

PHS675: Special Topics: Analysing RNA-Seq Data

Fall 2026

Homework 2: Configuring your biocluster accounts

This homework assignment is on a pass/fail basis. The purpose of this is to configure your biocluster accounts so that you have a suitable disk space to work in, and so that you can access the software we need for the class. Since this is merely a set of organizational steps to ensure you can do things in the class, all plagiarism and AI rules are waived. There is no need for you to understand precisely what is happening here, we just need this to work.

Please complete this before class on Tuesday, September 8th. Come to my office and get help if you can't get any steps to work.

Text highlighted in yellow is code you will need to run (unless in a step that you have established you can skip). You may want to run some other code too. Text in green is test code to check if things are working.

You should log into the biocluster server for the entirety of this assignment.

End goal:

When successfully completed, you will have:

- a folder structure on the storage cluster /seqdata-new with some data that we will be working on during class
- a conda package manager installed, and configured so that you can access the software

Tests to see if everything has worked (it won't yet!):

1. run the command

```
ls /seqdata-new/username/phs675/data/TH_M_20wk_Pancreas/raw_fastqs
```

(Replace username with your username on the server.) You should see a list of 34 files, all ending with the extension .fastq.gz

2. run the command

```
conda activate phs675
```

This should run without an error, and you should see (phs675) at the beginning of your prompt. (You can run

```
conda deactivate
```

to get back to "normal".)

Step 1: creating a folder structure on seqdata-new

Purpose:

1. As mentioned in class, the home folders on the server are restricted for disk space. We will use the data storage cluster /seqdata-new for our work. This ensures you have space to work in.
2. We will need to access some data files and work with them. This also creates a folder structure for working with those files, and copies the raw data files to the folder structure.

Run the following code, **replacing username with your actual biocluster user name**:

```
mkdir -p /seqdata-new/username/phs675/data/TH_M_20wk_Pancreas/raw_fastqs
```

As we saw in class, mkdir makes a new directory. The -p ("parent") option creates "parent" directories where needed. So in this case, if /seqdata-new/username does not exist, it will be created first, then /seqdata-new/username/data, etc.

Navigate to the new folder (replacing username again):

```
cd /seqdata-new/username/phs675/data/TH_M_20wk_Pancreas/raw_fastqs
```

Make sure this has completed successfully (you can use pwd) and then copy the data files to your new folder:

```
cp /seqdata-new/phs675/data/TH_M_20wk_Pancreas/raw_fastqs/* .
```

(The . at the end means "the current folder". It is the target for the file copy, so all the files will be copied to the current folder. Note the space before the .)

This may take a minute or two: you are copying about 10GB of data.

After this step, the first test code highlighted in green at the top of this document should run successfully.

Step 2: Installing and configuring conda

Purpose: We will need conda to get access to a shared environment, in which we have access to the software we will need for the rest of the course.

Background: conda is a package manager. This means it is a piece of software for managing the installation and versioning of other, often interdependent, software. The need for this primarily arises because we are working on a system that is shared by multiple users, who may all need different software installed on the system. In addition, each piece of software often depends on other software, and sometimes on specific versions of that software. This often results in different versions of the same software needing to be installed on the same system. Conda creates *environments*, and software is installed in specific environments. You can activate one environment at a time, which gives you access to the software (and specific version) installed in that environment.

Step 2.1: Installing conda

Run the following:

```
conda list
```

If you get a list of conda packages, beginning with a line that looks like

```
# packages in environment at ...
```

then you already have conda installed and can skip to step 2.2.

If you get a message such as "-bash: conda: command not found" then you need to do the following. Details for this step are at <https://www.anaconda.com/docs/getting-started/miniconda/install/linux-install> but are outlined below

Change to your home directory:

```
cd ~
```

Download the latest version of the miniconda installer:

```
curl -O https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
```

(the option is a capital letter O, not a zero). This will download the conda installer to your home directory. You can check it has downloaded with ls.

Run the installer using

```
bash ./Miniconda3-latest-Linux-x86_64.sh
```

- Press enter to view the user agreement and enter "yes" to accept it.
- Press enter to accept the default installation folder ("prefix"), and wait while the installation completes.
- Enter yes when asked if you want to initialize conda.
- When the installation is complete, exit from the server (exit) and log back in. If everything has worked, you should see (base) at the beginning of your prompt, indicating that you have the base environment activated.

Step 2.2: configuring conda

By default, environments are created in your home directory. As we mentioned previously, the home directory has limited space, and some software (packages) we install might take a lot of space. We are going to configure conda to use a folder in our directories on seqdata-new, and also specify a location that has a shared environment we can all use. This will ensure we all use the same versions of software for the class.

First, we will create some directories for conda to store software and metadata about the software. Run the following, **replacing username** with your actual biocluster user name:

```
mkdir -p /seqdata-new/username/.conda/pkgs
```

```
mkdir -p /seqdata-new/username/.conda/envs
```

The configuration file for conda is a file called .condarc and is located in your home directory. Open the nano editor to edit this file.

(If you have an existing configuration file, the editor will show that existing file. If this is a new file, it will be empty. If you have an existing file, generally you should put the text below *before* the existing text. If some items appear to overlap, it should be fairly obvious what to do, but you can check with an instructor if you need.)

```
nano ~/.condarc
```

Enter the following in the file. **Replace username with your actual biocluster user name.**

pkgs_dirs:

- /seqdata-new/username/.conda/pkgs
- /seqdata-new/phs675/.conda/pkgs

envs_dirs:

- /seqdata-new/username/.conda/envs
- /seqdata-new/phs675/.conda/envs

channels:

- conda-forge
- defaults
- bioconda

Write the file (Control-O, and then enter), and exit the editor (Control-X).

Exit the server with exit, and log back in.

To test if this worked, you can use

```
conda info --envs
```

which lists the environments that conda is aware of. At a minimum (if you had conda installed previously you may have other environments as well) you should see a base environment and a phs675 environment. The location of the phs675 environment should be

/seqdata-new/phs675/.conda/envs/phs675.

At this point, the second test code highlighted in green at the top of the document should work.