An introduction to effective use of enrichment analysis software

Hannah Tipney[*] *and Lawrence Hunter*

Center for Computational Pharmacology, University of Colorado Denver, Aurora, CO 80045, USA
[*]*Correspondence to*: Tel: +1 303 724 3369; E-mail: hannah.tipney@ucdenver.edu

Date received (in revised form): 28th January, 2010

Abstract

In recent years, there has been an explosion in the range of software available for annotation enrichment analysis. Three classes of enrichment algorithms and their associated software implementations are introduced here. Their limitations and caveats are discussed, and direction for tool selection is given.

Keywords: *software, enrichment, gene enrichment analysis, GSEA, pathway analysis, gene ontology, gene list*

What is enrichment analysis and why is it useful?

The final stage of many proteomic, genetic or metabolic analyses is the production of a list of 'interesting' biomolecules. Prominent examples of these include lists of genes ranked by differential or co-expression investigated in microarray experiments, lists of single nucleotide polymorphism (SNP)-containing genes ranked by *p*-values determined by genetic association to a phenotype of interest through a genome-wide association study, and computationally generated lists of putative transcription factor or miRNA targets ordered by probability. Unfortunately, such ranked lists tend to be devoid of structure and lacking in context. It is difficult to determine how, or even if, the genes and their protein products interact with each other or influence the biological processes under study, or even what their 'normal' behaviour might be, by just reviewing them. Extensive exploration of literature and databases is required to answer even rudimentary questions such as: 'What does this gene and its protein product do? How and where does it do it? Does it make sense to see it on this list? Does it interact with other genes/proteins? Does its behaviour change during disease, disorder or therapy?' Manual gene-by-gene searches, especially across large lists of genes, are overwhelming and frequently unachievable tasks. Equally, ranked lists of genes do little to replicate the intricate reality of biology, where genes and proteins work together in complex interacting groups to create functioning systems. Focusing on a collection of interesting genes or proteins as a whole is not only more biologically intuitive, but also tends to increase statistical power and reduce dimensionality. Understanding the functional significance of such lists of genes, although overwhelming, is therefore a critical task.

Annotation enrichment (sometimes called pathway analysis[1]) has become the go-to secondary analysis undertaken on collections of genes identified by high-throughput genomic methods owing to its ability to provide valuable insight into the collective biological function underlying a list of genes. By systematically mapping genes and proteins to their associated biological annotations (such as gene ontology [GO] terms[2] or pathway membership) and then comparing the distribution of the terms within a gene set of interest with the background distribution of these terms (eg all genes represented on a microarray chip), enrichment analysis can identify terms which are statistically over- or under-represented within the list of interest.[3] It is inferred that such enriched terms describe some important underlying biological process or behaviour. For example, if 10 per cent of the genes on the 'interesting' list are kinases, compared with 1 per cent of the genes in the human genome (the population background), by using common statistical methods (eg χ^2, Fisher's exact test, binomial probability or hypergeometric distribution), it is possible to determine that kinases are enriched in the gene list and therefore have important functions in the biological study undertaken.[3]

Three classes of enrichment algorithms — what does each one do?

The field of enrichment research has exploded, growing from 14 tools in 2005[1,4] to 68 cited in a recent survey.[5] This field is still very much under active development, however, with no one 'perfect' method or gold standard protocol guaranteed to give the best results. For this reason, it is useful to understand the current state of the art, the caveats and pitfalls associated with certain analyses and how to identify software tools best suited to a particular dataset.

Owing to the large number of available enrichment tools, it is helpful to use the nomenclature of Huang *et al.*[5] when discussing enrichment software. Huang *et al.* classify enrichment tools as belonging to at least one of three algorithmic categories: singular enrichment analysis (SEA), gene set enrichment analysis (GSEA) and modular enrichment analysis (MEA).

The most traditional enrichment approach, SEA, iteratively tests annotation terms one at a time against a list of interesting genes for enrichment. An enrichment *p*-value is calculated by comparing the observed frequency of an annotation term with the frequency expected by chance; individual terms beyond some cut-off (eg *p*-value ≤ 0.05) are deemed enriched.[5] This is a simple, useful and easy-to-use protocol. Tools belonging to this category (eg Onto-Express,[6] FuncAssociate 2.0,[7] GOStat,[8] BiNGO[9] and EasyGO[10]) predominantly rely on the GO[4,11] as a source of annotation terms. As SEA considers each term independently, however, it ignores the hierarchical relationships between GO terms.[11] This frequently results in output lists of enriched terms numbering in the hundreds because similar terms are treated as though they were unique, leading to redundancy. Semantic redundancy between terms can also dilute an enriched biological concept due to difficulties in identifying enrichment between different, yet semantically similar, terms.[4,5] A drawback to any method relying on a single knowledge or annotation source is that it will also inherit limitations of that source. In the case of the GO, although it currently contains 29,365 terms,[12] it is a work in progress[13] and its annotations remain incomplete and biased towards well-studied genes.[14]

GSEA-based methods, such as GSEA/P-GSEA[15,16] and GeneTrail,[17] are similar in character to SEA, but they consider *all* genes during analysis, not just those deemed as interesting or significant by some metric or threshold. GSEA methods work best in scenarios in which phenotypic classes or time points are assayed (eg tumour versus normal tissue, or treated versus untreated state) because the method requires a quantitative biological value (such as fold change or degree of differential expression) for each gene in order to rank them. A maximum enrichment score (MES) is calculated from the ranked list of all genes in a given annotation category and an enrichment *p*-value determined by comparing the ranked annotation MES to randomly generated MES distributions.[5,16] Simply, GSEA determines if those genes sharing a particular annotation (eg a biochemical pathway), known as a gene set, are randomly distributed throughout the larger ranked gene list and therefore not significantly associated with any phenotypic class, or if they tend to be over-represented towards the top or bottom of the longer ranked gene list, indicating an association between the gene set (ie genes sharing the annotation of interest) and the phenotypic classes under study.

Although many different annotation categories can be used by GSEA methods, including biological function (eg GO terms), physical position (eg chromosomal location), regulation (eg co-expression) or any other attribute for which prior knowledge is available, like SEA methods they are still considered one at a time and treated independently, with no consideration given to the semantic relationships which may exist between the different annotation terms. At times, it may be difficult to assign a single value to a gene; for example, multiple SNPs within a single gene may have differing *p*-values, or comparisons may have been made across many time points or conditions. In such instances, GSEA-based methods may be inappropriate. It should be noted, however, that recent modifications to GSEA methods to cope with genome-wide association study-derived datasets have been proposed,[18–21] and a novel GSEA method using mixed-effects models has successfully identified enriched GO terms in a time-course microarray dataset.[22,23] In addition, highly ranked genes (ie those with larger fold changes) contribute greatly to the enrichment *p*-value, the underlying assumption being that genes with greater deregulation (ie fold changes) contribute more to the observed phenotype. This assumption does not always hold true in real biology, however.[5]

The final algorithmic class defined by Huang *et al.*, MEA, is the only class to use the relationships that may exist between different annotation terms during enrichment. As mentioned previously, doing so can reduce redundancy and prevent the dilution of potentially important biological concepts. A number of tools (eg Ontologizer,[24] topGO[25] and GeneCodis[26]) claim to have improved sensitivity and specificity by considering relationships between GO annotation terms in this manner. The use of composite annotation terms may therefore be able to provide biological insight otherwise lacking in analyses treating single terms as independent entities.[5] These tools, however, still focus on a single annotation source — in this instance, the GO. Many additional types of information or attributes can be used during enrichment analysis and by incorporating a range of annotation types in concert, analysis can be more effective as increased coverage increases analytical power and can provide a more complete view of the biology underlying a gene set of interest.

Functional enrichment tools such as DAVID[3,27] and the recently released ConceptGen[28] do exactly that, not only considering relationships between annotation terms (both within an annotation source and between different sources), but also integrating annotation terms from a range of sources, including those representing protein–protein interactions, protein functional domains (eg InterPro), disease associations (eg OMIM), pathways (eg KEGG, BioCarta), sequence features, homology, expression patterns (eg GEO) and literature. By grouping similar, redundant and homogeneous annotation content from the same or different resources into annotation groups, the burden of associating similar and redundant terms is reduced, and the biological interpretation of gene sets moves from a gene-centric to a biological-module-centric approach, which may provide a better representation of a biological process.[4,5] These tools have also invested in novel visualisation methods to support effective exploration of results.[28] One consideration when using this seemingly comprehensive analysis protocol is that ‘orphan’ genes (ie terms without strong relationships to other terms) may be overlooked, requiring such genes to be investigated manually through other methods.[5]

Understanding and overcoming limiting factors aids effective analysis

Regardless of which specific tool or algorithmic class is used in an annotation enrichment procedure, a number of potentially limiting factors should be considered. First, the quality of any enrichment result is highly dependent on the quality of input. For SEA and MEA methods, this is the gene list defined as being interesting. For GSEA, it is the pre-computed annotated gene sets. In both instances, the possibility for bias and error needs to be guarded against. For example, a predefined gene set based on a curated pathway (eg KEGG) is likely to be incomplete. Identifying annotations that apply to all genes in a genome or on a microarray, and that are also appropriate, is also difficult. For example, chromosomal location is an annotation common to all genes, but this is unlikely to be an appropriate, and informative annotation for most analyses (barring cancer). Tools such as WhichGenes[29] and ConceptGen[28] can be used to aid gene set identification. In addition, for enrichment to be successful, it must be possible to map any gene or protein identifier used as input to the corresponding annotation source used by a particular enrichment method. In many instances, the NCBI EntrezGene database identifiers are recognised and recommended for use. It is advisable to check each tool's preferences,[5] however, due to the large impact that identifier selection can have on results. For example, annotation content can be significantly enriched using the DAVID Gene Concept procedure to re-agglomerate gene identifiers. This procedure can increase the number of GO terms assigned to genes by up to 20 per cent compared with annotations in each individual source.[30]

Secondly, the selection of an inappropriate background set can also heavily influence an enrichment protocol, resulting in concepts and genes appearing to be more significant than they actually are, or appearing significant (ie biased) when the bias is actually due to methodology rather than biology. It is imperative to think carefully about the 'world' from which an interesting subset of genes was taken. A good rule of thumb for background selection is only to include those genes or proteins that have a chance of making it into the 'interesting' set and exclude all others.[4,31] For example, during a microarray experiment the background set of genes should include only those genes for which corresponding probes are present on the chip. Without a corresponding probe, it is impossible for a gene to be identified as interesting, no matter what post-processing is undertaken.

Finally, assumptions of independence and random selection are frequently incorrect for biological systems; for example, many genes display dependent and correlated behaviour (discussed in detail by Tilford and Siemers[31]). Additionally, our current understanding, and therefore annotations, are also biased towards more extensively studied genes, proteins, pathways and disorders.[14] The mixed-effect model recently proposed by Wang *et al.* to identify annotation enrichment in time-course data, however, claims to model dependency accurately between genes.[22] As a footnote, it has been noted that — due to these limitations — enrichment analysis should be considered as an exploratory procedure rather than a definitive statistical solution, since the user is heavily involved in result assessment, in determining if results presented to them are of use, and in determining useful *p*-value cut-offs.[5]

Conclusions and additional resources

By focusing on sets of genes that share biologically important attributes, enrichment analysis can support the discovery of biological functions that may otherwise have been missed by moving the analysis of biological function from the level of single genes to that of biological processes (reviewed by Curtis *et al.*[1] and Khatri and Draghici[4]). Currently, no enrichment method has been identified as perfectly suitable under all analysis scenarios, and no gold-standard test set exists for effective comparative testing, making tool selection confusing.[5]

This unavoidably brief paper, and the citations and descriptions of tools within, aims to familiarise the user with the different enrichment resources available, the advantages and limitations of each and methodologies to ensure that analyses undertaken are comprehensive. A wealth of additional guidance is available for the interested reader. A detailed introduction to the field is provided by Tilford and Siemers,[31] while the statistical protocols underlying different approaches are comprehensively expanded upon and their performances compared by Ackermann and Strimmer.[32] For those requiring hands-on guidance, the protocols paper by Huang *et al.*[3] offers a tutorial-style introduction to enrichment using DAVID. An excellent review of currently available tools and remaining challenges in the field of enrichment is also provided by Huang *et al.*;[5] it includes a useful set of questions to help guide software selection depending on individual needs and experience, as does the paper of Tilford and Siemers.[31] It is also recommended to try a number of different tools on the same dataset to enable direct comparison by the user.[11]

Acknowledgments

The authors are supported by NIH R01 grants LM009254-04 (PI Hunter) and LM008111-05 (PI Hunter).

References

1. Curtis, R.K., Oresic, M. and Vidal-Puig, A. (2005), 'Pathways to the analysis of microarray data', *Trends Biotechnol*. Vol. 23, pp. 429–435.
2. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D. *et al*. (2000), 'Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium', *Nat. Genet*. Vol. 25, pp. 25–29.
3. Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009), 'Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources', *Nat. Protoc*. Vol. 4, pp. 44–57.
4. Khatri, P. and Draghici, S. (2005), 'Ontological analysis of gene expression data: Current tools, limitations, and open problems', *Bioinformatics* Vol. 21, pp. 3587–3595.
5. Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009), 'Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists', *Nucleic Acids Res*. Vol. 37, pp. 1–13.
6. Draghici, S., Khatri, P., Martins, R.P., Ostermeier, G.C. *et al*. (2003), 'Global functional profiling of gene expression', *Genomics* Vol. 81, pp. 98–104.
7. Berriz, G.F., Beaver, J.E., Cenik, C., Tasan, M. *et al*. (2009), 'Next generation software for functional trend analysis', *Bioinformatics* Vol. 25, pp. 3043–3044.
8. Beissbarth, T. and Speed, T.P. (2004), 'GOstat: Find statistically overrepresented gene ontologies within a group of genes', *Bioinformatics* Vol. 20, pp. 1464–1465.
9. Maere, S., Heymans, K. and Kuiper, M. (2005), 'BiNGO: A Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks', *Bioinformatics* Vol. 21, pp. 3448–3449.
10. Zhou, X. and Su, Z. (2007), 'EasyGO: Gene ontology-based annotation and functional enrichment analysis tool for agronomical species', *BMC Genomics* Vol. 8, p. 246.
11. Rhee, S.Y., Wood, V., Dolinski, K. and Draghici, S. (2008), 'Use and misuse of the gene ontology annotations', *Nat. Rev. Genet*. 9, pp. 509–515.
12. Gene Ontology website, http://www.geneontology.org [Accessed 19th January 2010].
13. Baumgartner, W.A., Jr, Cohen, K.B., Fox, L.M., Acquaah-Mensah, G. *et al*. (2007), 'Manual curation is not sufficient for annotation of genomic databases', *Bioinformatics* Vol. 23, pp. i41–i48.
14. Alterovitz, G., Xiang, M., Mohan, M. and Ramoni, M.F. (2007), 'GO PaD: The Gene Ontology Partition Database', *Nucleic Acids Res*. Vol. 35 (Database issue), pp. D322–D327.
15. Subramanian, A., Kuehn, H., Gould, J., Tamayo, P. *et al*. (2007), 'GSEA-P: A desktop application for gene set enrichment analysis', *Bioinformatics* Vol. 23, pp. 3251–3253.
16. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S. *et al*. (2005), 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', *Proc. Natl. Acad. Sci. USA* Vol. 102, pp. 15545–15550.
17. Backes, C., Keller, A., Kuentzer, J., Kneissl, B. *et al*. (2007), 'GeneTrail — Advanced gene set enrichment analysis', *Nucleic Acids Res*. Vol. 35 (Web Server issue), pp. W186–W192.
18. Hong, M.G., Pawitan, Y., Magnusson, P.K. and Prince, J.A. (2009), 'Strategies and issues in the detection of pathway enrichment in genome-wide association studies', *Hum. Genet*. Vol. 126, pp. 289–301.
19. Sohns, M., Rosenberger, A. and Bickeboller, H. (2009), 'Integration of *a priori* gene set information into genome-wide association studies', *BMC Proc*. Vol. 3 (Suppl. 7), p. S95.
20. Wang, K., Li, M. and Bucan, M. (2007), 'Pathway-based approaches for analysis of genomewide association studies', *Am. J. Hum. Genet*. Vol. 81(6), pp. 1278–1283.
21. Holden, M., Deng, S., Wojnowski, L. and Kulle, B. (2008), 'GSEA-SNP: Applying gene set enrichment analysis to SNP data from genome-wide association studies', *Bioinformatics* Vol. 24, pp. 2784–2785.
22. Wang, L., Chen, X., Wolfinger, R.D., Franklin, J.L. *et al*. (2009), 'A unified mixed effects model for gene set analysis of time course microarray experiments', *Stat. Appl. Genet. Mol. Biol*. Vol. 8(1): Article 47.
23. Wang, L., Zhang, B., Wolfinger, R.D. and Chen, X. (2008), 'An integrated approach for the analysis of biological pathways using mixed models', *PLoS Genet*. Vol. 4(7), p. e1000115.
24. Grossmann, S., Bauer, S., Robinson, P.N. and Vingron, M. (2007), 'Improved detection of overrepresentation of gene-ontology annotations with parent child analysis', *Bioinformatics* Vol. 23, pp. 3024–3031.
25. Alexa, A., Rahnenfuhrer, J. and Lengauer, T. (2006), 'Improved scoring of functional groups from gene expression data by decorrelating GO graph structure', *Bioinformatics* Vol. 22, pp. 1600–1607.
26. Nogales-Cadenas, R., Carmona-Saez, P., Vazquez, M., Vicente, C. *et al*. (2009), 'GeneCodis: Interpreting gene lists through enrichment analysis and integration of diverse biological information', *Nucleic Acids Res*. Vol. 37 (Web Server issue), pp. W317–322.
27. Dennis, G., Jr, Sherman, B.T., Hosack, D.A., Yang, J. *et al*. (2003), 'DAVID: Database for Annotation, Visualization, and Integrated Discovery', *Genome Biol*. Vol. 4, p. P3.
28. Sartor, M.A., Mahavisno, V., Keshamouni, V.G., Cavalcoli, J. *et al*. (2009), 'ConceptGen: A gene set enrichment and gene set relation mapping tool', *Bioinformatics* epub ahead of print, doi 1093/bioinformatics/btp683.
29. Glez-Pena, D., Gomez-Lopez, G., Pisano, D.G. and Fdez-Riverola, F. (2009), 'WhichGenes: A web-based tool for gathering, building, storing and exporting gene sets with application in gene set enrichment analysis'. *Nucleic Acids Res*. Vol. 37 (Web Server issue), pp. W329–W334.
30. Sherman, B.T., Huang da, W., Tan, Q., Guo, Y. *et al*. (2007), 'DAVID Knowledgebase: A gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis', *BMC Bioinformatics* Vol. 8, p. 426.
31. Tilford, C.A. and Siemers, N.O. (2009), 'Gene set enrichment analysis', *Methods Mol. Biol*. Vol. 563, pp. 99–121.
32. Ackermann, M. and Strimmer, K. (2009), 'A general modular framework for gene set enrichment analysis', *BMC Bioinformatics* Vol. 10, p. 47.