Computational approaches for the study of gene expression, genetic and epigenetic variation in human

by

James Wagner

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

in the School of Computer Science

MCGILL UNIVERSITY

August 2014

Abstract

Advances in high-throughput genomic technology seen in recent years have enabled the measurement of gene expression, DNA sequence polymorphisms, and epigenetic marks such as DNA methylation at thousands or millions of loci in tens or hundreds of samples. Inter-individual variation in gene expression and DNA methylation are present even when samples are drawn from an ostensibly healthy population. These measurements can be expected to vary due to underlying genetic variation, environmental effects, experimental noise, and, in the case of complex tissues, tissue composition heterogeneity among the individuals studied. The research described in this thesis results from the development of computational and statistical methods and their application to the analysis of three main high-throughput genomic experiments, all with the common goal of better characterizing variation of gene expression and DNA methylation in populations and generating hypotheses of the underlying causes of this variation. The first study involved a Hidden Markov Model based approach to detect statistically meaningful levels of allelic expression from experiments that generate a noisy measurement of allelic expression at heterozygous single nucleotide polymorphism (SNP) loci in a set of samples; we also described results seen when applying our approach to a set of lymphoblastoid cell line (LCL) samples. Next is an examination of the relationships between DNA methylation, gene expression and sequence variation in a set of human fibroblast samples, and results showing that information about chromatin accessibility and histone modifications are a more useful predictor of the directionality of these methylation-expression relationships than location of the CpG site relative to the gene alone. Finally is the identification and analysis of co-methylation modules present in adipose tissue samples, the relationship of these modules with Body Mass Index (BMI), DNA sequence variation, gene expression, open chromatin and histone modifications, and an approach to remove effects caused by tissue composition variation in the adipose tissue and re-characterize the relationships

present after correcting for these effects. Together, these studies represent an important contribution to the body of research seeking to better characterize and understand the sources of population level variation in various genetic and epigenetic properties, and introduce several useful tools and important considerations for researchers embarking on these kinds of studies.

Abrégé

Les avancées réalisées au cours des dernières années en matière de technologie génomique à haut débit ont rendu possible la mesure de l'expression des gènes, des polymorphismes de séquences d'ADN et des marques épigénétiques telles que la méthylation de l'ADN pour des centaines d'échantillons à des millions de loci. Il est intéressant de noter que même pour des échantillons prélevés d'individus d'une population en bonne santé, il existe une variation de l'expression des gènes et de la méthylation de l'ADN. Cette variation peut être expliquée par la variation génétique sous-jacente, l'environnement, le bruit expérimental et, dans le cas de tissus complexes, par la variation de la composition cellulaire des individus étudiés. La recherche décrite dans cette thèse est le produit de la mise au point de méthodes computationnelles et statistiques et de leur application afin d'analyser trois expériences génomiques à haut débit. Ces méthodes ont en commun les objectifs de procurer une meilleure caractérisation de la variation de l'expression des gènes et de la méthylation de l'ADN pour une population et de générer des hypothèses expliquant les causes sous-jacentes de cette variation. La première étude utilise un modèle de Markov caché afin de détecter les niveaux statistiquement significatifs d'expression allélique pour une expérience mesurant avec un certain bruit l'expression allélique à des loci hétérozygotes pour chacun des individus d'une population. Nous y décrivons également les résultats observés lors de l'application de notre approche à un ensemble d'échantillons d'une lignée de cellules lymphoblastoïdes (LCL). Ensuite, nous examinons la relation entre la méthylation de l'ADN, l'expression des gènes et la variation des séquences dans un ensemble d'échantillons de fibroblastes humains. Nos résultats démontrent que l'accessibilité à la chromatine et les modifications des histones sont plus importantes que la localisation de sites CpG par rapport à un gène afin de prédire une corrélation positive ou négative entre la méthylation et

l’expression d’un gène. Finalement, nous présentons des méthodes afin d'identifier et d'analyser les modules de co-méthylation dans des échantillons de tissus adipeux, la relation de ces modules avec l'indice de masse corporelle (IMC), la variation des séquences d'ADN, l’expression génique, la chromatine ouverte et les modifications des histones. Nous introduisons aussi une approche visant à éliminer les effets causés par la variation de la composition cellulaire dans des échantillons de tissu adipeux et à caractériser les relations présentes après la correction de ces effets. Ensemble, ces études représentent une contribution importante au corpus de recherche ayant pour but d’offrir une meilleure caractérisation et compréhension des sources de variations de diverses propriétés génétiques et épigénétiques dans une population. De plus, elles proposent plusieurs outils utiles et des considérations importantes pour les chercheurs exécutant des études rattachées à ce domaine.

Acknowledgements

I gratefully acknowledge my supervisor, Professor Mathieu Blanchette for the countless hours of feedback, brainstorming and moral support you have invested over the course of my PhD research. You provide the ideal that any scientist and mentor should aspire towards. My sincere thanks also go to co-supervisor Professor Tomi Pastinen and collaborator Professor Elin Grundberg for the passion for and insight into genetics and human health that you instilled into our research collaborations. Thank-you to Dr. Guillaume Bourque and Dr. Doina Precup, as well as Dr. Mathieu Blanchette and Dr. Tomi Pastinen for your participation in my PhD proposal examination and helpful suggestions offered. To my colleagues and lab-mates in the Blanchette, Pastinen, Grundberg, Ruths and Waldispuhl groups, too many to name, thank-you for the time and effort you invested at every stage in making every facet of my thesis research, from nitty-gritty code questions to major end-of-PhD practice presentations, a success.

Table of Contents

List of Tables

List of Figures

List of Acronyms

AE: allelic expression

aeQTL allelic expression quantitative trait locus

aeSNP: allelic expression single nucleotide polymorphism

CGI: CpG Island

DHS: DNase I Hypersensitive Site

emQTL: expression and methylation quantitative trait loci

eQTL: expression quantitiative trait locus

FDR: false discovery rate; GO: gene ontology

H3K27ac: histone 3 lysine 27 acetylation

H3K4me: histone 3 lysine 4 methylation

H3K27me3: histone 3 lysine 27 tri-methylation

HMM: hidden Markov model

LCL: lymphoblastoid cell line

LTOR-HMM: Left-to-right hidden Markov model

mQTL: methylation quantitative trait locus

TES: transcription end site

TSS: transcription start site

Chapter 1. Introduction

1.1. Overview

This thesis research involves the study of a number of entities in the human cell's nucleus, their properties and their relationships. These components include the genomic DNA, its epigenetic features such as DNA and chromatin modifications, and the level of gene expression. Together these entities form many of the key elements that will determine the cell's protein composition and hence, functionality. Abnormalities in sequence, gene expression or epigenetic modification can lead to disease. We study several datasets consisting of samples drawn from the general population, each study measuring different genetic or epigenetic features for each sample. Where necessary, computational and statistical tools were developed or applied to assess these datasets and the relationship between them in as robust of a fashion as possible, all with the goal of generating biologically interesting hypotheses about gene regulation or human phenotype. In this introductory section, I review in turn each key component of the human cell that will be studied in detail in this thesis, together with the high-throughput genomic technologies used for measuring them, I will then review a selection of key experiments that in recent years have utilized these high throughput technologies to measure the interactions between these components and their contribution to the determination of phenotypes.

1.2. Gene expression and its regulation

1.2.1. Overview of the genome and the central dogma

Genetic regulation can be defined as the highly complex process of integrating information stored in genetic sequences, epigenetic marks and environmental signals to effectively transcribe DNA sequences into messenger RNA (mRNA) and then translate mRNA into proteins, forming the functional workhorses of what will be a complex multi-cellular organism such as human. Given a starting point of 1953 with the model of the chemical structure of DNA and continuing to recent years during which models of gene regulation and its various components have been developed, as well as high-throughput methods to measure them, one can say that the research done to elucidate the mechanisms of gene regulation is one of the greatest human achievements of the second half of the 20th century.

This research endeavours in its own modest way to make a contribution to this mountain of research regarding the various components of gene expression and its regulation, including mRNA transcripts, DNA sequence, and epigenetic marks such as DNA methylation; the variability present in these components; and the relationships between these components.

The process of going from genomic DNA to the functional proteins of the cell can be roughly divided into 5 main steps: ((Carlberg and Molnár 2013), chapter 1) (i) gene transcription from genomic DNA to pre-mRNA, (ii) mRNA processing, (iii) mRNA transport, (iv) translation of the information of mRNA into protein and (v) further protein processing. As it is of the main relevance to considerations in this thesis, in this introductory section I focus mainly on the first step of this process, gene transcription.

1.2.2. Transcriptional gene regulation

At most of the 20,000+ coding and non-coding gene loci, RNA Polymerase II carries out the main enzymatic function of reading the DNA template and forming an mRNA. A variety of intracellular and extracellular signals help to positively or negatively regulate the initiation of RNA Polymerase II-mediated transcription, including thousands of transcription factors (TFs).

Transcription factors (TFs) are proteins that typically bind to short, degenerate nucleotide sequences in the region of the gene, referred to as transcription factor binding sites (TFBS). The promoter region is typically the area in the close vicinity of a few kilobases (kb) from the transcription start site, and is potentially bound by dozens of transcription factors at various points during transcription. In most genes, enhancer regions are found many kb or even megabases (Mb) upstream and downstream of the transcription start site. These regions are also hotspots for binding of various activating or repressive transcription factors. Levels of transcription for a given gene will depend in part on the levels of particular combinations of transcription factors bound to the promoters and enhancers.

Given the relative shortness (6-8 bp) and degeneracy of transcription factor binding sites, and the ~3,000 Mb size of the human genome, additional layers of control beyond the DNA sequence are required in order to achieve the tight levels of gene expression seen in the human cell. (Mohn and Schübeler 2009) illustrate this problem by pointing out that a random 6-mer (corresponding to a short TFBS motif) would be expected to appear by chance 781,200 times in the genome (under some simplistic assumptions about the nucleotide composition of the genome).

An important consideration is the fact that DNA does not carry its information solely in its sequences of A, C, G, and T nucleotides, but that it also has a 3-dimensional structure. This structure will entail that various parts of the genome can interact with each other despite a relatively large distance in terms of number

of base pairs separating these loci in the linear sequence. Of high relevance to these considerations is DNA structure. DNA does not exist as a naked sequence, but rather wrapped around nucleosomes, which are composed of histone proteins. This complex of DNA and proteins is termed chromatin. At a given locus in the genome, constituent nucleotides of the DNA sequence, as well as histones, are both subject to various modifications that can increase or decrease the accessibility of the locus for transcription factors, RNA Polymerase II and other transcription initiation and/or elongation machinery. These marks are termed epigenetic modifications, deriving from a Greek prefix “epi-” meaning “outside of”, and implying that these marks provide information beyond the sequence present in the genome. (Reviewed in (Mohn and Schübeler 2009; Jones 2012)).

The principal epigenetic marks considered in this thesis are DNA methylation and histone modifications and are reviewed briefly in the following section.

1.3. Epigenetics: beyond the sequence

I review here basic properties of two types of epigenetic marks that were investigated in our research: DNA methylation and histone modifications.

1.3.1. DNA methylation

DNA methylation is a covalent modification of DNA, the most well studied example of which in vertebrates takes place at the cytosine site in cytosine-guanine dinucleotides (CpG sites). It consists of adding a methyl ($-CH_3$) group to the 5 carbon of the pyrimidine ring of the cytosine base in DNA, converting this base from cytosine to 5-methylcytosine. The reaction is carried out by enzymes called DNA methyltransferases (DNMT) and involves transferring a methyl group from S-adenosyl methionine (SAM) to cytosine. Two families of DNMT are known in mammals: DNMT1 and DNMT3. DNMT3a and DNMT3b are regarded as *de novo* methyltransferases while DNMT1 plays a maintenance role in newly

replicated cells, scanning a newly synthesized genomic DNA sequence for methylated CpG sites in the mother strand and adding methyl groups to the corresponding CpG sites in the daughter strand. This viewpoint is of course an over-simplification, and (Jones and Liang 2009) present a revised model in which all of DNMT1, DNMT3a and DNMTb play a role in the maintenance of DNA methylation. The existence and process of active DNA demethylation is an even more active area of research. The TET group of proteins has received a great deal of attention and it is now established that proteins in this family can convert 5-methylcytosine to 5-hydroxymethylcytosine and 5-hydroxymethylcytosine to 5-carboxylmethylcytosine, which can be excised via base excision repair to revert to an unmethylated cytosine state (He et al. 2011; Ito et al. 2011).

CpG sites tend to be under-represented in genomes as a direct consequence of their propensity for methylation of the cytosine site and the vulnerability of methylated cytosines to deamination resulting in cytosine to thymine transitions. Methylation is the default state for a large proportion of cytosines present in CpG pairs, with the most important (but not the only) exception being CpG islands; the exact criteria for defining these regions is open to differences of opinion but in general consist of regions of several hundred or thousand base pairs with an enrichment for CpG dinucleotides relative to the genomewide average (Illingworth and Bird 2009). CpG sites in these regions are generally not methylated and hence, the distribution of DNA methylation from a sampling of CpG loci in a vertebrate genome is typically bimodal, with a low methylation mode corresponding to CpG sites within CpG islands, and a high methylation mode corresponding to CpG sites elsewhere. A third mode, though much smaller than the other two, could be assigned to hemi-methylated sites corresponding to imprinted regions in which either the maternally or paternally inherited copy of a locus is silenced early in development via DNA methylation, while the other copy remains unmethylated (Li et al. 1993).

Despite this “tri-modality” of DNA methylation, it should be noted that, for a given locus, a single cell can be methylated at both, one, or neither copies of this

locus. All cells in an individual, or even in a given cell sample taken from an individual, are not expected to follow an identical pattern of methylation (or lack thereof) at a given CpG site. Taking a sample of a group of cells from an individual and measuring the overall methylation level at a CpG site would lead to a continuous measurement that can be thought of as the fraction of CpG alleles that are methylated in the ensemble of cells in that sample. These overall methylation levels can vary between individuals in a population, and this variability is one of the key foci of this thesis.

The list of roles that methylated CpG sites are known to play in cellular function is a long, actively researched, and growing set of key regulatory processes. The particular role played by a CpG site depends heavily on the context of genetic sequence and other epigenetic modifications present in its vicinity. In the most general sense, it is a mark that is repressive to transcription, however it is also important to note that DNA methylation is not hypothesized to play a fully causative, repressive role in all contexts, but could also be a result of other factors leading to transcriptional activity. Some experimental evidence exists that DNA methylation reinforces a transcriptionally inactive state in some circumstances rather than being a straightforward cause or consequence thereof (Blattler and Farnham 2013).

Genomic imprinting and female X chromosome inactivation are two long-studied functions of DNA methylation. In the former, either the maternally or paternally inherited copy of a gene is silenced by copious methylation of CpG sites in its promoter region (Li et al. 1993). In the latter, methylation of one copy of a mammalian female's X chromosome results in this chromosome being largely transcriptionally inactive, to achieve the same overall level of transcription as is seen in males having only a single X chromosome (Mohandas et al. 1981).

Transposable elements comprise a large part of the human genome. In terms of sheer numbers of CpG sites involved, a large fraction of those methylated are in promoters of such elements (Yoder et al. 1997; Walsh et al.

1998). This methylation would lead to transcriptional inactivity as well as increased likelihood of C -> T mutagenesis over time, decreasing the likelihood of these elements continuing to mobilize in the genome, and increasing the overall genomic stability.

Beyond the examples outlined above, CpG methylation in mammals has been investigated the most in genes, particularly in the context of cancer where aberrant methylation is tied to inappropriate activation or repression of cell proliferation related genes. Promoter regions can be divided into two categories based on whether they contain CpG islands or not. Genes whose promoter sites contain CpG islands that are in the more common, unmethylated state are generally repressed via means other than DNA methylation, such as binding of Polycomb proteins. However, methylation of CpG island promoters is seen in regions where a long term fixing of a repressed state is needed, such as in female X chromosome inactivation and imprinted genes. Genes whose promoter region does not contain CpG islands show much more variability in their DNA methylation (Jones 2012).

CpG sites within the bodies of genes are also subject to variable DNA methylation. Exceptionally, this DNA methylation is typically positively correlated with expression of a gene when present in its body rather than near the transcription start site. Current hypotheses lean towards gene body methylation impeding transcription initiation at spurious start sites within the gene body, allowing transcription machinery to more effectively bind and initiate transcription at true start sites (Maunakea et al. 2010).

Enhancers are sites more distal (up to several hundred kb) from genes that also play a role in transcriptional regulation. The functions and effects of enhancer DNA methylation are less well researched than those for promoters. But recent efforts have found active enhancers to be neither completely unmethylated nor methylated, but to exist in states termed “low-methylation” regions (Stadler et al. 2011).

Research in past decades focused on DNA methylation, its patterns and effects at canonical genes, or in the context of diseases such as cancer. With additional high throughput methods for measuring DNA methylation at a wide range of CpG loci in the genome becoming available, some light has also been shed on quantifying DNA methylation distribution and variation in populations of healthy individuals, and its relationship to genetic variation, gene expression and other epigenetic marks. We outline some recent work done to investigate these relationships in section 1.5, and in chapter 3 describe our own results seen with a set of primary untransformed human fibroblasts, noting in particular the presence of both negative and positive correlations between DNA methylation and gene expression that depend less on position with respect to gene body or promoter and more with respect to histone marks in the region. In Chapter 4 we describe some recent efforts to identify clusters (or modules) of CpG sites with similar methylation patterns across a set of study samples (i.e. sites distributed across the genome having high methylation levels in approximately the same subset of individuals), as well as recent efforts to correct for tissue composition effects in measurements of DNA methylation in complex cell mixtures such as whole blood. We then move to outline characteristics of co-methylation modules found in our experiments with a set of primary human adipose tissue samples, and additional insights gained regarding co-methylation patterns and methylation quantitative trait loci (mQTLs) found when correcting for tissue composition effects.

1.3.2. Histone modifications

In eukaryotes, DNA is packaged into nucleosomes, which, on a general scale, tends to reduce access to DNA for the transcription machinery. Additional modifications to histones (i.e. the constituent proteins of nucleosomes) could either further restrict or alleviate access to DNA. Various amino acid residues within histones are subject to various modifications, including methylation, ubiquitination, acetylation and phosphorylation, leading to a combinatorial explosion of possible configurations of histone modifications present in a given

region. Recent efforts to study distributions of particular modifications have pointed towards various transcriptional states, such as active or inactive gene bodies, promoters and enhancers, being correlated with combinations of certain marks (Ernst and Kellis 2010).

Distribution of individual marks, their functions, and implications of a given combination of functions is a growing area of research and beyond the scope of our discussion, for more details, see: (Lee et al. 2010; Suganuma and Workman 2011). We highlight briefly a small number of marks that are most relevant to the discussions that follow. Nomenclature follows the standard format of: H3K4me2, which indicates di-methylation (me2) of lysine 4 (K4) on histone 3 (H3). Methylation of lysine residues 4, 27 and 36 on Histone 3 is one type of modification for which data are available in a wide variety of cell types. While these should only be interpreted as general guidelines rather than deterministic rules, H3K4me3 is typically associated with promoters of active genes, H3K4me2 is found within gene bodies of active genes, and H3K4me1 adjacent to active promoters in some cases and also with more distal enhancers of genes that are either active or poised for activation. Lysine 27 acetylation (H3K27ac) has been shown to be a mark that, together with H3K4me1, signals active enhancers as opposed to poised enhancers (Creyghton et al. 2010). H3K27me3 is indicative of inactive promoters, while H3K36me3 of active gene bodies.

While experiments analyzed in this thesis do not include novel results of specific histone modifications, we make use of public results from ENCODE (Myers et al. 2011) and Epigenomics RoadMap (Bernstein et al. 2010) experiments to better quantify the nature of expression correlated CpG sites (Chapter 3) and co-methylation modules (Chapter 4).

1.4. High-throughput genomic technologies

We describe in the following sections some of the high-throughput technologies upon which our results and analyses depend. We focus primarily on

the specific platforms used as examples. A review of some studies examining data obtained from these high-throughput platforms follows in Section 1.5.

SNP genotyping platforms enable the measurement of DNA sequence at hundreds of thousands or millions of known polymorphic loci distributed across the human (or other organism's) genome. The input to a genotyping experiment will be a sample of DNA derived from one individual; output will typically be, for each SNP across the genome, intensity for each allele present at the locus.

The Illumina BeadArray, which is used for research in this thesis, consists of locus specific 50mers, covalently linked to one of over 1,100,000 bead types. Each bead type is present on average in approximately 30 copies, enabling increased precision and outlier removal. The assay consists of the following main steps. For more details see: (Gunderson et al. 2005); (Gunderson et al. 2006); and http://supportres.illumina.com/documents/myillumina/f2a81381-1faa-45a5-bf4b-8d5d5e770dfe/inf_hd_gemini_assay_user_guide_11311007.pdf.

i) Performing Whole genome amplification of DNA.

ii) Hybridizing amplified genomic loci to an oligonucleotide probe array; the Illumina BeadArray consists of locus specific 50mers, covalently linked to one of over 1100000 bead types. Each bead type is present on average in approximately 30 copies, enabling increased precision and outlier removal.

iii) Washing away unhybridized and non-specifically hybridized DNA.

iv) Extending and staining hybridized DNA, using captured genomic DNA as the primer. Two main types of probes are used in Illumina BeadArray genotyping assays: Infinium I and Infinium II, which lead to different primer design in the extension step. In the case of Infinium I design (allele-specific primer extension), two distinct probes are included on the array for each locus, one for each of the two possible alleles. The 3' terminus is designed to match one of the two alleles present, and extension of this probe, and subsequent signal will only be found if the corresponding allele hybridizes. In Infinium II design (single base extension), only one probe per locus is required. The 3' terminus of the probe complements the base directly upstream of the

query, and different hybridization colours will be seen corresponding to different alleles present.

v) Imaging of BeadChips. Chips are scanned using a reader, which will excite the fluorophore of the extension product on the bead. Probes are linked to addresses and a decoding system is used to assign intensities for each allele, for each genomic locus.

vi) Intensity values for each allele, at each locus, are derived from the images obtained, and output to data files.

These intensity read outs are subject to quality control, filtering and normalization steps. Ratios between the intensities are used to "call" the genotype of the sample, using an approach such as Illuminus (Teo et al. 2007). A process called imputation (Marchini and Howie 2010) can also be applied to infer the genotypes at loci not measured in the given platform, by taking advantage of measurements available from a genetically similar set of reference individuals for which measurements at these loci are available, as facilitated, for example, by HapMap (Gibbs et al. 2003). This process enables further fine mapping as well as meta-analyses between studies using different genotyping chips. Hand in hand with imputation is phasing, which is the process of inferring the haplotypes, i.e. the sequences of alleles as they would appear on one of the chromosomes (see (Browning and Browning 2009) for an approach that integrates imputation and phasing into a single method). For any genotyping method applied, it should be noted that with an estimated 10 million SNPs in the human genome, a platform or combination of platforms may only cover ~10-20% of the SNPs. Though the haplotype block structure assures that many untyped SNPs can be imputed from observed SNPs and from genotyped SNPs in other populations, ultimately phasing and imputation are not trivial problems. The possibility of stratification of SNPs in sub-popuations that co-localize by chance with a phenotype of interest is also an an important caveat to any association study carried out with SNPs, as are quality control and accuracy of genotype calling. For a discussion of all of these considerations, see (Teo 2008).

1.4.1. Gene expression

Gene expression microarrays used in our work (Illumina HumanRef-8 and Human HT-12) share many basic steps in common with genotyping assays just described. Probe design is instead for 50-mer sequences specific to genes on the assay, with biotin nucleotides incorporated to allow imaging. Input consists of mRNA converted to complementary RNA or DNA (cRNA or cDNA), and output is intensity values for each probe (with one or more probes per annotated gene). For more details see: http://www.illumina.com/products/humanht_12_expression_beadchip_kits_v4.ilmn and (Fan et al. 2006). Careful normalization of probe intensities is essential as are considerations of possibilities such as are considerations of probe cross-hybridization (Wu et al. 2005) and RNA degradation (Opitz et al. 2010).

1.4.2. DNA Methylation

The Illumina HumanMethylation450 platform's steps and design are similar to those of Illumina's genotyping platforms. Bisulfite treatment of DNA will convert unmethylated cytosine residues to uracil, whereas methylated residues remain unaffected. Type I and Type II Infinium probes can be therefore designed in a similar fashion to those described for genotyping assays. This is achieved by having either two probes corresponding to unmethylated or methylated loci, or incorporating either a labelled G corresponding to the methylated C, or a labelled A base corresponding to the bisulfite-converted uracil which behaves as thymine in base pairing. The output for each CpG site (probe) on the assay are intensities for unmethylated (u_i) and methylated alleles (m_i). The overall methylation level at a site is estimated using a so-called beta value, determined as $\beta_i = \frac{m_i}{u_i+m_i+\alpha}$, where α is a regularizing parameter typically set to 100. For more details on this methylation platform see (Sandoval et al. 2011). While highly practical as an assay for reporting the methylation levels of a wide variety of CpG sites at relatively low cost, the HumanMethylation450 platform still contains a probe for only ~1% of CpG sites in the human genome, many of which must be excluded

for methylation QTL (mQTL) studies because of the presence of a SNP within the probe itself, or due to issues of mapping to multiple loci in the genome. The existence of two types of probes which yield different distributions of beta values must be kept in mind at each stage in the analysis process (Yousefi et al. 2013).

1.4.3. DNase I Hypersensitive Sites

Accessible chromatin, typified here as DNase I Hypersensitive Sites (DHSs) is a marker of an active regulatory element, be it an enhancer, promoter, silencer, insulator or locus control region. These regions are not wrapped tightly around histones and their sensitivity to DNase I cleavage enables their isolation, sequencing and detection in such an experiment. We make use of ENCODE derived DHS data (Thurman et al. 2012) from a number of cell lines. These data were processed according to either the University of Washington (John et al. 2011) or Duke (Boyle et al. 2008a) protocols. Both protocols involve treatment of intact nuclei with the enzyme DNase I to cleave exposed DNA. Protocols differ in terms of fragment size and other specifics, but in both cases DNA is isolated following DNase I treatment and the library is sequenced on an Illumina instrument. Sequencing reads are aligned to the genome, and the Hotspot algorithm (John et al. 2011) was used to detect peaks corresponding to a large number of reads, or regions of highly accessible DNase I hypersensitive sites. Briefly, the Hotspot algorithm uses a sliding window and the binomial distribution to estimate enrichment of sequence tags based on a local background model estimated around every tag. Hotspot also includes a false discovery rate (FDR) estimation procedure for thresholding hotspots and peaks, based on a simulation approach involving reads.

1.4.4. Histone modifications: Chromatin Immunoprecipitation and sequencing (ChIP-seq)

We utilized ENCODE and Roadmap Epigenomics public datasets for histone modification data in our experiments. ENCODE followed a protocol

detailed in (Landt et al. 2012) for their ChIP-seq experiments. ChIP-seq is a method for mapping the binding sites across the genome by a protein of interest, such as a transcription factor or post-translationally modified histone. ChIP-seq involves the treatment of cells with formaldehyde to cross-link proteins covalently to DNA. Cells are sonicated and the DNA digested to obtain chromatin of approximately 100-300 bp. The protein of interest and its bound DNA is enriched by purification with an antibody specific to that protein. Enriched DNA is purified and input to a sequencing machine. The sequencing reads are then mapped to the genome and can then be counted to obtain the number of reads found at a locus. Control data is critical as DNA breakage during sonication is not uniform. "Input DNA", i.e. a lysate of the same cell type being studied with a control antibody is typically used. After this, a peak-calling algorithm is called to identify regions enriched for mapped reads relative to input DNA and relative to their genomic background. MACS (Zhang et al. 2008) was applied to ENCODE ChIP-seq datasets.

1.5. Association studies for high-throughput datasets

I have thus far reviewed many of the biological entities such as DNA sequence and sequence polymorphism, DNA methylation, gene expression, open chromatin and histone modifications that are present in the cell and are essential for the phenotype and function seen in the organism, as well as some high-throughput platforms used for measuring the levels of each of these at multiple loci across the genome in an individual or set of individuals. As already mentioned at several points, each of these marks does not exist in isolation but rather these marks show correlations and relationships, interacting in complex ways. The nature and scope of these interactions, causal directions and implications for cellular function and disease are still very much active areas of research, and some experiments that have highlighted some of these interactions are reviewed in this section.

A useful and fascinating historical summary of the research done in past years and decades that made the experiments described in this thesis possible is presented by (Altshuler et al. 2008). In summary, research that sought to find loci in the genome associated with disease or other phenotype began with linkage analysis. This analysis is based on polymorphic variants in the DNA, or markers, which show correlated segregation with the trait of interest. In humans, linkage analysis took off with the ability to map microsatellites (tandem repeats of simple sequences) and systematically trace the transmission of chromosomal regions, as well as the trait of interest, in pedigrees. The feasibility of this approach was demonstrated with the localization of Huntington disease in 1983 (Gusella et al. 1983).

For complex traits that do not follow a classic Mendelian, single-gene form of causation, success with linkage analysis was limited. Association studies, which determined the correlation or relationship between a trait of interest and a panel of genetic variants in a set of unrelated individuals, began to show more and more promise. Proposals were developed (Risch and Merikangas 1996) to develop large panels of Single Nucleotide Polymorphisms (SNPs) to carry out association studies in a systematic, genome-wide manner. Soon thereafter, researchers published results demonstrating the feasibility of using microarray technology to interrogate thousands of SNPs distributed across the genome (Wang et al. 1998). These platforms would form the basis of genome wide association studies (GWAS) for detecting genetic variants correlating with a trait.

SNPs do not occur independently of their neighbours, and typically blocks of SNPs located close to each other will correlate and share a common set of alleles, termed a haplotype. Because of the low recombination rate in humans, a marker allele correlated with a trait in the population will typically show association with nearby marker alleles for many generations. This phenomenon is termed linkage disequilibrium (LD) (Wall and Pritchard 2003), and enabled the inference of genotypes of common SNPs from knowledge of only a few empirically measured SNPs (Johnson et al. 2001). The International HapMap

Project was launched in 2002, with the objective of characterizing SNP frequencies and levels of linkage disequilibrium across the genome in various human populations (Frazer et al. 2007).

Contemporary with developments in SNP genotyping microarray technology were gene expression microarray technologies (DeRisi et al. 1996). Together these converged to treating expression levels of genes as measured in a microarray experiment as the quantitative trait of interest, and associating these with genetic variation in cis (relatively near the gene) or trans (genome wide, potentially very far from the gene or on another chromosome). These studies were termed expression quantitative trait locus (eQTL) studies. Even taking into account the corrections for multiple hypothesis testing that are necessary for an experiment with such a large number of association tests (Kendziorski et al. 2006), early studies with HapMap LCLs confirmed that expression traits are mappable to genetic variation at levels much more substantial than expected by chance (Dixon et al. 2007; Göring et al. 2007; Stranger et al. 2007). Each human autosomal gene exists in two copies, and variations in cis of regulatory sequences adjacent to one copy of a gene may be expected to lead to unequal expression levels of the two copies of the gene, this phenomenon is termed allelic expression (AE). A microarray or RNA sequencing assay that measures expression levels for both copies of each gene in the platform can provide allelic expression levels, which can in turn be correlated to adjacent SNPs. Detection of allelic expression is covered in more detail in Chapter 2.

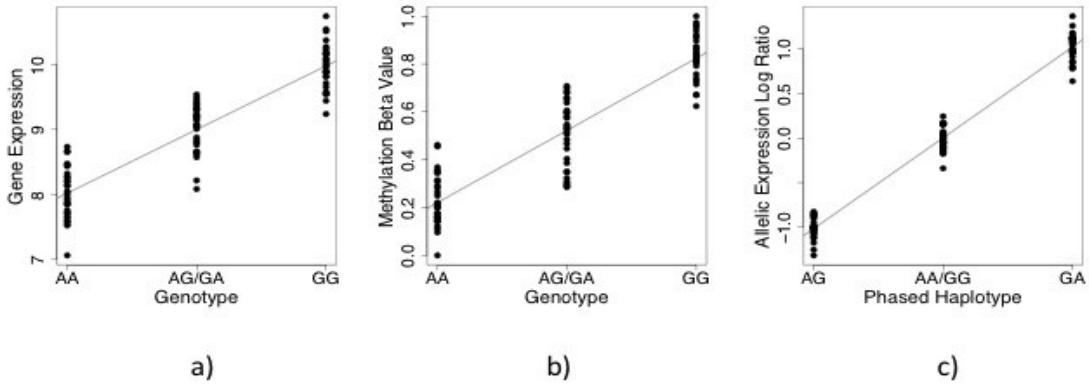

Figure 1.5-1 Comparison of QTL studies involved in this thesis.

(a) An eQTL study correlates the genotype of a particular SNP with expression intensities of a microarray probe corresponding to a particular gene. b) a methylation QTL (mQTL) study correlates genotype with methylation beta values for a probe corresponding to a CpG site. C) In an allelic expression study, expression intensities of both alleles for a heterozygous SNP are measured, and the log ratio obtained. Individuals who are heterozygous for a particular regulatory SNP are expected to show highly positive or negative allelic expression log ratios, whereas those that are homozygous will have a ratio closer to zero. Environmental or trans acting effects are expected to affect both alleles equally, thus the allelic expression measurement provides an internal control, and tighter regressions in which a larger proportion of expression variance is explained by cis regulatory loci.

Hypotheses of how cis-eQTLs can be expected to alter gene expression were recently summarized by (Gaffney 2013) and include 1) polymorphisms in binding site sequences changing the ability of activating or repressive transcription factors to bind to a site at or near the gene's promoter, through either directly altered binding site sequences (Kasowski et al. 2010) or locally altered chromatin structure (Degner et al. 2012) 2) Altering the likelihood of CpG sites in the vicinity of the gene to be methylated (Bell et al. 2011) 3) Co-transcriptional regulatory variation in the form of altered splicing (Kwan et al. 2007), mRNA decay (Pai et al. 2012) or polyadenylation (Yoon et al. 2012).

GWAS and eQTL studies do not exist in isolation; insights from the two can be combined to shed light both on genetic mechanisms of disease and the genetics of gene expression. Indeed, many other factors such as DNA methylation, binding of transcription factors, chromatin accessibility and histone modifications can and have been taken into account in association studies to dissect the various contributing factors of gene regulation.

Recent review articles highlight several key considerations of recent and future research, including the frequency of cis vs trans regulatory loci, similarities and differences between regulatory relationships in different tissues or cell types, pinpointing causative loci in disease or phenotype, finding master regulators and systematic responses via module or cluster based analyses, integrating results from multiple genomic and epigenomic experiments, and utilizing increasingly high throughput datasets with greater sample size and greater number of loci measured. (Gilad et al. 2008; Majewski and Pastinen 2011; Nica and Dermitzakis 2013). I review in the remainder of this section some key methods and results of several selected papers, chosen for their integration of more than one of the key considerations. Additional discussion of these and related works can be found in introductory sections to manuscripts presented in Chapters 3 and 4 (i.e. sections 3.3 and 4.3). An illustration of expression, methylation and allelic expression QTL mapping is provided in Figure 1.5.1.

1.5.1. Genetics of gene expression in multiple tissues and obesity (Emilsson et al. 2008)

With earlier eQTL studies focusing on lymphoblastoid cell lines (LCLs), this was one of the first large scale investigations utilizing primary, untransformed tissues. Investigators collected blood and subcutaneous adipose tissue from hundreds of Icelandic samples, measured biometric traits such as cell counts in blood and BMI. They carried out both cis and trans eQTL studies, finding considerable overlap between adipose and blood of cis-eQTLs (more than 50% of adipose eQTLs were also present in blood). 14.6% and 11.5% in adipose and

blood respectively had a cis-eQTL, adjusting for age, sex and BMI in blood increased by the number of mappable genes about 25%.

Researchers then transformed expression data into a coexpression network by performing a module detection algorithm similar to that of WGCNA (Langfelder and Horvath 2008), outlined in Section 4.1. A conserved module found between this human adipose experiment and a previously published experiment in mouse (Chen et al. 2008) was found to be enriched for GO terms related to macrophage function and correlated with BMI at FDR < 1%.

1.5.2. DNase I sensitivity QTLs (Degner et al. 2012)

As an example of integrating other data besides expression and DNA sequence variation to better dissect contributing factors to gene expression variation, (Degner et al. 2012) measured DNase I Hyper-sensitivity in 70 HapMap Yoruba LCLs, integrating these with expression and genotype data, to find genetic loci correlating with DNase I hypersensitivity (dsQTLs). They found nearly 9000 loci with DHS read count correlating significantly to genetic sequence within a 40 kb window, estimating that as many as 55% of eQTL loci are also dsQTLs. They also applied a Bayesian hierarchical model to obtain average properties of inferred causal sites (Veyrieras et al. 2008), finding that inferred causal variants are close to the transcription start site and that information about the presence of methylation QTLs (mQTLs) and transcription factor binding further pinpoint key dsQTLs.

1.5.3. Methylation QTLs (Bell et al. 2011)

Researchers carried out systematic analysis of relationships between gene expression, DNA methylation and sequence variation in LCLs. They utilized a predecessor of Illumina's HumanMethylation450, the HumanMethylation27 BeadChip. Even with this platform's strong bias for low variance CpG islands, a considerable enrichment for mQTLs was found, compared to what was expected

by chance. Likewise, though the fraction of mQTLs also shown to be eQTLs was modest, it was considerably better than expected by chance.

1.5.4. mQTLs in adipose and eQTLs in multiple tissues (Grundberg et al. 2012; Grundberg et al. 2013)

This work forms the base of datasets analyzed in Chapter 4, in the context of comethylation and correction for tissue composition effects in adipose tissue. (Grundberg et al. 2012) carried out cis and trans eQTL analyses in human adipose, skin and LCL samples taken from the Multiple Tissue Human Expression Resource (MuTHER) project (Nica et al. 2011). They found strong overlap in cis-eQTLs found, as well as evidence of cell type specific trans-QTLs associated with expression of multiple genes. They also demonstrated the utility of combining GWAS and eQTL data by obtaining GWAS SNPs from the National Human Genome Research Institute GWAS catalog (Welter et al. 2014), finding enriched overlaps between eQTLs for each tissue and GWAS traits corresponding to the cell type (i.e. an enriched overlap of immune function GWAS hits and LCL eQTLs).

(Grundberg et al. 2013) did cis-mQTL analysis with methylation values measured using the Illumina HumanMethylation450 platform in subcutaneous adipose tissue, from the same individuals as above. They found an enrichment of mQTLs and that 6% of mQTLs found also associated significantly with expression. mQTL findings were overlapped with RoadMap Epigenomics (Bernstein et al. 2010) data, showing mQTLs that overlap eQTLs restricted to adipose tissue or metabolic trait related loci were enriched in overlap with enhancer (H3K4me1) marks.

1.6. Thesis Roadmap and Rationale

The previous sub-sections have made clear that the regulation of gene expression is a complex phenomenon shaped by billions of years of evolution.

High throughput technologies to study this technology are by comparison but a few years old and much more work is needed to develop statistical and computational methods for the study of these data and the extraction of as much data from them as possible. This thesis work endeavours to make several contributions to this front.

This introductory chapter introduced some background and helped to put my research in the context of recent experiments in high-throughput genomic and epigenomic technologies and the variation of various measurements in human populations. The following three chapters cover the research contributions of this thesis. Chapter 2 describes an approach based on Hidden Markov Models (HMMs) used to estimate levels of allelic expression from a high-throughput array based experiment. Chapter 3 describes investigations of DNA methylation, gene expression and genetic sequence variation in human fibroblast, with the added context of histone modifications and open chromatin derived from public datasets. Chapter 4 describes research done on comethylation in human adipose tissue and its epigenomic properties, tissue composition effects that drive these relationships and a novel approach to remove these effects. The concluding Chapter 5 summarizes research contributions and possibilities for future work for this research.

1.7. Publications and Author Contributions

This thesis is composed of the full text and figures of three scientific articles, each of which has been published or about to be submitted for publication in a peer-reviewed journal. I am the first author of each article. These articles are as follows:

Chapter 2

Wagner JR, Ge B, Pokholok D, Gunderson KL, Pastinen T, Blanchette M. 2010. Computational Analysis of Whole-Genome Differential Allelic Expression Data in Human. *PLoS Comput Biol* **6**(7).

The development of the computational tools in this publication was done by me under Dr. Mathieu Blanchette's supervision and was applied to biological datasets generated by Dr. Bing Ge and Dr. Tomi Pastinen of McGill University / Genome Quebec, in collaboration with Dr. Dmitri Pokholok and Dr Kevin Gunderson at Illumina, Inc. I wrote the manuscript, with input from my supervisors.

Chapter 3

Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. 2014. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome Biology* **15**(2):R37.

Computational and statistical analysis of datasets was done by me under the supervision of Dr. Mathieu Blanchette. Datasets were generated by Dr. Bing Ge, Dr. Stephan Busche, Dr. Tony Kwan and Dr. Tomi Pastinen of Genome Quebec / McGill University. I wrote the manuscript, with input from my supervisors.

Chapter 4

Wagner JR, E Grundberg, Blanchette M. 2014. DNA co-methylation and tissue composition effects in human adipose tissue. *In preparation*. Development of the computational tools was done by me and Dr. Mathieu Blanchette. Analysis

of data was done by me under Dr. Mathieu Blanchette's supervision. Data were generated by Dr. Elin Grundberg of McGill University, during the course of her research at the Wellcome Trust Sanger Institute, and other analyses of these data were previously published in (Grundberg et al. 2012; Grundberg et al. 2013). We anticipate submission of this manuscript to a peer-reviewed journal such as Genome Biology during the fall of 2014. I wrote the manuscript, with input from my supervisors.

Chapter 2. Computational Analysis of Whole-Genome Differential Allelic Expression Data in Human

2.1. Preface

Allelic expression (AE, used interchangeably with allelic imbalance or AI in the following two chapters) is the process whereby the two alleles of a gene are expressed at unequal levels in an organism, cell or tissue. Our collaborators had developed a method to interrogate the level of AE at hundreds of thousands of SNP probes distributed across the genome, aggregated levels of AE for each consecutive group of SNPs located in an annotated gene or intergenic region, and associated these aggregate AE levels to cis regulatory SNPs (Ge et al. 2009). We demonstrate in this chapter results obtained by characterizing the level of AE in a set of lymphoblastoid cell lines (LCLs) using only the measurements themselves and their genomic coordinates, i.e. no aggregation is done based on gene annotations. The method of choice was a Left-to-Right Hidden Markov Model (LTOR-HMM). The following chapter shows some results of applying our method to AE measurements for a set of fibroblast samples, and correlating the HMM-post-processed measurements with SNPs in cis to obtain candidate regulatory SNPs, which were intersected with methylation QTLs (mQTLSs) to obtain a set of methylation-expression QTLs (meQTLs). We found improved ability to detect cis regulatory SNPs and meQTLs, compared to results obtained with a microarray and eQTL correlation experiment. I review in this preface several other works characterizing or measuring allelic expression prior to our work, which was published in PLoS Computational Biology in 2010 (Wagner et al. 2010).

2.1.1. Widespread monoallelic expression on human autosomes (Gimelbrant et al. 2007)

This was a study seeking to measure the incidence of what was termed "monoallelic expression" in LCL autosomes and could be regarded as cases of substantial allelic imbalance. A genotyping experiment was done with the Affymetrix Human Mapping 500 K dataset, utilizing reverse transcribed mRNA (complementary DNA or cDNA) and a separate experiment with the same platform using genomic DNA. Monoallelic expression in a given gene and a given clonal cell line was called if multiple informative SNPs were called homozygous in the cDNA but heterozygous in the gDNA. 2.2% of genes were called as monoallelically expressed with multiple informative SNPs per clone. The reverse transcription polymerase chain reaction (RT-PCR) was used to confirm specific cases.

2.1.2. BeadArrays for allelic expression (Serre et al. 2008)

This work adapted an Illumina genotyping BeadArray to measure intensity levels for both cDNA and a genomic DNA (gDNA) control. This was the first work that not only quantified allelic expression on a genome wide scale, but correlated these results to genotypes in the same individuals in order to find candidate cis-regulatory loci. Of 56 genes that were found to have informative allelic expression, 23 were also found to also map in cis to genetic variation.

2.1.3. Allele specific expression patterns in leukemia (Milani et al. 2009)

In this work, researchers performed a genome-wide assessment of allelic expression in 8000 genes from bone marrow taken from 197 Nordic children diagnosed with acute lymphoblastic leukemia (ALL). They also correlated methylation levels with allele specific expression. Again, genotyping was done of cDNA and normalized to genomic DNA. Allele specific expression was found in

16% of genes with informative SNPs, and methylation variation was found to correlate with allele specific expression.

2.1.4. Large scale LCL study for identifying aeSNPs (Ge et al. 2009)

This research carried out for this article by some of my collaborators and co-supervisor furnished the datasets that were analyzed in the course of this chapter's research. Whereas my research was focused on detecting allelic expression from this type of experiment without reference to genome annotations, this research was primarily focused with correlating SNP haplotypes with allelic expression levels to detect allelic expression QTLs (aeQTLs, referred to in this research as cis regulatory SNPs). Intensities for genomic DNA (gDNA) and mRNA reverse transcribed to complementary DNA (cDNA) were separately measured using the Illumina Human 1M platform in 53 LCL samples from the HapMap CEU cohort. SNPs went through various filtering steps and cDNA intensity levels were normalized to gDNA levels. SNPs were partitioned into windows based on annotated gene boundaries and aggregate allelic expression levels in each window were correlated to phased HapMap SNPs for the same sample set. As allelic expression is expected to specifically unmask cis regulatory relationships, only candidate cis regulatory SNPs within 250 kb of an allelic expression window were considered. Results were found to yield substantial overlap with previously published eQTL research (Dixon et al. 2007; Marioni et al. 2007) with the same cell lines but pointed to many cases of stronger associations being detected with the allelic expression based methods. The utility of allelic expression for finer mapping of SNPs in GWAS studies was also demonstrated by further dissecting results from an previously published autoimmune disorder GWAS studies (Barrett et al. 2008; Cooper et al. 2008; Hom et al. 2008) and finding candidate regulatory loci at a more precise level than previously published.

2.2. Abstract

Allelic imbalance (AI) is a phenomenon where the two alleles of a given gene are expressed at different levels in a given cell, either because of epigenetic inactivation of one of the two alleles, or because of genetic variation in regulatory regions. Recently, (Ge et al. 2009) have described the use of genotyping arrays to assay AI at a high resolution (~750,000 SNPs across the autosomes). In this paper, we investigate computational approaches to analyze this data and identify genomic regions with AI in an unbiased and robust statistical manner. We propose two families of approaches: (i) a statistical approach based on z-score computations, and (ii) a family of machine learning approaches based on Hidden Markov Models. Each method is evaluated using previously published experimental data sets as well as with permutation testing. When applied to whole genome data from 53 HapMap samples, our approaches reveal that allelic imbalance is widespread (most expressed genes show evidence of AI in at least one of our 53 samples) and that most AI regions in a given individual are also found in at least a few other individuals. While many AI regions identified in the genome correspond to known protein-coding transcripts, others overlap with recently discovered long non-coding RNAs. We also observe that genomic regions with AI not only include complete transcripts with consistent differential expression levels, but also more complex patterns of allelic expression such as alternative promoters and alternative 3′ ends. The approaches developed not only shed light on the incidence and mechanisms of allelic expression, but will also help towards mapping the genetic causes of allelic expression and identify cases where this variation may be linked to diseases.

2.3. Author Summary

Measures of gene expression, and the search for regulatory regions in the genome responsible for differences in levels of gene expression, is one of the key paths of research used to identify disease causing genes, as well as explain

differences between healthy individuals. Typically, experiments have measured and compared gene expression in multiple individuals, and used this information to attempt to map regulatory regions responsible. Differences in environment between individuals can, however, cause differences in gene expression unrelated to the underlying regulatory sequence. New genotyping technologies enable the measurement of expression of both copies of a particular gene, at loci that are heterozygous within a particular individual. This will therefore act as an internal control, as environmental factors will continue to affect the expression of both copies of a gene at presumably equal levels, and differences in expression are more likely to be explicable by differences in regulatory regions specific to the two copies of the gene itself. Differences between regulatory regions are expected to lead to differences in expression of the two copies (or the two alleles) of a particular gene, also known as allelic imbalance. We describe a set of signal processing methods for the reliable detection of allelic expression within the genome.

2.4. Introduction

In a diploid cell, each gene is present in two copies. The vast majority of microarray-based or RNA sequencing-based gene expression studies do not distinguish between the two copies and measure the sum of the expression of the two alleles. This hides the fact that the two alleles are not necessarily expressed at equal levels, a phenomenon called allelic imbalance (AI) (Pastinen and Hudson 2004). The complete shut down of one allele results in monoallelic expression (ME). The most drastic example of ME is X-chromosome inactivation, where, in females, one of the two copies of the X chromosome is inactivated and packaged into heterochromatin (Carrel and Willard 2005). Less drastic is random monoallelic expression, whereby a randomly selected copy of a gene or chromosomal region is silenced by epigenetic mechanisms (e.g. methylation). In contrast, imprinting results in parent-of-origin specific inactivation of the maternal or paternal allele, depending on the locus. While monoallelic expression

completely silences one of the two alleles, less drastic allelic expression differences can result from a heterozygous ***Aa*** regulatory site. For example, allele ***A*** of a transcription factor binding site may allow binding and result in normal expression of the target gene on that chromosome, while allele **a** may disrupt the binding site, resulting in lower expression. While the lower expression of allele ***a*** may be compensated by an increased transcription rate at allele ***A*** in heterozygous individuals, this may not be the case for individuals who are homozygous ***aa***, which may result in phenotypic variation. Researchers have tried to identify causative regulatory variants by measuring the total expression (i.e. expression of both copies) of a particular gene across multiple individuals, treating this as a Quantitative Trait Locus (eQTL), and mapping nearby cis-regulatory regions to the gene expression (reviewed in (Rockman and Kruglyak 2006). A key problem with this type of approach is that environmental differences across individuals can affect gene expression, making the mapping problem very challenging.

Instead, a focus on the relative expression of two alleles within the same cell has been suggested to factor out environmental sources of variation, allowing for more sensitive and specific detection of epigenetic and genetic phenomena related to local control of gene expression (Pastinen et al. 2004). Combining AI measurements obtained from a set of individuals with genotyping information about these same individuals, one can map cis-regulatory variants (Pastinen et al. 2005; Campino et al. 2008; Serre et al. 2008; Verlaan et al. 2009) or detect epigenetic variation in allelic expression (Gimelbrant et al. 2007; Pollard et al. 2008).

Past studies with the goal of detecting AI have typically relied upon panels of SNPs with relatively low density, located in only a subset of transcribed genes of the genome (Lo et al. 2003; Pant et al. 2006; Gimelbrant et al. 2007). A simple threshold for the ratios of expression of the two alleles at a heterozygous locus is usually established (e.g. 1.5 or 2-fold) and a gene is called as imbalanced based upon whether or not the SNP(s) within it exceed this threshold. Optimal AI

profiling in a genome-wide manner would require high-density sampling of expressed heterozygous sites in the genome. We recently generated the first large-scale, high-resolution assay of allelic expression (Ge et al. 2009). In this study, Illumina genotyping arrays were used to measure differential allelic expression at 755,284 polymorphic sites in lymphoblastoid cell lines (LCL) derived from 53 CEU samples included in the HapMap project (Frazer et al. 2007). Because of the noise in single point AI measurements made at each heterozygous locus, sophisticated analytical methods are required to make the most out of this data. In this paper, we develop signal processing approaches for the accurate identification and delineation of transcripts with allelic imbalance, either in a single individual at a time, or in a collection of samples.

To our knowledge, no hypothesis-free computational approaches have been proposed for the analysis of this type of data. Detection of AI in (Ge et al. 2009) relied heavily upon RefSeq, Vega, and UCSC gene annotations, and SNPs were first partitioned into windows corresponding to these annotated regions as well as intergenic regions and windows with significant AI were reported. Sophisticated bioinformatics approaches have been developed for a related, but simpler, problem in the past, that of detecting Copy Number Variants (CNV) or Loss Of Heterozygosity (LOH) in cancer cells using array-based Comparative Genomic Hybridization (CGH) (Shah et al. 2006; Marioni et al. 2007; Rueda and Diaz-Uriarte 2007; Shah 2008) or genotyping arrays (Nannya et al. 2005; Baross et al. 2007; Bengtsson et al. 2008; Li et al. 2008; Yau and Holmes 2008; Wu et al. 2009). These include the PennCNV program (Wang et al. 2007) and the QuantiSNP program (Colella et al. 2007), that use a Hidden Markov Model related to one of the approaches considered here. However, CNV or LOH regions have properties that make them easier to detect than regions of allelic imbalance: (i) the signal, coming from genomic DNA is generally quite strong, whereas gene expression can be very low; (ii) the number of copies of an allele is a small integer, whereas the allelic expression ratio is a real number; (iii) the regions affected are typically quite large, whereas AI can affect a single, short

gene, or even only part of a gene. The approaches listed above are thus not easily applicable to the detection of AI in gene expression. An alternate family of statistical approaches called changepoint methods has been proposed for segmenting array CGH data into regions exhibiting consistent signals (Fearnhead 2006; Browning 2008). These non-parametric, model-free approaches have the benefit of segmenting real-numbered data without enforcing discretization. However, they are difficult to generalize to a situation like ours, where signals come from a mixture of discrete (sites with no expression, sites with expression but no imbalance) and continuous (sites with real-valued imbalance) state space.

In this paper, we introduce a family of signal processing approaches for the analysis of AI data obtained from genotyping arrays. We consider both statistical approaches (Z-score computation) and machine learning approaches (Hidden Markov Models) to identify transcripts that show AI and to quantify the latter. We introduce a new type of left-to-right HMM for the joint prediction of allelic imbalance in the 53 samples considered. Our algorithms are evaluated using permutation testing and succeed at identifying regions with known AI. Our approaches reveal that more than 25% of transcripts (coding or non-coding) are subject to differential expression between the two alleles and that patterns of AI are varied and complex. The tools and data sets described here will help biologists and geneticists to identify regions of allelic imbalance, understand the mechanisms at play, identify the genetic or epigenetic causative agents, and associate expression polymorphisms with disease susceptibility.

2.5. Methods

2.5.1. Allelic Imbalance Data

Allelic imbalance was assayed using Illumina Infinium Human1M/Human1M-Duo SNP bead microarrays. These arrays, originally designed for genotyping, have probes for approximately 1.1 Million polymorphic sites from HapMap, of which 755284 were used for this study. Each probe estimates the abundance of each of the two possible alleles in the sample. Normally, genomic DNA is hybridized onto the chip and the genotypes are easily inferred from the probe intensities. We have previously described how one can take advantage of this technology to measure allelic expression in a high-resolution, genome-wide manner (Ge et al. 2009). Briefly, total RNA is extracted and cDNAs are synthesized based on a protocol on heteronuclear RNA, allowing us to measure unspliced primary transcripts (Verlaan et al. 2009). The cDNA sample is hybridized onto the array and each probe estimates the abundance of each of the two alleles in the sample. In parallel, genomic DNA from the same cell line is hybridized, which provides the basis for normalization of the cDNA hybridization while providing us with the genotype of each sample. Details for the full process of experimentally obtaining the raw imbalance information, as well as the sample information, can be obtained from (Ge et al. 2009).

Data obtained from technical replicates show that although the total expression level (sum of RNA abundance in both alleles) measured at a given SNP is highly reproducible (R^2 = 0.864), single point allelic expression ratios are much more noisy (R^2 = 0.632), especially for low expression levels (see 9). This suggests that careful data analysis is required to extract as much information as possible.

Let $a_i = [a_{i1}, a_{i2}]$ be the set of two alleles present at polymorphic site i in the population, for $i = [1 \ldots n]$ (the rare cases where three or more alleles exist at the same site are ignored in this study). For notational simplicity, we assume that

the genome consists of a single pair of chromosomes. In reality, the analysis that follows is repeated separately for each autosome. Genotype phasing consists of the decomposition of the genotype of an individual into its two homologous chromosomes. For individual k, let $x = x_1, x_2, \dots, x_n$ and $y = y_1, y_2, \dots, y_n$, be these two chromosomes, where $x_i, y_i \in a_i$. Phasing remains a computationally and statistically challenging problem (Browning 2008). In the case of HapMap individuals, phased genotypes are available, although they are not error free. Removal of SNPs not phased in CEU HapMap release R22 resulted in 755284 SNPs that were utilized in our study.

Let $X^K_{DNA}(a_{i1})$ and $X^K_{DNA}(a_{i2})$ be the intensity read outs obtained from the probes interrogating site i when hybridizing the genomic DNA of individual k. If individual k is heterozygous at site i (i.e. $x^k_i \neq y^k_i$), then we expect both $X^k_{DNA}(a_{i1})$ and $X^K_{DNA}(a_{i2})$ to be large. When it is homozygous, say for a_{i1}, (i.e. $x^k_1 = y^k_i = a_{i1}$), we expect $X^K_{DNA}(a_{i1})$ to be large and $X^K_{DNA}(a_{i2})$ to be small. The genotype of an individual can thus be deduced from the ratio of the two measurements.

Consider now $X^K_{RNA}(a_{i1})$ and $X^K_{RNA}(a_{i2})$, the intensity read outs obtained from the probes interrogating site i when hybridizing cDNA obtained from whole cell RNA extraction. When heterozygous site i sits in a transcribed region with no allelic imbalance, both $X^K_{RNA}(a_{i1})$ and $X^K_{RNA}(a_{i2})$ will be relatively large. Any difference between the two may indicate allelic imbalance. Regions that are not transcribed will obtain low values for both alleles. We consider the following pair of observations at each site i : $E^k_i = \log\left(\frac{X^k_{RNA}(a_{i1})+X^k_{RNA}(a_{i2})}{X^k_{DNA}(a_{i1})+X^k_{DNA}(a_{i2})}\right)$ which measures the total transcript abundance, and $R^k_i = \log\left(\frac{(\frac{X^k_{RNA}(a_{i1})}{X^k_{DNA}(a_{i1})})}{(\frac{X^k_{RNA}(a_{i2})}{X^k_{DNA}(a_{i2})})}\right)$

which measures the fold imbalance between the expression of the two alleles. Normalization with the DNA sample, which, for heterozygous sites, is known to be balanced, normalizes for probe sensitivity and biases.

Values for E and R were collected at 755284 sites. Those sites are not uniformly distributed in the genome, with genic regions (exonic and intronic) having roughly 1.3 times the SNP density as intergenic regions (one SNP per 3.5 kb in genic regions, one SNP per 4.5 kb in intergenic regions). Figure 2.5-1a shows the distribution of E over all genic and intergenic positions. The distribution of expression levels in gene regions is clearly bimodal: a good fraction of genes are not transcribed in LCL, and most but not all intergenic sites are not transcribed. Assuming that 50% of genes and 10% of intergenic sites are expressed, we can deconvolve these distributions to obtain the distribution of E for expressed and non-expressed regions (Figure 2.5-1b). For two individuals, experiments were done in triplicates. As seen in Figure 2.10-1a and b), the technical noise in the measurement of both E and R is quite significant. As expected, R values are particularly noisy at low expression levels.

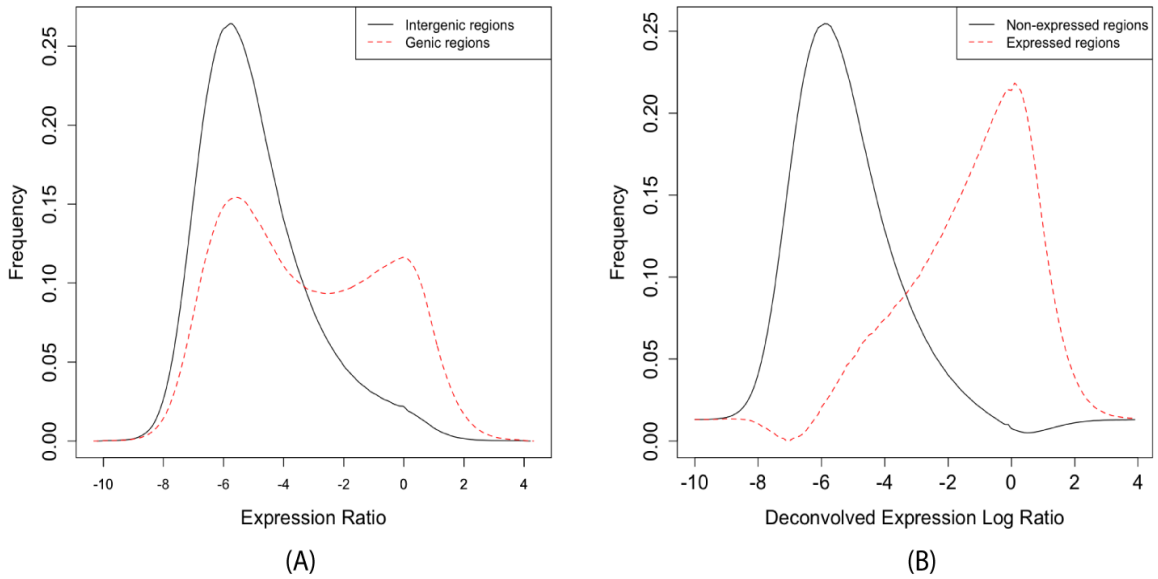

Figure 2.5-1 Distribution of *E* values.

(a) Distribution over genic/intergenic regions (b) deconvolutions to expressed/non-expressed regions.

2.5.2. Identification of transcripts with allelic imbalance

The main problem addressed in this study is the statistically robust identification of genomic regions with significant and consistent allelic imbalance. We start by noting that the data are too noisy for one to accurately call imbalance based on each SNP individually (e.g. by simply using on R_k^i), especially for regions whose expression level is relatively low. We thus consider approaches that take advantage of the fact that most regions with AI are relatively long and are expected to contain more than one SNP. Four main approaches were designed, implemented and compared. Each method aims to robustly assign a score $AI(i)$ to each SNP *i*, so that SNPs that belong to transcripts with significant allelic imbalance obtain large (positive or negative) scores. In all our AI detection algorithms, AI is detected without reference to any kind of gene annotation, contrasting with the annotation-driven approach used by (Ge et al. 2009), which allows us to identify regions of AI whose boundaries does not necessarily correspond to annotated genes. The first three approaches consider data from each sample individually while the last considers data from all samples jointly in order to improve the detection of AI in individual samples. The four approaches considered are first summarized below and then described in detail. The code implementing each algorithm is available at http://www.mcb.mcgill.ca/~blanchem/AI/code.zip.

1) Simple smoothing refers to the approach where the allelic imbalance log-ratio of a SNP is taken as the average of its own log-ratio and that of the m surrounding SNPs on either side.

2) The Z-Score approach involves binning SNPs based on their expression level, assigning each SNP a Z-Score based on its own allelic imbalance ratio, and then determining the Z-Scores of windows of consecutive SNPs and assigning this score to each SNP within the window.

3) The ergodic HMM approach models the AI data in a given individual as being generated by a Hidden Markov Model whose states correspond to different levels of total expression and allelic ratios.

4) The left-to-right HMM approach is an extension of the ergodic model that allows using the AI data from all individuals in order to assess the frequency of AI at each site, and then use those as site-specific priors on the transition probabilities to predict AI regions separately for each individual, but in the context of the data from other individuals.

2.5.3. Simple smoothing approach

Consider heterozygous site i and define window $W(i,m)$ to be the set consisting of m heterozygous sites to the left of i, m heterozygous sites to the right of m, and i itself. The simple smoothing approach estimates:

$$AI^{smoothing}(i) = \sum_{j \in W(i,m)} \frac{R_j}{2m+1}$$

.

Any site i with $\left|AI^{smoothing(i)}\right| > t^{smoothing}$ would then be reported as having imbalance, for some appropriate threshold $t^{smoothing}$. Based on False Discovery Rate assessment (described below), a value of $m = 4$ was determined to be the optimal window size and was used for all results reported.

2.5.4. Z-Score approach

At sites with no allelic imbalance, the value of R_i is modeled adequately using a normal distribution centered at 0. However, the variance is inversely correlated with the total expression E_i, as AI is difficult to estimate when the total expression is low (see Figure 2.10-1b). The range of possible values of E are subdivided into 100 bins of equal size and the mean μ_b and

variance σ_b^2 of R values were determined for SNPs belonging to every expression level bin b. A site-specific Z-Score $Z(i)$ is assigned to heterozygous site i as $Z(i) = (R_i - \mu_{bin(E_i)})/\sigma_{bin(E_i)}$. Homozygous sites, being uninformative with respect to allelic ratios, are excluded from the analysis. Consider now a collection of w consecutive heterozygous (ignoring possibly intervening homozygous sites) SNPs $i_1, i_2, \ldots, i_w$. We define the regional Z-score as $Z(i_1, i_2, \ldots, i_w) = \frac{\sum_{k=1\ldots w} Z(i_k)}{\sqrt{w}}$. Assuming the normality of noise R_i in measurements, $Z(i_1, i_2, \ldots, i_w)$ follows a Normal(0,1) distribution under the null hypothesis of absence of allelic imbalance.

Regional Z-Scores are first computed for every possible window of $w = 1 \ldots 50$ heterozygous sites. The region with the highest regional Z-score (in absolute value), Z^{max} is selected first and we set $AI^{zscore}(i) = Z^{max}$ for all sites heterozygous i within the region. This region is then masked out and the next highest scoring non-overlapping window is selected. The process is repeated until all heterozygous sites have a Z-Score assigned. We note that because the $AI^{zscore}(i)$ is obtained based on the *best* window that contains site i, there is a complex issue of multiple hypothesis testing that results in this measure not following a Normal(0,1) distribution under the null hypothesis (i.e. absence of AI). In consequence, one cannot easily translate $AI^{zscore}(i)$ into a p-value.

We also considered a variant of the Z-Score approach where each SNP is assigned the Z-Score of the *fixed-size* window centered around it. This approach, which can be seen as an improved version of our simple smoothing approach, indeed improves on the latter (based on permutation testing and comparison to transcripts with known AI - see below), but is far from being as accurate as the proposed Z-Score approach, because it leads to bleeding edges at transcript boundaries. We also investigated a version of the Z-Score approach where SNPs are not binned by expression level prior to Z-Score computation; this resulted in a small but significant decrease in accuracy, showing that the appropriate modeling of the dependency between the noise in allelic ratio and the total expression level is an important feature of our approach.

2.5.5. Simple sample ergodic hidden Markov model approach

The linear nature of the data in question lends itself well to a Hidden Markov Model (HMM) in which each data point corresponds to a particular SNP, the hidden states correspond to qualitative descriptions of the allelic imbalance (e.g. positive imbalance, negative imbalance, no imbalance), and emissions correspond to the total expression E_i and the allelic log-ratio R_i observed at site i.

We built an HMM consisting of a total of eight hidden states (see Figure 2.5-2a). Seven of these states correspond to SNPs belonging to expressed transcripts in the LCL sample in question, with various levels of imbalance: $S = \{S_{+++}, S_{++}, S_{+}, S_0, S_{-}, S_{--}, S_{---}\}$, corresponding to strongly positive imbalance (S_{+++}), moderately positive imbalance (S_{++}), slightly positive imbalance (S_{+}), balance (S_0), slightly negative imbalance (S_{-}), moderately negative imbalance (S_{--}) and strongly negative imbalance (S_{---}). There is also a state (S_N) that corresponds to SNPs located in regions that are predicted not to be transcribed, and for which allelic imbalance is meaningless. The emission probability for each state $s \in S$ is modeled with a pair of normal distributions for the E and R values, with parameters ($\mu_{E,s}$, $\sigma^2_{E,s}$), and ($\mu_{R,s}$, and $\sigma^2_{R,s}$) respectively. Whereas both total expression E and allelic imbalance measurements R are observed at heterozygous sites, only the expression is measured at homozygous sites. In the latter case, the imbalance data is left unobserved (i.e. all 8 states are equally likely to have generated the R observation). Homozygous SNPs can thus be included in the model training and predictions, and can help delineating regions of based on expression levels.

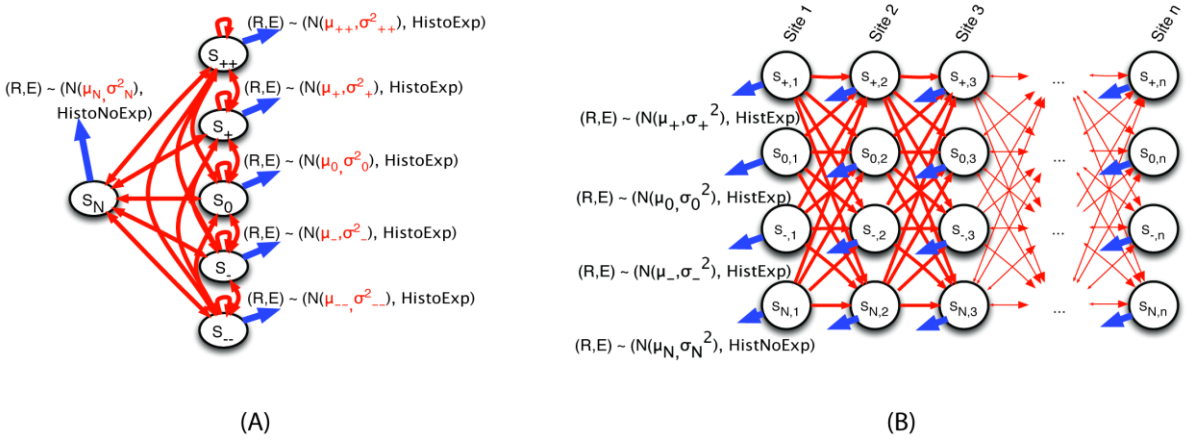

Figure 2.5-2 Architecture of the two Hidden Markov Models used in this study.

(a) Ergodic HMM architecture. HistoExp and HistoNoExp refer to the distributions depicted in Figure 2.5-1b. For readability, states S_{+++} and S_{---} are not shown. (b) Multi-sample left-to-right HMM architecture. States S_{+++}, S_{++}, S_{---}, and S_{--} are not shown for clarity. Only transition probabilities are trained. All copies of a given state have the same emission probability distribution, described on their left.

An HMM with a realistic correspondence to the data can in principle be built with $2K + 2$ states, where $K \geq 1$ represents the number of levels of positive (and negative) imbalance that the model represents. Larger values of K should in principle be favorable as they allow a finer discretization of allelic ratios. Models with $K \in [1,2,3,4]$ were trained and the false discovery rate measured and compared (see section 2.5.8) It was found that $K = 3$ performed better than $K = 1$ and $K = 2$, and similarly to $K = 4$ (Figure 2.10-2), so this value was used for both the ergodic and left-to-right models.

Certain parameters of the HMM are trained using the Baum-Welch algorithm, while others are fixed. For S_N, the emission probability distribution for E is modeled non-parametrically by the histogram of Figure 2.5-1b (black curve) whereas all expressing states share the same total expression distribution

from Figure 2.5-1b (red curve). These emission probability distributions are kept constant during the training procedure. The Baum-Welch algorithm (Baum et al. 1970) is used to find maximum likelihood estimators for $\mu_{R,s}$ and $\sigma^2_{R,s}$, for $s \in S$, as well as all transition probabilities and the initial state probability. The Baum-Welch algorithm is an expectation-maximization (EM) (Dempster et al. 1977) approach that alternates between the Expectation step (or E-step), in which the posterior probability over states is computed for each site using the Forward-Backward algorithm, and the Maximization step (or M-Step) where the parameters of the emission and transition probability distributions are adjusted to best reflect the observed data given these posterior probabilities. Formulas for updating the emission probability parameters and transition probabilities are adapted straightforwardly from (Mitchell 1997). We considered training one HMM per individual (which would allow the flexibility to model inter-experiment variation in noise, for example), or to train a single HMM based on the data from all individuals (which would have the benefit of being based on more data). The latter option produced slightly better results and this is the strategy we used for the rest of the study. We also considered filtering out sites with low total expression, as their allelic expression ratio may be less reliable. However, slightly better results were obtained without any filtering (allowing non-expressed SNPs to naturally be classified as belonging to state S_N). Training on the whole data set took less than Baum-Welch 20 iterations and 3 hours to converge on a standard desktop computer (convergence is defined as two consecutive iterations where no parameter or transition probability changed by more than 10^{-5} or 1% of their value). Restarts from different initial values converged to nearly the same values.

The Viterbi algorithm (Viterbi 1967) can then be used to identify, in each individual, predicted regions of different levels of positive or negative imbalance. The Forward-Backward algorithm (Rabiner 1989) yields an estimate of the posterior probability of each state at each site. In the latter case, a useful

summary score for each site is the posterior expected allelic expression log-ratio, which we use as an AI predictor: $AI^{ergodic}(i) = \sum_{s \in S} \Pr[S_i = s \mid E_{1\ldots n}, R_{1\ldots n}] \cdot \mu_s$.

Until now we have assumed homogenous transition probabilities, regardless of the distance in base pairs between consecutive SNPs along the chromosome. However, a more accurate model would factor in the distance between neighboring SNPs, to increase the probability of self-loops (i.e. staying in the same state) when the two sites are nearby but increase the probability of state change for two distant sites. Such an approach has been used previously in HMMs designed to detect CNVs (Colella et al. 2007). We obtained a unit transition probability matrix T as the d-th root of the transition matrix obtained via Baum-Welch training of the homogeneous model, where d is the average distance (in base pairs) between two consecutive SNPs in our data. Then, the transition probability matrix used for a pair of sites separated by l base pairs will be T^l, which is efficiently computed using the eigenvalue decomposition of T.

To ensure that our training procedure was not subject to overfitting, we used 2-fold cross validation (dividing the 53 samples into one 26-sample data set and one 27-samples data set) and trained our 8-state ergodic HMM separately on each half the samples. The parameters and transition probabilities obtained were nearly identical, and so were the FDR estimates obtained by running each HMM on the complementary data set, indicating that overfitting is not an issue.

2.5.6. Multi-sample left-to-right HMM approach

The previous HMM is called ergodic because it models an ergodic, homogeneous Markov chain over the state space (i.e. the set of transition probabilities is independent of the position along the genome). One limitation of this HMM is that it does not take full advantage of the fact that data exists for multiple individuals and that, while not all individuals are expected to have AI in exactly the same regions, one does expect AI hotspots where a significant fraction of the individuals would have imbalance. That would be the case, for

example, for genes where one allele is commonly or always silenced via epigenetic mechanisms, or when AI is due to a common regulatory variant. The approach proposed in this section aims at predicting AI regions separately in each individual, while taking into consideration the data observed in *all* individuals. In doing so, we still want to be able to identify AI regions that are unique to a given individual, but are hoping to improve the detection of regions with common AI. For example, AI regions containing only a few SNPs, or those where the imbalance is only moderate, may be missed when present in a single individual, but may be detectable if present in a large fraction of the population. In addition, we may be able to detect boundaries of AI regions with more accuracy when they are shared among individuals.

The approach utilized to address this is termed the *left-to-right HMM* (Rabiner 1989) (see Figure 2.5-2b), similar to profile HMMs (Eddy 1998). Each site has its own copy of the set of states and transitions can only occur between states associated with neighboring sites, from left to right. Each copy of a given state shares the same emission probability distributions that are modeled the same way as with the ergodic HMM. However, transition probabilities will vary across positions, making the model non-homogeneous (in contrast to our ergodic HMM approach). This configuration allows for greater fine-tuning at the level of each individual SNP or region, though at the cost of a substantially larger set of transition probabilities to be learned.

The training of our left-to-right HMM is a two stage process. In the first stage, emission probabilities, transition probabilities, and start probabilities are estimated for the ergodic version of the HMM using the Baum-Welch algorithm described above, using all available individuals. The parameters of the emission probabilities of the states in the left-to-right HMM will be set to those obtained on the ergodic training and will not be re-estimated. The obtained ergodic non-homogeneous distance-corrected transition probabilities will be used as prior for those of the left-to-right HMM.

In the second stage, we now switch to learning the transition probabilities of the left-to-right HMM. We assume that the data set from each individual is the result of an independent run of the HMM:

$\Pr\left((E^1,R^1),(E^2,R^2),\ldots,(E^k,R^k)\middle|HMM\right) = \prod_{i=1\ldots k}\Pr(E^i,R^i|HMM)$, and we seek to identify the set of transition probabilities of the left-to-right HMM that maximizes this joint likelihood. Consider a site i that is not imbalanced in any individual but where site $i+1$ is positively imbalanced in a large fraction of the individuals. The maximum likelihood estimator for the transition from state $S_0(i)$ to state $S_+(i+1)$ will be higher than at other positions where few individuals enter an imbalanced region. Now consider an individual where there is only weak evidence of AI starting at position $i+1$. When using an ergodic HMM for our predictions, the weak AI region will probably not be detected. However, in the left-to-right HMM, with the increased transition probability, the AI path becomes more likely, so provided that there is sufficient imbalance, the most likely path may now to go through one of the imbalanced state.

Estimating transition probabilities between two sites separated by l base pairs is done using a simple modification to the standard Baum-Welch algorithm, where the update rule for transitions is: $t'_{i,i+1}(a,b) = \frac{\sum_{j=1\ldots k}\left(\Pr\left(S_i^j=a,S_{i+1}^j=b\right)\right)+W\cdot T^l(a,b)}{\sum_{j=1\ldots k}(\Pr(S_i^j=a))+W}$ where T^l is the l-th power of the unit transition probability obtained previously and W indicates the pseudocount weight described in the following paragraph. The regularization obtained by using the ergodic transition probability as prior reduces the risks of overfitting while improving the convergence of the training procedure. In practice, based upon permutation tests and resulting FDR scores, a parameter of $W=1$ was determined to be optimal (data not shown).

Once the left-to-right HMM is trained using the data from all 53 individuals (which took 161 Baum-Welch iterations - less than 4 hours on a standard desktop computer), the standard Viterbi or Forward-Backward algorithms are

used to identify AI regions separately for each individual. As with the case of the ergodic HMM, we use the posterior expected allelic expression log-ratio $AI^{LtoR}(i)$ to summarize AI evidence at SNP i.

Overfitting is a possible issue with our left-to-right HMM, as the number of parameters estimated is much larger than for the ergodic HMM. We performed 5-fold cross-validation, training on 4/5 of the data and predicting on 1/5. Thanks to our regularization procedure, the predictions obtained were very similar to those obtained by training and testing on the full data set, with only a marginal decrease in FDR.

2.5.7. Cross-Hybridization

Upon study of some of the regions where AI was predicted in most or all individuals but where not known imprinted regions existed, we found that nearly half were a likely artifact of cross-hybridization. All these suspicious regions were the results of a segmental duplication, where a fragment of a gene was duplicated. Because the fragments still matched the genic region, sites within them will appear to be expressed (as they match the transcript of the paralogous region), and polymorphisms will cause mismatches between the probe and the true transcript, which will result in apparent AI. We thus used the human Blastz self-alignment from the UCSC Genome Browser (Kent et al. 2002; Kent et al. 2003) to filter out regions corresponding to recent duplications. A possible alternate approach would consist of using the results of the genomic DNA hybridization to identify probes that match more that one location in the genome, with the possible added benefit of detecting DNA possible copy-number variation.

2.5.8. False-Discovery Rate Estimation

Due to the relatively small number of “gold standard” regions known to exhibit AI, the best available option for comparison of the various models is through permutation tests. The goal was to preserve some of the structure of the

genome such that only SNPs with approximately equal expression levels and heterozygosity would be swapped, i.e., the only factor that is swapped freely is that of the allelic imbalance ratio. Permuted data sets were generated as follows. Sites were partitioned into five levels based on the number of individuals in which they are heterozygous. Five bins were also assigned based on the average level of expression seen across all individuals. Each SNP was then finally assigned to one of 25 bins, with one bin for each of the possible combinations of heterozygosity frequency and expression levels. Sites were randomly permuted within each bin, preserving the correspondence between sites in different individuals (in the case of the left-to-right HMM, the first stage of training of global HMM parameters was first done on non-permuted data, and then the second stage of model training was done on permuted data). Preserving expression levels and heterozygosity is important to create permuted data sets that are as realistic as possible, in particular with respect to the fact that expressed sites are found in contiguous genomic regions rather than dispersed randomly in the genome.

Each of the prediction methods described produces one AI score per site and per individual. For each method M, the number of regions of consecutive SNPs exceeding a given score threshold t, $N_{real}(t, M)$ and $N_{perm}(t, M)$ was determined in the real and permuted data, resulting in a False-Discovery Rate of $FDR(t, M) = \frac{N_{perm}(t,M)}{N_{real}(t,M)}$

2.6. Results

Each of our four approaches was applied to the data set and the AI predictions for each individual are available at http://www.mcb.mcgill.ca/~blanchem/AI/AIPredictions.zip.

2.6.1. Illustrative Case Studies

We use two examples to highlight the features of the data and the methods developed. Figure 2.6-1 gives a sample of the raw data and predictions made by each method in the BLK locus. BLK is a gene that has previously been described as allelically imbalanced in LCL (Ge et al. 2009). Interestingly, in this individual, two other neighbouring genes have strong allelic imbalance, with FAM167A showing expression on the opposite allele compared to BLK and GATA4 also obtaining strong and consistent signals. Although in this example the boundaries of allelic expression domains align nicely with known gene boundaries, this is not the case in general. As is obvious from the figure, the raw expression and allelic ratio data are quite noisy. The simple smoothing approach succeeds at identifying the main regions of allelic imbalance but does so much less reliably and precisely than the other three approaches. Notice that this individual has no heterozygous sites in the 5′ end of FAM167A. This results in different behaviours for each method. The ergodic approach assigns gradually decreasing expected allelic log-ratios in that region, while the Z-Score approach only predicts imbalance in the 3′ end of the gene. However, the left-to-right HMM has the benefit of considering data from other individuals, which have some heterozygous sites in the 5′ region of the gene, which allows it to predict strong and consistent negative allelic log-ratios over the whole gene, and a sharp transition entering the BLK transcript. A similar phenomenon is observed for GATA4.

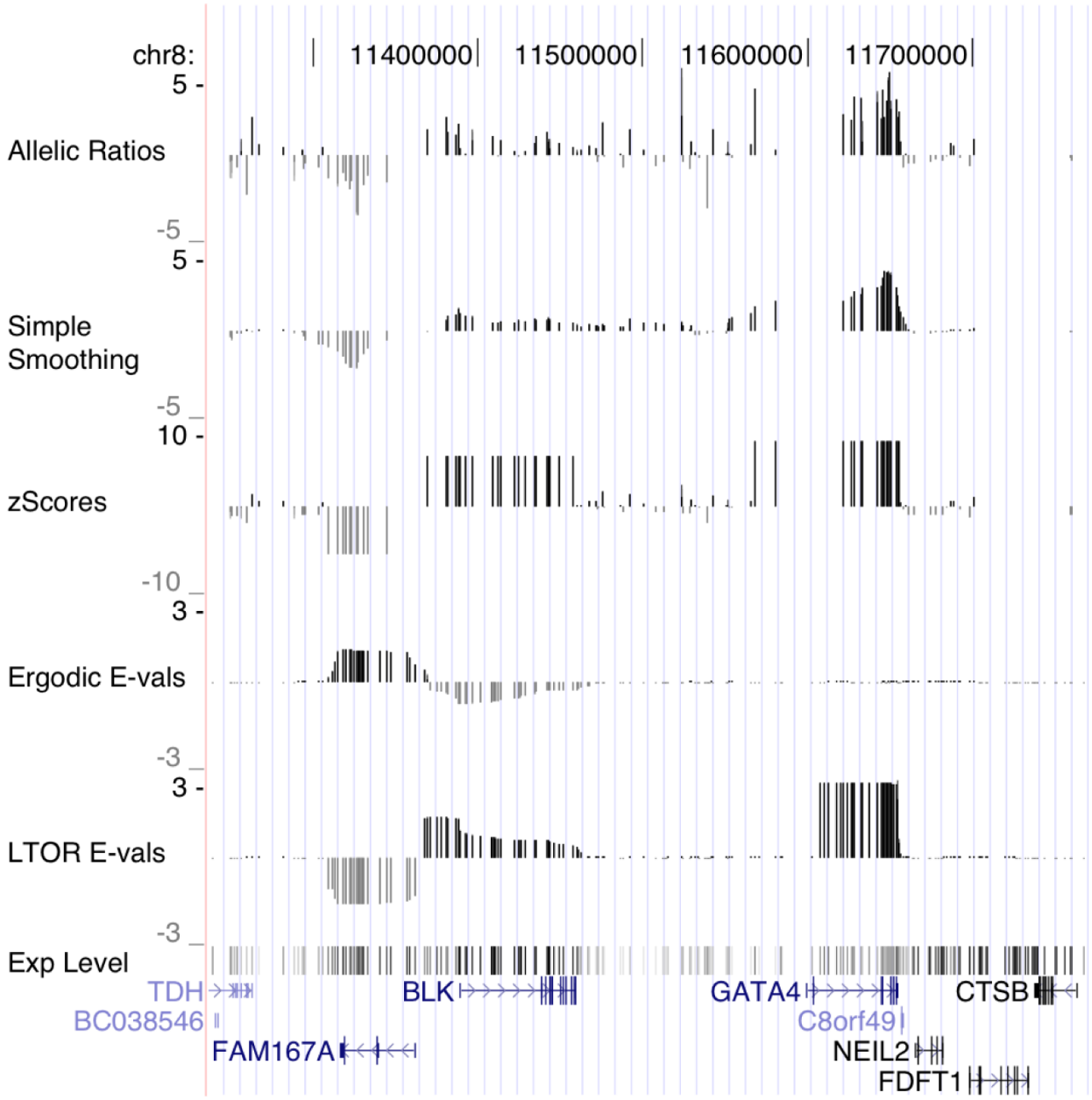

Figure 2.6-1 Raw data and predictions.

Example of genomic region with allelic imbalance. From top to bottom: Raw allelic log-ratio; Simple smoothing predictions; Z-score predictions; Ergodic 8-state predictions (expected allele log-ratio); Left-to-right 8-state HMM predictions (expected allele log-ratio); Raw total expression; UCSC known genes track. Data shown is for HapMap individual NA11840. Note: Allelic ratios at homozygous sites are not shown.

Figure 2.6-2 shows the set of predictions made by the Viterbi algorithm using the left-to-right HMM on the extended GATA3 locus, in all 53 samples. The region exhibits a large diversity of patterns of AI. In some cases, the region of AI closely matches an annotated gene (e.g. SFTMBT2 in several individuals). Often, AI regions do not overlap any known gene (e.g. the region located upstream of SFMBT2). Such regions, especially when they abut an annotated gene, may reflect the presence of alternative allele-dependent promoters. They may also represent completely novel unannotated transcripts. Another frequently observed pattern is the presence of AI within annotated transcripts, near the 5′ or 3′ end (e.g. the 3′ end of the ITIH5 gene). Finally, AI regions often encompass one or more complete genes (e.g. GATA3 and NM_207423), possibly because of epigenetic modification of one of the two alleles. We note based on analysis done in (Ge et al. 2009) that SFTMBT2 and ITIH5 show evidence of heritable allelic expression, whereas GATA3 does not show correlation with common genetic variants and could represent epigenetic modification of expression in LCLs.

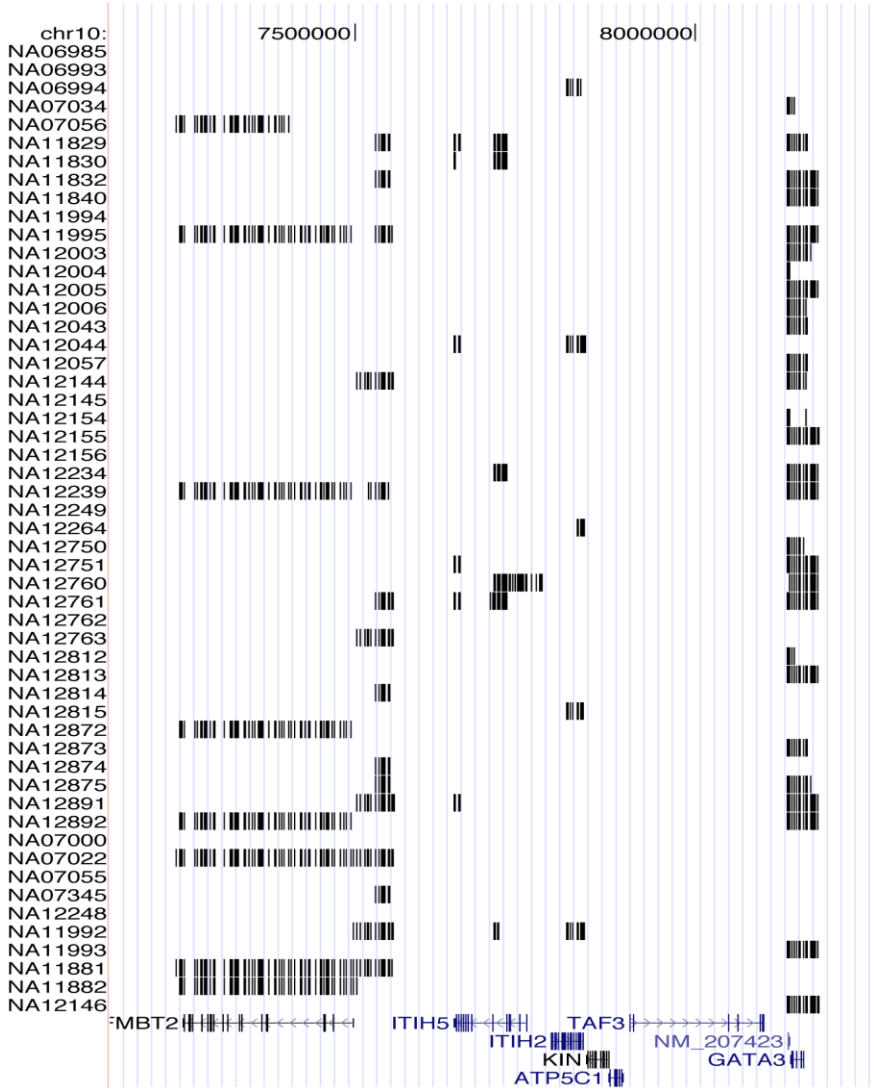

Figure 2.6-2 Allelic imbalance in 53 HapMap individuals in the GATA3 locus.

Each row reports the sites where AI has been predicted by the 8-state left-to-right HMM with the Viterbi algorithm. Each AI SNP is marked with a vertical black line; the impression of gray levels is an artifact of SNP density. Genes from RefSeq (Pruitt et al. 2005) are illustrated below.

2.6.2. Evaluation and Validation

The accuracy of the AI predictions made by each method was evaluated using both permutation testing (in order to assess the false discovery rate) and comparison to previously characterized AI transcripts.

Permutation Testing

We first estimated the false-discovery rate (FDR) of each method using a permutation test where genomic sites are randomly permuted, subject to some constraints (preservation of heterozygosity and expression level; see Methods). This randomized data set preserves the level of imbalance observed at each site, but randomly disperses sites in such a way that few regions are expected to exhibit strong and consistent allelic ratios over several consecutive sites (as real AI transcripts should). For each algorithm, the number of genomic regions with AI score above some threshold t in the real data was compared to the corresponding number on the permuted data - the ratio of these two numbers is an estimate of the FDR of the algorithm (note that the FDR could also be estimated at the individual SNP level, rather than at the region level; the conclusions are the same). Figure 2.6-3 shows the FDR curves obtained for each method, as a function of the number of predictions made. All methods are able to detect the most obvious cases of AI (roughly 200 regions per individual, where all methods have near-zero FDR). However, as our threshold decreases and the number of regions predicted increases, the performance of the four approaches become quite different. Setting 5% as an acceptable FDR, the simple smoothing, Z-Score, ergodic HMM, and left-to-right HMMs result in 360, 622, 662, and 954 predicted regions with AI. In other words, at that FDR level, the best approach, left-to-right HMM, is ~160% more sensitive than the simple smoothing approach and ~45% more sensitive than the second best approach, which is the ergodic HMM. Similar observations hold for other FDR thresholds. Therefore, the information obtained from the total expression levels, as well as the added site-specific transition probabilities are beneficial in terms of obtaining reliable AI predictions. This is particularly noteworthy for regions whose AI is weaker (those

ranking between the 500 to 1000th per individual), for which the FDR remains quite low with the left-to-right HMM but quickly increases with all other methods.

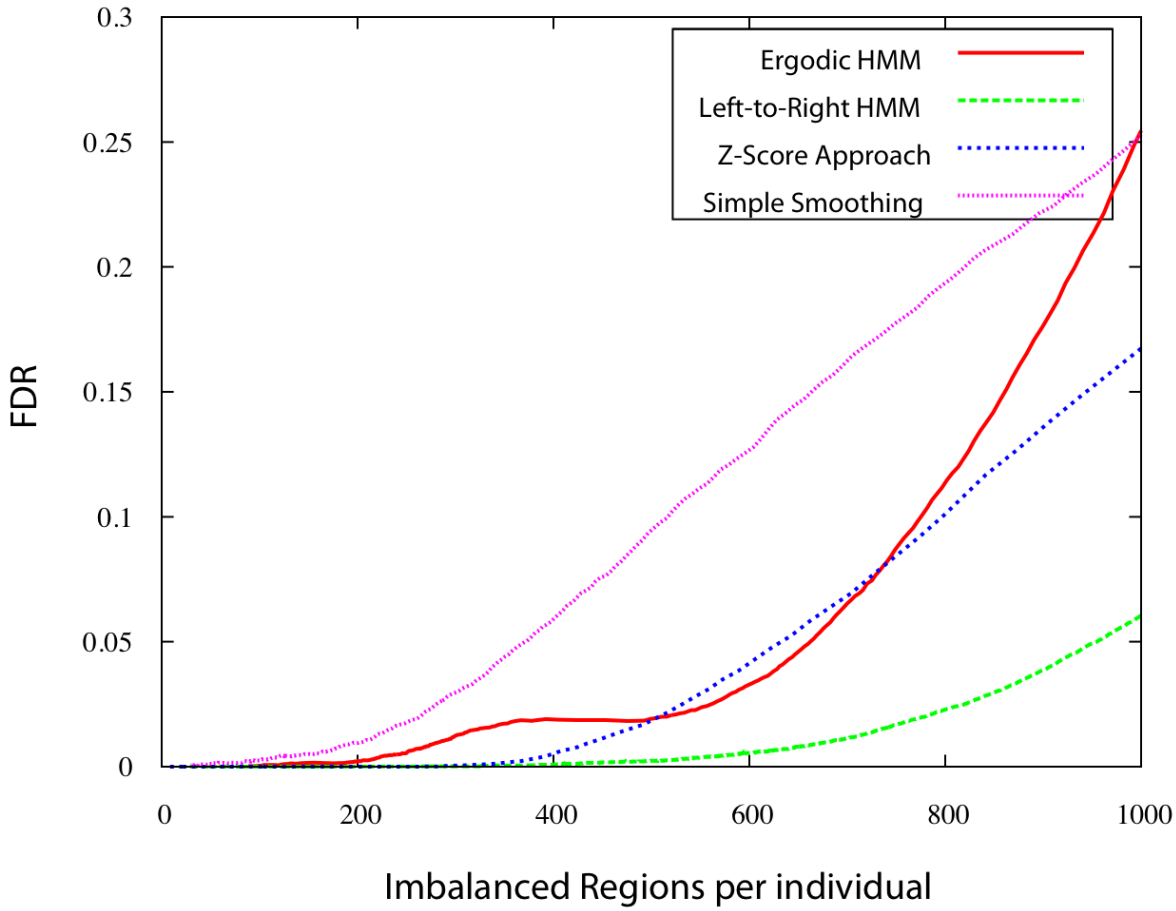

Figure 2.6-3 False Discovery Rates (FDR).

FDR obtained by permutation testing at thresholds resulting in different numbers of AI regions being predicted.

Comparison to Known AI Transcripts

Although no comprehensive set of validated AI transcripts exists to date, a set of 62 imprinted genes (containing 1099 SNPs in our data set) have been collected from the literature and posted on www.geneimprint.com. Most imprinted regions are easily detected by most methods, as they affect relatively large genomic regions and their allelic expression ratios are extremely large. Figure 2.6-4 shows how the enrichment of the overlap between imprinted genes and the

number of predictions made by each of the four methods varies as a function of the number of sites being predicted with AI. (The enrichment of the overlap between a set of predicted AI regions and a set of annotated regions is the ratio of the size of the overlap to the expected size of the overlap if AI regions had been selected randomly in the genome.) Imprinted SNPs are enriched 5 to 20-fold among the top predictions made by each algorithm (except the Z-Score approach, which assigns high scores to other types of regions). Focussing on the left-to-right HMM AI predictions at a 5% FDR threshold (which consist of roughly 40,000 SNPs per individual), we find that 67% (resp. 35%) of SNPs in imprinted regions are predicted to have AI in at least one (resp. five) individual. Manual inspection of imprinted genes that have gone undetected by any of our methods reveals genes that are short, contain few heterozygous SNPs, or are expressed at very low levels in LCL.

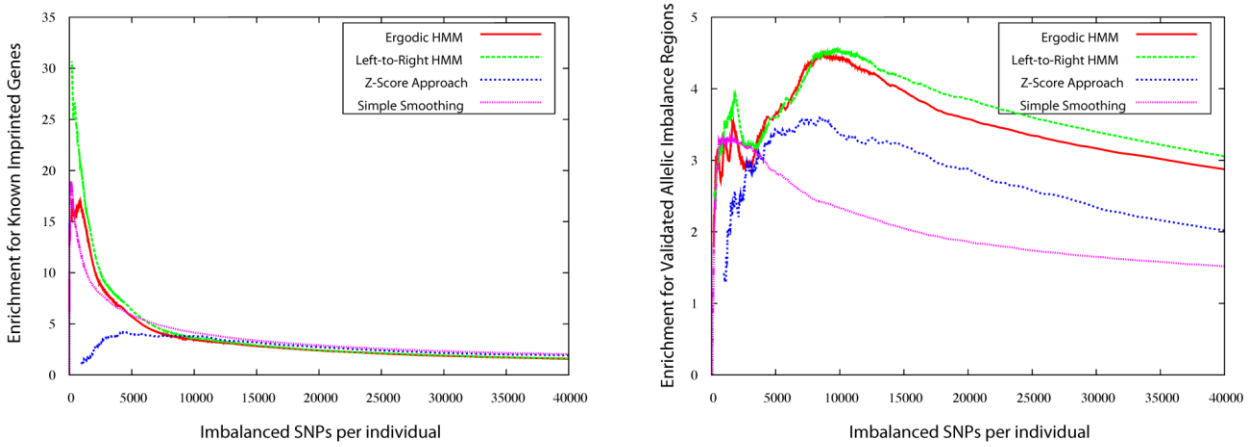

Figure 2.6-4 Enrichment for SNPs called as allelically imbalanced in imprinted and AI genes.

(a) Overlap with regions experimentally verified to be imprinted. (b) Overlap with experimentally validated imbalanced genes from (Verlaan et al. 2009).

Allelic imbalance resulting from cis-regulatory variation typically have allele ratios less extreme than imprinted genes and are thus more difficult to detect. A set of 61 transcripts (containing 1596 SNPs in our data set) with AI resulting from cis-regulatory variation in LCL have been identified and validated by (Verlaan et al. 2009). Figure 2.6-4b shows the fold-enrichment of these SNPs among those predicted as AI SNPs by each of our methods. Here, the predictions made by the two types of HMMs perform significantly better than the Z-Score and smoothing approaches, detecting approximately 50% and 100% more validated SNPs. Overall, our best approach is again the left-to-right HMM, which predicts 87% (resp. 70%) of the 1596 validated SNPS as imbalanced in at least one (resp. five) individual(s). Inspection of AI genes that were undetected showed that they exhibited little evidence of allelic imbalance by our method (see Figure 2.10-3). These represent likely false positives in the earlier study as well as more localized effects caused by few independent AI measurements and driving the association tests in previous analyses (Ge et al. 2009).

2.6.3. Distribution of AI in the Genome and Across Individuals

Our predictions allow a first glimpse into the diversity of allelic expression patterns in the human genome, although a comprehensive analysis of AI regions is beyond the scope of this study. We first observe that AI in LCL samples is widespread, with on average 9.7% (resp. 5.6%) of an individual's genes containing at least one (resp. all) imbalanced SNP (using the left-to-right HMM with a threshold corresponding to an FDR of 5%). Considered in total, 54.4% of genes show at least one imbalanced SNP in at least one individual, and 45.6% of genes have all of their SNPs showing allelic imbalance in at least one individual. Note that only approximately 50% of genes in total are detectably expressed in LCL (Cheung et al. 2003), and are hence candidates for being allelically imbalanced. Thus, the majority of expressed genes show AI in one or more individuals.

Figure 2.6-5 reports the distribution of AI regions across various types of genomic regions. While a substantial fraction (19%) of AI regions closely match annotated gene boundaries, most exhibit more complex relationships to annotated protein-coding gene transcripts, a larger portion of AI regions (28%) are within annotated genes but cover only a fraction of the transcript. In nearly half of those, allelic expression is found toward the 3′ end of the gene, possibly because of allele-specific transcription termination or mRNA degradation, or the presence of an allele-specific alternate transcription start site within the annotated gene. The presence of AI regions at the 5′ end of the transcript appears somewhat less frequent. 22% have little or no overlap with protein-coding genes, although this fraction is enriched for other types of transcripts such as LINC-RNAs (Khalil et al. 2009).

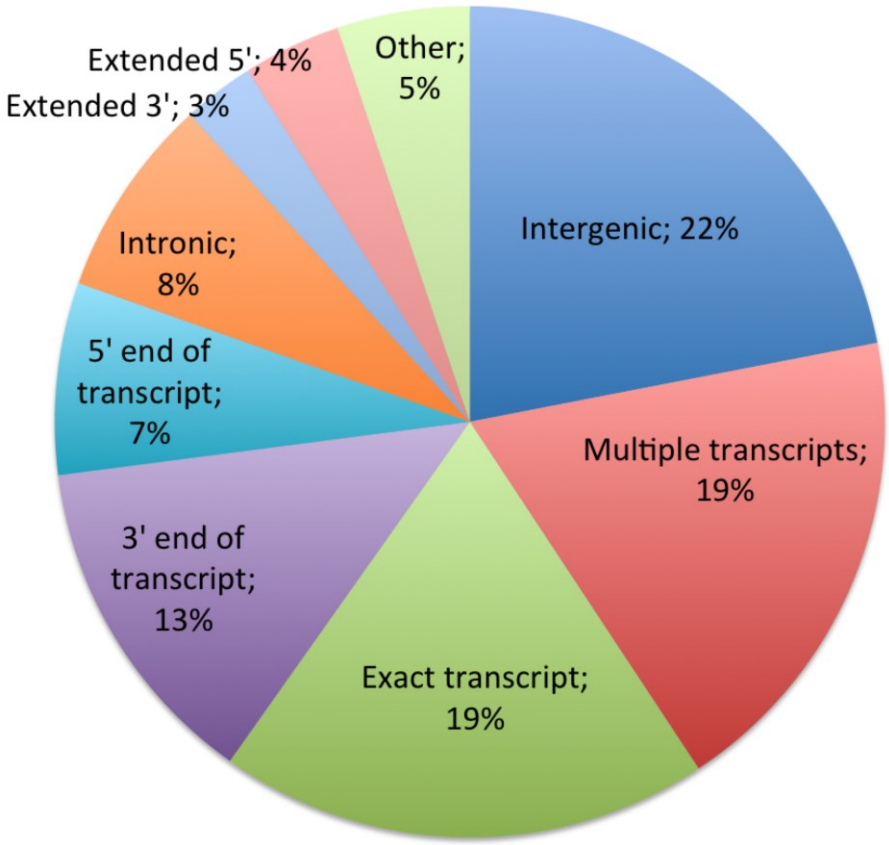

Figure 2.6-5 Classification of AI regions based on their overlap with annotated protein-coding genes.

The classification of an AI region is done based on a set of simple rules that allow for a sizable margin of error in the boundaries of the AI regions. Intergenic: Little or no overlap with annotated genes. Multiple transcripts: Overlaps several genes. Exact transcript: The left and right boundaries of the AI region match gene boundaries within 20 kb. 5′ (resp. 3′) end of transcript: AI region is at the 5′ end (resp. 3′ end) of the gene only. Intronic: AI region is within the gene but away from the gene boundaries. Extended 5′ (resp 3′): AI region extends upstream (resp. downstream) of the gene.

Our data set affords a first glimpse into the commonality of allelic imbalance at a given site across individuals. We calculated the number of individuals showing AI (based on the Viterbi predictions; see Figure 2.6-6). The

very long tail of this distribution indicates that a lot of AI is shared among a portion of the population. In fact, ~65% of an individual's AI regions are found in at least 10 other individuals. Allelic imbalance, whether caused by genetic or epigenetic causes, is thus highly structured in the human population. On the other hand, rare AI, defined as that seen in at most 10% of our individuals, constitutes approximately 20% of an individual's AI regions, while 4% are unique to that individual. We note however that because AI regions found in a large number of samples are easier to detect than those that are less common in the population, we may underestimate the proportion of AI that is found in a small number of individuals. We note that the left-to-right HMM predictions used for this analysis are potentially biased towards over-predicting sites with common AI and under-predicting those with rare AI. We thus repeated the analysis with the ergodic HMM approach, which does not suffer from this bias. The results were very similar, with only a very slight shift toward less frequent AI.

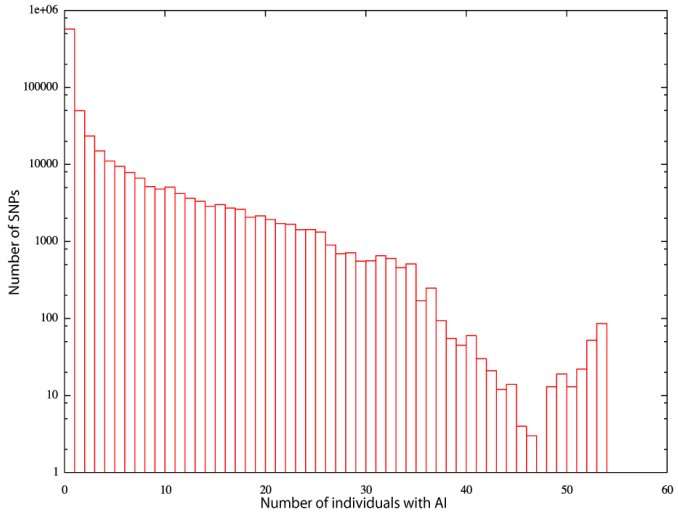

Figure 2.6-6 Commonality of allelic imbalance.

Number of SNPs in AI regions, as a function of the number of individuals with AI at the same site.

2.7. Discussion

The recent development of a genome-wide high-density assay of allelic imbalance based on genotyping arrays has resulted in a vast improvement in our understanding of this type of variation and in our ability to map this variation to causative regulatory SNPs (Ge et al. 2009). A relatively simple gene-based analysis was sufficient to identify a significant number of genes with allelic imbalance (Ge et al. 2009). However, taking full advantage of this technology requires advanced signal processing approaches to accurately detect, delineate and quantify allelic expression. Furthermore, relying too heavily on known gene annotation may hide the fact that most AI does not perfectly align with gene boundaries. Indeed, the approaches proposed here, which do not make use of gene annotations, reveal that allelic imbalance is widespread and exhibits complex patterns in relation to annotated genes. Although our approach was specifically applied to the analysis of data obtained from high-density genotyping arrays, it should be readily applicable to studies based on data obtained next generation RNA sequencing.

Detection of AI based on data from genotyping arrays proves challenging because of the significant noise in the allelic ratio measured at individual SNPs and because of the complex patterns of AI. To our knowledge, our study represents the first in-depth, statistical and computational analysis of a large scale, genome-wide allelic imbalance data set. Because of the noise level in allelic expression ratios at individual SNPs, one must rely on the fact that transcripts with allelic imbalance will generally contain several SNPs that are expected to show imbalance. Our Z-Score approach identifies regions where the allele ratio is significantly different from the expected one-to-one ratio. An aspect of the data that is not exploited by the Z-Score approach is that the total expression and allelic ratio are expected to be consistent across the transcript. Our two HMM approaches model this explicitly, and obtain better results in part because of this. An additional improvement in accuracy of AI detection is obtained by our left-to-right HMM, which considers jointly the data from all

individuals to serve as prior for the detection of AI in each one. This approach yields improved detection of AI regions that are shared among many individuals, while being able to detect those present in only one or a few samples. This relatively new type of machine learning problem, where a collection of sequences of observations are expected to have been derived from a common (but unknown) model but where each individual can significantly deviate from that model is a situation that may arise in a number of other situations where our left-to-right HMM approach may be useful, including for comparative genomics based gene predictions (Siepel et al. 2007) (where different species are expected to share some but not all of their exon structure).

Although a detailed biological analysis of allelic imbalance and its phenotypic consequences is beyond the scope of this paper, our predictions reveal that AI is widespread, with roughly 10% of genes showing evidence of AI in a given individual, and with the majority of genes expressed in LCLs showing AI in at least one of our 53 samples. Although roughly 60% of AI regions are clearly related to an annotated transcript, they often reflect the presence of alternative promoters, splicing, or transcription termination.

An increasing proportion of the genetic burden of disease is being associated with differences in gene regulation (Cookson et al. 2009). At the same time greater complexity of gene regulation and the transcriptome are being uncovered (Birney et al. 2007). Therefore, hypothesis-free methods for detecting allelic imbalance are a prerequisite to advancing our understanding of population variation in cis-regulatory control by heritable or epigenetic mechanisms.

2.8. Acknowledgments

We thank Javad Sadri for useful discussions, as well as three anonymous reviewers for their suggestions.

2.9. Author Contributions

Conceived and designed the experiments: JRW TP MB. Performed the experiments: JRW. Analyzed the data: JRW. Contributed reagents/materials/analysis tools: BG DP KLG TP. Wrote the paper: JRW MB.

2.10. Supplementary Figures

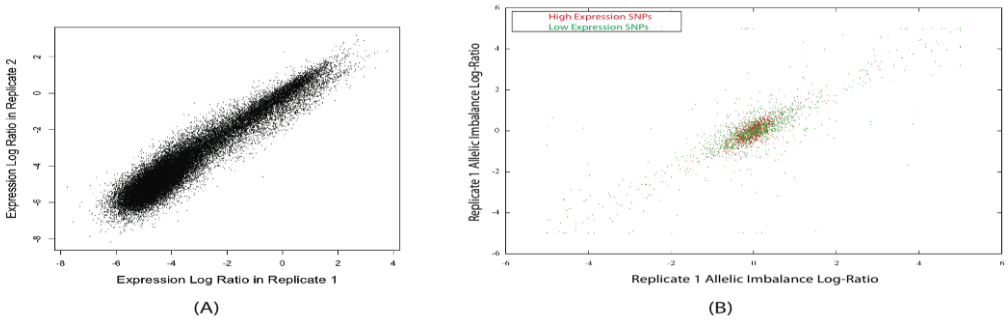

Figure 2.10-1 Analysis of the noise using technical replicates.

(a) Replicability of expression value *E*. (b) Replicability of allelic ratio *R*.

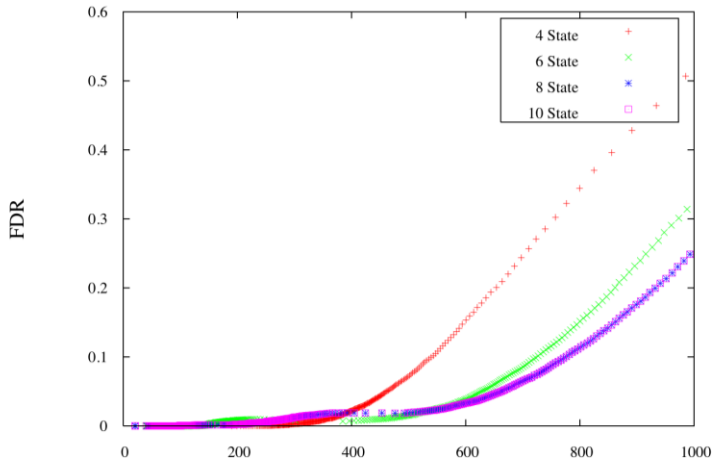

Figure 2.10-2 Performance of ergodic HMM with different levels of discretization.

False-discovery rate obtained by ergodic HMMs with 4, 6, 8, and 10 states (corresponding to 1, 2, 3 and 4 levels of positive and negative allelic imbalance).

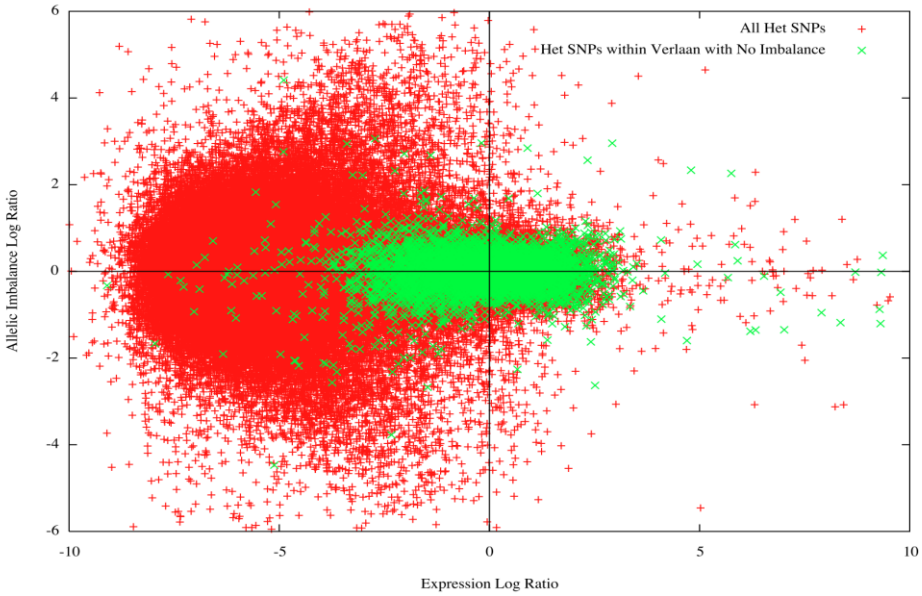

Figure 2.10-3 Analysis of AI data in false-negative regions.

Red: Genome-wide distribution of AI measurements (total expression vs allelic ratio). Green: AI measurements in genes identified as imbalanced by (Verlaan et al. 2009) but not predicted as such by our approach. These genes show no sign of imbalance in our data.

Chapter 3. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts

3.1. Preface

High-throughput platforms for measuring marks of gene expression, DNA sequence variation and epigenetic features continue to advance at an impressive rate. In the period of time research for this chapter was carried out (2012-13) and years prior, work showing links between these various features continued to be published, using ever higher sample sizes, number of loci interrogated, and more precise study definitions. Some of these works are outlined in Chapter 1. While not necessarily of the most impressive magnitude in terms of the number of individuals studied, our results represented an advance in terms of the number of relationships considered in our study with a primary, untransformed cell line (skin fibroblasts). We carefully considered genetic variation, gene expression and DNA methylation, and enriched these results by consideration of the epigenomic context in which correlated CpG sites were located. We also made use of our allelic expression (AE) Hidden Markov Model (HMM) developed in the previous chapter to enrich the sets of QTLs reported.

3.2. Abstract

DNA methylation plays an essential role in the regulation of gene expression. While its presence near the transcription start site of a gene has been associated with reduced expression, the variation in methylation levels across individuals, its environmental or genetic causes, and its association with gene expression remain poorly understood. We report the joint analysis of sequence variants, gene expression and DNA methylation in primary fibroblast

samples derived from a set of 62 unrelated individuals. Approximately 2% of the most variable CpG sites are mappable in *cis* to sequence variation, usually within 5 kb. Via eQTL analysis with microarray data combined with mapping of allelic expression regions, we obtained a set of 2,770 regions mappable in *cis* to sequence variation. In 9.5% of these expressed regions, an associated SNP was also a methylation QTL. Methylation and gene expression are often correlated without direct discernible involvement of sequence variation, but not always in the expected direction of negative for promoter CpGs and positive for gene-body CpGs. Population-level correlation between methylation and expression is strongest in a subset of developmentally significant genes, including all four *HOX* clusters. The presence and sign of this correlation are best predicted using specific chromatin marks rather than position of the CpG site with respect to the gene. Our results indicate a wide variety of relationships between gene expression, DNA methylation and sequence variation in untransformed adult human fibroblasts, with considerable involvement of chromatin features and some discernible involvement of sequence variation.

3.3. Introduction

Perhaps the best studied of epigenetic phenomena, the methylation of CpG dinucleotides has been known for many years to play a key role in X-chromosome inactivation (Payer and Lee 2008), transcriptional silencing of foreign DNA elements (Yoder et al. 1997) and imprinting of genes (Li et al. 1993), while aberrant DNA methylation is implicated in many types of cancer (Baylin et al. 1998). The relationship between methylation and gene expression is complex, with high levels of gene expression often associated with low promoter methylation (Kass et al. 1997) but elevated gene body methylation (Jones 1999), and the causality relationships have not yet been determined. In cell populations, the levels of DNA methylation across CpG sites in the genome is typically regarded as bimodal, with CpG-rich regions known as CpG Islands (CGIs), often associated with transcription start sites (TSSs), typically showing

hypomethylation, and other CpG sites showing hypermethylation (reviewed in (Jones 2012)).

Methylation has been shown to be highly variable across cell types with variable sites falling in two broad categories: those with inverse correlation between DNA methylation and chromatin accessibility, and those with variable chromatin accessibility and constitutive DNA hypomethylation (Thurman et al. 2012). As reviewed by Cedar and Bergman (Cedar and Bergman 2009), DNA methylation and histone modifications share many relationships from the time of embryonic development onwards, including hypothesized roles of DNA methylation preventing the tri-methylation of Histone 3 Lysine 4 (H3K4me3), a marker generally associated with active promoters, as well as H3K4me3 preventing DNA methylation (Hashimshony et al. 2003).

Methylation also varies between healthy individuals in a population. Relationships between DNA methylation, gene expression and various other genetic and epigenetic biomarkers have been examined previously. Recent studies have identified SNPs whose genotype correlates with DNA methylation (termed methylation quantitative trait loci, or meQTLs) in various human populations and cell types. Bell et al. (Bell et al. 2011) utilized the HumanMethylation27 BeadChips from Illumina to map associations between SNPs and methylation levels at 22,290 CpG dinucleotides in lymphoblastoid cell lines (LCLs), finding 180 CpG sites associated with nearby SNPs, and an enrichment for expression QTLs (eQTLS) amongst meQTLs. Gibbs et al. (Gibbs et al. 2010) used the same DNA methylation platform to study samples from four human brain regions in 150 individuals and reported hundreds of SNP-associated CpG sites in each brain tissue, with mQTLs typically located very close to the associated CpG site, and thousands of both mQTLs and eQTLs, but only modest overlaps between the two, averaging 13 CpG sites per tissue having a significant mQTL that was also an eQTL. Similar results were seen using 180 LCL lines derived from one African and one European population (Fraser et al. 2012). (Zhang et al. 2010) performed similar analyses using the same

methylation platform in 153 human adult cerebellum samples, finding 2046 CpG sites with mQTLs; they reported that in general CpG sites located in CpG islands are more likely to be mappable to a SNP than non-CpG-island sites. They also assessed the relationship between expression and methylation, with 20 of 112 CpG-gene pairs analyzed showing nominally significant correlations, with 5 of these 20 being positive correlations and the rest negative. At present, though it is known that there is a genetic component to both variable DNA methylation and gene expression, as well as genome-level differences in gene expression linked to DNA methylation, the combined relationships between the three factors remains poorly understood. Recent research (van Eijk et al. 2012) has examined the relationship between sequence, expression and DNA methylation as measured by the HumanMethylation27 assay in whole blood, finding numerous cases of methylation/expression relationships but focusing on the small number of cases in which a genetic component was also found. Drong et al. (Drong et al. 2013) report 149 CpG sites mappable to an mQTL when making use of differential methylation hybridization covering 27,718 genomic regions in 38 unrelated individuals, finding none of the mQTLs to also be eQTLs. Gutierrez-Arcelus et al. (Gutierrez-Arcelus et al. 2013) report positive and negative expression-methylation relationships at the inter-individual level in fibroblasts, T-cells and LCLs derived from a set of 204 umbilical cords from healthy newborns of European descent, with negatively correlated CpG sites enriched at ENCODE derived enhancer and promoter sites.

To further understand the relationship between genetics, gene expression, DNA methylation, and other epigenetic marks, we present analyses of DNA methylation, gene expression (both total and allelic) and DNA sequence polymorphisms, from a set of 62 fibroblast cell lines derived from healthy human individuals, augmented with publicly available histone mark and DNase I hypersensitivity (DHS) data. We show that:

a) Widespread relationships exist between DNA polymorphisms and DNA methylation (mQTLs).

b) Widespread relationships exist between DNA methylation and gene expression, especially in developmentally significant genes, including all four HOX clusters.

c) Thanks to the supplementing of expression quantitative trait locus (eQTL) data with mapping of allelic expression to adjacent SNPs, we obtain a large set of regions and genes mapping to a QTL which also functions as an mQTL, comprising 242 genes and 23 regions not overlapping with an annotated gene.

d) CpG sites where methylation correlates with gene expression in *cis* do not in general show strong overlap with annotated genes or promoter regions. Rather, CpG sites where this correlation is negative are most commonly seen in sites associated with active promoter marker H3K4me3 and DHS regions, while those with positive correlation are most commonly seen in the presence of the repressive chromatin marker H3K27me3.

3.4. Results

We report on the joint analysis of inter-individual variation in the levels of DNA methylation, total and allelic expression, and DNA sequence of 62 healthy parents of 31 parent-child trios of European descent. Here, we start by introducing each data set individually before discussing the relations among them.

3.4.1. DNA Methylation Assays

DNA methylation was assayed in forearm skin fibroblast samples using the Illumina 450K assay (Methods). For each sample, methylation was measured at approximately 485,000 CpG sites, but we only considered the approximately 392,000 sites uniquely mapped in autosomes and containing no known SNPs. Methylation levels are measured in populations of diploid cells using beta values

(Sandoval et al. 2011), which range from 0 (no methylation) to 1 (complete methylation of the two alleles). Methylation measurements were highly replicable, with the Pearson correlation coefficient between beta values of two replicates exceeding 0.99 in each of three pairs of biological replicates, while the average pairwise correlation coefficient between methylation from different samples levels ranges around 0.95 (Figure 3.11-1). Surrogate Variable Analysis (Leek and Storey 2007) was used to identify possible batch effects accounting for inter-individual methylation variation but none were detected, suggesting that the observed variation may mostly be due to stochastic, environmental, or genetic effects.

The Illumina 450K assay includes both type I probes utilizing two query probes per CpG locus (largely concentrated around genes' transcription start sites), and type II probes utilizing a single probe per locus (dispersed somewhat more uniformly across the genome; see Methods). The distributions of methylation beta values differ for type I and type II probes due to their localization biases but both are bimodal, with modes corresponding to CpG sites that are unmethylated in most cells of the sample (hypomethylated), and those that are methylated in most cells of the sample (hypermethylated) (Figure 3.4-1A (type II probes) and Figure 3.11-2 (type I probes)). Consistent with previous reports (Jones 2012), hypomethylated sites are mainly located in CpG islands and within 1.5 kb of the transcription start site (TSS) of a gene (53% of probes with mean beta value < 0.3 are located near a TSS, vs 34% of all probes; in the case of CpG Islands, it is 60% vs 32%), whereas hypermethylated sites are generally located in the rest of the genome (distal intergenic and gene body regions).

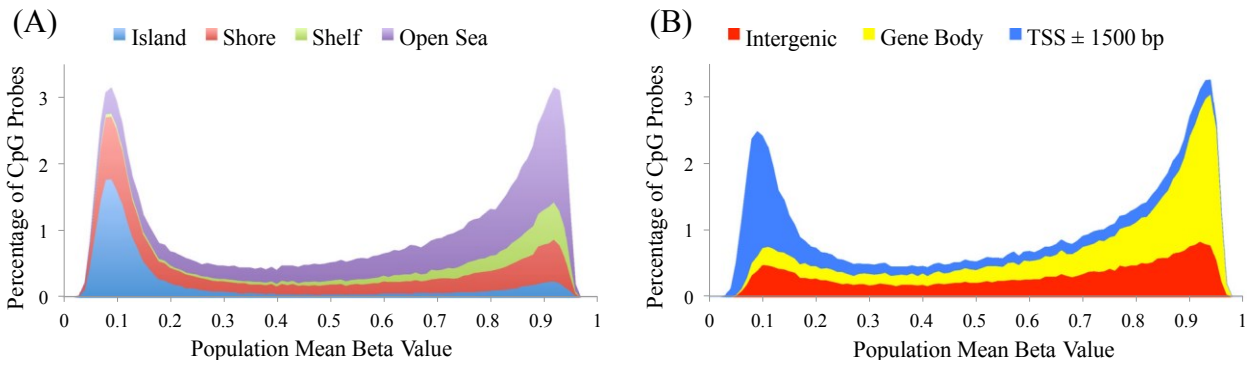

Figure 3.4-1 Fibroblast methylation beta values are bimodal and the two modes show different breakdown in terms of CpG islands and genes.

Distribution of methylation beta values in type II probes across the genome, partitioned by position relative to **(A)** CpG islands (with a shore defined by Illumina as less than 2 kb from an annotated CpG island, a shelf as 2 to 4 kb, and open sea as more than 4 kb) and **(B)** annotated genes.

Hypomethylated CpG sites are preferentially located in active regulatory regions characterized by DHS and H3K4me3, as measured by the ENCODE consortium in fibroblast cell lines (Myers et al. 2011) (Figure 3.4-2A (type II probes) and Figure 3.11-3A (type I probes)). Of hypomethylated CpG sites, 59% overlap with a DHS peak in the BJ foreskin fibroblast line, and 72% with an H3K4me3 peak. This is approximately twice the fraction seen among all CpG sites (29% and 34% respectively). On the contrary, hypermethylated sites show a considerable overlap with H3K36me3, an intragenic marker of active transcription (Rosenfeld et al. 2009), with 19% of sites with mean beta > 0.7 overlapping with a peak for this mark, compared to 9% among all sites. However, 62% of hypermethylated sites overlap none of the features considered in our analyses. Consistent with observations of low methylation in regions of DHS and active histone marks, genes with high expression levels show considerably lower methylation in the region proximal to the TSS (up to 1500 bp from the TSS) and higher methylation in the gene body region compared to genes with lower average expression levels (Figure 3.4-3A and Figure 3.11-4A), with probes

adjacent to genes in the top quartile of expression having mean beta < 0.3 81% of the time and mean beta > 0.7 only 11%. Those in the lowest quartile still have a plurality of hypomethylated probes near the TSS, but with numbers considerably diminished, i.e. 42% hypomethylated vs 30% hypermethylated.

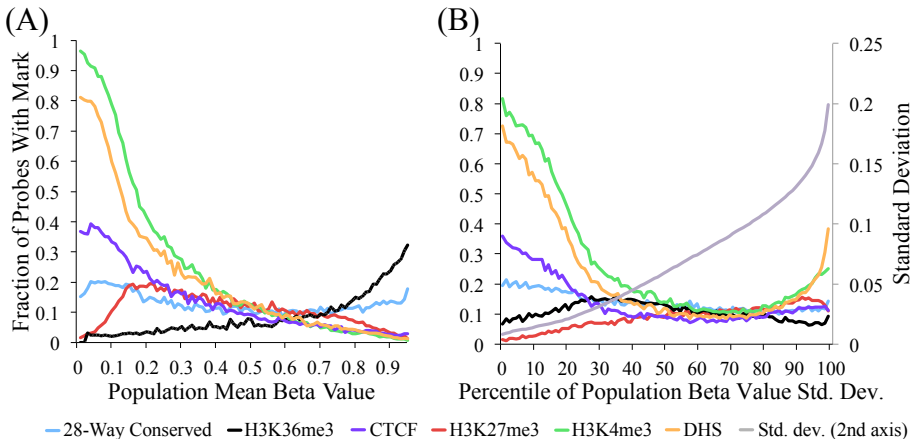

Figure 3.4-2 Mean and variance of beta values of CpG probes associate with several genome marks.

Proportion of type II CpG probes falling in various types of genomics regions identified by ENCODE, partitioned by **(A)** CpG probe mean beta value and **(B)** percentile of beta value standard deviation (Std. dev.). All data types, except for 28-way conservation, are derived from broad peaks in BJ human foreskin fibroblast cells.

We examined the levels of inter-individual variation of methylation probes, finding a drop in variation of probes located within 1500 bp of a TSS annotated for an actively expressed gene (Figure 3.4-3B and Figure 3.11-4B), with only 11% of probes near the TSS of a top quartile expression gene also being in the top quartile of methylation variation, compared to 30% for CpG sites adjacent to the TSS of a bottom quartile of expression gene. These results were corroborated by the finding that sites with low inter-individual methylation variation were enriched for DHS and H3K4me3, and, to a lesser degree, sequence conservation (Figure 3.4-2B and Figure 3.11-3B).

On the contrary, highly variable CpG probes (top 25%, std. dev > 0.0932) are usually located far away from the TSS (either in intergenic regions or in the gene body), or are located near the TSS of genes with low expression in fibroblasts and generally lack regulatory or evolutionary marks of function. The majority of these CpG sites show a unimodal distribution (Table 3.11-1). Genes whose TSS regions contain highly variable CpG probes were enriched for Gene Ontology (GO) terms related to multicellular organismal development (Table 3.11-2, worksheet 1), compared to the full set of genes having at least one CpG probe in the TSS region. Unexpectedly, extremely variable CpG probes (top 5%, standard deviation > 0.15) show a marked increase in their overlap with DHS and H3K4me3 marks. Genes collocated with these CpG probes are even more strongly enriched for having functions related to development, and include a large number of genes from the HOX clusters (see Discussion).

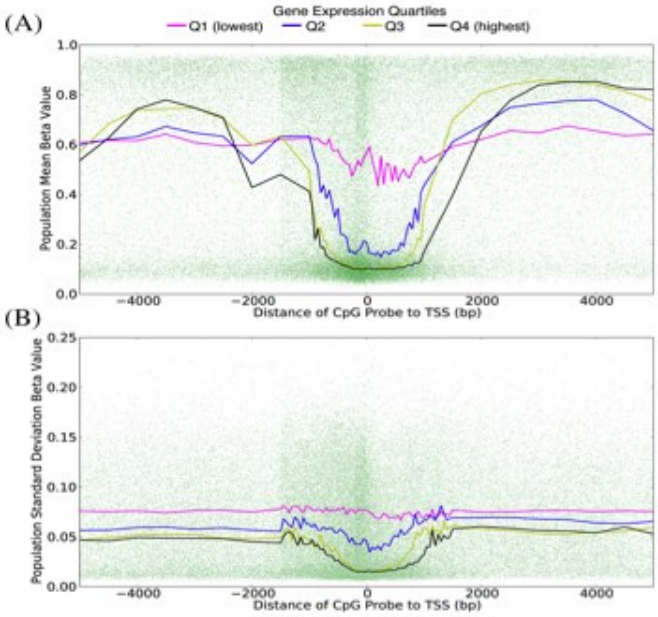

Figure 3.4-3 The mean and variance of beta values of CpG probes near transcription start sites depend on the gene's expression level.

Mean **(A)** and standard deviation **(B)** of type II CpG probes with respect to their position relative to TSSs of annotated genes. Each green dot corresponds to a CpG probe, and the four lines show the running median for probes based on the quartile of the expression level (from RNA-seq in four individuals) of the gene they are associated with.

3.4.2. Gene Expression Analysis

RNA expression levels for the 62 individuals were measured using the Illumina HumanRef8 microarray platform, giving expression levels for 21,916 probes mapping to a total of 16,952 genes. Only probes that showed moderate to high inter-individual expression variation (std. dev. > 0.1127, corresponding to a total of 9493 genes) were considered for further analyses. To complement total expression data, allelic expression (AE) was assayed at a set of approximately 900,000 SNP locations dispersed in annotated genes and intergenic regions of all autosomes using hybridization to genotyping arrays, as previously described (Ge et al. 2009) (see Methods). For each sample and each heterozygous SNP, the ratio of the expression level of each allele is estimated, after normalization to genomic DNA. Of 24,814 known canonical UCSC genes, 81% have at least one assayed SNP within their boundaries. A previously described (Wagner et al. 2010) hidden Markov model was used to reduce the noise in the data and estimate, for each SNP of each sample, the expected true allele expression log-ratio. We note that because this approach does not make use of gene annotation, it is able to detect AE at transcripts that do not, or only partially,

overlap annotated genes. However, detection power for genes that are short or contain a small number of SNPs is reduced.

As previously reported for other cell types (Ge et al. 2009), AE was seen to be widespread. We defined an aeSNP as a SNP whose expected $\log_2$ allele ratio is above 0.2 in at least two samples (which corresponds to 5% FDR; Methods), and found 74,624 aeSNPs within annotated gene regions (corresponding to 15.8% of genic/intronic SNPs), and 25,467 outside (corresponding to 5.4% of intergenic SNPs). aeSNPs were clustered into 3,327 aeRegions (consisting of two or more consecutive aeSNPs), of which more than 80% had full or partial overlap with an annotated gene (Figure 3.11-5), similar to results previously obtained in lymphoblasts (Wagner et al. 2010) (for full list of aeRegions, see Table 3.11-3).

3.4.3. Linking methylation and genetic variation

Inter-individual methylation variation is likely due to both genetic and environmental variation between samples. To determine the relationship between genetic variation and CpG methylation levels, we first genotyped our 62 samples (Methods). We then mapped CpG beta values to the imputed genotype at polymorphic sites within 250 kb (absolute value Spearman's rho above 0.452, which corresponds to a p-value of $6x10^{-6}$ and an FDR of 5% (Methods)). A set of 27,486 pairs (Table 3.11-4) were retained as significant, involving a total of 1,676 mappable CpG probes and 19,561 candidate methylation quantitative trait loci (mQTLs). Whole genome bisulfite sequencing (WGBS)-derived DNA methylation data were generated for four fibroblast cell lines (Table 3.11-5) and used to validate array methylation detected at mappable CpG loci. We observe high concordance between array and sequencing derived methylation for highly variable CpG sites, across the four cell lines (254 loci; median Pearson correlation coefficient = 0.84).

Remarkably, mappable CpG probes are 1.5-fold enriched in fibroblast DHS regions, but 1.75-fold depleted in highly conserved regions. While CpG probes found within CpG Islands are underrepresented in the set of highly variable CpG probes (Figure 3.4-1B), CpG Island probes are 1.66 fold enriched in mappable probes when compared to the set of highly variable CpG probes. Although mappable CpG probes represent only 1.7% of all highly variable CpG probes, they are approximately four times more frequent among extremely variable CpG probes relative to the set of highly variable probes (Figure 3.4-4). The majority of mappable CpG probes have a distribution of methylation levels that is unimodal, consistent with a moderate effect of genetic variation on methylation. However, bimodality and trimodality are much more frequent among this set of CpG probes than in highly variable CpG sites in general (29.7% and 4.8% of mappable probes, corresponding to 1.5- and 2.6-fold enrichments, respectively; Table 3.11-1). These correspond to cases where the impact of genetic variation is strong enough that classes of methylation levels are clearly distinct.

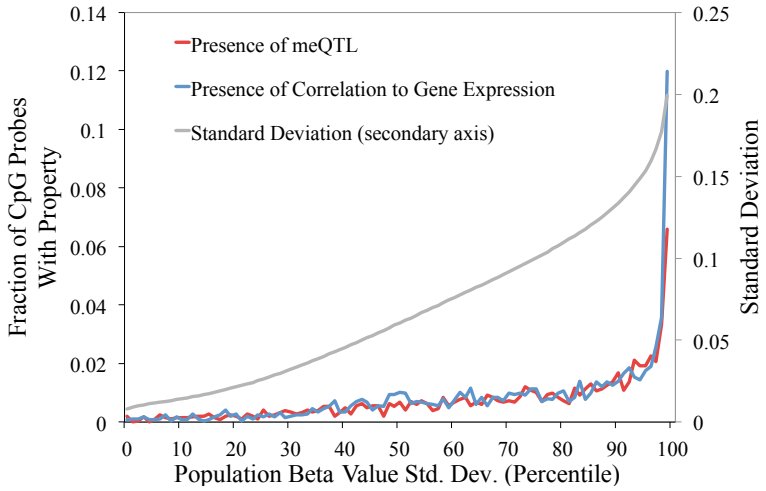

Figure 3.4-4 Variable CpG sites are more likely to be correlated with expression or sequence.

Proportion of probes being significantly correlated (5% FDR) to either an mQTL or a gene's expression levels, by percentile of population standard deviation.

The majority (67%) of mappable CpG probes have a significant mQTL within 5kb but in 6% of cases the closest significant mQTL lies more than 100kb away (Figure 3.4-5A). Despite their relative rarity, these distal regulators of methylation appear genuine, since even at these larger distances, such pairs are seen much more often than expected by chance (Figure 3.4-5B).

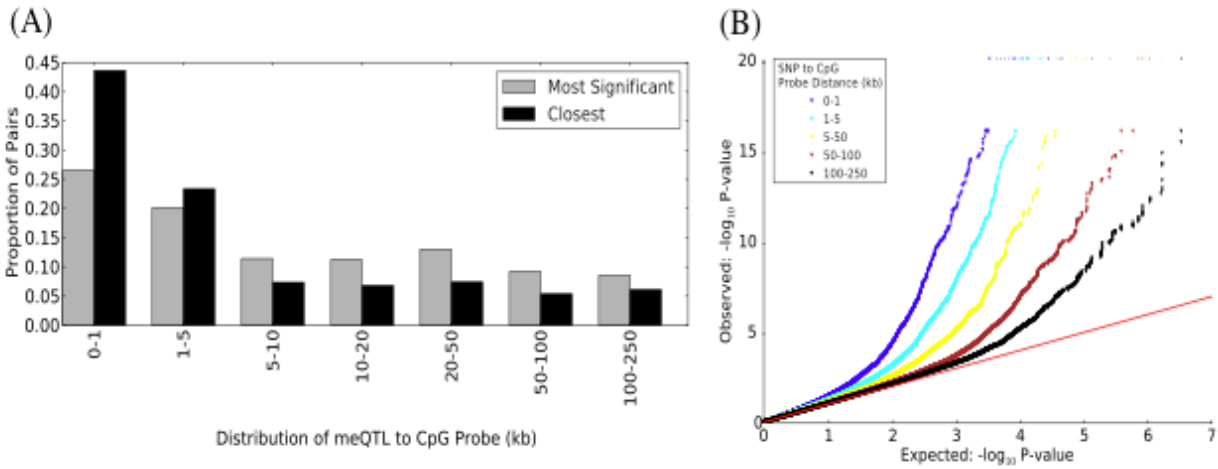

Figure 3.4-5 mQTLs are preferentially close to CpG sites.

(A) Distribution of the mQTL to CpG probe distances for all correlated SNP-CpG pairs at 5% FDR. For each CpG probe, when more than one SNP is significantly correlated, a single one is retained as having either the most significant correlation (gray bars) or being located closest to the CpG probe (black bars). **(B)** Quantile-quantile plot of SNP/CpG probe Spearman's rho *P*-values, grouped by pairwise distances. For each CpG probe included in the mQTL analysis, the most strongly correlated SNP within 250 kb was identified and the *P*-value obtained included in the set of *P*-values to be plotted for the distance bin in question. All SNPs in linkage disequilibrium with the selected SNP ($R^2 > 0.8$) were removed, and the next most strongly correlated SNP was taken, until all SNPs within the range of the CpG probe in question were considered. The number of significant mQTLs decays with distance, but is still more than expected by chance at distances greater than 100 kb.

3.4.4. Linking gene expression and genetic variation (eQTLs)

We sought expression QTLs (eQTLs) within 250 kb of each gene with variable expression (absolute value Spearman`s rho > 0.537, p-value < 1.4×10^{-5}, corresponding to a 5% FDR; see Methods). Such eQTLs were found for 420 (4.4%) genes and involved 9674 SNPs (Table 3.11-6). This is comparable to previous reports from (Veyrieras et al. 2008) (6.5% of genes mapping to an eQTL in lymphoblastoid cell lines, with a larger sample size of 210), but larger than the 2-3% seen by (Stranger et al. 2007) in four different HapMap populations. Consistent with previous reports (Stranger et al. 2007), genes with eQTLs were not enriched for any specific GO annotations. As previously reported (Veyrieras et al. 2008) eQTLs are most strongly over-represented near the TSS and transcription end site (TES) of genes, with a stronger enrichment within the gene body than outside (Figure 3.4-6).

These eQTL data were complemented with the mapping of allelic expression ratios in aeRegions to candidate regulatory allelic expression quantitative trait loci (aeQTLs) within 250 kb (Spearman rho > 0.452, p-

val=0.00029, corresponding to a 5% FDR; see Methods). A total of 95,949 aeQTL-aeRegion pairs were obtained (Table 3.11-7), involving a total of 2,360 (or 71%) aeRegions and 89,874 candidate aeQTLs (many of which being in linkage disequilibrium with each other). These mappable aeRegions had a significant overlap with 1452 annotated genes, three times more than the number of genes for which eQTLs were detected. 127 genes were found in both sets, corresponding to a 2.05-fold enrichment. Slightly larger overlap (2.92-fold enrichment) was observed in terms of the SNPs these genes mapped to. This significant but imperfect overlap by two methods is explained by multiple assay-specific factors: aeRegions are dependent on the presence of informative SNPs, are largely driven by primary transcript variation (intronic expressed SNPs) and in general allow for greater statistical power in terms of detecting statistically significant correlated SNPs (Suganuma and Workman 2011) whereas eQTL mapping (conducted on Illumina expression arrays) assesses both transcriptional and post-transcriptional variation and is skewed towards measuring exon-specific variation (Alter et al. 2000). Consequently, these methods can be used to complementarily capture different compartments of expression variation. Roughly 70% of mappable aeRegions have at least one candidate aeQTL within 5kb of one of their boundaries (Figure 3.4-7), which is comparable to results seen using eQTL analysis with known genes.

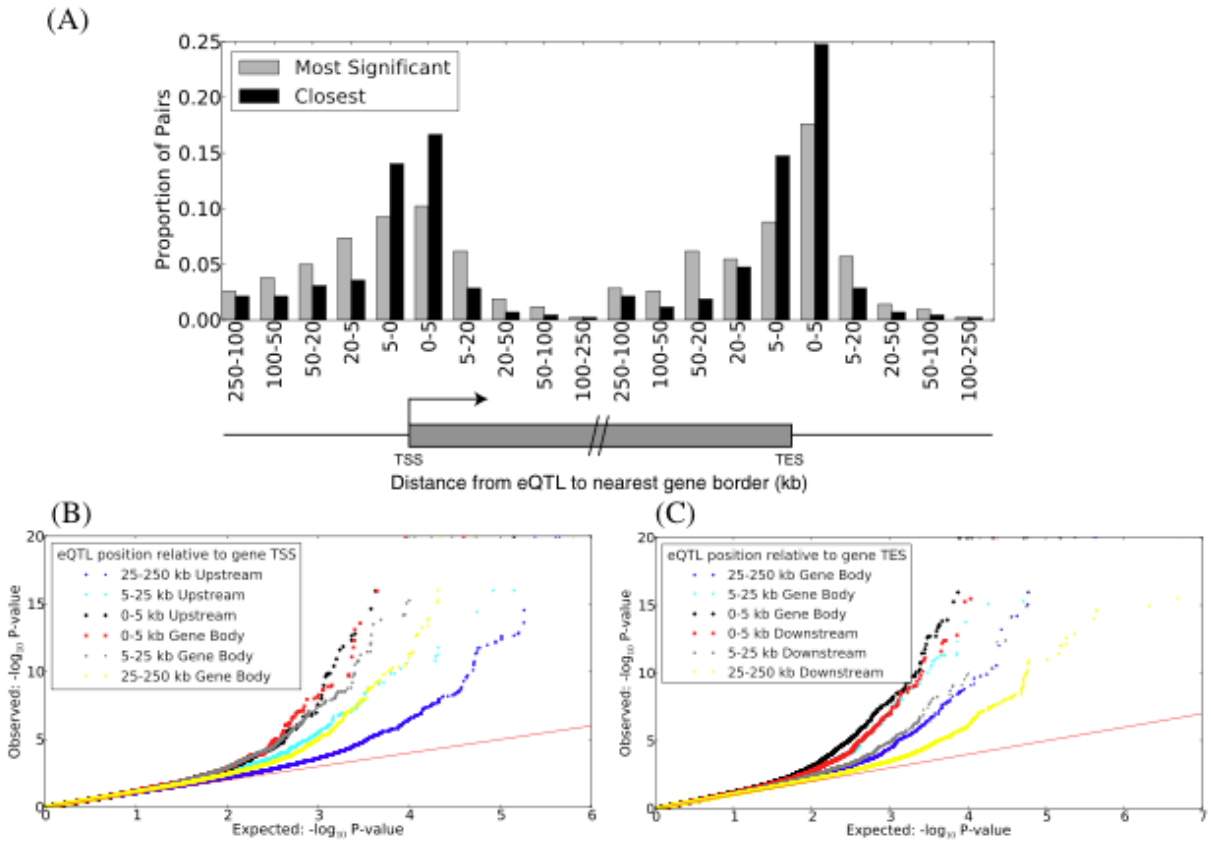

Figure 3.4-6 eQTLs are concentrated near the transcription start and end sites of genes.

(A) Distribution of the distance between eQTLs and the closest of the boundaries (TSS or TES) of the gene whose expression they correlated with, for all pairs at 5% FDR. When a gene's expression correlates significantly with more than one SNP, a single SNP is retained as having either the set of genotypes with the most significant correlation (gray bars) or being the most proximal to one of the two gene boundaries (TSS or TES). **(B,C)** Quantile-quantile plot of SNP/gene *P*-values, grouped by distances from the SNP to TSS **(B)** and TES**(C)**. Selection of *P*-values to be plotted followed a similar procedure to that in Figure 3.4-5B, with all SNPs located up to 250 kb on either side of the gene boundaries or within the gene body included for consideration.

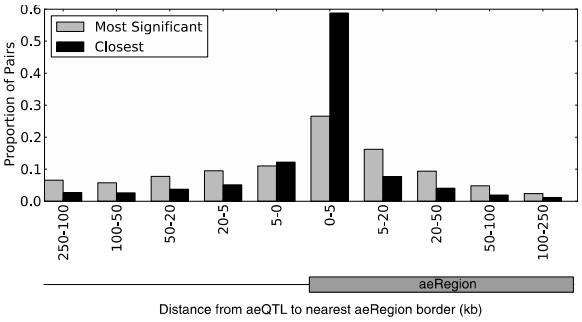

Figure 3.4-7 aeQTLs are concentrated near boundaries of aeRegions.

Distribution of the distance between aeRegion boundary and the SNP they correlate with (5% FDR). When an aeRegion's allelic expression correlates significantly with more than one SNP, a single SNP is retained as having either the set of genotypes with the most significant correlation (gray bars) or being the most proximal to one of the two aeRegion boundaries (black bars).

3.4.5. Linking gene expression to DNA methylation

We identified genes whose expression levels correlated with methylation levels of high-variance CpG probes located within their body or 250 kb on either end (absolute value Spearman's rho > 0.506, p-value $< 5.132 \times 10^{-5}$, resulting in an FDR of 5%, see Methods). This resulted in the identification of 587 genes with correlation to at least one of 1793 CpG probes (Table 3.11-8). Extremely variable CpG sites are strongly over-represented amongst sites correlated with gene expression (Figure 3.4-4), and correlated CpG sites are 1.6 fold and 3.2 fold enriched, respectively, for bimodal and trimodal sites relative to the set of highly variable CpG sites.

Remarkably, methylation-correlated genes are far from representing an unbiased sample of the genome, with 78 (13%) of them being known transcription factors (GO enrichment p-value = 8.23×10^{-16}) and 145 (24%) involved in multicellular organismal development (GO enrichment p-value = 6.1×10^{-22}) Table 3.11-2, worksheet 2). These include a number of genes from each of the four HOX clusters, together with several other key regulators of

development and cellular differentiation such as EN1, HAND2, TBX1, TBX2, TBX3, TBX5, and TBX15.

We sought to further characterize the CpG sites having methylation-expression correlations. Although about a quarter of methylation correlated genes had their closest correlated probe located within 1.5 kb of the TSS and 30% in their gene body, more than a third showed only correlation with distal intergenic probes (Figure 3.4-8). Since highly expressed genes have on average low DNA methylation near the TSS and higher DNA methylation at the gene body (Figure 3.4-3A), one might expect to see negative methylation-expression correlations for CpG probes located near a gene's TSS and positive correlations for CpG probes located in its body. However this is only partially verified, with one third of the former type of pairs showing a positive correlation and nearly half of the latter showing a negative correlation. Overall, strong enrichments were seen for both negatively and positively correlated probes in both the gene body and TSS region, compared to other regions 3' or more than 5 kb 5' of the gene (Figure 3.4-9).

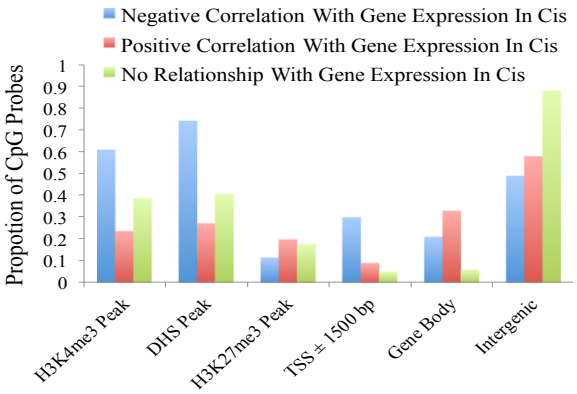

Figure 3.4-8 CpG sites where methylation positively or negatively correlates with expression differ with respect to chromatin marks.

Proportion of CpG probes having various chromatin marks in at least one of five ENCODE fibroblast cell lines or located at various positions with respect to genes, with CpG probes grouped into three categories based on the type of correlation seen with an adjacent gene expression values.

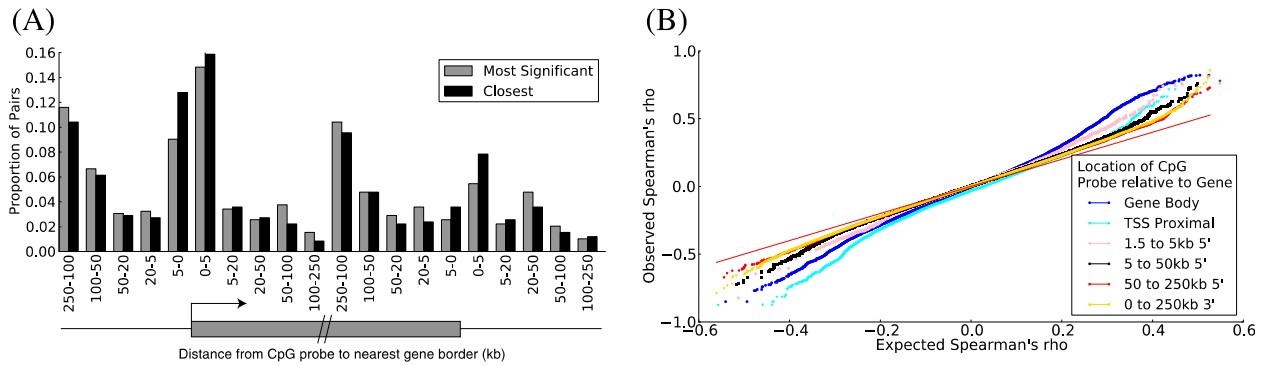

Figure 3.4-9 Positive and negative methylation/expression correlations are seen at all positions with respect to the gene.

(A) Distribution of the distance between expression-correlated CpGs and the closest of the boundaries (TSS or TES) of the gene whose expression they correlated with, for all pairs at 5% FDR. When a gene's expression correlates significantly with more than one CpG site, it is retained as having either the set of methylation beta values with the most significant correlation (gray bars) or being the most proximal to one of the two gene boundaries (TSS or TES) (black bars). **(B)** Quantile-quantile plot of methylation/expression rank based correlation (Spearman's rho), grouped by distances from the SNP to gene boundaries.

In order to find genomic features that may help distinguish CpG probes that correlate positively and negatively with gene expression, we turned to DHS and histone modification data obtained by the ENCODE consortium (Thurman et al. 2012), considering data from 5 human fibroblast cell lines. Though these cell lines were not derived from the same donors as used in this study, we found in general that they allowed a clear separation between the two types of CpG probes (Figure 3.4-8). CpG probes where methylation levels correlated negatively with gene expression are for the most part located in regions with marks of regulatory activity (H3K4me3 or DHS): marks that are less frequent among CpG probes that show no correlation with expression and even less frequent among those that show a positive correlation. In contrast, positively correlated probes were slightly more often seen with the inactive gene

associated marker H3K27me3 when compared with negatively correlated probes.

As illustrated in Figure 3.4-10, CpG sites in all types of genomic regions are more likely to be negatively correlated with gene expression if they are located in regions of DNase I HS in at least one of the five FB cell lines considered. A similar pattern was seen with the active transcription mark H3K4me3, with the notable difference that regions having this mark in all 5 fibroblast cell lines considered were *under*-represented for negatively correlated CpG marks, indicating perhaps that invariably active regions will also be subject to less consequential variability in terms of DNA methylation and expression. We also observe that regions containing H3K27me3 in at least one of the two fibroblast cell lines where this type of data was available are more likely to contain positively correlated CpG sites.

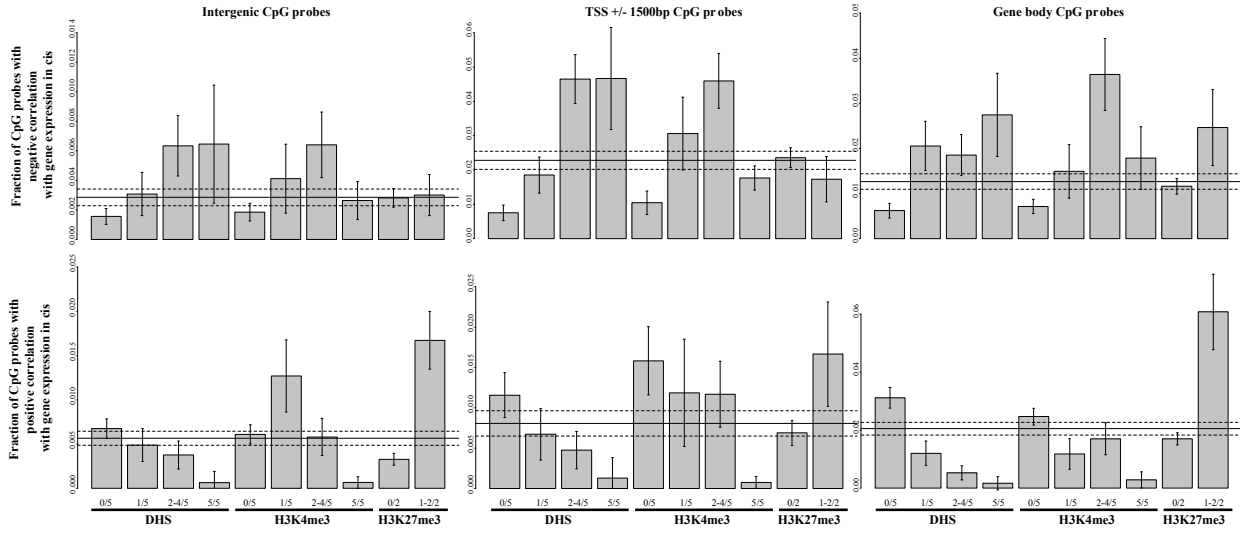

Figure 3.4-10 The proportion of CpG sites where methylation correlates with expression depends on the site location, DHS and histone marks.

Proportion of CpG probes showing correlation with gene expression, ±95% confidence interval, for probes located in intergenic regions (left), within 1.5 kb of the TSS (middle), or within the gene body (right), and showing either negative (top row) and positive (bottom row) correlation, depending on the presence of DHS, H3K4me3 and H3K27me3. For DHS and H3K4me3 marks, the individual bars are based on the number (out of five) of ENCODE fibroblast cell lines that have the mark in question.

In our samples, the four HOX clusters represent the densest centres of methylation/expression relationships in the genome. As seen in Figure 3.4-11A-D, each cluster is rich in both positive and negative methylation/expression correlations, involving CpG sites both within genes and intergenic regions, with many but not all negatively correlated sites lying in regions marked by H3K4me3 and/or DHS. Also of interest in HOXA and HOXD are the topological domains obtained from a recent Hi-C study in IMR-90 cell lines (Dixon et al. 2012) In HOXD, a 40kb region representing a boundary between the two domains contains the majority of CpG sites that have negative correlation with expression, whereas the boundary between two domains in HOXA also roughly delimits the positively and negatively CpG sites in this gene cluster. TBX1 and TBX3 represent other developmentally significant transcription factors having both positive and negatively correlated probes, whereas the latter largely coincide with DHS regions (Figure 3.4-11E,F).

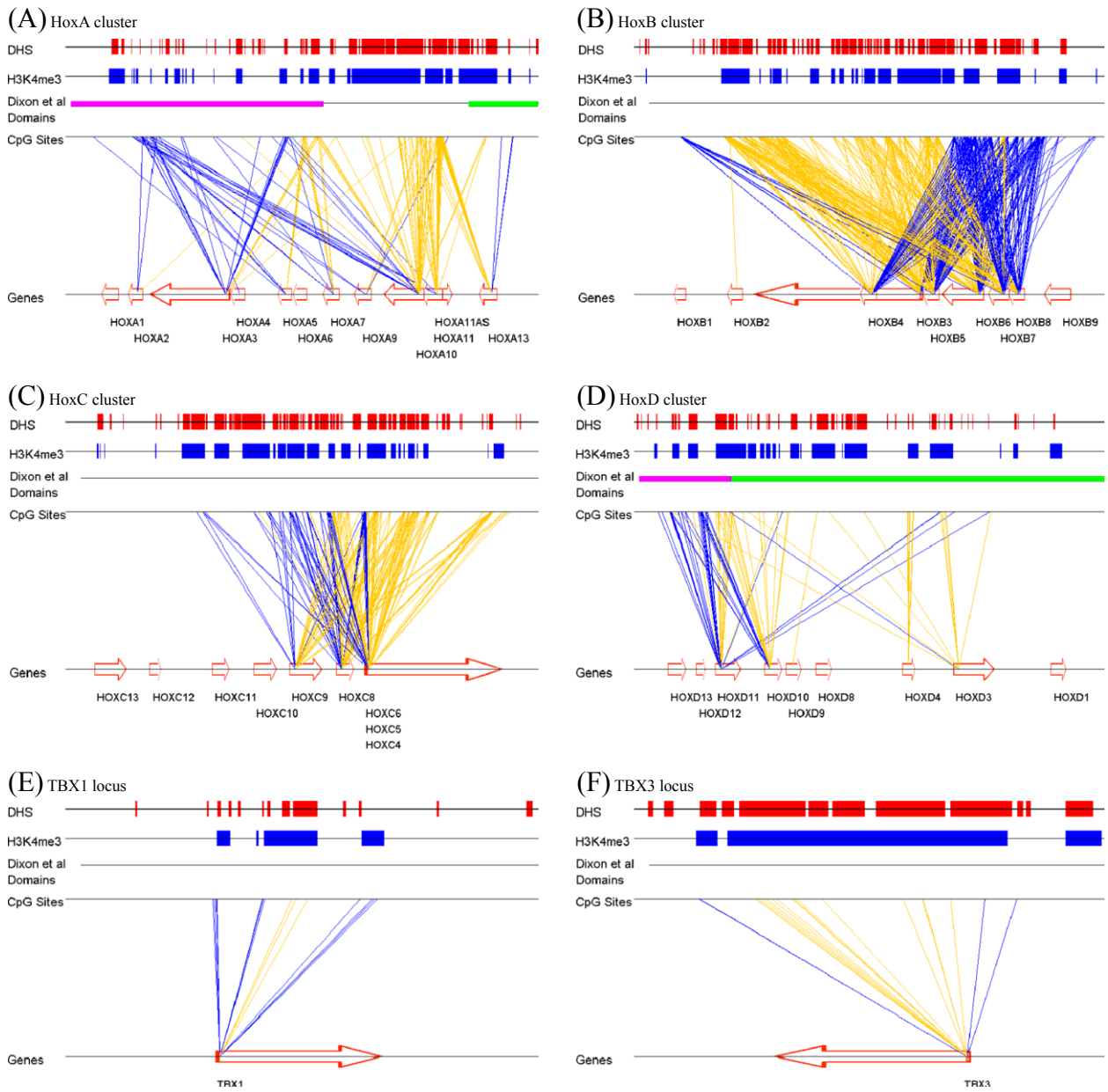

Figure 3.4-11 Methylation-expression relationships in genomic context.

Schematic of significant methylation-expression relationships for **(A-D)** the four HOX clusters, and **(E,F)** genes *TBX1* and *TBX3*. Gold and blue lines link the TSS of the gene and the CpG probes correlated to that gene's expression, with gold indicating negative correlation and blue indicating positive correlation. Red and blue blocks above indicate the presence of DHS or H3K4me3 marks in at least one of five ENCODE fibroblast cell lines. Where a domain boundary from (Dixon et al. 2012) was found, the domains are indicated with distinct colors.

3.4.6. Overlap between mQTLs and eQTLs

Three main types of relationships have so far been considered: methylation to sequence (mQTLs), expression to sequence (eQTLs and aeQTLs) and methylation to expression. To quantify the degree of overlap between the various relationships studied, we used genes, rather than CpG probes or SNPs, as the primary unit of interest. As seen in Figure 3.4-12, genes exhibiting two or three of the possible relationships form a relatively small but still non-negligible set. eQTLs and aeQTLs that were also mQTLs are termed in our report expression and methylation quantitative trait loci (emQTLs), and correspond to a total of 52 eQTL-mappable genes and 234 aeQTL-mappable aeRegions, that together form the set of emQTL-mappable loci obtained in our analyses. When emQTL-mappable aeRegions are broken into annotated genes they overlap with, and merged with the list of emQTL-mappable genes obtained via combining eQTLs and mQTLs, we obtain a set of 242 emQTL mappable genes, plus 23 emQTL mappable aeRegions not overlapping with any annotated genes. Compared to a random selection of SNPs matched for minor allele frequency, we find 5.9 times more mQTLs are also emQTLs than expected by chance.

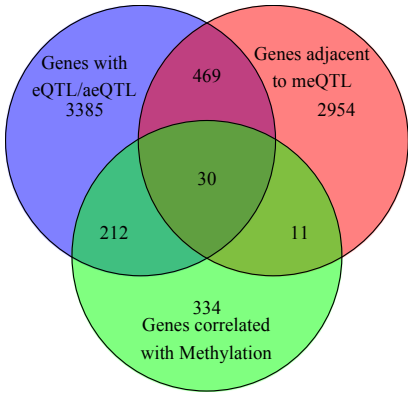

Figure 3.4-12 Overlap of genes with an eQTL, genes with expression correlated with methylation, and genes adjacent to mQTLs.

Number of genes corresponding to various categories or relationships.

One example of an emQTL-mappable gene is *C21orf56* (Figure 3.4-13A), which had previously been reported as having mappable CpG probes near the TSS (Bell et al. 2011). These probes overlap with DHS and H3K4me3 regions and are negatively correlated with expression. Also of note are positively correlated CpG probes located in the body of the gene, which are also mappable to a similar set of mQTLs.

Homeodomain transcription factor *PAX8,* transcription of which has been identified as an important biomarker in distinguishing various tumour types (reviewed in (Xiang and Kong 2013)) presented another particularly interesting case of overlap between the various types of relationships (Figure 3.4-13B), where CpG probes located near the gene's TSS were unexpectedly *positively* correlated with the gene expression and those located in its body were *negatively* correlated. A possible explanation may involve putative uncharacterized transcript DKFZP686E10196, antisense to and located within *PAX8*, whose expression would be negatively correlated with the CpG methylation at sites near its TSS (but in the body of *PAX8*) but positively correlated probes in the body of the transcript (but near the TSS of *PAX8*). Indeed, RNA-seq data obtained from 3 individuals with differing genotypes in the *cis*-associated emQTLs suggest that the expression of PAX8 and its antisense transcript are *positively* correlated, ruling out an interference between the two but instead hinting to a possible chromatin-linked role of DKFZP686E10196 activation in regulating PAX8 transcription. (For a recent review of antisense regulation, see (Faghihi and Wahlestedt 2009).)

Gene clusters of glutathione transferase families *GSTM* and *GSTT* also show multiple genes being mappable to similar sets of CpG probes and SNPs (Figure 3.4-13C,D), with active marks DHS and H3K4me3 located near negatively correlated CpG probes.

We estimated the proportion of gene expression variation that could be explained by either sequence variation alone or by a combination of sequence

variation and DNA methylation, using a simple linear model and 5-fold cross-validation (Methods). For each gene, the five SNPs (within 250kb) jointly explaining the largest portion of the expression variation on the training data were sequentially identified and regressed out. Independently of this we regressed out the five CpG sites explaining the largest portion of the expression variation. We found a total of 25.5% of gene expression variation to be explained by sequence variation, whereas methylation explained only 8.9% of expression variation. We applied a third model in which the top five SNPs were regressed out and then the top 5 CpGs were regressed from the residuals, finding in this case the variation explained by methylation dropped to 5.9%. This suggests that 5.9/8.9 = 66% of methylation facilitated gene expression variation was independent of sequence variation. These figures are considerably higher than the 1.2% and 3.3% variation of expression explained, respectively, by DNA sequence and DNA methylation found by (Li et al. 2013). in breast tumors indicative perhaps of much greater variation of gene expression brought about by other factors in the tumor micro-environment.

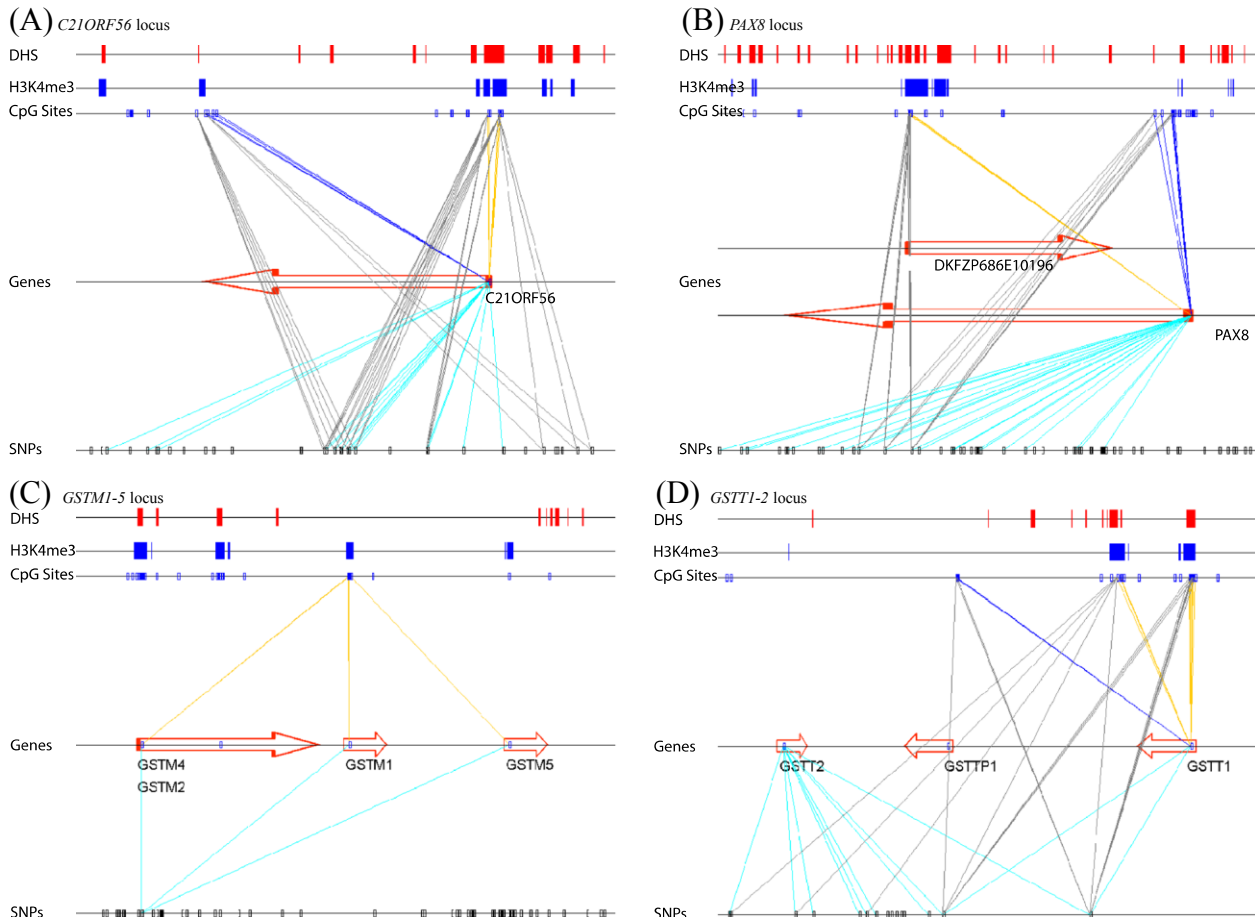

Figure 3.4-13 emQTL relationships in genomic context.

Schematic of methylation-sequence-expression relationships in the loci surrounding the **(A)***C21ORF56*, **(B)***PAX8*, **(C)***GSTM1-GSTM5*, and **(D)***GSTT1-GSTT2*genes. Annotations are similar to those in Figure 3.4-11, with added grey and cyan lines indicating mQTL and eQTL relationships, respectively.

3.5. Discussion

We have analyzed the inter-individual variability of and relationships between one of the most comprehensive set of biomarkers in untransformed adult cells to date, including a more expansive assay for DNA methylation

containing a large and diverse set of CpG dinucleotide probes; gene expression data; SNP data and allelic expression data, augmented with publicly available histone mark and DHS data from other cell cultures.

We chose primary skin fibroblasts as a model system. These comparatively easy to isolate and cultivate cells are a readily accessible source of patient material, and are in use as model system for complex diseases etiological studies of e.g. Parkinson's disease (Auburger et al. 2012). However, epigenomes are tissue-specific, hence the use of primary skin fibroblasts is limited to gaining insight into complex diseases of skin fibroblastic origin, or being a complementary tool with the requirement of additional studies of (the mostly more difficult to derive) primary patient tissue material. Of course, the limited sample size of this study reduced our ability to detect weak associations. However, complementing eQTL with mapping of allelic expression significantly increases the sensitivity of our expression mapping (Almlöf et al. 2012), resulting in the discovery of many more expression/methylation QTLs than reported before.

Although most CpG sites with variable methylation seem unrelated to variation in gene expression, a non-negligible portion show significant correlations. Remarkably, the properties of these relationships appear quite complex, and the location of CpG probes with respect to the gene provides relatively little information about the sign of the correlation. Instead, chromatin states, particularly those that are representative of active chromatin and transcribed regions (DHS and H3K4me3) were more strongly indicative of negative correlation. Using the publicly available ENCODE data, we found in general that negatively correlated probes most strongly overlapped with regions of constitutive DHS but variable H3K4me3 among the five fibroblast cell lines considered, whereas positively correlated probes most strongly overlapped with an indicator of inactive transcription, H3K27me3. Work published in the ENCODE paper on DHS (Thurman et al. 2012) indicated an inverse correlation of DNA methylation and DHS, and the authors provided evidence that DNA methylation

was excluded as a consequence of open chromatin, rather than DNA methylation preventing this opening from occurring. H3K4me3 was also previously found to be inversely correlated with DNA methylation (Cedar and Bergman 2009), with evidence for causality pointing in both directions. We have found further signs of intriguing links between all of these marks, and hope for experiments in the future more actively measuring these marks within the same cell lines to give better clues as to causality and to establish the constitutive and variable marks included in methylation-expression relationships.

Whether the associations between gene expression and methylation truly reflect variation in tissues or other differences acquired after sample collection is an important and challenging question. One possible source of post-sample collection variation is differences in cell proliferation rates. However, we have found that cell proliferation variation only explained 8% of the variance in methylation levels of expression-correlated CpG sites, and 13% of the variance in expression levels of methylation-correlated genes. Among mappable CpG sites and genes, the proportion of variance explained was negligible (0.7% and 5% respectively).

Relatively high overlap was seen with results from previous studies in terms of the rare genes where both expression and methylation could be linked to genetic variation (emQTLs). In particular, *C21orf56,* a gene for which we find many emQTLs in fibroblasts*,* also exhibits the same property in whole blood (van Eijk et al. 2012) and LCLs (Bell et al. 2011). Several other genes having emQTLs in whole blood (van Eijk et al. 2012) (*GSTM3*, *NAPRT1*, *SPG7* and *WBSCR27*) were also identified in our assay, indicating that genetic variation leading to both methylation and expression variation at the same locus is a relatively rare but reproducible phenomenon, the mechanism and implication of which merits further investigation. We report a total of 260 annotated genes or aeRegions that, to our knowledge, have not been previously reported as having emQTLs, including 23 aeRegions having no overlap to annotated genes. We attribute these discoveries to the usage of allelic expression assays as well as a gene

expression microarray experiments, together with use of the relatively recently developed Illumina Infinium HumanMethylation450 platform, interrogating methylation at a larger and more diverse set of CpG sites compared to most previous studies. As the effect of methylation on gene expression can in some cases involve cell-specific *trans*-acting factors (Cedar and Bergman 2009), additional emQTLs could be found if we were to extend our analyses to additional cells or tissues. Future studies with larger sample sizes, investigating more diverse sets of cell types and utilizing platforms with even more comprehensive coverage of CpG sites can only help to uncover a greater number and potentially more subtle cases of associated DNA methylation, gene expression and DNA sequence variation.

Relationships between gene expression and DNA methylation in a population setting have not been investigated as extensively as sequence-expression or sequence-methylation relationships. However, previous high-throughput gene expression studies in fibroblasts have revealed intriguing results. In a landmark paper (Chang et al. 2002) assessing gene expression in skin fibroblasts derived from various anatomical sites, genes involved in a) extracellular matrix formation, b) cell signalling or fate determination, and c) cell migration signals were found to be expressed in a positional dependent fashion. Most notably of all, clustering of the samples based solely on the expression levels of 51 HOX genes recapitulated their site of origin. (Koch et al. 2011) strengthened these results by also finding positional dependent DNA methylation at HOX loci in a set of skin fibroblast samples. In the present study, fibroblast samples drawn from the same site but from different individuals show considerable DNA methylation variation in CpG sites proximal to all four HOX clusters, and a subset of HOX genes are amongst those with the closest expression-methylation ties in the genome. However, the HOX genes with correlations to methylation reported differ from those previously found to have position dependent expression (Chang et al. 2002), indicating additional layers of complexity and additional factors affecting fibroblast HOX methylation/gene

expression beyond position in the body. Parents from several of the trios showed similar HOX expression and methylation profiles, indicating perhaps an environmental rather than a genetic origin for these characteristic patterns. Although this was not discussed in their paper, the data reported by (Gutierrez-Arcelus et al. 2013) also indicated that all 4 HOX clusters, as well as *PAX8*, showed high levels of methylation/expression correlations in each of the three cell types they studied. Future studies taking into account more carefully the environment and background of unrelated, healthy individuals will be paramount in understanding more clearly the factors at play in DNA methylation and gene expression of these fascinating loci. Overall, the inter-individual variability in gene expression seen in this fibroblast dataset, and the relationship of this variability to DNA methylation show intriguing parallels to results seen with positional gene expression and DNA methylation variability in fibroblasts.

Genetic and methylation variation jointly explain 31% of gene expression variation in our fibroblast samples. However the mechanisms involved appear complex and diverse, with a close interplay with other epigenetic marks. Further studies assaying inter-individual variation in histone marks and chromatin accessibility, ideally in an allele-specific manner, may bring the context necessary to the interpretation of sequence and methylation variation.

3.6. Conclusions

We report a comprehensive analysis of relationships between sequence variation, DNA methylation and gene expression in untransformed adult human fibroblast cells. Consistent with previous reports showing positional effects in fibroblast on HOX gene expression (Chang et al. 2002) and DNA methylation (Koch et al. 2011), we show inter-individual variation and correlation between DNA methylation and gene expression in fibroblast cells even when drawn from the same location in the body. CpG sites with positive and negative correlations to gene expression show distinctive patterns with respect to the histone marks

and chromatin accessibility seen in their genomic region in other fibroblast cell lines. We find in general the most remarkable relationships found with these data to be those involving gene expression and DNA methylation in developmentally significant regions having little or no discernible involvement of DNA sequence variation.

3.7. Materials and Methods

3.7.1. Description of cell lines and cell culture

Primary skin fibroblasts were obtained from Coriell (Camden, NJ, USA) and the McGill Cellbank (Montreal, QC, Canada). Cells were grown in alpha MEM Medium (SigmaAldrich, Oakville, ON, Canada) supplemented with 2 mmol/l L-glutamine, 100 U/ml penicillin, 100 mg/ml streptomycin, and 10% fetal bovine serum (SigmaAldrich) at 37°C with 5% CO_2 to 70-80% confluence, then harvested and stored at -80°C until RNA and DNA was extracted.

3.7.2. DNA and RNA extractions

Genomic DNA (gDNA) for SNP genotyping and DNA methylation analysis was extracted from cell lysates using the GenElute DNA Miniprep Kit (SigmaAldrich) and DNeasy Blood and Tissue Kit (Qiagen), respectively, according to manufacturer's protocol. DNA concentrations were determined using the Quant-iT PicoGreen kit (Invitrogen, Burlington, ON, Canada). Total RNA was extracted from cell lysates using the RNeasy Mini Kit (Qiagen) according to manufacturer's protocol, and treated with 6 U DNase I. RNA quality was confirmed to be high for all samples on the Agilent 2100 Bio-Analyzer (Agilent Technologies, Mississauga, ON, Canada), with an RNA integrity number (RIN) range of 8.1 to 10, and concentrations were determined using the Nanodrop ND-1000 (NanoDrop Technologies, Wilmington, DE, USA).

3.7.3. 450K methylation array

500 ng gDNA was used for bisulfite conversion employing the EZ DNA Methylation Kit (Zymo Research), according to manufacturer's protocols. The modified gDNA was processed as described in the Infinium Assay Methylation Protocol Guide Rev. C (November 2010), and analyzed on Infinium HumanMethylation450 BeadChips (Illumina, refer to http://www.illumina.com/documents/products/technotes/technote_hm450_data_analysis_optimization.pdf for more details), measuring DNA methylation at single CpG-site resolution based on genotyping of C/U polymorphisms. We excluded probes with ≥ 90% sequence similarity to multiple genomic locations, probes with sequence variants in the probe-binding region and probes located on sex chromosomes, leaving 392,904 probes for further analyses. For removal of variant-containing probes HapMap (release 28, 30 CEU trios) annotated variants were imputed with 1000 Genomes project variants (pilot), and probes mapping more than one variant were removed. As a measure of methylation we chose the beta-value, which theoretically ranges from 0, indicating no methylation at any allele, to 1.0 for complete methylation of both alleles.

Beta values of CpG probes were quantile normalized separately for type I and type II probes, with the reference distribution being the distribution of average per-probe beta values. Surrogate variable analysis (Leek and Storey 2007) was carried out using the sva package in Bioconductor (http://www.bioconductor.org/packages/release/bioc/html/sva.html) and identified no hidden variables responsible for variation in data, furthermore, following a methodology similar to that of Bell et al. (Bell et al. 2011), residuals obtained after regressing out up to 5 principal components were mapped to candidate mQTLs, and in none of the cases were a larger number of mQTLs or mQTL-mappable CpG probes obtained than with simply using quantile-normalized beta values, therefore quantile normalized beta values were used throughout for further correlation analyses.

3.7.4. Cell proliferation effects on expression and methylation

DNA concentrations from 8 individuals were used to obtain a set of 42 developmentally significant genes whose expression strongly ($R > 0.75$) correlates with DNA concentration. The first principal component of expression levels for the set of these 42 genes was obtained and used as a vector estimating the level of cell proliferation effects in the full set of individuals. For each methylation probe correlated either to gene expression or sequence variation, we carried out linear regression with the probe's beta values and the cell proliferation vector. The variance of the residuals was compared with the variance of the original methylation probe, and done so cumulatively across probes to obtain the total variation in methylation of correlated probes explained by cell proliferation effects (with separate categories for CpG sites correlated to DNA sequence and gene expression). The process was repeated with expression probes found to be correlated with eQTLs and/or methylation of adjacent CpG sites to obtain an estimation of the (total variation in expression of correlated genes explained by cell proliferation effects.

3.7.5. Whole genome bisulfite sequencing

Whole genome bisulfite sequencing was carried out for cell lines GM02316, GM02317, GM02456, and GM02555 as described (Browning and Browning 2009) with the modification that bisulfite conversion was carried out with the EZ DNA Methylation Kit (Zymo Research, Irvine, CA, USA) according to manufacturer's protocol. 100 bp paired-end sequencing was carried out on the Illumina HiSeq 2000 system; sequencing details are given in Table 3.11-5. Reads were mapped to the bisulfite converted reference genome using BWA and processed as described by Johnson et al (Boyle et al. 2008b).

3.7.6. Allelic expression measurement

Allelic expression measurement was carried out as described previously (Ge et al. 2009). In short, approximately 200 ng gDNA and 50-300 ng double-stranded cDNA were genotyped in parallel on Illumina Infinium HumanOmni1-Quad, or HumanOmni2.5-Quad microarrays. The cDNA synthesis protocol was applied on heteronuclear RNA, allowing the measure of unspliced primary transcripts. For cDNA synthesis approximately 150 µg of total RNA was enriched using the MicroPoly(A)Purist protocol (Ambion Inc., Streetsville, ON, Canada). First strand cDNA synthesis was carried out on 1 µg poly(A)-enriched RNA using random hexamers, and second strand cDNA synthesis was performed using the Superscript Double-Stranded cDNA Synthesis Kit (Invitrogen). Data were filtered removing non-expressed SNPs and SNPs where cDNA arrays were unable to discriminate between homozygous genotypes, and normalized to compensate observed intensity dependent shift in median beta values of cDNA vs gDNA. For filtered SNPs obtained in the assay, smoothed scores of allelic expression were assigned based upon an 8-state Left-to-Right Hidden Markov Model (LTOR-HMM) as described in (Wagner et al. 2010). Based upon tests in which a null distribution was simulated by permuting raw allelic expression ratios independently within each sample, a model trained and smoothed allelic expression scores obtained from the LTOR-HMM, a threshold of 0.2 in at least 2 samples was identified as identifying allelically expressed SNPs with an FDR of 5%. Consecutive aeSNPs (ae-SNPs) having a smoothed allelic expression value of the same sign and above this threshold in at least 2 individuals were aggregated into regions of allelic expression (aeRegions) and the mean smoothed AE score was obtained and assigned independently for each individual, in each aeRegion.

3.7.7. Genotyping

Imputation of HapMap genotypes and phasing of Infinium HumanOmni1 and HumanOmni2.5-derived genotyping data were done using Beagle (Browning

and Browning 2009). The SNPs used in correlation analysis throughout this study to obtain eQTLs, aeQTLs, mQTLs and emQTLs are all based upon this same set of SNPs.

3.7.8. Gene expression arrays and eQTL analysis

Gene expression levels for 58 of the 62 individuals were determined using the Illumina HumanRef-8 Expression BeadChip according to manufacturer's protocol, giving expression levels for 21,916 probes mapping to a total of 16,952 genes.

These expression values were quantile normalized, the genes filtered such that only those in the top 50% variance of expression were retained, and expression values of these genes correlated using Spearman's correlation coefficient to all SNPs within the gene boundaries or up to 250kb upstream of the TSS or downstream of the TES. Expression values of the top 50% variable genes were permuted and the correlation analysis repeated to obtain a null distribution of p-values, and a p-value of 1.4×10^{-5} was obtained as the cut-off yielding a 5% FDR.

3.7.9. Identifying allelic expression aeQTLs

All HapMap (release 28) SNPs at a distance of ±250 kb flanking each aeRegion and having minor allele frequency > 10% were correlated using Spearman's correlation coefficient to their respective aeRegion. For each aeRegion, allelic expression values were permuted amongst the samples and the regression repeated to obtain an overall null distribution used in determining the FDR of p-values. A p-value threshold of 0.0029 was set based upon an FDR of 5%.

3.7.10. Identifying methylation quantitative trait loci (mQTLs).

Only probes having variance across samples in the top 25% were kept for correlation analysis with SNPs. Spearman's rho was calculated between the highest 25% variance probes and HapMap SNPs at a distance of ±250 kb flanking each CpG probe and having minor allele frequency > 10%. For each variable CpG probe, the analysis was repeated with methylation values permuted across individuals, in order to obtain a p-value of $6x10^{-6}$ for an FDR cut-off of 5%.

3.7.11. Methylation-Expression Correlation

The same set of top 25% variable methylation probes and top 50% variable genes in the Illumina HumanRef-8 Expression BeadChip were used, obtaining the Spearman correlation coefficient between any methylation probe located within the body of an annotated gene or up to 250kb on either side. Expression levels for each gene were permuted across the samples and the same set of Spearman correlation coefficients obtained, in order to set the p-value cutoff of $5.132x10^{-5}$ for a 5% FDR.

A CpG probe was labeled as being in the "TSS" a gene if it was +/- 1500 bp from its TSS. It was labeled as "body" if it was not located within 1500 bp of any TSS but was within an annotated transcript. Finally, it was labeled as "intergenic" if it was neither "TSS" nor "body".

The percentage contribution of methylation and sequence variation to expression variation was assessed using 5-fold cross-validation and step-wise feature selection. For the training subset (80% of individuals), a linear model with expression of a particular gene as the response variable and genotypes of SNPs in the neighborhood of that gene as explanatory variables was selected using the stepAIC function in R (http://stat.ethz.ch/R-manual/R-patched/library/MASS/html/stepAIC.html), the model was then used to predict expression values in the testing subset (1/5) of individuals. The same procedure of training and predicting was used across all 5 folds, and the R^2 between the

expression values and the predicted expression values using the models was obtained as the percent of expression variation explained by sequence variation. The same procedure was repeated with the residuals of the gene expression values from the sequence-expression model as response variables and methylation beta values of CpG probes in the neighborhood as explanatory variables, in order to obtain the percentage of expression variation explained by methylation variation.

3.7.12. Gene Ontology (GO) term enrichment

Significantly overrepresented GO categories were obtained for variable CpG probes and genes correlated to DNA methylation using Fisher's Exact test via GOStat (Beissbarth and Speed 2004), using default parameters available on the web server.

In the case of enrichment for highly variable CpG sites, genes with at least one top 25% variable CpG site at TSS +/- 1500 bp were used as the test set; the set of all autosomal genes overlapping with at least one Illumina 450K CpG probe were used as the background set. In the case of methylation-expression correlation, the set of all genes whose expression correlated significantly at 5% FDR with at methylation of at least one CpG site were used in the test set; the set of all genes containing at least one CpG site within 250kb were used as the background set. P-values were calculated by the GOStat web server, whereas fold enrichment was determined by dividing the proportion of genes in the test set with a given GO term by the proportion of genes in the background set with the same GO term.

3.7.13. Overlap with DNase I Hypersensitivity and Histone Markers

Data were downloaded from the ENCODE Data Consortium Center at UCSC at http://genome.ucsc.edu/ENCODE/downloads.html on October 15, 2012, namely UW Dnase I HS and UW Histone broad peak data for fibroblast cell

lines: Ag04449, Ag04450, Bj, Hff and Hcfaa (only Ag04450 and Bj were available for H3k27me3). A genomic locus was defined as having a given mark if that mark was present in at least one of the three cell lines. For each variable gene, the sets of: a) positively correlated methylation probes, b) negatively correlated methylation probes, and c) all probes in a 250kb neighborhood were obtained. For each category, the average (across genes) proportion of probes overlapping each type of mark was determined.

3.7.14. Overlap between mQTLs and aeQTLs

The set of all SNPs that are categorized as correlated to both gene expression (*aeQTL* having relationship to an aeRegion and/or an eQTL correlated to a gene in the Ref8 array) and to DNA methylation (mQTL) at an FDR threshold of 5% in *both* of the respective analyses described above are categorized as methylation-regulatory SNPs (emQTLs).

3.8. Competing Interests

The authors declare they have no competing interests.

3.9. Authors' contributions

TP, MB, SB and JW conceived of the study. JW, MB, TP and SB wrote the article. TP, SB, BG and TK performed DNA methylation, gene expression, allelic expression and SNP genotyping experiments. JW, MB, and SB performed the computational analyses. All authors have read and approved the manuscript for publication.

3.10. Acknowledgments

We thank Mathieu Lavallée-Adam for useful discussions. This work was supported by Canadian Institute of Health Research grants MOP-111246 awarded to TP and EP1-120608 awarded to TP and MB, and by a Natural Science and Engineering Research Council of Canada Discovery grant to MB. TP is a recipient of a Canada Research Chair Tier 2 award. SB is a recipient of a postdoctoral fellowship from the Réseau de Medecine Genetique Appliquée (RMGA).

3.11. Supplementary Figures and Tables

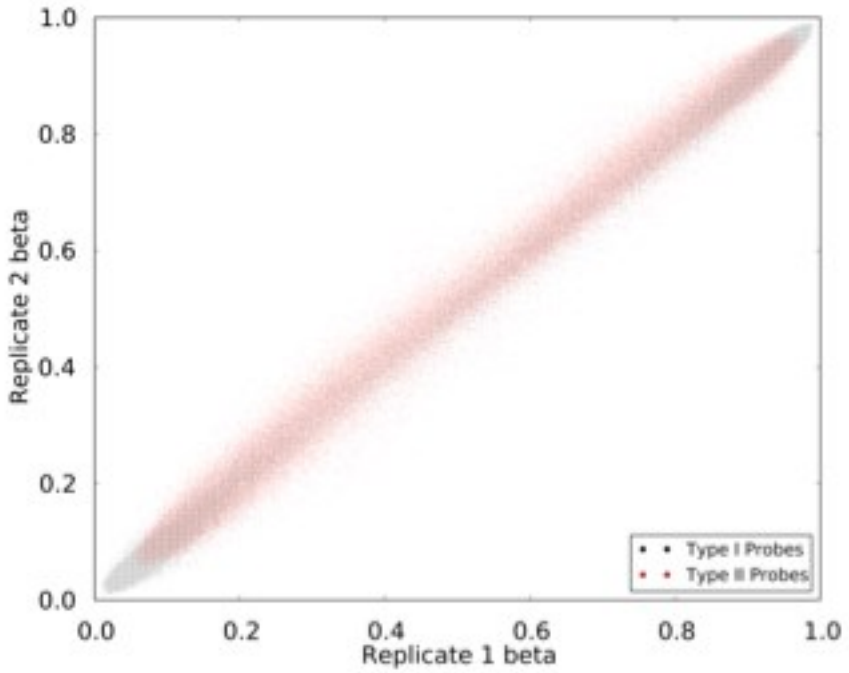

Figure 3.11-1 Replicability of beta values in samples GM02456.

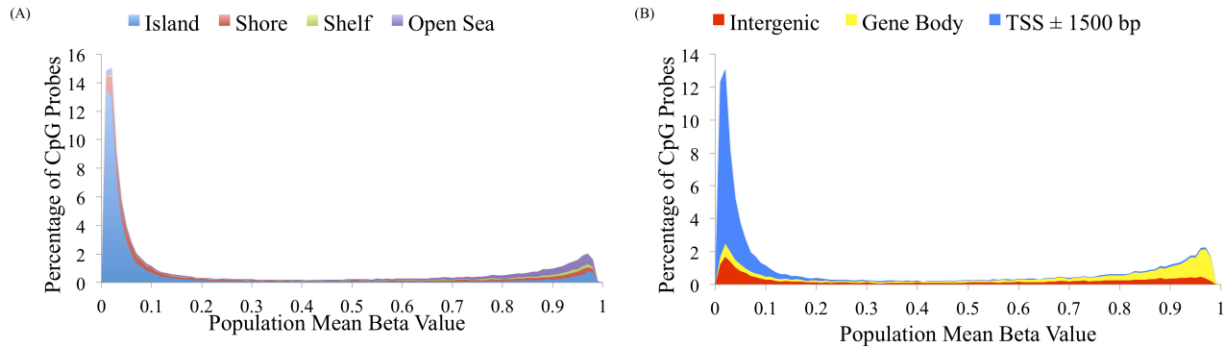

Figure 3.11-2 Distribution of methylation beta values in type I probes across the genome.

Values are partitioned by position relative to (A) CpG islands and (B) annotated genes.

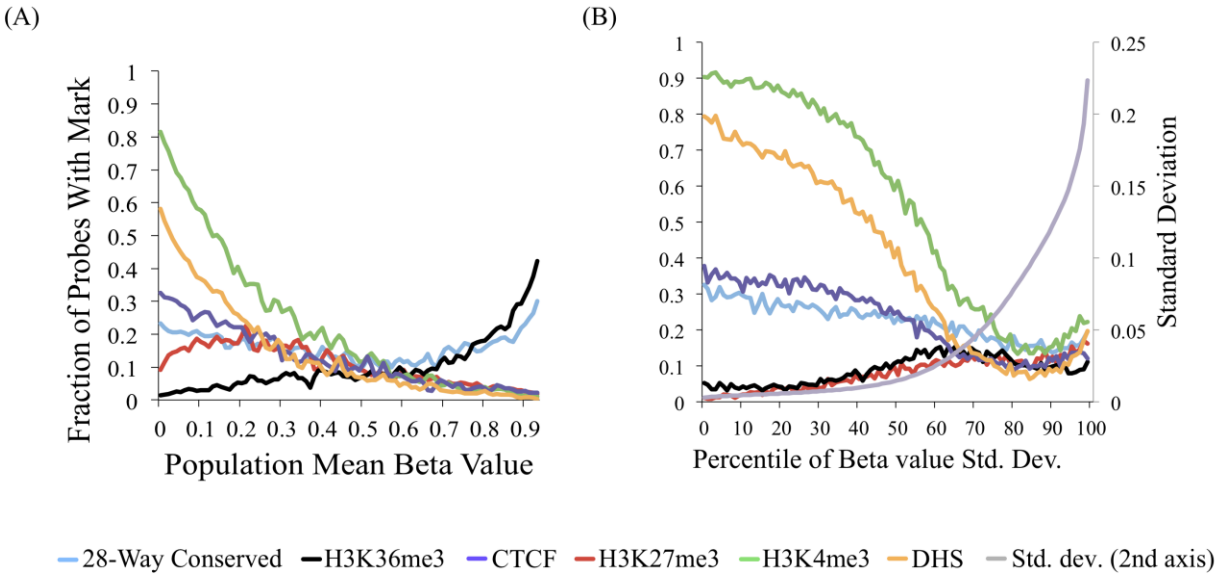

Figure 3.11-3 Proportion of type I CpG probes falling in various types of genomics regions identified by ENCODE.

Values are partitioned by (A) CpG probe mean beta value and (B) percentile of beta value standard deviation. All data types, except for 28-way conservation, are derived from broad peaks in BJ human foreskin fibroblast cells.

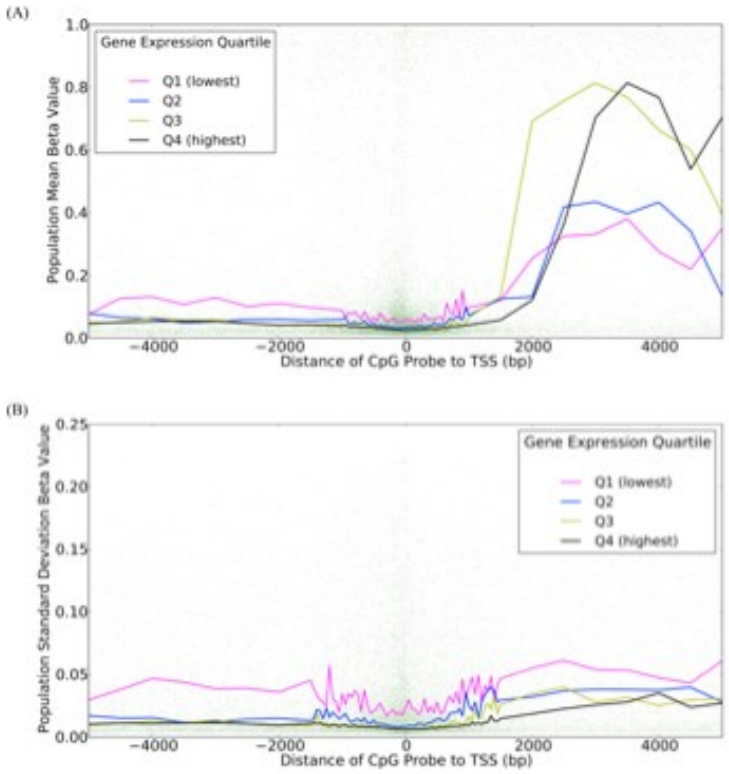

Figure 3.11-4 Mean (A) and standard deviation (B) of type I CpG probes with respect to their position relative to transcription start sites (TSSs) of annotated genes.

Each green dot corresponds to a CpG probe, and the four lines show the running median for probes based on the quartile of the expression level (from RNA-seq in four individuals) of the gene they are associated with.

	Top 25% Variable	Correlated to SNP	Correlated to Expression
Unimodal	78.4%	64.9%	62.7%
Bimodal	19.7%	29.7%	31.5%
Trimodal	1.8%	4.8%	5.7%

Table 3.11-1 Proportion of CpG sites determined to have various numbers of modes in the set of individuals in the present study.

Independently for each CpG site, we applied the kernel smoothing algorithm in R (http://stat.ethz.ch/R-manual/R-patched/library/stats/html/density.html), obtaining a set of 100 bins corresponding to a smoothed distribution of beta values. We then counted the modes, with a mode corresponding to a local maximum in the smoothed distribution having a y-value of at least 1.2 times the average.

Table 3.11-2 Enrichment/depletion of Gene Ontology terms.

Obtained using GoStat (Beissbarth and Speed 2004), for highly variable CpG probes (worksheet 1) and genes with expression correlated to DNA methylation. **For table contents see:** http://genomebiology.com/content/supplementary/gb-2014-15-2-r37-s6.xlsx

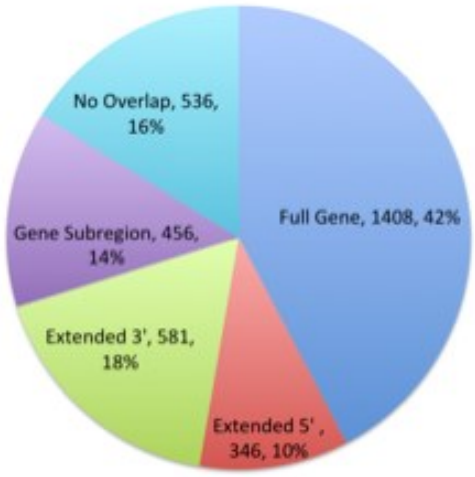

Figure 3.11-5 Overlap of aeRegions with annotated genes.

Table 3.11-3 Set of aeRegions.

For table contents see: http://genomebiology.com/content/supplementary/gb-2014-15-2-r37-s8.txt

Table 3.11-4 Significant mQTL-CpG probe pairs.

For table contents see: http://genomebiology.com/content/supplementary/gb-2014-15-2-r37-s9.txt

Table 3.11-5 Whole genome bisulfite sequencing (WGBS) statistics.

For table contents see: http://genomebiology.com/content/supplementary/gb-2014-15-2-r37-s10.txt

Table 3.11-6 Significant eQTL-Ref8 gene pairs.

For table contents see: http://genomebiology.com/content/supplementary/gb-2014-15-2-r37-s11.txt

Table 3.11-7 Significant aeQTL-aeRegion gene pairs.

For table contents see: http://genomebiology.com/content/supplementary/gb-2014-15-2-r37-s12.txt

Table 3.11-8 Signficant CpG probe-Ref8 gene methylation-expression correlation pairs.

For table contents see: http://genomebiology.com/content/supplementary/gb-2014-15-2-r37-s13.txt

3.12. Epilogue

Our research did not discuss in detail causal relationships between methylation and expression, either in the presence or absence of correlated genetic variation. Inference of causal relationships is an active area of research, for a discussion of some methods and caveats associated with the practice see (Li et al. 2010). Of practical note is recent research (van Eijk et al. 2012) which examined gene expression, DNA methylation and sequence variation in whole

blood. They applied a network edge orientation method based on partial correlations to make inferences about causality. In all cases, correlated SNPs were used as causal "anchors", as variability in the genetic sequences is for most practical purposes the cause of rather than consequence of DNA methylation or gene expression variation. Using these SNPs as anchors, and the resulting partial correlation-based structural equation modelling, hypotheses can be generated about either genetic variation causing expression variation which in turn causes methylation variation (S -> E -> M), or genetic variation causing methylation variation which in turn causes expression variation (S -> M -> E) at a particular genomic region. We applied the network edge orientation algorithm of (Aten et al. 2008), which had also been applied similar datasets in whole blood using the Illumina HumanMethylation27 platform (van Eijk et al. 2012). As noted in the Discussion section, several genes were found to be correlated to expression-methylation QTLs (emQTLs) in both our research and in that of (van Eijk et al. 2012). We also found agreement in causal inferences for these genes that overlapped in our study, with NAPRT1 having highest scores for an S -> M -> E model, and C21ORF56 and WBSCR27 having higher scores for an S -> E -> M model. In the case of GSTM3, we found only weak evidence of an S -> M -> E model, whereas (van Eijk et al. 2012) had found weak evidence of an S -> E -> M model. Clearly additional work, perhaps involving in vitro or in vivo verification is needed to refine the causal inference method and determine if there are indeed tissue specific differences in causal mechanisms or merely issues with the causal model at play.

Removal or otherwise taking into account probes overlapping with SNPs is an area of concern in methylation and expression QTL studies (Ramasamy et al. ; Veyrieras et al. 2012). In our research, CpG probes overlapping with a known SNP were removed. For the allelic expression data, probes were specifically designed to contain a SNP and this would not be expected to be a confounding factor. We had not specifically filtered probes containing SNPs in our eQTL analysis, so endeavoured to do so. Of 446 probes we had found to be mappable

to a SNP, 392 were found in the UCSC Genome Browser hg19 coordinates (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/illuminaProbes.txt.gz). The full list of “common SNPs (141)” was downloaded from UCSC hg19, containing all SNPs with minor allele frequency of at least 1%. Of 392 probes found to be mappable and with hg19 coordinates available on UCSC, 67 were found to overlap with a SNP in the common SNPs list. However, our study had only considered correlations of SNPs with minor allele frequency of at least 10%, whereas many of the SNPs contained within probes had minor allele frequency of less than 10% and thus less likely to be confounders in our eQTL study. Indeed, considering only those SNPS on the Common SNPs list with minor allelic frequency of at least 5%, only 43 mappable probes overlapped with these probes. The number dropped further to 33 when considering those with minor allele frequency > 10%. Loci used for figures or discussion in the paper, such as PAX8, C21ORF56, GSTT and GSTM loci were not found to overlap with any SNP in the list at any minor allele frequency, with the exception of the NAPRT1 gene which was mentioned in the Discussion section as a gene found both by ourselves and by (van Eijk et al. 2012) as mappable to SNPs which were also correlated with methylation (i.e. emQTLs using our notation). The probe for this gene overlaps with rs78452615, a SNP with minor allele frequency of 2.8% and therefore not expected to correlate strongly with the emQTLs in question. We do not expect any major conclusions drawn from our work to change in light of these findings.

Chapter 4. DNA co-methylation and tissue composition effects in human adipose tissue

4.1. Preface

Results of the previous chapter's research had demonstrated the complex patterns of inter-individual variation and co-variation with respect to DNA methylation, genetic variation, gene expression and chromatin. We had noted that CpG sites located on different chromosomes showed similar patterns of DNA methylation loci such as the HOX clusters in fibroblasts, and sought to characterize this DNA methylation covariation more systematically in a dataset with a larger sample size. Fortunately, DNA methylation data for adipose tissue in 581 samples recently published by our collaborator (Grundberg et al. 2013) afforded us exactly this opportunity. DNA co-methylation modules obtained from these adipose data were rich in many fascinating biological properties related to constituent cell types of adipose tissue but depleted for correlation to genetic variation in cis or trans.

These results together with knowledge of adipose tissue biology and its altered tissue composition in obese individuals led us to investigate further ways to correct for these tissue composition effects without reference to either methylation levels of constituent cell types or tissue composition of the adipose samples studied, neither of which were available. We developed a deconvolution approach taking as input only a parameter k indicating the desired number of cell types to infer and the matrix of methylation values for the study. This approach will infer the tissue composition levels in each cell type as well as the tissue composition of each sample. The residuals of these levels can be taken to obtain

the variation still present in the population when correcting for tissue composition effects.

In this preface, I review the main steps of the WGCNA clustering approach that is used in our work to find co-methylation modules, and then review other methods used for dealing with tissue composition and other sources of variance in high throughput genomic studies. More details on our specific deconvolution approach can be found in the methods section of this chapter.

4.1.1. WGCNA

To identify modules or clusters of co-methylated CpG sites from the Illumina HumanMethylation450 experiment, we used the Weighted Gene Co-expression Network Analysis (WGCNA) package in R (Langfelder and Horvath 2008), a program that has yielded hundreds publications in peer-reviewed journals exploring modules of co-methylation or co-expression. (Langfelder and Horvath 2008) identify 5 key steps that can be taken.

a) Construct a gene co-expression network using correlations between genes and topological overlap (shared neighbors) in the derived network

b) Identify modules using hierarchical clustering and a Dynamic Tree Cut

c) Relate modules to external information (clinical Data, SNPs, proteomics, Gene Ontology)

d) Study module relationships

e) Find the key drivers in interesting modules.

Steps c) through e) of these analyses were adapted to our specific needs in characterizing co-methylation modules obtained from a complex tissue such as adipose. The methods used and results obtained for these steps are detailed in the following sections of this chapter. I present here the computational and

statistical approaches used in a) and b), which relied on existing implementations of the WGCNA package. Choices of parameters and options offered by this software for carrying out steps a) and b) are considerable, and I outline only those used in the analyses done in this chapter. For further detail see (Langfelder and Horvath 2008), (Zhang and Horvath 2005), as well as http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/ and references cited therein. In order to maintain consistency with methods papers which utilize gene expression/co-expression as the property of interest, I refer to the units measured as 'genes' or 'expression probes' and the measurements 'expression levels'. In the analyses actually performed, the units measured are CpG sites or CpG probes in the Illumina HumanMethylation450 Platform, and the measurements are methylation beta values.

The input for step a) is a matrix of gene expression measurements for the samples or individuals in a given study. The output of a) and the input for b) is a gene co-expression network. The output of b) and what the researcher will study further will be a module assignment for each gene in the study, with the possibility that a given gene will not be assigned to any module. Other secondary properties can be obtained such as an "eigengene" summarizing the expression profile of each module.

Gene co-expression network construction

In experiments consisting of tens or hundreds of thousands of expression probes, a recommended first step, which we took advantage of, is to pre-assign probes to blocks of approximately 5000 probes each. A simple k-means with Euclidean distances approach is used to assign probes to blocks, with each block corresponding to a medoid output from the approach. Pairwise correlations will only be obtained between probes assigned to the same block, considerably reducing the total number of such correlations that need to be obtained. For each block, the correlation coefficient between each pair of probes is determined. Which correlation method to use is one parameter of this method and we choose

to use the biweight midcorrelation method, an approach robust to outliers, with formula and justification for its application given in (Song et al. 2012). Correlation coefficients are typically raised to a power greater than 1, such that those with values close to 0 will be pushed even closer to 0 compared to those with an absolute value close to 1. The power coefficient to use is also a parameter that can be set by the user. (Langfelder and Horvath 2008) recommend making this choice based on a power coefficient that generates a scale free network in terms of the distribution of vertex degrees. We use a power coefficient of 12 for the modules analyzed here, a relatively strict threshold to ensure a manageable number of modules, each relatively consistent in terms of various measured properties.

If the probes and the correlation coefficients are modelled as a network, each block can be regarded as one component of the network, with the vertices corresponding to probes and the edge weights corresponding to correlation coefficients raised to power 12. No edges are removed regardless of how close to 0 their weight becomes; therefore, each component is also a clique. In practice, utilizing a measure of topological overlap was found to give better results in terms of the modules, and this score is used as the edge weight in the network. We now move towards the identification of modules within each block, and the merging of similar modules between blocks.

Module identification

The topological overlap serves as an edge weight or measure of similarity between probes in the same block, and a matrix of pairwise similarities is encoded in the topological overlap matrix (TOM). 1 – TOM is the distance matrix between probes, and is used as input for a standard R hierarchical clustering algorithm (hclust). Each block is now represented as a hierarchical clustering tree, with each leaf corresponding to a probe. The Dynamical Tree Cut (Langfelder et al. 2008) is then applied to obtain a module assignment. This approach cuts the tree at different levels and assigns each group of probes to the

same module. An eigengene or eigenprobe summarizing the expression profile of each module is then obtained by performing a singular value decomposition or principal components analysis and describing the eigenprobe as the eigenvector of the first principal component. Modules in different blocks are merged depending on the similarity of their eigenprobes.

4.1.2. Correcting for tissue composition in high throughput experiments

Introductory remarks: Principal Components Analysis and Singular Value Decomposition

Principal components analysis (PCA) and the singular value decomposition (SVD) form the basis of several approaches discussed in this preface and widely used for summarizing variation levels. These variation levels can in turn be used for further association studies, or removed if believed to be the result of confounding variables. In short, in SVD, the *m x n* matrix M of gene expression measurements for *m* probes and *n sample*, is decomposed as follows: $M = U\Sigma W^T$ where Σ is an *m x n* diagonal matrix of singular values and W^T is an *n x n* matrix referred to as the the set of right singular vector, which can also be called the “eigengenes” in the context of a gene expression experiment, with the ones corresponding to the largest singular values accounting for the largest proportions of variance in the gene expression matrix (Alter et al. 2000; Leek and Storey 2007).

Surrogate Variables Analysis (SVA)

SVA (Leek and Storey 2007) is one of the most widely applied methods for correcting for unknown sources of variation and heterogeneity in a high throughput biological experiment. The model used consists of an expression matrix X consisting of *n* samples and m genes, and y, a vector of length n corresponding to an experimental variable of interest (eg case/control). The

model of a gene's expression as a function of the variable of interest is described as $x_{ij} = \mu_i + f_i(y_j) + e_{ij}$, where μ_i is the baseline level of expression, $f_i(y_j)$ is a function of the covariate of interest's effect on expression of gene i, and e_{ij} is random noise. Though this random noise is assumed to be independent between samples, unmodelled factors can potentially introduce dependence between noise levels. $x_{ij} = \mu_i + f_i(y_j) + \Sigma_{l=1}^{L} \gamma_{li} g_{li} + e_{ij}^*$ gives the relationship between primary variable of interest y and gene i, with $g_l = (g_{l1}, \ldots, g_{ln})$ a function of an unmodelled factor l, of which there are a total of L and γ_{li} is a gene-specific coefficient for the lth unmodeled factor. Explicitly including unmodelled factors that can affect expression of multiple genes, the noise factor can be modelled as one that is truly independent between genes, and hence is indicated in this model as e_{ij}^*. γ_{li} and g_{li} are replaced as mutually orthogonal vectors h_k and coefficients λ_{ki} such that $\Sigma_{l=1}^{L} \gamma_{li} g_{lj} = \Sigma_{k=1}^{K} \lambda_{ki} h_{kj}$ and $x_{ij} = \mu_i + f_i(y_j) + \Sigma_{l=1}^{L} \lambda_{li} g_{lj} + e_{ij}^* = \mu_i + f_i(y_j) + \Sigma_{k=1}^{K} \lambda_{ki} h_{kj} + e_{ij}^*$

The number l of unmodelled factors is determined via an approach whereby a residual matrix R is obtained by subtracting the effect of the primary variable on expression. A singular value decomposition is done, obtaining a set of n eigengenes. For each eigengene, the proportion of variance explained by that eigengene is determined. A null model is obtained by permuting the rows of R many times, repeating the singular value decomposition on each permuted version, and across all permutations, determine if this eigengene explains an equal or greater level of variance by chance with a frequency of at least alpha, a user defined parameter. The number of eigengenes passing this test and explaining this level of variance by chance is set as the number of surrogate variables.

Each significant eigengene, $\mathbf{e}_k$ is regressed on each gene x_i $(i = 1, \ldots, m)$ and a p-value obtained for the association. These p-values are used as the basis of a formula described in (Storey and Tibshirani 2003) to determine $\hat{\pi}_0$, an estimate of the proportion of genes with expression not truly associated with $\mathbf{e}_k$.

. $\hat{\mathrm{m}}_{\mathrm{i}} = (1 - \hat{\pi}_0) \times \mathrm{m}$ is the number of genes associated with this residual eigengene. A reduced expression matrix consisting only of the $\hat{\mathrm{m}}_{\mathrm{i}}$ genes most associated with this eigengene is obtained, and are expected to represent the expression of those genes containing the heterogeneity as represented by some value h_k. The eigengenes of this reduced expression matrix are calculated and the one most correlated with the residual eigengene is obtained, and used as one of the surrogate vectors h_k moving forward.

Though the empirical quality of results obtained via SVA is undeniable, and it has been used for a variety of expression and methylation datasets in many contexts, the transformation done with orthogonal surrogate variables rather than true variables results in a rather opaque data analysis. Typically, corrected values are used without further analysis. We develop a method that estimates tissue composition vectors per individual and mean methylation beta values per cell type that are readily interpretable and can be considered by the user in downstream data analysis steps (for example correlation between a given cell type's inferred composition proportion and genotype or phenotype) if so desired.

FaST-LMM-EWASher

FaST-LMM-EWASher, or factored spectrally transformed linear mixed model 'EWASher' (Suganuma and Workman 2011) is designed to correct for cell heterogeneity effects in the context of Epigenome Wide Association studies. It utilizes a linear mixed model and principal components analysis. The formula for a given phenotype y, matrix of methylation values G, number of probes M, a vector of methylation values at a given probe X_s, a vector of known covariates X is a vector of known covariates, random effects u, and fixed effects β and β_s:

$$y = X\beta + X_s\beta_s + \frac{1}{\sqrt{M}}Gu + \epsilon$$

This model is fit for each probe in the methylation study. If a tissue composition effect is present, it is expected to act as a confounding variable and inflate the number of significant effects present when fitting this model. These confounding effects can be expected to be modeled as a set of L top principal components that are a source of variation across many probes. With L principal components, the model is written as $\square = X\beta + X_s\beta_s + \Sigma_{i=1}^{L} A_i\lambda_i v_i + \frac{1}{\sqrt{M}} Gu + \epsilon$. Testing is done with various values of L and the smallest one that yields genomic control factor (Bacanu et al. 2000) close to 1 is chosen, indicative that most of the spurious associations between the trait of interest and the tissue composition have been removed.

This approach was designed specifically for correlating epigenetic measurements such as DNA methylation with a single phenotype of interest. We were interested in developing an approach that will return DNA methylation values representative of the cell type-specific inter-individual methylation variation present at a given probe which can then be used as a general purpose residual vector for correlation with genetic variation, gene expression, methylation at other loci, or a phenotype of interest.

PEER

PEER (Lee et al. 2010) utilizes a Bayesian model to infer various parameters for a gene expression matrix. The probability of an observed gene expression value in gene *g* and individual *j* is given by:

$$P\left(y_{g,j}|y_{g,j}^{(1)}, y_{g,j}^{(2)}, y_{g,j}^{(3)}, \tau_g\right) = Normal(Y_{g,j}|y_{g,j}^{(1)}, y_{g,j}^{(2)}, y_{g,j}^{(3)}, \frac{1}{\tau_g})$$

Where each of $y_{g,j}^{(1)}$, $y_{g,j}^{(2)}$ and $y_{g,j}^{(3)}$ correspond to models for a genotype effect, known factor and hidden factor model, and τ_g is a Gaussian noise variable with a gamma prior. Inference of parameters is done via variational Bayesian learning (Jordan et al. 1999), a generalization of the Expectation Maximization Algorithm.

Though utilizing a distinct approach to that of SVA, hidden factors are treated in a somewhat opaque factor by VBQTL and not straightforwardly mappable to potential real confounding effects like tissue composition effects. It is potentially a very useful general purpose algorithm for removing unwanted sources of variation in an experiment but was not designed specifically with tissue composition effects in mind.

Reference Free EWAS

A method for correcting for DNA tissue composition effects with reference to measurements was developed by (Gibbs et al. 2003). This method was expanded to infer cell specific effects in (Creyghton et al. 2010). This research was specifically developed to address concerns in an EWAS study, in which there is a phenotype of interest to be correlated with measurements of an epigenomic mark (in this example DNA methylation) at various sites in the genome. The standard model used in an unadjusted EWAS analysis in a study with m CpG sites, n individuals, and p covariates is $Y = B^*X^T + E^*$, where B^* is an *m x p* matrix of coefficients identifying the effect of a given covariate on methylation, Y is an *m* x *n* matrix of DNA methylation measurements. X is an *n* x *p* matrix of covariates and E* is an m x n matrix of errors. A refined model taking into account cell type specific effects mediated by the covariates would be as follows: $Y = BX^T + M\Omega^T + E$, where Ω is an *n* x *k* matrix of subject-specific cell proportions for k cell types, which itself is affected by covariates in the following model: $\Omega = X\Gamma + \Phi$, where X is the same matrix of covariates, Γ is a *p* x *k* coefficient matrix representing cell-proportion effects, and Φ is an *n* x *k* matrix of errors.

Taking into account the tissue composition effects, B is an *m* x *p* matrix of direct epigenetic effects (not mediated by effects on cell type), and M is an *m* x *k* matrix of cell-specific mean methylation values. The goal of the adjusted EWAS analysis is to estimate the direct effects matrix B. This is also done in a similar fashion to that of surrogate variable analysis, where an estimate of the optimal

dimension of tissue composition effects to subtract is estimated, in this case via random matrix theory, and surrogate variables are obtained after fitting the unadjusted model, performing singular value decomposition on the residuals and selecting d surrogate variables.

Like FaST-LMM-Ewasher, Reference Free EWAS was principally designed in the context of estimating the true relationship between DNA methylation and a phenotype of interest, by taking into account tissue composition levels that are also potentially affected by the phenotype. It does not give as part of its output a matrix of methylation values corrected to remove tissue composition effects but with cell type-specific methylation variation remaining.

4.2. Abstract

High throughput measurements of DNA methylation developed in recent years have led the development of epigenome wide association (EWAS) studies whose goal is to identify loci whose methylation levels correlate with a phenotype of interest. At a population level, DNA methylation and genotyping or sequencing data can also be integrated to find sequence variants correlating with methylation of CpG sites in cis or trans. Together with gene expression and epigenomic measurements, it is hoped these studies will better elucidate genetic and epigenetic mechanisms of transcriptional regulation and variation present both in healthy populations and in a disease of interest. Though DNA methylation analysis of readily extractable human tissues such as blood or adipose tissue offers an attractive opportunity, studies have elucidated more and more clearly that these measurements are subject to tissue composition effects, leading to potential confounding in EWAS studies and under-estimation of genetic effects in methylation quantitative trait loci (mQTL) studies. We report in this study the characterization of a set of CpG co-methylation modules from female human

adipose tissue samples, derived from 581 samples collected by the MuTHER Consortium as part of the TwinsUK study. We examine various lines of genomic and epigenomic evidence indicating these co-methylation modules, each comprised of hundreds to thousands of CpG sites distributed across all autosomes, are driven by tissue composition effects. We then introduce and apply an unsupervised cell mixture deconvolution method to attempt to infer tissue compositions of the individuals in the study and methylation profiles of constituent cell types of adipose. Residuals are then obtained from these inferred values to obtain estimated DNA methylation profiles of the individuals, with correction for tissue composition effects. These residuals are found to have improved mappability to mQTLs, a higher proportion of cis correlations when considering expression-methylation or methylation-methylation relationships, and a shifted landscape regarding the categories of genes represented when performing an EWAS study with BMI as the phenotype of interest. We posit that this work and our deconvolution method will be an important contribution to future methods that utilize DNA methylation together with various genomic measurements and phenotypes in samples drawn from complex tissues.

4.3. Introduction

DNA methylation has a long history of being studied as an epigenetic phenomenon, with earliest experiments shedding light on its role in female X chromosome inactivation (Riggs 1975) and genomic imprinting (Li et al. 1993). Another long studied feature of DNA methylation has been its increased presence at inactive promoters, and aberrant patterns of DNA hypo-methylation at oncogenes and, to a lesser extent, DNA hyper-methylation at tumour suppressor genes have been well studied as important cancer biomarkers (Baylin et al. 1998).

Recent experimental methods and platforms interrogating DNA methylation at a genome-wide scale, often in multiple individuals, have

elucidated new and exciting insights into the roles and mechanisms of DNA methylation (reviewed in (Jones 2012)), with genetic (Bell et al. 2011), epigenetic (Wagner et al. 2014) and environmental (Tobi et al. 2009) factors, and its use as a biomarker or treatment target in diseases such as cancer (Lengauer et al. 1997), (Amato 2007).

One key piece of the puzzle not yet considered is the question of how CpG sites co-vary across the genome, and in what genomic and epigenomic context this co-methylation takes place. This type of work can be expected to generate hypotheses regarding mechanisms of DNA methylation in the tissue studied, demonstrate a baseline level of variation present if the tissues are sampled from a healthy population, and generate hypotheses of changes that take place in a disease or other phenotype of interest.

Research into patterns of co-methylation variation across the genome, particularly in untransformed cell lines derived from healthy individuals, is still in its infancy. To date, most approaches rely on Illumina's 27K or 450K methylation arrays, which measure methylation at a corresponding number of (non-uniformly distributed) CpG sites in the human genome. An approach commonly used involves measuring methylation levels at a fixed set of sites in a set of cell lines following a case-control paradigm, finding co-methylation modules via the weighted gene correlation network analysis (WGCNA) approach, (Zhang and Horvath 2005), and finally then finding key modules with a strong differential methylation for the feature of interest. This approach was used by Busche et al. (Busche et al. 2013), who applied WGCNA to a set of pre-B ALL tumour samples, focusing on the genomic features of a module identified with substantial differential methylation in the subset containing a t(12;21) translocation. Relation between methylation and age has been studied by Bocklandt et al. (Bocklandt et al. 2011), who applied WGCNA to methylation levels at 450K sites in saliva samples to obtain a module with methylation levels substantially correlated with age, identifying in particular two probes capable of building a regression module explaining 73% of the variance in age. Along the same lines, Horvath et al.

(Horvath et al. 2012) employed WGCNA in a meta-analysis across multiple Illumina 27K and 450K datasets in brain and whole blood, identifying a preserved module with a correlation to age. The relation between co-methylation and co-expression modules was investigated by (van Eijk et al. 2012) in whole blood samples, who applied WGCNA separately to identify expression modules and methylation modules, then integrating these results, finding stronger signs of co-expression than co-methylation, and that, with noteworthy exceptions, overlaps between these two types of module are rare. (Akulenko and Helms 2013) measured DNA methylation in breast cancer tumour cells, also using the 27K platform, then obtained pairwise Pearson correlations between CpG sites, finding functional similarity between adjacent genes of correlated probes; clustering of CpG sites using affinity propagation also yielded clusters corresponding to gene groups of functional significance.

These studies considered results obtained from clustering or co-methylation networks at the level of enrichments for particular Gene Ontology (GO) or KEGG pathways, often limiting consideration to genes whose promoter or body overlaps with a modular CpG probe. We hypothesize that while these considerations are important, there is other vital information about the nature and impact of CpG co-methylation to be learned that goes beyond the promoter or bodies of genes. In particular, consideration of more distal CpG sites, correlation with gene expression and sequence polymorphisms in cis and in trans, chromatin features, and DNA sequence motifs leads to a richer understanding of the baseline levels of DNA methylation variation and co-variation present even in sets of untransformed cell lines derived from healthy individuals.

We report an in-depth characterization of co-methylation modules found in a set of adipose tissue samples derived from sets of twins studied as part of the Multiple Tissue Human Expression Resource (MuTHER) project (Nica et al. 2011), which is in turn part of the ongoing longitudinal study carried out by TwinsUK (Spector and Williams 2006). We found various lines of evidence contributing to the hypothesis that these co-methylation modules were driven in

large part by tissue composition effects in adipose tissue. Probes assigned to modules in our study were depleted for correlations to methylation QTLs (mQTLs). Correcting for or taking into consideration tissue composition effects was deemed to be imperative for fully understanding the DNA methylation variation present and its mappability to sequence variation in this tissue.

DNA methylation data for each of each constituent cell type of adipose tissue are not presently available, and for those for which it is available, it may not be expected to match that seen in the cell type in adipose tissue, or may have been subject to purification or culturing effects before measurement of methylation could take place.

We introduce a novel tissue composition deconvolution method to infer tissue composition levels and DNA methylation profiles of constituent cell types in adipose. Residuals obtained from these results can be regarded as DNA methylation levels corrected for these tissue composition effects. Based on results measured with published methylation data in whole blood and preadipocytes/fibroblasts we were able to find components corresponding to methylation levels in constituent cell types of adipose tissue.

In our analyses, these residuals were found to have improved correlation to genetic variation, with gains in correlation being much stronger when considering cis-relationships, and when involving probes assigned to modules in our original dataset, indicating an ability of our approach to correct for variation due to cell specific effects and better capture variation due to other sources such as mQTLs. Repeated analyses with WGCNA revealed much smaller co-methylation modules than before, with many modules disappearing and others being much smaller. The proportion of cis associations for both methylation-methylation and methylation-expression pairs increased considerably. Finally, the landscape of BMI-associated CpG sites was transformed from one in which immune related CpG sites predominated, putatively driven by macrophage infiltration, to one of CpG sites with genomic relationships to putative AP-1

binding sites, active regions of fibroblasts, and genes involved in wound healing; this transformed landscape is indicative of one in which obesity is associated with a fibrotic response and considerable DNA methylation in the fibroblast component of adipose tissue.

In summary, our work with adipose tissue derived from individuals studied by the MuTHER consortium represents one of the most extensive characterizations of co-methylation modules and their epigenomic and genomic contexts. In an effort to correct for tissue composition effects on our methylation measurements, we developed a method able to reveal stronger correlations to genetic variation as well as methylation-methylation and methylation-expression relationships expected to be present within component cell types. Together, results of co-methylation analyses of complex tissues done before and after tissue composition correction provide valuable insights and generate further hypotheses regarding the constituent cell types and their epigenetic similarities and differences.

4.4. Results

4.4.1. Datasets analyzed

We performed an integrated analysis of DNA methylation, SNP genotyping and gene expression data of adipose tissue samples from a cohort of 581 females, of which a total of 200 are derived from monozygotic twin pairs, 288 from dizygotic pairs, and 93 were lone individuals, generated as part of the ongoing experiments carried out by the MuTHER Consortium and first reported by (Grundberg et al. 2013). The samples were partitioned in two subsets, adipose-1 (*n*=290) and adipose-2 (*n*=291). Samples ranged in age from 39 to 85 and in BMI from 16 to 47. Samples were genotyped using a combination of Illumina platforms as described in (Grundberg et al. 2012; Grundberg et al. 2013) and assayed for gene expression (using the Illumina HumanHT-12 V3.0 expression BeadChip) and genome-wide DNA methylation (using Illumina

Infinium HumanMethylation450 arrays). Expression data were quantile normalized and corrected for batch effects using ComBat (Johnson et al. 2007), and probes overlapping known SNPs were excluded. For methylation data, type I and type II CpG probes were separately quantile normalized and corrected for batch, row, column, chip and plate effects using ComBat (Johnson et al. 2007), and probes overlapping known SNPs were excluded. The 186,194 methylation probes with the highest variance across the 581 samples (i.e. 50% top variable) were utilized for further analysis.

4.4.2. Identification and characterization of co-methylation modules

As noted previously (Grundberg et al. 2013; Gutierrez-Arcelus et al. 2013; Wagner et al. 2014), methylation shows substantial correlation with gene expression in populations, with positively and negatively correlated probes showing significant overlaps with specific chromatin marks. In many cases methylation variation can be explained by genetic variation, but this is only a partial explanation and many CpG sites with variable methylation levels show little to no association with genetic variation in cis. We sought to move beyond cis correlations of methylation with expression and genetic variation to a more global scale by seeking systematic patterns of co-methylation across the genome.

We first characterized the structure of the observed inter-individual variations in methylation levels. To this end, we used the WGCNA R package (Langfelder and Horvath 2008) to identify groups of CpG sites, called methylation modules, whose methylation co-varies across individuals. WGCNA has been used with success in the past to identify co-expression (Presson et al. 2008) and co-methylation modules (Horvath et al. 2012) in various large-scale data sets. Modules were identified separately in each the adipose-1 and adipose-2 subsets, with a power coefficient of 12 and minimum module size of 50 probes. Modules

with a high level of agreement between adipose-1 and adipose-2 were intersected (Table 4.7-1), and probes were reassigned or removed from modules where appropriate to ensure only those positively correlated with each other were assigned to a single module. Some basic properties of the 10 final adipose comethylation modules are reported in Table 4.4-1.

Module Number	Number of Probes	Average Standard Deviation	Average Pairwise R	Gene Body Probes (%)	TSS Adjacent Probes (%)
1	6173	0.0328	0.539	32.5	42.9
2	4916	0.0301	0.534	44.6	25.7
3	6064	0.0324	0.503	49.5	26.9
4	4964	0.0326	0.555	45.0	34.3
5	2312	0.0320	0.564	40.0	32.8
6	1521	0.0291	0.536	59.2	26.5
7	1273	0.0324	0.520	38.4	39.0
8	534	0.0336	0.519	47.2	24.9
9	160	0.0362	0.623	43.8	18.8
10	259	0.0364	0.599	50.2	17.8

Table 4.4-1 Basic properties of modules used in analyses.

Notably, most modules contain probes on all autosomes and these probes rarely cluster in the genome. In fact, of the 41.4 million pairs of probes with methylation-methylation R^2 above 0.36 (which corresponds to a <1% FDR), 99.6% are in trans (i.e. from different chromosome or more than 1 Mb apart on the same chromosome). Similarly, 99.6% of the correlated CpG methylation/gene

expression pairs (R^2 > 0.123, corresponding to a 5% FDR) occur in trans. This suggests that these correlated pairs are not causative but rather associated with some common source of external variation.

We then determined the extent to which methylation at CpG sites could be associated to genetic variants (mQTL), focussing on SNPs. Pearson correlation coefficients were obtained between each pair of SNPs and probes, independently for adipose=1 and adipose-2, and the minimum of the two R^2 values was retained for each CpG-SNP pair. A permutation test was used to determine that an R^2 of 0.1125 corresponds to a 5% false discovery rate. In total, 22,281 CpG sites mapped to at least one SNP. Remarkably, modular probes are more than four times less likely to be mappable to an mQTL, in either cis or in trans, than other highly variable non-modular probes (12.8% of non-modular probes vs 3.1% of modular probes are mappable). Again, this suggests an external cause of methylation variation at modular probes that muddies the association between methylation and genetic variants.

4.4.3. Co-methylation modules associate to cell-type specific methylation

We set out to better characterize the genomic and epigenomic properties of member probes of each of the major modules found in adipose tissue. In all cases, the properties of modular probes were contrasted against a background set consisting of the 50% most variable probes. Properties evaluated included:

Similarity to other methylation profiles. Methylation levels have been previously measured in various cell lines and tissues related to some of the expected constituents of adipose tissues, including pre-adipocytes, whole blood and dermal fibroblasts (Nazor et al. 2012). We thus assessed the overlap between probes from each module and sites that are hypo-methylated (beta < 0.3), i.e. putative active regulatory regions, in each of these samples. See

Figure 4.4-1.

Cell-type specific chromatin state. Histone modification data obtained by Chip-Seq, as well as DNase I Hypersensitivity assays, available from ENCODE (Myers et al. 2011) and Epigenomics RoadMap (Chadwick 2012) provide a rich perspective on cell-type specific regulatory regions. We measured the overlap between modular probes and these regions and assessed enrichment relative to background. See Figure 4.4-2.

Associated gene function enrichment. We used GREAT (McLean et al. 2010) (with default settings) to quantify enrichments for categories of genes overlapping or adjacent to modular CpG sites. Each of the eight largest adipose modules showed enrichment for GO terms linked to key functions of constituent cell types of adipose tissue.

DNA sequence motifs. We sought to determine if probes in a given module were co-located with specific sequence motifs, which may be binding sites for DNA binding proteins that would be modifying or modified by DNA methylation at those sites. For each module, motifs enriched in the 100bp flanking modular probes compared to background set were identified and classified using Homer (Heinz et al. 2010). See Figure 4.4-3.

Correlation to gene expression. Each CpG site's methylation level was correlated to the full set of gene expression measurements. With methylation tending to be a mark of inactivity or repression with respect to gene transcription, modules whose methylation level negatively correlates with expression of a gene category would be expected to correspond to these probes more often corresponding to enhancer or promoter regions of cell types in which this category of genes are actively expressed. Gene ontology enrichments for each module were obtained using Ontologizer (Bauer et al. 2008), using as foreground the set of expression probes with a negative correlation level (5% FDR) to at least one member probe in that module, and using as background the full set of expression probes correlated negatively at 5% to at least one methylation probe in the top 50% variable 450K set, modular or non-modular. (See Table 4.7-2)

Correlation to BMI. Obesity is associated with macrophage infiltration and otherwise altered landscape of tissue composition in adipose tissue. Body-Mass Index (BMI) measurements were correlated with methylation of each modular probe as well as each eigenprobe.

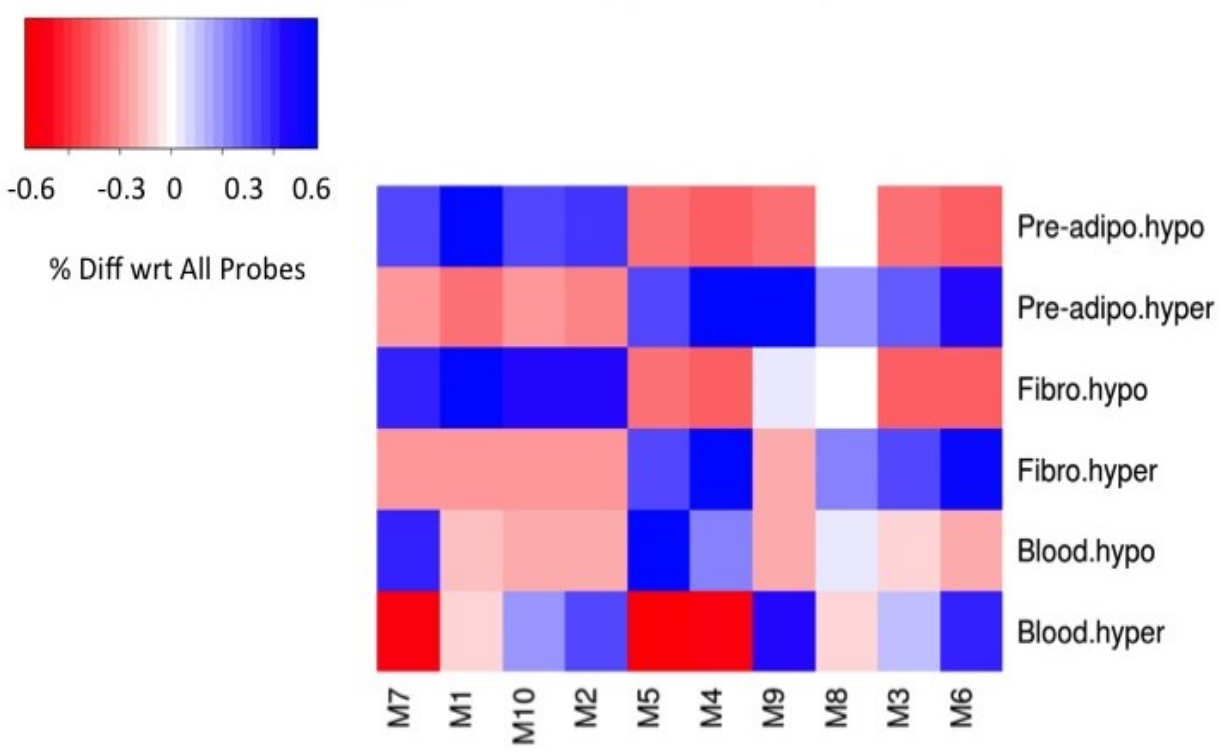

Figure 4.4-1 Module probes are enriched for hypo- and hyper-methylated probes in cell types related to constituent cell types of adipose

For each module, the colouring is based on the difference between the percentage of modular probes that are hypermethylated (Beta > 0.7) (resp. hypomethylated (Beta < 0.3)) in the given cell type (pre-adipocytes, fibroblasts and whole blood) and the corresponding percentage among the set of all top variable probes.

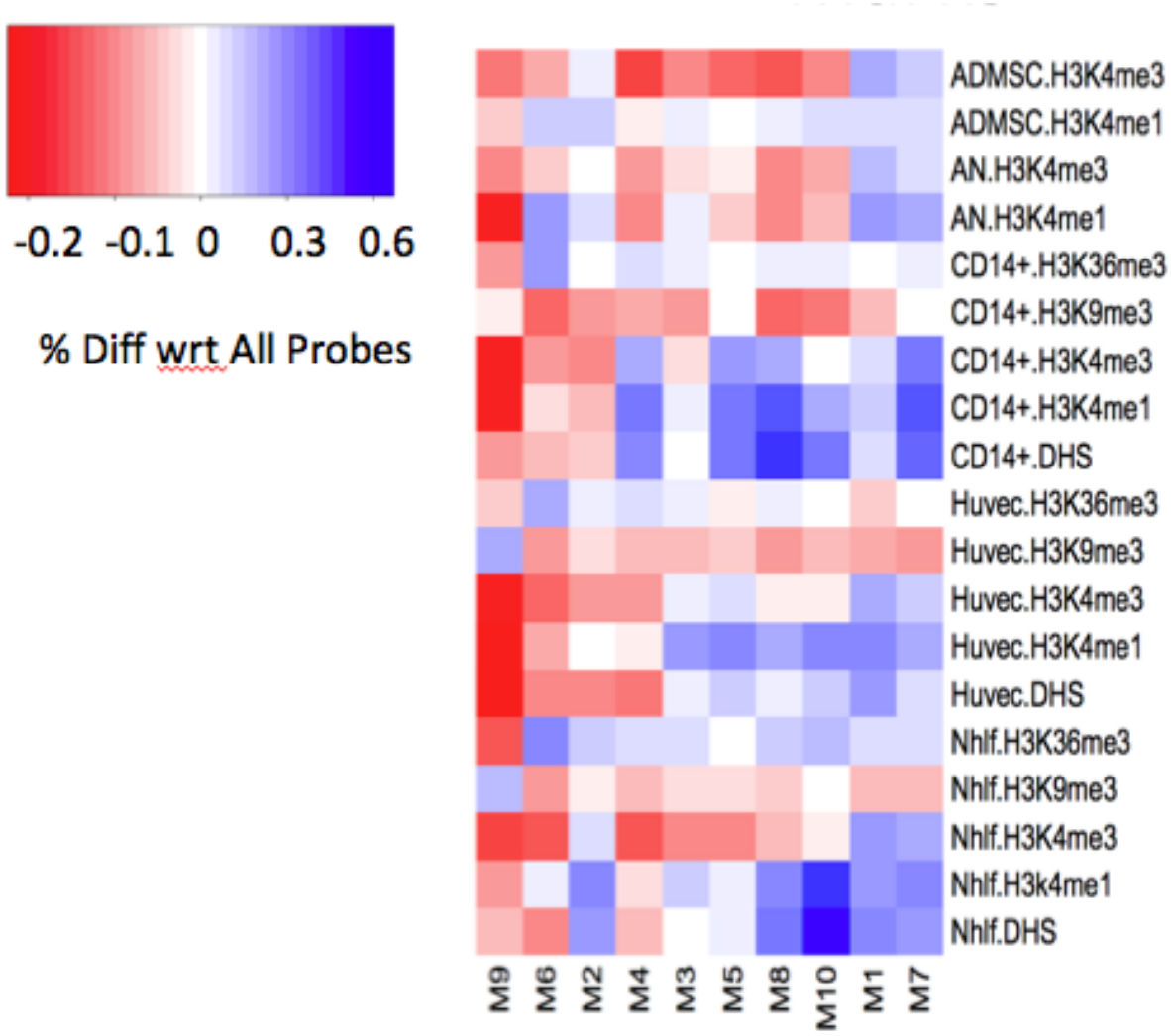

Figure 4.4-2 Modules are enriched and depleted with respect to histone marks and DNase I hypersensitivity (DHS).

For each module, the colouring is based on the difference between the percentage of modular probes that overlap with a peak for the given cell type and the corresponding percentage among of all top variable probes. Cell types: Nhlf: Normal Human Lung Fibroblast; Huvec: Human Umbilical Vein Endothelial Cells; AN: Adipose Nuclei; ADMSC: Adipose Derived Mesenchymal Stem Cells.

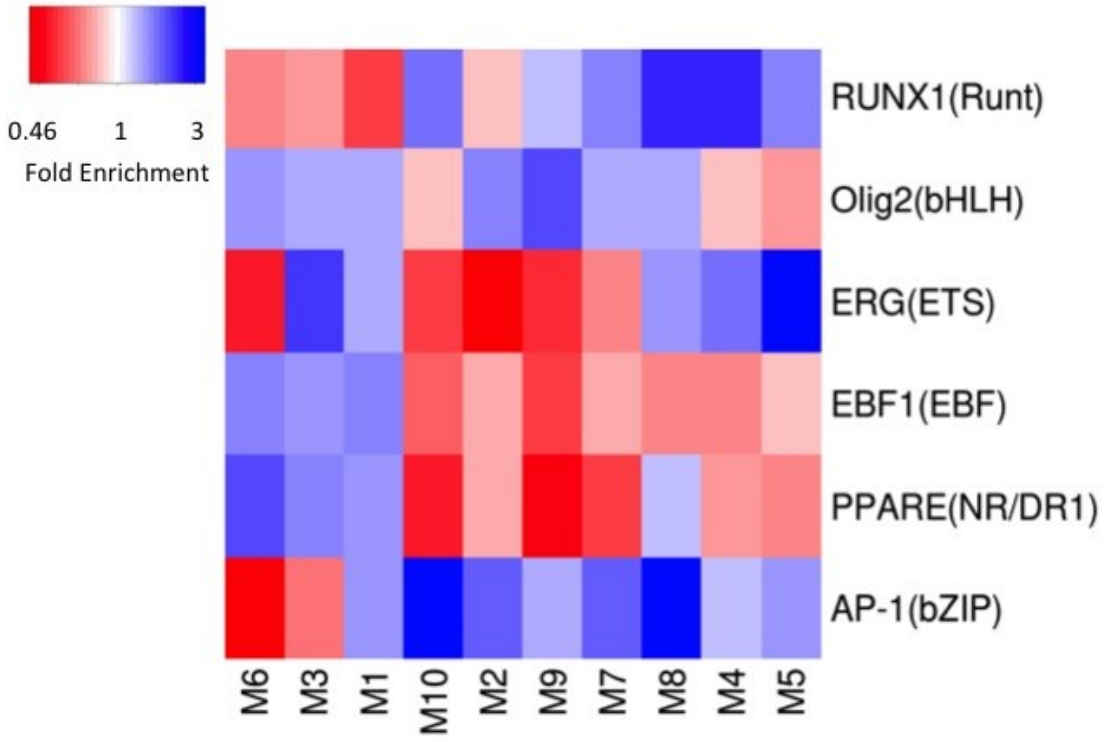

Figure 4.4-3 Modules are enriched and depleted with respect to transcription factor binding motifs in their neighbourhood.

Colorings for a given module correspond to the fold enrichment of the percentage within 100 bp of a motif for the transcription factor when compared to the proportion of all top variable CpG sites that are within 100 bp of that motif. Results were analyzed with HOMER (Heinz et al. 2010). Enrichments for AP-1 in Module 8 and Module 10 were even greater than 3, but were capped at this number for better color gradation.

Properties characterizing each module generally tend to point to a regulatory role in one or more of the key constituents of adipose tissue. The probes in the largest module, Module 1, tend to be associated to genes involved in adipocyte differentiation, with strong enrichment for GO terms such as Rho GTPase function, actin binding and beta catenin binding (Q values respectively: $2.3*10^{-7}$, $9.94*10^{-7}$, $2.9*10^{-3}$) all categories of genes related to signalling in adipocyte differentiation and adipose tissue formation (Cristancho and Lazar 2011). They are enriched for marks of enhancer and promoter activity in adipose nuclei and fibroblasts, for hypo-methylation in pre-adipocytes and fibroblasts, and negative methylation-expression correlation with genes of the extracellular space, also indicative of fibroblast related activity in regions of probes of this module. Conversely, they were depleted for hypo-methylation in whole blood, suggestive of a module consisting of active marks in the fibroblast-adipocyte lineage as opposed to blood cells.

Module 4, whose eigenprobe showed a strong anti-correlation to that of Module 1, complemented these results. Genes associated to these probes were often involved in leukocyte migration and activation, inflammation and immune response. These probes were also enriched for marks of regulatory activity in macrophage precursor CD14+ cells, and showed hypo-methylation in whole blood. These results are consistent with the observed enrichment for nearby binding motifs for the Runx1 transcription factor (2.1-fold enrichment, p-value<10^{-114}), which are involved in hematopoietic stem cell differentiation (de Bruijn and Speck 2004). Taken together, these results are consistent with this module corresponding to a subset of probes located within blood cell specific, especially macrophages, regulatory regions. Modules 1 and 4 together correspond on one hand, to probes in active, hypomethylated regions in cells of the adipose/fibroblast lineage, and on the other hand to probes in active

hypomethylated regions in macrophages or other immune cells, suggestive that two of the largest sets of co-methylated CpG sites in adipose tissue can be assigned to constituent cell types of adipose tissue, based on various genomic and epigenomic properties outlined here.

Like Modules 1 and 4, modules 2 and 5 have eigenprobes that are also strongly anti-correlated to each other. We had previously found HOX clusters as well as many other developmentally significant genes to be highly variable and co-variable with gene expression in fibroblast (Wagner et al. 2014). Genes associated to Module 2 are enriched for functions related to development and pattern specification (e.g. Hox genes and genes from the WNT pathway), but also collagen and the extracellular matrix, pointing to fibroblast-specific activity. Similarly to module 1, probes in this module are strongly enriched for active enhancer marks in fibroblasts, but also adipose derived mesenchymal stem cell. Weaker enrichments for marks related to adipocyte activity, or metabolism genes were seen in Module 2 when compared to module 1.

Like Module 2, probes in Module 5 are also associated to genes involved in development, but also to immune function. Its probes often carry marks of enhancer activity in CD14+ cells (similarly to module 4), but also endothelial cells. Work has previously shown expression of adipogenesis and immune related genes to be increased in endothelial cells of obese individuals (Villaret et al. 2010), and we likewise see enrichment for angiogenesis related genes in Module 5. Module 5 also shows a very strong enrichment for adjacency to binding sites of the ERG transcription factor. All things considered, modules 2 and 5 showed some correspondence to modules 1 and 4 in terms of also having some similar profiles with fibroblasts on one hand, and macrophages on the other hand, but showed other enrichments more characteristic of developmentally significant genes and transcription factors related to fibroblast proliferation on one hand and angiogenesis on the other, rather than to functional genes of fibroblasts and macrophages as seen in modules 1 and 4.

M6 showed a strong overlap with genes related to metabolism of fat, with considerable enrichment in regions related to lipid metabolism (lipid metabolic process q-value: $4.3*10^{-25}$), including many probes near the Fatty acid synthase (FASN) gene, a key gene in lipid metabolism; it also showed enrichment for adipose nuclei enhancer marks. Flanking sequences of module 6 probes showed modest enrichment for the PPAR-gamma transcription factor binding site, a factor active in the differentiation of adipocytes (Cristancho and Lazar 2011). This module also showed strong positive correlation with BMI, indicative all in all of a module driven by body type and tissue composition variation in the sample set, and corresponding most likely to regions active in adipocytes.

Modules 8 and 10 showed similar properties, including strong enrichment for H3K4me1 in fibroblasts, and were the two modules most strongly negatively correlated with BMI. Both are very enriched for AP-1 binding sites (3.5 and 5.5 fold respectively, p-values: 10^{-66} and 10^{-70}); an intriguing result given the demonstrated role of AP-1 in wound healing and variation in AP-1 activity in fibroblasts (Lallemand et al. 1997). Module 10 showed a stronger anti-correlation to expression of extracellular matrix related genes, whereas module 8 was more correlated to immune function genes.

All co-methylation modules discussed to date have evidence of regulatory activity in a subset of cell type constituents of adipose tissue. Module 9 is a clear exception, with its probes showing enrichment only for a H3K9me3 mark in all cell types considered, a mark typically associated with heterochromatin and repression of transcriptional activity (Kim and Kim 2012).

Module 3, whose eigenprobe had positive correlation to both modules 1 and 6, showed enrichment primarily for GO terms related GTP signalling (like module 1) but also some lipid transport terms. M3 showed modest enrichments for enhancer marks of endothelial cells and adipose nuclei enhancer marks, whereas its negatively correlated counterpart, module 7 is enriched for enhancer marks in all cell lines considered especially CD14+. Module 7 is also strongly

enriched for immune function but distinguishes itself from modules 4 and 5 by also being strongly enriched for adipose nuclei and fibroblast enhancer marks.

4.4.4. Isolating methylation variation caused by tissue composition variability.

Adipose tissue is expected to consist of a mixture of various cell types (adipocytes, fibroblasts, macrophages, etc.) (Eto et al. 2009). The proportion of each cell type may vary from sample to sample, resulting in groups of CpG sites with cell-type specific methylation patterns to have correlated methylation levels. However, these strong co-methylation signals are not intrinsic to methylation levels within individual cell types, and the variance caused by such tissue composition effects may mask more subtle signals such as weaker co-methylation signals and mQTLs.

We thus sought to develop methods for correcting for tissue composition effects. Our approach assumes that each sample is a linear combination of unknown proportions of *k* different "components" of unknown methylation profiles. It simultaneously estimates the proportion of each component in each sample, together with the methylation profile (mean and variance of every CpG probe) of each component (Methods). For each site and each sample, the residual obtained by subtracting the weighted sum of mean values from the observed value can be thought as the difference between the observed methylation level and the level predicted given the estimated tissue composition of that sample. The variance of the residual is thus the weighted sum of the variation that is present within each of the constituent cell types, rather than variation caused by tissue composition variation.

As the adipose tissue dataset had neither tissue composition information nor methylation values for pure constituent cell types available, we first applied our approach to an unpublished whole blood Illumina HumanMethylation450 dataset with n=167 samples and cell count information available. As with the adipose dataset, CpG sites on sex chromosomes or overlapping with SNPs were removed, and those in the top 50% with respect to variance were analyzed further. Each beta value was regressed with sex and age of the individuals, and the residuals from this analysis used as input to the deconvolution approach. Applying our deconvolution approach with k=2 yielded two components with weights correlating with the proportion of lymphocytes (R=0.77) and neutrophils (0.716) in each sample. Using Illumina HumanMethylation450 data for various pure blood cell types from (Reinius et al. 2012), component 1 inferred methylation values were found to correlate positively with CD4+ cells (average correlation=0.79), CD8+ cells (0.825), CD19+ cells (0.505) and CD56+ cells (0.725). On the other hand, component 2 correlated positively with neutrophils (0.56), granulocytes (0.55) and eosinophils (0.36).

Obviously, the variance of the residuals decreases as *k* increases, but the Bayesian Information Criterion (BIC) suggested that a value of *k*=3 was best supported by the methylation data (for each of the two twin subsets). We thus used this value to perform deconvolution of the methylation values. We estimate that tissue composition effects explain 20% of the CpG methylation variance, including 53% of that in modular probes (Table 4.7-3).

Of course, without access to purified cell components or measures of tissue composition in adipose, we are certain neither of a correct value of *k* nor that a particular *k* accounts for cell-composition induced variation and only this form of variation. We view the results obtained as a complement to those seen with using uncorrected methylation and expression data: the former providing insight into the key loci variable between cell types in primary adipose tissue, the latter providing at least a partial picture of DNA methylation and gene expression,

their variation and relationship with genetic variation in subsets of constituent cells in adipose.

The inferred component methylation profiles are similar to pure cell type profiles and their residuals lose most of the modules obtained in the original data.

To characterize each of the components, we compared the methylation profile of each component to a collection of recently published methylation data sets obtained from whole blood, pre-adipocytes and fibroblasts (Nazor et al. 2012) (Figure 4.4-4). We find that component 1 strongly resembles a whole blood methylation profile, while component 3 resembles a pre-adipocyte and fibroblast methylation profile. Component 2 does not seem to match well any of the considered cell types (note that no adipocyte methylation data sets were available), but correlates negatively with the whole blood profile. Interestingly, modular probes tend to be those with the strongest component-specific methylation.

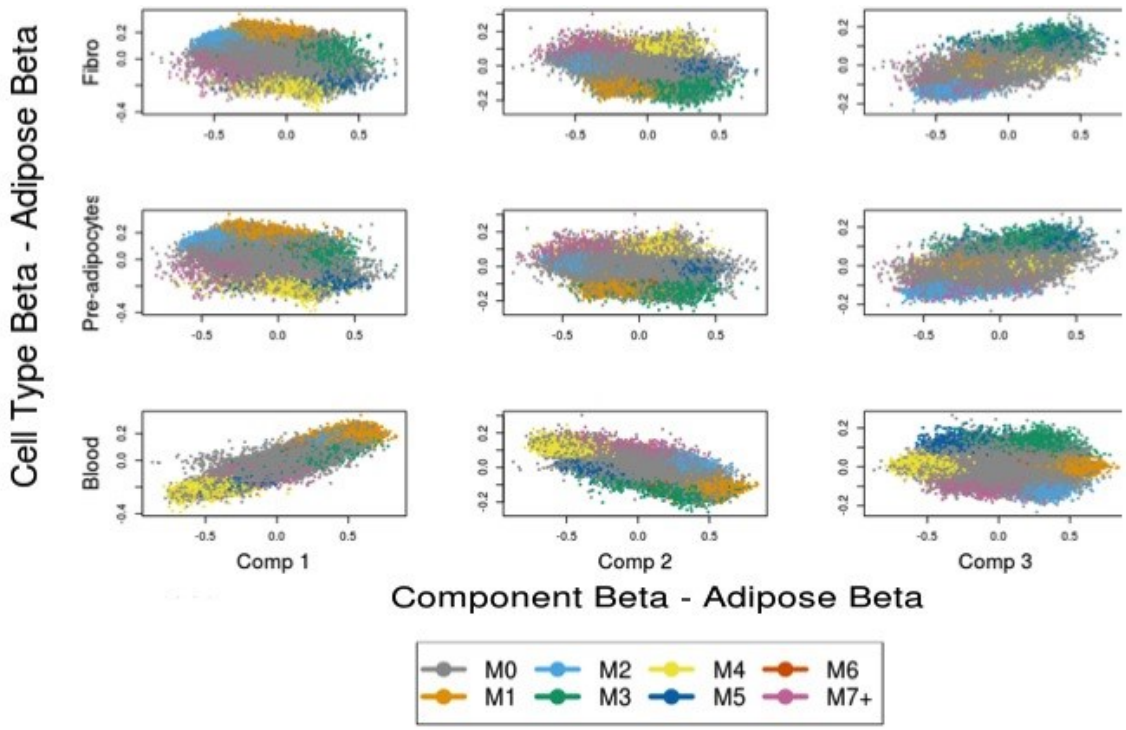

Figure 4.4-4 Inferred cell component methylation values correspond to measured methylation beta values from adipose constituent cell types.

The x-axis of each plot represents the difference between the inferred beta value for a component and the mean beta value observed in this study in adipose tissue, while the y-axis represents the difference between the observed beta value in a cell or tissue type and the beta value observed in this study in adipose tissue.

To further characterize the inferred component methylation profiles and their per-individual weights, we identified genes whose expression level correlated with component weights (Table 4.7-4). Component 1's identity as a representative of macrophage proportions was confirmed by its strong correlation with immune system related genes, while weights assigned to component 3 were found to be correlated to extracellular region related genes, indicative of its fibroblast functionality. Weight for component 2 correlated positively with mitochondrial genes, more characteristic of adipocyte and metabolic function.

4.4.5. Deconvolution enriches the set of mappable CpG sites and genes

If the variance in methylation levels at a given probe is in part caused by extrinsic factors such as cell type composition, one may expect that residuals would exhibit improved mappability to genetic variants, for the CpG sites whose variation is partly explained by genetic variation. We thus repeated the mQTL calculations using the methylation residuals, keeping the same R^2 cutoff of 0.1125 we had used previously, although a slightly lower one could be used with deconvoluted data while preserving a 5% FDR. Although most cis CpG-SNP pairs' correlation coefficients were unaffected by the deconvolution, a large number saw their correlation increase significantly, whereas almost none had a decrease (Figure 4.4-5). The most impressive gains were seen with CpG sites assigned to modules in the original dataset.

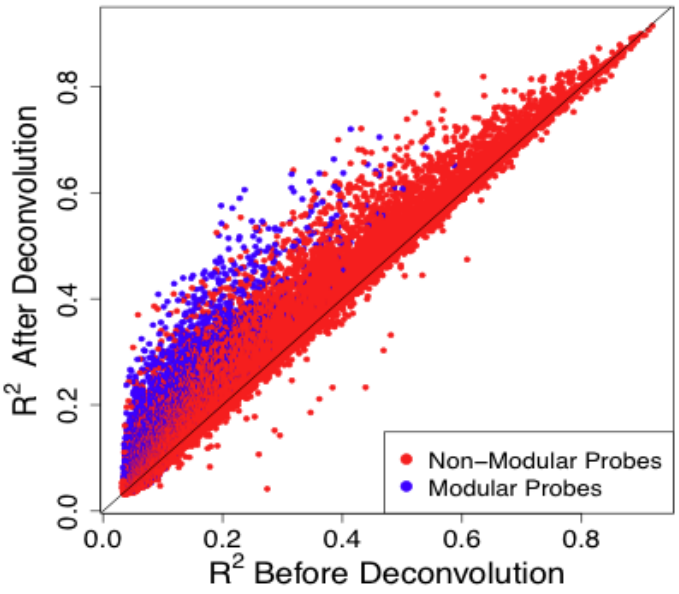

Figure 4.4-5 CpG gain in correlation with cis-mQTLs after deconvolution.

Each point represents the best cis correlation for a CpG site, using original beta values (x axis) and deconvolution residuals (y axis). Points are colored based on module membership or non-membership of the CpG site.

Figure 4.4-6A and Figure 4.4-6B illustrate this phenomenon for a module 1 CpG probe located near the transcription start site of CCDC50, and a SNP located approximately 40 kb from the CpG site, in an exon of this gene. Prior to deconvolution, individuals with high predicted levels for one of the three inferred cell components tended to cluster together in terms of their methylation profile, regardless of their genotype at the mQTL locus. Deconvolution removes this presumed tissue composition effect, and correlation improves considerably. Overall, the number of significant CpG-SNP cis pairs increased by 16% among non-modular probes, but by 325% among modular probes.

Deconvolution also lead to increased prediction of trans-mQTL associations. However single SNPs correlating in trans with many probes (i.e. a candidate “master regulator” of DNA methylation) still did not materialize after repeating these analyses, and the number of trans pairs remained modest compared to significant trans pairs. Figure 4.4-6C and Figure 4.4-6C illustrate the improved correlation seen for a SNP pair. CpG site cg14334824 was assigned to module 4 in our original WGCNA analysis, a module with ties to immune function and potentially corresponding to active regions in macrophage. It is located approximately 1.5 Mb from the ABL gene on chromosome 9. The correlated SNP is located 500 kb from the BCR gene on chromosome 22.

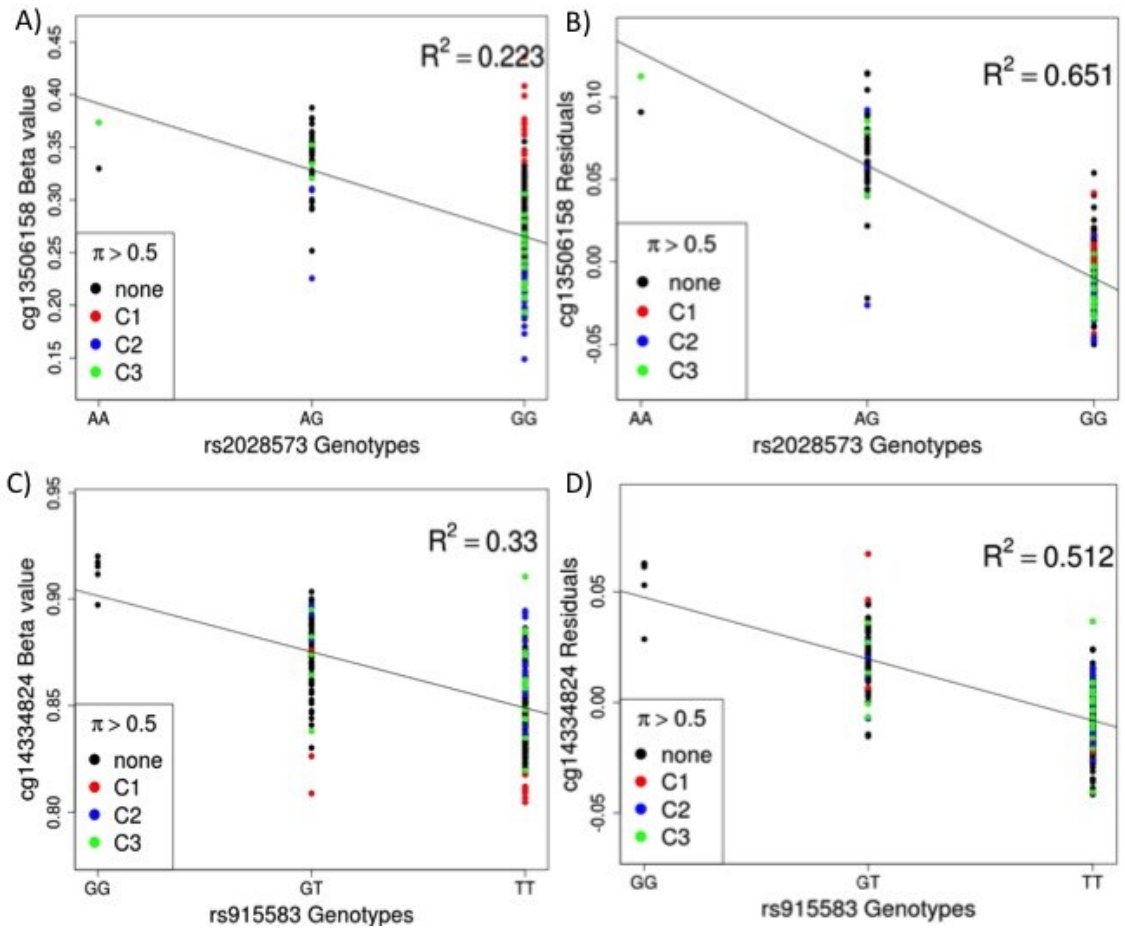

Figure 4.4-6 Improved mQTL relationships for specific examples.

Each point in each plot corresponds to one adipose-1 individual, colored red, blue or green if for one of the three components, its inferred proportion in that individual was greater than 0.5, and colored black otherwise.

4.4.6. Deconvolution greatly reduces the proportion of trans correlations between methylation values

A major effect of the deconvolution procedure is that methylation residuals are much less correlated to each other than the original methylation measurements were. At a relatively strict R^2 threshold of 0.36, (well beyond the threshold of 1% FDR in our permutation test-based approach) 99.8% of pairs of correlated CpG sites were in trans. After deconvolution (i.e. looking at the correlation of the residuals), less than 129,000 pairs remained, with only 71.3%

trans. This reduction was largely the result of a massive loss of correlation among pairs of modular probes from the same module.

We then sought to determine whether the remaining correlation between methylation residuals formed modules, using again WGCNA with the same parameters listed in Methods. The program identified five modules (see Table 4.4-2), each much smaller than those found in the original data. For the most part modules are subsets of existing modules, with some additional non-modular probes being re-assigned to residual modules in some cases.

Module RM1 of residual data strongly resembled Module 6 of the original data, in terms of having a strong overlap and similar properties with respect to enhancer mark enrichment. A strong positive correlation to BMI was preserved among residual probes, indicating that, although our results had captured tissue composition variation in other ways, obesity and the tissue composition effects driving DNA methylation covariation are not as tightly linked as expected.

The majority of residual module 3 probes were originally assigned to modules 8 or 10. A strong negative correlation with BMI was preserved in the residuals. Repeating motif discovery analysis with HOMER using flanking regions of residual module 3 probes revealed even stronger enrichments for the AP-1 motif than with the original module 8 or 10 (New enrichment: 6.8 fold, p-value < 10^{-109}).

Deconvoluted Module	Most Overlapping Original Module	Num Probes	Correlation with BMI	Most Enriched Chromatin Mark
RM1	6	390	0.50715500	Lung Fibroblast H3K36me3
RM2	9	167	-0.34914714	Huvec H3K9me3
RM3	8/10	243	-0.49749301	Fibroblast DHS
RM4	14	91	0.32999853	NA
RM5	4	58	0.26252661	Fibroblast DHS

Table 4.4-2 Properties of Residual Modules

4.4.7. Deconvolution strengthens cis methylation-expression relationships in modular loci

The relation between methylation and gene expression has been the focus of intense research recently, with methylation at promoter and enhancers typically associated with silencing of the target gene and methylation in the body of genes associated with active transcription. This relatively simple picture was muddied in our original methylation and expression data, where more than 400,000 CpG-gene pairs showed significant correlation (R^2>0.118, corresponding to an FDR < 5%), 99.6% of which were in trans. Increasing the correlation threshold only minimally changed this unusual distribution. Repeating the analysis on the residuals of the methylation and expression data yield a more familiar result. The number of pairs at the same correlation threshold is reduced to about a thousand, of which two thirds are in cis.

4.4.8. Deconvolution changes the profile of BMI correlated CpG Sites

In the original set of beta values, probes most positively correlated to BMI were in adipose related module 6, while those most negatively correlated were mostly from in AP-1 related modules 8 and 10. Gene Ontology analysis with GREAT (McLean et al. 2010) revealed a strong enrichment for "triglyceride metabolic process" (Q-val = 2.9×10^{-29}) among positively correlated probes, and for "immune system process" (Q-val = 4.0×10^{-27}) among negatively correlated probes. Repeating these experiments with the probes whose deconvolution residuals were positively correlated with BMI left the former results unchanged (triglyceride metabolic process, Q-val = 3.785109×10^{-30}). Using the original methylation values, negatively correlated probes in modules 8/10 corresponding to fibroblast enhancers, AP-1 binding sites and a fibrosis or wound healing response were present, but considerably outnumbered by module 4 and 5 probes related to macrophage infiltration of adipose tissue in obese individuals. However, the original enrichment for functions related to the immune system disappeared from the probes with negative residual correlation with BMI, leaving the strongest enrichment with "response to wounding" (Q-val = 2.9×10^{-6}). Deconvolution thus served to crystallize the methylation variation present within adipose fibroblasts and generated a useful hypothesis regarding epigenomic signatures of obesity in adipose fibroblasts, specifically at loci with regulatory roles in wound healing or fibrosis. Loss of correlations for modules other than 6, 8 and 10 is shown in Figure 4.4-7 and Table 4.7-5.

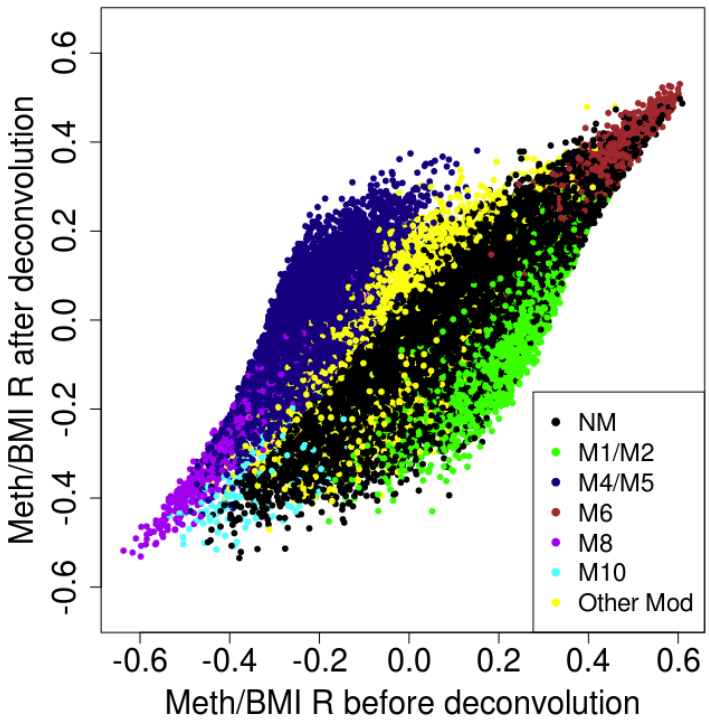

Figure 4.4-7 Deconvolution changes the BMI correlation profile of CpG sites.

4.5. Discussion

We report the first, to our knowledge, characterization of a set of co-methylation modules in a complex tissue derived from a general population, and a deconvolution method that estimates mean methylation profiles of constituent cell types and tissue composition of samples in an unsupervised fashion. Methylation beta values in a total of 581 human female adipose samples, many corresponding to monozygotic or dizygotic twin pairs, were obtained from the Illumina Infinium HumanMethylation450 platform and analyzed for co-methylation and their correlation with gene expression, genetic variation, and BMI. These same beta values were input to our deconvolution algorithm, and residuals obtained by subtracting the estimated cell type-specific methylation were subjected to similar analyses.

Applying the WGCNA module detection approach to the adipose methylation beta values for the top 50% variable sites of the platform, we found 10 large modules corresponding to hundreds or thousands of CpG sites each, distributed throughout the genome. Based on research of the last 10 years clearly identifying an altered cellular landscape of adipose tissue in obese individuals (Weisberg et al. 2003), we hypothesized that many of the co-methylation modules would be driven by tissue composition variation and correspond to CpG sites that are differentially methylated in a particular cell type or set of cell types present in adipose tissue. Two of the largest modules (M4 and M5) corresponded well to active, hypo-methylated regions in macrophage or other blood cells, while their negatively correlated counterparts (M1 and M2) showed enrichments for active marks in other constituent cell types of adipose tissue. These module pairs differed somewhat in that while M1/M4 were in the neighborhood of genes corresponding more to the functionality of the cell types in question (actin binding/extracellular matrix for fibroblast-related M1, immune response for macrophage related M4), M2/M5 were found more in the neighbourhood of genes involved in development and morphogenesis. Modest but statistically significant negative correlations between BMI and macrophage-related co-methylation modules were in line with reasonable expectations of obesity being tied to macrophage infiltration in adipose tissue (Weisberg et al. 2003), and complemented previous results of a conserved co-expression module of genes enriched for macrophage function found in human adipose tissue and conserved in mouse (Emilsson et al. 2008).

Despite dividing quite neatly along the lines of component cell types in adipose, and despite the considerable amount of research showing altered tissue composition in adipose tissue in obesity, these four large modules showed only borderline significant positive or negative correlations with BMI of the samples considered. Stronger positive correlations were seen with member probes of the comparatively small module M6, while negative correlations were seen with the even smaller modules 8 and 10. All things considered, these results were

surprising, and showed that although tissue composition would be expected to affect DNA methylation systematically, obesity/BMI may not be the only factor at play in changing tissue composition of adipose tissue.

Turning to these modules that showed strongest correlation with BMI, Module 6 was not a surprising result as the most strongly positively correlated. This module showed enrichment for regulatory activity at adipose nuclei and it would be expected that in more overweight or obese individuals, macrophage infiltration would lead to a reduced share of adipocytes with hypomethylation for enhancers of fat metabolism related genes. This would in turn entail higher average methylation at these sites in obese individuals, and the positive correlation present. Negatively correlated modules 8 and 10 showed strong enrichments for overlap with fibroblast specific enhancers and open chromatin, AP-1 binding sites, and the neighbourhood of genes related to wound healing. Deconvolution only strengthened this result, with residual module RM3 containing probes from original modules M8 and M10, and demonstrating even stronger enrichments for the properties listed above. Induction of AP-1 target genes in hypoxic conditions is well documented (Salnikow et al. 2002), as are hypoxia and subsequent fibrosis in adipose tissue of obese individuals (Halberg et al. 2009). Just how marked and widespread these enrichments proved to be was a surprising result.

Not all modules showed significant correlation with BMI, or straightforward translation to cell types. In particular, module pair M3/M7 were not easily translatable along the cell lines, showing enrichments or depletions along either all or none of the cell types considered. However, promoters or upstream regions of several key oncogenes such as RUNX3, MGMT and RASSF1 with known patterns of aberrant methylation in various cancers ((Kang et al. 2004), for example demonstrate hypermethylation of all three of these genes' promoters in prostate cancer) showed multiple probes that are members of either M3 or M7. The significance of this variability of oncogene region methylation is not clear, particularly given the relatively weak relationship these modules showed with

expression levels of genes measured. Levels of co-variance present in CpG sites located at or near promoters of genes playing key roles in stability of the genome and the cell are surprising and intriguing results, particularly given that the samples were drawn from adipose tissue in the general population, and not a cancer case/control study.

We showed via multiple lines of evidence that our deconvolution process uncovered true methylation variation present at CpG sites in constituent cell types. Increased proportions of cis methylation-methylation and methylation-expression correlations were examples of this. Perhaps most strikingly, residual CpG sites almost inevitably gained in correlations with SNPs located in cis, despite being tested against more than 2 million SNPs distributed throughout the genome. Failure to preserve biologically meaningful variation in a methylation experiment would have led to CpG sites being correlated with SNPs in an arbitrary manner with respect to gene position. Our analyses failed to show results of mQTLs correlating to multiple groups of CpG sites on different chromosomes. Our strongest result involving a modular (prior to deconvolution) CpG site correlating in trans consisted of a module 4 probe and a correlated SNP, each located within 2 Mb of ABL and BCR genes respectively, the genes present in the classically studied “Philadelpha chromosome” of chronic myelogenous leukemia (Lozzio and Lozzio 1975). While potentially coincidental given the distance of the CpG site and SNP from the genes, and the fact these samples were drawn from non-cancerous fat tissue, we still propose that chromosomal conformation and the possibility of translocation are key factors to consider in trans-mQTL, or potentially even trans-eQTL studies.

In conclusion, we report on many features of adipose co-methylation modules, and developed an approach to correct for tissue composition effects. We anticipate our work will be of interest to anyone wishing to dissect the relationships and correlations at play in studies seeking to measure epigenomic variation and its relationship to phenotype and genetic variation in complex primary tissues such as adipose, whole blood and brain.

4.6. Methods

4.6.1. Subjects and cell samples

Adipose cell samples in this study were the same as those used in (Grundberg et al. 2012; Grundberg et al. 2013). In short, 8 mm punch biopsies were used to obtain subcutaneous adipose samples in a total of 581 adult females, composed of 200 from monozygotic twin pairs, 288 from dizygotic twin pairs and 93 lone individuals, recruited as part of the TwinsUK longitudinal study cohort (Spector and Williams 2006). For purposes of this study, individuals were divided between sets adi-1 (N=290) and adi-2 (N=291), such that no pair of twin sisters was included in the same set.

4.6.2. Genotyping, DNA methylation and Gene Expression Assays

Experiments utilized data from these assays published in (Grundberg et al. 2012; Grundberg et al. 2013). Gene expression measurements were made using Illumina Human HT-12 v3 BeadChips with quantile normalization and quality control carried out in Illumina BeadStudio. Expression data are available on ArrayExpress under accession number E-TABM-1140 and were further corrected for batch effects using ComBat (Johnson et al. 2007).

Methylation was measured with the Illumina Infinium HumanMethylation450 BeadChip. BeadChips were scanned with the IlluminaHiScan SQ scanner, and raw data were imported to the GenomeStudio v.2010.3 software using methylation module 1.8.2 for the extraction of the image intensities. Data were filtered to remove probes containing known SNPs or mapping to multiple regions of the genome (Build hg19) using BLAT (Kent 2002) default parameters, as well as probes located on sex chromosomes. Raw and processed methylation data are available on ArrayExpress under accession number E-MTAB-1866. For this study we further corrected for row, chip, column, plate or batch effects using ComBat (Johnson et al. 2007). Unless otherwise

noted analyses were performed on probes whose quantile normalized, batch corrected beta values are in the top 50% variation across samples in this study. A total of 186194 probes remained after these filtering steps.

4.6.3. Module Identification

Top variable methylation probes from adi-1 and adi-2 were input independently to the blockwiseModules() function of the WGCNA R package (Langfelder and Horvath 2008). Default parameters were used with the exception of a minimum module size of 50, a power coefficient of 12, and utilizing bimod correlation coefficient. Modules discovered with adi-1 and adi-2 were found to strongly agree and adi-2 modules were used to confirm and refine the modules found with adi-1. Namely, a probe was only retained in the final module assignment if it was present both in that module in adi-1 and in the most strongly agreeing module in adi-2.

The first principal component eigenvector of a given module, which we term here the “eigenprobe”, is a vector of length equal to the number of samples input to WGCNA and provides a useful summary of the methylation profile of member probes in a profile. For each module, the majority of member probes were strongly positively correlated with the eigenprobe, while a minority of as much as 25% were strongly negatively correlated. We expect probes negatively correlated to the eigenprobe to show strongly divergent properties compared to the positively correlated majority, and thus considered these regions separately. Specifically, a strong negative correlation was found for eigenprobes of pairs M1/M4, M2/M5 and M3/M7. In each of these cases probes negatively correlated to their assigned module’s eigenprobe were reassigned to the negatively correlated counterpart. For the remaining modules, negatively correlated probes were assigned as non-modular. Analyses proceeded with all modules of size at least 100, listed in Table 4.4-1.

For deconvoluted data, module assignment was carried out with the same WGCNA parameters on residuals obtained by applying deconvolution with $k = 3$. Modules were intersected, and probes negatively correlated with the eigenprobe assigned for this stage of the analysis were reassigned as Non-Modular.

4.6.4. Methylation QTL Analysis

All pairs of top variable methylation probes were autosomal genotypes using the Pearson correlation coefficient. A 5% FDR threshold of $R^2 > 0.1125$ was obtained by permuting methylation values. Cis correlation is defined as CpG site within 1 Mb of the SNP.

4.6.5. Expression Methylation Correlations

All pairs of top variable methylation probes were associated with probes assayed on gene expression array. A Pearson correlation coefficient $R^2 > 0.118$ was defined as corresponding to a 5% FDR obtained by permuting expression values. Cis correlation defined was CpG site within 1 Mb of TSS for gene corresponding to probe.

4.6.6. Methylation-Methylation correlations

All pairs of top variable methylation probes were correlated with each other, and an FDR obtained by permuting methylation values. A Pearson correlation coefficient $R^2 > 0.36$ corresponded to an FDR < 1%. Cis correlation was defined as two CpG sites within 1 Mb of each other.

4.6.7. Histone and DHS

BroadPeak format files from ENCODE (Myers et al. 2011) ChipSeq and DNase I Hypersensitivity experiments were downloaded for all cell types available. Peaks were also obtained from Epigenomics RoadMap experiments (Bernstein et al. 2010) by inputting .wig files to the CisGenome peak caller algorithm (Ji et al. 2008) with input DNA for the appropriate cell type used as background. For each module, the proportion of probes overlapping with a peak from each of the chromatin experiments was tallied, and compared to the set of top variable probes in adipose.

4.6.8. Discriminative motif discovery in modules

For a given WGCNA module, flanking sequences of 100 bp on each side from the GRCh37 build were downloaded from the UCSC Genome Browser Database (Karolchik et al. 2003). This set of sequences were used as foreground for the HOMER discriminative motif discovery program (Heinz et al. 2010), with default settings, and background sequences consisting of sequences obtained in this manner for the full set of top variable probes. In cases where two CpG sites in a given foreground or background dataset were within 200 bp of one another, a single CpG site was selected at random.

4.6.9. Deconvolution

Inter-individual methylation co-variation in complex tissues such as adipose was assumed to be primarily driven by tissue composition variation between individuals. We developed a deconvolution method that, for a study with m methylation probes being analyzed and n samples, takes as input only an $m \times n$ methylation beta values and a value k representing the desired number of cell components to be inferred. The method assumes that a given observed beta value β_{ij} in probe i, sample j, consists of a weighted summation of respective methylation values in each of the k cell types:

$$\beta_{ij} = \sum_{c=1}^{k} \mu_{ic}\,\pi_{jc} + \sum_{c=1}^{k} \delta_{ijc}\,\pi_{jc} + e_{ij}$$

where μ_{ic} is the mean methylation of probe i in cell type c, π_{jc} is the proportion of cell type c in sample j, δ_{ijc} is the individual and probe specific differential methylation level in cell type c, following a Normal distribution with mean 0 and variance $\sigma^2{}_{ic}$ and e_{ij} is random noise. We develop an algorithmic procedure that will output inferred parameters $\breve{\mu}_{ic}$ and $\breve{\sigma}_{ic}^2$ for each probe and cell type and $\breve{\pi}_{jc}$ for each sample and cell type. That is, we find the set of parameters that correspond to a local optimum with respect to the maximum likelihood of the set of observed methylation values, which have the following probability distribution:

$$p(obs(\beta)) = \prod_{i \in probes} \prod_{j \in samples} p(\beta_{ij})$$

where $(\beta_{ij}) \sim Normal(\sum_{c=1}^{k} \breve{\mu}_{ic}\breve{\pi}_{jc}, \sum_{c=1}^{k} \breve{\sigma}_{ic}^2\breve{\pi}_{jc}^2)$.

We do this by setting initial random values for each of these parameters subject to reasonable constraints such as $0 < \breve{\mu}_{ic} < 1$. For each iteration of the algorithm, each probe's values of $\breve{\mu}_{ic}$ and $\breve{\sigma}_{ic}^2$ will be randomly increased or decreased, with each proposed change accepted if it leads to an overall higher probability given the observed beta values and the current values of $\breve{\pi}_{jc}$ across samples. Values of $\breve{\pi}_{jc}$ are then changed, with each change accepted if they lead to higher overall probabilities given observed beta values and current values of $\breve{\mu}_{ic}$ and $\breve{\sigma}^2_{ic}$ across probes.

The deconvolution process was run separately with adipose-1 and adipose-2 for each value of *k* from 2 to 12. The stopping criterion was that the log likelihood of the model changed by less than 0.01 in ten iterations. In order to balance the number of parameters inferred versus the likelihood of the inferred model, the Bayesian Information Criterion (Schwarz 1978) was calculated for each model, with a value of *k*=3 found to be the best choice for both sets.

Although the deconvolution process is unsupervised and could possibly converge to local optima, the component-specific methylation profiles obtained for each of the twin sets are remarkably similar (Figure 4.7-2), suggesting that a genuine source of biological variation was identified.

Residual variation $\mathrm{r_{ij}}$ for probe i in sample j can be can be obtained as:

$$r_{ij} = \sum_{c=1}^{k} \delta_{ijc}\,\pi_{jc} + e_{ij} = \beta_{ij} - \sum_{c=1}^{k} \breve{\mu}_{ic}\,\breve{\pi}_{jc}$$

Results reported correspond to those obtained with inferred $\breve{\mu}_{ic}$, $\breve{\pi}_{jc}$, and r_{ij} for adipose-1 and adipose-2 run separately with *k* = 3.

4.7. Supplementary Figures and Tables

Adi-1 Module ID	Adi-2 Module ID	Adi-1 Module Size	Adi-2 Module Size	Overlap size	Overlap Fold Enrich-ment	Hyper-geometric p-value
1	3	11762	8106	5819	11.9	0
2	4	9118	6108	4966	17.4	0
3	1	8425	9160	5703	14.4	0
3	8	8425	1250	770	14.3	0
4	2	7701	8589	5371	15.8	0
5	5	3329	3785	2281	35.3	0
6	6	1917	2417	1569	66.0	0
7	7	1612	1489	885	71.9	0
8	9	859	641	330	116.9	0
8	11	859	241	225	211.9	0
9	10	369	352	256	384.3	0
10	4	306	6108	106	11.06	7.58899E-79
10	9	306	641	154	153.1	0

Table 4.7-1 Intersections of adi-1 and adi-2 modules used in this study.

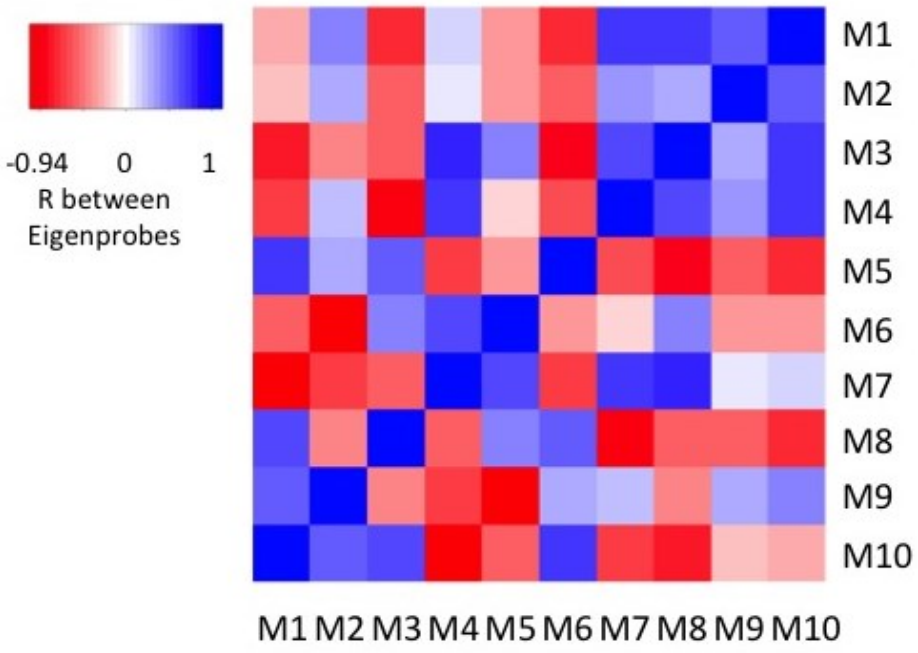

Figure 4.7-1 Pearson correlation coefficient between module eigenprobes

Module	Correlated Pairs	Unique Genes	Average # of Genes Correlated Per Probe	Top GO Term of Correlated Gene	Adjusted P-val of Top GO Term
1	173905	230	28.1	Extracellular Region	3.2E-10
2	10681	177	2.1	Cellular Metabolic Process	0.184
3	11940	213	1.9	Mitochondrion	1.24E-5
4	139042	463	27.7	Immune System Proces	4.88E-15
5	5997	180	2.5	Immune System Process	1.06E-3
6	120454	478	79.0	Mitochondrion	1.07E-19
7	2112	194	1.7	Immune System Process	8.58E-6
8	87729	679	163.7	Immune System Process	3.83E-12
9	2	2	0.01242236	NA	NA
10	13430	391	51.65384615	Extracellular Region	3.2E-10
NM	333611	1376	2.002539107	Transmembrane Receptor	0.99
All	890246	1565	4.609758974	NA	NA

Table 4.7-2 Correlations between gene expression and CpG sites, by module

Module	Average Standard Deviation (Beta value) Before Deconvolution	Percent Variance Removed By Deconvolution
NM	0.0279	13.6%
Modular	0.0320398	53%
Overall	0.0285	20.3%

Table 4.7-3 Methylation variance changes induced by deconvolution.

Component	Number of Genes Correlated (FDR < 10^{-5}, |R| > 0.3)	Top GO term of positively correlated genes (Adjusted p-val)	Top GO Term of negatively correlated genes (Adjusted p-val)
1	3219	Immune System (6.6×10^{-52})	Mitochondrion (2.5×10^{-25})
2	2867	Mitochondrion (4.3×10^{-63})	Immune System (1.6×10^{-38})
3	242	Extracellular Region Part (1.2×10^{-8})	Generation of Precursor Metabolites (1.3×10^{-5})

Table 4.7-4 Component weights correspond to gene expression profiles of cell specific expressed genes

Original Module	Eigenprobe correlation to BMI	% of Probes Correlated with BMI (FDR q-value < 0.01, |R| > 0.216)	After Deconvolution Percentage Remaining
1	0.306	68.4	3.0
2	0.281	47.1	1.3
3	0.097	14.8	2.5
4	-0.280	65.1	5.7
5	-0.295	62.8	3.6
6	0.555	99.6	77.3
7	-0.093	12.5	3.4
8	-0.543	98.7	58.7
9	-0.298	18.7	14.1
10	-0.396	75	88.8
NM	NA	13.8	3.3

Table 4.7-5 The majority of modular probes lose correlation to BMI after deconvolution, some module 10 probes gained correlations.

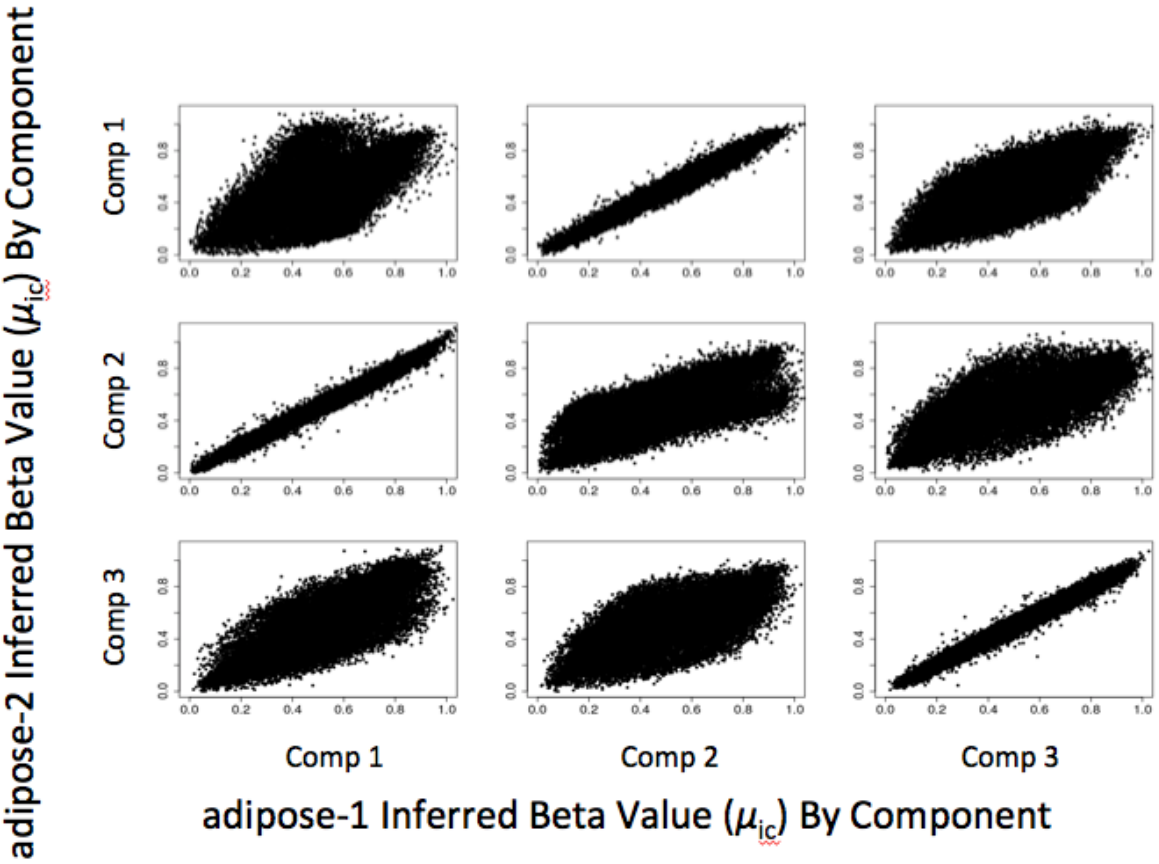

Figure 4.7-2 Replicability of inferred mean beta values in deconvolution

Chapter 5. Conclusion

5.1. Research Contributions

5.1.1. Chapter 2: Hidden Markov Models for Allelic Expression Detection

At the onset of this thesis research, research exploring allelic expression on a genome-wide scale utilizing array or sequencing methods had been ongoing for several years. Approaches to analyzing array based allelic expression data focused on making use of annotated gene boundaries to average or otherwise aggregate allelic expression levels at heterozygous SNPs within a given individual. We took advantage of the linear, sequential nature of allelic expression data to develop a Hidden Markov Model (HMM) to assign allelic expression levels for each locus measured in an individual. With our first HMM implementation (ergodic), the probability distribution of allelic expression states at a given locus will be dependent not only on its measured values, but also on those of the loci in its immediate neighborhood. We then developed a left-to-right HMM to learn a distinct set of transition probabilities for each locus in the genome. This approach has the advantage of utilizing information about a locus for the full set of individuals in the study, and was found to lead to improved detection of allelically expressed regions at a similar false discovery rate, compared to the ergodic HMM or other simple smoothing approaches tested. Visualizing specific examples showed crisper boundaries between allelic and non-allelic expression. We found this HMM based approach to be a promising application of a computational tool towards discovering regions of allelic expression without reference to gene boundary annotations, and found evidence of allelic expression present in intergenic regions, or in gene regions but not

corresponding precisely to annotated gene boundaries. Utilizing population information provided additional insight into the allelic expression present in any given individual and higher overall estimates of allelic expression as a whole. We applied our Left to Right HMM to allelic expression data in fibroblast, finding a higher number of allelic expression QTLs (aeQTLs or cis-regulatory SNPs) compared to a microarray-based eQTL study with the same samples.

5.1.2. Chapter 3: Relationships between DNA methylation, gene expression and sequence variation in human fibroblast

Chapter 3 was an exploration of the properties of DNA methylation in fibroblast. At the time of this research the Illumina Infinium 450K HumanMethylation array was a relatively new method and one that could interrogate the methylation status at a large number of loci in the genome in many individuals, for a relatively low cost compared to sequencing-based methods. Experimental results from this platform, together with those from gene expression arrays, allelic expression and SNP genotyping arrays afforded us one of the first opportunities to study the population level variation and covariation of DNA methylation, gene expression and sequence variation in a primary human cell type. Methylation was found to vary considerably across the genome and to correlate in both positive and negative directions in cis with gene expression at a subset of genes, including HOX loci and other developmentally significant transcription factors. Though chromatin accessibility and histone modification data were not available for the fibroblast samples in this study, these data were available for various fibroblast cell lines thanks to the ENCODE (Myers et al. 2011; Consortium 2012) project. Defining regions based on open chromatin and promoter related marks such as H3K4me3 corresponded better with CpG sites negatively correlated with gene expression than did taking account only the position of CpG sites with respect to the TSS. On the other hand, CpG sites positively correlated with gene expression corresponded better to marks associated with repression of gene expression such as H3K27me3 than they did to annotated gene bodies. Work done with chromatin marks and expression

correlated CpG sites showed the importance of moving beyond a paradigm of promoter DNA methylation in healthy human methylation variation and its relationship with expression, and corresponds well with a growing body of research stressing the importance of considering methylation to regions such as enhancers in healthy populations and cancer (Aran et al. 2013).

Genetic variation with an impact on methylation at nearby loci was statistically over-represented but still found to occur at only about 2% of highly variable methylation loci considered. Genetic variation with an impact on both methylation and expression was found to be even rarer. However, making use of allelic expression data and our Left-to-Right HMM approach for detecting and assigning levels of allelic expression to loci helped to enrich these data and enabled us to find considerably more regions with both methylation and expression variation being correlated with genetic variation. As was the case with the ENCODE open chromatin and histone modification data enriching the results seen with methylation-expression relationships; the increased number of expression-methylation QTLs being found when using allelic expression data without respect to gene boundaries demonstrated the importance of looking beyond gene boundaries, as excellent as the annotations are in the human genome, and considering other tools and/or public datasets to develop a fuller picture of the interactions and relationships present in the study at hand.

5.1.3. Chapter 4 DNA co-methylation and tissue composition effects in human adipose tissue

In chapter 4, we also considered relationships between methylation, gene expression and genetic variation in samples drawn from a general human population. The larger sample sizes of these datasets provided by the MuTHER Consortium afforded us the opportunity to carry out more extensive statistical tests regarding co-methylation between CpG probes in trans, as well as trans-mQTL analysis. On the other hand, it became apparent that special challenges would be posed by the fact these measurements were done in adipose, a

heterogeneous tissue with considerable tissue composition variation present in the population. Co-methylation analysis with data that had been normalized and corrected for batch effects, but not corrected for tissue composition effects revealed several modules consisting of hundreds to thousands of probes each. For both methylation-methylation or methylation-expression correlations, trans pairs dominated overwhelmingly, on the other hand, mQTL relationships were under-represented in modular probes. We developed a computational approach to infer tissue compositions of the adipose samples in our study, as well as methylation beta values for each of the constituent cell types inferred. This approach is unsupervised and takes as input only a matrix of methylation or expression values for the study in question, and a parameter k for the number of cell types to infer. Analyses were repeated with residuals obtained from this analysis and found to give stronger correlations with mQTLs, and much fewer trans methylation-methylation or methylation-expression correlations. Gene Ontology categories of genes in the neighborhood of CpG sites correlated to the Body Mass Index shifted from those of immune function towards extracellular matrix function, generating hypotheses and lending credence to the importance of the extracellular matrix and fibrosis as a consideration in obesity.

Our approach is useful in unmasking relationships between DNA methylation and DNA sequence that are under-estimated or missed when performed in complex tissue. Furthermore, application of our deconvolution approach to high-throughput datasets obtained from a complex tissue before doing comethylation, methylation-expression or epigenome wide analyses would be expected to generate hypotheses regarding these relationships that are more in line with regulatory variation present in particular constituent cell types, and less in line with simple tissue composition variation. Nevertheless, we propose that execution and presentation of these correlation-based analyses on data prior to deconvolution is still a worthwhile task, as it can still generate new hypotheses and insights regarding loci that are differentially methylated or expressed in particular constituent cell types in the context of tissue in its natural form, without

resorting to potentially distorting cell purification or cell sorting methods. Indeed, the most complete picture is obtained by presenting results of all analyses before and after deconvolution.

5.2. Future Work

5.2.1. Chapter 2

As with other genomic technologies, allelic expression has moved from an array based to sequencing based measurement (Pastinen 2010). A natural extension of the Hidden Markov Model based approach would be to sequencing based technologies. This would not necessarily be a trivial extension, given the higher number of loci to contend with in a full genome sequencing experiment, as well as the digital nature of read counts output by a sequencing experiment. Considerations of what coverage levels are adequate to draw meaningful conclusions regarding allelic expression are paramount, and in the case of a left to right Hidden Markov Model, questions regarding whether neighboring loci and/or the same locus in other individuals that have high read coverage can be helpful in particular loci with low read coverage in a given individual. Our LTOR HMM can be trivially parallelized by running the approach in parallel for each chromosome to learn parameters.

Our method presently considers only intensities related to the cDNA and gDNA output by the algorithm. As would be indicated results from ours and others, allele specific expression would be expected at some loci to be driven by heterozygosity of sequence variants in nearby binding sites. In other cases, it could be correlated with allele specific methylation in the same region. In a small but important number of regions, both methylation and sequence variation can be expected to be present in the same vicinity as allelic expression. The ideal would be to integrate all of these data, possibly even with other information such as

allele specific histone modification to develop a fully integrated picture of not only allelic expression, but other allele specific measurements, and generate hypotheses regarding the causal structure of these arrangements. Careful consideration of the model structure, the state space and the basis for emission and transition probabilities would be in order and highly dependent on the design of the experiment and the nature of the results set. Avoiding combinatorial explosion of the states possible in such a model and sensible validation of the results via an approach such as permutation testing based false discovery rate would also be in order.

5.2.2. Chapter 3

Given the datasets at hand, our work was a quite detailed consideration of the various relationships at play in the fibroblasts studied. A possible future work could be inference of causal mechanisms at loci involving correlations between methylation, expression and sequence variations. Several possible mechanisms involving the interplay between methylation variation as a cause or consequence of expression variation, or more complicated models of a reinforcing positive feedback loop between methylation and repressed transcription, were proposed by (Blattler and Farnham 2013). (van Eijk et al. 2012) utilized a local edge orienting method to try to infer such causal mechanisms using similar gene expression, DNA methylation and genotyping datasets for whole blood.

5.2.3. Chapter 4

Our approach for the unsupervised deconvolution of tissue composition effects has several potential extensions. While, as pointed out, the approach developed works in an unsupervised manner, it could potentially be made semi-supervised in cases where measurements of DNA methylation levels in cell types roughly corresponding to constituent cell types of the tissue studied. Although

these cell lines may not be expected to behave exactly as they would in the primary tissue, or are subject to purification or immortalization effects, they may serve as a better starting point than randomly selected methylation values.

We made use of the Bayesian Information Criterion (BIC) to select a single value of *k* for our experiments with adipose. While work could be proposed to use other statistical properties to automatically select a value of *k* and output a tissue composition and cell component methylation values corresponding to an "optimal" value of k, it is important to note that in any complex tissue, the number of cell types can be very high. Furthermore the definition of "cell type" can be rather fluid and hierarchical, as one hand types of cells can be categorized into lineages, and a particular cell type can show different patterns of expression or methylation depending on its tissue microenvironment or genetic sequence, thus leading to distinct "subtypes" of a particular cell in the sense of their genomic measurements, if not in their morphology. A possible extension could involve starting the algorithm with a very large level of k, merging cell components on the fly if they become too similar and/or if their proportion in all individuals becomes too low. Final methylation values for remaining cell types could be hierarchically clustered and mapped to appropriate cell sub-types, types and lineages based on known biological or epigenomic properties of the tissue and its constituent cell types.

With epigenomic studies only set to improve in terms of their quality, the number of loci interrogated, and the depth and variety of samples and tissues studied, these are exciting days for all of us wanting to unlock the secrets of gene regulation and its relationship with phenotype. Careful consideration of all possible relationships, augmenting analyses with data from public repositories, and inclusion of tissue composition effects in analyses are all important considerations that I hope this thesis has helped to highlight.

References

NCBI Infinium Probe Design. Vol 2014. NCBI.

Akulenko R, Helms V. 2013. DNA co-methylation analysis suggests novel functional associations between gene pairs in breast cancer samples. *Human Molecular Genetics* **22**(15): 3016-3022.

Almlöf JC, Lundmark P, Lundmark A, Ge B, Maouche S, Göring HHH, Liljedahl U, Enström C, Brocheton J, Proust C. 2012. Powerful identification of cis-regulatory SNPs in human primary monocytes using allele-specific gene expression. *PloS one* **7**(12): e52260.

Alter O, Brown PO, Botstein D. 2000. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences* **97**(18): 10101-10106.

Altshuler D, Daly MJ, Lander ES. 2008. Genetic mapping in human disease. *Science* **322**(5903): 881-888.

Amato RJ. 2007. Inhibition of DNA methylation by antisense oligonucleotide MG98 as cancer therapy. *Clinical genitourinary cancer* **5**(7): 422-426.

Aran D, Sabato S, Hellman A. 2013. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biol* **14**(3): R21.

Aten JE, Fuller TF, Lusis AJ, Horvath S. 2008. Using genetic markers to orient the edges in quantitative trait networks: the NEO software. *BMC Systems Biology* **2**(1): 34.

Auburger G, Klinkenberg M, Drost J, Marcus K, Morales-Gordo B, Kunz WS, Brandt U, Broccoli V, Reichmann H, Gispert S et al. 2012. Primary Skin Fibroblasts as a Model of Parkinson's Disease. *Molecular Neurobiology* **46**(1): 20-27.

Bacanu S-A, Devlin B, Roeder K. 2000. The power of genomic control. *The American Journal of Human Genetics* **66**(6): 1933-1944.

Baross A, Delaney AD, Li HI, Nayar T, Flibotte S, Qian H, Chan SY, Asano J, Ally A, Cao M et al. 2007. Assessment of algorithms for high throughput detection of genomic copy number variation in oligonucleotide microarray data. *Bmc Bioinformatics* **8**.

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM. 2008. Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature genetics* **40**(8): 955-962.

Bauer S, Grossmann S, Vingron M, Robinson PN. 2008. Ontologizer 2.0—a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics* **24**(14): 1650-1651.

Baum LE, Petrie T, Soules G, Weiss N. 1970. A MAXIMIZATION TECHNIQUE OCCURRING IN STATISTICAL ANALYSIS OF PROBABILISTIC FUNCTIONS OF MARKOV CHAINS. *Annals of Mathematical Statistics* **41**(1): 164-&.

Baylin SB, Herman JG, Graff JR, Vertino PM, Issa JP. 1998. Alterations in DNA methylation: A fundamental aspect of neoplasia. *Advances in Cancer Research, Vol 72* **72**: 141-196.

Beissbarth T, Speed TP. 2004. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics* **20**(9): 1464-1465.

Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK. 2011. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biology* **12**(1).

Bengtsson H, Irizarry R, Carvalho B, Speed TP. 2008. Estimation and assessment of raw copy numbers at the single locus level. *Bioinformatics* **24**(6): 759-767.

Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, Kellis M, Marra MA, Beaudet AL, Ecker JR et al. 2010. The NIH Roadmap Epigenomics Mapping Consortium. *Nature Biotechnology* **28**(10): 1045-1048.

Birney E Stamatoyannopoulos JA Dutta A Guigo R Gingeras TR Margulies EH Weng Z Snyder M Dermitzakis ET Stamatoyannopoulos JA et al. 2007. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**(7146): 799-816.

Blattler A, Farnham PJ. 2013. Cross-talk between Site-specific Transcription Factors and DNA Methylation States. *Journal of Biological Chemistry* **288**(48): 34287-34294.

Bocklandt S, Lin W, Sehl ME, Sanchez FJ, Sinsheimer JS, Horvath S, Vilain E. 2011. Epigenetic Predictor of Age. *Plos One* **6**(6).

Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, Furey TS, Crawford GE. 2008a. High-resolution mapping and characterization of open chromatin across the genome. *Cell* **132**(2): 311-322.

Boyle AP, Guinney J, Crawford GE, Furey TS. 2008b. F-Seq: a feature density estimator for high-throughput sequence tags. *Bioinformatics* **24**(21): 2537-2538.

Browning BL, Browning SR. 2009. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *American Journal of Human Genetics* **84**(2): 210-223.

Browning SR. 2008. Missing data imputation and haplotype phase inference for genome-wide association studies. *Human Genetics* **124**(5): 439-450.

Busche S, Ge B, Vidal R, Spinella J-F, Saillour V, Richer C, Healy J, Chen S-H, Droit A, Sinnett D et al. 2013. Integration of High-Resolution Methylome and Transcriptome Analyses to Dissect Epigenomic Changes in Childhood Acute Lymphoblastic Leukemia. *Cancer Research* **73**(14): 4323-4336.

Campino S, Forton J, Raj S, Mohr B, Auburn S, Fry A, Mangano VD, Vandiedonck C, Richardson A, Rockett K et al. 2008. Validating Discovered Cis-Acting Regulatory Genetic Variants: Application of an Allele Specific Expression Approach to HapMap Populations. *Plos One* **3**(12).

Carlberg C, Molnár F. 2013. *Mechanisms of Gene Regulation*. Springer Netherlands.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**(7031): 400-404.

Cedar H, Bergman Y. 2009. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics* **10**(5): 295-304.

Chadwick LH. 2012. The NIH roadmap epigenomics program data resource. *Epigenomics* **4**(3): 317-324.

Chang HY, Chi JT, Dudoit S, Bondre C, van de Rijn M, Botstein D, Brown PO. 2002. Diversity, topographic differentiation, and positional memory in human fibroblasts. *Proceedings of the National Academy of Sciences of the United States of America* **99**(20): 12877-12882.

Chen Y, Zhu J, Lum PY, Yang X, Pinto S, MacNeil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK. 2008. Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**(7186): 429-435.

Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen KY, Morley M, Spielman RS. 2003. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nature Genetics* **33**(3): 422-425.

Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J. 2007. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Research* **35**(6): 2013-2025.

Consortium EP. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**(7414): 57-74.

Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. 2009. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics* **10**(3): 184-194.

Cooper JD, Smyth DJ, Smiles AM, Plagnol V, Walker NM, Allen JE, Downes K, Barrett JC, Healy BC, Mychaleckyj JC. 2008. Meta-analysis of genome-wide association study data identifies additional type 1 diabetes risk loci. *Nature genetics* **40**(12): 1399-1401.

Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, Hanna J, Lodato MA, Frampton GM, Sharp PA. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences* **107**(50): 21931-21936.

Cristancho AG, Lazar MA. 2011. Forming functional fat: a growing understanding of adipocyte differentiation. *Nature reviews Molecular cell biology* **12**(11): 722-734.

de Bruijn MF, Speck NA. 2004. Core-binding factors in hematopoiesis and immune function. *Oncogene* **23**(24): 4238-4248.

Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, De Leon S, Michelini K, Lewellen N, Crawford GE. 2012. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**(7385): 390-394.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*: 1-38.

DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM. 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature genetics* **14**(4): 457-460.

Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KC, Taylor J, Burnett E, Gut I, Farrall M. 2007. A genome-wide association study of global gene expression. *Nature genetics* **39**(10): 1202-1207.

Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**(7398): 376-380.

Drong AW, Nicholson G, Hedman ÅK, Meduri E, Grundberg E, Small KS, Shin S-Y, Bell JT, Karpe F, Soranzo N. 2013. The presence of methylation quantitative trait loci indicates a direct genetic influence on the level of DNA methylation in adipose tissue. *PloS one* **8**(2): e55923.

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics* **14**(9): 755-763.

Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S. 2008. Genetics of gene expression and its effect on disease. *Nature* **452**(7186): 423-428.

Ernst J, Kellis M. 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature biotechnology* **28**(8): 817-825.

Eto H, Suga H, Matsumoto D, Inoue K, Aoi N, Kato H, Araki J, Yoshimura K. 2009. Characterization of structure and cellular components of aspirated and excised adipose tissue. *Plastic and reconstructive surgery* **124**(4): 1087-1097.

Faghihi MA, Wahlestedt C. 2009. Regulatory roles of natural antisense transcripts. *Nature reviews Molecular cell biology* **10**(9): 637-643.

Fan JB, Gunderson KL, Bibikova M, Yeakley JM, Chen J, Wickham Garcia E, Lebruska LL, Laurent M, Shen R, Barker D. 2006. [3] Illumina Universal Bead Arrays. *Methods in enzymology* **410**: 57-73.

Fearnhead P. 2006. Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing* **16**(2): 203-213.

Fraser HB, Lam LL, Neumann SM, Kobor MS. 2012. Population-specificity of human DNA methylation. *Genome Biology* **13**(2).

Frazer KA Ballinger DG Cox DR Hinds DA Stuve LL Gibbs RA Belmont JW Boudreau A Hardenbol P Leal SM et al. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**(7164): 851-U853.

Gaffney DJ. 2013. Global properties and functional complexity of human gene regulatory variation. *PLoS genetics* **9**(5): e1003501.

Ge B, Pokholok DK, Kwan T, Grundberg E, Morcos L, Verlaan DJ, Le J, Koka V, Lam KCL, Gagne V et al. 2009. Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. *Nature Genetics* **41**(11): 1216-U1278.

Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai S-L, Arepalli S, Dillman A, Rafferty IP, Troncoso J et al. 2010. Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain. *Plos Genetics* **6**(5).

Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Yang H, Ch'ang L-Y, Huang W, Liu B, Shen Y. 2003. The international HapMap project. *Nature* **426**(6968): 789-796.

Gilad Y, Rifkin SA, Pritchard JK. 2008. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in genetics* **24**(8): 408-415.

Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. 2007. Widespread monoallelic expression on human autosomes. *Science* **318**(5853): 1136-1140.

Göring HH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JB, Abraham LJ, Rainwater DL, Comuzzie AG. 2007. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature genetics* **39**(10): 1208-1216.

Grundberg E, Meduri E, Sandling JK, Hedman ÅK, Keildson S, Buil A, Busche S, Yuan W, Nisbet J, Sekowska M. 2013. Global analysis of DNA methylation variation in adipose tissue from twins reveals links to disease-associated variants in distal regulatory elements. *The American Journal of Human Genetics* **93**(5): 876-890.

Grundberg E, Small KS, Hedman ÅK, Nica AC, Buil A, Keildson S, Bell JT, Yang T-P, Meduri E, Barrett A. 2012. Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nature genetics* **44**(10): 1084-1089.

Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS. 2005. A genome-wide scalable SNP genotyping assay using microarray technology. *Nature genetics* **37**(5): 549-554.

Gunderson KL, Steemers FJ, Ren H, Ng P, Zhou L, Tsan C, Chang W, Bullis D, Musmacker J, King C. 2006. Whole-genome genotyping. *Methods in enzymology* **410**: 359-376.

Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, Watkins PC, Ottina K, Wallace MR, Sakaguchi AY et al. 1983. A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* **306**(5940): 234-238.

Gutierrez-Arcelus M, Lappalainen T, Montgomery SB, Buil A, Ongen H, Yurovsky A, Bryois J, Giger T, Romano L, Planchon A et al. 2013. Passive and active DNA methylation and the interplay with genetic variation in gene regulation. *eLife*.

Halberg N, Khan T, Trujillo ME, Wernstedt-Asterholm I, Attie AD, Sherwani S, Wang ZV, Landskroner-Eiger S, Dineen S, Magalang UJ. 2009. Hypoxia-inducible factor 1α induces fibrosis and insulin resistance in white adipose tissue. *Molecular and cellular biology* **29**(16): 4467-4483.

Hashimshony T, Zhang JM, Keshet I, Bustin M, Cedar H. 2003. The role of DNA methylation in setting up chromatin structure during development. *Nature Genetics* **34**(2): 187-192.

He Y-F, Li B-Z, Li Z, Liu P, Wang Y, Tang Q, Ding J, Jia Y, Chen Z, Li L. 2011. Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA. *Science* **333**(6047): 1303-1307.

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK. 2010. Simple Combinations of Lineage-Determining Transcription Factors Prime< i> cis</i>-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular cell* **38**(4): 576-589.

Hom G, Graham RR, Modrek B, Taylor KE, Ortmann W, Garnier S, Lee AT, Chung SA, Ferreira RC, Pant PK. 2008. Association of systemic lupus erythematosus with C8orf13–BLK and ITGAM–ITGAX. *New England Journal of Medicine* **358**(9): 900-909.

Horvath S, Zhang Y, Langfelder P, Kahn RS, Boks M, van Eijk K, van den Berg LH, Ophoff RA. 2012. Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol* **13**(10): R97.

Illingworth RS, Bird AP. 2009. CpG islands–'a rough guide'. *FEBS letters* **583**(11): 1713-1720.

Ito S, Shen L, Dai Q, Wu SC, Collins LB, Swenberg JA, He C, Zhang Y. 2011. Tet proteins can convert 5-methylcytosine to 5-formylcytosine and 5-carboxylcytosine. *Science* **333**(6047): 1300-1303.

Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH. 2008. An integrated software system for analyzing ChIP-chip and ChIP-seq data. *Nature biotechnology* **26**(11): 1293-1300.

John S, Sabo PJ, Thurman RE, Sung M-H, Biddie SC, Johnson TA, Hager GL, Stamatoyannopoulos JA. 2011. Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nature genetics* **43**(3): 264-268.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F. 2001. Haplotype tagging for the identification of common disease genes. *Nature genetics* **29**(2): 233-237.

Johnson WE, Li C, Rabinovic A. 2007. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**(1): 118-127.

Jones PA. 1999. The DNA methylation paradox. *Trends in genetics : TIG* **15**(1): 34-37.

Jones PA. 2012. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* **13**(7): 484-492.

Jones PA, Liang G. 2009. Rethinking how DNA methylation patterns are maintained. *Nature Reviews Genetics* **10**(11): 805-811.

Jordan MI, Ghahramani Z, Jaakkola TS, Saul LK. 1999. An introduction to variational methods for graphical models. *Machine learning* **37**(2): 183-233.

Kang GH, Lee S, Lee HJ, Hwang KS. 2004. Aberrant CpG island hypermethylation of multiple genes in prostate cancer and prostatic intraepithelial neoplasia. *The Journal of pathology* **202**(2): 233-240.

Karolchik D, Baertsch R, Diekhans M, Furey TS, Hinrichs A, Lu Y, Roskin KM, Schwartz M, Sugnet CW, Thomas DJ. 2003. The UCSC genome browser database. *Nucleic acids research* **31**(1): 51-54.

Kasowski M, Grubert F, Heffelfinger C, Hariharan M, Asabere A, Waszak SM, Habegger L, Rozowsky J, Shi M, Urban AE. 2010. Variation in transcription factor binding among humans. *science* **328**(5975): 232-235.

Kass SU, Landsberger N, Wolffe AP. 1997. DNA methylation directs a time-dependent repression of transcription initiation. *Current biology : CB* **7**(3): 157-165.

Kendziorski C, Chen M, Yuan M, Lan H, Attie A. 2006. Statistical methods for expression quantitative trait loci (eQTL) mapping. *Biometrics* **62**(1): 19-27.

Kent WJ. 2002. BLAT—the BLAST-like alignment tool. *Genome research* **12**(4): 656-664.

Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proceedings of the National Academy of Sciences of the United States of America* **100**(20): 11484-11489.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The human genome browser at UCSC. *Genome Research* **12**(6): 996-1006.

Khalil AM, Guttman M, Huarte M, Garber M, Raj A, Morales DR, Thomas K, Presser A, Bernstein BE, van Oudenaarden A et al. 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **106**(28): 11667-11672.

Kim J, Kim H. 2012. Recruitment and biological consequences of histone modification of H3K27me3 and H3K9me3. *ILAR Journal* **53**(3-4): 232-239.

Koch CM, Suschek CV, Lin Q, Bork S, Goergens M, Joussen S, Pallua N, Ho AD, Zenke M, Wagner W. 2011. Specific age-associated DNA methylation changes in human dermal fibroblasts. *PLoS One* **6**(2): e16679.

Kwan T, Benovoy D, Dias C, Gurd S, Serre D, Zuzan H, Clark TA, Schweitzer A, Staples MK, Wang H. 2007. Heritability of alternative splicing in the human genome. *Genome research* **17**(8): 1210-1218.

Lallemand D, Spyrou G, Yaniv M, Pfarr CM. 1997. Variations in Jun and Fos protein expression and AP-1 activity in cycling, resting and stimulated fibroblasts. *Oncogene* **14**(7): 819-830.

Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, Bernstein BE, Bickel P, Brown JB, Cayting P. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research* **22**(9): 1813-1831.

Langfelder P, Horvath S. 2008. WGCNA: an R package for weighted correlation network analysis. *Bmc Bioinformatics* **9**.

Langfelder P, Zhang B, Horvath S. 2008. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics* **24**(5): 719-720.

Lee J-S, Smith E, Shilatifard A. 2010. The language of histone crosstalk. *Cell* **142**(5): 682-685.

Leek JT, Storey JD. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *Plos Genetics* **3**(9): 1724-1735.

Lengauer C, Kinzler KW, Vogelstein B. 1997. DNA methylation and genetic instability in colorectal cancer cells. *Proceedings of the National Academy of Sciences* **94**(6): 2545-2550.

Li C, Beroukhim R, Weir BA, Winckler W, Garraway LA, Sellers WR, Meyerson M. 2008. Major copy proportion analysis of tumor samples using SNP arrays. *Bmc Bioinformatics* **9**.

Li E, Beard C, Jaenisch R. 1993. Role for DNA methylation in genomic imprinting. *Nature* **366**(6453): 362-365.

Li Q, Seo J-H, Stranger B, McKenna A, Pe'er I, LaFramboise T, Brown M, Tyekucheva S, Freedman ML. 2013. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**(3): 633-641.

Li Y, Tesson BM, Churchill GA, Jansen RC. 2010. Critical reasoning on causal inference in genome-wide linkage and association studies. *Trends in genetics* **26**(12): 493-498.

Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH, Lee MP. 2003. Allelic variation in gene expression is common in the human genome. *Genome research* **13**(8): 1855-1862.

Lozzio CB, Lozzio BB. 1975. Human chronic myelogenous leukemia cell-line with positive Philadelphia chromosome. *Blood* **45**(3): 321-334.

Majewski J, Pastinen T. 2011. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends in Genetics* **27**(2): 72-79.

Marchini J, Howie B. 2010. Genotype imputation for genome-wide association studies. *Nature Reviews Genetics* **11**(7): 499-511.

Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, Fiegler H, Andrews TD, Stranger BE, Lynch AG, Dermitzakis ET et al. 2007. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biology* **8**(10).

Maunakea AK, Nagarajan RP, Bilenky M, Ballinger TJ, D'Souza C, Fouse SD, Johnson BE, Hong C, Nielsen C, Zhao Y. 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. *Nature* **466**(7303): 253-257.

McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, Wenger AM, Bejerano G. 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nature biotechnology* **28**(5): 495-501.

Milani L, Lundmark A, Nordlund J, Kiialainen A, Flaegstad T, Jonmundsson G, Kanerva J, Schmiegelow K, Gunderson KL, Lönnerholm G. 2009. Allele-specific gene expression patterns in primary leukemic cells reveal regulation of gene expression by CpG site methylation. *Genome research* **19**(1): 1-11.

Mitchell TM. 1997. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill* **45**.

Mohandas T, Sparkes R, Shapiro L. 1981. Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. *Science* **211**(4480): 393-396.

Mohn F, Schübeler D. 2009. Genetics and epigenetics: stability and plasticity during cellular differentiation. *Trends in Genetics* **25**(3): 129-136.

Myers RM Stamatoyannopoulos J Snyder M Dunham I Hardison RC Bernstein BE Gingeras TR Kent WJ Birney E Wold B et al. 2011. A User's Guide to the Encyclopedia of DNA Elements (ENCODE). *Plos Biology* **9**(4).

Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang LL, Hangaishi A, Kurokawa M, Chiba S, Bailey DK, Kennedy GC et al. 2005. A robust algorithm for copy number detection using high-density oligonucleotide single

nucleotide polymorphism genotyping arrays. *Cancer Research* **65**(14): 6071-6079.

Nazor KL, Altun G, Lynch C, Tran H, Harness JV, Slavin I, Garitaonandia I, Müller F-J, Wang Y-C, Boscolo FS. 2012. Recurrent variations in DNA methylation in human pluripotent stem cells and their differentiated derivatives. *Cell stem cell* **10**(5): 620-634.

Nica AC, Dermitzakis ET. 2013. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences* **368**(1620): 20120362.

Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, Travers M, Potter S, Grundberg E, Small K. 2011. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS genetics* **7**(2): e1002003.

Opitz L, Salinas-Riester G, Grade M, Jung K, Jo P, Emons G, Ghadimi BM, Beißbarth T, Gaedcke J. 2010. Impact of RNA degradation on gene expression profiling. *BMC medical genomics* **3**(1): 36.

Pai AA, Cain CE, Mizrahi-Man O, De Leon S, Lewellen N, Veyrieras J-B, Degner JF, Gaffney DJ, Pickrell JK, Stephens M. 2012. The contribution of RNA decay quantitative trait loci to inter-individual variation in steady-state gene expression levels. *PLoS genetics* **8**(10): e1003000.

Pant PK, Tao H, Beilharz EJ, Ballinger DG, Cox DR, Frazer KA. 2006. Analysis of allelic differential expression in human white blood cells. *Genome research* **16**(3): 331-339.

Pastinen T. 2010. Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews Genetics* **11**(8): 533-538.

Pastinen T, Ge B, Gurd S, Gaudin T, Dore C, Lemire M, Lepage P, Harmsen E, Hudson TJ. 2005. Mapping common regulatory variants to human haplotypes. *Human Molecular Genetics* **14**(24): 3963-3971.

Pastinen T, Hudson TJ. 2004. Cis-acting regulatory variation in the human genome. *Science* **306**(5696): 647-650.

Pastinen T, Sladek R, Gurd S, Sammak A, Ge B, Lepage P, Lavergne K, Villeneuve A, Gaudin T, Brandstrom H et al. 2004. A survey of genetic and epigenetic variation affecting human gene expression. *Physiological Genomics* **16**(2): 184-193.

Payer B, Lee JT. 2008. X Chromosome Dosage Compensation: How Mammals Keep the Balance. In *Annual Review of Genetics*, Vol 42, pp. 733-772. Annual Reviews, Palo Alto.

Pollard KS, Serre D, Wang X, Tao H, Grundberg E, Hudson TJ, Clark AG, Frazer K. 2008. A genome-wide approach to identifying novel-imprinted genes. *Human Genetics* **122**(6): 625-634.

Presson AP, Sobel EM, Papp JC, Suarez CJ, Whistler T, Rajeevan MS, Vernon SD, Horvath S. 2008. Integrated weighted gene co-expression network analysis with an application to chronic fatigue syndrome. *BMC systems biology* **2**(1): 95.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* **33**(suppl 1): D501-D504.

Rabiner L. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE* **77**(2): 257-286.

Ramasamy A, Trabzuni D, Ryten M, Weale ME, Hardy J. Misinterpretation of eQTL data.

Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlén S-E, Greco D, Söderhäll C, Scheynius A, Kere J. 2012. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PloS one* **7**(7): e41361.

Riggs AD. 1975. X inactivation, differentiation, and DNA methylation. *Cytogenetic and Genome Research* **14**(1): 9-25.

Risch N, Merikangas K. 1996. The future of genetic studies of complex human diseases. *Science* **273**(5281): 1516-1517.

Rockman MV, Kruglyak L. 2006. Genetics of global gene expression. *Nature Reviews Genetics* **7**(11): 862-872.

Rosenfeld JA, Wang ZB, Schones DE, Zhao K, DeSalle R, Zhang MQ. 2009. Determination of enriched histone modifications in non-genic portions of the human genome. *Bmc Genomics* **10**.

Rueda OM, Diaz-Uriarte R. 2007. Flexible and accurate detection of genomic copy-number changes from aCGH. *Plos Computational Biology* **3**(6): 1115-1122.

Salnikow K, Kluz T, Costa M, Piquemal D, Demidenko ZN, Xie K, Blagosklonny MV. 2002. The regulation of hypoxic genes by calcium involves c-Jun/AP-1, which cooperates with hypoxia-inducible factor 1 in response to hypoxia. *Molecular and cellular biology* **22**(6): 1734-1741.

Sandoval J, Heyn HA, Moran S, Serra-Musach J, Pujana MA, Bibikova M, Esteller M. 2011. Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6**(6): 692-702.

Schwarz G. 1978. Estimating the dimension of a model. *The annals of statistics* **6**(2): 461-464.

Serre D, Gurd S, Ge B, Sladek R, Sinnett D, Harmsen E, Bibikova M, Chudin E, Barker DL, Dickinson T et al. 2008. Differential allelic expression in the human genome: A robust approach to identify genetic and epigenetic Cis-acting mechanisms regulating gene expression. *Plos Genetics* **4**(2).

Shah SP. 2008. Computational methods for identification of recurrent copy number alteration patterns by array CGH. *Cytogenetic and Genome Research* **123**(1-4): 343-351.

Shah SP, Xuan X, DeLeeuw RJ, Khojasteh M, Lam WL, Ng R, Murphy KP. 2006. Integrating copy number polymorphisms into array CGH analysis using a robust HMM. *Bioinformatics* **22**(14): e431-e439.

Siepel A, Diekhans M, Brejova B, Langton L, Stevens M, Comstock CLG, Davis C, Ewing B, Oommen S, Lau C et al. 2007. Targeted discovery of novel human exons by comparative genomics. *Genome Research* **17**(12): 1763-1773.

Song L, Langfelder P, Horvath S. 2012. Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics* **13**(1): 328.

Spector TD, Williams FM. 2006. The UK adult twin registry (TwinsUK). *Twin Research and Human Genetics* **9**(06): 899-906.

Stadler MB, Murr R, Burger L, Ivanek R, Lienert F, Schöler A, van Nimwegen E, Wirbelauer C, Oakeley EJ, Gaidatzis D. 2011. DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**(16): 9440-9445.

Stranger BE, Nica AC, Forrest MS, Dimas A, Bird CP, Beazley C, Ingle CE, Dunning M, Flicek P, Koller D et al. 2007. Population genomics of human gene expression. *Nature Genetics* **39**(10): 1217-1224.

Suganuma T, Workman JL. 2011. Signals and combinatorial functions of histone modifications. *Annual review of biochemistry* **80**: 473-499.

Teo YY. 2008. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Current opinion in lipidology* **19**(2): 133-143.

Teo YY, Inouye M, Small KS, Gwilliam R, Deloukas P, Kwiatkowski DP, Clark TG. 2007. A genotype calling algorithm for the Illumina BeadArray platform. *Bioinformatics* **23**(20): 2741-2746.

Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, Sheffield NC, Stergachis AB, Wang H, Vernot B et al. 2012. The accessible chromatin landscape of the human genome. *Nature* **489**(7414): 75-82.

Tobi EW, Lumey L, Talens RP, Kremer D, Putter H, Stein AD, Slagboom PE, Heijmans BT. 2009. DNA methylation differences after exposure to prenatal famine are common and timing-and sex-specific. *Human molecular genetics* **18**(21): 4046-4053.

van Eijk KR, de Jong S, Boks MPM, Langeveld T, Colas F, Veldink JH, de Kovel CGF, Janson E, Strengman E, Langfelder P et al. 2012. Genetic analysis of DNA methylation and gene expression levels in whole blood of healthy human subjects. *Bmc Genomics* **13**.

Verlaan DJ, Ge B, Grundberg E, Hoberman R, Lam KCL, Koka V, Dias J, Gurd S, Martin NW, Mallmin H et al. 2009. Targeted screening of cis-regulatory variation in human haplotypes. *Genome Research* **19**(1): 118-127.

Veyrieras J-B, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK. 2008. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS genetics* **4**(10): e1000214.

Veyrieras JB, Gaffney DJ, Pickrell JK, Gilad Y, Stephens M, Pritchard JK. 2012. Exon-Specific QTLs Skew the Inferred Distribution of Expression QTLs Detected Using Gene Expression Array Data. *Plos One* **7**(2).

Villaret A, Galitzky J, Decaunes P, Estève D, Marques M-A, Sengenès C, Chiotasso P, Tchkonia T, Lafontan M, Kirkland JL. 2010. Adipose tissue endothelial cells from obese human subjects: differences among depots in angiogenic, metabolic, and inflammatory gene expression and cellular senescence. *Diabetes* **59**(11): 2755-2763.

Viterbi AJ. 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on* **13**(2): 260-269.

Wagner JR, Busche S, Ge B, Kwan T, Pastinen T, Blanchette M. 2014. The relationship between DNA methylation, genetic and expression inter-individual variation in untransformed human fibroblasts. *Genome biology* **15**(2): R37.

Wagner JR, Ge B, Pokholok D, Gunderson KL, Pastinen T, Blanchette M. 2010. Computational Analysis of Whole-Genome Differential Allelic Expression Data in Human. *Plos Computational Biology* **6**(7).

Wall JD, Pritchard JK. 2003. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Reviews Genetics* **4**(8): 587-597.

Walsh CP, Chaillet JR, Bestor TH. 1998. Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature genetics* **20**(2): 116-117.

Wang DG, Fan J-B, Siao C-J, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280**(5366): 1077-1082.

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SFA, Hakonarson H, Bucan M. 2007. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Research* **17**(11): 1665-1674.

Weisberg SP, McCann D, Desai M, Rosenbaum M, Leibel RL, Ferrante AW. 2003. Obesity is associated with macrophage accumulation in adipose tissue. *Journal of Clinical Investigation* **112**(12): 1796-1808.

Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, Klemm A, Flicek P, Manolio T, Hindorff L. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic acids research* **42**(D1): D1001-D1006.

Wu C, Carta R, Zhang L. 2005. Sequence dependence of cross-hybridization on short oligo microarrays. *Nucleic acids research* **33**(9): e84-e84.

Wu L-Y, Zhou X, Li F, Yang X, Chang C-C, Wong STC. 2009. Conditional random pattern algorithm for LOH inference and segmentation. *Bioinformatics* **25**(1): 61-67.

Xiang L, Kong B. 2013. PAX8 is a novel marker for differentiating between various types of tumor, particularly ovarian epithelial carcinomas (Review). *Oncology letters* **5**(3): 735-738.

Yau C, Holmes CC. 2008. CNV discovery using SNP genotyping arrays. *Cytogenetic and Genome Research* **123**(1-4): 307-312.

Yoder JA, Walsh CP, Bestor TH. 1997. Cytosine methylation and the ecology of intragenomic parasites. *Trends in Genetics* **13**(8): 335-340.

Yoon OK, Hsu TY, Im JH, Brem RB. 2012. Genetics and regulatory impact of alternative polyadenylation in human B-lymphoblastoid cells. *PLoS genetics* **8**(8): e1002882.

Yousefi P, Huen K, Schall RA, Decker A, Elboudwarej E, Quach H, Barcellos L, Holland N. 2013. Considerations for normalization of DNA methylation data by Illumina 450K BeadChip assay in population studies. *Epigenetics* **8**(11): 1141-1152.

Zhang B, Horvath S. 2005. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology* **4**(1).

Zhang DD, Cheng LJ, Badner JA, Chen C, Chen Q, Luo W, Craig DW, Redman M, Gershon ES, Liu CY. 2010. Genetic Control of Individual Differences in Gene-Specific Methylation in Human Brain. *American Journal of Human Genetics* **86**(3): 411-419.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**(9): R137.