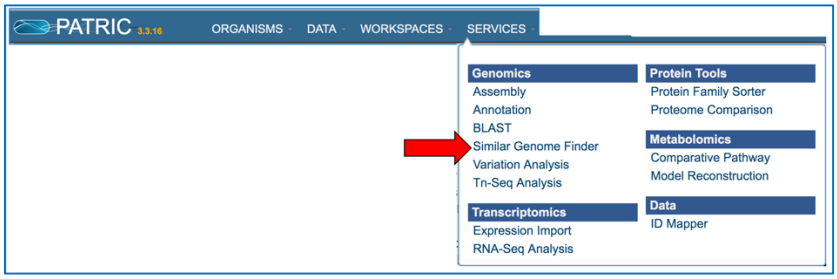
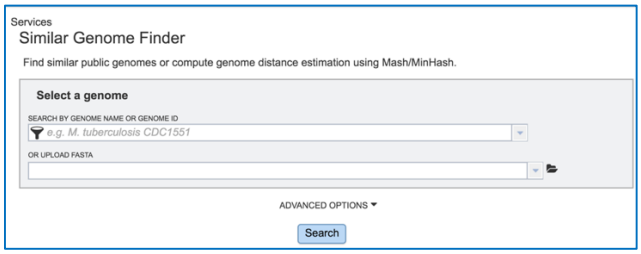
Finding close relatives to a select genome

When a researcher has a new genome sequence, one of the first things they want to identify is the closest relatives of their genome. PATRIC provides a new service that allows researchers to do this using Mash/MinHash[1]. Mash reduces large sequences and sequence sets to small, representative sketches, from which global mutation distances can be rapidly estimated. The MinHash dimensionality-reduction technique to include a pairwise mutation distance and P value significance test, enabling the efficient clustering and search of massive sequence collections.

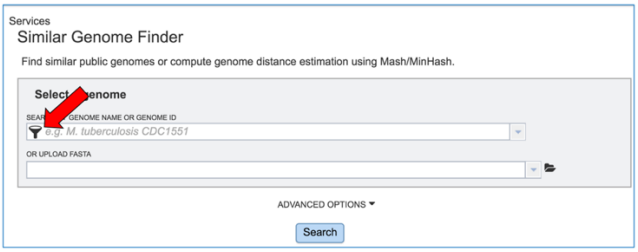
1. At the top of any PATRIC page, find the Services tab. Click on Similar Genome Finder



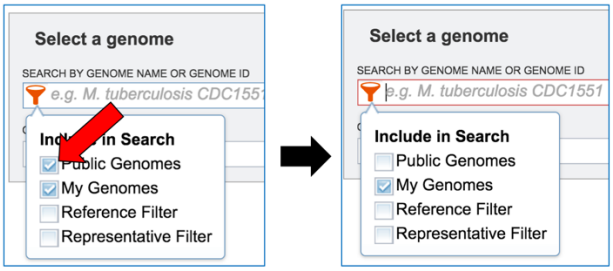
2. This will open up the Similar Genome Finder landing page.



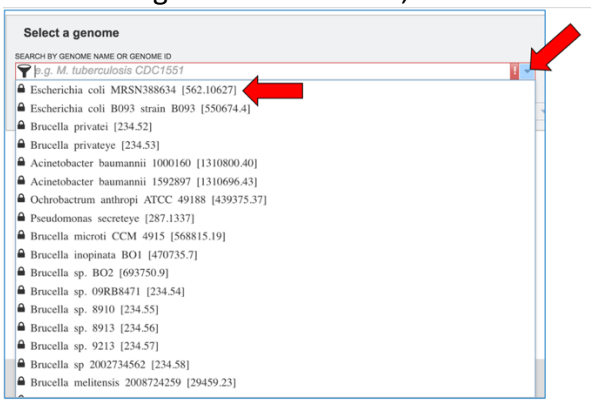
3. To load a genome that has been privately annotated in PATRIC, click on the filter icon that is at the left side of the text box under Search By Genome Name Or Genome ID.



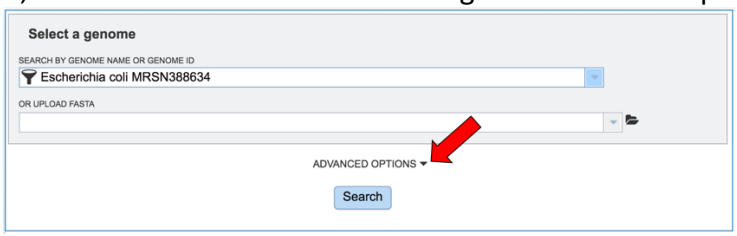
4. This will open a box that allows a researcher to search across all of the public genomes available in PATRIC, or across the genomes that they have annotated and that are stored in their private workspace. To select private genomes, deselect the Public Genomes box. This will leave the Public Genomes box selected.



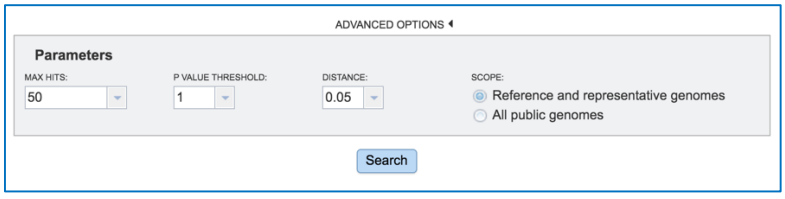
5. Click the down arrow at the right of the text box under Search By Genome Name Or Genome ID. This will open a drop-down box that shows all of the researcher's private genomes. Scroll down and find the genome of interest, and then click on it.



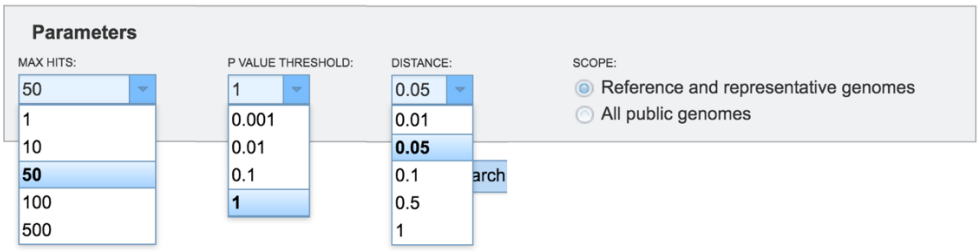
6. This will autofill the name of the genome in the text box. To see options that can be adjusted, click on the down arrow to the right of Advanced Options.



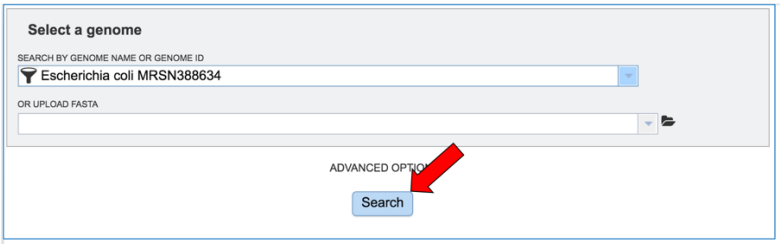
7. This will open a separate box that shows the parameters.



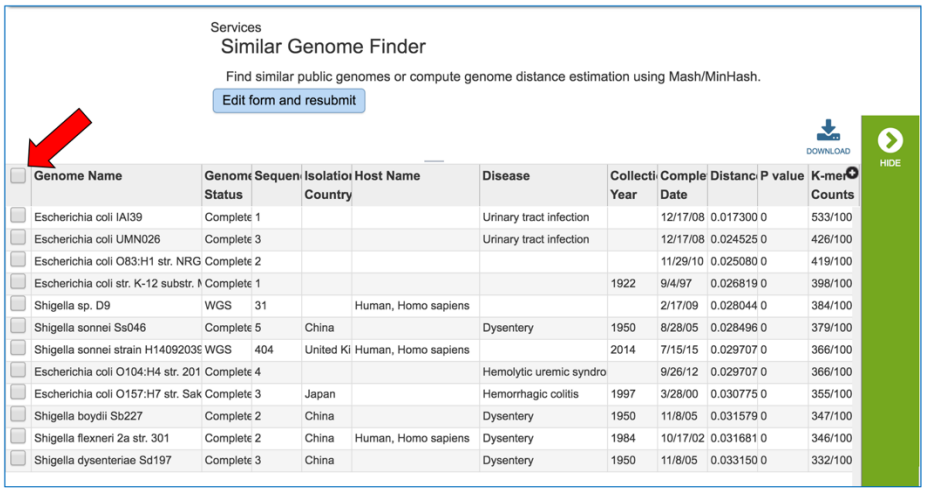
8. Researchers can adjust the number of hits that they want to see, adjust the P Value threshold, set the distance, and filter the search to compare against all the reference or representative genomes, or all the public genomes.



9. Once the parameters of interest have been selected, click the Search button at the bottom of the page.

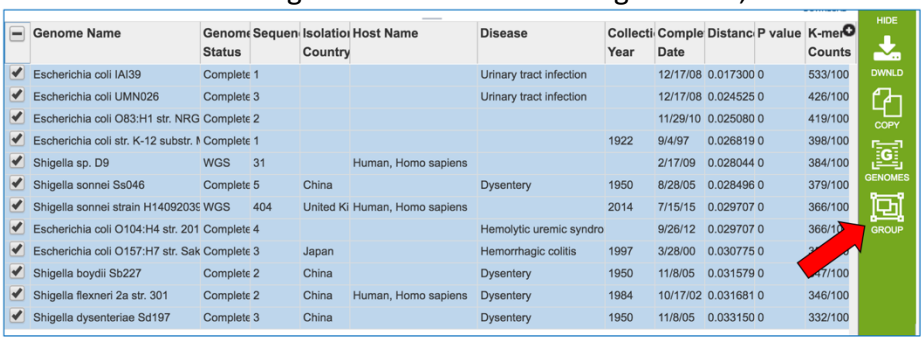


7. The tool will return the top hits to the selected genome. To select all those genomes and create them into a group for downstream analyses, click the check box to the right of the Genome Name column.



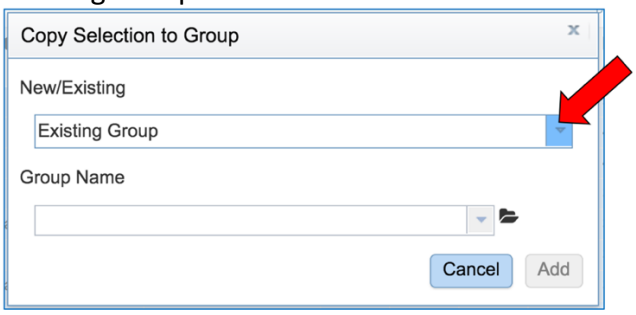
Genome Name	Genome Status	Sequen	Isolation Country	Host Name	Disease	Collecti Year	Comple Date	Distanc	P value	K-mer Counts
Escherichia coli IAI39	Complete	1			Urinary tract infection		12/17/08	0.017300	0	533/100
Escherichia coli UMN026	Complete	3			Urinary tract infection		12/17/08	0.024525	0	426/100
Escherichia coli O83:H1 str. NRG	Complete	2					11/29/10	0.025080	0	419/100
Escherichia coli str. K-12 substr. M	Complete	1				1922	9/4/97	0.026819	0	398/100
Shigella sp. D9	WGS	31		Human, Homo sapiens			2/17/09	0.028044	0	384/100
Shigella sonnei Ss046	Complete	5	China		Dysentery	1950	8/28/05	0.028496	0	379/100
Shigella sonnei strain H14092039	WGS	404	United Ki	Human, Homo sapiens		2014	7/15/15	0.029707	0	366/100
Escherichia coli O104:H4 str. 201	Complete	4			Hemolytic uremic syndro		9/26/12	0.029707	0	366/100
Escherichia coli O157:H7 str. Sak	Complete	3	Japan		Hemorrhagic colitis	1997	3/28/00	0.030775	0	355/100
Shigella boydii Sb227	Complete	2	China		Dysentery	1950	11/8/05	0.031579	0	347/100
Shigella flexneri 2a str. 301	Complete	2	China	Human, Homo sapiens	Dysentery	1984	10/17/02	0.031681	0	346/100
Shigella dysenteriae Sd197	Complete	3	China		Dysentery	1950	11/8/05	0.033150	0	332/100

8. This will auto select all the genomes. In the vertical green bar, click on the Group icon.

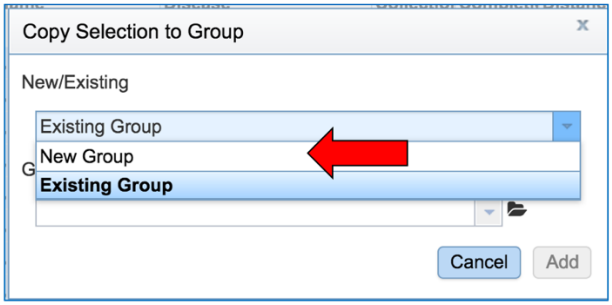


Genome Name	Genome Status	Sequen	Isolation Country	Host Name	Disease	Collecti Year	Comple Date	Distanc	P value	K-mer Counts
Escherichia coli IAI39	Complete	1			Urinary tract infection		12/17/08	0.017300	0	533/100
Escherichia coli UMN026	Complete	3			Urinary tract infection		12/17/08	0.024525	0	426/100
Escherichia coli O83:H1 str. NRG	Complete	2					11/29/10	0.025080	0	419/100
Escherichia coli str. K-12 substr. M	Complete	1				1922	9/4/97	0.026819	0	398/100
Shigella sp. D9	WGS	31		Human, Homo sapiens			2/17/09	0.028044	0	384/100
Shigella sonnei Ss046	Complete	5	China		Dysentery	1950	8/28/05	0.028496	0	379/100
Shigella sonnei strain H14092039	WGS	404	United Ki	Human, Homo sapiens		2014	7/15/15	0.029707	0	366/100
Escherichia coli O104:H4 str. 201	Complete	4			Hemolytic uremic syndro		9/26/12	0.029707	0	366/100
Escherichia coli O157:H7 str. Sak	Complete	3	Japan		Hemorrhagic colitis	1997	3/28/00	0.030775	0	355/100
Shigella boydii Sb227	Complete	2	China		Dysentery	1950	11/8/05	0.031579	0	347/100
Shigella flexneri 2a str. 301	Complete	2	China	Human, Homo sapiens	Dysentery	1984	10/17/02	0.031681	0	346/100
Shigella dysenteriae Sd197	Complete	3	China		Dysentery	1950	11/8/05	0.033150	0	332/100

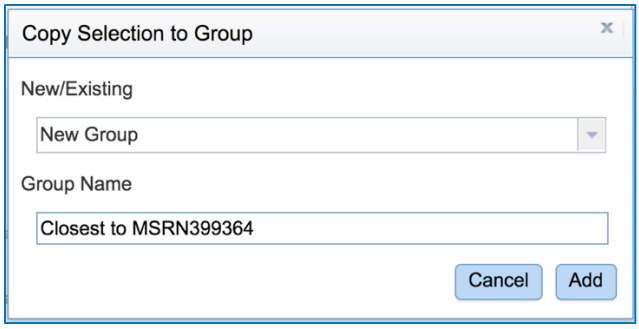
9. This will open a pop-up box. To save these into a new group, click on the down arrow that follows Existing Group.



10. Click on New Group.



11. Give a name to the group by entering it in the text box under Group Name. Save the group by clicking the Add button at the bottom of the box. The group you created will now be saved in your workspace.



References

1. Ondov BD, Treangen TJ, Melsted P et al. Mash: fast genome and metagenome distance estimation using MinHash, Genome biology 2016;17:132.