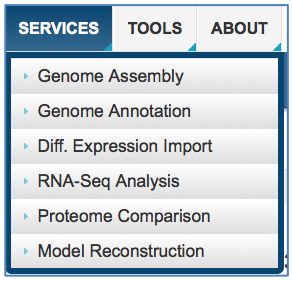
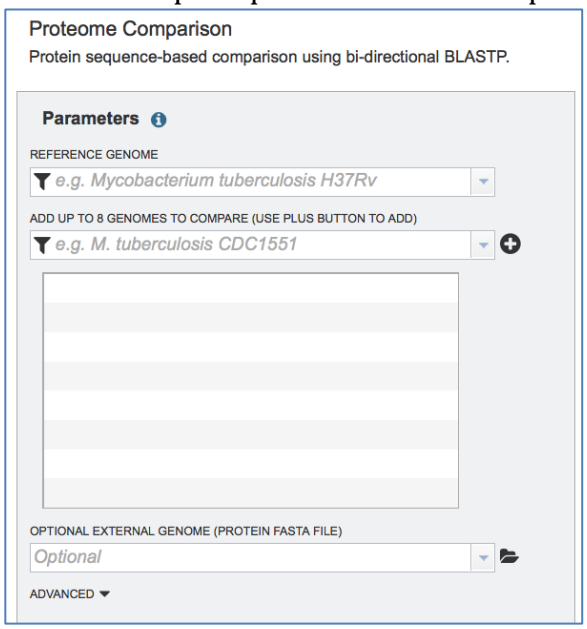
Proteome Comparison: A fine-grained tool for comparative genomics

In addition to the Protein Family Sorter that allows researchers to examine up to the protein families from up to 500 genomes at a time, PATRIC also provides a more detailed and fine-grained way to compare genomes. The Proteome Comparison tool is an enhancement of RAST's 'Sequence Based Comparison Tool' .[1] The Proteome Comparison tool allows users to readily identify insertions and deletions in up to nine target genomes compared with one reference genome, and examine the degree of homology among the genes. The tool colors each gene based on protein similarity using BLAST[2], and each gene is marked as being unique, a unidirectional best hit or a bidirectional best hit in comparison to the reference genome. The output also includes a whole-genome schematic colored by BLAST similarity and available as a Circos SVG image[3]. The resulting data table can also be downloaded for further analysis. The following workflow will demonstrate how to compare genomes using this tool

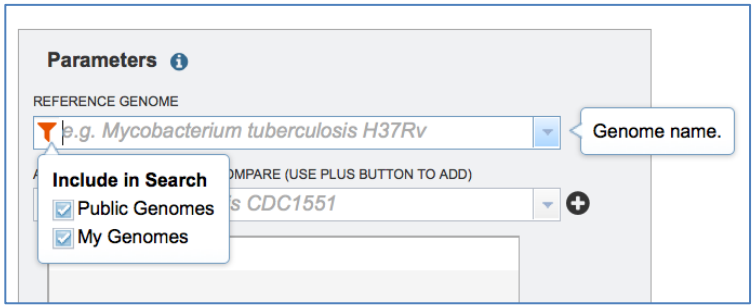
1. Click on the Services tab and then click on Proteome Comparison.



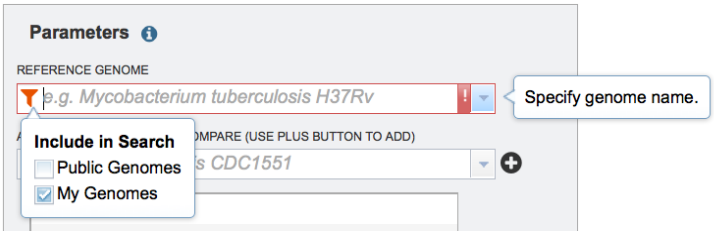
2. This will open up the Proteome Comparison landing page



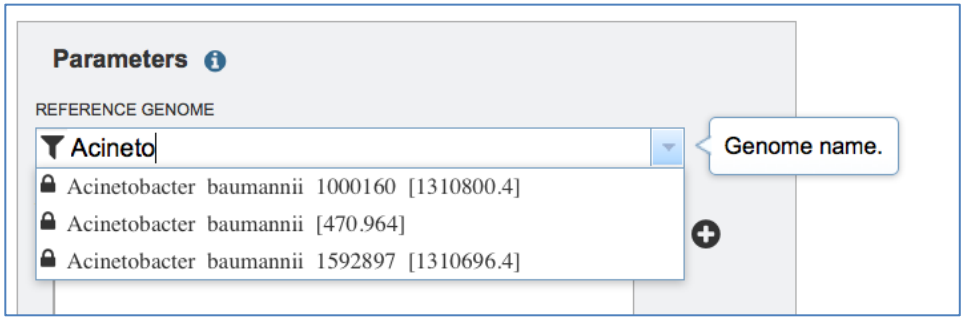
3. In the text box below the words “REFERENCE GENOME” click on the funnel icon. This will open up a pop-up window that has two things to select (Public or My Genomes).



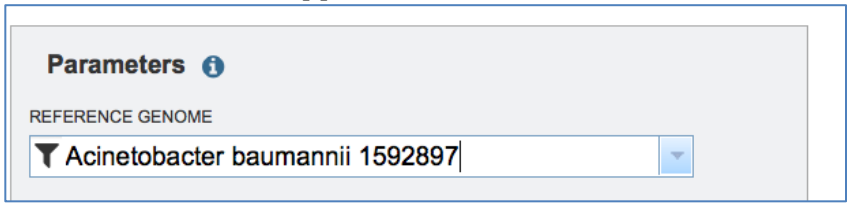
4. Deselect Public Genomes by clicking on the checkbox.



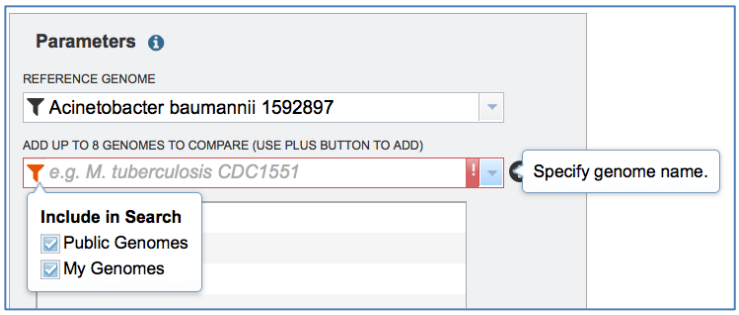
5. To see your private genomes, you will need to start typing the words you used to identify them when you annotated the genome. Here I start with “Acineto” and it will show me all of my private genomes that contain that text.



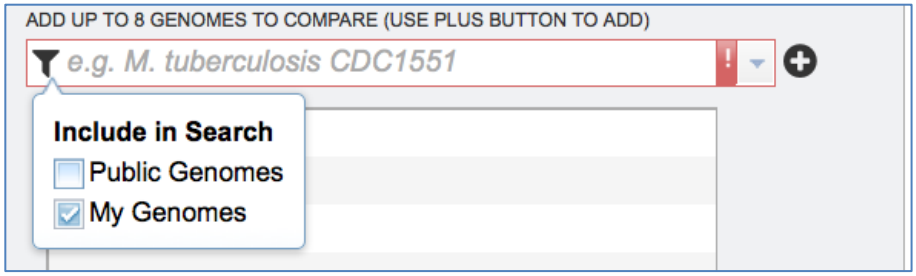
6. Select the Acinetobacter baumannii 1592897 genome by clicking on it. Be sure that the entire name appears in the text box for REFERENCE GENOME.



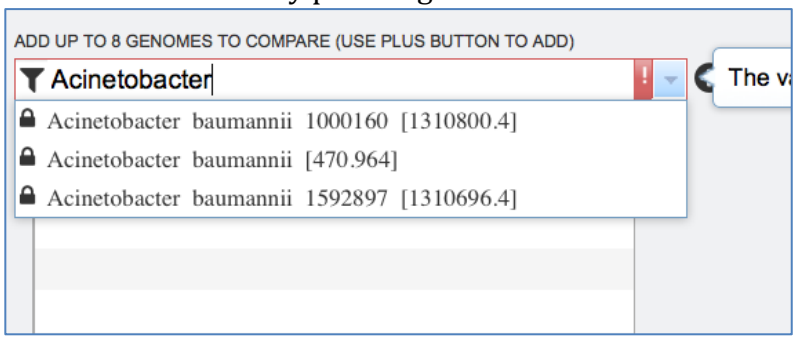
7. Now we need to add the comparison genomes. In the text box below the words “ADD UP TO 8 GENOMES TO COMPARE” click on the funnel icon. This will open up a pop-up window that has two things to select (Public or My Genomes).



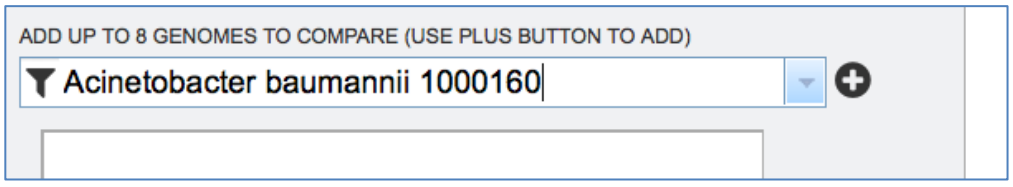
8. Deselect Public Genomes by clicking on the checkbox.



9. To see your private genomes, you will need to start typing the words you used to identify them when you annotated the genome. Here I start with "Acineto" and it will show me all of my private genomes that contain that text.

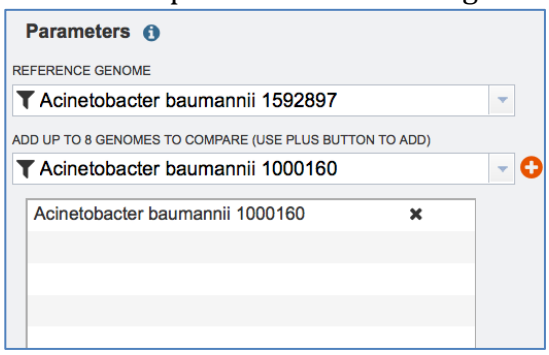


10. Select the Acinetobacter baumannii 1000160 genome by clicking on it. Be sure that the entire name appears in the text box for ADD UP TO 8 GENOMES TO COMPARE.

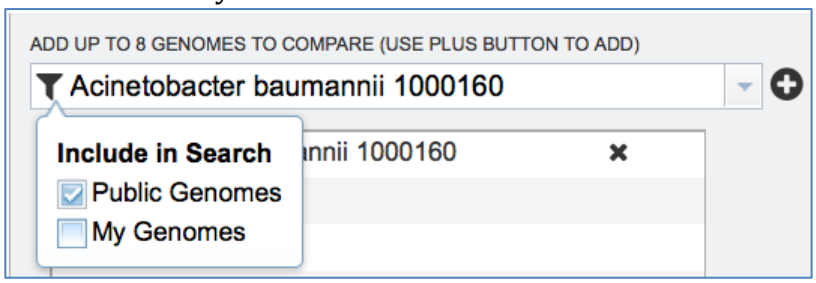


11. You will need to click on the Plus Sign Icon that is at the end of the text box. This will add the genome to the box below. This box contains all the genomes that you

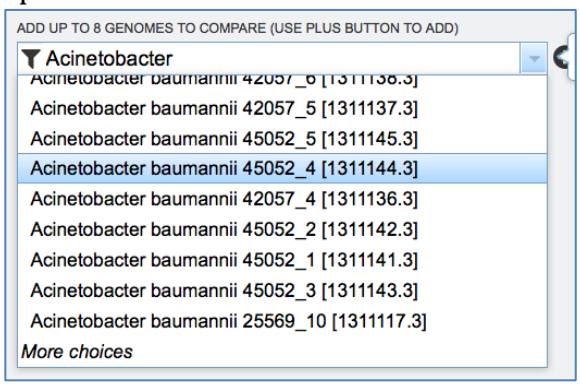
want to compare to the reference genome.



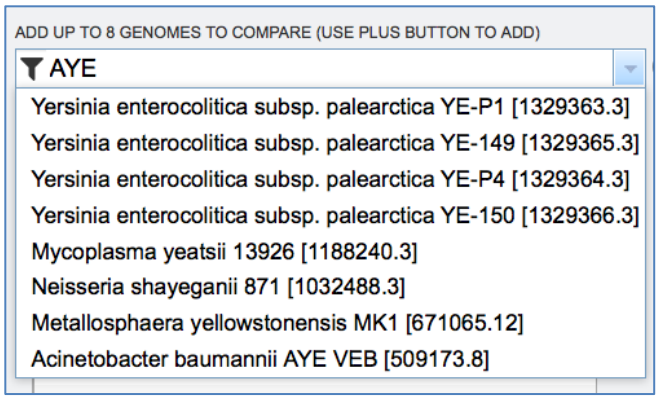
12. Now let's add some public genomes. You'll first have to click on the funnel icon, then deselect My Genomes and select Public Genomes.



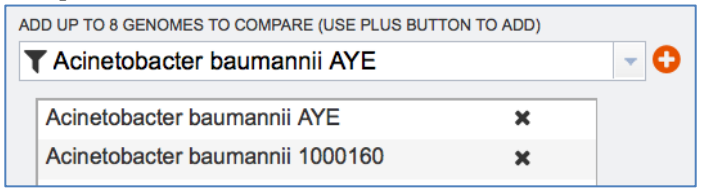
13. If you start typing the word Acinetobacter, you will see that a lot of choices pop up.



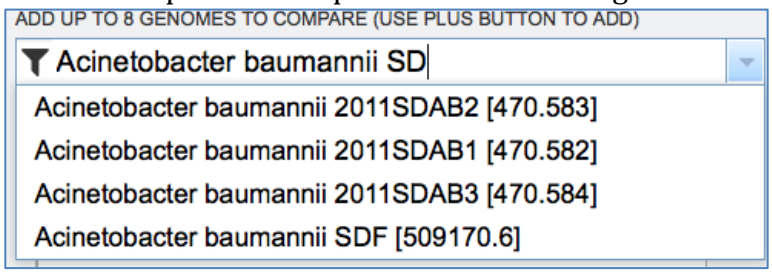
14. To find a specific genome, you will probably need to type something unique to it, perhaps the strain name. Here I typed in "AYE" and then clicked on the line that had "Acinetobacter baumannii AYE VEB [509173.8]".



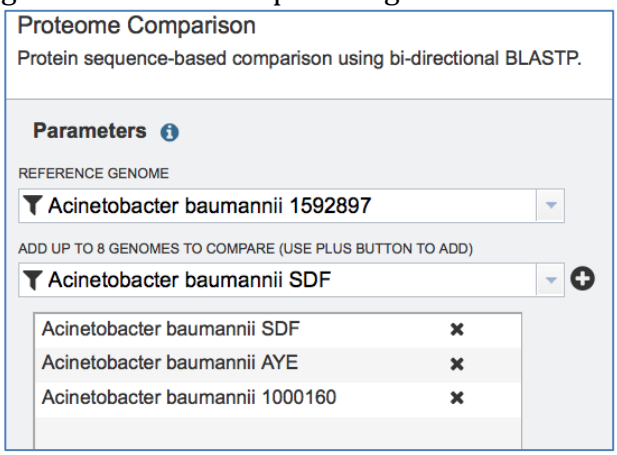
15. You will need to click the Plus Sign Icon to enter that genome into the comparison box below.



16. Use the previous step to include the SDF genome

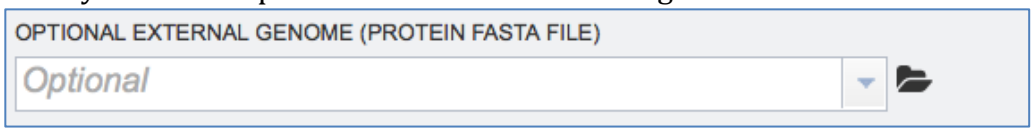


17. You should see one reference genome, and in this case three comparison genomes in the comparison genome box.

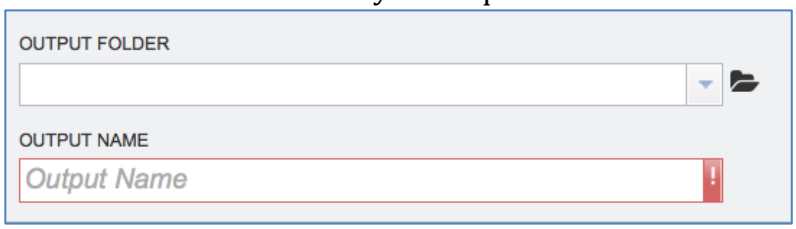


18. If you had an annotation that was not part of PATRIC, you could upload it at this time using the OPTIONAL EXTERNAL GENOME (PROTEIN FASTA FILE) function by

clicking on the folder at the end of the text box. Directions will pop up that will show you how to upload the file. We aren't doing that in this exercise.

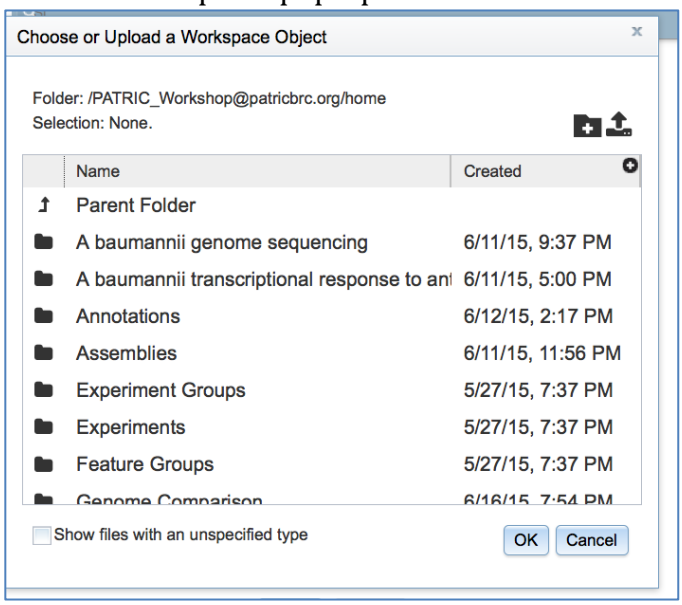


19. You will need to identify an output folder where the results will be placed.



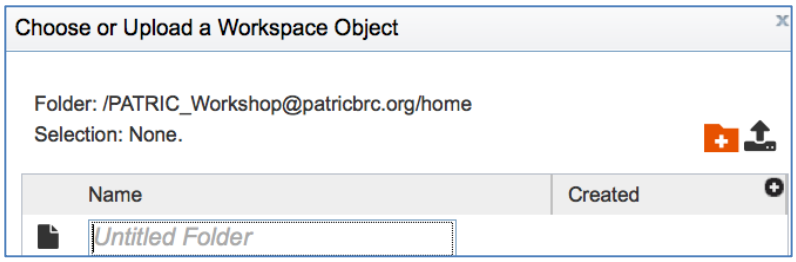
20. At the end of the text box under OUTPUT FOLDER, click on the Folder icon

21. This will open a pop-up window that shows the folders in your workspace.

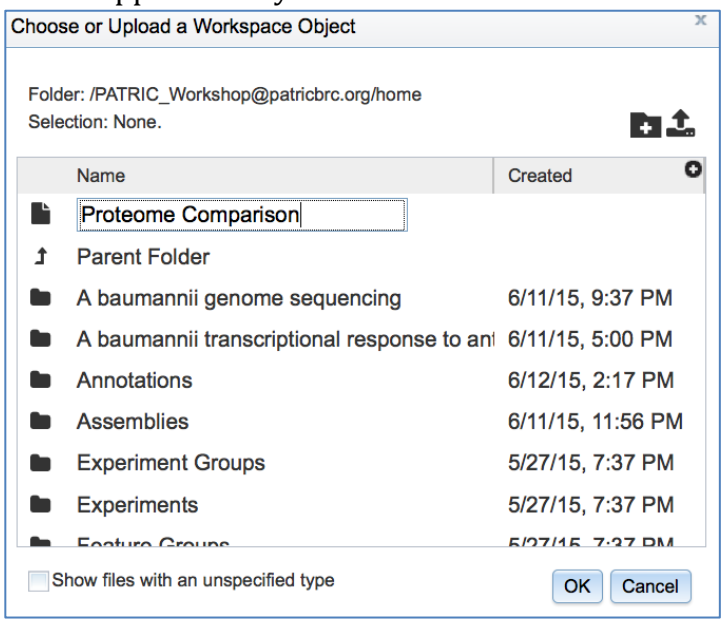


22. For this exercise, we want to name a new folder. In the upper right hand corner of the window you will see a Folder icon with a plus mark in it. Click on that.

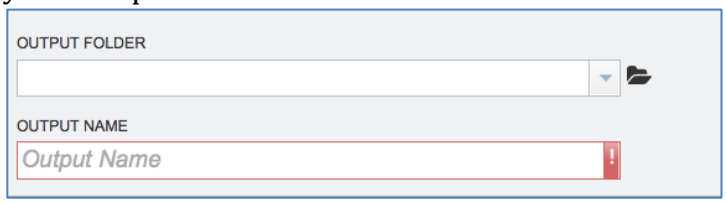
23. Below the folder icon, a Text icon and a text box will appear



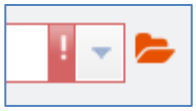
24. I entered the words "Proteome Comparison" in my text box. Once you name your folder, click the OK button at the bottom of the window. The pop-up window will disappear after you do this.



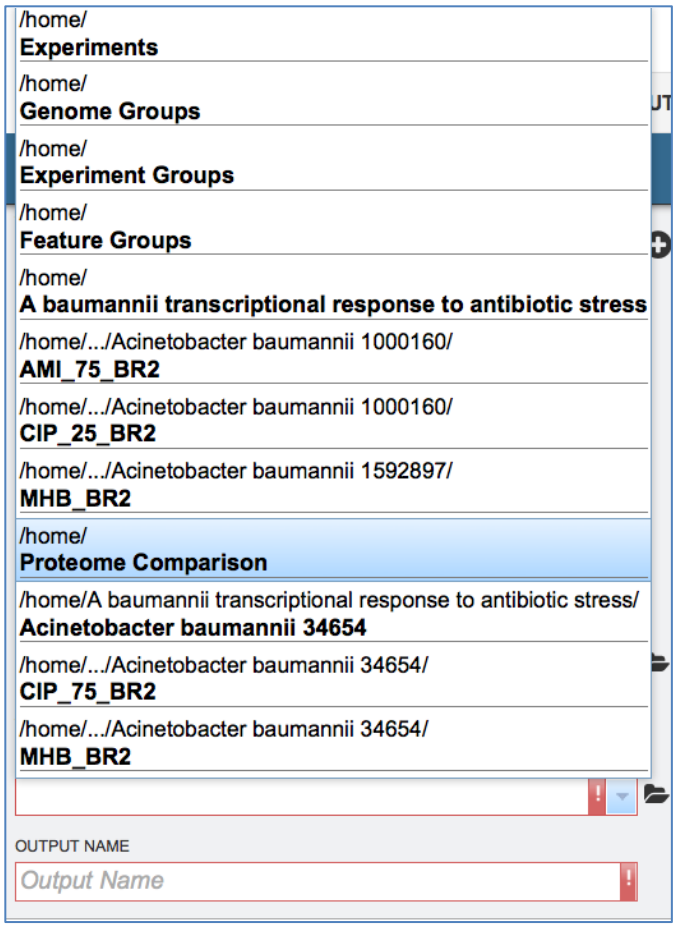
25. Now we need to choose that folder as your OUTPUT FOLDER for the results of your comparison.



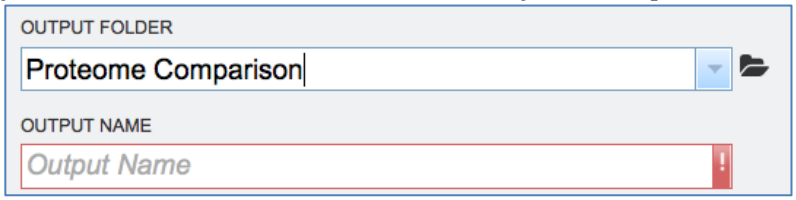
26. Click on the folder icon at the end of the text box.



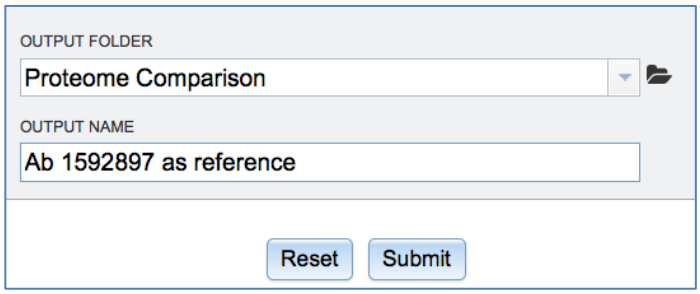
27. This will open up a pop-up window that has all of your folders. I have a lot of folders, and I selected the one I named "Proteome Comparison."



28. Clicking on it enters the name into the text box under OUTPUT FOLDER. Now you will need to choose what to name your comparison

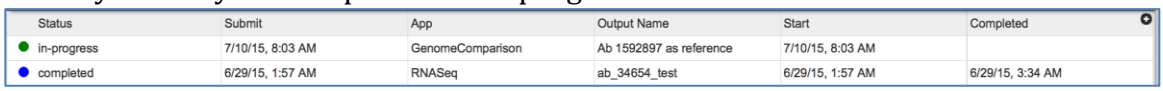


29. In the text box under OUTPUT NAME, enter text that you will recognize as specific to this comparison. I used "Ab 1592897 as reference". Then click the Submit button.



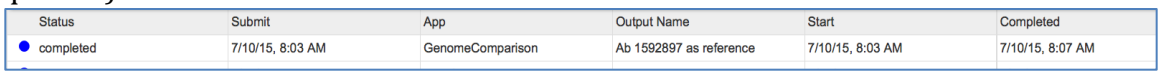
30. At the lower right hand corner of the page, you will see a box called Jobs with number following it. That tells you the number of submitted, in progress, queued and suspended jobs. Double click on the word "Jobs".

31. That will take you to the landing page for your PATRIC jobs. The green circle shows you that your comparison is in progress



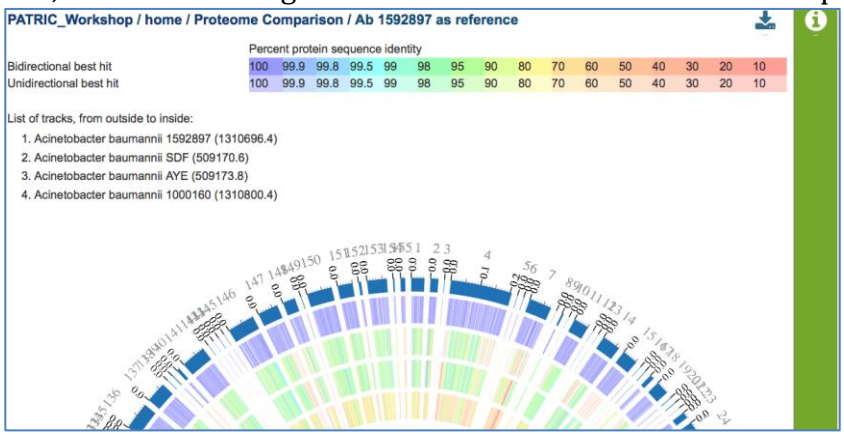
Status	Submit	App	Output Name	Start	Completed
in-progress	7/10/15, 8:03 AM	GenomeComparison	Ab 1592897 as reference	7/10/15, 8:03 AM	
completed	6/29/15, 1:57 AM	RNASeq	ab_34654_test	6/29/15, 1:57 AM	6/29/15, 3:34 AM

32. Once the job is complete, the circle will turn blue. Double click on that blue circle (Note: You can also access the comparison by clicking on WORKSPACE: HOME at the upper right hand corner of the page and finding the right folder, but this is quicker).

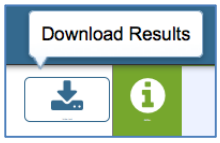


Status	Submit	App	Output Name	Start	Completed
completed	7/10/15, 8:03 AM	GenomeComparison	Ab 1592897 as reference	7/10/15, 8:03 AM	7/10/15, 8:07 AM

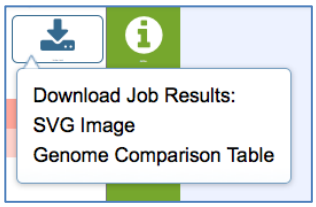
33. This will open up the results of your proteome comparison job. You will see a page that shows you the name, a list of the genomes compared, a figure that shows you're the percent sequence identity for both the uni- and bi-directional best BLAST hits, and a Circos diagram that shows the reference and comparison genomes.



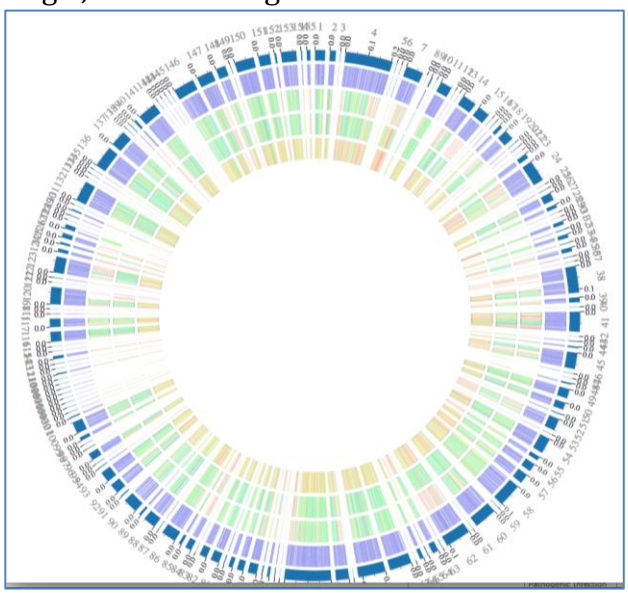
34. To further examine the results, you will need to download them. Click on the Download icon at the upper right corner, next to the green "i".



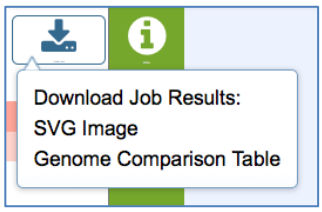
35. This will show you the download choices. Click on the SVG image.



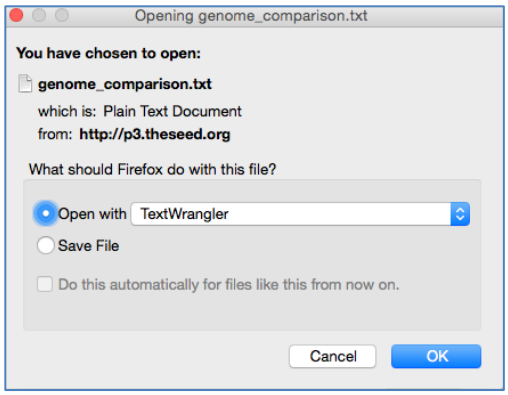
36. This will open up the visualization of the comparison results in Circos. The 1592857 genome is in a lot of contigs, and that is why you are seeing the fractured image. When you use a complete genome, the chromosome will be united into a single, uniform image.



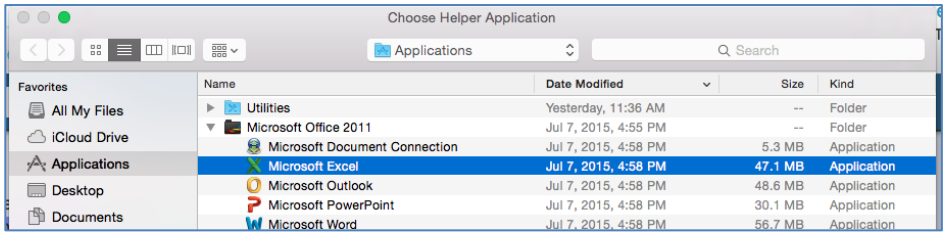
37. To see the data behind the image you will need to download the table. Return to the download icon and click on Genome Comparison Table



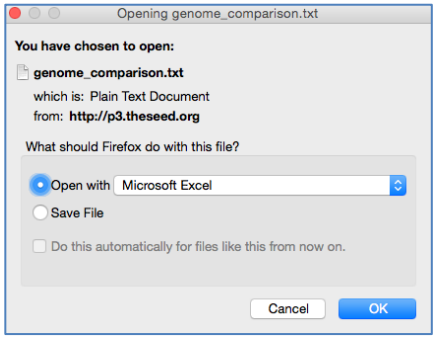
38. A pop-up window appears and I am given the choice to open it with TextWrangler, but I want to see it in excel so I click on the arrow following TextWrangler.



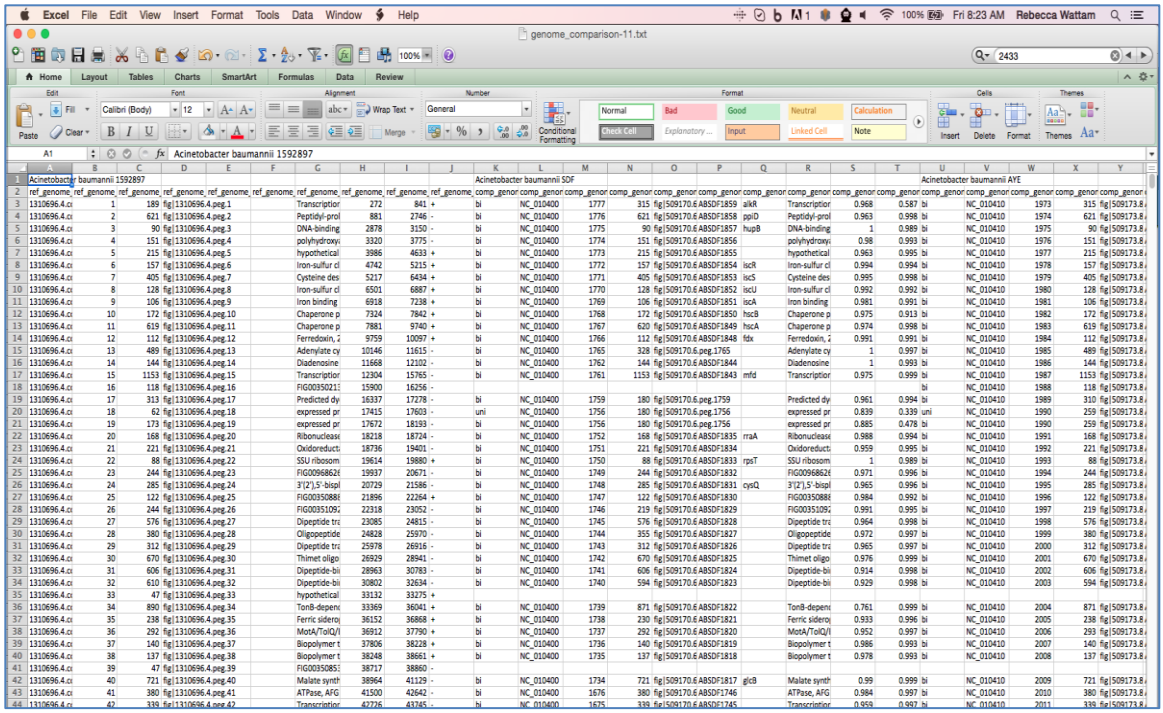
39. I choose Microsoft Excel.



40. Now click the OK button at the bottom of the pop-up window.



41. This will open up an excel file with all the results, and there are a lot of them. Don't be discouraged by all the numbers and columns, this is valuable information and once you explore it you will see how useful it is.. We will go over these in class, but you note that it begins with your reference genome , and then shows you the comparison genomes across Row 1. The second row has the pertinent column heading below those genomes. Click on each one of those column headings so that you can understand the data you will see below them for each individual gene. Using the Freeze Pane functionality in excel works very nicely here.



References

1. Overbeek, R., et al., *The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).* Nucleic Acids Res, 2014. **42**(Database issue): p. D206-14.
2. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25**(17): p. 3389-402.
3. Krzywinski, M., et al., *Circos: an information aesthetic for comparative genomics.* Genome Res, 2009. **19**(9): p. 1639-45.