Journal of Biomolecular Techniques • Volume 34(1); 2023 Apr

Functional Annotation Routines Used by ABRF Bioinformatics Core Facilities - Observations, Comparisons, and Considerations

Charles A. Whittaker[1] Alper Kucukural[2] Chris Gates[3] Owen Michael Wilkins[4,5] George W. Bell[6] John N. Hutchinson[7] Shawn W. Polson[8] Julie Dragon[9]

[1]Barbara K. Ostrom (1978) Bioinformatics and Computing Core Facility, Swanson Biotechnology Center, Koch Institute at the Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA,

[2]Bioinformatics Core, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA,

[3]BRCF Bioinformatics Core, University of Michigan, Ann Arbor, Michigan 48109, USA,

[4]Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire 03755, USA,

[5]Dartmouth Cancer Center, Dartmouth Hitchcock Medical Center, Lebanon, New Hampshire 03756, USA,

[6]Bioinformatics and Research Computing, Whitehead Institute, Cambridge, Massachusetts 02142, USA,

[7]Harvard T.H. Chan School of Public Health, Department of Biostatistics, Boston, Massachusetts 02115, USA,
[8]Bioinformatics Core, Center for Bioinformatics and Computational Biology, University of Delaware, Delaware Biotechnology Institute, Newark, Delaware 19713, USA,
[9]Vermont Integrative Genomics Resource and Vermont Biomedical Research Network Bioinformatic Core, University of Vermont, Burlington, Vermont 05405, USA

Association of Biomolecular Resource Facilities

Published on: Mar 27, 2023

DOI: https://doi.org/10.7171/3fc1f5fe.0b74b9db

ABSTRACT

The functional annotation of gene lists is a common analysis routine required for most genomics experiments, and bioinformatics core facilities must support these analyses. In contrast to methods such as the quantitation of RNA-Seq reads or differential expression analysis, our research group noted a lack of consensus in our preferred approaches to functional annotation. To investigate this observation, we selected 4 experiments that represent a range of experimental designs encountered by our cores and analyzed those data with 6 tools used by members of the Association of Biomolecular Resource Facilities (ABRF) Genomic Bioinformatics Research Group (GBIRG). To facilitate comparisons between tools, we focused on a single biological result for each experiment. These results were represented by a gene set, and we analyzed these gene sets with each tool considered in our study to map the result to the annotation categories presented by each tool. In most cases, each tool produces data that would facilitate identification of the selected biological result for each experiment. For the exceptions, Fisher's exact test parameters could be adjusted to detect the result. Because Fisher's exact test is used by many functional annotation tools, we investigated input parameters and demonstrate that, while background set size is unlikely to have a significant impact on the results, the numbers of differentially expressed genes in an annotation category and the total number of differentially expressed genes under consideration are both critical parameters that may need to be modified during analyses. In addition, we note that differences in the annotation categories tested by each tool, as well as the composition of those categories, can have a significant impact on results.

ADDRESS CORRESPONDENCE TO: Charles A. Whittaker, Koch Institute at MIT, 77 Massachusetts Ave. 76-189B, Cambridge, Massachusetts 02139 (E-mail: charliew@mit.edu; Phone: 617-324-0337; Fax: 617-324-2238)

Conflict of Interest Disclosure: The authors have no financial support or associations that pose a conflict of interest.

Keywords: gene expression profiling, gene ontology, knowledge bases

INTRODUCTION

The functional annotation of gene lists generated during gene expression (RNA-Seq, microarrays), chromatin state (ChIP-Seq, ATAC-Seq), or functional screening experiments is an essential service provided by bioinformatics cores. These analyses enable interpretation of the data and the generation of hypotheses for subsequent experiments. Numerous options are available to perform these studies, ranging from comprehensive and costly commercial packages to open-source tools. The statistical approaches involved fall into 3 general categories: over-representation analysis (ORA), functional class sorting (FCS), and topology-based analysis (TB). ORA starts with a list of differentially expressed genes and then uses Fisher's exact test to

determine if genes from any functional group overlap with that list to a larger degree than what would be anticipated by chance. This method is used by open-source tools such as Database for Annotation, Visualization and Integrated Discovery (DAVID),[1] g:Profiler,[2] Biological Networks Gene Ontology tool (BiNGO),[3] GO:TermFinder,[4] OntoExpress,[5] and FuncAssociate[6] as well as proprietary tools such as Metacore[7] and Ingenuity Pathway Analysis (IPA).[8] The input for FCS is the entire list of genes considered in an experiment ordered from high in one class to high in another using an experimentally derived statistic. FCS then applies the Kolmogorov–Smirnov statistic to determine if functionally grouped genes show coordinated association with 1 of the classes. This method is used by gene set enrichment analysis (GSEA).[9], [10] TB methods consider relationships and interactions between genes within annotation groups. This process emphasizes genes that play key regulatory roles in pathways that are under investigation, and as a result, it may be more sensitive in the discovery of biologically meaningful results. Examples of this method include ROntoTools,[11] and a commercial implementation of ROntoTools called iPathwayGuide,[12] Signaling Pathway Impact Analysis (SPIA),[13] and CogNet.[14]

Each type of analysis requires functional groupings of genes according to biological pathways or other characteristics. There are many sources for these annotations. The Gene Ontology (GO) Consortium initially published the first annotations more than 2 decades ago.[15] The aim of this Consortium was to create a dynamic, structured, controlled, and common vocabulary that captures information about gene products. The vocabulary included 3 independent ontologies: biological process, molecular function, and cellular component. In the GO, genes are associated with hierarchies of terms based on different types of evidence. The Disease Ontology database[16] takes a similar graph-based approach to describe disease states. Biological pathways are captured in databases such as Kyoto Encyclopedia of Genes and Genomes (KEGG),[17],[18] WikiPathways, [19] Pathway Interaction Database (PID),[20] and Reactome.[21] In addition, Metacore and IPA maintain their own proprietary databases of pathways, interactions, and functions. Many of these gene-to-term associations as well as experimentally derived gene lists are captured and presented in the Molecular Signature Database (MSigDB).[22]

The tools and annotation sources considered in this study are a fraction of what is available to researchers, and detailed reviews have been published. Xie et al., 2021 performed an analysis of 503 publications covering functional annotation tools or methods with a focus on popularity based on citation frequencies and performance using an analysis of validation approaches. They provide a Shiny application that aggregates reviews, methods papers, and benchmarking studies. Their results indicate that the 2 most popular tools are GSEA and DAVID. Furthermore, the large number of publications described in their work indicates that enormous effort has been expended to define the most highly performing functional annotation tools.[23] They also discuss the challenges facing benchmarking studies in this field and provide Jupyter workflows aimed at improving the results.

Comparative studies are challenging because different experiments can have dramatically different numbers of differentially expressed genes, and there is a lack of harmonization in gene set naming and content in the annotation repositories used by various tools. Nguyen et al., 2019 addressed these challenges in their benchmarking study by adjusting differential expression thresholds so that similar numbers of genes were included in each analysis, and they restricted their analysis to an annotation set of about 150 KEGG pathways.[24] These adjustments facilitated a comprehensive and quantitative comparison of 13 analysis tools and 86 (75 human/11 mouse) array-based gene expression experiments, and they were able to demonstrate improved performance of the TB methods. While these techniques worked well for their study, they are difficult to apply to the types of experiments encountered in core facilities. Rigorous statistical thresholds applied to experiments with variable design and complexity often result in variable numbers of differentially expressed genes that must be examined. Furthermore, the selection of annotation groups is often driven by experimental considerations and researcher requirements rather than algorithmic optimization.

Given the widespread use of functional annotation analyses in gene expression experiments and the number of analysis routines available, incorrect execution of these analyses is problematic. Wijesooriya et al., 2022 document routinely insufficient description of methods, inappropriate use of background lists in up to 95% of published examples, and failure to apply false discovery corrections in 43% of cases.[25] These authors emphasize the need for more rigorous standards in these workflows.

In functional analysis, there are biases that must be considered while performing any of the statistical approaches. A comprehensive functional description of all gene products remains incomplete. Previous GO annotation versions have been shown to be influenced by annotation bias, in which most annotations are for a small number of well-studied genes, or literature bias, in which a small number of articles contribute disproportionately many experimental annotations. These challenges have been noted,[26] and the problems affect a variety of applications such as GO enrichment analysis,[27] protein function prediction,[28] and gene network analysis.[29]

Here, we use a scenario-based approach to investigate different functional annotation routines. We were motivated by the observation that, within our research group, relatively little consensus existed with regard to best practices, and tool selection decisions were often based on cost, convenience, and researcher experience rather than any sort of a quantitative process. As a result, we prepared a framework to summarize our experiences with our preferred tools in a way that may be informative to other cores or researchers. Four example experimental comparisons that represent classes of experiments encountered by our cores were examined using a collection of 6 tools used by at least 1 member of the research group. The experimental comparisons range from a subtle experimental manipulation with limited differential gene expression to a comparison of different tissues with extensive differential gene expression. The tools include open-source and commercial applications that utilize ORA, TB, and FCS algorithms. We focused our comparisons on a single biological observation for each experiment while paying special attention to topics that were controversial

within our research group. Topics considered include background set size, differential expression thresholds, directionality of change, and commercial versus open-source tools.

MATERIALS AND METHODS

Data processing and differential expression analyses

The datasets considered in this study are listed in Table 1. Count data for G1vRev were obtained from the Gene Expression Omnibus (GEO, GSE83647), data for EvC are available from GEO (GSE55190), and count data were provided by the authors. The hpm3AvC FASTQ files were downloaded from GEO (GSE63966) and quantified using Salmon version 1.0.0[30] using the hg38 genome and an Ensembl version 98 annotation. SvC data were downloaded from Genotype-Tissue Expression (GTEx), GTEx Analysis v7.[31] Differential expression analysis was done using DESeq2.[32],[33] Version 1.24.0 and apeglm log fold change shrinkage was used for G1vRev, EvC, and hpm3AvC, and version 1.16.0 and normal log fold change shrinkage was used for SvC. Unless otherwise indicated, differentially expressed genes (DEGs) were defined as those having an absolute $\log_2$ fold change ≥ 1 and an False discovery rate (FDR)-adjusted P-value (padj) < 0.05. The final differential expression results used in this study are provided in Table S1.

Table 1

Datasets Considered in this Study

Dataset	Species	Comparison	DEG Count	Biological Concept Gene Set
G1Rev	Human	Aneuploid G1 arrested RPE-1 cells versus aneuploid RPE-1	102	HALLMARK_E2F_TARGETS
EvC	Mouse	Livers from Mapk8;Mapk9 null high-fat diet mice versus wild-type regular diet mice	1068	KEGG_PPAR_SIGNALING_PATHWAY
hpm3AvC	Human	hpm3 pleural cells treated with asbestos versus untreated cells	1289	HALLMARK_TNFA_SIGNALING_VIA_NFKB
SvC	Human	Skeletal muscle versus left ventricle cardiac muscle	6369	GO_SKELETAL_MUSCLE_CONTRACTION

Analysis tools

The tools considered in this study are listed in Table 2. Pre-ranked GSEA was run with the command line interface to GSEA 4.1 using the Wald Statistic from DESeq2 as ranking metric, weighted scoring scheme, and gene sets from the MsigDB v.7.2 collections hallmark, c2cp, c2cgp, c5cc, c5bp, c5mf. For the EvC experiment, mouse gene symbols were assigned to human orthologs using an orthology table provided by the Jackson Laboratory. Where orthology mapping was one-to-many, the human ortholog with the highest average expression across the dataset was selected as the representative. The Wald statistic rank files for each experiment are provided in Table S2. The normalized enrichment score (NES) and FDR q-value statistics were extracted from the output and aggregated into a single file for visualization and analysis. For GSEA concept mapping, the concept gene lists (Table S3) were used as input to the MSigDB Investigate Gene Sets tool in order to compute overlap with 6 gene set collections under investigation. The top 10 overlapping gene sets from each collection with FDR q-val < 0.05 were selected as concept-related gene sets (Table S4).

Table 2

Analysis Tools Considered in this Study

Tool	Cost	Interface	Statistic	Strategy
GSEA	Free	cli	FCS	Non-topology based
DAVID	Free	R	ORA	Non-topology based
g:profiler	Free	R	ORA	Non-topology based
Ingenuity Pathway Analysis	Paid	web	ORA	Non-topology based
MetaCore	Paid	web	ORA	Non-topology based
Advaita iPathwayGuide	Paid	web	ORA/TB	Topology based

DAVID was run using R (v 4.1.0) with RDAVIDWebService (v 1.28.0), org.Mm.eg.db (v 3.12.0), and org.Hs.eg.db (v 3.12.0). In some cases, these runs were supplemented with interactive runs with the DAVID web interface. Results for the annotation categories GOTERM_BP_DIRECT, GOTERM_MF_DIRECT, GOTERM_CC_DIRECT, KEGG_PATHWAY, REACTOME_PATHWAY, and BIOCARTA were collected for analysis and visualization. For DAVID concept mapping, genes in each concept set were mapped to Ensembl IDs and run using the default background for the appropriate species. For each annotation category in each run, up to the top 10 results with Bonferroni P-values less than 0.05 were flagged as related results (Table S4). In addition to the EASE (modified Fisher) P-values, DAVID presents Bonferroni, Benjamini P-values, and percent FDR for each result. There are relatively few significant results when filtering is done on any of these

adjusted P-values, suggesting that the DAVID statistical methods are more stringent than those in other tools. For the cell cycle analysis, the GOTERM_BP_ALL collection was also run.

g:Profiler was run using R (v 4.1.0) with gprofiler2 (v 0.2.0). The annotation sources used were GO:BP, GO:MF, GO:CC, KEGG, REAC, and WP. To identify g:profiler results related to the selected biological concepts, the gene sets were analyzed using annotated genes as a background scope. The control for multiple hypothesis testing was done with the g:profiler algorithm g:SCS. For each annotation category in each run, the top 10 results with g:SCS P-values less than 0.05 were flagged as related hits (Table S4). For the experimental data, DEG lists were obtained by application of abs(logFC) ≥ 1 and padj ≤ 0.05, and runs were carried out using custom backgrounds derived from the experimental data. Selected columns for all output from the R client runs were exported.

IPA was run using the web interface provided by the company with the default whole-genome background sets. The ontologies considered were Canonical Pathways and Diseases and Functions. To identify IPA annotations related to the selected biological concepts, the gene sets were analyzed using the default background. For both annotation categories in each run, the top 25 results with a P-value less than 0.05 were flagged as related hits (Table S4). For the experimental data, DEGs were submitted to the Core analysis tool and run with the default background. For Canonical Pathways, all results were exported; for Diseases and Functions, the top 500 were exported. For canonical pathways, IPA exports $-\log_{10}$(P-value) of the Fisher exact test raw P-value, and these were restored to linear space for plotting.

Metacore runs were executed using the web interface provided by the company and custom background sets derived from the input data. The ontologies considered were Pathway Maps, Process Networks, GO Processes, GO Molecular Functions, and GO Localizations. To identify Metacore annotations related to the selected biological concepts, the gene sets were analyzed using the default background. For each annotation category in each run, the top 10 results with an FDR less than 0.05 were flagged as related hits (Table S4). For the experimental data, custom background sets were prepared from the experimental data, and thresholds and DEGs were submitted for analysis. The top 50 results from each annotation category were extracted from the web output.

Advaita iPathwayGuide (iPG) was run using the web interface provided by the company. The ontologies considered were Pathway, GO:BP, GO:CC, and GO:MF. To identify Metacore annotations related to the selected biological concepts, the gene sets were analyzed using the default whole-genome background. For each annotation category in each run, the top 10 results plus ties with an FDR less than 0.05 were flagged as related hits (Table S4). For the experimental data, DEGs were selected, and the default whole-genome background was used. All results from each annotation category were exported from the web output.

Data analysis and visualization

For each of the 4 datasets and 6 analysis tools, the experimental results were combined and annotated with the results from the concept mapping procedure in which related terms are labeled as "hits" (Table S5). These assembled data were filtered, and plots were prepared using R version 4.2.1[34] with tidyverse_1.3.1. For the Fisher's test simulations, data frames were created in R, and tests were performed using the fisher.test function with the alternative parameter set to greater. Heat maps were prepared using Spotfire Analyst 10.10.3. The Fisher's test scripts as well as scripts used for plotting, DAVID analysis, and g:Profiler analysis all required data, and the supplemental tables are available in the GitHub repository for this study: https://github.com/GBIRG/JBT_FunctionalAnnotation.

RESULTS

Datasets considered in this study

A graphical representation of the analysis workflow is presented in Figure S1. This process is initiated with the selection of 4 experimental comparisons, and information about these is presented in Figure 1. These different datasets are representative of some of the different kinds of comparisons encountered by bioinformatics cores. These data are presented in the form of "case studies" that compare different analysis tools in the context of these diverse datasets. Members of our research group have experience with the analysis of these data, and that was leveraged to select a biological concept represented by an MSigDB gene set[22] that is altered in each comparison.

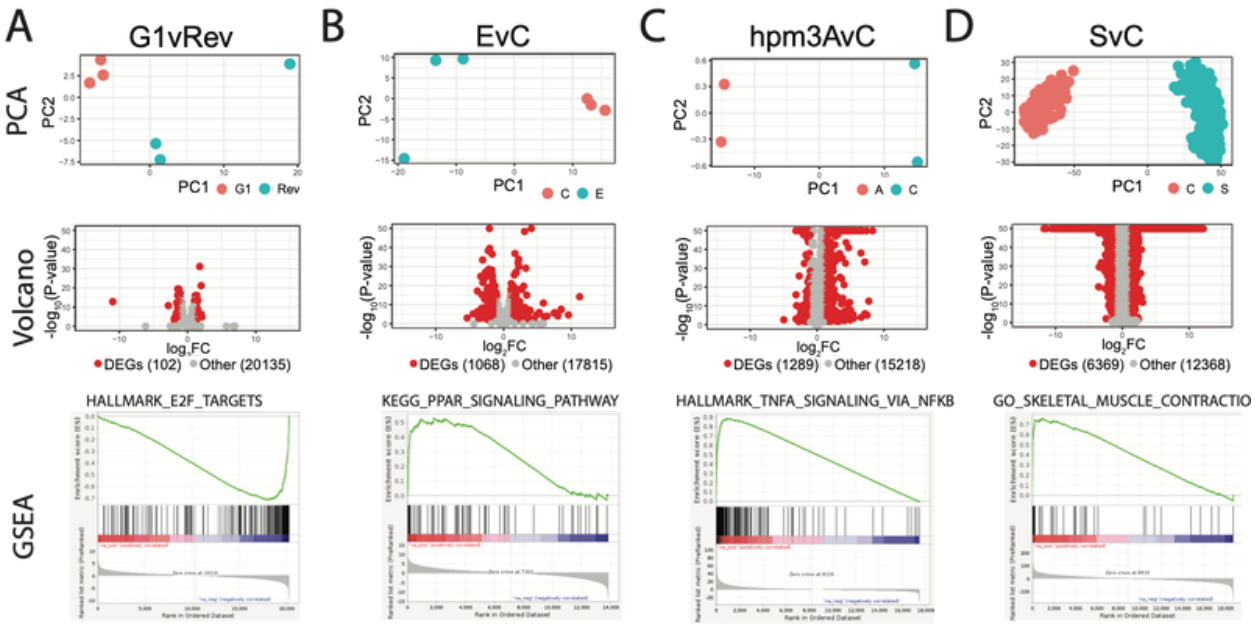


Figure 1

Experimental design, differential expression results, and selected biological results for the datasets considered in this study. The experiments represent a range of experimental designs (top panels) and differential expression results (middle panels, DEGs defined with absolute $\log_2$ fold change ≥2 and padj ≤ 0.05 in red). GSEA was used to associate each comparison with a biological result (bottom panels). (A) G1vRev has 3 replicates of each treatment and 102 DEGs. E2F target genes are enriched in the Rev condition. (B) EvC has 3 replicates of liver samples from mice with different genotypes and fed different diets and 1068 DEGs. Genes in the KEGG PPAR signaling gene set are enriched in the knockout/high-fat diet condition. (C) hpm3AvC as 2 replicates of each condition, principal component analysis (PCA) indicates that the majority of variability that exists in the data is associated with treatment. Genes in the TNF_SIGNALING_VIA_NFKB are enriched in response to asbestos treatment. (D) SvC has 564 skeletal muscle, 303 cardiac muscle replicates, and 6369 DEGs. Genes assigned to the GO_SKELETAL_MUSCLE category are enriched in the skeletal muscle condition.

G1vRev is an experimental manipulation of the human cell line RPE-1.[35] These cells were arrested at the G1/S boundary by thymidine treatment. Following arrest, cells were treated with Reversine to induce aneuploidy, and after 66 hours, RNA was isolated from these aneuploid cells (Rev). In addition, aneuploid cells were treated with Nocodozol to induce mitotic arrest, and detached dividing cells were removed by shaking. This process enriched the culture for senescent attached aneuploid cells (G1) that do not enter the cell cycle. There are 3 replicates for each condition (Figure 1A, top), and 102 DEGs were identified (A, middle), indicating a small degree of differential gene expression between G1 and Rev. The biological concept we considered is genes involved in cell cycle progression. These are more highly expressed in the Rev condition, and the result is represented by the gene set HALLMARK_E2F_TARGETS (Figure 1A, bottom).

The EvC dataset compares livers from Mapk8/9 double knockout mice fed a high-fat diet (E) with livers from wild-type mice fed a standard diet (C). There are 3 replicates of each condition (Figure 1B, top), and 1088 DEGs were identified (Figure 1B, middle), indicating substantial differences in gene expression between E and C. The original study describes the function of Jun terminal kinases (Mapk8 and Mapk9) in the liver and their role in insulin resistance. In this comparison, livers from the knockout mice fed a high-fat diet (E) more highly express genes involved in lipid metabolism. This observation is represented by the gene set KEGG_PPAR_SIGNALING_PATHWAY (Figure 1B, bottom).

Hpm3AvC compares primary human pleural cells treated with asbestos (A) with control-treated cells (C).[36] There are 2 highly correlated replicates for each condition (Figure 1C, top), and 1289 DEGs were identified (Figure 1C, middle), indicating a substantial degree of differential gene expression. Malignant mesothelioma is an aggressive cancer often involving pleural mesothelial cells that have been exposed to asbestos, and this carcinogenesis may be a result of asbestos-induced inflammation. In this comparison, genes involved in inflammation are more highly expressed in the asbestos-treated cells, and this observation is represented by the gene set HALLMARK_TNFA_SIGNALING_VIA_NFKB (Figure 1C, bottom).

SvC is a comparison of gene expression in skeletal (S) and cardiac (C) muscle. There are 564 replicates of skeletal muscle and 303 replicates of left ventricle (Figure 1D, top), and 6369 DEGs were identified (Figure 1D, middle). This highly replicated experiment comparing different types of muscle results in a large number of DEGs. The elevated expression of genes involved in skeletal muscle contraction was noted in the skeletal muscle samples, and this result is represented by the GO_SKELETAL_MUSCLE_CONTRACTION gene set (Figure 1D, bottom).

Analysis tools investigated in this study

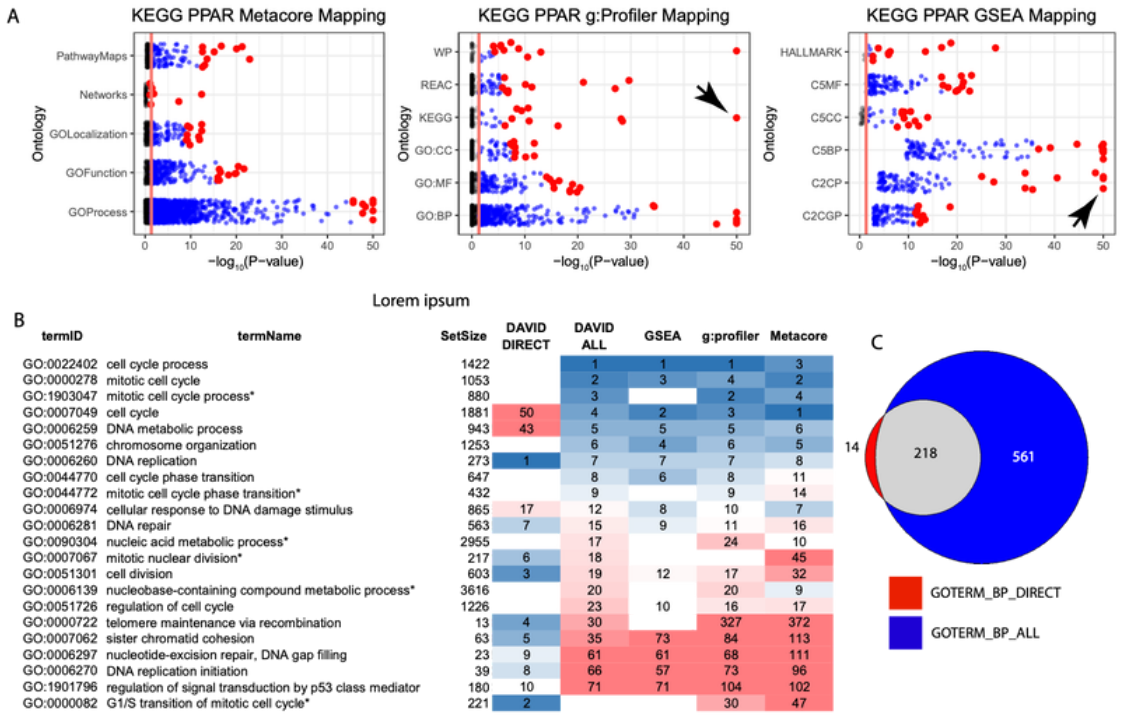
The 6 tools examined here (Table 2) are favored by study participants, and they include representatives from the 3 general algorithmic categories: FCS (GSEA), ORA (DAVID, g:profiler, IPA, Metacore), and TB (iPG). The commercial tools possess extensive functionality, and IPA and Metacore test their own manually curated and proprietary pathway ontologies. iPathwayGuide uses a unique TB analysis scheme that emphasizes highly connected genes in addition to differentially expressed lists. These commercial tools are commonly run interactively and produce heavily annotated results linked to proprietary and public data repositories.

Concept mapping

We sought to compare results produced by the selected tools on the different input datasets but found this process complicated because of inconsistencies across tools in both the naming and content of annotation categories. To facilitate comparisons, we developed a process called concept mapping. For each experiment, we first identified a biological result (concept) and selected an associated MsigDB gene set (Figure 1, GSEA panels, Supplemental Table S3). Next, for each tool, the concept gene sets were used as analysis input, producing the tool-specific results related to each concept. For each annotation collection, up to 10 tool-

specific results with the smallest P-values (plus ties) were annotated as concept-related sets. In order to be selected, the result had to have an adjusted P-value (when available) of 0.05 or smaller. In the case of IPA, which has 2 annotation collections, the top 25 results plus ties were annotated as concept related. This type of list analysis is not possible for GSEA, so instead, the MSigDB Compute Overlaps utility was used. In the case of iPathway guide, the “plus ties” expanded the results list for each concept because this tool appears to place a floor on the P-value, and as a result, many sets had the same small P-value. The concept mapping results for each tool are provided in Supplemental Table S4.

To illustrate the process in more detail, a partial example is shown in Figure 2A. The biological concept we selected for the EvC comparison was KEGG_PPAR_SIGNALING_PATHWAY. The results produced when this gene set is used as input to Metacore (left panel), g:Profiler (middle panel), and GSEA (right panel) are shown. In each panel, points are stratified on the Y-axis according to the ontology group, and random jitter has been introduced for display purposes. Red points indicate the top 10 results for each ontology group that were flagged as “concept-related” for subsequent analyses. The blue points are results that have a P-value < 0.05 but are outside the top 10, and black points are results with a P-value > 0.05. In the case of Metacore, this KEGG pathway is not specifically represented in the tested annotation categories, but numerous closely related gene sets were identified by our analysis, illustrating the utility of our concept mapping approach. By using category names alone, the association between KEGG_PPAR_SIGNALING_PATHWAY and these results would not be apparent. For g:Profiler and GSEA, the KEGG pathway is included and produces a highly significant result (black arrows). These concept-related annotations were identified for each ontology category and tool we consider in this study. In this process, we identify the most closely related results in each of the tools regardless of how the results are named or what genes they contain. We can then use these concept-related terms to consider the performance of the different tools in our experimental comparisons.



termID	termName	SetSize	DAVID DIRECT	DAVID ALL	GSEA	g:profiler	Metacore
GO:0022402	cell cycle process	1422		1	1	1	3
GO:0000278	mitotic cell cycle	1053		2	3	4	2
GO:1903047	mitotic cell cycle process*	880		3		2	4
GO:0007049	cell cycle	1881	50	4	2	3	1
GO:0006259	DNA metabolic process	943	43	5	5	5	6
GO:0051276	chromosome organization	1253		6	4	6	5
GO:0006260	DNA replication	273	1	7	7	7	8
GO:0044770	cell cycle phase transition	647		8	6	8	11
GO:0044772	mitotic cell cycle phase transition*	432		9		9	14
GO:0006974	cellular response to DNA damage stimulus	865	17	12	8	10	7
GO:0006281	DNA repair	563	7	15	9	11	16
GO:0090304	nucleic acid metabolic process*	2955		17		24	10
GO:0007067	mitotic nuclear division*	217	6	18			45
GO:0051301	cell division	603	3	19	12	17	32
GO:0006139	nucleobase-containing compound metabolic process*	3616		20		20	9
GO:0051726	regulation of cell cycle	1226		23	10	16	17
GO:0000722	telomere maintenance via recombination	13	4	30		327	372
GO:0007062	sister chromatid cohesion	63	5	35	73	84	113
GO:0006297	nucleotide-excision repair, DNA gap filling	23	9	61	61	68	111
GO:0006270	DNA replication initiation	39	8	66	57	73	96
GO:1901796	regulation of signal transduction by p53 class mediator	180	10	71	71	104	102
GO:0000082	G1/S transition of mitotic cell cycle*	221	2			30	47

Figure 2

A selected biological concept for each experiment was mapped to related gene sets, pathways, or categories by analyzing concept gene lists with each tool. (A) Demonstration of concept mapping for genes in the KEGG_PPAR_SIGNALING gene set using Metacore (left panel), g:Profiler (middle panel), and GSEA (right panel). For Metacore and g:Profiler, ORA was done using the gene list with default whole-genome background. For GSEA, related terms were identified using the MsigDB compute overlaps web utility. Arrows in g:Profiler and GSEA panels indicate the KEGG PPAR signaling pathway itself. This annotation is missing from Metacore. Red points are the 10 results in each annotation group with the smallest FDR corrected P-values, and blue are hits with P-value ≤ 0.05, but outside the top 10 and black are results with p > 0.05. For GSEA, the maximum number of results reported by the compute overlaps utility is 100, and in most cases, all values have P-values < 0.05. For display purposes, $-\log_{10}$(P-value) corrected P-values > 50 were reported as 50. (B) Analysis of genes in the HALLMARK_E2F_TARGETS gene set using 2 different DAVID annotation collections or GOTERM_BP, GSEA, g:Profiler, and Metacore. Twenty-two different GO categories are in the top 10 in at least 1 of the runs. The rank of each result for each tool is indicated. The results for DAVID ALL are similar to the other 3 tools, but DAVID direct results are distinct. (C) Venn diagram showing overlap between DAVID GOTERM_BP_DIRECT and DAVID GOTERM_BP_ALL terms in the results sets for E2F targets. With the exception of 14 terms in DIRECT that are missing from ALL, the rest are present in both, but they differ dramatically in the number of genes annotated to the term. For example, the term GO:0006281~DNA repair is #7 in direct due to 25/235 genes but #14 in ALL due to 52/530. Term GO:0007049~cell cycle is 4th in ALL with 116/1672 genes but 50th in DIRECT with 12/217.

DAVID allows users to test different subsets of GO terms that are distinguished by rules governing gene assignments to annotation groups. Figure 2B shows a comparison of the GO biological process results

produced by DAVID_DIRECT, which are annotations provided by the annotation source, and DAVID_ALL, which includes a more comprehensive collection of annotations. The table presents 22 GO biological process categories that are present in the top 10 significant results from DAVID, GSEA, g:Profiler, or Metacore. This subset of tools was considered because they are the 4 that specifically analyze the GO biological process annotation group. The size of the annotation category in MsigDB is indicated, and the rank of the result within each run is shown in the cells. Higher ranking results are shaded blue and lower ranking results red. These data show that the DAVID_ALL collection is similar to the GO annotations tested by the other 3 tools that directly test GO term annotations. The DAVID_DIRECT run results in high-ranking results in annotation categories that are generally smaller and perhaps more specific than those produced by DAVID_ALL and the other tools. Figure 2C shows the relationship between all results produced by DAVID_DIRECT and DAVID_ALL runs with the HALLMARK_E2F_TARGETS gene set as input. The detected terms overlap in 218 out of 232 cases. The numbers of genes annotated to each term varies in the DAVID_ALL and DAVID_DIRECT. In some cases, terms are annotated in a similar way (GO:0006297~nucleotide-excision repair, DNA gap filling), with 24 genes assigned to the term in both groups. In other cases, term annotation is dramatically different. For example, the GO:0006259~DNA metabolic process has 995 genes in DAVID_ALL and only 26 in DAVID_DIRECT. These data suggest that the alternative views of the GO presented by DAVID may present different and valuable perspectives on biological results. For this study, the results from DAVID_DIRECT were used to identify terms related to the biological concepts.

Concept-related annotations for each comparison

The DEGs identified in each of the 4 comparisons were analyzed using the 6 tools considered in this study (Figure 3). The 24 panels in this plot are arranged into 6 rows corresponding to tools and 4 columns for each comparison. Statistically significant results, usually defined by adjusted P-values, are shown as points in each panel; the Y-axis is the P-value with more significant results to the left, and random jitter has been introduced for display purposes. The concept-related annotations are highlighted in red. In most cases, a high density of red points is visible on the lefthand side of each panel, indicating that the tools detect annotations related to the biological concept selected for each comparison. The G1vRev comparison is a notable exception to this general conclusion. DAVID, g:Profiler, Metacore, and iPG all lack a high density of red points on the left side of the plots for this comparison. Furthermore, Metacore does not have any concept-related results with a P-value ≤ 0.05. IPA has concept-related results to the left, in part because the P-value included with IPA bulk data downloads is unadjusted, while all other tools report P-values with a false discovery adjustment. GSEA has high-density concept-related results on the lefthand side in each comparison, suggesting that the rank-based FCS algorithm used by GSEA may be more sensitive than the Fisher's test method used by all the other tools. In order to better understand the reasons for the less robust results in the G1vRev comparisons, it is helpful to consider some details about the Fisher's exact test.

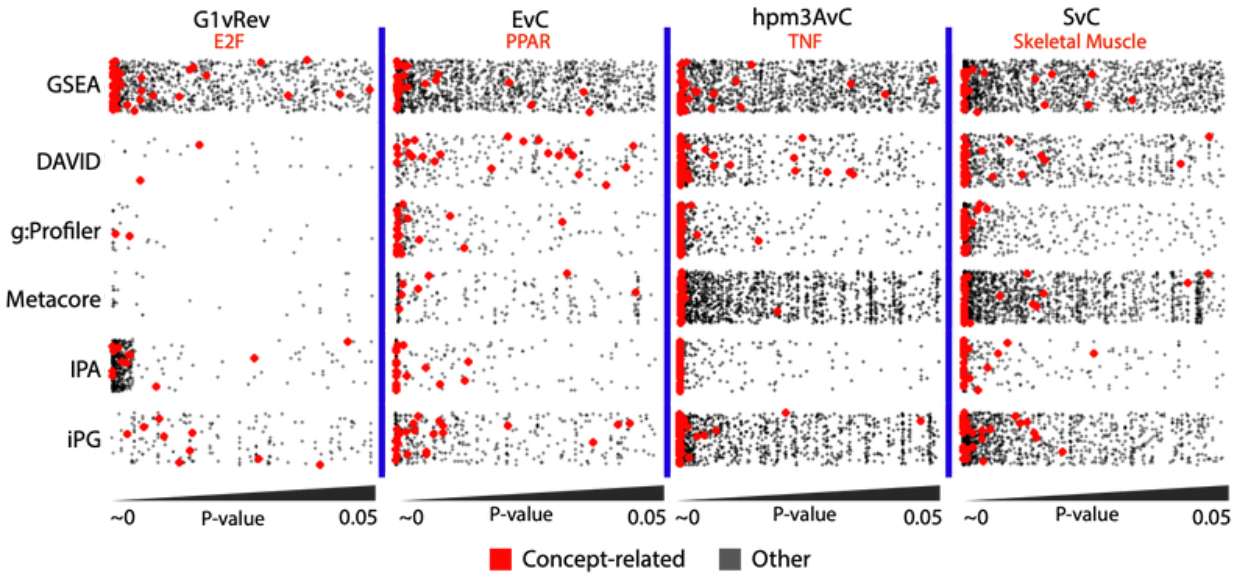


Figure 3

Summary of analysis results. The columns correspond to each experimental comparison, and the rows present data for each tool. Individual panels display results with P-value ≤ 0.05, with lowest P-value results to the left. Adjusted P-values are used in each case except for IPA in which raw P-values are plotted. The vertical jitter in the panels is random and added to separate data points for visualization purposes. Results related to the biological result are highlighted in red.

Overrepresentation analysis and Fisher's exact test

The tools that execute overrepresentation analysis generally make use of hypergeometric testing, and this method has been reviewed elsewhere.[37] The goal of the test is to assess if the number of DEGs in a set is greater than what would be expected to occur by chance given the size of the set, the number of DEGs, and the size of the background. As a result, the 4 input values required for the test are the number of DEGs in the annotation set being tested (Figure 4Aa), the number of DEGs that are not in the set (Figure 4Ab), the number of non-DEGs in the tested set (Figure 4Ac), and the number of non-set, non-DEGs, also known as background (Fig 4Ad). Consider the example in which 4 out of 100 DEGs are in a pathway that has 200 genes from a genome with 20000 genes. The test aims to allow researchers to conclude that the 4% differentially expressed pathway genes is a greater fraction than would be expected by chance given that 1% of the genome is in the pathway. In this example, the Fisher test produces a P-value of 0.021, suggesting that the pathway is overrepresented in the DEG list.

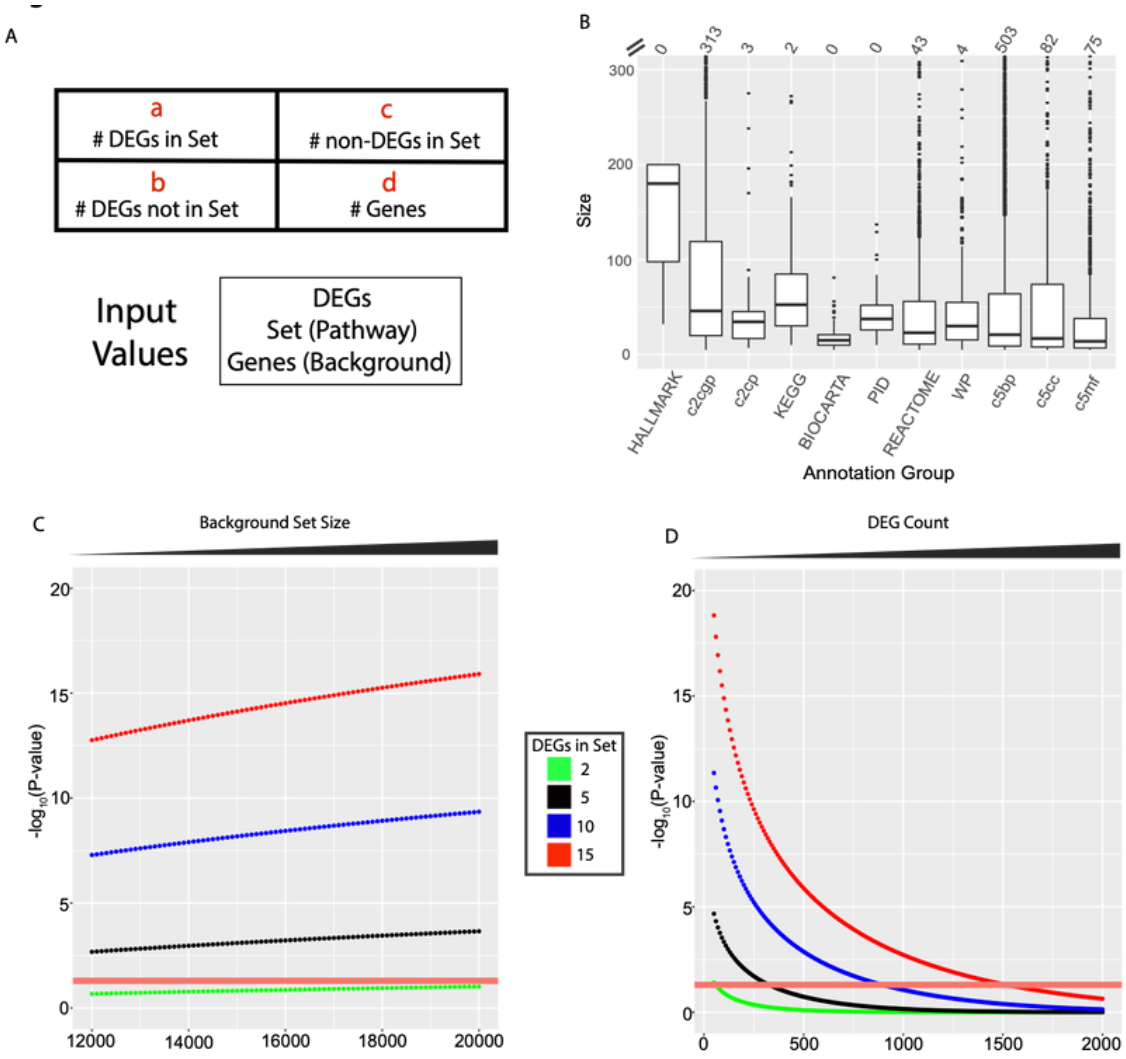


Figure 4

Consideration of input data used in the Fisher's exact tests. (A) The 2x2 contingency table contains input values (red letters) that correspond to counts of DEGs in the gene set being analyzed (a), counts of genes in set that are not differentially expressed (b), count of DEGs not in set (c), and non-differentially expressed/non-set background count (d). (B) Size analysis of the 16550 gene sets in the indicated collections of MsigDB v7.2. The count of outliers > 300 genes in size is indicated along the top of the plot. The overall average set size is 84 genes. (C) Fisher's exact tests using simulated data for a gene set of size 100 in which the number of DEGs in the set are 2 (green), 5 (blue), 10 (black), and 15 (red), and the background set size is varied from 12000 to 20000 genes. (D) Fisher's exact tests using simulated data for a gene set of size 100, and the number of DEGs in the set are 2 (green), 5 (blue), 10 (black), and 15 (red). A background set size of 17000 and the total number of DEGs is varied from 50 to 2000. In both C and D, the horizontal line indicated the $-\log_{10}$(P-value) threshold for significance of 1.3; values above this line have a P-value < 0.05.

The size of the annotation sets being considered is the only input parameter that is not dependent on experimental considerations. The box plot in Figure 4B presents the sizes of the gene sets in MsigDB stratified

by annotation groups. Most gene sets are less than 100 genes in size. The hallmark collection gene sets average about 200 genes, and there are several hundred gene sets in the c2 chemical and genetic perturbation (c2cgp) and GO biological process (c5bp) sets with greater than 300 genes. The remaining 3 input values to Fisher's test need to be defined by user decisions, and we noted a lack of consensus within our group regarding this process. As a result, we executed a series of simulations in order to inform a consensus about these decisions.

The definition of the background set was particularly controversial. It is possible that this should be the total number of genes encoded by the genome because all had an equal chance of being detected in an RNA-Seq experiment. Depending on the library preparation method, it could be best to include all genes with polyadenylated transcripts as represented by the protein coding gene list. Alternatively, it might be best to only consider genes that are detected in the experiment. To investigate this parameter, we tested a range of background set sizes in increments of 100, starting with 12000 to represent expressed genes and ending with 20000 to represent all protein coding genes (Figure 4C). The annotation set size was fixed at 100 genes, and tests were run with 2 (green), 5 (black), 10 (blue), and 15 (red) differentially expressed pathway genes out of a set out of 100 DEGs. In each case, the $-\log_{10}$(P-value) increases as the size of the background set increases, indicating that statistically significant results become easier to obtain as background set size increases. The curves for 5, 10, and 15 pathway DEGs never drop below the significance threshold, indicated by the horizontal red line at $-\log_{10}$(P-value) 1.3, indicating that these results are all significant across this range of background set sizes. The data for 2 pathway DEGs are never significant. This suggests that, although the upward trend exists, the slopes of the lines are generally subtle, and small adjustments to the background size, by excluding DEGs from the background set for example, are unlikely to have a consequential impact on the results.

In addition to background size, the thresholds that define DEGs also need to be selected by the researcher. A standard for these thresholds is an absolute $\log_2$ fold change ≥ 1 and an adjusted P-value ≤ 0.05, but these are somewhat arbitrary values that can be modified to adjust the stringency of the analyses. Another factor impacting the number of DEGs is the directionality of differential expression. It may be desirable to include both up- and downregulated genes in the counts because genes may be differentially regulated in the same biological pathway. That said, it could also be appropriate to consider the up- and downregulated genes separately. An example is a set of genes that are markers of a differentiation event, and all have higher expression in 1 of the phenotypic classes. Decisions about these parameters can impact both the count of DEGs and the numbers of genes found within the annotation set under analysis. Figure 4D is an analysis of these values. In this plot, a range of DEG counts from 50 to 2000 are tested for an annotation set with 100 genes and 2 (green), 5 (black), 10 (blue), and 15 (red) DEGs from that set using a background size of 17000. These curves are steep, and they often cross the horizontal significance threshold line in the middle range of DEG counts. For example, the blue line representing 10 DEGs in a set of 100 drops below significance near a DEG count of 1000. If all 10 of those genes were upregulated and differential expression was evenly distributed between the up and down direction, a biological conclusion might be made if only upregulated genes were

considered (10/500) compared to both directions simultaneously (10/1000). These data suggest that careful consideration of the definition of DEGs is important when running these analyses. It is notable that the Metacore web interface facilitates easy adjustment of differential expression thresholds, making this tool an excellent choice for the exploration of these parameters.

G1vRev and genes involved in the cell cycle

Our G1vRev comparison is the only one of our comparisons that does not result in numerous highly significant concept-related results for all 6 of the tools we tested (Figure 3). This comparison is a complex biological manipulation resulting in relatively subtle differences between conditions with only 102 DEGs using the standard cutoffs. Of the gene sets identified as related to the HALLMARK_E2F_SIGNALING concept, only 2 are significant in the DAVID and g:Profiler analyses, and none are significant using the Metacore tool. Therefore, analysis may benefit from the relaxation of fold change thresholds and the consideration of genes upregulated in only 1 of the conditions. These modifications to the selection of DEGs can be justified by careful consideration of the experiment.

The Rev condition is a mixed population consisting of about 40% arrested aneuploid cells and 60% aneuploid cells in various stages of the cell cycle. This cycling population is asynchronous, and it is estimated that about half of the cells are in the G1 phase of the cell cycle. To prepare the G1 condition, a repeated shake-off procedure was used following a Nocodozol block of mitosis. This takes advantage of the physical detachment of mitotic cells, allowing them to be removed from the culture, leaving only the arrested subpopulation that is not progressing through the cell cycle. Therefore, the G1 sample is a subpopulation of the Rev condition, resulting in substantial similarities between the conditions in this bulk RNA-Seq experiment. This, coupled with unavoidable variability driven by degree of aneuploidy, relative cell-state proportions, and other factors, results in divergent biological replicates and small numbers of DEGs. It is possible that genes with biologically relevant fold changes could fall below the standard thresholds, and in this case, they may be concentrated in just the Rev condition enriched for cycling cells.

To investigate this in more detail, we performed an analysis of this experiment using the modified differential expression selection criteria detailed in Figure 5A. The standard thresholds that result in selection of 102 genes are shown in panel a. Panel b shows 466 DEGs using a relaxed $\log_2$ fold change of 0.58 (1.5 fold in linear space). Panel c highlights a selection of 62 genes higher in the Rev condition (negative fold change) only using the standard fold change threshold. Panel d shows the 268 genes in this direction with the relaxed fold change threshold. Figure 5B shows the position of 26 HALLMARK_E2F_TARGETS target genes that have an absolute fold change > 0.4 in this comparison in a volcano plot. Twenty-three of these genes are higher in the Rev condition, but the fold changes are small, with only 5 genes meeting the standard $\log_2$ fold change threshold; all 5 of these are higher in the Rev condition.

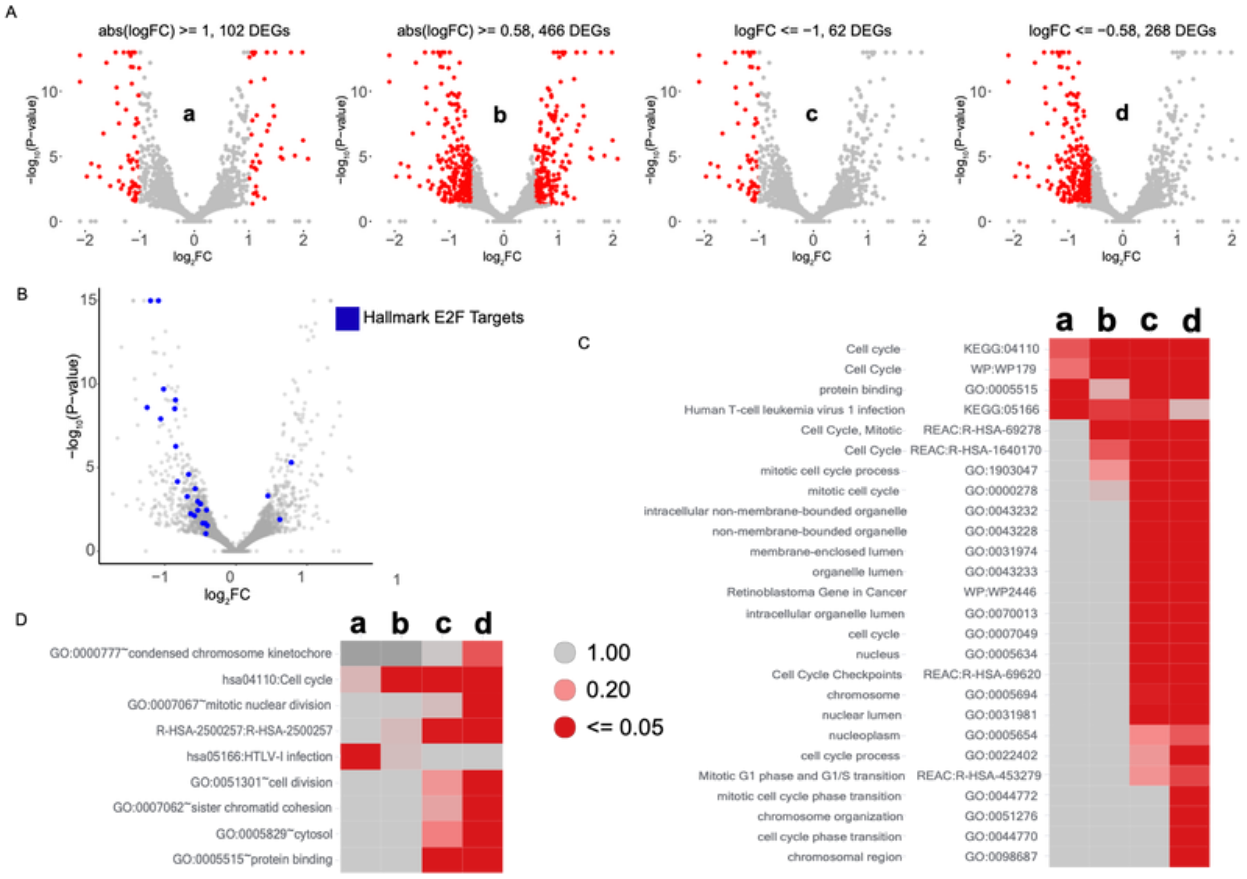


Figure 5

Cell cycle analysis of G1vRev experiment. (A) Differential expression analysis filtering thresholds used in this analysis. (B) Volcano plot highlighting differential expression of the 26 HALLMARK_E2F_TARGETS genes (blue points) with an absolute fold change > 0.4. (C) DAVID analysis of gene lists from each of the 4 differential expression sets. The 9 sets presented were identified as related terms in the DAVID concept mapping results and have a Bonferroni-adjusted P-value < 0.2 in at least one of the 4 tests. (D) g:Profiler analysis of gene lists from each of the 4 differential expression sets. The 26 sets presented were identified as related terms in the g:Profiler concept mapping results and have an adjusted P-value < 0.2 in at least 1 of the 4 tests. For both C and D, rows are ordered by sum of P-value from the 4 runs, so that sets most likely to be significant in all tests are presented on top.

The DEG lists prepared using different criteria were then analyzed with DAVID (Figure 5C) and g:Profiler (Figure 5D). Each heat map presents results for concept-related annotation categories HALLMARK_E2F_TARGETS concept mapping analyses that have a P-value < 0.2 in at least 1 of the analyses. The P-values used here are Bonferroni for DAVID and the g:SCS method for g:Profiler. The columns (a-d) correspond to the differential expression selections indicated in Figure 5A a to d, and colors are according to the indicated scale with ≤ 0.05 in dark red and values between 0.05 and 1 on a gradient from red to gray. These data show that for both of these tools, while relaxing the fold change threshold to 0.58 increases the number of categories detected as significant (a versus b and c versus d), the most substantial increase in numbers of concept-related results detected is observed when genes upregulated in the Rev condition are

considered separately (a versus c and b versus d). This is consistent with the presence of cycling cells in the Rev condition and suggests that an analysis of this experiment may benefit from using only genes higher in Rev because the majority of the cell cycle transcriptional signature is associated with this condition.

The HALLMARK_E2F_TARGETS gene set selected as the biological concept of interest for this experiment consists of about 200 genes that are targets of the E2F family of transcription factors. These genes are often involved in cell cycle–related processes, and the enrichment of this gene set is significant by GSEA (Figure 1A), with a normalized enrichment score of -3.03 and a nominal P-value of approximately 0. In this analysis, over half of the genes in this set were detected in the GSEA leading edge, indicating that, despite the relatively low fold changes, this gene set is robustly enriched in the Rev condition. This result highlights an aspect of GSEA that sets it apart from the other tools. Gene sets are scored based on relative rank in a gene list ordered from high in one condition to high in another, and users are not required to set fold change and P-value thresholds in order to analyze an experiment. In this work, the ranking metric used to order genes from high in one condition to high in the other was the Wald statistic, log2 fold change divided by standard error, and is related to fold change, so GSEA when run with this ranking metric considers genes upregulated in each condition separately.

EvC and PPAR signaling

The PPAR signaling pathway was found to be upregulated livers from Mapk8/Mapk9 double knockout mice fed a high-fat diet compared to livers from wild-type mice fed a standard diet. As a result, we selected the KEGG_PPAR_SIGNALING gene set as the biological concept for the EvC comparison (Figure 1B). This gene set was significantly enriched in the E condition, with a NES of 1.99 and a nominal P-value of approximately 0. This KEGG pathway is tested by iPG, GSEA, DAVID, and g:profiler, and it was significantly detected by all 4 (Figure 6A, red points). Although the manually curated Canonical Pathways in IPA and Pathway Maps in Metacore lack this specific pathway, our concept mapping approach identified a series of closely related annotation categories within the tested sets for each of these tools, and these related annotation sets are significant in our analysis (Figure 3). One of the most closely related Pathway Maps in Metacore is “Regulation of lipid metabolism_PPAR regulation of lipid metabolism.” This annotation category has 47 genes, and 16 of them are in KEGG_PPAR_SIGNALING. One of the top Canonical Pathways in IPA is “Fatty Acid **β**-oxidation I.” This consists of 33 genes, with 15 in KEGG_PPAR_SIGNALING. These results and others highlight the linkage between the KEGG_PPAR_SIGNALING gene set and pathways and functions involving lipids.

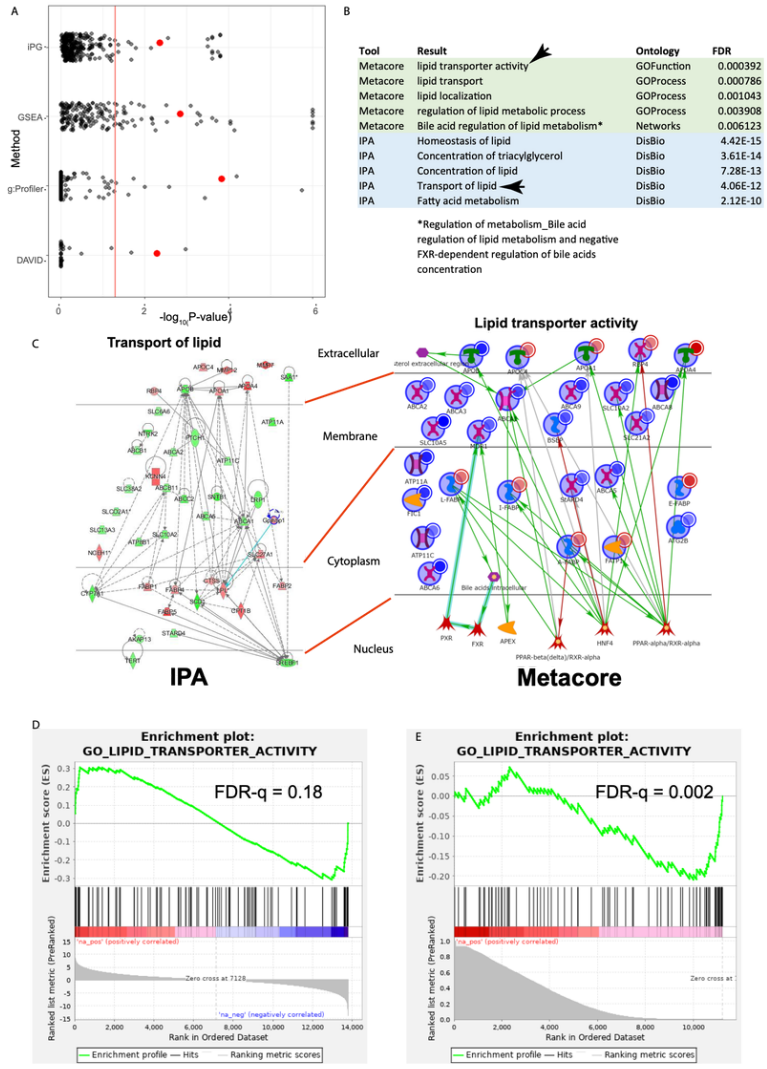


Figure 6

PPAR signaling and lipid metabolism analysis in the EvC experiment. (A) iPG, GSEA (canonical pathways collection), g:Profiler, and DAVID include the KEGG PPAR signaling pathway in the tested annotation, and in each case (red point) this result is significant in this analysis. P-values plotted were fdr_p for iPG, FDR_q_value for GSEA, Bonferroni DAVID, and g:SCS for g:Profiler. Vertical red line indicates the -$\log_{10}$(P-value) 1.3 significance threshold, and points to the right of the line have P-values < 0.05. Vertical jitter added to strip plots for display purposes. (B) Metacore and IPA do not test the KEGG PPAR pathway, but in concept mapping analysis, a series of lipid metabolism pathways were identified as related; the top 5 results for each of these are displayed. (C) IPA visualization of 43/174 genes. (D) Metacore visualization of 27/154 network. (E) GSEA plot of GO Lipid Transporter activity using Wald statistic as ranking metric and weighted scoring scheme. (F) GSEA plot of GO Lipid Transporter activity using adjusted P-value as ranking metric and classic scoring scheme.

One of the top GO molecular function results for the E versus C comparison in iPG is the term lipid transporter activity. This term is also a top result in the Metacore analysis of GO molecular function (Figure 6B). A related

term from IPA analysis of their “Diseases and Bio Functions” collection called “transport of lipid” also scores highly. These 2 results can be presented using the network visualization function available in these tools (Figure 6C). These types of visualization are also available in iPG, and similar plots can be prepared with open-source tools. In these examples, the interactions between the genes in each annotation category and subcellular localizations are indicated. In addition, fold change data from the EvC comparison are shown using colored shadings or icons associated with each gene. The accessibility and quality of these types of interactive visualizations are a distinguishing feature of the proprietary applications. Information about each gene and interaction between genes is hyperlinked in the visualization and are easily accessible. These data can be difficult to present in a static format, and for even these relatively small annotation categories, the visualization of individual gene names is challenging. If the focus is on general trends, these views can be informative. For example, in Figure 6C, it is clear that the pathways from both IPA and Metacore contain genes that are up- and downregulated in this comparison due to the green and red shading of genes in the IPA plot and the blue and red icons associated with genes in the Metacore plot.

This differential regulation of genes in these pathways highlights an aspect of GSEA analysis that can be adjusted for a different view of an experiment. The molecular function GO term in the Metacore analysis (lipid transporter activity) is captured in an MsigDB gene set, and this is not significantly enriched in the E condition by GSEA. Genes in this category are both up- and downregulated in this experiment, but pre-ranked GSEA using the Wald statistic as a ranking metric considers the different directions separately. As a result, neither direction reaches significance (Figure 6D). In cases like this, it may be beneficial to run pre-ranked GSEA using the adjusted P-value as a ranking metric. This tests for enrichment in genes that are different as opposed to up or down genes. In this case, the ranking metric is not related to the magnitude and direction of change, so the “classic” GSEA scoring scheme should be used instead of “weighted.” Figure 6E shows a GSEA plot using adjusted P-value as a ranking metric. The righthand side of the continuum are genes with low-adjusted P-value genes, and the GO_LIPID_TRANSPORTER_ACTIVITY set is significantly enriched on this side of the plot. Unlike runs ranked by the Wald statistic, the left side of the continuum are genes with P-values near 1, indicating a lack of differential expression, and the results here are difficult to interpret.

Hpm3AvC and IL-17 signaling

Treating human primary pleural cells with asbestos results in a robust inflammatory response. We selected the gene set HALLMARK_TNFA_SIGNALING_VIA_NFKB as the representative biological concept for this comparison (Figure 1C) in which this gene set is enriched in the A condition with a NES of 2.04 and a P-value of approximately 0. The concept mapping of this gene set revealed a strong relationship with terms relating to the IL-17 signaling pathway in 4/6 of the analysis tools. IL-17 signaling is represented by an 81 gene KEGG pathway called “IL-17 signaling pathway” that is directly tested by g:Profiler and iPG. Versions of this pathway are tested by Metacore and IPA result, but these annotations differ in composition. The Metacore version called “immune response_IL-17 signaling pathway” has 39 genes, while the IPA pathway “IL-17

Signaling" has 132 genes. In contrast, DAVID and GSEA lack IL-17 annotation sets that were identified by our concept mapping procedure (Figure 7A). GSEA and DAVID test a 15-gene Biocarta version of the IL-17 pathway, and GSEA also includes a Wikipathway version with 32 genes. Neither of these were flagged as related to HALLMARK_TNFA_SIGNALING_VIA_NFKB by our concept mapping procedure. Figure 7B shows analysis results for AvC with the 6 tools under consideration; red points indicate the 4 IL-17 signaling sets listed in Figure 7A, and blue points are other annotations that have some version of the text pattern "IL-17." The representations of the IL-17 signaling are highly significant, while other IL-17 annotation groups are less so. These different representations and notable absences of a biological process as well studied as IL-17 signaling highlight an important difference between the analysis tools. While these tools generally produce results that lead to similar biological interpretations of our comparisons (Figure 3), the annotation groups being tested by each can have substantial differences.

A

Tool	SetName	Ontology	Size	Rank in Ontology	P-value
Metacore	Immune response_IL-17 signaling pathways	PathwayMaps	39	6	5.06E-16
IPA	IL-17 Signaling	Canonical Pathways	132	4	1.26E-18
gProfiler	IL-17 signaling pathway	KEGG	81	4	7.09E-16
iPG	IL-17 signaling pathway	KEGG	81	6	1.92E-06
GSEA	N/A	N/A	N/A	N/A	N/A
DAVID	N/A	N/A	N/A	N/A	N/A

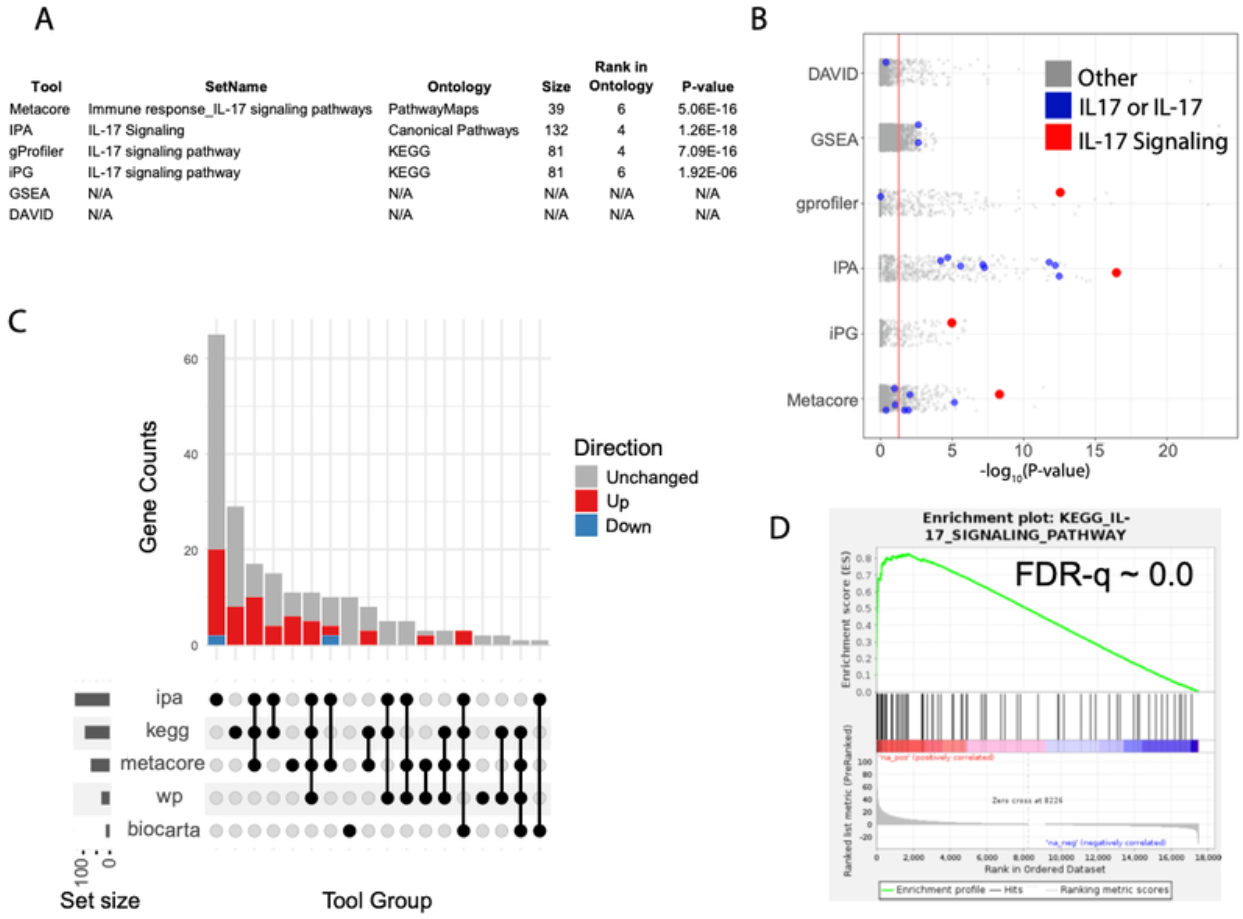


Figure 7

IL-17 signaling analysis in the hpm3AvC experiment. (A) Source and size of IL-17 signaling pathway annotations in study tools and the rank of the result in the TNF concept mapping analysis with each tool. (B) Results of IL-17 signaling (red points) and 19 sets whose names contain the IL-17 string (blue points). Vertical red line indicates the -$\log_{10}$(P-value) 1.3 significance threshold, and points to the right of the line have P-values < 0.05. (C) UpSet plot displaying the distribution of 201 different genes in the various IL-17 signaling sets. Of these genes, 61 are upregulated by asbestos treatment (red segments), 4 are downregulated (blue segments), and 136 do not meet differential expression thresholds (gray segments). Counts of genes in different annotation groups are indicated. (D) GSEA result for the KEGG IL-17 signaling pathway run as a custom set.

We wished to examine the heterogeneity of genes present across the different IL-17 signaling representations in the context of AvC differential expression. We identified 201 different genes present in at least 1 of the IPA, KEGG, Metacore, Biocarta, or Wikipathways IL-17 signaling annotation sets and summarized these results in an UpSet plot (Figure 7C). The lower-left plot shows the number of genes measured for each representation (row). The center matrix and upper bar plot show the number of genes considered across representations; each row represents a representation, and each column represents a set of genes. At the row-column coordinate, a gray node indicates this subset of genes was not measured by this representation, and a black node indicates

this protein set was measured; black nodes are vertically connected by intersection lines. The union of all representations contains a set of 201 distinct genes; at 132 genes, IPA's was the most inclusive gene set, but as the first column shows, 65 of 132 are only present in the IPA annotation. Each column is shaded according to whether the AvC differential expression was upregulated (red segment, 18 genes), downregulated (blue, 2), or not differentially expressed (gray, 45). The second largest group contains 29 genes unique to the KEGG pathway. Overall, almost half (94/201) of the genes included in these IL-17 annotations are found in these 2 non-overlapping groups, a suggestion indicating variable composition of these annotations. We note that, although there are no genes found in all 5 of these annotation categories, 84 genes are found in at least 2. In fact, the 3rd largest group is a set of 17 genes that are found in IPA, KEGG, and Metacore, and 10 of these genes are upregulated in response to asbestos. There are 3 genes in common between IPA, KEGG, Biocarta, and Metacore (CSF3, CXCL8, and IL6), and all 3 are upregulated in response to asbestos treatment. Given the discrepancies present in the composition of annotation categories, it may be beneficial to prepare custom annotation sets that may be missing from the collection being tested by a tool. This is trivial to accomplish in GSEA, in which custom gene sets can be created in gmx or gmt format and tested. Figure 7D shows one such result for a manually created gene set capturing the KEGG IL-17 signaling pathway and the AvC comparison. g:Profiler includes a function that allows the upload of custom gene sets in GMT format. The testing of custom gene sets by the other tools does not seem to be supported.

Skeletal muscle versus heart left ventricle

The comparison between the GTEx Skeletal Muscle (S) and Heart, Left Ventricle (C) involves a large number of biological replicates of 2 dramatically different tissues. A differential expression analysis results in a large number of DEGs, with 2488 genes more highly expressed in skeletal muscle and 3881 more highly expressed in cardiac muscle using standard thresholds. An analysis of this number of DEGs is difficult. IPA issues a warning when analysis is attempted with this number of genes and suggests filtering so that there are between 100 and 2000 DEGs under consideration. The DAVID R interface also encounters timeout failures when working with lists of this size. Although the other tools in our analysis do not provide warnings, it is likely that an analysis of large gene lists is problematic. For example, in a Metacore analysis of Biological Process with the full set of DEGs, all of the top 10 results were from general GO categories consisting of more than 2500 genes. These large and general categories are difficult to interpret. To investigate this observation further, we modeled Fisher’s exact test results for an annotation category with 100 genes in which 20, 40, 60, and 80 of those genes are detected in a range of DEG counts from 1000 to 6000 (Figure 8A). The majority of annotation sets contain fewer than 100 genes, so this sizing is relevant (Figure 4A). When 20 genes are differentially expressed out of 100 gene sets, statistical significance can be achieved only when there are fewer than about 2100 DEGs (Figure 8A, green curve above red horizontal line). For 40/100, significance cannot be achieved when there are more than about 4500 DEGs (Figure 8A, black line). For 60 and 80 genes in a pathway, significance is possible throughout this range of differentially expressed gene counts (Figure 8A, blue and red lines, respectively). These curves shift to the right for larger annotation sets; 200/1000 has a significance

threshold at about 3000 DEGS, and for 400/1000, the threshold is around 5500 genes (Figure 8B). These data indicate that when large DEG lists are under consideration, substantial fractions of genes in an annotation set must be differentially expressed in order to achieve low P-values, and larger annotation sets are more likely to be detected.

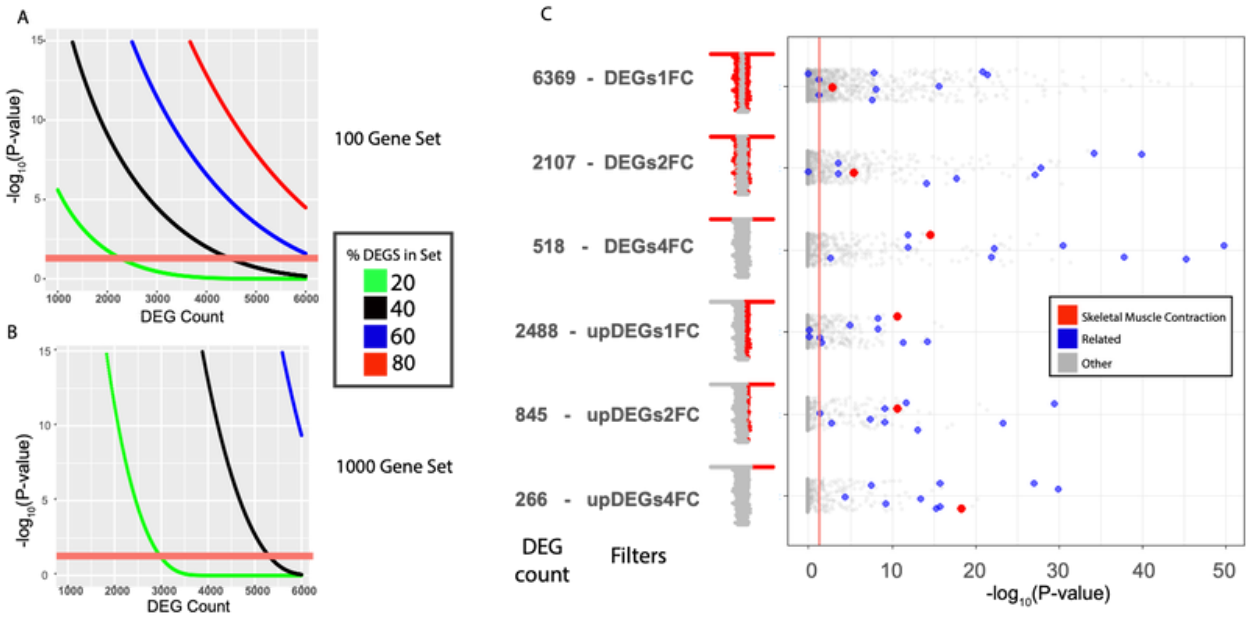


Figure 8

GO skeletal muscle analysis in the SvC experiment. (A) Simulated Fisher's exact tests for a 100 gene pathway when 20% (green), 40% (black), 60% (blue), and 80% (red) pathway genes are differentially expressed in the context of increasing numbers of DEGs ranging from 1000 to 6000. The horizontal line indicated the $-\log_{10}$(P-value) threshold for a significance of 1.3, and values above this line have P-value < 0.05. (B) The same as A except a 1000 gene annotation set is tested. (C) Significance of the GO skeletal muscle (red points) category and related terms (blue points) and other (grey points) under different filtering schemes using g:Profiler. Numbers of DEGs for each filter are indicated. The red/grey volcano plots indicated the filtering strategy. Vertical red line indicates the $-\log_{10}$(P-value) 1.3 significance threshold, and points to the right of the line have P-values < 0.05. Vertical jitter added to strip plots for display purposes.

Based on these results, it is likely that researchers must heed the IPA warning and focus on only the most robustly different genes when doing analysis. Figure 8C presents a g:Profiler analysis of the GO biological process ontology using different gene lists from the SvC comparison. The top 3 rows consider DEGs in both directions with absolute $\log_2$ fold change thresholds of 1, 2, and 4. The bottom 3 rows consider only genes more highly expressed in skeletal muscle with $\log_2$ fold change thresholds greater than 1, 2, and 4. The numbers of DEGs under consideration are indicated, and volcano plots highlight the thresholds in use. In all cases, selected DEGs had an adjusted P-value less than 0.05. The GO biological process category Skeletal Muscle Contraction was selected as the biological concept for this comparison (Figure 1D) and is plotted in

red; concept-related terms are in blue, and other terms are in gray. These data show that although the concept-related term is significant in all tests, similarly small P-values are obtained when a fold change threshold greater than 4 is applied or when only genes more highly expressed in skeletal muscle are considered. These results suggest that when the number of DEGs exceeds tool recommendations, stringent differential expression thresholds can be applied and still produce appropriate biological interpretations.

DISCUSSION

In this study, we present general agreement in the functional annotation results produced by 6 tools using 3 different algorithms (ORA, FCS, and TB) run on 4 datasets. In nearly all cases, the representative biological conclusion was readily apparent in the output. While our study was focused on practical parameters that should be considered when executing these analyses, the performance we observed is consistent with directed and large-scale evaluations that have been published. Tarca et al., 2013 reported a significant enrichment of target gene sets in 36 of 42 experiments examined by almost all of 16 tools they evaluated.[38] Geistlinger et al., 2021 describe a comprehensive framework for evaluating enrichment analysis methods and tools, discuss the statistical properties of these approaches, and apply their framework to 42 microarray and 15 RNA-Seq experiments. These authors observed general consistency in phenotypic relevance scores for the class of methods we consider in our study.[39] Ihnatova et al., 2018 performed an analysis of different topology-based analysis methods and described consistent detection of enriched pathways when sufficient numbers of pathway genes are differentially expressed. This study also includes an evaluation of various experimental designs with a range of DEG count, and suggestions regarding optimal tool selection under these different experimental parameters are provided.[40]

In our study, the exception to the similar performance profiles was observed in the G1vRev dataset. This is a comparison between 2 biologically similar conditions with substantial variability in replicates, and only about 100 DEGs are detected using default thresholds. Using ORA, the biological result of cell cycle is most readily detected when relaxed fold change thresholds are applied, and direction is considered separately. On the other hand, the biological result for our EvC comparison involves a lipid transport pathway that contains both up- and downregulated genes in response to the experimental manipulation. Here, consideration of the 2 groups of DEGs separately based on direction is counterproductive. These observations suggest that the application of a constant set of thresholds to ORA analyses may not be the most effective approach. Instead, analysts should experiment with these parameters and arrive at an approach guided by experimental design considerations.

An analysis of our G1vRev dataset also emphasized an important characteristic of GSEA. This FCS method does not require the application of thresholds to define DEGS; instead, it uses a list of genes ordered by some sensible differential expression statistic. As a result, the cell cycle biological concept was robustly detected without adjustment of the DEG thresholds. GSEA also performs well in comparisons that have large numbers of DEGs. When default thresholds are applied to our comparison of skeletal and cardiac muscle, more than 6000 DEGs are detected. We show that the majority of annotation categories are difficult to detect in situations

like these and that large and general annotation categories are more likely to achieve significance. Notably, strict fold change thresholds can be applied to reduce the number of DEGs without impacting detection of the relevant biological concept. GSEA, however, performs well in this comparison without adjustment of the thresholds used to define differential expression. This, combined with the G1vRev results, suggest that while careful adjustment of DEG thresholds may be required to obtain optimal ORA results, GSEA can often be executed with consistent parameters for a wide range of experimental comparisons.

Another component of ORA approaches is the definition of a background set. We show that background set sizes are unlikely to have a major impact on analysis results. In general, statistical significance is easier to achieve with larger background set sizes, so it may be beneficial to use the largest justifiable value for this parameter. Decisions about background set size should be carefully considered in order to minimize technology, detection, and biological biases.[41] Finally, the maintainer of the TB approach used by iPathwayGuide suggests that their tool makes use of biological information associated with the genes in an input list so that DEG lists containing genes with central and highly connected roles in pathways will be emphasized, and these impactful nodes will factor more prominently in the results.

Each tool considered in this study tests representations of gene categories and pathways from their own collection of annotations. In cases such as GO and KEGG, the underlying information is shared, and although it may be cumbersome to track specific versions in use, these annotations are probably stable and consistent between tools over time. Ontologies such as the IPA canonical pathways and the Metacore Pathway maps are manually curated, specific to the tool, and may represent valuable features of commercial software. We note that creating and testing custom gene lists that represent annotation categories that are missing, underappreciated, or relevant to a specific experiment is simple to accomplish with GSEA.

The differences in annotation collections and the naming conventions used by each tool made direct comparisons between the output of tools challenging. Our concept mapping approach attempted to address these challenges by identifying a gene set that represents a known biological result for each dataset and then using that gene set as an input to each tool in order to flag annotations within each tool that are related to the result. The necessity and utility of this process is illustrated by the AvC comparison. In this experiment, the selected biological concept was a MSigDB hallmark gene set called TNFA_SIGNALLING_VIA_NFKB. The concept mapping of these genes identified IL-17 signaling pathways as a response to asbestos treatment. Despite the fact that IL17 was characterized at the molecular level nearly 30 years ago,[42] a variety of representations of this pathway named in this fashion are present in the repositories. This may be caused by the numerous IL17 family members with a correspondingly complex array of receptors that have many biological roles in different kinds of inflammation.

The tools in this study are accessed with web, command line, or programmatic interfaces. For GSEA, our preferred method uses the Java command-line interface, but we have experience with the Java GUI and note that this mode is most accessible to our bioinformatics core clients. The GSEA/MSigDB system is well

supported and maintained, and full output from the runs is accessible in both text and html formats. The web interface to DAVID has been used by many of us over time. DAVID also supports programmatic analysis with Java, Perl, Python, and R clients;[43] however, there are restrictions on gene list sizes that may prevent the analysis of some datasets. The web interface to g:Profiler is easy to use, and the authors also provide a robust R client, making it accessible to both novice and experienced users.

The commercial tools Metacore, IPA, and iPathwayGuide all provide web interfaces with extensive analysis and visualization functionality. These commercial interfaces are highly interactive and produce optimal results when used in that manner. The resulting data can be explored by following hyperlinks and making adjustments to visualizations. Users can launch subsequent analyses to build networks or pathways or explore interacting molecules. We have presented only a subset of the functions provided by the commercial tools, and in aggregate, it is likely that these tools provide enough additional value over open-source methods to justify their expense. The licensing options for each commercial tool in our study vary but can be in the range of thousands of dollars. These licenses or accounts are often linked to a single user, making shared access difficult, so interactive use may require a designee from each core. It is noteworthy that iPathwayGuide provides the valuable ability to share interactive analysis results with unlicensed collaborators. During the course of this work, we had contact with technical support from all 6 tools and found them to be responsive and helpful, making all of them well-maintained and supported analytical options.

Representatives from the commercial tools were invited to participate in the study and were provided with the differential expression datasets. Our intention was to have companies work with their own tools to provide a “best case scenario” result, while the processing by GBIRG scientists would be a “real world scenario.” Evolving analytical goals and uneven participation led us to deemphasize this aspect of the work, but both Advaita (iPathwayGuide) and Clarivate Cortellis (Metacore) took advantage of the opportunity to examine the data, and results were consistent with those produced by GBIRG participants. We note that interaction with expert company representatives was educational, suggesting that the availability of this kind of support is another valuable aspect of the commercial tools.

The analyses presented here suggest that functional annotation of gene lists derived from genomics experiments is an artisanal process, with little evidence of a single optimal solution. Practically equivalent results can be obtained by all the tools we considered. We conclude that optimal analysis routines will consist of combined ORA and FCS methods, and the identification of high-priority results may be facilitated by TB methods. Furthermore, we feel that combining open-source tools with commercial products such as Metacore, IPA, or iPathway guide will leverage the available range of algorithmic and annotation resources and allow for a careful examination of biological annotations and pathways impacted by an experiment, leading to informative biological conclusions and productive hypothesis generation. We have sought to develop a novel framework evaluating functional annotation tool performance across a variety of analysis tools that is robust to inconsistent functional category naming schemes and variable gene content. Now that this framework has been

established and demonstrated to provide valuable insights, it may facilitate future studies that consider a greater number and variety of input datasets.

ACKNOWLEDGMENTS

We are indebted to the ABRF for providing us with the research group structure that encourages enriching professional interactions and continual improvement of our facilities. The GTEx Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health (NIH) and by the National Cancer Institute (NCI), NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal in November 2019. The authors are grateful to Matt Wampole, Shivanjali Joshi-Barr, and Kinsi Oberoi of Clarivate Analytics for their analyses and helpful discussions. OMW was supported in part by a Cancer Center Core Grant (P30CA023108) and Cancer Center Support (core) Grant (P30-CA14051) from the NCI and the NIH-funded Center for Quantitative Biology at Dartmouth (COBRE, 5P20GM130454-03). SWP was supported in part by the Delaware IDeA Network of Biomedical Research Excellence (INBRE) Program (NIH P20 GM103446).

SUPPLEMENTARY MATERIAL

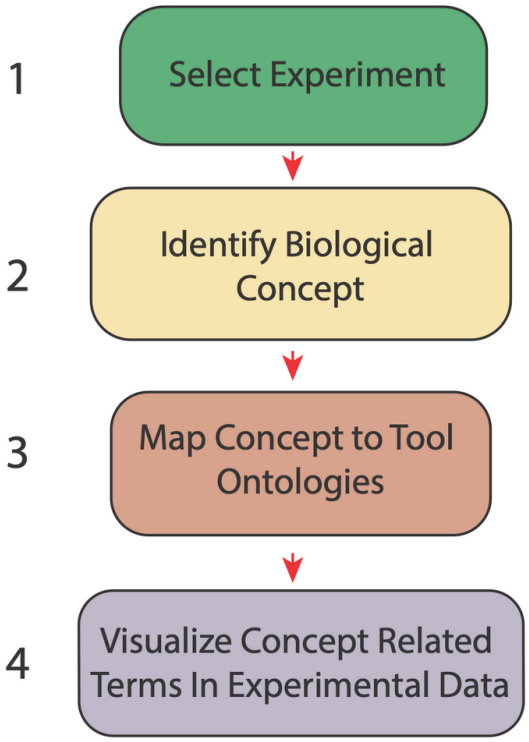


Figure S1 - Diagram of the analysis workflow. Experimental comparisons were performed using the same 4-step process. Step 1: Select experimental comparison for analysis. Four experiments were selected to represent a range of experimental designs. Step 2: Identify a single biological result (concept) for each experiment. These were results that were selected based on past experience with the data and are represented by a single MSigDB gene set. Step 3: Analyze the genes in the concept gene lists using each tool in order to identify results from each tool/ontology that are most closely related to the biological concept. Step 4: Analyze the experimental comparison with each tool and visualize the concept-related results for each tool produced in step 3 in the context of the experimental analysis. The results of step 4 are presented in Figure 3.

Table S1 - Differential expression data for each comparison considered in this study.

xlsx Supplemental_Table_S1.xlsx 3 MB

Table S2 - Ranking metrics for each comparison used in pre-ranked GSEA analyses.

xlsx Supplemental_Table_S2.xlsx 2 MB

Table S3 - Gene list for biological concepts

xlsx Supplemental_Table_S3.xlsx 18 KB

Table S4 - Concept-related annotation categories for each tool and concept list.

xlsx Supplemental_Table_S4.xlsx 123 KB

Table S5 - Analysis results for comparisons merged with concept-related annotations.

xlsx Supplemental_Table_S5.xlsx 12 MB

References

- Ahsan S, Drăghici S. Identifying significantly impacted pathways and putative mechanisms with iPathwayGuide. *Curr Protoc Bioinforma*. 2017;57(1). doi:10.1002/cpbi.24

  ↩

- Ashburner M, Ball CA, Blake JA, et al. Gene Ontology: tool for the unification of biology. *Nat Genet*. 2000;25(1):25-29. doi:10.1038/75556

  ↩

- Berriz GF, King OD, Bryant B, Sander C, Roth FP. Characterizing gene sets with FuncAssociate. *Bioinformatics*. 2003;19(18):2502-2504. doi:10.1093/bioinformatics/btg363

  ↩

- Boyle EI, Weng S, Gollub J, et al. GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes.

↩

- DeLuca DS, Levin JZ, Sivachenko A, et al. RNA-SeQC: RNA-seq metrics for quality control and process optimization. *Bioinformatics*. 2012;28(11):1530-1532. doi:10.1093/bioinformatics/bts196

↩

- Drăghici S, Khatri P, Martins RP, Ostermeier GC, Krawetz SA. Global functional profiling of gene expression. *Genomics*. 2003;81(2):98-104. doi:10.1016/S0888-7543(02)00021-6

↩

- Dragon J, Thompson J, MacPherson M, Shukla A. Differential susceptibility of human pleural and peritoneal mesothelial cells to asbestos exposure: differential susceptibility of human mesothelial cells. *J Cell Biochem*. 2015;116(8):1540-1552. doi:10.1002/jcb.25095

↩

- Dubovenko A, Nikolsky Y, Rakhmatulin E, Nikolskaya T. Functional analysis of OMICs data and small molecule compounds in an integrated “knowledge-based” platform. In: Tatarinova TV, Nikolsky Y, eds. *Biological Networks and Pathway Analysis*. Vol 1613. Methods in Molecular Biology. Springer New York; 2017:101-124. doi:10.1007/978-1-4939-7027-8_6

↩

- Falcon S, Gentleman R. Hypergeometric testing used for gene set enrichment analysis. In: *Bioconductor Case Studies*. Springer New York; 2008:207-220. doi:10.1007/978-0-387-77240-0_14

↩

- Geistlinger L, Csaba G, Santarelli M, et al. Toward a gold standard for benchmarking gene set enrichment analysis. *Brief Bioinform*. 2021;22(1):545-556. doi:10.1093/bib/bbz158

↩

- Gillis J, Pavlidis P. “Guilt by association” is the exception rather than the rule in gene networks. *PLoS Comput Biol*. 2012;8(3):e1002444. doi:10.1371/journal.pcbi.1002444

↩

- Gillis J, Pavlidis P. The impact of multifunctional genes on “guilt by association” analysis. *PLoS ONE*. 2011;6(2):e17258. doi:10.1371/journal.pone.0017258

↩

- Groß A, Hartung M, Prüfer K, Kelso J, Rahm E. Impact of ontology evolution on functional analyses. *Bioinformatics*. 2012;28(20):2671-2677. doi:10.1093/bioinformatics/bts498

↩

-

↩

- Ihnatova I, Popovici V, Budinska E. A critical comparison of topology-based pathway analysis methods. *PLOS ONE*. 2018;13(1):e0191154. doi:10.1371/journal.pone.0191154

↩

- Jassal B, Matthews L, Viteri G, et al. The reactome pathway knowledgebase. *Nucleic Acids Res*. 2022;48(D1):D49-D503. doi:10.1093/nar/gkz1031

↩

- Jiao X, Sherman BT, Huang DW, et al. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*. 2012;28(13):1805-1806. doi:10.1093/bioinformatics/bts251

↩

- Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res*. 2021;49(D1):D545-D551. doi:10.1093/nar/gkaa970

↩

- Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*. 2000;28(1):27-30. doi:10.1093/nar/28.1.27

↩

- Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Comput Biol*. 2012;8(2):e1002375. doi:10.1371/journal.pcbi.1002375

↩

- Krämer A, Green J, Pollard J, Tugendreich S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*. 2014;30(4):523-530. doi:10.1093/bioinformatics/btt703

↩

- Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database hallmark gene set collection. *Cell Syst*. 2015;1(6):417-425. doi:10.1016/j.cels.2015.12.004

↩

- Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8

↩

- Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics*. 2005;21(16):3448-3449. doi:10.1093/bioinformatics/bti551

↩

- Martens M, Ammar A, Riutta A, et al. WikiPathways: connecting communities. *Nucleic Acids Res*. 2021;49(D1):D613-D621. doi:10.1093/nar/gkaa1024

↩

- Monin L, Gaffen SL. Interleukin 17 family cytokines: signaling mechanisms, biological activities, and therapeutic implications. *Cold Spring Harb Perspect Biol*. 2018;10(4):a028522. doi:10.1101/cshperspect.a028522

↩

- Mootha VK, Lindgren CM, Eriksson KF, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*. 2003;34(3):267-273. doi:10.1038/ng1180

↩

- Nguyen TM, Shafi A, Nguyen T, Draghici S. Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol*. 2019;20(1):203. doi:10.1186/s13059-019-1790-4

↩

- Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods*. 2017;14(4):417-419. doi:10.1038/nmeth.4197

↩

- R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing; 2021. https://www.R-project.org

↩

- Raudvere U, Kolberg L, Kuzmin I, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res*. 2019;47(W1):W191-W198. doi:10.1093/nar/gkz369

↩

- Santaguida S, Richardson A, Iyer DR, et al. Chromosome mis-segregation generates cell-cycle-arrested cells with complex karyotypes that are eliminated by the immune system. *Dev Cell*. 2017;41(6):638-651.e5. doi:10.1016/j.devcel.2017.05.022

↩

- Schaefer CF, Anthony K, Krupa S, et al. PID: the Pathway Interaction Database. *Nucleic Acids Res*. 2009;37(suppl_1):D674-D679. doi:10.1093/nar/gkn653

↩

- Schriml LM, Arze C, Nadendla S, et al. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res*. 2012;40(D1):D940-D946. doi:10.1093/nar/gkr972

  ↩
- Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci*. 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102

  ↩
- Tarca AL, Bhatti G, Romero R. A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity. *PLoS ONE*. 2013;8(11):e79217. doi:10.1371/journal.pone.0079217

  ↩
- Tarca AL, Draghici S, Khatri P, et al. A novel signaling pathway impact analysis. *Bioinformatics*. 2009;25(1):75-82. doi:10.1093/bioinformatics/btn577

  ↩
- Timmons JA, Szkop KJ, Gallagher IJ. Multiple sources of bias confound functional enrichment analysis of global -omics data. *Genome Biol*. 2015;16(1):186. doi:10.1186/s13059-015-0761-7

  ↩
- Voichita C, Ansari S, Draghici S (2021). ROntoTools: R Onto-Tools suite. R package version 2.22.0.

  ↩
- Wijesooriya K, Jadaan SA, Perera KL, Kaur T, Ziemann M. Urgent need for consistent standards in functional enrichment analysis. *PLOS Comput Biol*. 2022;18(3):e1009935. doi:10.1371/journal.pcbi.1009935

  ↩
- Xie C, Jauhari S, Mora A. Popularity and performance of bioinformatics software: the case of gene set analysis. *BMC Bioinformatics*. 2021;22(1):191. doi:10.1186/s12859-021-04124-5

  ↩
- Yousef M, Ülgen E, Uğur Sezerman O. CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis. *PeerJ Comput Sci*. 2021;7:e336. doi:10.7717/peerj-cs.336

  ↩
- Zhu A, Ibrahim JG, Love MI. Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics*. 2019;35(12):2084-2092. doi:10.1093/bioinformatics/bty895

  ↩