

A NOVEL ASYMPTOTICALLY CORRECT STATISTIC FOR DETECTING
PAIRWISE AND HIGHER ORDER CONCORDANT EPISTASIS ACROSS
MULTIPLE QUANTITATIVE TRAITS

A Dissertation

Presented to the Faculty of the Weill Cornell Graduate School
of Medical Sciences

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

by

Mark C. Spurgeon

December 2018

©2018 Mark C. Spurgeon

A NOVEL ASYMPTOTICALLY CORRECT STATISTIC FOR DETECTING
PAIRWISE AND HIGHER ORDER CONCORDANT EPISTASIS ACROSS
MULTIPLE QUANTITATIVE TRAITS

Mark C. Spurgeon, Ph.D.

Cornell University 2018

Driven by the efficiency of DNA sequencing and related technologies, genome- and epigenome-wide association studies have already proven successful at producing specific results and general insights about the nature of genomic regulation. Discovery of expression Quantitative Trait Loci (eQTL), differentially methylated regions (DMRs), and other genomic and epigenetic features are proving integral to our understanding of how gene expression and DNA methylation (DNAm) are controlled throughout the human body and are changing how genomic and epigenetic data are analyzed in the study of cellular processes and complex diseases. Epistasis is the interaction among multiple genetic loci in their effect on gene expression. While epistasis is pervasive in biological systems and has the potential to account for heritability in traits that remain unexplained by the sum of main effects, the computational and statistical challenges of epistasis detection are daunting. We present the F-test of magnitude and concordance (Fomac) - a novel statistic that detects concordant epistasis across multiple datasets or co-expressed genes by constraining linear model parameters to be both significant and consistent. Simulations were carried out to compare the performance of Fomac to that of comparable methods for detecting single- and multi-trait epistasis, and they showed that Fomac is able to leverage concordant effects for improved statistical power. Fomac was also ap-

plied to gene expression from the Multiple Tissue Human Expression Resource (MuTHER) where a genome-wide analysis across 3 tissues identified 2754 examples of gene-wise Bonferroni-significant concordant epistasis. Epigenome-wide association studies (EWAS) are providing another angle from which to view genetic regulation. We performed an EWAS comparing the methylome of circulating monocytes in patients with and without Charcot foot (a devastating complication of diabetes.) Increased osteoclast activity has a role in the disease, and osteoclasts derived from monocytes are particularly well-suited for such a role. We observed that the methylome of these monocytes was significantly different in patients with and without Charcot foot, and identified specific genes with aberrant methylation. Together, the studies described in this dissertation serve the notion that by understanding relationships between and within omics data, we can both glean useful insights into specific regulatory mechanisms of the cell and apply patterns to accurately predict biological responses.

BIOGRAPHICAL SKETCH

Mark C. Spurgeon was born in Fairfax, Virginia and moved to Orchard Park, New York when he was three years old. After completing his high school education at Orchard Park High School, he studied physics & astronomy at the State University of New York College at Geneseo, graduating with a BA in physics. He joined the lab of Dr. Jason G. Mezey to study quantitative genomics at Weill Cornell Graduate School of Medical Sciences program in Physiology, Biophysics, and Systems Biology.

ACKNOWLEDGEMENTS

This doctorate was far from an individual pursuit. In fact, support from mentors and colleagues has been essential to my enjoyment and the personal development that I have undergone during this process. Whether these contributions have revealed where I must bolster my knowledge, injected creativity into my scientific endeavours, or encouraged me to persevere during challenging times, this achievement is shared.

My sincere gratitude to Dr. Jason Mezey for accepting me into the lab and providing a place for me to better myself and contribute to a team. I am extremely lucky to have found an adviser who is able to get the best out of me while still allowing me to follow my meandering curiosity through research. Your unwavering confidence in me has helped guide me through the challenges that I have faced during graduate school.

I am extremely grateful for the patience, time, energy, and attention that my committee members Dr. Fabien Campagne, Dr. Olivier Elemento, and Dr. Christina Leslie have committed to help guide me with my research.

I am thankful to have received instruction, advice, and support from Juan Rodriguez-Flores pertaining to my lab rotation project and later with our epigenetics project. I would like to thank our collaborators Charbel Abi-Khalil and Jennifer Pasquier at Weill Cornell Qatar for their patience and wisdom.

Research is collaborative, and that starts with lab members past and present: Francisco Agosto-Perez, Abishek Sainath Madduri, Monica Ramstetter, Sarah Brooks, Jin-Hyun Ju, Sushila Shenoy, Thomas Vincent, Afrah Shafquat, Zijun Zhao. Every one of you has made tangible contributions to my graduate work, and to my enjoyment of graduate school.

My friends and family have contributed in so many ways to my success

that it's difficult to disentangle my successes from theirs. I would like to thank everyone who helped foster my sense of empowered curiosity - especially my teachers, professors, and parents.

TABLE OF CONTENTS

Biographical Sketch	iii
Acknowledgements	iv
List of Tables	viii
List of Figures	ix
1 Introduction	1
1.1 The Omics Revolution	1
1.2 Genome-Wide Association Studies	3
1.2.1 Expression Quantitative Trait Loci	4
What have we learned from eQTL studies?	7
Challenges of eQTL Studies	9
1.2.2 Epistasis	10
Mathematical Formulation	13
Challenges of Epistasis Detection	14
Approaches to Epistasis Detection	16
Epistasis and Pleiotropy	17
Replication	19
Epistasis in Multiple Traits	22
1.3 Epigenome-Wide Association Studies	23
1.3.1 DNA Methylation	23
1.3.2 Role in Disease	24
1.3.3 Performing EWAS	25
1.4 Overview of Dissertation	26
1.4.1 Chapter 2 - <i>Fomac</i> (<i>F</i> -test of Magnitude and Concordance)	27
1.4.2 Chapter 3 - Connections Between CpG Methylation and Charcot Foot	27
2 A Novel Asymptotically Correct Statistic for Detecting Pairwise and Higher Order Concordant Epistasis Across Multiple Quantitative Traits	28
2.1 Introduction	29
2.2 Methods	32
2.2.1 The <i>Fomac</i> Framework	32
2.2.2 The <i>Fomac</i> Test Statistic	35
2.2.3 Correcting for Correlated Traits	36
2.2.4 Application to Epistasis/Generation of Epistasis Features	37
2.2.5 Comparing Performance	39
2.2.6 Simulations	40
2.2.7 Application to Human Data	41
2.3 Results	43
2.3.1 Simulations demonstrate sensitivity to inter-trait correla- tions and unbalanced designs	43

2.3.2	Performance Comparison	44
2.3.3	Human Data	45
	Quality Control	45
	Significant Tests	48
2.4	Discussion	49
3	Whole-methylome analysis of circulating monocytes in acute diabetic Charcot foot reveals the presence of differentially methylated genes involved in migration, differentiation and formation of osteoclasts	51
3.1	Introduction	52
3.2	Methods	53
3.2.1	Subjects	53
3.2.2	Monocytes Isolation and DNA/RNA Extraction	55
3.2.3	Enhanced Reduced Representation Bisulfite Sequencing and Data Processing	55
3.2.4	Differential Methylation	56
3.2.5	Gene Expression Data	60
3.3	Results	62
3.3.1	Comparing Diabetic Patients With and Without Neuropathy to Patients With Charcot Foot	63
3.3.2	Gene-mapped Differential Methylation in Patients with Diabetes and Charcot Foot Compared to Patients with Diabetes But no Charcot Foot	66
3.3.3	Association between DNA methylation and gene expression in Charcot foot patients	68
3.4	Discussion	69
A	Chapter 1 of Appendix	75
B	Chapter 3 of Appendix	77
	Bibliography	82

LIST OF TABLES

1.1	eQTL Datasets and Cell Lines Small sampling of available eQTL datasets for different tissues, sample sizes, and genotyping and gene-expression measurement platforms	9
2.1	Simulation Parameter Configurations	41
3.1	Baseline characteristics of the participants included in the study. Diabetes but no neuropathy (D), diabetes with neuropathy (DN), and diabetes with both neuropathy and Charcot foot (DCh). Data are represented as mean (standard deviation). p-values were calculated with ANOVA test.	54
3.2	Genome-wide methylation study on Charcot foot. In order to identify methylation differences specific to Charcot foot, circulating monocytes were isolated from blood of patients with diabetes but no neuropathy (D), diabetes with neuropathy (DN), and diabetes with both neuropathy and Charcot foot (DCh). CpG methylation data was produced by enhanced reduced representation bisulfite sequencing (ERRBS). Each group has n = 18 and patients are matched for age, gender, BMI, and HbA1c across all groups such that these covariates were not significantly different among groups (rightmost column). Methylation of a gene was calculated as the mean (weighted by coverage) observed CpG sites within 2kb upstream and 2kb downstream of transcription start site, as determined by Ensembl annotation. The bottom portion of the table provides summary statistics of the number of CpG sites within the mapping interval of a gene, and the total number of CpG sites with coverage of at least 10 (sites with coverage less than 10 were not used).	57
3.3	Differential methylation results for 4 groupings of patients. Results of 4 differential methylation approaches: methylome-wide differential methylation (p-value derived from multivariate linear regression fit using the first 3 principal components as dependent variables as well as batch and group as independent variables) and individual site/gene differential methylation for both CpG site (top 3 rows) and gene-mapped methylation (bottom 3 rows). For each approach four group comparisons were made, represented by the four columns of the table which are labeled based on the two patient groups that were compared. . .	63

LIST OF FIGURES

1.1	Types of eQTL Main Effects. Top row: scatterplots, each showing gene expression across the three genotype classes. Bottom row: genotype-phenotype (GP) maps visualizing the mean gene expression using color for each genotype class with blue indicating low expression, black indicating global mean expression, and yellow indicating high expression. An eQTL exists when the mean expression of at least one genotype class is different than the global mean (center and right columns), but no eQTL exists when all genotype classes have the same mean gene expression (left column). Main effects are decomposed into orthogonal additive (center) and dominance (right) effects.	5
1.2	Genotype-phenotype (GP) map showing color-coded mean expression values for each of the 9 genotype classes defined by the joint states of two genotypes (top right). Lack of main effects (below and to the left of GP map) means that the pattern visible in the GP map is due to epistasis.	12
1.3	Examples of Epistatic Effects (Using the genotype codings defined just above, epistasis effects were generated and visualized using GP maps)	15
1.4	Horizontal and Vertical Pleiotropy With and Without Epistasis. In horizontal pleiotropy a genetic variant (G) directly regulates multiple genes (P1 & P2). In vertical pleiotropy G regulates P1 directly and is associated with P2 by way of P1. Multiple genetic variants can also exhibit pleiotropy, as shown in the bottom two illustrations. In horizontal epistatic pleiotropy multiple genetic variants (G1 & G2) directly regulate P1 and P2 in a non-additive manner. In vertical epistatic pleiotropy G1 and G2 regulate P1 directly and are associated in a non-additive manner with P2 by way of P1.	19
1.5	Concordant and non-concordant epistatic effects	21
2.1	The Fomac Framework. Top Left: Fomac accepts a set of r traits and g genotypes. Top Right: Fomac can leverage highly parameterized models, so genotype variables can be readily transformed into several epistasis features. Bottom Left: Multivariate linear regression produces parameter estimates and estimates for distributions of these parameter estimates. Bottom Right: Fomac operates on these parameters to assess how concordant and significant epistatic effects are across traits.	33
2.2	Venn Diagram of Sample/Gene Overlap Across Datasets . . .	42
2.3	Simulation ROC Performance. Parameters: $r = 4, g = 2, n = 900$	46

2.4	Global QQ Plots for uncorrected (left) and main effect-corrected (right) p-values.	47
2.5	Gene-wise QQ Plots for unadjusted (top left) and main-effect adjusted p-values (top right)	50
3.1	CpG site and Gene-mapped differential methylation in patients with diabetes and CF compared to patients with diabetes but no CF. A. The first two principal components of autosomal gene methylation, as calculated by singular value decomposition. Samples are colored by group: diabetes in blue, diabetes with neuropathy in purple, and diabetes with CF in red. B. Wholemethylome signal as captured by the first three principal components (horizontal axes) of the displayed subset of patients. Violin plots representing the distribution of a particular patient group along a principal component are adjacent to their corresponding group. C. The first two principal components of autosomal gene methylation, as calculated by singular value decomposition. Samples are colored by group: diabetes in blue, diabetes with neuropathy in purple, and diabetes with CF in red. D. Whole-methylome signal as captured by the first three principal components (horizontal axes) of the displayed subset of patients. Violin plots representing the distribution of a particular patient group along a principal component are adjacent to their corresponding group. E. Chromosomal distribution of gene methylation differences. For each gene, the significance is displayed on the y-axis as the \log_{10} of the p-value. The results are ordered along the x-axis by chromosome, with each bar representing a different chromosome. The Bonferroni and BH p-value thresholds ($\alpha = 0.05$) were 2.3×10^{-6} (yellow line) and 5.8×10^{-3} (green dashed line), respectively. F. Chromosomal distribution of gene methylation differences. For each gene, the significance is displayed on the y-axis as the \log_{10} of the p-value. The results are ordered along the x-axis by chromosome, with each bar representing a different chromosome. The Bonferroni and Benjamini-Hochberg p-value thresholds ($\alpha = 0.05$) were 4.0×10^{-8} (yellow line) and 5.2×10^{-7} (green dashed line), respectively.	65

3.2	Gene analysis for gene-mapped differential methylation in patients with diabetes and Charcot foot compared to patients with diabetes but no Charcot foot. A. Graphic representation of the number of hypo- and hyper-methylated genes. B. Representation of the 10 top hypermethylated (top part, red) and hypomethylated (bottom part, green) genes in CF patients compared to patients with diabetes but no Charcot foot. C-D. Networks of altered genes mapped differential methylation created by IPA. The hypothetical networks generated by IPA based on the molecular relationships, interactions, and pathway associations between the methylated candidate genes are shown in a graphical representation.	67
3.3	Association between methylation and expression in Charcot foot. A. Schematic representation of CIS and TRANS association between methylated and expressed gene. B. Minimum network of protein-protein interaction using the 27 CF differentially methylated genes and their 24 expression associated genes (total of 51 genes) leading to 32 interactions using STRING.	70
A.1	Joint Methylation & Expression Analysis Pipeline Visualization of the pipeline which takes enhanced reduced representation bisulfite sequencing (ERRBS) CpG methylation data and integrates it with RNA-Seq gene expression data to produce data-driven candidate regulatory genes and sites	76
B.1	Gating strategy for monocyte sorting. A. Monocyte populations were gated (red population) using SSC/FSC. B-C Using FSC-W/FSC-h (B) and SSC-W/SSC-H (C), the doublets were excluded and only the living cells (blue population) were kept. D. Auto-fluorescent cells were excluded using Pacific-Blue channel. E. Final monocyte population was gated (orange population) as CD14(APC)+CD16(PE)+/-.	77

B.2	<p>A. Histogram of $\log_{10}(\text{coverage})$ at all CpG sites. The blue histogram represents the mean of all values in that bin over all samples. Each black line represents the counts from 1 of the 54 individual samples. B. Histogram of methylation -values at all CpG sites. The blue histogram represents the mean of all values in that bin over all samples. Each black line represents the counts from 1 of the 54 individual samples. C-D. The first two principal components of autosomal gene methylation for both (C) the CpG site and (D) gene-mapped analyses, as calculated by singular value decomposition. Samples are colored by batch: pink samples were collected in June 2014, orange samples were collected in February 2015, and gray samples were collected in March 2015.</p>	78
B.3	<p>Differential Methylation for all two-way comparisons of the three groups using the CpG site approach. First row (A, B, C): whole-methylome signal as captured by the first three principal components (horizontal axes) of the displayed subset of patients. Violin plots representing the distribution of a particular patient group along a principal component are adjacent to their corresponding group. Second row (D, E, F): chromosomal distribution of gene methylation differences. For each gene, the significance is displayed on the y-axis as the \log_{10} of the p-value. The results are ordered along the x-axis by chromosome, with each bar representing a different chromosome. Bonferroni and Benjamini-Hochberg p-value thresholds are displayed as blue and yellow (dashed) lines, respectively.</p>	79
B.4	<p>Differential Methylation for all two-way comparisons of the three groups using the gene-mapped approach. First row (A, B, C): whole-methylome signal as captured by the first three principal components (horizontal axes) of the displayed subset of patients. Violin plots representing the distribution of a particular patient group along a principal component are adjacent to their corresponding group. Second row (D, E, F): chromosomal distribution of gene methylation differences. For each gene, the significance is displayed on the y-axis as the \log_{10} of the p-value. The results are ordered along the x-axis by chromosome, with each bar representing a different chromosome. Bonferroni and Benjamini-Hochberg p-value thresholds are displayed as blue and yellow (dashed) lines, respectively.</p>	80

B.5 Quantile-quantile plots for all two-way comparisons of the three groups using both CpG site (A, B, C, D) and gene-mapped (E, F, G, H) differential methylation in patients with diabetes and CF compared to patients with diabetes but no CF. Bonferroni and Benjamini-Hochberg p-value thresholds are displayed as blue and yellow (dashed) lines, respectively. Differential methylation on permuted data was calculated 10 times for each QQ plot, resulting in 10 sets of null p-values that are plotted in cyan alongside the non-permuted p-values. 81

CHAPTER 1

INTRODUCTION

It is like a voyage of discovery into unknown lands, seeking not for new territory but for new knowledge. It should appeal to those with a good sense of adventure.

Frederick Sanger, Nobel Banquet, December 10, 1980

1.1 The Omics Revolution

In 1990, the Human Genome Project (HGP) set out to map the 3.4 billion bases that comprise the human genome. The first genome was published in 2004, providing researchers with the ability to address new fundamental questions in genomics[29]. Despite the fact that most of the original reference genome was pieced together using early-generation DNA sequencing and carried a rich \$2.3 billion price tag, the ambition of the HGP stoked a yearning for innovation in cost, speed, scale, and accuracy of sequencing technologies[87]. By 2014, the cost of sequencing a human genome had dropped to \$1000 and took only a couple of days[77].

Today, sequencing technologies continue to progress - a genome costing \$100 and ready in a few hours is expected to be available within a decade[83] - and the breadth of applications continues to grow. Pulled along by advances in DNA sequencing techniques, assays have been established for probing the transcriptome[131], epigenome, microRNAome[144], and too many more omes to name[8, 163, 73]. Microarray and more recently next-generation sequencing techniques have been adapted for high-throughput measurement of gene expression[141], microRNA[19], and DNA methylation[42]. Mass spectrometry

and nuclear magnetic resonance have been applied to produce metabolomic[38] and proteomic[1] data. Each type of data provides a snapshot of certain inter-related components of the complex regulatory network within cells, making it possible to interrogate this cellular machinery on a huge scale at both the population and single-cell levels[167]. The diverse quality, high production rates, and raw size of omics data situates genomics squarely in "big data" territory. Generally "big data" refers to datasets large enough that traditional data processing and analysis methods are insufficient[13]. The torrent of omics data has laid fertile ground for the development of computational and statistical methods designed to glean valuable insights into biology[126, 48], facilitate the transition from measurement to analysis[21], or feed back to improve the omics technologies themselves[40]. Substantial effort has already been put into integrating information across datasets[192], and across[205] and within[142] omes with the ultimate goal of filling in the gaps in our understanding of cellular regulation. Still, we face great challenges in acquisition, storage, distribution, and analysis of omics data. Sequencing technologies continue to improve and the demand for their services are growing - estimates range from 100 million to 2 billion human genomes being sequenced by 2025[213] - and that's only genomic data. The multitude of omics data are derived from different but biologically related layers of cellular regulation, meaning that there will be a greater demand to integrate information from multiple datasets[108]. Time-series measurements, measurements over multiple tissues, and biological replicates among other designs further increase the dimensionality of this already heterogeneous mass of data.

Uncovering the regulatory processes responsible for cellular and organismal function is a fundamental goal in biology[16]. These insights underpin

disease prevention and treatment of all types[116]. The richness of omics data has already allowed us to address a broad range of questions, from basic science inquiries like when early humans left Africa[197] to mechanisms of post-transcriptional regulation in Huntington's disease[88] and Parkinson's disease[89], and translational medicine advances like targeted therapies for cancer[71, 146]. Genomic data alone can be used to test phylogenetic hypotheses of agricultural significance[201]. Genomic data can be paired with readily available clinical data to identify causal/risk loci in complex phenotypes[228]. Transcriptomic (gene expression) data can be paired with genomic data to identify genetic markers that are highly predictive of gene expression[171]. Metabolomic data can be integrated with transcriptomic data to identify gene-to-metabolite networks[84]. Each of these approaches involves identifying relationships within and/or among omics datasets, and each interaction type can shed light on a slightly different part of cellular regulation. This dissertation will focus on (1) associations between and within the genome and transcriptome, (2) interactions between the methylome and clinical phenotypes, and touch on interactions between the methylome and transcriptome.

1.2 Genome-Wide Association Studies

As DNA sequencing throughput increased, it became possible to sequence thousands of genetic markers, including single nucleotide polymorphisms (SNPs) in a short period of time for many samples[151]. Genome-wide association studies emerged as a means to comprehensively test the genome for SNPs in close proximity to a genetic locus causal to qualitative or quantitative phenotypes (traits)[228]. Since it is rare for traits to feed back to alter the genetic locus it-

self, we assume the primary direction of regulation is that the locus causes disease/disease risk. We often refer to such an occurrence as a genetic effect or genomic association. An advantage of GWAS over previous association studies is that it represents an unbiased and comprehensive option that can be performed without a priori knowledge of causal genes[67]. Thousands of GWAS studies have been published[117] and have identified numerous genetic loci associated with complex traits like BMI[212], as well as risk loci and therapeutic targets for complex (non-Mendelian) diseases like coronary artery disease[222, 125, 206]. GWAS are allowing us to understand contributions to complex disease from SNPs across the genome[140]. Still, relatively little is known about the genes and genetic variants influencing susceptibility to common human diseases[5]. The majority of SNPs identified by GWAS are located in non-coding regions of the genome[143], which suggests a role in transcriptional regulation. Trait-associated SNPs are also more likely to influence the expression of one or more genes[169] than non-trait-associated SNPs, suggesting that there is a link between the genetic regulation of gene expression and that of complex traits[30].

1.2.1 Expression Quantitative Trait Loci

As DNA microarrays[200] and whole transcriptome shotgun sequencing (RNA-Seq)[231] technologies were developed, it became possible to test millions of genomic markers for association with tens of thousands of gene expression levels. These associations are called expression quantitative trait loci (eQTL) (Figure 1.1), and studying them is particularly insightful since intermediate phenotypes like gene expression are often separated from genetic influence by fewer regulatory steps compared to complex traits[68]. Therefore, a genetic effect should

be more easily detectable at the gene expression level than at the level of any complex (non-Mendelian) disease for example[255]. Given that GWAS have already identified many genomic loci involved in disease/disease risk, and differential expression analyses have long investigated the transcriptome's role in disease[6], understanding how a genomic locus regulates gene expression can help reveal mechanisms involved in a phenotype that is also impacted by that same genetic locus. In other words, eQTL can help us probe how the genome impacts disease through transcriptional regulation and other mechanisms.

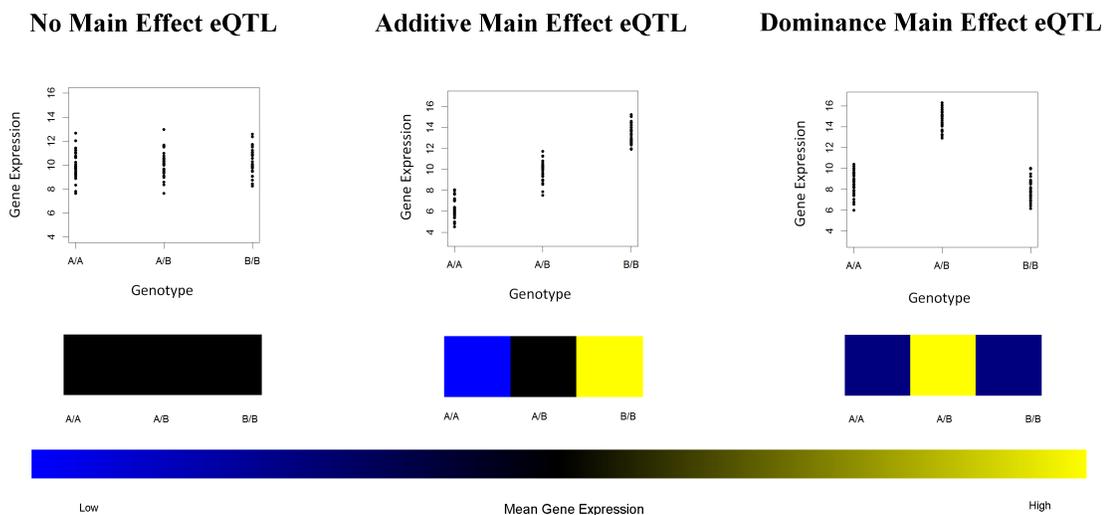


Figure 1.1: **Types of eQTL Main Effects.** Top row: scatterplots, each showing gene expression across the three genotype classes. Bottom row: genotype-phenotype (GP) maps visualizing the mean gene expression using color for each genotype class with blue indicating low expression, black indicating global mean expression, and yellow indicating high expression. An eQTL exists when the mean expression of at least one genotype class is different than the global mean (center and right columns), but no eQTL exists when all genotype classes have the same mean gene expression (left column). Main effects are decomposed into orthogonal additive (center) and dominance (right) effects.

Since genetic loci can take on one of three values in genetic variants, they can be represented as a 3-category variable or as 2 orthogonal ordinal "dummy" variables. In Figure 1.1, the center and right plots contain additive and dominance main effects, respectively. Main effect describes a statistical association between a single SNP and the trait. Decomposing the 3-category genotype variable into 2 "dummy" variables that capture these additive and dominance effects creates a suitable design for linear regression, which models gene expression as a linear combination of one or both of these "dummy" variables[69]. PLINK[184], a widely distributed software for efficient genomic computations, defines the additive coding of the g th genotype as

$$G_{g,a} = \begin{cases} 0 & \text{if } G_g = AA \\ 1 & \text{if } G_g = Aa \\ 2 & \text{if } G_g = aa \end{cases}$$

such that the additive coding captures how many reference alleles an individual has. Still another additive coding for the g th genotype is

$$G_{g,a} = \begin{cases} -1 & \text{if } G_g = AA \\ 0 & \text{if } G_g = Aa \\ 1 & \text{if } G_g = aa \end{cases}$$

which will capture the same main effect as PLINK's coding, but will diverge from PLINK's outcome when genotype interactions are considered. Still, this ordinal coding will not be sensitive to dominance effects where the mean expression of heterozygotes is greater than that of either homozygote. Therefore, dominance codings are often defined for the g th genotype as

$$G_{g,d} = \begin{cases} 0 & \text{if } G_g = AA \\ 1 & \text{if } G_g = Aa \\ 0 & \text{if } G_g = aa \end{cases}$$

or

$$G_{g,d} = \begin{cases} -1 & \text{if } G_g = AA \\ 1 & \text{if } G_g = Aa \\ -1 & \text{if } G_g = aa \end{cases}$$

Analogously to the additive case, both of these dominance codings will capture the same main effect but will diverge when genotype interactions are considered. Other dominance codings with identical properties that capture different scalings of the same effect exist in the literature(citation).

Most current fast implementations of genome-wide eQTL analyses either use this linear regression approach or a derivative of analysis of variance (ANOVA) to calculate p-values for all of the billions of possible genotype-gene expression combinations[65].

What have we learned from eQTL studies?

Many eQTL have been identified in plants[234, 41], mice[112], humans[226, 122, 180, 166, 10, 247, 26, 214, 50], and have also been specifically validated in yeast[18]. eQTL where the locus is proximal to the gene it regulates are called cis-eQTL, otherwise they are called trans-eQTL. Although cis-eQTL appear to be more abundant[66] and to have larger effect sizes[176] than trans-eQTL, cis-eQTL are also preferentially advantaged in both of these areas due to many

studies limiting the scope of their search to cis-genomic regions in order to limit multiple test correction. Cis-eQTL also appear to replicate more readily across independent datasets than trans-eQTL[190], suggesting that trans-eQTL may be more sensitive to experimental conditions like tissue heterogeneity[183] than cis-eQTL.

There is evidence of connections between loci involved in disease and eQTL: one study reported that 23.1% of catalogued GWAS hits for adult-onset neurological disorders showed up as a cis-eQTL signal in their analysis of various brain tissues[190]. For this reason, eQTL and GWAS have been integrated to help understand complex disease [154, 129, 171]. Systems genetics, which seeks a broad view of the molecular basis of complex traits, incorporates eQTL results with those from other layers of regulation to infer directional expression networks[27]. Studies involving eQTL show enrichment for eQTL in methylated DNA regions[12], regions of open chromatin[62], and transcription factor binding sites[37], which all serve to verify that results from eQTL studies are meaningful in the context of the integrative biological picture that we're trying to uncover.

At least 50 eQTL datasets are now available, ranging from consortia of hundreds of samples across at least 40 tissues[134] to single studies with thousands of samples of genome-wide genotypes and gene expression[239]. In the abundance of eQTL datasets (those with thousands of genomic loci and expression measurements for thousands of genes), numerous methods have emerged to facilitate the discovery of expression-impacting SNPs on a new whole-genome whole-transcriptome scale[94, 215].

Table 1.1: **eQTL Datasets and Cell Lines** Small sampling of available eQTL datasets for different tissues, sample sizes, and genotyping and gene-expression measurement platforms

Type	Study	<i>n</i>	G	E	Cell Lines
Adipose (Fat)	GTE _x	110	Array ^{1,2}	Array ^{1,6}	ASC52telo; ASC57telo; ASC-Bmi-1/hTERT
	MuTHER	776	Array ³⁻⁶	Array ²	
B cells (LCL)	MuTHER	777	Array ³⁻⁶	Array ²	COLO 829BL; HCC1007 BL; HCC1143 BL
	Geuvadis	344	WGS ^{11,12}	RNA-seq ⁶	
	HapMap (x7)	79-107	WGS ^{11,12}	Array ³	
Blood	GTE _x	167	Array ^{1,2}	Array ^{1,6}	D1.1; TALL-104; Jurkat, Clone E6-1
	DGN	922	Array ⁸	RNA-seq ⁶	
Breast	GTE _x	56	Array ^{1,2}	Array ^{1,6}	CCD-1059Sk; hTERT-HME1[ME16C]; MCF-12A
	Healthy TCGA	54	Array ^{9,10}	RNA-seq ^{3,4}	
Lung	GTE _x	123	Array ^{1,2}	Array ^{1,6}	MRC-5; BCi-NS1.1; HULEC-5a
	Healthy TCGA	40	Array ^{9,10}	RNA-seq ^{3,4}	
Skin	GTE _x	113	Array ^{1,2}	Array ^{1,6}	BJ-5ta; CCD 1102 KERT _r ; TIME
	MuTHER	667	Array ³⁻⁶	Array ²	
x14	GTE _x	>40	Array ^{1,2}	Array ^{1,6}	-

Type Sampled *in vivo* except for B cells (lymphoblastoid cell lines)

Study Name of the eQTL study

n Sample size (HapMap by population)

G Genotypes: Illumina (Ex¹, 5M², 1.2M³, 1M⁴, H300⁵, H610⁶, 2.5M⁷, Q⁸, H300⁹), Affy 6¹⁰, WGS (1000G¹¹, Hapmap¹²)

E Gene expression: Affy HG1.1¹, Illumina HT-12 V3², Sentrix V6³, Affy U133⁴, Agilent 244K⁵, RNA-Seq (HiSeq 2000⁶)

Cell Lines Commercially available from ATCC that will be used as a model of the tissue type

Challenges of eQTL Studies

Heritability is the proportion of phenotypic variation that is due to genetic variation, and is a central question in the modeling and prediction of complex phenotypes[238]. Most heritability in gene expression is not explained by detectable eQTLs[18] or epigenetic inheritance[183], suggesting that gene expression is influenced by many genetic loci of small effect, including many trans-eQTL.

As most eQTL studies transitioned to true genome-wide searches including both cis- and trans-loci, the challenges of working with "big data" took hold[95]. Genome-wide studies involving hundreds of thousands of SNPs and thousands of genes are understood to be a case of "large *p*, small *n*", which refers to when

the number of features p greatly exceeds the number of samples n [23]. Many computational[205] and statistical[185] approaches have popped up to deal with the difficulties of looking for genomic effects in such a huge search space[2].

1.2.2 Epistasis

While many eQTL involving a single locus impacting gene expression (referred to as marginal/additive effects) have been identified in a variety of biological systems [180, 156, 75], a large portion of heritability remains unexplained by these additive effects[140, 227]. It is clear that focusing on main effects is likely an oversimplification of the underlying biology since phenotypes are often affected by multiple genes in complex ways[43, 56].

Epistasis describes when the effect of a genetic locus on a trait is modulated by other genetic loci - a concept central to genetic regulation[178] - and known to play a role in molecular evolution[17], protein evolution[145], and the evolution of transcriptional networks[211]. The term has its roots in the early 20th century when Bateson described epistasis as an allelic effect at one locus being masked by the effect of another allele at a different locus[9] - a biological phenomenon and a well-developed area of research[7, 177, 74]. A decade later, Fisher defined epistasis as any statistical deviation from the sum of strictly additive genetic effects of two loci in their impact on a trait[55].

Despite the common name and similar sounding descriptions, the two definitions are not equivalent. Bateson's is a biological definition applying at the level of the individual whereas Fisher's is a statistical definition applied at the population level. We do know that it is possible for biological epistasis to be

present in the absence of statistical epistasis[159], but the degree to which statistical evidence of epistasis at the population level constitutes evidence of biological epistasis is still an active area of debate[54, 32, 160].

Statistical epistasis can be visualized through a contingency table that shows the mean phenotype value for all combinations of g genotypes, each of which can take on 3 values in diploid cells. The table therefore has g dimensions and 3 categories per dimension for a total of 3^g entries (2-way epistasis example in Figure 1.2). A genetic effect is therefore represented by a difference in the mean phenotype value among any of the 3^g genotype classes. Such an effect will be a combination of main effects and epistatic effects, and can be deconstructed into features that capture either purely main effects or purely epistatic effects (Figure 1.2) with the proper parameterization[102]. If distinguishing main effects from epistatic effects is not a priority, a general genetic effect can be tested for by employing an ANOVA on the 3^g groups[80].

From early on, accounting for epistatic interactions improved linkage studies in inflammatory bowel disease[25], type 2 diabetes[33], and other diseases[193]. Accounting for genetic interactions has already improved modeling of quantitative traits in yeast[57], and there is compelling evidence that epistasis plays a role in Alzheimer's disease[44, 20]. Epistasis has been utilized to successfully improve models of genetic regulation of clinical disease[148] and intermediate traits like gene expression (eQTL epistasis). Most quantitative trait analyses in animal[103, 96, 207] and plant[104] models suggest that statistical epistasis is widespread.

While epistasis can be detected in linkage or association analyses, and in qualitative or quantitative traits, this discussion specifically deals with eQTL

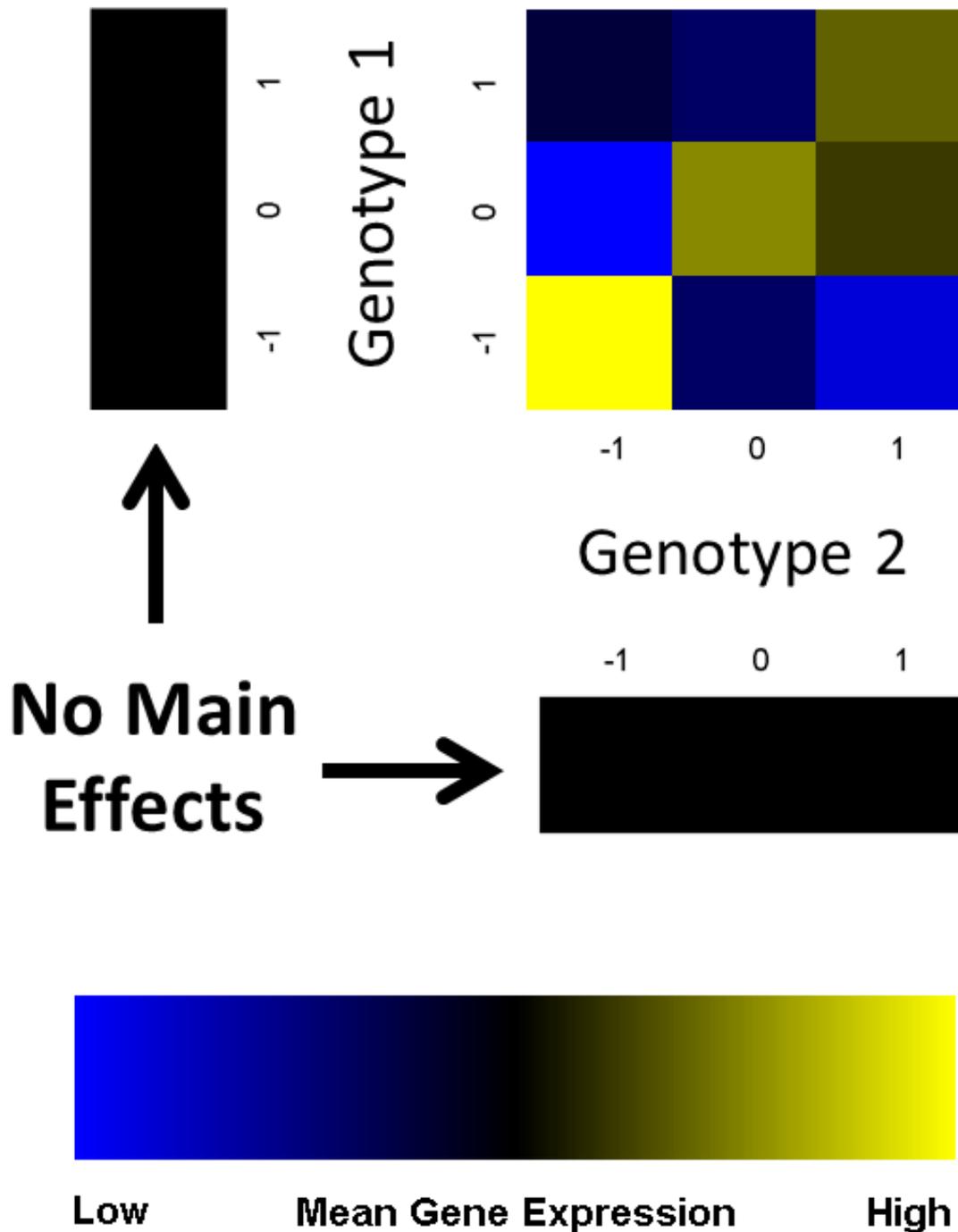


Figure 1.2: **Genotype-phenotype (GP) map** showing color-coded mean expression values for each of the 9 genotype classes defined by the joint states of two genotypes (top right). Lack of main effects (below and to the left of GP map) means that the pattern visible in the GP map is due to epistasis.

epistasis which is when the effect of a genetic locus on gene expression is modulated by other genetic loci[123].

Mathematical Formulation

Fisher's work gave rise to the mathematical definition of epistasis most commonly used by quantitative genomicists today - deviation from additivity (main effects) in the effects of multiple loci on a trait. As a baseline for this additivity, a set of features spanning all possible types of main effects is first defined. Consider two genetic loci G_1 and G_2 each of which can take on three possible values in a diploid organism. The loci are each transformed into two dummy variables which encode the same information as the original categorical coding. Define the additive coding of the g th genotype as

$$G_{g,a} = \begin{cases} -1 & \text{if } G_g = AA \\ 0 & \text{if } G_g = Aa \\ 1 & \text{if } G_g = aa \end{cases}$$

and the dominant coding of the g th genotype as

$$G_{g,d} = \begin{cases} -0.5 & \text{if } G_g = AA \\ 1 & \text{if } G_g = Aa \\ -0.5 & \text{if } G_g = aa \end{cases}$$

Consider a quantitative phenotype Y that we are interested in testing for association with these traits jointly. A full two-locus linear regression for main effects has the form

$$Y = \beta_0 + \beta_{1,a}G_{1,a} + \beta_{2,a}G_{2,a} + \beta_{1,d}G_{1,d} + \beta_{2,d}G_{2,d} + \epsilon$$

where β represents the strength of a genetic effect and ϵ is error. A full two-locus linear regression accounting for epistasis has the form

$$Y = \beta_0 + \beta_{1,a}G_{1,a} + \beta_{2,a}G_{2,a} + \beta_{1,d}G_{1,d} + \beta_{2,d}G_{2,d} + \beta_{a,a}G_{1,a}G_{2,a} \\ + \beta_{a,d}G_{1,a}G_{2,d} + \beta_{d,a}G_{1,d}G_{2,a} + \beta_{d,d}G_{1,d}G_{2,d} + \epsilon$$

If dummy coding and model parameterization is performed properly, a set of genomic features can be generated such that all epistatic features are orthogonal to all main effect features in the asymptotic limit for certain distributions of alleles[102]. In this scenario, epistatic features represent true epistasis (Figure 1.2) - independent of main effects from the involved genotypes. However, in many realistic settings, the distribution of samples dictates that epistatic features may not be perfectly orthogonal. This means that main effects may be detected as epistasis in these settings[35].

By fitting these two linear regressions and treating the first and second models as null and alternative, respectively, we can test for the presence of epistasis. A likelihood ratio test is constructed under the null hypothesis that the two models explain Y equally well. If the epistasis model explains Y better than the main effect model, epistasis is present and the likelihood ratio test will allow rejection of the null hypothesis. Alternatively a Wald test can be performed on epistasis parameters of the alternative model[72].

Challenges of Epistasis Detection

The sequencing technology enabling comprehensive eQTL studies has made the "large p , small n " problem relevant to genome-wide eQTL studies - but

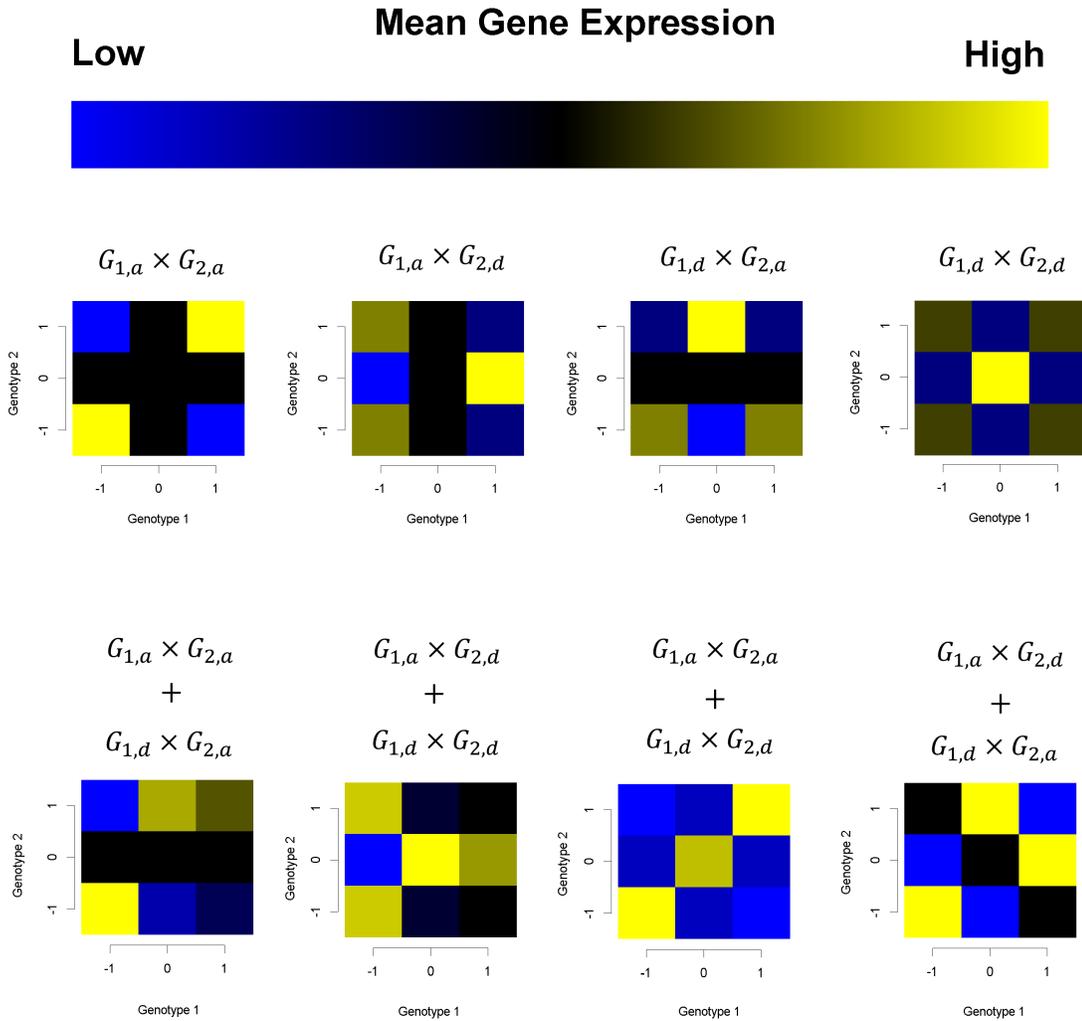


Figure 1.3: **Examples of Epistatic Effects** (Using the genotype codings defined just above, epistasis effects were generated and visualized using GP maps)

the same problem is exacerbated due to the combinatorial quality of epistasis (interaction) testing. Detecting interactions among variables is a well-known challenge in statistics and data mining[59]. If an eQTL dataset contains 1 million SNPs, it contains over 100 billion possible 2-way combinations of SNPS and over 100 quadrillion possible 3-way combinations. Such a combinatorial explosion presents challenges both computationally and statistically. Computa-

tionally, more tests means computation quickly becomes prohibitive as epistasis order (interaction order) is increased. Statistically, the problem comes down to reduced statistical power due to heavy multiple test corrections and epistasis requiring more heavily parameterized models. In order to balance false positive and false negative rates, multiple test correction must be employed; more tests overall means a heavier multiple test correction meaning that an effect must be detected more strongly in order to be identified. Since epistatic effects can take more forms than main effects, this necessitates either parameterizing a model more heavily or submitting a simple model to even more statistical tests[81]. The statistical challenges can be summarized in an analogy to needles (epistasis among loci) in a haystack (all possible interactions where epistasis could exist)[194]. Detection of epistasis encounters two primary difficulties: (1) Can one look for needles in the correct part of the haystack? (the "search" module of epistasis detection), and (2) If presented with a needle, can one distinguish it from hay? (the "identify" module of epistasis detection). Epistasis detection can therefore be thought about as a problem involving at least two modules, first for searching the genome broadly for epistasis and second for identifying the loci involved and the form of epistasis.

Approaches to Epistasis Detection

Many approaches have been developed to address the statistical challenges of epistasis detection. These can be labeled according to which module of epistasis detection they primarily address. Methods addressing the "Identify" module are designed to look at specific candidate associations, without consideration of the fact that the candidate is one in a sea of billions of potential associations[252].

These methods almost always utilize some sort of filtering in order to reduce the multiple test burden, but filtering is not their focus. An example is the additive-by-additive linear regression parameter that PLINK uses to test for a limited set of epistatic effects[22]. Methods classified as "Search/Identify" acknowledge the enormous size of the genetic feature space and are therefore designed to address both the "Search" and "Identify" modules in an inseparable manner using constructs like random forest[244] and multidimensional reduction[195]. "Search-Then-Identify" methods address both modules of epistasis detection but in a modular, separable way with an equal emphasis on the two modules. These methods usually filter the total possible number of tests down by either imposing knowledge derived from biology[82] or the data itself[149].

Another primary consideration is that most methods for detecting epistasis may not work for eQTL because they are designed to handle case/control traits and do not generalize to quantitative traits either in design or software implementation[240, 218].

Many methods have also been developed to address the substantial computational challenges of epistasis detection: FastEpistasis[203], FastANOVA[251], and EpiSNP[136] are just a few of the numerous examples.

Epistasis and Pleiotropy

Interactions among DNA, RNA, proteins, and other cellular constituents form the connections of highly complex regulatory networks. One feature of this complexity presents as functional and genetic redundancy, which are well-documented in biology [110]. For example, in *C. elegans* parallel and redun-

dant genetic pathways regulate dauer formation [179]. This regulation involves two groups of genes, regulated by the same pheromone, which positively regulate dauer formation through a functionally identical but physically separate pathway. In general, a set of genetic features (main and/or epistatic) that regulate expression of a particular gene can regulate at least one other gene - this is known as pleiotropy. More generally, pleiotropy involves mapping one component of the genome to multiple traits[172]. The existence of functional redundancy dictates that this regulation could be fully or partially redundant. In other words, a set of genetic features could regulate multiple genes in a similar manner, which would produce correlations among the genes and among the genetic models containing those genetic features. In its current conception, pleiotropy is categorized into two main forms: vertical and horizontal (Figure 1.4). Vertical pleiotropy describes when a locus has an effect on more than one phenotype in a causal chain. As such, vertical pleiotropy represents a combination of genetic and physiological effects. Horizontal pleiotropy describes when a locus affects more than one phenotype independently[223]. Among instances of horizontal pleiotropy is a specific type of genetic regulation - parsimonious pleiotropy[85] - which may present statistically when a locus has the same type of genetic effect on multiple phenotypes.

Pleiotropy and Epistasis do have a history of being detected independently: canonical correlation analysis (CCA) has been used to test the overall association between a genotype and multiple phenotypes[53] - and jointly: information from multiple phenotypes is used to constrain potential models of epistasis and to produce genetic networks that influence these quantitative traits [224].

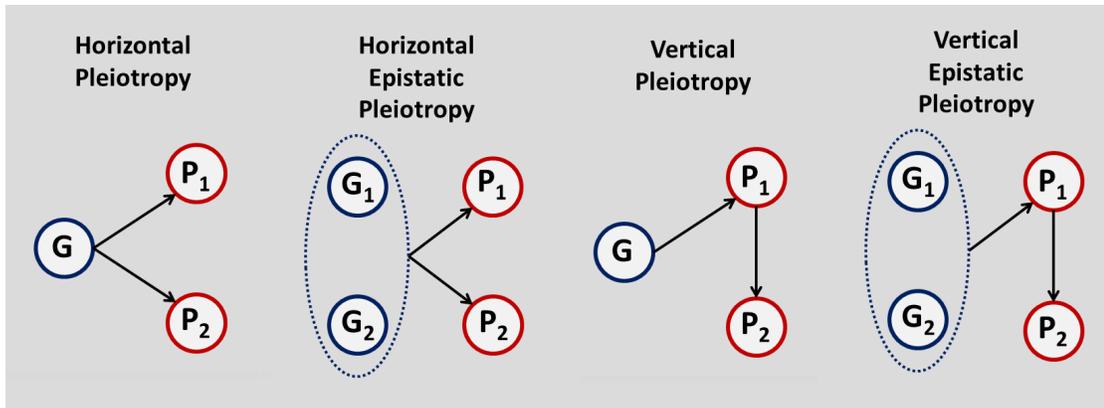


Figure 1.4: **Horizontal and Vertical Pleiotropy With and Without Epistasis.** In horizontal pleiotropy a genetic variant (G) directly regulates multiple genes (P1 & P2). In vertical pleiotropy G regulates P1 directly and is associated with P2 by way of P1. Multiple genetic variants can also exhibit pleiotropy, as shown in the bottom two illustrations. In horizontal epistatic pleiotropy multiple genetic variants (G1 & G2) directly regulate P1 and P2 in a non-additive manner. In vertical epistatic pleiotropy G1 and G2 regulate P1 directly and are associated in a non-additive manner with P2 by way of P1.

Replication

Despite success of eQTL studies and the prevalence of biological and statistical epistasis in complex traits[159], failure to reproduce early epistasis analysis results has placed renewed emphasis on replication in eQTL studies with[80] and without[109, 253] epistasis. Evidence is scarce on whether this lack of replication is due to spurious results or condition-specific genetic effects. Still, replication is a minimum standard for scientific results. Failure to replicate association study results hurts their generalizability and renders them less interpretable and less actionable[153]. Assessing replication can demonstrate that the identified effect is present in multiple independent datasets, reducing the chance that a dataset-specific artefact has produced a false positive. Biological differences

among datasets such as samples derived from different populations, different tissues, general batch effects across datasets, etc. serve to improve the potential for generalizability of any result produced that agrees across datasets.

There is also room for interpretation as to what is meant by "replication"[86]. One person could say that replication has occurred if an epistatic effect between a set of genotypes and a phenotype exists in multiple datasets. However, epistasis can take multiple forms - many of which could have different implications for an underlying mechanism. If epistatic effects of different forms were to show up in replicate datasets, this would indicate a higher chance of either a false positive or the datasets differing across some condition of importance. Therefore, it may be prudent to constrain the form of any epistatic effects to be consistent/concordant across conditions in order to demonstrate replication. Concordant epistasis exists when multiple quantitative traits contain similar/identical epistatic effects for a given set of genetic loci. In Figure 1.5, the marginal and two-way epistasis patterns are displayed for three different quantitative traits (may be different genes within the same dataset or the same gene across different datasets). Since marginal effects are absent but identical epistasis patterns are present, this represents epistasis that is concordant across conditions. These quantitative traits may represent expression measurements of different genes, or expression measurements of the same gene under different conditions (populations, tissues, etc).

Meta-analysis approaches can be performed to assess replication, but differing minor allele frequencies and other concerns between populations can make it impossible to replicate an effect that might have otherwise shown up in conditions of consistent minor allele frequencies across datasets. Hemani et al.[80]

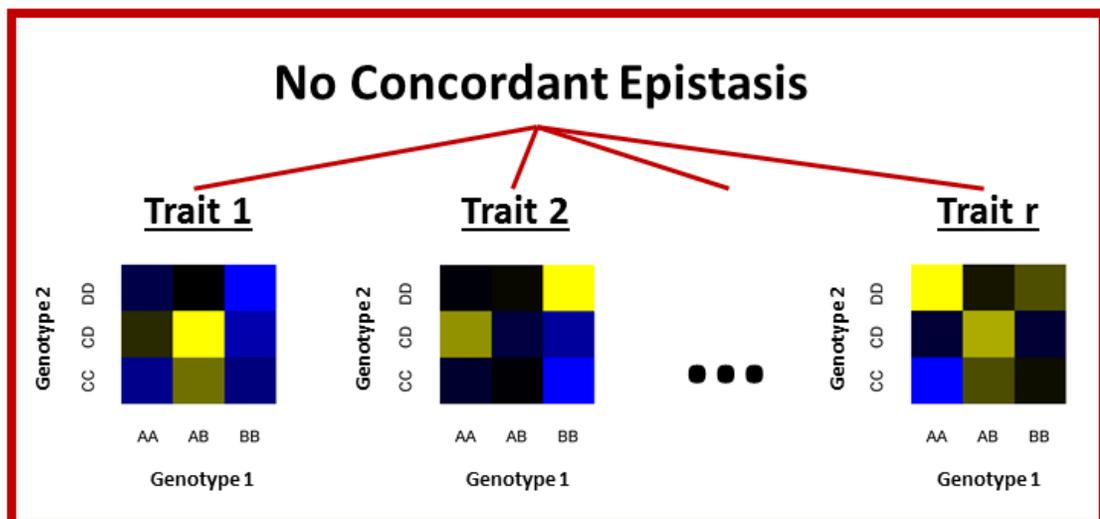
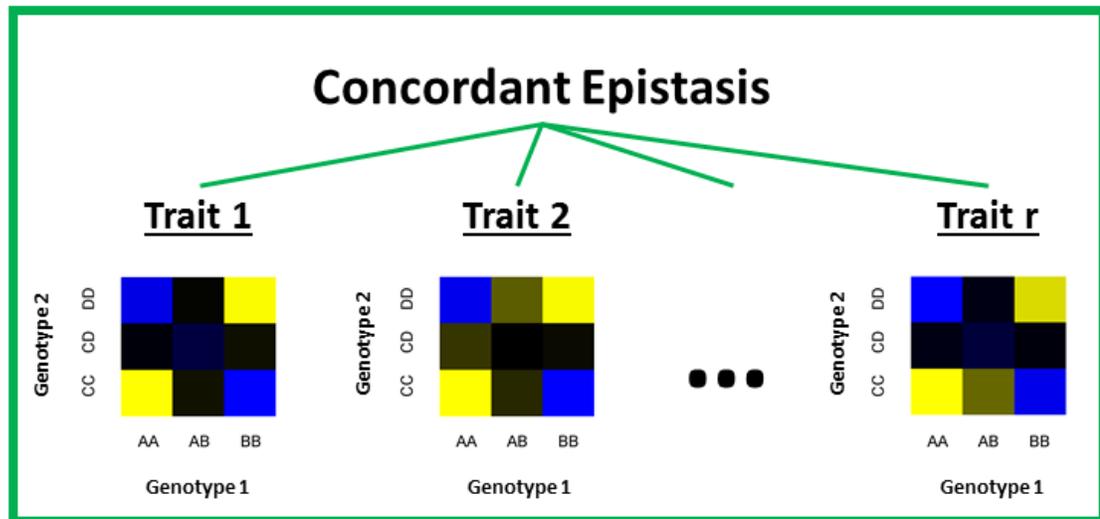


Figure 1.5: Concordant and non-concordant epistatic effects

found epistatic effects that replicated in 2 independent datasets. Their discovery power was limited to that of the original dataset since only SNPs identified as being globally significant in the original dataset were available to be validated in replication. Still, it is encouraging that in their study 22 of 30 interactions that were significant in all three datasets were concordant in effect.

Epistasis in Multiple Traits

The development of multivariate epistasis methods has been motivated by the need for methods to utilize/assess replication in eQTL epistasis studies as well as the plausibility of pleiotropy arising from the joint effect of two genetic loci. Multiple traits have also been incorporated into epistasis and eQTL mapping in order to find pleiotropy among genes [249, 224] or effects that replicate across datasets[113]. Considering multiple traits jointly has the potential to improve statistical power while providing useful information on the dependence structure of traits as it relates to genomic associations. Despite the numerous methods designed to detect 2-way and higher order epistasis[233] or genomic associations across multiple traits[63], few methods exist to detect epistasis across multiple traits. Results from testing for epistasis in individual traits can be meta-analyzed[58] or p-values can be combined across traits using Fisher's combined probability test or similar method[99]. Multivariate multiple regression also makes sense for this type of analysis because it can incorporate multiple traits into a model parameterized by multiple epistatic effects[161]. CAPE aims to detect epistasis by integrating information from multiple phenotypes that are influenced concordantly by genomic interactions[224]. MFRG tests for interaction between two genes jointly in multiple quantitative traits[249]. Zhang developed a Bayesian method for identifying epistasis and pleiotropy by treating co-expressed genes as a module[250]. Table 1 provides a list of methods for detecting epistasis in single and multiple traits.

1.3 Epigenome-Wide Association Studies

1.3.1 DNA Methylation

Advances in DNA sequencing have also transformed the study of epigenetics by enabling the measurement of features like chromatin interactions[121], histone modifications[173], and DNA methylation[28] using next generation sequencing technologies. Epigenetics is a broad term, encompassing many mitotically heritable elements which serve to modulate genetic function, and has therefore also been proposed to contribute to the missing heritability viewed in common human diseases[209, 220]. Epigenetics is known to play a significant role in development, cancer, cell differentiation, and development of cancer through aberrant cell differentiation[106]. Epigenetic alterations commonly alter chromatin accessibility, which in turn alters interactions between the DNA and regulatory elements[216]. One form of epigenetic alteration is DNA methylation (DNAm), which refers to methylation of the 5' position of a cytosine nucleotide[52]. DNAm is a key molecular mechanism in embryonic development, transcription, chromatin structure, X chromosome inactivation, genomic imprinting and chromosome stability[128, 196, 191, 45, 210]. In humans, the majority of DNAm in humans happens to CpG dinucleotide pairs[256]. CpGs have a statistically underrepresented[237, 60] and patterned distribution throughout the mammalian genome - an early indicator of the insights they would produce[208]. Epigenetic features involving CpGs include "CpG islands" (CGIs), which are regions of CpG site enrichment near promoters that generally remain unmethylated across conditions[36]. Approximately 60% of human genes are associated with CGIs[105]. Despite the many elements

present in the epigenome - known and unknown - numerous studies claim to have performed epigenome-wide association studies (EWAS) by looking at DNA methylation[188], presumably judging the methylome a sufficient proxy for the entire epigenome.

1.3.2 Role in Disease

Methylation variable position (MVP) is the methylome equivalent of a SNP in the genome, and displays differential methylation[189]. Differentially methylated regions (DMRs) consist of nearby adjacent CpG sites that are usually less than 1kb in length and have variable methylation between samples[165]. Both MVPs and DMRs represent epigenetic variation that could be associated with a condition like disease state, cell type, etc. via upstream and/or downstream regulation.

CpG islands have been observed to undergo aberrant hypermethylation during extended proliferation in vitro[150], suggesting that methylation may have a role to play in cancer treatment[11]. DNAm also plays a role in silencing repetitive DNA elements[51], and the loss of methylation of repetitive elements is an epigenetic hallmark of cancer[130]. Genome-wide loss of DNAm is also an early and frequent occurrence in cancer, and is associated with severity of cancer in many different tumour types[235]. Tissue- and cancer-specific DMRs appear more often in regions up to 2kb away from CGIs called "CpG island shores", and have been strongly related to both gene expression and disease[101], suggesting more complex patterns of functional CpG methylation distribution throughout the genome.

The epigenome appears to be highly tissue-specific, which has placed an emphasis on analysis accounting for cell-type heterogeneity[147, 61] and development of cell-type specific assays[91]. This tissue-specificity may be an advantage for clinical application of DNA methylation data. Biomarkers for disease or disease risk can be detected in epigenetic screens of cells from readily available tissue such as blood. Unlike the genome, the methylome can be greatly impacted by disease. Disease-distinguishing epigenetic elements (that may also be tissue-specific) can thus indicate the presence or absence of particular diseases (biomarker)[133].

1.3.3 Performing EWAS

An epigenome-wide association study (EWAS) is the methylome equivalent of the GWAS. In EWAS, millions of CpG methylation sites or thousands of DMRs throughout the methylome are tested for association with a trait such as disease or adiposity[230]. Like GWAS, EWAS are a means to comprehensively search for associations between genomic state and a trait. However, EWAS differ from GWAS in two primary ways. First, unlike the genetic variants, DNAm is extremely sensitive to cell type and therefore tissue type, so potential cell-type heterogeneity must be accounted for in the assay and/or analysis[257]. Second, epigenetic changes may contribute to a trait and/or occur downstream of the trait[182]. While this causal ambiguity may contribute positively to the utility of DNAm as a diagnostic or prognostic biomarker, it makes inferring regulatory causation more difficult compared to GWAS. Still, such inferences can be made with careful study designs or by leveraging other omics data. An EWAS for type I diabetes measured DNAm before and after diagnosis in order to ex-

plore disease aetiology[187]. More recently, results from EWAS and methylation quantitative trait loci (where SNPs affect DNAm) analyses were integrated to establish multiple lines of evidence that increased BMI was most likely causing increased DNAm levels rather than resulting from it[39].

Differential expression and differential methylation analyses have been integrated to identify instances of reduced methylation at gene promoters and increased gene expression, or vice versa[98]. A joint methylation and gene expression analysis pipeline that executes the same general procedure is show in Figure A.1.

1.4 Overview of Dissertation

My work has involved developing computational and statistical techniques for probing large groups of genetic variants, gene expression profiles, DNA methylation profiles, and clinical data for evidence of interactions within or across these omics readouts. I developed a novel asymptotically correct statistic (Fomac) for detecting low and high order statistical epistasis (interactions among genomic regions) in quantitative trait locus studies. Then, I probed genome-wide DNA methylation to find associations between methylated genomic regions, the whole methylome, and the disease Charcot foot.

1.4.1 Chapter 2 - *F*omac (*F*-test of Magnitude and Concordance)

Chapter 2 focuses on *F*omac, which is a novel statistic flexible enough to test for concordant epistasis involving an arbitrary number of traits and genetic loci. *F*omac is designed to detect instances where the epistatic parameters across several traits are both significant and in agreement, like in the top of Figure 1.5. Justifications for pursuing concordant epistasis, theoretical underpinnings of *F*omac, and application to simulated and real biological data will be covered.

1.4.2 Chapter 3 - Connections Between CpG Methylation and Charcot Foot

Chapter 3 covers an EWAS study that was performed using isolated monocytes from 54 patients with type II diabetes and varying levels of the disease Charcot foot to discover DNAm correlates of the disease. Differential methylation between disease and control was assessed in three ways, each demonstrating that the methylome was significantly different between patients with and without Charcot foot. Pathway analysis on differentially methylated genes unveiled a possible role of circulating monocytes in the pathogenesis of Charcot foot.

CHAPTER 2

A NOVEL ASYMPTOTICALLY CORRECT STATISTIC FOR DETECTING PAIRWISE AND HIGHER ORDER CONCORDANT EPISTASIS ACROSS MULTIPLE QUANTITATIVE TRAITS

Epistasis describes the interaction among multiple genetic loci in their association with a trait such as gene expression (eQTL epistasis), and is emerging as a potential solution to the problem of missing heritability and as a tool for uncovering the structure of genetic pathways. Despite the abundance of methods designed to address the computational burden and reduced statistical power of genome-wide testing for epistasis, detecting epistasis remains difficult and replication of epistatic effects remains rare. The need for flexible meta-analyses also grows with the demands of big data only set to expand in the near future. We propose a novel method Fomac: F-test of magnitude and concordance which jointly tests multivariate linear regression epistasis parameters for significance and consistency to identify concordant epistatic effects across traits. By specifically constraining the form of epistatic effects to be consistent across traits, we gain increased statistical power to identify these effects in high noise regimes. Fomac can be applied to sets of coexpressed genes within one sample (dataset) to discover pleiotropy, or applied across multiple samples with similar characteristics to find examples where epistatic effects are consistent across conditions (replication). We demonstrate in simulations capturing the effects of different sample sizes, epistasis orders, and number of traits on relative method performance that Fomac outperforms several comparable methods for detecting epistasis. The Fomac framework was applied across 3 genome-wide genotype and gene expression datasets of 618 individuals from the TwinsUK registry, and demonstrated a well-calibrated test statistic and several examples of tissue-

concordant epistasis. The fomac test-statistic is also compatible with many of the top frameworks already available, meaning it can act as a standalone or complementary tool for detecting concordant eQTL epistasis.

2.1 Introduction

Expression quantitative trait loci (eQTL) are instances where a genetic locus is associated with gene expression. While eQTL exhibiting main effects have been central to our understanding of the regulatory architecture of complex traits[190], it is clear that focusing only on main effects is likely an oversimplification of the underlying biology since most phenotypes are affected by multiple genes[43] in ways that are not strictly independent[56]. Indeed, a large portion of heritability in gene expression remains unexplained by these main effects[140]. Epistasis can be viewed on a population level as any statistical deviation from the sum of strictly additive genetic effects of two loci in their effects on a trait[55]. Epistasis has been proposed as a solution to the problem of missing heritability[140, 138, 47], as well as a tool for elucidating the structure of genetic pathways[111]. Epistasis is pervasive on both the organismal[204] and population[158, 223] levels. Still, the combinatorial explosion of considering genotypes combinations has caused greatly increased computational demands and reduced statistical power in epistasis studies[97]. The ever-growing nature of genomic data suggests that sample sizes will grow in the coming years - the challenge will be designing methods for leveraging the growing sample sizes of these data while remaining robust to high dimensional and heterogeneous data[93].

Many approaches have been applied to detect epistasis in gene expression

and complex traits, varying from parametric methods like linear regression[181] and bayesian inference[250, 241] to nonparametric methods like decision trees[244] and other method types[149]. Parametric methods have the advantage of being able to explicitly test for specific epistasis patterns while nonparametric methods have the flexibility to capture a wide range of effects[92]. The linear regression framework serves as the basis for a great many methods which have adapted the framework with different genotype codings and choices of parameters to include in the model[184, 54, 242]. Genotype coding refers to the choice made when transforming a genotype variable from categorical to ordinal - several different codings are represented in the literature, each capturing slightly different epistatic effects[161, 243, 164]. There is also disagreement about whether to use a single epistasis parameter or more heavily parameterized full epistasis model[31].

With the renewed emphasis on replication of eQTL results[100], studies have also found that epistatic effects can replicate[232] and be concordant[80] across multiple datasets. Assessing replication can demonstrate that the identified effect is present in multiple independent datasets, reducing the chance that a dataset-specific artefact has produced a false positive. Meta-analysis approaches can be performed, but differing minor allele frequencies and other concerns between populations can make it impossible to replicate an effect that might have otherwise shown up in conditions of consistent minor allele frequencies across datasets. Hemani et al.[80] found epistatic effects that replicated in 2 independent datasets. Further constraining the search to concordant effects has the potential to increase statistical power and identifies epistatic effects that are both replicable and robust to setting, which improves the chances that effects are generalizable and interpretable.

In this study we present Fomac, a novel statistic that detects concordant epistasis by leveraging the idea that when multiple traits are impacted by similar epistatic effects, the estimated parameters of models capturing these effects will be similar and will cluster together in parameter space. This asymptotically correct statistic makes use of the known null distribution of linear regression parameter estimates in order to calculate an f-statistic jointly evaluating the degree to which (1) the epistasis models are different than the null model (i.e. epistatic effects exist), and (2) the epistasis models are identical to each other (i.e. the epistatic effects are concordant). By requiring effects to be concordant across traits, we are able to detect concordant epistasis in higher noise conditions than comparable single- and multi-trait methods. The FOMAC test statistic can be applied to the same trait across different datasets to leverage replicating effects for epistasis discovery, or applied across different traits like coexpressed genes within a single dataset to discover pleiotropy.

Simulations were carried out to compare the performance of Fomac to that of comparable methods for detecting single- and multi-trait epistasis. We show that Fomac identifies more examples of concordant epistasis than the other methods in high-noise regimes and under 6 different simulation parameter settings. Fomac was also applied to gene expression from the Multiple Tissue Human Expression Resource (MuTHER)[168] where a genome-wide analysis across 3 tissues identified 2754 examples of gene-wise Bonferroni-significant concordant epistasis.

2.2 Methods

In its most general form the Fomac framework (Figure 2.1) accepts as input r datasets, each containing 1 trait and g genotype variables, and tests for g -way epistasis that is concordant across all datasets. However, in this general form notation is unwieldy, computation takes longer, and variations in allelic distributions between different populations reduce statistical power to detect genetic effects[157]. Therefore, it makes sense to match samples if possible. If samples are matched across all r traits - as they might be in analyses of time-series data, cross-tissue data, or pleiotropy analyses within a single dataset - parameter estimates correspond to those produced by multivariate regression, which involves more streamlined notation and computation compared to the case of unmatched samples. For both notational simplicity, computational efficiency, and to maximize statistical power, Fomac will be discussed here and applied to simulated and real data in settings where samples are matched across traits (fig 2). However, Fomac has the theoretical flexibility to handle multiple traits of different sample sizes and/or entirely different samples.

2.2.1 The *F*omac Framework

Fomac starts with r traits and g genotype variables. The multivariate multiple regression model has the form

$$Y = X\beta + E \tag{2.1}$$

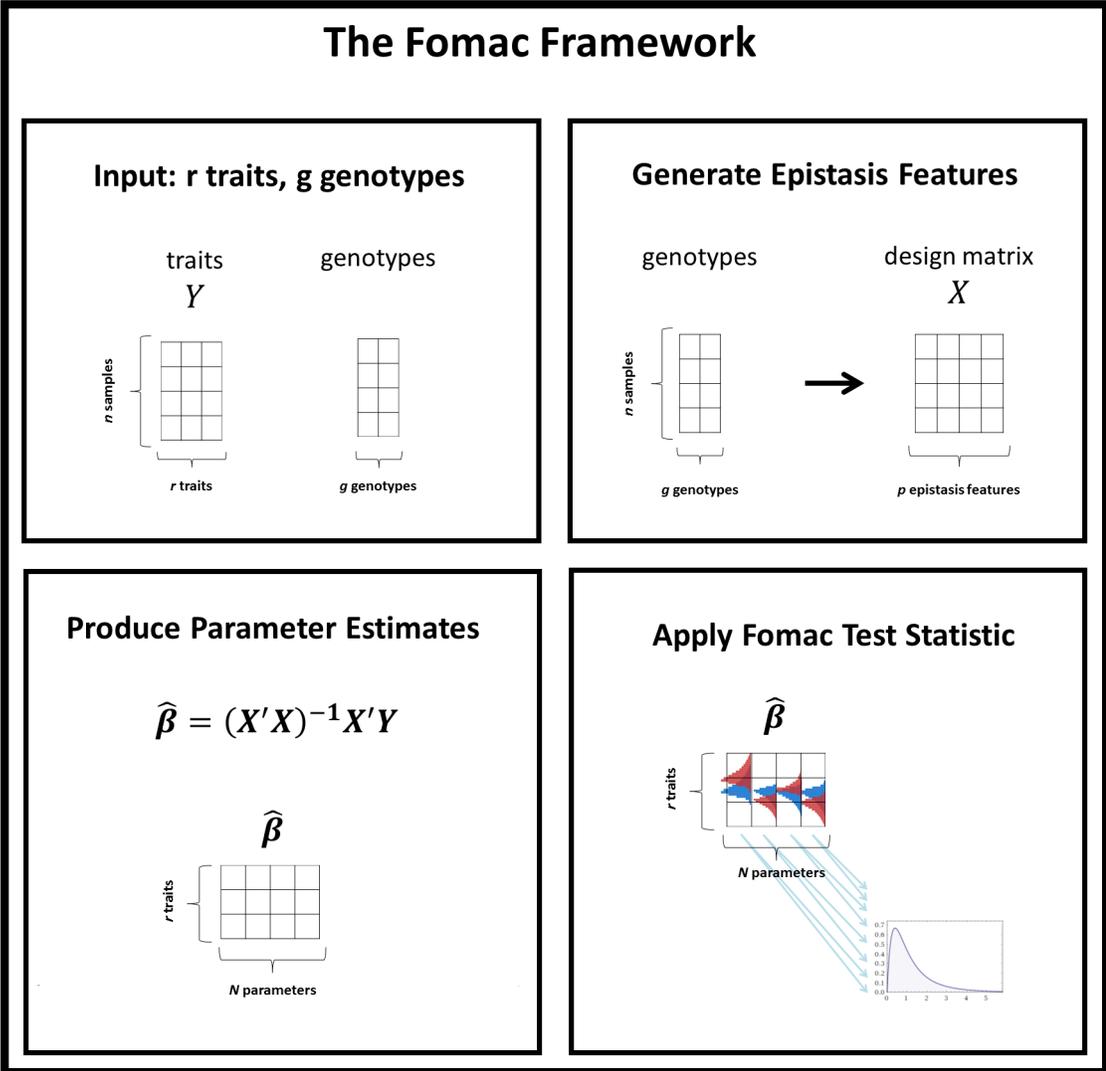


Figure 2.1: **The Fomac Framework.** Top Left: Fomac accepts a set of r traits and g genotypes. Top Right: Fomac can leverage highly parameterized models, so genotype variables can be readily transformed into several epistasis features. Bottom Left: Multivariate linear regression produces parameter estimates and estimates for distributions of these parameter estimates. Bottom Right: Fomac operates on these parameters to assess how concordant and significant epistatic effects are across traits.

where Y and is a known $n \times r$ matrix of traits, X is a known $n \times p$ matrix of p epistasis features, β is an unknown $p \times r$ matrix of p epistasis parameters across r traits, and E is a $n \times r$ random matrix with rows independently distributed as $N(0, \Sigma)$. In the case of epistasis detection X will contain features capturing interactions among the g genotype variables. Including intercept and main effect terms does not disrupt this procedure if the terms are orthogonal to epistasis features.

The maximum likelihood estimate of β is

$$\hat{\beta} = (X'X)^{-1}XY \quad (2.2)$$

where $\hat{\beta}$ is a $p \times r$ matrix of parameter estimates that follows a multivariate normal distribution[162]

$$\hat{\beta} \sim N(\beta, (X'X)^{-1} \otimes \Sigma) \quad (2.3)$$

and the maximum likelihood estimate of Σ is

$$\hat{\Sigma} = \frac{1}{n}(Y - X\hat{\beta})'(Y - \hat{\beta}) \quad (2.4)$$

which is used to estimate the covariance of $\hat{\beta}$ as

$$\hat{\Sigma}_{\hat{\beta}} = \frac{1}{n}(X'X)^{-1} \otimes ((Y - X\hat{\beta})'(Y - X\hat{\beta})) \quad (2.5)$$

Since we now have estimates for the parameters as well as estimates for both the mean and variance of each parameter under the null distribution, we can apply the Fomac test-statistic.

2.2.2 The *F*omac Test Statistic

The *F*-test of Magnitude and Concordance (Fomac) statistic is the ratio of two scaled chi-squared distributions. In its current form Fomac doesn't require the covariance structure of $\hat{\beta}$, and it is sufficient to know individual parameter variances. Define A as a vector of length p

$$A = \text{diag}(\hat{\Sigma}_{\hat{\beta}}) \quad (2.6)$$

Define $\bar{\hat{\beta}}_j$ as the mean of the j th parameter across all traits

$$\bar{\hat{\beta}}_j = \frac{1}{r} \sum_{i=1}^r \hat{\beta}_{i,j}^2 \quad (2.7)$$

Magnitude (numerator): The numerator measures how far the models' parameters are from the origin (where the null model is centered). The mean euclidean distance of Beta from the origin (over r replicates) is scaled by a factor accounting for two things: (1) the quantity involves the mean of models instead of just a single model, so must be scaled by the inverse square root of r , and (2) variance of the parameters needs to be unitized, so each parameter must be scaled by its estimated standard deviation. This quantity is chi-squared distributed with p degrees of freedom.

$$\sum_{j=1}^p \frac{1}{\sqrt{rA_j}} \bar{\hat{\beta}}_j \sim \chi_p^2 \quad (2.8)$$

Variance (denominator): The denominator measures how concordant the model parameters are by using their sample variance. For each parameter, vari-

ance is calculated across r replicates and scaled by the expected variance. Taking the sum of these scaled variances across all p parameters yields a quantity that is chi-squared distributed with $p(r - 1)$ degrees of freedom.

$$\sum_{j=1}^p \frac{1}{A_j} \sum_{i=1}^r \left(\hat{\beta}_{i,j} - \bar{\beta}_j \right)^2 \sim \chi_{p(r-1)}^2 \quad (2.9)$$

F-statistic: The numerator and denominator described above are chi-squared distributed. A ratio of chi-squared distributions, scaled by their respective degrees of freedom, has an f -distribution with degrees of freedom the same as those of the numerator and denominator, respectively. Therefore, taking the ratio of the previously established numerator and denominator, scaled by their degrees of freedom, results in an f -statistic that simplifies to

$$\frac{(r-1)}{\sqrt{r}} \sum_{j=1}^p \left[\frac{\sqrt{A_j} \bar{\beta}_j}{\sum_{i=1}^r \left(\hat{\beta}_{i,j} - \bar{\beta}_j \right)} \right] \sim F_{p,p(r-1)} \quad (2.10)$$

for which there is likely a more elegant linear algebra representation. Comparing the f -statistic against the upper tail of the f -distribution produces a p -value.

2.2.3 Correcting for Correlated Traits

We have determined via simulations that when traits are correlated, the test statistic deviates from the expected f -distribution by a multiplicative factor, such that

$$F_{expected} = \frac{F_{deviated}}{\text{median}(F_{deviated})}$$

Therefore, a heuristic has been built into the Fomac framework which automatically calculates this factor and corrects the test statistic. During the procedure, the test statistic is calculated 10000 times using the traits of interest and randomly chosen genotype sets of size g while being sure not to use any genotype sets that will subsequently be tested for concordant epistasis. Then, the median of this test statistic is calculated and used to scale all subsequent tests involving these traits.

2.2.4 Application to Epistasis/Generation of Epistasis Features

In order to specifically apply the FOMAC framework to epistasis detection, a design matrix capturing genotype interactions must be generated (Figure 2.1-top right). First, 2 ordinal codings must be defined in order to capture the 2 degrees of freedom captured by 3-category genotypes. Most commonly additive and dominant codings are applied to capture the additive effects of alleles and the independent effect of heterozygotes, respectively(citation). These codings can be constructed to produce epistatic features that are orthogonal to the main effects from which they are constructed, given certain distributions of samples among genotype classes. The codings used here are independent of main effects when genotype classes are balanced. The additive coding of the g th genotype is defined as

$$G_{g,a} = \begin{cases} -1 & \text{if } G_g = AA \\ 0 & \text{if } G_g = Aa \\ 1 & \text{if } G_g = aa \end{cases}$$

and the dominant coding of the g th genotype is defined as

$$G_{g,d} = \begin{cases} -0.5 & \text{if } G_g = AA \\ 1 & \text{if } G_g = Aa \\ -0.5 & \text{if } G_g = aa \end{cases}$$

To create a design matrix capturing all epistatic effects, all possible interactions between the involved genotypes are considered. In the case of two-way epistasis, this generates 4 epistatic features for the design matrix. An epistasis linear regression model has the form

$$Y = \beta_{a,d}G_{1,a}G_{2,d} + \beta_{d,a}G_{1,d}G_{2,a} + \beta_{d,d}G_{1,d}G_{2,d} + \epsilon$$

where β represents the strength of a genetic effect and ϵ is error.

The full epistasis linear regression model, which includes main effects (terms 2-5) and all orders of epistasis up to the maximum order (terms 6-9) has the form

$$Y = \beta_0 + \beta_{1,a}G_{1,a} + \beta_{2,a}G_{2,a} + \beta_{1,d}G_{1,d} + \beta_{2,d}G_{2,d} + \beta_{a,a}G_{1,a}G_{2,a} \\ + \beta_{a,d}G_{1,a}G_{2,d} + \beta_{d,a}G_{1,d}G_{2,a} + \beta_{d,d}G_{1,d}G_{2,d} + \epsilon$$

It is important to note that estimations of these epistasis parameters depend on the choice of coding and parameterization for this model. However, all fully parameterized (pure) epistasis models of this form will capture the same range of epistatic effects. In comparison, all full epistasis models which include main effects may not be equivalent. PLINK uses an epistasis model that includes additive main effects but is not fully parameterized since only the additive main effects and additive-by-additive epistasis terms are included in the regression.

2.2.5 Comparing Performance

Performance of Fomac was compared to other methods designed to detect 2-way epistasis in eQTL datasets: PLINK[184], CAPE[224], a random forest-based interaction-scoring method, multivariate and univariate versions of the linear regression model used by Fomac, and the combined p-values from the univariate regression[132]. Scores for the univariate method were p-values calculated by performing the Wald test[72] on all epistasis parameters from the same epistasis model used by Fomac. Scores for the multivariate method were p-values calculated from the Wilks lambda test statistic generated by fitting the same full epistasis model jointly to all traits. Since PLINK and the univariate linear regression model produced a p-value for each trait, only scores from the first trait were used in order to represent the performance of methods that use a single dataset for epistasis discovery. Additionally, p-values from PLINK and the univariate linear regression model were combined using Fishers combined probability test in order to represent the performance of methods that analyze multiple traits separately but afterwards consider results jointly across traits. The tree-based interaction-scoring method was a simple adaptation of SNPInterForest[244] implemented in the R statistical programming language. Briefly, the idea is to regress a trait on all genotypes using random forest, and look for instances where a set of genotypes co-occurs on the same branch more often than would be expected by random chance. Fomac was also tested in the 3- and 4-way epistasis settings, but PLINK and CAPE had to be excluded from the comparison since they do not accommodate greater than 2-way epistasis. Some methods were excluded from all comparisons because they returned scores for a group of genotypes or scores for each phenotype, thus rendering their output incompatible with the performance metrics used in this comparison[249].

2.2.6 Simulations

We simulated 50 datasets (sample size = 900), each containing 15 genotypes and 100 sets of 4 traits, which had been generated from an epistatic effect between a random pair of genotypes

$$Y_r = \beta_{a_1a_2}G_{a_1}G_{a_2} + \beta_{a_1d_2}G_{a_1}G_{d_2} + \beta_{d_1a_2}G_{d_1}G_{a_2} + \beta_{d_1d_2}G_{d_1}G_{d_2} + \epsilon_r \quad (2.11)$$

Where Y_r is the r th trait, $G_{i,d}$ is the additive coding of the i th randomly selected genotype coded -1 for heterozygous minor, 0 for homozygous, and 1 for heterozygous major, $G_{i,d}$ is the dominant coding of the i th randomly selected genotype in the set of interacting genotypes coded -0.5 for heterozygous minor, 1 for homozygous, and -0.5 for heterozygous major, and epsilon is a multivariate Normal error term with diagonal covariance matrix. Genotypes were simulated with evenly distributed allele-classes (resulting in minor allele frequency of 0.5), and no two genotypes had a correlation greater than 0.2.

For each dataset, there were 10500 possible cross-trait epistasis tests, 100 of which were condition positive. For every method, scores were used to produce a receiver operating characteristic (ROC) curve, and ROC curves from all 50 datasets were averaged to produce a single smoothed ROC curve for each method. Area under curve (AUC) was also calculated for all datasets, and the distribution of AUCs was used to produce a box plot for each method. Performance was assessed in this manner for 6 sets of parameters, spanning all combinations of number of traits (r) = 4/8 traits and epistasis order (g) = 2/3/4 genotypes (Table 2.1). In order to eliminate the increase in genotype class sparsity that would otherwise occur when increasing epistasis order (2 genotypes =

9 classes but 3 genotypes = 27 classes), sample sizes were scaled such that the ratio of sample size to number of possible genotype classes was 100 samples/class in all simulations.

Table 2.1: Simulation Parameter Configurations

Number of Traits (r)	Epistasis Order (g)	Sample Size (n)
4	2	900
8	2	900
4	3	2700
8	3	2700
4	4	8100
8	4	8100

2.2.7 Application to Human Data

We performed an analysis for concordant cross-tissue eQTL epistasis on adipose, lymphoblastoid cell line (LCL), and skin datasets obtained through the Multiple Tissue Human Expression Resource (MuTHER) project[75]. We kept all 618 samples that had genotypes and expression in all three tissues and genotypes (Figure 2.2A), and all 28827 genes that were measured and expressed in all three tissues (Figure 2.2B). Traits were adjusted for covariates by regressing each trait on age and appropriate tissue-specific covariates and keeping the residuals. Mean centering and variance standardizing had little effect on the resulting p-values (design matrix contains intercept term, which takes care of any nonzero mean), so these steps were skipped. In order to help keep genotype classes bal-

anced, we filtered out genotypes that had a minor class frequency (MCF) lower than 0.25 (fewer than 25% of the total samples in each of the three genotype classes). While this does not guarantee balanced classes, it limits the extent to which classes would be unbalanced. Genotypes were taken from the TwinsUK consortium, and the 18416 genotypes not requiring imputation and with MCF > 0.25 were kept.

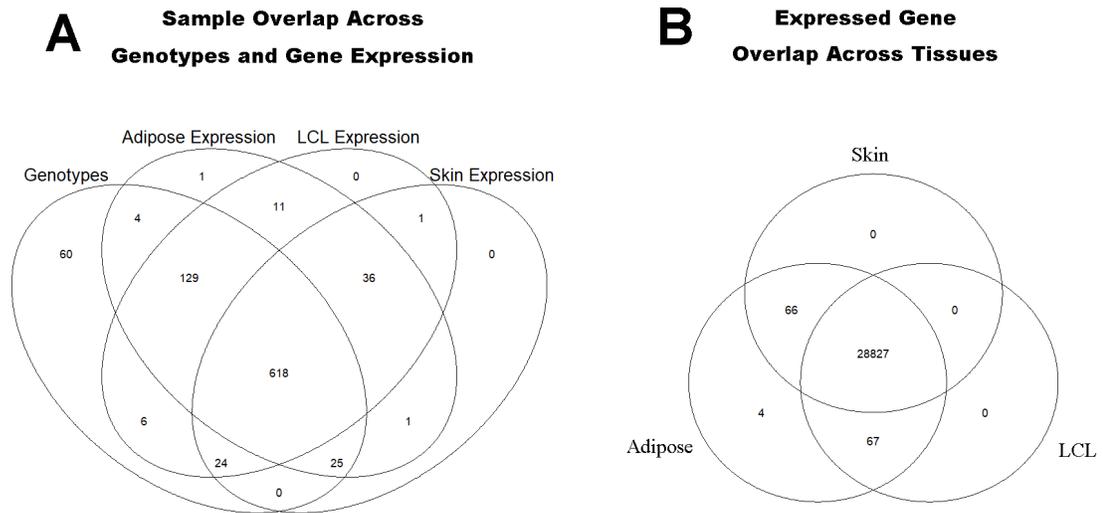


Figure 2.2: Venn Diagram of Sample/Gene Overlap Across Datasets

For each of 28827 sets of 3 traits (single gene measured 3 times in different tissues), main effects from all genotypes were calculated, and a set of 100 independent (Pearson correlation < 0.2) genotypes was assembled from the top main effect scores pooled from all tissues. The set of traits was then tested for concordant epistasis with all pairs of the set of independent genotypes. In very rare cases ($3.6 \times 10^{-3}\%$ of the time) $X'X$ was not invertible and a regression could not be performed, so not all genotype pairs were used for every gene. A total of 142,688,467 total tests were performed. Analogous tests for epistasis were also carried out with the only alteration occurring in the parameter estimation pro-

cess where the design matrix included the same main effect features that were used to detect main effects in the previous step.

2.3 Results

2.3.1 Simulations demonstrate sensitivity to inter-trait correlations and unbalanced designs

Simulations were performed to examine the test statistic's behavior under deviation from optimal conditions. We determined that (1) correlations among traits in standard sample space cause correlations among traits in parameter space as well, which causes the statistic to deviate from its expected f -distribution. This likely occurs because the statistic assumes that samples in parameter space are derived from independent multivariate Normal distributions, and this assumption is violated with correlated traits. We also determined that (2) if samples are distributed unevenly among the 9, 27, 81, etc. genotype classes for 2-, 3-, and 4-way epistasis, respectively, correlations appear among parameters in parameter space, which also causes the statistic to deviate from its expected f -distribution. This likely occurs because each full set of epistasis features is only orthogonal for a given distribution of samples among the genotype classes - the most commonly used sets of epistasis features usually exhibit orthogonality or near-orthogonality in settings where samples are nearly evenly distributed among the genotype classes (citations spanning all codings that have been used). In order to avoid skewing the test statistic, Fomac was applied in situations where samples were sufficiently evenly distributed among genotype classes to avoid

skewing the test statistic.

2.3.2 Performance Comparison

In our analysis on simulated data, we tested the ability of Fomac and comparable methods to detect epistasis when given multiple traits generated from concordant epistatic effects. We produced ROC curves averaged over 50 datasets (Figure 2.3 and calculated area under each curve to produce a distribution of ROC AUCs for each method (fig 4b, fig S1b). As assessed by mean AUC, methods ranked the same across all parameter configurations, with performance ordered (1) Fomac > (2) multivariate LM ~ combined univariate LM > (3) combined PLINK > (4) univariate LM > (5) PLINK > (6) random forest method > (7) CAPE. It is not entirely clear why CAPE is the only of the multivariate methods (Fomac, multivariate LM, CAPE) that performed worse than both univariate methods (PLINK, univariate LM). CAPE transforms traits into eigentraits and keeps only some top number of eigentraits, meaning that some information from the set of traits may be lost and will therefore be unavailable to the epistasis detection task - but the 3 eigentraits should still capture more information than just a single one of the original traits, especially since traits in the simulations were not highly correlated. The random forest method is expected to thrive in conditions with thousands of genotypes, when exhaustively testing all combinations becomes restrictive. In this setting, exhaustive testing is not a statistical or computational problem but the noise level is, so random forest has no advantage over the other methods. PLINK also suffers because it only tests for one of the four epistatic effects, and is therefore missing out on the other three effects that may exist. While containing the optimal parameterization for

detecting the effects that are present, the univariate LM is only using one of the traits (as most current methods do). Combined PLINK and combined univariate LM both do substantially better than when only one dataset is used to produce a score. This result also operates as a sanity check to make sure that at least one other method is able to leverage the concordant epistasis in these datasets.

2.3.3 Human Data

Quality Control

We used the Fomac framework to analyze gene expression from MuTHER[168] and genotypes from TwinsUK[155] datasets. Fomac was able to handle inter-correlated non-Gaussian traits, the presence of main effects, and tests that were strictly independent due to testing all possible combinations of a relatively small group of genotypes. While restricting genotypes to those with relatively evenly distributed classes is not ideal in practice, this step is necessary to demonstrate that the Fomac framework can handle all other characteristics of real biological data and return true p-values (with demonstrated uniform null distribution), as shown by global QQ plots (Figure 2.4). Since we expect concordant epistasis to produce correlations among traits, it is essential to be able to test correlated traits in a controlled manner. The heuristic for this correction is not well understood theoretically but works well in practice, in both simulated and real data scenarios. Since main effects were used to identify the group of candidate epistatic genotypes, it was essential that the epistasis features be orthogonal to main effects or else the resulting QQ plots would be inflated (lower p-values than expected under the null hypothesis) and we would risk identify-

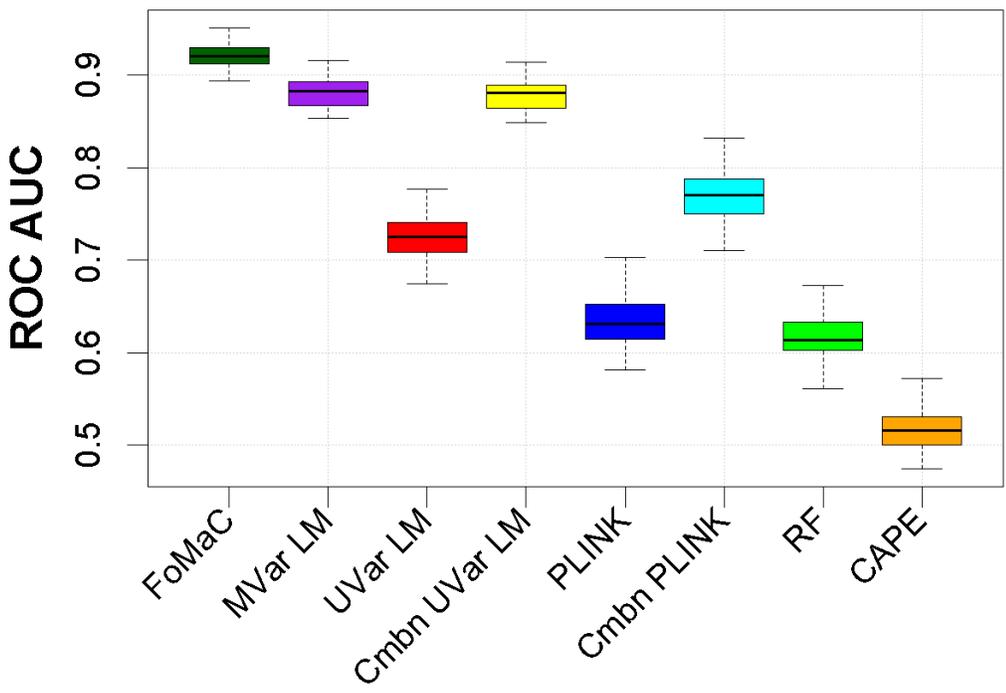
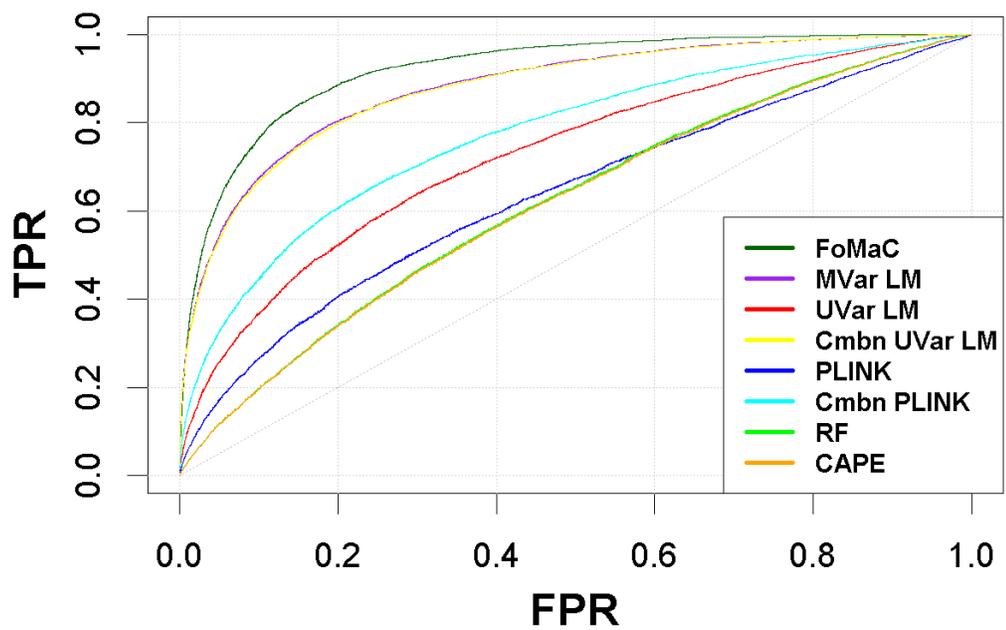


Figure 2.3: **Simulation ROC Performance.** Parameters: $r = 4, g = 2, n = 900$

ing spurious associations. Due to the choice of genotype coding, all 4 epistasis features in the model were theoretically orthogonal to the main effect codings in the case of evenly balanced genotypes classes, and nearly orthogonal in the case of nearly balanced genotype classes. This characteristic assured that main effects between genotypes and traits would not be detected by our epistasis tests, so that any identified effects were purely epistatic. The group size of candidate epistatic genotypes for a given gene (chosen either by main effects or proximity to gene) also impacted the results of this study. Group sizes less than 50 caused a noticeable dependence among tests for epistasis, which is why the group size of 100 was chosen. A big challenge was to balance using enough genotypes in a group while minimizing the total number of tests performed so as to maximize power to detect small epistatic effects among genetic loci exhibiting main effects.

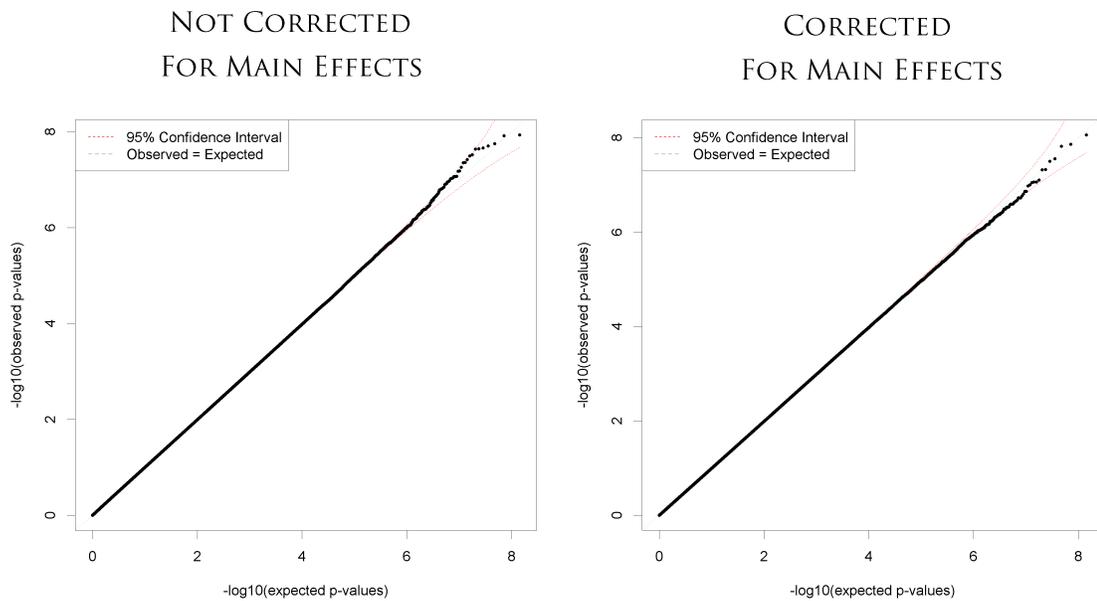


Figure 2.4: Global QQ Plots for uncorrected (left) and main effect-corrected (right) p-values.

Significant Tests

Of the 142,693,650 total tests performed genome-wide, none were globally significant at the Bonferroni or Benjamini-Hochberg level at $\alpha = 0.05$, which is consistent with the prevailing knowledge that exhaustive pairwise searches for epistasis severely limit statistical power. A common approach for addressing the limited power of genome-wide analyses is to employ a gene-wise cutoff[10, 120, 221, 34, 119]. Since each gene in an eQTL can be thought of as its own GWAS, the idea is to look next for tests that might not have had enough power to meet the global cutoff but show up as hits when only a single gene is considered[124]. Tests which meet the gene-wise Bonferroni cutoff in genes that have well-controlled QQ plots may still be useful. This analysis identified 2754 gene-wise Bonferroni-significant tests. While we don't suggest that all of these tests represent true positive signals, more tests are significant than would be expected by chance, suggesting that there could be a true signal present.

Of the 2754 gene-wise significant tests, 62 were significant both with and without correcting for main effects of the involved genotypes. 1304 had lower p-values after correction than before correction for main effects. 1 involved two *cis*-loci, 45 involved *cis-trans* loci, and 2708 involved two *trans*-loci. 2656 of these significant tests would not have been picked up by PLINK or Fish et al.[54] even when using the cutoff that would result from only considering one dataset at a time. Normally, one would need to perform r times as many statistical tests if testing datasets separately, so this cutoff would normally be more stringent.

An example of concordant epistasis discovered by Fomac is displayed in Figure 2.5. This gene-wise Bonferroni-significant concordant epistasis is not a hit before correcting for marginal effects ($p = 4.34 \times 10^{-3}$, green dot in top left QQ

plot of Figure 2.5), but afterwards the p-value for this concordant epistasis is $p = 8.75 \times 10^{-9}$, green dot in top right QQ plot of Figure 2.5). For reference, the global Bonferroni cutoff is 3.50×10^{-10} . The gene expression of ILMN_1749403 was corrected for batch effects, age, and main additive and dominance effects of genotypes rs1924458 and rs12708504, and the mean expression for each of the 9 genotype classes (min samples per genotype class was 35 here) was plotted (bottom three GP maps in Figure 2.5). An epistatic effect that is concordant across all three tissues (adipose, LCL, skin) is present visually. Since main effects have been corrected out of gene expression, this pattern must represent concordant pure epistasis.

2.4 Discussion

Here we presented Fomac, a novel statistic designed to detect instances of concordant epistasis among multiple quantitative traits. The versatility and utility of this statistic was demonstrated in simulations and real data analysis. In simulations, 6 parameter configurations were tested including higher order epistasis and many traits. In human data analysis, the Fomac framework produced a well-calibrated test statistic and many candidates for statistical epistasis that would not be identified by other approaches.

The large p , small n case presents an extreme challenge in epistasis detection, requiring at least two conceptual modules in the process of epistasis discovery. Since the approach outlined here only approaches the "identify" module, it depends on success in the initial "search" module. It seems that the primary limitation of this study has been the first module which I call the "search" module.

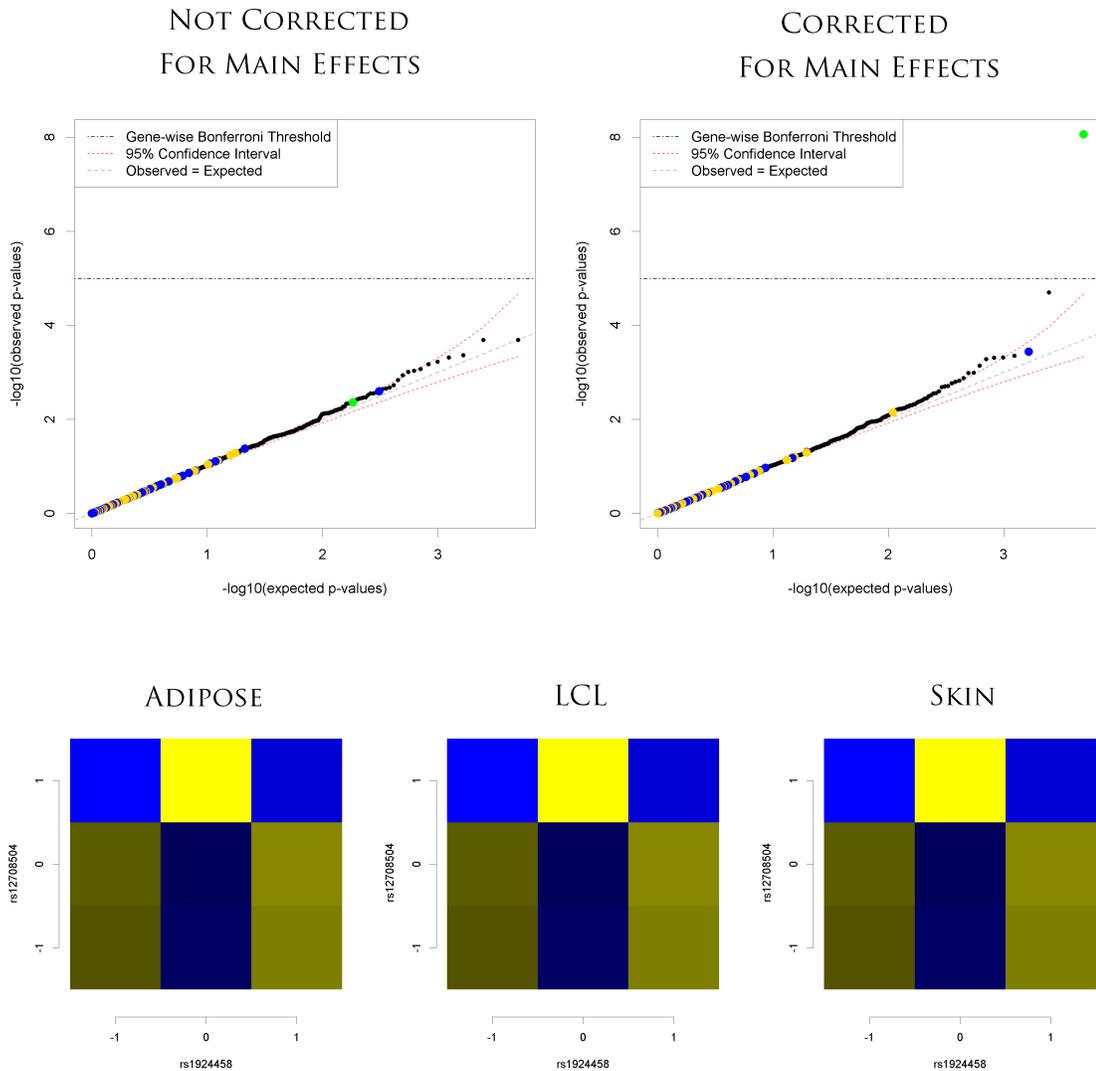


Figure 2.5: **Gene-wise QQ Plots for unadjusted (top left) and main-effect adjusted p-values (top right)**

This is actually good news because it means that Fomac can be combined with the plethora of methods designed to address the difficulties of searching for a needle in a haystack.

CHAPTER 3

WHOLE-METHYLOME ANALYSIS OF CIRCULATING MONOCYTES IN ACUTE DIABETIC CHARCOT FOOT REVEALS THE PRESENCE OF DIFFERENTIALLY METHYLATED GENES INVOLVED IN MIGRATION, DIFFERENTIATION AND FORMATION OF OSTEOCLASTS

Objectives: Increased differentiation of monocytes into osteoclasts, leading to bone resorption is the hallmark of Charcot foot (CF) disease.

Research design and methods: We studied the whole-methylome (WM) of circulating monocytes in 18 patients with type 2 diabetes (T2D) and acute CF, 18 T2D patients with equivalent neuropathy and 18 T2D patients without neuropathy, using the enhanced reduced representation bisulfite sequencing technique.

Results: WM analysis demonstrated that CF monocytes are differentially methylated compared to non-CF monocytes ($p = 5.1 \times 10^{-7}$). However, analysis comparing T2D patients to T2D patients with neuropathy did not show any significant differential methylation. 13 out of > 1.2 million individual CpG sites had p-values lower than the Benjamini-Hochberg cutoff when comparing CF patients to non-CF patients. Analysis of individual genes showed that 114 out of 21,542 genes were differentially methylated for CF versus non-CF. Out of those genes, 23 genes are involved in the migration process during monocytes differentiation into osteoclasts or indirectly involved in osteoclast formation through the regulation of inflammatory pathways. Finally, we demonstrated association between DNA methylation and gene expression in cis and trans-association. Interestingly, PPP2R5D was the only gene both differentially methylated and expressed (cis-association) in CF patients. We also identified 27 genes with significant changes in methylation between CF and non-CF that were significantly

associated with expression in 24 genes (*trans*-association).

Conclusion: In total, our findings unveil a possible role of circulating monocytes HM in the pathogenesis of CF, with the ultimate goal of finding means to modulate or prevent it.

3.1 Introduction

Charcot foot (CF) disease is a devastating complication of diabetes, associated with an increased risk of soft tissue infections, foot ulcers and amputations [198]. It is characterized by an exaggerated bone resorption [246], believed to be induced by an increased numbers of osteoclasts and their activity [137, 175]. Osteoclasts are derived from monocytes, mostly from CD14+ that have the highest potential to differentiate, following a differentiation pathway that results in mature functional osteoclasts whose role is to activate bone resorption [225].

Epigenetics modulate the differentiation of many adult cell types from progenitor or primary cells with whom they share the same DNA sequence, and play an important role in gene transcription through 3 major components: DNA methylation, non-coding RNAs, and post-translational changes of histone proteins [15]. Methylome, which is the set of nucleic acid methylation modifications in an organisms genome or in a particular cell [14], also participate in the pathophysiology of several diseases by controlling cellular differentiation processes and transcriptional activities of genes [182]. Therefore, we hypothesized that the methylome of circulating monocytes in patients with acute diabetic CF could be involved in the pathogenesis of the disease.

In here, we report the presence of differentially methylated genes involved in migration, differentiation and formation of osteoclasts in circulating monocytes of type 2 diabetes (T2D) patients with acute CF, compared to T2D patients without CF.

3.2 Methods

3.2.1 Subjects

Eighteen T2D patients with acute CF, matched for age, gender, BMI and HBA1C with 18 T2D patients with neuropathy but no CF and 18 T2D without neuropathy or CF were studied (Table 3.1). CF patients were recruited from the podiatry clinic at Hamad Medical Corporation (HMC), Doha- Qatar. All other patients were recruited from the department of endocrinology and diabetes at HMC.

Acute CF was diagnosed according to the American Diabetes Association and the American Podiatric Medical Association task force [198]. Patients had to have a red swollen foot with increased local temperature of more than 20 C compared to the contralateral foot with X-Ray evidence of acute CF. Foot temperature was measured using FLUKE Ti32 thermal imager (Fluke Corporation - USA). All patients with acute CF had neuropathy, diagnosis of which was based on the vibration perception threshold (Neurothesiometer NU-1, Horwell- UK) on the great toe being $>25V$ [245]. Among those, 5 had dislocations or subluxations, 4 had fractures, 5 had periosteal reactions and 4 had bone destruction on X-ray.

Table 3.1: **Baseline characteristics of the participants included in the study.** Diabetes but no neuropathy (D), diabetes with neuropathy (DN), and diabetes with both neuropathy and Charcot foot (DCh). Data are represented as mean (standard deviation). p-values were calculated with ANOVA test.

Baseline characteristics	Diabetes with Charcot N=18	Diabetes with neuropathy N=18	Diabetes without neuropathy N=18	P value
Age, years	52 (9.5)	55 (8.3)	54 (10.5)	0.66
Gender, M/F	13/5	13/5	13/5	-
BMI, kg/m ²	31 (5.9)	32.6 (7.4)	32.7 (6.7)	0.72
HbA1c, %	8.1 (1.7)	8.3 (2.2)	8.1 (1.7)	0.88
Fasting Glucose, mmol/L	9.1 (4.6)	10 (4.5)	9.9 (3.9)	0.79
Diabetes Duration, years	14.5 (6)	15.7 (6.6)	10.7 (11.9)	0.20
Systolic BP, mmHg	132 (15)	138 (17)	137 (19)	0.53
Diastolic BP, mmHg	74 (7.6)	73 (9.7)	78 (9.3)	0.21
eGFR, ml/min	77 (32.7)	81 (37.6)	101 (22.6)	0.07

The study was approved by the institutional review boards of Weill Cornell Medicine-Qatar and HMC (13-00031 and 14-14054, respectively). All participants provided written informed consent. The study was conducted in accordance with the 1964 Declaration of Helsinki and was registered at clinicaltrials.gov (NCT02316483).

3.2.2 Monocytes Isolation and DNA/RNA Extraction

10 mL of blood was withdrawn from peripheral venous puncture from each participant. Peripheral blood mononuclear cells (PBMCs) were first isolated from whole blood and stained with the mouse anti-human IgG2b CD14 APC and the mouse anti-human IgG1 CD16-PE (BD bioscience). Monocytes were then sorted using FACS Aria2 flow cytometer (BD Biosciences). Purity of the sorting was controlled after each sorting (Figure B.1). DNA and RNA from monocytes were extracted (Allprep DNA/RNA mini kit Qiagen) and stored at -80C, then shipped to the epigenomics core at Weill Cornell Medicine (WCM) and the New York Genome Center (NYGC) using dry ice for sequencing.

3.2.3 Enhanced Reduced Representation Bisulfite Sequencing and Data Processing

Enhanced Reduced Representation Bisulfite Sequencing (ERRBS) libraries, sequencing, data alignment and methylation calls were generated at the Epigenomics Core at WCM as described in Garrett-Bakelman et al [64]. The published protocol was modified as follows: samples were size selected on a 2% agarose cassette using a Pippin HT (Sage Science, Beverly, MA), and two size fragment lengths of 240375 bp and 375550 bp were recovered and further processed. Samples were checked for quality via two methods: (1) the distribution of CpG site coverage (Figure B.2A) - experiments that are suffering from PCR duplication bias will have a secondary peak to the right of the primary peak, and (2) distribution of methylation -values (Figure B.2B) this histogram should have a peak towards zero methylation and a peak towards methylation of one [4].

Two analysis approaches were carried out:

(1) CpG site analysis 1,240,581 CpG sites which had coverage of at least 10* across all patients were used.

(2) Gene-mapped analysis For each gene, CpG sites within an interval 2 kb upstream and 2 kb downstream of the transcription start site (TSS - taken from ENSEMBL annotation) were used to determine a methylation level for the gene since a majority of CpG islands are within 2kb of TSS [258]. The methylation of a gene was calculated as the mean methylation of measured sites within the interval, weighted by the coverage at each site [3]. Twenty-one thousand six hundred thirty-four genes had at least 1 CpG site within the mapping interval for all patients. More information regarding CpG sites and mapping to genes can be found in Table 3.2.

DNA from circulating monocytes was sequenced in three distinct batches. In order to determine if a batch effect should be considered in subsequent analysis. The first two principal components of the gene-mapped CpG data were plotted and colored by batch (Figure B.2C-D). It was determined that batch had affected the methylation measurements and should therefore be accounted for in statistical analysis.

3.2.4 Differential Methylation

Patients were placed in one of three groups:

D: group with diabetes but no neuropathy (n = 18)

Table 3.2: **Genome-wide methylation study on Charcot foot.** In order to identify methylation differences specific to Charcot foot, circulating monocytes were isolated from blood of patients with diabetes but no neuropathy (D), diabetes with neuropathy (DN), and diabetes with both neuropathy and Charcot foot (DCh). CpG methylation data was produced by enhanced reduced representation bisulfite sequencing (ERRBS). Each group has n = 18 and patients are matched for age, gender, BMI, and HbA1c across all groups such that these covariates were not significantly different among groups (rightmost column). Methylation of a gene was calculated as the mean (weighted by coverage) observed CpG sites within 2kb upstream and 2kb downstream of transcription start site, as determined by Ensembl annotation. The bottom portion of the table provides summary statistics of the number of CpG sites within the mapping interval of a gene, and the total number of CpG sites with coverage of at least 10 (sites with coverage less than 10 were not used).

	Type 2 diabetes with Charcot Foot (18 subjects, 5 females)		Type 2 diabetes with Neuropathy (18 subjects, 5 females)		Type 2 diabetes without Neuropathy (18 subjects, 5 females)		Group comparison
	Mean (s.d.)	Range	Mean (s.d.)	Range	Mean (s.d.)	Range	<i>p</i>
Number of CpG sites that map to a gene ²	1099669 (44842)	1000416 – 1168307	1141969 (56389)	1040115 – 1213237	1085492 (65237)	886300 – 1193351	6.5 × 10 ⁻³
Number of total CpG sites with coverage of at least 10	2706340 (113051)	2546276 – 2902798	2778313 (192791)	2441023 – 3008473	2595268 (180000)	2036173 – 2843919	1.1 × 10 ⁻²
Number of genes with mapped CpG sites	28729 (620)	27967 – 30278	28784 (1397)	25546 – 31268	29812 (1111)	28638 – 33160	6.2 × 10 ⁻³

DN: group with diabetes and neuropathy but no Charcot foot (n = 18)

DCh: group with diabetes and Charcot foot (n = 18)

In order to test for methylation differences that were specific to Charcot foot, patients were also grouped as either having Charcot foot (DCh) or not having Charcot foot (DDN, n = 36). DDN grouping is of primary interest for two reasons: (1) it will tend to identify genes that are different due to CF (which involves neuropathy) without identifying genes that are different due to purely neuropathic reasons, and (2) it will provide more total samples and therefore greater statistical power than testing for differences between CF and either one of the non-CF groups. This is true as long as the two subgroups being grouped together are not significantly different from each other. Therefore, in order to statistically justify this grouping, differential methylation was also tested for the three possible pairings of the three groups: D versus DN, D vs DCh, and DN vs DCh (Figures B.3 & B.4). If methylation differences are detected in the D vs DCh and DN vs DCh comparisons but not in the D versus DN comparison, then grouping non-CF patients together for comparison with CF is justified. Singular value decomposition [49] was performed for both methylation of the 1,220,216 autosomal CpG sites and methylation of the 21,049 autosomal genes. Sex chromosomes were excluded just for this portion of the analysis so that the samples did not stratify based on gender. The left-singular vectors represent independent methylation features that can capture large amounts of variance from the original data. Since the left-singular vectors capture much of the variance from the original data and represent a signal from many genes, they can be considered a good proxy for the whole methylome. In order to test for a difference in group methylomes the first three left-singular vectors from

(Figures 3.1B & 3.1D) were regressed on covariates (batch) and group using multivariate linear regression (this was done separately for the CpG site and gene-mapped approaches), which tests for a multivariate difference between group. Next, all 1,240,581 CpG sites and 21,634 individual genes were tested for differential methylation by regressing a site/gene on covariates (batch) and group using an analogous univariate linear regression methodology which tests for a univariate difference between group means. Both WM differential methylation and individual site/gene differential methylation for both CpG site and gene-mapped methylation were tested for 4 groupings of patients: DDN/DCh, D/DN, D/DCh and DN/DCh. P-values were produced using a likelihood ratio test of the model containing group versus the model not containing group. Linear regression has been shown to possess similar statistical power to Wilcoxon rank sum test, Kolmogorov-Smirnov test, permutation test, empirical Bayes method, and bump hunting method in simulated DNA methylation studies with sample sizes greater than $n = 12$ in each group [127]. Significance was assessed using a false discovery rate of $\alpha = 0.05$ for both Bonferroni and BH multiple test correction procedures. In order to compare the differential methylation findings to what would be expected by chance, each differential methylation analysis was carried out an additional 10 times on permuted data. A samples group status is permuted in order to disrupt any association between group and methylation. The association between methylation and batch is maintained since this association is required for appropriate covariate correction. These permutation analyses give a sense empirically of how far our original differential methylation p-values are from what we would observe if there were truly no association between group and methylation [46]. The results of this comparison between original and permuted data are displayed in quantile-quantile (QQ)

plots (Figure B.5).

3.2.5 Gene Expression Data

Data generation and filtering - DNA was synthesized from 10ng of good quality total RNA (RIN>7) using SMART-SEQ v4 Ultra Low Input RNA Kit (ClonTech) at the New York Genome Center (NYGC) according to the manufacturers protocol with 8 cycles of amplification. Resulting cDNA was cleaned up with a 1:1 volume ratio of AMPURE XP beads (Beckman) and evaluated on the Fragment Analyzer using a High Sensitivity DNA Assay (AATI). Full-length cDNA was sheared to an average size of 350bp fragments using Adaptive Focused Acoustics (AFA) technology (Covaris, LE220). Illumina-compatible libraries were prepared with KAPA Hyper Prep Kit (Roche) and Illumina dual indexed adapters according to the manufacturers specifications with 9 cycles of amplification. The libraries were quantified by picogreen assay and NGS assay (Fragment Analyzer, AATI) and sequenced on an Illumina HiSeq2500 sequencer (v4 chemistry, v2 chemistry for Rapid Run) using 2 × 50bp cycles. *RNA-sequencing analysis* - The reads were aligned with STAR (version 2.4.0c), and genes annotated in Gencode v18 were quantified with featureCounts (v1.4.3-p1). All the genes with less than one read across all the samples were not taken into the consideration which resulted with total set of 16007 genes for 30 samples. Normalization of expression was performed using the Bioconductor package DESeq2 using the rlogTransformation function [135].

Association between DNA methylation and gene expression in Charcot foot patients

RNA sequencing was done at NYGC. We tested the association between DNA methylation (independent variable) and \log_2 transformed gene expression (dependent variable) using linear model in which diseases status and batch effect were used as covariates. We included only BH significant genes on methylome-wide level (Table 3.1); $n=2,488$ genes) against their transcripts in our linear regression analysis. Since a total of 1,326 genes were common in both data sets, we performed p-value correction based on BH criteria for 1,326 genes. Subsequently, to identify potential trans effect between methylation and expression in CF, as well as to identify all the other associations relevant to CF condition we perform multiple test correction included all methylated and all expressed genes. As we did not identify a large number of significant CpG sites at methylome-wide level, only methylation-expression association on gene level were considered in these analyses. Gene ontology (GO) enrichment analysis [254] was performed using STRING database. STRING is a database of known and predicted proteinprotein interactions, direct (physical) and indirect (functional) associations [217].

3.3 Results

Combining diabetic patients with and without neuropathy

This analysis consisted of 4 ways of assessing differential methylation, produced by all combinations of 2 different ways of looking at methylation (individual CpG site methylation and gene-mapped methylation), and 2 different ways of looking at differences in methylation (wholomethylome and individual site/gene). For each way of assessing differential methylation, four comparisons were carried out: DDN/DCh, D/DN, D/DCh, and DN/DCh. The first grouping provides the primary result of interest, since it will identify differential methylation that is specific to CF and not due to either diabetes or the neuropathy that can afflict acute CF disease. However, in order to justify this grouping, it needs to be shown that diabetic patients with and without neuropathy can be considered similar enough to group together. The last three comparisons are therefore used to establish the similarity of the D and DN groups relative to the differences evident in both the D/DCh and DN/DCh comparisons.

Table 3.3 shows that the D/DN comparison identified 1 differentially methylated gene, 0 differentially methylated CpG sites, and was not able to stratify the top 3 principal components in either the CpG site ($p = 0.92$) or gene-mapped ($p = 9.6 \times 10^{-2}$, insignificant after multiple test correction) approaches. On the other hand, the D/DCh and D/DN comparisons identified several Bonferroni significant and several hundred BH significant differentially methylated genes, and both comparisons were able to stratify the top 3 principal components in both approaches (Table 3.3). These results demonstrate that patients with diabetes are not significantly different than patients with diabetes and neuropathy

when assessed for differential methylation in circulating monocytes using the discussed methods. Therefore, it is valid to group D and DN together to use as a single group to compare with CF patients using the same methods that were used to establish similarity of the non-CF patients.

Table 3.3: Differential methylation results for 4 groupings of patients.

Results of 4 differential methylation approaches: methylome-wide differential methylation (p-value derived from multivariate linear regression fit using the first 3 principal components as dependent variables as well as batch and group as independent variables) and individual site/gene differential methylation for both CpG site (top 3 rows) and gene-mapped methylation (bottom 3 rows). For each approach four group comparisons were made, represented by the four columns of the table which are labeled based on the two patient groups that were compared.

		DDN/DCh	D/DN	D/DCh	DN/DCh
CpG Site	Methylome-wide p-value	3.3×10^{-4}	0.92	1.0×10^{-2}	3.9×10^{-3}
	Bonferroni-significant sites	4	0	0	0
	BH-significant sites	13	0	0	0
Gene-Mapped	Methylome-wide p-value	1.9×10^{-5}	9.6×10^{-2}	4.6×10^{-4}	3.2×10^{-4}
	Bonferroni-significant genes	114	1	19	11
	BH-significant genes	2488	1	1494	1052

3.3.1 Comparing Diabetic Patients With and Without Neuropathy to Patients With Charcot Foot

WM analysis (which looks at linear combinations of many CpG sites or genes) demonstrates that CF monocytes are differentially methylated when compared to non-CF monocytes (DDN/DCh) in both the CpG site (Figure 3.1A-B) and

gene-mapped (Figure 3.1C-D) approaches. Individual genes reinforce this result with 114 out of 21,634 genes having p-values lower than the Bonferroni-significant cutoff of $p = 2.3 \times 10^{-6}$ for DDN/DCh (Figures 3.1E & B.5E). 13 out of 1,240,581 individual CpG sites had p-values lower than the BH cutoff of $p = 5.2^{-7}$ for DDN/DCh (Figures 3.1F & B.5A). This particular way of assessing differential methylation suffered from a very stringent multiple test correction since over a million CpG sites were tested individually. From a theoretical standpoint, power to detect a given effect size decreases as a researcher performs more tests to identify the effect. Therefore, it makes sense that fewer individual CpG sites than genes were identified as significantly differentially methylated.

Permutation analysis demonstrates that the effect of group on methylation vanishes when group is permuted. This analysis allows the structure of covariates with respect to methylation to remain intact while disrupting any association between group and methylation. These permutation analyses can be thought of as a way of accessing the empirical null distribution, which isn't always guaranteed to match the theoretical null distribution [16]. When the empirical null p-value distribution matches the theoretical null p-value distribution (uniform) or is skewed towards 1, the true empirical p-values are expected to be at least as low as the p-values that were calculated. All eight QQ plots (Figure B.5) show that permutation analyses return p values either matching the uniform distribution or skewed towards 1 (this is represented in the plots as the cyan line being below the dashed gray line). This result serves to justify the statistical methodology that was used. The QQ plots (Figures B.5B & B.5F) also serve as a way of further confirming that D and DN do not have different methylation patterns since inspection reveals that the original p-values overlap with

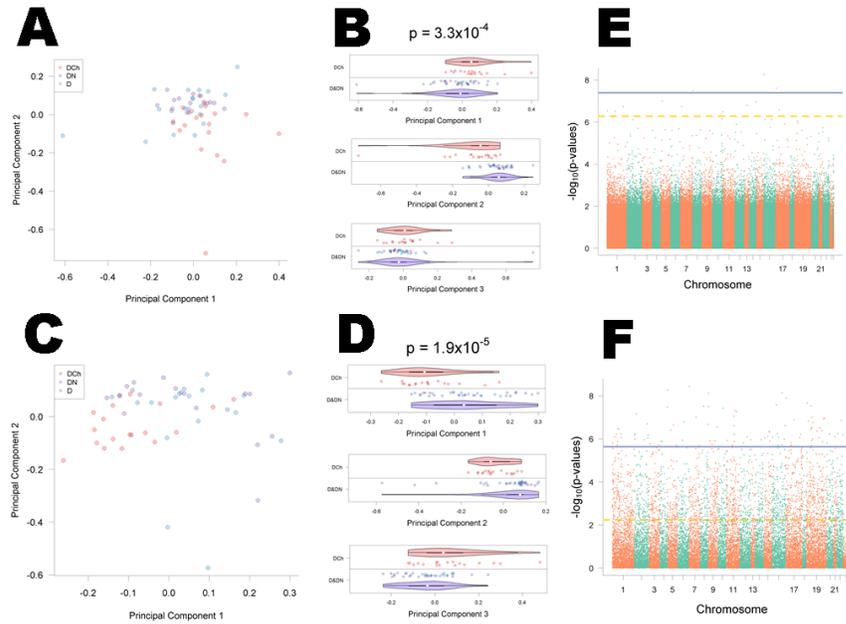


Figure 3.1: CpG site and Gene-mapped differential methylation in patients with diabetes and CF compared to patients with diabetes but no CF. **A.** The first two principal components of autosomal gene methylation, as calculated by singular value decomposition. Samples are colored by group: diabetes in blue, diabetes with neuropathy in purple, and diabetes with CF in red. **B.** Whole-methylome signal as captured by the first three principal components (horizontal axes) of the displayed subset of patients. Violin plots representing the distribution of a particular patient group along a principal component are adjacent to their corresponding group. **C.** The first two principal components of autosomal gene methylation, as calculated by singular value decomposition. Samples are colored by group: diabetes in blue, diabetes with neuropathy in purple, and diabetes with CF in red. **D.** Whole-methylome signal as captured by the first three principal components (horizontal axes) of the displayed subset of patients. Violin plots representing the distribution of a particular patient group along a principal component are adjacent to their corresponding group. **E.** Chromosomal distribution of gene methylation differences. For each gene, the significance is displayed on the y-axis as the \log_{10} of the p-value. The results are ordered along the x-axis by chromosome, with each bar representing a different chromosome. The Bonferroni and BH p-value thresholds ($\alpha = 0.05$) were 2.3×10^{-6} (yellow line) and 5.8×10^{-3} (green dashed line), respectively. **F.** Chromosomal distribution of gene methylation differences. For each gene, the significance is displayed on the y-axis as the \log_{10} of the p-value. The results are ordered along the x-axis by chromosome, with each bar representing a different chromosome. The Bonferroni and Benjamini-Hochberg p-value thresholds ($\alpha = 0.05$) were 4.0×10^{-8} (yellow line) and 5.2×10^{-7} (green dashed line), respectively.

the permutation p-values. The differential methylation of CF compared to non-CF foot using both CpG site and gene-mapped approaches also demonstrate that the way methylation was mapped to genes is at least as effective at capturing discriminatory information as looking at individual CpG sites. This isn't a reference to the number of significant sites or genes identified, instead it has to do with the actual p values. Comparing Figures 3.1E and 3.1F, it is evident that the mapped approach is producing p values at least as low as those produced from the CpG site approach. This finding matters because it means that this approach has largely been able to avoid the pitfall of combining discordant signals from nearby CpG sites, which would be detrimental to the discrimination task. Therefore, we focused the rest of the analysis on the gene-mapped differential methylation.

3.3.2 Gene-mapped Differential Methylation in Patients with Diabetes and Charcot Foot Compared to Patients with Diabetes But no Charcot Foot

114 out of 21,634 genes were differentially methylated in CF patients compared to non-CF monocytes when looking at the gene-mapped methylation. When we take a closer look to the 114 genes one by one, we notice that 37 genes out of the 114 Bonferroni-significant genes differently methylated for CF versus non-CF have an unknown function. Most of the 114 genes presented a hypermethylation (86%) in CF patients (Figure 3.2A). Figure 3.2B represents the 10 top hypermethylated (top part, red) and hypomethylated (bottom part, green) genes in CF patients.

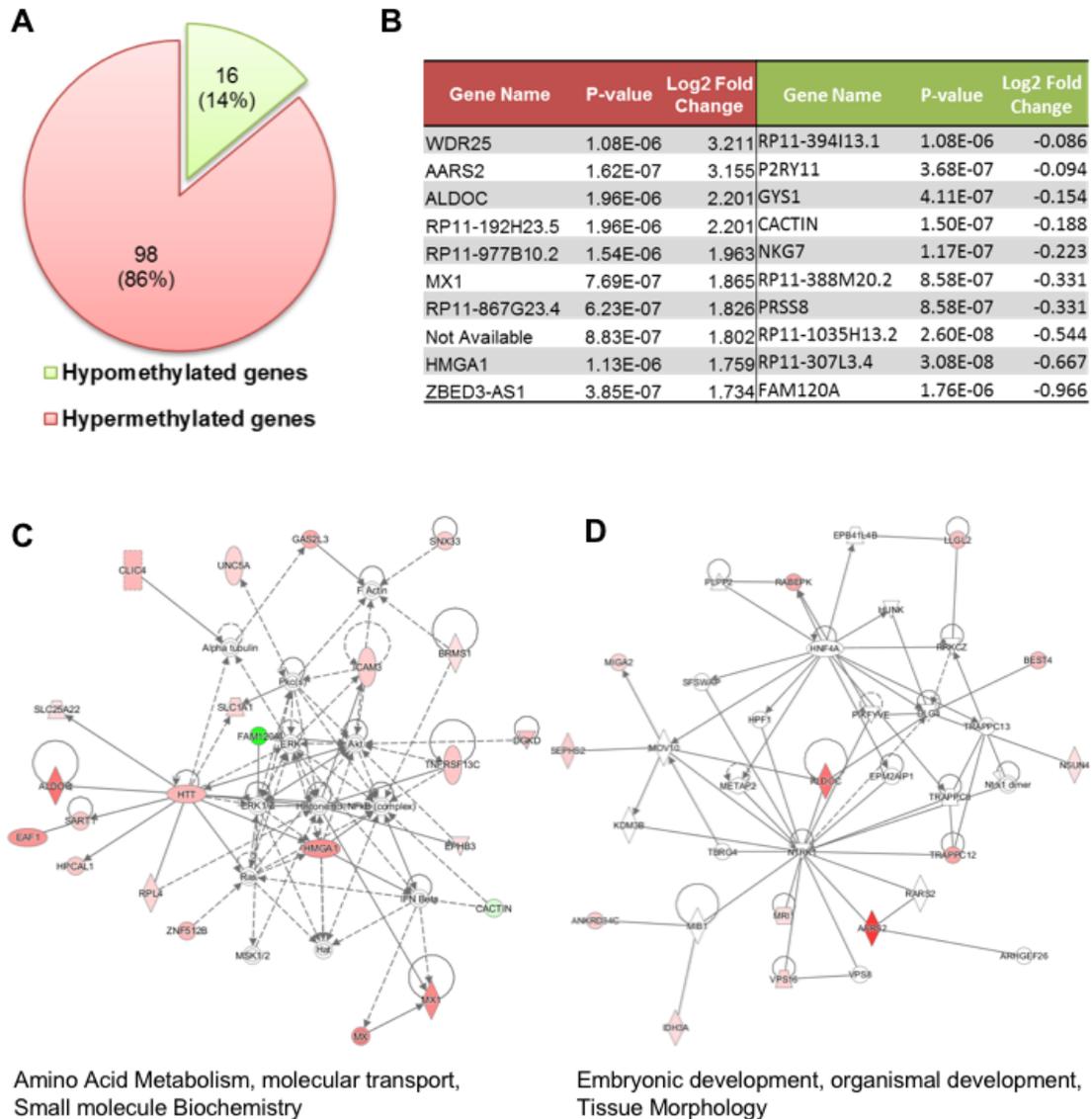


Figure 3.2: Gene analysis for gene-mapped differential methylation in patients with diabetes and Charcot foot compared to patients with diabetes but no Charcot foot. **A.** Graphic representation of the number of hypo- and hyper-methylated genes. **B.** Representation of the 10 top hypermethylated (top part, red) and hypomethylated (bottom part, green) genes in CF patients compared to patients with diabetes but no Charcot foot. **C-D.** Networks of altered genes mapped differential methylation created by IPA. The hypothetical networks generated by IPA based on the molecular relationships, interactions, and pathway associations between the methylated candidate genes are shown in a graphical representation.

IPA global analysis of the 114 Bonferroni-significant genes differently methylated for CF versus non-CF revealed significant enrichment of the category Amino Acid Metabolism, molecular transport, Small molecule Biochemistry (Figure 3.2C) as well as the category involving Embryonic development, organismal development, Tissue Morphology (Figure 3.2D). This observation indicates that circulating monocytes in CF patients seem to be ready for differentiation. Both enriched classes are coherent with the experimental design.

A total of 23 genes could be involved directly or indirectly in monocyte differentiation into osteoclast which represent around 15% of the total genes differentially methylated in CF patients (and around 30% of the gene with known function). One of the top hits in this list is MAPK11 that is known to enhance osteoclastogenesis and bone resorption in breast cancer [79].

3.3.3 Association between DNA methylation and gene expression in Charcot foot patients

Abnormal DNA methylation can result in aberrant gene expression [236]. Therefore, we investigated gene expression in our samples and try to associate it with the DNA methylation. First, we identified 2488 significant genes with BH correction (BH p-value cutoff = 5.7^{-3}). Of these 2488, 818 genes were down-regulated and 1670 upregulated. Gene methylation can have 2 different types of associations with gene expression: i) Local association (*cis*-) where one differentially methylated gene (A) will have an effect on the exact same gene (A) on expression level, or ii) Distal association (*trans*-) where one differentially methylated gene (A) will have an effect on a different gene (B) on expression level.

In order to test for *cis*- and *trans*-effect of the methylation on gene-expression, we performed linear-regression for 2488 genes, treating gene expression as dependent variable, and methylation as independent variable using disease status and batch effect as covariates. PPP2R5D was the only gene detected with a *cis*-association with expression. The correlation coefficient beta is 4.4, so expression and methylation of this gene are positively correlated. Both methylation and expression are increasing in CF patients. For the *trans*-association, we identified 27 genes out of 2488 genes with significant changes in methylation between CF and non-CF that were significantly associated with expression in 24 genes. Four (MTCL1P1, ITGAL, DHX40, GFOD2) of these genes were hypomethylated when comparing CF to non-CF samples, while the others were hypermethylated (Figure B.4). We created a minimum network of protein-protein interaction using the 27 CF differentially methylated genes and their 24 expression-associated genes (total of 51 genes) leading to 32 interactions (Figure 3.3B). Out of those 32 interactions, 7 had a positive correlation coefficient beta and 27 a negative one, suggesting suppression of expression associated with hypermethylation or gene sur-expression associated with hypomethylation.

3.4 Discussion

In this study, we demonstrated that methylation of circulating monocytes is involved in the pathogenesis of acute CF. The strength of our study relies on the fact that we used only one cell type from PBMCs. Several reports have already indicated that methylation changes could be cellspecific, and that several variations between cell types exist within the same individual [248]. Miao et al [152] demonstrated that monocytes and lymphocytes have distinct epigenomes

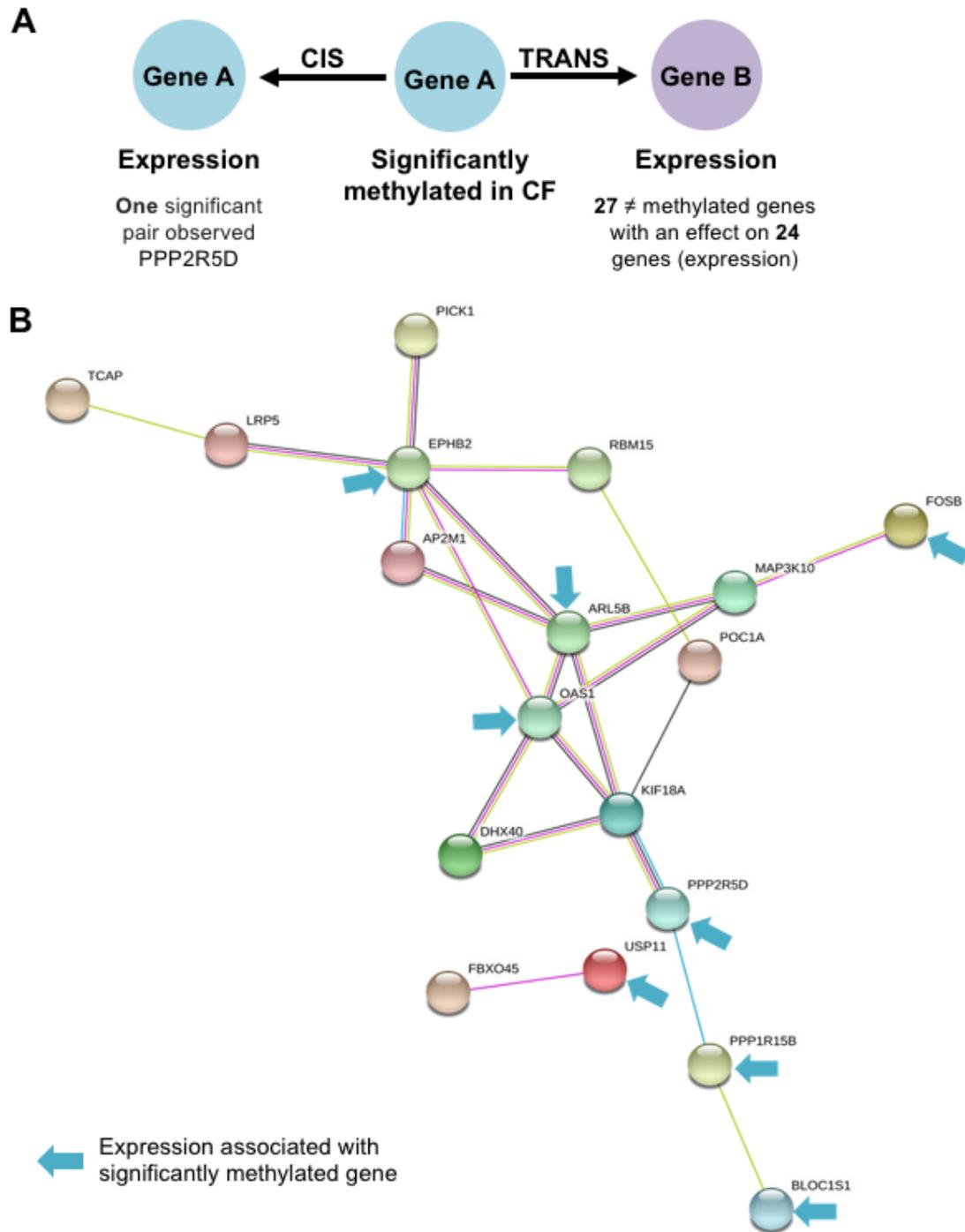


Figure 3.3: Association between methylation and expression in Charcot foot. **A.** Schematic representation of CIS and TRANS association between methylated and expressed gene. **B.** Minimum network of protein-protein interaction using the 27 CF differentially methylated genes and their 24 expression associated genes (total of 51 genes) leading to 32 interactions using STRING.

whereas patterns within a specific cell type are remarkably similar despite age or gender. It might be therefore inappropriate to assess disease-specific methylation changes using PBMCs as a surrogate endpoint. While several reports linked methylation changes to diabetes and its complications, most of them reported differential gene methylation in PBMCs and only a few assessed cell- or tissuespecific methylation changes in diabetes [174]. In pancreatic islets from T2D patients, Volkmar et al [229] identified 276 CpG loci affiliated to promoters of 254 genes differentially methylated comparing to non-diabetic islets. Nillsson et al [170] identified over 250 differentially methylated CpG loci in liver tissues of obese T2D patients as compared to non-diabetic individuals. In our study, we showed that the analysis of circulating monocytes whole-methylome was not able to differentiate between the groups of D and DN. CF is a neuropathic osteoarthropathy [198], and the fact that monocyte WM couldnt discriminate D and DN demonstrates that methylation differences that we uncovered in this study are specific to CF.

We identified several genes that are differentially methylated in circulating monocytes of patients with CF. Furthermore, most of those genes were involved in the migration process of monocytes and their differentiation into osteoclasts. The top hit genes were HMGA1 and MAPK11, both hypermethylated. The important role of P38B (MAPK11) in osteolytic bone destruction has been demonstrated in the context of breast cancer [79]. Upregulation of MCP1 expression by MAPK11 leads to the enhanced osteoclast differentiation and bone resorption. MPAK11 has also been shown to be a regulator of TNF gene expression in mononuclear phagocytes [139]. HMGA1 is a downstream nuclear target of the insulin receptor signaling pathway [24]. HMGA1 is a master regulator of tumor progression by driving inflammatory pathway and cell cycle progression genes

during tumorigenesis [202].

To strengthen our methylation data, we performed a transcriptomic study in order to link methylation to gene expression. We demonstrated that only PPP2R5D was significantly methylated and expressed (*cis*-association) in CF patients. The product of this gene belongs to the phosphatase-2A regulatory subunit B family that is known to be implicated in the negative control of cell growth and division. In our study, PPP2R5D was hypermethylated and over-expressed in monocytes of patients with CF, suggesting a decrease of their cell growth. Interestingly, monocytes have to decrease their division in order to differentiate into osteoclasts [219].

The study of trans-association revealed 27 CF-differentially methylated genes having an effect on 24 expression-associated genes. Among them, some play an important role in the length of long bones (POC1A) [199] or monocyte trafficking (FOSB) [118]; others are related to glycemic traits in type 1 diabetes and T2D [70, 186], insulin resistance (EPHB2) [114] and Golgi trafficking (ARL5B) [90]. We believe that POC1A and FOSB need to be more investigated in the context of CF disease. In fact, both are involved in bone-related disorders such as short stature, onychodysplasia, bone loss or osteosarcoma, rendering them potential candidate for further functional validation studies in CF disease.

To our knowledge, we are the first to report differential methylation changes in diabetes and related complications using the ERRBS. Reduced representation bisulfite sequencing (RRBS), and related methods such as ERRBS, are sequencing methods that enrich for CpG-rich parts of the genome, which enables sequencing of genomic regions where 5-methylcytosine modifications can alter gene expression via binding to gene promoters and bodies [64, 4, 76]. The ad-

vantage of this method over whole genome bisulfite sequencing (WGBS) is increased coverage depth and hence higher confidence variant calling for a fixed volume of sequencing reads. With respect to arraybased epigenome genotyping platforms such as the HELP array, ERRBS provides higher sensitivity for rare and population-specific variants [64, 76, 115].

Implications of epigenetics generally, and gene methylation in particular, are being increasingly used in clinical settings. While drugs that modulate DNA methylation of cancer cells are already used in oncology treatment [107], the pharmaco-epigenetic therapy in T2D and cardiovascular disease is currently limited to experimental studies. For example, inhibition of the methyltransferase SETD7 that is required for DNA methylation in macrophages resulted in a decrease in reactive oxygen species and up-regulation of anti-oxydant genes [78]. Similarly, in an experimental model of db/db mice with diabetic nephropathy, the angiotensin receptor blocker losartan reverses back the methylation of the histone H3K9 that is observed in mesangial cells under hyperglycemia.

We acknowledge the presence of few limitations in our study. First, our sample size is relatively small; thus, a higher number of study participants might have enabled us to detect more methylation calls. Additionally, we used circulating monocytes as surrogate markers for CF knowing that the disease is only limited to the foot. Despite the presence of those limitations, we were able to conduct a comprehensive analysis of circulating monocytes methylome in patients with acute diabetic CF and we demonstrated the presence of differentially methylated genes involved in migration, differentiation and formation of osteoclasts from circulating monocytes. Moreover, we were able to associate the difference in methylation with gene expression. Further studies are needed

to determine the timing of changes in methylation/expression, whether they precede evidence of acute CF disease or if they could be a downstream effect. Additionally, it would be important to assess whether the differential methylation in the acute stage will be present or not in chronic CF disease, and if those changes are reversible in patients who recover. Nevertheless, our findings could be used to elucidate the cause of CF, with the ultimate goal of finding means to modulate or prevent it. Finally, similar methodology could be used to evaluate the methylome of circulating monocytes in patients with conditions of increased bone resorption.

APPENDIX A
CHAPTER 1 OF APPENDIX

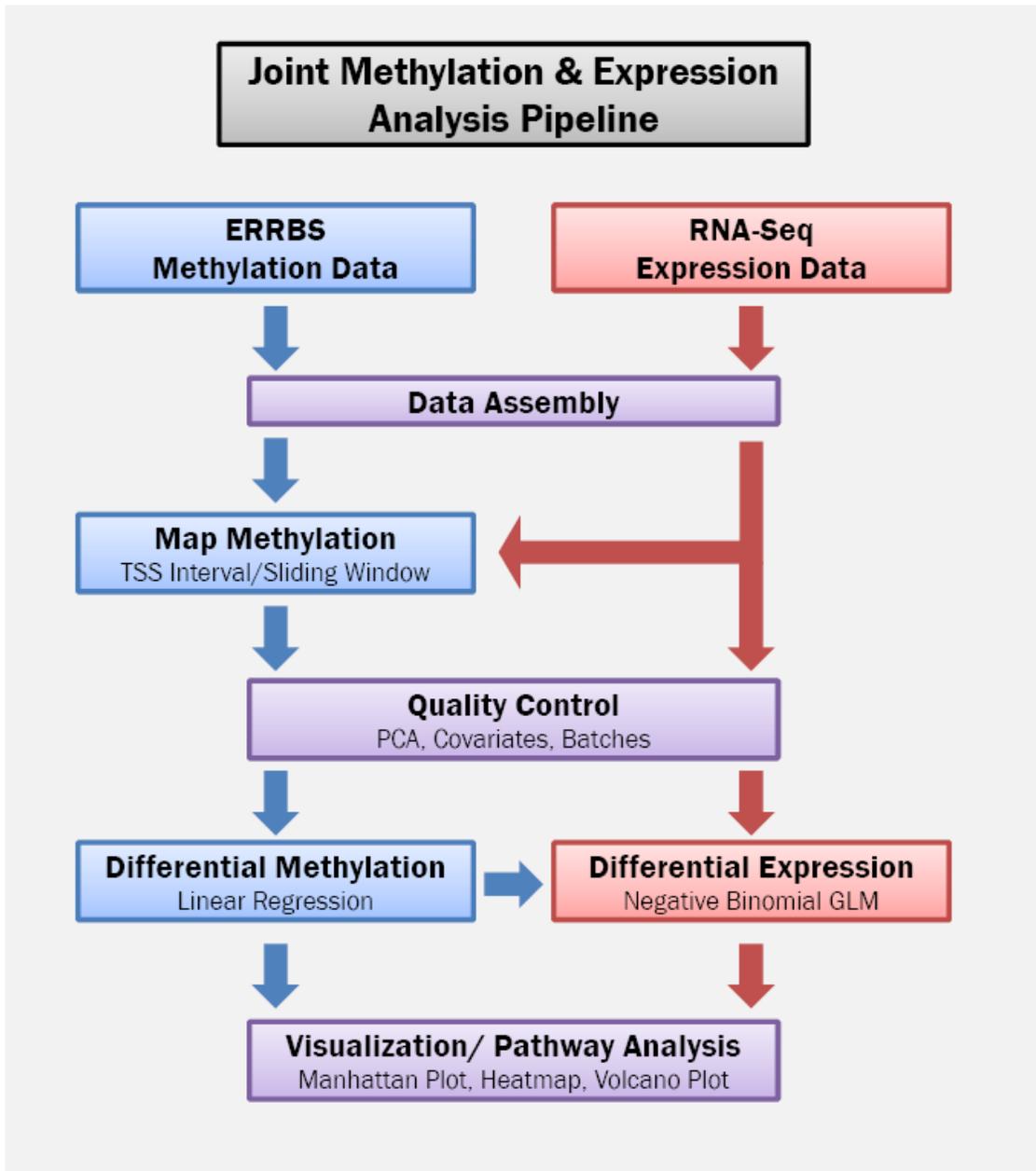


Figure A.1: **Joint Methylation & Expression Analysis Pipeline** Visualization of the pipeline which takes enhanced reduced representation bisulfite sequencing (ERRBS) CpG methylation data and integrates it with RNA-Seq gene expression data to produce data-driven candidate regulatory genes and sites

APPENDIX B
CHAPTER 3 OF APPENDIX

Supplementary Figure 1

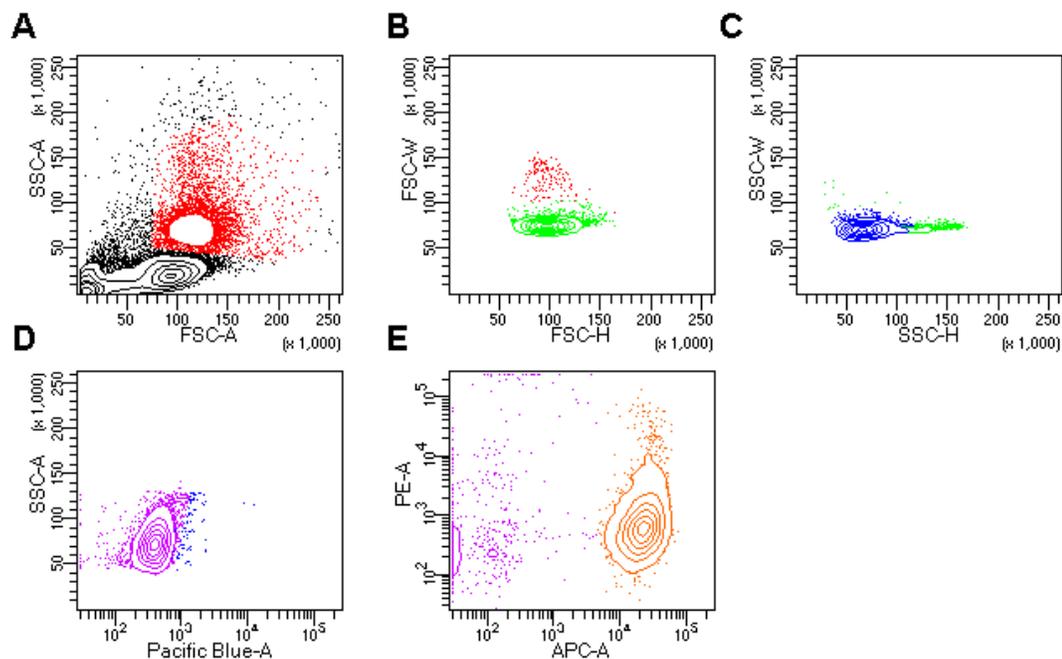


Figure B.1: Gating strategy for monocyte sorting. **A.** Monocyte populations were gated (red population) using SSC/FSC. **B-C** Using FSC-W/FSC-h (B) and SSC-W/SSC-H (C), the doublets were excluded and only the living cells (blue population) were kept. **D.** Auto-fluorescent cells were excluded using Pacific-Blue channel. **E.** Final monocyte population was gated (orange population) as CD14(APC)+CD16(PE)+/-.

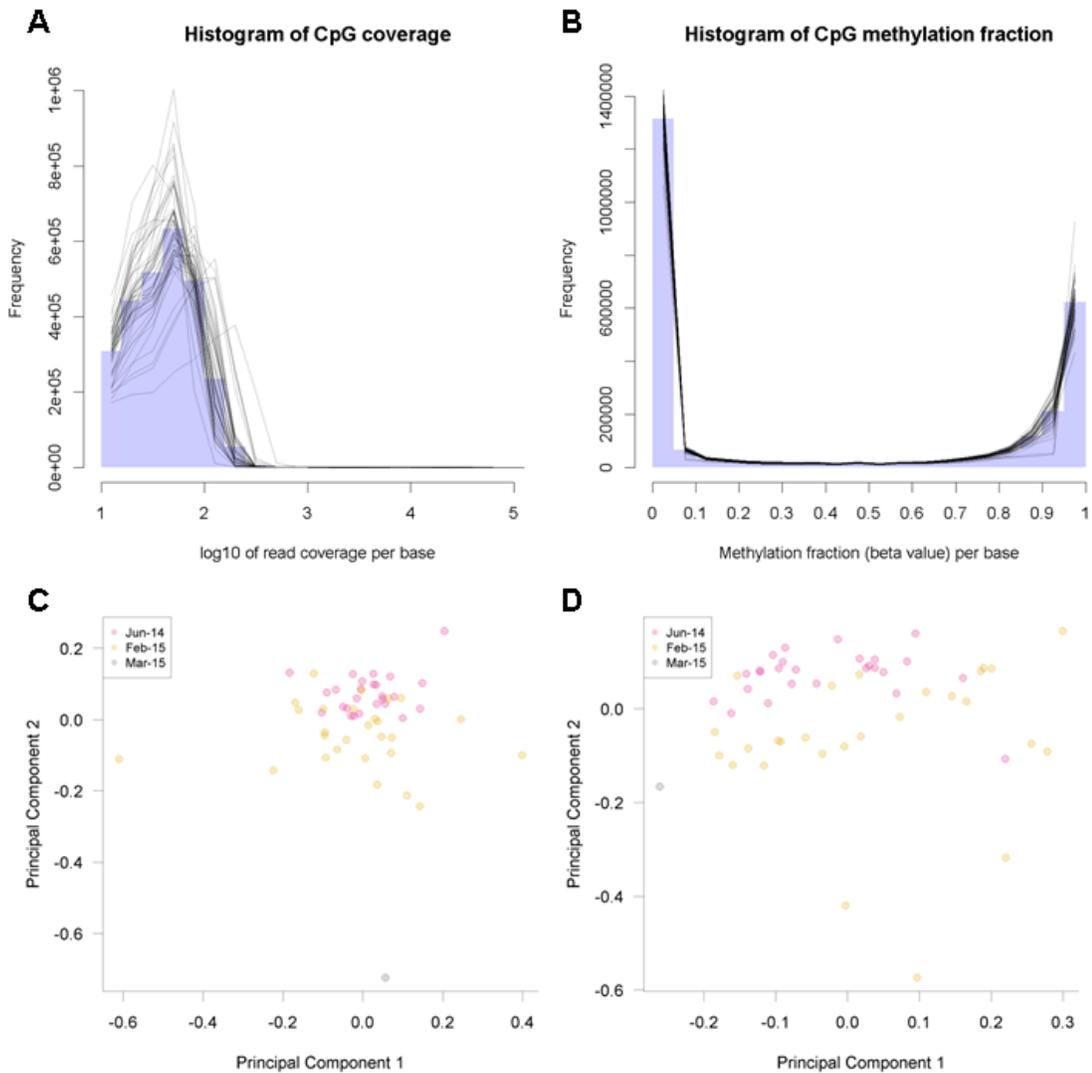


Figure B.2: **A.** Histogram of $\log_{10}(\text{coverage})$ at all CpG sites. The blue histogram represents the mean of all values in that bin over all samples. Each black line represents the counts from 1 of the 54 individual samples. **B.** Histogram of methylation -values at all CpG sites. The blue histogram represents the mean of all values in that bin over all samples. Each black line represents the counts from 1 of the 54 individual samples. **C-D.** The first two principal components of autosomal gene methylation for both (C) the CpG site and (D) gene-mapped analyses, as calculated by singular value decomposition. Samples are colored by batch: pink samples were collected in June 2014, orange samples were collected in February 2015, and gray samples were collected in March 2015.

Supplementary Figure 3

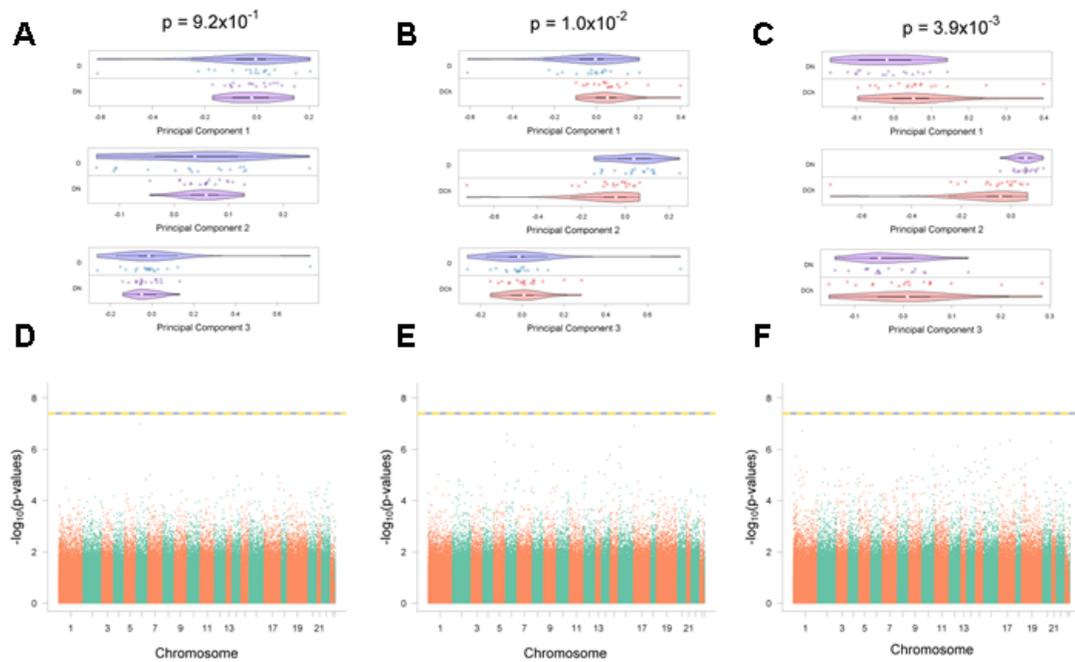


Figure B.3: Differential Methylation for all two-way comparisons of the three groups using the CpG site approach. First row (**A**, **B**, **C**): whole-methylome signal as captured by the first three principal components (horizontal axes) of the displayed subset of patients. Violin plots representing the distribution of a particular patient group along a principal component are adjacent to their corresponding group. Second row (**D**, **E**, **F**): chromosomal distribution of gene methylation differences. For each gene, the significance is displayed on the y-axis as the log₁₀ of the p-value. The results are ordered along the x-axis by chromosome, with each bar representing a different chromosome. Bonferroni and Benjamini-Hochberg p-value thresholds are displayed as blue and yellow (dashed) lines, respectively.

Supplementary Figure 4

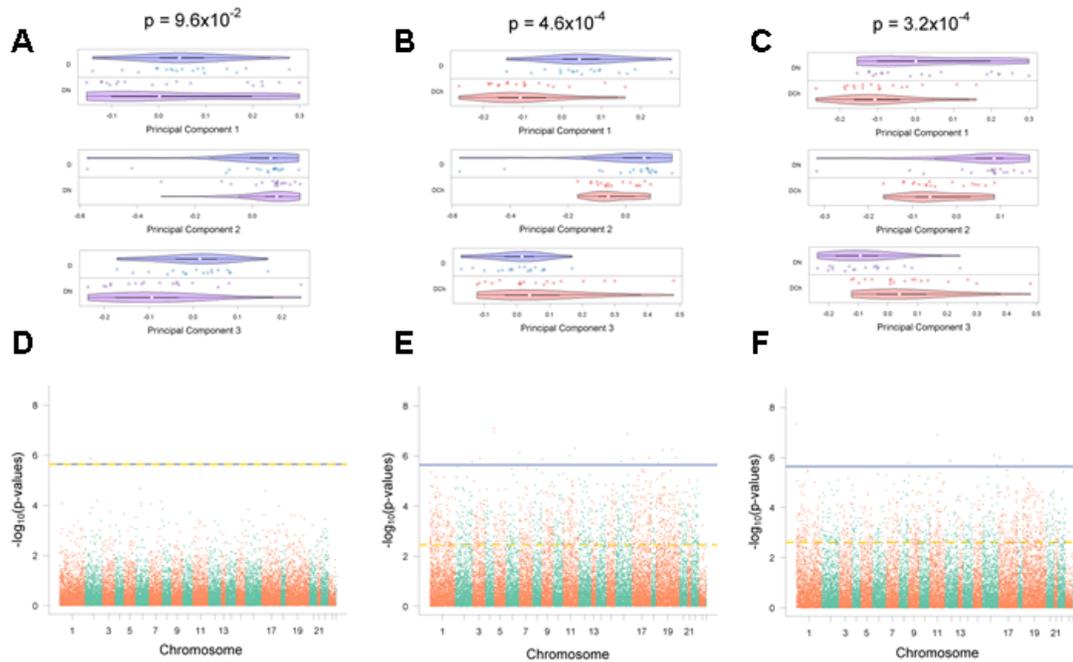


Figure B.4: Differential Methylation for all two-way comparisons of the three groups using the gene-mapped approach. First row (**A**, **B**, **C**): whole-methylome signal as captured by the first three principal components (horizontal axes) of the displayed subset of patients. Violin plots representing the distribution of a particular patient group along a principal component are adjacent to their corresponding group. Second row (**D**, **E**, **F**): chromosomal distribution of gene methylation differences. For each gene, the significance is displayed on the y-axis as the log₁₀ of the p-value. The results are ordered along the x-axis by chromosome, with each bar representing a different chromosome. Bonferroni and Benjamini-Hochberg p-value thresholds are displayed as blue and yellow (dashed) lines, respectively.

Supplementary Figure 5

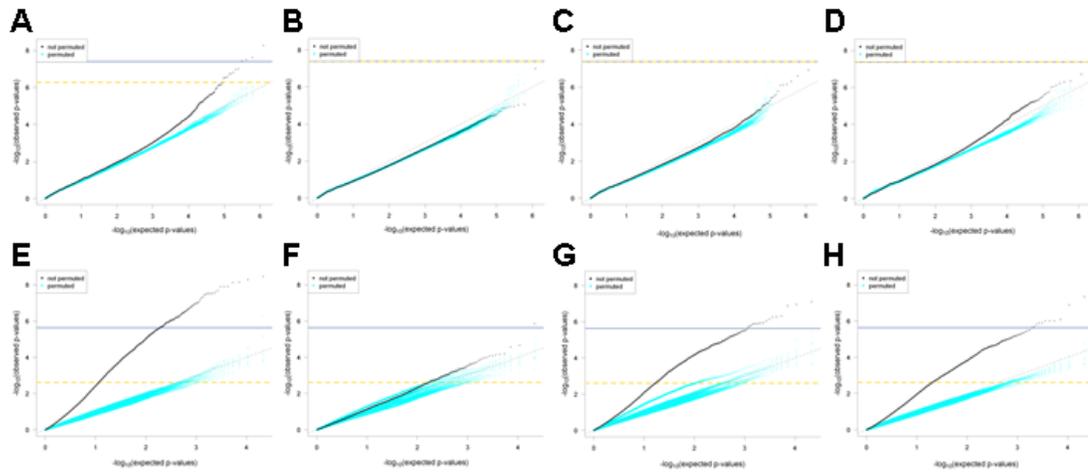


Figure B.5: Quantile-quantile plots for all two-way comparisons of the three groups using both CpG site (**A, B, C, D**) and gene-mapped (**E, F, G, H**) differential methylation in patients with diabetes and CF compared to patients with diabetes but no CF. Bonferroni and Benjamini-Hochberg p-value thresholds are displayed as blue and yellow (dashed) lines, respectively. Differential methylation on permuted data was calculated 10 times for each QQ plot, resulting in 10 sets of null p-values that are plotted in cyan alongside the non-permuted p-values.

BIBLIOGRAPHY

- [1] AEBERSOLD, R., AND MANN, M. Mass spectrometry-based proteomics. *Nature* 422, 6928 (2003), 198.
- [2] AGRAFIOTIS, D. K., LOBANOV, V. S., AND SALEMME, F. R. Combinatorial informatics in the post-genomics era. *Nature Reviews Drug Discovery* 1, 5 (2002), 337.
- [3] AKALIN, A., GARRETT-BAKELMAN, F. E., KORMAKSSON, M., BUSUTTIL, J., ZHANG, L., KHREBTUKOVA, I., MILNE, T. A., HUANG, Y., BISWAS, D., HESS, J. L., ET AL. Base-pair resolution dna methylation sequencing reveals profoundly divergent epigenetic landscapes in acute myeloid leukemia. *PLoS genetics* 8, 6 (2012), e1002781.
- [4] AKALIN, A., KORMAKSSON, M., LI, S., GARRETT-BAKELMAN, F. E., FIGUEROA, M. E., MELNICK, A., AND MASON, C. E. methylkit: a comprehensive r package for the analysis of genome-wide dna methylation profiles. *Genome biology* 13, 10 (2012), R87.
- [5] ALTSHULER, D., DALY, M. J., AND LANDER, E. S. Genetic mapping in human disease. *science* 322, 5903 (2008), 881–888.
- [6] ANDERS, S., AND HUBER, W. Differential expression analysis for sequence count data. *Genome biology* 11, 10 (2010), R106.
- [7] AVERY, L., AND WASSERMAN, S. Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends in genetics* 8, 9 (1992), 312–316.
- [8] BAKER, M. The 'omes puzzle. *Nature* 494, 7438 (2013), 416.
- [9] BATESON, W. Mendels principles of heredity. *Molecular and General Genetics MGG* 3, 1 (1910), 108–109.
- [10] BATTLE, A., MOSTAFAVI, S., ZHU, X., POTASH, J. B., WEISSMAN, M. M., MCCORMICK, C., HAUDENSCHILD, C. D., BECKMAN, K. B., SHI, J., MEI, R., ET AL. Characterizing the genetic basis of transcriptome diversity through rna-sequencing of 922 individuals. *Genome research* (2013).
- [11] BAYLIN, S. B. Dna methylation and gene silencing in cancer. *Nature Reviews Clinical Oncology* 2, S1 (2005), S4.

- [12] BELL, J. T., PAI, A. A., PICKRELL, J. K., GAFFNEY, D. J., PIQUE-REGI, R., DEGNER, J. F., GILAD, Y., AND PRITCHARD, J. K. Dna methylation patterns associate with genetic and gene expression variation in hapmap cell lines. *Genome biology* 12, 1 (2011), R10.
- [13] BEYER, M. A., AND LANEY, D. The importance of big data: a definition. *Stamford, CT: Gartner* (2012), 2014–2018.
- [14] BIRD, A. Dna methylation patterns and epigenetic memory. *Genes & development* 16, 1 (2002), 6–21.
- [15] BIRD, A. Perceptions of epigenetics. *Nature* 447, 7143 (2007), 396.
- [16] BOWER, J. M., AND BOLOURI, H. *Computational modeling of genetic and biochemical networks*. MIT press, 2004.
- [17] BREEN, M. S., KEMENA, C., VLASOV, P. K., NOTREDAME, C., AND KONDRASHOV, F. A. Epistasis as the primary factor in molecular evolution. *Nature* 490, 7421 (2012), 535.
- [18] BREM, R. B., STOREY, J. D., WHITTLE, J., AND KRUGLYAK, L. Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* 436, 7051 (2005), 701.
- [19] BUERMANS, H. P., ARIYUREK, Y., VAN OMMEN, G., DEN DUNNEN, J. T., AND AC'T HOEN, P. New methods for next generation sequencing based microrna expression profiling. *BMC genomics* 11, 1 (2010), 716.
- [20] BULLOCK, J. M., MEDWAY, C., CORTINA-BORJA, M., TURTON, J. C., PRINCE, J. A., IBRAHIM-VERBAAS, C. A., SCHUUR, M., BRETELER, M. M., VAN DUIJN, C. M., KEHOE, P. G., ET AL. Discovery by the epistasis project of an epistatic interaction between the *gstm3* gene and the *hhex/ide/kif11* locus in the risk of alzheimer's disease. *Neurobiology of aging* 34, 4 (2013), 1309–e1.
- [21] CAMPAGNE, F., DORFF, K. C., CHAMBWE, N., ROBINSON, J. T., AND MESIROV, J. P. Compression of structured high-throughput sequencing data. *PLoS one* 8, 11 (2013), e79871.
- [22] CHANG, C. C., CHOW, C. C., TELLIER, L. C., VATTIKUTI, S., PURCELL, S. M., AND LEE, J. J. Second-generation plink: rising to the challenge of larger and richer datasets. *Gigascience* 4, 1 (2015), 7.

- [23] CHEN, J., AND CHEN, Z. Extended bic for small-n-large-p sparse glm. *Statistica Sinica* (2012), 555–574.
- [24] CHIEFARI, E., NEVOLO, M. T., ARCIDIACONO, B., MAURIZIO, E., NOCERA, A., IIRITANO, S., SGARRA, R., POSSIDENTE, K., PALMIERI, C., PAONESSA, F., ET AL. Hmgal1 is a novel downstream nuclear target of the insulin receptor signaling pathway. *Scientific reports* 2 (2012), 251.
- [25] CHO, J. H., NICOLAE, D. L., GOLD, L. H., FIELDS, C. T., LABUDA, M. C., ROHAL, P. M., PICKLES, M. R., QIN, L., FU, Y., MANN, J. S., ET AL. Identification of novel susceptibility loci for inflammatory bowel disease on chromosomes 1p, 3q, and 4q: evidence for epistasis between 1p and ibd1. *Proceedings of the National Academy of Sciences* 95, 13 (1998), 7502–7507.
- [26] CHOY, E., YELENSKY, R., BONAKDAR, S., PLENGE, R. M., SAXENA, R., DE JAGER, P. L., SHAW, S. Y., WOLFISH, C. S., SLAVIK, J. M., COTSAPAS, C., ET AL. Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS genetics* 4, 11 (2008), e1000287.
- [27] CIVELEK, M., AND LUSIS, A. J. Systems genetics approaches to understand complex traits. *Nature Reviews Genetics* 15, 1 (2014), 34.
- [28] COKUS, S. J., FENG, S., ZHANG, X., CHEN, Z., MERRIMAN, B., HAUDENSCHILD, C. D., PRADHAN, S., NELSON, S. F., PELLEGRINI, M., AND JACOBSEN, S. E. Shotgun bisulphite sequencing of the arabidopsis genome reveals dna methylation patterning. *Nature* 452, 7184 (2008), 215.
- [29] CONSORTIUM, I. H. G. S., ET AL. Finishing the euchromatic sequence of the human genome. *Nature* 431, 7011 (2004), 931.
- [30] COOKSON, W., LIANG, L., ABECASIS, G., MOFFATT, M., AND LATHROP, M. Mapping complex disease traits with global gene expression. *Nature Reviews Genetics* 10, 3 (2009), 184.
- [31] CORDELL, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human molecular genetics* 11, 20 (2002), 2463–2468.
- [32] CORDELL, H. J., TODD, J. A., HILL, N. J., LORD, C. J., LYONS, P. A., PETERSON, L. B., WICKER, L. S., AND CLAYTON, D. G. Statistical modeling

of interlocus interactions in a complex disease: rejection of the multiplicative model of epistasis in type 1 diabetes. *Genetics* 158, 1 (2001), 357–367.

- [33] COX, N. J., FRIGGE, M., NICOLAE, D. L., CONCANNON, P., HANIS, C. L., BELL, G. I., AND KONG, A. Loci on chromosomes 2 (niddm1) and 15 interact to increase susceptibility to diabetes in mexican americans. *Nature genetics* 21, 2 (1999), 213.
- [34] CROTEAU-CHONKA, D. C., ROGERS, A. J., RAJ, T., MCGEACHIE, M. J., QIU, W., ZINITI, J. P., STUBBS, B. J., LIANG, L., MARTINEZ, F. D., STRUNK, R. C., ET AL. Expression quantitative trait loci information improves predictive modeling of disease relevance of non-coding genetic variation. *PloS one* 10, 10 (2015), e0140758.
- [35] CULVERHOUSE, R., SUAREZ, B. K., LIN, J., AND REICH, T. A perspective on epistasis: limits of models displaying no main effect. *The American Journal of Human Genetics* 70, 2 (2002), 461–471.
- [36] DEATON, A. M., AND BIRD, A. CpG islands and the regulation of transcription. *Genes & development* 25, 10 (2011), 1010–1022.
- [37] DEGNER, J. F., PAI, A. A., PIQUE-REGI, R., VEYRIERAS, J.-B., GAFFNEY, D. J., PICKRELL, J. K., DE LEON, S., MICHELINI, K., LEWELLEN, N., CRAWFORD, G. E., ET AL. Dnase i sensitivity qtls are a major determinant of human expression variation. *Nature* 482, 7385 (2012), 390.
- [38] DETTMER, K., ARONOV, P. A., AND HAMMOCK, B. D. Mass spectrometry-based metabolomics. *Mass spectrometry reviews* 26, 1 (2007), 51–78.
- [39] DICK, K. J., NELSON, C. P., TSAPROUNI, L., SANDLING, J. K., AÏSSI, D., WAHL, S., MEDURI, E., MORANGE, P.-E., GAGNON, F., GRALLERT, H., ET AL. Dna methylation and body-mass index: a genome-wide analysis. *The Lancet* 383, 9933 (2014), 1990–1998.
- [40] DRMANAC, R. The advent of personal genome sequencing. *Genetics in Medicine* 13, 3 (2011), 188.
- [41] DRUKA, A., POTOKINA, E., LUO, Z., JIANG, N., CHEN, X., KEARSEY, M., AND WAUGH, R. Expression quantitative trait loci analysis in plants. *Plant biotechnology journal* 8, 1 (2010), 10–27.

- [42] EADS, C. A., DANENBERG, K. D., KAWAKAMI, K., SALTZ, L. B., BLAKE, C., SHIBATA, D., DANENBERG, P. V., AND LAIRD, P. W. MethyLight: a high-throughput assay to measure dna methylation. *Nucleic acids research* 28, 8 (2000), e32–00.
- [43] EBBERT, M. T., RIDGE, P. G., AND KAUWE, J. S. Bridging the gap between statistical and biological epistasis in alzheimers disease. *BioMed research international* 2015 (2015).
- [44] EBBERT, M. T., RIDGE, P. G., WILSON, A. R., SHARP, A. R., BAILEY, M., NORTON, M. C., TSCHANZ, J. T., MUNGER, R. G., CORCORAN, C. D., AND KAUWE, J. S. Population-based analysis of alzheimers disease risk alleles implicates genetic interactions. *Biological psychiatry* 75, 9 (2014), 732–737.
- [45] EDEN, A., GAUDET, F., WAGHMARE, A., AND JAENISCH, R. Chromosomal instability and tumors promoted by dna hypomethylation. *Science* 300, 5618 (2003), 455–455.
- [46] EFRON, B. Bayesian inference and the parametric bootstrap. *The annals of applied statistics* 6, 4 (2012), 1971.
- [47] EICHLER, E. E., FLINT, J., GIBSON, G., KONG, A., LEAL, S. M., MOORE, J. H., AND NADEAU, J. H. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics* 11, 6 (2010), 446.
- [48] ELEMENTO, O., AND TAVAZOIE, S. A nonalignment approach for genome-scale discovery of dna and mrna regulatory elements using network-level conservation. *Comparative Genomics*, 349.
- [49] ELLIOTT, M. A., WALTER, G. A., SWIFT, A., VANDENBORNE, K., SCHOTLAND, J. C., AND LEIGH, J. S. Spectral quantitation by principal component analysis using complex singular value decomposition. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine* 41, 3 (1999), 450–455.
- [50] ELLIS, S. E., GUPTA, S., ASHAR, F. N., BADER, J. S., WEST, A. B., AND ARKING, D. E. Rna-seq optimization with eqtl gold standards. *BMC genomics* 14, 1 (2013), 892.
- [51] FABRIS, S., BOLLATI, V., AGNELLI, L., MORABITO, F., MOTTA, V., CUTRONA, G., MATIS, S., GRAZIA RECCHIA, A., GIGLIOTTI, V., GEN-

- TILE, M., ET AL. Biological and clinical relevance of quantitative global methylation of repetitive dna sequences in chronic lymphocytic leukemia. *Epigenetics* 6, 2 (2011), 188–194.
- [52] FENG, S., JACOBSEN, S. E., AND REIK, W. Epigenetic reprogramming in plant and animal development. *Science* 330, 6004 (2010), 622–627.
- [53] FERREIRA, M. A., AND PURCELL, S. M. A multivariate test of association. *Bioinformatics* 25, 1 (2009), 132–133.
- [54] FISH, A. E., CAPRA, J. A., AND BUSH, W. S. Are interactions between cis-regulatory variants evidence for biological epistasis or statistical artifacts? *The American Journal of Human Genetics* 99, 4 (2016), 817–830.
- [55] FISHER, R. A. Xv.the correlation between relatives on the supposition of mendelian inheritance. *Earth and Environmental Science Transactions of the Royal Society of Edinburgh* 52, 2 (1919), 399–433.
- [56] FLINT, J., AND MACKAY, T. F. Genetic architecture of quantitative traits in mice, flies, and humans. *Genome research* 19, 5 (2009), 723–733.
- [57] FORSBERG, S. K., BLOOM, J. S., SADHU, M. J., KRUGLYAK, L., AND CARLBORG, Ö. Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast. *Nature genetics* 49, 4 (2017), 497.
- [58] FRANKE, A., MCGOVERN, D. P., BARRETT, J. C., WANG, K., RADFORD-SMITH, G. L., AHMAD, T., LEES, C. W., BALSCHUN, T., LEE, J., ROBERTS, R., ET AL. Genome-wide meta-analysis increases to 71 the number of confirmed crohn’s disease susceptibility loci. *Nature genetics* 42, 12 (2010), 1118.
- [59] FREITAS, A. A. Understanding the crucial role of attribute interaction in data mining. *Artificial Intelligence Review* 16, 3 (2001), 177–199.
- [60] FRYXELL, K. J., AND MOON, W.-J. Cpg mutation rates in the human genome are highly dependent on local gc content. *Molecular Biology and Evolution* 22, 3 (2004), 650–658.
- [61] FUTSCHER, B. W., OSHIRO, M. M., WOZNIAK, R. J., HOLTAN, N., HANIGAN, C. L., DUAN, H., AND DOMANN, F. E. Role for dna methylation

- in the control of cell type-specific maspin expression. *Nature genetics* 31, 2 (2002), 175.
- [62] GAFFNEY, D. J., VEYRIERAS, J.-B., DEGNER, J. F., PIQUE-REGI, R., PAI, A. A., CRAWFORD, G. E., STEPHENS, M., GILAD, Y., AND PRITCHARD, J. K. Dissecting the regulatory architecture of gene expression qtls. *Genome biology* 13, 1 (2012), R7.
- [63] GALESLOOT, T. E., VAN STEEN, K., KIEMENEY, L. A., JANSSE, L. L., AND VERMEULEN, S. H. A comparison of multivariate genome-wide association methods. *PloS one* 9, 4 (2014), e95923.
- [64] GARRETT-BAKELMAN, F. E., SHERIDAN, C. K., KACMARCZYK, T. J., ISHII, J., BETEL, D., ALONSO, A., MASON, C. E., FIGUEROA, M. E., AND MELNICK, A. M. Enhanced reduced representation bisulfite sequencing for assessment of dna methylation at base pair resolution. *Journal of visualized experiments: JoVE*, 96 (2015).
- [65] GATTI, D. M., SHABALIN, A. A., LAM, T.-C., WRIGHT, F. A., RUSYN, I., AND NOBEL, A. B. Fastmap: fast eqtl mapping in homozygous populations. *Bioinformatics* 25, 4 (2008), 482–489.
- [66] GIBBS, J. R., VAN DER BRUG, M. P., HERNANDEZ, D. G., TRAYNOR, B. J., NALLS, M. A., LAI, S.-L., AREPALLI, S., DILLMAN, A., RAFFERTY, I. P., TRONCOSO, J., ET AL. Abundant quantitative trait loci exist for dna methylation and gene expression in human brain. *PLoS genetics* 6, 5 (2010), e1000952.
- [67] GIBSON, G. Hints of hidden heritability in gwas. *Nature genetics* 42, 7 (2010), 558.
- [68] GILAD, Y., RIFKIN, S. A., AND PRITCHARD, J. K. Revealing the architecture of gene regulation: the promise of eqtl studies. *Trends in genetics* 24, 8 (2008), 408–415.
- [69] GILBERT-DIAMOND, D., AND MOORE, J. H. Analysis of gene-gene interactions. *Current protocols in human genetics* 70, 1 (2011), 1–14.
- [70] GO, M. J., HWANG, J.-Y., KIM, Y. J., OH, J. H., KIM, Y.-J., KWAK, S. H., PARK, K. S., LEE, J., KIM, B.-J., HAN, B.-G., ET AL. New susceptibility loci in myl2, c12orf51 and oas1 associated with 1-h plasma glucose as predisposing risk factors for type 2 diabetes in the korean population. *Journal of human genetics* 58, 6 (2013), 362.

- [71] GOSS, K. L., AND GORDON, D. J. Gene expression signature based screening identifies ribonucleotide reductase as a candidate therapeutic target in ewing sarcoma. *Oncotarget* 7, 39 (2016), 63003.
- [72] GOURIEROUX, C., HOLLY, A., AND MONFORT, A. Likelihood ratio test, wald test, and kuhn-tucker test in linear models with inequality constraints on the regression parameters. *Econometrica: journal of the Econometric Society* (1982), 63–80.
- [73] GREENBAUM, D., LUSCOMBE, N. M., JANSEN, R., QIAN, J., AND GERSTEIN, M. Interrelating different types of genomic data, from proteome to secretome: 'oming in on function. *Genome research* 11, 9 (2001), 1463–1468.
- [74] GREGERSEN, J. W., KRANC, K. R., KE, X., SVENDSEN, P., MADSEN, L. S., THOMSEN, A. R., CARDON, L. R., BELL, J. I., AND FUGGER, L. Functional epistasis on a common mhc haplotype associated with multiple sclerosis. *Nature* 443, 7111 (2006), 574.
- [75] GRUNDBERG, E., SMALL, K. S., HEDMAN, Å. K., NICA, A. C., BUIL, A., KEILDSON, S., BELL, J. T., YANG, T.-P., MEDURI, E., BARRETT, A., ET AL. Mapping cis-and trans-regulatory effects across multiple tissues in twins. *Nature genetics* 44, 10 (2012), 1084–1089.
- [76] GU, H., SMITH, Z. D., BOCK, C., BOYLE, P., GNIRKE, A., AND MEISSNER, A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale dna methylation profiling. *Nature protocols* 6, 4 (2011), 468.
- [77] HAYDEN, E. C. Is the \$1,000 genome for real? *Nature News* (2014).
- [78] HE, S., OWEN, D. R., JELINSKY, S. A., AND LIN, L.-L. Lysine methyltransferase setd7 (set7/9) regulates ros signaling through mitochondria and nfe2l2/are pathway. *Scientific reports* 5 (2015), 14368.
- [79] HE, Z., HE, J., LIU, Z., XU, J., SOFIA, F. Y., LIU, H., AND YANG, J. Mapk11 in breast cancer cells enhances osteoclastogenesis and bone resorption. *Biochimie* 106 (2014), 24–32.
- [80] HEMANI, G., SHAKHBAZOV, K., WESTRA, H.-J., ESKO, T., HENDERS, A. K., MCRAE, A. F., YANG, J., GIBSON, G., MARTIN, N. G., METSPALU, A., ET AL. Detection and replication of epistasis influencing transcription in humans. *Nature* 508, 7495 (2014), 249.

- [81] HEMANI, G., THEOCHARIDIS, A., WEI, W., AND HALEY, C. Epigpu: exhaustive pairwise epistasis scans parallelized on consumer level graphics cards. *Bioinformatics* 27, 11 (2011), 1462–1465.
- [82] HEROLD, C., STEFFENS, M., BROCKSCHMIDT, F. F., BAUR, M. P., AND BECKER, T. Intersnp: genome-wide interaction analysis guided by a priori information. *Bioinformatics* 25, 24 (2009), 3275–3281.
- [83] HERPER, M. Illumina promises to sequence human genome for \$100but not quite yet. *Forbes*. Jan 29 (2017).
- [84] HIRAI, M. Y., YANO, M., GOODENOWE, D. B., KANAYA, S., KIMURA, T., AWAZUHARA, M., ARITA, M., FUJIWARA, T., AND SAITO, K. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in arabidopsis thaliana. *Proceedings of the National Academy of Sciences* 101, 27 (2004), 10205–10210.
- [85] HODGKIN, J. Seven types of pleiotropy. *International Journal of Developmental Biology* 42 (1998), 501–505.
- [86] HOLLOWAY, B., LUCK, S., BEATTY, M., RAFALSKI, J.-A., AND LI, B. Genome-wide expression quantitative trait loci (eqtl) analysis in maize. *BMC genomics* 12, 1 (2011), 336.
- [87] HOOD, L., AND ROWEN, L. The human genome project: big science transforms biology and medicine. *Genome medicine* 5, 9 (2013), 79.
- [88] HOSS, A. G., KARTHA, V. K., DONG, X., LATOURELLE, J. C., DUMITRIU, A., HADZI, T. C., MACDONALD, M. E., GUSELLA, J. F., AKBARIAN, S., CHEN, J.-F., ET AL. Micrnas located in the hox gene clusters are implicated in huntington’s disease pathogenesis. *PLoS genetics* 10, 2 (2014), e1004188.
- [89] HOSS, A. G., LABADORF, A., BEACH, T. G., LATOURELLE, J. C., AND MYERS, R. H. micrna profiles in parkinson’s disease prefrontal cortex. *Frontiers in aging neuroscience* 8 (2016), 36.
- [90] HOUGHTON, F. J., BELLINGHAM, S. A., HILL, A. F., BOURGES, D., ANG, D. K., GEMETZIS, T., GASNEREAU, I., AND GLEESON, P. A. Arl5b is a golgi-localised small g protein involved in the regulation of retrograde transport. *Experimental cell research* 318, 5 (2012), 464–477.

- [91] HOUSEMAN, E. A., ACCOMANDO, W. P., KOESTLER, D. C., CHRISTENSEN, B. C., MARSIT, C. J., NELSON, H. H., WIENCKE, J. K., AND KELSEY, K. T. Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC bioinformatics* 13, 1 (2012), 86.
- [92] HOWARD, R., CARRIQUIRY, A. L., AND BEAVIS, W. D. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3: Genes, Genomes, Genetics* (2014), g3–114.
- [93] HOWE, D., COSTANZO, M., FEY, P., GOJOBORI, T., HANNICK, L., HIDE, W., HILL, D. P., KANIA, R., SCHAEFFER, M., ST PIERRE, S., ET AL. Big data: The future of biocuration. *Nature* 455, 7209 (2008), 47.
- [94] HUANG, T., AND CAI, Y.-D. An information-theoretic machine learning approach to expression qtl analysis. *PloS one* 8, 6 (2013), e67899.
- [95] HUANG, T., LAN, L., FANG, X., AN, P., MIN, J., AND WANG, F. Promises and challenges of big data computing in health sciences. *Big Data Research* 2, 1 (2015), 2–11.
- [96] HUANG, W., RICHARDS, S., CARBONE, M. A., ZHU, D., ANHOLT, R. R., AYROLES, J. F., DUNCAN, L., JORDAN, K. W., LAWRENCE, F., MAGWIRE, M. M., ET AL. Epistasis dominates the genetic architecture of drosophila quantitative traits. *Proceedings of the National Academy of Sciences* 109, 39 (2012), 15553–15559.
- [97] HUANG, Y., WUCHTY, S., AND PRZYTYCKA, T. M. eqtl epistasis—challenges and computational approaches. *Frontiers in genetics* 4 (2013).
- [98] HUYNH, J. L., GARG, P., THIN, T. H., YOO, S., DUTTA, R., TRAPP, B. D., HAROUTUNIAN, V., ZHU, J., DONOVAN, M. J., SHARP, A. J., ET AL. Epigenome-wide differences in pathology-free regions of multiple sclerosis-affected brains. *Nature neuroscience* 17, 1 (2014), 121.
- [99] HWANG, D., RUST, A. G., RAMSEY, S., SMITH, J. J., LESLIE, D. M., WESTON, A. D., DE ATAURI, P., AITCHISON, J. D., HOOD, L., SIEGEL, A. F., ET AL. A data integration methodology for systems biology. *Proceedings of the National Academy of Sciences of the United States of America* 102, 48 (2005), 17296–17301.
- [100] INNOCENTI, F., COOPER, G. M., STANAWAY, I. B., GAMAZON, E. R., SMITH, J. D., MIRKOV, S., RAMIREZ, J., LIU, W., LIN, Y. S., MOLONEY,

- C., ET AL. Identification, replication, and functional fine-mapping of expression quantitative trait loci in primary human liver tissue. *PLoS genetics* 7, 5 (2011), e1002078.
- [101] IRIZARRY, R. A., LADD-ACOSTA, C., WEN, B., WU, Z., MONTANO, C., ONYANGO, P., CUI, H., GABO, K., RONGIONE, M., WEBSTER, M., ET AL. The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. *Nature genetics* 41, 2 (2009), 178.
- [102] JANA, S. Simulation of quantitative characters from qualitatively acting genes. *Theoretical and Applied Genetics* 41, 5 (1971), 216–226.
- [103] JARVIS, J. P., AND CHEVERUD, J. M. Mapping the epistatic network underlying murine reproductive fatpad variation. *Genetics* (2010).
- [104] JOHNSON, W. C., AND GEPTS, P. The role of epistasis in controlling seed yield and other agronomic traits in an andean × mesoamerican cross of common bean (*Phaseolus vulgaris* L.). *Euphytica* 125, 1 (2002), 69–79.
- [105] JONES, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* 13, 7 (2012), 484.
- [106] JONES, P. A., AND BAYLIN, S. B. The epigenomics of cancer. *Cell* 128, 4 (2007), 683–692.
- [107] JONES, P. A., ISSA, J.-P. J., AND BAYLIN, S. Targeting the cancer epigenome for therapy. *Nature Reviews Genetics* 17, 10 (2016), 630.
- [108] JOYCE, A. R., AND PALSSON, B. Ø. The model organism as a system: integrating ‘omics’ data sets. *Nature reviews Molecular cell biology* 7, 3 (2006), 198.
- [109] JU, J. H., SHENOY, S. A., CRYSTAL, R. G., AND MEZEY, J. G. An independent component analysis confounding factor correction framework for identifying broad impact expression quantitative trait loci. *PLoS computational biology* 13, 5 (2017), e1005537.
- [110] KAFRI, R., SPRINGER, M., AND PILPEL, Y. Genetic redundancy: new tricks for old genes. *Cell* 136, 3 (2009), 389–392.

- [111] KANG, M., ZHANG, C., CHUN, H.-W., DING, C., LIU, C., AND GAO, J. eqtl epistasis: detecting epistatic effects and inferring hierarchical relationships of genes in biological pathways. *Bioinformatics* 31, 5 (2014), 656–664.
- [112] KEMPERMANN, G., CHESLER, E. J., LU, L., WILLIAMS, R. W., AND GAGE, F. H. Natural variation and genetic covariance in adult hippocampal neurogenesis. *Proceedings of the National Academy of Sciences* 103, 3 (2006), 780–785.
- [113] KENDZIORSKI, C., CHEN, M., YUAN, M., LAN, H., AND ATTIE, A. D. Statistical methods for expression quantitative trait loci (eqtl) mapping. *Biometrics* 62, 1 (2006), 19–27.
- [114] KESAVAN, C., WERGEDAL, J. E., LAU, K.-H. W., AND MOHAN, S. Conditional disruption of igf-i gene in type 1 α collagen-expressing cells shows an essential role of igf-i in skeletal anabolic response to loading. *American Journal of Physiology-Endocrinology and Metabolism* 301, 6 (2011), E1191–E1197.
- [115] KHULAN, B., THOMPSON, R. F., YE, K., FAZZARI, M. J., SUZUKI, M., STASIEK, E., FIGUEROA, M. E., GLASS, J. L., CHEN, Q., MONTAGNA, C., ET AL. Comparative isoschizomer profiling of cytosine methylation: the help assay. *Genome research* 16, 8 (2006), 1046–1055.
- [116] KITANO, H. Computational systems biology. *Nature* 420, 6912 (2002), 206.
- [117] KLEIN, R. J., ZEISS, C., CHEW, E. Y., TSAI, J.-Y., SACKLER, R. S., HAYNES, C., HENNING, A. K., SANGIOVANNI, J. P., MANE, S. M., MAYNE, S. T., ET AL. Complement factor h polymorphism in age-related macular degeneration. *Science* 308, 5720 (2005), 385–389.
- [118] KOTLA, S., SINGH, N. K., KIRCHHOFER, D., AND RAO, G. N. Heterodimers of the transcriptional factors nfatc3 and fosb mediate tissue factor expression for 15 (s)-hydroxyeicosatetraenoic acid-induced monocyte trafficking. *Journal of Biological Chemistry* (2017), jbc-M117.
- [119] KUKURBA, K. R., PARSANA, P., BALLIU, B., SMITH, K. S., ZAPPALA, Z., KNOWLES, D. A., FAVÉ, M.-J., DAVIS, J. R., LI, X., ZHU, X., ET AL. Impact of the x chromosome and sex on regulatory variation. *Genome research* (2016), gr-197897.

- [120] LAMPARTER, D., MARBACH, D., RUEEDI, R., KUTALIK, Z., AND BERGMANN, S. Fast and rigorous computation of gene and pathway scores from snp-based summary statistics. *PLoS computational biology* 12, 1 (2016), e1004714.
- [121] LAN, X., WITT, H., KATSUMURA, K., YE, Z., WANG, Q., BRESNICK, E. H., FARNHAM, P. J., AND JIN, V. X. Integration of hi-c and chip-seq data reveals distinct types of chromatin linkages. *Nucleic acids research* 40, 16 (2012), 7690–7704.
- [122] LAPPALAINEN, T., SAMMETH, M., FRIEDLÄNDER, M. R., ACT HOEN, P., MONLONG, J., RIVAS, M. A., GONZALEZ-PORTA, M., KURBATOVA, N., GRIEBEL, T., FERREIRA, P. G., ET AL. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* 501, 7468 (2013), 506.
- [123] LAREAU, C., WHITE, B., OBERG, A., KENNEDY, R., POLAND, G., AND MCKINNEY, B. An interaction quantitative trait loci tool implicates epistatic functional variants in an apoptosis pathway in smallpox vaccine eqtl data. *Genes and immunity* 17, 4 (2016), 244.
- [124] LARSON, N. B., MCDONNELL, S., FRENCH, A. J., FOGARTY, Z., CHEVILLE, J., MIDDHA, S., RISK, S., BAHETI, S., NAIR, A. A., WANG, L., ET AL. Comprehensively evaluating cis-regulatory variation in the human prostate transcriptome by using gene-level allele-specific expression. *The American Journal of Human Genetics* 96, 6 (2015), 869–882.
- [125] LE SHU, M. B., AND YANG, X. Translating gwas findings to novel therapeutic targets for coronary artery disease. *Frontiers in cardiovascular medicine* 5 (2018).
- [126] LESLIE, C. S., ESKIN, E., COHEN, A., WESTON, J., AND NOBLE, W. S. Mismatch string kernels for discriminative protein classification. *Bioinformatics* 20, 4 (2004), 467–476.
- [127] LI, D., XIE, Z., LE PAPE, M., AND DYE, T. An evaluation of statistical methods for dna methylation microarray data analysis. *BMC bioinformatics* 16, 1 (2015), 217.
- [128] LI, E. Chromatin modification and epigenetic reprogramming in mammalian development. *Nature Reviews Genetics* 3, 9 (2002), 662.

- [129] LI, Q., SEO, J.-H., STRANGER, B., MCKENNA, A., PEER, I., LAFRAMBOISE, T., BROWN, M., TYEKUCHEVA, S., AND FREEDMAN, M. L. Integrative eqtl-based analyses reveal the biology of breast cancer risk loci. *Cell* 152, 3 (2013), 633–641.
- [130] LIAN, C. G., XU, Y., CEOL, C., WU, F., LARSON, A., DRESSER, K., XU, W., TAN, L., HU, Y., ZHAN, Q., ET AL. Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma. *Cell* 150, 6 (2012), 1135–1146.
- [131] LIANG, M., COWLEY, A. W., AND GREENE, A. S. High throughput gene expression profiling: a molecular approach to integrative physiology. *The Journal of physiology* 554, 1 (2004), 22–30.
- [132] LITTELL, R. C., AND FOLKS, J. L. Asymptotic optimality of fisher’s method of combining independent tests ii. *Journal of the American Statistical Association* 68, 341 (1973), 193–194.
- [133] LOFTON-DAY, C., MODEL, F., DEVOS, T., TETZNER, R., DISTLER, J., SCHUSTER, M., SONG, X., LESCHE, R., LIEBENBERG, V., EBERT, M., ET AL. Dna methylation biomarkers for blood-based colorectal cancer screening. *Clinical chemistry* 54, 2 (2008), 414–423.
- [134] LONSDALE, J., THOMAS, J., SALVATORE, M., PHILLIPS, R., LO, E., SHAD, S., HASZ, R., WALTERS, G., GARCIA, F., YOUNG, N., ET AL. The genotype-tissue expression (gtex) project. *Nature genetics* 45, 6 (2013), 580.
- [135] LOVE, M. I., HUBER, W., AND ANDERS, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology* 15, 12 (2014), 550.
- [136] LUECKE, G. R., WEEKS, N. T., GROTH, B. M., KRAEVA, M., MA, L., KRAMER, L. M., KOLTES, J. E., AND REECY, J. M. Fast epistasis detection in large-scale gwas for intel xeon phi clusters. In *Trust-com/BigDataSE/ISPA, 2015 IEEE* (2015), vol. 3, IEEE, pp. 228–235.
- [137] MABILLEAU, G., PETROVA, N., EDMONDS, M., AND SABOKBAR, A. Increased osteoclastic activity in acute charcots osteoarthopathy: the role of receptor activator of nuclear factor-kappab ligand. *Diabetologia* 51, 6 (2008), 1035–1040.
- [138] MAHER, B. Personal genomes: The case of the missing heritability. *Nature News* 456, 7218 (2008), 18–21.

- [139] MAHLKNECHT, U., WILL, J., VARIN, A., HOELZER, D., AND HERBEIN, G. Histone deacetylase 3, a class i histone deacetylase, suppresses mapk11-mediated activating transcription factor-2 activation and represses tnf gene expression. *The Journal of Immunology* 173, 6 (2004), 3979–3990.
- [140] MANOLIO, T. A., COLLINS, F. S., COX, N. J., GOLDSTEIN, D. B., HINDORFF, L. A., HUNTER, D. J., MCCARTHY, M. I., RAMOS, E. M., CARDON, L. R., CHAKRAVARTI, A., ET AL. Finding the missing heritability of complex diseases. *Nature* 461, 7265 (2009), 747–753.
- [141] MARDIS, E. R. The impact of next-generation sequencing technology on genetics. *Trends in genetics* 24, 3 (2008), 133–141.
- [142] MARGOLIN, A. A., NEMENMAN, I., BASSO, K., WIGGINS, C., STOLOVITZKY, G., DALLA FAVERA, R., AND CALIFANO, A. Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. In *BMC bioinformatics* (2006), vol. 7, BioMed Central, p. S7.
- [143] MAURANO, M. T., HUMBERT, R., RYNES, E., THURMAN, R. E., HAUGEN, E., WANG, H., REYNOLDS, A. P., SANDSTROM, R., QU, H., BRODY, J., ET AL. Systematic localization of common disease-associated variation in regulatory dna. *Science* (2012), 1222794.
- [144] MCCALL, M. N., KIM, M.-S., ADIL, M., PATIL, A. H., LU, Y., MITCHELL, C. J., LEAL-ROJAS, P., XU, J., KUMAR, M., DAWSON, V. L., ET AL. Toward the human cellular micrornaome. *Genome research* (2017).
- [145] MCCANDLISH, D. M., RAJON, E., SHAH, P., DING, Y., AND PLOTKIN, J. B. The role of epistasis in protein evolution. *Nature* 497, 7451 (2013), E1.
- [146] MCCARTHY, J. J., MCLEOD, H. L., AND GINSBURG, G. S. Genomic medicine: a decade of successes, challenges, and opportunities. *Science translational medicine* 5, 189 (2013), 189sr4–189sr4.
- [147] MCGREGOR, K., BERNATSKY, S., COLMEGNA, I., HUDSON, M., PASTINEN, T., LABBE, A., AND GREENWOOD, C. M. An evaluation of methods correcting for cell-type heterogeneity in dna methylation studies. *Genome biology* 17, 1 (2016), 84.
- [148] MCKINNEY, B., AND PAJEWSKI, N. Six degrees of epistasis: statistical network models for gwas. *Frontiers in genetics* 2 (2012), 109.

- [149] MCKINNEY, B. A., CROWE JR, J. E., GUO, J., AND TIAN, D. Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLoS genetics* 5, 3 (2009), e1000432.
- [150] MEISSNER, A., MIKKELSEN, T. S., GU, H., WERNIG, M., HANNA, J., SIVACHENKO, A., ZHANG, X., BERNSTEIN, B. E., NUSBAUM, C., JAFFE, D. B., ET AL. Genome-scale dna methylation maps of pluripotent and differentiated cells. *Nature* 454, 7205 (2008), 766.
- [151] METZKER, M. L. Sequencing technologies the next generation. *Nature reviews genetics* 11, 1 (2010), 31.
- [152] MIAO, F., WU, X., ZHANG, L., RIGGS, A. D., AND NATARAJAN, R. Histone methylation patterns are cell-type specific in human monocytes and lymphocytes and well maintained at core genes. *The Journal of Immunology* 180, 4 (2008), 2264–2269.
- [153] MICHAELSON, J. J., LOGUERCIO, S., AND BEYER, A. Detection and interpretation of expression quantitative trait loci (eqtl). *Methods* 48, 3 (2009), 265–276.
- [154] MIN, J. L., TAYLOR, J. M., RICHARDS, J. B., WATTS, T., PETTERSSON, F. H., BROXHOLME, J., AHMADI, K. R., SURDULESCU, G. L., LOWY, E., GIEGER, C., ET AL. The use of genome-wide eqtl associations in lymphoblastoid cell lines to identify novel genetic pathways involved in complex traits. *PloS one* 6, 7 (2011), e22070.
- [155] MOAYYERI, A., HAMMOND, C. J., HART, D. J., AND SPECTOR, T. D. The uk adult twin registry (twinsuk resource). *Twin Research and Human Genetics* 16, 1 (2013), 144–149.
- [156] MONTGOMERY, S. B., SAMMETH, M., GUTIERREZ-ARCELUS, M., LACH, R. P., INGLE, C., NISBETT, J., GUIGO, R., AND DERMITZAKIS, E. T. Transcriptome genetics using second generation sequencing in a caucasian population. *Nature* 464, 7289 (2010).
- [157] MOORE, J. H. Analysis of gene-gene interactions. *Current protocols in human genetics* 39, 1 (2003), 1–14.
- [158] MOORE, J. H. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human heredity* 56, 1-3 (2003), 73–82.

- [159] MOORE, J. H. A global view of epistasis. *Nature genetics* 37, 1 (2005), 13.
- [160] MOORE, J. H., AND WILLIAMS, S. M. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays* 27, 6 (2005), 637–646.
- [161] MORENO-GONZALEZ, J. Genetic models to estimate additive and non-additive effects of marker-associated qtl using multiple regression techniques. *Theoretical and Applied Genetics* 85, 4 (1992), 435–444.
- [162] MUIRHEAD, R. J. *Aspects of multivariate statistical theory*, vol. 197. John Wiley & Sons, 2009.
- [163] MÜLLEDER, M., CALVANI, E., ALAM, M. T., WANG, R. K., ECKERSTORFER, F., ZELEZNIAK, A., AND RALSER, M. Functional metabolomics describes the yeast biosynthetic regulome. *Cell* 167, 2 (2016), 553–565.
- [164] MUÑOZ, P. R., RESENDE, M. F., GEZAN, S. A., RESENDE, M. D. V., DE LOS CAMPOS, G., KIRST, M., HUBER, D., AND PETER, G. F. Unraveling additive from non-additive effects using genomic relationship matrices. *Genetics* (2014), genetics–114.
- [165] MURRELL, A., HEESON, S., AND REIK, W. Interaction between differentially methylated regions partitions the imprinted genes *igf2* and *h19* into parent-specific chromatin loops. *Nature genetics* 36, 8 (2004), 889.
- [166] MYERS, A. J., GIBBS, J. R., WEBSTER, J. A., ROHRER, K., ZHAO, A., MARLOWE, L., KALEEM, M., LEUNG, D., BRYDEN, L., NATH, P., ET AL. A survey of genetic human cortical gene expression. *Nature genetics* 39, 12 (2007), 1494.
- [167] NAWY, T. Single-cell sequencing. *Nature methods* 11, 1 (2013), 18.
- [168] NICA, A. C., PARTS, L., GLASS, D., NISBET, J., BARRETT, A., SEKOWSKA, M., TRAVERS, M., POTTER, S., GRUNDBERG, E., SMALL, K., ET AL. The architecture of gene regulatory variation across multiple human tissues: the mother study. *PLoS genetics* 7, 2 (2011), e1002003.
- [169] NICOLAE, D. L., GAMAZON, E., ZHANG, W., DUAN, S., DOLAN, M. E., AND COX, N. J. Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS genetics* 6, 4 (2010), e1000888.

- [170] NILSSON, E., MATTE, A., PERFILYEV, A., DE MELLO, V. D., KÄKELÄ, P., PIHLAJAMÄKI, J., AND LING, C. Epigenetic alterations in human liver from subjects with type 2 diabetes in parallel with reduced folate levels. *The Journal of Clinical Endocrinology & Metabolism* 100, 11 (2015), E1491–E1501.
- [171] OTTO, J. M., GIZER, I. R., DEAK, J. D., FLEMING, K. A., AND BARTHOLOW, B. D. A cis-eqtl in oprm 1 is associated with subjective response to alcohol and alcohol use. *Alcoholism: Clinical and Experimental Research* 41, 5 (2017), 929–938.
- [172] PAABY, A. B., AND ROCKMAN, M. V. The many faces of pleiotropy. *Trends in Genetics* 29, 2 (2013), 66–73.
- [173] PARK, P. J. Chip-seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics* 10, 10 (2009), 669.
- [174] PASQUIER, J., HOARAU-VÉCHOT, J., FAKHRO, K., RAFII, A., AND KHALIL, C. A. Epigenetics and cardiovascular disease in diabetes. *Current diabetes reports* 15, 12 (2015), 108.
- [175] PASQUIER, J., THOMAS, B., HOARAU-VÉCHOT, J., ODEH, T., ROBAY, A., CHIDIAC, O., DARGHAM, S. R., TURJOMAN, R., HALAMA, A., FAKHRO, K., ET AL. Circulating microparticles in acute diabetic charcot foot exhibit a high content of inflammatory cytokines, and support monocyte-to-osteoclast cell induction. *Scientific reports* 7, 1 (2017), 16450.
- [176] PETRETTO, E., MANGION, J., DICKENS, N. J., COOK, S. A., KUMARAN, M. K., LU, H., FISCHER, J., MAATZ, H., KREN, V., PRAVENEK, M., ET AL. Heritability and tissue specificity of expression quantitative trait loci. *PLoS genetics* 2, 10 (2006), e172.
- [177] PEZAWAS, L., MEYER-LINDENBERG, A., GOLDMAN, A., VERCHINSKI, B., CHEN, G., KOLACHANA, B., EGAN, M., MATTAY, V., HARIRI, A., AND WEINBERGER, D. Evidence of biologic epistasis between bdnf and slc6a4 and implications for depression. *Molecular psychiatry* 13, 7 (2008), 709.
- [178] PHILLIPS, P. C. Epistasis: the essential role of gene interactions in the structure and evolution of genetic systems. *Nature reviews. Genetics* 9, 11 (2008), 855.

- [179] PICKETT, F. B., AND MEEKS-WAGNER, D. R. Seeing double: appreciating genetic redundancy. *The Plant Cell* 7, 9 (1995), 1347.
- [180] PICKRELL, J. K., MARIONI, J. C., PAI, A. A., DEGNER, J. F., ENGELHARDT, B. E., NKADORI, E., VEYRIERAS, J.-B., STEPHENS, M., GILAD, Y., AND PRITCHARD, J. K. Understanding mechanisms underlying human gene expression variation with rna sequencing. *Nature* 464, 7289 (2010), 768.
- [181] PIRIYAPONGSA, J., NGAMPHIW, C., INTARAPANICH, A., KULAWONGANUNCHAI, S., ASSAWAMAKIN, A., BOOTCHAI, C., SHAW, P. J., AND TONGSIMA, S. iloci: a snp interaction prioritization technique for detecting epistasis in genome-wide association studies. In *BMC genomics* (2012), vol. 13, BioMed Central, p. S2.
- [182] PORTELA, A., AND ESTELLER, M. Epigenetic modifications and human disease. *Nature biotechnology* 28, 10 (2010), 1057.
- [183] PRICE, A. L., HELGASON, A., THORLEIFSSON, G., MCCARROLL, S. A., KONG, A., AND STEFANSSON, K. Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS genetics* 7, 2 (2011), e1001317.
- [184] PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I., DALY, M. J., ET AL. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* 81, 3 (2007), 559–575.
- [185] QI, J., ASL, H. F., BJÖRKEGREN, J., AND MICHOEL, T. krux: matrix-based non-parametric eqtl discovery. *BMC bioinformatics* 15, 1 (2014), 11.
- [186] QU, H., AND POLYCHRONAKOS, C. Reassessment of the type i diabetes association of the oas1 locus. *Genes and immunity* 10, S1 (2009), S69.
- [187] RAKYAN, V. K., BEYAN, H., DOWN, T. A., HAWA, M. I., MASLAU, S., ADEN, D., DAUNAY, A., BUSATO, F., MEIN, C. A., MANFRAS, B., ET AL. Identification of type 1 diabetes-associated dna methylation variable positions that precede disease diagnosis. *PLoS genetics* 7, 9 (2011), e1002300.
- [188] RAKYAN, V. K., DOWN, T. A., BALDING, D. J., AND BECK, S. Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics* 12, 8 (2011), 529.

- [189] RAKYAN, V. K., HILDMANN, T., NOVIK, K. L., LEWIN, J., TOST, J., COX, A. V., ANDREWS, T. D., HOWE, K. L., OTTO, T., OLEK, A., ET AL. Dna methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS biology* 2, 12 (2004), e405.
- [190] RAMASAMY, A., TRABZUNI, D., GUELFY, S., VARGHESE, V., SMITH, C., WALKER, R., DE, T., HARDY, J., RYTEN, M., WEALE, M. E., ET AL. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature neuroscience* 17, 10 (2014), 1418.
- [191] REIK, W., AND WALTER, J. Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics* 2, 1 (2001), 21.
- [192] RHODES, D. R., YU, J., SHANKER, K., DESHPANDE, N., VARAMBALLY, R., GHOSH, D., BARRETTE, T., PANDER, A., AND CHINNAIYAN, A. M. Oncomine: a cancer microarray database and integrated data-mining platform. *Neoplasia* 6, 1 (2004), 1–6.
- [193] RISCH, N. Linkage strategies for genetically complex traits. i. multilocus models. *American journal of human genetics* 46, 2 (1990), 222.
- [194] RITCHIE, M. D. Finding the epistasis needles in the genome-wide haystack. In *Epistasis*. Springer, 2015, pp. 19–33.
- [195] RITCHIE, M. D., HAHN, L. W., ROODI, N., BAILEY, L. R., DUPONT, W. D., PARL, F. F., AND MOORE, J. H. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics* 69, 1 (2001), 138–147.
- [196] ROBERTSON, K. D. Dna methylation and human disease. *Nature Reviews Genetics* 6, 8 (2005), 597.
- [197] RODRIGUEZ-FLORES, J. L., FAKHRO, K., AGOSTO-PEREZ, F., RAMSTETTER, M. D., ARBIZA, L., VINCENT, T. L., ROBAY, A., MALEK, J. A., SUHRE, K., CHOUCANE, L., ET AL. Indigenous arabs are descendants of the earliest split from ancient eurasian populations. *Genome research* (2016).
- [198] ROGERS, L. C., FRYKBERG, R. G., ARMSTRONG, D. G., BOULTON, A. J., EDMONDS, M., VAN, G. H., HARTEMANN, A., GAME, F., JEFFCOATE,

- W., JIRKOVSKA, A., ET AL. The charcot foot in diabetes. *Diabetes care* 34, 9 (2011), 2123–2129.
- [199] SARIG, O., NAHUM, S., RAPAPORT, D., ISHIDA-YAMAMOTO, A., FUCHS-TELEM, D., QIAOLI, L., COHEN-KATSENELSON, K., SPIEGEL, R., NOUSBECK, J., ISRAELI, S., ET AL. Short stature, onychodysplasia, facial dysmorphism, and hypotrichosis syndrome is caused by a *poc1a* mutation. *The American Journal of Human Genetics* 91, 2 (2012), 337–342.
- [200] SCHENA, M., SHALON, D., HELLER, R., CHAI, A., BROWN, P. O., AND DAVIS, R. W. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proceedings of the National Academy of Sciences* 93, 20 (1996), 10614–10619.
- [201] SCHOVILLE, S. D., CHEN, Y. H., ANDERSSON, M. N., BENOIT, J. B., BHANDARI, A., BOWSER, J. H., BREVIK, K., CAPPELLE, K., CHEN, M.-J. M., CHILDERS, A. K., ET AL. A model species for agricultural pest genomics: the genome of the colorado potato beetle, *leptinotarsa decemlineata* (coleoptera: Chrysomelidae). *Scientific reports* 8, 1 (2018), 1931.
- [202] SCHULDENFREI, A., BELTON, A., KOWALSKI, J., TALBOT, C. C., DI CELLO, F., POH, W., TSAI, H.-L., SHAH, S. N., HUSO, T. H., HUSO, D. L., ET AL. *Hmga1* drives stem cell, inflammatory pathway, and cell cycle progression genes during lymphoid tumorigenesis. *BMC genomics* 12, 1 (2011), 549.
- [203] SCHÜPBACH, T., XENARIOS, I., BERGMANN, S., AND KAPUR, K. Fastepistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics* 26, 11 (2010), 1468–1469.
- [204] SEGRE, D., DELUNA, A., CHURCH, G. M., AND KISHONY, R. Modular epistasis in yeast metabolism. *Nature genetics* 37, 1 (2005), 77.
- [205] SHABALIN, A. A. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics* 28, 10 (2012), 1353–1358.
- [206] SHAHID, S. U., SHABANA, N., REHMAN, A., AND HUMPHRIES, S. Gwas implicated risk variants in different genes contribute additively to increase the risk of coronary artery disease (cad) in the pakistani subjects. *Lipids in health and disease* 17, 1 (2018), 89.
- [207] SHAO, H., BURRAGE, L. C., SINASAC, D. S., HILL, A. E., ERNEST, S. R., O'BRIEN, W., COURTLAND, H.-W., JEPSEN, K. J., KIRBY, A., KULBOKAS,

- E., ET AL. Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. *Proceedings of the National Academy of Sciences* (2008), pnas-0810388105.
- [208] SHIMIZU, T. S., TAKAHASHI, K., AND TOMITA, M. CpG distribution patterns in methylated and non-methylated species. *Gene* 205, 1 (1997), 103–107.
- [209] SLATKIN, M. Epigenetic inheritance and the missing heritability problem. *Genetics* 182, 3 (2009), 845–850.
- [210] SMITH, Z. D., AND MEISSNER, A. Dna methylation: roles in mammalian development. *Nature Reviews Genetics* 14, 3 (2013), 204.
- [211] SORRELLS, T. R., BOOTH, L. N., TUCH, B. B., AND JOHNSON, A. D. Intersecting transcription networks constrain gene regulatory evolution. *Nature* 523, 7560 (2015), 361.
- [212] SPEAKMAN, J., LOOS, R., ORAHILLY, S., HIRSCHHORN, J., AND ALLISON, D. Gwas for bmi: a treasure trove of fundamental insights into the genetic basis of obesity. *International Journal of Obesity* (2018), 1.
- [213] STEPHENS, Z. D., LEE, S. Y., FAGHRI, F., CAMPBELL, R. H., ZHAI, C., EFRON, M. J., IYER, R., SCHATZ, M. C., SINHA, S., AND ROBINSON, G. E. Big data: astronomical or genomical? *PLoS biology* 13, 7 (2015), e1002195.
- [214] STRANGER, B. E., MONTGOMERY, S. B., DIMAS, A. S., PARTS, L., STEGLE, O., INGLE, C. E., SEKOWSKA, M., SMITH, G. D., EVANS, D., GUTIERREZ-ARCELUS, M., ET AL. Patterns of cis regulatory variation in diverse human populations. *PLoS genetics* 8, 4 (2012), e1002639.
- [215] SUN, W. A statistical framework for eqtl mapping using rna-seq data. *Biometrics* 68, 1 (2012), 1–11.
- [216] SUVÀ, M. L., RIGGI, N., AND BERNSTEIN, B. E. Epigenetic reprogramming in cancer. *Science* 339, 6127 (2013), 1567–1570.
- [217] SZKLARCZYK, D., MORRIS, J. H., COOK, H., KUHN, M., WYDER, S., SIMONOVIC, M., SANTOS, A., DONCHEVA, N. T., ROTH, A., BORK, P., ET AL. The string database in 2017: quality-controlled protein–protein as-

sociation networks, made broadly accessible. *Nucleic acids research* (2016), gkw937.

- [218] TANG, W., WU, X., JIANG, R., AND LI, Y. Epistatic module detection for case-control studies: a bayesian model with a gibbs sampling strategy. *PLoS genetics* 5, 5 (2009), e1000464.
- [219] TIEDEMANN, K., LE NIHOANNEN, D., FONG, J. E., HUSSEIN, O., BARRALET, J. E., AND KOMAROVA, S. V. Regulation of osteoclast growth and fusion by mtor/raptor and mtor/rictror/akt. *Frontiers in cell and developmental biology* 5 (2017), 54.
- [220] TREROTOLA, M., RELI, V., SIMEONE, P., AND ALBERTI, S. Epigenetic inheritance and the missing heritability. *Human genomics* 9, 1 (2015), 17.
- [221] TUNG, J., ZHOU, X., ALBERTS, S. C., STEPHENS, M., AND GILAD, Y. The genetic architecture of gene expression levels in wild baboons. *Elife* 4 (2015), e04729.
- [222] TURNER, A. W., WONG, D., DREISBACH, C. N., AND MILLER, C. L. Gwas reveal targets in vessel wall pathways to treat coronary artery disease. *Frontiers in cardiovascular medicine* 5 (2018), 72.
- [223] TYLER, A. L., ASSELBERGS, F. W., WILLIAMS, S. M., AND MOORE, J. H. Shadows of complexity: what biological networks reveal about epistasis and pleiotropy. *Bioessays* 31, 2 (2009), 220–227.
- [224] TYLER, A. L., LU, W., HENDRICK, J. J., PHILIP, V. M., AND CARTER, G. W. Cape: an r package for combined analysis of pleiotropy and epistasis. *PLoS computational biology* 9, 10 (2013), e1003270.
- [225] UDAGAWA, N., TAKAHASHI, N., AKATSU, T., TANAKA, H., SASAKI, T., NISHIHARA, T., KOGA, T., MARTIN, T. J., AND SUDA, T. Origin of osteoclasts: mature monocytes and macrophages are capable of differentiating into osteoclasts under a suitable microenvironment prepared by bone marrow-derived stromal cells. *Proceedings of the national academy of sciences* 87, 18 (1990), 7260–7264.
- [226] VEYRIERAS, J.-B., KUDARAVALLI, S., KIM, S. Y., DERMITZAKIS, E. T., GILAD, Y., STEPHENS, M., AND PRITCHARD, J. K. High-resolution mapping of expression-qtls yields insight into human gene regulation. *PLoS genetics* 4, 10 (2008), e1000214.

- [227] VISSCHER, P. M. Sizing up human height variation. *Nature genetics* 40, 5 (2008), 489–490.
- [228] VISSCHER, P. M., BROWN, M. A., MCCARTHY, M. I., AND YANG, J. Five years of gwas discovery. *The American Journal of Human Genetics* 90, 1 (2012), 7–24.
- [229] VOLKMAR, M., DEDEURWAERDER, S., CUNHA, D. A., NDLOVU, M. N., DEFRENCE, M., DEPLUS, R., CALONNE, E., VOLKMAR, U., IGOILLO-ESTEVE, M., NAAMANE, N., ET AL. Dna methylation profiling identifies epigenetic dysregulation in pancreatic islets from type 2 diabetic patients. *The EMBO journal* 31, 6 (2012), 1405–1426.
- [230] WAHL, S., DRONG, A., LEHNE, B., LOH, M., SCOTT, W. R., KUNZE, S., TSAI, P.-C., RIED, J. S., ZHANG, W., YANG, Y., ET AL. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* 541, 7635 (2017), 81.
- [231] WANG, Z., GERSTEIN, M., AND SNYDER, M. Rna-seq: a revolutionary tool for transcriptomics. *Nature reviews genetics* 10, 1 (2009), 57.
- [232] WEI, W.-H., HEMANI, G., AND HALEY, C. S. Detecting epistasis in human complex traits. *Nature Reviews Genetics* 15, 11 (2014), 722.
- [233] WEINREICH, D. M., LAN, Y., WYLIE, C. S., AND HECKENDORN, R. B. Should evolutionary geneticists worry about higher-order epistasis? *Current opinion in genetics & development* 23, 6 (2013), 700–707.
- [234] WEST, M. A., KIM, K., KLIEBENSTEIN, D. J., VAN LEEUWEN, H., MICHELMORE, R. W., DOERGE, R., AND CLAIR, D. A. S. Global eqtl mapping reveals the complex genetic architecture of transcript level variation in arabidopsis. *Genetics* (2006).
- [235] WIDSCHWENDTER, M., JIANG, G., WOODS, C., MÜLLER, H. M., FIEGL, H., GOEBEL, G., MARTH, C., MÜLLER-HOLZNER, E., ZEIMET, A. G., LAIRD, P. W., ET AL. Dna hypomethylation and ovarian cancer biology. *Cancer research* 64, 13 (2004), 4472–4480.
- [236] WILSON, A. G. Epigenetic regulation of gene expression in the inflammatory response and relevance to common diseases. *Journal of periodontology* 79, 8S (2008), 1514–1519.

- [237] WOJCIECHOWSKI, M., CZAPINSKA, H., AND BOCHTLER, M. CpG underrepresentation and the bacterial cpg-specific dna methyltransferase m. mpei. *Proceedings of the National Academy of Sciences* 110, 1 (2013), 105–110.
- [238] WRAY, N., AND VISSCHER, P. Estimating trait heritability. *Nature Education* 1, 1 (2008), 29.
- [239] WRIGHT, F. A., SULLIVAN, P. F., BROOKS, A. I., ZOU, F., SUN, W., XIA, K., MADAR, V., JANSEN, R., CHUNG, W., ZHOU, Y.-H., ET AL. Heritability and genomics of gene expression in peripheral blood. *Nature genetics* 46, 5 (2014), 430.
- [240] WU, X., DONG, H., LUO, L., ZHU, Y., PENG, G., REVEILLE, J. D., AND XIONG, M. A novel statistic for genome-wide interaction analysis. *PLoS genetics* 6, 9 (2010), e1001131.
- [241] XU, S. An empirical bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* 63, 2 (2007), 513–521.
- [242] YANG, C., HE, Z., WAN, X., YANG, Q., XUE, H., AND YU, W. SnpHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies. *Bioinformatics* 25, 4 (2008), 504–511.
- [243] YANG, C., WAN, X., YANG, Q., XUE, H., AND YU, W. Identifying main effects and epistatic interactions from large-scale snp data via adaptive group lasso. *BMC bioinformatics* 11, 1 (2010), S18.
- [244] YOSHIDA, M., AND KOIKE, A. Snpinterforest: a new method for detecting epistatic interactions. *BMC bioinformatics* 12, 1 (2011), 469.
- [245] YOUNG, M. J., BREDDY, J. L., VEVES, A., AND BOULTON, A. J. The prediction of diabetic neuropathic foot ulceration using vibration perception thresholds: a prospective study. *Diabetes care* 17, 6 (1994), 557–560.
- [246] YOUNG, M. J., MARSHALL, A., ADAMS, J. E., SELBY, P. L., AND BOULTON, A. J. Osteopenia, neurological dysfunction, and the development of charcot neuroarthropathy. *Diabetes Care* 18, 1 (1995), 34–38.
- [247] ZELLER, T., WILD, P., SZYMCAK, S., ROTIVAL, M., SCHILLERT, A., CASTAGNE, R., MAOUCHE, S., GERMAIN, M., LACKNER, K., ROSSMANN, H., ET AL. Genetics and beyond—the transcriptome of human monocytes and disease susceptibility. *PloS one* 5, 5 (2010), e10693.

- [248] ZHANG, B., ZHOU, Y., LIN, N., LOWDON, R. F., HONG, C., NAGARAJAN, R. P., CHENG, J. B., LI, D., STEVENS, M., LEE, H. J., ET AL. Functional dna methylation differences between tissues, cell types, and across individuals discovered using the m&m algorithm. *Genome research* (2013), gr-156539.
- [249] ZHANG, F., XIE, D., LIANG, M., AND XIONG, M. Functional regression models for epistasis analysis of multiple quantitative traits. *PLoS genetics* 12, 4 (2016), e1005965.
- [250] ZHANG, W., ZHU, J., SCHADT, E. E., AND LIU, J. S. A bayesian partition method for detecting pleiotropic and epistatic eqtl modules. *PLoS computational biology* 6, 1 (2010), e1000642.
- [251] ZHANG, X., ZOU, F., AND WANG, W. Fastanova: an efficient algorithm for genome-wide association study. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining* (2008), ACM, pp. 821–829.
- [252] ZHANG, Y. A novel bayesian graphical model for genome-wide multi-snp association mapping. *Genetic epidemiology* 36, 1 (2012), 36–47.
- [253] ZHANG, Y., HE, Q., ZHANG, R., ZHANG, H., ZHONG, W., AND XIA, H. Large-scale replication study identified multiple independent snps in ret synergistically associated with hirschsprung disease in southern chinese population. *Aging (Albany NY)* 9, 9 (2017), 1996.
- [254] ZHENG, Q., AND WANG, X.-J. Goeast: a web-based software toolkit for gene ontology enrichment analysis. *Nucleic acids research* 36, suppl_2 (2008), W358–W363.
- [255] ZHU, Z., ZHANG, F., HU, H., BAKSHI, A., ROBINSON, M. R., POWELL, J. E., MONTGOMERY, G. W., GODDARD, M. E., WRAY, N. R., VISSCHER, P. M., ET AL. Integration of summary data from gwas and eqtl studies predicts complex trait gene targets. *Nature genetics* 48, 5 (2016), 481.
- [256] ZILLER, M. J., MÜLLER, F., LIAO, J., ZHANG, Y., GU, H., BOCK, C., BOYLE, P., EPSTEIN, C. B., BERNSTEIN, B. E., LENGAUER, T., ET AL. Genomic distribution and inter-sample variation of non-cpg methylation across human cell types. *PLoS genetics* 7, 12 (2011), e1002389.

- [257] ZOU, J., LIPPERT, C., HECKERMAN, D., ARYEE, M., AND LISTGARTEN, J. Epigenome-wide association studies without the need for cell-type composition. *Nature methods* 11, 3 (2014), 309.
- [258] ZUO, T., LIU, T.-M., LAN, X., WENG, Y.-I., SHEN, R., GU, F., HUANG, Y.-W., DEATHERAGE, D., HSU, P.-Y., TASLIM, C., ET AL. Epigenetic silencing mediated through activated pi3k/akt signaling in breast cancer. *Cancer research* (2011), canres-3573.