BioDIGS: Getting Started

October 16, 2024

Table of Contents

# About this Book

This is a companion training guide for BioDIGS, a GDSCN project that brings a research experience into the classroom. Visit the BioDIGS (BioDiversity and Informatics for Genomics Scholars) website [here](https://biodigs.org/) for more information about this collaborative, distributed research project, including how you can get involved!

The GDSCN (Genomics Data Science Community Network) is a consortium of educators who aim to create a world where researchers, educators, and students from diverse backgrounds are able to fully participate in genomic data science research. You can find more information about its mission and initiatives [here](https://www.gdscn.org/home).



## 0.1 Target Audience

The activities in this guide are written for undergraduate students and beginning graduate students. Some sections require basic understanding of the R programming language, which is indicated at the beginning of the chapter.

## 0.2 Platform

The activities in this guide are demonstrated on NHGRI’s [AnVIL](https://anvilproject.org/) cloud computing platform. AnVIL is the preferred computing platform for the GDSCN. However, all of these activities can be done using your personal installation of R or using the online [Galaxy](usegalaxy.org) portal.

## 0.3 Data

The data generated by the BioDIGS project is available through the [BioDIGS website](biodigs.org), as well as through an [AnVIL workspace](https://anvilproject.org).

Data about the soil itself as well as soil metal content was generated by the [Delaware Soil Testing Program](https://www.udel.edu/canr/cooperative-extension/environmental-stewardship/soil-testing/) at the University of Delaware. Sequences were generated by the [Johns Hopkins University Genetic Resources Core Facility](https://grcf.jhmi.edu/) and by [PacBio](https://www.pacb.com/).

# 1 AnVIL Workspace

You can easily access the data on AnVIL by cloning the dedicated workspace. These sections guide you through creating an AnVIL account and accessing the workspace.

## 1.1 Create Google Account

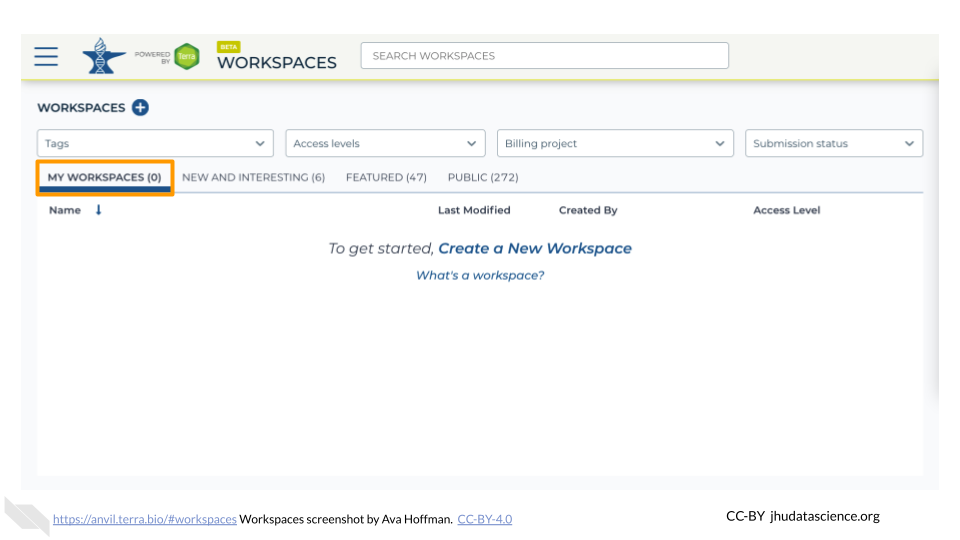
If you do not already have a Google account that you would like to use for accessing Terra, [create one now](https://accounts.google.com/SignUp).

If you would like to create a Google account that is associated with your non-Gmail, institutional email address, follow [these instructions](https://support.terra.bio/hc/en-us/articles/360029186611).

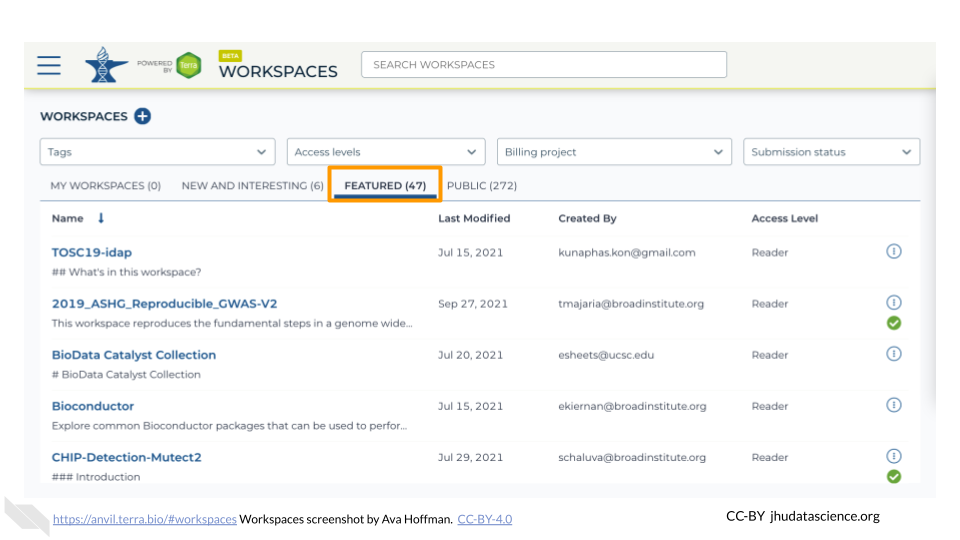
## 1.2 Clone the Workspace

## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output

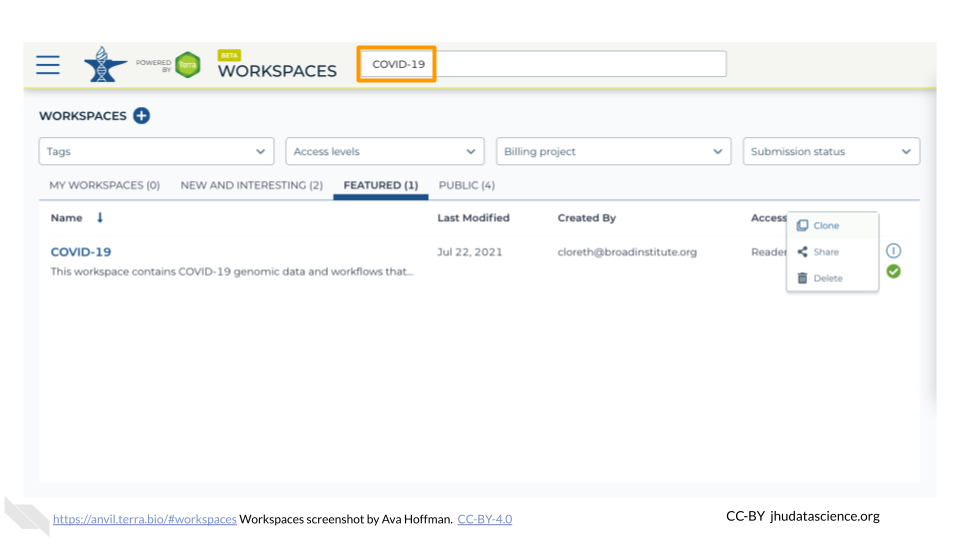
1. [Launch Terra](https://anvil.terra.bio/#workspaces)
2. Locate the Workspace you want to clone. If a Workspace has been shared with you ahead of time, it will appear in “MY WORKSPACES”. You can clone a Workspace that was shared with you to perform your own analyses. In the screenshot below, no Workspaces have been shared.

* 

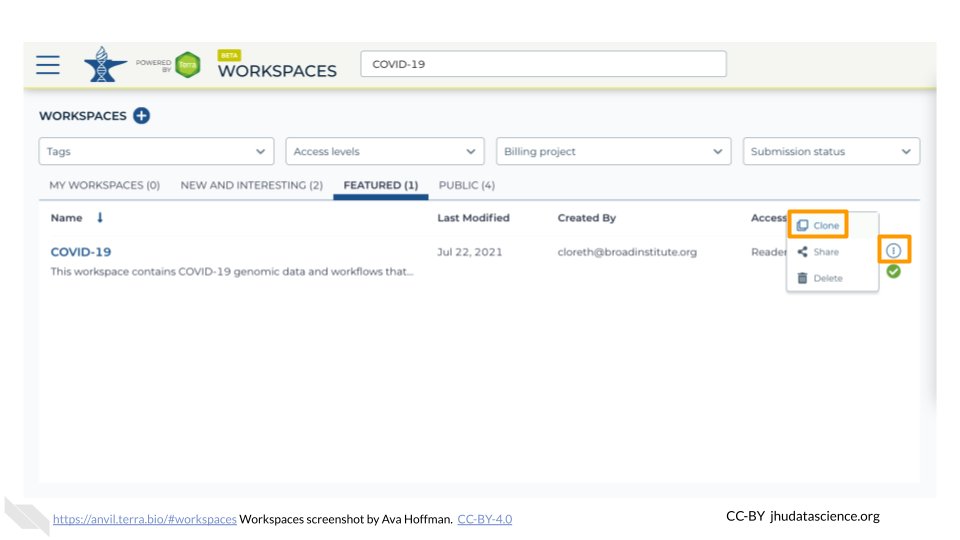
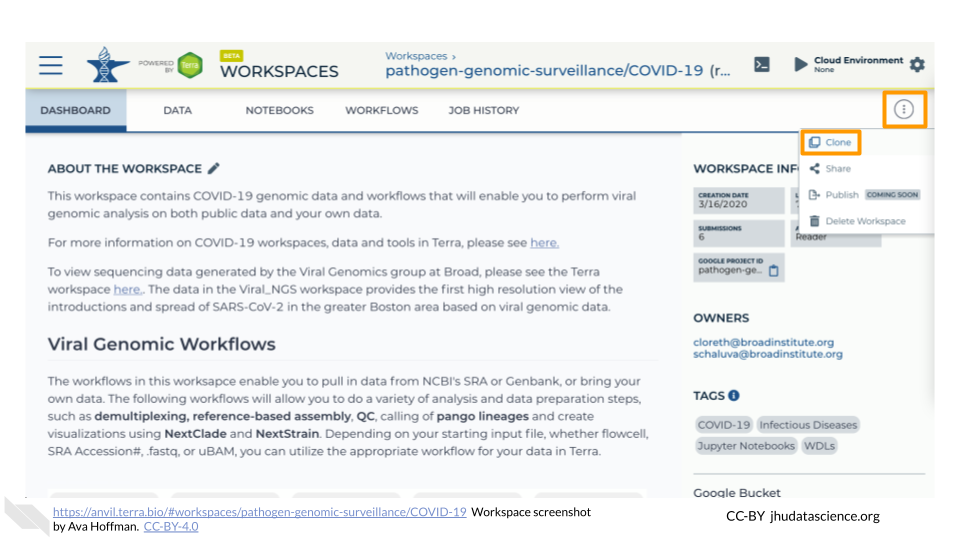
1. If a Workspace hasn’t been shared with you, navigate to the “FEATURED” or “PUBLIC” Workspace tabs.

* 

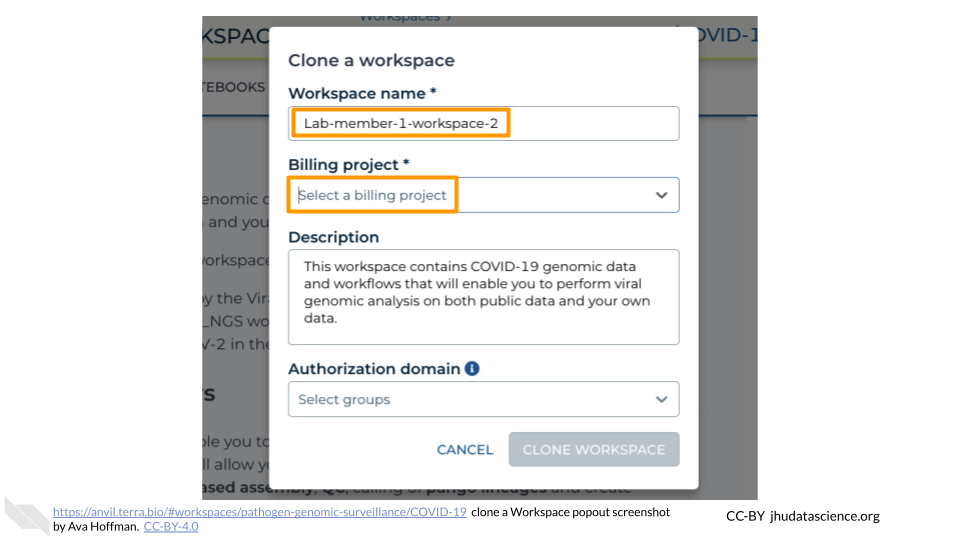
1. Use the search box to find the Workspace you want to clone.

* 

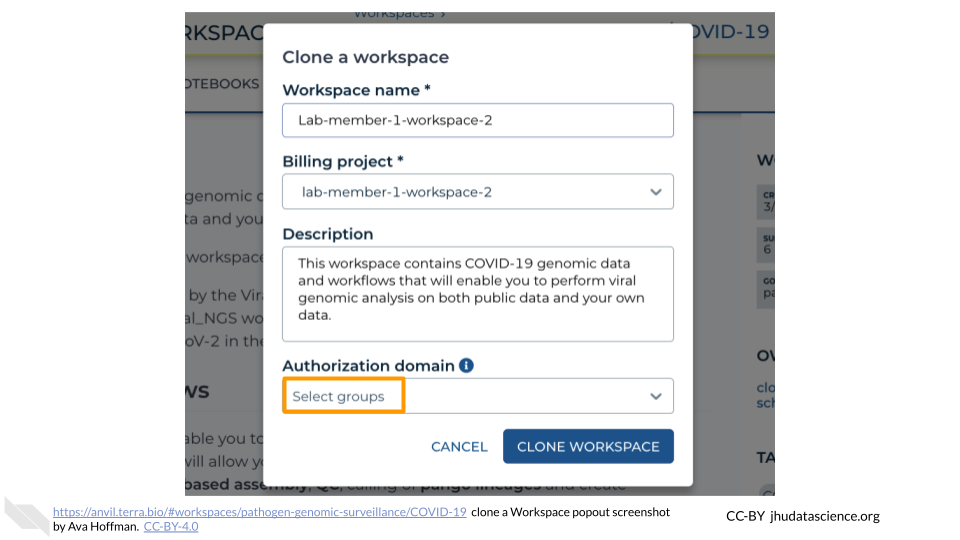
1. Click the teardrop button on the far right next to the Workspace you want to clone. Click “Clone”. You can also clone the Workspace from the Workspace Dashboard instead of the search results.

*  

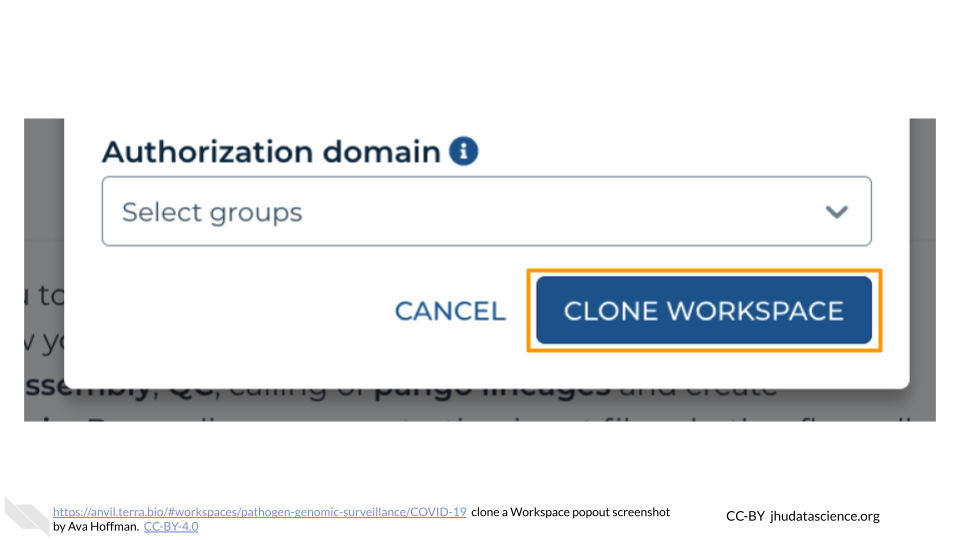
1. You will see a popup box appear. Name your Workspace and select the appropriate Terra Billing Project. **All activity in the Workspace will be charged to this Billing Project** (regardless of who conducted it). Remember that each Workspace should have its own Billing Project.

* 

1. If you are working with protected data, you can set the **Authorization Domain** to limit who can be added to your Workspace. Note that the Authorization Domain cannot be changed after the Workspace is created (i.e. there is no way to make this Workspace shareable with a larger audience in the future). Workspaces by default are only visible to people you specifically share them with. Authorization domains add an extra layer of enforcement over privacy, but by nature make sharing more complicated. We recommend using Authorization Domains in cases where it is extremely important and/or legally required that the data be kept private (e.g. protected patient data, industry data). For data you would merely prefer not be shared with the world, we recommend relying on standard Workspace sharing permissions rather than Authorization Domains, as Authorization Domains can make future collaborations, publications, or other sharing complicated.

* 

1. Click “CLONE WORKSPACE”. The new Workspace should now show up under your Workspaces.

* 

# 2 Using Galaxy on AnVIL

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.

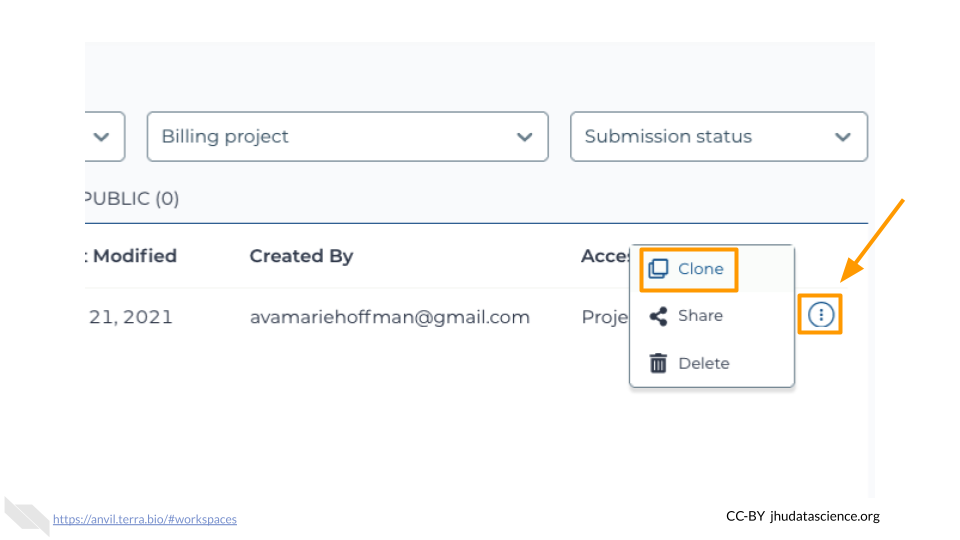
Use a web browser to go to the AnVIL website. In the browser type:

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.

Your instructor will give you information on which workspace you should clone. After logging in, click “View Workspaces”. Select the “Public” tab. In the top search bar type the activity workspace.

Clone the workspace by clicking the teardrop button (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”.



## 2.1 Video overview on using Galaxy

Here is a video tutorial that describes the basics of using Galaxy on AnVIL.

### 2.1.1 Objectives

* Start compute for your Galaxy on AnVIL
* Run tool to quality control sequencing reads
* Stop compute to minimize expenses

### 2.1.2 Slides

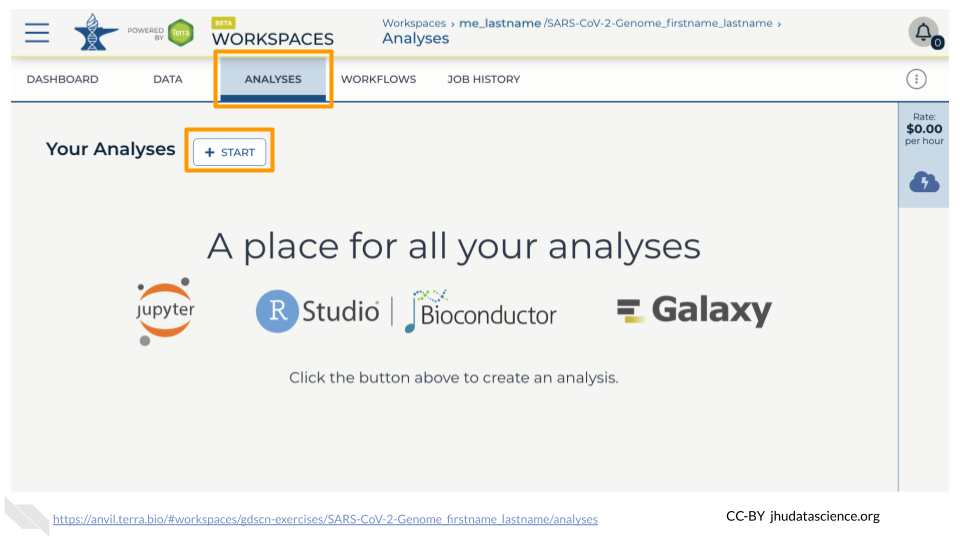
The slides for this tutorial are are located [here](https://docs.google.com/presentation/d/1yYCg4cPVBMMDghT17B4XzROieqyMH99Ex9nMm_Scm9Q/edit?usp=sharing).

## 2.2 Starting Galaxy

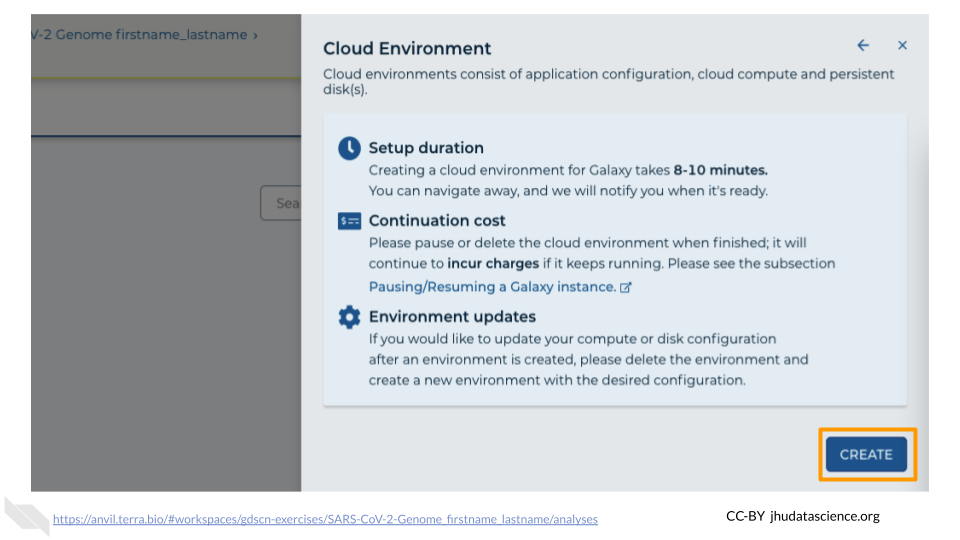
Galaxy is a free, relatively easy to use bioinformatics implementation package. It changes command line programs into GUI based programs and is a great tool for performing bioinformatics analysis without having to update software or worry too much about coding. In order to use Galaxy, we need to create 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.

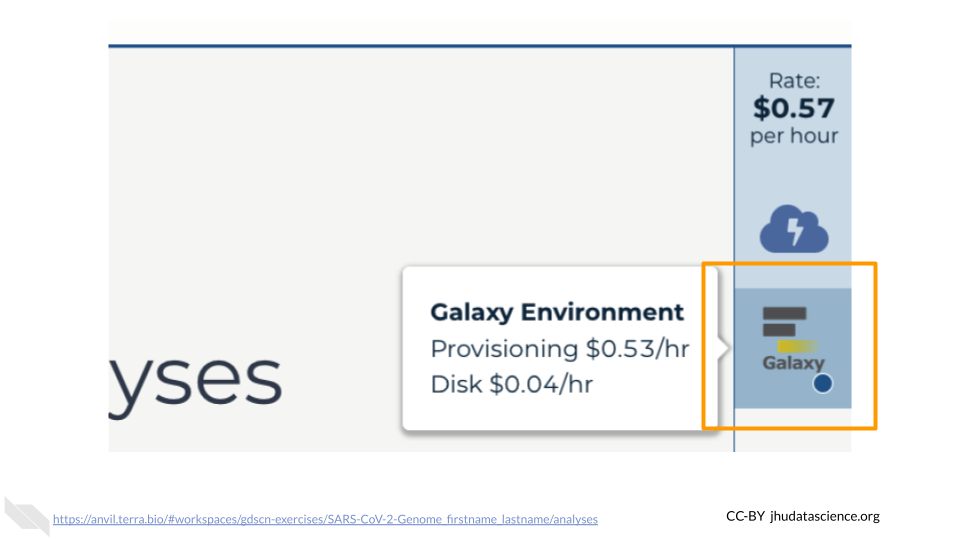
In your new Workspace, click on the “ANALYSES” tab. Next, click on “START”. You should see a popup window on the right side of the screen. Click on the Galaxy logo to proceed.



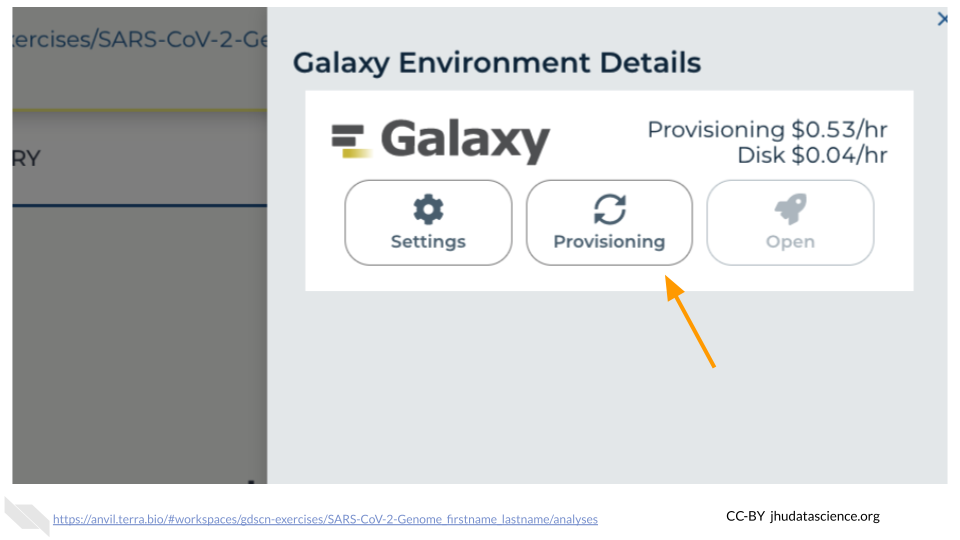
Click on “NEXT” and “CREATE” to keep all settings as-is.



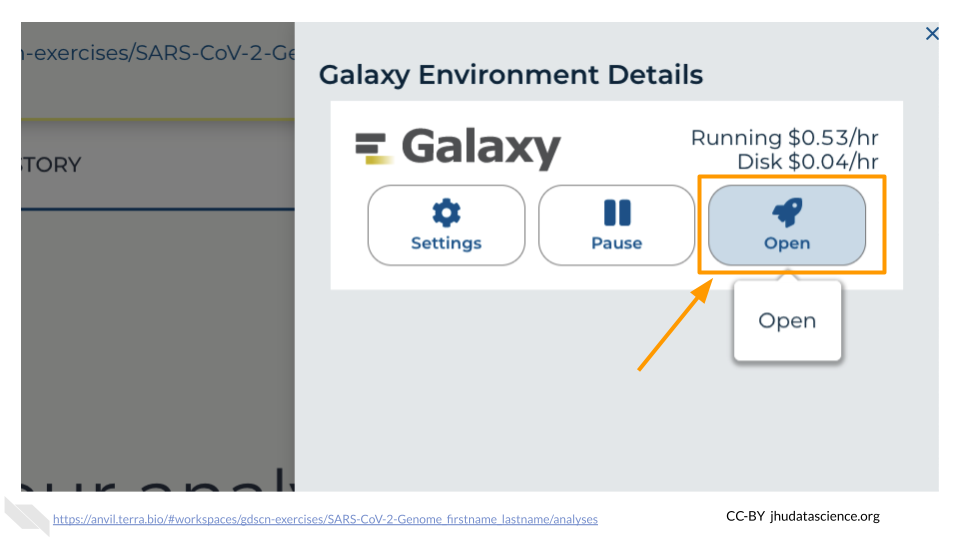
Click on the Galaxy icon.



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.

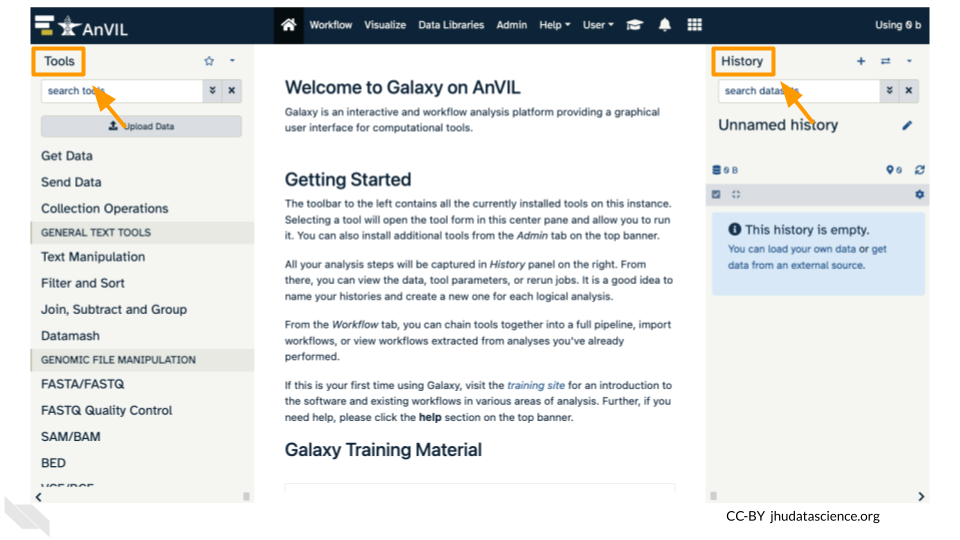
## 2.3 Navigating Galaxy

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.



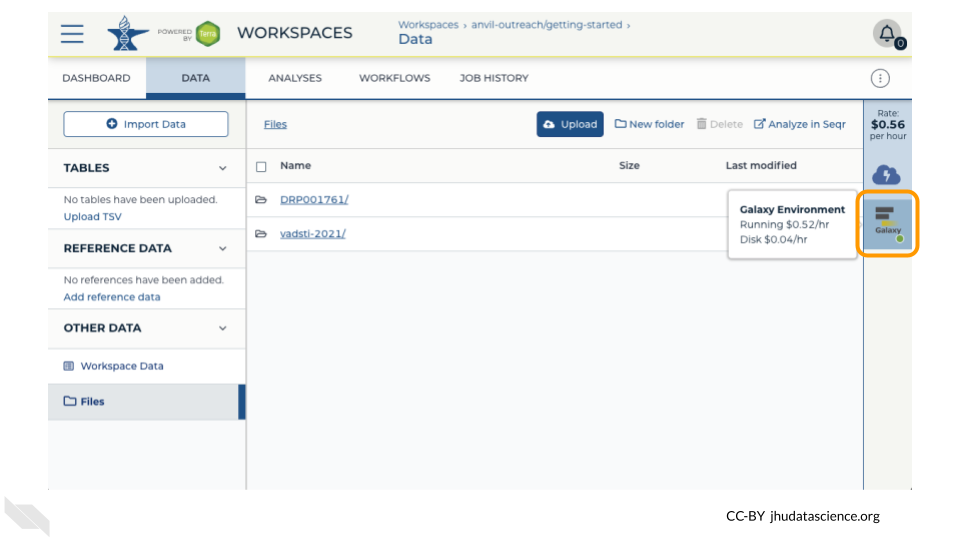
On the welcome page, there are links to tutorials. You may try these out on your own. If you want to try a new analysis this is a good place to start.

## 2.4 Deleting Galaxy

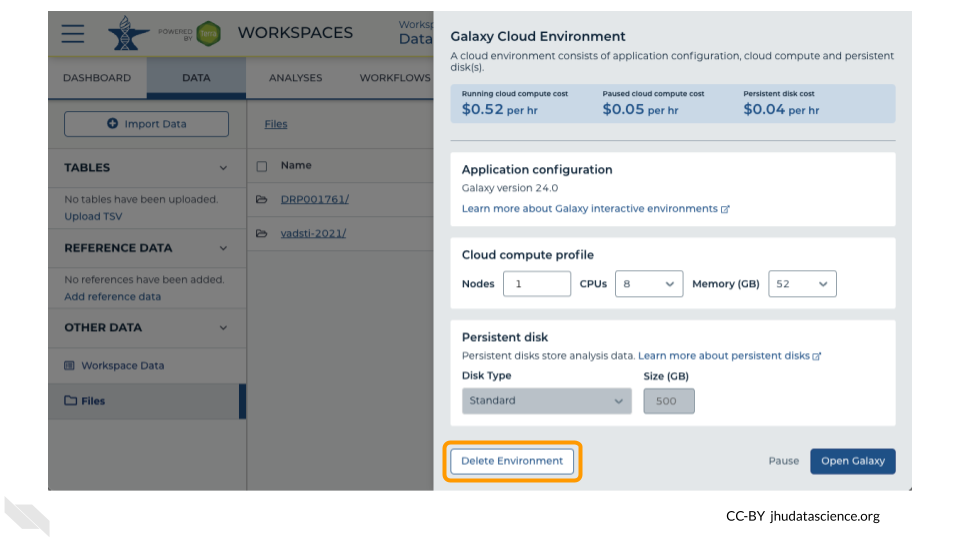
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output

Once you are done with your 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.

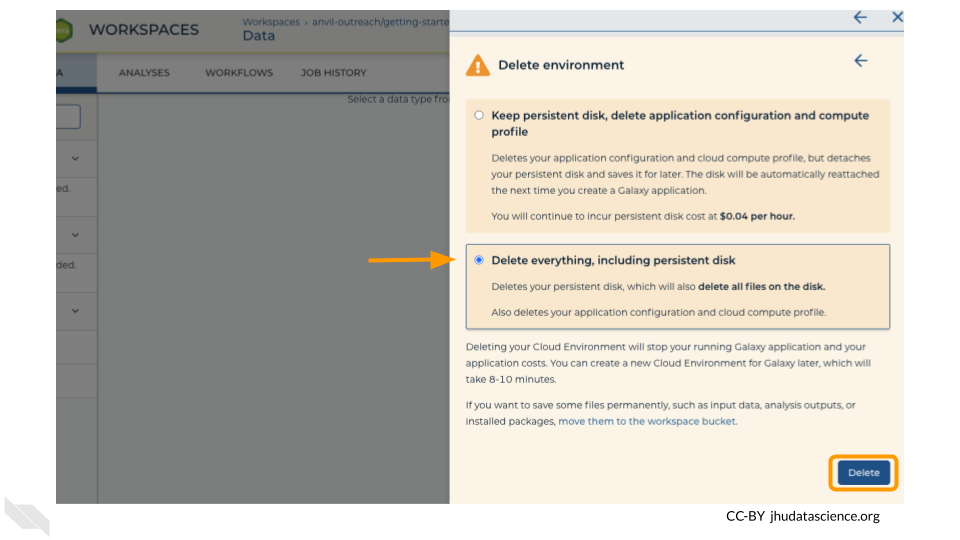
Return to AnVIL, and find the Galaxy logo that shows your cloud environment is running. Click on this logo.



Next, click on “Settings”. Click on “Delete Environment”.



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



# 3 Using RStudio on AnVIL

In the next few steps, you will walk through how to get set up to use RStudio 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.

Your instructor will give you information on which workspace you should clone.

## 3.1 Video overview of RStudio on AnVIL

Here is a video tutorial that describes the basics of using RStudio on AnVIL.

### 3.1.1 Objectives

* Start compute for your RStudio environment
* Tour RStudio on AnVIL
* Stop compute to minimize expenses

### 3.1.2 Slides

The slides for this tutorial are are located [here](https://docs.google.com/presentation/d/1eypYLLqD11-NwHLs4adGpcuSB07dYEJfAaALSMvgzqw).

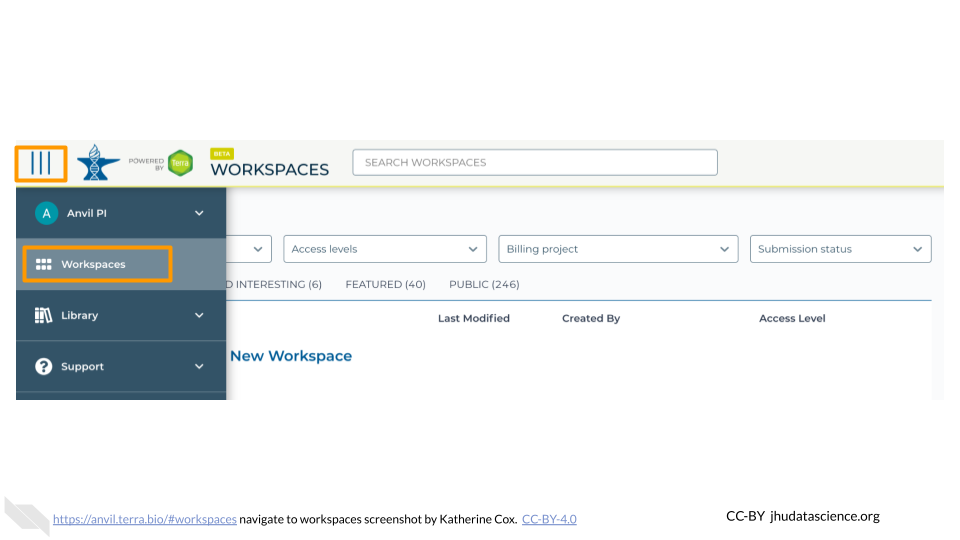
## 3.2 Launching RStudio

## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output

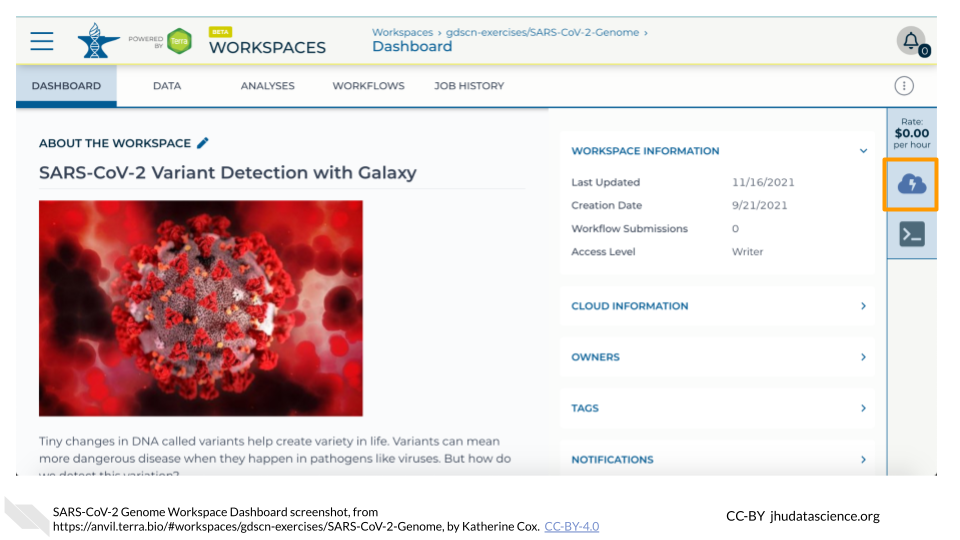
AnVIL is very versatile and can scale up to use very powerful cloud computers. It’s very important that you select a cloud computing environment appropriate to your needs to avoid runaway costs. If you are uncertain, start with the default settings; it is fairly easy to increase your compute resources later, if needed, but harder to scale down.

Note that, in order to use RStudio, you must have access to a Terra Workspace with permission to compute (i.e. you must be a “Writer” or “Owner” of the Workspace).

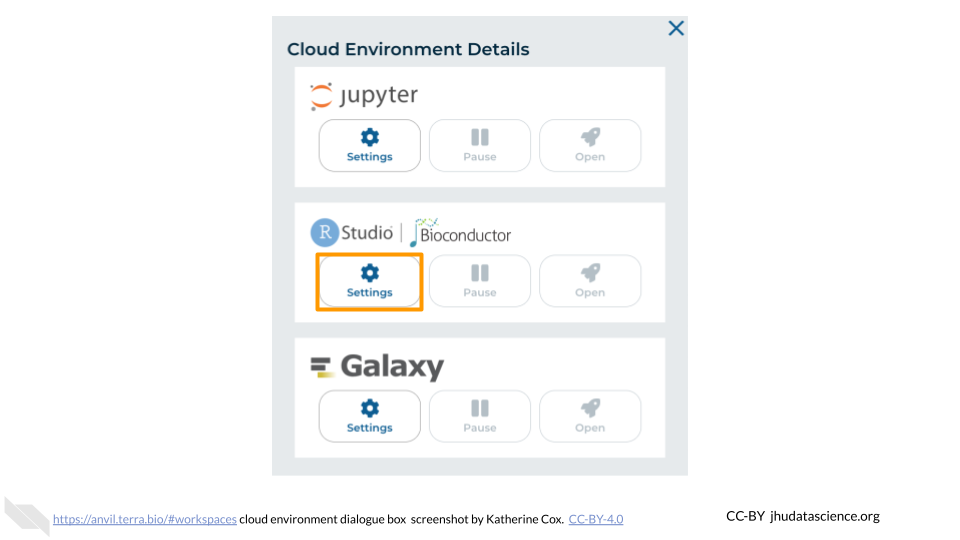
1. Open Terra - use a web browser to go to [anvil.terra.bio](https://anvil.terra.bio/)
2. In the drop-down menu on the left, navigate to “Workspaces”. Click the triple bar in the top left corner to access the menu. Click “Workspaces”.

* 

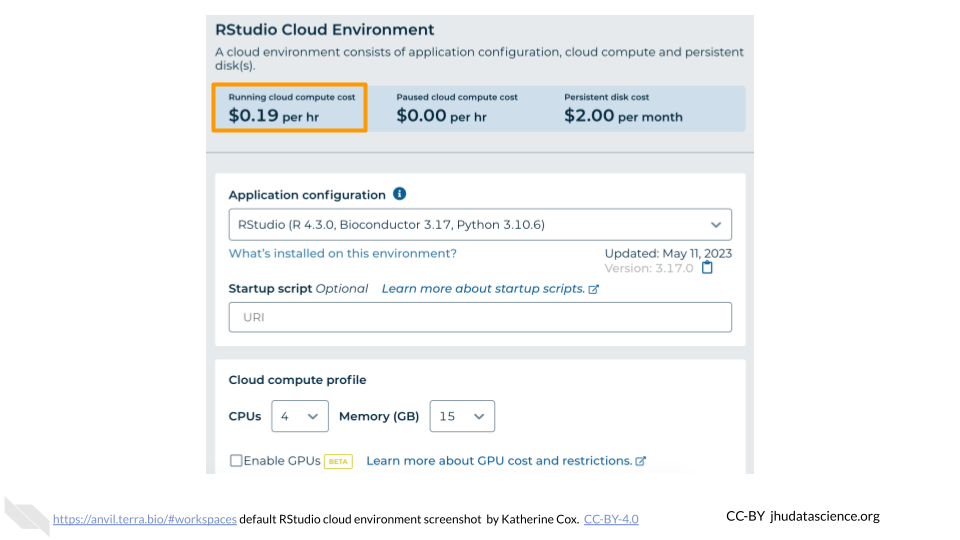
1. Click on the name of your Workspace. You should be routed to a link that looks like: https://anvil.terra.bio/#workspaces/<billing-project>/<workspace-name>.
2. Click on the cloud icon on the far right to access your Cloud Environment options. If you don’t see this icon, you may need to scroll to the right.

* 

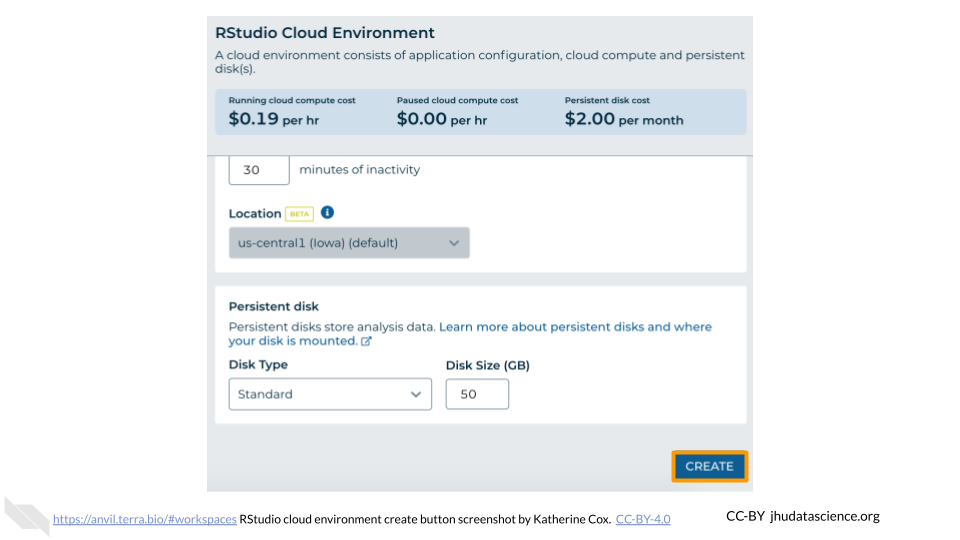
1. In the dialogue box, click the “Settings” button under RStudio.

* 

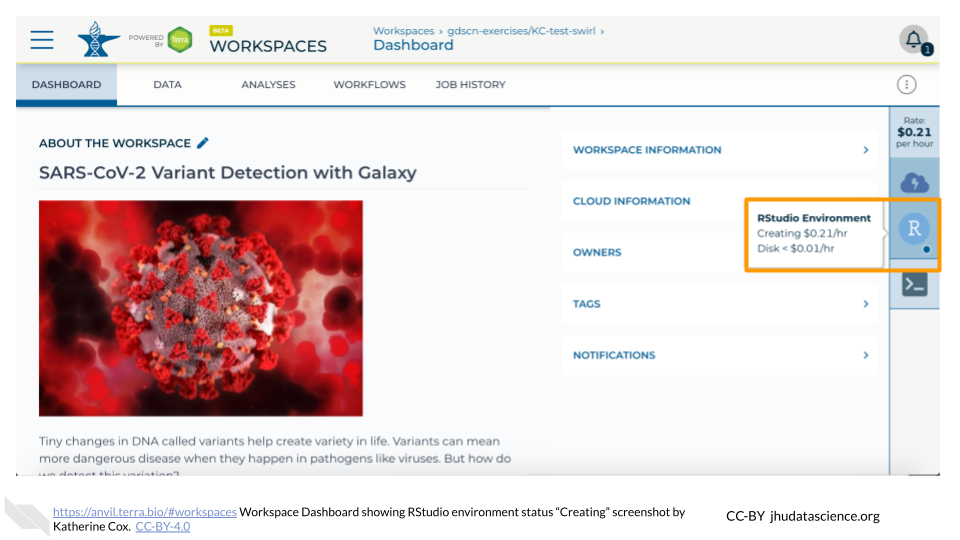
1. You will see some configuration options for the RStudio cloud environment, and a list of costs because it costs a small amount of money to use cloud computing.

* 

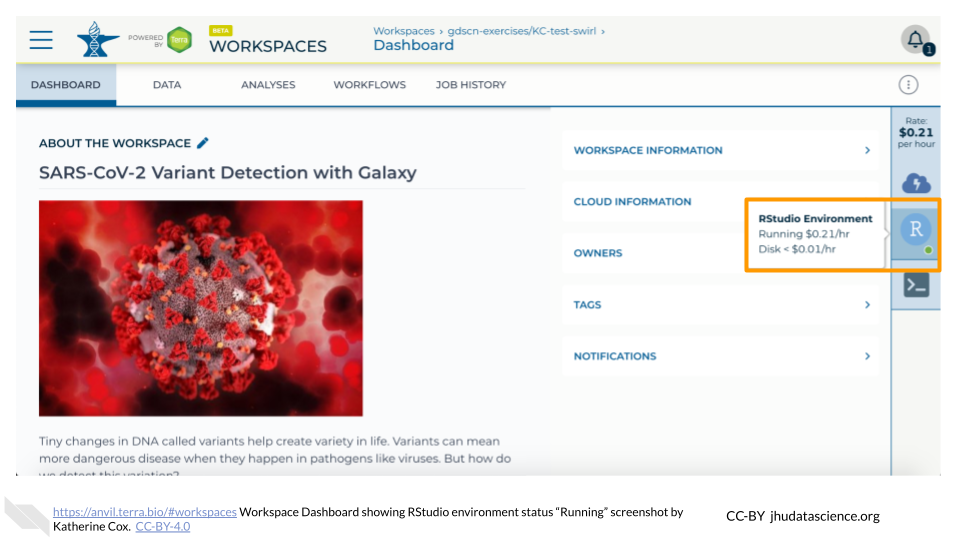
1. Configure any settings you need for your cloud environment. If you are uncertain about what you need, the default configuration is a reasonable, cost-conservative choice. It is fairly easy to increase your compute resources later, if needed, but harder to scale down. Scroll down and click the “CREATE” button when you are satisfied with your setup.

* 

1. The dialogue box will close and you will be returned to your Workspace. You can see the status of your cloud environment by hovering over the RStudio icon. It will take a few minutes for Terra to request computers and install software.

* 

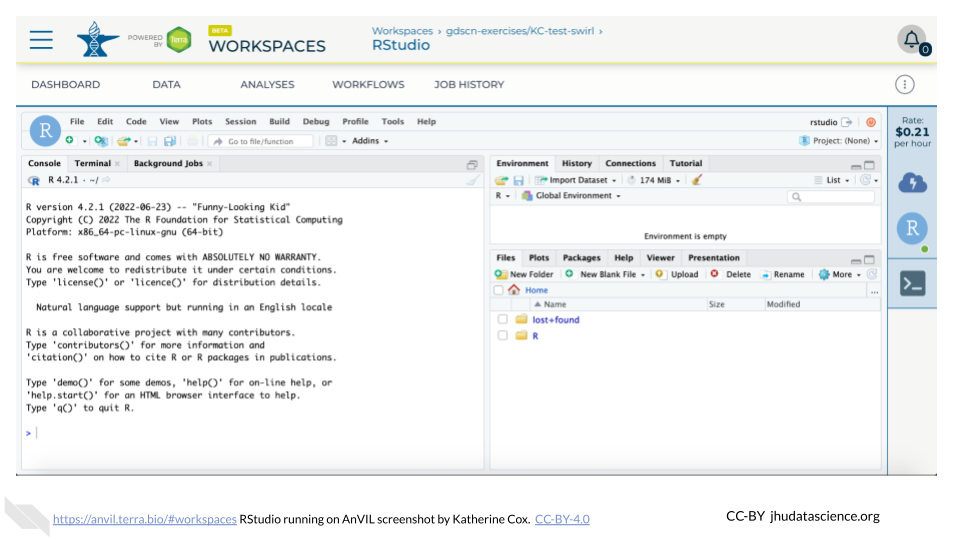
1. When your environment is ready, its status will change to “Running”. Click on the RStudio logo to open a new dialogue box that will let you launch RStudio.

* 

1. Click the launch icon to open RStudio. This is also where you can pause, modify, or delete your environment when needed.

* 

1. You should now see the RStudio interface with information about the version printed to the console.

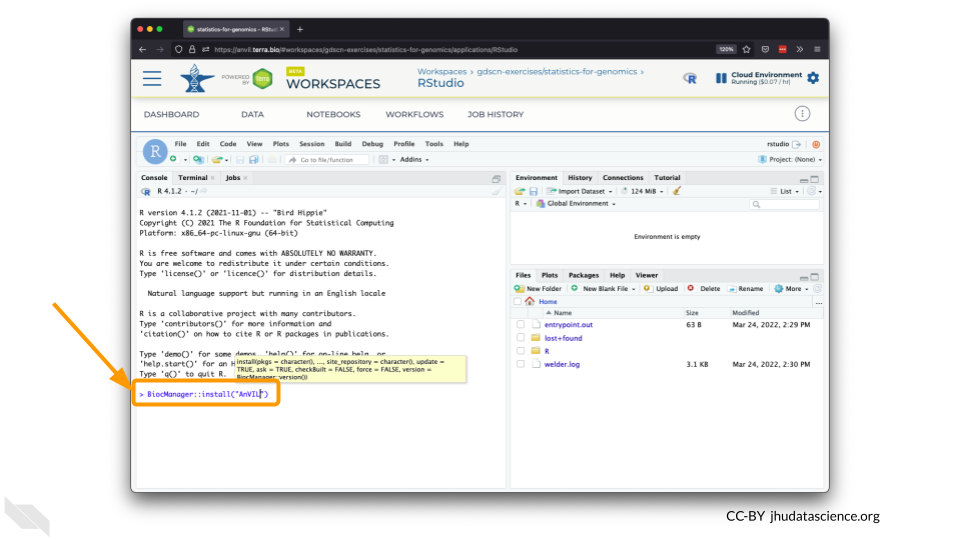
* 

## 3.3 Touring RStudio

## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output

Next, we will be using RStudio and the package Glimma to create interactive plots. See [this vignette](https://bioconductor.org/packages/release/bioc/vignettes/Glimma/inst/doc/limma_edger.html) for more information.

1. The Bioconductor team has created a very useful package to programmatically interact with Terra and Google Cloud. Install the AnVIL package. It will make some steps easier as we go along.

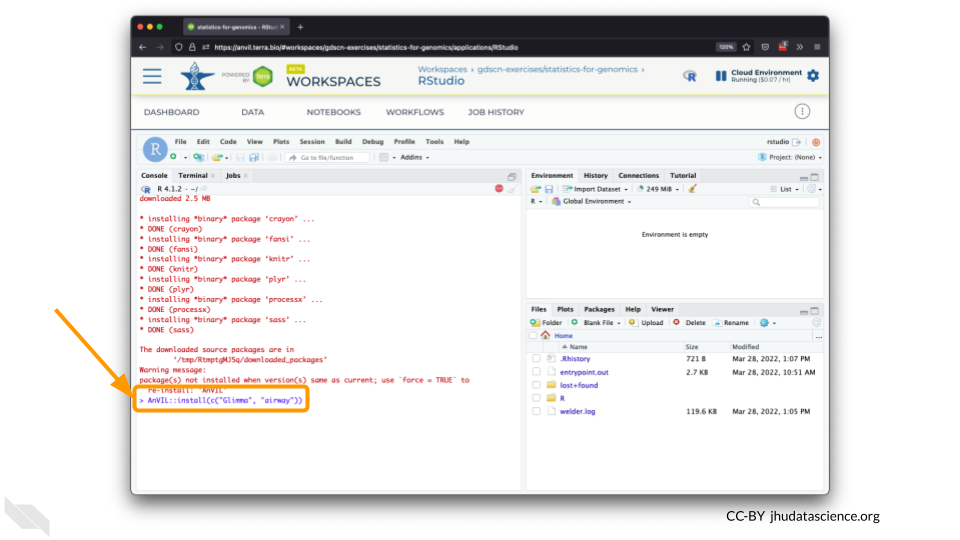
* 

1. You can now quickly install precompiled binaries using the AnVIL package’s install() function. We will use it to install the Glimma package and the airway package. The airway package contains a SummarizedExperiment data class. This data describes an RNA-Seq experiment on four human airway smooth muscle cell lines treated with dexamethasone.

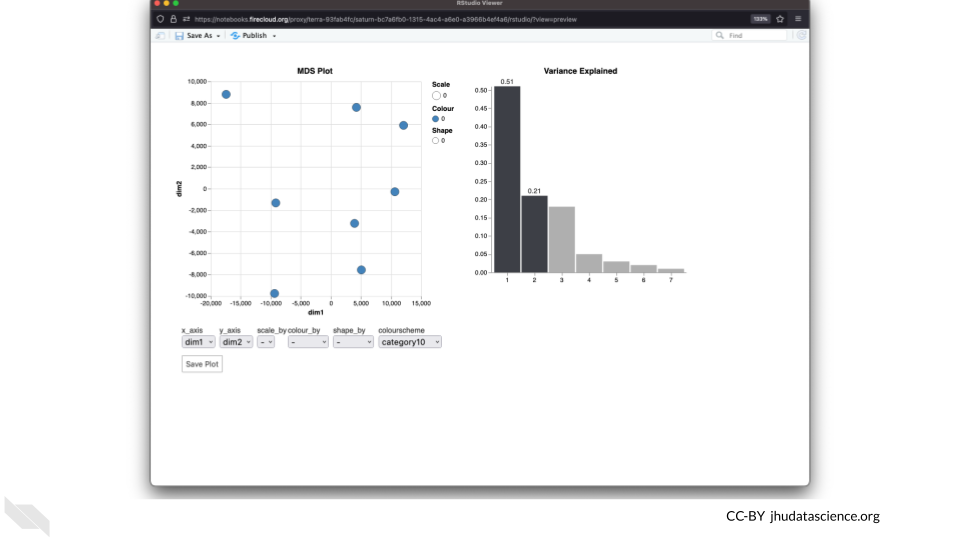
{Note: for some of the packages, you will have to install packaged from the CRAN repository, using the install.packages() function. The examples will show you which install method to use.}

![](resources/images/08-student\_using\_rstudio\_files/figure-docx//1BLTCaogA04bbeSD1tR1Wt-mVceQA6FHXa8FmFzIARrg\_g11f12bc99af\_0\_56.png)

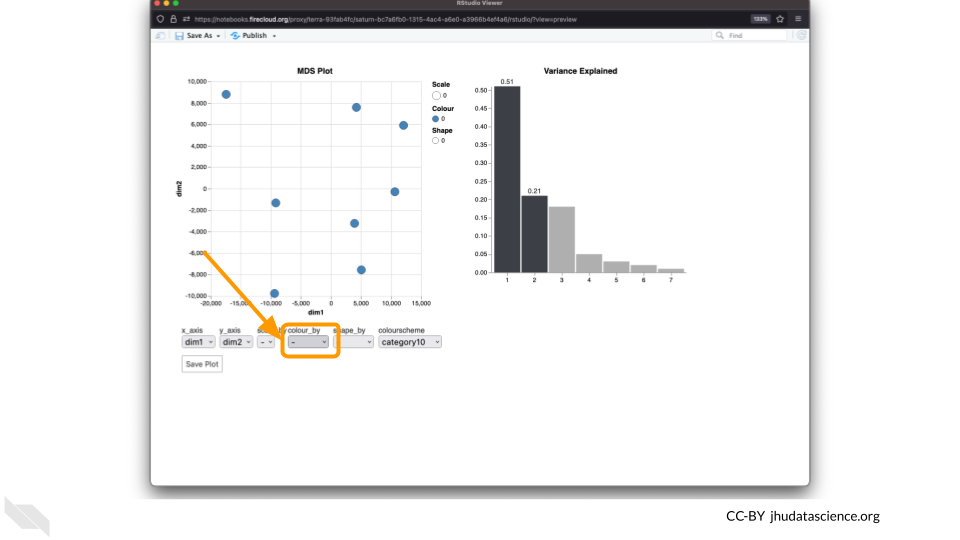
1. Load the example data.

* 

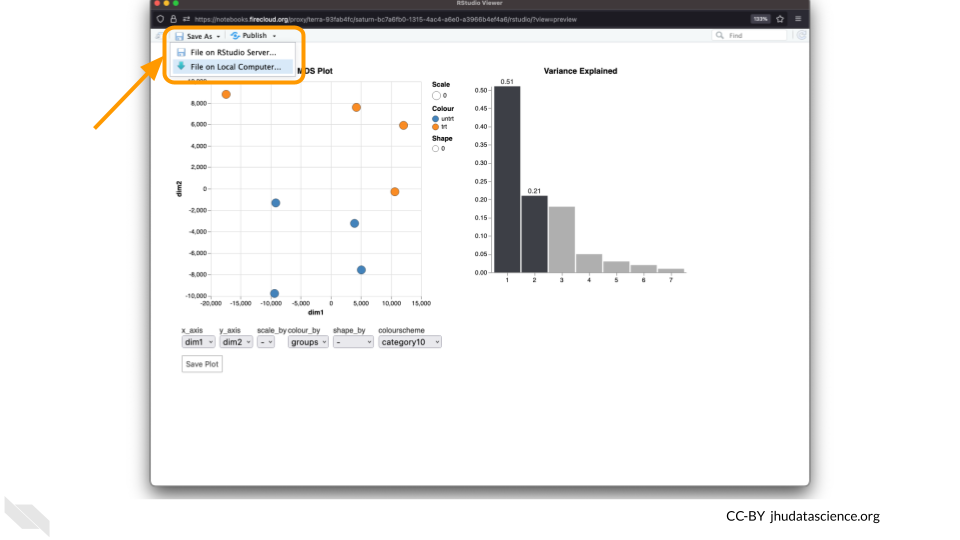
1. The multidimensional scaling (MDS) plot is frequently used to explore differences in samples. When this data is MDS transformed, the first two dimensions explain the greatest variance between samples, and the amount of variance decreases monotonically with increasing dimension. The following code will launch a new window where you can interact with the MDS plot.

* 

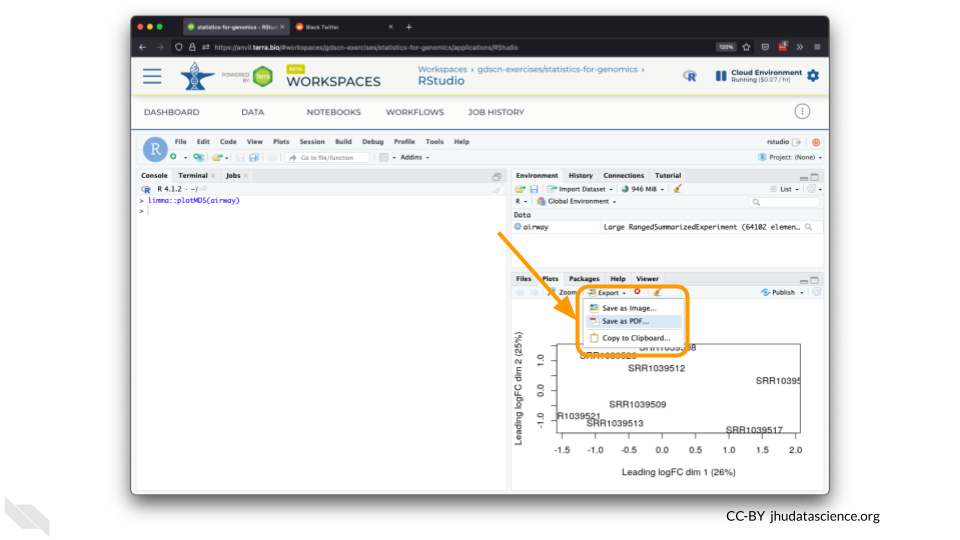
1. Change the colour\_by setting to “groups” so you can easily distinguish between groups. In this data, the “group” is the treatment.

* 

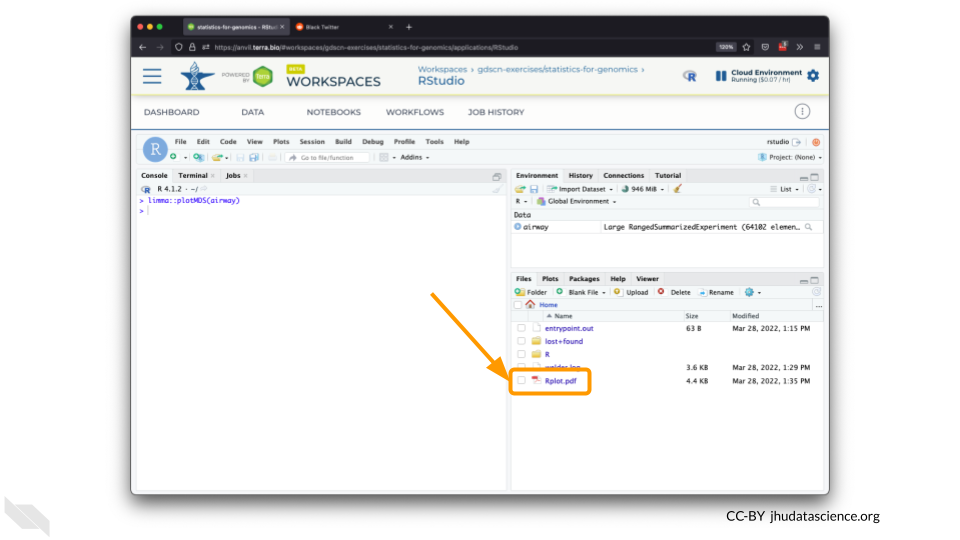
1. You can download the interactive html file by clicking on “Save As”.

* 

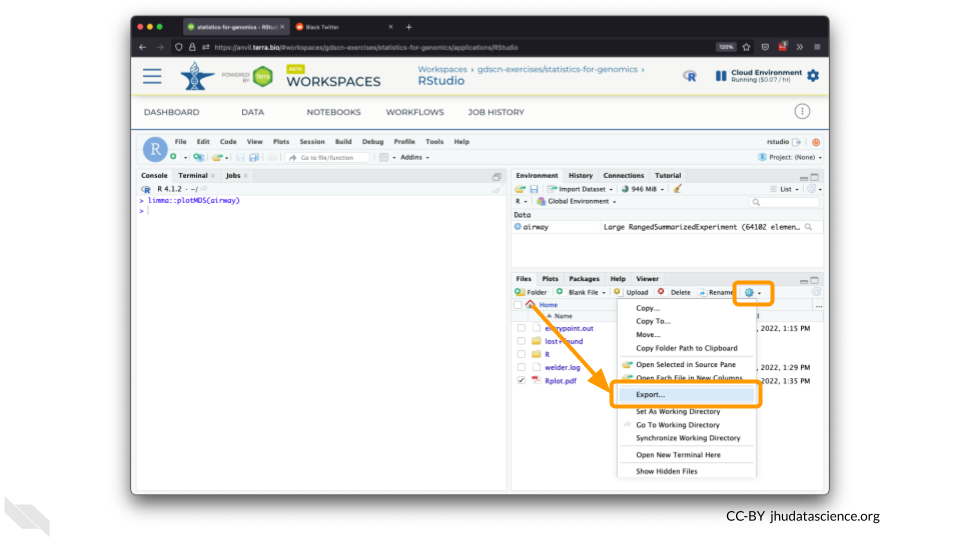
1. You can also download plots and other files created directly in RStudio. To download the following plot, click on “Export” and save in your preferred format to the default directory. This saves the file in your cloud environment.

* 

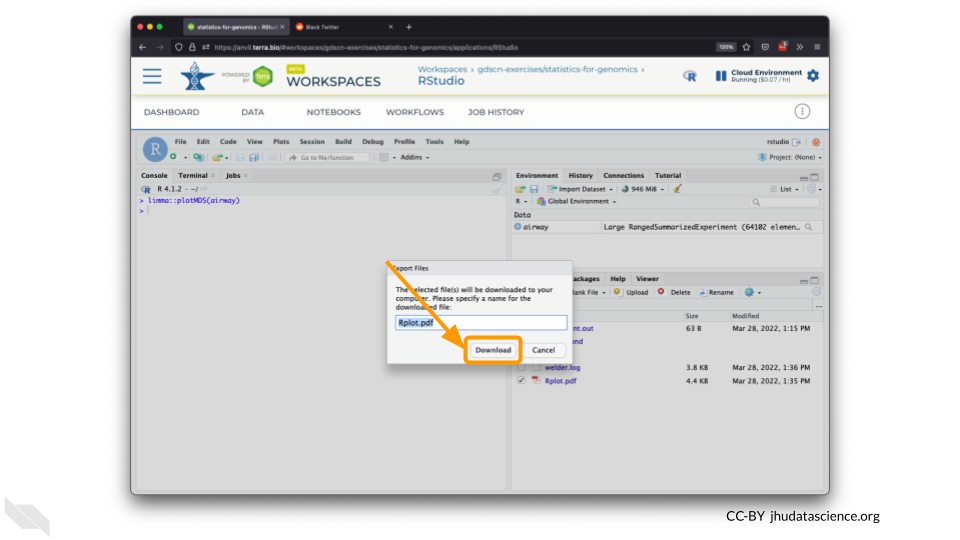
1. You should see the plot in the “Files” pane.

* 

1. Select this file and click “More” > “Export”

* 

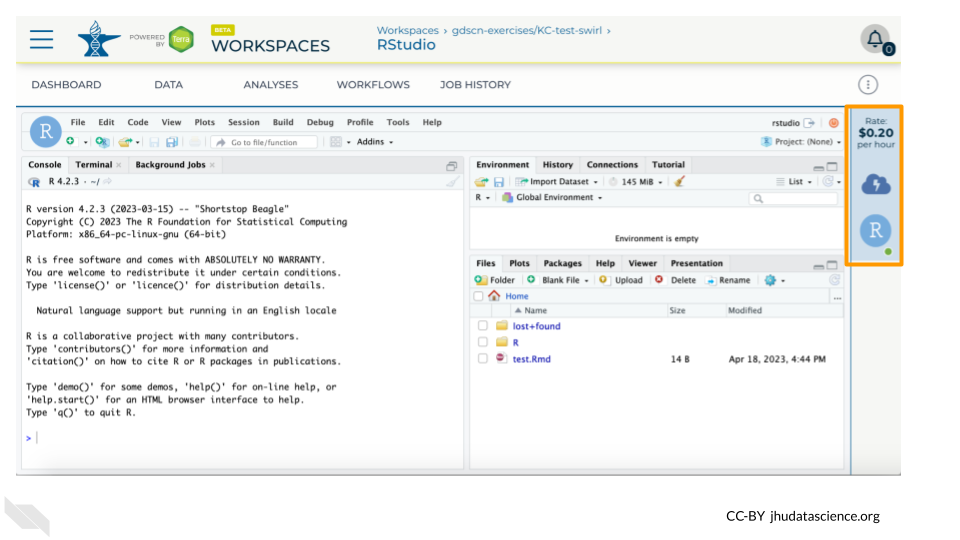
1. Select “Download” to save the file to your local machine.

* 

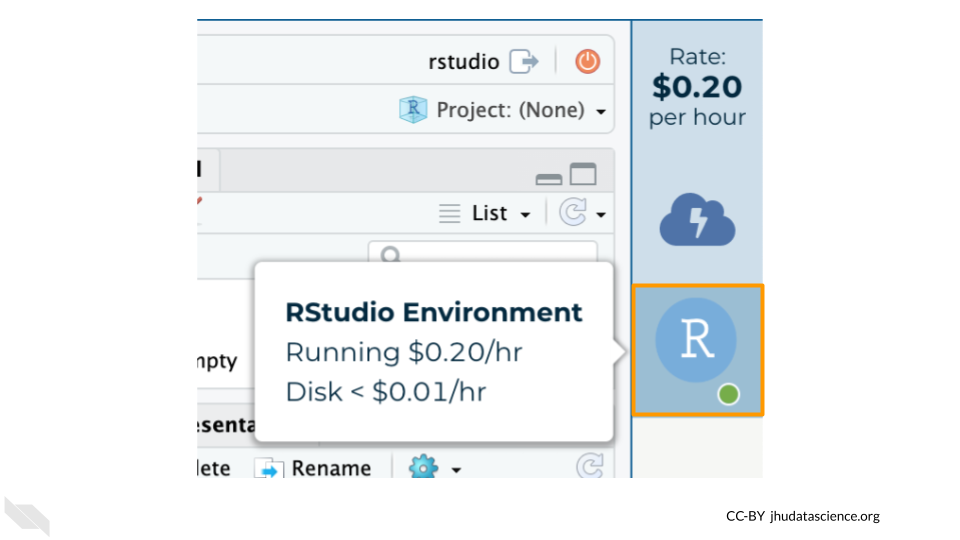
## 3.4 Pausing RStudio

## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output  
## Warning: Chunk option fig.alt is not supported for docx output

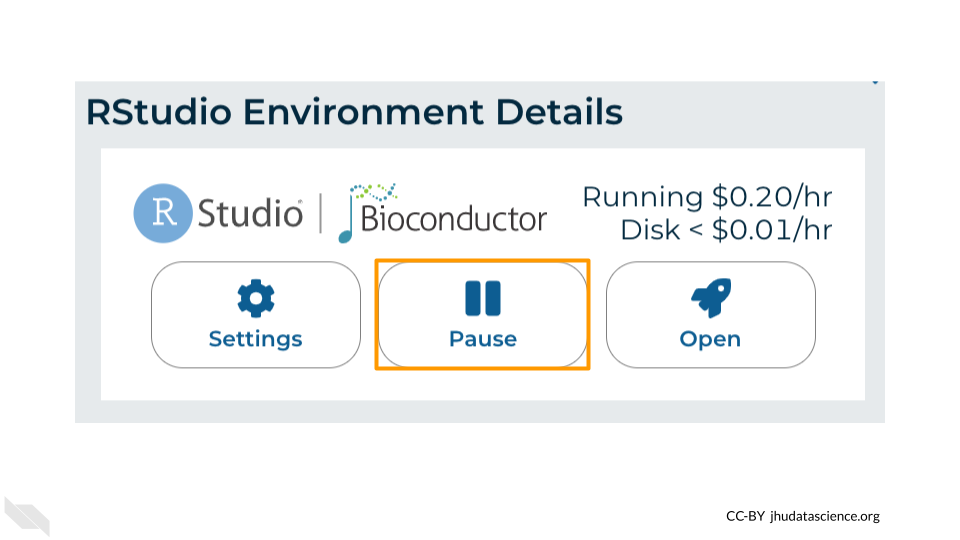
1. You can view costs and make changes to your cloud environments from the panel on the far right of the page. If you don’t see this panel, you may need to scroll to the right. Running environments will have a green dot, and paused environments will have an orange dot.

* 

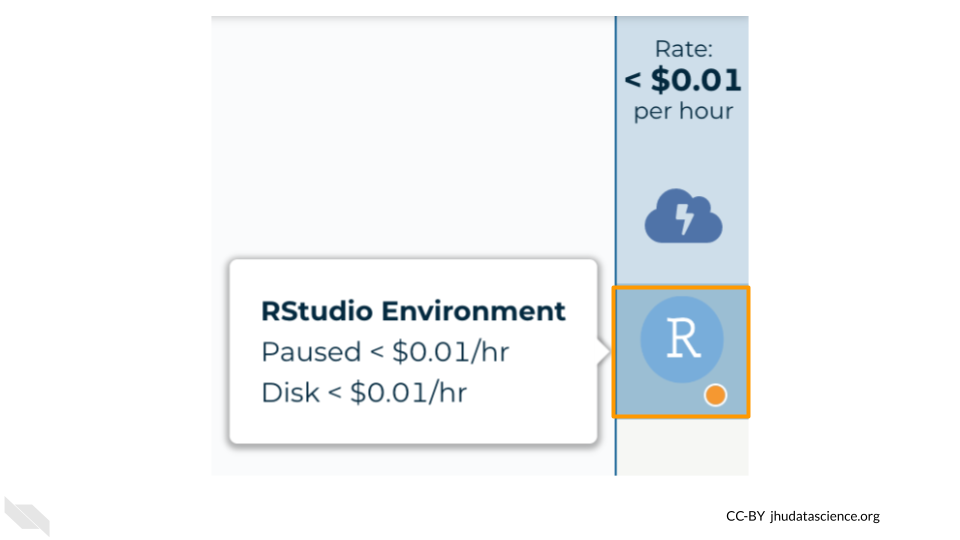
1. Hovering over the RStudio icon will show you the costs associated with your RStudio environment. Click on the RStudio icon to open the cloud environment settings.

* 

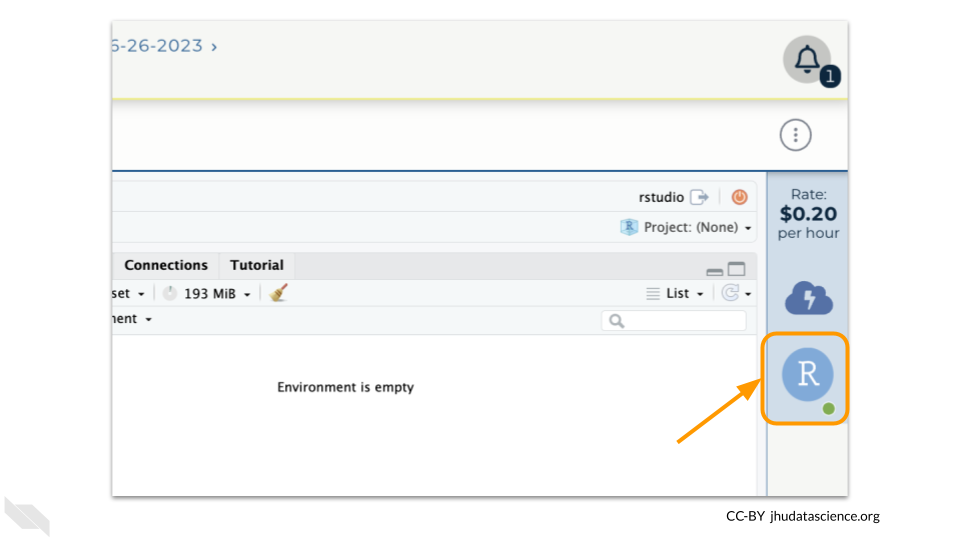
1. Click the Pause button to pause RStudio. This will take a few minutes.

* 

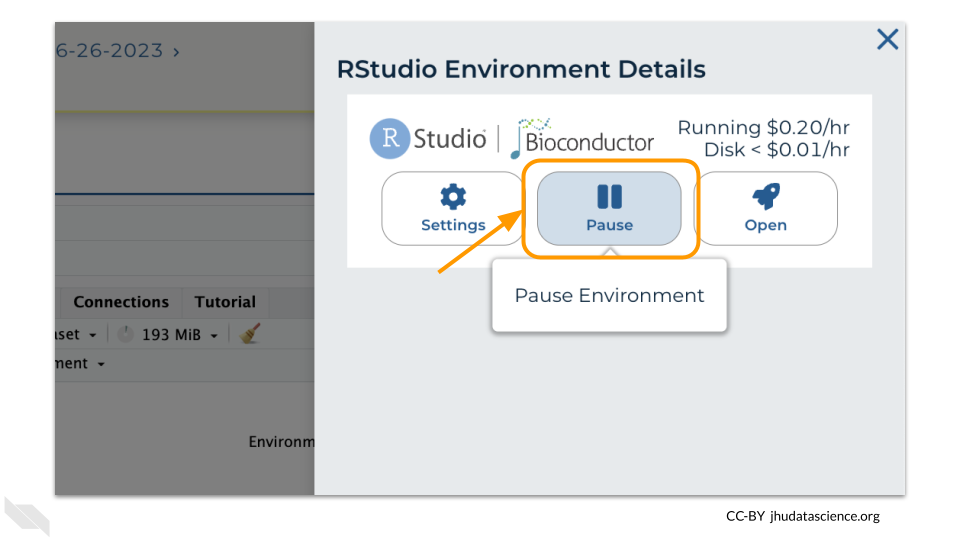
1. When the environment is paused, an orange dot will be displayed next to the RStudio icon. If you hover over the icon, you will see that it is paused, and has a small ongoing cost as long as it is paused. When you’re ready to resume working, you can do so by clicking the RStudio icon and clicking Resume.

* 

1. The right-hand side icon reminds you that you are accruing cloud computing costs. If you don’t see this icon, you may need to scroll to the right.

* 

1. You should minimize charges when you are not performing an analysis. You can do this by clicking on the RStudio icon and selecting “Pause”. This will release the CPU and memory resources for other people to use. Note that your work will be saved in the environment and continue to accrue a very small cost. This work will be lost if the cloud environment gets deleted. If there is anything you would like to save permanently, it’s a good idea to copy it from your compute environment to another location, such as the Workspace bucket, GitHub, or your local machine, depending on your needs.

* 

You can also pause your cloud environment(s) at <https://anvil.terra.bio/#clusters>.

# 4 References