frontiers | Frontiers in Bioinformatics

TYPE Brief Research Report
PUBLISHED 15 August 2025
DOI 10.3389/fbinf.2025.1634042

OPEN ACCESS

EDITED BY
Matthew Bashton,
Northumbria University, United Kingdom

REVIEWED BY
Nejc Umek,
University of Ljubljana, Slovenia
Wimalanathan Kokulapalan,
Atalanta Therapeutics, United States

*CORRESPONDENCE
Rhys A. Farrer,
✉ r.farrer@exeter.ac.uk

RECEIVED 23 May 2025
ACCEPTED 24 July 2025
PUBLISHED 15 August 2025

CITATION
Golden C, Studholme DJ and Farrer RA (2025) DIAMOND2GO: rapid Gene Ontology assignment and enrichment detection for functional genomics.
Front. Bioinform. 5:1634042.
doi: 10.3389/fbinf.2025.1634042

DIAMOND2GO: rapid Gene Ontology assignment and enrichment detection for functional genomics

Christopher Golden[1], David J. Studholme[2] and Rhys A. Farrer[1]*

[1]Medical Research Council Centre for Medical Mycology at the University of Exeter, Department of Biosciences, Faculty of Health and Life Sciences, Exeter, United Kingdom, [2]Biosciences, University of Exeter, Exeter, United Kingdom

DIAMOND2GO (D2GO) is a high-speed toolset for assigning Gene Ontology (GO) terms to genes or proteins based on sequence similarity. Leveraging the ultra-fast alignment capabilities of DIAMOND, which is 100 to 20,000 times faster than BLAST, D2GO enables rapid functional annotation of large-scale datasets. D2GO maps GO terms from pre-annotated sequences in the NCBI non-redundant database to query sequences. During benchmarking, D2GO assigned over 2 million GO terms to 98% of 130,184 predicted human protein isoforms in under 13 min on a standard laptop. In addition to annotation, D2GO includes an enrichment analysis tool that allows users to identify significantly overrepresented GO terms between subsets of sequences. We compared D2GO against two widely used tools, Blast2GO and eggNOG-mapper, and observed substantial differences in the number and type of annotations produced. These discrepancies reflect varying sensitivities and specificities across tools and suggest that using multiple methods in tandem may improve overall annotation coverage. D2GO is open-source and freely available under the MIT license at https://github.com/rhysf/DIAMOND2GO.

KEYWORDS

functional genomics, software, diamond, Gene Ontology, enrichment analyses

Introduction

The Gene Ontology (GO) provides a structured, controlled vocabulary for describing gene product functions across species (Ashburner et al., 2000). GO is organized into three primary categories: molecular function (MF), which describes the specific biochemical activities of gene products; cellular component (CC), which indicates where these activities occur in the cell; and biological process (BP), which captures broader physiological events involving multiple molecular activities (Ashburner et al., 2000). GO terms are arranged in a loosely hierarchical structure, where more specific "child" terms are linked to broader "parent" terms. For example, the MF for GO:0004375 glycine dehydrogenase (decarboxylating) activity is a more-specific child of GO:0003824 catalytic activity. A single child term may belong to several parent terms, reflecting the complex and interconnected nature of biological functions.

GO is developed and maintained by the GO Consortium (Ashburner et al., 2000), which curates the GO knowledgebase (Aleksander et al., 2023) as part of a larger initiative by the Open Biological and Biomedical Ontologies (OBO) Foundry (Smith et al., 2007).

The OBO Foundry oversees a wide range of ontologies, such as the Cell Ontology, the Foundational Model of Anatomy, and the Plant Ontology (Ashburner et al., 2000). Although the GO is a widely used framework for functional annotation, other resources such as Pfam (Protein Families) (Finn et al., 2014) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) also provide functional insights. Functional annotations are often inferred for newly predicted genes or proteins through sequence similarity, comparative genomics, or structural features.

Several tools have been developed to assign GO terms to protein or nucleotide sequences, with Blast2GO (B2GO) being one of the most widely used. B2GO applies sequence similarity searches using either BLAST or the faster DIAMOND algorithm (Buchfink et al., 2015) to identify homology between input sequences and experimentally annotated proteins (Conesa et al., 2005). B2GO supports searches against custom user-defined databases or established reference datasets such as the NCBI non-redundant (nr) database or UniProtKB/Swiss-Prot (Bairoch and Apweiler, 2000). These searches can be run locally or through remote services like CloudBLAST (Matsunaga et al., 2008), the NCBI QBLAST server, or Amazon Web Services (AWS) BLAST. GO term assignment in B2GO is based on a multi-step algorithm that integrates sequence similarity results with InterProScan (Hunter et al., 2009) domain predictions. The assignment process considers factors such as alignment quality, coverage of the query-hit match, the source and curation level of the database, the structure of the GO hierarchy, and the annotation evidence of the matched sequences.

In addition to the GO-term assignment, B2GO offers a suite of features for visualization and functional analysis. B2GO is integrated into the broader OmicsBox platform (Bioinformatics Software OmicsBox, 2024), which provides a user-friendly graphical user interface for genome and transcriptome analysis—particularly appealing to researchers without formal bioinformatics training. By default, B2GO performs similarity searches using online BLAST services, typically against the nr protein database. Although this approach provides access to a broad and curated reference set, it is often slow, with searches taking several minutes per sequence or batch, posing a significant bottleneck for large datasets. Although users can install a local database to improve performance, this option requires substantial storage, setup time, and technical expertise. For large-scale queries, such as complete proteomes from newly sequenced genomes, this limitation can significantly delay downstream analyses. Another important consideration is that B2GO/OmicsBox is no longer freely available. Although a 7-day free trial is offered, continued use requires a paid license, even for academic users.

Beyond B2GO, a range of other tools and approaches have been developed for the GO-term assignment. Some leverage machine learning models trained on diverse features such as predicted protein domains, GO-term co-occurrence patterns, and phylogenetic profiles (You et al., 2018; Fa et al., 2018). These methods aim to improve functional prediction accuracy by incorporating biological context beyond direct sequence similarity. eggNOG-mapper (Huerta-Cepas et al., 2017) is a widely used tool that assigns GO terms based on precomputed orthology relationships from the EggNOG database. By mapping query sequences to orthologous groups, it infers function through high-confidence evolutionary relationships. In contrast, Wei2GO (Reijnders, 2022) combines DIAMOND and HMMScan searches against the UniProtKB and Pfam databases to assign GO terms based on both sequence similarity and conserved domain architecture. Both eggNOG-Mapper and Wei2GO are freely available and open-access, making them accessible options for a wide range of genome annotation projects.

The Critical Assessment of Functional Annotation (CAFA) is a community-driven challenge designed to evaluate computational methods for protein function prediction. CAFA uses a time-delayed evaluation framework, in which predictions are submitted prior to the release of new experimental GO annotations, which are then used to benchmark a ground-truth set (Zhou et al., 2019). This approach enables objective and standardized comparison of annotation tools. In the most recent CAFA3 assessment, modest improvements in prediction performance were observed between 2016 and 2019 for MF and BP categories, but not for CC. The top-performing method in CAFA3 was GOLabeler, a machine learning-based approach that integrates features such as GO term frequency, sequence alignment, and amino acid trigrams (You et al., 2018). However, GOLabeler is currently not publicly available, and its official website has been offline since at least August 2024.

Once GO terms are assigned, they can be leveraged in a variety of downstream analyses using a wide range of dedicated tools (Shahzad et al., 2015). For example, PANTHER facilitates evolutionary and functional classification of protein-coding genes, allowing researchers to explore gene families, pathways, and biological processes across species (Thomas et al., 2022). The AmiGO web portal enables users to search, filter, visualize, and analyze GO annotations within the official GO database (Carbon et al., 2009). For more advanced modeling, GO-Causal Activity Modeling (GO-CAM) links individual GO terms into structured, interpretable networks that represent biological pathways and regulatory relationships (Thomas et al., 2019).

The release of DIAMOND in 2014 introduced a major leap in sequence alignment speed, achieving protein and translated DNA alignments 100 to 10,000 times faster than BLAST (Buchfink et al., 2015). This performance improvement offers a substantial advantage for time-consuming GO annotation workflows such as those used by B2GO. Additional gains in efficiency can be achieved by restricting searches to only those database sequences with existing GO annotations, reducing the search space to under 1 gigabase, and avoiding unnecessary alignments. By integrating the speed of DIAMOND with a lightweight, purpose-built annotation pipeline, we developed DIAMOND2GO (D2GO). D2GO can assign millions of GO terms to hundreds of thousands of proteins within minutes on a standard laptop (see Implementation section), offering an accessible and scalable solution for genome-wide functional annotation.

Implementation

All analyses in this study were conducted on a 2021 MacBook Pro equipped with an Apple M1 Max CPU and 64 GB RAM. The NCBI non-redundant database was downloaded on 14 May 2023 and pre-processed to remove non-printable ASCII characters. Associated gene and GO-term mappings were obtained from the

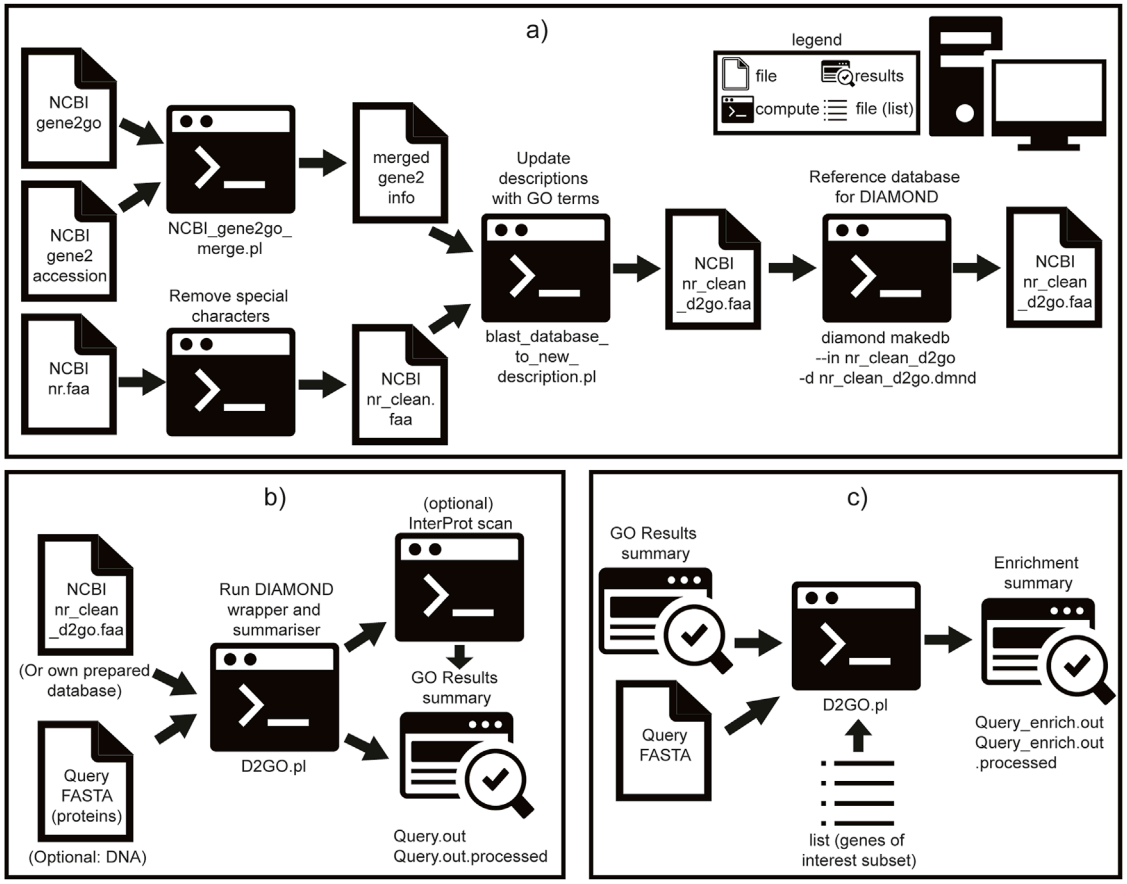

FIGURE 1
Schematic overview of the steps to construct a new database for DIAMOND2GO (D2GO), run D2GO, and perform enrichment analysis. Symbols are described in the embedded legend. All scripts and steps are included as part of software. **(a)** Pre-prepared D2GO database, **(b)** run D2GO, and **(c)** run D2GO enrichment.

NCBI gene2accession and gene2go files on 20 July 2023 (https://ftp.ncbi.nih.gov/gene/DATA/).

GO terms were merged with gene accessions using the D2GO utility script ncbi_gene2go_merge.pl and added to sequence descriptions via blast_database_to_new_description.pl. The resulting annotated database was indexed using DIAMOND makedb (Buchfink et al., 2015). This pre-processed and indexed database is included in the DIAMOND2GO GitHub repository as a Git Large File Storage (LFS) file; Git LFS must be installed as a prerequisite. Instructions for rebuilding this database are provided in Figure 1a. A Docker container is provided, with all required dependencies pre-installed, and is configured to run D2GO without additional setup. Instructions for using the container are included in the GitHub README. D2GO is freely available under the MIT license from https://github.com/rhysf/DIAMOND2GO.

D2GO acts as a lightweight wrapper around DIAMOND, adding mapping and output-parsing functionality. It currently supports both protein (BLASTP) and translated nucleotide (BLASTX) searches. The default settings include sensitivity set to ultra-sensitive, E-value cutoff: 10^{-10}, and max target sequences: 1. A schematic of the D2GO pipeline is shown in Figure 1b.

The D2GO functional annotation pipeline consists of four main steps: 1. DIAMOND alignment is performed using user-defined parameters (or defaults listed above). 2. Result summarization generates a tab-delimited file containing the gene name, species, associated GO terms, and evidence codes. GO terms tagged with NOT (e.g., NOT located_in) are excluded for clarity. Redundant GO terms per gene are collapsed by retaining only the match with the lowest E-value. 3. (Optional) InterProScan preparation: queries are filtered (optionally selecting only DIAMOND non-hits), STOP codons are removed, and sequences are batched into groups of 500. 4. (Optional) InterProScan annotation is run using the bundled iprscan5.pl script. The results are parsed and merged with the previous D2GO output to produce a combined annotation.

D2GO was tested on two datasets: 1. All predicted human proteins and splice variants from NCBI GenBank (GRCh38.p14, assembly: GCA_000001405.29) (Lander et al., 2001; International Human Genome Sequencing Consortium, 2004). 2. All predicted proteins of the chytrid fungus *Batrachochytrium salamandrivorans* (*Bsal*), downloaded from GenBank (assembly: GCA_002006685.2) (Wacker et al., 2023). The search parameters used were as follows: sensitivity set to ultra-sensitive, max target

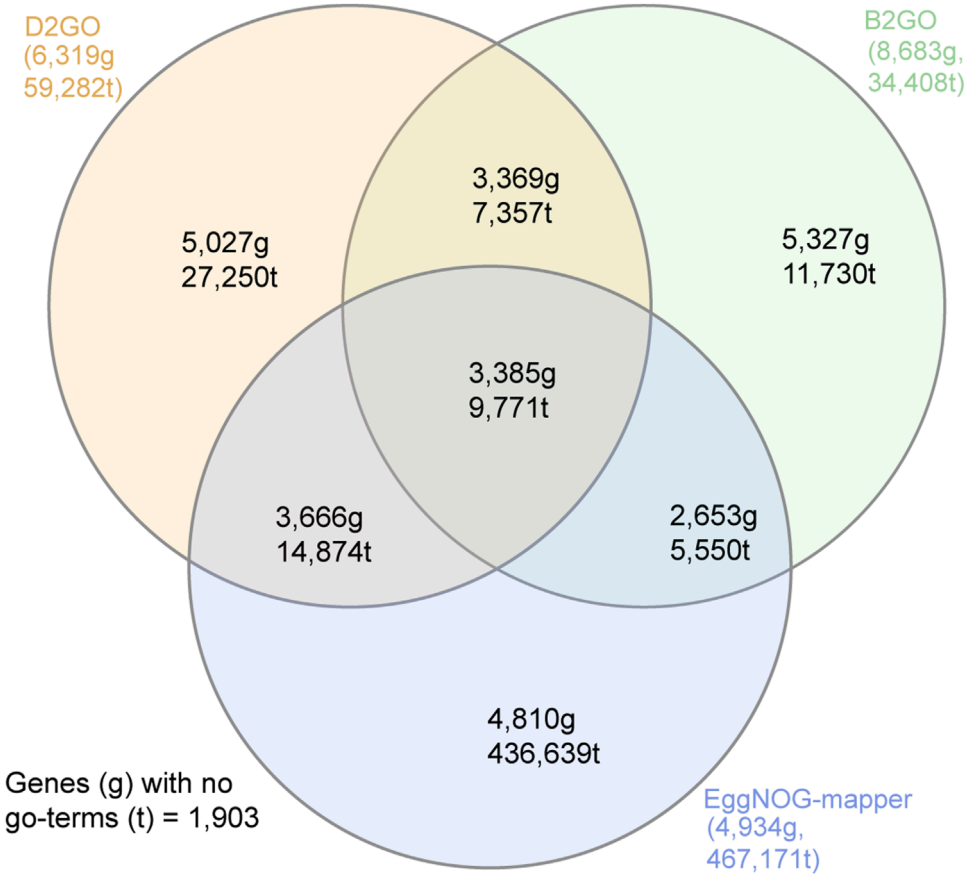

FIGURE 2
Venn diagram showing the number of *Bsal* genes (g) that were assigned GO terms by D2GO (recommended parameters), B2GO, and eggNOG-mapper. The total number of GO terms (t) that were assigned to *Bsal* genes by each tool is also shown.

sequences: 1, and E-value cutoff: 1e-5. For comparative benchmarking, B2GO v6.0.3 was run on the *Bsal* dataset as previously described (Wacker et al., 2023). eggNOG-mapper v2.1.12–3 was used with the following parameter values: m diamond--decorate_gff--excel--cpu 12 --override. InterProScan was integrated using the bundled iprscan5.pl script. A Venn diagram showing the overlap of results among tools was generated using InteractiVenn (Heberle et al., 2015). For GO-term enrichment analysis, significance was determined using a two-tailed Fisher's exact test followed by Storey–Tibshirani FDR correction (q-value <0.05) (Storey and Tibshirani, 2003).

Results

We present DIAMOND2GO (D2GO), a fast and open-access tool for assigning GO terms without subscription or license costs. To assess its performance, we annotated all 130,184 predicted human proteins, including splice variants. D2GO assigned 2,060,956 GO terms to 127,625 proteins (>98%) in just 12 min and 35 s on a laptop (see Implementation). In comparison, Blast2GO (B2GO) required several days on the same dataset, while EggNOG-mapper completed the task in 29 min and 38 s.

We evaluated D2GO alongside B2GO and eggNOG-mapper using the *Bsal* proteome (Wacker et al., 2023), running D2GO with eight different parameter sets (see Supplementary Table S1). Using the compare_go_tools.pl script (bundled with D2GO), we assessed both the number of genes with GO terms and the total number of GO terms assigned. D2GO assigned between 68,236 and 203,241 GO terms to between 5,756 and 7,729 genes (53%–71% of all 10,867 genes), depending on the parameters used. eggNOG-mapper assigned 467,171 GO terms to 4,934 genes (45%). B2GO assigned 34,408 GO terms to 8,683 genes (80%). Interestingly, eggNOG-mapper assigned the highest number of unique GO terms (88%–94% not found in other tools), followed by D2GO (45%–69%) and B2GO (28%–36%).

D2GO's annotation output varied depending on sensitivity and alignment parameters. Using e-value < 1e-5 and max target = 1, GO terms were assigned to 57% of genes. Switching to ultra-sensitive alignment increased coverage to 64% but increased runtime from <10 s to approximately 9.5 min. Based on this trade-off, we set ultra-sensitive, e-value = 1e-5, and max target = 1 as the default parameters, balancing speed and coverage. Gene and GO term overlaps with B2GO and eggNOG-mapper, using these defaults, are shown in Figure 2.

D2GO includes a wrapper for InterProScan, allowing extended annotation of sequences that lack DIAMOND hits. InterProScan added GO terms to an additional 7% of genes (increasing the total to 71%). However, the runtime increased substantially: ~3 h for only DIAMOND non-hit genes and ~17 h when run on all genes. Minimal additional GO term overlap was observed between D2GO and B2GO (Supplementary Table S1). Thus, although InterProScan expands coverage, it significantly reduces D2GO's speed advantage. We recommend using it only when completeness is prioritized over runtime.

To evaluate annotation quality, we examined results for a well-characterized gene: DNA polymerase alpha subunit pol12 (BSLG_001742). Both D2GO and B2GO identified the same three core GO terms. D2GO with InterProScan identified two additional parent terms (1 BP and 1 CC) (Supplementary Table S2). Even though B2GO matched a different *Bsal* gene (BSLG_01791) and D2GO was assigned to *Xenopus laevis* (due to *Bsal* absence in gene2go), the assigned GO terms were identical, highlighting D2GO's effectiveness despite differences in database content.

To test downstream usability, we used the D2GO utility test_enrichment.pl on all human genes with the term "polymerase" in their FASTA description (see Figure 1C). The top 10 enriched GO terms were all polymerase-related (Supplementary Table S3), confirming the tool's ability to detect biologically relevant signals and suggesting utility for less-characterized gene groups.

Discussion

D2GO offers a fast and accessible alternative to existing functional annotation tools such as B2GO and eggNOG-mapper. In our benchmark, D2GO demonstrated significantly improved speed over both tools while also being freely available, like eggNOG-mapper. This makes it particularly suitable for large-scale projects or rapid exploratory analysis where annotation time can be a limiting factor.

Without a gold standard for GO annotations, it remains difficult to determine absolute accuracy when discrepancies arise among tools. However, the variability we observed in GO-term assignment across D2GO, B2GO, and eggNOG-mapper underscores that each tool carries unique sensitivity and specificity profiles. For instance, we noted that D2GO's results were sensitive to user-defined parameters: increasing the number of accepted hits increased GO-term coverage but also yielded a higher number of uniquely assigned terms, potentially reflecting lower specificity. In contrast, restricting to top-hit matches or applying more stringent E-value thresholds led to fewer assignments and better consistency with other tools, indicating a trade-off between coverage and stringency. D2GO will also benefit from independent evaluation in future community-driven assessments, such as CAFA, to establish its predictive value under standardized benchmarks.

Our results suggest that consensus-driven annotation, where results from multiple tools are combined, may improve confidence in GO-term assignments, an approach that is increasingly advocated (Gaudet et al., 2011; Salzberg, 2019). In this context, D2GO could be integrated as a complementary tool within existing annotation pipelines. For example, workflows incorporating multiple annotation sources—either in a voting system or tiered schema—could leverage D2GO's speed for initial or broad annotation sweeps, followed by more conservative refinement with tools like B2GO or domain-specific curation. By enabling faster and higher-throughput GO annotations, D2GO can accelerate downstream applications such as functional enrichment analysis, pathway discovery, and systems biology investigations. These annotations play a critical role in diverse fields, including disease genomics and biomarker discovery (Škorjanc et al., 2025; Köhler et al., 2009).

D2GO's support scripts for enrichment analysis streamline integration into multi-omics workflows, making it a practical tool not only for standalone use but also for complementing more complex pipelines. With growing reliance on automated annotation in metagenomics, environmental sequencing, and transcriptomic analysis, D2GO's balance of speed, flexibility, and usability positions it as a valuable addition to the functional annotation toolkit.

In conclusion, D2GO expands the landscape of GO annotation tools by offering a fast, open-source alternative with customizable parameters and enrichment-ready output. Its utility lies not only in its speed but also in its potential to complement existing tools, inform consensus-based annotations, and support a wide range of biological and biomedical research workflows.

Data availability statement

Publicly available datasets were analyzed in this study. These data can be found at https://github.com/rhysf/Diamond2GO.

Author contributions

CG: writing – original draft, writing – review and editing, investigation, software. DS: methodology, investigation, supervision, writing – original draft, software, writing – review and editing. RF: writing – review and editing, funding acquisition, software, conceptualization, writing – original draft, methodology, supervision, investigation.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. RAF is part of the Medical Research Council Centre for Medical Mycology MR/N006364/2. RAF was supported by a Wellcome Trust Career Development Award (225303/Z/22/Z). DJS was supported by the UKRI BBSRC grant BB/V014609/1, "Deciphering pathogenicity and development in obligate downy mildew pathogen using small RNA approach".

Acknowledgments

The authors would like to thank Chris Desjardins for Perl code used in testing enrichment.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that Generative AI was used in the creation of this manuscript. Some portions of the text, including the Introduction, Implementation, Results, and Discussion, were revised using OpenAI's ChatGPT (model: GPT-4.5-turbo, July 2025 release). All outputs were critically reviewed, fact-checked, and edited by the authors to ensure accuracy.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2025.1634042/full#supplementary-material

SUPPPLEMENTARY TABLE S1
GO-terms assigned to Bsal proteins by D2GO using different settings, with or without InterPro (N=no, no_hits=run on genes with no DIAMOND hits, all=run on all genes). Overlap with B2GO is presented for comparison.

SUPPPLEMENTARY TABLE S2
GO-terms identified for the DNA-directed DNA polymerase alpha subunit pol12 (locus ID BSLG_001742) in Bsal by B2GO and D2GO (using InterProScan on all genes). GO-term sub-ontologies are Biological Process (BP), Cellular Component (CC) and Molecular Function (MF). Parental terms of other GO terms identified for this gene are shown.

SUPPPLEMENTARY TABLE S3
Top 10 significant GO terms for human genes that include the word 'polymerase' in the description.

References

Aleksander, S. A., Balhoff, J., Carbon, S., Cherry, J. M., Drabkin, H. J., Ebert, D., and The Gene Ontology Consortium (2023). The gene ontology knowledgebase in 2023. *Genetics* 224, iyad031. doi:10.1093/genetics/iyad031

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi:10.1038/75556

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* 28, 45–48. doi:10.1093/nar/28.1.45

Bioinformatics Software OmicsBox (2024). BioBam. Available online at: https://www.biobam.com/omicsbox/(Accessed February 1, 2024).

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nat. Methods.* 12, 59–60. doi:10.1038/nmeth.3176

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., et al. (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics* 25, 288–289. doi:10.1093/bioinformatics/btn615

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. doi:10.1093/bioinformatics/bti610

Fa, R., Cozzetto, D., Wan, C., and Jones, D. T. (2018). Predicting human protein function with multi-task deep neural networks. *PLOS ONE* 13, e0198216. doi:10.1371/journal.pone.0198216

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., et al. (2014). Pfam: the protein families database. *Nucleic Acids Res.* 42, D222–D230. doi:10.1093/nar/gkt1223

Gaudet, P., Livstone, M. S., Lewis, S. E., and Thomas, P. D. (2011). Phylogenetic-based propagation of functional annotations within the gene ontology consortium. *Briefings Bioinforma.* 12, 449–462. doi:10.1093/bib/bbr042

Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P., and Minghim, R. (2015). InteractiVenn: a web-based tool for the analysis of sets through venn diagrams. *BMC Bioinforma.* 16, 169. doi:10.1186/s12859-015-0611-3

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., et al. (2017). Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Mol. Biol. Evol.* 34, 2115–2122. doi:10.1093/molbev/msx148

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., et al. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Res.* 37 (Database), D211–D215. doi:10.1093/nar/gkn785

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. *Nature* 431, 931–945. doi:10.1038/nature03001

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30. doi:10.1093/nar/28.1.27

Köhler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dölken, S., Ott, C. E., et al. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am. J. Hum. Genet.* 85, 457–464. doi:10.1016/j.ajhg.2009.09.003

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., et al. (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860–921. doi:10.1038/35057062

Matsunaga, A., Tsugawa, M., and Fortes, J. (2008). "CloudBLAST: combining MapReduce and virtualization on distributed resources for bioinformatics applications," in *2008 IEEE fourth international conference on eScience*, 222–229.

Reijnders, MJMF (2022). Wei2GO: weighted sequence similarity-based protein function prediction. *PeerJ* 10, e12931. doi:10.7717/peerj.12931

Salzberg, S. L. (2019). Next-generation genome annotation: we still struggle to get it right. *Genome Biol.* 20, 92. doi:10.1186/s13059-019-1715-2

Shahzad, M., Ahsan, K., Nadeem, A., and Sarim, M. (2015). Gene ontology tools: a comparative study. *J. basic Appl. Sci.* 11, 619–629. doi:10.6000/1927-5129.2015.11.83

Škorjanc, A., Smrkolj, V., and Umek, N. (2025). GOReverseLookup: a gene ontology reverse lookup tool. *Comput. Biol. Med.* 191, 110185. doi:10.1016/j.compbiomed.2025.110185

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., et al. (2007). The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* 25, 1251–1255. doi:10.1038/nbt1346

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U. S. A.* 100, 9440–9445. doi:10.1073/pnas.1530509100

Thomas, P. D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L.-P., and Mi, H. (2022). PANTHER: making genome-scale phylogenetics accessible to all. *Protein Sci.* 31, 8–22. doi:10.1002/pro.4218

Thomas, P. D., Hill, D. P., Mi, H., Osumi-Sutherland, D., Van Auken, K., Carbon, S., et al. (2019). Gene ontology causal activity modeling (GO-CAM) moves beyond GO annotations to structured descriptions of biological functions and systems. *Nat. Genet.* 51, 1429–1433. doi:10.1038/s41588-019-0500-1

Wacker, T., Helmstetter, N., Wilson, D., Fisher, M. C., Studholme, D. J., and Farrer, R. A. (2023). Two-speed genome evolution drives pathogenicity in fungal pathogens of animals. *Proc. Natl. Acad. Sci. U. S. A.* 120, e2212633120. doi:10.1073/pnas.2212633120

You, R., Zhang, Z., Xiong, Y., Sun, F., Mamitsuka, H., and Zhu, S. (2018). GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. *Bioinforma. Oxf. Engl.* 34, 2465–2473. doi:10.1093/bioinformatics/bty130

Zhou, N., Jiang, Y., Bergquist, T. R., Lee, A. J., Kacsoh, B. Z., Crocker, A. W., et al. (2019). The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.* 20, 244. doi:10.1186/s13059-019-1835-8