PROBING THE ARCHITECTURE OF COMPLEX TRAITS: FUNCTIONAL GENOMICS METHODS AND APPLICATIONS

Matthew C. Weiser

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Curriculum in Bioinformatics and Computational Biology.

Chapel Hill
2015

Approved by:

Terrence S. Furey

Fernando Pardo-Manuel de Villena

Sayan Mukherjee

Ian J. Davis

Wei Sun

ABSTRACT

Matthew C. Weiser: Probing the architecture of complex traits: functional genomics methods and applications
(Under the direction of Terrence Furey)

Genome wide association studies (GWA) have had tremendous success in identifying genetic variants associated with complex traits. However, the majority of associated loci fall outside of protein coding regions, suggesting a role in regulatory function. This has highlighted a critical need for understanding the regulatory architecture of the genome. Recent advances in high-throughput sequencing technology have enabled transcriptional profiling and mapping of epigenetic features across a broad range of cell types and conditions, both in human and model organisms. As a result, an increasingly higher-resolution genome-wide annotation of regulatory elements is now available. Additionally, expression quantitative trait loci (eQTL) studies mapping the genetic basis of gene expression have identified single nucleotide polymorphisms (SNPs) whose allelic variation correlates with gene expression levels. In conjunction with epigenetic annotations, these results have greatly improved interpretability of variants implicated in complex traits. However, a more comprehensive model of epigenetic regulation in disease can only be obtained by directly assaying disease-relevant tissue in affected individuals. Moreover, traditional eQTL methods often perform a prohibitive number of statistical tests, and are underpowered for detecting weaker associations between SNPs and distally-located genes. In the following chapters I present a novel statistical method that reduces eQTL testing burden and improves power to detect genetic variants associated with expression levels of distal genes. Applying this method to data sets in yeast, mouse, and human, I identified thousands of new eQTL and highlighted candidate master regulators, which were consistently enriched across species for metabolic function. Additionally, I

present an analysis of the chromatin and transcriptional landscapes in colon tissue from 33 Crohn's disease and non-IBD individuals. In ten samples, I found evidence of a molecular signature consistent with metaplasia, the prevalence of which was highly over-represented in CD patients. In an analysis of the remaining individuals, I identified thousands of regulatory regions implicated in disease, many of which co-localize with differentially expressed genes, and highlighted several candidate driver transcription factors. Together, these methods and applications provide a richer understanding of genetic and epigenetic variants implicated in complex traits and disease, and provide hypotheses for future follow up studies.

ACKNOWLEDGEMENTS

First and foremost I'd like to thank my advisor Terrence Furey for guiding my research and providing constant feedback, mentorship and encouragement. Over the last five years I've learned a great deal from his thoughtful critiques and insightful approach to problem solving and technical troubleshooting. When I find myself stumped, I'll often begin by asking: "How would Terry approach this?" in an effort to emulate his scientific thinking. That being said, his fantasy baseball prowess so far eludes me, and I realize I still have much to learn in that respect…

I'd also like to thank my committee members Fernando Pardo-Manuel de Villena, Sayan Mukherjee, Ian Davis, and Wei Sun, for their helpful advice and guidance of this project. In particular, Sayan was instrumental in guiding the eQTL work presented in chapter 2, and provided constant essential feedback and statistical input. In addition, none of the Crohn's disease research detailed in chapter 3 would have been possible without the team effort of numerous collaborators, most notably Shehzad Sheikh in the UNC-CH Division of Gastroenterology. This project was driven by Shehzad's energy and expertise; his motivation and excitement throughout have been infectious, and his focus on integrating IBD genomics research in the clinic has been a constant reminder to me of the bigger picture. FAIRE and RNA extraction for this project was performed by Sheikh lab technicians Gregory Gipson, Adam Robinson, and Eric Lee, and I am grateful for their experienced lab hands. It is no exaggeration to say that I literally could not have done this research without their help, as I learned in my first year of graduate school that I am sadly not cut out for bench work. I'd also like to thank the individuals who participated in this study, for graciously donating their tissue to help further our understanding of this debilitating disease.

Yeast genotype and gene expression data discussed in chapter 2 were generously provided by Dr. Rachel Brem. Personal correspondence with Dr. Steve Haase and Dr. David Aylor provided essential context for the interpretation of eQTL results presented in chapter 2. Financial support for this dissertation was provided by NIH grants 5T32GM067553-09 and R01ES024983, as well as the University Cancer Research Fund (UCRF) from UNC-Chapel Hill, and NIH Grants U01-CA157703 and R01-CA166447.

Special thanks go out to the members of the Furey Lab, past and present, for providing new perspectives, helpful advice, and snippets of code, particularly: Jeremy Simon, Bryan Quach, Raulie Raulerson, Nur Shahir, Martin Buchkovich, and Karl Eklund, and Jeremy Wang. Finally, I'd like to thank all the friends and family who have been supportive every step of the way (as well as my dog, Marlo, who makes coming home from work every day feel like a party). In particular, mom and Gail for their constant support, advice, and love; Leah and Glenn for always making room for me on trips to visit; Amari, AJ, and Mia for keeping me up to date on all the cool stuff that old people like me don't know about, like Call of Duty and which Octonaut is best. And last but not least, Kat- your love, encouragement, emojis, and willingness to indulge my degenerate, late-night Taco Bell excursions have kept a smile on my face. Without you I'd still be hopelessly rummaging around in the dumpster.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AvA: all (SNPs) versus all (genes) association testing method

BH: Benjamini-Hochberg (false discovery rate correction)

bp: base pair

BY: Benjamini-Yekutieli (false discovery rate correction)

CC: Collaborative Cross

CD: Crohn's disease

ChIP: chromatin immunoprecipitation

CL: colon-like

DE: differential expression/differentially expressed

DNA: deoxyribonucleic acid

DNase: deoxyribonuclease

DRR: Differential regulatory region

ENCODE: encyclopedia of DNA elements

eQTL: expression quantitative trait loci

FAIRE: formaldehyde-assisted isolation of regulatory elements

FDR: false discovery rate

FWER: familywise error rate

GENEVAR: gene expression variation

GI: gastrointestinal

GO: gene ontology

GREAT: genomic regions enrichment of annotations

GTEx: genotype-tissue expression

GWA(S): genome wide association (study)

HapMap: International Haplotype Map Project

HTSF: high-throughput sequencing facility

IBD: inflammatory bowel disease

ICA: Independent Components Analysis

IL: ileum-like

kb: kilobase

LCL: lymhoblastoid cell line

LD: linkage disequilibrium

lncRNA: long non-coding RNA

LP: lamina propria

LPMC: lamina propria mononuclear cells

MAF: minor allele frequency

Mb: megabase

MHT: multiple hypothesis testing

modENCODE: model organism encyclopedia of DNA elements

NetLIFT: network-based, large-scale identification of distal-eQTL

PCA: principal components analysis

QC: quality control

RNA: ribonucleic acid

SNP: single nucleotide polymorphism

TAF: trans-acting factor

TES: transcription end site

TSS: transcription start site

UNC-CH: University of North Carolina at Chapel Hill

CHAPTER I

Introduction

The publication of the human genome sequence in 2001 [1] ushered in a new era in the biomedical sciences, paving the way for a comprehensive understanding of phenotype, development, disease, and evolution. However, initial results from the study immediately suggested that gene regulation and interaction played a more complex role in shaping complex traits than previously imagined. Surprisingly, the Human Genome Project estimated the number of protein coding genes at only 30,000-40,000 [1], a figure that has since been further reduced to ~21,000 [2], but was nevertheless already far fewer than pre-human-genome estimates of 60,000 or more [3,4]. Furthermore, only ~1% of the 3.2 billion nucleotides in the human genome was found to code for proteins [1]. In conjunction with results from genome wide association (GWA) studies, which have found that most trait-associated variation occurs in non-coding regions, this has highlighted an urgent need to identify all regulatory elements in both the human and model organism genomes, and understand their cell- and condition-specific role in shaping phenotype.

GENOME-WIDE CHARACTERIZATION OF REGULATORY ELEMENTS

In the years immediately following the release of the human genome, rapid technological advances and declining cost of sequencing facilitated cost-effective genome-wide assays measuring transcriptional output, DNA methylation, chromatin structure and interaction, DNA copy number, transcription factor occupancy, regulatory histone modifications, and more. Shortly upon the completion of the Human Genome Project, the Encyclopedia Of DNA Elements (ENCODE) project

was undertaken, with the aim of leveraging the power of high-throughput sequencing capabilities to annotate all functional elements of the human genome [5]. A similar project to study the regulatory architecture of model organisms *Drosophila melanogaster* and *Caenorhabditis elegans* was coordinated by the model Organism Encyclopedia Of DNA Elements (modENCODE) consortium in 2007 [6]. More recently, the Epigenome Roadmap Project [7] has created a compendium of "reference epigenomes" in adult, embryonic, healthy and diseased individuals. Currently, the ENCODE project has produced 1,640 data sets in 147 different human cell types [5], while modENCODE has generated 237 and 700 genome-wide data sets for *D. melanogaster* and *C. elegans*, respectively [8,9]. The Epigenome Roadmap Project has surveyed the epigenomes of 127 tissues and cell types [7], with a primary focus on five core histone modifications. Collectively, integrative analyses from these consortia have provided major insights into genome architecture, gene-gene regulatory relationships, chromatin and transcriptional landscapes, and evolutionary conservation, painting a rich, molecular portrait of how a genome functions.

In human, ENCODE identified nearly 400,000 regions with enhancer features and over 70,000 regions with promoter features; in total, 80.4% of genomic DNA was found to participate in regulatory activity in at least one cell type [5]. Genome-wide regions of accessible chromatin and presence of histone marks were measured with DNase I hypersensitive site sequencing (DNase-seq) and chromatin immunoprecipitation followed by sequencing (ChIP-seq), respectively, and signals at promoter regions were found to be strongly predictive of gene expression levels [5]. Additionally, combinatorial transcription factor binding as assayed by ChIP-seq was found to be cell-type and context-specific [10], with expression levels of target genes correlating strongly with both ChIP-seq derived binding signal [10] and DNase I hypersensitivity [11]. Initial results from the Epigenome Roadmap Project have annotated enhancer and promoter regions in each of 127 cell types, finding that an average of 5% of the genome for a given cell type was marked for either enhancer or promoter activity [12]. Motif analysis of "enhancer-only" regions identified cell-type

specific candidate regulators, and neighboring target gene sets were found to be enriched for cell-type specific function [12], providing an epigenetic framework for defining cell-type identity. These preliminary integrative -omics analyses have given key insights into genetic and epigenetic regulation of transcription, and have provided a model for using high-throughput data to screen for functional elements that play a role in the architecture of complex traits, particularly for those whose misregulation may contribute to disease.

GENOME WIDE ASSOCIATION STUDIES IN COMPLEX TRAITS

In contrast to the inter-omics analyses of ENCODE and the Epigenome Roadmap, genome wide association (GWA) studies have taken a functionally-agnostic approach to associate genetic variants with complex traits [13]. In these studies, thousands to millions of common single nucleotide polymorphisms (SNPs) are tested for statistical association with a binary (ex: disease versus no disease) or quantitative (ex: height) phenotype. SNPs that meet a genome-wide level of significance are thought to represent loci with direct association to the trait of interest, and may be prioritized for further functional studies. Since the first GWA study on age-related macular degeneration was conducted in 2005 [14], thousands more have followed; as of November 2013, the NHGRI GWA catalog contained a total of 11,912 genome-wide significant, trait-associated SNPs, obtained from a total of 1,751 curated publications [15].

Despite the success and widespread adoption of the GWA approach, very few trait-linked loci have been found within coding regions of genes [15], suggesting that the mechanism of association for many GWA loci is exerted via a regulatory influence on a nearby gene or genes. Thus fine-mapping and follow-up analyses are necessary for pinpointing the causal variants that potentially lie in linkage disequilibrium (LD) with an associated lead SNP, and identifying the mechanism of association with the trait. Results from functional genomics studies in tissue types of

interest, such as those from ENCODE, can therefore be of crucial importance in selecting which LD buddy SNPs to prioritize for time-consuming and expensive follow-up studies [16,17].

In order to attain sufficient power to detect meager effect sizes of common variants, GWA studies often perform analysis using thousands to hundreds of thousands of individuals. Still, for almost all traits studied, the combined effects of associated loci explain only a small fraction of trait heritability. Although some debate exists regarding the accuracy of heritability and effect size estimates [18], there are two predominant (and non-mutually exclusive) hypotheses for this "missing-heritability." The first contends that rare variants with minor allele frequencies (MAF) less than 0.01 account for a significant proportion of trait variance; the second claims that many common variants, all with small effect sizes, together account for the unexplained variance. Though it is not currently known to what extent these two hypotheses contribute to the missing heritability, an approach emphasizing functional effects of common variants with small effect sizes has been suggested as a potential way forward in the post GWA era [18], and may produce meaningful interpretations of genetic disease-association as increases in sample size provide diminishing returns in statistical power.

GENE EXPRESSION AS A QUANTITATIVE TRAIT: THE GENETIC BASIS OF TRANSCRIPTION

Borrowing from both the functional and agnostic models of association testing, expression quantitative trait loci (eQTL) analyses treat gene expression levels as a quantitative trait, and seek to identify genetic variants associated with transcription. Much as GWA studies perform association tests between genetic variants and a phenotype of interest, eQTL studies assay genotype and transcription across the same individuals, and systematically test for linkage between genetic markers and expression levels of thousands to tens of thousands of genes [19,20]. Significant associations provide an important information-bridge, providing a greater understanding of mechanism of association identified by transcriptome-phenotype and genotype-phenotype studies.

The first eQTL study was conducted using a cross of laboratory and wild-derived yeast strains [21]; since then, numerous other studies have been conducted in model organisms, including *Arabidopsis thaliana* [22], *C. elegans* [23], rat [24], and mouse [25]. Human eQTL databases such as GENe Expression VARiation (GENEVAR) [26] provide a data-integration and visualization platform for accessing results from multiple human eQTL studies conducted in adipose tissue, lymhoblastoid cell lines (LCL), T cells, skin, and fibroblasts [27–30]. Additionally, the Genotype-Tissue Expression (GTEx) project has recently concluded a pilot analysis [31] of a project involving 43 tissues and 175 individuals, and aims to scale up tissue collection to 900 donors in the coming years [32].

Results from the GTEx pilot analysis have shown significant enrichment for autoimmune-related GWA SNPs among eQTL identified in whole blood cell types, but not in tissues unrelated to disease, suggesting that eQTL results can be useful not only in highlighting functional relevance of GWA SNPs, but also in identifying relevant tissues for disease action [31]. Meanwhile, other studies have used eQTLs in LCLs to prioritize otherwise unknown candidate genes for GWA results in both childhood asthma [33] and Crohn's disease [34].

Although results from eQTL analyses and functional genomics studies have been invaluable in understanding how human genome relates to complex traits, many hurdles remain. In human eQTL studies, performing association tests for all pairs of SNPs and genes in involves billions of tests, leading to challenges in both computational burden and reduced power due to severe multiple hypothesis testing (MHT) corrections. One common solution is to restrict analyses to SNP-gene pairs that are located close to one another in genomic space – usually within 1Mb or less – thereby prioritizing discovery of "local" associations in which SNPs are thought to directly influence gene expression by altering binding affinity of transcriptional machinery. This reduces testing burden but ignores "distal" effects between SNPs and genes located on separate chromosomes, whereby a transcriptional association is mediated by differential expression of an intermediate

gene (presumably close to the eQTL SNP) that then alters the transcription rate of a distally-located target. Better methods for reducing the eQTL search space and identifying distal effects will improve functional annotation of the genome and increase our understanding of the genetic architecture of complex traits and disease.

Additionally, GWA studies have highlighted disease-associated loci, but do not directly provide information regarding tissue of interest or mechanism of effect. While the results from functional genomics studies can be of great use in identifying candidate genes interpreting disease-associated genetic variants, existing data is primarily limited to normal cell lines and tissues. A more comprehensive understanding of disease mechanism can be better obtained by assaying disease-relevant tissue in affected and unaffected individuals. Disease-associated regulatory elements, genes, and pairwise associations identified with this approach will enhance our understanding of molecular basis of disease, and when interpreted in conjunction with existing association studies and functional annotations, may provide novel candidate targets for treatment and/or predict therapeutic response.

In chapter II, I present a novel method for eQTL detection, Network-based, Large-scale Identification oF disTal-eQTL (NetLIFT), which reduced testing burden and outperformed the power of distal eQTL detection compared to existing methods [35]. I applied this method to gene expression and genotype data for yeast, mouse, and human, identifying thousands of novel distal eQTL, and showed a consistent enrichment of distal effects within metabolic pathways. In chapter III, I discuss unpublished work in which an integrative -omics approach in Crohn's disease identified regulatory regions and genes implicated in disease, highlighting functional regulatory relationships and candidate drivers. In chapter IV I discuss how these results improve the resolution of the current image of the functional genome, and contribute to a better understanding of the genetic basis of complex traits and disease.

CHAPTER II

Novel distal eQTL analysis demonstrates effect of population architecture on detecting and interpreting associations[1]

OVERVIEW

Mapping expression quantitative trait loci (eQTL) has identified genetic variants associated with transcription rates, and has provided insight for genotype-phenotype associations obtained from genome-wide association studies (GWAS). Traditional eQTL mapping methods present significant challenges for multiple testing burden, resulting in a limited ability to detect eQTL that reside distal to the affected gene. To overcome this, we developed a novel eQTL testing approach, NetLIFT, which performs eQTL testing based on the pairwise conditional dependencies between genes' expression levels. When applied to existing data from yeast segregants, NetLIFT replicated most previously-identified distal eQTL, and identified 46% more genes with distal effects compared to local effects. In liver data from mouse lines derived through the Collaborative Cross project, NetLIFT detected 5,744 genes with local eQTL while 3,322 genes had distal eQTL. This analysis revealed founder of origin effects for a subset of local eQTL that may contribute to previously described phenotypic differences in metabolic traits. In human lymphoblastoid cell lines, NetLIFT was able to detect 1,274 transcripts with distal eQTL that had not been reported in previous studies, while 2,483 transcripts with local eQTL were identified. In all species, we found no enrichment for transcription factors facilitating eQTL associations; instead, we find that most *trans*-

[1] A version of this work was previously published as Weiser M, Mukherjee S, Furey TS. Novel distal eQTL analysis demonstrates effect of population genetic architecture on detecting and interpreting associations. Genetics. 2014;198: 879–93.

acting factors were annotated for metabolic function, suggesting that genetic variation may indirectly regulate multi-gene pathways by targeting key components of feedback processes within regulatory networks. Furthermore, the unique genetic history of each population appears to influence the detection of genes with local and distal eQTL.

INTRODUCTION

Gene expression is highly heritable, indicating a strong genetic component [36,37]. Expression quantitative trait loci (eQTL) mapping strives to uncover the underlying genetic architecture of transcriptional regulation. An important concept in dissecting complex regulatory processes is to identify both local and distal variants that are associated with gene expression. Local eQTL are largely thought to regulate proximal genes by affecting the activity of regulatory elements that directly influence transcription rates, such as through alterations in genomic sequence that affect binding affinities of regulatory factors. In contrast, distal eQTL map to genomic locations far from the affected gene, possibly on different chromosomes, and likely act initially on the expression or function of some nearby, intermediate gene that then affects the associated target gene in *trans*. Notably, in genetically diverse populations such as humans, the reported effect sizes and significance levels for distal associations are weaker than for local eQTL [21,22,38]. This is likely attributable to the greater noise inherent in indirect effects that occur within the context of a protein-protein interaction network.

Initial eQTL discovery analyses performed association tests for all pairs of genomic variants and genes [39–41], leading to challenges in both sensitivity and interpretation. Although recent methods have greatly reduced the computational burden for this approach [42], the reduced statistical power due to multiple testing correction still present significant problems, especially in detecting distal eQTL. Using this technique, the reported frequency of distal effects has varied from 2% to 75% of all detected eQTL [40,43,44], and it remains unclear whether this is attributable to

differences in regulatory architecture or statistical power. Indeed, in several recent eQTL analyses using human data, distal eQTL mapping was either not performed or not reported [45,46], likely due to the inability to detect any distal eQTL whatsoever. Additionally, inferring the direction of effect of distal associations that result from protein interactions is difficult when dealing with gene expression data that is often noisy and highly correlated.

To detect distal eQTL with greater power, some recently-developed methods assume an underlying regulatory architecture in which the local regulation of an intermediate gene leads to widespread expression variation in a large set of target genes [47–50]. Modules of target genes are defined by factor analysis or gene-gene correlation statistics, and association testing is performed between genotypes and summary statistics of each module. In this setting, strong associations are thought to represent master regulators that exert broad, but potentially weak, effects in the regulatory network. These approaches reduce the multiple testing burden, as thousands of genes are replaced by a few dozen modules; however, there remain several drawbacks. First, if the regulatory activity of a *trans*-acting factor (TAF) affects only a handful of target genes, the initial clustering approach may not identify the small gene module. Secondly, the intermediate genes regulating the expression of gene modules are often not identified. Finally, expression for individual genes belonging to a module do not always correlate with the eQTL associated with the module, raising doubts about the validity of the results [47].

Others have developed methods focused on addressing interpretability and directionality of associations using randomization of genetic variables [51] and causal model selection tests [52] as a foundation for statistical inference. In these methods, conditional dependence between expression of genes and/or latent variables is used to probabilistically determine whether the association between the genetic variant and target gene is causal. In this study, we present a novel eQTL detection method: "Network-based, Large-scale Identification oF disTal eQTL" (NetLIFT), which, rather than performing causal model selection or randomization, uses pairwise partial

correlations derived from gene expression data to restrict distal association testing, thereby reducing the multiple testing burden and highlighting candidate regulatory genes. In this framework, statistically significant local associations are first identified, and then local eQTL variants are tested for distal associations only for genes whose expression values show evidence of direct effects. We show that NetLIFT identifies individual SNP-gene distal associations with greater power than traditional pairwise eQTL testing, scales well to large data sets, and provides interpretability regarding the mechanism of association by highlighting potential *trans*-acting factors. In simulation studies, NetLIFT better identified distal eQTL, especially those with small numbers of target genes, when compared with a traditional all-SNPs-vs-all-genes approach, a module-based approach (Independent Components Analysis, adapted from [50]), and a method designed to identify causal associations using randomization of genotype data [51]. Applying NetLIFT to a data set consisting of 112 yeast segregants [53], we recapitulated previously reported distal associations and putative regulators, while discovering several additional eQTL with plausible biological mechanisms of association. In mouse livers, we discovered founder of origin effects for a subset of local eQTL that drive differential expression of target genes in a subspecies-of-origin-specific manner, suggesting a possible role for these loci in transcriptomic and phenotypic differences between strains. Using data from human lymphoblast cell lines [45], we identified over one thousand distal associations not previously reported. We note that individuals from each of these three populations (yeast, mice, human) have unique genetic histories, and our analysis suggests that this influences the number and type of eQTL detected in each study.

MATERIALS AND METHODS

Description of the NetLIFT Model

The analysis workflow for the NetLIFT model is outlined in Figure 2.1, and was designed to parallel our understanding of the mechanism of *trans* regulatory effects. That is, if SNP s_i, affects the transcription of gene g_j in *trans*, we expect that s_i first directly affects the transcription level of an intermediate gene g_i, and that the transcription rate of g_i directly or indirectly affects the transcription rate of g_j. There are three main steps to the NetLIFT algorithm:

Step 1: Identify local eQTL:

Local association tests are performed for all variants that lie within an a priori defined window of each gene (Figure 2.1a). Allele counts are regressed on the genes expression values, using a univariate, additive linear model. Since some genes contain many more variants than others, we control the false positive rate in local testing by retaining only associations that meet a Bonferroni-corrected significance cutoff of 0.05. Significant associations represent variants that may have a direct effect on the transcription rate of nearby genes, likely by altering activity of *cis* regulatory elements.

Step 2: Estimate pairwise partial correlations for all genes:

Pairwise partial correlations are estimated for all gene pairs (Figure 2.1b) to identify genes with expression level dependencies. The distribution of connections for gene networks has been shown to follow a power-law distribution [54–57] with an overall small numbers of edges. Therefore, we estimate the partial correlation matrix **G** using a method that enforces sparsity on the entries of **G** via L1 regularization, and which has been shown to accurately identify network hubs [58,59].

Briefly, this method performs joint sparse regression on all p variables (genes) simultaneously, by minimizing the penalized loss function:

$$L = \frac{1}{2}\left(\sum_{i=1}^{p}\left\|\mathbf{g_i} - \sum_{j\neq i}\rho^{ij}\sqrt{\frac{\sigma^{jj}}{\sigma^{ii}}}\,\mathbf{g_j}\right\|^2\right) + \lambda \sum_{1\leq i<j\leq p}|\rho^{ij}|$$

where $\mathbf{g_i}$ and $\mathbf{g_j}$ are the expression vectors for genes i and j, ρ^{ij} denotes the partial correlation between genes i and j, and σ^{ii} and σ^{jj} are the i^{th} and j^{th} diagonal entries of the inverse covariance matrix. The L1 penalty λ controls the sparsity of the network, and was optimized by minimizing the BIC criterion outlined in [58].

For *p* genes, the resulting *p* x *p* matrix **G** consists of entries $G_{i,j}$ that represent the correlation between expression vectors $\mathbf{g_i}$ and $\mathbf{g_j}$, conditioned on the expression of all other genes' expression:

$$G_{ij} = \mathrm{corr}\left(g_i, g_j \middle| g_k, k \neq i, j\right).$$

G can be interpreted as an undirected network, where each node represents a gene, and an edge is drawn between two nodes if and only if the corresponding entry in the matrix **G** is nonzero.

Step 3: Distal eQTL testing:

Distal eQTLs are called by integrating the results from these two steps (Figure 2.1c). For each variant s_i that shows significant association to a local gene g_i, we test s_i for association with distal genes g_j that are nearby g_i in the partial correlation network defined by **G**. Since the edges of **G** only account for direct relationships between two genes, we exploit the network structure to search for second-degree (downstream) regulatory effects as well. Specifically, we require two conditions for s_i to be tested for a distal effect on g_j:

i) s_i must be strongly associated with expression of the putative *trans*-acting factor (TAF), g_i; and

ii) genes g_i and g_j must be separated in the partial correlation network by no more than two edges, i.e. either $G_{i,j} \neq 0$, or there exists a third gene g_k such that $G_{i,k} \neq 0$ and $G_{k,j} \neq 0$. Additionally, we incorporate a threshold whereby two-degree genes are tested only if the association between s_i and the intermediate gene g_k meets a user-defined

significance level (we selected $p < 0.2$ for this cutoff in all analyses presented here). Although longer-range interaction effects could be considered by testing genes at increased distances within the network, doing so would exponentially increase the number of tests performed at each distance cutoff. We sought to balance this tradeoff by limiting the edge distance to two.

If a locally-affected gene contains many significantly associated variants, only the variant with the strongest local association is tested with distal genes. Furthermore, we impose directionality in the ambiguous case where two directly connected genes both have local eQTL, by only recording the direction with the strongest distal association. We note that since **G** is a symmetric matrix representing an undirected network of correlated genes, we make no assumption regarding the direction of potential gene-gene effects, and therefore no assumption about how variant-to-gene effects may propagate through the network. Instead, we use the network structure only to select which variant-gene pairs to test for associations. Although significant associations do not provide conclusive evidence of *trans* associations, we expect that many of the distal eQTL will be acting in *trans*, potentially through the putative TAF identified by our method.

We note that the correlation-based network structure used to guide the distal association tests will likely lead to correlations among test statistics. The Benjamini-Yekutieli (BY) FDR correction holds rigorously under general dependence of test statistics [60]; however, this correction is generally considered to be overly-conservative. Instead, we use the standard Benjamini-Hochberg FDR [61], which in simulation studies was shown to perform comparably with the BY correction in the case of general dependency, and in particular for two sided t statistics [62].

ICA Method

The Independent Components Analysis (ICA) methodology was adopted from [50] and applied to the simulated data for comparison with NetLIFT. ICA identifies a predefined number of

hidden variables ("independent components") by factoring the gene expression data matrix, **X**, into a product of two matrices: **X**~**SA**. Each column of matrix **S** corresponds to an independent component or factor, and the *i*-th element of a column is the "activation" level of the *i*-th gene in that factor. These factors are meant to model some latent or underlying biological process. The *k*-th row of matrix **A** reflects the amount of activation of the *k*-th independent component across all individuals, A_{ij} is activation on the *j*-th individual for component *i*. Rows of **A** serve as the response vector when testing SNPs in a linear model. We used the fastICA function implemented in the R programming language to factor the expression data. This algorithm minimizes the statistical dependencies between the columns of **S**, so that each column of **S** defines groups of co-expressed genes. Since the method requires an a priori-defined number of components to use in factorization, we set this parameter to 14; the number of modules in each simulated expression data set. To assign individual genes to components, we used the fdrtool function, which models a column's scores as a mixture of null and alternative distributions. Each entry of the column is assigned an FDR corresponding to the likelihood of belonging to the null. For each component (column of **S**), a corresponding component-set was defined for genes with FDR < 0.05.

Association tests were performed by regressing allele counts on rows of **A**, which represent the activation of each component across individuals. SNP-component associations with Benjamini-Hochberg corrected FDR < 0.05 were considered significant. For each association between a true local eQTL and a component, we defined the number of true positives to be the number of component-set genes which were downstream of the locally-affected driver gene. False positives were defined as any other gene assigned to that component-set.

Trigger Method

The Trigger method is described in [51]. This method aims to infer causality of a genetic variant on expression of a gene by treating genetic variants as randomized variables, and leveraging

the causality equivalence theorem to identify the direction of effect. Briefly, let: s_i bet the genetic variant to be tested for association, and let g_i be a nearby gene. Trigger first tests for association between s_i and g_i (graphically: $s_i \rightarrow g_i$) using a standard likelihood ratio test. This gives $\Pr(s_i \rightarrow g_i)$. If the probability of a local association exceeds a defined threshold, the variant is then considered for distal association testing. A similar likelihood test is used for defining the probability of linkage between s_i and g_j, for all other genes g_j, under the condition that $s_i \rightarrow g_i$, (denoted $\Pr(s_i \rightarrow g_j \mid s_i \rightarrow g_i)$). Finally, we test whether s_i and g_j, are *independent*, given the expression of g_i: $\Pr(s_i \perp g_j \mid g_i \mid s_i \rightarrow g_i$ and $s_i \rightarrow g_j)$. The causality equivalence theorem can be used to show that:

$$\Pr(s_i \rightarrow g_i \rightarrow g_j) = \Pr(s_i \rightarrow g_i) \times \Pr(s_i \rightarrow g_j \mid s_i \rightarrow g_i) \times \Pr(s_i \perp g_j \mid g_i \mid s_i \rightarrow g_i \text{ and } s_i \rightarrow g_j),$$

so multiplying the probability estimates yields an estimate for direct effect of s_i on g_j. We use the R package "trigger" for implementation of this algorithm.

Data Simulation Procedure

A total of ten gene expression data sets were simulated, each with 500 genes and 250 samples. For each set of 500 genes, a network gene structure consisting of 14 disconnected gene modules of varying numbers of genes was imposed. Sizes of gene modules in each data set were as follows: 100 (x2), 50 (x2), 10 (x10), leaving 100 genes that were independent of any module. Module topologies are depicted in Figure 2.2. For each module, the hub gene's expression values for 250 samples were simulated first, by drawing from a standard normal distribution. Each successive downstream gene's expression was modeled as a linear combination of the upstream gene plus random error, using an effect size of ±1, and a random error drawn from a standard normal distribution, represented as follows:

$$\boldsymbol{g_{ds}} = \beta \boldsymbol{g_{us}} + \boldsymbol{\varepsilon}$$

where $\mathbf{g_{ds}}$ and $\mathbf{g_{us}}$ represent expression of the downstream and upstream genes, respectively, and $\varepsilon \sim N(0,1)$. Genes directly downstream of either the hub gene or a highly connected gene (defined as

a gene with degree greater than 20) were chosen to have effect sizes of 1, while all other effect sizes were assigned randomly as -1 or 1 with probability 0.3 and 0.7, respectively.

Next, for each gene, the total number of SNPs for that gene was drawn from a gamma(4,0.2) distribution and rounded to the next highest integer. Minor allele frequencies for each SNP were drawn from a uniform(0.05, 0.5) distribution; from these, diploid genotype frequencies encoded 0, 1, 2 were derived under the assumption of Hardy-Weinberg equilibrium.

For each module, a single gene, not necessarily the hub gene, was chosen to have a local eQTL effect. Since the network topology is undirected, local eQTL effects on non-hub driver genes may lead to spurious distal associations in the analysis. In order to investigate the sensitivity and specificity of the method under these potentially confounding circumstances, we assigned local eQTL effects to hub genes in some modules, and to genes downstream of the hub in others. Furthermore, thirty percent of the 100 independent genes were assigned at random to have local eQTL effects. If a gene was not chosen to have an eQTL, genotypes were assigned randomly to the 250 samples. For genes chosen to have an eQTL, the direction of effect was chosen to be positive or negative with probability 0.7 and 0.3, respectively. Genotype labels were assigned using a genetic algorithm that sought to maximize the effect size under the condition that the significance of association lie within a certain range (here, between 5e-05 and 1e-08). In cases where the eQTL was assigned to the hub gene, all genes in the module were considered as distal targets; however, to model cases where confounding associations may occur between the eQTL SNP and genes "upstream" of the locally-affected gene, we also assigned eQTL effects to non-hub genes.

The retrospective allele assignment allowed the specification of desired eQTL effect sizes and significance levels without the need to explicitly consider the pairwise correlations between genes when performing the genotype simulation. This procedure was carried out for 10 simulated data sets. Each data set consisted of gene expression networks for the same module topologies, and each module's expression was characterized by an identical underlying genetic architecture. We

defined true distal associations as those genes downstream of the locally-associated gene in the expression topology. Working code and a representative simulated data set is available for download at: http://fureylab.web.unc.edu/software/netlift/.

Yeast Data

Gene expression and genotype data, described previously [53] were obtained from R. Brem. 112 yeast segregants were mated from parent strains BY4716 and RM11-1a and grown in culture. Strains were genotyped at 2,957 markers and expression measurements were assayed for 6,216 ORFs. Genes with no available annotation information were removed, leaving a total of 5,647 genes for analysis.

Mouse Liver Data

Gene expression data was previously assayed on the Affymetrix Mouse Gene 1.0 ST array, and was obtained from GEO (accession number GSE22297) [63]. Expression values were normalized using the "rma-sketch" option in the Affymetrix Power Tools package. Probes containing SNPs were masked in the normalization procedure. Probesets that were expressed at a level above 6 on a log2 normalized scale in at least 87.5% of mice were retained, leaving a total of 9,377 probesets for further analysis. Genotypes for 181,752 markers from the "A" test array for the Mouse Diversity Array were obtained from D. Aylor.

Human Lymphoblastoid Cell Line Data

Gene expression data and HapMap phase 2 and 3 genotypes were obtained from http://eqtl.uchicago.edu. Normalization and processing were performed as described previously [45]. Additionally, the top 25% of transcripts ranked by expression level were retained for further

analysis, based on median expression level of the pre-quantile normalized data across all 69 individuals, leaving 9,810 transcripts that were retained for analysis.

RESULTS

Simulation Analysis

To assess the sensitivity and specificity of NetLIFT for identifying distal eQTLs, we applied the method to ten simulated data sets consisting of paired expression and genotype data (see Methods).

For comparison, we also tested three previously described eQTL detection methods: Independent Component Analysis (ICA), Trigger, and an All-vs-All pairwise testing approach (AvA) (Figure 2.3). The ICA method is primarily suited to identify eQTL that drive the expression of large numbers of distal genes; however, we note that the number of desired components must be defined according to some empirical criteria, and no specific intermediate gene is pinpointed as the *trans*-acting factor responsible for large scale variations. Therefore, this method does not identify local eQTLs.

We first compared the network structures inferred by NetLIFT's partial correlation analysis to the true simulated regulatory architecture. We found that NetLIFT estimates the gene-gene partial correlation structure with high sensitivity, but note that as module connectivity increases, specificity decreases (Table 2.1, Figure 2.4). However, since the network structure is used primarily to determine which SNP-gene tests to perform, the main effect of false network edges is a slight increase in testing burden. As a result, we were willing to tolerate a reduction in network accuracy so long as the sensitivity remained high.

For detection of local eQTL effects, NetLIFT, Trigger, and AvA both identified true positives with 100% success (FDR < 0.05, Table 2.2). The local eQTL false positive rate for NetLIFT was identical to AvA under this FDR; setting a stricter FDR cutoff of 0.001 resulted in only one false

positive for both methods. Additionally, we observed a large number of false positive local eQTL for Trigger, likely due to a lenient default thresholding criterion in the local eQTL testing step. Since we are particularly interested in this method's ability to detect distal eQTL, and since distal eQTL identification is conditional on local linkages for this method, we chose to retain the permissive threshold and focus primarily on results for distal associations.

Intra-module distal eQTL were predicted using each method simultaneously considering all genes and SNPs from all simulated modules. For each module, the true set of distal effects was defined as all SNP-gene associations between the module eQTL and genes downstream of the locally-affected gene. Thus, for modules where the eQTL acted on the hub gene, all combinations of the local eQTL SNP with non-hub genes were considered "true positives." For modules with eQTL acting on non-hub genes, the true positives were defined as the eQTL-gene pairs in which the associated genes were downstream of the locally-affected, driver gene. False positives were defined as eQTL-gene associations where the associated gene was not downstream of the locally-affected gene. Figure 2.5 details the performance of each of the four methods.

In this case, NetLIFT identified true distal associations at a higher rate for all module topologies (overall 77.9% detection rate), at the cost of a slightly elevated false positive rate. These false positives were mostly due to eQTL SNPs being linked distally to genes that were in the same module, but that were not downstream of the locally-affected gene. Since our network estimation step cannot infer directionality of expression effects, these false associations reflect our inability to distinguish true functional associations from those that are due to confounding gene expression correlations present in the data. However, we note that the estimation of direct gene-gene effects and subsequent testing procedure prevents many upstream genes from being tested against the eQTL SNP, reducing the overall burden of these false associations. Moreover, in a rank based test performed on FDR values, true positives were found to have higher significance values than the

false positives (p = 4.92e-96), again suggesting that the false positive count is strongly dependent on the FDR threshold chosen.

The AvA approach performed poorly, as most true associations were lost after correcting for multiple hypothesis testing. ICA performed well in large module settings, but poorly for small modules, suggesting that this approach is underpowered for detecting small co-regulated gene modules under the influence of a common variant. Trigger performed better than an AvA approach, though in general identified fewer than 12% of true distal associations. NetLIFT was the only method to consistently identify distal effects in all network topologies.

We next evaluated NetLIFT's performance in detecting "hotspot" eQTL loci, where a hotspot is defined as a locus that is associated with more transcripts than is expected by chance. To derive a FWER for each locus, we used the procedure described in [64], which permutes genotypes among samples but preserves the correlation structure present in the gene expression data. Performing association testing with the permuted genotype data sets yields a distribution of the expected maximum number of linkages under the null hypothesis of no eQTL associations. When restricting to a FWER of 0.05, NetLIFT identified the eQTL for all hub-based gene modules as hotspots in 10/10 simulated data sets, while the AvA approach identified these eQTL as hotspots only 20-60% of the time, and with many fewer linkages (Table 2.3).

To investigate whether a larger simulated data set affected the sensitivity and/or specificity of our method, we generated and analyzed an additional simulated data set consisting of 2,000 genes. We observed that the overall fraction of true and false positives remained similar in this analysis (data not shown). These simulation results indicate that in addition to scaling well to large data sets, NetLIFT may discover distal eQTL that are not readily identifiable with existing detection methods.

Analysis of 112 yeast segregants

We applied NetLIFT to previously analyzed paired genotype/gene expression data for 112 haploid yeast segregants [53]. After filtering for genes with available annotation, 5,647 genes and 2,956 variants were retained for analysis. Variants within 10kb of the gene's transcribed region were considered "local," and all other linkages were denoted as distal eQTL. At an FDR of 0.05, we identified a total of 1,124 (19.9%) and 1,642 (29.1%) genes with local and distal eQTL effects, respectively (Figure 2.6). Local and distal effects were observed to have a similar effect size and level of significance (Table 2.4). The large effect sizes for distal eQTL are in line with previously reported results, and are likely attributable to the extreme diversity between the two strains of yeast.

A GO analysis using all 143 genes identified as intermediate *trans*-acting factors (TAFs) for at least 10 downstream targets revealed enrichments for a wide range of functions, with top hits reserved for metabolic function and transport (Table 2.5). This corroborates previous findings where putative regulators located near hotspots were not found to be enriched for transcription factors; instead, evidence suggests that many *trans* regulators exert widespread transcriptional effects by mediating levels of key metabolites or regulating post-translational processes [44,65]. A comprehensive list of all putative regulators is provided in Table 2.6.

For most previously identified hotspots, NetLIFT correctly identified biologically validated regulators (Table 2.7). Several predicted novel regulators with more than 15 target genes were also found, many involved in metabolic and biosynthetic processes. In some cases, we provide regulatory evidence for novel drivers not identified previously for detected hotspots; furthermore, our results suggest that there may be numerous secondary drivers within previously identified hotspot regions, indicating that local association signals arising from two or more distinct loci may influence a similar set of distal target genes. One example is the hotspots on chromosome 2 where target genes are enriched for ribosome biogenesis and ncRNA processing (Table 2.7). Previous

results implicated *AMN1* and *MAK5* as *trans*-acting factors for subsets of the target genes; however, patterns of linkage to distinct regions within this locus suggest that additional regulators lie on chromosome 2 [21]. In addition to *AMN1*, NetLIFT implicated at least seven new candidate regulators on chromosome 2– *TBS1, ARA1, YSW1*, *TOS1*, *UMP1*, *NPL4,* and *YBR197C*– that were strongly linked with local eQTL ($p < 1.0e-05$) and were associated with highly overlapping sets of distally-associated genes (Figure 2.7). Notably, we fail to identify *MAK5*, as this putative regulator was shown to contain a loss of function mutation which has no effect on transcription [21]. By definition, distal effects arising from amino acid substitutions affecting protein function of the *trans*-acting factor will be undetectable using NetLIFT, as we specifically seek to identify distal effects that arise from local, *cis*-regulatory effects.

Given the strong enrichment for ribosome function among target genes linking to the chromosome 2 loci, we hypothesized that causal variants would significantly affect growth rates via widespread differential transcription originating from direct up-/down- local regulation of the candidate TAF. To investigate this, we used segregants' gene expression profiles to predict relative growth rate, using previously described methods [66]. We then tested each of the candidate regulators' distal eQTLs for association with the growth rate phenotype. After correction for multiple testing, we found that nearly all of the underlying variants attained significance at FDR < 0.05. We propose that differential expression of the putative regulators influences growth rate by perturbing common, growth-related pathways in *trans*.

We found numerous loci linking to small sets of target genes that are functionally related, as might be expected from the simulation results. *TEC1*, a transcription factor that targets filamentation genes, was found to have a significantly associated local variant that was distally linked to 16 genes enriched for pseudohyphal growth annotation ($p=1.03e-03$). Additionally, for 5 of these 16 genes (31.2%), the YEASTRACT database shows direct evidence of *TEC1* DNA binding and transcriptional regulation [67]. Of the 25 genes that mapped to the lead variant (defined as the

variant with strongest local effect on *TEC1*) in an all versus all test, only 4 (16%) showed direct evidence of *TEC1* binding and regulation, suggesting that NetLIFT is better able to identify biologically relevant associations.

We identify several putative regulators that are metabolic enzymes and whose target gene sets are enriched for metabolic and biosynthesis annotations. For example, a locus on chromosome 2 that acts as a local eQTL for *LYS2* was distally associated with 167 target genes enriched for the GO term "lysine biosynthetic process via aminoadipic acid" (p=1.27e-07). *LYS2* catalyzes the reduction of alpha-aminoadipate to alpha-aminoadipate semialdehyde (αAASA), the fifth step in the lysine biosynthesis pathway. Downstream of this reaction, glutamate-forming saccharopne dehydrogenase, which consists of the structural determinant *LYS9* and the regulatory product *LYS14*, converts αAASA to saccharopine. *LYS9* loss of function increases intracellular levels of αAASA, which induces the regulatory activity of *Lys14p* and results in the up-regulation of several genes in the pathway, including *LYS1*, *LYS9*, *LYS2*, *LYS4*, *LYS20*, and *LYS21* [68]. In a previous experiment, a mutant strain with loss of function for both *LYS2* and *LYS9* was shown to have decreased intracellular αAASA and lower levels of transcriptional activation of pathway genes, relative to the *LYS9* single mutant [69,70]. We hypothesize that strains harboring the genomic variant associated with decreased transcription of *LYS2* will have a similar reduction of intracellular αAASA concentration, and thus a decreased potential for transcriptional activation of *Lys14p*. Of the previously mentioned lysine biosynthesis genes that are targeted by *Lys14p*, we find four linked distally to the putative eQTL (*LYS1*, *LYS9*, *LYS20*, and *LYS21*). We note that the direction of effect between the eQTL and the downstream genes reflects what we expect under the proposed mechanism (Figure 2.8). Among the set of transcriptional targets are four additional genes whose promoters contain the *Lys14p* binding motif, TCCRNYGGA, one of which, *LYS12*, is involved in lysine biosynthesis and has a directional expression pattern matching the other *Lys14p* targets (Figure 2.8).

Analysis of 156 partially inbred mouse lines

To test how well NetLIFT scales to larger data sets, and for organisms with more complex mechanisms of gene regulation, we analyzed paired genotype and liver gene expression data from 156 partially inbred mice originating from eight founder mice (A/J, C57BL/6J, 129S1/SvImJ, NOD/LtJ, NZO/HlLtJ, CAST/EiJ, PWK/PhJ, and WSB/EiJ), part of the Collaborative Cross (CC) project [71,72] (Figure 2.9). Founder strains of the CC were chosen to provide a high level of genetic diversity, and represent three subspecies of origin: *Mus mus domesticus, Mus mus castaneus,* and *Mus mus musculus*. Wild-derived WSB/EiJ and classical inbred strains A/J, C57BL/6J, 129S1/SvImJ, NOD/LtJ, NZO/HlLtJ have a genetic background comprised mostly of the *Mus mus domesticus* subspecies, while the wild-derived CAST/EiJ and PWK/PhJ founder strains are primarily representative of the *Mus mus castaneus* and *Mus mus musculus* subspecies, respectively [71,72].

We filtered for probe sets expressed above background levels and retained 9,377 genes for analysis. PCA analysis revealed no batch effects in the data (Figure 2.10). Genotypes for the same mice were available for 171,761 markers. In a previous analysis, a total of 6,182 eQTL were discovered for 5,733 genes at a 5% genome-wide threshold; 75% of eQTL were within 10cM of the affected gene [63].

For eQTL testing, we defined local effects as those where variants were within 1Mb of the affected gene, based on the marker-to-gene distances for linkages reported previously for these data [63]. We detected a total of 5,744 genes (61%) with a local eQTL, and 3,322 (35%) with at least one distal eQTL (FDR < 0.05). Of the genes with a distal eQTL, 1,102 (12%) were linked to one SNP, 574 (6%) were linked to two SNPs, 400 (4%) were linked to three SNPs, and 1,246 (13%) were linked to four or more SNPs.

We next investigated patterns of large-scale effects on the regulatory architecture that are attributable to founder and/or subspecies of origin. For the 293 genes with a local eQTL that was linked to at least 5 genes on different chromosomes, genes inherited from a PWK genetic

background showed more extreme expression variation than genes inherited from the other founder strains (Figure 2.11). Mice from the CC have been shown to be phenotypically diverse for various immune related phenotypes [73,74], body weight [75], and behavior [75], with variance for some traits exceeding that observed in the founder strains [75]. One plausible reason for this is that epistatic interactions between alleles inherited from distinct subspecies (*castaneus*, *domesticus*, and *musculus*) may severely mis-regulate gene expression and homeostasis. To investigate whether allele inheritance from different subspecies of origin led to more extreme expression for particular combinations of locally-acting eQTL alleles and target genes, we mapped both eQTL SNPs and target genes to their subspecies of origin. Since alleles inherited from PWK mice appeared to be driving extreme expression variation in locally-affected genes, we reduced the locally-affected set of genes to a subset of 61 genes for which the *Mus musculus musculus*-derived PWK allele explained at least half of the overall genetic effect on expression (Figure 2.12a). We observed that for these SNPs, expression of distally-linked genes showed differential variation based on the combinatorial genetic backgrounds of the locally-associated variant and target gene (Figure 2.12b).

These transcriptomic differences may in turn affect phenotype. Body weight for wild-derived founder strains (CAST/EiJ, PWK/PhJ, WSB/EiJ) used in the Collaborative Cross is lower than in classical laboratory strains [63]. A GO analysis performed for the 142 distal genes linking to the PWK-driven eQTL revealed annotation for various terms related to metabolism and lipid processes (Table 2.8). This enrichment suggests a possible role for the candidate *trans* acting factors in regulating weight, via a broad but subtle effect on gene expression.

Analysis of 69 human individuals

RNA-seq data from lymphoblastoid cell lines and HapMap genotype data for 69 Nigerian individuals were recently interrogated for eQTLs [45]. For NetLIFT analysis, expression data was corrected for GC content and batch, and was normalized as described previously. We selected 9,810

Ensembl transcripts in the top quartile based on median expression level for further analysis. Genotype data for the same individuals, consisting of 9.5 million SNPs, were obtained from HapMap phase 2 and 3, release 27.

Using a local regulatory window of 200kb, similar to the original analysis [45], we identified 2,483 transcripts (25.3%) with a local eQTL effect (FDR < 0.10). Of the 929 transcripts previously identified as having local associations at the same FDR, we replicated 538. The remainder not found consisted of transcripts that we removed from the data set due to low median expression level, with the exception of 3 transcripts that were not identified in our analysis. In addition, we identified 1,945 novel local associations, likely attributable to greater power resulting from testing only the most highly expressed quartile of transcripts.

NetLIFT identified 1,274 transcripts (13.0%) with at least one distal eQTL (FDR < 0.10, Figure 2.13). None were reported in the previous analysis [45]. A traditional all SNPs-vs-all genes testing approach on this filtered set of genes and variants yielded only 5 significant distal associations at this FDR, indicating that our method is better powered for detecting these associations. A GO analysis for the 64 candidate regulators that were linked to at least 3 transcripts (FDR < 0.1) again suggested enrichment for metabolic and biosynthetic processes (Table 2.9).

DISCUSSION

Genome Wide Association Studies (GWAS) have so far identified thousands of quantitative trait loci associated with hundreds of complex traits [76]. However, the success of GWAS has been tempered by a lack of understanding of the mechanism of association for many variants. eQTL studies have shown excellent promise in highlighting potential biological mechanisms of SNP-phenotype associations, and prioritizing particular variants for follow up studies [40]. Furthermore, the correlation between significance levels of SNP-phenotype associations and eQTL associations may help to identify tissue types that play a key role in disease etiology [77]. Recently, gene-gene

interaction evidence has been incorporated in the GWAS setting to identify epistatic effects on phenotype [78], suggesting that correlation based testing may increase power to detect associated variants. We described here a novel method, NetLIFT, that addresses the problems of computational burden and power in traditional eQTL testing, by reducing the search space and using conditional dependencies between genes' expression to prioritize variant-gene testing. The reduced multiple testing correction penalty under our algorithm allows detection of weaker eQTL effects that are missed by currently available methods. Furthermore, our results provide immediate interpretability of the mechanism of association, by highlighting potential regulatory genes that mediate discovered distal effects. We note that in the current implementation of our code, runtime and memory usage increases nonlinearly as the number of genes increases, and that the major bottleneck in runtime is the estimation of the partial correlation matrix. Therefore, when the number of genes exceeds 10,000, users may wish to filter gene expression data sets by most highly expressed or most variable genes.

Importantly, we showed through simulations that NetLIFT can identify instances where distal eQTL only affect a small number of genes, not just the large hub genes found by other methods. Additionally, candidate regulators that are putatively affected in *cis* by the causal variant can be identified, highlighting potential mechanisms of association. We note that since our method seeks to identify distal effects that arise via alterations in the expression level of *trans*-acting factors located nearby the eQTL, we are unable to detect associations mediated by a loss-of-function coding variant in the *trans*-acting factor.

We demonstrated the ability of NetLIFT to identify distal eQTL in three very different data sets. In yeast segregants, we replicated numerous distal eQTL reported previously, as well as the biologically validated regulators for many of the associations. Additionally, we identified several novel biologically plausible distal associations. In inbred lines from genetically diverse founder mice, we detected an interesting pattern of eQTL effects driven by PWK-derived alleles, which may

provide clues as molecular underpinnings of downstream phenotypes such as reduced mouse size in the wild-type derived PWK mice. Lastly, in a set of 69 human individuals, NetLIFT was able to find over 1,200 gene transcripts with significant distal eQTL due to its increased power, whereas previously only 5 had been identified.

Intuitively, one might think that the best candidates for asserting regulatory influence on distal genes would be transcription factors that directly participate in controlling gene transcription rates. In accordance with previous results, though, we found no enrichment for transcription factor annotation among genes implicated by our method as *trans*-acting factors; instead, we find that many of these genes play a role in metabolic and biosynthesis pathways. This suggests that more commonly, the regulation of key genes in these pathways play a role in feedforward or feedback processes that then affect transcription rates of downstream target genes within the same pathway. These indirect effects are more subtle than the direct effects associated with local eQTL, but they can have significant effects on phenotypes, such as growth rates (seen in yeast) and size (seen in mouse).

Our results also highlight an often unaddressed topic in complex trait mapping; namely, that eQTL discovery and interpretability of mapping results is significantly influenced by the genetic and genomic diversity within the sample population. The two yeast strains from which the analyzed segregants were derived were extremely diverse, with an estimated sequence divergence of 0.5-1%. This, and overall genome complexity, likely contributed to many distal effects being found to be as strong as local effects, enabling their easier detection. Genetic incompatibilities between progenitors can result in atypical patterns of linkage disequilibrium, which present challenges in identifying causal versus linked markers. In an inbred mouse model, we were able to identify numerous distal linkages where expression variation in the distally-affected genes appears to be driven by differences in the genetic background at the local and distal loci. However, the resolution of the eQTL mapping is ultimately restricted by the randomization of the genome that is mediated

by recombination events. On the other hand, human studies typically involve genetically diverse individuals, whose genomes are randomized to a greater extent. Thus a model organism may allow for *accurate* eQTL mapping at the expense of *precision*, whereas in human populations we expect to identify eQTL with precision, but reduced accuracy.

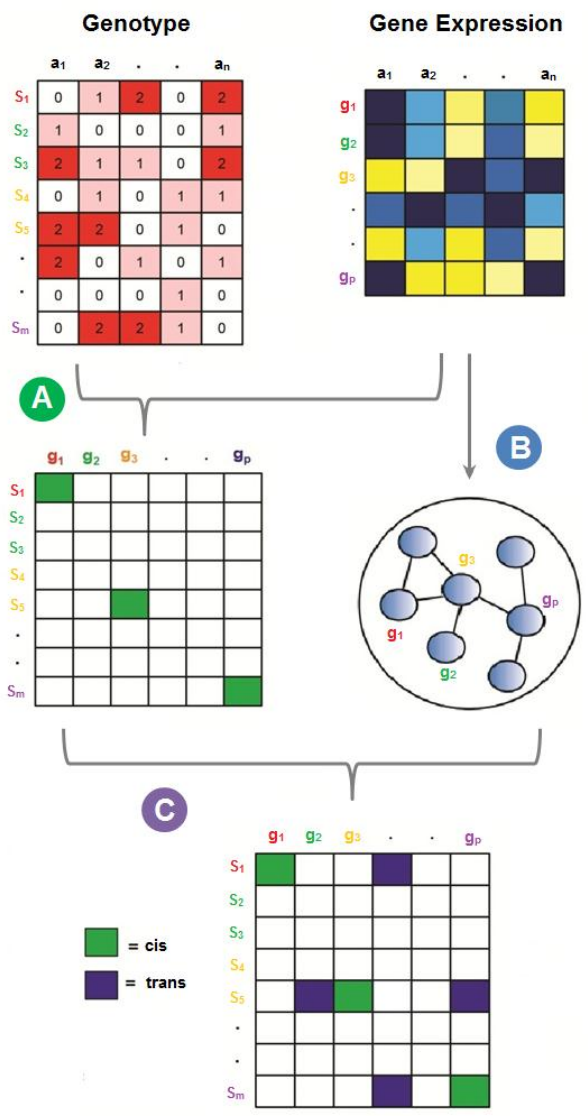
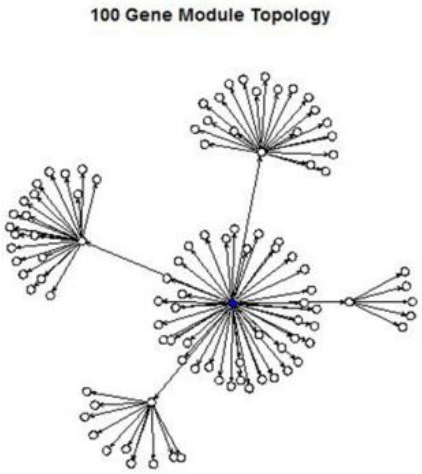
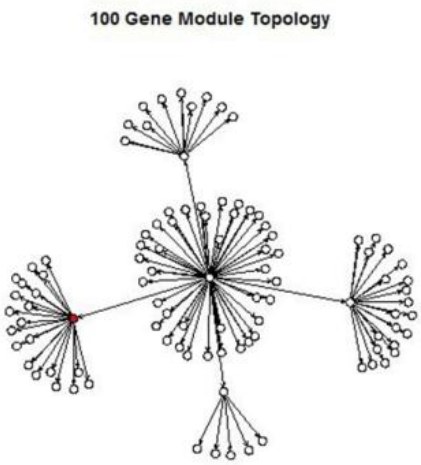
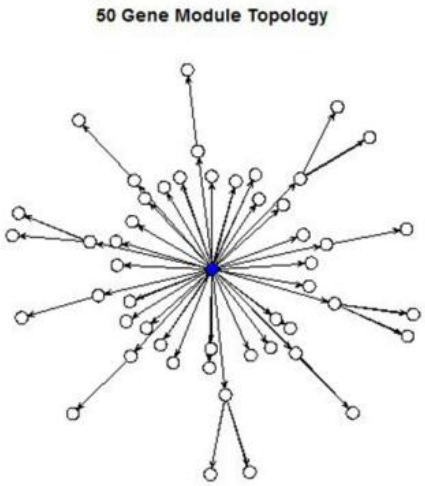
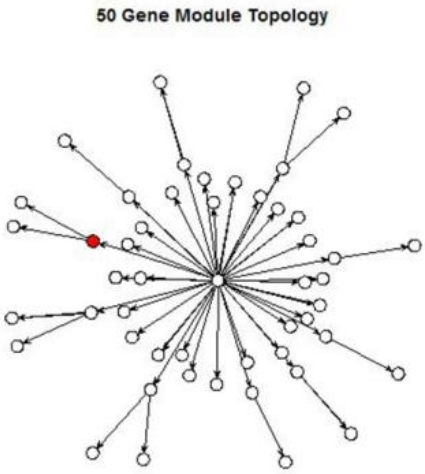


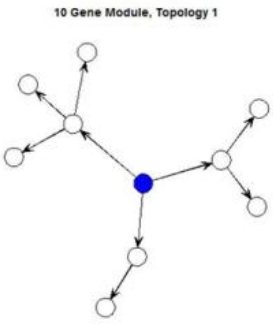
Figure 2.1. Schematic of the NetLIFT method. Top: genotypes for 'm' markers (s_1, s_2, … s_m) and 'p' genes (g_1, g_2, … g_p) are assayed for the same 'n' individuals (a_1, a_2, … a_n). Markers and genes that map to the same locus are color coded. Local eQTL mapping is performed for markers and nearby genes using an a priori defined genomic distance for local effects (A), yielding a local eQTL effect matrix (significant marker-gene associations depicted in green). A sparse partial correlation matrix is inferred from the expression data, representing a network of gene-gene interactions (B). Finally, significantly associated local eQTL markers are tested for distal eQTL effects on genes near the locally-affected gene in the interaction network (C).

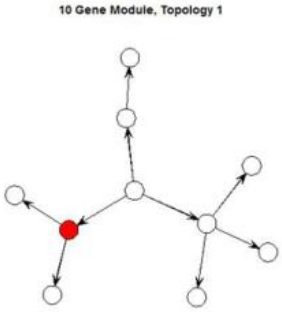

100 Gene Module Topology

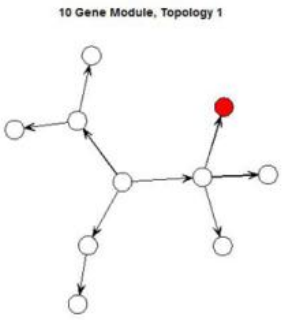

100 Gene Module Topology


50 Gene Module Topology


50 Gene Module Topology


10 Gene Module, Topology 1


10 Gene Module, Topology 1


10 Gene Module, Topology 1

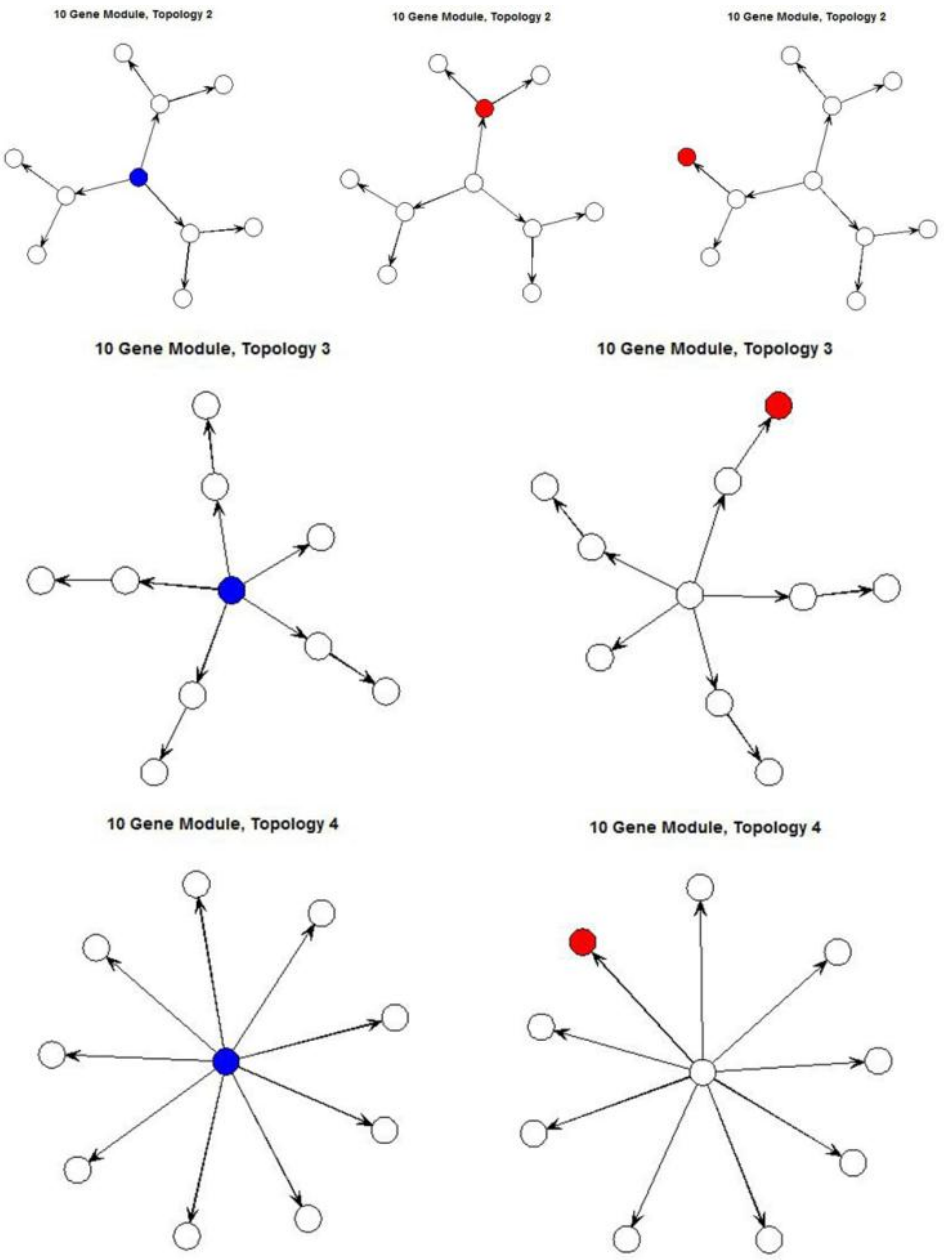


Figure 2.2. Simulated gene module topologies. Each module's expression effects were simulated by first generating the hub gene's expression; each successive downstream gene's expression values were simulated using the upstream gene's expression as a baseline (dependencies indicated by arrows). For each module, a single local eQTL effect was simulated for a SNP assigned to either the hub gene (blue), or to a gene downstream of the hub (red), but not both.

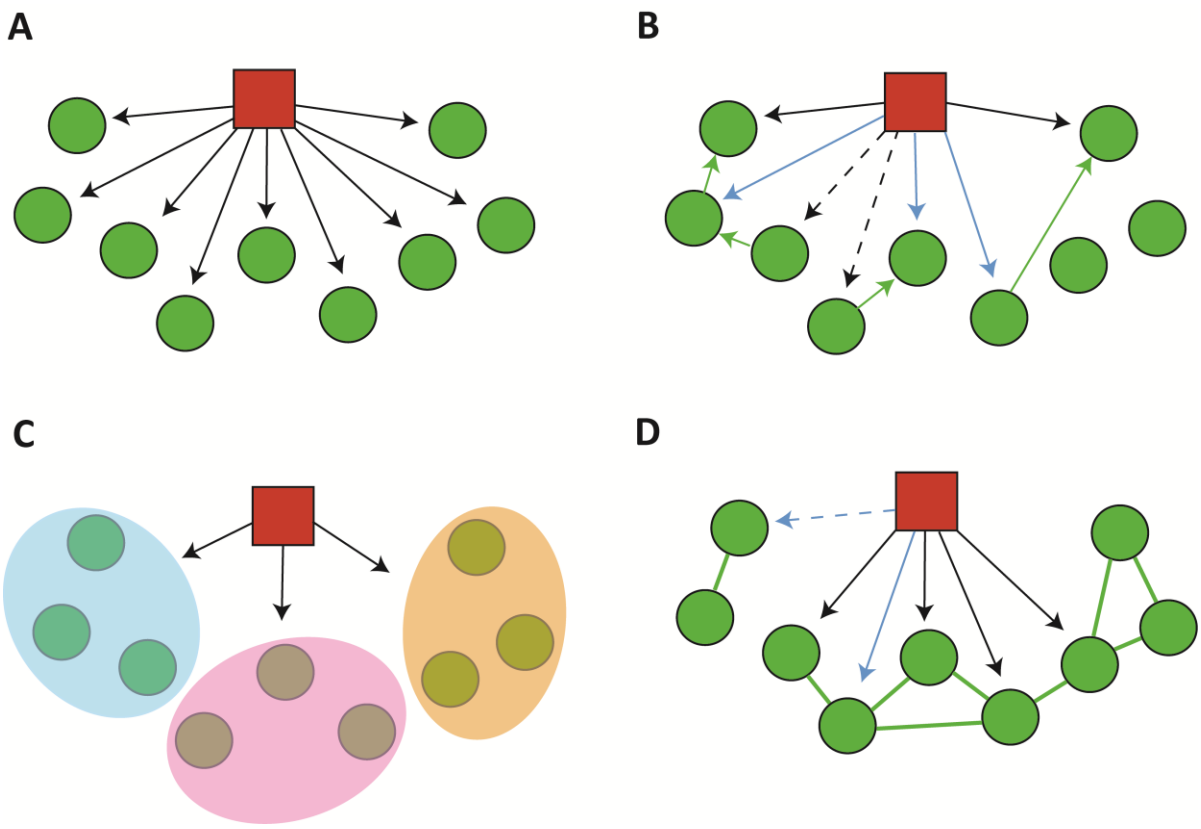


Figure 2.3. Illustration of eQTL detection methods. SNP is depicted as a red node; genes depicted in green. All vs All (A) performs a standard regression significance test for all pairs of SNPs and genes. Trigger (B) seeks to identify distal associations that are mediated by a locally associated variant-gene pair (local associations depicted with blue arrows). Genes downstream of the inferred direction of gene-gene effects (represented by green arrows) should be associated with the variant (true distal associations = solid black arrows), while genes upstream of gene effects will not show association (dashed black arrows). Independent Components Analysis (C) first factors expression data into Independent Components, then performs association tests between allele frequency and the activation levels of components across samples. NetLIFT (D) first performs local linkage tests for a SNP and nearby genes (blue arrows). For significant linkages (solid blue arrow), distal eQTL tests are performed for all genes in the network which are one- or two- edges removed from the locally affected gene (black arrows).

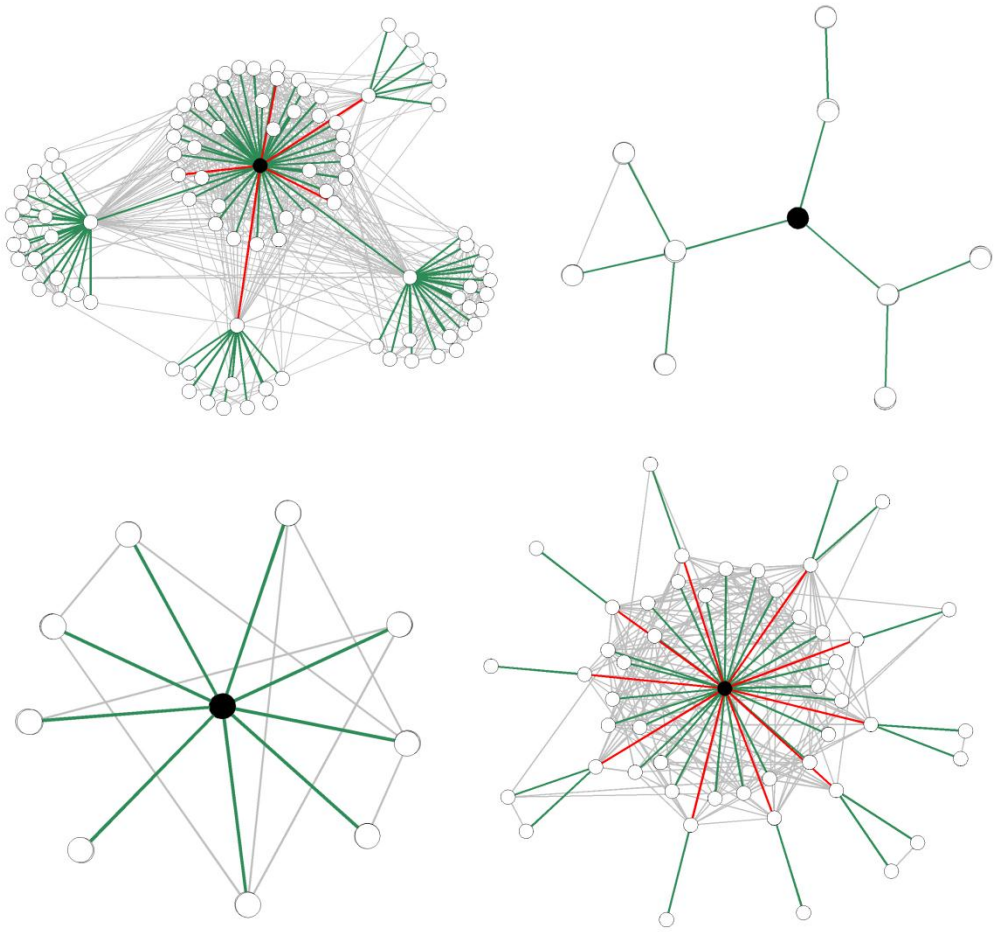


Figure 2.4. Partial correlation structure from network detection step, for representative 100 gene, 50 gene, and 10 gene modules. True positive correlations depicted with green edges, false negatives correlations in red, false positives in gray.

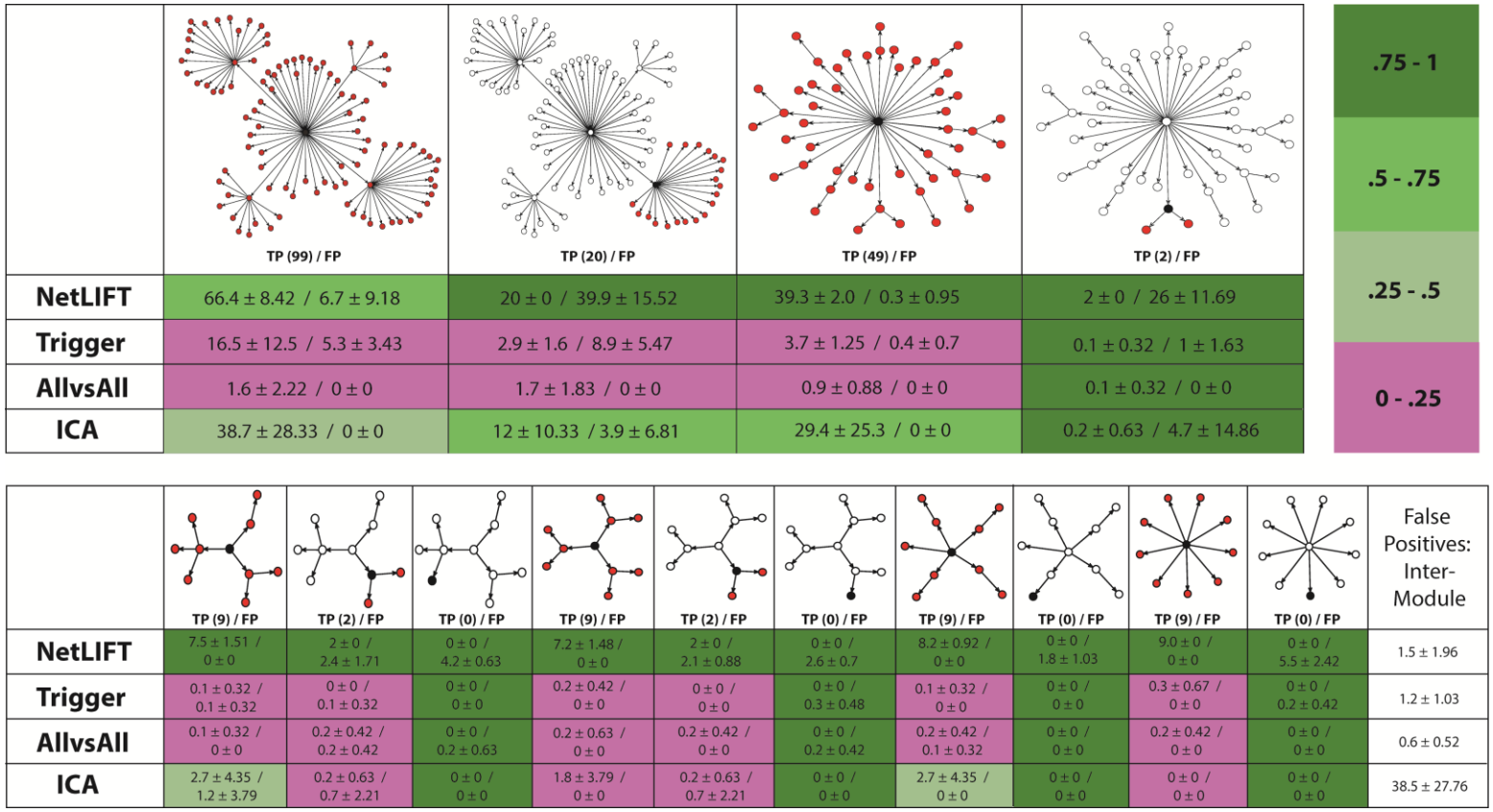


Figure 2.5. Number of detected distal associations, by module topology/method. Topology of each network module is depicted in header. Black nodes depict genes with an assigned local eQTL effect, and red nodes represent "true" distally-associated genes. Total number of "true" distal associations given in parentheses. Each cell value reports the mean and standard deviation of TP / FP, over the ten simulated data sets. Cells are colored according to fraction of true positives discovered. Rightmost column (bottom row) reports the number of false positive distal associations where the locally-regulated gene and target gene belonged to disjoint modules.

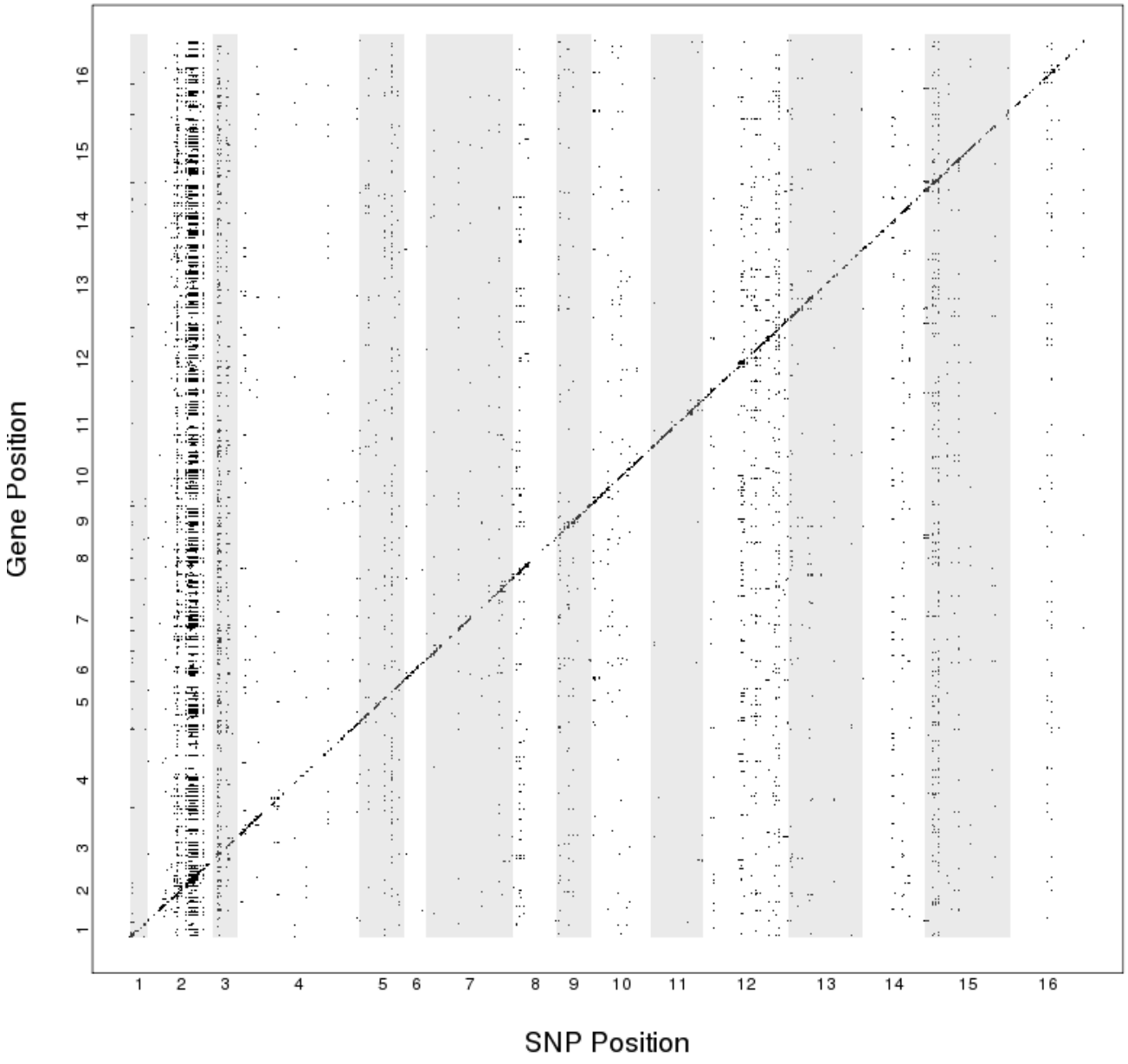


Figure 2.6. Local and distal eQTL linkages in yeast. X axis shows the genomic coordinates of marker variants; Y axis represents gene position. Each dot represents a significant marker-gene association at FDR < 0.05.

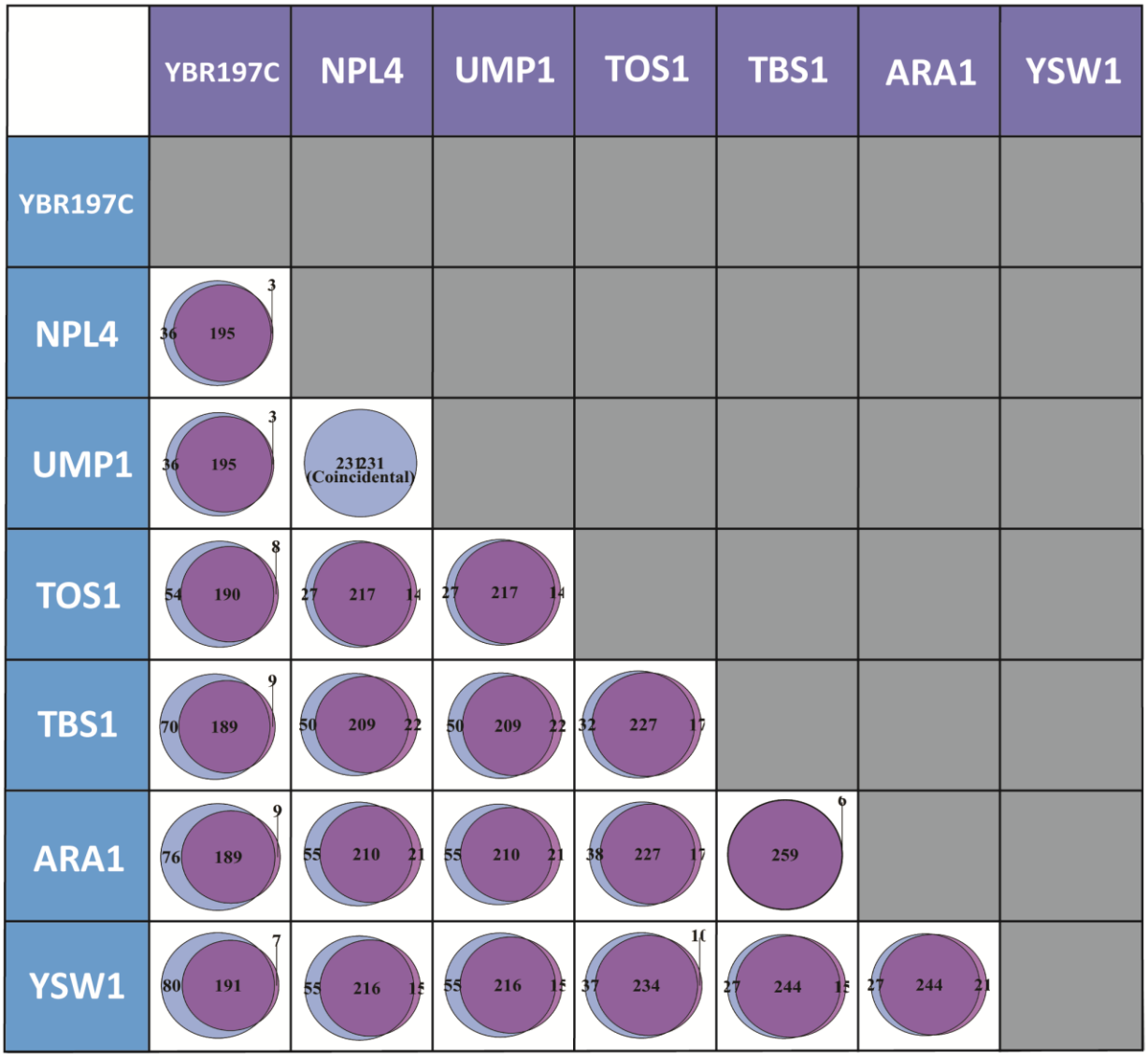


Figure 2.7. Pairwise overlap of target gene sets enriched for ribosomal annotation. Cell [i,j] shows the target gene overlap for between proposed regulators i, j.

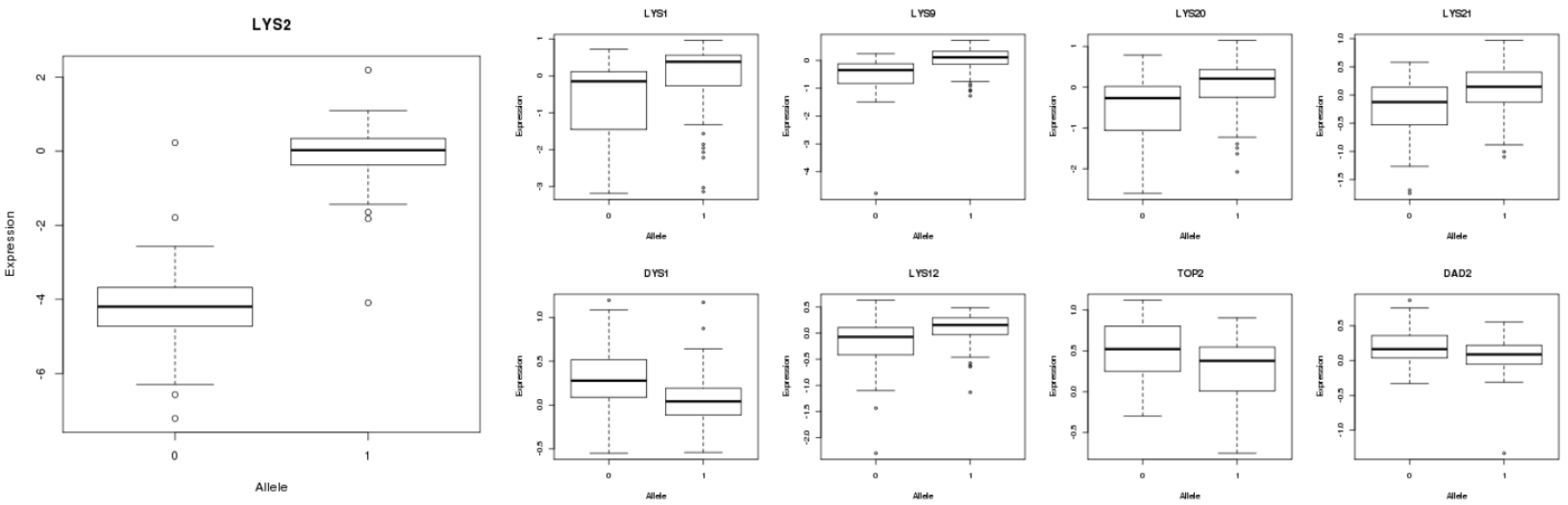


Figure 2.8. eQTL effects for LYS2 local regulatory variant and downstream genes. The allele associated with lower *LYS2* expression ("0") is associated with lower expression of known *Lys14p* targets *LYS2*, *LYS1*, *LYS9*, *LYS20*, and *LYS21*. The same allele also associates with higher expression of three non-*LYS* genes containing *Lys14p* binding motifs (*DYS1*, *TOP2*, *DAD2*), and the *Lys14p* motif-containing *LYS12*.

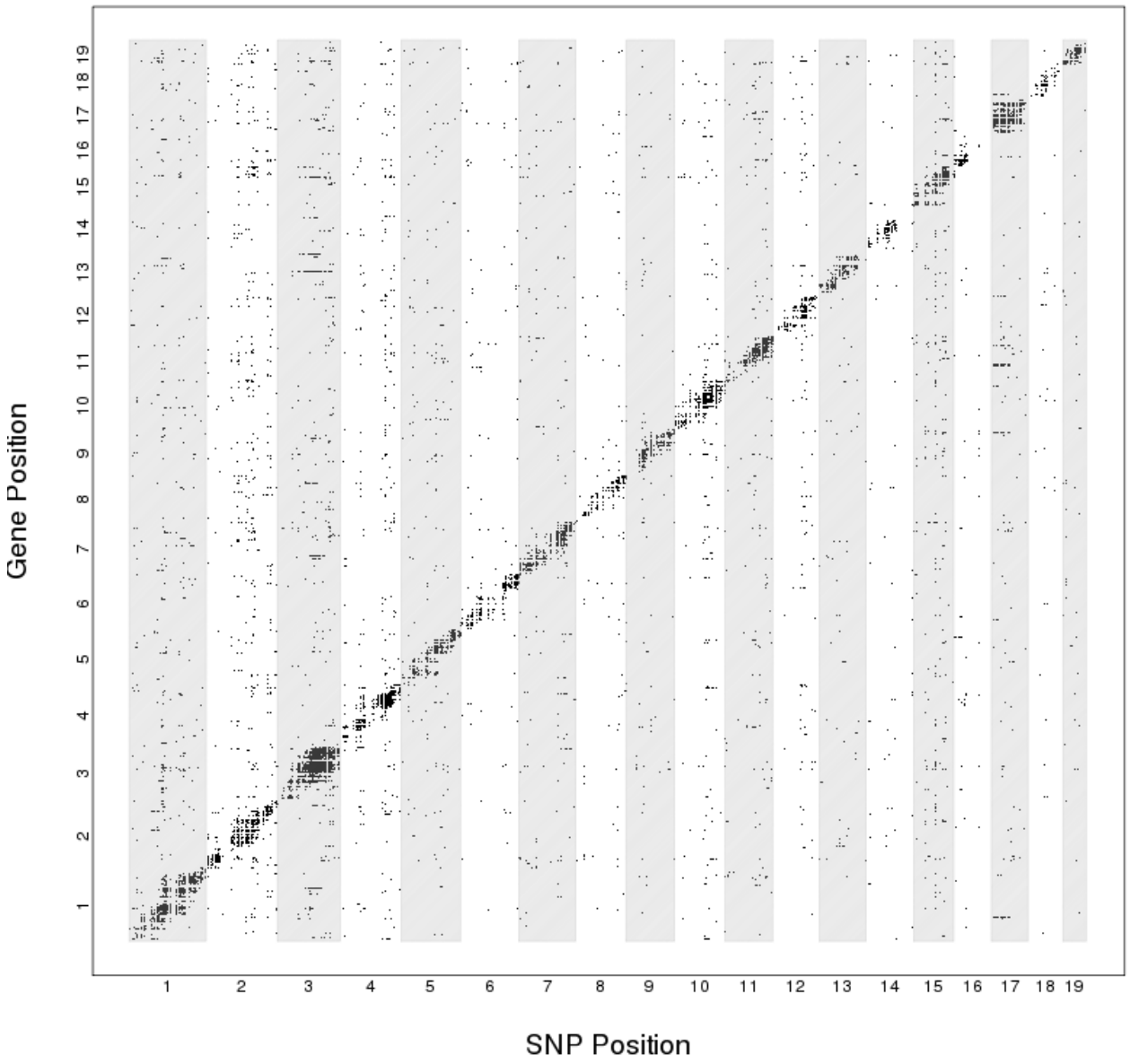


Figure 2.9. Distal eQTL associations in pre-Collaborative Cross mice. X axis gives the genomic coordinates of marker SNPs; Y axis represents gene position. Each dot represents a significant marker-gene association at FDR < 0.05, for markers that were at least 1Mb from the associated gene.

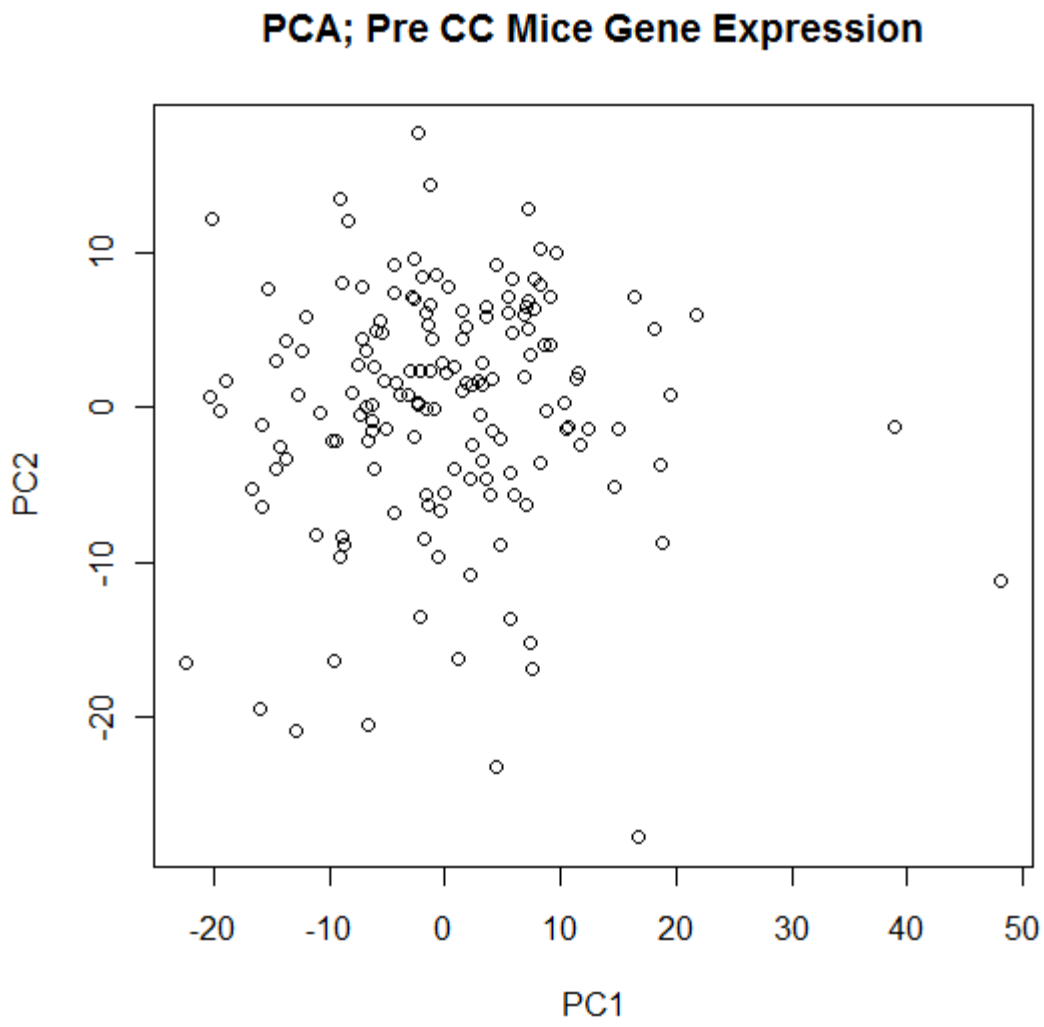


Figure 2.10. PCA analysis for pre-CC mice. Top two principal components for gene expression data in 156 pre-CC mice.

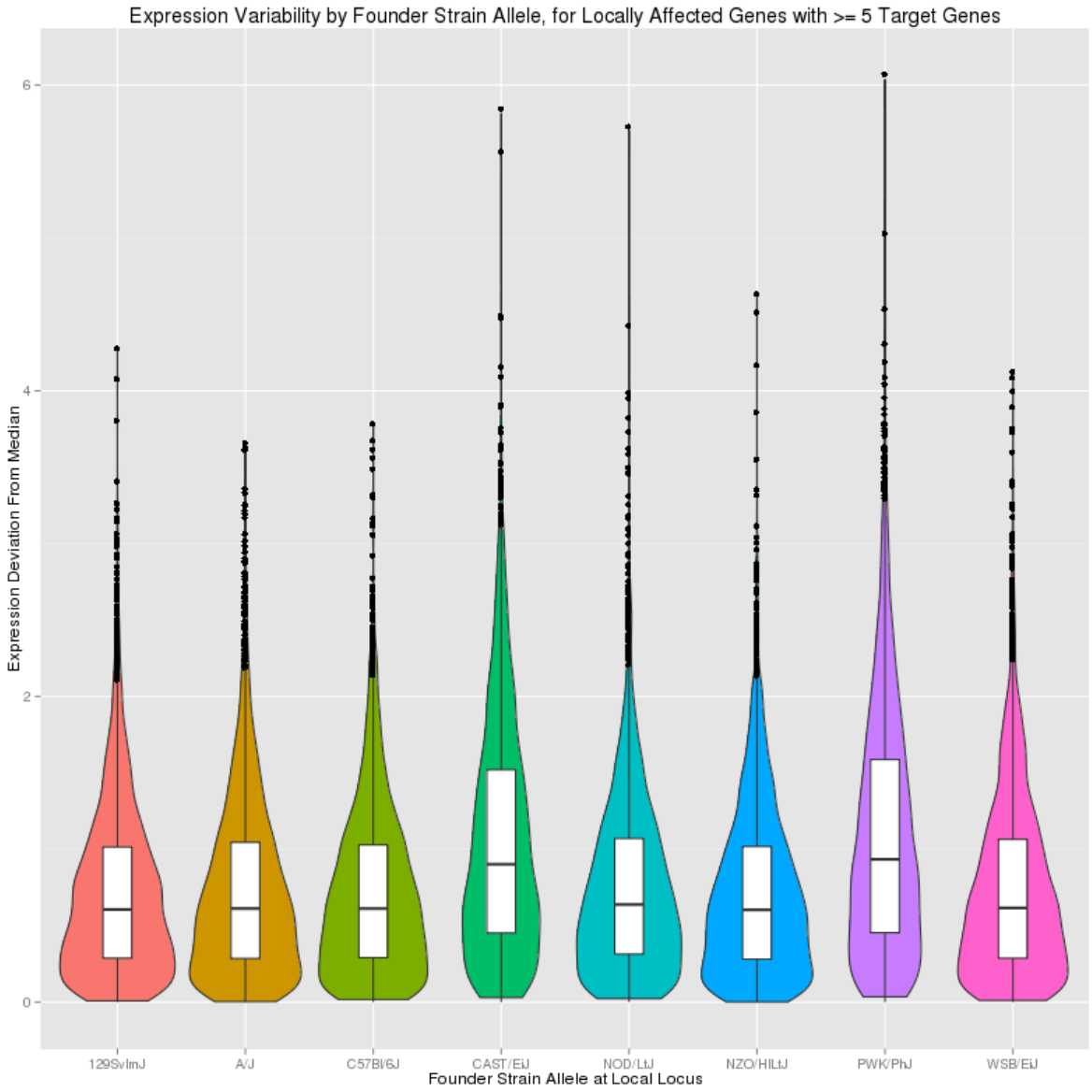


Figure 2.11. Expression variability by founder strain, for locally-regulated genes with at least 5 distal targets. Gene expression values were binned according to the genetic background of the locally-affected gene. Violin plot shows the level of variation compared to the overall sample expression medians, for each of the eight founder strains.

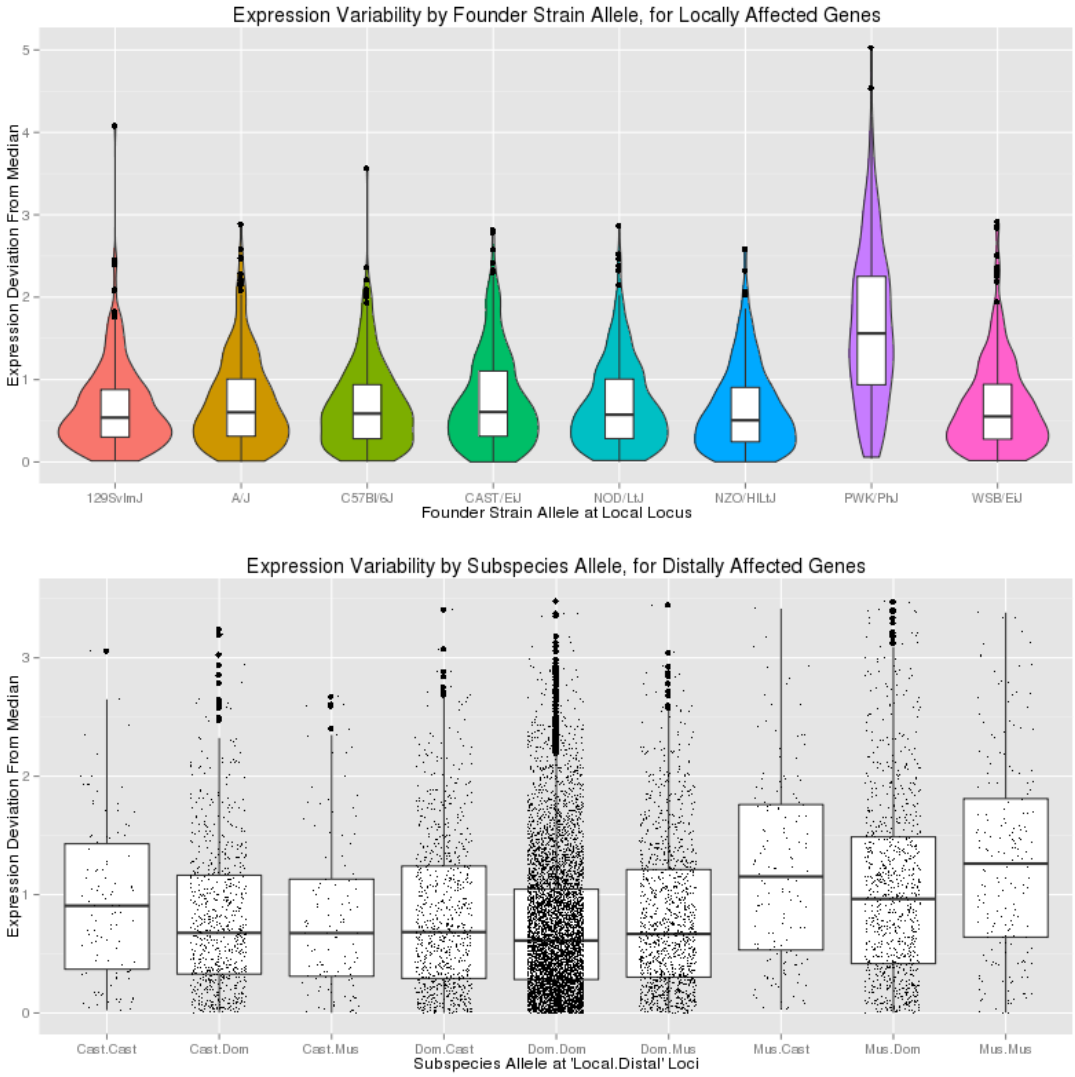


Figure 2.12. Expression variability for PWK-driven *trans*-acting factors and target genes, in pre-Collaborative Cross mice. Top: Distribution of absolute expression deviation from median, for putative *trans*-acting factors with a PWK-driven local eQTL, grouped by founder strain genetic background at the eQTL locus. Only putative *trans*-acting factors that were linked to at least 5 target genes on a different chromosome were considered. Bottom: Expression distribution for target genes of PWK-driven eQTL loci, stratified by subspecies of origin allele (*castaneus/domesticus/musculus*) at both the local and distal loci. Each boxplot represents the expression deviation for all target genes, for each possible combination of local/distal alleles.

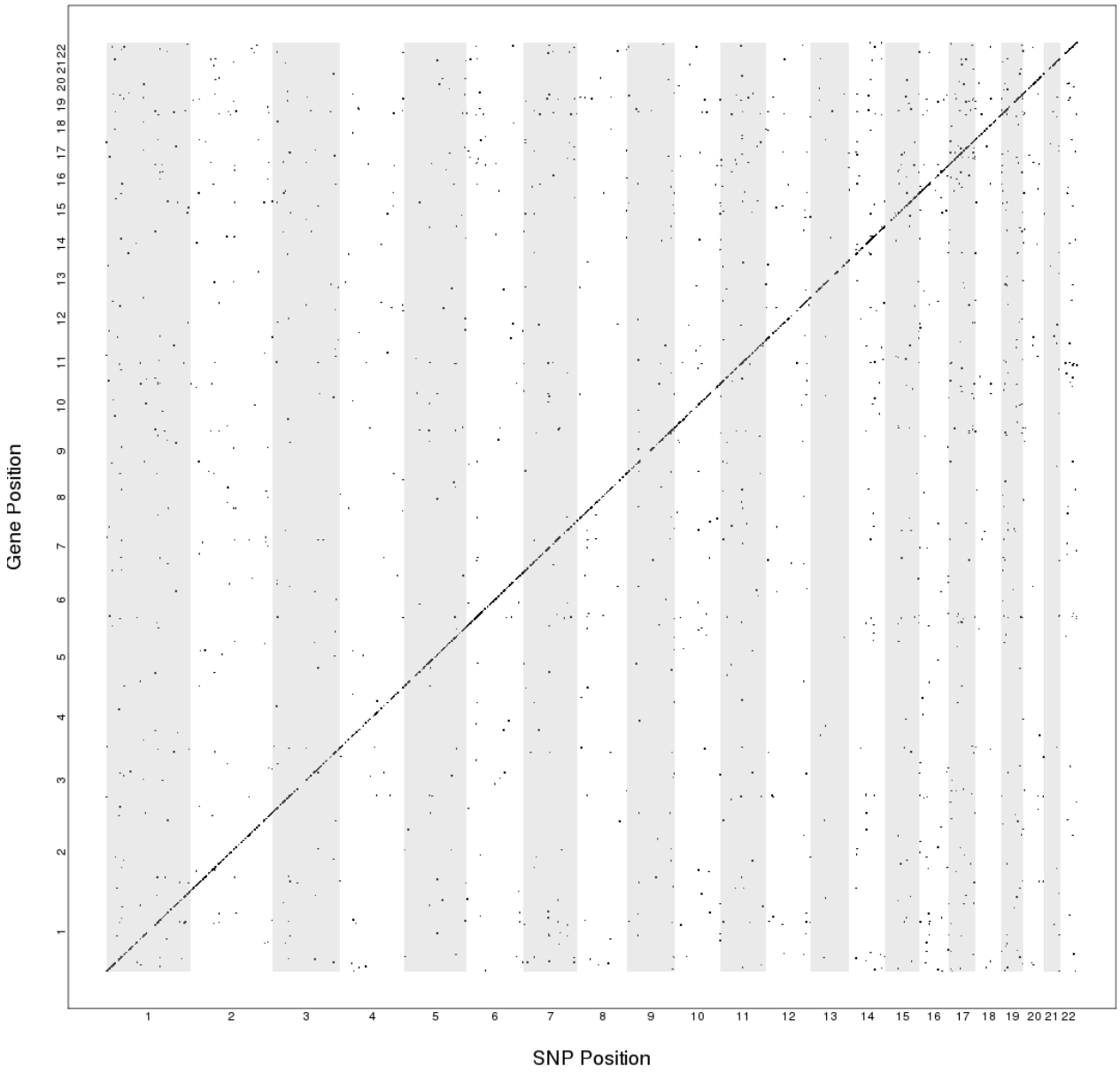


Figure 2.13. Local and distal eQTL linkages in human lymphoblastoid cell lines. X axis shows the genomic coordinates of SNPs; Y axis represents gene position. Each dot represents a significant marker-gene association at FDR < 0.1.

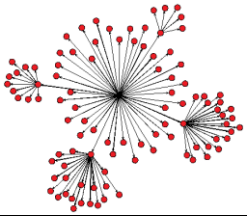
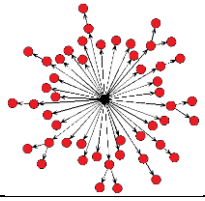
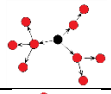
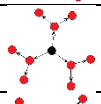
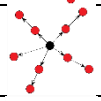
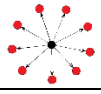
Module Topology	Mean(FracTP)± sd(FracTP)	Mean(FracFP)± sd(FracFP)
	0.94±0.018	0.092±0.0047
	0.79±0.015	0.16± 0.010
	1±0	0.13± 0.025
	1±0	0.14±0.022
	1±0	0.17±0.036
	1±0	0.22± 0.075

Table 2.1. Sensitivity and specificity of partial correlation detection, for simulated gene expression modules. Column 1 shows the mean and standard deviation for the fraction of true edges detected, for ten simulated data sets. Column 2 estimates the intra-module false-edge detection rate. For each module, the ratio of false positive edges detected to the total number of possible false edges is reported.

Method	True Positive	False Negative	False Positive
NetLIFT	442 (100%)	0	20
AllvsAll	442 (100%)	0	20
Trigger	442 (100%)	0	1653

Table 2.2. Detected local eQTL effects, by method. FDR cutoff was set to 0.05. Counts are pooled for all 10 simulated data sets.

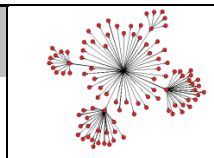
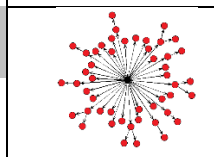
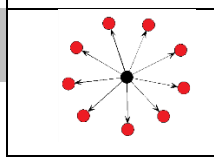
Method	Module	Num Associations Needed to Attain FWER 0.05	Mean Number of Associations Across Ten Simulations	Hotspot Detected
NetLIFT		3	66.4	10/10 (100%)
AllvsAll		1	1.6	4/10 (40%)
NetLIFT		3	39.3	10/10 (100%)
AllvsAll		1	0.9	6/10 (60%)
NetLIFT		3	9.0	10/10 (100%)
AllvsAll		1	0.2	2/10 (20%)

Table 2.3. Hotspot detection rate for gene modules with eQTL at hub gene, in ten simulated data sets. A null distribution of maximum linkage counts were derived from the permuted data sets, with upper 95th quantile for each method listed in column 3. The mean number of identified associations for each module across all ten (non-permuted) data sets is listed in column 4.

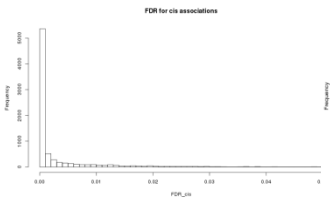
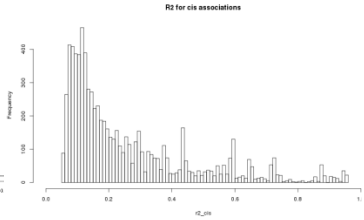
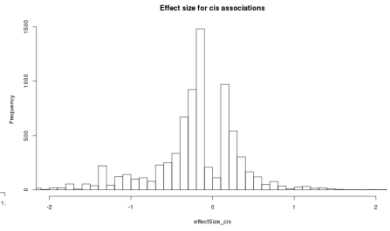
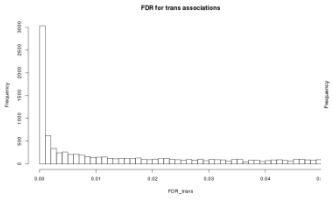
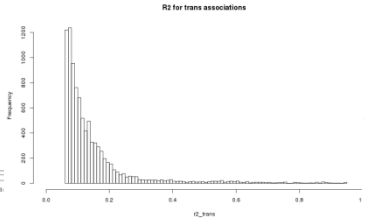
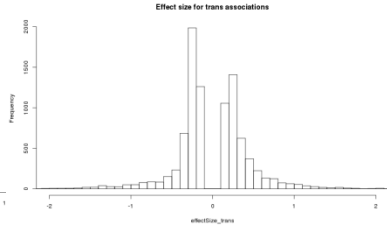
	Number (%)	FDR Distribution	R2 Distribution	Effect Size Distribution (β)
Local	1124 (19.9%)			
Distal	1642 (29.1%)			

Table 2.4. Distribution of eQTL effects for local, distal eQTL, in 112 haploid yeast segregants using NetLIFT method (FDR < 0.05).

Pvalue	Term
2.00E-06	asparagine catabolic process
5.89E-06	cellular response to nitrogen starvation
5.89E-06	cellular response to nitrogen levels
4.66E-05	asparagine metabolic process
4.90E-05	glutamine family amino acid catabolic process
0.000172	aspartate family amino acid catabolic process
0.001328	cellular response to nutrient levels
0.001784	response to nutrient levels
0.001784	cellular response to extracellular stimulus
0.001784	cellular response to external stimulus
0.002359	response to external stimulus
0.002359	response to extracellular stimulus
0.003704	cellular amino acid catabolic process
0.003936	developmental process involved in reproduction
0.004111	cellular response to starvation
0.005043	response to starvation
0.005191	amino acid transmembrane transport
0.005905	carbon catabolite regulation of transcription from RNA polymerase II promoter
0.005931	copper ion transport
0.007164	viral reproduction

Table 2.5. GO annotation enrichment for candidate regulators in yeast. GO analysis was performed for genes with >= 10 distal associations; top 20 enrichment terms reported in right column.

TAF	Chr	Start Pos	End Pos	SNP Pos	Growth FDR	eQTL FDR	# Distal Linkages
TDA8	1	13364	13744	7712	NA	0.000789	4
SEO1	1	7236	9017	11638	NA	0.000206	17
BDH2	1	33449	34702	23813	NA	0.007038	17
GPB2	1	39260	41902	29969	NA	2.49E-20	14
BDH1	1	35156	36304	36900	NA	0.000201	9
ACS1	1	42882	45023	42489	NA	4.09E-06	12
YAR028W	1	184886	185590	184405	NA	1.94E-36	4
UIP3	1	183764	184471	185122	NA	1.30E-22	1
MST28	1	188101	188805	187640	NA	8.18E-25	6
PHO11	1	225451	226854	229140	NA	0.017334	5
ECM13	2	136691	137464	142262	NA	5.88E-05	3
PET9	2	163044	164000	163042	NA	0.017557	1
ACH1	2	194125	195705	185438	NA	1.10E-05	4
FUS3	2	192454	193515	199101	NA	4.96E-05	2
PDR3	2	217473	220403	227290	NA	8.06E-26	15
UGA2	2	247012	248505	246129	NA	9.36E-06	1
SCO1	2	310564	311451	301671	NA	0.004078	12
GIP1	2	328369	330090	334022	NA	1.94E-05	52
YBR053C	2	339673	340749	343931	NA	0.020541	29
TRM7	2	364785	365717	368060	NA	0.001228	14
ECM2	2	368582	369676	368991	NA	1.78E-06	16
TAT1	2	376571	378430	376668	NA	3.55E-12	265
TIP1	2	372100	372732	376872	NA	1.06E-11	138
NRG2	2	370035	370697	376872	NA	4.41E-14	32
BAP2	2	373858	375687	380932	NA	0.001518	21
ECM33	2	393118	394854	392138	NA	0.006151	48
TEC1	2	409163	410623	401568	NA	6.37E-05	16
PBY1	2	432030	434291	427675	NA	0.00034	13
POL30	2	424984	425760	427676	NA	0.016151	40
RFC5	2	423759	424823	427683	NA	0.010096	46
LYS2	2	469742	473920	477206	NA	1.68E-38	167
RAD16	2	467242	469614	479164	NA	1.62E-06	25
AGP2	2	499646	501436	499895	NA	8.34E-40	35
TPS1	2	488899	490386	499895	NA	0.000468	28
YBR137W	2	513038	513577	506661	NA	4.06E-05	258
TBS1	2	541203	544487	537314	0.0036	1.86E-17	296
ARA1	2	539981	541015	537314	0.0036	2.22E-08	302
SUP45	2	530863	532176	537314	NA	5.92E-05	303
RTC2	2	536569	537459	537314	NA	0.00022	47
YSW1	2	537870	539699	548401	0.0036	7.80E-17	307
CNS1	2	549765	550922	548401	NA	8.96E-08	8

AMN1	2	556543	558192	555596	0.0036	8.51E-31	307
ICS2	2	553537	554304	555787	NA	0.005304	317
TOS1	2	563198	564565	565216	0.0036	1.71E-08	291
EXO5	2	565718	567475	565216	NA	1.27E-06	10
SSE2	2	573910	575991	565216	NA	0.013449	29
SEC66	2	578359	578979	569414	NA	0.014054	289
UMP1	2	581721	582167	584351	0.0036	2.05E-06	268
NPL4	2	576339	578081	584351	0.0036	2.58E-06	268
GDT1	2	602629	603471	592989	NA	6.47E-05	273
PCH2	2	600548	602355	603790	NA	2.33E-05	291
RPL21A	2	606265	607135	603790	NA	0.013744	50
RPS9B	2	604503	605503	603790	NA	0.000455	1
YBR197C	2	615198	615851	616262	0.006255	7.18E-15	237
COS111	2	629163	631937	620056	NA	0.002184	230
SDS24	2	651410	652993	658746	NA	1.45E-05	13
GPX2	2	707523	708011	697894	0.012127	0.002294	205
GLK1	3	50838	52340	43867	NA	0.00023	1
ATG22	3	54941	56527	64311	NA	0.004837	42
FRM2	3	74704	75285	75021	NA	5.98E-06	106
HIS4	3	65934	68333	76127	NA	1.26E-05	104
NFS1	3	92777	94270	90412	NA	1.92E-20	31
LEU2	3	91324	92418	92127	NA	3.42E-71	113
ILV6	3	104619	105548	105042	NA	5.51E-05	93
RPS14A	3	177496	178216	175799	NA	0.016242	49
MATALPHA1	3	200438	200965	201166	NA	3.66E-48	40
MATALPHA2	3	199542	200174	201166	NA	2.85E-34	28
RSC6	3	214990	216441	210748	NA	5.46E-09	22
AHC2	3	258880	259266	258303	NA	8.80E-11	2
BRE4	4	38868	42245	46292	NA	1.40E-57	4
HO	4	46272	48032	46292	NA	3.46E-54	6
TIM22	4	67984	68607	70901	NA	8.64E-14	6
UGA4	4	84271	85986	89821	NA	1.71E-06	9
HEM3	4	92763	93746	96259	NA	1.13E-26	21
MRPL11	4	98476	99225	96259	NA	0.002088	1
SFA1	4	159605	160765	161196	NA	1.04E-39	4
RPP1B	4	229906	230527	223324	NA	0.020328	3
STF1	4	229171	229431	226317	NA	0.009091	2
YDL124W	4	240259	241197	251013	NA	2.01E-14	9
NSE4	4	272389	273597	273846	NA	1.14E-26	5
YDL012C	4	431106	431515	437147	NA	6.22E-10	4
GAL3	4	463432	464994	465337	NA	1.81E-05	7
MRH1	4	508145	509107	509817	NA	9.19E-30	9
LYS14	4	509735	512107	509817	NA	6.38E-07	5

ENA2	4	531305	534580	527455	NA	2.44E-50	7
ENA1	4	535190	538465	527455	NA	1.35E-49	5
ENA5	4	527420	530695	527455	NA	2.14E-45	6
YDR134C	4	721069	721479	723155	NA	2.86E-16	7
MKC7	4	744309	746099	744522	NA	6.19E-06	4
YDR210W	4	871072	871299	864542	NA	9.80E-06	5
FCF1	4	1149946	1150515	1149761	NA	7.46E-08	18
YDR341C	4	1151798	1153621	1149761	NA	0.001413	4
APT2	4	1344510	1345055	1344670	NA	4.56E-19	3
ITR1	4	1443706	1445460	1455131	NA	4.70E-05	3
STL1	4	1507997	1509706	1500950	NA	0.001783	7
FDC1	4	1512085	1513596	1510883	NA	8.97E-21	4
YEL076C-A	5	4185	5114	5393	NA	0.008386	6
YEL077C	5	264	4097	5393	NA	0.009368	7
YEL076C	5	4464	5114	5394	NA	0.010973	2
DLD3	5	16355	17845	13213	NA	0.015912	7
YEL073C	5	7230	7553	17399	NA	8.65E-13	3
YEF1	5	75944	77431	79647	NA	0.000135	7
UTR2	5	78053	79456	79647	NA	0.000363	5
URA3	5	116167	116970	117056	NA	2.33E-62	28
EDC2	5	222638	223075	218250	NA	0.000758	1
CHO1	5	207643	208473	218250	NA	1.46E-07	3
PHM8	5	225888	226853	222998	NA	7.17E-06	5
KRE29	5	226857	228251	222998	NA	9.82E-05	1
JHD1	5	254655	256133	251267	NA	0.006914	2
SER3	5	322682	324091	321714	NA	0.006742	1
MET6	5	339860	342163	332264	NA	0.003088	37
LCP5	5	414477	415550	420595	NA	0.015375	102
NSA2	5	413390	414175	420595	NA	0.000103	5
FTR1	5	460521	461735	458085	NA	0.010821	7
YER158C	5	488852	490573	483538	NA	0.048671	9
YER160C	5	492851	498119	504714	NA	0.010776	19
SNZ3	6	11363	12259	5853	NA	3.57E-06	2
DDI2	6	9545	10222	5853	NA	0.000104	3
AAD6	6	14793	15431	5870	NA	7.59E-10	5
AAD16	6	14305	14763	5877	NA	1.09E-06	4
AGP3	6	17004	18680	15106	NA	6.16E-10	5
SNO3	6	10301	10969	18384	NA	2.11E-05	2
YFL054C	6	20847	22787	18384	NA	3.77E-19	1
DAK2	6	23423	25198	18384	NA	7.58E-09	1
YFL052W	6	28232	29629	38648	NA	1.07E-05	4
QCR6	6	224314	224757	232259	NA	0.019324	2
HXK2	7	23935	25395	16619	NA	0.012212	5

MTC3	7	73339	73710	69250	NA	0.000218	1
SIP2	7	97342	98589	98231	NA	1.13E-08	15
MCM6	7	117858	120911	117900	NA	1.43E-25	1
OLE1	7	398631	400163	402833	NA	0.001337	5
PRM8	7	402592	403305	402871	NA	3.82E-20	23
MST27	7	403690	404394	410146	NA	6.68E-08	3
RIM8	7	414106	415734	410146	NA	0.020258	9
TIF4632	7	406863	409607	415585	NA	5.89E-12	8
ERG26	7	495457	496506	502131	NA	0.000396	1
GSC2	7	548268	553955	557230	NA	0.002761	7
CLB1	7	703640	705055	711998	NA	0.001244	7
ECL1	7	784228	784863	790857	NA	0.006304	10
TDA10	7	909218	910090	913065	NA	1.52E-32	2
ZPR1	7	915246	916706	916675	NA	0.006239	27
HSV2	7	940872	942218	946196	NA	1.09E-08	5
BNS1	7	951897	952310	952041	NA	0.000953	12
YHB1	7	959908	961107	956838	NA	0.00082	3
MOS2	7	961364	962065	956838	NA	0.006376	4
CPD1	7	984971	985690	985414	NA	7.76E-22	4
SOL4	7	985977	986744	994478	NA	0.002401	3
NOP19	7	995644	996234	995892	NA	1.73E-12	2
SCW4	7	1048805	1049965	1051340	NA	5.31E-12	2
COS8	8	6400	7545	5842	NA	7.31E-10	5
ARN2	8	8298	10211	5843	NA	3.62E-30	3
YHL044W	8	13563	14270	14953	NA	2.69E-23	5
SPO11	8	62959	64155	56246	NA	5.36E-27	28
YHL012W	8	78932	80413	84437	NA	1.60E-08	38
ETP1	8	81960	83717	92978	NA	2.29E-10	24
GPA1	8	113494	114912	111682	NA	5.51E-11	29
ERG11	8	120086	121678	111682	NA	8.28E-09	27
YSC84	8	136874	138448	137221	NA	0.01866	6
MIP6	8	134547	136526	137227	NA	1.45E-08	17
DAP2	8	164971	167427	161987	NA	6.85E-18	3
PIH1	8	176958	177992	176670	NA	4.94E-05	2
YHR033W	8	175541	176812	185012	NA	1.58E-05	4
INM1	8	197391	198278	193175	NA	2.87E-13	3
DOG2	8	192798	193538	193175	NA	1.09E-32	4
YHR054C	8	213187	214251	203246	NA	4.96E-05	4
YHR214W	8	541651	542262	549634	NA	6.92E-05	1
YIL169C	9	23119	26106	21455	NA	1.40E-05	26
YIL168W	9	29032	29415	27026	NA	0.001944	3
YIL166C	9	30938	32566	33795	NA	6.20E-11	30
SUC2	9	37385	38983	38608	NA	0.00186	7

OM45	9	93619	94800	98949	NA	0.019546	3
RPL16A	9	98527	99416	98949	NA	0.001455	6
QDR1	9	134414	136105	133663	NA	1.30E-06	3
AYR1	9	126204	127097	133663	NA	9.19E-07	4
RPI1	9	136651	137874	141014	NA	7.61E-05	21
HIS5	9	142925	144082	141014	NA	3.31E-11	19
COX5B	9	155219	155762	154733	NA	3.19E-08	15
YIL082W-A	9	205632	210129	195965	NA	4.88E-18	14
YIL080W	9	205632	210354	195965	NA	6.29E-05	3
YIL089W	9	195596	196213	196145	NA	2.71E-19	3
AIM19	9	199643	200116	205191	NA	3.75E-10	10
YIL077C	9	214988	215950	214482	NA	0.01171	1
FIS1	9	241305	241772	251495	NA	2.11E-09	2
YIL055C	9	252040	253923	254745	NA	3.51E-05	3
YVH1	9	404870	405964	398074	NA	0.00101	2
MUC1	9	389569	393672	403134	NA	0.000351	2
DCG1	9	412033	412767	410028	NA	0.013272	1
YPS6	9	430494	432107	430910	NA	2.43E-16	1
GTT1	9	423806	424510	430910	NA	0.021989	1
YJL218W	10	21973	22563	22309	NA	5.48E-09	4
OPT1	10	33850	36249	24397	NA	0.000812	4
VTH2	10	11475	16124	24469	NA	0.000189	2
REE1	10	23133	23729	24739	NA	4.29E-14	18
YJL213W	10	32163	33158	34086	NA	0.000662	13
NUC1	10	40194	41183	40238	NA	0.010023	6
CPS1	10	97731	99461	99921	NA	1.18E-10	6
YJL171C	10	99698	100888	101187	NA	4.38E-09	1
YJL163C	10	111662	113329	111890	NA	6.30E-05	5
NCA3	10	193858	194871	204137	NA	4.58E-07	5
YJL113W	10	197912	203324	204137	NA	0.001428	11
YJL114W	10	197912	199156	204137	NA	0.002316	3
YJL107C	10	218848	220011	218798	NA	9.88E-09	4
PRM10	10	217700	218851	218798	NA	1.18E-11	1
SIP4	10	265920	268409	262593	NA	0.025365	17
IKS1	10	328112	330115	325500	NA	9.68E-05	11
YJL045W	10	356018	357922	353027	NA	0.018223	5
MHP1	10	361243	365439	353027	NA	0.005888	1
IRC18	10	376656	377330	372838	NA	6.20E-07	15
TAD2	10	380243	380995	380085	NA	1.46E-30	9
MHO1	10	452422	453438	450338	NA	0.005187	8
YJR015W	10	462713	464245	461201	NA	2.44E-50	2
SPC1	10	458069	458353	461201	NA	1.05E-31	2
BNA1	10	471130	471663	471555	NA	0.000767	6

OPI3	10	572307	572927	575236	NA	3.55E-05	3
TRP3	11	36700	38154	46635	NA	0.001134	6
YKL187C	11	89289	91541	97725	NA	6.63E-10	4
CWP1	11	260776	261495	261779	NA	5.62E-06	1
HEL1	11	471337	472992	468771	NA	2.49E-10	2
SPO14	11	500986	506037	510933	NA	0.000437	4
GAP1	11	514705	516513	522777	NA	2.36E-13	2
UTH1	11	519169	520266	522777	NA	0.000555	3
MET1	11	571254	573035	579819	NA	0.000611	1
MTD1	11	590037	590999	596215	NA	8.47E-11	6
TGL4	11	605275	608007	611765	NA	1.05E-12	2
YKR104W	11	656474	657394	649174	NA	2.45E-25	5
NFT1	11	652718	656374	649174	NA	7.37E-22	5
VBA5	11	658354	660102	649240	NA	7.12E-12	4
FLO10	11	645994	649503	649240	NA	2.63E-14	4
HSP104	12	88622	91348	86369	NA	3.55E-07	7
TPO1	12	84803	86563	86369	NA	9.35E-06	4
SSA2	12	95565	97484	92694	NA	0.000171	1
POM33	12	97996	98835	99261	NA	3.44E-08	9
PUF3	12	122074	124713	126934	NA	1.21E-41	22
YLL007C	12	134301	136298	131338	NA	4.66E-39	1
PSR1	12	129329	130612	131338	NA	3.61E-34	3
PCD1	12	441716	442738	433955	NA	1.34E-16	13
YLR152C	12	442959	444689	450042	NA	0.010395	14
ASP3-1	12	469318	470406	468981	NA	1.19E-67	50
ASP3-2	12	472970	474058	468981	NA	1.57E-49	44
YLR156W	12	472114	472458	468981	NA	4.33E-14	28
ASP3-4	12	486202	487290	489688	NA	3.75E-55	31
YLR159W	12	485346	485690	489688	NA	6.25E-11	11
YLR161W	12	488998	489342	489688	NA	5.77E-07	11
ASP3-3	12	482550	483638	489688	NA	3.07E-69	29
YLR173W	12	502423	504249	500493	NA	1.24E-17	11
PUS5	12	494496	495260	500493	NA	1.97E-10	12
YLR177W	12	511056	512942	501528	NA	1.81E-06	22
DPH5	12	501262	502164	501528	NA	9.92E-05	21
YLR179C	12	514110	514715	514835	NA	7.52E-25	39
HMX1	12	552726	553679	553064	NA	0.035763	15
BNA5	12	605760	607121	607076	NA	8.02E-08	29
TOP3	12	609785	611755	611854	NA	5.23E-05	5
MAP1	12	625170	626333	635380	NA	5.91E-20	16
HAP1	12	646417	650925	659357	NA	2.12E-13	29
GSY2	12	660718	662835	662627	NA	0.000234	33
YLR257W	12	658828	659793	662627	NA	0.002822	4

NEJ1	12	674429	675457	674651	NA	4.18E-19	21
EXG1	12	728957	730303	719857	NA	3.65E-05	23
ACO1	12	735214	737550	744436	NA	0.00399	1
CHS5	12	787664	789679	787505	NA	8.42E-13	7
FKS1	12	809997	815627	808623	NA	2.04E-08	9
GAS2	12	816094	817761	815480	NA	1.49E-49	24
BUD8	12	834351	836162	829265	NA	1.08E-09	5
TAL1	12	836349	837356	829693	NA	0.000655	6
STP3	12	871696	872727	881579	NA	2.20E-07	15
RPS29A	12	898651	898821	899898	NA	0.038439	15
DUS3	12	922442	924448	912976	NA	0.009912	51
CTR3	12	947251	947976	956366	NA	0.00645	18
PUN1	12	953350	954141	956366	NA	9.82E-07	64
YLR413W	12	951153	953180	956366	NA	0.003039	21
MRPL4	12	1014488	1015447	1006711	NA	0.00038	3
CAR2	12	1012498	1013772	1019347	NA	4.16E-07	11
ECM7	12	1022622	1023968	1023790	NA	0.000462	4
SST2	12	1039268	1041364	1042072	NA	0.013477	4
YRF1-4	12	1067085	1071233	1059929	NA	1.79E-12	6
YLR464W	12	1066570	1067499	1067121	NA	4.96E-13	15
YLR462W	12	1065954	1066562	1067121	NA	2.25E-06	8
YRF1-5	12	1072506	1077896	1067121	NA	1.78E-07	7
PHO84	13	24038	25801	28694	NA	0.000639	32
ATR1	13	38196	39824	46070	NA	0.006354	23
YML079W	13	110247	110852	110814	NA	1.25E-11	9
CYB2	13	165533	167308	163328	NA	5.51E-05	9
PPZ1	13	239458	241536	238291	NA	1.54E-14	11
YML002W	13	264541	266754	255486	NA	0.005168	6
MRPL39	13	251304	251516	255486	NA	0.021399	5
GLO1	13	261705	262685	268044	NA	3.46E-06	8
YML003W	13	263483	264355	273244	NA	3.00E-15	7
PLB2	13	277561	279681	277071	NA	1.15E-18	11
HXT2	13	288078	289703	298192	NA	0.0119	10
ARG7	13	395053	396378	395391	NA	0.019466	4
YMR155W	13	568550	570193	564148	NA	1.60E-06	7
YIM1	13	563095	564192	572643	NA	7.44E-11	9
RSN1	13	798517	801378	789466	NA	0.03545	6
YMR321C	13	917577	917894	922258	NA	4.60E-10	2
FET4	13	912878	914536	922258	NA	3.74E-10	1
SNO2	14	12208	12876	13845	NA	0.006868	3
SNZ2	14	13267	14163	13845	NA	7.92E-05	1
SPS19	14	259579	260457	258590	NA	4.09E-06	1
NAM9	14	368597	370057	371953	NA	2.55E-05	25

YNL134C	14	372453	373583	371953	NA	1.49E-07	5
YNL122C	14	398025	398372	402312	NA	0.02306	30
CYB5	14	416942	417304	410244	NA	0.007424	5
NOP2	14	510542	512398	502316	NA	0.005542	16
AQR1	14	503726	505486	502316	NA	7.21E-06	4
POR1	14	517996	518847	525061	NA	3.09E-05	7
COX5A	14	531727	532188	525061	NA	0.003086	8
YNL058C	14	515765	516715	525061	NA	5.47E-07	1
YNL040W	14	553382	554752	549682	NA	7.10E-14	6
YNL034W	14	570479	572317	577299	NA	1.22E-07	7
YNL022C	14	591429	592901	586789	NA	0.013636	6
YNL019C	14	598378	599232	591228	NA	4.66E-13	3
YNL018C	14	599938	601776	591228	NA	2.05E-11	15
MRP7	14	621316	622431	614342	NA	0.000592	1
BDS1	15	6175	8115	1152	NA	7.62E-32	9
YOL163W	15	9596	10105	7699	NA	1.91E-08	10
YOL162W	15	10118	10765	7861	NA	4.29E-09	4
ENB1	15	19490	21310	10529	NA	0.000299	12
HPF1	15	28702	31605	27928	NA	1.36E-11	10
ZPS1	15	34657	35406	37207	NA	0.004242	8
FRE7	15	40747	42609	43051	NA	7.33E-07	12
YOL153C	15	36821	38566	43051	NA	4.83E-05	15
NDJ1	15	116396	117454	108577	NA	5.12E-30	64
SKM1	15	104326	106293	113254	NA	3.58E-15	15
ZEO1	15	110297	110638	116709	NA	2.89E-08	26
SPO21	15	145334	147163	144659	NA	3.42E-23	43
ATG19	15	168727	169974	170945	NA	5.53E-08	25
PHM7	15	162356	165331	174364	NA	3.00E-21	107
YOL019W	15	288899	290554	290670	NA	2.01E-06	12
YOL014W	15	299694	300068	301074	NA	2.46E-17	6
AUS1	15	349679	353863	348934	NA	2.04E-06	6
CRS5	15	389213	389422	382531	NA	4.95E-17	11
ETT1	15	424848	426086	427159	NA	1.50E-11	10
RSB1	15	422669	423733	427159	NA	3.14E-09	9
RAT1	15	418631	421651	427159	NA	2.46E-06	4
CYT1	15	447441	448370	438824	NA	0.003489	16
YOR062C	15	442727	443533	438828	NA	2.65E-15	4
BAG7	15	578565	579794	581277	NA	0.012044	10
RDL1	15	849634	850053	850119	NA	4.37E-20	3
RRS1	15	868339	868950	861655	NA	0.01681	3
YOR304C-A	15	888518	888748	889464	NA	0.020566	11
FIT3	15	1060439	1061053	1065719	NA	0.046055	1
YOR389W	15	1074211	1076085	1065809	NA	0.017658	2

GLR1	16	375499	376950	368296	NA	6.30E-18	2
YPL067C	16	425248	425844	428612	NA	0.002836	3
SUR1	16	451906	453054	462646	NA	5.58E-07	34
SWI1	16	521011	524955	523450	NA	4.34E-19	40
IRC15	16	518732	520231	523450	NA	0.024603	7
SNF8	16	553624	554325	555416	NA	8.56E-23	2
YOP1	16	623524	624199	618581	NA	2.13E-08	3
AQY1	16	921856	922773	927500	NA	0.000119	3
ARR3	16	939918	941132	932535	NA	9.47E-12	4
OPT2	16	924300	926933	932535	NA	0.000184	6
ARR2	16	939275	939667	932535	NA	2.05E-05	1

Table 2.6. Comprehensive list of putative regulators identified in yeast. Columns 1-4: putative regulator, chromosome, and transcription start/stop annotation; column 5: position of local eQTL; column 6: FDR of eQTL variant with growth rate; column 7: FDR for eQTL association; column 8: number of distal genes linked to locally-acting eQTL, via putative trans-acting gene.

Method	eQTL Position	TAF	Previously Predicted Regulators	# Targets	GO Annotation Enrichment	GO pVal	FDR - Growth Rate Association
***	chrII:376668	TAT1	TRM7[79]	265	cytoplasmic translation	9.63E-37	NA
***	chrII:555596	AMN1	AMN1[44,79], MAK5[44]	307	ribosome biogenesis	2.90E-12	0.0036
***	chrII:697894	GPX2	None[44,79]	205	ncRNA processing	1.53E-17	0.012
***	chrIII:92127	LEU2	LEU2[44,79–81]	113	organic acid biosynthetic process	4.05E-25	NA
***	chrIII:105042	ILV6	ILV6[80,81]	93	organic acid biosynthetic process	2.45E-22	NA
***	chrIII:201116	MATALPHA1	MATALPHA1[44,79,80,82]	40	response to pheromone	1.78E-08	NA
***	chrV:117056	URA3	URA3[44,79–81]	28	'de novo' UMP biosynthetic process	8.66E-09	NA
***	chrVIII:111682	GPA1	GPA1[44,65,79,80,82]	29	conjugation	1.14E-15	NA
***	chrXII:659357	HAP1	HAP1[44,65,79,80,82]	29	steroid metabolic process	3.80E-09	NA
***	chrXII:1067121	YLR464W	YRF1-4[79], YRF1-5[79],YLR464[79]	15	telomere maintenance via recombination	1.81E-05	NA
***	chrXIV:371953	NAM9	MKT1[80], SAL1[80]	25	mitochondrial translation	1.55E-21	NA
***	chrXV:174364	PHM7	PHM7[80,81], IRA2[65,82]	107	cellular ketone metabolic process	8.89E-08	NA
***	chrXV:382531	CRS5	CAT5[44,79]	11	cellular respiration	3.77E-05	NA
**	chrI:11638	SEO1	NA	17	monocarboxylic acid metabolic process	1.11E-06	NA
**	chrII:376872	NRG2	NA	32	asparagine catabolic process	1.85E-06	NA
**	chrII:401568	TEC1	NA	16	pseudohyphal growth	1.03E-03	NA
**	chrII:477206	LYS2	NA	167	lysine biosynthetic process via aminoadipic acid	1.27E-07	NA
**	chrIV:96259	HEM3	NA	21	cytokinesis	5.47E-04	NA
**	chrIV:1149761	FCF1	NA	18	endonucleolytic cleavage involved in rRNA processing	4.02E-04	NA
**	chrV:420595	LCP5	NA	102	ncRNA metabolic process	1.90E-13	NA
**	chrV:504714	YER160C	NA	19	DNA integration	6.65E-24	NA
**	chrVII:402871	PRM8	NA	23	cellular zinc ion homeostasis	5.72E-06	NA
**	chrVII:916675	ZPR1	NA	27	ribosome biogenesis	2.56E-05	NA

**	chrIX:33795	YIL166C	NA	30	oligopeptide transport	2.22E-03	NA
**	chrIX:141014	RPI1	NA	21	L-asparagine biosynthetic process	1.34E-05	NA
**	chrX:24739	REE1	NA	18	formate metabolic process	3.32E-08	NA
**	chrX:262593	SIP4	NA	17	mitochondrial outer membrane translocase complex assembly	2.03E-04	NA
**	chrXII:126934	PUF3	NA	22	transposition, RNA-mediated	1.01E-06	NA
**	chrXII:468981	ASP3-1	NA	50	oxidation-reduction process	7.84E-07	NA
**	chrXII:956366	PUN1	NA	64	beta-alanine metabolic process	1.29E-04	NA
**	chrXIII:28694	PHO84	NA	32	negative regulation of catalytic activity	5.17E-05	NA
**	chrXVI:523450	SWI1	NA	40	regulation of DNA metabolic process	2.63E-04	NA
*	chrXIII:149075	NA	SMA2[80]	NA	NA	NA	NA

Table 2.7. Distal regulatory loci and candidate regulators identified in yeast. First column indicates eQTL identified by: previous methods and NetLIFT (***); NetLIFT only (**); previous methods only (*). Third and fourth columns list candidate regulators implicated by NetLIFT, previous methods, respectively. Fifth column gives the number of genes linked to the locus by NetLIFT. Top GO enrichment for linked transcripts listed in sixth column. For eQTL on chromosome 2 that were linked to genes with ncRNA and ribosomal annotation, association testing was performed for the marker and growth rate phenotype (far right column).

Pvalue	Term
0.00116742	malate metabolic process
0.00192771	progesterone metabolic process
0.00192771	negative regulation of nitric oxide biosynthetic process
0.002854168	organic acid metabolic process
0.003725601	carboxylic acid metabolic process
0.004640455	small molecule metabolic process
0.00524957	positive regulation of heart contraction
0.005659446	lipid transport
0.005687178	oxoacid metabolic process
0.006687313	phagocytosis, engulfment
0.006687313	complement activation, alternative pathway
0.007432993	steroid metabolic process
0.008282274	protein targeting to plasma membrane
0.009233885	monocarboxylic acid metabolic process
0.010029798	regulation of the force of heart contraction
0.010029798	C21-steroid hormone metabolic process
0.010037416	cellular response to lipid
0.011642566	lipid localization
0.011925326	natural killer cell differentiation
0.011925326	membrane invagination

Table 2.8. GO enrichments for distal genes linking to PWK-driver eQTL in pre-Collaborative Cross mice. GO analysis was performed for the pooled set of genes that linked to a PWK founder-driven eQTL with at least 5 distal effects; top 20 GO enrichments are reported in right column.

Pvalue	Term
8.27E-05	folic acid metabolic process
0.000759	folic acid-containing compound metabolic process
0.001212	one-carbon metabolic process
0.001766	pteridine-containing compound metabolic process
0.00537	histidine biosynthetic process
0.00537	glycyl-tRNA aminoacylation
0.00537	histidine metabolic process
0.00537	regulation of hippo signaling cascade
0.00537	imidazole-containing compound metabolic process

Table 2.9. GO term enrichment for putative *trans*-acting factors in human LBCs. GO analysis was performed for the set of putative *trans*-acting factors linked to >= 3 distal genes; enrichments at significance $p < 0.01$ reported in right column.

CHAPTER III

Integrative analysis of chromatin and transcriptional landscape in Crohn's disease colon tissue reveals metaplastic cell populations and highlights functional regulatory regions implicated in disease

OVERVIEW

Crohn's disease (CD) is a chronic inflammatory disease of the gastrointestinal tract, characterized by an inappropriate immune response to the enteric microbiota. Numerous genome wide association (GWA) studies have highlighted a strong genetic component to CD, and have to date identified 163 genetic susceptibility loci [83]. However, the majority of disease-associated variants lie in non-coding regions, suggesting their association with disease is mediated by transcriptional regulation rather than a direct effect on protein function. In order to identify causal genes and improve existing treatment regimens, it is crucial to better understand the genome-wide regulatory architecture of CD at the molecular level, and how it differs from normal tissue. To address this question, we performed formaldehyde-assisted isolation of regulatory elements (FAIRE-seq) and RNA-seq in whole colon tissue biopsies from 21 CD and 12 non-IBD individuals. We identified a subset of mostly-CD individuals that display an "Ileum-Like" signature consistent in both chromatin accessibility and transcription, suggesting a high prevalence of metaplastic cell populations in the cecum and ascending colon, particularly in the setting of CD. Among 22 individuals with expression and chromatin profiles representative of colon tissue ("Colon-Like" subclass), we performed an integrative analysis of FAIRE-seq and RNA-seq, finding 751 and 740 regulatory regions specific to CD and non-IBD, respectively, as well as 51/507 genes up/dowregulated among CD individuals. We identified regulatory regions that associate with both

disease status and expression of nearby genes, representing strong candidates for follow up functional studies. In a motif analysis, we identified several candidate regulators specific to each class, including NR2F6 in non-IBD tissue, suggesting a broad role for this transcription factor in maintaining an uninflamed phenotype. Additionally, we found a general enrichment of GWA SNPs in FAIRE-seq peak regions, but little evidence of increased enrichment in disease-specific chromatin peaks, suggesting that risk alleles may associate with disease via trans-acting mechanisms, and/or may be independent of chromatin conformation. These results provide valuable insights into the functional molecular basis of the disease and highlight specific candidate genes and regulators that may play a role in the complex etiology of CD.

INTRODUCTION

Crohn's disease (CD) is a chronic and debilitating form of inflammatory bowel disease (IBD) affecting over 500,000 Americans [84], with highest incidence reported in young adults between 10 and 20 years of age [85]. CD may affect any part of the gastrointestinal (GI) tract from mouth to anus, and is caused by an aberrant immune response to gut microbiota in a genetically susceptible host. This results in mucosal ulceration and inflammation of the GI tract, and symptoms such as persistent diarrhea, rectal bleeding, abdominal cramps, weight loss, and fatigue. Current treatment regimens consisting of immune-suppressants and corticosteroid combination therapy are effective in less than half of patients [86], and roughly 50% of CD patients require surgical resection within 10 years of diagnosis [87].

Numerous cell types have been implicated in the pathophysiology of CD, including T cells, dendritic cells, natural killer cells [83], and macrophages [88,89]. Though the exact mechanism of effect is unclear, disease onset is thought to result from disruption of the mucosal barrier in the intestine, upon which benign gut flora trigger an inflammatory response from immune cells in the lamina propria (LP) that are typically programmed for tolerance to bacterial antigens. In particular,

local macrophage populations that display a circulating monocyte phenotype appear to play a key role in initiating this immune response [88,90,91]. Furthermore, metaplastic Paneth cell populations have been previously reported at high frequencies in various regions of colon, relative to non-IBD individuals [92,93]. Paneth cells are typically confined to the small intestine, and presence in the colon may be primarily driven by chronic inflammation and associated repair and regeneration. However, a Paneth cell role in impaired innate immunity has also been suggested, via a known genetic susceptibility locus at the *NOD2* gene [94], highlighting a possible causal role for increased metaplasia seen in CD.

Major progress has been made in understanding the genetic component of CD in recent years. Genome wide association (GWA) studies have identified 163 loci linked to IBD [83], which is more than any complex disease to date. One locus residing in a gene desert has been shown to regulate gene expression of the colitis-linked gene *PTGER4* in lymphoblastoid cells [34], suggesting a possible mechanism of effect. Additionally, GWA SNPs have been found to be enriched in regulatory regions of numerous immune cell types [95], as defined by DNase I hypersensitive site sequencing (DNase-seq). However, the functional effect for the majority of associated loci remains unknown, as most lie outside of coding regions. A better understanding of the regulatory architecture of CD will be crucial in bridging the gap between genotype and phenotype, and will play an instrumental role in identifying novel therapeutic targets and more effective clinical treatment.

Formaldehyde-assisted isolation of regulatory elements followed by high-throughput sequencing (FAIRE-seq) has been used previously to identify cell type specific regulatory elements [96] as well disease-specific chromatin landscape changes in tissue [97,98]. Integrative -omics approaches combining paired RNA-seq data in the same individuals have highlighted specific mechanistic effects underlying disease, providing context for associations between phenotype and geneitc variation/chromatin profile [97]. In this study, we analyze FAIRE-seq and RNA-seq in

macroscopically uninflamed, ascending colon tissue biopsies from 33 CD and non-IBD individuals in order to better understand the regulatory architecture of CD. Small and total RNA derived from colon tissue has previously been used to identify prognostic microRNA markers that distinguish disease behavior phenotypes in CD [99], and has identified candidate mRNAs [100,101] and long non-coding RNAs (lncRNAs) [102] implicated in CD and IBD. These studies demonstrate the utility of sequencing-based assays performed in whole colon tissue as used to profile disease-relevant molecular signatures in CD. In our integrative analysis, we identify a subset of 10 individuals whose RNA and chromatin profiles suggest an "Ileum-Like" molecular signature, potentially due to inflammation-induced Paneth cell metaplasia. We then restrict analysis to 22 "Colon-Like" biopsies with paired chromatin and gene expression data, where we identify hundreds of candidate regulatory regions and differentially expressed genes specific to disease and non-IBD cohorts, and provide evidence for causal functional effects between them. Additionally, we show that GWA SNPs are enriched in regions of open chromatin in colon tissue, but find no evidence for disease-specific enrichment.

MATERIALS AND METHODS

Patient Population and tissue procurement

Biopsies were taken from resected colon tissue at time of surgery. For CD patients, tissue was biopsied from cecum or ascending colon, at macroscopically uninflamed regions as determined by a resident pathologist. Non-IBD individuals consisted of two individuals with diverticulitis; five colon cancer patients, one of which involved a Hartmann reversal after colon cancer; two individuals with colonic inertia; two adenoma cases, and one patient with a small intestine neuroendocrine tumor. For all non-IBD individuals, complete or partial colonic resection was performed and biopsies were taken from regions of cecum or ascending colon; for cancer patients,

biopsies were taken from sites distal to tumors, and for all individuals, tissue was macroscopically uninflamed at the site of biopsy.

Genotyping and personalized genome construction

For 32 individuals, genotypes were assayed on the Illumina Immunochip in two separate experiments. An additional individual was genotyped on the Illumina Omni Express platform. QC was performed using a standard Illumina cluster file, and imputation was performed with MaCH-Admix [103], using the full phase 1, release 3 vcf annotation for 1,000 Genomes as a reference panel. We constructed personalized genomes for all 33 genotyped individuals by starting with sex-specific genomes, identifying sites where the individual was homozygous for an allele that differed from the reference genome, and substituting the true genotype at that position.

RNA isolation and RNA-seq analysis pipeline

The Qiagen RNeasy kit was used to isolate RNA from flash frozen tissue. Libraries were prepared using the Illumina TruSeq polyA+ kit, and paired end sequencing was performed at the UNC-CH High-Throughput Sequencing Facility (HTSF), producing 50 bp reads. Reads were filtered requiring a quality score of 20 or greater in at least 90 percent of nucleotides, and artifactual reads were removed with the tagdust software [104]. Following QC steps, reads were aligned to sex-specific genomes, which incorporated SNP calls from genotype array and subsequent imputation. Alignment was performed using the "SNP-tolerant" GSNAP software [105], using a k-mer size of 15 and allowing for 2 mismatches per read. Aligned reads were blacklist-filtered using the DAC blacklisted region list defined by ENCODE [5], and were processed into RPKM values using an in-house script with current RefSeq gene annotations. Prior to analysis, RPKM values were incremented with a pseudocount of 1, log-normalized, and batch effects were removed using the ComBat function in the "sva" package in R [106]. Differentially expressed genes were called using a

two-sided t-test on normalized RPKM values, and gene ontology (GO) enrichments were computed using the hyperGTest function in the R package GOstats [107].

FAIRE-seq analysis pipeline

FAIRE was performed as described previously [108]. In samples for which RNA extraction was also performed, the same biopsy was used for both. Sequencing was performed at UNC-CH HTSF using the Illumina HiSeq 2000 platform, generating 50 bp single-end reads. Reads were filtered requiring a quality score of 20 or greater in at least 90 percent of nucleotides, and artifactual reads were removed with the tagdust software [104]. Additionally, no more than 5 reads were allowed to align to a single position. Post-QC, reads were aligned with SNP-tolerant GSNAP software [109] to personalized genomes, constructed as described above, using k-mer size of 15 and allowing 1 mismatch per read. Post-alignment blacklist filtering was performed as described for RNA-seq reads, and peak calls were performed using F-seq [110] using a feature length of 500 and hg19 bff background file created using 50 bp sequences. The full genome was tiled into 300 bp windows overlapping by 100bp, and raw FAIRE-seq read overlaps were computed for each region. Windows overlapping with simple repeat regions and ENCODE DAC blacklist regions were masked from downstream analysis. Window counts were normalized by total aligned read counts for each sample, and batch effect correction was performed in R using ComBat [106].

Preliminary peak union sets were created separately for class-specific cohorts (Ileum-Like/Colon-Like, CD/non-IBD), as well as for the full set of samples by concatenating the list of top 50,000 peaks for each sample in that cohort. Next, consistent peaks were defined for each of the three union sets, by requiring that at least 30% of samples within that cohort have a peak at a given site. Peaks within 10bp were merged using the bedtools merge command with –d 10 option, yielding a final union set of peaks for each cohort. Differential regulatory regions (DRRs) were called using a two-sided t-test performed on normalized window counts for all windows that

intersected a consistent peak within the respective cohort. When necessary, a single representative was selected from each group of overlapping DRRs by selecting the region with highest overall signal.

The average FAIRE-seq signal at transcription start sites of differentially expressed genes was computed using the bedtools coverage command, with aligned reads and TSS annotations as input. Signal values at each base pair were aggregated across all TSSs of interest and averaged separately within cohorts.

To compute enrichment of DRRs near differentially expressed genes, the number of DRRs falling within 50kb of a differentially expressed gene was recorded. For comparison, ten sets consisting of an identical number of regions, selected at random from the full union set of consistent peaks, were used as a measure of the expected value under the null hypothesis.

ChIP-seq analysis

Aligned ChIP-seq reads for histone marks H3K27ac, H3K27me3, H3K36me3, H3K4me1, H3K4me3, and H3K9me3 were downloaded from the Epigenome Roadmap project data portal [7], for both sigmoid colon and small intestinal tissue. Aligned reads for H3K27ac in inflamed and uninflamed sigmoid colon tissue were processed as described in [111] and were downloaded from Gene Expression Omnibus (GEO), accession number 51425. ChIP-seq signal over differential regulatory regions (DRRs) defined by FAIRE-seq analyses were computed by extracting base-pair resolution overlap of aligned ChIP-seq reads with the genomic regions of interest, for 3kb intervals centered at the DRR midpoints. Overlap counts at each index were normalized by sequencing depth and averaged across all regions of interest.

Motif enrichment analysis

The Discriminative Regular Expression Motif (DREME) package [112] and motif comparison tool TomTom [113] were used to discover enriched motifs within DRRs and compare with known canonical motifs, respectively. To control for local sequence variation, 300bp flanking regions on either side of DRRs for a given class were used as control sequences in the DREME enrichment analysis. Motif comparison in TomTom was restricted to the HOCOMOCO (v9) human database. All other options for both DREME and TomTom were set to default parameters.

GWA enrichment analysis

To compute GWA SNP enrichment at FAIRE-seq peaks in colon tissue, we first downloaded 163 Crohn's disease associated SNPs from the NHGRI GWAS catalog [15], and expanded to all SNPs in LD with this seed set at a cutoff of an $R^2 > 0.8$, as defined by a CEU reference panel from 1,000 genomes. The total fraction of overlap for the set of GWA hits and LD buddies was computed for each sample, using the top 50,000 F-seq peaks in that sample. To compute overlap enrichment relative to a null distribution, we mapped each of the 163 seed SNPs to a random SNP in the genome, matched for number of LD buddies, MAF, distance to nearest TSS, distance to nearest TES, and exonic/genic annotation. A total of 1,000 null sets of SNPs and LD buddies were generated this way, and overlap for each of the null sets with each sample's 50,000 FAIRE-seq peak set was computed, yielding a background distribution of expected overlap specific to each sample's chromatin landscape.

RESULTS

To investigate the role of genotype, chromatin accessibility, and gene expression on the etiology of Crohn's disease, we performed genotyping and high-throughput sequencing assays on macroscopically-uninflamed tissue specimens biopsied from colon resections of 21 CD individuals

and 12 non-IBD patients (Table 3.1). Genotypes for 32 individuals were assayed using the Illumina Immunochip; one additional individual was genotyped on the Omni Express platform. Imputation for all samples was performed with MaCH-Admix, and SNPs passing QC were used to construct personalized genomes (see methods). We measured the genome-wide DNA accessibility using formaldehyde-assisted isolation of regulatory elements followed by high-throughput sequencing (FAIRE-seq). Gene expression levels were assayed using the Illumina HiSeq 2500 platform. Following sequencing, reads were aligned to personalized, sex-specific genomes. The F-seq software [110] was used to call peaks of open chromatin from FAIRE-seq alignments in each sample (see methods), and RNA-seq alignments were processed into log-normalized RPKM values.

Clustering analysis of RNA-seq and FAIRE-seq from colon tissue reveals "Ileum-Like" molecular signatures in patient subset

Post-alignment values for reads per kilobase per million mapped reads (RPKM) were called with a personalized script using RefSeq gene annotations, and were subsequently log-normalized and batch-corrected using the ComBat function in the R "sva" package [106]. We noted that hierarchical clustering and PCA analysis of the normalized counts revealed a subset of genes driving a striking separation between samples (Figure 3.1a, b). When classified into two groups based on this stratification, a two sided t-test for differential expression (DE) revealed a total of 468 DE genes at FDR < 0.05. A GO analysis [107] of these genes showed strong enrichment for transport, drug response, and metabolic annotation (Table 3.2), suggesting a potential effect of treatment-specific response. Additionally, many of these genes were implicated in ileum-specific gene expression signatures. We compared our data to a previous study that conducted transcriptomic analyses on several regions of the gut [114]; of the 1,099 genes that were previously found to be differential between ileum and transverse colon (FDR < 0.05), 188 were present in our set of 468, all of which agreed in direction of effect (p=9.2E-128, hypergeometric test) (Figure 3.1c).

To investigate whether these effects were consistent with the chromatin data derived from the same biopsies, we used a 100bp sliding offset to tile the entire genome into overlapping 300bp windows, and aligned FAIRE-seq reads were converted to signal counts in each window, for all samples. Window counts showed a slight evidence of batch effect and were corrected using the ComBat software (Figure 3.2). We then restricted to CD individuals only, and stratified the normalized FAIRE-seq samples into two classes according to "Ileum-Like" (IL, n=8) and "Colon-Like" (CL, n=12) labelings derived from RNA-seq clustering. Using a two sided t-test, we identified windows significantly more open in either class, at a p-value cutoff of 0.01. In an effort to avoid false positives, we required that resulting differential regulatory regions (DRRs) overlap an F-seq peak call in at least 30% of samples within their respective class. Adjacent windows were merged, resulting in a total of 3,077 DRRs, 1,971 of which were more open in CL and 1,106 were more open in IL subclasses. When compared with regions selected at random from the union set of peaks across all samples, DRRs for each class were enriched within 50kb of genes up in the same class, and depleted within 50kb of the opposite class (Figure 3.3a); additionally, when using t-statistics from the DE analysis as a measure of up/down regulation, we found that active DRRs were more often associated with upregulation of nearby differentially expressed genes than with downregulation (Figure 3.3b). To compare our CL/IL-specific regions with known regulatory regions identified in ileum and colon tissue, we used chromatin immunoprecipitation sequencing (ChIP-seq) data from the Epigenome Roadmap project [7], and computed the aggregate ChIP signal for six histone marks around the DRRs for each class. Colon-Like DRRs were enriched for the enhancer marks H3K27ac and H3K4me1 in ChIP-seq data from colon, but not in small intestine (Figure 3.3c). Similarly, DRRs for the Ileum-Like subclass were enriched for the same enhancer marks in small intestine tissue, but no enrichment was observed in colon. We note that no enrichment is observed for the promoter mark H3K4me3 in either class/tissue pair, and that DRRs are depleted near transcription start sites (Figure 3.4), indicating that the regions of differential

chromatin accessibility are primarily located in enhancer regions. Previous studies have shown that cell-type specific chromatin changes are predominantly found in distal regions and are characterized by enhancer-specific histone marks [12,96,115–117].

Interestingly, all but one of the samples classified as Ileum-Like was a CD patient; the lone non-IBD individual was affected with diverticulitis, a condition characterized by the formation of inflamed pouches in the colon. Thus, given the common underlying factor of inflammation in the IL samples, one possibility for the distinct molecular profiles observed is an inflammation-driven presence of metaplasia in the ascending colon. This hypothesis is consistent with previous results showing that Paneth cells, typically found in ileum, are commonly found in the colon of IBD individuals, and are infrequently observed in non-IBD colon [92,93]. Therefore, we propose that conditions of chronic inflammation in the colons of the Ileum-Like individuals – and resulting tissue damage, cellular turnover, and/or metaplasia – may be the underlying cause of this molecular phenotype. Additionally, we note that visual inspection of disease behavior and location in CD patients (Table 3.1) revealed no obvious correlation with IL/CL stratification, suggesting no clear association between metaplasia and disease severity.

Inflammatory pathways are enriched for differential expression between "Colon-Like" Crohn's samples and non-IBD individuals

To investigate genes implicated in CD pathogenesis in the colon, we removed the 10 samples with ileal gene signatures and focused exclusively on the 11 CD and 11 non-IBD "Colon-Like" individuals with paired RNA-seq and FAIRE-seq data. A two sided t-test on log-normalized RPKM values identified 51 and 507 genes up/downregulated in CD, respectively, at $p < 0.05$ (Figure 3.5a). Hierarchical clustering of these genes was similarly reflective of the phenotypic breakdown in samples (Figure 3.5b). We note that when corrected for multiple testing, none of these genes attained an FDR < 0.05; this overall low significance of expression differences may be attributable

to the heterogeneous nature of the disease and its underlying causes, differences in cell populations obtained from individual biopsies, a subtle expression change in the context of uninflamed colon tissue, or some combination thereof. Nonetheless, a GO analysis performed for genes upregulated in CD ($p < 0.05$) revealed strong enrichments for antibacterial response, immune response, response to stress, and inflammation (Table 3.3). A similar GO enrichment analysis for downregulated genes returned top hits for differentiation, stem cell development, and proliferation, suggesting that processes involving normal epithelial cell turnover may be downregulated in CD relative to non-IBD colon tissue (Table 3.4).

We next asked whether chromatin profiles at the transcription start sites (TSSs) of differentially expressed genes correlated with expression profiles between classes. As expected, we observed a higher overall FAIRE signal in non-IBD samples near TSSs of the genes upregulated in non-IBD tissue. Surprisingly, however, no differential signal was observed at the TSSs of the 51 DE genes upregulated in CD (Figure 3.5c). This may be attributable to small sample size and/or known technical issues with FAIRE-seq in highly-accessible regions such as promoters, whereby DNA that is highly depleted in nucleosomes can become overly sonicated and therefore is not sequenced at high levels (personal communication, Terrence S. Furey). We note that H3K4me3 signal at the TSS for these genes is markedly higher in macrophage and monocyte cell types, and lower in epithelial cell derived cell lines (Figure 3.6). This suggests that upregulation of these genes in CD colon tissue is driven by the macrophage cell population present in the lamina propria, where resident macrophages remain transcriptomically poised for in inappropriate immune response to bacteria, even in macroscopically uninflamed tissue.

Colon tissue FAIRE-seq identifies phenotype-specific regulatory regions that co-localize near differentially expressed genes and show differing enhancer potential under inflammatory conditions

To investigate whether the chromatin landscape in colon tissue differs between CD and non-IBD individuals, we used the same "Colon-Like" cohorts as in the RNA-seq analysis, consisting of 11 CD and 11 non-IBD individuals. We used a two sided t-test on FAIRE signal for genome-wide 300bp windows, and restricted differential regions to those that overlapped an F-seq peak in at least 30% of their respective class. At a p-value cutoff of 0.01, we identified a total of 751 and 740 differential regulatory regions specific to CD and non-IBD, respectively (Figure 3.7a). We used the ontology enrichment software GREAT [118] to associate nearby genes with the class-specific regions, and found significant enrichment for genes up/down-regulated in macrophages, peripheral blood mononuclear cells, dendritic cells, T cells, and B cells in response to environmental stimuli (Figure 3.7b). Of particular interest, we discovered enrichment near non-IBD regions for the term "Genes up-regulated in comparison of control macrophages versus macrophages treated with interferon alpha," while the top result for CD DRRs was "Genes down-regulated in comparison of untreated macrophages versus those cultured with M-CSF and IFNG." Taken together, the direction of regulatory effects from these annotations are consistent with the fact that the aberrant immune response in CD is driven by infiltrating macrophages that initiate an inappropriate immune response to the host microbiota [88,90,91].

Next, we sought to identify candidate transcription factors that play a role in regulating IBD-specific genes by binding to regions of differential chromatin. We used the DREME package in the MEME software suite [112] for- de-novo identification of sequences present in the class-specific DRRs; results were compared to the HOCOMOCO database of transcription factor motifs and scored using the Tomtom package in the MEME suite [113]. At an FDR cutoff of 0.2, we identified 3 significantly enriched motifs for each class. In CD regions, top hits were for BARX2, HXC6, and SRY,

while in non-IBD regions, enrichments were identified for NR2F6, ATF5, and CDX1 (Figure 3.7c). Interestingly, NR2F6, the top hit for the non-IBD regions, has been shown to compete with the transcription factor NFAT at promoter of the pro-inflammatory cytokine gene *IL17a* in $CD4^+$ T cells [119], thereby reducing transcriptional output. Enrichment of NR2F6 binding sites in non-IBD FAIRE regions may suggest a more general role for maintaining an anti-inflammatory phenotype in immune cells present in non-IBD tissues.

To determine whether the DRRs for either class were enriched for inflammation-specific epigenetic changes, we downloaded ChIP-seq data for H3K27ac marks in sigmoid colon, which was assayed in both inflamed (colonectomy, ulcerative colitis) and uninflamed (colectomy due to carcinoma) individuals [111]. We then computed the aggregate ChIP-seq signal across the full set of DRRs for each class, normalized for overall read count. In non-IBD regions, we found enrichment of H3K27ac marks in uninflamed colon, which was reduced in the presence of inflammation, suggesting that under inflammatory conditions, these regions undergo a reduction in their regulatory potential. However, no enrichment of H3K27ac marks was observed in either condition for CD regions (Figure 3.7d). These results suggest that DRRs that are closed in CD may be open and actively regulating gene expression in uninflamed, non-IBD colon tissue, but that an inflamed microenvironment reduces the regulatory capacity. On the other hand, the decreased accessibility of these regions in CD individuals may reflect a baseline state that retains the epigenetic features of inflammation, even in the macroscopically uninflamed setting.

Finally, we investigated the spatial relationship between chromatin landscape changes and differential expression. For each class, we computed the fraction of DRRs within 50kb of a DE gene ($p < 0.05$). When compared with regions selected at random from the union set of peaks across all 22 individuals, we found that CD DRRs tended to be closer to genes upregulated in CD than expected by chance, and were depleted near genes upregulated in non-IBD individuals (Figure 3.7e). Similarly, non-IBD DRRs were located closer to genes upregulated in non-IBD, and were

depleted near TSSs of CD genes. These results suggest a functional link between chromatin conformation and gene expression in CD, whereby regions of open chromatin tend to associate with upregulation of nearby genes.

Colon tissue open chromatin is enriched for Crohn's Disease GWA SNPs

IBD risk loci have been shown to be enriched in both colon tissue eQTL data [120] and DNaseI hypersensitive sites in immune cells [95,121]. Here, we asked whether GWA hits for CD are enriched in regions of open chromatin in colon tissue, and whether this enrichment is preferential for regions found only in CD individuals. We obtained a list of 163 Crohn's Disease-annotated SNPs from the NHGRI-EBI GWA catalog and mapped them to their respective sets of LD buddies, using the HapMap Central European ancestry (CEU) reference panel and an R^2 threshold of 0.8, resulting in a total of 3,179 SNPs. For each sample, the top 50,000 peaks called with F-Seq [110] were overlapped with the candidate GWA set, resulting in an overall overlap score. To create a null distribution for comparison, we created 1,000 sets of randomly-chosen SNPs, matched to the 163 SNP CD seed set for number of LD buddies, MAF, distance to nearest TSS, distance to nearest TES, and exonic/genic annotation. These null sets were expanded to sets of LD buddies based on the same criteria as the CD GWA SNPs, and were then overlapped with each of the sample peak sets. In most colon tissue samples, we observed an enrichment of CD GWA SNPs in peak regions of open chromatin compared to the expected scores under random chance (Figure 3.8a), suggesting that whole colon is a relevant tissue type for investigating the mechanism of disease. However, this enrichment was not specific to CD samples, and was similarly observed in non-IBD individuals. When restricting to peaks found exclusively in CD or non-IBD individuals, we observed only modest enrichment of GWA SNPs in CD-specific peak regions (Figure 3.8b), suggesting that the chromatin conformation alone at GWA loci may not be sufficient to contribute to disease. Additionally, when pooling all peaks across samples and performing overlap analysis stratified by consistency of the

peak call across samples, we found enrichment at all cutoffs, but no increase of enrichment as peak consistency increased (Figure 3.9).

Correlation between FAIRE-seq signal and expression level of DE genes identifies candidate functional regulatory elements implicated in disease

In order to prioritize the most significant regulatory regions for follow up study, we annotated linkages between differential regions of open chromatin and nearby target genes that showed evidence of differential expression. For each DRR, we pooled all 22 samples and computed the Spearman rank-based correlation between FAIRE signal and expression level of all DE genes within 1Mb. To assign statistical significance for each chromatin/expression association, we computed Spearman correlations under 1,000 permutations of the sample labels. Furthermore, since the number of DE genes located proximal to a DRR will vary randomly, some DRR will be tested against many more genes than others. Therefore, we corrected for multiple testing bias separately for each DRR using a Benjamini-Hochberg FDR correction. We identified a total of 72 regulatory-region-to-gene mappings, under the criteria that the target gene be differentially expressed at $p < 0.1$, and the FDR for the chromatin-to-expression linkage be less than 0.1. Of these linkages, 25 and 47 were specific to CD and non-IBD-specific DRR, respectively. We note that only 63.9% (46 out of 72) of the associations between DRRs and DE genes appeared to be associated with upregulation; the remaining associations implied a downregulation of the target gene (Table 3.5). These region-gene pairs represent excellent candidates for follow-up experimental validation, based on their high likelihood of functional association with CD.

DISCUSSION

In 10 out of 33 tissue samples biopsies from colon, we found strong evidence of ileal molecular signatures, underscoring the previously-described heterogeneity of colon tissue,

particularly among IBD individuals. Among CD samples, no evidence of association was observed between presence of metaplasia and disease subtypes. Of note, the chromatin profiles for "Ileum-Like" individuals differed only at gene-distal, enhancer specific regions with lower overall FAIRE-seq signal, which were globally undetectable in principal components and clustering analysis. Without matched RNA-seq data from the same individuals, identifying this molecular stratification would have been extremely difficult. This result highlights the importance in performing thorough QC at the level of tissue composition when studying IBD in colon tissue, particularly when biopsies are taken from cecum and ascending colon, where metaplasia has been found to be most prevalent [92,93].

Overall, the disease-specific effects we identified in uninflamed colon tissue were far more subtle than the effects distinguishing the "Colon-Like"/"Ileum-Like" classes. This was particularly evident for gene expression changes, where immune-related pathways were only subtly upregulated in colon tissue of CD individuals. One potential explanation is that in a macroscopically uninflamed setting, relevant immune cell types in CD display chromatin landscapes that are molecularly "poised" for initiating a transcriptional immune response, but remain relatively inactive until inflammation occurs. Alternatively, the cell population in CD tissues may consist of slightly higher proportions of disease-specific cell types, such as infiltrating macrophages that are not programed for immune tolerance. Though appreciable expression differences may exist between the macrophage populations within CD/non-IBD tissues, similarities in expression profiles of additional cell types, such as epithelial cells, fibroblasts, and neurons, may dampen the signal distinguishing disease from normal.

In the tissue biopsies molecularly defined as "Colon-Like," we found general enrichment of CD associated GWA SNPs. Interestingly, CD-specific regions of open chromatin showed little enrichment of GWA SNPs relative to regions open only in non-IBD, and enrichment did not increase with consistency of chromatin accessibility across individuals. This highlights the disease-relevance

of regulatory regions defined by FAIRE-seq in whole colon tissue, and suggests that chromatin conformation at a GWA locus may be invariant with respect to presence of a risk allele.

We identified hundreds of regulatory regions more accessible in either CD or non-IBD, which localize near genes annotated for immune cell function. In particular, genes near non-IBD DRRs were associated with upregulation in control macrophages relative to macrophages stimulated with interfereon alpha, a cytokine transcribed at higher rates in CD lamina propria mononuclear cells (LPMC) [122], which include lymphocytes, monocytes and macrophages. Meanwhile, the top enrichment for genes near CD DRRs involved genes upregulated in macrophages by interferon gamma, which is also upregulated in Crohn's disease LPMC [122]. Using a loose threshold of $p < 0.05$, we defined a set of 51 and 507 genes that were up/downregulated, respectively, in CD. Upregulated genes were enriched near DRR regions more open in CD, while downregulated genes tended to co-localize with DRRs that were closed in CD. In conjunction, these result show that alterations in chromatin accessibility are bidirectional in CD; furthermore, regions that "open" in CD more often result in upregulation of nearby inflammation-related genes, while those that "close" tend to downregulate genes that characterize an uninflamed or immune-tolerant phenotype.

Applying motif enrichment analysis to the sequences under DRRs, we identified NR2F6 as a potential regulator in the setting of non-IBD. NR2F6 has been shown to produce an anti-inflammatory response in T cells by reducing transcription of the inflammatory cytokine IL-17A. The loss of chromatin accessibility at these regions in CD raises the intriguing possibility that NR2F6 plays a further role in maintaining immune tolerance by regulating transcription of additional genes.

In summary, we verified that FAIRE-seq performed in whole colon tissue is suitable to identify chromatin landscape changes associated with Crohn's disease. Among hundreds of candidate regulatory regions associated with disease, we highlighted dozens that link to nearby

differentially expressed genes, representing ideal candidates for future follow-up studies. These results provide key insights into the functional mechanisms underlying Crohn's disease, and provide a better understanding of the molecular bridge that connects genotype and phenotype in CD.

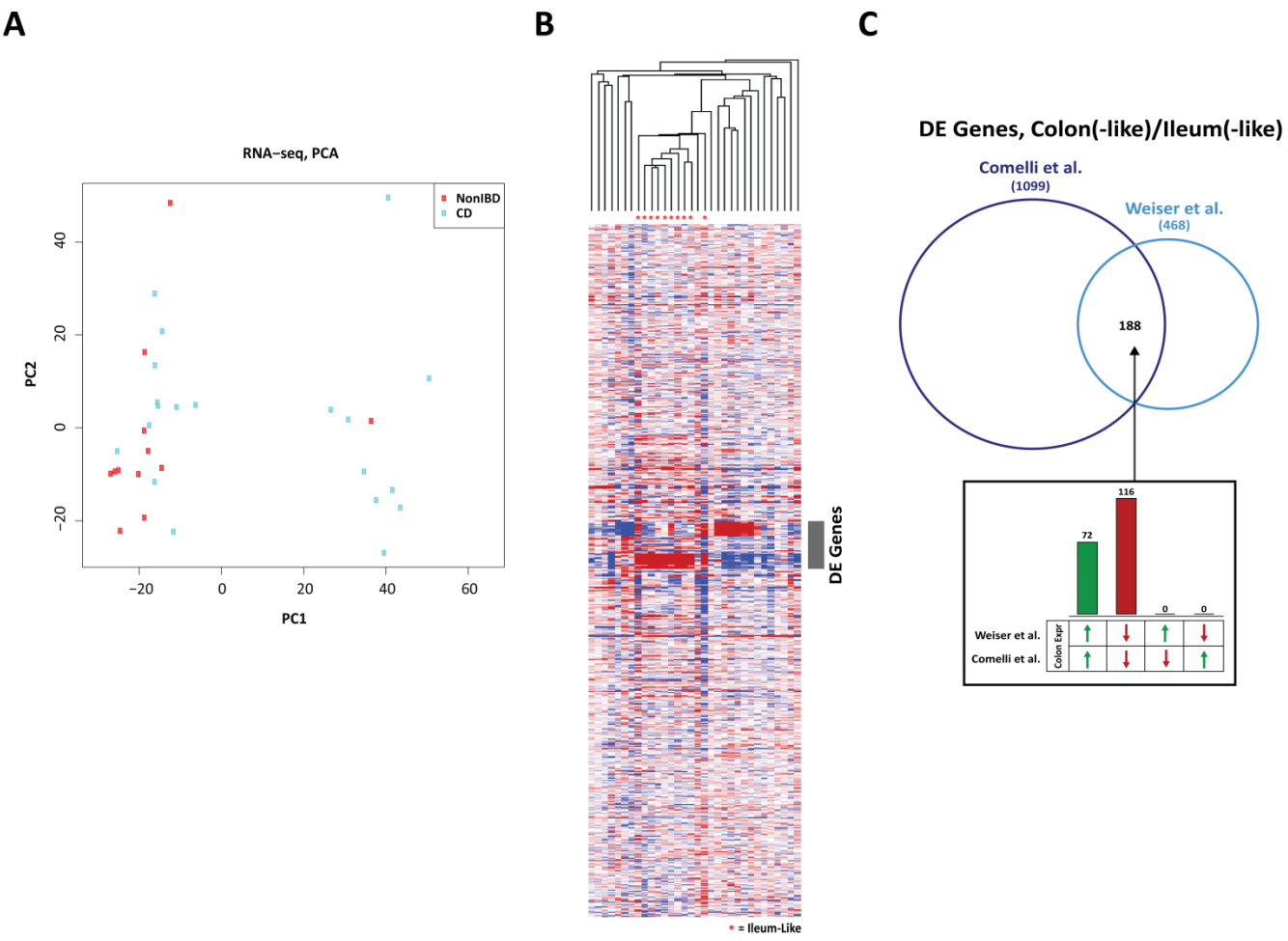


Figure 3.1. Gene expression signatures in colon tissue reveal molecular subtypes corresponding to colon- and ileum-specific transcription. A. Principal components analysis (PCA) stratifies samples into two distinct subtypes. **B.** Unsupervised hierarchical clustering of log-normalized RPKM values. When stratified into subtypes based on PCA, transcripts differentially expressed between "Colon-Like" and "Ileum-Like" strongly separate the molecular classes. C. Overlap and direction of effect for differentially expressed genes compared to Comelli et al. [114] analysis comparing gene expression profiles in ileum and transverse colon.

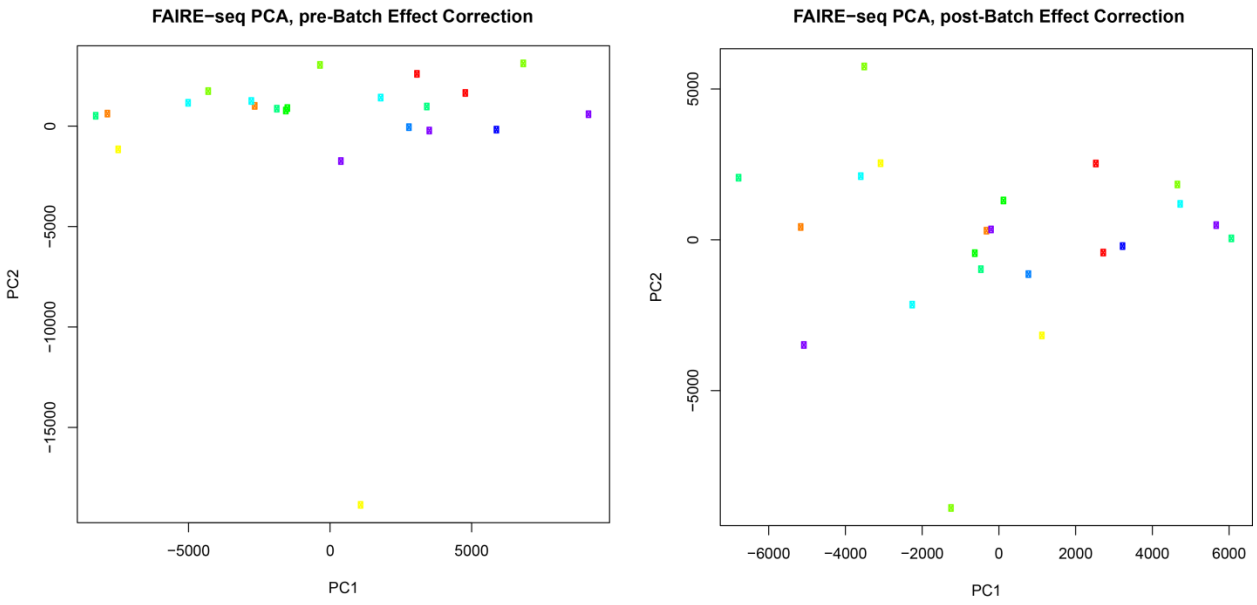


Figure 3.2. Principal components analysis of genome-wide FAIRE-seq signal. A. First and second principal components, for 300bp windows normalized by aligned read count, colored by batch. **B.** Top two principal components after batch effect correction.

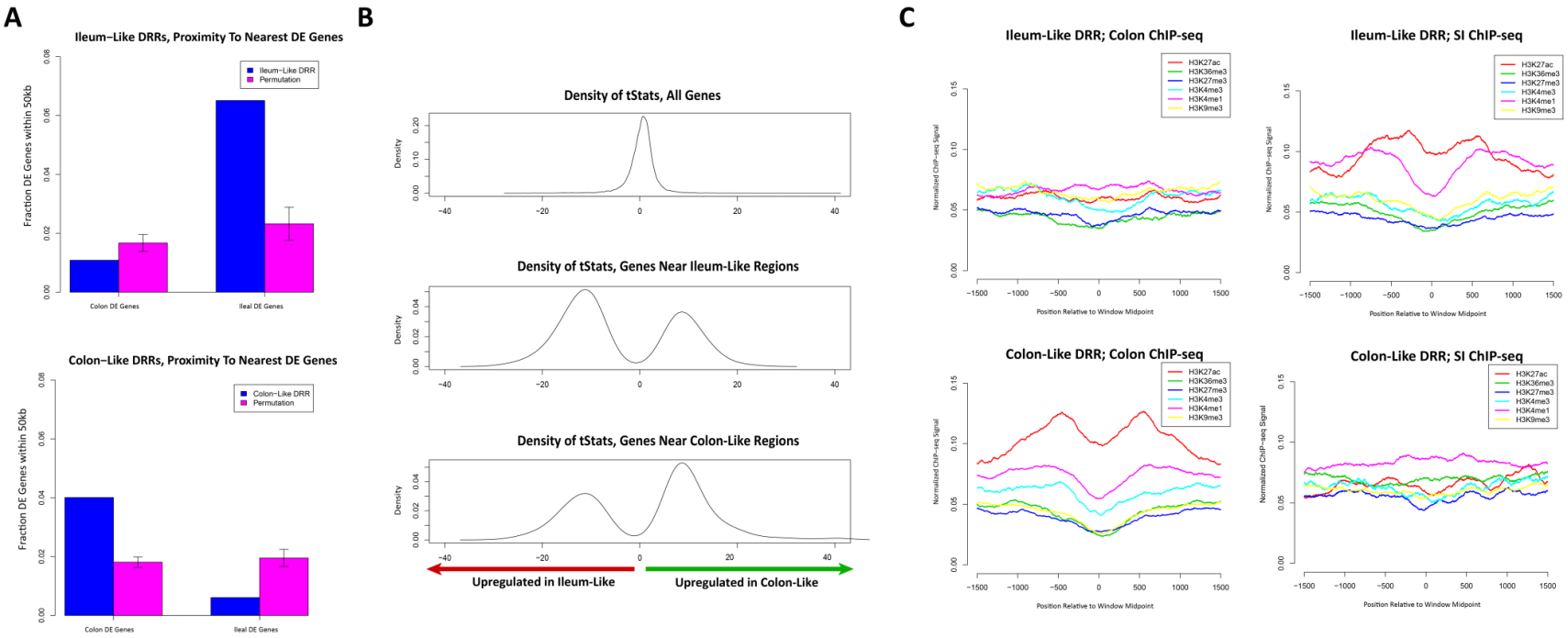


Figure 3.3. Molecular profiles defined by FAIRE-seq correspond to tissue classifications defined by RNA. A. Fraction of genes upregulated in either Colon-Like or Ileum-Like classes that co-localize with differential regulatory regions (DRRs) defined in each class. Actual rate of co-localization for pairwise analyses (blue); co-localization rates expected under random chance (pink). **B.** Density distributions for t-statistics of differentially expressed genes near DRRs defined for each class. DRRs for each class tend to co-localize with a differentially expressed genes that are upregulated in the same class. **C.** Aggregate ChIP-seq signal at DRRs for each class, in both small intestine and colon tissue. Ileum-Like DRRs are marked for H3K27ac enhancer activity in small intestine, but not colon. Colon-Like DRRs show cell type specific H3K27ac marks in colon, but not small intestine.

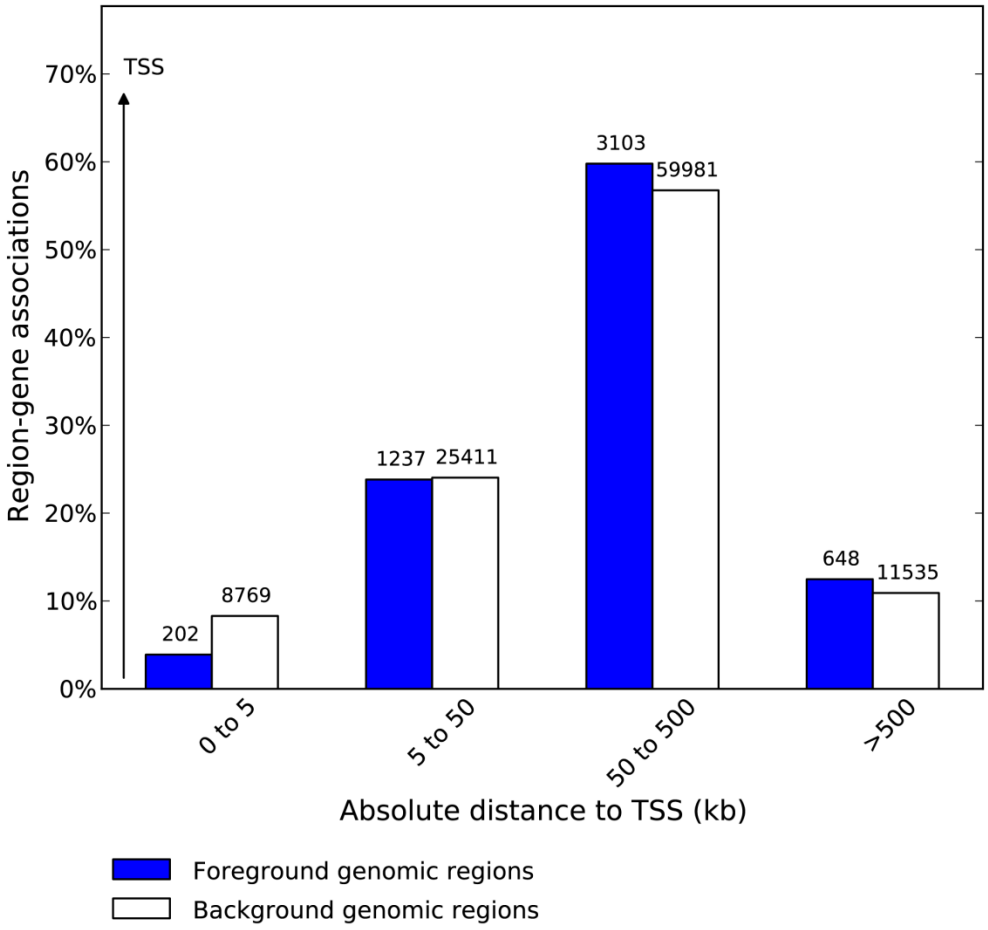


Figure 3.4. Absolute distance to nearest TSS, for differential regulatory regions (DRRs) specific to Colon-Like and Ileum-Like classes. DRRs specific to either class (blue) are depleted near TSSs relative to the union set of peak regions (white).

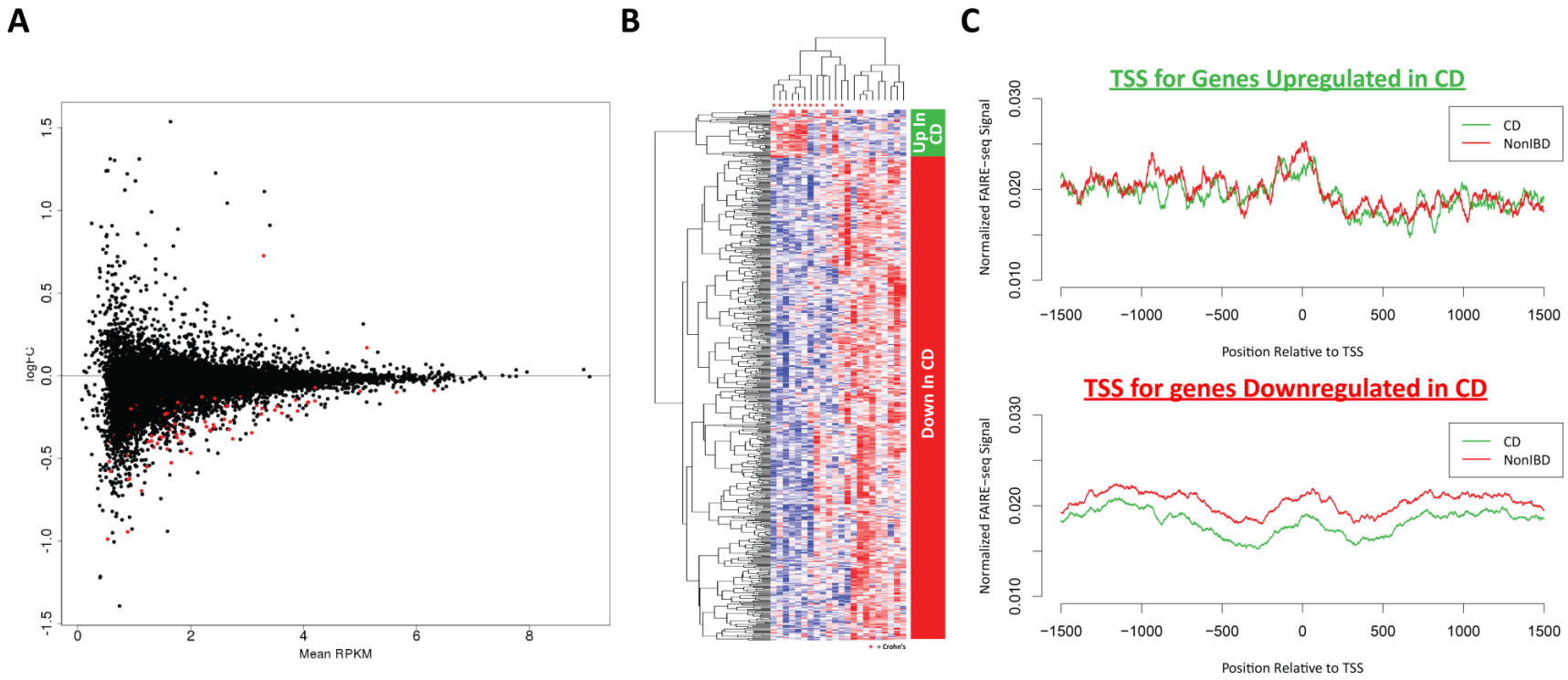


Figure 3.5. Differential gene expression between CD, non-IBD individuals, in Colon-Like subset. A. MA plot of log-normalized RPKM values. Differentially expressed genes ($p < 0.05$) in red. **B.** Unsupervised hierarchical clustering of 51 and 507 genes up/downregulated, respectively, in CD. **C.** Aggregate FAIRE-seq signal at TSSs of upregulated genes (top); downregulated genes (bottom), by disease classification.

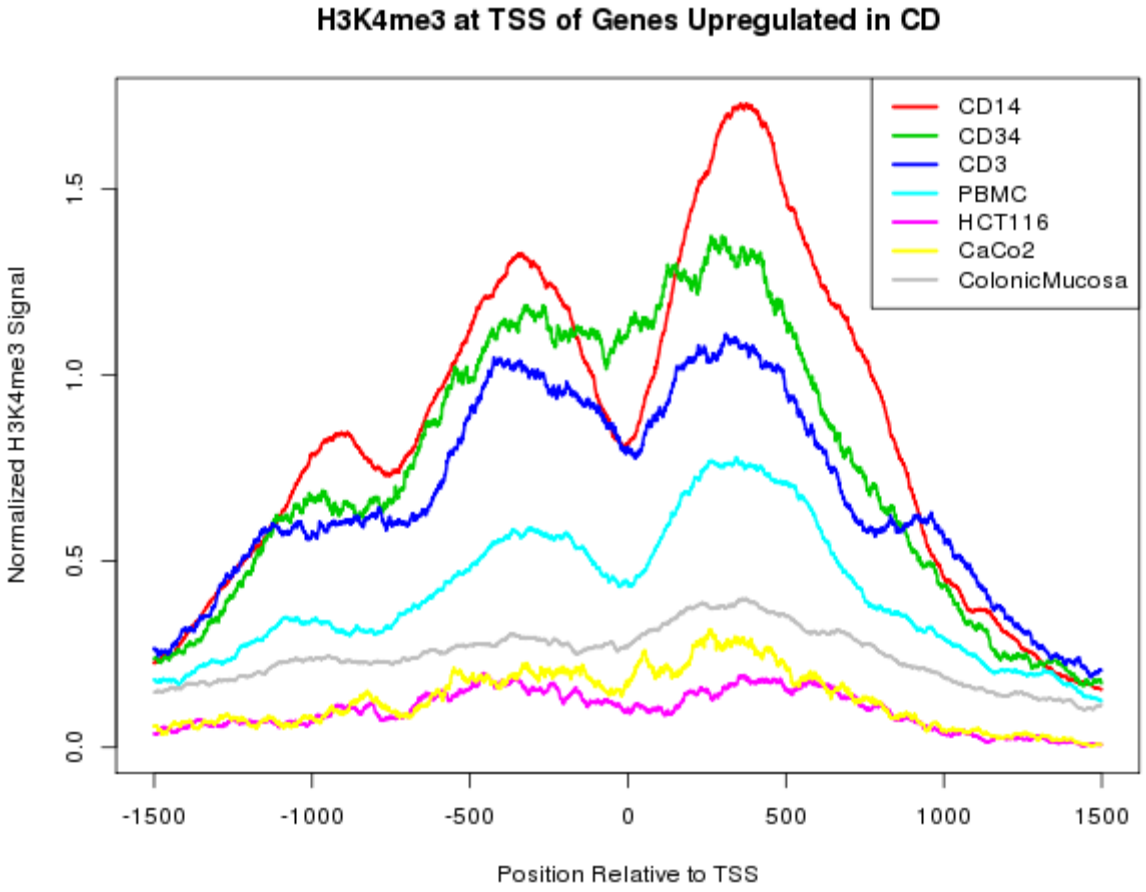


Figure 3.6. H3K4me3 signal at TSS of genes upregulated in CD. Normalized signal for macrophage cell types (CD14, CD34), T cells (CD3), peripheral blood monocytes (PBMC), and epithelial-derived cells (HCT116, CaCo2, ColonicMucosa).

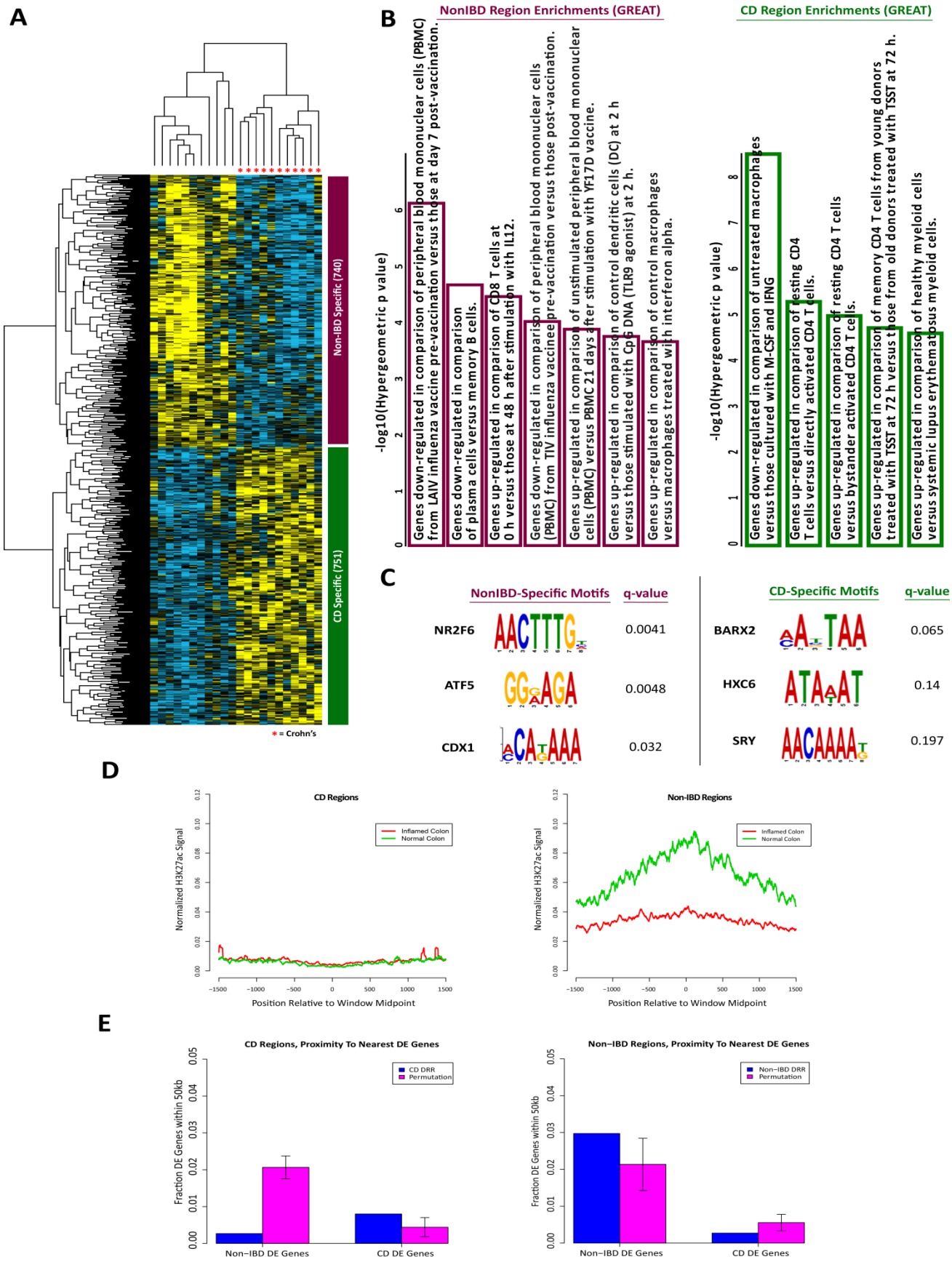

A
Non-IBD Specific (740)
CD Specific (751)
* = Crohn's
B
NonIBD Region Enrichments (GREAT)
-log10(Hypergeometric p value)
Genes down-regulated in comparison of peripheral blood mononuclear cells (PBMC) from LAIV influenza vaccine pre-vaccination versus those at day 7 post-vaccination.
Genes down-regulated in comparison of plasma cells versus memory B cells.
Genes up-regulated in comparison of CD8 T cells at 0 h versus those at 48 h after stimulation with IL12.
Genes down-regulated in comparison of peripheral blood mononuclear cells (PBMC) from TIV influenza vaccine pre-vaccination versus those post-vaccination.
Genes up-regulated in comparison of unstimulated peripheral blood mononuclear cells (PBMC) versus PBMC 21 days after stimulation with YF17D vaccine.
Genes up-regulated in comparison of control dendritic cells (DC) at 2 h versus those stimulated with CpG DNA (TLR9 agonist) at 2 h.
Genes up-regulated in comparison of control macrophages versus macrophages treated with interferon alpha.
CD Region Enrichments (GREAT)
-log10(Hypergeometric p value)
Genes down-regulated in comparison of untreated macrophages versus those cultured with M-CSF and IFNG
Genes up-regulated in comparison of resting CD4 T cells versus directly activated CD4 T cells.
Genes up-regulated in comparison of resting CD4 T cells versus bystander activated CD4 T cells.
Genes up-regulated in comparison of memory CD4 T cells from young donors treated with TSST at 72 h versus those from old donors treated with TSST at 72 h.
Genes up-regulated in comparison of healthy myeloid cells versus systemic lupus erythematosus myeloid cells.
C
NonIBD-Specific Motifs
q-value
NR2F6
0.0041
ATF5
0.0048
CDX1
0.032
CD-Specific Motifs
q-value
BARX2
0.065
HXC6
0.14
SRY
0.197
D
CD Regions
Non-IBD Regions
Inflamed Colon
Normal Colon
Normalized H3K27ac Signal
Position Relative to Window Midpoint
E
CD Regions, Proximity To Nearest DE Genes
Non–IBD Regions, Proximity To Nearest DE Genes
CD DRR
Non–IBD DRR
Permutation
Fraction DE Genes within 50kb
Non–IBD DE Genes
CD DE Genes

Figure 3.7. Differential chromatin accessibility analysis for CD, non-IBD highlights disease-specific pathways and regulatory mechanisms. A. Unsupervised clustering of 751 and 740 differential regulatory regions (DRRs) for CD and non-IBD classes. **B.** GREAT analysis enrichments for genes co-localizing with DRRs in non-IBD (left); CD (right) highlight pathways in disease-related, immune-specific cells. **C.** Significantly enriched motifs within each set of DRRs. **D.** ChIP-seq signal at CD (left) and non-IBD (right) DRRs in inflamed (red) and normal (green) colon. **E.** Fraction of DRRs that co-localize with genes upregulated in either CD or non-IBD classes. Actual rate of co-localization for pairwise analyses (blue); co-localization rates expected under random chance (pink).

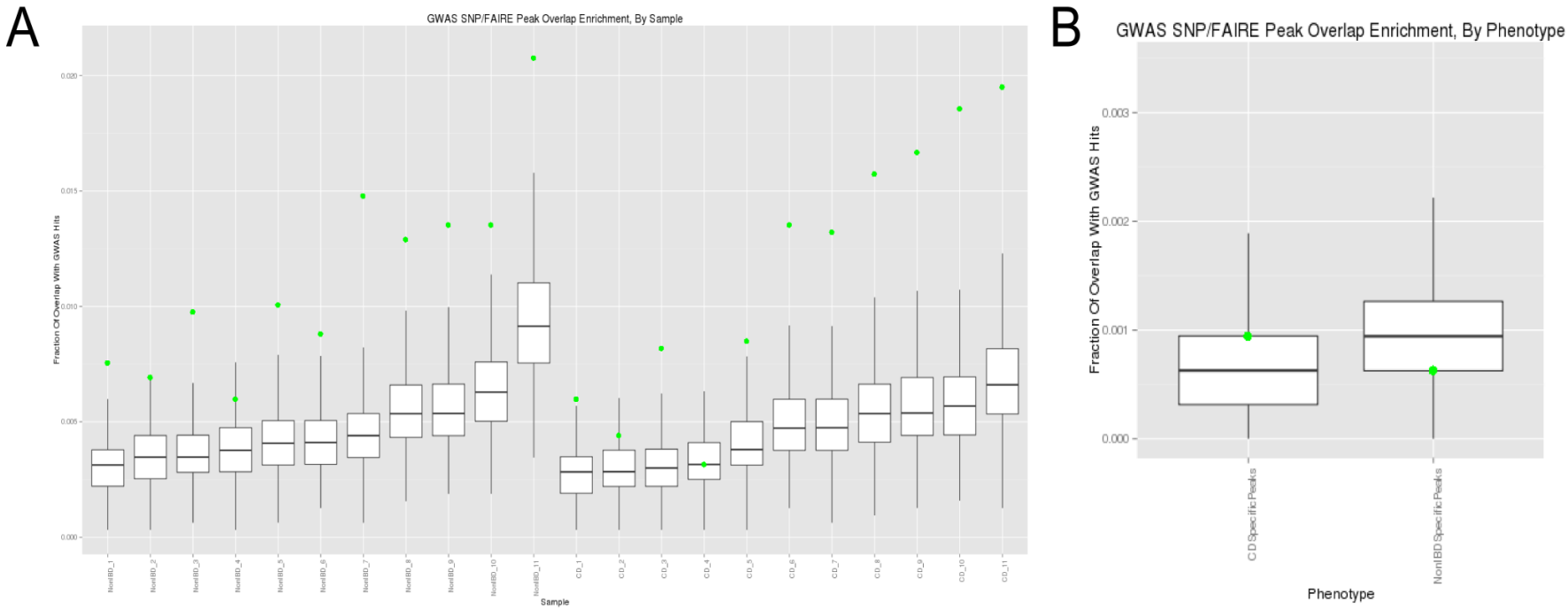


Figure 3.8. Genome Wide Association (GWA) SNP overlap with FAIRE-seq open chromatin regions in colon tissue. A. Green dots represent overlap rate of CD GWA SNPs and LD buddies for top 50,000 peaks defined in each CD and non-IBD individual. Boxplot for each individual represents distribution of overlap counts for 1,000 null sets of paired SNPs. **B.** Overlap rate for GWA SNPs in peaks specific to CD (left) and non-IBD (right) cohorts.

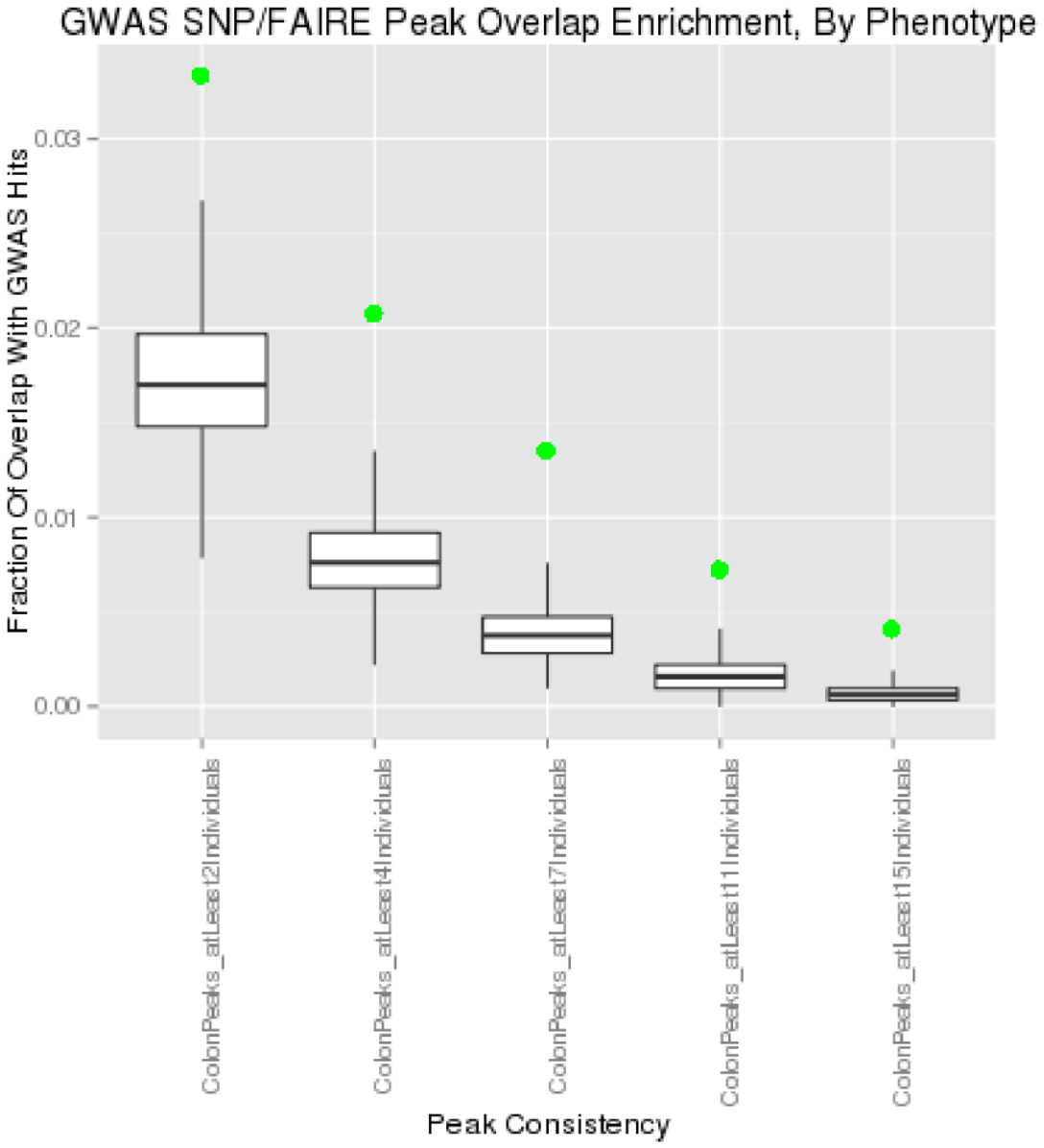


Figure 3.9. Overlap of genome wide association (GWA) SNPs for CD, with peak calls stratified by consistency. Rate of overlap is enriched in all pooled peak union sets, but does not increase in peaks found in higher fraction of individuals.

ID	Disease Status	Phenotype	Genotype	FAIRE	RNA	FAIRE Subclass	RNA Subclass
22	nonIBD	Colon cancer	X	X	X		C
23	nonIBD	Colon cancer (Hartmann reversal)	X	X	X		C
25	nonIBD	Diverticulitis	X	X	X		I
27	nonIBD	Diverticulitis	X	X	X		C
30	nonIBD	Colon cancer	X	X	X		C
32	nonIBD	Colon cancer	X	X	X		C
36	nonIBD	Colonic inertia	X	X	X		C
39	nonIBD	Colon cancer	X	X	X		C
43	nonIBD	Colonic inertia	X	X	X		C
48	nonIBD	Adenoma	X	X	X		C
49	nonIBD	Adenoma	X	X	X		C
50	nonIBD	SI Neuroendocrine tumor	X	X	X		C
20	CD	A1L1B2	X	X	X	I	I
21	CD	A2L2B2	X	X	X	I	I
29	CD	A2L3B3	X	X	X	I	I
51	CD	A2L3B2	X	X	X	C	C
54	CD	A1L2B1	X	X	X	C	C
62	CD	A2L3B3	X	X	X	C	C
63	CD	A1L3B1	X	X	X	C	C
64	CD	A2L3B3	X		X		I
405	CD	A2L2B1	X	X	X	C	C
407	CD	A2L2B3p	X	X	X	C	C
408	CD	A2L2B1	X	X	X	C	C
413	CD	A2L2B1	X	X	X	I	I
420	CD	A2L2B2p	X	X	X	C	C
422	CD	A1L3B2	X	X	X	I	I
424	CD	A2L3B3	X	X	X	I	I
429	CD	A2L3B1	X	X	X	C	C
431	CD	A1L2B1	X	X	X	C	C
433	CD	A2L3B1	X	X		C	
434	CD	A2L3B3	X	X	X	I	I
440	CD	A2L3B2	X	X	X	C	C
450	CD	A2L1B3	X	X	X	I	I

Table 3.1. Data availability, clinical phenotype, and molecular subtype designations for patient cohort. Montreal classification is listed under phenotype for CD individuals; clinical designation and reason for resection is provided for non-IBD individuals. Molecular subclass designations are given by: C="Colon-Like"; I="Ileum-Like." For all CD and non-IBD patients, biopsy was taken from macroscopically uninflamed regions, in either cecum or ascending colon; for cancer resections, biopsy was performed at sites distal to tumor.

Pvalue	Term
8.50E-11	organic anion transport
1.22E-10	anion transport
5.71E-07	response to drug
1.31E-06	drug metabolic process
1.45E-06	exogenous drug catabolic process
3.41E-06	terpenoid metabolic process
9.35E-06	small molecule metabolic process
1.17E-05	organic acid transport
1.17E-05	carboxylic acid transport
1.51E-05	ion transport
1.74E-05	lipid metabolic process
2.20E-05	one-carbon metabolic process
2.20E-05	triglyceride catabolic process
2.64E-05	transmembrane transport
3.10E-05	digestion
3.13E-05	regulation of systemic arterial blood pressure by hormone
3.13E-05	neutral lipid catabolic process
3.13E-05	acylglycerol catabolic process
3.21E-05	bicarbonate transport
3.21E-05	plasma lipoprotein particle assembly

Table 3.2. Top 20 GO analysis results for genes differentially expressed between Ileum-Like and Colon-Like patient subsets.

Pvalue	Term
7.60E-09	immune response
7.77E-09	defense response
1.12E-07	immune system process
1.25E-07	response to external biotic stimulus
1.25E-07	response to other organism
1.78E-07	response to biotic stimulus
4.35E-07	defense response to other organism
1.24E-06	response to bacterium
4.77E-06	inflammatory response
1.20E-05	immune effector process
1.28E-05	humoral immune response
1.36E-05	antibacterial humoral response
1.44E-05	response to external stimulus
1.87E-05	antimicrobial humoral response
2.01E-05	innate immune response
3.80E-05	response to type I interferon
3.80E-05	type I interferon signaling pathway
3.80E-05	cellular response to type I interferon
0.000111	defense response to bacterium
0.000115	response to stress

Table 3.3. Top 20 GO terms for genes upregulated in CD samples.

Pvalue	Term
2.34E-05	stem cell differentiation
2.70E-05	mating
7.02E-05	stem cell development
9.40E-05	muscle cell proliferation
9.45E-05	copulation
9.50E-05	tissue development
0.000262	penile erection
0.000284	negative regulation of developmental process
0.000337	mesenchymal cell differentiation
0.000581	spongiotrophoblast differentiation
0.000581	endothelin receptor signaling pathway
0.000591	smooth muscle cell proliferation
0.000607	reproductive structure development
0.000638	reproductive system development
0.000668	regulation of cyclic nucleotide metabolic process
0.000975	mesenchymal cell development
0.000984	negative regulation of cell differentiation
0.00106	regulation of myeloid leukocyte differentiation
0.001172	sensory perception of pain
0.001187	single organism reproductive process

Table 3.4. Top 20 GO terms for genes downregulated in CD samples.

DRR Chr	DRR StartPos	DRR Stop Pos	Gene	Gene TSS	Gene TES	Gene LogFC
chr1	234659100	234659400	MIR4753	2.35E+08	235353432	-0.38345
chr12	96883000	96883300	NTN4	96184536	96184537	-0.21654
chr17	7589800	7590100	MIR4521	8090262	8090263	-0.94453
chr17	66510000	66510300	ABCA8	66951533	66951534	-0.30759
chr2	28112400	28112700	FTH1P3	27616443	27616444	-0.32693
chr11	14657100	14657400	SPON1	13984183	13984184	-0.38118
chr2	224858400	224858700	SERPINE2	2.25E+08	224896196	-0.39724
chr14	74256800	74257100	DNAL1	74111577	74111578	-0.21512
chr1	16876000	16876300	NBPF11	16940100	16940101	-0.14593
chr7	100143500	100143800	GAL3ST4	99766373	99766374	-0.49859
chr19	6076900	6077200	ZNF557	7069470	7069471	-0.15419
chr20	17517700	17518000	SNX5	17949634	17949635	-0.07242
chr3	24084300	24084600	THRB	24536313	24536314	-0.57633
chr1	28100600	28100900	SCARNA1	28160911	28160912	-0.21578
chr12	40501300	40501600	CNTN1	41302158	41302159	-0.6701
chr19	21950400	21950700	ZNF431	21324839	21324840	-0.12263
chr8	103423000	103423300	KLF10	1.04E+08	103666193	-0.18031
chr13	19447400	19447700	ZMYM5	20437776	20437777	-0.13375
chr7	128200500	128200800	RBM28	1.28E+08	127983963	-0.13014
chr12	15358600	15358900	PTPRO	15475190	15475191	-0.40886
chr21	11017700	11018000	BAGE2	11098925	11098926	0.229919
chr21	11017700	11018000	BAGE3	11098925	11098926	0.229919
chr7	32780100	32780400	RP9	33149002	33149003	-0.12237
chr20	17517700	17518000	PET117	18118498	18118499	-0.07982
chr21	11017700	11018000	BAGE5	11098925	11098926	0.216943
chr21	11017700	11018000	BAGE4	11098925	11098926	0.216969
chr17	25642200	25642500	NOS2	26127555	26127556	0.592232
chr13	79962700	79963000	SPRY2	80915086	80915087	-0.10617
chr10	74837100	74837400	MRPS16	75012451	75012452	-0.07296
chr15	31781000	31781300	ARHGAP11B	30918878	30918879	-0.15675
chr7	100143500	100143800	AP4M1	99699129	99699130	-0.09966
chr14	97263000	97263300	ATG2B	96829678	96829679	0.065612
chr16	21357900	21358200	EEF2K	22217591	22217592	-0.13601
chr17	12159600	12159900	ZNF18	11900689	11900690	-0.12128
chr7	131012100	131012400	PODXL	1.31E+08	131241377	-0.18672
chr10	64564000	64564300	ADO	64564515	64564516	-0.10134
chr1	234659100	234659400	IRF2BP2	2.35E+08	234745272	-0.07284
chr6	56607100	56607400	BMP5	55740375	55740376	-0.19978
chr11	10955600	10955900	SNORD97	10823155	10823156	-0.06888
chr19	21950400	21950700	ZNF708	21512212	21512213	-0.08677

chr15	32836600	32836900	GREM1	33010204	33010205	-0.2245
chr16	28722700	28723000	EIF3CL	28415162	28415163	-0.09543
chr7	86785900	86786200	ABCB1	87342639	87342640	-0.20763
chr1	232764800	232765100	NTPCR	2.33E+08	233086370	-0.10903
chr16	28722700	28723000	EIF3C	28699878	28699879	-0.09825
chr5	54455700	54456000	GZMK	54320106	54320107	0.60093
chr3	15643400	15643700	METTL6	15469042	15469043	-0.09204
chr7	85416900	85417200	SEMA3D	84751247	84751248	-0.62259
chr15	24738800	24739100	SNORD116-24	25339182	25339183	-0.52605
chr15	24738800	24739100	SNORD116-7	25307478	25307479	-0.40905
chr15	24738800	24739100	SNORD116-5	25307478	25307479	-0.40905
chr15	24738800	24739100	SNORD116-2	25299355	25299356	-0.39398
chr15	24738800	24739100	SNORD116-3	25302005	25302006	-0.41332
chr15	24738800	24739100	SNORD116-9	25302005	25302006	-0.41332
chr15	24738800	24739100	SNORD116-22	25335068	25335069	-0.35292
chr15	24738800	24739100	SNORD116-8	25315577	25315578	-0.40899
chr16	57026500	57026800	NLRC5	57050985	57050986	0.237747
chr15	24738800	24739100	SNORD116-16	25327913	25327914	-0.31242
chr15	24738800	24739100	SNORD116-19	25328733	25328734	-0.26733
chr15	24738800	24739100	SNORD116-17	25328733	25328734	-0.26733
chr15	24738800	24739100	SNORD116-14	25325287	25325288	-0.29379
chr2	24664200	24664500	TP53I3	24308085	24308086	0.099092
chr11	3936100	3936400	SNORA54	2985123	2985124	-0.13343
chr15	24738800	24739100	SNORD116-15	25326432	25326433	-0.32917
chr5	125346200	125346500	PHAX	1.26E+08	125936607	-0.08546
chr11	104730900	104731200	CASP1	1.05E+08	104905885	0.191626
chr15	24738800	24739100	SNORD116-6	25310171	25310172	-0.30651
chr11	104730900	104731200	CASP5	1.05E+08	104893896	0.287258
chr14	106869200	106869500	ELK2AP	1.06E+08	106139145	0.523947
chr14	20178400	20178700	RPPH1	20811570	20811571	0.0382
chr17	44458500	44458800	MGC57346	43697711	43697712	-0.14022
chr15	24738800	24739100	SNORD116-23	25336931	25336932	-0.30978

Table 3.5. Differential regulatory regions (DRRs) and candidate target genes. Annotations for CD-open (green) and closed (red) DRRs that link to a differentially expressed (DE) gene within 1Mb, using cutoffs FDR < 0.05 for DRR-to-gene expression linkage, $p < 0.01$ for DRR significance, and $p < 0.05$ for DE significance.

CHAPTER IV

Discussion

Despite each containing the same copy of DNA, the 200 different cell types in a single human body display a broad range of phenotypic variation and specialization. This remarkable level of diversity is achieved by regulating the cell's information content, controlling how and when genes are transcribed. Thus understanding a genome's relationship with phenotype requires a comprehensive model centered on the regulatory architecture.

Recent advances in sequencing technology have enabled a high-throughput approach to studying the molecular signatures of cell lines and tissues, both in human and model organisms. These studies have painted a vast molecular portrait of gene expression patterns, transcription factor activity, chromatin architecture and interactions, histone modifications and DNA methylation across hundreds of different cell types and conditions. Integrative analyses of these -omics data sets have increased our understanding of the functional genome, identifying regulatory regions, gene-gene interactions, and SNPs associated with transcriptional profiles. However, more powerful statistical methods for performing associating tests at the genome-wide scale are necessary to identify genetic variants and epigenetic factors with small or distal effects on target genes. Additionally, understanding the regulatory architecture and how it relates to disease will require an integrative approach in a disease-relevant tissue or cell type. Incorporating multi-dimensional data in disease-specific conditions will not only increase power for identifying trait-specific variants with small effect sizes, but will also provide immediate context for mechanism of effect and suggest novel gene targets for therapy.

In chapter 2, I described a novel method, Network-based, Large-scale Identification oF disTal eQTL (NetLIFT), for detecting genetic variants associated with expression of genes located distally in genomic space. NetLIFT uses pairwise partial correlations between gene expression levels to construct a network representing likely gene-gene interactions, and performs association testing of genetic variants and genes that are most likely to be regulatory targets. By reducing the search space and number of association tests performed, this method increased power to detect distal eQTL, and I was able to discover thousands of previously unidentified SNP-gene associations. Additionally, patterns of linkage revealed insights into gene regulatory networks and effects on phenotype. Among genes predicted to mediate expression effects on distal targets, I found a consistent enrichment for metabolic functional annotation, suggesting that feedback mechanisms within these pathways regulate co-expressed, functionally-related modules of genes. When I applied NetLIFT to gene expression data from mouse liver, patterns of local and distal linkages suggested that many expression effects were dependent on the combinatorial genetic background at the local and distal sites. Furthermore, target genes related to specific genetic background were enriched for annotations related to body weight, suggesting a potential role for these genes in modulating previously described weight differences between strains [63]. The increased power of this method will enable detection of more genetic variants with eQTL effects, and highlight their associations with phenotype. Annotating SNP-gene effects with greater resolution may also lend mechanistic interpretation to SNPs linked to complex traits and disease.

In the preceding chapter, I use FAIRE-seq and RNA-seq data from colon tissue of Crohn's disease (CD) and non-IBD individuals to identify regulatory regions and genes associated with disease. I identified a striking stratification of samples into two molecular subtypes, and showed that the features associated with the "Ileal-Like" subtype display markers specific to small intestinal tissue. Metaplasia of small intestinal cells in the colon has been previously described in CD [92,93]; however, this represents the first study to describe the molecular signature characterizing these

local tissue abnormalities. In an analysis of tissue biopsies classified as "Colon-Like," I identified gene expression signatures specific to disease and non-IBD classes. Notably, transcriptional changes were moderate in effect size, potentially due to heterogeneous nature of the tissue, reduced overall effect in the setting of uninflammed tissue, or both. I then investigated the chromatin landscape in disease/non-IBD cohorts, and found over 700 regulatory regions specific to each class. I showed that these regions are preferentially located near differentially expressed genes, providing evidence of functional association. Additionally, nearby genes were enriched for annotation specific to disease-relevant immune cells, and in particular reflected transcriptional changes identified in macrophages stimulated with pro-inflammatory cytokines. A motif analysis of these regions highlighted three transcription factors relevant to each cohort, implicating possible driver roles for these genes in either maintaining immune tolerance or producing an inflammatory response. These results demonstrated the suitability of using molecular assays performed in whole colon tissue as a means to study the functional regulatory architecture of CD, and provide evidence of the functional effect for many putative disease-associated regulatory regions.

More than ten years have passed since the first genome wide association (GWA) study was conducted. Since then, a flood of additional research has annotated thousands of genetic variants associated with hundreds of complex traits and diseases. However, our understanding of how most of these variants are functionally related to a trait of interest is still only in its infancy. The fact that nearly 90% lie in non-coding regions presents an enormous challenge for the post-GWA era. At a single GWA locus, there are often dozens of candidate SNPs, each in strong linkage disequilibrium with the lead associated SNP. Determining the causal variant and the mechanism of effect requires first identifying a target gene and relevant tissue; once this has been established, reporter assays and electrophoretic mobility shift assays (EMSAs) can be used to experimentally validate allelic effects on gene expression and protein binding, respectively, at the putative enhancer/silencer region. However, this process is extraordinarily time consuming and expensive. Thus, annotations

derived from large-scale consortia such as the ENCODE project and the Epigenome Roadmap are instrumental in prioritizing SNPs based on their localization within DNA annotated for regulatory function. Statistical models have shown strong enrichment for trait-associated SNPs within regulatory regions [17,95,121]; meanwhile, results from eQTL studies have shown that GWA SNPs often associate with transcription. Incorporating this regulatory information with GWA results will help to address the frustrating question of missing heritability seen in complex traits. By directly integrating regulatory information in a GWA modeling approach, one study found that the number of traits rising to genome-wide significance rose by 5% [121]. Increasing the resolution of association mapping by directly and indirectly incorporating additional genomics data will likely allow for the discovery of variants with smaller effect sizes that play a role in disease.

Though the importance of incorporating genomic information cannot be understated, it is equally important to select the appropriate tissue for the trait of interest, and to conduct regulatory studies in both diseased and normal tissue. Cell lines and tissues included in ENCODE and Epigenome Roadmap typically represent immortalized cell lines or otherwise "normal" tissue, and may differ from the in-vivo chromatin states present in disease, particularly at regions that are specific to disease. In a previous study, FAIRE-seq was used to characterize chromatin landscape changes specific to tumor tissue in renal cell carcinoma [97], and highlight transcriptional changes associated with mutations in chromatin modifying proteins. In the preceding chapter, I used a similar approach in whole colon tissue biopsies of CD and non-IBD individuals to FAIRE-seq and RNA-seq in CD to annotate possible causal genes as well as their associated enhancer regions and transcription factor drivers. Disease-focused inter-omics analyses such as these will be instrumental in interpreting existing results from GWA, and generating new leads.

Continual advances in sequencing technology will further facilitate this aim. The recently described assay for transposase-accessible chromatin using sequencing (ATAC-seq) was originally shown to identify genome wide regulatory regions using as few as 500 cells [123], far fewer than

the 1-50 million cells typically required for FAIRE-seq and DNase-seq. ATAC-seq has since been successfully applied to single cells [124], and prevents tantalizing new opportunities for studying traits in tissues with limited cell numbers. In CD, this technology would make possible epigenome profiling specifically targeted to immune cell populations in the lamina propria. For instance, local macrophage populations could be isolated from both diseased and normal individuals, and regulatory regions specific to the macrophage populations could be defined. This would avoid the problem of signal convolution when using FAIRE-seq applied to whole tissue, allowing for greater power in identifying relevant regions of interest, and better biological interpretability for their effect in a specific cell type.

In conclusion, annotating the functional genome remains a major challenge for the post-GWA era (perhaps even more challenging than watching Mega Vizura miss 14 free throws in a row). Understanding the genetic basis of complex traits will only be possible by investigating their equally complex underlying regulatory architecture. Although large scale catalogues of genomic annotations have greatly complemented what we know about genetic variants associated with complex traits and disease, increased resolution will be required to identify variants with smaller effect sizes. Furthermore, investigation of disease-relevant tissue will be crucial to identifying regulatory elements specific to a trait of interest. In this dissertation, I developed and applied methods for better annotating the regulatory architecture of the genome. These advances offer insight into the complex interplay of genetics, epigenetics, transcriptional regulation, and phenotype, and have provided numerous hypotheses for future functional validation.

REFERENCES

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. Macmillian Magazines Ltd.; 2001;409: 860–921. doi:10.1038/35057062

2. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22: 1760–74. doi:10.1101/gr.135350.111

3. Fields C, Adams MD, White O, Venter JC. How many genes in the human genome? Nat Genet. 1994;7: 345–6. doi:10.1038/ng0794-345

4. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J. Gene index analysis of the human genome estimates approximately 120,000 genes. Nat Genet. 2000;25: 239–40. doi:10.1038/76126

5. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489: 57–74. doi:10.1038/nature11247

6. Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, et al. Unlocking the secrets of the genome. Nature. 2009;459: 927–30. doi:10.1038/459927a

7. Romanoski CE, Glass CK, Stunnenberg HG, Wilson L, Almouzni G. Epigenomics: Roadmap for regulation. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2015;518: 314–316. doi:10.1038/518314a

8. Roy S, Ernst J, Kharchenko P V, Kheradpour P, Negre N, Eaton ML, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010;330: 1787–97. doi:10.1126/science.1198374

9. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 2010;330: 1775–87. doi:10.1126/science.1196914

10. Gerstein MB, Kundaje A, Hariharan M, Landt SG, Yan K-K, Cheng C, et al. Architecture of the human regulatory network derived from ENCODE data. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2012;489: 91–100. doi:10.1038/nature11245

11. Natarajan A, Yardimci GG, Sheffield NC, Crawford GE, Ohler U. Predicting cell-type-specific gene expression from regions of open chromatin. Genome Res. 2012;22: 1711–22. doi:10.1101/gr.135129.111

12. Consortium RE, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. Nature Publishing Group, a division of

Macmillan Publishers Limited. All Rights Reserved.; 2015;518: 317–330. doi:10.1038/nature14248

13. Stranger BE, Stahl EA, Raj T. Progress and promise of genome-wide association studies for human complex trait genetics. Genetics. 2011;187: 367–83. doi:10.1534/genetics.110.120907

14. Klein RJ, Zeiss C, Chew EY, Tsai J-Y, Sackler RS, Haynes C, et al. Complement factor H polymorphism in age-related macular degeneration. Science. 2005;308: 385–9. doi:10.1126/science.1109557

15. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42: D1001–6. doi:10.1093/nar/gkt1229

16. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark road from association to function. Am J Hum Genet. 2013;93: 779–97. doi:10.1016/j.ajhg.2013.10.012

17. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. Genome Res. 2012;22: 1748–59. doi:10.1101/gr.136127.111

18. Manolio T a, Collins FS, Cox NJ, Goldstein DB, Hindorff L a, Hunter DJ, et al. Finding the missing heritability of complex diseases. Nature. Nature Publishing Group; 2009;461: 747–53. doi:10.1038/nature08494

19. Jansen R. Genetical genomics: the added value from segregation. Trends Genet. 2001;17: 388–391. doi:10.1016/S0168-9525(01)02310-1

20. Li J, Burmeister M. Genetical genomics: combining genetics with gene expression analysis. Hum Mol Genet. 2005;14 Spec No: R163–9. doi:10.1093/hmg/ddi267

21. Brem RB, Yvert G, Clinton R, Kruglyak L. Genetic dissection of transcriptional regulation in budding yeast. Science. 2002;296: 752–5. doi:10.1126/science.1069516

22. West MAL, Kim K, Kliebenstein DJ, van Leeuwen H, Michelmore RW, Doerge RW, et al. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. Genetics. 2007;175: 1441–50. doi:10.1534/genetics.106.064972

23. Li Y, Alvarez OA, Gutteling EW, Tijsterman M, Fu J, Riksen JAG, et al. Mapping determinants of gene expression plasticity by genetical genomics in C. elegans. PLoS Genet. 2006;2: e222. doi:10.1371/journal.pgen.0020222

24. Petretto E, Mangion J, Dickens NJ, Cook SA, Kumaran MK, Lu H, et al. Heritability and tissue specificity of expression quantitative trait loci. Flint J, editor. PLoS Genet. Public Library of Science; 2006;2: e172. doi:10.1371/journal.pgen.0020172

25. Hasin-Brumshtein Y, Hormozdiari F, Martin L, van Nas A, Eskin E, Lusis AJ, et al. Allele-specific expression and eQTL analysis in mouse adipose tissue. BMC Genomics. 2014;15: 471. doi:10.1186/1471-2164-15-471

26. Yang T-P, Beazley C, Montgomery SB, Dimas AS, Gutierrez-Arcelus M, Stranger BE, et al. Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies. Bioinformatics. 2010;26: 2474–6. doi:10.1093/bioinformatics/btq452

27. Grundberg E, Small KS, Hedman ÅK, Nica AC, Buil A, Keildson S, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet. 2012;44: 1084–9. doi:10.1038/ng.2394

28. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, et al. Patterns of cis regulatory variation in diverse human populations. PLoS Genet. 2012;8: e1002639. doi:10.1371/journal.pgen.1002639

29. Nica AC, Parts L, Glass D, Nisbet J, Barrett A, Sekowska M, et al. The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. PLoS Genet. 2011;7: e1002003. doi:10.1371/journal.pgen.1002003

30. Dimas AS, Deutsch S, Stranger BE, Montgomery SB, Borel C, Attar-Cohen H, et al. Common regulatory variation impacts gene expression in a cell type-dependent manner. Science. 2009;325: 1246–50. doi:10.1126/science.1174148

31. Ardlie KG, Deluca DS, Segre A V., Sullivan TJ, Young TR, Gelfand ET, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science (80-). 2015;348: 648–660. doi:10.1126/science.1262110

32. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013;45: 580–585. doi:10.1038/ng.2653

33. Moffatt MF, Kabesch M, Liang L, Dixon AL, Strachan D, Heath S, et al. Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. Nature. 2007;448: 470–3. doi:10.1038/nature06014

34. Libioulle C, Louis E, Hansoul S, Sandor C, Farnir F, Franchimont D, et al. Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. PLoS Genet. Public Library of Science; 2007;3: e58. doi:10.1371/journal.pgen.0030058

35. Weiser M, Mukherjee S, Furey TS. Novel distal eQTL analysis demonstrates effect of population genetic architecture on detecting and interpreting associations. Genetics. 2014;198: 879–93. doi:10.1534/genetics.114.167791

36. Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen K-Y, Morley M, et al. Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet. Nature Publishing Group; 2003;33: 422–5. doi:10.1038/ng1094

37. Schadt EE, Monks SA, Drake TA, Lusis AJ, Che N, Colinayo V, et al. Genetics of gene expression surveyed in maize, mouse and man. Nature. Nature Publishing Group; 2003;422: 297–302. doi:10.1038/nature01434

38. Doss S, Schadt EE, Drake TA, Lusis AJ. Cis-acting expression quantitative trait loci in mice. Genome Res. 2005;15: 681–91. doi:10.1101/gr.3216905

39. Holloway B, Luck S, Beatty M, Rafalski J-A, Li B. Genome-wide expression quantitative trait loci (eQTL) analysis in maize. BMC Genomics. 2011;12: 336. doi:10.1186/1471-2164-12-336

40. Mehta D, Heim K, Herder C, Carstensen M, Eckstein G, Schurmann C, et al. Impact of common regulatory single-nucleotide variants on gene expression profiles in whole blood. Eur J Hum Genet. Macmillan Publishers Limited; 2012; doi:10.1038/ejhg.2012.106

41. Alberts R, Chen H, Pommerenke C, Smit AB, Spijker S, Williams RW, et al. Expression QTL mapping in regulatory and helper T cells from the BXD family of strains reveals novel cell-specific genes, gene-gene interactions and candidate genes for auto-immune disease. BMC Genomics. 2011;12: 610. doi:10.1186/1471-2164-12-610

42. Shabalin AA. Matrix eQTL: Ultra fast eQTL analysis via large matrix operations. Bioinformatics. 2012;28: 1353–1358. doi:10.1093/bioinformatics/bts163

43. Göring HHH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, et al. Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. Nat Genet. 2007;39: 1208–16. doi:10.1038/ng2119

44. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, et al. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat Genet. 2003;35: 57–64. doi:10.1038/ng1222

45. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. Macmillan Publishers Limited. All rights reserved; 2010;464: 768–72. doi:10.1038/nature08872

46. Lappalainen T, Sammeth M, Friedländer MR, 't Hoen PAC, Monlong J, Rivas MA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013;501: 506–511. doi:10.1038/nature12531

47. Kompass KS, Witte JS. Co-regulatory expression quantitative trait loci mapping: method and application to endometrial cancer. BMC Med Genomics. 2011;4: 6. doi:10.1186/1755-8794-4-6

48. Bottolo L, Petretto E, Blankenberg S, Cambien F, Cook SA, Tiret L, et al. Bayesian detection of expression quantitative trait loci hot spots. Genetics. 2011;189: 1449–59. doi:10.1534/genetics.111.131425

49. Duarte CW, Zeng Z-B. High-confidence discovery of genetic network regulators in expression quantitative trait loci data. Genetics. 2011;187: 955–64. doi:10.1534/genetics.110.124685

50. Rotival M, Zeller T, Wild PS, Maouche S, Szymczak S, Schillert A, et al. Integrating genome-wide genetic variations and monocyte expression data reveals trans-regulated gene modules in humans. Barsh GS, editor. PLoS Genet. Public Library of Science; 2011;7: e1002367. doi:10.1371/journal.pgen.1002367

51. Chen LS, Emmert-Streib F, Storey JD. Harnessing naturally randomized transcription to infer regulatory relationships among genes. Genome Biol. 2007;8: R219. doi:10.1186/gb-2007-8-10-r219

52. Neto EC, Broman AT, Keller MP, Attie AD, Zhang B, Zhu J, et al. Modeling causality for pairs of phenotypes in system genetics. Genetics. 2013;193: 1003–13. doi:10.1534/genetics.112.147124

53. Brem RB, Kruglyak L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. Proc Natl Acad Sci U S A. 2005;102: 1572–7. doi:10.1073/pnas.0408709102

54. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5: 101–13. doi:10.1038/nrg1272

55. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. Nature. 2000;407: 651–4. doi:10.1038/35036627

56. Yook S-H, Oltvai ZN, Barabási A-L. Functional and topological characterization of protein interaction networks. Proteomics. 2004;4: 928–42. doi:10.1002/pmic.200300636

57. Lorenz WW, Alba R, Yu Y-S, Bordeaux JM, Simões M, Dean JFD. Microarray analysis and scale-free gene networks identify candidate regulators in drought-stressed roots of loblolly pine (P. taeda L.). BMC Genomics. 2011;12: 264. doi:10.1186/1471-2164-12-264

58. Peng J, Wang P, Zhou N, Zhu J. Partial Correlation Estimation by Joint Sparse Regression Models. J Am Stat Assoc. 2009;104: 735–746. doi:10.1198/jasa.2009.0126

59. Allen JD, Xie Y, Chen M, Girard L, Xiao G. Comparing statistical methods for constructing large scale gene networks. Holme P, editor. PLoS One. Public Library of Science; 2012;7: e29348. doi:10.1371/journal.pone.0029348

60. Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat. 2001;29: 1165–1188. Available: http://projecteuclid.org/euclid.aos/1013699998

61. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J R Stat Soc Ser B. 1995;57: 289 – 300. doi:10.2307/2346101

62. Romano JP, Shaikh AM, Wolf M. Control of the false discovery rate under dependence using the bootstrap and subsampling. TEST. 2008;17: 417–442. doi:10.1007/s11749-008-0126-6

63. Aylor DL, Valdar W, Foulds-Mathes W, Buus RJ, Verdugo RA, Baric RS, et al. Genetic analysis of complex traits in the emerging Collaborative Cross. Genome Res. 2011;21: 1213–22. doi:10.1101/gr.111310.110

64. Breitling R, Li Y, Tesson BM, Fu J, Wu C, Wiltshire T, et al. Genetical genomics: spotlight on QTL hotspots. Churchill GA, editor. PLoS Genet. Public Library of Science; 2008;4: e1000232. doi:10.1371/journal.pgen.1000232

65. Litvin O, Causton HC, Chen B-J, Pe'er D. Modularity and interactions in the genetics of gene expression. Proc Natl Acad Sci U S A. 2009;106: 6441–6. doi:10.1073/pnas.0810208106

66. Airoldi EM, Huttenhower C, Gresham D, Lu C, Caudy AA, Dunham MJ, et al. Predicting cellular growth from gene expression signatures. Rzhetsky A, editor. PLoS Comput Biol. Public Library of Science; 2009;5: e1000257. doi:10.1371/journal.pcbi.1000257

67. Teixeira MC, Monteiro PT, Guerreiro JF, Gonçalves JP, Mira NP, dos Santos SC, et al. The YEASTRACT database: an upgraded information system for the analysis of gene and genomic transcription regulation in Saccharomyces cerevisiae. Nucleic Acids Res. 2014;42: D161–6. doi:10.1093/nar/gkt1015

68. Becker B, Feller A, El Alami M, Dubois E, Pierard A. A nonameric core sequence is required upstream of the LYS genes of Saccharomyces cerevisiae for Lys14p-mediated activation and apparent repression by lysine. Mol Microbiol. 1998;29: 151–163. doi:10.1046/j.1365-2958.1998.00916.x

69. RAMOS F, DUBOIS E, PIERARD A. Control of enzyme synthesis in the lysine biosynthetic pathway of Saccharomyces cerevisiae. Evidence for a regulatory role of gene LYS14. Eur J Biochem. 1988;171: 171–176. doi:10.1111/j.1432-1033.1988.tb13773.x

70. Feller A, Ramos F, Pierard A, Dubois E. In Saccharomyces cerevisae, feedback inhibition of homocitrate synthase isoenzymes by lysine modulates the activation of LYS gene expression by Lys14p. Eur J Biochem. 1999;261: 163–170. doi:10.1046/j.1432-1327.1999.00262.x

71. Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, Beatty J, et al. The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet. 2004;36: 1133–7. doi:10.1038/ng1104-1133

72. Collaborative Cross Consortium. The genome architecture of the Collaborative Cross mouse genetic reference population. Genetics. 2012;190: 389–401. doi:10.1534/genetics.111.132639

73. Phillippi J, Xie Y, Miller DR, Bell TA, Zhang Z, Lenarcic AB, et al. Using the emerging Collaborative Cross to probe the immune system. Genes Immun. 2013; doi:10.1038/gene.2013.59

74. Ferris MT, Aylor DL, Bottomly D, Whitmore AC, Aicher LD, Bell TA, et al. Modeling host genetic regulation of influenza pathogenesis in the collaborative cross. PLoS Pathog. 2013;9: e1003196. doi:10.1371/journal.ppat.1003196

75. Philip VM, Sokoloff G, Ackert-Bicknell CL, Striz M, Branstetter L, Beckmann MA, et al. Genetic analysis in the Collaborative Cross breeding population. Genome Res. 2011;21: 1223–38. doi:10.1101/gr.113886.110

76. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A. 2009;106: 9362–7. doi:10.1073/pnas.0903103106

77. Kang HP, Morgan AA, Chen R, Schadt EE, Butte AJ. Coanalysis of GWAS with eQTLs reveals disease-tissue associations. AMIA Summits Transl Sci Proc. 2012;2012: 35–41. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3392070&tool=pmcentrez&rendertype=abstract

78. Ma L, Clark AG, Keinan A. Gene-based testing of interactions in association studies of quantitative traits. Williams SM, editor. PLoS Genet. Public Library of Science; 2013;9: e1003321. doi:10.1371/journal.pgen.1003321

79. Gat-Viks I, Meller R, Kupiec M, Shamir R. Understanding gene sequence variation in the context of transcription regulation in yeast. PLoS Genet. 2010;6: e1000800. doi:10.1371/journal.pgen.1000800

80. Zhu J, Zhang B, Smith EN, Drees B, Brem RB, Kruglyak L, et al. Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat Genet. 2008;40: 854–61. doi:10.1038/ng.167

81. Zhu J, Sova P, Xu Q, Dombek KM, Xu EY, Vu H, et al. Stitching together multiple data dimensions reveals interacting metabolomic and transcriptomic networks that modulate cell regulation. Levchenko A, editor. PLoS Biol. Public Library of Science; 2012;10: e1001301. doi:10.1371/journal.pbio.1001301

82. Smith EN, Kruglyak L. Gene-environment interaction in yeast gene expression. Mackay T, editor. PLoS Biol. Public Library of Science; 2008;6: e83. doi:10.1371/journal.pbio.0060083

83. Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491: 119–24. doi:10.1038/nature11582

84. Kappelman MD, Moore KR, Allen JK, Cook SF. Recent trends in the prevalence of Crohn's disease and ulcerative colitis in a commercially insured US population. Dig Dis Sci. 2013;58: 519–25. doi:10.1007/s10620-012-2371-5

85. Polito JM, Childs B, Mellits ED, Tokayer AZ, Harris ML, Bayless TM. Crohn's disease: influence of age at diagnosis on site and clinical type of disease. Gastroenterology. 1996;111: 580–6. Available: http://www.ncbi.nlm.nih.gov/pubmed/8780560

86. Colombel JF, Sandborn WJ, Reinisch W, Mantzaris GJ, Kornbluth A, Rachmilewitz D, et al. Infliximab, azathioprine, or combination therapy for Crohn's disease. N Engl J Med. 2010;362: 1383–95. doi:10.1056/NEJMoa0904492

87. Peyrin-Biroulet L, Loftus E V, Colombel J-F, Sandborn WJ. The natural history of adult Crohn's disease in population-based cohorts. Am J Gastroenterol. American College of Gastroenterology; 2010;105: 289–97. doi:10.1038/ajg.2009.579

88. Thiesen S, Janciauskiene S, Uronen-Hansson H, Agace W, Högerkorp C-M, Spee P, et al. CD14(hi)HLA-DR(dim) macrophages, with a resemblance to classical blood monocytes, dominate inflamed mucosa in Crohn's disease. J Leukoc Biol. 2014;95: 531–41. doi:10.1189/jlb.0113021

89. Karaiskos C, Hudspith BN, Elliott T, Rayment NB, Avgousti V, Sanderson JD. Defective macrophage function in crohn's disease: role of alternatively activated macrophages in inflammation. Gut. 2011;60: A143–A144. doi:10.1136/gut.2011.239301.304

90. Sheikh SZ, Plevy SE. The role of the macrophage in sentinel responses in intestinal immunity. Curr Opin Gastroenterol. 2010;26: 578–82. doi:10.1097/MOG.0b013e32833d4b71

91. Kamada N, Hisamatsu T, Okamoto S, Chinen H, Kobayashi T, Sato T, et al. Unique CD14 intestinal macrophages contribute to the pathogenesis of Crohn disease via IL-23/IFN-gamma axis. J Clin Invest. 2008;118: 2269–80. doi:10.1172/JCI34610

92. Tanaka M, Saito H, Kusumi T, Fukuda S, Shimoyama T, Sasaki Y, et al. Spatial distribution and histogenesis of colorectal Paneth cell metaplasia in idiopathic inflammatory bowel disease. J Gastroenterol Hepatol. 2001;16: 1353–1359. doi:10.1046/j.1440-1746.2001.02629.x

93. Simmonds N, Furman M, Karanika E, Phillips A, Bates AWH. Paneth cell metaplasia in newly diagnosed inflammatory bowel disease in children. BMC Gastroenterol. 2014;14: 93. doi:10.1186/1471-230X-14-93

94. Grimm MC, Pavli P. NOD2 mutations and Crohn's disease: are Paneth cells and their antimicrobial peptides the link? Gut. 2004;53: 1558–60. doi:10.1136/gut.2004.043307

95. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337: 1190–5. doi:10.1126/science.1222794

96. Song L, Zhang Z, Grasfeder LL, Boyle AP, Giresi PG, Lee B-K, et al. Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity. Genome Res. 2011;21: 1757–67. doi:10.1101/gr.121541.111

97. Simon JM, Hacker KE, Singh D, Brannon AR, Parker JS, Weiser M, et al. Variation in chromatin accessibility in human kidney cancer links H3K36 methyltransferase loss with widespread RNA processing defects. Genome Res. 2014;24: 241–50. doi:10.1101/gr.158253.113

98. Davie K, Jacobs J, Atkins M, Potier D, Christiaens V, Halder G, et al. Discovery of Transcription Factors and Regulatory Regions Driving In Vivo Tumor Development by ATAC-seq and FAIRE-seq Open Chromatin Profiling. McKinnon P, editor. PLOS Genet. 2015;11: e1004994. doi:10.1371/journal.pgen.1004994

99. Peck BCE, Weiser M, Lee SE, Gipson GR, Iyer VB, Sartor RB, et al. MicroRNAs Classify Different Disease Behavior Phenotypes of Crohn's Disease and May Have Prognostic Utility. Inflamm Bowel Dis. 2015; doi:10.1097/MIB.0000000000000478

100. Costello CM, Mah N, Häsler R, Rosenstiel P, Waetzig GH, Hahn A, et al. Dissection of the inflammatory bowel disease transcriptome using genome-wide cDNA microarrays. PLoS Med. 2005;2: e199. doi:10.1371/journal.pmed.0020199

101. Noble CL, Abbas AR, Lees CW, Cornelius J, Toy K, Modrusan Z, et al. Characterization of intestinal gene expression profiles in Crohn's disease by genome-wide microarray analysis. Inflamm Bowel Dis. 2010;16: 1717–28. doi:10.1002/ibd.21263

102. Mirza AH, Berthelsen CH, Seemann SE, Pan X, Frederiksen KS, Vilien M, et al. Transcriptomic landscape of lncRNAs in inflammatory bowel disease. Genome Med. 2015;7: 39. doi:10.1186/s13073-015-0162-2

103. Liu EY, Li M, Wang W, Li Y. MaCH-admix: genotype imputation for admixed populations. Genet Epidemiol. 2013;37: 25–37. doi:10.1002/gepi.21690

104. Lassmann T, Hayashizaki Y, Daub CO. TagDust--a program to eliminate artifacts from next generation sequencing data. Bioinformatics. 2009;25: 2839–40. doi:10.1093/bioinformatics/btp527

105. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26: 873–81. doi:10.1093/bioinformatics/btq057

106. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28: 882–3. doi:10.1093/bioinformatics/bts034

107. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007;23: 257–8. doi:10.1093/bioinformatics/btl567

108. Simon JM, Giresi PG, Davis IJ, Lieb JD. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. Nat Protoc. Nature Publishing Group; 2012;7: 256–67. doi:10.1038/nprot.2011.444

109. Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26: 873–81. doi:10.1093/bioinformatics/btq057

110. Boyle AP, Guinney J, Crawford GE, Furey TS. F-Seq: a feature density estimator for high-throughput sequence tags. Bioinformatics. 2008;24: 2537–8. doi:10.1093/bioinformatics/btn480

111. Mokry M, Middendorp S, Wiegerinck CL, Witte M, Teunissen H, Meddens CA, et al. Many inflammatory bowel disease risk loci include regions that regulate gene expression in immune cells and the intestinal epithelium. Gastroenterology. 2014;146: 1040–7. doi:10.1053/j.gastro.2013.12.003

112. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics. 2011;27: 1653–9. doi:10.1093/bioinformatics/btr261

113. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007;8: R24. doi:10.1186/gb-2007-8-2-r24

114. Comelli EM, Lariani S, Zwahlen M-C, Fotopoulos G, Holzwarth JA, Cherbut C, et al. Biomarkers of human gastrointestinal tract regions. Mamm Genome. 2009;20: 516–27. doi:10.1007/s00335-009-9212-7

115. Marstrand TT, Storey JD. Identifying and mapping cell-type-specific chromatin programming of gene expression. Proc Natl Acad Sci U S A. 2014;111: E645–54. doi:10.1073/pnas.1312523111

116. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473: 43–9. doi:10.1038/nature09906

117. Waki H, Nakamura M, Yamauchi T, Wakabayashi K, Yu J, Hirose-Yotsuya L, et al. Global mapping of cell type-specific open chromatin by FAIRE-seq reveals the regulatory role of the NFI family in adipocyte differentiation. PLoS Genet. 2011;7: e1002311. doi:10.1371/journal.pgen.1002311

118. McLean CY, Bristor D, Hiller M, Clarke SL, Schaar BT, Lowe CB, et al. GREAT improves functional interpretation of cis-regulatory regions. Nat Biotechnol. 2010;28: 495–501. doi:10.1038/nbt.1630

119. Hermann-Kleiter N, Meisel M, Fresser F, Thuille N, Müller M, Roth L, et al. Nuclear orphan receptor NR2F6 directly antagonizes NFAT and RORγt binding to the Il17a promoter. J Autoimmun. 2012;39: 428–40. doi:10.1016/j.jaut.2012.07.007

120. Hulur I, Gamazon ER, Skol AD, Xicola RM, Llor X, Onel K, et al. Enrichment of inflammatory bowel disease and colorectal cancer risk variants in colon expression quantitative trait loci. BMC Genomics. 2015;16: 138. doi:10.1186/s12864-015-1292-z

121. Pickrell JK. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. Am J Hum Genet. 2014;94: 559–73. doi:10.1016/j.ajhg.2014.03.004

122. Fais S, Capobianchi MR, Silvestri M, Mercuri F, Pallone F, Dianzani F. Interferon expression in Crohn's disease patients: increased interferon-gamma and -alpha mRNA in the intestinal lamina propria mononuclear cells. J Interferon Res. 1994;14: 235–8. Available: http://www.ncbi.nlm.nih.gov/pubmed/7861027

123. Buenrostro JD, Giresi PG, Zaba LC, Chang HY, Greenleaf WJ. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods. Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved.; 2013;10: 1213–8. doi:10.1038/nmeth.2688

124. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523: 486–90. doi:10.1038/nature14590