BIOINFORMATICS

Vol. 00 no. 00 2017
Pages 1–3

A network approach to exploring the functional basis of gene-gene epistatic interactions in disease susceptibility

Danny Kit-Sang Yip [1,*], Landon L. Chan [1,2,*], Iris K. Pang [3], Wei Jiang [4], Nelson L. S. Tang [5], Weichuan Yu[4] and Kevin Y. Yip [1,6,7,8,†]

[1]Department of Computer Science and Engineering, [2]Faculty of Medicine, [3]School of Life Sciences, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong
[4]Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong
[5]Department of Chemical Pathology, [6]Hong Kong Bioinformatics Centre, [7]CUHK-BGI Innovation Institute of Trans-omics, [8]Hong Kong Institute of Diabetes and Obesity, The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Individual genetic variants explain only a small fraction of heritability in some diseases. Some variants have weak marginal effects on disease risk, but their joint effects are significantly stronger when occurring together. Most studies on such epistatic interactions have focused on methods for identifying the interactions and interpreting individual cases, but few have explored their general functional basis. This was due to the lack of a comprehensive list of epistatic interactions and uncertainties in associating variants to genes.

Results: We conducted a large-scale survey of published research articles to compile the first comprehensive list of epistatic interactions in human diseases with detailed annotations. We used various methods to associate these variants to genes to ensure robustness. We found that these genes are significantly more connected in protein interaction networks, are more co-expressed, and participate more often in the same pathways. We demonstrate using the list to discover novel disease pathways.

Contact: kevinyip@cse.cuhk.edu.hk

1 INTRODUCTION

Genome-wide association studies (GWAS) have systematically identified genetic variants within important susceptibility loci in various diseases (Easton *et al.*, 2007; Fellay *et al.*, 2007; Frayling *et al.*, 2007; Plenge *et al.*, 2007; Visscher *et al.*, 2012; The Wellcome Trust Case Control Consortium, 2007). However, for some complex diseases, the identified variants account for only a small portion of disease susceptibility, leading to the question of what causes this "missing heritability" (Eichler *et al.*, 2010; Fuchsberger *et al.*, 2016; Manolio *et al.*, 2009). For example, only 20-25% of the estimated heritability from pedigree studies of Crohn's disease could be explained by the identified common variants from GWAS (Lander, 2011). Several explanations for the missing heritability have been proposed, including the insufficient sample size for detecting common variants with small effects, the presence of rare variants with large effects, and the inflated heritability estimation in pedigree studies attributed to non-additive effects such as epistasis (Gibson, 2012).

In order to evaluate the extent of missing heritability due to non-common variants, a linear-mixed-model-based approach called Genomic-Relatedness-based Restricted Maximum-Likelihood (GREML) was proposed (Yang *et al.*, 2010). Using this approach, in many diseases and traits the proportion of heritability explained by all genotyped SNPs was found to be much larger than the proportion explained only by the identified common variants (Lee *et al.*, 2011; Yang *et al.*, 2010; Visscher *et al.*, 2012). For instance, GREML estimated that the genotyped SNPs altogether could explain 34% of the heritability for Crohn's disease (Golan *et al.*, 2014). Various improved estimation methods were subsequently proposed (Bulik-Sullivan *et al.*, 2015; Golan *et al.*, 2014; Speed *et al.*, 2012, 2017). Results based on these methods also demonstrated increased explainable heritability by using all SNPs. On the other hand, there is still a large gap between the pedigree-based heritability and the SNP-based heritability. To further explore the heritability explained by rare variants with large effects, an improved method based on GREML was proposed (Yang *et al.*, 2015). This method can estimate the heritability explained by imputed variants, which include a large percentage of rare variants. Applying this method in the study of genetic factors of height and BMI, the explainable heritability was found to be substantially increased by the rare variants.

Another explanation for missing heritability that has attracted much attention is the presence of epistatic genetic interactions (Zuk *et al.*, 2011), in which the joint effect of two or more genetic variants on disease susceptibility is significantly stronger than the expected total effect of the individual variants if they were independent (Cordell, 2009). These interactions were not thoroughly studied in early GWAS, which instead mainly focused on the effects of individual

*The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First authors.

†to whom correspondence should be addressed

genetic variants. In order to identify epistatic interactions, the statistical significance of many combinations of genetic variants needs to be determined. This is both statistically and computationally challenging, since it is common to investigate millions of genetic variants in a single study, which lead to trillions of variant pairs, not to mention higher-order groups of more than two variants. In the past few years, many methodological advancements have been made to enhance the ability of detecting these epistatic interactions (Steen, 2012), including the invention of statistical models (Purcell *et al.*, 2007; Zhang and Liu, 2007), pre-selection of variants with potential interactions (Emily *et al.*, 2009), pre-grouping of variants (Zhang *et al.*, 2014), sampling of variants (Prabhu and Pe'er, 2012), better computational algorithms (Wan *et al.*, 2010b), and the use of computer hardware to accelerate the calculations (Hu *et al.*, 2010; Kam-Thong *et al.*, 2011; Yung *et al.*, 2011).

Using these methods, previous studies have identified various epistatic interactions that are statistically significant in explaining disease susceptibility. However, the extent to which missing heritability can be explained by epistatic interactions remains unclear (Yang *et al.*, 2017). Whether epistatic interactions represent a general phenomenon in human with biological importance also remain controversial (Aschard *et al.*, 2012; Hemani *et al.*, 2014a; Wood *et al.*, 2014; Hemani *et al.*, 2014b).

Some methods have used existing biological knowledge in the discovery of epistatic interactions. Most notably are methods that use functional pathways and networks to pre-select SNPs that could be interacting epistatically (Sun *et al.*, 2014; Wei *et al.*, 2014). For example, a framework was proposed (Liu *et al.*, 2012) to generate potential interacting SNP pairs based on functional data such as KEGG (Kanehisa *et al.*, 2017) and STRING (Szklarczyk *et al.*, 2015). The Biofilter (Bush *et al.*, 2009; Pendergrass *et al.*, 2013) pipeline was proposed to integrate multiple pathway and interaction network databases to build SNP-SNP interaction models. These knowledge-driven filtering methods assume that epistatic SNP-SNP interactions are correlated with functional interactions of the corresponding affected genes, yet none of these studies has systematically proved the presence of such correlations.

The lack of systematic investigation of the functional basis of epistatic interactions in human diseases was due to the absence of a comprehensive list of such interactions from published studies. The fact that a genetic variant does not always affect its closest gene also adds uncertainty to gene-based functional analysis methods.

The functional basis of epistatic interactions has been much more systematically studied in the baker's yeast *Saccharomyces cerevisiae* (Dixon *et al.*, 2009). There are high-throughput methods that study the growth rate of yeast cells with a large number of double knock-outs of two genes, as compared to the corresponding growth rates of the two single knock-outs (Tong *et al.*, 2004; Pan *et al.*, 2006; Decourty *et al.*, 2008). The data produced have helped formulate two main theories underlying negative genetic interactions (i.e., double knock-outs with a more severe phenotype than the expectation of the two single knock-outs), namely the between-pathway and within-pathway theories (Kelley and Ideker, 2005; Boone *et al.*, 2007). In the between-pathway theory, there are two pathways that perform similar or complementary functions. If genes (and consequently their gene products) in only one pathway are defective, the damage to the cell is tolerable since the other pathway is still intact. On the other hand, if genes in both pathways are defective, the resulting damage would be much more severe, leading to epistatic interactions between genes from the two pathways. In the within-pathway theory, mutations that affect a single gene in a pathway or protein complex can be tolerated, but if multiple genes are affected, the whole pathway/complex may break down, resulting in a much more serious phenotype.

Here we test if the within-pathway theory can also be applied to explain statistically significant epistatic interactions associated with human diseases. We present a list of published epistatic interactions between single nucleotide polymorphisms (SNPs) in various diseases from an extensive literature survey. To our knowledge, this is the first comprehensive list of SNP-SNP interactions in human diseases. In order to study the functional basis of these interactions, we associated the SNPs in these interactions with corresponding genes they likely affect. We used a variety of association methods to ensure robustness of our results. We also removed gene pairs close to each other on the primary genomic sequence, in order to eliminate potential effects caused by genetic linkage (Hemani *et al.*, 2014a; Wood *et al.*, 2014; Hemani *et al.*, 2014b). We then explored various functional relationships between the two genes in each resulting pair, including protein-protein interactions, co-expression, and co-occurrence in annotated biological pathways. Furthermore, we describe an algorithm for identifying additional genes that may be involved in the disease pathways from the combined epistatic and functional interaction network. Finally, we discuss several biologically interesting cases discovered by this algorithm that are well-supported by the literature.

2 METHODS

2.1 Compilation of the list of epistatic interactions

We used PubMed to search for research articles that describe epistatic SNP-SNP interactions as follows. We used "epistasis" and "SNP-SNP interaction" as keywords for the search, restricting the results to "human" for the species. From the results, we selected around 1,000 papers for manual checking (Supplementary File 4). Specifically, we first scanned the paper titles to identify the ones that likely report SNP-SNP interactions, such as those containing the keywords "epistasis", "gene-gene interaction", "SNP-SNP interaction" or "association studies". After this quick filtering, 310 papers remained. For these potentially relevant papers, we then scanned the main text to look for SNP-SNP interactions, based on various exclusion criteria such as containing only simulated data or non-human disease studies (Supplementary File 4). From the extracted SNP-SNP interaction pairs, we further filtered out pairs within 1Mbp from each other, which is a stringent criterion for eliminating possible effects of genetic linkage. For each resulting pair of SNP-SNP interactions, we recorded the associated diseases/phenotypes, the computational methods used for identifying them, and measures of their statistical significance (Supplementary File 1).

2.2 Associating the SNPs to potentially affected genes

Since most functional data are gene-centric, it is much more feasible to study the functional basis of SNP-SNP interactions by associating each SNP with the genes that it likely affects. Currently there is not a consensus as to the best way to perform such associations, but if a SNP overlaps a gene or is close to it, it is reasonable to assume that the gene could be affected by the SNP (Petersen *et al.*, 2013). We therefore associated a SNP to a gene by its genomic proximity using several different methods previously considered in the literature to ensure the robustness of our results. Specifically, we assigned a SNP to 1) the closest gene, 2) all genes within 2kbp, 3) 10kbp or 4) 25kbp from it, and 5) all genes within the same linkage disequilibrium (LD) block as the SNP. The LD blocks were downloaded from DistiLD (Pallejà *et al.*, 2012), which were defined for the hg19 human genome. For the other four methods, we performed the associations using both hg19 and hg38 human

reference genomes to evaluate the influence of the choice of the reference. As a result, we had 9 sets of SNP-to-gene associations. The genes considered were taken from Gencode (Harrow *et al.*, 2012) (v19 for hg19 and v21 for hg38) protein-coding genes. For each of the above methods, if in a SNP-SNP interaction at least one of the two SNPs could not be associated with a gene, the pair was removed from our list. The final result is a list of gene pairs which we will refer to as the list of gene-gene epistatic interactions for each SNP-to-gene association method.

2.3 Collection of biological networks

To study the functional relationships between the genes on our epistatic interaction lists, we collected three types of biological networks, namely protein-protein interactions (PPI), co-expression, and annotated pathways. We collected all human PPIs in the Human Protein Reference Database (HPRD) (Prasad *et al.*, 2009) and Reactome (Croft *et al.*, 2014), and considered each PPI as an unweighted, undirected edge in the network. For co-expression, we obtained the mutual ranks of co-expression values for each gene pair from the COXPRESdb (Obayashi *et al.*, 2013), and considered each pair as an undirected edge weighted by the mutual rank in the co-expression network. Finally, we downloaded annotated pathways included in Gene Set Enrichment Analysis (GSEA) (Mootha *et al.*, 2003; Subramanian *et al.*, 2005), including the Gene Ontology terms of molecular functions and biological processes (The Gene Ontology Consortium, 2015), BioCarta gene sets (Nishimura, 2001), Kyoto Encyclopedia of Genes and Genomes (KEGG) gene sets (Kanehisa *et al.*, 2017) and canonical pathways. Canonical pathways include the Sigma Aldrich pathways (Merck, 2017), Signaling Transduction KE pathways (Iyengar, 2003), SuperArray pathways (Burkhalter *et al.*, 2011), Signaling Gateway pathways (Li *et al.*, 2003) and Pathway Interaction Database (Schaefer *et al.*, 2009). Each pathway contained a set of directed unweighted edges with their meanings depending on the corresponding pathways.

For each network, we standardized the gene names based on the HGNC database of human gene names (Gray *et al.*, 2013).

2.4 Studying the functional relationships between the genes in the gene-gene epistatic interactions

We used two different methods to study the functional relationships between the genes in the gene-gene epistatic interactions, namely i) statistical testing, and ii) network neighborhood search (Figure 1).

2.4.1 Statistical testing We performed four sets of statistical tests to see whether the two genes in epistatic interaction pairs are, compared to random gene pairs, significantly:

1. More often connected in the PPI network (PPI-connectedness test)
2. Closer to each other in the PPI network (PPI-distance test)
3. More co-expressed in the co-expression network (co-expression test)
4. More often in the same biological pathway (same-pathway test)

We performed these tests by comparing the gene pairs on the list of epistatic interactions with 100,000 other random gene pairs (Figure 1a), and repeated it 10 times to ensure robustness of the results. Since the gene pairs on the list of epistatic interactions were formed by the corresponding associated SNP pairs, a gene would be more likely to be on the list by chance if it contains more SNPs. Correspondingly, in the random gene pairs, each gene was sampled with a probability proportional to the number of SNPs associated to it by the association method considered. In addition, as with the gene pairs on the epistatic interaction list, we also required each random gene pair to be formed by a random SNP pair at least 100Mb apart.

For the PPI-connectedness test, we encoded each gene pair with value 1 if the two genes were connected in the PPI network, and with value 0 if they were not connected. The number of gene pairs having these two values for the epistatic interactions and for the random gene pairs thus formed a

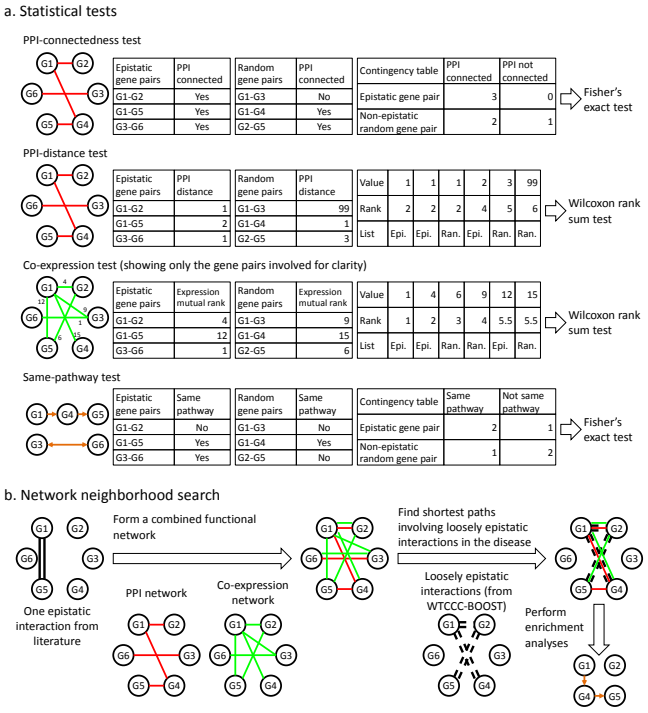


Fig. 1: Methods used for studying the functional relationships between genes in the epistatic interactions. Abbreviations: Epi. - epistatic interaction gene pair; Ran. - random gene pair

2x2 contingency table. We then used a one-tailed Fisher exact test to compute the p-value that the gene pairs on the list of epistatic interactions were significantly more connected than the random gene pairs.

For the PPI-distance test, we encoded each pair of genes with their shortest-path distance in the network (for two genes are not connected in the network, a maximum value larger than the longest path in the network was given). This procedure produced two vectors of distance values, one for the epistatic interactions, and one for the random gene pairs. We then used a one-tailed Wilcoxon rank-sum test to compute the p-value that the gene pairs on the list of epistatic interactions were significantly closer in the PPI network than the random gene pairs.

For the co-expression test, we used a procedure similar to the one for the PPI-distance test, to compute the p-value that the gene pairs on the list of epistatic interactions had significantly higher mutual co-expression ranks than the random gene pairs.

Finally, for the same-pathway test, we used a procedure similar to the one for the PPI-connectedness test, to compute the p-value that the gene pairs on the list of epistatic interactions were significantly more often to co-occur in at least one annotated pathway than random gene pairs.

To ensure the generality of our findings, we further repeated each set of tests two times, once with the genes in the random gene pairs sampled from the whole set of genes, and once with these genes sampled from only genes that have at least one interaction in the corresponding network.

We also used a permutation-based approach to performing these four tests. The details are provided in the Supplementary Materials.

2.4.2 Network neighborhood search In the Results section, we will show that most of the results of the above statistical tests were highly significant, suggesting that the gene pairs on the list of epistatic interactions are functionally related in the three types of biological networks. Since existing

biological networks are incomplete and mostly static (i.e., not containing context-specific information), we wondered whether integrating the information about epistatic interactions and functional interactions would be useful in identifying disease-related pathways.

To explore this possibility (Figure 1b), we collected GWAS data from the Wellcome Trust Case-Control Consortium (WTCCC) study of five common diseases/phenotypes (Crohn's disease, hypertension, rheumatoid arthritis, type 1 diabetes mellitus, and type 2 diabetes mellitus) with 14,000 cases and 3,000 shared controls (The Wellcome Trust Case Control Consortium, 2007). We used BOOST (Wan *et al.*, 2010a) to perform an all-against-all calculation, to compute the p-value for each pair of genetic variants to have an epistatic interaction associated with the disease/phenotype. Applying a loose threshold ($p<4.89E-6$, corresponding to a chi-square value > 30) to this full list, we obtained a set of loosely significant epistatic interactions. We associated these SNPs with genes they likely affect in the same ways as described above, leading to a network of genes with loosely epistatic interactions for the disease phenotype. Since the SNP pairs only have weakly significant p-values, only a fraction of them are expected to play crucial roles in the diseases/phenotypes.

We next formed a combined functional network consisting of all the edges in the PPI and co-expression (binarized based on Pearson correlation cutoff of 0.5) networks, while the annotated pathways were excluded for validating the results. Next, for each pair of genes on our list of epistatic interactions for these 5 diseases/phenotypes from the literature survey, we identified all shortest paths between the two genes in the combined functional network with the direct edge between the two genes excluded, and then retained only genes on these shortest paths with a loosely epistatic interaction with at least one other gene on these paths. As a result, for each initial gene pair on our list of epistatic interactions, we obtained a cluster of genes that were densely connected with each other in terms of epistatic interactions and functional (PPI and co-expression) interactions. We expect the functional data to help identify the subset of loosely epistatic interactions most relevant to the diseases/phenotypes.

We then performed an enrichment analysis of each cluster to check whether the genes in the cluster were enriched in annotated biological pathways. Specifically, for each cluster and each pathway, we formed a 2x2 contingency table for the genes in the cluster or not, and in the pathway or not, where the background set contains all genes contained in the PPI network, co-expression network or GSEA pathways. Based on this contingency table, we computed a corrected chi-square statistic (Huang *et al.*, 2008) and the corresponding p-value. Among the statistically significant cases, we only considered the ones with a significant enrichment of cluster genes in the pathway, but not the significantly depleted cases. These raw p-values were then corrected by Bonferroni correction, based on the total number of GSEA pathway terms (2,451) and the number of clusters identified based on the respective SNP-gene association method with at least 3 genes.

2.5 Testing on an RNAi data set

To further evaluate the generality of the functional relationships between genes with epistatic interactions, we considered an RNAi data set for studying epistasis among cancer genes (Wang *et al.*, 2014). The data set contained 847 gene pairs with significant epistatic interactions among 1508 pairs tested with combinatorial RNAi screening (with successful experiments from an original list of 66×29=1914 pairs). We used these 847 pairs as positive and the 1508-847=661 pairs as negative to perform the four types of statistical tests to see if the positive pairs are significantly more related by the functional relationships.

3 RESULTS

3.1 List of gene-gene epistatic interactions in human diseases/phenotypes

Based on our literature survey, we identified 83 to 2,449 gene-gene epistatic interactions in human diseases/phenotypes depending on the way of associating SNPs to genes (Table 1, Supplementary File 1). Most of the interactions are originated from SNP-SNP interactions between SNPs in the dbSNP database (Sherry *et al.*, 2011). The remaining cases involve particular alleles/genotypes of the genes, or only the interacting genes with no information of the genetic variants. Most of the gene pairs involve genes from different chromosomes (95-99%), while all the other pairs have the two genes at least 1Mbp apart from each other.

Table 1. Number of gene-gene epistatic interactions in human diseases and disease phenotypes based on our literature survey using different methods for associating SNPs to genes. Abbreviations: diff. - different; chr. - chromosome; CD - Crohn's Disease; HT - Hypertension; RA - Rheumatoid Arthritis; T1DM - Type 1 Diabetes Mellitus; T2DM - Type 2 Diabetes Mellitus

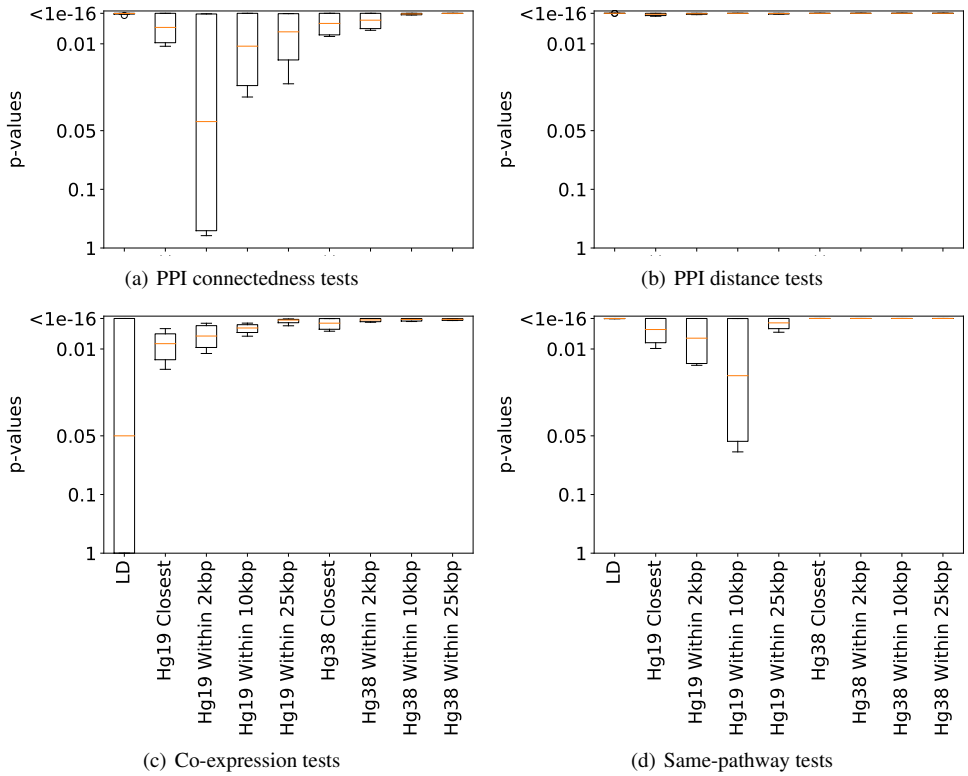
Reference genome	SNP-gene association	Number of gene pairs on epistatic interaction list							
		Total	Diff. chr.	Same chr., >1Mbp apart	CD	HT	RA	T1DM	T2DM
hg19	LD	2,449	2,411	38	62	1	51	0	77
hg19	Closest	104	102	2	5	1	3	2	7
hg19	Within 2kbp	83	80	3	4	1	1	2	7
hg19	Within 10kbp	121	118	3	5	1	5	2	13
hg19	Within 25kbp	255	252	3	7	1	12	12	16
hg38	Closest	107	104	3	5	1	3	3	7
hg38	Within 2kbp	85	81	4	4	1	1	3	7
hg38	Within 10kbp	132	126	6	5	1	4	5	13
hg38	Within 25kbp	270	262	8	7	1	11	17	16

3.2 Genes with epistatic interactions in human diseases/phenotypes are functionally related in various ways

We then performed the four types of statistical tests on the gene pairs with epistatic interactions. With the 9 SNP-gene association methods, 2 ways to sample random gene pairs (considering all genes or only genes with interactions in the functional network), and 10 sets of random gene pairs, each test resulted in 180 p-values. Based on the distributions of these p-values, we found that genes with epistatic interactions were functionally related in various ways (Figure 2).

We found that genes with epistatic interactions are significantly more connected and closer in the PPI network (Figures 2a,b). For the PPI connectedness tests, most p-values were smaller than 0.01 except when SNPs were associated with all genes within 2kb based on hg19. This small distance threshold caused many SNPs to be not associated to any genes and were thus excluded from the statistical tests, leading to insignificant p-values. Interestingly, the p-values were generally more significant with the hg38 reference than with hg19, suggesting that updates to the reference genome also improved this functional analysis. For the PPI distance tests, all p-values were highly significant regardless of the setting, demonstrating the reliability of the results.

For the co-expression tests (Figure 2c), the p-values were less than 0.01 in over 90% of the cases. We observed an issue with the LD block-based SNP-gene association, that in one single LD block there could be many genes. In one extreme case, there was a



(a) PPI connectedness tests (b) PPI distance tests

(c) Co-expression tests (d) Same-pathway tests

Fig. 2: Box plots of the p-values obtained from (a) the PPI connectedness tests, (b) PPI distance tests, (c) co-expression tests, and (d) same-pathway tests.

large LD block over 7Mb in size on chromosome 6 containing 650 genes. An epistatic SNP-SNP interaction involving a SNP in this LD block led to a large number of corresponding gene-gene pairs, many of which are not expected to have real epistatic interactions. Other than this setting, the p-values were in general significant in all other settings.

For the same-pathway tests (Figure 2d), again most of the p-values were highly significant, with only some insignificant p-values when SNPs were associated with nearby genes based on the old hg19 reference.

To further ensure the robustness of our results, we also used a permutation-based approach to perform the above four types of statistical tests (Supplementary methods). The results (Figure S1) confirmed statistical significance of the PPI connectedness, co-expression and pathway co-occurrence of the genes with epistatic interactions. This conclusion was generally robust with largely stable results across different SNP-gene association methods and the way of sampling random gene pairs. On the other hand, the PPI distances of gene pairs with epistatic interactions were only significantly smaller than background gene pairs in some settings but not in some others.

We also performed these four tests for the cancer genes with epistatic interactions, using the remaining gene pairs among these genes tested in the combinatorial RNAi experiments as the negative set. The results showed that the gene pairs having epistatic interactions were significantly more connected and closer in the PPI network, and more often co-occurring in the same pathways, with p-values of 1.2E-8, 5.2E-8 and 2.8E-7, respectively. On the other hand, the epistatically interacting gene pairs were only marginally more co-expressed than the negative pairs, with a p-value of 0.21.

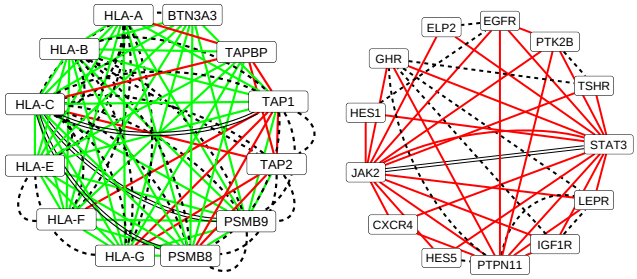
Taking all the results together, genes with epistatic interactions are generally related in terms of protein-protein interactions, biological pathways and gene expression, although the level of significance varies among data sets, statistical testing methods and testing configurations.

3.3 Identifying disease-related pathways by neighborhood searching in the combined epistatic-functional network

Next we applied the neighborhood searching method to identify potential disease-related pathways from the combined epistatic-functional network. The full list of results is provided in Supplementary File 2. For each cluster, we performed an enrichment analysis to check if the genes in the clusters were enriched in certain biological pathways. Interestingly, although the information of these pathways was not used in the neighborhood search, many clusters exhibited significant enrichment of the pathways (Supplementary File 3), confirming that the neighborhood searching method was able to identify biological pathways based on the epistatic and functional interactions. Although this result is not surprising since we

have used both an epistatic interaction reported to be strongly related to the disease as well as a set of loosely epistatic interactions as input, the strong relationship between the identified gene clusters and the diseases suggest that the loosely epistatic interactions included in the cluster are the ones more relevant to the diseases. Furthermore, we found some terms that were enriched only when we restricted the genes to those having loosely epistatic interactions with each other. For example, the cluster of genes identified from the JAK2-STAT3 pair (more details below) was enriched in the KEGG pathway "hsa04920:Adipocytokine signaling pathway" only when this restriction was applied, showing that the loosely epistatic interactions contain some supplementary information not fully contained in the functional interactions.

Here we show two interesting gene clusters identified that have strong supports from the literature (Figure 3). A third cluster is discussed in the Supplementary Materials due to space limit.



(a) (b)

Fig. 3: Results of neighborhood searching from the combined epistatic-functional network, based on the epistatic interactions (a) between HLA-C and PSMB8/PSMB9/TAP1 in type 1 diabetes mellitus, with SNPs associated with genes within 10kb using hg19 reference, and (b) between JAK2 and STAT3 in Crohn's disease, with SNPs associated with genes within the same LD block. Double black lines indicate literature-reported epistatic interactions, black dotted lines indicate WTCCC-BOOST loosely epistatic interactions, red lines indicate PPIs, and green lines indicate co-expression.

Figure 3a shows a cluster identified from type 1 diabetes mellitus (T1DM), an autoimmune disease marked by the destruction of insulin-producing β-cells in the pancreatic islets. This example involves the epistatic interactions between HLA-C (due to the SNPs rs2524089 and rs2524095) and PSMB8 (previously called LMP7), PSMB9 (previously called LMP2) and TAP1 (due to the SNPs rs9276815, rs9276825 and rs9276832) (Wan *et al.*, 2010a). These genes are linked to each other and to the other genes in the cluster with loosely epistatic interactions, PPI and co-expression. Most of these genes have been individually reported to be associated with T1DM risk (Sia and Weinem, 2005; Noble *et al.*, 2010). The genes in this cluster are enriched in many pathways, such as antigen processing and presentation, interferon signaling and endocytosis, all with Bonferroni corrected p-values $<$1.8E-10.

The high density of epistatic and functional interactions between these genes suggests that they belong to a pathway highly relevant to T1DM. Indeed, these genes encode proteins that are part of the MHC-I antigen processing and presentation pathway, a process critical for the activation of CD8 T cell-mediated adaptive immune responses. Autoreactive CD8 T cells are key players in the killing of pancreatic β-cells, resulting in autoimmune diabetes. Among the genes we identified, HLA-A, -B, -C, -E, -F and -G are gene paralogues encoding for the MHC-I heavy chain, which forms part of the antigen presentation complex displayed on the surface of most cells. MHC-I molecules bind peptide antigens generated by protein degradation in the proteasome. The β-subunits of the immunoproteasome are encoded by two genes in the cluster, PSMB8 and PSMB9. Peptide antigens are transported from the cytosol to ER by the transporter associated with antigen processing 1 (TAP1) and 2 (TAP2), which form part of the MHC-I peptide-loading complex. Thus, each of the functions described above (proteasomal activity, antigen processing and antigen presentation) is aided by the expression of two or more genes. If one of the genes has a mutation, another gene with a similar function may compensate for the mutated gene in order to maintain normal functions, as has been described in genetic knockout mice lacking PSMB8 alone, PSMB9 alone, or both (Kincaid *et al.*, 2012). Therefore, having one gene mutated may have a minimal effect on the overall MHC-I antigen presentation pathway, but if multiple functions are altered by genetic mutations, the net effect is expected to be more severe across the whole pathway, which may explain the epistatic interactions between HLA-C and TAP1.

Another gene in the cluster, BTN3A3, also called CD277 is a member of the butyrophilin (BTN) family. The functions of proteins encoded by the BTN gene cluster are not well understood although polymorphisms in the BTN-gene cluster have been reported to associate with susceptibility to T1DM (Viken *et al.*, 2009). BTN3A3 has a closely related isoform, BTN3A1, which is known to bind and present pyrophosphate antigens to $\gamma\delta$ T cells (Vavassori *et al.*, 2013), suggesting that BTN3A proteins are functionally important in antigen presentation. Based on our results, it will be interesting to investigate the link between BTN3A3 and the MHC-I antigen presentation pathway in the development of T1DM. Thus, using a combined epistatic functional network approach, our analyses provide evidence supporting the key role of the MHC-I antigen processing and presentation pathway in conferring susceptibility to T1DM (Sia and Weinem, 2005).

Figure 3b shows a cluster identified from an epistatic interaction in Crohn's disease (CD) between JAK2 (due to the SNP rs10758669) and STAT3 (due to the SNP rs744166) (Polgar *et al.*, 2012). CD is a sub-form of inflammatory bowel diseases (IBD) that result in chronic inflammation of the gastrointestinal tract. The cluster involves 11 other genes that form loosely epistatic interactions and protein-protein interactions with JAK2 and STAT3. As expected, many of the genes in this cluster are in the JAK-STAT signaling pathway (Bonferroni corrected p-value $<$ 1.7E-9). STAT3 belongs to the STAT family of transcription factors activated by engagement of growth factors, interferons or cytokines on cell surface receptors. Receptor engagement activates the JAK family of receptor-associated tyrosine kinases, including JAK2, leading to the recruitment, activation and translocation of STAT3 to the nucleus to regulate target gene transcription.

The JAK-STAT pathway is essential for the differentiation of T helper 17 (Th17) cells and the suppressive functions of regulatory T cells, which are key players in the pathogenesis of CD (Chaudhry *et al.*, 2009; Patel and Kuchroo, 2015). Many genes in this cluster are involved in receptor-mediated activation of the JAK/STAT

pathway. For example, CXCR4 is a chemokine receptor found on both T cells and intestinal epithelial cells. CXCR4 binds to CXCL12 and signals the activation of JAK2 and STAT3 (Ahr *et al.*, 2005). CXCR4 is more highly expressed in patients with IBD (Werner *et al.*, 2011), suggesting its involvement in the pathogenesis and progression of the disease (Mrowicki *et al.*, 2014). Another gene in the cluster, PTPN11 encodes for the protein tyrosine phosphatase SHP2, which mediates tyrosine dephosphorylation of JAK2 to control the activity of the JAK/STAT pathway (Xu and Qu, 2008). Genetic mutations in PTPN11 are associated with increased susceptibility to IBD in animal experiments (Coulombe *et al.*, 2013) and human studies (Marcil *et al.*, 2013; The Wellcome Trust Case Control Consortium, 2007).

Other genes in the cluster encode for the epidermal growth factor receptor (EGFR) receptor, insulin-like growth factor I receptor (IGF-IR) and growth hormone receptor (GHR), which have been shown to trigger JAK/STAT activation upon receptor engagement (Sugimoto, 2008; Wieduwilt and Moasser, 2008; Zong *et al.*, 2000). These growth factor receptors and their ligands are being investigated as potential therapeutic targets for IBD because of their roles in mediating signals involving mucosal repair and intestinal inflammation (Barahona-Garrido *et al.*, 2009). Taken together, we identified a pathway involved in the activation of the JAK/STAT signaling, which helps explain the epistatic interaction between JAK2 and STAT3 in CD.

4 DISCUSSION

In this paper, we have demonstrated that epistatic interactions in human diseases can be studied using biological networks. By associating the SNP-SNP epistatic interactions to corresponding genes, we have shown that these genes are significantly more connected to each other in the protein-protein interaction network, more co-expressed, and more often appear in the same annotated pathways. These genes are also significantly closer to each other in the protein-protein interaction network in some settings, although the results are less significant in other settings. Based on these initial findings, we have further demonstrated that the neighborhoods around these genes in the combined epistatic-functional network can be used to identify disease pathways.

The list of epistatic interactions we compiled from an extensive review of research articles serves as a resource for studying epistatic interactions in human diseases. We have overcome the issue of associating SNPs with genes by using a variety of association methods and showing that the results are largely immune to the choice of method. We provide all these association results for anyone interested in studying epistatic interactions to choose the most suitable set based on the research problem.

Our current list of epistatic interactions includes SNP pairs identified by a variety of methods. Since the list is not very long, considering potential issues with statistical power we did not separately analyze the subsets produced by different methods. These method details are provided in Supplementary File 4, which can be used for extracting any subset of particular interests in future studies.

Our list of epistatic SNP-SNP interactions will need to be updated as more cases are published. Our literature survey procedures, especially the paper exclusion criteria, serve as guidelines for these future updates.

ACKNOWLEDGEMENT

We thank Yingying Wei for helpful discussions.

Funding: NLST, WY and KYY are partially supported by the HKRGC Theme-based Research Scheme T12-402/13N.

REFERENCES

Ahr, B. et al (2005). Identification of the cytoplasmic domains of CXCR4 involved in jak2 and STAT3 phosphorylation. *Journal of Biological Chemistry*, **280**, 6692–6700.

Aschard, H. et al (2012). Inclusion of gene-gene and gene-environment interactions unlikely to dramatically improve risk prediction for complex diseases. *The American Journal of Human Genetics*, **90**, 962–972.

Barahona-Garrido, J. et al (2009). Growth factors as treatment for inflammatory bowel disease: A concise review of the evidence toward their potential clinical utility. *The Saudi Journal of Gastroenterology*, **15**, 208–212.

Boone, C., Bussey, H. and Andrews, B.J. (2007). Exploring genetic interactions and networks with yeast. *Nature Reviews Genetics*, **8**(6), 437–449.

Bulik-Sullivan, B.K. et al (2015). LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature Genetics*, **47**, 291–295.

Burkhalter, R.J. et al (2011). Integrin regulation of β-catenin signaling in ovarian carcinoma. *Journal of Biological Chemistry*, **286**(26), 23467–23475.

Bush, W.S., Dudek, S.M. and Ritchie, M.D. (2009). Biofilter: A knowledge-integration system for the multi-locus analysis of genome-wide association studies. In *Pacific Symposium of Biocomputing*, pages 368–379.

Chaudhry, A. et al (2009). CD4+ regulatory T cells control TH17 responses in a stat3-dependent manner. *Science*, **326**, 986–991.

Cordell, H.J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Review Genetics*, **10**(6), 392–404.

Coulombe, G. et al (2013). Epithelial tyrosine phosphatase SHP-2 protects against intestinal inflammation in mice. *Molecular and Cellular Biology*, **33**(11), 2275–2284.

Croft, D. et al (2014). The reactome pathway knowledgebase. *Nucleic Acids Research*, **42**, D472–D477.

Decourty, L. et al (2008). Linking functionally related genes by sensitive and quantitative characterization of genetic interaction profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(15), 5821–5826.

Dixon, S.J. et al (2009). Systematic mapping of genetic interaction networks. *Annual Review of Genetics*, **43**, 601–625.

Easton, D.F. et al (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, **447**(7148), 1087–1093.

Eichler, E.E. et al (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, **11**(6), 446–450.

Emily, M. et al (2009). Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*, **17**(10), 1231–1240.

Fellay, J. et al (2007). A whole-genome association study of major determinants for host control of HIV-1. *Science*, **317**(5840), 944–947.

Frayling, T.M. et al (2007). A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, **316**(5826), 889–894.

Fuchsberger, C. et al (2016). The genetic architecture of type 2 diabetes. *Nature*, **536**(7614), 41–47.

Gibson, G. (2012). Rare and common variants: Twenty arguments. *Nature Reviews Genetics*, **13**, 135–145.

Golan, D., Lander, E.S. and Rosset, S. (2014). Measuring missing heritability: Inferring the contribution of common variants. *Proceedings of the National Academy of Sciences of the United States of America*, **111**(49), E5272–E5281.

Gray, K.A. et al (2013). Genenames.org: The HGNC resources in 2013. *Nucleic Acids Research*, **41**, D545–D552.

Harrow, J. et al (2012). GENCODE: The reference human genome annotation for the ENCODE project. *Genome Research*, **22**, 1760–1774.

Hemani, G. et al (2014a). Detection and replication of epistasis influencing transcription in humans. *Nature*, **508**(7495), 249–253.

Hemani, G. et al (2014b). Hemani et al. reply. *Nature*, **514**(7520), E5–E6.

Hu, X. et al (2010). SHEsisEpi, a GPU-enhanced genome-wide SNP-SNP interaction scanning algorithm, efficiently reveals the risk genetic epistasis in bipolar disorder. *Cell Research*, **20**(7), 854–857.

Huang, D.W., Sherman, B.T. and Lempicki, R.A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, **4**(1), 44–57.

Iyengar, R. (2003). A composite schematic of gpcr signaling. *Science Signaling*.

Kam-Thong, T. et al (2011). EPIBLASTER-fast exhaustive two-locus epistasis detection strategy using graphical processing units. *European Journal of Human Genetics*, **19**(4), 465–471.

Kanehisa, M. et al (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, **45**, D353–D361.

Kelley, R. and Ideker, T. (2005). Systematic interpretation of genetic interactions using protein networks. *Nature Biotechnology*, **23**(5), 561–566.

Kincaid, E.Z. et al (2012). Mice completely lacking immunoproteasomes display major alterations in antigen presentation. *Nature Immunology*, **13**(2), 129–135.

Lander, E.S. (2011). Initial impact of the sequencing of the human genome. *Nature*, **470**(7333), 187–197.

Lee, S.H. et al (2011). Estimating missing heritability for disease from genome-wide association studies. *The American Journal of Human Genetics*, **88**, 294–305.

Li, Z. et al (2003). Directional sensing requires g$\beta\gamma$-mediated PAK1 and PIXα-dependent activation of Cdc42. *Cell*, **114**(2), 215–227.

Liu, Y. et al (2012). Gene, pathway and network frameworks to identify epistatic interactions of single nucleotide polymorphisms derived from GWAS data. *BMC Systems Biology*, **6**, S15.

Manolio, T.A. et al (2009). Finding the missing heritability of complex diseases. *Nature*, **461**(7265), 747–753.

Marcil, V. et al (2013). Association between the PTPN2 gene and crohn's disease: Dissection of potential causal variants. *Inflammatory Bowel Diseases*, **19**, 1149–1155.

Merck (2017). IUBMB-sigma-nicholson metabolic pathway charts.

Mootha, V.K. et al (2003). Pgc-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature Genetics*, **34**(3), 267–273.

Mrowicki, J. et al (2014). The role of polymorphisms of genes CXCL12/CXCR4 and MIF in the risk development IBD the polish population. *Molecular Biology Reports*, **41**, 4639–4652.

Nishimura, D. (2001). BioCarta. *Biotech Software & Internet Report*, **2**(3), 117–120.

Noble, J.A. et al (2010). HLA class i and genetic susceptibility to type 1 diabetes – results from the type 1 diabetes genetics consortium. *Diabetes*, **59**, 2972–2979.

Obayashi, T. et al (2013). COXPRESdb: A database of comparative gene coexpression networks of eleven species for mammals. *Nucleic Acids Research*, **41**, D1014–D1020.

Pallejà, A. et al (2012). DistiLD database: Diseases and traits in linkage disequilibrium blocks. *Nucleic Acids Research*, **40**, D1036–D1040.

Pan, X. et al (2006). A DNA integrity network in the yeast saccharomyces cerevisiae. *Cell*, **124**, 1069–1081.

Patel, D.D. and Kuchroo, V.K. (2015). Th17 cell pathway in human immunity: Lessons from genetics and therapeutic interventions. *Immunity*, **43**, 1040–1051.

Pendergrass, S.A. et al (2013). Genomic analyses with biofilter 2.0: Knowledge driven filtering, annotation, and model development. *BioData Mining*, **6**, 25.

Petersen, A. et al (2013). Assessing methods for assigning snps to genes in gene-based tests of association using common variants. *PLOS ONE*, **8**(5), e62161.

Plenge, R.M. et al (2007). TRAF1-C5 as a risk locus for rheumatoid arthritis - a genomewide study. *The New England Journal of Medicine*, **357**, 1199–1209.

Polgar, N. et al (2012). Investigation of JAK2, STAT3 and CCR6 polymorphisms and their genevgene interactions in inflammatory bowel disease. *International Journal of Immunogenetics*, **39**, 247–252.

Prabhu, S. and Pe'er, I. (2012). Ultrafast genome-wide scan for SNPvSNP interactions in common complex disease. *Genome Research*, **22**, 2230–2240.

Prasad, T.S.K. et al (2009). Human protein reference database - 2009 update. *Nucleic Acids Research*, **37**, D767–D772.

Purcell, S. et al (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, **81**, 559–575.

Schaefer, C.F. et al (2009). PID: the pathway interaction database. *Nucleic acids research*, **37**(suppl 1), D674–D679.

Sherry, S.T. et al (2011). dbSNP: The NCBI database of genetic variation. *Nucleic Acids Research*, **29**, 308–311.

Sia, C. and Weinem, M. (2005). Genetic susceptibility to type 1 diabetes in the intracellular pathway of antigen processing - a subject review and cross-study comparison. *The Review of Diabetic Studies*, **2**, 40–52.

Speed, D. et al (2012). Improved heritability estimation from genome-wide SNPs. *The American Journal of Human Genetics*, **91**, 1011–1021.

Speed, D. et al (2017). Reevaluation of SNP heritability in complex human traits. *Nature Genetics*, **49**, 986–992.

Steen, K.V. (2012). Travelling the world of gene-gene interactions. *Briefings in Bioinformatics*, **13**, 1–19.

Subramanian, A. et al (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**(43), 15545–15550.

Sugimoto, K. (2008). Role of STAT3 in inflammatory bowel disease. *World Journal of Gastroenterology*, **14**(33), 5110–5114.

Sun, S. et al (2014). Analysis pipeline for the epistasis search statistical versus biological filtering. *Frontiers in Genetics*, **5**, 350.

Szklarczyk, D. et al (2015). STRING v10: Proteinprotein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, **43**, D447–D452.

The Gene Ontology Consortium (2015). Gene ontology consortium: Going forward. *Nucleic Acids Research*, **43**, D1049–D1056.

The Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**(7145), 661–678.

Tong, A.H.Y. et al (2004). Global mapping of the yeast genetic interaction network. *Science*, **303**(5659), 808–813.

Vavassori, S. et al (2013). Butyrophilin 3A1 binds phosphorylated antigens and stimulates human $\gamma\delta$ t cells. *Nature Immunology*, **14**(9), 908–916.

Viken, M.K. et al (2009). Reproducible association with type 1 diabetes in the extended class i region of the major histocompatibility complex. *Genes and Imuunity*, **10**(4), 323–333.

Visscher, P.M. et al (2012). Five years of GWAS discovery. *The American Journal of Human Genetics*, **90**, 7–24.

Wan, X. et al (2010a). BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, **87**(3), 325–340.

Wan, X. et al (2010b). Detecting two-locus associations allowing for interactions in genome-wide association studies. *Bioinformatics*, **26**(20), 2517–2525.

Wang, X. et al (2014). Widespread genetic epistasis among cancer genes. *Nature Communications*, **5**, 4828.

Wei, W.H., Hemani, G. and Haley, C.S. (2014). Detecting epistasis in human complex traits. *Nature Reviews Genetics*, **15**, 722–733.

Werner, L. et al (2011). Reciprocal regulation of CXCR4 and CXCR7 in intestinal mucosal homeostasis and inflammatory bowel disease. *Journal of Leukocyte Biology*, **90**(3), 583–590.

Wieduwilt, M.J. and Moasser, M.M. (2008). The epidermal growth factor receptor family: Biology driving targeted therapeutics. *Cellular and Molecular Life Sciences*, **65**, 1566–1584.

Wood, A.R. et al (2014). Another explanation for apparent epistasis. *Nature*, **514**(7520), E3–E5.

Xu, D. and Qu, C.K. (2008). Protein tyrosine phosphatases in the JAK/STAT pathway. *Frontiers in Bioscience*, **13**, 4925–4932.

Yang, J. et al (2010). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, **42**, 565–569.

Yang, J. et al (2015). Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nature Genetics*, **47**, 1114–1120.

Yang, J. et al (2017). Concepts, estimation and interpretation of SNP-based heritability. *Nature Genetics*, **49**, 1304–1310.

Yung, L.S. et al (2011). GBOOST: A GPU-based tool for detecting genevgene interactions in genomevwide case control studies. *Bioinformatics*, **27**(9), 1309–1310.

Zhang, F., Boerwinkle, E. and Xiong, M. (2014). Epistasis analysis for quantitative traits by functional regression model. *Genome Research*, **24**, 989–998.

Zhang, Y. and Liu, J.S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, **39**(9), 1167–1173.

Zong, C.S. et al (2000). Mechanism of STAT3 activation by insulin-like growth factor i receptor. *Journal of Biological Chemistry*, **275**, 15099–15105.

Zuk, O. et al (2011). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(4), 1193–1198.

Supplementary methods

4.1 Permutation-based tests for studying the functional relationships between the genes in the gene-gene epistatic interactions

In addition to the statistical tests described in the main text, we also used a permutation-based approach to evaluate the statistical significance of the four functional relationships between genes in the gene-gene epistatic interactions.

For the PPI-connectedness test, we used the number of connected pairs in the PPI network as the test statistic. We first computed this number for the gene pairs with epistatic interactions in the actual PPI network. We then produced 10,000 permuted networks and computed the test statistic for each of them. The p-value was then defined as the fraction of permuted networks with a test statistic the same or larger than the number in the actual PPI network. The permuted networks were constructed as follows. Each time, we randomly selected two PPI pairs (p1, p2) and (p3, p4), and swapped their connections to become (p1, p3), (p2, p4) (Milo *et al.*, 2002). The degree distribution of the network was preserved by this operation. We used the first 100,000 times of edge swapping as burn-out to construct the first permuted network. We then continued the procedure and obtained one permuted network for every 10,000 additional edge swaps, and repeated it until getting 10,000 permuted networks.

For the PPI-distance test, we used the same procedure to produce permuted PPI networks, but used a different test statistic as follows. We first computed the network distance between the two genes in each epistatically interacting pair. If the two genes were not connected in the network, their distance was set to a value larger than the maximum distance of all gene pairs in the network. All these distance values were then sorted, and the mean of the 50% smallest distance values was used as the test statistic. We checked the results to confirm that in all our actual calculations, the test statistic never involved the artificial distance value assigned for the disconnected pairs. We chose this test statistic instead of the mean or median of all the distance values because the mean of all the distance values would depend on the fraction of disconnected pairs in the network, which was the subject of the PPI-connectedness test but not this PPI-distance test, while the median would give very similar values for all networks due to the small-world property of PPI networks.

For the co-expression test, we constructed permuted networks by re-assigning mutual ranks of co-expression of all gene pairs randomly. We constructed 10,000 permuted networks, and used the mean of the mutual rank among the epistatically interacting pairs as the test statistic.

For the same-pathway test, we constructed permuted pathways by assigning random genes to pathways, preserving the total number of pathways and the number of genes in each pathway. We constructed 10,000 permuted pathways and used the number of epistatically interacting gene pairs with the two genes co-occurring in at least one pathway as the test statistic.

As in the original tests described in the main text, we also tried the different ways to assign SNPs to genes, and repeated the tests considering either all human genes as the background or only the genes in the original networks as the background.

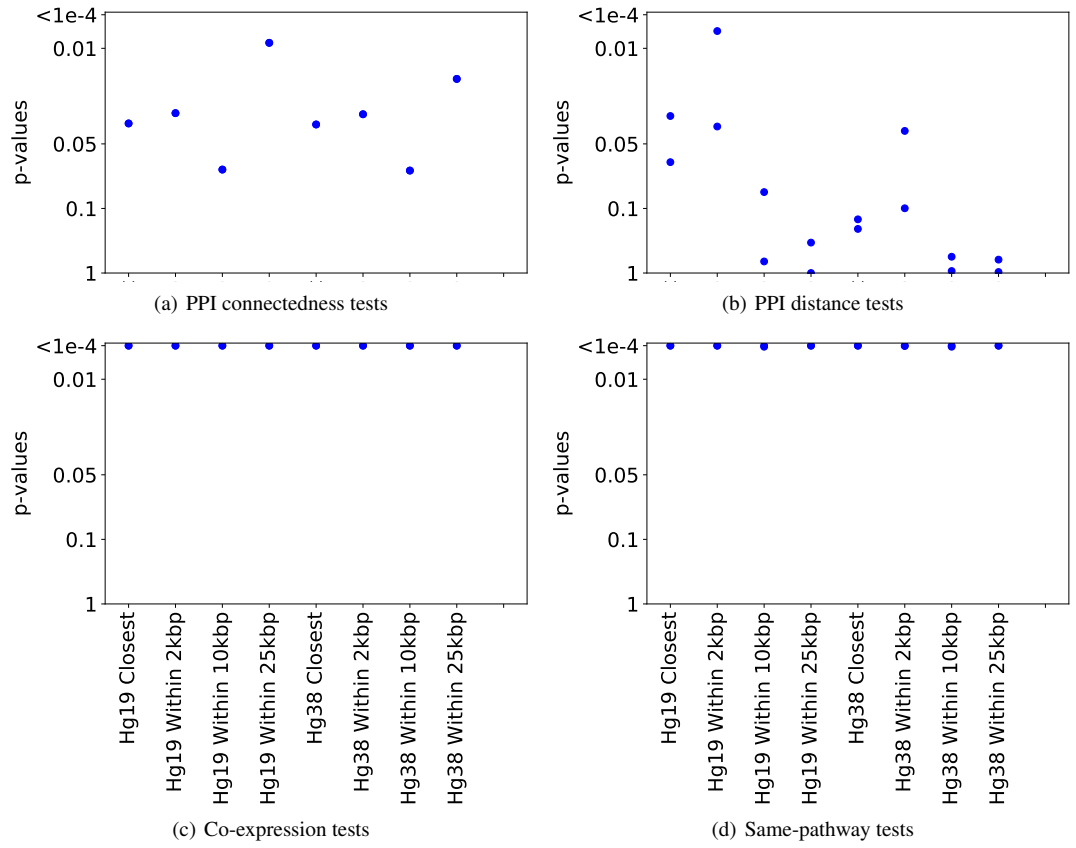
Supplementary results

The results of the permutation-based statistical tests are shown in Figure S1.

Epistatic and functional interactions between NOD2 and TLE1 (due to the SNP rs6559629) have been implicated in the pathogenesis of CD (Nimmo *et al.*, 2011), but the underlying mechanisms remain largely unknown. As shown in Figure S2, we found that NOD2 and TLE1 form a cluster with 21 other genes from different chromosomes and different LD blocks that are enriched in various pathways, including cytokine signaling in immune system (Bonferroni corrected p-value = 2.3E-5) and programmed cell death (Bonferroni corrected p-value $<$ 1E-14). NOD2 is a member of the nucleotide-binding oligomerization domain (NOD)-like receptor (NLR) family, which functions as intracellular microbial sensor of the innate immune system to regulate inflammation and cytokine production. TLE1 is a transcriptional co-repressor that binds to transcription factors to regulate a wide range of cellular processes including cell growth and differentiation (Ali *et al.*, 2010) and inflammatory signaling (Ramasamy *et al.*, 2016).

Our analysis suggests that interactions between NOD2 and TLE2 involve effector molecules of the TGF-β1-dependent pathway including the SMAD and SMURF family of proteins. SMAD3 phosphorylation, an important step in the initiation of TGF-β1-mediated suppression of intestinal inflammation, is significantly decreased in IBD patients compared to normal controls (Monteleone *et al.*, 2001). Controlling TGF-β1-associated and SMAD-associated signaling is a novel therapeutic strategy for the treatment of CD (Monteleone *et al.*, 2015). Engagement of the TGF-β1-signaling pathway by NOD2 may be mediated through physical interaction with ERBB2IP (Kufer *et al.*, 2006), an adaptor protein shown to inhibit TGF-β signaling by sequestering SMAD2/3 (Dai *et al.*, 2007). Activated SMAD2/3 form a multi-protein complex with SMAD4 and Runx3 to control the transcription of target genes in the nucleus (Chuang *et al.*, 2017). Runx3 is capable of mediating gene repression by recruiting the TLE-1 transcriptional co-repressor to target promoters (Yarmus *et al.*, 2006). Several Runx3 SNPs have been identified to associate with increased risks for CD (Weersma *et al.*, 2008; Yamazaki *et al.*, 2013). The Smad complexes can also bind to another transcriptional repressor, FoxG1, which in turns antagonizes the effect of TGF-β signaling (Seoane *et al.*, 2004). FoxG1 is also known to recruit co-repressors including TLE1 to mediate gene repression (Dastidar *et al.*, 2012).

In addition, NOD2 activation is linked to signaling through another innate receptor, TLR2, as both receptors can be activated by closely related structures derived from bacterial peptidoglycan (Inohara *et al.*, 2003). NOD2 exerts a synergistic or negative effect on TLR2 activation in human monocytes depending on the dose of ligand stimulation, but this regulatory mechanism is lost in NOD2-deficient patients with CD (Borm *et al.*, 2008). Moreover, dysregulated expression of TLR2 has been found in the intestinal mucosa of IBD patients (Frolova *et al.*, 2008). TLR2 can also be activated to promote inflammation in the presence of extracellular HMGB1, a nuclear protein secreted by stressed or dying cells in response to cellular damage (Park *et al.*, 2004). HMGB-1 is a robust biomarker for inflammatory bowel diseases that correlates with disease progression and therapeutic outcomes (Hu *et al.*, 2015).



(a) PPI connectedness tests (b) PPI distance tests

(c) Co-expression tests (d) Same-pathway tests

Fig. S1: Box plots of the p-values obtained from (a) the PPI connectedness tests, (b) PPI distance tests, (c) co-expression tests, and (d) same-pathway tests based on the permutation-based approach.

Another gene in the cluster encodes the NLRC4 protein, which belongs to the same NLR protein family as NOD2 and controls innate immune response to bacterial pathogens. NLRC4 forms a multimeric complex called the inflammasome in response to specific bacterial ligands, leading to activation of caspase-1 and production of pro-inflammatory cytokines, including IL-1β and IL-18. Caspase-1 activity, and the levels of IL-1β and IL-18 are elevated in clinical samples of IBD patients (McAlindon *et al.*, 1998; Siegmund, 2002), suggesting that targeting the NLR-mediated inflammatory pathway maybe a therapeutic option for IBD. Thus, our results suggest that epistatic interactions between NOD2 and TLE1 in CD may involve the counterbalance between NLR-mediated chronic inflammation in the gut and TGF-β/Smad signaling-dependent suppressive activities through transcriptional repression involving the transcriptional co-repressor, TLE1.

SUPPLEMENTARY REFERENCES

Ali, S.A. et al (2010). Transcriptional corepressor TLE1 functions with runx2 in epigenetic repression of ribosomal RNA genes. *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 4165–4169.

Borm, M.E.A. et al (2008). The effect of NOD2 activation on TLR2-mediated cytokine responses is dependent on activation dose and NOD2 genotype. *Genes and Immunity*, **9**(3), 274–278.

Chuang, L.S.H., Ito, K. and Ito, Y. (2017). *Roles of RUNX in Solid Tumors*, pages 299–320. Springer.

Dai, F. et al (2007). Erbin inhibits transforming growth factor β signaling through a novel smad-interacting domain. *Molecular and Cellular Biology*, **27**, 6183–6194.

Dastidar, S.G. et al (2012). Transducin-like enhancer of split-1 (TLE1) combines with forkhead box protein g1 (FoxG1) to promote neuronal survival. *Journal of Biological Chemistry*, **287**, 14749–14759.

Frolova, L. et al (2008). Expression of toll-like receptor 2 (TLR2), TLR4, and CD14 in biopsy samples of patients with inflammatory bowel diseases: Upregulated expression of TLR2 in terminal ileum of patients with ulcerative colitis. *Journal of Histochemistry and Cytochemistry*, **56**, 267–274.

Hu, Z. et al (2015). Role of high-mobility group box 1 protein in inflammatory bowel disease. *Inflammation Research*, **64**, 557–563.

Inohara, N. et al (2003). Host recognition of bacterial muramyl dipeptide mediated through NOD2. implications for crohn's disease. *Journal of Biological Chemistry*, **278**, 5509–5512.

Kufer, T.A. et al (2006). Role for erbin in bacterial activation of nod2. *Infection and Immunity*, **74**, 3115–3124.

McAlindon, M.E., Hawkey, C.J. and Mahida, Y.R. (1998). Expression of interleukin 1β and interleukin 1β converting enzyme by intestinal macrophages in health and inflammatory bowel disease. *Gut*, **42**, 214–219.

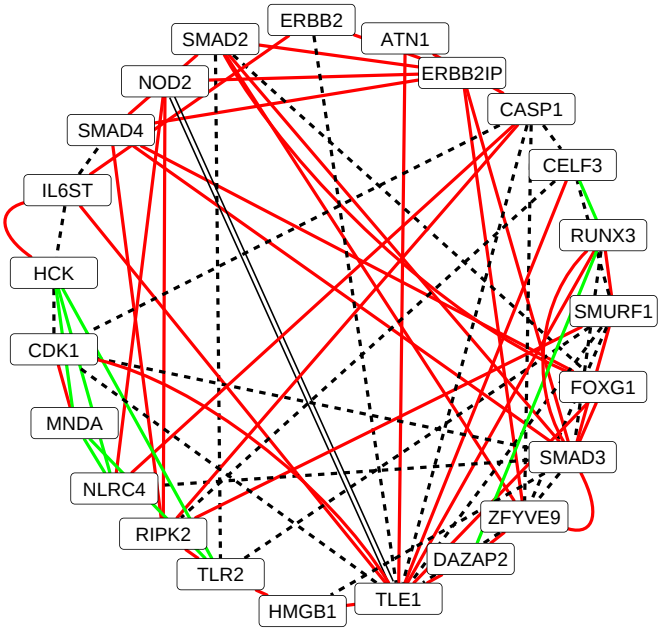


Fig. S2: Results of neighborhood searching from the combined epistatic-functional network, based on the epistatic interactions between NOD2 and TLE1 in Crohn's disease, with SNPs associated with genes within the same LD block. Double black lines indicate literature-reported epistatic interactions, black dotted lines indicate WTCCC-BOOST loosely epistatic interactions, red lines indicate PPIs, and green lines indicate co-expression.

Milo, R. et al (2002). Network motifs: Simple building blocks of complex networks. *Science*, **298**, 824–827.

Monteleone, G. et al (2001). Blocking smad7 restores TGF-β1 signaling in chronic inflammatory bowel disease. *The Journal of Clinical Investigation*, **108**, 601–609.

Monteleone, G. et al (2015). Mongersen, an oral SMAD7 antisense oligonucleotide, and crohn's disease. *New England Journal of Medicine*, **372**, 1104–1133.

Nimmo, E.R. et al (2011). TLE1 modifies the effects of NOD2 in the pathogenesis of crohn's disease. *Gastroenterology*, **141**, 972–981.e2.

Park, J.S. et al (2004). Involvement of toll-like receptors 2 and 4 in cellular activation by high mobility group box 1 protein. *Journal of Biological Chemistry*, **279**, 7370–7377.

Ramasamy, S. et al (2016). Tle1 tumor suppressor negatively regulates inflammation in vivo and modulates NF-kB inflammatory pathway. *Proceedings of the National Academy of Sciences of the United States of America*, **113**(7), 1871–1876.

Seoane, J. et al (2004). Integration of smad and forkhead pathways in the control of neuroepithelial and glioblastoma cell proliferation. *Cell*, **117**, 211–223.

Siegmund, B. (2002). Interleukin-1β converting enzyme (caspase-1) in intestinal inflammation. *Biochemical Pharmacology*, **64**, 1–8.

Weersma, R.K. et al (2008). Runt-related transcription factor 3 is associated with ulcerative colitis and shows epistasis with solute carrier family 22, members 4 and 5. *Inflammatory Bowel Diseases*, **14**, 1615–1622.

Yamazaki, K. et al (2013). A genome-wide association study identifies 2 susceptibility loci for crohn's disease in a japanese population. *Gastroenterology*, **144**, 781–788.

Yarmus, M. et al (2006). Groucho/transducin-like enhancer-of-split (TLE)-dependent and -independent transcriptional regulation by runx3. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 7384–7389.