ABSTRACT
Summary: Disease ontology (DO) annotates human genes in the context of disease. DO is important annotation in translating molecular findings from high-throughput data to clinical relevance. DOSE is an R package providing semantic similarity computations among DO terms and genes which allows biologists to explore the similarities of diseases and of gene functions in disease perspective. Enrichment analyses including hypergeometric model and gene set enrichment analysis are also implemented to support discovering disease associations of high-throughput biological data. This allows biologists to verify disease relevance in a biological experiment and identify unexpected disease associations. Comparison among gene clusters is also supported.
Availability and implementation: DOSE is released under Artistic-2.0 License. The source code and documents are freely available through Bioconductor (http://www.bioconductor.org/packages/release/bioc/html/DOSE.html).
Supplementary information: Supplementary data are available at Bioinformatics online.
Contact: gcyu@connect.hku.hk or tqyhe@jnu.edu.cn
Received on May 7, 2014; revised on October 9, 2014; accepted on October 14, 2014

1 INTRODUCTION
Characterizing disease-disease relationships and mining gene-disease associations provides insights in analyzing high-throughput data to elucidate molecular mechanisms of complex diseases. Understanding similarities among diseases and among genes in disease context helps in early diagnosis, drug repurposing, and new drug development. Investigating gene-disease associations with gene lists obtained by high-throughput experiments helps exploring biological questions in disease context and discovering unanticipated functions.

Disease ontology (DO) provides a consistent description of genes in disease perspectives. To provide researchers with more accessible of disease knowledge, the DO database (Schröd et al., 2012) supplies a web browser for users to explore DO vocabularies while disease and gene annotations database (Peng et al., 2013) supplies a web interface for mapping genes and diseases. DO is organized as a directed acyclic graph, laying the foundation for computation of disease knowledge using semantic similarity algorithms. There are many generic quality tools for computation of semantic measures including SML, SimPack, SemMF, OWLSim and Similarity Library (http://goo.gl/3xCuJ6). These generic libraries can be employed to analyze DO semantic similarities. DOSim (Li et al., 2011) was designed specific for DO, but the authors fail to maintain the package. Functional DO (FunDO) (Osborne et al., 2009) implemented hypergeometric test to assess significant of DO associations with a gene list. However, FunDO doesn’t allow users to customize the background set of genes and thus may introduce biases in the results.

To address the shortcoming of lack of R/Bioconductor package that designed for computation of semantic and enrichment analyses based on DO, we present DOSE, that allows measuring semantic similarity among DO terms and genes using several information-content and graph-structure based algorithms. For evaluating functional associations with gene lists of high-throughput genomic and proteomic studies, DOSE supports hypergeometric test and gene set enrichment analysis (GSEA), which incorporate expression level measurements to extract disease relevance of biological experiments. More importantly, DOSE provides several DO-specific visualization functions to produce highly customizable, publication-quality figures of similarity and enrichment analyses that are not available elsewhere. With these visualization tools, the results obtained by DOSE are more interpretable.

2 IMPLEMENTATION
DOSE provides doSim function to compute semantic similarity among DO terms. DOSE implemented four information content based algorithms proposed by Resnik (Resnik, 1999), Lin (Lin, 1998), Jiang and Conrath (Jiang and Conrath, 1997) and Schlicker (Schlicker et al., 2006), respectively, and one graph based algorithm proposed by Wang (Wang et al., 2007) to measure the semantic similarity among DO terms.

These algorithms were extended from in-house developed R package GOsemSim (Yu et al., 2010). By mapping genes to DO annotations, geneSim function measures the semantic similarities among genes based on their annotated DO terms. Four combine strategies were implemented in DOSE for aggregating semantic similarity scores of multiple DO terms associated with genes,
3 RESULTS AND CONCLUSION

DOSE was developed using the R statistical computing language and is released within Bioconductor project. It provides five algorithms for DO and gene semantic similarity measurements (Fig. 1A); hypergeometric test for identifying significant disease association of gene list (Fig. 1B and C); GSEA for interpreting genome wide expression profiles in disease context (Fig. 1D) and comparison of significant disease associations among different gene sets (Fig. 1E). R scripts to generate Figure 1 are presented in Supplemental File.

The DOSE package presented here makes use of semantic similarity approaches and enrichment analyses to facilitate users the abilities to visualize semantic similarities, significant gene-disease associations, and gene set comparison.

Funding: This work was supported by the National Natural Science Foundation of China (21271086 to Q.-Y.H.) and Fundamental Research Funds for the Central Universities (21613414 to G.Y.).

REFERENCES


Li, J. et al. (2011) DOSim: an R package for similarity between diseases based on disease ontology. BMC Bioinformatics, 12, 266.


