HHS Public Access

Author manuscript

Pac Symp Biocomput. Author manuscript; available in PMC 2018 January 01.

Published in final edited form as:

Pac Symp Biocomput. 2018 ; 23: 524–535.

Convergent downstream candidate mechanisms of independent intergenic polymorphisms between co-classified diseases implicate epistasis among noncoding elements§

Jiali Han,
Center for Biomedical Informatics and Biostatistics (CB2) and Departments of Medicine and of Systems and Industrial Engineering, The University of Arizona, Tucson, AZ 85721, USA

Jianrong Li,
Center for Biomedical Informatics and Biostatistics (CB2) and Departments of Medicine and of Systems and Industrial Engineering, The University of Arizona, Tucson, AZ 85721, USA

Ikbel Achour,
Center for Biomedical Informatics and Biostatistics (CB2) and Departments of Medicine and of Systems and Industrial Engineering, The University of Arizona, Tucson, AZ 85721, USA

Lorenzo Pesce,
Computation Institute, Argonne National Laboratory and University of Chicago, Chicago, IL 60637, USA

Ian Foster,
Computation Institute, Argonne National Laboratory and University of Chicago, Chicago, IL 60637, USA

Haiquan Li, and
CB2, BIO5 Institute, UACC, and Dept of Medicine, The University of Arizona, Tucson, AZ 85721, USA

Yves A. Lussier
CB2, BIO5 Institute, UACC, and Dept of Medicine, The University of Arizona, Tucson, AZ 85721, USA

Abstract

Eighty percent of DNA outside protein coding regions was shown biochemically functional by the ENCODE project, enabling studies of their interactions. Studies have since explored how convergent downstream mechanisms arise from independent genetic risks of one complex disease. However, the cross-talk and epistasis between intergenic risks associated with distinct complex diseases have not been comprehensively characterized. Our recent integrative genomic analysis

§This work was supported in part by The University of Arizona Health Sciences CB2, the BIO5 Institute, NIH (U01AI122275, HL132532, CA023074, 1UG3OD023171, 1R01AG053589-01A1, 1S10RR029030)

Correspondence to: Haiquan Li; Yves A. Lussier.

Jiali Han, Jianrong Li and Ikbel Achour are joint-first-authors.

unveiled downstream biological effectors of *disease-specific* polymorphisms buried in intergenic regions, and we then validated their genetic synergy and antagonism in distinct GWAS. We extend this approach to characterize convergent downstream candidate mechanisms of distinct intergenic SNPs *across distinct* diseases *within* the same clinical classification. We construct a multipartite network consisting of 467 diseases organized in 15 classes, 2,358 disease-associated SNPs, 6,301 SNP-associated mRNAs by eQTL, and mRNA annotations to 4,538 Gene Ontology mechanisms. Functional similarity between two SNPs (similar SNP pairs) is imputed using a nested information theoretic distance model for which p-values are assigned by conservative scale-free permutation of network edges without replacement (node degrees constant). At FDR≤5%, we prioritized 3,870 intergenic SNP pairs associated, among which 755 are associated with distinct diseases sharing the same disease class, implicating 167 intergenic SNPs, 14 classes, 230 mRNAs, and 134 GO terms. Co-classified SNP pairs were more likely to be prioritized as compared to those of distinct classes confirming a noncoding genetic underpinning to clinical classification (odds ratio ~3.8; $p \leq 10^{-25}$). The prioritized pairs were also enriched in regions bound to the same/interacting transcription factors and/or interacting in long-range chromatin interactions suggestive of epistasis (odds ratio ~ 2,500; $p \leq 10^{-25}$). This prioritized network implicates complex epistasis between intergenic polymorphisms of co-classified diseases and offers a roadmap for a novel therapeutic paradigm: repositioning medications that target proteins within downstream mechanisms of intergenic disease-associated SNPs. Supplementary information and software: http://lussiergroup.org/publications/disease_class

Keywords

SNP; Intergenic; Noncoding; Disease class; Biological similarity; Enrichment

1. Introduction

Human diseases can be classified via multiple criteria: cell type, tissue, organ, system, topological body region, pathophysiological, epidemiological characteristics, and etiological causes. Thus, in clinical classification of diseases, genetic disorders have conventionally relegated to a subset of the classification pertaining to its etiology. The advent of genomic assays now offers the opportunity to utilize unbiasedly a broad number of molecules of life to redefine the architecture of clinical classifications.

For example, cancers pertaining to distinct organ and cell types have been shown to share common somatic mutations [1] or transcriptomes and sometimes respond to the same therapy in spite of their distinct conventional classification, suggesting a new systems oncology etiology to cancer pathophysiology. We have previously shown that the miRNome of tumors classify the primary cancers by organ of origin as expected, while their paired metastases remarkably classify according to their progression (oligometastatic vs. polymetastatic) regardless of the primary site and metastatic site [2]. Recently, Genome-Wide Association Studies (**GWAS**) have implicated the same polymorphisms to distinct diseases of the same clinical class (e.g., cardiovascular system). Many distinct autoimmune diseases are found to have the same polymorphisms relating to the major histocompatibility complex region of chromosome 6, along with some other chromosome regions involving signaling in immune

response (e.g., cytokine, interleukin, and interferon) [3, 4]. These same polymorphisms have also been associated with distinct traits of the metabolic syndrome [5].

In addition to studying each disease class separately, studies have also been conducted at a system level to unveil mechanisms that link individual diseases to a disease class. A disease class is likely to be driven by common genes and even common biological sub-networks, thus rendering a cluster structure or modularity in the biological network that separates it from other classes [6]. The modularity for disease classes has been observed in various types of molecular networks based on their risks identified in shared intragenic regions, including disease-gene networks [7–10], drug-target networks [11], transcription factor networks [6, 12], and protein-protein interaction networks [13]. Ohn broadened the similarity between diseases by looking into correlated polymorphisms by GWAS p-values [14]. In addition, two studies leveraged trans- Expression Quantitative Trait Loci (**eQTL**) analyses studies respectively limited to the immune systems and node-degree properties [15, 16]. On the other hand, traditional genetic-interaction studies such as PLINK [17] and BOOST [18], as well as recent integrative functional studies on non-coding disease variants [19, 20] such as GWAS3D [21] and CEPID [22] may also provide insight into how distinct diseases of the same disease class co-classify together. In spite of the genetic, genomic, and biological network studies generally conducted for specific disease classes, the biological mechanisms of the majority of disease-associated intergenic polymorphisms remain obscure as well as their contribution to explaining these risks at the disease class level.

We recently reported that downstream functional effects of distinct intergenic Single Nucleotide Polymorphisms (**SNP pairs**) associated with the same complex disease are likely to converge at some levels of biology such as sharing downstream transcripts or regulating functionally similar biological pathways or processes [23]. Our collaborators, Moore and Denny research groups, confirmed genetic synergy or antagonism between the top prioritized convergent intergenic SNP-pairs in a GWAS of Alzheimer's and a Phenome-Wide Association Study (PheWAS) of rheumatoid arthritis [23]. However, this study did not address the convergent mechanism of SNP pairs between distinct diseases associated with the same clinical classification (**co-classified**).

Here, the downstream functional similarity between two SNPs (**similar SNP pairs**) is imputed using a multiscale information theoretic distance model for which p-values are assigned by conservative permutation resampling of network edges without replacement (node degrees constant). We hypothesized that we could extend this approach to identify downstream mechanisms of **intergenic SNPs with distinct co-classified diseases**, by integrating the classification information of the NHGRI diseases/traits and reanalyzing the results, to infer the noncoding genetic architecture of disease classes, which has implications for drug repositioning and mitigation of risks for multiple diseases within the same class.

2. Methods

2.1. Main Datasets

We surveyed Lead SNPs (SNPs investigated in GWAS) from two datasets, the National Human Genome Research Institute (**NHGRI**) GWAS catalog [24] and the eQTL association

dataset named SNP and Copy Number Variant Annotation (SCAN) database [25]. The NHGRI GWAS catalog provides a comprehensive resource by systematically cataloging and summarizing the key characteristics of reproducible trait/disease-associated SNPs from currently published GWAS [24]. The NHGRI GWAS catalog comprises 7,236 associations between 574 diseases/traits and 6,432 distinct SNPs (6/7/2012). The SCAN database contains 4,189,682 eQTL associations between 833,004 distinct SNPs and 11,860 mRNA at $P \leq 10^{-4}$ from lymphoblastic cell lines. The integration of these two datasets yields 2,358 Lead SNPs in common (1,092 intergenic SNPs), along with their traits/diseases and mRNA information. The 574 NHGRI diseases/traits were classified into 15 organ & clinical systems disease classes according to Maurano et al. [6] along with curation (Suppl. Tab. 1).

A pairwise analysis was conducted on all possible combinations of two Lead SNPs inherited in distinct haplotypes (pairs of SNPs in strong linkage disequilibrium (**LD**) were removed from our study). The HapMap CEU LD dataset [26] was used to determine LD level and the exclusion criterion of $r^2 \geq 0.8$. Since our major interest is in intergenic variants (i.e., located between genes), the pairs in which both SNPs are intragenic (i.e., located within genes) were also excluded. The definition of "intergenic" and "intragenic" are derived from dbSNP (Build 138 on 2/21/2014) [27], which considers a SNP in a gene region to be intragenic if it is within 2kb upstream (5′ side) or 0.5 kb downstream (3′ side) of that gene. ~2.8 million pairwise combinations were derived from these Lead SNPs with $r^2 < 0.8$, associated with 467 diseases, 6,301 mRNAs, 1,635 molecular functions (**MF**), and 2,903 biological processes (**BP**). Among them, 1,977,927 pairs contain at least one intergenic SNP (named as intergenic Lead SNP pairs): 595,053 intergenic-intragenic and 1,382,874 intergenic-intergenic. 800,438 pairs are intragenic-intragenic. Among the intergenic Lead SNP pairs, 211,808 are associated with same disease classes (i.e., each SNP in one pair is associated with a specific disease class) while 1,766,119 are associated with distinct ones.

2.2. Calculation of SNP similarity

The prioritization process was applied to the intergenic Lead SNP pairs based on their convergence of eQTL-associated biological mechanisms. Three approaches were exploited to determine such shared (convergent) candidate mechanisms: (1) eQTL-associated mRNA overlap, (2) molecular function (MF) similarity of eQTL-associated mRNA, and (3) biological process (BP) similarity. We extracted MFs and BPs of each mRNA associated with a SNP from gene ontology (GO) annotations [28, 29] to calculate the similarity of a SNP pair [23] (Table 1 & Figure 1).

2.3. Network permutation to establish the p-values for observed mRNA overlap and ITS scores between two SNPs

To determine the statistical significance of imputed biologically convergent mechanisms of SNP pairs, permutation of the eQTL network was conducted for mRNA overlap, molecular function similarity, and biological process, separately. We also included the eQTL associations of SNPs not known to be associated with any diseases to create a null distribution of SNP mRNA overlap (**statistical mRNA overlap**) and ITS. When examining the significance of each of the three mechanisms, we controlled the original node degree (ND) of each specific SNP and each specific mRNA. Specifically, we kept the number of

mRNAs associating with one SNP the same, or vice versa, during the resampling of the bipartite eQTL network (shuffling the associations between SNPs and mRNAs). Deep permutations at 100,000 times were conducted on the Argonne Lab Beagle supercomputer to reach a sufficient power (20 million core hours). P-values were derived from the imputed results of the observed eQTL network and the set of permuted networks. False Discovery Rate (**FDR**) was used to adjust for multiplicity, and the SNP pairs with FDR<0.05 are termed prioritized Lead SNP pairs.

For MF and BP similarity calculations, a similar permutation procedure was conducted as done for mRNA overlap, except that SNPs and mRNAs without corresponding GO annotations were removed and only those with BP or MF associations remained in the bipartite network for resampling. We further investigated the significance of overlapped GO terms from the SNP-GO-SNP triplets for every pair of SNPs based on the same set of permutations and prioritized the overlapped terms between pairs of SNPs with a FDR<0.05. The whole procedure of permutations was conducted multiple times for different eQTL association cutoffs ranging from $P \leq 10^{-4}$ to $P \leq 10^{-6}$ and at three levels of node degrees: ND≥1, ND≥3, and ND≥5.

Through such stringent scale-free network controls, not only will the SNP pairs associated with same mRNAs be prioritized, but also the pairs in which two SNPs are associated with distinct mRNAs, if biological similarity exists.

2.4. Internal Validation: enrichment studies of co-classified intergenic SNP pairs among prioritized pairs

To demonstrate whether the shared biological mechanisms of intergenic Lead SNP pairs are relevant to the underlying biology of disease classes, we assessed whether they are more likely to be found related to the same disease class than those across distinct classes. One-tailed Fisher's Exact Test (**FET**) was applied for the enrichment study, and odds ratios of significant mRNA overlapping, MF, and BP similarities for SNP pairs associated with the same disease classes were calculated by FET at multiple eQTL p-value cutoffs and three levels of node degrees.

2.5. External Validation: ENCODE regulatory elements and chromatin interaction enrichment of co-classified prioritized intergenic SNP pairs

The potential mechanisms at play for the prioritized SNP pairs were also investigated. We evaluated whether regulatory mechanisms were more likely to occur in prioritized intergenic SNP pairs associated with the same disease class as compared to their counterparts (distinct classes or insignificant). We integrated Encyclopedia of DNA Elements (**ENCODE**) data [19] of Lead SNPs and conducted Fisher's Exact Test to assess the enrichment of molecular regulations within prioritized SNP pairs of the same disease classes. Three possible shared regulatory mechanisms are assessed for pairs of SNPs located in distinct regions, including (1) binding with same transcription factor (via ChIP-seq), (2) binding with distinct transcription factors (via **ChIP-seq**) connecting through protein-protein interaction (**PPI**), and (3) within the anchor regions of long-range chromatin interactions (via **ChIA-PET**[34]). We compared the enrichment of regulatory mechanisms with two conventional methods,

which prioritized SNP pairs by (1) any intergenic Lead SNP pairs and (2) intergenic Lead SNP pairs with at least one mRNA overlap (**non-statistical mRNA overlap**) in eQTL associations, respectively. To avoid loss of information when calculating regulatory functions between Loci in ENCODE, every Lead SNP was extended to its strongly associated LD SNPs based on the RegulomeDB database [35] (inheritable haplotype).

3. Results and Discussion

3.1. Overall results and visualization

Prioritization of convergent downstream mechanisms of SNPs required extensive conservative scale-free permutation resampling of network edges (node degrees constant), shown substantially more conservative than conventional theoretical statistics or similarity-scores cutoffs (Suppl. Fig. 1). We prioritized 3,870 intergenic Lead SNP pairs (1,378 intergenic-intergenic; 2,492 intergenic-intragenic) at FDR<0.05 that share at least one of the three imputed biological mechanisms, of which 755 pairs are found within the same disease class (280 intergenic-intergenic pairs; 475 intergenic-intragenic; 80 were associated with the same diseases). Without additional prioritization, the network relates these 755 pairs with as many as 1,683 mRNAs and 2,060 GO terms. After convergent mechanism prioritization, these SNP pairs implicate 14 disease classes, 277 Lead SNPs (167 intergenic, 98 noncoding intragenic, 12 protein-coding), 230 mRNAs, and 134 GO mechanisms. A simplified network shows only the 755 prioritized intergenic Lead SNP pairs and their related disease classes, leaving out the mRNAs and GO-terms for simplicity (Fig. 2). 14 of the 15 studied disease classes harbor convergent biological processes and molecular functions perturbed by a set of intergenic SNPs with similar downstream effects, presenting an apparent modularity for each class. We further show a sub-network of prioritized biological mechanisms for the prioritized SNP pairs associated with the same classes in Fig. 3. The convergent connections among intergenic SNPs of distinct diseases within the same disease class suggest the investigation of an unusual form of pleiotropy: distinct intergenic risks of co-classified disease sharing common downstream mechanisms that could affect the same target transcripts that may relate to the emergence of both diseases in the same pathophysiological classification (e.g., Fig. 4 showing the detail of co-classified diseases associated through SNP pair similarity in Fig. 2, only cancer and cardiovascular system shown).

3.2. Enrichment of shared biological mechanisms in prioritized intergenic SNP pairs of distinct co-classified diseases (Methods 2.4)

We investigated whether intergenic Lead SNP pairs, with each SNP associated with two distinct co-classified diseases, were more likely to share a biological mechanism (prioritized) than SNP pairs associated with distinct diseases classified in distinct pathophysiological classes. Enrichment analyses were performed for the 755 prioritized SNP pairs associated with same classes among 3,870 prioritized intergenic Lead SNP pairs at different eQTL p-value cutoffs ($10^{-6} \leq$ eQTL p-value$\leq 10^{-4}$; 100,000 permutation resampling, SNP pair FDR<0.05) and different node degrees SNP node degree (count of mRNAs associated with that SNP at the eQTL p-value cutoff). As shown in Fig. 5, odds ratios (**ORs**) range from 1.4 to 3.8 (x-axis: $5.1\times10^{-6} \leq$p-value≤ 0.02), 1.4 to 3.4 ($6.5\times10^{-26} \leq P \leq 2.1\times10^{-2}$), and 1.9 to 3.7 ($8.3\times10^{-4} \leq P \leq 2.2\times10^{-7}$) for mRNA overlapping,

MF similarity, and BP similarity, respectively. This internal validation supports the hypothesis that biological mechanisms are more likely to be shared within a class of diseases and may define in part a common pathophysiology of otherwise distinct diseases.

3.3. Enrichment of ENCODE regulatory elements and chromatin interaction in prioritized intergenic SNP pairs of distinct co-classified diseases (Methods 2.5)

ENCODE data provides an opportunity to question if convergent candidate mechanisms of prioritized SNP pairs of co-classified diseases imputed by eQTL associations may be attributed to common regulatory elements (e.g., transcriptional factors) or long-range chromatin interactions. If so, this could be suggestive of possible epistasis between disease risks of distinct co-classified diseases, in other words, a disease class epistasis. We identified substantial enrichment in three types of regulatory elements: shared transcription factor (Fig. 6**panel A**), interacting transcription factors (Fig. 6**panel B**), and long-range chromatin interactions in the region of the SNPs in the pair (Fig. 6**panel C**). However, the effect size (odds ratio) of enrichment of regulatory elements in SNP pairs associated with distinct co-classified diseases shown in the figure is about 30 percent smaller than that of our previously published enrichment of SNP pairs associated with the *same* disease (not shown [23]). Taken together, these results indicate that common regulatory mechanisms of intergenic SNPs strongly underpin the pathogenesis of a disease and to a moderate degree some mechanisms are also shared by distinct, yet pathophysiologically co-classified diseases.

4. Limitations and future studies

First, we only reported eQTLs derived from LCL cell lines. Studies on 44 tissues in the GTEx project are ongoing and will be reported elsewhere. SNPs with marginal p-values [36] will also be investigated using the proposed method to unveil their pairwise synergy. Second, gene ontology annotations are biased by human interest. Even though the biases were controlled partially by the scale-free persisted permutations, some biases may still exist and induce false positive results. Alternative unbiased approaches may be worth incorporating in the future such as the information-theoretic framework to address the accuracy of the GO annotation [37–39]. Third, the permutations on large eQTL networks are expensive; we are working on more efficient implementations and strategies. Fourth, the validation in a GWAS of epistasis between convergent intergenic SNPs associated with distinct co-classified diseases is not possible retrospectively as clinical phenotypes are generally obtainable for only one disease in a GWAS. A prospective study for the validation is cost-prohibitive; we are thus planning a collaboration with eMERGE researchers to conduct a PheWAS. Finally, the SNPs prioritized in this study are statistically associated with but not necessarily functionally causal to a disease (or co-classified diseases) thus other polymorphisms inherited in the same loci must be considered. Of note, our approach incorporated this calculation through the Linkage Disequilibrium parameter (**Methods section 2.5**). Also, further systematic investigation on the relationship between functional synergy and genetic interaction of SNPs prone to same or co-classified diseases will provide insight into the mechanisms of disease classes.

Beyond the modularity within classes, related disease classes are obviously also interconnected through shared genes and gene ontology annotations in Fig. 3. This study focused on intergenic SNPs prioritized across distinct diseases of the same class, leaving out thousands of SNP pairs prioritized across classes. Indeed, cross-class biomodularity merits its own publication and additional analyses due to its complexity.

5. Conclusion

Using the quantified measurement of SNP biological similarity we recently developed, we identified 755 intergenic SNP pairs associated by convergent eQTL function to distinct, yet pathophysiologically co-classified diseases. We found that these independently inherited (LD $r^2<0.01$) intergenic SNP pairs were more likely to be enriched in (i) shared transcription factors, (ii) interacting transcription factors, and (iii) long-range chromatin interactions. A common genetic architecture of the pathophysiology of co-classified diseases is unsurprising; however, a common noncoding intergenic architecture for clinical classification harbors many new questions. For example, is epistasis occurring between distinct disease risks, and if so, can some disease risks protect against other diseases through antagonism of long-range chromatin interactions implicating noncoding intergenic regions? Additionally, can we implicate new drug targets or reposition drugs through the shared intergenic interactions between distinct co-classified diseases? While the prioritized intergenic SNP pairs associated with each disease class reassuringly recapitulates the pathophysiological classification of disease of complex inheritance, does this implicate that complex diseases are fundamentally distinct from Mendelian ones through these noncoding interactions? Indeed, GWAS identified about half the variants in intergenic regions. However, the array platforms are seeded biasedly with half the probes in intergenic regions (selection bias). This proposes that more than 80% of the complex-disease associated variants could be located in intergenic regions, suggesting that if the heritability gap is attributable to genetic interactions, the majority of these would occur with intergenic noncoding regions. On the other hand, our study aligns further intergenic genetic signal with that of the central dogma of molecular biology, as we provide for each prioritized SNP pair falsifiable hypotheses of convergent mechanisms implicating coding regions (eQTL mRNAs).

This prioritized network implies complex epistasis between intergenic polymorphisms of co-classified diseases and offers a roadmap for a novel therapeutic paradigm: repositioning medications that target proteins within downstream mechanisms of intergenic disease-associated SNPs.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Drs. Nancy Cox and Eric R. Gamazon for sharing their eQTL associations, Dr. Roger Luo for curating the disease classification, and Dr. Colleen Kenost for proofreading the manuscript.

References

1. Inaki K, Liu ET. Trends in Genetics. 2012; 28:550–559. [PubMed: 22901976]
2. Lussier YA, Xing HR, et al. PloS one. 2011; 6:e28650. [PubMed: 22174856]
3. Seldin MF. Journal of autoimmunity. 2015; 64:1–12. [PubMed: 26343334]
4. Zenewicz LA, Abraham C, et al. Cell. 2010; 140:791–797. [PubMed: 20303870]
5. Vattikuti S, Guo J, Chow CC. PLoS genetics. 2012; 8:e1002637. [PubMed: 22479213]
6. Maurano MT, Humbert R, et al. Science. 2012; 337:1190–1195. [PubMed: 22955828]
7. Bulik-Sullivan B, Finucane HK, et al. Nature genetics. 2015; 47:1236–1241. [PubMed: 26414676]
8. Goh KI, Cusick ME, et al. Proc Natl Acad Sci U S A. 2007; 104:8685–8690. [PubMed: 17502601]
9. Jiang X, Liu B, et al. FEBS letters. 2008; 582:2549–2554. [PubMed: 18582463]
10. Lee Y, Li J, et al. Summit on Translational Bioinformatics. 2010:31.
11. Yildirim MA, Goh KI, et al. Nature biotechnology. 2007; 25:1119.
12. Karczewski KJ, Dudley JT, et al. Proc Natl Acad Sci U S A. 2013; 110:9607–9612. [PubMed: 23690573]
13. Sam L, Liu Y, et al. Pacific Symposium on Biocomputing. 2007:76. [PubMed: 17992746]
14. Ohn JH. J Am Med Inform Assoc. 2017:ocx026.
15. Fehrmann RS, Jansen RC, et al. PLoS genetics. 2011; 7:e1002197. [PubMed: 21829388]
16. Li H, Pouladi N, et al. Journal of biomedical informatics. 2015; 58:226–234. [PubMed: 26524128]
17. Purcell S, Neale B, et al. The American Journal of Human Genetics. 2007; 81:559–575. [PubMed: 17701901]
18. Wan X, Yang C, et al. The American Journal of Human Genetics. 2010; 87:325–340. [PubMed: 20817139]
19. ENCODE Project Consortium. Nature. 2012; 489:57. [PubMed: 22955616]
20. Lee Y, Gamazon ER, et al. PLoS genetics. 2012; 8:e1002998. [PubMed: 23133393]
21. Li MJ, Wang LY, et al. Nucleic acids research. 2013; 41:W150–W158. [PubMed: 23723249]
22. Li MJ, Li M, et al. Genome biology. 2017; 18:52. [PubMed: 28302177]
23. Li H, Achour I, et al. NPJ genomic medicine. 2016; 1:16006. [PubMed: 27482468]
24. Welter D, MacArthur J, et al. Nucleic acids research. 2013; 42:D1001–D1006. [PubMed: 24316577]
25. Gamazon ER, Zhang W, et al. Bioinformatics. 2009; 26:259–262. [PubMed: 19933162]
26. Gibbs RA, Belmont JW, et al. 2003
27. Sherry ST, Ward M-H, et al. Nucleic acids research. 2001; 29:308–311. [PubMed: 11125122]
28. Ashburner M, Ball CA, et al. Nature genetics. 2000; 25:25. [PubMed: 10802651]
29. Gene Ontology Consortium. Nucleic acids research. 2010; 38:D331–D335. [PubMed: 19920128]
30. Lin D. International Confernece on Machine Learning. 1998:296–304.
31. Tao Y, Sam L, et al. Bioinformatics. 2007; 23:i529–i538. [PubMed: 17646340]
32. Regan K, Wang K, et al. J Am Med Inform Assoc. 2012; 19:306–316. [PubMed: 22319180]
33. Li H, Lee Y, et al. J Am Med Inform Assoc. 2012; 19:295–305. [PubMed: 22278381]
34. Djebali S, Davis CA, et al. Nature. 2012; 489:101. [PubMed: 22955620]
35. Boyle AP, Hong EL, et al. Genome research. 2012; 22:1790–1797. [PubMed: 22955989]
36. Lee Y, Li H, et al. J Am Med Inform Assoc. 2013; 20:619–629. [PubMed: 23355459]
37. Alterovitz G, Xiang M, et al. Nucleic acids research. 2006; 35:D322–D327. [PubMed: 17098937]
38. Chen JL, Liu Y, et al. BMC bioinformatics. 2007; 8:S7.
39. Clark WT, Radivojac P. Bioinformatics. 2013; 29:i53–i61. [PubMed: 23813009]

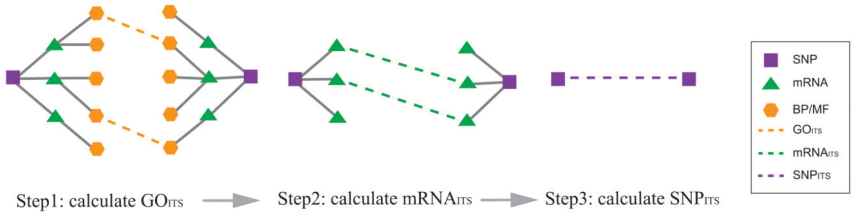

Fig. 1.

Nested Information theoretic calculations. The similarity between SNP pairs is calculated by three nested steps subsequently (I) similarity between two gene ontology terms (GO_{ITS}), (II) similarity between two genes ($mRNA_{ITS}$) using GO term similarities, and (III) similarity between two SNPs (SNP_{ITS}) using mRNA similarities.

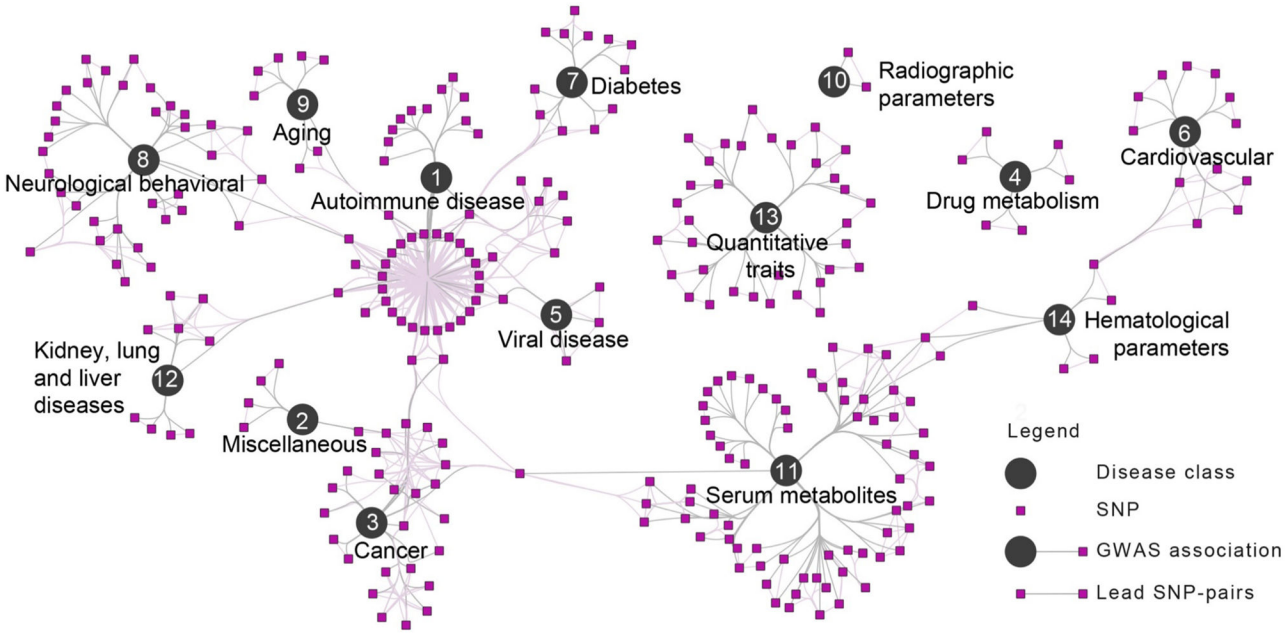

Fig. 2.

The network of 755 prioritized intergenic SNP pairs within disease class at FDR<0.05. 80 SNP pairs are within the same disease (previously published), 675 are within the same disease class but across distinct diseases (new). 3,115 SNP pairs prioritized cross-class are not shown. 19 SNPs were associated with two distinct diseases in distinct classes by GWAS and shown.

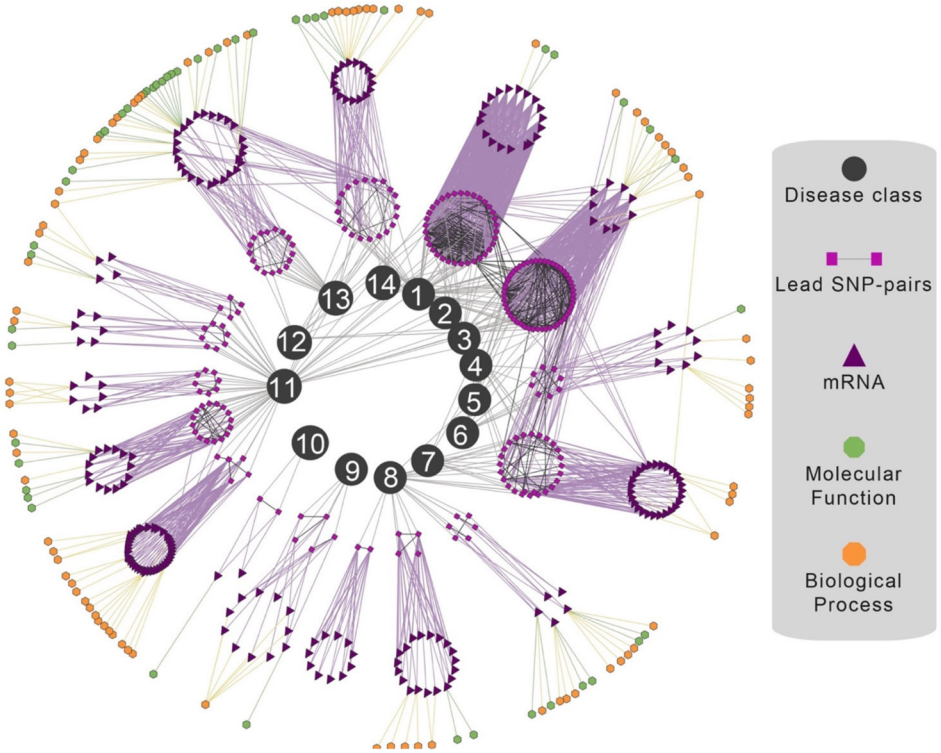

Fig. 3.

The subset of the prioritized network of disease class mechanisms containing 230 mRNAs shared between 428 SNP pairs and their associated GO mechanisms (48 GO-MFs, and 86 GO-BPs). Biological modularity of shared groups of mRNAs is associated with distinct SNPs themselves associated with distinct co-classified diseases. Not shown are the biomodules where 327 SNP-pairs are associated by distinct mRNAs to distinct but similar pathways (Methods 2.2). Names of classes are defined in Fig. 2.

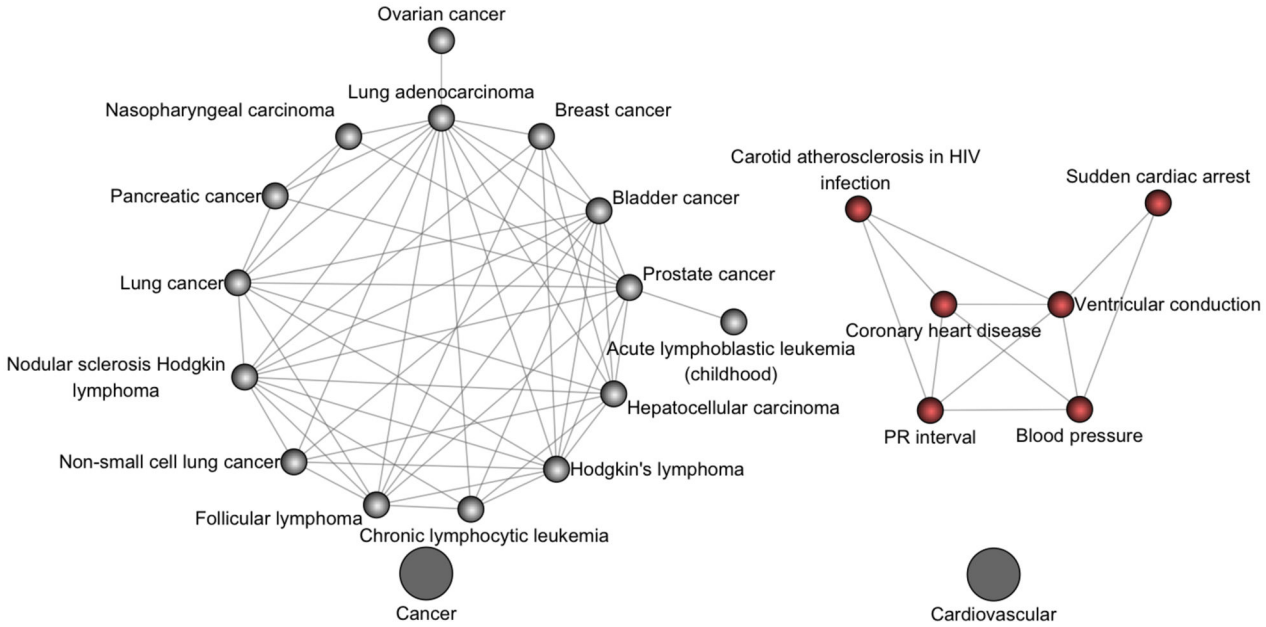

Figure 4.

Details of implicated co-classified diseases through SNP pair similarity confirming shared genetic underpinning and biological mechanisms. Two classes, cancers (Fig. 2–3 #3) and cardiovascular disease (Fig. 2–3; #6), shown. Disease-pairs are related by at least one out of 755 prioritized pairs of Lead SNPs, each associated with a disease in the pair respectively. Previous studies have shown somatic mutations and transcriptomes can reclassify cancers molecularly. Here a new property is presented: common mechanisms of noncoding intergenic regions.

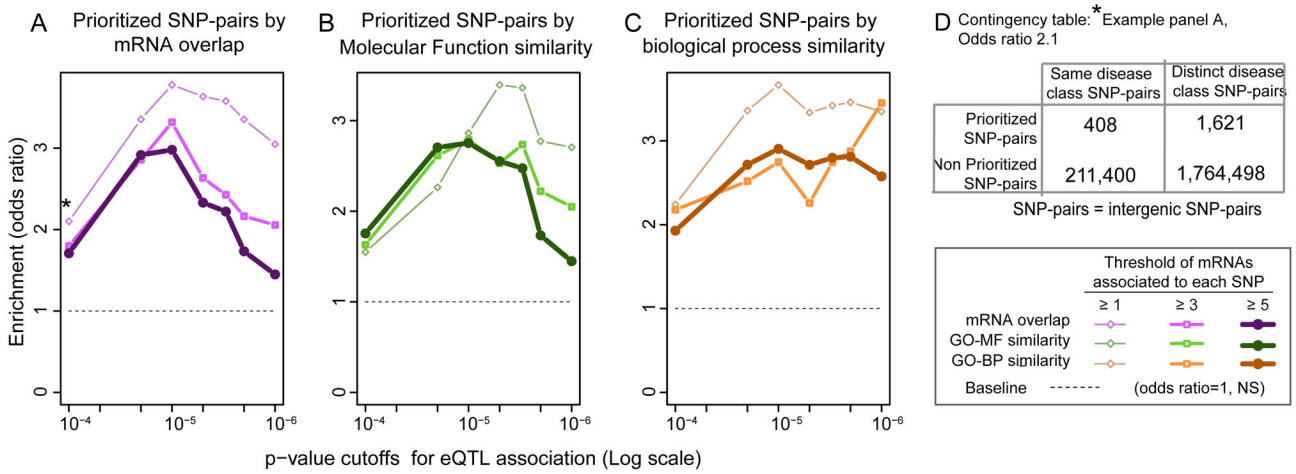

Fig. 5.

Enrichment of shared biological mechanisms among 755 intergenic Lead SNP pairs associated with the same disease classes (Method 2.1, LD cutoff r^2<0.8), remains similar with more stringent LD cutoff (r^2<0.01, not shown) and also remains the same when excluding the previously published 80 SNP pairs associated with the same diseases (results not shown). The subset of 280 prioritized SNP pairs comprising only intergenic-intergenic pairs also remains significant (Suppl. Fig. 2).

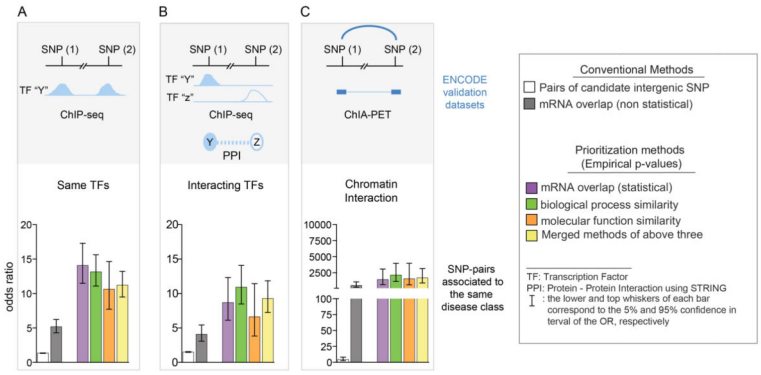

Fig. 6.

Enrichment of common ENCODE-derived regulatory mechanisms in genomic regions of the prioritized intergenic Lead SNP pairs for disease classes. More stringent LD cutoff (R^2<0.01) yielded similar results (not shown).

Table 1

Biological similarity calculations between two SNPs using nested Information Theoretic Similarity (ITS)

Nested calculations (3 steps)
1. Calculate the Information Theoretic Similarity (ITS) between two GO terms (GO_{ITS}) associated with the two SNPs through mRNAs using Lin's method [30, 31].
2. Based on $GO_{ITS,}$ calculate the information theoretic similarity between two distinct mRNAs ($mRNA_{ITS}$), each associated with a SNP within a SNP Pair, using a modified Tao's approach [31–33].
3. Determine the semantic biological similarity between two SNPs (SNP_{ITS}) within a SNP pair using the $mRNA_{ITS}$ of pairs of mRNAs associated with the two SNPs respectively, using Li's nested ITS approach we recently published [23]. The SNP_{ITS} values range from 0 to 1, with 0 corresponding to no similar downstream effects and 1 corresponding to identical downstream effects (e.g., either the same mRNAs or distinct mRNAs with the equivalent GO terms). The similarity measurement between SNPs can capture relationships between SNPs including the ones without any common mRNAs in their eQTL associations.