Bioinformatics, 35(11), 2019, 1923–1930
doi: 10.1093/bioinformatics/bty882
Advance Access Publication Date: 18 October 2018
Original Paper

Systems biology

Disease gene identification based on generic and disease-specific genome networks

Yonghyun Nam[1], Jong Ho Jhee[1], Junhee Cho[1], Ji-Hyun Lee[2] and Hyunjung Shin[1,*]

[1]Department of Industrial Engineering, Ajou University, Yeongtong-gu, Suwon 16499, South Korea and [2]DR. Noah Biotech, Yeongtong-gu, Suwon 16229, South Korea

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on July 4, 2018; revised on October 11, 2018; editorial decision on October 12, 2018; accepted on October 17, 2018

Abstract

Summary: Immune diseases have a strong genetic component with Mendelian patterns of inheritance. While the tight association has been a major understanding in the underlying pathophysiology for the category of immune diseases, the common features of these diseases remain unclear. Based on the potential commonality among immune genes, we design *Gene Ranker* for key gene identification. Gene Ranker is a network-based gene scoring algorithm that initially constructs a backbone network based on protein interactions. Patient gene expression networks are added into the network. An add-on process screens the networks of weighted gene co-expression network analysis (WGCNA) on the samples of immune patients. Gene Ranker is disease-specific; however, any WGCNA network that passes the screening procedure can be added on. With the constructed network, it employs the semi-supervised learning for gene scoring.

Results: The proposed method was applied to immune diseases. Based on the resulting scores, Gene Ranker identified potential key genes in immune diseases. In scoring validation, an average area under the receiver operating characteristic curve of 0.82 was achieved, which is a significant increase from the reference average of 0.76. Highly ranked genes were verified through retrieval and review of 27 million PubMed literatures. As a typical case, 20 potential key genes in rheumatoid arthritis were identified: 10 were *de facto* genes and the remaining were novel.

Availability and Implementation: Gene Ranker is available at http://www.alphaminers.net/GeneRanker/

Contact: shin@ajou.ac.kr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Identification of key disease-genes, whose products play a central role in monogenic or genetically complex diseases, is a major aim of human genome research (Venter *et al.*, 2001). The key genes may also be bio-markers, target genes (proteins) or drug targets. Thus, key gene identification is important to developing new treatments for patients and discovering new drugs for treating diseases. To identify key genes, *in silico* methods are applied as a preliminary procedure prior to *in vitro* and *in vivo* approaches (Hurle *et al.*, 2013). To evaluate a disease, *in silico* approaches use computational dry runs that simulate and search all possible combinations of genes, diseases and drugs from databases, which have become more widely available. There have been numerous successful trials to identify key genes triggered by *in silico* approaches, such as XPC, BAP1 and *Scotin* in lung cancer; MYB and PCM1 in leukemia (Barski *et al.*, 2013); ATM, PPP1R13L, FOSL2, ERBB4 and HIP1 in breast cancer; and TP53, JUN, PIK3R1, HTT and HNF1A in diabetes mellitus (Ganegoda *et al.*, 2015).

For key gene identification, various machine learning algorithms have been widely used. Gardiner *et al.* (2012) used structural

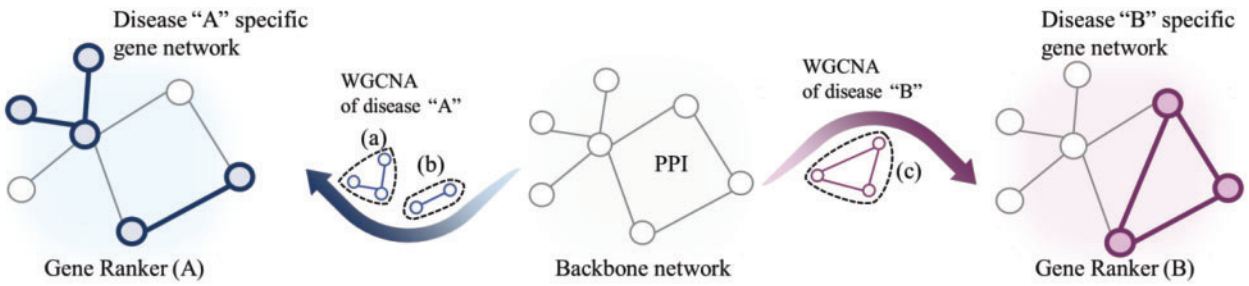

Fig. 1. Schematic description for main idea of the proposed method: The center panel is the backbone (base) network constructed by PPI network; left and right panels are specialized for diseases 'A' and 'B,' respectively. Gene Ranker (A) is incorporated with a specific WGCNA based on the backbone network

equation modeling (SEM) to validate kidney injury molecule-1 as a biomarker of chronic kidney diseases. SEM is a simple method for imputing relationships between unobserved constructs from observable variables, providing interpretation on the relations between genes. Zhao and Li (2010) proposed 'drugCIPHER,' a framework based on pharmacological and genetic correlations throughout the genome. The system predicts key gene–drug interactions. Based on the protein interaction, a linear regression model was constructed using a combination of target validity, drug treatment similarity and chemical similarity. Bakheet and Doig (2009) proposed a method for identifying new potential target genes for drug discovery with support vector machines. Protein sequence properties and amino acid composition of protein, hydrophobicity and PEST were used. Zhao *et al.* (2011) provided a gene scoring model. The score was calculated by Katz-centrality, which is similar to the well-known PageRank algorithm. The score was based on gene expression values in addition to proximity in protein–protein interactions (PPIs).

Most approaches including those described above identify new target genes by obtaining knowledge from disease–gene (or protein) relations and gene–gene relations (Campillos *et al.*, 2008; Chiang and Butte, 2009). For immune and inflammatory diseases (disorder), it is more promising to use information of the underlying genetic relation when detecting key genes. Immune diseases are generally complex genetic diseases in which genes and the environment interact in unknown manners. Particularly, immune diseases have a strong genetic component with Mendelian patterns of inheritance (Gregersen and Olsson, 2009) and tend to appear in families (Ermann and Fathman, 2001). Therefore, recent epidemiologic and pathogenesis studies have suggested numerous commonalities between the pathogenesis of prototypic autoimmune diseases such as rheumatoid arthritis (Abou-Raya and Abou-Raya, 2006). To model an immune disease, a gene network-based model seems to be the most suitable approach. Immune diseases are mostly caused by genetic propensity and are likely to share genes. In network representation, the genes tend to reside in the nearby neighborhood in the network, making it easy to depict the shared genes associated with the same or similar diseases.

In this study, we propose a network-based key gene scoring method for immune diseases. The method is denoted as Gene Ranker throughout this paper. It accommodates both numerous *general* genes and *immune disease-specific* genes. Gene Ranker employs a network structure to better feature the relations between genes. The PPI network is used as a backbone network. Then, a weighted gene co-expression network analysis (WGCNA) is incorporated as an add-on network to demonstrate the peculiarities of respective immune diseases. However, any WGCNA network in the category of immune diseases can be added on, regardless of the given disease, if it passes a selection procedure. This is designated because genetic sharing is well-known to underlie the pathophysiology of immune or inflammatory diseases (Cotsapas and Hafler, 2013). In addition, by using PPI as the backbone network, we can expect some more (plausible) genes among those that interact with the known disease genes (a WGCNA network cannot be regarded as containing all genes related to an immune disease). On the other hand, it is also helpful to overcome sparsity of individual WGCNA networks (Grigoriev, 2001) (usually, most machine learning algorithms produce unstable results for sparse networks).

With the constructed network, Gene Ranker employs a semi-supervised learning for gene scoring. Because few genes have been identified to play a key role in immune diseases, most machine learning algorithms are not applicable because of a lack of labeled data. In contrast, semi-supervised learning can conduct learning with few labeled data. Based on the resulting scores, Gene Ranker recommends the genes, each of which is a potential key gene of an immune disease. Figure 1 describes the concept of this study. There is a backbone network in the center, as well as three different disease-specific networks. (a) and (b) are obtained from WGCNA of disease A, while (c) is from disease B. The network in the left panel is specialized for disease A by overlaying the add-on networks (a) and (b) onto the backbone network. Gene Ranker is specified for disease A and denoted as Gene Ranker (A). In the same manner, Gene Ranker (B) is constructed. If network (a) is helpful for predicting key genes associated to disease B, Gene Ranker (B) can incorporate (a) as its member network. More detailed descriptions regarding network construction, integration with selected networks and scoring are explained in the following sections. In the experiment, the highly ranked genes are validated by retrieval and review of existing studies by text mining of nearly 27 million PubMed literatures. Additionally, as a typical case, 20 potential key genes for rheumatoid arthritis are provided; half of the genes are publicly known to be related to the disease although there has been no information on which genes are more important than others. However, in the proposed method, the priorities of those genes are measured by means of scoring. The remaining 10 genes were novel genes identified in this study.

2 Materials and methods

The proposed method consists of three steps as shown in Figure 2: (i) network construction, (ii) network selection and integration and (iii) key gene scoring. Gene Ranker constructs a disease-specified gene network by adding patients' gene-expression data to the PPI network. Next, gene scores are calculated by a graph-based semi-supervised learning. The resulting gene scores and ranks can be used to identify the priorities of known de facto disease genes and recommend potential candidate genes.

2.1 Gene network construction

A gene network is a graph that represents genes (or proteins) as nodes and connections between pairs of genes as edges. Depending

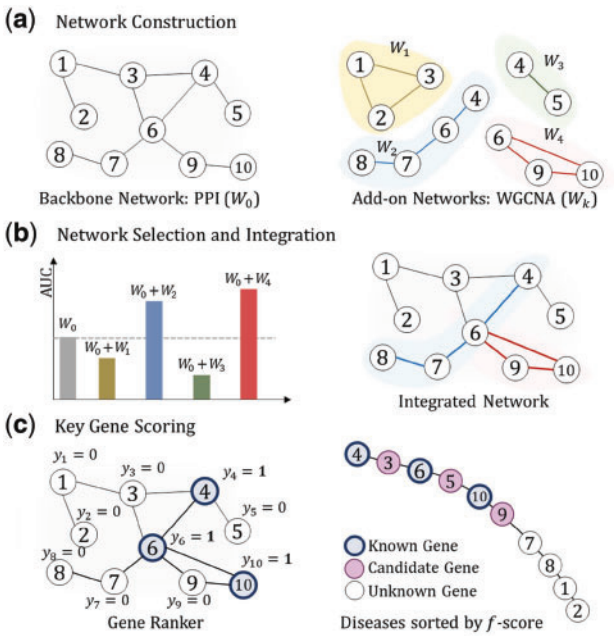

Fig. 2. Overall procedure of the proposed method: (**a**) Backbone network is constructed by PPI. It shows representative gene–gene relation. Add-on networks are constructed by the correlation coefficients between the expression values of two genes using WGCNA. The edge of add-on network reflects the deviation of expression values for disease-specific gene. (**b**) For the screening process, Gene Ranker determines which WGCNA network is helpful for gene scoring of the given disease. The criterion is $AUC(W_0 + W_k) > AUC(W_0)$. After screening, integration of the selected networks is performed. (**c**) To obtain gene scores, Gene Ranker employs SSL. In the graph, all known genes (marked in blue) are set as '1' for the labeled set, and the remaining genes (marked in white) are set as '0' for the unlabeled set. The scoring curve shows the predicted *f*-score (gene scores). The inserted genes between known genes such as Genes 3, 5 and 9 are regarded as strong candidate genes (marked in pink) (Color version of this figure is available at *Bioinformatics* online.)

on how the value of the edge is obtained, different types of gene networks can be constructed. In the proposed method, a PPI network is used as the backbone (base) network, in which the edges indicate the presence or absence ('1' or '0' respectively) of physical contact among two or more proteins. Meanwhile, as an add-on network, WGCNA is employed. WGCNA is conducted to measure pairwise correlations between gene-expression values. Therefore, an edge of the WGCNA network denotes the correlation coefficients between the expression values of two genes. If PPI stands for representative gene–gene relation, WGCNA reflects the difference of co-expression for disease-specific genes (note that it does not necessarily mean the difference of the gene expression). The difference in WGCNA network specifies disease specific co-expression genes. The PPI network contains general information with gene coverage of all but 16 593 genes of nearly 25 000. In contrast, for the WGCNA network, the gene set varies depending on the inclusion/exclusion of genes used for analysis. If it has a low gene coverage, the network is sparse (see Fig. 2a).

2.2 Network selection and integration

Network selection: To construct a disease-specific gene network, it is required to determine which WGCNA network is helpful for gene scoring of a given disease. The criteria for network selection are simple. If the performance of an add-on network improves that of the reference performance, the add-on network is added to the PPI network. In this study, performance is measured as the area under the receiver operating characteristic curve (AUC) and the reference value is obtained from the PPI network. To technically rephrase this, the criterion is $AUC(W_0 + W_k) > AUC(W_0)$, where W_0 denotes the PPI network and W_k is the kth WGCNA network to be added. This indicates that the gene set of the add-on WGCNA and its gene–gene correlation values contain gene–gene information specific to the given disease. As shown in Figure 2b, the second (W_2) and fourth (W_4) networks meet the selection criteria.

Network integration: After screening the redundant or unnecessary networks from among all WGCNA networks, integration for the selected networks is performed. Unit weights are used as coefficients for the linear combination of networks $\sum_{k=0}^{K} \boldsymbol{\beta}_k \boldsymbol{W}_k$. Note that in the pseudocode (Fig. 3), $\boldsymbol{W}_k$ is successively added; therefore, $\boldsymbol{\beta} = 1/(\mathrm{K}+1)\mathbf{1}^{\mathrm{T}}$, where $\mathbf{1}^T$ is a column vector of rank $(K+1)$. More sophisticated approaches could be used for differently weighted coefficients (e.g. Shin *et al.*, 2007, 2009; Tsuda *et al.*, 2005); however, it would incur huge complication relative to the gain in accuracy, as the WGCNA networks in this study are large and the respective networks show large variability in network density (most network integration methods expect the networks to be integrated to have similar densities, so that the combining coefficients only concern the importance of the information in the individual network).

2.3 Gene ranker: key gene scoring

With the integrated network, disease-specific gene scoring is performed. The resulting scores are used to determine the priorities of genes as key genes. For known disease-specific genes, the ranks of involvement reveal which gene should be tested as a target gene prior to testing other genes. In contrast, for unknown genes, the ranks are reduced from the large set of genes to a smaller list of potential candidate target genes.

Scoring employs a graph-based semi-supervised learning (SSL), as disease-genes are sparsely known. Using few labeled data, most machine learning algorithms cannot perform well, whereas SSL can deal with such difficulty and perform prediction by propagating the label information to unlabeled nodes along with edges (Chapelle *et al.*, 2009; Zhu and Goldberg, 2009). If the graph is disconnected, propagation cannot continue. This motivates us to adopt the PPI network as the backbone and incorporate add-on networks.

The following is the scoring procedure and formulation. Let $y = (y_1, \ldots, y_{n=l+u})^{\mathrm{T}}$ denote the label set of genes, $f = (f_1, \ldots, f_n)^{\mathrm{T}}$ denote the set of resulting gene scores, and $\boldsymbol{W}$ as the similarity matrix of the integrated network. If we have l known genes and u unknown genes, we set $y = (y_1 = 1, \ldots, y_l = 1, y_{l+1} = 0, \ldots y_u = 0)^{\mathrm{T}}$. Then SSL uses the graph Laplacian matrix $\boldsymbol{L}$, and obtains score $\boldsymbol{f}$ by minimizing the following objective function:

$$\min \ (\boldsymbol{f} - \boldsymbol{y})^{\mathrm{T}}(\boldsymbol{f} - \boldsymbol{y}) + \mu \boldsymbol{f}^{\mathrm{T}} \boldsymbol{L} \boldsymbol{f}. \tag{1}$$

The graph Laplacian $\boldsymbol{L}$ is defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, where $D = diag(d_i)$ and $d_i = \sum_i w_{ij}$. The score $f = (f_1, \ldots, f_l, f_{l+1}, \ldots, f_{n=l+u})^{\mathrm{T}}$ is the output solution of (1). The parameter μ trades off *loss* and *smoothness* (Chapelle *et al.*, 2009; Zhu and Goldberg, 2009).

As described above, the integrated network is represented as a linear combination of K WGCNA networks with coefficient $\beta = 1/(K+1)$. By replacing the single Laplacian $\boldsymbol{L}$ in (1) with the integrated function $\sum_{k=0}^{K} \boldsymbol{\beta}_k \boldsymbol{W}_k$, the closed form solution is obtained as

$$\boldsymbol{f} = \left(\boldsymbol{I} + \mu \sum_{k=0}^{K} \boldsymbol{\beta} \boldsymbol{W}\mathrm{k} \right)^{-1} \boldsymbol{y}, \ \boldsymbol{\beta} = \frac{1}{\mathrm{K}+1} \mathbf{1}^{\mathrm{T}} . \tag{2}$$

1: PPI network $\mathbf{W}_0 = \{w_{ij}\}_{i,j=1}^{n}$
WGCNA network $\mathbf{W}_k,\ k = 1, \cdots, K$
Label vector $\mathbf{y} = \{\boldsymbol{y}_l, \boldsymbol{y}_u\} = \{y_1, \ldots y_l, y_{l+1}, \cdots, y_{n=l+u}\}$
Score vector $\boldsymbol{f} = \{f_1, \ldots, f_n\}$
Smoothness control (user parameter) μ
Number of candidate genes (user parameter) δ

2: **Initialize** $\mathbf{W} \leftarrow \mathbf{W}_0,\ \boldsymbol{y}_l = \mathbf{1},\ \boldsymbol{y}_u = \mathbf{0},\ \boldsymbol{f} = \mathbf{0}$
3: **for** $k = 1{:}\ K$
4: **if** $\mathrm{AUC}(\boldsymbol{W}_0 + \boldsymbol{W}_k) \geq \mathrm{AUC}(\boldsymbol{W}_0)$ **then**
5: $\boldsymbol{\beta} \leftarrow \frac{1}{K+1}\mathbf{1}$
6: $\boldsymbol{W} \leftarrow \boldsymbol{\beta}(\boldsymbol{W} + \boldsymbol{W}_k)$
7: **end if**
8: **end for**
9: $d_{ii} = \sum_j w_{ij}$, $\boldsymbol{D}$ is a diagonal matrix of $d_{ii},\ i = 1,2, \cdots, n.$
10: $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$
11: $\boldsymbol{f} = (\boldsymbol{I} + \mu \boldsymbol{L})^{-1}\boldsymbol{y}$
12: Sort $\boldsymbol{f}$ in descending order
13: Scale $\boldsymbol{f}$
$\boldsymbol{f}' = (\boldsymbol{f} - \min(\boldsymbol{f}))/(\max(\boldsymbol{f}) - \min(\boldsymbol{f}))$

14: **if** $n_l + \delta > \mathrm{g}$ **then** (g is the last index in $\boldsymbol{f}'$ for labeled vector)
15: $\boldsymbol{f}' = \{f_1', \ldots, f_g'\}$
16: **else**
17: $\boldsymbol{f}' = \{f_1', \ldots, f_{n_l+\delta}'\}$
18: **end if**
19: **return** $\boldsymbol{f}'$

Fig. 3. Pseudocode for Gene Ranker

The resulting score $\boldsymbol{f}$ is rearranged as follows:

$$f' = \frac{f - \min(f)}{\max(f) - \min(f)}.$$

To identify the top-tier genes, the total n genes are sorted in descending order. Next, the ranks are endowed to the labeled and unlabeled genes. Every gene can be a candidate target gene, and a user-specified parameter δ limits the maximum number. Figure 2c shows the scoring results: known genes are marked with a blue outline, unknown genes are marked as white in the network and genes are marked in pink if they are chosen as candidate genes. On the score curve, the inserted genes between known genes such as Genes 3, 5 and 9 are regarded as strong candidate genes. For more information, see the pseudocode in Figure 3.

3 Results

3.1 Data

The data were collected for 27 immune diseases. An immune disease (disorder) is defined as a dysfunction of the immune system, which occurs when the body tissues are attacked by its own immune system. These diseases can be characterized based on the components of the immune system affected, whether the immune system is overactive or underactive, and whether the condition is congenital or acquired. Examples of (auto-)immune diseases include rheumatoid arthritis, systemic lupus erythematosus, Sjogren syndrome, asthma, HIV infection, multiple sclerosis, Hashimoto thyroiditis, juvenile (type 1) diabetes, Addison disease and pulmonary fibrosis, among others. We acquired these terms from the immune disease category of MeSH (the abbreviation for medical subject headings of the National Library of Medicine). It has a controlled vocabulary thesaurus for disease names in the form of a taxonomy. When considering up to the second level of the taxonomy, there are 27 descriptors for immune diseases out of the 4663 listed diseases. In this study, we only present six (of 27) immune diseases: rheumatoid arthritis, asthma, HIV infection, Crohn's disease, multiple sclerosis and inflammatory bowel disease.

First, to construct the PPI backbone network, 16 593 genes and 244 767 interaction information were used to constitute the nodes and edges, respectively. The interaction information was obtained from PharmDB (Lee *et al.*, 2015), a tripartite pharmacological network database containing human diseases, drugs and proteins, which compiles and integrates nine existing interaction databases (see the second row in Table 1). The PPI network has a value of '1' or '0,' which indicates the presence or absence of a relation between genes, respectively. Across the 27 immune diseases, 2420 genes were known to be associated with at least one immune disease. This amounts to approximately 15% of the genes included in the backbone network. Second, to build the disease-specific WGCNA networks, 186 gene expression datasets were collected from the GEO database, which includes control and test groups. Among them, 98 datasets were available. For rheumatoid arthritis, 6 WGCNA networks were constructed, 15 for asthma, 4 for HIV infection, 2 for Crohn's disease, 17 for multiple sclerosis and 1 for inflammatory bowel disease. The remaining 53 WGCNA networks were obtained from 21 other immune diseases. Third, to validate the genes detected by scoring, we examined previous studies. From the PubMed database, 27 931 712 literatures were collected and searched for related studies conducted by other researchers. Table 1 lists the details of these data (Supplementary Table A1 in Appendix A shows the number of WGCNA networks for each immune disease).

3.2 Results of network construction for gene ranker

Changes in network density: Using the network selection and integration procedures described in Section 2.2, a final network was constructed for each immune disease. Hereafter, the resulting network is denoted as Gene Ranker (·), where the parentheses specify a disease when required. The 98 WGCNA networks were selectively added to the backbone network. For respective diseases, different WGCNA networks were chosen. For rheumatoid arthritis, 28 networks were selected and incorporated into the backbone network. One of the distinguishable merits of Gene Ranker is the increase in network density. Individual WGCNA networks were sparse. If WGCNA networks were only used, there may be disconnections between genes; hence, the application of machine learning algorithm would not have been promising. Additionally, the density of Gene Ranker was significantly increased. Figure 4 shows the densities of Gene Rankers for the six diseases. For each disease panel, the black outlined circles indicate the densities of the WGCNA networks (ranging from 0 to 0.8%), while the dotted red lines indicate their average. The number of outlined circles varies in the panel because the number of selected WGCNA networks varies by disease. The blue circle is the density of Gene Ranker. For rheumatoid arthritis, the average density over 28 distinct WGCNA networks was 0.067%. However, the value increased to 1.431% in Gene Ranker (Rheumatoid Arthritis), which was approximately 22 times denser than the individual WGCNA networks. A similar improvement in network density was observed for other diseases. Thus, Gene Ranker could collect more information by holding a large number of genes and the relations between them in the network.

Composition of member networks: We then examined the composition of member networks in Gene Ranker. Although Gene Ranker is disease-specific, the member networks were not limited to their own WGCNA networks, which were only designated for the disease.

Table 1. Data description: sources for backbone network, add-on networks, immune diseases, immune specific gene relations and validation for studies

Backbone network	
Genes	**16 593** genes
Protein–protein interaction	**244 767** relations
Sources: Entrez Gene, DIP, PharmDB, MINT and PharmGKB	
Add-on network	
Gene expression data	**186** (WGCNA) datasets of 27 diseases
Source: GEO	
Diseases	
Immune diseases	**27** diseases (in MeSH)
Arthritis, Juvenile	Leukemia, Lymphocytic, Chronic, B-Cell
Arthritis, Rheumatoid*	Leukemia, T-Cell
Asthma*	Lymphoma
Burkitt Lymphoma	Lymphoma, B Cell
Colitis, Ulcerative	Lymphoma, Follicular
Inflammatory Bowel Diseases*	Lymphoma, T-Cell
Crohn's Disease*	Lymphoma, T-Cell, Peripheral
Dermatitis, Allergic Contact	Common Variable Immunodeficiency
Dermatitis, Atopic	Multiple Myeloma
Diabetes Mellitus, Type 1	Multiple Sclerosis*
Glomerulonephritis, IGA	Sjogren's Syndrome
HIV Infections*	Urticaria
Hypersensitivity	Waldenstrom Macroglobulinemia
Monoclonal Gammopathy of Undetermined Significance	
Disease–gene relation	**2420** genes across 27 diseases
Sources: MeSH, ChEMBL, DCDB, DrugBank, T3DB, TTD and CTD	
Validations	
Literatures/Studies	**27 931 712** literatures of 27 diseases and 16 593 proteins
Source: PubMed (US National Library of Medicine National Institutes of Health)	

Note: Diseases with superscripts are used for validation in this paper.

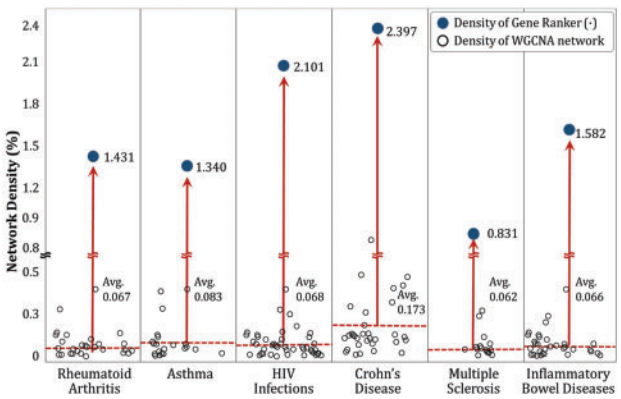

Fig. 4. Densities of networks: Gene Ranker versus WGCNA networks

Given a disease, those WGCNA networks specific for the disease are denoted as endogenous networks. In contrast, those that are not specific for the given disease are denoted as exogenous WGCNA networks. The pie chart in Figure 5 depicts the proportion of endogenous and exogenous networks when a disease is specified. The number in the inner circle of the pie represents the total number

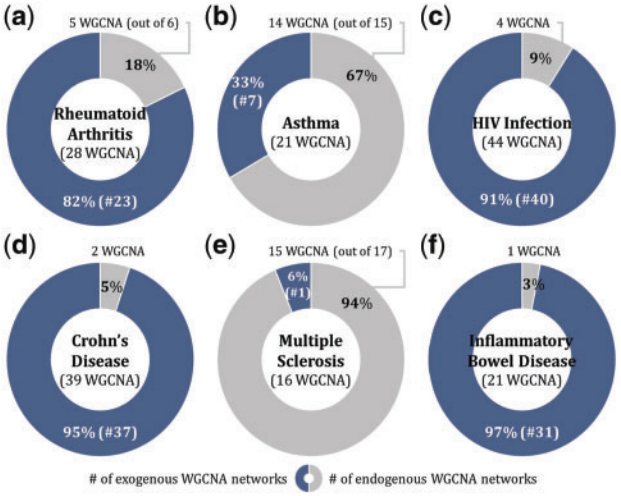

Fig. 5. Composition of member networks of Gene Ranker

of networks selected for Gene Ranker. The pies charts reveal a common pattern: if the endogenous WGCNA networks are not sufficient, Gene Ranker compliments the vacancy of information from the exogenous networks. If the endogenous WGCNA networks are sufficient, Gene Ranker takes most of the information from its own networks. For instance, in Gene Ranker (Rheumatoid Arthritis), five of six networks were taken from its own WGCNA, while the remaining 23 networks were selected from those of other diseases. It was approximately 82% of its composition. On the contrary, for Gene Ranker (Asthma), 67% of its members were chosen from the endogenous WGCNA networks, while only 33% of its members were chosen from the exogenous networks.

3.3 Performance results of gene ranker

The objective of Gene Ranker is to align genes in order of the strength of their association with a given disease. The higher the score that a gene gains, the closer its association with the disease. Therefore, top-tier genes on the scoring curve were regarded as key genes. To verify this, we performed a blind test. Some de facto genes for a certain disease (known to be associated with the disease) were assumed to be unknown. After gene scoring by Gene Ranker, we tested how many de facto genes were highly ranked. In the experiment, 10% of de facto genes were assumed to be unknown by setting the labels as '0,' and the labels of 90% of de facto genes were set as '1.' Note that there are very large numbers of unknown genes (not by assumption), and their labels were set as '0' by default. For instance, in the case of rheumatoid arthritis, there were 351 de facto genes. In the experiment, 316 genes were labeled (set as '1') and 35 plus 16 242 were unlabeled (set as '0'). Ten sets of experiments per disease were carried out in a similar manner for a 10-fold cross-validation, and this procedure was repeated 10 times.

Figure 6 shows the effect of adding the WGCNA networks to the backbone network. The figure shows the case of rheumatoid arthritis. The AUC of the backbone network was 0.745. As the WGCNA networks were added on the backbone network, the performance gradually increased, reaching an AUC of 0.815. The red dots indicate endogenous networks, while the black dots indicate exogenous networks. Notably, there was a sudden up-rise on the curve when the first WGCNA network was added to the backbone network. This indicates that the PPI network presents the overall ubiquity among genes on Gene Ranker, whereas the WGCNA

networks add the peculiarity of disease-specific genes. Additionally, note that the exogenous networks contribute to improving the performance of Gene Ranker. This reserves a conjecture for an extended study such that the immune diseases may be associated to one another by sharing a common genetic information through genetic interaction. It is notable that there are some studies in line with our conjecture (Abou-Raya and Abou-Raya, 2006; Ermann and Fathman, 2001; Gregersen and Olsson, 2009; Mariani, 2004). Additional performance information for other diseases is provided in Supplementary Figure B1 in the Appendix.

Figure 7 shows the comparative results of the overall performance of Gene Ranker. The *x*-axis represents diseases and the *y*-axis indicates the AUC values. For a disease, the bundle of bars indicates the performance of the PPI only network (gray bar), PPI plus endogenous network (blue bar) and PPI plus endogenous and exogenous networks (red bar).

The average AUC of the PPI network is 0.766 for the six diseases shown in the figure. When adding disease-specific (endogenous) WGCNA networks onto the PPI network, the performance was increased by 5.61% (from 0.766 to 0.809). This value was further increased when integration was conducted for the selected networks from endogenous and exogenous networks. The highest AUC was obtained from the last one. This amounts to an average 0.828 AUC. The results can be interpreted as follows: the increase from the gray bar to the blue bar can be explained as the endogenous WGCNA networks enhancing Gene Ranker by adding disease-specific traits onto the PPI network. In contrast, the increase from the blue bar to red bar indicates that the exogenous networks from other diseases may compliment Gene Ranker with common traits genetically shared by immune diseases.

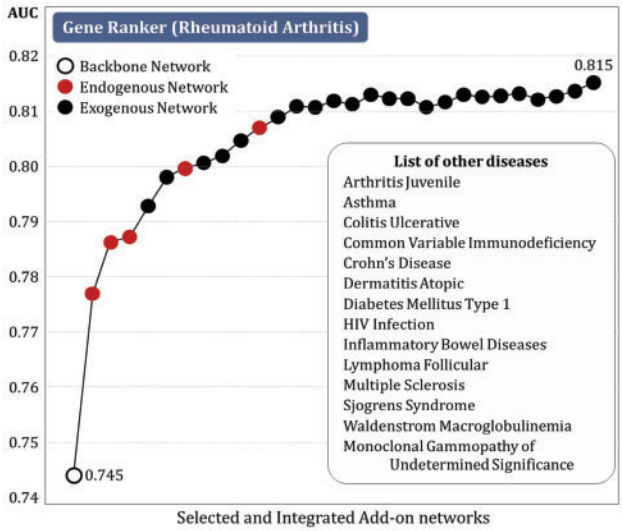

Fig. 6. Effect of adding WGCNA networks to the backbone network for rheumatoid arthritis: Gene Ranker (Rheumatoid Arthritis) is constructed by PPI plus 28 WGCNA across 15 diseases. The list of other diseases is shown in the inner box

3.4 Implication and validation of gene ranker

To validate Gene Ranker, we performed gene scoring to six Gene Rankers. All de facto genes for a disease were set as '1' for the labeled set, and the remaining genes were set as '0.' For rheumatoid arthritis, it has 351 de facto genes. After gene scoring by Gene Ranker, the resulting gene scores and ranks were obtained. With the scoring results, some genes were validated through retrieval and text mining by reviewing nearly 27 million literatures in the PubMed database.

The results in Figure 8 exemplify a typical output of Gene Ranker (Rheumatoid Arthritis). In Figure 8a, the mesh grid presents the subnetwork of genes in the case of rheumatoid arthritis. The genes are laid on the grid in the *xy* plane, and the vertical axis indicates the scores of the genes, i.e. the value of f in (2). Higher peaks (in black) of scores were observed for known genes such as OTC, B3GNT9 and C1orf167, among others. Most of the lower peaks (in blue) were from unknown genes. Compared to genes on the floor, these genes showed relatively huge potential to be key genes in rheumatoid arthritis. Figure 8b shows the score curve for rheumatoid arthritis. In the figure, the solid line stands for the values of gene score f of the top 500 ranked genes. The red circles on the line correspond to the potential key genes. On the score curve, the inserted genes (marked as red) between known genes are regarded as strong candidate genes. Table 2 summarizes the ranking of de facto genes and potential candidate genes. Some discovered genes were validated by PubMed literature reviews, which are marked with a superscript in the list. Below are some quotes from the literature. Additional validation of diseases is provided in Supplementary Appendix C.

- *UCN3*[a] and *UCN2*[c]: These data indicate the role of endogenous CRF, *UCNs* and CRFR2 in the osteoarthritic and *rheumatoid arthritis* joint microenvironment (Pérez-García *et al.*, 2011).
- *TIE1*[b]: The present study shows that unique ASV derived from receptors that play key roles in angiogenesis, namely, VEGF receptor type 1 and for the first time *TIE1*, can markedly reduce the *rheumatoid arthritis* severity (Jin *et al.*, 2008).

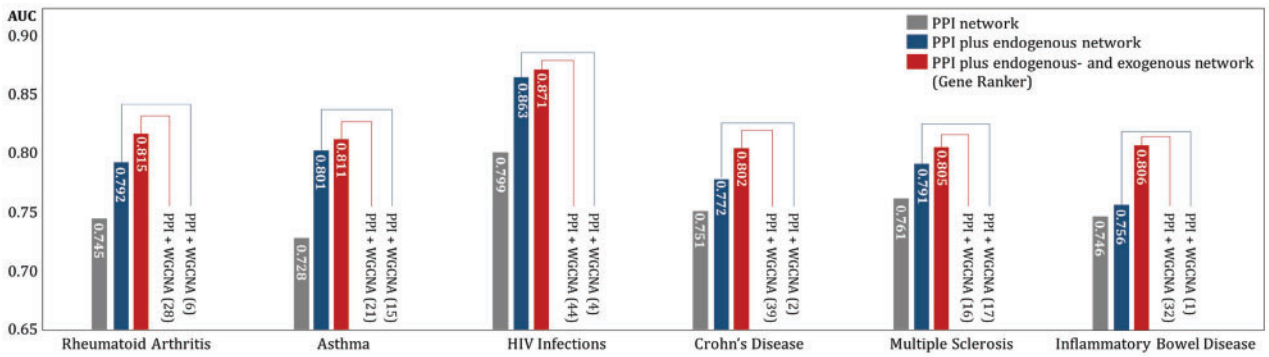

Fig. 7. Comparative result of overall performance of Gene Ranker

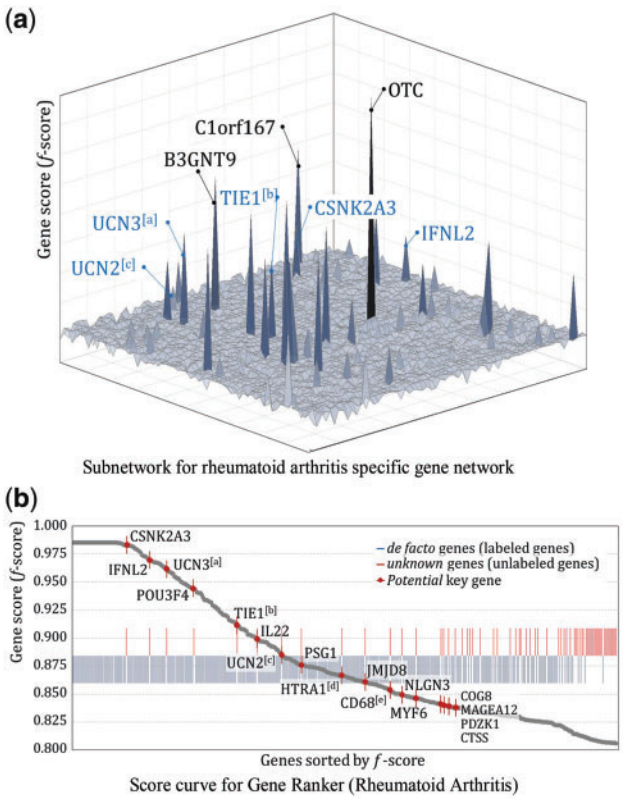

Fig. 8. Output of Gene Ranker (Rheumatoid Arthritis)

Table 2. Ranking of de facto and potential genes: list of top 10 ranked known and unknown genes

Ranking of de facto genes	
(1) OTC	(6) NKAPL
(2) B3GNT9	(7) PSG5
(3) C1orf167	(8) NUBPL
(4) KIAA1109	(9) HOXD11
(5) CST5	(10) SLC22A11
Ranking of candidate genes	
(1) CSNK2A3	
(2) IFNL2	
(3) UCN3[a]	PMID: 21360527
(4) POU3F4	
(5) TIE1[b]	PMID: 18593464, 14991531, 12010571
(6) IL22	
(7) UCN2[c]	PMID: 21360527
(8) PSG1	
(9) HTRA1[d]	PMID: 24907345, 20533271
(10) CD68[e]	PMID: 12634940, 19660107, 15647425

Note: The candidate genes are inserted between known genes. The genes with superscripts are validated by studies in the literature.

- *HTRA1[d]*: This study offers new insights into the molecular regulation of HTRA1 expression and its role in RA pathogenesis, which may have significant impacts on clinical therapy for RA and possibly other HTRA1-related diseases, including osteoarthritis, age-related macular degeneration and cancer (Hou *et al.*, 2014).
- *CD68[e]*: CD68 expression was also associated with erosions and radiological progression in rheumatoid arthritis (Salvador *et al.*, 2005).

4 Discussion

In this study, we developed an *in silico* method called Gene Ranker for key gene identification in immune diseases. Immune and inflammatory diseases have a strong genetic component with Mendelian patterns of inheritance (Gregersen and Olsson, 2009). Additionally, a major concept of the underlying pathophysiology of autoimmune diseases has been evaluated in genome-wide association scans, which have identified a degree of genetic sharing among autoimmune diseases such as rheumatoid arthritis and multiple sclerosis, among others (Cotsapas and Hafler, 2013). Although the common features of these diseases remain unclear, they are generally grouped in the immune category (Mariani, 2004). Based on the potential commonality among immune genes, we designed Gene Ranker. This algorithm was initially constructed based on the network of protein interactions, and then the patients' gene expression networks were added onto the PPI network. In the add-on process, disease-specific networks were obtained from disease-wise WGCNA analyses of samples from patients with a specific immune disease. The PPI network revealed the overall ubiquity among genes, whereas the patients' WGCNA networks added the peculiarities of the respective immune diseases. However, any of these networks can be used in common in the category of immune diseases, regardless of the given disease, if they pass the screening procedure. Thus, the possibility of *pan immune disease genes* was reflected. In the experiment, we demonstrated the superiority of Gene Ranker, and the resulting genes (the highly ranked ones) were validated through retrieval and text mining by reviewing nearly 27 million literatures in the PubMed database.

One advantage of Gene Ranker is that it can predict key genes even when there are few known genes, which is the primary difficulty in evaluating immune diseases. It is particularly advantageous for a new drug-target discovery, which has not been thoroughly studied because of lack of known facts about the disease–gene association. From a methodological perspective, Gene Ranker is scalable and evolvable, and can be instantly and easily updated Gene Ranker with the most recent knowledge.

Some aspects of this study remain as future work. First, we should further expand the scope of the study to other disease categories (i.e. not limited to immune diseases). Second, a more sophisticated method is required when integrating networks to determine the importance of individual networks.

Acknowledgements

The author would like to gratefully acknowledge support from the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF-2018R1D1A1B07043524), ICT R&D program of MSIP/IITP (No. 2017-0-00887).

Funding

This study was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2018R1D1A1B07043524, No. 2017-0-00887).

Conflict of Interest: none declared.

References

Abou-Raya,A. and Abou-Raya,S. (2006) Inflammation: a pivotal link between autoimmune diseases and atherosclerosis. *Autoimmun. Rev.*, **5**, 331–337.

Bakheet,T.M. and Doig,A.J. (2009) Properties and identification of human protein drug targets. *Bioinformatics*, **25**, 451–457.

Barski,L. *et al.* (2013) Comparison of diabetic ketoacidosis in patients with type-1 and type-2 diabetes mellitus. *Am. J. Med. Sci.*, **345**, 326–330.

Campillos,M. *et al.* (2008) Drug target identification using side-effect similarity. *Science*, **321**, 263–266.

Chapelle,O. *et al.* (2009) Semi-supervised learning (Chapelle,O. et al. eds.; 2006)[book reviews]. *IEEE Trans. Neural Netw.*, **20**, 542-542.

Chiang,A.P. and Butte,A.J. (2009) Systematic evaluation of drug-disease relationships to identify leads for novel drug uses. *Clin. Pharmacol. Therap.*, **86**, 507.

Cotsapas,C. and Hafler,D.A. (2013) Immune-mediated disease genetics: the shared basis of pathogenesis. *Trends Immunol.*, **34**, 22–26.

Ermann,J. and Fathman,C.G. (2001) Autoimmune diseases: genes, bugs and failed regulation. *Nat. Immunol.*, **2**, 759.

Ganegoda,G.U. *et al.* (2015) ProSim: a method for prioritizing disease genes based on protein proximity and disease similarity. *BioMed Res. Int.*, **2015**, 1.

Gardiner,L. *et al.* (2012) Structural equation modeling highlights the potential of Kim-1 as a biomarker for chronic kidney disease. *Am. J. Nephrol.*, **35**, 152–163.

Gregersen,P.K. and Olsson,L.M. (2009) Recent advances in the genetics of autoimmune disease. *Annu. Rev. Immunol.*, **27**, 363–391.

Grigoriev,A. (2001) A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. *Nucleic Acids Res.*, **29**, 3513–3519.

Hou,Y. *et al.* (2014) The inhibitory effect of IFN-γ on protease HTRA1 expression in rheumatoid arthritis. *J. Immunol.*, **193**, 130–138.

Hurle,M. *et al.* (2013) Computational drug repositioning: from data to therapeutics. *Clin. Pharmacol. Therap.*, **93**, 335–341.

Jin,P. *et al.* (2008) Novel splice variants derived from the receptor tyrosine kinase superfamily are potential therapeutics for rheumatoid arthritis. *Arthritis Res. Ther.*, **10**, R73.

Lee,J.-H. *et al.* (2015) PharmDB-K: integrated bio-pharmacological network database for traditional Korean medicine. *PLoS One*, **10**, e0142624.

Mariani,S.M. (2004) Genes and autoimmune diseases—a complex inheritance: highlights of the 54th Annual Meeting of the American Society of Human Genetics; October 26-30, 2004; Toronto, Ontario, Canada. *Medscape Gen. Med.*, **6**, 18.

Pérez-García,S. *et al.* (2011) Mapping the CRF–urocortins system in human osteoarthritic and rheumatoid synovial fibroblasts: effect of vasoactive intestinal peptide. *J. Cell. Physiol.*, **226**, 3261–3269.

Salvador,G. *et al.* (2005) p53 expression in rheumatoid and psoriatic arthritis synovial tissue and association with joint damage. *Ann. Rheumatic Dis.*, **64**, 183–187.

Shin,H. *et al.* (2007) Graph sharpening plus graph integration: a synergy that improves protein functional classification. *Bioinformatics*, **23**, 3217–3224.

Shin,H. *et al.* (2009) Protein functional class prediction with a combined graph. *Exp. Syst. Appl.*, **36**, 3284–3292.

Tsuda,K. *et al.* (2005) Fast protein classification with multiple networks. *Bioinformatics*, **21**, ii59–ii65.

Venter,J.C. *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.

Zhao,J. *et al.* (2011) Ranking candidate disease genes from gene expression and protein interaction: a Katz-centrality based approach. *PLoS One*, **6**, e24306.

Zhao,S. and Li,S. (2010) Network-based relating pharmacological and genomic spaces for drug target identification. *PLoS One*, **5**, e11764.

Zhu,X. and Goldberg,A.B. (2009) Introduction to semi-supervised learning. *Synth. Lect. Artif. Intell. Mach. Learn.*, **3**, 1–130.