# Analysis of *trans* eSNPs infers regulatory network architecture

Anat Kreimer

Submitted in partial fulfillment of the

requirements for the degree

of Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

ABSTRACT

Analysis of *trans* eSNPs infers regulatory network architecture

Anat Kreimer

eSNPs are genetic variants associated with transcript expression levels. The characteristics of such variants highlight their importance and present a unique opportunity for studying gene regulation. eSNPs affect most genes and their cell type specificity can shed light on different processes that are activated in each cell. They can identify functional variants by connecting SNPs that are implicated in disease to a molecular mechanism. Examining eSNPs that are associated with distal genes can provide insights regarding the inference of regulatory networks but also presents challenges due to the high statistical burden of multiple testing. Such association studies allow: simultaneous investigation of many gene expression phenotypes without assuming any prior knowledge and identification of unknown regulators of gene expression while uncovering directionality.

This thesis will focus on such distal eSNPs to map regulatory interactions between different loci and expose the architecture of the regulatory network defined by such interactions. We develop novel computational approaches and apply them to genetics-genomics data in human. We go beyond pairwise interactions to define network motifs, including regulatory modules and bi-fan structures, showing them to be prevalent in real data and exposing distinct attributes of such arrangements. We project eSNP associations onto a protein-protein interaction network to expose topological properties of eSNPs and their targets and highlight different modes of distal regulation. Overall, our work offers insights concerning the topological structure of human regulatory networks and the role genetics plays in shaping them.

# Tables of Contents

# List of Figures

# List of Tables

# Acknowledgements

Trying to remember how this journey began, I roll back to this specific night in Tel Aviv, sitting at my brother's apartment sipping whiskey, when that phone call arrived. On the line was Prof. Andrea Califano, welcoming me to Columbia. I don't remember what we talked about, just the screaming and jumping after the call ended. Well, as you can imagine, this started a roller-coaster that has changed my life.

Many have accompanied me throughout these five years and deserve gratitude and recognition. First and foremost, my PhD advisor and mentor Prof. Itsik Pe'er. Thank you for believing in my potential and teaching me with great patience and expertise. You developed my scientific skills and introduced me to many aspects of the academic world. To my thesis committee – Prof. Dennis Vitkup, Prof. Richard Friedman, Prof. Yufeng Shen and Prof. Christina Leslie: thank you for your investment and your thoughtful comments and suggestions. To my colleagues and lab-mates at Columbia, particularly Shai, Vlada, El-ad, Gal, Pier, Snehit, Sasha and Eimear: thank you for your advice and friendship.

On a more personal note, I want to thank my husband, Itamar for always encouraging me and pushing me forward. Thank you for agreeing to all me crazy ideas and making this possible! To my parents - Joseph and Bella and my brother Avi, for believing in me and always being proud and supportive. And last, but not least, to my new love - Danielle, for never letting me sleep but always making me smile.

# Chapter 1: Introduction

In the last decade many genetic variants (eSNPs) have been reported as associated with expression of transcripts, holding the promise for functional dissection of regulatory structure of human transcription. There are several approaches by which eSNPs can be explored. First, they can be characterized by different categories: their genomic location, functional role, distance from associated transcript (*cis/trans* eSNPs), and similarities and differences across tissues, cells and conditions. Second, eSNPs can be integrated with results from genome-wide association studies (GWAS) to predict their specific regulatory role in disease and human traits. Third, their analysis with respect to gene networks and functional annotations addresses questions regarding the organization of transcription, the causality in association and can pinpoint the regulatory mechanisms through which eSNPs act.

eSNPs are found to effect the expression of most genes [1], stressing the importance of studying and characterizing such variants. eSNPs can affect the expression of a close, usually defined as up to 1MB (*cis*), or a distal (*trans*) gene. For example, a synonymous SNP may have a local effect on the expression level of its host gene, while a non-synonymous SNP in a transcription factor may have a distal effect on its targets. *Cis* eSNPs are enriched in exons comparing to introns [2]. A large fraction of *cis* eSNPs are found in close proximity of the transcription start site (TSS), approximately 50kb on either sides of the TSS [3] and are enriched in promoters and transcription factor binding sites, suggesting that many directly impact protein-DNA interactions [3]. There are significantly less *cis*-eSNPs that affect central and critical genes, along with a trend of reduced effect sizes as variant frequency increases, providing evidence that purifying

selection and buffering have limited the deleterious impact of regulatory variation on the cell [4]. Finally, there is a significant overlap between SNPs that are associated with gene expression levels and essential epigenetic marks, i.e., methylation, [5] DNase I sensitivity [6] and histone marks [7] levels, as well as miRNA expression levels [8].

eSNPs are cell [9-11] and tissue [12-14] type specific, thus they can be telling regarding different mechanisms that are distinct or shared. This phenomenon is stronger for *trans* eSNPs [9] and can be used to detect different pathways and interactions that could suggest functional processes that are common or specific for pairs of cell types. For example when comparing *trans* associations between B-cells and monocytes, Fairfax *et al.* find LYZ as a monocyte-specific master regulator of a large gene set. Although in general, shared *cis* eSNPs have the same directional effect on the gene expression in each analyzed cell type [10], there is an enrichment for shared *cis* eSNPs with opposing directional effects in each cell, i.e., cell type–specific directionality [9]. eSNPs are condition-specific and they depend on the time and duration of the stimulus [15]. These condition-specific eSNPs were found to be more distal to the transcriptional start site and, in some cases, showed reversal of effect between conditions. Moreover, stimulation reveals novel *trans*-eSNPs with simultaneous effects involving many genes [15].

Although genome wide association studies (GWAS) [16] have linked numerous genetic loci to various human diseases and traits, pinpointing the causal variants and understanding the underlying mechanisms of these phenotypes is still limited. Since GWAS SNPs are known to be statistically enriched in eSNPs [17], one approach for addressing these challenging questions is to integrate these two types of data while using eSNP data to interpret GWAS signals [18, 19].

This approach has been validated by several studies. For example, Dubois *et al.* [20] found that 20 of the 38 loci that had associated risk variants for celiac disease are also correlated with variation in the expression of a nearby gene. In an analysis of the genetics of migraine, genotypic correlation to expression of a candidate gene suggests a regulatory basis for this trait [21]. An approach combining eSNPs in metabolically active tissues with pathways enriched for relevant GWAS SNPs provided a potential powerful framework for identifying biological mechanisms underlying GWAS findings [22]. Finally, expression quantitative trait loci (eQTL) meta-analysis that was performed in peripheral blood samples from thousands of individuals identified and replicated *trans* eQTLs that were previously associated with complex traits at genome-wide significance. The observed regulation patterns indicated that such approach provides insight into the downstream effects of many trait-associated variants [23].

Integrating GWAS SNP data with biological networks can illuminate mechanisms underlying disease. Studies that project GWAS SNPs on the protein-protein interaction (PPI) network conclude that disease associated loci encode directly interacting proteins beyond chance expectation, suggesting that risk variants may act on suites of proteins involved in the same process [24-26]. Previous works that integrated co-expression networks with disease variants, found enrichment of these variants in their co-expression modules, implicating that these modules represent causal effects [27]. This approach highlights the potential use in network analyses to reconstruct molecular phenotypes for the identification of the genetic association signal derived from pathways, rather than small effects from individual genes [28]. Overall, examining genetic variants that are associated with a specific disease unravels functional gene

networks [29-33]. Ultimately, the goal is to pinpoint the causal variants for human traits and provide a functional explanation to how they exert these phenotypical changes.

There is a high statistical burden of multiple testing when considering association in *trans*, therefore most of the studies still focus on *cis* association. While *cis* regulation is extremely important in understanding the mechanisms of transcription, it is limited, by its local nature, in the insights it can provide regarding interactions, pathways and the overall architecture of gene regulation [34]. Constructing regulatory networks based on eSNP data in different biological contexts (e.g., specific cell type or disease) can shed light on questions regarding the role of genetics in shaping the organization of gene regulation, how it changes under different environments and conditions and by which mechanisms. Multiple studies that have taken this approach report intriguing findings. *Trans*-eSNPs seem to be organized in a modular fashion, when a single variant is associated with the expression of multiple genes [4, 9, 35, 36]. This single variant usually has a *cis* effect on the expression of a gene, which in turn has a *trans* effect on the gene set [4, 36]. These co-regulated gene sets are enriched in functional annotations and correspond to known pathways [15], they are cell type specific [9] and condition dependent [15], highlighting processes that are relevant under a specific biological setting. For example, Fairfax *et al.* report findings of coding polymorphisms in CYP1B1, P2RY11, and IDO2 that modulate activity and develop *trans* network effects that can be observed only upon stimulation [15].

In the following chapters, I will describe our work on eSNPs in *trans*, which aims at providing insight on regulatory interactions between different loci and the architecture of the regulatory network that such interactions define.

In chapter 2 we present a computational framework that goes beyond pairwise interactions to define network motifs [36]. We show that considering transcripts, each weakly associated to a single 'main' SNP, exposes high confidence regulatory modules structures. We represent the dependencies between the transcripts in the module and the main SNP by a graphical model. When applied to genetics-genomics data in the liver, we observe that the modules are prevalent in real data and exhibit unique characteristics. In chapter 3, we extend this framework to combine every two basic module structures, i.e., modules composed of two genes, that share the same gene pairs, exposing a bi-fan structure in the human regulatory network [22]. This structure is a known building block of model organisms' regulatory networks [37]. In chapter 4 we take a step forward and integrate eSNP associations with a PPI network. We show that projecting these interactions onto the PPI network exposes topological properties of eSNPs and their targets, unravels different modes of *trans* regulation and highlights a mechanism by which the gene expression is altered [38]. In chapter 5, we summarize our main findings and discuss the limitations of our approaches and future directions.

There is a very large number of eSNP studies being performed in human cohorts and the vast majority of their analyses are based on considering a single SNP associated with a single transcript and mainly in *cis* [1, 6, 39, 40]. While this analysis only captures a fraction of the complexity of genetics of regulation, the advantage is that these approaches provide some statistical guarantees on the associations discovered. There is a smaller number of studies that build networks from eSNP data [4, 9, 34, 35]. While these papers provide much more comprehensive models, the main issue is that they do not provide the same strength of statistical

assurance in their findings. The main advantage of our approach [22, 41] is that it provides a framework for analysis of eSNP data which is very different from the typical analyses and bridges these two approaches while establishing statistical guarantees on our inferred results using permutations.

There are number of works integrating SNP data with biological networks. Many of these works focus on GWAS SNPs [24-26] while some of them rely on co-expression networks [38], that are derived from gene expression data, the same data that is used for finding eSNPs. Our approach [38] utilizes two independent sources of information: eSNP associations, derived from sequencing-ascertained variants, and an established PPI network [42], aiming to address the gap between association, causality and mechanism. Overall, our work offers insights concerning the architecture of the human regulatory network and the role genetics plays in shaping it.

# Chapter 2: Inference of modules associated to eQTLs

*Summary:* Cataloging the association of transcripts to genetic variants in recent years holds the promise for functional dissection of regulatory structure of human transcription. Here, we present a novel approach, which aims at elucidating the joint relationships between transcripts and single-nucleotide polymorphisms (SNPs). This entails detection and analysis of modules of transcripts, each weakly associated to a single genetic variant, together exposing a high confidence association signal between the module and this 'main' SNP. To explore how transcripts in a module are related to causative loci for that module, we represent such dependencies by a graphical model.

We applied our method to the existing data on genetics of gene expression in the liver. The modules are significantly more, larger and denser than found in permuted data. Quantification of the confidence in a module as a likelihood score, allows us to detect transcripts that do not reach genome-wide significance level. Topological analysis of each module identifies novel insights regarding the flow of causality between the main SNP and transcripts. We observe similar annotations of modules from two sources of information: the enrichment of a module in gene subsets and locus annotation of the genetic variants. This and further phenotypic analysis provide a validation for our methodology [36].

# 2.1 Introduction

Variation in genomic DNA can affect function in multiple ways, most typically by alteration of the expressed quantity or sequence content of local transcripts. This premise motivated extensive studies over the last decade, cataloging the influence of human genetic variants on gene expression, most often in *cis* [43, 44]. Local gene expression level is formally considered as a quantitative trait that is directly modified by allelic variation in regulatory elements [45, 46]. Such modifications of transcriptional regulation have been documented to affect health-related traits as diverse as asthma [47] and low density lipoprotein (LDL) cholesterol concentration [48].

Yet, for large fraction of single-nucleotide polymorphisms (SNPs) with well supported associations to disease phenotypes [49] which are neither coding, nor linked to coding SNPs in *cis*, no *cis*-regulatory effect have been reported in studies conducted thus far. A compelling biological hypothesis is that such a SNP does change the transcriptome state or program in order to exert its phenotypic impact, and this regulation is mediated by a transcript in *cis*, but in the particular tissue examined, the changes to transcription level of the mediator gene are too minute to guarantee detection in small association cohorts. This hypothesis leads to an approach for mapping expression quantitative trait loci (eQTLs) that is focused on downstream effects of a regulatory SNP across multiple genes in *trans*, rather than the *cis*-transcript that may mechanistically mediate the effect. A related approach had been successful in simpler organisms [50], motivating this work.

Data on both gene expression and SNP variation across multiple individuals, often termed genetic genomics have facilitated identification of thousands of expression single-nucleotide

polymorphisms (eSNPs) [17, 51]. Approaches that combine these two types of data along with additional factors including the previously inferred biological network structure [52], modularity of gene expression [53], pathway analysis [54] and enzymatic activity [55] had been proposed. However, tying genetic variation in specific loci to phenotypes is still an active field of research.

In this study, we focus on the modularity of gene regulatory networks, a major organizing principle of biological systems [56]. A module is the fundamental unit of a biological network that consists of a set of elements (e.g. genes) working jointly to fulfill a distinct function. Several studies have used this property to gain better understanding of the regulatory mechanisms [57] that are affected by genetic variation. Litvin et al. [50] characterize how genetic variants in multiple loci combine to influence the expression of clusters of co-expressed genes in yeast. Ghazalpour et al. [53] used co-expression networks to study the genetics of complex physiological traits that are relevant to the metabolic syndrome. Schadt et al. [52] used previously reconstructed regulatory networks of genes in mouse and human [58] to support the existing Genome Wide Association Studies (GWAS) results [16]. Known pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG) [59] were used by Zhong et al. [54] for the same purpose. Common to all these studies are three steps. The first two are independent: (i) construction of a network from gene expression data; and (ii) detection of association between genetic variants and expression traits; the final step is (iii) integration of genetic association into the network.

However, it is artificial to separate the stages of network construction based on expression data only from a single SNP–transcript association mapping. Ideally, one would combine information

from multiple transcripts with genetics in a unified analysis. This motivates complementary approaches to analysis of eSNPs. Specifically, our premise is that the modular organization of gene regulation can be used to pinpoint eSNPs that affect multiple, rather than single genes. Therefore, we developed a method that focuses on groups of transcripts (modules) that are each associated with a single genetic variant.

We present a novel approach that entails analyzing modules of transcripts, each associated to a single genetic variant. These modules are constructed based on both available types of data: transcript expression and genotypes. We combine these transcripts into modules that each share an associated SNP, which we denote as the 'main' SNP of that module. This step utilizes the modular organization of gene regulation. We filter the modules according to a confidence score. This score allows us to identify groups of transcripts that are associated to a SNP even if their individual association is not genome-wide significant. We examine the topology of modules, accounting for independent co-association, which is not merely the result of co-expression. This step allows us to infer the flow of causality between the main SNP and the transcripts in the module. We distinguish direct versus indirect SNP– transcript associations through another intermediate transcript whose expression level is co-associated to the same SNP. The main SNP can possibly have *cis*- or *trans*-effects on the transcripts in the module. A local *cis*-effect on a transcript that is either included or excluded from a module can in turn have a modular *trans*-regulatory effect on the other transcripts in the module by virtue of its changed expression levels or altered produced protein (e.g. a mutation in transcription factor).

Regulatory effects can be categorized by *cis*- and *trans*-effects. The *cis*-effects of eSNPs are often due to changes within the promoter, enhancer or other regulatory regions of a gene that may change the expression of that gene. *Trans*-effects of the main SNP on module transcripts can be the outcome of two potentially overlapping scenarios: First, a *cis* main SNP that is located within or close by the coding region of one of the genes in the module can alter the produced protein. The altered protein may then have a *trans*-regulatory effect on the other transcripts in the module by virtue of its differential expression level despite the protein itself being potentially unmodified. Second, a *trans* main SNP that is located within or close by the coding region of a gene that is not a part of the module can alter the produced protein. This distant altered protein may then have a *trans*-effect on the other transcripts in the module by virtue of its modified sequence, despite potentially maintaining its expression level.

All methods previously introduced group transcripts by a shared associated marker and determine intra-cluster interactions by using the correlation of gene expression levels. To our knowledge, this is the first work where a confidence score is assigned to each module and direct/ indirect interactions are determined between pairs of transcripts within a module illustrating the dependence/independence of their expression levels conditioned on the main SNP. We are thus able to go beyond traditional clustering-related methods that are based on expression only, and in fact, examine the joint association and the topology of the modules and not merely their content. For completion, we further search for regulatory hierarchical structure within each module: we examine SNPs whose association to transcript levels in a module is conditioned on the main SNP, and denote those as ‘secondary’ SNPs. This step is illustrated as a decision tree where samples in each module are split, first by the genotype of the main SNP and then by the genotype

of the secondary SNP. We applied our method to data regarding genotype and gene expression in the liver across 371 samples. This data had been previously analyzed in other means [52]. We observe known relationships from the literature between a module and its associated genetic variants, thereby providing support to our methodology.

# 2.2 Results

### *2.2.1 Computational framework for detecting transcriptional modules*

We set out to develop a statistical–computational framework to elucidate the regulatory structure by which genetic variants affect transcription. Specifically, we aim to examine the hypothesis that SNPs can have a modular effect on gene expression. Our method detects transcriptional modules, each including transcripts that are associated with the same main SNP. It is important to distinguish the modules that we find from co-expression clusters. Specifically, we represent each module as a graph, where nodes are transcripts, and for each possible pair of transcripts an edge correspond to a scenario where at least one of the transcripts remains significantly associated to the SNP when conditioned on its counterpart.

An initial step of detecting association pairs of SNP and transcript, showed as many such pairs as expected under the null hypothesis of no such true association. However, we were still motivated to search for modules, as the same associated SNPs were shared by many transcripts. Briefly, we collated association pairs that share a SNP into triplets and larger modules. Such modules are more numerous, bigger, denser in association and more functionally enriched than expected by chance.

In detail, we devised a three-step procedure for detecting the modules regulated by eQTLs. The first step detects 67,540 association pairs of a SNP *s* and a transcript *t* whose expression level is putatively associated with *s* (nominal association $P<10^{-5}$, see Materials and Methods section 2.4.2 for details). The distribution of the number of pairs in the permuted data (Figure 2-1a) demonstrates that the observed number of association pairs is consistent with the null

expectation ($P \approx 0.07$). We eliminate 623 pairs that include transcripts whose association statistic is strongly distorted, as observed by permutation (see Materials and Methods section 2.4.2 for details). We proceed with analyzing the remaining 66,917 association pairs.

Association pairs are binned by SNP *s*, and give rise to 10,354 modules (see Materials and Methods section 2.4.3), ranging in size from 2 to 91 transcripts who are associated to the same main SNP of the module (Figure 2-2). Only 518 modules are large, i.e. with 10 or more transcripts. There are significantly more modules—10,354 (Figure 2-1b) than those found in the permuted data (average 2,322 across permutations; SD 208). Specifically, there are significantly more large modules—518 (Figure 2-1c) than those found in the permuted data (average 220; SD 42). While the observed number of significantly associated pairs of transcript and SNP is consistent with the null expectation, we find that there are significantly more modules than those found in the permuted data. This finding is consistent with the premise that gene regulation is modularly organized.

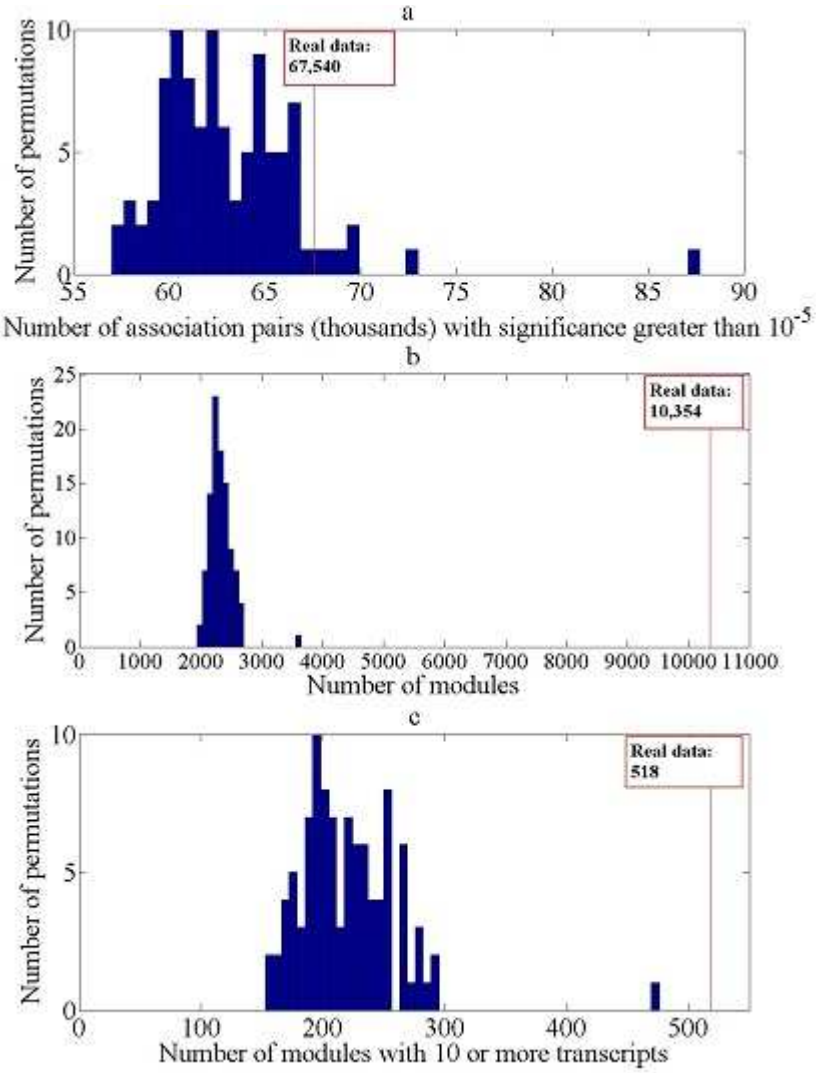


**Figure 2-1. Histogram for the number of association pairs, modules and large modules.**
The number of (a) association pairs (b) modules and (c) large modules in real data compared with 100 permuted data sets. Although only 93 out of the 100 permutated data sets have fewer association pairs than in the real data, all of them have fewer (large) modules.

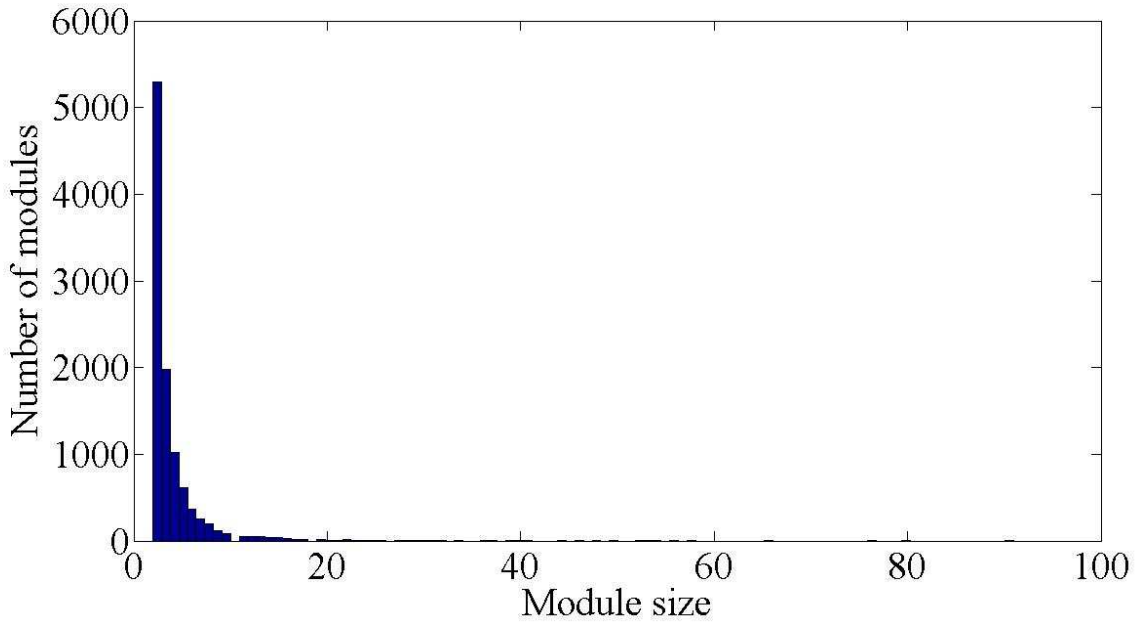


**Figure 2-2. modules sizes distribution.**

### *2.2.2 Modules' topology*

The set of pairs includes 137,889 possible triplets (*s*,*t*,*t'*) where (*s*,*t*) and (*s*,*t'*) are association pairs. Focusing on co-associated pairs of transcripts, we find that for 129,130 of these triplets, association for at least one of the pairs, (*s*,*t*) remains significant ($P<0.05$) even upon conditioning on the transcript level of *t'* (see Materials and Methods section 2.4.5). These triplets are further sub-divided into the 101,762 'bi-directional' triplets versus the remaining 27,368 'uni-directional' (for definitions see Materials and Methods section 2.4.5).

We describe independence of associations in each module M as a graph *G(M)* (see Materials and Methods section 2.4.5), when examining the topology of the modules, we notice that for most modules, nearly all association pairs are mutually independent (Figure 2-3 and Figure 2-4).

Furthermore, considering all possible pairs of transcripts in a module, the fraction of them which were connected by edges is 87.7% (averaged across all modules; SD 13.3%). This is significantly more than those found in permuted data (average 12.5%; SD 6.2%). Specifically, both bi-directional (average 79.4%, SD 18.9% versus average 2.3%; SD 3.1%), as well as uni-directional edges (average 8.3%; SD 6.3% versus 10.2%; SD 5.2%) are enriched in real compared with permuted data (Figure 2-3). This is consistent with the main SNP affecting expression levels of most transcripts in its module in a simultaneous rather than a cascaded manner. This also addresses concerns of artifactual modules that are possibly just clusters of co-expressed genes rather than truly independent association to the main SNP.

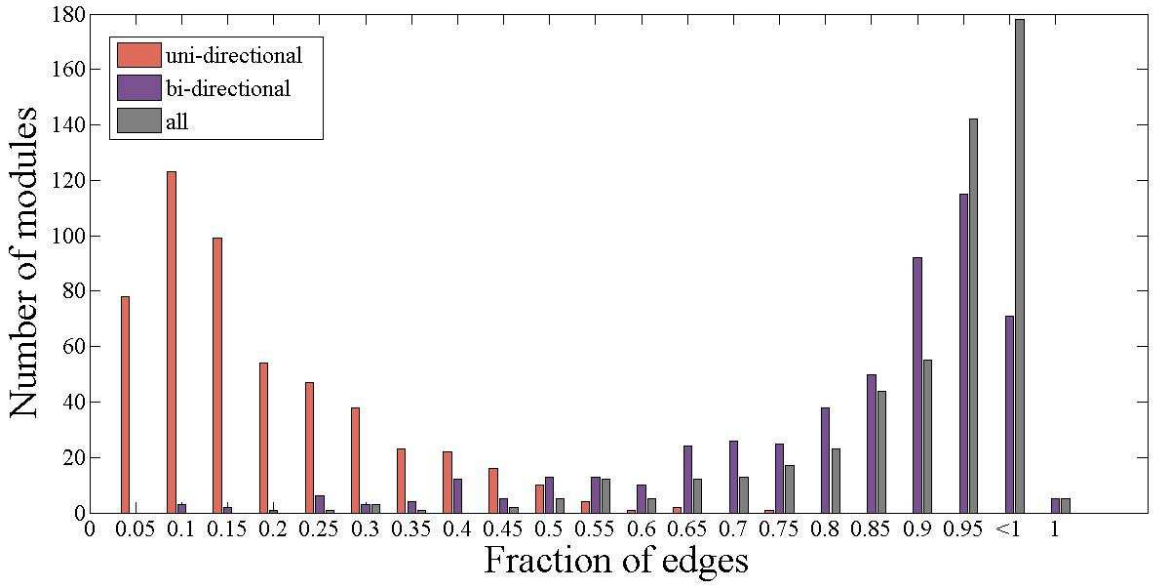


**Figure 2-3. The distribution of the number of modules with different fractions of edges.**
This figure shows the distribution of the number of modules with different fractions of uni, bi- and all edges represented in pink, purple and gray, respectively in each one of the 518 large modules.

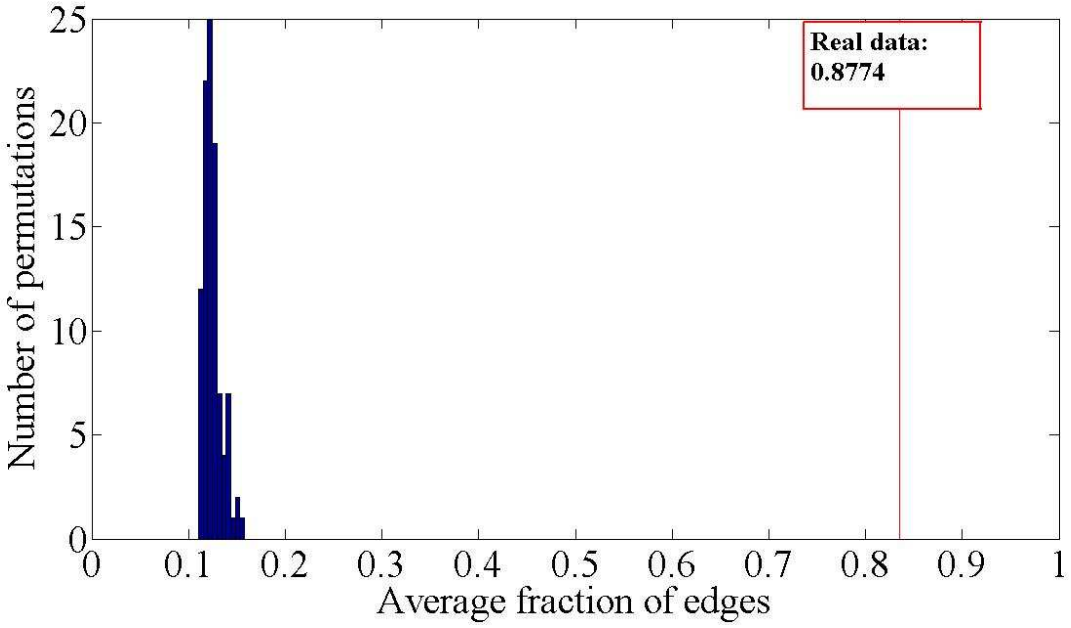


**Figure 2-4. The average fraction of edges.**
Sum of directed and bidirectional out of all possible edges, for 518 modules with 10 or more transcripts in 100 permutations.

### *2.2.3 Module's score and filtering*

To establish a measure of confidence in the resulting modules, we assign a score to each module, considering the module size and the strength of associations between the main SNP and each of the transcripts in the module. This score is justified as a log-likelihood-ratio that compares two hypotheses (see Materials and Methods section 2.4.3). We provide an empirical P-value interpretation by scaling the scores of modules in the real data, compared with the average score of the modules in permutations. We further prune the large modules, defining a subset of 114 high confidence modules with FDR<0.02 (Figure 2-5).

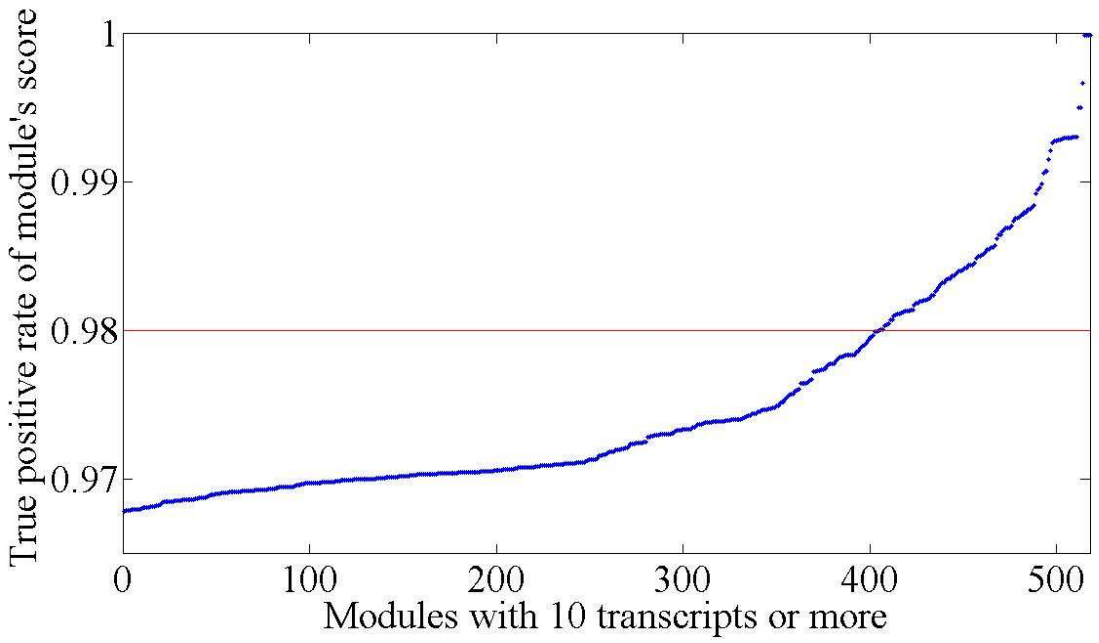


**Figure 2-5. Scaling score for modules.**
For 518 modules with more than 10 transcripts. The red line indicates the FDR threshold of 0.02.

We notice that in most of the modules there are few transcripts that are expressed in an opposite direction to the majority of transcripts in the module. This suggests that the main SNP affects the majority of transcripts in the same direction. We verify this observation by quantifying the percentage of positive and negative correlation of the main SNP with the transcripts in each module (Figure 2-6).

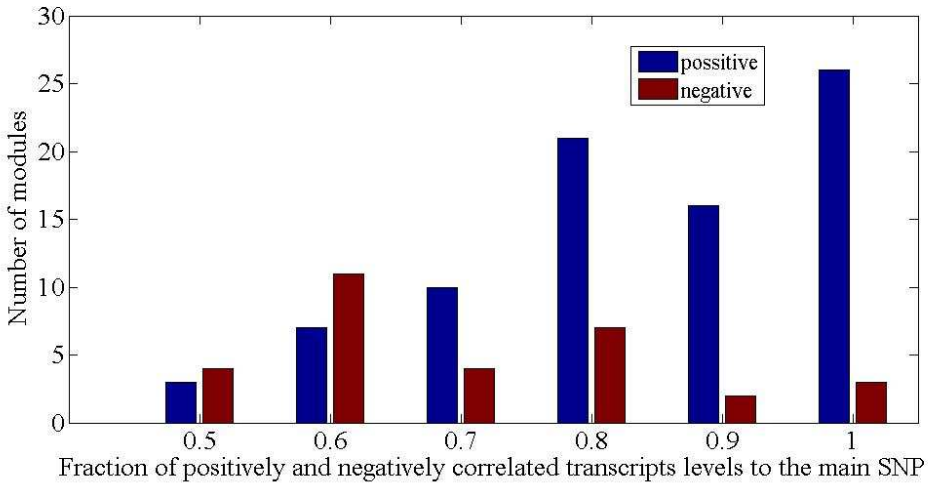


**Figure 2-6. Direction of effect the main SNP has on transcripts in the module.**
The distribution of the number of modules with different fractions of positively (blue) and negatively (red) correlated transcript levels to the main SNP for each one of the 114 modules.

### *2.2.4 Cis/trans-effects*

Some of the previous studies have optimized power to detect *cis*-regulatory variation by using different P-value threshold for defining *cis* eSNPs [49], based on strong priors in their favor [17]. Here, we set a fixed threshold of $10^{-5}$ for both *cis*, and *trans* association, putting them on equal footing for the detection of modules.

There are 110 modules with *trans* main SNP, the remaining 4 modules have *cis* main SNP (see Materials and Methods section 2.4.6 for definitions). We systematically sought potential *cis*-effects of main SNPs that were not strong enough to be captured by our first-pass analysis. To examine this, we record the gene closest (see Materials and Methods section 2.4.6) to each main SNP. In two modules, the main SNPs did not have a close gene from our data. The main SNPs of the remaining 112 modules have 94 unique closest genes, which we call 'main genes'. Out of all main SNPs, 88 are at least 1Mb apart from one another (see Materials and Methods section 2.4.7

for additional details). We record the P-value for the linear regression between each main SNP and the expression levels of its closest gene. In total, 24 main SNPs were nominally ($P$<0.05) *cis* associated to their respective closest gene, with 14 unique associated genes ($P$=1.76×10$^{-4}$, see Materials and Methods section 2.4.8) and with 10 unique associated SNPs that are at least 1Mb apart from one another ($P$=8.1×10$^{-3}$, see Materials and Methods section 2.4.8). These main SNPs are *trans* main SNPs. These results support our suggested *trans*-effect model.

### *2.2.5 Independent cross validation by similar annotations from two sources of information and phenotypic analysis*

We characterized high confidence modules by considering two sources of information:
(i) the enrichment of transcripts in a module for membership in gene-sets from the Gene Ontology [60], NCBI Gene and KEGG [59] databases.
Of the 114 modules, 26 (22.8%) were reported as enriched in any category. This contrasts with modules in 100 permuted data sets, where 12.8±2.7% of the modules show any functional enrichment (Figure 2-7) and
(ii) locus annotation of the main and secondary SNPs of each module, as reflected in the existing literature, Ensembl [61] and wikigenes [62].
These sources are independent for modules with *trans* main SNP. We observe similar annotations of modules from the two sources of information. This independent cross validation provides support for our methodology.
Additional support comes from intersecting the 94 main loci with the 2,626 unique genes (2,212 among the 18,873 transcripts available for analysis in this work) reported to house GWAS SNPs

[16]. We find an overlap of 21 genes (hypergeometric $P$=1.1×10$^{-3}$). We discard 19 modules whose set of transcripts have a 90% overlap with other modules, resulting in 95 distinct modules (see Materials and Methods section 2.4.7 and Tables 2-1a and 2-1b for full listing of all 95 modules). We present details of the annotation analysis for three modules: the largest with an annotated *cis*-SNP, and two of the four largest modules overall.

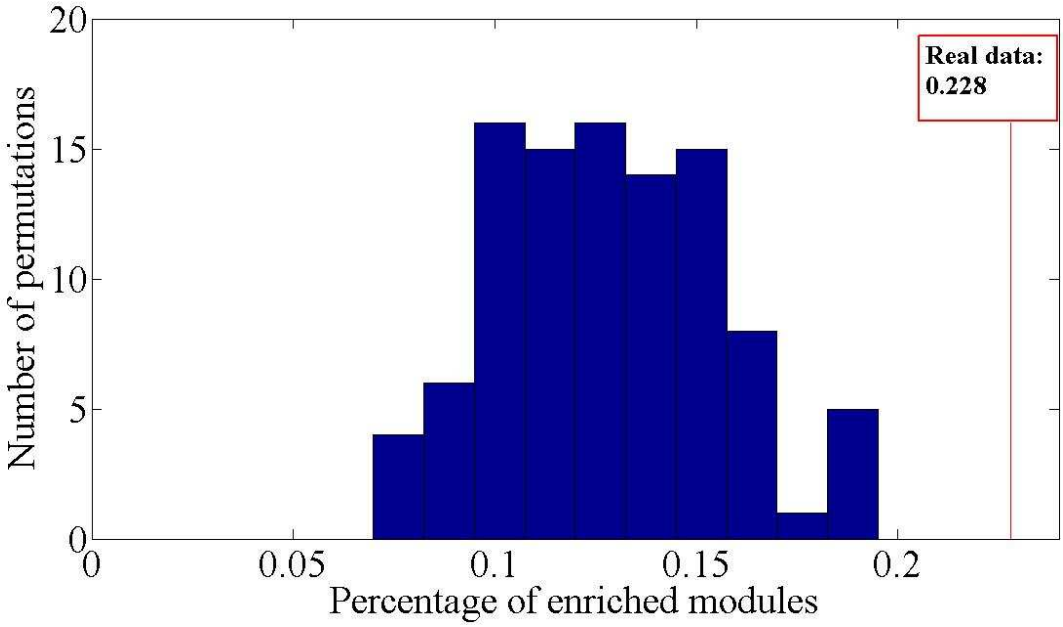


**Figure 2-7. The percentage of enriched (large) modules in real data compared to 100 permuted datasets.**

**Table 2-1. Modules' annotations (separate file).**
File: Table5a-Filtered_modules_GO_enrichment.xlsx. 95 modules full information: module number (decreasing size), #transcripts and Entrez IDs, (a) transcripts' enrichment, main SNP number and position (a) Closest gene to main SNP: name, position and description, secondary SNPs and number of correlated transcripts (a) Closest gene to secondary SNP: name, position and description and the fraction of edges. We indicate the enrichment of the big module when other modules are included within it. Biological information regarding transcripts, was extracted from genecards [62]. SNPs locations were extracted from Ensebml [61]. We represent a group of similar modules by one module that is highly enriched in gene sets. We denote the Entrez Ids, main SNP and fraction of edges for all similar modules in the group.

### *2.2.6 Comparison with standard approach to module construction*

We implemented the standard approach of grouping genes according to their associated SNP. We used a standard, stricter FDR cutoff of 10% for association–pairs [52]. We show this approach to produce fewer modules, smaller modules, limiting its use for finding modules. Moreover, our approach finds modules that are more enriched for functional annotation categories, compared with the standard approach, supporting our modules being genuine.

Specifically, the standard approach produced 22,015 association pairs, 3,387 modules, 75 with 10 transcripts or more (Figure 2-8). The largest module has 27 transcripts. We examine the enrichment of these modules in GO categories and KEGG pathways: 4 out of the 75 modules had significant biological enrichment in at least one category (5.3% comparing with 22.8% functional enrichment in our modules).

*Support for modules filtering step*: All four modules that were found by the standard method and were functionally enriched are contained in one of our final 95 modules. This provides a support for our module scoring and filtering step.

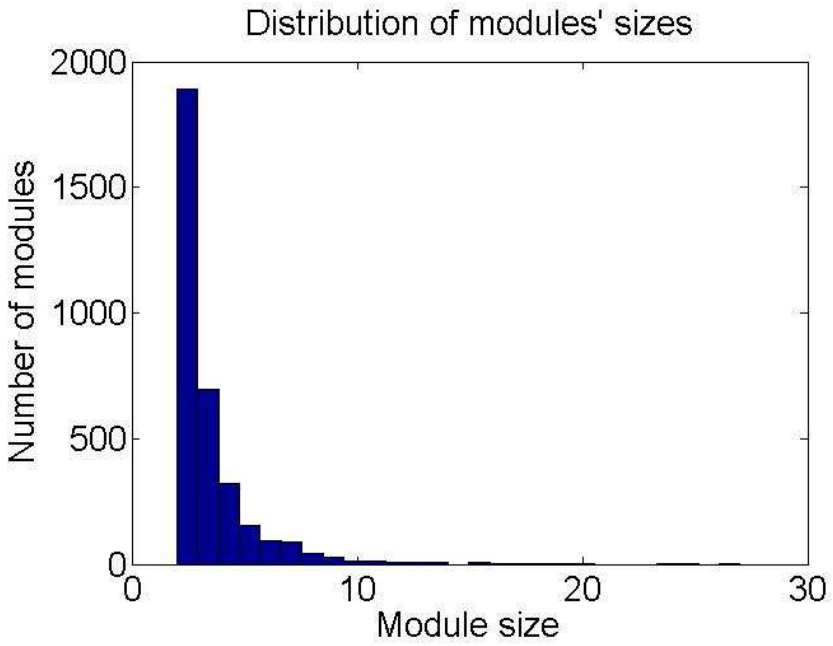


**Figure 2-8. Modules sizes distribution using the standard approach for modules reconstruction.**

### *2.2.7 Analysis of specific modules*

We present a positive control for our method using module #29 with 16 transcripts and *cis* main SNP. The main SNP rs9267658 partitions the samples into three groups: 277 samples that are homozygous C (C/C), 89 C/T samples and 5 T/T samples. The secondary SNP for the C/T subgroup of samples is rs4902609 and is associated with eight transcripts. This module is enriched for Major histocompatibility Complex (MHC) genes (FDR 0.0049), with related annotation for relevant KEGG pathways (allograft rejection—FDR 0.0046, antigen processing and presentation—FDR 0.0041, cell adhesion molecules—FDR 0.0088) and autoimmune diseases (graft-versus-host disease—FDR 0.0027, type I diabetes mellitus—FDR 0.0021, thyroid disease—FDR 0.0023, viral myocarditis—FDR 0.0036 and asthma— FDR 0.045). The main SNP resides within the MHC region [63]. The module includes three transcripts in *cis* to the main SNP that play a central role in the immune system: HLA-DRB5, HLA-DRB4 are MHC

class II and HLA-G is MHC class I. The closest gene to rs4902609, RAD51L1 is a tumor suppressor gene, whose *trans*-association to the MHC transcripts may relate to previous reports on links between autoimmunity and cancer [64] (Figure 2-9).

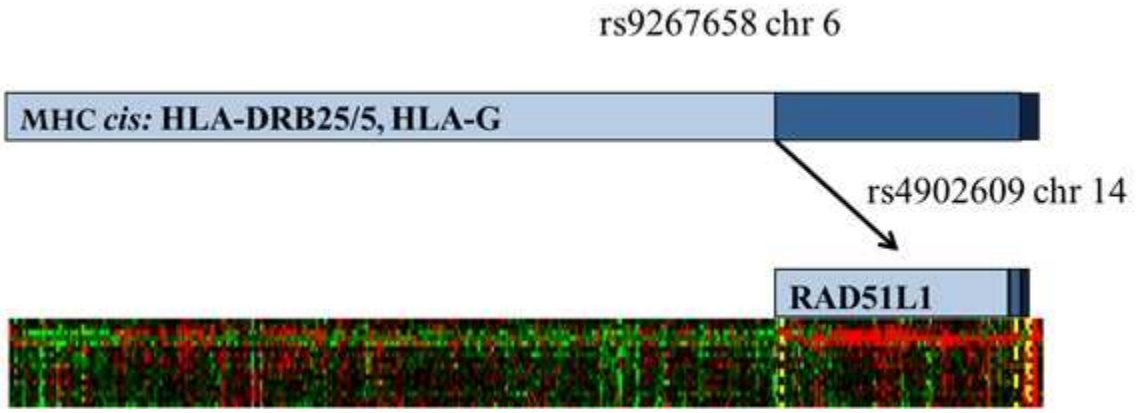


**Figure 2-9. Module of size 16 transcripts and their expression levels over 371 samples.**
The heatmap of expression levels (red/black/green) across samples (columns) and genes (rows) is segmented (top) into SNP– genotype splits—light, medium and dark blue represent carriers of 0, 1 or 2 minor alleles, respectively. Closest genes to the main and secondary SNP are listed.

The largest module (#1) has 91 transcripts. The main SNP rs10818053 partitions the samples into 303 T/T samples, 65 T/C samples and 3 C/C samples. The secondary SNPs are rs6433115 for the major-allele homozygotes and rs2122013 for heterozygotes. This module is enriched for transcripts involved in oxidation reduction (FDR $5.9\times10^{-15}$), lipid metabolic processes (FDR $1.9\times10^{-5}$) and genes expressed in the mitochondrion (FDR 0.015). In terms of pathways, it is enriched for drug metabolism pathways (FDR $7.7\times10^{-5}$) and primary bile acid biosynthesis (FDR $3.4\times10^{-4}$) that occurs in the liver. The closest gene to the main SNP, TLR4 cooperates to mediate the innate immune response to bacterial lipopolysaccharide (LPS). TLR4 activation mediates liver inflammatory response [65] and is responsible for oxidized phospholipid-mediated inhibition of TLR signaling [66]. Secondary SNP rs6433115 for the T/T subgroup is associated

with 26 transcripts and is within the span of LRP2. Secondary SNP rs2122013 for the T/C subgroup is associated with 35 transcripts and is closest to MTX2 gene (Figure 2-10). LRP2 is a lipoprotein that is also involved in the cellular uptake of drugs, including lipid-based formulations [67]. MTX2 is involved in the import of proteins into the mitochondrion [62]. This module may be related to the effect of drugs on lipid metabolism [68] and the possible role of the mitochondrion in such pathways [69].

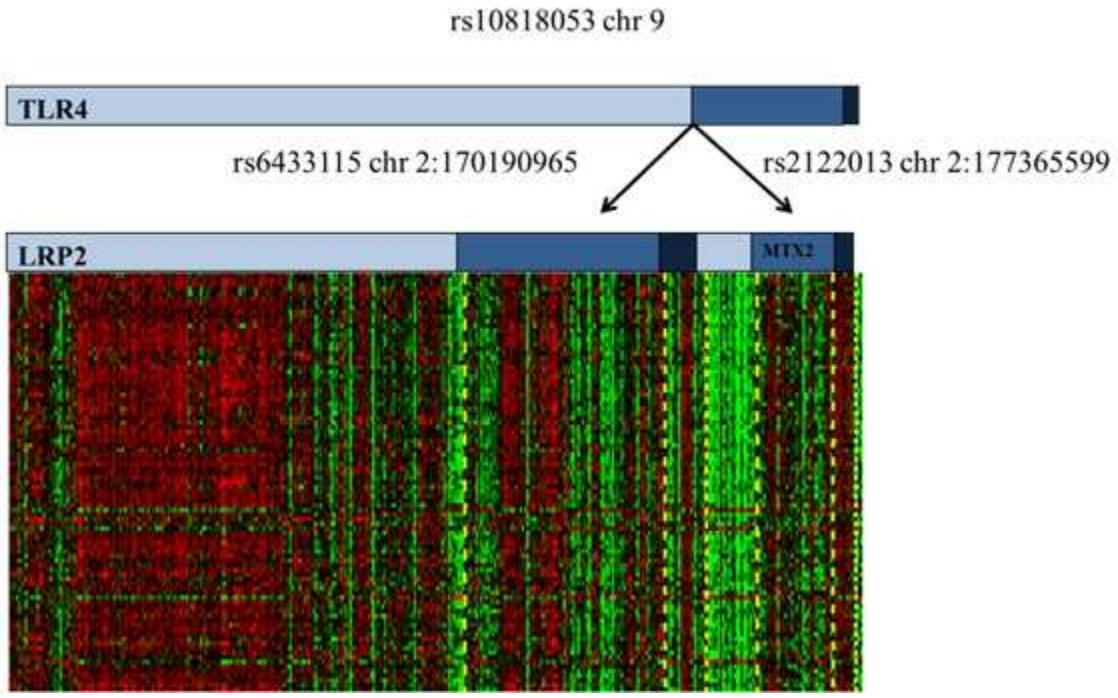


**Figure 2-10. The largest module of 91 transcripts and their expression levels over 371 samples.** See Figure 2-9 legend for further details.

Since mutations in TLR4 are associated with liver damage, we investigate the main SNP's association to drug sensitivity. Data for liver risk in the 371 samples [52], genotype of the main SNP rs10818053 and liver risk in 371 samples are detailed in Table 2-2. The clinical data presented by Schadt et al. [52] for liver risk, are binary entries describing (according to

clinicians' diagnosis) if there is a risk to the patient's liver if treated by drugs. We present preliminary analysis showing that these minor–minor and major–minor allele samples are enriched for liver risk more than is expected by chance (Hypergeometric $P<0.012$) which implies that individuals carrying C/C or T/C alleles in the main SNP's locus may be prone to liver sensitivity for drug treatment. This analysis provides the first support for our method from non-expression traits.

| ***rs10818053 genotype*** / ***Liver risk*** | Minor–minor C/C | Major–minor T/C | Major–major T/T | Total no. of samples |
|---|---|---|---|---|
| Positive | 2 | 13 | 39 | 54 |
| Negative | 1 | 52 | 264 | 317 |
| Total no. of samples | 3 | 65 | 303 | 371 |

**Table 2-2. Data for liver risk in 371 samples.**
Separated by minor– minor, major–minor and major–major allele samples, respectively and genotype of rs10818053.

Module #4 has 50 transcripts. The main SNP rs1477511 partitions the samples into 288 T/T samples, 76 T/G samples and 7 G/G samples. The secondary SNPs are rs6464842 for the first subgroup and rs861508 for the second subgroup and are associated with 7 and 9 transcripts, respectively (Figure 2-11). This module is enriched in transcripts that regulate cellular (FDR 0.0036) and metabolic processes (FDR 0.013), specifically cell proliferation and differentiation (FDR $5.2\times10^{-5}$). It is enriched for ErbB (FDR $1.5\times10^{-3}$) and Mitogen activated protein kinase (MAPK) signaling pathways (FDR $5.2\times10^{-3}$). The closest gene to the main SNP, STK11IP interacts with LKB1 which regulates cell polarity and functions as a tumor suppressor [62]. LKB1 is a serine/threonine kinase which is inactivated by mutation in the Peutz– Jeghers

polyposis and cancer predisposition syndrome (PJS) [70], with correlation to the putative function of the module. We observe a significant P-value (<0.031) between the expression levels of LKB1 and the genotype of rs1477511. Mutations in CNTNAP2, where rs6464842 resides, have been implicated in multiple neurodevelopmental disorders, including attention deficit hyperactivity disorder (ADHD) and schizophrenia. With correlation to CNTNAP2 function, the ErbB signaling was suggested to impair working memory and executive functions that are affected in schizophrenia, ADHD and other psychiatric disorders [71]. SPANXC which is the closest gene to rs861508 resides in a region that confers susceptibility to prostate cancer. ErbB and MAPK signaling are known to have an important role in cancer [72, 73].

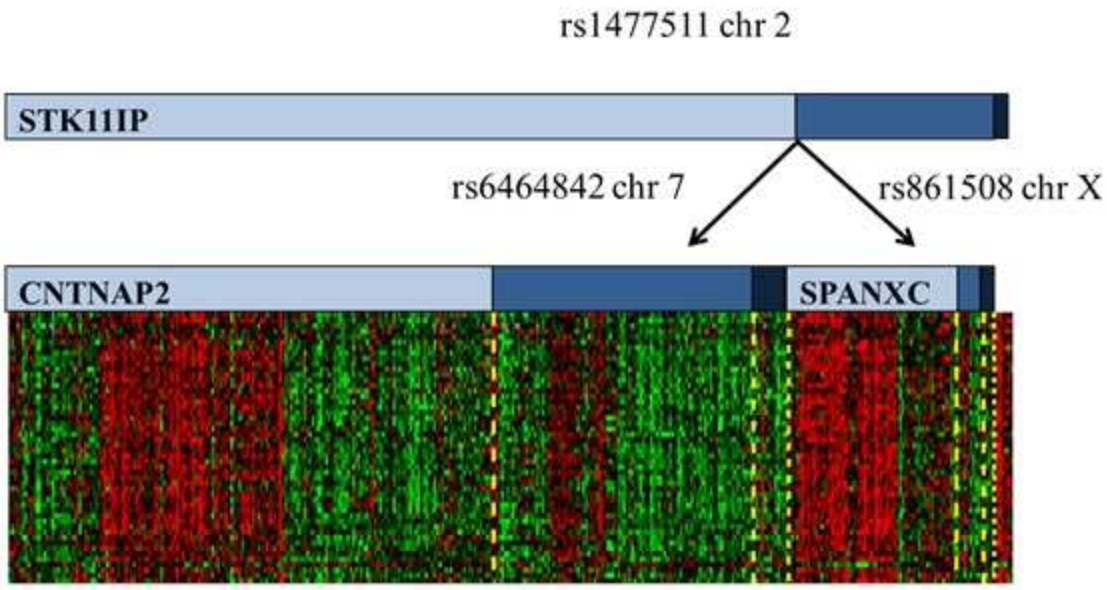


**Figure 2-11. Module of size 50 transcripts and their expression levels over 371 samples.**
See Figure 2-9 legend for further details.

Finally, we present a second support for our method from non-expression traits. Module #101 with 10 transcripts is the only module where the main SNP maps to a locus associated with oxidative damage control: rs1453226 at OXR1 indicated to be involved in protection from oxidative damage [62]. The transcripts in this module are slightly enriched for oxoacid metabolic process (FDR 0.04). Therefore, we decided to investigate its association to alcohol risk. Data for alcohol risk in the 371 samples [52], genotype of the main SNP rs1453226 and alcohol risk in minor–minor allele samples are detailed in Table 2-3. It is challenging to provide clinical support, since the clinical data presented by Schadt et al. [52] is very sparse. We present preliminary analysis showing that these samples are enriched for alcohol risk more than is expected by chance (Hypergeometric $P<0.03483$), which implies that individuals carrying A/A alleles in the main SNP's locus may be prone to sensitivity for alcohol use.

| ***rs1453226 genotype*** / ***Alcohol risk*** | Minor–minor A/A | Major–major G/G and Major–minor G/A | Total no. of samples |
|---|---|---|---|
| Positive | 4 | 15 | 19 |
| Negative | 4 | 93 | 97 |
| Unknown | 28 | 227 | 255 |
| Total no. of samples | 36 | 335–195 and 140, respectively | 371 |

**Table 2-3. Data for alcohol risk in the 371 samples.**
Genotype of rs1453226 and alcohol risk in minor–minor allele samples.

## 2.3 Discussion

We presented a three-step approach to the analysis of eSNPs and their relation to phenotypes that goes beyond documenting associations of each to expression levels, by applying a module score filtering procedure, and complements co-expression networks by unraveling module topology. As a first step, we assemble transcripts associated to the same main eSNP into the modules. We then filter the reported modules by a confidence score, and finally associate subgroups of transcripts within a module with additional variants conditioned on the genotype of the main SNP.

We apply our method to data on human liver expression and SNP genotypes [52]. We find that the number of association pairs of eSNP and transcript is consistent with the null expectation, whereas assembled modules are significantly more numerous, bigger and denser than those observed in the permuted data. This indicates modules are not random clusters of correlated-expression genes, but rather show truly independent association to their main SNP. We compare our results with a standard approach that maps transcript-eQTL pairs with a standard FDR (e.g. 10%) and forms groups consisting of transcripts that share an eQTL. We observe smaller number of modules, smaller in size and significantly less enriched in Biological categories.

Our method detects 95 distinct modules; out of those, only one has a main SNP in cis to module transcripts. Among the remaining 94 *trans* main eSNPs, we observe enrichment for milder, not genome-wide significant *cis*-effects that explain the *trans*-effect of the main SNPs on transcripts in the associated modules. We characterize modules by two sources of information that are independent for modules with a *trans* main SNP: enrichment in subsets of genes and locus

annotation of the main and secondary SNPs. We observe similar annotations from both sources of information. Thus, providing support for our method. We present detailed analysis of four modules: annotation analysis for three of the four modules: one with a *cis* main SNP and two with *trans* main SNPs, and phenotypic analysis for two of the four modules.

This study holds the promise for extension beyond its current limitations. The current analysis focuses on transcripts that are directly regulated by a variant. Mining the data for additional transcripts that are downstream along the same pathway of regulation, e.g. by consideration of co-expressed genes with milder association to the main SNP can complement reverse engineering of the regulatory program [50]. Furthermore, both the raw data sets [52] and supporting databases [59-61] in this work are noisy and limited. Potential increase in sample size for eQTL data may enable detection of eSNP associations at more significant P-values for even milder effects. Likewise, as the functional annotation continues to build up, better understanding of modules would be facilitated.

Future studies could extend the approach presented here to investigate how modules correlate with phenotype, for example, using the data on enzymatic activity that was presented by Yang et al. [55]. As data becomes available, comparison of modular structure between healthy and affected samples, as well as across different tissue types is likely to improve understanding of disease and developmental regulatory processes. It remains a significant challenge to validate the results presented here by experimental means, and analysis of independent data may provide such validation by replication.

# 2.4 Materials and Methods

### *2.4.1 Data details and processing*

The DeLiver data set by Merck had been described elsewhere [52]. Briefly, the raw data set consists of 653,894 SNPs and 25,917 expression probes (log-transformed values) with an Entrez gene ID assayed for 385 samples. We remove 99 expression probes that are mapped to the Y chromosome. Multiple probes that are mapped to the same gene had been averaged if correlated ($r>0.75$) or discarded otherwise, resulting in 18,883 genes with unique Entrez IDs. 5,055 genes had variable levels of liver expression across the individuals ($SD>0.2$). Standard filters have been applied to the SNP data: Minor allele frequency>0.05, SNP missingness rate <0.1 and individual missingness rate <0.1 [74]. After filtering, the data for analysis consists of 371 samples (200 males, 171 females) with 557,456 SNPs and 5,055 genes.

For each individual *i*, we denote the expression levels of each transcript *t* by *X(i,t)*, and the genotype for each SNP *s* by *G(i,s)*.

### *2.4.2 Step 1—nominal association testing*

We test for association between pairs (*s*,*t*) of any SNP *s* and transcript *t* using linear regression and record the results between every (*s*,*t*) pair with nominal $P<10^{-5}$. To eliminate transcripts whose association statistic is strongly distorted, we repeated the analysis 1,000 times with permuted data, obtained by randomly switching the samples' labels, discarding recurrently observed transcripts as follows. A small fraction of observed association pairs tend to recur in permuted data sets more than expected (Table 2-4). Specifically, 2,979 of the observed association pairs detected in the real data appear exactly once in the 1,000 permuted data sets

(<676 expected), and 520 recur twice (<7 expected). This suggests a bias in the test statistic for these pairs, and we discard all 623 pairs that appear in two permutations or more from subsequent analysis.

| ***#permutations*** | ***Number of pairs*** | ***Expected #permutations*** |
|---|---|---|
| 1 | 2,979 | < 676 |
| 2 | 520 | < 7 |
| 3 | 87 | < 1 |
| 4 | 14 | < 1 |
| 5 | 2 | < 1 |

**Table 2-4. Distribution of observed and expected association pairs in 1000 permutations.**

When considering association pairs detected in the real and permuted data, we note that over dispersion of the test-statistic exists in both. In the real data, $10^{-4.61}$ of (*s*,*t*) pairs attain a test statistic theoretically corresponding to a $P=10^{-5}$ (Figure 2-12), whereas in the 100 permutations using all SNPs in the data, only $10^{-4.65}$ of such pairs attain this level. We use the nominal $P=10^{-5}$ as a threshold, keeping in mind that this P-value is not genome-wide significant, and 69,172 random association pairs are expected to pass this threshold by chance alone. This justifies the use of such a threshold, as our methodology relies on having a variety of association pairs, that only when cross-compared across transcripts would yield a meaningful result.

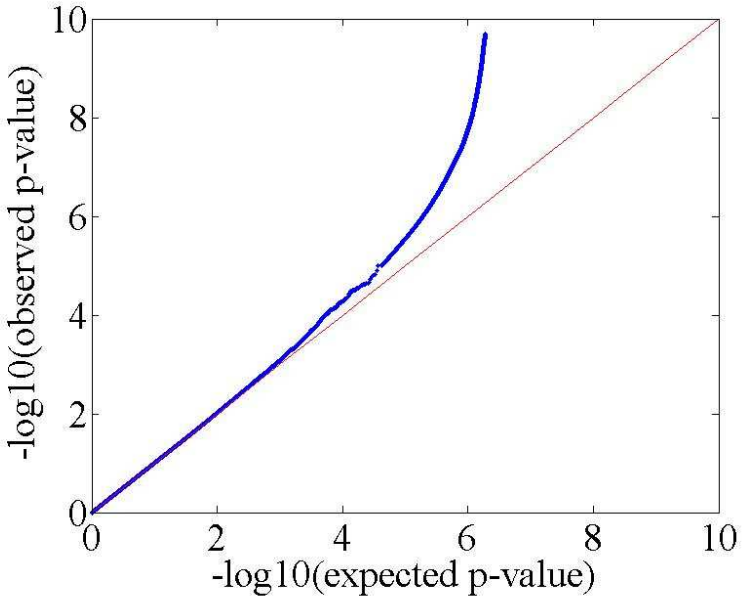


**Figure 2-12. QQ plot for association pairs in real data.**
X-axis denotes -log 10 of the expected p-value. Y-axis denotes -log10 of the observed p-value which represent 100 transcripts that were sampled randomly, and 1/100 of each p-value range was sampled. Also, out of all p-values better than $10^{-5}$, 1/100 were sampled randomly.

### *2.4.3 Step 2—module construction, scoring and filtering*

The putatively associated transcripts are binned by their SNP *s*, each bin hereby referred to as a module. This associated SNP s is denoted as the 'main' SNP. We consider each module in turn. Let *M* be a module of size *k*, with a set of transcripts $\{t_1, \ldots, t_k\}$ and a main SNP *s*. For each transcript $t_i$ we consider the P-value denoted Pval($t_i$) of the association test between the main SNP *s* and its expression level. We compute the empirical false positive rate (EFPR) for each such P-value by permutation: We use 100 permutations to tally the average number of P-values better than Pval($t_i$) across the permuted data sets divided by the analogous number in the real data. This ratio is the EFPR corresponding to Pval(*ti*). We follow a similar procedure to calculate the analogous ratio for module size k: EFPR(*k*) is defined as the ratio of the average number of

modules with size bigger than *k* across the permuted data sets and the analogous number in the real data. The score *S(M)* of the module *M*

$$S(M) = -\sum_{i=1}^{k} log(\mathrm{EFPR}(pval(t_i))) - \log(\mathrm{EFPR}(k))$$

is justified as a log-likelihood-ratio that compares two hypotheses

$$Likelihood\_ratio = \frac{Likelihood(H_0)}{Likelihood(H_1)} = \frac{\overline{Pr^{perm}(k)}}{Pr^{data}(k)} \cdot \prod_{i=1}^{k} \frac{\overline{Pr^{perm}(pval(t_i))}}{Pr^{data}(pval(t_i))}$$

$H_0$ denotes the null hypothesis that a module size and the strength of associations within the module follow the same distribution in the real and permuted data. $H_1$ denotes the alternative hypothesis, i.e. that a module size in the real data would be larger than in the permuted data, as well as the strength of the associations within it.

In order to assign significance to the obtained scores, we again use 100 permutations. We score each of the modules in the permuted data sets against the other 99 (a 'leave one out' procedure) in a similar process to the one described for computing the scores of modules in the real data. We thereby provide an empirical P-value interpretation by scaling the scores of modules in the real data, compared with the average score of modules in permutations, i.e. the true positive rate (TPR) of the score of a module.

### *2.4.4 Step 3—finding secondary SNPs*

We split the samples by the genotype of the main SNP into three subsets of samples with genotypes AA, Aa and aa, respectively (where A and a are the major and minor alleles, respectively). AA and Aa are the two larger subsets of samples. In each of those two subsets, we then turn to find the corresponding two subset-specific SNPs that best explain the expression of the largest group of genes in each subset, and denote these 'secondary' SNPs [50, 75]. To search for secondary SNPs, we test each SNP for association only to the transcripts within the module, and only within the current subset of samples. We discard pairs of transcript and SNP in recurrently observed association pairs by using 1,000 permutations and removing all association pairs that appear in one permutation or more (empirical FDR<0.001). We consider all SNPs that comply with three criteria: (i) maximal-size subgroup of transcripts (with minimum of five transcripts), (ii) F-test for independent association of transcript pairs and (iii) minimal product of association P-values. More specifically: For each module, and each genotype group we first list all SNPs that achieve an association nominal P-value of $10^{-5}$ or better with a large subgroup of transcripts (five transcripts or more). We consider only those whose subgroup is maximal as candidate secondary SNPs. We test all possible pairs of transcripts in the subgroup for conditional association (see Materials and Methods section 2.4.5), and discard a candidate secondary SNP if any pair fails the test. Out of this list, we seek the SNP with the minimal product of association P-values with its subgroup of transcripts. These steps control for false discovery, because the phenomena of big and edge dense modules does not exist in permutations.

### *2.4.5 Analysis of dependencies within modules*

For each module, we consider all possible ordered triplets (*t,t',s*) of two transcripts *t*, *t'* whose levels are significantly associated with the same main SNP *s*. We define bi-directional triplets where association is mutually independent, i.e. for both association pairs remain nominally significant given the respective other transcript versus 'uni-directional' triplets where association is directionally independent (Figure 2-13a). Formally, we test whether the association model provides significantly better fit to the data than the null model.

Null Model: $X(i,t) = \alpha_0 + \alpha_1 \cdot X(i,t') + \varepsilon_1$

Association Model: $X(i,t) = \beta_0 + \beta_1 \cdot X(i,t') + \beta_2 \cdot G(i,s) + \varepsilon_2$

We use the F-test for better fit symmetrically, attempting to explain the expression levels of either *t* by *t'* or the converse, with or without genotypes (testing the significance of $\beta_2$ being non-zero coefficient would yield the same results). We describe independence of associations in each module *M* as a graph *G(M)*, whose vertices correspond to transcripts. A directed/bidirectional edge connects transcripts with directionally/mutually independent association with the main SNP (Figure 2-13b).

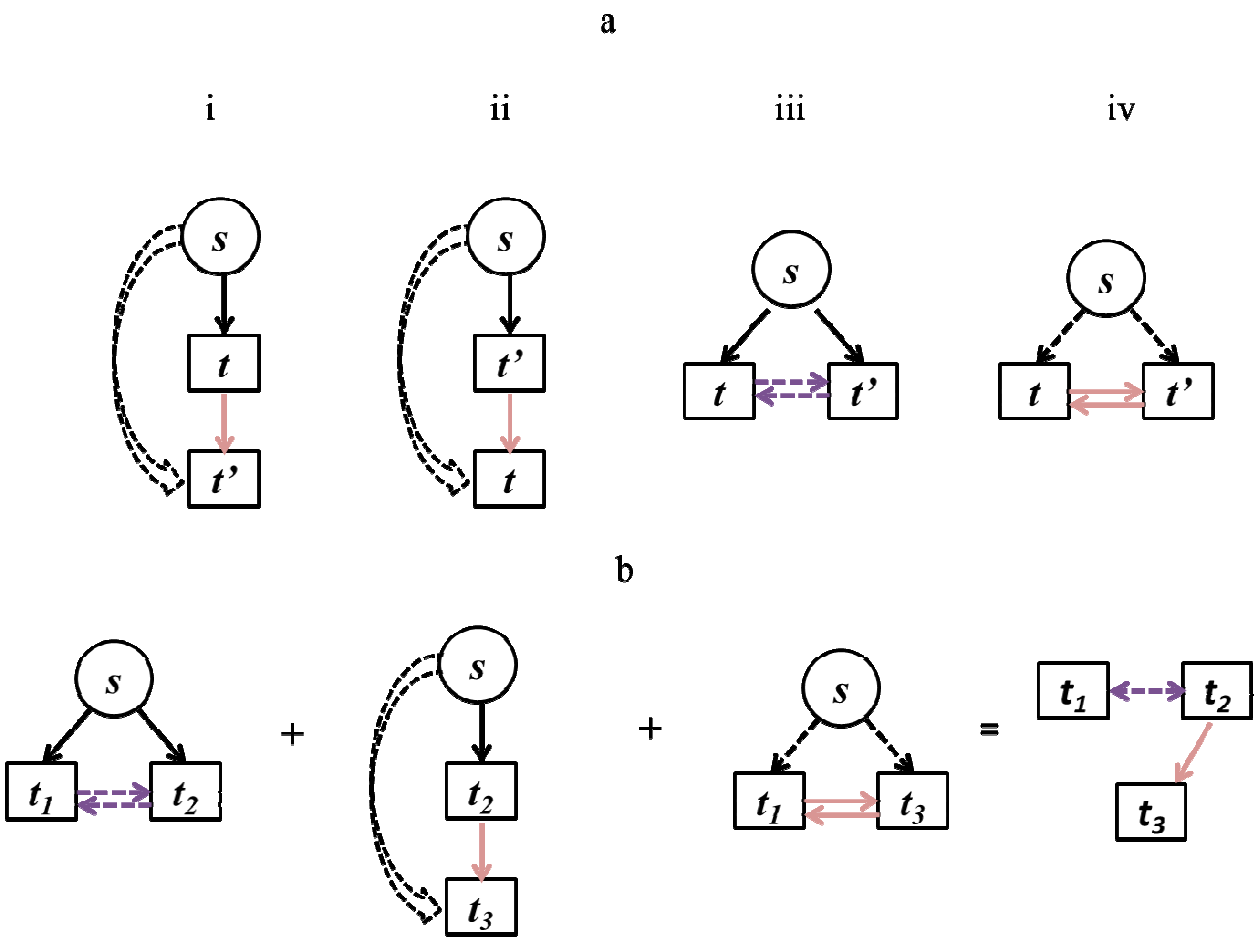


**Figure 2-13. Graphical representation of modules.**
**(a) Graphical illustration of a triplet with two transcripts *t* and *t'* and a main SNP *s*.** The dashed/full black line represents dependent/ independent association between a SNP and a transcript, respectively. The uni/bi-directional pink/purple line represents an edge that connects transcripts with directionally/mutually independent association to the main SNP (i) unidirectional triplet—the association pair (*s*, *t*) remains significant ($P<0.05$) even upon conditioning on the transcript level *t'*, but not vice versa. (ii) unidirectional triplet (*s*,*t'*) remains significant even upon conditioning on the transcript level *t*, but not vice versa. (iii) bi-directional triplet (*s*,*t*) remains significant even upon conditioning on the transcript level *t'* and (*s*, *t'*) remains significant even upon conditioning on the transcript level *t*. (iv) dependent triplet (*s*,*t*) and (*s*,*t'*) are insignificant ($P>0.05$) when conditioning on the transcript level *t'* and *t* respectively. **(b) Graphical representation of intra-module interactions.** We consider a module with three transcripts: $t_1$, $t_2$, $t_3$ and a main SNP *s*. A bi-directional dashed purple edge is placed between transcripts $t_1$ and $t_2$, representing the mutually independent association of both $t_1$ and $t_2$ with the main SNP *s*. A directed solid pink edge is placed between transcripts $t_2$ and $t_3$, representing the dependent association of (*s*, $t_3$) on the transcript levels of $t_2$. No edge is placed between transcripts $t_1$ and $t_3$, representing the mutually dependent association of both $t_1$ and $t_3$ with the main SNP *s*.

### *2.4.6 Module annotation*

The enrichment of a module in gene subsets from the Gene Ontology (GO) [60], and KEGG [59] databases was calculated using DAVID [76, 77]. The enrichment of real and permuted modules in gene subsets from the NCBI gene database was calculated using LitVAn [75]. We report only modules with annotations that have a significant FDR of 0.05 or better. Depending on context, we discuss the proximity of a gene to a SNP in several ways: A SNP may be 'in the span of the gene', i.e. the SNP resides between the ENSEMBL [61] transcription start site and stop codon of the gene; 'closest to the gene', i.e. this gene spans the closest among all spanned sites on either direction; or 'close to the gene'—means the SNP is within 1Mb of a site spanned by the gene. We define a *cis* main SNP when the main SNP is within 1Mb of one or more transcripts in the module. We define a *trans* main SNP when the main SNP is 1Mb or further of all the transcripts in the module.

### *2.4.7 Filtering modules using different criteria*

There are 94 main SNPs have a close gene with a unique Entrez ID and 88 main SNPs that are at least 1MB apart from one another (Table 2-5). We filter all modules that have minimum of 90% overlap with another module, resulting in 95 distinct modules (Table 2-6).

**Table 2-5. Distance filters and *cis* effects (separate file).**
File: Table2-5.xlsx. Presents for each module its serial number, rs#, size, chromosome, SNP position, # group according to rs location, closest gene Entrez ID – a zero entry means there is no closest gene, transcription start site, transcription end site, whether the SNP is located within the transcript region and the *cis* effect p-value.

**Table 2-6. The percentage of overlap between every two modules (separate file).**
File: Table2-6.txt. The percentage is calculated out of the number of transcripts in the smaller module.

### *2.4.8 Enrichment of cis-effects for main SNPs*

We model the examination of *cis*-effects for main SNPs as a binomial experiment. For each main SNP, we record one closest gene. Conservatively, unique genes are tested for association to exactly one main SNP, a binomial experiment Bin(n=number of unique genes, $P$=0.05) with significant number of successes. We then record main SNPs that are at least 1Mb apart from one another and test them for association to exactly one closest gene, a binomial experiment Bin(n=number of main SNPs that are at least 1Mb apart from one another, $P$=0.05) with significant number of successes (Table 2-5 and section 2.2.4 in Results).

# Chapter 3: Co-regulated transcripts associated to cooperating eSNPs define bi-fan motifs in human gene networks

*Summary:* Associations between the level of single transcripts and single corresponding genetic variants, eSNPs, have been extensively studied and reported. However, most expression traits are complex, involving the cooperative action of multiple SNPs at different loci affecting multiple genes. Finding these cooperating eSNPs by exhaustive search has proven to be statistically challenging.

In this paper we utilized availability of sequencing data with transcriptional profiles in the same cohorts to identify two kinds of usual suspects: eSNPs that alter coding sequences or eSNPs within the span of transcription factors (TFs). We utilize a computational framework for considering triplets [36], each comprised of a SNP and two associated genes. We examine pairs of triplets with such cooperating source eSNPs that are both associated with the same pair of target genes. We characterize such quartets through their genomic, topological and functional properties.

We establish that this regulatory structure of cooperating quartets is frequent in real data, but is rarely observed in permutations. eSNP sources are mostly located on different chromosomes and away from their targets. In the majority of quartets, SNPs affect the expression of the two gene targets independently of one another, suggesting a mutually independent rather than a directionally dependent effect. Furthermore, the directions in which the minor allele count of the SNP affects gene expression within quartets are consistent, so that the two source eSNPs either both have the same effect on the target genes or both affect one gene in the opposite direction to the other. Same-effect eSNPs are observed more often than expected by chance. Cooperating quartets reported here in a human system might correspond to bi-fans, a known network motif of four nodes previously described in model organisms. Overall, our analysis offers insights regarding the fine motif structure of human regulatory networks [22].

# 3.1 Introduction

Markers associated with changes in gene expression, called eSNPs have been extensively mapped using high throughput genomic data [1, 36, 45, 57, 71, 78-80]. They allow effectively delineating regulatory associations between each eSNP source and each of its regulated target transcripts. Taken together, these source-target links comprise a regulatory network that abstracts both the genes at source loci as well as their targets as nodes.

Regulatory networks have been characterized as featuring specific motifs as their fundamental building blocks [37, 81]. These motifs occur significantly more than expected by chance and suggest respective functional mechanisms. Specifically, studies in model organisms highlighted the bi-fan motif which consists of two regulators regulating two genes as having a functional role, e.g. of a filter and synchronizer of feedback loop signals [37, 82]. While previously studied networks are often derived from TF-DNA or protein-protein binding experiments, this work utilizes genetics-genomics data to study the bi-fan motif across a human regulatory network.

Model organisms, amenable to pervasive experimental methods, suggest regulatory networks to commonly include structures more complex than single SNP – single gene links, e.g. mapping genetic interactions in yeast [83, 84]. In humans, where experimental approaches are more limited, eSNPs provide natural perturbations that inform us of similar regulatory links and systems. Concerted analysis of a multitude of eSNPs allows better understanding of the interactions that establish their network structure. Statistically, epistatic interaction is defined as the deviation from additivity in a linear model involving two or more loci [85, 86].

Unfortunately, finding such association signal for statistical interaction between a pair of SNPs in even a single phenotype has proven computationally difficult [84, 87-89]. Association analysis across all pairs of SNPs vs. all transcripts exacerbates this tractability problem.

While structures of multiple eSNPs to one transcript offer one lens for genetic-genomic analysis, a complementary perspective is provided by regulatory modules, where a single eSNP is associated to multiple genes [4, 9, 36]. Modularity of gene regulatory networks was shown to be a major organizing principle of biological systems [44], with modules often defining functional units of a biological network: each such units consists of a set of elements (e.g. genes) working jointly to perform a distinct function.

Analysis of single eSNP-single transcript interactions indicates that variation in genomic DNA can affect transcription in multiple ways. Level of transcripts *in cis* of an eSNP may be altered due to allelic variation in *cis*-regulatory elements [90], while *trans* association can, for example, be the result of an eSNP in a transcription factor that regulates the expression of its distal targets transcripts. Associations in *cis* are easier to detect because of favorable testing burden. Unfortunately, such associations are limited in their capacity to inform us regarding the network of regulatory interactions between one gene and another, as both the eSNP and the transcript are from the same locus. In contrast, *trans* eSNPs can identify downstream effects and previously un-annotated regulatory pathways. Moreover, when considering independent association between more than a single eSNP and more than a single gene, the genomic distances between eSNP sources and their gene targets require special attention. In the case of examining a pair of proximal eSNPs, their frequent co-inheritance would induce statistical dependence (linkage

disequilibrium) between them. Thus, for most independent pairs of eSNPs that cooperate in regulating the same transcript, at least one of them will have a *trans* effect.

In our previous work [36], we studied eSNPs associated with simplest modular unit of two transcripts, together creating a *triplet.* We focus on mutually independent triplets, whereby the eSNP association with either of the two transcript remains nominally significant given the respective other transcript, as well as and directionally independent triplets, where only one of these association signals remains nominally significant given the level of the other transcript. We established the occurrence of such triplets in real data significantly more than expected by chance.

In this study, we devise a computational framework for examining pairs of triplets that share the same associated two genes. We hypothesize that such eSNP-transcript *quartets* will highlight true eSNP associations, and demonstrate that by analyzing their distinct topological and functional properties. These properties differ significantly from those of spurious quartets with candidate association signals. Moreover, we replicated those properties in an independent dataset with a larger number of samples [1], supporting the robustness of our findings. In particular, the two eSNPs in a quartet tend to have independent, but consistent effect on the pair of genes they co-regulate.

# 3.2 Results

### *3.2.1 Computational framework for associating pairs of SNPs with pairs of genes*

Definition and discovery of quartets

We used a publicly available classic dataset of 50 fully sequenced Yoruban samples [91] along with their transcription profiles from RNA-seq data [40], bearing in mind that such available cohorts are limited in size. Due to this small sample size, we have limited power in detecting association. Therefore, most candidate eSNPs can only be designated as such with various levels of uncertainty. We demonstrate the ability to recapitulate the observed phenomena in a larger dataset [1] using the same method.

We evaluated two categories of candidate eSNPs that reside within regions along the genome with known regulatory potential, i.e., within the span of known exons and TFs (including introns) (Figure 3-1; see Materials and Methods section 3.4.2). These eSNPs can be associated with the expression of both local and distal genes. We consider all mutually independent and directionally independent triplets (Figure 3-2a, see [36] for details). Going beyond the associations of a single eSNP *source* requires the examination of pairs of triplets that share the same *target* transcripts. We call this arrangement a *quartet* (Figure 3-2b). We aim to study quartets with *cooperating* eSNP *sources*, i.e. SNPs that carry independent information towards predicting the level of each one of the two transcripts, and no third intermediate SNP can explain the expression to either gene better (Figure 3-2c; see Materials and Methods section 3.4.4). We note that such *cooperating quartets* may overlap in their genes, introducing double-counting of the same effect in different quartets. To ensure our analysis involves quartets with distinct targets, we filtered this set of cooperating quartets further and focused on the quartets that have

two unique gene targets. In this workflow, no post-filter quartet has the same pair of gene targets as any other (Figure 3-2c).

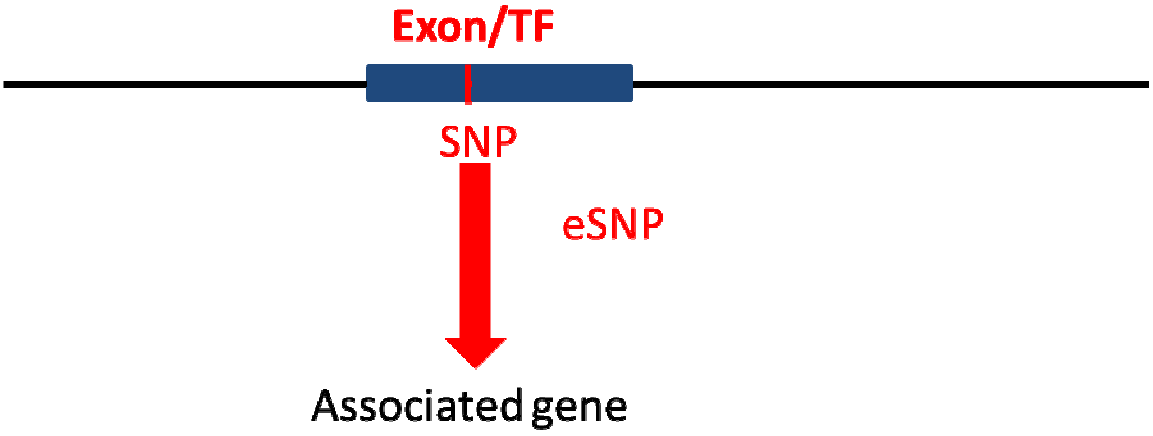


**Figure 3-1. Association testing.**
Illustrating the association testing between pairs of SNPs within known regulatory regions and genes. If the regulatory element has a SNP within the boundaries of an exon or a TF then we check for association ($P < 10^{-4}$ denoted by a red edge) using linear regression between the minor allele count of the SNP and any gene.

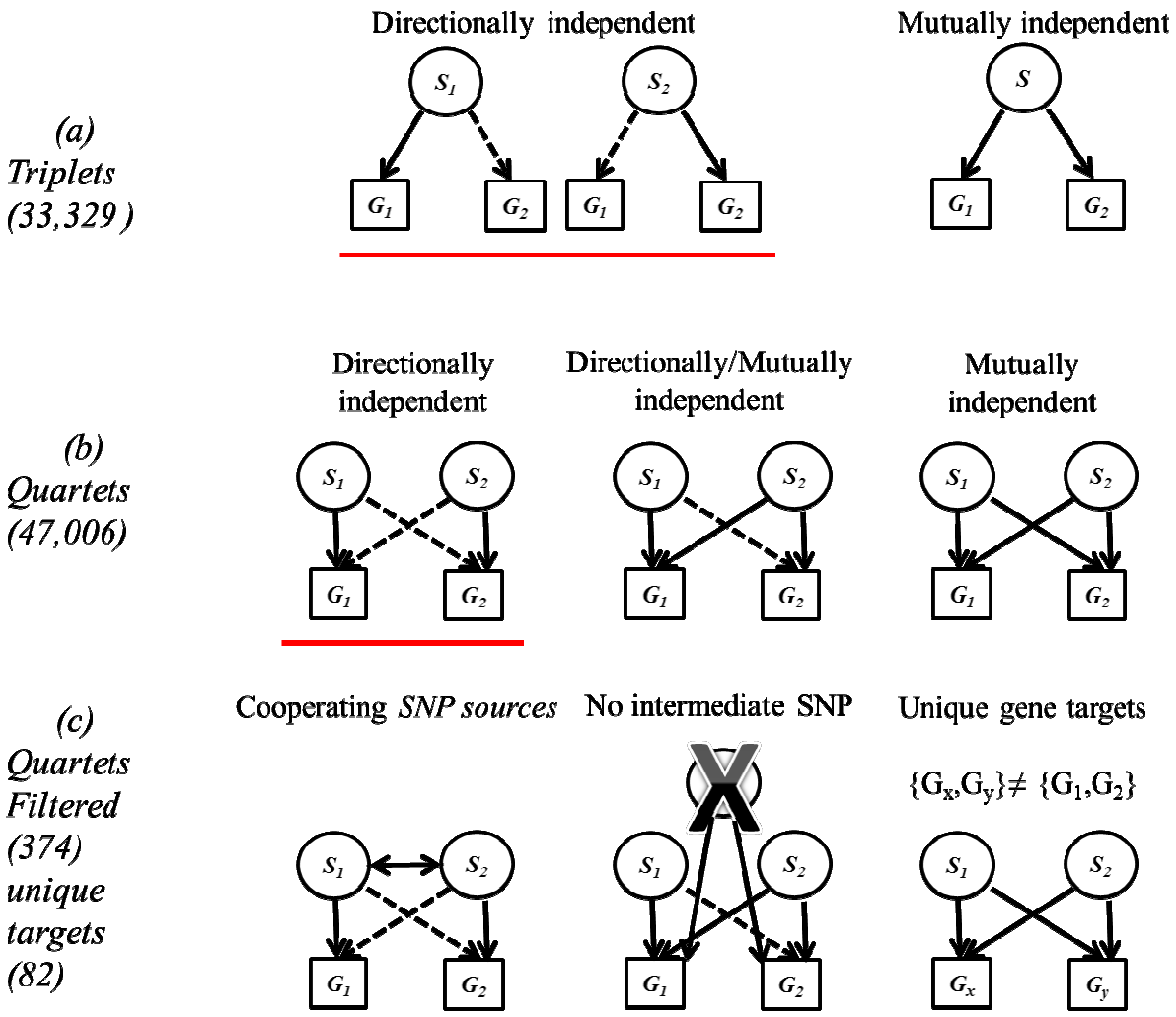


**Figure 3-2. A diagram explaining the framework for creating and filtering quartets.**
(a) We include mutually independent and directionally dependent triplets. A solid line represents mutually independent association. A dashed line represents directionally independent association. (b) Quartets are assembled from triplets in (a) with the same associated gene targets. Quartets are assembled either from two directionally independent triplets (red underline), two mutually independent triplets or one directionally independent triplet and one mutually independent triplet. (c) We filter the quartets using three criteria: (1) Restricting our analysis to quartets with cooperating eSNPs sources, i.e., SNPs that carry independent information towards predicting the expression of each one of the two genes. (2) Removing quartets where a third intermediate SNP can explain the expression to either transcript better. (3) Focus on quartets that have two unique gene targets, i.e., after filtering, no quartet has the same pair of gene targets.

Evidence for the validity of quartets

We choose an association testing threshold of $10^{-4}$ (Figure 3-3) by the number of quartets produced, aiming at FDR < 5% when comparing to the number of quartets in permutations. We examined the number of triplets in real data vs. 100 permuted data sets where sample labels had been switched. In permuted data sets, an average number of 33,329 triplets exceeded association p-value threshold of $10^{-4}$ (Figure 3-4). We therefore considered a comparable set of triplets, the same number of top results in real data, which corresponded to an association p-value threshold of $10^{-4.52}$ (Figure 3-4). This step creates an equal starting point for permuted vs. real datasets when approaching further analysis. We next examined the number of quartets formed by such triplets in real vs. permuted datasets. We observe that the number of 47,006 quartets in real data is consistent with chance expectations (empirical p-value = 0.07, Figure 3-5). Out of 47,006 quartets, there are 4,009 quartets with unique gene targets.

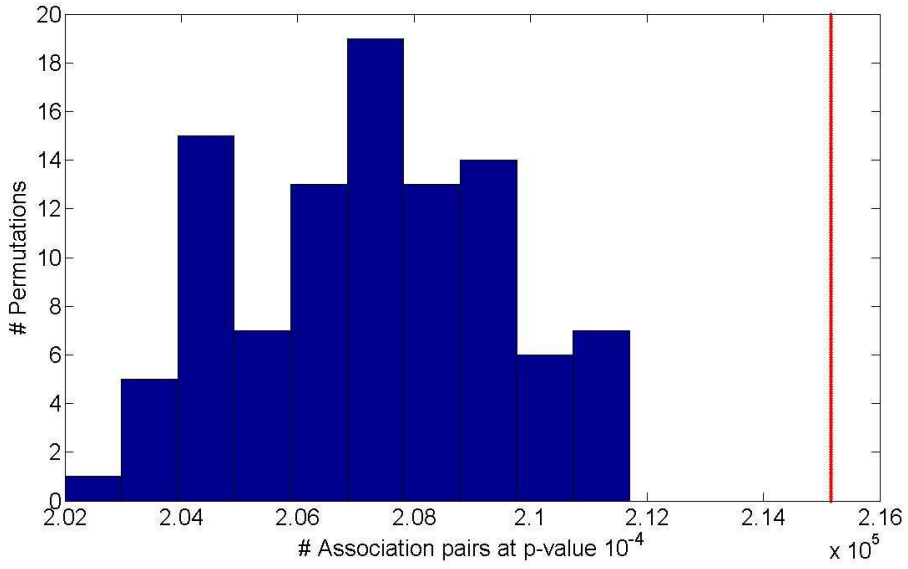


**Figure 3-3. Histogram of the number of association pairs in 100 permutations for a p-value cutoff $10^{-4}$.**
The red line indicates this number in the real data.

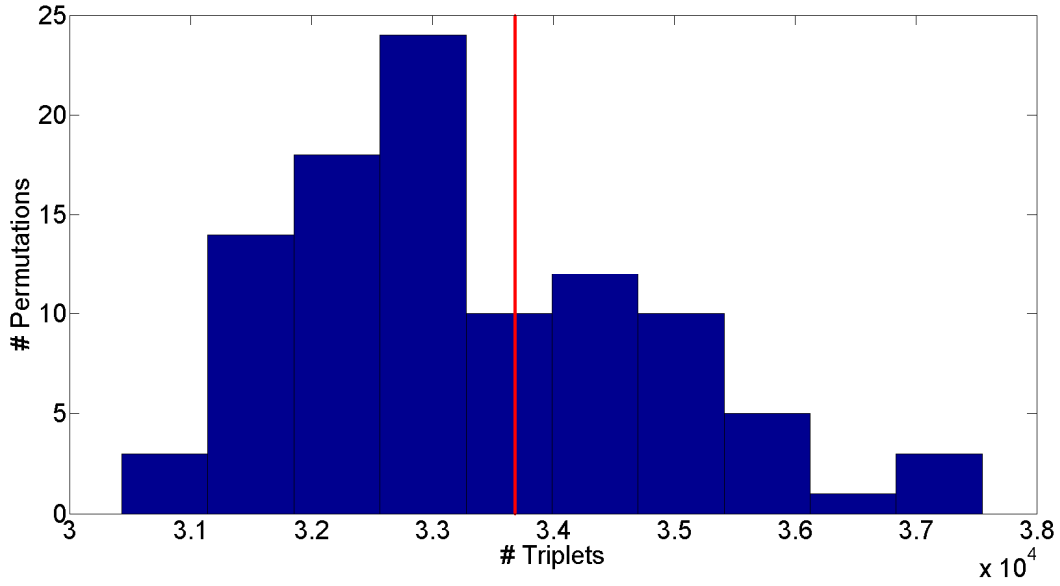


**Figure 3-4. Histogram of the number of triplets in 100 permutations, at association p-value of $10^{-4}$.** The red line indicates the observed number of triplets in real data at association p-value $10^{-4.52}$.

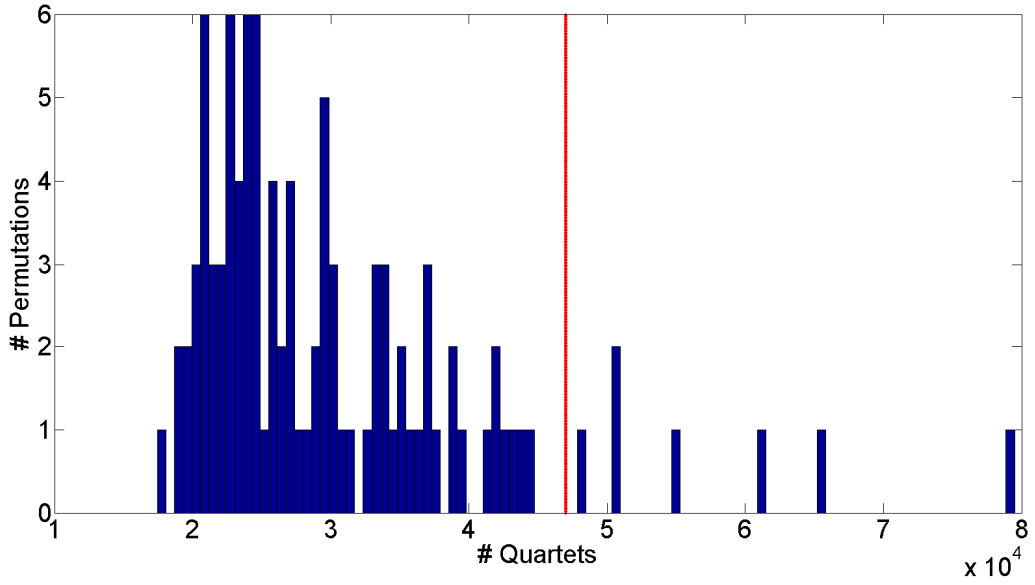


**Figure 3-5. Histogram of the number of quartets in 100 permutations, at association p-value of $10^{-4}$.** The red line indicates the observed number of quartets in real data at association p-value $10^{-4.52}$.

Interestingly, when examining cooperating quartets, we observe 374 such quartets in real data (0.8%) comparing to a mean of 19.18 in permutations (0.063% out of a mean of 30,250 quartets) (Figure 3-6). These results establish that the regulatory structure of cooperating quartets is nearly exclusive to real data, as it is rarely emerges in permutations. Out of 374 cooperating quartets with cooperating eSNP sources we focus on the 82 quartets that have two unique gene targets (Table 3-1). These include 2.05% of the total of 4,009 quartets with unique gene targets. Such unique cooperating quartets are more common in real data than in permuted data both in absolute number as well as in their relative fraction: permutations include only 3.71 such quartets on average (empirical FDR < 5% , Figure 3-7) 0.097% of an average of 3,819 quartets with unique gene targets.

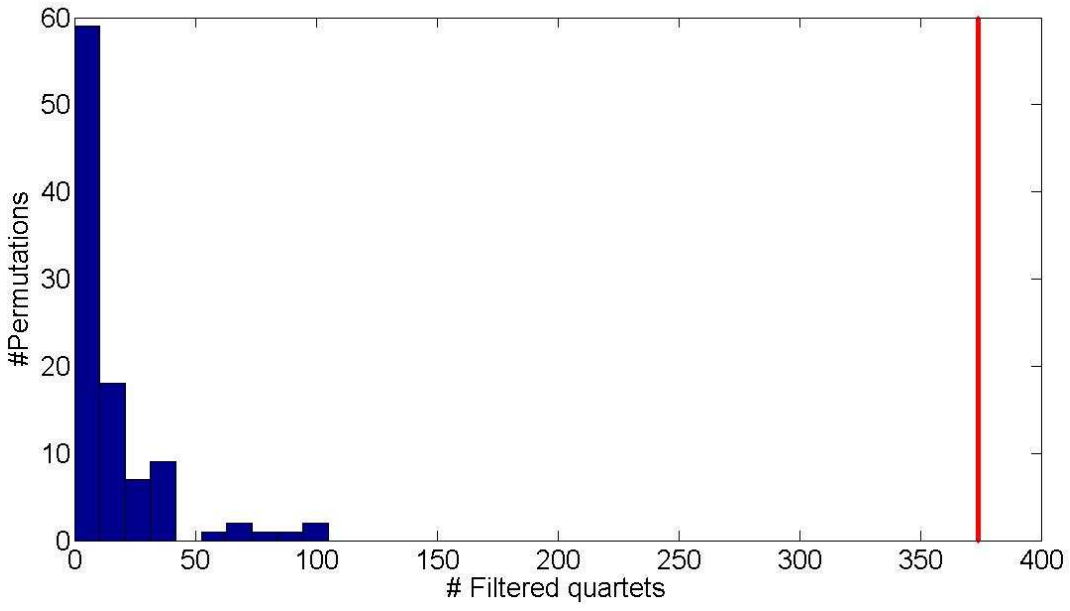


**Figure 3-6. Histogram of the number of filtered quartets in 100 permutations, at association p-value of $10^{-4}$.**
The red line indicates the observed number of filtered quartets in real data at association p-value $10^{-4.52}$.

**Table 3-1. A comprehensive description of 82 cooperating quartets (separate file).**
File Table3-1.xls 3000000000 correspond to different chromosomes.

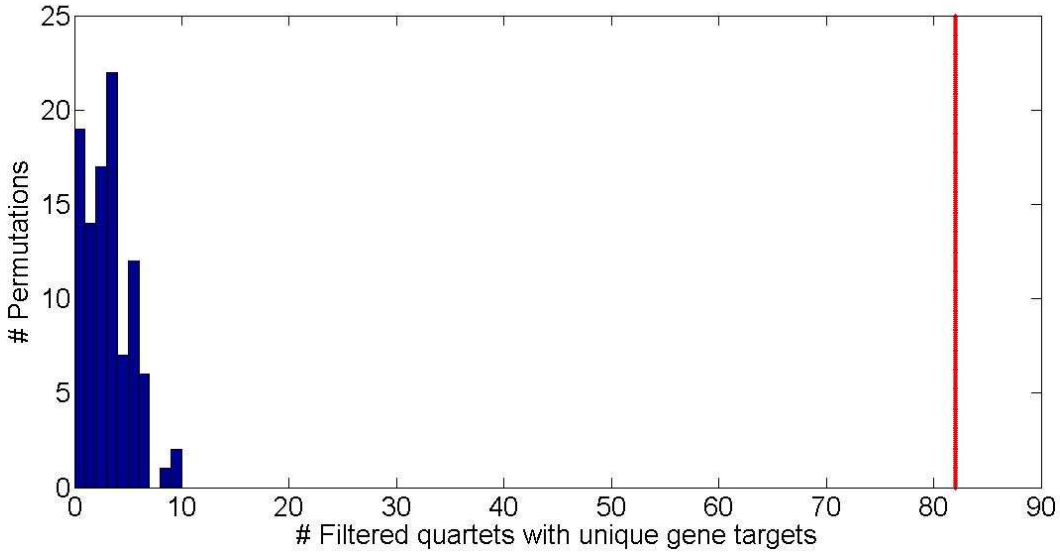


**Figure 3-7. Histogram of the number of filtered quartets with unique gene targets in 100 permutations.**
The red line indicates the observed number of filtered quartets with unique gene targets in real data (empirical FDR < 5%).

Cooperating quartets are a motif of the human regulatory network analogous to the bi-fan motif found in *e.coli* [37, 82]. We set out to characterize these cooperating quartets and study their functional, genomic and topological properties. In the next section we compare such quartet properties to permuted data, highlighting the quartets observed in real data as a true phenomenon, as opposed to quartets observed by chance.

Quartets in real data have distinct properties

Since the number of quartets in each permutation is low (Figure 3-7), we combine all quartets across all permutations and treat them as a "permuted set" of 342 quartets. From this point we compare the 82 quartets in real data vs. those in the permuted set to uncover properties that are unique to real structures.

### *3.2.2 Distribution of genomic properties of eSNP sources and their gene targets*

We first record genomic annotation categories of eSNP sources (Figure 3-8). eSNP sources tend to be one in exon and one in TF (Figure 3-8a upper panel; Fisher's exact $p < 1.9\times10^{-8}$ compared to the permuted set, see Figure 3-8a lower panel), or both in exons (Fisher's exact $p < 0.013$ compared to permuted set). We notice that most eSNP sources are located on different chromosomes (74% Figure 3-8b upper panel). For comparison, there are only 3.8% of eSNP sources on different chromosomes in the permuted set (13 out of 342; Figure 3-8b lower panel). An eSNP is said to be in *cis* of a target if it resides within the span of the target, and in *trans* otherwise. We characterize the *cis/trans* regulation of the four pairs of eSNP sources and their gene targets in each quartet by binning quartet data into three *cis*/*trans* categories: (1) two *cis* relationships (2) one *cis* relationship (3) two *trans* relationships. We notice that only a fraction of quartets involves *cis* regulation (Figure 3-8c upper panel), compared to none in the permuted set (Figure 3-8c lower panel). The target genes are located mostly (83%) on different chromosomes which is consistent with empirical expectation based on permutation. They are observed to be co-expressed significantly ($P < 4\times10^{-11}$) more often than in real data when comparing the absolute value of the correlation coefficient.

These results highlight unique properties of cooperating eSNPs and their distances from target transcripts. Specifically, we show that pairs of eSNP sources are located on different chromosomes.

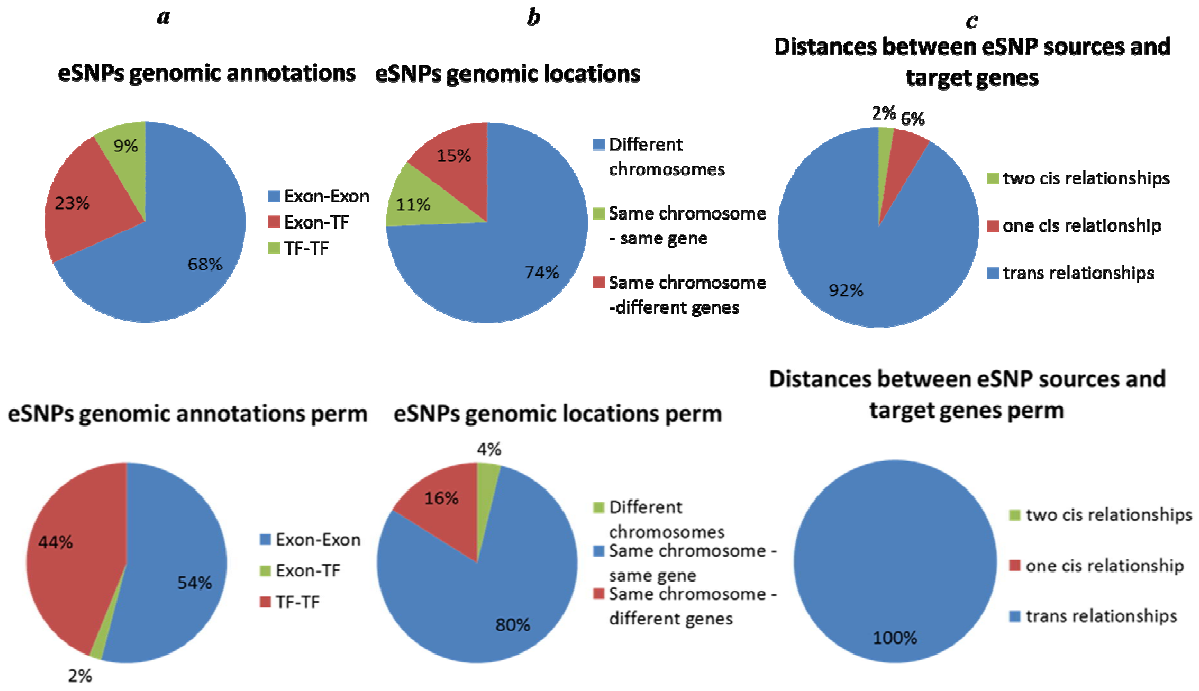


**Figure 3-8. Distribution of genomic properties of eSNP sources.**
Upper panel: real data. Lower panel: permuted data. By (a) genomic annotation (b) relative genomic location (c) distances between them and their targets. An eSNP is said to be *in cis* if it resides within the span of the target gene and *in trans* otherwise.

### *3.2.3 Characterizing dependencies within cooperating quartets*

We examine the dependency across association signals for each quartet source, i.e., whether the effect is mutually independent or directionally dependent. Dependencies within a quartet are therefore either (1) pair of mutually independent associations (2) one directionally dependent association and one mutually independent association, or (3) a pair of directionally dependent associations. We observe that 82% (67 out of 82) of the quartets are composed of a pair of mutually independent associations (Figure 3-9a). This is significantly more than expected

according to the permuted set, that includes mostly quartets with a pair of directionally dependent associations (Fisher's exact $P < 2.3\times10^{-35}$, Figure 3-9b). These results suggest that the eSNP sources affect the expression levels of both transcripts in a mutually independent manner rather than through directional dependence.

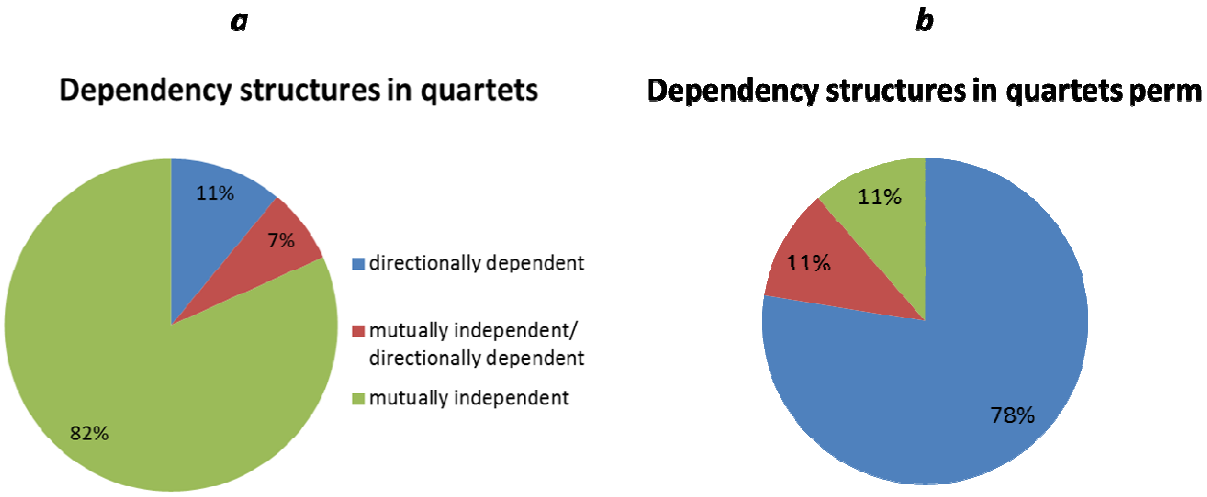


**Figure 3-9. Dependency structures in quartets.**
In (a) real data (b) permutations. Quartets are either comprised of a pair of mutually independent association signals, one directionally dependent association and one mutually independent association, or a pair of directionally dependent association signals.

### *3.2.4 Identifying direction of effect between eSNP sources and gene targets*

We were interested in examining the direction of SNP effects on gene expression. Within quartets we orient all SNP effects by using the convention of up (down) regulation to mean positive (negative) correlation between the number of copies of the minor SNP allele and the expression level of the associated gene. Out of the $2^4$=16 up/down configurations that are theoretically possible between two sources and two targets, we observe only eight configurations in real data – the ones with an even number of "up" effects (Figure 3-10). Consideration of the symmetry between the two sources, as well as the one between the two targets, highlights a sense

in which these eight categories involve *consistent* directions of effect, as we now explain. It is natural to classify the categories into four pairs, each defined by two binary criteria. The first criterion considers whether the two source SNPs have the same directions of effect on one gene as they do on the other or whether directions of effect on the second gene are opposite to the first one. The second criterion distinguishes whether the effect of one SNP on the two target genes is in the same direction as the effect the other SNP has on them, or whether directions of effect of the second SNP are opposite (Figure 3-10).

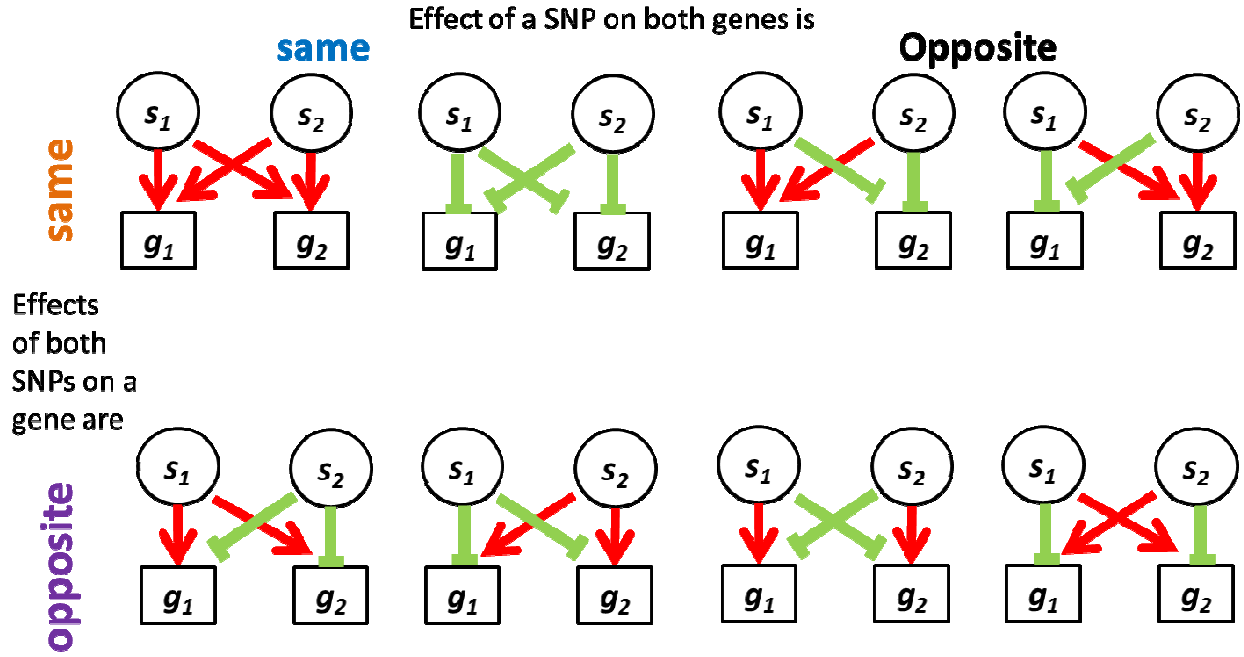


**Figure 3-10. Categories for direction of effect between eSNP sources and gene targets.**
The effect of a SNP on both genes can be the same (e.g., both genes upregulated) or opposite (i.e., one gene is upregulated and one downregulated). The effect of both SNPs on a gene can be the same (e.g., both downregulate the gene) or opposite (i.e., one SNP upregulate the gene and the other SNP downregulate it).

In contrast to the real data, where all quartets are consistent (Figure 3-11a), 30% (101 of 342) of quartets in the permuted set are *inconsistent* quartets (Figure 3-11b), meaning that the effects of

the two SNPs on one of the targets go in the same direction, while their effects on the other target are opposite (Figure 3-12).

We hypothesized that quartets in real data may be practically forced to be consistent due to correlation patterns across the expression levels of their targets. Specifically, a source SNP would the same (opposite) effect on both target genes due to their expression being correlated (anti-correlated). Indeed, we observe this pattern across all quartets in the real data but not always in the permuted set.

There are a couple of statistical challenges involved in comparison of real quartets to those observed in permutations (see Materials and Methods section 3.4.5). When these are addresses, specifically by analyzing eSNPs sources from the same quartet but from different chromosomes, we observe them to be enriched for same-direction effects compared to their permuted set counterparts (Figure 3-11c and 3-11d) and the gene targets to be located on different chromosomes. We listed all characterizing features of cooperating quartets (Table 3-1).

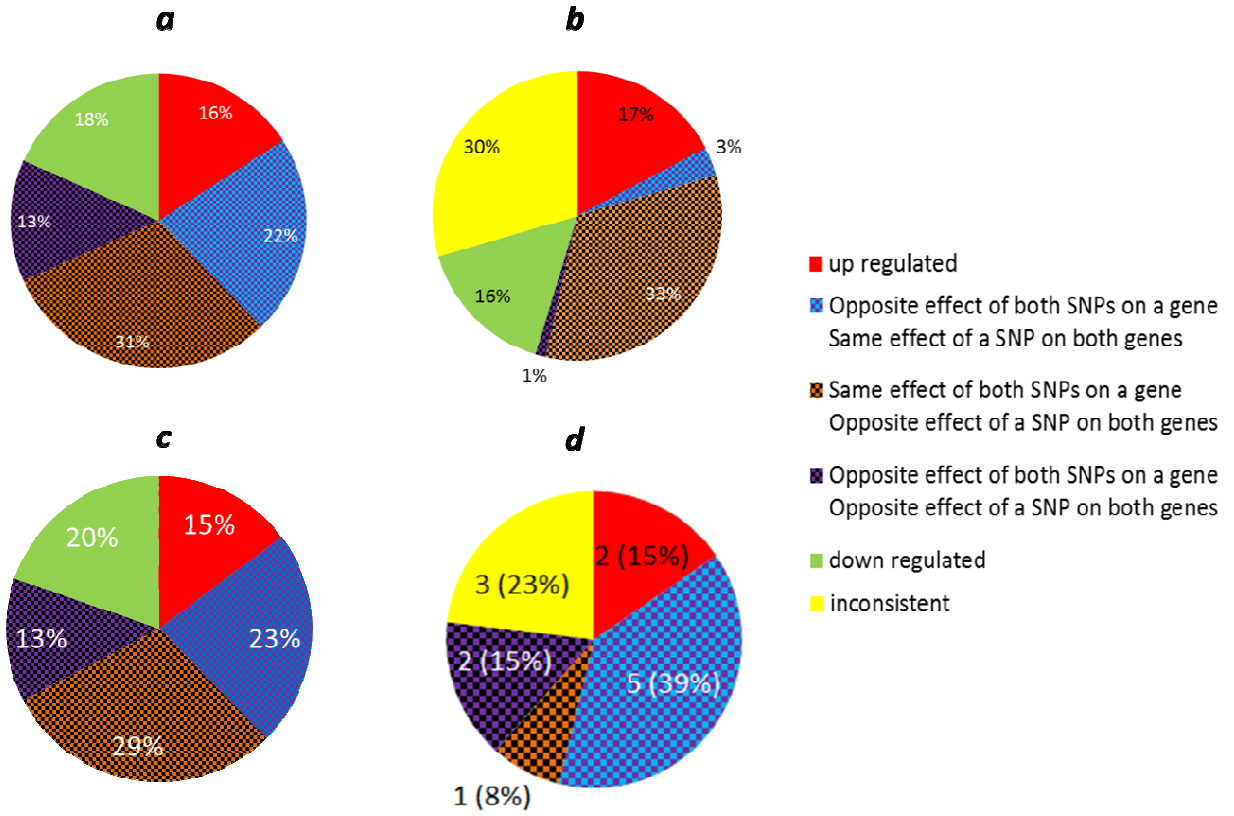


**Figure 3-11. Direction of effect for eSNP sources association with gene targets expression.**
In (a) real data (b) permutations (c) real data when the eSNP sources are located on different chromosomes (d) permutations when the eSNP sources are located on different chromosomes. Both SNPs can have either the same or opposite effect on gene targets. The effect of a SNP on both genes is either the same or opposite.

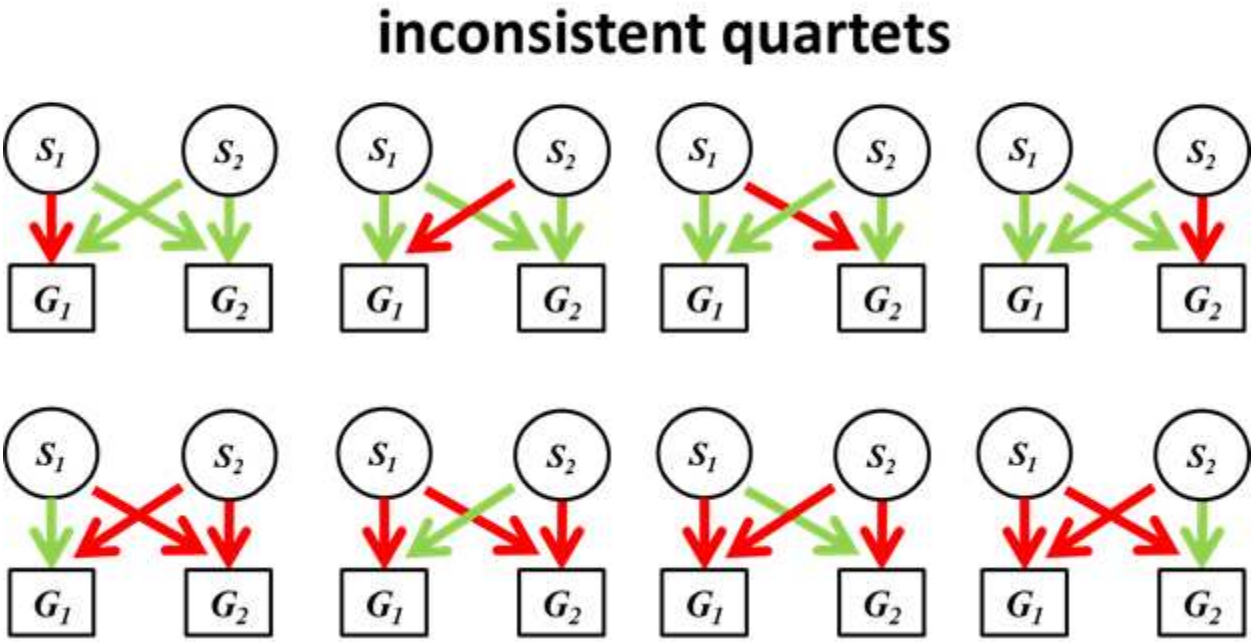


**Figure 3-12. All eight patterns of inconsistent quartets.**

### *3.2.4 HLA quartet*

A particularly illustrative sub-group of 7 quartets includes those with eSNP sources and gene targets along the MHC region of chromosome 6 (Table 3-1). This is significantly more (Fisher's exact $P < 0.0014$) than 4 out 342 (~1%) in the permuted set. The eSNP sources collapse to reference alleles of rs9274634, rs1129740, rs1142334, rs9274389 and rs2808143 and non-reference alleles of rs1130034, rs8227, rs1130116 and rs9272851 downregulating HLA-DQA1 and HLA-DQB1 and upregulating HLA-DQA2 and HLA-DQB2. These common variants are shared by specific assembled sequences and are associated with co-expression of DQA1-DQB1 and anti-correlated to DQA2-DQB2. All the genes containing eSNP sources and target genes are collapsed into the following four HLA genes: HLA-DQB1, HLA-DQA1, HLA-DQB2 and HLA-DQA2 (Figure 3-13). All four genes are involved with the MHC class II receptor activity

(enrichment FDR < $1.4 \cdot 10^{-12}$), and serve as an example how quartet structures create functional units.

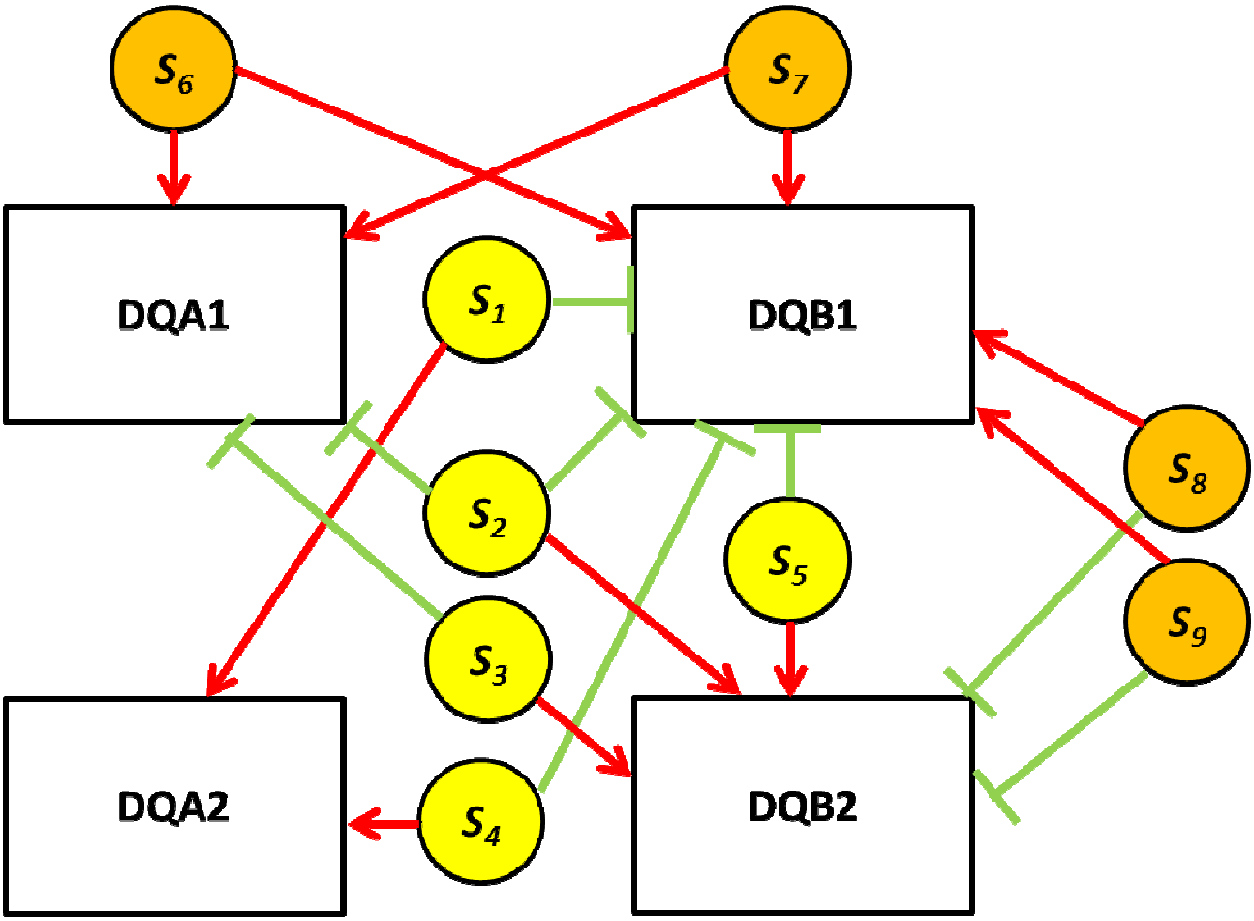


**Figure 3-13. HLA quartet.**
An example of examining eSNP sources and gene targets on the same chromosome together. Assembled quartets at the HLA locus highlight a 9-SNP haplotype associated with co-expression of DQA1-DQB1 and anti-correlated to DQA2-DQB2. $s_1$, $s_2$, $s_3$, $s_4$ and $s_5$ (yellow circles) correspond to rs2808143, rs1129740, rs9274634, rs9274389 and rs1142334 respectively. $s_6$, $s_7$, $s_8$ and $s_9$ (orange circles) correspond to rs8227, rs1130034, rs9272851 and rs1130116 respectively. Red edges indicate up-regulation, green edges indicate down-regulation.

### *3.2.5 Functional enrichment of quartets*

We perform a gene set enrichment analysis to examine if the pair of gene targets shares a GO category significantly more than pairs in the permuted set. In this case we observe a higher number of shared descriptors which is not significant in this dataset (Fisher exact p-value < 0.14). Interestingly, when we focus the enrichment analysis on pairs of genes that harbor cooperating SNP sources, we observe a significant difference (Fisher exact p-value < $1.5\times10^{-6}$). This supports our ability to detect SNPs that cooperate together to perform a joint function. We were intrigued to examine if our approach could be applied to understand gene regulatory networks underlying complex diseases. We therefore utilized the GWAS catalog [16] to find all genes that harbor a GWAS SNPs in our dataset. We then intersected this list with the genes that harbor cooperating SNPs in real data and compared to permutations. We observe a significant overlap of GWAS loci with at least one eSNP source, for quartets with sources that reside on different genes (Fisher exact p-value < 0.017). This indicates that our approach could shed light on regulatory circuits that are involved in complex disease. For example, in quartet #35 (Table 3-1) eSNP sources rs16877111 and rs7925000 are on chr5 and chr11 respectively. The eSNP sources reside in genes CMYA5 and RPL27A which are obesity GWAS loci. The gene targets HIST1H1D and HIST1H2AH are part of a histone cluster on chr 6.

### *3.2.6 Replication of quartet properties in a larger dataset*

Since our initial study was underpowered, we attempted to replicate the discovered properties of cooperating quartets in a larger, more recent dataset. We hypothesized that the fraction of true positives among signals of association to be higher is such a dataset, thereby pointing to true characteristics of quartets, rather than potential artifacts of false positive signals. We repeat our

analysis in the Geuvadis [80] dataset for each of its five populations: Utah European (CEU; n=91), Finnish (FIN, n=95), British (GBR; n=94), Italian (TSI; n=93) and Yoruban (YRI; n=89) as well as on the combined set of all European samples (n=373). We observe that the number of association signals achieving p-value $<10^{-5}$ is enriched in true positive associations (~5 fold more associations than expected). Overall, we replicate all properties (Same effect of both eSNPs, distal regulation, eSNP sources on different chromosomes, gene targets on different chromosomes and consistency of quartets) that were found in the smaller dataset, most of them at higher frequencies (Table 3-2). This provides an additional support from an independent dataset to the validity of quartets and their characteristics.

| Pop | #associations $10^{-5}$ | #expected associations at $10^{-5}$ | #filtered quartets | Same effect of both eSNPs (%64) | Distal regulation (92%) | $S_1$ –$S_2$ diff chr (75%) | $G_1$-$G_2$ diff chr (83%) | Consistency (100%) |
|---|---|---|---|---|---|---|---|---|
| EUR | 50048 | 10287 | 21674 | 82% | 78% | 88% | 89% | 99.3% |
| CEU | 54232 | 10155 | 43341 | 99% | 88% | 77% | 92% | 99.9% |
| FIN | 43111 | 10334 | 16663 | 90% | 82% | 88% | 84% | 99.7% |
| GBR | 43396 | 10267 | 18398 | 98% | 84% | 92% | 82% | 99.9% |
| TSI | 44562 | 10251 | 16171 | 93% | 86% | 93% | 90% | 100% |
| YRI | 94671 | 14698 | 51115 | 96% | 85% | 87% | 91% | 99.9% |

**Table 3-2: Replication of quartets' properties in the Geuvadis dataset [1].**
For each property in the first row we indicate the percentage in the original, smaller dataset.

## 3.3 Discussion

Discovering the building blocks of regulatory network has been an active field of research in the last decade [37, 92]. Specifically, the human regulatory network was the focus of a multiple recent studies involving diverse data types [81, 93]. In this work we devised a computational framework to study characteristics of cooperating quartets comprised of a pair of cooperating eSNP sources that reside either in exons or in the span of TFs, and a pair of associated target transcripts.

Our results establish that the regulatory structure of cooperating quartets is nearly exclusive to real data, and exhibits unique functional, genomic and topological characteristics. Cooperating quartets reported here in a human system might correspond to bi-fans, a known network motif of four nodes, previously described in model organisms [37].

Most cooperating quartets involve pairs of eSNP sources located on different chromosomes, away from their targets, which are themselves mostly located on different chromosomes. These quartets typically comprise of a pair of mutually independent association signals. All quartets are consistent in terms of the direction of eSNP effects on correlated and anti-correlated transcripts. We identify a separate sub-group of quartets with eSNP sources and gene targets all involving 4 MCH Class II genes from chromosome 6, highlighting a functional unit built from the quartet motif.

This study holds the promise for extension beyond its current limitations. First, our focus on causal variants localized to the single-base resolution imposed relying on a dataset of fully sequenced individuals along with their transcription profiles. Such cohort sizes are limited in size, reducing the power to detect association and allowing us to observe only the strongest

effects. Potential increase in sample size for eQTL data would enable detection of eSNP associations and regulatory motifs at greater significance and confidence. Second, the current analysis focuses on discovering a network motif where pairs of transcripts are co-regulated by a pair of variants. Mining the data for additional motifs can elucidate other structures in the human regulatory network. Overall, both the raw datasets [40, 91] and supporting databases [47, 50, 54, 62] in this work were noisy and limited. As functional annotation continues to build up, better understanding of motifs would be facilitated.

In this and in our previous work [36] we define network motifs showing them to be prevalent in real data, explaining the organization of *trans* regulation. Comparison of such structures between healthy and affected samples and across different tissues is likely to improve understanding of disease and developmental regulatory processes. Future studies could expand this approach to focus on complex disease circuits by using this framework on a dataset that is focused on GWAS SNPs and find quartets where the eSNP sources are also known GWAS loci.

The vast majority of eQTL studies involve analyses that are based on considering a single SNP associated with a single transcript, primarily *in cis* [1, 6, 39, 40]. While these analyses capture only a fraction of genetic contribution to changes in the regulatory landscape, the advantage is high statistical power for detecting associations. A complementary effort focuses on building networks from eSNP data [4, 9, 34, 35]. While these studies provide much more comprehensive models, they lack the same strength of statistical assurance in their findings. The main advantage of our approach is that it provides a unique framework for analyzing eSNP data by bridging these two approaches, establishing statistical guarantees on our inferred results using permutations. Applying such analysis to different datasets can shed light on the architecture of the human regulatory network and the role genetics plays in shaping it.

# 3.4 Materials and Methods

### *3.4.1 Data details and processing*

We analyze a cohort of 50 Yoruban samples, for which genotypes of SNVs that are fully ascertained from sequencing data [91] along with RNA-seq data[40] are publicly available. Briefly, the raw dataset consists of 10,553,953 genotyped SNVs and expression measurements (quantile-quantile normalized values) of 18,147 genes with Ensembl gene ID across these 50 samples. Standard filters have been applied to the genetic data: Minor allele frequency > 0.05, SNP missingness rate < 0.1 and individual missingness rate < 0.1 [74]. After filtering, data for analysis consists of 50 samples with 7,206,056 SNPs. The Geuvadis [80] dataset that we use for replication consists of five populations: Utah European (CEU; n=91), Finnish (FIN, n=95), British (GBR; n=94), Italian (TSI; n=93) and Yoruban (YRI; n=89) as well as on the combined set of all European samples (n=373). After filtering all SNPs with Minor allele frequency < 0.05 and focusing only on SNPs in exons and TFs, there are 42,810, 43,561, 43,279, 43,214, 61,960 and 43,365 for CEU, FIN, GBR, TSI, YRI and EUR respectively.

### *3.4.2 Association testing*

For association analysis, we consider only SNPs that reside within candidate regulatory regions along the genome. In Kreimer et al. [38] we detect enrichment in *trans* association signals for eSNPs in exons and in TFs in this dataset. For TFs, the number of multiple associated transcripts is significantly higher for TFs in the real dataset than in permuted data sets. For exons, there is an excess of the number of eSNPs within exons indicating true positive results. We test for association between a SNP and every gene; we consider SNPs within the span of known exons

and TFs (including introns) [94]. We test for association using linear regression performed by the --assoc command in PLINK [74] .

### *3.4.3 Obtaining a random distribution of association test-statistics*

Examining the random distribution of association tests is helpful in evaluating the empirical significance of results. This is achieved by generating 100 permutations that shuffle the sample IDs. This allows repeating the analysis of genotypes vs. expression on permuted data while maintaining the correlation structure among the genotype profiles and among the expression profiles, separately.

### *3.4.4 Creating and filtering quartets*

We assemble quartets from directionally and mutually independent triplets that consist of a SNP and two associated genes. A mutually independent triplet is when both of the association pairs remain nominally significant given the respective other gene and a directionally independent triplet is where only one of the association pairs remain nominally significant given the other gene. Two triplets that share the same associated genes define a quartet. We then filter these quartets further using the following rules:

1. We are only interested in quartets where both SNPs carry significant information in predicting the expression of gene 1 and gene 2. i.e. $\alpha_1, \alpha_{2,} \beta_1, \beta_2$ should be significantly different than zero.

   $g_1$ – represents the expression of gene 1.

   $g_2$ – represents the expression of gene 2.

   $s_1$ – represents the minor allele count SNP 1.

   $s_2$ – represents the minor allele count SNP 2.

$$g_1 = \alpha_0 + \alpha_1 \cdot s_1 + \alpha_2 \cdot s_2 + \varepsilon_1$$

$$g_2 = \beta_0 + \beta_1 \cdot s_1 + \beta_2 \cdot s_2 + \varepsilon_2$$

2. Moreover, we are interested in examining quartets that have no intermediate third SNP ($s_3$) that can explain the expression better.

   The third intermediate SNP should satisfy the following:

   1. On the same chr
   2. $r^2(s_1, s_3) \geq 0.5$ and $r^2(s_2, s_3) \geq 0.5$
   3. $s_3$ should be in a triplet with the two genes.
   4.

$$g_1 = \alpha_0 + \alpha_1 \cdot s_1 + \beta_1 \cdot s_2 + \gamma_1 \cdot s_3 + \varepsilon_1$$

$$g_2 = \beta_0 + \alpha_2 \cdot s_1 + \beta_2 \cdot s_2 + \gamma_2 \cdot s_3 + \varepsilon_2$$

$$\gamma_1, \gamma_2 \neq 0$$

***3.4.5 Statistical challenges in comparing real vs. permuted quartets.***

There are a couple of statistical challenges involved in comparison of real quartets to those observed in permutations. One bias is that of proximal eSNP sources in permutations. This leads for example to the artifact of enrichment of opposite direction eSNP sources in real data, comparing to the proximal, hence correlated effect eSNPs in permutations (Figures 3-11a and 3-11b). A second challenge is due to the rarity of eSNP sources on different chromosomes in permutations. This makes it statistically hard for comparing characteristics of sub-groups between real and permuted data (Figures 3-11d).

# Chapter 4: Variants in exons and in transcription factors affect gene expression *in trans*

*Summary:* In recent years many genetic variants (eSNPs) have been reported as associated with expression of transcripts in *trans*. However, the causal variants and regulatory mechanisms through which they act remain mostly unknown. In this paper we follow two kinds of usual suspects: SNPs that alter coding regions or transcription factors, identifiable by sequencing data with transcriptional profiles in the same cohort. We show these interpretable genomic regions are enriched for eSNP association signals, thereby naturally defining source-target gene pairs. We map these pairs onto a protein-protein interaction (PPI) network and study their topological properties.

For exonic eSNP sources, we report source-target proximity and high target degree within the PPI network. These pairs are more likely to be co-expressed and the eSNPs tend to have a *cis* effect, modulating the expression of the source gene. In contrast, transcription factor source-target pairs are not observed to have such properties, but instead a transcription factor source tends to assemble into units of defined functional roles along with its gene targets, and to share with them the same functional cluster of the PPI network.

Our results suggest two modes of *trans* regulation: transcription factor variation frequently acts via a modular regulation mechanism, with multiple targets that share a function with the transcription factor source. Notwithstanding, exon variation often acts by a local *cis* effect, delineating shorter paths of interacting proteins across functional clusters of the PPI network [38].

# 4.1 Introduction

Creating the complete human regulatory map is an active field of study. Many previous studies have used genomic analyses of gene expression, binding motifs, epigenetic marks and other local features to infer regulatory interactions [73, 95-98]. In recent years it has been established that genetic variation can contribute an additional angle to this investigation [45, 57, 78, 79]. Formally, transcription level is considered as a quantitative trait that is altered by allelic variation with thousands of single nucleotide polymorphism (SNPs) reported as associated with changes in gene expression [36, 45, 71, 80]. Such markers, called expression SNPs (eSNPs) are further found to contribute to variation of disease phenotypes and other clinically relevant traits [17, 36, 48].

Variation in genomic DNA can affect transcription in multiple ways. Most intuitively perhaps, level of transcripts *in cis* of an eSNP may be altered due to allelic variation in regulatory elements [90]. Alternatively, such levels may be auto-regulated by changes in protein structure that reflect variation of the sequence content of local transcripts. Therefore, *cis* eSNPs have been studied extensively. However, *cis* associations are limited in their ability to inform us regarding the network of regulatory interactions between one gene and another. This motivates more focused study of the effects of genetic variants on expression of distal transcripts (*trans* associations). Unfortunately, while *trans* eSNPs can identify downstream effects and previously un-annotated regulatory pathways, they are harder to statistically and biologically justify than *cis* eSNPs. From a statistical perspective, since *trans* eSNPs can be associated with any distal transcript, the multiple testing burden dramatically increases, thus only a small number of results is detected. From a biological perspective, more complex mechanisms are needed to explain

*trans* associations. An example of such a mechanism is an eSNP with local *cis* effect on a gene which codes for a transcription factor known to regulate other genes *in trans*. Indeed, across multiple eSNP studies [10, 57], even when statistically significant *trans* or *cis* eSNPs associations are detected aplenty, the regulatory mechanisms by which they alter gene expression remain mostly unknown.

A large fraction of SNPs identified by genome-wide association studies (GWAS) [45] have been reported to be associated with disease phenotypes [17] despite being neither coding, nor linked to coding SNPs *in cis*. Furthermore, since large-scale genetic studies have been predominantly based on SNP arrays, SNP alleles that are reported as associated, in studies of either disease [45] or gene expression [57], are often merely tags for causal variants, whose identity is challenging to track down. More generally, the multitude of phenotypes for eSNPs represents an opportunity for tackling the central question of causation in association.

Protein-protein interaction (PPI) networks capture various experimental data, such as from yeast two-hybrid systems [99], regarding the physical binding of proteins, and are often used to examine how these interactions are involved in a specific biological function. Recently, improved data on signal transduction and metabolic and molecular networks have contributed to the fidelity and accuracy of the reconstructed PPI networks. However, the data represented by these networks can sometimes be partial and noisy. PPI networks have been modeled as theoretical graphs and their topological properties extensively studied [100-102]. This provided insights pertaining to functional, structural and evolutionary characterization of these networks, primarily in model organisms. Genetic interactions in yeast were studied in the context of protein

complexes network [103], motivating the investigation of genetic variants that alter gene expression (as interactions) with respect to the human PPI network[26]. Studies of PPI networks in the context of genetic variation have thus far focused on GWAS-detected SNPs that are associated with common traits and disease, reporting that genes that harbor such SNPs frequently code for interacting proteins [24, 26, 104-106] .Yet, such studies only considered the PPI-network nodes that correspond to the associated SNP, without a PPI network node that would correspond to the phenotype.

Here, we perform a comprehensive study of *trans* genetic associations and their large-scale properties as manifested on a PPI network. We use SNPs from sequencing data [91] that are candidates to be causal based on their genomic location, and then project their association to gene expression on a PPI network. We hypothesized that genes involved in true eSNP associations have distinct PPI-network properties that differ significantly from spurious genes with candidate association signals. To address this hypothesis, we focus on *trans* association of eSNPs in exons and transcription factors (TFs), analyzing their properties as reflected on the PPI-network topology and annotations of the genes involved. Our focus on expression quantitative traits allows consideration of paths along the PPI network, whose links with genetic variation had previously only been studied with respect to SNPs, rather than the transcripts they modulate.

Our results suggest that a significant fraction of eSNPs in exons act *in trans* through mild effects *in cis*, with a regulation mechanism that is mediated by PPI paths that are shorter than expected by chance and tend to traverse across functional clusters of the PPI network. These paths

highlight zinc ion binding genes as a possible mechanism of transcript-eSNP feedback across the PPI network. In comparison to such coding eSNPs, we observe that TFs harboring eSNPs and their associated genes create units of genes that are functionally enriched for biological annotations. This suggests a different, modular regulatory mechanism for such TF eSNPs. Altogether, our analysis offers insights concerning a variety of mechanisms by which genetic variation at functional loci shapes the structure of human regulatory networks.

# 4.2 Results

### *4.2.1 Computational framework for mapping trans associations onto the PPI network*

We were interested in pinpointing directly associated variants rather than indirectly imputed ones. We thus used a publicly available dataset of 50 fully sequenced Yoruban samples [91] along with their transcription profiles from RNA-sequencing data [40], bearing in mind that such available cohorts are limited in size. Due to this small sample size, we have limited power in detecting association. Therefore, most candidate eSNPs can only be designated with various levels of uncertainty.

We were intrigued to examine *trans*-eSNPs interactions with respect to an independent space of interactions, that is, a PPI network. Therefore, we evaluated two categories of candidate eSNPs that reside within regions along the genome with known regulatory potential and can be mapped onto a PPI network, that is, exons and TFs (see Materials and Methods section 4.3.2). Examining the distribution of *P*-values across these two categories of candidate *trans*-eSNPs , we observed that candidate eSNPs within exons show evidence of including true positive eSNPs (Figure 4-1a), as been previously shown [2]. By contrast, eSNP candidates in TFs show association signal distributions consistent with random expectation (Figure 4-1b). We further examine if TF candidate eSNPs exhibit qualities that are different from random. We hypothesized that a single TF will be associated with multiple transcripts via eSNPs. To address this hypothesis, we created 1,000 permuted sets of pairs of TF and transcript (see Materials and Methods section 4.3.3). We observed that the number of multiple associated transcripts is significantly higher (Wilcoxon rank sum test $P < 0.05$) in the real dataset (973 out of 1,000 permuted sets, empirical *P*-value = 0.027). Following these two observations, we focused on eSNPs within exons as the first subject

of our investigation, and compared them to eSNPs within the span of transcription factor genes. We set out to characterize and compare these two modes of *trans* regulation.

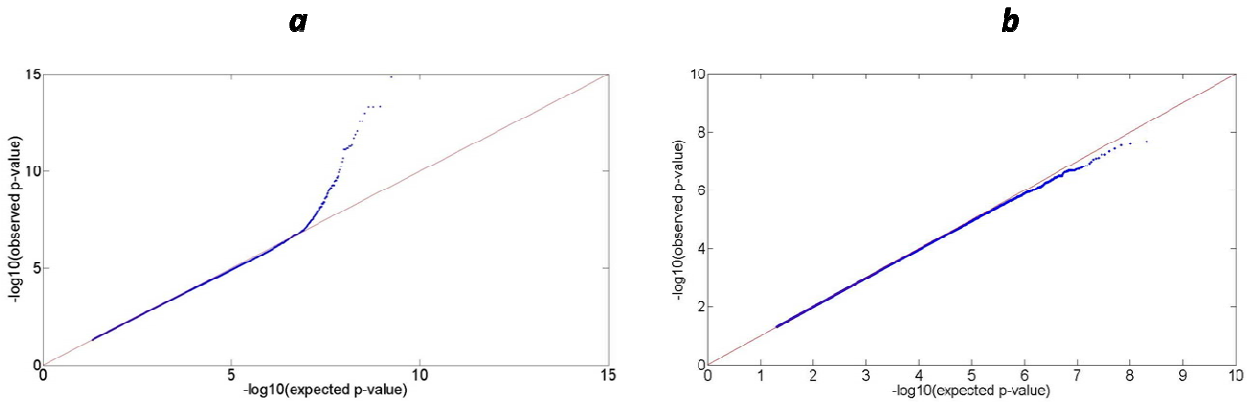


**Figure 4-1. QQ plot for association pairs of SNPs within known regulatory regions and genes.**
(a) eSNPs in exons and (b) eSNPs in TFs. X-axis denotes -log 10 of the expected p-value. Y-axis denotes -log10 of the observed p-value. The red line denotes expectation by chance (Y=X).

For each candidate eSNP that is associated with levels of a transcript in *trans*, we denoted this transcript as the 'target' of the eSNP. When this eSNP was located within an exon or in the span of a TF, we defined this gene as 'source'. We attempted to characterize eSNPs interactions on the molecular level by mapping these pairs of source-target genes onto a PPI network (Figure 4-2) and studied their functional annotations and topological properties.

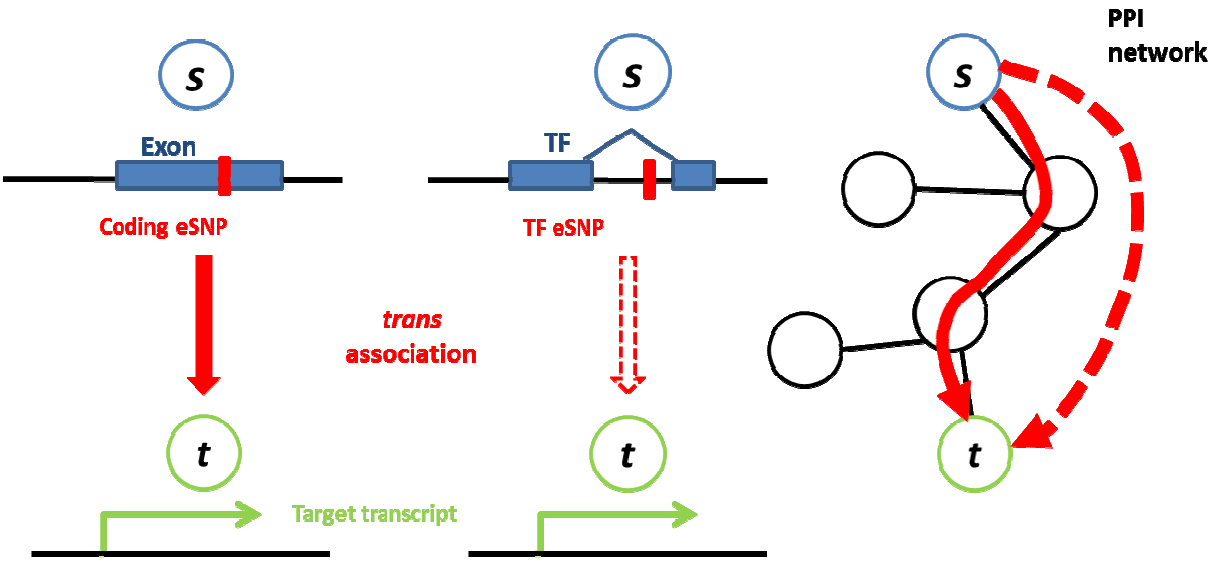


**Figure 4-2. *trans* associations on a protein-protein interaction network.**
*Trans* association marked by solid and dashed red straight arrows. An eSNP (red tick mark) that resides within a known exon (left) or TF (middle) maps to the PPI network (right). The source gene (blue s) is associated in trans with the levels of a target transcript (green t). PPI network edges are denoted in black, and define the shortest path between the exon source and its target (solid red curved arrow). The association between an eSNP within a TF source and its gene target is denoted by a dashed red curved arrow. eSNP, expression single nucleotide polymorphism; PPI, protein-protein interaction; TF transcription factor.

### *4.2.2 Identifying topological properties of exonic eSNP interactions*

We first considered pairs of exon eSNP source and target that demonstrated an association signal which was significant exome-wide for a particular transcript (association $P < 10^{-7}$). We observed such pairs to be significantly closer ($P = 0.03$) on the PPI network when compared with randomly permuted candidate eSNPs (see Materials and Methods section 4.3.4). Beyond pairwise properties of sources and targets, we further attempted to characterize each by their single-node features. Specifically, the targets of exon eSNPs had significantly higher ($P = 0.003$) degree than expected based on random pairs.

We reasoned that the cutoff of association *P*-value we used ($P < 10^{-7}$) was in many ways arbitrary, as we were interested in the statistical properties of the set of results rather than the significance of a particular result amid the testing burden. We therefore considered multiple *P*-value thresholds of eSNP association and at each threshold evaluated topological properties of eSNP source and target pairs, while assessing significance vis-à-vis randomly permuted sets of candidate eSNPs in exons (see Materials and Methods section 4.3.4). We observed that the lower the association *P*-values for source-target pairs, the more their topological properties differed compared with random pairs (Table 4-1). For example, for source-target pairs of exon eSNP, the average target degree among the 52 pairs exceeding an association *P*-value cutoff of $10^{-6.5}$ was 16.42, but it reached as much as 22.22 among the more focused set of 22 pairs that exceeded association *P*-value cutoff $10^{-6.8}$. These averages were each significant ($P = 0.02$ and 0.006, respectively) when compared with permuted pairs of exon eSNPs, whose target degree was only 9.36 on average. These trends are consistent with properties of true positives being diluted by false positives at less significant *P*-value thresholds. We quantified such trends by regressing each topological property on the negative log10 of the association *P*-value (Figure 4-3). We confirmed that for exonic source-target pairs, network distance decreased and the target degree increased with the significance of association (Spearman rank correlation coefficients r = -0.98 and 0.97, respectively; permutation *P*-value $P = 0.001$ and 0.002, respectively - see Materials and Methods section 4.3.4).

**Table 4-1. Topological properties and statistical differences of exonic eSNPs on the PPI network in real and permuted data (separate file).**

File Table4-1.xlsx. Exon source with their corresponding eSNP targets, for each P-value smaller than $10^{-6}$, where a source-target pair on the PPI network was added, we recorded the differences between topological properties of random and real pairs using Wilcoxon rank sum test. The table includes for each P-value the number of unique pairs on the PPI network, the rank sum test P-values and the mean value for each one of the topological properties (distance and source and target degrees) for real and random pairs.

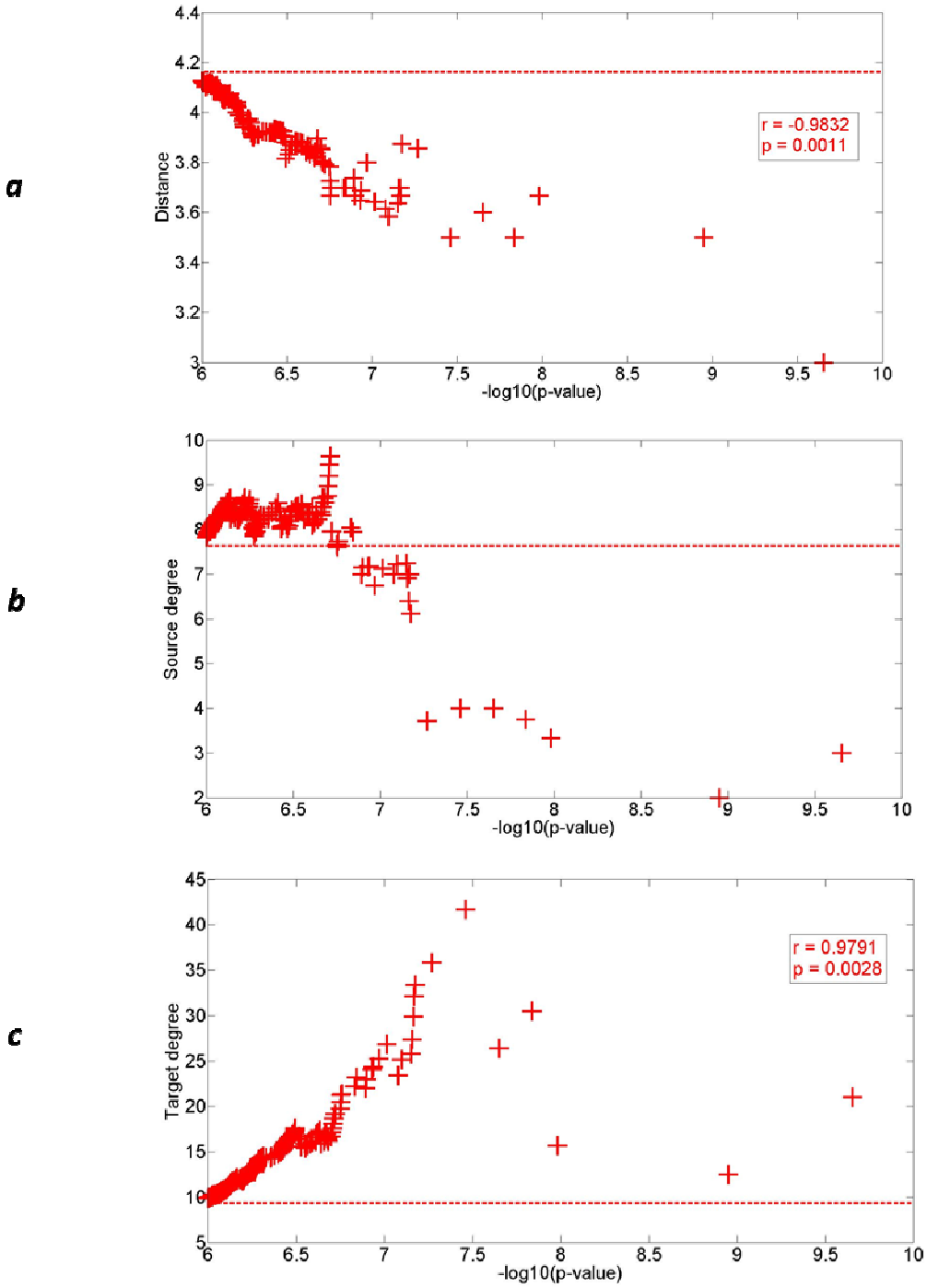


**Figure 4-3. Topological properties on a protein-protein interaction network versus exonic source-target association significance.**
Averages for (a) distance between source and target, (b) source degree and (c) target degree are evaluated across source-target pairs of candidate exon eSNPs at varying association p-value thresholds (+). The average of randomly permuted pairs (dashed horizontal line) is shown for permuted pairs and Spearman's rank correlation coefficient (denoted r) is listed when significant at P <0.05 (denoted p).

These results highlight unique properties of part of the transcripts whose *trans* regulation is due to coding variation. Specifically, we show that loci implicated by eSNPs encode for proteins that physically interact in a non-random fashion. Furthermore, target proteins are likely to interact with significantly more nodes of the PPI network than expected by chance.

### *4.2.3 Characterization of exon and transcription factor sources and targets*

Based on these results, for further analysis, we focused on the maximal *P*-value cutoff of $10^{-6.463}$, for which all topological properties showed significant difference between true source-target pairs of exon eSNPs and random ones (Wilcoxon rank sum test $P < 0.05$), (Figure 4-4 and Tables 4-2 and 4-1).

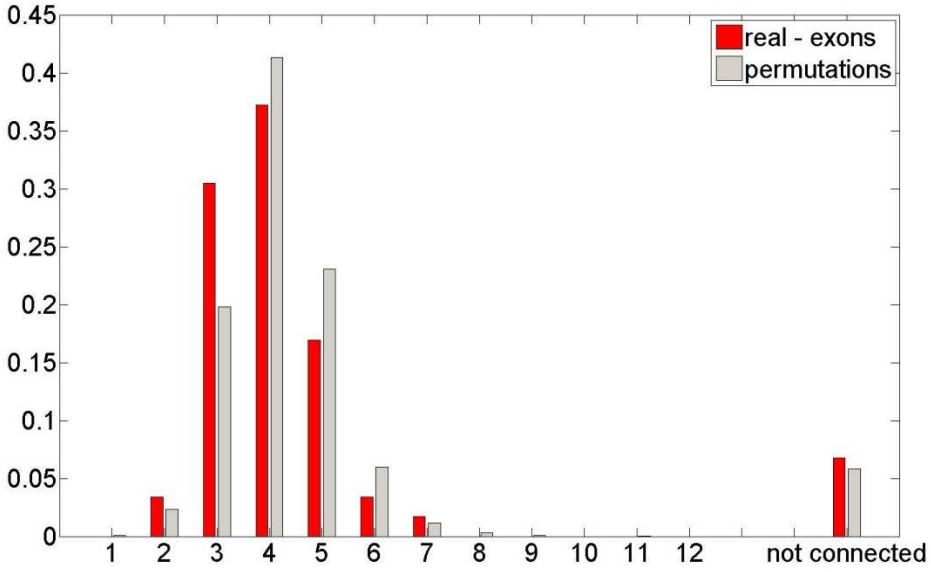


**Figure 4-4. Histogram in percentage for the distances between pairs of exon source and target.**
In real (red) and permuted (grey) data, for p-value=$10^{-6.463}$.

| ***Distance btw. pairs*** | ***Real – 59 pairs*** | ***Random – 18,675 pairs*** |
|---|---|---|
| 2 | 2 | 431 |
| 3 | 18 | 3,701 |
| 4 | 22 | 7,723 |
| 5 | 10 | 4,308 |
| 6 | 2 | 1,118 |
| 7 | 1 | 212 |
| 28 (not connected) | 4 | 1,092 |

**Table 4-2. Distances between real exon source and target and between random pairs.**

There were 343 pairs of source and target and 295 unique pairs, 59 of them on the network. Of these pairs, 318 (92.71%) were on different chromosomes and 25 (7.29%) were on the same chromosome, at least 1 Mb apart. At this cutoff there were 333 unique eSNPs in exons, 286 unique gene sources and 267 unique gene targets (Table 4-3). When comparing the effect sizes (absolute values of betas in the linear regression) of 929 previously published *cis* expression quantitative trait loci (eQTLs) [40] with the distribution of exonic and TF *trans* eSNPs effect sizes, we found that the *trans* effect sizes (mean 1.198) were significantly higher than those of corresponding *cis* effects (mean 0.964; Wilcoxon rank sum test *P*-value $<2.25 \times 10^{-49}$ and $3.56 \times 10^{-54}$ for exonic and TF eSNPs, respectively; Figure 4-5). We binned eSNPs and SNPs in exons by first, middle and last exons (Figure 4-6). We also examined the position of the eSNP along the transcript and compared these results to SNPs in exons (Figure 4-7). We observed that these *trans* exonic eSNPs tended to be located along middle exons, rather than in first or last exons (Fisher's exact test *P*-value <0.009). We further observed that they tended to lie farther away down the transcript (Wilcoxon rank sum test $P = 0.0058$). These results were different from what was observed for *cis* eQTLs. Montgomery *et al*. [39] reported that eQTLs with higher confidence were located in the first and last exons significantly more than in middle exons.

**Table 4-3. Genomic description of eSNPs in exons and TFs (separate file).**
File Table4-3.xlsx. For all TF and exonic source-target pairs we give the eSNP rs number, eSNP chromosome, eSNP location, source gene ID, target gene ID, target chromosome and association P-value. For eSNPs in TF, we indicate whether they are within an exon.

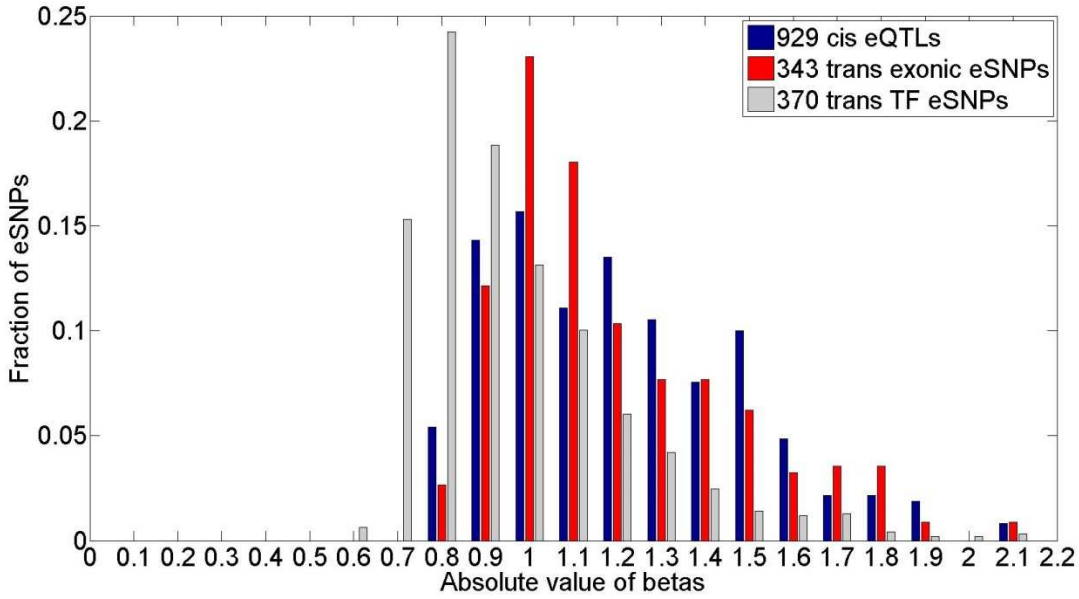


**Figure 4-5. Comparing effect sizes.**
(absolute value of betas) between previously published 929 cis eQTLs and 343 and 370 exonic and TF trans eSNPs respectively.

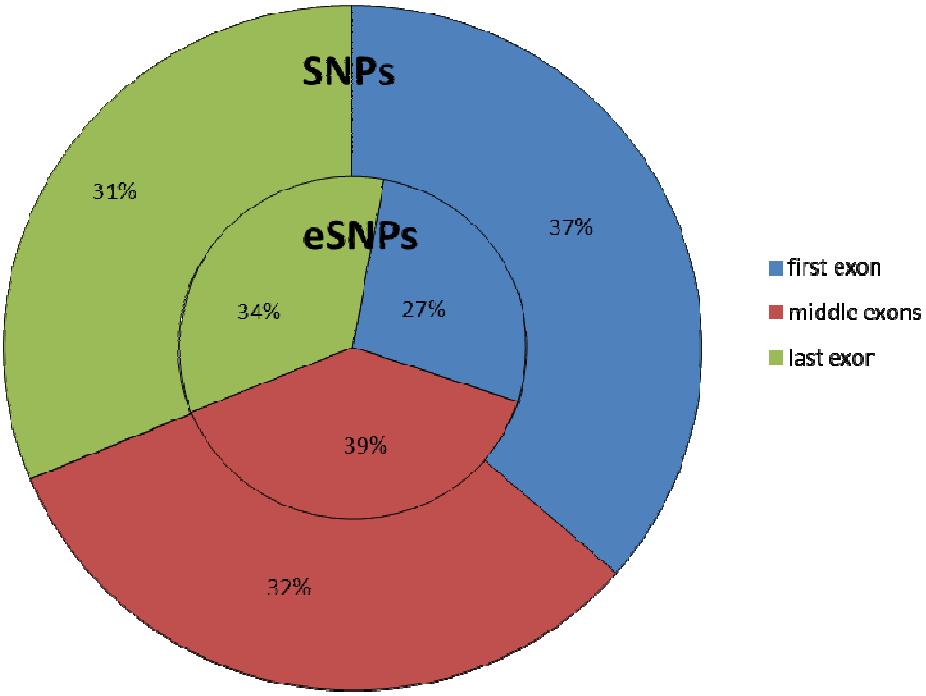


**Figure 4-6. Distribution of SNPs and trans eSNPs in exons.**

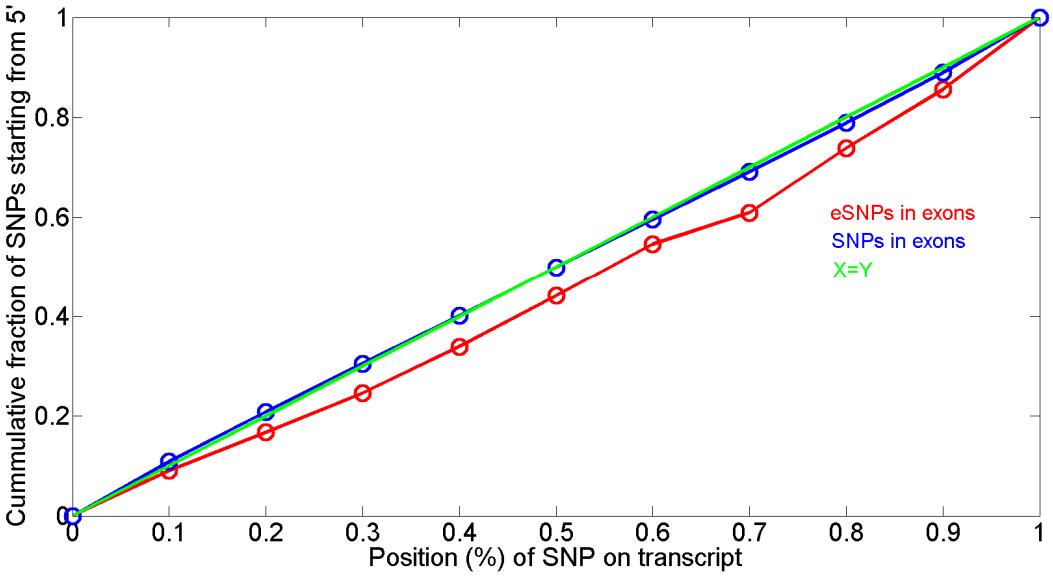


**Figure 4-7. Cumulative fraction of the position of exonic eSNPs (red) and SNPs (blue) on the transcript.**
Wilcoxon rank sum test p-value between the position of exonic eSNPs and SNPs on transcript < 0.0058

The combined set of exon sources was enriched for major histocompatibility complex protein genes (false discovery rate (FDR) <0.046) with concordance to findings in previous studies, indicating human leukocyte antigen SNPs were 10-fold enriched for *trans*-eSNPs [34]. We further observed that the set of target genes was enriched for multitude functional processes (see Table 4-4 for full list of annotations). The three highest scoring functional annotations of the target set, macromolecule modification, phosphatidylinositol-3,5-bisphosphate binding and protein modification process, provide additional support for the role of exonic eSNP targets as network hubs [107].

**Table 4-4. Functional enrichment analysis of combined sets of exon sources, exon targets and TF targets (separate file).**
File: Table4-4.xlsx. Gene sets include only genes that map to an Entrez ID.

For further investigation and comparison, we considered source-target pairs of TF candidate eSNPs, a set with similar order of magnitude, corresponding to association signals passing the *P*-value cutoff of $10^{-6}$. There were 370 such pairs of TF source-target, 193 of them unique, 58 of which were on the network. Of these pairs, 359 (97.03%) were on different chromosomes and 11 (2.97%) were on the same chromosome, at least 1 Mb apart. There were 358 unique eSNPs in TFs, 77 unique TF sources and 192 unique targets (Table 4-3). Out of the 358 unique eSNPs in TFs, 15 were in exons, significantly more than expected by chance (hypergeometric *P*-value $<1.8\times10^{-4}$). When we examined the combined set of TF targets, we observed that this gene set was enriched for various annotation categories (see Table 4-4 for full list of annotations).

### *4.2.4 Co-expression of targets and cis-effects on the source gene*

To further establish the association between the source and target genes, we examined the co-expression between eSNP source and target for all candidate pairs of associated genes in this dataset by evaluating Spearman's rank-correlation coefficient *r*. For pairs of exon-source eSNPs and their corresponding targets, the absolute value of *r* was significantly higher than expected from the entire distribution of co-expression measurements in this dataset (Wilcoxon rank sum test $P < 5.4 \times 10^{-5}$; Materials and Methods section 4.3.6). By contrast, for pairs of TF-source eSNPs and their corresponding targets, there was no significant difference in terms of co-expression. We observed the fraction of non-synonymous SNPs to be 0.082 out of exon eSNPs, which was higher than their overall fraction 0.071 among all exonic SNPs [108] (Fisher exact P approximately 0.1). For each eSNP we examined *cis* effects that were too mild to be detected at genome-wide significance threshold by testing for its association with the expression of its source gene (see Materials and Methods section 4.3.7). In total, 50 pairs of exonic eSNP and source gene were nominally ($P < 0.05$) *cis* associated, out of 286 such unique sources ($P = 3.6 \times 10^{-15}$). We estimated how many of the SNPs in exons have a *cis*-effect (linear regression *P*-value <0.05) on the expression of their host gene. We found that out of 97,135 exonic SNPs, 9,661 showed *cis*-effect on their host gene at the nominal significance level ($P < 0.05$). Compared to this background distribution, the observed 50 out of 286 *trans* eSNPs having such *cis*-effects is significantly more than expected by chance (Fisher's exact test *P*-value$< 9.6 \times 10^{-5}$). This provides additional support for the *cis*-effect phenomena. For comparison, we did not observe a nominally significant *cis* effect between TF eSNP and its source gene more than expected by chance (3 out of the 66 TF sources in this dataset). These results suggest a mechanism where

exonic variation often operates in *trans* eSNPs via alteration of gene expression in *cis*, and the source and target genes have correlated expression.

### *4.2.5 Modular organization of eSNPs in TFs*

TFs are known to control the transcription of multiple genes; we were therefore interested in whether we observed the same phenomena in TF variation. Each TF source forms, along with its targets, a set of genes that we called a unit. We observed that these units tended to be enriched for functional annotation categories. Specifically, for the 33 TF sources with two target genes or more (Tables 4-5 and 4-6), 26 out of 33 define units that are functionally enriched (two or more annotated genes, FDR <0.05; Materials and Methods section 4.3.8) [13] in KEGG [47] and GO [62] categories (Table 4-7). Interestingly, eSNP targets did not tend to share exon sources. Specifically, out of 286 unique sources, 278 had a single target, 7 (*AKNA*, *CDK7*, *BLK*, *ATP5G1*, *RPL8*, *TRAPPC12*, *MUC2*) of the remaining ones had two, and one (*HLA-C*) had three (Table 4-3). The difference between the number of associated targets in TF and exon variation was statistically significant (Wilcoxon rank sum test $P < 3.4\times10^{-4}$). These results support the hypothesis that TF variation frequently acts via a modular regulation mechanism, with multiple targets that share a function with the TF source.

| ***Unit size*** | ***Number of units*** |
|---|---|
| 2 | 35 |
| 3 | 16 |
| 4 | 8 |
| 5 | 2 |
| 6 | 2 |
| 8 | 1 |
| 10 | 2 |
| 11 | 1 |
| 17 | 1 |

**Table 4-5: Units size distribution of TF source and their gene targets.**
Uunits include only genes that map to an Entrez ID. We include the TF in the module.

| *Unit number* | *Unit size* | *TF source* | *Genes in the unit* |
|---|---|---|---|
| 1 | 3 | RUNX1 | CLN5, TCL1A |
| 2 | 3 | DMRT1 | EIF3H, GPATCH8 |
| 3 | 3 | GTF2F2 | GOLGB1, ATP2C1 |
| 4 | 3 | HSF2 | ABCC1, CCDC102A |
| 5 | 3 | NFIX | ORC2, CCDC91 |
| 6 | 3 | TCF4 | PNMT, PPHLN1 |
| 7 | 3 | TCF12 | 1-Dec, AZI2 |
| 8 | 3 | TFDP2 | SEMA6B, EXOC3L4 |
| 9 | 3 | MBTPS1 | SOCS2, PPAN |
| 10 | 3 | MTF2 | DNAH17, BTG3 |
| 11 | 3 | ATF7IP | UNC13A, RILP |
| 12 | 3 | PHTF2 | PPP1R16B, OTOGL |
| 13 | 3 | TFB2M | ZNHIT1, CAMKK1 |
| 14 | 3 | TCF7L1 | KCNQ4, SETDB2 |
| 15 | 3 | GABPB2 | PSMC6, ABHD5 |
| 16 | 3 | NFXL1 | HIST1H2BB, CHEK2 |
| 17 | 4 | ATF3 | NR3C1, SNORD26, INPP5E |
| 18 | 4 | NFYC | RBM5, CACNG4, STARD9 |
| 19 | 4 | GTF2A1L | CUL1, RNF24, LAMP3 |
| 20 | 4 | CNOT1 | PSMC2, PCGF6, THAP3 |
| 21 | 4 | WWTR1 | NFYB, FAM118B, FAM78B |
| 22 | 4 | TRERF1 | ADPRH, DNAJC5, NBPF23 |
| 23 | 4 | BACH2 | MYO7A, SNRPD1, SIRT7 |
| 24 | 4 | TFAP2D | PLA2G6, C8orf55, TMEM159 |
| 25 | 5 | BRF1 | WWC1, GRHL1, DDX54, DDX51 |
| 26 | 5 | AKNA | TCFL5, CCT5, SLC25A39, ALG8 |
| 27 | 6 | MITF | NEDD9, ARHGAP11A, TMEM51, MAGOHB, MIR589 |
| 28 | 6 | TCF7L2 | BOK, TLE4, NOP58, NAT10, TOR3A, |
| 29 | 8 | TEAD1 | ST3GAL3, DYNLT1, DBNL, GCNT4, PHF7, HNRNPA1L2, BTBD19 |

| | | | |
|---|---|---|---|
| 30 | 10 | STAT4 | ACO1, GNRHR, GYPC, PTRH2, MBOAT7, OBFC1, CORO6, UHMK1, PPTC7 |
| 31 | 10 | TCERG1L | SLC25A20, GATA2, ZNF3, LRPPRC, ABCA12, PCYOX1L, LBH, C16orf74, MIR1909 |
| 32 | 11 | MYT1L | APBB2, CDC25A, COL1A2, MMP7, SH3BP2, CWC27, NCAPH2, HNRPLL, ZMAT2, RPS26P6 |
| 33 | 17 | CAMTA1 | NFKBIE, QDPR, SKP1, CDK2AP1, TAOK2, GNB5, NECAP1, TMBIM4, PTRH2, VASH2, TMEM121, ZFP91, NHLRC2, H3F3C, C1orf190, SNORA81 |

**Table 4-6: TF units' content and sizes.**
TF source and gene targets (two or more).

**Table 4-7. TF units' functional enrichment (separate file).**
File Table4-7.xlsx. Gene sets include only genes that map to an Entrez ID.

### *4.2.6 Support for eSNPs in TFs from different data sources*

We systematically looked for pairs of TF source-target that were experimentally validated as binding. We found such enrichment, with 6 out of 34 TF source-target pairs compared to 551 out of 6,904 random pairs (Fisher's exact test $P$ <0.05, see Materials and Methods section 4.3.9) in a database reporting binding of TFs to DNA, based on chromatin immunoprecipitation (ChIP)-X experiments [109]. We used the data in [6] to find the closest DNaseI hypersensitive site (DHS) window to the gene target, and examined whether the TF eSNP was associated with the DHS levels in this window. We found that 33 of 370 such pairs of TF eSNP and gene target were significantly associated (P<0.05) indicating significant enrichment ($P < 5.5\times10^{-4}$) of this phenomenon. This enrichment was not an artifact of TF eSNP ascertainment: we tested the association of 29,212 TF SNPs to DHS levels in a randomly picked DHS window; as expected by chance, 1,400 of these SNPs showed such association at the nominal significance level, $P$ <0.05. Compared to this background distribution, the observed set of 33 out of 370 *trans* eSNPs having such association was significantly larger than expected by chance (Fisher's exact test *P*-

value < $6\times10^{-4}$). This shows that even in a small sample size where the number of true positives is diluted with false positives, we still recover a true signal.

### *4.2.7 Distribution of TF sources and targets in PPI functional clusters*

We were intrigued by potential connections between source-target pairs and cluster properties in the PPI network. Therefore, we partitioned the PPI network into clusters of genes, optimizing the modularity measure [110] (see Materials and Methods section 4.3.10). Out of the resulting 249 PPI clusters with two genes or more, 225 (90%) demonstrated functional enrichment for a biological category (Table 4-8). TF source-target pairs were found in the same PPI clusters more than expected by chance: 26 out of 58 TF pairs compared with 26,966 out of 100,000 random pairs (Fisher's exact test $P$ <0.0043; see Materials and Methods section 4.3.11).

**Table 4-8. Functional enrichment analysis of clusters in the PPI network (separate file).**
File Table4-8.xlsx. Gene sets include only genes that map to an Entrez ID.

### *4.2.8 Specific example of TF eSNP*

As an illustration for our results, we show an example (Figure 4-8a) of a specific source and its gene target, examining transcription factor 7-like 2; T-cell specific, HMG-box (*TCF7L2*) and its transcript target transducin-like enhancer of split 4 (*TLE4*). There was a significant *cis* effect ($P$ <0.012) of the associated intronic eSNP rs7087006 with the expression of *TCF7L2*, but the co-expression correlation of the source and target was not statistically significant in this dataset. *TCF7L2* and its five targets (unit number 28, Table 4-6) comprise a unit that was enriched (two out of six) for cell proliferation (FDR <0.03; Table 4-7). This TF plays a key role in the Wnt signaling pathway, activating v-myc avian myelocytomatosis viral oncogene homolog (*MYC*) expression in the presence of catenin (cadherin-associated protein), beta 1, 88kDa (*CTNNB1*).

The gene target *TLE4* within the PPI network is a transcriptional co-repressor that represses transactivation mediated by *TCF7L2* and *CTNNB1*. These annotations implicate that *TCF7L2, TLE4* and *MYC* act as the network motif incoherent type-1-feed-forward loop (a pulse generator and response accelerator) [92] where the two arms of the feed-forward loop act in opposition: *TCF7L2* activates *MYC* (in the presence of *CTNNB1*) but also represses *MYC* by activating the repressor *TLE4* (via an eSNP). We note that *TCF7L2* harbors the common allele most strongly associated with increased risk of type 2 diabetes. Correspondingly, *TLE4* was recently discovered as a T2D locus [81]. Specifically, *TLE4* encodes a protein that forms complexes with *TCF* proteins, including *TCF7L2*, to modulate transcription at target sites [111]. The source and target are part of the same PPI network cluster, which is enriched (1,257 out of 4,627) for regulation of transcription (FDR $<2.4 \times 10^{-88}$, Table 4-8; Figure 4-8a). This demonstrates a case of shared function between a source TF and its target.

### *4.2.9 Distribution of exonic sources and targets in PPI functional clusters*

By contrast, only 19 (32%) of exon eSNP sources were found in the same PPI network cluster as their respective single targets, consistent with chance expectation (see Materials and Methods section 4.3.11). Yet, as such pairs were linked by relatively shorter paths (Figure 4-3a), it follows that coding variants affect transcription *in trans* not in a modular way but rather in a linear fashion that defines shorter paths than expected by chance. We recorded the proteins along such paths (Table 4-9) and evaluated the enrichment of functional annotation for each path (Table 4-10).

| ***Path number*** | ***Path length*** | ***Genes in path (from source to target)*** |
|---|---|---|
| 1 | 3 | HLA-C, LILRB1, HLA-A |
| 2 | 3 | HLA-C, LILRB1, HLA-G |
| 3 | 4 | CYBA, 4687, CSNK2A1, HNRNPC |

| | | |
|---|---|---|
| 4 | 4 | DVL3, PPP2CA, TP53, DAXX |
| 5 | 4 | GATA3, ETS1, NR3C1, COPS6 |
| 6 | 4 | HLA-DQB1, CD4, PIK3R1, AKT1 |
| 7 | 4 | PITX2, KAT5, CDK1, AMPH |
| 8 | 4 | PTPRA, KCNA2, DLG1, PAX6 |
| 9 | 4 | RPS14, SMAD2, TSC2, MAPKAPK2 |
| 10 | 4 | TPI1, CFL1, ATXN1, KIAA2026 |
| 11 | 4 | SIP1, SNRPD2, EGFR, MET |
| 12 | 4 | MAP4K4, ITGB1, CRKL, EPOR |
| 13 | 4 | ERC1, YWHAG, LUC7L2, UNC119 |
| 14 | 4 | CLASP2, FEZ1, PRKCZ, GSK3A |
| 15 | 4 | GGA3, TSG101, NR3C1, SUMO4 |
| 16 | 4 | TES, ACTN1, GRIN2A, PTPN4 |
| 17 | 4 | PSMC3IP, NR3C1, PRKDC, EIF2S2 |
| 18 | 4 | PIDD, EFEMP2, TP53, PLK3 |
| 19 | 4 | MIF4GD, UBQLN4, IMPDH2, SUMO4 |
| 20 | 4 | STK11IP, SMAD4, MAPK13, MAPKAPK3 |
| 21 | 5 | BLK, BCL2, CDK2, PRKAR1A, C2orf88 |
| 22 | 5 | DYNC1H1, YWHAG, ARAF, TH1L, FRMD5 |
| 23 | 5 | STX2, STXBP1, PRKCA, TIAM1, MAPK8IP1 |
| 24 | 5 | RBPJ, HMGB1, C14orf1, NSF, NAPG |
| 25 | 5 | MUC4, ERBB2, PTPN18, GAB1, MAPK4 |
| 26 | 5 | MYO5A, DYNLL1, MTA1, CCNH, CDK2 |
| 27 | 5 | PIN1, CHPF, SMAD9, LNPEP, TNKS2 |
| 28 | 5 | RAB5A, TSC2, SMAD2, HDAC1, DNMT3B |
| 29 | 5 | RAC2, CUL1, SMAD3, GGA1, M6PR |
| 30 | 5 | ENC1, TGFBR1, FBXO34, SKP1, FBXL8 |
| 31 | 5 | MADD, PIDD, CRADD, LRIF1, RNF10 |
| 32 | 5 | NRXN1, SYT1, GOLM1, NIPSNAP3A, EPHX2 |
| 33 | 5 | PRDX6, RARA, COPS2, COPS6, WIPI2 |
| 34 | 5 | CAMKK2, CALM1, CAMK2G, GRIN2B, AP4M1 |
| 35 | 5 | MAST3, PTEN, CSNK2A2, SMURF1, NAA16 |
| 36 | 5 | PPIL2, HSP90AA1, WASL, SH3GL3, C11orf68 |
| 37 | 5 | PTRH2, AES, AR, CDC25A, PIM1 |
| 38 | 5 | DNAJB11, PTN, BCCIP, RAD51, DMC1 |
| 39 | 5 | KLHDC5, COIL, SMN1, BCL2, PPP3CA |
| 40 | 5 | HIF3A, HIF1A, CREBBP, MED25, MED15 |
| 41 | 5 | COL18A1, KDR, SRC, PRKACA, TPH1 |
| 42 | 5 | IQCG, BAG6, SMN1, KPNB1, UBR5 |
| 43 | 6 | CSF3, CSF3R, GRB2, EPHB6, SAT1, SAT2 |
| 44 | 6 | MUC2, PLEKHM1, EIF2S2, CSNK2A1, CDK1, NES |
| 45 | 6 | CLIP2, DYNLL1, TP53BP1, EP300, MYBL2, ZNF622 |
| 46 | 6 | PRPF4B, YWHAG, PRKCA, ITGB2, HP, C1RL |
| 47 | 6 | BRE, GFI1B, PSMA3, CDKN1A, RAB1A, ZNF593 |
| 48 | 6 | EDEM1, CANX, SMURF2, NEK6, CDK7, GTF2H2 |
| 49 | 6 | MAML1, CREBBP, EWSR1, RALYL, ZNF408, ZNF330 |
| 50 | 6 | SEC23B, SEC24D, LMO4, MERTK, BMPR2, PDZRN3 |
| 51 | 6 | CECR2, UXT, AR, RB1, TRIM27, FXYD6 |
| 52 | 6 | FBXO30, SMAD1, MAPK1, NEK2, NDC80, SPC25 |

| | | |
|---|---|---|
| 53 | 7 | EIF4EBP2, EIF4E, PML, RELA, BRCA1, PSAP, CELSR1 |
| 54 | 7 | TNKS1BP1, TNKS, FNBP1, CDC42, WAS, CIB1, IFI6, |
| 55 | 8 | IRAK4, TRAF6, TRAF2, TCEA2, CENPT, PPCDC, DBI, TSPO |

**Table 4-9. Exon paths lengths and genes in path from source to target.**

**Table 4-10. Functional enrichment of exon paths, between source and target (separate file).**
File Table4-10.xlsx. Gene sets include only genes that map to an Entrez ID.

### *4.2.10 Specific example of exonic eSNP*

We show an example (Figure 4-8b) of exon source and its gene target, examining the path between gene source p53-induced death domain protein (*PIDD*) and gene target polo-like kinase 3 (*PLK3*); path number 18, Tables 4-9 and 4-10). This path was enriched for the p53 signaling pathway (FDR <0.01, Table 4-10). *PIDD* promotes apoptosis downstream of the tumor suppressor as a component of the DNA damage/stress response pathway that connects p53 to apoptosis. The gene target *PLK3* is a serine/threonine kinase that plays a role in regulation of cell cycle progression and potentially in tumorgenesis. Epidermal growth factor-containing fibulin-like extracellular matrix protein 2 (*EFEMP2*)and tumor protein p53 (*TP53*) reside along the shortest path between *PIDD* and *PLK3* (Figure 4-8b). There is evidence from ChIP-ChIP and ChIP-seq experiments that *TP53* has binding sites in the promoter of *PLK3* [109] and it is annotated as a zinc ion binding protein. Furthermore, the combination of a pair of genes with TF-DNA and PPI edge between them is a known network motif (mixed-feedback loop) [82], suggesting a mechanism by which the expression of the target gene is altered. In support of this, the co-expression correlation of the source and target genes was significant (Spearman rank-correlation test $r = 0.3223$, $P < 0.02$). The exon gene source and target reside in different PPI

network clusters: *PIDD* resides in a cluster that is enriched for regulation of cell death (FDR $<4.5\times10^{-6}$, Table 4-8) and *PLK3* resides in a cluster that is enriched for regulation of transcription (FDR $<2.4\times10^{-88}$, Table 4-8).

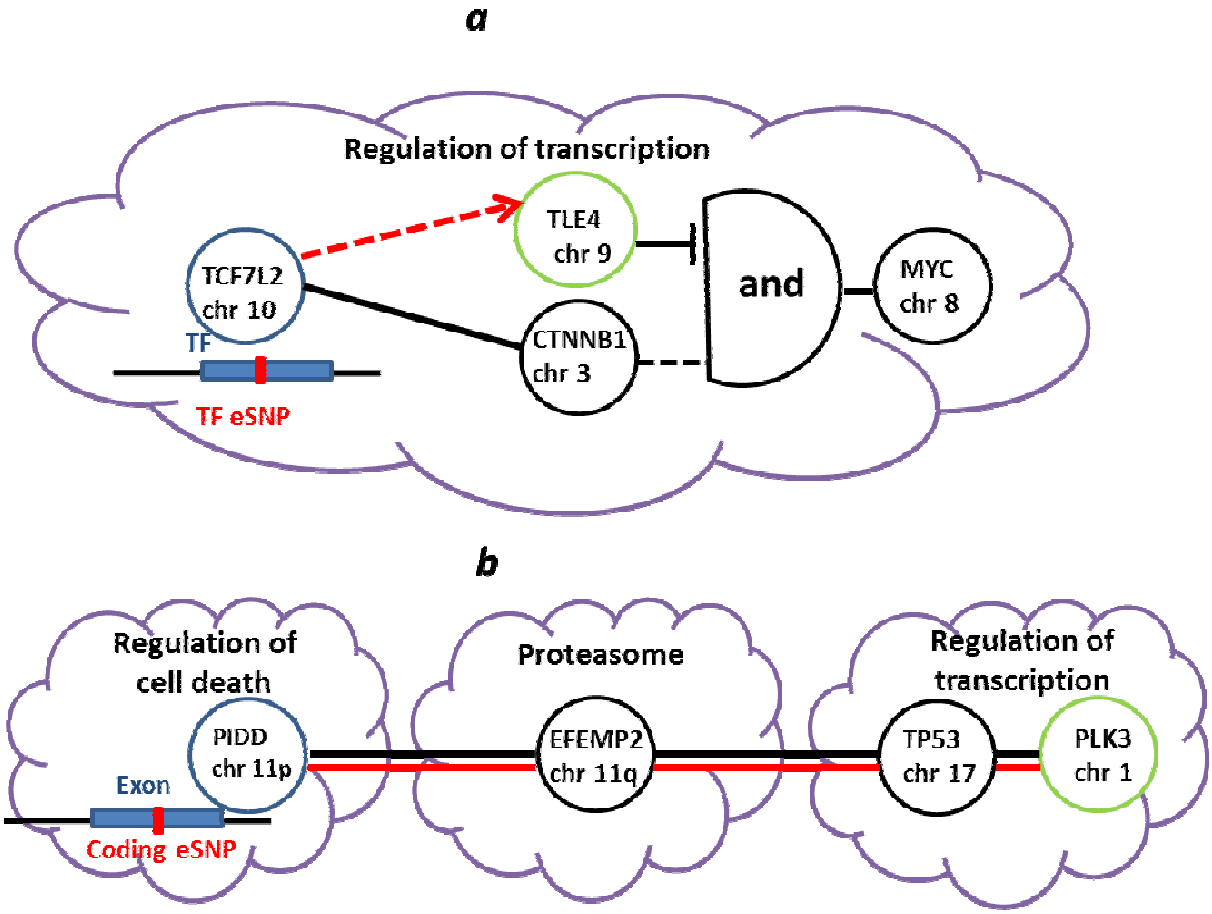


**Figure 4-8. Examples of transcription factors and exon source-target pairs.**
An eSNP (red tick mark) along a source gene (blue circle), either in an exon or TF (blue rectangle), is associated (solid red line for exon, dashed for TF) with levels of transcription of the target gene (green circle). The source and target genes interact via nodes (black circles) and edges (black solid lines) in the PPI network. Each node belongs to a PPI cluster (purple cloud) with a functional annotation. (a) Network motif I1-FFL [92]: TCF7L2 activates MYC (in the presence of CTNNB1) but also represses MYC by activating the repressor TLE4 (via an eSNP) . (b) The shortest path on the PPI network between PIDD source and its gene target PLK3. Binding sites of TP53 were found in the promoter of PLK3. TP53 is annotated as a zinc ion binding protein. There was a significant correlation between the expression of the source and target genes. TCF7L2, transcription factor 7-like 2; T-cell specific; TLE4 transducin-like enhancer of split 4; MYC, v-myc avian myelocytomatosis viral oncogene; catenin (cadherin-associated protein), beta 1, 88kDa (CTNNB1); PIDD, p53- induced death domain protein; PLK3, polo-like kinase 3; EFEMP2, Epidermal growth factor-containing fibulin-like extracellular matrix protein 2; TP53, tumor protein p53.

### *4.2.11 Mechanistic interpretation of exonic eSNPs*

These results beg a mechanistic explanation that would clarify how the network interaction at the protein level is leading to the observed changes in transcript levels. Fortunately, examination of the genes along the reported paths provides a plausible answer, as they are strongly enriched for zinc ion binding proteins. Specifically, when we examined the enrichment for annotations of genes along shortest paths in the real dataset, we observed 410 enriched categories (minimum of 10 genes from a category, FDR <0.05; Table 4-11; Materials and Methods section 4.3.12). For comparison, across 1,000 permuted datasets we observed a total of 1,870 categories satisfying the same enrichment criteria. We focus on the six categories that were enriched in real data and not in permutations: ion binding, metal ion binding, cation binding and intracellular, zinc ion binding and transition metal ion binding (Table 4-11). We compared two properties in real versus permuted datasets: first, the number of genes from each category (empirical *P*-values 0.005 and 0.014 for zinc ion binding and transition metal ion binding respectively); and second the number of paths where we observed at least one gene from each category (empirical *P*-values 0.016 and 0.038 for zinc ion binding and transition metal ion binding respectively). These results were replicated in a second permuted dataset. For comparison, only 7 and 10 out of the 404 joint categories achieve an empirical *P*-value lower than 0.05 for these two properties respectively. These results indicate that the genes in real paths were enriched for zinc ion binding, which is associated with regulation of transcription, suggesting a possible mechanism by which the expression level of the target transcript is modified.

**Table 4-11. Enriched annotations (minimum 10 genes, FDR <0.05) of genes along real and permuted data shortest paths, and gene names for the six categories that were enriched in real shortest paths (separate file).**
File Table4-11.xlsx.

## 4.3 Discussion

We present a computational approach to study the characteristics of *trans* regulation. We observed that candidate eSNPs within exons exhibited an overabundance of significant association signals. We consequently focused on eSNPs that resided within an exon of a source gene, and were associated with the expression level of a different gene target. We observed that candidate eSNPs within TFs were associated with a higher number of transcripts than expected by chance. We subsequently examined eSNPs that resided within the span of source TFs. We mapped these pairs of source and target onto a PPI network and analyzed their topological properties.

We applied our approach to publicly available genetics and genomics [40] data from the same samples. We demonstrated that, by combining association data with information on PPI, it is possible to unravel topological properties for the two *trans* association types. We found that for an eSNP exon source and its gene target, the stronger the association, the closer the source-target distance and the higher the target degree in the PPI network. Expression analysis showed these source-target pairs to be frequently co-expressed, and that these exon eSNPs often had significant *cis* effects on the expression of the source genes. The observed phenomenon of exonic variation leaving a signature on PPI paths raises speculations regarding the mechanisms of transcription regulation. Previous studies have indirectly tackled these speculations regarding the connection between eSNP regulation and the PPI space. Specifically, Rossin *et al*. found that PPI connections between loci defined in GWAS of a specific disease were more densely connected than chance expectation [26], and Nicolae *et al*. [17] observed that SNPs found in GWAS were more likely to be eSNPs. The comprehensiveness of our work relied on combining

eQTL data with the PPI network and not merely GWAS data, as described in previous studies [105]. This allowed us to examine source-target connections across the network, rather than be limited to studying the source nodes as in GWAS-PPI analyses. The novel observation is that the genetic variation that modifies PPI network properties is associated with a normal expression landscape and not only with extreme cases of disease.

We attempted to go beyond topological results and shed light on the regulatory mechanism by which gene expression of the target gene is altered in these shorter paths. We systematically compared genes along real and permuted shortest paths and found enrichment for ion zinc binding proteins, suggesting a plausible mechanism by which the expression level of the target transcript is modified. More generally, the paths of interacting protein pairs, from a source protein to the target protein, were consistent with concatenation of two pathways (Figure 4-9). The prefix of the path was consistent with a regulatory pathway, leading to some regulatory protein (TF or other) that affects expression of the target. The suffix of the path may match a self feedback loop in reverse: from the target protein back to the same regulatory protein [37].

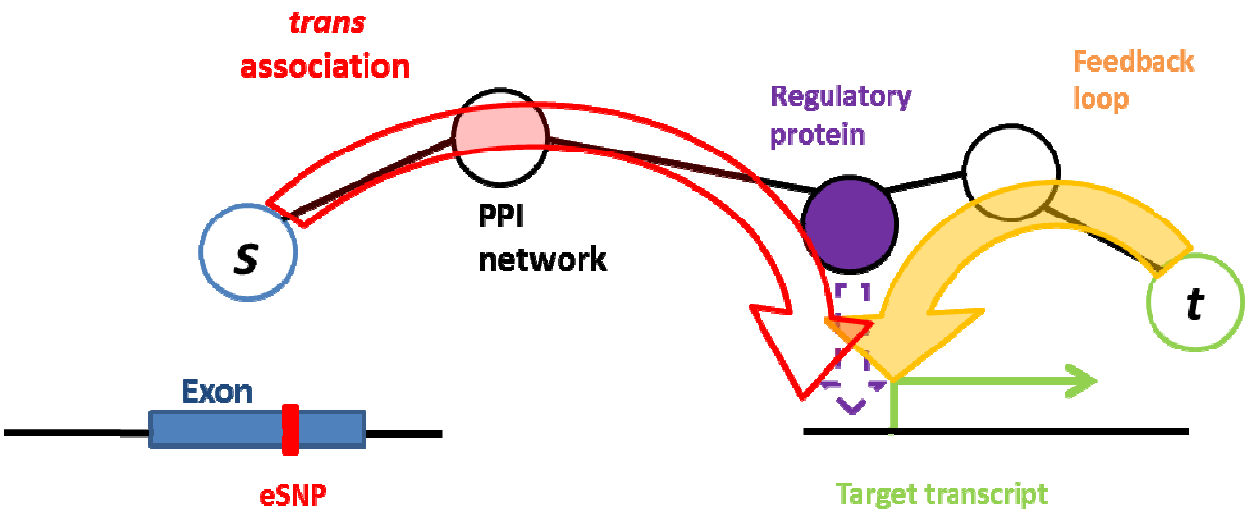


**Figure 4-9. Mechanistic interpretation of exonic eSNPs.**
A path of interacting protein pairs (black circles and connectors) along the PPI network, from a source protein (blue) to the target transcript and protein (green), is consistent with concatenation of two pathways: the prefix of the path is consistent with a regulatory pathway (red), leading to some regulatory protein (purple node), that (directly or indirectly) affects expression of the target (purple arrow), thus being observed as a trans-eQTL signal. The suffix of the path may match a self feedback loop in reverse: from the target protein back to the same regulatory protein (orange arrow).

We demonstrated it is possible to characterize regulatory variation in TFs. We observed that eSNP TF sources and their gene targets create units of genes that are enriched for functional annotations. When decomposing the PPI network to clusters, we observed that these source-target pairs tend to reside within the same cluster.

The design choices for a study of this kind convey a few methodological limitations. First, because we were interested in detecting putatively causal variants based on their exact genomic location, we used a dataset of fully sequenced individuals along with their transcription profiles. Such cohort sizes are limited in size, reducing the power to detect association and allowing us to see only the strongest effects. Second, we were interested in understanding the mechanisms

underlying eSNPs interactions. This required the use of a well-established interaction network. We examined our results on a PPI network, rather than a TF-DNA interaction network or co-expression network derived from this dataset, to establish a broad and independent network of interactions. Overall, both the raw datasets [40, 91] and supporting databases [42, 47, 50, 54, 62, 109] in this work were noisy and limited. That we observed statistically significantly plausible results in such a small dataset combined with noisy databases is encouraging. Potentially, an increase in sample size may enable detection of eSNP associations at more significant *P*-values for even milder effects.

Over the last decade, causal interpretation of genetic association signals for common variants and common traits had been impeded by two hurdles. First, many of the signals had been obtained as indirect association to proxy genetic markers, without access to the directly and causally associated variant. Second, often the trait under investigation was not understood at the molecular mechanistic level well enough to decipher the connection between variant and phenotype. This work bridges the gap between association and causality by considering both direct association to sequencing-ascertained variants, as well as expression quantitative traits. The ability to tie together these loose ends of genetic association using an interaction map constitutes a notable stride towards understanding the thousands of such connections that recent genetics have discovered.

Our main findings suggest two modes of *trans* regulation via genetic variation in exons and TFs. Exonic variation possibly acts through mild *cis* effects that alter the expression of the source gene and delineates shorter paths between functional clusters (Figure 4-10a), and exonic eSNP

targets might play an important role in the PPI network as hubs. TF variation frequently acts via a modular regulation mechanism, with multiple targets that share a function with the TF source (Figure 4-10b).

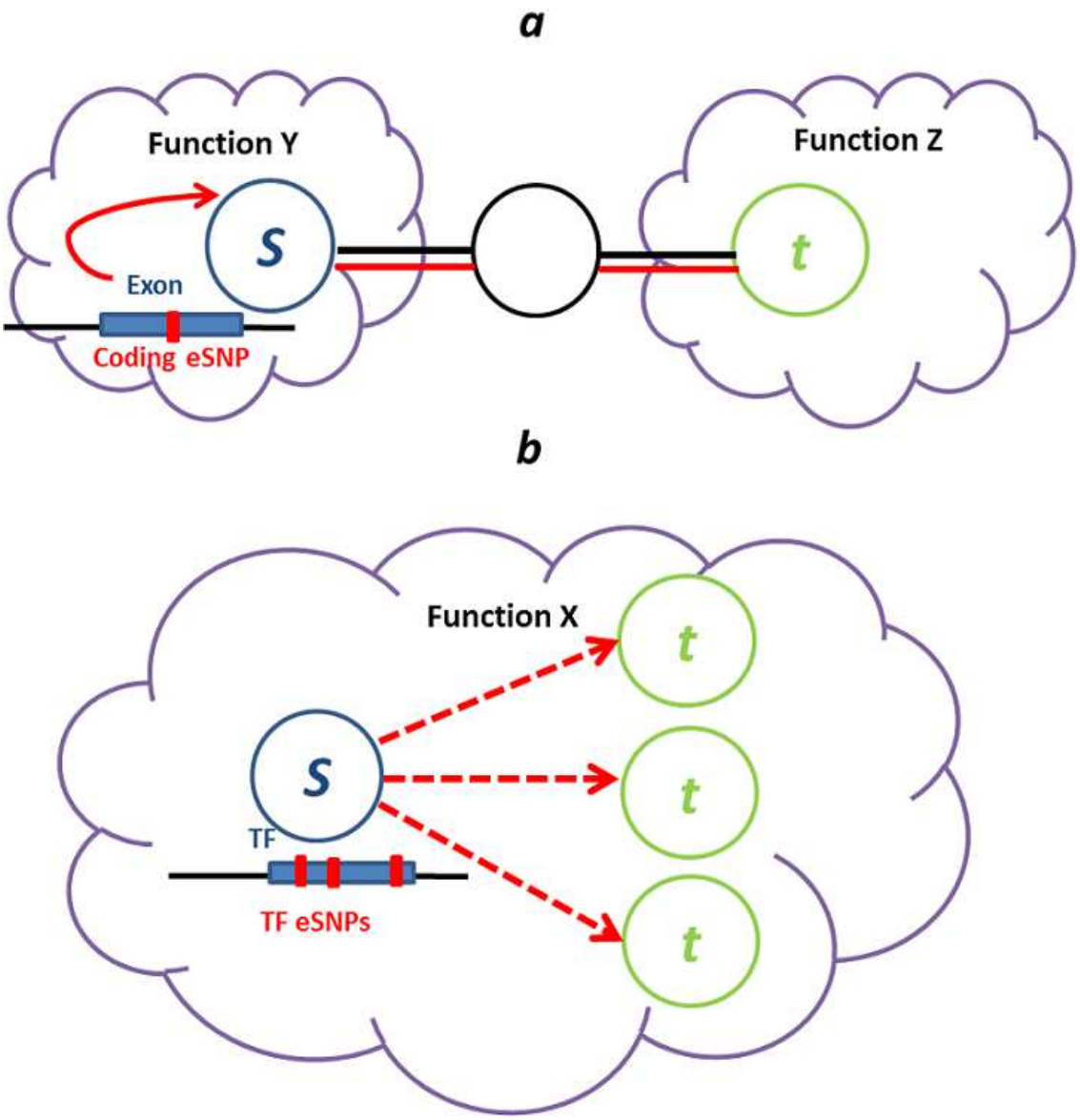


**Figure 4-10. Summary illustration - two suggested modes of *trans* regulation.**
(a) Exon variation often acts by a local cis effect, delineating shorter paths of interacting proteins across functional clusters of the PPI network. (b) TF variation frequently acts via a modular regulation mechanism, with multiple targets that share a function with the TF source. (See Figure 4-8 legend for further details).

Future studies could extend the approach presented here to investigate how genetic variation in different meaningful genomic locations (for example, enhancers, insulators, miRNAs) correlates with gene targets. Datasets that combine sequenced variants coupled with gene expression and phenotypic traits are limited in human, but available for other model organisms [112, 113]. It would be insightful to combine this type of study with phenotypic data, to see how t*rans* association tracks with phenotypes. Specifically, applying our approach to samples under various conditions (for example, disease), could improve understanding of condition-specific regulatory processes [26]. Moreover, considering genetics-genomics data across different tissues along with a tissue-specific PPI network [114] could be telling regarding the underlying regulatory mechanisms characterizing these tissues.

# 4.4 Materials and Methods

### *4.3.1 Data details and processing*

We analyzed a cohort of 50 Yoruban samples, for which genotypes of SNPs that are fully ascertained from sequencing data [91] along with RNA-sequencing data [40] are publicly available. Briefly, the raw dataset consists of 10,553,953 genotyped SNPs and expression measurements (quantile-quantile normalized values) of 18,147 genes with Ensembl gene ID across these 50 samples. Standard filters have been applied to the genetic data: minor allele frequency >0.05, SNP missingness rate <0.1 and individual missingness rate <0.1 [46]. After filtering, data for analysis consist of 50 samples with 7,206,056 SNPs.

### *4.3.2 Association testing*

For association analysis, we considered only SNPs that resided within candidate regulatory regions along the genome. For *trans* association, we tested for association between a SNP and every gene; we considered SNPs within the span of known exons and TFs (including introns) [94]. We tested for association using linear regression.

### *4.3.3 Obtaining a random distribution of association test statistics*

Examining the random distribution of association tests was helpful in evaluating the empirical significance of results. This was achieved by generating 100,000 random pairs of sources and targets for exonic and TF variation separately. We used a strict randomization process of edges switching. We picked a source gene from all sources in the real data; we then picked a target gene from all targets in the real data with a *P*-value cutoff of $10^{-6}$. When evaluating the number

of targets per TF source, we created 1,000 sets of random TF source and gene target pairs; each set contained 370 such pairs corresponding to 370 TF source-target pairs at a *P*-value cutoff of $10^{-6}$ in the real data.

### *4.3.4 Identifying topological trends across association P-values*

For exons, we observed the emergence of true positive associations between *P*-values $10^{-6}$ and $10^{-7}$ (Figure 4-1). Therefore, we focused on *P*-values $<10^{-6}$ and sorted all source-target pairs according to the significance of their association signal. We considered each prefix of this list, that is, each subset of source-target pairs exceeding a particular threshold, for significance of association signal. For each such subset, we reported each one of the topological properties defined above averaged over the subset. We calculated Spearman's correlation coefficient between significance thresholds and each of these cumulative averages. In a similar process, we randomly chose an equal number of arbitrary source-target pairs on the PPI network. Adding these pairs one by one created a distribution of analogous cumulative averages for permuted pairs. We recorded the Spearman correlation coefficient for these 100,000 permuted distributions. We calculated the empirical *P*-value for the significance of the observed correlation coefficients by counting the number of times when permuted r > real r and divided this by the number of permutations.

### *4.3.5 Identifying topological properties of source-target pairs projected on the PPI network*

We used the PPI network provided by the Human Protein Reference Database [42]. The undirected network contains 9,671 nodes and 37,041 edges. For each node, we calculated its

degree: the number of edges incident on the node. We defined a distance between every two nodes as the number of edges on the shortest path between them. All pair-wise shortest paths were determined using the Floyd-Warshall algorithm [15]. In cases where the network had more than one connected component, nodes from two different components were defined to have a distance of twice the maximal distance obtained within the components.

### *4.3.6 Expression analysis*

We calculated all pairwise co-expression correlations for all gene pairs in the dataset using Spearman rank-correlation test, and therefore obtained the distribution of the correlation coefficient *r*. To determine whether the distribution of *r* between source-target pairs differed from its background distribution, we employed the Wilcoxon ranked-sum test.

### *4.3.7 Enrichment of eSNPs for cis effects*

We examined whether eSNPs that were associated with a target's expression level also affected expression levels of the corresponding source. We tested this by considering, for each source-target pair, the one eSNP most associated to the expression for the target. We tallied the source-target pairs for which this eSNP was also significantly associated ($P$ <0.05) with the expression level of the source. Under the null, the number of such pairs is a random variable that is binomially distributed. Bin (n = number ofunique source genes, $P$ =0.05).

### *4.3.8 Unit and path annotation*

We defined units of genes by considering a TF source and its gene targets. We examined shortest paths within the PPI network between eSNP exon source and its gene target. The enrichment of units and paths with gene subsets from the Gene Ontology [62], and KEGG [47] databases was

calculated by Genatomy [13]. We reported only units or paths with annotations that had a significant FDR of 0.05 or better. The description of genes in units or paths is cited from the National Center for Biotechnology Information Gene database and GeneCards [115].

#### *4.3.9 Finding transcription factor source-target pairs in the experimental database*

The ChIP Enrichment Analysis (ChEA) database [109] represents a collection of interactions describing the binding of transcription factors to DNA, collected from ChIP-X (ChIP-chip, ChIP-sequencing, ChIP-positron emission tomography and DNA adenine methyltransferase identification) experiments. For each TF source and target, we examined if they were present in ChEA. We repeated the same procedure for 100,000 permuted pairs of a random TF source and a random gene target. We then compared, using Fisher's exact test, the number of pairs in ChEA between real and permutation pairs, out of all pairs where the TF source was included in the database.

#### *4.3.10 Finding PPI network decomposition to clusters*

The decomposition of the PPI network to clusters was computed by using the Louvain algorithm presented in *[95]*. This is a heuristic method that is based on modularity optimization. The method consists of two phases and partitions the network into clusters such that the number of edges between clusters is significantly less than expected by chance. The method provides a mathematical measure for modularity with network-size normalized values, ranging from 0 (low modularity) to 1 (maximum modularity). This method has been previously applied to various biological networks *[116]* and specifically to a PPI network *[117]*.

### *4.3.11 Significance of source and target residing in the same PPI cluster*

For each exon and TF source-target pair, we recorded whether both source and target resided in the same PPI cluster. We repeated the same procedure with 100,000 permuted unique source-target pairs from nodes on the PPI network. We then compared the number of cluster co-occurrences between real data and permutations using the Fisher exact test.

### *4.3.12 Comparing shortest paths annotation content*

We recorded all genes along the shortest paths between exonic sources and targets, both in real and permuted data. We then looked for enrichment in this set of genes (at least 10 genes per category, FDR <0.05). We created sets of 1,000 permuted 55 shortest paths (from the 17,564 shortest paths in permutations) that followed the exact length distribution of the 55 real paths. For each one of the six categories that was not enriched in permutations, we performed two analyses: first, we counted how many genes from each category appeared in the real paths (with repetitions, that is if gene X from category Y appeared in two shortest paths we counted it twice); and second, we counted how many of the 55 paths had at least one gene from this category. We repeated the same procedures for the 1,000 permuted sets. For each category, we then counted how many of the 1,000 permutations achieved equal or greater numbers than seen for the real data (empirical *P*-value).

# Chapter 5: Conclusions

Variants that are associated with changes in gene expression (eSNPs) are known to play a role in many human traits [17], making them the subject of recent research efforts. Here, we focus on eSNPs in *trans*, as they provide insight on regulatory interactions between different loci and the structure of the regulatory network that such interactions define. First, we present a novel approach for defining network motifs, including regulatory modules [36]. Second, we extend this approach to discover bi-fan structures [22]. Third, we devise a computational framework where we project eSNP associations onto a PPI network to characterize properties of eSNPs and their targets [38] . Overall, our work offers insights concerning the topological structure of human regulatory networks and the effect genetic variation has on shaping them.

We assemble modules of transcripts each associated to the same main SNP; then assign a confidence score to each module, lastly we determine intra-module topology from the dependencies between the transcripts in the module and the main SNP [36]. We show these modules to be high confidence structures. We apply our method to data on human liver expression and SNP genotypes [52] and find that *trans* regulation exhibits a modular structure with a single variant that is associated with a set of genes and shares the same annotation descriptors with them. This regulation structure is usually mediated by a *cis* effect of the main SNP on the expression of a close gene, and the direction of effect on the genes in the module is mostly consistent (either up or down regulation). There are significantly more modules, and they are bigger, denser and more enriched in annotations than those observed in the permuted data, providing support for our methodology.

We extend this approach to define quartet structures comprised of a pair of two main SNPs that are associated to the same pair of transcripts [22]. We uncover a bi-fan motif in the human regulatory network [22], which was previously described as a building block of model organisms' regulatory networks [37]. This regulatory structure is nearly exclusive to real data, and exhibits unique characteristics. Most human bi-fans involve pairs of eSNPs located on different chromosomes, away from their targets which are likewise located on different chromosomes. All quartets are consistent in terms of the direction of eSNP effects on correlated and anti-correlated transcripts and there is enrichment for eSNPs with the same-direction effects, i.e., the directional effect of both eSNPs on a transcript is the same. We replicate these characteristics in a larger dataset [1].

Finally, we present a computational framework that integrates eSNPs within exons with a PPI network [38]. We then compare eSNPs in exons with eSNPs in TFs to uncover characteristics of *trans* regulation. We applied our approach to publicly available genetics and genomics [40, 91] data from the same samples. Our findings suggest two distinct modes of *trans* regulation:
Exon variation possibly acts through mild *cis* effects that alter the expression of the source gene. The exonic source and target, which are frequently co-expressed, seem to be connected by shorter paths between functional clusters and the target degree is higher. Moreover, we find enrichment for ion zinc binding proteins, suggesting a plausible mechanism by which the expression level of the target transcript is modified. TF variation frequently acts via a modular regulation mechanism, with multiple targets that share a function with the TF source.

The advances in sequencing and RNA-seq technologies and the drop of prices make this an exciting time for genetics-genomics research, but there are still some substantial limitations to overcome. First, the traditional use of SNP arrays for genotyping in large scale genetic studies is limiting to findings that are predominantly tags for causal variants. Cohorts that include both RNA-seq for gene expression and sequencing-ascertained variants for genotyping are still limited in size [1, 39, 40], reducing power for eSNP associations discovery. Second, findings in eSNPs studies are commonly supported by annotation data bases [42, 47, 54, 62] that are noisy, partial and in many cases publication biased. In the recent ENCODE effort [118] it was established that most of disease associated variants are located within regulatory regions [119], highlighting the importance of improving whole genome annotation and not merely focusing on the coding regions. Finally, statistical and computational approached are helpful in shortlisting candidate loci that have high susceptibility to affect phenotypes. Such findings should be accompanied by experimental validations, which are costly and time consuming.

A recent conference I attended “The biology of genomes” provided a good snapshot of the field and where it is headed. There is an effort to produce and make publicly available datasets with large number of samples. A good example is the Geuvadis dataset [1] which includes RNA-seq and genotyping data for more than 450 samples from different populations. The GTEx consortium [13] is collecting and producing RNA-seq and genotype data across multiple tissues. This resource will provide insights into tissue specific regulatory mechanisms. Another important question would be to characterize eQTLs within a specific tissue but in different cell types. A natural extension to the study of eSNPs is to focus on SNPs that are associated with other cellular phenotypes, e.g., long intergenic non-coding RNAs (lincRNAs) expression [120]

and protein levels [121] or focusing on different type of variants that are associated with gene expression, e.g., Short Tandem Repeat (STR) [122]. The findings from such studies will complement and extend the understanding of biological processes. The future goal of this field would be to find and characterize causal variants, understand the mechanisms through which they act and ultimately move from bench to bedside and develop personalized treatment.

The code for all methods presented in this thesis can be found in the following link: http://www.columbia.edu/~ak2996/Software.htm

# References


1. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-11.

2. Ben-Shitrit, T., et al., *Systematic identification of gene annotation errors in the widely used yeast mutation collections.* Nat Methods, 2012. **9**(4): p. 373-8.

3. Gaffney, D.J., et al., *Dissecting the regulatory architecture of gene expression QTLs.* Genome Biol, 2012. **13**(1): p. R7.

4. Battle, A., et al., *Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals.* Genome Res, 2014. **24**(1): p. 14-24.

5. Bell, J.T., et al., *DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines.* Genome Biol, 2011. **12**(1): p. R10.

6. Degner, J.F., et al., *DNase I sensitivity QTLs are a major determinant of human expression variation.* Nature, 2012. **482**(7385): p. 390-4.

7. McVicker, G., et al., *Identification of genetic variants that affect histone modifications in human cells.* Science, 2013. **342**(6159): p. 747-9.

8. Gamazon, E.R., et al., *Genetic architecture of microRNA expression: implications for the transcriptome and complex traits.* Am J Hum Genet, 2012. **90**(6): p. 1046-63.

9. Fairfax, B.P., et al., *Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles.* Nat Genet, 2012. **44**(5): p. 502-10.

10. Yosef, N., et al., *ANAT: a tool for constructing and analyzing functional protein networks.* Sci Signal, 2011. **4**(196): p. pl1.

11. Brown, C.D., L.M. Mangravite, and B.E. Engelhardt, *Integrative modeling of eQTLs and cis-regulatory elements suggests mechanisms underlying cell type specificity of eQTLs.* PLoS Genet, 2013. **9**(8): p. e1003649.

12. Fu, J., et al., *Unraveling the regulatory mechanisms underlying tissue-dependent genetic variation of gene expression.* PLoS Genet, 2012. **8**(1): p. e1002431.

13. *The Genotype-Tissue Expression (GTEx) project.* Nat Genet, 2013. **45**(6): p. 580-5.

14. Nica, A.C., et al., *The architecture of gene regulatory variation across multiple human tissues: the MuTHER study.* PLoS Genet, 2011. **7**(2): p. e1002003.

15. Fairfax, B.P., et al., *Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression.* Science, 2014. **343**(6175): p. 1246949.

16. Hindorff, L.A., 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**(23): p. 9362-7.

17. Stranger, B.E., et al., *Genome-wide associations of gene expression variation in humans.* PLoS Genet, 2005. **1**(6): p. e78.

18. Nica, A.C., et al., *Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations.* PLoS Genet, 2010. **6**(4): p. e1000895.

19. Montgomery, S.B. and E.T. Dermitzakis, *From expression QTLs to personalized transcriptomics.* Nat Rev Genet, 2011. **12**(4): p. 277-82.

20. Dubois, P.C., et al., *Multiple common variants for celiac disease influencing immune gene expression.* Nat Genet, 2010. **42**(4): p. 295-302.

21. Anttila, V., et al., *Genome-wide association study of migraine implicates a common susceptibility variant on 8q22.1.* Nat Genet, 2010. **42**(10): p. 869-73.

22. Kreimer, A. and I. Pe'er, *Co-regulated transcripts associated to cooperating eSNPs define bi-fan motifs in human gene networks.* Accepted for publication in PLOS Genetics, 2014.

23. Westra, H.J., et al., *Systematic identification of trans eQTLs as putative drivers of known disease associations.* Nat Genet, 2013. **45**(10): p. 1238-43.

24. Purohit, V., B. Gao, and B.J. Song, *Molecular mechanisms of alcoholic fatty liver.* Alcohol Clin Exp Res, 2009. **33**(2): p. 191-205.

25. Raj, T., et al., *Alzheimer disease susceptibility loci: evidence for a protein network under natural selection.* Am J Hum Genet, 2012. **90**(4): p. 720-6.

26. Levkovitz, L., et al., *A novel HMM-based method for detecting enriched transcription factor binding sites reveals RUNX3 as a potential target in pancreatic cancer biology.* PLoS One, 2010. **5**(12): p. e14423.

27. Voineagu, I., et al., *Transcriptomic analysis of autistic brain reveals convergent molecular pathology.* Nature, 2011. **474**(7351): p. 380-4.

28. de Jong, S., et al., *A gene co-expression network in whole blood of schizophrenia patients is independent of antipsychotic-use and enriched for brain-expressed genes.* PLoS One, 2012. **7**(6): p. e39498.

29. Gilman, S.R., et al., *Diverse types of genetic variation converge on functional gene networks involved in schizophrenia.* Nat Neurosci, 2012. **15**(12): p. 1723-8.

30. Gilman, S.R., et al., *Rare de novo variants associated with autism implicate a large functional network of genes involved in formation and function of synapses.* Neuron, 2011. **70**(5): p. 898-907.

31. Califano, A., et al., *Leveraging models of cell regulation and GWAS data in integrative network-based association studies.* Nat Genet, 2012. **44**(8): p. 841-7.

32. Karczewski, K.J., et al., *Coherent functional modules improve transcription factor target identification, cooperativity prediction, and disease association.* PLoS Genet, 2014. **10**(2): p. e1004122.

33. Setty, M., et al., *Inferring transcriptional and microRNA-mediated regulatory programs in glioblastoma.* Mol Syst Biol, 2012. **8**: p. 605.

34. Fehrmann, R.S., et al., *Trans-eQTLs reveal that independent genetic variants associated with a complex phenotype converge on intermediate genes, with a major role for the HLA.* PLoS Genet, 2011. **7**(8): p. e1002197.

35. Boel Brynedal, T.R., Barbara E Stranger, Robert Bjornson, Benjamin M Neale, Benjamin F Voight, Chris Cotsapas, *Cross-phenotype meta-analysis reveals large-scale trans-eQTLs mediating patterns of transcriptional co-regulation.* arXiv preprint arXiv:1402.1728, 2014.

36. Kreimer, A., et al., *Inference of modules associated to eQTLs.* Nucleic Acids Res, 2012. **40**(13): p. e98.

37. Milo, R., et al., *Network motifs: simple building blocks of complex networks.* Science, 2002. **298**(5594): p. 824-7.

38. Kreimer, A. and I. Pe'er, *Variants in exons and in transcription factors affect gene expression in trans.* Genome Biol, 2013. **14**(7): p. R71.

39. Montgomery, S.B., et al., *Transcriptome genetics using second generation sequencing in a Caucasian population.* Nature, 2010. **464**(7289): p. 773-7.

40. Pickrell, J.K., et al., *Understanding mechanisms underlying human gene expression variation with RNA sequencing.* Nature, 2010. **464**(7289): p. 768-72.

41. Xiao, S., et al., *Small-molecule RORgammat antagonists inhibit T helper 17 cell transcriptional network by divergent mechanisms.* Immunity, 2014. **40**(4): p. 477-89.

42. Kheradpour, P., et al., *Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay.* Genome Res, 2013. **23**(5): p. 800-11.

43. Cheung, V.G. and R.S. Spielman, *Genetics of human gene expression: mapping DNA variants that influence gene expression.* Nat Rev Genet, 2009. **10**(9): p. 595-604.

44. Cookson, W., et al., *Mapping complex disease traits with global gene expression.* Nat Rev Genet, 2009. **10**(3): p. 184-94.

45. Rockman, M.V. and L. Kruglyak, *Genetics of global gene expression.* Nat Rev Genet, 2006. **7**(11): p. 862-72.

46. Yvert, G., et al., *Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors.* Nat Genet, 2003. **35**(1): p. 57-64.

47. Moffatt, M.F., et al., *Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma.* Nature, 2007. **448**(7152): p. 470-3.

48. Kathiresan, S., et al., *Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans.* Nat Genet, 2008. **40**(2): p. 189-97.

49. Nicolae, D.L., et al., *Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.* PLoS Genet, 2010. **6**(4): p. e1000888.

50. Litvin, O., et al., *Modularity and interactions in the genetics of gene expression.* Proc Natl Acad Sci U S A, 2009. **106**(16): p. 6441-6.

51. Gilad, Y., S.A. Rifkin, and J.K. Pritchard, *Revealing the architecture of gene regulation: the promise of eQTL studies.* Trends Genet, 2008. **24**(8): p. 408-15.

52. Schadt, E.E., et al., *Mapping the genetic architecture of gene expression in human liver.* PLoS Biol, 2008. **6**(5): p. e107.

53. Ghazalpour, A., et al., *Integrating genetic and network analysis to characterize genes related to mouse weight.* PLoS Genet, 2006. **2**(8): p. e130.

54. Zhong, H., et al., *Integrating pathway analysis and genetics of gene expression for genome-wide association studies.* Am J Hum Genet, 2010. **86**(4): p. 581-91.

55. Yang, X., et al., *Systematic genetic and genomic analysis of cytochrome P450 enzyme activities in human liver.* Genome Res, 2010. **20**(8): p. 1020-36.

56. Hartwell, L.H., et al., *From molecular to modular cell biology.* Nature, 1999. **402**(6761 Suppl): p. C47-52.

57. Ihmels, J., et al., *Comparative gene expression analysis by differential clustering approach: application to the Candida albicans transcription program.* PLoS Genet, 2005. **1**(3): p. e39.

58. Schadt, E.E., et al., *An integrative genomics approach to infer causal associations between gene expression and disease.* Nat Genet, 2005. **37**(7): p. 710-7.

59. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.

60. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

61. Birney, E., et al., *An overview of Ensembl.* Genome Res, 2004. **14**(5): p. 925-8.

62. Hoffmann, R., *A wiki for the life sciences where authorship matters.* Nat Genet, 2008. **40**(9): p. 1047-51.

63. Lee, H.S., et al., *Several regions in the major histocompatibility complex confer risk for anti-CCP-antibody positive rheumatoid arthritis, independent of the DRB1 locus.* Mol Med, 2008. **14**(5-6): p. 293-300.

64. Mantovani, A., et al., *Cancer-related inflammation.* Nature, 2008. **454**(7203): p. 436-44.

65. Zhai, Y., et al., *Cutting edge: TLR4 activation mediates liver ischemia/reperfusion inflammatory response via IFN regulatory factor 3-dependent MyD88-independent pathway.* J Immunol, 2004. **173**(12): p. 7115-9.

66. Erridge, C., et al., *Oxidized phospholipid inhibition of toll-like receptor (TLR) signaling is restricted to TLR2 and TLR4: roles for CD14, LPS-binding protein, and MD2 as targets for specificity of inhibition.* J Biol Chem, 2008. **283**(36): p. 24748-59.

67. Wasan, K.M., et al., *Impact of lipoproteins on the biological activity and disposition of hydrophobic drugs: implications for drug discovery.* Nat Rev Drug Discov, 2008. **7**(1): p. 84-99.

68. Thomas, E.A., et al., *Antipsychotic drug treatment alters expression of mRNAs encoding lipid metabolism-related proteins.* Mol Psychiatry, 2003. **8**(12): p. 983-93, 950.

69. Canto, C., et al., *AMPK regulates energy expenditure by modulating NAD+ metabolism and SIRT1 activity.* Nature, 2009. **458**(7241): p. 1056-60.

70. Smith, D.P., et al., *LIP1, a cytoplasmic protein functionally linked to the Peutz-Jeghers syndrome kinase LKB1.* Hum Mol Genet, 2001. **10**(25): p. 2869-77.

71. Kwon, O.B., et al., *Neuregulin-1 regulates LTP at CA1 hippocampal synapses through activation of dopamine D4 receptors.* Proc Natl Acad Sci U S A, 2008. **105**(40): p. 15587-92.

72. Dhillon, A.S., et al., *MAP kinase signalling pathways in cancer.* Oncogene, 2007. **26**(22): p. 3279-90.

73. Hynes, N.E. and H.A. Lane, *ERBB receptors and cancer: the complexity of targeted inhibitors.* Nat Rev Cancer, 2005. **5**(5): p. 341-54.

74. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

75. Akavia, U.D., et al., *An integrated approach to uncover drivers of cancer.* Cell, 2010. **143**(6): p. 1005-17.

76. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.

77. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.

78. Kaplan, T., et al., *Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development.* PLoS Genet, 2011. **7**(2): p. e1001290.

79. Listgarten, J., et al., *Correction for hidden confounders in the genetic analysis of gene expression.* Proc Natl Acad Sci U S A, 2010. **107**(38): p. 16465-70.

80. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans.* Nature, 2013. **501**(7468): p. 506-11.

81. Gerstein, M.B., et al., *Architecture of the human regulatory network derived from ENCODE data.* Nature, 2012. **489**(7414): p. 91-100.

82. Wu, C., et al., *Induction of pathogenic TH17 cells by inducible salt-sensing kinase SGK1.* Nature, 2013. **496**(7446): p. 513-7.

83. Costanzo, M., et al., *The genetic landscape of a cell.* Science, 2010. **327**(5964): p. 425-31.

84. Hannum, G., et al., *Genome-wide association data reveal a global map of genetic interactions among protein complexes.* PLoS Genet, 2009. **5**(12): p. e1000782.

85. Fisher, R.A., *XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance.* Transactions of the Royal Society of Edinburgh, 1919. **52**: p. 399-433.

86. Moore, J.H. and S.M. Williams, *Epistasis and its implications for personal genetics.* Am J Hum Genet, 2009. **85**(3): p. 309-20.

87. Prabhu, S. and I. Pe'er, *Ultrafast genome-wide scan for SNP-SNP interactions in common complex disease.* Genome Res, 2012. **22**(11): p. 2230-40.

88. Storey, J.D., J.M. Akey, and L. Kruglyak, *Multiple locus linkage analysis of genomewide expression in yeast.* PLoS Biol, 2005. **3**(8): p. e267.

89. Hemani, G., et al., *Detection and replication of epistasis influencing transcription in humans.* Nature, 2014.

90. Stranger, B.E., et al., *Relative impact of nucleotide and copy number variation on gene expression phenotypes.* Science, 2007. **315**(5813): p. 848-53.

91. Hastie, T., R. Tibshirani, and J.H. Friedman, *The elements of statistical learning : data mining, inference, and prediction*. 2nd ed. ed. 2009, New York: Springer. xxii, 745 p.

92. Lee, Y., et al., *Induction and molecular signature of pathogenic TH17 cells.* Nat Immunol, 2012. **13**(10): p. 991-9.

93. Karczewski, K.J., et al., *Cooperative transcription factor associations discovered using regulatory variation.* Proc Natl Acad Sci U S A, 2011. **108**(32): p. 13353-8.

94. Fujita, P.A., et al., *The UCSC Genome Browser database: update 2011*. Nucleic Acids Res, 2011. **39**(Database issue): p. D876-82.

95. Pique-Regi, R., et al., *Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data.* Genome Res, 2011. **21**(3): p. 447-55.

96. Gertz, J. and B.A. Cohen, *Environment-specific combinatorial cis-regulation in synthetic promoters.* Mol Syst Biol, 2009. **5**: p. 244.

97. Cox, R.S., 3rd, M.G. Surette, and M.B. Elowitz, *Programming gene expression with combinatorial promoters.* Mol Syst Biol, 2007. **3**: p. 145.

98. Raveh-Sadka, T., et al., *Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast.* Nat Genet, 2012. **44**(7): p. 743-50.

99. Freilich, S., et al., *Decoupling Environment-Dependent and Independent Genetic Robustness across Bacterial Species.* PLoS Comput Biol, 2010. **6**(2): p. e1000690.

100. Patwardhan, R.P., et al., *Massively parallel functional dissection of mammalian enhancers in vivo.* Nat Biotechnol, 2012. **30**(3): p. 265-70.

101. Freilich, S., et al., *The large-scale organization of the bacterial network of ecological co-occurrence interactions.* Nucleic Acids Res, 2010. **38**(12): p. 3857-68.

102. Sharon, E., et al., *Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters.* Nat Biotechnol, 2012. **30**(6): p. 521-30.

103. Freilich, S., et al., *Metabolic-network-driven analysis of bacterial ecological strategies.* Genome Biol, 2009. **10**(6): p. R61.

104. Melnikov, A., et al., *Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay.* Nat Biotechnol, 2012. **30**(3): p. 271-7.

105. Friedman, J.H. and C.B. Roosen, *An introduction to multivariate adaptive regression splines.* Stat Methods Med Res, 1995. **4**(3): p. 197-217.

106. Weingarten-Gabbay, S. and E. Segal, *The grammar of transcriptional regulation.* Hum Genet, 2014. **133**(6): p. 701-11.

107. Chang, X., et al., *Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of 'date' and 'party' hubs.* Sci Rep, 2013. **3**: p. 1691.

108. Davidson, E.H., et al., *A genomic regulatory network for development.* Science, 2002. **295**(5560): p. 1669-78.

109. Yosef, N., et al., *Dynamic regulatory network controlling TH17 cell differentiation.* Nature, 2013. **496**(7446): p. 461-8.

110. Stein, G.Y., et al., *Met kinetic signature derived from the response to HGF/SF in a cellular model predicts breast cancer patient survival.* PLoS One, 2012. **7**(9): p. e45969.

111. Yosef, N., et al., *A complex-centric view of protein network evolution.* Nucleic Acids Res, 2009. **37**(12): p. e88.

112. Peters, A. and N. Yosef, *Understanding Th17 cells through systematic genomic analyses.* Curr Opin Immunol, 2014. **28C**: p. 42-48.

113. Yosef, N. and A. Regev, *Impulse control: temporal dynamics in gene transcription.* Cell, 2011. **144**(6): p. 886-96.

114. Kleinewietfeld, M., et al., *Sodium chloride drives autoimmune disease by the induction of pathogenic TH17 cells.* Nature, 2013. **496**(7446): p. 518-22.

115. Arvey, A., et al., *Sequence and chromatin determinants of cell-type-specific transcription factor binding.* Genome Res, 2012. **22**(9): p. 1723-34.

116. Yosef, N., et al., *A supervised approach for identifying discriminating genotype patterns and its application to breast cancer data.* Bioinformatics, 2007. **23**(2): p. e91-8.

117. Yosef, N., A. Kaufman, and E. Ruppin, *Inferring functional pathways from multi-perturbation data.* Bioinformatics, 2006. **22**(14): p. e539-46.

118. Bernstein, B.E., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.

119. Maurano, M.T., et al., *Systematic localization of common disease-associated variation in regulatory DNA.* Science, 2012. **337**(6099): p. 1190-5.

120. Ian McDowell, C.G., GTEx Consortium, Athma A Pai, Timothy E Reddy, Barbara E Engelhardt, *Identification of long ntergenic non-coding RNA QTLs in four tissue types*

*reveals association with metabolic phenotypes.* Biology of genomes conference. Cold spring harbor., 2014.

121. Alexis Battle, Z.K., Sidney Wang, Timothee Flutre, Michael Ford, Amy Mitrano, Yoav Gilad, Jonathan Pritchard, *Genome-wide mass ometry, ribosomal profiling and RNA-sequencing reveal genetic variants associated with postnscriptional gene regulation.* Biology of Genomes conference. Cold spring harbor., 2014.

122. Melissa Gymrek, S.G., Barak Markus, Jenny Chen, Perla I Villarreal, Dina Zielinski, Jonathan Pritchard, Yaniv Erlich, *The contribution of STRs to the genetic architecture of gene expression.* Biology of Genomes conference. Cold spring harbor., 2014.