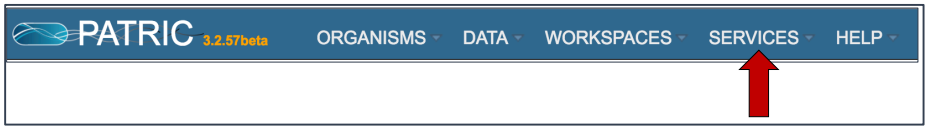
Proteome Comparison: Comparing annotated proteins across genomes.

PATRIC's Proteome Comparison tool can be used to readily identify insertions and deletions in up to nine target genomes that are compared with one reference, which can be a researcher's private genome in PATRIC, a genome that has been annotated outside PATRIC, any of the publicly available genomes in PATRIC, or a set of proteins that you have saved in PATRIC as a feature group.

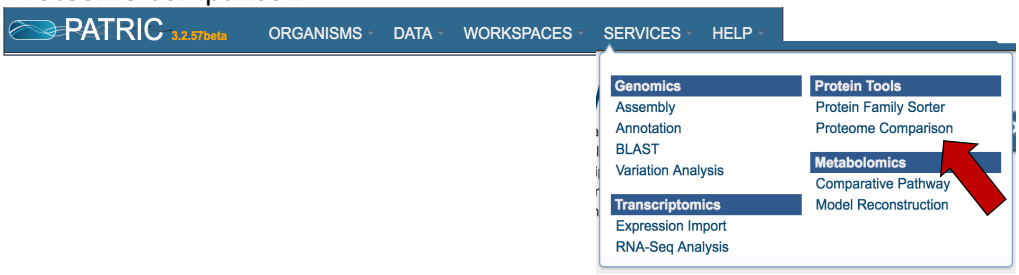
The Proteome Comparison tool is based on the original Sequence-based Comparison tool that was part of RAST[1]. This tool colors each gene based on protein similarity using BLASTP and marks each gene as either unique, a unidirectional best hit or a bidirectional best hit when compared to the reference genome. The output includes a whole-genome schematic that is colored based on BLAST. A table that details all the results can be downloaded for further analysis, as can a scalable vector graphic (svg) diagram of the results that is publication quality.

I. Finding the tool.

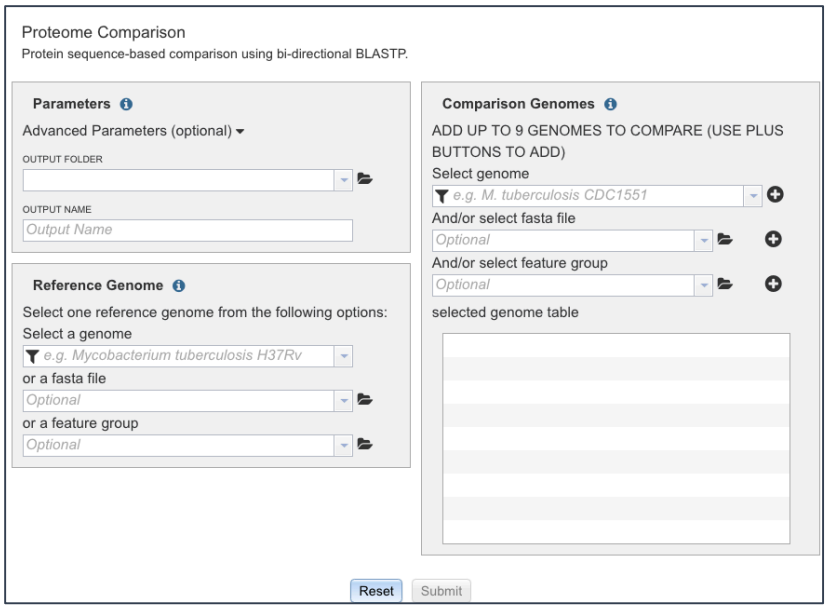
1. At the top of any PATRIC page, find the Services tab and click on it.



2. Click on Proteome Comparison

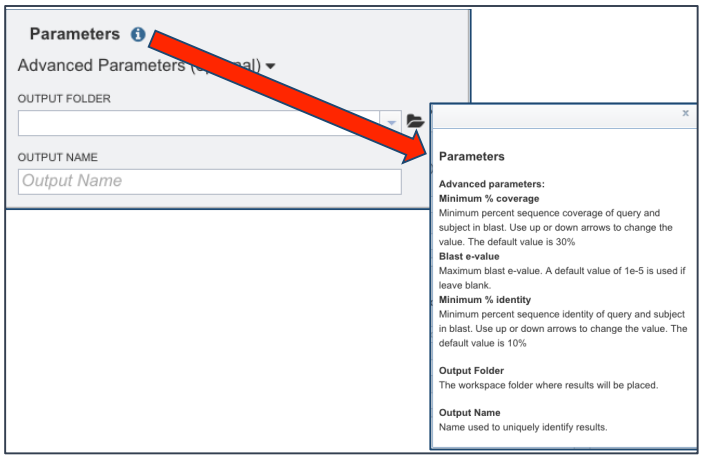


3. This will open up the landing page for where you can submit a Proteome Comparison job. This tool is a best bidirectional BLAST comparison of the annotated proteins from up to nine different genomes.

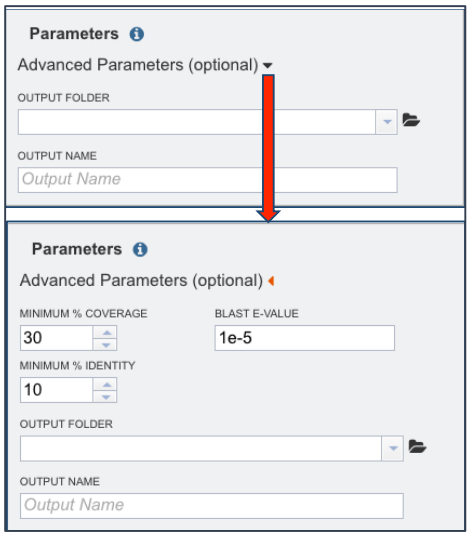


II. Setting parameters and selecting an output folder

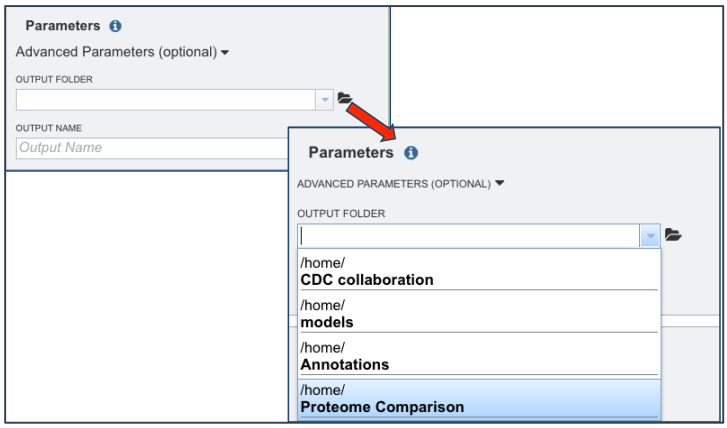
1. To see what the advanced parameters are in the proteome comparison, click on the information icon (Blue I as indicated by the Red Arrow). This opens a pop-up window that describes what can be adjusted.



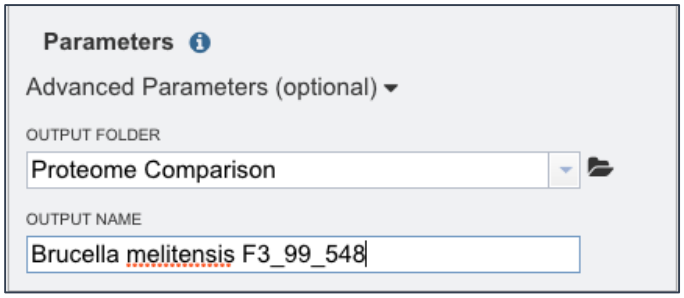
2. The box to adjust the advanced parameters can be opened by clicking on the down arrow that follows Advanced Parameters (optional) as indicated by the Red Arrow. Researchers can adjust the minimum percent coverage, the minimum percent identity and the BLAST E-value.



3. Next the researcher must select an output folder where the proteome comparison job will be placed. To do this, click on the folder icon that follows the text box under the words Output Folder (Red Arrow) and click on preferred folder.

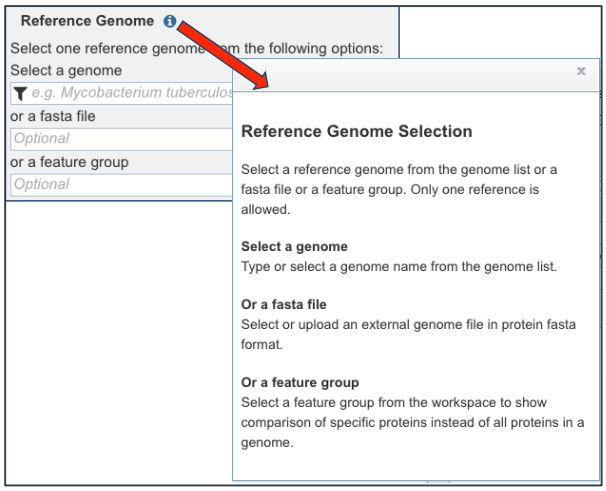


4. Provide a distinctive name for the proteome comparison in the text box underneath the words Output Name.

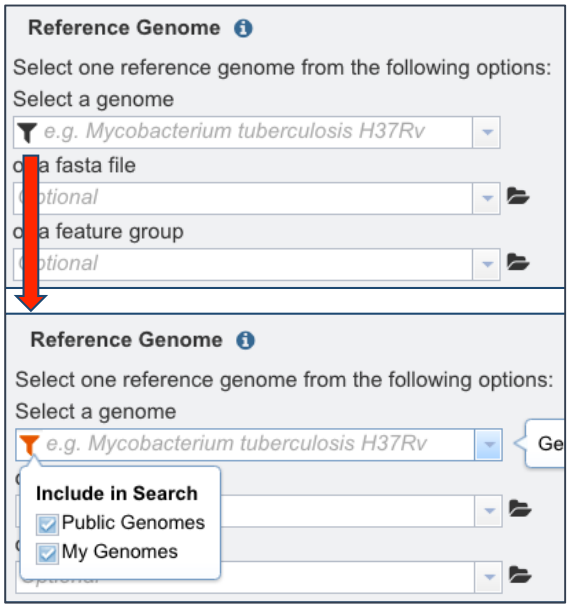


III. Selecting the Reference Genome - Genome

1. The proteome comparison tools allows researchers to select genomes or a specific feature group that contains a set of proteins to serve as a reference that other genomes will be BLASTed against. To see and understand the available options, click on information icon (Blue I as indicated by the Red Arrow). This opens a pop-up window that describes the types of selections that can be made.

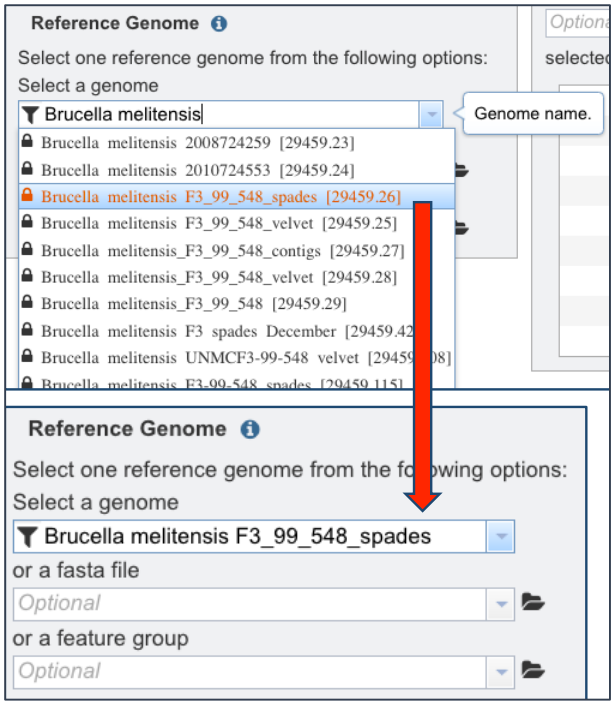


2. In this example, a private genome is selected. To do this, first click on the filter icon (Red Arrow under Reference Genome). This will open a box that allows a researcher to search across all of the public genomes available in PATRIC, or across the genomes that they have annotated and that are stored in their private workspace.



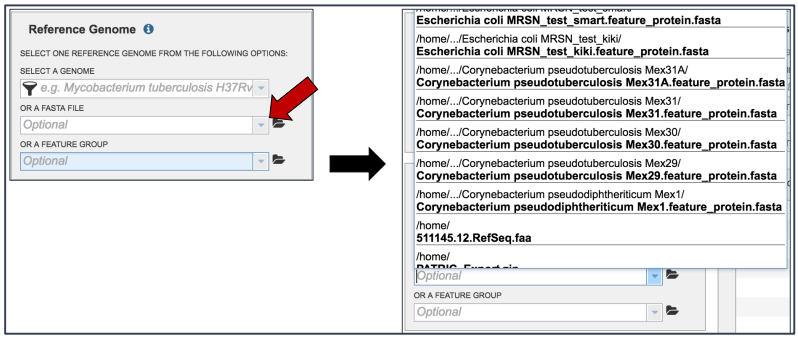
3. Click on the box in front of Public Genomes (Red Arrow) to deselect that box.

4. In the text box below Reference Genome, start typing some words that will identify the reference. Once the user starts typing in the text box, a list appeared below the box, providing the user with possible choices that match the text. Clicking on a name (Red Arrow) will auto-fill the name in the text box under Reference Genome (lowest panel).

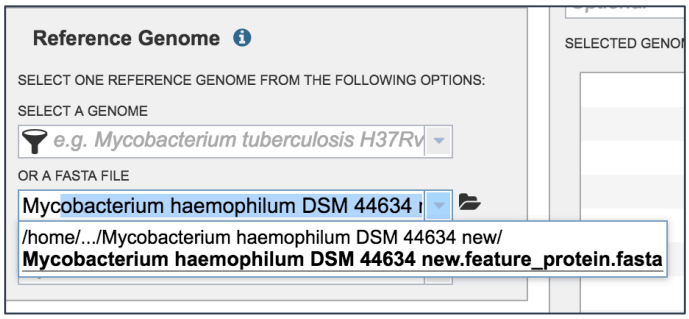


IV. Selecting a Reference Genome – FASTA file

1. Researchers can use the drop-down box to find a fasta file of interest that follows the text box by clicking on the down arrow (red arrow).

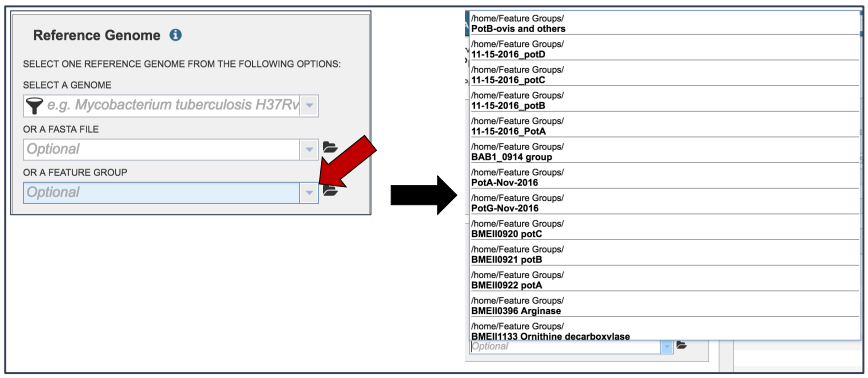


2. However, if a researcher has a number of fasta files in their workspace, the easiest way to find one of interest is to begin typing the name of the group. This will generate a drop down box that shows the choices matching that text.

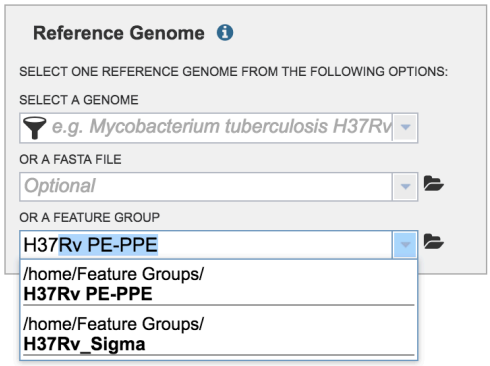


V. Selecting a Reference Genome – Feature group

1. Researchers can use the drop-down box to find a feature group of interest that follows the text box by clicking on the down arrow (red arrow).

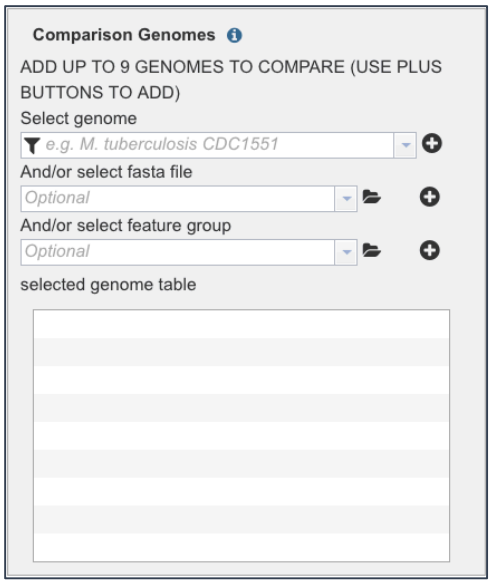


2. However, if a researcher has a number of feature groups, the easiest way to find one of interest is to begin typing the name of the group. This will generate a drop down box that shows the choices matching that text.

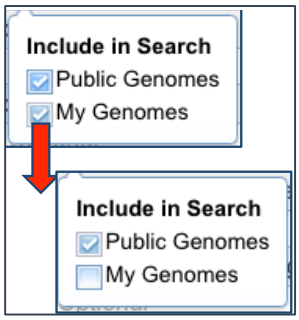


VI. Selecting the Comparison Genomes.

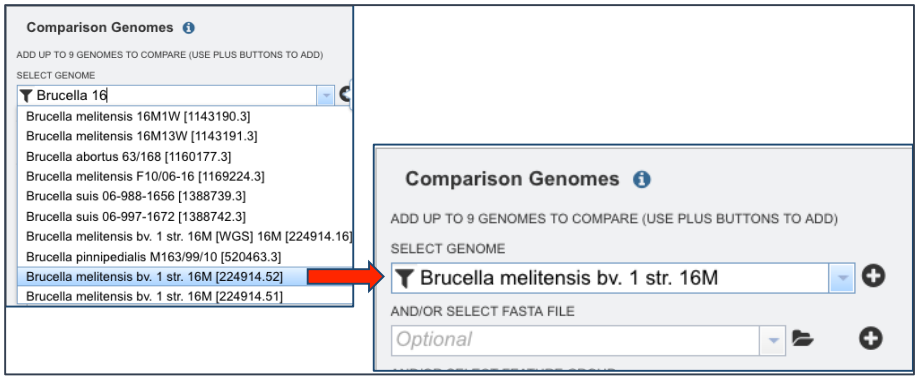
1. Locate the panel for selecting comparison genomes, or a set of proteins, that will be BLASTed against the selected reference.



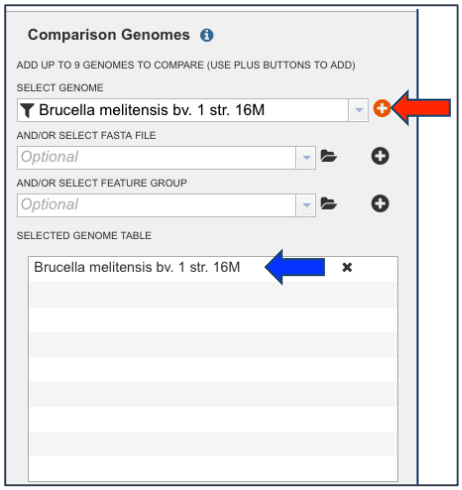
2. Comparison genomes can be public or private. To select genomes that are publicly available, deselect the check box in front of Private Genomes.



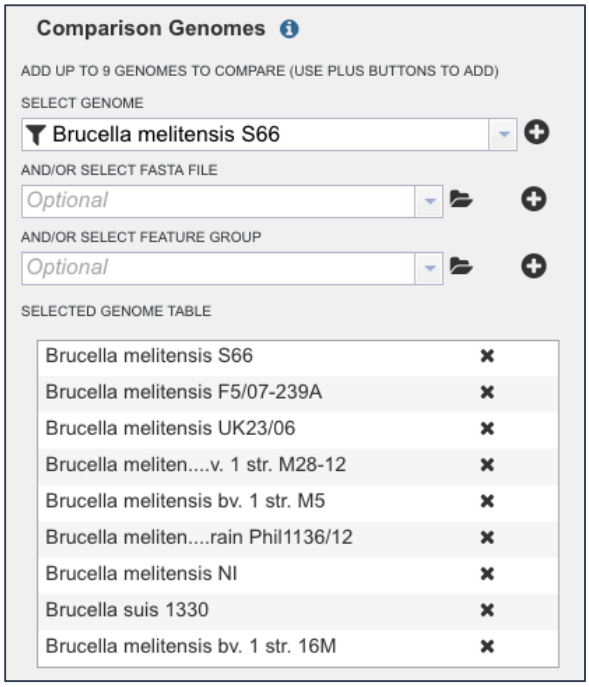
3. Start typing a name into the text box. Once enough text has been entered to see the genome of interest, click on that (Red Arrow 1 in Panel A below). This will auto-fill the text box with the selected name (Panel B). If the choice is correct, click the "+" icon (Red arrow 2). The genome will then appear in the box below (Panel C).



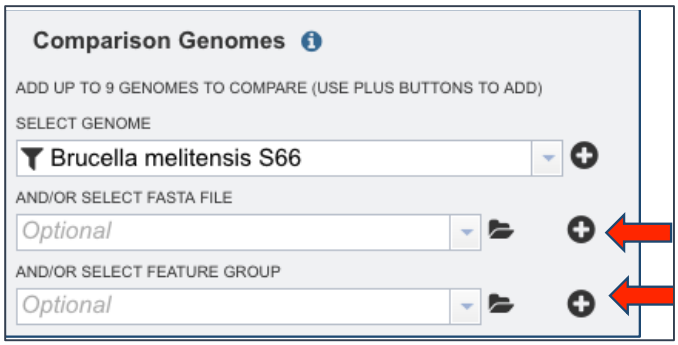
4. To finally selecting the genome, click on the + icon at the end of the text box that has the name of the selected genome (Red Arrow). Clicking on the icon will move the genome to the Selected Genome table (Blue Arrow).



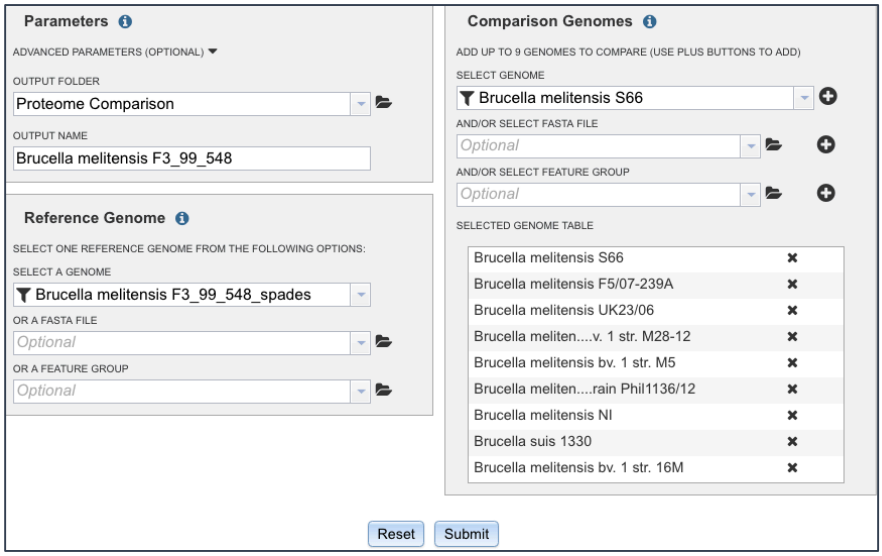
5. Repeat step 3 to add as many genomes (up to 9) to compare to the reference genome you have selected.



6. Genomes that have been annotated by a different service, or a specific group of proteins, can also be used in the comparison. Those can be included as indicated by the red arrows as seen below.

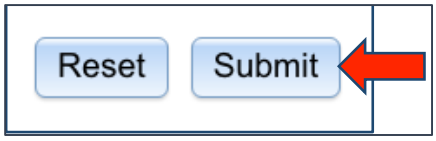


7. Once the nine genomes (or feature groups) are included, the Comparison Genome box will show all the members you have selected.

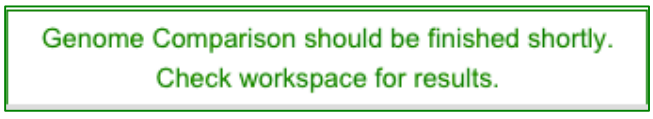


VII. Submitting the proteome comparison job.

1. Click the Submit button at the bottom of the page (Red Arrow)



2. A message will appear that confirms that the job has been submitted. This message is temporal and will disappear after several seconds.

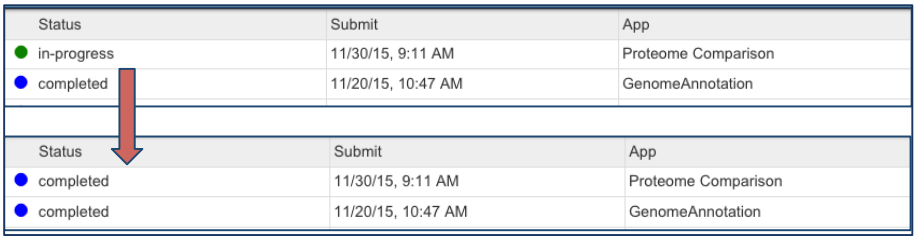


3. To check the status of the annotation job by clicking on the Jobs indicator at the bottom left of the PATRIC page.

4. Clicking on Jobs opens the Jobs Status page, which shows the status of the proteome comparison job. The statuses of all the previous service jobs that have submitted to PATRIC are also available.

Status	Submit	App	Output Name	Start	Completed
in-progress	11/30/15, 9:11 AM	Proteome Comparison	B_melitensis_comparison_16M_reference	11/30/15, 9:11 AM	
completed	11/20/15, 10:47 AM	GenomeAnnotation	Brucella melitensis_F3_99_548	11/20/15, 10:47 AM	11/20/15, 10:52 AM

5. You will be able to see when your job is complete.



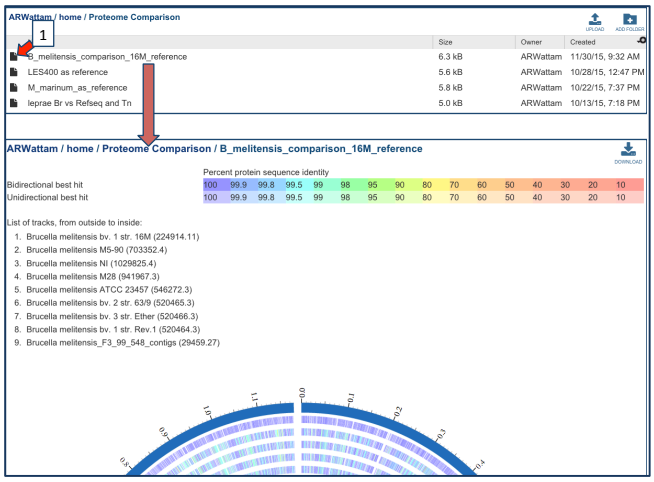
Status	Submit	App
in-progress	11/30/15, 9:11 AM	Proteome Comparison
completed	11/20/15, 10:47 AM	GenomeAnnotation

Status	Submit	App
completed	11/30/15, 9:11 AM	Proteome Comparison
completed	11/20/15, 10:47 AM	GenomeAnnotation

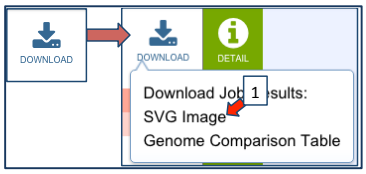
VIII. Accessing the Proteome Comparison job.

1. To access the results of the job, click on Workspace Home that is found at the top left of any PATRIC page. This will open up the workspace, where all folders are visible. Find the folder where the comparison job was placed and click on the folder icon that precedes the name (Red arrow 1).

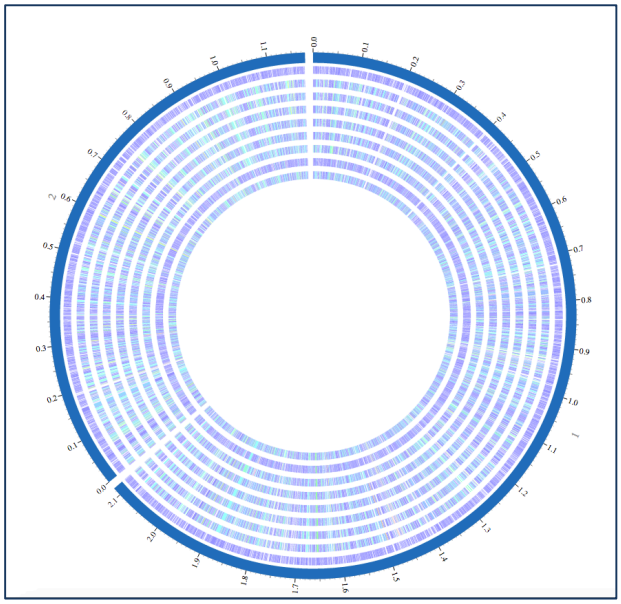
2. This will open a page that shows all the comparisons available in that folder. Click on the icon in front of the job name (Red arrow 1). This opens up the landing page for that job that shows a diagram showing the sequence identity, a list of the names of the genomes in the job, and will also show a circular image that shows the relatedness. This will take a few seconds to load.



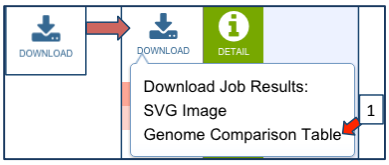
3. To see the entire image, find the download icon at the top of the page on the left and click on SVG image (Red arrow 1).



4. This will download the publication quality SVG image that shows the percent identity across all the proteins in the comparison genomes compared to the reference genome.



5. To examine the data underneath the visualization, click on Genome Comparison Table (Red arrow 1).



6. This will open up a text file, that can be opened with excel. The document contains a lot of information, and users will be able to see all the genes in the comparison genomes that have the best BLAST hits to the reference genome. The first row will show the genome names in specific columns, and the column heading in the second row show the additional information. Data begins with the genome that was used as a reference (2A-J) and includes the following: accession number for the contig in the reference genome (Column A); the order number of this gene in the genome (B); size in amino acids (C); PATRIC locus tag (D); RefSeq locus tag (E); gene name (F); functional annotation (G); start location for the gene on the contig (H); end of the gene on the contig (I); and strand that the gene is located on (J). This is followed by information on the comparison genomes. This data in columns K-T for row 2 (for the first comparison genome) include: data on the type of BLAST hit (Column K, bi- or uni-directional, or missing); contig that the gene is located on (L); the order number of this gene in the genome (M); size in amino acids (N); PATRIC locus tag (O); RefSeq locus tag (P); gene name (Q); functional description (R); percent identity of the BLAST hit (S); and sequence coverage compared to the reference (T). This pattern is repeated for all comparison genomes.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
Brucella melitensis bv. 1 str. 16M										Brucella melitensis M5-90					
ref_genome_	ref_genome_	ref_genome_	ref_genome_patric_id	ref_genome_	ref_genome_	ref_genome_	ref_genome_	ref_genome_	ref_genome_	comp	comp_genor	comp_genor	comp_genor	comp_genor	comp_genor
NC_003317	1	341	fig|224914.11.peg.1	BMEI0001	hemE	Uroporphyri	192	1217	+	bi	CP001851	2045	341	fig|703352.4	BM590_A2050
NC_003317	2	178	fig|224914.11.peg.2	BMEI0002		Protoporphy	1214	1750	+	bi	CP001851	2044	166	fig|703352.4	BM590_A2049
NC_003317	3	421	fig|224914.11.peg.3	BMEI0003	rho	Transcriptior	2132	3397	+	bi	CP001851	2043	421	fig|703352.4	BM590_A2048
NC_003317	4	124	fig|224914.11.peg.4	BMEI0004		Uncharacteri	3482	3856	-	bi	CP001851	2042	124	fig|703352.4	BM590_A2047
NC_003317	5	475	fig|224914.11.peg.5	BMEI0005		Uncharacteri	3984	5411	-	bi	CP001851	2041	475	fig|703352.4	BM590_A2046
NC_003317	6	442	fig|224914.11.peg.6	BMEI0006	trmE	GTPase and t	5526	6854	+	bi	CP001851	2040	442	fig|703352.4	BM590_A2045
NC_003317	7	636	fig|224914.11.peg.7	BMEI0007		tRNA uridine	7081	8991	+	bi	CP001851	2039	634	fig|703352.4	BM590_A2044
NC_003317	8	213	fig|224914.11.peg.8	BMEI0008	gidB	16S rRNA (gu	8988	9629	+	bi	CP001851	2038	213	fig|703352.4	BM590_A2043
NC_003317	9	265	fig|224914.11.peg.9	BMEI0009		Chromosome	9682	10479	+	bi	CP001851	2037	265	fig|703352.4	BM590_A2042
NC_003317	10	293	fig|224914.11.peg.10	BMEI0010		Chromosome	10523	11404	+	bi	CP001851	2036	293	fig|703352.4	BM590_A2041
NC_003317	11	347	fig|224914.11.peg.11	BMEI0011	holA	DNA polyme	11498	12541	-	bi	CP001851	2035	347	fig|703352.4	BM590_A2040
NC_003317	12	264	fig|224914.11.peg.12	BMEI0012		ABC transpor	12703	13497	-	bi	CP001851	2034	264	fig|703352.4	BM590_A2039
NC_003317	13	294	fig|224914.11.peg.13	BMEI0013		ABC transpor	13494	14378	-	bi	CP001851	2033	294	fig|703352.4	BM590_A2038
NC_003317	14	86	fig|224914.11.peg.14	BMEI0014		hypothetical	14558	14818	+						
NC_003317	15	335	fig|224914.11.peg.15	BMEI0015		ABC transpor	15022	16029	-	bi	CP001851	2032	335	fig|703352.4	BM590_A2037
NC_003317	16	69	fig|224914.11.peg.16	BMEI0016		FIG00450157	16050	16259	+						
NC_003317	17	340	fig|224914.11.peg.17	BMEI0017		Alkanal mon	16256	17278	+	bi	CP001851	2031	340	fig|703352.4	BM590_A2036
NC_003317	18	389	fig|224914.11.peg.18	BMEI0018		Conserved hy	17312	18481	-	bi	CP001851	2030	389	fig|703352.4	BM590_A2035
NC_003317	19	352	fig|224914.11.peg.19	BMEI0019		Transcriptior	18732	19790	+	bi	CP001851	2029	352	fig|703352.4	BM590_A2034
NC_003317	1367	288	fig|224914.11.peg.1367			FIG00451027	1269994	1270860	-						
NC_003317	1368	128	fig|224914.11.peg.1368	BMEI1221		transposition	1270857	1271243	-						
NC_003317	1369	209	fig|224914.11.peg.1369	BMEI1222		transposition	1271273	1271902	-						
NC_003317	1370	704	fig|224914.11.peg.1370	BMEI1223		Transposase	1271799	1273913	-	bi	CP001851	736	704	fig|703352.4	BM590_A0741
NC_003317	1371	298	fig|224914.11.peg.1371	BMEI1224		hypothetical	1273913	1274809	-						
NC_003317	1374	139	fig|224914.11.peg.1374	BMEI1225		FIG00451364	1275189	1275608	+						
NC_003317	1375	264	fig|224914.11.peg.1375	BMEI1226		Antitoxin Hig	1275633	1276427	+						
NC_003317	1376	522	fig|224914.11.peg.1376	BMEI1227		hypothetical	1276369	1277937	+	bi	CP001851	732	522	fig|703352.4	BM590_A0738
NC_003317	1377	78	fig|224914.11.peg.1377			FIG00451722	1277958	1278194	-	bi	CP001851	731	78	fig|703352.4.peg.731	
NC_003317	1378	148	fig|224914.11.peg.1378	BMEI1229		exonuclease,	1278969	1279415	+						
NC_003317	1379	253	fig|224914.11.peg.1379	BMEI1230		hypothetical	1279636	1280397	-						
NC_003317	1380	89	fig|224914.11.peg.1380	BMEI1231		ETC complex	1280792	1281061	-	bi	CP001851	729	89	fig|703352.4	BM590_A0734
NC_003317	1381	196	fig|224914.11.peg.1381	BMEI1232		L-lactate per	1281667	1282257	+	bi	CP001851	728	208	fig|703352.4	BM590_A0733
NC_003317	1382	339	fig|224914.11.peg.1382	BMEI1233		L-lactate per	1282306	1283325	+	bi	CP001851	727	339	fig|703352.4	BM590_A0732
NC_003317	1383	44	fig|224914.11.peg.1383			FIG00450959	1283478	1283612	-	bi	CP001851	726	39	fig|703352.4	BM590_A0731
NC_003317	1384	51	fig|224914.11.peg.1384			FIG00450294	1283707	1283862	+	bi	CP001851	725	51	fig|703352.4	BM590_A0730
NC_003317	1385	219	fig|224914.11.peg.1385	BMEI1235		3-oxoacyl-[ac	1284069	1284728	+	bi	CP001851	724	219	fig|703352.4	BM590_A0729
NC_003317	1386	261	fig|224914.11.peg.1386	BMEI1236		Hypothetical	1284801	1285586	+	bi	CP001851	723	261	fig|703352.4	BM590_A0728
NC_003317	1387	289	fig|224914.11.peg.1387	BMEI1237		UDP-glucose	1285583	1286452	-	bi	CP001851	722	289	fig|703352.4	BM590_A0727

References

1. Overbeek, R., et al., *The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).* Nucleic acids research, 2014. **42**(D1): p. D206-D214.