Towards Prediction and Prioritization of disease genes by the modularity of human phenome-genome assembled network

Jeffrey Q Jiang[1*†], Andreas W M Dress[1] and Ming Chen[2]

[1]CAS-MPG Partner Institute for Computational Biology, Shanghai 200031, China

[2]College of Life Sciences, Zhejiang University, Hangzhou 310058, China

Summary

Empirical clinical studies on the human interactome and phenome not only illustrates prevalent phenotypic overlap and genetic overlap between diseases, but also reveals a modular organization of the genetic landscape of human disease, provding new opportunities to reduce the complexity in dissecting the phenotype-genotype association. We here introduce a network-module based method towards phenotype-genotype association inference and disease gene identification. This approach incorporates protein-protein interaction network, phenotype similarity network and known phenotype-genotype associations into an assembled network. We then decomposes the resulted network into modules (or communities) wherein we identified and prioritized the disease genes from the candidates within the loci associated with the query disease using a linear regression model and concordance score. For the known phenotype-gene associations in the OMIM database, we used the leave-one-out validation to evaluate the feasibility of our method, and successfully ranked known disease genes at top 1 in 887 out of 1807 cases. Moreover, applying this approach on 850 OMIM loci characterized by an unknown molecular basis, we propose high-probability candidates for 81 genetic diseases.

1 Introduction

Deciphering genotypes underlying specific phenotype, especially human disease, is one of the principal goals for genetics research and is of vital importance of biomedicine. While many human genetic diseases are caused by multiple genes, it has been increasingly recognized that, because the mutations of these genes lead to disease with overlapping clinical phenotypes, these genes are likely to be functionally related[3, 14, 32], and such functional relatedness can be exploited to identify novel disease genes[21, 11]. This discovery spurs the transition from traditional genetic mapping, typically, positional cloning and linkage analysis, to a new *guilt-by-association* paradigm [11] for disease gene identification.

Indeed, this concept has been applied to search for or prioritize disease genes by various computational methods, including, for example, functional relatedness based on Gene Ontology annotations (GO)[8, 29], gene expression profiles[18, 26] and protein-protein interaction networks (PPIs)[32, 21]. In particular, PPIs are a strong manifestion of such functional relatedness among these genes. In fact, several genetically heterogeneous hereditary diseases, such as

*To whom correspondence should be addressed. E-mail: qiangjiang2006@gmail.com

†Present address: Department of Computer Science, City University of Hong Kong

Hermansky-Pudlak syndrome and Fanconi anemia, are caused by mutations of the genes whose protein products interact to each other[6, 16]; in breast cancer, more than a half of proteins from about 100 mutated cancer genes formed a tight cluster in a PPI network [15]. The premise of the PPI-based method is the assumption that a PPI network-neighbor of a disease-causing gene is more likely (than randomly chosen gene) to be related to the disease [21, 11]. Therefore, for prioritizing or predicting disease-causing genes, the goal is that, given one or more small genomic regions (e.g. locus predicted by a linkage analysis), or some disease-causing genes, how to rank a small number of candidate genes based on their likelihood to be disease-causing derived from a PPI network. This requires new systematical network-based methods to quantifying the association between a gene and a disease.

An recent endeavor[32] in the genome-wide inference of disease genes have shown that a simple linear regression model efficiently capture the underlying architecture of the human interactome and phenome networks and suggests a global concordance of the topology between the phenotype network and the gene network. Although it achieves remarkable success, it is not available to rank candidates and select plausible susceptibility genes for a query disease in a short time due to the huge computational time for a network with large size. As suggested in a recent investigation[22], fortunately, human disease shows a modular organization on the genetic landscape. Hence, an intuitive idea towards the above-mentioned method is clustering disease based on their phenotypic similarities and making predictions for query disease within the same phenotype cluster. Inspired by this idea, we introduce a network module-based method that integrates PPI network, phenotype similarity network and known phenotype-gene associations, and then decomposes the resulted assembled network into modules (or communities) wherein we identified or prioritized the disease genes from the candidates within the loci associated with the query disease using the linear regression model and concordance score proposed in [32].

We first verify the hypothesis of our method that human disease shows a modular organization on the genetic landscape and the modularity is significantly correlated with disease classification. In addition, to evaluate the feasibility of our method, we used the *leave-one-out* validation on the known phenotype-genotype associations in OMIM database—Online Mendelian Inheritance in Man database [23], and successfully ranked known disease genes at top 1 in 887 out of 1807 cases. Moreover, applying this approach on 850 OMIM loci characterized by an unknown molecular basis, we propose high-probability candidates for 81 genetic diseases.

2 Materials and methods

Fig. 1 schematically illustrates our method.

2.1 Phenotype similarity network

A phenotype network consisting of disease phenotypes as nodes and phenotypic similarity as edges was constructed by van Driel et al. [30] using the OMIM database[23]. Considering some OMIM records have since been moved to other records, we removed them from the network to avoid technical error. We obtained a network with 4146 phenotypes and 29489 edges using a similarity of 0.5 as the threshold. However, selection of the threshold had no significant effect

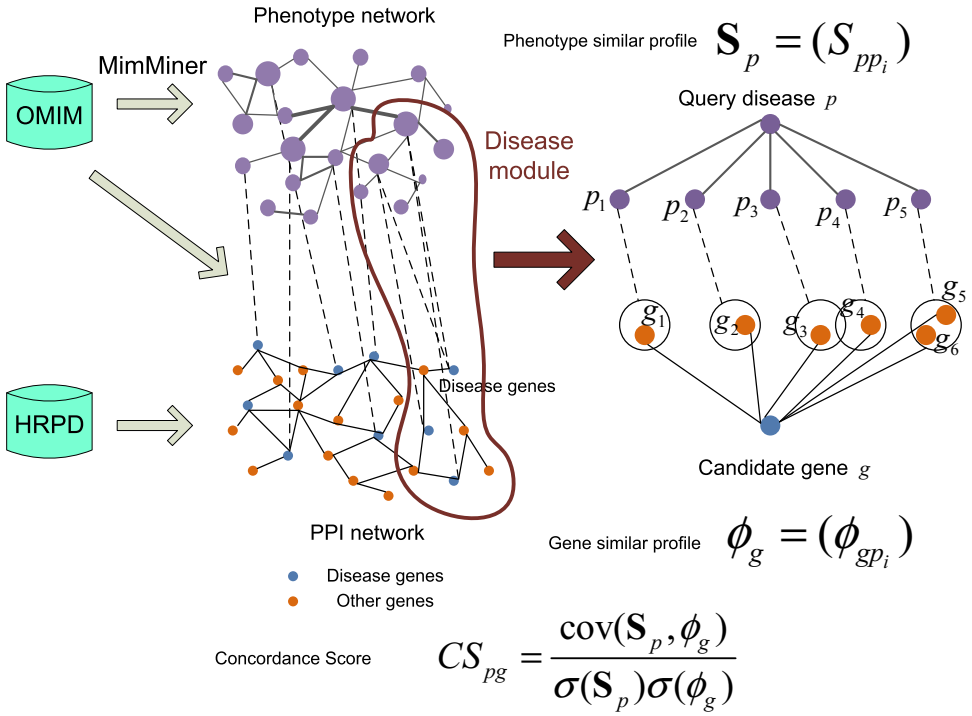


Figure 1: Sketch map of network module-based method for phenotype-gene association.

on the modularity of the disease network (Table 1). In order to link phenotypes in the network to known disease genes, OMIM genotype-phenotype associations for 1807 genes and 2265 disease phenotypes were downloaded on December 2 2008 from Ensembl [10] using the data mining tool BioMART[13].

We used the classification of disease phenotypes which was manually established in [9] to evaluate the significance of modularity in the phenotype network. 2929 phenotypes were classified into 22 primary disorder classes based on the physiological system affected. We did not merge several phenotypes into a single disorder as Goh et al. did[9], and we obtained 1184 phenotypes within 21 disease classes in our phenotype network.

2.2 PPI network

We obtain 34364 manually curated PPIs between 8919 human proteins from the HRPD database [27] and call this resulted PPI network ”GeneNet1”. In order to avoid biased towards better-studied proteins, we also obtain 33049 predicted human PPIs between 7185 (4116 are absent from HPRD) proteins from the OPHID database[4], which is built by mapping PPIs from high-throughput screen of model organisms to human. The extended protein network, called ”GeneNet2”, combines HPRD, OPHID and two other curated PPI databases: BIND[1] and MINT[5], yielding a network of 72431 unique pairwise binary interactions between 14433 human proteins.

Table 1: The threshold of disease phenotype similarity has no significant influence on the modularity

Threshold	Nodes	Edges	Modularity
0.45	4,681	56,752	0.717
0.475	4,425	40,482	0.748
0.5	4,146	29,489	0.783
0.525	3,783	21,778	0.807
0.55	3,381	16,349	0.824
0.575	3,008	12,520	0.833
0.6	2,575	9,592	0.833
0.625	2,202	7,479	0.831
0.65	1,888	5,890	0.833

2.3 Computation of dyadicity D and heterophilicity H

Dyadicity (D) and heterophilicity (H) are two network properties of nodes which were recently proposed [25] for quantifying whether nodes with similar characteristics have a tendency to link to each other. We used these parameters to investigate whether phenotypes in a specific physiological class tend to cluster together in our phenotype network. The value of a phenotype depends on whether it belongs (1) or does not belong (0) to a disease class. Thus three types of links between phenotypes exists: 1-1, 1-0, and 0-0; the number of these links are termed m_{11}, m_{10} and m_{00}, respectively. The two parameters dyadicity D and heterophilicity H are defined as:

$$D := \frac{m_{11}}{\bar{m}_{11}} \quad H = \frac{m_{10}}{\bar{m}_{10}}$$

If $D \gg 1$ and $H \ll 1$, phenotypes in the specific disease class must have a clear clustering tendency within the network.

The expected value of $\bar{m}_{11}$ and $\bar{m}_{10}$ is computed next. If we take cancer as an example, we can call n_1 the number of phenotypes belonging to cancer and n_0 the number of other phenotypes. $N = n_1 + n_0$ is the total number of phenotypes and is the total number of edges in the network. Let $p := \frac{2M}{N(N-1)}$ represent the connectance that indicates the average probability that two phenotypes are connected in the network. The value of a phenotype depends on whether it belongs to a cancer class (1), or does not (0). The three varieties of link styles between phenotypes are 1-1, 1-0, and 0-0, and the number of these links can be labeled as m_{11}, m_{10} and respectively. If any phenotype in the network has an equal chance of being cancer, the expected values of m_{11} and m_{10} are $\bar{m}_{11}$ and $\bar{m}_{10}$ respectively[25]

$$\bar{m}_{11} = \binom{n_1}{2} \times p = \frac{n_1(n_1-1)}{2}p$$

$$\bar{m}_{10} = \binom{n_1}{1}\binom{n_0}{1} \times p = n_1(N-n_2)p$$

Statistically significant deviations of m_{11} and m_{10} from their expected values of $\bar{m}_{11}$ and $\bar{m}_{10}$ imply that cancer phenotypes are not distributed randomly in the phenotype network.

Dyadicity $D > 1$ ($D < 1$) indicates that phenotypes in the disease class tend to connect more (less) densely among themselves than expected for a random configuration. Similarly, heterophilicity $H > 1$ ($H < 1$) means that phenotypes in the disease class have more (fewer) connections to phenotypes in other classes than expected randomly. If $D \gg 1$ and $H \ll 1$, phenotypes in the specific disease class must have a clear clustering tendency within the network.

2.4 Extracting the modules of the phenotype network

We detected modules in the phenotype network using the spectral algorithm based on modularity Q defined as (more details, see [20] and reference therein)

$$Q := \sum_{i=1}^{m} \left[e_{ii} - (\sum_{j} e_{ij})^2 \right]$$

where m is the number of modules, e_{ii} are the fraction of the edges that connect two nodes inside a module i, and e_{ij} are the fraction of the edges connecting nodes of module i and j. The modularity Q of a partition is high when the number of intra-module edges is much larger than expected for a random partition. We identified modules by maximizing the modularity Q so that there were many intra-module edges and few between-module edges. However the method could not identify the hierarchical structure of the modules. Therefore, we decomposed all modules which had more than 100 phenotypes into sub-modules.

The number of final modules which are based in the secondary level of modularity may affect the results. We managed to reduce the effect by visually inspecting each sub-network with more than 100 phenotypes in the first level modules while automatically decomposing the phenotype network using our previous algorithm[12].

2.5 Computing p-value for the disease class enrichment of modules in the phenotype networks

We then used the disease classification dataset to see if the disease phenotypes within a single module tended to fall within the same disease class. We utilized the method introduced in [31] that computes a p-value for the functional enrichment of modules in PPI networks. Take cancer as an example. For a given module M we randomly selected a set of phenotypes which had the same number of members as M, and counted how many of them are cancer. The p-value was calculated as the probability that the number of cancer phenotypes in a random group would be equal to or greater than what we observed in M. We used 100,000 simulations to obtain the p-values.

2.6 Regression model and the concordance score

In each module, we used the regression model proposed in [32]. They assumes that additivity of the contribution to phenotype similarity from different disease genes and is defined as

$$S_{pp'} = C_p + \sum_{g \in G(p)} \sum_{g' \in G(p')} \beta_{pg} e^{-d^2_{gg'}} \quad (1)$$

where $S_{pp'}$ is the similarity score between a query phenotype p and another phenotype p', and $d_{gg'}$ is the topological distance between gene g and g' on the PPI network. $G(p)$ denotes all disease genes belonging to the phenotype p. The Gaussian kernel $e^{-d^2_{gg'}}$ is used to transfer gene-gene distance to gene-gene closeness. C_pi s a constant, and β_{pg} is the coefficient of this regression model, respectively. C_p could be explained as the basal similarity between p and other phenotypes whose causative genes are not connected to those of p in the protein network, and β_{pg} represents the level of the gene g contributing to the similarity of the phenotype p to any other phenotype p'. We consider the topological distance $d_{gg'}$ as the graph theory SP length between genes g and g' in the protein network.

To quantify the association between a phenotype and a gene, we define the closeness of gene g to phenotype p' as the summation of gene-gene closeness from gene g to all disease genes of phenotype p', as

$$\Phi_{gp'} = \sum_{g' \in G(p')} e^{-d^2_{gg'}}$$

Hence, equation (1) can be written as

$$\mathbf{S}_p = C_p + \sum_{g \in G(p)} \beta_{pg} \mathbf{\Phi}_g \quad (2)$$

where the vectors $\mathbf{S}_p = (S_{pp_1}, \ldots, S_{pp_n})$ and $\mathbf{\Phi}_g = (\Phi_{gp_1}, \ldots, \Phi_{gp_n})$ are described the similarities between the query phenotype p and all n phenotype in the same module and the closeness between gene g and all these n phenotypes respectively. Thus, in this linear regression model, we define the Pearson correlation coefficient as the concordance score

$$CS_{pg} = \frac{\mathrm{cov}(\mathbf{S}_p, \mathbf{\Phi}_g)}{\sigma(\mathbf{S}_p)\sigma(\mathbf{\Phi}_g)} \quad (3)$$

where cov and σ mean covariance and standard deviation, respectively. This concordance score measures the consistency between the position of gene g in the protein network and the variations of phenotype similarity for phenotype p in the whole phenotype network. It is then used to rank all the candidate genes for a specific phenotype.

2.7 Benchmark tests

A leave-one-out cross-validation procedure is used to assess the performance of our method. In this procedure, we remove the direct link between true disease gene g and phenotype p, and see if the method can recover this link (rank gene g at the top of the N test genes). This is carried out by taking known disease gene g as unknown when calculating $\Phi_{g_i p}$, the closeness from test gene g_i to query phenotype p. For phenotypes with more than one known causative genes, we modified the definition of a successful prediction: for a test case (p, g) in which p has k known disease genes, if gene g is among the top k-ranked genes, we consider it a successful prediction.

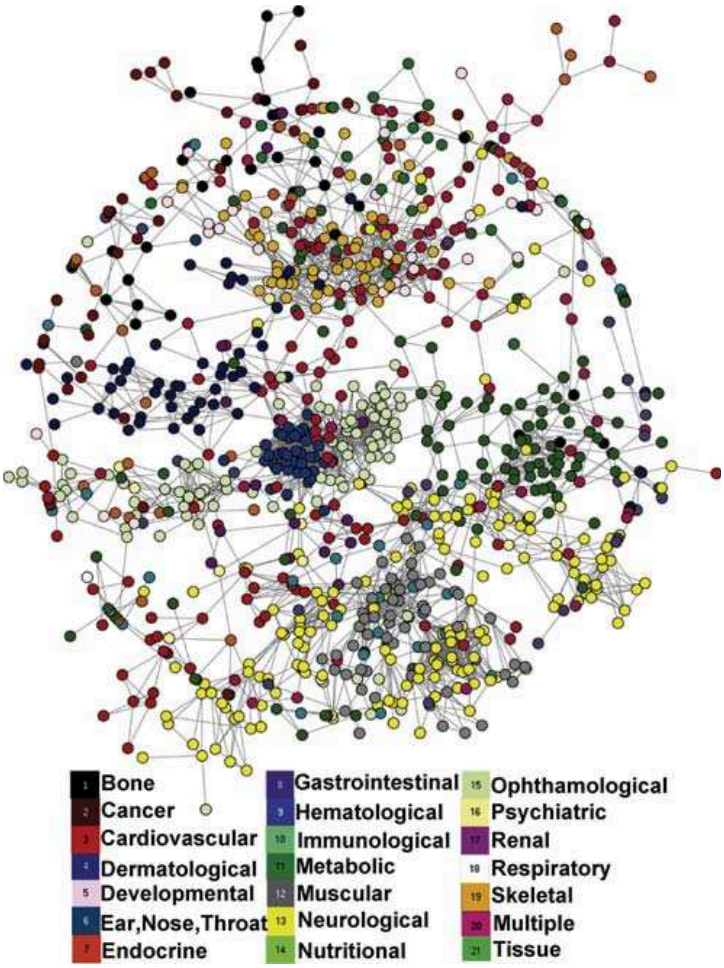


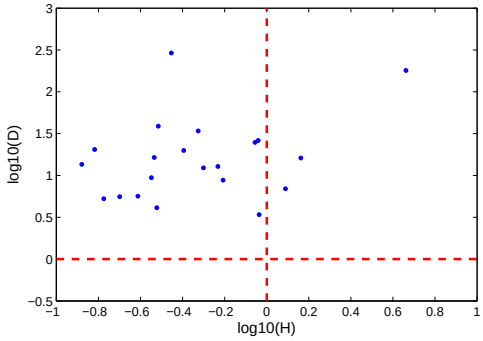
Figure 2: Phenotype network with 21 disease classes. Different colors indicate different disease classes and unclassified phenotypes are not shown.

3 Results

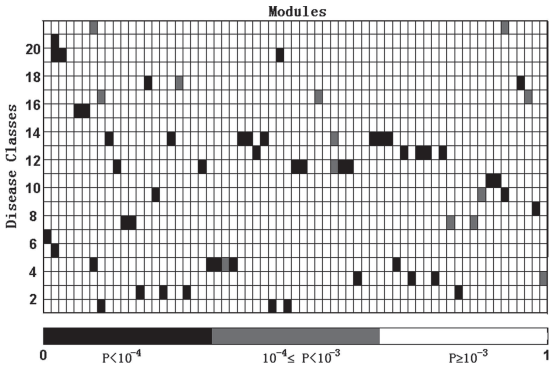
3.1 Modularity of human phenotype network

Visualization of the human phenotype network using 21 disease classes indicated that phenotypes within the same disease class are clustered into densely connected groups (Fig.2). Most of the disease classes are dyadic ($D \gg 1$) and heterophobic ($H \ll 1$), revealing a highly modular structure (Fig. 3(a)). However, a few disease classes are heterophilic, suggesting that they tend to have phenotypes that overlap among different categories of diseases. These diseases include developmental, skeletal and 'ear, nose, throat'. For instance, developmental diseases, in which a delay occurs in physical or mental development, tend to overlap with other diseases. This is logical because most developmental disorders would be expected to affect multiple tissues. An interesting observation is that although phenotypes in the 'ear, nose, throat' class have strong heterophilicity with other disease classes, the dyadicity of this class is very large ($\sim$180). This suggests that the 'ear, nose, throat' class may be a densely connected part of the network even though it has many connections to phenotypes in other classes.

We used the spectral algorithm to decompose the phenotype network into modules based on modularity Q. The maximal modularity Q equals 0.78, which indicates a distinctly modular



(a) Log-log plot of dyadicity and heterophilicity of 21 disease classes. Most of the disease classes are in the upper-left area ($\log_{10} D \gg 0$ and $\log_{10} H \ll 0$), revealing a highly modular structure.



(b) p-value profile for 21 disease classes in modules. Each small square in the grid shows the statistical significance of the enrichment of a specific function in a module. Significance levels are indicated by different grey scales.

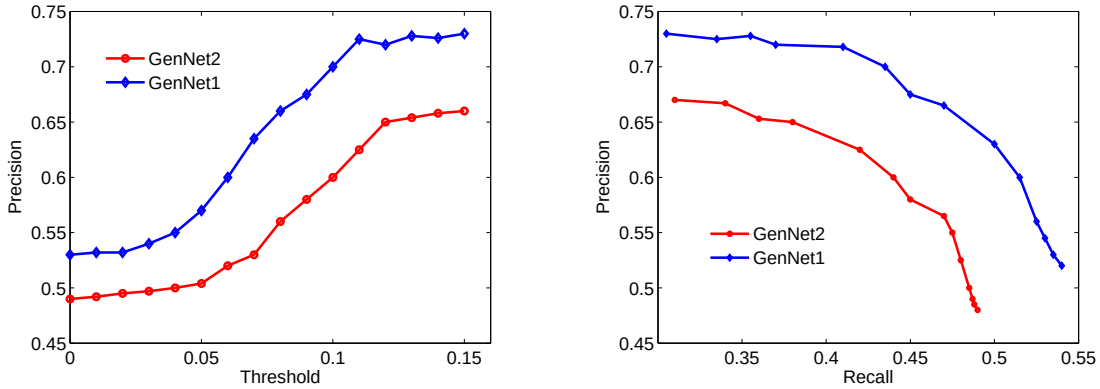
Figure 3: Phenotype network shows a modular organization on the genetic landscape and the modularity is significantly correlated to disease classes.

structure rather than a random network. The network was partitioned into 28 modules in the first partition. In order to identify the hierarchical structure of the modules, we decomposed all modules which had more than 100 phenotypes into sub-modules. If the sub-network of phenotypes in a module had a clear secondary modular structure ($Q \geq 0.5$), we used the sub-modules instead of the first level one. This yields 231 modules, 214 of which were based on the secondary level of modularity (see Supplementary File 1 for more details).

3.2 Modularity is significantly correlated with disease classification

The p-value profile for 21 disease classes in each module with more than five phenotypes is demonstrated in Fig.3(b). Almost all modules were significantly enriched with one or two (three of few modules) disease classes when we used 10^{-3} or 10^{-4} as the cutoff for statistical significance of each class.

van Driel et al. [30] constructed a set of phenotype similarities by text-mining all records that describe genetics disorders in the OMIM database using medical subject headings (MeSH). The nature of the similarity measure ensures that two phenotypes will be connected in the network if they have similar clinical traits. As expected, phenotypes in a disease class tend to group in the network. However, they can be divided into many different modules. For example, phenotypes in the neurological disease class are distributed into about ten modules; of them, one module contains primarily ataxia phenotypes, such as spinocerebellar ataxia and cerebellar ataxia; and one module contains mostly Charcot-Marie-Tooth disease phenotypes. Thus modules generally are subclasses of the primary disease classes. In addition, in several cases, some phenotypes in different disease classes may be grouped together because they have similar clinical traits. These results indicate that the network method can not only provide a computational validation of the disease classification which was determined manually by Goh et al. [9], but also provide a more specific classification of disease phenotypes.



(a) Score threshold plotted against precision. (b) Recall plotted against precision.

Figure 4: Performance of the leave one-out cross validation.

3.3 Benchmark test:leave-one-out validation

To examine how well our method reflects the biological truth, we took each of the 1807 known gene-phenotype association as one test case, and for each case we generate an artificial loci[14, 26].

For each gene-disease link, we simulate a linkage locus around the true disease gene by including 181 neighboring genes as negative controls to simulate the median size of linkage intervals for OMIM phenotype loci with unknown molecular basis [14](see Supplementary Figure S3 for the distribution of gene numbers in all the disease loci). This strategy for resembling known disease loci in the OMIM database has been widely used in previous studies[14, 32]. The 181 test genes are then treated equally by assuming links to the disease under study and go through the network prediction procedure. We then calculate the concordance score for each test gene, and rank the test genes according to the score. If the known disease gene is ranked as top 1, we consider it a *successful prediction*, and we define the *precision* as the proportion of successful predictions among all predictions. We set a threshold and only make prediction when the highest score of all test genes in a case is no less than it, and define *recall* as the fraction of true disease genes predicted among all disease genes[14]. A *leave-one-out cross-validation* (see Materials and Methods) shows our method can at least rank known disease genes at top 1 in 887 out of 1807 test cases, achieving a precision of 0.49 and a recall of 0.49. With the increasing of the threshold, the precision becomes larger while the recall becomes smaller. For high-scoring candidates, the precision can approach 0.73 while maintaining a high recall of 0.31 (Fig.4). Compared with GeneNet1, the precision using GeneNet2 is smaller, varying from 0.49 to 0.65. The reason is that there are many genes in network GeneNet2 have been not well-studied and their association with disease have been not identified.

3.4 Prioritization candidates in disease loci

The above results indicate that our method can efficiently predict the human disease genes from genetic loci. Therefore, we applied our procedure to 850 OMIM phenotype entries with at least one mapped disease locus but unknown molecular basis. We obtained predictions for 81 loci, only top 20 of which are summarized in Table 2. All the predictions are available in

Table 2: Candidate genes found for top 20 of 81 OMIM loci with unknown molecular basis.

OMIM ID	locus	locus size	HUGO	Ensembl ID
119540	2q32	75	COL3A1	ENSG00000168542
			COL5A2	ENSG00000204262
121210	8q13-21	188	C8orf46	ENSG00000169085
			EFCBP1	ENSG00000123119
			RALYL	ENSG00000184672
			STMN2	ENSG00000104435
			GDAP1	ENSG00000104381
			C8orf34	ENSG00000165084
			STAU2	ENSG00000040341
130080	12p13	277	A2M	ENSG00000175899
			VWF	ENSG00000110799
			C1S	ENSG00000182326
			MFAP5	ENSG00000197614
			EMP1	ENSG00000134531
			CD163	ENSG00000177575
			TNFRSF1A	ENSG00000067182
			TSPAN9	ENSG00000011105
			CSDA	ENSG00000060138
			LTBR	ENSG00000111321
			CD9	ENSG00000010278

Supplementary Table S2. We obtained predictions for 81 loci, 67 of which where only from the GeneNet1, 5 only from the GeneNet2 and 9 from both. Interestingly, in 4 of the latter cases, the list of candidates from the two networks contained at least one common gene.

Notably, for three OMIM phenotypes (163000, familial multiple nevi flammei; 268700, saccharopinuria; 300195; AMMECR1) our predictions include the actual disease genes that, although not yet correctly annotated in OMIM, have been found to be mutated in patients.

For 22 loci, at least one of the candidates obtained from either network was already known to be involved in phenotypes similar to those described for the locus. These genes represent the most obvious candidates and our results should be considered as further, independent evidence for their possible involvement in the disease. However, it must be noted that some of them were previously excluded, either by the direct identification of crossovers or by the negative results of mutation screenings. Nevertheless, since mutations have most likely been searched only within the annotated exons, we think that the decision to definitively rule out the involvement of such candidates should be taken cautiously. Moreover, even silent exonic mutations, although often considered innocuous polymorphisms, can have severe effects on proteins by disrupting splicing patterns[24, 7].

Table 3: Comparison of the fold enrichment of several methods, where PPI: protein-protein interaction network, FA: gene functional annotation, GE: gene expression microarray data..

Methods	Data Source	Fold enrichment
Lage et al [14]	PPI	23.1
Perez-Inrateeta et al [26]	FA	19.4
Fredudenberg and Propping[8]	FA	13.3
Oti et al [21]	PPI	10.0
Turner et al [29]	FA	5.2
Wu et al [32]	PPI	53.5
Our method	**PPI**	**37.2**

In most cases only few candidates are given for a locus, thus providing extremely focused working hypotheses for the identification of the actual disease genes, which in many cases are made even stronger by the available sequence or functional information. For instance, one of the two candidates provided for the OMIM phenotype entry 607221 (partial epilepsy with pericentral spikes, located on 4p15) corresponds to KCNIP4[19]. This protein has been show to specifically modulate the activity of Kv4 A-type potassium channels, which are well known regulators of membrane excitability[2] and have been recently involved in epilepsy[28].

Even when the number of candidates for a particular locus is substantially higher, our results may provide a strong restriction of the experimental search field, which can be further narrowed by additional evidences. For instance, the phenotype with OMIM ID 130080 (Ehlers-Danlos syndrome, type VIII), is mapped to 12p13, containing 277 genes. In this case, the GeneNet1 and GeneNet2 networks provide 8 and 4 candidates, respectively. Interestingly, the candidate with the lowest associated score is the Alpha-2-macroglobulin precursor (A2M), whose absence was previously reported in a patient with Ehlers-Danlos syndrome[17].

3.5 Fold enrichment: comparison with other methods

Various methods [14, 26, 21, 8, 29, 32] have been proposed for prioritizing candidate genes, but few of them have reported the precision within their publications. Traditionally, the power of these methods is measured by their ability to enrich known disease genes over random selection, say, fold enrichment [14]: If a method successfully ranks known disease genes in the top m% of all candidate genes in n% of the linkage intervals, there is a n/m-fold enrichment on average. We compared these methods by computing their fold enrichment (listed in Table 3) that illustrate our method's potential.

4 Discussions and Conclusions

The success of our method can be attributed to utilizing the modular nature of human genetic disease that paved a new way for phenotype-gene association studies from several aspects. First, there is now good evidence from bioinformatic analysis that human genetic diseases can be clustered on the basis of their phenotypic similarities and that such a clustering represents

true biological relationships of the genes involved. Second, one may use such phenotypic similarity to predict and then test for the contribution of apparently unrelated genes to the same functional module. Third, one can use bioinformatics to make predictions about new genes for diseases that form part of the same phenotype cluster. This is done by starting from the known disease genes and then searching for genes that share one or more functional attributes such as gene expression pattern, co-evolution, or gene ontology. Ultimately, one may expect that a modular view of disease genes should help the rapid identification of additional disease genes for multifactorial diseases once the first few contributing genes (or environmental factors) have been reliably identified.

Certainly, our approach can be improved in the following directions. First, our method is limited to genes with known protein interactions (about one-third of the entire human genome). Further expanding the protein network to embrace less reliable protein interactions (such as the OPHID network) or non-physical functional associations may increase the power to detect less-studied disease genes in practice. Second, our method suffers from the imprecision and subjectiveness in quantifying phenotype similarity. The continuing endeavor for standardizing and quantifying phenotypic description would further enhance our method. Third, like other methods for disease gene finding, our method cannot tell where the causative genetic variants are in high-rank genes. With the recent progress in the prioritization of candidate genetic variants for human diseases, it is expected that by prioritizing candidate genes and genetic variants at the same time, the two may benefit each other and facilitate the discovery of disease genes and causative genetic variants therein.

Our method also illustrates well the power of the integration of different types of networks. We suggest that the ongoing large-scale mapping of human interaction networks and systematic collection of human phenotypic data are valuable for biomedical research, and the increasing coverage and quality of human interaction network, as well as more standardized and objective phenotype descriptions will facilitate the discovery of new disease genes.

Acknowledgement

We would like to thank Dr Brunner HG and his laboratory for the generosity of providing us with the phenotype network data, and Xiao-yin Zhu and Dong-lin Huang for data extraction and mining.

References

[1] Bader GD, Betel D, Hogue CWV (2003) BIND: the biomolecular interaction network database. *Nucleic Acids Res* **31**, 248-250

[2] Birnbaum SG, Varga AW, Yuan LL, Anderson AE, Sweatt JD, et al. (2004) Structure and function of Kv4-family transient potassium channels. *Physiol Rev* **84**: 803-833.

[3] Brunner HG and van Driel MA (2005) From syndrome families to functional genomics. *Nat. Rev. Genet.* **5**, 545-551

[4] Brown KR, Jurisica I (2005) Online predicted human interaction database. *Bioinformatics* **21**, 2076-2082

[5] Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, Castagnoli L, Cesareni G (2007) MINT: the molecular INTeraction database. *Nucleic Acids Res* **35**, D572-D574

[6] Di Pietro SM, Dell'Angelica EC (2005) The cell biology of Hermansky-Pudlak syndrome: recent advances. *Traffic* **6**, 525-533

[7] Faustino NA, Cooper TA (2003) Pre-mRNA splicing and human disease. *Genes Dev* **17**: 419-437.

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

[9] Goh, K.I., et al.(2007) The human disease network. *Proc. Natl. Acad. Sci. USA* **104**, 8685-8690

[10] Hubbard, T. et al. (2002) The Ensembl genome database project. *Nucleic Acids Res.* **30**, 38-41.

[11] Ideker T and Sharan R (2008) Protein networks in disease. *Genome Res.* **18**, 644-652

[12] Jiang Q., Dress A. W. M., Yang G.(2009) A spectral clustering-based framework for detecting community structures in complex networks. *Appl. Math. Lett.* **22**, 1479-1482

[13] Kasprzyk, A. et al.(2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.* **14**, 160-169.

[14] 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.

[15] Lin J, Gan CM, Zhang X, et al.(2007) A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Res.* **17**, 1304-1318

[16] Mace G, Bogliolo M, Guervilly JH, Dugas du Villard JA, Rosselli F (2005) 3R coordination by Fanconi anemia proteins. *Biochimie* **87**, 647-658

[17] Mahour GH, Song MK, Adham NF, Rinderknecht H (1978) Alpha2-macroglobulin deficiency in a patient with Ehlers-Danlos syndrome. *Pediatrics* **61**: 894-897.

[18] Masseroli, M., Galati, O., Pinciroli, F.(2005) GFINDer: genetic disease and phenotype location statistical analysis and mining of dynamically annotated gene lists. *Nucleic Acids Res.*, **33**, W717-W723.

[19] Morohashi Y, Hatano N, Ohya S, Takikawa R, Watabiki T, et al. (2002) Molecular cloning and characterization of CALP/KChIP4, a novel EF-hand protein interacting with presenilin 2 and voltage-gated potassium channel subunit Kv4. J Biol Chem 277: 14965-14975.

[20] Newman, M.E.J. (2006) Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA*, **103**, 8577-8582

[21] Oti M, Snel B, Huynen MA, Brunner HG (2006) Predicting disease genes using protein-protein interactions. *J Med Genet* **43**, 691-698

[22] Oti M, Brunner HG.(2007) The modular nature of genetic diseases. *Clin Genet.* **71**(1), 1-11

[23] Online Mendelian Inheritance in Man, OMIM (TM). (2009) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). World Wide Web `http://www.ncbi.nlm.nih.gov/omim`

[24] Pagani F, Raponi M, Baralle FE (2005) Synonymous mutations in CFTR exon 12 affect splicing and are not neutral in evolution. *Proc Natl Acad Sci USA* **102**: 6368-6372.

[25] Park J and Barabasi AL. (2007) Distribution of node characteristics in complex networks. *Proc. Natl. Acad. Sci. USA* **104**, 17916-17920

[26] Perez-Iratxeta, C., Bork, P., Andrade, M.A.(2002) Association of genes to genetically inherited diseases using data mining. *Nat. Genet.*, **31**, 316-319.

[27] Peri S, et al (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res* **13**, 2363-2371

[28] Singh B, Ogiwara I, Kaneda M, Tokonami N, Mazaki E, et al. (2006) A Kv4.2 truncation mutation in a patient with temporal lobe epilepsy. *Neurobiol Dis* **24**: 245-253.

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

[30] van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G. and Leunissen, J.A. (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.* **14**, 535-542.

[31] Wang Z. and Zhang J. (2007) In search of the biological significance of modular structures in protein networks. *PLoS Comput. Biol.* **3**, e107

[32] Wu X, Jiang R, Zhang MQ, Li S(2008) Network-based global inference of human disease genes. *Mol. Sys. Biol.* **4**:189.

Supplemantary File for "Towards Prediction and Prioritization of disease genes by the modularity of human phenome-genome assembled network

Jeffrey Q Jiang[1*†], Andreas W M Dress[1] and Ming Chen[2]

[1]CAS-MPG Partner Institute for Computational Biology, Shanghai 200031, China

[2]College of Life Sciences, Zhejiang University, Hangzhou 310058, China

1 Detect the modules/submodules of the phenotype network

From a combinatorial point of view, a network is a simple graph, i.e., a pair $G = (V, E)$ consisting of a set V, called its vertex set and a subset E of the set $\binom{V}{2} := \{e \in V : |e| = 2\}$ called its edge set, associated with a symmetric weight matrix $W = (w)_{uv \in V} \in \Re_{\geq 0}^{V \times V}$. Further, given a network G, a *community structure* is a partition Π of V into a disjoint union of non-empty subsets $V_1, \ldots, V_m$ whose vertices are, intuitively speaking, "more densely connected" to one another than to the other vertices of G.

Module ID	Size	Modularity	Module ID	Size	Modularity
Modu 1	246	0.128	Modu 12	192	0.698
Modu 2	289	0.659	Modu 13	130	0.766
Modu 3	471	0.423	Modu 15	147	0.501
Modu 4	256	0.336	Modu 17	100	0.697
Modu 6	221	0.557	Modu 18	114	0.575
Modu 7	234	0.412	Modu 19	169	0.534
Modu 8	111	0.301	Modu 24	120	0.661
Modu 11	269	0.683	Modu 28	191	0.725

Table 1: Decomposition of large modules into secondary level modules.

A famous quantitative measure for evaluating the "goodness of fit" of a partition Π to G, the modularity function $Q = Q(\Pi)$, was proposed by M.E.J. Newman and M. Girwan in [3] and is represented as

$$Q(\Pi) := \sum_{j=1}^{m} \left[\frac{\mathcal{F}(V_i, V_i)}{\mathcal{F}(V, V)} - \left(\frac{\mathcal{F}(V_i, V)}{\mathcal{F}(V, V)} \right)^2 \right] \tag{1}$$

where $\mathcal{F}(V', V")$ is defined, for any two subsets $V', V"$ of V, by $\mathcal{F}(V', V") := \sum_{u \in V', v \in V"} w_{uv}$.

It was illustrated [4, 1, 2] that a high Q-value indicates that the partition Π represents a "good" community structure for G and, so, much work has been devoted in recent years to designing

*To whom correspondence should be addressed. E-mail: qiangjiang2006@gmail.com
†Present address: Department of Computer Science, City University of Hong Kong

methods proposed towards this end for iteratively improving the Q-measure. The method, however, could not identify the hierarchical structure of the modules and thus we here decomposed all modules which had more than 100 phenotypes into sub-modules. Obviously, the number of final modules depends on the secondary level of modularity that we identified. To reduce the effect, we visually inspect each sub-network with more than 100 phenotypes in the first level modules while automatically decompose the phenotype network using our previous algorithm [2].

The network was partitioned into 28 modules in the first level. We found 16 modules with at least 100 phenotypes, of which 11 modules had a significant secondary level of modularity using Q-measure (Table S1). We scrutinized each sub-network of phenotypes in the modules using the spring-embedded layout in Cytoscap software to check if each sub-network has a clear modular structure. In the resulting spring-embedded layout, nodes with edges between them tend to be situated near each other, whereas nodes without edges between them tend to be spread apart. Finally, we found that the modularity 0.5 is an appropriate threshold value for extracting secondary level modules. Following this way, we identified 231 modules in the end, most of which (214 of 231) are based on the secondary level of modularity. Thus, we believed that this decomposition method will reveal the actual modularity of the phenotype network.

2 Candidate genes for OMIM loci with unknown molecular basis

We give the predicted candidate genes for 81 OMIM loci with unknown molecular basis in the following table.

OMIM ID	locus	locus size	Ensembl ID
119540	2q32	75	ENSG00000168542
			ENSG00000204262
121210	8q13-21	188	ENSG00000169085
			ENSG00000123119
			ENSG00000184672
			ENSG00000104435
			ENSG00000104381
			ENSG00000165084
			ENSG00000040341
130080	12p13	277	ENSG00000175899
			ENSG00000110799
			ENSG00000182326
			ENSG00000197614
			ENSG00000134531
			ENSG00000177575
			ENSG00000067182
			ENSG00000011105
			ENSG00000060138

continued on next page

continued from previous page

OMIM ID	locus	locus size	Ensembl ID
			ENSG00000111321
			ENSG00000010278
142700	13q22	36	ENSG00000102554
145410	22q11.2	271	ENSG00000185214
			ENSG00000100030
			ENSG00000099972
154275	17q11.2-q24	881	ENSG00000067191
			ENSG00000108878
156232	2q24-q32	296	ENSG00000128710
			ENSG00000175879
			ENSG00000115290
			ENSG00000128713
156600	13q31-q32	99	ENSG00000080166
			ENSG00000184564
162820	7q22-qter	712	ENSG00000106123
			ENSG00000146904
163000	5q13-q22	327	ENSG00000164252
			ENSG00000145715
164210	14q32	414	ENSG00000011114
			ENSG00000100697
			ENSG00000100664
			ENSG00000184916
177720	16q23-q24	170	ENSG00000174990
180020	6q25-q26	133	ENSG00000120278
			ENSG00000164674
181430	12q15-q23.1	215	ENSG00000111046
			ENSG00000139289
183600	2q31	126	ENSG00000128710
			ENSG00000128713
			ENSG00000128652
185000	9q34.1	135	ENSG00000148346
203650	3q26.3-q27.3	150	ENSG00000114770
			ENSG00000163898
213200	9q34-qter	290	ENSG00000148408
			ENSG00000176884
213600	14q	1215	ENSG00000171723
214900	15q	1074	ENSG00000140505
218400	6q21-q22	215	ENSG00000152661
225000	11q23-q24	314	ENSG00000186318
			ENSG00000109846
			ENSG00000149591
			ENSG00000168334
255160	3p22.2-p21.32	84	*continued on next page*

continued from previous page

OMIM ID	locus	locus size	Ensembl ID
			ENSG00000010282
			ENSG00000183873
259450	17p12	37	ENSG00000141052
			ENSG00000109099
268700	7q31.3	39	ENSG00000008311
300046	Xq23-q24	116	ENSG00000068366
			ENSG00000131725
300148	Xp22.13-p21.1	117	ENSG00000131828
300195	Xq22.3	40	ENSG00000188153
300324	Xq22.2-q26	327	ENSG00000068366
300489	Xq13.1-q21	182	ENSG00000147166
			ENSG00000131171
309610	Xp11-q21	481	ENSG00000085224
			ENSG00000086758
			ENSG00000147162
			ENSG00000124486
			ENSG00000179222
			ENSG00000102316
			ENSG00000147202
			ENSG00000180182
			ENSG00000131263
310440	Xq28	151	ENSG00000130821
			ENSG00000184343
			ENSG00000124334
			ENSG00000013563
			ENSG00000013563
311510	Xq28	151	ENSG00000185825
			ENSG00000184216
			ENSG00000180879
314580	Xq13-q21	182	ENSG00000147166
			ENSG00000067177
600131	8q24	247	ENSG00000167632
			ENSG00000180155
600175	12q23-q24	445	ENSG00000152137
			ENSG00000196091
			ENSG00000111245
600593	4p16	160	ENSG00000163132
600624	18q21.1-q21.3	148	ENSG00000197632
			ENSG00000057149
			ENSG00000206075
600792	14q12	43	ENSG00000100473
			ENSG00000176165

continued on next page

continued from previous page

OMIM ID	locus	locus size	Ensembl ID
600964	10pter-p11.2	312	ENSG00000107537
600977	17p13-p12	319	ENSG00000109047 ENSG00000091622
601202	17p13	282	ENSG00000109047
601251	17p	494	ENSG00000109047
601362	10p14-p13	71	ENSG00000107485
601676	1p31	156	ENSG00000132855 ENSG00000116761 ENSG00000116791
601764	19q	1002	ENSG00000105409 ENSG00000142290 ENSG00000105711 ENSG00000105290 ENSG00000063180 ENSG00000167614 ENSG00000131409 ENSG00000105695 ENSG00000167619 ENSG00000198597 ENSG00000083842 ENSG00000169169 ENSG00000105767 ENSG00000104863 ENSG00000105223 ENSG00000160460 ENSG00000105737
601846	19p13.3	238	ENSG00000077009 ENSG00000171119 ENSG00000176533 ENSG00000196415 ENSG00000099875 ENSG00000125733
602067	6q23	74	ENSG00000135541 ENSG00000112319 ENSG00000118526 ENSG00000051620
603165	1q21	334	ENSG00000143412 ENSG00000143369 ENSG00000143631 ENSG00000203782 ENSG00000189334 ENSG00000169469

continued on next page

continued from previous page

OMIM ID	locus	locus size	Ensembl ID
603204	15q24	97	ENSG00000169783 ENSG00000198794
603511	7q	1069	ENSG00000146926 ENSG00000146809 ENSG00000128591 ENSG00000154415 ENSG00000004799 ENSG00000106436 ENSG00000135218 ENSG00000131558 ENSG00000146904 ENSG00000128573
603786	4p	361	ENSG00000163697
604288	2q14-q22	278	ENSG00000121989 ENSG00000150540
604364	22q11-q12	474	ENSG00000166862 ENSG00000100095 ENSG00000182902 ENSG00000128266 ENSG00000183597
604454	2p13	111	ENSG00000159399 ENSG00000169604
604499	11p	637	ENSG00000180210 ENSG00000110169 ENSG00000148965
604781	19p13.2-p13.1	434	ENSG00000105131 ENSG00000105141 ENSG00000171954
604801	1q42	194	ENSG00000143632 ENSG00000163050 ENSG00000116962 ENSG00000119280 ENSG00000135776
605021	16p13	366	ENSG00000078328 ENSG00000167971 ENSG00000138834
605285	10q23.2	26	ENSG00000173267
605480	4p16-p15.2	239	ENSG00000072832 ENSG00000185774 ENSG00000153012 ENSG00000168824 ENSG00000074211

continued on next page

continued from previous page

OMIM ID	locus	locus size	Ensembl ID
605582	6q12-q16	209	ENSG00000065833 ENSG00000146242
605642	1q21	334	ENSG00000143436
605711	2p14-p13	142	ENSG00000124370
605751	16p12-q12	407	ENSG00000087258 ENSG00000087250 ENSG00000174938 ENSG00000129636 ENSG00000103404 ENSG00000103540 ENSG00000166501
605809	17p13	282	ENSG00000188265 ENSG00000170175 ENSG00000108515
606070	5q	1116	ENSG00000120729 ENSG00000113296 ENSG00000113758 ENSG00000113083 ENSG00000164294 ENSG00000184347 ENSG00000113140 ENSG00000169271 ENSG00000081189 ENSG00000157510 ENSG00000164176 ENSG00000113578 ENSG00000113327 ENSG00000085365 ENSG00000044115 ENSG00000120738 ENSG00000134352 ENSG00000175745 ENSG00000081853 ENSG00000168938 ENSG00000038427 ENSG00000094755 ENSG00000133710 ENSG00000171992 ENSG00000184349 ENSG00000113721 ENSG00000145730 ENSG00000129625

continued on next page

continued from previous page

OMIM ID	locus	locus size	Ensembl ID
			ENSG00000172901
			ENSG00000131711
			ENSG00000072682
			ENSG00000113657
			ENSG00000113580
			ENSG00000164347
			ENSG00000048162
			ENSG00000131730
			ENSG00000167768
			ENSG00000172867
606257	12p11.2-q14	548	ENSG00000139648
			ENSG00000170421
			ENSG00000111405
606483	10q24.1-q25.1	179	ENSG00000120057
			ENSG00000179477
			ENSG00000179148
			ENSG00000167741
606545	17p13.2-p13.1	210	ENSG00000108515
			ENSG00000129194
			ENSG00000125414
			ENSG00000181885
			ENSG00000128710
606708	2q31	126	ENSG00000175879
			ENSG00000128713
			ENSG00000128652
606744	18p11.31-q11.2	169	ENSG00000101558
607086	11q23.3-q24	250	ENSG00000149591
			ENSG00000204633
			ENSG00000068976
			ENSG00000173442
			ENSG00000133315
			ENSG00000175591
			ENSG00000110092
			ENSG00000149257
607088	11q13	354	ENSG00000149782
			ENSG00000186642
			ENSG00000165457
			ENSG00000175564
			ENSG00000085733
			ENSG00000174807
			ENSG00000062282
607221	4p15	86	ENSG00000185774

continued on next page

continued from previous page

OMIM ID	locus	locus size	Ensembl ID
			ENSG00000167768
			ENSG00000172867
			ENSG00000170421
			ENSG00000161849
			ENSG00000170426
			ENSG00000139648
			ENSG00000111405
607936	12q13	355	ENSG00000111424
			ENSG00000170477
			ENSG00000185069
			ENSG00000172819
			ENSG00000135423
			ENSG00000139209
			ENSG00000123364
			ENSG00000161850
608096	12q22-q23.3	163	ENSG00000185046
608224	12q24.32-qter	71	ENSG00000177084
			ENSG00000177169
608318	14q32	414	ENSG00000140093
			ENSG00000128578
608423	7q32.1-q32.2	71	ENSG00000128591
			ENSG00000174697
608762	9q32-q33	161	ENSG00000136935
608816	6p21	395	ENSG00000124507

Table 4: Candidate genes for OMIM loci with unknown molecular basis.

3 The distribution of gene numbers in all the disease loci

We count the numbers of disease genes in all the known loci in OMIM database and give its distributions in Fig.S1.

References

[1] S. Fortunato, Community detection in graphs. Physics Reports 486(2010), 75-174.

[2] J. Q. Jiang, A. W. M. Dress, G. Yang, A spectral clusteringbased framework for detecting community structures in complex networks. Appl. Math. Lett. 22(2009) 1479-1482.

[3] M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks. Phys. Rev. E 69(2004), 026113.

[4] M.E.J. Newman, Fast algorithm for detecting community structure in networks, Phys. Rev. E 69 (2) (2004) 066133.

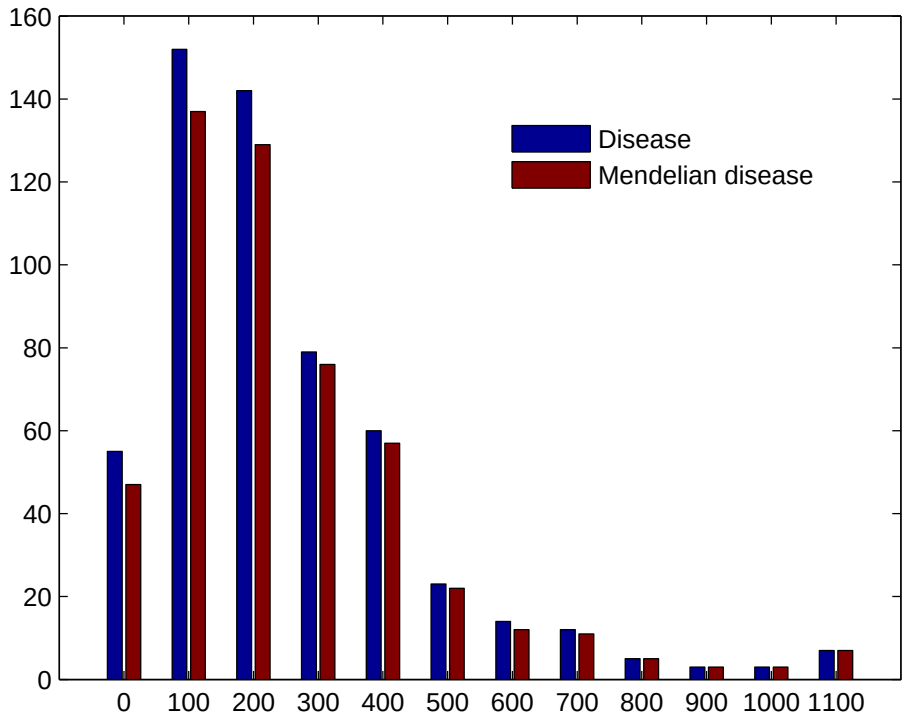


Figure 1: The distribution of gene numbers in all the disease loci.