SURVEY AND SUMMARY

Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists

Da Wei Huang, Brad T. Sherman and Richard A. Lempicki*

Laboratory of Immunopathogenesis and Bioinformatics, Clinical Services Program, SAIC-Frederick, Inc., National Cancer Institute at Frederick, Frederick, MD 21702, USA

ABSTRACT

Functional analysis of large gene lists, derived in most cases from emerging high-throughput genomic, proteomic and bioinformatics scanning approaches, is still a challenging and daunting task. The gene-annotation enrichment analysis is a promising high-throughput strategy that increases the likelihood for investigators to identify biological processes most pertinent to their study. Approximately 68 bioinformatics enrichment tools that are currently available in the community are collected in this survey. Tools are uniquely categorized into three major classes, according to their underlying enrichment algorithms. The comprehensive collections, unique tool classifications and associated questions/issues will provide a more comprehensive and up-to-date view regarding the advantages, pitfalls and recent trends in a simpler tool-class level rather than by a tool-by-tool approach. Thus, the survey will help tool designers/developers and experienced end users understand the underlying algorithms and pertinent details of particular tool categories/tools, enabling them to make the best choices for their particular research interests.

INTRODUCTION

The traditional biological research approaches typically study one gene or a few genes at a time. In contrast, high-throughput genomic, proteomic and bioinformatics scanning approaches (such as expression microarray, promoter microarray, proteomics, ChIP-on-CHIPs, etc.) are emerging as alternative technologies that allow investigators to simultaneously measure the changes and regulation of genome-wide genes under certain biological conditions. Those high-throughput technologies usually generate large 'interesting' gene lists as their final outputs. However, the biological interpretation of large, 'interesting' gene lists (ranging in size from hundreds to thousands of genes) is still a challenging and daunting task. Over the last few decades, bioinformatics methods, using the biological knowledge accumulated in public databases [e.g. Gene Ontology (1)], make it possible to systematically dissect large gene lists in an attempt to assemble a summary of the most enriched and pertinent biology. A number of high-throughput enrichment tools, including, but not limited to Onto-Express, MAPPFinder, GoMiner, DAVID, EASE, GeneMerge and FuncAssociate, etc. (2–10), were independently developed during 2002 and 2003 as initial studies to address the challenge of functionally analyzing large gene lists. Since then, the enrichment analysis field has been very productive, resulting in more, similar tools becoming publicly available. In 2005, approximately 14 such tools were collected and reviewed by Khatri *et al.* (11) and by Curtis *et al.* (12), respectively. The activity in the field has continually grown stronger as the number of new enrichment tools (with distinct new ideas and features) has significantly increased. Approximately 68 such tools have been collected in this survey (2–10,13–73) (Table 1 and Supplementary Data 1).

During the past several years, bioinformatics enrichment tools have played a very important and successful role contributing to the gene functional analysis of large gene lists for various high-throughput biological studies, which is clearly evidenced by thousands of publications citing these tools (based on Google Scholar as of September 2008). However, these bioinformatics

*To whom correspondence should be addressed. Tel: +1 301 846 5093; Fax: +1 301 846 6762; Email: rlempicki@mail.nih.gov

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

enrichment tools are still in an actively growing and improving stage, without unified methods or one 'gold' standard. As more enrichment tools emerge in the scientific community, the individual tool-developing group or end user finds it more and more difficult to comprehensively track the usefulness of all of the existing works to his or her research. This confusing plethora of tools has resulted in several issues: (i) difficulty in comprehensively comparing and remembering the algorithms/features in a tool-by-tool manner among the overwhelmingly large number of tools available (approximately 68 current tools); (ii) a chance that some good work may be overlooked; (iii) redundant efforts in developing ideas that already exist, because of developers' difficulties in grasping the breadth of the field; (iv) out-of-date ideas being used in newly released tools because of the developers' lack of awareness of the latest methods; and (v) difficulties for end users in deciding, among so many overwhelming choices, which enrichment tools are most suitable to their analytic needs.

This survey includes four sections to address the situations listed earlier: First, it will identify 68 enrichment tools that are currently available, and further describe the rationales behind them. That way, the tool designers, developers and end users will be made aware of most, if not all, of the existing tools. Secondly, tools will be uniquely classified, according to their underlying algorithms, into three major categories. Thus, readers can more easily and quickly grasp the key spirit of the 68 tools by following the categorical logic instead of trying to search through a tool-by-tool layout. Thirdly, the paper will focus on several important, but largely unanswered, questions and issues associated with the field. We hope that the questions/issues to be discussed will drive more attention, independent thinking, and discussion in the field, thereafter leading to better solutions in the near future. Finally, the paper will conclude with the current status and trends in the field.

GENERAL PRINCIPLE OF ENRICHMENT ANALYSIS AND 68 AVAILABLE TOOLS

A biological process is typically made up of a group of genes, as opposed to an individual gene alone. The principal foundation of enrichment analysis is that if a biological process is abnormal in a given study, the co-functioning genes should have a higher (enriched) potential to be selected as a relevant group by the high-throughput screening technologies. Such a rationale can make the analysis of large gene lists move from an individual gene-oriented view to a relevant gene group-based analysis. Because the analytic conclusion is based on a group of relevant genes instead of on an individual gene, it increases the likelihood for investigators to identify the correct biological processes most pertinent to the biological phenomena under study. For example, 10% of the user's genes selected by a microarray experiment are kinases, as opposed to 1% of the genes in the human genome (this is the gene population background) that are kinases. The enrichment can therefore be quantitatively measured by some common and well-known statistical methods, including Chi-square, Fisher's exact test, Binomial probability and Hypergeometric distribution (more discussion of enrichment *P*-value in a later section of this paper). Thus, a conclusion may be obtained for the particular example, that is, kinases are enriched in the user's study and therefore play an important role in the study. Fortunately, annotation databases, such as Gene Ontology (GO) (1), collecting biological knowledge in a format of gene-to-annotation, are very suitable for high-throughput bioinformatics scanning for the enrichment analysis. The tools systematically map a large number of interesting genes in a list to the associated biological annotation terms (e.g. GO Terms or Pathways), and then statistically examine the enrichment of gene members for each of the annotation terms by comparing the outcome to the control (or reference) background. Thereafter, the annotation terms with enriched gene members can be identified from tens of thousands of other annotation terms in a high-throughput fashion (11,12). The enriched annotation terms associated with the large gene list will give important insights that allow investigators to understand the biological themes behind the large gene list.

Approximately 68 bioinformatics tools (Table 1 and Supplementary Data 1) (2–10,13–73), aligned with the above analytic scenarios and purposes, are collected in this study. Regardless of their distinct features, the general procedure of the tools can be described as having three major layers: data support (backend annotation database); data mining (algorithm and statistics); and result presentation (interface and exploration) (Figure 1). Each of the layers may greatly impact the comprehensiveness of analytic results, as discussed in later sections of this paper. The general features associated with each tool, such as tool home page, publication link, general database scope [see SerbGO (74), which searches detailed annotation coverage across tools], pathway presentation, etc., can be found in Supplementary Data 1, in order to help end users/developers look up tools for their research interests. Moreover, the capability, sensitivity and backend databases can be very different from tool to tool. It is not uncommon for users to try multiple tools with similar analytic capability for the same dataset in order to obtain maximum satisfactory analytic results (75).

CLASSIFICATION OF ENRICHMENT TOOLS

When the tool developer or end user is searching for particular features among the many tools available, it is not an easy task to digest the features for all 68 tools without appropriate classification. Based on the difference of algorithms, this survey classifies the 68 current enrichment tools into three classes: singular enrichment analysis (SEA); gene set enrichment analysis (GSEA); and modular enrichment analysis (MEA). A complete list of tools and their defining classes can be found in Table 1 and Supplementary Data 1. Notably, some tools with diverse capabilities belong to more than one class. The general features and limitations associated with each class are discussed in the following sections and are compared in Table 2.

Table 1. List of 68 enrichment tools

| Enrichment tool name	Year of release	Key statistical method	Category
FunSpec	2002	Hypergeometric	Class I
Onto-express	2002	Fisher's exact; hypergemetic; binomial; chi-square	Class I
EASE	2003	Fisher's exact (modified as EASE score)	Class I
FatiGO/FatiWise/FatiGO+	2003	Fisher's exact	Class I
FuncAssociate	2003	Fisher's exact	Class I
GARBAN	2003	Hypergeometric	Class I
GeneMerge	2003	Hypergeometric	Class I
GoMiner	2003	Fisher's exact	Class I
MAPPFinder	2003	Z-score; hypergeometric	Class I
CLENCH	2004	Hypergeometric; chi-square; binomial	Class I
GO::TermFinder	2004	hypergeometric	Class I
GOAL	2004	Permutation	Class I
GOArray	2004	Hypergeometric; Z-score; permutation	Class I
GOStat	2004	Fisher's exact; chi-squre	Class I
GoSurfer	2004	Chi-square	Class I
OntologyTraverser	2004	Hypergeometric; Fisher's exact	Class I
THEA	2004	Hypergeometric	Class I
BiNGO	2005	Hypergeometric; binomial	Class I
FACT	2005	Adopt GeneMerge and GO::TermFinder statistical modules	Class I
gfinder	2005	Fisher's exact	Class I
Gobar	2005	Hypergeometric	Class I
GOCluster	2005	Hypergeometric	Class I
GOSSIP	2005	Fisher's exact	Class I
L2L	2005	Binomial; hypergeometric	Class I
WebGestalt	2005	Hypergeometric	Class I
BayGO	2006	Bayesian; Goodman and Kruskal's gamma factor	Class I
eGOn/GeneTools	2006	Fisher's exact	Class I
Gene Class Expression	2006	Z-statistics	Class I
GOALIE	2006	Hidden Kripke model	Class I
GOFFA	2006	Fisher's inverse chi-square	Class I
GOLEM	2006	Hyerpgeometric	Class I
JProGO	2006	Fisher's exact; Kolmogorov–Smirnov test; student's t-test; Wilcoxon's test; hypergeometric	Class I
PageMan	2006	Fisher's exact; chi-square; Wilcoxon	Class I
STEM	2006	Hypergeometric	Class I
WEGO	2006	Chi-square	Class I
EasyGO	2007	Hypergeometric; chi-square; binomial	Class I
g:Profiler	2007	Hypergeometric	Class I
ProbCD	2007	Yule's Q; Goodman-Kruskal's gamma; Cramer's T	Class I
GOEAST	2008	Hypergeometric	Class I
GOHyperGAll	2008	Hypergeometric	Class I
CatMap	2004	Permutations	Class II
Godist	2004	Kolmogorov–Smirnov test	Class II
GO-Mapper	2004	Gaussian distribution; EQ-score	Class II
iGA	2004	Permutations; hypergeometric; t-test; Z-score	Class II
GSEA	2005	Kolmogorov–Smirnov-like statistic	Class II
MEGO	2005	Z-score	Class II
PAGE	2005	Z-score	Class II
T-profiler	2005	t-Test	Class II
FuncCluster	2006	Fisher's exact	Class II
FatiScan	2007	Fisher's Exact	Class II
FINA	2007	Fisher's exact	Class II
GAzer	2007	Z-statistics; permutation	Class II
GeneTrail	2007	Hypergeometric; Kolmogorov–Smirnov	Class II
MetaGP	2007	Z-score	Class II
Ontologizer	2004	Fisher's exact	Class III
POSOC	2004	POSET (a discrete math: finite partially ordered set)	Class III
topGO	2006	Fisher's exact	Class III
GO-2D	2007	Hypergeometric; binomial	Class III
GENECODIS	2007	Hypergeometric; chi-square	Class III
GOSim	2007	Resnik's similarity	Class III
PalS	2008	Percent	Class III
ProfCom	2008	Greedy heuristics	Class III
GOTM	2004	Hypergeometric	Class I,II
ermineJ	2005	Permutations; Wilcoxon rank-sum test	Class I,II
DAVID	2003	Fisher's Exact (modified as EASE score)	Class I,III
GOToolBox	2004	Hypergeometric; Fisher's exact; Binomial	Class I,III
ADGO	2006	Z-statistic	Class II,III
FunNet	2008	Unclear	Unclear
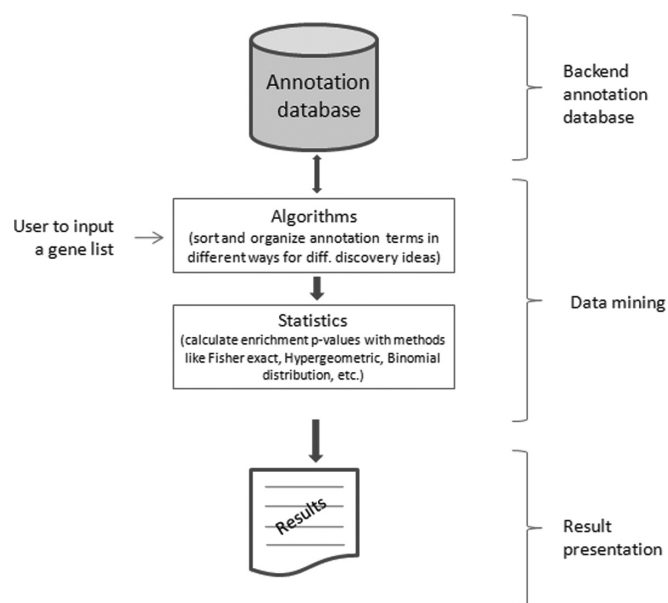
Figure 1. The infrastructure of typical enrichment tools. Even though the enrichment analysis tools have distinct features, they can be generally described as three major layers: backend annotation database; data mining; and result presentation. Each of the layers, rather than statistical methods alone, greatly influences the analytic results.

Class 1: Singular enrichment analysis (SEA)

The most traditional strategy for enrichment analysis is to take the user's preselected (e.g. differentially expressed genes selected between experimental versus control samples by *t*-test with a *P*-value ≤ 0.05 and fold change ≥ 1.5) 'interesting' genes, and then iteratively test the enrichment of each annotation term one-by-one in a linear mode. Thereafter, the individual, enriched annotation terms passing the enrichment *P*-value threshold are reported in a tabular format ordered by the enrichment probability (enrichment *P*-value). The enrichment *P*-value calculation, i.e. number of genes in the list that hit a given biology class as compared to pure random chance, can be performed with the aid of some common and well-known statistical methods (11,12,76), including Chi-square, Fisher's exact test, Binomial probability and Hypergeometric distribution, etc. (Table 1). More discussion regarding the enrichment *P*-value can be found in a later section of this paper.

Even though the strategy and output format of SEA are simple, SEA is indeed a very efficient way to extract the major biological meaning behind large gene lists, which may be generated from any type of high-throughput genomic studies or bioinformatics software packages. Most of the earlier tools (such as GoMiner, Onto-Express, DAVID and EASE) and a lot of the recently released tools (such as GOEAST and GFinder), adopted this strategy and demonstrated significant success in many genomic studies. However, the common weakness of tools in this class is that the linear output of terms can be very large and overwhelming (from hundreds to thousands). Therefore, the data analyst's focus and interrelationships of relevant terms can be diluted. For example, relevant GO terms like apoptosis, programmed cell death, induction of apoptosis, anti-apoptosis, regulation of apoptosis, etc., are spread out at different positions in a large linear output. It is difficult to focus on interrelationships of relevant biology terms among hundreds or thousands of other terms. In addition, the quality of pre-selected gene lists could largely impact the enrichment analysis, which makes SEA analysis unstable to a certain degree when using different statistical methods or cutoff thresholds.

Class 2: Gene set enrichment analysis (GSEA)

GSEA carries the core spirit of SEA, but with a distinct algorithm to calculate enrichment *P*-values as compared to SEA (35). People in the field give great attention and expectation to the GSEA strategy. The unique idea of GSEA is its 'no-cutoff' strategy that takes all genes from a microarray experiment without selecting significant genes (e.g. genes with *P*-value ≤ 0.05 and fold change ≥ 1.5). This strategy benefits the enrichment analysis in two aspects: 1) it reduces the arbitrary factors in the typical gene selection step that could impact the traditional enrichment analysis; and 2) it uses all information obtained from microarray experiments by allowing the minimally changing genes, which cannot pass the selection threshold, to contribute to the enrichment analysis in differing degrees. The maximum enrichment score (MES) is calculated from the rank order of all gene members in the annotation category. Thereafter, enrichment *P*-values can be obtained by matching the MES to randomly shuffled MES distributions (a Kolmogorov–Smirnov-like statistic) (35). Other enrichment tools in the GSEA class using the 'no-cutoff' strategy, such as ErmineJ (31), FatiScan (55), MEGO (36), PAGE (29), MetaGF, Go-Mapper (22) and ADGO (45), etc., employ parametric statistical approaches such as *z*-score, *t*-test, permutation analysis, etc. These approaches directly take experimental values (e.g. fold change) of all genes into the calculation for each annotation term. Collectively, recent GSEA tools which integrate the total experimental values into the functional data mining are an interesting trend with a lot of potential as a complement to traditional SEA (47,77–79).

However, tools in the GSEA class are also associated with some common limitations. First, the 'no-cutoff' strategy is the key advantage of GSEA, but is also becoming its major limitation in many biological studies. The GSEA method requires a summarized biological value (e.g. fold change) for each of the genome-wide genes as input. Sometimes, it is a difficult task to summarize many biological aspects of a gene into one meaningful value when the biological study and genomic platform are complex. For example, each gene derived from a SNP microarray could associate with a set of SNPs, which vary in size, *P*-values, physical distances, disease regions, LD (Linkage Disequilibrium) strength and SNP-gene locations (e.g. in exon, or in intron) from gene to gene. It is still a very experimental procedure to summarize such diverse aspects of biology into one comprehensive value. Similar challenges may be found in many of the emerging genomic platforms (e.g. SNP, Exon, Promoter microarray). The situations in the examples fully or partially fail in the GSEA-required input data structure requirement.

Table 2. Categorization of enrichment analysis tools

Tool category	Description	Indication and limitation	Sub-type of algorithms	Methods	Example tool
Class I: singular enrichment analysis (SEA)	Enrichment *P*-value is calculated on each term from the pre-selected interesting gene list. Then, enriched terms are listed in a simple linear text format. This strategy is the most traditional algorithm. It is still dominantly used by most of the enrichment analysis tools.	Capable of analyzing any gene list, which could be selected from any high-throughput biological studies/technologies (e.g. Microarray, ChIP-on-CHIP, ChIP-on-seqence, SNP array, EXON array, large scale sequence, etc.). However, the deeper inter-relationships among the terms may not be fully captured in linear format report.	Global reference background	Fisher's exact hypergeometric chi-square binomial	GoStat, GoMiner, GOTM, BinGO, GOtoolBox, GFinder, etc.
			Local reference background	Fisher's Exact hypergeometric chi-square binomial	DAVID, Onto-Express, GARBAN, FatiGO, etc.
			Neural network	Bayesian	BayGO
Class II: gene set enrichment analysis (GSEA)	Entire genes (without pre-selection) and associated experimental values are considered in the enrichment analysis. The unique features of this strategy are: (i) No need to pre-select interesting genes, as opposed to Classes I and II; (ii) Experimental values integrated into *P*-value calculation.	Suitable for pair-wide biological studies (e.g. disease versus control). Currently, may be difficult to be applied to the diverse data structures derived by a complex experimental design and some of the new technologies (e.g. SNP, EXON, Promoter arrays).	Based on ranked gene list	Kolmogorov–Smirnov-like	GSEA, CapMap, etc.
			Based on continuous gene values	*t*-Test permutation Z-score	FatiScan, ADGO, ermineJ, PAGE, iGA, GO-Mapper, GOdist, FINA, T-profiler, MetaGP, etc.
Class III: modular enrichment analysis (MEA)	This strategy inherits key spirit of SEA. However, the term–term/gene–gene relationships are considered into enrichment *P*-value calculation. The advantage of this strategy is that term–term/gene–gene relationship might contain unique biological meaning that is not held by a single term or gene. Such network/modular analysis is closer to the nature of biological data structure.	Capable of analyzing any gene lists, which could be selected from any high-throughput biological studies/technologies, like Class I. Emphasis on network relation-ships during analysis. 'Orphan' gene/term (with little relationships to other genes/terms), that some-times could be very interesting, too, may be left out from the analysis.	Composite annotations	Measure enrichment on joint terms	ADGO, GeneCodis, ProfCom, etc.
			DAG Structure	Measure enrichment by considering parents-child relationships	topGO, Ontologizer, POSOC, etc.
			Global annotation relationship	Measure term–term global similarity with Kappa Statistics Czekanowski-Dice Pearson's correlation	DAVID, GoToolBox, etc.

For another example, many clinical microarray studies involve multiple factors/variants simultaneously, such as disease/normal, ages, sex, drug treatment/control, reagent batch effects, animal batch effect, etc. In such complex situations, sophisticated statistical methods, like ANOVA, time series analysis, survival analysis, etc., will be more powerful to handle multi-variances, multiple time points and batch effects, etc. simultaneously for data-mining interesting gene lists. In many similar cases, the upstream data processing and comprehensive gene selection statistics cannot be simply avoided or replaced by GSEA. Moreover, the genes ranked in higher positions (usually with higher differences, e.g. fold change) are the major force driving (highly weighted) the enrichment P-values in GSEA. Thus, the underlying assumption is that the genes with large regulations (e.g. fold changes) are contributing more to the biology. Obviously, this is not always true in real biology. Biologists know that small changes of some signal transduction genes can result in larger downstream biological consequences. In contrast, some big changes in metabolic genes may be just a consequence of other small, but important, signal regulation events. Depending on the questions that the researcher is asking, the mildly changed signal transduction genes may be more interesting/important than those largely regulated genes.

The GSEA and SEA methods have been available in the community for many years. Surprisingly, no comprehensive and systematic side-by-side comparisons are available yet. A recent study ran the same datasets with DAVID methods (a SEA/MEA method) versus ErmineJ (a GSEA method) (60). As expected, the results from both methods were highly consistent with each other. The consistency makes sense because the major driving force of the enrichment calculation in GSEA is the largely changing genes. In addition, those genes most likely have better chances to be selected in the traditional gene selection procedures, thus resulting in very similar results between the SEA and GSEA methods.

Class 3: Modular enrichment analysis (MEA)

MEA inherits the basic enrichment calculation found in SEA and incorporates extra network discovery algorithms by considering the term-to-term relationships. Recent tools, such as Ontologizer (69), topGO (41), GENECODIS (59), ADGO (45) and ProfCom (68), claimed to improve discovery sensitivity and specificity by considering inter-relationships of GO terms in the enrichment calculations, i.e. using genes of composite (joint) annotation terms as a reference background. The key advantage of this approach is that the researcher can take advantage of term–term relationships, in which joint terms may contain unique biological meaning for a given study, not held by individual terms. Moreover, when using heterogeneous annotation content, the annotation terms are highly redundant, and also have strong interrelationships regarding different aspects for the same biological process. Building such relationships is one step closer to the true nature of biology during data mining. GoToolBox (18) developed functions to cluster related GO terms or genes, which provides the gene functional annotation in a network context. However, the functions only work for a small scope and only for GO terms. DAVID (60,61) recently provided a new tool that is able to organize and condense a wide range of heterogeneous annotation content, such as GO terms, protein domains, pathways and so on, into term or gene classes. This organization is accomplished by using *Kappa* statistics to mine the complex biological co-occurrences found in multiple heterogeneous annotation content. Combined with traditional enrichment P-value calculations, the new approach allows the enrichment analysis to progress from term-centric or gene-centric to biological module-centric analysis. These methods take into account the redundant and networked nature of biological annotation content in order to concentrate on building the larger biological picture rather than focusing on an individual term or gene. Such data-mining logic seems closer to the nature of biology in that a biological process works in a network manner. However, the obvious limitation of MEA is that 'orphan' terms or genes (without strong relationships to neighbor terms/genes) could be left out from the analysis. Thus, it is important to examine those terms or genes that are left out during analysis when using MEA (60). In addition, the quality of the pre-selected gene list impacts the analytic results, just as it does in SEA analysis.

REMAINING QUESTIONS AND CHALLENGES IN THE FIELD

1. Realistically positioning the role of enrichment *P*-values in the current data-mining environment

The high-throughput enrichment data-mining environment is extremely complicated. Variations of the user gene list size, the deviation of the number of genes associated with each annotation, the gene overlap between annotations, the incompleteness of annotation content, the strong connectivity/dependency among genes, unbalanced distributions of annotation content, and high/low frequency of annotation content are examples of sources leading to this complexity and variation. None of the statistical methods mentioned in Table 1 is perfectly suitable for all situations. The complex situations found in the biological data-mining environment determine the discovery sensitivity and specificity (1—false-positive rate) of those statistical methods that are not yet in an optimal state, as discussed by Goeman *et al.* (73,80,81). Therefore, in real-life practice, many data analysts may treat the resulting enrichment *P*-values as a scoring system that plays a advisory role: i.e. rank and suggest possible relevant annotation terms, as opposed to an absolute, decision-making role (82). The analysts themselves are still playing critical roles in making the final decisions in terms of the most relevant, enriched annotation terms that are highlighted by the enrichment analysis tool. Even though annotation terms may be associated with very significant enrichment *P*-values, it is not uncommon that analysts discard/ignore some of the enriched annotation terms (such as terms with enrichment *P*-values <0.001) because they are not 'making sense' to a given study, based on *a priori* biological knowledge. The analogous example of this type of situation is

like that of a *Google* search, which returns some results that are not relevant to the user's original query. It is up to the user, based on his or her knowledge of the situation, to make the final judgment about the results. Collectively, current enrichment analysis is more of an exploratory procedure, with the aid of enrichment *P*-value, rather than a pure statistical solution. The notion that the enriched terms should make sense based on *a priori* biological knowledge of the study is the most important guideline to help users in adjusting analytic thresholds and thereby answering questions such as, 'Should my enrichment *P*-value cutoff be 0.05 or 0.01?' or 'Should I always consider the term with a significant enrichment *P*-value like 0.001?' or 'Which enrichment tool(s) could be more sensitive to my dataset?'

The most popular and traditional statistical methods used in the enrichment calculation are Fisher exact, Chi-square, Hypergeometric distribution and Binomial distribution, as collected in Table 1 and Supplementary Data 1. It is believed on a principal level that Binomial probability is good for analysis with a large population background. The Fisher exact test, Chi-square test and the Hypergeometric distribution are better for analysis with a smaller population background (12) (see subsection #4 for more discussion about population background). Given the weakness of the typical statistical methods, some alternative mathematical approaches were recently proposed in an attempt to improve the enrichment *P*-value calculations. These approaches include (but are not limited to) mid-*P*-value by Rivals *et al.* (76), finite partially ordered set approach (POSET) by POSOC (83,84), hidden Kripke model (HKM) by GOLie, greedy heuristics by ProfCom (68), Fisher's inverse chi-squared by GOFAA (50), master-target test/mutually exclusive target–target/intersecting target–target tests by GeneTools (42), EASE Score by EASE (8), Yule's Q by ProbCD (73), Fold Change by GoMiner (39) and Bayesian by BayGO (52). However, it is still too early to state definitively whether some of the improved alternative statistical methods really stand out over the traditional statistical approaches. Given the very complex data-mining environments discussed throughout the manuscript, all current statistical methods are working largely at the edge of their intended capability. Indeed, the specificity of enrichment analysis is more impacted by non-statistical layers than it is by statistical methods alone. In this sense, it is not realistic to guide users to choose enrichment tools simply according to statistical methods that are based purely on statistical advantages/disadvantages. Thus, we do not extensively discuss the differences between statistical methods, since such a discussion could potentially mislead a user's judgment. It is in the user's best interests to try many statistical methods on the same dataset and to compare the results whenever possible. Obviously, the need for new, more robust statistical methods to overcome the limitations of the current methods is still in high demand by the field.

2. Understanding the limitation of multiple testing correction on enrichment *P*-values

According to standard statistical principles, the more annotations that are tested, the greater the chance of an increase in the family-wide false-positive rate (85,86). To control the family-wide false-positive rate in the result list, the review article by Khatri *et al.* (11,12) indicates that the multiple test correction of enrichment *P*-values must be performed on the functional annotation categories being tested at the same time. Indeed, the majority of the tools performed such corrections with methods such as Bonferroni, Benjamini–Hochberg, Holm, Q-value, Permutation, etc. (Supplementary Data 1). Given the extremely complicated gene functional data-mining environment as discussed in the previous section, a critical question is how much of an improvement in discovery sensitivity and specificity (1—false-positive rate) is achieved by applying such corrections in real-life practice?

Even though many enrichment tools implement such corrections, only a few tools systematically provide evidence regarding the improvements of discovery results with and without such corrections in real-life analytic environments, rather than believing the benefits based on the statistical principle alone. Recently, GOSSIP (27) comprehensively compared the discovery sensitivity and specificity across various correction techniques provided by various tools with real-life datasets. It was concluded that the common multiple testing correction techniques, known to be overly conservative approaches if there are thousands or even more annotation terms involved in the analysis, may not improve specificity as much as people had believed those techniques would. In fact, the sensitivity may actually be negatively affected because of the conservative nature of these corrections (27).

Given the complexity of biological data-mining environments, the enrichment *P*-values derived from the common statistical methods can be very fragile, and are influenced not only by the statistical methods themselves, but also greatly by the algorithms, data sources, the individual biological process itself and so on. The specificity of the discovery is indeed greatly impacted by the non-statistical layers, which cannot be simply fixed by multiple test corrections. Great efforts regarding sensitivity and specificity issues involved in the enrichment analysis may require that improvements are made on the fundamental, non-statistical layers first (Figure 1). Then, the power of various statistical approaches including the multiple test correction can be utilized fully in the enrichment analysis. More than a dozen of the enrichment tools, including recent ones such as EasyGO (66) and g:Profiler (64), as well as the earlier ones such as GoMiner (10), have not implemented multiple test corrections (Supplementary Data 1), but are still widely used by the community in real-life data-mining projects. In summary, the multiple test correction is only a partial solution, not a resolution of the specificity problem in current enrichment analysis platforms.

3. Cross-comparing enrichment analysis results derived from multiple gene lists

A larger gene list can have higher statistical power, resulting in a higher sensitivity (more significant *P*-values) to slightly enriched terms, as well as to more specific terms.

On the other hand, the sensitivity is decreased toward largely enriched terms and broader terms. Thus, the size of the gene list impacts the absolute enrichment *P*-values, making it difficult to directly compare the absolute enrichment *P*-values across gene lists. Regardless of the challenges, cross-comparisons sometimes are necessary and important when studying the changes/trends among multiple time course datasets. Tools, such as GOBar (32), Go-Mapper (22), GOAlie, PageMan (51), high-throughput GoMiner (39), and the most recent, GOEAST (70), are intended to provide some of these capabilities to display multiple time course datasets simultaneously. However, users should keep the *P*-value comparison issue in mind when using these tools. The issue is even more critical, particularly when the sizes of gene lists are dramatically different from each other. More comprehensive and appropriate algorithms regarding the comparisons are still in high demand in the field.

4. Setting up the 'right' gene reference background

As noted in our previous example, 10% of the user's genes selected by a microarray experiment are kinases, as opposed to 1% of the genes in the human genome (this is the gene population background) that are kinases. The enrichment can therefore be quantitatively measured. A conclusion may be obtained for the particular example, that is, kinases are enriched in the user's study, and therefore play important roles in the study. However, 10% alone cannot lead to such a conclusion without comparison to the gene reference background (i.e. 1%). Thus, the different gene reference background settings may greatly impact the enrichment *P*-values, even when using the same statistical method and annotation content (12). For example, tools such as GOToolBox (18), GOstat (14), GoMiner (10), FatiGO (13) and GOTM (24), use the total genes in the genome as a global reference background. They tend to give more significant *P*-values, as compared to the tools (e.g. Onto-Express) using a narrowed-down set of genes (e.g. genes only existing on a microarray) as a gene reference background. In addition, DAVID (61) tends to be more conservative by using genes existing on the array and found to be associated with terms in the corresponding annotation categories, as the gene reference background. Many tools further allow users to upload a customized gene list as a gene reference background (Supplementary Data 1). Even though there is no 'gold' standard for the reference background, a general guideline is to set up the reference background as the pool of genes that could be selected for the studied annotation category (12). For example, the total genes found on a microarray chip seem to be the 'right' reference background, if the analysis gene list is derived from a microarray study conducted with the given chip. However, it is not perfect, since some genes on the chip could have little or no chance to be selected during the study, due to a low expression level that falls below the microarray detection range, and/or 'bad' probe design, etc. Even though the gene reference background directly impacts enrichment *P*-value, it will impact the *P*-values of all terms in a relatively similar manner within the same analysis. For the same dataset,

analyzed with different gene reference backgrounds, the output rank/order of the enrichment terms will remain relatively the same, even though the terms may be associated with different *P*-values. Such stable order/rank of enrichment terms in the output is more important than their absolute *P*-values so that the annotation exploration and conclusion on the same dataset will be similar and comparable when using different gene reference backgrounds. In this sense, another important principle of setting a gene reference background is to use a consistent gene reference background within the same analysis.

5. Extending backend annotation databases

Due to its enriched content and suitable data structure for high-throughput data mining, GO (1) is the only backend data source used in most, if not all, of the earlier enrichment tools, as well as in some of the more recent tools (Supplementary Data 1). However, many different biological aspects are being maintained and annotated by different independent resources; these aspects have not only a significant amount of overlapping information, but also a significant amount of unique data, due to the differing focus of the specialized groups. No one, single source is able to maintain all of the biological aspects, such as GO for the biological process, molecular functions or cellular components; Pfam for protein domains; BIND for protein–protein interactions; KEGG for pathways; TRANSFAC for gene regulations; GNF for gene–tissue expressions; OMIM for gene–disease associations; and so on (65,87,88). In this sense, a comprehensive backend database integrated with diverse and heterogeneous data sources will allow the enrichment tools to more comprehensively mine the large gene lists on broad-based annotation content covering different biological aspects, rather than on GO content alone. Obviously, the improvement of the annotation database alone can significantly improve the comprehensiveness of the data mining. Otherwise, the power of advanced data-mining algorithms and statistics cannot be fully utilized in the enrichment analysis.

Many tools are still using GO as the only backend database in the enrichment analysis (Supplementary Data 1). However, some recent tools or new releases of early-generation tools, such as Onto-Express (62), DAVID (61), WebGestalt (40), Fatigo+ (56), FACT (30), g:Profiler (64), GAzer (63) and GeneTrail (57), etc., extended their backend bio-databases by integrating wide-range heterogeneous data content (e.g. GO, KEGG pathways, protein domains, disease association, tissue expression, etc.) in order to increase the comprehensiveness of the enrichment analytic results. The WebGestalt, DAVID and Onto-Express groups independently reported their efforts in detail, with the resulting collections including GeneKeyDB, the DAVID Knowledgebase and OT, respectively (65,87,88). Each group described the steps involved in integrating and constructing such large bio-databases, particularly for the purposes of high-throughput gene functional analysis. Moreover, the databases of L2L (34) and DAVID (61) include gene expression data from publicly available SAGE, EST and microarray

studies. Thus, the user's dataset may be aligned with this data with similar conditions during functional analysis. Regarding species coverage, although the backend databases of several of the enrichment tools may cover a wide range of species, the support for a less popular species (i.e. rice) may not be as robust as that of more popular species (i.e. human, mouse, rat, yeast, fly). Given this situation, several enrichment tools were specifically designed for these less popular species, such as WEGO for rice (54); easyGO for crops (66); FINA for prokaryotes (58); CLENCH for Arabidopsis (21); JProGo for prokaryotes (48); BayGo for *Xylella fastidiosa* (52). Collectively, the quality, integration, and coverage of databases designed for high-throughput gene functional analysis have recently made notable progress, compared to that in earlier works. While the database improvement is an endless task, the current improvements have already significantly benefited individual groups and tools, as well as provided better backend bio-sources to the field for future tool development (65,87,88). The tools that still use GO as their only backend database should consider the integration of a wider collection of bio-databases in order to reflect the need and progress of the field.

6. Efficiently mapping users' input gene identifiers to the available annotation

If the gene identifier (ID) cannot be efficiently mapped to its corresponding annotation content, the subsequent data mining will be largely impaired. Thus, the comprehensiveness of mapping ID-to-ID and ID-to-annotation content in the database is essential as the first step to maximally translate gene lists into possible annotation content for further high-throughput enrichment analysis algorithms (12). However, this is not a simple and trivial issue when the identifiers representing gene/proteins are highly redundant, and are maintained by independent bioinformatics organizations. Even though the identifier cross-mapping issues were effectively addressed within each major bioinformatics organization, such as NCBI Entrez Gene (89), UniProt UniRef (90) and PIR-NREF (91), respectively, the weaker referencing capability across organizations still exists. For example, UniProt does not cover RefSeq IDs and NCBI Entrez Gene does not reference PIR ID at all. When different annotation databases use one system as their major gene identifier systems, e.g. GeneRif adopts NCBI IDs as major associated identifiers, and InterPro uses UniProt/SwissProt as major associated identifiers (65), some annotation content does not favor certain types of user input IDs. Thus, for a given type of ID, without special attention to this issue, important annotation content could be easily left out of the high-throughput analysis without the user's awareness, resulting in an incomplete or even failed enrichment analysis. Unfortunately, the enrichment tools, in general, have poorly documented how they handle the ID-to-ID and ID-to-annotation mapping issues. Most of the tools have likely adopted the existing work of another major group such as the NCBI Entrez Gene database (89). In such a case, although a tool may claim to support many ID systems, it does not mean that all types of IDs are fully integrated into the backend annotation database, due to the cross-organization issues discussed earlier. Some recent efforts, such as Onto-Translate (62), MatchMiner (92), IDConverter (93) and DAVID ID Converter (61), have made large improvements in an effort to help the ID-to-ID and ID-to-annotation mapping issue. With these aforementioned works, users may easily translate one type of ID to another. Moreover, they not only provide the improved cross-referencing capability but also enrich annotation content. For example, after gene IDs were re-agglomerated by a procedure called the DAVID Gene Concept, 10–20% more GO terms were able to be assigned to corresponding genes in the DAVID Knowledgebase, as compared to annotations in each individual source (65).

7. Enhancing the exploratory capability and graphical presentation

Due to the limitations of current enrichment analysis, the analysis of large gene lists, in the authors' opinion, is still more of an exploratory procedure rather than a single statistical solution at this time. Data analysts still play the most important role in interpreting the analytic results and collecting information from different views to make the final decision of which enriched annotation categories/biology are most relevant for the study in question. Such decisions are usually made with the aid of the enrichment *P*-values derived from the enrichment analysis, the previously known knowledge of expected biology relevant to experiments, and more importantly, the various data collected through exploration of the genes and annotation categories.

Flexibility in allowing users to define the analytic scope, e.g. GO levels, can make the analysis more focused in terms of a user's interests. Many tools, such as GOMiner (10), Onto-Express (62), DAVID (61) and FatiGO (56), support this type of flexibility. In addition, many tools, providing comprehensive links to primary annotation resources regarding annotation categories or gene reports, allow users to quickly and efficiently gather relevant information concerning items of interest. A Directed Acyclic Graph (DAG) maintains the structure of GO annotation terms (1). Even though all tools adopt GO in their enrichment analysis, most tools break down the structured nodes into flat terms during the calculation of enrichment *P*-values, and thereafter list the results in an easily readable tabular format. This simplified linear format and efficient organization of data for easy interpretation is widely used by most of the enrichment tools. Moreover, a number of tools, such as Onto-Express (62), easyGO (66), GoMiner (10), eGOn (42), GoSurfer (25), GOFFA (50) and GeneTrail (57), are able to display the enrichment analysis results on the DAG or a tree structure so that users may easily explore the enrichment results in neighboring nodes. Onto-Express further provides recalculation functions for 'drill down' analysis of a particular branch of the DAG. In contrast, POSOC (83) made an important note, that is, that DAG, as a structure, holds GO orientations, but lacks the power for biological inference, since a lot of functionally related

terms may be maintained in different DAG branches (83). Thus, more and more recent tools, such as Onto-Express (62), DAVID (61), POSOC (83), BayGO (52), FatiGO + (56), MAPPFinder (7), FuncCluster (43) and FunNet, have started to integrate BioCarta, KEGG, or other pathway visualizations in order to more efficiently examine the user's genes in a network context. In addition, some high-throughput pathway visualization tools, such as PathMAPA, Pathway Miner, Pathway Processor, ArrayXPath, Pathway Express, PathwayExplorer, KOBAS and VAMPIRE, are very useful, but are not included in this review because of their focuses on pathway analysis alone. Interestingly, biological module/classes of annotation terms, provided by PalS (67), DAVID (61) and GoToolBox (18), present heterogeneous annotation terms or genes in a group scope. This focuses the analysis on the larger biological picture and reduces the efforts involved in mining too many individual and redundant terms or genes. In addition, DAVID provides a simple 2D view visualization (61) that is able to efficiently display the related and heterogeneous many-genes-to-many-terms relationships, identified by the DAVID classification functions (60), on one well-organized page. Using such visualizations, users can efficiently examine the inter-relationships of highly related heterogeneous annotations and genes to pinpoint important commonalities and differences.

8. Evaluating the analytic capability of new enrichment tools

Sixty-eight enrichment tools, and potentially more that are missing from this collection, have already made the field very crowded. Many of the tool publications present minimal cross-comparisons to other tools. An appropriate standard evaluation procedure would make the analytic capability more comparable among tools, particularly for new tools. In addition, a good standard could make some new tools really stand out, as well as prevent redundant work from appearing in publications. Such standards should include, but not be limited to: a set of common datasets (gene lists) with expected and known biology in different, difficult levels for analysis; important aspects (e.g. backend database, enrichment *P*-values, speed, exploratory capability, graphic presentation, etc.) for cross-comparisons; emphasis on differences and advantages over other competing methods; etc. There is no detailed proposal as of yet, but obviously a standard is needed in the field.

9. Choosing the most appropriate enrichment tools from the various choices

Choosing the most suitable enrichment tool or tools largely depends on the users' research needs, IT experiences and the questions being asked. A precise guideline is most likely not possible since the research goals are very diverse from project to project. Before choosing a tool, a user may ask questions such as, 'Is the GO data source enough or are more (such as pathway, protein domain, protein–protein interactions, etc.) needed?'; 'Is the SEA linear enrichment report enough or do I really need MEA to look into inter-relationships?'; 'Is my experimental design simple enough to fit into the GSEA input requirement or is a comprehensive statistical method necessary for gene selection?'; 'What is my IT capability to handle R, standalone tools, or web tools?'; etc. Thereafter, tools that maximally meet the user's requirements can be logically selected. Table 2 compares the strength and limitation of each tool class. Instead of looking up individual tools among the overwhelming choices, it is recommended that the researchers locate the desired tool class (i.e. SEA, GSEA and MEA) first, then further narrow down to individual tools within that class. Supplementary Data 1 lists some of the aspects that users may be interested in, for every tool. In addition, a protocol paper regarding enrichment analysis by Huang *et al.* (82) could be useful for beginning users. SerbGO is a good site to search and compare detailed features and annotation coverage among tools. It is not recommended that the researchers choose tools simply according to the underlying enrichment statistical methods. As discussed in previous sections, the behavior of most statistical methods in current enrichment tools is working with large uncertainties.

Moreover, successful analytic works in higher-quality publications could serve as important examples to guide end users in the choice of 'well-used' tools and to follow analytic procedures for similar situations. Importantly, it is not unusual that different tools have similar capabilities and functions, but output very different results due to the variations in the implementations of the various important aspects. Thus, it is recommended that the user test multiple tools, which even offer similar analytic capability, in order to obtain the most satisfactory results (75).

CONCLUSIONS AND PERSPECTIVES

Due to the complexity of biological data-mining situations, in its current state, the analysis of large gene lists with the current enrichment tools is still more of an exploratory data-mining procedure rather than a pure statistical solution. The best analytic conclusions are made with the aid of the investigator's bio-knowledge, integrated annotation databases, computing algorithms and the enrichment *P*-values derived from statistical methods.

A large, linear list of enriched annotation terms in output reports may not satisfy researchers as much as it did years ago. The next generation of enrichment tools will strive for an integrative and comprehensive data-mining environment that will not only provide a more efficient means to identify the individual enriched annotations with improved databases, algorithms and statistical methods, but also comprehensively address the internal relationships of many enriched heterogeneous annotations. Tools with such capabilities could make the analysis more focused and understandable in a network context. Many of the most recently reported tools fall into the class II and III categories, which suggests such a trend in the field (Table 1 and Supplementary Data 1).

Finally, it can be expected that the activities and passions of developing new enrichment tools will continue, due to the unmet needs and limitations of current

enrichment analytic methods. A standard for evaluating new tools will facilitate the growth of the field.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

FUNDING

REFERENCES

1. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
2. Khatri,P., Draghici,S., Ostermeier,G.C. and Krawetz,S.A. (2002) Profiling gene expression using onto-express. *Genomics*, **79**, 266–270.
3. Robinson,M.D., Grigull,J., Mohammad,N. and Hughes,T.R. (2002) FunSpec: a web-based cluster interpreter for yeast. *BMC Bioinformatics*, **3**, 35.
4. Berriz,G.F., King,O.D., Bryant,B., Sander,C. and Roth,F.P. (2003) Characterizing gene sets with FuncAssociate. *Bioinformatics*, **19**, 2502–2504.
5. Castillo-Davis,C.I. and Hartl,D.L. (2003) GeneMerge—post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, **19**, 891–892.
6. Dennis,G., ,, Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.*, **4**, P3.
7. Doniger,S.W., Salomonis,N., Dahlquist,K.D., Vranizan,K., Lawlor,S.C. and Conklin,B.R. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol.*, **4**, R7.
8. Hosack,D.A., Dennis,G.Jr, Sherman,B.T., Lane,H.C. and Lempicki,R.A. (2003) Identifying biological themes within lists of genes with EASE. *Genome Biol.*, **4**, R70.
9. Martinez-Cruz,L.A., Rubio,A., Martinez-Chantar,M.L., Labarga,A., Barrio,I., Podhorski,A., Segura,V., Sevilla Campo,J.L., Avila,M.A. and Mato,J.M. (2003) GARBAN: genomic analysis and rapid biological annotation of cDNA microarray and proteomic data. *Bioinformatics*, **19**, 2158–2160.
10. Zeeberg,B.R., Feng,W., Wang,G., Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C., Lababidi,S. *et al.* (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.
11. Curtis,R.K., Oresic,M. and Vidal-Puig,A. (2005) Pathways to the analysis of microarray data. *Trends Biotechnol.*, **23**, 429–435.
12. Khatri,P. and Draghici,S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.
13. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
14. Beissbarth,T. and Speed,T.P. (2004) GOstat: find statistically over-represented Gene Ontologies within a group of genes. *Bioinformatics*, **20**, 1464–1465.
15. Boyle,E.I., Weng,S., Gollub,J., Jin,H., Botstein,D., Cherry,J.M. and Sherlock,G. (2004) GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*, **20**, 3710–3715.
16. Breitling,R., Amtmann,A. and Herzyk,P. (2004) Iterative Group Analysis (iGA): a simple tool to enhance sensitivity and facilitate interpretation of microarray experiments. *BMC Bioinformatics*, **5**, 34.
17. Breslin,T., Eden,P. and Krogh,M. (2004) Comparing functional annotation analyses with Catmap. *BMC Bioinformatics*, **5**, 193.
18. Martin,D., Brun,C., Remy,E., Mouren,P., Thieffry,D. and Jacq,B. (2004) GOToolBox: functional analysis of gene datasets based on Gene Ontology. *Genome Biol.*, **5**, R101.
19. Masseroli,M., Martucci,D. and Pinciroli,F. (2004) GFINDer: Genome Function INtegrated Discoverer through dynamic annotation, statistical analysis, and mining. *Nucleic Acids Res.*, **32**, W293–300.
20. Pasquier,C., Girardot,F., Jevardat de Fombelle,K. and Christen,R. (2004) THEA: ontology-driven analysis of microarray data. *Bioinformatics*, **20**, 2636–2643.
21. Shah,N.H. and Fedoroff,N.V. (2004) CLENCH: a program for calculating Cluster ENriCHment using the Gene Ontology. *Bioinformatics*, **20**, 1196–1197.
22. Smid,M. and Dorssers,L.C. (2004) GO-Mapper: functional analysis of gene expression data using the expression level as a score to evaluate Gene Ontology terms. *Bioinformatics*, **20**, 2618–2625.
23. Volinia,S., Evangelisti,R., Francioso,F., Arcelli,D., Carella,M. and Gasparini,P. (2004) GOAL: automated Gene Ontology analysis of expression profiles. *Nucleic Acids Res.*, **32**, W492–499.
24. Zhang,B., Schmoyer,D., Kirov,S. and Snoddy,J. (2004) GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using Gene Ontology hierarchies. *BMC Bioinformatics*, **5**, 16.
25. Zhong,S., Storch,K.F., Lipan,O., Kao,M.C., Weitz,C.J. and Wong,W.H. (2004) GoSurfer: a graphical interactive tool for comparative analysis of large gene sets in Gene Ontology space. *Appl. Bioinformatics*, **3**, 261–264.
26. Al-Shahrour,F., Minguez,P., Vaquerizas,J.M., Conde,L. and Dopazo,J. (2005) BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res.*, **33**, W460–464.
27. Bluthgen,N., Brand,K., Cajavec,B., Swat,M., Herzel,H. and Beule,D. (2005) Biological profiling of gene groups utilizing Gene Ontology. *Genome Inform.*, **16**, 106–115.
28. Boorsma,A., Foat,B.C., Vis,D., Klis,F. and Bussemaker,H.J. (2005) T-profiler: scoring the activity of predefined groups of genes using gene expression data. *Nucleic Acids Res.*, **33**, W592–595.
29. Kim,S.Y. and Volsky,D.J. (2005) PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics*, **6**, 144.
30. Kokocinski,F., Delhomme,N., Wrobel,G., Hummerich,L., Toedt,G. and Lichter,P. (2005) FACT–a framework for the functional interpretation of high-throughput experiments. *BMC Bioinformatics*, **6**, 161.
31. Lee,H.K., Braynen,W., Keshav,K. and Pavlidis,P. (2005) ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, **6**, 269.
32. Lee,J.S., Katari,G. and Sachidanandam,R. (2005) GObar: a gene ontology based analysis and visualization tool for gene sets. *BMC Bioinformatics*, **6**, 189.

33. Maere,S., Heymans,K. and Kuiper,M. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*, **21**, 3448–3449.

34. Newman,J.C. and Weiner,A.M. (2005) L2L: a simple tool for discovering the hidden significance in microarray expression data. *Genome Biol.*, **6**, R81.

35. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.

36. Tu,K., Yu,H. and Zhu,M. (2005) MEGO: gene functional module expression based on gene ontology. *Biotechniques*, **38**, 277–283.

37. Wrobel,G., Chalmel,F. and Primig,M. (2005) goCluster integrates statistical analysis and functional interpretation of microarray expression data. *Bioinformatics*, **21**, 3575–3577.

38. Young,A., Whitehouse,N., Cho,J. and Shaw,C. (2005) OntologyTraverser: an R package for GO analysis. *Bioinformatics*, **21**, 275–276.

39. Zeeberg,B.R., Qin,H., Narasimhan,S., Sunshine,M., Cao,H., Kane,D.W., Reimers,M., Stephens,R.M., Bryant,D., Burt,S.K. *et al.* (2005) High-throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of Common Variable Immune Deficiency (CVID). *BMC Bioinformatics*, **6**, 168.

40. Zhang,B., Kirov,S. and Snoddy,J. (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Res.*, **33**, W741–748.

41. Alexa,A., Rahnenfuhrer,J. and Lengauer,T. (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.

42. Beisvag,V., Junge,F.K., Bergum,H., Jolsum,L., Lydersen,S., Gunther,C.C., Ramampiaro,H., Langaas,M., Sandvik,A.K. and Laegreid,A. (2006) GeneTools—application for functional annotation and statistical hypothesis testing. *BMC Bioinformatics*, **7**, 470.

43. Henegar,C., Cancello,R., Rome,S., Vidal,H., Clement,K. and Zucker,J.D. (2006) Clustering biological annotations and gene expression data to identify putatively co-regulated biological processes. *J. Bioinform. Comput. Biol.*, **4**, 833–852.

44. Lewin,A. and Grieve,I.C. (2006) Grouping Gene Ontology terms to improve the assessment of gene set enrichment in microarray data. *BMC Bioinformatics*, **7**, 426.

45. Nam,D., Kim,S.B., Kim,S.K., Yang,S., Kim,S.Y. and Chu,I.S. (2006) ADGO: analysis of differentially expressed gene sets using composite GO annotation. *Bioinformatics*, **22**, 2249–2253.

46. Pereira,G.S., Brandao,R.M., Giuliatti,S., Zago,M.A.Jr and Silva,W.A. (2006) Gene class expression: analysis tool of Gene Ontology terms with gene expression data. *Genet. Mol. Res.*, **5**, 108–114.

47. Rubin,E. (2006) Circumventing the cut-off for enrichment analysis. *Brief Bioinform.*, **7**, 202–203.

48. Scheer,M., Klawonn,F., Munch,R., Grote,A., Hiller,K., Choi,C., Koch,I., Schobert,M., Hartig,E., Klages,U. *et al.* (2006) JProGO: a novel tool for the functional interpretation of prokaryotic microarray data using Gene Ontology information. *Nucleic Acids Res.*, **34**, W510–515.

49. Sealfon,R.S., Hibbs,M.A., Huttenhower,C., Myers,C.L. and Troyanskaya,O.G. (2006) GOLEM: an interactive graph-based gene-ontology navigation and analysis tool. *BMC Bioinformatics*, **7**, 443.

50. Sun,H., Fang,H., Chen,T., Perkins,R. and Tong,W. (2006) GOFFA: Gene Ontology For Functional Analysis – A FDA Gene Ontology tool for analysis of genomic and proteomic data. *BMC Bioinformatics*, **7**, S23.

51. Usadel,B., Nagel,A., Steinhauser,D., Gibon,Y., Blasing,O.E., Redestig,H., Sreenivasulu,N., Krall,L., Hannah,M.A., Poree,F. *et al.* (2006) PageMan: an interactive ontology tool to generate, display, and annotate overview graphs for profiling experiments. *BMC Bioinformatics*, **7**, 535.

52. Vencio,R.Z., Koide,T., Gomes,S.L. and Pereira,C.A. (2006) BayGO: Bayesian analysis of ontology term enrichment in microarray data. *BMC Bioinformatics*, **7**, 86.

53. Verspoor,K., Cohn,J., Mniszewski,S. and Joslyn,C. (2006) A categorization approach to automated ontological function annotation. *Protein Sci.*, **15**, 1544–1549.

54. Ye,J., Fang,L., Zheng,H., Zhang,Y., Chen,J., Zhang,Z., Wang,J., Li,S., Li,R., Bolund,L. *et al.* (2006) WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.*, **34**, W293–297.

55. Al-Shahrour,F., Arbiza,L., Dopazo,H., Huerta-Cepas,J., Minguez,P., Montaner,D. and Dopazo,J. (2007) From genes to functional classes in the study of biological systems. *BMC Bioinformatics*, **8**, 114.

56. Al-Shahrour,F., Minguez,P., Tarraga,J., Medina,I., Alloza,E., Montaner,D. and Dopazo,J. (2007) FatiGO + : a functional profiling tool for genomic data. Integration of functional annotation, regulatory motifs and interaction data with microarray experiments. *Nucleic Acids Res.*, **35**, W91–96.

57. Backes,C., Keller,A., Kuentzer,J., Kneissl,B., Comtesse,N., Elnakady,Y.A., Muller,R., Meese,E. and Lenhof,H.P. (2007) GeneTrail—advanced gene set enrichment analysis. *Nucleic Acids Res.*, **35**, W186–192.

58. Blom,E.J., Bosman,D.W., van Hijum,S.A., Breitling,R., Tijsma,L., Silvis,R., Roerdink,J.B. and Kuipers,O.P. (2007) FIVA: Functional Information Viewer and Analyzer extracting biological knowledge from transcriptome data of prokaryotes. *Bioinformatics*, **23**, 1161–1163.

59. Carmona-Saez,P., Chagoyen,M., Tirado,F., Carazo,J.M. and Pascual-Montano,A. (2007) GENECODIS: a web-based tool for finding significant concurrent annotations in gene lists. *Genome Biol.*, **8**, R3.

60. Huang da,W., Sherman,B.T., Tan,Q., Collins,J.R., Alvord,W.G., Roayaei,J., Stephens,R., Baseler,M.W., Lane,H.C. and Lempicki,R.A. (2007) The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.*, **8**, R183.

61. Huang da,W., Sherman,B.T., Tan,Q., Kir,J., Liu,D., Bryant,D., Guo,Y., Stephens,R., Baseler,M.W., Lane,H.C. *et al.* (2007) DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res.*, **35**, W169–W175.

62. Khatri,P., Voichita,C., Kattan,K., Ansari,N., Khatri,A., Georgescu,C., Tarca,A.L. and Draghici,S. (2007) Onto-Tools: new additions and improvements in 2006. *Nucleic Acids Res.*, **35**, W206–W211.

63. Kim,S.B., Yang,S., Kim,S.K., Kim,S.C., Woo,H.G., Volsky,D.J., Kim,S.Y. and Chu,I.S. (2007) GAzer: gene set analyzer. *Bioinformatics*, **23**, 1697–1699.

64. Reimand,J., Kull,M., Peterson,H., Hansen,J. and Vilo,J. (2007) g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res.*, **35**, W193–200.

65. Sherman,B.T., Huang da,W., Tan,Q., Guo,Y., Bour,S., Liu,D., Stephens,R., Baseler,M.W., Lane,H.C. and Lempicki,R.A. (2007) DAVID Knowledgebase: a gene-centered database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. *BMC Bioinformatics*, **8**, 426.

66. Zhou,X. and Su,Z. (2007) EasyGO: Gene Ontology-based annotation and functional enrichment analysis tool for agronomical species. *BMC Genomics*, **8**, 246.

67. Alibes,A., Canada,A. and Diaz-Uriarte,R. (2008) PaLS: filtering common literature, biological terms and pathway information. *Nucleic Acids Res.*, **36**, W364–W367.

68. Antonov,A.V., Schmidt,T., Wang,Y. and Mewes,H.W. (2008) ProfCom: a web tool for profiling the complex functionality of gene groups identified from high-throughput data. *Nucleic Acids Res.*, **36**, W347–W351.

69. Bauer,S., Grossmann,S., Vingron,M. and Robinson,P.N. (2008) Ontologizer 2.0 - A multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics.*, **24**, 1650–1651.

70. Zheng,Q. and Wang,X.J. (2008) GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. *Nucleic Acids Res.*, **36**, W358–W363.

71. Frohlich,H., Speer,N., Poustka,A. and Beissbarth,T. (2007) GOSim—an R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics*, **8**, 166.

72. Zhu,J., Wang,J., Guo,Z., Zhang,M., Yang,D., Li,Y., Wang,D. and Xiao,G. (2007) GO-2D: identifying 2-dimensional cellular-localized functional modules in Gene Ontology. *BMC Genomics*, **8**, 30.

73. Vencio,R.Z. and Shmulevich,I. (2007) ProbCD: enrichment analysis accounting for categorization uncertainty. *BMC Bioinformatics*, **8**, 383.

74. Mosquera,J.L. and Sanchez-Pla,A. (2008) SerbGO: searching for the best GO tool. *Nucleic Acids Res.*, **36**, W368–371.

75. Rhee,S.Y., Wood,V., Dolinski,K. and Draghici,S. (2008) Use and misuse of the gene ontology annotations. *Nat. Rev. Genet.*, **9**, 509–515.

76. Rivals,I., Personnaz,L., Taing,L. and Potier,M.C. (2007) Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics*, **23**, 401–407.

77. Nilsson,B., Hakansson,P., Johansson,M., Nelander,S. and Fioretos,T. (2007) Threshold-free high-power methods for the ontological analysis of genome-wide gene-expression studies. *Genome Biol.*, **8**, R74.

78. Yang,D., Li,Y., Xiao,H., Liu,Q., Zhang,M., Zhu,J., Ma,W., Yao,C., Wang,J., Wang,D. *et al.* (2008) Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories. *Bioinformatics*, **24**, 265–271.

79. Jiang,Z. and Gentleman,R. (2007) Extensions to gene set enrichment. *Bioinformatics*, **23**, 306–313.

80. Goeman,J.J. and Buhlmann,P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, **23**, 980–987.

81. Gold,D.L., Coombes,K.R., Wang,J. and Mallick,B. (2007) Enrichment analysis in high-throughput genomics - accounting for dependency in the NULL. *Brief Bioinform.*, **8**, 71–77.

82. Huang,D.W., Sherman,B.T. and Lempicki,R.A. (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, doi: 10.1038/nprot.2008.211.

83. Joslyn,C.A., Mniszewski,S.M., Fulmer,A. and Heaton,G. (2004) The gene ontology categorizer. *Bioinformatics*, **20**, i169–177.

84. Barriot,R., Sherman,D.J. and Dutour,I. (2007) How to decide which are the most pertinent overly-represented features during gene set enrichment analysis. *BMC Bioinformatics*, **8**, 332.

85. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.

86. Dudoit,S., Popper,J. and Boldrick,S. (2003) Multiple hypothesis testing in microarray experiments. *Stat. Sci.*, **18**, 71–103.

87. Draghici,S., Sellamuthu,S. and Khatri,P. (2006) Babel's tower revisited: a universal resource for cross-referencing across annotation databases. *Bioinformatics*, **22**, 2934–2939.

88. Kirov,S.A., Peng,X., Baker,E., Schmoyer,D., Zhang,B. and Snoddy,J. (2005) GeneKeyDB: a lightweight, gene-centric, relational database to support data mining environments. *BMC Bioinformatics*, **6**, 72.

89. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.

90. The UniProt Consortium (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.

91. Wu,C.H., Yeh,L.S., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al.* (2003) The protein information resource. *Nucleic Acids Res.*, **31**, 345–347.

92. Bussey,K.J., Kane,D., Sunshine,M., Narasimhan,S., Nishizuka,S., Reinhold,W.C., Zeeberg,B., Ajay,W. and Weinstein,J.N. (2003) MatchMiner: a tool for batch navigation among gene and gene product identifiers. *Genome Biol.*, **4**, R27.

93. Alibes,A., Yankilevich,P., Canada,A. and Diaz-Uriarte,R. (2007) IDconverter and IDClight: conversion and annotation of gene and protein IDs. *BMC Bioinformatics*, **8**, 9.