frontiers | Frontiers in Genetics

TYPE Original Research
PUBLISHED 02 May 2023
DOI 10.3389/fgene.2023.1181391

OPEN ACCESS

EDITED BY
Wei Lan,
Guangxi University, China

REVIEWED BY
Xiaoshu Zhu,
Yulin Normal University, China
Liang Cheng,
Harbin Medical University, China

*CORRESPONDENCE
Jianwei Li,
lijianwei@hebut.edu.cn

RECEIVED 07 March 2023
ACCEPTED 11 April 2023
PUBLISHED 02 May 2023

CITATION
Li J, Li Z, Wang Y, Lin H and Wu B (2023), TLSEA: a tool for lncRNA set enrichment analysis based on multi-source heterogeneous information fusion . *Front. Genet.* 14:1181391. doi: 10.3389/fgene.2023.1181391

TLSEA: a tool for lncRNA set enrichment analysis based on multi-source heterogeneous information fusion

Jianwei Li[1,2]*, Zhiguang Li[1], Yinfei Wang[1], Hongxin Lin[1] and Baoqin Wu[1]

[1]Institute of Computational Medicine, School of Artificial Intelligence, Hebei University of Technology, Tianjin, China, [2]School of Electronic and Information Engineering, Hebei University of Technology, Tianjin, China

Long non-coding RNAs (lncRNAs) play an important regulatory role in gene transcription and post-transcriptional modification, and lncRNA regulatory dysfunction leads to a variety of complex human diseases. Hence, it might be beneficial to detect the underlying biological pathways and functional categories of genes that encode lncRNA. This can be carried out by using gene set enrichment analysis, which is a pervasive bioinformatic technique that has been widely used. However, accurately performing gene set enrichment analysis of lncRNAs remains a challenge. Most conventional enrichment analysis methods have not exhaustively included the rich association information among genes, which usually affects the regulatory functions of genes. Here, we developed a novel tool for lncRNA set enrichment analysis (TLSEA) to improve the accuracy of the gene functional enrichment analysis, which extracted the low-dimensional vectors of lncRNAs in two functional annotation networks with the graph representation learning method. A novel lncRNA–lncRNA association network was constructed by merging lncRNA-related heterogeneous information obtained from multiple sources with the different lncRNA-related similarity networks. In addition, the random walk with restart method was adopted to effectively expand the lncRNAs submitted by users according to the lncRNA–lncRNA association network of TLSEA. In addition, a case study of breast cancer was performed, which demonstrated that TLSEA could detect breast cancer more accurately than conventional tools. The TLSEA can be accessed freely at http://www.lirmed.com:5003/tlsea.

KEYWORDS

lncRNA, functional enrichment analysis, heterogeneous network representation learning, lncRNA–lncRNA association network, random walk with restart, web server

1 Introduction

The central principle of molecular biology has proposed that RNA is an intermediary between protein-coding genes and proteins. However, genes encoding proteins only account for 1.5% of the human genome, and more than 98% of the human genome does not encode proteins. Most of these non-protein-coding genes were transcribed into non-coding RNAs (ncRNAs) (Guttman et al., 2009; Xue et al., 2017; DiStefano, 2018). These ncRNAs were often considered as “noise” of genome transcription and were not associated with any biological functions for decades. According to the length of the nucleotide sequence, ncRNAs

can be further divided into small ncRNAs (<200 nucleotides) and long non-coding RNAs (>200 nucleotides, lncRNAs) (Chen et al., 2016; McDonel and Guttman, 2019). Although lncRNAs are not directly translated into proteins, their complex and diverse functions have helped gain insights into several biological processes in humans. As a novel class of ncRNAs, their functional studies received great interest, and considerable progress has been made in exploring lncRNA biology. Since the discovery of the lncRNAs H19 and XIST in the early 1990s (Tsang and Kwok, 2007; Li et al., 2013), substantial evidence has suggested that lncRNAs play an important role in the regulation of several processes, including epigenetic process, cell cycle, cell differentiation, and transcription mode (Gloss and Dinger, 2016; Kashi et al., 2016; Kopp and Mendell, 2018). With the rapid development of scientific methodology and experimental technology, researchers have identified thousands of lncRNAs that play important roles in many basic and key biological processes in eukaryotes, from nematodes to humans (Munos, 2009; Lalevée and Feil, 2015). Moreover, lncRNAs also play a key regulatory role in the occurrence and development of complex human diseases (Engreitz et al., 2016; Fang and Fullwood, 2016), such as breast cancer (Niknafs et al., 2016), non-small-cell lung cancer (Hua et al., 2019), gastric cancer (Liu et al., 2015), and cardiovascular diseases (Uchida and Dimmeler, 2015). Mutations or disorders of lncRNAs are closely related to many human diseases. For example, MALAT1 (or NEAT2) is upregulated in non-small-cell lung cancer and can be used as a biomarker for early cancer prognosis (Gutschner et al., 2013), and the use of lncRNA HOTAIR has been explored as a potential biomarker for detecting recurrence of hepatocellular carcinoma (Topel et al., 2020). With the advancement of high-throughput sequencing technology, more lncRNA gene sets have been produced as a result of data analysis using high-throughput experiments. However, it is still challenging to find how the associations between lncRNAs in one set can be used to develop a comprehensive understanding of the biological regulatory functions of lncRNA gene sets. Moreover, it is important to assess how the regulatory function of lncRNA lists of interest submitted by users on a large scale can be more accurately analyzed in the face of a large amount of omics data (Wang and Krishnan, 2014; Fillinger et al., 2019). We believe that the two aforementioned issues can be solved using the lncRNA set function enrichment analysis method, which identifies the importance of biological functions that are overrepresented in a long list with respect to their role in the whole human genome. The lncRNA set function enrichment analysis method has become an important research area in the field of lncRNA regulatory function research.

Gene set enrichment analysis was used to determine whether a group of genes with common characteristics (such as differential expression) were enriched on a certain functional pathway based on a gene set rather than a single gene, which would increase the reliability of gene function prediction (Del Giacco and Cattaneo, 2012). During the calculation, the gene set functional enrichment analysis method integrated data from different levels and sources and provided important insights for constructing characteristic gene modules and molecular regulatory networks in different physiological and pathological states. To date, dozens of gene set functional enrichment analysis methods have been developed that can be divided into four categories based on their data sources and execution algorithms.

The first category is over-representation analysis (ORA) methods, which are early and conventional enrichment analysis methods. Such methods intersect a group of genes of the user's interest (called a gene list) with the background gene sets, count common genes as hit numbers, and evaluate whether the background gene set is significantly enriched in the gene list using statistical methods (Khatri et al., 2012). At present, there are many online tools and software that provide overexpression analysis, such as DAVID (Jiao et al., 2012), GOstats (Beissbarth and Speed, 2004), and GenMAPP (Doniger et al., 2003). In addition, LncSEA is an online tool that can enrich and analyze lncRNA lists using over-representation analysis methods. ORA methods are robust, reliable, and widely used. However, their limitations are also obvious, which hamper their application. The second category is functional class scoring (FCS) methods. Many FCS methods have been proposed, of which GSEA (Mootha et al., 2003; Subramanian et al., 2005) is the most commonly used one. FCS methods treat each lncRNA equally and in isolation, and the feature information of each gene within the background gene set and the associations with other genes are both neglected, which could be an obstacle to seeking more insightful biological processes for researchers. The third category is path topology (PT) methods. In the biological pathways, genes usually affect the biological processes of cells through complex relationships. Pathway-Express (Draghici et al., 2007) was the first PT method. SPIA (Tarca et al., 2009) introduced the concept of regulation intensity of each regulation relationship in a pathway based on retaining influencing factors. TopoGSA (Glaab et al., 2010) adopted the centripetality parameters of pathways while comparing differences between pathways. Currently, only the KEGG database (Kanehisa and Goto, 2000; Kanehisa et al., 2014) provides a comprehensive path topology. The fourth category is network topology (NT) methods. The key idea of NT methods is to convert the functional enrichment analysis problem of the gene list of interest into the functional enrichment analysis problem of gene pairs based on functional annotation networks. The most comprehensive and typical example of this category is the network ontology analysis (NOA) method (Wang et al., 2011). NT methods adopt gene importance and association information at the system level, overcome the defect that core genes are ignored due to small differential expression, and can make more accurate and reliable predictions. Therefore, NT methods are recommended when there are suitable gene function annotation networks, and they have become one of the mainstream methods associated with functional enrichment analysis at present.

Owing to the construction and integration of lncRNA-associated networks, NT methods based on gene network topology cannot be directly applied to lncRNA functional enrichment analysis. Inspired by the miRNA similarity network based on disease association, we constructed two lncRNA similarity networks based on miRNA–lncRNA associations and lncRNA–disease associations. These multi-source functional annotation networks provide an important basis for lncRNA functional enrichment analysis.

In this study, we aimed to develop a novel tool for lncRNA set enrichment analysis (TLSEA) to improve the accuracy of lncRNA set enrichment analysis. A flowchart of the TLSEA model is shown in Figure 1. First, two lncRNA functional similarity networks were constructed; the first network was based on lncRNA–miRNA

10.3389/fgene.2023.1181391

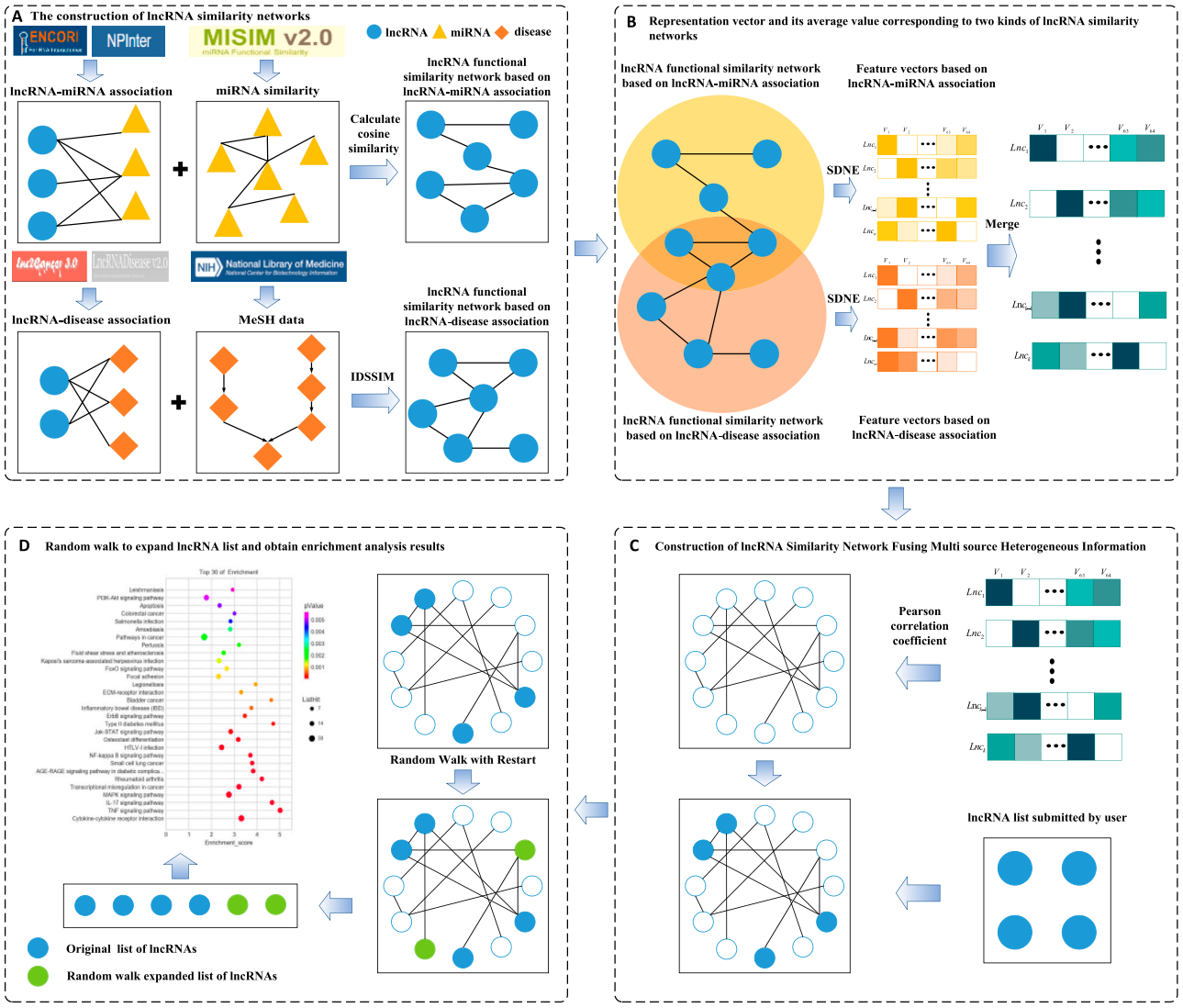

FIGURE 1
Flowchart of the TLSEA model. **(A)** Two lncRNA similarity networks were constructed based on lncRNA–miRNA associations and lncRNA–disease associations, respectively. **(B)** The SNDE method was used to extract 64-dimensional vectors from two networks, respectively. If an lncRNA appeared in two similar networks, its average value was taken as the final low-dimensional vector value. **(C)** The Pearson correlation coefficient was adopted to obtain the lncRNA similarity network that integrated multi-source heterogeneous information, and the differential expression list was used as the seed node of the network. **(D)** The expanded lncRNA list was obtained with network random walk with the restart method and was merged with the original lncRNA list; functional enrichment analysis was performed on it.

associations by integrating the miRNA functional similarity network, and the second network was based on the lncRNA–miRNA association network and based on lncRNA–disease associations by integrating the disease semantic similarity network and lncRNA–disease association network, respectively. Second, to fuse a variety of heterogeneous lncRNA functional annotation networks and extract more feature information of lncRNA nodes, two sets of 64-dimensional vectors were created with the graph embedding algorithm structural deep network embedding (SDNE) based on the two lncRNA functional similarity networks and were merged into a new set of 64-dimensional vectors. If an lncRNA appeared in only one of the two functional annotation networks, its vector was retained as the merged vector. If one lncRNA appeared in both functional annotation networks, the average value of the two corresponding vectors was taken as the merged vector value. After merging, a feature matrix was obtained, in which each lncRNA node corresponded to a row vector. Third, the lncRNA–lncRNA association network was constructed by calculating the similarity of each corresponding lncRNA vector pair. Fourth, based on the lncRNA–lncRNA association network, a novel lncRNA functional enrichment analysis model could perform a more comprehensive and accurate enrichment analysis on the lncRNA list submitted by users from the two aspects of regulation function and disease association. It employed a network random walk with the restart method to enrich the lncRNA list prior to functional enrichment analysis. Our model mapped the nodes in the lncRNA list to the lncRNA–lncRNA association network as random walk seed nodes,

and the lncRNA nodes closely associated with the subnet in the lncRNA–lncRNA association network were identified. Finally, both the lncRNA list submitted by the user and the expanded list were merged into a new lncRNA list, and functional enrichment analysis was performed.

2 Materials and methods

2.1 Datasets

Our study mainly included lncRNA–miRNA association data, lncRNA–disease association data, and lncRNA expression profile data. The lncRNA–miRNA association data, which were previously confirmed experimentally, were downloaded from the NPInter v4.0 (Teng et al., 2020) (see Supplementary Table S1) and ENCORI (Li et al., 2014) (see Supplementary Table S2) databases, and lncRNA symbols were obtained from the HGNC database (Seal et al., 2022). Thereafter, the lncRNA–miRNA association data obtained from the two databases were merged, and duplicate pairs were removed. LncRNA–miRNA pairs with non-standard naming formats were deleted, and the naming formats of both miRNA precursors and mature bodies were standardized. Finally, 18,033 validated lncRNA–miRNA associations were obtained between 1,002 lncRNAs and 437 miRNAs. The miRNA similarity data that were previously confirmed by experiments were downloaded from MISIM v2.0 (Li et al., 2019) and included 1,044 miRNAs (see Supplementary Table S3). The lncRNA–disease association data were downloaded from the LncRNADisease v2.0 (Bao et al., 2019) (see Supplementary Table S4) and Lnc2Cancer v3.0 (Gao et al., 2019) (Supplementary Table S5) databases. Thereafter, lncRNA functional similarity was calculated using disease semantic similarity; for this, all disease names were standardized according to the MeSH vocabulary (Baumann, 2016). The lncRNA–disease associations that did not conform to HGNC were removed. Finally, 2,230 validated lncRNA–disease associations were obtained, involving 777 lncRNAs and 257 diseases. LncRNA expression profile data were downloaded from the NONCODE database (Fang et al., 2018). After converting NONCODE ID to HGNC and removing lncRNAs with no expression, we obtained 303 lncRNAs from 24 tissues or organs.

2.2 Disease semantic similarity network

Using a previous method that was based on improved disease semantic similarity (Fan et al., 2020), we adopted IDSSIM, a model to calculate the functional similarity of lncRNAs in TLSEA. Primarily, IDSSIM introduced the IC contribution factor into the semantic value calculation, which considered both the hierarchical structure of the directed acyclic graph (DAG) and the specificities of diseases. IDSSIM was superior to conventional models, such as LNCSIM1, LNCSIM2 and ILNCSIM. No consideration of the hierarchical structure of the directed acyclic graph was included in LNCSIM1. ILNCSIM and LNCSIM2 considered only the specificities of diseases.

The semantic similarity between two diseases could be calculated using their DAG, which was constructed by mapping the names of the two diseases to MeSH names. For a disease A, its DAG is expressed as $DAG_A = \{T_A, E_A\}$, where T_A is a collection of ancestor nodes of disease A and E_A is the set of all edges in the DAG. The disease term $t \in T_A$ in DAG_A had a semantic contribution to disease A, which was defined as the semantic value $SV_A^1(t)$ of t to disease A, calculated in LNCSIM1 (Chen et al., 2015) using the following formula:

$$SV_A^1(t) = \begin{cases} 1 & t = A \\ max\left(\Delta \times SV_A^1(t') \middle| t' \in C(t)\right) & t \neq A \end{cases}, \quad (1)$$

where $C(t)$ is a subset of t and Δ represents the semantic contribution factor of the edge connecting the linking disease term t with its child disease term t' in E_A, which is usually set to 0.5 (Wang et al., 2010). In natural language processing, inverse document frequency is used to evaluate the importance of words in a document. From formula (1), we conclude that the higher the frequency of a disease, the lower its speech contribution.

In addition, LNCSIM2 adopts another common formula to calculate the contribution of the disease term $t \in T_A$ in DAG_A to the semantic value $SV_A^2(t)$ of disease A:

$$SV_A^2(t) = -\log\frac{DAGs(t)}{D}, \quad (2)$$

where D is the number of diseases in MeSH and DAGs (t) is the number of DAGs that contain disease term t.

Using Equations 1, 2, the advantages of LNCSIM1 and LNCSIM2 methods were combined to calculate the semantic similarity of diseases. The contribution of the disease term t$\in T_A$ in DAG_A to the semantic value of disease A was calculated using the following equation:

$$SV_A^3(t) = \begin{cases} 1 & t = A \\ max\left((\Delta + P_t) \times SV_A^3(t') \middle| t' \in C(t)\right) & t \neq A \end{cases}. \quad (3)$$

Here, P_t is the IC contribution factor, which was calculated as follows:

$$P_t = \frac{\max_{k \in K}(DAGs(k)) - DAGs(t)}{D}, \quad (4)$$

where K is the set of all the diseases in the MeSH. It should be noted that for disease term t, the P_t value changed with the continuously updated versions of the MeSH.

The semantic value of disease A, $SV(A)$, was then calculated as the sum of the contributions of all disease terms in DAG_A to disease A:

$$SV(A) = \sum_{t \in T_A} SV_A^3(t). \quad (5)$$

Based on the intersection of the disease term set of diseases A and B, the semantic similarity of diseases A and B, $DSS(A, B)$, was defined as follows:

$$DSS(A, B) = \frac{\sum_{t \in T_A \cap T_B}\left(SV_A^3(t) + SV_B^3(t)\right)}{SV(A) + SV(B)}, \quad (6)$$

where T_A is a collection of ancestor nodes of disease A. $SV(A)$ is the sum of the contributions of all disease terms for disease A in DAG_A. $DSS(A, B)$ is the disease semantic similarity between diseases A and B.

10.3389/fgene.2023.1181391

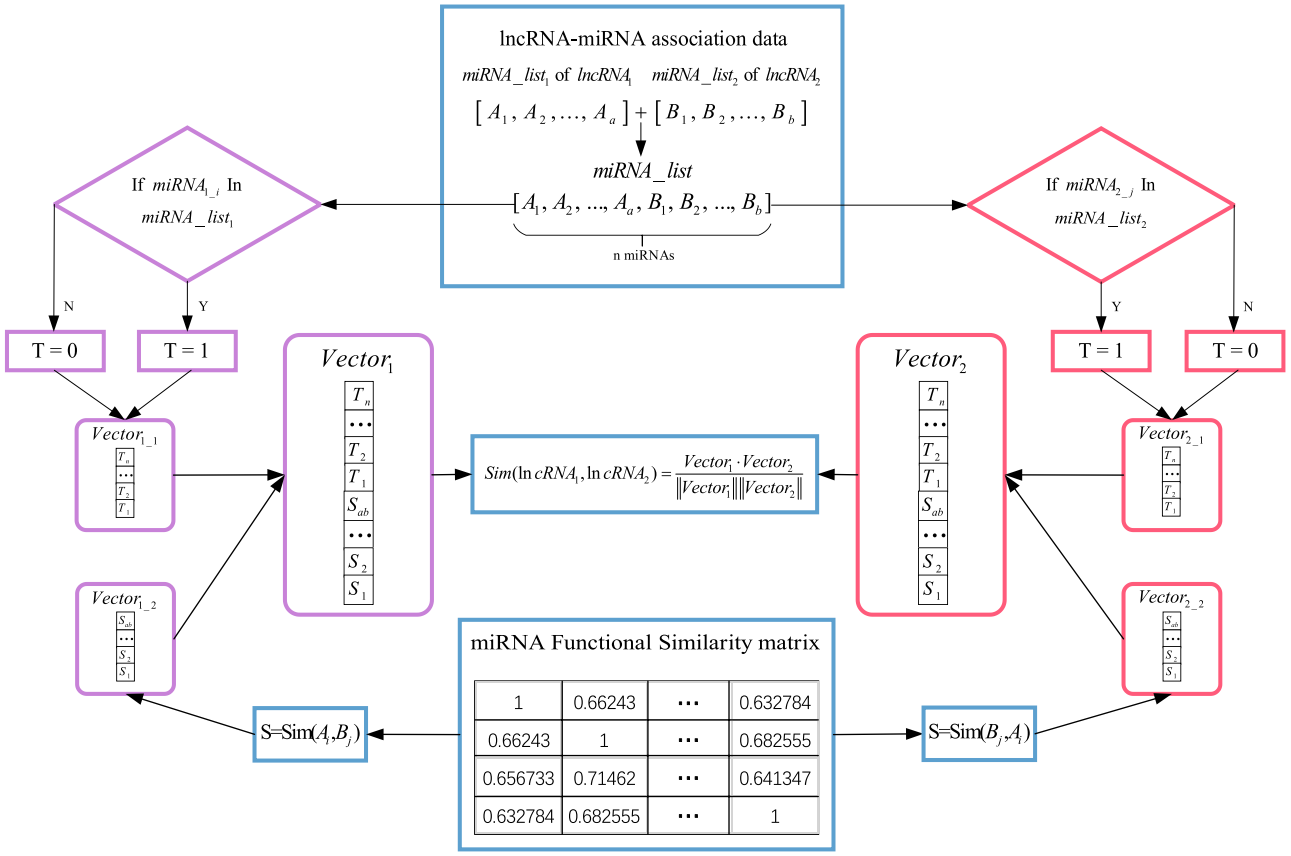

FIGURE 2
Flowchart of lncRNA similarity calculation based on lncRNA–miRNA associations.

2.3 LncRNA functional similarity network based on lncRNA–miRNA associations

Previous studies have confirmed that lncRNAs with more common target miRNAs may have a higher similarity. Based on this assumption, an lncRNA functional similarity network was constructed by integrating lncRNA–miRNA association data with miRNA similarity data in the present study. LncRNA–miRNA association data were downloaded from the NPInter v4.0 and ENCORI databases, and miRNA similarity data were obtained from MISIM v2.0. We constructed lncRNA feature vectors based on lncRNA–miRNA associations and calculated the association scores of the two lncRNAs using cosine correlation.

As shown in Figure 2, we calculated the functional similarity between $lncRNA_1$ and $lncRNA_2$ by utilizing the data collected from the databases to generate the characteristics of the two lncRNAs based on the shared target miRNAs and the functional similarity between the target miRNAs. As associations between lncRNAs and miRNAs include multiple relationships, one lncRNA can target more than one miRNA, and conversely, multiple lncRNAs may target the same miRNA. Moreover, lncRNAs with the same target genes have generally similar functions (Tay et al., 2014). Based on this evidence, the miRNAs associated with $lncRNA_1$ and $lncRNA_2$ were first sorted into lists $miRNA\_list_1$ and $miRNA\_list_2$, respectively. After removing duplicate elements, $miRNA\_list_1$ and $miRNA\_list_2$ were merged to create a new list $miRNA_{list}$, which contained all miRNAs interacting with $lncRNA_1$ and $lncRNA_2$. The number of $miRNA\_list$ was n. Then, two vectors, $Vector_{1\_1}$ and $Vector_{2\_1}$, were utilized as the first parts of the feature vectors of $lncRNA_1$ and $lncRNA_2$ to describe the common target miRNAs of $lncRNA_1$ and $lncRNA_2$. For each miRNA in the $miRNA\_list$, if it existed in the $miRNA\_list_1$, T was 1, and otherwise, 0. The values of T were added to $Vector_{1\_1}$ and $Vector_{2\_1}$. Thus, $Vector_{1\_1}$ and $Vector_{2\_1}$ are vectors composed of 1 and 0. The more the two lncRNAs that had common target miRNAs, the more the elements with the same position in $Vector_{1\_1}$ and $Vector_{2\_1}$ had the same value of 1. The higher the similarity between two vectors, the higher their functional similarity. Subsequently, miRNAs in $miRNA\_list_1$ and $miRNA\_list_2$ were extracted to form miRNA set A and miRNA set B, and the similarity values of Sam (a_i, B) and Sam (b_j, A) were calculated to form $Vector_{1\_2}$ and $Vector_{2\_2}$, respectively. $Vector_{1\_2}$ and $Vector_{2\_2}$ denote the functional similarity of the miRNAs that were targeted by $lncRNA_1$ and $lncRNA_2$. The higher the functional similarity of the associated miRNAs, the higher the similarity between $Vector_{1\_2}$ and $Vector_{2\_2}$. Finally,

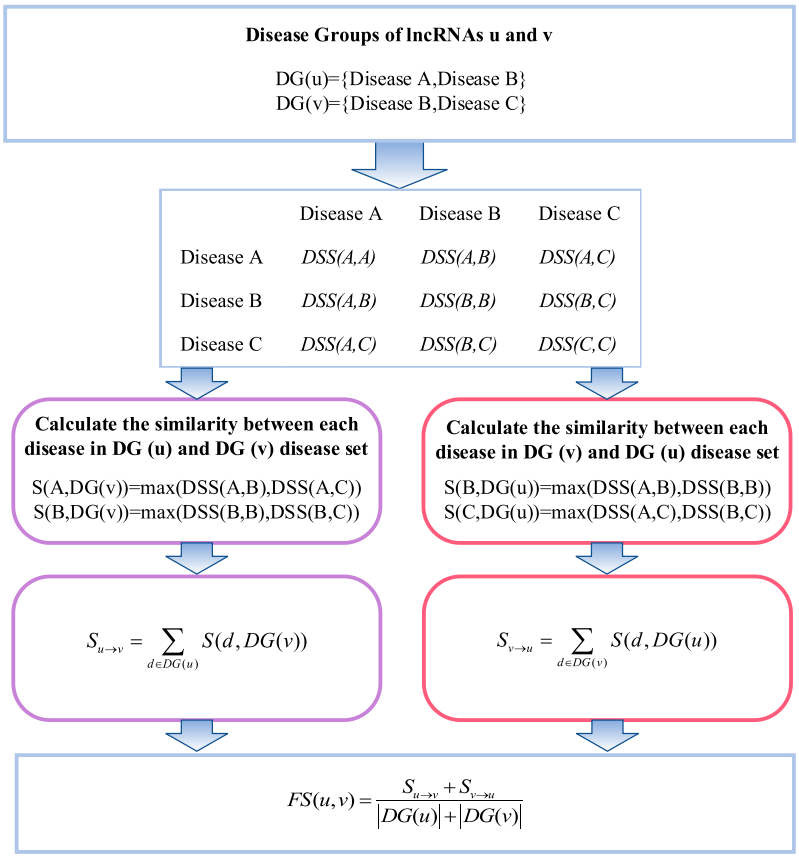

FIGURE 3
Flowchart of lncRNA function similarity calculation.

$Vector_{1\_1}$ and $Vector_{1\_2}$ are merged into $Vector_1$. Following this approach, $Vector_2$ was created. Ultimately, $Vector_1$ and $Vector_2$ represented $lncRNA_1$ and $lncRNA_2$, respectively. The cosine similarity formula was adopted to calculate the functional similarity scores of the $lncRNA_1$ and $lncRNA_2$ as follows:

$$Sim(lncRNA_1, lncRNA_2) = \frac{Vector_1 \cdot Vector_2}{\|Vector_1\|\|Vector_2\|}. \quad (7)$$

2.4 LncRNA functional similarity network based on lncRNA–disease associations

The calculation of lncRNA functional similarity is that lncRNAs related to similar diseases may have similar functions. LncRNA functional similarity can be calculated by integrating the semantic similarity of diseases and known lncRNA–disease association data. The flowchart of lncRNA functional similarity based on lncRNA–disease associations is shown in Figure 3, where $DG(u)$ and $DG(v)$ were defined as all disease sets related to lncRNA u and lncRNA v, respectively. The semantic similarity of each disease in $DG(u)$ and $DG(v)$ was used to calculate the lncRNA functional similarity between lncRNA u and lncRNA v.

Specifically, the similarity coefficient between one disease in the disease set corresponding to lncRNA u and all disease sets of lncRNA v was first calculated as follows:

$$S(d_u, DG(v)) = \max_{d\in DG(v)}(DSS(d_u, d)), \quad (8)$$

$$S(d_v, DG(u)) = \max_{d\in DG(u)}(DSS(d_v, d)), \quad (9)$$

where d_u and d_v represent disease in $DG(u)$ and $DG(v)$, DSS is the disease semantic similarity, and $S(d_u, DG(v))$ is the similarity between disease d_u and disease group $DG(v)$.

Thereafter, the coefficients of the disease set of lncRNA u and the disease set of lncRNA v were accumulated as

$$S_{u\to v} = \sum_{d\in DG(u)} S(d, DG(v)), \quad (10)$$

$$S_{v\to u} = \sum_{d\in DG(v)} S(d, DG(u)). \quad (11)$$

Finally, the functional similarity between lncRNA u and lncRNA v was defined as

$$FS(u, v) = \frac{S_{u\to v} + S_{v\to u}}{|DG(u)| + |DG(v)|}, \quad (12)$$

where the operator |.| represents the total number of diseases corresponding to the disease sets.

2.5 LncRNA functional similarity network based on expression profiles

Functionally interacting genes tend to exhibit similar expression profiles, thereby providing a theoretical basis for the calculation of lncRNA similarity with lncRNA expression profile data. Therefore, lncRNA functional similarity imputation methods based on expression profiles have been used in lncRNA function research. Analysis of lncRNA characteristics indicated that lncRNAs have significant tissue specificity and are conserved in mammals. The expression profiles of lncRNAs vary between different tissues and change at different growth stages in the same tissue or organ (Zhu et al., 2014). In the present study, the lncRNA expression profiles of 24 tissues and organs were downloaded from the NONCODE database. Each item had 24 dimensions, representing the expression profiles of this lncRNA in 24 tissues or organs. In the present study, some items missing the expression profile were deleted, and the naming format of the lncRNA was standardized. Finally, 303 lncRNA expression profiles were obtained. The correlation between these lncRNAs was analyzed using the Spearman correlation analysis method, and the Spearman correlation coefficient between two lncRNAs was adopted to determine their similarity.

2.6 Graph embedding methods

Currently, many graph embedding methods have been proposed to discover novel proper mapping functions to convert graph data that are usually high-dimensional sparse matrices to low-dimensional dense vectors. They maintained the proximity of these low-dimensional vector representations to solve the conundrum, which was difficult to consider using machine learning algorithms. Hence, graph embedding methods, such as node classification, link prediction, and association mining, have been used for mining biological information.

Existing graph embedding models are generally divided into five categories according to their algorithm principles: graph embedding based on matrix decomposition, graph embedding based on random walk, graph embedding based on self-encoder, graph embedding based on graph neural networks (GNNs) (Scarselli et al., 2009), and graph embedding based on other methods. In our study, we adopted four types of prevailing graph embedding algorithms to obtain low-dimensional dense vectors of graph data: DeepWalk and Struc2Vec (based on random walk), SDNE (based on self-encoder), and LINE (based on other methods).

DeepWalk is a graph embedding method based on Word2vec. It is an extension of the language model and unsupervised learning from word sequences to graphs. First, the neighbor nodes of the nodes in the network were randomly generated to form a fixed-length random walk sequence, and then, the generated fixed-length node sequence was mapped into a low-dimensional embedded vector using the skip-gram model. The generated vector encoded the relationship between nodes in the low-dimensional vector space, which was used to capture neighborhood similarity and community structure and extracted the potential characteristics of the nodes. This method can learn the relationship information of node pairs and realize incremental learning of dynamic graphs; its time complexity is $O(\log|V|)$. However, the performance of this method in a weighted graph was poor, as it could only maintain the second-order similarity of the graph, and the explicit objective function was not used in the optimization process, which limited the ability of the model to maintain the network structure, which would affect the integrity of the context information.

SDNE utilized the depth self-encoder and the first-order and second-order similarities of the graph to obtain the final embedded vectors through the highly non-linear function and the optimization objective function, which can effectively capture the highly non-linear network structure. SDNE includes supervised and unsupervised components that maintain the first-order and second-order similarities of the nodes. The supervised component introduced Laplacian feature mapping as the objective function of first-order similarity so that the generated embedding can capture local structure information. The unsupervised component modifies the L2 reconstruction loss function as the objective function of the second-order similarity so that the generated embedding can obtain the global structural features. The joint optimization of the first- and second-order similarities enhanced the robustness of the model on the sparse graph, and the generated embedding preserved the global and local structure information simultaneously. However, SDNE was inefficient in realizing the embedding of network nodes with higher orders of magnitude and could not realize the incremental update of graphs.

LINE also defined and optimized first-order and second-order similarity functions. First-order similarity was used to keep the point product of the adjacency matrix close to the embedded representation, and second-order similarity was adopted to maintain the similarity of the context nodes. LINE optimized the objective functions of the first-order and second-order similarities to minimize the distance between the node pair probability distribution generated by the adjacency matrix and the probability distribution generated by the embedded inner product through KL divergence, realized the optimization of graph embedding, and spliced the generated embedding vectors. The edge sampling strategy of LINE overcomes the limitations of random gradient descent and makes it applicable to large-scale graph embedding. However, the single optimization of the first-order and second-order representations and the simple splicing operation also limit the representation ability of LINE.

Unlike conventional graph embedding models, Struc2Vec focuses on the roles of different nodes in the network. The features of the nodes represent their locations and relationships with other nodes. However, many existing algorithms only express the nodes as vectors according to the distance relationships and do not consider the other structural features of the nodes. Most graph embedding models believe that the more common the neighbors of two nodes, the more similar the two nodes are, and it is natural to reduce their distance in the embedding space. However, this method cannot be used to identify node pairs with similar structures. In fact, some nodes have similar topological structures, but are too far away to have common neighbors. Struc2Vec ignores the attributes of nodes and edges and their positions in the network to evaluate the structural similarity between nodes; however, the limitations were still present.

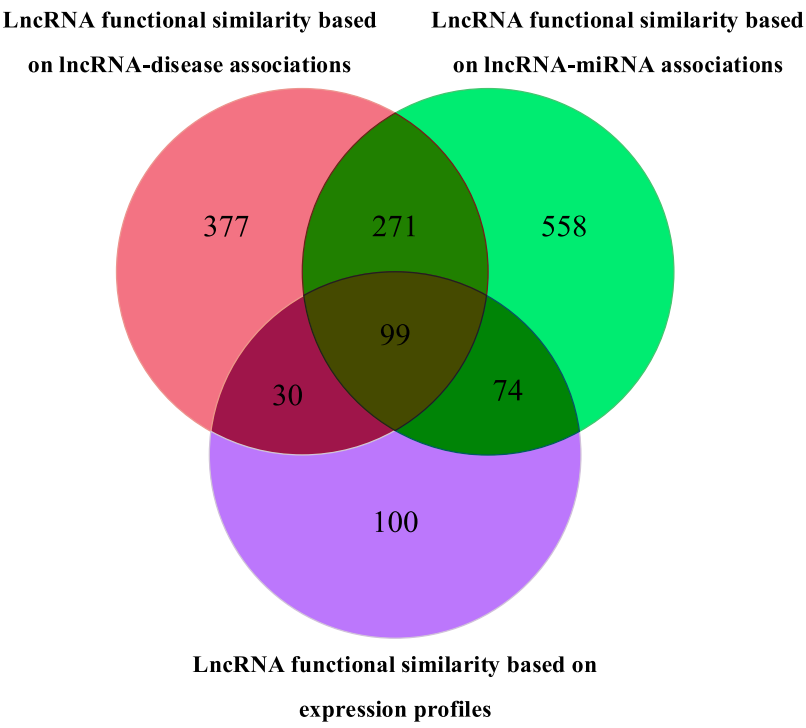

FIGURE 4
Venn diagram of three lncRNA sets in lncRNA feature vector selection.

2.7 Server construction

In this study, we developed a web server named TLSEA for lncRNA functional enrichment analysis, which was based on the fusion of heterogeneous information obtained from multiple sources. The flask framework was adopted for data processing and calculation. At the front end of the TLSEA, the framework of "HTML + CSS + Bootstrap" was employed, and Ploty.js and JQuery were used for graphical visualization and application logic, respectively. All computational algorithms were implemented in Python using the NumPy and Pandas packages. In total, lncRNA pathways of 385 diseases were identified (see Supplementary Table S6). TLSEA is unrestricted (without a login procedure), compatible with most web browsers, and accessible at http://www.lirmed.com:5003/tlsea.

3 Results

3.1 LncRNA feature vector selection

Three lncRNA networks were constructed using different similarity networks, as shown in Figure 4. The first lncRNA functional similarity network was constructed by integrating lncRNA–miRNA association data and miRNA similarity data, which included 1,002 lncRNAs (named LncRNAset1). The second lncRNA functional similarity network was constructed by integrating lncRNA–disease association data and disease semantic similarity network and included 777 lncRNAs (named LncRNAset2) and the third lncRNA functional similarity network by lncRNA expression profile data, which included 303 lncRNAs (named LncRNAset3).

To gain better low-dimensional feature vectors of lncRNAs, we extracted 16-, 32-, 64-, and 128-dimensional feature vectors with four types of prevailing graph embedding algorithms (DeepWalk, Struc2Vec, SDNE, and LINE). In this study, a random forest classifier was selected to evaluate the performance of the four algorithms. LncRNAset1, LncRNAset2, and LncRNAset3 were fed into the classifier to train it. Four types of random forest classifiers, namely, R1(X), R2(X), R3(X), and R4(X), were employed based on the 16-, 32-, 64-, and 128-dimensional lncRNA feature vectors extracted by the DeepWalk, Struc2Vec, SDNE, and LINE methods in turn. Finally, the effectiveness of each type of lncRNA feature vector was validated using 10-fold cross validation. Accuracy (ACC) values are listed in Table 1. The numbers in the vector names represent feature vector dimensions. For example, DeepWalk16 represents the DeepWalk algorithm with 16-dimensional feature vectors.

As shown in Table 1, the feature vectors extracted by SDNE64 yielded the best classification results for the two lncRNA functional similarity networks of lncRNA feature vectors, which are marked in bold. Therefore, we believe that the 64-dimensional vector extracted by SDNE could retain more information and be selected in this study.

3.2 Construction of an lncRNA–lncRNA association network

First, two sets of 64-dimensional lncRNA feature vectors based on the lncRNA–miRNA association network and lncRNA–disease

TABLE 1 ACC results of different lncRNA feature vectors.

Types of lncRNA feature vectors	ACC based on lncRNA–disease association	ACC based on lncRNA–miRNA association
DeepWalk16	0.8638	0.8066
LINE16	0.8869	0.8184
SDNE16	0.9815	0.9078
Struc2Vec16	0.8588	0.8246
DeepWalk32	0.8467	0.8101
LINE32	0.9351	0.8220
SDNE32	0.9812	0.9061
Struc2Vec32	0.8693	0.8186
DeepWalk64	0.8854	0.8000
LINE64	0.9509	0.8063
SDNE64	0.9824	0.9131
Struc2Vec64	0.8840	0.7996
DeepWalk128	0.8907	0.7324
LINE128	0.9559	0.7888
SDNE128	0.9822	0.9045
Struc2Vec128	0.8931	0.7380

association network were merged into a new set. If one lncRNA appeared in only one of the two association networks, its vector was retained as the merged vector. If it appeared in both functional annotation networks, the average value of the two corresponding vectors was considered as the merged vector value. After merging, a novel lncRNA feature matrix was obtained in which each lncRNA node of the two association networks corresponded to a row vector. The Pearson correlation coefficient ρ was used to evaluate the closeness of the relationship between two lncRNA feature vectors, which was calculated using the following formula:

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2\right]^{\frac{1}{2}}}, \quad (13)$$

where $\bar{x}$ is the average values for all x and $\bar{y}$ is the average values for all y.

Integrating the lncRNA functional similarity based on lncRNA–disease association and lncRNA functional similarity based on lncRNA–miRNA association network, an lncRNA–lncRNA association network (see Supplementary Table S7) was constructed by merging lncRNA-related heterogeneous information obtained from multiple sources and various lncRNA-related similarity networks, which included 1,409 lncRNAs.

3.3 Overview of the TLSEA web server

In the TLSEA, users only need to submit a list of lncRNAs of interest. Users can utilize the TLSEA to calculate the p-values of the original lncRNA list and the lncRNA list after random walk expansion. As shown in Figures 5, 6, the web interface of the TLSEA was designed according to the following workflow. First, the user inputs an lncRNA list of interest and then selects the similarity coefficient for expansion. The larger the similarity coefficient, the more similar the expanded lncRNA list is to the original lncRNA list. The unified lncRNA naming format used in the TLSEA enrichment analysis is the HGNC symbol. If the lncRNA names did not match the HGNC symbols, users needed to convert them to this format using the LncBook 2.0 database (Li et al., 2022) or other conversion tools before analyzing the data with the TLSEA. If the user chooses the similarity coefficient as "None," it implies that only the original lncRNA list was used for enrichment analysis. Finally, the user clicked the "Run" button to complete the task. If the similarity coefficient was not "None," the TLSEA would additionally display the expanded lncRNA list and provide a button to export it. The enrichment analysis results are shown in Figure 6. TLSEA could also visualize the results with one bubble chart by clicking the "Results Visualization" button.

3.4 Case studies

To further evaluate the application of the TLSEA model in practical situations, we used the TLSEA to analyze the functions of differentially expressed genes in breast cancer. For the case study, we first downloaded a list of differentially expressed lncRNAs ($log_2 FC > 1$; $P_{adj} < 0.05$) of breast cancer from the TCGA project and CircRNAnet database as the input list of the TLSEA. Then, the similarity coefficient was set to "None," which meant that the enrichment analysis was performed based on the original differential expression lncRNA list, and the "Run" button was clicked to implement the enrichment analysis. The TLSEA would

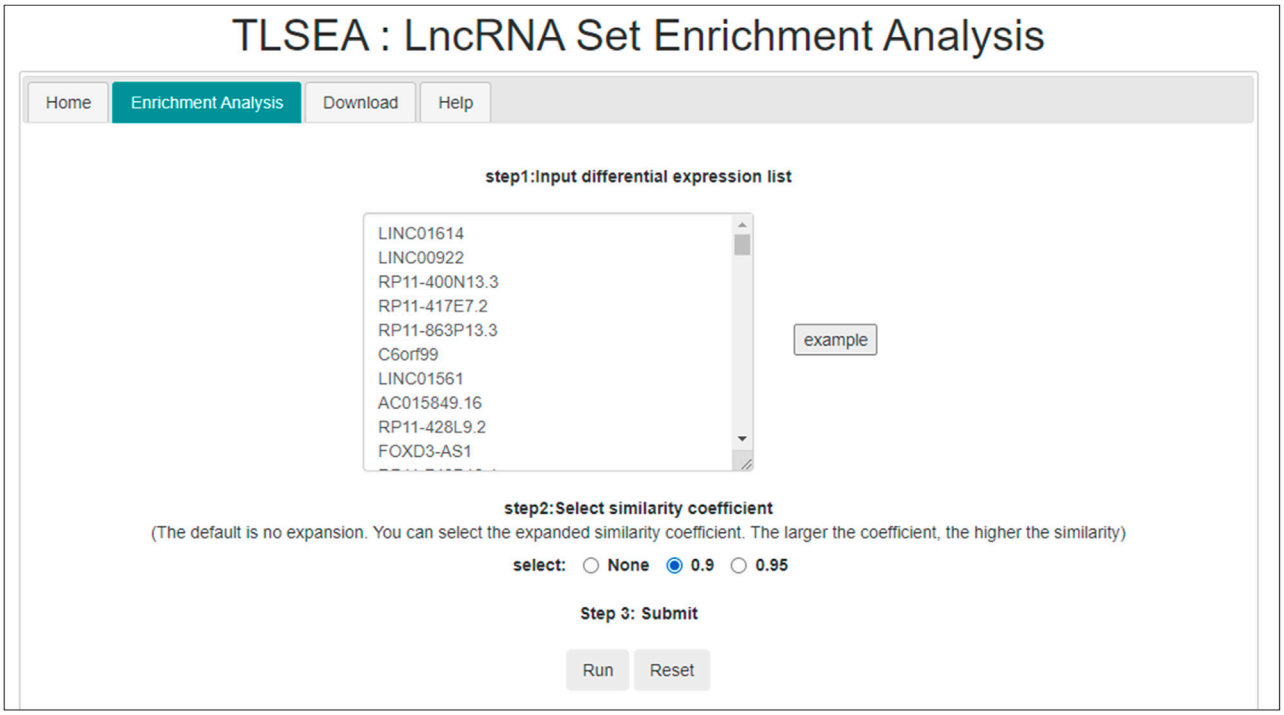

FIGURE 5
Enrichment analysis page of TLSEA.

output all disease pathways with a p-value <0.01 and provide the results of visual enrichment analysis. The top 15 significances of the enrichment analysis results are displayed on the results page. Subsequently, the similarity coefficient was set to 0.9, which meant that the TLSEA would screen the lncRNA–lncRNA association network based on multi-source heterogeneous information fusion in advance, retaining only the edges whose similarity values exceeded 0.9. Subsequently, the differentially expressed lncRNAs were used as seed nodes, and the random walk with restart method was performed on the lncRNA–lncRNA association network. After all nodes converged, the nodes whose random walk probabilities were not 0 were identified as expanded lncRNAs and used to obtain the expanded lncRNA list. Finally, the TLSEA performed an enrichment analysis of these expanded lncRNAs.

The breast cancer disease lncRNA set included 185 lncRNAs and only 30 lncRNAs from the original differentially expressed lncRNA list, with a hit rate of 16.22%. The case study results showed that 73 lncRNAs from the expanded lncRNA list were hit after performing the random walk with the restart method, and the hit rate increased to 39.46%, as shown in Figure 7. The p-value of breast cancer with the original differentially expressed lncRNA list was $1.34e^{-13}$, and the p-value of breast cancer with the expanded list after performing random walk with the restart method was $1.37e^{-20}$. The expanded list was significantly enriched in breast cancer compared with the original list. In addition, the p-values of the top 10 diseases in the enrichment analysis results of the original list were significantly improved, as shown in Table 2. Experimental findings proved that TLSEA could effectively improve the accuracy of enrichment analysis of the lncRNA list.

After statistical analysis, 43 additional lncRNAs were calculated based on the expanded list, which were not found based on the original lncRNA list. All of them were validated in the literature; their names and corresponding PMIDs are listed in Table 3.

The enrichment results calculated by the TLSEA not only significantly improved the p-values of the diseases in the enrichment results but also mined new diseases that were not enriched by the original lncRNA list. In this study, the expanded lncRNA list of breast cancer was significantly enriched in head and neck squamous cell carcinoma, with a p-value of $1.38e^{-6}$. According to the lncSEA database (Chen et al., 2021), there were a total of 12 lncRNAs related to head and neck squamous cell carcinoma, and nine of them were hit in our case study. In contrast, only two lncRNAs were hit based on the original lncRNA list, as shown in Table 4. The experimental results of this case have previously shown that there is a certain relationship between breast cancer and head and neck squamous cell carcinoma (Croce, 2022), which also indicates that the TLSEA can detect some new diseases ignored by conventional enrichment analysis methods.

4 Discussion

Gene set enrichment analysis is a pervasive bioinformatic technique used to detect the underlying biological pathways and

10.3389/fgene.2023.1181391

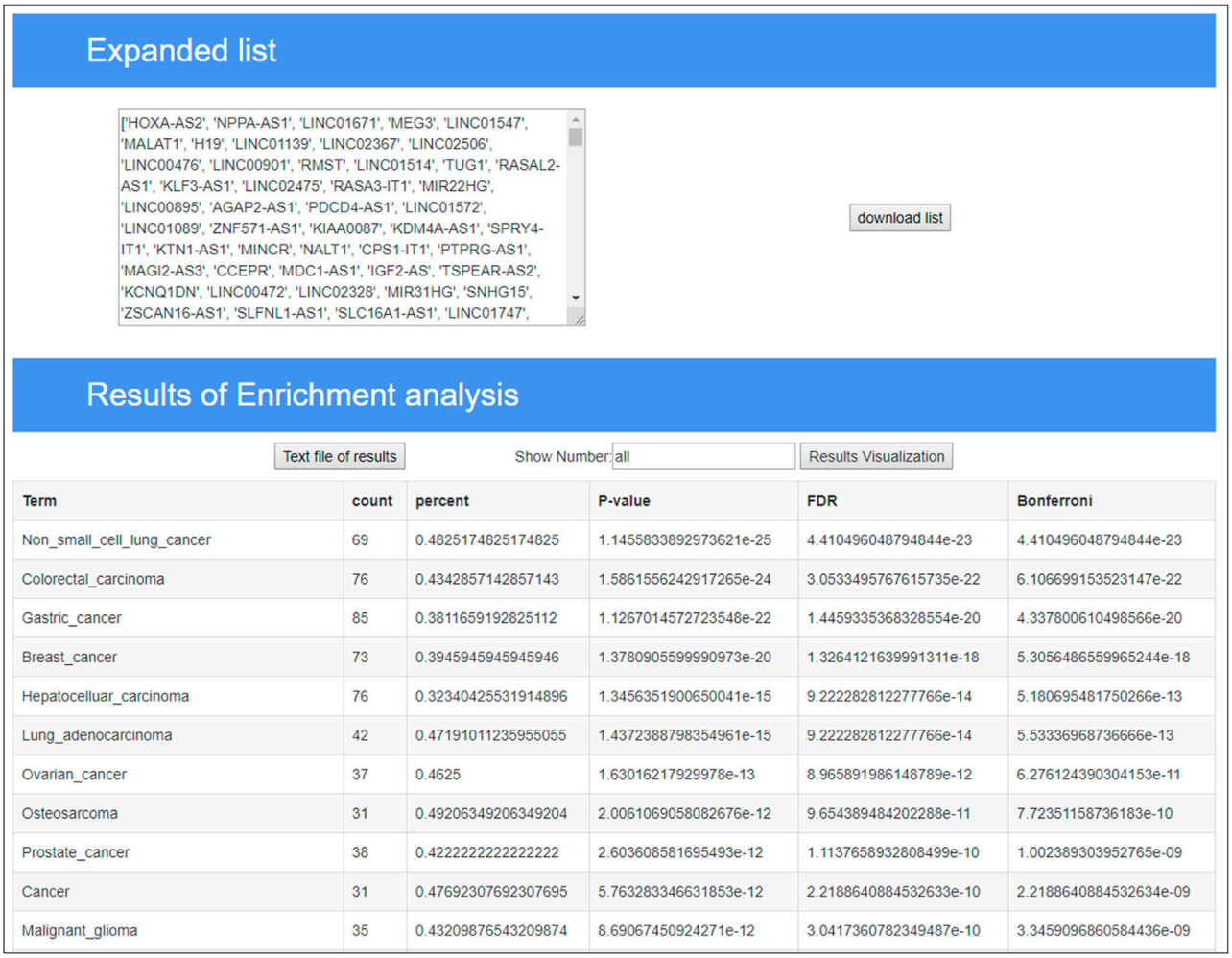

FIGURE 6
Enrichment analysis result page of TLSEA.

TABLE 2 Comparison of *p*-values of the origin list and expanded list of the top 10 diseases in TLSEA.

Disease	*p*-values of the origin list	*p*-values of the expanded list
Non-small-cell lung cancer	$5.48e^{-14}$	$1.15e^{-25}$
Breast cancer	$1.34e^{-13}$	$1.38e^{-20}$
Gastric cancer	$6.27e^{-13}$	$1.13e^{-22}$
Colorectal carcinoma	$5.26e^{-11}$	$1.59e^{-24}$
Nasopharyngeal carcinoma	$2.00e^{-8}$	$4.85e^{-10}$
Prostate cancer	$2.49e^{-8}$	$2.60e^{-12}$
Esophageal squamous cancer	$3.64e^{-8}$	$2.33e^{-11}$
Thyroid cancer	$6.52e^{-8}$	$3.00e^{-8}$
Esophageal cancer	$1.49e^{-7}$	$3.15e^{-8}$
Squamous cell carcinoma	$1.87e^{-7}$	$1.19e^{-7}$

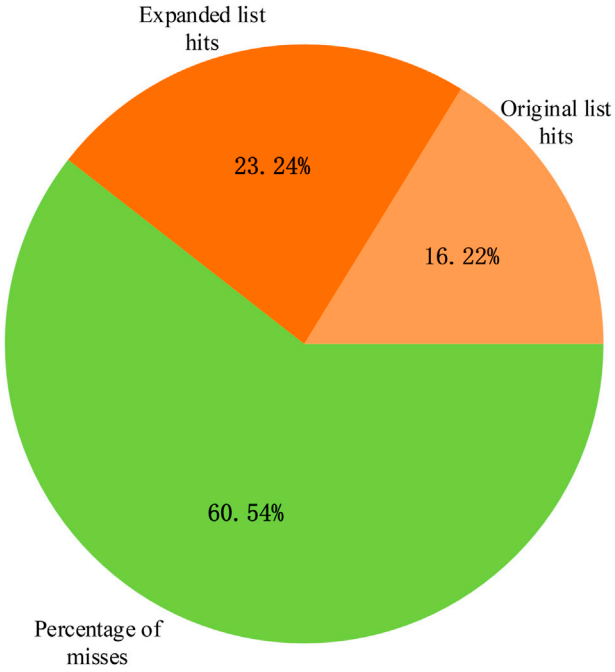

FIGURE 7
Percentage of lncRNA hit terms after performing random walk with the restart method.

functional categories of a given gene list. In this study, we developed an lncRNA set enrichment analysis tool, the TLSEA, based on a multi-source heterogeneous information fusion. Conventional algorithms, such as hypergeometric and binomial tests, do not explicitly consider the rich association information among input lncRNAs in the lncRNA list, which is a hindrance in obtaining more insightful biological processes. We introduced the interaction information between miRNAs and lncRNAs and that between diseases and lncRNAs to construct a novel lncRNA–lncRNA association network and expanded the lncRNA list using random walk with restart. Using a case study, TLSEA demonstrated that the expanded lncRNA list can detect more insightful pathways than the original lncRNA list. Additionally, TLSEA provides a simple and user-friendly interface for analyzing, browsing, and downloading detailed information from lncRNA set enrichment analysis, which can help researchers understand the mechanisms of disease and develop effective diagnosis and treatment.

Generally, the expanded lncRNA list is more significantly enriched in the corresponding disease pathway than the original lncRNA. However, sometimes, selecting a small threshold for the similarity coefficient may lead to the introduction of some unrelated lncRNAs, which would cause poor enrichment analysis results. Therefore, how to provide a more suitable expansion strategy for users will be the future subject of a follow-up article on TLSEA. In addition, the insufficient disease pathway data will limit the comprehensiveness of enrichment analysis and calculation results. As the size of disease pathway data grows, TLSEA will supplement and

TABLE 3 List of 43 additional lncRNAs and corresponding PMIDs of breast cancer based on the expanded lncRNA list.

lncRNA	PMID	lncRNA	PMID
HOXA-AS2	28545023	LINC02099	27597120
LINC00472	33668040	CERNA2	32248842
NORAD	34190442	SNHG15	32141559
SNHG7	33099915	MIR31HG	34076993
RMST	29215701	HIF1A-AS2	30635931
CYTOR	33842324	LINC00636	26929647
RASSF1-AS1	31062660	LINP1	27111890
FGF14-AS2	31486497	EGOT	26159853
NNT-AS1	32691576	PTPRG-AS1	34326372
LINC00598	28339037	LINC00901	25435812
NBAT1	26378045	CASC2	29523222
DIRC3	25122612	MALAT1	30349115
LINC01089	31417284	MAGI2-AS3	32730644
STXBP5-AS1	34764730	LINC-ROR	29041978
SNHG16	32122142	MEG3	33845141
TUG1	33380806	SPRY4-IT1	31736268
IRAIN	25465188	H19	33324070
FOXC2-AS1	29562954	JADRR	24097061
LINC02130	28003470	CASC22	24879036
LINC00339	31781497	PDCD4-AS1	33248413
KLF3-AS1	29453409	LINC01671	28003470
LINC00993	31921620		

TABLE 4 Enrichment analysis results of the original lncRNA list and the expanded lncRNA list on head and neck squamous cell carcinoma.

lncRNA	PMID	Original list	Expanded list
C5orf66-AS1	30280186	Hitting	Hitting
CYTOR	35963855	Not hitting	Hitting
SPRY4-IT1	29575229	Not hitting	Hitting
FAM3D-AS1	Unconfirmed	Not hitting	Hitting
HAND2-AS1	29575229	Not hitting	Hitting
H19	27994496	Not hitting	Hitting
LUCAT1	29575229	Not hitting	Hitting
HEIH	29575229	Not hitting	Hitting
HOTAIR	31297902	Hitting	Hitting
EPB41L4A-AS2	29490660	Not hitting	Not hitting
LNC-JPH1-7	27323410	Not hitting	Not hitting
LNC-LCE5A-1	25904139	Not hitting	Not hitting

include more multi-source heterogeneous information on lncRNAs. The future version of TLSEA will include more categories of lncRNAs and integrate additional functional information into the knowledge base, annotate more species, and develop a more efficient expanding method for lncRNA lists of interest.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material; further inquiries can be directed to the corresponding author.

Author contributions

JL conceived and designed the study; ZL and HL developed the algorithm and performed the statistical analysis; ZL and BW wrote the codes; ZL and YW drafted the original manuscript; and JL and ZL revised the manuscript. All authors have read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China under Grant Nos. 62072154 and 62202330.

Acknowledgments

The authors thank the members in their groups for their valuable discussions.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2023.1181391/full#supplementary-material

References

Bao, Z., Yang, Z., Huang, Z., Zhou, Y., Cui, Q., and Dong, D. (2019). LncRNADisease 2.0: An updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res.* 47 (D1), D1034–D1037. doi:10.1093/nar/gky905

Baumann, N. (2016). How to use the medical subject headings (MeSH). *Int. J. Clin. Pract.* 70 (2), 171–174. doi:10.1111/ijcp.12767

Beissbarth, T., and Speed, T. P. (2004). GOstat: Find statistically overrepresented gene ontologies within a group of genes. *Bioinformatics* 20 (9), 1464–1465. doi:10.1093/bioinformatics/bth088

Chen, J., Shishkin, A. A., Zhu, X., Kadri, S., Maza, I., Guttman, M., et al. (2016). Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biol.* 17, 19. doi:10.1186/s13059-016-0880-9

Chen, J., Zhang, J., Gao, Y., Li, Y., Feng, C., Song, C., et al. (2021). LncSEA: A platform for long non-coding RNA related sets and enrichment analysis. *Nucleic Acids Res.* 49 (D1), D969–d980. doi:10.1093/nar/gkaa806

Chen, X., Yan, C. C., Luo, C., Ji, W., Zhang, Y., and Dai, Q. (2015). Constructing lncRNA functional similarity network based on lncRNA-disease associations and disease semantic similarity. *Sci. Rep.* 5, 11338. doi:10.1038/srep11338

Croce, M. V. (2022). An introduction to the relationship between lewis x and malignancy mainly related to breast cancer and head neck squamous cell carcinoma (HNSCC). *Cancer Invest.* 40 (2), 173–183. doi:10.1080/07357907.2021.2016800

Del Giacco, L., and Cattaneo, C. (2012). Introduction to genomics. *Methods Mol. Biol.* 823, 79–88. doi:10.1007/978-1-60327-216-2_6

DiStefano, J. K. (2018). The emerging role of long noncoding RNAs in human disease. *Methods Mol. Biol.* 1706, 91–110. doi:10.1007/978-1-4939-7471-9_6

Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C., and Conklin, B. R. (2003). MAPPFinder: Using gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.* 4 (1), R7. doi:10.1186/gb-2003-4-1-r7

Draghici, S., Khatri, P., Tarca, A. L., Amin, K., Done, A., Voichita, C., et al. (2007). A systems biology approach for pathway level analysis. *Genome Res.* 17 (10), 1537–1545. doi:10.1101/gr.6202607

Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., et al. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature* 539 (7629), 452–455. doi:10.1038/nature20149

Fan, W., Shang, J., Li, F., Sun, Y., Yuan, S., Liu, J. X., et al. (2020). IDSSIM: an lncRNA functional similarity calculation model based on an improved disease semantic similarity method. *BMC Bioinformatics* 21 (1), 339. doi:10.1186/s12859-020-03699-9

Fang, S., Zhang, L., Guo, J., Niu, Y., Wu, Y., Li, H., et al. (2018). NONCODEV5: A comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.* 46 (D1), D308–D314. doi:10.1093/nar/gkx1107

Fang, Y., and Fullwood, M. J. (2016). Roles, functions, and mechanisms of long non-coding RNAs in cancer. *Genomics Proteomics Bioinforma.* 14 (1), 42–54. doi:10.1016/j.gpb.2015.09.006

Fillinger, S., de la Garza, L., Peltzer, A., Kohlbacher, O., and Nahnsen, S. (2019). Challenges of big data integration in the life sciences. *Anal. Bioanal. Chem.* 411 (26), 6791–6800. doi:10.1007/s00216-019-02074-9

Gao, Y., Wang, P., Wang, Y., Ma, X., Zhi, H., Zhou, D., et al. (2019). Lnc2Cancer v2.0: Updated database of experimentally supported long non-coding RNAs in human cancers. *Nucleic Acids Res.* 47 (D1), D1028–D1033. doi:10.1093/nar/gky1096

Glaab, E., Baudot, A., Krasnogor, N., and Valencia, A. (2010). TopoGSA: Network topological gene set analysis. *Bioinformatics* 26 (9), 1271–1272. doi:10.1093/bioinformatics/btq131

Gloss, B. S., and Dinger, M. E. (2016). The specificity of long noncoding RNA expression. *Biochim. Biophys. Acta* 1859 (1), 16–22. doi:10.1016/j.bbagrm.2015.08.005

Gutschner, T., Hämmerle, M., Eissmann, M., Hsu, J., Kim, Y., Hung, G., et al. (2013). The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res.* 73 (3), 1180–1189. doi:10.1158/0008-5472.Can-12-2850

Guttman, M., Amit, I., Garber, M., French, C., Lin, M. F., Feldser, D., et al. (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458 (7235), 223–227. doi:10.1038/nature07672

Hua, Q., Jin, M., Mi, B., Xu, F., Li, T., Zhao, L., et al. (2019). LINC01123, a c-Myc-activated long non-coding RNA, promotes proliferation and aerobic glycolysis of non-small cell lung cancer through miR-199a-5p/c-Myc axis. *J. Hematol. Oncol.* 12 (1), 91. doi:10.1186/s13045-019-0773-y

Jiao, X., Sherman, B. T., Huang da, W., Stephens, R., Baseler, M. W., Lane, H. C., et al. (2012). DAVID-WS: A stateful web service to facilitate gene/protein list analysis. *Bioinformatics* 28 (13), 1805–1806. doi:10.1093/bioinformatics/bts251

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28 (1), 27–30. doi:10.1093/nar/28.1.27

Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2014). Data, information, knowledge and principle: Back to metabolism in KEGG. *Nucleic Acids Res.* 42, D199–D205. doi:10.1093/nar/gkt1076

Kashi, K., Henderson, L., Bonetti, A., and Carninci, P. (2016). Discovery and functional analysis of lncRNAs: Methodologies to investigate an uncharacterized transcriptome. *Biochim. Biophys. Acta* 1859 (1), 3–15. doi:10.1016/j.bbagrm.2015.10.010

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: Current approaches and outstanding challenges. *PLoS Comput. Biol.* 8 (2), e1002375. doi:10.1371/journal.pcbi.1002375

Kopp, F., and Mendell, J. T. (2018). Functional classification and experimental dissection of long noncoding RNAs. *Cell* 172 (3), 393–407. doi:10.1016/j.cell.2018.01.011

Lalevée, S., and Feil, R. (2015). Long noncoding RNAs in human disease: Emerging mechanisms and therapeutic strategies. *Epigenomics* 7 (6), 877–879. doi:10.2217/epi.15.55

Li, J. H., Liu, S., Zhou, H., Qu, L. H., and Yang, J. H. (2014). starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* 42, D92–D97. doi:10.1093/nar/gkt1248

Li, J., Zhang, S., Wan, Y., Zhao, Y., Shi, J., Zhou, Y., et al. (2019). MISIM v2.0: A web server for inferring microRNA functional similarity based on microRNA-disease associations. *Nucleic Acids Res.* 47 (W1), W536–W541. doi:10.1093/nar/gkz328

Li, Y., Zhuang, L., Wang, Y., Hu, Y., Wu, Y., Wang, D., et al. (2013). Connect the dots: A systems level approach for analyzing the miRNA-mediated cell death network. *Autophagy* 9 (3), 436–439. doi:10.4161/auto.23096

Li, Z., Liu, L., Feng, C., Qin, Y., Xiao, J., Zhang, Z., et al. (2022). LncBook 2.0: Integrating human long non-coding RNAs with multi-omics annotations. *Nucleic Acids Res.* 51, D186–D191. doi:10.1093/nar/gkac999

Liu, Y., Zhao, J., Zhang, W., Gan, J., Hu, C., Huang, G., et al. (2015). lncRNA GAS5 enhances G1 cell cycle arrest via binding to YBX1 to regulate p21 expression in stomach cancer. *Sci. Rep.* 5, 10159. doi:10.1038/srep10159

McDonel, P., and Guttman, M. (2019). Approaches for understanding the mechanisms of long noncoding RNA regulation of gene expression. *Cold Spring Harb. Perspect. Biol.* 11 (12), a032151. doi:10.1101/cshperspect.a032151

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., et al. (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* 34 (3), 267–273. doi:10.1038/ng1180

Munos, B. (2009). Lessons from 60 years of pharmaceutical innovation. *Nat. Rev. Drug Discov.* 8 (12), 959–968. doi:10.1038/nrd2961

Niknafs, Y. S., Han, S., Ma, T., Speers, C., Zhang, C., Wilder-Romans, K., et al. (2016). The lncRNA landscape of breast cancer reveals a role for DSCAM-AS1 in breast cancer progression. *Nat. Commun.* 7, 12791. doi:10.1038/ncomms12791

Scarselli, F., Gori, M., Tsoi, A. C., Hagenbuchner, M., and Monfardini, G. (2009). The graph neural network model. *IEEE Trans. Neural Netw.* 20 (1), 61–80. doi:10.1109/tnn.2008.2005605

Seal, R. L., Tweedie, S., and Bruford, E. A. (2022). A standardised nomenclature for long non-coding RNAs. *IUBMB Life*. doi:10.1002/iub.2663

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* 102 (43), 15545–15550. doi:10.1073/pnas.0506580102

Tarca, A. L., Draghici, S., Khatri, P., Hassan, S. S., Mittal, P., Kim, J. S., et al. (2009). A novel signaling pathway impact analysis. *Bioinformatics* 25 (1), 75–82. doi:10.1093/bioinformatics/btn577

Tay, Y., Rinn, J., and Pandolfi, P. P. (2014). The multilayered complexity of ceRNA crosstalk and competition. *Nature* 505 (7483), 344–352. doi:10.1038/nature12986

Teng, X., Chen, X., Xue, H., Tang, Y., Zhang, P., Kang, Q., et al. (2020). NPInter v4.0: An integrated database of ncRNA interactions. *Nucleic Acids Res.* 48 (D1), D160–D165. doi:10.1093/nar/gkz969

Topel, H., Bagirsakci, E., Comez, D., Bagci, G., Cakan-Akdogan, G., and Atabey, N. (2020). lncRNA HOTAIR overexpression induced downregulation of c-Met signaling promotes hybrid epithelial/mesenchymal phenotype in hepatocellular carcinoma cells. *Cell Commun. Signal* 18 (1), 110. doi:10.1186/s12964-020-00602-0

Tsang, W. P., and Kwok, T. T. (2007). Riboregulator H19 induction of MDR1-associated drug resistance in human hepatocellular carcinoma cells. *Oncogene* 26 (33), 4877–4881. doi:10.1038/sj.onc.1210266

Uchida, S., and Dimmeler, S. (2015). Long noncoding RNAs in cardiovascular diseases. *Circ. Res.* 116 (4), 737–750. doi:10.1161/circresaha.116.302521

Wang, D., Wang, J., Lu, M., Song, F., and Cui, Q. (2010). Inferring the human microRNA functional similarity and functional network based on microRNA-associated diseases. *Bioinformatics* 26 (13), 1644–1650. doi:10.1093/bioinformatics/btq241

Wang, J., Huang, Q., Liu, Z. P., Wang, Y., Wu, L. Y., Chen, L., et al. (2011). NOA: A novel network Ontology analysis method. *Nucleic Acids Res.* 39 (13), e87. doi:10.1093/nar/gkr251

Wang, W., and Krishnan, E. (2014). Big data and clinicians: A review on the state of the science. *JMIR Med. Inf.* 2 (1), e1. doi:10.2196/medinform.2913

Xue, M., Zhuo, Y., and Shan, B. (2017). MicroRNAs, long noncoding RNAs, and their functions in human disease. *Methods Mol. Biol.* 1617, 1–25. doi:10.1007/978-1-4939-7046-9_1

Zhu, S., Hu, X., Han, S., Yu, Z., Peng, Y., Zhu, J., et al. (2014). Differential expression profile of long non-coding RNAs during differentiation of cardiomyocytes. *Int. J. Med. Sci.* 11 (5), 500–507. doi:10.7150/ijms.7849