Received: 19 February 2020 | Revised: 8 July 2020 | Accepted: 27 July 2020

DOI: 10.1002/mgg3.1456

ORIGINAL ARTICLE

Molecular Genetics & Genomic Medicine Open Access WILEY

Genome-wide pathogenesis interpretation using a heat diffusion-based systems genetics method and implications for gene function annotation

Yuan Quan[1,2] | **Qing-Ye Zhang**[2] | **Bo-Min Lv**[2] | **Rui-Feng Xu**[1] | **Hong-Yu Zhang**[2]

[1]School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen Graduate School, Shenzhen, China

[2]Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan, China

Correspondence
Yuan Quan, School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen 518055, China.
Email: quanyuan725@163.com

Hong-Yu Zhang, Hubei Key Laboratory of Agricultural Bioinformatics, College of Informatics, Huazhong Agricultural University, Wuhan 430070, China.
Email: zhy630@mail.hzau.edu.cn

Funding information
This work was supported by the National Natural Science Foundation of China (grant 31670779) and the Fundamental Research Funds for the Central Universities (Grant 2662017PY115).

Abstract

Background: Genetics is best dedicated to interpreting pathogenesis and revealing gene functions. The past decade has witnessed unprecedented progress in genetics, particularly in genome-wide identification of disorder variants through Genome-Wide Association Studies (GWAS) and Phenome-Wide Association Studies (PheWAS). However, it is still a great challenge to use GWAS/PheWAS-derived data to elucidate pathogenesis.

Methods: In this study, we used HotNet2, a heat diffusion-based systems genetics algorithm, to calculate the networks for disease genes obtained from GWAS and PheWAS, with an attempt to get deeper insights into disease pathogenesis at a molecular level.

Results: Through HotNet2 calculation, significant networks for 202 (for GWAS) and 167 (for PheWAS) types of diseases were identified and evaluated, respectively. The GWAS-derived disease networks exhibit a stronger biomedical relevance than PheWAS counterparts. Therefore, the GWAS-derived networks were used for pathogenesis interpretation by integrating the accumulated biomedical information. As a result, the pathogenesis for 64 diseases was elucidated in terms of mutation-caused abnormal transcriptional regulation, and 47 diseases were preliminarily interpreted in terms of mutation-caused varied protein-protein interactions. In addition, 3,802 genes (including 46 function-unknown genes) were assigned with new functions by disease network information, some of which were validated through mice gene knockout experiments.

Conclusions: Systems genetics algorithm HotNet2 can efficiently establish genotype-phenotype links at the level of biological networks. Compared with original GWAS/PheWAS results, HotNet2-calculated disease-gene associations have stronger biomedical significance, hence provide better interpretations for the pathogenesis of genome-wide variants, and offer new insights into gene functions as well. These results are also helpful in drug development.

KEYWORDS

drug discovery, gene function annotation, Genome-Wide Association Studies, pathogenesis, Phenome-Wide Association Studies, systems genetics

1 | INTRODUCTION

Interpreting the pathogenesis of various diseases and assigning biological functions on genes have been the central focus of modern genetic research. The rapid development of biological technology in the past decade has allowed the identification of gene-disease associations at the genome-wide level. Thanks to the wide use of single-nucleotide polymorphism (SNP) microarrays and next-generation DNA sequencing platforms (Altshuler, Daly, & Lander, 2008; Consortium IH, 2005), Genome-Wide Association Studies (GWAS) have been instrumental in identifying DNA sequence variants responsible for human genetic diseases or physiological traits of interest. GWAS have rapidly become a standard method for disease gene discovery (Cantor, Lange, & Sinsheimer, 2010; Visscher, Brown, McCarthy, & Yang, 2012; Wang, Barratt, Clayton, & Todd, 2005). In the last decade, over 4,200 GWAS-related papers have been published, which in total reported ~157,000 variant-trait associations, as listed in the GWAS catalog of the National Human Genome Research Institute (https://www.ebi.ac.uk/gwas/, September 2019) (Buniello et al., 2019). These variant-trait association data have proven useful for improving our scientific understanding of disease mechanisms and facilitating the diagnosis of diseases (Lander, 2011; Plenge, Scolnick, & Altshuler, 2013). Besides, researchers have developed an alternative strategy called Phenome-Wide Association Studies (PheWAS) to complement GWAS (Hebbring, 2014). In contrast to the phenotype-to-genotype approach in GWAS, PheWAS use a genotype-to-phenotype strategy to explore the variant-disease associations (Hebbring, 2014; Pendergrass et al., 2011; Pendergrass & Ritchie, 2015). PheWAS integrate the massive data in electronic health record (EHR) from the patient group with genotype data (Denny et al., 2010), and used International Classification of Disease Codes (ICD-9) to perform the phenotyping for calculating genotype-phenotype associations (Denny et al., 2010). Following the first PheWAS study published in 2010 (Denny et al., 2010), thousands of statistically significant variants associated with hundreds of human diseases have been identified over several years (Denny et al., 2013).

Although conventional GWAS and PheWAS have identified thousands of genotype-phenotype links, most of them are difficult for clarifying. This may be caused by several reasons. First, most of the disease-associated variants (80% or more) identified by GWAS or PheWAS are located in non-coding regions and their corresponding effector genes are difficult to determine (Manolio, 2013). Second, traditional GWAS or PheWAS usually use very strict *P*-values (~10^{-8}) as thresholds in the statistics in order to reduce false-positive rates. As a result, only low *P*-value SNPs that are truly associated with diseases can be screened out, leaving out the high *P*-value SNPs that may be also related to the diseases (Manolio et al., 2009). Together, it is still a great challenge to use the routine GWAS or PheWAS results to elucidate pathogenesis for genetic diseases and annotate gene functions.

Systems genetics is a thriving research area that intends to elucidate associations of genotype with complex phenotype from the perspective of biological networks (Civelek & Lusis, 2014). Of these networks, biological molecules are represented as nodes and the relationships between them are represented as edges. In the past few decades, with the rapid accumulation of omics data, the nodes in networks can be genes, mRNAs, proteins, metabolites, or a mixture of them (Civelek & Lusis, 2014). Among these networks, the gene- or protein-level networks were investigated most extensively.

Based on "guilty by association" rule, it can be speculated that if a gene is directly adjacent to a known pathogenic gene in the protein interaction network, then this gene will also be considered a potential pathogenic gene (Brun et al., 2003; Chua, Sung, & Wong, 2006; Hishigaki, Nakai, Ono, Tanigami, & Takagi, 2001; Oti, Snel, Huynen, & Brunner, 2006; Samanta & Liang, 2003; Schwikowski, Uetz, & Fields, 2000; Vazquez, Flammini, Maritan, & Vespignani, 2003). However, such a simple inference may lead to false-positive/negative predictions. That is, genes that are not related to certain diseases may be identified by biologically unimportant links within networks (Noble, Kuang, Leslie, & Weston, 2005). The genes that are truly associated with certain diseases may be missed due to their long connection paths to known pathogenic genes (Cowen, Ideker, Raphael, & Sharan, 2017).

To overcome the above problems, in recent years researchers have proposed a variety of efficient network construction strategies (Cowen et al., 2017). These methods include random walks, information diffusion, and electrical resistance, which essentially are variants of network propagation (Cowen et al., 2017). These models have been widely used in systems genetics, for gene/protein function prediction (Noble et al., 2005; Sharan, Ulitsky, & Shamir, 2007), biology module identification (Mitra, Carvunis, Ramesh, & Ideker, 2013), disease characterization (Cho, Kim, & Przytycka, 2012; Ideker & Sharan, 2008), and drug target prediction (Csermely, Korcsmáros, Kiss, London, & Nussinov, 2013). Recently, using the random walk with restart algorithm, Fang, De Wolf, Knezevic, Burnham, and Osgood (2019)

constructed an approach to prioritize potential drug targets based on the GWAS data. This strategy has been successfully applied to target discovery of 30 immune-related traits and has been validated by a variety of methods (Fang et al., 2019).

HotNet diffusion-oriented subnetworks (HotNet2) algorithm is one representative approach of network propagation (Leiserson et al., 2015). HotNet2 considers both the heats (reflecting genetic or biological importance) of individual genes and the topology of protein-protein interactions (PPIs) that is based on an insulated heat diffusion kernel algorithm (Leiserson et al., 2015). This method can reveal functionally interacted genes within disease networks, and the gene-disease associations obtained by these networks have relatively higher reliability. Thus, the corresponding effector genes of conventional GWAS- or PheWAS-derived non-coding variants can be determined efficiently. In addition, the consideration of PPI topology enables HotNet2 calculation to identify high *P*-value disease genes that are missed by conventional GWAS or PheWAS. Therefore, HotNet2 calculation is expected to overcome the limitations of routine GWAS and PheWAS. In this study, we first performed HotNet2 calculation on GWAS and PheWAS data to construct disease networks and evaluated the calculation results. Then, through analyzing GWAS-derived disease networks, we elucidated the pathogenesis for hundreds of genome-wide variants and annotated new functions for a variety of genes. In addition, because genetic disease genes are important sources of drug targets (Quan, Wang, Chu, & Zhang, 2018; Rastegar-Mojarad, Ye, Kolesar, Hebbring, & Lin, 2015; Sanseau et al., 2012; Wang & Zhang, 2013), the reliable gene-disease associations calculated by HotNet2 are very helpful for drug development.

2 | METHODS

2.1 | Ethical compliance

The study protocol was reviewed and approved by the Servicebio Institutional Animal Use Committee (IACUC) (animal ethics permit no. 18-0253). The protocol and experimental methods comply with the Helsinki Declaration.

2.2 | Gene mapping of PheWAS-derived variants

By the variant-to-gene mapping method proposed by Nelson et al., we first obtained the strongly linked variants of PheWAS-derived SNPs using LD analysis based on the 1000 Genomes Project pilot sequence genotypes for all the populations ($r^2 \geq 0.8$) (Consortium GP, 2010; Consortium IH, 2005; Nelson et al., 2015). Next, the genes that are potentially regulated by the PheWAS-derived loci were identified through the combinatorial use of different types of information, such as physical proximity to the gene (plus or minus 5 kb on the gene region), and eQTL information derived from eqtl.chicago.edu (http://eqtl.uchicago.edu/Home.html, accessed on 28 December 2015) and RegulomeDB (http://www.regulomedb.org/, accessed on 25, December 2015) (Boyle et al., 2012). It has been recognized that GWAS-derived SNPs are enriched in regulatory DNA regions marked by DHSs (Maurano et al., 2012). Considering the similarity between GWAS and PheWAS, we extracted the genomic coordinates of PheWAS-derived variants overlapped with DHS peaks recorded in RegulomeDB (Dehghan et al., 2009; Maurano et al., 2012; Nelson et al., 2015). For SNPs located in the transcription factor (TF)-binding sites (with RegulomeDB score ≤4), we obtained their host genes by determining whether an SNP falls inside a DHS peak region that is either located in the vicinity of the TSS or within an enhancer of a gene (detected by chromosome conformation capture methods including 5C and ChIA-PET; Maurano et al., 2012). Variants that can change the amino acid sequences encoding corresponding genes were obtained by Ensembl Variant Effect Predictor (http://asia.ensembl.org/info/docs/tools/vep/index.html). Finally, we only reserved genes that correspond to a determined UniProt ID (https://www.uniprot.org/) in Ensembl database (http://grch37.ensembl.org/index.html) for our analyses.

2.3 | Standardization of diseases

The Unified Medical Language System (UMLS), which includes a comprehensive set of medical concepts, was used to standardize disease annotation of genes. We used the natural language processing program MetaMap to convert disease terms of genes or indications of agents to the corresponding disease concepts (Aronson, 2001). We selected Medical Subject Headings (MeSH) as the vocabulary and limited the semantic type to "Pathologic Function," "Injury or Poisoning," and "Anatomical Abnormality" to obtain the disease-related concepts (Liu et al., 2014). MeSH classifies each disease to a narrow disease type using a hierarchical system. For example, "Alzheimer disease" is simply a broader term for "Alzheimer disease 15." All subtype disease concepts were converted to the appropriate broader term using a Perl module UMLS::Interface. Disease terms of genes that could not be mapped to any disease concept were excluded from subsequent analyses. Using the disease classes provided by Pharmaprojects, which cover 704 disease classes (similarity threshold: 0.75) (McInnes, Pedersen, & Pakhomov, 2009), 18,292 GWAS-derived disease genes (covering 617 types of diseases), and 7,212 PheWAS-derived disease genes (covering 296 types of diseases) were obtained. During this process,

we used the Lin to evaluate the disease term similarity of all disease concepts. Lin, which is calculated using the information content and path of concepts, has shown good performance for similarity measurement (Nelson et al., 2015). The Lin is calculated using the following equation:

$$Lin = \frac{IC(lcs)}{IC(\text{concept1}) + IC(\text{concept2})}$$

where *IC* is the negative log of the probability of the concept, which is pre-calculated by the Perl module by summing the probability of the concept occurring in some text plus the probability of its descendants occurring in some text, and *lcs* is the least common subsuming concept of concept1 and concept2.

The same standardized procedure was performed for the DisGeNET data, resulting in 725,589 pairs of gene-disease associations.

2.4 | Calculation of disease-related networks

In this study, we applied the HotNet2 algorithm to calculate the disease-related networks. The algorithm requires two types of data input: initial heat vectors of genes and PPI information. During the HotNet2 calculation, the negative logarithms of the *P*-values of GWAS-derived or PheWAS-derived loci were used as initial heat vectors for the corresponding disease genes. Considering the computational efficiency of HotNet2 (because too much gene input is difficult to obtain a significant disease network and time-consuming, the number of disease genes cannot be too large), we only collected disease genes with $P \leq 1 \times 10^{-5}$ as input for GWAS. When a gene-disease association corresponded to multiple *P*-values, the minimum *P*-value (maximal negative logarithm of *P*-value) was used as the initial heat vector. The PPI network was obtained from HINT, iRefIndex, and Multinet (Leiserson et al., 2015). The previously used parameters and procedures for HotNet2 calculation (https://github.com/raphael-group/hotnet2) were applied (Leiserson et al., 2015).

2.5 | Permutation test

To assess whether the identified TF-disease pairs were random results, a permutation test was performed. About 1,916 TF-disease pairs were generated by random shuffling between network-enriched 253 TFs with 186 types of disease. Then, we calculated the DisGeNET-supported frequency of 1,916 TF-disease pairs derived from the 10,000 random tests. The frequency distributions were compared with the real frequency of TF-disease pairs derived from the network enrichment.

2.6 | Calculation of protein complex binding energy

The wild-type FGB/FGA complex crystal structure was downloaded from protein data bank (PDB id: 3GHG) and the mutant-type protein complex structure (named mutant-3GHG) was built by virtual mutation and theoretical optimization using Discovery Studio 3.5. Then, the wild-3GHG and mutant-3GHG complex interaction energy changes were calculated by FiberDock (Mashiach, Nussinov, & Wolfson, 2010).

2.7 | Knockout experiment of mice

C10orf88 KO C57BL/6 mice (*C10orf88*+/-) and wild type C57BL/6 mice (WT) were prepared by Cyagen Biosciences. All mice used in these studies were 8–12 weeks of age.

2.8 | Western blotting assay and quantitative real-time PCR assay

The retina of mice was separated from the eyeball, total protein and RNA were extracted from retina using the RIPA buffer (Beytime), and the MiniBEST Universal RNA Extraction Kit (TaKaRa, Dalian, China) for western blotting assay and quantitative real-time PCR assay, respectively. The primary antibodies used in western blotting assay were as follows: rabbit polyclonal antibodies against GAPDH (Proteintech), VEGF (Proteintech), TGFB1 (Proteintech), and PRKCB2 (CST). The antibody of GADPH was used as an internal control. The primer sequences used in the qRT-PCR assay are listed in Table S17.

3 | RESULTS AND DISCUSSION

3.1 | Disease network calculation and evaluation

3.1.1 | Disease network calculation

In this study, GWAS data were collected from Systematic Target OPportunity assessment by Genetic Association Predictions (STOPGAP, https://github.com/StatGenPRD/STOPGAP) database, which contained 482,843 gene-disease associations after disease standardization (Table S1) (see Methods) (Shen, Song, Slater, Ferrero, & Nelson, 2017). To guarantee the reliability of PheWAS data, we selected Denny et al.'s research as the data source in this study, in which the phenotype-associated SNPs implicated by GWAS were regarded as mediators of human phenotypes (https://

phewascatalog.org/phewas) (Denny et al., 2013). According to the SNP-to-gene mapping procedure used by Nelson et al., (2015), 7,213 potential disease genes were identified from PheWAS data. Through standardization of diseases (see Methods), 296 disease phenotypes were obtained for PheWAS-derived disease genes, constituting 275,661 gene-disease associations (Table S2).

Next, we used the HotNet2 algorithm to construct disease networks for the GWAS- or PheWAS-identified disease genes (Figure 1) (see Methods). During HotNet2 calculation, the negative logarithms of *P*-values of GWAS- or PheWAS-derived SNPs were used as initial heat vectors for corresponding disease genes. The PPI networks were obtained from HINT, iRefIndex, and Multinet (Leiserson et al., 2015). Previously used parameters and procedures for HotNet2 calculation (https://github.com/raphael-group/hotnet2) were applied in this study (Leiserson et al., 2015). As a result, significant subnetworks for 202 types of diseases were successfully identified from GWAS data; more than 50% of the networks contained 21–50 proteins (Figure 2a and Table S3). For PheWAS data, 167 disease subnetworks were constructed by HotNet2; more than 40% of the networks constituted 51–100 proteins (Figure 2b and Table S4). Next, the HotNet2-calculated disease subnetworks were evaluated through analyzing their performance in disease gene enrichment, drug enrichment, and tissue-specific expression.

3.1.2 | Network evaluation by disease gene enrichment

Recently, Piñero et al., (2017) scored genetic disease genes by integrating information from multiple sources and recorded these scores in the database DisGeNET (http://www.disgenet.org). The 202 GWAS-derived disease networks defined 15,833 gene-disease pairs, 33.76% (5,346/15,833) of

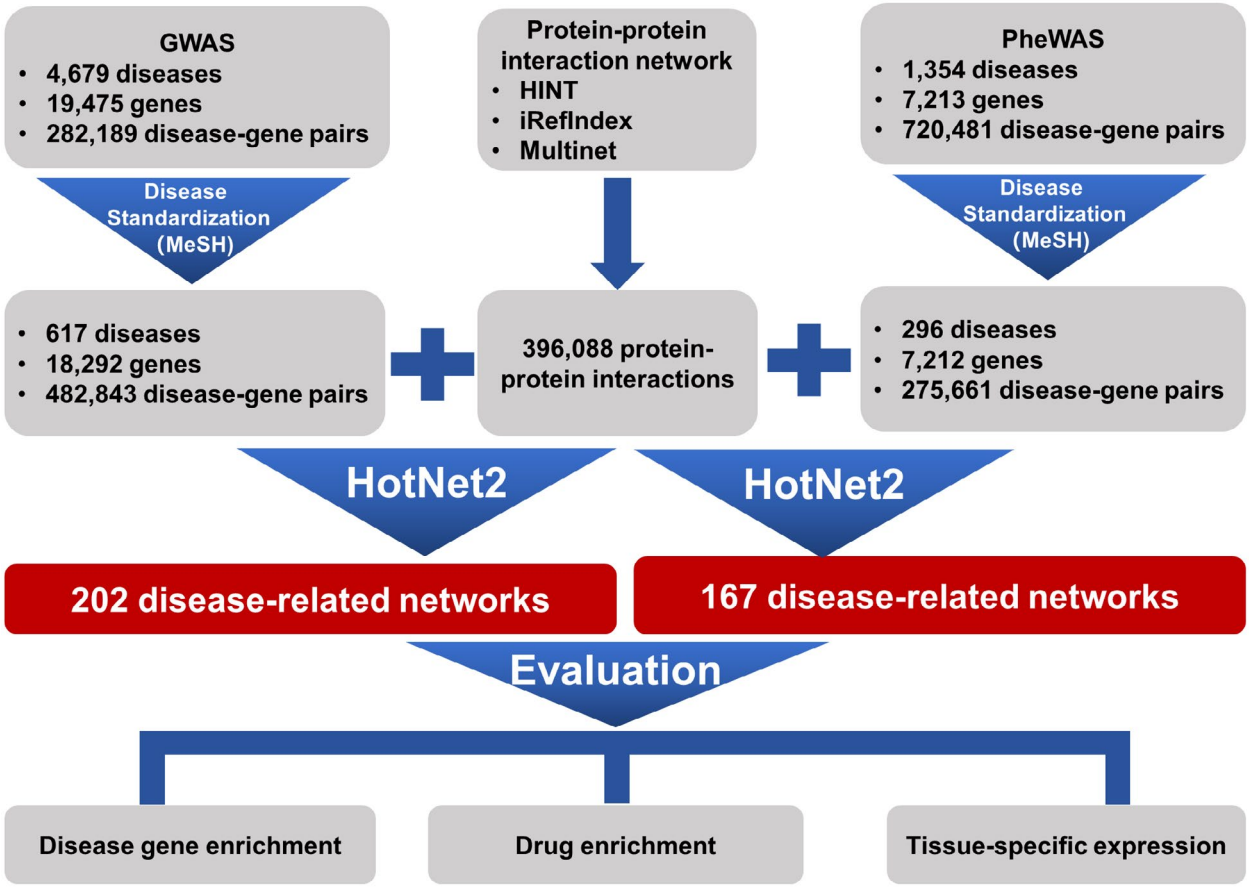

FIGURE 1 Data processing pipeline for systems genetics-based disease network construction and evaluation. GWAS-derived disease genes were collected from Systematic Target OPportunity assessment by Genetic Association Predictions (STOPGAP, https://github.com/StatGenPRD/STOPGAP) database and PheWAS data were derived from work by Denny et al., (2013). The genetic disease genes were standardized using MetaMap, where MeSH thesaurus was selected as the vocabulary source for UMLS. During HotNet2 calculation, negative logarithms of *P*-values of GWAS-derived or PheWAS-derived loci were used as initial heat vectors for the corresponding disease genes. The protein-protein interaction network was obtained from HINT, iRefIndex, and Multinet (Leiserson et al., 2015). As a result, significant subnetworks for 202 types of diseases were successfully identified from GWAS data, and for 167 types of diseases from PheWAS data. After that, the disease networks were evaluated by three different methods.

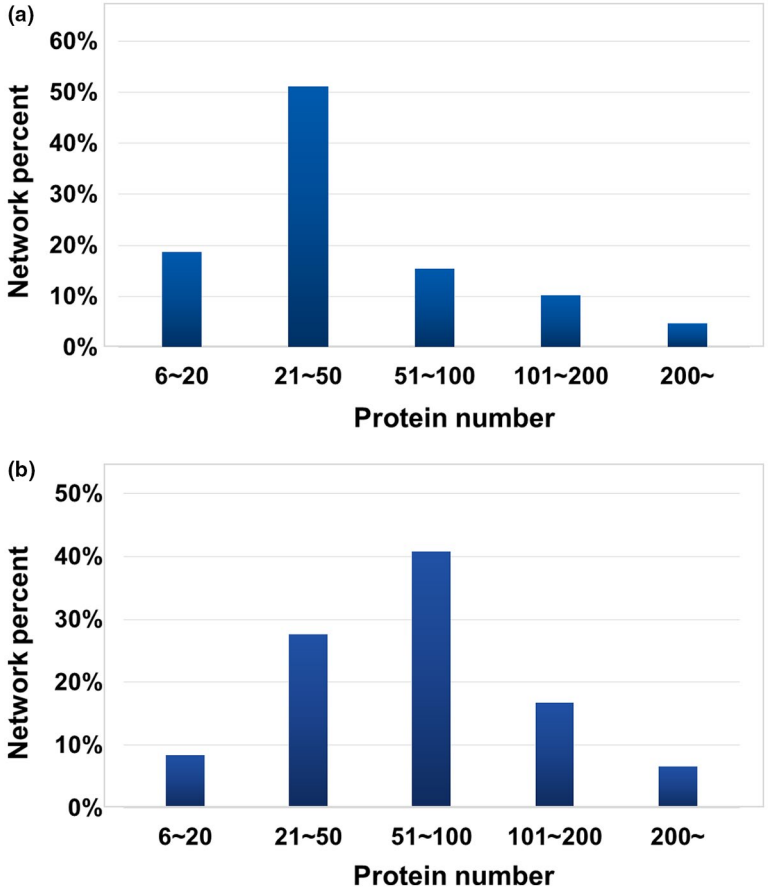

FIGURE 2 Distribution patterns of protein numbers of HotNet2-calculated networks. (a) for GWAS-derived disease networks; (b) for PheWAS-derived disease networks.

which have been documented in DisGeNET (Table S5). For 167 PheWAS-derived disease networks, 13,259 gene-disease pairs were defined, 15.91% (2,110/13,259) of which have been documented in DisGeNET (Table S6). In comparison, only 16.38% (79,101/482,843) of original GWAS-derived and 7.18% (19,780/275,661) of original PheWAS-derived disease genes are supported by DisGeNET records ($p = 0$ for GWAS, $p = 1.05 \times 10^{-270}$ for PheWAS, hypergeometric test) (Tables S5 and S6), suggesting that HotNet2-calculated disease networks established stronger links between genes and diseases. It is of great interest to note that the calculated networks contain some high *P*-value genes, which may be abandoned by the original GWAS or PheWAS (with *P*-values greater than 10^{-8}) but, indeed, are strongly associated with corresponding diseases according to the DisGeNET records (Tables S5 and S6). For instance, HotNet2-calculated networks of vitamin D deficiency involve the gene *PACS1* (OMIM accession number: 607492) and networks of encephalitis cover the gene *ATM* (OMIM accession number: 607585), which are truly associated with genetic diseases according to DisGeNET (with a confidence score of 0.2400 and 0.7087, respectively, much higher than the average score of all genes [0.0151]), but not significant in traditional GWAS ($p = 5.93 \times 10^{-6}$) or PheWAS ($p = 5.61 \times 10^{-3}$).

In addition, a large number of previous studies have shown that ohnolog genes, which are linked with whole-genome duplication events of human genome, are closely associated with human genetic disease because of their dose-sensitive property (Caspermeyer, 2017; Makino & McLysaght, 2010; Sekine & Makino, 2017; Xie, Yang, Wang, McLysaght, & Zhang, 2016). This property means that abnormal changes in gene copy number will induce phenotypic variety and disease occurrence. Based on this principle, we speculated that HotNet2-calculated disease genes will be enriched with ohnologs. From the work by Makino & McLysaght, (2010), 7,294 ohnolog genes were extracted, which accounts for 27.43% of the human genome. It was found that 2,219 (43.28%) of 5,127 genes covered by GWAS-derived networks and 1,196 (43.78%) of 2,732 genes involved in PheWAS-derived networks are ohnologs (Tables S5 and S6), exhibiting a strong enrichment of ohnologs in HotNet2-calculated disease networks ($p = 3.17 \times 10^{-165}$ for GWAS network, $p = 4.86 \times 10^{-84}$ for PheWAS network, hypergeometric test).

3.1.3 | Network evaluation by drug enrichment

Because the reliability of gene-disease associations is crucial for target identification and thus influences the efficiency of drug discovery (Quan et al., 2019), the calculated disease networks were further evaluated by drug enrichment analysis.

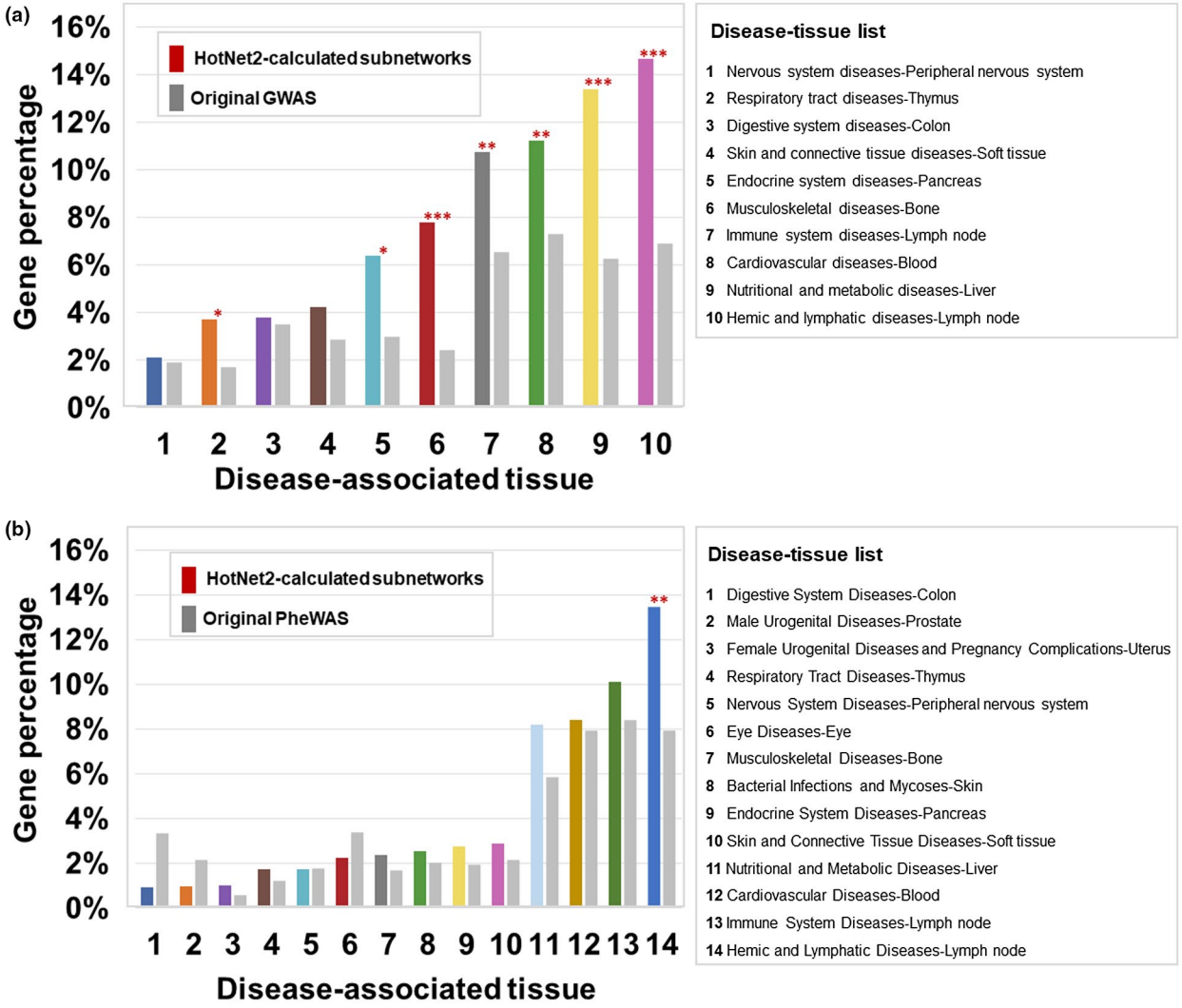

FIGURE 3 Tissue-specific gene enrichment for disease-related networks. We divided the standardized disease phenotypes into 24 categories according to the NCBI MeSH database (https://www.ncbi.nlm.nih.gov/mesh/). Only consider the disease categories that gene numbers were more than 150 because of the credibility of the statistics. The tissue-specific gene expression data were obtained from TiGER (http://bioinfo.wilmer.jhu.edu/tiger/). (a) For GWAS-derived networks, the gene expression for 7 of 10 disease categories showed more significant enrichment in relevant pathogenic tissues compared to the original GWAS data (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$, hypergeometric test); (b) For PheWAS-derived networks, the gene expression for 1 of 14 disease categories showed more significant enrichment in relevant pathogenic tissues compared to the original PheWAS data (*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$, hypergeometric test).

The information for drugs, targets, and related clinical activities were derived from the SCG-Drug database (http://zhanglab.hzau.edu.cn/scgdrug) (Quan et al., 2019). SCG-Drug integrates associations between chemical agents and their corresponding targets from the Drug-Gene Interaction database (DGIdb, http://dgidb.genome.wustl.edu/), Therapeutic Target Database (TTD, http://bidd.nus.edu.sg/group/cjttd/), and DrugBank (http://www.drugbank.ca/). And the agent clinical activity annotations of this database were collected from TTD, DrugBank, and ClinicalTrials (http://clinicaltrials.gov/). The clinical activities of the agents were standardized using MetaMap (Aronson, 2001; Quan et al., 2019). From GWAS-derived disease subnetworks, we obtained 22,424 potential agent-disease pairs. About 14.53% (3,258/22,424) of these pairs were supported by clinical tests (Table S7), significantly higher than the supported ratio of original GWAS-derived agents (6.42%, $p = 0$, hypergeometric test). For PheWAS-derived networks, although only 3.74% (758/20,286) of potential agent-disease pairs were supported by clinical tests (Table S8), their performance is also better than the original PheWAS (3.45%, $p = 8.14 \times 10^{-3}$, hypergeometric test). These results strongly supported the effectiveness of HotNet2 calculation in enriching disease genes.

3.1.4 | Network evaluation by tissue-specific expression

Finally, we examined the tissue-specific expression of network-containing genes. The standardized disease phenotypes

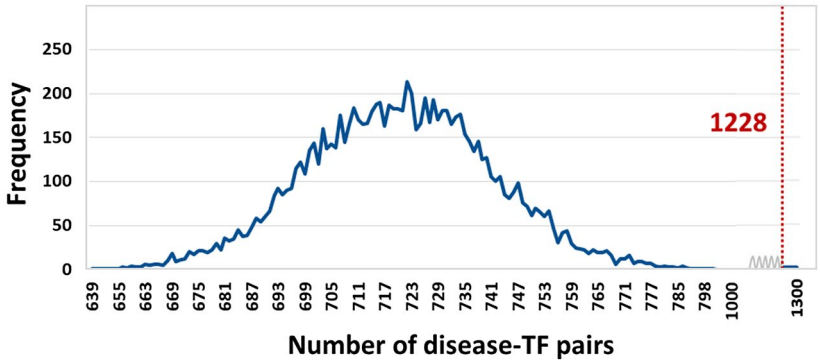

FIGURE 4 The frequency of evidence-supported TF-disease pairs. We obtained TF-target regulatory relationships from TRRUST (http://www.grnpedia.org/trrust/), PAZAR (http://www.pazar.info/), and AnimalTFDB (http://bioinfo.life.hust.edu.cn/AnimalTFDB/). Using the hypergeometric test, 202 disease-related networks were enriched to 253 TFs, resulting in 1,916 potential TF-disease pairs. Through comparing gene-disease relationships recorded in the comprehensive database DisGeNET (http://www.disgenet.org), 64.09% (1,228/1,916) of potential TF-disease pairs can be supported by this database. By permuting TF-disease pairs 10,000 times and again cross-referencing random TF-disease pairs, this supported ratio is only 39.58% (720 ± 81).

were divided into 24 categories according to the NCBI MeSH database (https://www.ncbi.nlm.nih.gov/mesh/). Each disease category corresponds to a specifically defined pathogenic tissue. Through retrieving the tissue-specific gene expression database (TiGER, http://bioinfo.wilmer.jhu.edu/tiger/) (Liu, Yu, Zack, Zhu, & Qian, 2008), we can calculate the proportion of a disease gene set specifically expressed in a certain tissue. Here, for every disease category, we performed the enrichment analysis by comparing the tissue-specific proportions of gene sets contained in the GWAS- or PheWAS-derived networks with the gene sets of original GWAS or PheWAS data. As a result, it was found that 7 of 10 disease categories have significant enrichment in relevant pathogenic tissues for GWAS-derived disease networks ($p < 0.05$, hypergeometric test) (Figure 3a), validating the HotNet2 calculation on GWAS data. In comparison, the PheWAS-derived networks exhibit a worse performance. Only 1 of 14 disease categories have significant enrichment in relevant pathogenic tissues (Figure 3b).

In summary, the above results demonstrated that HotNet2 calculation can efficiently enrich disease-associated genes for both GWAS and PheWAS data. In addition, GWAS-derived networks perform better than PheWAS-derived counterparts in various evaluations. The reason for this phenomenon may be that the performance of HotNet2 calculation strongly depends on the quality of initial heat vectors, which is defined by the reliability of the gene-disease associations. The good performance of GWAS-derived networks can be largely attributed to the high quality of original GWAS results, which, indeed, display more biomedical relevance than original PheWAS results (see the above analysis). Therefore, in the following part of this study, GWAS-derived disease networks are used for pathogenesis interpretation and gene function annotation.

3.2 | Genome-wide pathogenesis interpretation

3.2.1 | Data profile

The HotNet2-identified disease genes from GWAS data correspond to 5,716 SNPs, of which 20.33% (1,162) are original GWAS SNPs and the rest belong to linkage disequilibrium (LD) SNPs (Table S9). Based on SNP annotations retrieved from Ensembl (Ensembl genome browser 84, http://www.ensembl.org/index.html), it was found that most (70.10%) of these SNPs are *trans* variants. The others are *cis*-SNPs, which fall inside intronic regions (accounting for 19.56%), upstream or downstream of genes (8.85%), coding regions (3.73%), 5′ or 3′ untranslated regions (1.31%), and noncoding transcript regions (0.21%; Table S9).

3.2.2 | Pathogenesis interpretation for trans variants and implications for drug repositioning

The pathogenesis of the *trans* variants is expected to be interpreted in terms of transcriptional regulation. As is well known, transcription factors (TFs) and regulatory elements (such as promoters and enhancers) constitute gene regulatory networks, which influence the transcription of genes and play a key role in biological regulation (Davidson et al., 2002; Walhout, 2006). A mutation within a given transcriptional regulatory region is possible to either weaken or strengthen the binding affinity of TFs, leading to downregulation or upregulation of the related genes. These abnormal expressions can cause serious diseases (Fuxman Bass et al., 2015). Therefore, to understand the links between non-coding

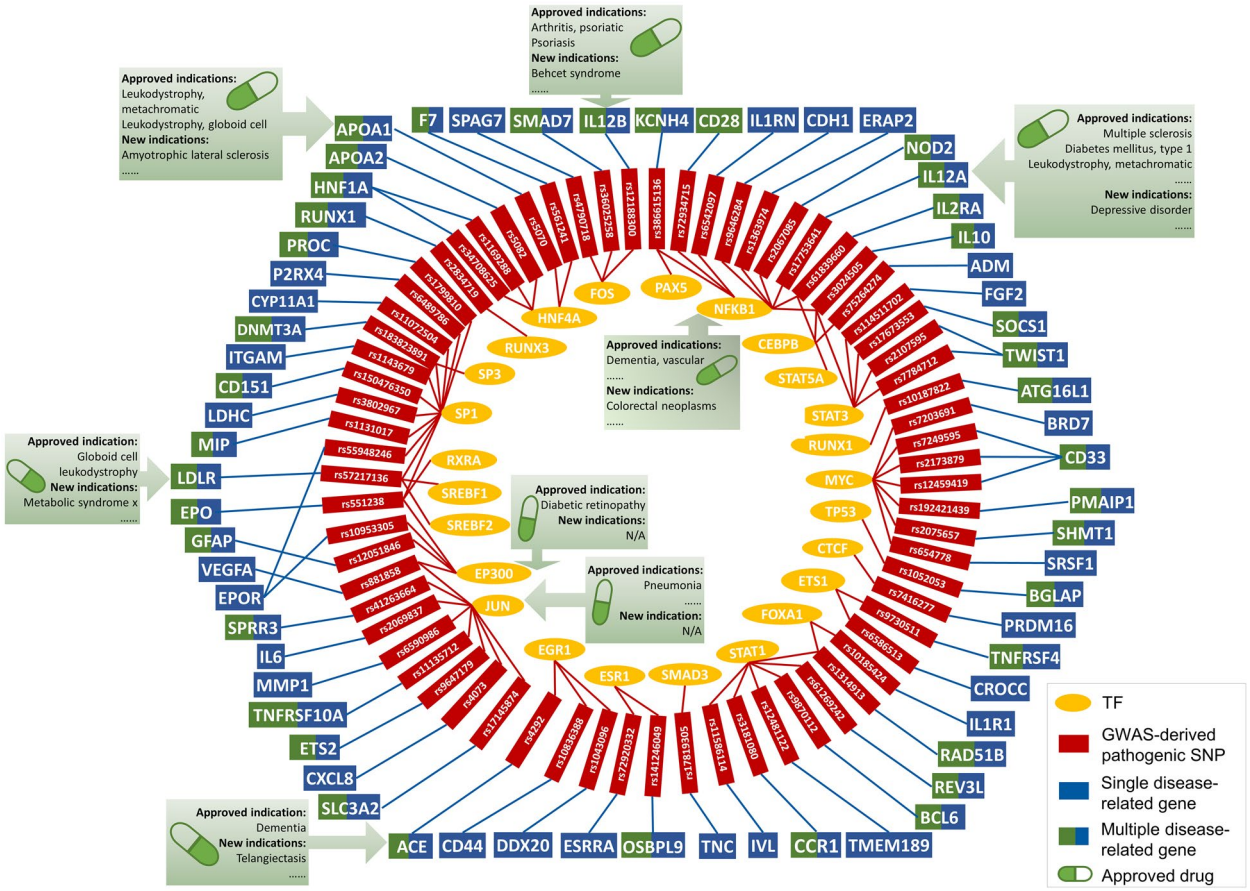

FIGURE 5 Pathogenesis interpretation for trans SNPs by TF-SNP-gene trios. By searching the TF-target gene database, 1,916 TF-disease pairs were identified, 64.09% (1,228/1,916) of which are supported by DisGeNET. Then, 25 TFs were identified for 69 GWAS-derived pathogenic loci (including LD SNPs) from the database RegulomeDB (http://www.regulomedb.org/). Together, there were 57 groups of TF-SNP-gene trios that had coherent genetic annotations, and some trios are responsible for multiple diseases. The latter has direct implications for drug repositioning through targeting TFs (e.g., NFKB1, EP300, JUN) or downstream genes (e.g., IL12B, IL12A, APOA1, LDLR, ACE).

variants and diseases, TFs were primarily enriched for the 202 calculated disease networks. Through searching TF-target gene databases, such as TRRUST (http://www.grnpedia.org/trrust/) (Han et al., 2018), PAZAR (http://www.pazar.info/; Portales-Casamar et al., 2007), and AnimalTFDB (http://bioinfo.life.hust.edu.cn/AnimalTFDB/; Hu et al., 2019), 253 TFs were successfully identified for 186 types of disease networks, resulting in 1,916 TF-disease pairs. The identified TFs were validated by functional analysis, which showed that 64.09% (1,228/1,916) of TF-disease pairs have been documented in DisGeNET (Table S10). However, if the TFs were randomly assigned with diseases (for 10,000 times), the DisGeNET-supported pairs got significantly rarer (720 ± 81) ($p < 1\times10^{-4}$, Permutation test; Figure 4; see Methods). These results demonstrated that these TFs are disease-relevant and deserved further investigation.

Moreover, in living organisms, many TFs perform their regulatory functions by interacting with other TFs to form complexes (Zhu, Shendure, & Church, 2005). Therefore, we speculated that TFs that regulate the same disease are more likely to interact with each other. Of the above network-enriched TFs, there are 6,405 TF-TF pairs that potentially regulate the same disease-related networks, covering 253 TFs and 181 diseases (Table S11). By querying functional protein association networks database STRING (https://string-db.org/; Franceschini et al., 2013), 739 of 6,405 (11.54%) TF-TF pairs have physical interactions that are supported by experimental evidence (experimental confident score >400) (Table S11). This supported ratio is only 5.78% (1,843/31,878) for random TF-TF pairs, which is significantly lower than network-enriched TF-TF pairs ($p = 4.58 \times 10^{-92}$, hypergeometric test). This result implied that TFs, indeed, tend to form complexes through physical interactions that result in regulation of the same disease, indicating the usefulness of systems genetics to elucidate pathogenesis underlying genetic diseases.

Then, we retrieved the SNP-TF binding information from the database RegulomeDB (http://www.regulomedb.org/) (Boyle et al., 2012). As a result, 25 TFs were identified for 69 GWAS-derived pathogenic SNPs (including LD SNPs), regulating 64 genes (Figure 5, Table S12). Together, there were 57 groups of TF-SNP-gene trios that have coherent

genetic annotations, covering 64 types of diseases (Table S12). These results are helpful to elucidate the pathogenesis of GWAS-derived loci in terms of transcriptional regulation. For instance, SNP NC_000006.11:g.43806609G>A (dbSNP identifier: rs881858) has been reported to induce glomerulonephritis by GWAS (Köttgen et al., 2010). The present analysis reveals the binding relationship between rs881858 and activator E1A binding protein p300 (EP300). Using the deltaSVM method (Lee et al., 2015), we can predict that this SNP will strengthen the binding of EP300 and thus upregulate the expression of the downstream target gene-vascular endothelial growth factor A (*VEGFA*, OMIM accession number: 192240) (Figure 6). This inference is, indeed, supported by the experimental observation that *VEGFA* is linked with glomerulonephritis through upregulated expression (Abe-Yoshio et al., 2008).

It is interesting to note that some trios are responsible for multiple diseases. For instance, TF (SREBF1)-SNP (NC_000019.9:g.11201124T>C, dbSNP identifier: rs57217136)-gene (*LDLR*, OMIM accession number: 606945) trio is linked with globoid cell leukodystrophy and metabolic syndrome X. Because LDLR has served as a successful drug target for the treatment of globoid cell leukodystrophy, it is reasonable to infer that the targeted drugs, for example, atorvastatin, also have the potential to combat metabolic syndrome X, which is, indeed, validated by clinical evaluation (according to the annotations in SCG-Drug). This drug repositioning method could be used for other TF-SNP-gene trios. In total, there were 15 TF-SNP-gene-drug quartets with consistent genetic annotations and approved therapeutic activities (Figure 5). According to the additional genetic implications of TF-SNP-gene trios, we assigned approved drugs that target TFs or genes with new activities, resulting in 62 new drug-disease associations. Based on the SCG-Drug database, 19 (30.65%) of these predictions are supported by clinical tests (Table S13).

3.2.3 | Pathogenesis interpretation for coding regions and implications for drug repositioning

As to the pathogenesis interpretation for *cis* variants, the GWAS-derived loci in coding regions are of apparent interest, which comprise 205 SNPs, including 146 missense and 59 synonymous mutations (Table S9). For missense mutations, we speculated that these mutations may cause diseases by affecting the PPI of networks. Through searching Protein Data Bank (PDB, https://www.rcsb.org/) (Sussman et al., 1998), 32 mutations can be mapped onto 3D protein structures. Using structure analysis with Molecular Operating Environment (MOE) simulation software (Vilar, Cozza, & Moro, 2008), we found that 27 mutations (corresponding to 47 diseases) were located in the surface area of proteins, whereas only five mutations were in the protein core. Therefore, the pathogenesis of 27 GWAS-identified mutations may be elucidated in terms of protein-protein interaction (Table S14). For example, the missense variant NP_005132.2:p.Arg478Lys (dbSNP identifier: rs4220) is located in the exon region of fibrinogen beta chain (*FGB*, OMIM accession number: 134830) and causes an Arg478Lys mutation, which is responsible for von Willebrand disease (Dehghan et al., 2009). In the calculated von Willebrand disease network, FGB interacts with fibrinogen alpha chain (FGA, OMIM accession number: 134820). The crystal structure for FGB/FGA complex is available in the PDB. Through calculating the binding energy of FGB and FGA and the influence of Arg478Lys mutation (see Methods), we found that the mutation resulted in an energy loss in the FGB-FGA interaction (Figure 7). This is consistent with the previous report that the loss of FGA function leads to von Willebrand disease (Kunicki et al., 2004).

Although synonymous mutation does not change the amino acid sequence, it may alter the codon usage that is positively

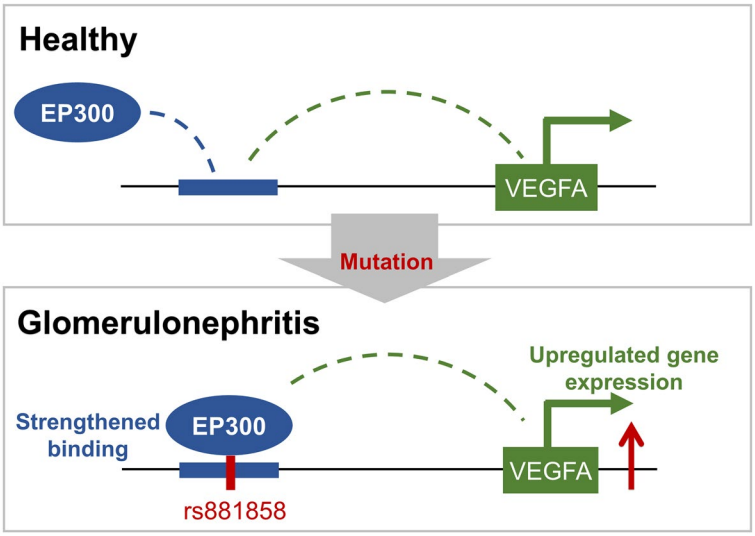

FIGURE 6 Pathogenesis interpretation for SNP rs881858. Glomerulonephritis-related SNP rs881858 can strengthen the binding of activator E1A binding protein p300 (EP300) and thus upregulate the expression of the downstream gene vascular endothelial growth factor A (*VEGFA*). *VEGFA* is, indeed, linked with glomerulonephritis through its upregulated expression.

correlated with tRNA abundance (Kanaya, Yamada, Kudo, & Ikemura, 1999; Percudani, Pavesi, & Ottonello, 1997; Qian, Yang, Pearson, Maclean, & Zhang, 2012). Therefore, if a mutation alters the codon usage from low to high frequency, it will lead to enhanced expression of corresponding proteins (GOF) and vice versa (LOF) (Duret, 2002; Goetz & Fuglsang, 2005). For instance, synonymous mutation NC_000017.10:g.61566031G>A (dbSNP identifier: rs4343) introduces an ACG-to-ACA mutation in angiotensin I converting enzyme (*ACE*, OMIM accession number: 106180). Given that inhibiting ACE (e.g., by enalapril) can treat hypertension (in SCG-Drug records), we predicted that rs4343 leads to the upregulation of *ACE* expression. Interestingly, for the rs4343 mutation, the codon frequency is significantly elevated from 6.1 (ACG) to 15.1 (ACA) (per thousand) according to Codon Usage Database (http://www.kazusa.or.jp/codon/) (Nakamura, Gojobori, & Ikemura, 1999), which implies an increased expression of *ACE* and is consistent with our prediction. In addition, asthma-associated synonymous mutation NG_016779.1:g.5169 T>C (dbSNP identifier: rs2069763) caused the codon of *Interleukin 2* (*IL2*, OMIM accession number: 147680) *t*o change from TCT to TCC, and the corresponding codon usage frequency increased from 15.2 to 17.7 (per thousand) (Nakamura et al., 1999). Therefore, we predicted that this mutation will result in the upregulated expression of *IL2* to cause asthma. According to the diseases associated over-expressed and under-expressed gene database (OUGene, http://www.csbio.sjtu.edu.cn/bioinf/OUGene/) (Pan & Shen, 2016), *IL2*, indeed, leads to asthma by upregulating its expression. The above results demonstrated that it is feasible to explore the pathogenic mechanism of synonymous mutations based on codon usage preferences.

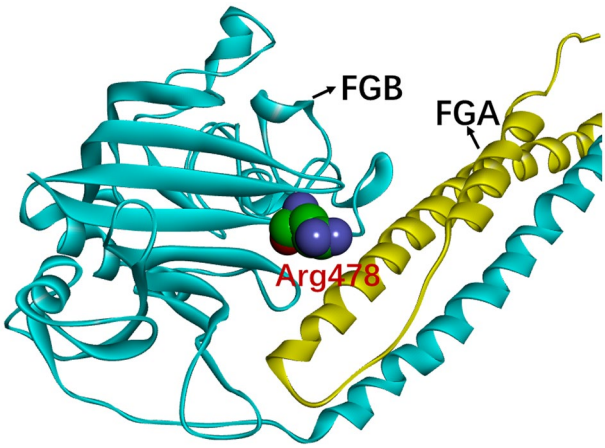

FIGURE 7 Pathogenesis interpretation for SNP rs4220 in terms of protein-protein interaction. The von Willebrand disease-related SNP rs4220 causes an Arg478Lys mutation in the fibrinogen beta chain (FGB). In the calculated von Willebrand disease network, FGB interacted with the fibrinogen alpha chain (FGA). Calculating the binding energy of FGB and FGA and the influence of Arg478Lys mutation, we found that the mutation caused an energy loss in the FGB-FGA interaction (the binding energy score of FiberDock increases from −273.74 to −270.57). This is consistent with a previous report that the loss of function of FGA leads to von Willebrand disease (Kunicki et al., 2004).

The present results are also valuable for drug repositioning. The disease-associated SNPs located in protein-coding regions correspond to 200 genes, 124 of which are involved in multiple disease networks and 18 have been successfully used as drug targets. Therefore, the drugs targeting these genes can be assigned to new functions. For instance, apolipoprotein E (*APOE*, OMIM accession number: 107741) is involved in both metachromatic leukodystrophy and telangiectasis networks. The agent targeting APOE (i.e., simvastatin) has been approved for treating metachromatic leukodystrophy. It is thus inferred that simvastatin might be used for the treatment of telangiectasis, which has, indeed, been validated in clinical trials (according to SCG-Drug records). Using this method, we predicted 39 new agent-disease associations, 13 (33.33%) of which are supported by clinical trials or literature (Table S15).

3.3 | Gene function annotation

3.3.1 | Principle

The above analyses demonstrated the stronger reliability of GWAS-derived disease networks identified by systems genetics methods. Therefore, we can infer the gene functions from the disease information of networks. That is, if a gene is involved in a disease network, we can predict that this gene is associated with the disease. Using this method, new functions for 3,802 genes (including 3,756 common genes and 46 function-unknown genes) were annotated (Table S16).

3.3.2 | New function annotation for common genes

The networks have identified some new disease associations for 3,756 common genes, which have not yet been documented in SCG-drug. SCG-drug is a very comprehensive database that integrates eight representative disease gene databases (Quan et al., 2019), such as Genetic Association Database (GAD, https://geneticassociationdb.nih.gov/), Online Mendelian Inheritance in Man (OMIM, http://omim.org/), Clinvar (http://www.ncbi.nlm.nih.gov/clinvar/), Orphanet (http://www.orpha.net/consor/cgi-bin/index.php), DisGeNET (http://www.disgenet.org/web/DisGeNET/menu/rdf), INtegrated TaRget gEne PredItion (INTREPID), GWASdb (http://jjwanglab.org/gwasdb), and The Human Gene Mutation Database (HGMD, http://www.hgmd.cf.ac.uk/ac/index.php) (Table S16). Therefore, the predicted new

functions are of apparent genetic interest. For example, the gene *histone deacetylase 6* (*HDAC6*, OMIM accession number: 300272) is associated with melanoma and prostatic neoplasms according to DisGeNET records (Piñero et al., 2017). However, the HotNet2 calculation indicated that this gene is included in the osteoporosis-related network. It is thus predicted that *HDAC6* is also associated with osteoporosis. Indeed, according to the Mouse Genome Informatics database's annotations (MGI, http://www.informatics.jax.org/) (Shaw, 2016), knocking out the gene *HDAC6* for mice can result in the phenotype of increased bone mineral content. The reverse phenotype (decreased bone mineral content) of this phenotype is, indeed, one of the important symptoms in osteoporosis. In addition, many new-predicted gene functions, such as dwarfism for gene *hedgehog acyltransferase* (*HHAT*, OMIM accession number: 605743), pulmonary emphysema for gene *aminoacyl tRNA synthetase complex interacting multifunctional protein 1* (*AIMP1*, OMIM accession number: 603605), pneumonia for gene *ubiquitin specific peptidase 25* (*USP25*, OMIM accession number: 604736), hypertension for gene *guanylate cyclase 1 soluble subunit beta 1* (*GUCY1B3*, OMIM accession number: 139397), sarcoma for gene *RAN binding protein 2* (*RANBP2*, OMIM accession number: 601181), and blood coagulation disorders for gene *protein C*, *inactivator of coagulation factors Va and VIIIa* (*PROC*, OMIM accession number: 612283), have been validated by gene knockout experiments of mice (as annotated in MGI database) (Shaw, 2016).

3.3.3 | Function annotation for function-unknown genes

Using the same principle, we further predicted functions for 46 function-unknown genes, covering 67 diseases (Table S16). For instance, gene *chromosome 10 open reading frame 88* (*C10orf88*) was predicted to be associated with retina-related diseases (including macular edema, retinitis pigmentosa, retinal diseases, and diabetic retinopathy). To test this inference, we tried to establish a *C10orf88* knockout mouse model. However, the knockout of *C10orf88* induced homozygous mice to die during the embryonic period, therefore only heterozygote was obtained and was used in the following test. Neovascularization is a pathological

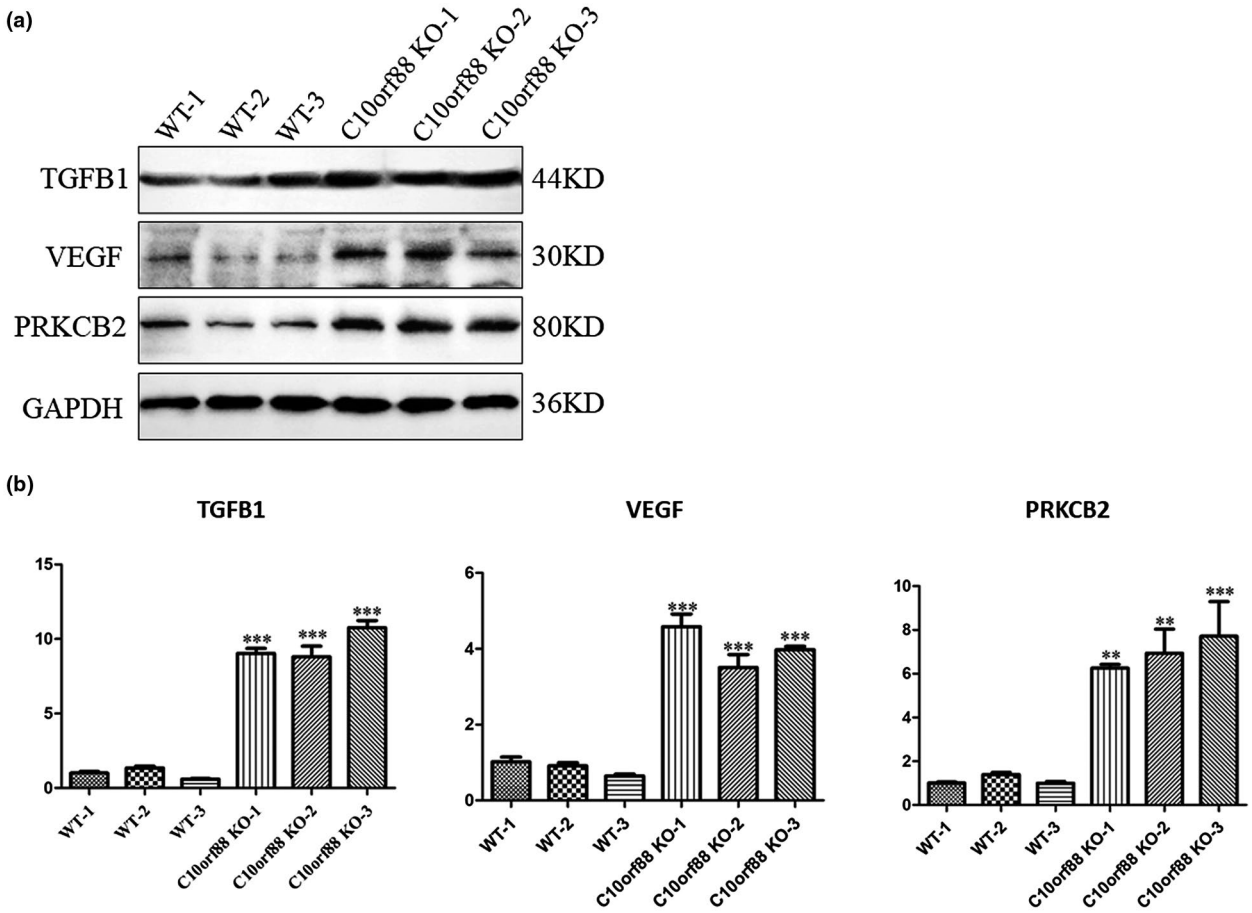

FIGURE 8 The expression of *TGFB1*, *VEGF*, and *PRKCB2* in *C10orf88*+/- mice and WT mice retinal. (a) The protein level of *TGFB1*, *VEGF*, and *PRKCB2* in *C10orf88*+/- mice and WT mice retinal measured by Western blot; (b) The mRNA level of *TGFB1*, *VEGF*, and *PRKCB2* in *C10orf88*+/- mice and WT mice retinal measured by qRT-PCR.

hallmark of numerous retinal diseases, from diabetic retinopathy (DR) to age-related related macular degeneration (AMD). *Vascular endothelial growth factor* (*VEGF*, OMIM accession number: 192240) is an angiogenic factor involved in collateral vessel formation, inflammation, tumor progression, and retinal conditions (Campbell & Doyle, 2019). Anti-VEGF drugs have been widely used for the intravitreal treatment of retinal conditions (Pham et al., 2019). *protein kinase C beta 2* (*PRKCB2*, OMIM accession number: 176970) is primarily involved in mediating a variety of functional and structural abnormalities in vascular tissues. In diabetic retinopathy, both *PRKCB2* and *VEGF* are considered very important parameters to do the diagnosis of microangiopathy (Kumar et al., 2012). The loss of retinal pericytes is another hallmark of retinal diseases. The expression of gene *transforming growth factor beta 1* (*TGFB1*, OMIM accession number: 190180) could induce TGFB-induced Gene Human Clone 3 (*BIGH3*) to promote pericyte apoptosis (Betts-Obregon et al., 2016). In this study, the protein and mRNA levels of *VEGF*, *TGFB1*, and *PRKCB2* were measured by Western blot and qRT-PCR, respectively (see Methods). As shown in Figure 8a,b, the expression of *VEGF*, *TGFB1*, and *PRKCB2* were higher in *C10orf88*+/− mice than WT mice, which are biomarkers of retinal-related diseases (Betts-Obregon et al., 2016; Choi et al., 2017; Wallace, 2016). These results indicated that the knockout of gene *C10orf88* could change the vascular microcirculation and induce inflammation reaction in retinal. Together, the association of *C10orf88* with retina-related diseases was preliminarily verified.

4 | CONCLUSION

In summary, the rapid progress in GWAS and PheWAS has accumulated thousands of disease-susceptible loci. However, it remains a great challenge to use these data to get new insights into pathogenesis interpretation and gene function annotation. In this study, we demonstrate that systems genetics methods, for example, HotNet2 algorithm, can efficiently facilitate the interpretation of traditional genetics-derived disease loci and the annotation of genetics-derived disease genes. And because of the higher quality of GWAS data, our results showed that GWAS-derived disease networks have a stronger biological significance compared to PheWAS. Additionally, because systems genetics establishes genotype-phenotype relationships at the level of biological networks rather than single elements, it is expected to find applications in overcoming other limitations of traditional GWAS and PheWAS. Furthermore, HotNet2-calculated disease networks also contribute to drug discovery. Since the HotNet2 algorithm only needs to input both initial heat and PPI data, this methodology can be readily extended to other genetic data with the strength information of gene-disease associations. However, our study still has several limitations at this stage. First, it is not appropriate to input too large number of initial disease genes in the HotNet2 calculation, which may cause some causative genes whose GWAS or PheWAS-derived P-values are too high to be abandoned in theory. Second, because this study covers various disease categories, we only chose the broad-spectrum human PPIs as the initial input in HotNet2 calculations. It is well known that the organs and tissues corresponding to various diseases are different. Therefore, if researchers can use organ- or tissue-specific PPIs as HotNet2 input in a certain disease focused-research based on prior knowledge, it is possible to identify the disease networks more in line with biological reality.

5 | COMPETING INTERESTS

The authors declare that they have no competing interests.

ACKNOWLEDGMENTS

We are grateful to Zhi-Hui Luo for his assistance in disease standardization.

AUTHORS' CONTRIBUTIONS

H.-Y.Z. conceived and designed the study, Y.Q. ran the computational pipeline, Y.Q., Q.-Y.Z., and H.-Y.Z. analyzed the data, interpreted the results, and wrote the paper. B.-M.L. was responsible for mice and cell experiments also participated in the writing of the paper. R.-F.X. helped in preparing the manuscript. All authors revised and approved the final manuscript.

DATA AVAILABILITY STATEMENT

The complete GWAS data are available from the STOPGAP database (https://github.com/StatGenPRD/STOPGAP). The complete PheWAS data were derived from work by Denny *et al*. (https://phewascatalog.org/phewas) (Denny et al., 2013). The disease gene scores of DisGeNET are obtained from http://www.disgenet.org. The information for drugs, targets, and related clinical activities was derived from the SCG-Drug database (http://zhanglab.hzau.edu.cn/scgdrug). The basic information on the various databases involved in this study is available in Table S18.

REFERENCES

Abe-Yoshio, Y., Abe, K., Miyazaki, M., Furusu, A., Nishino, T., Harada, T., … Kohno, S. (2008). Involvement of bone marrow-derived endothelial progenitor cells in glomerular capillary repair in habu snake venom-induced glomerulonephritis. *Virchows Archiv*, *453*(1), 97–106. https://doi.org/10.1007/s00428-008-0618-5

Altshuler, D., Daly, M. J., & Lander, E. S. (2008). Genetic mapping in human disease. *Science*, *322*(5903), 881–888. https://doi.org/10.1126/science.1156409

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium*, *2001*, 17–21.

Betts-Obregon, B. S., Mondragon, A. A., Mendiola, A. S., LeBaron, R. G., Asmis, R., Zou, T., … Tsin, A. T. (2016). TGFβ induces BIGH3 expression and human retinal pericyte apoptosis: a novel pathway of diabetic retinopathy. *Eye*, *30*(12), 1639–1647. https://doi.org/10.1038/eye.2016.179

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., … Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Research*, *22*(9), 1790–1797. https://doi.org/10.1101/gr.137323.112

Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guenoche, A., & Jacq, B. (2003). Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biology*, *5*(1), R6.

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., … Parkinson, H. (2019). The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, *47*(D1), D1005–1012. https://doi.org/10.1093/nar/gky1120

Campbell, M., & Doyle, S. L. (2019). Current perspectives on established and novel therapies for pathological neovascularization in retinal disease. *Biochemical Pharmacology*, *164*, 321–325. https://doi.org/10.1016/j.bcp.2019.04.029

Cantor, R. M., Lange, K., & Sinsheimer, J. S. (2010). Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *American Journal of Human Genetics*, *86*(1), 6–22. https://doi.org/10.1016/j.ajhg.2009.11.017

Caspermeyer, J. (2017). The estimation of Alzheimer's disease causative genes by applying an evolutionary approach to medicine. *Molecular Biology and Evolution*, *34*(9), 2425–2426. https://doi.org/10.1093/molbev/msx201

Cho, D.-Y., Kim, Y.-A., & Przytycka, T. M. (2012). Chapter 5: network biology approach to complex diseases. *PLoS Computational Biology*, *8*(12), e1002820. https://doi.org/10.1371/journal.pcbi.1002820

Choi, J. A., Chung, Y. R., Byun, H. R., Park, H., Koh, J. Y., & Yoon, Y. H. (2017). The anti-ALS drug riluzole attenuates pericyte loss in the diabetic retinopathy of streptozotocin-treated mice. *Toxicology and Applied Pharmacology*, *315*, 80–89. https://doi.org/10.1016/j.taap.2016.12.004

Chua, H. N., Sung, W. K., & Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, *22*(13), 1623–1630. https://doi.org/10.1093/bioinformatics/btl145

Civelek, M., & Lusis, A. J. (2014). Systems genetics approaches to understand complex traits. *Nature Reviews Genetics*, *15*(1), 34–48. https://doi.org/10.1038/nrg3575

Consortium GP (2010). A map of human genome variation from population-scale sequencing. *Nature*, *467*(7319), 1061–1073.

Consortium IH (2005). A haplotype map of the human genome. *Nature*, *437*(7063), 1299–1320.

Cowen, L., Ideker, T., Raphael, B. J., & Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, *18*(9), 551–562. https://doi.org/10.1038/nrg.2017.38

Csermely, P., Korcsmáros, T., Kiss, H. J., London, G., & Nussinov, R. (2013). Structure and dynamics of molecular networks: a novel paradigm of drug discovery: a comprehensive review. *Pharmacology & Therapeutics*, *138*(3), 333–408. https://doi.org/10.1016/j.pharmthera.2013.01.016

Davidson, E. H., Rast, J. P., Oliveri, P., Ransick, A., Calestani, C., Yuh, C. H., …Otim, O. (2002). A genomic regulatory network for development. *Science*, *295*(5560), 1669–1678. https://doi.org/10.1126/science.1069883

Dehghan, A., Yang, Q., Peters, A., Basu, S., Bis, J. C., Rudnicka, A. R., … Folsom, A. R. (2009). Association of novel genetic Loci with circulating fibrinogen levels: a genome-wide association study in 6 population-based cohorts. *Circulation: Cardiovascular Genetics*, *2*(2), 125–133. https://doi.org/10.1161/CIRCGENETICS.108.825224

Denny, J. C., Bastarache, L., Ritchie, M. D., Carroll, R. J., Zink, R., Mosley, J. D. et al (2013). Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nature Biotechnology*, *31*(12), 1102–1110.

Denny, J. C., Ritchie, M. D., Basford, M. A., Pulley, J. M., Bastarache, L., Brown-Gentry, K. et al (2010). PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene-disease associations. *Bioinformatics*, *26*(9), 1205–1210.

Duret, L. (2002). Evolution of synonymous codon usage in metazoans. *Current Opinion in Genetics & Development*, *12*(6), 640–649.

Fang, H. A.ULTRA-DD Consortium, De Wolf, H., Knezevic, B., Burnham, K. L., & Osgood, J. et al (2019) A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nature Genetics*. *51*(7):1082–1091.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A. et al (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, *41*, D808–D815.

Fuxman Bass, J. I., Sahni, N., Shrestha, S., Garcia-Gonzalez, A., Mori, A., Bhat, N. et al (2015). Human gene-centered transcription factor networks for enhancers and disease variants. *Cell*, *161*(3), 661–673.

Goetz, R. M., & Fuglsang, A. (2005). Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from Escherichia coli. *Biochemical and Biophysical Research Communications*, *327*(1), 4–7.

Han, H., Cho, J. W., Lee, S., Yun, A., Kim, H., Bae, D. et al (2018). TRRUST v2: an expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Research*, *46*(D1), D380–D386.

Hebbring, S. J. (2014). The challenges, advantages and future of phenome-wide association studies. *Immunology*, *141*(2), 157–165.

Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., & Takagi, T. (2001). Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, *18*(6), 523–531.

Hu, H., Miao, Y. R., Jia, L. H., Yu, Q. Y., Zhang, Q., & Guo, A. Y. (2019). AnimalTFDB 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic Acids Research*, *47*(D1), D33–D38. https://doi.org/10.1093/nar/gky822

Ideker, T., & Sharan, R. (2008). Protein networks in disease. *Genome Research*, *18*, 644–652. https://doi.org/10.1101/gr.071852.107

Kanaya, S., Yamada, Y., Kudo, Y., & Ikemura, T. (1999). Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate

analysis. *Gene*, *238*(1), 143–155. https://doi.org/10.1016/S0378-1119(99)00225-5

Köttgen, A., Pattaro, C., Böger, C. A., Fuchsberger, C., Olden, M., Glazer, N. L., … Fox, C. S. (2010). New loci associated with kidney function and chronic kidney disease. *Nature Genetics*, *42*(5), 376–384. https://doi.org/10.1038/ng.568

Kumar, B., Gupta, S. K., Srinivasan, B. P., Nag, T. C., Srivastava, S., & Saxena, R. (2012). Hesperetin ameliorates hyperglycemia induced retinal vasculopathy via anti-angiogenic effects in experimental diabetic rats. *Vascular Pharmacology*, *57*(5–6), 201–207. https://doi.org/10.1016/j.vph.2012.02.007

Kunicki, T. J., Federici, A. B., Salomon, D. R., Koziol, J. A., Head, S. R., Mondala, T. S., … Peake, I. R. (2004). An association of candidate gene haplotypes and bleeding severity in von Willebrand disease (VWD) type 1 pedigrees. *Blood*, *104*(8), 2359–2367. https://doi.org/10.1182/blood-2004-01-0349

Lander, E. S. (2011). Initial impact of the sequencing of the human genome. *Nature*, *470*(7333), 187–197. https://doi.org/10.1038/nature09792

Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., & Beer, M. A. (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nature Genetics*, *47*(8), 955–961. https://doi.org/10.1038/ng.3331

Leiserson, M. D. M., Vandin, F., Wu, H.-T., Dobson, J. R., Eldridge, J. V., Thomas, J. L.,… Raphael, B. J. (2015). Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, *47*(2), 106–114. https://doi.org/10.1038/ng.3168

Liu, C.-C., Tseng, Y.-T., Li, W., Wu, C.-Y., Mayzus, I., Rzhetsky, A., … Zhou, X. J. (2014). DiseaseConnect: a comprehensive web server for mechanism-based disease-disease connections. *Nucleic Acids Research*, *42*, W137–W146. https://doi.org/10.1093/nar/gku412

Liu, X., Yu, X., Zack, D. J., Zhu, H., & Qian, J. (2008). TiGER: a database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, *9*, 271. https://doi.org/10.1186/1471-2105-9-271

Makino, T., & McLysaght, A. (2010). Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(20), 9270–9274. https://doi.org/10.1073/pnas.0914697107

Manolio, T. A. (2013). Bringing genome-wide association findings into clinical use. *Nature Reviews Genetics*, *14*(8), 549–558. https://doi.org/10.1038/nrg3523

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., … Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, *461*(7265), 747–753. https://doi.org/10.1038/nature08494

Mashiach, E., Nussinov, R., & Wolfson, H. J. (2010). FiberDock: Flexible induced-fit backbone refinement in molecular docking. *Proteins.*, *78*(6), 1503–1519.

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., … Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science*, *337*(6099), 1190–1195. https://doi.org/10.1126/science.1222794

McInnes, B. T., Pedersen, T., & Pakhomov, S. V. (2009). UMLS-Interface and UMLS-Similarity: Open source software for measuring paths and semantic similarity. *AMIA Annu Symp Proc.*, *2009*, 431–435.

Mitra, K., Carvunis, A. R., Ramesh, S. K., & Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, *14*(10), 719–732. https://doi.org/10.1038/nrg3552

Nakamura, Y., Gojobori, T., & Ikemura, T. (1999). Codon usage tabulated from the international DNA sequence databases; its status 1999. *Nucleic Acids Research*, *27*(1), 292. https://doi.org/10.1093/nar/27.1.292

Nelson, M. R., Tipney, H., Painter, J. L., Shen, J., Nicoletti, P., Shen, Y., … Sanseau, P. (2015). The support of human genetic evidence for approved drug indications. *Nature Genetics*, *47*(8), 856–860. https://doi.org/10.1038/ng.3314

Noble, W. S., Kuang, R., Leslie, C., & Weston, J. (2005). Identifying remote protein homologs by network propagation. *FEBS Journal*, *272*(20), 5119–5128. https://doi.org/10.1111/j.1742-4658.2005.04947.x

Oti, M., Snel, B., Huynen, M. A., & Brunner, H. G. (2006). Predicting disease genes using protein-protein interactions. *Journal of Medical Genetics*, *43*(8), 691–698. https://doi.org/10.1136/jmg.2006.041376

Pan, X., & Shen, H. B. (2016). OUGENE: a disease associated over-expressed and under-expressed gene database. *Science Bulletin*, *61*(10), 752–754. https://doi.org/10.1007/s11434-016-1059-1

Pendergrass, S. A., Brown-Gentry, K., Dudek, S. M., Torstenson, E. S., Ambite, J. L., Avery, C. L., … Ritchie, M. D. (2011). The use of phenome-wide association studies (PheWAS) for exploration of novel genotype-phenotype relationships and pleiotropy discovery. *Genetic Epidemiology*, *35*(5), 410–422. https://doi.org/10.1002/gepi.20589

Pendergrass, S. A., & Ritchie, M. D. (2015). Phenome-wide association studies: leveraging comprehensive phenotypic and genotypic data for discovery. *Current Genetic Medicine Reports*, *3*(2), 92–100. https://doi.org/10.1007/s40142-015-0067-9

Percudani, R., Pavesi, A., & Ottonello, S. (1997). Transfer RNA gene redundancy and translational selection in Saccharomyces cerevisiae. *Journal of Molecular Biology*, *268*(2), 322–330.

Pham, B. A., Thomas, S. M., Lillie, E., Lee, T., Hamid, J., Richter, T., … Tricco, A. C. (2019). Anti-vascular endothelial growth factor treatment for retinal conditions: a systematic review and meta-analysis. *British Medical Journal Open*, *9*(5), e022031. https://doi.org/10.1136/bmjopen-2018-022031

Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., … Furlong, L. I. (2017). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research*, *45*(D1), D833–D839. https://doi.org/10.1093/nar/gkw943

Plenge, R. M., Scolnick, E. M., & Altshuler, D. (2013). Validating therapeutic targets through human genetics. *Nature Reviews Drug Discovery*, *12*(8), 581–594. https://doi.org/10.1038/nrd4051

Portales-Casamar, E., Kirov, S., Lim, J., Lithwick, S., Swanson, M. I., Ticoll, A., … Wasserman, W. W. (2007). PAZAR: a framework for collection and dissemination of cis-regulatory sequence annotation. *Genome Biology*, *8*(10), R207. https://doi.org/10.1186/gb-2007-8-10-r207

Qian, W., Yang, J. R., Pearson, N. M., Maclean, C., & Zhang, J. (2012). Balanced codon usage optimizes eukaryotic translational efficiency. *PLoS Genetics*, *8*(3), e1002603. https://doi.org/10.1371/journal.pgen.1002603

Quan, Y., Luo, Z.-H., Yang, Q.-Y., Li, J., Zhu, Q., Liu, Y.-M., … Zhang, H.-Y. (2019). Systems chemical genetics-based drug discovery: Prioritizing agents targeting multiple/reliable disease-associated

genes as drug candidates. *Frontiers in Genetics*, *10*, 474. https://doi.org/10.3389/fgene.2019.00474

Quan, Y., Wang, Z. Y., Chu, X. Y., & Zhang, H. Y. (2018). Evolutionary and genetic features of drug targets. *Medicinal Research Reviews*, *38*(5), 1536–1549. https://doi.org/10.1002/med.21487

Rastegar-Mojarad, M., Ye, Z., Kolesar, J. M., Hebbring, S. J., & Lin, S. M. (2015). Opportunities for drug repositioning from phenome-wide association studies. *Nature Biotechnology*, *33*(4), 342–345. https://doi.org/10.1038/nbt.3183

Samanta, M. P., & Liang, S. (2003). Predicting protein functions from redundancies in large-scale protein interaction networks. *Proceedings of the National Academy of Sciences of the United Science of America*, *100*(22), 12579–12583. https://doi.org/10.1073/pnas.2132527100

Sanseau, P., Agarwal, P., Barnes, M. R., Pastinen, T., Richards, J. B., Cardon, L. R., & Mooser, V. (2012). Use of genome-wide association studies for drug repositioning. *Nature Biotechnology*, *30*(4), 317–320. https://doi.org/10.1038/nbt.2151

Schwikowski, B., Uetz, P., & Fields, S. (2000). A network of protein-protein interactions in yeast. *Nature Biotechnology*, *18*(12), 1257–1261. https://doi.org/10.1038/82360

Sekine, M., & Makino, T. (2017). Inference of causative genes for Alzheimer's disease due to dosage imbalance. *Molecular Biology and Evolution*, *34*(9), 2396–2407. https://doi.org/10.1093/molbev/msx183

Sharan, R., Ulitsky, I., & Shamir, R. (2007). Network-based prediction of protein function. *Molecular Systems Biology*, *3*, 88. https://doi.org/10.1038/msb4100129

Shaw, D. R. (2016). Searching the Mouse Genome Informatics (MGI) resources for information on mouse biology from genotype to phenotype. *Current Protocols in Bioinformatics*, *56*(1), 1-7.

Shen, J., Song, K., Slater, A. J., Ferrero, E., & Nelson, M. R. (2017). STOPGAP: a database for systematic target opportunity assessment by genetic association predictions. *Bioinformatics*, *33*(17), 2784–2786. https://doi.org/10.1093/bioinformatics/btx274

Sussman, J. L., Lin, D., Jiang, J., Manning, N. O., Prilusky, J., Ritter, O., & Abola, E. E. (1998). Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. *Acta Crystallographica Section D Biological Crystallography*, *54*(Pt 6 Pt 1), 1078–1084. https://doi.org/10.1107/S0907444998009378

Vazquez, A., Flammini, A., Maritan, A., & Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nature Biotechnology*, *21*(6), 697–700. https://doi.org/10.1038/nbt825

Vilar, S., Cozza, G., & Moro, S. (2008). Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. *Current Topics in Medicinal Chemistry*, *8*(18), 1555–1572.

Visscher, P. M., Brown, M. A., McCarthy, M. I., & Yang, J. (2012). Five years of GWAS discovery. *American Journal of Human Genetics*, *90*(1), 7–24. https://doi.org/10.1016/j.ajhg.2011.11.029

Walhout, A. J. (2006). Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. *Genome Research*, *16*(12), 1445–1454. https://doi.org/10.1101/gr.5321506

Wallace, D. K. (2016). Retinopathy of Prematurity: Anti-VEGF treatment for ROP: which drug and what dose? *Journal of American Association for Pediatric Ophthalmology and Strabismus*, *20*(6), 476–478.

Wang, W. Y., Barratt, B. J., Clayton, D. G., & Todd, J. A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, *6*(2), 109–118. https://doi.org/10.1038/nrg1522

Wang, Z. Y., & Zhang, H. Y. (2013). Rational drug repositioning by medical genetics. *Nature Biotechnology*, *31*(12), 1080–1082.

Xie, T., Yang, Q. Y., Wang, X. T., McLysaght, A., & Zhang, H. Y. (2016). Spatial colocalization of human ohnolog pairs acts to maintain dosage-balance. *Molecular Biology and Evolution*, *33*(9), 2368–2375. https://doi.org/10.1093/molbev/msw108

Zhu, Z., Shendure, J., & Church, G. M. (2005). Discovering functional transcription-factor combinations in the human cell cycle. *Genome Research*, *15*(6), 848–855. https://doi.org/10.1101/gr.3394405

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the Supporting Information section.

How to cite this article: Quan Y, Zhang Q-Y, Lv B-M, Xu R-F, Zhang H-Y. Genome-wide pathogenesis interpretation using a heat diffusion-based systems genetics method and implications for gene function annotation. *Mol Genet Genomic Med*. 2020;8:e1456. https://doi.org/10.1002/mgg3.1456