doi:10.1093/jmcb/mjv008
Published online February 13, 2015

**Article**

# Walking on multiple disease-gene networks to prioritize candidate genes

Rui Jiang[1,2,*]

[1] MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST; Department of Automation, Tsinghua University, Beijing 100084, China
[2] Department of Statistics, Stanford University, Stanford, CA 94305, USA
[*] Correspondence to: Rui Jiang, E-mail: ruijiang@tsinghua.edu.cn

**Uncovering causal genes for human inherited diseases, as the primary step toward understanding the pathogenesis of these diseases, requires a combined analysis of genetic and genomic data. Although bioinformatics methods have been designed to prioritize candidate genes resulting from genetic linkage analysis or association studies, the coverage of both diseases and genes in existing methods is quite limited, thereby preventing the scan of causal genes for a significant proportion of diseases at the whole-genome level. To overcome this limitation, we propose a method named pgWalk to prioritize candidate genes by integrating multiple phenomic and genomic data. We derive three types of phenotype similarities among 7719 diseases and nine types of functional similarities among 20327 genes. Based on a pair of phenotype and gene similarities, we construct a disease-gene network and then simulate the process that a random walker wanders on such a heterogeneous network to quantify the strength of association between a candidate gene and a query disease. A weighted version of the Fisher's method with dependent correction is adopted to integrate 27 scores obtained in this way, and a final *q*-value is calibrated for prioritizing candidate genes. A series of validation experiments are conducted to demonstrate the superior performance of this approach. We further show the effectiveness of this method in exome sequencing studies of autism and epileptic encephalopathies. An online service and the standalone software of pgWalk can be found at http://bioinfo.au.tsinghua.edu.cn/jianglab/pgwalk.**

**Keywords:** disease-gene network, gene prioritization, random walk, data fusion

## Introduction

The identification of causal genes is the primary step toward the prevention, diagnosis, and treatment of human inherited diseases. With the accumulation of functional genomic data, genetics studies are often combined with bioinformatics approaches to facilitate more precise pinpointing of potential causal genes (Moreau and Tranchevent, 2012). For example, in a linkage analysis, disease genes are mapped by measuring recombination against a panel of markers that spread over the entire genome (Ott et al., 2011). The result defines a candidate region of ∼5 M basepairs, containing dozens of candidate gens. In a genome-wide association (GWA) study, a case population is compared with a control one to locate genes related to a query disease in the resolution of ∼10 M basepairs, containing ∼100 candidate genes (McCarthy et al., 2008). The subsequent problem following these genetic studies is then how to rank the candidate genes according to their strength of association with the query disease. Furthermore, resorting to the recent advances in the whole-exome sequencing technique, dozens or hundreds of *de novo* mutations can be screened for a query disease (Bamshad et al., 2011). The question is then how to infer causal genes from genes that contain such mutations.

There have been quite a few bioinformatics approaches developed for the prioritization of candidate genes, and most of these methods can be classified into two categories, those using genomic data only and those combining both phenomic and genomic data. Specifically, methods using genomic data only are usually designed based on the 'guilt-by-association' principle, which suggests that genes associated with a disease should share common functions (Altshuler et al., 2000). Therefore, with the knowledge of a set of seed genes that are known to be associated with a query disease under investigation, candidate genes could be ranked according to their functional similarity to the seed genes. Particularly, the functional similarities have been quantified in existing studies based on a variety of genomic data, including the gene expression (Emilsson et al., 2008), gene ontology (Tiffin et al., 2005), protein sequences (Adie et al., 2005), protein–protein interactions (PPI) (Köhler et al., 2008), and many others (Freudenberg and Propping, 2002; Turner et al., 2003; Lopez-Bigas and Ouzounis, 2004). Moreover, studies have also

Received September 26, 2014. Revised December 11, 2014. Accepted January 11, 2015.
© The Author (2015). Published by Oxford University Press on behalf of *Journal of Molecular Cell Biology*, IBCB, SIBS, CAS. All rights reserved.

Downloaded from https://academic.oup.com/jmcb/article/7/3/214/890187 by guest on 21 August 2022

demonstrated that the integration of multiple genomic data could result in more accurate results of prioritization (Aerts et al., 2006). Nevertheless, genetic bases for a significant proportion of human diseases remain completely unknown, as shown in the OMIM (Online Mendelian Inheritance in Man) database (Hamosh et al., 2005). The scope of applications of such methods based on the guilt-by-association principle is therefore largely restricted due to the difficulty of selecting a suitable set of seed genes.

To overcome this limitation, the second class of methods have been proposed based on the 'guilt-by-indirect-association' principle, which suggests that genes associated with phenotypically similar diseases should share common functions (Chen et al., 2011). Therefore, with knowledge of phenotype similarities between diseases and functional similarities between genes, candidate gens could be ranked by making use of known annotations of diseases genes in the database in a global way. For example, Lage et al. (2007) derived phenotype similarities based on the unified medical language system (UMLS) and adopted a Bayesian model to integrate such similarities and a protein–protein interaction (PPI) network. Wu et al. (2008) adopted phenotype similarities derived from the medical subject headings (MeSH) and quantified the strength of association between a disease and a gene using correlation between phenotype similarities and gene proximities in a PPI network. Wu et al. (2009) further proposed to perform a local alignment of a phenotype network against a PPI network. Li and Patra (2010) adopted a random walk with restart model on an integrated network composed of both diseases and genes. Jiang et al. (2011) further derived a gene semantic similarity network from the gene ontology and showed the advantage of such a network over a PPI one. Vanunu et al. (2010) proposed to simulate how disease status propagated through candidate genes. Chen et al. (2011) proposed to quantify the strength of association between a disease and a gene using the maximum information flow in a phenome-interactome network. However, these methods are often restricted by the availability of the phenotype similarity data and the coverage of the gene similarity data. For example, the most widely used phenotype similarity data as published in van Driel et al. (2006) cover only 5080 diseases, about two-third of diseases recorded in the OMIM database till November 2014. Similarly, it is estimated that the human genome contains >20000 genes, whereas the human protein reference database (Keshava Prasad et al., 2009), as the most widely used PPI data, covers only 9429 (<50%) genes.

Targeting on overcoming these limitations, we propose to prioritize candidate genes by integrating three types of phenomic data (the human phenotype ontology, medical subject headings and unified medical language system) and nine types of genomic data (the gene expression, gene ontology, KEGG pathway, protein sequence, protein domain, protein-protein interaction, signaling network, transcriptional regulatory and microRNA regulation). We first derive three types of phenotype similarities between diseases and nine types of functional similarities between genes from these data. Then, we construct a disease-gene network based on a pair of disease similarity and gene similarity, and then simulate the process that a random walker wanders on such a heterogeneous network to quantitatively measure the strength of association between a candidate gene and a query disease. We further adopt a weighted version of the Fisher's method with dependent correction to integrate 27 scores obtained this way and calibrate a $q$-value for prioritizing candidate genes. We conduct a series of validation experiments to demonstrate the superior performance of our approach and show the effectiveness of this method in exome sequencing studies about neurological diseases.

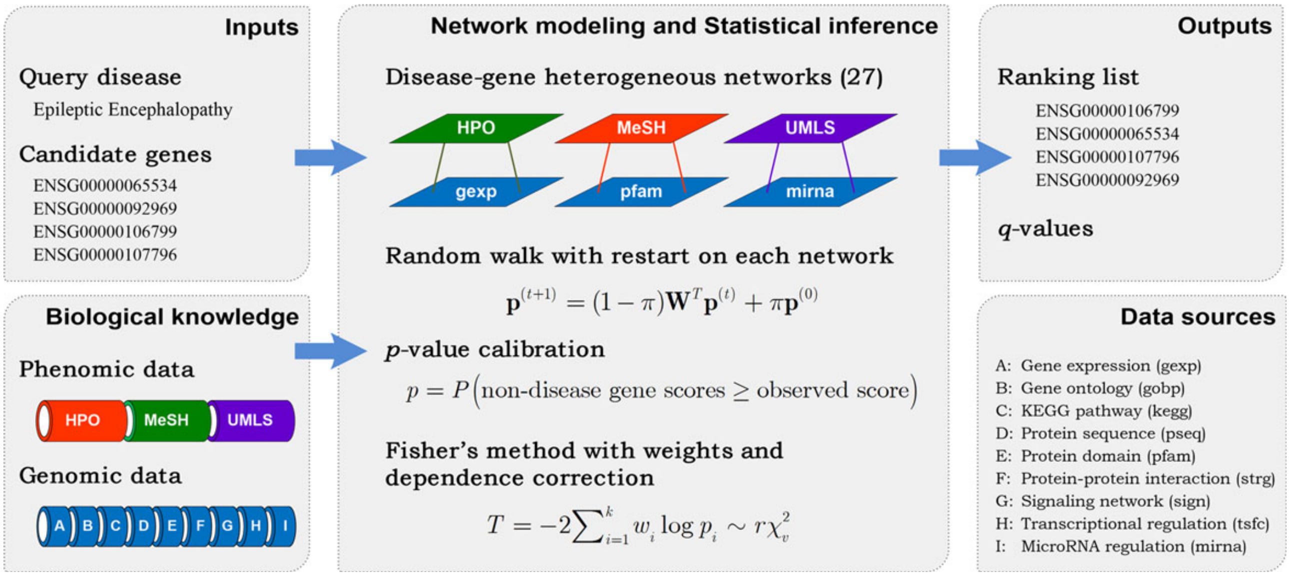
Inputs
Query disease
Epileptic Encephalopathy
Candidate genes
ENSG00000065534
ENSG00000092969
ENSG00000106799
ENSG00000107796
Biological knowledge
Phenomic data
HPO
MeSH
UMLS
Genomic data
A
B
C
D
E
F
G
H
I
Network modeling and Statistical inference
Disease-gene heterogeneous networks (27)
HPO
MeSH
UMLS
gexp
pfam
mirna
Random walk with restart on each network
$\mathbf{p}^{(t+1)} = (1-\pi)\mathbf{W}^T\mathbf{p}^{(t)} + \pi\mathbf{p}^{(0)}$
$p$-value calibration
$p = P\left(\text{non-disease gene scores} \geq \text{observed score}\right)$
Fisher's method with weights and dependence correction
$T = -2\sum_{i=1}^{k} w_i \log p_i \sim r\chi_v^2$
Outputs
Ranking list
ENSG00000106799
ENSG00000065534
ENSG00000107796
ENSG00000092969
$q$-values
Data sources
A: Gene expression (gexp)
B: Gene ontology (gobp)
C: KEGG pathway (kegg)
D: Protein sequence (pseq)
E: Protein domain (pfam)
F: Protein-protein interaction (strg)
G: Signaling network (sign)
H: Transcriptional regulation (tsfc)
I: MicroRNA regulation (mirna)

**Figure 1** Diagram of pgWalk. Given a query disease and a list of candidate genes, pgWalk calculates the statistical significance that a candidate gene is causative for the query disease by integrating three phenomic data and nine genomic data, thereby providing a means of prioritizing the candidate genes.

Downloaded from https://academic.oup.com/jmcb/article/7/3/214/890187 by guest on 21 August 2022

## Results

### *Overview of pgWalk*

Our method, named pgWalk, takes a query disease and a set of candidate genes as inputs and produces a ranking list of the genes according to their strength of association with the query disease. As illustrated in Figure 1, we first derive three phenotype similarity matrices for 7719 human diseases and nine functional similarity matrices for 20327 human genes (Table 1). The phenotype similarity matrices are derived from the human phenotype ontology (HPO) (Robinson et al., 2008), medical subject headings (MeSH) (Lipscomb, 2000) and unified medical language system (UMLS) (Lindberg et al., 1993). Functional similarity matrices are derived from the gene expression (gexp) (Su et al., 2004), gene ontology (gobp) (Ashburner et al., 2000), KEGG pathway (kegg) (Kanehisa and Goto, 2000), protein sequence (pseq) (Apweiler et al., 2004), protein domain (pfam) (Bateman et al., 2004), protein-protein interaction (strg) (Snel et al., 2000), signaling network (sign) (Cui et al., 2007), transcriptional regulation (tsfc) (Matys et al., 2003), and microRNA regulation (mirna) (Betel et al., 2008). Then, we construct disease networks from phenotype similarity matrices and gene networks from gene similarity matrices by adopting a nearest neighbor strategy, and we generate a number of 27 disease-gene networks by connecting a disease network and a gene network using known associations between diseases and genes. Next, we simulate the process that a random walker wanders in each of these disease-gene heterogeneous networks, obtain steady-state probabilities for candidate genes, and calibrate the probabilities to obtain *p*-values that measure the strength of association between the query disease and the candidate genes. After that, we integrate these *p*-values to obtain a single statistical significance using Fisher's method with dependence correction and taking weights of the data sources into consideration. Finally, we apply a multiple testing correction procedure to calculate *q*-values from the combined *p*-values, for the purpose of controlling the positive false discovery rate of the results, and we sort the candidate genes according to their *q*-values to produce the ranking list as output.

**Table 1 Coverage of individual data sources.**

| Data sources | Coverage | |
|---|---|---|
| | **Disease/gene** | **Ratio (%)** |
| Phenomic data | 7719 | 100.00 |
| Human phenotype ontology (HPO) | 6376 | 82.60 |
| Medical subject headings (MeSH) | 7719 | 100.00 |
| Unified medical language system (UMLS) | 7719 | 100.00 |
| Genomic data | 20327 | 100.00 |
| Gene expression (gexp) | 12462 | 61.31 |
| Gene ontology (gobp) | 15602 | 76.76 |
| KEGG pathway (kegg) | 6468 | 31.82 |
| Protein sequence (pseq) | 14196 | 69.84 |
| Protein domain (pfam) | 17091 | 84.08 |
| Protein–protein interaction (strg) | 12432 | 61.16 |
| Signaling network (sign) | 5995 | 29.49 |
| Transcriptional regulatory pattern (tsfc) | 20314 | 99.94 |
| MicroRNA regulatory pattern (mirna) | 17552 | 86.35 |

The three phenotype similarity and nine gene similarity matrices cover 7719 human genetic disorders and 20327 human genes, respectively.

### *Phenotype overlap implies genotype overlap*

We first validated the basic assumption of our approach by exploring whether genes associated with phenotypically similar diseases exhibited functional similarities across different genomic data sources. Given a type of phenomic data and a type of genomic data, we quantified the phenotype similarity between a pair of diseases as the cosine value calculated by the text mining technique based on the phenomic data, and we measured the genotype similarity of the two diseases as the average pairwise similarity scores of their associated genes under the genomic data source. With these definitions, we calculated the phenotype similarity between each pair of the 3933 diseases with associated genes, partitioned the resulting scores into 10 bins of equal size, identified disease pairs belonging to each bin, and calculated the average genotype similarity of disease pairs in each bin.

As an illustration, relationships between the phenotype similarity derived from MeSH and the nine genotype similarities are shown in Figure 2. Taking KEGG pathway as an example, for disease pairs with weak phenotype similarity (0.0–0.1), their genotype similarity is also weak (0.001142 on average). For disease pairs with strong phenotype similarity (0.9–1.0), their genotype similarity is also strong (0.5682 on average). In the middle of the spectrum, for disease pairs with medium phenotype similarity (0.5–0.6), their genotype similarity is also at the medium level (0.0910 on average). For the eight other genomic data sources, we observe similar patterns. These results suggest that genes associated with phenotypically similar diseases indeed exhibit functional similarities across different gnomic data sources. In other words, the guilt-by-indirect-association principle is valid. We further regressed the mean genotype similarity of each bin against the corresponding mean phenotype similarity. Results show that the coefficients of determination ($r^2$) are 0.8910 for the gene expression, 0.8344 for gene ontology, 0.8609 for KEGG pathway, 0.8839 for protein sequence, 0.7995 for protein domain, 0.8614 for protein–protein interaction, 0.8845 for signaling network, 0.8115 for transcriptional regulation, and 0.8724 for microRNA regulation. These results further suggest that the phenotype overlap implies the genotype overlap. We also repeated the above analysis for phenotype similarities derived from HPO and UMLS and found similar results.

### *Performance in validation experiments*

We performed three large-scale leave-one-out cross-validation experiments to validate the effectiveness of pgWalk using the 4606 annotated associations between 3933 diseases and 3028 genes. We fist simulated the situation of a traditional linkage analysis or association study, in which the objective was to prioritize candidate genes in a linkage interval. In each validation run, we focused on one disease-gene pair in an annotated association, took the disease as the query disease and the gene as the test gene, collected a set of control genes that are located within a 10 Mb region centered at the test gene, and ranked the test gene against the control genes using our method. In this procedure, we removed all annotated associations regarding the query disease to simulate the circumstance that the genetic basis of the query disease was completely unknown. We summarized

Downloaded from https://academic.oup.com/jmcb/article/7/3/214/890187 by guest on 21 August 2022

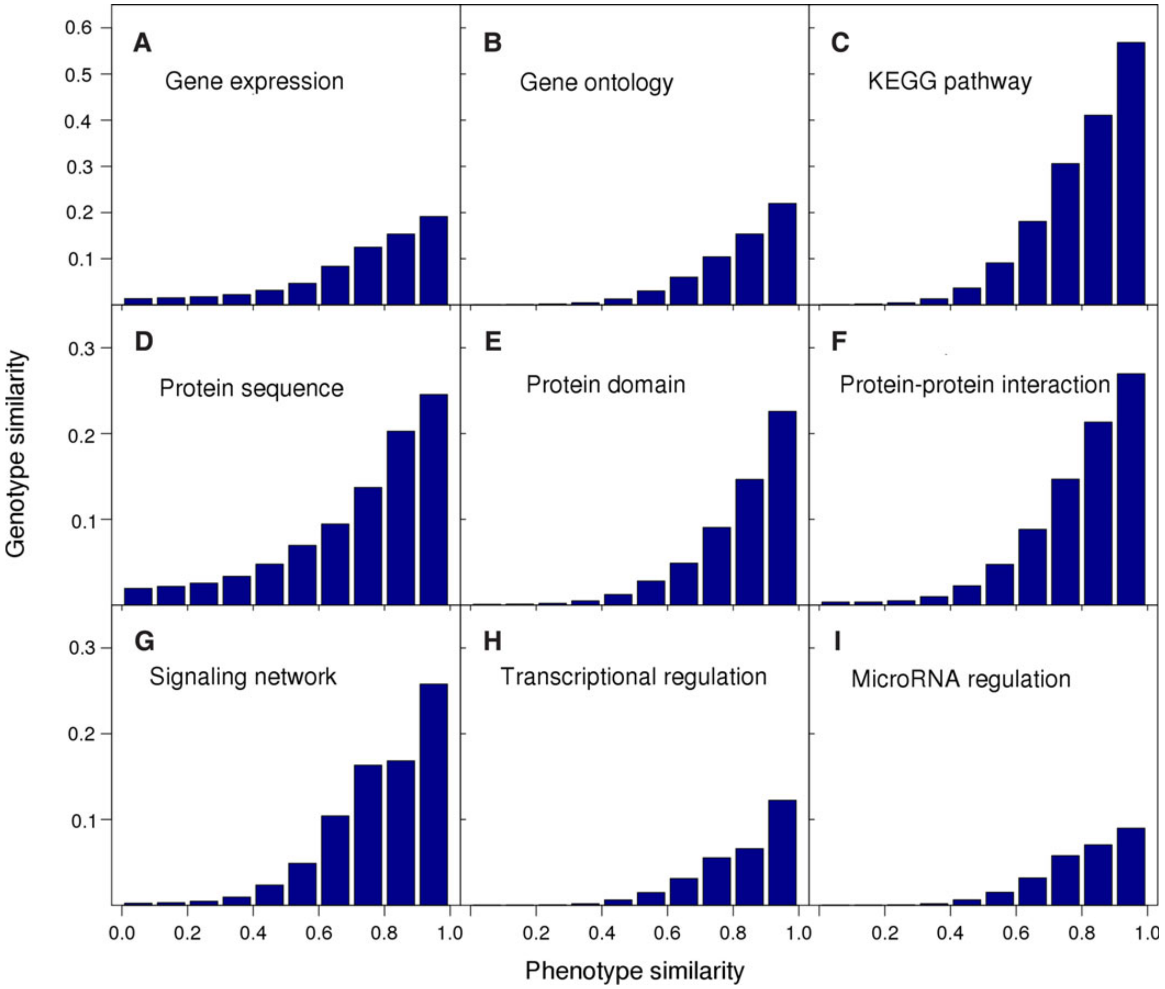

**Figure 2** Phenotype overlap implies genotype overlap. For a pair of two diseases, we define their phenotype similarity as the cosine value calculated by the text mining technique and their genotype similarity under a certain genomic data source as the average pairwise similarity of their associated genes derived from the genomic data. Results are obtained using phenotype similarities derived from MeSH.

ranks of the test genes in Figure 3A. In a total of 4606 validation runs, the average length of a candidate gene list is 104. Our method ranks 2537 test genes first and 3867 among top 10 in their corresponding candidate lists. In contrast, with a random guess procedure, one can only expect 4606/104≈44.29 test genes ranked first and 442.88 enriched among top 10. These results suggest the capability of our method in identifying disease genes from a linkage interval.

We then derived three criteria to quantify the performance of our method. Dividing the number of test genes ranked among top 10 by the total number of candidate genes, we obtained a criterion called the ratio of top ranked test genes (TOP). Dividing the rank of a test gene by the total number of test and control genes in a validation run, we obtained the rank ratio of the test gene. Averaging rank ratios of all test genes, we obtained a criterion called the mean rank ratio (MRR). At a certain threshold of the rank ratio, we defined the sensitivity and the specificity as the fraction of test and control genes ranked above and below the threshold, respectively. Varying the threshold, we plotted the rank operating characteristic (ROC) curve (sensitivity versus 1-specificity) and further calculated the area under this curve as a criterion called the AUC score. As shown in Table 2, TOP, MRR, and AUC for validation experiment against a linkage interval are 83.96%, 9.44%, and 90.99%, respectively, further suggesting the effectiveness of our method.

The number of control genes in a linkage interval may have variation, thereby introducing biases in assessing the capability of our method in enriching test genes at top positions (e.g. ranking a test gene among top 10 against 20 control genes is much easier than ranking it among top 10 against 100 control genes). We therefore performed another validation experiment (i.e. nearest neighbors) by ranking each test gene against 99 control genes that were closest to the test gene in the same chromosome. From Figure 3B, we observe the capability of our method in ranking 2472 test genes first and 3791 among top 10 in their corresponding candidate lists. From Table 2, we observe that the TOP, MRR, and AUC are 82.31%, 8.94%, and 91.95%, respectively, further suggesting the effectiveness of our method.

Downloaded from https://academic.oup.com/jmcb/article/7/3/214/890187 by guest on 21 August 2022

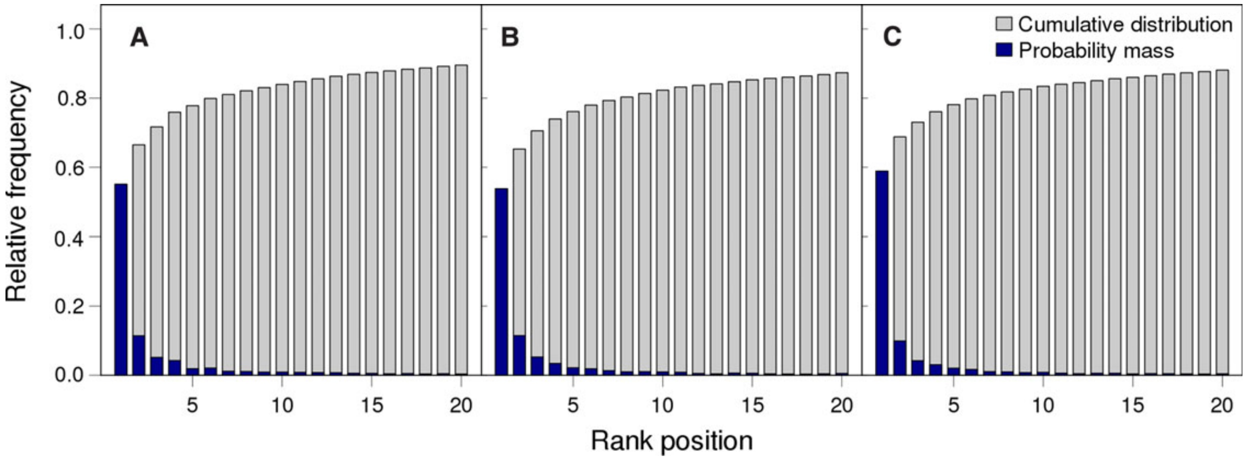

**Figure 3** Enrichment of top ranking test genes. Probability mass and cumulative distribution of top 20 genes for validation against a linkage interval (**A**), nearest neighbors (**B**), and random controls (**C**).

We simulated the situation of exome sequencing studies, in which genetic variants are sequenced across the whole exome. In each validation run, we collected a set of 99 control genes that were selected at random from the entire genome, and ranked a test gene against the random controls. We summarized ranks of the test genes in this validation in Figure 3C. In a total of 4606 validation runs, our method ranks 2712 test genes first and 3842 among top 10 in their corresponding candidate lists, strongly supporting the capability of our method in identifying disease genes from random controls. The low MRR (8.11%) and high TOP (83.41%) and AUC (92.80%) shown in Table 2 further confirm the effectiveness of our method.

We further compared the performance of pgWalk with three state-of-the-art methods (Aerts et al., 2006; Li and Patra, 2010; Vanunu et al., 2010) and show the superiority of our approach (Supplementary text and Table S1).

*Performance for diseases of different inheritance styles*

We assessed performance of pgWalk for diseases with different inheritance styles. We first classified the 3933 diseases into 58 complex diseases and 3875 Mendelian diseases according to the Genetic Association database (Becker et al., 2004) and found that our method could recover disease genes for both groups (Table 2 and Supplementary Figure S1). In the validation against a linkage interval, the MRR and AUC are 5.13% and 95.48% for complex diseases, respectively and 9.50% and 90.93% for Mendelian diseases, respectively. Nevertheless, a two-sided Wilcoxon rank sum test suggests that the rank ratios of test genes in these two categories of disease are not significantly different (raw $p$-value $= 0.746$). We then identified 695 autosomal dominant, 729 autosomal recessive, 263 X-linked and 3 Y-linked disorders according to the OMIM database. Results show that our method can recover disease genes for all these categories of disorders. In the validation against a linkage interval, the MRRs are 7.26%, 8.45%, 13.37%, and 14.48% for autosomal dominant, autosomal recessive, X-linked, and Y-linked disorders, respectively, while the AUCs are 93.15%, 91.99%, 87.22%, and 87.49%, respectively. Two-sided Wilcoxon rank sum tests suggest that the rank ratios of test genes are different ($p$-value $= 2.16 \times 10^{-3}$ after Bonferroni correction for multiple comparisons) for the two autosomal inheritance styles, not different ($p$-value $= 0.339$ after Bonferroni correction) for the two sex-linked inheritance styles, and different ($p$-values $< 0.05$ after Bonferroni correction) for the autosomal and sex-linked disorders. We further identified 46 immune diseases and 261 neurological disorders and found that our method was also capable of recovering disease genes for these two classes of diseases. In the validation against a linkage interval, the MRR and AUC are 6.09% and 94.32% for immune diseases, respectively, and 6.28% and 94.12% for neurological disorders, respectively. A two-sided Wilcoxon rank sum test suggests that the rank ratios of test genes for these two categories of diseases are not different (raw $p$-value $= 6.45 \times 10^{-2}$). We also repeated the above analysis using validation experiments against nearest neighbors and random controls and observed similar results.

We then classified the 3933 diseases into seven categories according to the number of genes annotated as associated with the diseases in the OMIM database (Hamosh et al., 2005). Results show that our method can recover disease genes for all these groups in that the MRRs are all around 10%, and the AUCs are all around 90% (Table 2 and Supplementary Figure S1). A pairwise two-sided Wilcoxon rank sum test suggests that the rank ratios of test genes for most group are not different ($p$-values $>0.01$ after Bonferroni correction), except that the performance for diseases associated with two genes is different from that for diseases associated with 11 or more genes ($p$-values $= 6.80 \times 10^{-4}$ after Bonferroni correction). We resorted to a linear regression model to analyze the relationship between the rank ratio of a disease gene and the number of genes associated with the disease that the gene was relevant to. Results show that the data can hardly fit the model ($r^2 = 5.40 \times 10^{-4}$), suggesting that the prioritization procedure does not depend on the amount of prior information regarding a disease.

Downloaded from https://academic.oup.com/jmcb/article/7/3/214/890187 by guest on 21 August 2022

**Table 2** Performance of pgWalk in the leave-one-out cross-validation experiments.

| Category | Association | | | Linkage interval | | | Nearest neighbors | | | Random controls | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Case | Disease | Gene | TOP | MRR | AUC | TOP | MRR | AUC | TOP | MRR | AUC |
| Inheritance style | | | | | | | | | | | | |
| Complex | 60 | 58 | 60 | 91.67 | 5.13 | 95.48 | 91.67 | 4.42 | 96.52 | 91.67 | 3.70 | 97.27 |
| Mendelian | 4546 | 3875 | 2991 | 83.85 | 9.50 | 90.93 | 82.18 | 9.00 | 91.89 | 83.30 | 8.17 | 92.74 |
| Autosomal dominant | 951 | 695 | 689 | 86.65 | 7.26 | 93.15 | 85.49 | 6.84 | 94.08 | 86.65 | 6.50 | 94.43 |
| Autosomal recessive | 916 | 729 | 818 | 87.55 | 8.45 | 91.99 | 85.37 | 7.98 | 92.93 | 85.37 | 7.16 | 93.75 |
| X-linked | 270 | 263 | 181 | 74.81 | 13.37 | 87.22 | 71.11 | 12.38 | 88.47 | 77.04 | 10.10 | 90.78 |
| Y-linked | 7 | 3 | 6 | 100.00 | 14.48 | 87.49 | 85.71 | 10.32 | 90.54 | 85.71 | 5.29 | 95.67 |
| Immune | 78 | 46 | 70 | 85.90 | 6.09 | 94.32 | 85.90 | 5.85 | 95.09 | 85.90 | 5.44 | 95.51 |
| Neurological | 293 | 261 | 237 | 89.08 | 6.28 | 94.12 | 89.08 | 5.83 | 95.09 | 89.42 | 5.66 | 95.28 |
| Associated genes | | | | | | | | | | | | |
| 1 | 3546 | 3546 | 2579 | 83.90 | 9.41 | 91.03 | 81.75 | 8.87 | 92.02 | 82.63 | 8.39 | 92.51 |
| 2 | 375 | 217 | 303 | 85.07 | 8.93 | 91.46 | 85.33 | 8.52 | 92.38 | 86.93 | 7.28 | 93.64 |
| 3 | 147 | 53 | 137 | 88.44 | 5.14 | 95.29 | 87.76 | 4.71 | 96.23 | 88.44 | 4.01 | 96.96 |
| 4 | 87 | 25 | 82 | 87.36 | 9.38 | 91.05 | 85.06 | 9.36 | 91.53 | 85.06 | 6.76 | 94.16 |
| 5 | 82 | 20 | 75 | 85.37 | 10.41 | 90.09 | 81.71 | 9.43 | 91.44 | 84.15 | 8.20 | 92.71 |
| 6–10 | 207 | 56 | 178 | 84.06 | 11.30 | 89.05 | 83.57 | 10.89 | 89.97 | 87.44 | 7.01 | 93.91 |
| ≥11 | 162 | 16 | 149 | 75.93 | 12.45 | 87.89 | 79.63 | 12.29 | 88.54 | 81.48 | 9.67 | 91.23 |
| Associated SNPs | | | | | | | | | | | | |
| 1 | 876 | 876 | 789 | 80.48 | 11.22 | 89.20 | 77.97 | 10.74 | 90.13 | 78.42 | 10.05 | 90.84 |
| 2 | 473 | 439 | 439 | 86.47 | 8.62 | 91.85 | 83.72 | 8.05 | 92.85 | 84.99 | 7.74 | 93.17 |
| 3 | 269 | 245 | 254 | 88.48 | 6.96 | 93.47 | 87.73 | 6.46 | 94.45 | 88.85 | 5.96 | 94.96 |
| 4 | 228 | 187 | 222 | 84.21 | 8.04 | 92.37 | 82.46 | 7.53 | 93.35 | 83.77 | 7.36 | 93.55 |
| 5 | 176 | 152 | 170 | 88.64 | 8.19 | 92.29 | 88.07 | 7.32 | 93.59 | 88.64 | 5.90 | 95.04 |
| 6–10 | 398 | 342 | 364 | 87.19 | 7.75 | 92.68 | 86.18 | 7.32 | 93.58 | 87.19 | 6.91 | 94.01 |
| 11–20 | 353 | 236 | 329 | 87.54 | 7.31 | 93.11 | 87.82 | 6.84 | 94.08 | 89.52 | 5.67 | 95.26 |
| 21–50 | 278 | 176 | 244 | 94.24 | 6.82 | 93.66 | 93.88 | 6.32 | 94.61 | 94.60 | 3.56 | 97.40 |
| ≥51 | 152 | 102 | 149 | 92.76 | 5.41 | 95.01 | 92.76 | 4.93 | 96.01 | 93.42 | 4.61 | 96.35 |
| **All diseases** | **4606** | **3933** | **3028** | **83.96** | **9.44** | **90.99** | **82.31** | **8.94** | **91.95** | **83.41** | **8.11** | **92.80** |

Diseases are classified into different categories according to their inheritance styles, numbers of annotated genes associated with a disorder in the OMIM database, and numbers of annotated SNPs associated with a disorder in the Swiss-Prot database. Numbers for evaluation criteria (TOP, MRR, and AUC) are percentages.

We further identified 25074 SNPs associated with 2709 diseases in our study according to the Swiss-Prot database (Apweiler et al., 2004) and classified these diseases into nine categories according to the number of SNPs annotated as associated with the diseases. Results demonstrate that our method can also recover disease genes for all these groups (most MRRs less than 10% and most AUCs greater than 90%, as shown in Table 2 and Supplementary Figure S1). A pairwise two-sided Wilcoxon rank sum test suggests that the rank ratios of test genes for most group are not different ($p$-values $>0.01$ after Bonferroni correction), except that the performance for diseases associated with only one SNP is different from those for diseases associated with 11 or more SNPs ($p$-values $<0.01$ after Bonferroni correction). We further performed a linear regression analysis regarding the rank ratio of a disease gene and the number of SNPs associated with the disease that the gene was relevant to, and we found that the data could hardly fit the model ($r^2 = 7.16 \times 10^{-3}$), again suggesting that the prioritization procedure does not depend on the amount of prior information about a disease.

### *Data fusion improves prioritization performance*

We compared the performance of pgWalk with that of individual disease-gene networks (features) and summarized the results in Table 3 and Supplementary Figure S2. Taking the validation experiment against a linkage interval as an example, among the 27 features derived from the three phenomic and nine genetic data sources, the combination of UMLS and the gene ontology (gobp) yields the highest performance (TOP = 79.53%) in terms of TOP, while that of HPO and the signaling network (sign) yields the lowest performance (TOP = 43.86%). In terms of MRR and AUC, the combination of UMLS and the protein–protein interaction network (strg) yields the highest performance (MRR = 12.65% and AUC = 87.98%), while that of HPO and the microRNA regulation (mirna) yields the lowest performance (MRR = 30.59% and AUC = 69.66%). The improvements of pgWalk over individual data sources in terms of MRR are then between 25.38% and 69.14%. We further plotted ROC curves for pgWalk and individual features in Figure 4, from which we clearly observe that the curve of pgWalk climbs much faster toward the upper left corner of the plot than do individual features, suggesting the superior capability of this data fusion approach in achieving high sensitivity while maintaining high specificity. Similar results are obtained for validation experiments against nearest neighbors and random controls. These results clearly demonstrate the improvement of pgWalk over individual features in the prioritization accuracy and suggest the power of data fusion.

More importantly, the coverage of pgWalk also benefits from data fusion. For example, HPO covers only 6376 diseases, and the signaling network covers only 5995 genes (Table 1). Transcriptional regulation (tsfc), though covers 20314 genes, can only achieve low prioritization performance (MRR around 29% and AUC about 71%). With data fusion, however, pgWalk covers

Downloaded from https://academic.oup.com/jmcb/article/7/3/214/890187 by guest on 21 August 2022

**Table 3** Performance of individual disease-gene networks and pgWalk in the leave-one-out cross-validation experiments.

| Phenomic data | Genomic data | Linkage interval (%) | | | Nearest neighbors (%) | | | Random controls (%) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | TOP | MRR | AUC | TOP | MRR | AUC | TOP | MRR | AUC |
| HPO | gexp | 50.87 | 24.87 | 75.55 | 47.03 | 24.33 | 76.13 | 47.57 | 23.81 | 76.68 |
| | gobp | 69.24 | 15.36 | 85.11 | 66.80 | 14.69 | 85.72 | 67.22 | 13.71 | 86.70 |
| | keg | 46.92 | 20.68 | 80.26 | 45.01 | 19.21 | 81.78 | 45.31 | 17.89 | 83.16 |
| | pseq | 50.89 | 21.48 | 78.93 | 48.15 | 20.89 | 79.52 | 48.89 | 20.30 | 80.12 |
| | pfam | 54.97 | 25.45 | 74.86 | 52.08 | 24.97 | 75.23 | 52.95 | 24.43 | 75.77 |
| | strg | 67.80 | 14.48 | 86.13 | 66.13 | 13.66 | 86.98 | 66.46 | 12.88 | 87.80 |
| | sign | 43.86 | 20.78 | 80.17 | 42.21 | 19.57 | 81.46 | 42.27 | 19.17 | 81.93 |
| | tsfc | 47.37 | 29.50 | 70.73 | 44.55 | 29.29 | 71.38 | 44.88 | 29.17 | 71.47 |
| | mirna | 46.96 | 30.59 | 69.66 | 44.29 | 29.85 | 70.26 | 44.96 | 29.63 | 70.50 |
| MeSH | gexp | 57.97 | 22.56 | 77.89 | 54.32 | 21.92 | 78.58 | 54.91 | 21.59 | 78.95 |
| | gobp | 79.24 | 13.19 | 87.30 | 76.86 | 12.54 | 87.88 | 77.68 | 11.55 | 88.86 |
| | kegg | 52.41 | 18.25 | 82.74 | 51.09 | 16.85 | 84.23 | 51.50 | 15.55 | 85.56 |
| | pseq | 58.68 | 18.88 | 81.56 | 56.04 | 18.38 | 82.06 | 56.64 | 17.87 | 82.58 |
| | pfam | 62.87 | 23.61 | 76.72 | 60.29 | 23.13 | 77.07 | 61.16 | 22.47 | 77.71 |
| | strg | 75.90 | 12.82 | 87.81 | 74.10 | 12.00 | 88.68 | 74.75 | 11.05 | 89.68 |
| | sign | 49.33 | 18.45 | 82.55 | 47.83 | 17.10 | 84.01 | 47.85 | 16.80 | 84.37 |
| | Tsfc | 54.06 | 28.25 | 72.00 | 51.50 | 28.01 | 72.68 | 51.78 | 27.85 | 72.81 |
| | mirna | 53.47 | 29.07 | 71.20 | 50.80 | 28.40 | 71.71 | 50.87 | 28.26 | 71.86 |
| UMLS | gexp | 58.21 | 22.58 | 77.88 | 54.08 | 21.92 | 78.58 | 54.75 | 21.64 | 78.89 |
| | gobp | 79.53 | 12.95 | 87.55 | 77.12 | 12.27 | 88.15 | 77.66 | 11.39 | 89.03 |
| | kegg | 52.65 | 17.88 | 83.12 | 51.37 | 16.44 | 84.66 | 51.50 | 15.33 | 85.80 |
| | pseq | 58.99 | 18.91 | 81.53 | 57.10 | 18.37 | 82.07 | 57.10 | 17.82 | 82.63 |
| | pfam | 63.31 | 23.48 | 76.86 | 60.64 | 23.02 | 77.17 | 61.03 | 22.44 | 77.74 |
| | strg | 75.71 | 12.65 | 87.98 | 73.95 | 11.85 | 88.83 | 74.58 | 11.01 | 89.71 |
| | sign | 49.37 | 18.19 | 82.81 | 47.70 | 16.80 | 84.31 | 48.11 | 16.61 | 84.57 |
| | tsfc | 54.32 | 28.15 | 72.09 | 51.87 | 27.93 | 72.76 | 51.93 | 27.78 | 72.88 |
| | mirna | 53.28 | 29.07 | 71.20 | 50.39 | 28.41 | 71.70 | 50.87 | 28.28 | 71.84 |
| **pgWalk** | | **83.96** | **9.44** | **90.99** | **82.31** | **8.94** | **91.95** | **83.41** | **8.11** | **92.80** |
| Greedy selection | | 83.50 | 9.60 | 90.82 | 82.67 | 9.08 | 91.13 | 83.87 | 8.23 | 91.98 |
| Equal weight | | 79.22 | 11.41 | 89.00 | 76.60 | 10.94 | 89.92 | 77.42 | 10.15 | 90.73 |

All numbers are percentages. pgWalk clearly induces an improvement in the prioritization accuracy.

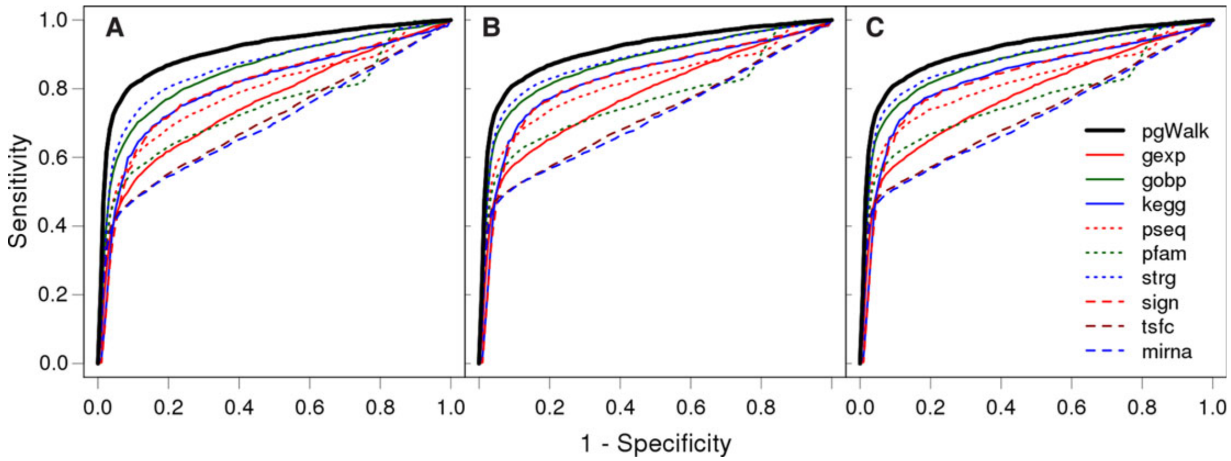

**Figure 4** ROC curves for pgWalk and individual disease-gene networks in the validation experiment against a linkage interval. Curves are plotted for nine genomic data sources under the phenomic data from HPO (**A**), MeSH (**B**), and UMLS (**C**). The curve of pgWalk climbs much faster toward the upper left corner, suggesting its superiority in achieving high sensitivity while maintaining high specificity.

7719 diseases and 20327 genes, much more than most individual data sources, and thus makes it feasible to perform a whole-genome scan for disease genes for a query disease. This advantage is very important for recent advances such as exome sequencing studies in researches regarding human diseases.

*Feature weighting improves data fusion effectiveness*

We compared the performance of our feature weighting method with two other strategies, equal weight and greedy selection. In the former, we assigned equal weights to all of the 27 features, and hence this fusion scheme was equivalent to a dependent and

Downloaded from https://academic.oup.com/jmcb/article/7/3/214/890187 by guest on 21 August 2022

unweighted version of the Fisher's method. In the later, we applied a sequential backward selection algorithm that started from a set containing all the features and repeatedly removed one feature at a time until the performance began to drop.

As shown in Table 3, we observe that both the feature weighting and greedy selection strategies achieve obviously high performance than the equal weight strategy. For example, in the validation against a linkage interval, the equal weight strategy achieves a TOP of 79.22%, an MRR of 11.41%, and an AUC of 89.00%, apparently lower than those of either the feature weighting strategy (TOP = 83.96%, MRR = 9.44%, and AUC = 90.99%) and the greedy selection strategy (TOP = 83.50%, MRR = 9.60%, and AUC = 90.82%). This observation confirms the necessity of incorporating a feature selection mechanism into the data fusion procedure.

We also observe that the greedy selection strategy, though effective by itself, performs slightly worse than the feature weighting strategy. This phenomenon can be explained as follows. In the feature weighting strategy, each feature is weighted by a parameter in the interval [0, 1]. In the greedy selection strategy, however, each feature is either selected or excluded. Therefore, the weighting strategy is more general than the greedy selection one and provides more flexible control over the selection of important features. We further analyzed the relationship between features selected by the greedy strategy and their weights assigned by the weighting strategy. Results show that features selected by the greedy strategy are typically assigned large weights, and vice versa (Supplementary Figure S3).

*Contributions of individual data sources*

We applied a hierarchical cluster analysis to the correlation coefficient matrix of the 27 disease-gene networks (features). Results, as shown in Figure 5, demonstrate the existence of positive correlations between features (mean correlation coefficient = 0.2631). Moreover, for a fixed type of genomic data source (e.g. pseq), correlations between the features (e.g. hpo-pseq, mesh-pseq, and umls-pseq) are typically strong (blocks along the main diagonal). For a fixed type of phenomic data source, however, we do not observe such phenomena. It is interesting to see that features related to the protein sequence (pseq) are highly correlated with those related to the protein domain (pfam). This observation is consistent with the fact that a protein family usually consists of proteins with local sequence similarities. We also observe that features related to the protein–protein interaction (strg) show medium correlations with features related to the signaling network (sign). This observation is consistent with the fact that ~1/3 edges in the signaling network are physical interactions.

Considering the existence of correlations between features, the prediction power of an individual disease-gene network may not reflect its real contribution to the final performance of our method. We therefore evaluated relative contribution of a feature

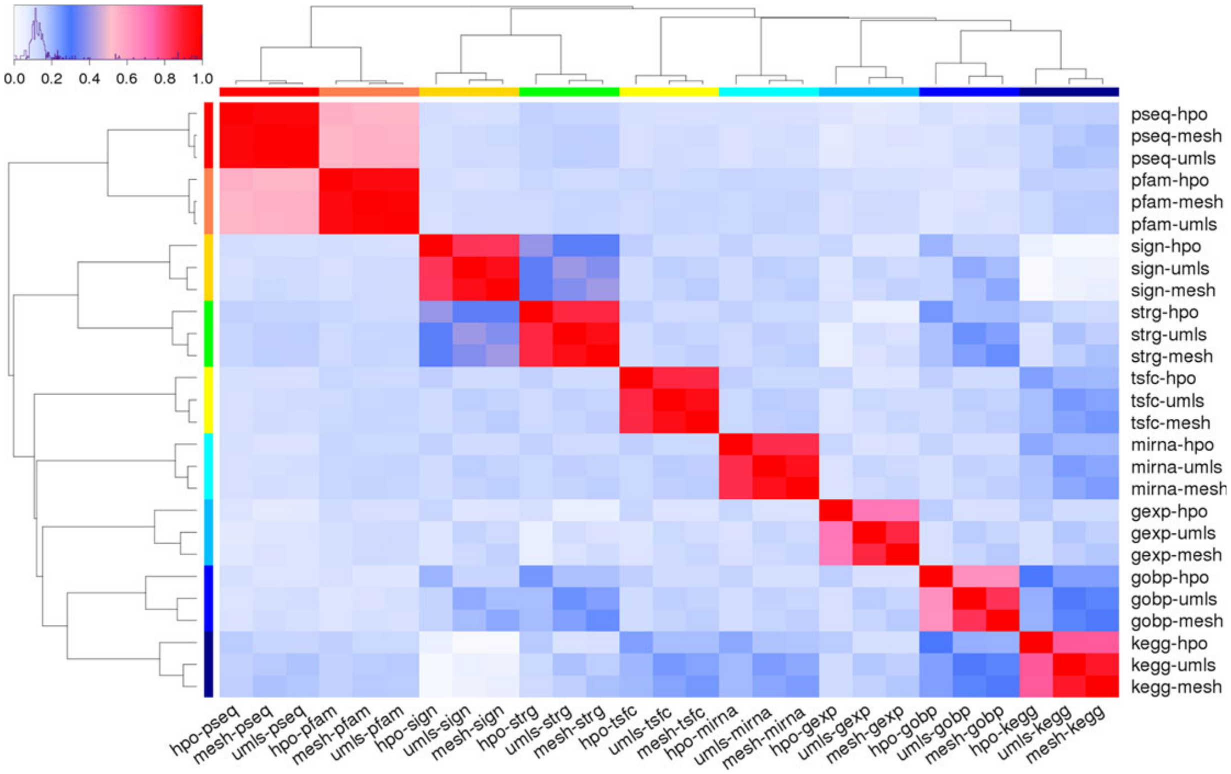

**Figure 5** Cluster analysis of correlation coefficients among individual disease-gene networks. Positive correlations widely exist between disease-gene networks (features). For a fixed type of genomic data source, correlations between the features are typically strong (blocks along the main diagonal). Features related to protein sequences (pseq) are highly correlated with features related to protein domains (pfam). Features related to protein–protein interaction network (strg) show medium correlations with features related to signaling network (sign).

Downloaded from https://academic.oup.com/jmcb/article/7/3/214/890187 by guest on 21 August 2022

by removing it from the Fisher's method and repeating the validation experiments. As shown in Figure 6, the gene ontology (gobp) and protein-protein interaction (strg), no matter combined with which phenomic data source, have positive contributions, because the removal of a feature with these two genomic data sources involved results in increase in MRR and decrease in AUC in all the three validation experiments. The protein sequence (pseq), when combined with MeSH or UMLS, has positive contributions across the three validation experiments. The KEGG pathway (kegg), when combined with UMLS, has positive contributions across the three validation experiments. The gene expression (gexp), when combined with HPO, has positive contributions in all the three validation experiments. The signaling pathway (sign), when combined with UMLS, has positive contributions in the validation against a linkage interval and nearest neighbors. The two regulation data (tsfc and mirna), no matter combined with which phenomic data, have weak contributions (either positive or negative) across all the three validation experiments. It is also interesting to see that the protein domain (pfam) has negative contributions in most cases. However, using this data source alone yields higher performance than the two regulatory information (Table 3). We conjecture this inconsistency is due to the fact that features derived from the protein domain are strongly correlated with those derived from the sequence (Figure 5), and thus information in the former has been included in the later.

We further evaluated relative contributions of a data source by repeating the validation experiments with all features derived from the data source removed. Results show that all the 12 data sources have positive contributions in two or more validation experiments (Supplementary Figure S4). Particularly, the three phenomic data can be ordered according to their contributions (from the most to the least) as UMLS, Mesh and HPO. Among the nine genomic data sources, the gene ontology has the largest contribution, followed by the protein–protein interaction, and contributions of the seven other data sources only show subtle differences.

### *Pairwise interactions between data sources*

We evaluated pairwise interactive effects of the 27 disease-gene networks (features). Focusing on the validation against a linkage interval, we first obtained contributions of individual features by repeating the experiment with a feature removed and calculating

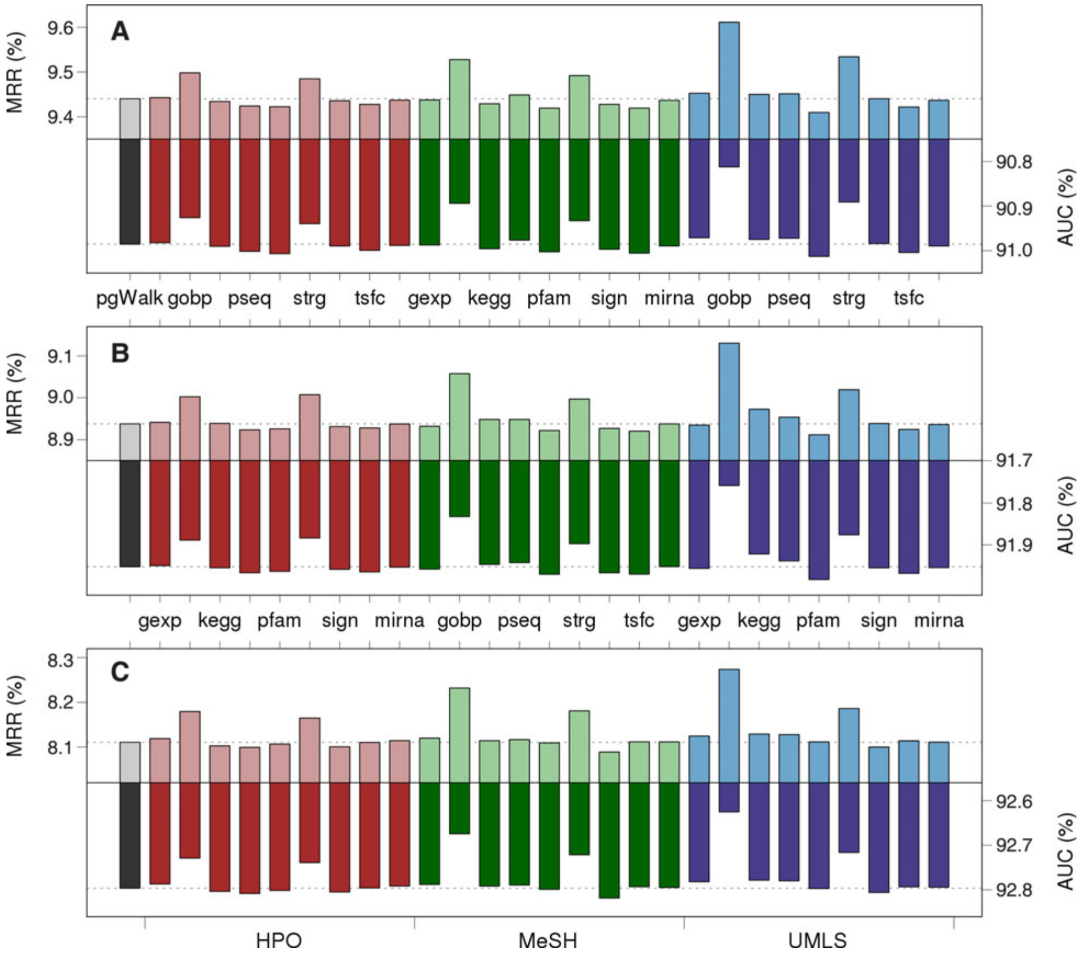

**Figure 6** Contributions of individual disease-gene networks. Results are obtained by excluding individual disease-gene networks in the calculation of the combined *p*-value. Bars are MRR (light colored) and AUC (dark colored) for validation against a linkage interval (**A**), nearest neighbors (**B**), and random controls (**C**).

Downloaded from https://academic.oup.com/jmcb/article/7/3/214/890187 by guest on 21 August 2022

the change of MRR after and before the removal of the feature. Then, we repeated the experiment with a pair of features excluded to obtain contributions of the feature pair. Finally, we calculated a raw interaction score for a feature pair as the contribution of the pair subtracting the maximum contributions of the two individual features, and we divided raw scores by the maximum of such scores to obtain the final scores. We then applied a hierarchical cluster analysis to the resulting matrix that obtained interaction scores between the features and showed the results in Figure 7.

We observe that interactions related to a small set of features (region 1, the first seven rows and columns in the figure) are stronger than those among the other features, and a one-sided Wilcoxon rank sum test strongly supports this observation ($p$-value $= 2.29 \times 10^{-33}$). We notice that five features in this set are related to UMLS. We also observe positive interactions between four features related to HPO and seven features related to either MeSH or UMLS (region 2). We then regress the strength of interaction between two features against their correlation coefficient and obtain a model with very small $r^2$ ($3.79 \times 10^{-3}$), suggesting that interactions between features are not related to their correlations. We further measure the importance of a feature by summing over all interaction scores related to the feature and find the resulting quantity is strongly correlated with the weight of the feature (Pearson's correlation coefficient $= 0.7445$, $p$-value $= 8.46 \times 10^{-6}$). As for individual features, we observe that the gene ontology and protein–protein interaction have positive interactions with most other genomic data sources, no matter which type of phenomic data is used, revealing the importance of these two data sources. We finally measure the importance of a data source by summing over all scores to which the data source contributes and order the nine genomic data sources according to their importance (from the most to the least) as the gene ontology, protein–protein interaction, KEGG pathway, protein sequence, microRNA regulation, gene expression, signaling network, transcriptional regulation, and protein domain. These results complement the analysis in the previous section about contributions of individual data sources.

We further evaluated pairwise interactions of the 12 data sources by adopting an approach similar to the above one, except that we removed all disease-gene networks related to a data source when considering its importance. Results show that every data source has positive interactions with some other data (Supplementary Figure S5), suggesting the necessary of including all data sources in our method. For genomic data, the gene ontology and protein–protein interaction have a strong positive interaction. For phenomic data, UMLS and MeSH have a strong positive interaction. Moreover, medium positive interactions exist among the three genomic data sources and four genomic data (the gene ontology, protein–protein interaction, KEGG pathway, protein sequence), revealing the collaborative effects between phenomic and genomic data.

Downloaded from https://academic.oup.com/jmcb/article/13/3/214/890187 by guest on 21 August 2022

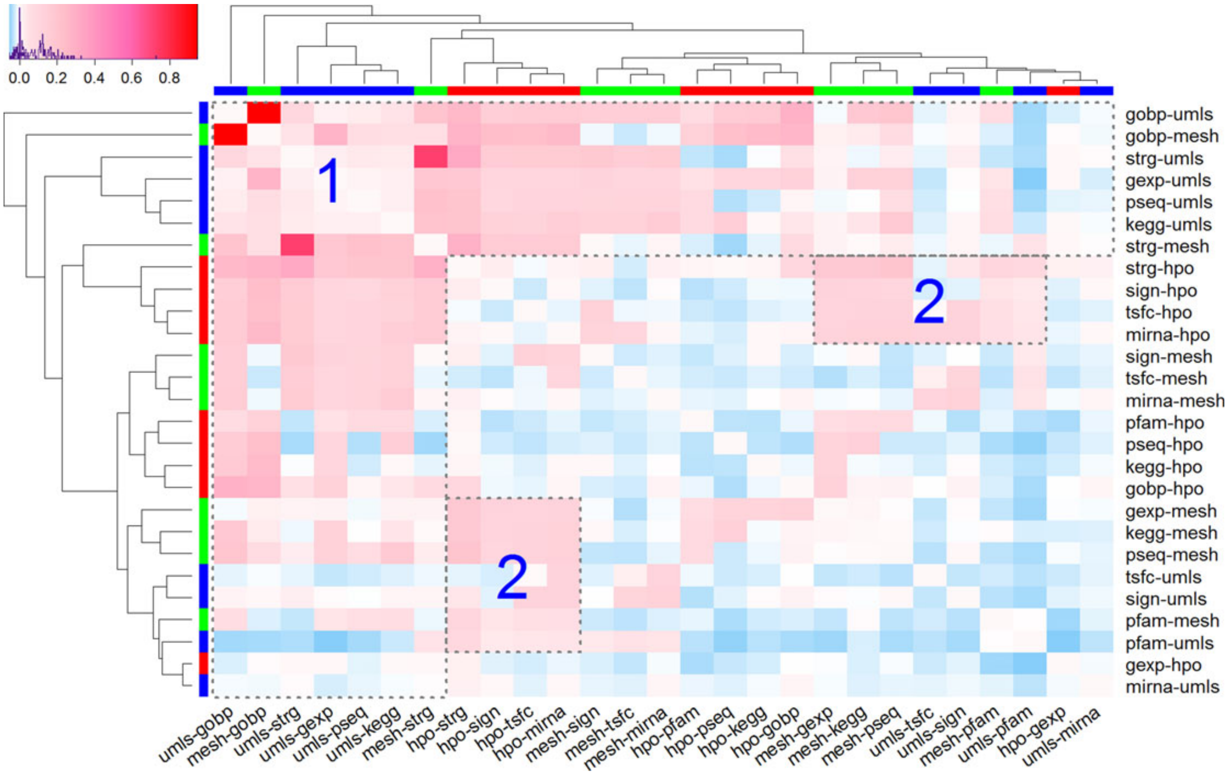

**Figure 7** Pairwise interactions among individual disease-gene networks. Results are obtained by excluding a pair of disease-gene networks (features) in the calculation of the combined $p$-value. Interactions related to a small set of features (region 1) are stronger than those among other features, and five features in this set are related to UMLS. Region 2 contains positive interactions among four features related to HPO and seven features related to either MeSH or UMLS.

*Applications to exome sequencing studies*

The recent advancement in exome sequencing studies has demonstrated that individual *de novo* mutations occurring in individual genes could be the major cause of Mendelian diseases such as Schinzel–Giedion syndrome (Hoischen et al., 2010), Kabuki syndrome (Li et al., 2011) and Bohring–Opitz syndrome (Hoischen et al., 2011), while the collection of *de novo* mutations affecting different genes might explain a proportion of common complex diseases such as autism (O’Roak et al., 2012) and epileptic encephalopathies (Epi4K Consortium & Epilepsy Phenome/Genome Project, 2013). We therefore apply our method to the exome sequencing data of these two types of complex diseases to demonstrate the power of our method in diagnosing disease genes.

Autism (MIM: 209850) is a neurological and developmental disorder that usually appears during childhood, especially the first 3 years of life. Although the exact cause of autism is unknown, many investigations have suggested that autism is a complex genetic disease with a strong genetic basis. Recent studies based on exome sequencing have shown that highly disruptive nonsense and splice site *de novo* mutations in brain-expressed genes exhibit strong associations with autism, revealing potential large impacts of *de novo* mutations on the pathogenesis of this disease. From the literature (PMID 22495309) (O’Roak et al., 2012), we collected 153 unique candidate genes that contained 154 unique *de novo* mutations, and 34 of these genes were reported as likely functional in the literature. When looking at the results produced by our method with the assumption that genetic bases of this disease were completely unknown (Table 4), we found that all genes ranked among top 3 were reported as functional, yielding a *p*-value of 0.01 according to the one-sided Fisher’s exact test against the alternative hypothesis that the probability of observing three functional genes among top 3 is significantly higher than the random guess. Moreover, 4 genes ranked among top 5, 8 genes ranked among top 10, and 13 genes ranked among top 20 were reported as likely functional, yielding *p*-values of $8.64 \times 10^{-3}$, $8.87 \times 10^{-5}$, and $9.52 \times 10^{-6}$, respectively. At the *q*-value cutoff value 0.01, all three candidate genes were reported as likely functional, yielding a *p*-value of 0.01. At the *q*-value cutoff value 0.25, 7 out of 9 candidate genes were reported as likely functional, yielding a *p*-value of $3.84 \times 10^{-4}$. All these results strongly support the capability of our method in identifying disease genes for this complex disease.

**Table 4** Top 20 candidate genes for autism (MIM: 209850).

| Rank | Chromosome | Gene | *p*-value | *q*-value | Functional |
|---|---|---|---|---|---|
| **1** | **14** | **CHD8** | **8.84E-07** | **1.35E-04** | **Yes** |
| **2** | **10** | **PTEN** | **9.43E-06** | **7.22E-04** | **Yes** |
| **3** | **3** | **NLGN1** | **2.75E-05** | **1.40E-03** | **Yes** |
| 4 | 1 | ST3GAL3 | 1.67E-03 | 6.40E-02 | No |
| **5** | **X** | **RPS6KA3** | **4.72E-03** | **1.45E-01** | **Yes** |
| 6 | 3 | EIF4G1 | 6.91E-03 | 1.76E-01 | No |
| **7** | **14** | **PSEN1** | **9.88E-03** | **2.08E-01** | **Yes** |
| **8** | **3** | **CTNNB1** | **1.09E-02** | **2.08E-01** | **Yes** |
| **9** | **8** | **CHD7** | **1.22E-02** | **2.08E-01** | **Yes** |
| **10** | **5** | **PCDHB4** | **2.10E-02** | **3.22E-01** | **Yes** |
| 11 | 3 | LAMB2 | 2.66E-02 | 3.70E-01 | No |
| **12** | **17** | **CHD3** | **3.98E-02** | **5.08E-01** | **Yes** |
| 13 | 15 | STARD9 | 4.82E-02 | 5.67E-01 | No |
| **14** | **12** | **MDM2** | **5.99E-02** | **6.23E-01** | **Yes** |
| 15 | 2 | TTN | 6.10E-02 | 6.23E-01 | No |
| 16 | 17 | MYH10 | 8.14E-02 | 7.32E-01 | No |
| **17** | **19** | **NOTCH3** | **8.16E-02** | **7.32E-01** | **Yes** |
| **18** | **16** | **TSC2** | **8.61E-02** | **7.32E-01** | **Yes** |
| **19** | **3** | **RUVBL1** | **1.02E-01** | **7.94E-01** | **Yes** |
| 20 | 8 | BMP1 | 1.10E-01 | 7.94E-01 | No |

A total of 153 candidate genes were collected, among which 34 were reported as likely functional in the literature (PMID 22495309). pgWalk ranked 8 such functional genes among top 10.

**Table 5** Top 20 candidate genes for epileptic encephalopathies (MIM: 615369).

| Rank | Chromosome | Gene | *p*-value | *q*-value | Functional |
|---|---|---|---|---|---|
| **1** | **15** | **GABRB3** | **7.34E-08** | **1.31E-05** | **Yes** |
| **2** | **2** | **SCN1A** | **2.66E-07** | **1.90E-05** | **Yes** |
| **3** | **12** | **SCN8A** | **3.19E-07** | **1.90E-05** | **Yes** |
| **4** | **2** | **SCN2A** | **6.64E-07** | **2.70E-05** | **Yes** |
| **5** | **20** | **KCNQ2** | **7.53E-07** | **2.70E-05** | **Yes** |
| 6 | 16 | GNAO1 | 1.45E-06 | 4.03E-05 | No |
| **7** | **12** | **GRIN2B** | **1.57E-06** | **4.03E-05** | **Yes** |
| **8** | **9** | **STXBP1** | **2.71E-06** | **5.41E-05** | **Yes** |
| **9** | **8** | **KCNQ3** | **2.72E-06** | **5.41E-05** | **Yes** |
| **10** | **9** | **KCNT1** | **3.33E-06** | **5.96E-05** | **Yes** |
| **11** | **5** | **GABRA1** | **5.72E-06** | **9.30E-05** | **Yes** |
| **12** | **X** | **CDKL5** | **1.04E-05** | **1.56E-04** | **Yes** |
| 13 | 4 | GABRB1 | 2.94E-05 | 3.88E-04 | No |
| **14** | **X** | **ALG13** | **3.04E-05** | **3.88E-04** | **Yes** |
| 15 | 5 | GPR98 | 3.25E-05 | 3.88E-04 | No |
| **16** | **9** | **GRIN1** | **6.42E-05** | **7.19E-04** | **Yes** |
| **17** | **19** | **CACNA1A** | **1.70E-03** | **1.79E-02** | **Yes** |
| **18** | **10** | **ANK3** | **2.94E-03** | **2.93E-02** | **Yes** |
| **19** | **X** | **FLNA** | **3.41E-03** | **3.21E-02** | **Yes** |
| 20 | 1 | NFASC | 6.06E-03 | 5.42E-02 | No |

A total of 179 candidate genes were collected, among which 19 were reported as likely functional in the literature (PMID 23934111). pgWalk ranked 9 such functional genes among top 10.

Epileptic encephalopathies (MIM: 615369) refer to a group of severe childhood epilepsy disorders for which the cause remains largely unknown. Recently, exome sequencing was successfully applied to the study of this group of complex diseases, showing strong statistical evidence on the association of several *de novo* mutations with epileptic encephalopathies (PMID 23934111) (Epi4K Consortium & Epilepsy Phenome/Genome Project, 2013). From the sequencing data of 264 probands and their parents in this study, we collected 179 unique candidate genes that contained 192 unique *de novo* mutations, and 19 of these genes were reported as likely functional. When looking at the results produced by our method with the assumption that genetic bases of this disease were completely unknown (Table 5), we found that all genes ranked among top 5 were reported as functional, yielding a *p*-value of $8.03 \times 10^{-6}$ according to the one-sided Fisher’s exact test against the alternative hypothesis that the probability of observing five functional genes among top 5 is significantly higher than the random guess. Moreover, 9 genes ranked among top 10 and 16 genes ranked among top 20 were reported as likely functional, yielding *p*-values of $2.06 \times 10^{-9}$ and $1.64 \times 10^{-16}$, respectively. At the *q*-value cutoff value 0.001, 13 out of 16 candidate genes were reported as likely functional, yielding a

Downloaded from https://academic.oup.com/jmcb/article/7/3/214/890187 by guest on 21 August 2022

$p$-value of $6.89 \times 10^{-13}$. At the $q$-value cutoff value 0.05, 16 out of 19 candidate genes were reported as likely functional, yielding a $p$-value of $3.35 \times 10^{-17}$. All these results strongly support the capability of our method in identifying disease genes for this complex disease.

### *Whole-genome scan of disease genes*

We finally performed a whole-genome scan of causative genes using our method for a total of 7719 diseases included in either of the phenomic data, with the focus on a total of 20327 genes contained in either of the genomic data. Prediction results, together with an online service and the standalone software of pgWalk, are available at http://bioinfo.au.tsinghua.edu.cn/jianglab/pgwalk.

## Discussion

In this paper, we prioritize candidate genes by integrating three types of phenomic data and nine types of genomic data. Specifically, we construct a disease-gene network based on a pair of phenotype data and genomic data, and we score the strength of association between a candidate gene and a query disease by simulating the process that a random walker wanders on such a heterogeneous network. We further adopt a weighted version of the Fisher's method with dependent correction to integrate 27 scores obtained this way and calibrate a final $q$-value for prioritizing the candidate genes. We conducted a series of validation experiments to demonstrate the superior performance of this approach and further show the effectiveness of our method in exome sequencing studies about neurological diseases.

The success of our method can be attributed to several aspects. First, our method relies on the combination of three types of phenomic data and nine types of genomic data, thereby utilizing more comprehensive information than those that use genomic data only or the combination of a single phenomic data and a single genomic data. Second, our method is designed based on the random walk model, which has been demonstrated to be one of the most effective methods in gene prioritization. Third, we ground the data integration strategy on a carefully designed statistical model and systematically consider crucial issues such as $p$-value calibration, dependence correction, feature weighting, and multiple testing correction.

Certainly, our method can further be extended from the following aspects. First, although protein-coding genes have received the most attention in the study of disease-related genetic risk factors, non-coding functional elements such as lncRNAs have been proved to be of great importance in the development of diseases. How to extend our method to infer the relationship of these elements and a query disease would be one of our future direction. Second, the problem of biasedness toward well-studied genes has been recognized in many studies. This bias issue can be alleviated with the integration of multiple types of data, because the data integration strategy does not depend on a single type of data to make inference. However, how to explicitly eliminate the influence of bias is still an open question worth exploration. Finally, our method provides a means for solving two basic questions in dealing with heterogeneous data, the comparability of different types of data and the integration of multiple types of data. Therefore, it is natural to adopt our method to address problems such as the inference of disease-related genetic variants (Wu et al., 2014). How to make such an extension is another focus of our future direction.

## Materials and methods

### *Data sources*

We focused our study on 7719 diseases extracted from the OMIM database (accessed in November 2014) and 20327 genes obtained from the Ensembl database (accessed in November 2014). Using the tool BioMart (Haider et al., 2009), we extracted 4606 associations between 3933 diseases and 3028 genes, and we calculated that on average each disease was associated with 1.17 genes, and each gene was relevant to 1.52 diseases.

For diseases, we derived a pairwise phenotype similarity matrix for 6376 diseases by applying a text mining technique to the OMIM records of these diseases with the use of the human phenotype ontology (HPO) as the standard vocabulary. In a similar way, we derived a second similarity matrix based on the medical subject headings (MeSH) and a third one relying on the unified medical language system (UMLS), both for 7719 diseases. Based on each of these matrices, we constructed a disease network by keeping only 20 nearest neighbors for each disease. The coverage of each such network is shown in Table 1.

For genes, we derived a pairwise expression similarity matrix (gexp for short) for 12462 genes based on gene expression data that measured whole genome transcripts across 79 human tissues. We derived a semantic similarity matrix (gobp) for 15602 genes based on the biological process domain of the gene ontology and associated annotations for human genes (both released on November 22, 2014). We derived a pathway similarity matrix (kegg) for 6468 genes based on 283 KEGG pathway annotated for human (released on March 11, 2014). We derived a sequence similarity matrix (pseq) between 14196 genes based on a total of 20272 human protein sequences from the Swiss-Prot database (release 2014_01). We derived a domain similarities matrix (pfam) for 17091 genes relying on a total of 14831 protein domains from the Pfam database (version 27.0). We derived a network similarity matrix (strg) for 12432 genes relying on a total of 403514 interactions between 13747 human proteins extracted from the STRING database (version 9.1). We derived a signaling similarity matrix (sign) for 5995 genes relying on a total of 62937 interactions between 6305 human genes downloaded from Edwin Wang's lab (version 6). We derived a transcriptional regulation similarity matrix (tsfc) for 20314 genes based on high quality position specific scoring matrices for 218 vertebrate transcription factors obtained from the TRANSFAC database (release 2013.1). We derived a microRNA regulation similarity matrix (mirna) for 17552 genes based on high-quality predictions of microRNA targets obtained from miRanda (release 2010.8). Putting together, we obtained a total of 20327 genes that were present in at least one of the nine data sources. We then constructed a gene network based on each of these matrices by keeping only 100 nearest neighbors for each gene. The coverage of the resulting networks is shown in Table 1.

Downloaded from https://academic.oup.com/jmcb/article/7/3/214/890187 by guest on 21 August 2022

*Construction of disease networks*

We adopted a text mining technique to derive phenotype similarity matrices and constructed disease networks accordingly. First, focusing on HPO and associated annotations for 6376 human diseases (Robinson et al., 2008), we collected 10777 concepts in the annotations and characterized each disease using a numeric vector of such number of dimensions. Here, an element in the vector was the information content of the corresponding concept, calculated as the negative logarithm of its occurrence frequency in the annotations. Considering the directed acyclic graph (DAG) structure of the ontology, we added the occurrence frequency of a concept to its parents recursively. For a pair of diseases, we calculated the cosine of the angle between the corresponding vectors to obtain their similarity scores. Note that although there have been quite a few methods for calculating semantic similarity based on an ontology, it has been shown recently that the cosine measure, though simple, often produces reasonably good results (Gan, 2014). Applying the above method to every pair of diseases, we obtained a phenotype similarity matrix for human diseases. Although this matrix can be regarded as the weight matrix of a fully connected disease network, such a network may contain a large number of low confident edges between diseases of small phenotype similarities. We therefore kept only $\alpha$ (defaulting to 20) neighboring diseases of the highest similarity scores for each disease and obtained a nearest neighbor network, referred to as HPO. According to the literature (Jiang et al., 2011), the final result is quite robust to the parameter $\alpha$.

Then, we extracted 7719 disease records from the OMIM database and split sentences in the TX and CS fields of these records into words. Mapping these words onto MeSH concepts by using the MetaMap program (Aronson, 2001), we obtained 5632 concepts for describing human diseases. Counting the occurrence frequency of each of these concepts in an OMIM record, we obtained a high dimensional numeric vector for the record. We then calculated pairwise phenotype similarity between diseases as the cosine of the angle between corresponding vectors and further constructed a disease network referred to as MeSH using the aforementioned nearest neighbor strategy. Finally, we replaced the MeSH concepts with 7745 UMLS ones and repeated the above procedure to construct a disease network referred to as UMLS.

*Construction of gene networks*

We constructed nine gene networks based on a variety of genomic data, including the gene expression, gene ontology, pathway membership, protein sequence, protein domain, protein–protein interaction, signaling network, transcriptional regulation, and microRNA regulation.

We characterized each human gene using a 79-dimensional numeric vector that represented expression levels of the gene across the same number of tissues, relying on whole-genome microarrays for 44775 transcripts across human tissues (Su et al., 2004). For a pair of genes, we calculated the absolute value of the Pearson's correlation coefficient of the corresponding vectors to obtain their raw similarity scores. Considering that such raw scores may include noise in the original expression data, we further applied an exponential transformation to convert raw scores into final similarity scores, as

$$\varphi_{gh}^{(\mathrm{gexp})} = \exp\left[-\lambda\left(\frac{1-\omega_{gh}^{(\mathrm{gexp})}}{\sigma_{gh}^{(\mathrm{gexp})}}\right)^{2}\right],$$

where $\varphi_{gh}^{(\mathrm{gexp})}$ was the final score for two genes $g$ and $h$, $\omega_{gh}^{(\mathrm{gexp})}$ the raw score, $\sigma^{(\mathrm{gexp})}$ the standard deviation of raw scores for all gene pairs, and $\lambda$ a tuning parameter with defaulting value 1. With this transformation, the highest raw score (1.0) kept highest, while the lowest raw score (0.0) became $\exp(-\lambda(\sigma_{gh}^{(\mathrm{gexp})})^{-2})$, which was close to zero because the standard deviation $\sigma^{(\mathrm{gexp})}$ was typically small. Applying the above method to every pair of genes, we obtain a gene similarity matrix. With a similar reasoning as for diseases, we kept only $\beta$ (with default value 100) neighboring gene of the highest similarity scores for each gene and obtained a nearest neighbor network (gexp). According to the literature (Jiang et al., 2011), the final result is quite robust to the parameter $\beta$.

We collected 26784 concepts from the biological process domain of the gene ontology (Ashburner et al., 2000) and characterized each human gene using a numeric vector of such number of dimensions. Here, each element in a vector was the information content of the corresponding concept. We calculated the raw similarity scores between a pair of genes as the cosine of the angle between the corresponding vectors and applied the exponential transformation ($\lambda = 0.1$) to convert raw scores into final similarity scores. We further constructed a gene network (gobp) using the nearest neighbor strategy.

We collected 238 human pathways from the KEGG database (Kanehisa and Goto, 2000) (with disease-related ones discarded to avoid biases toward well-studied diseases) and characterized each human gene using a binary vector of such number of dimensions. We then calculated the raw similarity scores between a pair of genes as the cosine of the angle between the corresponding vectors, applied the exponential transformation ($\lambda = 1$) to obtain final similarity scores, and further constructed a gene network (kegg) using the nearest neighbor strategy.

We calculated pairwise local sequence alignments of 20274 human protein sequences extracted from the Swiss-Prot database (Apweiler et al., 2004) using the Smith-Waterman algorithm implemented in SSEARCH (Li et al., 2012). We then constructed a sequence similarity network of these proteins by connecting two proteins with an undirected edge if their alignment e-value is less than a predefined threshold ($10^{-4}$). Next, we calculated the shortest path distance ($\delta_{gh}^{(\mathrm{pseq})}$) for every pair of proteins ($g$ and $h$) in this network and converted it into a similarity value in the range of 0 and 1 ($\omega_{gh}^{(\mathrm{pseq})} = 1 - \delta_{gh}^{(\mathrm{pseq})}/\max\delta_{gh}^{(\mathrm{pseq})}$). Finally, we applied the exponential transformation ($\lambda = 1$) to obtain the similarity score and further constructed a gene network (pseq) using the nearest neighbor strategy. Note that the construction of a sequence similarity network in this procedure greatly reduced the sensitivity to the parameters involved and thus enhanced the robustness of this method.

We obtained a total of 14831 domains from the Pfam database (Version 27.0) (Bateman et al., 2004) and characterized each

Downloaded from https://academic.oup.com/jmcb/article/7/3/214/890187 by guest on 21 August 2022

human protein using a binary vector of such number of dimensions. For a pair of two genes, we calculated the cosine of the angle between the corresponding vectors to obtain their raw similarity scores and further applied the exponential transformation ($\lambda = 1$) to obtain final similarity scores. We further constructed a gene network (pfam) using the nearest neighbor strategy.

We extracted a total of 403514 interactions among 13747 proteins from the STRING database (Version 9.1) (Snel et al., 2000) and constructed a protein–protein interaction network accordingly. Then, we calculated the shortest path distance ($\delta_{gh}^{(\mathrm{strg})}$) for every pair of proteins ($g$ and $h$) in this network and converted it into a value in the range of 0 and 1 ($\omega_{gh}^{(strg)} = 1 - \delta_{gh}^{(strg)}/\max\delta_{gh}^{(strg)}$). Finally, we applied the exponential transformation ($\lambda = 1$) to obtain the similarity score and constructed a gene network (strg) using the nearest neighbor strategy.

We identified a total of 62937 signaling actions (33398 activation, 7960 inhibition, and 21579 physical interaction) between 6305 genes from a manually curated human signaling network downloaded from Edwin Wang's lab (Version 6) (Cui et al., 2007) and transformed these actions into a graph by assigning directed edges to activation or inhibition relationships and undirected edges to physical interactions. Then, we calculated the shortest path distance ($\delta_{gh}^{(\mathrm{sign})}$) for every pair of proteins ($g$ and $h$) in this network and converted it into a value in the range of 0 and 1 ($\omega_{gh}^{(\mathrm{sign})} = 1 - \delta_{gh}^{(\mathrm{sign})}/\max\delta_{gh}^{(\mathrm{sign})}$). Finally, we applied the exponential transformation ($\lambda = 1$) to obtain the similarity score and constructed a gene network (sign) using the nearest neighbor strategy.

We extracted 218 high confidence position specific scoring matrices for the same number of vertebrate transcription factors from the TRANSFAC database (Matys et al., 2003) and searched 1000 basepairs upstream for each human gene using the program MATCH to identify potential binding sites for each transcription factor. Then, we characterized each gene using a numeric vector of 218 dimensions, with each element indexing the number of potential binding sites for the corresponding transcription factor and calculated the raw similarity scores between a pair of genes as the cosine of the angle between the corresponding vectors. Finally, we applied the exponential transformation ($\lambda = 1$) to obtain the similarity score and constructed a gene network (tsfc) using the nearest neighbor strategy.

We extracted 249 microRNAs collected in the miRanda database (Betel et al., 2008) and characterized each gene using a binary vector of such number of dimensions, with each element denoting whether the gene had been predicted as a target of the corresponding microRNA. We then calculated the raw similarity scores between a pair of genes as the cosine of the angle between the corresponding vectors. Finally, we applied the exponential transformation ($\lambda = 1$) to obtain the similarity score and constructed a gene network (mirna) using the nearest neighbor strategy.

### *Walking on a disease-gene network to score association strength*

Given a disease network, a gene network and known associations between diseases and genes, we constructed a heterogeneous network whose nodes included both diseases and genes, and we simulated the process that a random walker wandered on such a disease-gene network to score the strength of association between a disease and a gene (Li and Patra, 2010).

In detail, a disease-gene network included a disease layer, a gene layer, and interconnections between these two layers. The disease layer, which can be selected as one of the three aforementioned disease networks, is composed of diseases and their relationships. The gene layer, which can be selected as one of the nine gene networks constructed before, is composed of gene and their connections. Interconnections, which connect diseases and genes, are obtained from known associations between diseases and genes. Given a query disease of interest, a random walker starts a journey in a disease-gene network with some initial probability $\mathbf{p}^{(0)}$. Then, in each step of the journey, the walker may select to start a new journey with probability $\pi$ or move on with probability $1 - \pi$. When moving on, the walker may select to jump from the disease layer to the gene layer or vice versa with probability $\tau$ or choose to wander in either the disease or the gene layer with probability $1 - \tau$. When wandering about, the walker moves to one of its direct neighbors. After a number of steps, the probability that the walker stays in each node of the disease-gene network would reach a steady state $\mathbf{p}^{(\infty)}$, which gives a measure of the strength of association between the query disease and genes in the gene layer.

In mathematics, a heterogeneous network is denoted by a triple $\mathbf{H} = (\mathbf{D}, \mathbf{G}, \mathbf{A})$, where $\mathbf{D} = (d_{ij})_{m\times m}$ is the weight matrix of the disease layer, $\mathbf{G} = (g_{ij})_{n\times n}$ that of the gene layer, $\mathbf{A} = (a_{ij})_{m\times n}$ the adjacency matrix of the interconnections, and $m$ and $n$ the numbers of diseases and genes, respectively. Applying row-normalization to $\mathbf{D}$, we obtain a transition matrix $\mathbf{U} = (u_{ij})_{m\times m}$, where $u_{ij} = d_{ij}/\sum_{j=1}^{m} d_{ij}$ denotes the probability that the walker moves from the $i$-th disease to the $j$-th disease when it stays in the former. Similarly, we obtain three other transition matrices: $\mathbf{V} = (v_{ij})_{n\times n}$ with $v_{ij} = g_{ij}/\sum_{j=1}^{n} g_{ij}$ denoting the probability that the walker moves from the $i$-th gene to the $j$-th gene when it stays in the former, $\mathbf{R} = (r_{ij})_{m\times n}$ with $r_{ij} = a_{ij}/\sum_{j=1}^{n} a_{ij}$ ($r_{ij} = 0 \quad \text{if } \sum_{j=1}^{n} a_{ij} = 0$) being the probability that the walker jumps from the $i$-th disease to the $j$-th gene when it stays in the former, and $\mathbf{S} = (s_{ij})_{n\times m}$ with $s_{ij} = a_{ij}/\sum_{j=1}^{m} a_{ij}(s_{ij} = 0 \quad if\ \sum_{j=1}^{m} a_{ij} = 0)$ being the probability that the walker jumps from the $i$-th gene to the $j$-th disease when it stays in the former. We then define matrix $\mathbf{T}$ as

$$\mathbf{T} = \begin{pmatrix} (1-\tau)\mathbf{U} & \tau\mathbf{R} \\ \tau\mathbf{S} & (1-\tau)\mathbf{V} \end{pmatrix},$$

and perform row-normalization to obtain the transition matrix for the heterogeneous network as $\mathbf{W} = (w_{ij})_{(m+n)\times(m+n)}$, where $w_{ij} = t_{ij}/\sum_{j=1}^{m+n} t_{ij}$ and $\tau$ the probability of jumping from the disease layer to the gene layer or vice versa.

Let $\mathbf{u}^{(0)} = (u_i^{(0)})_{m\times 1}$ and $\mathbf{v}^{(0)} = (v_i^{(0)})_{n\times 1}$ be initial probabilities for the disease and the gene layers, respectively. We obtain $\mathbf{u}^{(0)}$ by assigning probabilities proportional to disease similarities to neighbors of the query disease and 0 otherwise, and we set $\mathbf{v}^{(0)}$ to zeros to simulate the situation that genetic basis for the query disease is completely unknown. Let $\mathbf{p}^{(0)} = ((\mathbf{u}^{(0)})^T, (\mathbf{v}^{(0)})^T)^T$ contains initial probabilities for the heterogeneous network and $\mathbf{p}^{(t)}$

Downloaded from https://academic.oup.com/jmcb/article/7/3/214/890187 by guest on 21 August 2022

contains probabilities that the walker stays at each node at time $t$, we have the iterative formula

$$\mathbf{p}^{(t+1)} = (1-\pi)\mathbf{W}^T\mathbf{p}^{(t)} + \pi\mathbf{p}^{(0)}.$$

Repeating the iteration a number of steps until $\mathbf{p}^{(t)}$ is stable (e.g. the $L_1$ norm of $\Delta\mathbf{p} = \mathbf{p}^{(t+1)} - \mathbf{p}^{(t)}$ is less than a small positive number $\varepsilon$), we obtain the steady-state probability $\mathbf{p}^{(\infty)}$, which can be decomposed into a disease part $\mathbf{u}^{(\infty)} = (u_i^{(\infty)})_{m\times 1}$ and a gene part $\mathbf{v}^{(\infty)} = (v_i^{(\infty)})_{n\times 1}$. The later one, $\mathbf{v}^{(\infty)}$, can then be used to score the strength of association between the query disease and genes. It has been show that the random walk model is not sensitive to the parameters involved in the model (Li and Patra, 2010; Jiang et al., 2011). We therefore set default values for the parameters as $\tau = 0.5$, $\pi = 0.7$, and $\varepsilon = 10^{-4}$. An alternative approach is to solve the linear equation $\mathbf{p}^{(\infty)} = (1-\pi)\mathbf{W}^T\mathbf{p}^{(\infty)} + \pi\mathbf{p}^{(0)}$ with respect to the steady-state probability $\mathbf{p}^{(\infty)}$ and obtain $\mathbf{p}^{(\infty)} = \pi(\mathbf{I} - (1-\pi)\mathbf{W}^T)^{-1}\mathbf{p}^{(0)}$ directly. In literature, the simulation method is more frequently used, while the matrix inversion method is suitable for the situation where $\mathbf{W}$ is fixed.

Although the steady-state probability itself could serve as a score to characterize the strength of association between the query disease and a gene, the calibration of a $p$-value to indicate the statistical significance of such a steady-state score would be helpful in many applications. For this purpose, we simulate the distribution of the steady-state scores for all disease-gene pairs that are not included in annotated associations and derive a $p$-value as the proportion of scores in this distribution that is greater than or equal to the score that needs to be calibrated. In other words, we define

$$p = P(\text{scores of non-annotated disease-gene pairs} \geq \text{the score in calibration}).$$

The meaning of such a $p$-value is therefore the probability of observing stronger association scores under the null hypothesis that a gene is not associated with a disease.

### *Integration of multiple association scores*

We adopted Fisher's method to integrate $p$-value derived from different disease-gene networks to obtain a single $p$-value, with efforts on both the weighting scheme of the data sources and the correction of dependence between the $p$-values.

Specifically, given the $p$-values to be combined, denoted by $p_1,\ldots,p_K$, where $K = 27$ is the total number of data sources, we define a statistic as

$$U = \sum_{i=1}^{K} w_i V_i$$

where $w_i$ is the weight of the $i$-th data source and $V_i = -2\log p_i$. It is clear that under the null hypothesis, $p_i \sim \text{Uniform}(0,1)$ and $V_i \sim \chi_2^2$. In the dependent and weighted case, we assume that under the null hypothesis $U$ follows a scaled chi-squared distribution with scale $\eta$ and degrees of freedom $v$. Resorting to the method of moments, we derive the population mean and variance as

$$\mathrm{E}[\eta\chi_v^2] = \eta v \text{ and } \mathrm{Var}[\eta\chi_v^2] = 2\eta^2 v,$$

and the corresponding sample mean and variance as

$$\mathrm{E}[U] = 2\sum_{i=1}^{K} w_i \text{ and } \mathrm{Var}[U] = \sum_{i=1}^{K}\sum_{j=1}^{K} w_i w_j \mathrm{cov}(V_i, V_j).$$

Matching these quantities for the population and the sample, we obtain

$$\hat{\eta} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{K} w_i w_j \mathrm{cov}(V_i, V_j)}{4\sum_{i=1}^{K} w_i} \text{ and } \hat{v} = \frac{2}{\hat{\eta}}\sum_{i=1}^{K} w_i.$$

Covariances cov($V_i$,$V_j$) can be estimated using a normal model as follows. Suppose $p_i = \Phi(1 - z_i)$, where $\Phi(\bullet)$ is the cumulative distribution function of the standard normal distribution and $Z_i$ a statistic that has a standard normal distribution under the null hypothesis. As suggested in the literature (Yang, 2010), let

$$\hat{\rho}_{ij} = \mathrm{Cor}(Z_i, Z_j) \text{ and } \tilde{\rho}_{ij} = \hat{\rho}_{ij}\left(1 + \frac{1-\hat{\rho}_{ij}^2}{2n-1}\right).$$

The covariance is then calculated as

$$\mathrm{Cov}(V_i, V_j) \approx a_1\tilde{\rho}_{ij} + a_2\tilde{\rho}_{ij}^2 + a_3\tilde{\rho}_{ij}^3 + a_4\tilde{\rho}_{ij}^4,$$

where $a_1 = 3.263119$, $a_2 = 0.709866$, $a_3 = 0.026589$, $a_4 = -0.709866/n$, $n$ the sample size for obtaining $Z_i$.

Nevertheless, the determination of optimal weights is far from trivial. We therefore adopt an empirical strategy to obtain a set of weights that reflect the relative goodness of individual data sources. We first resort to a cross-validation experiment to measure the performance of a disease-gene network and quantify its effectiveness using a criterion called the mean rank ratio (MRR, see results part for details). Then, we calculate a raw weight as $\tilde{w}_i = \exp(-\gamma r_i / \min_{i=1}^{K} r_i)$, where $r_i$ is the MRR of the $i$-th data source and $\min_{i=1}^{K} r_i$ the minimum MRR of all data sources. Finally, we normalize over all such raw weights to obtain the final weight as $w_i = \tilde{w}_i / \sum_{i=1}^{K} \tilde{w}_i$. Obviously, a data source with higher performance would have larger weight, and that with lower performance would have smaller weight. We set the parameter $\gamma = 2.5$ by default in our study. A grid search show that our method is quite robust to this parameter, and 2.5 is near to the optimal value.

We further apply multiple testing corrections to the combined $p$-values by controlling the positive false discovery rate (pFDR) of candidate genes through their $q$-values (Storey, 2003). Existing studies have shown the significant improvement in the test power of this method over the traditional approach of Benjamini–Hochberg that controls the false discovery rate (FDR) (Benjamini and Hochberg, 1995). It is possible that some data sources are absent for a candidate gene. To deal with this

Downloaded from https://academic.oup.com/jmcb/article/7/3/214/890187 by guest on 21 August 2022

problem, we ignore the missing data source in the Fisher's method and decrease the total number of $p$-values to be combined accordingly.

## Supplementary material

Supplementary material is available at *Journal of Molecular Cell Biology* online.

## Acknowledgements

The author would like to thank the anonymous reviewers for suggested revisions that significantly improved the presentation of this paper.

## Funding

This work was supported by the National Basic Research Program of China (2012CB316504), the National High Technology Research and Development Program of China (2012AA020401), and the National Natural Science Foundation of China (61175002).

**Conflict of interest:** none declared.

## References

Adie, E.A., Adams, R.R., Evans, K.L., et al. (2005). Speeding disease gene discovery by sequence based candidate prioritization. BMC Bioinformatics *6*, 55.

Aerts, S., Lambrechts, D., Maity, S., et al. (2006). Gene prioritization through genomic data fusion. Nat. Biotechnol. *24*, 537–544.

Altshuler, D., Daly, M., and Kruglyak, L. (2000). Guilt by association. Nat. Genet. *26*, 135–138.

Apweiler, R., Bairoch, A., Wu, C.H., et al. (2004). UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. *32*, D115–D119.

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

Ashburner, M., Ball, C.A., Blake, J.A., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. *25*, 25–29.

Bamshad, M.J., Ng, S.B., Bigham, A.W., et al. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. *12*, 745–755.

Bateman, A., Coin, L., Durbin, R., et al. (2004). The Pfam protein families database. Nucleic Acids Res. *32*, D138–D141.

Becker, K.G., Barnes, K.C., Bright, T.J., et al. (2004). The genetic association database. Nat. Genet. *36*, 431–432.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B *57*, 289–300.

Betel, D., Wilson, M., Gabow, A., et al. (2008). The microRNA.org resource: targets and expression. Nucleic Acids Res. *36*, D149–D153.

Chen, Y., Jiang, T., and Jiang, R. (2011). Uncover disease genes by maximizing information flow in the phenome-interactome network. Bioinformatics *27*, i167–i176.

Cui, Q., Ma, Y., Jaramillo, M., et al. (2007). A map of human cancer signaling. Mol. Syst. Biol. *3*, 152.

Emilsson, V., Thorleifsson, G., Zhang, B., et al. (2008). Genetics of gene expression and its effect on disease. Nature *452*, 423–428.

Epi4K Consortium & Epilepsy Phenome/Genome Project. (2013). De novo mutations in epileptic encephalopathies. Nature *501*, 217–221.

Freudenberg, J., and Propping, P. (2002). A similarity-based method for genome-wide prediction of disease-relevant human genes. Bioinformatics *18(Suppl 2)*, S110–S115.

Gan, M. (2014). Correlating information contents of gene ontology terms to infer semantic similarity of gene products. Comput. Math. Methods Med. *2014*, 891842.

Haider, S., Ballester, B., Smedley, D., et al. (2009). BioMart Central Portal—unified access to biological data. Nucleic Acids Res. *37*, W23–W27.

Hamosh, A., Scott, A.F., Amberger, J.S., et al. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. *33*, D514–D517.

Hoischen, A., van Bon, B.W., Gilissen, C., et al. (2010). De novo mutations of SETBP1 cause Schinzel-Giedion syndrome. Nat. Genet. *42*, 483–485.

Hoischen, A., van Bon, B.W., Rodriguez-Santiago, B., et al. (2011). De novo nonsense mutations in ASXL1 cause Bohring-Opitz syndrome. Nat. Genet. *43*, 729–731.

Jiang, R., Gan, M., and He, P. (2011). Constructing a gene semantic similarity network for the inference of disease genes. BMC Syst. Biol. *5(Suppl 2)*, S2.

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

Keshava Prasad, T.S., Goel, R., Kandasamy, K., et al. (2009). Human Protein Reference Database—2009 update. Nucleic Acids Res. *37*, D767–D772.

Köhler, S., Bauer, S., Horn, D., et al. (2008). Walking the interactome for prioritization of candidate disease genes. Am. J. Hum. Genet. *82*, 949–958.

Lage, K., Karlberg, E.O., Storling, Z.M., et al. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol. *25*, 309–316.

Li, Y., and Patra, J.C. (2010). Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics *26*, 1219–1224.

Li, Y., Bogershausen, N., Alanay, Y., et al. (2011). A mutation screen in patients with Kabuki syndrome. Hum. Genet. *130*, 715–724.

Li, W., McWilliam, H., Goujon, M., et al. (2012). PSI-Search: iterative HOE-reduced profile SSEARCH searching. Bioinformatics *28*, 1650–1651.

Lindberg, D.A., Humphreys, B.L., and McCray, A.T. (1993). The Unified Medical Language System. Methods Inf. Med. *32*, 281–291.

Lipscomb, C.E. (2000). Medical Subject Headings (MeSH). Bull. Med. Libr. Assoc. *88*, 265–266.

Lopez-Bigas, N., and Ouzounis, C.A. (2004). Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res. *32*, 3108–3114.

Matys, V., Fricke, E., Geffers, R., et al. (2003). TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. *31*, 374–378.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., et al. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat. Rev. Genet. *9*, 356–369.

Moreau, Y., and Tranchevent, L.C. (2012). Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat. Rev. Genet. *13*, 523–536.

O'Roak, B.J., Vives, L., Girirajan, S., et al. (2012). Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature *485*, 246–250.

Ott, J., Kamatani, Y., and Lathrop, M. (2011). Family-based designs for genome-wide association studies. Nat. Rev. Genet. *12*, 465–474.

Robinson, P.N., Kohler, S., Bauer, S., et al. (2008). The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am. J. Hum. Genet. *83*, 610–615.

Snel, B., Lehmann, G., Bork, P., et al. (2000). STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res. *28*, 3442–3444.

Storey, J.D. (2003). The positive false discovery rate: a Bayesian interpretation and the $q$-value. Ann. Stat. *31*, 2013–2035.

Su, A.I., Wiltshire, T., Batalov, S., et al. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl Acad. Sci. USA *101*, 6062–6067.

Tiffin, N., Kelso, J.F., Powell, A.R., et al. (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates. Nucleic Acids Res. *33*, 1544–1552.

Turner, F.S., Clutterbuck, D.R., and Semple, C.A. (2003). POCUS: mining genomic sequence annotation to predict disease genes. Genome Biol. *4*, R75.

van Driel, M.A., Bruggeman, J., Vriend, G., et al. (2006). A text-mining analysis of the human phenome. Eur. J. Hum. Genet. *14*, 535–542.

Downloaded from https://academic.oup.com/jmcb/article/7/3/214/890187 by guest on 21 August 2022

Vanunu, O., Magger, O., Ruppin, E., et al. (2010). Associating genes and protein complexes with disease via network propagation. PLoS Comput. Biol. *6*, e1000641.

Wu, X., Jiang, R., Zhang, M.Q., et al. (2008). Network-based global inference of human disease genes. Mol. Syst. Biol. *4*, 189.

Wu, X., Liu, Q., and Jiang, R. (2009). Align human interactome with phenome to identify causative genes and networks underlying disease families. Bioinformatics *25*, 98–104.

Wu, J., Li, Y., and Jiang, R. (2014). Integrating multiple genomic data to predict disease-causing nonsynonymous single nucleotide variants in exome sequencing studies. PLoS Genet. *10*, e1004237.

Yang, J.J. (2010). Distribution of Fisher's combination statistic when the tests are dependent. J. Stat. Comput. Simul. *80*, 1–12.

Downloaded from https://academic.oup.com/jmcb/article/7/3/214/890187 by guest on 21 August 2022