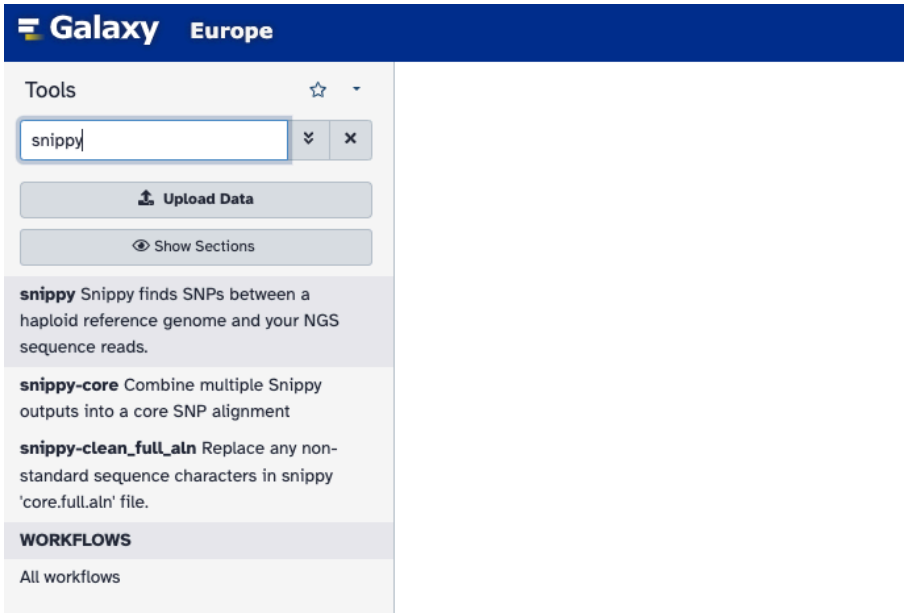
Performing read alignments using snippy

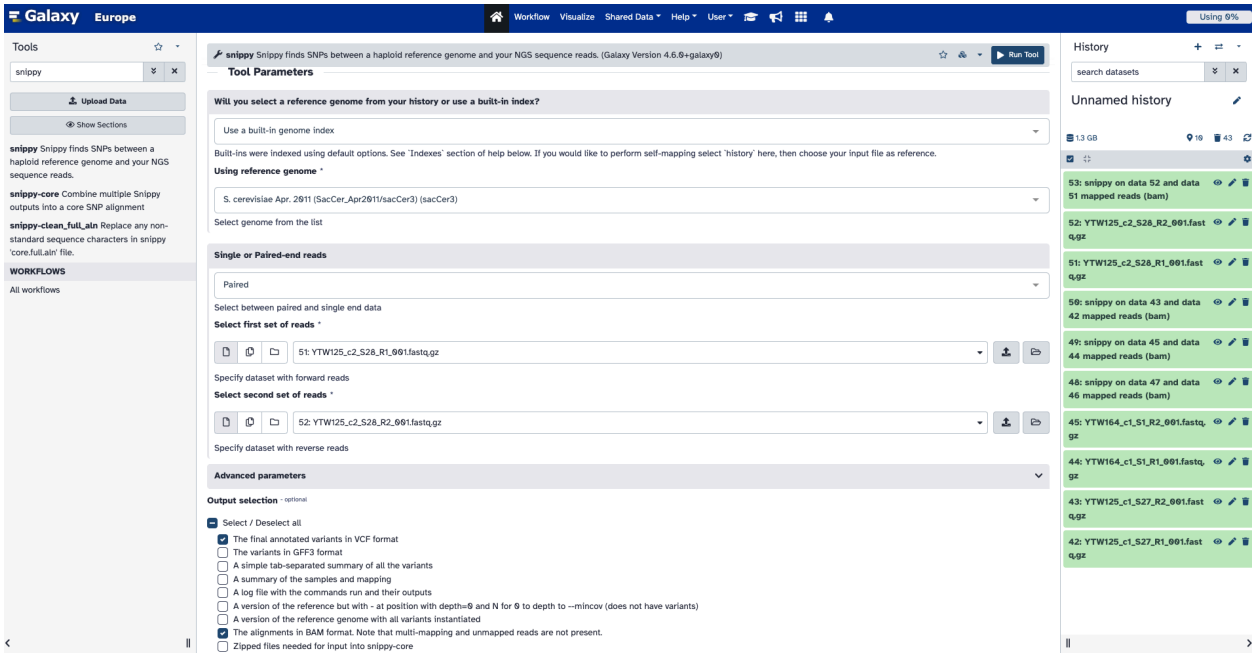
Snippy is a tool for performing alignments of sequencing reads to a reference genome. For more information, it has a github and tutorial.

Snippy is not on usegalaxy.org so you need to sign up for an account on usegalaxy.eu

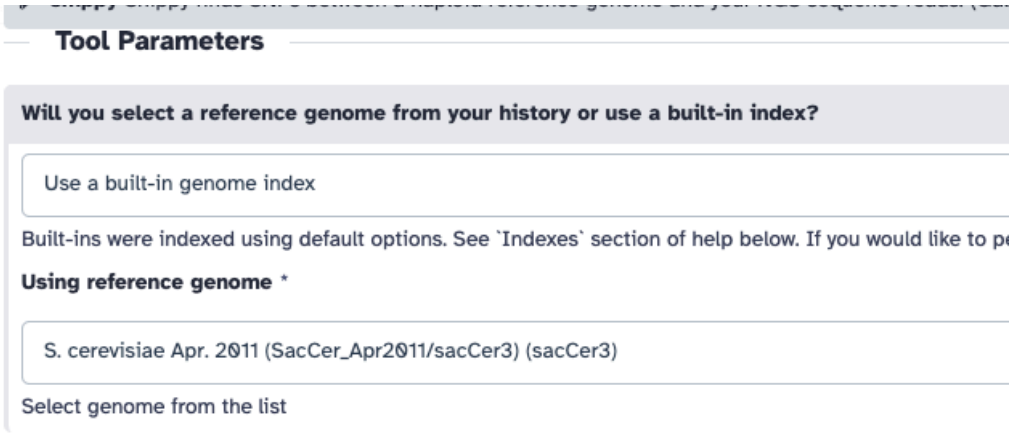
Search for “snippy” in the tools bar (left side) and click on “snippy”



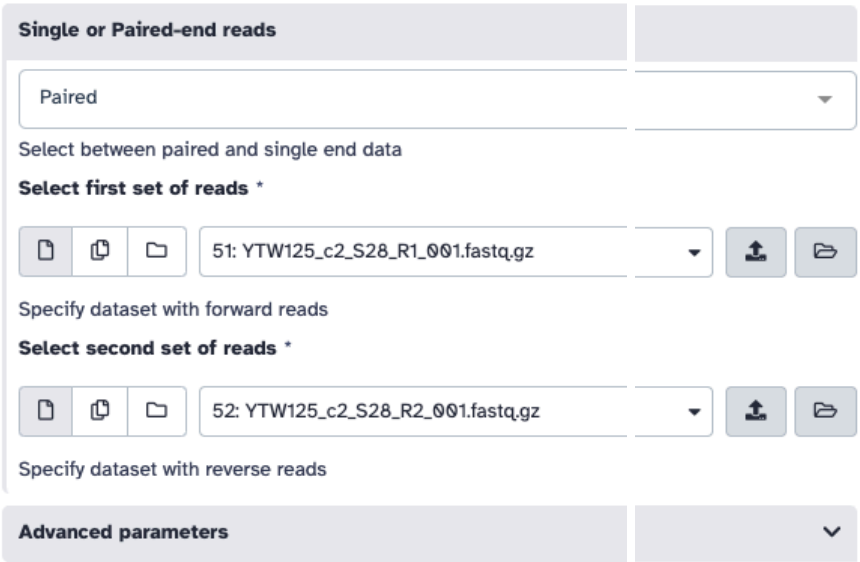
Now the center of the window will populate with the tool parameters.



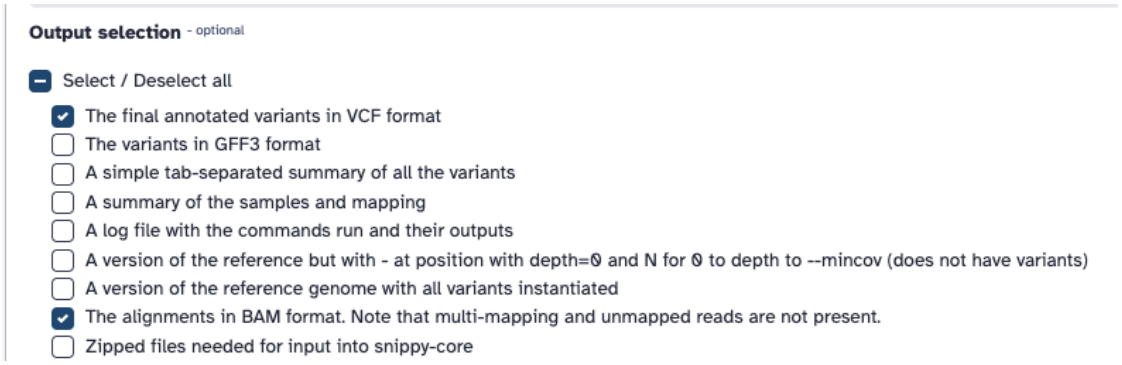
- Will you select a reference genome from your history or use a built-in index?
    - Leave as “use a built-in genome index”
    - Change the reference genome to sacCer3 (you can type to search)



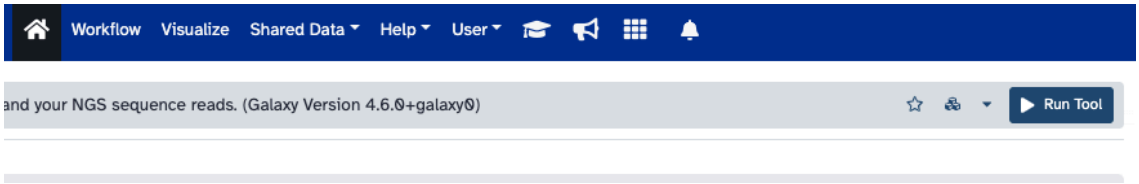
- Single or Paired-end reads
    - Select “paired”
    - Decide if you want to run multiple alignments at once with the same parameters, or just alignment. Click on the corresponding icon under “Select the first set of reads” to have a single file, multiple files, or a folder of files.
        - If you need to upload files, that button is to the right of the box (upload dataset or browse already uploaded datasets)
        - If you are running multiple alignments, make sure that the file names match up to the “R1” or “R2” so that it can match them.
    - Do the same for the “Select the second set of reads” to select the pair(s).



- Advanced parameters
    - You can set mapping quality etc. here, but leaving as defaults is fine for yEvo
- Output selection
    - Choose the file(s) you want it to create.
    - For yEvo, you want minimally “The final annotated variants in VCF format” and “The alignments in BAM format. Note that multi-mapping and unmapped reads are not present.”



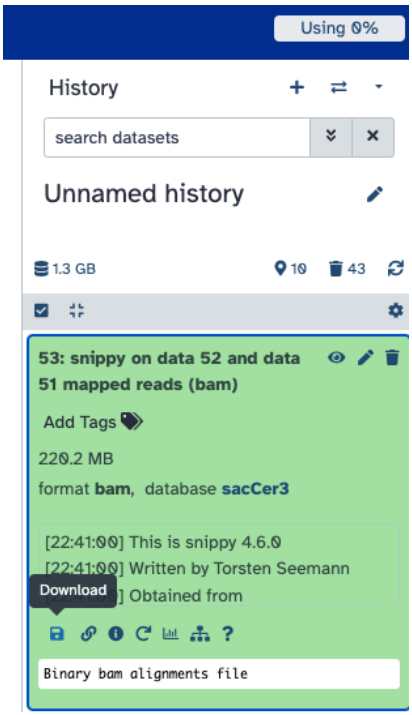
- Additional options
    - You can have it email you when it’s done
- Click “Run tool” (at the bottom of the parameters, or the top right of the central window).



Your alignment(s) will now start running.

- You can see them in the right side panel. If they are waiting to run, they will be yellow. If they are finished, they will be green. If there is an error, it will turn red and stop running that job, but if you are running multiple, the others will continue running.
- Make sure to write down the job number that is running on each file! The output files will be named by number, so you want to keep track of what number refers to what file (e.g. if uploading Read 1 file was “job 51” and uploading Read 2 file was “job 52”, then the output file will be named as “snippy on file 51 and 52,” so you need to know what those files are!! This will remain in your history unless you clear it, so you can go back and find out.)

- When an alignment is done, it will turn green. You can click on it to expand it, and then click the little floppy disk icon to download the file. I recommend immediately renaming it something that is more meaningful than the automatic galaxy naming.
    - You will need to download each file individually, the bam and the vcf.



Determining which mutations are new in the evolved clone

For comparing an evolved clone against the ancestor yeast strain, you will need to run snippy on your evolved clone, and one person in the class will need to run it on the ancestor so that everyone has ancestor files to use for visualizing in IGV (.vcf to look at mutation list, and .bam and .bai to open in IGV). Note that snippy will call any difference between the sequence and the sacCer3 reference genome, not the ancestor strain! So it will call many “mutations” that are not new, and therefore you need to look for mutations that are present in the .vcf for the evolved clone but not the ancestor strain.

Using scripts in languages like R will enable you to filter quickly, but you can also filter for unique mutations in Excel. One way that will identify most unique calls in the evolved strain is to paste ancestor and evolved into the same excel sheet, add a column with the strain name to differentiate them, and then filter for unique genome locations in that column; this would fail if there were different mutations at the same location but should capture most mutations.

Then proceed with your filtered list of mutations to visualize them in IGV.