Journal of Biomedical Informatics 58 (2015) 226–234

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

eQTL networks unveil enriched mRNA master integrators downstream of complex disease-associated SNPs

Haiquan Li [a,b], Nima Pouladi [a,b], Ikbel Achour [a,b], Vincent Gardeux [a,b], Jianrong Li [a,c], Qike Li [a,d], Hao Helen Zhang [d,e], Fernando D. Martinez [b,f], Joe G.N. 'Skip' Garcia [a,c], Yves A. Lussier [a,b,c,d,*]

[a] Department of Medicine, University of Arizona, Tucson, AZ, USA
[b] Bio5 Institute, University of Arizona, Tucson, AZ, USA
[c] Cancer Center, University of Arizona, Tucson, AZ, USA
[d] Interdisciplinary Program in Statistics, University of Arizona, Tucson, AZ, USA
[e] Department of Mathematics, University of Arizona, Tucson, AZ, USA
[f] Department of Pediatrics, University of Arizona, Tucson, AZ, USA

ARTICLE INFO

Article history:
Received 21 August 2015
Revised 15 October 2015
Accepted 20 October 2015
Available online 30 October 2015

Keywords:
Translational bioinformatics
Centrality
Complex diseases
eQTL
Single Nucleotide Polymorphism (SNP)
Master integrator
Computational genomics
Genomics
Transcriptome
mRNA
Network biology
Big data
Computational biology
Computational medicine
Complex disease
Genetics
Systems biology
Systems medicine
Signal integration

ABSTRACT

The causal and interplay mechanisms of Single Nucleotide Polymorphisms (SNPs) associated with complex diseases (complex disease SNPs) investigated in genome-wide association studies (GWAS) at the transcriptional level (mRNA) are poorly understood despite recent advancements such as discoveries reported in the Encyclopedia of DNA Elements (ENCODE) and Genotype-Tissue Expression (GTex). Protein interaction network analyses have successfully improved our understanding of both single gene diseases (Mendelian diseases) and complex diseases. Whether the mRNAs downstream of complex disease genes are central or peripheral in the genetic information flow relating DNA to mRNA remains unclear and may be disease-specific. Using expression Quantitative Trait Loci (eQTL) that provide DNA to mRNA associations and network centrality metrics, we hypothesize that we can unveil the systems properties of information flow between SNPs and the transcriptomes of complex diseases. We compare different conditions such as naïve SNP assignments and stringent linkage disequilibrium (LD) free assignments for transcripts to remove confounders from LD. Additionally, we compare the results from eQTL networks between lymphoblastoid cell lines and liver tissue. Empirical permutation resampling ($p < 0.001$) and theoretic Mann–Whitney U test ($p < 10^{-30}$) statistics indicate that mRNAs corresponding to complex disease SNPs via eQTL associations are likely to be regulated by a larger number of SNPs than expected. We name this novel property *mRNA hubness* in eQTL networks, and further term mRNAs with high hubness as *master integrators*. mRNA master integrators receive and coordinate the perturbation signals from large numbers of polymorphisms and respond to the personal genetic architecture integratively. This genetic signal integration contrasts with the mechanism underlying some Mendelian diseases, where a genetic polymorphism affecting a single protein hub produces a divergent signal that affects a large number of downstream proteins. Indeed, we verify that this property is independent of the hubness in protein networks for which these mRNAs are transcribed. Our findings provide novel insights into the pleiotropy of mRNAs targeted by complex disease polymorphisms and the architecture of the information flow between the genetic polymorphisms and transcriptomes of complex diseases.

1. Introduction

Genome-wide association studies (GWAS) have successfully discovered many genetic variants associated with the development and progression of diverse classes of complex diseases. However, existing knowledge about the mechanisms of complex diseases is limited. One important reason may be that GWASs focus primarily on finding mere associations of genetic variants with traits, while ignoring the underlying architecture of complex diseases. Many common single nucleotide polymorphisms (SNPs) are associated with complex diseases, but each SNP contributes only a small amount to the disease risk. The observation of additive and synergistic effects implies the presence of crosstalk and network of complicated regulatory mechanisms of SNPs [1–3] that can be perceived as local perturbations. However, identifying and

* Corresponding author at: 1657 E Helen Street, Tucson, AZ 85721, USA. Tel.: +1 520 626 0245; fax: +1 520 626 4824.

http://dx.doi.org/10.1016/j.jbi.2015.10.010
1532-0464/

characterizing these systematic mechanisms of SNPs, which cause the underlying complex diseases, remains a challenge.

During the past decade, network analyses have increased our understanding about the underlying mechanisms of the behavior of biological systems and their higher order properties such as their topological ones [4,5]. Centrality, as a basic topological property, measures the connectivity of a node to other nodes, which represents the relative impact of that node in the network upon its perturbation. When applying this framework to a biological network: (i) proteins are modeled as nodes in protein–protein interaction (PPI) networks, while (ii) both mRNAs and SNPs are nodes in expression quantitative trait loci (eQTL) networks. Edges relating the molecules of life are protein interactions in PPIs and statistical associations between SNPs and mRNAs in eQTL networks. The degree of a node is the number of connections it has to other nodes, and nodes with the highest degree are called "hubs". In PPI networks, genetic loss of function of a hub protein can be embryologically lethal (essential genes) as shown in knock down mouse models [6,7], while genetic gain of function of the corresponding hub proteins (e.g. transcriptional factors) are known to increase cancer progression [8]. Hub proteins of non-essential genes are enriched in Mendelian diseases, with large effect sizes [7]. Missense mutations of hub protein may lead to a more subtle alteration of the biological function and has been more frequently seen in Mendelian diseases [9]. On the other hand, for other complex diseases (e.g. diabetes mellitus, hypertension), the individual impact from a polymorphism usually is not fatal, therefore, genes associated with these diseases are generally found peripherally in the PPI network [10,11].

Complex disease single nucleotide polymorphisms are considered upstream discrete inputs in genetic models with outputs being multiple quantitative continuous phenotypes that characterize a disease [12]. In genetic models combined with downstream protein interaction networks, the polymorphisms or mutations of a gene may initiate an edgetic perturbation and cause a Mendelian disorder [13,14]. While edgetic effects are much more subtle in complex diseases, previous studies of biological networks identified an enrichment of intragenic SNPs in the shortest paths of complex disease associated host genes [15]. Our prior research also revealed the enrichment of the polymorphisms coding for hub proteins in PPI networks of complex diseases [16]. This finding suggests some convergence of complex and Mendelian diseases at the protein interaction level [17]. We applied this clue to Alzheimer's disease and indeed found the candidate genes sharing molecular mechanisms between Mendelian and complex diseases coupled with shortest PPI path strategies [18]. Nevertheless, whether hubness is an intrinsic property of biological networks underpinning complex polymorphisms is not widely accepted [19].

GWAS demonstrate the difference between complex diseases and single gene diseases, as dozens of polymorphisms rarely explain more than 10% of the phenotypes of the former while a single polymorphism usually determines the key phenotypes of the latter. However, there is a lack of information models that take into account the downstream transcriptomic effect of complex diseases genetics to provide formal insight about their genetic architecture. To our knowledge, previous studies focus on PPI networks and divergent centrality of hub proteins. For instance, a hub in a gene co-expression network, such as a microRNA (e.g. miR-204 [20]), can downregulate the expression of many downstream genes, acting as a divergent inhibitor (Fig. 1A). A hub in a protein signaling network, such as signal receptor binding protein Ras [21,22]), may amplify a signal it receives and activate a series of downstream processes through physical protein interactions and post-translational modifications (Fig. 1B). In a complex disease scenario, a master regulator SNP may affect the expression of many downstream mRNAs [1] (Fig. 1C). However, the mRNAs dysregulated by SNPs related to complex diseases may behave differently than other divergent hubs (Fig. 1D). An mRNA hub (e.g. HLA-DQA1 [23]) may combine the perturbation signals from various polymorphisms, consequently affecting the regulation of other downstream mRNAs such that a disease phenotype develops.

Increasing our knowledge of how genetic signals are integrated (multivariate discrete inputs) is crucial to our understanding of their downstream impact on disease phenotypes (multivariate continuous outputs). Biological signal integration has been the subject of numerous studies, particularly in the context of crosstalk

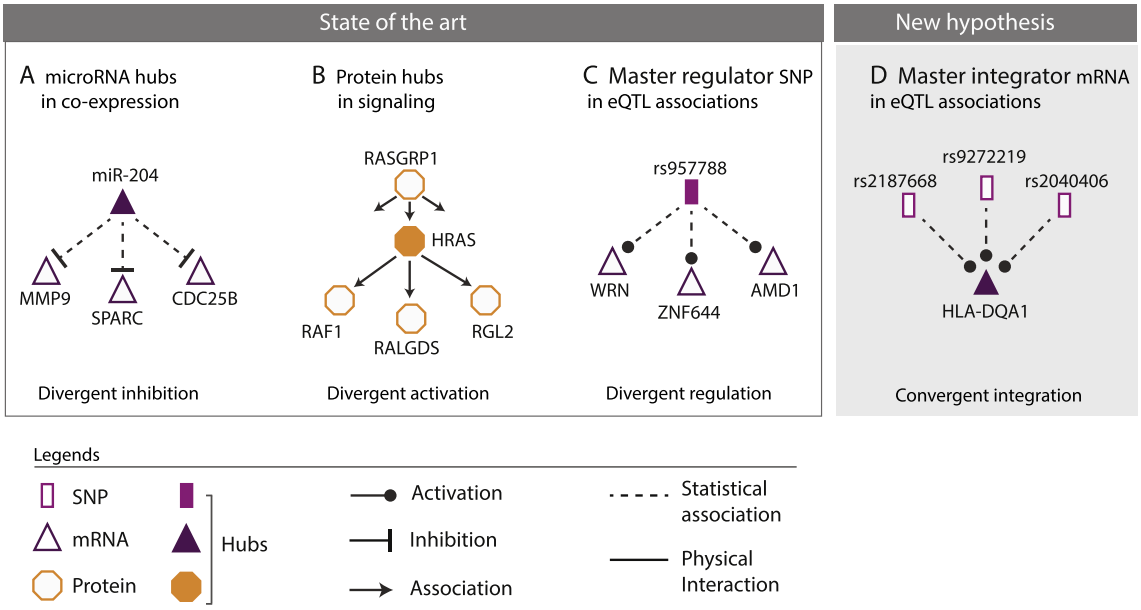

Fig. 1. Divergent and convergent (integrator) hubs in biological networks. Existing studies concentrate on divergent hubs, which inhibit or activate excessive numbers of downstream molecules of life, as exemplified by microRNAs (Panel A), essential signaling proteins (Panel B), and master regulator SNPs (Panel C). Genes susceptible to complex diseases may interact and enrich downstream mRNA master integrators, which integrate an excessive number of genetic perturbations, finally leading to a disease phenotype (Panel D). Of note, divergent hubs have an excessive number of out degrees while convergent hubs (master integrators) have an excessive number of in degrees in the directional networks.

between individual signaling pathways [24]. Unsurprisingly, cells must integrate a large number of signals in order to maintain their homeostasis and regulate complex cellular processes. Recent studies shifted their signal integration focus from canonical pathways to more complex biological networks [25]; however, to our knowledge eQTL networks have not been studied for signal integration.

Few studies report convergent integration for disease genes from different scales of gene regulation networks [5]. Some network studies indicate the prevalence of the divergent and convergent topology in biological networks as compared to alternatives, such as densely connected modules [26]. Particularly, these studies focus more on divergent hubs than on convergent integrative ones [27]. In this regard, we hypothesize that networks of genomic loci affecting expression levels of mRNAs would provide novel insight into systems properties of complex diseases. We concentrate specifically on expression Quantitative Trait Loci association (eQTL) [28] networks and investigate whether mRNAs related to complex disease SNPs are more central than the remaining mRNAs in the network of eQTL associations. We term the mRNAs related to an excessive number of SNPs (e.g. top 20%) *mRNA master integrators*, and use the eQTL data of two human cell types from lymphoblastoid (LCL) and liver tissues to examine the enrichment of mRNA master integrators in the eQTL networks. We further hypothesize that mRNA master integrators play important roles in integrating multiple perturbation signals from genetic polymorphisms and are closely related to the physiopathology of complex diseases. Our aim is to reveal a novel type of mRNA master integrators that appears to characterize in part systems biology of complex diseases.

2. Methods

The flow of the methodology is illustrated in Fig. 2 and the abbreviations are in Table 1. Three types of datasets were employed, with subsequent measurement of mRNA node degree in eQTL networks and then investigation of the mRNA master integrator enrichment in genes related to complex disease classes and specific complex diseases. An additional protein interaction dataset was used to compare the findings in eQTL networks and protein interaction networks.

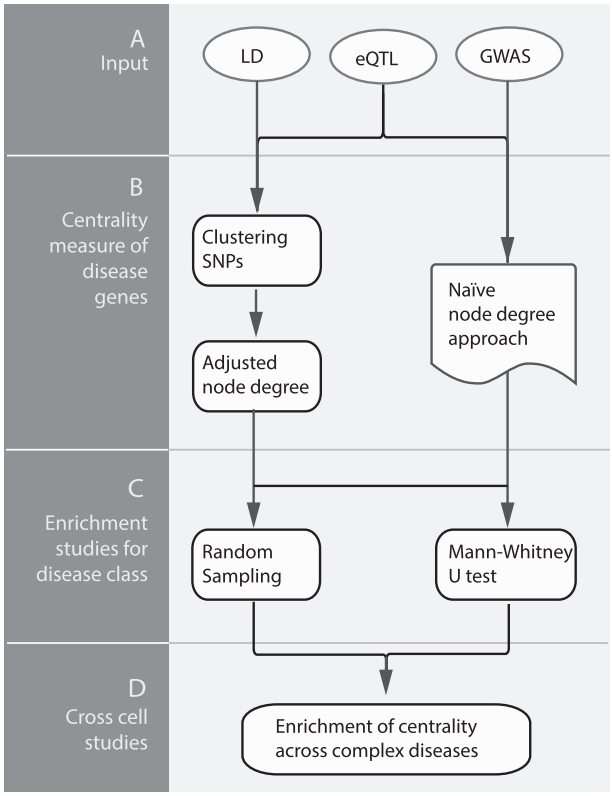

Fig. 2. Schematic diagram of our approach.

Table 1
Definitions and descriptions of abbreviations and terms.

Abbreviation	Definition; descriptions
Centrality	The impact of a node in a network; consequence of the dysregulation of a molecule of life (node) on the function of the biological network
Edge	A connection between two nodes in a graph or network; e.g. biophysical interaction (edges) between proteins (nodes)
eQTL	Expression Quantitative Trait Loci; e.g. statistical associations between the a SNP value and the level of expression of a mRNA
GWAS	Genome-wide Association Study; e.g. statistical association between a SNP value and a trait (e.g. response to therapy) or disease
Hub; hubness	A highly connected node in a network
LCL	Lymphoblastoid cell line
LD	Linkage disequilibrium; non-random association of alleles (e.g. SNP values) at different genomic locations
mRNA master integrator	mRNA associated with a large number of distinct SNPs in an eQTL network
Node	An entity in a network: a molecular of life (e.g. SNP, mRNA. protein) in a biological network
Pleiotropy	One gene that influences two or more unrelated phenotypes (e.g. diseases)
Polymorphism	Genetic variations between individual subjects
PPI	Protein–protein interaction (network)
SNP	Single Nucleotide Polymorphism; DNA sequence variation occurring in which a single nucleotide differs among individual subjects

2.1. Datasets

Five datasets were employed in this study:

(i) Two SNP-mRNA eQTL datasets derived from lymphoblastoid cell lines (LCL; B lymphoblast) and liver tissues from Caucasian populations were downloaded from SCAN.db [23] on October 11, 2010 and August 15, 2013, respectively. The LCL dataset consisted of 4,189,682 associations between 833,004 distinct SNPs and 11,860 mRNAs (p-value $\leqslant 10^{-4}$), and the liver dataset contained 314,545 associations between 139,814 SNPs and 19,641 mRNAs (p–value $\leqslant 10^{-5}$).

(ii) SNP-complex disease (or trait) associations were downloaded from NHGRI GWAS catalog on May 2012. It comprised 7236 associations between 6432 SNPs and 574 complex diseases/traits. These traits were classified into 15 disease classes according to the Maurano et al. [29] curation, with few additional traits manually curated by the authors (Supplementary Table 1: http://www.lussierlab.net/publications/eQTL_centrality/Table-S1-disease-Class.xls).

(iii) Associations of alleles at two loci were assessed by SNP linkage disequilibrium (LD) [30], which were downloaded from Hapmap on April 19, 2009 [31]. Only Caucasian LD data was used.

(iv) Protein interaction data were downloaded from String-DB [22]. The dataset combined v8.2 and v6.3 for high confidence protein interactions and was used in our previous study of protein centrality [16].

2.2. Measurement of centrality in an eQTL association network

An eQTL association network is dependent on a cutoff that determines the minimal level of statistical significance for each

association. The network relates SNPs to mRNA and is thus structured as a bipartite graph. We defined the hubness centrality measure of mRNAs according to their respective node degrees in the eQTL bipartite graph, where the node degree of an mRNA was defined as the number of SNPs associated with the mRNA with respect to the significance cutoff.

The measure for hubness of an mRNA may be biased by connection of LD SNPs in the same "LD block" to identical mRNAs; we assumed these should not be repeatedly counted. Therefore, we defined two different measures for qualifying the hubness of a given mRNA: (i) the "unadjusted node degree" that does not take into account the LD bias and (ii) the "adjusted node degree" that clusters the LD SNPs into "LD blocks" (according to a LD cutoff) before calculating the number of independent SNPs associated with a given mRNA. Fig. 3 shows the algorithm used for clustering the SNPs into the so-called "LD blocks".

mRNA hubness was defined as the normalized node degree (either unadjusted or adjusted) across all mRNAs. The normalization procedure first ranked the node degree of all mRNAs in ascending order and then divided the order number (starting from 1) by the total number of mRNAs. An mRNA with a hubness score of at least 80% (top 20%) was defined as a *master integrator* in the eQTL bipartite network.

2.3. mRNA master integrator enrichment test

We annotated the SNPs to complex disease classes according to the GWAS and class curation (see Section 2.1). At a given eQTL cutoff, all SNPs without eQTL associations were excluded from the enrichment analysis. mRNAs associated with any SNP of a complex disease in the eQTL associations were annotated as the disease-class related mRNAs. Some mRNAs annotated to a disease class overlapped with another class.

The enrichment of mRNA master integrators in a disease class was investigated in two ways: (i) The hubness of each mRNA was compared with that of all other mRNAs in the eQTL bipartite network, using a one tail Mann–Whitney *U* test (function "wilcox.test" in R) to examine whether the median mRNA hubness in this class was significantly larger than that in the background (consisting of the scores of the hubness of all mRNAs not in the class). Enrichment of the mRNA master integrators of complex diseases as a whole was done similarly. (ii) The proportion of mRNA master integrators in the class was examined, with a deviation of expected value 20% as a direct indicator of master integrator enrichment. We assessed the significance of the proportion of mRNA master integrators in each complex disease class through randomization. A bipartite network was constructed such that the two layers were composed of SNPs and mRNAs, respectively. We then calculated the proportion of master integrators for each class of diseases. Afterwards, we generated 1000 control distributions for each disease class by (1) resampling SNPs without replacement equal in numbers to those of GWAS studies subsumed by the disease class and (2) randomly resampling mRNAs equal in number to that in the observed eQTL network. Following this, we computed the proportion of master integrators in each of the randomly generated gene sets, thus giving rise to a null distribution. Subsequently, we calculated the number of times the randomly generated master integrator proportion values exceeded that of the observed ones, and assigned a *p*-value to the observed value accordingly. We repeated the same process separately for all of the different classes of diseases.

2.4. Comparison of eQTL centrality to protein centrality

To study the centrality of mRNAs and their coding protein products comparatively, we computed the hubness of proteins from PPI networks. Per convention, proteins with top proportions (20%) of hubness were defined as hubs. Then, we compared eQTL mRNA hubness and corresponding protein hubness of mRNAs associated with complex diseases, each disease class, and each specific disease (See Section 2.1) using Fisher's Exact Test (FET) for enrichment and Spearman method for correlation at various centrality cutoffs.

3. Results

eQTL analysis shows how the expression values of a large number of mRNAs are altered and associated with each class of complex diseases. We observed that in LCL cell lines, 6301 (53%), 1095 (9.4%) and 160 (2.1%) of all mRNAs are related to complex diseases at the eQTL *p*-value cutoff values of 10^{-4}, 10^{-5} and 10^{-6}, respectively. We also observed the same pattern in liver tissue eQTL data, in which 1638 out of 19,641 (8.3%) and 362 out of 12,851 (2.8%) remained associated with complex diseases at cutoff values of 10^{-5} and 10^{-6}, respectively. After examining the proportion of master integrators (the *top 20% of* mRNAs with highest node connectivity values) among these mRNAs at different *p*-value cutoffs, we noticed a higher ratio than the expected value of 20%. Specifically, there are 31.3% (Fig. 4A), 54.7%, and 65.6% of master integrators for LCL cell lines at eQTL *p*-value cutoff of 10^{-4}, 10^{-5} and 10^{-6}, respectively, and 47.3% and 53.9% for liver tissue at *p*-value $\leqslant 10^{-5}$ and 10^{-6}, respectively. Importantly, we found that across the range of these different *p*-value cutoffs the median hubness of mRNAs within a complex disease class is significantly higher than those not related to complex disease (the background mRNAs; controls) by using Mann–Whitney (MW) *U* test (*p*-value $\leqslant 9.7 \times 10^{-40}$ and *p*-value $\leqslant 2.6 \times 10^{-56}$ for LCL and liver, respectively) (Fig. 4A and C; data of other cutoffs not shown). Further analysis corroborated our findings. QQ-plots reveal that the

Algorithm Adjust_Node_Degree (*mRNA*, τ)
Begin
1) Create an empty list of clusters
2) Randomly select an SNP from the SNPs associated with the *mRNA* in the eQTL association network but not included in any cluster, and calculate its average LD to all existing clusters using the following equation (*r* being the correlation of LD between two SNPs):

$$average\_ld(snp_k, cluster_i) = \frac{1}{|cluster_i|} \sum_{snp_j \in cluster_i} r^2(snp_k, snp_j)$$

3) Identify the cluster i^* for which the average LD to SNP_k is maximal. If the value is at least τ then add SNP_k to the cluster, else build a new cluster
4) If any SNP has not yet been considered and is not in any cluster, go back to step 2
5) The number of clusters induced becomes the adjusted node degree of the given *mRNA*

End

Fig. 3. Clustering algorithm used for creating the "LD blocks" of SNPs with high LD values. It is a greedy algorithm inspired from average linkage algorithm [32]. The LD cutoff τ represents the minimum average LD (r^2) of an SNP cluster.

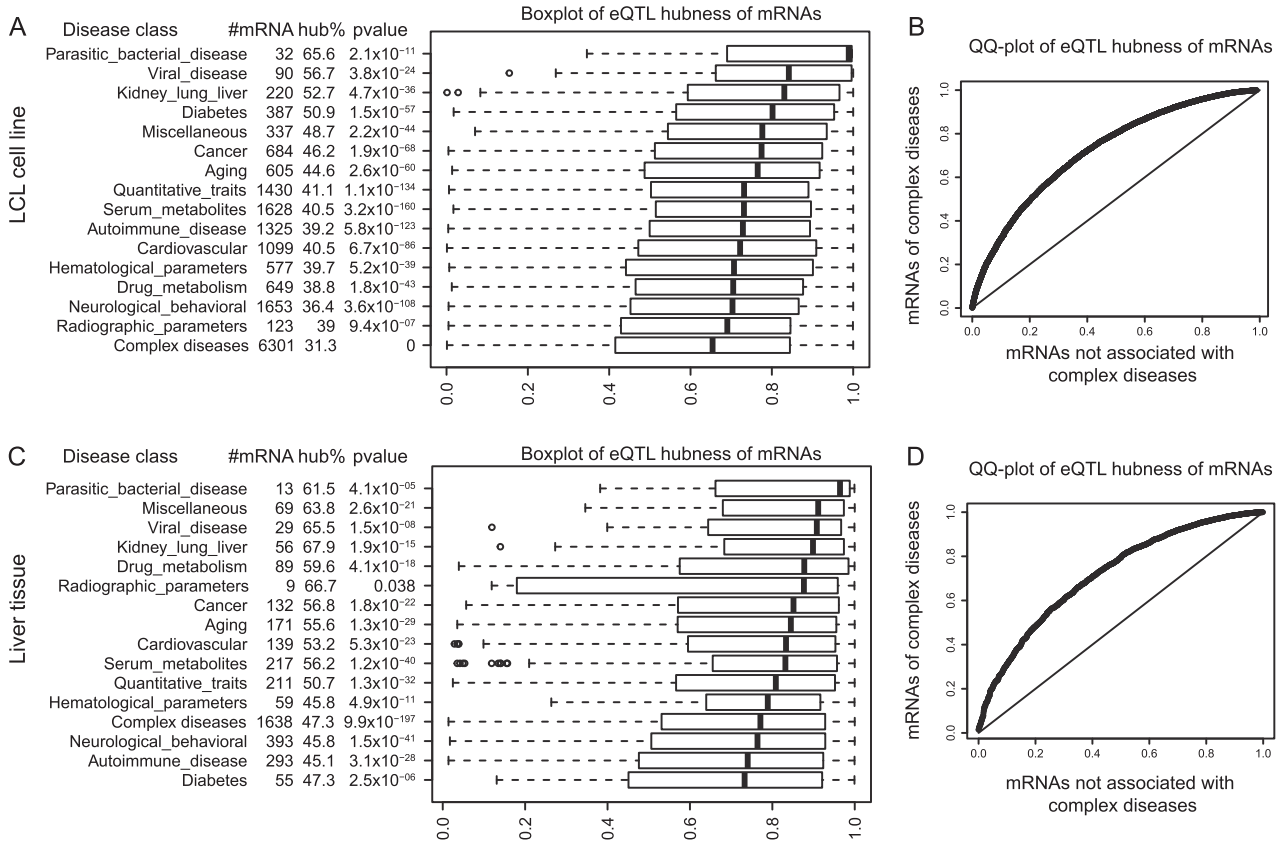

Fig. 4. Enrichment of mRNA master integrators in eQTL association networks. mRNAs of complex diseases tended to possess a large node degree (superior mRNA hubness) than those not related to complex diseases in both LCL cell lines (Panels A and B) and liver tissue (Panels C and D). The enrichment of mRNA master integrators is retained in each disease class level (Panels A and C). The proportion of master integrators (mRNA hubs) in each disease class is significantly more than the expected 20% (Panels A and C; third column of legend): for LCL eQTL $p \leqslant 10^{-4}$, 36.4–65.6% of mRNAs are hubs in the eQTL network (Panel A; $10^{-161} < p < 10^{-6}$; one tail MW p-value significance, class ranked by median mRNA hubness) while for liver tissue eQTL $p \leqslant 10^{-5}$, the corresponding proportion is 45.1–67.9% (Panel C; $10^{-40} < p \leqslant 0.038$; one tail MW test).

distributions of mRNA hubness of these two groups are significantly different in both the LCL and liver tissues (Fig. 4B and D). More importantly, we observed that all complex disease classes tended to comprise mRNAs with high degree of hubness in both LCL cell lines and liver tissue (Fig. 4A and C). We also assessed the enrichment of master integrators in each class of complex diseases through randomization (See Section 2). Interestingly, the results from randomization-based analysis matched those generated from Mann–Whitney *U* test (Fig. 5). These findings imply that the enrichment of mRNAs with high connectivity values is an inherent property of complex diseases irrespective of specific disease class.

Due to the possibility of variants affecting mRNA being in linkage disequilibrium, thus confounding the results of master integrator enrichment, we clustered SNPs associated with the same mRNAs at various range of r^2 cutoffs, specifically at values 0.8, 0.5, 0.3, 0.1 and 0.01, and retrieved the independent SNPs (See Section 2). Even with this measure, we observed that our results remain reproducible in the two cell lines under study (Fig. 6).

Furthermore, we observed that shared mRNAs across various classes of complex diseases do not confound our findings. Once we removed from 12% to 44% of mRNAs that are related to at least two disease classes, the majority of disease classes still showed a higher degree of enrichment with master integrators (data not shown). The data indicates that enrichment of distinct disease classes with highly connected mRNAs is an inherent property of the *transcriptome* network inferred from the associations between mRNAs and complex disease-associated SNPs, both intragenic and intergenic.

We also focused on the master integrators of individual complex diseases to see whether our results were due to a mere aggregate effect. Specifically, we ranked the complex diseases according to the enrichment of mRNA hubness as compared to the background. The top 10 diseases/traits were tissue/cell line relevant, in addition to a 50% overlap (Tables 2 and 3). Moreover, the membership of the top 10 was roughly consistent across various *p*-value cutoffs. This indicates that the enrichment of mRNA hubness converged on disease class levels despite the heterogeneity of individual complex diseases.

Finally, we studied the interconnection between mRNA hubness and protein hubness. Overall, there is a slight correlation between eQTL mRNA hubness and protein interaction hubness (correlation = 0.037, Spearman; OR = 1.27, $p = 0.0015$, FET). No consistent enrichment was observed for any disease class between mRNA master integrators and protein centralities (either hubness). For specific diseases/traits, only a few traits related to specific protein functions, such as proinsulin levels and fasting glucose-related traits, indicated a moderate of enrichment between master integrators and protein hubs. These observations suggest polymorphisms perturbing mRNA expressions and protein interactions likely involve different mechanisms.

4. Discussion

Previous studies have focused primarily on the assessment of the centrality on one scale of 'omic data by defining nodes as genetic variants, mRNAs, or proteins. These studies do not investigate the centrality of mRNAs in disease network across two scales

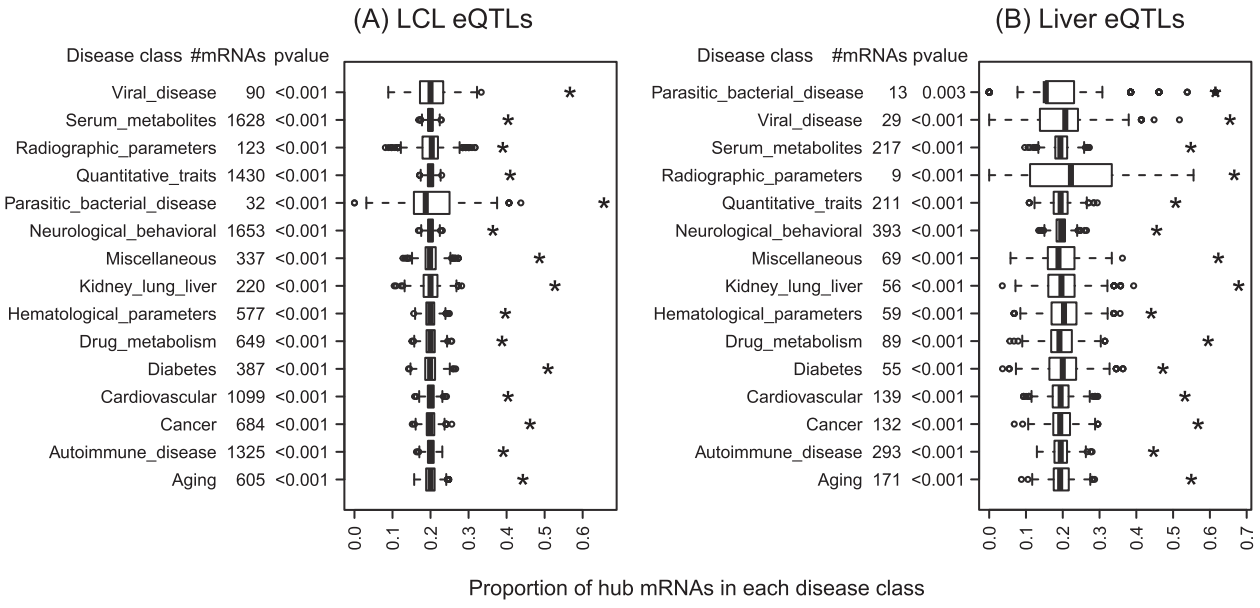

Fig. 5. Random sampling studies reemphasize the enrichment of master integrators in eQTL association networks. Stars (∗) represent the observed proportions of master integrators in the eQTL network for the disease classes. The statistical significance of the proportion of master integrators was studied using random sampling from the same eQTL association network as that in theoretical statistics, with eQTL p-values $\leqslant 10^{-4}$ for LCL (Panel A) and $\leqslant 10^{-5}$ liver tissues (Panel B) (Methods). Empirical studies yielded the same enrichment results as those given by the Mann–Whitney U test (Fig. 4), which indicates that theoretical statistics work for this study. Boxplots show the null distribution of the proportion of master integrators (mRNAs with top 20% of connectivity to SNPs) generated through randomization.

of cell biology, namely genetic variants and the expression values of mRNAs. In eQTL networks, mRNA regulation works through the cumulative effect of multiple upstream genetic variants. The findings in the current study contrast mRNA master integrators, which respond to multiple genetic polymorphisms, with hubs in co-expression and protein interaction networks (Fig. 1). Indeed, disrupting an mRNA master integrator has a local effect while disrupting a co-expression of a protein interaction hub has far reaching consequences to deregulate the network. In addition, deregulation of hubs in undirected protein interaction networks may cause the deregulation of their direct interaction neighbors. Paradoxically, in directed bipartite eQTL networks where SNPs may regulate directly or indirectly mRNA expression, mRNA master integrators are highly constrained mRNAs for which the regulation is coordinated by the summative effect of multiple upstream genetic signals (eQTL SNPs). In other words, complex diseases with high percentage of mRNA master integrators (in eQTL networks), such as systematic lupus erythematosus, are less sensitive to a single SNP perturbation than those with lower percentage of mRNA master integrators.

In this work, we derived mRNA networks of two cell lines/tissues from their corresponding eQTL data since this type of data makes the connection between two biological scales possible. We then computed the hubness centrality of mRNAs in these eQTL networks. In doing so, we find that mRNAs associated with complex disease SNPs are more likely to possess higher node degrees. In other words, mRNA associated with complex disease polymorphism via eQTL studies are perturbed by a large number of distinct polymorphisms, thus showing high *responsiveness* to the genetic architecture.

These complex disease-related mRNA master integrators in eQTL networks may correspond to the mRNAs that are susceptible to diseases due to multiple perturbations. Our observation is robust and reproducible across various ranges of eQTL p-value cutoffs, different tissue/cell types, and distinct values of linkage disequilibrium. We further corroborated our findings through randomization. The robustness of our results shows that the enrichment of mRNA master integrators in eQTL networks of complex diseases is likely an inherent property of complex disease classes.

Our results provide new insights to the genetic architecture of complex diseases. In accordance with our findings, a large number of genetic variants may perturb complex diseases, which in turn alter the expression of many mRNAs, both of which have small effect sizes. Our results further suggest that the larger number of genetic variations may perturb the expression of complex disease genes, thus making the interconnectivity between genetic variants and gene expression more complicated than expected. In other words, this study suggests an intermediate level of leveraging signal integration between SNP and diseases by mRNA master integrators and implies the robustness of the human organism upon individual perturbations. Because distinct SNPs, associated with distinct diseases in GWAS, may influence the expression level of the same mRNA, this observation provides insight into the possible role of mRNA master integrators as a mechanism of pleiotropy.

In our measurement of hubness centrality of complex disease-related genes, we computed both the number of associated SNPs regardless of LD and the number of independent SNPs in eQTL networks derived from a clustering approach. The rationale for this strategy is that both measures correspond to some biological mechanisms. Studies have shown that even SNPs in LD may cooperatively regulate the downstream mRNAs, such as by working as a part of two enhancers to regulate the target genes [8]. In other words, LD SNPs, although not distinguishable in eQTL associations, are not completely functionally redundant. On the other hand, independent SNPs without sufficient LD are inherited separately and are thus more likely to possess independent, distinct function mechanisms, with respect to the same mRNA. Therefore, they are distinct sources of perturbation for the expression of mRNAs and, consequently, distinct sources of the underlying complex disease.

However, our findings should be interpreted with caution because of the predictions derive from five datasets. First, we simply defined the mRNAs associated with trait-associated SNPs in eQTL studies as related mRNAs of complex diseases, but did not

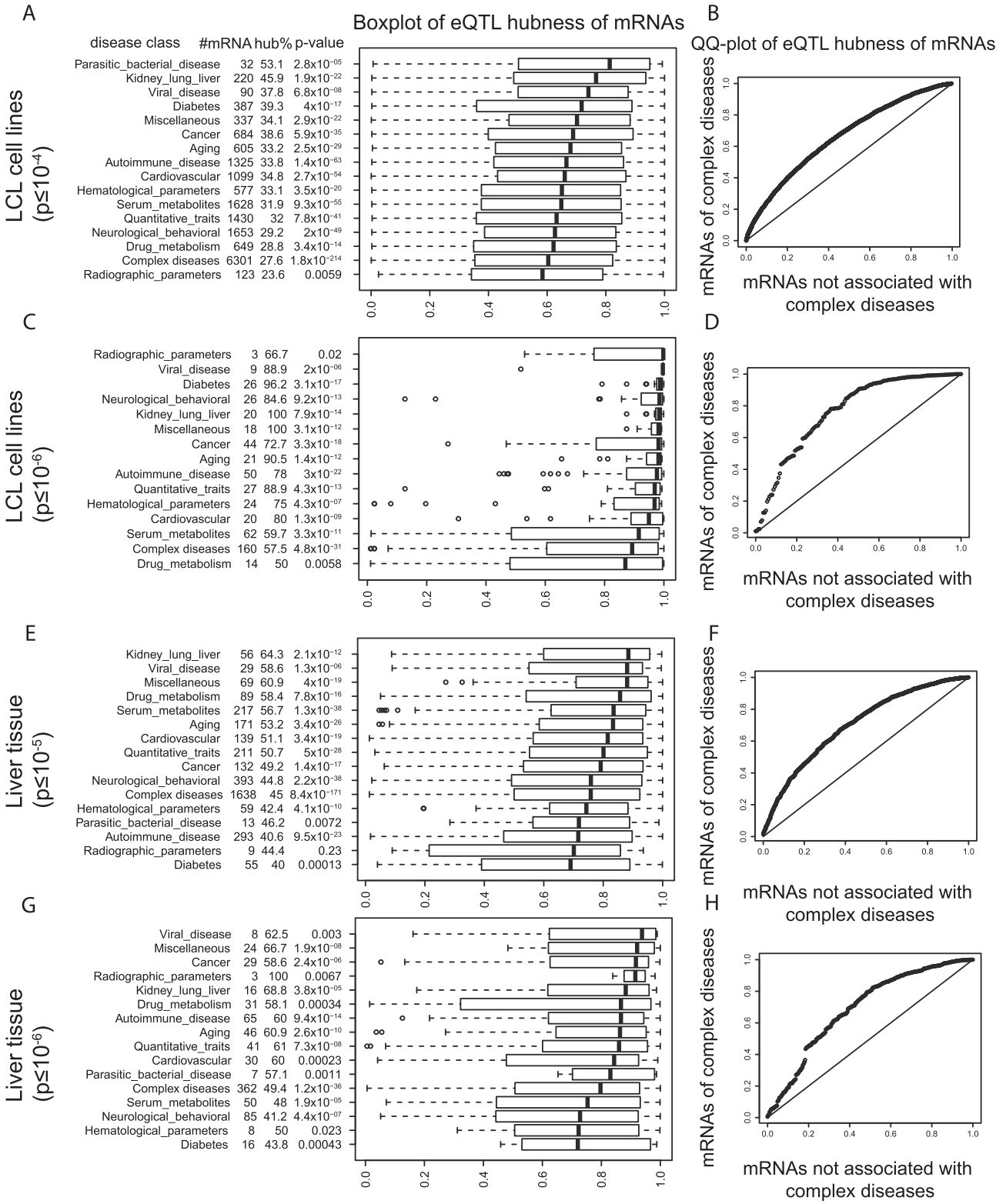

Fig. 6. Enrichment of master integrators in mRNAs associated with complex disease SNPs under the linkage disequilibrium control (LD $r^2 < 0.01$). Master integrators was enriched in mRNAs related to complex disease by eQTL and GWAS associations even under the control of linkage disequilibrium (number of SNPs with LD $r^2 < 0.01$) in LCL cell lines (Panel A to D) and liver tissue (Panel E to H). The enrichment remained in each disease class level (Panel A, C, E and G). For LCL cell lines, the p-value for Mann–Whitney U test is at most 0.02 but down to 1.4×10^{-63} for any disease class at an LCL p-value cutoff of 10^{-4} (Panel A) and 10^{-6} (Panel C). For liver tissue with eQTL $p \leqslant 10^{-5}$ and 10^{-6}, hubness centrality was enriched in all disease classes except radiographic parameter traits ($1.3 \times 10^{-38} \leqslant p \leqslant 0.023$; one tail MW test; Panel E and G). Analysis of hubness centrality among complex disease-related mRNAs also shows that the enrichment of mRNA master integrators is stronger with increasing strength of eQTL associations. The boxplot of the hubness for mRNAs under eQTL p-value cutoffs of 10^{-4}–10^{-6} demonstrated this trend: mRNA hubness shifts from the left (small node degrees) to the right (Panel A vs C, Panel E vs G); complex disease-related mRNAs render larger deviations on hubness between the two groups of distributions (Panel B, D, F, and H) for more stringent eQTL. For instance, 25 out of 26 mRNAs associated with diabetes and all 20 mRNAs associated with kidney, lung and liver diseases are master integrators at LCL eQTL $p \leqslant 10^{-6}$.

Table 2
Top 10 complex diseases derived from *LCL cell lines* with enriched master integrators under LD control. Numbers of master integrators (#hub) indicate those among the top 20% of scores of adjusted node degree of independent SNPs ($r^2 < 0.01$). The significance of the enrichment for the mRNA master integrators as compared to other mRNAs not related to the disease was yielded by Mann–Whitney U test. Complex diseases are ranked according to the significance under eQTL $p \leqslant 10^{-6}$. Bolded diseases are those consistent with that of liver tissues. Of note, 8 out of 10 are autoimmune diseases.

Disease class	Complex disease	eQTL $p \leqslant 10^{-6}$		eQTL $p \leqslant 10^{-5}$		eQTL $p \leqslant 10^{-4}$	
		#hub	Significance*	#hub	Significance*	#hub	Significance*
Autoimmune disease	**Systemic lupus erythematosus**	29	1.4×10^{-16}	56	4.6×10^{-26}	93	3.4×10^{-19}
Autoimmune disease	Multiple sclerosis	19	1.3×10^{-12}	35	1.0×10^{-13}	107	1.0×10^{-13}
Autoimmune disease	**Rheumatoid arthritis**	20	8.5×10^{-14}	27	7.6×10^{-14}	62	1.5×10^{-15}
Diabetes	Type 1 diabetes	19	3.4×10^{-13}	22	2.9×10^{-9}	62	1.4×10^{-5}
Autoimmune disease	**Asthma**	25	4.0×10^{-17}	26	1.9×10^{-15}	35	2.3×10^{-9}
Autoimmune disease	Ulcerative colitis	18	1.6×10^{-12}	21	7.0×10^{-12}	44	3.8×10^{-9}
Serum metabolites	Hypothyroidism	17	6.6×10^{-12}	20	1.2×10^{-11}	39	1.3×10^{-14}
Autoimmune disease	Systemic sclerosis	19	3.4×10^{-13}	24	1.4×10^{-14}	21	1.5×10^{-6}
Autoimmune disease	**Immunoglobulin A**	18	1.4×10^{-12}	18	1.3×10^{-10}	19	8.6×10^{-8}
Autoimmune disease	**Inflammatory bowel disease**	17	6.5×10^{-12}	16	1.6×10^{-9}	15	1.3×10^{-7}

Table 3
Top 10 complex diseases derived from *liver* tissue with enriched master integrators under LD control. Number of master integrators (#hub) corresponds to those among the top 20% of scores of adjusted node degree of independent SNPs ($r^2 < 0.01$). The significance of the enrichment for the mRNA master integrators as compared to other mRNAs not related to the disease was yielded by Mann–Whitney *U* test. Complex diseases are ranked according to the significance at eQTL $p \leqslant 10^{-6}$. Bolded diseases are consistent with those of LCL cell lines. Of note, 6 out of 10 are autoimmune diseases.

Disease class	Complex disease	eQTL $p \leqslant 10^{-6}$		eQTL $p \leqslant 10^{-5}$	
		#hub	Significance*	#hub	Significance*
Autoimmune disease	**Systemic lupus erythematosus**	16	3.2×10^{-10}	31	3.5×10^{-13}
Neurological behavior	Cognitive performance	5	1.6×10^{-4}	24	2.6×10^{-11}
Autoimmune disease	**Asthma**	10	1.6×10^{-5}	15	1.7×10^{-6}
Quantitative traits	Skin pigmentation	6	3.2×10^{-5}	17	1.9×10^{-12}
Autoimmune disease	**Immunoglobulin A**	8	4.7×10^{-6}	11	1.8×10^{-6}
Autoimmune disease	Celiac disease	8	1.5×10^{-5}	11	1.3×10^{-4}
Autoimmune disease	**Inflammatory bowel disease**	7	1.2×10^{-5}	9	1.3×10^{-5}
Cancer	Hodgkin's lymphoma	6	8.3×10^{-5}	9	1.9×10^{-5}
Autoimmune disease	**Rheumatoid arthritis**	5	2.1×10^{-4}	10	4.3×10^{-5}
Cancer	Nodular sclerosis Hodgkin lymphoma	6	3.7×10^{-5}	8	6.6×10^{-5}

confirm through case-control transcriptome studies that these mRNAs are indeed observed as dysregulated in transcriptome measurements. Second, we counted the SNPs associated with an mRNA in measuring the hubness of the mRNA, regardless of the SNPs to disease association. In the future, we will focus studies on intragenic SNPs leveraging high confidence predictions of the mechanism associated with the polymorphisms using bioinformatics software such as MutationTaster [33] and SNPdryad [34]. Third, we analyzed eQTL data derived from only two cell types; more cell lines or tissue eQTL, such as those from GTEx [35], should be investigated to reinforce reproducibility in the future. Fourth, our genetic variant data covers a range of common SNPs limited to those with significant eQTL associations. Fifth, the current approach was based on eQTL studies that did not measure alternative splicing transcripts of a gene. One interpretation of our results could be that mRNA hubness is confounded with distinct alternative splicing variants specific to each eQTL-associated SNP [36]. As alternative splicing eQTL data of a recent study has become available [37], we intend to follow-up to characterize further mRNA hubness mechanisms. Additionally, more computational methods, such as other permutation and clustering strategies including adaptive and non-random clustering, could have been be used to substantiate the findings. Although enrichment was diminished when using over-conservative control (using SNPs in eQTL networks unrelated to the disease class as a control rather than all SNPs regardless of their eQTL association), further verification is needed [33,34].

5. Conclusions

Network theory has increased our knowledge of the higher order characteristics of various classes of human diseases, both complex and Mendelian. Among the distinct properties of disease networks, protein interaction hubness centrality of nodes is of biological and clinical importance due to its correlation with lethality; however, studies have seldom investigated the network properties of regulatory networks that are perturbed by intergenic SNPs. Using eQTL network centrality metrics, our study shows that mRNAs associated with SNPs of complex diseases are systematically more likely to be master integrators than mRNAs associated with non-disease SNPs in significant eQTL associations. Further, we confirm this pattern within each complex disease class and verify that these mRNA master integrators are independent of the hubs of the proteins coded by these mRNAs. Our findings provide novel insights into the possible pleiotropy of mRNAs targeted by complex disease polymorphisms and the architecture of the information flow between the genetic polymorphisms and transcriptomes of complex diseases. Despite the limitations of our study, our findings are still of clinical importance, as they indicate that the mRNA expression values of the genes contributing to the development and progression of complex disease are associated with an increasing number of genetic variations. The findings outlined here highlight the importance of developing combinatorial therapy approaches with the ultimate goal of improving quality of life for patients.

Conflict of Interest

The authors declare they have no direct or indirect conflict of interest.

Acknowledgments

The study was supported in part by the USA NIH Grant UL1TR000050 (University of Illinois CTSA), K22LM008308, and NCI P30CA023074 grant of the University of Arizona Cancer Center, USA. We thank Colleen Kenost and Jacob Smith for their assistance with proofreading the manuscript. We acknowledge the contribution of Dr. Roger Luo for verifying the annotation of the diseases into classes.

References

[1] D.L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M.E. Dolan, N.J. Cox, Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS, PLoS Genet. 6 (2010) e1000888.
[2] J.F. Degner, A.A. Pai, R. Pique-Regi, J.-B. Veyrieras, D.J. Gaffney, J.K. Pickrell, et al., DNase [thinsp] I sensitivity QTLs are a major determinant of human expression variation, Nature 482 (2012) 390–394.
[3] X. Zhang, E.L. Moen, C. Liu, W. Mu, E.R. Gamazon, S.M. Delaney, et al., Linking the genetic architecture of cytosine modifications with human complex traits, Hum. Mol. Genet. ddu313 (2014).
[4] P. Jia, S. Zheng, J. Long, W. Zheng, Z. Zhao, DmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks, Bioinformatics 27 (2011) 95–102.
[5] L. Zhang, S. Kim, Learning gene networks under SNP perturbations using eQTL datasets, PLoS Comput. Biol. 10 (2014) e1003420.
[6] K.-I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, A.-L. Barabási, The human disease network, Proc. Nat. Acad. Sci. 104 (2007) 8685–8690.
[7] H. Jeong, S.P. Mason, A.-L. Barabási, Z.N. Oltvai, Lethality and centrality in protein networks, Nature 411 (2001) 41–42.
[8] S. Wachi, K. Yoneda, R. Wu, Interactome–transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues, Bioinformatics 21 (2005) 4205–4208.
[9] T. Ideker, R. Sharan, Protein networks in disease, Genome Res. 18 (2008) 644–652.
[10] A.-L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: a network-based approach to human disease, Nat. Rev. Genet. 12 (2011) 56–68.
[11] F. Barrenas, S. Chavali, P. Holme, R. Mobini, M. Benson, Network properties of complex human disease genes identified through genome-wide association studies, PloS One 4 (2009) e8090.
[12] M.N. Cantor, Y.A. Lussier, Mining OMIM™ for insight into complex diseases, Stud. Health Technol. Inform. 107 (2004) 753.
[13] Q. Zhong, N. Simonis, Q.R. Li, B. Charloteaux, F. Heuze, N. Klitgord, et al., Edgetic perturbation models of human inherited disorders, Mol. Syst. Biol. 5 (2009).
[14] M.N. Cantor, I.N. Sarkar, O. Bodenreider, Y.A. Lussier, Genestrace: phenomic knowledge discovery via structured terminology, in: Pacific Symposium on Biocomputing, World Scientific, 2005, pp. 116–127.
[15] H. Li, Y. Lee, J.L. Chen, E. Rebman, J. Li, Y.A. Lussier, Complex-disease networks of trait-associated single-nucleotide polymorphisms (SNPs) unveiled by information theory, J. Am. Med. Inform. Assoc. 19 (2012) 295–305.
[16] Y. Lee, H. Li, J. Li, E. Rebman, I. Achour, K.E. Regan, et al., Network models of genome-wide association studies uncover the topological centrality of protein interactions in complex diseases, J. Am. Med. Inform. Assoc. 20 (2013) 619–629.
[17] L. Sam, Y. Liu, J. Li, C. Friedman, Y.A. Lussier, Discovery of protein interaction networks shared by diseases, in: Pacific Symposium on Biocomputing, 2007, pp. 76.
[18] K. Regan, K. Wang, E. Doughty, H. Li, J. Li, Y. Lee, et al., Translating Mendelian and complex inheritance of Alzheimer's disease genes for predicting unique personal genome variants, J. Am. Med. Inform. Assoc. 19 (2012) 306–316.
[19] S. Chavali, F. Barrenas, K. Kanduri, M. Benson, Network properties of human disease genes with pleiotropic effects, BMC Syst. Biol. 4 (2010) 78.
[20] Y. Lee, X. Yang, Y. Huang, H. Fan, Q. Zhang, Y. Wu, et al., Network modeling identifies molecular functions targeted by miR-204 to suppress head and neck tumor metastasis, PLoS Comput. Biol. 6 (2010) e1000730.
[21] S.L. Campbell, R. Khosravi-Far, K.L. Rossman, G.J. Clark, C.J. Der, Increasing complexity of Ras signaling, Oncogene 17 (1998) 1395–1413.
[22] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, et al., The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored, Nucl. Acids Res. 39 (2011). D561-D8.
[23] E.R. Gamazon, W. Zhang, A. Konkashbaev, S. Duan, E.O. Kistner, D.L. Nicolae, et al., SCAN: SNP and copy number annotation, Bioinformatics 26 (2010) 259–262.
[24] R. Linding, Multivariate signal integration, Nat. Rev. Mol. Cell Biol. 11 (2010) 391.
[25] C.T. Pawson, J.D. Scott, Signal integration through blending, bolstering and bifurcating of intracellular information, Nat. Struct. Mol. Biol. 17 (2010) 653–658.
[26] A. Vazquez, R. Dobrin, D. Sergi, J.-P. Eckmann, Z. Oltvai, A.-L. Barabási, The topological relationship between the large-scale attributes and local interaction patterns of complex networks, Proc. Nat. Acad. Sci. 101 (2004) 17940–17945.
[27] G. Balazsi, A.-L. Barabási, Z. Oltvai, Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*, Proc. Nat. Acad. Sci. 102 (2005) 7841–7846.
[28] M.V. Rockman, L. Kruglyak, Genetics of global gene expression, Nat. Rev. Genet. 7 (2006) 862–872.
[29] M.T. Maurano, R. Humbert, E. Rynes, R.E. Thurman, E. Haugen, H. Wang, et al., Systematic localization of common disease-associated variation in regulatory DNA, Science 337 (2012) 1190–1195.
[30] D.E. Reich, M. Cargill, S. Bolk, J. Ireland, P.C. Sabeti, D.J. Richter, et al., Linkage disequilibrium in the human genome, Nature 411 (2001) 199–204.
[31] R.A. Gibbs, J.W. Belmont, P. Hardenbol, T.D. Willis, F. Yu, H. Yang, et al., The international HapMap project, Nature 426 (2003) 789–796.
[32] F. Murtagh, Complexities of hierarchic clustering algorithms: state of the art, Comput. Stat. Quart. 1 (1984) 101–113.
[33] J.M. Schwarz, C. Rödelsperger, M. Schuelke, D. Seelow, MutationTaster evaluates disease-causing potential of sequence alterations, Nat. Methods 7 (2010) 575–576.
[34] K.-C. Wong, Z. Zhang, SNPdryad: predicting deleterious non-synonymous human SNPs using only orthologous protein sequences, Bioinformatics btt769 (2014).
[35] M. Melé, P.G. Ferreira, F. Reverter, D.S. DeLuca, J. Monlong, M. Sammeth, et al., The human transcriptome across tissues and individuals, Science 348 (2015) 660–665.
[36] Y. Lee, E.R. Gamazon, E. Rebman, Y. Lee, S. Lee, M.E. Dolan, et al., Variants affecting exon skipping contribute to complex traits, PloS Genet. 8 (2012) e1002998.
[37] J. Monlong, M. Calvo, P.G. Ferreira, R. Guigó, Identification of genetic variants associated with alternative splicing using sQTLseekeR, Nat. Commun. (2014) 5.