Using Ontology Fingerprints to evaluate genome-wide association study results

Lam C. Tsoi[1], Michael Boehnke[2], Richard L. Klein[3,4], W. Jim Zheng[1,*]

[1]Department of Biostatistics, Bioinformatics & Epidemiology, Medical University of South Carolina, Charleston, SC; [2]Department of Biostatistics and Center for Statistical Genetics, School of Public Health, University of Michigan, Ann Arbor, MI; [3]Division of Endocrinology, Metabolism, and Medical Genetics, Department of Medicine, Medical University of South Carolina; [4]Research Service, Ralph H. Johnson Department of Veterans Affairs Medical Center, Charleston, SC

Abstract

We describe an approach to characterize genes or phenotypes via ontology fingerprints which are composed of Gene Ontology (GO) terms overrepresented among those PubMed abstracts linked to the genes or phenotypes. We then quantify the biological relevance between genes and phenotypes by comparing their ontology fingerprints to calculate a similarity score. We validated this approach by correctly identifying genes belong to their biological pathways with high accuracy, and applied this approach to evaluate GWA study by ranking genes associated with the lipid concentrations in plasma as well as to prioritize genes within linkage disequilibrium (LD) block. We found that the genes with highest scores were: ABCA1, LPL, and CETP for HDL; LDLR, APOE and APOB for LDL; and LPL, APOA1 and APOB for triglyceride. In addition, we identified some top ranked genes linking to lipid metabolism from the literature even in cases where such knowledge was not reflected in current annotation of these genes. These results demonstrate that ontology fingerprints can be used effectively to prioritize genes from GWA studies for experimental validation.

Introduction

Genome-wide association (GWA) studies have become a feasible and important method to identify loci that are associated with a particular phenotype . Assessing quantitatively the likely importance of genes identified as significant to disease risk based on biological facts is essential to proceed efficiently toward experimental validation processes and, ultimately, to define the causal relationships between genes and phenotypes.

Various text-mining methods have been developed to extract information from the biomedical literature for gene annotation . In addition, GO provides a standardized characterization of gene functions . Despite the fact that biomedical literatures were written without GO in mind, it has been shown that GO terms that can be identified in PubMed abstracts tend to occur frequently in the literature . Therefore, GO as a standardized terminology provides a semantic grounding to mine the PubMed literature.

Here we describe a comprehensive analysis combining text mining of PubMed abstracts and GO with quantitative measure to assemble ontology fingerprints for genes and phenotypes, and a method to calculate a similarity score between two ontology fingerprints. We further describe how comparing the ontology fingerprints of a phenotype with that of genes identified in a GWA study can be used to prioritize genes for follow-up investigation, including fine mapping and functional studies.

Methods

Data

We used the June 13th, 2007 version of GO and 2007 version of PubMed abstracts for this study. The PubMed abstracts and the genes annotated were obtained from the NCBI "pubmed2gene" file. Abstracts that contained GO terms were also annotated by mapping each term to the abstracts using exact string match. Since GO is a Directed Acyclic Graph (DAG) , abstracts containing a GO term were also labeled with all the parents of that GO term in the GO hierarchy as well. In addition, each abstract was labeled with a GO term only once regardless of how many times the term occurred. Because we were attempting to decipher human gene-phenotype relationships, the ontology fingerprints were derived from abstracts linked to human genes. In total, we retrieved 178,687 abstracts, and we constructed ontology fingerprints for all 25,357 human genes. There were 5,001 ontology terms that mapped to PubMed abstracts linked to human genes.

Enrichment test

To test whether a GO term appeared more often in PubMed abstracts linked to a gene than in the rest of the PubMed abstracts linked to other human genes,

$$p - value = \frac{1}{2}P(A_{obs} = e) + P(A_{obs} > e) \qquad \text{Equation 1}$$

we performed a hypergeometric test, resulting in a list of GO terms with p-values for each gene. Due to the discreteness of the hypergeometric distribution, the mid p-value was used in the calculation :

For each gene and ontology pair, A_T is the total number of abstracts considered, while A_O and A_G denotes the number of abstracts linked to the ontology term and gene respectively; number of abstracts that linked to both the ontology term and the gene is labeled as e. A_{obs} is the random variable of observing the number of abstracts linked to both the ontology term and the gene. The p-value was then adjusted to remove insignificant GO terms (See Supplementary information for details).

We also performed the same test on each phenotype-ontology pair. While each gene or phenotype has a list of ontology terms serving as ontology fingerprints defined as ontology terms with p-value <1, collectively the terms and the quantification reflect the characteristics of the gene or phenotype.

Similarity score calculation

The ontology fingerprint characterizes the cellular component, molecular function, or biological process of a gene or a phenotype with a quantitative measure. By comparing how similar the ontology fingerprints between a gene and a phenotype are, we can infer to what extent a gene may be related to the phenotype. We calculate a similarity score using a modified

$$S_j = \frac{\sum_{i=1}^{O} log(q_i) log(r_{ij})}{max\left\{1, \sum_{i=1}^{O} [I(q_i < 1) I(r_{ij} = 1)]\right\}} \qquad \text{Equation2}$$

version of the inner product:

$i = 1, 2, ..., O$ represents the ontology terms, and the r_{ij} and q_i represent the adjusted p-values of the i^{th} ontology term of the gene j and the phenotype term, respectively. We took the logarithm of the probabilities to prevent underflow during computation. In the numerator, ontology terms that have adjusted p-values equal 1.0 for either the gene or phenotype (i.e. not in either of the gene's or phenotype's fingerprint) will have a score of zero for that ontology term i, and thus make no contribution. Each similarity score is then normalized by $\sum_{i=1}^{O} I(q_i < 1) I(r_{ij} = 1)$, which is the number of ontology terms in the fingerprint of the phenotype but not in that of gene j. The normalization intends to give more weight on a gene's ontology fingerprint that has a higher degree of overlapping terms with the phenotype's ontology fingerprint. If all of the ontology terms of a phenotype overlap with those of a gene, 1 is used in the denominator. Note from Equation 2 that an ontology term with low adjusted p-values for both the phenotype and the gene would contribute significantly to the similarity score. Therefore, the equation considers both the number of GO terms in the ontology fingerprints and the significance level indicated by the p-value. A p-value threshold (λ) was selected and applied to calculate similarity score between genes and phenotypes (See Supplementary information for detail).

Significant genes identified from GWA study

We applied our approach to a GWA study that investigated the influences of loci on lipid concentrations, HDL, LDL, and triglyceride . Genes within or overlap with the top linkage disequilibrium (LD) blocks of best SNPs for each trait were obtained as significantly associated with the corresponding trait (top 199, 201 and 200 LD blocks for LDL, HDL and TG respectively). Independent loci were defined as having low correlation ($r^2 < 0.2$) with any other higher ranking SNP. The p-value of the most significant SNP within each block was used.

Results

Ontology Fingerprints

We computed the association of genes or phenotypes with GO terms by using the hypergeometric enrichment test. The p-values from the test (raw p-values) were then adjusted, taking into consideration the number of ontology terms associated with the genes or phenotypes. The purpose of the adjustment was to reduce the impact of insignificant ontology terms on the ontology fingerprints of genes or phenotypes that have been extensively studied. The resulting ontology terms with adjusted p-values collectively served as the ontology fingerprint for the gene or phenotype, with the p-value for each term reflecting the significance of the term's enrichment among the abstracts associated with the gene or phenotype. Only terms with adjusted p-values < 1.0 were used to define the ontology fingerprints for the gene or phenotype. Table 1 illustrates a small portion of the ontology fingerprint for the gene *VEGFA*, which encodes vascular endothelial growth factor A. This ontology fingerprint serves as a comprehensive, quantitative characterization of the gene using well-defined ontology terms.

Table 1. Eight out of the 279 GO terms in the ontology fingerprint for *VEGFA*. Full list is shown in Supplementary Table 1.

GO id	GO term	Adjusted p-value
GO#GO_0008083	Growth Factor	1.00×10^{-323}
GO#GO_0001525	Angiogenesis	1.00×10^{-323}
…	…	…
GO#GO_0008283	Cell Proliferation	1.52×10^{-6}
GO#GO_0006928	Cell Motility	1.71×10^{-6}
…	…	…
GO#GO_0004714	Transmembrane Receptor Protein Tyrosine Kinase	2.60×10^{-1}
GO#GO_0002253	Activation of Immune Response	2.64×10^{-1}
…	…	…
GO#GO_0042098	T Cell Proliferation	9.35×10^{-1}
GO#GO_0003773	Heat Shock Protein	9.58×10^{-1}
…	…	…

Similarity Scores between Genes and Phenotypes

By comparing the genes' and phenotypes' ontology fingerprints, we calculated similarity scores to quantify the relevance of particular genes to phenotypes. We tested our approach by using 10 randomly selected KEGG pathways as phenotype domains for evaluation. The AUCs for the 10 pathways are shown in Table 2 (column "Ontology Fingerprint AUC"). We compared our approach to a similar text-mining approach which uses "concept profiles" to evaluate the association between different biological concepts . Table 2 shows how well the ontology fingerprint approach and this Anni 2.0 system correctly associated genes with their corresponding KEGG pathways. Specifically, our ontology fingerprint-based method has higher AUC for associating genes with their corresponding pathways than Anni 2.0. 1. We attribute such significant improvement to the employment of Gene Ontology, a well-developed controlled vocabulary to characterize the biological features of genes and phenotypes, the hypergeometric test, which highly increases the sensitivity for detecting the associated ontology terms, and our scoring method, which emphasizes on the number of ontology terms characterizing both the gene and the phenotype.

Table 2. Ontology Fingerprints-derived similarity scores can correctly assign genes to their corresponding pathways. The area under ROC curves for each of 10 KEGG pathways are shown. The middle column shows the results from the Ontology fingerprint method, while the right column is the result from the Anni 2.0; * represents the difference between the two methods is significant at 0.0001 level by the Wilcoxon rank-sum test.

Pathway	Ontology Fingerprint AUC	Anni 2.0 AUC	p-value from Wilcoxon Test
Apoptosis	0.96	0.85*	5.56×10^{-19}
Biosynthesis of steroids	0.75	0.73	0.66
Fatty acid metabolism	0.88	0.86	0.14
Focal Adhesion	0.94	0.87*	4.06×10^{-11}
Galactose metabolism	0.90	0.78*	7.64×10^{-9}
Glycolysis	0.80	0.72*	1.86×10^{-6}
MAP kinase signaling	0.90	0.78*	2.21×10^{-14}
Prostate cancer	0.95	0.91*	3.80×10^{-8}
Renal cell carcinoma	0.93	0.81*	1.65×10^{-12}
Sphingolipid metabolism	0.89	0.72*	2.09×10^{-9}

Using Ontology Fingerprints to Prioritize Genes from GWA Studies

We applied our method to evaluate the results from a GWA analysis studying the genetic variants influencing plasma lipid concentrations, including High-density lipoprotein (HDL), Low-density lipoprotein (LDL), and Triglyceride (TG). Among genes strong associations with lipid concentration, many are not clearly identified in their annotation as being relevant to lipid metabolism. Within the top-ranked genes are quite a few well-known cholesterol related genes, including cholesterol ester transfer protein, plasma (*CETP*), low density lipoprotein receptor (*LDLR*), lipoprotein lipase (*LPL*). Simply based on the gene annotations alone, there are 10, 8, and 12 genes related to the lipid mechanism among the top 20 genes with highest similarity scores. For the remaining genes that do not have Entrez Gene annotation to be associated with the lipid metabolism, we found that there are additional 3, 9 and 7 genes that could potentially influence the

HDL, LDL and TG concentrations respectively by tracing back to the GO terms and the literatures that contributed to the similarity scores. One example is transferrin (*TF*), which is ranked by the similarity score among the top 20 genes for HDL. While current annotation of *TF* does not show any relevance to lipid or lipid metabolism, we found that Cubilin (*CUBN*), an endocytic receptor, can act as a receptor for both transferrin and apolipoprotein A1 . Another example is thyroid hormone receptor beta (*THRB*). *THRB* was found to negatively regulate the lipoprotein lipase inhibitor , and the agonist of *THRB* is associated with a decrease of triglyceride concentration in rats . Neither the relationship of *THRB* to nor its influence on the concentration of triglycerides in humans is established, so the annotation for this gene shows no direct link to lipid metabolism. Our results indicate that the ontology fingerprint method can identify genes relevant to the phenotypes revealed through GWA study (The top 20 ranked genes are listed in supplementary Table 2).

Conclusion

Even though several text mining approaches have been developed to identify relationships between genes and phenotypes, our approach is significantly different in several aspects: 1) a hypergeometric enrichment test was used to focus on identifying overrepresented ontology terms for genes and phenotypes in relevant PubMed abstracts; 2) ontology fingerprints with quantitative measures, rather than individual ontology term annotations, were used to capture comprehensive characteristics of genes and phenotypes; 3) a method to calculate similarity scores between ontology fingerprints evaluated the relevance between genes and phenotypes.

*The Supplementary information can be found at:
http://genomebioinfo.musc.edu/OntoFinger/

Acknowledgement

This work is supported by grants IRG 97-219-08 from the ACS, a pilot project of Grant 5 P20 RR017696-05 and PhRMA Foundation Research Starter Grant (WJZ), DK62370 and HG00376 (MB), and NLM training grant 5-T15-LM007438-02 (LCT).

References