Protein Family Analysis: Protein Family Sorter

1. To look for similarities and differences at the protein level across different genomes in PATRIC, you can use the Protein Family Sorter. Go to the Tools tab across the top of any PATRIC page and click on it. This will open up a list of tools (A below). Click on the Protein Family Sorter. This will take you to the landing page for that tool. If you are logged in, you will see the groups you have created in the box under "1. Select organism(s)."



2. The groups that you have created should be visible. Your can select one, or many to compare. In this example, I am selecting a group of 42 genomes that I created, and a group with a single, private genome that I annotated in PATRIC (Blue arrow 1 below). I have a lot of groups, so you can only see the one checked below. Once you have made your selections, click on the Search button (Blue arrow 2).



3. **Pan Genome.** This will take you to a page that contains a filter for genomes on the left, and a table of protein families on the right. When this page loads, you are seeing the pan genome for the group you have selected. It contains all the protein families across all the genomes.

Filter By			Tab	e Heatman								
45 genom	ee (Channe Genome !	Celection)	470	7 families found								
Present in	n all families	(albaabdu)	470	7 Tannines Tounu								
Abse	ent from all families			Workspace	View		Download	Tools Columns				Help
	Eicher/Plixed	Comme	100	Add Feature(s)	8 FASTA DNA		Table •	Rethway Summary III Multiple Seq Alignment Show/Hide Default				User Guides
0 0	Condition frame			·	PASTA Prote	in 🔊	FASTA •	MAP IDs to •			•	
0 0	 Brucella abortus 	N WGS		ID		Proteins	Genomes *	Product Description	Min AA Length	Max AA length	Mean	Std
0 0	 Brucella abortus 	519 Compreta	11	FIG00136922		45	45	Succinate dehvdrogenase flavoprotein subunit (EC 1.3.99.1)	613	614	613	0.15
0 0	Brucella abortus	b Complete		FIG00515945		47	45	Methyltransferase (EC 2.1.1)	79	337	287	64.68
0 0	 Brucella abortus 	o WGS		FIG00004244		52	45	PropionvI-CoA carboxylase biotin-containing subunit (EC 6.4.1.3)	45	667	577	181.04
0 0	Brucella abortus	b WGS		FIG00138938		47	45	tRNA(Cytosine32)-2-thiocytidine synthetase	119	322	281	42.50
0 0	Brucella abortus	b WGS		FIG00450724		45	45	FIG00450725: hypothetical protein	135	143	142	1.22
0 0	Brucella abortus	b WGS		FIG00138924		55	45	Amino acid regulated cytosolic protein	94	486	391	133.23
0 0	Brucella abortus	b WGS		FIG00450712		46	45	competence lipoprotein ComL, putative	104	287	280	29.75
0 0	Brucella abortus	b WGS		FIG00007816		96	45	Thiamin ABC transporter, transmembrane component	40	548	355	151.19
0 0	 Brucella abortus 	st WGS		FIG01304195		45	45	ABC transporter involved in cytochrome c biogenesis, ATPase component ComA	205	218	214	2.15
0 0	Brucella canis AT	C Complete		FIG01344846		46	45	Purine/pyrimidine phosphoribosyl transferase (EC 2.4.2)	62	352	277	35.39
0 0	Brucella ceti B1/9	H WGS		FIG00450701		45	45	EIG00450702: hypothetical protein	57	58	57	0.21
0 0	Brucella ceti M13	/ WGS		FIG00004220		46	45	Dipeptidyl carboxypeptidase Dcp (EC 3.4.15.5)	207	682	665	77.69
0 0	Brucella ceti M45	0 WGS		FIG01304998		85	45	RND efflux system, membrane fusion protein CmeA	140	459	377	58.06
~ ~	- Decents and MCA	4 1000		FIG00024898		46	45	Erythritol transcriptional regulator EryO	46	401	310	42.85
Filter by o	one or more keyword			FIG00905558		91	45	Oxidoreductase (EC 1.1.1)	154	378	352	34.56
		_		FIG00109595		45	45	Non-specific DNA-binding protein Dps / Iron-binding ferritin-like antioxidant protein / Ferroxidase (EC 1.16.3.1)	160	165	161	2.17
				FIG01266244		49	45	Ribosomal protein S12p Asp88 (E. coli) methylthiotransferase	87	438	400	90.23
				FIG00138192		92	45	glutamine synthetase family protein	210	476	448	53.55
			1	FIG00450767		45	45	FIG00450768: hypothetical protein	39	268	54	32.82
Perfec	ct Families (One prote	in per	10	FIG00012403		45	45	Thiamin ABC transporter, substrate-binding component	329	335	333	1.73
genome)			14	Page 1 of	236 🕨 🕅			Show 20 per page Apply Apply to ALL tables			Displaying familie	== 1 - 20 of 4707
 Non P All Far Number c 	Perfect Families milies of Proteins per Family											
Number o	of Genomes per Famil	y										
	to	_										
		Filter										

4. **Core Genome.** To find the core genome (the protein families that all genomes have at least one member in), you will need to click on the circle under the column called "Present in all families" in the row called "Genome Name" (Blue arrow 1 below). This will auto select "Present in all families" for all the genomes in your selection. It will also resort the table to show the protein families that meet this condition (highlighted by the red box).



5. Accessory Genome-One genome at a time. You can also get the accessory genome, but only one genome at a time (you would have to combine the results in excel later to get the final number). To do this, first check the circle under the column called "Absent in all families" in the row called "Genome Name" (Blue arrow 1 below). This will auto select "Absent in all families" for all the genomes in your selection. It will also resort the table to show the protein families that meet this condition (highlighted by the red box below, where you can see that the number of families found is 0). Then you will need select to the circle under the column called "Present in all families" for a single genome (Blue arrow 2 below). This will show the protein families that are unique to just that genome. It will also resort the table to show the protein families that are unique to just that genome. It will also resort the table to show the protein families that are unique to just that genome. It will also resort the table to show the protein families that are unique to just that genome. It will also resort the table to show the protein families that are unique to just that genome. It will also resort the table to show the protein families that meet this condition (highlighted by the red box in the lowest panel).



6. Searching for specific names. You can use the text box near the bottom of the filter to find protein families that are specifically named. This will filter on the name of the family, not necessarily on the product description of the individual genes, so the results from the Feature Table and the Protein Family table will not necessarily match. To find specific genes, you can enter a name (NOT a locus tag) in the filter box on the left (Blue arrow 1) and then click the filter button (Blue arrow 2). This will filter the results to show the protein families that match the search.

	Table Heatmap				
Filter by one or more keywords	18 families found				
virB	Workspace	View	Download	Tools	Columns
	Add Feature(s	FASTA DNA	FASTA •	MAP IDs to •	Show/Hide • Default
	ID ID	Proteins	Genomes *	Product Description	
Perfect Families (One protein per	FIG00002012	46	45	Inner membrane protein forms channel for type IV se	ecretion of T-DNA complex (VirB3)
genome)	FIG00001517	46	45	Forms the bulk of type IV secretion complex that spa	ns outer membrane and periplasm (VirB9)
Non Perfect Families	FIG00038155	46	45	Major pilus subunit of type IV secretion complex (Vir	B2)
Il Families	FIG00002640	43	43	Bores hole in peptidoglycan layer allowing type IV se	cretion complex assembly to occur (VirB1)
Number of Proteins per Family	FIG00002241	51	37	Integral inner membrane protein of type IV secretion	complex (VirB6)
to	FIG00001460	33	32	ATPase provides energy for both assembly of type IV	secretion complex and secretion of T-DNA complex
	FIG00851149	30	30	Inner membrane protein of type IV secretion of T-DN	IA complex, VirB6
Number of Genomes per Family 2	FIG01304041	28	28	Inner membrane protein forms channel for type IV se	ecretion of T-DNA complex, VirB8
to 🖌 🚽	FIG00048546	27	27	Minor pilin of type IV secretion complex (VirB5)	
Filter	FIG01304715	22	22	ATPase required for both assembly of type IV secreti	on complex and secretion of T-DNA complex, VirB1
r inter	FIG00055840	22	21	ATPase provides energy for both assembly of type IV	secretion complex and secretion of T-DNA complex
	FIG00001567	20	19	Inner membrane protein forms channel for type IV se	ecretion of T-DNA complex (VirB10)
	FIG00001707	15	14	Inner membrane protein forms channel for type IV se	ecretion of T-DNA complex (VirB8)
	FIG01303993	11	11	ATPase required for both assembly of type IV secreti	on complex and secretion of T-DNA complex, VirB4
	FIG00978077	2	2	Type IV secretory pathway, VirB4 components	
	FIG00001036	2	2	Type II/IV secretion system ATP hydrolase TadA/VirE	311/CpaF, TadA subfamily
	FIG01960941	1	1	Inner membrane protein forms channel for type IV se	ecretion of T-DNA complex (VirB8)
	FIG01304692	1	1	Minor pilin of type IV secretion complex (VirB5)	

7. **Heatmap view of protein families.** You can see the presence or absence of protein families across particular genomes by looking at the Heatmap view. To see this, click on the Heatmap tab (Blue arrow 1 below) and this will open the heatmap view page. Protein families are listed on the x-axis across the top of this view, and genomes along the y-axis.



8. To see the names of the individual protein families and/or genomes, you will have to move the sliders to expand the view. In the box at the of the heat map, move the x-axis slider to the right as indicated by the arrow, and the y-axis slider down. This will open the heatmap so that you can read both the names of the genomes and the protein families. Cells that are yellow indicate that there is a single protein that belongs to that protein family annotated in that genome. Darker yellow cells indicate that the genome has two proteins that are part of that family. Orange cells indicate 3 or more proteins, and when there is a black cell it means that the genome does not have a protein annotated in that family.



9. **Legend.** You can click on the legend (Red asterisk) to see what the different colors of the cells mean, but basically, black cells indicate a gene absence and all other colors indicate gene presence.



10. **Clustering.** You can group the protein families and genomes by similarity using the Cluster function. Click on the word cluster (Blue arrow 1 below) and this will resort to heatmap to show similar patterns. The default settings for the clustering button are based on the Pearson Correlation Algorithm and a Pairwise Average-linkage Clustering Type. In the screenshot of the heatmap below you can see similar patterns in two closely related members of the *Brucella ceti* clade that none of the other genomes share.



11. **Anchoring.** PATRIC also provides a function where you can examine the presence and absence of protein families across all the genomes, but as they appear in the order of a single genome. This is called "anchoring" and it is a good way to look for genomic islands. To anchor the protein families click on the down arrow that is part of the text box following "Advanced Clustering" (Blue arrow 1). This will open a dropdown box that shows all the genomes in your group. Scroll down until you find the genome that you want to use as your anchor and click on that name (Blue arrow 2). In this example, I chose *Brucella inopinata* BO1.

Table	Heatma	p	
4707 fa	milies fou	und	1
Flip Axis	Cluster	Advanced Clustering	
			Brucella ceti M490/95/1
			Brucella ceti M644/93/1
			Brucella ceti str. Cudo
			Brucella inopinata BO1
			Brucella melitensis ATCC 23457
			Brucella melitensis biovar Abortus 2308
			Brucella melitensis bv. 1 str. 16M
			Brucella melitensis bv. 1 str. 16M [WGS]
			Brucella melitensis bv. 1 str. Rev.1
			Brucella melitensis bv. 2 str. 63/9
			Brucella melitensis bv. 3 str. Ether
			Prusalla malitansis E2 00 E48 contias

12. This will resort the protein families to occur in the order that the genes occur on the BO1 genome. You can use the scroll bar at the bottom of the heatmap to follow

along the genome (from A to B below) until you see an area of interest. In this example, I chose an area that few genomes seem to have, as is indicated by the black cells in panel B.



13. You can mouse over individual cells (Red arrow 1 below) to see the names of both the genome and the protein family in the blue header box above the heatmap (Red arrow 2).



14. Downloading data from Heatmap. To see the unique protein families, you can use your mouse to draw a box around the area of interest in the heat map (Red arrow 1 in Panel A). A pop-up window will appear that allows you to download the heatmap data, download the proteins from your selection, show the proteins from your selection, add the proteins to a group, or cancel (Panel B). To see the proteins, click on "Show Proteins" (Red Arrow 2 below).

A			В						
	Selected Area from Heatmap								
	Genomes selected: 1 Family selected: 35 Members: 7								
	Download Heatmap Data •	Download Proteins •	2 Show Proteins	Sort Protein Families	Add Proteins To Group	Cancel			

15. A new window will load that shows you the data behind your selection, including the name of the genome, accession number, PATRIC and RefSeq locus tags, the size of the protein, and the product description.

35 features found in 35 protein families							
Workspace View	Download	То	pols	Columns			
Add Feature(s) Add Feature(s)	Table ▼ FASTA ▼	Arthway Summary Arthway Summary	Multiple Seq Alignment	Show/Hide • Defa	ult		
Genome Name	Accession PA	TRIC ID	RefSeq Locus Tag 🔺	Length (AA)	Product Description		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.679	BIBO1 0646	763	Nitrogen regulation protein NtrY (EC 2.7.3)		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.1089	BIBO1 1024	92	hypothetical protein		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.1114	BIBO1 1049	39	FIG00451771: hypothetical protein		
Brucella inopinata BO1	NZ_ADEZ010000 fig	470735.4.peg.1165	BIBO1 1094	327	Signal peptide peptidase A (SppA), a serine protease		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.1345	BIBO1 1261	79	FIG00451842: hypothetical protein		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.1376	BIBO1 1293	43	hypothetical protein		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.1672	BIBO1 1576	48	FIG00451723: hypothetical protein		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.1862	BIBO1 1761	65	FIG00451561: hypothetical protein		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.2131	BIBO1 2018	252	hypothetical protein		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.2243	BIBO1 2123	415	hypothetical protein		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.2245	BIBO1 2125	342	hypothetical protein		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.2328	BIBO1 2201	40	FIG00451963: hypothetical protein		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.2689	BIBO1 2545	37	FIG00451720: hypothetical protein		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.2999	BIBO1 2765	418	FIG00452104: hypothetical protein		
Brucella inopinata BO1	NZ_ADEZ010000 fig	470735.4.peg.3002	BIBO1 2767	70	FIG00451836: hypothetical protein		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.3140	BIBO1 2877	561	COG0028: Thiamine pyrophosphate-requiring enzymes		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.1191		57	FIG00451563: hypothetical protein		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.2051		44	FIG00452160: hypothetical protein		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.2130		70	FIG00452065: hypothetical protein		
Brucella inopinata BO1	NZ ADEZ010000 fig	470735.4.peg.2324		52	FIG00451856: hypothetical protein		

16. **Pathway Summary.** You can do a number of things with these genes. You could save them to a group, get their nucleotide or amino acid sequences, download the information, see if they were important in any metabolic pathways, or generate a multiple sequence alignment for them. To generate an alignment, click on the box

in front of "Genome Name" (Arrow 1 below). This will select all the genes in the table. Then you need to click on the Pathway Summary button at the top of the table (Arrow 2).

Workspace Add Feature(s)	View FASTA DNA FASTA Protein	Download Image: Table Image: FASTA	Pathway Summary MAP IDs to	2 Multiple Seq Alignment	
Genome Name		Accession	PATRIC ID	RefSeq Locus Tag 🔺	
Brucella inopinata BO	<u>01</u>	NZ ADEZ010000	fig 470735.4.peg.75	BIBO1 0072	
Brucella inopinata BO	<u>01</u>	NZ ADEZ010000	fig 470735.4.peg.113	BIBO1 0109	
Brucella inopinata BC	<u>01</u>	NZ ADEZ010000	fig 470735.4.peg.132	BIBO1 0125	
Brucella inopinata BC	<u>01</u>	NZ ADEZ010000	fig 470735.4.peg.184	BIBO1 0175	
Brucella inopinata BO	<u>01</u>	NZ ADEZ010000	fig 470735.4.peg.258	BIBO1 0249	
Brucella inopinata BO	01	NZ ADEZ010000	fig 470735.4.peg.287	BIBO1 0274	
Brucella inopinata BO	01	NZ ADEZ010000	fig 470735.4.peg.338	BIBO1 0323	

17. This will open a table that shows all the pathways that the genes you selected are present in (Panel A). To find the pathway that has the most of the selected genes present in it, double-click on the "# of Genes Selected" column header (Panel B). This will reorder the pathways to show the one that has the most of the selected genes in it at the top (Panel C). Click on that pathway (Blue arrow 1).

	Workspace	View	Download	Tools	Columns
8	Add Feature(s)	S FASTA DNA	Table • FASTA •	Pathway Summary 🔠 Multiple	Seq Alignment Show/Hide -
	Pathway Name			# of Genes Selected	# of Genes Annotated
	Synthesis and degra	dation of ketone bodies		8	
	Streptomycin biosyn	thesis		7	
	Butanoate metabolis	<u>sm</u>		5	36
	Caprolactam degrad	ation		1	9
	1,4-Dichlorobenzene	e degradation		2	18
	Fatty acid elongation	n in mitochondria		1	10
	Nicotinate and nicot	inamide metabolism		1	11
	Geraniol degradation	<u>1</u>		1	11
			# of Gene	es Selected	B
	Workspace	View	Download	Tools	Columns
8	Add Feature(s)	FASTA DNA	Table ▼ ≥ FASTA ▼	APathway Summary 🗱 Multiple	e Seq Alignment Show/Hide -
	Pathway Name	1		# of Genes Selected	# of Genes Annotated
	Butanoate metabolis			5	36
	Synthesis and degra	dation of ketone bodies		2	8
	1,4-Dichlorobenzene	degradation		18	
	Propanoate metaboli	sm		2	34
	Fatty acid metabolisi	<u>m</u>		1	29
	Inositol phosphate n	netabolism		1	12
	Tyrosine metabolism	1		1	19
	Nicotinate and nicoti	namide metabolism		1	11
	Porphyrin and chloro	phyll metabolism		1	37

18. This will open up a page that contains a summary of the EC numbers present in this genome on the left, and the KEGG map for that pathway on the right.

KEGG Map	He	eatmap				
EC Table				~~	KEGG Map	
Annotation S	iource :	PATRIC .	Tol	al # of Genomes : 1	Legend 🔸 💾 Save Map 🔒 Print	Мар
EC Number	Genor	me Cc Feature 🔻	Genome Count	Occurrence	BUTANOATE METABOLISM (SS) Buterse 3 dial Of 111 72 0 (S) 2 Academic	Т
1.1.1	1	L <u>2</u>	0	1		
1.1.1.157	1	L <u>2</u>	0	1		
1.1.1.30	1	L <u>2</u>	0	1	Discetyl O 111 303 2-(e.Hyticocythis) Glycolysis	
1.1.1.35	1	L <u>3</u>	0	1	(R,R)-Butans-2,3-diol ○ ← 1.1.1.4 → ○ ← 4.1.15 → 0 ← 2.2.1.6 ○ ← 2.2.1.6	
1.1.99.2	1	1	0	1	(V)-2-Acetolacitate	
1.2.4.1	1	. 1	0	1		
1.3.8.1	1	5	0	1		
2.2.1.6	1	L <u>2</u>	0	2	Funarate Malaste (R)-Malaste	
2.3.1	1	L <u>2</u>	0	1	Butanoate 0 - 2.7.2.7 Dutanoyi-P	
2.3.1.9	1	L <u>5</u>	0	1		
2.6.1.19	1	L 1	0	1		
2.8.3.8	1	L <u>1</u>	0	1	Alania, sparste and entry approximation of the sparse of t	
4.1.1.15	1	1	0	1		
4.1.3.4	1	L <u>1</u>	0	1	4 Amino- White and a set of the	
5.1.2.3	1	. ⊻ I 1	0	1	LGhuhanat & 41.1.15 + O 4 26.1.96 + 4.1.1.61 + O 4 53.33 + 4.1.170 - O 4 42.1. + O glinbary.COA	
6.2.1.16	1	L <u>1</u>	0	1	Successory Successory 42.1.55 Crotonoyl-CoA 1	
					Vitamin B6 metabolism)d/ (R)-3-Hydrogy 42.1.17	
					$\psi^{\text{outandyl-co.n}}$ ψ (R)-3;Hydroxy, $\bigcirc 4$ 3,11,75 $\rightarrow \bigcirc 4$ 2,31.	
					butanoykoxylbutanoste Poly-β-hydroxybutynate 11.99.2	
					S.1.122 Pyrovate	
					(R)-3-Hydroxybutanosts	
					T Alazine, separate and 1.11.30 Acetoacety4-CoA \downarrow (glutanea methodism) \downarrow	
					Synthesis and degradation	
					of latone bodies"	

19. Clicking on the legend opens it. In this example, green boxes indicate that this genome has at least one gene annotated with that EC number. The blue boxes indicate those genes that were part of your selection from the heat map, and white boxes indicate that these genes are not present in this genome.

