**Protein Set Analysis Report**

Hao-Bo Guo & Hong Qin, 02-19-2019

**Methods**

A method for network-based association study (NetBAS) was developed to reveal the relationships between categorical or quantitative data and interested gene or protein sets1. Particular categorical data for proteins include the gene ontology (GO) annotations2 and pathways3. Using the protein-protein interaction (PPI) network, GO/pathway enrichment analysis performed by NetBAS can predict the biological significance of a protein set based on the overrepresentation of functional GO/pathway terms involved in the proteins that directly interact with the protein set. Here we define a functional neighbor network of the protein set as the subnetwork formed by all proteins that carry a certain functional GO or pathway term and are first neighbors that form direct physical interactions with the protein set. Z-scores using the size or edge number of the neighbor network from known PPI network compared with null permutations are estimated to quantify the enrichment degrees of various GO/pathway terms. NetBAS is effective for a small protein set and can even be applied in GO annotations of a single protein.

The gene list studied in present work has five human (*H. sapiens*) genes: GBK1, RAB1B, ARFGAP1, USO1 and JAGN1. The InWeb\_IM4 PPI is selected, which comprises 592,685 interactions (edges) among 16,641 proteins (vertices). The modular enrichment tool, DAVID5, and two other PPI-network-based tools, WebGestalt6 and pathfindR7 have been used for comparisons. The gene ontology (GO) for all genes are obtained from the GO knowledgebase2,8 and the pathway data are downloaded from the ConcensusPathDB9,10.

**Results**



**Figure 1**. (**A**) Top ten enriched biological process (BP) GO terms by NetBAS (left, z-scores labeled) and the corresponding enrich scores (p-values) by DAVID (right). The functional neighbor network of the protein set for the top enriched BP (**B**), CC (**C**), MF (**D**) GO terms and KEGG (**E**), BioCarta (**F**) and Reactome (**G**) pathways. Each dot represents a protein in the PPI network: red dots are proteins that belong to the protein set but are not involved in the functional GO/KEGG term, green dots are proteins that belong to the protein set and carry the functional GO/KEGG term, and blue dots are proteins that are not from the protein set but carry the functional GO/KEGG term.

Because the protein set is relatively small (5 proteins), the network-based tool WebGestalt did not recognize any enriched GO terms for this set. DAVID showed 36 BP (biological), 10 CC (cellular component) and 0 MF (molecular function) terms in the enriched list. Nevertheless, the NetBAS method gives 19 highly enriched GO terms with z-scores > 10. Fig.1 gives the top 10 enriched BP terms by NetBAS compared with DAVID. For the BP terms, both NetBAS and DAVID yield the same top enriched term, GO:0006888 (Fig. 1A and 1B), which refers to “ER to Golgi vesicle-mediated transport”. Fig. 1B shows that the functional neighbor network of the top-enriched BP term comprises 293 edges (interactions) compared to an average of 20.7 in 1000 null models, indicating that the biological function described by this GO term is highly enriched in the functional neighbor network. The top-enriched CC and MF term by NetBAS, however, have not been recognized by DAVID (in fact, DAVID reveals none enriched MF terms). Here we showed that based on the PPI network, the CC term GO:0033116 (Fig. 1C, 106 edges) and MF term GO:0017112 (Fig. 1D, 28 edges) are considerably enriched in the corresponding functional neighbor networks, compared to those from the null models (average 5.5 for the top-CC and 2.2 for top-MF terms averaged from 1000 null models, respectively).

The PPI-based pathway enrichment tool pathfindR revealed none pathways associated with the gene list, and DAVID only recognized 1 KEGG, 1 BioCarta and 3 Reactome pathways, respectively. Here, the top-enriched KEGG term by DAVID (hsa04144: Endocytosis) is also listed in NetBAS with a rank of 3. The top enriched KEGG path by NetBAS, which is missed by DAVID, is hsa04130 (Fig. 1E). Both NetBAS and DAVID give the same top enriched BioCarta term (Fig. 1F). However, the top enriched Reactome term (R-HSA-199977: ER to Golgi Anterograde Transport) by NetBAS is missed by DAVID (Fig. 1G), which gives a dense neighbor network with 296 edges compared to 18.5 from null models, and leads to a large z-score of 67.211. The other Reactome term (R-HSA-948021: Transport to the Golgi and subsequent modification") is also highly enriched by NetBAS but missed by DAVID, and has a z-score of 66.726. The top enriched Reactome term by DAVID (R-HSA-6807878: COPI-mediated anterograde transport) is ranked 3 by NetBAS (z-score = 58.94). The top 10 enriched BP, CC and MF GO terms as well as KEGG, BioCarta and Reactome pathway terms have been summarized in a separate spread sheet. The configuration figures of the functional neighbor network of the protein set associated with these top 10 GO/pathway terms are also plotted.

*[Note and the spread sheet (UAB.Sztul.top10.enriched.terms.xlsx)]:*

GO.ID/ID: the ID for GO or pathway terms; GO.term/pathway: details of the GO/pathway terms;

Zscore: the enrichment score calculated using z=(obs-mean)/sd, where obs is the size (total edge number) of the observed functional neighbor network from the original PPI network, mean is the average size of the neighbor networks from 1000 null models of the PPI network and sd is the standard deviation.

**References**

1 Hao-Bo Guo, Hong Qin, Network-based association study of protein sets and applications in gene ontology enrichment analysis, in preparation.

2 The Gene Ontology, C. The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res*, in press, doi:10.1093/nar/gky1055 (2019).

3 Kanehisa, M., Sato, Y., Furumichi, M., Morishima, K. & Tanabe, M. New approach for understanding genome variations in KEGG. *Nucleic Acids Res* **47**, D590-D595, doi:10.1093/nar/gky962 (2019).

4 Li, T. *et al.* A scored human protein-protein interaction network to catalyze genomic interpretation. *Nat Methods* **14**, 61-64, doi:10.1038/nmeth.4083 (2017).

5 Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).

6 Wang, J., Vasaikar, S., Shi, Z., Greer, M. & Zhang, B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. *Nucleic Acids Res* **45**, W130-W137, doi:10.1093/nar/gkx356 (2017).

7 Ulgen, E., Ozisik, O. & Sezerman, O. U. pathfindR: An R Package for Pathway Enrichment Analysis Utilizing Active Subnetworks. *bioRxiv*, 272450 (2018).

8 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).

9 Kamburov, A., Wierling, C., Lehrach, H. & Herwig, R. ConsensusPathDB--a database for integrating human functional interaction networks. *Nucleic Acids Res* **37**, D623-628, doi:10.1093/nar/gkn698 (2009).

10 Herwig, R., Hardt, C., Lienhard, M. & Kamburov, A. Analyzing and interpreting genome data at the network level with ConsensusPathDB. *Nat Protoc* **11**, 1889-1907, doi:10.1038/nprot.2016.117 (2016).