Supplemental Materials

Index

*Headings for each section of supplementary content are hyperlinked to facilitate document navigation.

Supplementary Data

S1. All keywords accumulated for DUF34 protein family during the comprehensive published data capture process.

S2. Curated sets of tools organized by: (a) orthology-/homology-limited and phyletic patterning tools; (b) gene neighborhood; and (c) synteny tools.

S3. All curated yields of specialized corpus search tools: PubMed, EuropePMC, PubTator, Scinapse.

S4. Simple Comparison of Text-based vs. Sequence-based Publication Retrieval.

S5. Publication Retrieval Method Comparison — Single-sequence vs. HMM-based PaperBLAST Methods.

S6. Venn Diagram Data for the Comparison of the Two Major PaperBLAST Retrieval Methods.

S7. Publication Retrieval Method Comparison — Single-sequence vs. HMM-based vs. Idealized "QCC" Cycle Method.

S8. Venn Diagram Data Comparing PaperBLAST Retrieval Methods to "QCC Cycle" Approach.

S9. Comparison of the PaperBLAST Results of Several, Disparately Related DUF34 Homolog Sequences. **S10.** Final Summary Venn Diagram of all Methods' Results Relevant to DUF34.

Supplementary Tables

S1. All 65 gene families characterized by the Laboratory of Valérie de Crécy-Lagard, PhD.

S2. Results of investigation comparing sequence-based search tool PaperBLAST (two methods) to those of the idealized "QCC" method.

Supplementary Elaborations

S1. Explanation of XynX "productive ignorance" example.

Supplementary Figures

S1. Seed Information Assessment: establish understanding of starting information in preparation for the capture of published data.

S2. Retrieve sequences using known names/aliases.

S3. Family-level analyses.

S4. Example of over-propagation of root family attribution in EggNOG Database as a result of fusions.

S5. WorldWideScience.org-, ScienceResearch.com-generated diagrams of keywords.

S6. Coincidental homonyms are challenges for text-based search engines, even those as specialized as PubTator.

S7. FlaGs example output.

S8. EFI-GNT example output.

S9. Annotree example: single family, PF02591, in bacteria.

S10. COGNAT example output.

S11. SubtiWiki (CoreWiki) beta feature of the database's Genomic Neighborhood Comparison viewer of the DUF34 homolog of B. subtilis, YqfO (BSU_25170).

S12. Example tree output of MicrobesOnline.

S13. Screengrab demonstrating different "filters" for target family recognition, particularly used differentially with multiple targets (MicrobesOnline).

S14. Screengrab demonstrating navigation to family-specific trees in MicrobesOnline via the "Gene Info" tab of each entry.

S15. "Families" (COGs) in STRING search tool options.

S16. KO identifiers being supplied to KEGG orthology tool (multiple families, limited genome customization).

S17. Example CAGECAT output using DUF34 and COG1579 homolog sequences: MSMEG_4307 (NCBI-ProteinID: ABK73705); MSMEG_4306 (NCBI-ProteinID: ABK70599).

S18. Example output for FunCoup using "YbgI" of E. coli.

S19. BV-BRC protein family sorter, general output example (a) with additional viewer examples.

- S20. BV-BRC protein family sorter output, Families.
- **S21.** EggNOG Phylogenetic Profile tool example output using COG0327, COG3323, and COG1579.
- S22. KBase workflow for generating a "gene tree" (as regarded by KBase) from several select genomes.
- S23. Genomic Context Visualizer example output.
- **S24.** GizmoGene example output.

Supplementary References

Supplementary Data

*Supplementary data files in submission packet are also accessible via the accompanying FigShare.

Data S1. All keywords accumulated for DUF34 protein family during the comprehensive published data capture process.

Data S2. Curated sets of tools organized by: (a) orthology-/homology-limited and phyletic patterning tools; (b) gene neighborhood; and (c) syntemy tools.

Data S3. All curated yields of specialized corpus search tools: PubMed, EuropePMC, PubTator, Scinapse. Each search engine was queried using the 10 selected keywords thought to represent the most commonly associated names/aliases associated with the target protein family (i.e., "NGG1 interacting factor 3", "NIF3", "NIF3L1", "GTP cyclohydrolase 1 type 2", "DUF34", "YbgI", "PF01784", "COG0327", "YqfO", and "COG3323"). Respective yields were exported from each engine's results and, subsequently, manually reviewed for relevance rated at three levels: 1) "Focal", a true positive hit in which the target homolog was mentioned in either the title and/or the abstract; 2) "Non-focal", a true positive hit in which the target homolog, while not mentioned in the title or abstract, was mentioned in another subsection or supplemental data of the publication; and 3) "False Positive", a curated and confirmed false positive hit in which the target homolog was falsely identified among any of the yielded publication's text. These three ratings constitute the three left-most columns of each subset of exported hits per search engine (respective publication's row assigned a "yes" in the column corresponding to one of the three rating categories for which it was found affirmative, reflecting the curation results or relevance status). Tables reflecting the curation summaries of each raw export (i.e., each keyword/query per search engine) were generated to document curation process (see "Export Subsets, Curation"). The total number of raw yields varied per search engine due to the variable productivity of individual engines and select keywords within those search engines. Raw yields (i.e., returned publications) for each keyword per search engine were also included in this supplementary dataset, accompanying the export summaries; queries returning no hits were unable to be exported and, therefore, are not represented within a raw file (see export summaries for which queries resulted in no hits, and, therein, which query yields were left out of the raw data files; Data S3 table includes a key for all raw export files). The results of this survey are illustrated in Figure 4 of the main text.

Supplementary Tables

Table S1. All 65 gene families characterized by the Laboratory of Valérie de Crécy-Lagard, PhD. (Asterisks indicate homologs in human)

Pathway	Name	UniProt Representative	COG/DUF/PFAM	PMID
RNA Modification				
Agmitidine Synthase	TiaS	O59476	COG1571	18844986;
				20139989
Archaeosine Synthesis	QueF-like	A3MSP1	COG0780 QueF-paralog	22032275;
-				28383498
Archaeosine Synthesis	Gat/QueC fusion	Q981C9	COG0449/COG0603	22032275
Archaeosine Synthase	ArcS	Q58428	COG1549, DUF5591, TGT	20129918
-			paralog	
5-methylaminomethyl-2- thiouridine synthase	MnmC1/C2 fusion	P77182	COG4121/COG0579	17673083
Pre-Q ₀ Reductase	QueF	Q46920	COG2904/0780	14660578;
(Queuosine Biosynthesis	Quei	Q10720	00022004/0700	15767583
Queuosine/Archaeosine	QueC	O31675	COG0603	14660578
Biosynthesis	Quee	031075	000003	14000370
Queuosine/Archaeosine	QueD	O31676	COG0720	14660578
Biosynthesis	QueD	031070	000720	14000378
Queuosine/Archaeosine	QueE	O31677	COG0602	14660578
Biosynthesis	Quer	0310//		14000378
	Quall	00W710	DUE208	20120540
Queuosine Biosynthesis	QueH	Q9WZJ0	DUF208	28128549
PreQ0/preQ1 transport	YhhQ	P37619	COG1738	28208705
Queuosine (Q) hydrolase	CD1682/QueK	Q186N9	COG1957 paralog	31481610
Queuine lyase	CD1684/QueL	Q186P0	COG1244 paralog	31481610
Q transport	CT140	O84142	COG1738 paralog	31481610
Q Synthesis	CT193	O84196	COG0343 paralog	31481610
Queuosine Salvage	Qng1*	Q9HDZ9	DUF2419	24911101;
				36610787
tRNA-ac4C Biosynthesis	TmcA	D4GW73	COG1444	19478918
tRNA-m1A22	TrmK	P54471	COG2384	18420655;
Biosynthesis				17852564
tRNA-m1Psi54 synthesis	PusY	D4GTL8	COG1901/DUF358	22274953;
tRNA-t6A37 Synthesis (Universal)	TsaC *	P45748	COG0009	19287007
tRNA-t6A37 Synthesis	TC2 S5	D22570	DE02491	10297007
	TsaC2, Sua5	P32579	PF03481	19287007
(Universal)		005516	0001014	21205040
tRNA-t6A37 Synthesis	TsaB	O05516	COG1214	21285948
(Bacteria)				
tRNA-t6A37 Synthesis	TsaD	P05852	COG0533	21285948
(Bacteria)				
tRNA-t6A37 Synthesis	TsaE	P0AF67	COG0802	22378793
(Bacteria)				
tRNA-t6A37 Synthesis	Kae1*	P36132	PF00814	21285948
(Eukarya/Archaea Cyto)				
tRNA Dihydrouridine	DusABC*	P32695	COG0042	11983710
Synthase (Bacteria)				
Wybutosine Biosynthesis	TYW1, WyeA*	Q08960	PF00258/PF04055/PF0860 8	16162496
Wyosine Derivative	Taw22	Q9V2G1	COG2520	20382657
Biosynthesis (Archaea)				
tRNA Psi 32/32 Synthase	YjbO	O31613	COG0564 paralog	32629984
(B. subtilis)	J		Finning	
rRNA Psi Sythase (B.	YhcT	P54604	COG0564 paralog	32629984
subtilis)		10.001	20000 Palmog	2202000
tRNA m6t6A	TrmO*	P28634	COG1720	25063302
DNA Modification	1	1 2000 1	2001120	20000002
dADG modification	DpdA	A0A3A3JBR7	COG0343 TGT paralog	30159947
dADG modification	DpdB	A0A3A3NHD9	pfam14072 DndB paralog	30139947 30159947
dADG modification	DpdC	A0A3A3IHU3	DUF308	30139947 30159947
		AUASASIHUS	DUL200	3013994/

Folate Synthesis (Bacteria,	PTPS-III	A0A098MZ39	COG0720 paralog	19395485
FolB Shunt)				
Folate FolC-like (Chlamydiae)	CT611	O84617	COG1478	17645794
GTP Cyclohydrolase I type 2	FolE2	P94398	COG1469	17032654
Folate, New PABA	CT610	O84616	COG5424	25006229
Synthesis Iron-sulfur Cluster Repair	V	DOADEQ	COC0254 Coort revela	20490192
Pterin Biosynthesis	YgfZ* FolB2/MptD	P0ADE8	COG0354, GcvT paralog	20489182
(Archaea)	FoIB2/MptD	Q57851	COG2098/DUF372/381	22931285
Pterin Biosynthesis (Archaea)	FolK2/MptE	Q59028	COG1634/DUF115	22931285
Tetrahydromonapterin Synthesis	FolX	P0AC19	COG1539	19897652
Tetrahydromonapterin	FolM	P0AFS3	COG1028 paralog	19897652
Synthesis Thiamine Metabolism	T AE	0046320	COC0912	25014715
	TenAE YeiR*	Q9ASY9	COG0812 COG0523 paralog	25014715
Zinc Homeostasis		P33030 A8M783	DUF62	15690043 32776704
R,S SAM hydrolase Oxoproline Metabolism	Sare_1364 YbgJ	A8M/83 P0AAV4	COG20149	28830929
Oxoproline Metabolism Oxoproline Metabolism	YbjK	P0AAV4 P75745		28830929
		P75745 P75746	COG1984	
Oxoproline Metabolism	YbgL		COG1540 paralog	28830929
Riboflavin	YbiA/RibX	P30176	COG3236	25431972
Thiamine Metabolism	At4g29530	Q9SU92	PF06888 paralog	26537753
Thiamine Metabolism (Plants)	TenAC	F4KFT7	IPR036412/IPR016084 /IP R004305	27677881
Phosphopanteine Hydrolase	At4g32180*	Q8L5Y9	DUF89_II (PanK fusion)	27322068
B12 Metabolism	PF0295	Q8U404	COG2103/DUF71 paralog	23013770
PLP Homeostasis	YggS/PLPHP/PLPBP*	P67080	COG0325	26872910
Protein Modification			1	
Beta-Lysylation of Elongation Factor P	YjeA	P0A8N7	COG226 paralogs	20070887; 21841797
Beta-Lysylation of Elongation Factor P	YjeK	P39280	COG0535 paralogs	20070887; 21841797
Diphthamide Biosynthesis	Dph6*	Q12429	COG2103/DUF71 paralog	23013770
Deoxyhypusine	HVO 2299	D4GWG4	COG0010 paralogs	28053595
Biosynthesis	HVO_2299	D40w04	COG0010 paralogs	28033393
Carbon Source Catabolism				
Mannitol-phosphate Dehydrogenase/Phosphata se	MtlD	Q6FBP5	PF13419/PF08125	24800891
Sugar-phosphate and Nucleotide Hydrolase	PH1575	059272	DUF89_I	27322068
Sugar-phosphate Hydrolase	YMR027W	Q04371	DUF89_III	27322068
3-Oxo-tertronate Kinase	YgbK	Q46889	DUF1537 paralog	27402745; 27294475
D-Threonate Kinase	DtnK	Q8ZRS5	DUF1537 paralog	27294473 27402745; 27294475
D-Threonate 4-phosphate	PdxA2	P58718	COG1995 paralog	27402745;
Dehydrogenase	()			27294475
Unpublished (papers in-pr		025000	COC1242	
nm5U34 Synthase	YtqA VtaP	O35008	COG1242 COG2519	
nm5U34 Methylase	YtqB DrdM	O34614		latter - : //la *
dPreQ1 Methylase	DpdM	M4PNV1	PF03692 (paralog?)	https://biorxi v.org/cgi/con
				tent/short/20

				<u>23.04.13.536</u> <u>721v1</u>
dPreQ1 Formyltransferase	DpdN	NA, ncbi: YP_010114479.1	PF00551 paralog	https://biorxi v.org/cgi/con tent/short/20 23.04.13.536 721v1
CDG Decarboxylase	DpdL	M4SNA7	COG0720 paralog	https://biorxi v.org/cgi/con tent/short/20 23.04.13.536 721v1
preQ ₀ /preQ ₁ Transporter	Bifidobacterium breve CREST homolog	A0A0M3T8W5	PF03006/COG1272 paralog	
preQ ₀ /preQ ₁ Transporter	Bartonella henselae MFS homolog	WP_011180872.1	PF07690/COG2814 paralog	
preQ ₀ /preQ ₁ /q	Acidobacteria bacterium DMT homolog	A0A2V9U0M9	PF07168 paralog	

Table S2. Results of investigation comparing sequence-based search tool PaperBLAST (two methods) to those of the idealized "QCC" method. "Focal" and "False Positive" publication labels have been bolded in the table.

	Method			
	PaperBLAST			QCC
	HMM	Single Sequence	e/UniProt ID	
Relevance	PF01784	H. sapiens (Q9GZT8)	E. coli (P0AFP6)	
Non-Focal				Ahmed 2011
Focal	Akiyama 2003	Akiyama 2003		Akiyama 2003
Non-Focal				Alalouf 2011
Non-Focal	Alam 2011	Alam 2011	Alam 2011	Alam 2011
Non-Focal	Anderson 2010	Anderson 2010		
Non-Focal				Alderman 2019
Non-Focal				Alqazlan 2020 (thesis)
Non-Focal				Amon 2010 (thesis)
Non-Focal				Antelmann 2005
Non-Focal				Antoniali 2017
Non-Focal	Araújo 2020	Araújo 2020		Araújo 2020
Non-Focal		•		Ashburner 1999
Non-Focal				Aurass 2009
Non-Focal				Avican 2021
Focal				Bagautdinov 2008
Non-Focal				Bai 2017
Non-Focal				Bashir 2020
Non-Focal				Baysal 2013
Non-Focal	Belvin 2019	Belvin 2019		Belvin 2019
Non-Focal				Bermudez 2015
Non-Focal				Bhasin 2008
Non-Focal				Bhatraju 2015
Non-Focal	Bleichert 2020	Bleichert 2020		Bleichert 2020
Non-Focal				Boot 2016
Non-Focal				Bootsma 2010 (patent)
Non-Focal				Boswell 2018
Non-Focal				Breker 2014
Non-Focal				Brosnahan 2019
Non-Focal				Brosnahan 2020 (thesis)
False Positive	Brady 1989		Brady 1989	<u> </u>
Non-Focal				Bruce 2012 (patent)

Non-Focal				Burnstein 2016
Non-Focal	Byrne 2014		Byrne 2014	Burnstein 2010 Byrne 2014
Non-Focal	Byffle 2014		Byffie 2014	Camp 2016
Non-Focal				Cardenas 2009
Non-Focal				
Non-Focal				Chang 2016 Charlton 2015
Non-Focal				Chauhan 2015
Non-Focal				Chen 1999
Focal				Chen 2010a
Non-Focal				Chen 2010b
Focal	Chen 2014	Chen 2014	Chen 2014	Chen 2014
Non-Focal				Cherkas 2016
Non-Focal				Chiesa 2020
Focal	Choi 2013	Choi 2013		Choi 2013
Non-Focal				Chung 2013
Non-Focal				Codrich 2019
Non-Focal				Comas 2008
Non-Focal				Conn 2017
Focal				Constantine 2009
Non-Focal	Cox 2020		Cox 2020	
Non-Focal	Crapoulet 2006	Crapoulet 2006		Crapoulet 2006
Non-Focal				Cury 2020
Non-Focal	Czubat 2020	Czubat 2020		
Non-Focal	Da Costa 2018	Da Costa 2018		Da Costa 2018
Non-Focal				Daas 2016
Non-Focal				Daas 2018
Non-Focal				Dalia 2011 (thesis?)
Non-Focal				Damianou 2020
Non-Focal	Danelishvili 2005	Danelishvili 2005		Danelishvili 2005
Non-Focal	Danensiiviii 2005			Daou 2019
Non-Focal				Dapas 2019 (thesis)
Non-Focal				Dapas 2019 (mesis) Dartigalongue 2001
Non-Focal				Degnan 2005
Non-Focal	D. C. 1. 2015	D. G. 1: 2015		DeLoney 2002
Non-Focal	De Sordi 2015	De Sordi 2015	Dí M-:/- 2000	Día- Malía 2000
Non-Focal	Díaz-Mejía 2009		Díaz-Mejía 2009	Díaz-Mejía 2009
Non-Focal				Dickey 2013 (thesis)
Non-Focal				Ding 2015
Non-Focal				Dionisio 2016
Non-Focal				Ditse 2017
Non-Focal	Dong 2008	Dong 2008		Dong 2008
Non-Focal				Drummelsmith 2007
Non-Focal				Ducret 2021
Non-Focal				Dudin 2017
Non-Focal				Duncan 2017
Non-Focal				Dunman 2001
Non-Focal				Duzyj 2014
Non-Focal				Edwards 2013
Non-Focal				El Bounkari 2009
Non-Focal	Esser 2008		Esser 2008	Esser 2008
Non-Focal				Facciuolo 2013
Non-Focal				Facciuolo 2016
Non-Focal				Falkenberg 2019
Non-Focal				Fang 2010
Non-Focal	Fenner 2019	Fenner 2019		Fenner 2019
Non-Focal				Fontes 2018 (thesis)
Non-Focal				Freiberg 2016
Focal	Fujishiro 2014	Fujishiro 2014	Fujishiro 2014	Fujishiro 2014
Non-Focal	1 4/15/11/0 2017	1 4/10/11/2017	1 4/101110 2017	Fujishiro 2015
11011-1 UCal	L	1		1 4/15/11/0 2015

Non-Focal				Fujishiro 2016
Non-Focal				Fulishina 2020
Non-Focal	Furuta 2010		Furuta 2010	Furuta 2010
Non-Focal	Furuta 2010		Furuta 2010	Galperin 2010
Non-Focal	Gangaiah 2013		Gangaiah 2013	Galperni 2010
Non-Focal	Gangaiah 2013		Gangaiah 2013	Gangaiah 2014
Non-Focal	Galigatali 2014		Galigaiali 2014	Gao 2019
Non-Focal	Gaudet 2011		Gaudet 2011	040 2019
Non-Focal	Gawinecka 2012	Gawinecka 2012		Gawinecka 2012
Non-Focal	Gawinecka 2012	Gawinecka 2012		Geislet 1992
Non-Focal				Geyer 2019
Non-Focal	Ghaemmaghami 2003	Ghaemmaghami 2003		Ghaemmaghami 2003
Non-Focal		Onachinaghann 2003		Ghersi 2011
Non-Focal				Giannangelo 2018 (thesis)
Non-Focal				Gifford 2000
Non-Focal				Giotti 2019
Non-Focal				Giovannelli 2017
Non-Focal				Girinathan 2018
Non-Focal				Giuffrida 2006
Focal	Godsey 2007	Godsev 2007	Godsev 2007	Godsey 2007
Non-Focal	Gousey 2007	Cousey 2007	Gousey 2007	Gooderham 2008 (thesis)
Non-Focal				Graf 2018
Non-Focal	Gupta 2017	Gupta 2017		Gupta 2017
Non-Focal	Ouplu 2017	Supta 2017		Gupta 2017 Gupta 2018
Focal				Hadano 2001
Non-Focal				Hadj-Hamou 2010 (thesis)
Non-Focal	Han 2006		Han 2006	Han 2006
Non-Focal	Happonen 2019	Happonen 2019	Happonen 2019	Happonen 2019
Non-Focal				Harrison 2021
Non-Focal				Hart 2018
Non-Focal				Hayles 2013
Non-Focal				He 2019
Non-Focal	Hendrickson 2010	Hendrickson 2010		
Non-Focal				Hermans 2008 (patent)
Non-Focal				Hews 2019
Non-Focal				Hladik 2019
Non-Focal	Ho 2015	Ho 2015		Но 2015
Non-Focal				Но 2016
Non-Focal				Hossain 2021
Non-Focal	Huang 2015	Huang 2015	Huang 2015	
Non-Focal				Huang 2019
Non-Focal				Huston 2017 (lecture)
Non-Focal				Huttlin 2010
Non-Focal				Iyer 2016
Non-Focal				Jackson 2013
Non-Focal				Jama 2014 (thesis)
Non-Focal				Jiang 2013
Non-Focal	Jiang 2023	Jiang 2023	Jiang 2023	
Non-Focal				Johnson 2018
Non-Focal				Johnson 2019 (preprint)
Non-Focal				Jones 2006
Non-Focal	L 1:2020			Jostes 2019 (thesis)
Non-Focal	Joshi 2020		Joshi 2020	
Non-Focal				Jothi 2006
Non-Focal				Kai 2016
Non-Focal	V 1:0001	V 1:0001		Kalari 2013
KOOOL	Kanesaki 2021	Kanesaki 2021	1	
Focal Focal				Karniely 2006

Non-Focal				Kaysser 2010
Non-Focal				Kaysser 2010 Kerns 2020 (thesis)
Non-Focal				Kim 2018
Non-Focal	Kölbel 2019	Kölbel 2019		Kim 2018 Kölbel 2019
		Kolbel 2019	IZ 11 2004	
Non-Focal	Kolker 2004	Karan 2012	Kolker 2004 Kuan 2013	Kolker 2004
Focal	Kuan 2013	Kuan 2013	Kuan 2015	Kuan 2013
Non-Focal				Kumar 2018
Non-Focal				Kusonmano 2018
Non-Focal	L L L . D . 2000		L L L L D 2000	Kwon 2005
Non-Focal	Labandeira-Rey 2009		Labandeira-Rey 2009	Labandeira-Rey 2009
Focal	Ladner 2003		Ladner 2003	Ladner 2003
Focal	-			Lamba 2013 (poster abstract)
Non-Focal	Lasserre 2006			
Non-Focal	Leang 2005	Leang 2005	Leang 2005	Leang 2005
Non-Focal				Lee 2008 (thesis)
Non-Focal	Li 2010	Li 2010	Li 2010	Li 2010
Non-Focal	Li 2017	Li 2017		Li 2017
Non-Focal	Liang 2013	Liang 2013	Liang 2013	Liang 2013
Non-Focal				Lie 2013
Non-Focal				Lin 2004
Non-Focal				Lin 2019
Non-Focal				Liu 2013
Non-Focal				Liu 2014
Non-Focal				Liu 2019
Non-Focal				Lively 2010
Non-Focal	Livengood 2008		Livengood 2008	Livengood 2008
Non-Focal				Long 2020
Non-Focal				López-Madrigal 2013
Non-Focal				Lu 2015
Non-Focal				Luo 2005 (thesis)
Non-Focal				Luo 2007
Non-Focal				Lv 2021
Non-Focal				Mahdi 2013
Non-Focal				Makarova 2010
Non-Focal				Malan-Müller 2020
Focal				Malik 2020 (preprint)
Non-Focal	Malik-Kale 2008	Malik-Kale 2008	Malik-Kale 2008	Malik-Kale 2008
Focal	Manan 2018	Manan 2018	Manan 2018	Manan 2018
Non-Focal				Manzano-Marin 2019
Non-Focal	Manzourolajdad 2015	Manzourolajdad 2015		Manzourolajdad 2015
Focal	Martens 1996	Martens 1996		Martens 1996
Non-Focal				Martin 2002
Non-Focal				Martin 2018
Non-Focal			1	Martinot 2018
Non-Focal		1	ł	Massart 2015
Non-Focal		1	ł	McBrearty 2019 (thesis)
Non-Focal			1	McKee 2018
Non-Focal				McKenzie 2012
Non-Focal	McKenzie 2015	McKenzie 2015		McKenzie 2012
False Positive	McLeod 2017	McLeod 2017		
Non-Focal	Medrano 2009	101000 2017	Medrano 2009	Medrano 2009
Non-Focal			1/10/1/01/02/007	Meibom 2004
Non-Focal	Melian 2021	Melian 2021	Melian 2021	
Focal				Merla 2004
Non-Focal				Meshram 2019
				Minias 2015
Non-Focal				
Non-Focal				Minias 2018
Non-Focal	Managa 2019	Managa 2019		Moreira 2008
Non-Focal	Moreno 2018	Moreno 2018		Moreno 2018

Non-Focal				Morgonstern 2017
Non-Focal				Morgenstern 2017 Munoz 2020
Non-Focal				Nalls 2009
Non-Focal	Nagui 2015		Nagyi 2015	Naqvi 2015
	Naqvi 2015		Naqvi 2015	Naqvi 2013
Non-Focal Non-Focal	Naqvi 2018		Naqvi 2018	Nelson 2007
Non-Focal	Neville 2020	Navilla 2020	Neville 2020	INEISOII 2007
	Neville 2020	Neville 2020	Neville 2020	Ni-1 2015
Non-Focal				Niehaus 2015
Non-Focal				Nowlan 2020
Non-Focal				Nowlan 2021
Non-Focal				O'Connor 2012
Non-Focal				Ogura 2010
Non-Focal				Ogura 2016
Non-Focal Focal				Ogura 2018 Ogura 2019
Non-Focal				O'Hanlon 2011
Non-Focal	D 2005	D 2005		Oja 2018
Non-Focal	Page 2005	Page 2005		Page 2005
Non-Focal	Demant 2016	Dame (2016		Pan 2012
Non-Focal	Parente 2016	Parente 2016		Parente 2016
Non-Focal				Park 2007
Non-Focal				Passalacqua 2012
Non-Focal				Patel 2015
Non-Focal	D 1 2017	D 12017		Pattrick 2019
Non-Focal	Paul 2017	Paul 2017		Paul 2017
Non-Focal	D : 2005	D : 2005		Peng 2019
Non-Focal	Pereira 2005	Pereira 2005		
Non-Focal	Pereira 2011	Pereira 2011		Pereira 2011
Non-Focal				Perez-Samper 2018
Non-Focal	D : 11 2001			Phuphisut 2018
Non-Focal	Pompesiello 2001		Pompesiello 2001	Pompesiello 2001
Non-Focal	Porcelli 2013	Porcelli 2013	Porcelli 2013	Porcelli 2013
Non-Focal	D			Prahlad 2020
Non-Focal	Prunetti 2016		Prunetti 2016	Prunetti 2016
False Positive	Prusty 2013		Prusty 2013	
Non-Focal				Pulido 2020 (thesis)
Non-Focal				Punekar 2016
Focal				Qijing 2007 (translated)
Non-Focal				Qiu 2019
Non-Focal				Qu 2020
Non-Focal				Quigley 2014
Non-Focal				Rachel 2012
Non-Focal				Rahmani-Badi 2016
Non-Focal				Rajathei 2013
Non-Focal				Raman 2006
Non-Focal				Ramaniuk 2018
Non-Focal	D10001	D 1 2021	D 1 2021	Rankin 2002
Focal	Reed 2021	Reed 2021	Reed 2021	Reed 2021
Non-Focal	Reinders 2006	Reinders 2006		Reinders 2006
Non-Focal				Resch 2016
Non-Focal				Rijlaarsdam 2016
Non-Focal				Risler 2012
Non-Focal	Doomay 2000		Doomar 2000	Rollins 2006
Non-Focal	Rooney 2009	Doomay 2011	Rooney 2009	Rooney 2009
Non-Focal	Rooney 2011	Rooney 2011	Rooney 2011	Rooney 2011
Non-Focal				Rosenbaum 2011
Non-Focal				Ryan 2005
Non-Focal				Rylander 2011
Non-Focal	1			Saez 2019

Focal	Saikatendu 2006	Saikatendu 2006		Saikatendu 2006
Non-Focal				Sailani 2015
Non-Focal	Sainsbury 2012		Sainsbury 2012	Sainsbury 2012
Non-Focal	Sumsoury 2012		Sumsoury 2012	Samper 2019
Non-Focal				Samudrala 2009
Non-Focal				Sarker 2012
Non-Focal				Schlenker 2011 (thesis)
Non-Focal				Schneeweiss 2020
Non-Focal				Schrader 2016
Non-Focal				Selby 2017
Non-Focal				Selim 2021
Focal	Sergeeva 2018		Sergeeva 2018	Sergeeva 2018
Non-Focal	Shahbaaz 2013		Shahbaaz 2013	Shahbaaz 2013
Non-Focal	Shanbaaz 2013		Shanoadz 2013	Shakya 2020 (preprint)
Non-Focal				Shimada 2011
Non-Focal				Shin 2011
Non-Focal				Shulami 1999
Non-Focal				Shulami 2007
Focal	Shulami 2014		Shulami 2014	Shulami 2007 Shulami 2014
Focal Non-Focal	Silulaliii 2014		Silulaini 2014	Sigdel 2011
Non-Focal				Signer 2011 Simon 2016
Non-Focal				Simonis 2018
Non-Focal				
				Simpson 2015 Skottman 2005
Non-Focal				
Non-Focal	0 1 2010		0 1 2010	Skunca 2013
Non-Focal	Spinola 2010		Spinola 2010	Spinola 2010
Non-Focal	Stabler 2010			Stabler 2010
Non-Focal				Starkey 2009
Non-Focal				Stenger 2019 (thesis)
Non-Focal				Stern 2020
Non-Focal				Steyn 2017 (thesis)
Non-Focal				Strillacci 2020
Non-Focal				Su 2012
Non-Focal				Suzuki 2006
Non-Focal				Ta 2010
Non-Focal	-			Tajkarimi 2017
Focal	Tascou 2000	Tascou 2000		Tascou 2000
Focal	Tascou 2003	Tascou 2003		Tascou 2003
Non-Focal				Thankam 2018
Non-Focal				Thorenoor 2010
Non-Focal				Tian 2014
Non-Focal				Tian 2020
Focal	Tomoike 2009		Tomoike 2009	Tomoike 2009
Non-Focal				Tran 2018
Non-Focal				Tsuiko 2016
Non-Focal				Turlin 2006
Non-Focal				Uxa 2019
Non-Focal	van Diemen 2017	van Diemen 2017		van Diemen 2017
Non-Focal	Van Dyke 2016		Van Dyke 2016	Van Dyke 2016
Non-Focal				van Ouwerkerk 2019
Non-Focal				Volha 2019 (thesis)
Non-Focal			Wagley 2014	Wagley 2014
Non-Focal				Waldor 2013
Non-Focal				Wang 2010
False Positive	Wang 2012	Wang 2012		Wang 2012
Non-Focal				Wang 2015
Non-Focal				Wang 2018
Non-Focal				Wang 2021
Non-Focal				Watkins 2008

Non-Focal			Watkins 2010
Non-Focal			Wei 2001
Non-Focal	Wei 2016	Wei 2016	Wei 2016
Non-Focal			Weinzierl 2008
Non-Focal			Wilkinson 2018 (thesis)
Non-Focal	Williams 2013	Williams 2013	Williams 2013
Non-Focal			Willis 2010 (thesis)
Non-Focal			Winer 2011
Non-Focal			Wittchen 2018
Non-Focal			Wu 2019
Non-Focal			Xi 2011
Non-Focal			Xia 2017
Non-Focal			Xiang 2012
Focal			Xie 2019
Non-Focal	Xu 2021	Xu 2021	
Non-Focal			Yan 2014
Non-Focal			Yan 2020
Non-Focal			Yang 2013
Non-Focal			Yao 2016
Focal			Yu 2015
Non-Focal			Yu 2018
Non-Focal			Yuan 2006
Non-Focal			Zamanian-Daryoush 2020
Non-Focal			Zhao 2018
Non-Focal			Zheng 2020
Non-Focal			Zuccotti 2008

Supplementary Elaborations

Elaboration S1. Explanation of XynX "productive ignorance" example.

XynX, a DUF34 homolog of *Geobacillus stearothermophilus* T-6, was identified and described by a couple researcher groups over the course of several papers between the years 1994 and 2016 [1-6]. Upon identification, because of its genomic location and position within an operon of biosynthetic interest, it was named with respect to neighboring genes. This naming and system-level characterization were completed almost entirely independent of the preceding literature of the protein family of which it was a recognizable member. Additionally, and although the genomic information of the organism had been purportedly annotated and respective data made publicly available, it was-upon review of Reed et al. Biomolecules 2021 [7]—discovered that the genome and related annotation information was not as accessible as had been claimed. It had been observed that the complete genome and any annotated, encoded proteins had been intermittently submitted and updated by the submitter over the course of 15 years (1993-2008). Although the EuropePMC Data subsection for the earliest publication [1] lists 51 links to individual UniProt entries and a single **ENA** link to 57 nucleotide records (accessed May 20, 2022; https://europepmc.org/article/MED/8031084): the original publisher does not provide links to this information (now hosted website for the parent journal, at the ASM: https://doi.org/10.1128/aem.60.6.1889-1896.1994); this paper does not provide any systematic accession identifiers for the data allegedly being reported (i.e., the genome); and only 56 of the 57 EuropePMC reported genes are actually listed in the ENA database under the queryable publication-associated ID, 8031084.

It is hypothesized that this EuropePMC data being reported as being associated with this publication [1] is mistakenly attributed to this particular record. At this time, it is suspected that the accessioned data referenced in many of these online publications has been retroactively linked to these other works from the Gat et al. 1994 record and, subsequently, the Shulami et al. 2011 publication in the Journal of Bacteriology [5] (https://doi.org/10.1128/JB.00222-11; first cited by Alalouf et al. 2011 in the Journal of Biological Chemistry (PMID: 21994937) [4], methods section) with which a GenBank record was submitted for the 77,747 bp (linear) of the G. stearothermophilus T-6 genome that was reported to contain "the xylan, xylose, arabinan, and arabinose utilization region" relevant to the study. The record, GenBank: DQ868502.2, is titled "Geobacillus stearothermophilus strain T-6 genomic sequence", suggesting that the genome is completely sequenced; GenBank record is dated "26-JUL-2016". This record replaced the earlier GenBank record, DQ868502.1, titled "Geobacillus stearothermophilus strain T-6 xylan utilization gene cluster, complete sequence; and NAD(P)H-dependent flavin oxidoreductase (orfD) gene, partial cds" dated 01-FEB-2007. As per Gat et al., 1994, the earliest genomic record (annotated with only a few CDSs, but largely unannotated) was submitted to GenBank, DDBJ, and EMBL (ENA) with the accession identifier Z29080. The first in-text description of xynX and its encoded protein of the same name was in Shulami et al., Applied and Environmental Microbiology 2007 [3] and described as a "xylan utilization gene". In Alalouf et al., 2011 [4], XynX is described as "a putative regulatory gene with unknown function". It's putative cooperative regulation of xvnA expression alongside XvlR and CodY is described in more detail in Shulami et al., Journal of Biological Chemistry 2014 [5]. No description of the family to which XynX belongs is mentioned, no attempt to identify the associated protein family is made, despite assigning the protein a name and describing a functional association. Although these characterizations (and alias) were assigned with ignorance of the protein family, the functional association of XynX as a DUF34 member was not dissimilar from subsequent characterizations, for example, in Bacillus subtillis [8], and, therefore, cannot be discounted in its contributions to describing functional associations at the family level.

Supplementary Figures

Figure S1. Seed Information Assessment: establish understanding of starting information in preparation for the capture of published data.

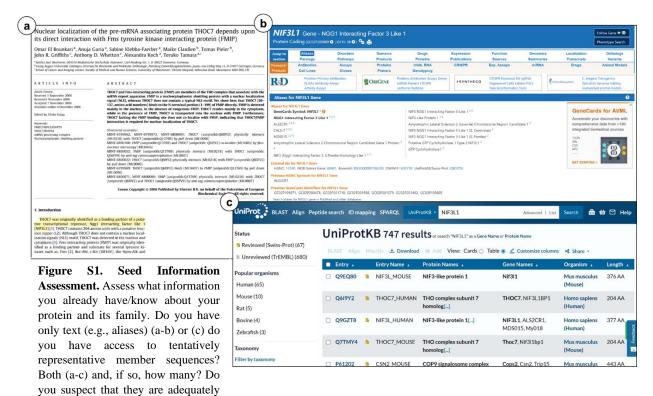
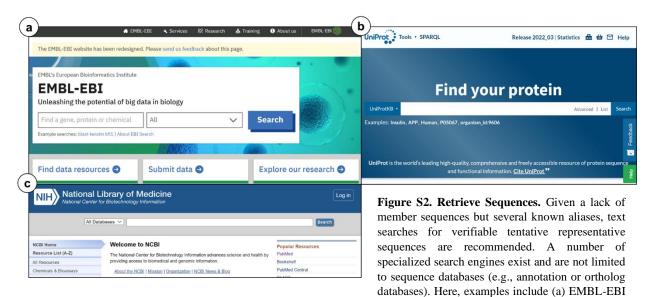


Figure S2. Retrieve sequences using known names/aliases.

representative of the family?



broad search, (b) UniProtKB, and (c) NIH National

Library of Medicine all-database search.

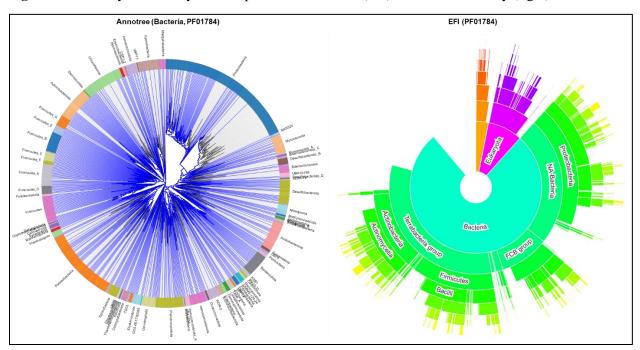


Figure S3a. Family-level analysis examples from Annotree (left) and EFI Taxonomy (right).

Figure S3b. Family-level analyses using various protein family classification databases (i.e., Pfam, CDD, OrthoMCL, InterPro, OrthoInspector, EggNOG, ENSEMBL tools), of which can be used to approximate architectural subgroups in concert with using representative sequences as inputs into the sequence-based search tools (e.g., i.e., PaperBLAST) for corpus retrieval.

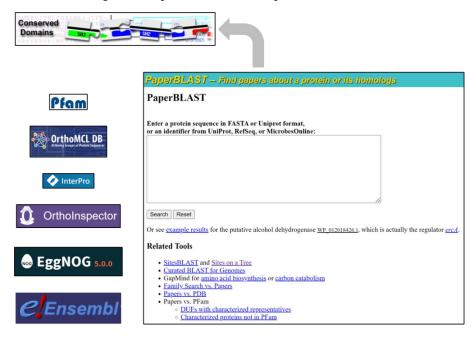


Figure S4. Example of over-propagation of root family attribution in EggNOG Database as a result of fusions. EggNOG aggregates data from other resources and often these collections accrue family members that are domain architectural variants and contain non-family sequence annotations. These annotations are not filtered out or flagged in any way and require the user to discern these erroneous aggregates on their own. (a) EggNOG root cluster of the DUF34 protein family. Added notation of screenshot: red bars underline the incorrect annotations aggregated from KEGG, one biologically incorrect based on published data, a second resulting from a fusion sequence of COG0327 and COG1579 [PMID: 34572495], and a third that is a product of a gene fusion of a eukaryotic DUF34 homolog and CIAO (COG2319). (b) Link-out to the first of three incorrect KO entries listed EggNOG listed KO, K22391, for the enzyme activity of GTP cyclohydrolase I; KO entry lists the corresponding COG, though in this case, no additional COGs are aggregated in KEGG Orthology. (c) Link-out to the second of three listed incorrect KOs of DUF34, that for an uncharacterized bacterial protein family, COG1579 (K07164), an annotation recognized as being associated with the DUF34 family by EggNOG—an association that is likely through the known fusion with the Wollinella succinogenes DUF34 homolog [PMID: 34572495] that may or may not have functionally diverged-while KO does not identify the source of this association. (d) Link-out to the third of three incorrect KOs listed in the entry card for LCOG0327 in EggNOG, CIAO (COG2319), also a product of fusion sequences within the DUF34 family. (e) COG0327 (child of root, LCOG0327) "Duplications profile" view.

a.

LCOG0327	GTP cyclohydrolase I [EC:3.5.4.16],uncharacterized protein,cytosolic iron-sulfur protein assembly protein 9672 proteins 8821 specie
oot)	CIAO1
fam domain	NIF3 (95.24%), zf-RING_7 (0.29%), Methyltransf_18 (0.07%)
mart domain	SIGNAL (0.32%), TRANS (0.20%), PAS (0.02%)
O slim	GO:0006351 (0.55%), GO:0006355 (0.55%), GO:0048856 (0.53%)
EGG pathway	ko01240 (1.35%), map00790 (1.35%), map01100 (1.35%)
EGG module	M00126 (1.35%)
EGG ortholog	<u>K22391</u> (1.35%), <u>K07164</u> (0.10%), <u>K24730</u> (0.01%)
EGG gene symbol	E3.5.4.16 (1.35%), K07164 (0.10%), CIAO1 (0.01%)
EGG gene name	GTP cyclohydrolase I [EC:3.5.4.16] (1.35%), uncharacterized protein (0.10%), cytosolic iron-sulfur protein assembly protein CIAO1 (0.01%)
CHILDREN OG	
COG0327	
COG1579	
COG3323	
D4X3B	
D6XEM	
D6XEM D72WI D9YGS	

b.

ORTHOLOGY: K22391

Entry		
	К22391 КО	All links
Symbol	E3.5.4.16	Ontology (4)
Name	GTP cyclohydrolase I [EC:3.5.4.16]	KEGG BRITE (2)
Pathway	map00790 Folate biosynthesis	GO (1) COG (1)
	map01100 Metabolic pathways	Pathway (7) KEGG PATHWAY (6)
	map01240 Biosynthesis of cofactors	KEGG MODULE (1)
Module	M00126 Tetrahydrofolate biosynthesis, GTP => THF	Chemical reaction (1)
Reaction	R00424 GTP 7,8-8,9-dihydrolase	KEGG ENZYME (1) KEGG REACTION (5)
	R00428 GTP 8,9-hydrolase R04639 2-amino-4-hydroxy-6-(erythro-1,2,3-trihydroxypropyl)	KEGG RCLASS (7)
	dihydropteridine triphosphate hydrolase	Gene (832) KEGG GENES (398)
	R05046 formamidopyrimidine nucleoside triphosphate amidohydrolase	KEGG MGENES (412)
	R05048 2,5-diaminopyrimidine nucleoside triphosphate mutase	RefGene (16) OC (6)
Brite	KEGG Orthology (KO) [BR:ko00001]	Literature (1)
	09100 Metabolism	PubMed (1) All databases (857)
	09108 Metabolism of cofactors and vitamins 00790 Folate biosynthesis	
	K22391 E3.5.4.16; GTP cyclohydrolase I	Download RDF
	Enzymes [BR:ko01000]	
	3. Hydrolases	
	3.5 Acting on carbon-nitrogen bonds, other than peptide bonds	
	3.5.4 In cyclic amidines 3.5.4.16 GTP cyclohydrolase I	
	K22391 E3.5.4.16; GTP cyclohydrolase I	
	BRITE hierarchy	
Other DBs	COG: COG0327 GO: 0003934	
Genes	LIMN: HKT17_05100	
	HPY: HP_0959	
	HEO: C694_04940	
	HEO: C694_04940 HPJ: jhp_0893	
	HEO: C694_04940 HPJ: jhp_0893 HPA: HPAG1_0943	
	HEO: C694_04940 HPJ: jhp_0893 HPA: HPAG1_0943 HPS: HPSH_05055	
	HEO: C694_04940 HPJ: jhp_0893 HPA: HPAG1_0943 HPS: HPSH_05055 HHP: HPSH112_04975	
	HEO: C694_04940 HPJ: jhp_0893 HPA: HPAG1_0943 HPS: HPSH105055 HHP: HPSH112_04975 HHQ: HPSH165_04880	
	HEO: C694_04940 HPJ: jhp_0893 HPA: HPAG1_0943 HPS: HPSH_05055 HHP: HPSH112_04975	
	HEO: C694_04940 HPJ: jhp_0893 HPA: HPAG1_0943 HPS: HPSH_05055 HHP: HPSH112_04975 HHQ: HPSH126_04880 HHR: HPSH417_04675	
	HEO: C694_04940 HPJ: jhp_0893 HPA: HPAG1_0943 HPS: HPSH_05055 HHP: HPSH112_04975 HHQ: HPSH112_04975 HHQ: HPSH169_04880 HHR: HPSH417_04675 HPG: HPG27_907	
Reference	HEO: C694_04940 HPJ: jhp_0893 HPA: HPAG1_0943 HPS: HPSH_05055 HHP: HPSH112_04975 HHQ: HPSH169_04880 HHR: HPSH417_04675 HPG: HPG27_907 > show all Taxonomy_UniProt	_
Reference Authors	HEO: C694_04940 HPJ: jhp_0893 HPA: HPAG1_0943 HPS: HPSH_05055 HHP: HPSH112_04975 HHQ: HPSH12_04975 HHR: HPSH417_04675 HPG: HPG27_907 > show all Taxonomy_UniProt	_
	HEO: C694_04940 HPJ: jhp_0893 HPA: HPAG1_0943 HPS: HPSH_05055 HHP: HPSH112_04975 HHQ: HPSH112_04975 HHQ: HPSH112_04675 HPG: HPG27_907 > show all Taxonomy UniProt PMID:23825549 Choi HP, Juarez S, Ciordia S, Fernandez M, Bargiela R, Albar JP,	
Authors	HEO: C694_04940 HPJ: jhp_0893 HPA: HPAG1_0943 HPS: HPSH_050855 HHP: HPSH112_04975 HHQ: HPSH169_04880 HHR: HPSH417_04675 HPG: HPG27_907 > show all Taxonomy UniProt PMID:23825549 Choi HP, Juarez S, Ciordia S, Fernandez M, Bargiela R, Albar JP, Mazumdar V, Anton BP, Kasif S, Ferrer M, Steffen M Biochemical Characterization of Hypothetical Proteins from Helicobacter pylori.	

c.

KEGG	ORTHOLOGY: K07164	Help
Entry	К07164 КО	All links
Symbol	K07164	Ontology (2)
Name	uncharacterized protein	KEGG BRITE (1)
Brite	KEGG Orthology (KO) [BR:ko00001] 09190 Not Included in Pathway or Brite 09194 Poorly characterized 99997 Function unknown K07164 K07164; uncharacterized protein BRITE hierarchy	COG (1) Gene (23853) KEGG GENES (257 KEGG MGENES (17 RefGene (3280) OC (20) All databases (238
Other DB	COG: COG1579	Download RDF
Genes	HPY: HP_0958 HEO: C694_04935 HPJ: jhp_0892 HPA: HPAG1_0942 HPS: HPSH_05050 HHP: HPSH112_04970 HHQ: HPSH417_04670 HPG: HPSH417_04670 HPF: HPF12_0954 *> show all	

All links
Ontology (2) KEGG BRITE (1) COG (1) Gene (23853) KEGG GENES (2733) KEGG MGENES (17980) RefGene (3280) OC (20) All databases (23855)
Download RDF

KECC	ORTHOLOGY: K24730	
Entry	К24730 КО	All links
Symbol	CIA01, CIA1	Ontology (2)
Name	cytosolic iron-sulfur protein assembly protein CIAO1	KEGG BRITE (2)
Brite	<pre>KEGG Orthology (KO) [BR:ko00001] 09180 Brite Hierarchies 09183 Protein families: signaling and cellular processes 04990 Domain-containing proteins not elsewhere classified K24730 CIAO1, CIAI; cytosolic iron-sulfur protein assembly protein Domain-containing proteins not elsewhere classified [BR:ko04990] WD40 repeat (MDR) domain-containing proteins Other WDR domain-containing proteins K24730 CIAO1, CIAI; cytosolic iron-sulfur protein assembly protein C</pre>	Gene (1314) KEGG GENES (952) KEGG MGENES (362) All databases (1316) Download RDF
	BRITE hierarchy	
Other DBs	COG: COG2319	
Genes	HSA: 9391(CIAO1) PTR: 741901(CIAO1) PPS: 106975574(CIAO1) GGO: 101151871(CIAO1) PON: 100443810(CIAO1) NLE: 100600141(CIAO1) HMH: 116483514(CIAO1) MCC: 705318(CIAO1) MCF: 102120380(CIAO1) MTHB: 126934314 > show all Taxonomy_UniProt	
Reference	PMID:9556563	
Authors Title	Ciao 1 is a novel WD40 protein that interacts with the tumor suppressor protein WT1.	
Journal	DOI:10.1074/jbc.273.18.10880	
Sequenc	e [hsa:9391]	

e.

lacteria)			
fam domain	NIF3 (94.82%), zf-RING_7 (0.34%), Methyltransf_18 (0.07%)		
mart domain	SIGNAL (0.30%), TRANS (0.18%), PAS (0.02%)		
iO slim	G0:0006281 (0.05%), G0:0051604 (0.01%), G0:0065003 (0.01%)		
EGG pathway	ko01240 (1.59%), map00790 (1.59%), map01100 (1.59%)		
EGG module	M00126 (1.59%)		
EGG ortholog	K22391 (1.59%), K07164 (0.12%)		
EGG gene symbol	E3.5.4.16 (1.59%), K07164 (0.12%)		
EGG gene name	GTP cyclohydrolase I [EC:3.5.4.16] (1.59%), uncharacterized protein (0.1	2%)	
PARENT OGS		60SJM	
6V4A8		60ZGF	
6WAY1		62NRV	
6YOMR		68Q8E	
EAAIX		68U7W 693DK	
COG0327		69Z0V	
COG3323 LCOG0327		6AANE	
		6AQA8	
LCOG1579 G members Taxo	nomic profile Functional profile Duplications profile	Tree and alignment	
LCOG1579 G members Taxor	nomic profile Functional profile Duplications profile		
LCOG1579 G members Taxo Download list			DUPLICATIONS
LCOG1579 G members Taxo Download list TAXONOMY Microgenomates group	nomic profile Functional profile Duplications profile		DUPLICATIONS 1
LCOG1579 G members Taxo Download list TAXONOMY Microgenomates group Rheinheimera (67575)			DUPLICATIONS 1
LCOG1579 G members Taxo Download list TAXONOMY Microgenomates group Rheinheimera (67575) Vibrio (662)			DUPLICATIONS 1 1
LCOG1579 G members Taxo Download list TAXONOMY Microgenomates group Rheinheimera (67575) Vibrio (682) Vibrionaceae (641)	becterium RIFCSPLOWO2, 01, FULL 46, 13 (1817759)		DUPLICATIONS 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
COGIS79 G members Taxo Download list TAXONOMY Microgenomates group Rheinheimera (67373) Vibriancecce (641) Xenorhabdus eapokensi	beaterium RIFCS910W02, 01, FULL 46, 13 (1817750) x (1873482)		DUPLICATIONS 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
COGIS79 G members Taxo Download list TAXONOMY Wicrogenomates group Rhinhelmera (67373) Vibrio (662) Vibrionecee (641) Kikronhaddus espoknsil	becterium BIFCS9E0W02_01_FULL_46_13 (1817750) s (1873482) 0)		DUPLICATIONS 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
COGIS79 G members Taxo Download list TAXONOMY Microgenomates group Michine (662) Vibrien (662) Vibrien (662) Vibrien (662) Vibrien (662) Conducting (62 Condidentus Conducting (62 Conducting (62 Conducting (62 Condidentus Conducting (62 Conducting	bacterium NIFCSFLOWO2. 01. FULL 48, 13 (1817759) x (1873482) D3) Accumulibacter (7619654)		DUPLICATIONS 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
COGIS79 G members Taxo Download list TAXONOMY RheInheimera (67573) Vibrio (6642) Vibrio (6642) Vibrionacede (641) Xenorhabdus eapokensi Nitrosemondelis (126u medosymbilot Candi Bathyr	besterium RIFCSPOWO2, 01, FULL, 46, 13 (1817750) x (1873482) 20) Accumulikecter (2619054) moleius asptemdierum str. Mysijin kevall (1203921)		DUPLICATIONS 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
Constant G members Taxon Download list TAXONOMY Microgenomates group: Rheinheimera (67575) Wibrion (682) Wibrion (682) Wibrion discussed (41) unclassified Condidates (12) unclassified Condidates (12) Consultation of Bathyu Consultation	besterium RIFCSPOWO2, 01, FULL, 46, 13 (1817750) x (1873482) 20) Accumulikecter (2619054) moleius asptemdierum str. Mysijin kevall (1203921)		DUPLICATIONS 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
LCOG1579 G members Taxo Download list TAXONOMY Microgenomates group Microgenomates (6737) Vikiro (662) Vikironcece (641) Xenorhabdus eapokensi Nitrosomondules (202 conexibect models) Conexibect models)	besterium RIFCSPOWO2, 01, FULL, 46, 13 (1817750) x (1873482) 20) Accumulikecter (2619054) moleius asptemdierum str. Mysijin kevall (1203921)		DUPLICATIONS 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
COG1579 G members Taxon Download list TAXONOMY Microgenometes group Microgenometes group Wibrionecee (647) Wibrionecee (647) Wibrionecee (647) Micromohdule sepokensi Nifrsenonadiele (120 unclassified Candidatus literatorial de althy Conesibacter woele DS Paenibacillus (44249) Becillus (138)	besterium RIFCSPOWO2, 01, FULL, 46, 13 (1817750) x (1873482) 20) Accumulikecter (2619054) moleius asptemdierum str. Mysijin kevall (1203921)		DUPLICATIONS 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
LCOG1579 G members Taxo Download list TAXONOMY Microgenomates group Rheinhelmere (67575) Vibrie (682) Vibrienceces (641) Xinonheddus epokensi Nitrosomonadaes (200 Constibuter wessel D5 Peenihedilus (4149) Bacillus (1368) Clastridiacees (1399)	becterium RIFCSPLOWO2, 01, FULL, 46, 13 (1817750) s (1873-482) DJ Accumulibacter (2619054) modialus septemdierum str. Mysijin knoll (1203921) M 14684 (489383)		DUPLICATIONS 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
COG1579 G members Taxon Download list TAXONOMY Microgenometes group Microgenometes group Wibrionecee (647) Wibrionecee (647) Wibrionecee (647) Micromohdule sepokensi Nifrsenonadiele (120 unclassified Candidatus literatorial de althy Conesibacter woele DS Paenibacillus (44249) Becillus (138)	becterium RIFCSPLOWO2, 01, FULL, 46, 13 (1817750) s (1873-482) DJ Accumulibacter (2619054) modialus septemdierum str. Mysijin knoll (1203921) M 14684 (489383)		DUPLICATIONS 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
LCOG1579 G members Taxo Download list TAXONOMY Microgenomates group Rheinhelmere (67575) Vibrie (682) Vibrienceces (641) Xinonheddus epokensi Nitrosomonadaes (200 Constibuter wessel D5 Peenihedilus (4149) Bacillus (1368) Clastridiacees (1399)	becterium RIFCSPLOWO2, 01, FULL, 46, 13 (1817750) s (1873-482) DJ Accumulibacter (2619054) modialus septemdierum str. Mysijin knoll (1203921) M 14684 (489383)		DUPLICATIONS 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

d.

Figure S5. WorldWideScience.org-, ScienceResearch.com-generated diagrams of keywords. Visualized yields illustrate coincidental homonyms of NIF3 (DUF34).

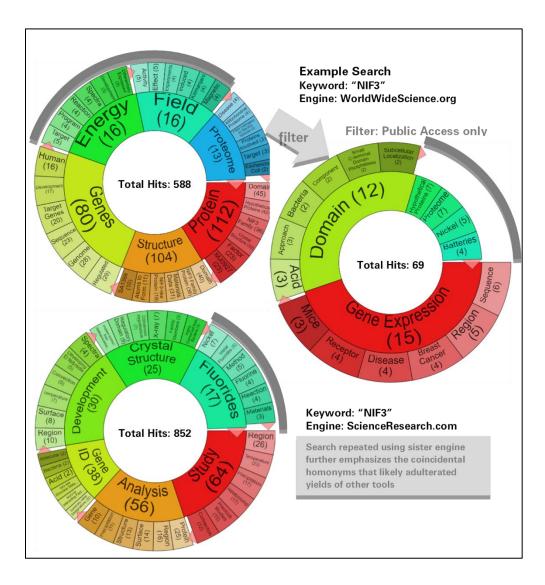
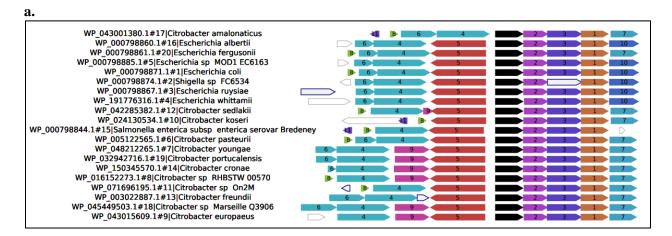


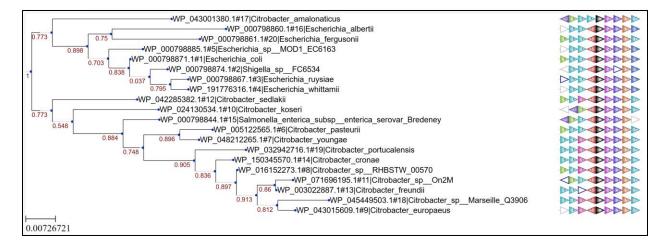
Figure S6. Coincidental homonyms are challenges for text-based search engines, even those as specialized as PubTator. Top screenshot shows PubTator top hits for "GTP cyclohydrolase 1 type 2" search, while the bottom screenshot shows the top hits for alias synonym "GTP cyclohydrolase I type 2". These searches should hypothetically retrieve identical results; however, in addition to the false synonyms being retrieved due to the automatic generalization of the original search term to "GTP cyclohydrolase", the engine simultaneously retrieves a correct result in only one of the two case uses (the bottom instance, "GTP cyclohydrolase I type 2"), suggesting that the tool is unable to discern singular "I" usage as the Roman numeral for 1, which is nomenclature practice frequently used in the naming of genes/proteins.

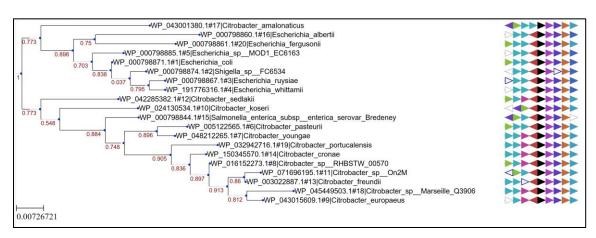
PubTator	"GTP cyclohydrolase 1 type 2"	NIH) NLM
	Showing 1 to 15 of 145 publications. < Page 1 of 10 >	ANONYMOUS +3 LOGIN
False Positive	2023 Endothelial cell vasodilator dysfunction mediates progressive pregnancy-induced hypertension in endothelial cell tetrahydrobiopterin deficient mice. Chusiphichai S, Dickmon Y Douglas G + Vascul Pharmacol ≧ Bocht. [2]pund	Collections
False Positive	279400400005 Time-of-Day-Dependent Effects of Bromocriptine to Ameliorate Vascular Pathology and Metabolic Syndrome in SHR Rats Held on High Fat Diet. Erothi M, Zhang Y., Cincotta M+ int J MalSci ≹BootM. \$20ppmd	FILTERS
False Positive	Apolipoprotein A-I mimetic peptide inhibits atherosclerosis by increasing tetrahydrobiopterin via regulation of GTP- cyclohydrolase 1 and reducing uncoupled endothelial nitrie oxide synthase activity. Nmg DS, Mol.—Ou DS • Alherosclerosis ≧ Bocom: "Depend	
False Positive	2021 2021 Overexpression of Riboflavin Excretase Enhances Riboflavin Production in the Yeast Candida famata. Toyrulnyk AO, Fedrowych DVSbirry AA Methods Mot Blot ≧ Biocold, CDpend	
PubTator	"GTP cyclohydrolase I type 2"	NIH) NLM
	Showing 1 to 15 of 223 publications.	ANONYMOUS HOLDGIN
False Positive		Collections Coll
True Positive	2 PMG3457495 - PMG849502 2021 Comparative Genomic Analysis of the DUF34 Protein Family Suggests Role as a Metal Ion Chaperone or Insertase. Reed Cl, Hutinet G, de Crécy-Lagard V + Biomelecules ▲ Biocom.	FILTERS O full-text only
False Positive	Time-of-Day-Dependent Effects of Bromocriptine to Ameliorate Vascular Pathology and Metabolic Syndrome in SHR Rats Held on High Fat Diet. Erroki M, Zhang Y., Chechta AH • Int J Mol Sci ▲ Becota, Cippend	

Figure S7. FlaGs (WebFlaGs) example outputs for NCBI BLASTp results for COG0327 (DUF34) family member of *E. coli*, YbgI (POAFP6); list of RefSeq Accession IDs was used as the WebFlaGs input. a) Shows the PDF output showing the gene neighborhoods of the input sequences; cluster information is also available in text format accompanying this graphical output (not shown). b) Flanking genes tree graphical result option 1. c) Flanking genes tree graphical result option 2. d) MAFFT sequence alignment tree graphical result. The latter panels, b-d, are available as components constituting the exportable zipped folder available upon FlaGs output receipt.

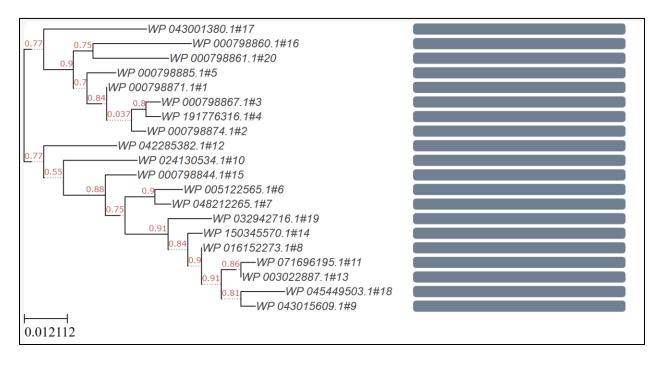


b.



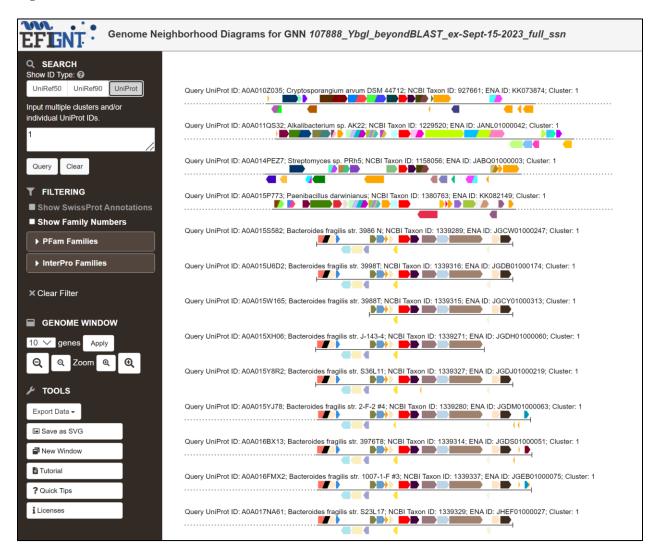






c.

Figure S8. EFI-GNT



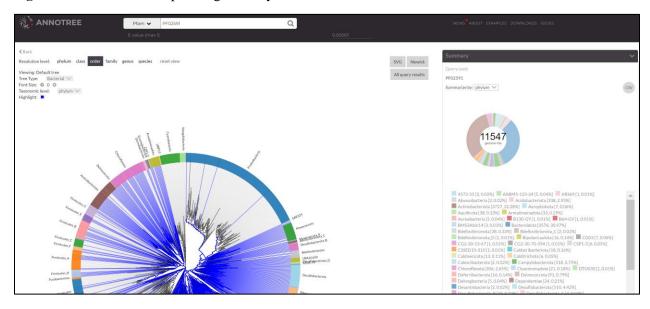
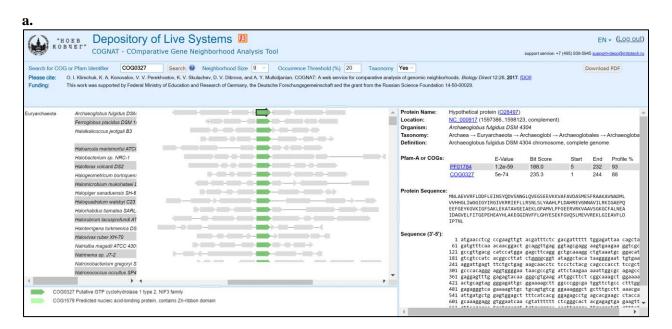


Figure S9. Annotree example: single family, PF02591, in bacteria.

Figure S10. COGNAT COG-based gene neighborhood and physical clustering tool example output for COG0327 (DUF34 protein family). a) shows the raw output, while b) shows a portion of the pdf export made available via link-out download.



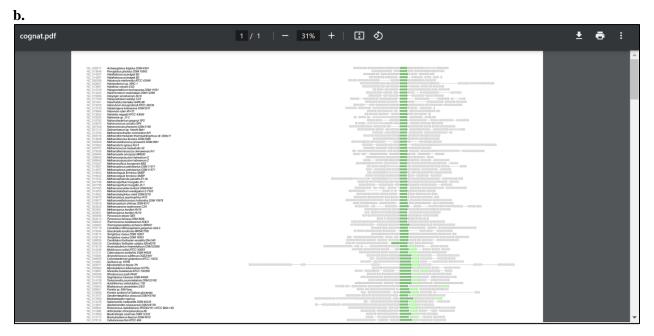


Figure S11. SubtiWiki (CoreWiki) beta feature of the database's Genomic Neighborhood Comparison viewer of the DUF34 homolog of B. subtilis, YqfO (BSU_25170).

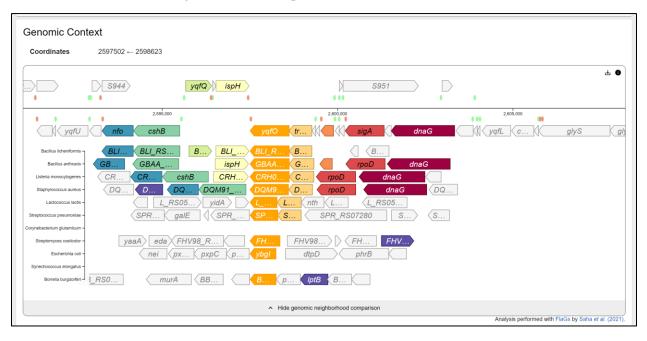
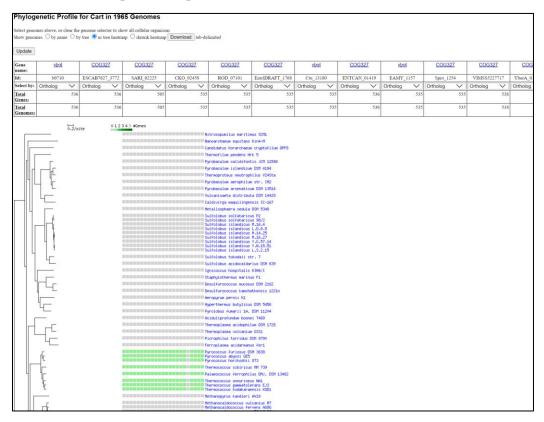


Figure S12. Example tree output of MicrobesOnline.



Ŭ		creengrab		rating un			•	•	•		•			
differentially with multiple targets								(MicrobesOnline).						
Phylog	Phylogenetic Profile for Cart in 1965 Genomes													
	Select genomes above, or clear the genome selector to show all cellular organisms Show genomes O by name O by tree S as tree heatmap O shrunk heatmap Download tab-delimited Update													
Gene name:	<u>ybgl</u>	<u>COG327</u>	<u>COG327</u>	<u>COG327</u>	<u>COG327</u>	<u>COG327</u>	<u>ybgl</u>	<u>COG327</u>	<u>ybgl</u>	<u>COG327</u>	<u>COG327</u>			
Id:	b0710	ESCAB7627_3772	SARI_02225	CKO_02458	ROD_07101	EntclDRAFT_1768	Ctu_13180	ENTCAN_01419	EAMY_1157	Spro_1254	VIMSS5227717			
Select by:	>= 100 bits ~	>= 200 bits ~	>= 300 bits ~	>= 400 bits ~	\geq 500 bits \vee	COG327 V	PTHR13799 ~	PF01784 ~	SSF102705 V	TIGR00486 V	Ortholog ~			
Total Genes:	536	536	505	535	535	535	535	536	535	535	538			
Total Genomes:	536	536	505	535	535	535	535	536	535	535	538			
	0.3	d 2/site)1234>#Genes		Nitrosop	umilus maritimus SC	11							

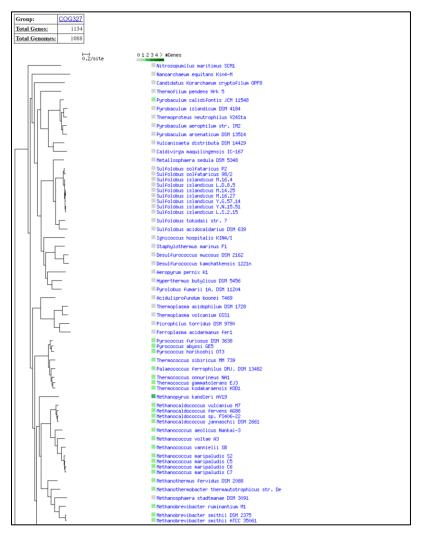
· C · 1

• . •

. 1 1

1

Figure S14. Screengrab demonstrating navigation to family-specific trees in MicrobesOnline via the "Gene Info" tab of each entry.



011 0

1 1

· ·

1.00

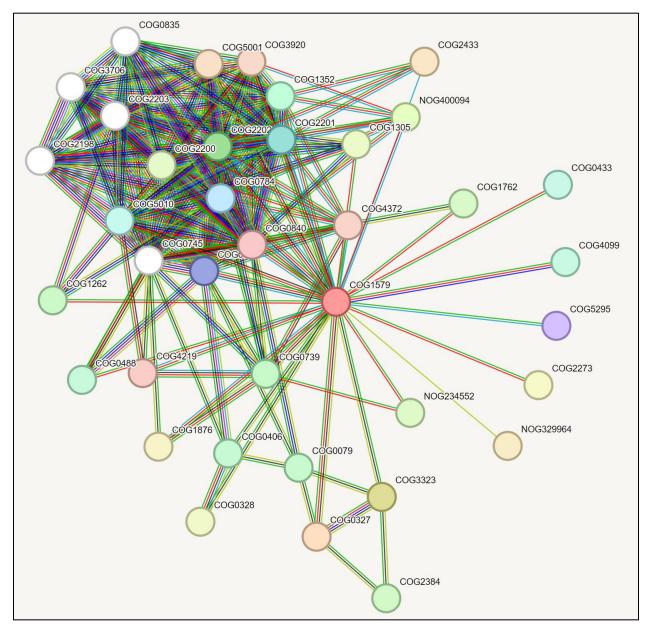
Nanoarchaeum equitans Kin4-M

Figure S15. "Families" (COGs) in STRING search tool options.

a. Input, multiple protein names (release, 12.0)

Version: 12.0					LOGIN	REGISTER	SURVEY
🕸 STRING				Search	Download	Help	My Data
Protein by name Multiple proteins	>	SEARCH					
Proteins by sequences	>		Protein Families I	by Multiple N	lames		
Proteins with Values/Ranks Protein families ("COGs")	>		List Of Names:	(examp	les: <u>#1</u> <u>#2</u> <u>#3</u>)		
 by family name by protein name by protein sequence multiple protein names 			or, upload a file:	E	Zarowse		
multiple sequences Pathway / Process / Disease ^{New}	>		Organisms:		Ψ.		
Add organism ^{New}	>		SEA	ARCH			
Organisms	>						
Examples	>						
Random entry	>						

b. STRING network output for the physically-interacting proteins DUF34 family protein P9WFM1 (COG0327) and P9WLH3 (COG1579) of *Mycobacterium tuberculosis* (strain ATCC 25618 / H37Rv) via "multiple family names" type query; high-confidence (0.700) minimum required interaction score; no more than 50 interactors, max number of interactors to show for first shell; no more than 5 interactors, max number of interactors to show for second shell.



c. STRING output for the co-occurrence view for COG0327 (DUF34 family homolog, P9WFM1) and COG2072 (P9WLH3) (phylogeny collapses/expansions that are shown were generated by default).

① opisthokonta (1129 taxa)						0	BE	N	E	CC	0	C	Cl	JR	R	Eľ	1	CE														
Top protocolacteria (399 taxa)	search within tree Go	COG1579	COG0327	C0G3323	NOG400094	C0G2202	C0G1352	C0G2201	COG0784	COG5295	COG0840	C0G4219	C0G4372	COG3920	COG5001	C0G2433	NOG329964	COG1876	C0G2273	COG1305	C0G2200	N0G234552	C0G1762	C0G2384	C0G1262	COG0739	COG0079	COG0406	C0G0488	COG04099	COG5010	
To protochard (2009) [159 taxa)		-																														•
Image: Proceeding (167 tara) X X X X X X X X X X X X X X X X X X X		× × ×	< < < <	× × ×	¥ -	× × ×	× : × : × :	¥ : ¥ : ¥ :	× 1 × 1 × 1		× ×	× × ×	× × ×	× × ×	× × ×	× × × ×	-	× × ×	× 1 × 1 × 1		X X X X	 K K K 	× × ×	KK	KK	× × × ×	× × ×	× : × : × :	× 1 × 1 × 1		 X X	
Acidobacteria (45 taxa)		<u> </u>	¥	<u>-</u>	-	¥	<u> </u>	¥ .	Ý ,	<u> </u>		<u> </u>	¥	<u> </u>	¥	_	_	¥ .	¥ ,	< <u>-</u>	¥	<u> </u>	¥	<u>-</u>	<u>-</u>	¥	¥	<u> </u>	Ý ,		< ¥	
Fusebacteriales (41 taxa) X X X		¥	¥	¥	-	¥	<u> </u>	¥ .	<u>×</u>	-	• -	¥	¥	¥	¥	¥	¥	¥	<u>×</u> ·	< <u> </u>	¥	¥	¥	-	¥	~	¥	<u> </u>	<u> </u>	< <u>×</u>	-	
 		¥	≚	-	Ξ.	~	<u>×</u> :	ř.	~ `	 	_	–	¥	-	≚	Ξ.	Ξ.	¥	ž i	< <u>×</u>	–	–	≚	Ξ.	*	¥	¥	¥ :	<u>×</u> 2	<u> </u>	 ✓ 	
 Eusimicrobia (31 taxa) Y <liy< li=""> <liy< li=""> <liy< li=""></liy<></liy<></liy<>		–	-	¥	-	¥.	¥ :	¥ :	¥ -	< <u>×</u>	_	–	¥.	¥.	-	-	-	-	Ť,	*	-	¥	–	-	-	-	ž,	¥ :	ž i	_ ×	-	
Intermotogae (31 taxa) Y <td></td> <td>Ť</td> <td>Ť</td> <td>-</td> <td>Ξ.</td> <td>1</td> <td>Ť</td> <td>Ť</td> <td></td> <td>Ť</td> <td></td> <td>Ť</td> <td>Ť</td> <td>Ť</td> <td>Ť</td> <td>¥</td> <td>-</td> <td>5</td> <td><u> </u></td> <td></td> <td>Ť</td> <td>-</td> <td>Ť</td> <td>-</td> <td>Ť</td> <td>-</td> <td>Ť</td> <td>Ť</td> <td><u> </u></td> <td>(×</td> <td></td> <td></td>		Ť	Ť	-	Ξ.	1	Ť	Ť		Ť		Ť	Ť	Ť	Ť	¥	-	5	<u> </u>		Ť	-	Ť	-	Ť	-	Ť	Ť	<u> </u>	(×		
Aquificae (19 taxa) Y		ž	ž	5	Ξ.	ž	Ť	Ť	Ť	< <u>×</u>	Ť	~	¥	Ť	ž	Ξ.	Ξ.	¥ .	Ť	<u> </u>	Ť	Ť	Ť	Ξ.	Ť	*	ž	Ť	Ľ.		Ť	eria
Aquificae (19 taxa)		-	-	-	-	-	Ĵ				Ĵ	-	5	-	Ť	5	-	-			Ť	Ť	Ţ		Ť	Ĵ	Ť	Ľ.			Ĵ	act
 Thermodesulfobacteriaceae (10 taxa) 		Ť	Ţ	-	-	Ĵ	Ĵ.		<u> </u>	Ť	Ĵ	-	Ť	Ţ	Ť	Ť	-	_			Ť	-	Ť		Ť	Ť	Ľ	<u> </u>		Ĩ	Ě	8
Deferribacterales (7 taxa)		Ľ	Ċ	5	-	Ż	Ċ				÷	_	Ĵ	-	Ĵ	-	Ξ.	-			Ĵ		Ĵ		Ĵ	Ċ	Ľ	Ĵ.		1	Ľ	
 Nitrospinae Tectomicrobia group (6 taxa) 		Ľ	Ľ	Ĵ	Ξ.	5	ζ.	2	5		Ľ	-Ξ	Ē.	-	Ţ	Ξ.	Ξ.	Ţ.,	Ξ.		÷	Ξ.	Ţ	Ξ.	Ĵ,	Ĵ.	Ĵ.	Ĵ.		Ĵ	Ľ	
Coprothermobacter (2 taxa)		Ŷ	Ç.	<u> </u>	Ξ.	Ľ	Ç.,		5		Ţ	- 2	2	2	7	Ξ.	Ξ.	2	γ,	/ -/	7	- 2	7	Ξ.	J	Ĵ	Ĵ.		, ,	/	Ĵ	
Caldisericum (2 taxa) Y		2	Ţ	Ξ.	Ξ.	2			<u> </u>			- 🗸	1	Ξ.	2	Ξ.	Ξ.	Ξ.	2 3		1	- 2	J	Ξ.	<u> </u>	Ĵ	Ĵ,	2			2	
Chrysiogenaceae (2 taxa) Caldithrix (2 taxa)		~	Ĵ	7	Ξ.	-	<u> </u>	_				Ĵ	1	Ξ.	Ξ.	Ξ.	Ξ.	Ξ.	Ξ,	/		- 2	Ĵ	Ξ.	Ξ.	Ĵ	<u> </u>	<u> </u>				
Caldithrix (2 taxa) -		2	Ĵ	Ĵ	Ξ.	2	Ξ.		2			<u> </u>	2	2	~	2	Ξ.	Ξ.			Ţ	- 2	Ĵ	Ξ.	7	Ĵ	2	<u>,</u>			-	
Dictyoglomus (2 taxa) -		~	4	<u> </u>	2	2	¥ .	ý.,	2.		. ¥	~	_	~	7	_	_	_	× .		_	_	Ŷ	_	7	2	ý.	Ç.,	2.		÷.,	
Image: Design of the start		-	-	_	_	-	2	<u> </u>	ý.,	_		_	~	-	_	_	_	_	-		_	_	_	_	_	Ŷ	-	,				
Yiridiplantae (124 taxa)		_	4	¥	2	_	<u> </u>	~		<u></u>		-	_	_	¥	2	<u> </u>	¥ .	ý.,	<u>_</u>	¥	-	-	2	2	2	2	ý .	<u>_</u>			- 1
E Stramenopiles (24 taxa) Y<		_	_	_	<u> </u>	_	_ 3	¥ .	_			-	-	_	¥	-	<u> </u>	¥	¥ .		_	-	-	2	-	¥	¥	~ .				
Image: Second secon		-	¥	_	<u> -</u>	_	<u> </u>					-	-	-	_	-	_	_		<u> </u>	-	-	-	-	-	≚	¥	~ .				
 Amoebozoa (8 taxa) Trypanosomatidae (6 taxa) Amoebozoa (8 taxa) Trypanosomatidae (6 taxa) Anodophra (4 taxa) Metamonada (3 taxa) Amoebozoa (3 taxa) Amoebozoa (3 taxa) Amoebozoa (8 taxa) Amoeboxoa (8 taxa) Amoebo		-	-	_	<u>-</u>	_	<u> </u>			< _		-	-	-	_	-	_	_		< _	-	-	-	-	-	2	¥	<u> </u>				
Trypanosomatidae (6 taxa) Trypanosomatidae (6 taxa) Anotophyta (6 taxa) Anotophyta (4		-	~	¥	<u>-</u>	<u> </u>	<u> </u>					-	-	-	_	-	_	_		< _	-	-	-	-	-	_	¥	~ .				
Thecamonas trahens Guillardia theta -Emiliania huxleyi Plasmolophora brasicae Naegleria gruberi Image: DPANN group (11 taxa) Y Y Y Y Image: DPANN group (11 taxa) Y Y Y Y Image: DPANN group (11 taxa) Y Y Y Y Y Y Image: DPANN group (11 taxa) Y Y Y		-	~	_	_	-	<u> </u>					-	-	-	_	-	_	-			-	-	-	-	-	÷.	_	~ .				ota
Thecamonas trahens Guillardia theta Emiliania huxleyi Plasmodiophora brassicae Naegleria gruberi E <i>Luryarchaeota (324 taxa)</i> Y Y Y Y PARNM group (118 taxa) Y Y		-	-	-	÷.	-						-	-	-	-	-	-	-		< _	-	-	-	-	-	÷.	~	~ .	-			aryc
Thecamonas trahens Guillardia theta Emiliania huxleyi Plasmodiophora brassicae Naegleria gruberi Image: Display transformed a gruberi		-	-	~	-	-			- 2	<u> </u>		-	-	-	-	-	-	-			-	-	-	-	-	-	¥	<u> </u>				- R
Emiliania huxleyi	-Thecamonas trahens	-	-	-	-	-	÷ .					-	-	-	-	-	-	-			-	-	-	-	-	-	~	-				
Plasmodiophora brassicae Naegleria gruberi Image: University of the second sec	1	-	~	-	-	-						-	-	-	-	-	-	-			-	-	-	-	-	~	~	-				
Naegleria gruberi • Euryarchaeota (324 taxa) • TACK group (108 taxa) • Tack group (118 taxa)	—Emiliania huxleyi	-	~	-	-	-					-	-	-	-	-	-	-	-			-	-	-	-	-	~	~	× .				
Euryarchaeota (324 taxa)	Plasmodiophora brassicae	-	-	-	-	-					-	-	-	-	-	-	-	-			-	-	-	-	-	-	-	•				
• TACK group (108 taxa) • × × × × × × × × × × × × × × ×	└─Naegleria gruberi	-	-	-	-	-					-	-	-	-	-	-	-	-			-	-	-	-	-	-	-	•				
□ DPANN group (11 taxa) × </td <td>+ Euryarchaeota (324 taxa)</td> <td>-</td> <td>¥</td> <td>-</td> <td>-</td> <td>-</td> <td>¥ 3</td> <td>¥ .</td> <td>¥ -</td> <td>-</td> <td>• ¥</td> <td>-</td> <td>-</td> <td>¥</td> <td>-</td> <td>-</td> <td>-</td> <td></td> <td>¥ 2</td> <td>< <u>×</u></td> <td>-</td> <td>-</td> <td>¥</td> <td>-</td> <td>¥</td> <td>¥</td> <td>¥</td> <td>¥ .</td> <td></td> <td></td> <td>-</td> <td></td>	+ Euryarchaeota (324 taxa)	-	¥	-	-	-	¥ 3	¥ .	¥ -	-	• ¥	-	-	¥	-	-	-		¥ 2	< <u>×</u>	-	-	¥	-	¥	¥	¥	¥ .			-	
→ environmental samples (5 taxa)		-	¥	-	-	-	¥ :	¥ .	¥ -	-	• ¥	-	-	-	-	-	-		¥ 2	< <u>×</u>	-	-	-	-	¥	¥	¥	¥ .			-	
		-	¥	-	-	-			- 2	-	• 🔟	-	-	-	-	-	-	-	- 2	< ¥	-	-	-	-	-	¥	¥	¥ ,			-	ea
		-	¥	-	-	-	¥ :	¥ .	¥ -	-	• ¥	-	-	-	-	-	-	-	- 2	< <u>×</u>	-	-	-	-	¥	-	¥	¥ ,			-	cha
→ Asgard group (3 taxa)		-	-	-	-	-			- 2	-	-	-	-	-	-	-	-	-	- 2	< <u>×</u>	-	-	-	-	-	-	-	¥ ,			-	Ar
Hydrothermarchaeota archaeon		-	~	-	-	-			¥ -	-	·	-	-	¥	-	-	-	-	- 2	-	-	-	-	-	-	-	¥				-	

Figure S16. KO identifiers being supplied to KEGG orthology tool (multiple families, limited genome customization).

a. Output for TsaE (K06925, ubiquitous protein family) and COG1579 (K07164) in KEGG Orthology K number search.

Grp	Genus	Organism	K06925 (tsaE)[6863]	K07164 (K07164)[2322]
B.GamE	Escherichia	eco	b4168	
B.GamE	Escherichia	ecj	JW4126	
B.GamE	Escherichia	ecd	ECDH10B_4363	
B.GamE	Escherichia	ebw	BWG_3880	
B.GamE	Escherichia	ecok	ECMDS42_3610	
B.GamE	Escherichia	ece	Z5775	
B.GamE	Escherichia	ecs	ECs_5144	
B.GamE	Escherichia	ecf	ECH74115_5684	
B.GamE	Escherichia	etw	ECSP_5268	
B.GamE	Escherichia	elx	CDCO157_4830	
B.GamE	Escherichia	eoi	ECO111_5046	
B.GamE	Escherichia	eoj	ECO26_5334	
B.GamE	Escherichia	eoh	ECO103_4961	
B.Gam	Escherichia	ecoo	ECRM13514_5432	
B.GamE	Escherichia	ecoh	ECRM13516_5179	
B.GamE	Escherichia	esl	O3K_22805	
B.GamE	Escherichia	eso	030_02580	
B.Gam	Escherichia	esm	O3M_22710	
B.GamE	Escherichia	eck	 EC55989_4723	
B.GamE	Escherichia	ecg	E2348C_4491	
B.GamE	Escherichia	eok	G2583 4995	
B.GamE	Escherichia	elr	EC055CA74_23945	
B.GamE	Escherichia	elh	ETEC_4514	
B.GamE	Escherichia	ecw	EcE24377A 4725	
B.GamE	Escherichia	eun	UMNK88_5106	
B.GamE	Escherichia	еср	ECP_4413	
B.GamE	Escherichia	ena	ECNA114_4384	
B.GamE	Escherichia	ecos	EC958_4655	
B.GamE	Escherichia	ecv	APECO1 2223	
B.GamE	Escherichia	ecoa	APECO78_01770	
B.GamE	Escherichia	ecx	EcHS_A4410	
B.GamE	Escherichia	ecm	EcSMS35_4639	
B.GamE	Escherichia	ecy	ECSE_4465	
B.GamE	Escherichia	ecr	ECIAI1_4401	
B.GamE	Escherichia	ecq	ECED1 4953	
	Escherichia	eum	ECUMN 4701	
B.GamE	Escherichia	ect	ECIAI39_4632	
B.GamE	Escherichia	eoc	CE10 4907	
B.GamE	Escherichia	ebr	ECB 04035	
B.GamE	Escherichia	ebl	ECD 04035	
B.GamE	Escherichia	ebe	B21 03997	
	Escherichia	ebd	ECBD_3866	
	Escherichia	eci	UTI89_C4768	
	Escherichia	eih	ECOK1_4682	
	Escherichia	ecz	ECS88_4754	
	Escherichia	ecc	c5252	
	Escherichia	elo	EC042_4641	
	Escherichia	eln	NRG857 21185	
	Escherichia	ese	ECSF 4054	
	Escherichia	ecl	EcolC 3845	
	Escherichia	eko	EKO11 4144	
	Escherichia	ekf	KO11 22465	
	Escherichia	eab	ECABU_c47260	
	Escherichia	edb	EcDH1 3825	
	Escherichia	edj	ECDH1ME8569 402	
	Escherichia	elu	UM146 21075	
e. Guille	Loononia	Ciu	0.1110_21070	

b. Input for KEGG Orthology K number search (ortholog table).

KEGG	Databases	Tools	Auto annotation	Kanehisa Lab
K CGG	-		OGY) Database al systems by functional	orthologs
KEGG2 PATHWA	Y BRITE MOD	ULE KO A	nnotation Taxonomy	Synteny Mapper
Search KO	→ for			Go
Enter K numb K06925, K07 Filter Orth	164	161 K00162 K0016 athway Map I	53 K00627 K00382 brite Map module Get	title Get entry Clear

Figure S17. Example CAGECAT output using DUF34 and COG1579 homolog sequences: MSMEG_4307 (NCBI-ProteinID: ABK73705); MSMEG_4306 (NCBI-ProteinID: ABK70599). This figure was generated after using cblaster then clinker applications as part of a tandem suite of tools, namely the "compared to query" visualization.

Query Cluster NA:1-2375	V	ABK73705.1	Instructions:
Mycolicibacterium smegmatis atR40 (Cluster 88, 3.12 score) NZ_SIUA01000017.1:36687-38556			 Drag cluster names to re-order clusters Hover over a locus to reveal box and handles; drag box to move the locus, drag handles to resize it, double click to flip
Mycolicibacterium smegmatis MC2 51 (Cluster 45, 3.12 score) NZ_JAJD01000113.1:90791-92660	<		 Click genes to anchor the figure around them Right click genes to edit their labels Click cluster names and legend entries to rename them
Mycolicibacterium smegmatis atR37 (Cluster 86, 3.12 score) NZ_SITZ01000004.1:94107-95976			Click legend circles to change gene colours Right click legend entries to remove them and hide links Click scale bar text to change its length
Mycolicibacterium smegmatis atR33 (Cluster 84, 3.12 score) NZ_SITY01000003.1:94107-95976			fixed with the state of th
Mycolicibacterium smegmatis atR19 (Cluster 82, 3.12 score) NZ_SITX01000004.1:94107-95976			generation of gene cluster comparison figures Cameron I.M. Gilchrist, Yit-Heng Chool Bioinformatics. doi: https://doi.org/10.1093/bioinformatics/btabu Load data: Choose File No file chosen
Mycolicibacterium smegmatis atR17 (Cluster 80, 3.12 score) NZ_SITW01000004.1:94107-95976			Save Data Save SVG Plot arrangement

Figure S18. Example output for FunCoup using "YbgI" of *E. coli*. While other interaction/correlational data is present, operons are included in each subset.

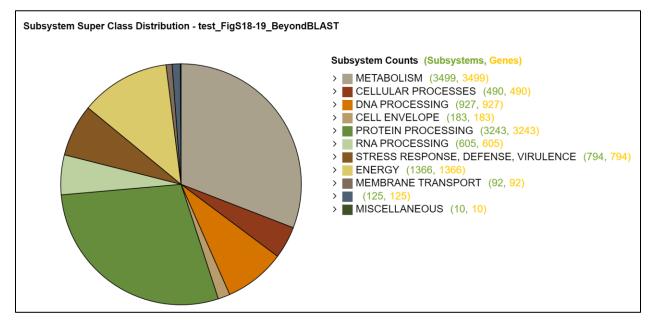
Query interaction	ons					
Interaction partners	Confidence	Network	Evidence types SCL NIC MMR MMR MMR MMR MMR MMR MMR MMR MMR MM	HSA MMU MMU CFA GGA GGA BTA SCE SCE SCE SCE ATH ATH ATH FA PFA	DDI SSO SSC MJA OSA In Gold st.	
vbgi (1 of 2)						
▼ <u>rplA (1 of 2)</u>	1.000	Complex			000000	?
	0.021	Operons				
	0.016	Metabolic				
	0.005	PPI				
	0.000	Signaling				
rpIA (2 of 2)	1.000	Complex				
rpIM (1 of 2)	1.000	Complex				
rpIM.(2 of 2)	1.000	Complex				
dnaJ (1 of 2)	1.000	Complex				
dnaJ (2 of 2)	1.000	Complex				
rpsM (1 of 2)	1.000	Complex				

Figure S19. BV-BRC protein family sorter, general output example (a) with additional viewer examples: b) subsystems; c) pathways – KEGG Map; d) pathways – Heatmap; e) families.

a. Initial Output Files (all)

12	>BV-BRC ^{3.30.5}	All Data Types Find a gene, genome, microarray, etc									
	ecy / home / Protein Con		_FigS18-19	_BeyondBLAST (7 items	5)				Pathways	Families	Subsystems
Job	ID	10638982									
Sta	rt time	4/28/23, 4:13 PM	28/23, 4:13 PM								
End	l time	4/28/23, 4:15 PM									
Rur	1 time	2m37s									-
	Name				*	Size	Owner	Members	Cre	ated	0
t	Parent folder							-			
	report.txt					192 B	me	Only me	4/2	8/23, 4:15 P	М
Ê	test_FigS18-19_BeyondE		4.7 MB	me	Only me	4/2	8/23, 4:15 P	Μ			
Ê	test_FigS18-19_BeyondE	BLAST_pathway	/s_tables.jsc		7.1 MB	me	Only me	4/2	8/23, 4:15 P	М	
	test_FigS18-19_BeyondE	BLAST_proteinfa	ams_tables.		1.6 MB	me	Only me	4/2	8/23, 4:15 P	Μ	
	test_FigS18-19_BeyondE	BLAST_subsyste		6.2 MB	me	Only me	4/2	8/23, 4:15 P	Μ		
	test_FigS18-19_BeyondE	BLAST_subsyste		7.3 MB	me	Only me	4/2	8/23, 4:15 P	М		
Ê	test_FigS18-19_BeyondE	BLAST_subsyste	ems_variant	t_mtx.tsv		68.1 kB	me	Only me	4/2	8/23, 4:15 P	Μ

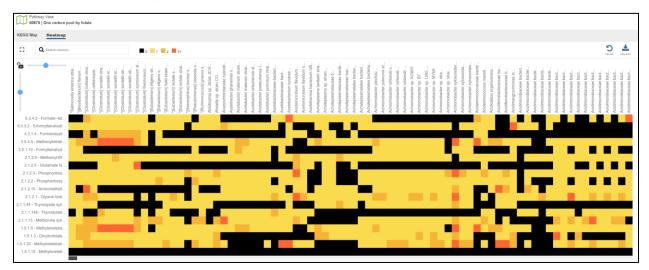
b. Subsystems (overview). Subsystems (via the "Subsystems" tab adjacent to the "Subsystems Overview" tab) can then be selected individually or together (>1) for either mapping for the former case or grouping for the latter.



c. Pathways – KEGG Map

KEGG Map	Heatm	пар		
EC 🝷 Number	Genome Count	Feature Count	Genome Count Not Present	Occurrer 🖸
6.3.4.3	7631	9292	11437	1
6.3.3.2	17313	18447	1755	1
.3.1.4	1698	1885	17370	1
.5.4.9	18628	25262	440	1
.5.1.10	11427	14100	7641	1
.1.2.9	18561	20223	507	1
.1.2.5	1095	1237	17973	2
.1.2.3	18382	21280	686	1
2.1.2.2	17991	19076	1077	2
2.1.2.10	14419	16499	4649	2
.1.2.1	18489	26998	579	1
2.1.1.45	16352	18417	2716	1
2.1.1.148	4392	4624	14676	1
2.1.1.13	13426	18567	5642	1
1.5.1.5	18608	25238	460	1
1.5.1.3	17086	27209	1982	3
1.5.1.20	15417	19322	3651	1
1.5.1.15	1	1	19067	1

d. Pathways – Heatmap

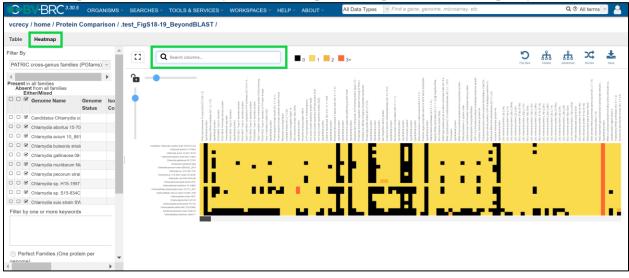


e. Families (output)

	3-BV- E	BRC ^{3.30.5} organisms	S 👻 SEARCHES 👻 TOOLS &	SERVICES	S ▼		All Data Type	s 🔻 Find a	gene, genor
[ι	iser] / hor	ne / Protein Comparison	/ .test_FigS18-19_Beyond	BLAST /					
Pa	athways	EC Number							
									T
00	WNLOAD KEY	WORDS	ADV Search						FILTERS
	Pathway ID	Pathway Class	Pathway Name	Genome Count	EC Number Count	Gene Count	Genome EC Count	EC Conservatio (%)	Gene Conservatio (%)
	00240	Nucleotide Metabolism	Pyrimidine metabolism	20	29	487	321	55.34	22.46
	00051	Carbohydrate Metabolism	Fructose and mannose metabolism	20	19	119	91	23.95	8.08
]	00670	Metabolism of Cofactors and Vitamir	One carbon pool by folate	20	14	121	135	48.21	16.39
]	00770	Metabolism of Cofactors and Vitamir	Pantothenate and CoA biosynthesis	20	15	58	55	18.33	11.76
]	00020	Carbohydrate Metabolism	Citrate cycle (TCA cycle)	20	14	289	204	72.86	19.02
00230		Nucleotide Metabolism	Purine metabolism	20	41	455	287	35	16.45
]	00982	Xenobiotics Biodegradation and Met	Drug metabolism - cytochrome P450	2	1	6	2	10	5
]	00540	Glycan Biosynthesis and Metabolism	Lipopolysaccharide biosynthesis	20	16	172	188	58.75	22.92
]	00040	Carbohydrate Metabolism	Pentose and glucuronate interconver	20	21	51	41	9.76	7.07
]	00562	Carbohydrate Metabolism	Inositol phosphate metabolism	20 4		35 30		37.5	10
]	00620	Carbohydrate Metabolism	Pyruvate metabolism	20	28	232	164	29.29	13.09

Figure S20. BV-BRC protein family sorter output, Families.

a. Heatmap view of Families output. Green boxes highlight the text entry box and the heatmap tab selection. Text box provides search ability that is dependent upon annotation quality in database.



b. Table view of Families output. Purple boxes highlight the multiple-select checkboxes that a user may select for viewing in the accompanying heatmap or for export/grouping.

01	3V-BRC 3.30.8	ORGAN	ISMS -	SE	ARCHES -	TOOLS & SERVI	CES 🔻	WORKSPACES	S * HELP * ABOUT * All Data	Types - Fin	d a gene, genome,	microarray, e	lc		Q 🕲 All terms 👻
/crecy	y / home / Protein	Compari	son / .	test_	FigS18-19	_BeyondBLAS	т/								
ſable	Heatmap														
lter By			-											Ø	Nothing selected.
PATRI	C cross-genus families	. ,	-		D	Pro	teins	Genomes	Description	Min AA Length	Max AA Length	Mean	Std O Dev	HIDE	Select one or more items on the left to see their
Present	in all families				PGF_03004613		20	20	Histidyl-tRNA synthetase (EC 6.1.1.21)	389	485	427	18	GUIDE	details and possible actions.
	ither/Mixed				PGF_00982259		20	20	Aspartyl-tRNA synthetase (EC 6.1.1.12) @ Aspa	582	599	585	3		
104	8 Genome Name	Genome Status	lso Co		PGF_02454577		18	18	SSU ribosomal protein S20p	94	99	98	1		
000	Candidatus Chlamvdia		00		PGF_00003741		13	13	FKBP-type peptidyl-prolyl cis-trans isomerase FI	186	255	249	18		
	Chlamydia abortus 15-7				PGF_00413203		19	19	tRNA (cytidine(34)-2'-O)-methyltransferase (EC	151	161	155	3		
			- 1		PGF_10345259		20	19	Thioredoxin	70	170	105	17		
	Chlamydia avium 10_8				PGF_00740092		19	19	hypothetical protein	238	246	239	2		
			- 1		PGF_03753407		19	19	tRNA threonylcarbamoyladenosine biosynthesis	111	154	140	13		
					PGF_00421600		15	15	DNA polymerase III, epsilon chain	249	249	249	0		
	onanyaa maraara		- 1		PGF_00420739		18	18	Cytosolic acyl-CoA thioester hydrolase family pri-	120	161	153	12		
	Chlamydia pecorum str				PGF 00061920		18	18	Apolipoprotein N-acyltransferase	521	542	537	7		
	Chlamydia sp. H15-195		- 1		PGF_00063944		19	19	UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosan	280	296	282	4		
	Chlamydia sp. S15-834		- 1		PGF 00014043		13	13	Inclusion membrane protein-54	393	406	401	3		
	Chlamydia suis strain S				- PGF_07504595		19	19	3-hydroxyacyl-[acyl-carrier-protein] dehydratase	102	154	150	12		
Filter by	y one or more keyword	s			PGF 00033729	1	20	20	Acyl-[acyl-carrier-protein]UDP-N-acetylglucosa	257	280	278	5		
				Ū.	- PGF_05165078		17	17	Methionyl-tRNA formyltransferase (EC 2.1.2.9)	306	321	318	4		
					PGF 01214589		12	12	Inclusion membrane protein-44	293	325	316	11		
					PGF 01211937		1	1	hypothetical protein	87	87	87	0		
Per	fect Families (One pro	tein per			PGF 00016431		19	19	LSU ribosomal protein L3p (L3e)	169	235	218	12		
renome			• •		200 of 3965								3 20 >		

Figure S21. EggNOG Phylogenetic Profile tool example output using the site-provided example input data. At the time of writing, the tool seemed to be experiencing loading and/or query processing errors, failing to generate the intended product.

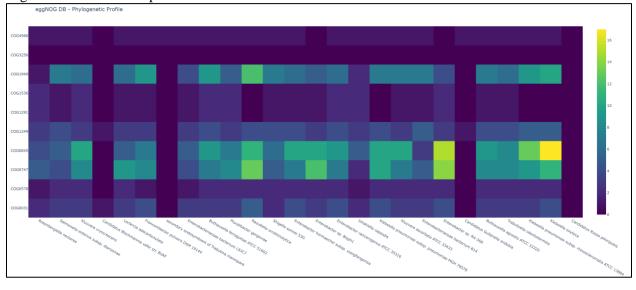


Figure S22. KBase workflow for generating a "gene tree" (as regarded by KBase) from several select genomes. Figure was directly adapted from a KBase 2020 phylogenomics online workshop diagram. Circles indicate data for input/outputs. Squares indicate applications within the KBase "*narrative*" (i.e., "narratives" in KBase are curated workflows of applications, as well as respective inputs and outputs [data objects], that can be created by users or, alternatively, made available to users by developers for implementation as example templates for common workflows).

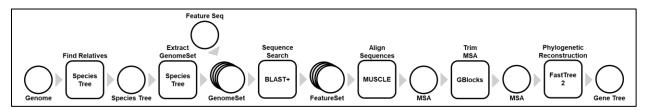


Figure S23. Genomic Context Visualizer (GeCoViz)

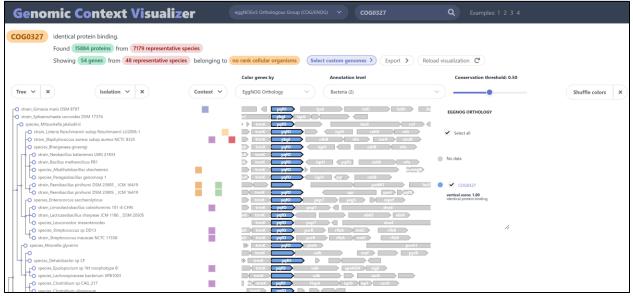


Figure S24. GizmoGene example output for the DUF34 homolog of *Mycobacterium tuberculosis* H37Rv_CG (BV-BRC feature ID: fig|1773.25616.peg.2443) using a custom set of BV-BRC genomes (62 Representative genomes, bacteria).



Supplementary References

- 1. Gat, O.; Lapidot, A.; Alchanati, I.; Regueros, C.; Shoham, Y. Cloning and DNA sequence of the gene coding for Bacillus stearothermophilus T-6 xylanase. *Appl. Environ. Microbiol.* **1994**, *60*, 1889–1896, doi:10.1128/AEM.60.6.1889-1896.1994.
- 2. Shulami, S.; Gat, O.; Sonenshein, A.L.; Shoham, Y. The Glucuronic Acid Utilization Gene Cluster from Bacillus stearothermophilus T-6. *J. Bacteriol.* **1999**, *181*, 3695–3704, doi:10.1128/JB.181.12.3695-3704.1999.
- Shulami, S.; Zaide, G.; Zolotnitsky, G.; Langut, Y.; Feld, G.; Sonenshein, A.L.; Shoham, Y. A Two-Component System Regulates the Expression of an ABC Transporter for Xylo-Oligosaccharides in *Geobacillus stearothermophilus*. *Appl. Environ. Microbiol.* 2007, 73, 874–884, doi:10.1128/AEM.02367-06.
- 4. Alalouf, O.; Balazs, Y.; Volkinshtein, M.; Grimpel, Y.; Shoham, G.; Shoham, Y. A New Family of Carbohydrate Esterases Is Represented by a GDSL Hydrolase/Acetylxylan Esterase from *Geobacillus stearothermophilus. J. Biol. Chem.* **2011**, 286, 41993–42001, doi:10.1074/jbc.M111.301051.
- 5. Shulami, S.; Shenker, O.; Langut, Y.; Lavid, N.; Gat, O.; Zaide, G.; Zehavi, A.; Sonenshein, A.L.; Shoham, Y. Multiple Regulatory Mechanisms Control the Expression of the Geobacillus

stearothermophilus Gene for Extracellular Xylanase. J. Biol. Chem. 2014, 289, 25957–25975, doi:10.1074/jbc.M114.592873.

- 6. Daas, M.J.A.; van de Weijer, A.H.P.; de Vos, W.M.; van der Oost, J.; van Kranenburg, R. Isolation of a genetically accessible thermophilic xylan degrading bacterium from compost. *Biotechnol. Biofuels* **2016**, *9*, 210, doi:10.1186/s13068-016-0618-7.
- 7. Reed, C.J.; Hutinet, G.; de Crécy-Lagard, V. Comparative Genomic Analysis of the DUF34 Protein Family Suggests Role as a Metal Ion Chaperone or Insertase. *Biomolecules* **2021**, *11*, 1282, doi:10.3390/biom11091282.
- 8. Kanesaki, Y.; Ogura, M. RNA-seq analysis identified glucose-responsive genes and YqfO as a global regulator in Bacillus subtilis. *BMC Res. Notes* **2021**, *14*, 450, doi:10.1186/s13104-021-05869-1.