, m$ with h being on the dense grid $\{h \in d/H : d = 0, 1, \ldots, H\}$. We next form estimates $\hat{\alpha}(r), \hat{\beta}_j(r), j = 1, \ldots, m$ for any $r \in [0, 1]$ by linear interpolation. This is much faster than the penalized function-on-scalar regression available through the refund package.[52]

Importantly, in this framework, using a saturated model in which all the variation (but the mean) is explained by the covariates results in removing all variation and is equivalent to quantile normalization. In our notation, quantile normalized quantile functions are

$$q_i^{\text{quantile}}(r) = \hat{\alpha}(r) \tag{2.6}$$

where $\hat{\alpha}$ is the mean of the empirical quantile functions. This corresponds to the maximum variation that can be removed in our model. In contrast, including no covariates makes the model comparable to no normalization at all. By choosing covariates which only measure unwanted technical variation, functional normalization will only remove the variation explained by these technical measurements and will leave biological variation intact. Functional normalization allows a sensible tradeoff

between not removing any technical variation at all (no normalization) and removing too much variation, including global biological variation, as it can occur in quantile normalization.

2.2.4 Functional normalization for 450k arrays

We apply the functional normalization model to the methylated (M) and unmethylated (U) channels separately. Since we expect the relationship between the methylation values and the control probes to differ between Type I and Type II probes, functional normalization is also applied separately by probe type to obtain more representative quantile distributions. We address probes mapping to the sex chromosomes separately, see below. This results in 4 separate applications of functional normalization, using the exact same covariate matrix, with more than 100,000 probes in each normalization fit. For functional normalization, we pick $H = 500$ equidistant points (see notation in previous section). As covariates, we use the first $m = 2$ principal components of the summary control measures as described above. We do this because the control probes are not intended to measure biological signal since they are not designed to hybridize to genomic DNA. Our choice of $m = 2$ is based on empirical observations on several datasets.

Following the ideas from quantile normalization for 450k arrays,[12,13] we normalize the probes mapping to the sex chromosomes (11,232 and 416 probes for the X and Y chromosomes respectively) separately from the autosomal probes. For each

of the two sex chromosomes, we normalize males and females separately. For the X chromosome we use functional normalization and for the Y chromosome we use quantile normalization since the small number of probes on this chromosome violates the assumptions of functional normalization which results in instability.

Functional normalization only removes variation in the marginal distributions of the two methylation channels associated with control probes. This preserves any biological global methylation difference between samples.

We have found (see Results) that we get slightly better performance of functional normalization if we apply it to data that have been background corrected with noob.[17]

2.2.5 Data

2.2.5.1 Data description

The Ontario study.

The Ontario study consists of samples from 2200 individuals from the Ontario Familial Colon Cancer Registry (OFCCR)[53] who had previously been genotyped in a case-control study of colorectal cancer in Ontario.[54] The majority of these samples are lymphocytes derived from whole blood. We use various subsets of this dataset for different purposes.

The Ontario-EBV dataset. Lymphocyte samples from 100 individuals from the Ontario study were transformed into immortal lymphoblastoid cell lines (LCL)

using EBV transformation. We divided the 100 EBV-transformed samples into two equal equally sized datasets (discovery and validation). For the discovery dataset, we matched the 50 EBV-transformed samples to 50 other lymphocyte samples assayed on the same plates. For the validation dataset, we matched the 50 EBV-transformed samples to 50 other lymphocyte samples assayed on different plates.

The Ontario-Blood dataset. From the Ontario study, we first created a discovery-validation design where we expect only a small number of loci to be differentially methylated. For the discovery dataset, we selected all cases and controls on 3 plates showing little evidence of plate effects among the control probes, which yielded a total of 52 cases and 231 controls. For the validation dataset, we selected 4 plates where the control probes did show evidence of a plate effect and then selected cases and controls from separate plates, to maximize a confounding effect of plate. This yielded a total of 175 cases and 163 controls.

The Ontario-Sex dataset. Among 10 plates for which the control probes demonstrated differences in distribution depending on plate, we selected 101 males from a set of 5 plates and 105 females from another set of 5 plates, attempting to maximize the confounding effect of batch on sex.

The Ontario-Replicates dataset. Amongst the lymphocyte samples from the Ontario study, 19 samples have been assayed 3 times each. One replicate is a hybridization replicate and the other replicate is a bisulfite conversion replicate. The 57 samples have been assayed on 51 different slides across 11 different plates.

The TCGA-KIRC datasets. From The Cancer Genome Atlas (TCGA) we have access to kidney renal clear cell carcinoma and normal samples, assayed on two different methylation platforms. We use the level 1 data which contains the IDAT files. For the 450k platform, TCGA has assayed 300 tumor samples and 160 normal samples. For the discovery set, we select 65 tumor samples and 65 matched normals from slides showing little variation in the control probes. These 130 samples were assayed on 3 plates. For the validation dataset we select the remaining 95 normal samples together with all 157 cancer samples part of the same TCGA batches as the 95 normals. These samples were spread over all 9 plates, therefore maximizing potential batch effects. For the 27k platform, TCGA has assayed 219 tumor samples and 199 normals. There is no overlap between the individuals assayed on the 450k platform and the individuals assayed on the 27k platform.

The TCGA-AML datasets. Also from TCGA, we used data from 194 acute myeloid leukemia (AML) samples, where each sample was assayed twice: first on the 27K Illumina array and subsequently on the 450K array. Every sample but 2 have been classified according to the French-American-British (FAB) subtype classification scheme[41] which classifies the tumor into one of 8 subtypes. The 2 unclassified samples were removed post-normalization. We use the data known as level 1, which contains the IDAT files.

WGBS EBV data. Hypomethylated blocks and small differentially methylated regions (DMRs) between transformed and quiescent cells were obtained from a pre-

vious study.[39] Only blocks and DMRs with a family-wise error rate equal to 0 was retained (see the reference). A total of 228,696 probes on the 450K array overlap with the blocks and DMRs.

The Ontario methylation data has been deposited in dbGAP under accession number phs000779.v1.p1. This data is available to researchers under the following constraints: that (1) use of data is limited to research on cancer (2) local IRB approval and (3) approval of either Colon Cancer Family Registries (`http://coloncfr.org/collaboration`) or Mount Sinai Hospital (Toronto) Research Ethics Board Approval.

The TCGA data (KIRC and AML) is available through the TCGA Data Portal (`https://tcga-data.nci.nih.gov/tcga/`

The WGBS EBV data is available through NCBI GEO under the accession number GSE49629.

Our method is available as the "preprocessFunnorm" function in the minfi package through the Bioconductor project (`http://www.bioconductor.org/packages/release/bioc/html/minfi.html`). The code in this package is licensed under the open source license Artistic-2.0.

2.2.5.2 Data processing

Data was available in the form of IDAT files from the various experiments (see above). We used minfi[13] and illuminaio[50] to parse the data and used the various normalization routines in their reference implementations (see below).

We performed the following quality control on all datatsets. As recommended in *et al.*,[12] for each sample we computed the percentage of loci with a detection p-value greater than $p = 0.01$, with the intention of excluding a sample if the percentage was higher than 10%. We used the minfi[13] implementation of the detection p-value. We also used additional quality control measures[13] and we interactively examined the arrays using the shinyMethyl package;[55] all arrays in all datasets passed our quality control.

We performed the following filtering of loci, after normalization. We removed $17,302$ loci which contains a SNP with an annotated minor allelle frequency greater than or equal to 1% in the CpG site itself or in the single-base extension site. We used the UCSC Common SNPs table based on dbSNP 137; this table is included in the minfi package. We removed $29,233$ loci which have been shown to cross-hybridize to multiple genomic locations.[44] The total number of loci removed is $46,535$, ie 9.6% of the array. We chose to remove these loci post-normalization as done previously,[17,56] reasoning that while these probes may lead to spurious associations, we believe they are still subject to technical variation and should therefore contain information useful for normalization.

2.2.6 Comparison to normalization methods

We have compared functional normalization, "Funnorm", to the most popular normalization methods used for the 450k array. This includes the following between-array

normalization methods: (1) "Quantile": stratified quantile normalization as proposed by Touleimat *et al.*[12] and implemented in minfi,[13] (2) "dasen": background adjustment and between-sample quantile normalization of M and U separately[16] and (3) "noob": a background adjustment model using the out-of-band control probes followed by a dye bias correction,[17] implemented in the methylumi package. We also consider two within-array normalization methods: (4) "SWAN": Subset-quantile within-array normalization[14] and (5) "BMIQ": Beta-mixture quantile normalization.[15] Finally, we consider (6) "Raw" data: no normalization, ie. we only matched up the Red and the Green color channels with the relevant probes according to the array design (specifically, it is the output of the "preprocessRaw" function in minfi.)

In its current implementation, noob yielded missing values for at most a couple of thousand loci (less than 1%) per array. This is based on excluding loci below an array-specific detection limit. We have discarded those loci from our performance measures, but only for the noob performance measures. In its current implementation, BMIQ produced missing values for all type II probes in 5 samples for the TCGA AML dataset. We have excluded these samples for our performance measures, but only for our BMIQ performance measures.

For clarity, in figures we focus on the top-performing methods which are Quantile, Raw and noob. The assessments of the other methods, dasen, BMIQ and SWAN are available in Supplementary Materials.

2.2.7 Comparison to SVA, RUV and ComBat

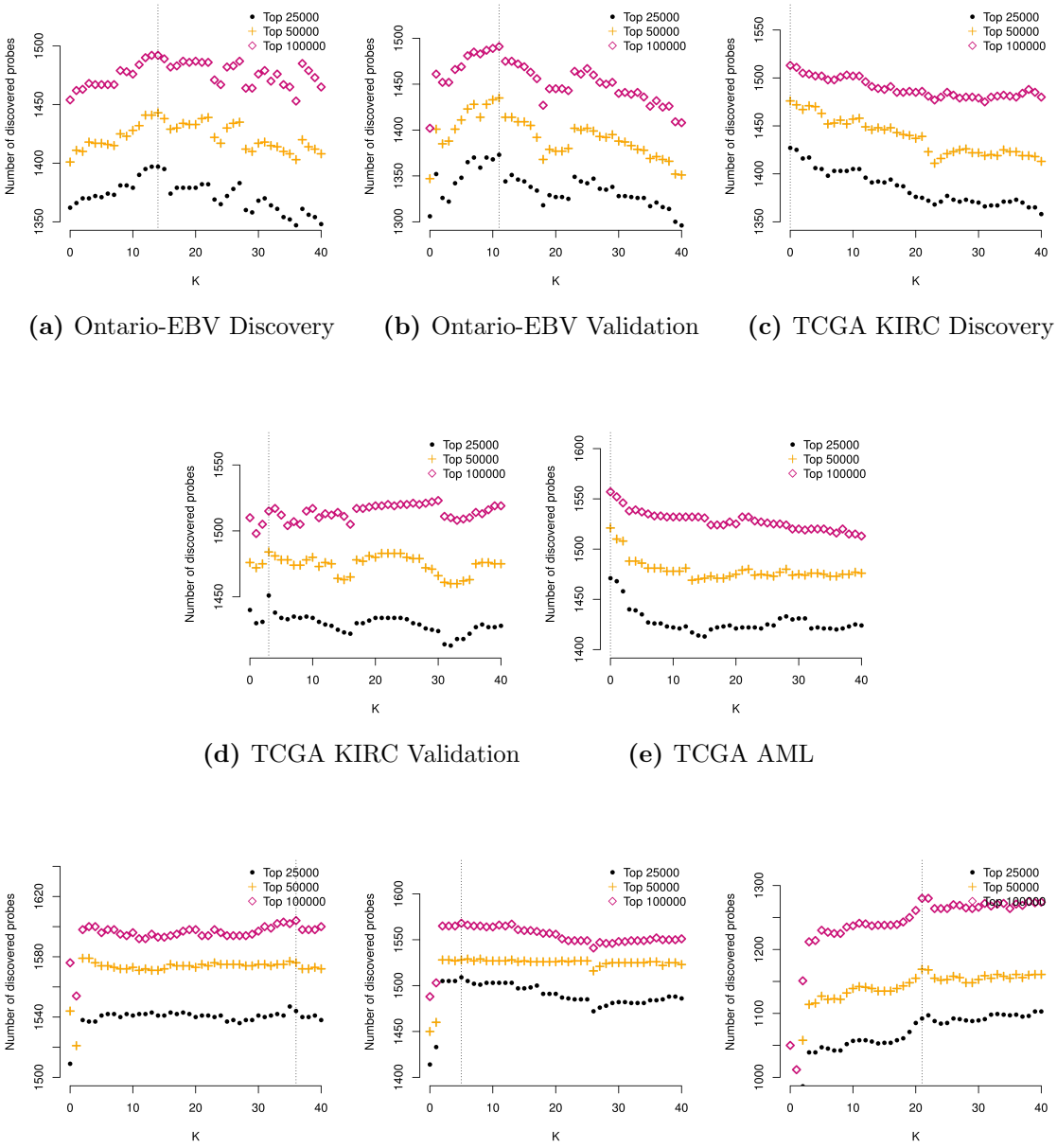
We used the reference implementation of SVA in the "sva" package.[46] We applied SVA to the M-values obtained from the raw data. Surrogate variables were estimated using the iteratively re-weighted surrogate variable analysis algorithm,[27] and were estimated separately for the discovery and validation cohorts. In the analysis of the Ontario-EBV dataset, SVA found 21 and 23 surrogate variables respectively for the discovery and the validation cohorts. In the analysis of the Ontario-Blood dataset, SVA found 18 and 21 surrogate variables respectively for the discovery and the validation cohorts. In the analysis of the TCGA KIRC dataset, SVA found 29 and 32 surrogate variables respectively for the discovery and the validation cohorts. In the analysis of the TCGA AML dataset, SVA found 24 surrogate variables.

The RUV-2 method was originally developed for gene expression microarrays.[29] The method involves a number of domain specific choices. To our knowledge, there is no publicly available adaption of RUV-2 to the 450k platform, so we adapted RUV-2 to the 450k array. The core of the method is implemented in software available at a personal website (`http://www.stat.berkeley.edu/~johann/ruv/`). As negative genes (genes not associated with the biological treatment group), we selected the raw intensities in the green and red channels of the 614 internal negative control probes available on the 450k array.

To determine the number k of factors to remove (see Gagnon-Bartsch and Speed[29]

for details on this parameter), we followed the approach described in.[29] First, for each value $k = 0, 1, \ldots, 40$, we performed a differential analysis with respect to sex. Second, we considered as positive controls the probes that are known to undergo X inactivation (see Section Sex validation analysis) and probes mapping to the Y chromosome. Third, for the top ranked $m = 25000, 50000$ and 100000 probes, we counted how many of the positive control probes are present in the list. Finally, we picked the value of k to be the value for which these counts are maximized. The different tuning plots are presented in Figure 2.2. The optimal k was 14 and 11 for the discovery and the validation cohorts of the Ontario-EBV dataset respectively. In the analysis of the Ontario-Blood dataset, the optimal k was 0 and 3 respectively for the discovery and the validation cohorts. In the analysis of the TCGA KIRC dataset, the optimal k was 36 and 5 respectively for the discovery and the validation cohorts. In the analysis of the TCGA AML dataset, k was selected to be 0 (which is equivalent to the Raw method).

We used the reference implementation of ComBat in the sva package.[46] Because ComBat cannot be applied to datasets for which the phenotype of interest is perfectly confounded with the batch variable, we could only run ComBat for the AML and KIRC datasets.



(a) Ontario-EBV Discovery **(b)** Ontario-EBV Validation **(c)** TCGA KIRC Discovery

(d) TCGA KIRC Validation **(e)** TCGA AML

(f) Ontario-Blood Discovery **(g)** Ontario-Blood Validation **(h)** Ontario-Gender

Figure 2.2: RUV tuning plots. Selection of the tuning parameter for RUV.

2.2.8 Identification of differentially methylated positions

To identify differentially methylated positions (DMPs), we used F-statistics from a linear model on the beta values from the array. The linear model was applied on a probe-by-probe basis. In most cases, the model included case/control status as a factor. In the 27K data, we adjusted for batch by including a plate indicator (given by TCGA) in the model.

2.2.9 Discovery-Validation comparisons

To measure the consistency of each normalization method at finding true DMPs, we compared results obtained on a discovery-validation split of a large dataset. Comparing results between two different subsets of a large dataset is an established idea and have been applied to the context of 450k normalization.[15,47] We extended this basic idea in a novel way by introducing an *in silico* confounding of treatment (case/control status) by batch effects as follows. In a first step, we selected a set of samples to be the discovery cohort, by choosing samples where the treatment variable is not visibly confounded by plate effects. Then the validation step is achieved by selecting samples demonstrating strong potential for treatment confounding by batch, for example by choose sample from different plates (see Data descriptions). The extent to which it

is possible to introduce such a confounding is dataset dependent. In contrast to earlier work,[47] we normalized the discovery and the validation cohort separate, to more realistically mimic an independent replication experiment. The idea of creating *in silico* confounding between batch and treatment has been previously explored in the context of genomic prediction.[40]

We quantified the agreement between validation and discovery in two ways: by an ROC curve and a concordance curve.

For the ROC curve, we used the discovery cohort as the gold-standard. Because the validation cohort is affected by a batch effect, a normalization method that is robust to batch effects will show better performance in the ROC curve. Making this ROC curve required us to choose a set of DMPs for the discovery cohort. The advantage of the ROC curve is that the plot displays immediately interpretable quantities such as specificity and sensitivity.

For the concordance curve, we compared the top k DMPs from the discovery and the validation set each, and displayed the percentage of the overlap for each k. These curves do not require us to select a set of DMPs for the discovery cohort. Note that these curves have been previously used in the context of multiple-laboratory comparison of microarray data.[57]

2.2.10 Sex validation analysis

On the 450k array, 11232 and 416 probes map to the X and Y chromosome respectively. Because some genes have been shown to escape X inactivation,[45] we only considered genes for which the X-inactivation status is known to ensure an unbiased sex prediction. From,[45] 1678 probes undergo X-inactivation, 140 probes escape X-inactivation, and 9414 probes have either variable or unknown status.

For the ROC curves, we defined the true positives to be the 1678 probes undergoing X-inactivation and the probes mapping to the Y chromosome (416 probes) as true positives; by removing the probes that have been shown to cross-hybridize,[44] we were left with 1877 probes. For the true negatives, we considered the 140 probes escaping X-inactivation and the autosomal probes that do not cross-hybridize. The rest of the probes were removed from the analysis.

2.2.11 Sample size simulation

To assess the performance of Funnorm for different small sample sizes, we devised the following simulation scheme for the Ontario-EBV dataset. First, we kept the discovery dataset intact to ensure a reasonable gold standard in the discovery-validation ROC curves; we only simulated different sample sizes for the validation subset. For sample sizes $n = 10, 20, 30, 50, 80$, we randomly chose half of the samples from the EBV-transformed samples, and the other half from the lymphocyte samples. For

instance, for $n = 10$ samples, we randomly picked 5 samples from each of the treatment groups. We repeated this subsampling $B = 100$ times, which generated 100 discovery-validation ROC curves for each n. For a fixed n, we considered the mean of the $B = 100$ ROC curves as well as the 0.025 and 0.975 quantiles to mimic a 95% confidence interval.

2.3 Results

.

2.3.1 Control probes may act as surrogates for batch effects

The 450k array contains 848 control probes. These probes can roughly be divided into negative control probes (613), probes intended for between array normalization (186) and the remainder (51) which are designed for quality control, including assessing the bisulfite conversion rate (see Methods and Supplementary Materials). Important for our proposed method, none of these probes are designed to measure biological signal.

Figure 2.3a shows a heatmap of a simple summary (see Methods) of these control probes, for 200 samples assayed on 4 plates (Ontario dataset). Columns are the control measure summaries and rows are samples. The samples have been pro-

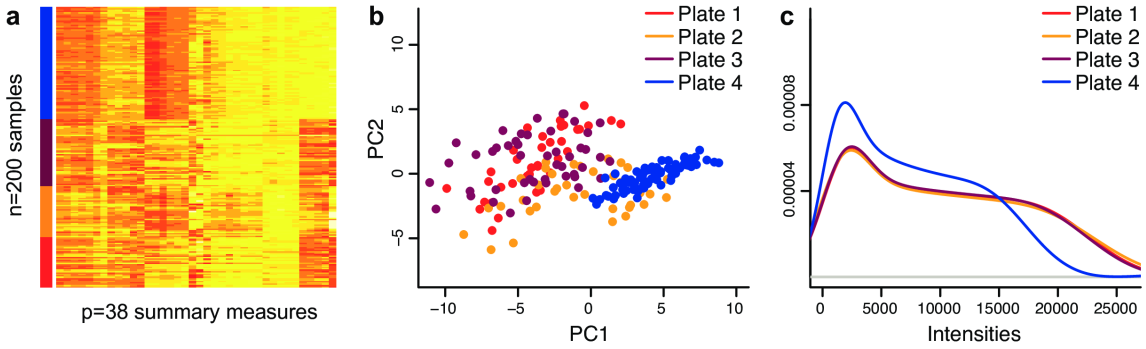


Figure 2.3: Control probes acts as surrogates for batch effects. (a) A heatmap of a summary (see Methods) of the control probes, with samples on the y-axis and control summaries on the x-axis. Samples were processed on a number of different plates indicated by the color label. Only columns have been clustered. (b) The first two principal components of the matrix depicted in (a). Samples partially cluster according to batch, with some batches showing tight clusters and other being more diffuse. (c) The distribution of methylated intensities averaged by plate. These three panels suggests that the control probe summaries partially measure batch effects.

cessed on different plates, and we observe a clustering pattern correlated with plate. Figure 2.3b shows the first two principal components of the same summary data and there is evidence of clustering according to plate. Figure 2.3c shows how the marginal distributions of the methylated channel vary across plates. This suggests that the summarized control probes can be used as surrogates for unwanted variation. This is not a new observation; the use of control probes in normalization has a long history in microarray analysis.

2.3.2 Functional normalization

We propose functional normalization (see Methods), a method which extends the idea of quantile normalization. Quantile normalization forces the empirical marginal distributions of the samples to be the same, which removes all variation in this statistic. In contrast, functional normalization only removes variation explained by a set of covariates, and is intended to be used when covariates associated with technical variation are available and are independent of biological variation. We adapted functional normalization to data from the 450k array (see Methods), using our observation that the control probe summary measures are associated with technical variability and batch effects. As covariates, we recommend using the first $m = 2$ principal components of the control summary matrix, a choice with which we have obtained consistently good results; this is discussed in greater depth below. We have also examined the contributions of the different control summary measures in several different datasets,

and we have noted that the control probe summaries given the most weight varied across different datasets. We have found (see below) that we can improve functional normalization slightly by applying it to data which have already been background corrected using the noob method.[17]

Functional normalization, like most normalization methods, does not require the analyst to provide information about the experimental design. In contrast, supervised normalization methods such as SVA,[26,27] ComBat,[28] SNM[25] and RUV[29] require the user to provide either batch parameters or an outcome of interest. Like functional normalization, RUV also utilizes control probes as surrogates for batch effects, but builds the removal of batch effects into a linear model, which returns test statistics for association between probes and phenotype. This limits the use of RUV to a specific statistical model, and methods such as clustering, bumphunting[13,37] or other regional approaches[38] for identifying differentially methylated regions (DMRs) cannot readily be applied.

2.3.3 Funnorm improves the replication between experiments, even when a batch effect is present

As a first demonstration of the performance of our algorithm, we compare lymphocyte samples from the Ontario data set to EBV-transformed lymphocytes samples from the same collection (see Methods). We have recently studied this transformation[39]

and have shown that EBV transformation induces large blocks of hypomethylation encompassing more than half the genome, similar to what is observed between most cancers and normal tissues. This introduces a global shift in methylation as shown by the marginal densities in Figure 2.1.

We divided the dataset into discovery and validation cohorts (see Methods), with 50 EBV transformed lymphocytes and 50 normal lymphocytes in each cohort. As illustrated in Figure 2.14a we attempted to introduce *in silico* unwanted variation confounding EBV transformation status in the validation cohort (see Methods), to evaluate the performance of normalization methods in the presence of known confounding unwanted variation. This has been previously done by others in the context of genomic prediction.[40] We normalized the discovery cohort, identified the top k differentially methylated positions (DMPs) and asked: "how many of these k DMPs can be replicated in the validation cohort". We normalized the validation cohort separately from the discovery cohort to mimic a replication attempt in a separate experiment. We identified DMPs in the validation cohort using the same method and the result is quantified using an ROC curve where the analysis result on the discovery cohort is taken as the gold standard.

To enable the comparison between normalization methods, we fix the number of DMPs across all methods. Because we know from previous work[39] that EBV transformation induces large blocks of hypomethylation covering more than half of the genome, we expected to find a large number of DMPs, and we set $k = 100,000$. The

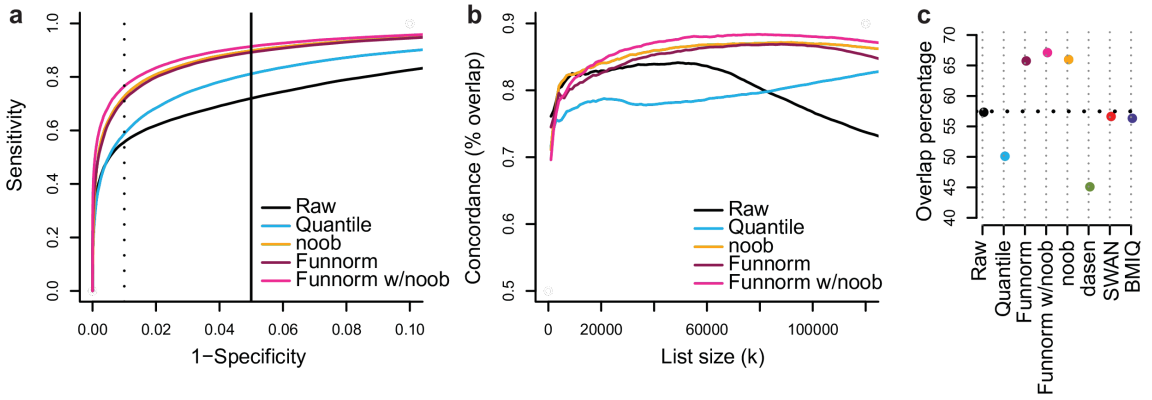


Figure 2.4: Improvements in replication for the EBV dataset. (a) ROC curves for replication between a discovery and a validation dataset. The validation dataset was constructed to show *in silico* batch effects. The dotted and solid lines represent respectively the commonly used false discovery rate cutoffs of 0.01 and 0.05. (b) Concordance curves showing the percent overlap between the top k DMPs in the discovery and validation cohort. Additional normalization methods assessed in Figure 2.15. Functional normalization shows a high degree of concordance between datasets. (c) The percentage of top 100,000 DMPs which are replicated between the discovery and validation cohort and also inside a differentially methylated block or region from Hansen *et al.*[39]

resulting ROC curves are shown in Figure 2.4a. In this figure we show, for clarity, what we have found to be the most interesting alternatives to functional normalization in this setting: Raw data, Quantile normalization as suggested by Touleimat *et al.*[12] and implemented in minfi[13] and the noob background correction.[17] Figure 2.15a,b contains results for additional normalization methods: BMIQ,[15] SWAN[14] and dasen.[16] Note that each normalization method will results in its own set of gold-standard DMPs and these ROC curves therefore measures the internal consistency of each normalization method. We note that functional normalization (with noob background correction) outperforms raw, quantile and noob normalizations when the specificity is above 90% (which is the relevant range for practical use).

We also measured the agreement between the top k DMPs from the discovery cohort with the top k DMPs from the validation cohort by looking at the overlap percentage. The resulting concordance curves are shown in Figure 2.4b, with additional methods in Figure 2.15c, and show functional normalization outperforming the other methods.

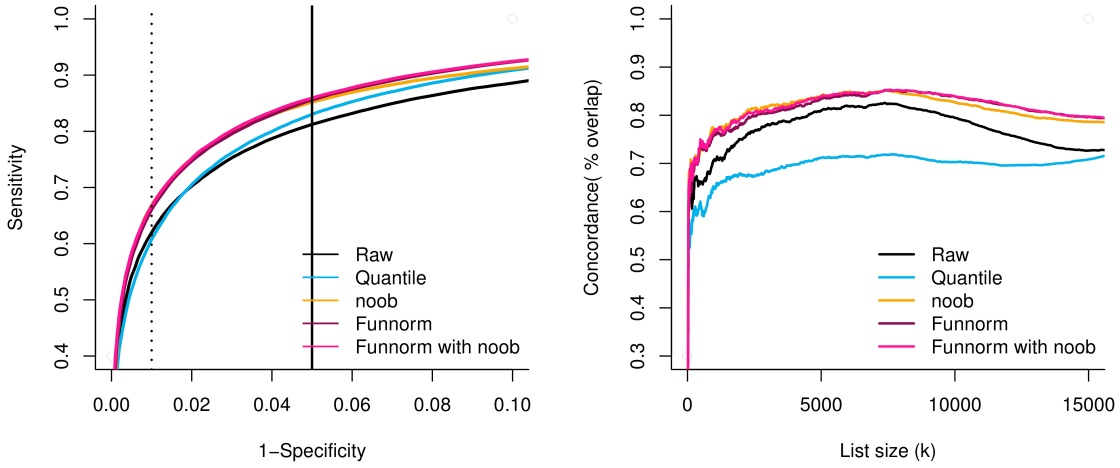
We can assess the quality of the DMPs replicated between the discovery and validation cohorts by comparing them to the previously identify methylation blocks and differentially methylated regions.[39] In Figure 2.4c, we present the percentage of the initial $k = 100,000$ DMPs that are both replicated and present among the latter blocks and regions. We note that these previously reported methylation blocks represent large scale, regional changes in DNA methylation and not regions where

every single CpG is differentially methylated. Nevertheless, such regions are enriched for DMPs. This comparison shows that functional normalization achieves a greater overlap with this external dataset, with an overlap of 67% compared to 57% for raw data, while other methods, but noob, perform worse than the raw data.

2.3.4 Replication between experiments in cancer study (TCGA-KIRC datasets)

We applied the same discovery-validation scheme to measure performance, used for the analysis of the Ontario-EBV study, on kidney renal clear cell carcinoma samples (KIRC) from TCGA. In total, TCGA has profiled 300 KIRC cancer and 160 normal samples on the 450K platform. Therefore we defined a discovery cohort containing 65 cancers and 65 normals and a validation cohort of 157 cancers and 95 normals (see Methods).

Our *in silico* attempt at introducing unwanted variation associated with batch for this experiment succeeded in producing a validation cohort where the cancer samples have greater variation in background noise (Figure 2.14b). This difference in variation is a less severe effect compared to the difference in mean background noise we achieved in the Ontario-EBV dataset (Figure 2.14a). As for the dataset containing EBV transformed samples, we expect large scale hypomethylation in the cancer samples and therefore we again consider $k = 100,000$ loci. The resulting ROC



(a) **(b)**

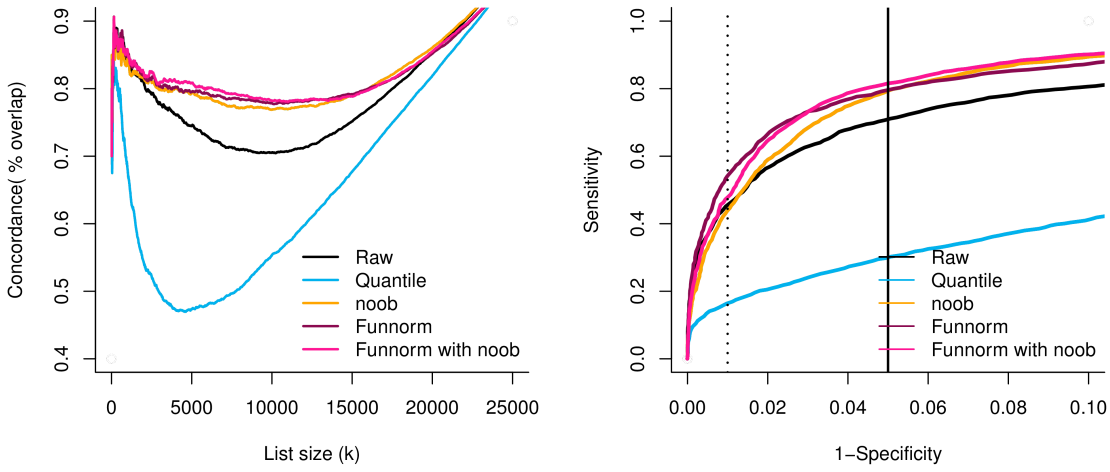
Figure 2.5: Improvements in replication for the TCGA KIRC dataset. (a) ROC curves for replication between a discovery and a validation dataset. The validation dataset was constructed to show *in silico* batch effects. (b) Concordance plots between an additional cohort assayed on the 27k array and the validation dataset. Additional normalization methods assessed in Figure 2.16. Functional normalization shows a high degree of concordance between datasets.

curves are shown in Figure 2.5a, with additional methods in Figure 2.16a,b.Functional normalization and noob are best and do equally well. Again, the gold-standard set of probes that is used to measure performance in these ROC curves differs between normalization methods, and hence these ROC curves reflect the degree of consistency between experiments within each method.

To further compare the quality of the DMPs found by the different methods, we used an additional dataset from TCGA where the same cancer was assayed with the Illumina 27k platform (see Methods). We focused on the 25,978 CpG sites that were assayed on both platforms and asked about the size of the overlap for the top k DMPs. For the validation cohort, with the most unwanted variation, this is depicted in Figure 2.5b and Figure 2.16c for additional methods; for the discovery cohort, with least unwanted variation, results are presented in Figure 2.16d. Functional normalization, together with noob, shows the best concordance in the presence of unwanted variation in the 450k data (the validation cohort) and is comparable to no normalization in the discovery cohort.

2.3.5 Funnorm preserves subtype heterogeneity in tumor samples (TCGA-AML datasets)

To measure how good our normalization method is at preserving biological variation among heterogeneous samples while removing technical biases, we use 192 acute



(a) **(b)**

Figure 2.6: Improvements in replication of tumor subtype heterogeneity. In the AML dataset from TCGA, the same samples have been assayed on 450k and 27k arrays. (a) Concordance plots between results from the 450k array and the 27k array. (b) ROC curves for the 450k data, using the results from the 27k data as gold standard.

myeloid leukemia samples from TCGA for which every sample has been assayed on both the 27K and the 450K platforms (see Methods). These two platforms assay 25,978 CpGs in common (but note the probe design changes between array types), and we can therefore assess the degree of agreement between measurements of the same sample on two different platforms, assayed at different time points. The 450k data appears to be affected by batch and dye bias, see Figure 2.17.

Each sample was classified by TCGA according to the French-American-British (FAB) classification scheme,[41] which proposes 8 tumor subtypes, and methylation differences can be expected between the subtypes.[42,43] Using data from the 27k

arrays, we identified the top k DMPs which distinguish the 8 subtypes. In this case, we are assessing the agreement of subtype variability, as opposed to cancer-normal differences. The analysis of the 27k data uses unnormalized data but adjusts for sample batch in the model (see Methods). Using data from the 450k arrays, we first processed the data using the relevant method, and next identified the top k DMPs between the 8 subtypes. The analysis of the 450k data does not include sample batch in the model, which allows us to see how well the different normalization methods remove technical artifacts introduced by batch differences. While both of the analyses are conducted on the full set of CpGs, we focus on the CpGs common between the two platforms and ask "what is the degree of agreement between the top k DMPs identified using the two different platforms". Figure 2.6a shows that functional normalization and noob outperforms both quantile normalization and raw normalization for all values of k, and functional normalization is marginally better than noob for some values of k. Figure 2.18a shows the results for additional methods. We can also compare the two datasets using ROC curves, with the results from the 27k data as gold standard (Figure 2.6b and Figure 2.18b). As DMPs for the 27k data we used the 5,451 CpGs that demonstrate an estimated false-discovery rate less than 5%. On the ROC curve functional normalization outperforms noob, Quantile and Raw data for the full range of specificity.

2.3.6 Replication between experiments with small changes

To measure the performance of functional normalization in a setting where there are no global changes in methylation, we used the Ontario-Blood dataset which assays lymphocytes from individuals with and without colon cancer. We expect a very small, if any, impact of colon cancer on the blood methylome. As above, we selected cases and controls to form discovery and validation cohorts, and we introduced *in silico* unwanted variation that confounds case-control differences in the validation dataset only (see Methods). The discovery and validation datasets contain respectively 283 and 339 samples. For $k = 100$ loci, both functional and quantile normalization show good agreement between discovery and validation datasets, wheres noob and Raw data show an agreement which is not better than a random selection of probes (Figure 2.7a, Figure 2.19a).

2.3.7 Funnorm improves X and Y chromosomes probes prediction in blood samples

As suggested previously,[16] one can benchmark performance by identifying DMPs associated with sex. One copy of the X chromosome is inactivated and methylated in females, and the Y chromosome is absent. On the 450k array, 11,232 and 416 probes

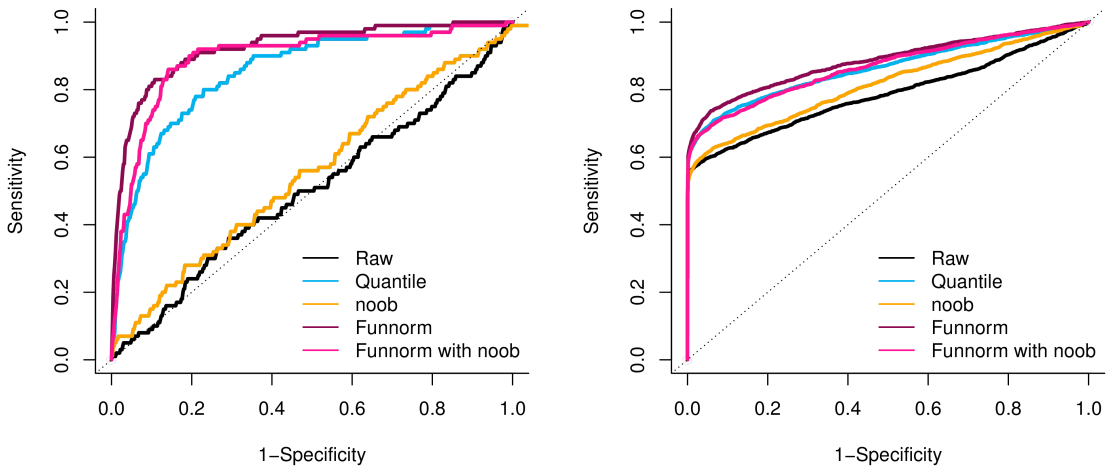


Figure 2.7: Performance Improvements on blood samples dataset. (a) An ROC curve for replication of case-control differences between blood samples from colon cancer patients and blood samples from normal individuals, the Ontario-Blood dataset. The validation dataset was constructed to show an *in silico* batch effect. (b) An ROC curve for identification of probes on the sex chromosomes for the Ontario-Sex dataset. Sex is confounded by an *in silico* batch effect. Both evaluations show good performance of functional normalization.

are annotated to be on the X and Y chromosomes, respectively. For this analysis it is sensible to remove regions of the autosomes which are similar to the sex chromosomes to avoid artificial false positives that are independent of the normalization step. We therefore remove a set of 30,969 probes which have been shown to cross-hybridize between genomic regions.[44] Because some genes have been shown to escape X inactivation,[45] we only consider genes for which the X-inactivation status is known to ensure an unbiased sex prediction (see Methods).

We introduced *in silico* unwanted variation by selecting 101 males and 105 females from different plates (see Methods), thereby confounding plate with sex. Results show that functional normalization performs well (Figure 2.7b, Figure 2.19b).

2.3.8 Funnorm reduces technical variability

From the Ontario-Replicates lymphocyte dataset (see Methods), we have 19 individuals assayed in technical triplicates dispersed among 51 different chips. To test the performance of each method to remove technical variation, we calculated the probe-specific variance within each triplicate, and averaged the variances across the 19 triplicates. Figure 2.8 presents box plots of these averaged probe variances of the all methods. All normalization methods improve on Raw data, and functional normalization is in the top 3 of the normalization methods. Dasen in particular does well on this benchmark, which shows that improvements in reducing technical variation do not necessarily lead to similar improvements in the ability to replicate associations.

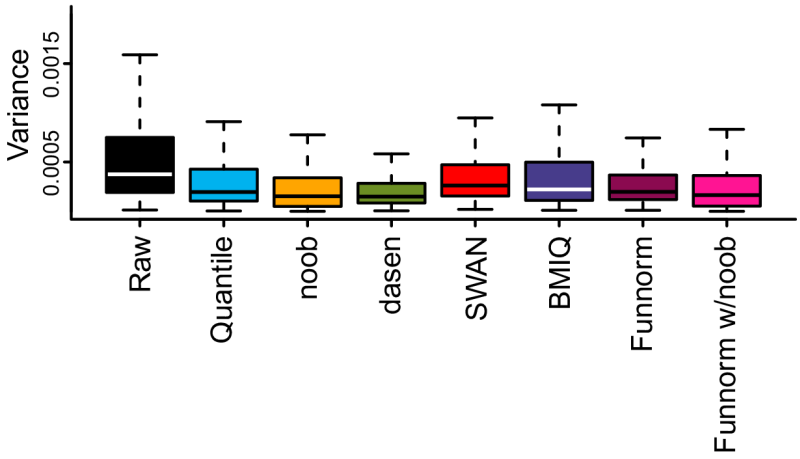


Figure 2.8: Variance across technical triplicates. Boxplots of the probe-specific variances estimated across 19 individuals assayed in technical triplicates. All normalization methods improve upon Raw data, and Funnorm performs well.

Each 450k array is part of a slide of 12 arrays, arranged in 2 columns and 6 rows (see Figure 2.9). Figure 2.9a-c shows an effect of column and row position on quantiles of the beta value distribution, across several slides. This effect is not present in all quantiles of the beta distribution, and it depends on the dataset which quantiles are affected. Figure 2.9d-f shows that functional normalization corrects for this spatial artifact.

2.3.9 Number of principal components

As described above, we recommend using functional normalization with the number of principal components set to $m = 2$. Figure 2.10 shows the impact of varying the number of principal components on various performance measures we have used

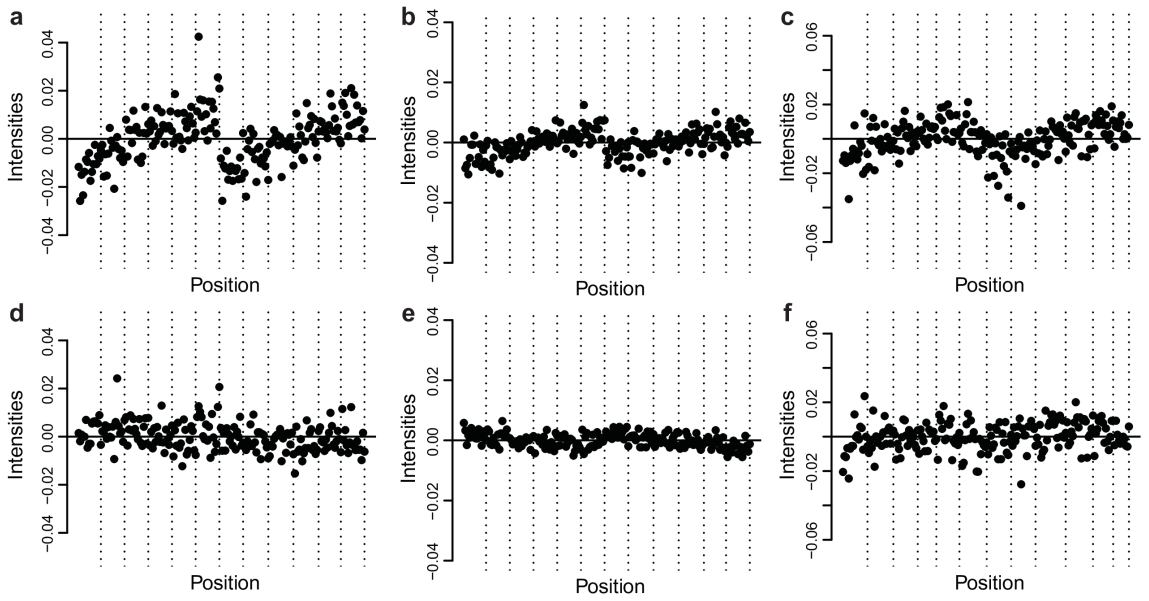


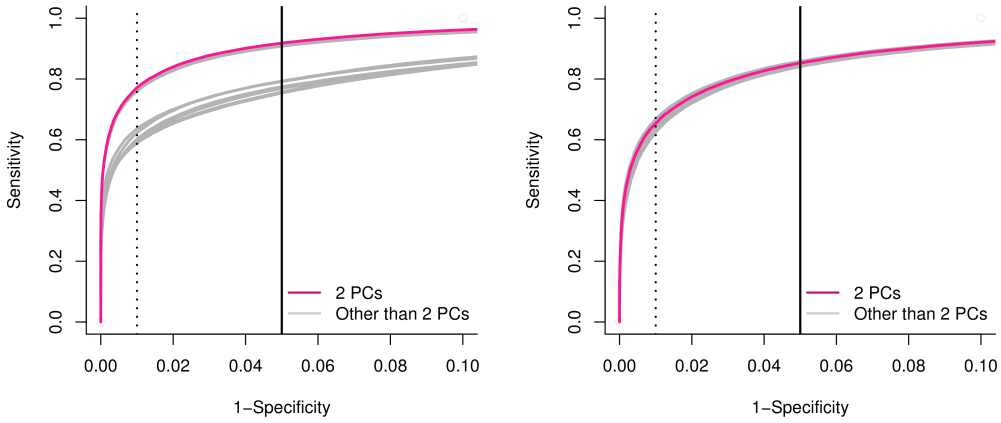
Figure 2.9: Spatial location affects overall methylation. Quantiles of the beta distributions adjusted for a slide effect. The 12 vertical stripes are ordered as rows 1-6 in column 1 followed by rows 1-6 in column 2. (a) 10^{th} percentile for Type II probes for the unnormalized AML dataset. (b) 15^{th} percentile for Type I probes for the unnormalized AML dataset. (c) 85^{th} percentile for Type II probes for the unnormalized Ontario-Lympho dataset. Panel (a-c) show that the top of the slide has a different beta distribution from the bottom. (d-f) Like (a-c) but after functional normalization, which corrects this spatial artifact.

throughout, and shows that $m = 2$ is a good choice for the datasets we have analyzed.. It is outperformed by $m = 6$ in the analysis of the KIRC data and by $m = 3$ in the analysis of the AML data, but these choices perform worse in the analysis of the Ontario-EBV data. While $m = 2$ is a good choice across datasets, we leave m to be a user-settable parameter in the implementation of the algorithm. This analysis assumes we use the same m for the analysis of both the discovery and validation dataset. We do this to prevent overfitting and to construct an algorithm with no user input. It is possible to obtain better ROC curves by letting the choice of m vary between discovery and validation, because one dataset is confounded by batch and the other is not.

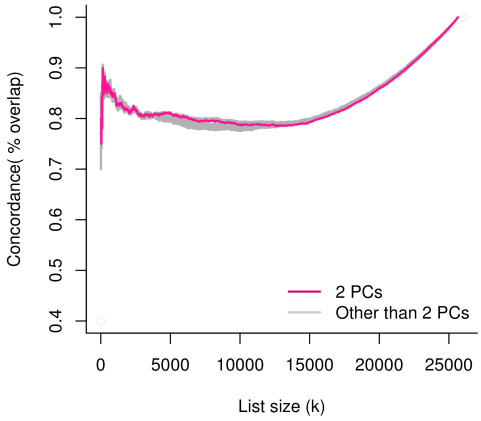
2.3.10 Comparison to batch effect removal tools

Batch effects are often considered to be unwanted variation remaining after an unsupervised normalization. In the previous assessments, we have comprehensively compared functional normalization to existing normalization methods and have shown great performance in the presence of unwanted variation. While functional normalization is an unsupervised normalization procedure, we were interested in comparing its performance to supervised normalization methods such as SVA,[26,27] RUV[29] and ComBat.[28] We adapted RUV to the 450k array (see Methods) and used reference implementations for the other two methods.[46]

We applied these three batch removal tools to all datasets analyzed previously. We



(a) Ontario-EBV **(b)** TCGA KIRC



(c) TCGA AML

Figure 2.10: The impact of the number of principal components. (a) Like Figure 2.4a. (b) Like Figure 2.5a. (c) Like Figure 2.6a, but showing the difference between using $m = 2$ components and other choices of $m \leq 10$.

let SVA estimate the number of surrogate variables, and allowed this estimation to be done separately on the discovery and the validation datasets, which allowed for the best possible performance by the algorithm. For RUV, we selected negative control probes on the array as "negative genes" and probes mapping to the X and Y chromosomes as "positive genes" in the language of RUV (see Methods for details). These negative and positive genes were used to select the number of unwanted factors, as per the recommendations in Gagnon-Bartsch and Speed.[29] Figure 2.11 compares the three methods to functional normalization and Raw data for our evaluation datasets. The three methods have the greatest difficulty with the TCGA AML and the Ontario-Blood datasets compared to functional normalization. Functional normalization is still a top contender for the Ontario-EBV and the TCGA KIRC datasets, although RUV does outperform functional normalization slightly on Ontario-EBV. This shows that the unsupervised functional normalization outperforms these three supervised normalization methods on multiple datasets.

2.3.11 The effect of normalization strategy on effect size estimates

To assess the impact of normalization on the estimated effect sizes, we computed estimated methylation differences on the Beta scale between cases and controls for the Ontario-EBV and KIRC datasets. Figure 2.12 shows the distribution of effect sizes

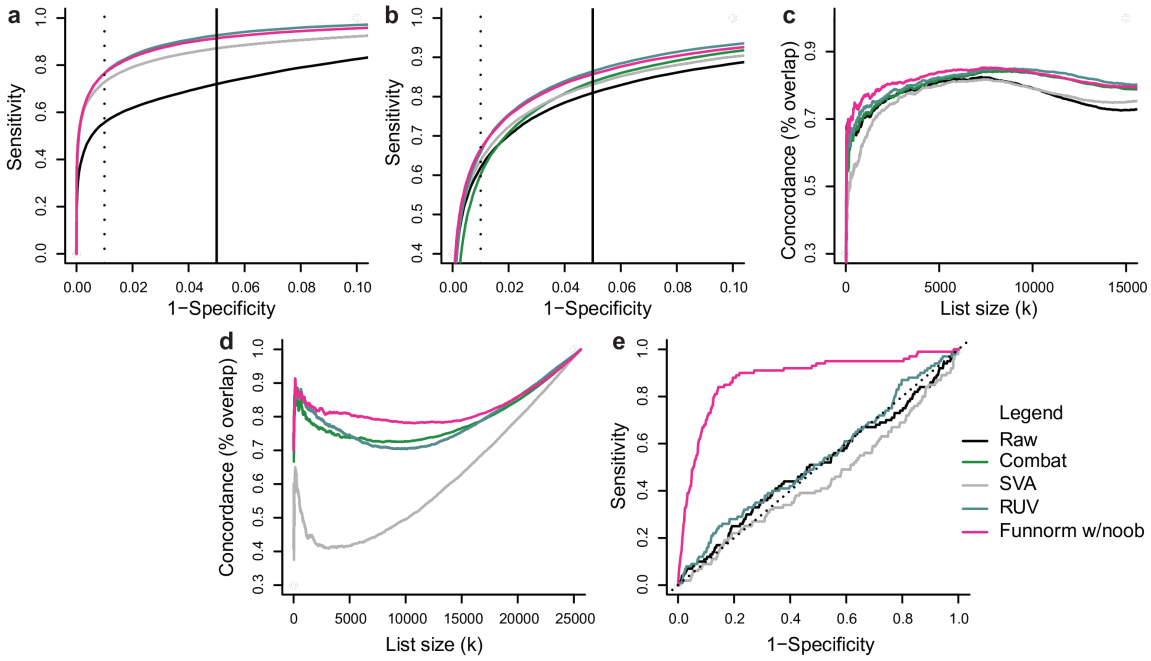


Figure 2.11: Comparison to batch effect removal tools SVA, RUV and ComBat (a) Like Figure 2.4a; an ROC curve for the Ontario-EBV dataset. (b) Like Figure 2.5a, an ROC curve for the TCGA KIRC dataset. (c) Like Figure 2.5b, a concordance curve between the validation cohort from 450k data and the 27k data for TCGA KIRC dataset. (d) Like Figure 2.6a, a concordance plots between results from the 450k array and the 27k array for TCGA AML dataset. (e) Like Figure 2.7a, an ROC curve for Ontario-Blood dataset.

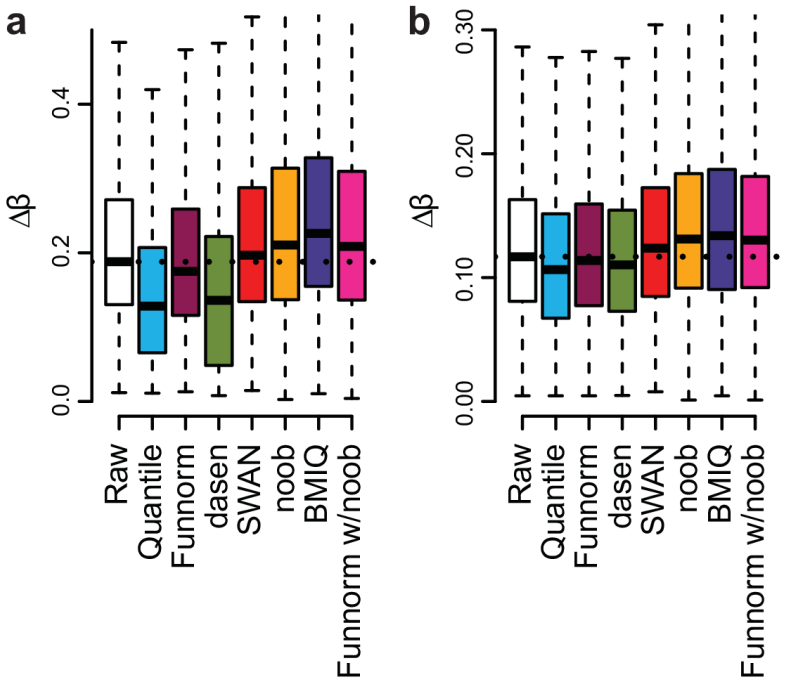


Figure 2.12: Effect size of the top replicated loci. Boxplots represent the effect sizes for the top k loci from the discovery cohort that are replicated in the validation cohort. The effect size is measured as the difference on the Beta value scale between the two treatment group means. (a) Boxplots for the top $k = 100,000$ loci replicated in the Ontario-EBV dataset. (b) Boxplots for the top $k = 100,000$ loci replicated in the TCGA KIRC dataset.

for the top loci in the discovery datasets that are replicated in the validation datasets.

The impact of normalization method on these distributions is dataset dependent.

2.3.12 The performance of Funnorm for smaller sample sizes

To assess the performance of Funnorm with small sample sizes, we repeated the analysis of the Ontario-EBV dataset with different sample sizes by randomly subsampling

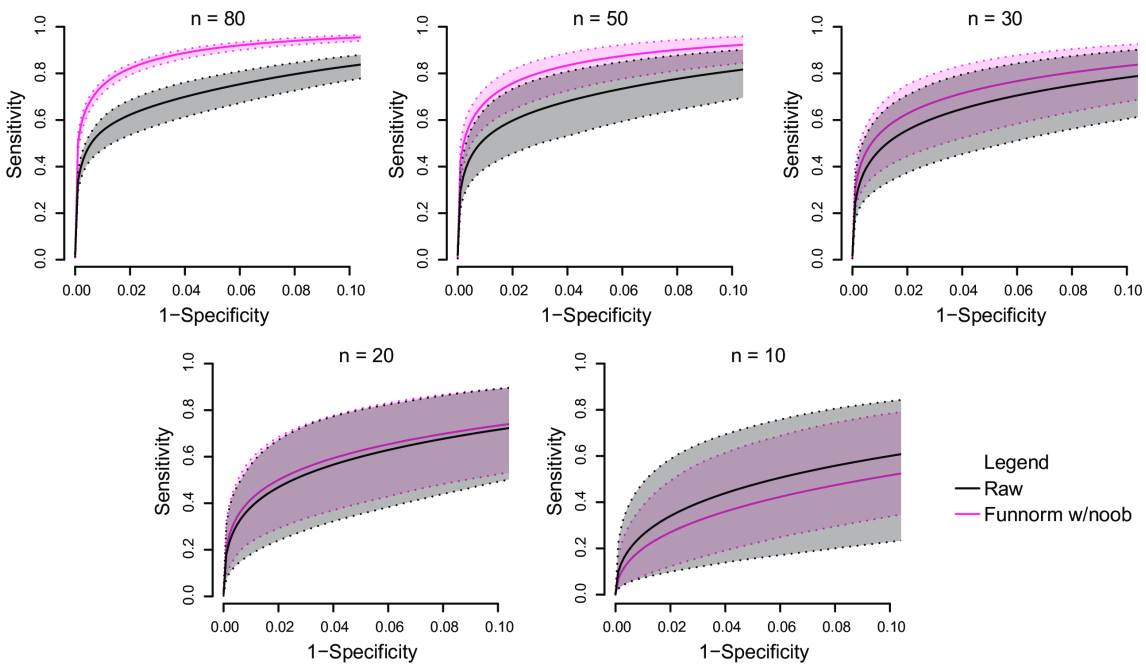


Figure 2.13: Sample size simulation for the Ontario-EBV dataset. Partial discovery-validation ROC curves for the Ontario-EBV dataset similar to Figure 2.4a but for random subsamples of different sizes $n = 10, 20, 30, 50, 80$. Each solid line represents the mean of the ROC results for $B = 100$ subsamples of size n. The dotted lines represent the 0.025 and 0.975 percentiles.

an equal number of arrays from the two treatment groups multiple times. For instance, for sample size $n = 30$, we randomly drew 15 lymphocyte samples and 15 EBV-transformed samples. We repeated the subsampling $B = 100$ times and calculated 100 discovery-validation ROC curves. Figure 2.13 shows the mean ROC curve together with the 0.025 and 0.975 percentiles for both the raw data and the data normalized with Funnom with noob, for different sample sizes. At a sample size of 20, Funnorm very slightly outperforms Raw and Funnorm improves on Raw with sample sizes $n \geq 30$.

2.4 Discussion

We have presented functional normalization, an extension of quantile normalization, and have adapted this method to Illumina 450k methylation microarrays. We have shown that this method is especially valuable for normalizing large-scale studies where we expect substantial global differences in methylation, such as in cancer studies or when comparing between tissues, and when the goal is to perform inference at the probe level. Although an unsupervised normalization method, functional normalization is robust in the presence of a batch effect, and performs better than the three batch removal tools ComBat, SVA and RUV on our assessment datasets. This method fills a critical void in the analysis of DNA methylation arrays.

We have evaluated the performance of our method on a number of large scale cancer studies. Critically, we define a successful normalization strategy as one that enhances the ability to reliably detect associations between methylation levels and phenotypes of interest, across multiple experiments. Various other metrics for assessing the performance of normalization methods have been used in the literature on preprocessing methods for Illumina 450k arrays. These metrics include assessing variability between technical replicates,[14,15,17,18,47] and comparing methylation levels to an external gold standard measurement such as bisulfite sequencing.[12,15,18] We argue that a method which yields unbiased and precise estimates of methylation in a single sample does not necessarily lead to improvements in estimating the differences

between samples, yet the latter is the relevant end-goal for any scientific investigation. This is a consequence of the well-known bias-variance trade-off.[48] An example of this trade-off for microarray normalization is the performance of the RMA method[49] for analysis of Affymetrix gene expression microarrays. This method introduces bias into the estimation of fold-changes for differentially expressed genes; however, this bias is offset by a massive reduction in variance for non-differentially expressed genes, leading to the method's proven performance. Regarding reducing technical variation, we show in Figure 2.8 that methods which show the greatest reduction in technical variation does not necessarily have the best ability to replicate findings, and caution the use of this assessment for normalization performance.

In our comparisons, we have separately normalized the discovery and the validation dataset, to mimic replication across different experiments. We have shown that functional normalization was always amongst the top performing methods, whereas other normalization methods tended to perform well on some, but not all, of our test datasets. As suggested by Dedeurwaerder *et al.*,[18] our benchmarks showed the importance of comparing performance to Raw data, which outperformed (using our metrics) some of the existing normalization methods. In several datasets, we have observed that the within-array normalization methods SWAN and BMIQ had very modest performance compared to Raw data and between-array normalization methods. This suggests that within-array normalization methods does not lead to improvements in the ability to replicate findings between experiments.

Our closest competitor is noob,,[17] which includes both a background correction and a dye-bias equalization. We outperformed noob substantially in the Ontario-Blood and Ontario-Sex datasets and we performed slightly better on the TCGA AML dataset. The best performance was obtained by using functional normalization after the noob procedure.

Our method relies on the fact that control probes carry information about unwanted variation from a technical source. This idea was also used by Gagnon-Bartsch and Speed[29] to design the batch removal tool RUV. As discussed in the Results section, the RUV method is tightly integrated with a specific statistical model, requires the specification of the experimental design, and cannot readily accommodate regional methods[13,37,38] nor clustering. In contrast, functional normalization is completely unsupervised and returns a corrected data matrix which can be used as input into any type of downstream analysis, such as clustering or regional methods. Batch effects are often considered to be unwanted variation remaining after an unsupervised normalization, and we conclude that functional normalization removes a greater amount of unwanted variation in the preprocessing step. It is interesting that this is archived merely by correcting the marginal densities.

However, control probes cannot measure unwanted variation arising from factors representing variation present in the samples themselves, such as cell type heterogeneity, which is known to be an important confounder in methylation studies of samples containing mixtures of cell types.[34] This is an example of unwanted varia-

tion from a biological, as opposed to technical, source. Cell type heterogeneity is a particular challenge in EWAS studies of whole blood, but this requires other tools and approaches to address.

Surprisingly, we showed that functional normalization improved on the batch removal tools ComBat, SVA and RUV applied to raw data, in the datasets we have assessed. It is a very strong result that an unsupervised normalization method improves on supervised normalization procedures, which requires the specification of the comparison of interest.

While we have shown that functional normalization performed well in the presence of unwanted variation, we still recommend that any large-scale study considers the application of batch removal tools such as SVA,[26,27] ComBat[28] or RUV[29] after using functional normalization, due to their proven performance and their potential for removing unwanted variation which cannot be measured by control probes. As an example, Jaffe and Irizarry[34] discusses the use of such tools to control for cell type heterogeneity.

The analysis of the Ontario-Blood dataset suggests that functional normalization has potential to improve the analysis in a standard EWAS setting, in which only a small number of differentially methylated loci are expected. However, if only very few probes are expected to change, and if those changes are small, it becomes difficult to evaluate the performance of our normalization method using our criteria of successful replication.

The main ideas of functional normalization can readily be applied to other microarray platforms, including gene expression and miRNA arrays, provided that the platform of interest contains a suitable set of control probes. We expect the method to be particularly useful when applied to data with large anticipated differences between samples.

2.4.1 Reproducibility

A machine readable document detailing our analyses are available at GitHub (`https://github.com/Jfortin1/funnorm_repro/tree/master/repro_document`).

2.5 Supplementary Material

In the following text, we describe exhaustively the transformations that we applied to the control probes to create the control probe summaries that were used throughout the paper as covariates in the functional normalization. Note that we chose the transformations by considering the recommendations made in the GenomeStudio Methylation manual.

- For "Bisulfite Conversion I" probes, 3 probes (C1,C2,C3) are expected to have high signal in the green channel in case the bisulfite conversion reaction was successful, and similarly 3 additional probes (C4,C5,C6) are expected to have high signal in the red channel. We therefore consider these 6 intensities and

take the mean as a single summary value.

- For "Bisulfite Conversion II" probes, 4 probes are expected to have high intensities in the red channel in case the bisulfite conversion reaction was successful. Therefore we consider the mean of these 4 intensities as a single summary value.

- For the "Extension" control probes, 2 probes must be monitored in the red channel (A,T) and 2 probes must be monitored in the green channel (C,G). We consider the 4 raw intensities as output values for a total of 4 summary values.

- For the "Hybridization" probes, the 3 probes have to be monitored in the green channel. We consider the raw intensities as output values, corresponding to low, medium and high hybridization signals, for a total of 3 summary values.

- For the "Staining" probes, we select the green intensity of the probe that is expected to have high intensity in the green channel, and similarly for the probe that is expected to have high intensity in the red channel. This results in 2 summary values.

- For "Non-polymorphic" controls, we consider the 2 probes that are expected to be high in the green channel (C and G) and the two probes that are expected to be high in the red channel (A and T). We consider the 4 raw intensities as output for a total of 4 summary values.

- "Target removal" probes have to be monitored only in the green channel. We

use the raw intensities for the 2 probes as output values for a total of 2 summary values.

- For "Specificity II" probes, we monitor the 3 probes in both the green and red channels, for a total of 6 output values. The green channel is expected to be low and the red channel to be high, so we also form the ration of the mean of the green channel to the mean of the red channel as output value. This results in 7 summary values.

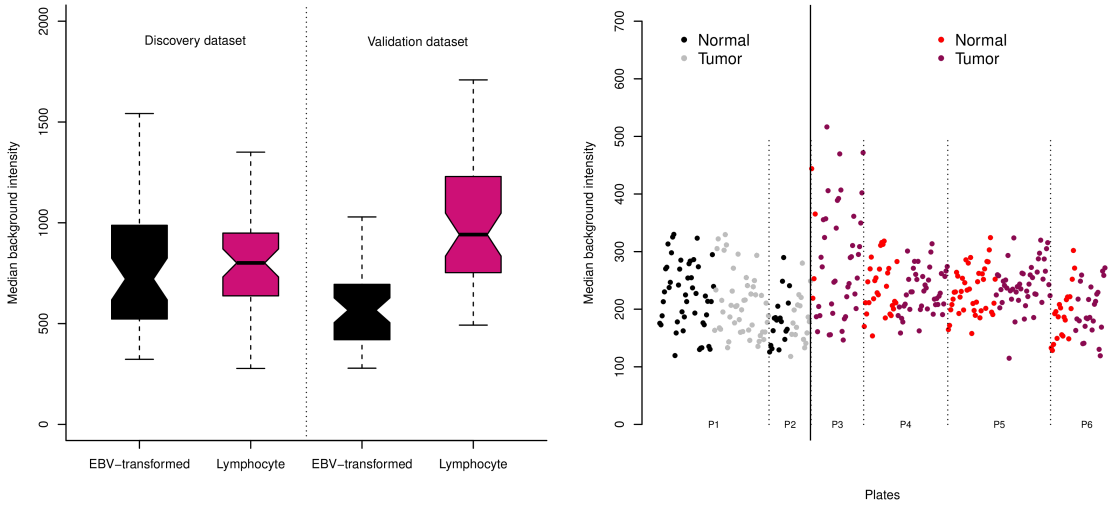
- For "Specificity I" probes, 3 probes are expected to have high signal in the green channel, and 3 different probes are expected to have high signal in the red channel. For each trio, we take the raw intensities in the corresponding "good" channel, giving us 6 output values. For each trio, we also consider the signal-to-noise ratio by taking the ratio of the mean of the 3 intensities measured in the good channel (high signal expected) and the mean of the three intensities in the opposite channel (low signal expected). This gives us 2 ratio summaries. We also consider the mean of these two ratios, giving us 1 additional output value. This results in a total of 9 summary values.

- "Normalization" probes: probes targeting A bases (32) and T bases (61) have to be monitored in the red channel and probes targeting G (32) and C (61) have to be monitored in the green channel. For each type (A,C,T,G), we consider the mean of the intensities in their corresponding channel, that we denote by

$normA$, $normC$, $normT$ and $normG$ (4 output values). Moreover, we consider the ratio $(normC + normG)/(normA + normT)$ as a surrogate for dye bias computed with positive controls (1 additional output value), for a total of 5 summary values.

- For the Out-of-band probes (Oob), we first take the 1st, 50th and 99th percentiles of the 92,596 green intensities (3 output values). Because the variation seen in the green Oob probes is similar to the that of the red Oob probes, we omit the latter. Nevertheless, we consider the ratio of the median of the 92,596 green intensities and the median of the 178,406 red intensities, as a surrogate for dye bias. This results in a total of 4 summary values.

2.6 Supplementary Figures



(a) **(b)**

Figure 2.14: Illustration of *in silico* batch effects. (a) Distribution of background intensity for the Ontario-EBV dataset, showing an *in silico* introduced difference in the validation dataset. (b) Distribution of background intensity for the TCGA KIRC dataset, showing an *in silico* introduced difference in variation in the validation dataset.

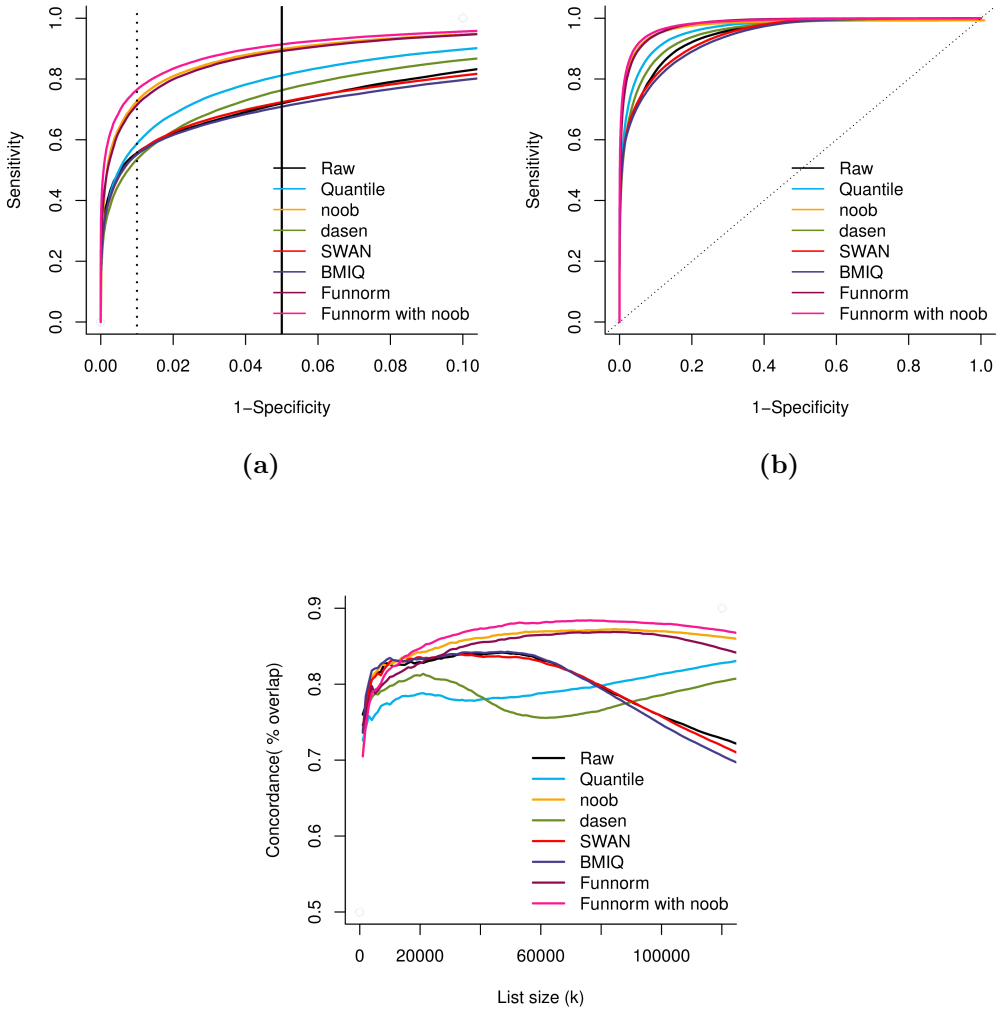


Figure 2.15: Improvements in replication for the EBV dataset, all methods. (a) Like Figure 2.4b, but for all examined normalization methods. (b) Like (a) but for the full range of specificity. (c) Like Figure 2.4c but for all examined normalization methods.

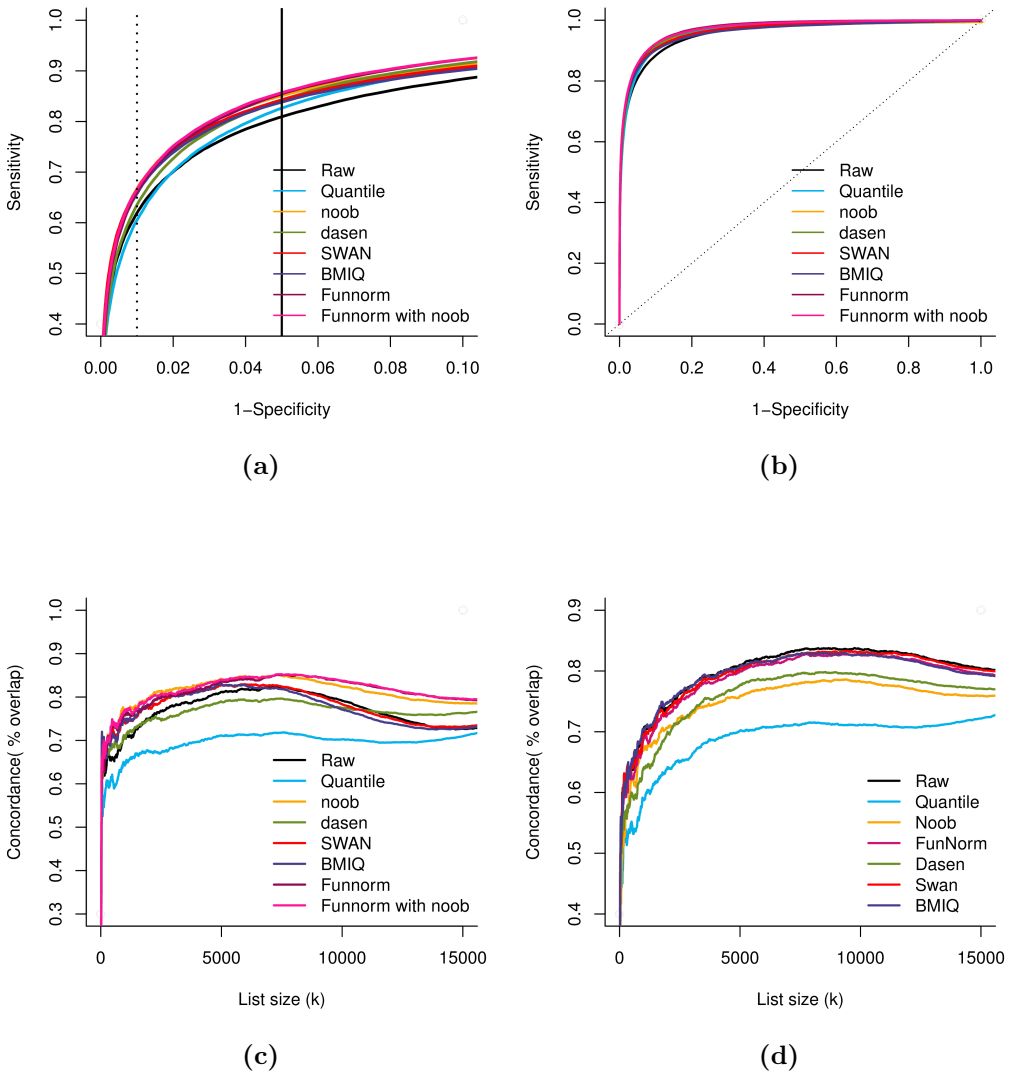
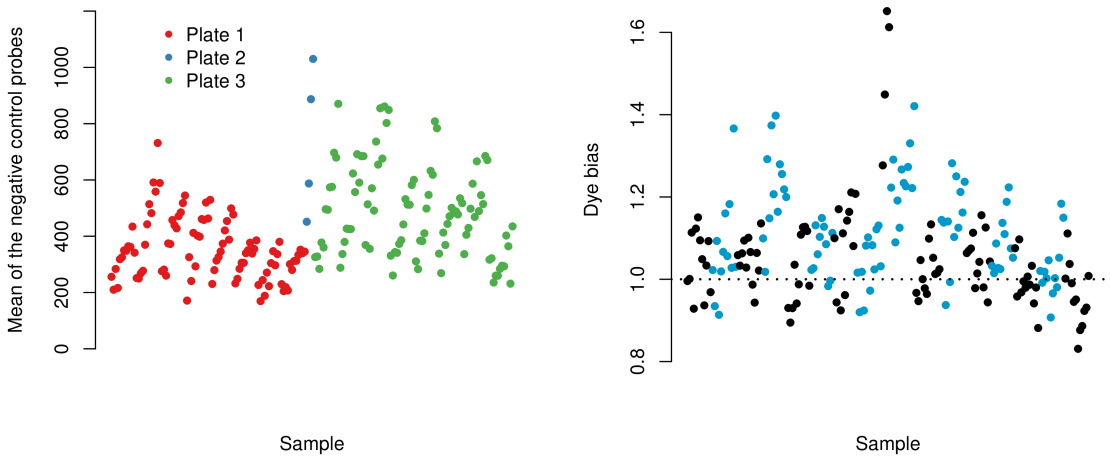
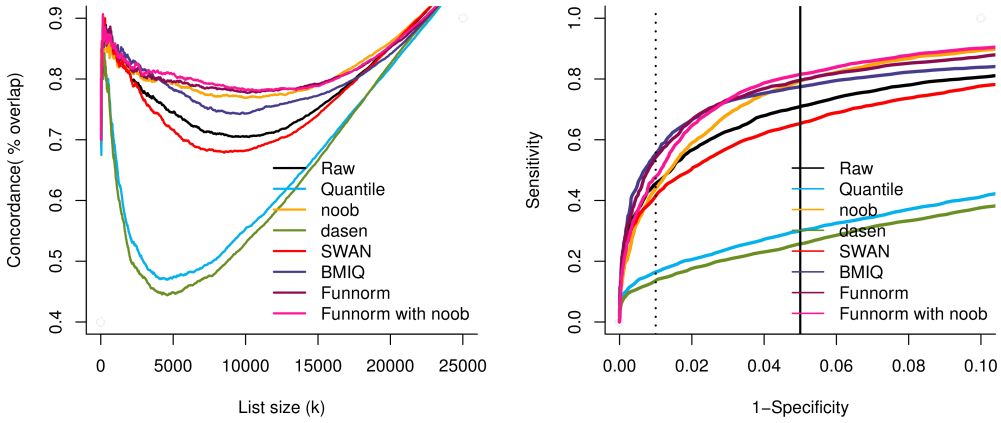


Figure 2.16: Improvements in replication for the TCGA KIRC dataset, all methods. (a) Like Figure 2.5a but for all normalization methods we assess. (b) Like (a) but for the full range of specificity. (c) Like Figure 2.5b but for all normalization methods we assess. (d) Like (c) but comparing to the discovery dataset instead of the validation dataset; the discovery dataset is less affected by unwanted variation.

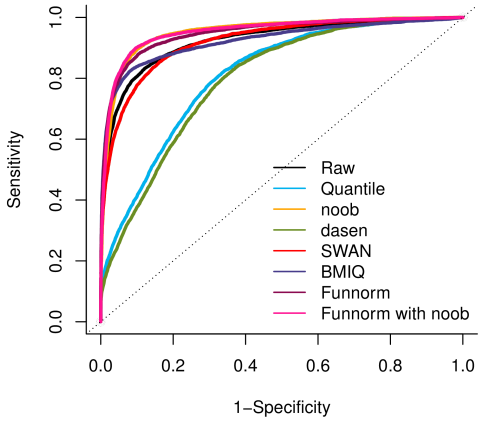


(a) **(b)**

Figure 2.17: Plate effects and dye bias for the AML dataset (a) The means of the negative control probes is correlated with the processing plate (96 samples) indicating that background intensity is affected by batch. (b) We measure dye bias by taking the ratio of the negative control probes in the green channel and the negative control probes in the red channel. A value of 1 means that there is no dye-bias. We plot the dye bias (y-axis) for samples ordered by plate and then by slide (x-axis). The plate order is the same as (a). We use two different alternating colors to differentiate the slides. We observe that dye bias is orthogonal to plate effect and highly slide dependent.

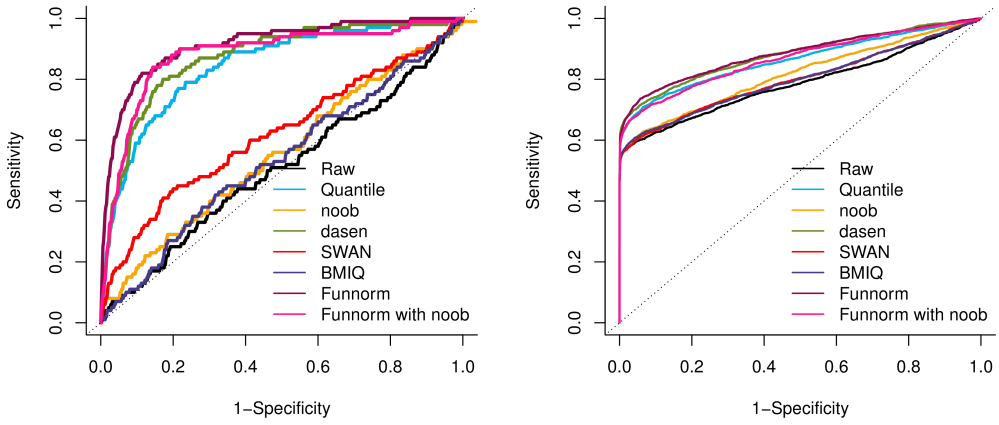


(a) **(b)**



(c)

Figure 2.18: Improvements in replication of tumor subtype heterogeneity. (a) Like Figure 2.6a. (b) Like Figure 2.6b but for all normalization methods. (c) Like (b) but for the full range of specificity



(a) Ontario-Blood **(b)** Ontario-Gender

Figure 2.19: Improvements in blood samples, all methods Like Figure 2.7, but for all normalization methods we assess.

Chapter 3

Normalization of structural MRI images

This chapter describes work under revision at *NeuroImage*, with contributions from co-authors Elizabeth Sweeney, John Muschelli, Ciprian Crainiceanu, and supervised by Dr. Russell Taki Shinohara.

3.1 Introduction

In recent years, there has been an increase in the number of multi-site neuroimaging studies, including the Human Connectome Project (HCP), the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Australian Imaging, Biomarkers and Lifestyle Flagship Study of Aging (AIBL). In structural magnetic resonance imaging

(MRI) studies, larger samples of subjects yield more power to detect structural variations in different subgroups, for example changes in the hippocampal volume associated with Alzheimer's disease (AD) and mild cognitive impairment (MCI). However, because MRI intensities are acquired in arbitrary units, it has often been found that the differences in MRI intensities between scanning parameters and studies are larger than the biological differences observed in these images. For instance,[102] shows that in the ADNI and AIBL studies, which have highly standardized protocols, striking differences in the raw intensities are observed between imaging sites.

Since the raw image intensities are non-comparable across sites and between subjects, intensity normalization is paramount before performing any between-subject comparisons or population-level modeling. The challenge of intensity normalization has been largely addressed in the literature,[102–109] with several methods reviewed in.[110] Recently, a novel intensity normalization method, called White Stripe,[102] was developed to bring raw image intensities to a biologically interpretable intensity scale. The method applies a z-score transformation to the whole brain using parameters estimated from a latent subdistribution of normal-appearing white matter (NAWM). The use of NAWM for normalization makes the method suitable for many studies of brain abnormalities, as in the case of multiple sclerosis (MS) lesions. While the method has been shown to make the white matter (WM) comparable across subjects, it was noted that residual across-subject variability was still present in the grey matter (GM).

In this work, we investigate between-scan technical variability that is left uncorrected by intensity normalization. We show that while common intensity normalization methods successfully correct for global intensity shifts associated with scanner site, substantial between-scan technical variation remains. This technical variation can be due to scanning parameters, scanner manufacturers, scanner field strength, and other factors. We refer to any post-normalization inter-scan variation that is not biological in nature as a "scan effect".

To correct for scan effects, we propose Removal of Artificial Voxel Effect by Linear regression (RAVEL). RAVEL is a tool for removing unwanted variation present after intensity normalization. RAVEL is inspired by the batch effect correction tools SVA[26,27] and RUV[29] used broadly in genomics. In the analysis of gene expression and other genomic data, residual noise after intensity normalization is referred to as batch effects, because experiments are often performed in batches run on different dates. If not accounted for, batch effects have been shown to lead to spurious associations.[85] To make a parallel with brain imaging studies, batch effects are comparable to scan effects, where a single scan plays the role of a batch.

We use the linear model introduced in[26] to decompose the variation of the normalized intensities into a biological component of interest (variation associated with clinical covariates) and an unknown, unwanted variation component to be estimated from the data. The unwanted variation component encapsulates both technical variation and biological variation that is not of interest in the study. We register the

different scans to a common template to allow the use of voxel-wise linear models, and estimate the unwanted variation component from regions of the brain that are not expected to be associated with the clinical covariates of interest. This follows the methodology of the RUV batch effect correction tool[29] which was later discussed in[111] for RNA sequencing. Unlike intensity-normalization methods, RAVEL utilizes all images in the study to leverage information about unwanted variability. Here, we use voxels that are consistently labelled as cerebrospinal fluid (CSF) across subjects as a control region; these voxels are not expected to be associated with disease.[112]

We evaluate the performance of RAVEL using a large subset of the ADNI database consisting of more than 900 subjects. We demonstrate our method by using the T1-weighted (T1-w) images from subjects with AD and MCI, as well as healthy controls. We follow the work of[1] to benchmark RAVEL against two intensity normalization procedures without any scan effect correction: the popular histogram matching algorithm and White Stripe. We focus on showing that RAVEL improves the replicability of the biological findings. Critically, we show that a reduction of technical variation does not result in removing biological variability. Namely, making intensity densities more similar does not necessarily improve sensitivity to biological changes; on the contrary, overmatching of distributions can result in the removal of biologically relevant signal. To show improvement in terms of biological findings, we first demonstrate that the top voxels associated with AD in the RAVEL-corrected dataset are more replicable across independent subsets of subjects. We measure the replicability of the

results by randomly splitting the ADNI dataset into discovery and validation cohorts multiple times. Then, we show that the top voxels associated with AD after RAVEL correction are more enriched for brain regions known to undergo structural changes in AD. Finally, we show that the average hippocampal intensity after RAVEL correction performs better than intensity-normalized-only images in discriminating between AD patients and healthy controls, and between MCI patients and healthy controls. This shows that RAVEL-corrected T1-w intensities are more biologically meaningful than intensity-normalized-only images for group comparisons, and also potentially promising for the development of biomarkers.

Although we apply RAVEL in the context of T1-w MRI of the brain, our method is generalizable to many imaging modalities. Furthermore, the flexibility in the choice of the control voxels makes RAVEL applicable to any disease or pathology.

3.2 Materials and methods

3.2.1 Study population

Our dataset consists of a subset of 917 subjects downloaded from the ADNI database (`adni.loni.usc.edu`). For each subject, we selected a study visit at random. We obtained 506, 184 and 227 subjects from the ADNI, ADNI-2 and ADNI-GO phases, respectively. We present summary statistics of the study population in Table 3.1.

The selected scans were acquired at 83 different imaging sites, with a median number of 10 patients per site. The scans are also well-balanced for disease status across sites. We report the different scanning parameters in Table 3.2.

		Healthy	MCI	AD
	n	261	439	217
	% Female	48	36	47
	Median Age [Q_1,Q_3]	76 [72-79]	75 [70-80]	76 [71-81]
Manufacturer	% GE	42	45	47
	% Philips	11	10	14
	% Siemens	47	45	39
Field Strength	% 1.5T	85	88	88
	% 3T	15	12	12

Table 3.1: Summary statistics of the ADNI sample.

3.2.2 Imaging sequences and preprocessing

We considered T1-w imaging acquired on T1.5 and T3 scanners according to the ADNI standardized protocol.[113] The analysis was performed in R,[114] using the packages oro.nifti,[115] fslr,[116] ANTsR[117] and WhiteStripe.[118]

We applied the N4 inhomogeneity correction algorithm[119] to each image. We nonlinearly registered all T1-w images to a high-resolution T1-w image atlas,[120] using the symmetric diffeomorphic image registration algorithm[121] implemented in the ANTs suite. We used non-linear registration in order to define a brain control region aligned across subjects and to find spatially coherent nuisance patterns for removal.

Manufacturer	Field (T)	Model	Sequence	Coil	FA (°)	TI (ms)	TR (ms)	TE (ms)	Cont	AD	MCI
GE	1.5	GENESIS_SIGNA	MPRAGE	HEAD	8	1000	(10.2,10.4)	4.1	15	15	38
GE	1.5	SIGNA EXCITE	MPRAGE	8HRBRAIN	8	1000	(8.6,9.2)	(3.8,4.1)	71	57	107
GE	1.5	SIGNA EXCITE	MPRAGE	8NVHEAD_A	8	1000	9	3.9	1	0	0
GE	1.5	SIGNA EXCITE	MPRAGE	HEAD	8	(1000,1043)	(9,11)	(3.9,5)	10	13	19
GE	1.5	SIGNA HDx	MPRAGE	8HRBRAIN	8	1000	(8.6,9.2)	(3.8,4.1)	2	8	13
GE	1.5	SIGNA HDx	MPRAGE	HEAD	8	1000	8.6	3.8	0	0	2
GE	1.5	Signa HDxt	IR-SPGR	8HRBRAIN	8	600	9.2	3.9	2	0	6
GE	3	DISCOVERY MR750	IR-SPGR	8HRBRAIN	11	400	7.3	3	0	2	4
GE	3	SIGNA EXCITE	MPRAGE	8HRBRAIN	8	900	7	3	1	0	3
GE	3	Signa HDxt	IR-SPGR	8HRBRAIN	11	400	(7,7.2)	(2.8,3)	7	7	4
Philips	1.5	Achieva	MPRAGE	SENSE-Head-8	8	0	8.6	4	3	4	10
Philips	1.5	Gyroscan Intera	MPRAGE	HEAD	8	0	8.6	4	3	4	2
Philips	1.5	Gyroscan Intera	MPRAGE	SENSE-Head	8	0	8.6	4	2	0	1
Philips	1.5	Intera	MPRAGE	HEAD	8	0	8.6	4	5	4	6
Philips	1.5	Intera	MPRAGE	SENSE-Head	(8,90)	0	(8.5,3000)	(4,12)	13	13	22
Philips	1.5	Intera	MPRAGE	SENSE-Head-6	8	0	8.6	4	1	1	1
Philips	1.5	Intera Achieva	MPRAGE	SENSE-Head-8	8	0	8.6	4	1	0	0
Philips	3	Ingenia	MPRAGE	MULTI COIL	9	0	6.8	3.2	0	3	0
Philips	3	Ingenia	MPRAGE	SENSE-Head	9	0	6.8	3.2	0	1	0
Philips	3	Intera	MPRAGE	SENSE-Head	8	0	6.8	3.2	1	0	0
Philips	3	Intera	MPRAGE	SENSE-Head-8	(8,9)	0	6.8	3.2	0	1	2
Siemens	1.5	Avanto	MPRAGE	PA	8	1000	2400	3.5	24	13	31
Siemens	1.5	Sonata	MPRAGE	HE	8	1000	(2400,3000)	3.5	15	15	28
Siemens	1.5	Sonata	MPRAGE	PA	8	1000	(2400,3000)	3.5	7	8	22
Siemens	1.5	SonataVision	MPRAGE	PA	8	1000	2400	3.5	1	1	5
Siemens	1.5	Symphony	MPRAGE	HE	8	1000	3000	(3.6,3.9)	8	9	18
Siemens	1.5	Symphony	MPRAGE	PA	(2,8)	(0,1000)	(3.9,3000)	(1.3,3.7)	38	25	57
Siemens	3	TrioTim	MPRAGE	PA	9	900	2300	3	25	9	32
Siemens	3	Verio	MPRAGE	PA	9	900	2300	3	5	4	6

Table 3.2: Scanning parameters for the ADNI data subset. We report the different scanning parameters of the ADNI dataset for the images used in our analysis. The different scanning parameters are flip angle in degrees (FA), inversion time in milliseconds (TI), repetition time in milliseconds (TR) and echo time in milliseconds (TE). We report the range of the parameters if more than one value was reported across subjects. For each scanner configuration (row of the table), we report the number of healthy controls (Cont), the number of patients with AD (AD) and the number of patients with MCI (MCI) that we included in our subset of the ADNI database.

To remove extra-cerebral tissue from each scan, we first created a brain mask on the template using the skull-stripping algorithm FSL BET[122] using the fslr package and subsequently applied this resulting brain mask to all N4-corrected and registered images. The preprocessing pipeline is summarized at the top of Figure 3.1.

In addition to the template brain segmentation, we performed a 3-class tissue segmentation by running the FSL FAST segmentation algorithm[123] on the N4-corrected, registered and skull-stripped images, for each subject separately.

3.2.3 RAVEL methodology

The RAVEL correction procedure adapts the linear model introduced in SVA[26,27] to intensity-normalized MRI images. The method removes unwanted variation in the normalized intensities by modeling the residual unwanted variation across subjects. For the optimal performance of RAVEL, we use intensities normalized with White Stripe (see Figure 3.2a). We model the $m \times n$ matrix $\mathbf{V}^{WS}$ of registered and White Stripe-normalized voxel intensities, for m voxels and n subjects, as a decomposition of a biological component of interest and an unwanted component as follows:

$$\mathbf{V}^{WS} = \alpha\mathbf{1}^T + \beta\mathbf{X}^T + \gamma\mathbf{Z}^T + \mathbf{R}. \tag{3.1}$$

where $\alpha\mathbf{1}^T$ represents the average scan in the sample, $\beta\mathbf{X}^T$ accounts for the known clinical covariates of interest (e.g. AD status, age, gender), and $\gamma\mathbf{Z}^T$ accounts for

unknown, unwanted factors. We refer to $\mathbf{V}^{WS}$ as the $m \times n$ matrix of intensities, α as the $m \times 1$ vector of baseline intensities, $\mathbf{X}$ as the $n \times p$ matrix of clinical covariates, β as the $m \times p$ coefficient matrix associated with $\mathbf{X}$, $\mathbf{Z}$ as the $n \times b$ matrix of unwanted factors, γ as the $m \times b$ coefficient matrix associated with $\mathbf{Z}$, and $\mathbf{R}$ as the $m \times n$ matrix of residuals. In this model, α, β, γ and $\mathbf{Z}$ are unknown parameters that need to be estimated from the data. In the case the unwanted factors $\mathbf{Z}$ are known, the problem is reduced to simple linear regression models fit at each voxel separately.

As in RUV,[29] we use a subset of the voxels not associated with disease to estimate the unwanted factors $\mathbf{Z}^T$. We refer to such voxels as "control voxels". An association between CSF intensities and disease status is highly unlikely,[112] and therefore CSF voxels are good candidates for inferring $\mathbf{Z}^T$. We perform a subject-specific tissue segmentation of the T1-w image and choose control voxels as voxels classified as CSF for all subjects in the study. We denote by $\mathbf{V}_c^{WS}$ the subset of $\mathbf{V^{WS}}$ confined to the control voxels. For the control voxels, Equation 3.1 simplifies to

$$\mathbf{V}_c^{WS} = \alpha_c \mathbf{1}^T + \gamma_c \mathbf{Z}^T + \mathbf{R}_c. \tag{3.2}$$

because of the absence of association between the control voxels and $\mathbf{X}$. To estimate $\mathbf{Z}^T$, we perform a singular value decomposition (SVD) of $\mathbf{V}_C^{WS}$ as follows

$$\mathbf{V}_c^{WS} = \mathbf{UDW}^T. \tag{3.3}$$

and define $\hat{\mathbf{Z}}^T$ to be the first b right-singular vectors $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_b\}$ of $\mathbf{W}$. The choice of b is discussed in the next section. Note that for $b = 1$, the estimator $\hat{\mathbf{Z}}^T$ will closely estimate the average CSF intensity for each subject. We obtain the estimates $\hat{\gamma}_i$ in Equation 3.1 by performing a linear regression at each voxel separately, using our estimate of $\mathbf{Z}^T$ in the equation. We define the RAVEL-corrected voxel i for subject j as

$$v_{ij}^{\text{RAVEL}} = v_{ij}^{WS} - \hat{\gamma}_i \hat{\mathbf{Z}}^T$$

where v_{ij}^{WS} is the White Stripe-normalized intensity for the i-th voxel and for the j-th subject. In summary, RAVEL aims to identify patterns of variation in the control voxels across subjects, and then assess the degree to which this variation explains the brain-wide intensity distributions. In practice, this works well if the space spanned by the unwanted factors estimated from the control voxels also spans the unwanted variation space for all voxels. A schematic of the RAVEL method is presented in Figure 3.1.

3.2.4 Estimation of the number of unwanted factors

We select the optimal number of unwanted factors b to include in Equation 3.1 by maximizing the discovery-validation replication rate described in section 3.2.7. Nor-

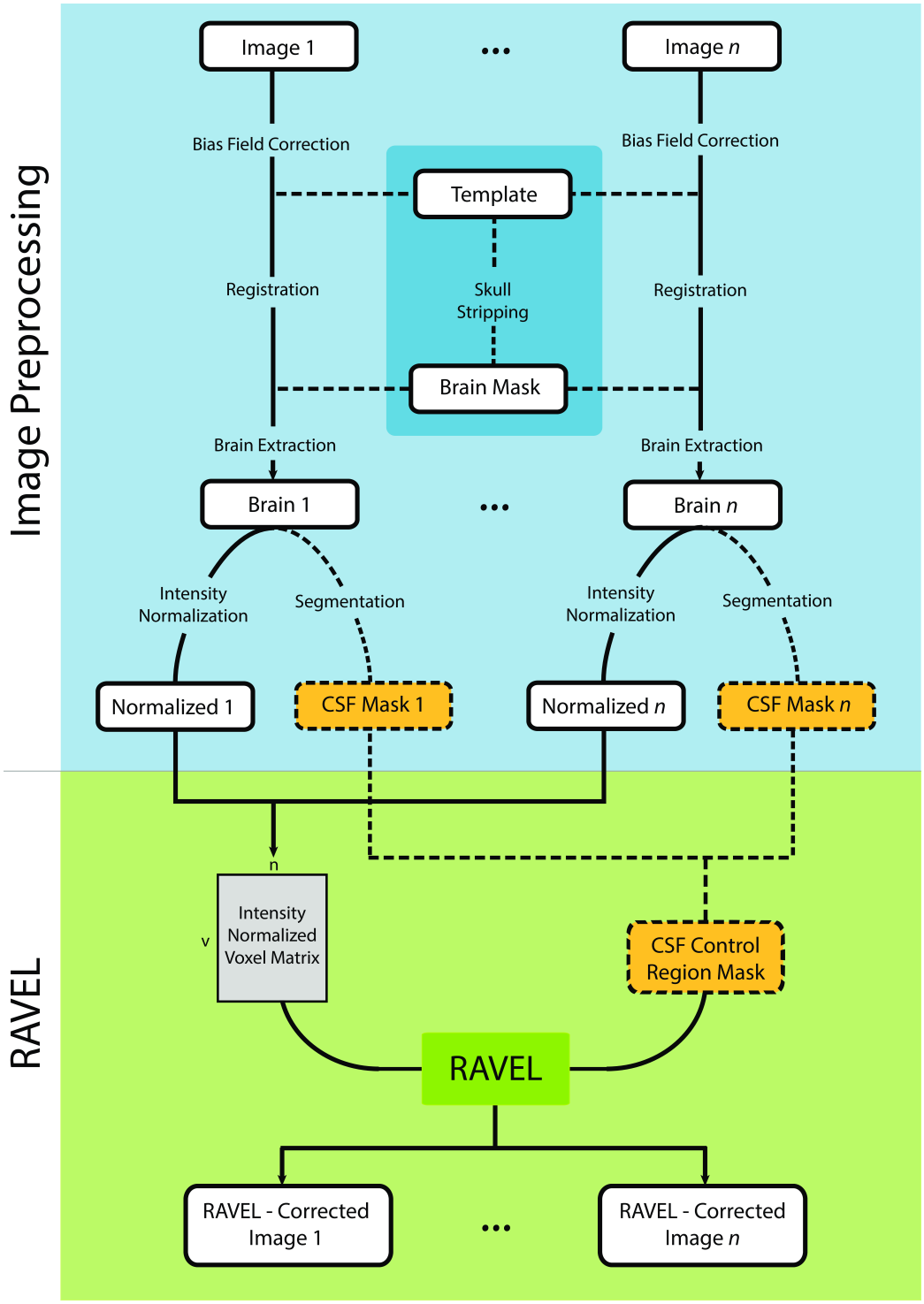


Figure 3.1: Schematic showing the RAVEL pipeline. The steps shown in the blue region are standard preprocessing steps that can be run in parallel. The green region shows the RAVEL algorithm.

malized intensities for which the top voxels associated with disease have better replication between independent experiments are more robust to technical artifacts, like site effect and differences in protocol.

Other approaches have been proposed to select b. Among others,[29] use voxels that are known to be associated with a clinical outcome to optimize b, called "positive control voxels". These authors perform a sensitivity analysis for the parameter b, and b is chosen to optimize the number of positive control voxels that fall into the top voxels associated with the outcome. The downside of using this approach is that positive controls must be identified in advance, which is not possible for discovery studies.

Alternatively, the estimation of b could be done in an unsupervised manner by thresholding the percentage of variance explained by the first b singular vectors. This approach, which is agnostic of the outcome, can potentially provide additional safeguards against overfitting, but could also decrease the performance of RAVEL by adding noise.

3.2.5 Comparison to intensity normalization methods

We compare RAVEL to two intensity normalization procedures without scan effect correction: White Stripe, as implemented in,[118] and the popular histogram matching

method proposed by[103] and further refined in.[110] The histogram matching method matches the histograms of each subject to a reference population histogram using a piecewise linear transformation. We implemented the algorithm in R and we made the code available at `https://github.com/Jfortin1/RAVEL/blob/master/R/hm.R`. For better performance, we removed the background voxels before running the histogram matching algorithm. We used healthy subjects to form a reference population histogram distribution, as described in.[102]

3.2.6 Identification of voxels associated with clinical covariates

Here we describe how we perform the voxel-wise analysis of the intensity distributions. For a clinical covariate x, (e.g. disease status, age, gender), we perform a simple linear regression at each voxel of the T1-w voxel intensity v on the clinical covariate x, and consider the standard Wald t-statistic as a measure of the strength of association. We obtain a t-statistic for each of the m voxels, that is a list $\{t_1, t_2, \ldots, t_m\}$, and we rank the t-statistics in a decreasing order to get a list of rank indices $\{r_1, r_2, \ldots, r_m\}$ where r_j is such that $t_{r_j} = t_{(m-j)}$, the latter being $(m-j)$-th order statistic. For a chosen threshold q, we call the top q ranked voxels the "top voxels associated with x".

3.2.7 Evaluating the replicability of the top voxels associated with AD

To evaluate the replicability of the biological findings, that is the chance that an independent experiment will produce consistent results,[124] we devised a discovery-validation cohorts scheme inspired by.[1] We randomly split the full dataset into two equally sized subsets that we call discovery and validation cohorts, assigning AD and healthy patients equally between the two cohorts.

For each of the two cohorts separately, we perform a differential analysis as described in Section 3.2.6 to obtain two lists of ranked voxels using the differential t-statistics: $\mathbf{r}^{Dis} = \{r_1^{Dis}, r_2^{Dis}, \ldots, r_p^{Dis}\}$ and $\mathbf{r}^{Val} = \{r_1^{Val}, r_2^{Val}, \ldots, r_p^{Val}\}$, for the discovery and validation cohorts respectively. The agreement between the two lists $\mathbf{r}^{Dis}$ and $\mathbf{r}^{Val}$ serves as a measure of replicability. More specifically, we are interested in the agreement of the top-ranked voxels since those are likely more relevant and more representative of a true biological signal. For a given integer k, we look at the proportion of overlap, denoted $O(k)$, of the top k voxels from each list by

$$O(k) = \frac{|\{r_1^{Dis}, r_2^{Dis}, \ldots, r_k^{Dis}\} \cap \{r_1^{Val}, r_2^{Val}, \ldots, r_k^{Val}\}|}{k}$$

A concordance at the top (CAT) plot[125] is a plot showing $O(k)$ for several values of k. To quantify uncertainty of the overlap measure $O(k)$, we repeat the random

discovery-validation cohort splitting one hundred times, and present the mean curve along with a 95% confidence band.

3.2.8 Pseudo-ROC curves and enrichment curves

In this section, we review the methodology behind pseudo-ROC curves[126] and enrichment curves. We use these curves to evaluate the performance of the different normalization and scan effect removal methods by using prior information about structural changes associated with AD. In several neuroimaging studies, prior information about a specific disease allows us to expect a set of voxels to be associated with disease. For instance, a large proportion of the hippocampus and parahippocampal voxels are known to be associated with AD and MCI (see Table 3.3 for references). In the absence of a gold standard, these voxels can play the role of a proxy for a gold standard. We refer to these voxels as a silver standard, that is a gold standard with some contamination.

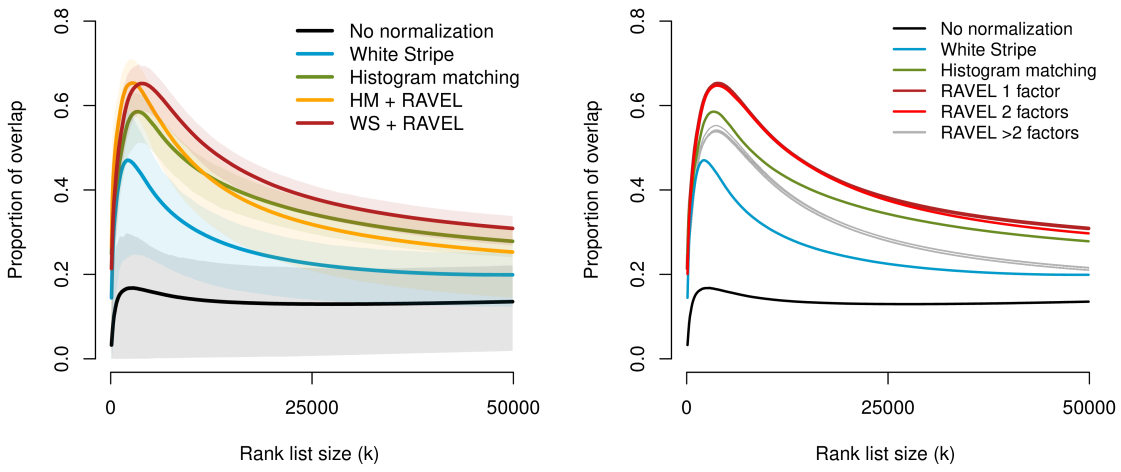
In the context of genomics, silver standards have been previously used to compare the performance of different classification methods[126] and normalization methods.[1,127126] show that receiver operating characteristic (ROC) curves based on a silver standard, called "pseudo-ROC curves", preserve the relative ranking of different classification methods with respect to ROC curves based on a gold standard. A sufficient condition for the validity of the pseudo-ROC curves ranking is that the contamination of the silver standard, with respect to the gold standard, occurs independently of the

misclassification errors of the different methods compared. In the Results section, we use the t-statistics measuring the association of the voxel intensities with AD to classify voxels as either associated with AD or not. To estimate the sensitivity and specificity of each normalization method, we use voxels from 5 regions known to be associated with AD from an extensive search of the literature (see Table 3.3) as a silver standard.

Brain region	References
Hippocampus	128–130
	131–133
	134–136
	137–139
	140–142
	143, 144
Amygdala	145–147
	129, 134, 136
	142, 143, 148
	144, 149
Parahippocampal gyrus	129, 131, 134
	136, 140, 150
Enthorinal region	133, 135, 151
	138, 143
	150, 152
Fornix and S. Terminalis	134, 153, 154

Table 3.3: Brain regions previously reported to undergo a structural change in the progression of AD.

A second approach to benchmark different normalization/scan effect correction methods is to count the number of candidate voxels that fall into the list of the top



(a) **(b)**

Figure 3.2: CAT plots with additional methods. (a) Like Figure 3.5b, but distinguish between RAVEL run on intensities normalized by White Stripe (default) and RAVEL run on intensities normalized by histogram matching. (b) Like Figure 3.5b, but for different numbers of unwanted factors in the RAVEL model. The pink line is for RAVEL with 2 factors, and the grey lines represent RAVEL with 3 to 15 factors. We can observe that the choice of 1 or 2 factors in the RAVEL model optimizes the replication of the voxels associated with AD.

k voxels associated with disease. We refer to the curve that depicts the counts for different values of k as the "enrichment curve".

3.3 Results

We compared RAVEL to three normalization strategies: raw image intensities (no normalization), White Stripe,[102] and histogram matching.[110] We recall that RAVEL correction was performed on the White Stripe-normalized intensities for better per-

formance (see Figure 3.2a).

3.3.1 RAVEL reduces inter-subject variability

We used a subset of the CSF intensities as control voxels to estimate factors of unwanted variation in the RAVEL model. We obtained 9869 CSF control voxels; we recall that a voxel is qualified as a CSF control if it is classified as CSF for all subjects. As expected, the CSF control voxels were located primarily in the center of the ventricles (Figure 3.3a). Maximizing the discovery-validation replication rate, we only kept the first singular vector as the unwanted factor term $\mathbf{Z}^T$ in Equation 3.1, corresponding to $b = 1$ (see Figure 3.2b). Unsurprisingly, the singular vector is highly correlated with the mean CSF intensity for each subject (correlation of 95.7%).

In Figure 3.3b, we depict the coefficient $\hat{\gamma}$ at each voxel. We notice that the distribution of $\hat{\gamma}$ varies across brain tissues, for instance darker red in WM and yellow in CSF. This shows that the method allows an unsupervised tissue-specific normalization. This prevents over-normalization in situations where the technical variation of the CSF intensities is not representative of the variation of other tissues.

In Figure 3.4, we show the histograms of intensities before and after RAVEL correction. The first row shows the unnormalized image histograms and the second row shows the histograms for the images normalized with White Stripe. The last row depicts the histograms for the White Stripe-normalized images with RAVEL correction. In accordance with the findings of,[102] the White Stripe-normalized images

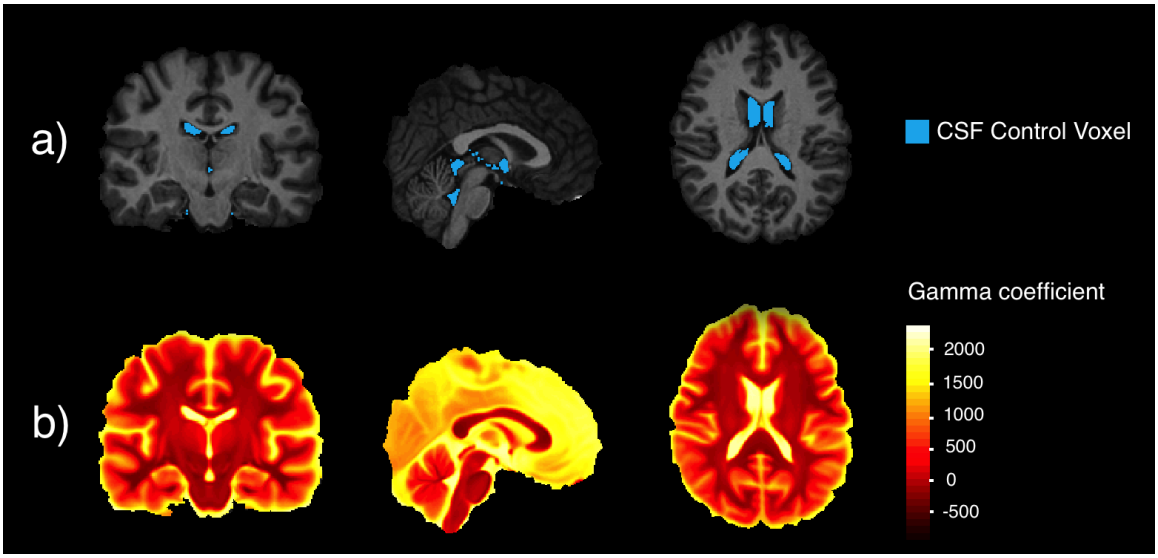


Figure 3.3: Estimation of technical variability using CSF control voxels. (a) The voxels selected in the RAVEL model as control voxels for CSF are shown in blue overlaid on the template; the control voxels were selected as voxels classified as CSF for every subject. (b) Heatmap of the RAVEL coefficient $\hat{\gamma}$ from Equation 3.1 depicted on the template, using $b = 1$ in Equation 3.1. The coefficient depends on the brain tissue, with a high coefficient for voxels in CSF (yellow regions), a moderate coefficient in GM (orange and lighter red) and a low coefficient for WM (darker red).

show good comparability of the WM across subjects. This can be seen by the similar WM densities centered around zero (Figure 3.4 second row, third column). For GM, the White Stripe densities are less clustered and show more variability, which is even more exaggerated for the CSF intensities. This shows that scaling and centering using a NAWM stripe is not enough to make GM and CSF intensities comparable across subjects. This can be explained by differential WM to GM and WM to CSF contrast ratios across images and protocols. In the third row, one can see that RAVEL substantially corrects for the extra variability in CSF and GM intensities that is not accounted for by intensity normalization. RAVEL also preserves the comparability of the WM intensities. The histograms for each tissue class cluster together well and show similar characteristics (mean, scale and range).

The main source of variation in the unnormalized images is from scanning site; on average, 67.8% of the variation in the intensities is explained by scanning site (R^2 averaged across voxels). Interestingly, we observed much less variation explained by scanning site for both intensity-normalized datasets (18% for both White Stripe and histogram matching) and for RAVEL (18%). We randomly permuted the scanning site variable 100 times and obtained a null distribution of the average R^2 with range of $[16.1\%, 16.5\%]$. This implies that after intensity normalization alone, the variability between different sites is close to the within-site variability. However, as shown in Figure 3.4, RAVEL removes additional technical variability in comparison to intensity normalization alone.

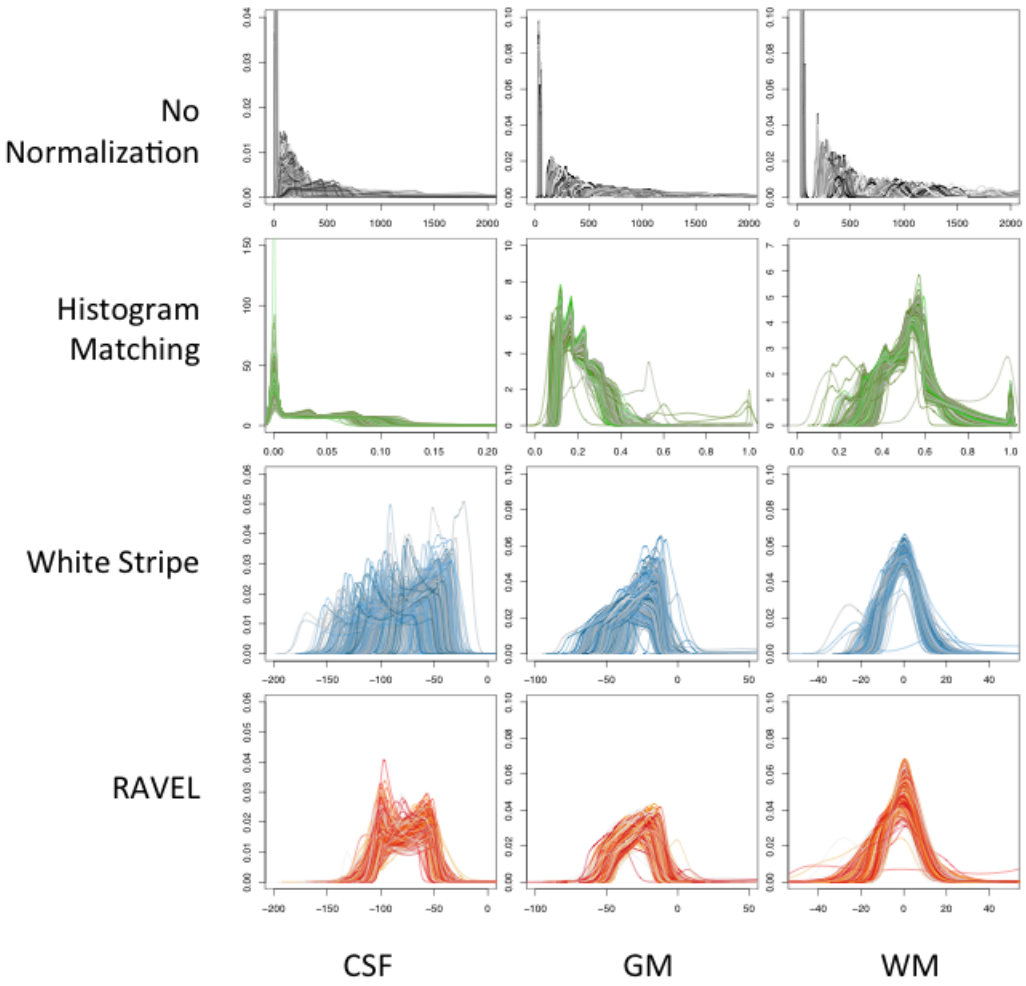


Figure 3.4: Effect of RAVEL on the histograms of intensities. Rows correspond to different preprocessing steps, and columns to different brain tissues. Each curve represents the corresponding histogram of intensities for one subject.

3.3.2 RAVEL improves replicability of large MRI studies

The study of large epigenetic data has shown that the ability to reduce technical variation does not necessarily lead to a better detection of features associated with the outcome of interest.[1,155] A good normalization method should both reduce technical variability and enhance the replicability and robustness of biological findings. Here, we evaluate the performance of RAVEL in terms of estimating brain regions associated with AD.

We randomly split the ADNI dataset into discovery and validation cohorts one hundred times, and we present in Figure 3.5b the mean CAT curves with 95% confidence bands. As expected, the raw images intensities show very poor replication of the results (maximum of 0.17), while RAVEL improves replication of the findings substantially (up to 0.65) upon intensity normalization methods alone.

The replicated voxels fall into regions that are known to be associated with AD. In Figure 3.5a, we show voxels associated with AD that were replicated among the top 50,000 voxels for all random splittings. No normalization led to zero voxels replicated across splittings. This is not surprising since raw image intensities are expressed in arbitrary units. White Stripe replicated 1541 voxels, while histogram matching and RAVEL replicated 3758 and 4897 voxels respectively (Figure 3.5c). In addition, RAVEL is the most powerful method for finding replicated voxels in the hippocampus

and amygdala, two structures known to be associated with AD. The number of replicated voxels for the hippocampus are the following: 0 for no normalization, 396 for White Stripe, 1693 for histogram matching and 2405 for RAVEL. For the amygdala, we obtained the following counts: 0 for no normalization, 323 for White Stripe, 368 for histogram matching and 518 for RAVEL.

White Stripe and histogram matching, by correcting for inter-subject variability in the white matter, substantially increased the number of replicated voxels associated with AD in comparison to no normalization. RAVEL led to a 3-fold increase in the number of replicated voxels with respect to White Stripe. This was achieved by additionally modeling brain-wide unwanted variability using a CSF control region. This is consistent with the idea that while CSF is not interesting on its own with respect to disease, it can be used powerfully to distinguish signal from noise in the entire brain.

3.3.3 RAVEL uncovers known regions associated with AD

The discovery-validation scheme allowed us to evaluate the replicability of the top voxels associated with AD. In the current section, we aim to evaluate the validity of the results by comparing the top voxels to brain regions known to undergo a structural change in the progression of AD. Those structural changes include, among others, GM

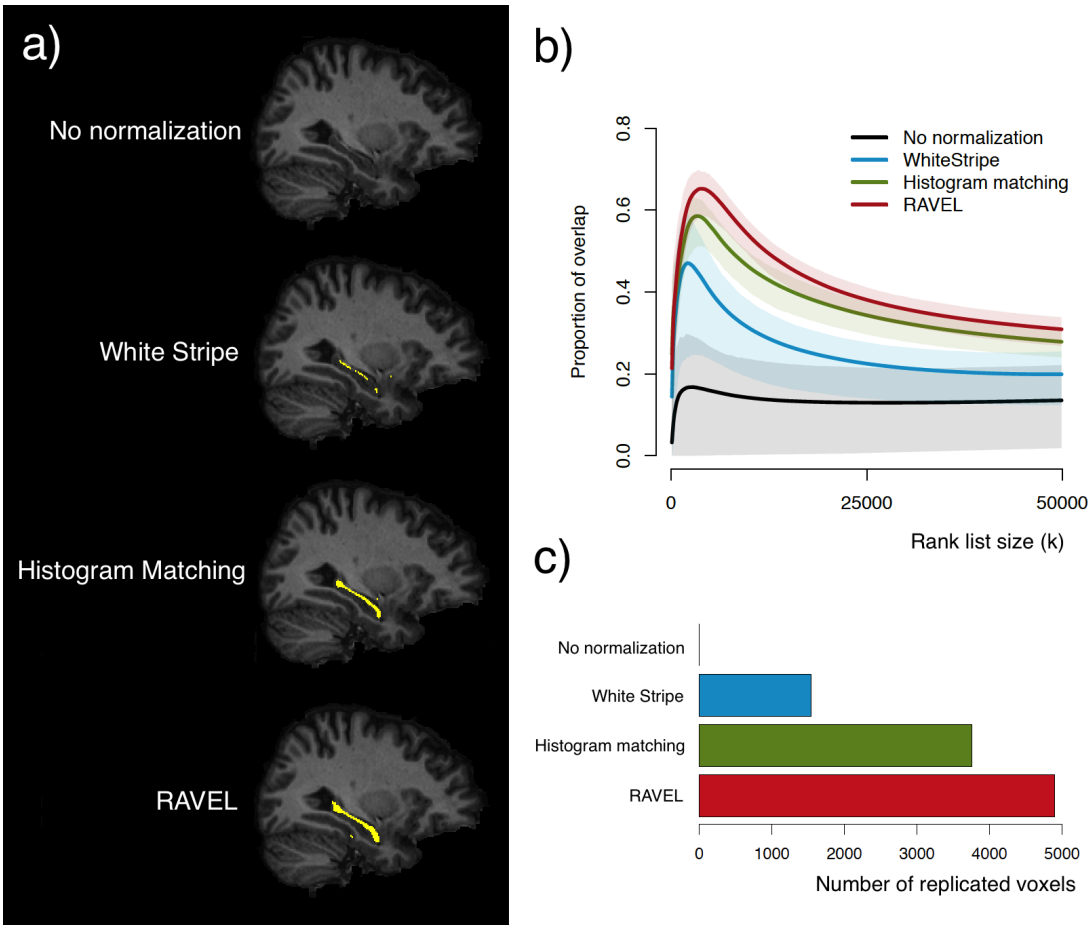
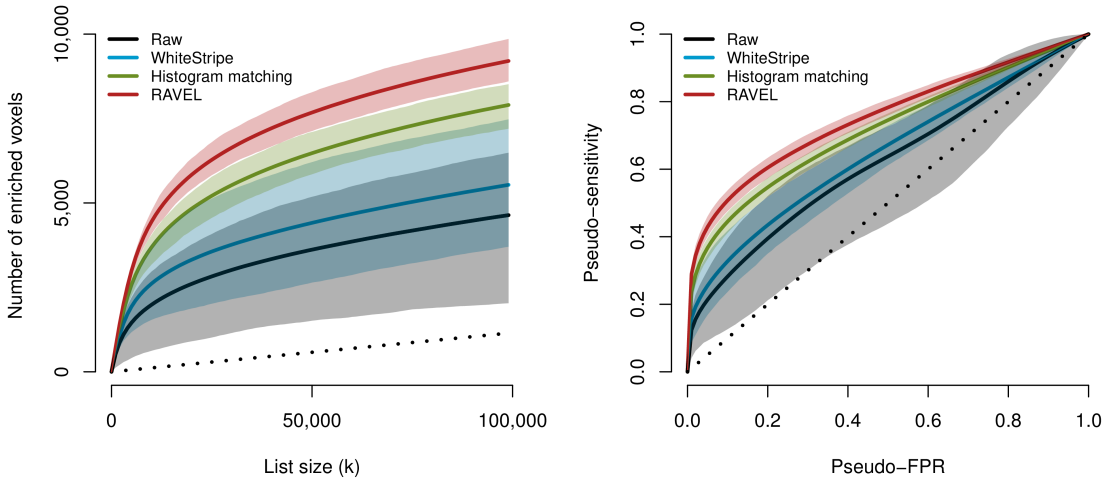


Figure 3.5: RAVEL improves replicability of voxels associated with AD. (a) In template space, we depict in yellow the voxels that are replicated across all random splittings, from the list of the top 50,000 associated with AD. (b) Mean CAT curves for association with AD with 95% confidence bands. (c) Number of voxels replicated for each method in (a). RAVEL shows excellent performance at replicating the discovery of regions of the brain associated with AD.

and WM atrophy, neuronal loss, amyloid senile plaques, loss of fiber tract integrity and tau lesions. In the context of AD, these changes have been described in the hippocampal formation and several parahippocampal structures. The list includes, but is not limited to, the hippocampus, the amygdala, the enthorinal cortex, the fornix, the stria terminalis and the parahippocampal gyrus. Table 3.3 lists several studies that have reported structural changes in these regions.

Using the template parcellation map,[120] we considered 67,983 voxels that are part of the regions listed in Table 3.3. These voxels represent 3.5% of the template and are candidates for association with AD. We use these voxels as a silver standard to evaluate the performance of the different normalization methods and RAVEL. For different values of k, we count the number of the top k voxels associated with AD that are part of the silver standard, which are said to be enriched for the truth. The enrichment curves, depicted in Figure 3.6a (solid lines), show the number of enriched voxels for different values of k, for each normalization method. The dotted line at the bottom represents the number of voxels expected by chance only. To account for variability in the enrichment curves, we nonparametrically bootstrapped with replacement by subject to recalculate the top voxels associated with AD and recompute the curves. The shaded regions of Figure 3.6 represent bootstrapped 95% confidence bands. We observe that RAVEL discovers significantly more voxels that are truly associated with AD than the competing methods. The top voxels associated with RAVEL are also more stable than other methods, as measured by narrower 95%



(a) **(b)**

Figure 3.6: The top voxels associated with AD are enriched for the hippocampus and parahippocampal regions (a) For the top k voxels associated with AD (x-axis), the solid lines display the number of voxels out of the k voxels falling into five structures known to be associated with the progression of AD: the hippocampus, amygdala, enthorinal cortex, fornix and stria terminalis and parahippocampal gyrus. The dotted line represents the number of voxels expected by chance only. The shaded areas represent 95% confidence bands computed using 100 bootstrapped samples. (b) From the t-statistics measuring the association of the voxel intensities with AD, we present the pseudo-ROC curves for classifying a voxel as a member of the five regions described in (a). RAVEL shows significantly better sensitivity and specificity than the other methods for detecting hippocampus and parahippocampal changes associated with AD.

confidence bands. Notably, RAVEL offers a substantial improvement with respect to intensity normalization with White Stripe alone.

Next, we obtained pseudo-ROC curves to measure the specificity and sensitivity of RAVEL for detecting a true association between voxel intensities and AD. In Figure 3.6b, we present the pseudo-ROC curves for classifying voxels as associated with AD or not, using the differential analysis (voxel-wise) t-statistics as a measure of association. The voxels from the regions listed in Table 3.3 are used as a silver standard. As with the enrichment curves, we present bootstrapped 95% confidence bands. RAVEL outperforms histogram matching, White Stripe and raw image intensities for the full range of specificity.

In Figure 3.7, we show in template space the negative log p-value at each voxel for association between the intensities and AD status.

3.3.4 RAVEL-corrected intensities improve prediction of AD and MCI

We investigated the potential use of T1-w RAVEL-corrected intensities as biomarkers for disease identification and progression. We first compared the average hippocampal intensity between AD patients and healthy controls. We used the template parcellation map to identify 9847 voxels labelled as hippocampus. Using the mean intensity of the hippocampus as a score, we classified each subject as either having AD or being

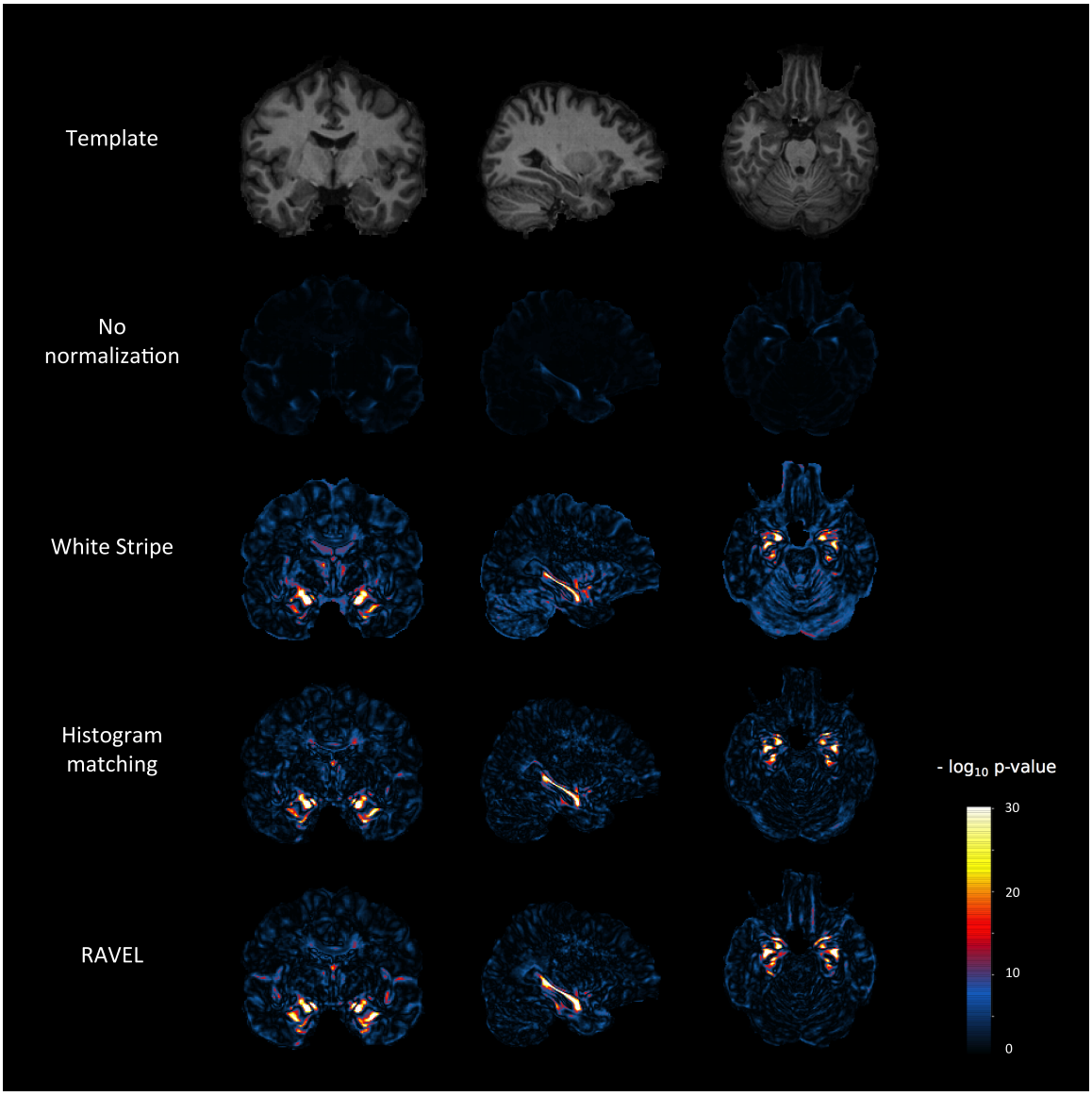


Figure 3.7: Voxel-level p-value maps from AD vs. healthy patient differential analysis. At each voxel, we computed a t-statistic for testing a difference in intensities between AD and healthy patients. For each normalization method, we report the negative log p-values from the t-test. We include at the top of the figure the template for anatomical reference.

healthy, thresholding the scores at different levels. The corresponding ROC curves are presented in Figure 3.8a. We obtained an area under the curve (AUC) of 81.7% for RAVEL (95% CI [77.6, 85.4]), as opposed to 74.9% for histogram matching ([70.4, 79.2]), 64.4% for White Stripe ([58.9, 69.0]), and 57.0% for no normalization ([52.1, 62.0]). We obtained the 95% confidence intervals by bootstrapping the samples with replacement 1000 times. Similarly, we used the average hippocampal intensity to distinguish between MCI patients and healthy controls; the corresponding ROC curves are presented in Figure 3.8b. We obtained an AUC of 67.3% for RAVEL (95% CI [63.1, 71.3]), as opposed to 63.4% for histogram matching ([59.6, 67.7]), 59.0% for White Stripe ([54.8, 63.4]), and 52.9% for no normalization ([48.4, 57.3]). This shows that RAVEL-corrected intensities are more representative of true biological variation than intensity-normalized intensities alone, indicating that the development of biomarkers using MRI studies in many neurological and psychiatric disorders could benefit from the RAVEL scan effect correction tool.

3.4 Discussion

In this work, we have presented the scan effect correction tool RAVEL, to correct for inter-scan unwanted variability in MRI studies that is present after intensity normalization. We have shown that RAVEL, applied after normalizing the intensities with White Stripe, substantially improves the replicability of the regions of the brain found

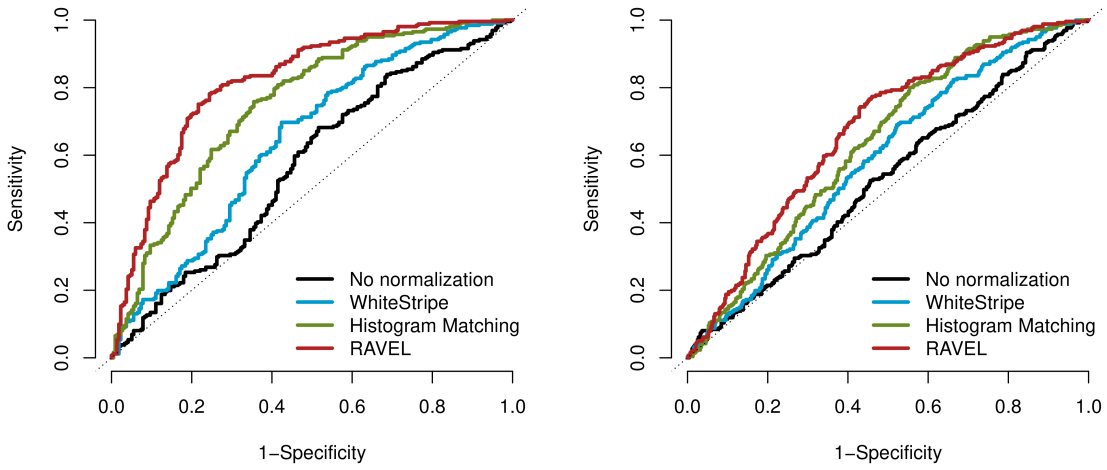


Figure 3.8: RAVEL improves the prediction of AD and MCI. (a) The mean hippocampus intensity was used to predict AD. The AUC is 81.7 % for RAVEL, 74.9% for histogram matching, 64.4% for White Stripe and 57.0% for no normalization, with 95% CIs [77.6, 85.4], [70.4, 79.2], [58.9, 69.0] and [52.1, 62.0] respectively. (b) The mean hippocampus intensity was used to predict MCI. The AUC is 67.3% for RAVEL, 63.4% for histogram matching, 59.0% for White Stripe and 52.9% for no normalization with 95% CIs [63.1, 71.3], [59.6, 67.7], [54.8, 63.4] and [48.4, 57.3] respectively.

to be the most associated with AD. RAVEL, inspired by the batch effect correction tools SVA and RUV, infers the unwanted variation in the images by using regions of the brain that are not associated with disease. After registering all images to a common template, we used voxels that were labelled as CSF for all images as control voxels. We used a linear regression model at each voxel to regress out the variation in the intensities explained by variation in the control CSF voxels intensities. We used an SVD to reduce the dimensionality of the control voxels, and selected the number of components to include in the regression models by maximizing the replication rate of biological findings between independent subsets of the data.

We have shown that while common intensity normalizations remove a large part of the unwanted site effects for T1-w imaging, significant unwanted variation remains uncorrected. We encapsulated this post-normalization residual variability using the term *scan effect*. We have shown that the scan effect correction tool RAVEL successfully improves the comparability of the images in a large subset of the ADNI database by removing this extra variability. We measured the performance of RAVEL and other methods by estimating the replicability of the top voxels associated with AD in independent subsets of the ADNI dataset. To do so, we randomly divided the ADNI dataset into discovery and validation cohorts several times, and computed the top-replicated voxels for each random split. We have also shown that the top voxels associated with AD in our analysis and replicated in the discovery-validation division are more enriched for brain regions known to be associated with AD than

those found using intensity-normalized data only. This shows that RAVEL is a potent method for improving the discovery of brain regions associated with disease. Finally, we have also shown that the RAVEL correction improves the prediction of AD and MCI compared to healthy controls, using the mean hippocampal intensity as a predictor. This suggests that RAVEL is a promising method that may facilitate the development of biomarkers using MRI intensities. Furthermore, with the recent emphasis on multivariate pattern analysis for biomarker development,[156–161] RAVEL promises to produce more generalizable biomarkers that are less susceptible to biases associated with scanner and site imbalances.

The idea of using a control region of the brain which is not associated with disease is not new. In,[162–166] the regions of interest were divided by the mean signal intensity of a CSF region to correct for potential inter-subject variation.[102] used a NAWM stripe to estimate a scaling and shifting parameter in their z-score normalization method. In,[167] in the context of estimating quantitative T_1 maps (qT_1) from conventional MRI, the authors proposed an adaptation of the z-score normalization method by using a combination of NAWM and cerebellar gray matter (CBGM), where the NAWM was used for the scaling parameter and the CBGM was used for the shifting parameter. In,[168] the authors used the median GM intensity for the shifting parameter, and the difference between the median intraconal orbital fat intensity and the median GM intensity for the scaling parameter. In,[169] the authors use the whole brain to estimate the scale and shift parameters. We note that the different versions of the

z-score transformation used in[102, 167–169] only leave room for the choice of two control regions at maximum, corresponding to the mean and scale parameters. While this improves comparability between subjects in comparison to the unnormalized intensities, as shown in Figure 3.5b, we have shown that RAVEL improves dramatically upon a z-score transformation only.

There are several limitations to our method. If control regions are misspecified, i.e. the regions do not carry any information about the technical variability across subjects, or worse yet, if the control regions are inadvertently associated with the outcome of interest, the RAVEL correction may remove biological signals of interest. In both cases, however, cross-validation using the concordance curves from the discovery-validation scheme allows the user to estimate directly the performance of RAVEL on their dataset.

Another limitation is the use of nonlinear registration to align voxels across subjects. The registration step is necessary to apply the voxel-wise linear models from Equation 3.1. Because patients with AD and MCI have different volumes of WM, GM and CSF in comparison with healthy controls, misregistration error might be associated with the outcome of interest. However, this is a problem inherent to any cross-subject voxel analysis, and remains an active subject of research in image analysis. While voxels that are associated with disease can be a consequence of differential misregistration, this does not change the results of the present work, as misregistered voxels should be detected by intensity normalization method, after scan effect correc-

tion. It may also be possible to approximate RAVEL corrections using mean values in reference regions; indeed, in the ADNI dataset, the mean T1-w intensity in CSF after White Stripe correction was highly correlated with the first RAVEL factor. Thus, in the case of the well-controlled ADNI protocol, adjusting by regression on the mean in CSF would yield similar results. In cases where there is more heterogeneity in acquisitions, and in imaging modalities that are more difficult to calibrate, additional RAVEL factors are likely and using the mean in the reference region may not perform well.

A first extension of the presented methodology is to precede the RAVEL correction tool by a variant of the White Stripe intensity normalization method. For instance, as used in,[169] a whole-brain z-transformation might be used instead, where the mean and scaling parameters are estimating using all brain intensities. Subsequently, the RAVEL correction model can be applied using additional control regions, and mask erosion could be performed to improve the homogeneity of the selected control regions.

Although we have shown the performance of RAVEL in the context of T1-w MRI of the brain, RAVEL is a promising scan effect correction tool for many imaging modalities, such as quantitative images, maps derived from diffusion tensor imaging (DTI), functional imaging and many other modalities. Furthermore, the choice of the control regions, left to the user, makes the method applicable to virtually any disease and pathology. The RAVEL software can be found at `https://github.com/Jfortin1/RAVEL`.

Abbreviations

AD: Alzheimer's disease; ADNI: Alzheimer's Disease Neuroimaging Initiative; ANTs: Advanced Normalization Tools; AUC: area under the curve; BET: Brain Extraction Tool; CAT: concordance at the top; CBGM: cerebellar gray matter; CSF: cerebrospinal fluid; DTI: diffusion tensor imaging; FAST: FMRIB's Automated Segmentation Tool; FMRIB: Oxford Centre for Functional MRI of the Brain; FSL: FMRIB Software Library; GM: grey matter; MCI: mild cognitive impairment; MRI: magnetic resonance imaging; NAWM: normal-appearing white matter; NIFTI: The Neuroimaging Informatics Technology Initiative; RAVEL: Removal of Artificial Voxel Effect by Linear regression; ROC: receiver operating characteristic; RUV: removing unwanted variation; SVA: surrogate variable analysis; SVD: singular value decomposition; T1-w: T1-weighted; WM: white matter; WMPM: white matter parcellation map.

Acknowledgements

We would like to thank Paul Yushkevich and Sandhitsu Das for insightful discussions concerning biomarkers in AD.

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Insti-

tute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60million, 5-year public private partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California San Francisco. ADNI is the result of the efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow-up duration of each group is specified in the protocols for ADNI-1, ADNI-2 and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see `www.adni-info.`

org.

Chapter 4

Reconstruction of A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data

This chapter describes work published in a separate form in the journal *Genome Biology*, supervised by my advisor Kasper D. Hansen.

4.1 Introduction

Hi-C, a method for quantifying long-range physical interactions in the genome, was introduced by Lieberman-Aiden *et al.*,[58] and reviewed in Dekker *et al.*[59] A Hi-C assay produces a so-called genome contact matrix which – at a given resolution determined by sequencing depth – measures the degree of interaction between two loci in the genome. In the last 5 years, significant efforts have been made to obtain Hi-C maps at ever increasing resolutions.[60–65] Currently, the highest resolution maps are 1kb.[64] Existing Hi-C experiments have largely been performed in cell lines or for samples where unlimited input material is available.

In Lieberman-Aiden *et al.*,[58] it was established that at the megabase scale, the genome is divided into two compartments, called A/B compartments. Interactions between loci are largely constrained to occur between loci belonging to the same compartment. The 'A' compartment was found to be associated with open chromatin and the 'B' compartment with closed chromatin. Lieberman-Aiden *et al.*[58] also showed that these compartments are cell-type specific, but did not comprehensively describe differences between cell types across the genome. In most subsequent work using the Hi-C assay, the A/B compartments have received little attention; the focus has largely been on describing smaller domain structures using higher resolution data. Recently, it was shown that 36% of the genome changes compartment during mammalian development[65] and that these compartment changes are associated with gene

expression; they conclude "that the A and B compartments have a contributory but not deterministic role in determining cell-type-specific patterns of gene expression".

The A/B compartments are estimated by an eigenvector analysis of the genome contact matrix after normalization by the observed-expected method.[58] Specifically, boundary changes between the two compartments occur where the entries of the first eigenvector change sign. The observed-expected method normalizes bands of the genome contact matrix by dividing by their mean. This effectively standardizes interactions between two loci separated by a given distance by the average interaction between all loci separated by the same amount. It is critical that the genome contact matrix is normalized in this way, for the first eigenvector to yield the A/B compartments.

Open and closed chromatin can be defined in different ways using different assays such as DNase hypersensitivity or ChIP sequencing for various histone modifications. While Lieberman-Aiden *et al.*[58] established that the 'A' compartment is associated with open chromatin profiles from various assays, including DNase hypersensitivity, it was not determined to which degree these different data types measure the same underlying phenomena, including whether the domain boundaries estimated using different assays coincide genome-wide.

In this manuscript, we show that we can reliably estimate A/B compartments as defined using Hi-C data by using Illumina 450k DNA methylation microarray data[2] as well as DNase hypersensitivity sequencing,[66,67] single-cell whole-genome bisulfite

sequencing (scWGBS)[68] and single-cell assay for transposase-accessible chromatin (scATAC) sequencing.[69] Data from the first two assays are widely available on a large number of cell types. In particular, the 450k array has been used to profile a large number of primary samples, including many human cancers; more than 20,000 samples are readily available through the Gene Expression Omnibus (GEO) and The Cancer Genome Atlas (TCGA).[70] We show that our methods can recover cell type differences. This work makes it possible to study A/B compartments comprehensively across many cell types, including primary samples, and to further investigate the relationship between genome compartmentalization and transcriptional activity or other functional readouts.

As an application, we show how the somatic mutation rate in prostate adenocarcinoma is different between compartments and we show how the A/B compartments change between several human cancers; currently The Cancer Genome Atlas does not include assays measuring chromatin accessibility. Furthermore, our work reveals unappreciated aspects of the structure of long-range correlations in DNA methylation and DNase hypersensitivity data. Specifically, we observe that both DNA methylation and DNase signal are highly correlated between distant loci, provided that the two loci are both in the closed compartment.

4.2 Results and discussion

4.2.1 A/B compartments are highly reproducible and are cell-type specific

We obtained publicly available Hi-C data on EBV-transformed lymphoblastoid cell lines (LCLs) and fibroblast cell lines and estimated A/B compartments by an eigenvector analysis of the normalized Hi-C contact matrix (Methods). The contact matrices were preprocessed with ICE[71] and normalized using the expected-observed method.[58] As in Lieberman-Aiden *et al.*,[58] we found that the eigenvector divides the genome into two compartments based on the sign of its entries. These two compartments have previously been found to be associated with open and closed chromatin; in the following, we will use open to refer to the 'A' compartment and closed to refer to the 'B' compartment. The sign of the eigenvector is arbitrary; in this manuscript, we select the sign so that positive values are associated with the closed compartment (Methods). In Figure 4.1 we show estimated eigenvectors at 100kb resolution from chromosome 14 across 2 cell types measured in multiple laboratories with widely different sequencing depth, as well as variations in the experimental protocol. We observed a very high degree of correspondence between replicates of the same cell type; on chromosome 14, the correlation between eigenvectors from experiments in

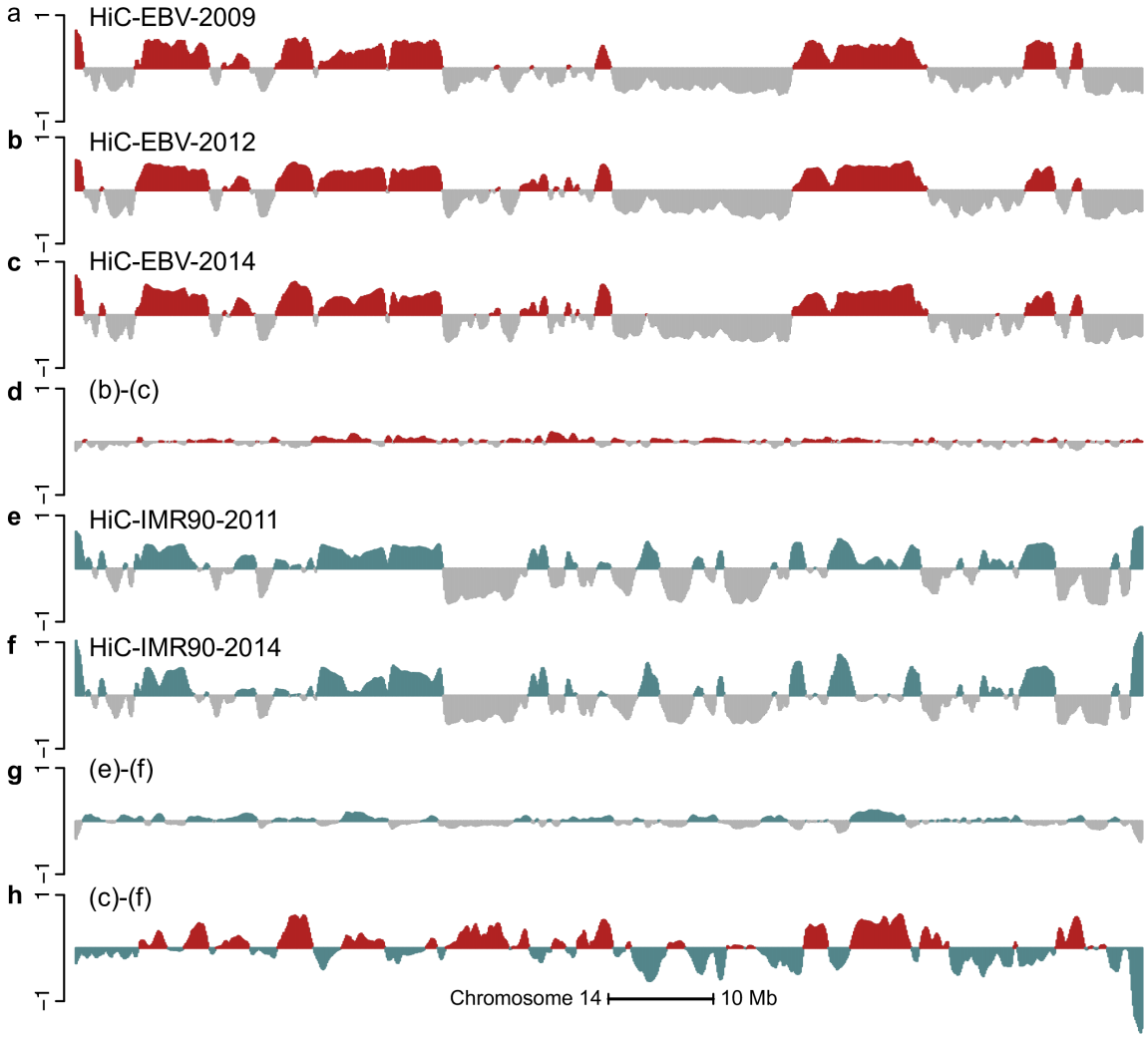


Figure 4.1: A/B compartments are reproducible and cell-type specific. The figure displays data on all of chromosome 14 at 100kb resolution. The first eigenvector for the expected-observed normalized (a) HiC-EBV-2009, (b) HiC-EBV-2012, (c) HiC-EBV-2014 datasets. (d) The difference between (b) and (c). The first eigenvector for the expected-observed normalized (e) HiC-IMR90-2013, (f) HiC-IMR90-2014 datasets and (g) their difference. (h) The difference between (c) and (f), which is greater than the technical variation depicted in (d) and (g). This establishes that Hi-C compartments are highly reproducible between experiments in different laboratories and that compartments are cell type specific.

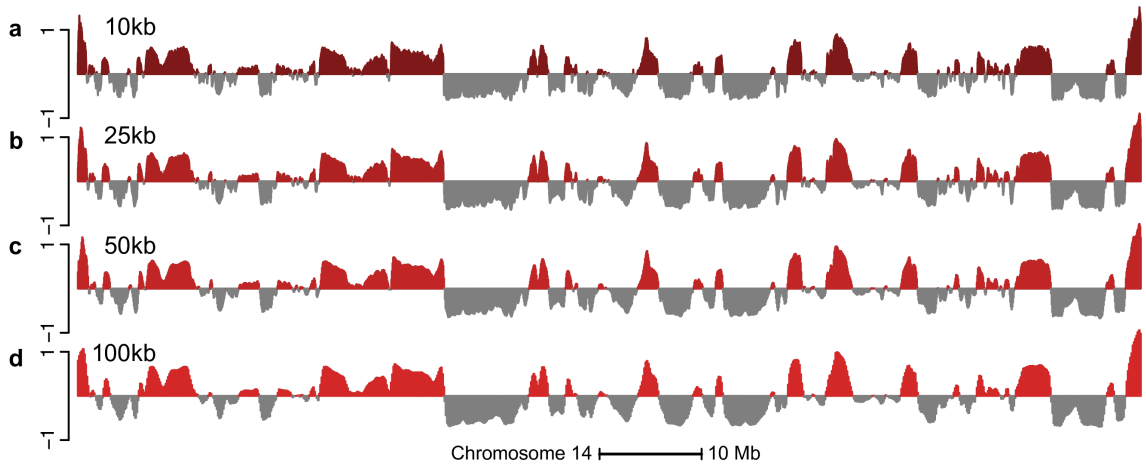


Figure 4.2: A/B compartments revealed by Hi-C data do not change at resolutions higher than 100kb. The figure displays data on all of chromosome 14 at different resolutions. The four different tracks represent the first eigenvector of the HiC-IMR90-2014 dataset at resolutions (a) 10kb, (b) 25kb, (c) 50kb and (d) 100kb.

the same cell type is greater than 0.96 (ranges from 0.96 to 0.98). The agreement, defined as the percentage of genomic bins which are assigned to the same compartment in two different experiments, is greater than 92% (ranges from 92.6% to 96.0%) on chromosome 14. These measures vary little between chromosomes.

Using high-resolution data does not change the estimated A/B compartments as seen in Figure 4.2. Note that the Hi-C datasets have been processed into unadjusted contact matrices using different alignment and filtering pipelines (see Methods for details); this shows that the choice of alignment and filtering method has negligible impact on A/B compartments estimation.

Figure 4.1 shows the A/B compartments are cell-type specific, with a variation between cell types which exceeds technical variation in the assay; this has been previously noted.[58,65] The correlation between eigenvectors from different cell types is

around 0.60, in contrast to 0.96+ between eigenvectors from the same cell type.

ICE normalization removes any marginal dependence of the contact matrix on GC content by forcing the marginal sums of the contact matrix to be constant.[71] Despite this, Imakaev *et al.*[71] found high correlation (0.80) between the first eigenvector of the contact matrix and GC content of the underlying bin, and interpreted this as a biological association and not technical bias. To further investigate whether this dependence is a result of technical bias or a biological association, we computed the dependence for multiple experiments (Figure 4.3). Like the eigenvector itself, we found that the dependence shows little variation between experiments done on the same cell line but in different labs, and some variation between cell lines (Figures 4.3 and 4.4). This comparison includes two cell line experiments performed in the same laboratory with the same experimental protocol. That the effect of GC content depends on the cell line suggests that the relationship at least partly reflects biology. Various biological entities are correlated with GC content, including gene density;[72] it is therefore not inconceivable that open and closed chromatin has a biological association with GC content. It is possible to computationally adjust for the dependence on GC content by regressing out the fitted loess curve displayed in Figure 4.3; like Imakaev *et al.*[71] we currently believe that doing so will remove some biological signal.

In the remainder of the manuscript, we use the most recent data, ie. HiC-EBV-2014 and HiC-IMR90-2014, to represent eigenvectors and A/B compartments derived from Hi-C data in these cell types.

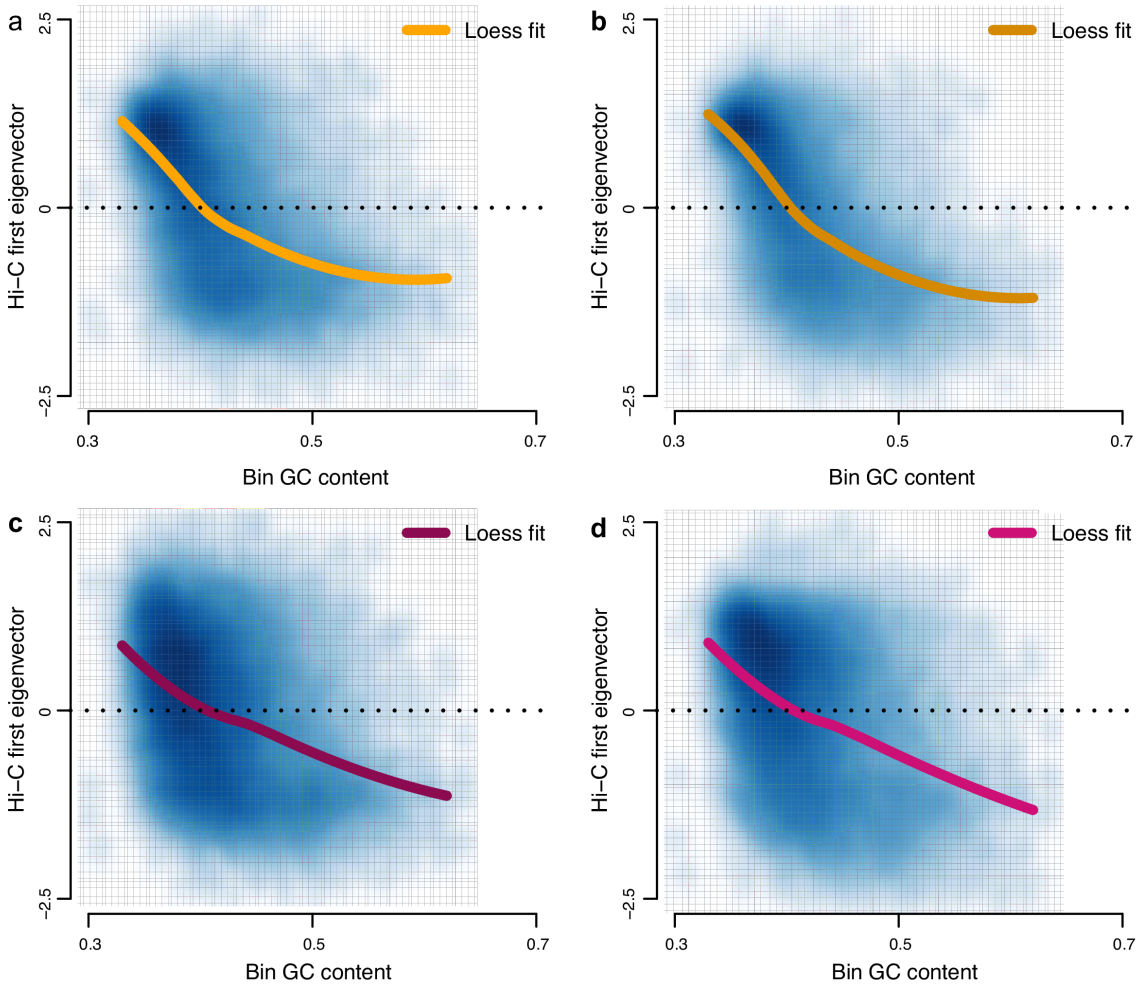
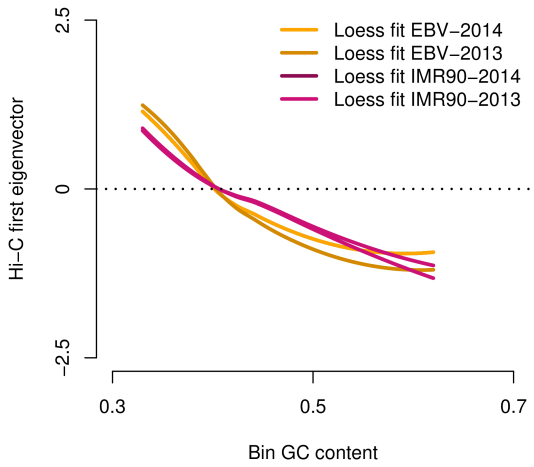


Figure 4.3: Association between the Hi-C eigenvectors and GC content is reproducible and cell-type specific. The four plots represent the genome-wide eigenvector value for each bin against the GC content of the bin at resolution 100kb. The four datasets are (a) HiC-EBV-2014, (b) HiC-EBV-2013, (c) HiC-IMR90-2014, and (d) HiC-IMR90-2013. Note that the experiments (a) and (c) are from the same laboratory, while the experiments (b) and (d) are from different labs. The orange lines represent the loess fit for GC content. We observe that loess curves from the same cell type (the two at the top, and then the two at the bottom) are more similar to each other than across cell types, despite different protocols and experiment years. This reflects the cell-type specificity of the GC content association with the Hi-C first eigenvector, reflecting a functional association rather than a pure technical bias.



(a)

Figure 4.4: Association between the Hi-C eigenvectors and GC content is reproducible and cell-type specific. The four loess fits come from Figure 4.3. The relationship between the Hi-C eigenvectors and GC content is cell-type specific.

4.2.2 Predicting A/B compartments from DNA methylation data

To estimate A/B compartments using epigenetic data other than Hi-C, we first concentrate on DNA methylation data assayed using the Illumina 450k microarray platform. Data from this platform is widely available across many different primary cell types. To compare with existing Hi-C maps, we obtained data from 288 EBV-transformed lymphoblastoid cell lines from the HapMap project.[73]

DNA methylation is often described as related to active and inactive parts of the genome. Most established is high methylation in a genic promoter leading to silencing

of the gene.[74] As a first attempt to predict A/B compartments from DNA methylation data, we binned the genome and averaged methylation values across samples and CpGs inside each bin. Only CpGs more than 4kb away from CpG islands were used; these are termed open sea CpGs (Methods). We found that high levels of average methylation were associated with the open compartment and not the closed compartment; this might be a consequence of averaging over open sea probes. Figure 4.5 depicts data from such an analysis for lymphoblastoid cell lines on chromosome 14 at a 100kb resolution and shows some agreement between estimated compartments from Hi-C and this analysis, with a correlation of 0.56 and a compartment agreement between datasets of 71.7% on this chromosome. In this analysis, we implicitly assume that there is no variation in compartments between different individuals in the same cell type.

Surprisingly, we found that we could improve considerably on this analysis by doing an eigenvector analysis of a suitably processed between-CpG correlation matrix (Figure 4.5). This matrix represents correlations between any two CpGs measured on the 450k array, with the correlation being based on biological replicates of the same cell type. The correlation eigenvector shows strong agreement with the Hi-C eigenvector; certainly higher than with the average methylation vector (Figure 4.5). Quantifying this agreement, we found that the correlation between the two vectors is 0.85 and the compartment agreement is 83.8% on chromosome 14. Genome-wide, the correlation is 0.71 and the agreement is 79% (Table 4.1). Again, this analysis

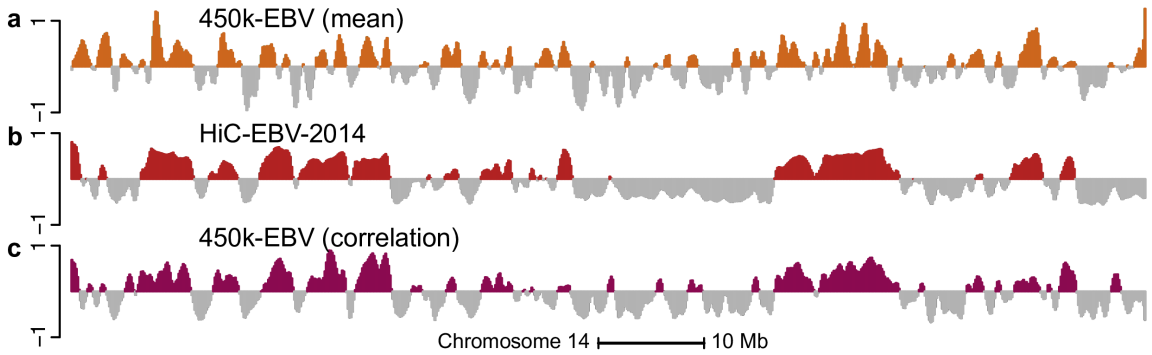


Figure 4.5: The methylation correlation signal is a better predictor of A/B compartments than the average methylation signal. The figure displays data on all of chromosome 14 at 100kb resolution. (a) The smoothed, average methylation signal on the beta-value scale for the 450k-EBV dataset. The signal has been centered by the mean and the sign has been reversed so that values close to one correspond to low methylation values. (b) The first eigenvector of the HiC-EBV-2014 Hi-C dataset. (c) The smoothed first eigenvector of the binned correlation matrix of the 450k-EBV dataset. We see that (c) correlates better with (b) than (a).

implicitly assumes lack of variation in compartments between biological replicates.

Table 4.1: Correlation and agreement between Hi-C and 450k-based eigenvector estimates of genome compartments. Thresholding refers to excluding genomic bins where the entries of the relevant eigenvector has an absolute value less than 0.01

	Chr 14		Genome	
	EBV	Fibro	EBV	Fibro
No threshold				
Correlation	0.85	0.75	0.70	0.74
Agreement	83.7%	79.1%	79.0%	79.5%
Bins retained	100%	100%	100%	100%
Threshold, Methylation				
Correlation	0.87	0.78	0.74	0.77
Agreement	88.8%	83.4%	87.9%	87.9%
Bins retained	86%	85%	78%	79%
Threshold, Methylation and Hi-C				
Correlation	0.89	0.84	0.77	0.81
Agreement	93.0%	90.5%	92.6%	92.8%
Bins retained	76%	67%	66%	64%

Closely examining differences between the 450k-based predictions and the Hi-C-based estimates, we found that almost all disagreements between the two methods occur when an entry in one of the two eigenvectors is close to zero; in other words, where there is uncertainty about the compartment in either one of the two analyses. Excluding bins where the 450k-based prediction is close to zero, that is bins that have an absolute eigenvector value less than 0.01, we got an agreement of 88.8% (14.2% of the bins excluded). Excluding bins where either the 450k-based prediction is close to zero or the Hi-C eigenvector is close to zero, we got an agreement of 93% (24.8% of the bins excluded).

Our processing of the correlation matrix is as follows (see Methods for details); the rationale behind our choices will be explained later in the manuscript. First, in our correlation matrix, we only included so-called 'open sea' CpGs; these CpGs are more than 4kb away from CpG islands. Next, we binned each chromosome into 100kb bins and computed which open sea CpGs are inside each bin; this varies between bins due to the design of the 450k microarray. To get a single number representing the correlation between two bins, we took the median of the correlations of the individual CpGs located in each bin. We obtained the first eigenvector of this binned correlation matrix and gently smoothed the signal by using two iterations of a moving average with a window size of 3 bins.

The sign of the eigenvector is chosen so that the sign of the correlation between the eigenvector and column sums of the correlation matrix is positive; this ensures

that positive values of the eigenvector are associated with the closed compartment (see Methods).

4.2.3 Long-range correlations in DNA methylation data predict A/B compartment changes between cell types

To examine how well the predictions based on long-range correlations in 450k data capture differences between cell types, we obtained publicly available 450k data from 62 fibroblast samples,[75] and compared them to Hi-C data from the IMR90 cell lines. Note that the fibroblast cell lines assayed on the 450k platform are from primary skin in contrast to the IMR90 cell line, a fetal lung fibroblast. Figure 4.6 and Table 4.1 show our ability to recover the A/B compartments in fibroblasts; it is similar to our performance for EBV-transformed lymphocytes.

To firmly establish that the high correlation between our predicted compartments using DNA methylation and Hi-C data is not due to chance, we compared the predicted compartments in EBV transformed lymphocytes and fibroblasts to Hi-C data from different cell types, including the K562 cell line which serves as a somewhat independent negative control. In Figure 4.7, we show the correlation and agreement between the two sets of predicted compartments and Hi-C data from the three cell types. There is always a decent agreement between predicted compartments of any

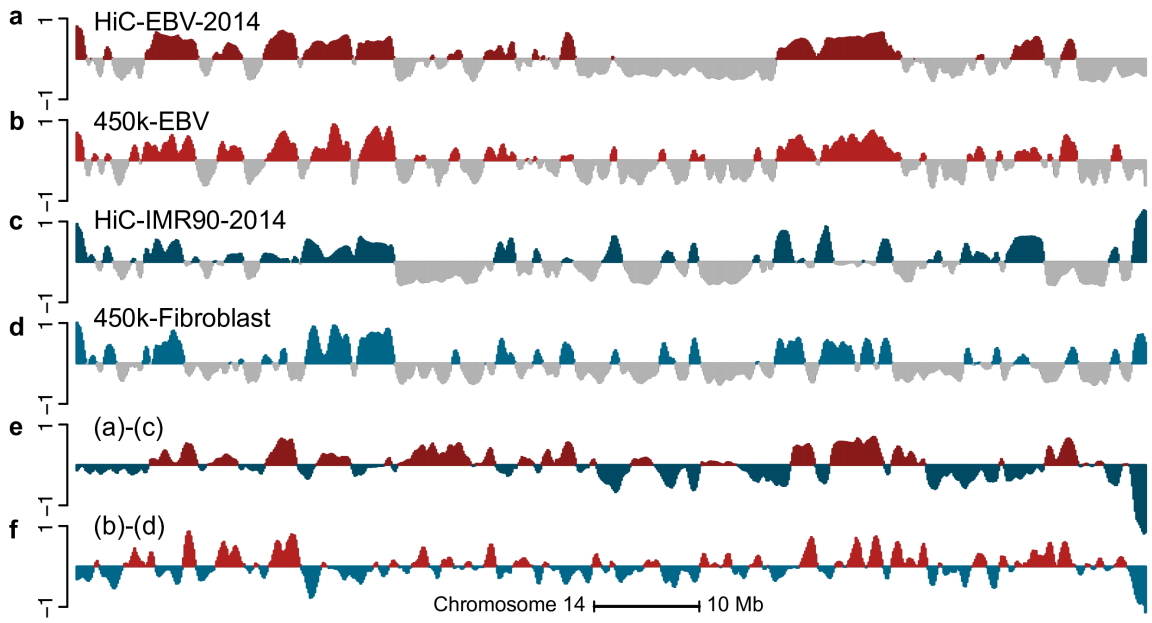


Figure 4.6: Cell-type specific A/B compartments using Hi-C data are predicted using DNA methylation data. The figure displays data on all of chromosome 14 at 100kb resolution. (a) The first eigenvector of the HiC-EBV-2014 dataset. (b) The smoothed first eigenvector of the binned correlation matrix of the 450k-EBV dataset. (c) The first eigenvector of the HiC-IMR90-2014 Hi-C dataset. (d) The smoothed first eigenvector of the binned correlation matrix of the 450k-Fibroblast dataset. (e) The difference between (a) and (c). (f) the difference between (b) and (d). The high correlation between (e) and (f) supports that the correlation eigenvectors of the 450k data can be used to find differences between compartments in the two cell types.

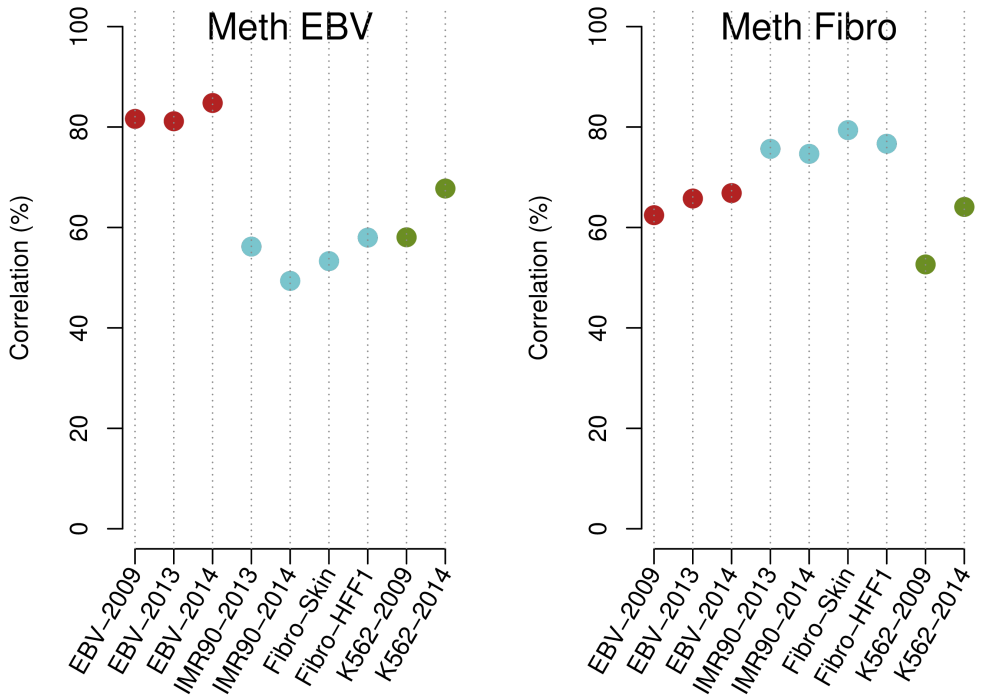


Figure 4.7: Compartment predictions based on 450k data are cell-type specific. Correlations between the eigenvectors from 9 different Hi-C datasets across 3 different cell types with both the 450k-EBV and 450k-Fibroblast datasets. We observe higher correlation between Hi-C data and DNA methylation data when the comparison is being made for the same cell type.

two cell types, but the agreement is consistently higher when the prediction is from data from the same cell type as the Hi-C data.

How to best quantify differences in A/B compartments is still an open question. Lieberman-Aiden *et al.*[58] used 0 as a threshold to differentiate the two compartments. Considering the difference of two eigenvectors derived in different cell types, it is not clear that functional differences exist exactly when the two eigenvectors have opposite signs; instead, functional differences might be associated with changes in

the magnitude of the eigenvectors reflecting a genomic region being relatively more open or closed. We note that the genomic region highlighted as cell-type specific, and validated by FISH, in Lieberman-Aiden *et al.*,[58] is far away from zero in one condition and has small values fluctuating around zero in the other condition.

Following this discussion, we focus on estimating the direction of change in eigenvectors between different cell types. Figure 4.6 shows estimated differences between Hi-C and 450k eigenvectors for two cell types. Large differences between the two vectors are replicated well between the two data types, but there is disagreement when the eigenvectors are close to zero. This is to be expected; there is technical variation in such a difference even between Hi-C experiments (Figure 4.1). Using the data displayed in Figure 4.1 we found that the technical variation in the Hi-C data is such that 98% of genomic bins have an absolute value less than 0.02. Using this cutoff for technical variation, we found that the correlation between the two difference vectors displayed in Figure 4.6 is 0.85 when restricted to the 24% of genomic bins where both vectors have an absolute value greater than 0.02. The signs of the differential vectors are also in high agreement; they agree in 90% of the genomic bins exceeding the cutoff for technical variation. In contrast, the correlation is 0.61 when the entire chromosome is included, reflecting that the technical noise is less correlated than the signal.

Large domains of intermediate methylation have been previously described,[76] as well as long blocks of hypomethylation associated with colon cancer and EBV transfor-

mation.[10,11,39] We obtained previously characterized[76] partially methylated domains (PMDs) in IMR90 and found a significant overlap with closed compartments from the HiC-IMR90-2014 dataset (odds ratio: 13.6) as well as closed compartments from the 450k-Fibroblast dataset (odds ratio: 16.4). Likewise, we obtained previously characterized blocks of hypomethylation associated with EBV transformation[39] and found a significant overlap with closed compartments from the HiC-EBV-2014 dataset (odds ratio: 11.9) and 450k-EBV dataset (odds ratio: 9.4). This confirms previously described overlap between Hi-C compartments and these types of methylation domains by Berman *et al.*[11]

4.2.4 The structure of long-range correlations in DNA methylation data

To understand why we are able to predict open and closed compartments using the 450k array, we studied the structure of long-range correlations in DNA methylation data. In Figure 4.8, we show density plots of binned correlations on chromosome 14, stratified in two ways. The first stratification separates correlations between bins which are both in the open compartment, both in the closed and finally cross-compartment correlations. This stratification shows that we have a large number of intermediate correlation values (0.2-0.5), but only between bins which are both in the closed compartment. The second stratification separates open sea probes and

CpG resort probes (probes within 4kb of a CpG island, see Methods). This stratification shows that we only have intermediate correlation values for open sea probes; CpG resort probes are generally uncorrelated. In conclusion, we have the following structure of the binned correlation matrix: most of the matrix contains correlation values around zero (slightly positive), except between two bins both in the closed compartment, which have an intermediate correlation value of 0.2-0.5. This shows why an eigen analysis of the binned correlation matrix recovers the open and closed compartments; see Figure 4.9 for an illustration.

The lack of decay of correlation with distance extends even to trans-chromosomal correlations, again with a clear difference between correlations within the open compartment and the closed compartment (Figure 4.10).

To understand what drives the correlation between loci within the closed compartment, we carefully examined the DNA methylation data in these genomic regions. Figure 4.11 shows a very surprising feature of the data, which explains the long-range correlations. In this figure, we have arbitrarily selected 10 samples and we plot their methylation levels across a small part of chromosome 14, with each sample having its own color. Data from both EBV-transformed lymphocytes and fibroblasts are depicted. While the same coloring scheme has been used for both cell types, there is no correspondence between the samples assayed in the different experiments. The figure shows that the 10 samples have roughly the same ranking inside each region in the closed compartment. This illustrates a surprising genome-wide ranking between

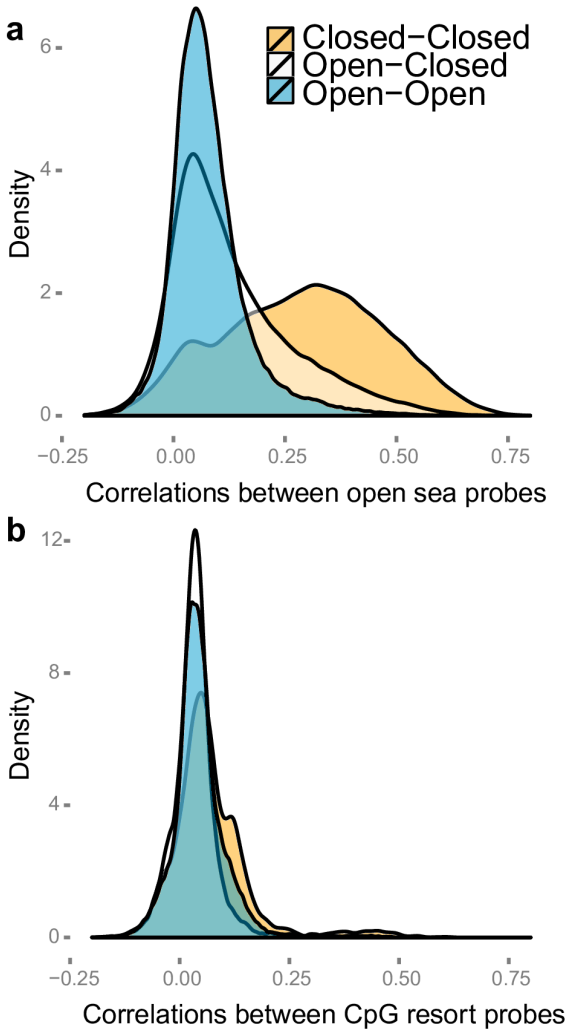


Figure 4.8: Densities of the correlations of the 450k methylation probes. Chromosome 14 was binned at resolution 100 kb and we display the binned, stratified correlations for the 450k-EBV dataset. Each plot shows one density curve for each type of interaction: between two bins in open compartments, between two bins in closed compartments and between a bin in the open compartment and the closed compartment. (a) Binned correlations for open sea probes only. (b) Binned correlations for CpG resort probes only. Most correlations are around zero, except correlations between two open sea probes in the closed compartment. The open and closed compartments were defined using the HiC-EBV-2014 dataset.

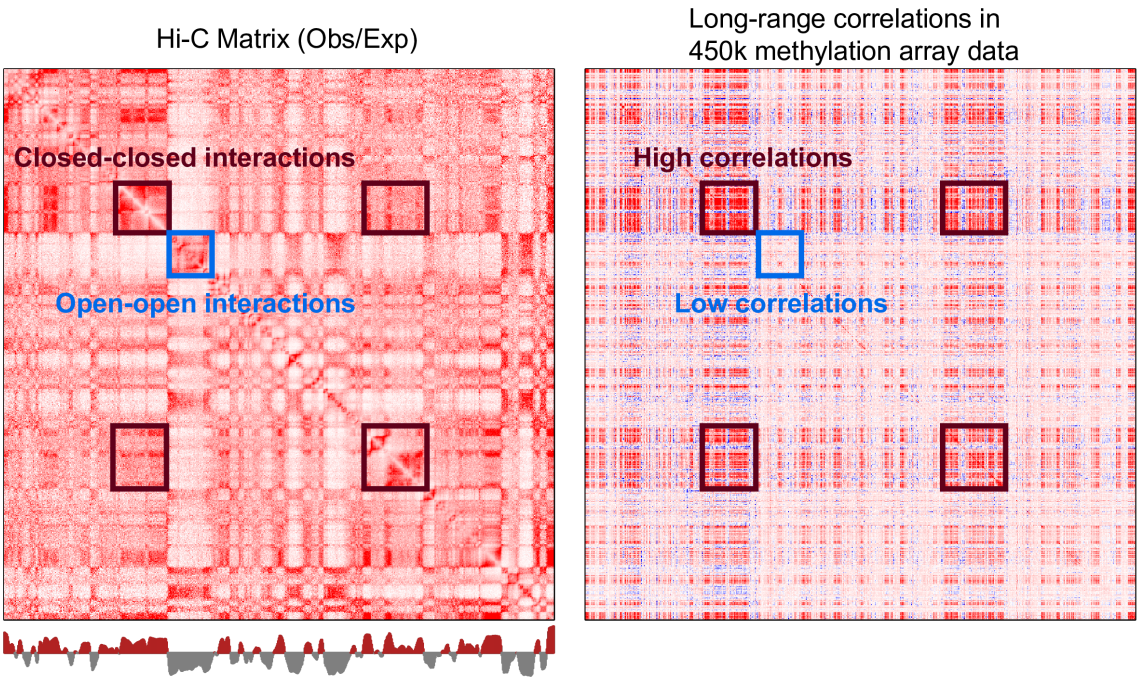


Figure 4.9: The relationship between a Hi-C contact matrix and a binned DNA methylation correlation matrix. Depicted are the observed-expected normalized genome contact matrix for the HiC-IMR90-2014 dataset together with the binned correlation matrix for the 450k-Fibroblast dataset. Both matrices depict chromosome 14 at resolution 100kb. There is a relationship between A/B compartments in the Hi-C data and regions with low and high correlations.

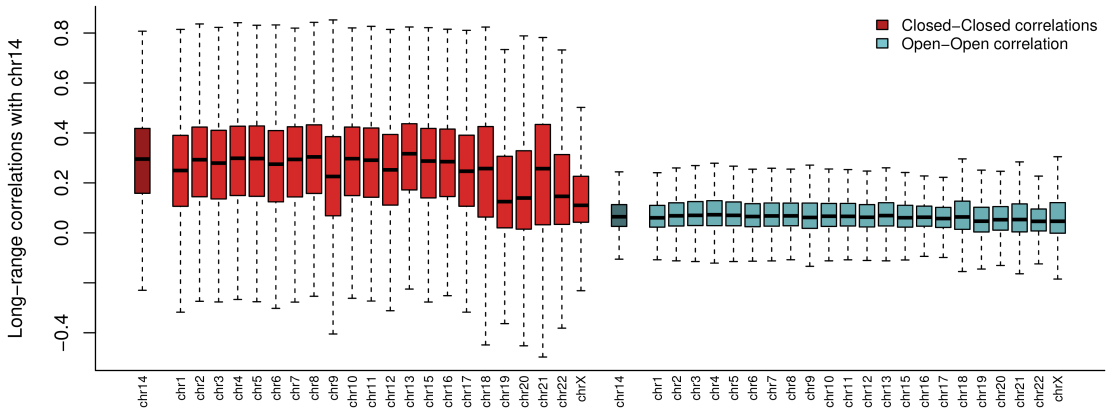


Figure 4.10: Between-chromosome correlations of DNA methylation. Each boxplot shows the binned correlations between bins on chromosome 14 and bins on other chromosomes for the 450k-EBV dataset. The boxplots are stratified by whether the correlation is inside the open compartment or inside the closed compartment (open to close compartment correlations are not depicted). The open and closed compartments were defined using the HiC-EBV-2014 dataset.

samples in the closed compartment.

To gain more insights into whether this ranking is caused by technical artifacts or whether it reflects real differences between the biological replicates, we obtained data where the exact same HapMap samples were profiled in two different experiments using the Illumina 27k methylation array. This array design is concentrated around CpG islands, but we determined that 5,599 probes are part of the 450k array and annotated as open sea probes. For these probes, we determined which were part of the closed compartment and we computed the sample-specific average methylation in this compartment as a proxy for the observed ranking described above. In Figure 4.12a, we show that the genome-wide correlation of these measurements between hybridization duplicates from the same experiment is high (0.927). In Figure 4.12b, we show that the these measurements replicate well between different experiments (correlation of 0.744).

For the 450k-Fibroblast experiment, we had access to the raw IDAT files and therefore to the control probes located on the array. For this dataset, we examined if the striking global ranking between different samples using the open sea probes in the closed compartment could be explained by technical factors such as bisulfite conversion. To test this, we regressed the mean (and median) methylation levels against each of the following 5 variables: chip and well variables (surrogates for batch), Bisulfite I and Bisulfite II control probes and negative control probes (background noise). None of these variables were significantly associated with the mean of the

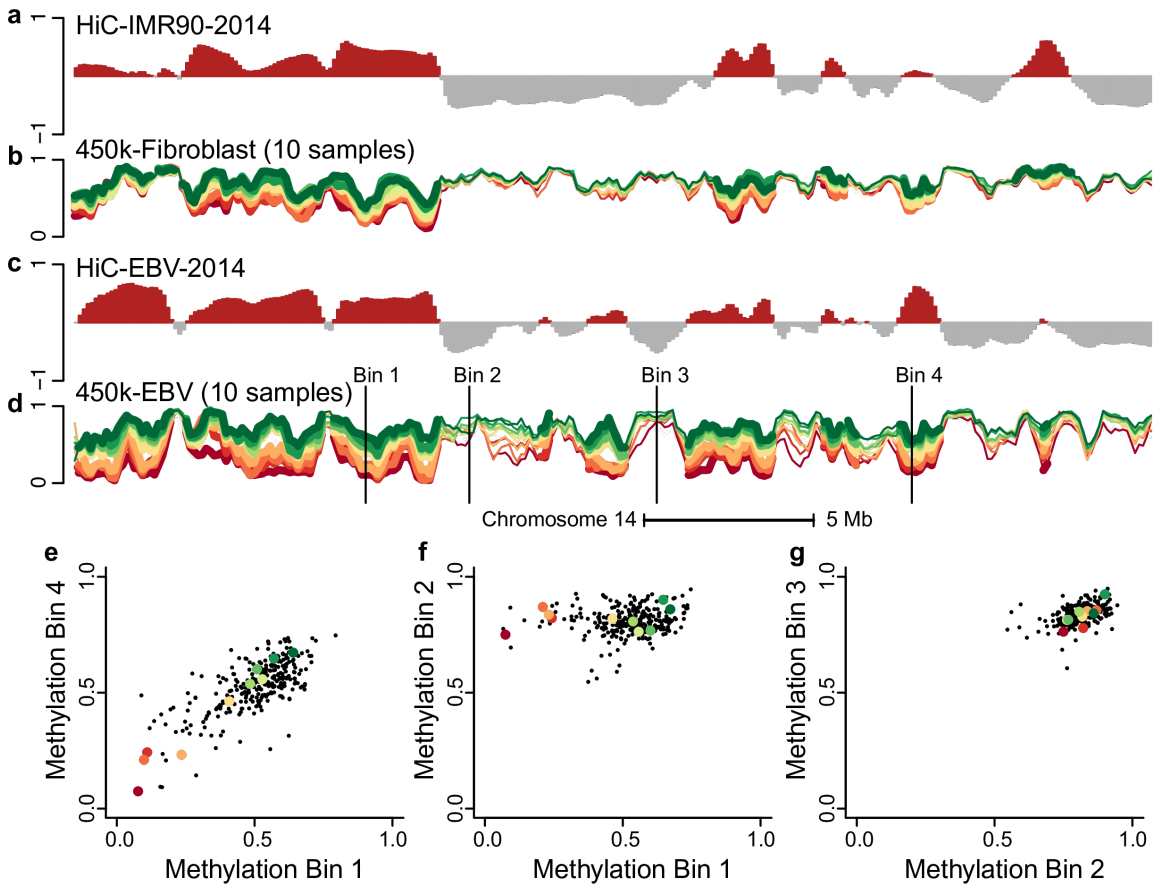
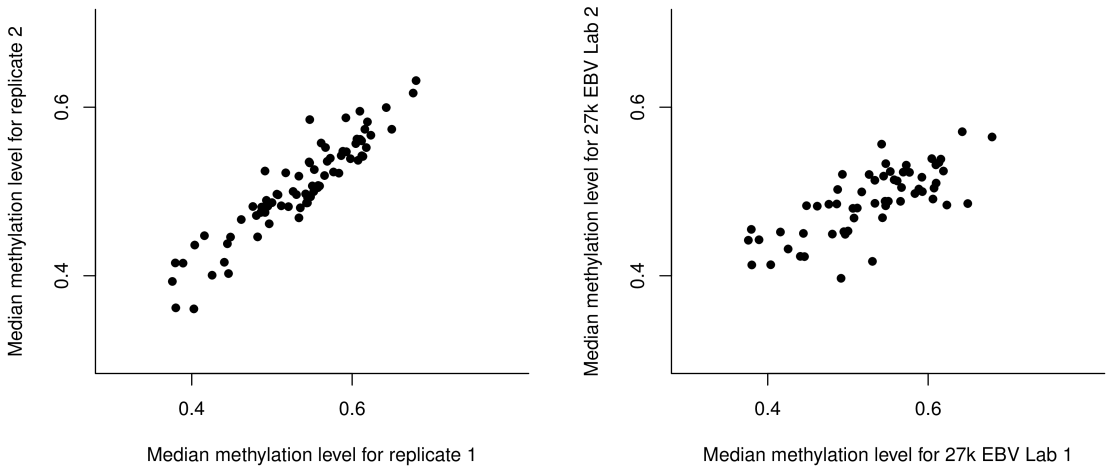


Figure 4.11: Comparison of the methylation levels and the Hi-C compartments signal for chromosome 14. The figure displays data from 36.4 to 69.8 Mb on chromosome 14 at 100kb resolution. (a) The first eigenvector from the HiC-IMR90-2014 dataset. (b) Average methylation on the beta scale for 10 selected samples from the 450k-Fibroblast dataset; each sample is a line and divergent colors are used to distinguish the different levels of methylation in the different samples. (c) The first eigenvector from the HiC-EBV-2014 data. (d) Like (b) but for 10 samples from the 450k-EBV dataset; the samples from the two datasets are unrelated. On (d) we depict 4 different bins; scatterplots between methylation values in different bins across all samples in the dataset are shown in (e-g) where (e) are two bins in the closed compartment, (g) are one bin in the open and one bin in the closed compartment and (g) two bins in the open compartment. The figure shows that samples have roughly the same ranking inside each closed compartment.



(a) **(b)**

Figure 4.12: Sample ranking based on methylation levels in the closed compartments replicate across experiments. We computed the average methylation level of open sea probes in the closed compartment. The compartments were defined using the HiC-EBV-2014 data. (a) Comparison between hybridization replicates from the 27k-London dataset. (b) Comparison between the same samples assayed in two different experiments, the 27k-Vancouver and the 27k-London experiments.

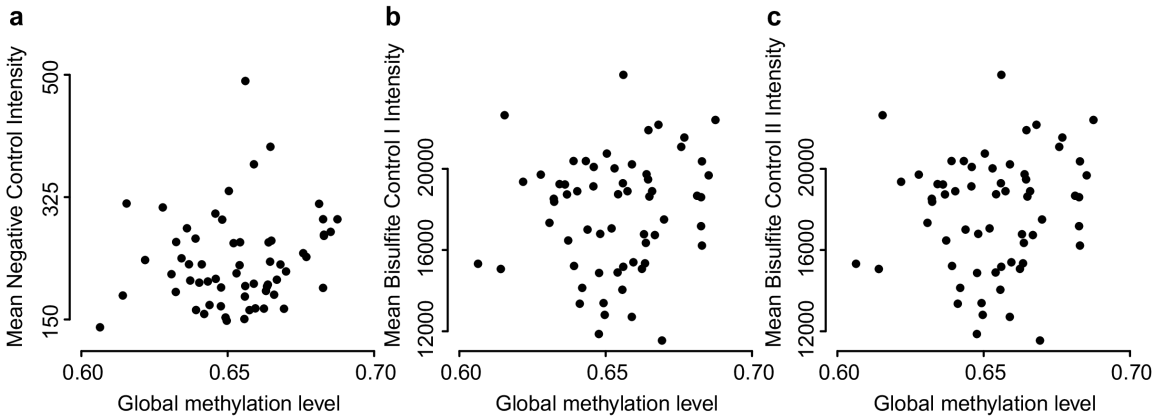


Figure 4.13: Mean methylation levels in the 450k-Fibroblast dataset are not associated with technical control probes. For each of the 62 samples from the 450k-Fibroblast dataset, we computed the average methylation level for the open sea probes and looked for association with the array technical control probes. (a) Average intensity of the negative control probes against the average methylation level (b) same as (a) but for bisulfite conversion efficiency control probes Infinium I (c) same as (b) but for bisulfite conversion efficiency control probes Infinium II. We conclude that the average methylation levels are not associated with known technical covariates.

median methylation levels (all p-values greater than 0.09 and R^2 less than 16%), see Figure 4.13. We conclude that the global ranking cannot be explained by technical issues.

Finally, using the 27k data, we show that the eigenvector replicates between a 450k experiment and a 27k experiment using the same cell type (EBV), but different samples (correlation of 0.89). As a control, we compared to a 450k-derived eigenvector for a different cell type (fibroblast) and observed weak correlation (0.40). We note that the eigenvector derived from the 27k experiment is based on far fewer probes; we do not recommend using 27k data to estimate compartments. This result shows that the estimated genome compartments do not depend on the design of the microarray

and suggests that our observations are common across methylation assays.

4.2.5 The impact of GC content on long-range correlations in DNA methylation data

To examine the impact of GC content on the distribution of correlations, we computed this distribution as a function of both the GC content of the probe and a 1kb window around the probe (Figure 4.14a,b), and did not observe any dependence of the distribution of probe-specific correlations on GC content. The same was true when we examined the distribution of correlations as a function of the methylation level of the probe (Figure 4.14c). This is in sharp contrast to the well-known high degree of association between methylation and GC content in a 1kb around the probe (Figure 4.14d). In Figure 4.14, we have only displayed open sea probes, and we note that these probes cover a wide range of GC content and methylation values. These results strongly suggest that the low correlations observed for CpG resort probes are not a technical artifact caused by GC content or probe-level methylation.

Because the Hi-C based eigenvectors are associated with GC content, it is expected to see such an association for 450k-derived eigenvectors. To estimate how much of the correlation between Hi-C and methylation is due to GC content, we applied a GC content stratified permutation procedure similar to what Imakaev *et al.*[71] used. Briefly, we sorted the Hi-C and methylation eigenvectors by GC content and

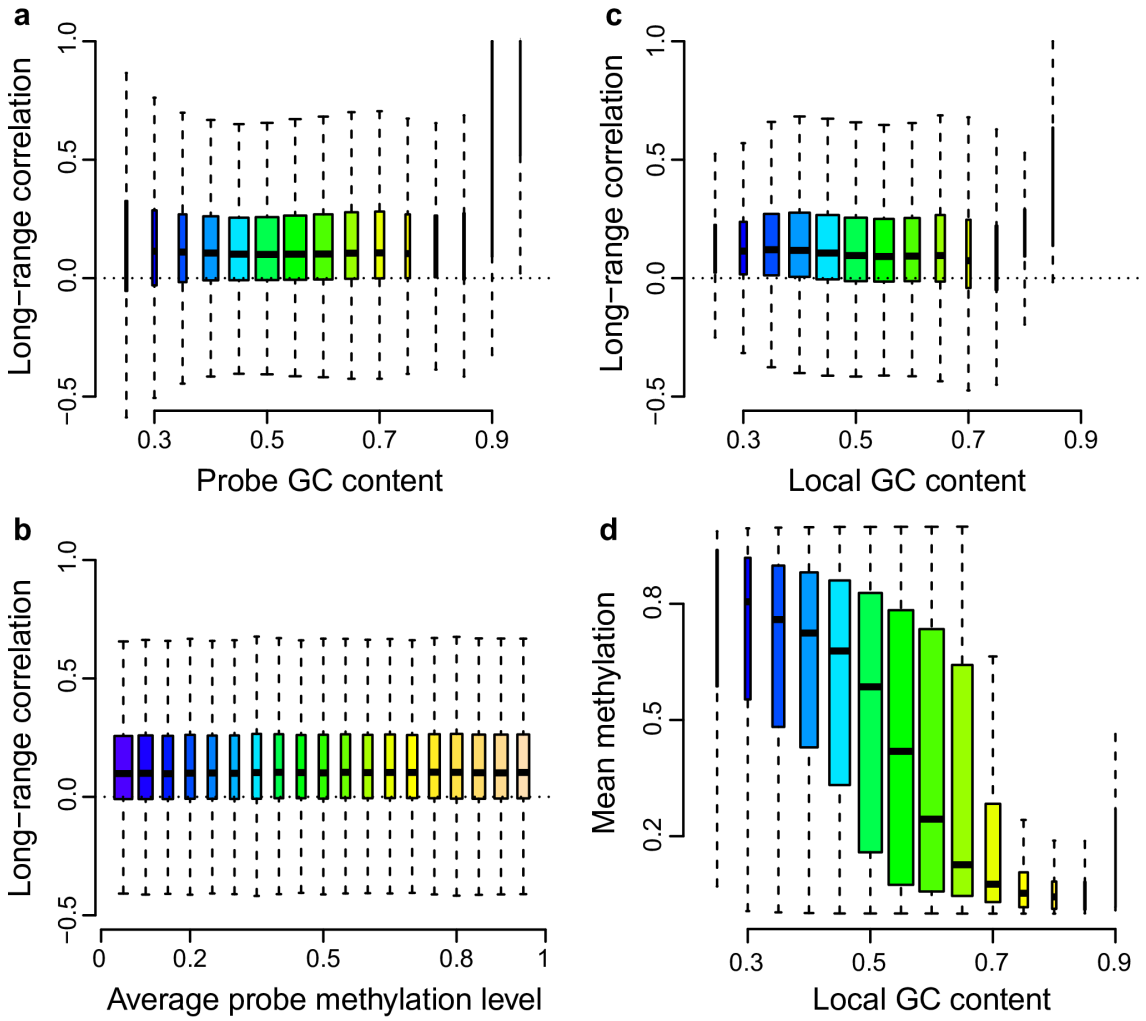


Figure 4.14: Relationship between long-range correlations, GC content and methylation levels for the 450k-EBV dataset. Only open sea probes were included in the analysis. (a) Long-range correlations of the methylation levels stratified by probe GC content (b) Same as (a), but GC content was measured in a 1kb window around the probe (c) Long-range correlations of the methylation levels stratified by average probe methylation (d) Relationship between mean methylation level and GC content. While regions with high GC content tend to have low methylation, as for instance CpG islands, we do not observe any relationship between the open sea probes GC content and the long-range correlations. We conclude that GC content is not a bias of our methylation correlation analysis.

permuted neighbors within a 5-bin window (to keep GC content roughly unchanged) and recalculated the correlation between the two eigenvectors. We generated 100 such permutations. While the genome-wide correlation between the Hi-C and methylation eigenvectors is high before permutation (0.74), the correlation drops to 0.21 after permuting (0.20 and 0.22 for the 2.5 and 97.5 percentiles respectively). We conclude that GC content by itself fails to explain the high correlation between the Hi-C and methylation eigenvectors. Based on these results, and the reasoning above, we caution that removing the GC content effect might remove biological signal. Nevertheless, we examined whether adjusting for GC content in both Hi-C and 450k eigenvectors would change the association between the two vectors. Before loess correction, the genome-wide correlation between the two eigenvectors for the EBV data is 0.71 with a domain agreement of 79%. After GC content adjustment, the residual eigenvectors are still highly correlated (0.69) with a domain agreement of 77%. This shows that adjusting for GC content does not diminish our ability to estimate A/B compartments using 450k methylation data.

4.2.6 Sometimes compartment prediction fails using DNA methylation data

We caution that it is not always possible to estimate A/B compartments using data from the 450k DNA methylation array. As an example, we present an analysis of 305

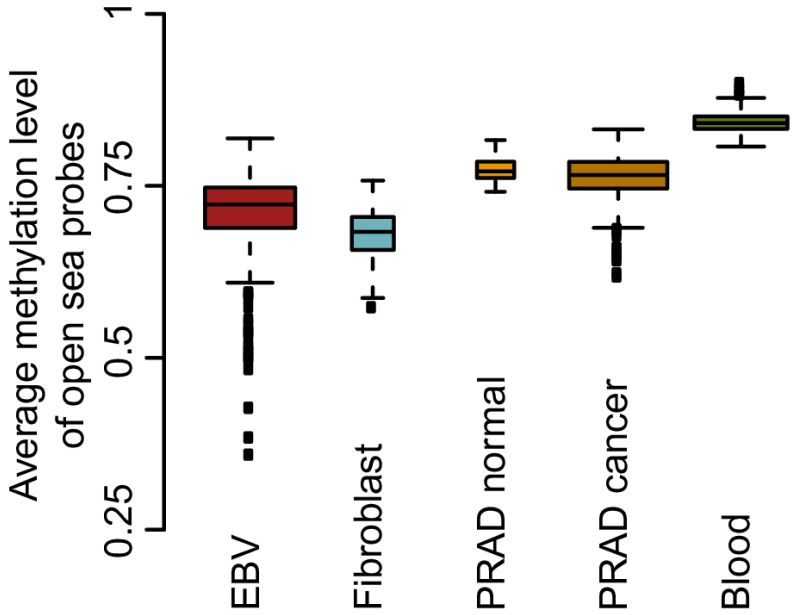


Figure 4.15: Between-sample variability in marginal methylation. For each dataset, the boxplot shows the distribution of average methylation levels of the open sea probes on the Beta value scale. We are able to estimate compartments for all datasets but the 450k-Blood dataset.

whole blood samples described previously.[77] The first eigenvector from this dataset is shown in Figure 4.16. It is immediately clear that this eigenvector looks different from the other datasets we present; it seems to be oscillating more rapidly. While compartments are cell-type specific, in our experience compartments from any two cell types are somewhat correlated, reflecting that large parts of the genome do not change compartment. For example, the correlation between HiC-EBV-2014 and HiC-IMR90-2014 is 0.66 with a domain agreement of 73.4%. In contrast, this 450k dataset from whole blood has a correlation and domain agreement of 0.27 and 59.7% with HiC-EBV-2014 and 0.27 and 59.6% with HiC-IMR90-2014. The data were were quantile normalized and adjusted for cell-type composition as described in,[77] but we also obtained and preprocessed the raw data to exclude that data processing was the cause of the poor performance. We note that the percent variance explained by the first eigenvector was only 57%, in contrast to 85% for the 450k-EBV dataset and 74% for the 450k-Fibroblast dataset. Based on our insights above, we hypothesized that the poor performance might be related to the lack of between-sample variability in marginal methylation, as shown in Figure 4.15. However, one dataset on primary prostate shows a similar degree of between-sample variability in marginal methylation and our method works for this dataset (see below).

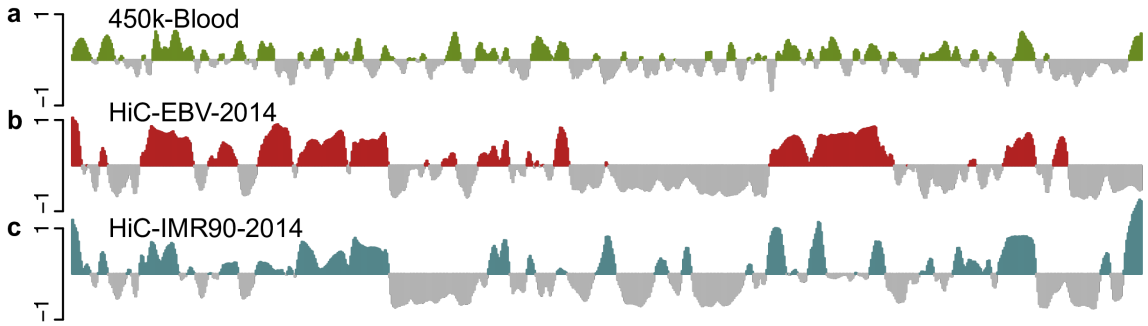


Figure 4.16: The methylation correlation signal of the 450k-Blood dataset does not correlate well with other datasets. The figure displays data on all of chromosome 14 at 100kb resolution. (a) The smoothed first eigenvector of the binned correlation matrix of the 450k-Blood dataset. (b) The first eigenvector of the HiC-EBV-2014 dataset. (c) The first eigenvector of the HiC-IMR90-2014 dataset. We see that (c) does not correlate well with (b) and (a).

4.2.7 Notes on processing of the DNA methylation data

We have analyzed a wide variety of DNA methylation data both from the Illumina 450k and Illumina 27k microarrays. For each dataset, it varies which kind of data is publicly available (raw or processed). If possible, we have preferred to process the data ourselves starting from the Illumina IDAT files. However, for several datasets, we had to use the original authors' preprocessing pipeline; see the Methods section for details.

We examined the impact of preprocessing methods on the estimated eigenvectors by using both functional normalization,[1] quantile normalization adapted to the 450k array[13] and raw (no) normalization; we did not find any substantial changes in the

results. The agreement between the eigenvectors using the different preprocessing methods is greater than 94% and we note that the agreement with Hi-C data is best using functional normalization. This might be caused by the ability of functional normalization to preserve large differences in methylation between samples,[1] which is what we observe in the closed compartment.

4.2.8 An application to prostate cancer

We applied these methods to Illumina 450k data on prostate adenocarcinoma (PRAD) from The Cancer Genome Atlas (TCGA). Quality control shows both normal and cancer samples to be of good quality. Since the normal prostate samples represent uncultured primary samples, we confirmed that this dataset has the same information in its long-range correlation structure as established above (Figure 4.17; compare with Figure 4.11).

We obtained a list of curated somatic mutations from TCGA and used them to compute simple estimates of the somatic mutation rate in each 100kb bin of the genome (ie. the elevated mutation rate in the cancer samples compared to normals). Since the list of somatic mutations was obtained using whole-exome sequencing (WXS) we identified the capture assay used in these experiments and used the capture regions from this specific assay to compute somatic mutation rates for each 100kb genomic bin by computing the number of somatic mutations per base captured in that bin. Because the capture assay is biased towards coding regions, the somatic

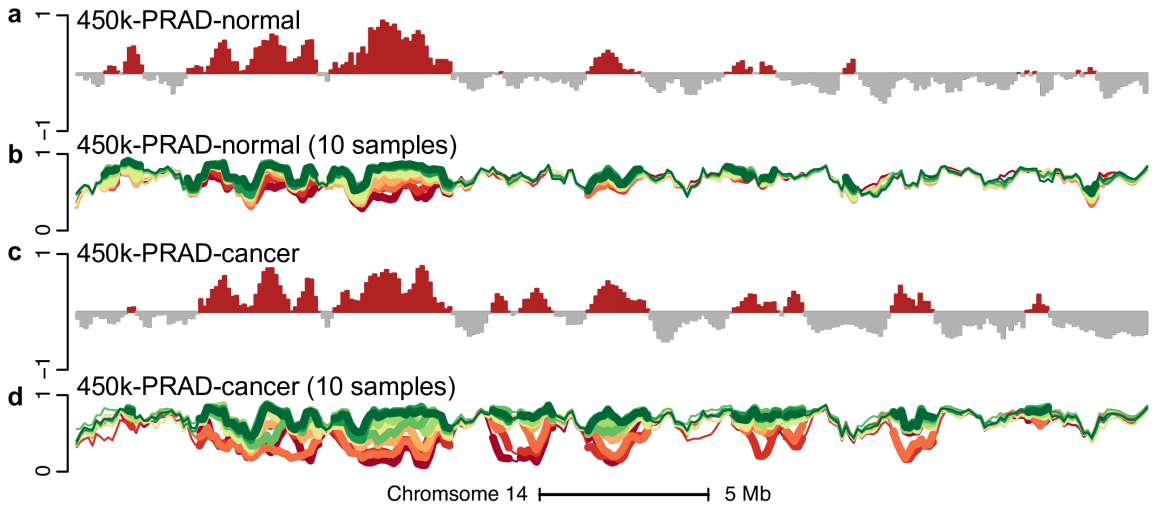


Figure 4.17: Comparison of the methylation levels and the Hi-C compartments signal for the 450k-PRAD datasets. Similar to Figure 4.11, but for the 450k-PRAD-cancer/normal datasets. (a) The first eigenvector of the binned methylation correlation matrix for the 450k-PRAD-normal dataset. (b) Average methylation signal on the beta scale for 10 selected samples for the 450k-PRAD-normal dataset. (c) Like (a) but for the 450k-PRAD-cancer dataset. (d) Like (b) but for the 450k-PRAD-cancer dataset.

mutation rates we computed can roughly be interpreted as the somatic mutation rate in coding regions per genomic bin. Many genomic bins have a somatic mutation rate of zero, and the number of bases captured varies between bins. In Figure 4.18, we display this somatic mutation rate vs. the value of the first eigenvector of the cancer data. In this figure, we display two smoothed loess curves; one curve includes bins with a mutation rate of zero, the other excludes them. Both curves show an elevated somatic mutation rate in the closed compartment of the cancer samples. This confirms previous observations about the relationship between mutation rates and open and closed chromatin,[78] including cancer.[79,80] To our knowledge, this is the first time a cancer-specific map of open and closed compartments based on primary samples has been derived; existing analyses depend on chromatin assays performed in ENCODE and Epigenomics Roadmap samples.[79,80]

While open and closed chromatin are cell-type specific, it is not surprising that a large percent of the genome (74%) are in the same compartment in both normal and cancer samples. To illustrate the added value of of a cancer-specific map of open and closed chromatin, we focused on the somatic mutation rate of bins which change compartment between normal and cancer. These bins are displayed in color in Figure 4.18. In Table 4.2, we computed the average somatic mutation rate across these bins. First, as shown above, the somatic mutation rate across the the part of the genome which is open in both cancer and normal was 54.1 compared to 97.2 for part of the genome which is closed in both cancer and normals. Focusing on

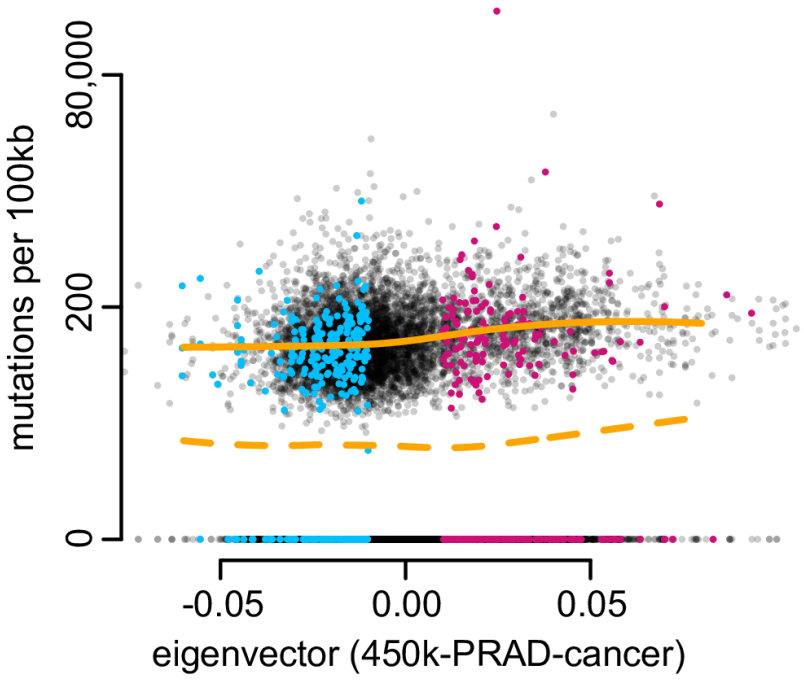


Figure 4.18: Relationship between A/B compartments and somatic mutation rate in prostate cancer. Somatic mutation rate for prostate cancer calculated using whole exome sequencing data from TCGA displayed against the first eigenvector of the 450k-PRAD-cancer dataset. The y-axis uses the hyperbolic arcsine scale which is similar to the logarithm for values greater than one. A large number of genomic bins have a mutation rate of zero. The pink line is a loess curve fitted to all the data and the orange line is a loess curve fitted only to bins with a strictly positive mutation rate. We observe an increase in somatic mutation rate in the closed compartment, as expected. Colored points represent bins which confidently change compartments between normal samples and cancer samples, blue is closed to open and red is open to closed. A bin confidently changes compartment if its associated eigenvector value has a magnitude greater than 0.01 (but with different signs) in both datasets.

the parts of the genome which change compartments, we observed that the somatic mutation rate in the parts of the genome which changes from closed to open in cancer was 58.0, close to the somatic mutation rate of 54.1 in the open compartment. Conversely, the somatic mutation rate for the parts of the genome changing from open to closed in cancer was 83.9, closer to the somatic mutation rate of 97.2 in the closed compartment. This result suggests that the somatic mutation rate of a genomic region which changes compartment depends only on the compartment status of the cancer samples. One possible explanation for this, is that changes chromatin accessibility happens relatively early in cancer development and that this change affect the somatic mutation rate; this is highly speculative. Our result illustrates the added value of obtaining cancer-specific maps of open and closed chromatin.

Table 4.2: Number of somatic mutations per 100kb in PRAD stratified by compartment.

		Normal	
		Open	Closed
Cancer	Open	54.1	58.0
	Closed	83.9	97.2

4.2.9 Compartments across human cancers

Using the method we have developed in this manuscript, it is straightforward to estimate A/B compartments across a wide variety of human cancers using data from TCGA. Figure 4.19 displays the smoothed first eigenvectors for chromosome 14 at

100kb resolution for eleven different cancers. Regions of similarity and differences are readily observed. We emphasize that TCGA does not include assays measuring chromatin accessibility such as DNase or various histone modifications. The extent to which these differences are associated with functional differences between these cancers is left for future work. Estimated compartments for all these cancer datasets are available online (see Methods).

4.2.10 Compartment prediction using DNase hypersensitivity data

Lieberman-Aiden *et al.*[58] established a connection between A/B compartments and DNase data, mostly illustrated by selected loci. Based on these results, we examined the degree to which we can predict A/B compartments using DNase hypersensitivity data. This data, while widely available from resources such as ENCODE, does not encompass as wide a variety of primary samples as the Illumina 450k methylation array.

We obtained DNase-seq data on 70 samples[81] from EBV-transformed lymphocytes from the HapMap project, as well as 4 experiments on the IMR90 cell line performed as part of the Roadmap Epigenomics project.[82] We computed coverage vectors for each sample and adjusted them for library size.

For each sample, we computed the signal in each 100kb genomic bin. To obtain

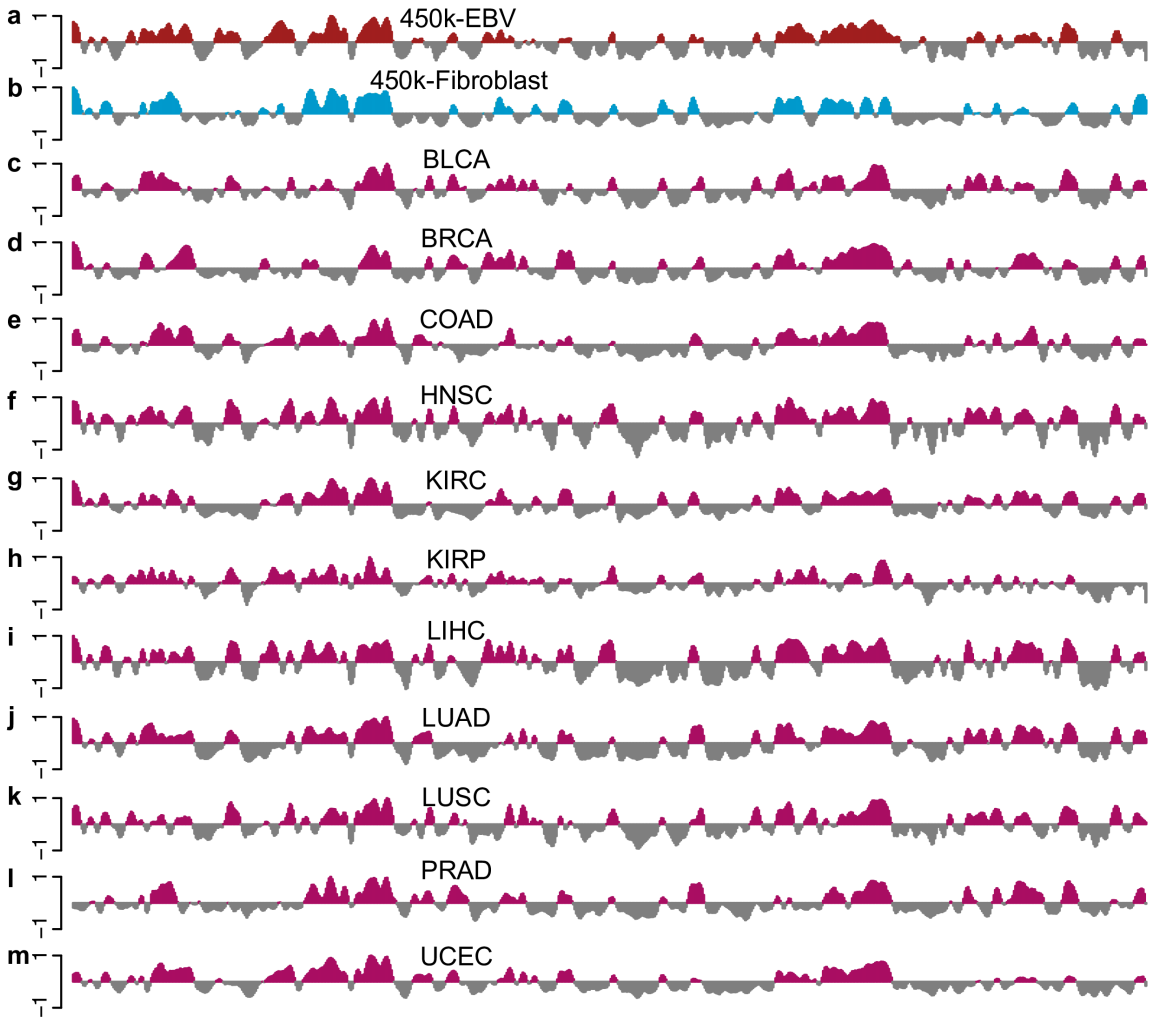


Figure 4.19: Estimated A/B compartments across several human cancers. The figure displays data on all of chromosome 14 at 100kb resolution. Each track represents the first eigenvector of the methylation correlation matrix for the corresponding dataset. The datasets depicted in (a) and (b) are the 450k-EBV and 450k-Fibroblast datasets. The datasets in (c-m) are cancer samples from TCGA for different cancers: (c) BLCA, (d) BRCA, (e) COAD, (f) HNSC, (g) KIRC, (h) KIRP, (i) LIHC, (j) LUAD, (k) LUSC, (l) PRAD, and (m) UCEC.

the average DNase signal, we averaged the signal across samples. The resulting mean signal is skewed towards positive values in the open compartment, and we therefore centered the signal by the median. The median was chosen as this has the best compartment agreement with Hi-C data. Figure 4.20 shows the result of this procedure, slightly modified for display purposes (the sign was changed to let high values be associated with the closed compartment; additionally very low values were thresholded). A good visual agreement is observed for both cell types; the correlation between Hi-C and the average DNase signal on chromosome 14 is 0.68 for EBV and 0.75 for IMR90 with a compartment agreement of 82% for EBV and 82% for IMR90.

Inspired by the success of considering long-range correlations for the 450k data, we examined whether this approach is useful for DNase data. We therefore computed the Pearson correlation matrix of the binned DNase signal; in contrast to the 450k data, we did not bin the correlation matrix as the signal matrix was already binned. The first eigenvector of this correlation matrix is highly skewed; we centered it by its median. Figure 4.20 shows the result of this procedure. For chromosome 14, we obtained a correlation between this centered eigenvector and the Hi-C eigenvector of 0.75 for EBV and 0.76 for IMR90 and a compartment agreement of 86% for EBV and 80% for IMR90. These results are similar to what we obtained using the average DNase signal.

We observed an association between GC content and average DNase signal (Figure 4.21); this is expected. There is a small between-sample variation in GC content

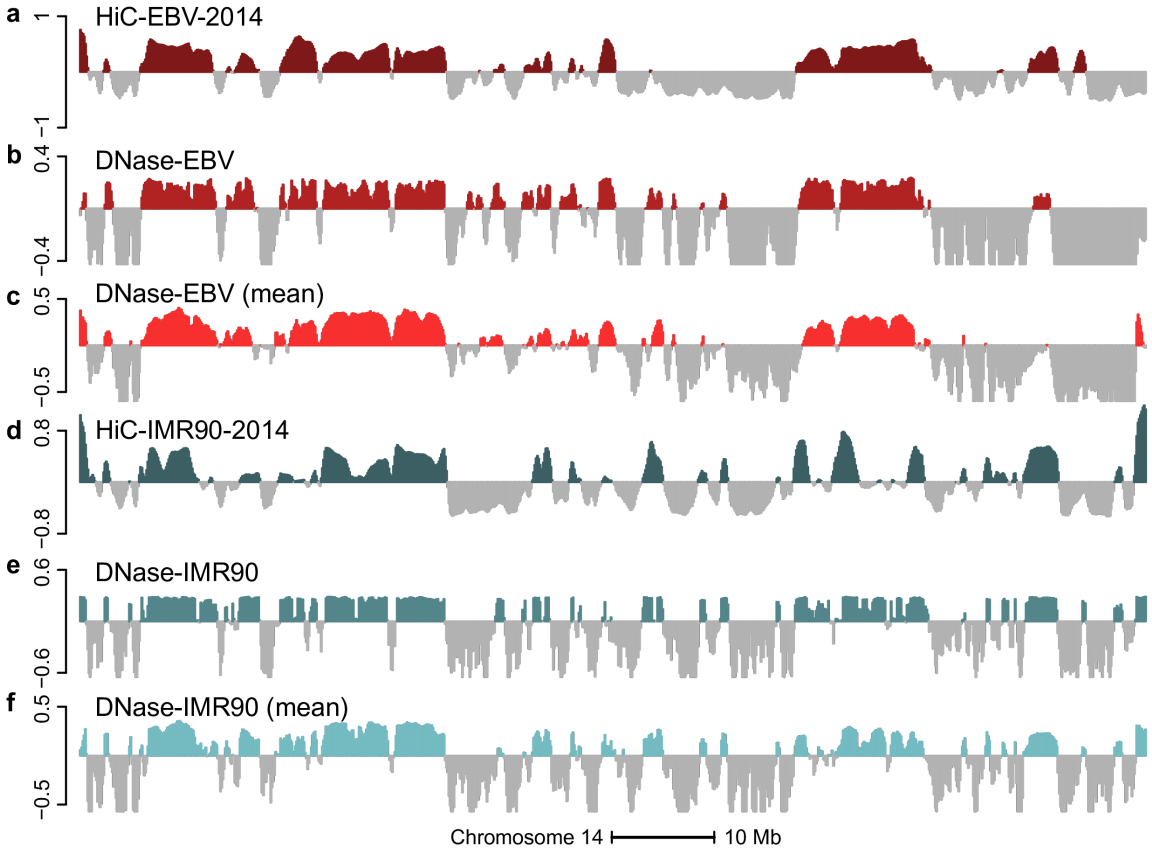
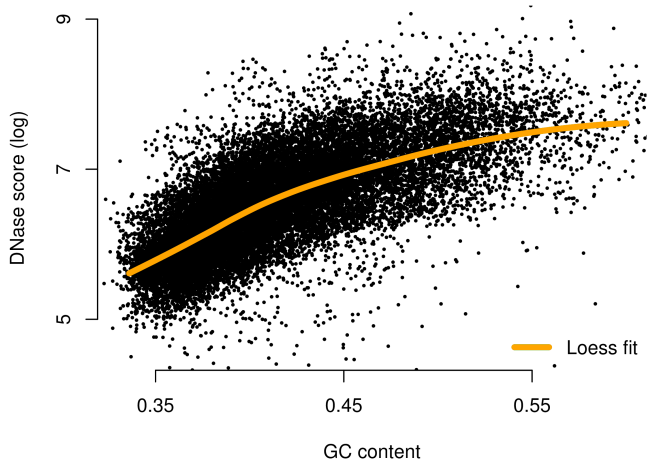


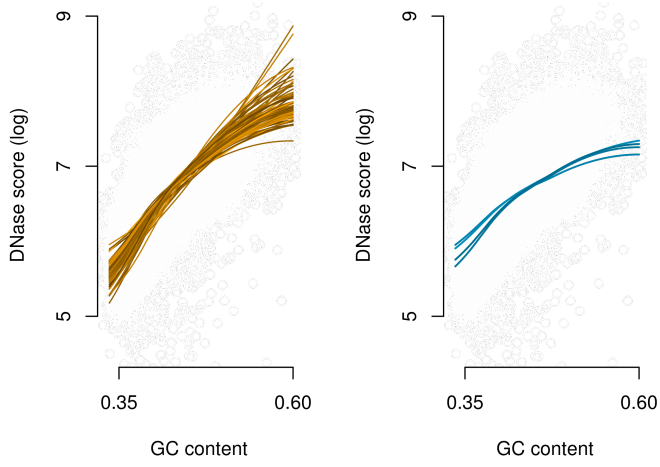
Figure 4.20: DNase data can predict A/B compartments revealed by Hi-C. The figure displays data on all of chromosome 14 at 100kb resolution. (a) The first eigenvector of the HiC-EBV-2014 dataset. (b) The smoothed first eigenvector of the correlation matrix of the binned DNase-EBV dataset after median centering (c) Average DNase signal across samples after binning and median subtraction. The sign of the signal was reversed for display purpose. (d) The first eigenvector of the HiC-IMR90-2014 dataset. (e) The smoothed first eigenvector of the correlation matrix of the binned HiC-DNase-IMR90 dataset after median centering. (f) Average DNase signal across samples after binning and median subtraction. The sign of the signal was reversed for display purpose. Both the average signal and correlation eigenvector are highly predictive of the Hi-C compartments for both cell types.

effect. It is easy to remove this GC content effect by estimating the effect using loess and subsequently regressing it out. Doing so led to much worse results when estimating compartments using the average DNase signal, but the results obtained using our correlation method were only slightly negatively impacted. To be precise, for the average DNase signal on chromosome 14, we got a correlation 0.35 for EBV and 0.69 for IMR90 with a compartment agreement of 69% for EBV and 78% for IMR90. For our correlation-based method, we got a correlation of 0.68 for EBV and 0.78 for IMR90 and a compartment agreement of 78% for EBV and 81% for IMR90.

To examine why the correlation-based approach works for DNase data, we performed the same investigation as for the 450k datasets. In Figure 4.22, we show the distribution of correlations stratified by compartment type. As for the DNA methylation data, the DNase data has high positive correlations between bins in the closed compartment, although the correlations in the DNase data are much higher. For DNA methylation data, correlations were close to zero between loci when at least one locus was in the open compartment. In contrast, the DNase data shows an almost uniform distribution of correlation values when one of the two loci are in the open compartment. In the same figure, we display the distribution of correlations when we used a sample-specific GC content effect correction; this correction changes the correlation substantially and suggests that some of the correlation structure is driven by GC content. Nevertheless, correcting for this effect slightly decreased our power to estimate the Hi-C compartments.



(a)



(b) **(c)**

Figure 4.21: Relationship between DNase scores and GC content. (a) Relationship between the log DNase score of one individual from the DNase-EBV dataset and GC content at the bin level, genome-wide (100kb resolution); loess fit is in orange. (b) Loess fits, as described in (a), for all 70 individuals from the DNase-EBV dataset (c) Loess fits for 4 replicates from the IMR90 cell line (DNase-IMR90 dataset)

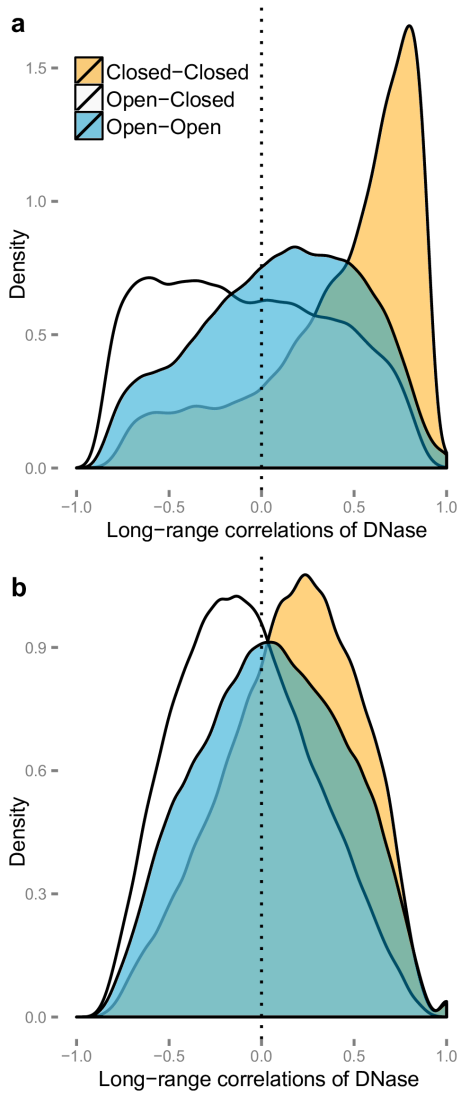


Figure 4.22: Densities of the correlations of DNase data. Chromosome 14 was binned at resolution 100kb. Depicted is the correlations of this data for the DNase-EBV dataset, stratified by compartment type. The open and closed compartments were defined using the HiC-EBV-2014 dataset. (a) represents the correlations without GC content correction (b) represents the correlations after GC content correction. This figure is similar to Figure 4.8.

Above, we have examined correcting for a sample specific GC content effect. It is also possible to directly regress out a GC content effect on the estimated eigenvector. Doing so, on both DNase data and Hi-C, does not decrease the correlation between the two eigenvectors. As discussed earlier in this manuscript, we do not recommend doing this, as we believe it might remove some biological signal.

4.2.11 Compartment prediction using single-cell epigenetic data

Experimental techniques for measuring epigenetics in a single cell are in rapid development. We have applied our methods to data from the few genome-wide, single-cell epigenetic experiments available. This includes data both on chromatin accessibility[69] and DNA methylation.[68]

Chromatin accessibility is measured by a single cell variant of an assay called assay for transposase-accessible chromatin (ATAC) sequencing,[83] which generates data similar to DNase hypersensitivity. From Cusanovich *et al.*[69] data is available on mixtures of two cell lines, GM12878 and HL60, but not on pure samples of one cell type. First, we developed a simple method for assigning single cells from this mixture to one of the two known cell lines, based on average accessibility of known, cell-type specific hypersensitive sites; this is a much more simple method than what is suggested in Cusanovich *et al.*[69] Using our method, we observed two distinct clusters of cells, and

most cells can easily be assigned unambiguously to a cell type using an arbitrary but seemingly sensible cutoff (Methods, Figure 4.23a). This yielded data on 2677 cells from the GM12878 cell line from one experiment. We next applied our correlation-based approach to this data; now the correlation is between single cells within the same cell line. Furthermore, the data consists of accessibility quantified over 195,882 hypersensitive sites the original authors derived from ENCODE data, with the accessibility of each site being a value of 0, 1, or 2. We summarized this data in 100kb bins (see Methods), not unlike our treatment of bulk DNase-seq data. On chromosome 14, we observed a correlation of 0.84 and a compartment agreement of 81% between the first eigenvector of this data and the first eigenvector from HiC-EBV-2014 data (Figure 4.23b,c). We observed that the three different types of correlations have different distributions, very different from other data types (Figure 4.23d). Closed-closed correlations are skewed towards negative values, while open-open correlations are shifted towards positive values.

Single-cell DNA methylation can be measured using a form of whole-genome bisulfite sequencing (WGBS) as described in Smallwood *et al.*[68] Due to technical limitations of the assay, the number of assayed cells is small. We have data on 20 individual mouse embryonic stem cells (mESC) cultured in serum conditions, with corresponding Hi-C data from a different source.[60] We generated a binned methylation matrix by averaging methylation values for open sea CpGs and discarded bins with little or no data (see Methods). We next applied our correlation-based approach to this data,

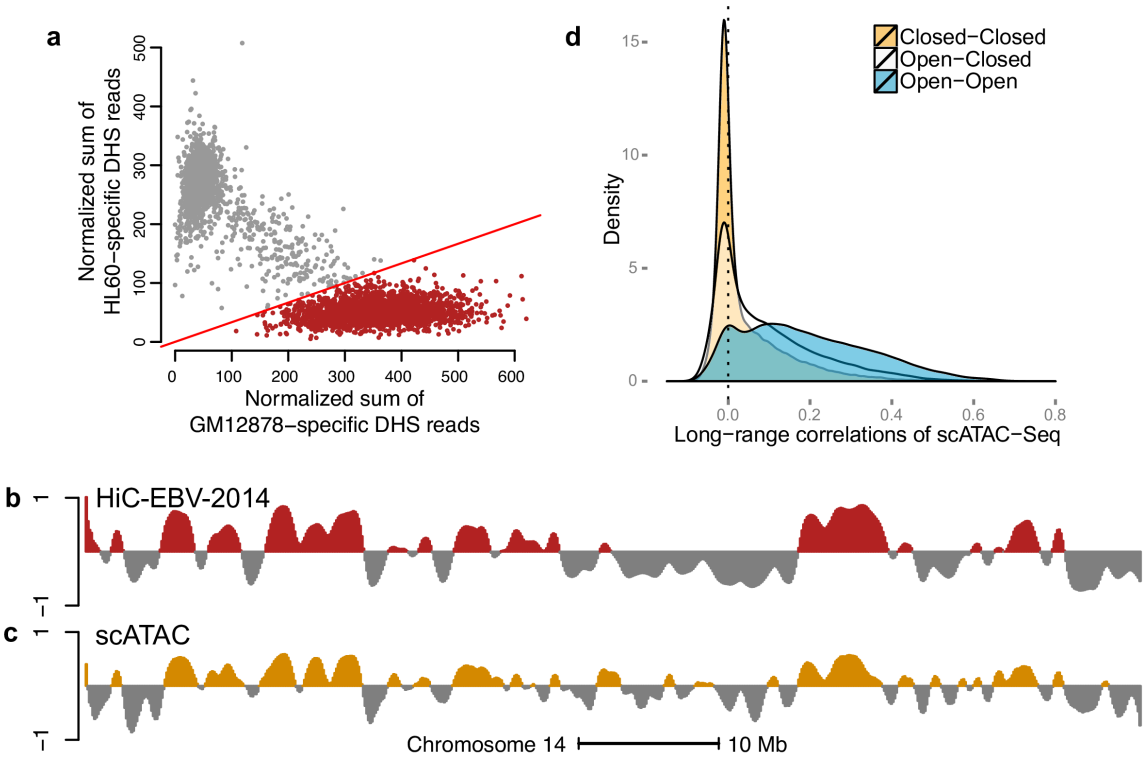


Figure 4.23: Single-cell ATAC-seq data. Data from a single experiment on a mixture of the GM12878 and HL60 cell lines described in.[69] (a) ENCODE DNAse-seq data was used to define hypersensitive sites (DHS) specific to these two cell lines. For each of these two sets of sites we computed the average number of ATAC-seq reads normalized by the total number of reads mapped to known DHS sites. The figures shows two distinct clusters; we arbitrarily selected the line $y = 1/3x$ to delineate cells from the GM12878 cell line (red points); this defines the scATAC-EBV data containing 2677 cells. (b) Estimated compartments on chromosome 14 at a resolution of 100kb using the HiC-EBV-2014 data. (b) Estimated compartments for the scATAC-EBV data. (d) Density of correlations for scATAC-EBV. We observe that the three different types of correlations have different distributions. Closed-closed correlations are skewed towards negative values, while open-open correlations are shifted towards positive values.

computing a correlation matrix across these 20 cells. On mouse chromosome 12, we observed a correlation of 0.61 and a domain agreement of 81%, using existing Hi-C data on the mouse embryonic stem cell line J1[60] (Figure 4.24a-c). An analysis of the pattern of correlation between loci in open and closed compartments showed some difference between the two distributions (Figure 4.24d), although both open-open and closed-closed are highly correlated in contrast to other data types. In contrast to what we observed for 450k data, loci in the open domain are still substantially positively correlated. We note that[68] show substantial between-cell heterogeneity in genomewide methylation across these 20 cells, depicted in Figure 4.24e. However, this heterogeneity of genomewide methylation was not observed for mouse ovulated metaphase II oocytes (MII) cells (Figure 4.24e); the correlation distribution is substantially different for this dataset (Figure 4.24d) and the first eigenvector of the correlation matrix only explains 19% of the variance, in contrast to 99% of the variance explained for mESC cells (Figure 4.24c). We do not have Hi-C data available for this cell type, but based on these observations we are doubtful that the first eigenvector accurately reflects the A/B compartments in this cell type.

4.3 Conclusion

In this work, we show how to estimate A/B compartments using long-range correlations of epigenetic data. We have comprehensively evaluated the use of data from

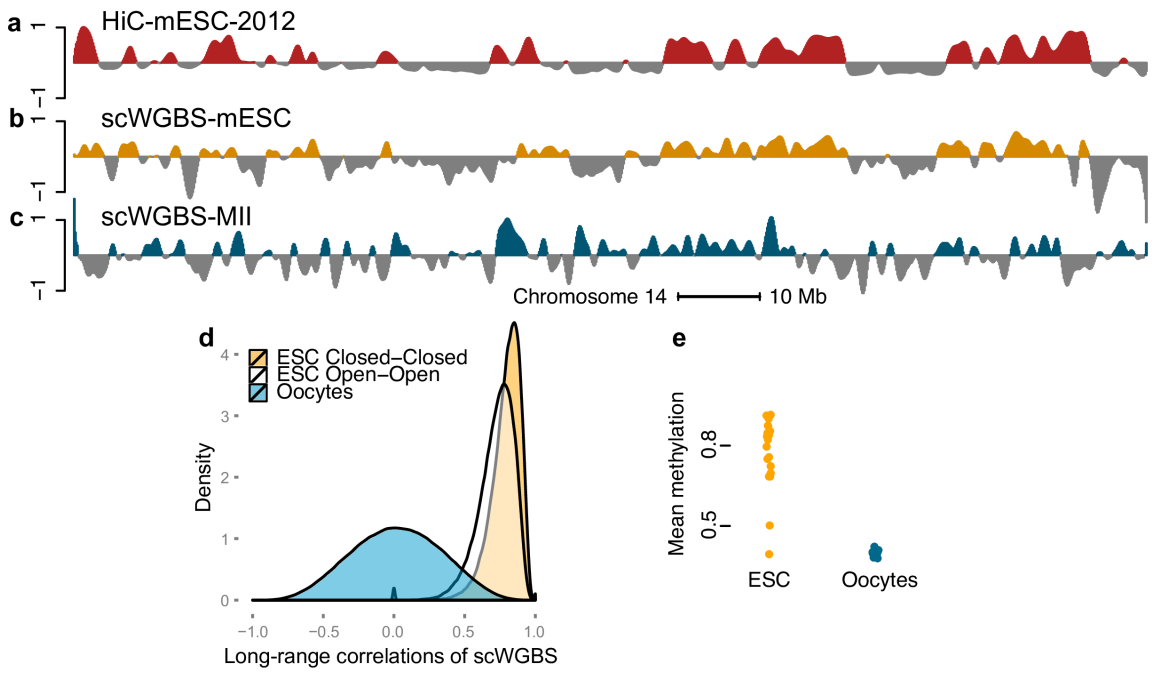


Figure 4.24: Single-cell WGBS data. Depicted is data from experiments in mouse embryonic stem cells. (a) Estimated compartments using the HiC-mESC-2012 data on chromosome 12 at a resolution of 100kb. (b) Estimated compartments using single-cell whole-genome bisulfite sequencing data from 20 mESC cells grown on serum. (c) The first eigenvector of a correlation matrix obtained using single-cell whole-genome bisulfite sequencing data from 12 ovulated metaphase II oocytes cells. (d) Density of correlations for data on mESC and MII cells; compartments are estimated using the HiC-mESC-2012 data. The two cell types have very different patterns. (e) Genome-wide methylation for 20 mouse ESC cells and 12 ovulated metaphase II oocytes (MII) cells. Substantial heterogeneity is observed for the former but not the later.

the Illumina 450k DNA methylation microarray for this purpose; such data is widely available on many primary cell types. Using data from this platform, we can reliably estimate A/B compartments in different cell types, as well as changes between cell types.

This result is possible because of the structure of long-range correlations in this type of data. Specifically, we found that correlations are high between two loci both in the closed compartment and low otherwise, and do not decay with distance between loci. This result only holds true for array probes measuring CpGs located more than 4kb from CpG islands, so-called open sea probes. This high correlation is the consequence of a surprising ranking of DNA methylation in different samples across all regions belonging to the closed compartment. We have replicated this result in an independent experiment using the Illumina 27k DNA methylation microarray.

We have furthermore established that A/B compartments can be estimated using data from DNase hypersensitivity sequencing. This can be done in two ways: first by simply computing the average DNase signal in a genomic region, and second by considering long-range correlations in the data, like for 450k array data. Again, we exploited the structure of long range correlations in this type of epigenetic data and, like the case for DNA methylation data, we found that correlations between loci both in the closed compartment are high, whereas correlations between other loci are approximately uniformly distributed. Again, this correlation is caused by a ranking of DNase signal in different samples across all regions belonging to the closed

compartment. Surprisingly, our method works both for biological replicates (EBV-transformed lymphocytes) but also on technical between-lab replicates of the same cell line (IMR90).

Finally we have established that our method works on single-cell epigenetic data, including single-cell ATAC-seq and single-cell WGBS. These experimental techniques are in their infancy; it is likely that additional data will allow us to tune aspects of our method to this type of data. Now, the correlation is between single cells as opposed to biological replicates of bulk cells. This potentially allows our method to be used on rare types of cells. During review of this paper, Buenrostro *et al.*[84] appeared in press, with the same conclusion as outs: single-cell ATAC-seq can reveal features of the Hi-C contact matrix.

Recently, clusters of DNA methylation under genetic control (GeMes) was described.[77] These clusters of highly correlated CpGs are different from the compartments described here. This work described 2,100 such clusters in whole blood ranging in size from 6bp to 50bp. Only 5 of these are greater than 10kb and 1,953 are smaller than 1kb.

Our approach is based on computing the first eigenvector of a (possibly binned) correlation matrix. It is well-known that this eigenvector is equal to the first left-singular vector from the singular value decomposition of the data matrix. The right-singular vector of this matrix is in turn equal to the first eigenvector of the sample correlation matrix; also called the first principal component. This vector has been

shown to carry fundamental information about batch effects.[85] Because of this relationship, we are concerned that our method might fail when applied to experiments that are heavily affected by batch effects; we recommend careful quality control of this issue before further analysis.

We have examined the impact of GC content on our method. It has previously been established that GC content is associated with A/B compartments.[71] This association can be removed computationally but we, and Imakaev *et al.*,[71] are concerned that it might remove biological signal. Nevertheless, our correlation-based method shows good agreement between compartments estimated using Hi-C data and estimated using other epigenetic data, whether or not the GC content effect is removed. We have also established that GC content itself is not the main driver of long-range correlations.

The reason our method works is because of a surprising, consistent ranking of different samples across all regions belonging to the closed compartment (and only the closed compartment). By comparison with additional 27k methylation array experiments, we have shown that this ranking is not a technical artifact caused by (for example) hybridization conditions.

We caution that while we have had success with our method on many datasets, we have seen failures as we described in our analysis of the dataset on whole blood measured on 450k. This raises the issue of when and why the method fails. In recent work, we studied colon cancer and EBV transformation of lymphocytes using WGBS.[10,39] In

these two systems, we observed global hypomethylation as well as an increased variation in global methylation levels in colon cancer and EBV-transformed lymphocytes when compared to normal-matched samples from the same person. However, we saw minimal variation in global methylation between 3 normal samples in both systems. This type of observation is the same as what we see for the single-cell WGBS data on mESC and MII cells (Figure 4.24e); there is substantial heterogeneity in global methylation for mESC and not for MII where the method fails. The same observation is reflected in Figure 4.16 where we, as expected, see a substantial variation in cancer, EBV-transformed lymphocytes and cultured fibroblasts, and substantially less variation in samples from whole blood. However, our method does work on normal prostates which also show minimal variation in global methylation, suggesting that this is not the explanation for the failure. More work is needed to firmly establish whether this ranking holds true for most primary tissues or might be a consequence of oncogenesis, manipulation in culture or a kind of unappreciated batch effect, affecting a well-defined compartment of the genome. We note that the cause of the ranking does not matter; as long as the ranking is present it can be exploited to reconstruct A/B compartments.

The functional implications of A/B compartments have not been comprehensively described; we know they are associated with open and closed chromatin,[58] replication timing domains,[63,86] changes during mammalian development and are somewhat associated with gene expression changes.[65] Our work makes it possible to more com-

prehensively study A/B compartments, especially in primary samples. We have illustrated this with a brief analysis of the relationship between A/B compartments and somatic mutation rate in prostate adenocarcinoma.

4.4 Methods

4.4.1 Infinium HumanMethylation450 BeadChip

We use the standard formula $\beta = M/(M+U+100)$ for estimating percent methylation given (un)methylation intensities U and M. Traditionally, the term M-value is used for the logit transform of the beta value, and we do the same.

With respect to CpG density, the 450k array probes fall into 4 categories that are related to CpG islands. CpG Island probes (30.9% of the array) are probes located in CpG islands, shore probes (23.1%) are probes within 2 kbs of CpG islands, and shelf probes (9.7%) are probes between 2 kbs and 4 kbs from CpG islands. Open sea probes (36.3%) are the rest of the probes. We use the term CpG resort probes to refer to the union of island, shore and shelf probes; in other words non-open sea probes.

4.4.2 Methylation Data

Also see Table 4.3.

The 450k-Fibroblast dataset. The study contains 62 samples from primary skin fibroblasts from.[75] The raw data (IDAT files) are available on GEO under the accession number [GEO:GSE52025].

The 450k-EBV dataset. The study contains 288 samples from EBV-transformed lymphoblastoids cell lines (LCL)[73] from three HapMap populations: 96 African-American, 96 Han Chinese-American and 96 Caucasian. The data are available on GEO under the accession number [GEO:GSE36369].

The 450k-Blood dataset. The study contains 305 samples from whole blood.[77] The data are available on GEO under the accession number [GEO:GSE54882].

The 27k-EBV Vancouver dataset. The study contains 180 samples from EBV-transformed lymphoblastoid cell lines (LCL)[87] from two HapMap populations: 90 individuals from Northern European ancestry (CEU), and 90 individuals from Yoruban (West African) ancestry (YRI). The processed data are available on GEO under the accession number [GEO:GSE27146].

The 27k-EBV London dataset. The study contains 77 EBV-transformed lymphoblastoid cell lines (LCL) assayed in duplicates.[88] Individuals are from the Yoruba HapMap population, and 60 of them are also part of the 27k-EBV Vancouver dataset. The raw data (IDAT files) are available on GEO under the accession number [GEO:GSE26133].

The 450k-PRAD-normal, 450k-PRAD-cancer datasets. At the time of download, the dataset contained 340 prostate adenocarcinoma cancer samples from

Table 4.3: Methylation data sources.

Dataset	Cell type	n	Platform	Acc.	Reference
450k-Fibroblast	Fibroblast (skin)	62	450k	GSE52025	[73]
450k-EBV	LCL (EBV)	288	450k	GSE36369	[75]
450k-Blood	Whole blood	305	450k	GSE54882	[77]
27k-EBV Vancouver	LCL (EBV)	180	27k	GSE27146	[87]
27k-EBV London	LCL (EBV)	160	27k	GSE26133	[88]
450k-PRAD-normal	Prostate (normal)	49	450k	TCGA	[70]
450k-PRAD-cancer	Prostate (cancer)	340	450k	TCGA	[70]

The Cancer Genome Atlas (TCGA)[70] along with 49 matched normal samples. We used the Level 1 data (IDAT files) available through the TCGA Data portal.[89]

The PMDs-IMR90 dataset. The partially methylated domain (PMD) boundaries from IMR90[90] are available at.[91]

The EBV hypomethylation blocks dataset. Hypomethylated blocks between EBV-transformed and quiescent B-cells were obtained form a previous study.[39] Only blocks with a family-wise error rate equal to 0 were retained (see the reference).

4.4.3 Processing of the methylation data

For the 450k-Fibroblast and 450k-PRAD datasets, we downloaded the IDAT files containing the raw intensities. We read the data into R using the illuminaio package.[50] For data normalization, we use the minfi package[13] to apply the noob background subtraction and dye-bias correction[17] followed by functional normalization.[1] We have previously shown[1] that functional normalization is an adequate between-array normalization when global methylation differences are expected between individuals. For

the 450k-EBV dataset, only the methylated and unmethylated intensities were available, and therefore we did not apply any normalization. For the 450k-Blood dataset, data were quantile normalized and then adjusted for estimated cell proportions and sex as described in.[77] For the 27k-EBV London dataset, IDAT files were available, and we applied the noob background correction and dye-bias correction as implemented in the methylumi package.[17] For the 27k-EBV Vancouver dataset, IDAT files were not available and therefore we used the provided quantile normalized data as discussed in.[87]

For quality control of the samples, we used the packages minfi and shinyMethyl[13,55] to investigate the different control probes and potential batch effects. All arrays in all data sets passed the quality control. After normalization of the 450k array, we removed 17,302 loci that contain a SNP with an annotated minor allele frequency greater than or equal to 1% in the CpG site itself or in the single-base extension site. We used the UCSC Common SNPs table based on dbSNP 137. The table is included in the minfi package.

For the analysis of the 27k array data, we only considered probes that are also part of the 450k array platform (25,978 probes retained in total) and applied the same probe filtering as discussed above.

4.4.4 Construction of 450k correlation matrices

For each chromosome, we start with a $p \times n$ methylation matrix M of p normalized and filtered loci and n samples. We use M-values as methylation measures. We compute the $p \times p$ matrix of pairwise probe correlations $C = \text{cor}(M')$, and further bin the correlation matrix C at a predefined resolution k by taking the median correlation for between CpGs contained in each of two bins. Because of the probe design of the 450k array, some of the bins along the chromosome do not contain any probes; these bins are removed. As discussed in the Results section, the correlations of the open sea probes are the most predictive probes for A/B compartments, and therefore the correlation matrix is computed using only those probes (36.3% of the probes on the 450k array). The inter-chromosomal correlations are computed similarly.

4.4.5 Hi-C Data

Samples are described in Table 4.4.

4.4.6 Processing of the Hi-C data

For the datasets HiC-EBV-2014, HiC-K562-2014 and HiC-IMR90-2014 from,[64] we used the raw observed contact matrices that were constructed from all read pairs that map to the human genome hg19 with a MAPQ ≥ 30. These contact matrices

Table 4.4: Hi-C data sources.

Dataset	Cell line	Cell type	Acc.	Reference
HiC-EBV-2009	GM06990	LCL (EBV)	GSE18199	[58]
HiC-EBV-2013	GM12878	LCL (EBV)	GSE48592	[94]
HiC-EBV-2014	GM12878	LCL (EBV)	GSE63525	[64]
HiC-IMR90-2013	IMR90	Fibroblast (lung)	GSE43070	[61]
HiC-IMR90-2014	IMR90	Fibroblast (lung)	GSE63525	[64]
HiC-Fibro-Skin	-	Fibroblast (skin)	GSE41763	[93]
HiC-Fibro-HFF1	HFF-1	Fibroblast (skin)	E-MTAB-1948	[62]
HiC-K562-2009	K562	Leukemia	GSE18199	[58]
HiC-K562-2014	K562	Leukemia	GSE63525	[64]
HiC-mESC-2012	J1	mESC	GSE35156	[60]

are available in the supplementary files of the GEO deposition [GEO:GSE63525]. For the HiC-IMR90-2013 dataset from,[61] we used the online deposited non-redundant read pairs that were mapped with Bowtie[92] to human genome hg18 using only the first 36 bases. For the HiC-EBV-2009 and HiC-K562-2009 datasets from Lieberman-Aiden *et al.*,[58] we used the mapped reads deposited on GEO under the accession number [GEO:GSE18199]. Reads were mapped to human genome hg18 using Maq, as described. For the Fibro-Skin dataset from,[93] we merged the reads from two individuals with normal cells (Father and Age-Matched control). We used the processed reads of the GEO deposition [GEO:GSE41763] that were mapped using Bowtie2 to the hg18 genome in an iterative procedure called ICE previously described in.[71] For the HiC-mESC-2012 dataset, we used the mapped reads deposited on GEO under the accession number [GEO:GSE35156]; reads were mapped to the mm9 genome.

For the HiC-EBV-2012 dataset from[94] and the HiC-Fibro-HFF1 dataset from,[62] we downloaded the SRA experiments containing the FASTQ files of the raw reads.

We mapped each end of the paired reads separately using Bowtie to the hg18 genome with the –best mode enabled. We kept only paired reads with both ends mapping to the genome.

For all datasets but the Hi-C datasets from,[64] we used the liftOver tool from UCSC to lift the reads to the human genome hg19 version for consistency with the 450k array. Reads from[64] were already mapped to the hg19 genome.

4.4.7 Construction of Hi-C matrices

As a first step, we build for each chromosome an observed contact matrix C at resolution k whose (i, j)'th entry contains the number of paired-end reads with one end mapping to the i-th bin and the other end mapping to the j-th bin. The size of the bins depends on the chosen resolution k. We remove genomic bins with low coverage, defined as bins with a total count of reads less than 10% of the total number of reads in the matrix divided by the number of genomic bins. This filtering also ensures that low mappability regions are removed.

To correct for coverage and unknown sources of biases, we implemented the iterative correction procedure called ICE[71] in R. This procedure forces bins to have the same experimental visibility. We apply the normalization procedure on a chromosome basis and noted that for each Hi-C dataset, the iterative normalization converged in less than 50 iterations. For the purpose of estimating A/B compartments, we further normalize the genome contact matrix by the observed-expected procedure,[58] where

Table 4.5: DNase-Seq data sources.

Dataset	Cell line	n	Acc.	Reference
DNase-EBV	LCL (EBV)	70	GSE31388	[81]
DNase-IMR90	Fibroblast (lung)	4	GSE18927	[82]

each band of the matrix is divided by the mean of the band. This procedure accounts for spatial decay of the contact matrix.

4.4.8 DNase-Seq data

Also see Table 4.5.

The DNase-EBV dataset. The study contains 70 biological replicates of EBV-transformed lymphoblastoid cell lines (LCL)[81] from the HapMap Yoruba population. The data are deposited on GEO under the accession number [GEO:GSE31388] and raw files are available at.[95]

The DNase-IMR90 dataset. The dataset is composed of 4 technical replicates of the IMR90 fetal lung fibroblast cell line available on GEO under the accession number [GEO:GSE18927].

Processing of the DNase-Seq data and construction of the correlation matrices

For the DNase-EBV dataset from,[81] we downloaded the raw reads in the HDf5 format for both the forward and reverse strands. We converted the reads to bedGraph, lifted the reads to the hg19 genome and converted the files to bigWig files using the UCSC tools. For the DNase-IMR90 dataset, we used the raw data already provided in the bigWig format. Reads were mapped to the hg19 genome. For both datasets, data were read into R by using the rtracklayer package.[96] To adjust for library size, we normalized each sample by dividing the DNase score by the total number of reads. For each sample, we constructed a normalized DNase signal at resolution 100kb by taking the integral of the coverage vector in each bin. This was done using BigWig files and the rtracklayer package in R.[96] All DNase datasets have the same read length within experiment (EBV/IMR90). This results in a $p \times n$ signal data matrix where p is the number of bins for the chromosome, and n the number of samples. We defined the average DNase signal as the across-sample mean of the signal matrix. The DNase correlation matrix is the $p \times p$ Pearson correlation matrix of the signal matrix.

4.4.9 GC content correction of the DNase data

For GC content correction of the DNase data, we fitted a loess curve of the DNase signal against the bin GC content for each sample differently and regressed out the fitted relationship.

4.4.10 Single-cell ATAC-seq data

Single-cell ATAC-seq data was obtained from GEO under the accession number [GEO:GSE68103] described in.[69] We used data processed by the authors, specifically the file `GSM1647124_CtlSet1.dhsmatrix.txt.gz`. This experiments represents data on a mixture of two cell lines: GM12878 and HL60. We use data processed by the authors of the paper, which consists of a matrix of accessibility across 195882 known hypersensitive sites (from ENCODE) and 4538 cells. Each hypersensitive site is furthermore characterized as being specific to GM12878, specific to HL60 or common across the two cell types. To classify each cell to a cell type, we computed the total number of reads in each of the cell type specific hypersensitive sites. This yields two numbers per cell. These number are further normalized by (1) the total number of reads in all hypersensitive sites scaled to 2000 reads (slightly more than the median number of reads per cell) and (2) the number of cell type specific hypersensitive sites scaled to 50,000 sites. The final scale is the number of reads mapped for a cell with a read depth of 2,000 and a cell type with 50,000 hypersensitive sites. These numbers

Table 4.6: Single-cell epigenetic data sources.

Dataset	Cell type	Organism	n	Acc.	Reference
scATAC-EBV	LCL (EBV)	Human	2679	GSE68103	[69]
scWGBS-mESC	mESC	Mouse	20	GSE56879	[68]
scWGBS-MII	MII	Mouse	12	GSE56879	[68]

are displayed in Figure 4.23a. Cells are assigned to the GM12878 cell type if they have more than 3 times as many normalized reads for this cell type, compared to HL60; in other words if they are below the $y = 1/3x$ line in the figure. Subsequently we discarded hypersensitive sites which had no reads in any of the cells and obtained 631 bins at a resolution of 100kb on chromosome 14. Eigenvectors were computed and smoothed as described below.

4.4.11 Single-cell WGBS data

Single-cell whole genome bisulfite sequencing data was obtained from GEO under the accession number [GEO:GSE56879] described in.[68] We used data processed by the authors, specifically the files `GSM1370555_Ser_X.CpG.txt.gz` where `X` takes value 1 to 20. These files describe the single CpG methylation levels of 20 individual cells for mouse embryonic stem cells cultured in serum conditions. We removed CpGs within 4kb of a CpG Island (using the CpG Islands defined in[97]), as we did for the 450k methylation array data. We next binned the genome in 100kb bins and computed, for each bin, the average methylation value across all CpGs in the bin. Bins with a total coverage of less than 100 was removed from analysis. This resulted

in a binned methylation matrix which was used to compute an empirical correlation matrix. Eigenvectors were computed and smoothed as described below.

4.4.12 Eigenvector analysis

To obtain eigenvectors of the different matrices from Hi-C, DNA methylation and DNase data, we use the non-linear iterative partial least squares (NIPALS) algorithm implemented in the mixOmics package in R.[98] Each eigenvector is smoothed by a moving average with a 3-bin window, with the following exceptions. For the 450k data, we used two iterations of the moving average smoother. For the single-cell epigenetic data, we used a window size of 5 bins with two iterations of the moving average smoother for ATAC-seq and 3 iterations for WGBS.

When we compare eigenvectors from two different types of data, we only consider bins which exists in both data types; some bins are filtered out in a data-type dependent manner for example because of absence of probes or low coverage. This operation slightly reduces the number of bins we consider in each comparison.

Because the sign of the eigenvector is arbitrarily defined, we use the following procedure to define a consistent sign across different chromosomes, datasets and data types. For Hi-C data and DNase data, we correlate the resulting eigenvector with the eigenvector from Lieberman-Aiden *et al.*;[58] changing sign if necessary to ensure a positive correlation. For DNA methylation data, we use that the long-range correlations are significantly higher for the closed-closed interactions. We therefore ensure that

the eigenvector has a positive correlation with the column sums of the binned correlation matrix; changing sign if necessary. This procedure results in positive values of the eigenvector being associated with closed chromatin and the 'B' compartment as defined in Lieberman-Aiden *et al.*[58] (in this paper they ensure that negative values are associated with the closed compartment).

To measure the similarity between two eigenvectors, we use two measures: correlation and compartment agreement. The correlation measure is the Pearson correlation between the smoothed eigenvectors. The compartment agreement is defined as the percentage of bins that have the same eigenvector sign, interpreted as the percentage of bins that belong to the same genome compartment (A or B) as predicted by the two eigenvectors. Occasionally, this agreement is restricted to bins with an absolute eigenvector value greater than 0.01 to discard uncertain bins.

Because open-chromatin regions have very high DNase signal in comparison to closed chromatin regions, the DNase signal distribution is highly skewed to the right; therefore we center both the average signal and the first eigenvector by subtracting their respective median, before computing the correlation and agreement.

4.4.13 Somatic mutations in PRAD

We obtained a list of somatic mutations in PRAD from the TCGA data portal.[89] Several lists exists; we used the Broad Institute curated list: `broad.mit.edu__IlluminaGA_curated_DNA_sequencing_level2.maf`. To obtain capture regions, we queried the

CGHub website[99] and found that all samples were profiled using the same capture design described in the file `whole_exome_agilent_1.1_refseq_plus_3_boosters.targetIntervals.be` obtained from the CGHub bitbucket account.

Somatic mutation rates in each 100kb genomic bin were computed as the number of mutations inside each bin, divided by the length of the capture regions inside the bin.

4.4.14 Data

Estimated compartments for TCGA cancer data are available at `https://github.com/Jfortin1/TCGA_AB_Compartments`. We processed 450k IDAT files from TCGA with NOOB[17] followed by functional normalization[1] as implemented in the minfi[13] package. Compartments were estimated using `compartments()` of minfi version 1.15.11.

4.4.15 Software

Software for performing the analysis of 450k methylation arrays described in this manuscript have been added to the minfi package[13] version 1.15.11 or greater, available through the Bioconductor project.[100,101] The main function is `compartments()`.

Chapter 5

Interactive visualization of DNA methylation data

This chapter describes work published in a separate form in the journal *F1000Research*, with co-authors Elana J. Fertig and Kasper D. Hansen.

5.1 Introduction

The recent release of the R package *shiny*[170] has substantially lowered the barriers to interactive visualization in R, opening the door to interactive exploration of high-dimensional genomic data.

DNA methylation is an epigenetic mark, and changes in DNA methylation have been associated with various diseases, such as cancer.[171] For DNA methylation data,

thousands of samples from the state-of-the-art Illumina 450k methylation array[2] have been generated and are accessible online from The Cancer Genome Atlas (TCGA) and through the Gene Expression Omnibus (GEO). This array has a series of probes used to measure a methylation and an unmethylation signal for a series of loci. Probes are designed using two main chemistries resulting in a challenging array design, essentially a mix of a two color and a one color array discussed in Bibikova *et al.*[2] Analysis of data from this array requires careful quality control and pre-processing that account for these distinct chemistries. The assessment of these steps could benefit from an interactive visualization tool.

Our solution is *shinyMethyl*, an interactive visualization package for 450k arrays, based on the packages *minfi*[13] and *shiny*.[170] The goal of shinyMethyl is two-fold; (1) to help with quality assessment and (2) to help with assessing the effect of pre-processing. We use pre-computation to enable interactive visualization of thousands of samples to circumvent computational bottlenecks during data exploration. The pre-computation can happen on a large computing server and the resulting data object can be used for interactive visualization on a laptop. Quality control and pre-processing large 450k datasets become easy and intuitive with *shinyMethyl*.

5.2 Methods

The first step of *shinyMethyl* is pre-computation of various summaries of the 450k array data, using the function `shinySummarize`. This pre-computation is run on raw (not pre-processed) data and – optionally – pre-processed data, resulting in either one or two summary objects, as described below. These summary objects, called `shinyMethylSet`, are saved in a platform-independent format. The interactive interface is then launched via the function `runShinyMethyl`. The function requires a `shinyMethylSet` containing the summary data from the raw data. In addition, the function accepts as a second argument a `shinyMethylSet` that contains summaries from pre-processed data, in which case both raw and pre-processed data will be displayed in the interactive interface. Figure 5.1 illustrates the *shinyMethyl* workflow.

5.2.0.1 Raw data summarization

Summarizing the raw data uses the *minfi*[13] and *illuminaio*[50] R packages to parse Illumina IDAT files into a *minfi* object called `RGChannelSet`. `shinySummarize` operates on this `RGChannelSet` and the summarization object created by this function is 35x smaller than the full data representation in *minfi*; 1,000 samples use 205 MB. Specifically, the summarized data contain the quantile distributions of the raw intensities for the unmethylated (U) and methylated (M) channels, copy numbers (CN = M + U), Beta values (Beta) and M values (M-Val). The object contains also the raw control

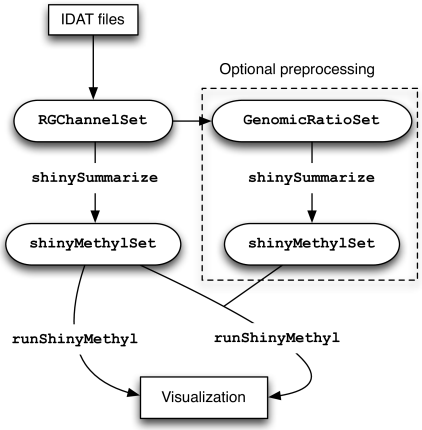


Figure 5.1: The workflow of *shinyMethyl*. IDAT files are parsed using *minfi* and *illuminaio* into a `RGChannelSet`. This object is summarized using `shinySummarize`. Optionally, the data are pre-processed and the pre-processed data are summarized. For visualization, `runShinyMethyl` is used on either one or two sets of summarized data.

probes intensities and the results of the principal component analysis performed on the autosomal Beta values. The function also extracts the phenotype variables stored in the `RGChannelSet`. The summarization is done separately by probe types (I and II, see Bibikova *et al.*[2]) and for sex chromosomes. An S4 class, called `shinyMethylSet`, is used to represent the data in R, and this object is independent of the operating system. The *shinyMethyl* interface is launched by passing the `shinyMethylSet` to the function `runShinyMethyl`. An example of the interface is shown in Figure 5.2.

5.2.0.2 Pre-processed data summarization (optional)

Summarizing pre-processed data in *shinyMethyl* operates on an S4 object in *minfi* termed `GenomicRatioSet`. The summaries of the pre-process data are stored in an ad-

ditional `shinyMethylSet`. Again, the summarized data object is substantially smaller than the full data representation in *minfi*. If this `shinyMethylSet` is also included in the `runShinyMethyl` command, the summaries of the pre-processed data are automatically added to the *shinyMethyl* interface. This option represents a powerful diagnostic tool to assess the global performance of a normalization method, such as plate effect correction (Figure 5.2), or preservation of the expected biological differences between different tissues or conditions (Figure 5.3).

5.2.1 Quality control assessment

Once the DNA methylation data have been summarized, *shinyMethyl* offers three interactive plots for quality control. These plots react conjointly to the user mouse: (1) a density plot of the M/Beta values, (2) a QC plot proposed in *minfi* and (3) a plot of control probes intensities. The samples are colored by a phenotype variable selected by the user. The three plots together allow the user to select aberrant samples, whose array identifiers are saved into a csv file for exclusion in subsequent analyses (outside of *shinyMethyl*). An example of quality control panel is presented in Figure 5.2 in which summaries from the TCGA head and neck squamous cell carcinoma (HNSCC) samples are colored by batch; *shinyMethyl* allows to observe significant batch effects, a source of obscure variation that has critical consequences in downstream analysis.[22]

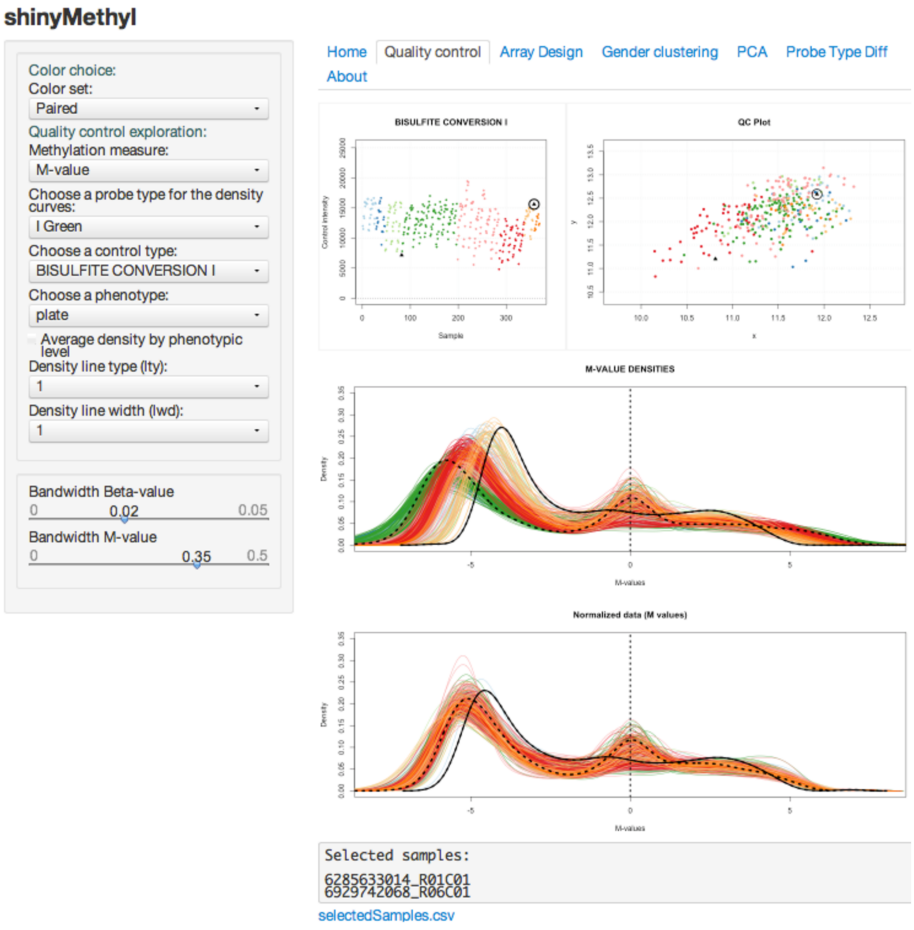


Figure 5.2: The shinyMethyl user interface for quality control. The interface shows an example interactive visualization of batch effects and quality control (TCGA head and neck squamous cell carcinoma, HNSCC dataset). The interface is divided into a user menu and a plotting area. (a) A menu containing a number of user-settable visualization parameters. The "phenotype" is set to "plate" which makes the color scheme reflect batch. The four plots (b-e) are interactive and react simultaneously to the user mouse clicks, so that samples selected on one plot are immediately highlighted on the additional plots. The solid lines in black represents the sample(s) currently selected by the user and match the dot circled in black on (b,c). The dashed lines in black represents another sample, previously selected by the user and match the black dot without the circle. (b) Average negative control probes intensities; (c) the median intensity of the M channel against the median intensity of the U channel; (d-e) M-value densities for Infinium I probes before and after functional normalization.

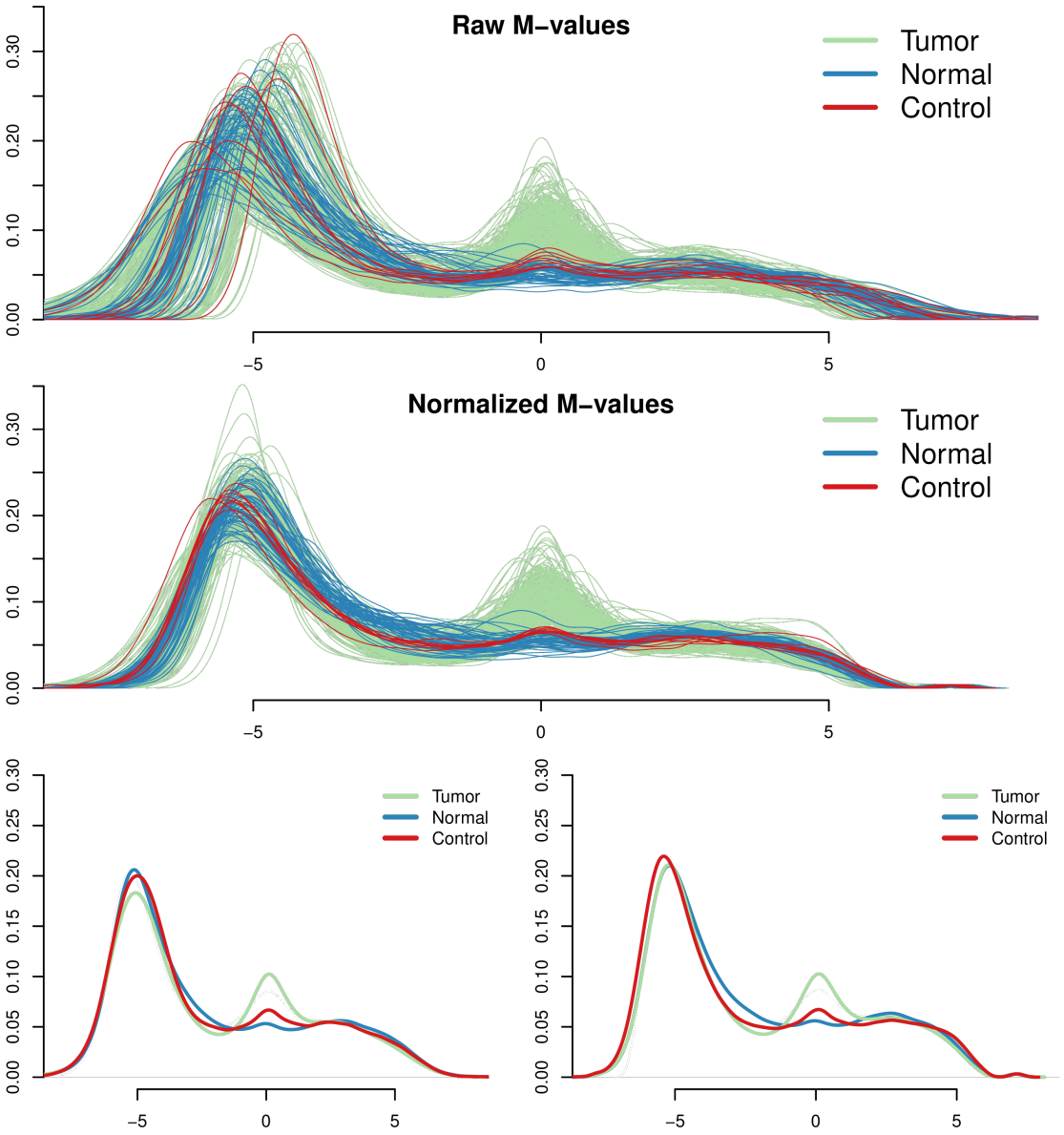


Figure 5.3: Visualization of cancer/normal differences in the TCGA dataset, before and after normalization. In the first two plots are shown the densities of the M-values for Type I green probes before (a) and after (b) functional normalization as presented in the shinyMethyl interactive interface. Green and blue densities represent tumor and normal samples respectively, and red densities represent 9 technical replicates of a control cell line. The last two plots show the average density for each sample group before and after normalization. Functional normalization preserves the expected marginal differences between normal and cancer, while reducing the variation between the technical controls (red lines).

5.2.2 Sex prediction

The sex of the samples can be accurately predicted by using the intensities of the probes mapping to the sex chromosomes in the M and U channels.[13] *shinyMethyl* implements this prediction algorithm and allows the user to interactively specify a cutoff to cluster samples by sex. The array identifiers of the samples for which the predicted sex does not agree with the user-provided sex phenotype are displayed within the interface and can be saved into a csv file for further analysis. From the HNSCC TCGA dataset (described in Example data), one sample shows discrepancy, indicating possible mislabeling (Figure 5.4).

shinyMethyl also performs a principal component analysis (PCA) on the 20,000 most variable autosomal probes. This analysis enables the observation of associations between phenotype and methylation levels. An additional panel displays the physical arrays colored by phenotype. This coloring allows the user to discern potential confounding between phenotype and study design.

5.2.3 Example data

The data package *shinyMethylData* contains the summarized data for 369 HNSCC cancer samples from TCGA. It is available from the Bioconductor project (http://www.bioconductor
All analyses were performed on raw IDAT intensity files available from Level I data in the TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga). Both raw intensities

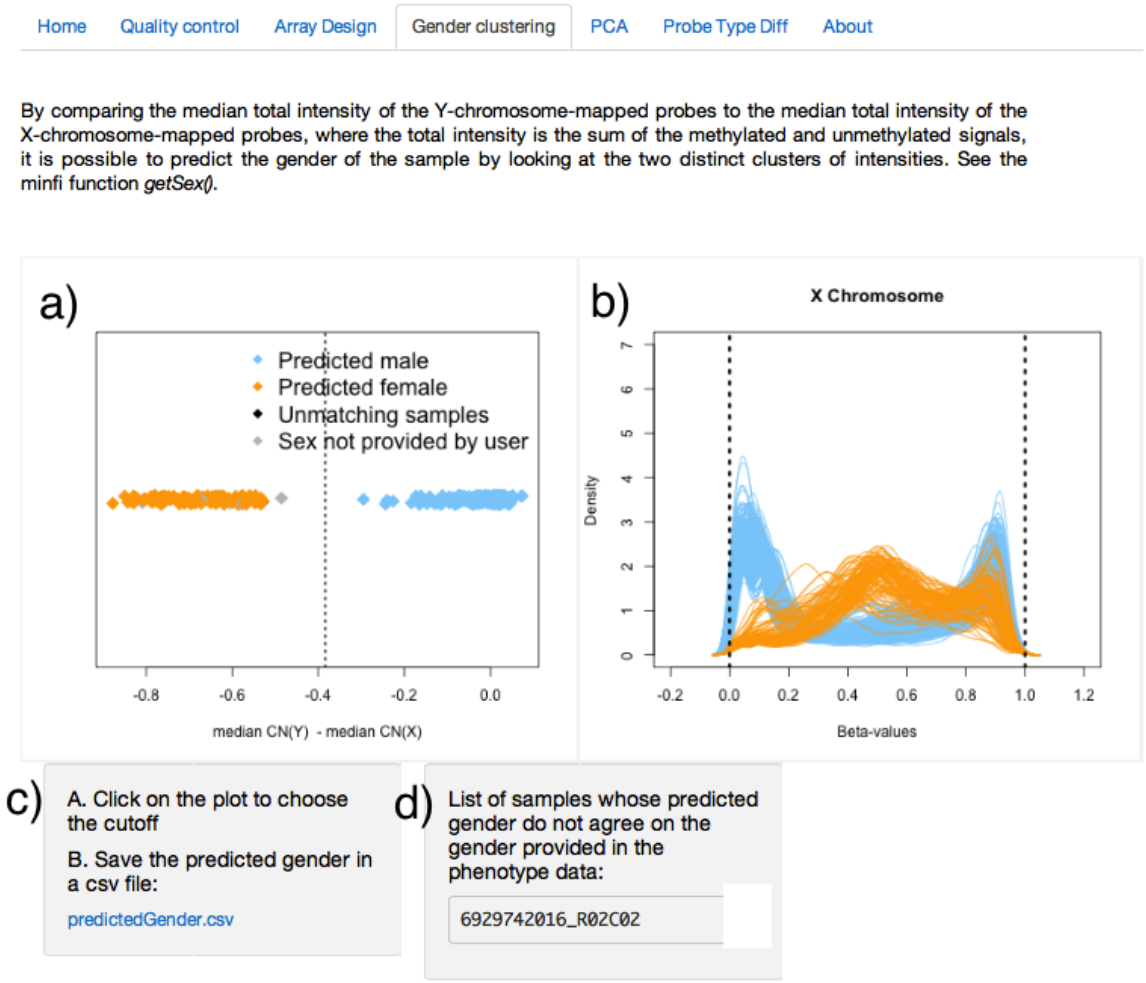


Figure 5.4: Sex prediction interface. The difference of the median copy number intensity for the Y chromosome and the median copy number intensity for the X chromosome can be used to separate males and females. In a), the user can select the vertical cutoff (dashed line) manually with the mouse to separate the two clusters (orange for females, blue for males). Corresponding Beta-value densities appear in b) for further validation. The predicted sex can be downloaded in a csv file in c), and samples for which the predicted sex differs from the sex provided in the phenotype will appear in d).

and normalized methylation values obtained by functional normalization using control probes and a slide covariate [1] are included. The `shinyMethylSet` objects containing respectively the raw and normalized data can be accessed by `summary.tcga.raw` and `summmary.tcga.norm`.

5.3 Discussion

shinyMethyl makes the quality control and pre-processing of 450k methylation array data fast and intuitive through an interactive application in R. We also show, by example, how to use *shiny* to develop interactive visualization interfaces. Our example will facilitate future developments of interactive visualization tools for the processing of high-dimensional genomic data in subsequent Bioconductor[36] packages.

5.3.1 Software Availability

shinyMethyl is an R package available from the Bioconductor project (http://www.bioconductor.org)

Latest source code: https://github.com/jfortin1/shinyMethyl

Source code at the time of publication:https://github.com/F1000Research/shinyMethyl/releases

Archived source code as at the time of publication: http://dx.doi.org/10.5281/zenodo.10748

Bibliography

[1] J.-P. Fortin, A. Labbe, M. Lemire, B. Zanke, T. Hudson, E. Fertig, C. Greenwood, and K. D. Hansen, "Functional normalization of 450k methylation array data improves replication in large cancer studies," *Genome Biology*, vol. 15, no. 11, p. 503, 2014.

[2] M. Bibikova, B. Barnes, C. Tsan, V. Ho, B. Klotzle, J. M. Le, D. Delano, L. Zhang, G. P. Schroth, K. L. Gunderson, J.-B. Fan, and R. Shen, "High density dna methylation array with single cpg site resolution," *Genomics*, vol. 98, no. 4, pp. 288–95, 2011.

[3] V. K. Rakyan, T. A. Down, D. J. Balding, and S. Beck, "Epigenome-wide association studies for common human diseases." *Nature Reviews Genetics*, vol. 12, no. 8, pp. 529–541, 2011.

[4] Y. Liu, M. J. Aryee, L. Padyukov, M. D. Fallin, E. Hesselberg, A. Runarsson, L. Reinius, N. Acevedo, M. Taub, M. Ronninger, K. Shchetynsky, A. Scheynius, J. Kere, L. Alfredsson, L. Klareskog, T. J. Ekström, and A. P. Feinberg,

BIBLIOGRAPHY

"Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis." *Nature Biotechnology*, vol. 31, no. 2, pp. 142–147, 2013.

[5] A. P. Feinberg and B. Vogelstein, "Hypomethylation distinguishes genes of some human cancers from their normal counterparts." *Nature*, vol. 301, no. 5895, pp. 89–92, 1983.

[6] M. A. Gama-Sosa, V. A. Slagel, R. W. Trewyn, R. Oxenhandler, K. C. Kuo, C. W. Gehrke, and M. Ehrlich, "The 5-methylcytosine content of DNA from human tumors." *Nucleic Acids Research*, vol. 11, no. 19, pp. 6883–6894, 1983.

[7] S. E. Goelz, B. Vogelstein, S. R. Hamilton, and A. P. Feinberg, "Hypomethylation of DNA from benign and malignant human colon neoplasms." *Science*, vol. 228, no. 4696, pp. 187–190, 1985.

[8] A. P. Feinberg and B. Tycko, "The history of cancer epigenetics," *Nature Reviews Cancer*, vol. 4, no. 2, pp. 143–153, 2004.

[9] P. A. Jones and S. B. Baylin, "The Epigenomics of Cancer," *Cell*, vol. 128, no. 4, pp. 683–692, 2007.

[10] K. D. Hansen, W. Timp, H. C. Bravo, S. Sabunciyan, B. Langmead, O. G. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep, E. Briem, K. Zhang, R. A. Irizarry,

and A. P. Feinberg, "Increased methylation variation in epigenetic domains across cancer types." *Nature Genetics*, vol. 43, no. 8, pp. 768–775, 2011.

[11] B. P. Berman, D. J. Weisenberger, J. F. Aman, T. Hinoue, Z. Ramjan, Y. Liu, H. Noushmehr, C. P. E. Lange, C. M. van Dijk, R. A. E. M. Tollenaar, D. Van Den Berg, and P. W. Laird, "Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina-associated domains." *Nature Genetics*, vol. 44, no. 1, pp. 40–46, 2012.

[12] N. Touleimat and J. Tost, "Complete pipeline for Infinium Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation." *Epigenomics*, vol. 4, no. 3, pp. 325–341, 2012.

[13] M. J. Aryee, A. E. Jaffe, H. Corrada-Bravo, C. Ladd-Acosta, A. P. Feinberg, K. D. Hansen, and R. A. Irizarry, "Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA Methylation microarrays." *Bioinformatics*, 2014, in press.

[14] J. Maksimovic, L. Gordon, and A. Oshlack, "SWAN: Subset quantile Within-Array Normalization for Illumina Infinium HumanMethylation450 BeadChips." *Genome Biology*, vol. 13, no. 6, p. R44, 2012.

[15] A. E. Teschendorff, F. Marabita, M. Lechner, T. Bartlett, J. Tegner, D. Gomez-Cabrero, and S. Beck, "A beta-mixture quantile normalization method for cor-

recting probe design bias in Illumina Infinium 450 k DNA methylation data," *Bioinformatics*, vol. 29, no. 2, pp. 189–196, 2013.

[16] R. Pidsley, C. C. Y. Wong, M. Volta, K. Lunnon, J. Mill, and L. C. Schalkwyk, "A data-driven approach to preprocessing Illumina 450K methylation array data," *BMC Genomics*, vol. 14, no. 1, p. 293, 2013.

[17] T. J. Triche, D. J. Weisenberger, D. Van Den Berg, P. W. Laird, and K. D. Siegmund, "Low-level processing of Illumina Infinium DNA Methylation BeadArrays." *Nucleic Acids Research*, vol. 41, no. 7, p. e90, 2013.

[18] S. Dedeurwaerder, M. Defrance, M. Bizet, E. Calonne, G. Bontempi, and F. Fuks, "A comprehensive overview of Infinium HumanMethylation450 data processing," *Briefings in Bioinformatics*, 2013.

[19] R. A. Irizarry, C. Ladd-Acosta, B. Carvalho, H. Wu, S. A. Brandenburg, J. A. Jeddeloh, B. Wen, and A. P. Feinberg, "Comprehensive high-throughput arrays for relative methylation (CHARM)." *Genome Research*, vol. 18, no. 5, pp. 780–790, 2008.

[20] M. J. Aryee, Z. Wu, C. Ladd-Acosta, B. Herb, A. P. Feinberg, S. Yegnasubramanian, and R. A. Irizarry, "Accurate genome-scale percentage DNA methylation estimates from microarray data." *Biostatistics*, vol. 12, no. 2, pp. 197–210, 2011.

[21] Z. Wu and M. J. Aryee, "Subset quantile normalization using negative control

features." *Journal of Computational Biology*, vol. 17, no. 10, pp. 1385–1395, 2010.

[22] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data." *Nature Reviews Genetics*, vol. 11, no. 10, pp. 733–739, 2010.

[23] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–64, Apr 2003.

[24] W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron, "Variance stabilization applied to microarray data calibration and to the quantification of differential expression," *Bioinformatics*, vol. 18 Suppl 1, pp. S96–104, 2002.

[25] B. H. Mecham, P. S. Nelson, and J. D. Storey, "Supervised normalization of microarrays," *Bioinformatics*, vol. 26, no. 10, pp. 1308–1315, 2010.

[26] J. T. Leek and J. D. Storey, "Capturing heterogeneity in gene expression studies by surrogate variable analysis," *PLoS Genetics*, vol. 3, no. 9, pp. 1724–1735, 2007.

[27] ——, "A general framework for multiple testing dependence," *Proceedings of the National Academy of Sciences*, vol. 105, no. 48, pp. 18 718–18 723, 2008.

[28] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods." *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[29] J. A. Gagnon-Bartsch and T. P. Speed, "Using control genes to correct for unwanted variation in microarray data," *Biostatistics*, vol. 13, no. 3, pp. 539–552, 2012.

[30] K. N. Harper, B. A. Peters, and M. V. Gamble, "Batch effects and pathway analysis: two potential perils in cancer studies involving dna methylation array analysis," *Cancer Epidemiol Biomarkers Prev*, vol. 22, no. 6, pp. 1052–60, 2013.

[31] E. A. Houseman, W. P. Accomando, D. C. Koestler, B. C. Christensen, C. J. Marsit, H. H. Nelson, J. K. Wiencke, and K. T. Kelsey, "DNA methylation arrays as surrogate measures of cell mixture distribution." *BMC Bioinformatics*, vol. 13, no. 1, p. 86, 2012.

[32] C. M. Montaño, R. A. Irizarry, W. E. Kaufmann, K. Talbot, R. E. Gur, A. P. Feinberg, and M. A. Taub, "Measuring cell-type specific differential methylation in human brain tissue," *Genome Biology*, vol. 14, no. 8, p. R94, 2013.

[33] J. Guintivano, M. J. Aryee, and Z. A. Kaminsky, "A cell epigenotype specific

model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression." *Epigenetics*, vol. 8, no. 3, pp. 290–302, 2013.

[34] A. E. Jaffe and R. A. Irizarry, "Accounting for cellular heterogeneity is critical in epigenome-wide association studies," *Genome Biology*, vol. 15, no. 2, p. R31, 2014.

[35] E. A. Houseman, J. Molitor, and C. J. Marsit, "Reference-free cell mixture adjustments in analysis of DNA methylation data." *Bioinformatics*, vol. 30, no. 10, pp. 1431–1439, 2014.

[36] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang, "Bioconductor: open software development for computational biology and bioinformatics." *Genome Biology*, vol. 5, no. 10, p. R80, 2004.

[37] A. E. Jaffe, P. Murakami, H. Lee, J. T. Leek, M. D. Fallin, A. P. Feinberg, and R. A. Irizarry, "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies." *International Journal of Epidemiology*, vol. 41, no. 1, pp. 200–209, 2012.

[38] T. Sofer, E. D. Schifano, J. A. Hoppin, L. Hou, and A. A. Baccarelli, "A-

clustering: a novel method for the detection of co-regulated methylation regions, and regions associated with exposure." *Bioinformatics*, vol. 29, no. 22, pp. 2884–2891, 2013.

[39] K. D. Hansen, S. Sabunciyan, B. Langmead, N. Nagy, R. Curley, G. Klein, E. Klein, D. Salamon, and A. P. Feinberg, "Large-scale hypomethylated blocks associated with Epstein-Barr virus-induced B-cell immortalization." *Genome Research*, vol. 24, no. 2, pp. 177–184, 2014.

[40] H. S. Parker and J. T. Leek, "The practical effect of batch on genomic prediction," *Statistical Applications in Genetics and Molecular Biology*, vol. 11, no. 3, p. Article 10, 2012.

[41] J. M. Bennett, D. Catovsky, M. T. Daniel, G. Flandrin, D. A. Galton, H. R. Gralnick, and C. Sultan, "Proposals for the classification of the acute leukaemias. French-American-British (FAB) co-operative group," *British Journal of Haematology*, vol. 33, no. 4, pp. 451–8, 1976.

[42] M. E. Figueroa, S. Lugthart, Y. Li, C. Erpelinck-Verschueren, X. Deng, P. J. Christos, E. Schifano, J. Booth, W. van Putten, L. Skrabanek, F. Campagne, M. Mazumdar, J. M. Greally, P. J. M. Valk, B. Löwenberg, R. Delwel, and A. Melnick, "DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia," *Cancer Cell*, vol. 17, no. 1, pp. 13–27, 2010.

[43] A. Akalin, F. E. Garrett-Bakelman, M. Kormaksson, J. Busuttil, L. Zhang,

I. Khrebtukova, T. A. Milne, Y. Huang, D. Biswas, J. L. Hess, C. D. Allis, R. G. Roeder, P. J. M. Valk, B. Löwenberg, R. Delwel, H. F. Fernandez, E. Paietta, M. S. Tallman, G. P. Schroth, C. E. Mason, A. Melnick, and M. E. Figueroa, "Base-pair resolution DNA methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia," *PLoS Genetics*, vol. 8, no. 6, p. e1002781, 2012.

[44] Y.-a. Chen, M. Lemire, S. Choufani, D. T. Butcher, D. Grafodatskaya, B. W. Zanke, S. Gallinger, T. J. Hudson, and R. Weksberg, "Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray," *Epigenetics*, vol. 8, no. 2, pp. 203–209, 2013.

[45] L. Carrel and H. F. Willard, "X-inactivation profile reveals extensive variability in x-linked gene expression in females," *Nature*, vol. 434, no. 7031, pp. 400–4, Mar 2005.

[46] J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, and J. D. Storey, "The sva package for removing batch effects and other unwanted variation in high-throughput experiments." *Bioinformatics*, vol. 28, no. 6, pp. 882–883, 2012.

[47] M. C. Wu, B. R. Joubert, P.-F. Kuan, S. E. Håberg, W. Nystad, S. D. Peddada, and S. J. London, "A systematic assessment of normalization approaches for the infinium 450k methylation platform," *Epigenetics*, vol. 9, no. 2, 2013.

BIBLIOGRAPHY

[48] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Springer, 2009.

[49] R. A. Irizarry, B. M. Bolstad, F. Collin, L. M. Cope, B. Hobbs, and T. P. Speed, "Summaries of affymetrix genechip probe level data," *Nucleic Acids Research*, vol. 31, no. 4, p. e15, 2003.

[50] M. L. Smith, K. A. Baggerly, H. Bengtsson, M. E. Ritchie, and K. D. Hansen, "illuminaio: An open source IDAT parsing tool for Illumina microarrays," *F1000Research*, vol. 2, no. 264, 2013.

[51] P. T. Reiss, L. Huang, and M. Mennes, "Fast function-on-scalar regression with penalized basis expansions," *Int J Biostat*, vol. 6, no. 1, p. Article 28, 2010.

[52] C. M. Crainiceanu, P. T. Reiss, J. Goldsmith, L. Huang, H. Lan, and F. Scheipl, "refund: Regression with functional data," 2013.

[53] M. Cotterchio, G. McKeown-Eyssen, H. Sutherland, G. Buchan, M. Aronson, A. M. Easson, J. Macey, E. Holowaty, and S. Gallinger, "Ontario familial colon cancer registry: methods and first-year response rates." *Chronic diseases in Canada*, vol. 21, no. 2, pp. 81–86, 2000.

[54] B. W. Zanke, C. M. T. Greenwood, J. Rangrej, R. Kustra, A. Tenesa, S. M. Farrington, J. Prendergast, S. Olschwang, T. Chiang, E. Crowdy, V. Ferretti, P. Laflamme, S. Sundararajan, S. Roumy, J.-F. Olivier, F. Robidoux, R. Sladek,

A. Montpetit, P. Campbell, S. Bezieau, A. M. O'Shea, G. Zogopoulos, M. Cotterchio, P. Newcomb, J. McLaughlin, B. Younghusband, R. Green, J. Green, M. E. M. Porteous, H. Campbell, H. Blanche, M. Sahbatou, E. Tubacher, C. Bonaiti-Pellié, B. Buecher, E. Riboli, S. Kury, S. J. Chanock, J. Potter, G. Thomas, S. Gallinger, T. J. Hudson, and M. G. Dunlop, "Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24." *Nature Genetics*, vol. 39, no. 8, pp. 989–994, 2007.

[55] J.-P. Fortin, E. Fertig, and K. Hansen, "shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R," *F1000Research*, vol. 3, no. 175, 2014.

[56] B. R. Joubert, S. E. Håberg, R. M. Nilsen, X. Wang, S. E. Vollset, S. K. Murphy, Z. Huang, C. Hoyo, Ø. Midttun, L. A. Cupul-Uicab, P. M. Ueland, M. C. Wu, W. Nystad, D. A. Bell, S. D. Peddada, and S. J. London, "450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy." *Environmental health perspectives*, vol. 120, no. 10, pp. 1425–1431, 2012.

[57] R. A. Irizarry, D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. N. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martínez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang,

S. Q. Ye, and W. Yu, "Multiple-laboratory comparison of microarray platforms," *Nature Methods*, vol. 2, no. 5, pp. 345–50, 2005.

[58] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome." *Science*, vol. 326, no. 5950, pp. 289–293, 2009.

[59] J. Dekker, M. A. Marti-Renom, and L. A. Mirny, "Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data," *Nature Reviews Genetics*, vol. 14, no. 6, pp. 390–403, 2013.

[60] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions." *Nature*, vol. 485, no. 7398, pp. 376–380, 2012.

[61] F. Jin, Y. Li, J. R. Dixon, S. Selvaraj, Z. Ye, A. Y. Lee, C.-A. Yen, A. D. Schmitt, C. A. Espinoza, and B. Ren, "A high-resolution map of the three-dimensional chromatin interactome in human cells." *Nature*, vol. 503, no. 7475, pp. 290–294, 2013.

[62] N. Naumova, M. Imakaev, G. Fudenberg, Y. Zhan, B. R. Lajoie, L. A. Mirny,

and J. Dekker, "Organization of the mitotic chromosome." *Science*, vol. 342, no. 6161, pp. 948–953, 2013.

[63] B. D. Pope, T. Ryba, V. Dileep, F. Yue, W. Wu, O. Denas, D. L. Vera, Y. Wang, R. S. Hansen, T. K. Canfield, R. E. Thurman, Y. Cheng, G. Gulsoy, J. H. Dennis, M. P. Snyder, J. A. Stamatoyannopoulos, J. Taylor, R. C. Hardison, T. Kahveci, B. Ren, and D. M. Gilbert, "Topologically associating domains are stable units of replication-timing regulation," *Nature*, vol. 515, no. 7527, pp. 402–405, 2014.

[64] S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden, "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping." *Cell*, vol. 159, no. 7, pp. 1665–1680, 2014.

[65] J. R. Dixon, I. Jung, S. Selvaraj, Y. Shen, J. E. Antosiewicz-Bourget, A. Y. Lee, Z. Ye, A. Kim, N. Rajagopal, W. Xie, Y. Diao, J. Liang, H. Zhao, V. V. Lobanenkov, J. R. Ecker, J. A. Thomson, and B. Ren, "Chromatin architecture reorganization during stem cell differentiation." *Nature*, vol. 518, no. 7539, pp. 331–336, 2015.

[66] G. E. Crawford, I. E. Holt, J. Whittle, B. D. Webb, D. Tai, S. Davis, E. H. Margulies, Y. Chen, J. A. Bernat, D. Ginsburg, D. Zhou, S. Luo, T. J. Vasicek, M. J. Daly, T. G. Wolfsberg, and F. S. Collins, "Genome-wide map-

ping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS)." *Genome Research*, vol. 16, no. 1, pp. 123–131, 2006.

[67] A. P. Boyle, S. Davis, H. P. Shulha, P. Meltzer, E. H. Margulies, Z. Weng, T. S. Furey, and G. E. Crawford, "High-resolution mapping and characterization of open chromatin across the genome." *Cell*, vol. 132, no. 2, pp. 311–322, 2008.

[68] S. A. Smallwood, H. J. Lee, C. Angermueller, F. Krueger, H. Saadeh, J. Peat, S. R. Andrews, O. Stegle, W. Reik, and G. Kelsey, "Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity," *Nature Methods*, vol. 11, no. 8, pp. 817–20, 2014.

[69] D. A. Cusanovich, R. Daza, A. Adey, H. A. Pliner, L. Christiansen, K. L. Gunderson, F. J. Steemers, C. Trapnell, and J. Shendure, "Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing," *Science*, vol. 348, no. 6237, pp. 910–914, 2015.

[70] TCGA, "The cancer genome atlas." [Online]. Available: http://cancergenome.nih.gov

[71] M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny, "Iterative correction of Hi-C data reveals hallmarks of chromosome organization." *Nature Methods*, vol. 9, no. 10, pp. 999–1003, 2012.

[72] D. Mouchiroud, G. D'Onofrio, B. Aïssani, G. Macaya, C. Gautier, and G. Bernardi, "The distribution of genes in the human genome." *Gene*, vol. 100, pp. 181–187, 1991.

[73] H. Heyn, S. Moran, I. Hernando-Herraez, S. Sayols, A. Gomez, J. Sandoval, D. Monk, K. Hata, T. Marques-Bonet, L. Wang, and M. Esteller, "DNA methylation contributes to natural human variation." *Genome Research*, vol. 23, no. 9, pp. 1363–1372, 2013.

[74] A. M. Deaton and A. Bird, "CpG islands and the regulation of transcription." *Genes & Development*, vol. 25, no. 10, pp. 1010–1022, 2011.

[75] J. R. Wagner, S. Busche, B. Ge, T. Kwan, T. Pastinen, and M. Blanchette, "The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts," *Genome Biology*, vol. 15, no. 2, p. R37, 2014.

[76] R. Lister, M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. C. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q.-M. Ngo, L. Edsall, J. Antosiewicz-Bourget, R. Stewart, V. Ruotti, A. H. Millar, J. A. Thomson, B. Ren, and J. R. Ecker, "Human DNA methylomes at base resolution show widespread epigenomic differences," *Nature*, vol. 462, no. 7271, pp. 315–322, 2009.

[77] Y. Liu, X. Li, M. J. Aryee, T. J. Ekström, L. Padyukov, L. Klareskog, A. Vandiver, A. Z. Moore, T. Tanaka, L. Ferrucci, M. D. Fallin, and A. P. Feinberg,

"GeMes, clusters of DNA methylation under genetic control, can inform genetic and epigenetic analysis of disease." *American Journal of Human Genetics*, vol. 94, no. 4, pp. 485–495, 2014.

[78] K. D. Makova and R. C. Hardison, "The effects of chromatin organization on variation in mutation rates in the genome," *Nature Reviews Genetics*, vol. 16, no. 4, pp. 213–223, 2015.

[79] B. Schuster-Böckler and B. Lehner, "Chromatin organization is a major influence on regional mutation rates in human cancer cells," *Nature*, vol. 488, no. 7412, pp. 504–507, 2012.

[80] P. Polak, R. Karlić, A. Koren, R. Thurman, R. Sandstrom, M. S. Lawrence, A. Reynolds, E. Rynes, K. Vlahoviček, J. A. Stamatoyannopoulos, and S. R. Sunyaev, "Cell-of-origin chromatin organization shapes the mutational landscape of cancer," *Nature*, vol. 518, no. 7539, pp. 360–364, 2015.

[81] J. F. Degner, A. A. Pai, R. Pique-Regi, J.-B. Veyrieras, D. J. Gaffney, J. K. Pickrell, S. De Leon, K. Michelini, N. Lewellen, G. E. Crawford, M. Stephens, Y. Gilad, and J. K. Pritchard, "Dnase sensitivity qtls are a major determinant of human expression variation," *Nature*, vol. 482, no. 7385, pp. 390–4, 2012.

[82] B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, T. S. Mikkelsen, and J. A. Thomson, "The

nih roadmap epigenomics mapping consortium," *Nat Biotechnol*, vol. 28, no. 10, pp. 1045–8, Oct 2010.

[83] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." *Nature Methods*, vol. 10, no. 12, pp. 1213–1218, 2013.

[84] J. D. Buenrostro, B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf, "Single-cell chromatin accessibility reveals principles of regulatory variation." *Nature*, vol. 523, no. 7561, pp. 486–490, 2015.

[85] J. T. Leek, R. B. Scharpf, H. C. Bravo, D. Simcha, B. Langmead, W. E. Johnson, D. Geman, K. Baggerly, and R. A. Irizarry, "Tackling the widespread and critical impact of batch effects in high-throughput data." *Nature Reviews Genetics*, vol. 11, no. 10, pp. 733–739, 2010.

[86] T. Ryba, I. Hiratani, J. Lu, M. Itoh, M. Kulik, J. Zhang, T. C. Schulz, A. J. Robins, S. Dalton, and D. M. Gilbert, "Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types," *Genome Research*, vol. 20, no. 6, pp. 761–770, 2010.

[87] H. B. Fraser, L. L. Lam, S. M. Neumann, and M. S. Kobor, "Population-

specificity of human dna methylation," *Genome Biology*, vol. 13, no. 2, p. R8, 2012.

[88] J. T. Bell, A. A. Pai, J. K. Pickrell, D. J. Gaffney, R. Pique-Regi, J. F. Degner, Y. Gilad, and J. K. Pritchard, "Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines," *Genome Biology*, vol. 12, no. 1, p. R10, 2011.

[89] "The tcga data portal." [Online]. Available: https://tcga-data.nci.nih.gov/tcga/

[90] R. Lister, M. Pelizzola, Y. S. Kida, R. D. Hawkins, J. R. Nery, G. Hon, J. Antosiewicz-Bourget, R. O'Malley, R. Castanon, S. Klugman, M. Downes, R. Yu, R. Stewart, B. Ren, J. A. Thomson, R. M. Evans, and J. R. Ecker, "Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells," *Nature*, vol. 471, no. 7336, pp. 68–73, 2011.

[91] "Data from salk institute." [Online]. Available: http://neomorph.salk.edu/ips_methylomes/data.htm

[92] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome." *Genome Biology*, vol. 10, no. 3, p. R25, 2009.

[93] R. P. McCord, A. Nazario-Toole, H. Zhang, P. S. Chines, Y. Zhan, M. R. Erdos,

BIBLIOGRAPHY

F. S. Collins, J. Dekker, and K. Cao, "Correlated alterations in genome organization, histone methylation, and DNA-lamin A/C interactions in Hutchinson-Gilford progeria syndrome." *Genome Research*, vol. 23, no. 2, pp. 260–269, 2013.

[94] S. Selvaraj, J. R Dixon, V. Bansal, and B. Ren, "Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing." *Nature Biotechnology*, vol. 31, no. 12, pp. 1111–1118, 2013.

[95] "Data from u. chicago." [Online]. Available: http://eqtl.uchicago.edu/dsQTL_data/RAW_DATA_HDF5

[96] M. Lawrence, R. Gentleman, and V. J. Carey, "rtracklayer: an R package for interfacing with genome browsers," *Bioinformatics*, vol. 25, no. 14, pp. 1841–1842, 2009.

[97] H. Wu, B. Caffo, H. A. Jaffee, R. A. Irizarry, and A. P. Feinberg, "Redefining cpg islands using hidden markov models," *Biostatistics*, vol. 11, no. 3, pp. 499–514, 2010.

[98] S. Dejean, I. Gonzalez, K.-A. L. C. with contributions from Pierre Monget, J. Coquery, F. Yao, B. Liquet, and F. Rohart, *mixOmics: Omics Data Integration Project*, 2014, r package version 5.0-3. [Online]. Available: http://CRAN.R-project.org/package=mixOmics

BIBLIOGRAPHY

[99] "Cg hub website." [Online]. Available: https://cghub.ucsc.edu

[100] R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang, "Bioconductor: open software development for computational biology and bioinformatics." *Genome Biology*, vol. 5, no. 10, p. R80, 2004.

[101] W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Oleś, H. Pagès, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan, "Orchestrating high-throughput genomic analysis with Bioconductor." *Nature Methods*, vol. 12, no. 2, pp. 115–121, 2015.

[102] R. T. Shinohara, E. M. Sweeney, J. Goldsmith, N. Shiee, F. J. Mateen, P. A. Calabresi, S. Jarso, D. L. Pham, D. S. Reich, C. M. Crainiceanu, Australian Imaging Biomarkers Lifestyle Flagship Study of Ageing, and Alzheimer's Disease Neuroimaging Initiative, "Statistical normalization techniques for magnetic resonance imaging," *Neuroimage Clin*, vol. 6, pp. 9–19, 2014.

[103] L. G. Nyúl and J. K. Udupa, "On standardizing the mr image intensity scale," *Magn Reson Med*, vol. 42, no. 6, pp. 1072–81, Dec 1999.

[104] L. G. Nyúl, J. K. Udupa, and X. Zhang, "New variants of a method of mri scale standardization," *IEEE Trans Med Imaging*, vol. 19, no. 2, pp. 143–50, Feb 2000.

[105] N. L. Weisenfeld and S. K. Warfield, "Normalization of joint image-intensity statistics in mri using the kullbackleibler divergence," *Biomedical Imaging: Nano to Macro, 2004 IEEE International Symposium on (101104IEEE)*, 2004.

[106] F. Jager, Y. Deuerling-Zheng, B. Frericks, F. Wacker, and H. Hornegger, "A new method for mri intensity standardization with application to lesion detection in the brain," *Vision Modeling and Visualization*, pp. 269–276, 2006.

[107] A. Madabhushi, J. K. Udupa, and G. Moonis, "Comparing mr image intensity standardization against tissue characterizability of magnetization transfer ratio imaging," *J Magn Reson Imaging*, vol. 24, no. 3, pp. 667–75, Sep 2006.

[108] K. K. Leung, M. J. Clarkson, J. W. Bartlett, S. Clegg, C. R. Jack, Jr, M. W. Weiner, N. C. Fox, S. Ourselin, and Alzheimer's Disease Neuroimaging Initiative, "Robust atrophy rate measurement in alzheimer's disease using multi-site serial mri: tissue-specific intensity normalization and parameter selection," *Neuroimage*, vol. 50, no. 2, pp. 516–23, Apr 2010.

[109] R. T. Shinohara, C. M. Crainiceanu, B. S. Caffo, M. I. Gaitán, and D. S. Reich, "Population-wide principal component-based quantification of blood-brain-

barrier dynamics in multiple sclerosis," *Neuroimage*, vol. 57, no. 4, pp. 1430–46, Aug 2011.

[110] M. Shah, Y. Xiao, N. Subbanna, S. Francis, D. L. Arnold, D. L. Collins, and T. Arbel, "Evaluating intensity normalization on mris of human brain with multiple sclerosis," *Med Image Anal*, vol. 15, no. 2, pp. 267–82, Apr 2011.

[111] J. T. Leek, "svaseq: removing batch effects and other unwanted noise from sequencing data," *Nucleic Acids Res*, vol. 42, no. 21, Dec 2014.

[112] K. Luoma, R. Raininko, P. Nummi, and R. Luukkonen, "Is the signal intensity of cerebrospinal fluid constant? intensity measurements with high and low field magnetic resonance imagers," *Magn Reson Imaging*, vol. 11, no. 4, pp. 549–55, 1993.

[113] C. R. Jack, M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L Whitwell, C. Ward *et al.*, "The alzheimer's disease neuroimaging initiative (adni): Mri methods," *Journal of Magnetic Resonance Imaging*, vol. 27, no. 4, pp. 685–691, 2008.

[114] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2014. [Online]. Available: http://www.R-project.org/

[115] B. Whitcher, V. J. Schmid, and A. Thornton, "Working with the DICOM and

NIfTI data standards in R," *Journal of Statistical Software*, vol. 44, no. 6, pp. 1–28, 2011. [Online]. Available: http://www.jstatsoft.org/v44/i06/

[116] J. Muschelli, E. M. Sweeney, M. A. Lindquist, and C. M. Crainiceanu, "fslr: Connecting the fsl software with r," *The R Journal*, vol. 7, no. 1, pp. 163–175, Feb 2015.

[117] B. B. Avants, B. m. Kandel, J. T. Duda, and P. A. Cook, "Antsr: Ants in r," 2015. [Online]. Available: https://github.com/stnava/ANTsR

[118] T. Shinohara and J. Muschelli, "Whitestripe: White matter normalization for magnetic resonance images using whitestripe," 2015. [Online]. Available: https://cran.r-project.org/web/packages/WhiteStripe/index.html

[119] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: improved n3 bias correction," *IEEE Trans Med Imaging*, vol. 29, no. 6, pp. 1310–20, Jun 2010.

[120] K. Oishi, A. Faria, and S. Mori, "Jhu-mni-ss atlas," 05 2010.

[121] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Med Image Anal*, vol. 12, no. 1, pp. 26–41, Feb 2008.

[122] S. M. Smith, "Fast robust automated brain extraction," *Hum Brain Mapp*, vol. 17, no. 3, pp. 143–55, Nov 2002.

[123] Y. Zhang, M. Brady, and S. Smith, "Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm," *IEEE Trans Med Imaging*, vol. 20, no. 1, pp. 45–57, Jan 2001.

[124] J. T. Leek and R. D. Peng, "Opinion: Reproducible research can still be wrong: adopting a prevention approach," *Proc Natl Acad Sci U S A*, vol. 112, no. 6, pp. 1645–6, Feb 2015.

[125] R. A. Irizarry, D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. N. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martínez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S. Q. Ye, and W. Yu, "Multiple-laboratory comparison of microarray platforms," *Nature Methods*, vol. 2, no. 5, pp. 345–50, 2005.

[126] R. W. Bourgon, "Chromatin immunoprecipitation and high-density tiling microarrays: a generative model, methods for analysis, and methodology assessment in the absence of a gold standard," Ph.D. dissertation, University of California, Berkeley, 2006.

[127] R. Schmid, P. Baum, C. Ittrich, K. Fundel-Clemens, W. Huber, B. Brors, R. Eils, A. Weith, D. Mennerich, and K. Quast, "Comparison of normaliza-

tion methods for illumina beadchip humanht-12 v3," *BMC Genomics*, vol. 11, p. 349, 2010.

[128] N. C. Fox, E. K. Warrington, P. A. Freeborough, P. Hartikainen, A. M. Kennedy, J. M. Stevens, and M. N. Rossor, "Presymptomatic hippocampal atrophy in alzheimer's disease. a longitudinal mri study," *Brain*, vol. 119 (Pt 6), pp. 2001–7, Dec 1996.

[129] E. Mori, Y. Yoneda, H. Yamashita, N. Hirono, M. Ikeda, and A. Yamadori, "Medial temporal structures relate to memory impairment in alzheimer's disease: an mri volumetric study," *J Neurol Neurosurg Psychiatry*, vol. 63, no. 2, pp. 214–21, Aug 1997.

[130] C. R. Jack, Jr, R. C. Petersen, Y. C. Xu, P. C. O'Brien, G. E. Smith, R. J. Ivnik, B. F. Boeve, S. C. Waring, E. G. Tangalos, and E. Kokmen, "Prediction of ad with mri-based hippocampal volume in mild cognitive impairment," *Neurology*, vol. 52, no. 7, pp. 1397–403, Apr 1999.

[131] P. J. Visser, P. Scheltens, F. R. Verhey, B. Schmand, L. J. Launer, J. Jolles, and C. Jonker, "Medial temporal lobe atrophy and memory dysfunction as predictors for dementia in subjects with mild cognitive impairment," *J Neurol*, vol. 246, no. 6, pp. 477–85, Jun 1999.

[132] C. R. Jack, Jr, R. C. Petersen, Y. Xu, P. C. O'Brien, G. E. Smith, R. J. Ivnik, B. F. Boeve, E. G. Tangalos, and E. Kokmen, "Rates of hippocampal atrophy

correlate with change in clinical status in aging and ad," *Neurology*, vol. 55, no. 4, pp. 484–89, Aug 2000.

[133] Y. Xu, C. R. Jack, Jr, P. C. O'Brien, E. Kokmen, G. E. Smith, R. J. Ivnik, B. F. Boeve, R. G. Tangalos, and R. C. Petersen, "Usefulness of mri measures of entorhinal cortex versus hippocampus in ad," *Neurology*, vol. 54, no. 9, pp. 1760–7, May 2000.

[134] D. J. Callen, S. E. Black, F. Gao, C. B. Caldwell, and J. P. Szalai, "Beyond the hippocampus: Mri volumetry confirms widespread limbic atrophy in ad," *Neurology*, vol. 57, no. 9, pp. 1669–74, Nov 2001.

[135] A. T. Du, N. Schuff, D. Amend, M. P. Laakso, Y. Y. Hsu, W. J. Jagust, K. Yaffe, J. H. Kramer, B. Reed, D. Norman, H. C. Chui, and M. W. Weiner, "Magnetic resonance imaging of the entorhinal cortex and hippocampus in mild cognitive impairment and alzheimer's disease," *J Neurol Neurosurg Psychiatry*, vol. 71, no. 4, pp. 441–7, Oct 2001.

[136] C. M. C. Bottino, C. C. Castro, R. L. E. Gomes, C. A. Buchpiguel, R. L. Marchetti, and M. R. L. Neto, "Volumetric mri measurements can differentiate alzheimer's disease, mild cognitive impairment, and normal aging," *Int Psychogeriatr*, vol. 14, no. 1, pp. 59–72, Mar 2002.

[137] G. Chételat, B. Desgranges, V. De La Sayette, F. Viader, F. Eustache, and

J.-C. Baron, "Mapping gray matter loss with voxel-based morphometry in mild cognitive impairment," *Neuroreport*, vol. 13, no. 15, pp. 1939–43, Oct 2002.

[138] C. Pennanen, M. Kivipelto, S. Tuomainen, P. Hartikainen, T. Hänninen, M. P. Laakso, M. Hallikainen, M. Vanhanen, A. Nissinen, E.-L. Helkala, P. Vainio, R. Vanninen, K. Partanen, and H. Soininen, "Hippocampus and entorhinal cortex in mild cognitive impairment and early ad," *Neurobiol Aging*, vol. 25, no. 3, pp. 303–10, Mar 2004.

[139] H. Wolf, A. Hensel, F. Kruggel, S. G. Riedel-Heller, T. Arendt, L.-O. Wahlund, and H.-J. Gertz, "Structural correlates of mild cognitive impairment," *Neurobiol Aging*, vol. 25, no. 7, pp. 913–24, Aug 2004.

[140] G. Chételat, B. Landeau, F. Eustache, F. Mézenge, F. Viader, V. de la Sayette, B. Desgranges, and J.-C. Baron, "Using voxel-based morphometry to map the structural changes associated with rapid conversion in mci: a longitudinal mri study," *Neuroimage*, vol. 27, no. 4, pp. 934–46, Oct 2005.

[141] B. H. Ridha, J. Barnes, J. W. Bartlett, A. Godbolt, T. Pepple, M. N. Rossor, and N. C. Fox, "Tracking atrophy progression in familial alzheimer's disease: a serial mri study," *Lancet Neurol*, vol. 5, no. 10, pp. 828–34, Oct 2006.

[142] T. F. D. Farrow, S. N. Thiyagesh, I. D. Wilkinson, R. W. Parks, L. Ingram, and P. W. R. Woodruff, "Fronto-temporal-lobe atrophy in early-stage alzheimer's

disease identified using an improved detection methodology," *Psychiatry Res*, vol. 155, no. 1, pp. 11–9, May 2007.

[143] J. L. Whitwell, S. A. Przybelski, S. D. Weigand, D. S. Knopman, B. F. Boeve, R. C. Petersen, and C. R. Jack, Jr, "3d maps from multiple mri illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to alzheimer's disease," *Brain*, vol. 130, no. Pt 7, pp. 1777–86, Jul 2007.

[144] S. P. Poulin, R. Dautoff, J. C. Morris, L. F. Barrett, B. C. Dickerson, and Alzheimer's Disease Neuroimaging Initiative, "Amygdala atrophy is prominent in early alzheimer's disease and relates to symptom severity," *Psychiatry Res*, vol. 194, no. 1, pp. 7–13, Oct 2011.

[145] S. A. Scott, S. T. DeKosky, and S. W. Scheff, "Volumetric atrophy of the amygdala in alzheimer's disease: quantitative serial reconstruction," *Neurology*, vol. 41, no. 3, pp. 351–6, Mar 1991.

[146] S. A. Scott, S. T. DeKosky, D. L. Sparks, C. A. Knox, and S. W. Scheff, "Amygdala cell loss and atrophy in alzheimer's disease," *Ann Neurol*, vol. 32, no. 4, pp. 555–63, Oct 1992.

[147] T. H. Vereecken, O. J. Vogels, and R. Nieuwenhuys, "Neuron loss and shrinkage in the amygdala in alzheimer's disease," *Neurobiol Aging*, vol. 15, no. 1, pp. 45–54, 1994.

[148] D. Horínek, P. Petrovický, J. Hort, J. Krásenský, J. Brabec, M. Bojar, M. Vanecková, and Z. Seidl, "Amygdalar volume and psychiatric symptoms in alzheimer's disease: an mri analysis," *Acta Neurol Scand*, vol. 113, no. 1, pp. 40–5, Jan 2006.

[149] M. I. Miller, L. Younes, J. T. Ratnanather, T. Brown, H. Trinh, D. S. Lee, D. Tward, P. B. Mahon, S. Mori, M. Albert, and BIOCARD Research Team, "Amygdalar atrophy in symptomatic alzheimer's disease based on diffeomorphometry: the biocard cohort," *Neurobiol Aging*, vol. 36 Suppl 1, pp. S3–S10, Jan 2015.

[150] U. A. Khan, L. Liu, F. A. Provenzano, D. E. Berman, C. P. Profaci, R. Sloan, R. Mayeux, K. E. Duff, and S. A. Small, "Molecular drivers and cortical spread of lateral entorhinal cortex dysfunction in preclinical alzheimer's disease," *Nat Neurosci*, vol. 17, no. 2, pp. 304–11, Feb 2014.

[151] T. Gómez-Isla, J. L. Price, D. W. McKeel, Jr, J. C. Morris, J. H. Growdon, and B. T. Hyman, "Profound loss of layer ii entorhinal cortex neurons occurs in very mild alzheimer's disease," *J Neurosci*, vol. 16, no. 14, pp. 4491–500, Jul 1996.

[152] H. Braak and K. Del Tredici, "Alzheimer's disease: pathogenesis and prevention," *Alzheimers Dement*, vol. 8, no. 3, pp. 227–33, May 2012.

[153] M. M. Mielke, N. A. Kozauer, K. C. G. Chan, M. George, J. Toroney, M. Zer-

rate, K. Bandeen-Roche, M.-C. Wang, P. Vanzijl, J. J. Pekar, S. Mori, C. G. Lyketsos, and M. Albert, "Regionally-specific diffusion tensor imaging in mild cognitive impairment and alzheimer's disease," *Neuroimage*, vol. 46, no. 1, pp. 47–55, May 2009.

[154] Y. Liu, G. Spulber, K. K. Lehtimäki, M. Könönen, I. Hallikainen, H. Gröhn, M. Kivipelto, M. Hallikainen, R. Vanninen, and H. Soininen, "Diffusion tensor imaging and tract-based spatial statistics in alzheimer's disease and mild cognitive impairment," *Neurobiol Aging*, vol. 32, no. 9, pp. 1558–71, 2011.

[155] S. Dedeurwaerder, M. Defrance, M. Bizet, E. Calonne, G. Bontempi, and F. Fuks, "A comprehensive overview of infinium humanmethylation450 data processing," *Brief Bioinform*, vol. 15, no. 6, pp. 929–41, Nov 2014.

[156] C. Davatzikos, K. Ruparel, Y. Fan, D. G. Shen, M. Acharyya, J. W. Loughead, R. C. Gur, and D. D. Langleben, "Classifying spatial patterns of brain activity with machine learning methods: application to lie detection," *Neuroimage*, vol. 28, no. 3, pp. 663–8, Nov 2005.

[157] F. De Martino, G. Valente, N. Staeren, J. Ashburner, R. Goebel, and E. Formisano, "Combining multivariate voxel selection and support vector machines for mapping and classification of fmri spatial patterns," *Neuroimage*, vol. 43, no. 1, pp. 44–58, Oct 2008.

[158] P. Vemuri, J. L. Gunter, M. L. Senjem, J. L. Whitwell, K. Kantarci, D. S.

Knopman, B. F. Boeve, R. C. Petersen, and C. R. Jack, Jr, "Alzheimer's disease diagnosis in individual subjects using structural mr images: validation studies," *Neuroimage*, vol. 39, no. 3, pp. 1186–97, Feb 2008.

[159] R. C. Craddock, P. E. Holtzheimer, 3rd, X. P. Hu, and H. S. Mayberg, "Disease state prediction from resting state functional connectivity," *Magn Reson Med*, vol. 62, no. 6, pp. 1619–28, Dec 2009.

[160] C. Davatzikos, P. Bhatt, L. M. Shaw, K. N. Batmanghelich, and J. Q. Trojanowski, "Prediction of mci to ad conversion, via mri, csf biomarkers, and pattern classification," *Neurobiol Aging*, vol. 32, no. 12, pp. 2322.e19–27, Dec 2011.

[161] B. Gaonkar and C. Davatzikos, "Analytic estimation of statistical significance maps for support vector machine based multi-variate image analysis and classification," *Neuroimage*, vol. 78, pp. 270–83, Sep 2013.

[162] J. Pujol, C. Junqué, P. Vendrell, J. M. Grau, J. L. Martí-Vilalta, C. Olivé, and J. Gili, "Biological significance of iron-related magnetic resonance imaging changes in the brain," *Archives of neurology*, vol. 49, no. 7, pp. 711–717, 1992.

[163] R. Bakshi, R. H. B. Benedict, R. A. Bermel, S. D. Caruthers, S. R. Puli, C. W. Tjoa, A. J. Fabiano, and L. Jacobs, "T2 hypointensity in the deep gray matter of patients with multiple sclerosis: a quantitative magnetic resonance imaging study," *Arch Neurol*, vol. 59, no. 1, pp. 62–8, Jan 2002.

[164] C. W. Tjoa, R. H. B. Benedict, B. Weinstock-Guttman, A. J. Fabiano, and R. Bakshi, "Mri t2 hypointensity of the dentate nucleus is related to ambulatory impairment in multiple sclerosis," *J Neurol Sci*, vol. 234, no. 1-2, pp. 17–24, Jul 2005.

[165] S. D. Brass, R. H. B. Benedict, B. Weinstock-Guttman, F. Munschauer, and R. Bakshi, "Cognitive impairment is associated with subcortical magnetic resonance imaging grey matter t2 hypointensity in multiple sclerosis," *Mult Scler*, vol. 12, no. 4, pp. 437–44, Aug 2006.

[166] M. Neema, A. Arora, B. C. Healy, Z. D. Guss, S. D. Brass, Y. Duan, G. J. Buckle, B. I. Glanz, L. Stazzone, S. J. Khoury, H. L. Weiner, C. R. G. Guttmann, and R. Bakshi, "Deep gray matter involvement on brain mri scans is associated with clinical progression in multiple sclerosis," *J Neuroimaging*, vol. 19, no. 1, pp. 3–8, Jan 2009.

[167] A. Mejia, E. M. Sweeney, B. Dewey, G. Nair, P. Sati, C. Shea, D. S. Reich, , and R. T. Shinohara, "Statistical estimation of t1 relaxation time using conventional magnetic resonance imaging," *UPenn Biostatistics Working Papers*, vol. Working Paper 37, 2015.

[168] R. Ghassemi, R. Brown, S. Narayanan, B. Banwell, K. Nakamura, and D. L. Arnold, "Normalization of white matter intensity on t1-weighted images of pa-

tients with acquired central nervous system demyelination," *J Neuroimaging*, vol. 25, no. 2, pp. 184–90, 2015.

[169] E. M. Sweeney, R. T. Shinohara, N. Shiee, F. J. Mateen, A. A. Chudgar, J. L. Cuzzocreo, P. A. Calabresi, D. L. Pham, D. S. Reich, and C. M. Crainiceanu, "Oasis is automated statistical inference for segmentation, with applications to multiple sclerosis lesion segmentation in mri," *Neuroimage Clin*, vol. 2, pp. 402–13, 2013.

[170] RStudio and Inc., *shiny: Web Application Framework for R*, 2014, r package version 0.10.0. [Online]. Available: http://CRAN.R-project.org/package=shiny

[171] A. P. Feinberg and B. Vogelstein, "Hypomethylation distinguishes genes of some human cancers from their normal counterparts." *Nature*, vol. 301, no. 5895, pp. 89–92, 1983.

Jean-Philippe Fortin

Biographical Information

Date of birth: January 27, 1989

Place of birth: Quebec City, Canada

Email: fortin946@gmail.com

Website: jfortinbiostats.com

Education

Ph.D. Biostatistics, 2016
Johns Hopkins University, Baltimore, USA
Advisor: Kasper D. Hansen

B.Sc. Honors Mathematics, 2012
McGill University, Montréal, Canada
First class honors and distinction

Diploma of collegial studies, CEGEP Sainte-Foy, 2008

Honors and Awards

June B. Culley Award, 2016
for outstanding achievement on the schoolwide examination paper
Johns Hopkins University

John Van Ryzin Award, 2014
for the most outstanding ENAR Distinguished Student Paper

Poster Award, 1st place, 2014
Hopkins Genetics Research Day, Johns Hopkins University

PhD Comprehensive Examination Award, 2013
for outstanding achievement on the first-year comprehensive examination
Johns Hopkins University

Best Research Trainee Presentation Award, 2012
Canadian Human and Statistical Genetics Meeting

Poster Award, 2nd place, 2012
Annual Undergraduate Research Conference, McGill University

Dean's Honors List, 2011
McGill University

Governor General's Academic Medal, 2008

Scholarships

Travel Scholarship, 2014
Bioconductor Conference 2014, Boston, USA

FQRNT Scholarship (ES M), 2013-2014
Fonds québécois de la recherche sur la nature et les technologies (FQRNT)

NSERC Scholarship (PGS M) , 2012-2013
Natural Sciences and Engineering Research Council of Canada (NSERC)

Travel Scholarship, 2013
Bioconductor Conference 2013, Seattle, USA

Undergraduate Research Scholarship, 2011
Institut des sciences mathématiques (ISM), Montréal

Narcissa Farrand Entrance Scholarship, 2011
McGill University

Publications

Published

Kathleen Oros Klein, Stepan Grinek, Sasha Bernatsky, Luigi Bouchard, Antonio Ciampi, Ines Colmegna, **Jean-Philippe Fortin**, Long Gao, Marie-France Hivert, Marie Hudson, Michael S. Kobor, Aurélie Labbe, Julia L. MacIsaac, Michael J. Meaney, Alexander M. Morin, Kieran J. O'Donnell, Tomi Pastinen, Marinus H. Van Ijzendoorn, Gregory Voisin, Celia M.T. Greenwood (2015). "funtooNorm: an R package for normalization of DNA methylation data when there are multiple cell or tissue types". *Bioinformatics*.

Jean-Philippe Fortin, Kasper D Hansen (2015). "Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data." *Genome Biology*.

Jean-Philippe Fortin, Aurélie Labbe, Mathieu Lemire, Brent W Zanke, Thomas J Hudson, Elana J Fertig, Celia MT Greenwood, Kasper D Hansen (2014)."Functional normalization of 450k methylation array data improves replication in large cancer studies." *Genome Biology*.

Jean-Philippe Fortin, Elana J Fertig, Kasper D Hansen (2014). "shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R." *F1000Research*.

Jean-Philippe Fortin, Samantha Rudinsky (2013). "Asymptotic eigenvalue distribution of random lifts." *Waterloo Mathematics Review*.

Under Revision

Jean-Philippe Fortin, Elizabeth M Sweeney, John Muschelli, Ciprian M Crainiceanu and Russell T Shinohara (2015). "Removing inter-subject technical variability in magnetic resonance imaging studies." Under revision at *NeuroImage*.

Teaching

Instructor, *450k Array Analysis Workshop London*, University College London, UK Tutorial on the minfi software for the analysis of methylation data, May 2014

Instructor, *Bioconductor Conference*, Boston, USA Tutorial on the minfi software for the analysis of methylation data, July 2014

Instructor, *Bioconductor Conference*, Seattle, USA Tutorial on the minfi and shinyMethyl packages for the analysis of methylation data, July 2013

Lead teaching assistant / Lab Instructor Led labs with interactive exercises on theory, design, interpretation, computation and presentation of results. Beta-tested and proctored quizzes and exams. All teaching assistantships were at the Johns Hopkins Bloomberg of Public Health.

Statistical Methods in Public Health III (PH.140.623), *Winter 2016*

Statistical Methods in Public Health I (PH.140.622), *Fall 2015*

Statistics for Genomics (PH.140.688) *Spring 2015*

Statistical Methods in Public Health III (PH.140.623), *Winter 2015*

Statistics for Genomics (PH.140.688) *Spring 2014*

Teaching assistant

Statistical Reasoning in Public Health I-II (PH.140.611-612), *Summer 2015*
Graduate Summer Institute of Epidemiology and Biostatistics

Statistical Theory I-II (PH.140.731) *Fall 2014*

Statistical Methods in Public Health I-III (PH.140.621-3)

Peer Review Activities

Peer Reviewer for *Bioinformatics*.

Presentations

RAVEL: Removal of Artificial Voxel Effect by Linear regression
PennSIVE group, University of Pennsylvania, Philadelphia, USA, Talk, 2015

Removing inter-subject technical variability in large neuroimaging studies
5th Annual Hopkins Imaging Conference, Johns Hopkins University, Baltimore, USA, 2015.

Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data
9th Annual Symposium and Poster Session on Genomics and Bioinformatics, Johns Hopkins University, Baltimore, USA, Invited Talk, 2015.

Normalization guidelines: Don't throw the baby out with the bathwater!
Genomics for Students, Johns Hopkins University Baltimore, USA, Talk, 2015.

Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data
450k Workshop, Queen Mary University of London, London, UK, Invited Talk, 2015.

Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data
NCI Symposium on Chromosome Biology, Bethesda, USA, Poster, 2015.

Functional normalization of 450k methylation array data improves replication in large cancer studies
ENAR spring meeting, Miami, USA, Talk, 2015.

Functional normalization of 450k methylation array data improves replication in large cancer studies
Delta Omega Poster Competition, Johns Hopkins University, Baltimore, USA, Poster, 2014.

shinyMethyl: interactive tool for 450K arrays
450K Workshop, University College London, London, UK, Poster, 2014.

About Hi-C and other high-dimensional beverages
Genomics Working Group, Johns Hopkins University, Baltimore, USA, Talk, 2014.

Functional normalization (FunNorm): A better alternative to quantile normalization for methylation data
ENAR spring meeting, Baltimore, USA , Poster, 2014.

Functional normalization of 450k methylation array data improves replication in large cancer studies
Hopkins Genetics Day, Johns Hopkins University, Baltimore, USA, Poster, First place, 2014.

Functional normalization for methylation data
Program in Quantitative Genomics Conference, Harvard School of Public Health, Boston, USA, Poster, 2013.

Adaptive Resistant Regression Method (ARRm): A Better Alternative to Quantile Normalization for Methylation Data
Joint Statistical Meetings (JSM), Montréal, Canada, Talk, 2013.

Functional Normalization (FunNorm): A New Approach for Batch Correction of DNA Methylation Data
PennSIVE group, University of Pennsylvania, Philadelphia, USA, Talk, 2013.

Illumina 450k: a Microarray for the Study of DNA Methylation
Genomics for Students, Johns Hopkins University, Baltimore, USA, Talk, 2013.

Statistical Methods for Analysis of Illumina Infinium Methylation Arrays in a Case-Control Study of Colorectal Cancer
Third Annual Lady Davis Institute Scientific Retreat, Montréal, Canada, Poster, 2012.

Normalization Method for Illumina Infinium Methylation Arrays in a Case-Control Study of Colorectal Cancer
Canadian Human and Statistical Genetics Meeting, Niagara Falls, Canada, Selected Talk. Best Research Trainee Presentation Award, 2012.

Asymptotic Eigenvalue Distribution of Random Lifts
Ottawa Undergraduate Research Poster Competition, University of Ottawa, Poster, 2012.

Asymptotic Eigenvalue Distribution of Random Lifts
Annual Faculty of Science Undergraduate Research Conference, McGill University, Montréal, Canada, Poster, Second prize, 2011.

Geometric Interpretation of the Uniformly Minimum-Variance Unbiased Estimator Canadian Undergraduate Mathematics Conference, Université Laval, Québec, Canada, Talk, 2011.

Représentations de nombres premiers par des formes quadratiques binaires
Colloque pan-québécois des étudiants de l'Institut des sciences mathématiques, Université de Montréal, Montréal, Canada, Talk, 2011.