

Research Article

Annotating the Function of the Human Genome with Gene Ontology and Disease Ontology

Yang Hu,¹ Wenyang Zhou,¹ Jun Ren,¹ Lixiang Dong,² Yadong Wang,³
Shuilin Jin,⁴ and Liang Cheng⁵

¹*School of Life Science and Technology, Harbin Institute of Technology, Harbin 150001, China*

²*School of Software, Harbin Institute of Technology, Harbin 150001, China*

³*School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China*

⁴*Department of Mathematics, Harbin Institute of Technology, Harbin 150001, China*

⁵*College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China*

Correspondence should be addressed to Yadong Wang; ydwang@hit.edu.cn, Shuilin Jin; jinsl@hit.edu.cn, and Liang Cheng; liangcheng@hrbmu.edu.cn

Received 2 June 2016; Revised 24 July 2016; Accepted 27 July 2016

Academic Editor: Xing Chen

Copyright © 2016 Yang Hu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Increasing evidences indicated that function annotation of human genome in molecular level and phenotype level is very important for systematic analysis of genes. In this study, we presented a framework named Gene2Function to annotate Gene Reference into Functions (GeneRIFs), in which each functional description of GeneRIFs could be annotated by a text mining tool Open Biomedical Annotator (OBA), and each Entrez gene could be mapped to Human Genome Organisation Gene Nomenclature Committee (HGNC) gene symbol. After annotating all the records about human genes of GeneRIFs, 288,869 associations between 13,148 mRNAs and 7,182 terms, 9,496 associations between 948 microRNAs and 533 terms, and 901 associations between 139 long noncoding RNAs (lncRNAs) and 297 terms were obtained as a comprehensive annotation resource of human genome. High consistency of term frequency of individual gene (Pearson correlation = 0.6401, $p = 2.2e - 16$) and gene frequency of individual term (Pearson correlation = 0.1298, $p = 3.686e - 14$) in GeneRIFs and GOA shows our annotation resource is very reliable.

1. Introduction

The human genome is the complete set of nucleic acid sequence for human beings [1]. Researches on sequence of the human genome aim at exploring the functions of genes [2–5]. Human genes consisting of sequences could play diverse roles based on their functions in molecular level in balancing the body. Once the balance is lost by lack or enhancement of the functions of genes, diseases could be induced [6–9].

Previous studies focused on identifying the functions of the protein-coding genes in molecular level based on their encoded proteins. For example, through investigating p53 protein, Brain and Jenkins [10] exposed that TP53 gene is potentially capable of inhibiting mammalian replicative DNA synthesis by blocking the DNA strand separation step during replication origin recruitment. Based on a case

control study, Benzon Larsen et al. [11] determined that ADH polymorphisms, which modify the rate of ethanol oxidation to acetaldehyde, were associated with breast cancer risk.

As a growing number of protein-coding genes identified, lots of functional terms emerged. For ease of comparing the functions of genes, these terms needed to be normalized. To this end, ontology was introduced to standardize the functional terms of genes. Among existing ontologies, Gene Ontology (GO) [12] is one of the earliest and most frequently used vocabularies, which focuses on describing biological process (BP), molecular function (MF), and cell component (CC) of genes. Since appearing in 2000, a large number of databases recording the functions of genes were annotated to the GO. The functional annotation of human protein-coding genes was provided at GO Annotation (GOA) databases [13], which involves a nonredundant set of annotations to

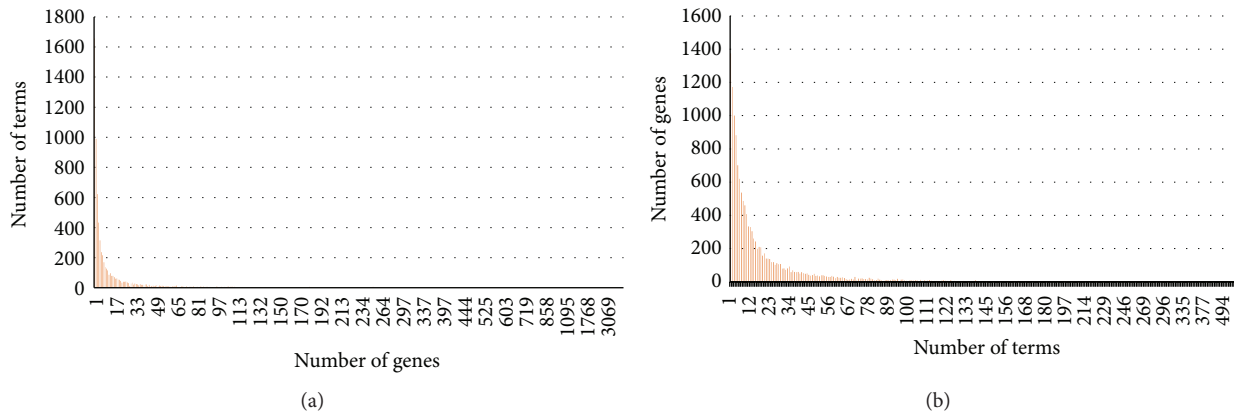


FIGURE 1: Distribution of functional terms and genes in the annotation results. (a) Histogram of the number of genes associated with individual functional term. (b) Histogram of the number of functional terms associated with individual gene.

the human proteome. In comparison with the GO, Disease Ontology (DO) [14] focuses on standardizing the functional terms of genes at phenotype level. And disease terms in Gene Reference into Function (GeneRIF) [15] were annotated to the DO [16–18].

Recently, large-scale sequence analysis at genomic and transcriptomic level has shown that more than 98% of genome sequence cannot encode protein [19, 20], and microRNA genes and long noncoding RNA (lncRNA) genes constitute a large portion of them [21]. In comparison with protein-coding genes, the functions of microRNA genes and lncRNA genes are difficult to be identified [22]. However, these noncoding genes play an important role at molecular level and phenotype level [23–27]. For example, at molecular level, qPCR and in silico hybridization revealed that miR-124 and miR-155 can be directly involved in the transcriptional regulation of Runt-related transcription factor 2 (RUNX2) and receptor activator of nuclear factor kappa-B ligand (RANKL) genes [28]. At phenotype level, Huang et al. identified that underexpression of miR-345 is associated with prostate cancer [29]. At present, microRNA- and lncRNA-related diseases in HMDD [30] and lncRNADisease [31] have been manually annotated by Medical Subject Headings (MeSH) [32]. And several recent works proved more relationship between miRNA and diseases would be detected yet [33–35].

Although a few of databases have been annotated to gene functional vocabularies, a comprehensive annotation resource recording the functions of human genes had not yet appeared. For example, in our knowledge, no databases of noncoding genes were annotated to functional vocabularies at molecular level. This may be caused by the lack of resources that record the functions of protein-coding genes and non-coding genes simultaneously. Fortunately, GeneRIFs [15] provides a brief (up to 255 character) functional description of each gene in the NCBI database, and these functional descriptions could be annotated to vocabularies, such as DO and GO.

In this paper, we presented a framework, Gene2Function, to annotate the function of human genome with GO and DO.

TABLE 1: The statistical information of associations between genes and terms.

The number of genes	The number of terms	The number of associations between genes and terms
mRNA		
13,148	7,182	288,869
MicroRNA		
948	533	9,496
lncRNA		
139	297	901

After annotating GeneRIF, a comprehensive resource involving protein-coding genes, microRNA genes, and lncRNA genes could be obtained. The resource could be accessed from <http://www.bio-annotation.cn/gene2function/>.

2. Results

2.1. Mapping Genes to Gene Ontology and Disease Ontology. After annotating GeneRIFs by GO and DO (see Section 3), 288,869 associations between 13,148 mRNAs and 7,182 terms, 9,496 associations between 948 microRNAs and 533 terms, and 901 associations between 139 lncRNAs and 297 terms were obtained. The statistical information is shown in Table 1.

Figure 1(a) demonstrates the histogram of the number of genes associated with terms of GO and DO in the annotation results. 1,657 functional terms (23.0%) are associated with only one gene, while 3,924 functional terms (54.5%) are associated with more than three genes. The histogram of the number of terms associated with individual gene is shown in Figure 1(b). 1,375 genes (9.9%) are associated with only one functional terms, while 10,273 genes (74.3%) are associated with more than three genes.

The top ten terms ordered by the number of gene annotations and the top ten genes ordered by the number of term annotations are shown in Tables 2 and 3, respectively. Not surprisingly, several general terms in the top layer of the

TABLE 2: The top ten terms ordered by the number of gene annotations.

Term ID	Term name	Number of genes
GO:0005623	Cell	7,524
GO:0005488	Binding	5,011
GO:0065007	Biological regulation	4,846
GO:0023052	Signaling	4,466
GO:0032502	Developmental process	3,521
GO:0009058	Biosynthetic process	3,346
DOID:162	Cancer	3,139
GO:0006351	Transcription, DNA-templated	3,121
DOID:305	Carcinoma	3,069
GO:0040007	Growth	3,011

TABLE 3: The top ten genes ordered by the number of term annotations.

HGNC gene ID	Gene symbol	Number of functional terms
HGNC:11998	TP53	828
HGNC:11892	TNF	792
HGNC:6018	IL6	683
HGNC:12680	VEGFA	669
HGNC:11766	TGFB1	664
HGNC:3236	EGFR	560
HGNC:7176	MMP9	521
HGNC:391	AKT1	517
HGNC:7794	NFKB1	494
HGNC:6025	CXCL8	473

DAG have a larger number of genes associated with them, such as cell, binding, and developmental process (Table 2). The most prevalent disease terms appearing in the annotation result is cancer, which is associated with 3,139 genes (22.7% of all the terms). When we look at the genes associated with many terms, TP53 is the most prevalent genes appearing in the annotation result, which is associated with 828 terms (11.5% of all the genes).

2.2. Comparing with Existing Ontology Annotation Resources.

To validate the performance of our annotation result, we compared the result with the previous prevalent annotation resources GOA [13], in which human gene is manually annotated to GO. To ensure the exact evaluation, DO annotations of GeneRIFs were discarded, and annotations Inferred from Electronic Annotations (IEA) of GOA were removed.

In total, we obtained 196,423 associations between 4,613 GO terms and 13,107 genes in GeneRIFs and 168,246 associations between 13,920 GO terms and 16,724 genes in GOA. Only 10,658 associations and 3,375 GO terms appeared in both annotation resources. In comparison, both of them have more common genes (11,816).

Figures 2(a) and 2(b) demonstrate the histogram of the number of genes per GO term, and the histogram of the number of GO terms per gene in annotations of GeneRIFs and

TABLE 4: Data sources.

Data source	Web site (date of download)
GeneRIF	http://www.ncbi.nlm.nih.gov/gene/about-generif (Jun 2016)
HGNC	http://www.genenames.org/ (Jun 2016)
GO & GOA	http://geneontology.org/ (Jun 2016)
DO	http://disease-ontology.org/ (Jun 2016)

GOA, respectively. Obviously, more GO terms (4,545) could be annotated to only one gene in GOA than that (1,114) in GeneRIFs. In contrast, more genes (1,671) could be annotated to only one term in GeneRIFs than that (1,499) in GOA.

In order to evaluate the consistency, we compared the term frequency of individual gene and gene frequency of individual term in GeneRIFs and GOA. As a result, term frequency of individual gene in GeneRIF was significant positively correlated with it in GOA (Pearson correlation $\gamma^2 = 0.6401$, $p = 2.2e - 16$; Figure 2(c)), and gene frequency of individual term in GeneRIF was also significantly positively correlated with it in GOA (Pearson correlation $\gamma^2 = 0.1298$, $p = 3.686e - 14$; Figure 2(d)). Considering that GOA is most frequency used annotation resource, annotations of GeneRIFs should be also reliable.

2.3. *A Network Visualization Based on the Functional Annotation of the Human Genome.* Information in the annotation result can be used to describe the relationship among multiple genes or multiple terms. To this end, we create a bipartite network that describes the relationships between three genes (RNF2, RNF8, and RPS6) and 79 terms (Figure 3). Within this network, 33 terms are annotated to RNF2, 37 terms are annotated to RNF8, and 37 terms are annotated to RPS6. At the centre of the figure, 6 terms involving translation, execution phase of apoptosis, breast cancer, biological regulation, binding, and apoptotic process are related to all of these three genes. Using our annotation result, one can create this type of bipartite network as needed.

3. Materials and Methods

3.1. Data Collection

3.1.1. *GeneRIF.* GeneRIF was downloaded in June 2016 (Table 4). It involves five columns for describing tax identifier, NCBI gene ID, PubMed Unique Identifier (PMID), updated date, and function description. After extracting function descriptions of human genes, 650,079 descriptions remained.

3.1.2. *Normalized Gene Symbol Vocabulary.* The Human Genome Organisation Gene Nomenclature Committee (HGNC) [36] is responsible for approving unique symbols and names for human loci, including protein-coding genes and noncoding genes, to allow unambiguous scientific communication. In this paper, genes in GeneRIFs were normalized to HGNC gene symbols.

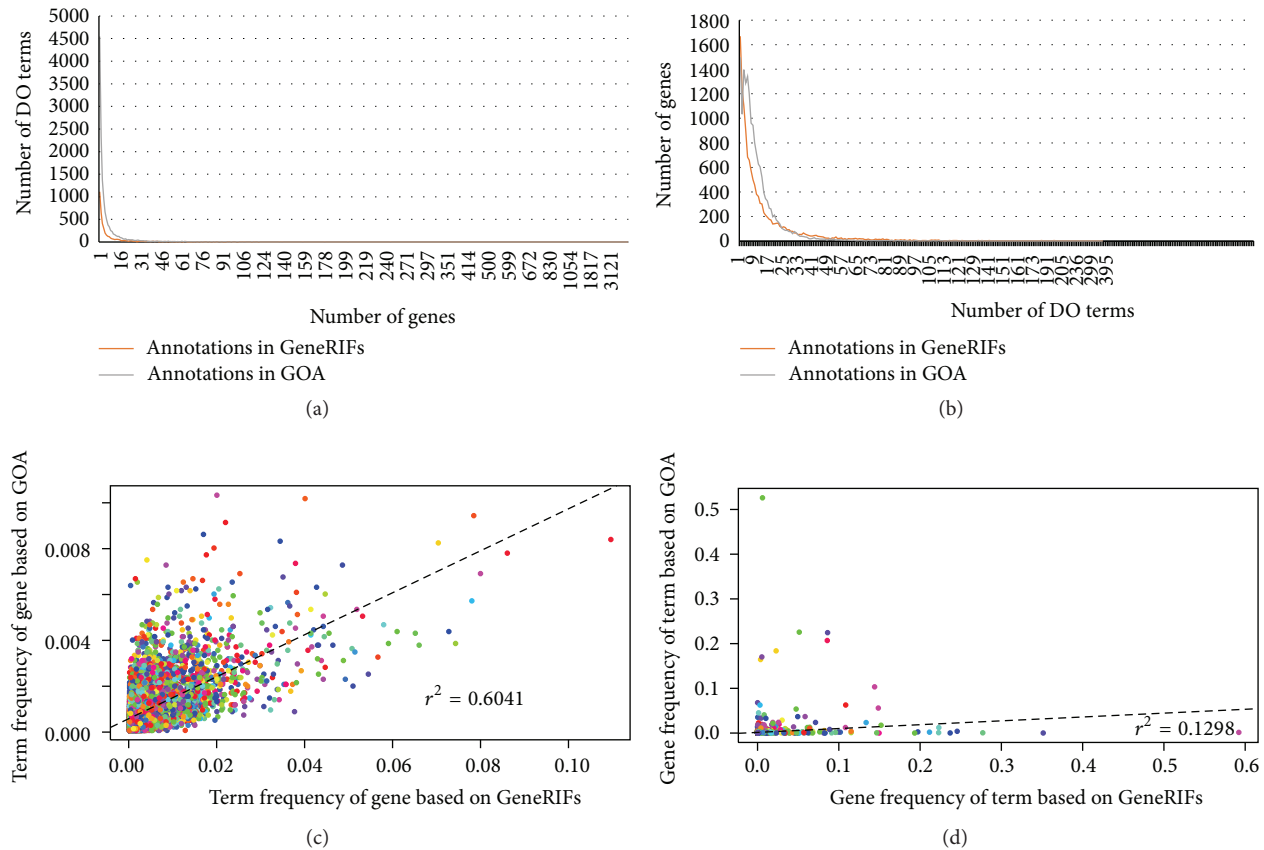


FIGURE 2: The comparison of annotations in GeneRIFs and with annotations in GOA. (a) Histogram of the number of genes associated with individual GO term. (b) Histogram of the number of DO terms associated with individual gene. (c) The correlation between term frequency of gene by GeneRIFs and GOA. (d) The correlation between gene frequency of term by GeneRIFs and GOA.

3.1.3. Ontologies and Annotations. As shown in Figure 4, GO organized BP terms in the Directed Acyclic Graph (DAG) by “IS_A” relationship. Currently, GO contains 55,565 “IS_A” relationships between 28,654 BP terms, 12,375 “IS_A” relationships between 10,159 MF terms, and 5,618 “IS_A” relationships between 3,907 CC terms. GOA was compared with our annotation result. After removing IEA and getting rid of duplicate records of GOA, 168,246 associations between 13,920 GO terms and 16,724 genes remained.

DO is a first ontology to organize terms around human disease, which describes each disease by a unique identifier, a disease name, and its synonymous. In the current version, it involved 7,124 “IS_A” relationships between 6,920 disease terms.

3.2. Method for Annotating Human Genome. As shown in Figure 5(a), we presented a framework, Gene2Function, to annotate the function of human genome. Firstly, a raw text of GeneRIF with functional description should be annotated by a text mining tool named Open Biomedical Annotator (OBA) [37], which provided an ontology-based web service that annotates public datasets with biomedical ontology concepts

based on their textual metadata. As a result, the functional description will be mapped to the corresponding ontologies, such as GO and DO. Then, the Entrez gene identifier will be converted into a normalized gene symbol. Here, HGNC was exploited for normalizing and labelling the locus type of gene, such as protein-coding genes, microRNA genes, and lncRNA genes. Finally, each GeneRIF could be annotated to a triple involving gene symbol, locus type, and functional description.

All the GeneRIFs could be annotated based on the annotation framework. Figure 5(b) gives an example of annotating a GeneRIF with GO. “Enzyme activity” is a synonym of “catalytic activity (GO:0003824),” which was identified by OBA. And Entrez gene identifier “9” was converted into “NAT1 (HGNC:7645)” based on HGNC. Through the annotation framework, the annotation triple “mRNA, NAT1, catalytic activity” could be obtained.

4. Discussion

The importance of the functional annotations of genes had been reflected in the previous annotation resource, such as

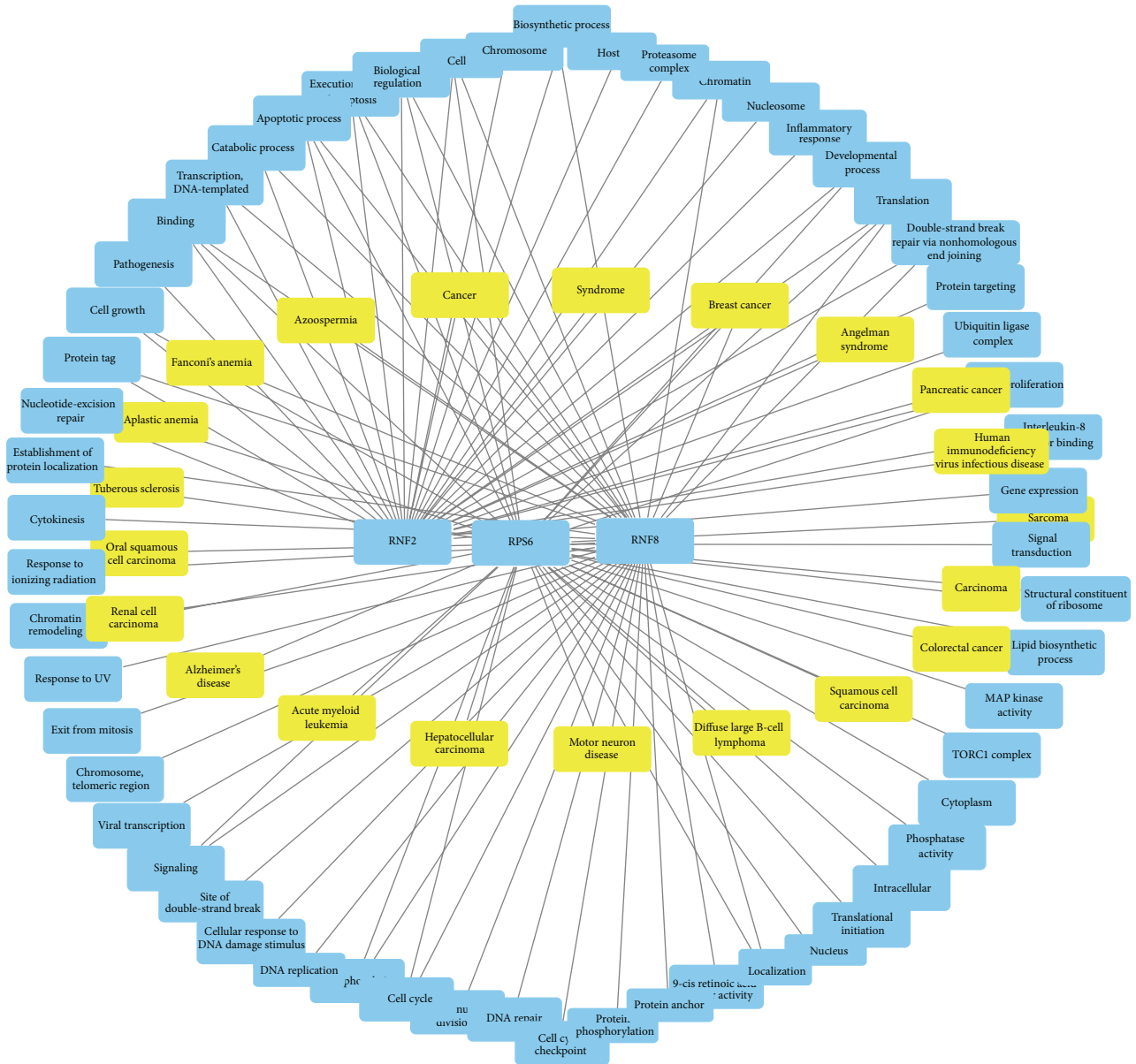


FIGURE 3: A bipartite network demonstrating the relationship between genes and terms. Rectangles with yellow represent DO terms, three rectangles with blue in the center of the figure indicate DO terms, and other rectangles are GO terms. An edge is placed between a gene and a term of GO and DO if the gene relates with the term.

GOA. Unfortunately, functional annotation resources of non-coding RNA are very few, which lead to the lack of a comprehensive annotation resource involving protein-coding genes, microRNA genes, and lncRNA genes. With the largest number of noncoding genes in the human genome, it is urgent to provide functional annotation of these genes. In this study, we presented a framework, Gene2Function, for annotating GeneRIFs. As a result, a comprehensive functional annotation resource of human genome was obtained based on the framework, which could be accessed at <http://www.bio-annotation.cn/gene2function/>. To evaluate the reliability, our annotation result was compared with a prevalent resource GOA. Subsequently, a network visualization of connectivity

of genes by their functional terms shows the usability of the annotation result.

The annotation framework is based on a text mining tool OBA [37]. Under the framework, the functional terms of descriptions of GeneRIFs were annotated to GO and DO terms. And gene symbols were mapped to a normalized vocabulary of human gene HGNC [36], which makes it easy to distinguish the locus type of gene, such as protein-coding RNA, microRNA, and lncRNA.

The consistency test of the GeneRIFs and GOA (Figures 2(c) and 2(d)) shows the reliability of our annotation result. Because of a small amount of common associations between

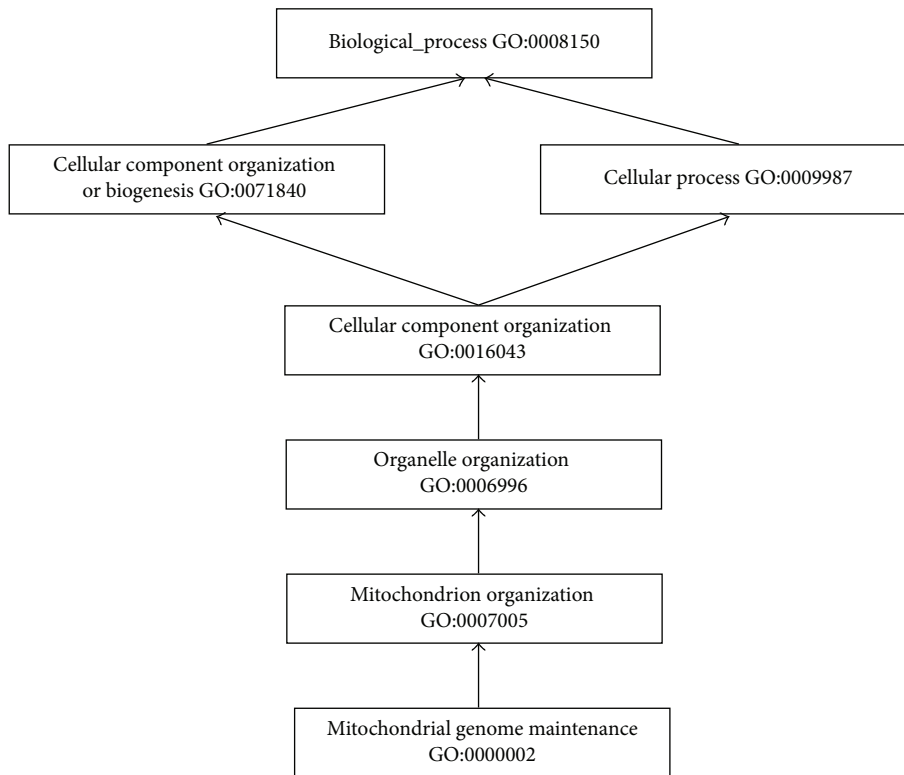


FIGURE 4: A subgraph of the DAG for BP term “Mitochondrial genome maintenance (GO:0000002).” The arrow symbol represents an “IS_A” link of GO. For example, “Mitochondrial genome maintenance (GO:0000002)” is linked to “Mitochondrion organization (GO:0007005)” by an “IS_A” relationship.

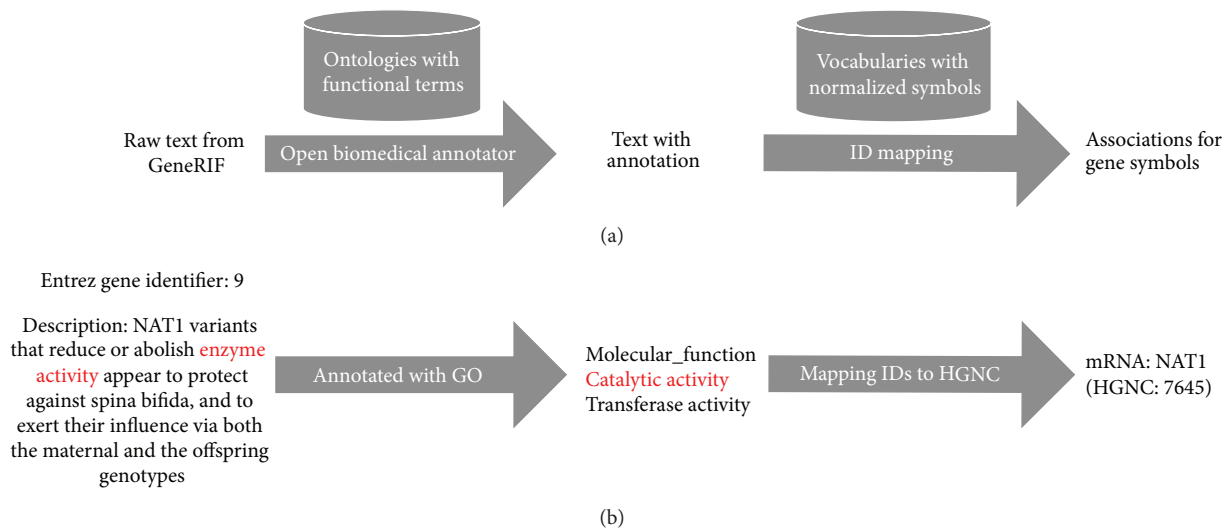


FIGURE 5: Diagram of functional annotation of human genome. (a) A framework to annotate functional description of human genome to ontologies. (b) An example of annotating a GeneRIF.

genes and GO terms in both annotation resources, they could be complementary in the usage of protein-coding RNA annotation. More GO terms were annotated in GOA (see Section 2) suggesting it is more deep and serious than our annotation results. In comparison, advantage of GeneRIFs is that not only protein-coding genes but also microRNA genes

and lncRNA genes could be annotated with GO and other function terms (Table 1).

Competing Interests

The authors declare that they have no competing interests.

References

- [1] J. C. Venter, M. D. Adams, E. W. Myers et al., "The sequence of the human genome," *Science*, vol. 291, no. 5507, pp. 1304–1351, 2001.
- [2] V. Ambros, "The functions of animal microRNAs," *Nature*, vol. 431, no. 7006, pp. 350–355, 2004.
- [3] S. M. Hollenberg, C. Weinberger, E. S. Ong et al., "Primary structure and expression of a functional human glucocorticoid receptor cDNA," *Nature*, vol. 318, no. 6047, pp. 635–641, 1985.
- [4] W. E. Holmes, J. Lee, W.-J. Kuang, G. C. Rice, and W. I. Wood, "Structure and functional expression of a human interleukin-8 receptor," *Science*, vol. 253, no. 5025, pp. 1278–1280, 1991.
- [5] R. R. Schumann, S. R. Leong, G. W. Flaggs et al., "Structure and function of lipopolysaccharide binding protein," *Science*, vol. 249, no. 4975, pp. 1429–1431, 1990.
- [6] H. Shi, G. Zhang, M. Zhou et al., "Integration of multiple genomic and phenotype data to infer novel miRNA-disease associations," *PLoS ONE*, vol. 11, no. 2, Article ID e0148521, 2016.
- [7] M. Zhou, Y. Sun, Y. Sun et al., "Comprehensive analysis of lncRNA expression profiles reveals a novel lncRNA signature to discriminate nonequivalent outcomes in patients with ovarian cancer," *Oncotarget*, vol. 7, no. 22, pp. 32433–32448, 2016.
- [8] M. Zhou, W. Xu, X. Yue et al., "Relapse-related long non-coding RNA signature to improve prognosis prediction of lung adenocarcinoma," *Oncotarget*, vol. 7, no. 20, pp. 29720–29738, 2016.
- [9] H. Shi, J. Xu, G. Zhang et al., "Walking the interactome to identify human miRNA-disease associations through the functional link between miRNA targets and disease genes," *BMC Systems Biology*, vol. 7, article 101, 2013.
- [10] R. Brain and J. R. Jenkins, "Human p53 directs DNA strand reassociation and is photolabeled by 8-azido ATP," *Oncogene*, vol. 9, no. 6, pp. 1775–1780, 1994.
- [11] S. Benzon Larsen, U. Vogel, J. Christensen et al., "Interaction between ADH1C Arg²⁷²Gln and alcohol intake in relation to breast cancer risk suggests that ethanol is the causal factor in alcohol related breast cancer," *Cancer Letters*, vol. 295, no. 2, pp. 191–197, 2010.
- [12] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.
- [13] E. Camon, M. Magrane, D. Barrell et al., "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology," *Nucleic Acids Research*, vol. 32, pp. D262–D266, 2004.
- [14] W. A. Kibbe, C. Arze, V. Felix et al., "Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data," *Nucleic Acids Research*, vol. 43, no. 1, pp. D1071–D1078, 2015.
- [15] Z. Lu, K. B. Cohen, and L. Hunter, "Generif quality assurance as summary revision," in *Proceedings of the Pacific Symposium on Biocomputing (PSB '07)*, pp. 269–280, January 2007.
- [16] J. D. Osborne, J. Flatow, M. Holko et al., "Annotating the human genome with Disease Ontology," *BMC Genomics*, vol. 10, supplement 1, article S6, 2009.
- [17] W. Xu, H. Wang, W. Cheng et al., "A framework for annotating human genome in disease context," *PLoS ONE*, vol. 7, no. 12, Article ID e49686, 2012.
- [18] K. Peng, W. Xu, J. Zheng et al., "The Disease and Gene Annotations (DGA): an annotation resource for human disease," *Nucleic Acids Research*, vol. 41, no. 1, pp. D553–D560, 2013.
- [19] P. Kapranov, A. T. Willingham, and T. R. Gingeras, "Genome-wide transcription and the implications for genomic organization," *Nature Reviews Genetics*, vol. 8, no. 6, pp. 413–423, 2007.
- [20] E. S. Lander, L. M. Linton, B. Birren et al., "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.
- [21] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and promising identification of human MicroRNAs by incorporating a high-quality negative set," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 1, pp. 192–201, 2014.
- [22] Y. Huang, N. Liu, J. P. Wang et al., "Regulatory long non-coding RNA and its functions," *Journal of Physiology and Biochemistry*, vol. 68, no. 4, pp. 611–618, 2012.
- [23] J. Sun, H. Shi, Z. Wang et al., "Inferring novel lncRNA-disease associations based on a random walk model of a lncRNA functional similarity network," *Molecular BioSystems*, vol. 10, no. 8, pp. 2074–2081, 2014.
- [24] M. Zhou, X. Wang, H. Shi et al., "Characterization of long non-coding RNA-associated ceRNA network to reveal potential prognostic lncRNA biomarkers in human ovarian cancer," *Oncotarget*, vol. 7, no. 11, pp. 12598–12611, 2016.
- [25] J. Sun, M. Zhou, H. Yang, J. Deng, L. Wang, and Q. Wang, "Inferring potential microRNA-microRNA associations based on targeting propensity and connectivity in the context of protein interaction network," *PLoS ONE*, vol. 8, no. 7, Article ID e69719, 2013.
- [26] Y. A. Huang, X. Chen, Z. H. You, D. S. Huang, and K. C. Chan, "ILNCSIM: improved lncRNA functional similarity calculation model," *Oncotarget*, vol. 7, no. 18, pp. 25902–25914, 2016.
- [27] X. Chen, Y. A. Huang, X. S. Wang, Z. H. You, and K. C. Chan, "FMLNCSIM: fuzzy measure-based lncRNA functional similarity calculation model," *Oncotarget*, 2016.
- [28] A. P. Daiwile, S. Sivanesan, A. Izzotti et al., "Noncoding RNAs: possible players in the development of fluorosis," *BioMed Research International*, vol. 2015, Article ID 274852, 10 pages, 2015.
- [29] Y.-A. Huang, Z.-H. You, X. Chen, K. Chan, and X. Luo, "Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding," *BMC Bioinformatics*, vol. 17, no. 1, article 148, 2016.
- [30] Y. Li, C. Qiu, J. Tu et al., "HMDD v2.0: a database for experimentally supported human microRNA and disease associations," *Nucleic Acids Research*, vol. 42, no. 1, pp. D1070–D1074, 2014.
- [31] G. Chen, Z. Wang, D. Wang et al., "LncRNADisease: a database for long-non-coding RNA-associated diseases," *Nucleic Acids Research*, vol. 41, no. 1, pp. D983–D986, 2013.
- [32] C. E. Lipscomb, "Medical subject headings (MeSH)," *Bulletin of the Medical Library Association*, vol. 88, no. 3, pp. 265–266, 2000.
- [33] X. Zeng, X. Zhang, and Q. Zou, "Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks," *Briefings in Bioinformatics*, vol. 17, no. 2, pp. 193–203, 2016.
- [34] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016.
- [35] Q. Zou, J. Li, Q. Hong et al., "Prediction of microRNA-disease associations based on social network analysis methods," *BioMed*

Research International, vol. 2015, Article ID 810514, 9 pages, 2015.

- [36] E. A. Bruford, M. J. Lush, M. W. Wright, T. P. Sneddon, S. Povey, and E. Birney, "The HGNC database in 2008: a resource for the human genome," *Nucleic Acids Research*, vol. 36, no. 1, pp. D445–D448, 2008.
- [37] C. Jonquet, N. H. Shah, and M. A. Musen, "The open biomedical annotator," *AMIA Summits on Translational Science Proceedings*, vol. 2009, pp. 56–60, 2009.