Novel applications of rank aggregation methods in the realm of disease gene prioritisation

G van Zyl*, JH van Vuuren

Stellenbosch Unit for Operations Research in Engineering, Department of Industrial Engineering, Stellenbosch University, Stellenbosch, South Africa

Private Bag X1, Matieland, Stellenbosch, 7602, South Africa

Abstract

With the development of more sophisticated low-level data integration strategies, the approach towards integrating heterogeneous proteomic data in the context of disease gene prioritisation which once dominated, namely by means of rank aggregation (RA), may soon become obsolete. In this paper, novel applications of RA methods in the context of disease gene prioritisation are explored. To this end, a novel disease gene prioritisation framework is proposed for leveraging the desirable properties of both low-level data integration strategies and high-level data integration strategies, whilst simultaneously avoiding the significant drawbacks typically encountered when implementing the respective approaches in a disease gene prioritisation problem setting. The proposed framework is demonstrated practically and found to correctly include all ten known disease proteins considered for verification purposes among the highest ranking disease protein candidates. In addition, the framework is successfully applied to obtain three sets of putative disease genes ranked according to their likelihood of contributing to disease — a number of which are validated by retrieving evidence of their involvement in the origins of diseases from the literature.

Keywords: Disease gene prioritisation, Rank aggregation, Graph-based semi-supervised learning, Protein-protein interaction networks, Heterogeneous data integration

1. Introduction

The identification of genes that predispose an individual to disease (disease genes in short) has become one of the fundamental problems in medical sciences and systems biology [1]. The most accurate approaches towards detecting disease genes require performing costly and time-consuming validation experiments involving both known disease genes and candidate genes [2]. Such experiments are furthermore complicated by the fact that the majority of human diseases are associated with simultaneous deleterious mutations in multiple genes [3, 4].

Given current estimates that humans have more than 20 000 protein-encoding genes [5], testing even relatively small subsets of pairs of candidate genes is prohibitively expensive [4]. To mitigate this problem, numerous predictive analytical and computational approaches aimed at ranking putative disease genes according to their likelihood of contributing to disease have been proposed. These approaches are collectively referred to as gene prioritisation techniques.

In recent years, protein-protein interaction (PPI) data — typically captured in undirected, possibly weighted graphs in which vertices represent proteins and edges denote interactions between proteins — have been particularly prevalent in such computational approaches towards disease gene prioritisation [6, 7, 8]. Disease gene prioritisation methods based on PPI data rely on the widely-accepted hypothesis that "the network neighbour of a disease protein is likely to cause the same or a similar disease" [9, 10, 11, 12, 13].

In addition, the majority of pre-eminent PPI-based disease gene prioritisation methods also exploit additional biological data, such as phenotype similarity data, gene expression data and protein pathway data, which have been shown to significantly improve the performance of disease gene prioritisation models [14, 15, 16, 17, 18, 19, 20, 21, 22, 23].

Leveraging additional biological data from a multitude of sources, however, necessitates the integration of various heterogeneous data types. To this end, numerous approaches towards the integration of data from multiple omics platforms have been proposed [24, 25, 26, 27]. Both Nguyen and Ho [17] and Nam et al. [14], for example, applied graph-based semi-supervised learning (SSL) algorithms to homogenous graphs (in which all vertices represent proteins) by distilling heterogeneous biological data from multiple sources into edge weights denoting the degree of functional similarity between proteins.

Another widely-implemented approach towards data integration in the realm of PPI-based disease gene prioritisation entails constructing heterogeneous graphs (in which vertices may represent different types of biological entities, including proteins, genes, phenotypes, diseases and protein complexes) and applying a graph-based algorithm to the integrated graph [16, 18, 19, 23, 28, 29].

The above-mentioned approaches towards integrating data from various sources are examples of low-level integration strategies in which various data sets are integrated *before* training the disease gene prioritisation model.

*Corresponding author

Email address: ghietevanzyl@gmail.com (G van Zyl)

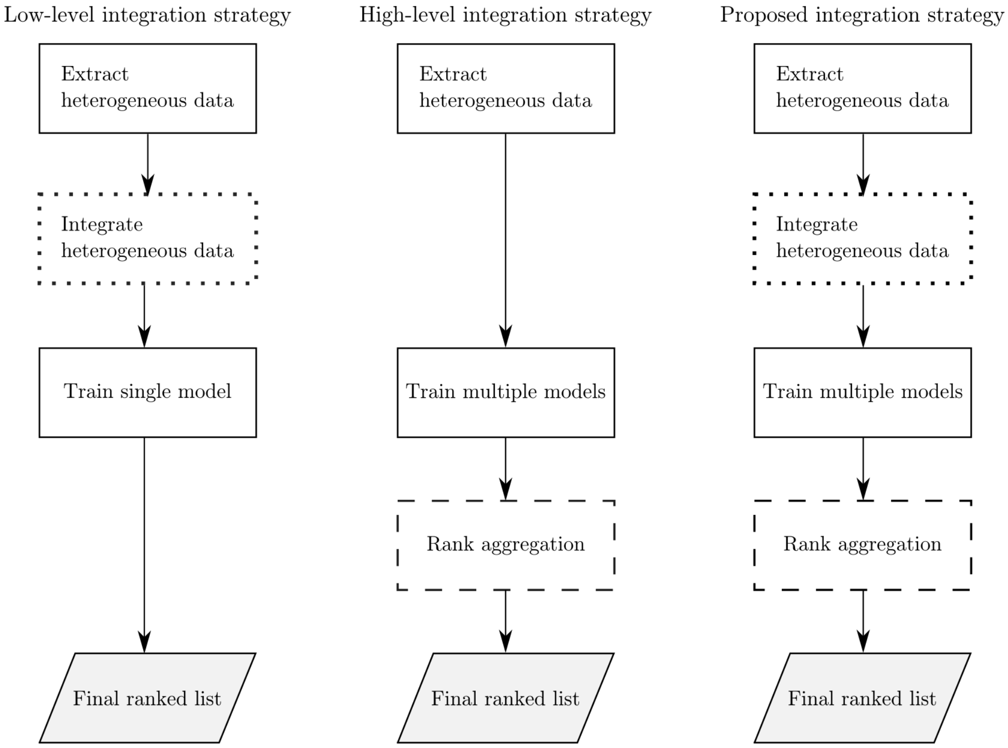
Low-level integration strategy
High-level integration strategy
Proposed integration strategy
Extract heterogeneous data
Integrate heterogeneous data
Train single model
Final ranked list
Extract heterogeneous data
Train multiple models
Rank aggregation
Final ranked list
Extract heterogeneous data
Integrate heterogeneous data
Train multiple models
Rank aggregation
Final ranked list

Figure 1: A high-level view of various approaches towards the integration of heterogeneous proteomic data adopted in disease gene prioritisation efforts.

The majority of PPI-based disease gene prioritisation methods developed to date, however, employ high-level integration strategies in which data integration is performed *after* model-development and training. In the context of disease gene prioritisation, high-level integration strategies typically involve training separate models using different data sets (describing different protein features) to obtain multiple sets of ranked proteins which are subsequently combined into one unifying prediction by means of rank aggregation (RA) [15, 30, 31, 32, 33, 34, 35, 36]. The distinction between low-level and high-level integration strategies in the context of the disease gene prioritisation problem is elucidated in the first and second flow-diagrams of Figure 1.

The ubiquity of RA in disease gene prioritisation efforts may be attributed to a number of desirable properties shared by rank-based algorithms, such as their robustness to outliers [37] and the fact that, provided that the relative orderings of items in a ranked list are preserved, rankings are invariant to transformation and normalisation [38, 39, 40]. In addition, ranked lists are an intuitive representation of gene prioritisation model outputs.

Another useful property of RA methods is that RA algorithms allow for the synthesis of individual results (i.e. separate ranked lists corresponding to different data sets) to be integrated without modelling the interdependencies between the various raw data. Intuitively, modelling the interrelations between the raw data may pose a significant challenge if current understanding of the complexities of the system being modelled is limited.

This same property may, however, be considered a weakness of RA methods if sufficient knowledge of the interrelations between these data exist and this knowledge can be leveraged in order to gain valuable insights that ultimately result in improved model performance. In the context of disease gene prioritisation, for example, training models corresponding to a particular protein feature in isolation disregards the potentially informative relationships between the various proteomic data [16].

In addition, the majority of RA methods employed in disease gene prioritisation studies consider the rankings in each of the input lists to be equally informative. That is, each of the protein features used to generate these lists are considered equally relevant — an assumption which rarely holds [16, 17, 34, 41].

Although a number of RA algorithms which afford certain lists greater influence over the final aggregate rankings have been proposed, this is typically achieved by employing some weighting scheme specified by the practitioner. With respect to disease gene prioritisation, however, research on the relative importance of different types of biological data is still in its infancy. Consequently, confidently selecting suitable weights for the various protein features may prove challenging, if not impossible.

For the above-mentioned reasons, RA is not recommended for the purposes of integrating heterogeneous proteomic data as in existing disease gene prioritisation efforts employing RA. Rather, we propose a novel application of RA methods in the context of disease gene prioritisation which is unaffected by the aforementioned weaknesses whilst also benefiting from RA methods' more desirable properties.

To this end, we have designed a two-stage RA procedure which may be implemented in combination with the graph-based SSL disease gene identification framework in [42] leveraging various proteomic data (integrated by means of a low-level integration strategy) and performing disease gene classification within a graph-based SSL paradigm. We recommend using this RA procedure in order to condense the outputs of the disease gene identification framework, namely by ranking multiple lists of disease protein candidates ranked, according to their likelihood of contributing to disease, into a final list of globally ranked disease gene candidates. The framework in [42] is hereafter referred to as the disease gene *identification* framework and the combined framework, comprising both the disease gene identification framework in [42] and the two-stage RA procedure proposed in this paper, is referred to as the disease gene *prioritisation* framework.

The proposed disease gene prioritisation framework employs both low-level data integration strategies and RA techniques, as shown in the third flow-diagram of Figure 1. To the best of our knowledge, this is the first PPI-based disease gene prioritisation framework to incorporate both low-level and high-level data integration strategies within the same framework. Our focus in this paper is on the RA procedure employed in the proposed disease gene prioritisation framework.

During the first stage of the RA procedure, thirty nine sets of ranked disease protein candidates (each containing separate ranked lists corresponding to five different graph-based SSL algorithms) are aggregated into thirty nine ranked lists, each of which contains disease protein candidates that are likely to contribute to the same or phenotypically similar diseases.

Thereafter, during the second stage of the RA procedure, the thirty nine ranked lists generated during the first stage are aggregated into one final list of globally ranked disease protein candidates. Finally, the encoding genes of the highest ranking proteins in this final globally ranked list are identified in order to obtain a set of strong candidate disease genes which may be recommended for further validation experiments and functional studies.

2. Preliminaries and theory

Various approaches towards solving the RA problem have been proposed in the literature. These approaches may be classified broadly as distribution-based algorithms, optimisation algorithms, Bayesian methods and heuristics [43].

A recent comparative study of fourteen RA methods in genomic applications [44] found that three heuristic methods developed within a Markov chain modelling framework, namely MC1, MC2 and MC3, consistently ranked among the best-performing RA algorithms, outperforming even the more recent and computationally expensive Bayesian methods [44]. The MC1, MC2 and MC3 algorithms furthermore demonstrated superior performance in web search applications [45]. The findings in [44, 45], combined with the intuitive nature and computational efficiency of heuristic RA algorithms [43], motivated their implementation in the disease gene prioritisation framework proposed in this paper. The MC1, MC2 and MC3 algorithms are considered in more detail in the remainder of this section.

2.1. Notation

Let $\mathcal{T} = \{1, 2, \ldots, |\mathcal{T}|\}$ denote an unordered list. A permutation of $\mathcal{T}$, denoted by $\Omega(\mathcal{T}) = (t_1, t_2, \ldots, t_{|\mathcal{T}|})$, such that $R_\Omega(t_i) < R_\Omega(t_j)$ for any $i < j$, is an ordered list of the elements in $\mathcal{T}$ that does not allow for ties. Here, $R_\Omega(t)$ denotes the rank of element t under the ranking mechanism Ω.

Moreover, the top-k list $\mathcal{S} = (s_1, s_2, \ldots, s_{|\mathcal{S}|})$ comprises the top-most $k = |\mathcal{S}|$ elements in $\Omega(\mathcal{T})$ and it is implicitly assumed that an element in $\mathcal{T}$ which does not appear in $\mathcal{S}$ has a rank lower than k.

2.2. Markov chain-based methods

In Markov chain-based RA algorithms [43], the union of the elements from the given top-k lists forms the state space of an ergodic Markov chain. The transition matrix of this Markov chain is constructed in such a way such that the chain's steady state distribution will have larger probabilities corresponding to states (the elements to be ranked) that are ranked higher. These Markov chain RA algorithms, called MC1, MC2 and MC3, are differentiated by their transition probability matrix construction methodologies.

Let $\mathcal{S}_1, \ldots, \mathcal{S}_L$ be L top-k lists with corresponding underlying ranking mechanisms $\Omega_1, \ldots, \Omega_L$. Then, $\mathcal{S} = \cup_{\ell=1}^{L} \mathcal{S}_\ell$ comprises all the elements to be ranked and makes up the state space of the Markov chain. Each element to be ranked is, therefore, represented by a state in the Markov chain.

For each state $u \in \mathcal{S} \cap \mathcal{S}_\ell^c$, where $\mathcal{S}_\ell^c$ is the complement of $\mathcal{S}_\ell$, define $R_{\Omega_\ell}(u) = k_\ell + 1$. The transition probability p_{uv} from state u to state v may then be computed using information about pairwise rankings (e.g. state v is ranked higher than state u in the individual list $\mathcal{S}_\ell$).

The MC1 [46] transition probability matrix is constructed as follows: For each $u \in \mathcal{S}$, the probability that the chain transitions from state u to state v in one time period is

$$p_{uv} = \begin{cases} 1/|\mathcal{S}| & \text{if } R_{\Omega_\ell}(u) > R_{\Omega_\ell}(v) \text{ for at least one of} \\ & \text{the input lists,} \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

where $v \neq u$ and $v \in \mathcal{S}$.

The MC2 [46] algorithm defines the probability that the chain transitions from state u to state v as

$$p_{uv} = \begin{cases} 1/|\mathcal{S}| & \text{if } R_{\Omega_\ell}(u) > R_{\Omega_\ell}(v) \text{ for a majority of} \\ & \text{the input lists,} \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

where $v \neq u$ and $v \in \mathcal{S}$.

The final Markov chain-based RA algorithm, namely MC3 [46], constructs the transition probability matrix as follows: For each $u \in \mathcal{S}$ let

$$p_{uv} = \frac{\sum_{\ell=1}^{L} I\left(R_{\Omega_\ell}(u) > R_{\Omega_\ell}(v)\right)}{L|\mathcal{S}|}, \tag{3}$$

where $v \neq u$, $v \in \mathcal{S}$, and $I(\cdot)$ is the indicator function that evaluates to 1 if its argument condition is satisfied, or 0 otherwise.

For each of the above-mentioned algorithms, the probability that the chain remains in its current state is defined as

$$p_{uu} = 1 - \sum_{\substack{v \neq u, \\ v \in \mathcal{S}}} p_{uv}. \tag{4}$$

Finally, in order to ensure that a Markov chain is ergodic, the transition probabilities may be modified as follows

$$p'_{uv} = (1 - \alpha) p_{uv} + \alpha / |\mathcal{S}|, \tag{5}$$

where α is a tuning parameter which is usually assigned a small value.

3. Methodology

The disease gene prioritisation framework proposed in this paper is outlined in Figure 2. This section contains a concise summary of the disease gene identification component of the framework in Sections 3.1 – 3.5. This is followed by a more detailed description of the RA component of the disease gene prioritisation framework, indicated by the shaded area in Figure 2, in Sections 3.6 – 3.8.

3.1. Approximating the human interactome

First, an approximation of the human interactome, hereafter referred to as the full PPI graph, is constructed using PPI data extracted from the iRefIndex database [47]. Thereafter, known disease proteins in the full PPI graph are identified by means of genotype-phenotype associations in the Online Mendelian Inheritance in Man (OMIM) [48] and DisGeNET [49] databases.

3.2. Constructing feature vectors

Additional biological data related to protein domains, protein pathways and protein complexes are retrieved from the Pfam [50], Reactome [51] and CORUM [52] databases, respectively.

In combination with the PPI data, these additional biological data are leveraged in order to construct feature vectors capturing descriptive data for each protein in the full PPI graph. More specifically, each feature vector characterise the corresponding protein in terms of twelve attributes. The first eight attributes describe the degree and betweenness centrality of the protein in the full PPI graph, the number of known disease proteins in that protein's 1st- and 2nd-order neighbourhoods, the proportions of known disease proteins in the protein's sets of 1st- and 2nd-order neighbours, the number of disease domains associated with the protein (a domain is considered a disease domain if it is associated with a known disease protein) and the number of disease domains relative to the total number of domains associated with that protein.

The ninth and tenth attributes describe the corresponding protein in terms of the protein pathways with which that protein is associated. To this end, each pathway is first assigned two 'scores' capturing the number of disease proteins associated with that pathway and the ratio of disease proteins relative to the total number of proteins associated with that pathway, respectively. The ninth and tenth protein attributes are then computed by summing the first and second 'scores' of each of the pathways with which that protein is associated, respectively.

The last two attributes incorporating protein complex data are generated similarly to these pathway attributes. That is, two 'scores' are once again computed for each protein complex based on the number and proportion of known disease proteins in that protein complex. The eleventh and twelfth attributes are subsequently computed by taking the sum of the first and second 'scores' recorded for each of the protein complexes to which the corresponding protein belongs, respectively.

3.3. Constructing PPI communities

Thirty nine phenotype modules (highly connected subgraphs which are themselves sparsely connected to each other) are computationally extracted from a human phenotype-phenotype network constructed by van Driel et al. [53], by employing the community detection algorithm proposed by White and Smyth [54], called Spectral-1.

Thereafter, the full PPI graph was partitioned into (overlapping) subgraphs corresponding to these thirty nine phenotype modules, hereafter referred to as PPI communities, where proteins that belong to the same community are expected to be associated with phenotypically similar diseases [55, 56, 57, 58, 59, 60].

Since proteins in the same PPI community are assumed to be associated with phenotypically similar diseases, the 'phenotype feature' of proteins belonging to the same PPI community are highly similar and thus have very little predictive power (within that community). Therefore, phenotype-phenotype associations need not be considered when calculating the degree of similarity between proteins in the same PPI community.

For each PPI community a set of known disease proteins is extracted from the genotype-phenotype associations in Section 3.1. Since no proteins have definitively been classified as non-disease proteins, techniques from the realm of positive-unlabelled (PU) learning are employed in order to generate a set of 'reliable' non-disease proteins for each PPI community. These known disease proteins and 'reliable' non-disease proteins comprise the positive and negative data sets, respectively.

Subsequently a test : validation : training ratio of 80 : 16 : 4 is used in order to generate ten different combinations of testing, validation and training data for each PPI community. Hereafter, a particular combination of testing, validation and training data, is referred to as a sample of the associated PPI community.

3.4. PPA graphs

The heterogeneous data captured in the feature vectors of Section 3.2 are distilled into protein-protein association (PPA) graphs by applying the PG-LEARN [61] algorithm — a state-of-the-art graph construction hyperparameter learning algorithm developed expressly for graph-based SSL applications. A separate PPA graph is constructed for every sample of each of the thirty nine PPI communities in order to obtain 390 distinct integrated data sets. By capturing the heterogeneous data in PPA

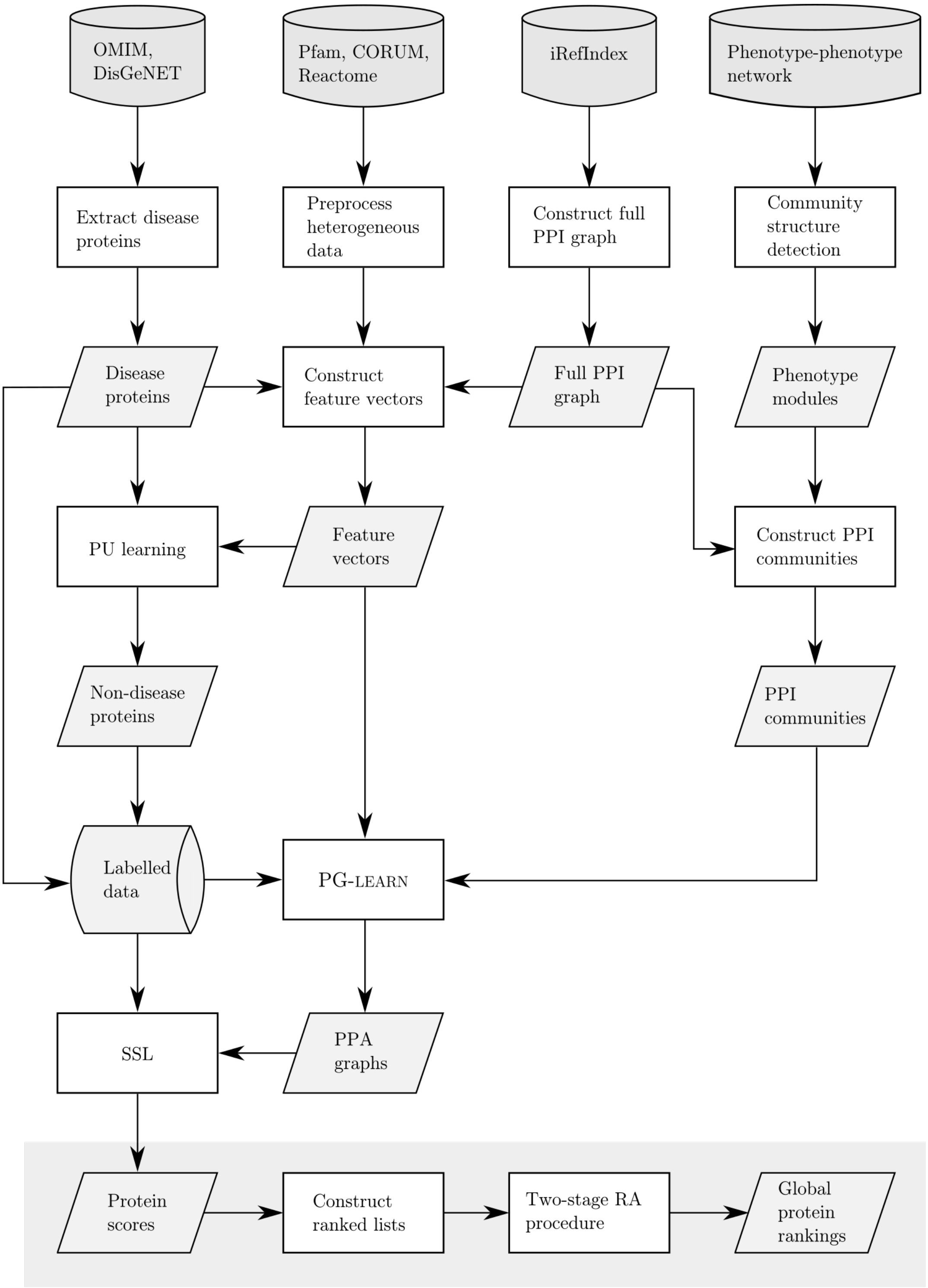

Figure 2: Proposed framework for the prioritisation of putative disease proteins by means of graph-based SSL.

graphs that closely approximate the true structure of the underlying data, the interrelation between the various proteomic data can be leveraged in order to gain additional insights that may result in improved model performance.

It should be highlighted that, in effect, two different low-level data integration strategies are employed in order to incorporate the various data types into these PPA graphs. The selection of these integration strategies was guided by the properties of the data to be incorporated into the PPA graphs.

Recall that phenotype similarity data were leveraged by constructing thirty nine PPI communities corresponding to disease modules extracted from the human phenotype-phenotype network. The adoption of this approach was motivated by the modular nature of the human phenotype-phenotype network [55, 56, 57, 58, 59, 53, 60] which has been exploited in order to gain new insights into a number of diseases [62, 58, 63, 64, 65, 66, 67].

The remaining data types (the proteomic data captured in the feature vectors) are then integrated using the PG-LEARN algorithm which not only allows for the relative importance of various data types (features) to be modulated, but maximises model performance by computing near-optimal values for the corresponding graph construction hyperparameter values. This leverages a crucial capability of low-level data integration strategies in disease gene prioritisation applications due to the currently limited understanding of the significance of various proteomic data in respect of disease protein identification.

This novel combination of low-level integration strategies allows for the relative importances of the various types of proteomic data to be optimised for the individual PPI communities. This is significant since the true relative importance of different types of proteomic data for the purposes of identifying putative disease genes depends on the particular disease phenotype being considered.

3.5. Protein scores

Five different graph-based semi-supervised learning (SSL) algorithms are applied to each PPA graph, namely the learning with local and global consistency (LGC) algorithm [68], the Gaussian random fields and harmonic functions (GFHF) algorithm [69], the Laplacian regularised least squares (LapRLS) algorithm [70], the greedy gradient-based MaxCut (GGMC) algorithm [71] and the kernel-induced label propagation by mapping (Kernel LP) [72] algorithm. The values assigned to a protein appearing in one of these PPA graphs by the respective algorithms are hereafter referred to as that protein's LGC score, GFHF score, LapRLS score, GGMC score and Kernel LP score, respectively

3.6. Generating ranked lists

The input lists provided to the MC1, MC2 and MC3 algorithm are generated using the protein scores obtained by means of the LGC, GFHF, LapRLS, GGMC and Kernel LP algorithms (described in Section 3.5). More specifically, for each of the unlabelled proteins in PPI community C, the arithmetic mean of the ten LGC scores assigned to a protein $\boldsymbol{x}_i$ corresponding to the ten different samples of PPI community C is computed. This value is hereafter referred to as the mean LGC score of $\boldsymbol{x}_i$ with respect to PPI community C. Similarly, the mean GFHF scores, mean LapRLS scores, mean GGMC scores and mean Kernel LP scores for each unlabelled protein in PPI community C are obtained and used to generate five ordered lists, corresponding to the five graph-based SSL algorithms, for PPI community C.

In each of these ordered lists the proteins of PPI community C are ranked in descending order of their associated mean scores. That is, the more confident a graph-based SSL algorithm is that a particular protein contributes to disease, the earlier that protein appears in the ordered list corresponding to that graph-based SSL algorithm.

It should be highlighted that, by ranking proteins according to these mean scores, as opposed to only one score corresponding to a single sample of the PPI community, the risk of obtaining an outlier ranking due to an unfortunate selection of labelled training data is mitigated.

Subsequently, five input lists corresponding to PPI community C are obtained by taking the hundred highest ranking proteins in each of these lists. This process is repeated thirty nine times, resulting in five top-k lists, each containing a hundred proteins, for every PPI community.

3.7. Validation method

In order to validate the proposed RA procedure, a number of known disease proteins are inserted into the top-k lists. Intuitively, one would expect to observe a number of these known disease proteins among the top-ranked putative disease proteins in the lists of aggregated rankings obtained.

Known disease proteins that appear exclusively in test sets and never in the training or validation sets of any of the PPA graphs in which they appear (hereafter referred to as candidate verification proteins), are extracted and their mean LGC scores, mean GFHF scores, mean LapRLS scores, mean GGMC scores and mean Kernel LP scores are computed.

Ten of these proteins are selected to make up the final set of verification proteins. The final verification proteins were selected at random from the pool of candidate verification proteins, subject to the condition that each PPI community must contain at least one of these verification proteins.

The verification proteins are subsequently inserted into the top-k lists corresponding to the PPI communities in which they appear. As for the unlabelled proteins, the position at which a verification protein is inserted into a particular top-k list is determined by that protein's mean score obtained for the corresponding PPI community and graph-based SSL algorithm. Finally, the top-k lists are truncated in order once again to obtain a hundred proteins per top-k list.

3.8. Aggregating ranked lists of putative disease proteins

Unlike previous applications of RA in the context of disease gene prioritisation problems, RA is *not* employed as part of a high-level approach towards integrating heterogeneous proteomic data. Rather, novel applications of RA methods in the context of disease gene prioritisation are explored.

To this end, the two-stage RA procedure illustrated in Figure 3 was designed and is implemented here. RA is performed using the implementations of the MC1, MC2 and MC3 algorithms contained in the R package TopKLists [73].

During the first stage of the RA procedure, each of the thirty nine PPI communities is considered in isolation. For each PPI community, the five top-k lists (corresponding to the LGC, GFHF, LapRLS, GGMC and Kernel LP algorithms, respectively) associated with that PPI community are provided as the input lists for the MC1, MC2 and MC3 algorithms. This yields three separate lists of ranked putative disease proteins per PPI community. These lists of aggregated rankings are truncated in order to obtain three ranked lists, corresponding to the three

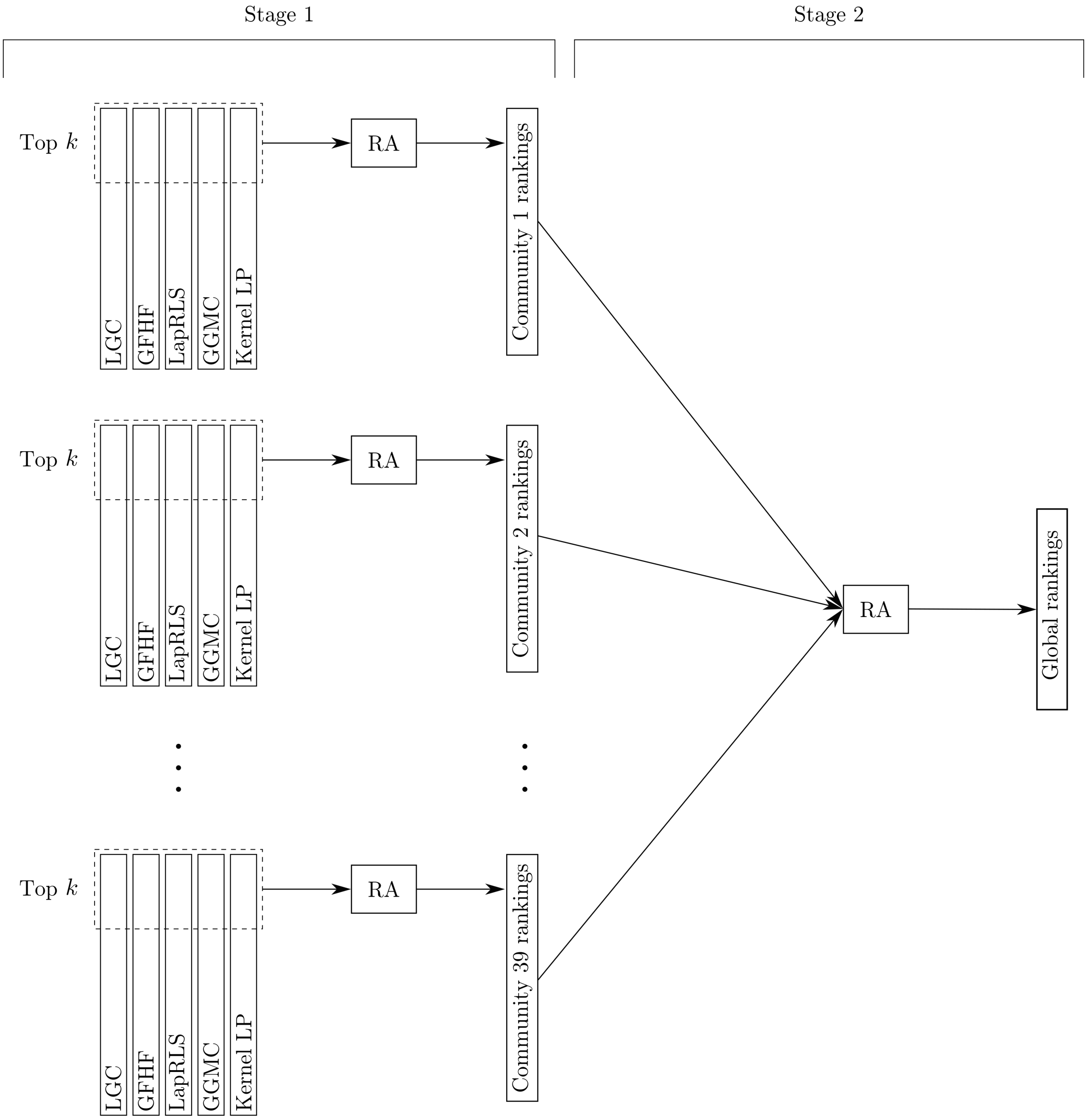

Figure 3: The two-stage RA procedure for constructing globally ranked lists of putative disease proteins.

RA algorithms, comprising the hundred strongest disease protein candidates belonging to a particular PPI community.

The intuition underlying this approach is that different models have different biases and that, if the errors that occur due to these biases are uncorrelated, the models are expected to compensate for each other's weaknesses in such a manner that the aggregate ranking obtained by the RA algorithm more closely resembles the truth than any of the individual ranked lists generated by a particular model.

Note that, whilst RA algorithms' inability to capture the interrelations between different data sets is a notable weakness when integrating heterogeneous proteomic data, the fact that RA methods are capable of synthesising multiple data sets without modelling the interdependencies between the input data sets is an advantage when consolidating outputs obtained by separate models trained on the same data set.

Thereafter, the MC1, MC2 and MC3 algorithms are employed during the second stage of the RA procedure in order to aggregate the sets of thirty nine truncated lists (corresponding to the thirty nine PPI communities) generated during the first stage of the RA procedure. This yields three globally ranked lists, corresponding to the MC1, MC2 and MC3 algorithms, respectively.

To the best of our knowledge, this is the first application of RA methods (in a disease gene prioritisation setting) for the purpose of aggregating the rankings of a protein based on the relative likelihood of its involvement in the expression of *diverging* phenotypes.

4. Results and discussion

4.1. PPI community rankings

The RA top 20 lists corresponding to PPI communities 10, 31 and 34 are provided as examples in Tables 1–3 (proteins that received perfectly consistent rankings across all three RA algorithms are italicised). Each of these ordered lists contain the twenty highest ranking putative disease proteins identified by the MC1, MC2 and MC3 algorithms corresponding to a particular PPI community.

Table 1: The twenty highest ranking putative disease proteins of PPI community 10 identified by the MC1, MC2 and MC3 algorithms, respectively.

Rank	MC1	MC2	MC3
1	*P25101*	*P25101*	*P25101*
2	*A0A0S2Z3S6*	*A0A0S2Z3S6*	*A0A0S2Z3S6*
3	*D2KUA6*	*D2KUA6*	*D2KUA6*
4	O00170	A0A024QZA9	O00170
5	A0A024QZA9	O00170	A0A024QZA9
6	Q99697	A0A024R3C5	Q99697
7	A0A024R3C5	Q15052	A0A024R3C5
8	Q15052	Q99697	Q15052
9	*B4E0R1*	*B4E0R1*	*B4E0R1*
10	*A0A1X7SBR3*	*A0A1X7SBR3*	*A0A1X7SBR3*
11	*G4XH65*	*G4XH65*	*G4XH65*
12	*A0A024R274*	*A0A024R274*	*A0A024R274*
13	A0A0S2A4E4	P16278	P16278
14	P16278	A0A0S2A4E4	A0A0S2A4E4
15	*Q99PU7*	*Q99PU7*	*Q99PU7*
16	*A0A0E3SU01*	*A0A0E3SU01*	*A0A0E3SU01*
17	*D9YZU5*	*D9YZU5*	*D9YZU5*
18	*A0A024QZD2*	*A0A024QZD2*	*A0A024QZD2*
19	*P27815*	*P27815*	*P27815*
20	Q92736	Q7Z407	Q7Z407

Table 2: The twenty highest ranking putative disease proteins of PPI community 31 identified by the MC1, MC2 and MC3 algorithms, respectively.

Rank	MC1	MC2	MC3
1	*P16278*	*P16278*	*P16278*
2	*A0A024R5K8*	*A0A024R5K8*	*A0A024R5K8*
3	*A0A140VJQ0*	*A0A140VJQ0*	*A0A140VJQ0*
4	*Q68DB7*	*Q68DB7*	*Q68DB7*
5	*A0A024R6R4*	*A0A024R6R4*	*A0A024R6R4*
6	*L7RTG7*	*L7RTG7*	*L7RTG7*
7	*A0A140VJJ7*	*A0A140VJJ7*	*A0A140VJJ7*
8	*A0A024RA94*	*A0A024RA94*	*A0A024RA94*
9	*Q9H7N4*	*Q9H7N4*	*Q9H7N4*
10	*A0A024R906*	*A0A024R906*	*A0A024R906*
11	*O14944*	*O14944*	*O14944*
12	*A0A024RDS3*	*A0A024RDS3*	*A0A024RDS3*
13	*Q9NXW9*	*Q9NXW9*	*Q9NXW9*
14	*Q61221*	*Q61221*	*Q61221*
15	*Q12018*	*Q12018*	*Q12018*
16	*O15315*	*O15315*	*O15315*
17	*O75150*	*O75150*	*O75150*
18	*P29400*	*P29400*	*P29400*
19	*I3WAC9*	*I3WAC9*	*I3WAC9*
20	*D3DWC4*	*D3DWC4*	*D3DWC4*

Table 3: The twenty highest ranking putative disease proteins of PPI community 34 identified by the MC1, MC2 and MC3 algorithms, respectively.

Rank	MC1	MC2	MC3
1	*Q5JT25*	*Q5JT25*	*Q5JT25*
2	*A5K5E5*	*A5K5E5*	*A5K5E5*
3	*Q24120*	*Q24120*	*Q24120*
4	*Q8VHP6*	*Q8VHP6*	*Q8VHP6*
5	*P17371*	*P17371*	*P17371*
6	*Q9JXV4*	*Q9JXV4*	*Q9JXV4*
7	*P18627*	*P18627*	*P18627*
8	*P29400*	*P29400*	*P29400*
9	*D2KUA6*	*D2KUA6*	*D2KUA6*
10	*Q805P6*	*Q805P6*	*Q805P6*
11	Q9UBL3	A6NMZ7	Q9UBL3
12	A6NMZ7	Q969K3	A6NMZ7
13	Q969K3	Q9UBL3	Q969K3
14	*A8TX70*	*A8TX70*	*A8TX70*
15	*A8K287*	*A8K287*	*A8K287*
16	*A0A024R1D8*	*A0A024R1D8*	*A0A024R1D8*
17	*P06756*	*P06756*	*P06756*
18	*P52757*	*P52757*	*P52757*
19	*P25101*	*P25101*	*P25101*
20	*Q9Z2Q6*	*Q9Z2Q6*	*Q9Z2Q6*

In Table 1, it may be seen that, with respect to PPI community 10, the MC1, MC2 and MC3 algorithms assigned equal ranks to the proteins occupying positions one to three, nine to twelve and fifteen to nineteen.

The proteins in Table 1 which were not assigned perfectly consistent rankings by all three RA algorithms did, however, receive fairly similar rankings, with the largest discrepancy between the ranks assigned to a particular protein appearing in all three RA top 20 lists in Table 1 corresponding to protein Q99697 which was ranked sixth, eighth and sixth, by the MC1, MC2 and MC3 algorithms, respectively. In addition, protein Q92736, ranked twentieth by the MC1 algorithm, does not appear in RA top 20 lists generated by the MC2 or MC3 algorithms (it is ranked twenty fifth and twenty first by the MC2 and MC3 algorithms, respectively).

An even higher degree of consistency was observed for the rankings in the RA top 20 lists corresponding to PPI communities 31 and 34, shown in Tables 2 and 3, respectively. More specifically, the rankings in the RA top 20 lists corresponding to PPI community 31, are perfectly consistent, and the only discrepancies in the RA rankings corresponding to PPI community 34 appear in positions eleven, twelve and thirteen.

The variations in the rankings assigned to a particular protein by the different RA algorithms arise due to the distinct approaches adopted by the MC1, MC2 and MC3 algorithms towards constructing the transition probability matrix, as described in Section 2.2.

The method employed by the MC1 algorithm, for example, encourages transitioning to any protein (state) with a higher ranking in any of the input lists with equal probability, or remaining at the current protein. Consequently, the MC1 algorithm is sensitive to proteins that achieve particularly high mean scores, and therefore ranks, corresponding to any of the five graph-based SSL algorithms.

The MC2 algorithm, on the other hand, encourages transitioning to a protein with a ranking at least as high as the rank of the current protein in at least half of the input lists. Consequently, the MC2 algorithm is capable of prioritising proteins which are often identified as putative disease proteins by the various SSL models, but which do not necessarily achieve exceptionally high rankings in any of the corresponding input lists.

The MC3 algorithm, on the other hand, evaluates the probability of transitioning to a particular protein, from the current protein, as proportional to the number of input lists that rank this new protein higher than the current protein.

The number of discrepancies in the rankings corresponding to the different RA algorithms and a particular PPI community may furthermore yield some insight into the discriminatory abilities exhibited by the underlying SSL models (corresponding to the LGC, GFHF, LapRLS, GGMC and Kernel LP algorithms and a particular PPI community) used to generate the various input lists.

That is, if the individual models' predictions more closely approximate the truth, one would expect a high degree of similarity between the input lists provided to the RA algorithms. Logically, the task of RA becomes more straightforward when the ranks in the input lists provided are fairly consistent and different algorithms are likely to produce similar rankings.

This notion, that different RA algorithms are likely to produce highly consistent rankings if the underlying SSL models exhibit significant discriminatory abilities, is supported by the superior AUC ROC values achieved by the SSL models corresponding to PPI community 31, compared to those achieved for PPI communities 10 and 34. These AUC ROC values are shown in Table 4.

Table 4: The AUC ROC values achieved by the LGC, GFHF, LapRLS, GGMC and Kernel LP models corresponding to PPI communities 10, 31 and 34.

	PPI community		
Algorithm	10	31	34
LGC	0.9568	0.9917	0.7914
GFHF	0.9731	0.9999	0.9833
LapRLS	0.9991	0.9999	0.9857
GGMC	0.9697	0.9999	0.9831
Kernel LP	0.9159	0.9943	0.9839

An apparent contradiction to this notion is, however, observed when comparing the degree of consistency in the rankings generated by the different RA algorithms and the average AUC ROC values achieved by the SSL models corresponding to PPI communities 10 and 34. That is, since the average AUC ROC value achieved by the SSL models corresponding to PPI community 34 is 0.9455 and the average AUC ROC value achieved by the SSL models corresponding to PPI community 10 is 0.9629, one would expect to observe a higher degree of consistency among the aggregated rankings obtained for PPI community 10. From Tables 1 and 3, however, it would appear that a larger proportion of the proteins associated with PPI community 34 are assigned consistent rankings than for PPI community 10.

Upon closer inspection, the AUC ROC values recorded in Table 4 show that the SSL models corresponding to PPI community 34 largely outperform those corresponding to PPI community 10 and that the inferior average AUC ROC value achieved by the models for PPI community 34 is largely due to the unusually poor AUC ROC value of 0.7914 achieved by the LGC model. Indeed, when excluding the worst-performing models for both PPI communities 10 and 34, the average AUC ROC values achieved by the remaining models are 0.9747 and 0.9840, respectively. Hence, this example does not contradict the expected relationship between consistency among the protein rankings obtained by different RA algorithms and the performance levels achieved by the SSL models used to generate the input lists. Rather, it demonstrates the robustness of RA methods with respect to the weaknesses of individual models and emphasises the advantage of employing an RA-based approach rather than ranking disease protein candidates based on the average of their protein scores corresponding to the different SSL algorithms.

Finally, note that, in some instances, the rank assigned to a particular protein in different PPI communities varies significantly. Protein P16278, for example, is ranked first in the RA top 20 list corresponding to PPI community 31 by all three RA algorithms. In the RA top 20 list generated for PPI community 10, however, protein P16278 is ranked fourteenth, thirteenth and thirteenth by the MC1, MC2 and MC3 algorithms, respectively. Similarly, P29400 is ranked eighth in the RA top 20 lists corresponding to PPI community 34, but eighteenth in the RA top 20 lists corresponding to PPI community 31.

These discrepancies may be attributed to the multi-functional nature of proteins which enables them to play different roles in a variety of biological processes. That is, the role played by a protein may differ in distinct biological scenarios. One way in which such scenarios may be distinguished is by interaction partners. In other words, a protein may contribute to the expression of a disease phenotype when interacting with one set of proteins, but have no harmful effects when interacting with a different set of proteins. Hence, if a protein is classified as a disease protein in PPI community *A* but not in PPI community *B*, it may indicate that the protein contributes to a disease phenotype in the biological scenario corresponding to PPI community *A* but not when functioning in the biological scenario corresponding to PPI community *B*. In short, the likelihood that a protein contributes to the expression of a disease phenotype may vary, depending on the proteins with which it interacts and, therefore, the PPI community in which it appears.

4.2. Global rankings

This section is devoted to a description of the results obtained when implementing the second phase of the RA procedure proposed in Section 3.8. The hundred highest ranking putative disease proteins in the full PPI graph retrieved by the MC1, MC2 and MC3 algorithms, are shown in the second, third and fourth columns of Table A.5. Proteins that were assigned the same global ranking by all three RA algorithms are once again italicised.

In the table it may be seen that perfectly consistent rankings across the MC1, MC2 and MC3 algorithms were retrieved for the proteins ranked first to fifth, nineteenth, twenty first and twenty fifth.

Overall, the number of perfectly consistent global rankings is, however, relatively small compared to those observed for the PPI community rankings corresponding to PPI communities 10, 31 and 34 considered in Section 4.1. This finding was to be expected since, during the first stage of the RA procedure, the top-k input lists are all associated with the same PPI community. The input lists utilised during the second phase of the RA procedure, on the other hand, comprise proteins belonging to different PPI communities. Recall that the significance of a given protein's contribution to the expression of a disease phenotype may vary depending on the biological scenario in which it is considered. Consequently, divergent rankings may be assigned to the same protein appearing in different lists of PPI community rankings, as highlighted in Section 4.1 in respect of proteins P16278 and P29400. Nevertheless, the global rankings assigned to a particular protein by the MC1, MC2 and MC3 algorithms are generally notably similar, albeit not necessarily identical.

This observation is echoed by the considerable overlap between the proteins appearing in the lists of global rankings corresponding to the different RA algorithms. More specifically, ninety two of the hundred proteins comprising the list of globally ranked proteins obtained by means of the MC1 algorithm also appear in the list of one hundred globally ranked proteins corresponding to the MC2 algorithm. Likewise, the lists of the hundred globally ranked proteins corresponding to the MC1 and MC3 algorithms share ninety five common proteins. Moreover, the lists of globally ranked proteins returned by means of the MC2 and MC3 algorithms share ninety five common proteins. In total, a hundred and nine unique proteins appear in the union of the three lists of globally ranked proteins.

The discrepancies between the rankings assigned to a particular protein by the different RA algorithms may once again be attributed to the different underpinning mechanisms of the MC1, MC2 and MC3 algorithms.

In this context the MC1 algorithm is sensitive to proteins with high rankings in the PPI community ranking lists (obtained during the first stage of the nested RA procedure), but that do not necessarily appear in a large proportion of these lists.

The global rankings obtained by the MC1 algorithm may, therefore, be pertinent when seeking a putative disease protein belonging to a relatively small number of PPI communities. Genes involved in the expression of autosomal recessive disorders, for example, are typically located at the periphery of the full PPI graph [74]. Intuitively, proteins located at the periphery of the PPI graph are likely to appear in a smaller number of PPI communities, and therefore input lists, than are proteins that occupy more central positions in the full PPI graph.

The MC2 algorithm, on the other hand, favours proteins that appear in a comparatively large proportion of the PPI community ranking lists. One may interpret the presence of a putative disease protein in a large proportion of the PPI community rankings as an indication that the protein is likely to contribute to disease independent of its interaction partners, i.e. the context in which the protein is considered. In other words, a protein belonging to a large number of PPI community ranking lists likely participates in various biological processes, and disruptions in any of these biological processes are likely to result in the expression of a disease phenotype.

The last RA algorithm employed in this framework, the MC3 algorithm, strikes a balance between the two above-mentioned algorithms. This is reflected by the larger number of shared proteins between the lists of global rankings obtained by means of the MC1 and MC3, and MC2 and MC3 algorithms, compared to the ninety two shared proteins appearing in the global rankings corresponding to the MC1 and MC2 algorithms.

The steady state probabilities of the thirty highest ranking proteins in the lists of globally ranked proteins obtained by means of the MC1, MC2 and MC3 algorithms are plotted in Figures 4, 5 and 6, respectively. In addition, each steady state probability is annotated by the UniProt accession number of the corresponding protein and the ten known disease proteins inserted into the initial top-k lists (corresponding to the various SSL algorithms) as verification proteins are indicated in boldface. The ten known disease proteins selected to make up the set of verification proteins are: O14832, P49715, P25101, Q92979, O00170, Q92736, P16278, P24043, P29400 and Q99697.

It may be seen that all ten verification proteins appear among the thirty highest ranking proteins in the full PPI graph corresponding to each of the three RA algorithms. More specifically, for a cut-off rank of 5, 10, 20 and 30, the MC1 algorithm identifies four, seven, eight and ten verification proteins, respectively, whereas both the MC2 and MC3 algorithms identify four, seven, nine and ten verification proteins, respectively. This leads to an appreciable level of confidence in the putative disease proteins retrieved by the respective algorithms.

In order to obtain ranked lists of putative disease genes, the genes responsible for encoding the hundred highest ranking disease protein candidates identified by the three RA algorithms were extracted from the full PPI graph. The HGNC symbols of the predicted disease genes corresponding to the MC1, MC2 and MC3 algorithms are shown in the second, third and fourth columns of Table A.6.

4.3. Validation from the literature

The top hundred highest ranking putative disease proteins in the lists of globally ranked proteins obtained by the MC1, MC2 and MC3 algorithms were validated by finding evidence of their encoding genes' involvement in disease in the literature. To this end, the text-mining tool PubTator was employed in order to retrieve PubMed abstracts related to disease phenotypes and the putative disease genes shown in Table A.6 (excluding the genes corresponding to the verification proteins). Of the ninety nine unique putative disease genes considered in this paper, twenty three could be linked to a disease phenotype via a PubMed abstract ID. The HGNC symbols of the putative disease genes

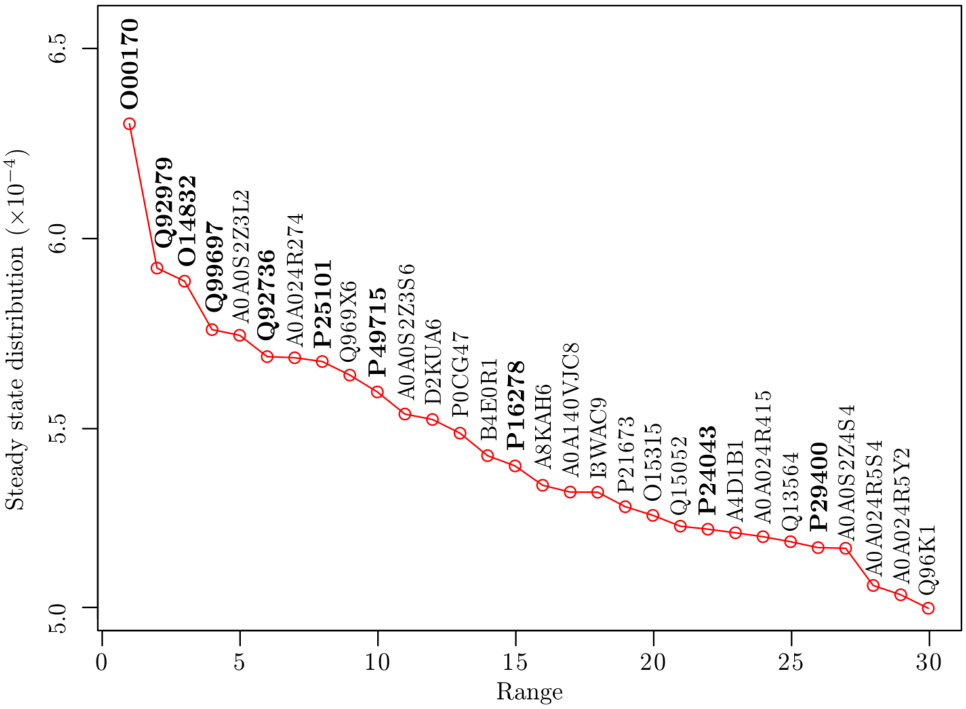

Figure 4: The thirty highest ranking putative disease proteins in the full PPI graph obtained by means of the MC1 algorithm.

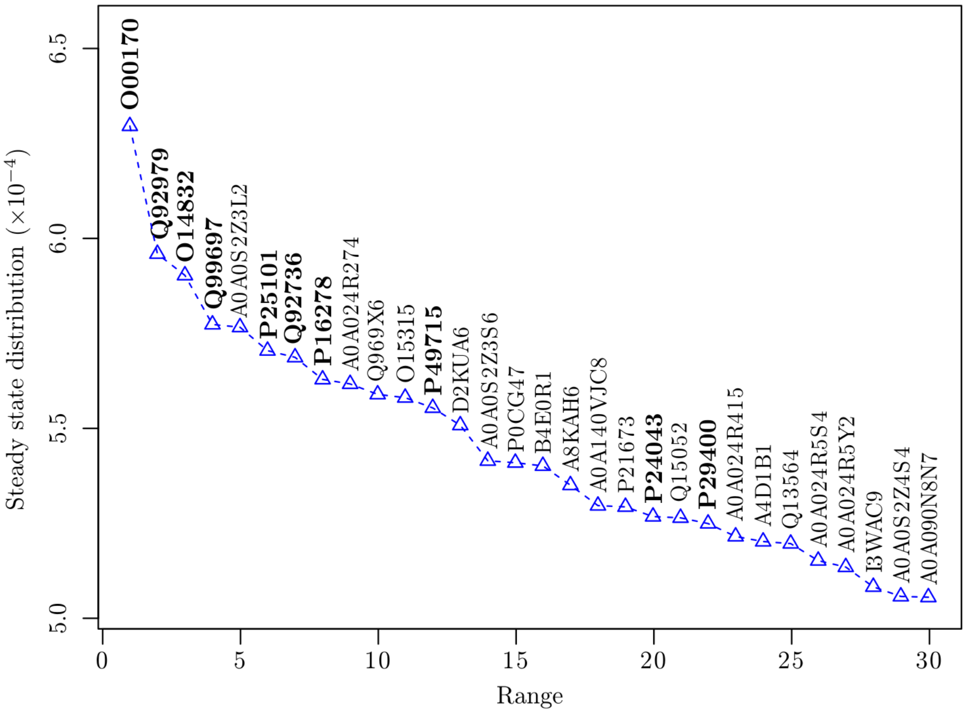

Figure 5: The thirty highest ranking putative disease proteins in the full PPI graph obtained by means of the MC2 algorithm.

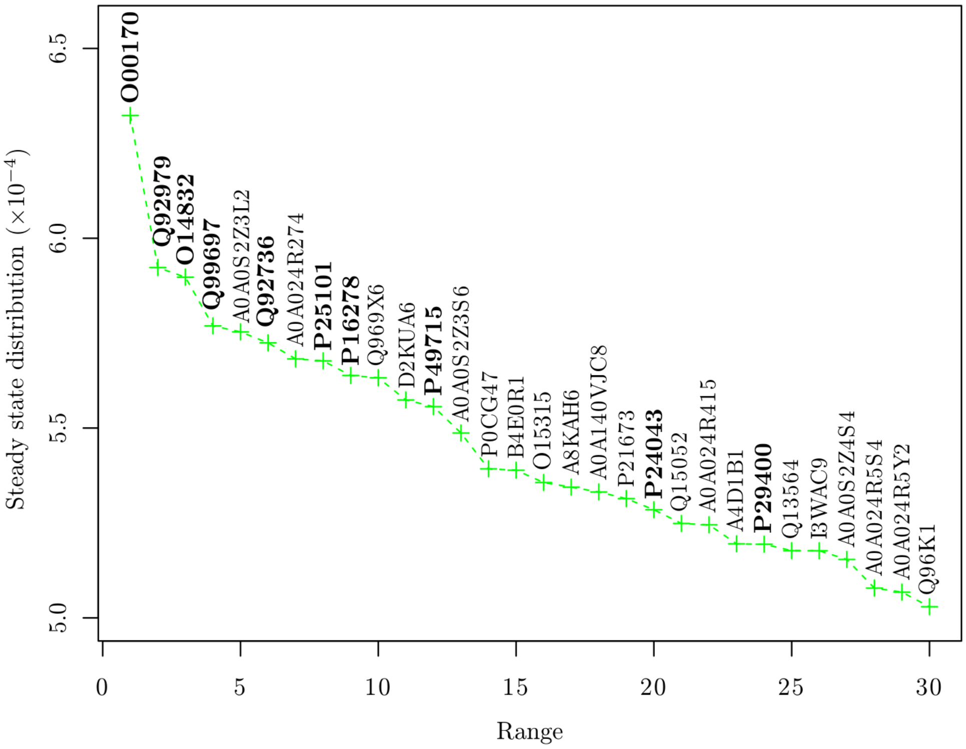

Figure 6: The thirty highest ranking putative disease proteins in the full PPI graph obtained by means of the MC3 algorithm.

which could be validated from the literature are listed in Table B.7. The corresponding disease phenotypes and PubMed abstract IDs are also provided.

Note that, in respect of the majority of novel disease genes identified by a computational disease gene prediction approach, one would not necessarily expect to recover evidence of these genes' involvement in disease from the literature. Intuitively, once a disease gene candidate is associated with a disease phenotype, that gene is typically studied extensively and, if sufficient evidence supporting a link between that gene and the expression of a disease phenotype exists, that gene is added to the various databases of known disease genes. These databases are generally updated on a regular basis. Consequently, the majority of disease genes for which strong experimental evidence of their involvement in disease exists, are already included in the databases of known disease genes. The product proteins of these genes are therefore likely included in the set of labelled disease proteins rather than the set of unlabelled proteins to be classified and subsequently validated.

5. Conclusions

A novel disease gene prioritisation framework capable of leveraging the desirable properties of both low-level data integration strategies and RA methods, whilst simultaneously avoiding the significant drawbacks typically encountered when implementing the respective approaches in a disease gene prioritisation problem setting, was presented in this paper.

Unlike existing disease gene prioritisation methods employing RA, a low-level approach was adopted towards the integration of the heterogeneous proteomic data leveraged by the framework and novel applications of RA in the context of disease gene prioritisation were proposed. More specifically, a two-stage RA procedure was designed and implemented in order to condense multiple ranked lists of disease protein candidates obtained by means of graph-based SSL into one list of globally ranked disease gene candidates.

The high degree of consistency between the ranked lists generated by the MC1, MC2 and MC3 algorithms, the high ranks assigned to each of the known disease proteins included in the verification set and the extraction of evidence from the literature which validates a number of the highest ranking disease gene candidates suggest that promising results may be obtained when utilising the novel applications of RA in the context of disease gene prioritisation explored in this paper.

In future work, the framework may be extended in order to relate the highest ranking disease protein candidates (and their encoding genes) to specific disease phenotypes.

References

[1] E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous, B. S. Pickard, Suspects: Enabling fast and effective prioritization of positional candidates, Bioinformatics 22 (2006) 773–774.

[2] L. R. Cardon, J. I. Bell, Association study designs for complex diseases, Nature Reviews Genetics 2 (2001) 91–99.

[3] K. M. Dipple, E. R. B. McCabe, Modifier genes convert "simple" mendelian disorders to complex traits, Molecular Genetics and Metabolism 71 (2000) 43–50.

[4] N. Risch, K. Merikangas, The future of genetic studies of complex human diseases, Science 273 (1996) 1516–1517.

[5] M. Pertea, A. Shumate, G. Pertea, A. Varabyou, F. P. Breitwieser, Y.-C. Chang, A. K. Madugundu, A. Pandey, S. L. Salzberg, Chess: A new human gene catalog curated from thousands of large-scale rna sequencing experiments reveals extensive transcriptional noise, Genome Biology 19 (2018) 1–14.

[6] S. Vlaic, T. Conrad, C. Tokarski-Schnelle, M. Gustafsson, U. Dahmen, R. Guthke, S. Schuster, Modulediscoverer: Identification of regulatory modules in protein-protein interaction networks, Scientific Reports 8 (2018) 433–443.

[7] S. D. Ghiassian, J. Menche, A.-L. Barabási, A disease module detection (diamond) algorithm derived from a systematic analysis of connectivity patterns of disease proteins in the human interactome, PLoS Computational Biology 11 (2015) e1004120.

[8] Y. Silberberg, M. Kupiec, R. Sharan, Gladiator: a global approach for elucidating disease modules, Genome medicine 9 (2017) 48–61.

[9] X. Wang, X. Wei, B. Thijssen, J. Das, S. M. Lipkin, H. Yu, Three-dimensional reconstruction of protein networks provides insight into human genetic disease, Nature Biotechnology 30 (2012) 159–172.

[10] D. S. Lee, J. Park, K. A. Kay, N. A. Christakis, Z. Oltvai, A. L. Barabási, The implications of human metabolic network topology for disease comorbidity, Proceedings of the National Academy of Sciences 105 (2008) 9880–9885.

[11] H. Yu, P. M. Kim, E. Sprecher, V. Trifonov, M. Gerstein, The importance of bottlenecks in protein networks: Correlation with gene essentiality and expression dynamics, PLoS Computational Biology 3 (2007) 59–72.

[12] T. J. Lopes, M. Schaefer, J. Shoemaker, Y. Matsuoka, J. F. Fontaine, G. Neumann, M. A. Andrade-Navarro, Y. Kawaoka, H. Kitano, Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases, Bioinformatics 27 (2011) 2414–2421.

[13] T. Ideker, N. J. Krogan, Differential network biology, Molecular Systems Biology 8 (2012) 565–580.

[14] Y. Nam, J. H. Jhee, J. Cho, J. H. Lee, H. Shin, Disease gene identification based on generic and disease-specific genome networks, Bioinformatics 35 (2018) 1923–1930.

[15] S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, et al., Gene prioritization through genomic data fusion, Nature Biotechnology 24 (2006) 537–554.

[16] A. Valdeolivas, L. Tichit, C. Navarro, S. Perrin, G. Odelin, N. Levy, et al., Random walk with restart on multiplex and heterogeneous biological networks, Bioinformatics 35 (2018) 497–505.

[17] T. P. Nguyen, T. B. Ho, Detecting disease genes based on semi-supervised learning and protein-protein interaction networks, Artificial Intelligence in Medicine 54 (2012) 63–71.

[18] P. Yang, X. Li, M. Wu, C. K. Kwoh, S. K. Ng, Inferring gene-phenotype associations via global protein complex network propagation, PLoS One 6 (2011) 1–11.

[19] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, R. Sharan, Associating genes and protein complexes with disease via network propagation, PLoS Computational Biology 6 (2010) 1–9.

[20] R. Barshir, O. Shwartz, I. Y. Smoly, E. Yeger-Lotem, Comparative analysis of human tissue interactomes reveals factors leading to tissue-specific manifestation of hereditary diseases, PLoS Computational Biology 10 (2014) 1–12.

[21] M. Li, J. Zhang, Q. Liu, J. Wang, F. X. Wu, Prediction of disease-related genes based on weighted tissue-specific networks by using dna methylation, BMC Medical Genomics 7 (2014) 54–70.

[22] O. Magger, Y. Y. Waldman, E. Ruppin, R. Sharan, Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks, PLoS Computational Biology 8 (2012) 1–10.

[23] X. Yao, H. Hao, Y. Li, S. Li, Modularity-based credible prediction of disease genes and detection of disease subtypes on the phenotype-gene heterogeneous network, BMC Systems Biology 5 (2011) 79–98.

[24] A. Serra, M. Fratello, D. Greco, R. Tagliaferri, Data integration in genomics and systems biology, in: 2016 IEEE Congress on Evolutionary Computation, Vancouver, 2016, pp. 1272–1279.

[25] Y. Li, F.-X. Wu, A. Ngom, A review on machine learning principles for multi-view biological data integration, Briefings in Bioinformatics 19 (2018) 325–340.

[26] S. Huang, K. Chaudhary, L. X. Garmire, More is better: Recent progress in multi-omics data integration methods, Frontiers in Genetics 8 (2017) 84–96.

[27] M. Kim, F. Farnoud, O. Milenkovic, Hydra: Gene prioritization via hybrid distance-score rank aggregation, Bioinformatics 31 (2015) 1034–1043.

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

[29] J. Luo, S. Liang, Prioritization of potential candidate disease genes by topological similarity of protein–protein interaction network and phenotype data, Journal of Biomedical Informatics 53 (2015) 229–236.

[30] X. Chen, G. Yan, W. Ren, J. Qu, Modularized random walk with restart for candidate disease genes prioritization, Systems Biology (Stevenage) 353 (2009) 360–366.

[31] D. Chisanga, S. Keerthikumar, S. Mathivanan, N. Chilamkurti, Integration of heterogeneous 'omics' data using semi-supervised network labelling to identify essential genes in colorectal cancer, Computers and Electrical Engineering 67 (2018) 267–277.

[32] M. Krauthammer, C. A. Kaufmann, T. C. Gilliam, A. Rzhetsky, Molecular triangulation: Bridging linkage and molecular-network information for identifying candidate genes in alzheimer's disease, National Academy of Sciences 101 (2004) 15148–15153.

[33] S. Picart-Armada, S. J. Barrett, D. R. Wille, A. Perera-Lluna, A. Gutteridge, B. H. Dessailly, Benchmarking network propagation methods for disease gene identification, PLoS Computational Biology 15 (2018) 1–12.

[34] P. Radivojac, K. Peng, W. T. Clark, B. J. Peters, A. Mohan, S. M. Boyle, et al., An integrated approach to inferring gene–disease associations in humans, Proteins: Structure, Function, and Bioinformatics 72 (2008) 1030–1037.

[35] M. Ruffalo, M. Koyutürk, R. Sharan, Network-based integration of disparate omic data to identify "silent players" in cancer, PLoS Computational Biology 11 (2015) 1–20.

[36] X. Wu, R. Jiang, M. Q. Zhang, S. Li, Network-based global inference of human disease genes, Molecular Systems Biology 4 (2008) 889–907.

[37] A.-L. Boulesteix, M. Slawski, Stability and aggregation of ranked gene lists, Briefings in Bioinformatics 10 (2009) 556–568.

[38] E. M. Conlon, J. J. Song, A. Liu, Bayesian meta-analysis models for microarray data: A comparative study, BMC Bioinformatics 8 (2007) 80–101.

[39] A. C. Tan, D. Q. Naiman, L. Xu, R. L. Winslow, D. Geman, Simple decision rules for classifying human cancers from gene expression profiles, Bioinformatics 21 (2005) 3896–3904.

[40] I. Fishel, A. Kaufman, E. Ruppin, Meta-analysis of gene expression data: A predictor-based approach, Bioinformatics 23 (2007) 1599–1606.

[41] T. A. Peterson, D. Park, M. G. Kann, A protein domain-centric approach for the comparative analysis of human and yeast phenotypically relevant mutations, BMC Genomics 14 (2013) 5–17.

[42] G. Van Zyl, J. H. Van Vuuren, Graph-based semi-supervised learning for the detective of putative disease genes, 2020. Manuscript submitted for publication (under review).

[43] S. Lin, Rank aggregation methods, Wiley Interdisciplinary Reviews: Computational Statistics 2 (2010) 555–570.

[44] X. Li, X. Wang, G. Xiao, A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications, Briefings in Bioinformatics 20 (2019) 178–189.

[45] M. E. Renda, U. Straccia, Web metasearch: Rank vs. score based rank aggregation methods, in: ACM Symposium on Applied Computing (SAC), New York (NY), 2003, pp. 841–846.

[46] S. Lin, Space oriented rank-based data integration, Statistical Applications in Genetics and Molecular Biology 9 (2010) 1–23.

[47] S. Razick, G. Magklaras, I. M. Donaldson, irefindex: A consolidated protein interaction database with provenance, BMC Bioinformatics 9 (2008) 405–417.

[48] McKusick-Nathans Institute of Genetic Medicine John Hopkins University, Online Mendelian Inheritance in Man, OMIM, [Online], [Cited April 2017], Available at `https://omim.org/`, 2015.

[49] J. Piñero, Á. Bravo, N. Queralt-Rosinach, A. Gutiérrez-Sacristán, J. Deu-Pons, E. Centeno, et al., Disgenet: A comprehensive platform integrating information on human disease-associated genes and variants, Nucleic Acids Research (2016) 833–839.

[50] S. El-Gebali, J. Mistry, A. Bateman, S. R. Eddy, A. Luciani, S. C. Potter, et al., The Pfam protein families database in 2019, Nucleic Acids

Research 47 (2018) 427–432.

[51] A. Fabregat, S. Jupe, L. Matthews, K. Sidiropoulos, M. Gillespie, P. Garapati, et al., The reactome pathway knowledgebase, Nucleic Acids Research 46 (2017) 649–655.

[52] M. Giurgiu, J. Reinhard, B. Brauner, I. Dunger-Kaltenbach, G. Fobo, G. Frishman, et al., CORUM: The comprehensive resource of mammalian protein complexes — 2019, Nucleic Acids Research 47 (2018) 559–563.

[53] M. A. Van Driel, J. Bruggeman, G. Vriend, H. G. Brunner, J. A. Leunissen, A text-mining analysis of the human phenome, European Journal of Human Genetics 14 (2006) 535–542.

[54] S. White, P. Smyth, A spectral clustering approach to finding communities in graphs, in: Proceedings of the SIAM International Conference on Data Mining, Newport Beach (CA), 2005, pp. 12–18.

[55] M. Oti, H. G. Brunner, The modular nature of genetic diseases, Clinical Genetics 71 (2007) 1–11.

[56] A. Bauer-Mehren, M. Bundschus, M. Rautschka, M. A. Mayer, F. Sanz, L. I. Furlong, Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases, PLoS One 6 (2011) 20284–20295.

[57] J. Freudenberg, P. Propping, A similarity-based method for genome-wide prediction of disease-relevant human genes, Bioinformatics 18 (2002) 110–115.

[58] S. Jones, X. Zhang, D. W. Parsons, J. C. H. Lin, R. J. Leary, P. Angenendt, et al., Core signaling pathways in human pancreatic cancers revealed by global genomic analyses, Science 321 (2008) 1801–1806.

[59] J. Lim, T. Hao, C. Shaw, A. J. Patel, G. Szabó, J. F. Rual, et al., A protein–protein interaction network for human inherited ataxias and disorders of purkinje cell degeneration, Cell 125 (2006) 801–814.

[60] N. A. Zaghloul, N. Katsanis, Functional modules, mutational load and human genetic disease, Trends in Genetics 26 (2010) 168–176.

[61] X. Wu, L. Zhao, L. Akoglu, A quest for structure: Jointly learning the graph structure and semi-supervised classification, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, ACM, Turin, 2018, pp. 87–96.

[62] A. D. D'Andrea, M. Grompe, The fanconi anaemia/brca pathway, Nature Reviews Cancer 3 (2003) 23–37.

[63] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, et al., Dynamic modularity in protein interaction networks predicts breast cancer outcome, Nature Biotechnology 27 (2009) 199–207.

[64] M. Chen, M. W. Deem, Hierarchy of gene expression data is predictive of future breast cancer outcome, Physical Biology 10 (2013) 6–56.

[65] D. He, Z. P. Liu, L. Chen, Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach, BMC Genomics 12 (2011) 592–599.

[66] S. Tripathi, M. W. Deem, Hierarchy in gene expression is predictive of risk, progression, and outcome in adult acute myeloid leukemia, Physical Biology 12 (2015) 16016–16089.

[67] F. Ye, D. Jia, M. Lu, H. Levine, M. W. Deem, Modularity of the metabolic gene network as a prognostic biomarker for hepatocellular carcinoma, Oncotarget 9 (2018) 15015–15026.

[68] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, B. Schölkopf, Learning with local and global consistency, in: Proceedings of Advances in Neural Information Processing Systems, Vancouver, 2004, pp. 321–328.

[69] X. Zhu, Z. Ghahramani, J. D. Lafferty, Semi-supervised learning using gaussian fields and harmonic functions, in: Proceedings of the 20th International conference on Machine learning (ICML-03), Washington (DC), 2003, pp. 912–919.

[70] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: A geometric framework for learning from labeled and unlabeled examples, Journal of Machine Learning Research 7 (2006) 2399–2434.

[71] J. Wang, T. Jebara, S. F. Chang, Semi-supervised learning using greedy max-cut, Journal of Machine Learning Research 14 (2013) 771–800.

[72] Z. Zhang, L. Jia, M. Zhao, G. Liu, M. Wang, S. Yan, Kernel-induced label propagation by mapping for semi-supervised classification, IEEE Transactions on Big Data 5 (2018) 148–165.

[73] M. Schimek, E. Budinska, K. Kugler, V. Svendova, J. Ding, S. Lin, TopKLists: A comprehensive r package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists, Statistical Applications in Genetics and Molecular Biology 14 (2015) 311–316. URL: http://www.degruyter.com/doi/10.1515/sagmb-2014-0093.

[74] K. I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, A. L. Barabási, The human disease network, Proceedings of the National Academy of Sciences 104 (2007) 8685–8690.

Appendix A. Global rankings tables

Appendix B. Evidence from the literature supporting the putative disease genes identified

Table A.5: The hundred highest ranking putative disease proteins in the full PPI graph identified by the MC1, MC2 and MC3 algorithms.

Rank	MC1	MC2	MC3
1	*O00170*	*O00170*	*O00170*
2	*Q92979*	*Q92979*	*Q92979*
3	*O14832*	*O14832*	*O14832*
4	*Q99697*	*Q99697*	*Q99697*
5	*A0A0S2Z3L2*	*A0A0S2Z3L2*	*A0A0S2Z3L2*
6	Q92736	P25101	Q92736
7	A0A024R274	Q92736	A0A024R274
8	P25101	P16278	P25101
9	Q969X6	A0A024R274	P16278
10	P49715	Q969X6	Q969X6
11	A0A0S2Z3S6	O15315	D2KUA6
12	D2KUA6	P49715	P49715
13	P0CG47	D2KUA6	A0A0S2Z3S6
14	B4E0R1	A0A0S2Z3S6	P0CG47
15	P16278	P0CG47	B4E0R1
16	A8KAH6	B4E0R1	O15315
17	A0A140VJC8	A8KAH6	A8KAH6
18	I3WAC9	A0A140VJC8	A0A140VJC8
19	*P21673*	*P21673*	*P21673*
20	O15315	P24043	P24043
21	*Q15052*	*Q15052*	*Q15052*
22	P24043	P29400	A0A024R415
23	A4D1B1	A0A024R415	A4D1B1
24	A0A024R415	A4D1B1	P29400
25	*Q13564*	*Q13564*	*Q13564*
26	P29400	A0A024R5S4	I3WAC9
27	A0A0S2Z4S4	A0A024R5Y2	A0A0S2Z4S4
28	A0A024R5S4	I3WAC9	A0A024R5S4
29	A0A024R5Y2	A0A0S2Z4S4	A0A024R5Y2
30	Q96IK1	A0A090N8N7	Q96IK1
31	A0A024QZN4	Q96IK1	B2RCZ7
32	A0A024R3C5	A0A024R1Q5	A0A024QZN4
33	B2RCZ7	B2RCZ7	A0A024R3C5
34	A0A024RBC0	A0A024R3C5	A0A024RBC0
35	A0A0S2Z4D1	A0A024RBC0	A0A0S2Z4D1
36	D9IAI1	A0A024QZN4	A0A090N8N7
37	A0A024R7M6	Q6NTF7	A0A024R7M6
38	P0CK57	A0A024R7M6	D9IAI1
39	P20333	P0CK57	A0A024R1Q5
40	A0A024R104	D9IAI1	P0CK57
41	B4DFJ1	B4DFJ1	P20333
42	Q99571	P20333	B4DFJ1
43	P05556	A0A024R104	A0A024R104
44	A0A024R794	A0A140VK84	Q99571
45	P11501	P05556	P05556
46	A0A090N8N7	Q99571	P11501
47	Q13029	P11501	A0A024QZY5
48	A0A140VJP2	A0A024QZY5	A0A140VJP2
49	A0A024QZU0	B4DJ51	A0A024R794
50	Q5BJQ6	A0A024R794	Q13029

Table A.5 (cont.): The hundred highest ranking putative disease proteins in the full PPI graph identified by the MC1, MC2 and MC3 algorithms.

Rank	MC1	MC2	MC3
51	A0A024QZY5	Q13029	A0A024QZU0
52	A0A024RDS3	A0A024QZA9	Q96IK1
53	A0A024R3M4	A0A024QZU0	Q5BJQ6
54	E9KL48	A0A024RDS3	Q6NTF7
55	A0A024R1Q5	Q5BJQ6	A0A024RDS3
56	F6M2K3	A0A140VJP2	O00767
57	A0A024R2F2	F6M2K3	A0A140VK84
58	P25454	E9KL48	P79114
59	O14944	A0A0S2Z4D1	F6M2K3
60	P18627	A0A024R2F2	E9KL48
61	P79114	P18627	A0A140VJK4
62	A0A140VJK4	O14944	Q8NEY8
63	Q8NEY8	A0A140VJK4	A0A024R2F2
64	A0A024RBH0	B4DQY3	O14944
65	A0A024R5F9	P79114	P25454
66	A0A0E3SU01	Q5XUX1	A0A024RBH0
67	A0A024QZA9	A0A024RBH0	A0A024R3M4
68	Q5XUX1	E9KL35	P18627
69	A0A024R7G8	A0A0E3SU01	E9KL35
70	E9KL35	O00767	Q5XUX1
71	P27313	A0A024R7G8	A0A0E3SU01
72	Q9NUI1	P25454	B4DJ51
73	A0A024R9X3	A0A024R3M4	A0A024R7G8
74	Q96IK1	P27313	Q9NUI1
75	O75056	Q9NUI1	A0A024R5F9
76	B4DJ51	A0A024R9X3	P27313
77	A0A140VK84	A0A024R5M8	A0A024R9X3
78	P28028	A0A024QZT0	A0A024QZQ1
79	A0A024QZQ1	Q8NEY8	O75056
80	A0A024QYW3	O75056	O75223
81	Q12018	Q96IK1	A0A024QZT0
82	O75223	A0A024QZQ1	P28028
83	Q6NTF7	Q9HC38	A0A024QYW3
84	Q9HC38	P28028	Q12018
85	A0A0S2Z3G9	A0A024QYW3	Q08297
86	Q08297	Q12018	P06241
87	A0A024RCS9	O75223	Q9HC38
88	P06241	A0A024RCS9	A0A0S2Z3G9
89	Q8IZ13	A0A024R5F9	A0A024RCS9
90	P30050	P06241	A0A024QZA9
91	A0A0K0K1J1	A0A0S2Z3G9	P30050
92	A0A024R0T0	P30050	B3VK56
93	B3VK56	P35803	Q6ZUT1
94	Q6ZUT1	Q8WZ73	A0A024R0T0
95	A0A024R1Q8	Q643R0	J9JIF5
96	Q99576	B3VK56	A0A0K0K1J1
97	Q643R0	B2RC27	P35803
98	A0A024QZF2	Q6ZUT1	A0A024QZF2
99	J3KNF8	J9JIF5	A0A024R1Q8
100	B2R9L7	A0A024R1Q8	A0A024R1L9

Table A.6: The encoding genes of the hundred highest ranking putative disease proteins in the full PPI graph obtained by means of the MC1, MC2 and MC3 algorithms.

Table A.6 (cont.): The encoding genes of the hundred highest ranking putative disease proteins in the full PPI graph obtained by means of the MC1, MC2 and MC3 algorithms.

Rank	MC1	MC2	MC3
1	AIP	AIP	AIP
2	EMG1	EMG1	EMG1
3	PHYH	PHYH	PHYH
4	PITX2	PITX2	PITX2
5	ATP2A2	ATP2A2	ATP2A2
6	RYR2	EDNRA	RYR2
7	SMAD4	RYR2	SMAD4
8	EDNRA	GLB1	EDNRA
9	UTP4	SMAD4	GLB1
10	CEBPA	UTP4	UTP4
11	CYBB	RAD51B	PPARG
12	PPARG	CEBPA	CEBPA
13	UBB	PPARG	CYBB
14	ITGB2	CYBB	UBB
15	GLB1	UBB	ITGB2
16	HSPB2	ITGB2	RAD51B
17	APP	HSPB2	HSPB2
18	INS	APP	APP
19	SAT1	SAT1	SAT1
20	RAD51B	LAMA2	LAMA2
21	ARHGEF6	ARHGEF6	ARHGEF6
22	LAMA2	COL4A5	PNKD
23	CD36	PNKD	CD36
24	PNKD	CD36	COL4A5
25	NAE1	NAE1	NAE1
26	COL4A5	USP8	INS
27	TLR2	MAP2K5	TLR2
28	USP8	INS	USP8
29	MAP2K5	TLR2	MAP2K5
30	BOD1	PPP1R17	BOD1
31	VCL	BOD1	ETHE1
32	DRD2	NAGA	VCL
33	ETHE1	ETHE1	DRD2
34	KITLG	DRD2	KITLG
35	STK11	KITLG	STK11
36	PEBP1	VCL	PPP1R17
37	GATAD2A	APOBEC3H	GATAD2A
38	BDLF4	GATAD2A	PEBP1
39	TNFRSF1B	BDLF4	NAGA
40	CNTN1	PEBP1	BDLF4
41	TINF2	TINF2	TNFRSF1B
42	P2RX4	TNFRSF1B	TINF2
43	ITGB1	CNTN1	CNTN1
44	TNPO3	FN3KRP	P2RX4
45	HSP90AA1	ITGB1	ITGB1
46	PPP1R17	P2RX4	HSP90AA1
47	PRDM2	HSP90AA1	PRPF4B
48	MAT2B	PRPF4B	MAT2B
49	RIPK1	CALM1	TNPO3
50	CSTF1	TNPO3	PRDM2
51	PRPF4B	PRDM2	RIPK1
52	ZMYM2	BCKDK	BOD1
53	TIRAP	RIPK1	CSTF1
54	GLUD1	ZMYM2	APOBEC3H
55	NAGA	CSTF1	ZMYM2
56	NCOA6	MAT2B	SCD
57	VHL	NCOA6	FN3KRP
58	RAD51B	GLUD1	MYO10
59	EREG	STK11	NCOA6
60	LAG3	VHL	GLUD1
61	MYO10	LAG3	GSTA1
62	GSTA1	EREG	PPHLN1
63	PPHLN1	GSTA1	VHL
64	CDK17	NDUFAF7	EREG
65	B4GAT1	MYO10	RAD51B
66	BDNF	FBXW9	CDK17
67	BCKDK	CDK17	TIRAP
68	FBXW9	RACK1	LAG3
69	RAD23A	BDNF	RACK1
70	RACK1	SCD	FBXW9
71	PUUVSSGP1	RAD23A	BDNF
72	DECR2	RAD51B	CALM1
73	CNGA1	TIRAP	RAD23A
74	BOD1	PUUVSSGP1	DECR2
75	SDC3	DECR2	B4GAT1
76	CALM1	CNGA1	PUUVSSGP1
77	FN3KRP	POLD3	CNGA1
78	BRAF	VDAC2	SIRT1
79	SIRT1	PPHLN1	SDC3
80	PLP2	SDC3	GGCT
81	CDC53	BOD1	VDAC2
82	GGCT	SIRT1	BRAF
83	APOBEC3H	GLOD4	PLP2
84	GLOD4	BRAF	CDC53
85	ACTN4	PLP2	RAD51B
86	RAD51B	CDC53	FYN
87	RGL2	GGCT	GLOD4
88	FYN	RGL2	ACTN4
89	ZBED8	B4GAT1	RGL2
90	RPL12	FYN	BCKDK
91	CST3	ACTN4	RPL12
92	ZNF576	RPL12	NOS1
93	NOS1	GPM6B	NKAPD1
94	NKAPD1	RFFL	ZNF576
95	RPL23	TPX2	CXXC4
96	TSC22D3	NOS1	CST3
97	TPX2	RSPO3	GPM6B
98	RRAS	NKAPD1	RRAS
99	CYB5B	CXXC4	RPL23
100	OXTR	RPL23	CBY1

Table B.7: Putative disease genes and the associated disease phenotypes for which PubMed IDs were obtained.

Gene	Phenotype
UTP4	Cirrhosis [10820129, 12417987, 24147052, 11045837, 15768832, 20385600, 16225863], Intrahepatic Cholestasis [16225863], Malignant Neoplasm of Colon and/or Rectum [28350096]
UBB	Diabetes Insipidus [11124436, 10911966], Neuroblastoma [11726544], Alzheimer's Disease [11980917, 16213790, 15198995, 16432153, 10674623, 10911966, 17531157, 12200043, 31654319, 15519283, 22797007, 12893422, 22730000, 11124436, 14597671, 17237936, 18760506, 17052186, 10666673, 12055595, 11726544, 21059367, 27240056], Nervous System Disorder [12007022], Progressive Supranuclear Palsy [12871580, 10674623, 15050720], Dementia [12893422], Argyrophilic Grain Disease, Lewy Body Disease, Pick Disease of the Brain [14597671], Machado-Joseph Disease [15198995], Tauopathies [15198995, 17052186, 21762696, 17405812, 18760506, 14597671, 21059367], Inclusion Body Myositis [15452314, 17052186], Down Syndrome [16432153, 12055595, 10666673], Neurodegenerative Disorders [16714280, 21762696, 10674623, 14597671, 27966098], Huntington's Disease [17052186, 15198995, 18760506, 16226348], Multiple System Atrophy [17237936], Myopathy [17405812,17931355, 17931355], Cleft Palate [17468296], Memory Impairment [21059367], Parkinson's Disease [21731658, 17052186, 16213790], Bronchopneumonia [22730000], Malignant Neoplasm of Cervix [24367661], Malignant Neoplasm of Lung [25820571], Malignant Neoplasm of Ovary [29130934], Malignant Neoplasms [29130934, 31096896, 24367661, 29130938, 22057347, 29130934], Exfoliation Syndrome [30738879], Malignant Neoplasm of Stomach [31081222]
SAT1	Rheumatoid Arthritis [10852253, 24578214], Malignant Neoplasm of Breast [10978316, 19727732, 19727732, 16207710], Tick-Borne Encephalitis, Venezuelan Equine Encephalomyelitis [12083816], Keratosis Follicularis Spinulosa Decalvans [12215835, 24313295], Malignant Neoplasm of Lung [1327507, 9717831, 17711504, 16757480, 9717831, 1327507, 17711504, 12697027], Malignant Neoplasm of Colon and/or Rectum [14506281, 9052989, 17237273, 18430370, 30555259, 16262603, 16854216, 15799689, 16262603, 17237273], Iron Deficiency [15845361], Malignant Neoplasm of Ovary [15905201], Mental Disorder [16389195, 25959060, 23958961, 19051286, 27229768], Urinary Tract Infection [16903846], Dermatitis [17374397], Adenocarcinoma [17711504], Parvoviridae [18753240], Malignant Neoplasm of Prostate [18974881, 10646846], Depressive Disorder [19152344, 25959060, 19851986], Multiple Myeloma [19531770, 19555670], Diabetes [19763711, 31739742, 22705322, 19615702, 30913011, 31383894, 28413567], Pterygium [19956562], Parkinson's Disease [20837543], Lymphoma [20930513, 19531770], Hyperoxaluria [21093948], Substance Abuse, Depressive Disorders [21152090], Anxiety Disorders [21152090, 23958961, 18759322], Pancreatitis [21318871], Glioblastoma [22179681, 25277523, 25277523, 22179681], Adiposis Dolorosa [22300160], Carcinogenesis [22579641, 16010435, 16262603, 25277523], Squamous Cell Carcinoma [27263493, 22718136], Malignant Neoplasms [22718136, 10646846, 30214636, 19531770], Arteriosclerosis [22761463], Keratosis Pilaris [22816986], Behcet Syndrome [23768751], Bipolar Disorder [23768751, 23958961, 27229768], Burkitt Lymphoma [23891576], Tumor Cell Invasion [23986438, 27901475, 29926315], Lipodystrophy [24129368], Ichthyosis Follicularis Atrichia Photophobia Syndrome, Olmsted Syndrome [24313295], Substance Abuse [24735382, 27191827], Brain Neoplasms [25277523, 31399646], Depressive Disorder [25959060, 19152344], Lymphoma, Pneumonia [27755248, 28520818], Congestive Heart Failure [28035653], Malignant Neoplasm of Esophagus [28243981], Autism Spectrum Disorders [28695149], Asthma [29729200], Malignant Neoplasm of Stomach [29926315, 19686286, 19686286, 28808875], Neoplasm Metastasis [29926315, 25893668, 27901475], Schizophrenia [29958750, 19162121], Ankylosing Spondylitis, Hyperostosis [30031587], Metabolic Syndrome [30031587, 17490981], Zika Virus Infection [30040423], Crohn's Disease [30215805],Ablepharon-Macrostomia Syndrome [30278484], Tuberculosis [30295760, 22205804, 26825909, 28779754, 24824246, 30652492, 22205804], Anhedonia [30326340], Sarcoidosis [30652492], Anophthalmia and Pulmonary Hypoplasia, Hyperglycemia [30677401], Foot-and-Mouth Disease [30807778, 28224715, 28298597], Oropharyngeal Disorders [31092573], Gestational Trophoblastic Neoplasms [31219560], Cardiovascular Diseases [31739742, 28403191], Metabolic Diseases [31779686],

Table B.7 (cont.): Putative disease genes and the associated disease phenotypes for which PubMed IDs were obtained.

Gene	Phenotype
	Obesity [31462690, 20882379, 29715464, 31440954, 28231762, 18940394, 22134720], Malignant Neoplasm of Liver [9397163, 22579641, 30100754, 22579641, 9397163], Neoplasms [9717831, 27012811, 17711504, 22718136, 25277523, 17987291, 27698118, 10646846, 18974881, 31399646, 17096347, 9397163, 29957732]
ARHGEF6	Pulmonary Hypertension [15242552], Glioblastoma [16320026], Mental Retardation [17304053, 11017088, 19377476, 21989057, 11337747], Inflammatory Bowel Diseases [25479423, 28333213], Schizophrenia [29759351]
NAE1	Complete Trisomy 21 Syndrome [15192323, 17611268], Malignant Neoplasm of Liver [15201980, 31817100, 29846044, 31002342], Alzheimer's Disease [17611268], Down Syndrome [17611268, 15192323], Myocardial Infarction [21386696], Oropharyngeal Disorders [22895816], Anophthalmia and Pulmonary Hypoplasia [28535453], Melanoma [29233905], Congestive Heart Failure [29632206], Lymphoma [29910671], Malignant Neoplasm of Pancreas [31404297]
TNFRSF1B	Encephalomyelitis [10679120, 30498033, 31785393, 28052249, 31220564], Lupus [10703622, 30185417, 15674653, 17028114, 11196716, 10395102, 11600223, 11607787, 16306881, 11197692, 12739039, 11169260, 10643707, 11762942, 19684152, 20516030, 12739039], Hypertensive Disease [10942422, 16216983, 16003175, 23337087, 19557004], Diabetes [10946317, 18566344, 22068019, 25200302, 31073629, 31813103, 28843039, 26786322, 9174153, 30333553, 15787661, 11315843, 28367848, 21849023, 16979382, 11882518, 15217754], Hyperlipidemia [10958645, 16054550], Narcolepsy [11144293, 12601524, 11285131], Takayasu Arteritis, Vasculitis [11169260], Rheumatoid Arthritis [11212177, 12730509, 31609517, 16142859, 30723161, 12610797, 11508576, 15252214, 12209506, 25850964, 20401725, 16871413, 12233877, 26071216, 24777778, 1320571, 29748156, 12858434, 10765919, 25311255, 29889832, 14872483, 16277675, 12209507, 18309487, 28150360, 18565259, 15603867, 14687710, 29760711, 15022314, 12913922], Sciatic Neuropathy [11240015], Neuropathy [11315843], Hypertensive Disease [11315843, 15925743, 19557004], Cerebral Infarction [11358448], Idiopathic Pulmonary Fibrosis [11371414, 21144722], Central Visual Impairment [11472434], Trichohepatoenteric Syndrome [11588035], Encephalitis [11607787, 11169260], Ovarian Cancer [11705863], Neoplasm Metastasis [11705863, 25010932, 18764880, 17143529, 15342406, 20646319], Fatty Liver Disease [11732005], Epithelial Hyperplasia [11781288], Colitis [11781288, 11904678, 15274667, 30457980, 15842589, 25387791], Arterial Occlusive Diseases [11854734], Crohn's Disease [11904678, 12049175, 18248655, 11196680, 15784704, 11781288, 29848778, 24121042, 26071216, 15842589], Psoriasis [12011375, 22111980, 26071216, 25537528], Arthritis [12011375, 16947419, 29375132, 15252214, 29339199, 29342501, 21346237, 18415772, 29218573, 29618659, 29339199], Pigmentary Disorder [12032115], Lymphoma [12149223, 12149223, 16173964, 16198418, 23672298, 29928327, 29930163, 29930163, 8445947, 24757092, 29922294], Polycystic Ovary Syndrome [12161545], Hyperandrogenism [12161545, 19039234], Septicemia [12500222, 15526005, 23029405], Pancreatitis [12741461], Behcet Syndrome [12770792, 18415772], Degenerative Polyarthritis [12913922, 16871413, 11508576, 1320571, 30826358, 31261789], Malaria [14504653, 23408847], Adenocarcinoma [14613990, 15146559], Myelodysplastic Syndrome [14728878], Pleural Effusion Disorder [14997042], Hypoalbuminemia [15044820], Mesothelioma [15084380], Periodontitis [15142217, 21593780], Endometriosis [15212671, 19238748, 18510047, 14506926, 24844917], Malignant Neoplasm of Colon and/or Rectum [15743036, 16477629, 28123565, 24762198, 27504605, 28412748, 18088549, 21994466, 20566746, 21994466, 18088549, 24762198, 28412748, 16477629], Dystrophia Myotonica 2 [15787661], Neurodegenerative Disorders [15827736, 29604364], Cerebral Ischemia [15829914], Anterior Uveitis [15851552], Eating Disorders [15866349], Malignant Neoplasm of Prostate [15948150, 17143529, 17143529, 15948150], Neurilemmoma [16001231], Limb Ischemia [16095891, 23065828], Malignant Neoplasm of Liver [16109524], Anemia [16142859], Neuroblastoma [16215672, 9136990, 19142969, 23314735, 24253178, 20404560, 8656283, 19142969, 16215672, 14654552, 9136990], Malignant Neoplasm of Liver [16477629, 31848339],

Table B.7 (cont.): Putative disease genes and the associated disease phenotypes for which PubMed IDs were obtained.

Gene	Phenotype
	Obesity [16503147, 11782876, 18685868, 12161545, 10841005, 16216983, 16979382], Retinoblastoma [16555252, 17973327, 16555252], Metabolic Syndrome [16645020, 10942422, 16979382], Impaired Glucose Tolerance [16732051, 16979382], Osteoarthritis [16871413, 16282562], Osteoporosis [17002564, 16502120, 15071724], Myeloid Leukemia [17018605, 18403643], Intestinal Diseases [17030185, 15274667], Incontinentia Pigmenti Achromians, Invasive Pulmonary Aspergillosis [17207711], Medulloblastoma [17550129], Chordoma [17700442], Palindromic Rheumatism [17763205], Common Variable Immunodeficiency [17825894], Myocardial Infarction [17852784, 21362018, 29547707], Osteoporotic Fractures [18038243, 15071724], Depressive Disorder [18081157, 21938001, 26278479, 28761093, 28761093, 28839357, 17094069, 16458261, 24047966, 28867280, 24094876, 28867280, 28839357], Malignant Neoplasm of Lip and/or oral Cavity [18206417], Epiretinal Membrane [18398367, 12032115], Myeloproliferative Disease [18403643, 17018605], Motor Axonal Neuropathy [18717726], Neoplasms [31555271, 11705863, 29892300, 30685298, 14997042, 28560678, 28789455, 9685865, 15084380, 28096513, 29032004, 29295954, 11389056, 19142969, 19738070, 30628200, 25010932, 17973327, 18974393, 30127886, 29920311, 31040894, 28819854, 27888699, 20224295, 16555252, 23576602, 15342406, 27626702, 29234328, 20646319, 17143529, 30356161, 29802567, 29623079, 21995493, 29435163, 27464624, 24511008, 18088549], Malignant Neoplasm of Lung [31559061, 30989732, 29080421, 31848339], Spondylarthritis [31609517, 29618659], CNS Disorder [31785393], Malignant Neoplasm of Breast [31839752], Melanoma [8908597, 12789270, 8305739, 21418516]
P2RX4	Epilepsy [12941474, 12941474, 19084381, 19084381], Cystic Fibrosis [14701827], Muscular Dystrophy [15006691], Tuberculosis [17034577], Anxiety Disorders [17197037], Chlamydia Infections [17785807], Bronchopulmonary Dysplasia, Borderline Personality Disorder [18543274, 23602648], Schizophrenia [18614336], Bipolar Disorder [18614336, 18543274], Hypertensive Disease [18852390], Uncinate Epilepsy [19084381], Amyloidosis [19562525], Lupus, Rheumatoid Arthritis [20493226], Neuroblastoma [21081501], Hepatitis C [21899776], Degenerative Polyarthritis [22715356, 24145861], Macular Degeneration [23303206], Narcolepsy [23497937, 21170044, 23725858], Major Depressive Disorder [23602648, 18543274], Depressive Disorder [23602648, 31656696, 28751018, 28751018, 23602648], Amyotrophic Lateral Sclerosis [23771221, 17990272, 20084016], Febrile Convulsions [24703484], Septicemia [25318479, 29875325], Neoplasms [26505137, 30209547, 21157381, 30006588], GVH Disease [26538394], Chronic Obstructive Airway Disease [26541524], Dengue Fever [26969484], Arteriosclerosis [27355755], Eosinophilic Disorder [27863396], Hepatitis [27940204], Mycobacterium Infections [28233049], Influenza [28351919], Lesion of Brain [28394853], Pulmonary Fibrosis [28415591], Squamous Cell Carcinoma [28430665], Brain Neoplasms [28536012, 21157381], Brain Diseases [28554730], Alzheimer's Disease [28554730, 27792010], Malignant Neoplasms [28668500, 25515510, 31285280], Ischemic Stroke [28751018], Arthritis [28797095], Impaired Cognition [28840058], Lung Diseases [28878780, 28415591], Metabolic Syndrome [28937704], Kidney Failure [29021225], Intraocular Pressure Disorder [29085298], Diabetes [29210478, 30544633, 30537520], Fatty Liver Disease [29270247], Fenestration [29326590], Non-Hodgkin's Lymphoma [29392934], Substance Abuse [29477921, 28306606], Parkinson's Disease [29692728, 31054067], Status Epilepticus [29749377], Fibrosis [29802948], Hashimoto's Encephalitis [29973381], Vascular Diseases [30010623], Periodontitis [30103845], Osteopenia [30103845, 28298636], Migraine Disorders [30146940], Central Nervous System Sensitization [30165876], Glioblastoma [30362154], Carcinogenesis [30362154, 19017759], Asthma [30387030, 27863396], Polycystic Kidney Disease [30417216, 15325248], Encephalomyelitis [30500565, 29973381], Endothelial Dysfunction [30544633, 31412063], Toxoplasmosis [30653952], Astrocytoma [30684150, 27867013], Malignant Neoplasm of Lung [30762755], Hyperglycemia [30821687], Multiple Sclerosis [31015145, 30500565, 28326637, 29341465, 30908981], Agnosia For Pain [31045747], Coronary Arteriosclerosis, Myocardial Ischemia [31051227], Neurodegenerative Disorders [31054067], Cardiac Arrhythmia [31152337], Intestinal Diseases [31202513],

Table B.7 (cont.): Putative disease genes and the associated disease phenotypes for which PubMed IDs were obtained.

Gene	Phenotype
	Congestive Heart Failure [31152337, 28840058, 16497176], Corneal Allograft Rejection [31197223], Malignant Neoplasms [31285280, 25515510, 28668500], Hyperactive Behavior [31396053, 29343707], Mental Disorders [31396053, 18268501, 29122639], Adenocarcinoma [31401785], Malignant Neoplasm of Liver [31401785, 26517690, 30343397], Oedema Auricular [31627451]
ITGB1	Congestive Heart Failure [11884376], Anoxia [12200131], Neoplasm Metastasis [12517798, 22695923, 23336515, 22829201, 28613134, 30611077, 15500293, 28401009, 26728244, 27004522, 25894721, 30399378, 30688657, 27742688, 25089569, 25741138, 29746931, 31523184, 24762228, 22943849, 15024036, 29948648, 24004467, 23441154, 22382453, 30581390, 23562787, 29431199, 31423176, 28560430, 29703238, 28542982, 26903137], Astrocytosis, Gliosis [12851778], Carcinogenesis [1337297, 30849478, 30611077, 26497667, 19829083, 24931361], Adenocarcinoma [14612932, 16732726, 27289231, 23677397], Lymphoma [14623330, 15731179, 30350200, 29301866, 11264180, 12393420], Thalassemia [15054814, 30506348], Melanoma [15292257, 18632734, 11996105, 12376466, 12218055, 17352405, 28284838, 28476030, 28476030], Endometrial Neoplasms [15645131], Myocardial Infarction [15978110, 28367125], Malignant Neoplasm of Urinary Bladder [16103120, 11996105, 22386417, 29113179, 31092844, 31486485, 28042869, 22386417, 25092917, 30915742, 28498731, 29113179], Intracranial Arteriovenous Malformation [16385340], Alzheimer's Disease [16448724, 30918899], Hemangioma [16456130], Sclerosis [16459165], Malignant Neoplasm of Urinary Bladder [16820945, 28498731], Cardiomyopathy [17186162, 18340010, 11884376, 27693578, 18340010], Cardiac Arrest [17313560], Kidney Diseases [17514628, 30271355], Bullous Pemphigoid [17515951], Galloway Mowat Syndrome [18594871], Chylothorax [18973153], Colitis [19103643], Glomerulonephritis [19662603], Hydrocephalus [19726708], Diabetes [20187441, 25371288], Endometriosis [21063030, 26357653], Basal Cell Carcinoma [21067603], Neoplasm Invasiveness [21224397], Ameloblastoma [21255062, 17498594], Malignant Neoplasm of Esophagus [21426305], Malignant Neoplasm of Head and/or Neck [21463917, 30642292, 28613134, 31632090, 30779921], Ulcerative Colitis [22486997, 25386078, 29596443], Meningioma [23238945], Venous Thromboembolism [23358226], Malignant Neoplasm of Thyroid [23388428, 31392080], Lymphatic Metastasis [23441154], Malignant Neoplasm of Lung [23441154, 26728244, 29636624, 30719123, 31598171, 12020426, 31598171, 23441154, 21053345, 31696479, 21478906, 26509557, 27207836, 28537888, 22829201, 29673969, 31796058, 30688657], Malignant Neoplasm of Lung [23441154, 26728244, 31598171, 30719123, 29636624], Gestational Trophoblastic Disease [23455756], Hydatidiform Mole [23455756, 23455756], Uterine Fibroids [23482612], Hepatitis C [23498955], Crohn's Disease [23625284], Malignant Neoplasm of Ovary [23877403, 26497667, 22388103, 29218693, 22388103, 23877403, 12517798], Degenerative Polyarthritis [24289792, 30854241], Malignant Neoplasm of Stomach [24870620, 25741138, 21618249, 26903137, 28542982, 27832972, 27832972, 24870620, 16773720],Malignant Neoplasm of Larynx [30519312, 29930482, 29930482], Malignant Neoplasm of Breast [30771534, 29636624, 29774124, 30502358, 28213554, 26728650, 31579239, 24805830, 29948648, 28160423, 27197172, 30719702, 28160423, 26728650, 30771534, 28361350, 28969074, 19074826, 27563827, 26675717], Cholera [30808871], Schizophrenia [30813134], Urinary Stress Incontinence [31059065], Cerebral Infarction [31081104], Phlebosclerosis [31214872], Malignant Neoplasm of Colon and/or Rectum [31215867, 29515334, 15757908, 31519588, 31801971, 24807392, 31676872, 31519588, 24498407, 30863155, 29636624, 31423176, 30784076, 24777809, 25894721, 25387809, 31704841, 31704841, 31801971], Ankylosing Spondylitis [31471299], Multiple Myeloma [31586190, 18615555, 19057841, 26848618, 18850009], Malformations of Cortical Development [31657647], Rabies [31666383], Tumour Budding [31801971, 31676872]

Table B.7 (cont.): Putative disease genes and the associated disease phenotypes for which PubMed IDs were obtained.

Gene	Phenotype
PRDM2	Malignant Neoplasm of Liver [10508492, 18712668, 16706806, 10862032, 17963297, 20675009, 11259086, 28339081, 12557265], Malignant Neoplasms [10508492, 29367689, 11748455, 10369808, 17034532, 19746436, 8921366, 29228717], MSI-High [11135439], Malignant Neoplasm of Colon and/or Rectum [11259086, 10688904, 11280725, 14760116, 19843671, 25987089], Endometrial Carcinoma [11280725, 10987271, 28528974], Melanoma [12082534], Lymphoma [12472571, 22300346, 16039715, 8921366, 12002276, 11544182, 12002276, 8921366, 15201966, 29352181], Malignant Neoplasm of Head and/or Neck [12631603], Treatment Related Leukaemia [12888905], Malignant Neoplasm of Stomach [14534544, 11259086, 15069684, 14534544, 15309726], Malignant Neoplasm of Thyroid [14668725, 25722013, 17103461], Medulloblastoma [14688019], Neoplasms [14688019, 12631603, 11280725, 16706806, 14633678, 19799859, 21369371, 10508492, 28718376, 27757741, 16953217, 12082534, 29228717, 11259086, 11748455, 14668725, 22363126, 23098508, 9242555, 15711769, 28560012, 9766644, 28528974, 17103461, 27713401, 22300346, 20159667, 22614009, 29367689, 19602237, 11719434, 28243945, 11135439, 11544182, 17922684, 18712668, 20675009, 24115813, 17052263, 20878080, 18488713, 12002276, 15809732, 25987089, 29575614, 10987271, 10688904, 21369371], Marinesco-Sjogren Syndrome [14760116], Oligodendroglioma [15711769], Paraganglioma [15809732], Pheochromocytoma [15809732, 14668725], Blast Phase [16953217, 18246120, 19602237], Squamous Cell Carcinoma [17034532, 24115813, 24993551, 22363126], Carcinogenesis [17052263, 15069684, 25987089, 14668725, 21678463, 9766644, 11544182, 17103461, 11135439, 22363126, 10688904, 20159667, 18712668], Adenocarcinoma [17693662], Malignant Neoplasm of Lung [17693662, 11719434, 9766644, 9766644, 17693662], Malignant Neoplasm of Ovary [17922684], Retinoblastoma [18037365, 19746436, 10369808, 20675009, 8921366, 15488642, 23098508, 8643684, 9006946, 27830966, 16674107], Osteosarcoma [18488713], Hyperactive Behavior [18712668], Neuroblastoma [18819746, 9766644, 20878080, 20878080, 18819746], CML Progression [19602237], Malignant Neoplasm of Cervix [20159667, 29575614], Osteoporosis [20508921], Myelodysplastic Syndrome [20828818], Ganglioneuroma [20878080], Malignant Neoplasm of Prostate [21369371, 19746436, 17052263, 19746436, 21369371], Parkinson's Disease [21469201, 26227905], Congenital Contractural Arachnodactyly, Opisthorchis Viverrini-Related Cholangiocarcinoma [23098508], Malignant Neoplasm of Esophagus [24115813], Prolactinoma [25884948], Malignant Neoplasm of Urinary Bladder [26039340], Substance Abuse [27573876, 27657733, 18316681], Glioblastoma [27757741], Myeloid Leukemia [27830966, 16953217, 19602237], Malignant Neoplasm of Endometrium [28528974], Meningioma [28560012, 22614009, 28560012, 28243945], Malignant Neoplasms [29228717, 17034532, 29367689, 11719434, 11748455, 19746436, 8921366, 14633678, 10369808, 31741141, 11135439, 9766644, 10508492], Obesity [29367689], Intracranial Aneurysm [30823506], Malignant Neoplasm of Breast [9766644, 16356493, 11259086, 10508492, 21503890, 11259086, 9766644, 15069684]
BOD1	Intellectual Disability [27166630], Malignant Neoplasm of Breast [30091683], Herpes Zoster Disease [31064637]
APOBEC3H	Immunologic Deficiency Syndromes [16571802, 30060196, 29267382], Malignant Neoplasm of Lung [26459911, 27650891], Multiple Malignancies [27016308, 29290613], HIV Infection [29153851, 26559750, 25721876, 25411794, 22023594, 21167246], Malignant Neoplasm of Liver [31322199], Hepatitis B [31400856]
SCD	Hypertriglyceridemia [12401889], Malignant Epithelioma [12419843], Malignant Neoplasms [12419843, 12376462, 16316942, 23139775, 15708362, 31284458, 30558661, 28448568, 24368438, 31119400, 23013158, 31838050, 31481234, 20876744, 28442322, 29869888, 29938858, 25880005, 31019378, 25675381, 29484133, 30248655, 30065049, 21954435, 26451612, 29530061, 25528629, 15609334], Congenital Chromosomal Disease [14995083], Hyperinsulinism [15030794, 16741579, 16284748], Diabetes [15030794, 21733300, 30248655, 15662557, 31838050, 15662557, 15030794, 23015358, 28356733, 30958562, 25930966, 27004414, 30190473, 24985009, 29089222, 16908084, 31543975],

Table B.7 (cont.): Putative disease genes and the associated disease phenotypes for which PubMed IDs were obtained.

Gene	Phenotype
	Adenocarcinoma [15609334, 28797843, 18813799, 28368399], Neoplastic Cell Transformation [15708362], Anaplastic Carcinoma [16316942, 12376462], Attention Deficit Hyperactivity Disorder [16893529], Coinfection [17456467], Hypercholesterolemia [17456467, 28750643], Congestive Ophthalmopathy, Myopathic Ophthalmopathy [17614770], Hyperlipidemia [18340007], Dyslipidemias [18340007, 29074585, 17456467, 24759262, 26879377, 31356236, 29922430], Anoxia [18832746], Schizophrenia [19195843], Endothelial Dysfunction [19954369], Congenital Long Qt Syndrome [20233272], Perinatal Depression In Mother [20395685], Hemoglobinopathies [20712578], Depressive Disorder [20863572, 26513616], Leg Ulcer [20872960], Impaired Glucose Tolerance [21741058, 21661758, 29965978], Zinc Deficiency [23213233], Recurrent Tumor [23376425], Tumor Necrosis [23612727], Dermatologic Disorders, Insulin Resistance [21661758], Malignant Neoplasm of Ovary [23676551], Anemia [29232169, 26450553, 31099901, 15048161, 21288648, 29459493, 31385153, 27667587], Metabolic Diseases [29258511, 31673045, 29043930, 29074585, 24295027, 15030794, 29089222, 29605251], Dyssomnias [29324574], Vitamin A Deficiency [29409806], Meibomian Gland Dysfunction [29463801, 27569371], Malignant Neoplasms [29484133, 31119400, 29530061, 30558661, 30065049, 23139775, 23013158, 27861513, 21954435, 26451612, 24368438, 15609334, 31481234, 31019378, 30248655, 20876744, 28442322, 25880005, 28448568, 25675381, 29938858, 31284458, 20713121, 31838050, 25528629, 15708362, 29869888], Bacteremia [29531653], Malignant Neoplasm of Endometrium [29552293], Lipodystrophy [29684791], Myocardial Ischemia [29864776], Cardiac Arrest [29864776, 30486705, 21831960], Skin Toxicity [29869888], Neoplasm Metastasis [29989648, 31733993, 23633458, 29530061, 31678511], Virus Diseases [30118512], Sleep Apnea [30135591], Hypertrophic Cardiomyopathy [30152798, 28559533, 29864776, 12109867], Anterior Segment Ischemia [30153804], Hodgkin's Disease [30195881], Gestational Diabetes [30216387], Patulous Eustachian Tube [30287116], Synovial Hypertrophy [30418113], Mitral Valve Prolapse Syndrome [30486705], Parkinson's Disease [30527540, 30635723], Pervasive Development Disorder [30544006], Autism Spectrum Disorders [30544006, 30180836], Vitamin D Deficiency [30553405], Melanoma [30558661, 30184109], Tachycardia [30590464, 31026510, 30737990], Rheumatoid Arthritis [30706710], Ventricular Fibrillation [30737990], Dengue Fever [30853381], Arteriosclerosis [30927246, 23747827, 21045115, 31119852, 30927246, 18832746, 29864776, 29074585, 14521667], Tumor Initiation [30930246], Splenic Sequestration [31069977], Immunosuppression [31073775], Neoplasms [31083642, 31019378, 29326439, 21954435, 23612727, 25675381, 27468719, 28143772, 26813308, 31847887, 27861513, 23019225, 24309934, 23633458], Renal Carnitine Transport Defect [31099901, 29459493], Hypertensive Disease [31129993], Liver and Intrahepatic Biliary Tract Carcinoma [31199678, 28647567], Depressive Disorder [31286994, 28602692], Ischemic Cardiomyopathy [31375934], Bipolar Disorder [31455761], Congestive Heart Failure [31502385, 27476098, 26670611, 26133158], Anorexia Nervosa [31540208], Substance Abuse [31593753], Impaired Cognition [31597710, 31810489], Presenile Dementia [31597710, 31810489, 30635723, 28646686], Hyperglycemia [31690632], Hamartoma [31738477], Malignant Neoplasm of Colon and/or Rectum [31777589, 27992526, 31481234, 29074607, 31847887, 26451612, 26647913, 31678511, 26813308, 28894242, 29530061, 28661203, 31339921, 31119400, 23633458, 27468719, 27861513, 29530061, 26813308, 29074607, 27992526], Glioblastoma [31798454, 29354058], Amyloidosis [31810489], Dementia [31810489, 30635723, 31597710, 28646686], Mild Cognitive Disorder [31810489, 31597710, 28646686, 31517023], Carcinogenesis [31838050, 30063922, 26391970, 28145413, 15609334, 28143772]
PPHLN1	Intrahepatic Cholangiocarcinoma, Primary Cholangiocarcinoma of Intrahepatic Biliary Tract [25608663]
EREG	Anaplasia [14581411], Malignant Fibrous Histiocytoma [15274392], Neurilemmoma [16462207], Truncus Arteriosus [16469638], Angiofibroma [18292222], Intraepithelial Neoplasia [18497965], Neoplasm Metastasis [19138957, 23374602, 30738695], Erythropoietic Protoporphyria, Ferrochelatase Deficiency [19267999],

Table B.7 (cont.): Putative disease genes and the associated disease phenotypes for which PubMed IDs were obtained.

Gene	Phenotype
	Gastrointestinal Stromal Tumors [19298600], Malignant Neoplasm of Prostate [20651988], Malignant Neoplasm of Ovary [21769422, 15313392], Tuberculosis [22170233, 30634928, 30634928, 24898387], Malignant Neoplasms [22491422], Malignant Neoplasm of Stomach [22508389, 30738695, 25203737, 30738695, 28960674], Malignant Neoplasm of Liver [22516259, 22409860, 27296289], Adenocarcinoma [22964644], Behcet Syndrome [23625463], Cholesteatoma [23826119], Rheumatoid Arthritis [24309559], Glioblastoma [24330607], Liver Regeneration Disorder [24812054], Congenital Contractural Arachnodactyly [24935374], Respiration Disorders [25402004], Arthritis [25556244], Skin Toxicity [25707609], Herpes Nos, Kaposi Sarcoma [25979343], Hematologic Neoplasms [25990537], Carcinogenesis [26215578, 26894620, 11156386, 28274874, 25990537, 20498653, 12702554], Malignant Neoplasm of Breast [26215578, 17962208, 28274874, 31674634], Tumor Cell Invasion [26381405, 18620900, 22964644, 31793126, 31661003, 31234944], Neurodegenerative Disorders [27130034], Adenocarcinoma [27270421], Cleft Palate [28282589], Amyotrophic Lateral Sclerosis [28366802], Nasal Polyps [28398769], Malignant Neoplasm of Lung [28472347, 15152945, 19138957, 22964644, 26894620, 29109770, 18620900], Diabetes [28499214, 29130334, 29130334, 28499214], Temporomandibular Joint Disorders [28783046], Malignant Neoplasm of Bone [28915604, 25004126], Neurofibromatosis 1 [29032173], Immunologic Deficiency Syndromes [29109770], Tumor Progression [29353521, 20636398, 28465351, 28733611, 30738695, 28686677], Diabetic Nephropathy [29385323], Endotoxemia [29621998], Polycystic Ovary Syndrome [29734187], Arteriosclerosis [29744301], Asthma [30036600, 31443605], Malignant Neoplasm of Colon and/or Rectum [30252132, 27270421, 27272216, 22409860, 24335920, 31519572, 23374602, 27344184, 29199273, 21206494, 17664471, 19738126, 25520391, 26284333, 27002940, 24800946, 21283802, 26341080, 23099994, 24335920, 30252132], Atypical Teratoid Rhabdoid Tumor [30302601], Immune System Diseases [30400011], Degenerative Polyarthritis [30544699], Malignant Neoplasm of Lymph Node [30738695], Oestrogen Receptor Positive Breast Cancer [30967627], Myocardial Infarction [31062344], Squamous Cell Carcinoma [31234944, 18497965, 30302873], Neoplasms [31383134, 28839211, 29318941, 22030282, 26284333, 28255348, 27272216, 23374602, 28986856, 31234944, 31571868, 10891365, 25520391, 26894620, 30677557, 29109770, 29792309, 28351146, 24335920, 23549083, 26215578, 29086481, 29786605, 28841547, 21283802, 19925653, 31519572, 29138565, 30771581, 31808011, 27002940, 21161326, 27422777, 15520187, 22409860, 24687921, 31421407, 19738126, 26370161, 31744896, 25677871, 19138957, 30639548, 30738695, 27270421], Malignant Neoplasms [31683673, 23549083, 27744564, 30677557, 31234944, 29130334, 25990537, 29572902, 21769422, 31234944, 31683673], Amyloidosis [31696167], Dermatitis [31761787]
LAG3	Agammaglobulinemia [11433390], Myocardial Infarction [16413964], Hodgkin Disease [16757686], Mitral Valve Stenosis [17020785], Cardiomyopathy [21492761, 22886719, 29663959], Psoriasis [25006012], Immunologic Deficiency Syndromes [25549835, 29987244], Immunosuppression [25944800, 31623599], Pseudohyperkalemia Cardiff [26095288], Viremia [26449164], Arthritis [26996070], Myelodysplastic Syndrome [27565576], Parkinson's Disease [27708076, 30279468, 30485547, 31847878], Arteriosclerosis, Coronary Heart Disease [27777974], Squamous Cell Carcinoma [27835902, 29847156, 30519331], Immune Suppression [28028751], Lupus [28052118, 27911796], Hepatitis B [28072682, 31390978, 28072682, 29033936], Adenocarcinoma [28132868], Neuroblastoma [28270499], Hepatitis A [28325534], Rheumatoid Arthritis [28511719], Allergy To Peanuts [28689791], Diabetes [28783703], Autoimmune Diseases [28783703, 28934318, 28258692], Degenerative Polyarthritis [28800255], Hepatitis [28928158], Neoplasm Metastasis [28935468], Hepatitis C [29033936, 8871676, 26595560, 26095288, 31216427], Epithelioma [29045526], Cirrhosis [29087397], Septicemia [29135922, 31440257], Mesothelioma [29163783, 31078776], Systemic Scleroderma [29245183], Malignant Neoplasm of Lung [29396238, 31053602, 28132868, 28935468], Lymphomas [29427592], Malignant Neoplasm of Ovary [29616115, 31851060, 28197366],

Table B.7 (cont.): Putative disease genes and the associated disease phenotypes for which PubMed IDs were obtained.

Gene	Phenotype
	Malignant Neoplasms [29616115, 31851060, 30279468, 27297395, 31291131, 31681578, 31219974, 31168847, 28258692, 30990738, 30511409, 31434339, 31604537, 16618772, 31077581], Coronary Artery Disease [29729899, 27777974], Malignant Neoplasm of Colon and/or Rectum [29854709, 30713804, 31077581, 31219974, 29900067, 31287991, 30861269, 31484656, 31219974, 31484656], Progressive Neoplastic Disease [29939877], Malignant Neoplasm of Liver [30145359, 29666149, 28928158, 28648905, 23261718, 28648905, 29900067], Tumor Immunity [30147330, 28935468, 30872779], Malignant Neoplasm of Stomach [30223387], Hematologic Neoplasms [30233564, 31186046], Glioblastoma [30248181, 31548728, 30248181, 31548728], Prion Diseases [30279468], HIV Infection [30379878, 29987244, 30379878, 25154740, 30653605], Keratitis [30619285, 31217250], Myocarditis [30721928], Encephalomyelitis [30770703], Multiple Sclerosis [30770703, 31134049, 15674389, 17020785], Neoplasms [30880064, 29353075, 30580966, 31440257, 31244852, 31516753, 31007846, 23261718, 16618772, 27835902, 30872779, 30560130, 29602773, 16266981, 30333318, 28270499, 19261614, 31667169, 30952743, 30101129, 30587557, 28115575, 30248181, 28132868, 30272332, 30377566, 28928158, 31672704, 28920005, 15577684, 20442311, 29047105, 29427592, 30482746, 30285868, 29045526, 28935468, 28258692, 31695700, 31053602, 30524900, 10601994, 30971442, 31681578, 30233564, 31219974], Melanoma [30880064, 31219974, 29602773, 29939877, 30880064, 29599411], Intraocular Pressure Disorder [30944129], Memory Impairment [30989321], Lymphoma [31007846, 29296517, 31681578, 30116392, 30209120, 31681578, 29296517, 20228263, 28154084, 31783023, 31612643, 27565576, 28977875, 11781252, 31672704, 28901003, 29427592], Smoldering Myeloma [31053880], Pleurisy [31078776], Malignant Neoplasm of Urinary Bladder [31174611], Herpes Simplex Infections [31217250, 30619285], Malignant Neoplasm of Pancreas [31219974], Influenza [31236615], Tuberculosis [31306460, 25549835, 25549835, 28880895], Tumor Progression [31413911], Multiple Myeloma [31445183, 20568250], Sarcoma of Soft Tissue [31516753], Asthma [31534137, 31794836, 28886264, 30292924], Malaria [31584094], Pleural Effusion Disorder [31639551], Inflammatory Myofibroblastic Tumor [31667169], Dendritic Cell Sarcoma [31667169, 23888072], Classical Hodgkin's Lymphoma [31697809], HIV Infection [31827152, 24906112], Malignant Neoplasm of Breast [31841754, 29045526, 31434339, 31681578, 29983831, 29983831, 31841754, 31077581, 16618772, 30511409, 27297395, 31604537, 30990738, 31168847, 31721169, 28258692, 31219974, 30279468, 30223387, 31291131, 29963107]
SDC3	Muscular Dystrophy [11968010], Obesity [17698399, 19820907, 17018662, 29666642], Hyperandrogenism [19820907], Neoplasms [23060448, 23351331], Neoplasm Metastasis [23060448, 28638231], Malignant Neoplasm of Breast [23351331], Hepatitis C [28404852], Malignant Neoplasm of Liver [28557334], Malignant Neoplasm of Pancreas [28638231], Malignant Neoplasms [28638231, 29968393], Metabolic Syndrome [29666642], Malignant Neoplasm of Colon and/or Rectum [29968393], Malignant Neoplasms [29968393, 28638231], Lymphoma [30226583], Alzheimer's Disease [30718543, 31608281], Periodontitis, Inflammatory Disorder [31300004], Rheumatoid Arthritis [31300004, 16052590]
GGCT	Neoplasms [10340908, 21508379, 11358908, 29362917, 9586664, 29429592, 11397402, 22653386, 29316360], Azoospermia [11549683], Neurodegenerative Disorders [11807295], Malignant Neoplasm of Male Breast [12602915], Hypoaldosteronism [14614232], Malignant Neoplasm of Liver [15042566], Male Infertility [15044606, 17684052], Malignant Neoplasm of Endometrium [15721279], Endometrial Carcinoma [15721279, 31757677], Peroxisome Biogenesis Disorder [16100771, 20441557], Malignant Neoplasm of Testis [16172197], Schizophrenia [16556465], Lymphoma [16765912], Diabetes [17130574, 25501168, 20187968, 28323045], Rheumatoid Arthritis [17431729], Malignant Neoplasm of Ovary [17465232, 11437399, 29429592, 30011933, 29429592], Malignant Neoplasms [17465232, 11437399, 29429592, 21508379, 20527979, 26828272, 22653386, 25941902, 31519583, 26339607, 29316360, 31757677, 22144684, 26818177, 29736310, 19963113], Severe Myopia [17557158], Pharyngitis [18217553], Bulimia Nervosa [19852950], Progeria [19938095], Malignant Neoplasm of Colon and/or Rectum [19963113, 27500968],

Table B.7 (cont.): Putative disease genes and the associated disease phenotypes for which PubMed IDs were obtained.

Gene	Phenotype
	Myotonic Dystrophy [20080938], Mucocutaneous Lymph Node Syndrome [20374367], Carcinogenesis [20527979], Deficiency of Monooxygenase [20529578], Tumor Cell Invasion [21508379, 31757677, 26339607], Alopecia [21981665, 26178169, 15824176], Immunodeficiency 13 [22184408], Malignant Neoplasm of Lung [22653386, 25941902, 24204934, 11397402, 25941902, 22653386], Benign Prostatic Hyperplasia [22653589, 23184046], Systemic Scleroderma [23027890], Vascular Diseases [23098893], Ankylosing Spondylitis [23441776], Hepatitis C [23978570], Autism Spectrum Disorders [24453138], Immune Thrombocytopenic Purpura [25158149], Osteosarcoma [25170932, 30011933, 21508379, 30011933, 25170932], Asthma [25233048], Malignant Neoplasm of Mouth [25639284], Malignant Neoplasm of Pancreas [25639356], Multiple Sclerosis [25937052], Uterine Fibroids [26773178], Malignant Neoplasms [26828272, 26339607, 29316360, 21508379, 22144684, 14652007, 25941902, 31757677, 26818177, 19963113, 22653386, 29736310, 20527979, 31519583], Malignant Neoplasm of Stomach [27905872, 30485682, 30485682, 27905872], Squamous Cell Carcinoma [28287835, 17634560], Melanoma [28799406], Malignant Neoplasm of Breast [29362917, 11437399, 20527979, 25256603, 22101789, 16920725, 12023982, 22101789, 29362917], Tumor Progression [29429592, 30011933], Autosomal Dominant Tubulointerstitial Kidney Disease [29513881], Malignant Neoplasm of Thyroid [29552790], Cryptorchidism [29914823, 15757859], Gastric Cancer [30485682], Malignant Neoplasm of Lymph Node [30485682, 29429592], Baratela-Scott Syndrome [30554721], Hypercholesterolemia, Hyperglycemia, Depressive Disorder [30728411], Malignant Neoplasm of Urinary Bladder [30952730], Glioblastoma [31345573, 26828272], Neuronal Intranuclear Inclusion Disease [31413119, 31178126], Leukoencephalopathy [31433517], Lupus [31565862, 20571895, 17339498], Malignant Neoplasm of Prostate [7728763, 31519583, 31345573, 11935317, 12084187, 22653589, 12376504, 18181049, 29316360, 10234512, 15824176, 23184046, 27357524, 15810021, 11702204, 22653589, 7728763]
FYN	Parkinson's Disease [11193173, 31148150, 14720204, 27671864, 18650345, 29246765, 26924014, 11733371, 28990084, 29241709, 28460160, 19018246, 23223297], Epilepsy [11226670, 28922833], Neuroblastoma [12450793], Glioblastoma [15994925], Malignant Neoplasm of Stomach [16367923], Catalepsy [16407246], Osteoporosis [17320499], Asthma [17703099], Adenocarcinoma [17943724], Substance Abuse [18849153, 11121167], Malignant Neoplasm of Prostate [18990162, 17943724, 26624980, 17943724, 31530281], Malignant Neoplasm of Prostate [18990162, 26624980, 17943724], Bipolar Disorder [19330793, 19468241], Dementia [19953343], Lupus [19955046], Tumor Cell Invasion [20087650, 27349276], Shy-Drager Syndrome, Parkinsonian Disorders, Cerebellar Ataxia [20493840], Mesothelioma [22354875], Multiple Sclerosis [23469041], Gastrointestinal Stromal Tumors [23716303], Cerebral Atrophy [24243499], Arteriosclerosis [24626634], Neoplasms [24882577, 29140740, 29413048, 23716303, 7723281, 17943724, 31077609], Down Syndrome [24927707], Diabetes [25371288, 24098138], Liver Cirrhosis [25380136], Lymphoma [26437031, 28978570, 26848862, 18337055, 24413734], Tumor Progression [26549256], Neoplasm Metastasis [26624980, 31077609, 24882577], Obesity [26646899], Crest Syndrome [26669670], Malignant Neoplasm of Urinary Bladder [26786295], Angioimmunoblastic Lymphadenopathy [27177312], Tauopathies [28033507], Myocardial Infarction [28713962], Autism Spectrum Disorders [28922833], Schizophrenia [28991256, 11121167, 19102774, 31498476, 19501919, 30285260, 19468241], Malignant Neoplasm of Breast [29066500, 24882577], Malignant Neoplasm of Breast [29066500, 29348460, 24882577, 30789269, 27349276, 19404734, 24882577, 29066500], Malignant Neoplasm of Thyroid [29140740], Lewy Body Disease [29246765, 26924014], Irritable Bowel Syndrome [29446765], Alzheimer's Disease [29875655, 30549331, 24927707, 15708437, 14999081, 30735733, 24852829, 28033507, 22927204, 29467305], Clostridium Difficile Infection [29981838, 30885591], Cognition Disorders, Neuroblastoma [30130557], Amyloidosis [30549331, 29467305], Hypothyroidism [30595370], Pulmonary Fibrosis [30658076], Memory Impairment [30716615], Squamous Cell Carcinoma [30784933], Carcinogenesis [30784933, 29140740, 29066500], Anophthalmia and Pulmonary Hypoplasia [31232492],

Table B.7 (cont.): Putative disease genes and the associated disease phenotypes for which PubMed IDs were obtained.

Gene	Phenotype
	Malignant Neoplasm of Lung [31285693, 29937990, 29360191, 29937990, 31285693, 30746237], Osteoblastoma [31467230], Parkinson's Disease [31846630]
GLOD4	Malignant Neoplasm of Liver [11642406, 12528892], Complications of Diabetes [27776915]
RPL12	Malignant Neoplasm of Prostate [17013881], Brucellosis [17931756]
RAD51B	Meigs Syndrome [11746973], Autoimmune Diseases [15942943], Hamartoma [15942943, 11978964], Neoplasms [15942943, 21852249, 19602464, 26608380, 27651161, 16778173], Fibrosarcoma [16778173], Chondroid Hamartoma [18276084, 11978964], Mental Disorders, Substance Abuse [20098672], Malignant Neoplasm of Head and/or Neck [20512145, 21368091], Glioblastoma [20610542], Biliary Cirrhosis [21399635, 26394269, 22961000, 21399635], Glioblastoma [22017238], Malignant Neoplasm of Male Breast [23001122, 27149063, 23001122], Lipoblastoma [23890983], Malignant Neoplasm of Ovary [24139550, 26351136, 26351136, 27197191, 24139550], Epithelial Ovarian Cancer [24190013, 26261251], Melanoma [25600502], Carcinogenesis [25600502, 27651161], Perivascular Epithelioid Cell Neoplasms [25651471], Malignant Neoplasms [26261251, 26351136, 24139550, 24190013, 27334422, 27683114, 27651161], Uterine Fibroids [26787895, 9892177, 11135437, 15942943, 26787895], Hereditary Breast and Ovarian Cancer Syndrome [26898890], Malignant Neoplasm of Prostate [26964030, 26339569, 23535732, 23535732, 26964030, 27197191, 29892016], Malignant Neoplasm of Breast [27149063, 28983784, 21533530, 26351136, 25600502, 27197191, 27848153, 23001122, 21593217, 19330030, 23535729, 22232737, 23593120, 29255180, 27149063, 22454379, 29059683, 24139550, 21791674, 25751625, 21852249, 27467053, 24729084, 20095854, 23593120, 28983784, 19330030, 21844186], Adenocarcinoma [27197191], Malignant Neoplasm of Cervix [27334422, 25779941, 27334422], Hyperactive Behavior, Malignant Neoplasm of Stomach, Malignant Neoplasm of Lymph Node [27651161], Malignant Neoplasms [27651161, 27334422, 27683114], Squamous Cell Carcinoma [27683114], Coronary Artery Disease [27744395], Chemical and Drug Induced Liver Injury [28043905], Parkinson's Disease [28117402], Superficial Ulcer [28361912], Lupus [28714469, 26502338], Myocardial Ischemia [29024686], Allergic Reaction [29083406, 29785011], Glycogen Storage Disease Type II [29197628], Malignant Neoplasm of Lung [29207658], Macular Degeneration [29487693, 23455636, 24526414, 29197628, 26691988, 26691988], Rheumatoid Arthritis [30166627, 24022229, 27744395, 24390342, 30423114, 28361912, 24532676], Eczema, Hypothyroidism, Respiratory Tract Diseases [30595370], Asthma [31095684, 30929738, 31619474, 29785011, 27182965, 31036433], Malignant Neoplasm of Colon and/or Rectum [31209889, 27197191, 31209889], Lymphoma [31361614, 27903959], Fanconi Anemia [31584931], Myopia [31697570]
RFFL	Hypertensive Disease [21357277]
TSC22D3	Lymphoma [11160940, 21121775], Fibromyalgia [18468809], Multiple Myeloma [18499442], Obesity [19849801, 27178044], Cushing's Syndrome, Osteoporosis [19875485], Rheumatoid Arthritis [20496421, 25047643], Neoplasms [21750716, 29695779, 31501614, 23315031, 21546924], Post-Traumatic Stress Disorder [22137507, 22981834, 31346158], Autoimmune Diseases [22369971, 30371949], Hyperinsulinism, Testicular Diseases, Male Infertility [22556341], Clostridium [22792400], Gonorrhea [22981834], Asthma [23160983], Non-Obstructive Azoospermia [23494955], Cardiovascular Diseases [24747114], Dermatitis [26077873], Psoriasis [26077873, 31572404], Immuno Suppression, Melanoma [27465291], CNS Disorder [27889894], Nodule [28363169], Myocardial Infarction [28499885], Osteopenia [28771604], Malignant Neoplasm of Thyroid [29467389], Ascites [30124596], Kidney Diseases [30301736], Alzheimer's Disease [30338290, 27889894, 30740047], Depressive Disorder [30602137], Inflammatory Bowel Diseases [30971930], Arthritis [30971930, 20496421], Colitis [30971930, 30083167], Lupus [31379872, 28601944, 31379872], Malignant Neoplasms [31501614, 28259749, 29695779, 31715130], Dyslexia, Personality Disorders [31632253], Uterine Fibroids [31665442], Inflammatory Abnormality of the Eye [31770752], Septicemia [31840802, 30124596]