OXFORD

Briefings in Bioinformatics, 2022, **23(5)**, 1–13

https://doi.org/10.1093/bib/bbac355

Problem Solving Protocol

Systematic assessment of pathway databases, based on a diverse collection of user-submitted experiments

Annika L Gable, Damian Szklarczyk, David Lyon, João F Matias Rodrigues and Christian von Mering

Corresponding author: Christian von Mering, Department of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, 8057 Zurich, Switzerland. Tel: +41 44 635 31 47;
Fax: +41 44 635 68 64; E-mail: mering@imls.uzh.ch

Abstract

A knowledge-based grouping of genes into pathways or functional units is essential for describing and understanding cellular complexity. However, it is not always clear a priori how and at what level of specificity functionally interconnected genes should be partitioned into pathways, for a given application. Here, we assess and compare nine existing and two conceptually novel functional classification systems, with respect to their discovery power and generality in gene set enrichment testing. We base our assessment on a collection of nearly 2000 functional genomics datasets provided by users of the STRING database. With these real-life and diverse queries, we assess which systems typically provide the most specific and complete enrichment results. We find many structural and performance differences between classification systems. Overall, the well-established, hierarchically organized pathway annotation systems yield the best enrichment performance, despite covering substantial parts of the human genome in general terms only. On the other hand, the more recent unsupervised annotation systems perform strongest in understudied areas and organisms, and in detecting more specific pathways, albeit with less informative labels.

Keywords: gene set enrichment, pathways, Gene Ontology, benchmark, STRING, functional annotation

Introduction

When interpreting genome-wide functional genomics experiments, scientists often rely on statistical tests and visualization methods to gain initial systems-level insights. Functional enrichment analysis, also termed ‘gene set enrichment analysis’, is a widely used approach to accomplish this. It is typically applied to transcriptomics or proteomics profiling experiments, but is also used for perturbation screens, genome-wide association data, bioinformatics inferences and other data modalities. For the analysis, sets of genes derived from gene annotation frameworks such as the Gene Ontology [1] are tested for statistical enrichment (or overrepresentation), for example in one experimental condition versus the other. The gene sets considered for these tests typically correspond to either pathways, protein families, subcellular compartments, disease genes or other knowledge-based groupings. Testing for sets of genes instead of single genes increases statistical power because it can capture small but coherent changes in gene-level metrics, even in the presence of substantial noise or when no individual gene is by itself changing consistently [2].

Functional enrichment can be performed by a wide variety of convenient GUI-based enrichment tools, including DAVID [3], GSEA [2], Enrichr [4] and Panther [5], among others. Each tool uses one or several pathway systems or annotated gene set collections to test against.

One of the caveats in functional enrichment is that the results depend on the quality and completeness of the annotation systems used [6–8]. It is well established that there is a relatively small number of richly annotated genes, whereas the majority of genes is only sparsely annotated [7, 9, 10]. This annotation bias is likely due to study biases that can be explained and predicted for any given gene [11], and it has been shown to impact further biomedical research [7, 10]. As a recent example, Maertens, *et al.* [12] reveal systematic biases in popular annotation systems such as GO Biological Process and show that a majority of genes associated with clinical outcomes in cancer are understudied.

As of yet, there have been no comprehensive evaluations assessing the most commonly used functional annotation systems, with regard to their role as providers of gene sets for functional enrichment analysis. This may be due to the fact

Annika L. Gable is a postdoctoral researcher in the Department of Molecular Life Sciences at the University of Zurich. Her research focuses on protein–protein interaction networks, functional gene set enrichment and pathway annotation systems.
Damian Szklarczyk is a senior staff scientist at the Swiss Institute of Bioinformatics and University of Zurich. His research focuses on protein–protein interaction networks, gene orthology relations, phylogeny and functional annotation.
David Lyon is a senior data scientist at MonteRosa Therapeutics. In his previous position as senior staff scientist at the Swiss Institute of Bioinformatics and University of Zurich, his research focused on protein–protein interaction networks, functional gene set enrichment and proteomics.
João F. Matias Rodrigues is a senior staff scientist in the Department of Molecular Life Sciences at the University of Zurich. His research focuses on microbial metagenomics.
Christian von Mering is a professor and group leader at the University of Zurich and Swiss Institute of Bioinformatics. With his research group, he develops widely used tools and databases for microbial metagenomics, proteomics, gene orthology and protein–protein interaction networks, and collaborates with life science researchers. More details about his research group are available at https://www.mls.uzh.ch/en/research/von-mering/research.html.
Received: March 9, 2022. **Revised:** July 13, 2022. **Accepted:** July 30, 2022

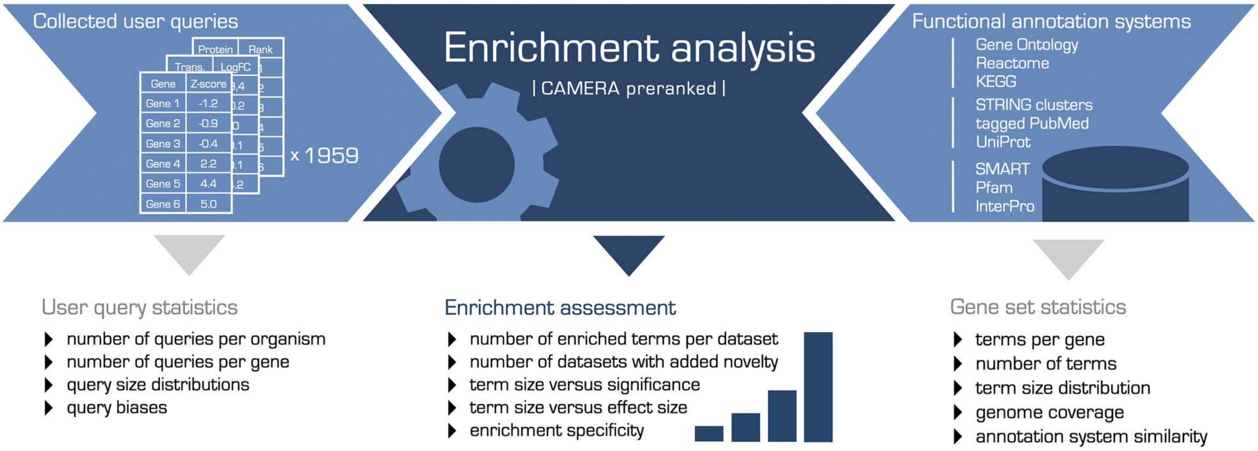

Figure 1. Assessing differences between functional pathway annotation systems. A diverse collection of 3651 enrichment query datasets is collected, with permission, from STRING users. After filtering for redundancy, 1959 datasets are resubmitted in a standardized setting for enrichment testing against established classification systems. These 11 pathway annotation systems, including the three Gene Ontology categories, are compared based on their structural properties and their performance in enrichment analysis using the CAMERA preranked method.

that these annotation databases are very heterogeneous and enrichment testing may not have been the primary objective during their original design. Instead, they each focus on a certain context, such as detailed metabolic pathway maps (KEGG [13]), curated human pathways (Reactome [14]), comprehensive annotation of protein function (UniProt [15]), identification of protein families and domains (SMART [16], InterPro [17], Pfam [18]) or formalized ontologies of gene function or cellular location on multiple levels of granularity (Gene Ontology [19]).

Three studies so far have compared functional annotation databases in specific use cases. However, these focus either on cell signaling databases only and/or base their comparison on a single use case, considering one specific disease or class of diseases:

Tieri & Nardini [20] compare the databases Reactome, KEGG, Nature Pathway Interaction Database (PID) and InnateDB based on the number of rheumatoid arthritis pathways, given verified disease gene or protein identifiers and the coherence of the results.

Chowdhury and Sarkar [21] compare 24 different cell signaling databases based on the number of pathways and entities contained, the functional areas covered, and the sources and data representation models that they are based upon.

Maleki *et al.* [6] is the only publication so far that assesses functional annotation databases for use in functional enrichment analysis. They compare the three Gene Ontology hierarchies, KEGG, BioCarta and Reactome in their similarity, number of pathways and gene set sizes. Enrichment-specific comparisons are performed for juvenile idiopathic arthritis based on three GEO datasets for this disease. They specifically derive a metric that is meant to inform the choice of a certain functional annotation system for a specific use case. However, their method relies on determining the list of genes relevant to the biological question in advance, which is not a viable approach in a more exploratory setting, and may potentially exacerbate existing research biases.

In addition to the three studies comparing annotation databases, Bateman *et al.* [8] assess not individual databases but instead the collections available in MSigDB, the gene set database that is the default option for the enrichment tool GSEA. Two of the collections cover gene sets from functional annotation databases, including Gene Ontology, and selected sets from BioCarta, KEGG, Reactome, PID and others. The comparison and performance evaluation are based on a limited number of drug response expression experiments in human cancer cell lines.

Thus, to our knowledge, there are no assessments so far that evaluate a broad range of annotation databases as providers of gene sets for functional enrichment, or that base their evaluation on more than one type of input data. Furthermore, the available assessments are almost exclusively focused on human diseases, despite functional enrichment being applied across all phenotypes and domains of life.

Here, we take advantage of the broad and diverse user base of the STRING database, in order to assess and compare 11 functional gene annotation systems. With the users' permissions, we have collected query datasets submitted anonymously to the STRING functional enrichment tool over a period of 7 months. This amounts to a benchmark collection of nearly 2000 unique, real-life query inputs, originally submitted specifically for the purpose of genome-scale enrichment testing (see Figure 1).

The STRING database (https://string-db.org) is an online resource used mainly to explore protein–protein associations, covering over 14 000 different organisms in the current version (11.5) [22]. As an accessory feature, it also allows functional enrichment analysis on user-uploaded multi-gene queries. In version 11.0, STRING considerably expanded this functionality by adding a genome-scale enrichment mode for gene–value lists and by introducing two additional functional annotation systems that are based on the STRING database itself: STRING network clusters and tagged PubMed publications. The network clusters are the result of unsupervised, hierarchical clustering of the STRING protein–protein association networks, while the tagged publications sets comprise the genes or proteins discussed in each paper based on automatic named-entity recognition.

Here, our goals are to systematically assess (i) differences in coverage depth and focus among the pathway collections, (ii) biases in user query interests and to what extent these biases may match pathway annotation biases, (iii) quantitative differences in enrichment findings among pathway systems for a given set of queries and (iv) to which extent the unsupervised annotation systems may provide complementary insights and help to address understudied organisms and biological processes.

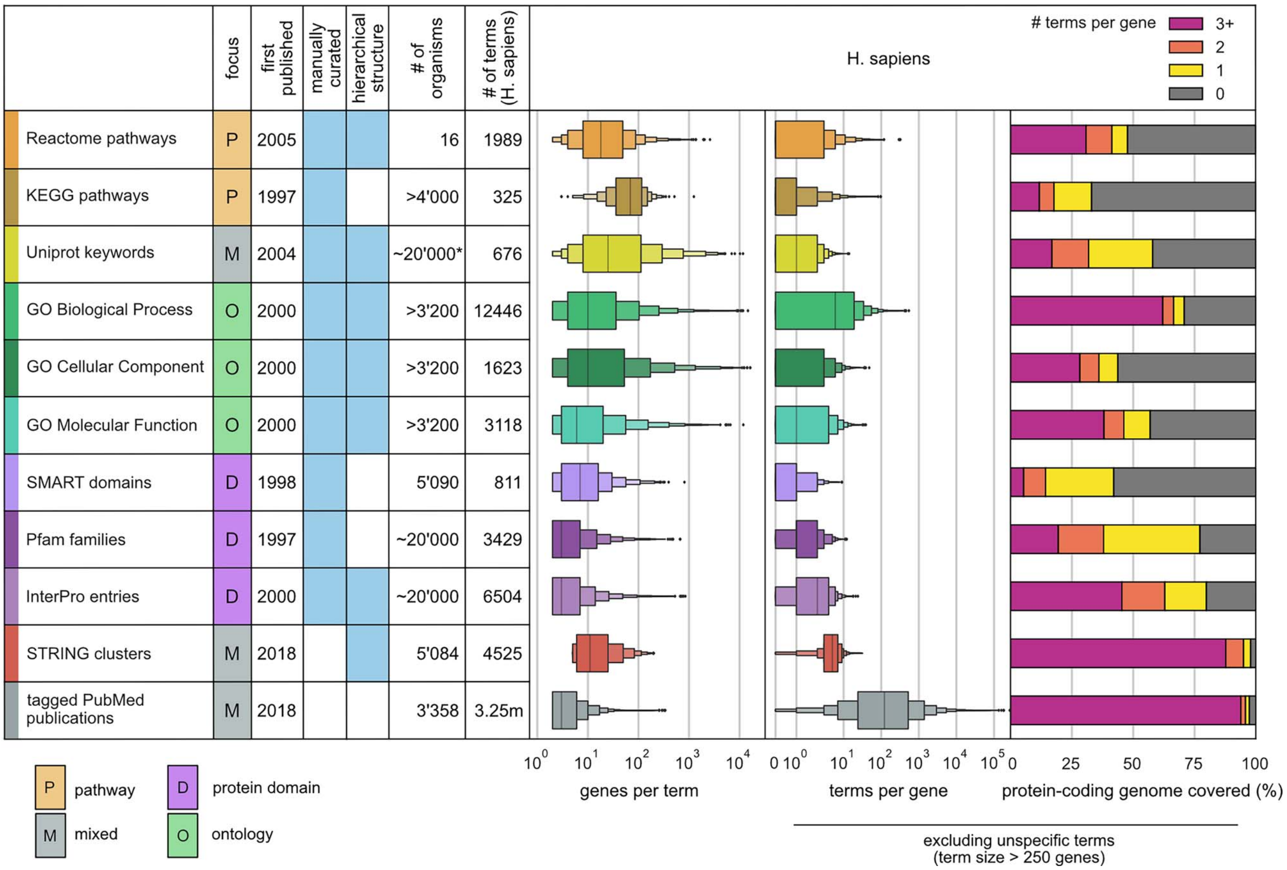

Figure 2. Functional pathway annotation systems assessed here. Basic characteristics of each annotation system are shown alongside four metrics that are specific to the human annotations: the number of terms, term size distribution, gene coverage and genome coverage. Terms here refer to pathways, keywords or any other type of gene set. A version that does not exclude large terms for the right-most two panels is shown in Figure S2 (see Supplementary Data available online at http://bib.oxfordjournals.org/). (* = reference proteomes in UniProtKB).

Methods

Functional pathway annotation systems

All analyses and comparisons are based on the functional annotation systems as mapped to and integrated into STRING v11: Reactome pathways, KEGG pathways, UniProt keywords, the three Gene Ontology (GO) hierarchies (Biological Process, Cellular Component and Molecular Function), SMART protein domains, Pfam protein families, InterPro protein families/domains, STRING clusters and tagged PubMed publications (see also Figure 2). All annotations are used as provided by the corresponding databases, except for the removal of 23 terms or keywords. These removed terms/keywords are the three root terms of the Gene Ontology ('Biological Process', 'Cellular Component' and 'Molecular Function'), and 20 of the UniProt keywords: the 10 root terms, plus the 10 'technical terms' (i.e. the children of KW-9990). The version numbers and release dates of all functional annotation systems are shown in Table S1 (see Supplementary Data available online at http://bib.oxfordjournals.org/).

User query data collection

Since version 11, STRING has provided genome-wide functional enrichment testing for all of its organisms. In addition to simple gene set overrepresentation testing, STRING also accepts gene or protein lists with experimental summary statistics such as fold changes, *P*-values, t-statistics, scores or ranks, without requiring any cutoffs. Upon submission of such a genome-scale query, users are asked for their consent to anonymized, aggregated use of their submission with an opt-out option. For this study, all 3651 submissions with consent from between July 2019 and February 2020 were used. The only information collected was the gene–value pairs provided by the user. No metadata was collected, and the user-provided data has been analyzed in aggregated form only.

Query data processing

We developed an analysis pipeline to re-process the user-submitted datasets for the assessment of functional annotation databases (detailed in Figure S1, see Supplementary Data available online at http://bib.oxfordjournals.org/). From all user inputs, we removed gene–value pairs with nonfinite or 'NA' values. After this first filter, we set the following criteria for inclusion: (i) at least 500 genes can be mapped to STRING identifiers, (ii) no values are above |1e200|, (iii) at least 10 unique input values occur and (iv) the most frequently occurring value in any given user input makes up less than 80% of all values of this input. The last two conditions were put in place in order to exclude inputs with too many repeated values, since they are not suitable for rank-based functional enrichment testing.

In addition, we aimed to identify and remove redundant submissions (a notable fraction of users submit the same or very similar queries multiple times). To achieve this, we kept only the largest user input of groups of datasets that either contained the exact same set of genes (± 2 genes), or correlated with each other with a Spearman ρ^2 correlation of 0.8 or more (either between the original or the absolute values). Specifically, the two types of

Spearman ρ^2 values and symmetric difference were calculated for each pair of user inputs. Next, we performed a separate single linkage agglomerative clustering for each of the three metrics (using $1 - \rho^2$ as a distance metric for the Spearman values). The minimum distance for each cluster was 1–0.8 for the Spearman distances and 2 for the symmetric difference. If two or more inputs fell into the same cluster in any of the three clusterings, only the largest input of each cluster was retained. For Spearman correlation, only input pairs with 100 or more genes in common were calculated, otherwise correlation was set to 0.

Functional enrichment testing

Although STRING possesses its own highly performant functional enrichment tool, we chose an already established functional enrichment method for this assessment study to ensure an independent evaluation. We used CAMERA preranked [37], an enrichment tool provided by the limma Bioconductor package (v3.42.0), used under R v3.6.3. In contrast to the standard CAMERA method, CAMERA preranked is a univariate method, meaning that it uses only a list of gene–value pairs from a genome-scale experiment, instead of an entire expression matrix. This type of analysis mode matches the user query data from our benchmark collection, where users typically provide *P*-values, fold changes or gene scores for each gene. We restricted the enrichment testing to pathways /gene sets (here: ‘terms’) that overlap with the user’s query input data by at least three genes.

The *P*-values obtained were corrected for multiple testing within each functional annotation category using the Benjamini–Hochberg procedure, as implemented within CAMERA. We considered all terms with a corrected *P*-value lower than 0.05 as significantly enriched.

For Figure 4, an additional nonredundant version of the enrichment results was produced by removing all detected terms that were subsets of another detected term within the same pathway annotation system.

Added novelty of enrichment results

In order to assess complementarity between the pathway annotation systems, terms detected as enriched by only one of the annotation systems were considered. The genes associated with each term were filtered to those actually present in the user query, and the nonredundant enrichment results were analyzed. A term was labeled as not detected by other annotation systems if at least 50% of the filtered genes of this term were not contained in any single enriched term in all other annotation systems. The number of user inputs with at least one such complementary term was plotted for each annotation database.

Coverage depth analysis

In order to calculate the coverage depth, for each functional annotation database we counted how many different terms each human protein-coding gene has been assigned to. Since very large pathways with hundreds of genes typically provide little functional insights but disproportionately affect coverage statistics, we analyzed coverage for all terms, as well as for specific terms only (terms annotating 250 genes or less).

Term size versus effect size/significance analysis

Since CAMERA provides an enrichment *P*-value but not an effect size, we separately calculated the effect size for each enriched term as (mean deviation of the term’s input ranks from the mean input value)/(maximum possible deviation from the mean input rank). Thus, the effect size of each tested annotation term ranges between 0 and 1.

For every enrichment test performed, we then plotted the effect sizes, and the *P*-values calculated by cameraPR against the respective term size. No restrictions of term size were imposed. For the resulting density plots, the enrichment results for the tagged PubMed publications were downsampled by a factor of 100 for better comparability with the results from other annotation systems.

Generating the STRING clusters

In order to generate the STRING clusters, the entire STRING protein–protein association network (v11.0) of each organism was used without any interaction score confidence cutoff. An overview of the generation process of the STRING clusters is shown in Figure S7A (see Supplementary Data available online at http://bib.oxfordjournals.org/). HPC-CLUST [23] was used to perform average linkage clustering on the distance matrix, where the distance between each protein pair was equal to 1.0 - STRING association score. Protein pairs that did not have association scores in STRING were assigned the prior probability of association (which is lower than the lowest association score stored in the database).

All hierarchical clusters from the clustering procedure were retained, that is, the hierarchical structure was not cut at a specific distance threshold to produce a ‘flat’, nonoverlapping, clustering. This means that a protein will be the member of a cluster and of all its parent clusters. However, to increase the usefulness as a functional annotation resource, we disregarded all clusters that were smaller than five proteins. To reduce the redundancy between parent and child clusters, we also retained only clusters that were at least five proteins larger than their child clusters, propagating this rule from the smallest clusters up to the root. Clusters above a size of 200 proteins were disregarded and not used in functional enrichment (neither in STRING’s functional enrichment tools nor in the functional enrichment method used in this benchmark). The full set of clusters without the cutoff at 200 is available at https://string-db.org/cgi/download.

Annotating the STRING clusters

For each STRING cluster generated above, we created an accession ID (unique within a given organism) and a description analogous to other functional annotation databases. The description was generated automatically by searching for consensus annotations based on similar terms in other functional annotation databases. Here, uniqueness was not enforced. The annotation reflects the degree of overlap (defined by F1 score) between the two best-matching terms from the Gene Ontology (Biological Process, Molecular Function, Cellular Location), Reactome, KEGG pathways, UniProt keywords, InterPro, SMART and Pfam.

We calculated the F1 score between STRING clusters and all other annotation terms. The term with the highest F1 score for a given STRING cluster was chosen as its description if 80% or more of the STRING cluster genes are contained in that term. For lower overlaps, the two terms with the highest F1 scores were selected as the description. If the overlap was lower than 40%, the cluster was labeled ‘mostly uncharacterized, incl. [term1] and [term2]’, and if the overlap was lower than 20%, the cluster was labeled, ‘uncharacterized, incl. [term1] and [term2]’ (see Figure S7B, see Supplementary Data available online at http://bib.oxfordjournals.org/). In the case of tied overlaps, the functional databases were given priority over the orthology databases. Within hierarchical databases, the smaller (more specific) term was chosen in case of equal overlap. Finally, all else being equal, the shorter textual annotation was preferred. The degree of characterization of the human STRING clusters is shown in Figure S7B and C (see Supplementary Data available online at http://bib.oxfordjournals.org/).

Tagged PubMed publications

The tagged PubMed publications were generated using the text mining channel of the STRING database itself, which maps the protein and gene names found in the abstracts and, in many cases, full texts of PubMed publications to the STRING protein space. Each publication thus becomes an annotation term, with the members being the proteins or genes mentioned in this publication as detected by the named entity recognition pipeline for STRING v11.0 [24]. Only publications with at least two genes tagged are used.

Directional similarity between annotation systems

Jaccard indices between all *H. sapiens* terms of all 11 functional annotation systems were calculated. For each annotation system pair, we defined the directional similarity as the average maximum Jaccard index between system A with system B.

$$\frac{\sum_{i=1}^{n} \max\limits_{1\leq j\leq m} j(A_i B_j)}{\|A\|}$$

This average depends on the direction of comparison, so that the directional similarity of A with B is not equal to the overlap of B with A.

Directional similarity was also applied in order to compare the 11 annotation systems based on enrichment results obtained on the set of *H. sapiens* query datasets. For each user query, the enriched genes of each detected term were used in the directional similarity calculation. The resulting similarity matrices were averaged over all query datasets.

Results

Established pathway annotation systems differ widely in number and sizes of their pathways/terms

As a first step in our assessment of the nine established and two novel functional gene annotation systems, we used three simple metrics (Figure 2): the number of terms in each database, the distributions of their term sizes (i.e. how many genes are annotated with this term) and the distribution of gene coverage depth, defined as the number of terms each gene appears in. All comparisons shown here were made for the annotations of protein-coding human genes. As gene groups can carry different designations (‘pathway’, ‘process’, ‘cluster’, ‘keyword’, etc.) the neutral word ‘term’ is used throughout.

The number of terms that are contained in each classification system varies by several orders of magnitude (Figure 2). At one end of the spectrum lies the KEGG pathways database, which tends to have a relatively small number of pathways per organism (but these pathways are in turn relatively large). Conversely, the number of PubMed publications tagged with at least two human genes vastly outnumbers all of the other classification systems (albeit at the expense of considerable redundancy).

The number of distinct terms annotated for a given gene also varies widely. Annotation systems with a deep hierarchy (such as Gene Ontology and Reactome) cover each gene multiple times, while homology-driven annotation systems using protein domains may only list a handful of terms per gene. The high variability in the depth of annotation within the system has been linked to over-estimating the significance of the enriched terms [9]. In this measure, among the functional annotation systems, the SMART domains and the STRING network clusters stand out by giving each gene an approximately equal weight in the annotation, as evidenced by the narrow distribution of coverage depths (Figure 2).

The annotation coverage depth of a given gene is not random: It is correlated with properties of the protein it codes for, such as tissue diversity, protein length, average tissue abundance, the number of literature mentions and even protein orderedness (Figure S6, see Supplementary Data available online at http://bib.oxfordjournals.org/).

Most of the human genome is not annotated in detail

None of the annotation systems fully cover, in an informative way, the entire space of human protein-coding genes. This is particularly visible when excluding relatively unspecific, large pathways, here arbitrarily defined as encompassing more than 250 genes: KEGG pathways then annotate just 33% of the human protein-coding genome, but also SMART, GO Cellular Component and Reactome each describe less than half of all genes. At the other extreme, the STRING network clusters and the tagged PubMed publications both annotate more than 97% of genes at least once (Figure 2), albeit in many cases with less clarity and authority than the manually curated classification systems.

Query interests of users are varied and show a clear bias toward universally expressed genes

While the above reveals clear differences between the functional annotation systems, it is unclear to what extent each of the systems is aligned with typical user queries. Here, a broad survey of user datasets submitted to the STRING database (originally for the purpose of rank-based enrichment testing) enabled a quantitative assessment. Of the 3651 original user queries submitted to the STRING database over a period of seven months, 1959 remained after removing redundant or improperly formatted/sized queries. Of these, 1188 (60.6%) were requesting enrichment searches for *H. sapiens*, which is arguably the organism where most of the global annotation efforts have been directed, and which we focus on here. A detailed breakdown of the user queries per organism is shown in Figure S3 (see Supplementary Data available online at http://bib.oxfordjournals.org/).

Of the queries submitted for *H. sapiens*, few cover the entire genome – presumably reflecting technical limitations during data generation, and/or pre-filtering of query data by users. Only about half of the human genes appear in more than 25% of user queries (see Figure S3, see Supplementary Data available online at http://bib.oxfordjournals.org/). The three most commonly submitted genes were PKM, ENO1 and GAPDH (see Figure S4A, see Supplementary Data available online at http://bib.oxfordjournals.org/). All three are highly expressed genes involved in glycolysis, one of the processes present indiscriminately in all types of cells, but even these appeared in less than two thirds of the queries. Whether or not a given gene appears frequently in user queries appears to be nonrandom: on a global network visualization of the entire human genome, frequently queried genes often cluster together (Figure 3A), that is, have a clear tendency to interact or to appear in related functional areas. Other genes are rarely queried by users: most notably the olfactory receptors, many

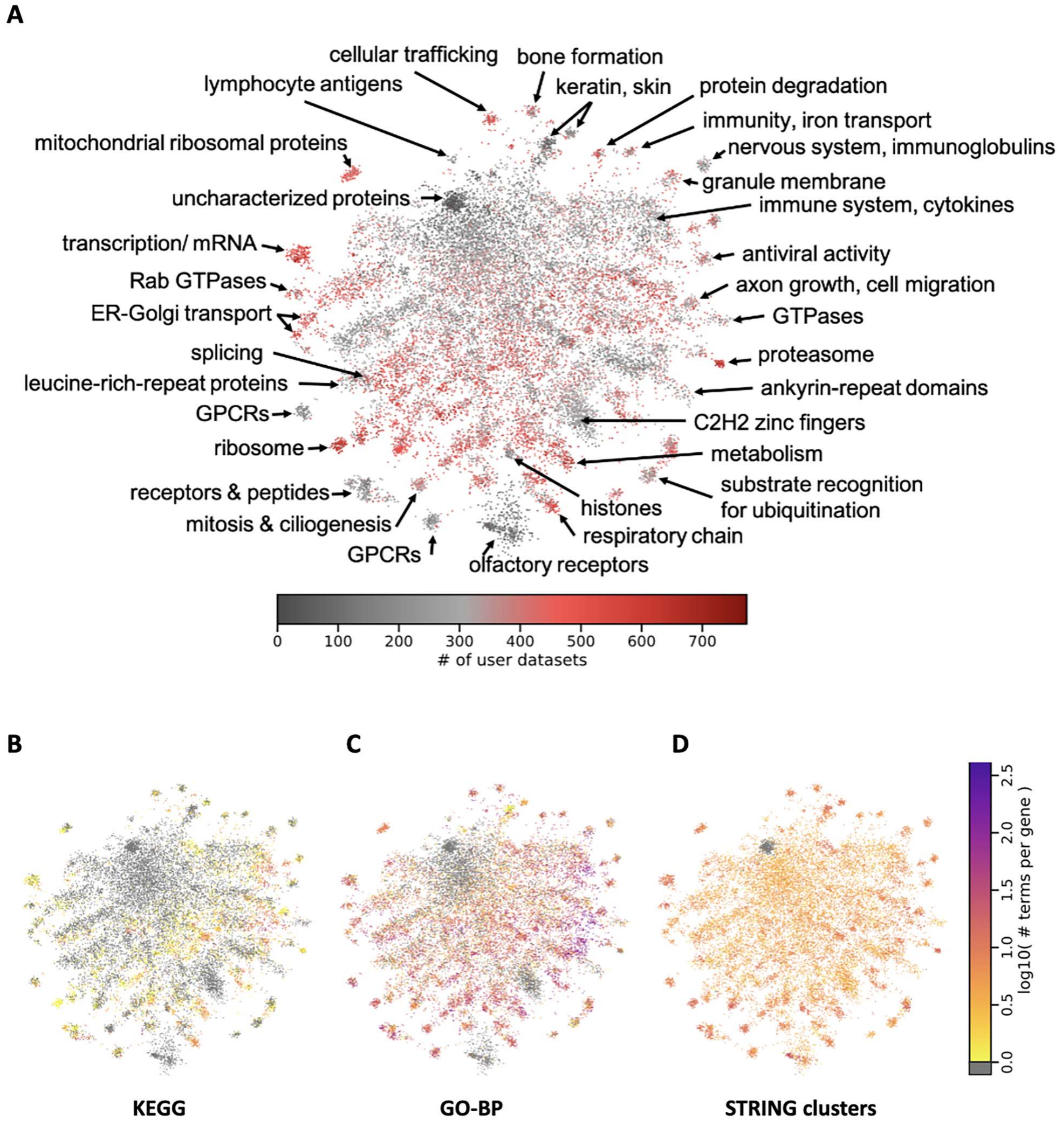

Figure 3. Differences in human genome coverage. For illustration, the entire human genome is laid out in a nonsupervised manner using the t-SNE algorithm [25]. The similarity between the points (genes) corresponds to the combined STRING interaction score. Clusters on the projection are manually labeled according to broadly shared functions. (**A**) Frequency of gene occurrence among the 1188 human user-submitted enrichment queries. (**B–D**) Coverage of the human genome by the KEGG pathways, GO Biological Process and STRING clusters annotation systems. Shown is the number of terms per gene, for terms up to a size of 250 genes. Genes not covered by the respective database are shown in gray. For further annotation systems, and a version including all term sizes, refer to Figures S9 and S10 (see Supplementary Data available online at http://bib.oxfordjournals.org/).

of the G-protein-coupled receptors (GPCRs), keratin/skin genes, immunoglobulins and genes associated with the nervous system. The most frequently studied genes are involved in transcription/RNA, ribosome, proteasome and metabolism. These trends seem to broadly follow the cell type promiscuity of the processes that the genes are involved in. In general, genes that often appear in user queries tend to be more widely expressed across tissues, more abundantly expressed in a cell, more often mentioned in the literature and tend to code for larger proteins (see Figure S4B–D, see Supplementary Data available online at http://bib.oxfordjournals.org/).

Annotated pathway coverage depth reflects study bias/research interests

Visually, pathway annotations systems also differ in their relative coverage of the human protein-coding genome (Figure 3B–D, and Figure S9, see Supplementary Data available online at http://bib.oxfordjournals.org/). As is the case for the user interests, the annotation coverage is distributed nonrandomly on the genome – with closely connected, functionally associated genes sharing similar annotation coverages, as perhaps expected.

As shown in Figure S5 (see Supplementary Data available online at http://bib.oxfordjournals.org/), the GO Biological Process and the automatically parsed biomedical publications in PubMed show the strongest correlation with user interest (occurrence frequency in the user input datasets, Spearman's rho = 0.36 and 0.37, respectively). On the other end of the spectrum, the domain-centered databases and the STRING network clusters correlate markedly less. This is perhaps to be expected, since the protein families in the domain annotation systems and the hierarchical clustering of the entire protein-coding genome are less subject to annotation based on research trends.

Several large areas of the genome are poorly covered by almost all the annotation systems. These areas include (i) the olfactory receptors, (ii) the C2H2 zinc finger transcription factors and (iii) a large group of 'uncharacterized' or 'poorly characterized' genes. The STRING network clusters (Figure 3D) have a unique position here, covering almost a complete protein-coding genome while maintaining a relatively small variation in the depth of annotation between the covered proteins. This combination of breadth and balanced depth is a unique feature among all the annotation systems, and avoids any potential study bias giving equal representation of both well- and poorly studied genes in the annotated sets, while still providing functionally associated gene sets on various levels of specificity (Figure 3B–D).

The gene sets from automatically parsed biomedical literature (tagged PubMed publications) stand out by having a raw coverage that is several orders of magnitude higher and at the same time more varied than any of the other annotation databases (Figure S9, see Supplementary Data available online at http://bib.oxfordjournals.org/). Around half of all human protein coding genes are discussed in more than 100 articles, with around 300 genes mentioned in more than 10′000 publications. While this creates a lot of redundancy, it also has an advantage: Each gene is mentioned together with several potential pathway partners, in varying combinations. This increases the chance that a 'near optimal' gene set granularity for a given gene set enrichment task is among these articles. On the other hand, the high redundancy means that the resulting *P*-values have to be heavily corrected for multiple testing. The published literature also exhibits a shared bias with the average user interests, but there are noted differences: for example, while the GPCRs are underrepresented in the user queries, they are fairly well represented among the publications. This discrepancy may be caused by the low physiological expression levels of GPCRs on the one hand and their importance in various biological processes and as drug targets on the other hand [26].

Overlap and complementarity between annotation systems

For a more quantitative comparison of the functional space covered by the various annotation systems, we calculated their directional overlap matrix (Figure S8A, see Supplementary Data available online at http://bib.oxfordjournals.org/). As expected, the domain-annotation databases (InterPro, SMART, Pfam) cluster closely together. Terms from SMART mostly find an equivalent in InterPro and Pfam, but not necessarily vice versa. Nearly all systems find relatively good equivalents in the tagged PubMed publications, likely due to the sheer number of these literature-derived terms. In the reverse, most of the PubMed terms fail to have an equivalent in the more established annotation systems, highlighting its complementary aspect. KEGG pathways stand out in that no other annotation database finds a large number of equivalent terms in KEGG – likely due to KEGG's pathway size distribution being shifted toward larger terms compared to the other resources.

Reactome pathways, GO-BP, UniProt keywords and STRING clusters have the highest discovery power for most datasets

When inspecting the number of detected terms and pathways per user query, the STRING clusters generate the highest number of terms detected as enriched. This holds true across all organisms that we received data for, except *Saccharomyces cerevisiae*, where GO Biological Process detects more terms (Figure S11A, see Supplementary Data available online at http://bib.oxfordjournals.org/). However, deeply hierarchical annotation systems are favored by such a comparison, which is why we re-analyzed the data with terms removed which are perfect subsets of other enriched terms (Figure 4A). Now, the picture becomes more variable across organisms: The human-centric Reactome database provides the most nonredundant terms in the human datasets, with GO Biological Process close behind. In other organisms, UniProt keywords and STRING clusters provide an equal or higher number of enriched terms. While Reactome performs well on human data, many organisms have not as yet been annotated by it.

The tagged PubMed publications stand out with the highest variability of enrichment results. While more than half of all user queries don't result in any literature enrichment (Figure S11B, see Supplementary Data available online at http://bib.oxfordjournals.org/), the remaining half receive up to 100′000 unique publications significantly associated with their query dataset, even after multiple testing correction.

Given the frequently observed annotation overlaps between pathways/terms within given annotation systems (e.g.in hierarchical annotation systems, or within the parsed publications), a simple count of enriched pathways is not ideal for a comparison. Several other measures are arguably of more practical relevance. For example, the performance on 'difficult' queries might be revealing, that is, on those queries for which some or all of the annotation systems fail to report any enrichment. Figure S11B (see Supplementary Data available online at http://bib.oxfordjournals.org/) shows a comparison where each annotation system is scored only once per user query—it is either successful in reporting one or more enrichments or it is not. Differences in this score arise almost exclusively due to the 'difficult' queries. Remarkably, by that metric the STRING network clusters proved to be the most successful—they most often reported significant pathway enrichments even when other annotation systems came up empty. This is observed in *H. sapiens*, as well as in many of the less well-studied organisms. In general, STRING clusters, KEGG pathways and UniProt keywords most consistently provided at least one significant term for a given experimental input.

STRING clusters and tagged PubMed publications define functional terms where no other annotation terminology can provide information

In cases where a query input has a clear enrichment signal, that signal is typically detected by multiple pathway systems. This is reassuring, but arguably those parts of the signal that are detected by only one classification system warrant special attention. Defining a detected pathway as unique to one classification system

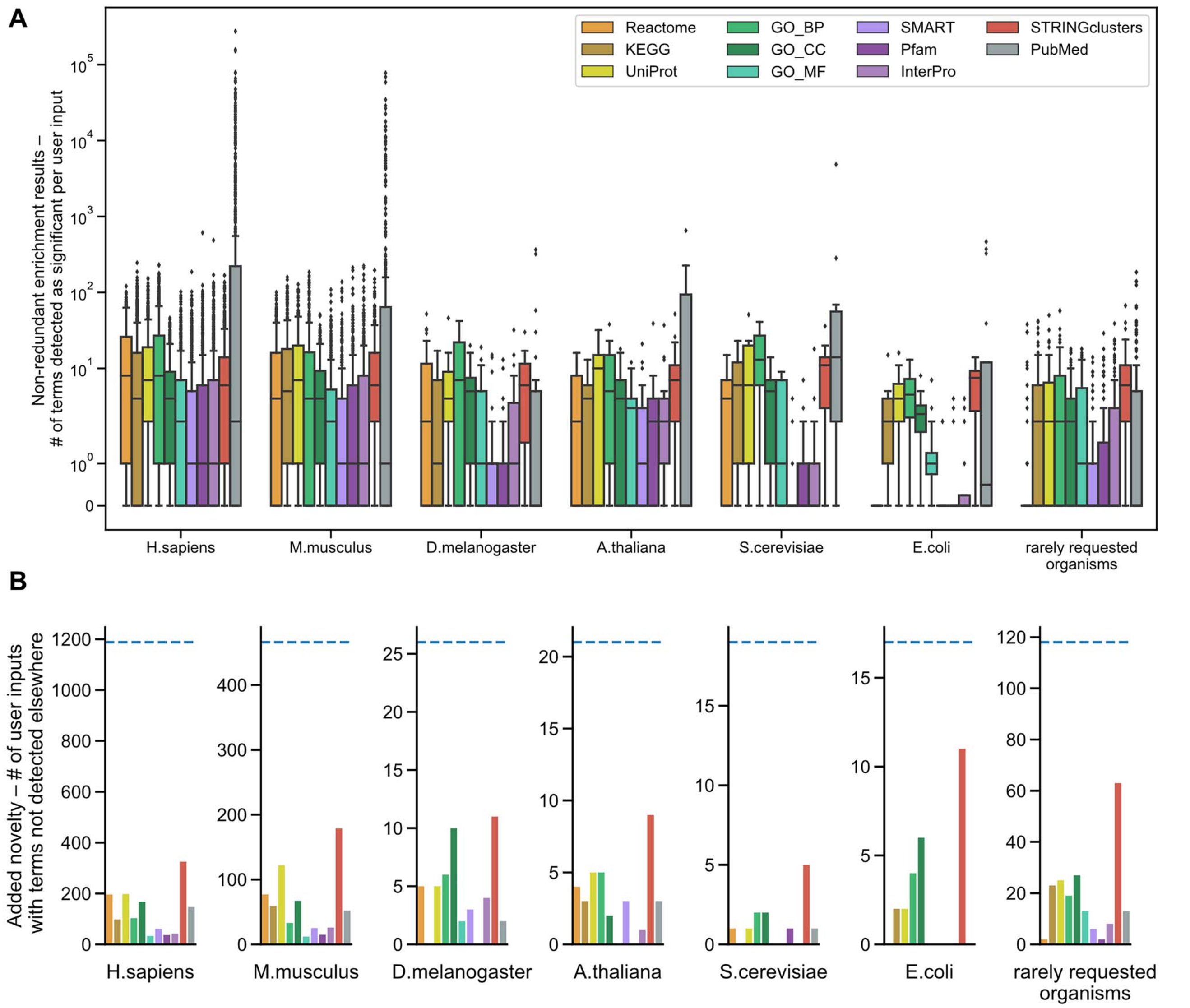

Figure 4. Enrichment performance. Comparison of functional annotation databases based on the enrichment results of all user data query submissions. Only nonredundant terms are counted here, that is those gene sets that are not subsets of other gene sets from the same annotation system. (For full results, see Figure S11, see Supplementary Data available online at http://bib.oxfordjournals.org/). User queries for which no enrichment in any category was detected are omitted. For brevity, results from *Rattus norvegicus* and *Danio rerio* are omitted. (**A**) Number of annotation terms reported as significantly enriched, per user input (symlog scale). (**B**) Added novelty provided by a given annotation system: number of user inputs for which only one annotation system detects any term as significantly enriched. The blue line represents the total number of user inputs for each species.

if less than half of its constituent genes have been detected elsewhere, about 68% of user inputs querying *H. sapiens* lead to at least one such unique detection.

In Figure 4B, we show how these unique detections are distributed between the annotation systems. We only count those user queries where the enriched term is detected by only one of the annotation systems for any given query dataset. Thus, we gain insight into the added novelty or complementarity each annotation system delivers to the interpretation. Strikingly, we see that more users receive a complementary result from the STRING clusters than from any other annotation system, across all but *S. cerevisiae*. Typically, tagged PubMed publications are enriched in less than half of the user queries, but if they are, they deliver a large number of terms not detected by other annotation systems. Effectively, both STRING clusters and tagged PubMed publications assemble genes into ‘pathways’ in an unsupervised fashion; hence they can capture functional relationships that are still actively worked on and not yet consolidated into canonical pathway knowledge.

As an example, Figure S14 (see Supplementary Data available online at http://bib.oxfordjournals.org/) highlights one of the more speculative STRING network clusters. Its automatically derived description is ‘mostly uncharacterized, incl. Retrotransposon gag protein, and Magnesium transporter NIPA’, whereby the latter two terms are derived from protein domains found in only 4 of the 32 cluster proteins. A separate overrepresentation analysis of the cluster shows that there are no terms from GO Biological Process or KEGG pathways enriched, and other databases merely show an involvement in transmembrane transport. However, the enrichment for publications listed in PubMed clearly hints at an involvement in the epigenetic process of imprinting. While the six most significantly enriched publications mention imprinting in their title, the seventh publication mentions the Prader-Willi syndrome region, which contains several imprinted genes. Overall, of the top 200 significantly enriched publications, 135 contain 'imprint' or 'Prader-Willi' in their title. In this way, STRING clusters and tagged PubMed publications can be used in tandem to identify as-of-yet unannotated functional clusters, complementing the existing annotation systems.

Enrichment results from pathway annotation systems cluster together, while literature gene sets provide complementary annotations

One of the questions one might ask when choosing which pathway annotation system(s) to use for functional enrichment analysis would be how similar to each other, or, in other words, how complementary the results generated by using different annotation systems would be. To answer this question, we generated an average directional overlap matrix for the enrichment results of our 1188 *H. sapiens* datasets (Figure S8B, see Supplementary Data available online at http://bib.oxfordjournals.org/). We observed that, as expected, the domain-annotation databases Pfam, InterPro and SMART formed the densest cluster, while the other, function-annotating databases, formed a separate cluster. Only the tagged PubMed publications did not cluster with any of the other annotation systems, perhaps pointing to a strong complementarity. Interestingly, these three clusters of annotation systems were observed more clearly in the heatmap based on enrichment results than in the one based purely on the annotation system structures (Figure S8, see Supplementary Data available online at http://bib.oxfordjournals.org/).

Annotation terms of medium size achieve the highest significance

Another complication in the performance assessment lies in the varying pathway size distributions. As an example, the KEGG pathways system contains relatively few pathways but nevertheless performs well in the above task of reporting at least one enriched pathway per input (Figure 4B, Figure S11B, see Supplementary Data available online at http://bib.oxfordjournals.org/). However, the reported pathways are often quite large. This helps statistically in the enrichment testing but may provide relatively nonspecific insights to the user. Arguably, the smaller a pathway (i.e. the fewer genes are annotated in it), the more informative a reported enrichment becomes [27], providing a clearer guide to follow-up experiments.

The relationship between enrichment detection success and pathway size is not trivial. Figure 5A shows a global summary for all submitted queries for *H. sapiens*: If a pathway is too small, it may not be detectable statistically unless the experimentally detected signal is very specific to that gene set. Conversely, if a pathway is too large, it may no longer comprise a single functional unit in the cell, diluting the signal over functionally unrelated genes (Figure 5B). Overall, the strongest statistical signal is achieved by terms encompassing around 100 genes. This effect is visible in all pathway-centered databases: KEGG pathways, Reactome pathways, Uniprot keywords as well as the three GO systems, and the STRING clusters (Figure S12A, see Supplementary Data available online at http://bib.oxfordjournals.org/). Due to the consistency of this finding, we speculate that this reflects the average breadth of the experimental signal, rather than stemming from the structure of the annotation systems. In the domain-family databases as well as in the publications, there is no single ideal term size. Across all databases, the terms that are enriched are, on average, somewhat larger than the terms that are not enriched (see Figure S12B, see Supplementary Data available online at http://bib.oxfordjournals.org/).

STRING clusters achieve the highest detection specificity

In Figure 5B, we noted how larger term sizes are more likely to be detected as enriched, and, on the other hand, how these larger terms show a relatively smaller effect size. In order to investigate how these trends affect the specificity of the enrichment results produced by the different pathway annotation systems, we assessed the relationship between the number of unique enriched genes and the term sizes, for all user queries (shown in Figure 5C). We consider more enriched genes at a small term size to signify more specific pathway annotations, and thus more fine-grained and informative enrichment results. While specificity curves can look vastly different between individual input datasets, STRING clusters provide the most specific terms for the typical input, while Reactome, UniProt, GO Biological Process and GO Cellular component provide most results with term sizes larger than 200.

Discussion

In this study, we assessed and compared 11 functional pathway annotation systems (Table 1). We highlighted their differences regarding the number of pathways (annotation terms) they contain, annotation term sizes and gene coverage depth, focusing on the human genome. We showed that fairly large parts of the human genome are still lacking informative annotations, even in well-maintained annotation systems such as the Gene Ontology. We further used a compendium of nearly 2000 diverse, user-provided quantitative datasets to assess the performance of the different annotation systems for the task of functional enrichment analysis. We observed that most databases provide at least some enriched annotation terms for a given genome-scale query of human genes, while the performance in other species is very variable across databases. We also showed a strong correlation between the annotation term size and the likelihood of a term to be detected as significantly enriched in the benchmark data. We illustrated how the two novel, STRING-specific annotation systems – STRING clusters and tagged PubMed publications – avoid some of the biases of the established gene annotation resources and can be used to potentially discover as yet uncurated pathways.

In our analysis of coverage depth, that is, the number of annotation terms assigned to a given gene, we observed a strong variability between genes in almost all annotation systems. Highly studied, more abundant, longer and more universally expressed genes are more likely to be annotated. The degree to which these gene/transcript/protein properties affect the annotation depth varies considerably among the databases. The problem of missing gene annotations on the one hand and heavy annotation of a few genes on the other hand has been described and identified as problematic for functional gene set enrichment repeatedly in recent years [7, 9–12]. While these previous studies have described the issue in terms of number of publications and number of GO terms per gene, we here also analyze the coverage distribution of several other databases. Among the 11 annotation systems assessed, we found that the tagged PubMed publications and GO Biological Process terms have the largest variability in the number of annotations a gene receives, confirming the previous studies.

On the other hand, we also found that KEGG pathways, Reactome, GO Cellular Component and SMART all annotate less than

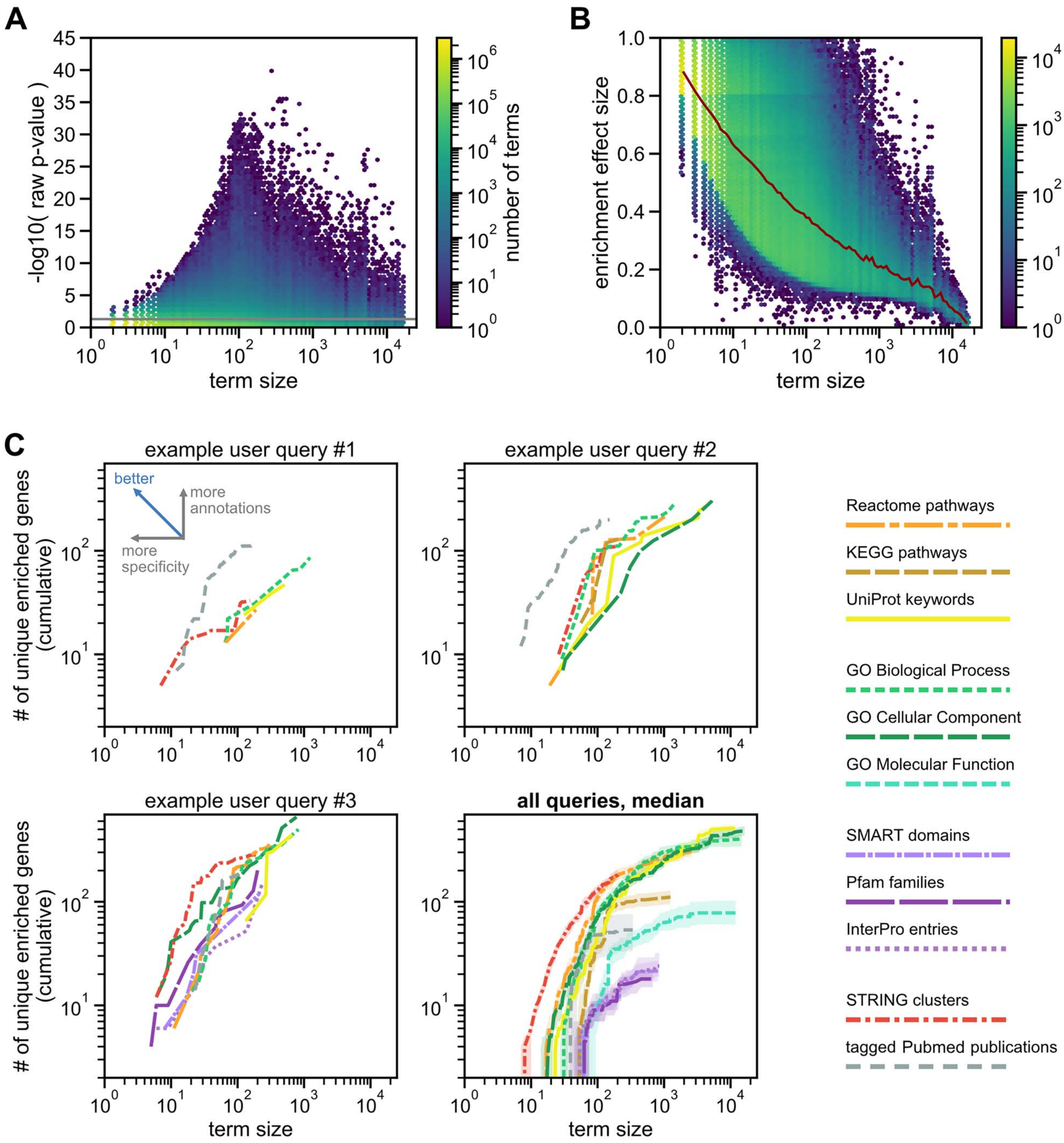

Figure 5. Specificity of enriched annotation terms in human. Relationship between annotation term size and enrichment results, based on our set of 1188 enrichment queries for *H. sapiens*. (**A**, **B**) Density plots of enrichment tests. For better visualization, no limitations on term sizes tested were imposed during enrichment testing, and tagged PubMed publication results were randomly downsampled. (**A**) Relationship between annotation term size and raw *P*-value. Gray line represents $P = 0.05$. (**B**) Relationship between enrichment effect size and term size. Spearman's rank correlation = −0.76. Red trendline: binned means. Shown are only terms with raw *P*-value < 0.05. (**C**) Unique enriched proteins per user input query, cumulated over all annotation term sizes. The lines in the combined plot represent the median of all user queries, while the shading represents the 5-percentile band around the median (47.5 and 52.5%). Enrichment at small term sizes points toward more specific terms, while enrichment at large term sizes points toward more general terms.

half of the human protein-coding genome with specific terms (i.e. terms assigned to less than 250 genes). For KEGG pathways, Reactome and SMART, this holds true even when including larger terms, while GO Cellular Component reaches a high coverage of the genome, albeit in part due to very large terms. For example, 83% of the protein-coding genome is annotated with 'cell'. The low coverage of some annotation systems may reflect intended restrictions in scope, as well as study biases in the respective scientific disciplines. To some degree the low coverage could be expected, since KEGG and Reactome focus mainly on metabolic pathways, and many other, more specific functions are not included. In general, any skews in annotation coverage may potentially turn into a vicious cycle: If genes are not annotated, they may be overlooked in functional interpretations, whereas

Table 1. Result summary

		Reactome	KEGG	UniProt	GO Biological Process	GO Cellular Component	GO Molecular Function	SMART	Pfam	InterPro	STRING clusters	tagged PubMed publications
% genome coverage	All terms	53	36	96	85	90	76	50	83	87	**98**[a]	97
	Terms ≤ 250	48	33	58	71	44	57	42	77	80	**98**[a]	97
Coverage evenness	All terms	0.8	1.8	2.1	0.4	0.8	0.7	**4.0**[a]	3.4	2.1	**4.0**[a]	0.2
	Terms ≤ 250	1.0	2.0	2.8	0.5	1.3	1.2	**4.4**[a]	3.2	2.0	4.0	0.2
Enrichment performance	Human	**8**[a]	4	7	**8**[a]	4	2	1	1	1	6	2
	Mouse	4	5	**7**[a]	4	4	2	1	1	1	6	1
	All others	0	3	4	4	3	2	0	0	0.5	**5**[a]	0
Added novelty	All terms	196	98	198	103	168	33	61	37	42	**325**[a]	147
Enrichment specificity	All terms	306	111	**512**[a]	404	483	78	22	18	24	183	54
	Terms ≤250	**190**[a]	96	119	155	133	37	17	16	18	183	54
	Terms ≤50	40	0	18	21	25	0	0	0	0	**70**[a]	22

All values are for human data unless stated otherwise. Coverage evenness = 1/variance(log(term_count/gene)). Enrichment performance = median number of enriched terms per user input, highly overlapping terms removed. Added novelty = number of enriched terms per user dataset that are not detected by another annotation system. Enrichment specificity = median number of unique enriched genes per user dataset. [a]Best performance on a measured variable.

annotated genes may be specifically mentioned in publications, and in turn more likely to be annotated for even more functions, thus producing a rich-get-richer effect [11, 12].

In contrast, the STRING clusters are relatively agnostic to the degree of studiedness, having been derived by a nonsupervised clustering of a functional association network that is itself nonsupervised (but scored and benchmarked). In these network clusters, all protein-coding genes are annotated at a similar coverage depth, which may help in combating the known problem of annotation bias. The protein–protein interactions in association networks use several information sources which are partially orthogonal to curated pathways. The STRING clusters can thereby provide annotations of potential human pathways which are not included in any of the established pathway annotation resources, and are especially valuable in as of yet underannotated organisms. A potential drawback of these clusters is that they rely on the consensus of existing functional gene annotations to receive a meaningful description. Here, however, the tagged PubMed publications, if available, can assist the user in the functional interpretation of enriched STRING clusters. Since all other annotation databases are mapped to the STRING identifier space, the fraction of the protein-coding genome covered by the other annotation terms may be underestimated to some degree.

All of the established domain family and pathway annotation databases assessed here rely on manual curation to at least some degree. Manual curation ensures good annotation quality, but it also has its drawbacks: First, a vast number of new discoveries are published every day. This means on the one hand that the rate at which published discoveries are integrated into the annotations is limited by the rate at which the literature can be reviewed by curators [28]. Second, curation efforts are often directed at specific, manually selected pathways guided by current interests in the research community [1, 29], meaning that some areas of the genome remain relatively unannotated, exacerbating study bias [10–12].

To address annotation delays, community efforts such as Wikipathways and others [28, 31] can play an important role. Study biases in annotation potentially could be ameliorated by adjusting curation strategies. Additionally, there are also approaches to account for some of the biases in pathway annotation databases during the data analysis step [9, 32, 33].

These issues of manually curated databases can be circumvented to some degree by relying on automatically generated pathway annotation systems such as the STRING clusters and tagged PubMed publications tested here. Almost the entire protein-coding genome is included in the network clusters, enabling the discovery of unannotated functional clusters and their inclusion in the interpretation of experiments, as a step toward reducing the influence of annotation bias and missing annotations. Similarly, the tagged PubMed publications (updated weekly on string-db.org) allow a direct enrichment for genes mentioned in research publications without the intermediate step of curation. It can take years or even decades until a proposed new pathway in a publication is integrated into curated pathway annotation systems [30]. With automated text mining, which is updated weekly, resources such as STRING build on new knowledge as soon as it is published.

In our assessment of the annotation databases in the context of functional enrichment, we used all genome-scale user queries submitted within a defined time span, instead of small numbers of manually selected benchmark datasets or simulated data as was done elsewhere [6, 8, 20, 21]. The major advantage of this approach is that it provides thousands of real-world scenarios of user requests, only filtered by whether they are too small or too similar to other datasets. The collection was not restricted to a certain method of data acquisition, or research area, or even specific organisms. As a drawback, we did not have a way of determining the quality and accuracy of the enrichment testing results that each user received for their dataset since we did not collect any metadata about the type and purpose of the submitted data. Furthermore, we assume that each dataset corresponds to a biological experiment or to an otherwise meaningful ranking where a functional or structural difference would be expected, and not, for example, a randomized control query or a comparison among replicates, which could expose technical rather than biological signals. We showed that the user datasets tend to contain universally expressed, large and well-studied genes more often than others. These and other trends tend to be shared between the annotation databases and the user queries. Enrichment tools should aim to inform users about potential biases and how they may influence the analysis. In this benchmark however, the query data, including any technical or biological biases they may have, serve as a realistic benchmark dataset for the functional annotation databases.

In the human genome, which is the most queried and arguably the most important research subject, the GO Biological Process and Reactome classifications provided the overall best performance among the manually curated annotation systems. Among the non-curated, automated systems, the STRING clusters provided the best coverage and the most annotations for user inputs that did not receive detected enrichments in any other annotation systems, whereas the tagged PubMed publications offered the most added novelty (both in terms of pathways not covered elsewhere, and in terms of the number and diversity of descriptions in the actual publications). These two latter systems, being non-curated and sensitive to weak signals in the underlying data, perform well in less annotated organisms and pathways, but the discovered enriched terms may suffer from lack of immediate interpretability in comparison to the manually curated systems. For a complete picture in enrichment testing, it is probably best to combine a broad array of both curated and noncurated classification systems, and to carefully browse the results while being mindful of their complementary strengths.

Key Points

- Pathway annotation systems differ widely in the number and sizes of their pathways/terms/gene sets and their organism coverage.
- There are notable study biases in most annotation systems with respect to which genes are annotated and to what depth. This has implications for the discoverability of less well-known biological functions.
- The size of an annotation gene set influences its likelihood of being detected as enriched. In *Homo sapiens*, pathways with a size of 100–200 genes tend to offer the best tradeoff between detectability and specificity.
- Where available, established pathway annotation systems such as Reactome and Gene Ontology provide excellent enrichment performance.
- Unsupervised annotation systems can help to include not-yet curated gene sets and interpret datasets from less well-studied functional areas and organisms.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Data availability

The functional gene set annotation data and aggregated query data statistics underlying this article are available in Zenodo, at https://doi.org/10.5281/zenodo.6325375. Additionally, three example user datasets can be found there. The individual user datasets cannot be shared publicly due to privacy obligations. However, specific additional summary statistics can be created upon request, and readers may send us their code to run on our user query datasets. Starting in 2023, new, individual STRING enrichment query datasets with reuse permission will be available inside the Github repository containing the benchmark code. The Snakemake benchmark pipeline is available at https://github.com/meringlab/annotation_system_assessment.

Funding

Swiss National Science Foundation (grant no. 310030_192569); Alexion Pharmaceuticals Inc. Funding for Open Access charges: University of Zurich.

References

1. Carbon S, Douglass E, Dunn N, *et al.* The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res* 2019;**47**: D330–8.
2. 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 USA* 2005;**102**:15545–50.
3. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009;**4**:44–57.
4. Kuleshov MV, Jones MR, Rouillard AD, *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* 2016;**44**:W90–7.
5. Mi H, Ebert D, Muruganujan A, *et al.* PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res* 2021;**49**:D394–D403.
6. Maleki F, Rezaei E, Ovens K, *et al.* Gene set databases: a fountain of knowledge or a siren call? ACM-BCB 2019- proc. 10th ACM Int. Conf. Bioinforma. *J Bioinform Comput Biol* 2019;**17**: 269–78.
7. Tomczak A, Mortensen JM, Winnenburg R, *et al.* Interpretation of biological experiments changes with evolution of the Gene Ontology and its annotations. *Sci Rep* 2018;**8**:5115.
8. Bateman AR, El-Hachem N, Beck AH, *et al.* Importance of collection in gene set enrichment analysis of drug response in cancer cell lines. *Sci Rep* 2014;**4**:4092.
9. Glass K, Girvan M. Annotation enrichment analysis: an alternative method for evaluating the functional properties of gene sets. *Sci Rep* 2014;**4**:1–9.
10. Haynes WA, Tomczak A, Khatri P. Gene annotation bias impedes biomedical research. *Sci Rep* 2018;**8**:1362.
11. Stoeger T, Gerlach M, Morimoto RI, *et al.* Large-scale investigation of the reasons why potentially important genes are ignored. *PLoS Biol* 2018;**16**:e2006643.
12. Maertens A, Tran VH, Maertens M, *et al.* Functionally enigmatic genes in cancer: using TCGA data to map the limitations of annotations. *Sci Rep* 2020;**10**:4106.
13. Goto S, Bono H, Ogata H, *et al.* Organizing and computing metabolic pathway data in terms of binary relations. *Pac Symp Biocomput Pac Symp Biocomput* 1997;175–86.
14. Joshi-Tope G, Gillespie M, Vastrik I, *et al.* Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 2005;**33**:D428–32.
15. Apweiler R, Bairoch A, Wu CH, *et al.* UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2004;**32**:115D–9.
16. Schultz J, Milpetz F, Bork P, *et al.* SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci U S A* 1998;**95**:5857–64.
17. Apweiler R, Attwood TK, Bairoch A, *et al.* InterPro–an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics* 2000;**16**:1145–50.

18. Sonnhammer ELL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins Struct Funct Genet* 1997;**28**:405–20.
19. Ashburner M, Ball CA, Blake JA, *et al.* Gene ontology: tool for the unification of biology. *Nat Genet* 2000;**25**:25–9.
20. Tieri P, Nardini C. Signalling pathway database usability: lessons learned. *Mol Biosyst* 2013;**9**:2401–7.
21. Chowdhury S, Sarkar RR. Comparison of human cell signaling pathway databases—evolution, drawbacks and challenges. *Database* 2015;**2015**:126.
22. Szklarczyk D, Gable AL, Nastou KC, *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* 2021;**49**:D605–12.
23. Matias Rodrigues JF, Von Mering C. HPC-CLUST: distributed hierarchical clustering for large sets of nucleotide sequences. *Bioinformatics* 2014;**30**:287–8.
24. Szklarczyk D, Gable AL, Lyon D, *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 2019;**47**:D607–D613 .
25. van der Maaten L, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
26. Corin K, Tegler LT, Koutsopoulos S. G-protein-coupled receptor expression and purification. Protein Downstr. *Methods Mol Biol* 2021;**2178**:439–67.
27. Karp PD, Midford PE, Caspi R, *et al.* Pathway size matters: the influence of pathway granularity on over-representation (enrichment analysis) statistics. *BMC Genomics* 2021;**22**:191.
28. Naithani S, Gupta P, Preece J, *et al.* Involving community in genes and pathway curation. *Database* 2019;**2019**:146.
29. Jassal B, Matthews L, Viteri G, *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res* 2020;**48**:D498–503.
30. Hanspers K, Riutta A, Summer-Kutmon M, *et al.* Pathway information extracted from 25 years of pathway figures. *Genome Biol* 2020;**21**:273.
31. Martens M, Ammar A, Riutta A, *et al.* WikiPathways: connecting communities. *Nucleic Acids Res* 2021;**49**: D613–21.
32. Mi G, Di Y, Emerson S, *et al.* Length bias correction in gene ontology enrichment analysis using logistic regression. *PLoS ONE* 2012;**7**:e46128.
33. Gaudet P, Dessimoz C. Gene ontology: pitfalls, biases, and remedies. *Gene Ontol Handb* 2017;**1446**:189–205.
34. Dosztányi Z, Csizmók V, Tompa P, *et al.* The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J Mol Biol* 2005;**347**:827–39.
35. Palasca O, Santos A, Stolte C, *et al.* TISSUES 2.0: an integrative web resource on mammalian tissue expression. *Database* 2018;**2018**:bay028.
36. Wang M, Herrmann CJ, Simonovic M, *et al.* Version 4.0 of PaxDb: protein abundance data, integrated across model organisms, tissues, and cell-lines. *Proteomics* 2015;**15**: 3163–8.
37. Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res* 2012;**40**: 1–12.