Xue *et al. BMC Systems Biology* 2019, **13**(Suppl 2):34
https://doi.org/10.1186/s12918-019-0697-8

BMC Systems Biology

RESEARCH Open Access

Predicting disease-related phenotypes using an integrated phenotype similarity measurement based on HPO

Hansheng Xue[1,2], Jiajie Peng[1*] and Xuequn Shang[1*]

From The 17th Asia Pacific Bioinformatics Conference (APBC 2019)
Wuhan, China. 14–16 January 2019

Abstract

Background: Improving efficiency of disease diagnosis based on phenotype ontology is a critical yet challenging research area. Recently, Human Phenotype Ontology (HPO)-based semantic similarity has been affectively and widely used to identify causative genes and diseases. However, current phenotype similarity measurements just consider the annotations and hierarchy structure of HPO, neglecting the definition description of phenotype terms.

Results: In this paper, we propose a novel phenotype similarity measurement, termed as *DisPheno*, which adequately incorporates the definition of phenotype terms in addition to HPO structure and annotations to measure the similarity between phenotype terms. *DisPheno* also integrates phenotype term associations into phenotype-set similarity measurement using gene and disease annotations of phenotype terms.

Conclusions: Compared with five existing state-of-the-art methods, *DisPheno* shows great performance in HPO-based phenotype semantic similarity measurement and improves the efficiency of disease identification, especially on noisy patients dataset.

Keywords: Human phenotype ontology, Semantic similarity, Phenotype similarity

Background

With the high-speed development of next generation sequencing (NGS) techniques, large-scale biological and medical data is generated exponentially, which greatly contributes to Mendelian disease and cancer diagnosis [1–3]. However, it is still difficult to make accurate clinic diagnosis solely based on sequencing technologies, because of the complex and incomprehensible relationships between genetic variants and clinical phenotypes [4].

Some observable features of patients, such as behaviors and biomedical properties, are defined as patient phenotypes, which are usually determined by both genetic and environmental factors [5]. Currently, patient phenotypes are widely used to improve efficiency of disease diagnosis by analysing the complex relationships between clinic phenotypes and phenotypes of known diseases.

Human Phenotype Ontology (HPO) is a widely used ontology resource, which provides a standardized vocabulary of phenotypic abnormalities encountered in human disease [6]. HPO contains multiple types of information of phenotype, such as frequency modifier and definitions of phenotype terms. Besides, phenotype terms in HPO are organized as a directed acyclic graph (DAG) to describe the phenotypic characteristics and their relationships (An example is illustrated in Fig. 1). Based on HPO, researchers start to calculate phenotype similarity, which recently has been widely utilized to improve efficiency of disease diagnosis, and phenotype semantic similarity has become a rising research area [7, 8].

In phenotype semantic similarity area, previous researchers have proposed various HPO-based similarity

*Correspondence: jiajiepeng@nwpu.edu.cn; shang@nwpu.edu.cn
[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
Full list of author information is available at the end of the article

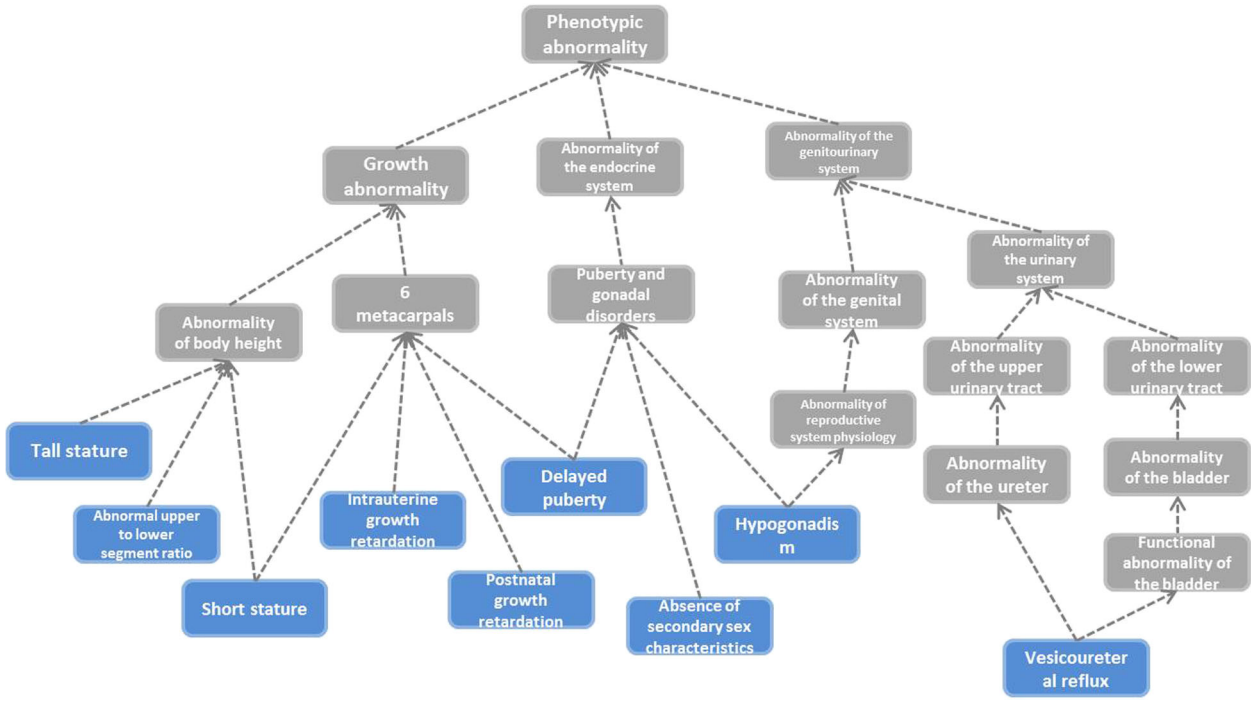

Fig. 1 An illustrative example of Human Phenotype Ontology (HPO). An example of Phenotypic abnormality (HP:0000118) forming a directed acyclic graph (DAG), in which nodes represent phenotype terms and edges represent "subclass of " relationships between phenotype terms

measurements. Most of existing semantic similarity measurements are based on Information Content (IC), such as Phenomizer [9], OWLSim [10] and PhenomeNet [11]. In detail, Phenomizer measures any two phenotype terms similarity based information content of phenotype ontology, which is similar as Masino et al. [12]. PhenomeNet and OWLSim extend simGIC [13] to calculate phenotype similarity of two phenotype sets. However, IC-based similarity measurements ignore the associated relationships of phenotype terms. Besides IC-based measurements, most existing measurements are similar to GO-based similarity measurements and neglect the unique topological structure of HPO [14–22]. And the main difference between HPO and GO is the biological knowledge representing by their structure. In the low-level of GO structure, sibling terms are often similar to each other. In contrast, sibling terms in the low-level of HPO structure are hard to prove that they have associations at the gene level or share any disease symptoms. For instance, phenotype terms "Split hand (HP:0001171)" and "Areflexia of upper limbs (HP:0012046)" are two leaf terms in HPO, but between them, there is no known gene-level associations nor shared disease symptoms [23].

Thus, it is essential to propose a novel and unique HPO-based semantic similarity measurement which designs for considering topological information of HPO. We designed a new path-constrained IC-based phenotype term semantic similarity measurement, termed as *PhenoSim*, which considers the unique DAG structure of HPO [23]. In addition, some practical online or offline tools have been developed for biological researchers, including HPOSim [24] and PhenoSimWeb [25]. HPOSim provides an offline R package, which implements seven common ontology-based similarity measurements, including Jiang [26], Lin [27], Wang [28] and Schlicker [29]. PhenoSimWeb is an easy-to-use online application which implements five phenotype measurements and provides an intuitive visualization interface.

Although above methods are widely used to calculate phenotype semantic similarity, none of them make the best of phenotype ontology information, such as definition description of phenotype term and phenotype annotation information. *PhenoSim* proposed a phenotype similarity measurement based on topological structure of HPO, but it neglects text description and association information of phenotype term. Current HPO-based methods adopt gene or disease annotations to represent information content of phenotype term. However, this method cannot describe phenotype term fully and accurately, since many annotations associated with a phenotype are still unknown [30–32]. Therefore, it is essential and necessary to explore a novel phenotype similarity

measurement that make the best of phenotype ontology information, such as hierarchical structure, term annotation and text description of phenotype.

In this paper, we propose a novel phenotype similarity measurement, named *DisPheno*, which integrates hierarchy structure and phenotype term definition of HPO. Compared with existing methods, the main contributions of our work can be summarized as:

- To the best of our knowledge, *DisPheno* is the first HPO-phenotype similarity measurement integrating term annotation, hierarchical structure and text description.
- *DisPheno* applies Point-wise Mutual Information to calculate phenotype annotations and integrates into phenotype-set similarity measurement.
- The evaluation results show that *DisPheno* outperforms some state-of-the-art approaches.

Methods

In order to improve the performance of identifying disease-related phenotypes, we propose a novel phenotype similarity measurement, termed as *DisPheno*, which is a optimized method of a path-constrained information content-based similarity measurement. DisPheno mainly contains four steps. First, it annotates phenotype ontology information content using both genes and diseases. Second, it reconstructs topological structure of phenotype term using TF-IDF method [33]. Third, it computes semantic similarity between two phenotype term t_i and t_j considering information content(IC), distance between terms and DAG structure. Finally, it computes phenotype term associations using Point-wise Mutual Information (PMI) method [34] and calculates phenotype set similarity. The framework of *DisPheno* is shown in Fig. 2. and the detailed steps will be introduced as follows.

Step 1. Annotating phenotype ontology information content

Most of current phenotype similarity measurement are based on information content(IC), and the types of annotating phenotype term mainly contains gene annotation and disease annotation. Existing phenotype similarity measurement are annotated using gene or disease, and our method integrates these two types of annotations. In annotating part, we use a weighted coefficient w to adjust

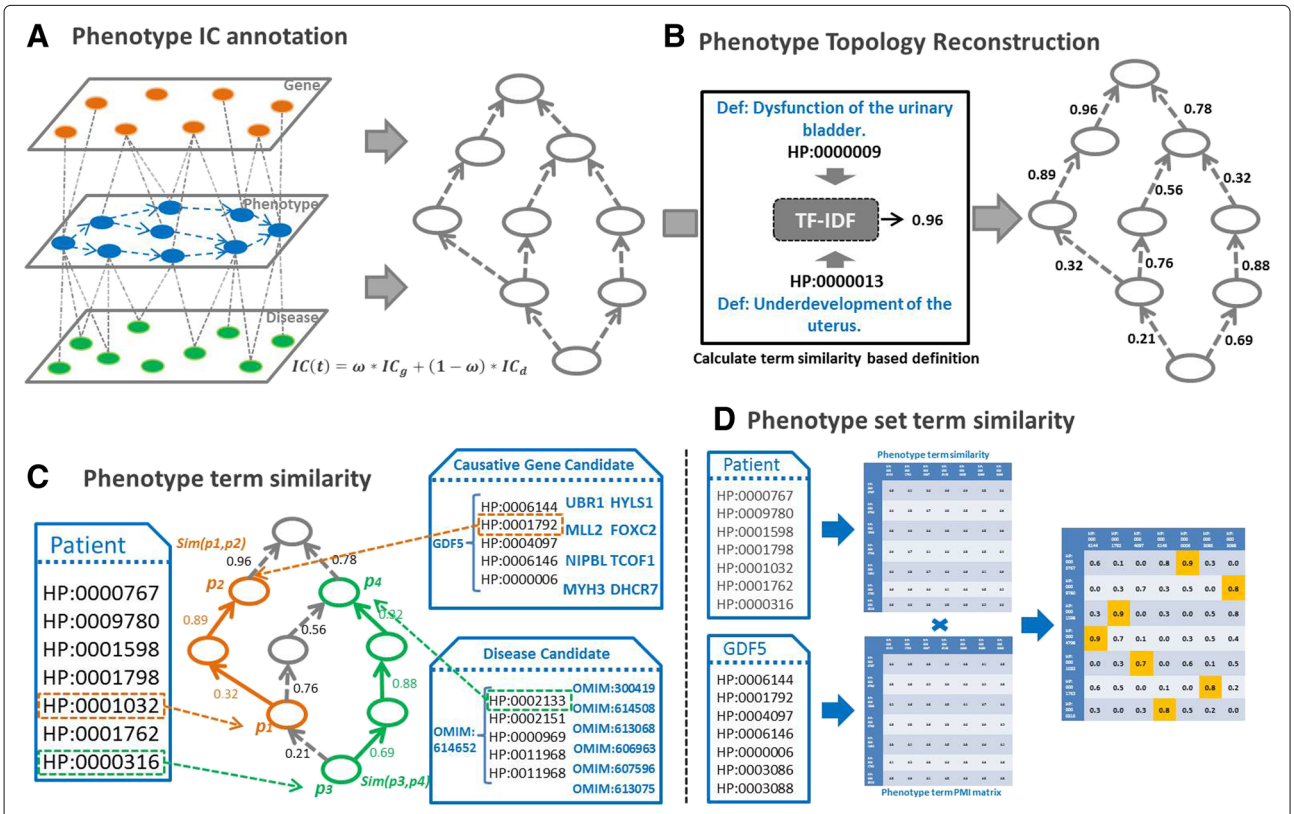

Fig. 2 The workflow of DisPheno. It mainly contains four parts: **a** Annotating phenotype ontology information content using both gene annotation and disease annotation; **b** Reconstructing topological structure of phenotype term by calculating phenotype term definition similarity using TF-IDF; **c** Measuring phenotype semantic similarity based on HPO by integrating term definition-similarity; **d** Calculating phenotype term association and set similarity by measuring phenotype term associations using Point-wise Mutual Information(PMI)

the ratio of two types of annotations. The IC of phenotype term t can be described as follows:

$$IC(t) = w * IC_{gene} + (1 - w) * IC_{disease}$$

$$IC_{gene}(t) = ln\left(\frac{G}{G_t}\right) \qquad IC_{disease}(t) = ln\left(\frac{D}{D_t}\right)$$

where $IC_{gene}(t)$ represents the information content of phenotype term t annotated by genes, G and G_t represent the size of genes annotated to the root and term t respectively ($IC_{disease}(t)$ is similar to gene annotation). Finally, we can comprehensively integrate the relationships between phenotypes and genes / diseases into the information content of phenotype terms.

Step 2. Reconstructing topological structure of phenotype term

Human Phenotype Ontology (HPO) provides a directed acyclic graph (DAG) to describe the phenotype term and associations. However, the edge of DAG has no weight and just indicate the hierarchical relationship. To further describe the relationship between phenotype terms, we try to turn original DAG into a weighted directed acyclic graph (WDAG). In our model, we calculate the cosine similarity between the definitions of phenotype terms using TF-IDF method and try to add weights for edges of original DAG.

To calculate the phenotype term similarity, we need to convert the term definition into vector by TF-IDF firstly. TF-IDF is short for term frequency-inverse document frequency, which is often used in data mining and information retrieval to measure the importance of a document in a collection or corpus [33].

Given a phenotype term definition $t = \{p_1, p_2, ..., p_n\}$, p_i represents a specify word, and the term frequency of p_i is $tf(p_i, t) = n_i/|t|$, where n_i represents the times that word p_i occurs in phenotype term definition t, and $|t|$ is the number of words in t. And the inverse document frequency of word p_i is $idf(p_i, T) = \log \frac{|T|}{|\{t \in T: p_i \in t\}|}$, where $|T|$ is the total number of phenotype term in the HPO corpus and $|\{t \in T : p_i \in t\}|$ is the number of phenotype term where the word p_i appears. Thus, the Term frequency-Inverse document frequency(TF-IDF) can be calculated as:

$$TF - IDF(p_i, t, T) = TF(p_i, t) * IDF(p_i, T)$$

After translating the phenotype term definitions into TF-IDF vectors by calculating the word TF-IDF scores, we can calculate the term similarity between pair-wise phenotype term using cosine similarity based on the TF-IDF vectors. Then, we can obtain a phenotype term similarity matrix $S \in R^{n*n}$, where n is the number of total phenotype terms. Finally, we add the phenotype term similarity into the DAG and we can reconstruct the unweighted directed acyclic graph into a weighted directed acyclic graph (WDAG). And the reconstructed WDAG will be used in the process of calculating phenotype term similarity.

Step 3. Measuring phenotype semantic similarity

Most phenotype similarity measurements are based on information content, they just consider the information content of most informative common ancestor or phenotype terms. They neglect the effects of hierarchy structure and text description of phenotype terms.

In our previous research, *PhenoSim* has proposed a path-constrained information content-based phenotype similarity measurement. The core idea is to consider the structural accessibility of phenotype terms. In detail, if there is a directed path between any two phenotype terms t_i and t_j in the hierarchy structure of HPO, we consider that these two terms are highly similar to each other and "reachable". Otherwise, these two phenotype terms are "unreachable" in the DAG structure of HPO.

Based on this measurement, we propose a novel method, termed as *DisPheno*, which considering the distance between term t_i and t_j and the pathway on the weighted directed acyclic graph. Thus, we define a novel phenotype-based similarity measurement as:

$$sim(t_i, t_j) = \begin{cases} WIC(t_{MICA}) * \left(1 - \frac{dist(t_i, t_j)}{mostDepth}\right) & reachable \\ 0 & otherwise \end{cases}$$

where $WIC(t_{MICA}) = \min(IC(t_i), IC(t_j)) * W(t_i, t_j)$, $(mostDepth - dist(t_i, t_j))/mostDepth$ implies the influences of distance between t_i and t_j, and $W(t_i, t_j)$ is the weight product from t_i to t_j among weighted directed acyclic graph of HPO. Specifically, *mostDepth* describes the longest path in the hierarchy structure of HPO, or the maximum number of edges that leaf node reaches the root node.

Step 4. Computing phenotype term association and set similarity

Before calculating the phenotype set similarity, we need to measure the association among all phenotype terms. Current phenotype set similarity measurements all adopt the average value of maximum phenotype term similarity between phenotype term and phenotype set as the phenotype set similarity. In our model, we introduce the phenotype association relationships and use Pointwise Mutual Information(PMI) to compute the phenotype term associations.

Assuming that if two term t_i and t_j belongs to same causative gene (or disease) in the gene-to-phenotype (or disease-to-phenotype) association file, we hold that term t_i and t_j are associated. The pair-wise association between phenotype terms can be calculated as:

$$PMI(t_i, t_j) = \log\left(\frac{p(t_i, t_j)}{p(t_i) * p(t_j)}\right)$$

where $p(t_i, t_j)$ is the probability that term t_i and t_j appear on the same gene or disease annotation set simultaneously, $p(t_i)$ and $p(t_j)$ are total probability of term t_i and t_j in the phenotype annotation set.

Given a patient and a candidate gene(or disease), the corresponding phenotype sets are T_p and T_c respectively. The phenotype set similarity between specific patient and candidate genes (or diseases) are the average value of pair-wise phenotype terms similarities between T_p and T_c:

$$Sim_{set}(T_p \rightarrow T_c) = \frac{1}{N_p}\sum_{t_i \in T_p} \max_{t_j \in T_c}(sim(t_i, t_j) * PMI(t_i, t_j))$$

$$Sim_{set}(T_c \rightarrow T_p) = \frac{1}{N_c}\sum_{t_j \in T_c} \max_{t_i \in T_p}(sim(t_j, t_i) * PMI(t_j, t_i))$$

where phenotype similarity $sim(t_i, t_j)$ is measured in previous step and N described the number of phenotype terms in set T. Due to the similarity score relies on the order of the phenotype-set and the above two equation are asymmetric, we use the following equation to eliminate the asymmetry affects. The symmetrical phenotype similarity measurement are described as:

$$Sim_{sym}(T_p, T_c) = \frac{1}{2}(Sim_{set}(T_p \rightarrow T_c) + Sim_{set}(T_c \rightarrow T_p))$$

Where Sim_{sym} is the average value of set similarities of two phenotype sets with different order. Phenotype term and set similarity measurement are the key of identifying true disease from candidate disease set. By modifying existing HPO-based similarity, we can further improve the efficiency of disease diagnosis.

Results

Data preparation

The experimental datasets were downloaded from Human Phenotype Ontology (HPO) official website (https://hpo.jax.org/), which contain 10,838 phenotype terms, 99,186 disease-to-phenotype annotations and 61,784 gene-to-phenotype annotations.

To evaluate the performance of our method, we used the patients that simulated in our previous work *PhenoSim*, which mainly contains "patients with known causative genes" and "patients with known diseases" two parts. Taking into account the clinical situation, we generated dataset with noise phenotype terms, named noisy dataset, and imprecision phenotype terms, named imprecision dataset. The optimal and noisy datasets used in this paper are same as our previous paper [35]. The details of simulating patients are described as follows.

Optimal dataset Each simulated patient was assigned one selected disease, and then we randomly added phenotype terms that selected disease associated with into this stimulated patient. In detail, if the randomly generated number was not greater than the known penetrance of the phenotype that disease associated with, this phenotype will be assigned to this simulated patient. The process was repeated for 100 times, then we obtained final optimal simulated patients.

Noisy dataset The noisy dataset is an extension of optimal, which considers the real clinic dataset. Before simulating noisy dataset, we firstly generated a noisy phenotype-set that much larger than the number of optimal phenotypes for every selected disease. The noise phenotype can be defined as the term which is not associated to this disease. After generating noisy phenotype-set, half number of noisy phenotype terms are selected and added into the phenotype set of simulated patients. Finally, we repeated this process for optimal patients and we generated the noisy simulated patients.

Noisy & Imprecision dataset Besides noisy phenotypes, clinical datasets usually contain imprecision phenotypes which attributes to the limitation of medical technology. The imprecision data is described as a kind of phenotype terms that one of their ancestors is associated with the disease d instead of the explicit phenotype term itself. In this noisy & imprecision dataset, we randomly selected half of the optimal terms and replaced them with one of their ancestors. Then we added noisy phenotype terms into the imprecision dataset, and the number of noise terms is half of the imprecision dataset. Finally, optimal, noisy and noisy & imprecision data all account for one-third of the whole dataset.

Performance evaluation on optimal dataset

We utilized the same evaluation criterion with PhenoSim to validate the prediction performance of *DisPheno* [12]. The main idea is to rank the candidate diseases of each simulated patient. We calculated the phenotype similarity value between the patient and each candidate diseases using *DisPheno*, then ranked all the candidate diseases in descending order by their similarity values. Higher the true disease's rank is, the better the algorithm's performance. Finally, we compared *DisPheno* with other five existing state-of-the-art measures on all the simulated datasets.

"Optimal patients with known causative gene" dataset contains 3300 simulated patients and each patient corresponds to one causative gene. We tested *DisPheno* and other five methods on this optimal dataset and compared the rank of true disease. Specifically, there is mapping relationship between causative genes and diseases. Because the HPO-based similarity measurements are usually used on disease diagnosis, we ranked the candidate diseases for each simulated patient instead of causative genes. In the cumulative rank distribution figure, we can find that *DisPheno* performed much better than the

other methods (Fig. 3). 28.78% of true candidate diseases rank first using *DisPheno* which is the highest percentage among all methods. The percentage of rank among top-3 using *DisPheno* is 49.86%, while the ratio of other methods are 42.94% (PhenoSim), 35.69% (Masino), 28.55% (Lin), 30.69% (Jiang) and 28.42% (Schlicker) respectively. In addition, 60.43% candidate diseases rank among top-5 using *DisPheno* and it is 10.98% higher than *PhenoSim* (49.45%) which is the second best method.

"Optimal patients with known disease" dataset contains 24,000 simulated patients and each patient corresponds to one disease. We tested the performance of all six approaches on this optimal dataset (see Table 1). Although the percentage of top-1 using *DisPheno* (83.12%) is less than the ratio of Schlicker (96.36%), 99.10% of candidate diseases rank among top-3 which is the highest compared with other methods. Although top-1 percentage is not highest, *DisPheno* shows great performance on disease identification. In the clinical cancer diagnosis or disease prediction, it usually provides scientists with several top candidates instead of the single highest one.

In the optimal datasets, *DisPheno* performs better than other five methods. And it also shows great performance and latent capacity on predicting disease and disease diagnosis. Considering that clinical phenotype set often contains lots of noise data, we further validate the performance of *DisPheno* on the simulated patient with noisy phenotype terms.

Performance evaluation on noisy dataset

"Noisy patients with known causative gene" dataset contains noisy phenotypes which are not annotated phenotype terms of the causative gene. We applied *DisPheno* and other five measures on the noisy dataset. Our method performed the best in all the six measurements (Fig. 4). The ratio of true diseases rank among top-5 using *DisPheno* reaches the highest (57.38%), which is 11.20% higher than the second highest method PhenoSim (46.18%). The percentage of other methods perform on this dataset are 36.85% (Masino), 10.67% (Lin), 6.80% (Jiang) and 14.61% (Schlicker). *DisPheno* shows great performance on noisy patient with known causative gene, it indicates good application prospect on clinical diagnosis.

"Noisy patients with known disease" dataset contains noisy phenotypes which are not annotated phenotype terms of the disease. We applied *DisPheno* and other five approaches on the noisy dataset, and our method performed the best in all the six measurements (Fig. 5). On the noisy patients with known diseases, the performance of *DisPheno* is far superior than the other five

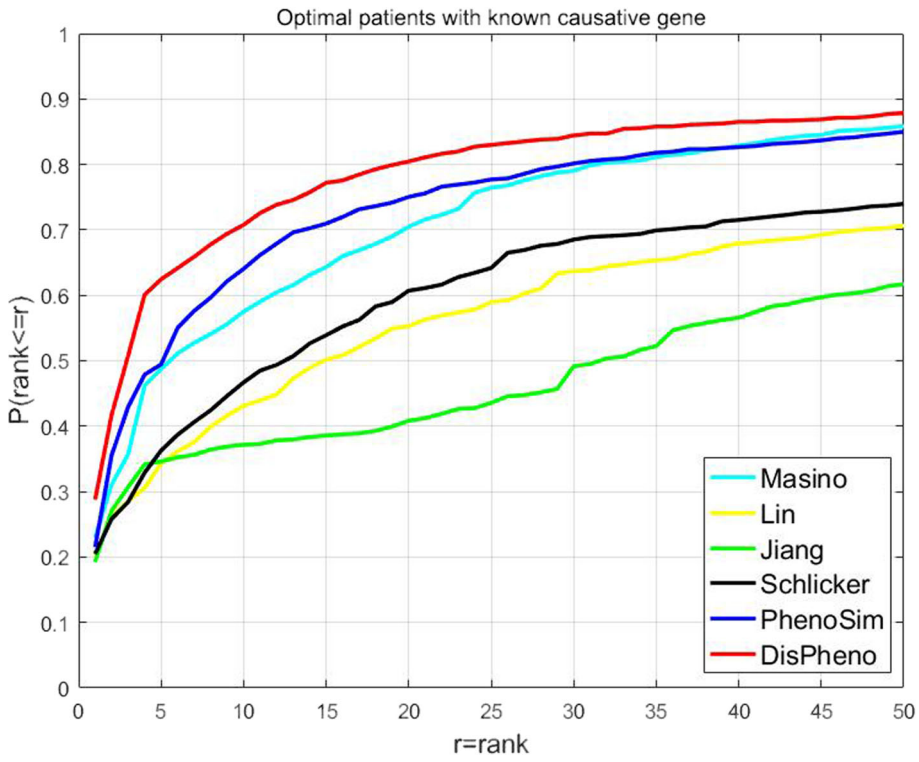

Fig. 3 Cumulative rank distribution of optimal patient dataset with the known causative gene. The x-axis is the rank threshold and the y-axis is the cumulative probability of true disease rank

Table 1 The percentage of cumulative rank distribution

Method	Top-1	Top-3	Top-5	Top-10
DisPheno	83.12%	99.10%	99.71%	99.87%
PhenoSim	79.50%	98.62%	99.45%	99.83%
Masino	82.48%	97.43%	98.63%	99.16%
Lin	95.68%	97.94%	98.63%	99.35%
Jiang	95.43%	98.17%	99.11%	99.69%
Schlicker	96.36%	98.31%	98.93%	99.53%

DisPheno was compared with other five methods on the optimal patient with the known disease

algorithms. 56.48% of candidate diseases rank the highest using *DisPheno*. Instead, the ratio of other five methods are 42.74% (PhenoSim), 20.04% (Masino), 0.5% (Lin), 0.32% (Jiang) and 1.82% (Schlicker). The second highest is PhenoSim, which is 13.74% less than *DisPheno*. The great gap shows the performance of our method in disease identification, especially on noisy simulated patient dataset.

Overall, *DisPheno* performs better than other five similarity measurements on the stimulated datset with noise phenotype terms, and it shows great robustness. It implies huge potential on clinical disease diagnosis.

Performance evaluation on noisy & imprecision dataset

Except noisy phenotype terms, clinical datasets often contains imprecision phenotypes. In this part, we performed *DisPheno* on the noisy and imprecision patient dataset with known disease to evaluate the performance respectively.

Compared with other five methods, *DisPheno* shows good and stable performance on simulated patients with noisy and imprecision phenotypes (see Table 2). The percentage of true disease rank among top-10 using *DisPheno* reaches 22.34%, which is much higher than others. It indicates that *DisPheno* would perform well on the clinical datasets and it shows great prospects on disease diagnosis.

Effects of parameters on DisPheno model

In this part, we test the various parameters on *DisPheno* model. In the first part of our model, we utilize both gene and disease annotations. We run *DisPheno* multiple times by varying the parameter w from 0.0 to 1.0 to test the performance of different weighted coefficients. Figure 6 shows that *DisPheno* achieves the best performance when the weighted coefficient is equal to 0.5 or 0.9.

Besides, we also run different parts of *DisPheno* to evaluate the contribution of different components in the model. Compared with previous algorithm *PhenoSim*, this novel model mainly adds four parts to improve the

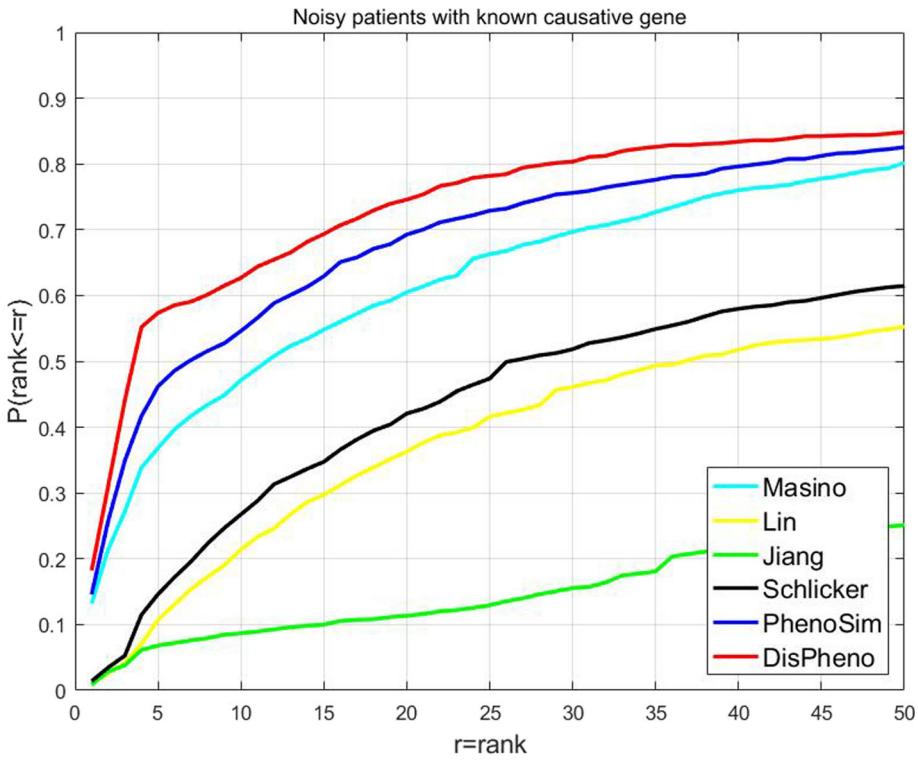

Fig. 4 Cumulative rank distribution of noisy patient dataset with the known causative gene. The x-axis is the rank threshold and the y-axis is the cumulative probability of true disease rank

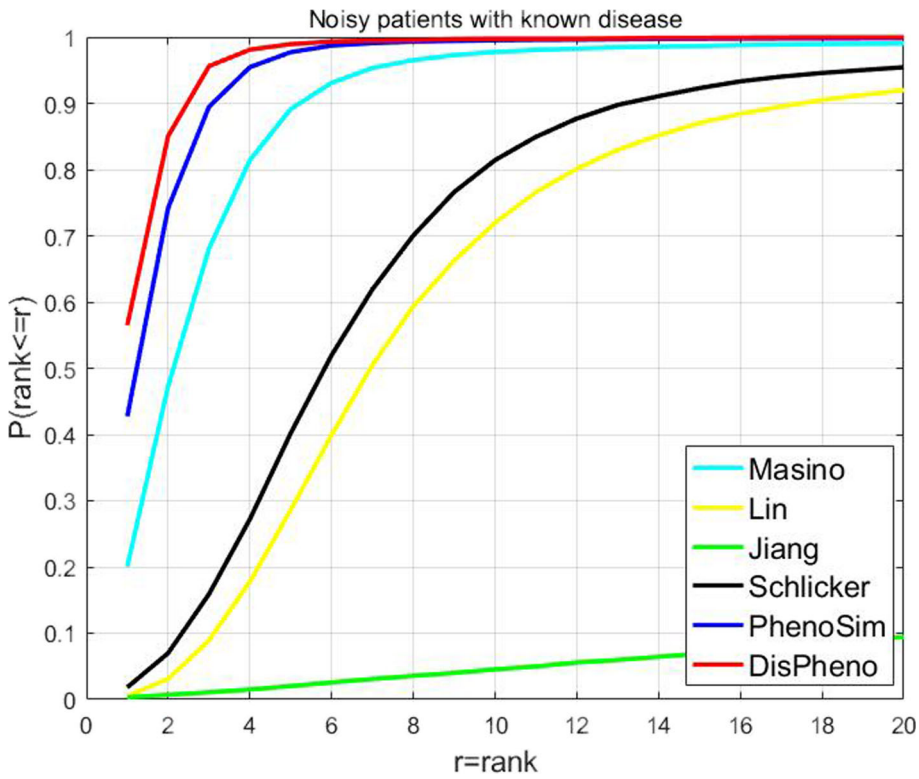

Fig. 5 Cumulative rank distribution of noisy patient dataset with the known disease. The x-axis is the rank threshold and the y-axis is the cumulative probability of true disease rank

performance of identifying true disease. First part is utilizing both gene and disease to annotate phenotype terms, named as *Anno*. Second part of our model mainly consider the effect of the distance between two phenotype terms, thus we add $(1 - dist(t_i, t_j)/mostDepth)$ in the process of calculating phenotype term similarity, named as *Depth*. Besides, we utilize TF-IDF and Cosine Similarity to measure the similarity between any two phenotype terms based on their definitions. We then add term definition similarities into phenotype topological structure, and convert original directed acyclic graph into a weighted directed acyclic graph. This part is named as *Weight*. In the part of calculating phenotype term similarity, we calculate PMI matrix to measure the association of phenotype terms. This step is named as *PMI*. We run our model with different single part to evaluate the performance of *DisPheno*. Figure 7 shows that each part of *DisPheno* contributes to improve the performance of identifying true disease from disease candidate sets. From this experimental results, we can find that the phenotype annotation method, distance between two phenotype, definition of phenotype term and association of phenotype sets are all critical to phenotype similarity measure and it could significantly improve the performance of disease diagnosis.

Table 2 The percentage of cumulative rank distribution

Method	Top-10	Top-20	Top-30	Top-40	Top-50
DisPheno	22.34%	36.00%	44.55%	52.55%	58.07%
PhenoSim	3.86%	11.76%	20.69%	29.03%	36.49%
Masino	7.12%	23.76%	38.21%	48.89%	56.57%
Lin	2.14%	8.16%	15.16%	21.77%	27.98%
Jiang	1.66%	2.57%	3.45%	4.32%	5.25%
Schlicker	1.89%	6.88%	13.67%	20.34%	26.78%

DisPheno was compared with other five methods on the noisy & imprecision patient with the known disease

Performance evaluation on gene and disease similarity

To further test the performance of *DisPheno*, we also apply our method on similarity measurement of gene and disease. Each gene or disease can be annotated by a set of phenotype terms. Therefore, gene or disease similarity measurement can be translated into a task of measuring phenotype set similarity. We run our method *DisPheno* on a gene set and a disease set. Both of the two sets contain 20 genes or diseases. We use venn diagram to show

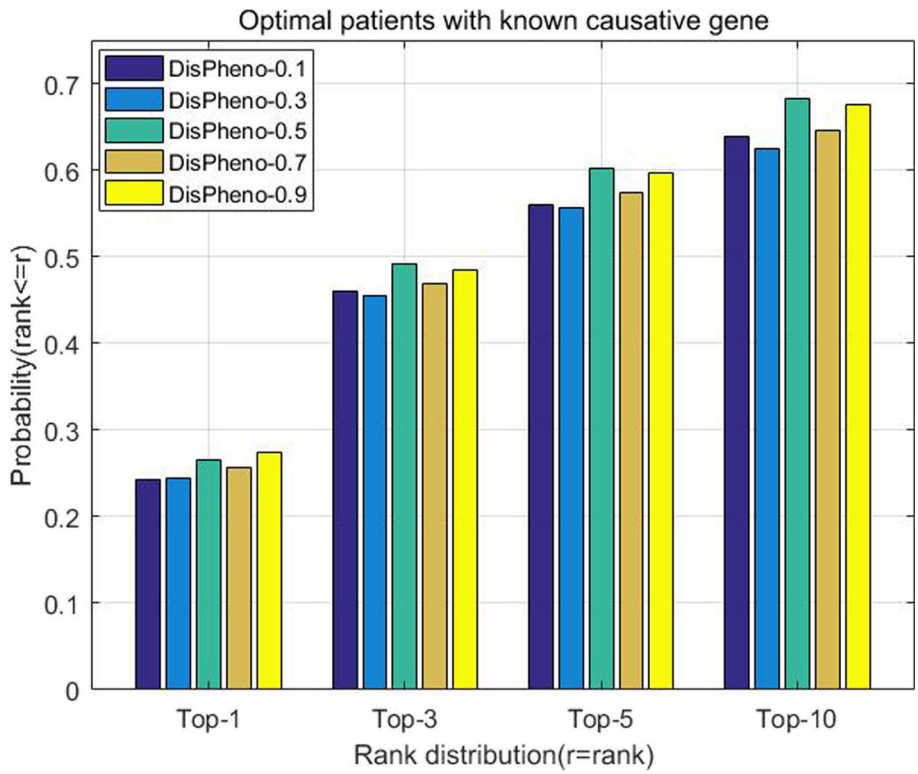

Fig. 6 The histogram of cumulative rank distribution with different weighted coefficient between gene and disease annotations. The x-axis is the four different top ranks and the y-axis is the cumulative probability of true disease rank. When the weighted coefficient is 0.5 or 0.9, DisPheno achieves better performance on the optimal patients with known causative gene dataset

the experimental results of five measurements (DisPheno, PhenoSim and other three methods randomly selected from Masino, Jiang, Lin and Schlicker). In detail, we firstly rank gene or disease pairwise similarities calculated by all five methods. Then, we calculate the intersection of top-20 gene pairs or disease pairs, and visualize the result by venn diagram.

The venn diagram (Fig. 8) shows that *DisPheno* is slightly better than other similarity measurements. We compare DisPheno and PhenoSim with other three methods which randomly selected from four phenotype similarity measurements. In the task of gene similarity calculation, the top-20 gene pairs of *DisPheno* are all part of others. In contract, *PhenoSim* contains 2 or 4 gene pairs which do not belong to any intersection. Similarity, *DisPheno* has fewer single disease-pairs than others in the task of measuring disease similarity.

Besides, we used the visualization tool of *PhenoSimWeb* to visualize the disease and gene set similarity [25]. PhenoSimWeb is an online application which can be used to calculated phenotype, gene and disease similarity. It also can predict disease and causative gene based on the input phenotype set. PhenoSimWeb contains other useful tools, such as text description translator and visualization interface. And the visualization interface of disease set similarity calculated by *DisPheno* is shown in Fig. 9. The main panel is the terms association network, where nodes represent disease terms and edges represent similarities between diseases. The upper left is the mini control panel, where you can adjust threshold and visual layout. The lower left part is the overall distribution of similarity scores. The upper right shows the neighborhood of selected disease term "OMIM:601894". This visualization webpage provides user a clear and convenient way to analysis the results of disease similarity.

PhenoSimWeb is an online phenotype similarity calculating and visualizing application, which currently contains five phenotype similarity measurements, including PhenoSim, Masino, Jiang, Lin and Schlicker. And in this paper, we propose a novel HPO-based phenotype similarity method. We will add our method *DisPheno* into the online tool *PhenoSimWeb* and enrich phenotype similarity measurement of this web application in the future.

Conclusions

The high-speed development of biological techniques such as next generation sequencing has greatly improved efficiency of cancer prediction and disease diagnosis.

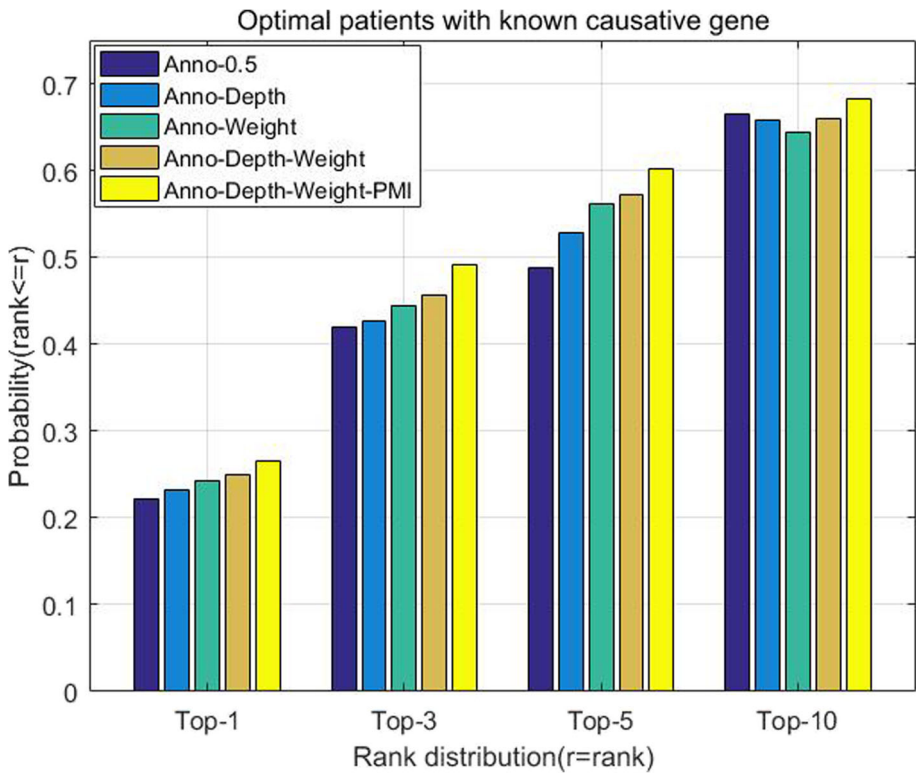

Fig. 7 The histogram of cumulative rank distribution with different parts of *DisPheno*. The x-axis is the four different top ranks and the y-axis is the cumulative probability of true disease rank. The navy blue bar is *PhenoSim* method with part *Anno* and the weighted coefficient is 0.5. The sky blue contains part *Depth* based on previous step. The green one contains part *Weight*. The orange one combines part *Depth* and *Weight*. The yellow bar is the method *DisPheno* which performs better than others

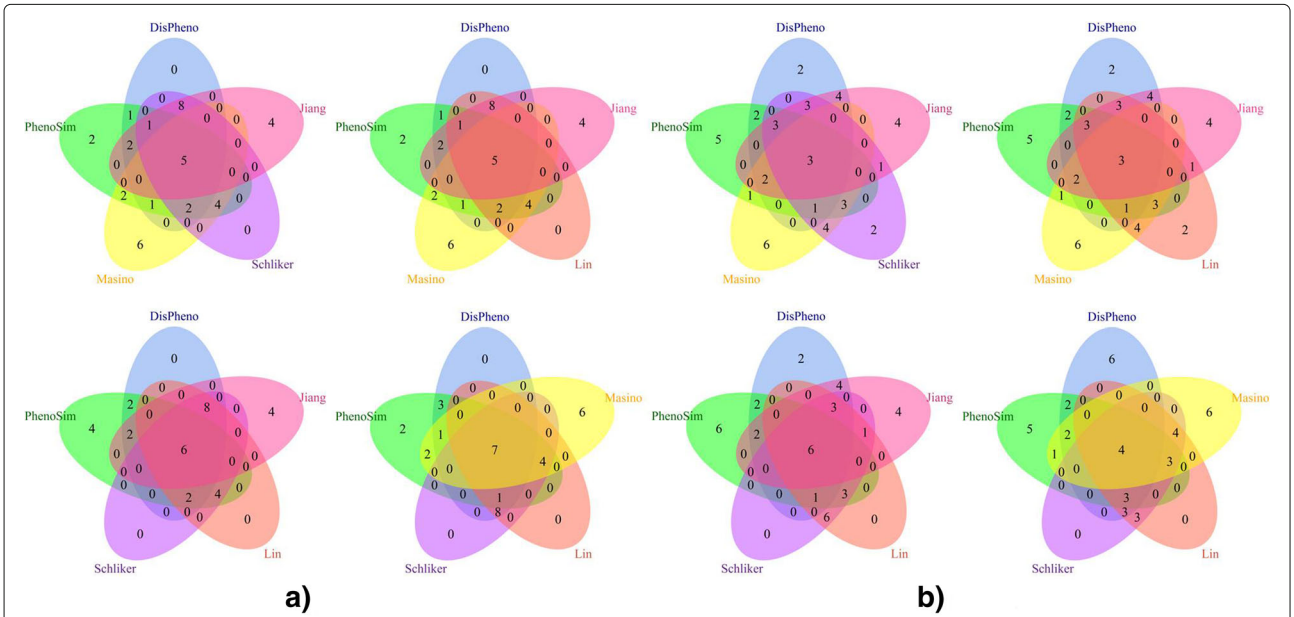

Fig. 8 The venn diagram of top-20 gene and disease pairwise similarity. The blue and green are *DisPheno* and *PhenoSim*. The purple, tomato, yellow and red are *Schlicker*, *Lin*, *Masino* and *Jiang* respectively. From the intersection of venn figure, *DisPheno* performs better than other methods on task of gene and disease similarity measurement. For instance, the upper-left venn diagram shows that there are 5 pairwise genes are included in all methods' results. All top-20 pairwise genes of DisPheno are contained by others. In contrast, there are 2 (PhenoSim), 4 (Jiang) and 6 (Masino) pairwise genes not belongs to any intersections. **a** Gene **b** Disease

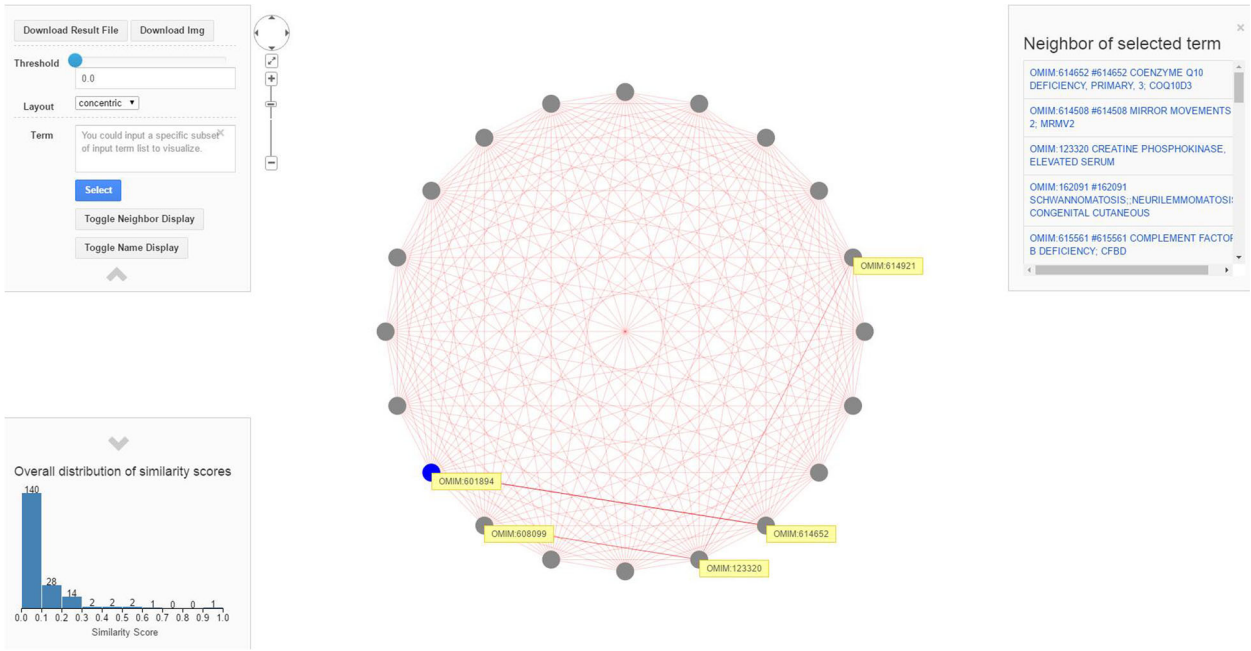

Fig. 9 Disease similarity using visualization interface of *PhenoSimWeb*. We ran *DisPheno* on a disease dataset and calculated the pair-wise similarity of these diseases. Then, we visualized the results using the visualization interface of PhenoSimWeb. The main panel is the diseases association network, where nodes represent disease terms and edges represent similarities between diseases. The upper left is the mini control panel, where you can adjust threshold and choose different visual layout. The lower left part is the overall distribution of similarity scores. The upper right shows the neighborhood of selected disease term "OMIM:601894"

However, intricate phenotype ontology and high genetic heterogeneity have stunted further improvement of disease identification. As an useful and powerful tool, HPO-based phenotype semantic similarity could fill this gap and accelerate the disease diagnosis effectively. In this paper, we proposed an unique and novel phenotype similarity measurement, called *DisPheno*, which integrates multiple types of information: hierarchical structure, phenotype term annotation and text description. Compared with existing five state-of-art methods on the optimal and noisy datasets, our method performs much better than the others. In summary, *DisPheno* accelerates the efficiency of disease identification significantly and it also shows greatly potentiality in practical clinical studies.

Abbreviations
DAG: Directed acyclic graph; HPO: Human phenotype ontology; IC: Information Content; PMI: Point-wise mutual information; TF-IDF: Term frequency-inverse document frequency

Acknowledgments
We thank all anonymous reviewers.

Funding
Publication of this article was sponsored by National Natural Science Foundation of China (No. 61702421, 61332014, 61772426), China Postdoctoral Science Foundation (No. 2017M610651), Fundamental Research Funds for the Central Universities (No. 3102018zy033), Top International University Visiting Program for Outstanding Young Scholars of Northwestern Polytechnical University.

Availability of data and materials
Not applicable.

About this supplement
This article has been published as part of *BMC Systems Biology Volume 13 Supplement 2, 2019: Selected articles from the 17th Asia Pacific Bioinformatics Conference (APBC 2019): systems biology*. The full contents of the supplement are available online at https://bmcsystbiol.biomedcentral.com/articles/supplements/volume-13-supplement-2.

Authors' contributions
JP and XS designed the algorithm framework; HX implemented the algorithm; JP and HX wrote this manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Published: 5 April 2019

References
1. De Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, Kroes T, et al. Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med. 2012;367(20):1921–9.

2. Yang Y, Muzny DM, Xia F, Niu Z, Person R, Ding Y, et al. Molecular findings among patients referred for clinical whole-exome sequencing. Jama. 2014;312(18):1870–9.
3. Study TDDD. Large-scale discovery of novel genetic causes of developmental disorders. Nature. 2015;519(7542):223–8.
4. Zemojtel T, Köhler S, Mackenroth L, Jäger M, Hecht J, Krawitz P, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. Sci Transl Med. 2014;6(252): 252ra123.
5. Groza T, Köhler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, et al. The human phenotype ontology: semantic unification of common and rare disease. Am J Hum Genet. 2015;97(1):111–24.
6. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The Human Phenotype Ontology: a tool for annotating and analyzing human hereditary disease. Am J Hum Genet. 2008;83(5):610–5.
7. Peng J, Hui W, Shang X. Measuring phenotype-phenotype similarity through the interactome. BMC Bioinformatics. 2018;19(S5):114.
8. Peng J, Li Q, Shang X. Investigations on factors influencing HPO-based semantic similarity calculation. J Biomed Semant. 2017;8(1):34.
9. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet. 2009;85(4):457–64.
10. Washington NL, Haendel MA, Mungall CJ, Ashburner M, Westerfield M, Lewis SE. Linking human diseases to animal models using ontology-based phenotype annotation. PLoS Biol. 2009;7(11):e1000247.
11. Hoehndorf R, Schofield PN, Gkoutos GV. PhenomeNET: a whole-phenome approach to disease gene discovery. Nucleic Acids Res. 2011;39(18):e119.
12. Masino AJ, Dechene ET, Dulik MC, Wilkens A, Spinner NB, Krantz ID, et al. Clinical phenotype-based gene prioritization: an initial study using semantic similarity and the human phenotype ontology. BMC Bioinformatics. 2014;15(1):1.
13. Pesquita C, Faria D, Bastos H, Ferreira AE, Falcao AO, Couto FM. Metrics for go based protein semantic similarity: a systematic evaluation. BMC Bioinformatics. 2008;9(5):4.
14. Peng J, Zhang X, Hui W, Lu J, Li Q, Liu S, et al. Improving the measurement of semantic similarity by combining gene ontology and co-functional network: a random walk based approach. BMC Syst Biol. 2018;12(1):18.
15. Peng J, Li H, Liu Y, Juan L, Jiang Q, Wang Y, et al. InteGO2: a web tool for measuring and visualizing gene semantic similarities using gene ontology. BMC Genomics. 2016;17(5):530.
16. Cheng L, Jiang Y, Wang Z, Shi H, Sun J, Yang H, et al. DisSim: an online system for exploring significant similar diseases and exhibiting potential therapeutic drugs. Sci Rep. 2016;6:30024.
17. Peng J, Uygun S, Kim T, Wang Y, Rhee SY, Chen J. Measuring semantic similarities by combining gene ontology annotations and gene co-function networks. BMC Bioinformatics. 2015;16(1):1.
18. Peng J, Wang H, Lu J, Hui W, Wang Y, Shang X. Identifying term relations cross different gene ontology categories. BMC Bioinformatics. 2017;18(16):573.
19. Teng Z, Guo M, Liu X, Dai Q, Wang C, Xuan P. Measuring gene functional similarity based on group-wise comparison of GO terms. Bioinformatics. 2013;29(11):1424–1432.
20. Caniza H, Romero AE, Heron S, Yang H, Devoto A, Frasca M, et al. GOssTo: a stand-alone application and a web tool for calculating semantic similarities on the Gene Ontology. Bioinformatics. 2014;30(15):2235–6.
21. Garla VN, Brandt C. Semantic similarity in the biomedical domain: an evaluation across knowledge sources. BMC Bioinformatics. 2012;13(1):261.
22. Batet M, Sánchez D, Valls A. An ontology-based measure to compute semantic similarity in biomedicine. J Biomed Inform. 2011;44(1):118–25.
23. Peng J, Xue H, Shao Y, Shang X, Wang Y, Chen J. A novel method to measure the semantic similarity of HPO terms. Int J Data Min Bioinform. 2017;17(2):173–88.
24. Deng Y, Gao L, Wang B, Guo X. HPOSim: an R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. PloS ONE. 2015;10(2):e0115692.
25. Peng J, Xue H, Hui W, Lu J, Chen B, Jiang Q, Shang X, Wang Y. An online tool for measuring and visualizing phenotype similarities using hpo. BMC Genomics. 2018;19(S6):571.
26. Jiang JJ, Conrath DW. Semantic similarity based on corpus statistics and lexical taxonomy. 1997709008. arXiv preprint cmp-lg/9.
27. Lin D. An information-theoretic definition of similarity. In: Proceedings of the Fifteenth International Conference on Machine Learning, ICML '98. San Francisco: Morgan Kaufmann Publishers Inc.; 1998. p. 296–304.
28. Wang JZ, Du Z, Payattakool R, Philip SY, Chen CF. A new method to measure the semantic similarity of GO terms. Bioinformatics. 2007;23(10): 1274–81.
29. Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics. 2006;7(1):1.
30. Garcelon N, Neuraz A, Salomon R, Bahi-Buisson N, Amiel J, Picard C, et al. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. Orphanet J Rare Dis. 2018;13(1):85.
31. Malone BM, Perkins AD, Bridges SM. Integrating phenotype and gene expression data for predicting gene function. BMC Bioinformatics. 2009;10 Suppl 11(Suppl 11):S20.
32. Kumar AA, Van LL, Alaerts M, Ardeshirdavani A, Moreau Y, Laukens K, et al. pBRIT: Gene Prioritization by Correlating Functional and Phenotypic Annotations Through Integrative Data Fusion. Bioinformatics. 2018;34(13): 2254–2262.
33. Jing LP, Huang HK, Shi HB. Improved feature selection approach TFIDF in text mining. In: Proceedings. International Conference on Machine Learning and Cybernetics vol. 2. Beijing: IEEE; 2002. p. 944–6.
34. Church KW, Hanks P. Word association noms, Mutual Information, and lexicography. Comput Linguis. 1990;16(1):76–83.
35. Peng J, Xue H, Shao Y, Shang X, Wang Y, Chen J. Measuring phenotype semantic similarity using Human Phenotype Ontology. In: bioinformatics and biomedicine (BIBM), 2016 IEEE international conference on Shenzhen. IEEE; 2016. p. 763–6.