Getting started with Galaxy on Ceres (22 September 2017)

Contacts: ARS-Galaxy.Support@ars.usda.gov

There are a few compelling reasons why we think you're going to like using Galaxy on Ceres:

1. It is a great bioinformatics alternative for those who prefer a graphical user interface.
2. You don't have to worry with defining queues and number of nodes in a batch command file. We have tried hard to establish appropriate default parameters to take advantage of Ceres's parallel processing power.
3. FTP transfer is tightly integrated in the Ceres Galaxy framework, so file transfer is intuitive and you don't run into data limit bottlenecks as quickly.
4. Galaxy makes it easy to share your analysis with bioinformatic support should the need arise. In addition, you can share with collaborators assuming that they have a Ceres account.
5. There is plenty of external documentation covering almost all conventional bioinformatic analyses. In fact, many workflows (see below) probably already exist that you can use directly on your data.
6. If you don't see a tool in the current interface, it probably is in the Toolshed (see below). You can then request that an administrator install it.

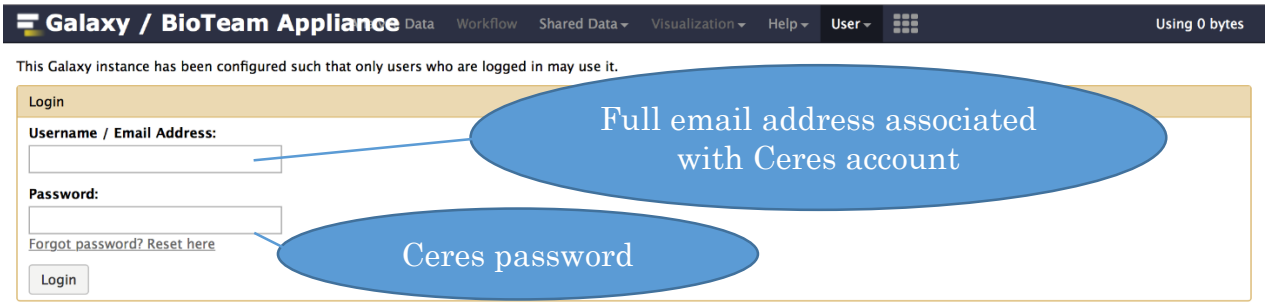
In order to use Ceres Galaxy, you need an account on Ceres. Go to https://e.arsnet.usda.gov/sites/OCIO/scinet/accounts/SitePages/SCINetAccountRequest.aspx to start that process. New accounts will automatically get a Galaxy user name and directory. The user name should match the email you used when registering for your Ceres account and the password will match your Ceres password.

A quick note on the terms "Ceres" and "SCINet": "SCINet" is the network through which you access the high-performance computer named "Ceres". Generally, we will refer to Ceres accounts in this tutorial because most people think of having an account on a computer, but the account is, more broadly, a network account on SCINet.

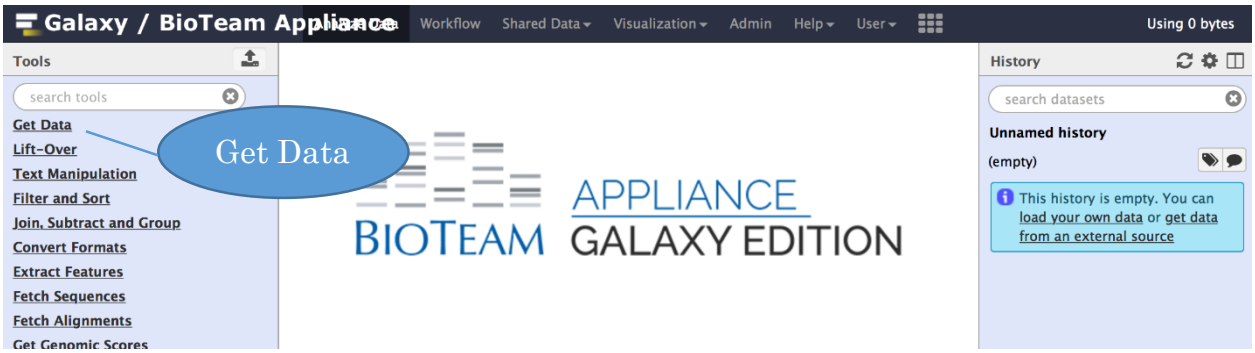
Logging on

Ceres Galaxy is public facing, so to access it you just type or paste "https://galaxy.scinet.science" into the address bar of a web browser. We recommend Firefox, Chrome, or Safari. MAKE SURE TO TYPE THE "https://" prefix and note the "s".

You will be presented with the following screen:

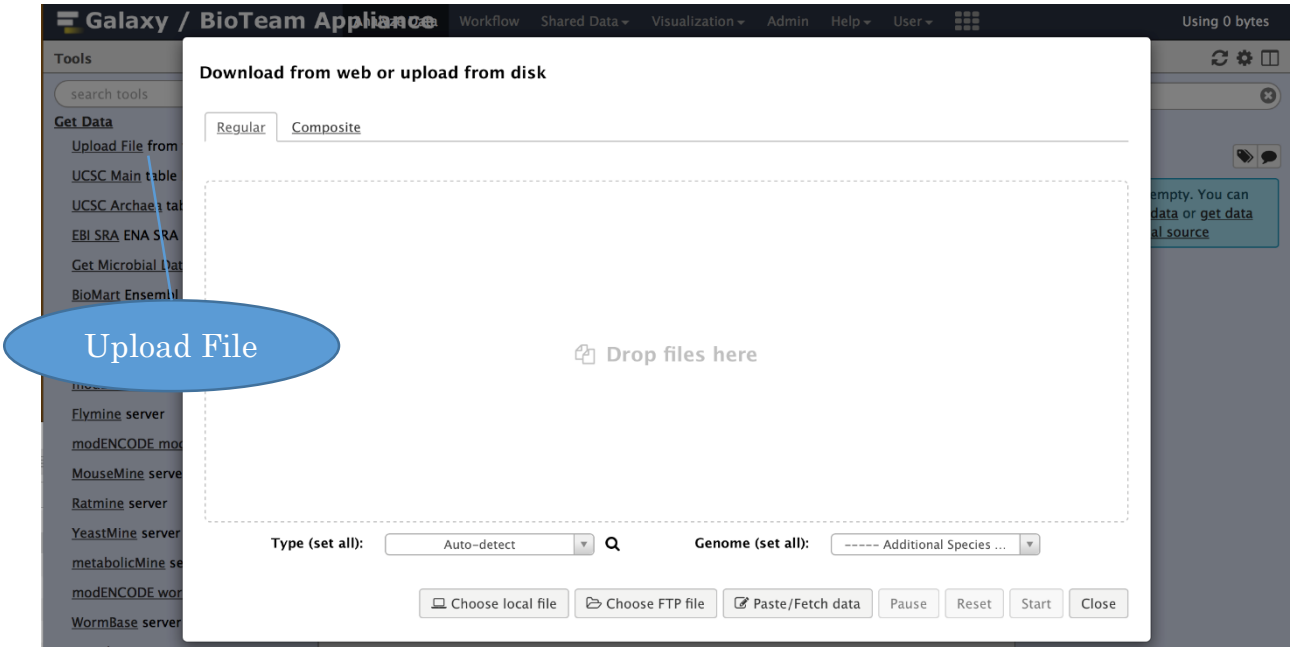


Your user name should match the full email address you used when registering for your Ceres account and the password should match your Ceres password. After logging in, the standard Galaxy home screen should appear as:



The *Tools* bar in the left window frame is where you can load, manipulate, and analyze data. The central window frame is where you will see options and parameters programs that you will be running. The right *History* frame shows all the imported files and programs you have run. In order to begin, you need to upload data. Often data of general interest, such a genome sequences or annotations, are already available in Galaxy under the *Get Data* tab. More commonly, you will be using data that you have generated.

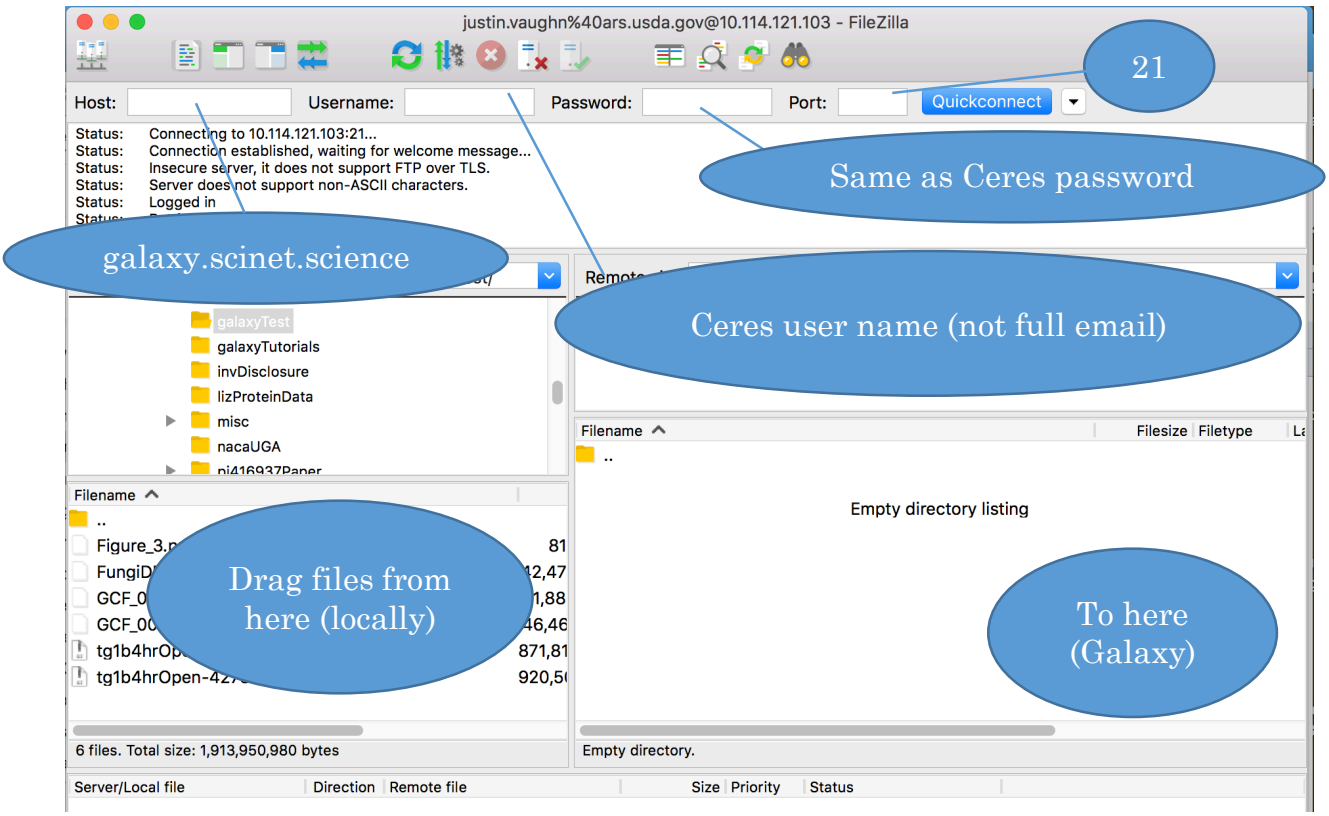
You will import files into Galaxy by clicking on the *Tools* sub-heading *Get Data* and then *Upload File* under *Get Data*. The following screen should appear:



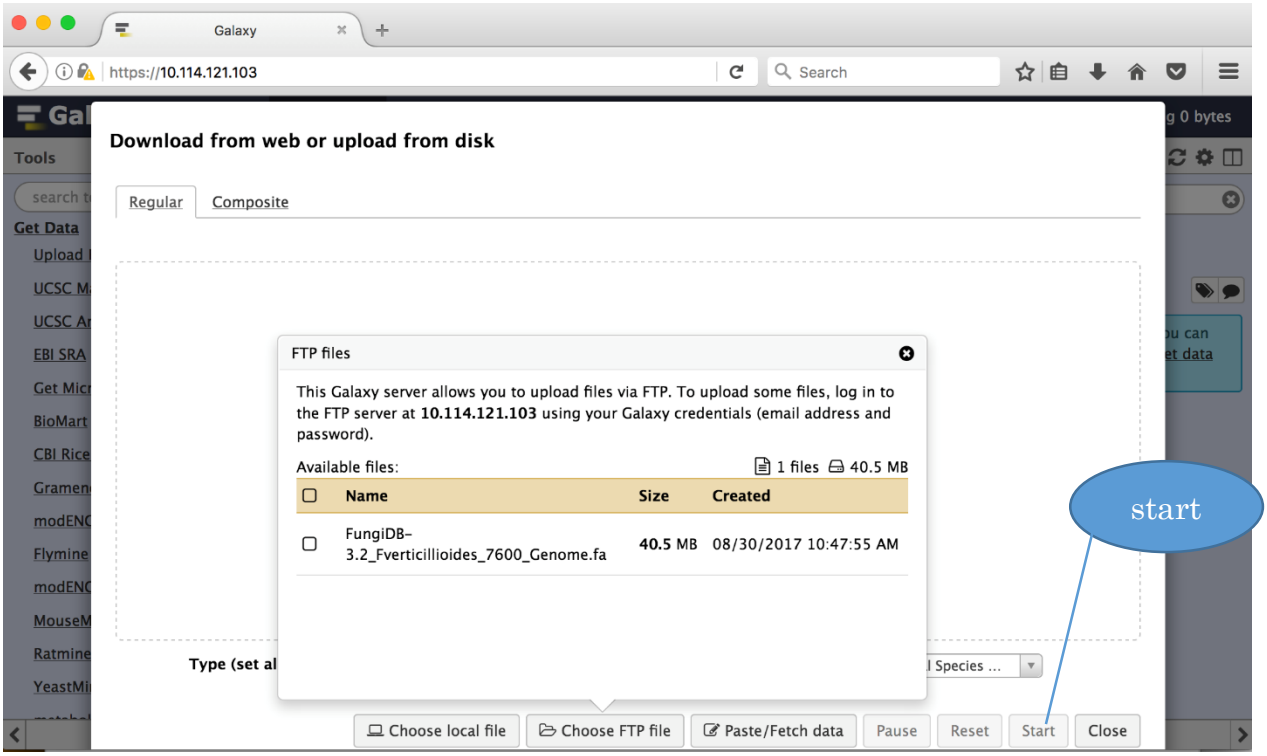
Galaxy offers a method to import data directly from your computer as *Choose local file* button. Feel free to experiment with this direct method, but we find it can be slower and more fickle than FTP transfer. To that end, we will mainly focus on uploading files using a file transfer program, such as Filezilla. If you click on *Choose FTP file*, Galaxy will look in your FTP folder on Ceres for files you have uploaded, so you must first upload your data to that folder.

FTP transfer to Ceres Galaxy

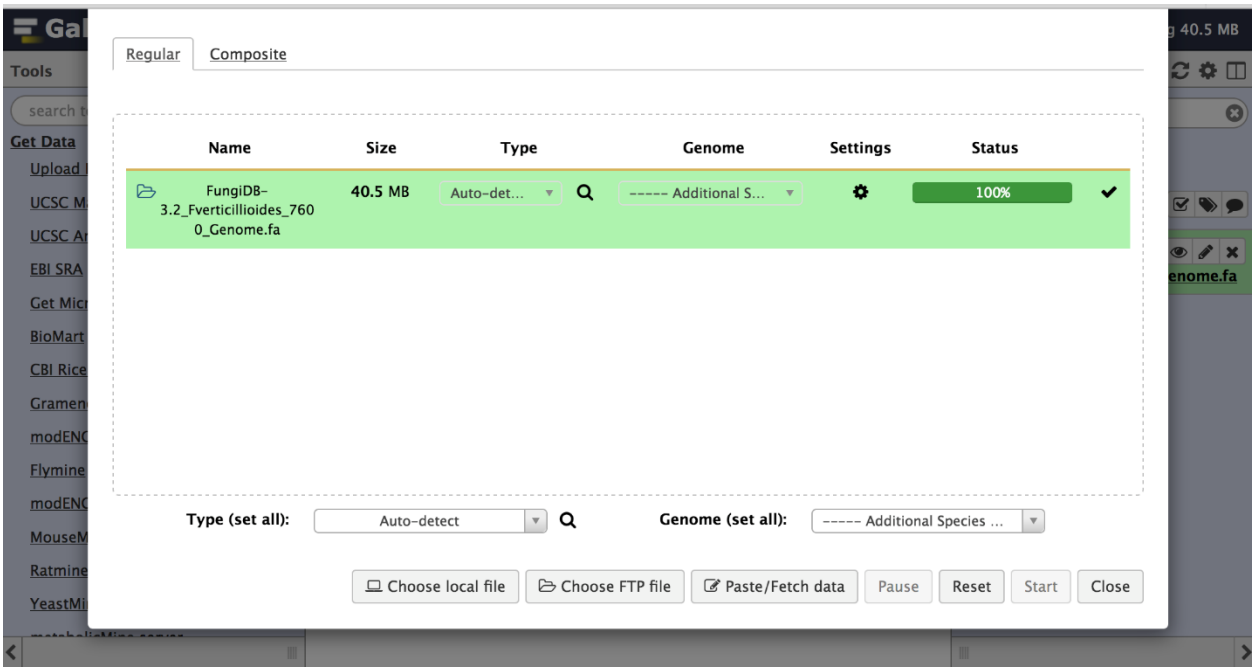
We recommend using Filezilla (https://filezilla-project.org/), which will work on Mac, Linux, or Windows. Cyberduck is another good option. Once installed you can access the directory were you need to move files by entering the Host [galaxy.scinet.science], Username, and Password in the blanks supplied. Enter "21" in the Port blank. Click "Quickconnect". (Filezilla will remember this information under the arrow next to "Quickconnect".) Once logged in, you will be automatically sent to the directory where files need to be deposited.



In Filezilla, you can simply drag-and-drop files from your computer (on the left) to your Galaxy folder (on the right). (Galaxy will uncompress many file-types on-the-fly, so it is best to transfer files in their compressed form.) Once you have done this, go back to the "Upload File" screen and click "Choose FTP site". You will see a screen like this:

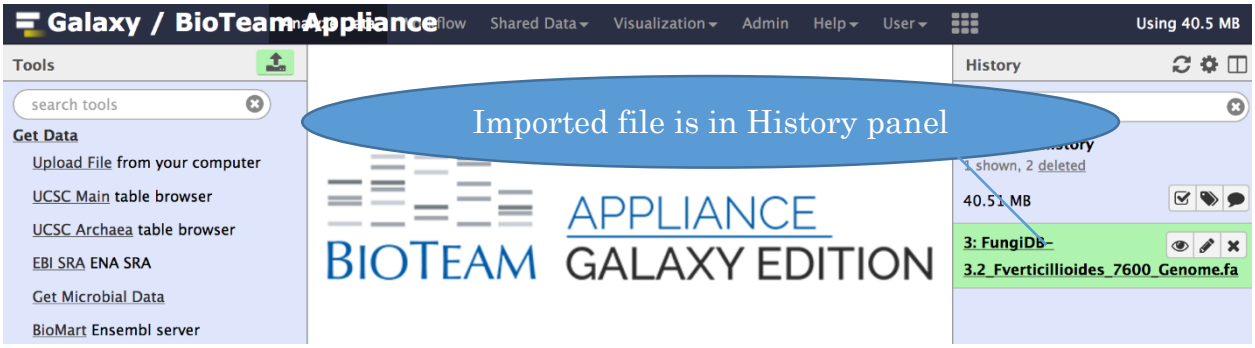


Select the file you want to import and hit "Start". Once imported, your screen will look like this:



NOTE: IMPORTING THE DATA INTO GALAXY WILL REMOVE THE FILE THAT YOU MOVED VIA FTP.

You can close this window and the file will appear in your History as so:



This is a generic way to import files and can be used regardless of file format. You are now ready to process your data.

NOTE: If you already have data on Ceres and want to make this data "visible" to Galaxy, you can use the *galaxy* folder in your home directory, which is automatically created when your galaxy account is established. Copy files into this directory, either via command-line ("cp" command) or a Filezilla-like tool. These files will appear as if you had uploaded them via FTP and can be imported using the approach described above.

Using Galaxy

We have tried to focus on aspects of using Galaxy that are specific to USDA-ARS's Ceres installation. There are hundreds of tutorials and videos to introduce you to the Galaxy framework, which is essentially the same no matter where it is installed. We suggest that you start at https://galaxyproject.org/learn/. A nice interactive introduction is also available at Help > Interactive Tours, or https://galaxy.scinet.science/tours.

A common initial hang-up is getting your uploaded data in the right format. For sequencing data, you usually need to make clear what quality-scale you are using. For illumina reads, review the following link: https://galaxyproject.org/support/fastqsanger/. In addition, it is common for sample data from sequencing centers to be spread across multiple lanes. If you only have a few samples, you can concatenate these in Galaxy using Concatenate – Text Manipulation tool. If you have many samples, it will probably be best to concatenate those files prior to uploading using the commandline function *cat*. For example, for files 'samp1_L1_R1.fq', 'samp1_L2_R1.fq' use 'cat *R1.fq >sample1_R1.fq'. Also note, .gz compressed files can be concatenated using the same approach without having to uncompress them first.

You will usually want to structure your data into Collections for batch processing and downstream analysis. See https://galaxyproject.org/tutorials/collections/ or, for a worked example, https://depot.galaxyproject.org/hub/attachments/documents/presentations/gcc2014/C

hilton.pdf. Alternatively, most tools will allow you to run the same process on multiple datasets of the same format without combining them as Collections.

Some links to common analysis are given below. Many of these analysis already exist as published workflows and can be used directly. See https://usegalaxy.org/workflow/list_published for a searchable list. To use, download the workflow of interest, click on the "Workflow" tab at the top of your main screen, and then import the workflow according to the instructions. An example of usage is available here: http://sepsis-omics.github.io/tutorials/modules/workflows/. For those wanting to develop their own worflows, a graphic editor is available, as described here: https://galaxyproject.org/tutorials/g101/#opening-workflow-editor.

RNA-seq – general overview (https://galaxyproject.org/tutorials/rb_rnaseq/) and galaxy specific pipeline (https://galaxyproject.org/tutorials/nt_rnaseq). Also check out https://sites.google.com/site/princetonhtseq/tutorials/rna-seq.

SNP-calling – https://galaxyproject.org/tutorials/var_dip/ for diploid genomes and https://galaxyproject.org/tutorials/var_hap/ for haploid genomes.

IMPORTANT: If you receive the job error "This job was terminated because it ran longer than the maximum allowed job run time." It means that the tool has not been appropriately configured to run on Ceres using the scale of data that you have provided. Please contact ars-galaxy.support@ars.usda.gov, and we can optimize these parameters for you and for future users.

Sharing your data and analysis

If you need some consultation on your results or on parameter settings, it can be very useful to share your analysis with someone so that you can both be viewing the same thing. To that end, we will give this aspect special emphasis. You can see a short introduction to this option at https://moin.galaxyproject.org/Learn/Share.

Can't find a tool you need?

We have loaded a core subset of all tools that would be available on the public Galaxy server (https://usegalaxy.org). If you do not see a tool you need and cannot use a good alternative, you should search in the Toolshed (https://toolshed.g2.bx.psu.edu). After identifying the relevant tool, email a Ceres Galaxy administrator: **ARS-Galaxy.Support@ars.usda.gov**. If there are multiple options, we encourage the use of tools developed by either "devteam" or "iuc". Depending on the nature of the tool you are requesting, we may ask you to supply a sample dataset for testing purposes, so anticipate having that data available.