Conceptualization of molecular findings by mining gene annotations

Vicky Chen, MS, and Xinghua Lu, MD, PhD
University of Pittsburgh
Department of Biomedical Informatics

Abstract

Background: Contemporary genome-scale studies often return a long list of genes of potential interest. A challenging task is to reveal the major functional themes the genes are involved in at a conceptual level. The Gene Ontology (GO) is an ontology representing molecular biology concepts related to genes and their products, and current annotations from the GO Consortium be highly specific. There is a need for tools that reveal the major functional themes through mining and representing semantic information of genes in an objective and quantitative manner.

Methods: In this study, we utilized the organization of the GO to derive a more abstract representation of the major biological processes of a list of genes based on their annotations. We cast the task as follows: given a list of genes, identify non-disjoint, functionally coherent subsets, such that the functions of the genes in a subset are summarized by an informative GO term that accurately captures the semantic information of the original annotations.

Results: We evaluated different metrics for assessing information loss when merging GO terms, and different statistical schemes to assess the functional coherence of a gene set. We found that the best discriminative power was achieved by using a combination of the information-content-based measure as the information loss metric, and graph-based statistics derived from a Steiner tree connecting genes in an augmented GO graph.

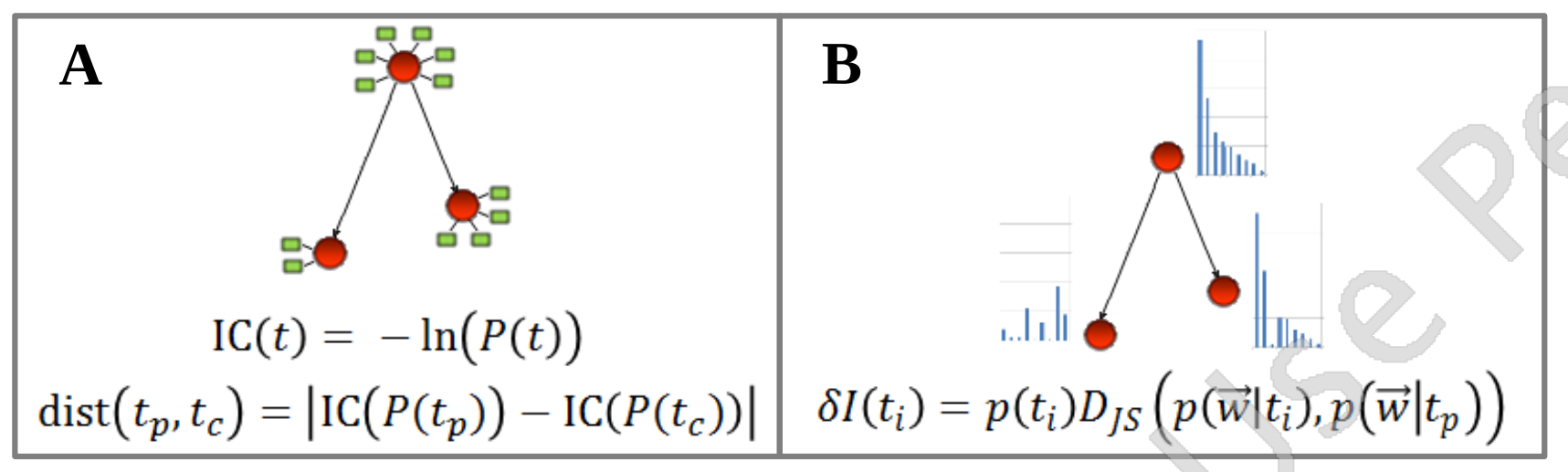
Conclusions: Our methods provide an objective and quantitative approach to capturing the major directions of gene functions in a context-specific fashion.

GOGrapher

We developed a Python software package that can represent:

- Gene Ontology structure
- Gene, protein, and PubMed annotations
- Semantic or gene-based term distances

Each node represents a GO term, each edge represents the relationship between two terms, and edge weights represent term distances. Semantic distance denotes the different in semantic meaning of concepts. Gene-based distance denotes the amount of genes shared between concepts.



$$\mathrm{IC}(t) = -\ln(P(t))$$

$$\mathrm{dist}(t_p, t_c) = |\mathrm{IC}(P(t_p)) - \mathrm{IC}(P(t_c))|$$

$$\delta I(t_i) = p(t_i) D_{JS}\left(p(\vec{w}|t_i), p(\vec{w}|t_p)\right)$$

Figure 1. Examples of term weight measures
A) The distance measured using information content (IC) is greater for terms that have a smaller number of shared genes. B) The distance measured through KL divergence for information bottleneck (IB) is greater for word distributions that differ more.

Measuring Information Loss

We used two statistical schemes to calculate the information loss from grouping genes into a subset:

- Lowest Common Ancestor tree length
- Steiner tree length

We also tried augmenting graphs by connecting nodes with shared genes.

Functional Coherence

Functional coherence is a measure of how closely related the functions associated with a group of genes are. We use this to determine the probability of a given amount of information loss occurring at random. The means and variances used to calculate the probability were generated from repeated sampling of random genes of different set sizes.

Term Distance Comparison

We investigated the characteristics of the different distance measures to determine which best assessed information loss. We found that the edge weight distributions are dominated by edges with short distances. The IB distribution is dominated by outliers while IC has a more consistent decrease in distance over level.

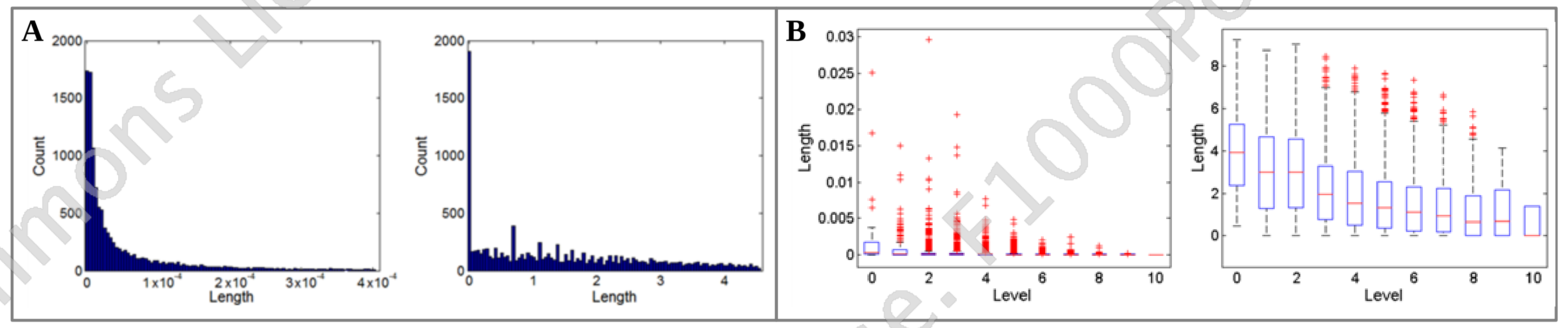


Figure 2. Distribution of edge weights
Both A and B are organized with the IB-based edge weight plot on the left and the IC-based edge weight plot on the right. A) Distribution of the shortest 90% of edges in the entire graph. B) Boxplots of the edge weight distribution organized according to the level of edge, where level 0 contains edges that connected to root.

Evaluating Metrics

We tried information loss and distance metrics combinations to determine which had the highest discriminative power. Human KEGG pathways were used as coherent gene sets and randomly drawn gene sets were non-coherent. The results indicate that the combination of IC and Steiner tree on an augmented GO graph performs the best.

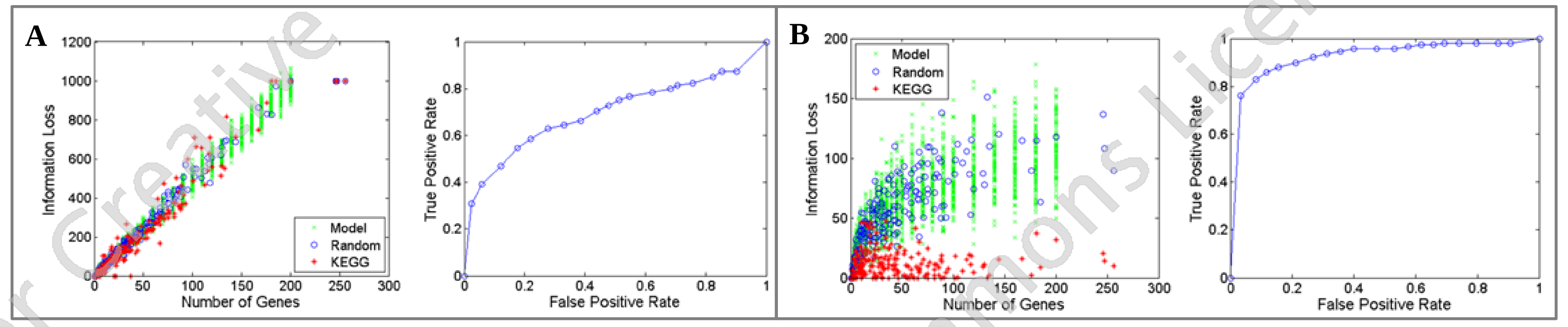


Figure 3. Example distributions of statistics and discriminative power of coherence models
A) Statistics derived from KEGG gene sets (red), the matched random gene sets (blue), and random gene sets for coherence model building (green) were plotted. On the right is the corresponding ROC curve. B) Scatterplot of the graph-based statistics and ROC curve of the model using IC and Steiner tree length based on an augmented GO graph.

Finding multiple function aspects in KEGG pathways

We investigated if it made sense to treat a large KEGG pathway as coherent. One pathway we looked at was the MAPK signaling pathway. Our model returned multiple subsets reflecting different functions these genes participate in. Therefore, it is sensible that our model treated the full gene set as not-coherent.

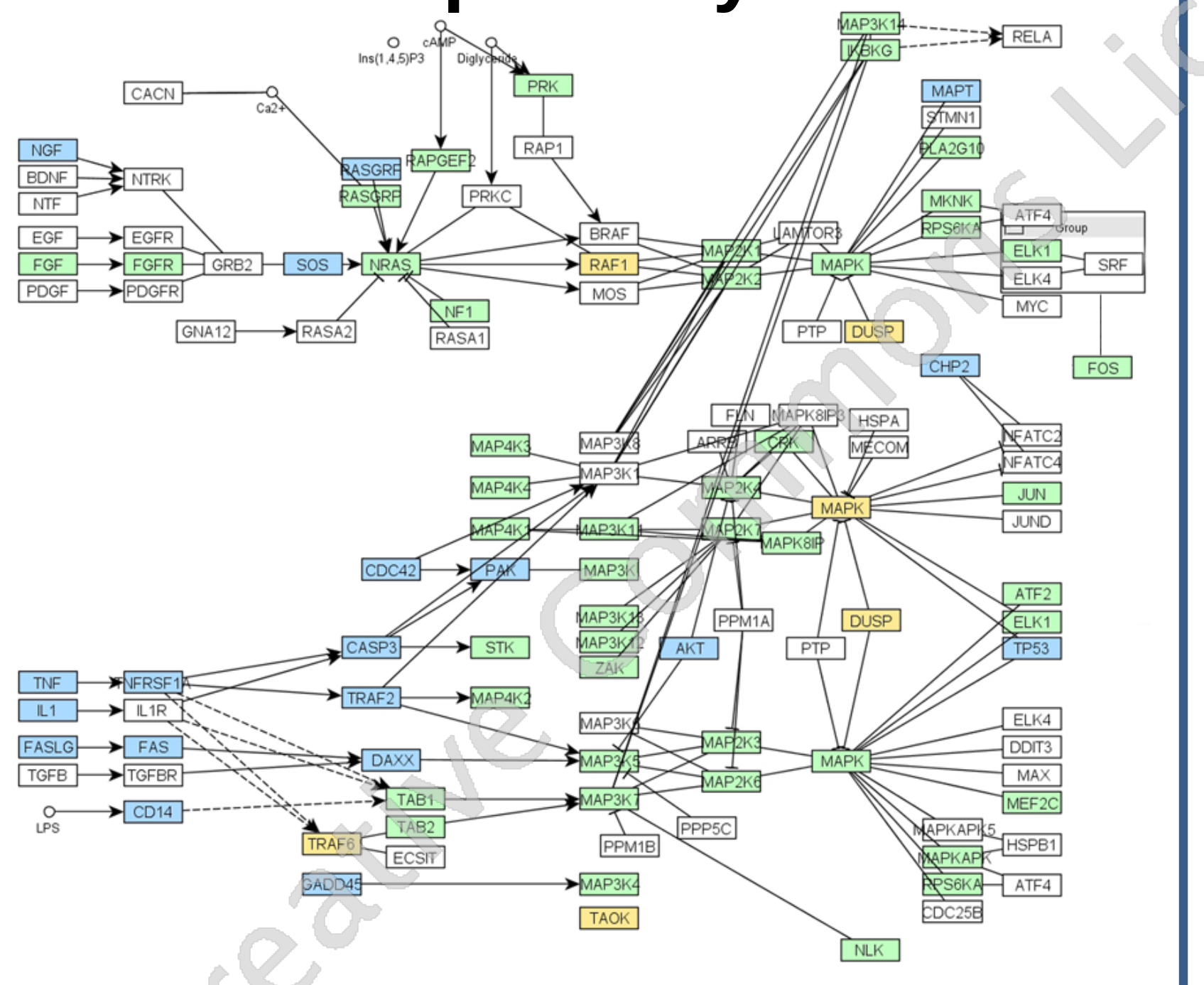


Figure 4. A KEGG pathway containing genes involved in multiple processes
The *MAPK signaling pathway* (hsa04010) is shown. Two functionally coherent subsets are highlighted. The genes summarized by GO:0023014 (Signal transduction by phosphorylation) are in green and GO:0006915 (Apoptotic process) are in blue. Genes involved in both biological processes are in yellow.

Application in Real World Data Analysis

We identified a list of differentially expressed genes in ovarian cancer to compare our methods with original GO annotations, GO enrichment analysis, and GO slim mapping. We plotted the distribution of the terms based on their level. The results show that summarizing terms identified by our method tend to be more specific than GO slims, and more general than original and enriched annotations while covering more genes.

Protein-protein interaction evaluation

To support the notion that the our model measures functional coherence, we assessed the functional relatedness of the proteins in a subset using the within module PPI ratio. We applied our model using 3 different p-value cutoff thresholds. The results indicate that our metric agrees with another metric reflecting the functional coherence of genes.

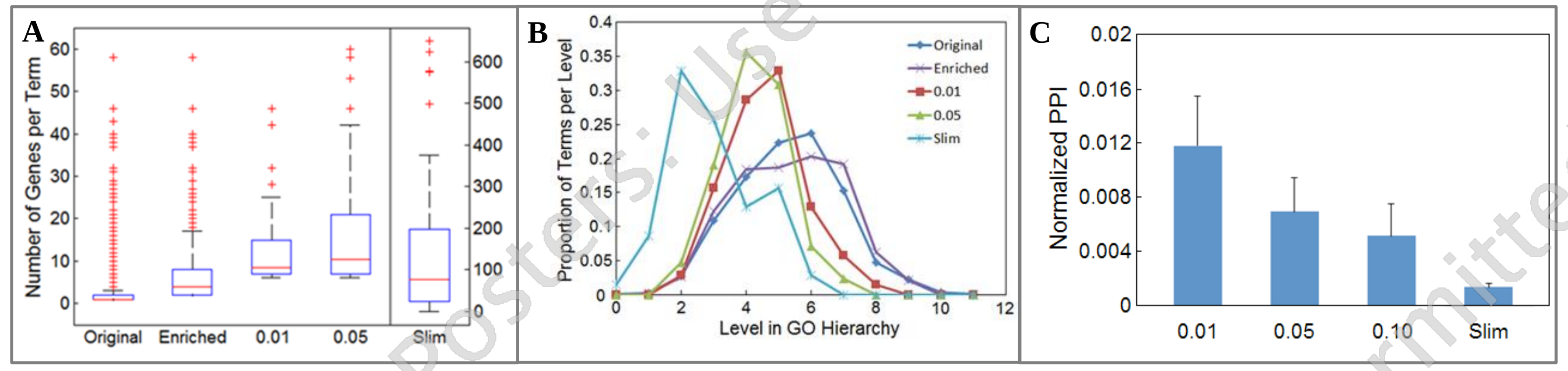


Figure 5. Distribution of genes and average within module PPI associated with summarizing GO terms
A) Boxplots of the distribution of number of genes associated per term under five different conditions: original GO annotation, enriched GO annotations, our method with a p-value ≤ 0.01 and 0.05 thresholds, and the Generic GO slim. B) Plot of the proportion of terms per level in the GO hierarchy under the five conditions. C) Plots of the average within module PPI ratio for the summarizing terms that resulted from different thresholds of merging the AMIGO GO Slim tool. The whisker denotes the calculated standard error.

Acknowledgements

This work has been supported in part by the NLM training grant 5T15LM00709 and research grant R01 LM 011155.