WhatIsMyGene: Back to the Basics of Gene Enrichment

Kenneth Hodge[1], Thammakorn Saethang[2]

[1*]Unaffiliated,

[2]Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, 10900, Thailand

*Correspondence: whatismygene@gmail.com

Oct 31, 2023

WIMG Abstract

Since its inception over 20 years ago, gene enrichment has been largely associated with curated gene lists (e.g. GO) that are constructed to represent various biological concepts; the cell cycle, cancer drivers, protein-protein interactions, etc. Researchers expect that a comparison of their own lab-generated lists with curated lists should produce insight. Despite the abundance of such curated lists, we here show that they rarely outperform existing individual lab-generated datasets when measured using standard statistical tests of study/study overlap. This demonstration is enabled by the WhatIsMyGene database, which we believe to be the single largest compendium of transcriptomic and micro-RNA perturbation data. The database also houses voluminous proteomic, cell type clustering, lncRNA, epitranscriptomic (etc.) data. In the case of enrichment tools that do incorporate specific lab studies in underlying databases, WIMG generally outperforms in the simple task of reflecting back to the user known aspects of the input set (cell type, the type of perturbation, species, etc.), enhancing confidence that unknown aspects of the input may also be revealed in the output. A limited number of GO lists are included in the database. However, these lists are assigned backgrounds, meaning that GO lists that are replete with abundant entities do not inordinately percolate to the highest ranking positions in output. We delineate a number of other features that should make WIMG indispensable in answering essential questions such as "What processes are embodied in my gene list?" and "What does my gene do?"

Main

Modern transcriptomic protocols typically result in the identification of 20,000 RNA species. Proteomic datasets may now contain over 8,000 proteins. In most cases, researchers are interested in subsets of identified entities, e.g. the set of transcripts that are strongly up-regulated when interferon is applied to a particular cell line. Once such a subset has been selected, the researcher seeks to compare the subset with curated gene lists (CGLs) that presumably reflect important biological processes. In the case of the interferon experiment, it is likely that a CGL with a term such as "innate immunity" would best match the experimental subset (ES). In many cases, the researcher may have little expectation as to the result of such a comparison. A great deal of effort has been devoted to the construction of curated lists. At the time of writing, the GO consortium [1] offers 42,887 terms (lists), with KEGG[2], PANTHER[3], REACTOME[4] (34788843), GSEA (via "Hallmark" gene sets[5]) and others, each offering thousands of terms.

CGLs represent biology's best attempts to identify gene sets that recurrently appear over multiple experiments. However, problems may arise when various statistical tests are applied in order to isolate

and rank the CGLs that best match an experimental result. These issues include the potential absence of important genes in CGLs[6], the over-representation of particular genes over many CGLs[6-9], derivation of CGLs from a small subset of relevant studies and cell types[6, 10], significant changes in the composition of CGLs over time (the "reproducibility crisis"[8]), absence of human oversight in generating CGLs[8], distortion due to the effects of multifunctional genes in CGLs[11], and the questionable assumption that biological entities are independently expressed[12, 13]. The blame for some of these deficiencies need not rest solely on the curating bodies, as experimentalists may choose to research easy, predictable, or "hot" genes in pursuit of job security and other mundane concerns[14-16]. To the above concerns we would add the possibility that the typical ES may best be described in terms of modules of genes that do not share obvious functions (as opposed to one or two dominant themes that are easily subsumed by CGLs). A final important concern is the simple fact that CGLs lack backgrounds. That is, the complete set of genes from which a CGL subset was drawn, a critical input for statistical analysis, is unknown. Typically, analysis tools operate under the assumption that a CGL's background is simply the total number of known genes for an organism, or the total number of genes found in all CGLs. Gene enrichment tutorials may emphasize that the background of the experimental subset must be carefully considered, but seem unconcerned with the fact that CGLs themselves do not have backgrounds.

Numerous attempts have been made to resolve the above concerns by generating improved algorithms for gene enrichment analysis. We promote a solution that resides not primarily in algorithms, but in the composition of the enrichment database: a "study/study" approach of matching one's own experimental subset with subsets generated by other experimenters[10]. RNA-seq results, for example, do not ordinarily suffer from over-representation of popular genes, the purposeful ignoring of rare or uninteresting genes, fads in the field, or unusual biases in the appearance of multifunctional genes. The experimental subsets identified via RNA-seq experiments are not expected to "evolve" over the years. Furthermore, an RNA-seq experiment's background can be easily ascertained. Below, we further explore some difficulties presented by the usage of CGLs, and detail how WhatIsMyGene, via sheer database size and a variety of analysis tools and approaches, may greatly aid in the generation of biological insight.

Results

The absence of CGL backgrounds is consequential

It could be argued that the decision to assume either a large or small background for a CGL would not have significant repercussions on output. That is, when ranking the relevance of various CGLs against an experimental subset, rankings will not change, or will change insubstantially, depending on the selection of CGL backgrounds.

We first illustrate the dependence of output *P*-values on CGL backgrounds. In Fig. 1a, we fix the experimental background, but allow the CGL background to vary, applying Fisher's exact test over 4 combinations of ES sizes and ES/CGL intersections. For simplicity, the ES size and the CGL size are equal in all cases. In the image, case 2 illustrates a "typical" ES/CGL result, where significance is seen only when the actual intersection is considerably less than the expected intersection. Such a result would usually be excluded from a supplemental dataset of ranked CGLs. Case 1 is of greater interest. Here, if the CGL has a background greater than 8600, the ES/CGL intersection would be significant. However, a CGL background below 4600 would also be significant due to anti-correlation between ES and CGL. One may imagine a case where the ES represents genes down-regulated on a particular drug treatment, and

the CGL involves genes up-regulated in Alzheimer's disease. If the CGL "ground-truth" background is 20,000, an Alzheimer's treatment may have been identified. However, if the CGL ground-truth background is 4,000, the drug treatment may actually accelerate the progression of Alzheimer's disease. Comparing case 3 and case 4, case 4 settings initially outperform case 3 settings because of one extra gene in the case 4 ES/CGL intersection. However, the case 4 line levels off at the background setting of the experimental dataset (10,000). In all cases, the choice of CGL background strongly influences the measure of significance.

One may ask how it is possible that the proper background for a CGL would be 4,000. We present several simple cases as illustrative of the possibility. First, consider 2 hypoxia-related studies, having backgrounds of 10,000 and 20,000 genes. Shall we assume that 10,000 genes are common to both sets? No, because the 10,000 member set may contain genes not found in the 20,000 member set. Perhaps the joint background is 9,000. We wish to include 10 more hypoxia studies as the basis for a CGL. It is apparent that incorporation of each additional experimental set lowers the background of the CGL. The general effect of this process is to enrich for abundant genes found in all datasets, and eliminate rare ones. We may now have a CGL with an effective background of 4,000. The problem is exacerbated by the inclusion of genes derived from studies with particularly small backgrounds, genes from "targeted microarrays", or genes highlighted in studies that predate the big-data era.

Next, we generate a frequency table of genes found in 2,604 GO[1] CGLs. Using a whole-human list of proteins ranked according to abundance[17], frequencies of genes are plotted against their abundances (Table S1). Fig. 1b makes clear that abundant genes are indeed enriched in these CGLs; $P = 5.4*10^{-73}$ (Pearson). Not surprisingly, the single most common gene over these CGLs is TNF, with 695 appearances. The gene is an outlier, however, given its moderate abundance ranking (the 6206th most abundant gene in a list of 17861). Examples of genes that are both frequent and abundant are CTNNB1, RHOA, ANXA1, and HSPA1A.

Assessing GO-CGLs according to abundance rankings needn't be limited to ES/GO-CGL combinations. Fulcher et al ranked GO-CGLs according to their tendencies to produce high "category false positive rates"[13] against mouse and human transcriptomic data. Using a list of mouse brain proteins ranked according to abundance[18], we first generate "abundance rankings" for 2604 GO-CGLs. The rankings reflect the mean abundance of genes in each CGL; CGLs with highly abundant genes receive high rankings. Fig. 1c shows a strong relationship between GO-CGL false-positive rates and our estimates of GO-CGL abundance (Table S2).

Finally, we invoke a supplemental table from an actual study in which the mouse brain proteome was measured after RAB3GAP1 disruption[19]. We plot the supplementary table *P*-values representing GO-CGL enrichment against our GO-CGL abundance rankings (Table S3). Fig 1d shows that GO-CGLs with high abundance rankings strongly correlate with the most significant *P*-values. An exhaustive survey of supplementary gene enrichment tables is beyond the scope of this work. Fulcher et al[13], in fact, performed this exercise for 31 brain transcriptomic studies, finding excessive representation of particular GO-CGLs over multiple phenotypes. The fact that multiple enrichment tools were used over these 31 studies suggests that bias, to some extent, lies outside the mere algorithms utilized by these tools. One should consider the possibility that numerous such tables are skewed in this way due to the absence of backgrounds for CGLs.

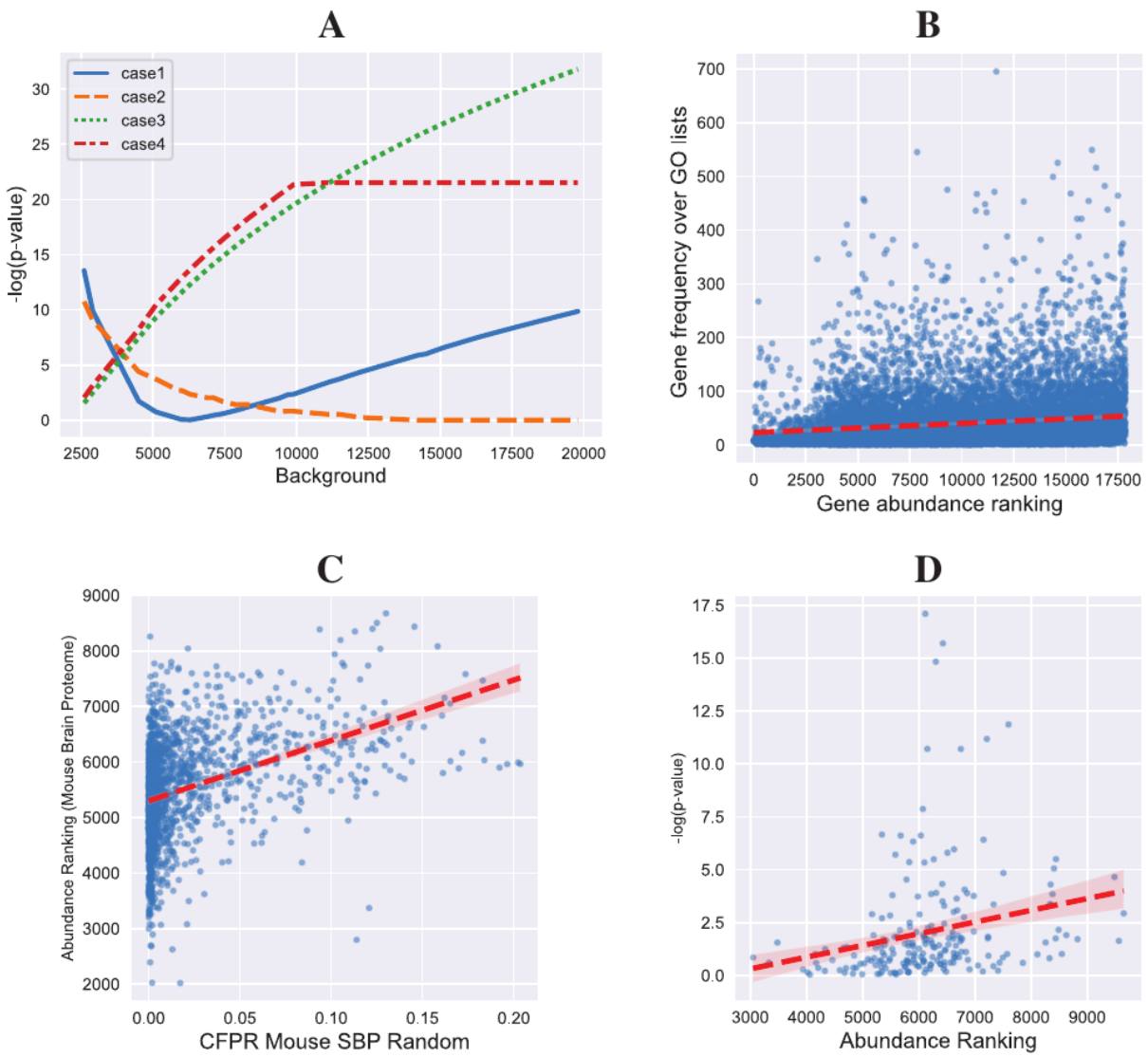

Figure 1. Assumptions regarding CGL background sizes may affect gene enrichment results. Fig 1a. Dependency of significance measurement on CGL background. Intersection, CGL size, ES size, and experimental background for Fisher's exact test are as follows: Case 1 (40,500,500,20000), Case 2 (2,200,200,20000), Case 3 (20,200,200,20000), Case 4 (21,200,200,10000). The results assume that all genes comprising the CGL background are found in the experimental background. The possibility of testing over multiple CGLs is ignored in the significance calculations. Fig. 1b. Abundance vs frequency of GO-CGL genes. The frequency of gene appearance over 2604 GO-CGLs is plotted against abundance rankings derived from the whole human proteome; a high rank means high abundance (R^2 = .155; P = $5.4*10^{-73}$). Fig. 1c. GO-CGL false positive rates correlate with transcriptomic abundance (R^2 =.39, P = $1.6*10^{-55}$). Fig 1d. CGL enrichment tables may be biased toward abundance. 210 GO-CGLs are assigned P-values in a supplementary table[19]. These values are plotted against abundance rankings for each GO-CGL (R^2 =.23, P = .0008). It is apparent that CGLs containing abundant genes tend to receive the most significant P-values.

It might be instructive to examine the contents of a CGL with a particularly low abundance ranking. Numerous GO CGLs are devoted to ribosomal components; e.g. GO:0022625, "cytosolic large ribosomal subunit." This CGL is indisputably an unabridged collection of large ribosomal components. Such genes are, of course, quite abundant over all cell types and organisms. Shall we, however, assign a background equivalent to the count of all genes in the genome to this CGL?

Study/study comparisons outperform study/GO-CGL comparisons

At the time of writing, the abridged WIMG database contains 70300 gene lists, of which 2604 are GO CGLs suitable for the following analysis. We performed Fisher's exact test using the 124 most recently entered ESs against the 70300 gene lists ($8.7*10^6$ *P*-values), observing the rankings of the best-performing GO CGLs against each of the 124 lists (Table S4). In all cases, an ES could be found that outranked (i.e. generated a more significant *P*-value than) the highest ranking GO-CGL. Over the 124 tests, the best performing GO-CGL appeared as high as the 3rd position, but as low as the 2268th position. The average position of the highest ranked GO-CGL over the 124 tests was 468; i.e. typically, 467 other lists outperformed the best GO-CGL. Given the 2604 lists, one would naively assign a 50% probability of finding at least one GO-CGL within the top 27 ranks for each of the 124 tests, but this occurred in only 19 of the 124 cases ($P = 5.4*10^{-16}$, binomial distribution) under liberal background settings (see below).

Because of the parent-child structure of GO-CGLs, redundancy between GO-CGLs is expected. One could thus argue that it is unreasonable to expect typical GO-CGLs to breach the 27th rank in our test. It could also be argued that since CGLs capture general processes over various species, cell types, and experimental conditions, it should not be surprising that very specific ES/ES combinations can be found that outrank CGL/ES combinations. Below, we show that CGLs can indeed perform strongly against ES/ES competition.

WIMG lists

Currently, the WIMG database includes 265 "WIMG Lists." These are created by three methods. First, studies representative of a well-defined process (e.g. hypoxia) may be gathered. Genes strongly altered throughout these studies are gathered into up- and down-regulation lists. These genes must be derived from unabridged datasets or supplemental tables, not via text-mining or scanning of papers. The second method involves clustering. As an example, *P*-values representing all disease-related ES/ES combinations may be generated. Clustering is then performed on a resulting disease/disease matrix that would currently contain $3*10^7$ *P*-values. Genes most representative of a particular cluster are accumulated into a list. The third method simply labels specific ESs as "WIMG exemplars." For example, we note that genes down-regulated upon Raver2 knockdown in one study[20] repeatedly generate extreme *P*-values when compared against numerous other ESs. This process is not captured by GO-CGLs or by the other two classes of WIMG lists. We have labeled 24 ESs as "WIMG exemplars."

We repeated the above test of 124 ESs against 70300 gene lists, this time observing the rankings of the 124 best-performing WIMG lists. Here, a WIMG list received top ranking in 10 of the 124 tests. The lowest ranking was 907, with an average ranking of 60. Given the 265 lists, one would expect a WIMG list to be found within the top 265.3 rankings in half of the 124 studies. However, WIMG lists were found in this region in 118 cases ($P = 7.7*10^{-24}$, binomial distribution). Ignoring expectation values, the top-ranked WIMG list still outranked the top-ranked GO-CGL in 100 of the 124 tests. We would argue that even over a mere 124 ESs, certain processes recur, and these are well-captured in WIMG lists.

The WIMG database contains 160 lists based on WGCNA clustering[21]. Typically, studies that utilize this procedure involve single-cell analysis, requiring large matrices to generate reliable gene-gene co-expression patterns. One may thus not expect such lists to overlap strongly with, for example, an ES involving the application of a cancer drug to a cell line. Despite this disadvantage, we repeated the 124 ES test against WGCNAs. Here, the highest ranking was 2, and the lowest 2044, with an average ranking of 558. WGCNAs breached the ranking expectation value on 72 occasions (*P*=.01, binomial test). As with most WIMG lists, the WGCNA method extracts patterns from raw data, as opposed to requiring that all genes within a list conform to easily-defined categories. This may explain WGCNA's relatively strong performance compared to GO-CGLs.

For clarity, all WIMG lists were generated prior to entry of the 124 ESs into the database. The composition of all WIMG lists is available in Table S5. The results of the above GO/WGCNA/WIMG comparison are available in Table S4. We further discuss results of these comparisons in the "WIMG Background Options" section below.

WIMG study/study comparisons are particularly informative

Despite a plethora of gene enrichment databases[22], few databases contain extensive ES data from which ES/ES comparisons may be drawn. Enrichr[23] and GSEA are two exceptions. We set out to compare the ability of WIMG, Enrichr, and GSEA to "predict" two known features of the input ES; the cell type and the perturbation type (knockout, disease, cell type, etc.). Enrichr was excluded from this test because of the absence of a single output list of top-ranked studies. 201 differential expression gene lists derived from studies appearing in the GEO database[24] in December 2021 were generated. Each list, containing no more than 250 genes, was entered into the GSEA and WIMG tools, and the top ranking 5 output ESs or CGLs were examined for the presence of the cell type and perturbation type associated with the input. In terms of perturbation, WIMG made correct predictions 25.2% of the time, while GSEA scored at 8.6% ($P = 1.02*10^{-11}$, t-test, unequal variance). For cell type, WIMG scored at 35% and GSEA at 13.9% ($P = 7.5*10^{-12}$) (see Table S6 and Methods).

The GSEA/WIMG comparison could be critiqued for a number of reasons. In particular, the composition of the December 2021 GEO database may not mirror the composition of the GSEA database, which certainly contains a larger ratio of CGLs to ESs than the WIMG database. We would point readers to another result involving the WIMG "keyword" feature. Each WIMG study is tagged with multiple keywords. Some of the keywords apply to many studies (e.g. brain, blood, disease), while others apply to a small fraction of studies (parasite, hdac, UPR). The appearance of these keywords over the top 500 ranked studies is compared to their appearance over the remainder of the database using a simple binomial test. The tool is unique to WIMG, so comparisons to other tools cannot be made. Table 1 offers some examples of the concordance between input data and output results.

STUDY	WIMG keywords: negative log(p)>10	NOTES
upregulated in ovcar3 line on boris oe vs kd (GSE166767)	female, epithelial, line, ovary, adherent	5/5 WIMG studies involve ovcar3
downregulated in a549 line on il17d oe (GSE191111)	carcinoma, lung, epithelial, line, adherent, male, female, breast	5/5 WIMG studies involve a549 cell line

enriched on dmd clip (vs input)(GSE128318)	rip, epithelial, adherent, line	note the correct prediction of an RNA-IP experiment
upregulated in gingival epithelial cells on irf1 kd (both w/hsv infection)(GSE184463)	line, adherent, epithelial, kd	WIMG predicted a knockdown in 5/5 studies
upregulated in mouse follicular B cells on cic ko (GSE163455)	lymphocyte, b-cell, blood	WIMG 4th ranked study involves cic chip-seq
downregulated in tregs on stimulation w/anti-tnfr2 agonistic antibody vs no stimulation (GSE161835)	lymphocyte, blood, treatment	4/5 WIMG studies involve t-cell activation
upregulated in ipsc-derived astrocytes from schizophrenia-affected twin vs unaffected (GSE191248)	stem, neural, fibroblast	2/5 WIMG studies involve astrocytes
upregulated in mouse e13.5 heart on prdm16 ko (GSE179386)	heart, mouse, cell type, embryo, brain, ko, neuron	2/5 WIMG studies involve mouse embryonic heart
downregulated in mouse cd11c+ cells on high cholesterol diet vs control (Apolipoprotein E derived from CD11c+ cells ameliorates atherosclerosis)	blood, mouse, lymphocyte	WIMG dendritic keyword -log(p) = 6.47
upregulated in mouse aml line on bcl11a oe (both w/trib1 oe)(GSE147797)	Blood, mouse, lymphocyte, bone marrow, ko, macrophage	WIMG leukemia keyword -log(p) = 7
upregulated in ovary-metastatic tissue vs primary gastric adenocarcinoma (GSE191139)	cell type, brain, nerve, ovary	WIMG cancer_tissue keyword -log(p) = 8
downregulated in ovary-metastatic tissue vs primary gastric adenocarcinoma (GSE191139)	disease, cancer tissue, stomach, blood	WIMG metastasis keyword -log(p) = 6.98
upregulated in huvecs on shp2 kd (GSE191056)	endothelial, umbilical, ifn	4/5 WIMG studies involve endothelial cells
upregulated in hnecs on 14d hrv16 infection (GSE190344)	ifn, cytokine, virus, RNA virus, infection, treatment, umbilical	WIMG 5th ranked study involves nasal tissue upon rhinovirus infection
downregulated in CD11c+ cells from 9m vs 5d mouse brain (GSE190713)	mouse, ko, hippocampus, brain	2/5 WIMG studies involved Alzheimer's
downregulated in mouse follicular B cells on cic ko (GSE163455)	Blood, lymphocyte, b-cell, mouse, ko, spleen	4/5 WIMG studies involve B-cells
downregulated in mouse cerebellum on aptx ko (GSE174841)	brain, mouse, cerebellum, ko, cortex	note that cerebellum is a significant keyword
downregulated in gingival epithelial cells on irf1 kd (both w/hsv infection)(GSE184463)	ifn, cytokine, breast, epithelial, treatment, gland, line	WIMG study #5 involves irf1
downregulated in cd138+ multiple myeloma cells at relapse vs diagnosis (GSE162403)	blood, cell type, lymphocyte, pbmc	WIMG study #1 relates to CML; study #3 to AML
upregulated in keloid tissue vs healthy (GSE190626)	cancer_tissue, fibroblast, disease	WIMG #1 ranked study involves keloid tissue
downregulated in rat striatum with (vs w/o) physical exercise following stroke (GSE190355)	brain, hippocampus, mouse, rat, ko, striatum	note WIMG keyword "rat"; note WIMG keyword "striatum"
upregulated in mescs on ubn2 kd (GSE177026)	embryo, stem, mouse, muscle, ko	5/5 WIMG studies involve mescs
upregulated in squamous vs non-squamous lung tumors (GSE190265)	skin, breast, cancer_tissue, keratinocyte, esophagus, lung,	5/5 WIMG studies involve lung squamous cell carcinoma

	gland, head/neck, bladder, tongue	
upregulated in mouse MLL-AF9 leukemia stem cells on tifab oe (GSE178853)	blood, leukemia	note the "leukemia" keyword; 4/5 WIMG studies relate to leukemia
upregulated in HO-1-N-1 line on C. parapsilosis treatment vs heat-killed C. parapsilosis treatment (GSE169278)	epithelial, adherent, female, line	top ranked WIMG study involves fungus (aspergillus) infection
downregulated in patient-derived rhabdomyosarcoma lines on ehmt1 kd (GSE179049)	female, line	4/5 WIMG studies relate to sarcoma
upregulated in mouse b-cells heterozygous for tbk1 (GSE188659)	blood, lymphocyte, mouse, b-cell	2/5 WIMG studies involve mutation; 1 involves haploinsufficiency
upregulated in mouse microglia on tamoxifen-induced Cre activity (vs no induction)(GSE190207)	ifn, cytokine, virus, RNA virus, infection, treatment, macrophage, mouse, microglia	note the WIMG microglia keyword

Table 1: WIMG output (columns 2 and 3) strongly mirrors study input (column 1).

We would ask readers to peruse Table S6 and enter their own ESs into the WIMG "Fisher" tool for further evidence of the ability to predict the nature of the input set. Researchers, of course, are not interested in learning that their pancreatic adenocarcinoma study is indeed a cancer study. However, WIMG's ability to correctly mirror *known* aspects of the input to users should enhance confidence in its ability to properly reflect back *unknown* or unexpected aspects of the study. If a knockout results in output that mirrors the trends seen in glioma, or an antiretroviral treatment reverses Alzheimer's trends, or an exercise regimen reverses trends seen in aging studies, researchers may wish to take note.

Tackling gene set overlap

After submitting an ES for analysis, the CGL output list may, depending on the tool, contain hundreds or even thousands of ranked CGLs. Unfortunately, this does not mean hundreds of insights have been generated, as highly ranked outputs tend to be quite similar in content. This is the gene set overlap problem. A number of solutions have been employed[7, 25-27]. However, gene set overlap may still remain strong. For example, utilizing one of these solutions, we enter a list of genes upregulated in nasal epithelial cells on ifn-lambda treatment[28]. Selecting a redundancy reduction feature, the first and second ranked outputs, involving the response to type I interferon and the response to gamma interferon, still contain 16 genes in common. Applying Fisher's exact test to these two sets, we derive $P = 10^{-224}$.

WIMG takes an extreme approach to eliminating redundancy. Using the site's "Third Set" tool, the user enters two gene sets that are known to intersect. The tool seeks a third gene set that overlaps with the first, "central" set, but does not overlap significantly with the second set (i.e. $P>.05$). Inputting the ifn-lambda experimental data and genes found in GO:0034340 (response to type I interferon), and setting the "experiment" filter to "external list", which eliminates all ESs from consideration, the tool is actually unable to find a second GO-CGL that intersects significantly with the central set but not the secondary set. Removing the "external list" restriction, however, the tool is indeed able to find numerous studies that overlap with the ifn-lambda study but not the GO-CGL. The second-ranked study involves genes upregulated in the hela line on ifn-gamma vs ifn-beta (a type I interferon) treatment[29] (Fig 2a). This result "makes sense", perhaps heightening confidence that the top-ranked study, involving a novel αVβ3 inhibitor, relates specifically to ifn-gamma metabolism[30].

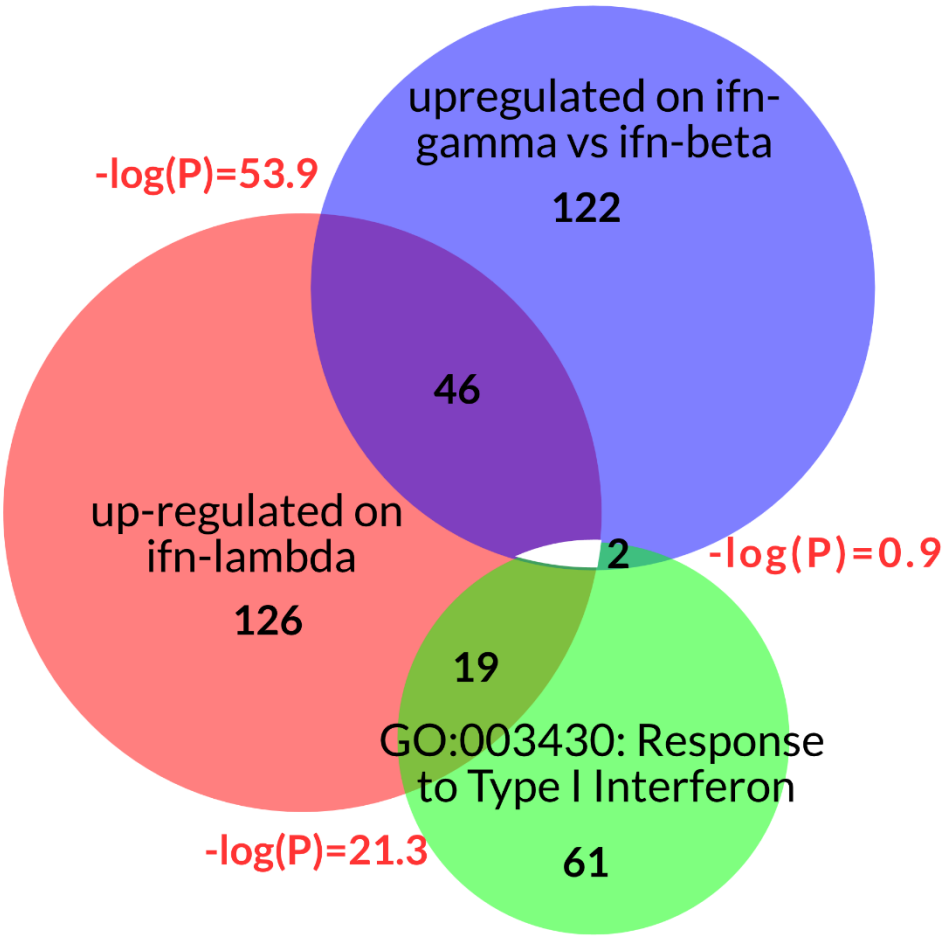

Fig 2. WIMG redundancy reduction. The central IFN-lambda gene set and the secondary GO:0034340 gene set were entered into the "Third Set" tool. The tool found a third gene set that intersects strongly with the central set, but not the second set. *P*-values for each of the three intersections are given.

The general inability to find CGLs that do not themselves overlap, but do overlap with a central ES, may have necessitated the construction of redundancy reduction algorithms that cannot entirely eliminate significant intersection between output sets. When ESs are allowed as outputs, however, a thoroughly restrictive approach may produce results.

Note that the tool has utility beyond redundancy reduction. If a user has reason to believe that a dominant set of genes is "blocking" an otherwise interesting result, the third set tool effectively removes the dominant group, as well as gene lists that contain content similar to that of the dominant group. The blocking set could range from a specific list of genes typically involved in innate immunity, which could be hiding subtleties involved in the response to a particular virus, to a more general list of genes which evince disproportionate differential expression over a large range of phenotypes[31].

One important drawback to the WIMG approach is the fact that two sets (as opposed to one set) must first be identified for input, meaning that the "Third Set" approach may serve as a complement to existing tools.

The cut-off problem

Two basic algorithms are most commonly employed for gene set analysis; over-representation analysis (ORA) and functional class scoring (FCS)[26, 32]. WIMG uses an ORA-based algorithm. When discussing the weaknesses of ORA against FCS, it is commonly pointed-out that ORA requires input of a subset of all identified genes, whereas FCS requires only that all genes be ranked according to some criteria. The truncation of ORA input lists may be considered somewhat artificial; genes differentially expressed at $P = .049$ may make the cut, but genes with $P = .051$ are ignored. A related issue is the fact that ORA tools treat each gene equally; genes differentially expressed with extreme significance are no more valuable than genes that barely make the cut.

To some extent, however, WIMG resolves the above critiques. This is because 1) 79% of the gene lists in the WIMG database are sorted and 2) users may elect to shorten a WIMG gene list such that only the most significantly altered genes remain. When large numbers of genes show statistically significant differential expression, lists are sorted according to log(fold change) divided by significance. Otherwise, lists are sorted by log(fold-change) alone. Users can examine both full and truncated WIMG lists, looking for *P*-value alterations as the list shortens. If, for example, the most significant *P*-value is found against a truncated version of a WIMG list, one can conclude the excluded genes are fairly irrelevant. In the case of FCS, a "leading edge" subset of the complete input list is generated. As with any ORA tool, the WIMG user is free to alter the length of an input list. However, with WIMG, database lists may be truncated as well, reflecting the view that input lists and database lists should be treated equivalently.

We intersected 1,006 WIMG ESs and CGLs, all pruned to exactly 200 genes, against the remainder of the WIMG database, generating $7*10^7$ P-values. For each of the 1,006 lists, the single most significant *P*-value against the remainder of the database was recorded. We then repeated the same exercise, this time pruning the 1,006 lists to 100 genes each. We refer to this as the "halving test." A large, "well-focused" list of genes should always outperform a short, equally focused list in tests of significance, as increased intersection between an input and a database set will overwhelm minor alterations in the number of genes in these two sets. Nevertheless, in 6.8% of cases significance increased when the 1,006 lists were halved in size, suggesting that the remaining 100 genes in the 200 member set are of minimal importance, at least with respect to whatever insights can be gleaned from the two gene lists in question.

Most WIMG GO-CGLs are sorted according to "perturbability." In brief, a frequency table of genes appearing over all WIMG ESs that involve a perturbation (e.g. a disease or knockout) is generated. Currently, EGR1 is the most commonly perturbed gene in the database, appearing in 7% of ESs. Genes in

GO-CGLs are then sorted according to these frequencies. If EGR1 is present in a GO-CGL, it will rank first in the list. Housekeeping genes, while common and critical to many of the processes that GO-CGLs seek to embody, receive the lowest rankings. Just as housekeeping genes and other rarely altered genes tend not to be found in lab-based ESs, they may also be excluded from WIMG GO-CGLs via truncation.

We repeated the halving test, this time focusing on 1,004 sorted GO-CGLs with at least 200 members (which were subsequently trimmed to precisely 200 members). GO-CGLs were not allowed to intersect with GO-CGLs, but were allowed to intersect with WIMG-CGLs. In this case, significance increased 34% of the time when GO-CGLs were halved. The pruning of GO-CGLs may actually improve their relevance to ESs.

WIMG does not offer an option to simultaneously test an input set against multiple database gene set lengths, due simply to the processing required. Currently, one would need to generate an initial ranked output list with the "Fisher" tool, trim the database via the "Restrict IDs" option, generate a second ranked list, and compare the two lists.

WIMG background options

We have shown that the absence of CGL backgrounds creates difficulties in assessing the significance of ES/CGL overlaps. The above construction of "abundance rankings" suggests, however, that backgrounds could actually be assigned to CGLs, thus improving their relevance to ESs. The subject will discussed in depth in a forthcoming paper, as methods for deriving backgrounds, validation of these backgrounds, and an examination of the consequences of background-adjustment requires considerable exposition.

Whenever possible, WIMG records the "actual" background of genes involved in a study. If the study involves RNA-seq, for example, this is simply the count of all genes that pass quality control. Thus, the quality of any method that estimates the background of a set of genes can be assessed by comparing these estimates to actual backgrounds. Using a "reverse-Fisher" approach (see Methods) we assigned backgrounds to 44,620 WIMG ESs based on cell type, species, and the gene composition of the individual ESs. Over these studies, we find high correlation between the actual backgrounds and the background implied by the gene composition (R^2 =.126, P=6*10^{-158}, Pearson). Fig. 3 plots the actual and implied background of a subset of 1% of these lists. The only requirement for entry into this test was that we had already generated abundance ranks for the specific cell types involved in the tests (e.g. an MCF7 study would require an MCF7-based list of genes ranked by abundance). Studies were entered as we encountered them in the GEO database or other sources. A second, naïve, method of background estimation, that of simply doubling the abundance ranking, is also utilized. The reverse-Fisher method predicted actual backgrounds with R^2=.54, P=4.7*10^{-23}, while doubling of the abundance ranking gave R^2=.52, P=5*10^{-21}. We attribute the improvement in R values versus the large scale test to proper matching of tissue types. Note that background prediction remained accurate even when the actual backgrounds were relatively small, as seen in proteomic data sets and occasional RNA-seq results.

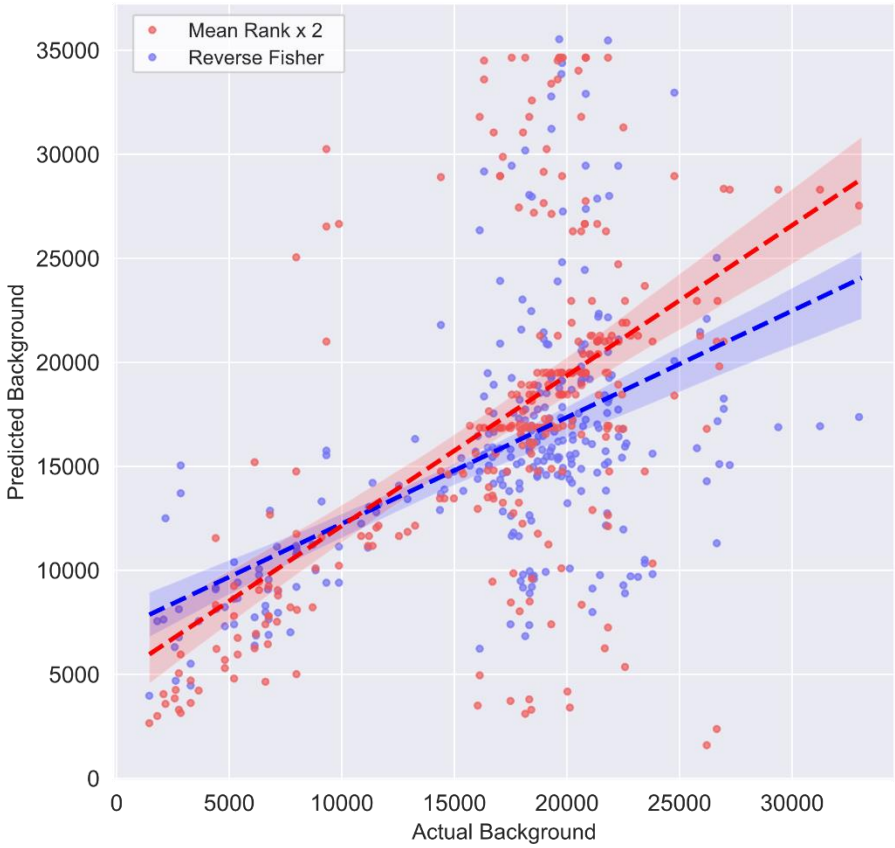

Fig. 3. Based on gene list composition and tissue type, backgrounds can be predicted with accuracy.

Though refinement of these background-adjustment methods would undoubtedly be beneficial, WIMG has proceeded to generate backgrounds for GO-CGLs and other gene lists (e.g. gene clusters generated from single-cell transcriptomic matrices) accordingly. The assumption is that if background-adjustment offers reasonable background estimates for studies with known backgrounds, it should do the same for gene lists with unknown backgrounds. The typical GO-CGL received a median background of 10254, and a mean background of 15300, with a low of 643 (GO:0072562, "blood microparticle") and a high of 50148 (GO:0008237, "metallopeptidase activity"). Table S7 gives our estimated backgrounds for 2605 GO-CGLs. The underlying data is found in Table S9.

WIMG users may over-ride the backgrounds that WIMG attaches to GO-CGLs (see supplemental WIMG manual). Returning to the GO/WGCNA/WIMG comparison tests, background-adjusted GO-CGLs, WIMG lists, and WGCNA lists all tended to fare considerably worse against ES competition versus the non-adjusted (background = 20000) lists used in our initial tests. Nevertheless, background-adjusted WIMG lists outperformed expectations on 99 of 124 occasions, while background-adjusted GO-CGLs outperformed on only 8 occasions. In Table S4, the single most significant *P*-value derived from each of the 124 ESs against the remainder of the database is shown. One may term ESs associated with extreme significance as particularly "relevant" (conversely, ESs with weak maximal significance would be similar to ESs generated simply by scrambling gene lists). Interestingly, background-adjusted WIMG list rankings correlated with these maximal *P*-values ($P = .013$, t-test), which was not the case for non-adjusted lists. Following background-adjustment, WIMG lists correlate more strongly with the "relevance" of studies.

We considered the possibility that our typical sorting method for input ES data, that of dividing a gene's log(fold-change) by the significance of its alteration, could alter the rankings of GO-CGLs vs ESs. We thus examined the results of first excluding all genes with significance >.05, and then sorting the remaining genes according to fold-change in a new test involving 132 ES input lists. In both cases, background-adjustment was utilized for GO-CGLs, WGCNA lists, and WIMG lists. However, the sorting procedure made little difference in rankings. Over the 132 tests, GO-CGLs outperformed expectations on only four occasions using WIMG-typical sorting; the figure increases to five using the second sorting approach. WIMG lists outperformed expectations on 123 occasions using the first approach, and 121 times using the second. It should be noted that in these tests, WGCNA lists failed to outperform expectations. The results of these tests are available in Table S4 sheet "test2."

Discussion

We have shown that GO-CGLs, and presumably other similarly constructed lists, perform poorly when ranked against ES competition. This is not simply because "averaged-out" exemplars of particular processes cannot be expected to fare well against very specific ES/ES combinations, as WIMG lists outperformed ES competition on numerous occasions. WIMG evinces a strong ability to reflect back known aspects of input data to users. Put differently, the tool maximizes true positive identifications. WIMG also tackles standard enrichment obstacles such as gene-set overlap and the problem of arbitrary cutoffs in ORA algorithms. We suggest the sorting of genes in CGLs and the assignment of backgrounds to CGLs as means to improve their relevance to real-life experimental subsets.

As with any gene enrichment tool, WIMG should not be considered a one-stop solution. We have already pointed out some features that may suffer in comparison to other tools. To these, we add the fact that our emphasis on database bulk may come at the cost of quality control and speed of output. Major sources of data include the GEO and GREIN[33] databases; no batch correction of test vs control sets beyond that offered on these sites is conducted. We do tend to minimize covariates (e.g. prefer analysis of the effect of a drug on a single cell line vs multiple cell lines) for the simple reason that the percentage of differentially altered genes is often much higher this way. Another source is supplemental datasets, which we take at "face value." Datasets with obvious gender bias (e.g. an excess of males in a test group vs a control group), unusually small backgrounds, and targeted arrays (e.g. those focused only on cancer genes) are generally excluded from the database, however. Our assumption is that noisy sets in the database generally will not percolate to the top output ranks; this could be proven false.

Another concern involves benchmarking. A number of attempts to compare performances of various gene enrichment tools have been made[34-36]. In all cases of which we are aware, however, CGLs are used as "gold standards" for analysis; a tool that best matches a leukemia ES to a cancer CGL is declared winner. Given our numerous concerns with the usage of CGLs, we refrain from this exercise.

WIMG can be classed as an ORA approach. The question arises of whether FCS overcomes the above concerns regarding CGLs with undefined backgrounds. Briefly, let us assume a cancer ES with background of 20000 and a cancer CGL with an effective background of 4000. One would then expect that 80% of truly relevant cancer genes could be absent in the CGL; such genes actually subtract from the FCS running sum. There is no reason, however, that FCS algorithms cannot be altered to accommodate CGLs and ESs with unequal backgrounds.

Future directions for WIMG include a more dynamic interface with the ability to generate relevant graphics. Ideally, the run-time of some tools should decrease. New features may include an option to analyze input data via an FCS approach. Most importantly, we believe that the sheer volume of ESs in the WIMG database creates an opportunity for AI-based analysis of input ESs and the database as a whole.

WIMG tools may be accessed at WhatIsMyGene.com. The site is funded by donations.

Methods

10532 GO-CGLs were downloaded from gsea-msigdb.org [37]. Gene sets with fewer than 32 or more than 1000 annotations were then excluded. All sets were then sorted according to the frequency with which genes are perturbed within the WIMG database. Each GO-CGL was assigned an "abundance ranking" by first assigning each gene within the set a ranking based on the gene's abundance within the whole human proteome or mouse brain proteome, and then calculating the average rank for all genes within the set. To avoid confusion, we should point out that the WIMG database actually contains 3803 external CGLs at the time of writing, but only the 2604 that have been sorted were used for downstream analysis. The GO list abundance rankings are available as Table S8.

It should be noted that the data used in fig. 1d is atypical in the sense that the authors included non-significant GO terms in their table (s4 "GOCC_UE-RAB3GAP1"). Had only significant GO terms been listed, the correlation between abundance and significance would have weakened. The exercise of correlating GO term significance with GO list abundance rankings can, of course, be performed for any input set of choice by utilizing any GO enrichment tools.

To compare rankings of GO-CGLs, WGCNA clusters, and WIMG lists against input ESs, the following steps were taken. First, studies entered into the GEO database between October and December 2021 were selected. This relatively short time-frame is intended to minimize the possibility of "cherry picking" studies that would tend to prove our points. Individual input gene lists derived from these studies were then entered into the WIMG Fisher tool. The ranks of the single highest-ranking GO-CGL, WGCNA cluster, and WIMG list were recorded, and the exercise repeated with the next study. Expected top-rankings for GO-CGLs were calculated simply by dividing the total number of gene lists in the WIMG database at the time of the test by the number of GO-CGLs in the WIMG database; the same applies for WGCNA clusters and WIMG lists. Test #1 was conducted over two background settings (all CGL backgrounds are fixed at 20000 and all CGL backgrounds are estimated using "reverse Fisher"). Test #2 was conducted using two methods of sorting input data (dividing log(fold-change) by significance and by excluded all insignificant genes and then sorting by fold-change).

Regarding the GSEA-WIMG comparison (Table S6), in the case of an experimental "cell type" designation, we required that a comparison between two cell types be made (as opposed to "liver cluster 17").

For WIMG lists, clustering is performed using a matrix of Fisher *P*-values generated by intersecting relevant studies. Optimizing cluster numbers is performed with the R fviz_nbclust function from the factoextra package. Clustering is performed with Cluster 3.0[38] using k-Means with the Euclidean distance similarity metric. The resulting .kgg file contains the names of studies according to cluster. Gene IDs

from within these studies are then tabulated. The most frequent genes in a cluster are compiled into a WIMG list and a background is assigned to the list.

For an explanation of the "reverse Fisher" method of background prediction, we will assume that a study is to be tested against a CGL. Note that Fisher's exact test requires inputs that are the intersections of 4 sets: A) the intersection of the ES with the CGL, B) the intersection of the CGL with identified genes not in the ES, C) the intersection of the ES with genes not in the CGL, D) and the intersection of genes not in the ES against genes not in the CGL. In the case of C), the value is taken to be the count of genes in the ES minus the ES/CGL intersection. D is thus the only unknown intersection. Beginning with a large value of D and halting at a value of D that renders the equation unsolvable, the equation is repeatedly solved for *P*. The value D at which *P* reaches a maximum is taken to be the correct D. The final background of a CGL is then the sum of D and the count of all genes in the CGL. The reverse Fisher algorithm was used in estimating backgrounds for 2605 GO-CGLs (Table S7).

All graphics were generated using the Matplotlib 3.7.1[39] and Seaborn 0.12.2[40] libraries executed in the Python 3.9 environment.

Competing Interests

The authors declare no competing interests.

Funding

The study is not funded by any institution.

Author's Contributions

KH performed data collection and analysis, programming, and writing. TS performed programming, data analysis, and graphics generation.

Supplemental Material:

Supplemental File 1: The WIMG Manual. Contains information regarding WIMG database content, data-entry protocols, and a description of the tools and filters found on the WIMG website.

Table S1: Data used in the construction of Fig 1b. "Final product" sheet shows gene IDs, their abundance rankings in various tissues, and frequencies of their appearance in GO-CGLs. Whole human protein data, used in Fig 1b, is highlighted in blue. "Go Frequencies" sheet shows frequencies of genes over 2604 GO-CGLs. "Tissue Abundances" shows abundance-ranked transcripts or proteins over 77 mouse and human tissues. The gene IDs are ensemble-gene IDs stripped of the initial "ENSG". The sources of the abundance data are given in the headers.

Table S2: Data used in the construction of Fig 1c. Fulcher's "CFPR_Mouse_SBPrandom" scores for 1498 GO-CGLs were merged with abundance rankings for the same GO-CGLs. Mouse brain proteome abundance rankings were used in Fig 1c; whole human and human brain rankings are also given in the table.

Table S3: Data used in the construction of Fig 1d. 210 GO-CGL's and their associated *P*-values in the relevant mouse brain proteomic study were merged with mouse brain abundance rankings for these GO-CGLs.

Table S4: Two tables detail the rankings of the top performing GO-CGLs, WGCNA clusters, and WIMG lists overlapping 124 (test1) and 132 (test2) studies entered in the GEO database between October and December of 2021. A "read me" sheet offers further explanation.

Table S5: The gene composition of all WIMG lists.

Table S6: Comparison of GSEA and WIMG outputs over 201 studies. For each study, the top 5 ranking outputs were examined for concordance with input cell type and perturbation. The table also lists WIMG keywords that occurred at log(*P*)<-10 (binomial distribution).

Table S7: Background estimates of 2605 GO-CGLs, based on whole human proteome abundance data.

Table S8: Abundance rankings for 2605 GO-CGLs. For clarity, we should point out that these rankings strongly correlate with the background-estimates in Table S7. Here, however, rankings are derived according to the whole human proteome, the human brain proteome, the human brain transcriptome, and the mouse brain proteome.

Table S9: Data underlying a comparison of actual backgrounds versus two background estimate methods over 288 studies.

REFERENCES

1. Gene Ontology C, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The Gene Ontology knowledgebase in 2023. Genetics. 2023;224(1).
2. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023;51(D1):D587-D92.
3. Thomas PD, Campbell MJ, Kejariwal A, Mi H, Karlak B, Daverman R, et al. PANTHER: a library of protein families and subfamilies indexed by function. Genome Res. 2003;13(9):2129-41.
4. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022;50(D1):D687-D92.
5. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417-25.
6. Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics. 2005;21(18):3587-95.
7. Tarca AL, Draghici S, Bhatti G, Romero R. Down-weighting overlapping genes improves gene set analysis. BMC Bioinformatics. 2012;13:136.
8. Tomczak A, Mortensen JM, Winnenburg R, Liu C, Alessi DT, Swamy V, et al. Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. Sci Rep. 2018;8(1):5115.
9. Haynes WA, Tomczak A, Khatri P. Gene annotation bias impedes biomedical research. Sci Rep. 2018;8(1):1362.
10. Davies MN, Meaburn EL, Schalkwyk LC. Gene set enrichment; a problem of pathways. Brief Funct Genomics. 2010;9(5-6):385-90.
11. Gillis J, Pavlidis P. The impact of multifunctional genes on "guilt by association" analysis. PLoS One. 2011;6(2):e17258.

12. Tamayo P, Steinhardt G, Liberzon A, Mesirov JP. The limitations of simple gene set enrichment analysis assuming gene independence. Stat Methods Med Res. 2016;25(1):472-87.
13. Fulcher BD, Arnatkeviciute A, Fornito A. Overcoming false-positive gene-category enrichment in the analysis of spatially resolved transcriptomic brain atlas data. Nat Commun. 2021;12(1):2669.
14. Edwards AM, Isserlin R, Bader GD, Frye SV, Willson TM, Yu FH. Too many roads not taken. Nature. 2011;470(7333):163-5.
15. Stoeger T, Gerlach M, Morimoto RI, Nunes Amaral LA. Large-scale investigation of the reasons why potentially important genes are ignored. PLoS Biol. 2018;16(9):e2006643.
16. Rocha JJ, Jayaram SA, Stevens TJ, Muschalik N, Shah RD, Emran S, et al. Functional unknomics: Systematic screening of conserved genes of unknown function. PLoS Biol. 2023;21(8):e3002222.
17. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics. 2015;15(18):3163-8.
18. Voskuhl RR, Itoh N, Tassoni A, Matsukawa MA, Ren E, Tse V, et al. Gene expression in oligodendrocytes during remyelination reveals cholesterol homeostasis as a therapeutic target in multiple sclerosis. Proc Natl Acad Sci U S A. 2019;116(20):10130-9.
19. Liu Y, Tian F, Li S, Chen W, Gong W, Xie H, et al. Global effects of RAB3GAP1 dysexpression on the proteome of mouse cortical neurons. Amino Acids. 2021;53(9):1339-50.
20. Padonou F, Gonzalez V, Provin N, Yayilkan S, Jmari N, Maslovskaja J, et al. Aire-dependent transcripts escape Raver2-induced splice-event inclusion in the thymic epithelium. EMBO Rep. 2022;23(3):e53576.
21. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics. 2008;9:559.
22. Mubeen S, Hoyt CT, Gemund A, Hofmann-Apitius M, Frohlich H, Domingo-Fernandez D. The Impact of Pathway Database Choice on Statistical Enrichment Analysis and Predictive Modeling. Front Genet. 2019;10:1203.
23. Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14:128.
24. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res. 2013;41(Database issue):D991-5.
25. Simillion C, Liechti R, Lischer HE, Ioannidis V, Bruggmann R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. BMC Bioinformatics. 2017;18(1):151.
26. Maleki F, Ovens K, Hogan DJ, Kusalik AJ. Gene Set Analysis: Challenges, Opportunities, and Future Research. Front Genet. 2020;11:654.
27. Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb). 2021;2(3):100141.
28. Salka K, Abutaleb K, Chorvinsky E, Thiruvengadam G, Arroyo M, Gomez JL, et al. IFN Stimulates ACE2 Expression in Pediatric Airway Epithelial Cells. Am J Respir Cell Mol Biol. 2021;64(4):515-8.
29. Wandel MP, Pathe C, Werner EI, Ellison CJ, Boyle KB, von der Malsburg A, et al. GBPs Inhibit Motility of Shigella flexneri but Are Targeted for Degradation by the Bacterial Ubiquitin Ligase IpaH9.8. Cell Host Microbe. 2017;22(4):507-18 e5.
30. Godugu K, Hay BA, Glinsky GV, Mousa SA. Discovery of novel thyrointegrin alphavbeta3 antagonist fb-PMT (NP751) in the management of human glioblastoma multiforme. Neurooncol Adv. 2023;5(1):vdac180.
31. Crow M, Lim N, Ballouz S, Pavlidis P, Gillis J. Predictability of human differential gene expression. Proc Natl Acad Sci U S A. 2019;116(13):6491-500.
32. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.

33. Mahi NA, Najafabadi MF, Pilarczyk M, Kouril M, Medvedovic M. GREIN: An Interactive Web Platform for Re-analyzing GEO RNA-seq Data. Sci Rep. 2019;9(1):7580.
34. 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.
35. Guala D, Sonnhammer ELL. A large-scale benchmark of gene prioritization methods. Sci Rep. 2017;7:46598.
36. Geistlinger L, Csaba G, Santarelli M, Ramos M, Schiffer L, Turaga N, et al. Toward a gold standard for benchmarking gene set enrichment analysis. Brief Bioinform. 2021;22(1):545-56.
37. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545-50.
38. de Hoon MJ, Imoto S, Nolan J, Miyano S. Open source clustering software. Bioinformatics. 2004;20(9):1453-4.
39. Hunter JD. Matplotlib: A 2D graphics environment. Computing in Science \& Engineering. 2007;9:90--5.
40. Waskom ML. Seaborn: statistical data visualization. Journal of Open Source Software. 2021;6:3021.