GDSCN Book: SARS with Galaxy on AnVIL

January 31, 2025

Table of Contents

# Overview

This book provides resources for instructors to engage students in a cloud-based Galaxy activity on AnVIL, focused on SARS-CoV-2 variant detection.

There is a growing need for undergraduate students to learn cutting-edge concepts in genomics data science, including performing analysis on the cloud instead of a personal computer. This lesson aims to introduce a mutant detection bioinformatics pipeline based on a publicly available genetic sample of SARS-CoV-2. Students will be introduced to the sequencing revolution, variants, genetic alignments, and essentials of cloud computing prior to the lab activity. During the lesson, students will work hands-on with the point-and-click Galaxy interface on the [AnVIL](https://anvilproject.org/) cloud computing resource to check data, perform an alignment, and visualize their results.

## Skills Level

*Genetics*  
**Novice**: no genetics knowledge needed

*Programming skills*  
**Novice**: no programming experience needed

## Learning Objectives

Learning objectives for this activity come from the [Genetics Core Competencies](https://genetics-gsa.org/education/genetics-learning-framework/):

* Gather and evaluate experimental evidence, including qualitative and quantitative data
* Generate and interpret graphs displaying experimental results
* Critique large data sets and use bioinformatics to assess genetics data
* Tap into the interdisciplinary nature of science

## GDSCN Collection

This exercise is part of a collection of teaching resources developed through the *Genomic Data Science Community Network* (GDSCN). GDSCN works towards a vision where researchers, educators, and students from diverse backgrounds are able to fully participate in genomic data science research. Learn more about GDSCN by visiting <https://www.gdscn.org/home> or reading the [article in Genome Research](https://doi.org/10.1101/gr.276496.121).

Please check out our full collection of AnVIL and related resources: <https://hutchdatascience.org/AnVIL_Collection/>

# 1 Student Activity Guide

This chapter contains the student instructions for the SARS-CoV-2 Variant Detection with Galaxy activity.

## 1.1 Introduction

Galaxy is a free, relatively easy to use bioinformatics implementation package. It changes command line programs into GUI based programs.

### 1.1.1 Before You Start

Make sure you have an AnVIL account.

In order to run your analyses, you will use the [AnVIL cloud computing platform](https://anvilproject.org/). The AnVIL (Analysis Visualization and Informatics Lab-space) platform is specially designed for analyzing biological data, and is used by scientists doing all sorts of biological research.

**AnVIL in a nutshell**

* Behind the scenes, AnVIL relies on Google Cloud Platform to provide computing infrastructure. Basically, AnVIL lets you “rent” computers through the internet. The analysis is run on the rented computer. AnVIL lets you see the results in your browser.
* AnVIL uses [Terra](https://anvil.terra.bio/) to provide many computational tools useful for biological data analysis, such as [RStudio](https://www.rstudio.com/products/rstudio/), [Galaxy](https://usegalaxy.org/), and [Jupyter Notebooks](https://jupyter.org/). Terra takes care of installing these tools so you can use them right away.

### 1.1.2 Create Google Account

First, you will need to set up a (free) Google account.

A Google account usually looks like “myname@gmail.com”. Alternatively, you can enable Google for an existing non-Gmail email address using [these instructions](https://support.terra.bio/hc/en-us/articles/360029186611).

If you do not already have a Google account that you would like to use for accessing AnVIL, [create one now](https://accounts.google.com/SignUp).

### 1.1.3 Log In to Terra

Next, make sure you can log in to Terra – you will use Terra to perform computations on AnVIL.

You can access Terra by going to [anvil.terra.bio](https://anvil.terra.bio/). Open Terra, and you should be prompted to sign in with your Google account.

### 1.1.4 Share Username

Make sure your instructor has your Google account username (e.g. myname@gmail.com), so they can give you access to everything you need.

* Make sure there are no typos!
* If you have multiple Google accounts, make sure you give them the username that you will be using to log in to anvil.terra.bio.

It is *very important* that you share the Google account you will be using to access AnVIL with with your instructor! Otherwise, the instructor cannot add you to Billing Projects or Workspaces, and you will be unable to proceed with your assignments.

### 1.1.5 Objectives

This activity will teach you how to use the AnVIL platform to:

1. Get started working on AnVIL
2. Launch the Galaxy tool
3. Import data into the Galaxy tool
4. Examine fastq sequence data files
5. Align sequence data to a reference genome
6. View the aligned data and reference genomes interactively

## 1.2 Getting Started

### 1.2.1 Set Up

In the next few steps, you will walk through how to get set up to use Galaxy on the AnVIL platform. AnVIL is centered around different “Workspaces”. Each Workspace functions almost like a mini code laboratory - it is a place where data can be examined, stored, and analyzed. The first thing we want to do is to copy or “clone” a Workspace to create a space for you to experiment.

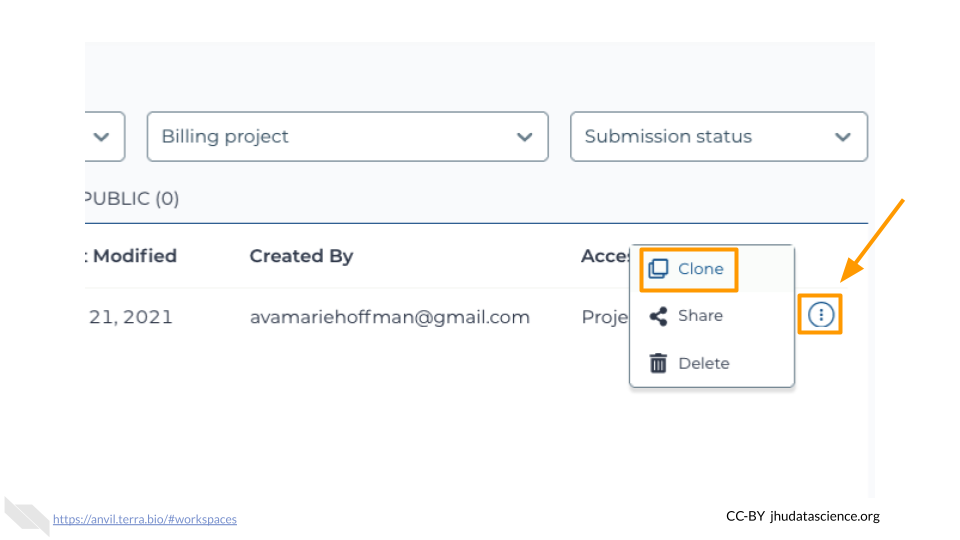
Use a web browser to go to the AnVIL website. In the browser type:

anvil.terra.bio

**Tip** At this point, it might make things easier to open up a new window in your browser and split your screen. That way, you can follow along with this guide on one side and execute the steps on the other.

After logging in, click “View Workspaces”. Select the “Public” tab. In the top search bar type the activity workspace “SARS-CoV-2-Genome”. You can also go directly to the following link: <https://anvil.terra.bio/#workspaces/gdscn-exercises/SARS-CoV-2-Genome>.

Clone the workspace by clicking the teardrop button (). And selecting “Clone”.



In the first box, give your Workspace clone a name by adding an underscore (“\_”) and your name. For example, “SARS-CoV-2-Genome\_Ava\_Hoffman”. Next, select the Billing project provided by your instructor. Leave the bottom two boxes as-is and click “CLONE WORKSPACE”.

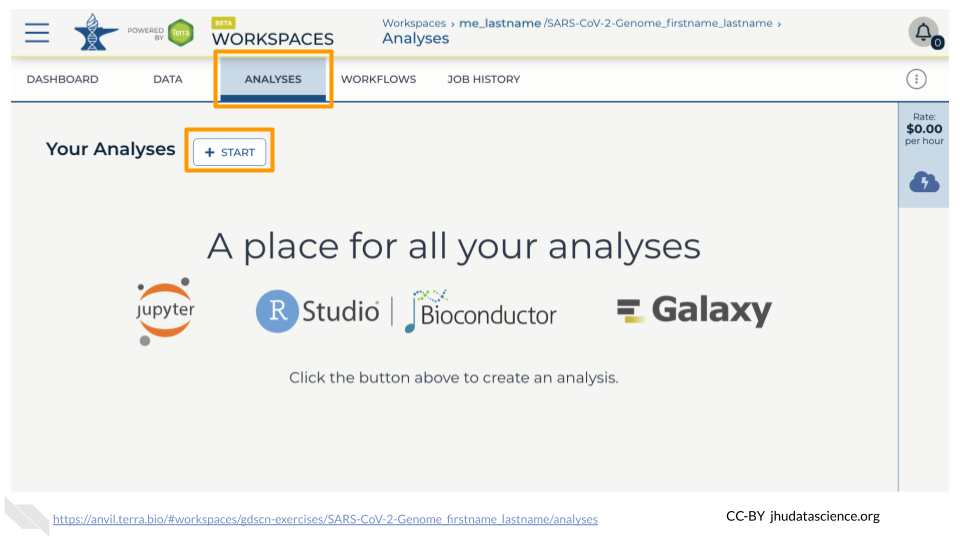


### 1.2.2 Starting Galaxy

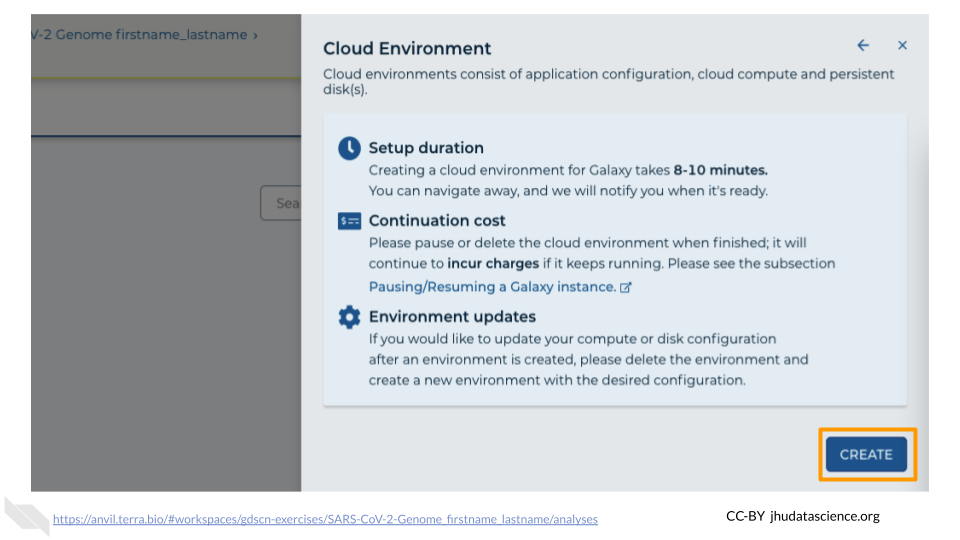
Galaxy is a great tool for performing bioinformatics analysis without having to update software or worry too much about coding. In order to use Galaxy, we need to create a cloud environment. This is like quickly renting a few computers from Google as the engine to power our Galaxy analysis.

Currently, you will need to use Chrome or Safari as your browser for Galaxy cloud environments to work.

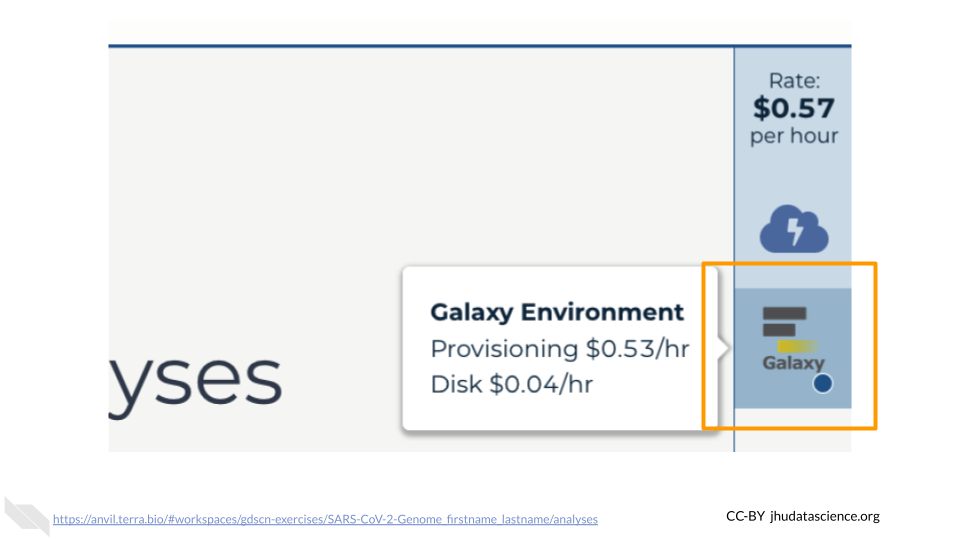
In your new Workspace, click on the “ANALYSES” tab. Next, click on “START”. You should see a popup window on the right side of the screen. Click on the Galaxy logo to proceed.



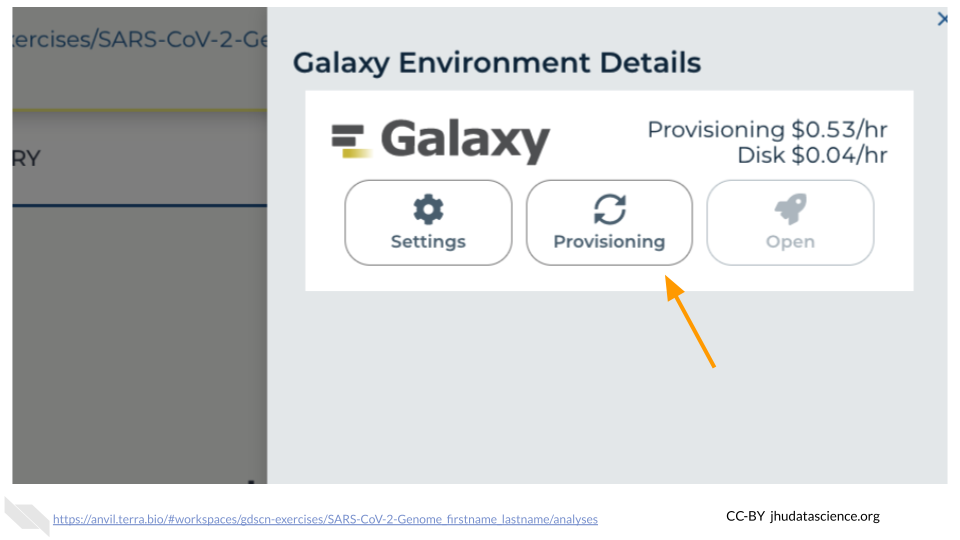
Click on “NEXT” and “CREATE” to keep all settings as-is.



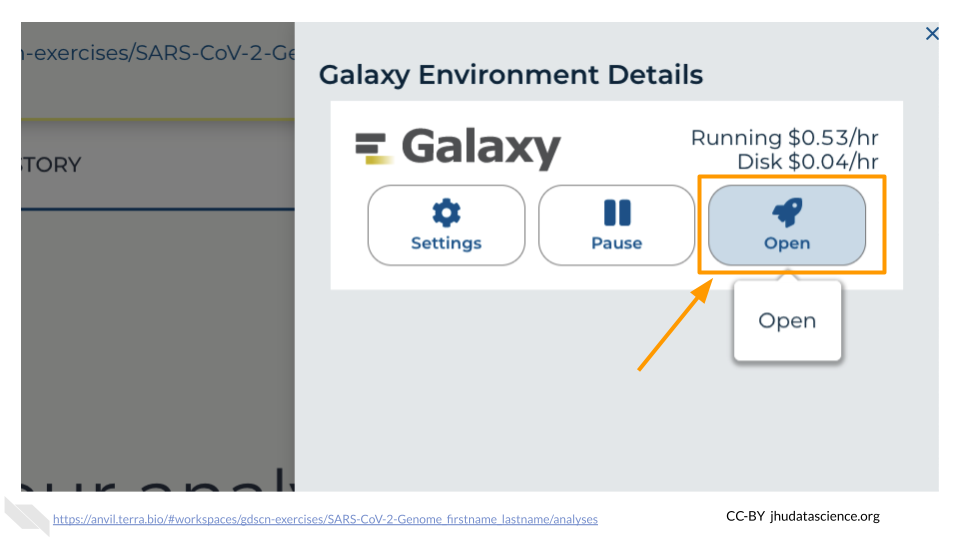
Click on the Galaxy icon.



You will see that the environment is still being set up.



This will take 8-10 minutes. When it is done, click “Open”. You might need to refresh the page.



Remember that you can refresh your browser or navigate away at any time. This is because the connection to the environment is in the cloud, not on your personal computer.

You can also follow along with the first ~2 minutes of [this video](https://jhudatascience.org/AnVIL_Book_Getting_Started/starting-galaxy.html) to start Galaxy on AnVIL.

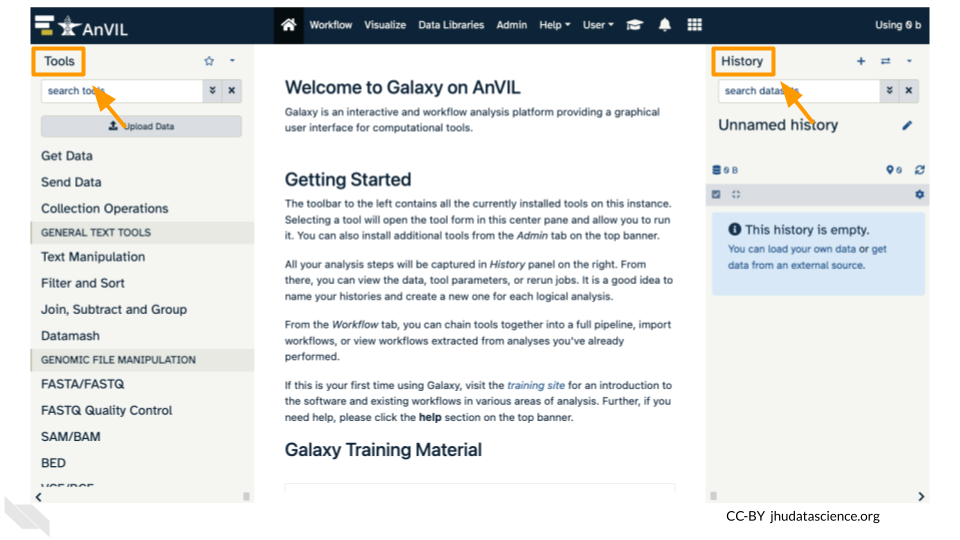
### 1.2.3 Navigating Galaxy

Notice the three main sections.

**Tools** - These are all of the bioinformatics tool packages available for you to use.

**The Main Dashboard** - This contains flash messages and posts when you first open Galaxy, but when we are using data this is the main interface area.

**History** - When you start a project you will be able to see all of the documents in the project in the history. Now be aware, this can become very busy. Also the naming that Galaxy uses is not very intuitive, so you must make sure that you label your files with something that makes sense to you.

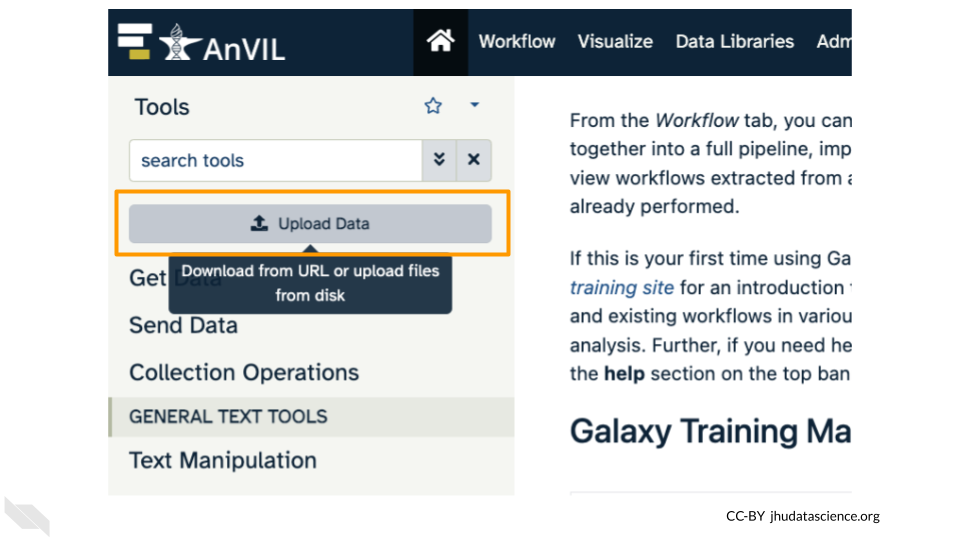


On the welcome page, there are links to tutorials. You may try these out on your own. If you want to try a new analysis this is a good place to start.

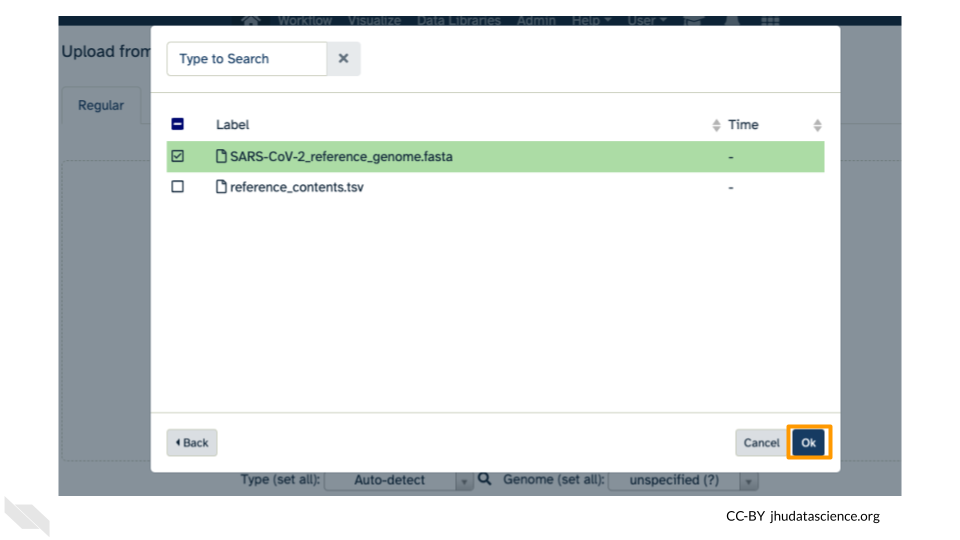
## 1.3 Exercise One: Importing Data into Galaxy

Luckily, we linked to the original data when we cloned our Workspace! We have three files we will need for our activity. These are (1) the reference genome for SARS-CoV-2, and both forward (2) and reverse (3) reads for our sample. There are two sets of reads for our sample because the scientists who collected it used paired-end sequencing. The reference genome ends in “.fasta” because it has already been cleaned up by scientists. The sample we are looking at ends in fastq because it is raw data from the sequencer.

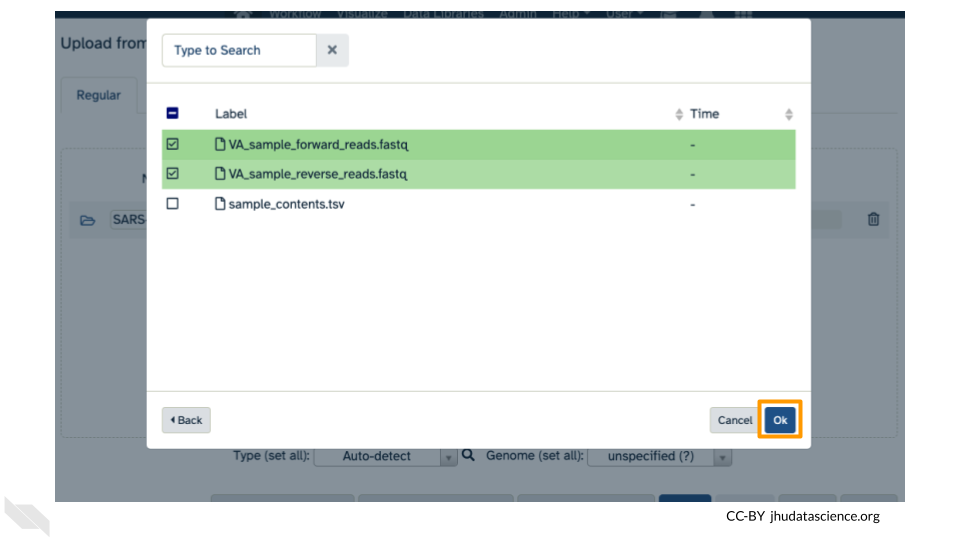
1. Click on “Upload Data” in the Tools pane.

* 

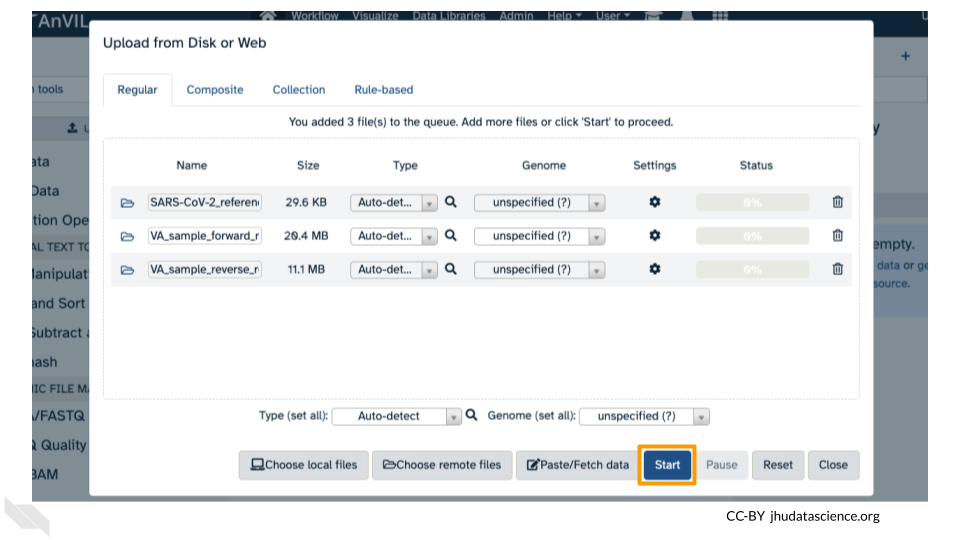
1. Click on “Choose remote files” at the bottom of the popup. Double-click the Workspace folder, then “Tables/” then “reference/”. Click the reference .fasta file so that it is highlighted in green and click “Ok”.

* 

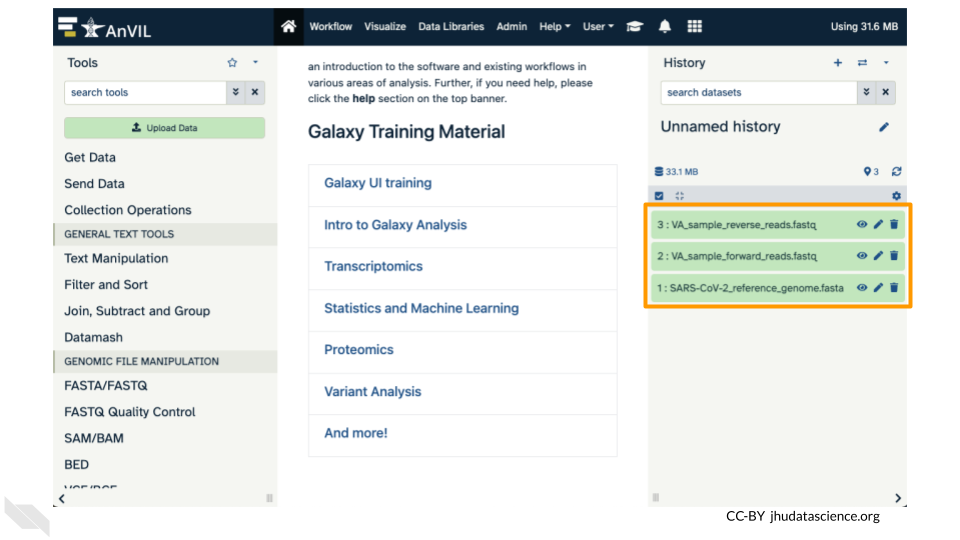
1. Now that your reference has been added, click “Choose remote files” again to add the two sample files. Double-click the Workspace folder, then “Tables/” then “samples/”. Click the two sample fastq files so that they are highlighted in green and click “OK”.

* 

1. Click “Start” and once complete, you can click “Close”.

* 

1. Confirm your upload worked by looking at the file names in the History pane.

* 

## 1.4 Exercise Two: Examining fastq Files in Galaxy

*Now we have some data in our account we can look at it. In this exercise we will see data in fastq format. This is the typical output from an Illumina Sequencer, but also the standard format for most alignment software.*

### 1.4.1 Examining Inputs

Use your mouse and click on the eye icon () of the first fastq file (**VA\_sample\_forward\_reads.fastq**). In the Main screen you will see something like this:



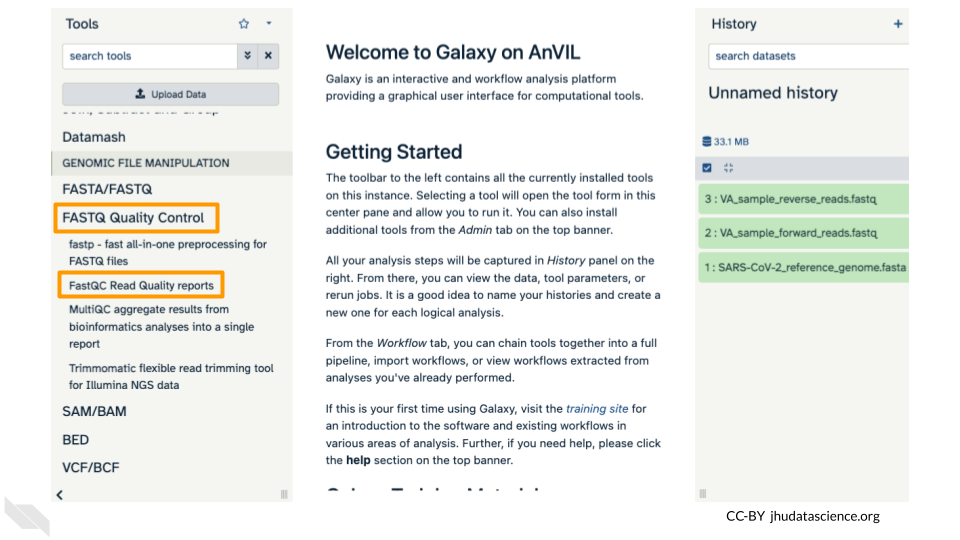
QUESTIONS:

1. How many lines in a .fastq file represent an individual read?
2. What does each line represent?
3. Why is the final line for each read (the quality score) important?

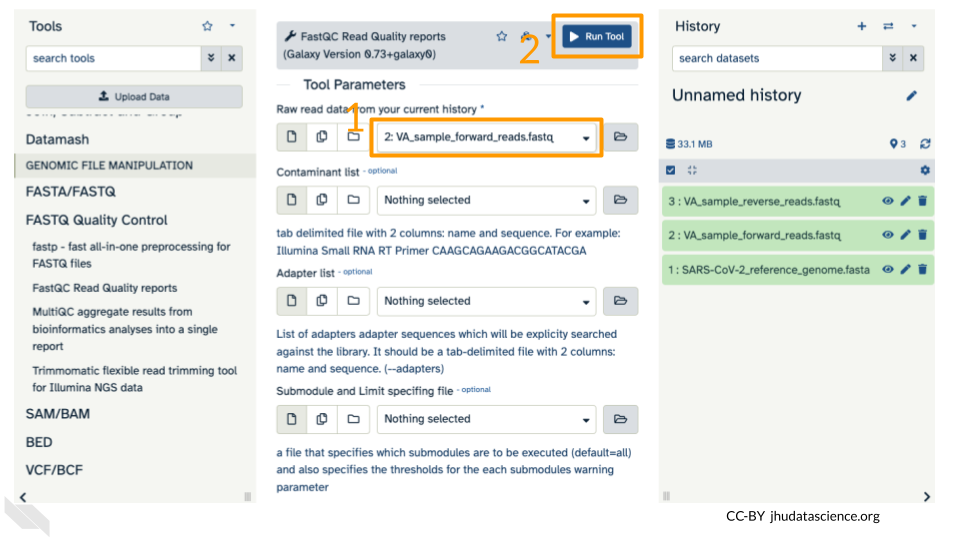
### 1.4.2 Quality Scoring

FastQC is a tool which aims to provide simple quality control checks on raw sequence data coming from high throughput sequencing pipelines. It provides a set of analyses which you can use to get a quick impression of whether your data has any problems of which you should be aware before doing any further analysis.

Find the FastQC tool in the GENOMIC FILE MANIPULATION: FASTQ Quality Control tool folder. You will see something like this in the tools:

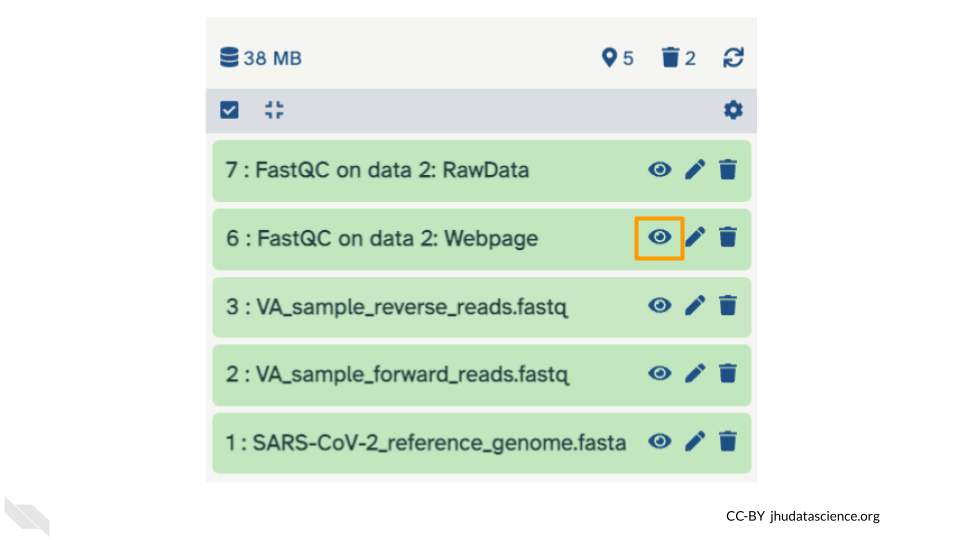


In the first drop down menu, make sure it has your first fastq file (**VA\_sample\_forward\_reads.fastq**) loaded. Leave everything else as-is and click on the blue “Run Tool” button.

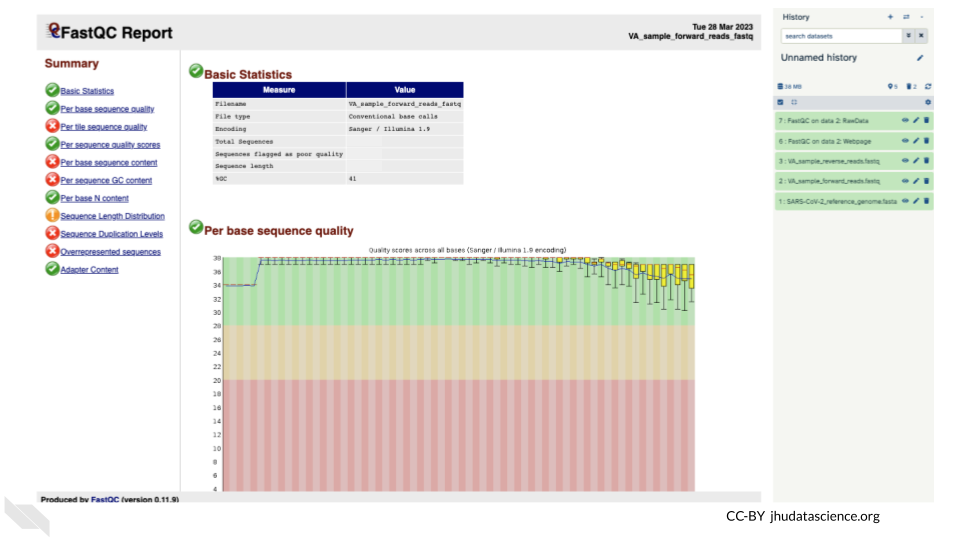


The main dash will highlight in green if everything is okay. In the history, you will see the new files turn yellow, then green. If the job fails it will show an error.

Click on the eye icon () in the new file in the history “FASTQC on data2 Webpage”.



You will open up a summary report for the sequencing file:



QUESTIONS:

1. Explore “Basic Statistics”. How many total reads are there? Have any been flagged as poor quality? What is the sequence length?
2. Explore “Per base sequence quality”. Based on the Basic Statistics, is 28-40 a good or bad quality score?
3. Is it okay to proceed based on the per base sequence quality?

**Breakout Box: Learn more about quality scores**

You may be wondering how the fourth line of the .fastq files relates to the quality score above. To save space, the sequencer records an [ASCII character](http://drive5.com/usearch/manual/quality_score.html) to represent scores 0-42. For example 10 corresponds to “+” and 40 corresponds to “I”. FastQC knows how to translate this. This is often called “Phred” scoring.

What does 0-42 represent? These numbers, when plugged into a formula, tell us the probability of an error for that base. This is the formula, where Q is our quality score (0-42) and P is the probability of an error:

Q = -10 log10(P)

Using this formula, we can calculate that a quality score of 40 means only 0.00010 probability of an error!

## 1.5 Exercise Three: Alignment

*Given that our data has passed some quality checks, we will try to align the data to the reference genome. In this case it is simple, a viral genome. A human sequencing project will generate much larger data sets. There are many aligners, but we will start off looking at a simple aligner BWA-MEM. This example uses paired data.*

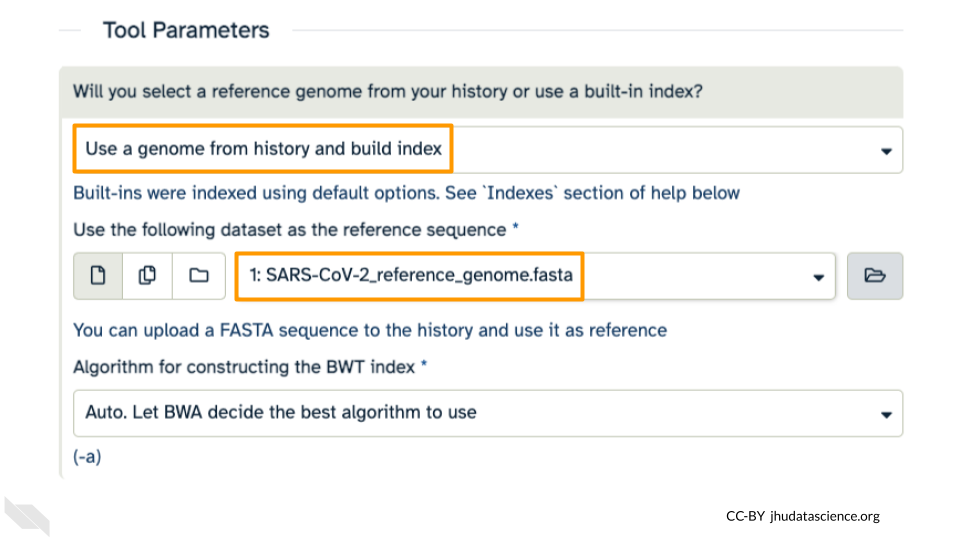
We will use our two SARS data files, which are ready for alignment.

* **VA\_sample\_forward\_reads.fastq**
* **VA\_sample\_reverse\_reads.fastq**

Go to GENOMICS ANALYSIS and expand the Mapping menu. Select “Map with BWA-MEM”. This program will align your reads to your SARS reference genome. Some of our reads are >100 base pairs so we will use the MEM option.



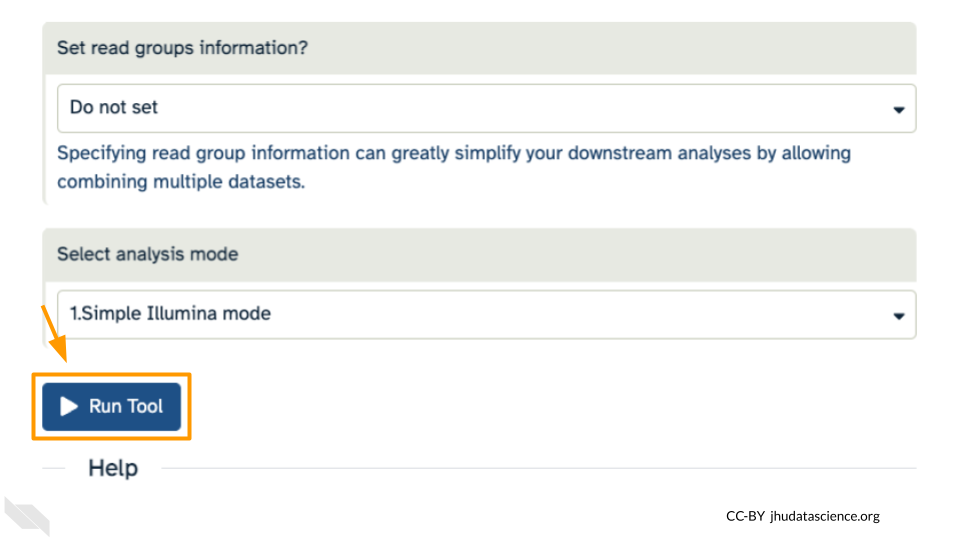
First, choose your reference. In the first drop down box change it to “Use a genome from history and build index”. Then choose the SARS reference fasta file that you uploaded as the reference earlier.



Under the “Single or Paired-end reads” ensure the “Paired” option is selected. Now choose your forward and reverse fastq files.



Leave other options as-is. *You can learn more about what the alignment software BWA-MEM is doing if you scroll down below the Run Tool button*. Click “Run Tool”.



The output file is a BAM file, which lists where each read aligns to the reference genome and whether there are any differences. You can click the eye button to preview the results, but the results are not easy to interpret visually (much like the fastq files). Instead you will use a genome viewer in the next step.

QUESTIONS:

1. What is alignment software (for example, BWA-MEM) actually doing?
2. Here we are using paired fastq (“paired end”) data. What is an advantage of using paired data?

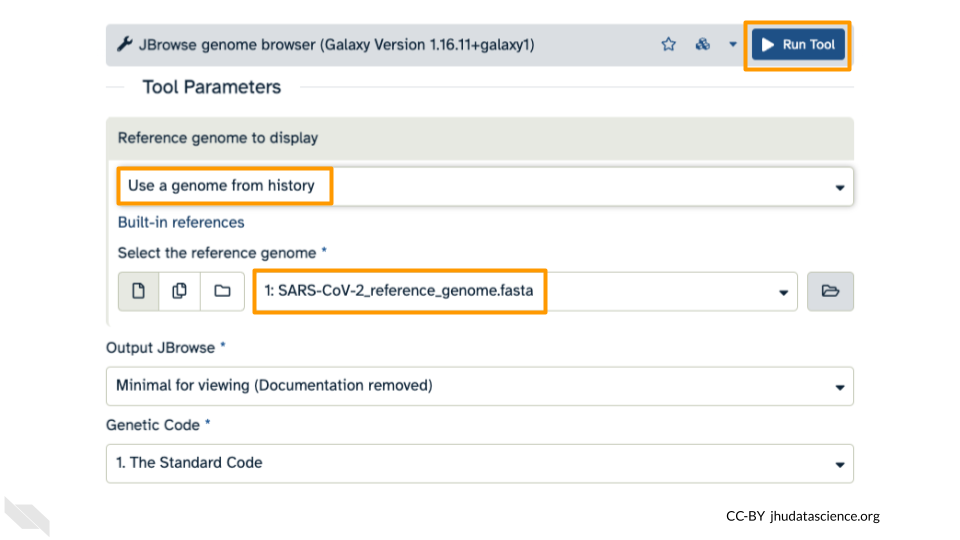
## 1.6 Exercise Four: Viewing aligned data

We have aligned our data but it is currently a table of where the reads align. This is hard to read, so we will use [JBrowse](https://jbrowse.org/jb2/) to view the data.

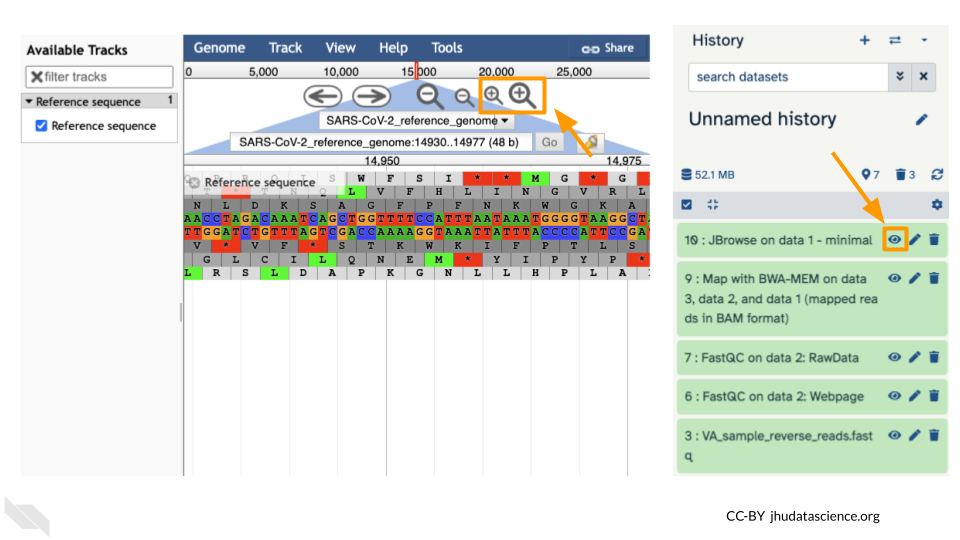
Scroll down in the Tools menu to STATISTICS AND VISUALIZATION. Under “Graph/Display Data”, select “JBrowse”.



Under “Reference genome to display” ensure that “Use a genome from history” is selected. Below this, make sure that the **SARS-CoV-2\_reference\_genome.fasta** file is selected. Click “Run Tool”.

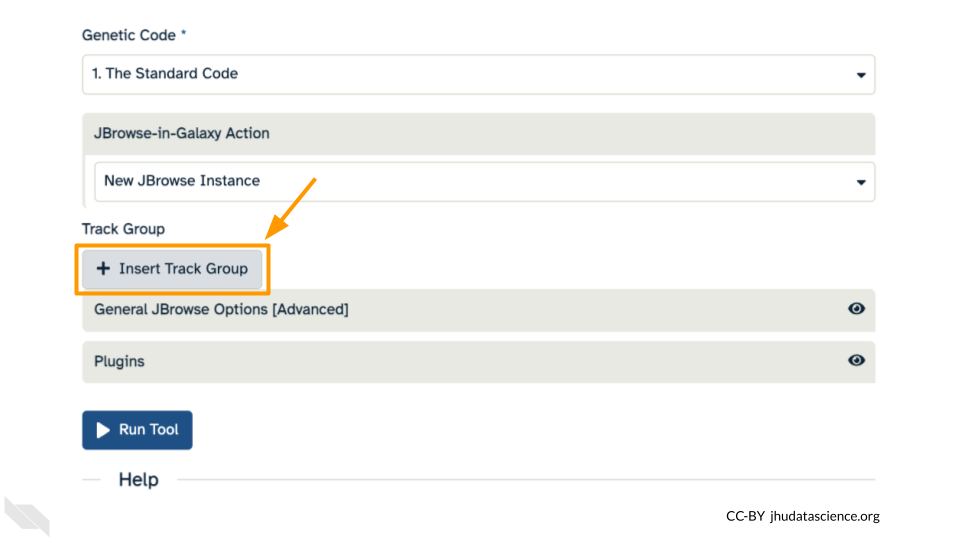


You should see a new JBrowse item appear in your history. Click the eye icon () to open JBrowse. You will need to click on the magnifying glasses to zoom in, but you should see the A,C,G, and Ts and their corresponding colors that make up the SARS-CoV-2 genome!

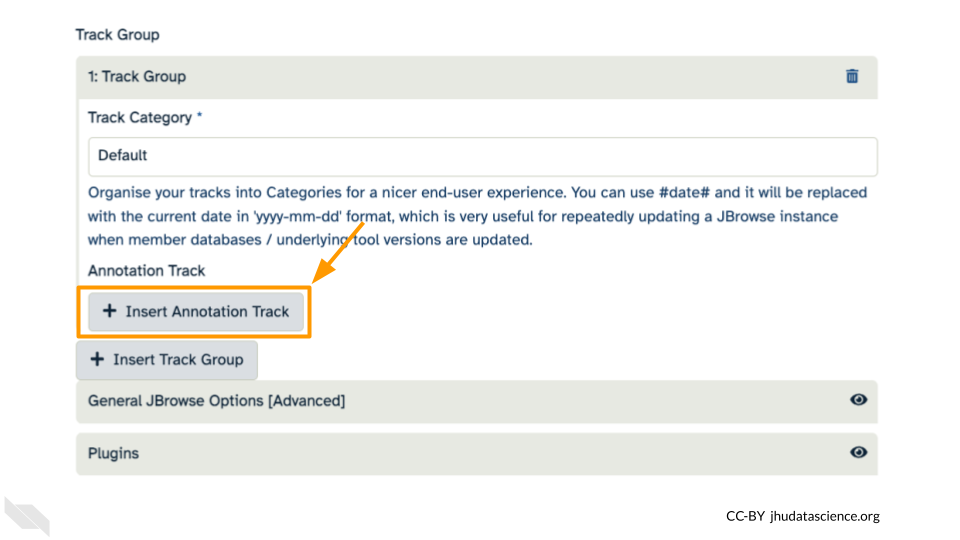


This is interesting, but it doesn’t let us compare the genome to the sample we have. We suspect there may be some differences that indicate our sample is the delta variant. Go back to STATISTICS AND VISUALIZATION: Graph/Display Data and select “JBrowse”. Just like before, under “Reference genome to display”, ensure that “Use a genome from history” is selected. Below this, make sure that the **SARS-CoV-2\_reference\_genome.fasta** file is selected.

This time, we’ll add our alignment data from [Exercise Three](#alignment). Click the “+ Insert Track Group” button.



Click “+ Insert Annotation Track” to add our alignment data.



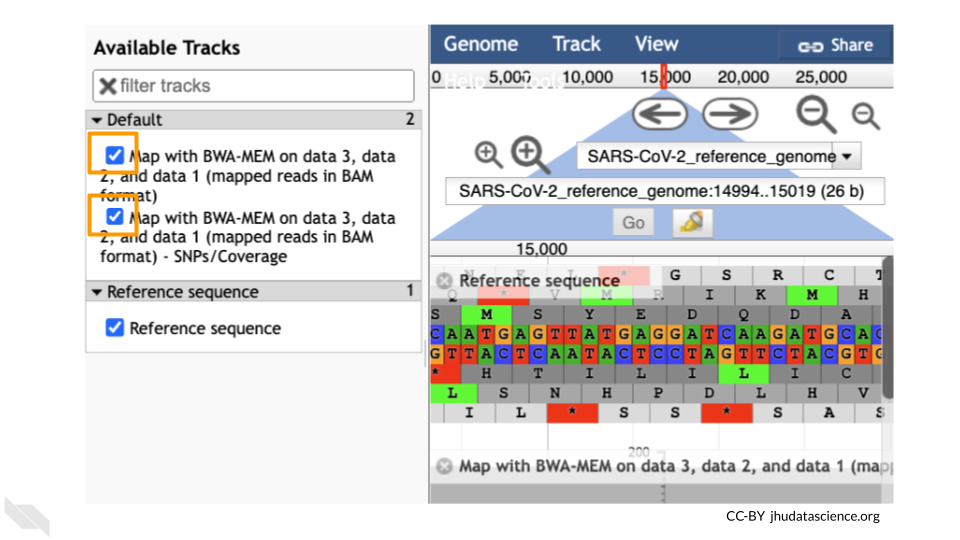
You’ll keep everything the same except the following:

* **Track Type**: BAM Pileups
* **Autogenerate SNP Track**: Yes
* **Maximum size of BAM chunks**: Add one more zero: 50000000

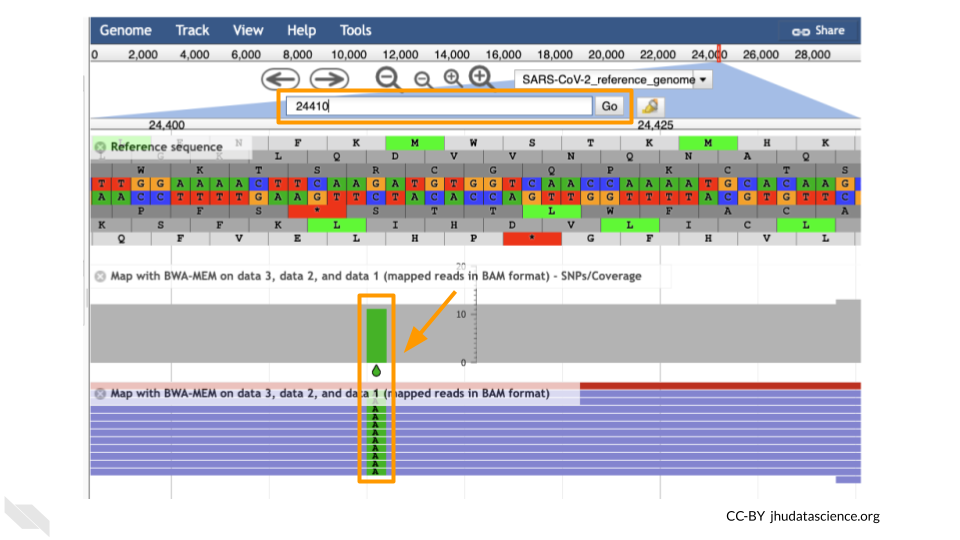


These arguments tell JBrowse what kind of data we are using, as well as some memory options. Click “Run Tool”.

You should see a new JBrowse item appear in your history. Click the eye icon () to open JBrowse. Make sure that all boxes are checked on the left side: “Available Tracks”. The tracks will show up in the order that you click on them.



Let’s look at an example mutation in our sample. Type in the reference position “24410” and click “Go”. You should see a bunch of “A”s highlighted in green throughout our sample. The reference sequence (top line) is a “G” but all of the reads are an “A”. This means that our sample is genetically different from the established SARS-CoV-2 reference genome! Researchers often call these single base differences “SNPs” - Single Nucleotide Polymorphisms.

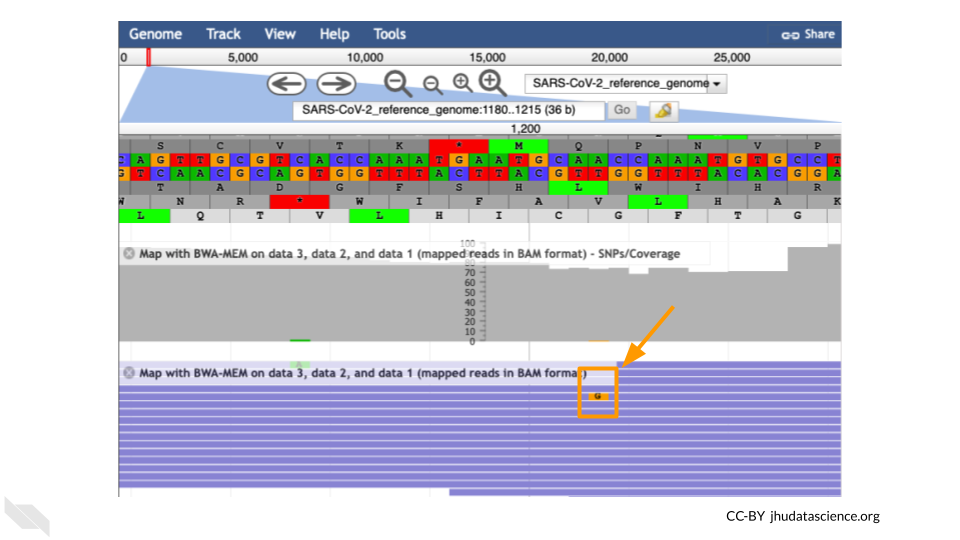


QUESTIONS:

1. How long is the SARS-CoV-2 genome? Hint: zoom out and scroll to the end of the genome.
2. Locate position 23,603. This is the site of an important mutation in the spike protein of the delta variant “P681R”. In this mutation, the amino acid proline is replaced by arginine. Is this mutation present at position 23,603 in our sample? Based on the evidence, do you think this sample is a delta variant?

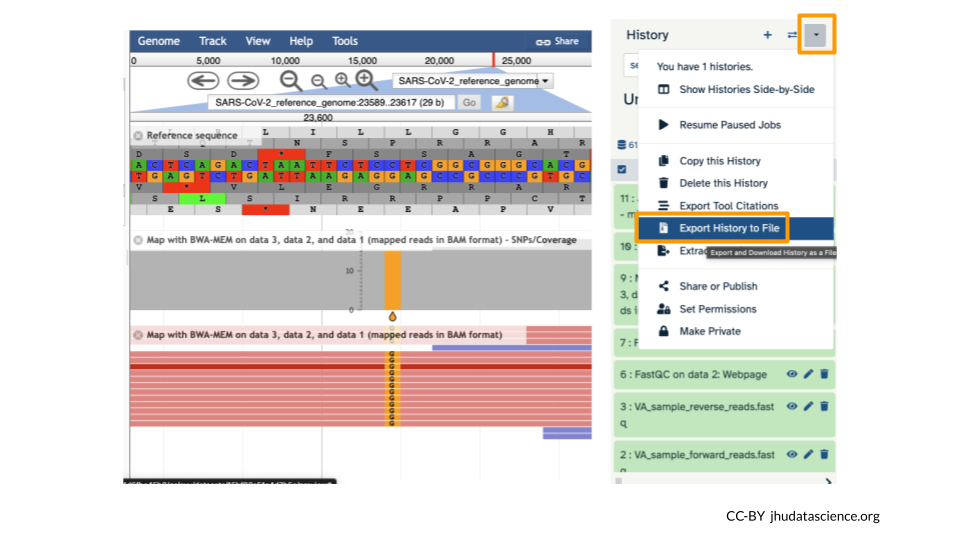
**Breakout Box: Sequencing errors**

It’s possible to make mistakes in the data preparation before we get to the data analysis. Sometimes this happens when the samples are being prepared in the lab and sometimes this happens because the sequencer makes a mistake. This is one reason why quality scores are helpful. With millions of reads of data, it’s more likely that we see a “SNP” that is actually an accident. Multiple copies of the same areas of our data (“read depth”) help us be sure it’s a real SNP. When we compare across lots of aligned reads of the same area, we can determine the actual sequence by consensus. For example, we can be reasonably confident that the “G” at position 1,203 shown below is just a sequencing or lab mistake.

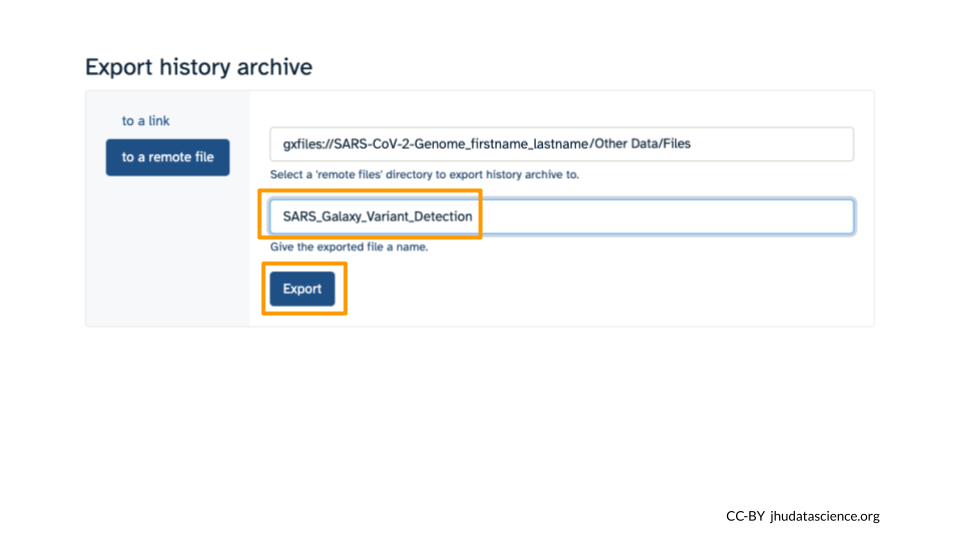


## 1.7 Export Your History

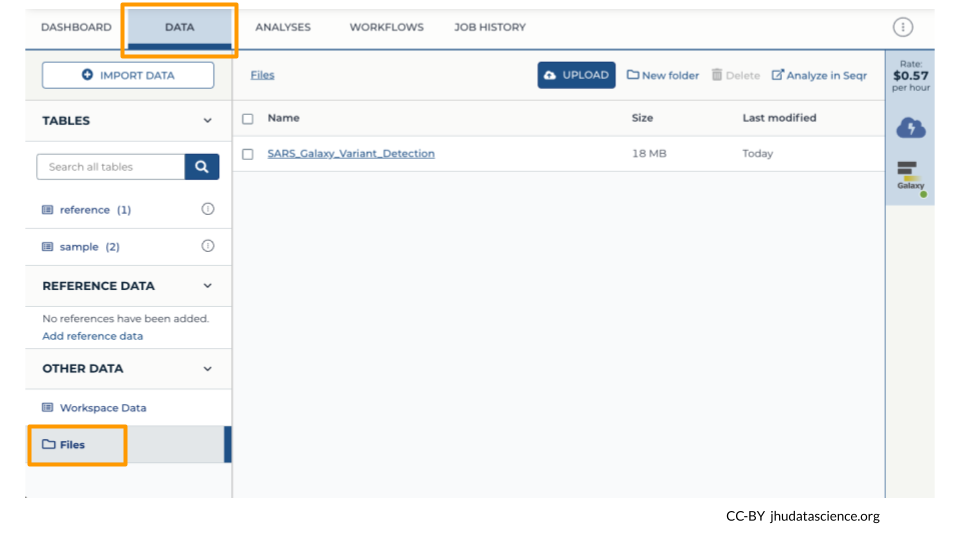
It’s a good idea to export your “History” so that your collaborators can see what you did. Click on the History Menu and click on “Export History to File”.



Make sure you select “to a remote file”. Then, click to select where to export your History. On the popup menu, select your Workspace name, then select “Other Data”. Finally, select “Files”. Then click “Select this folder”. Make sure the export directory looks correct. Next, name your history “SARS Galaxy Variant Detection” and click “Export”.



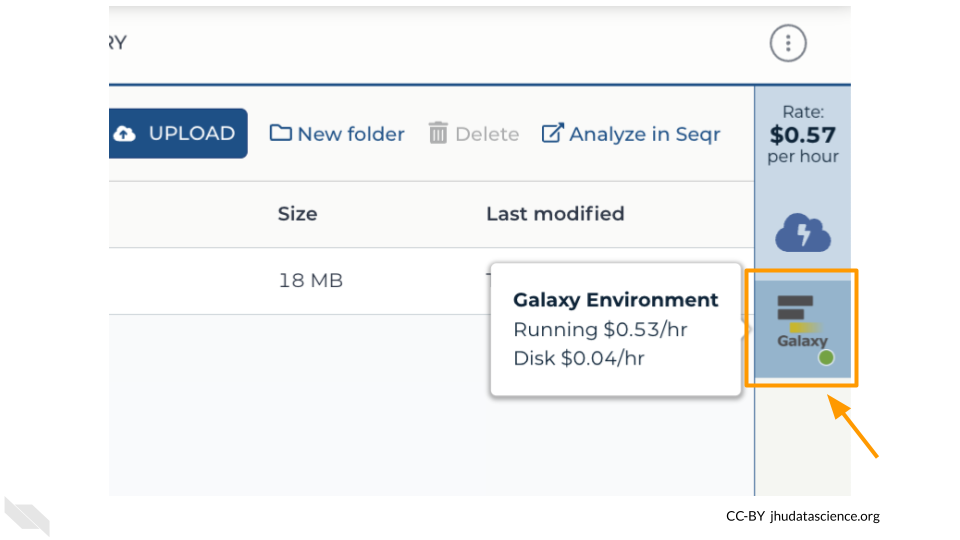
Back at your Workspace, click on the “Data” tab, and the Files folder. You should now see the History export in your files. If you click on the file, you can download it or view it in Google Cloud Storage Browser.



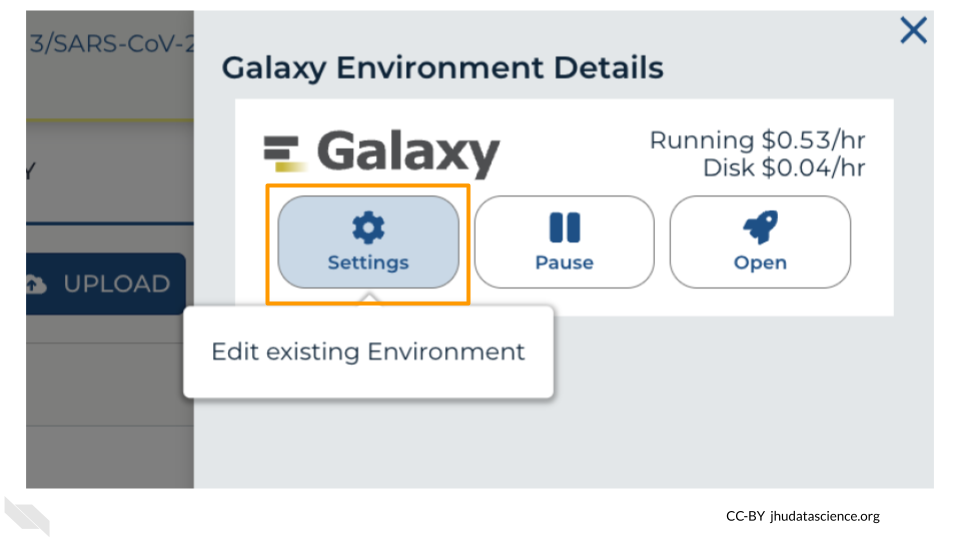
## 1.8 Wrap-up

Once you are done with the activity, you’ll need to shut down your Galaxy cloud environment. This frees up the cloud resources for others and minimizes computing cost. The following steps will delete your work, so make sure you are completely finished at this point. Otherwise, you will have to repeat your work from the previous steps.

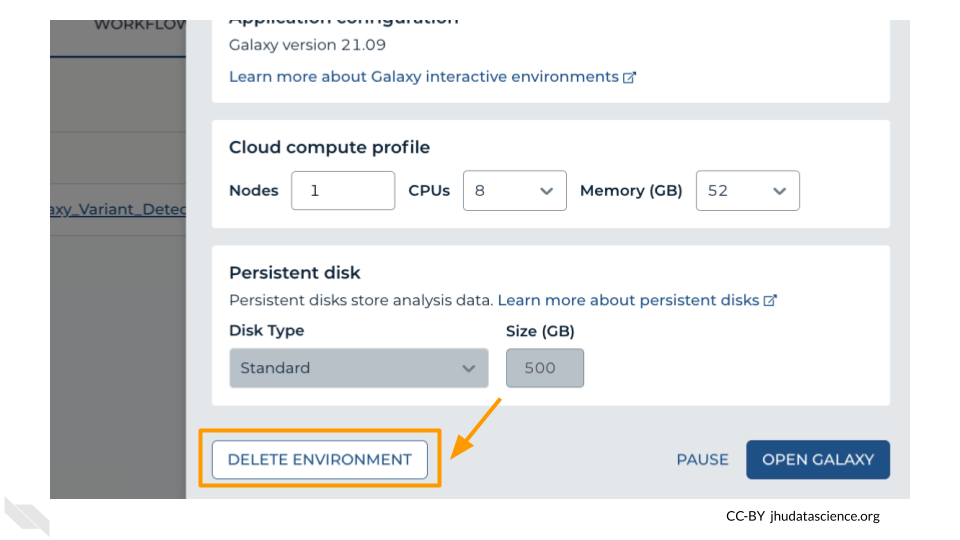
Return to AnVIL, and find the Galaxy logo that shows your cloud environment is running. Click on this logo:



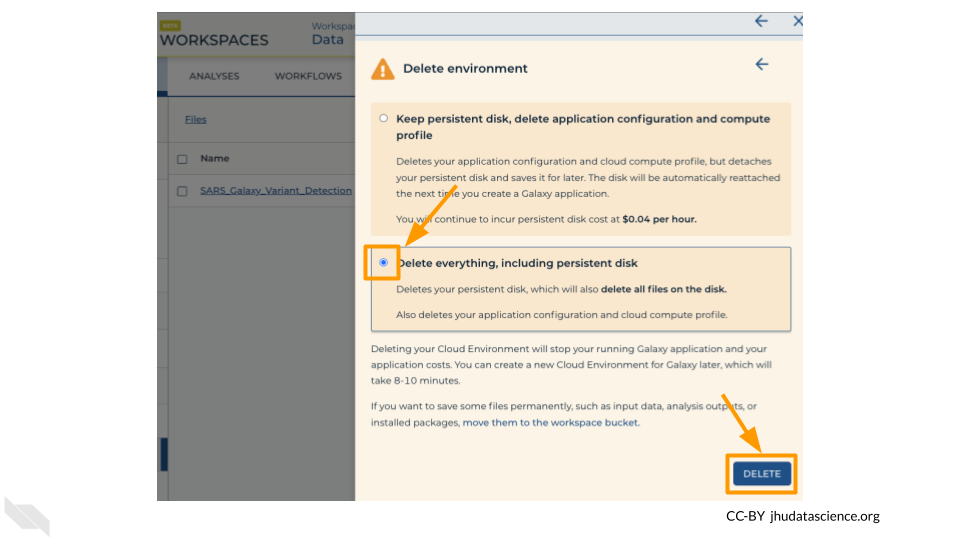
Click “Settings”.



Next, scroll down and click on “DELETE ENVIRONMENT”:



Finally, select “Delete everything, including persistent disk”. Make sure you are done with the activity and then click “DELETE”.



# References