W330–W336 *Nucleic Acids Research, 2022, Vol. 50, Web Server issue* *Published online 31 May 2022*
https://doi.org/10.1093/nar/gkac441

FUNAGE-Pro: comprehensive web server for gene set enrichment analysis of prokaryotes

Anne de Jong *, **Oscar P. Kuipers and Jan Kok**

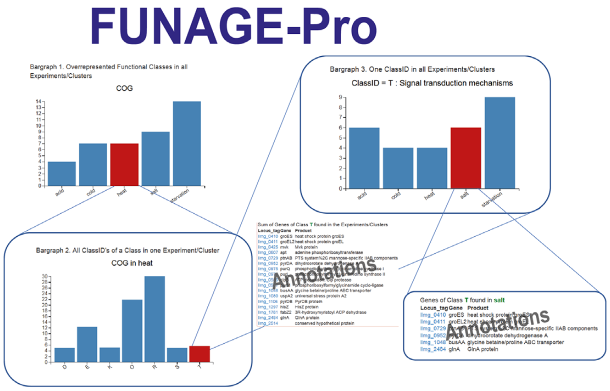
Department of Molecular Genetics, University of Groningen, Groningen Biomolecular Sciences and Biotechnology Institute, the Netherlands

Received February 27, 2022; Revised April 20, 2022; Editorial Decision May 09, 2022; Accepted May 10, 2022

ABSTRACT

Recent advances in the field of high throughput (meta-)transcriptomics and proteomics call for easy and rapid methods enabling to explore not only single genes or proteins but also extended biological systems. Gene set enrichment analysis is commonly used to find relations in a set of genes and helps to uncover the biological meaning in results derived from high-throughput data. The basis for gene set enrichment analysis is a solid functional classification of genes. Here, we describe a comprehensive database containing multiple functional classifications of genes of all (>55 000) publicly available complete bacterial genomes. In addition to the most common functional classes such as COG and GO, also KEGG, InterPro, PFAM, eggnog and operon classes are supported. As classification data for features is often not available, we offer fast annotation and classification of proteins in any newly sequenced bacterial genome. The web server FUNAGE-Pro enables fast functional analysis on single gene sets, multiple experiments, time series data, clusters, and gene network modules for any prokaryote species or strain. FUNAGE-Pro is freely available at http://funagepro.molgenrug.nl.

GRAPHICAL ABSTRACT



INTRODUCTION

High-throughput RNA-seq and proteomics experiments are widely employed and generate an immense quantity of raw data. Several statistical tools have been developed to process the raw data, among which are those used to study gene- or protein expression. Differential expression analysis of genes or proteins in two or more experiments typically results in sets of features that need to be explored. Investigating the function of a set of genes or proteins will give insight in the biological meaning of observed differences between samples. It is not always easy or even possible to find the broader biological effect based on the function of single genes or proteins. Associating genes or proteins through their biological function will reveal changes in functional classes such as metabolic pathways, protein families, operons or orthologous classes (see Table 1).

A common method used to signify the overrepresentation of a functional class in a set of genes is gene set enrichment analysis. This methodology is widely used in diverse studies. Methods or web servers for gene set enrichment analysis have been developed for a limited number of model-organisms (7). For most bacteria the user

*To whom correspondence should be addressed. Tel: +31 503632047; Email: anne.de.jong@rug.nl

Table 1. Overview of the functional classification used for FUNAGE-Pro

Functional class	Abbreviation	Web server	Reference
Gene Ontology	GO	http://geneontology.org	(1)
Metabolic pathways	KEGG	https://www.kegg.jp	(2)
Cluster of Orthologous Groups	COG	http://www.ncbi.nlm.nih.gov/COG	(3)
Cluster of Orthologous Groups	eggnog_COG	http://eggnog5.embl.de	(4)
Protein Families	PFAM	http://pfam.xfam.org	(5)
Protein Families	InterPro	https://www.ebi.ac.uk/interpro	(6)
Operons	Operon	http://genome2d.molgenrug.nl	This work
Keywords	KEYWORDS	http://genome2d.molgenrug.nl	This work

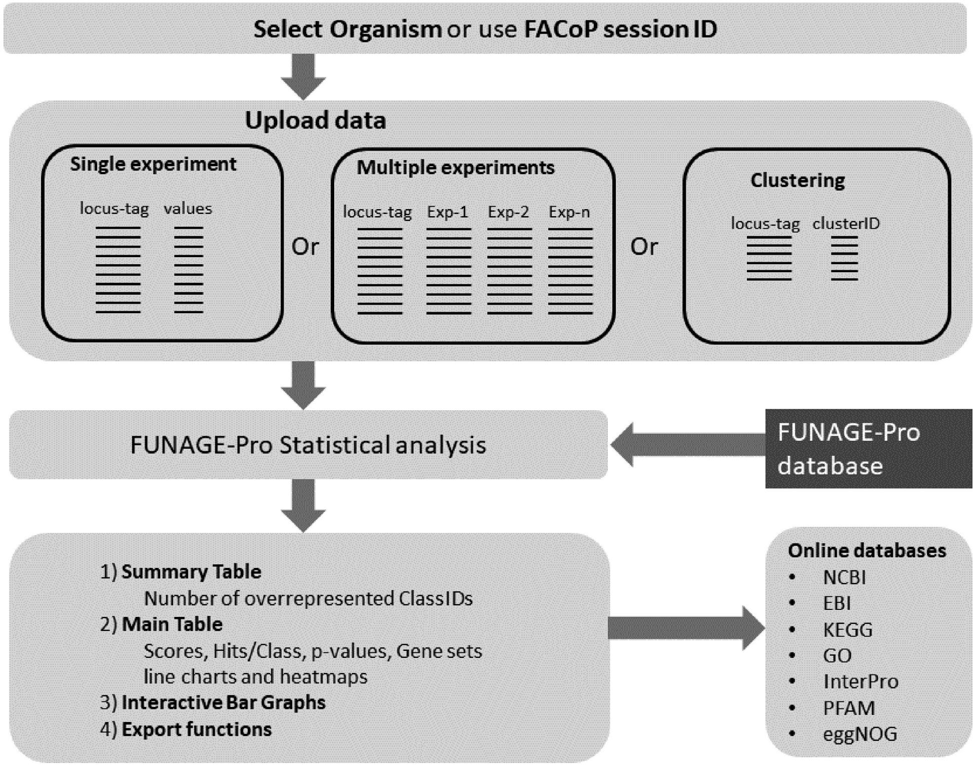


Figure 1. Flow chart of FUNAGE-Pro.

needs to provide 'hard to get' gene classification data to enable gene set enrichment analysis. Statistical packages for gene set enrichment analysis are available in R (8–10) or online web servers such as DAVID (11), MSigDB (12) and Enrichr (13). Unfortunately, gene set enrichment analysis web servers for prokaryotes do not exist or are depreciated, such as JProGO (14). We have built a database based on eight functional classes (see Table 1) to facilitate easy gene set enrichment analysis on any prokaryote with a complete genome available at NCBI (15). We also developed an annotation tool, FACoP, for fast classification of genes or proteins in Whole Genome Sequences (WGS) or novel complete genomes. Gene set enrichment analysis can be done using the FUNAGE-Pro web server (http://funagepro.molgenrug.nl). It can also be performed in conjunction with FACoP (http://facop.molgenrug.nl). For large projects, we offer a stand-alone bioinformatics toolbox to add protein classification to new or draft genomes. It includes an R-script for the statistical analyses.

MATERIALS AND METHODS

Architecture of FUNAGE-Pro

FUNAGE-Pro can handle multiple types of input, varying from a single list of locus-tags to the data of multiple experiments and clustering results. The analysis core is written in R and can be used as stand-alone analysis pipeline. The associated interactive web server enables easy and detailed mining of results. The generated hyperlinks give access to online databases to retrieve background information on the classes listed in the results tables or graphs. Figure 1 shows a global overview of the architecture of FUNAGE-Pro.

Building a classification database

The primary resource for FUNAGE-Pro is a mirror of all complete bacterial genomes in the NCBI RefSeq (15) and Genbank databases (16). All protein sequences specified by >55 000 bacterial genomes were mapped against the reviewed and manually curated (Swiss-Prot) prokaryote database of UniProt (17). This mapping was done based

Table 2. Number of differentially expressed genes (DEGs)[a]

Time point	Upregulated	Downregulated
P0	877	763
P2	591	172
P3	869	591
P4	1054	870
P5	1120	949
P6	1190	1069

[a] Obtained by comparing gene expression in the culture at the indicated time points with gene expression in the culture at time point **P1** (the common reference).

Table 3. Summary table. Number of overrepresented class-IDs per time point

Class-ID\time point	P0	P2	P3	P4	P5	P6
COG	10	13	12	3	10	10
GO	57	67	38	9	26	32
IPR	71	79	70	12	49	33
KEGG	11	11	6	2	3	2
KEYWORDS	13	19	12	7	15	13
Pfam	28	32	32	5	22	16
REGULON	37	36	20	12	16	28
eggNOG_COG	14	19	15	5	14	12
operons	26	42	31	7	34	21

on DIAMOND (18) best hit with an e-value cutoff of 0.01. Using UniProt protein annotation, the following functional classes were assigned to each protein: GO, KEGG, PFAM, InterPro, COG and eggnog_COG (Table 1). Proteins were assigned to COG functional categories by matching the description of the COG functional categories to the description of the InterPro annotation. A special class of KEYWORDS was generated based on the 'KW' field of the UniProt protein database, excluding over- and underrepresented keywords. Operons were built from overlapping gene pairs. A gene pair consists of two genes transcribed from the same strand, with an intergenic spacing smaller than 150 bases and not containing a predicted transcription terminator. The classification data of genomes can be downloaded per genome from the FUNAGE-Pro web server.

Classifying Whole Genome Sequences (WGS) or novel genomes

For protein classification of novel genomes or Whole Genome Sequences (WGS), we developed the fast online tool FACoP: Functional Annotation and Classification of Proteins of Prokaryotes. It is based on the annotation method described above. This web server is freely available at http://facop.molgenrug.nl and accepts up to 10,000 proteins for classification. The results are stored on the server and can be used as input for FUNAGE-Pro via the FACoP session ID.

Input data and automated threshold detection

A reference genome is selected from the FUNAGE-Pro database as a starting point for analysis by FUNAGE-Pro. Alternatively, the session ID from the FACoP web server can be used as input. The second step is to load the data to be analyzed, such as a table with differentially expressed genes (DEGs) with locus-tags as identifiers. Many DEG analysis methods require an arbitrary cutoff for the ratio and/or p-value to identify the genes considered as differentially expressed. FUNAGE-Pro can automatically determine optimal settings by benchmarking cutoff values to find the maximum number of overrepresented GO (Gene Ontology) classes. Optimal settings are those positive and negative cutoff values that give rise to the highest number of overrepresented GO classes.

Input data types

The input for FUNAGE-Pro is a bacterial genome selected from the FUNAGE-Pro database or annotated by FACoP. The gene set data consists of (i) a single list of genes, or (ii) the results of multiple experiments, or (iii) clustering data derived e.g. from k-means clustering or network reconstructions. The NCBI changed the locus-tag names of many genomes to the RefSeq nomenclature, but FUNAGE-Pro also supports Genbank 'old locus-tag' names. Additionally, our Genome2D server (http://genome2d.molgenrug.nl/) (19) enables interconverting the Genbank and RefSeq locus-tags. To analyse proteomics data, the protein names need to be converted to locus-tags using the Uniport ID mapper: https://www.uniprot.org/uploadlists/.

Gene set enrichment statistics used in FUNAGE-Pro

Statistical methods used for gene set enrichment analysis are 'Hypergeometric Testing' (20), 'Z-score and Permutation' (13) or 'Linear Models' (21). These are incorporated in various Bioconductor R packages (22). The upper cumulative hypergeometric probability is used in FUNAGE-Pro to find overrepresented class-IDs. This statistical test uses four values to calculate a P-value of a class-ID; (i) '*population size*'**N** is the total number of annotated genes in the genome, (ii) '*population identified as a success*'**k** is the number of genes with a significant differential expression value, (iii) '*sample size*' **n** is the number genes in a class-ID and (iv) '*sample identified as a success*'**m** is the number of significant values in the class-ID.

$$P = \sum_{x=m}^{n} \frac{\binom{x}{k}\binom{N-x}{n-k}}{\binom{N}{n}}$$

Subsequently, Benjamini–Hochberg multiple testing correction is applied to calculate the final P-value.

$$P_{(i)}^{BH} = \min_{i \le k} \{P_{(k)}.m/k\}$$

In addition to the P-values, a ranking score from 0–9 (bad–good) was added using the formula: $-\log_2(P\text{-value}) \times$ Hits/ClassSize. This ranking score is only used for visualization purposes by converting this value to a blue color gradient. Based on our experience from many analyses,

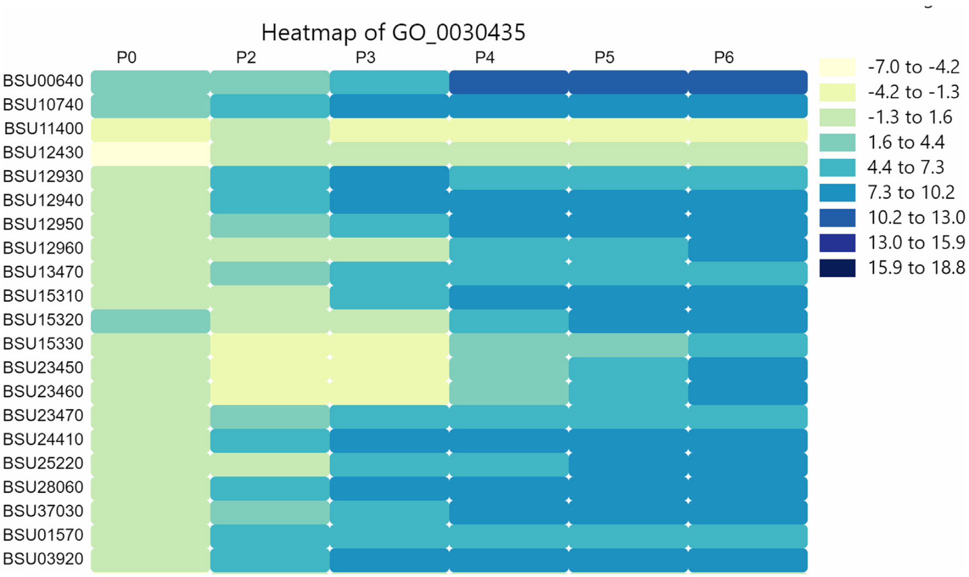


Figure 2. FUNAGE-Pro heatmap of GO:0030435 ('sporulation resulting in formation of a cellular spore'). The gene ontology class GO:0030435 (266 genes) is overrepresented during most time points and expression of most genes in this class increases over time. Negative differential expression of gene values is shown in yellow, positive values are shown in blue.

Table 4. Cluster 14, a cluster with the highest overrepresentation of sporulation-associated Gene Ontology (GO) class-IDs

ClassID	Class description	Locus-tag	gene	Protein description
GO:0003690	double-stranded DNA binding	BSU19950	*sspC*	small acid-soluble spore protein C
		BSU29570	*sspA*	small acid-soluble spore protein A
GO:0042601/GO:0030436	endospore-forming forespore	BSU08550	*sspK*	small acid-soluble spore protein K
		BSU17990	*sspO*	small acid-soluble spore protein O
GO:0009847	spore germination	BSU03700	*gerKA*	spore germination protein KA
		BSU07780	*yfkR*	spore germination protein YfkR
		BSU07790	*yfkQ*	hypothetical protein
		BSU22920	*ypeB*	sporulation protein YpeB
		BSU33050	*gerAA*	spore germination protein A1
		BSU33060	*gerAB*	spore germination protein A2
		BSU33070	*gerAC*	spore germination protein A3
		BSU35800	*gerBA*	spore germination protein B1
GO:0016021	integral component of membrane	BSU10980	*yitG*	MFS transporter
		BSU22040	*ypbQ*	hypothetical protein
		BSU25010	*yqgE*	hypothetical protein
		BSU38390	*ywbA*	permease IIC component YwbA

we observed that classIDs that occur in >50 proteins describe very general functions and are less important for final results. For very small classes with only one or two members, statistics is not reliable. For these two reasons we lowered the scoring value if $n \leq 2$ or $n \geq 50$ with two points.

FUNAGE-Pro web server

The FUNAGE-Pro web server (http://funagepro.molgenrug.nl/) is a user-friendly interface and produces easy interpretable interactive results (Figure 1). The FACoP web server allows fast annotation and classification of genomes and is directly linked to the FUNAGE-Pro web server.

RESULTS

Functional analysis of a sporulating bacterium using FUNAGE-Pro

The FUNAGE-Pro web server offers easy and fast functional analyses of transcriptome or proteome data. To show the power of FUNAGE-Pro, an RNA-seq time series data set (GEO: GSE108659) published by Krawczyk *et al.* (23) was used. This dataset describes the sporulation process of 8 different *Bacillus* strains followed over time. We selected *Bacillus subtilis* 168 and used this part of the data set as a showcase, as that allows checking the outcome of FUNAGE-Pro by existing literature (see supplementary file S1.SporulationData.xlsx). The analysis of the stress response to novel antimicrobial coatings in a clinical methicillin-resistant *Staphylococcus aureus* (MRSA) strain

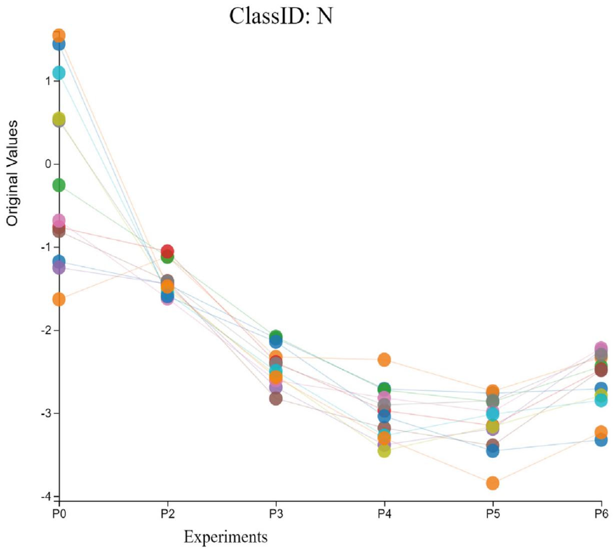


Figure 3. FUNAGE-Pro line chart of the class COG:N ('cellular processes and signalling; cell motility'). Gene expression time series profile of 11 genes of the COG:N class are shown in different colors. The identity of the colored genes can be rapidly obtained by 'mousing over'.

was performed as an example of a less-well studied organism (see supplementary FACoP_example.zip). The transcriptome of *Bacillus subtilis* 168 was studied during seven stages of the morphological phases in sporulation (23): P0, resuspension of cells in sporulation medium; P1, onset of sporulation before asymmetric division; P2, visible asymmetric septum; P3, ongoing engulfment of the forespore compartment by the mother cell compartment of sporulating cells; P4, completed engulfment (engulfed phase-dark forespores are surrounded by the mother-cell cytoplasm); P5, maturation (ongoing dehydration of the forespore core, seen as a transition of phase-dark forespores into phase-bright forespores); and P6, phase-bright forespores (sporulation almost completed, mother-cells contain phase-bright dehydrated forespores). The T-REx web server (24) was employed to perform differential gene expression statistics (input data for T-Rex; SporulationData.xlsx). The number of DEGs was obtained by comparing gene expression in the culture at the various time points with that of the common reference, the culture at time point P1. The results are shown in Table 2. The table of all DEGs observed at all time points and the result of the *k*-means clustering were used as input for the FUNAGE-Pro web server (supplementary files S2.DEG.txt and S3.Clusters.txt, respectively).

FUNAGE-Pro reveals sets of genes involved in sporulation

Next to the table with DEGs (supplementary file S2.DEG.txt), the *B. subtilis* 168 RefSeq entry was taken from the FUNAGE-Pro database. FUNAGE-Pro analysis was started by selecting, respectively, the 'Auto detect cutoff values' and 'Experiments' options. The output of FUNAGE-Pro consists of three parts: (i) a 'Summary Table' showing the number of overrepresented class-IDs per time point (Table 3), (ii) a 'Main Table', further divided into nine tables of class categories, including line charts and heatmaps and (iii) an 'Interactive Bar Graphs', which contains the same information as the 'Main Table' but allows interactive and more visual mining of the results. FUNAGE-Pro revealed that at time point P0, at which cells were resuspended in sporulation medium, core processes changed such as UMP biosynthesis and nitrate metabolism. Specific classification can be visualized using line graphs or heatmaps (see Figure 2).

The COG class analysis revealed a response of cell motility (COG:N) at time point P2: many genes involved in flagella are downregulated, indicating that cells strongly react to the change from rich medium to sporulation medium (Figure 3). At sporulation stage P3, genes involved in sporulation are differentially expressed (*gerAA, gerKA, yraG* and *spoIID*), while the *YheDC* operon encoding endospore coat-associated proteins is activated at P4. Interestingly, three genes co-located on the genome, *cotP* (specifying spore coat protein P) and two uncharacterized genes *ydgA* and *ydgB* show highly correlated expression profiles. This result suggests that *ydgA* and *ydgB* may also play a role in sporulation in *B. subtilis*.

Analysis of the InterPro class revealed (Figure 4) that genes encoding spore-coat (IPR019593) and endospore-associated (IPR026838) proteins and sporulation lipoproteins (IPR019076) are overrepresented at the later time points P5 and P6.

In addition to a comprehensive visualization of results in bar charts, FUNAGE-Pro allows zooming into different time points or categories. This option allows getting a quick overview of the behavior of genes in different time points (or experiments).

Cluster analysis reveals sporulation-associated clusters

FUNAGE-Pro was used to identify biological function(s) that are overrepresented in the gene sets derived from the *k*-means clustering by T-Rex. The results showed that 'Cluster 1' contains genes associated with the flagellum; their expression decreased over time. This association is based on the overrepresentation of six GO-IDs associated to flagellum; GO:0071973, bacterial-type flagellum-dependent cell motility; GO:0003774, motor activity; GO:0009425, bacterial-type flagellum basal body; GO:0030694, bacterial-type flagellum basal body rod; GO:0044780, bacterial-type flagellum assembly; GO:0009288, bacterial-type flagellum (Figure S1). The most interesting clusters with respect to sporulation are Clusters 2, 11 and 14. The latter contained the most overrepresented GO classes associated with the sporulation process (Table 4). Also, other classifications such as COG, operon, PFAM, InterPro and keywords showed that Cluster 14 contains the highest overrepresentation of sporulation-associated class-IDs. Genes involved in early-stage sporulation and endospore formation are overrepresented in Cluster 9 (data not shown).

DISCUSSION

The results presented here show that FUNAGE-Pro can quickly and fully automatically perform gene set enrichment analyses on data derived from RNA-seq analysis

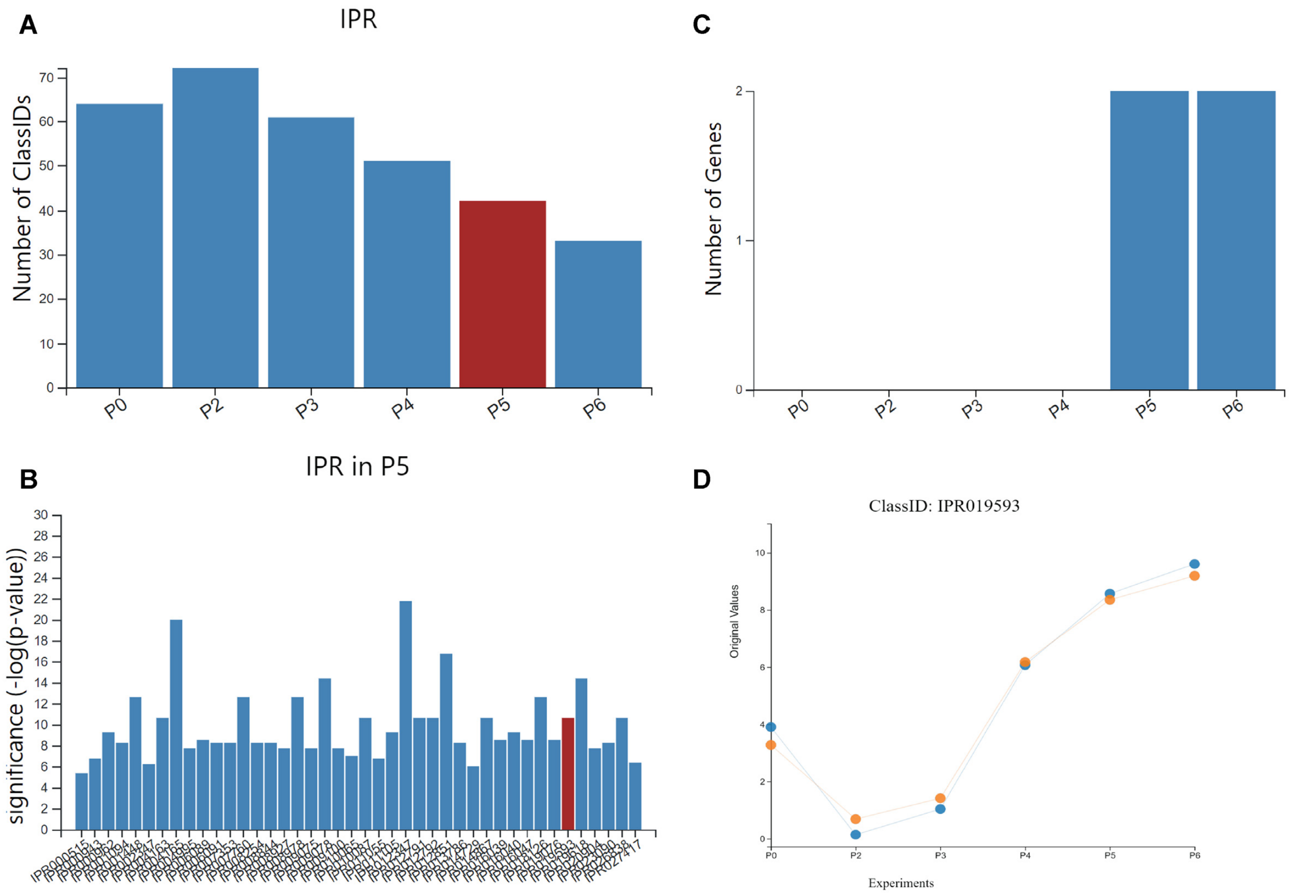


Figure 4. FUNAGE-Pro bar graph analysis of the InterPro (IPR) class. (**A**) Bar size represents the number of over-represented InterPro-IDs at each time point (P0 – P6). Details of the selected time point P5 (red) are shown in bar graph B. (**B**) All over-represented InterPro-IDs at time point P5 (IPR019593 (red)) were selected for further analysis, shown in (C). (**C**) The number of DEGs containing the IPR019593 protein domain is shown for all time points. (**D**) Line chart of the differential expression values at each time point of the two genes constituting IPR019593, *cotZ* (BSU11740, spore coat protein Z) and *cotY* (BSU11750, spore coat protein Y).

pipelines. In general, gene set enrichment analysis is limited to GO classification. Using more types of gene classifications (COG, KEGG, InterPro), such as provided in FUNAGE-Pro, will give more in-depth information on the biological function of gene sets. We provide a basic operon prediction tool to build operon classes for complete genomes, improving the power of FUNAGE-Pro. Functional analysis of clusters derived from e.g. *k*-means clustering is also supported. FUNAGE-Pro produces a lot of data, but the results can be easily grasped by using the interactive tables and graphs. FUNAGE-Pro is uniquely suited for the analysis of differential gene expression data due to availability of classification data of all fully sequenced bacterial genomes. The possibility to quickly annotate and classify proteins specified by novel genomes using FACoP further amplifies the utility of FUNAGE-Pro. Analysis of the data from a *B. subtilis* sporulation RNA-seq chronotranscriptomics study revealed that biological processes can be efficiently examined using the user-friendly FUNAGE-Pro web server. Although FUNAGE-Pro was developed for prokaryotes, it has also recently been successfully used for the analysis of yeast RNA-seq data using the stand-alone version. This option has not been implemented in the web server. FUNAGE-Pro will probably perform well for a wider taxonomic scope. The FUNAGE-Pro web server is continuously updated and expanded using the feedback of users located all over the world.

DATA AVAILABILITY

R, Perl and Python scripts of FUNAGE-Pro and FACoP can be downloaded from github: https://github.com/annejong/FUNAGEPro.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Antonina Krawczyk for making available the raw and processed transcriptome data from her study on *B. subtilis* sporulation. We also wish to thank Dr Auke van Heel and Dr Sjoerd van de Meulen for their continuous support.

FUNDING

Department of Molecular Genetics, Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, the Netherlands.

Conflict of interest statement. None declared.

REFERENCES

1. The Gene Ontology Consortium (2021) The gene ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
2. Kanehisa,M., Furumichi,M., Sato,Y., Ishiguro-Watanabe,M. and Tanabe,M. (2021) KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.*, **49**, D545–D551.
3. Galperin,M.Y., Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, **43**, D261–D269.
4. Huerta-Cepas,J., Szklarczyk,D., Heller,D., Hernández-Plaza,A., Forslund,S.K., Cook,H., Mende,D.R., Letunic,I., Rattei,T., Jensen,L.J. *et al.* (2019) eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.*, **47**, D309–D314.
5. Mistry,J., Chuguransky,S., Williams,L., Qureshi,M., Salazar,G.A., Sonnhammer,E.L.L., Tosatto,S.C.E., Paladin,L., Raj,S., Richardson,L.J. *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
6. Blum,M., Chang,H.-Y., Chuguransky,S., Grego,T., Kandasaamy,S., Mitchell,A., Nuka,G., Paysan-Lafosse,T., Qureshi,M., Raj,S. *et al.* (2021) The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*, **49**, D344–D354.
7. Kuleshov,M.V., Diaz,J.E.L., Flamholz,Z.N., Keenan,A.B., Lachmann,A., Wojciechowicz,M.L., Cagan,R.L. and Ma'ayan,A. (2019) modEnrichr: a suite of gene set enrichment analysis tools for model organisms. *Nucleic Acids Res.*, **47**, W183–W190.
8. Zhang,Y., Topham,D.J., Thakar,J. and Qiu,X. (2017) FUNNEL-GSEA: FUNctioNal ELastic-net regression in time-course gene set enrichment analysis. *Bioinformatics*, **33**, 1944–1952.
9. Napolitano,F., Carrella,D., Gao,X. and di Bernardo,D. (2020) gep2pep: a bioconductor package for the creation and analysis of pathway-based expression profiles. *Bioinformatics*, **36**, 1944–1945.
10. Rahmatallah,Y., Zybailov,B., Emmert-Streib,F. and Glazko,G. (2017) GSAR: bioconductor package for gene set analysis in R. *BMC Bioinf.*, **18**, 61.
11. Jiao,X., Sherman,B.T., Huang,D.W., Stephens,R., Baseler,M.W., Lane,H.C. and Lempicki,R.A. (2012) DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, **28**, 1805–1806.
12. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. U.S.A.*, **102**, 15545–15550.
13. Chen,E.Y., Tan,C.M., Kou,Y., Duan,Q., Wang,Z., Meirelles,G.V., Clark,N.R. and Ma'ayan,A. (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinf.*, **14**, 128.
14. Scheer,M., Klawonn,F., Münch,R., Grote,A., Hiller,K., Choi,C., Koch,I., Schobert,M., Härtig,E., Klages,U. *et al.* (2006) JProGO: a novel tool for the functional interpretation of prokaryotic microarray data using gene ontology information. *Nucleic Acids Res.*, **34**, W510–W515.
15. Li,W., O'Neill,K.R., Haft,D.H., DiCuccio,M., Chetvernin,V., Badretdin,A., Coulouris,G., Chitsaz,F., Derbyshire,M.K., Durkin,A.S. *et al.* (2021) RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic Acids Res.*, **49**, D1020–D1028.
16. Sayers,E.W., Cavanaugh,M., Clark,K., Pruitt,K.D., Schoch,C.L., Sherry,S.T. and Karsch-Mizrachi,I. (2021) GenBank. *Nucleic Acids Res.*, **49**, D92–D96.
17. The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
18. Buchfink,B., Xie,C. and Huson,D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*, **12**, 59–60.
19. Baerends,R.J.S., Smits,W.K., de Jong,A., Hamoen,L.W., Kok,J. and Kuipers,O.P. (2004) Genome2D: a visualization tool for the rapid analysis of bacterial transcriptome data. *Genome Biol.*, **5**, R37.
20. Backes,C., Keller,A., Kuentzer,J., Kneissl,B., Comtesse,N., Elnakady,Y.A., Müller,R., Meese,E. and Lenhof,H.-P. (2007) GeneTrail—advanced gene set enrichment analysis. *Nucleic Acids Res.*, **35**, W186–W192.
21. Oron,A.P., Jiang,Z. and Gentleman,R. (2008) Gene set enrichment analysis using linear models and diagnostics. *Bioinformatics*, **24**, 2586–2591.
22. Huber,W., Carey,V.J., Gentleman,R., Anders,S., Carlson,M., Carvalho,B.S., Bravo,H.C., Davis,S., Gatto,L., Girke,T. *et al.* (2015) Orchestrating high-throughput genomic analysis with bioconductor. *Nat. Methods*, **12**, 115–121.
23. Krawczyk,A.O., de Jong,A., Eijlander,R.T., Berendsen,E.M., Holsappel,S., Wells-Bennik,M.H.J. and Kuipers,O.P. (2015) Next-Generation whole-genome sequencing of eight strains of Bacillus cereus, isolated from food. *Genome Announc.*, **3**, e01480-15.
24. de Jong,A., van der Meulen,S., Kuipers,O.P. and Kok,J. (2015) T-REx: transcriptome analysis webserver for RNA-seq expression data. *BMC Genomics*, **16**, 663.