Le and Pham *BMC Systems Biology* (2017) 11:61
DOI 10.1186/s12918-017-0437-x

BMC Systems Biology

SOFTWARE Open Access

# HGPEC: a Cytoscape app for prediction of novel disease-gene and disease-disease associations and evidence collection based on a random walk on heterogeneous network

Duc-Hau Le[1,2] and Van-Huy Pham[3*]

## Abstract

**Background:** Finding gene-disease and disease-disease associations play important roles in the biomedical area and many prioritization methods have been proposed for this goal. Among them, approaches based on a heterogeneous network of genes and diseases are considered state-of-the-art ones, which achieve high prediction performance and can be used for diseases with/without known molecular basis.

**Results:** Here, we developed a Cytoscape app, namely HGPEC, based on a random walk with restart algorithm on a heterogeneous network of genes and diseases. This app can prioritize candidate genes and diseases by employing a heterogeneous network consisting of a network of genes/proteins and a phenotypic disease similarity network. Based on the rankings, novel disease-gene and disease-disease associations can be identified. These associations can be supported with network- and rank-based visualization as well as evidences and annotations from biomedical data. A case study on prediction of novel breast cancer-associated genes and diseases shows the abilities of HGPEC. In addition, we showed prominence in the performance of HGPEC compared to other tools for prioritization of candidate disease genes.

**Conclusions:** Taken together, our app is expected to effectively predict novel disease-gene and disease-disease associations and support network- and rank-based visualization as well as biomedical evidences for such the associations.

**Keywords:** Cytoscape app, Disease-gene association, Disease-disease association, Random walk with restart algorithm, Heterogeneous network, Gene prioritization, Disease prioritization

## Background

The goal of gene and disease prioritization, one of the challenging issues in biomedicine, is to predict the most promising genes and diseases associated with a disease of interest. Many network-based methods have been proposed for this purpose [1, 2]. Among them, methods based on a heterogeneous network of genes and diseases are proven to outperform those solely based on a homogeneous network of genes/proteins [3–5]. In addition, these methods can not only prioritize candidate genes but also diseases; therefore, not only novel disease-gene relationships but also novel disease-disease associations can be identified. Moreover, prediction of novel genes associated with a disease, of which molecular basis is unknown, can be performed. In parallel with the proposed methods, a number of tools have been developed. However, they only focus on prediction of disease-gene associations [6, 7].

In a recent study, we have developed a tool, namely GPEC [8], which uses a random walk with restart (RWR) algorithm on a homogeneous network of genes/proteins to prioritize candidate genes. This RWR-based

* Correspondence: phamvanhuy@tdt.edu.vn
[3]Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam
Full list of author information is available at the end of the article

method is state-of-the-art among ones solely based on protein interaction network [9]. However, it can only prioritize candidate genes of diseases with known molecular basis and cannot directly figure out novel disease-disease associations.

Recently, a variant of RWR algorithm on a heterogeneous network, namely RWRH, has been proposed and used to identify novel disease-gene and disease-disease associations on a heterogeneous network of genes and diseases [3]. This method was proven to overcome limitations of the RWR-based method. More importantly, the RWRH algorithm can be extended to use any network of genes/proteins in the heterogeneous network. Indeed, a recent RWRH-based method has used a semantic similarity network of genes instead of the protein interaction network [10] and shown to outperform the original one [3]. We also note that there is no tool which employs this method available in public domain [11]. Therefore, we develop a tool, namely HGPEC, for identification of novel disease-gene and disease-disease associations. This tool can make use those advances of the RWRH-based method.

A common issue of gene prioritization tools is collection of biomedical evidence for novel promising associations between highly ranked genes and the disease of interest [6, 7]. For instance, network-based tools such as PRINCIPLE [12] and NetworkPrioritizer [13] just provide rankings for candidate genes but do not support evidences for associations between highly ranked genes and the disease of interest. In GPEC, we employed gene ontology [14], KEGG pathway [15], GeneRIF [16], PubMed [17], and OMIM [18] to support novel promising disease-gene associations. Note that, recent studies have demonstrated roles of shared known disease-associated genes, protein complexes, pathways and disease ontologies [19–22] in disease-disease associations. Therefore, in HGPEC, we additionally used protein complexes from CORUM [23] and terms from Disease Ontology [24] to support novel promising disease-disease associations.

To demonstrate functions of HGPEC, we first showed its ability in predicting novel genes and diseases associated with breast cancer. To this end, we selected top 20 ranked candidate genes and top 20 ranked candidate diseases, then performed visualization and evidence collection. Visualization results showed that most of the top ranked candidate genes are directly connected to known breast cancer-associated genes. Also, the top ranked candidate diseases are directly connected to either breast cancer or known breast cancer-associated genes. In addition to visualization, we collected evidences for promising associations between the top ranked candidate genes/diseases and breast cancer. Evidence collection results showed that each of the promising associations between the top ranked candidate genes and breast cancer is supported by at least two data sources. Meanwhile, seventeen out of the top 20 ranked candidate diseases have at least one gene, one pathway, one protein complex or one disease ontology term shared with breast cancer. Three remaining ones are highly phenotypically similar to breast cancer since they are directly connected to breast cancer in the phenotypic disease similarity network. Second, we compared the overall prediction performance of HGPEC with other tools, GPEC [8] and PRINCIPLE [12]. Simulation result on 330 diseases showed that HGPEC is much superior to these tools for prediction of disease-associated genes.

## Methods

Ranking/prioritization of candidate genes/diseases is to predict novel genes/diseases associated with a disease of interest. In this section, we first provide a summary of the RWRH-based method, which is used for ranking candidate genes/diseases in HGPEC. Then, we describe the implementation and databases used in HGPEC.

### RWRH-based method

The heterogeneous network of genes and diseases can be represented as a connected weighted graph G($V$, $E$) with a set of nodes $V = \{v_1, v_2, \ldots, v_N\}$, a set of links $E = \{(v_i, v_j) | v_i, v_j \in V\}$ and a $N \times N$ adjacency matrix $W$ of link weights. Figure 1 shows an illustrative heterogeneous network of genes and diseases. Given a disease of interest $d_1$, the

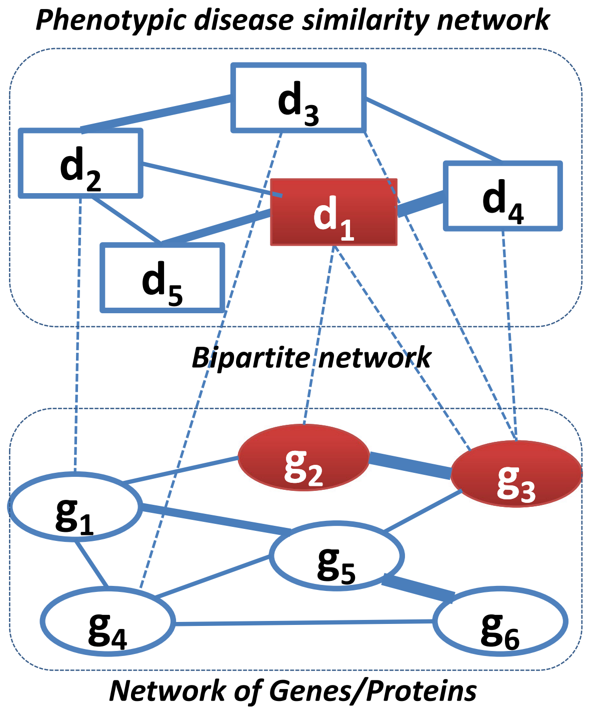

Phenotypic disease similarity network
d3
d2
d1
d4
d5
Bipartite network
g2
g3
g1
g5
g4
g6
Network of Genes/Proteins

**Fig. 1** An illustrative heterogeneous network of genes and diseases. A phenotypic disease similarity network and a network of genes/proteins are connected by known disease-gene associations (i.e., a bipartite network) to form a heterogeneous network of genes and diseases

rankings of candidate genes/diseases are based on their relative importance to a set of source $S \subseteq V$ (including $d_1$ and known $d_1$-associated genes). The relative importance measures how much a candidate gene/disease is associated with $d_1$. Here, we introduce the RWRH algorithm for such task. This algorithm was proposed for prediction of disease-associated genes on a heterogeneous network of genes and diseases [3, 10, 25], drug-target interaction prediction on a heterogeneous network of drugs and targets [26] and identification of novel disease-microRNA associations based on heterogeneous network of diseases and microRNAs [27].

RWRH mimics a walker that moves from a current node to a randomly selected adjacent node or goes back to source nodes with a back-probability $\gamma \in (0, 1)$ in a heterogeneous network. RWRH was formally defined as follows:

$$P^{t+1} = (1-\gamma) W^{'} P^{t} + \gamma P^{0}$$

where $P^t$ is a $N \times 1$ probability vector of all nodes in the network at a time step $t$ of which the $i$th element represents the probability of the walker being at node $v_i \in V$, and $P^0$ is the $N \times 1$ initial probability vector. $W^{'}$is the transition matrix of the graph, the $(i, j)$ element in $W^{'}$, denotes a probability with which a walker at $v_i$ moves to $v_j$ among $V \backslash \{v_i\}$. All nodes in the network are eventually ranked according to the steady-state probability vector $P^{\infty}$. The steady-state of each node represents its relative importance to the set of source nodes $S$.

For the heterogeneous network of diseases and genes, the transition matrix $W^{'}$was defined as follows:

$$W^{'} = \begin{bmatrix} W^{'}_{G} & W^{'}_{GD} \\ W^{'}_{DG} & W^{'}_{D} \end{bmatrix}$$

where $W^{'}_{G}$ and $W^{'}_{D}$ are intra-subnetwork transition matrices of the network of genes/proteins and the phenotypic disease similarity network, respectively. $W^{'}_{GD}$, $W^{'}_{DG}$ are inter-subnetwork transition matrices. Let $\lambda$ be the jumping probability the random walker jumps from the network of genes/proteins to the phenotypic disease similarity network or vice versa. Then, these matrices were defined as follows:

$$\left(W^{'}_{GD}\right)_{i,j} = p\left(d_j, |g_i\right) = \begin{cases} \dfrac{\lambda (W_{GD})_{ij}}{\sum_j (W_{GD})_{ij}} & if \sum_j (W_{GD})_{ij} \neq 0 \\ 0 & otherwise \end{cases}$$

$$\left(W^{'}_{DG}\right)_{i,j} = p\left(g_j, |d_i\right) = \begin{cases} \dfrac{\lambda (W_{GD})_{ji}}{\sum_j (W_{GD})_{ji}} & if \sum_j (W_{GD})_{ji} \neq 0 \\ 0 & otherwise \end{cases}$$

$$\left(W^{'}_{G}\right)_{i,j} = \begin{cases} \dfrac{(W_G)_{ij}}{\sum_j (W_G)_{ij}} & if \sum_j (W_{GD})_{ij} = 0 \\ \dfrac{(1-\lambda)(W_G)_{ij}}{\sum_j (W_G)_{ij}} & otherwise \end{cases}$$

$$\left(W^{'}_{D}\right)_{i,j} = \begin{cases} \dfrac{(W_D)_{ij}}{\sum_j (W_D)_{ij}} & if \sum_j (W_{GD})_{ji} = 0 \\ \dfrac{(1-\lambda)(W_D)_{ij}}{\sum_j (W_D)_{ij}} & otherwise \end{cases}$$

where $W_G$ , $W_D$ and $W_{GD}$ are adjacency matrices of the gene/protein, the phenotypic disease similarity and the bipartite networks, respectively.

By letting η be the parameter to weight the importance of each network, the initial probability vector was defined as follows:

$$P^0 = \begin{cases} (1-\eta)\dfrac{1}{|S|} & if\, v_i \in S \\ \eta & if\, v_i \equiv d_1 \\ 0 & otherwise \end{cases}$$

In case we are interested in a disease class/group, which contains set of diseases ($D$), we additionally define $P^0$ as follows:

$$P^0 = \begin{cases} (1-\eta)\dfrac{1}{|S|} & if\, v_i \in S \\ \eta\dfrac{1}{|D|} & if\, v_i \in D \\ 0 & otherwise \end{cases}$$

All remaining diseases in the phenotypic disease similarity network are specified as candidate diseases, whereas candidate genes can be specified by users in different ways such as all remaining genes, neighbors of known associated genes, etc...

## Implementation

HGPEC is developed based on the RWRH-based method as an app of Cytoscape v3.x, which is a platform for data integration, network analysis and visualization [28]. Therefore, it can work on any operating system such as Windows, Linux and Mac OS X, where Cytoscape is designed to work on. Figure 2 shows the implementation of HGPEC. In particular, HGPEC runs on a heterogeneous network consisting of a phenotypic disease similarity network, a network of genes/proteins and a bipartite network (Part A of Fig. 2). Given a disease of interest, training data including the given disease and its known associated genes is specified (Part B of Fig. 2). Candidate gene and disease sets are then provided. In which, candidate disease set includes non-training diseases. Meanwhile, candidate gene

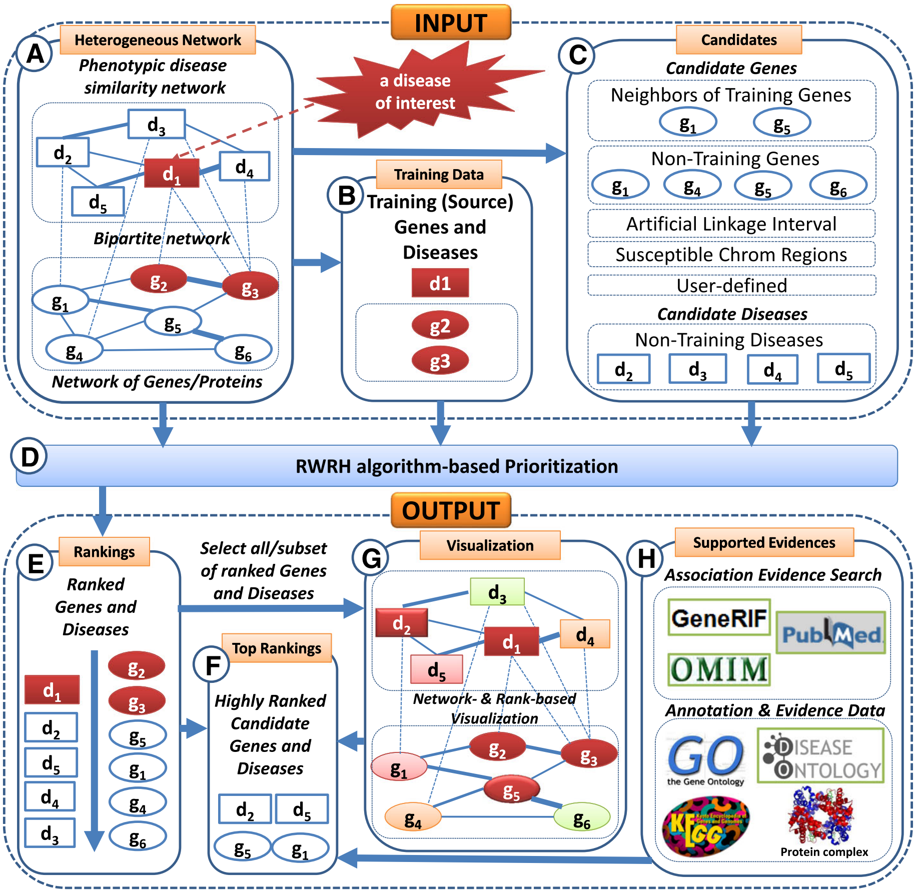

INPUT
A
Heterogeneous Network
Phenotypic disease similarity network
$d_3$
$d_2$
$d_1$
$d_4$
$d_5$
Bipartite network
$g_2$
$g_3$
$g_1$
$g_5$
$g_4$
$g_6$
Network of Genes/Proteins
a disease of interest
B
Training Data
Training (Source) Genes and Diseases
d1
g2
g3
C
Candidates
Candidate Genes
Neighbors of Training Genes
$g_1$ $g_5$
Non-Training Genes
$g_1$ $g_4$ $g_5$ $g_6$
Artificial Linkage Interval
Susceptible Chrom Regions
User-defined
Candidate Diseases
Non-Training Diseases
$d_2$ $d_3$ $d_4$ $d_5$
D
RWRH algorithm-based Prioritization
OUTPUT
E
Rankings
Ranked Genes and Diseases
$d_1$ $d_2$ $d_5$ $d_4$ $d_3$
$g_2$ $g_3$ $g_5$ $g_1$ $g_4$ $g_6$
Select all/subset of ranked Genes and Diseases
F
Top Rankings
Highly Ranked Candidate Genes and Diseases
$d_2$ $d_5$
$g_5$ $g_1$
G
Visualization
$d_3$ $d_2$ $d_1$ $d_4$ $d_5$
Network- & Rank-based Visualization
$g_2$ $g_3$ $g_1$ $g_5$ $g_4$ $g_6$
H
Supported Evidences
Association Evidence Search
GeneRIF
PubMed
OMIM
Annotation & Evidence Data
GO the Gene Ontology
DISEASE ONTOLOGY
KEGG
Protein complex

**Fig. 2** Implementation of HGPEC. Implementation of HGPEC consists of following steps: i) Constructing a heterogeneous network of genes and diseases (Part *A*); ii) Giving a disease of interest ($d_1$) and identifying training (source) genes and diseases (Part *B*); iii) Providing candidate genes and diseases (Part *C*); iv) Prioritizing all genes and diseases in the network (Part *D*); and v) Displaying all ranked genes and diseases (Part *E*) and selecting highly ranked candidate genes and diseases for further investigation (Part *F*). Novel promising disease-gene and disease-disease associations can be supported with network- and rank-based visualization (Part *G*) as well as evidences and annotations from biomedical data (Part *H*)

set can be easily constructed in several ways such as neighbors of training genes in the network, neighbors of training genes in the same chromosome, non-training genes in the network, susceptible chromosome regions/bands and freely defined by user (Part C of Fig. 2). The RWRH-based method then uses the training data to rank all candidate genes and diseases in the heterogeneous network (Part D of Fig. 2). Ranked genes and diseases are shown for further investigation (Part E of Fig. 2). For instance, highly ranked candidate genes and diseases (Part F of Fig. 2) can be further investigated by: i) network- and rank-based visualization (Part G of Fig. 2) and ii) supporting evidences including annotations for genes/diseases and evidences for novel promising disease-gene and disease-disease associations with preinstalled and automatically retrieved biomedical data (Part H of Fig. 2). Beside preinstalled data of gene and protein, gene ontology annotation and KEGG pathway like those in GPEC, we additionally preinstalled other biomedical data such as protein complex from CORUM [23] and disease ontology [24]. These data is used to annotate and support evidences for novel promising gene-disease and disease-disease associations. In addition, such associations can be further supported with evidence searched from GeneRIF, PubMed and OMIM. In HGPEC, the network of genes/proteins can be freely provided by users. By default, we loaded a human protein interaction network containing 10,486 genes and 50,791 interactions collected from ftp://ftp.ncbi.nlm.nih.gov/gene/GeneRIF/interactions.gz. This is a collection of human protein interactions from BIND [29], BioGRID [30] and HPRD [31]. Meanwhile, the phenotypic disease similarity network was collected from MimMiner [32] and the bipartite network are known disease-gene associations collected from either DisGeNET [33] or OMIM (http://ftp.ncbi.nlm.nih.gov/gene/DATA/mim2gene_medgen) [18]. These networks were also preinstalled in the app.

## Results and discussion

### Case study: Prediction of novel breast cancer-associated genes and diseases

Here, we show the ability of HGPEC in identifying novel genes and diseases associated with breast cancer (OMIM ID: 114480). Particularly, after ranking, sets of highly ranked candidate genes and diseases were further analyzed to find evidences about their associations with breast cancer. These associations were shown in a network-based view as well as evidences and annotations from biomedical data. To complete this task, we performed five following steps (see Fig. 3 and more detail in User manual in Additional file 3):

First, we constructed a heterogeneous network of genes and diseases. This network includes: i) a phenotypic disease similarity network containing 5080 diseases and 19,729 interactions, which was extracted from a phenotypic disease similarity matrix data collected from MimMiner [32] where only five interactions having largest weight to each disease were selected; ii) the default human protein interaction network and iii) the bipartite network containing known disease-gene associations collected from OMIM [18].

Second, we selected breast cancer (OMIM ID: 114480) for investigation. There are 21 known breast cancer-associated genes in the human protein interaction network. These genes and the disease of interest are played as training genes and disease.

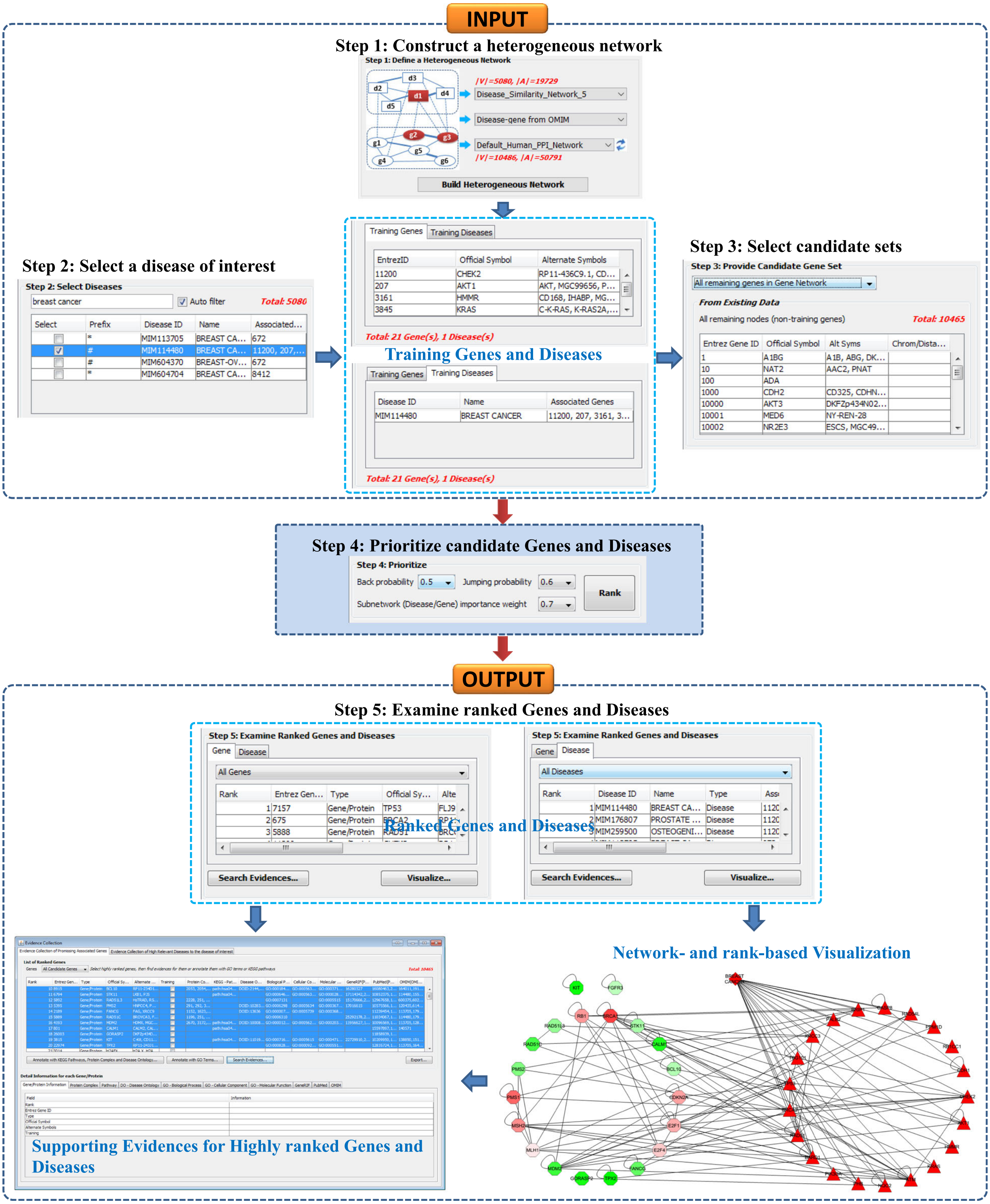
Third, we selected all remaining genes (i.e., non-known breast cancer-associated genes) in the human protein interaction network as candidate genes and all remaining diseases in the phenotypic disease similarity network as candidate diseases. As a result, the candidate gene and disease sets include 10,465 genes and 5079 diseases, respectively.

Fourth, all genes and diseases in the heterogeneous network are ranked by applying the RWRH-based method with back-probability (i.e., $\gamma$), jumping probability (i.e., $\lambda$) and subnetwork importance weight (i.e., $\eta$) were set to 0.5, 0.6 and 0.7, respectively.

Finally, the associations between highly ranked candidate genes/diseases and breast cancer are then investigated by two means: i) Network- and rank-based visualization and ii) Collection of evidences including annotations for genes/diseases and evidences of promising associations.

For network- and rank-based visualization, we first investigated topological relationships between highly ranked candidate genes and breast cancer. To this end, we selected top 20 ranked candidate genes and 21 known breast cancer-associated genes for visualization. Fig. 4a and b show that most highly ranked genes are directly connected to known genes (only two candidate genes, *KIT* and *FGFR3*, are isolated). In addition, we explored topological relationships between highly ranked candidate diseases and breast cancer. More specifically, we selected top 20 ranked candidate diseases, breast cancer and its 21 known associated genes for visualization. Fig. 4c shows that highly ranked candidate diseases are directly connected to either breast cancer or the known breast cancer-associated genes. This means that candidate diseases which are phenotypically similar to or share known associated genes with the disease of interest are highly ranked.

For collection of evidences, we first collected annotations for highly ranked candidate genes and evidences for promising associations between them and breast cancer. In particular, we annotated the top 20 ranked genes with pathways, protein complexes, disease and gene ontology terms. Then, we collected evidences for promising associations between these genes and breast cancer from GeneRIF [16, 34], PubMed [35] and OMIM [18, 36]. As a result, at least one data source provides evidences for such the associations. In addition, all collected annotations and evidences for genes and promising disease-gene associations can be exported for further use (See Table S1 in Additional file 1). Second, we collected annotations and evidences for promising associations between highly ranked candidate diseases and breast cancer. To this end, we also annotated top 20 ranked candidate diseases with pathways, protein complexes, disease and gene ontology terms. Based on reports that common associated genes, protein complexes, pathways and annotated disease ontology terms can play important roles in disease-disease associations [19–22], we additionally checked whether or not these candidate diseases share genes, pathways, protein complexes and disease ontology terms with breast cancer. Table 1 shows that twelve of them (i.e., ranks: 1, 2, 6, 10, 11, 12, 13, 14, 15, 16, 19 and 20) have at least one gene, pathway, protein complex and disease ontology term shared with breast cancer. Meanwhile, five of them (i.e., ranks: 3, 4, 8, 17 and 18) have at least one pathway, protein complex and disease ontology term shared with breast cancer, but they do not share any gene with breast cancer. This means that if we only based on shared genes to associate these diseases with breast cancer, we could not find any association. However, other biomedical data such as pathway, protein complex and disease ontology can provide evidences for these associations. Finally, three remaining ones (Ranks: 5, 7 and 9) do not share any gene, pathway, protein complex or disease ontology term with breast cancer, but they have high rankings. This is because they are phenotypically similar to breast cancer as they are directly connected to it in the phenotypic disease similarity network (see Fig. 4c). Similarly, evidences of the promising associations between the selected candidate diseases and breast cancer can be collected from GeneRIF, PubMed and OMIM based on associations between their known associated genes and breast cancer. In addition, all collected annotations and evidences for diseases and


INPUT
Step 1: Construct a heterogeneous network
Step 1: Define a Heterogeneous Network
d1
d2
d3
d4
d5
g1
g2
g3
g4
g5
g6
Build Heterogeneous Network
Step 2: Select a disease of interest
Step 2: Select Diseases
Training Genes and Diseases
Training Genes
Training Diseases
Step 3: Select candidate sets
Step 3: Provide Candidate Gene Set
From Existing Data
Step 4: Prioritize candidate Genes and Diseases
Step 4: Prioritize
Rank
OUTPUT
Step 5: Examine ranked Genes and Diseases
Step 5: Examine Ranked Genes and Diseases
Gene
Disease
Ranked Genes and Diseases
Search Evidences...
Visualize...
Evidence Collection
Supporting Evidences for Highly ranked Genes and Diseases
Network- and rank-based Visualization
BREAST CANCER
KIT
FGFR3
RB1
BRCA1
STK11
CALM1
RAD51L3
RAD51C
PMS2
BCL10
PMS1
CDKN2A
MSH2
E2F1
MLH1
E2F4
MSH2
FANCG
GORASP2
TP53

Step 1:
|V|=5080, |A|=19729 → Disease_Similarity_Network_5
→ Disease-gene from OMIM
→ Default_Human_PPI_Network
|V|=10486, |A|=50791

Step 2:
breast cancer ☑ Auto filter *Total: 5080*

| Select | Prefix | Disease ID | Name | Associated... |
|---|---|---|---|---|
| ☐ | * | MIM113705 | BREAST CA... | 672 |
| ☑ | # | MIM114480 | BREAST CA... | 11200, 207,... |
| ☐ | # | MIM604370 | BREAST-OV... | 672 |
| ☐ | * | MIM604704 | BREAST CA... | 8412 |

Training Genes:

| EntrezID | Official Symbol | Alternate Symbols |
|---|---|---|
| 11200 | CHEK2 | RP11-436C9.1, CD... |
| 207 | AKT1 | AKT, MGC99656, P... |
| 3161 | HMMR | CD168, IHABP, MG... |
| 3845 | KRAS | C-K-RAS, K-RAS2A,... |

*Total: 21 Gene(s), 1 Disease(s)*

Training Diseases:

| Disease ID | Name | Associated Genes |
|---|---|---|
| MIM114480 | BREAST CANCER | 11200, 207, 3161, 3... |

*Total: 21 Gene(s), 1 Disease(s)*

Step 3:
All remaining genes in Gene Network
All remaining nodes (non-training genes) *Total: 10465*

| Entrez Gene ID | Official Symbol | Alt Syms | Chrom/Dista... |
|---|---|---|---|
| 1 | A1BG | A1B, ABG, DK... | |
| 10 | NAT2 | AAC2, PNAT | |
| 100 | ADA | | |
| 1000 | CDH2 | CD325, CDHN... | |
| 10000 | AKT3 | DKFZp434N02... | |
| 10001 | MED6 | NY-REN-28 | |
| 10002 | NR2E3 | ESCS, MGC49... | |

Step 4:
Back probability 0.5 Jumping probability 0.6
Subnetwork (Disease/Gene) importance weight 0.7

Step 5 (Gene): All Genes

| Rank | Entrez Gen... | Type | Official Sy... | Alte |
|---|---|---|---|---|
| 1 | 7157 | Gene/Protein | TP53 | FLJ9 |
| 2 | 675 | Gene/Protein | BRCA2 | RP1 |
| 3 | 5888 | Gene/Protein | RAD51 | BRC |

Step 5 (Disease): All Diseases

| Rank | Disease ID | Name | Type | Ass |
|---|---|---|---|---|
| 1 | MIM114480 | BREAST CA... | Disease | 1120 |
| 2 | MIM176807 | PROSTATE ... | Disease | 1120 |
| 3 | MIM259500 | OSTEOGENI... | Disease | 1120 |

**Fig. 3** A workflow for prediction of novel breast cancer-associated genes and diseases. This task is completed after five following steps: *1)* Construct a heterogeneous network by selecting a phenotypic disease similarity network and a network of genes/proteins. *2)* Select breast cancer (OMIM ID: 114480, a disease of interest), then identify training genes (i.e., known breast cancer-associated genes) and training disease (i.e., breast cancer). *3)* Select a set of candidate genes; all remaining diseases in the network are selected as candidate diseases by default. *4)* Rank/prioritize all candidate genes and diseases by the RWRH-based method. *5)* Examine ranked genes and diseases by two means: ii) network- and rank-based visualization and ii) collection of annotations and association evidences for highly ranked candidate genes and diseases

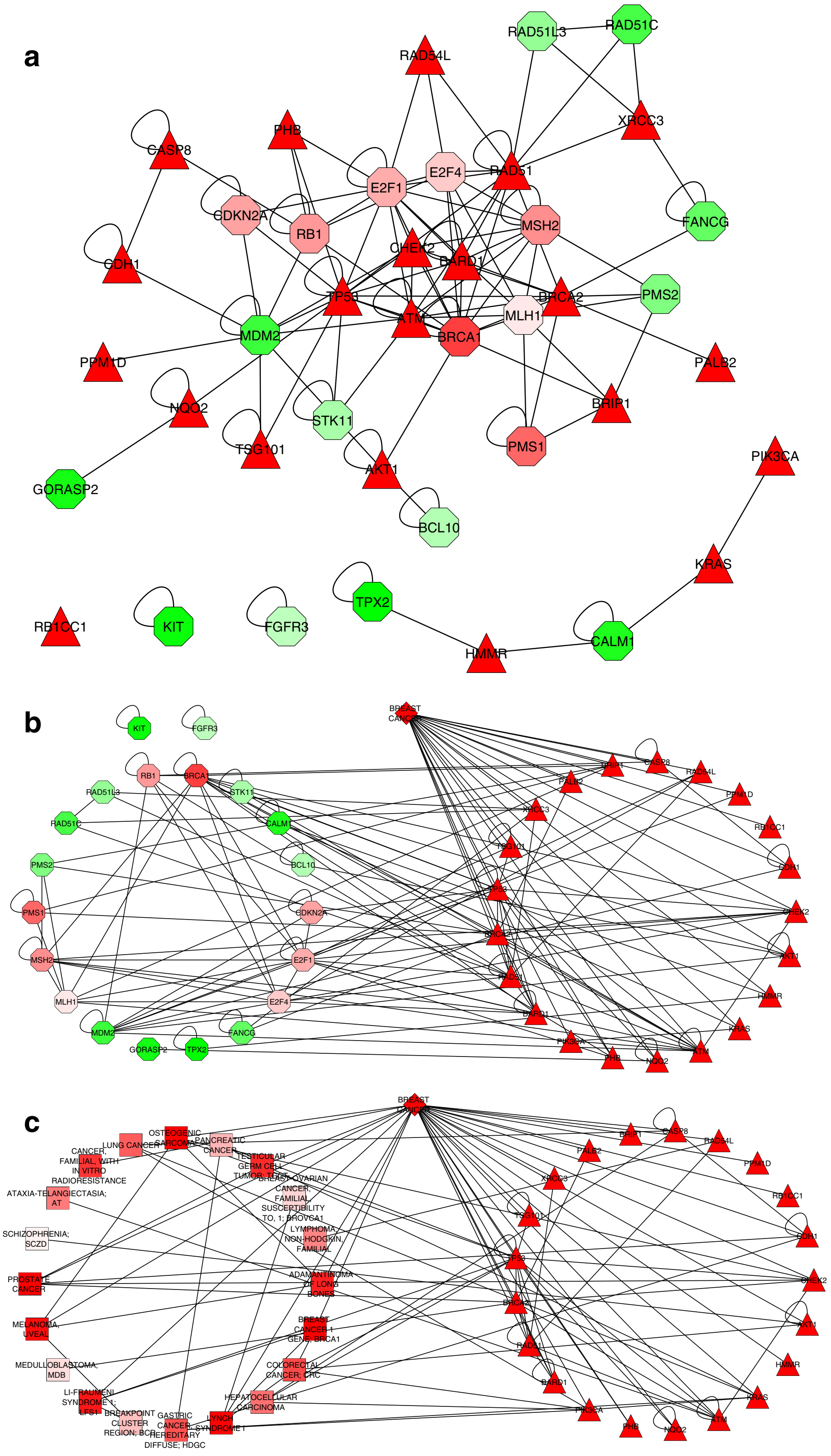


**Fig. 4** (See legend on next page.)

(See figure on previous page.)
**Fig. 4** Topological relationships between highly ranked candidate genes/diseases and breast cancer. **a** Topological relationships between 20 highly ranked candidate genes and known breast cancer-associated genes in the human protein interaction network. **b** Topological relationships between 20 highly ranked candidate genes and breast cancer. **c** Topological relationship between 20 highly ranked candidate diseases and breast cancer. Nodes in octagon, rectangle, triangle and rhombus shape are candidate genes, candidate diseases, training genes and training disease, respectively. Nodes with high rankings are in *red*, relative high are in *pink*, medium are in *white* and *light green*, low are in *green*

promising disease-disease associations can be exported for further use (See Table S2 in Additional file 2). Moreover, all of the collected annotations and association evidences can be viewed in more detail in below panels (see User manual in Additional file 3).

### Comparison to other network-based tools for prioritization of candidate disease gene

Many web-based tools, which are based on different computational methods, have been introduced for disease gene prediction [6, 7]. These tools only focus on prioritization of candidate genes. In addition, some tools require users uploading their own data. Recently, a number of Cytoscape apps have been designed for disease gene prioritization. The underlying methods of these tools are network-based since they can utilize functions of network integration, analysis and visualization of Cytoscape. Indeed, PRINCIPLE [12] is a tool for associating genes with diseases via network propagation algorithm PRINCE [37]. Given a query disease, PRINCIPLE prioritizes candidate disease genes based on their closeness in a protein interaction network to genes causing phenotypically similar disorders to the query disease. Therefore, this tool cannot directly figure out novel disease-disease associations. In addition, novel disease-gene associations predicted by this tool are not provided with biomedical evidences. Another Cytoscape app, NetworkPrioritizer [13], which is also designed for prioritization of candidate disease genes. This tool computes a number of important centrality measures to rank nodes based on their relevance for network connectivity and provides different methods to aggregate and compare rankings. Based on the final rankings, novel disease-associated genes can be predicted. However, it has the same limitation as in PRINCIPLE because there is no function in NetworkPrioritizer which helps user to search evidences for predicted disease-gene associations. In addition, it is not designed to find novel disease-disease associations. As aforementioned, we have recently developed

**Table 1** Evidence of associations between top 20 ranked diseases and breast cancer

| Rank | OMIM ID | Name | # Shared genes | # Shared protein complexes | # Shared pathways | # Shared disease ontology |
|---|---|---|---|---|---|---|
| 1 | MIM176807 | PROSTATE CANCER | 3 | 10 | 19 | 71 |
| 2 | MIM259500 | OSTEOGENIC SARCOMA | 2 | 14 | 23 | 55 |
| 3 | MIM113705 | BREAST CANCER 1 GENE; BRCA1 | 0 | 14 | 0 | 27 |
| 4 | MIM120435 | LYNCH SYNDROME I | 0 | 1 | 2 | 35 |
| 5 | MIM155720 | MELANOMA, UVEAL | 0 | 0 | 0 | 0 |
| 6 | MIM151623 | LI-FRAUMENI SYNDROME 1; LFS1 | 1 | 14 | 23 | 71 |
| 7 | MIM102660 | ADAMANTINOMA OF LONG BONES | 0 | 0 | 0 | 0 |
| 8 | MIM273300 | TESTICULAR GERM CELL TUMOR; TGCT | 0 | 3 | 8 | 34 |
| 9 | MIM211410 | CANCER, FAMILIAL, WITH IN VITRO RADIORESISTANCE | 0 | 0 | 0 | 0 |
| 10 | MIM114500 | COLORECTAL CANCER; CRC | 3 | 20 | 64 | 99 |
| 11 | MIM137215 | GASTRIC CANCER, HEREDITARY DIFFUSE; HDGC | 3 | 8 | 60 | 93 |
| 12 | MIM211980 | LUNG CANCER | 3 | 10 | 62 | 118 |
| 13 | MIM114550 | HEPATOCELLULAR CARCINOMA | 3 | 27 | 62 | 81 |
| 14 | MIM208900 | ATAXIA-TELANGIECTASIA; AT | 1 | 4 | 3 | 37 |
| 15 | MIM605027 | LYMPHOMA, NON-HODGKIN, FAMILIAL | 1 | 1 | 3 | 5 |
| 16 | MIM260350 | PANCREATIC CANCER | 2 | 14 | 45 | 54 |
| 17 | MIM151410 | BREAKPOINT CLUSTER REGION; BCR | 0 | 0 | 1 | 10 |
| 18 | MIM604370 | BREAST-OVARIAN CANCER, FAMILIAL, SUSCEPTIBILITY TO, 1; BROVCA1 | 0 | 14 | 0 | 27 |
| 19 | MIM155255 | MEDULLOBLASTOMA; MDB | 1 | 3 | 3 | 36 |
| 20 | MIM181500 | SCHIZOPHRENIA; SCZD | 1 | 3 | 40 | 84 |

a Cytoscape app, GPEC [8], for disease gene prediction and evidence collection based on the RWR-based algorithm. This app was shown more useful than the above Cytoscape apps since it has functions for collecting biomedical evidences for predicted disease-gene associations. However, it also cannot directly predict novel disease-disease associations. In addition, like the above tools, it can only work with diseases with known molecular basis. Therefore, HGPEC is introduced to overcome all of these limitations. In addition, HGPEC is designed based on a state-of-the-art network-based method (i.e., RWRH-based method), which was shown to outperform the methods used in GPEC as well as PRINCIPLE. To compare overall performance of HGPEC with that of GPEC and PRINCIPLE, we used the human protein interaction network and set the best settings for the three methods as reported in previous studies [3, 37, 38] (i.e., back-probability and weight parameter were set to 0.5 in GPEC and PRINCIPLE, respectively. Meanwhile, back-probability, jumping probability and subnetwork importance weight were set to 0.5, 0.6 and 0.7 for HGPEC, respectively). Due to using leave-one-out cross validation method, we selected a set of 330 diseases with at least two known associated genes to compare the performance of these tools in terms of AUC (i.e., area under the ROC curve) values. Figure 5 shows that HGPEC (AUC = 0.987) performs much better than GPEC (AUC = 0.788) and PRINCIPLE (AUC = 0.789).

## Conclusions

HGPEC employs the random walk with restart algorithm in a heterogeneous network of genes and diseases. It is developed to overcome the limitations of existing disease gene prediction tools. Beside the capability of prioritization of candidate genes, HGPEC can also rank candidate diseases. Therefore, it can discover not only novel gene-disease associations but also new disease-disease associations. In addition, it can identify novel genes associated with diseases without known molecular basis. Moreover, it is also convenient for users with freedom input of network of genes/proteins. Furthermore, novel promising gene-disease and disease-disease associations can be supported with network- and rank-based visualization as well as evidences and annotations collected from biomedical data. A case study on prediction of novel breast cancer-associated genes and diseases was performed to show the abilities of HGPEC. In addition, we also showed that HGPEC is much better than other tools (i.e., GPEC and PRINCIPLE) in prioritizing candidate disease genes. Note that, disease similarity network (i.e., diseasome) can be constructed based on shared disease gene [19], shared pathways [21], shared miRNA [39], shared protein complex [40], shared disease ontology [22] and disease comorbidity [41]. Therefore, in our future study, the phenotypic disease similarity network will be replaced by any diseasome, which are able to be provided freely by users. Moreover, we are going constantly to upgrade HGPEC so that it will be compatible with latest Cytoscape series and therefore become more popular.

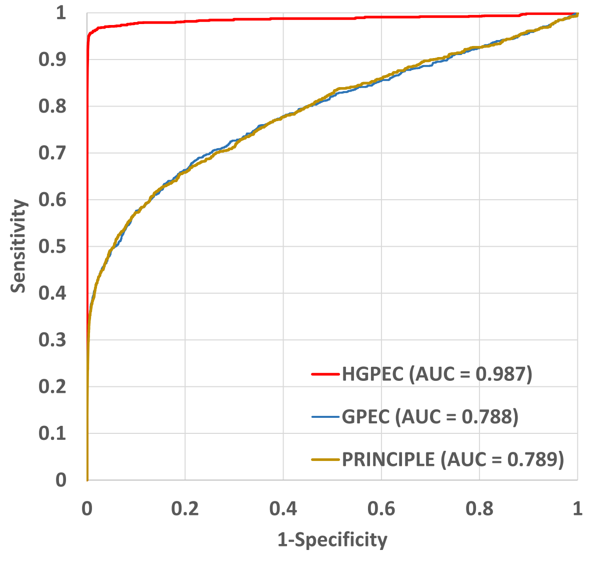


**Fig. 5** Performance comparison

## Availability and requirements

- **Project name:** HGPEC
- **Project home page:** https://sites.google.com/site/duchaule2011/bioinformatics-tools/hgpec
- **Operating system(s):** Windows/Linux/MacOS
- **Programming language:** Java
- **Other requirements:** Java 1.7 or higher, Cytoscape 3.x (Cytoscape 3.3 or higher)
- **License:** None
- **Any restriction to use by non-academics:** None

## Additional files

**Additional file 1: Table S1.** All collected annotations and evidences for associations between top 20 ranked candidate genes and breast cancer. (TXT 622 kb)

**Additional file 2: Table S2.** All collected annotations and evidences for associations between top 20 ranked candidate diseases and breast cancer. (TXT 278 kb)

**Additional file 3:** User manual. (PDF 2710 kb)

### Acknowledgements

We greatly appreciate the help of Ba-Su Pham and Anh-Minh Dao in partially porting the code from Cytoscape 2.x to 3.x platform.

### Funding

This research is funded by Foundation for Science and Technology Development of Ton Duc Thang University (FOSTECT), website: http://fostect.tdt.edu.vn, under Grant FOSTECT.2016.BR.09.

### Availability of data and materials

The app and User manual can be downloaded at https://sites.google.com/site/duchaule2011/bioinformatics-tools/hgpec

### Authors' contributions

DHL conceived of the study. DHL and VHP wrote the program code and the manuscript. All authors read and approved the final manuscript.

### Competing interests

The author declares that there is no conflict of interests regarding the publication of this article.

### Consent for publication

Not applicable.

### Ethics approval and consent to participate

Not applicable.

### Author details

[1]Vinmec Research Institute of Stem Cell and Gene Technology, 458 Minh Khai, Hai Ba Trung, Hanoi, Vietnam. [2]Thuyloi University, 175 Tay Son, Dong Da, Hanoi, Vietnam. [3]Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City, Vietnam.

Received: 3 April 2017 Accepted: 31 May 2017
Published online: 15 June 2017

### References

1. Barabasi A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56–68.
2. Wang X, Gulbahce N, Yu H. Network-based methods for human disease gene prediction. Briefings in Functional Genomics. 2011;10(5):280–93.
3. Li Y, Patra JC. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics. 2010;26(9):1219–24.
4. Chen Y, Jiang T, Jiang R. Uncover disease genes by maximizing information flow in the phenome-interactome network. Bioinformatics. 2011;27(13):i167–76.
5. Guo X, Gao L, Wei C, Yang X, Zhao Y, Dong A. A computational method based on the integration of heterogeneous networks for predicting disease-Gene associations. PLoS One. 2011;6(9):e24171.
6. Tranchevent L-C, Capdevila FB, Nitsch D, De Moor B, De Causmaecker P, Moreau Y. A guide to web tools to prioritize candidate genes. Brief Bioinform. 2010;12(1):22–32.
7. Oti M, Ballouz S, Wouters MA. Web tools for the prioritization of candidate disease genes. In Silico Tools for Gene Discovery. 2011;760:189–206.
8. Le D-H, Kwon Y-K: GPEC: a Cytoscape plug-in for random walk-based gene prioritization and biomedical evidence collection. Comput Biol Chem 2012, 37(0):17-23.
9. Navlakha S, Kingsford C. The power of protein interaction networks for associating genes with diseases. Bioinformatics. 2010;26(8):1057–63.
10. Jiang R, Gan M, He P. Constructing a gene semantic similarity network for the inference of disease genes. BMC Syst Biol. 2011;5(Suppl 2):S2.
11. Moreau Y, Tranchevent L-C. Computational tools for prioritizing candidate genes: boosting disease gene discovery. Nat Rev Genet. 2012;13(8):523–36.
12. Gottlieb A, Magger O, Berman I, Ruppin E, Sharan R. PRINCIPLE: a tool for associating genes with diseases via network propagation. Bioinformatics. 2011;27(23):3325–6.
13. Kacprowski T, Doncheva NT, Albrecht M. NetworkPrioritizer: a versatile tool for network-based prioritization of candidate disease genes or other molecules. Bioinformatics. 2013;29(11):1471–3.
14. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene ontology consortium. Nat Genet. 2000;25(1):25–9.
15. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. 2009;38(suppl 1):D355–60.
16. Mitchell JA, Aronson AR, Mork JG, Folk LC, Humphrey SM, Ward JM: Gene Indexing: Characterization and Analysis of NLM's GeneRIFs. In: Proceedings of AMIA 2003 Symposium. American Medical Informatics Association; 2003.
17. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, et al. Database resources of the National Center for biotechnology information. Nucleic Acids Res. 2011;39(suppl 1):D38–51.
18. Amberger J, Bocchini CA, Scott AF, Hamosh A. McKusick's online Mendelian inheritance in man (OMIM®). Nucleic Acids Res. 2009;37(suppl 1):D793–6.
19. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. Proc Natl Acad Sci. 2007;104(21):8685–90.
20. Wang Q, Liu W, Ning S, Ye J, Huang T, Li Y, et al. Community of protein complexes impacts disease association. Eur J Hum Genet. 2012;20(11):1162–7.
21. Li Y, Agarwal P. A pathway-based view of human diseases and disease relationships. PLoS One. 2009;4(2):e4346.
22. Li J, Gong B, Chen X, Liu T, Wu C, Zhang F, et al. DOSim: an R package for similarity between diseases based on disease ontology. BMC Bioinformatics. 2011;12(1):266.
23. Ruepp A, Brauner B, Dunger-Kaltenbach I, Frishman G, Montrone C, Stransky M, et al. CORUM: the comprehensive resource of mammalian protein complexes. Nucleic Acids Res. 2008;36(suppl 1):D646–50.
24. Schriml LM, Arze C, Nadendla S, Chang Y-WW, Mazaitis M, Felix V, et al. Disease ontology: a backbone for disease semantic integration. Nucleic Acids Res. 2012;40(D1):D940–6.
25. Le DH, Dang VT. Ontology-based disease similarity network for disease gene prediction. Vietnam Journal of Computer Science. 2016;3:197-205. https://link.springer.com/article/10.1007/s40595-016-0063-3.
26. Chen X, Liu M-X, Yan G-Y. Drug-target interaction prediction by random walk on the heterogeneous network. Mol BioSyst. 2012;8(7):1970–8.
27. Le DH. Disease phenotype similarity improves the prediction of novel disease-associated microRNAs. 2015 2nd National Foundation for Science and Technology Development Conference on Information and Computer Science (NICS). 2015;76-81. http://ieeexplore.ieee.org/document/7302226/.
28. Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. Cytoscape 2.8: new features for data integration and network visualization. Bioinformatics. 2011;27(3):431–2.
29. Bader GD, Betel D, Hogue CWV. BIND: the Biomolecular interaction network Database. Nucleic Acids Res. 2003;31(1):248–50.
30. Breitkreutz B-J, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, et al. The BioGRID interaction Database: 2008 update. Nucleic Acids Res. 2008;36(suppl_1):D637–40.
31. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, et al. Human protein reference Database–2009 update. Nucleic Acids Res. 2009;37(suppl_1):D767–72.
32. van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JAM. A text-mining analysis of the human phenome. Eur J Hum Genet. 2006;14(5):535–42.
33. Piñero J, Bravo À, Queralt-Rosinach N, Gutiérrez-Sacristán A, Deu-Pons J, Centeno E, et al. DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. Nucleic Acids Res. 2017;45(D1):D833–9.
34. Osborne J, Lin S, Kibbe W, Zhu L, Danila M, Rex C. GeneRIF is a more comprehensive, current and computationally tractable source of gene-disease relationships than OMIM. Bioinformatics Core, Northwestern University: Technical Report; 2007.
35. Chang AA, Heskett KM, Davidson TM. Searching the literature using medical subject headings versus text word with PubMed. Laryngoscope. 2006;116(2):336–40.
36. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2005;33(suppl_1):D514–7.
37. Vanunu O, Magger O, Ruppin E, Shlomi T, Sharan R. Associating genes and protein complexes with disease via network propagation. PLoS Computional Biology. 2010;6(1):e1000641.
38. Kohler S, Bauer S, Horn D, Robinson P. Walking the Interactome for prioritization of candidate disease genes. Am J Hum Genet. 2008;82(4):949–58.
39. Lu M, Zhang Q, Deng M, Miao J, Guo Y, Gao W, et al. An analysis of human MicroRNA and disease associations. PLoS One. 2008;3(10):e3420.
40. Markou M, Singh S. Novelty detection: a review - part 2: neural network based approaches. Signal Process. 2003;83(12):2499–521.
41. Lee DS, Park J, Kay KA, Christakis NA, Oltvai ZN. BarabÃ¡si AL: the implications of human metabolic network topology for disease comorbidity. Proc Natl Acad Sci. 2008;105(29):9880–5.