Bacterial Feature Finder (BaFF) – A system for extracting features overrepresented in sets of prokaryotic organisms

Supplementary Material

Javier López-Ibáñez, Laura T. Martín, Mónica Chagoyen & Florencio Pazos*

Computational Systems Biology Group. Systems Biology Program. Spanish National Centre for Biotechnology (CNB-CSIC). c/ Darwin, 3. 28049 Madrid. Spain. Phone: +34-915854669. Fax: +34-915854506.

pazos@cnb.csic.es

1.- Statistical Tests

Enrichment analysis statistics were computed using functions from the R *stats* package [1].

For quantitative features (continuous values), a two Sample, two-sided Kolmogorov-Smirnov test (R function: **`ks.test()`**) was used for checking if both sets of values (input set and background) have the same distribution. We report the p-value of rejecting the hypothesis that both distributions are different.

For qualitative features (discrete values), we compute the probability of obtaining a given annotation randomly from the background set using the cumulative hypergeometric density function (R function: **`phyper()`**). In these cases a p-value correction for multiple testing is also performed, with Benjamini's False Discovery Rate (FDR) correction [2] (R function: **`p.adjust()`**).

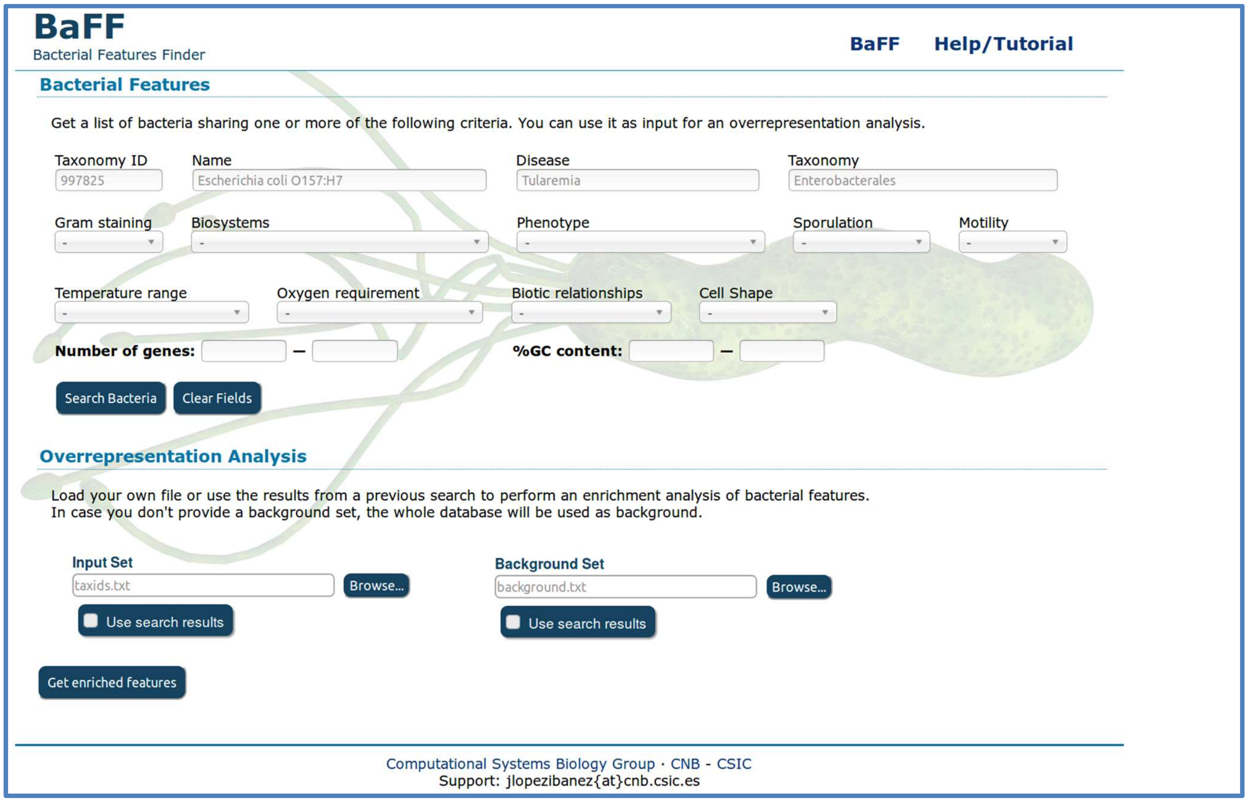
[1] R Core Team (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[2] Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society, Series B. 57: 289-300.

2.- Features and sources

Feature	Source Database	URL
- Number of genes	EnsemblGenomes	https://img.jgi.doe.gov/cgi-bin/m/main.cgi http://ensemblgenomes.org/info/genomes
- GC content	NCBI Genome	https://www.ncbi.nlm.nih.gov/genome/
- Biosystems	NCBI biosystems	https://www.ncbi.nlm.nih.gov/biosystems
- COG Functional Classes (25)	Microbesonline	http://www.microbesonline.org/cgi-bin/genomeInfo.cgi
- Gram staining - Sporulation - Motility - Shape - Biotic Relationships - Phenotype - Oxigen requirement - Temperature range - Disease - Host	Integrated Microbial genomes (IMG)	https://img.jgi.doe.gov/
- Taxonomy	NCBI Taxonomy	https://www.ncbi.nlm.nih.gov/Taxonomy/

3.- Screenshots of the interface



Bacterial Features

Get a list of bacteria sharing one or more of the following criteria. You can use it as input for an overrepresentation analysis.

Taxonomy ID: 997825
Name: Escherichia coli O157:H7
Disease: Tularemia
Taxonomy: Enterobacterales

Gram staining: -
Biosystems: -
Phenotype: -
Sporulation: -
Motility: -

Temperature range: -
Oxygen requirement: -
Biotic relationships: -
Cell Shape: -

Number of genes: —
%GC content: —

Overrepresentation Analysis

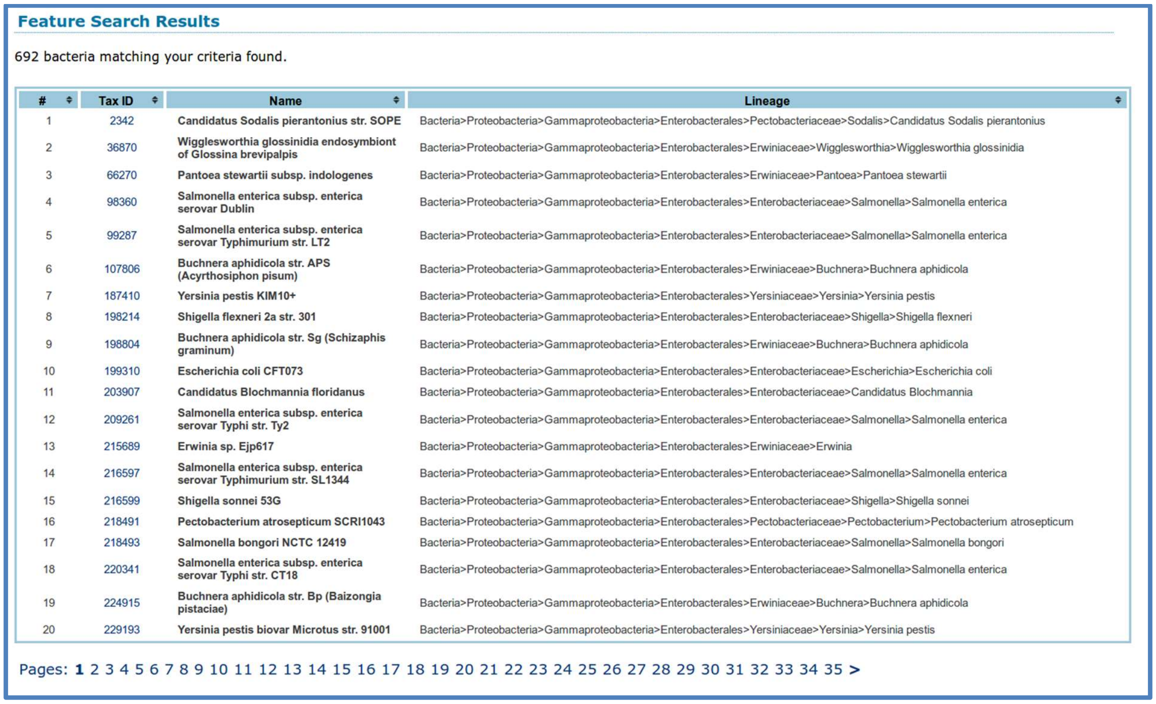
Load your own file or use the results from a previous search to perform an enrichment analysis of bacterial features.
In case you don't provide a background set, the whole database will be used as background.

Input Set: taxids.txt
□ Use search results

Background Set: background.txt
□ Use search results

Computational Systems Biology Group · CNB - CSIC
Support: jlopezibanez{at}cnb.csic.es

Main interface



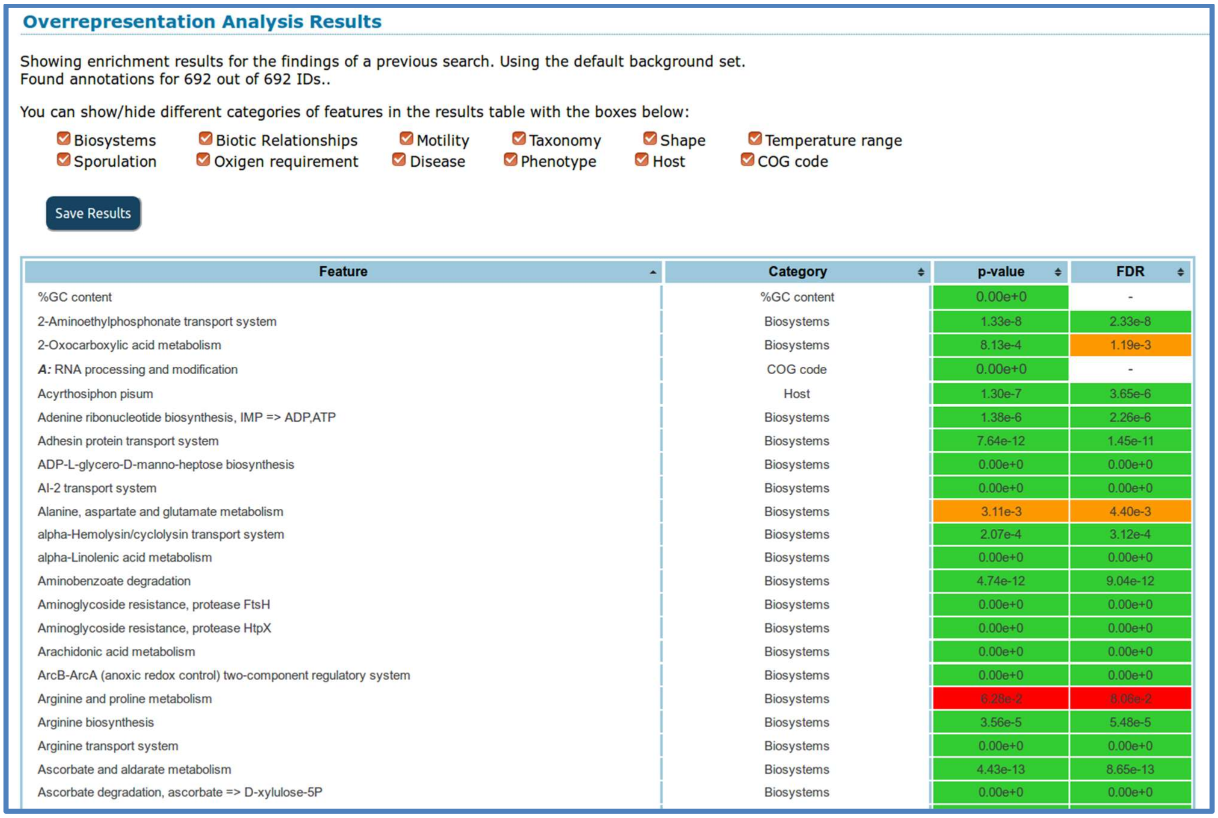
Feature Search Results

692 bacteria matching your criteria found.

#	Tax ID	Name	Lineage
1	2342	Candidatus Sodalis pierantonius str. SOPE	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Pectobacteriaceae>Sodalis>Candidatus Sodalis pierantonius
2	36870	Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Erwiniaceae>Wigglesworthia>Wigglesworthia glossinidia
3	66270	Pantoea stewartii subsp. indologenes	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Erwiniaceae>Pantoea>Pantoea stewartii
4	98360	Salmonella enterica subsp. enterica serovar Dublin	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Enterobacteriaceae>Salmonella>Salmonella enterica
5	99287	Salmonella enterica subsp. enterica serovar Typhimurium str. LT2	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Enterobacteriaceae>Salmonella>Salmonella enterica
6	107806	Buchnera aphidicola str. APS (Acyrthosiphon pisum)	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Erwiniaceae>Buchnera>Buchnera aphidicola
7	187410	Yersinia pestis KIM10+	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Yersiniaceae>Yersinia>Yersinia pestis
8	198214	Shigella flexneri 2a str. 301	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Enterobacteriaceae>Shigella>Shigella flexneri
9	198804	Buchnera aphidicola str. Sg (Schizaphis graminum)	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Erwiniaceae>Buchnera>Buchnera aphidicola
10	199310	Escherichia coli CFT073	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Enterobacteriaceae>Escherichia>Escherichia coli
11	203907	Candidatus Blochmannia floridanus	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Enterobacteriaceae>Candidatus Blochmannia
12	209261	Salmonella enterica subsp. enterica serovar Typhi str. Ty2	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Enterobacteriaceae>Salmonella>Salmonella enterica
13	215689	Erwinia sp. Ejp617	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Erwiniaceae>Erwinia
14	216597	Salmonella enterica subsp. enterica serovar Typhimurium str. SL1344	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Enterobacteriaceae>Salmonella>Salmonella enterica
15	216599	Shigella sonnei 53G	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Enterobacteriaceae>Shigella>Shigella sonnei
16	218491	Pectobacterium atrosepticum SCRI1043	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Pectobacteriaceae>Pectobacterium>Pectobacterium atrosepticum
17	218493	Salmonella bongori NCTC 12419	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Enterobacteriaceae>Salmonella>Salmonella bongori
18	220341	Salmonella enterica subsp. enterica serovar Typhi str. CT18	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Enterobacteriaceae>Salmonella>Salmonella enterica
19	224915	Buchnera aphidicola str. Bp (Baizongia pistaciae)	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Erwiniaceae>Buchnera>Buchnera aphidicola
20	229193	Yersinia pestis biovar Microtus str. 91001	Bacteria>Proteobacteria>Gammaproteobacteria>Enterobacterales>Yersiniaceae>Yersinia>Yersinia pestis

Pages: **1** 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 **>**

Paginated results of a database search



Overrepresentation Analysis Results

Showing enrichment results for the findings of a previous search. Using the default background set.
Found annotations for 692 out of 692 IDs..

You can show/hide different categories of features in the results table with the boxes below:

☑ Biosystems ☑ Biotic Relationships ☑ Motility ☑ Taxonomy ☑ Shape ☑ Temperature range
☑ Sporulation ☑ Oxigen requirement ☑ Disease ☑ Phenotype ☑ Host ☑ COG code

Feature	Category	p-value	FDR
%GC content	%GC content	0.00e+0	-
2-Aminoethylphosphonate transport system	Biosystems	1.33e-8	2.33e-8
2-Oxocarboxylic acid metabolism	Biosystems	8.13e-4	1.19e-3
A: RNA processing and modification	COG code	0.00e+0	-
Acyrthosiphon pisum	Host	1.30e-7	3.65e-6
Adenine ribonucleotide biosynthesis, IMP => ADP,ATP	Biosystems	1.38e-6	2.26e-6
Adhesin protein transport system	Biosystems	7.64e-12	1.45e-11
ADP-L-glycero-D-manno-heptose biosynthesis	Biosystems	0.00e+0	0.00e+0
AI-2 transport system	Biosystems	0.00e+0	0.00e+0
Alanine, aspartate and glutamate metabolism	Biosystems	3.11e-3	4.40e-3
alpha-Hemolysin/cyclolysin transport system	Biosystems	2.07e-4	3.12e-4
alpha-Linolenic acid metabolism	Biosystems	0.00e+0	0.00e+0
Aminobenzoate degradation	Biosystems	4.74e-12	9.04e-12
Aminoglycoside resistance, protease FtsH	Biosystems	0.00e+0	0.00e+0
Aminoglycoside resistance, protease HtpX	Biosystems	0.00e+0	0.00e+0
Arachidonic acid metabolism	Biosystems	0.00e+0	0.00e+0
ArcB-ArcA (anoxic redox control) two-component regulatory system	Biosystems	0.00e+0	0.00e+0
Arginine and proline metabolism	Biosystems	6.28e-2	8.06e-2
Arginine biosynthesis	Biosystems	3.56e-5	5.48e-5
Arginine transport system	Biosystems	0.00e+0	0.00e+0
Ascorbate and aldarate metabolism	Biosystems	4.43e-13	8.65e-13
Ascorbate degradation, ascorbate => D-xylulose-5P	Biosystems	0.00e+0	0.00e+0

Results of the analysis of enriched features.