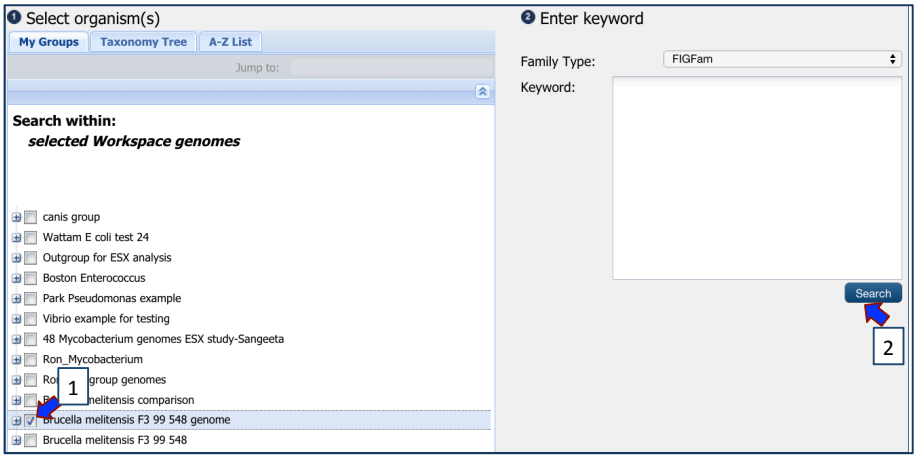
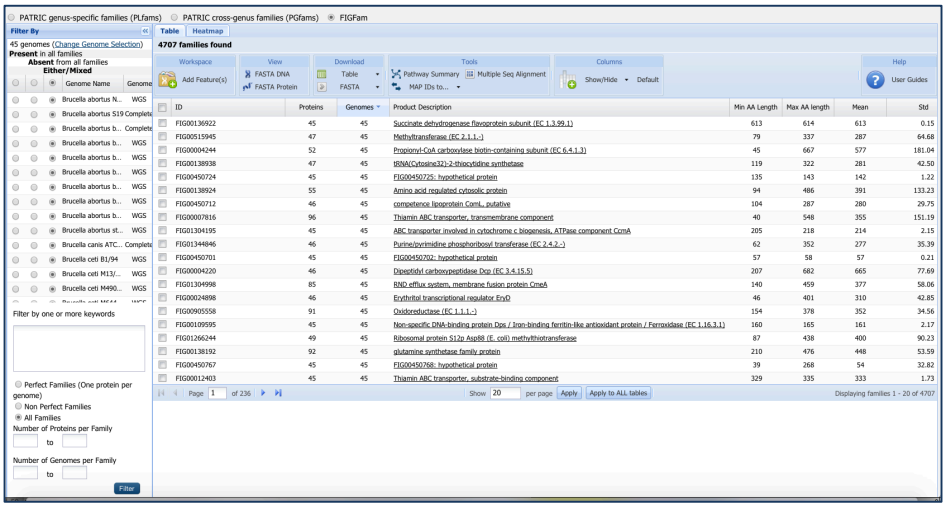
Protein Family Analysis: Protein Family Sorter

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)."

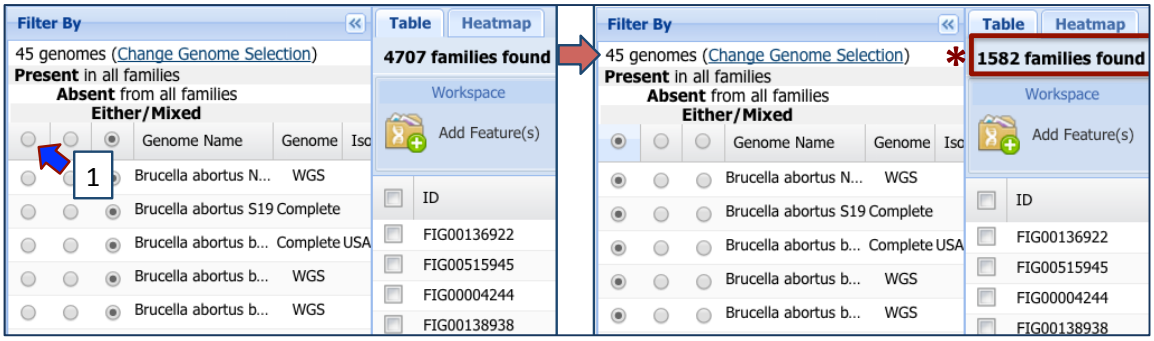
2. The groups that you have created should be visible. Your can select one, or many to compare. In this example, I am selecting a group of 42 genomes that I created, and a group with a single, private genome that I annotated in PATRIC (Blue arrow 1 below). I have a lot of groups, so you can only see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



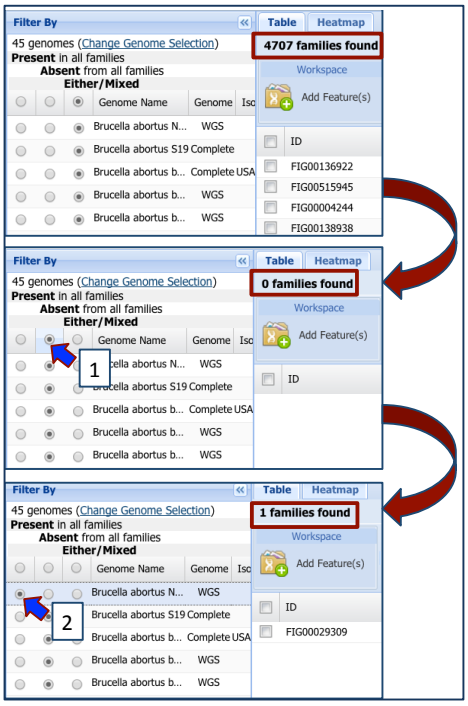
3. **Pan Genome.** This will take you to a page that contains a filter for genomes on the left, and a table of protein families on the right. When this page loads, you are seeing the pan genome for the group you have selected. It contains all the protein families across all the genomes.



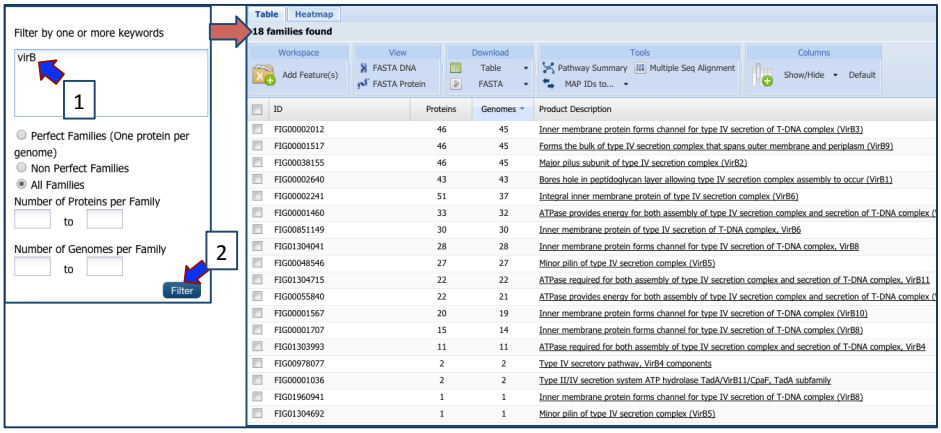
4. **Core Genome.** To find the core genome (the protein families that all genomes have at least one member in), you will need to click on the circle under the column called “Present in all families” in the row called “Genome Name” (Blue arrow 1 below). This will auto select “Present in all families” 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).



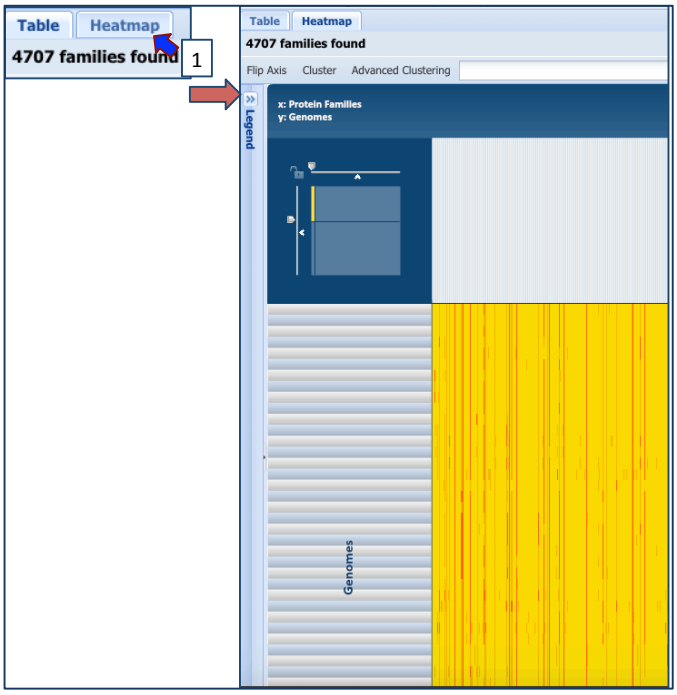
5. **Accessory Genome-One genome at a time**. You can also get the accessory genome, but only one genome at a time (you would have to combine the results in excel later to get the final number). To do this, first check the circle under the column called “Absent in all families” in the row called “Genome 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 meet this condition (highlighted by the red box in the lowest panel).



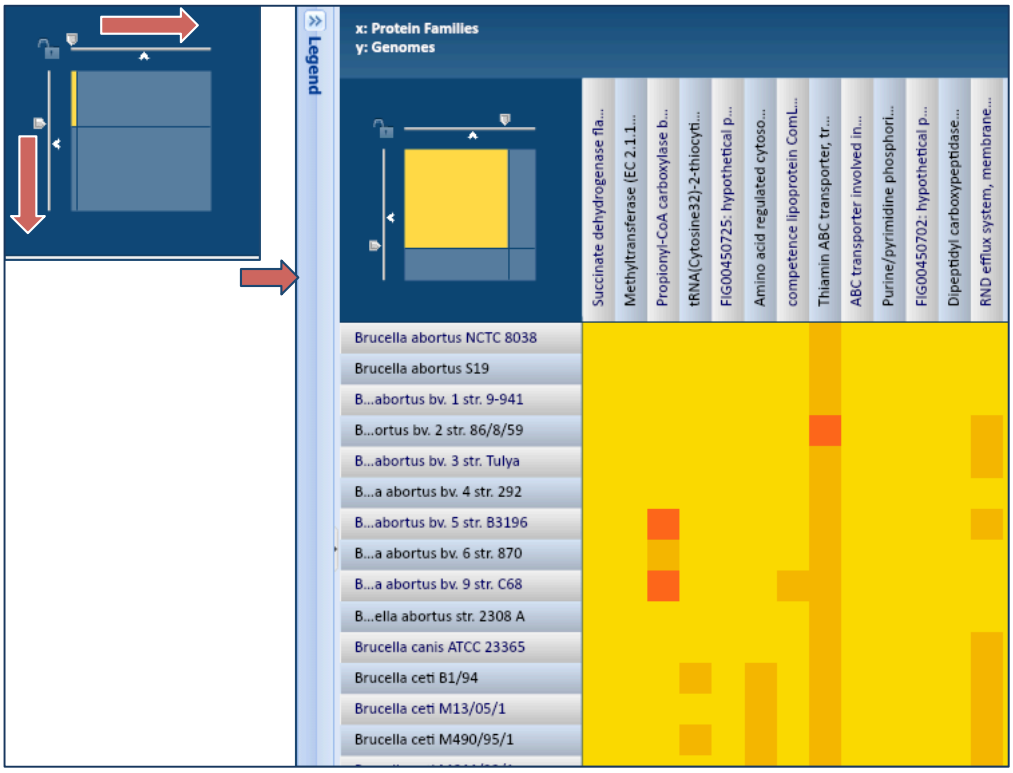
6. Searching for specific names. You can use the text box near the bottom of the filter to find protein families that are specifically named. This will filter on the name of the family, not necessarily on the product description of the individual genes, so the results from the Feature Table and the Protein 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.



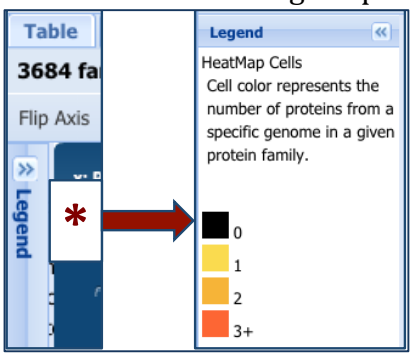
7. **Heatmap view of protein families.** You can see the presence or absence of protein families across particular genomes by looking at the Heatmap view. To see this, click on the Heatmap tab (Blue arrow 1 below) and this will open the heatmap view page. Protein families are listed on the x-axis across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



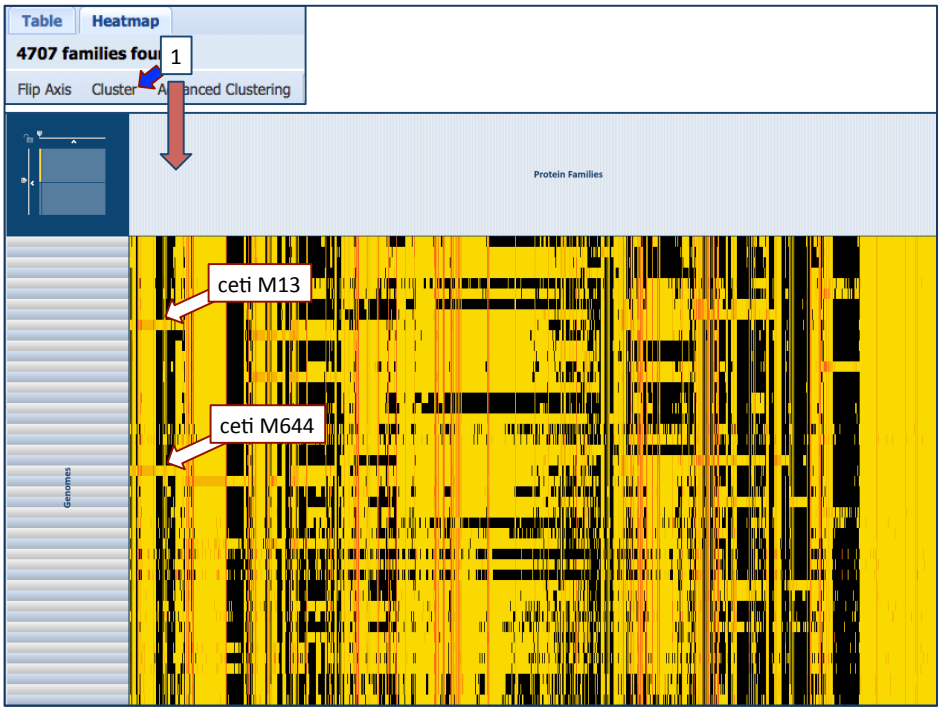
8. To see the names of the individual protein families and/or genomes, you will have to move the sliders to expand the view. In the box at the of the heat map, move the x-axis slider to the right as indicated by the arrow, and the y-axis slider down. This will open the heatmap so that you can read both 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.



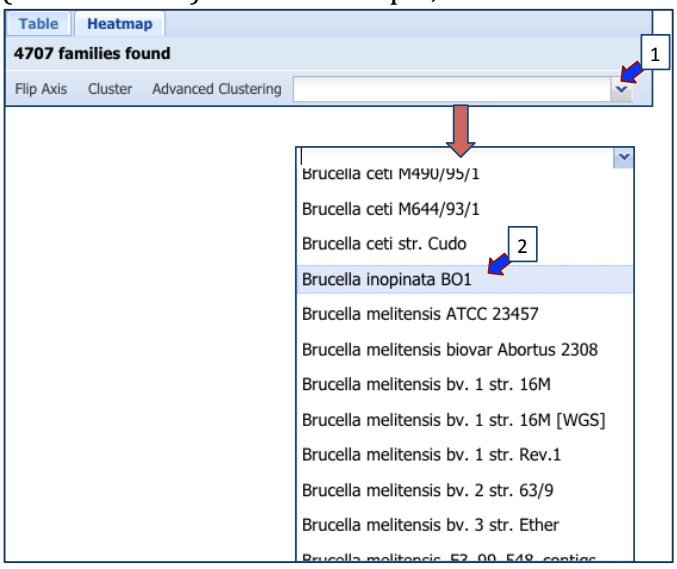
9. **Legend.** You can click on the legend (Red asterisk) to see what the different colors of the cells mean, but basically, black cells indicate a gene absence and all other colors indicate gene presence.



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.

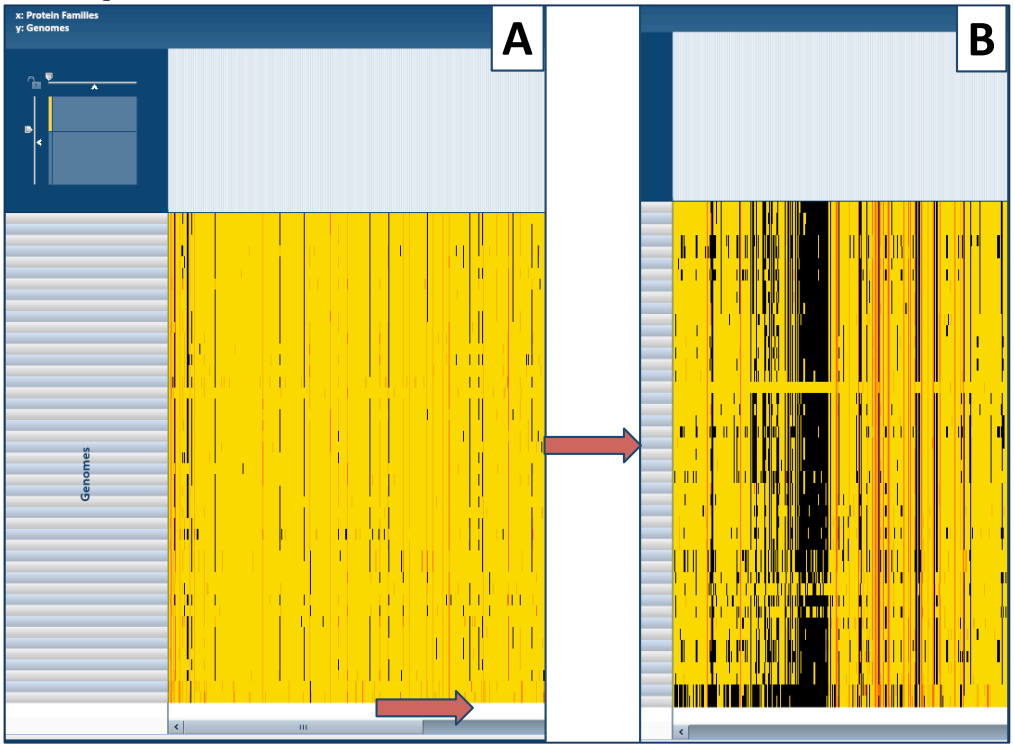


11. **Anchoring.** PATRIC also provides a function where you can examine the presence and absence of protein families across all the genomes, but as they appear in the order of a single genome. This is called "anchoring" and it is a good way to look for genomic islands. To anchor the protein families 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.

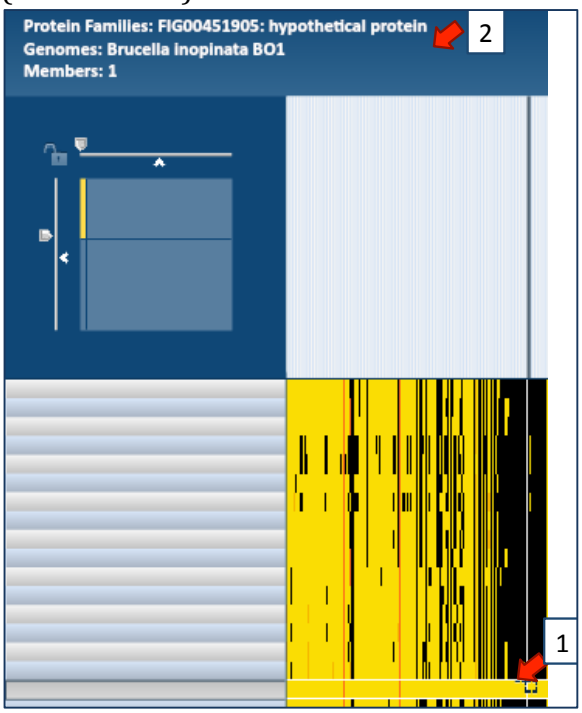


12. This will resort the protein families to occur in the order that the genes occur on the BO1 genome. You can use the scroll bar at the bottom of the heatmap to follow

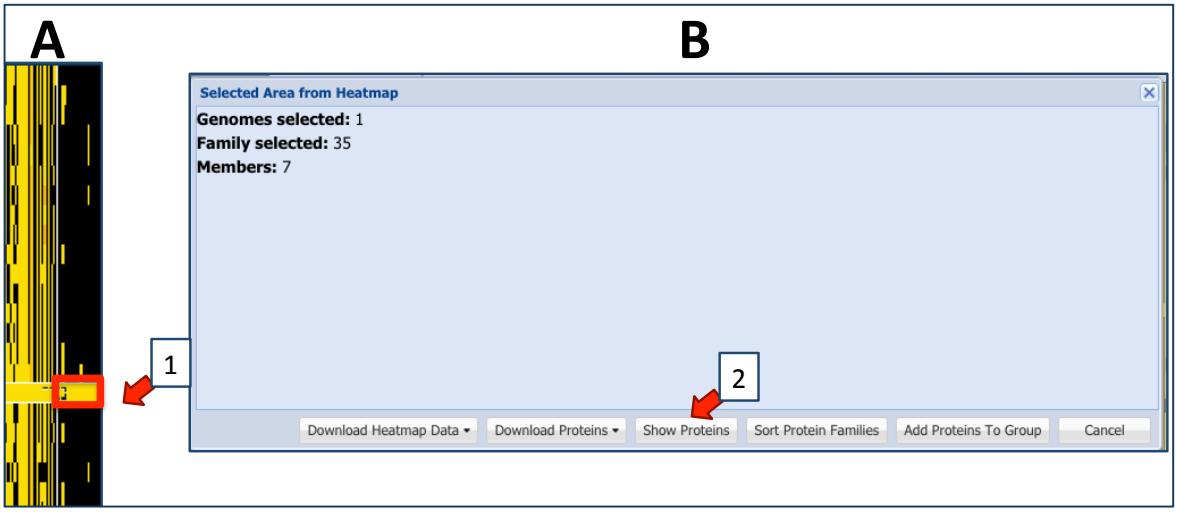
along the genome (from A to B below) until you see an area of interest. In this example, I chose an area that few genomes seem to have, as is indicated by the black cells in panel B.



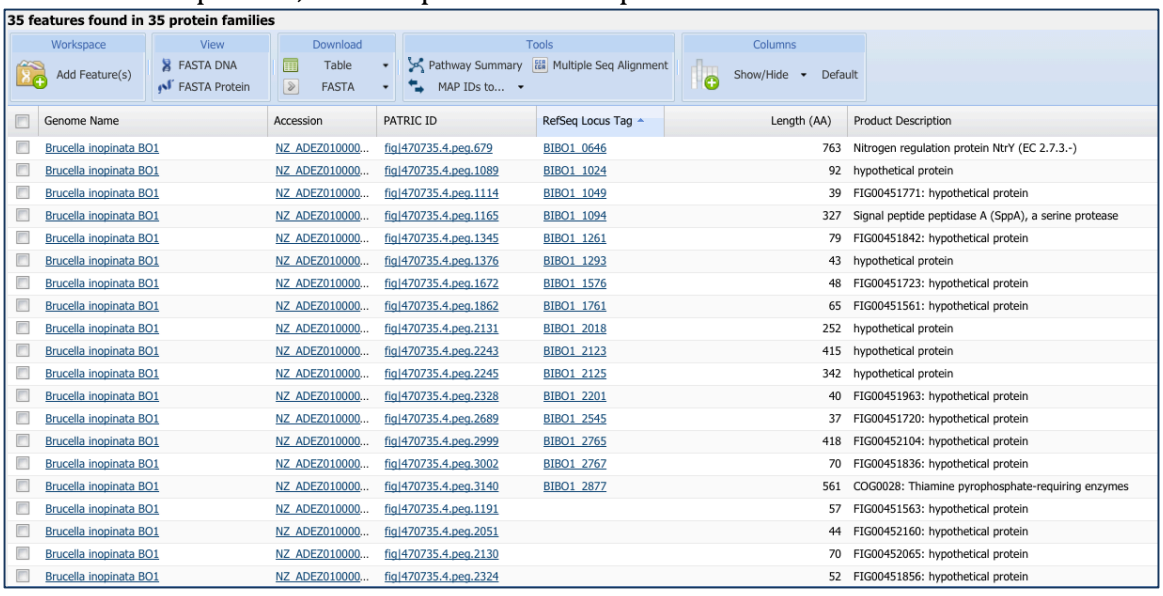
13. You can mouse over individual cells (Red arrow 1 below) to see the names of both the genome and the protein family in the blue header box above the heatmap (Red arrow 2).



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).



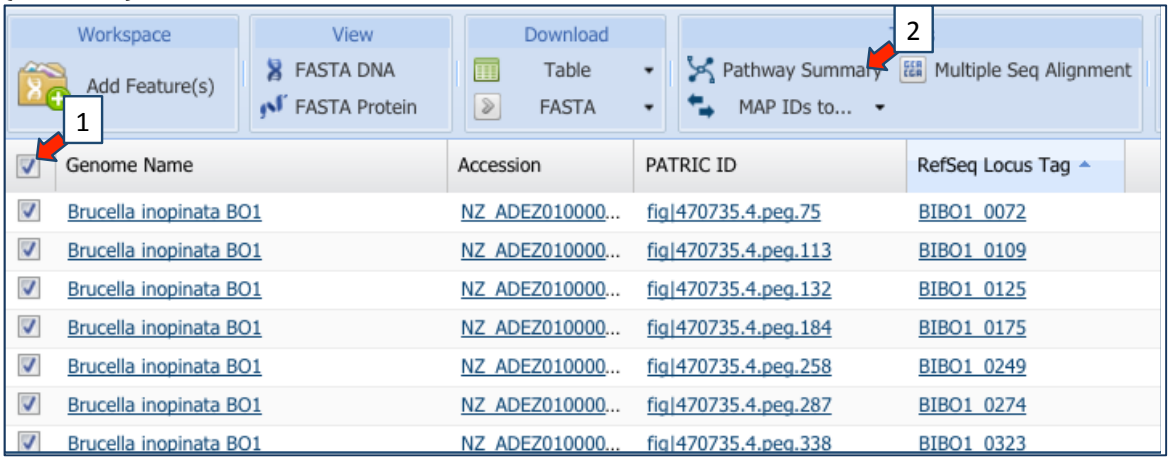
15. A new window will load that shows you the data behind your selection, including the name of the genome, accession number, PATRIC and RefSeq locus tags, the size of the protein, and the product description.



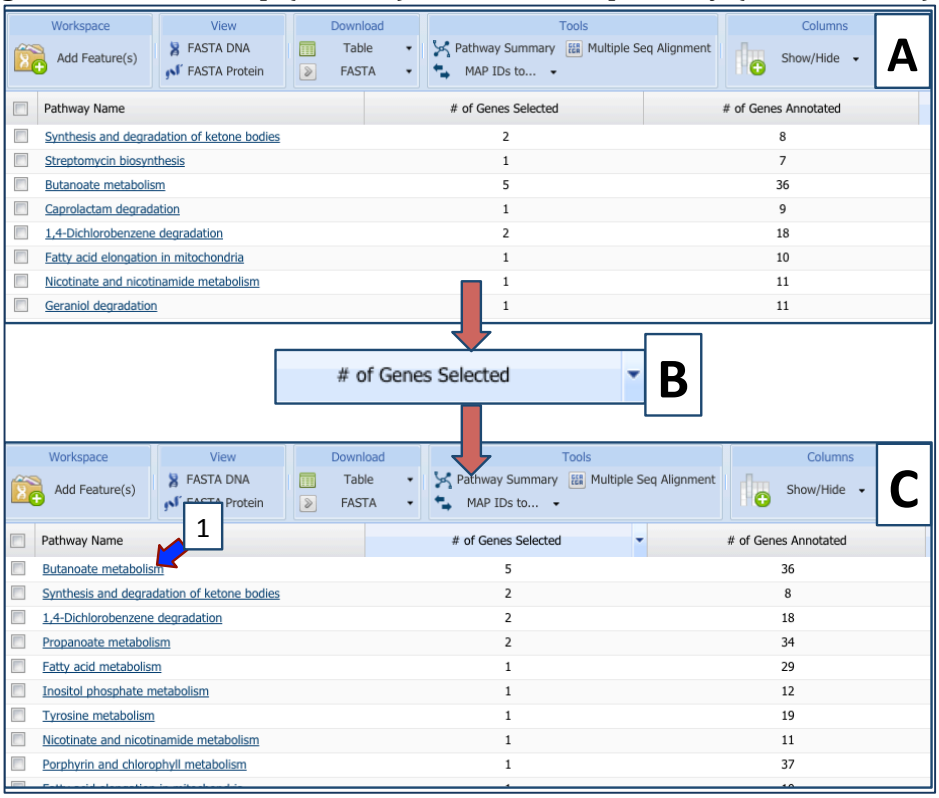
Genome Name	Accession	PATRIC ID	RefSeq Locus Tag	Length (AA)	Product Description
Brucella inopinata BO1	NZ_ADEZ010000...	fig|470735.4.peg.679	BIBO1_0646	763	Nitrogen regulation protein NtrY (EC 2.7.3.-)
Brucella inopinata BO1	NZ_ADEZ010000...	fig|470735.4.peg.1089	BIBO1_1024	92	hypothetical protein
Brucella 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

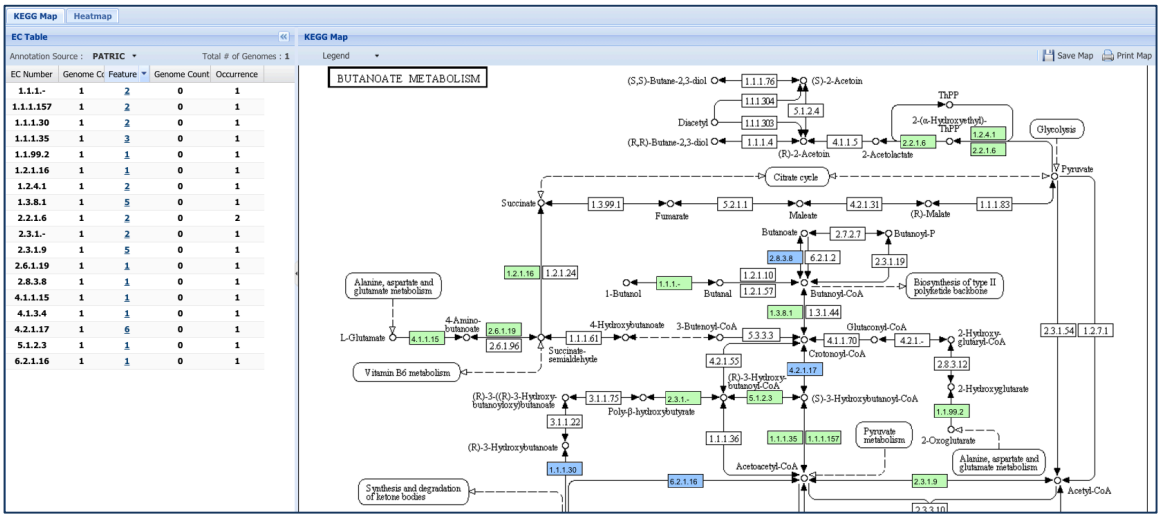
in front of "Genome Name" (Arrow 1 below). This will select all the genes in the table. Then you need to click on the Pathway Summary button at the top of the table (Arrow 2).



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).



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.



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.