BIOINFORMATICS ORIGINAL PAPER Vol. 27 no. 19 2011, pages 2692–2699
doi:10.1093/bioinformatics/btr463

Systems biology

Advance Access publication August 8, 2011

Inferring disease and gene set associations with rank coherence in networks

TaeHyun Hwang[1], Wei Zhang[1], Maoqiang Xie[2], Jinfeng Liu[3] and Rui Kuang[1,*]

[1]Department of Computer Science and Engineering, University of Minnesota Twin Cities, Minneapolis, MN 55455, USA, [2]College of Software, Nankai University, Tianjin, China and [3]Department of Bioinformatics and Computational Biology, Genentech Inc., South San Francisco, CA 94080, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: To validate the candidate disease genes identified from high-throughput genomic studies, a necessary step is to elucidate the associations between the set of candidate genes and disease phenotypes. The conventional gene set enrichment analysis often fails to reveal associations between disease phenotypes and the gene sets with a short list of poorly annotated genes, because the existing annotations of disease-causative genes are incomplete. This article introduces a network-based computational approach called rcNet to discover the associations between gene sets and disease phenotypes. A learning framework is proposed to maximize the coherence between the predicted phenotype–gene set relations and the known disease phenotype-gene associations. An efficient algorithm coupling ridge regression with label propagation and two variants are designed to find the optimal solution to the objective functions of the learning framework.

Results: We evaluated the rcNet algorithms with leave-one-out cross-validation on Online Mendelian Inheritance in Man (OMIM) data and an independent test set of recently discovered disease–gene associations. In the experiments, the rcNet algorithms achieved best overall rankings compared with the baselines. To further validate the reproducibility of the performance, we applied the algorithms to identify the target diseases of novel candidate disease genes obtained from recent studies of Genome-Wide Association Study (GWAS), DNA copy number variation analysis and gene expression profiling. The algorithms ranked the target disease of the candidate genes at the top of the rank list in many cases across all the three case studies.

Availability: http://compbio.cs.umn.edu/dgsa_rcNet

Contact: kuang@cs.umn.edu

Received on April 11, 2011; revised on June 28, 2011; accepted on August 2, 2011

1 INTRODUCTION

Determination of the molecular cause of diseases is a major focus in genomics research since early 1960s (McKusick, 2007). Recently, powered by the advanced high-throughput genomic technologies, numerous large-scale genome-wide disease studies such as genome-wide association studies (Johnson and O'Donnell, 2009; The Wellcome Trust Case Control Consortium, 2007), DNA copy number detections (Shlien and Malkin, 2009) and gene expression profiling (van't Veer and Bernards, 2008) were conducted toward this goal. Typically, the objective of a study is to perform a high-throughput scanning for a list of genes that are involved with the disease under study, and then a standard follow-up enrichment analysis or its variants and extensions is applied to analyze the gene set, based on the statistical significance of the overlap between the genes and gene functional annotations or associations with disease phenotypes. Examples of the well-known tools are DAVID (Huang *et al.*, 2009), GSEA (Subramanian *et al.*, 2005), GOToolBox (Martin *et al.*, 2004) and many others. However, in many cases, since the existing annotations of disease-causative genes is far from complete (McKusick, 2007), and a gene set might only contain a short list of poorly annotated genes, enrichment-based approaches often fail to reveal the associations between gene sets and disease phenotypes.

The availability of large phenotypic and molecular networks provides a new opportunity to study the association between diseases and the gene sets identified from the high-throughput genomic studies. The human disease phenotype network (van Driel *et al.*, 2006) provides information on phenotype similarities computed by text mining of the full text and clinical synopsis of the disease phenotypes in OMIM (McKusick, 2007). Large molecular networks such as the human protein–protein interaction network (Chuang *et al.*, 2007) or functional linkage network (Linghu *et al.*, 2009) provide functional relations among genes or proteins. Based on the observation that genes associated with the same or related diseases tend to interact with each other in the gene network, many network-based approaches are proposed to utilize the disease modules and gene modules in the networks to prioritize disease genes, a task of ranking genes for studying genetic diseases (Franke *et al.*, 2006; Hwang and Kuang, 2010; Köhler *et al.*, 2008; Linghu *et al.*, 2009; Li and Patra, 2010; Vanunu *et al.*, 2010; Wu *et al.*, 2008).

In this article, we propose a general network-based approach to infer associations between disease phenotypes and gene sets, utilizing the disease phenotype network and the gene network. We formulate the problem as a gene set query problem. By querying the networks with a given gene set, a user expects to retrieve a list of disease phenotypes with the highest predicted association with the gene set. The principle is that, if genes are ranked by their relevance to the query gene set, and disease phenotypes are ranked by their relevance to the hidden target disease phenotypes of the query gene set, the known associations between the most relevant genes and phenotypes tend to be over-represented compared with random cases. We formulate a simple learning framework maximizing Rank

*To whom correspondence should be addressed.

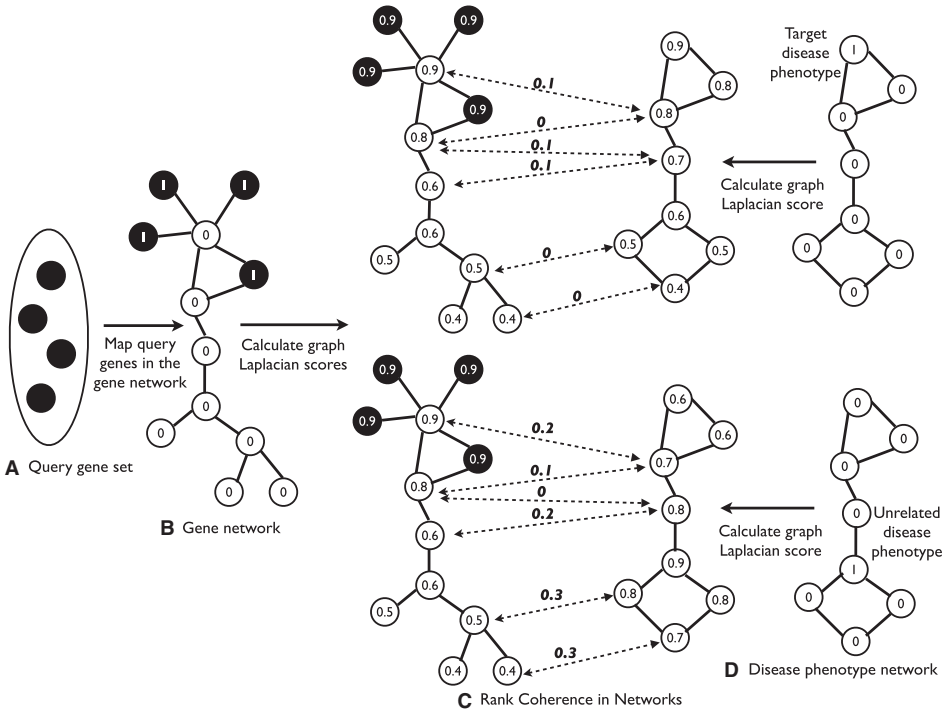

Fig. 1. Illustration of Rank Coherence in Networks. A query gene set of four genes is given in (**A**). The four genes are mapped in the gene network and the corresponding nodes are marked with 1 in (**B**). The graph Laplacian scores are then computed to quantify the relevance between each gene (including the query genes) and the query gene set. In (**D**), if a disease phenotype is selected as the candidate target phenotype and marked with 1, the graph Laplacian scores can also be derived to quantify the relevance between each disease phenotype and the selected phenotype. We contrast the case with the target phenotype selected (the upper case) to the case with an unrelated phenotype selected (the bottom case) in (**D**) to demonstrate the difference in the relevance scores. As showed in (**C**), the known OMIM associations are represented as the dashed edges connecting the associated phenotype and gene. Based on the coherence assumption, similarly ranked genes and phenotypes in the two networks should be highly connected with each other in the upper case, and the connectivity will be close to random in the bottom case. The edges connecting associated genes and phenotypes are labeled by the discrepancy between their ranking scores. Clearly, the phenotype ranking given by the relevance to the target phenotype is more coherent than the ranking given by the relevance to an unrelated phenotype.

Coherence in Networks (rcNet) with respect to the known disease phenotype–gene associations in OMIM.

2 METHODS

2.1 Overview

Figure 1 illustrates the general idea of Rank Coherence in Networks. We first measure the relevance between the query gene set and all the genes with graph Laplacian scores (Fig. 1A and B). The Laplacian scores can be considered as the result of using the query gene set as the seed to perform random walk with restart (or label propagation) in the gene network (Bengio *et al.*, 2006). Thus, the relevance is global because the full network is used to derive the Laplacian scores. The global relevance between a target disease phenotype and all disease phenotypes can be similarly computed as the Laplacian scores with random walk on the disease phenotype network (Fig. 1D). Our assumption is that, between the rankings given by the query gene set and the target disease phenotype, the top-ranked genes and the top-ranked phenotypes should be highly connected by known associations, quantified by Rank Coherence in Networks (Fig. 1C). In a real problem, the target disease phenotypes are unknown. The rcNet algorithms are designed to search for the phenotype(s) with the best rcNet score against the query gene set. We propose two strategies. The first approach relaxes the combinatorial problem as ridge regression to find a closed-form solution for selecting the target disease phenotypes. The second approach in two variants enumerates all phenotype configurations to find the best match of the query gene set.

2.2 Problem definition

We formulate a graph query problem for disease phenotype and gene set association discovery. Given a heterogenous network consisting of the gene network, phenotype network and association network, we query the network with a gene set to retrieve a phenotype (or several) predicted to have association with the query gene set. We define $\mathbf{G_{(n\times n)}}$, $\mathbf{P_{(m\times m)}}$, and $\mathbf{A_{(n\times m)}}$ as the adjacency matrix of the gene network, the disease network, and the disease–gene association network, respectively, where $\mathbf{n}$ is the number of genes and $\mathbf{m}$ is the number of disease phenotypes in the networks. The query gene set is represented by a binary vector $\mathbf{g}=[\mathbf{g_1},\mathbf{g_2},...,\mathbf{g_n}]^{\mathbf{T}}$ denoting the gene membership against the gene set, i.e. each $\mathbf{g_i}=\mathbf{1}$ if gene $\mathbf{i}$ is in the query gene set, otherwise $\mathbf{0}$. Similarly, the list of target phenotype(s) is given by another binary vector $\mathbf{p}=[\mathbf{p_1},\mathbf{p_2},...,\mathbf{p_m}]^{\mathbf{T}}$ and phenotype $\mathbf{j}$ is a target phenotype if $\mathbf{p_j}=\mathbf{1}$. Our objective is to find the $\mathbf{p}$ that gives the best rank coherence with the query gene set $\mathbf{g}$.

2.3 Computing graph Laplacian scores

To fully utilize network topological information, we compute the global relevance score between the query gene set $\mathbf{g}$ and all the genes based on the graph Laplacian of the gene network $\mathbf{G_{(n\times n)}}$. The laplacian scores exploit

modular information in a network to capture long range interactions between the nodes in a graph. We first normalize $\mathbf{G}$ as $\bar{\mathbf{G}}=\mathbf{D}_{\mathbf{G}}^{-\frac{1}{2}}\mathbf{G}\mathbf{D}_{\mathbf{G}}^{-\frac{1}{2}}$, where $\mathbf{D}_{\mathbf{G}}$ is a diagonal matrix with diagonal elements $\mathbf{D}_{\mathbf{G}\mathbf{i},\mathbf{i}}=\sum_{\mathbf{j}}\mathbf{G}_{\mathbf{i},\mathbf{j}}$. A vector $\tilde{\mathbf{g}}$ of graph Laplacian scores is derived from the following optimization problem (Zhou *et al.*, 2004),

$$\mathbf{min}_{\tilde{\mathbf{g}}}\sum_{\mathbf{i},\mathbf{j}}\bar{\mathbf{G}}_{\mathbf{i},\mathbf{j}}(\tilde{\mathbf{g}}_{\mathbf{i}}-\tilde{\mathbf{g}}_{\mathbf{j}})^{\mathbf{2}}+\frac{\mathbf{1}-\alpha}{\alpha}\sum_{\mathbf{i}}(\tilde{\mathbf{g}}_{\mathbf{i}}-\mathbf{g}_{\mathbf{i}})^{\mathbf{2}}. \quad (1)$$

In Equation (1), the first term is a smoothness penalty, which forces connected genes to receive similar scores, and the second term ensures the consistency with the query gene set. The Laplacian scores combine the neighboring information in the network with the consistency with the query gene set to provide a global relevance measure between each gene and the query gene set. Parameter $\alpha\in(\mathbf{0},\mathbf{1})$ balances the contributions from the two penalties. The closed-form solution of Equation (1) is

$$\tilde{\mathbf{g}}=(\mathbf{1}-\alpha)(\mathbf{I}-\alpha\bar{\mathbf{G}})^{-\mathbf{1}}\mathbf{g}. \quad (2)$$

Empirically, to avoid computing the inverse of $(\mathbf{I}-\alpha\bar{\mathbf{G}})$, an iterative algorithm can efficiently compute the closed-form solution with the following update rule at each time step t,

$$\tilde{\mathbf{g}}^{\mathbf{t}}=(\mathbf{1}-\alpha)\mathbf{g}+\alpha\bar{\mathbf{G}}\tilde{\mathbf{g}}^{\mathbf{t}-\mathbf{1}}. \quad (3)$$

Similarly, graph Laplacian scores can be derived to measure the relevance between the phenotypes and the target phenotypes $\mathbf{p}$ with optimization of

$$\mathbf{min}_{\tilde{\mathbf{p}}}\sum_{\mathbf{i},\mathbf{j}}\bar{\mathbf{P}}_{\mathbf{i},\mathbf{j}}(\tilde{\mathbf{p}}_{\mathbf{i}}-\tilde{\mathbf{p}}_{\mathbf{j}})^{\mathbf{2}}+\frac{\mathbf{1}-\beta}{\beta}\sum_{\mathbf{i}}(\tilde{\mathbf{p}}_{\mathbf{i}}-\mathbf{p}_{\mathbf{i}})^{\mathbf{2}}, \quad (4)$$

with the closed-form solution

$$\tilde{\mathbf{p}}=(\mathbf{1}-\beta)(\mathbf{I}-\beta\bar{\mathbf{P}})^{-\mathbf{1}}\mathbf{p}, \quad (5)$$

where $\bar{\mathbf{P}}$ is the normalized $\mathbf{P}$ and $\beta\in(\mathbf{0},\mathbf{1})$ is the balancing parameter. Computing the Laplacian scores is equivalent to a weighted summation of performing random walk on the graph from one step to infinite step. Note that G and P could also be normalized as a stochastic matrix, which is similar to the normalized graph Laplacian $\bar{G}$ and $\bar{P}$. Thus, G and P are allowed to be directed graphs. One can use other scoring functions such as counting the direct neighbors of the query gene set, or measuring the shortest path from the query gene set to other genes as suggested in (Wu *et al.*, 2008). However, empirically, the direct-neighbor function tends to generate very sparse information, and the shortest path function does not fully explore the neighborhood information.

2.4 Rank coherence in networks

The rcNet measures whether the query gene set $\mathbf{g}$ and a phenotype set $\mathbf{p}$ show coherent associations with the known disease–gene associations. Specifically, given the graph Laplacian scores $\tilde{\mathbf{g}}$, which rank the genes by their relevance to the query gene set $\mathbf{g}$, and the graph Laplacian scores $\tilde{\mathbf{p}}$, which rank the disease phenotypes by their relevance to the hidden target phenotypes $\mathbf{p}$, Rank Coherence in Networks $\mathbf{rcNet}(\tilde{\mathbf{g}},\tilde{\mathbf{p}},\mathbf{A})$ measures whether the associations given by A are connecting genes and phenotypes with similar scores in $\tilde{\mathbf{g}}$ and $\tilde{\mathbf{p}}$. We propose two different approaches to define rcNets. The first approach adopts a ridge regression model coupled with label propagations to compute a closed-form solution of $\mathbf{p}$, relaxed to real numbers. The second approach uses simpler measures and enumerate all possible $\mathbf{p}$ to find the best fitting for $\mathbf{g}$.

2.4.1 A ridge regression model Under the assumption that the Laplacian score of a phenotype can be reconstructed by the linear combination of the Laplacian scores of its gene neighbors in A, we can formulate the following least-square cost function,

$$\Omega=||\mathbf{A}\tilde{\mathbf{p}}-\tilde{\mathbf{g}}||^{\mathbf{2}}. \quad (6)$$

Eventually, we are interested in deriving $\mathbf{p}$. After replacing $\tilde{\mathbf{g}}$ with Equation (2) and $\tilde{\mathbf{p}}$ with Equation (5), we have the following regularization framework,

$$\Omega(\mathbf{p})=||(\mathbf{1}-\beta)\mathbf{A}(\mathbf{I}-\beta\bar{\mathbf{P}})^{-\mathbf{1}}\mathbf{p}-(\mathbf{1}-\alpha)(\mathbf{I}-\alpha\bar{\mathbf{G}})^{-\mathbf{1}}\mathbf{g}||^{\mathbf{2}}+\kappa||\mathbf{p}||^{\mathbf{2}}, \quad (7)$$

where $||\mathbf{p}||^{\mathbf{2}}$ is a 2-norm regularizer and κ is a small constant. Equation (7) takes the standard form of ridge regression, and thus the closed-form solution $\mathbf{p}^*$ can be derived by

$$\mathbf{p}^*=(\mathbf{1}-\alpha)(\bar{\mathbf{A}}^{\mathbf{T}}\bar{\mathbf{A}}+\kappa\mathbf{I})^{-\mathbf{1}}\bar{\mathbf{A}}^{\mathbf{T}}(\mathbf{I}-\alpha\bar{\mathbf{G}})^{-\mathbf{1}}\mathbf{g}, \quad (8)$$

where $\bar{\mathbf{A}}=(\mathbf{1}-\beta)\mathbf{A}(\mathbf{I}-\beta\bar{\mathbf{P}})^{-\mathbf{1}}$. Note that the solution $\mathbf{p}^*$ is a real vector, which can be seen as an approximation of the binary vector $\mathbf{p}$. A simple post-processing is to select one or a few phenotypes that are assigned with significantly larger scores as the phenotypes associated with the gene set. The full algorithm to solve the ridge regression model is given below.

dgsa_rcNet$(\mathbf{g},\bar{\mathbf{G}},\bar{\mathbf{P}},\mathbf{A},\alpha,\beta,\psi)$

1 $\mathbf{p}=\mathbf{0}$
2 $\tilde{\mathbf{g}}=(\mathbf{1}-\alpha)(\mathbf{I}-\alpha\bar{\mathbf{G}})^{-\mathbf{1}}\mathbf{g}$ (equation (3))
3 $\bar{\mathbf{A}}=(\mathbf{1}-\beta)\mathbf{A}(\mathbf{I}-\beta\bar{\mathbf{P}})^{-\mathbf{1}}$
4 $\mathbf{p}^*=(\bar{\mathbf{A}}^{\mathbf{T}}\bar{\mathbf{A}}+\kappa\mathbf{I})^{-\mathbf{1}}\bar{\mathbf{A}}^{\mathbf{T}}\tilde{\mathbf{g}}$
5 $\mathbf{p}(\mathbf{p}^*>\psi)=\mathbf{1}$ *(target selection with threshold ψ)*
6 **return** $(\mathbf{p})$

The steps at line 2, 3 and 4 require cubic matrix inversion algorithms. Thus, the time complexity of rcNet algorithm is $\mathbf{O}(\mathbf{m}^{\mathbf{3}}+\mathbf{n}^{\mathbf{3}})$.

2.4.2 Enumeration methods The ridge regression model provides an approximation solution, but if we are only interested in retrieving the most relevant disease phenotype. We can simply go through each phenotype and compute a score against the query gene set $\mathbf{g}$ for each case. Finally, the phenotype with the largest score is chosen as the target phenotype. We propose two functions to measure rcNet for this approach,

$$\mathbf{rcNet}_{\mathbf{corr}}(\tilde{\mathbf{g}},\tilde{\mathbf{p}},\mathbf{A})=\mathbf{corr}(\mathbf{A}\tilde{\mathbf{p}},\tilde{\mathbf{g}}), \quad (9)$$

$$\mathbf{rcNet}_{\mathbf{lap}}(\tilde{\mathbf{g}},\tilde{\mathbf{p}},\mathbf{A})=-\sum_{\mathbf{i},\mathbf{j}}\mathbf{A}_{\mathbf{i},\mathbf{j}}(\tilde{\mathbf{p}}_{\mathbf{i}}-\tilde{\mathbf{g}}_{\mathbf{j}})^{\mathbf{2}}. \quad (10)$$

Function $\mathrm{rcNet}_{\mathrm{corr}}$ simply uses the Pearson's correlation coefficient to check the consistency between $\mathbf{A}\tilde{\mathbf{p}}$ and $\tilde{\mathbf{g}}$, similar to the concordance score used by CIPHER (Wu *et al.*, 2008). Function $\mathrm{rcNet}_{\mathrm{lap}}$ checks if the neighboring genes and phenotypes in the association network are assigned similar scores, and the smaller the disagreement, the higher the relevance. The full algorithm to solve the two enumeration models is given below.

dgsa_rcNet_enum.$(\mathbf{g},\bar{\mathbf{G}},\bar{\mathbf{P}},\mathbf{A},\alpha,\beta)$

1 $\tilde{\mathbf{g}}=(\mathbf{1}-\alpha)(\mathbf{I}-\alpha\bar{\mathbf{G}})^{-\mathbf{1}}\mathbf{g}$
2 $\mathbf{p}=\mathbf{0},\mathbf{s}=\mathbf{0}$
3 **for** $\mathbf{i}=\mathbf{1}$ **to** $\mathbf{n}$
4 $\quad\mathbf{p}_{\mathbf{i}}=\mathbf{1}$
5 $\quad\tilde{\mathbf{p}}=(\mathbf{1}-\beta)(\mathbf{I}-\beta\bar{\mathbf{P}})^{-\mathbf{1}}\mathbf{p}$
6 $\quad\mathbf{s}_{\mathbf{i}}=\mathbf{corr}(\mathbf{A}\tilde{\mathbf{p}},\tilde{\mathbf{g}})$ or $-\sum_{\mathbf{i},\mathbf{j}}\mathbf{A}_{\mathbf{i},\mathbf{j}}(\tilde{\mathbf{p}}_{\mathbf{i}}-\tilde{\mathbf{g}}_{\mathbf{j}})^{\mathbf{2}}$
7 $\quad\mathbf{p}_{\mathbf{i}}=\mathbf{0}$
8 $\mathbf{j}=\mathbf{argmax}_{\mathbf{i}}\ \mathbf{s}_{\mathbf{i}}$
9 $\mathbf{p}_{\mathbf{j}}=\mathbf{1}$
10 **return** $(\mathbf{p})$

Inside the for-loop between line 3 and 7, the rcNet score is computed for each configuration of $\mathbf{p}$. The overall time complexity of this algorithm is also $\mathbf{O}(\mathbf{m}^{\mathbf{3}}+\mathbf{n}^{\mathbf{3}})$ if $(\mathbf{1}-\beta)(\mathbf{I}-\beta\bar{\mathbf{P}})^{-\mathbf{1}}$ is precomputed. Note that this is the computational cost by which we only want to retrieve one phenotype. If we want to explore all possible configurations of $\mathbf{p}$, the total cost is

exponential in **m**. This enumeration strategy is similar to CIPHER (Wu *et al.*, 2008). The advantages are the conceptual simplicity and the optimality of the exact solution. The disadvantages are the computational cost incurred by the repeated calculation of the association score for each possible combination of the individual phenotypes, and the inflexibility to extend to more general problem of finding multiple target phenotypes.

3 RESULTS

The rcNet algorithms are compared to other methods with leave-one-out cross-validation on Online Mendelian Inheritance in Man (OMIM) associations and prediction of an independent set of recently discovered disease-gene- associations. The rcNet algorithms are then applied to validate findings in datasets from Genome-Wide Association Study (GWAS), DNA copy number analysis and microarray gene expression profiling.

3.1 Preparing networks

The disease phenotype network is an undirected graph with 5080 vertices representing OMIM disease phenotypes and edges with weights in $[\mathbf{0},\mathbf{1}]$. The edge weights measure the similarity between two phenotypes by their overlap in the text and the clinical synopsis in OMIM records, calculated by text mining (van Driel *et al.*, 2006).

The disease–gene associations are represented by an undirected bipartite graph with edges connecting phenotype nodes with their causative gene nodes. Two versions (May-2007 Version and May-2010 Version) of OMIM associations were used in the experiments. May-2007 Version contains 1393 associations between 1126 disease phenotypes and 916 genes, and May-2010 Version contains 2469 associations connecting 1786 disease phenotypes and 1636 genes. May-2010 version was used in the experiments on the DNA copy number and gene expression datasets. May-2007 version was used in the experiments on the GWAS datasets since many of the GWAS disease genes were already included in May-2010 Version. Specifically, we selected the GWAS disease gene sets that do not overlap with May-2007 Version as the test cases.

Two gene networks were used in the experiments. The first one was derived from the human protein–protein interaction (PPI) network obtained from Human Protein Reference Database (HPRD) (Peri *et al.*, 2003). The PPI network contains 34 364 binary undirected interactions between 8919 genes. This smaller network was used in validations on the OMIM data due to the computation limitation. A larger human functional linkage network (Huttenhower *et al.*, 2009) was used in the experiments on the GWAS, DNA copy number and gene expression datasets. This network contains 24 433 genes and ∼60 million weighted edges. To reduce the computational complexity, we applied a cutoff 0.6 on the edge weights to generate a sparser network with ∼7 million weighted edges.

3.2 Comparison with other methods

The rcNet algorithms were compared with CIPHER (Wu *et al.*, 2008) and Random Walk with Restart (label propagation) methods (Hwang and Kuang, 2010; Köhler *et al.*, 2008; Li and Patra, 2010; Vanunu *et al.*, 2010), since those methods reported the best performance for disease gene prioritization. We adopted CIPHER with direct neighbor (CDN) or shortest path (CSP) for disease phenotype and gene set association analysis by averaging the correlations across the genes in the query gene set. The Random Walk algorithm described in Li and Patra (2010) (RWR) was chosen as the label propagation method for comparison, because it is straightforward to use the model for disease phenotype and gene set association analysis. The two hyperparameters α and β for rcNet were chosen from $\{\mathbf{0.1},\mathbf{0.3},\mathbf{0.5},\mathbf{0.7},\mathbf{0.9}\}$, and a fixed small number $\kappa=\mathbf{10^{-5}}$ was used for ridge regression in all experiments. The three balancing parameters for RWR were also chosen from $\{\mathbf{0.1},\mathbf{0.3},\mathbf{0.5},\mathbf{0.7},\mathbf{0.9}\}$. For all the methods, a leave-one-out cross-validation on the OMIM May-2007 Version was performed for parameter tuning, and then the methods were applied to predict the associations in the independent set of associations in OMIM May-2010 Version.

3.2.1 Evaluation measures In all the experiments, a query gene set was used to rank all the 5080 disease phenotypes. The higher the target phenotype in the ranking, the better the performance. We measured the performance by the area under the curve of receiver operating characteristic (AUC). Since we are most interested in whether the target phenotype is near the top, we report the AUC up to the first 50, 100 and 250 (top 5% phenotypes) false positives. Specifically, we calculated the AUC in the part of ROC curve in the range from 0 to 50, 100 or 250 false positives. Another important evaluation is how well a method selects highly coherent top-ranked genes and top-ranked phenotypes since high coherence implies a good utilization of known associations. Specifically, the top genes and phenotypes ranked by the query gene set and the target disease phenotype contribute the largest penalties in the cost functions. Thus, connections between them cancel out the large scores and result in a smaller penalty. To quantify the connectivity, the top-**r** disease genes and the top-**l** disease phenotypes with known OMIM disease–gene associations are selected to measure *fold enrichment*, which is calculated as $\frac{\mathbf{k}}{(\mathbf{r}*\mathbf{l})*\mathbf{e}}$, where **k** is the number of observed OMIM associations between the **r** genes and the **l** disease phenotypes, and **e** is the probability of observing a random association between a gene node and phenotype node, estimated from the density of the OMIM disease phenotype–gene associations. Higher fold enrichment indicates higher coherence between top-ranked genes and disease phenotypes, i.e. highly connected with OMIM disease phenotype–gene associations.

3.2.2 Leave-one-out cross-validation For each disease phenotype, the genes associated with the phenotype in OMIM were used as the query gene set to retrieve the disease phenotype. Note that the associations between the query gene set and all disease phenotypes including the target disease phenotype were removed in the experiment for leave-one-out cross-validation. In the experiments with RWR, as suggested by Li and Patra (2010), the disease phenotype network was pruned by taking the five nearest neighbors of each node to reduce the computational complexity in leave-one-out cross-validation. The upper panel of Table 1 reports the average AUC_{50}, AUC_{100} and AUC_{250} across all the query cases in the leave-one-out cross-validation. The detailed parameter tuning results are reported in Supplementary Table S1. $rcNet_{corr}$ and $rcNet_{lap}$ achieved the best results with ∼5% and 6% better ranking compared with the best of the others ($P<1.31e-04$ with paired Wilcoxson test). rcNet performed similar with RWR, while CIPHER DN and CIPHER SP achieved lower scores. Figure 2A shows a global comparison of the ranking by plotting the number of query cases with the target disease phenotype ranked above a certain rank. $rcNet_{corr}$ and $rcNet_{lap}$ generally achieved better rankings at any ranking threshold in the experiments. For example, $rcNet_{corr}$

Table 1. Performance comparison in leave-one-out cross-validation and new association prediction with OMIM data.

Methods	*rcNet*	*rcNet*$_{corr}$	*rcNet*$_{lap}$	*RWR*	*CDN*	*CSP*
Leave-one-out cross-validation						
AUC$_{50}$	**0.160**	**0.195**	**0.198**	0.143	0.139	0.154
AUC$_{100}$	**0.207**	**0.254**	**0.257**	0.201	0.197	0.195
AUC$_{250}$	**0.273**	**0.340**	**0.343**	0.282	0.268	0.252
Predicting novel disease phenotype-gene associations in test set						
AUC$_{50}$	**0.117**	**0.143**	**0.141**	0.115	0.077	0.062
AUC$_{100}$	**0.151**	**0.191**	**0.191**	0.154	0.103	0.096
AUC$_{250}$	**0.204**	**0.271**	**0.269**	0.215	0.170	0.158

The tables report the average AUC$_{50}$, AUC$_{100}$ and AUC across all the query cases for each method. The experiments are on leave-one-out cross-validation and a independent test set. Thus, no *p*-value is associated with the AUCs.

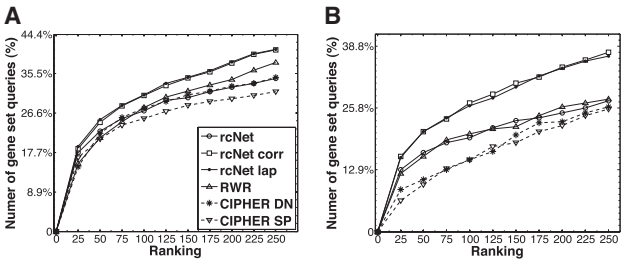

Fig. 2. Ranking comparison. This figure reports the number of query cases, on which a method ranked the target disease phenotype among the top $\mathbf{k} \in [\mathbf{1}, \mathbf{100}]$ phenotypes. (**A**) Leave-one-out; (**B**) Independent test set.

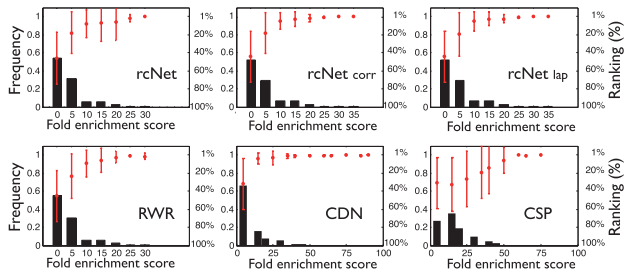

Fig. 3. Analysis of target-phenotype rank versus fold enrichment. The histogram is the distribution of the fold enrichment scores between the top-20 genes ranked by the query gene set and the top-20 disease phenotypes ranked by each disease phenotype. The mean and variance of the target-phenotype rank of the query cases in the same bin are plotted above the histogram.

and rcNet$_{lap}$ ranked ∼290 out of 1126 queries (26%) above rank 50, while RWR and CIPHERs ranked ∼230 out of 1126 queries (20%) above that rank.

3.2.3 Ranking performance and association enrichment Figure 3 plots for each method the ranking performance versus fold enrichment in OMIM associations. Based on the coherence assumption, the top-ranked genes and the top-ranked phenotypes by the query gene set and the target disease phenotype should be highly connected with each other by a good method. In other words, the higher the target-phenotype ranking the higher the fold enrichment. The positive correlation is observed in all the methods. Since CIPHER DN only relies on the direct neighbors in the ranking, the association enrichment is not significant for around 65% of the query cases. CIPHER SP generated more significant fold enrichment scores. However, since CIPHER SP only utilized the shortest path to rank the disease phenotypes, there is weak correlation between its ranking performance and the fold enrichment in the lower range [0,20]. The rcNet algorithms and RWR consistently generates more cases with significant fold enrichment and correlated ranking performance than CIPHER DN/SP. To better quantify the relation between ranking performance and fold enrichment, we also report the number of target disease phenotypes ranked within top 100 with the corresponding fold enrichment score higher than 10 in Supplementary Table S4. The rcNet algorithms identified more such cases. The results support the rank coherence assumption and suggest that label propagation is a better measure than direct neighbors or shortest path to distinguish a target phenotype from unrelated ones, because the information of gene and phenotype neighborhoods are better utilized.

3.2.4 Predicting associations in test set With the best parameters learned in leave-one-out cross-validation, all the methods were applied to predict the target disease phenotype of the new disease genes added into OMIM between May, 2007 and May, 2010. This test set contains 387 disease phenotypes with new associations in OMIM since May, 2007, excluding 11 new disease phenotypes whose disease genes have no interaction in the gene network. In this experiment, the task is to predict the target disease phenotype of the newly annotated disease genes, i.e. to query a set of new disease genes of a disease phenotype to retrieve the phenotype based on the disease–gene associations in May-2007 Version. Table 1 reports the average AUC$_{50}$, AUC$_{100}$ and AUC$_{250}$. Figure 2B reports the number of query cases with the target disease phenotype ranked above a certain rank. The significances of the pair-wise comparisons between the methods are reported in Supplementary Table S3B. rcNet$_{corr}$ and rcNet$_{lap}$ performed the best, followed by rcNet and RWR. CIPHER DN and SP did not produce comparable results with the other methods. The results further support the better performance of the rcNet algorithms compared with the other methods.

3.3 Robustness under bias and noise in networks

To test the robustness of the rcNet algorithms to bias and noise in the networks, we repeated the experiments on perturbed gene network and phenotype network. It is known that well-studied disease proteins tend to have more interactions in the PPI network and this degree bias could potentially lead to superior performance of the network-based methods. An extended PPI network with the same degree of interactions for each protein was generated to assess the influence of the bias as suggested by Wu *et al.* (2008). The extended PPI network was combined from HPRD, OPHID, BIND and MINT database contain 72 431 undirected binary interactions between 14 433 human proteins. The results are reported in Supplementary Table S5. The rcNet algorithms consistently outperformed CIPHER SP. With the replacement by the unbiased PPI network, rcNet performed similarly as in the original experiment, while rcNet$_{corr}$, rcNet$_{lap}$ and CIPHER SP performed worse. We also introduced noise into the phenotype network to assess the robustness of

Table 2. Ranking the target disease phenotype of the disease susceptibility genes identified from GWAS

Category	Disease/trait	PubMed index	OMIM index	Gene set size	Rank by rcNet (pval)	Rank by $rcNet_{corr}$ (pval)	Rank by $rcNet_{lap}$ (pval)
Cancer	Prostate cancer	20 676 098*	176 807	15	2 (0.001)	2 (0.001)	2 (0.001)
	Breast cancer	20 872 241*	113 705	26	7 (0.002)	51 (0.07)	43 (0.012)
	Basal cell carcinoma (cutaneous)	18 849 993	605 462	5	7 (0.005)	189 (0.04)	228 (0.049)
	Basal cell carcinoma (cutaneous)	18 849 993	604 451	5	90 (0.004)	202 (0.026)	256 (0.028)
	Urinary bladder cancer	18 794 855	109 800	1	14 (0.026)	48 (0.011)	60 (0.011)
	Acute lymphoblastic leukemia (childhood)	20 670 164*	159 555	3	19 (0.011)	51 (0.008)	45 (0.004)
	Lung cancer	20 304 703*	211 980	12	22 (0.03)	587 (0.098)	1610 (0.427)
	Lung adenocarcinoma	20 871 597*	211 980	6	52 (0.075)	838 (0.201)	1815 (0.445)
	Chronic lymphocytic leukemia	20 062 064*	151 430	14	57 (0.058)	318 (0.066)	306 (0.031)
	Neuroblastoma (high-risk)	19 412 175	600 613	1	143 (0.001)	110 (0.017)	138 (0.039)
Immunological	Systemic lupus erythematosus	20 169 177*	152 700	10	46 (0.032)	178 (0.034)	161 (0.013)
	Leprosy	20 018 961	246 300	4	78 (0.019)	62 (0.017)	64 (0.013)
	Leprosy	20 018 961	607 572	4	272 (0.003)	54 (0.014)	55 (0.014)
Endocrine	Type 2 diabetes	20 862 305*	125 853	9	97 (0.171)	718 (0.144)	1912 (0.385)
	Type 1 diabetes	19 966 805*	222 100	26	331 (0.249)	690 (0.126)	191 (0.034)
Gastrointestinal	Crohns disease	17 684 544	266 600	2	60 (0.165)	1396 (0.268)	3012 (0.57)

The disease categories in the first column are based on the definition in Goh *et al.* (2007). Multiple GWASs for a disease/trait are marked with '*'.

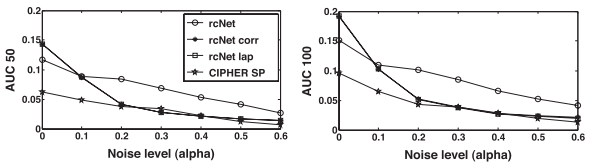

Fig. 4. Performance of ranking target disease phenotypes in noise disease phenotype networks. The figures plot **AUC** scores varying with different noise level in the disease phenotype network.

the methods to potentially inaccurate disease phenotype similarity network as suggested by Wu *et al.* (2008). As the noise level in the phenotype network increases, all the algorithms start to perform worse (Fig. 4). The rcNet algorithms outperformed CIPHER SP at all noise levels. Interestingly, while the performance of $rcNet_{corr}$, $rcNet_{lap}$ and CIPHER SP dropped sharply under the presence of even relatively low noise, there is slower drop in the performance of rcNet. These experiments on the extended PPI network and the noise disease similarity networks suggest that rcNet is robust to bias and noise in the networks compared with the other methods. Note that since the perturbed gene network and phenotype network are too dense to run RWR and CIPHER DN could not handle weighted edges, RWR and CIPHER DN were not included in this analysis.

3.4 Predicting disease phenotypes of disease susceptibility genes from GWAS

The goal of GWAS is to discover disease susceptibility loci/genes that could be useful for assessing or predicting an individual's risk of disease. However, it is often challenging to assess how a set of novel disease susceptibility genes potentially influence susceptibility in disease, especially when the set of genes have no or little previously known disease implications, or function and pathway annotations. In this case study, we collected new disease susceptibility genes from GWAS, whose roles in disease susceptibility are not previously understood, and applied rcNet algorithms to predict the disease phenotype of the disease susceptibility genes. We extracted all the disease susceptibility genes discovered in GWAS based on a recent survey of all studies reported in the GWAS catalog as of December 2010 (Hindorff *et al.*, 2009). After filtering out the genes already included in OMIM May-2007 Version, we selected 217 diseases/traits with novel susceptibility genes that are not associated with any disease phenotype in OMIM May-2007 Version, and 31 of the 217 diseases/traits could be matched with OMIM phenotypes in the disease network. Subsequently, the 31 diseases/traits and their susceptibility genes were used in this experiment.

We queried the set of disease susceptibility genes of each of the 31 diseases/traits to rank the 5080 OMIM disease phenotypes. The ranking results of a subset of the 31 diseases/traits are reported in Table 2. The full results are in Supplementary Table S6. Among the 31 queries, 14 cases ranked the target diseases within top 2% (ranked within top 100). Notable examples are prostate cancer, breast cancer, basal cell carcinoma, bladder cancer, acute lymphoblastic leukemia, systemic lupus erythematous and leprosy. In these cases, the rcNet algorithms ranked the target disease phenotype of the query gene set within top 1% ($P < 0.05$). The *P*-values were calculated by rerunning the algorithms on the network with randomized OMIM associations 1000 times. Figure 5 shows the example that rcNet accurately ranked the breast cancer phenotypes, by querying with 26 novel breast cancer susceptibility genes from GWAS. One interesting observation is that the target disease phenotype OMIM:113705 'BREAST CANCER 1 GENE; BRCA1' is only directly connected with three top-ranked disease phenotypes, OMIM: 114480 'BREAST CANCER', OMIM:151623 'LI-FRAUMENI SYNDROME 1:LFS1', and OMIM:'259500: OSTEOGENIC SARCOMA', and only 5 of the 26 query genes directly interact with the top-ranked disease genes. Direct neighbor expansions in both the gene network and the phenotype network resulted in four OMIM disease–gene associations. This observation suggests that, simply exploring the direct neighbors of the query gene set and the target disease phenotype in the networks, a method might fail to infer disease–gene set associations, due to the low

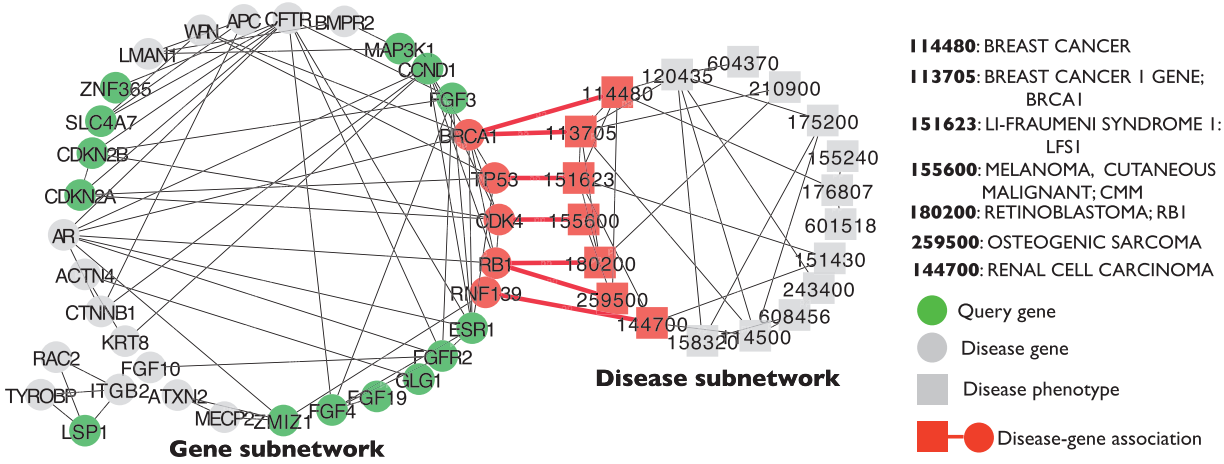

Fig. 5. Querying with breast cancer susceptibility genes from GWAS by rcNet. By querying with the 26 novel breast cancer susceptibility genes from GWAS, rcNet ranked the 20 disease genes in the gene subnetwork at the top. The gene subnetwork also includes 14 out the 26 query genes, which are connected with the top 20 genes. Similarly, the top 20 disease phenotypes ranked by OMIM113705 breast cancer disease phenotype are included in the disease subnetwork. Five of the 20 top-ranked disease genes are connected to 7 of the top 20 disease phenotypes given by seven OMIM disease–gene associations, compared with the expected 0.87 association between 34 random genes and 20 random phenotypes.

statistical significance of the sparse connectivity between the genes and the disease phenotypes. Specifically, in this example, the fold enrichment for four associations is 7.53, which is significantly <12.35 fold enrichment obtained by rcNet. Another interesting example is the inference of the association between leprosy and its susceptibility genes from GWAS (pubmed 20018961). In OMIM May-2007 Version, leprosy has no causative genes, and the leprosy susceptibility genes from GWAS also have no association with any diseases. The lack of known associations in both the target disease phenotype and the gene set poses a hard case that gene set enrichment analysis based on overrepresentation will fail to reveal, but the rcNet algorithms ranked leprosy within top 2%.

Interestingly, previous studies showed that disease susceptibility genes from GWAS catalog have less modularities in the gene network compared with the known disease genes in OMIM. Furthermore, phenotypically similar diseases such as immunological and gastrointestinal diseases do not tend to share their disease genes (Baranzini, 2009; Barrenas *et al.*, 2009). Those previous studies also hypothesized that, due to the unique topological characteristics of the disease susceptibility genes discovered in GWAS, the existing network-based methods would fail to reveal the associations between the disease susceptibility genes and the disease. However, our experiments suggest that, by incorporating the global topological information in the networks and the known OMIM associations, the rcNet algorithms successfully discovered the elusive associations in many cases.

3.5 Predicting disease phenotypes of genes with copy number changes or differential expression

We collected 13 human DNA copy number change datasets from a recent human cancer copy number study from Beroukhim *et al.* (2010). Genes in the detected copy number change regions were used as the query gene set to predict their target disease phenotypes. The target cancer phenotypes are ranked within top 2% for six cancers by rcNet and seven by $rcNet_{corr}$. In nine cases, at least one algorithm ranked the target disease within top 100. Beroukhim *et al.* (2010) concluded that more than three-quarters of the statistically significantly altered copy number regions contain potential cancer-causing genes that are not previously validated targets of cancer somatic copy number alternations. This suggests that enrichment analysis of the genes will not reveal any disease association, but the rcNet algorithms found many associations with the network information. Similarly, we collected 13 human cancer microarray gene expression dataset from GEO. The differentially expressed genes were used to query for their target diseases. rcNet ranked 7 within top 5%, and 12 within top 10%. Although the result is only moderately encouraging, it validates the hypothesis that the neighboring information of the differentially expressed genes provides clue of association with the target disease phenotype. The full results are reported in Supplementary Tables S7 and S8.

4 RELATED WORK

The rcNet algorithms are different from the gene set enrichment analysis with statistical methods such as Hypergeometric statistics, McNemar's test, permutation test or other non-parametric methods (Huang *et al.*, 2009; Martin *et al.*, 2004; Subramanian *et al.*, 2005), because the rcNet algorithms use the topological information in the disease phenotype network and the gene network to analyze the association between a gene set and all phenotypes simultaneously. The simultaneous analysis of all phenotypes provides a global dependence, and thus richer and more reliable information for computing the association scores are used to rank the phenotypes. The rcNet algorithms share more algorithmic similarity with the disease gene prioritization methods, which were proposed for a different purpose. CIPHER (Wu *et al.*, 2008) scores each gene against a disease phenotype based on the correlation between their relevances with all the phenotypes, where the relevance between the gene and a phenotype is calculated based on the distance between the gene and the genes associated with the phenotype. The methods proposed by Köhler *et al.* (2008), Vanunu *et al.* (2010) and Linghu *et al.* (2009) applied random walk (label propagation)

or simpler neighborhood weighting to exploit the gene networks for ranking genes for a disease phenotype, based on the seed genes mapped from the disease phenotype. One limitation is that the phenotype network and the sparse known associations are not fully utilized in the global analysis. The label propagation algorithms proposed by Hwang and Kuang (2010) and Li and Patra (2010) explore a heterogeneous network combining the gene network, the phenotype network and the associations to explore gene modules, phenotype modules and the phenotype–gene association biclusters. Since the two methods make full use of the information in the networks, it is difficult to interpret the results and to tune the best parameters for combining the information.

5 DISCUSSION

Analysis of the gene sets from genome-wide high-throughput screening is a continuing challenge in many disease studies. Statistics from OMIM (January 2011) show that 3745 of the 6675 disease phenotypes are still unknown for their molecular basis. Accordingly, enrichment analysis will fail to find any associations between the 3745 disease phenotypes and any query gene set. rcNet is a helpful tool, with which researchers can validate their findings from high-throughput studies, especially, to validate novel associations between complex diseases and a query gene set with no known associations. rcNet can also help identify closely related phenotypes of the target disease of the query gene set. Since rcNet algorithms utilize both the disease similarities and the gene interactions, some phenotypically similar disease phenotypes will be ranked at the top. These disease phenotypes might provide additional information to investigate the target disease in the study.

Compared with the other methods that also utilize the gene network and the disease network, rcNet is more flexible in handling the network data, because rcNet is capable of handling weighted associations and weighted edges in denser gene network and disease network. rcNet does not rely on deciding direct neighbors or shortest path as CIHPER or PRINCE (Vanunu *et al.*, 2010). rcNet also does not suffer the converge problem as RWR does on a dense large heterogeneous network. The ridge regression model coupled with label propagation provides an approximation of jointly finding association between a gene set and multiple disease phenotypes, which is difficult to achieve with enumeration-based strategies. The network perturbation analysis validated that rcNet is robust under the presence of network bias and noise, which explains that rcNet outperformed the two variances, rcNet_{corr} and rcNet_{lap} in the case studies. A potential limitation of the current study is that it is difficult to distinguish the closely related phenotypes from false positives, because it is possible that some of the top-ranked phenotypes are not similar to or share any common disease genes with the target disease phenotype in the disease network. Interpretation of these phenotypes will not be straightforward. A possible solution is to identify subnetworks and use information from the gene cluster and the phenotype cluster for finding explanations.

ACKNOWLEDGEMENT

We sincerely thank Thanh Le and Xiang Yao for their contribution to the development of rcNet webtool.

Funding: The project was supported by internal funding from University of Minnesota. The grant information is complete in the statement.

Conflict of Interest: none declared.

REFERENCES

Baranzini,S.E. (2009) The genetics of autoimmune diseases: a networked perspective. *Curr. Opin. Immunol.*, **21**, 596–605.

Barrenas,F. *et al.* (2009) Network properties of complex human disease genes identified through genome-wide association studies. *PLoS One*, **4**, e8090.

Bengio,Y. *et al.* (2006) Label propagation and quadratic criterion. In Chapelle,E.O. *et al.* (eds). *Semi-Supervised Learning*. MIT Press, Cambridge, MA.

Beroukhim,R. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.

Chuang,H.-Y. *et al.* (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.

Franke,L. *et al.* (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.*, **78**, 1011–1025.

Goh,K. *et al.* (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

Hindorff,L.A. *et al.* (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362.

Huang,D. *et al.* (2009) Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Huttenhower,C. *et al.* (2009) Exploring the human genome with functional maps. *Genome Res.*, **19**, 1093–1106.

Hwang,T. and Kuang,R. (2010) A heterogeneous label propagation algorithm for disease gene discovery. In*Proceedings of SIAM International Conference on Data Mining*. pp. 583–594.

Johnson,A. and O'Donnell,C. (2009) An open access database of genome-wide association results. *BMC Med. Genet.*, **10**, 6.

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

Linghu,B. *et al.* (2009) Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biol.*, **10**, R91.

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

Martin,D. *et al.* (2004) GOToolbox: functional analysis of gene datasets based on gene ontology. *Genome Biol.*, **5**, R101.

McKusick,V. (2007) Mendelian inheritance in man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.

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.

Shlien,A. and Malkin,D. (2009) Copy number variations and cancer. *Genome Med*., **1**, 62.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.

van't Veer,L. and Bernards,R. (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, **452**, 564–570.

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

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

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

Zhou,D. *et al.* (2004) Learning with local and global consistency. In *Advanced Neural Information Processing Systems*, Vol. 16, MIT Press, Cambridge, MA, pp. 321–328.