Python Introductory Education Series

Alexandra L Emmons Ph.D. & Joe Wu Ph.D.
BTEP/GAU/CCR/NCI/NIH - email ncibtep@mail.nih.gov
Bioinformatics Training and Education Program

Table of Contents

Course overview

Lesson 1 slides

Lesson 1

Lesson 2

Lesson 3

Lesson 4

Starting Jupyter Lab through Tunneling

Practice questions

Lesson 2 practice 78

Lesson 3 practice 80

Lesson 4 practice 83

Finding help

Course Overview

Welcome to the Python Introductory Education Series (PIES) course. This course is composed of four lessons (see schedule below) and is meant to help those with no or limited experience in Python get started using this general purpose scripting language for data analyses. Each one-hour lesson will be followed by an optional one-hour help session. At the end of this course series, participants should

- Have obtained a broad overview of Python, including:
  - Familiarity with tools used to write Python code
  - Knowledge of Python command syntax
  - Ability to find help for Python commands
  - Knowledge of where to find Python packages
  - Familiarity with self-learning resources
- Be able to describe Python data types and structures and provide examples of where some of the data structures are used.
- Know how to use loops and iterators to perform repetitive tasks.
- Know how to work with tabular data
- Be able to construct data visualizations

Lesson schedule:

- Tuesday June 3, 2025, 2 – 3 PM: Getting Started with Python
- Thursday June 5, 2025, 2 – 3 PM: Python Data Types, Variable Assignment, Conditionals, Loops and Iterators
- Tuesday June 10, 2025, 2 – 3 PM: Data Wrangling using Python
- Thursday June 12, 2025, 2 – 3 PM: Data Visualization using Python

A Biowulf account is required for this class. Visit the Biowulf User Dashboard *(https://hpcnihapps.cit.nih.gov/auth/dashboard/)* to unlock an inactive account. For instructions on obtaining a Biowulf account, visit https://hpc.nih.gov/docs/accounts.html *(https://hpc.nih.gov/docs/accounts.html)*.

Example Data

Download data used in this course

Lesson 1 slides

Getting Started with Python

Joe Wu, PhD
NCI/CCR Bioinformatics Training and Education Program
ncibtep@nih.gov

Lesson 1 Learning Objectives

After this class, participants will be able to:

- Describe Python and provide rationale for using Python
- List tools for interacting with Python
- Sign onto Biowulf, start a Jupyter Lab session, and become familiar with the Jupyter Lab interface.
- Describe Python command syntax
- Describe where to get and how to install external packages
- Get help for Python commands

Why use Python?

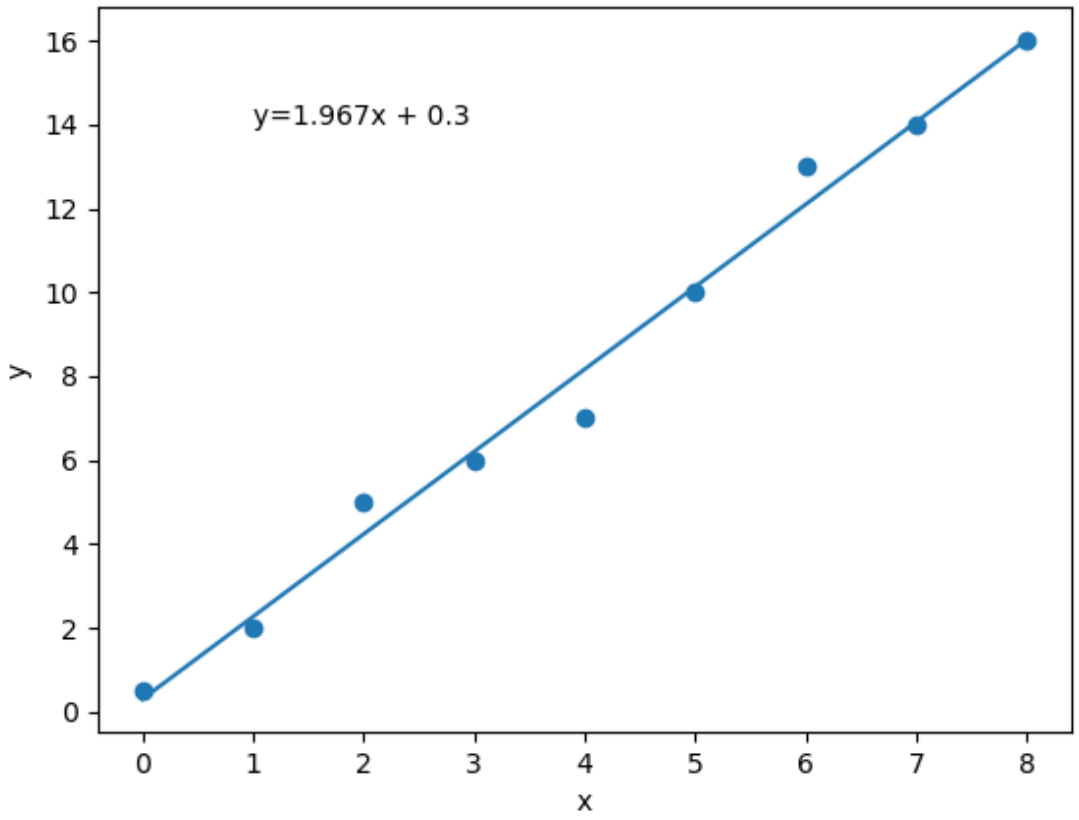
- General purpose scripting language
    - Analyze and visualize large datasets
    - Reusability and reproducibility
    - Versioning and keeping track of changes is possible when analyzing data using scripts
    - Easy to learn
- External packages that enhances functionality
    - Python Package Index *(https://pypi.org)*
    - Anaconda *(https://www.anaconda.com/)*
    - Biopython *(https://biopython.org)*
- Large community support

Python enables Elegant Data Visualization

An abundance of external packages make scientific computing and data presentation easy. For instance, the packages matplotlib *(https://matplotlib.orghttps://matplotlib.org)* and seaborn *(https://seaborn.pydata.org/)* are good tools for generating data visualizations. With a few lines of code, scientists can generate scatter plots to view relationship between variables and/or heatmaps that can reveal distinct clusters in a dataset.

Generating a Scatter Plot using Matplotlib

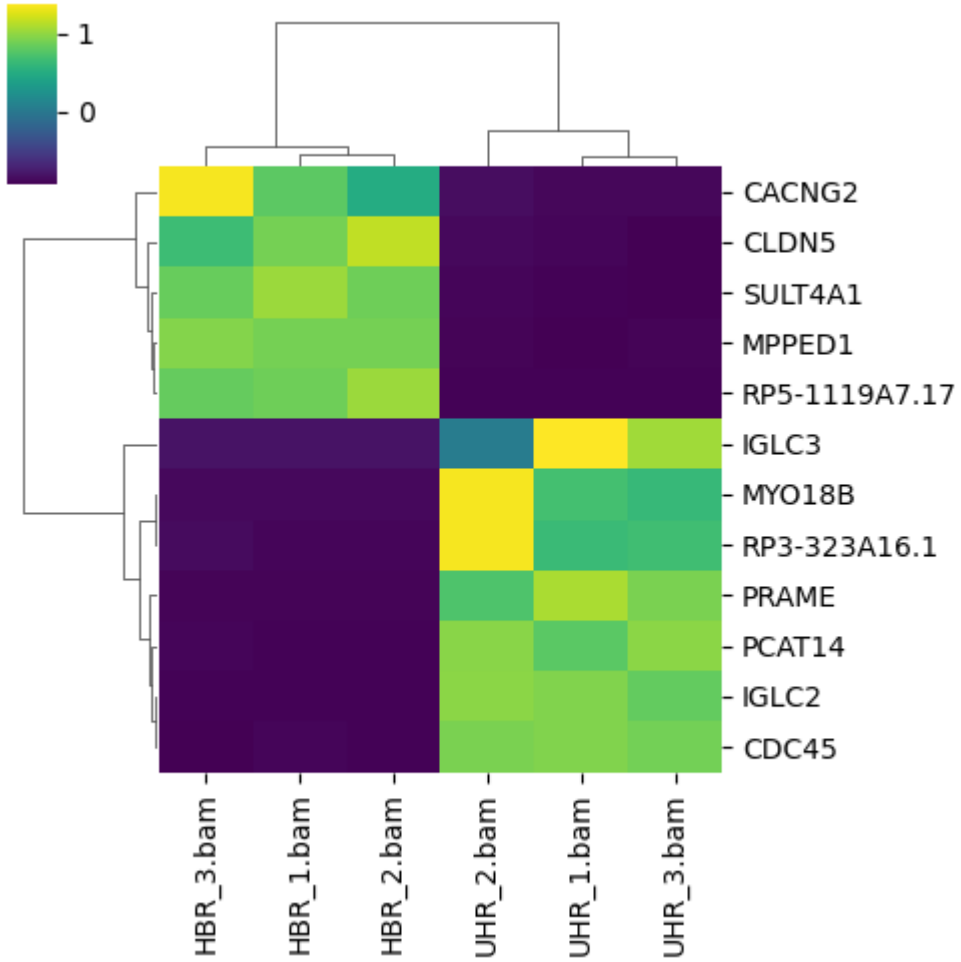
```
import matplotlib.pyplot as plt
import numpy

x=numpy.array([0,1,2,3,4,5,6,7,8])
y=numpy.array([0.5,2,5,6,7,10,13,14,16])
plt.scatter(x,y)
slope, intercept=numpy.polyfit(x,y,1)
plt.plot(x,slope*x+intercept)
plt.text(1,14,'y='+str(round(slope,3))+'x' ' + ' + str(round(intercep
plt.xlabel('x')
plt.ylabel('y')
```



Generating a Gene Expression Heatmap using Seaborn

```
import pandas
import seaborn
import matplotlib.pyplot as plt
counts1=pandas.read_csv("./hbr_uhr_top_deg_normalized_counts.csv", ir
```

```
seaborn.clustermap(counts1,z_score=0,cmap="viridis", figsize=(5,5))
plt.suptitle("Gene expression heatmap",y=1.1)
plt.show()
```



Tools for Interacting with Python

- Python can be run at the command prompt
- Ipython *(https://ipython.org)*
- Run python script at the command prompt
- Integrated Development Environments such as:
    - Spyder *(https://www.spyder-ide.org/)*
    - Pycharm *(https://www.jetbrains.com/pycharm/)*
- Visual Studio Code from Microsoft has extensions that support Python scripting
- R Studio
- Juptyer Lab/Notebook

Python at the Command Prompt

Assuming Python is installed, just type `python` at the command prompt to start using Python. Hit control-d to exit back to the command prompt. The downside to this is that users cannot save the commands into a script.

```
(base) [wuz8@cn4303 pies_data]$ python
Python 3.12.10 | packaged by conda-forge | (main, Apr 10 2025, 22:21:13) [GCC 13.3.0] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>> print("hello")
hello
>>> import numpy as np
>>> print(np.pi)
3.141592653589793
>>> print(np.sqrt(25))
5.0
>>> _
```

Ipython

Ipython *(https://ipython.org)* enables users to run Python commands interactively at the terminal. It features autocomplete of commands and allows for saving of commands to a python script using `%save` followed by the name of the script.

```
(base) [wuz8@cn4303 pies_data]$ ipython
Python 3.12.10 | packaged by conda-forge | (main, Apr 10 2025, 22:21:13) [GCC 13.3.0]
Type 'copyright', 'credits' or 'license' for more information
IPython 9.1.0 -- An enhanced Interactive Python. Type '?' for help.
Tip: IPython supports combining unicode identifiers, eg F\vec<tab> will become F⃗, useful for physics equations. Play with \dot \d
dot and others.

In [1]: print("hello")
hello

In [2]: import numpy as np

In [3]: print(np.pi)
3.141592653589793

In [4]: print(np.sqrt(25))
5.0

In [5]: %save pies_class_2025_ipython.py
The following commands were written to file `pies_class_2025_ipython.py`:
print("hello")
import numpy as np
print(np.pi)
print(np.sqrt(25))
```

Hit `control-d` to exit Ipython and return to the command prompt.

While using Ipython is better than just running commands on the terminal, it still is not very efficient in terms of saving work. Also, users will not be able to view plots on HPC systems such as Biowulf since these do not support inspection of graphical outputs.

> **Note**
>
> The `pies_class_2025_ipython.py` script can be run from the command line. To run a Python script from command line, just do `python` followed by name of the script. Python scripts can also be submitted as a job to the Biowulf batch system.

```
python pies_class_2025_ipython.py
```

```
hello
3.141592653589793
5.0
```

Using Python through IDE

Integrated Development Environments or IDEs are ideal for scripting in Python as well as other languages. See https://ritza.co/comparisons/pycharm-vs-spyder-vs-jupyter-vs-visual-studio-vs-anaconda-vs-intellij.html *(https://ritza.co/comparisons/pycharm-vs-spyder-vs-jupyter-vs-visual-studio-vs-anaconda-vs-intellij.html)* for a breakdown of of common ones such as Spyder, Pycharm, VS Code, R Studio, and Jupyter Lab. Essentially, IDE enable users to write scripts, access as well as view data, and view plots. Some IDEs enable users to generate analysis report that details steps of an analysis as well as the tool and the code use.

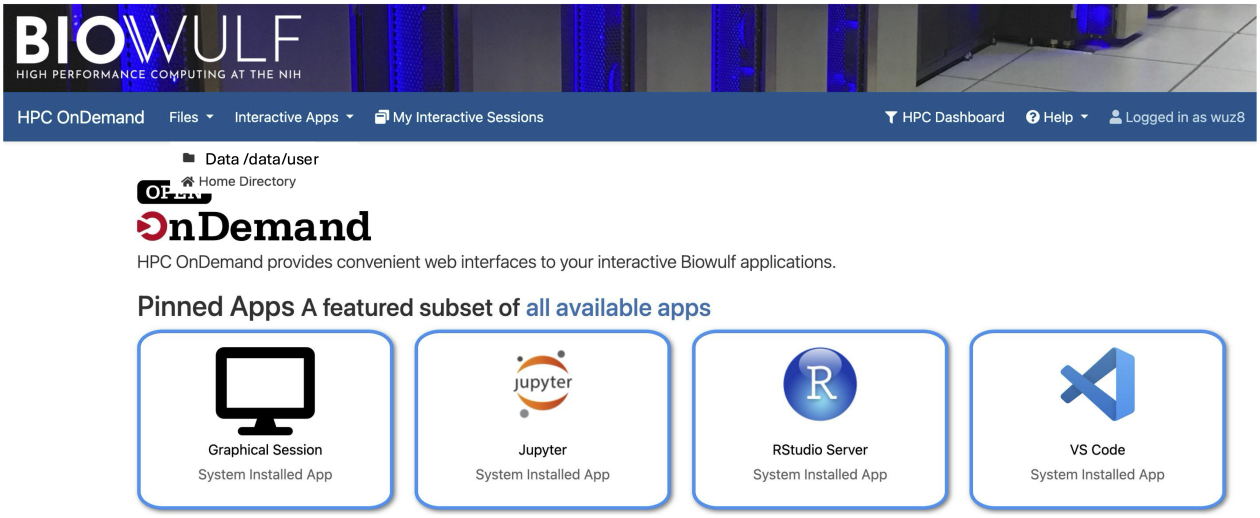
Accessing Python at NIH

- Biowulf (HPC OnDemand *(https://hpcondemand.nih.gov/)* is recommended).
- Use Python locally on government furnished personal computer via NIH Anaconda Professional License *(https://nih.sharepoint.com/sites/CIT-ApplicationRepository/SitePages/Anaconda.aspx)*. This will require users to install Anaconda to local computer. See BTEP's Topic Spotlight on NIH's Anaconda Professional license *(https://bioinformatics.ccr.cancer.gov/btep/getting-started-with-an-nih-anaconda-business-license/)* to learn more.
- NCI scientists also can use Python through Posit Workbench. Fill out the form at https://forms.office.com/pages/responsepage.aspx?id=eHW3FHOX1UKFByUcotwrBnYgWNrH6QdOsCsoiQ9eiaZUQ1ZZODJKT0FERUdHOVZYUkJaMzA2 *(https://forms.office.com/pages/responsepage.aspx?id=eHW3FHOX1UKFByUcotwrBnYgWNrH6QdOsCsoiQ9eiaZUQ1ZZODJKT0FERUdHOVZYUkJaMzA2* to request access.

Signing onto Biowulf HPC OnDemand

- Open a web browser on local computer (Google Chrome is recommended) and go to https://hpcondemand.nih.gov/ *(https://hpcondemand.nih.gov/)*, which is the URL for Biowulf's HPC OnDemand.
- Once at HPC OnDemand, sign in with participant's NIH PIV card credentials.

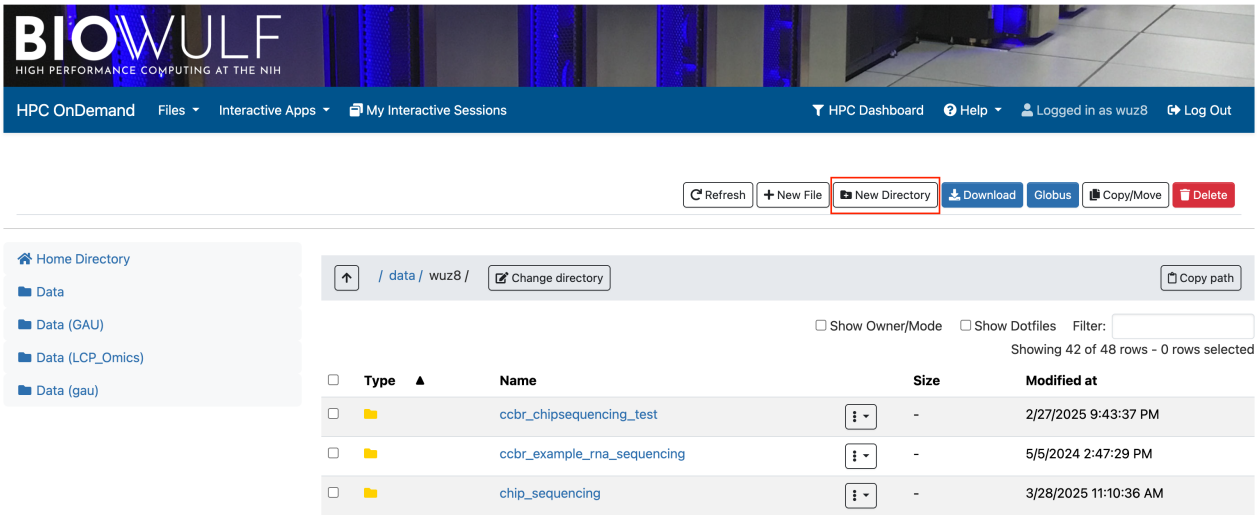
- After signing in, users will see links to applications available through HPC OnDemand such as Jupyter. User's Biowulf file system can be accessed via OnDemand as well (just click Files in the menu bar).



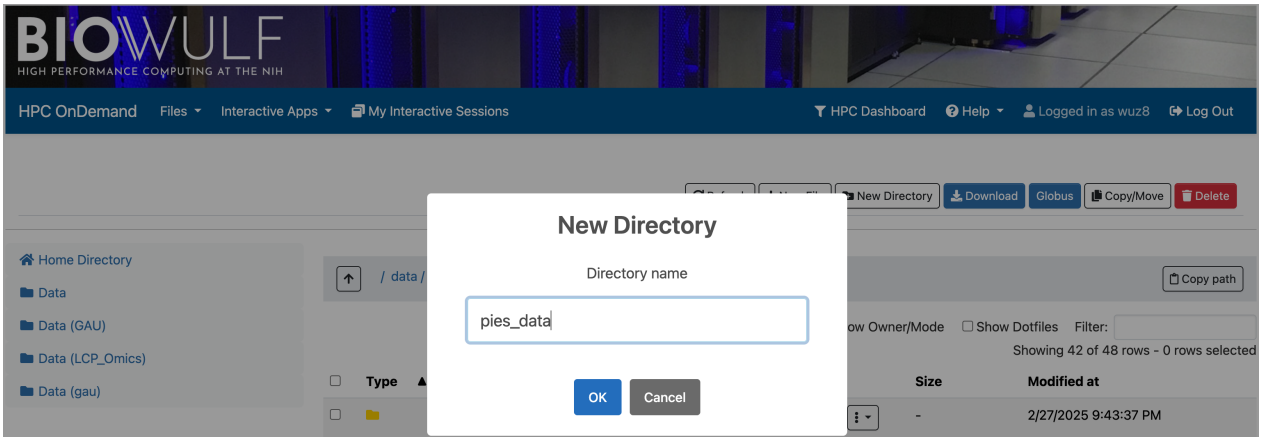
Get the Example Data

Goto the course overview section in the class documents and scroll to the bottom and click on "Download data used in this course" to download some example data to local computer. Take note of where it downloads, but typically it should go into the user's local computer `Downloads` folder. The data comes as zip file. While Macs will automatically unzip, Windows users may need to right click on the file to uncompress it. After the download and uncompressing is finished, participants will see a folder called `pies_data`.

Next, click "Files" in the HPC OnDemand menu and choose the folder labeled `/data/user`, where user is the participant's Biowulf user ID. The subsequent page will show the content (ie. files and folders) of the participant's Biowulf `data` directory. Click on the "New Directory" tab to make a to make a folder to store the example data and Jupyter Notebook for this course series.

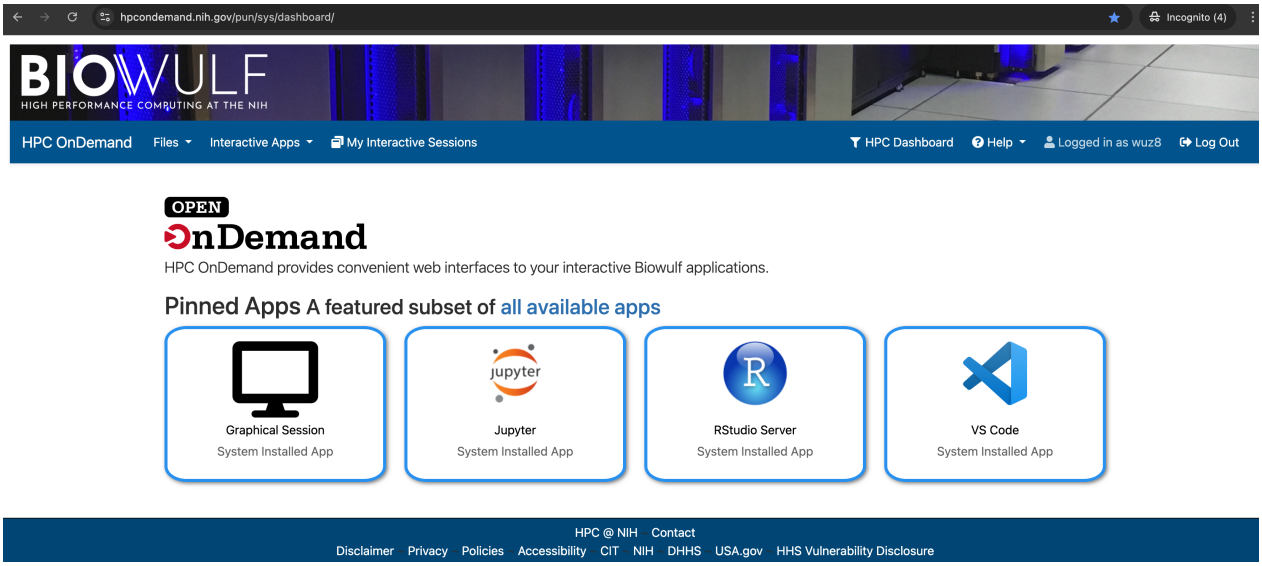


Name the class directory `pies_data`. Click "Ok" when ready.

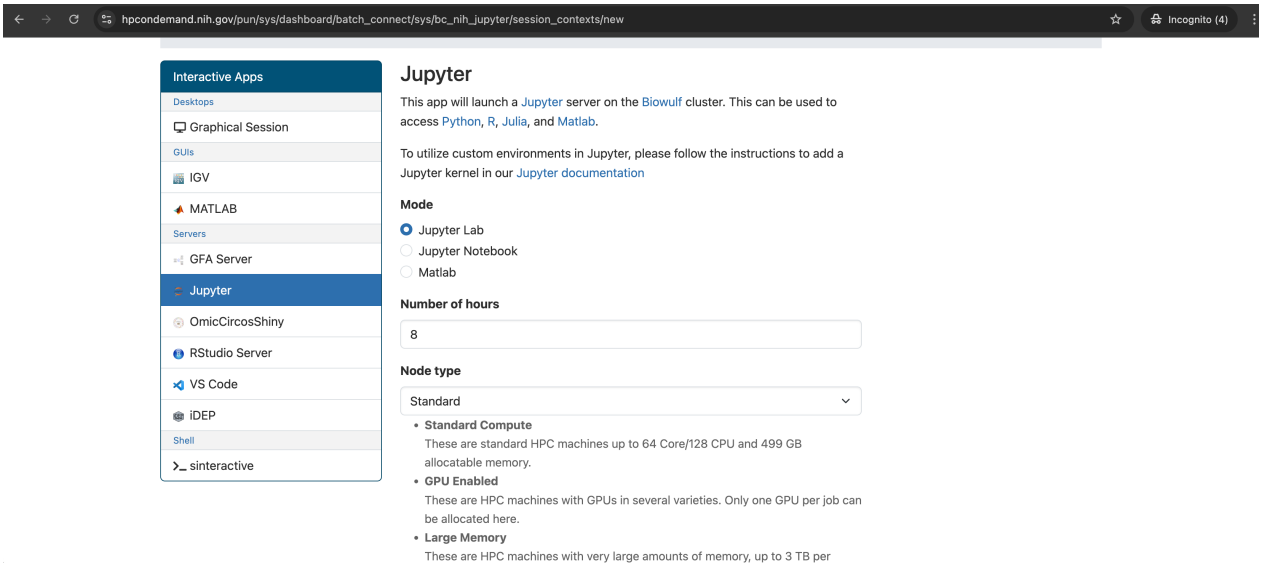


Start a Jupyter Lab Session

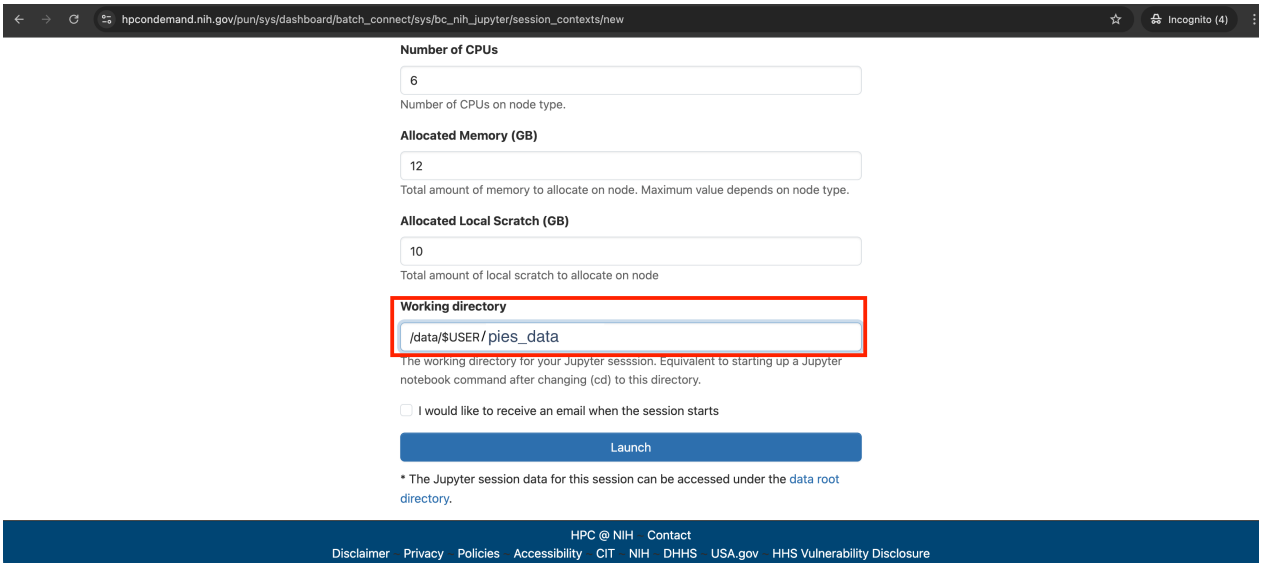
Navigate back to the HPC OnDemand website by clicking "HPC OnDemand" at the top left corner. Then click on the "Jupyter" tab to launch a Jupyter Lab session.



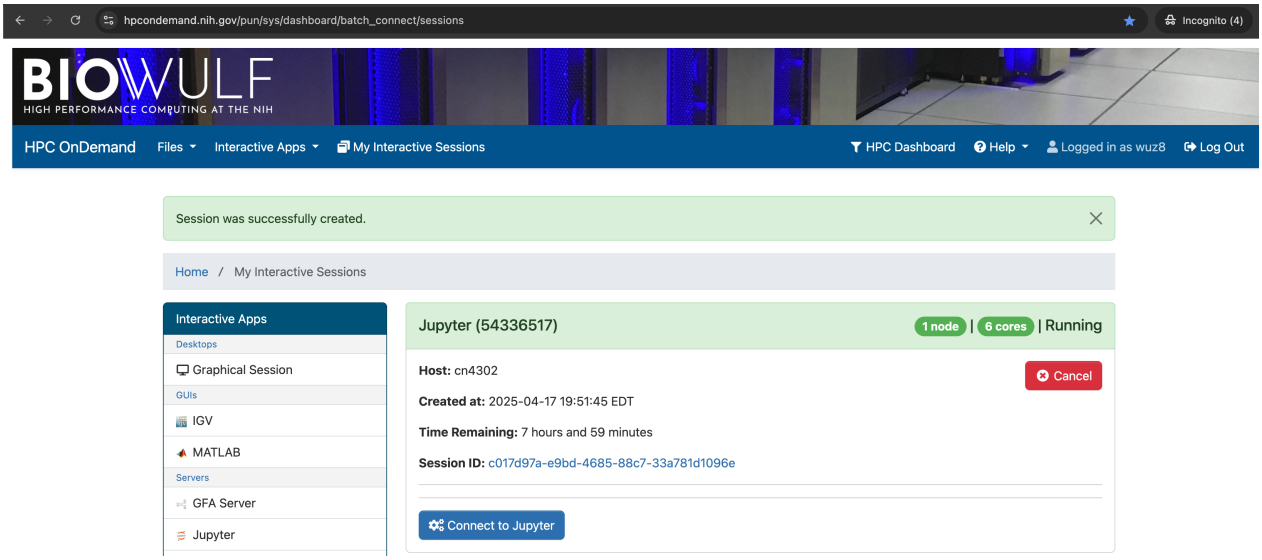
- The subsequent page allows users to specify compute resources. Leave these as is for this class. Also, be sure that the radial button for Jupyter Lab is selected.



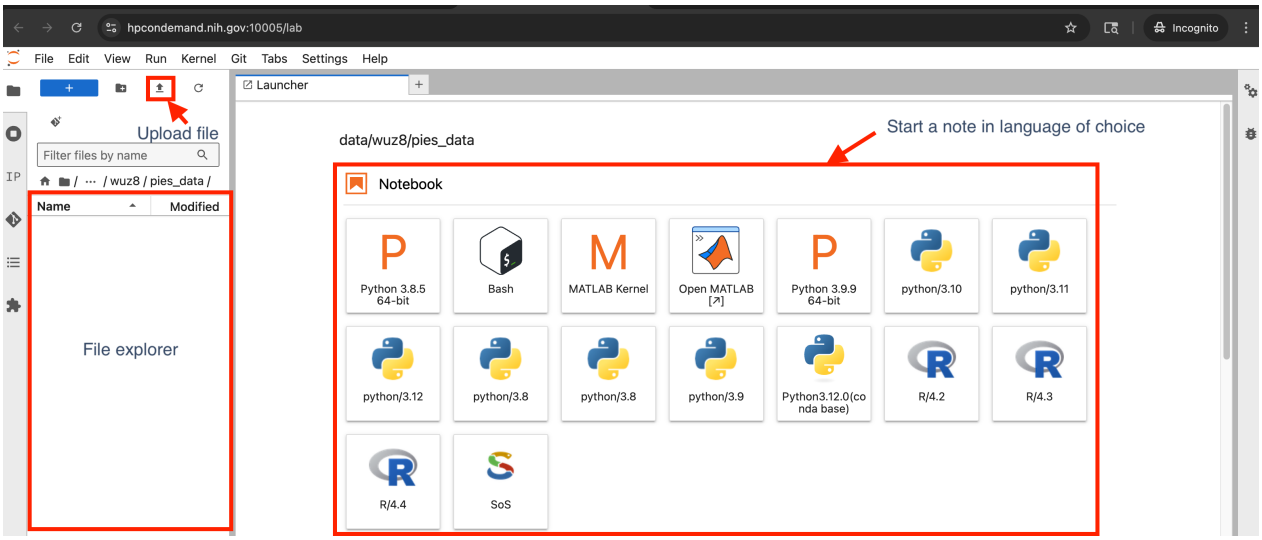
- Make sure to specify for Jupyter to start in the `/data/$USER/pies_data` directory, where $USER is a variable that points to the participants Biowulf user ID.



Click on "Connect to Jupyter" when the Jupyter Lab session has been granted.



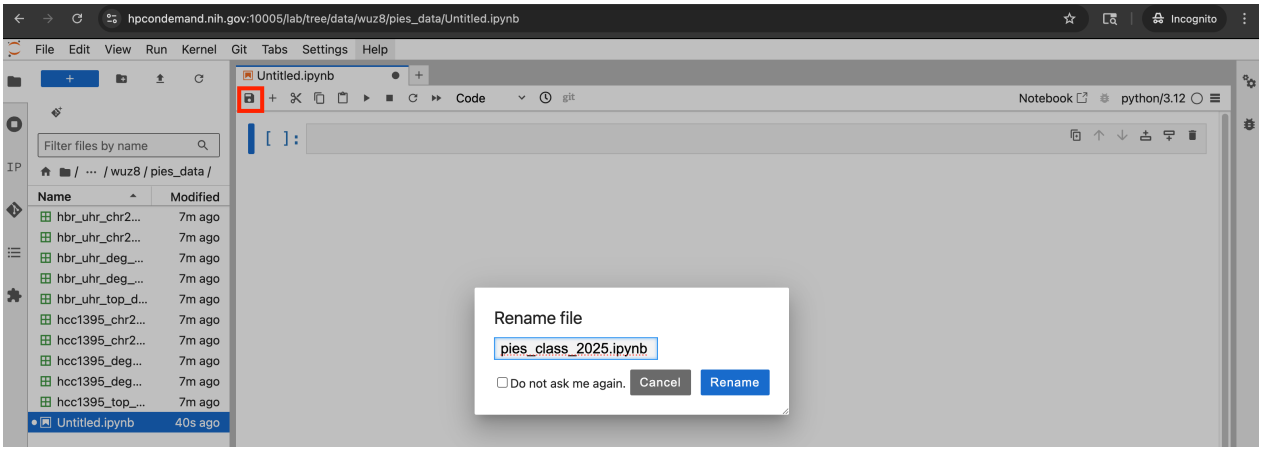
Users will see an interface that looks like the image shown below. The left hand panel is the file explorer. Users can navigate through files and folders that are available in the directory in which Jupyter Lab was started. The launcher panel contains quick links for initiating a Jupyter Notebook in the user's language of choice.



Note the file explorer is empty. The next step then will be to open the `pies_data` folder on the participants local `Download` directory and select all of the `csv` files in the folder and drag and drop in to the Jupyter Lab file explorer or use the upload button.

Create a new Jupyter Notebook

Create a new Jupyter Notebook in Python 3.12 (click on the "python/3.12" tile). The new notebook has the name "Untitled.ipynb". Click on the disk icon in the notebook menu bar to rename it pies_class_2025.



Tip

For a detailed overview of Jupyter Lab, see BTEP's Documenting Analysis Steps using Jupyter Lab *(https://bioinformatics.ccr.cancer.gov/docs/analysis-documentation-jupyter/index.html)*

Python Command Syntax

Arguments and options for Python commands are enclosed in parentheses. In general, the anatomy is `command(argument, option)`.

For example, the command below is `print` and it will display the argument, "Hello BTEP" as output.

```
print("Hello BTEP")
```

```
Hello BTEP
```

To get help for a Python command, use `help`.

For instance:

```
help(print)
```

```
Help on built-in function print in module builtins:

print(*args, sep=' ', end='\n', file=None, flush=False)
    Prints the values to a stream, or to sys.stdout by default.

    sep
      string inserted between values, default a space.
    end
      string appended after the last value, default a newline.
    file
      a file-like object (stream); defaults to the current sys.std
out.
    flush
      whether to forcibly flush the stream.
```

From the `print` command's help information, line breaks can be added using `\n`. Try the following to print three sentences, one in each line.

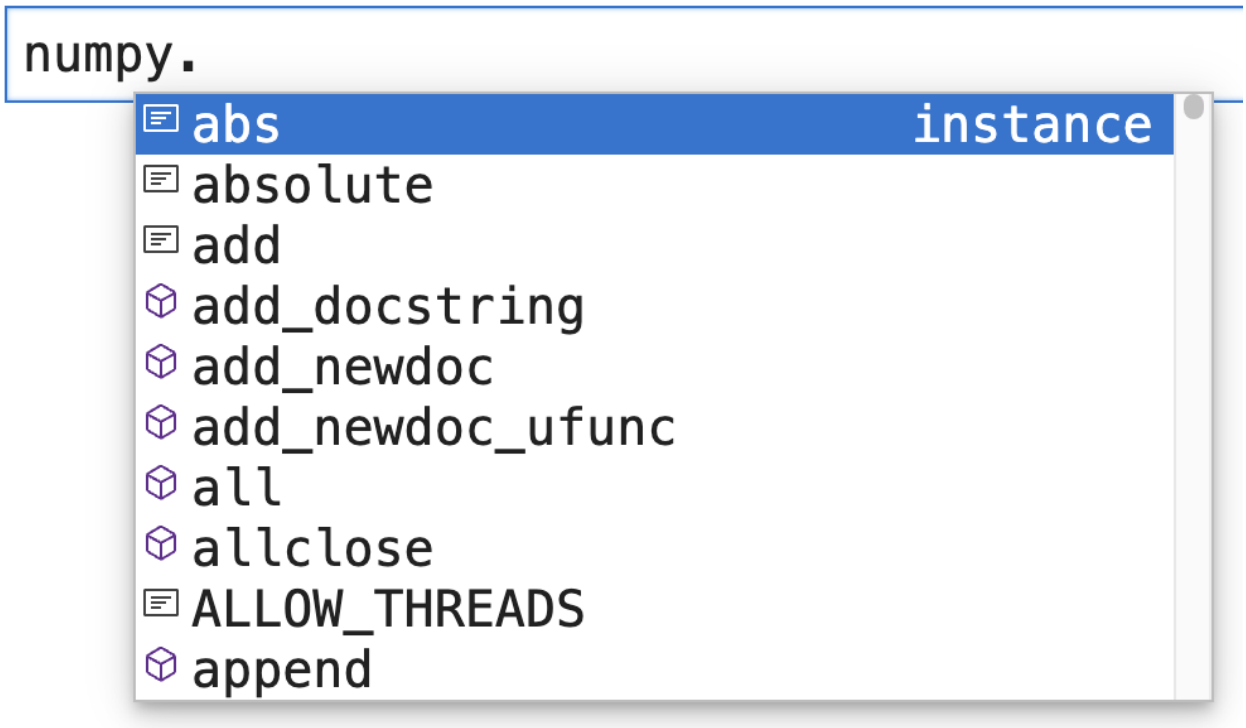
```
print("Python can make data analysis more efficient.\n"
"It helps with reusability and reproducibility.\n"
"There is strong community support.\n"
"External packages are available for data wrangling and visualizatio
```

```
Python can make data analysis more efficient.
It helps with reusability and reproducibility.
There is strong community support.
External packages are available for data wrangling and visualization
```

> What is different with `numpy.array` used in the earlier example to generate numeric arrays?

{{Sdet}}{{Ssum}}Answer{{Esum}}

`numpy` is a Python package that has many subcommands. To call a subcommand from a package, use the general syntax of `package.subcommand`.



{{Edet}}

`numpy` has a subcommand `divide`. How can that be called?

{{Sdet}}{{Ssum}}Answer{{Esum}}

`numpy.divide`

{{Edet}}

What does the `divide` subcommand from `numpy` do?

{{Sdet}}{{Ssum}}Answer{{Esum}}

`help(numpy.divide)`

{{Edet}}

Installing external packages

Python external packages are found at the Python Package Index *(https://pypi.org)*. To install a package from PyPi, just use `pip install package_name`, where package_name can be any package of choice. For instance, to install scipy, do:

```
pip install scipy
```

Note

Package management for Python needs to be done at the terminal.

`pip` is the package installer for Python. If `pip` is not available with the user's Python installation, see https://pip.pypa.io/en/stable/installation/ *(https://pip.pypa.io/en/stable/installation/)* to learn how to get it.

To uninstall a package, do `pip uninstall package_name`.

To update a package, use `pip install --upgrade package_name`.

`pip freeze` will pull up a list of currently installed Packages installed via `pip`. To find if a specific package is installed do `pip freeze | grep package_name`.

Those who chose to use the package manager Anaconda can install via the command line using `conda install package_name`. Again, package_name is the user's package of choice. Package managers offer the benefit of reducing issues that arise from versioning, dependency, and security when installing software. See https://docs.conda.io/projects/conda/en/stable/user-guide/tasks/manage-pkgs.html *(https://docs.conda.io/projects/conda/en/stable/user-guide/tasks/manage-pkgs.html)* to learn more about installing, updating, and uninstalling packages using Conda. For working locally on government furnished personal computer, researchers are recommended to use the NIH Anaconda Professional License *(https://nih.sharepoint.com/sites/CIT-ApplicationRepository/SitePages/Anaconda.aspx)*. Biowulf also has a guide on managing Anaconda environments on the cluster. See https://hpc.nih.gov/docs/diy_installation/conda.html *(https://hpc.nih.gov/docs/diy_installation/conda.html)*.

Python Data Types, Loops and Iterators

Learning Objectives

After this class, participants will

- Be able to describe Python data types and structures
- Become familiar with variable assignment
- Be able to use conditional operators and if-else statements
- Understand how loops and iterators can be used automate processes
- Be able to load packages
- Know how to import tabular data
- Know how to view tabular data

Start a Jupyter Lab session

Before getting started, make sure to start a Jupyter Lab session with the default resources via HPC OnDemand *(https://hpcondemand.nih.gov/pun/sys/dashboard/)*.

> **Hint**
>
> Be sure to start the Jupyter Lab session in `/data/$USER/pies_data'. Where $USER is the environmental variable that points to the participant's Biowulf user ID.

Next, click on `pies_class_2025.ipynb` in the file explorer to open the Jupyter Notebook used for this class.

Python Data Types and Data Structures

An important step to learning any new programming language and data analysis is to understand its data types and structures. Common data types and structures that will be encountered include the following.

- Text (str)
- Numeric
    - int (ie. integers)
    - float (ie. decimals)
- Boolean (True or False)
    - conditionals
    - filtering criteria
    - command options

- Data frames
- Lists
- Arrays
- Tuples
- Range
- Dictionaries

Identifying Data Type and Structure in Python

The command `type` can be used to identify data types and structures in Python.

```
type(100)
```

This will return `int` for integer as 100 is an integer.

```
int
```

```
type(3.1415926)
```

This will return `float` as 3.1415926 has decimals.

```
float
```

```
type("bioinformatics")
```

This will return `str` for string as the word bioinformatics is a text string.

```
str
```

Variable Assignments

In Python, variables are assigned to values using "=".

```
test1_score=100
test1_score
```

```

```

```
100
```

```
mole=6.02e23
mole
```

```
6.02e+23
```

```
btep_class="Python Introductory Education Series"
btep_class
```

```
'Python Introductory Education Series'
```

The command `type(btep_class)` will return `str` because the variable `btep_class` is a text string.

```
type(btep_class)
```

```
str
```

It is also possible assign a variable to another variable.

```
test2_score=test1_score
test2_score
```

```
100
```

Change the value of test2_score to 60.

```
test2_score=60
```

```
test2_score
```

```
60
```

```
test1_score
```

```
100
```

```
print("The student got a", test2_score, "on exam 2.")
```

Definition

Immutable objects in Python are variables whose values cannot be changed after they have been created. This includes integers, floats, strings, and tuples. In the above example, test2_score was initially set to test1_score. However, upon changing test2_score to 60, the value of test1_score does not change. Thus, demonstrating that integers are immutable.

Conditionals

Conditionals evaluate the validity of certain conditions and operators include:

- ==: is equal to?
- >: is greater than?
- >=: is greater than or equal to?
- <: is less than?
- <=: is less than or equal to?
- `!=`: is not equal to?
- `and`
- `or`

The command below will evaluate if test1_score is equal to test2_score.

```
test1_score==test2_score
```

Because test1_score is 100 and test2_score is 60, the result from the above command will be false.

```
False
```

> If the expression of gene A is 25 and gene B is 100 as obtained from bulk RNA sequencing, how would you test if gene B has a higher expression value than gene A using conditionals in Python?

{{Sdet}}{{Ssum}}Answer{{Esum}}

```
gene_a=25
gene_b=100
gene_b > gene_a
```

```
True
```

{{Edet}}

If statements are also conditionals and are used to instruct the computer to do something if a condition is met. To have the computer do something when the condition is not met, use `elif` (else if) or `else`.

The command below will accomplish the following:

- Use `if` to evaluate if test1_score>=90, if yes then indicate using `print` that someone got an A!
- Use `elif` (which stands for else if) to evaluate if test2_score>=80, if yes then use the `print` statement to indicate that someone does not have to take the final!
- Finally, `else` will print for all other conditions that someone failed the class.

```
if test1_score>=90:
    print("You get an A!")
elif test2_score>=80:
    print("You don't have to take the final!")
else:
    print("You failed the class!")
```

Tip

The `print` command can be used to print variables by not enclosing in quotes.

A ":" is required after `if`, `elif`, and `else`. The command(s) to execute when conditions are met are placed on a separate line but tab indented.

Data Frames

Often, in bioinformatics and data science, data comes in the form of rectangular tables, which are referred to as data frames. Data frames have the following property.

- Study variable(s) form the columns
- Observation(s) form rows

- Can have a mix of data types (strings and numeric) between columns but **each column/ study variable can contain only one data type**
- Limited to one value per cell

A popular package for working with data frames in Python is Pandas *(https://pandas.pydata.org)*.

To load a Python package use the `import` command followed by the package name (ie. pandas).

```
import pandas
```

Sometimes the name of the package is long, so users might want to shorten it by creating an alias. The alias `pd` is often used for the Pandas package. To add an alias, just append `as` followed by the user defined alias to the package import command.

```
import pandas as pd
```

Importing Tabular Data with Pandas

This exercise will use the `read_csv` function of Pandas to import a comma separated value (csv) file called `hbr_uhr_chr22_rna_seq_counts.csv`, which contains RNA sequencing gene expression counts from the Human Brain Reference (hbr) and Universal Human Reference (uhr) study *(https://rnabio.org/module-01-inputs/0001/05/01/RNAseq_Data/)*.

```
hbr_uhr_chr22_counts=pandas.read_csv("./hbr_uhr_chr22_rna_seq_counts
```

> **Note**
>
> If a Python package was imported using an alias (ie. `pd` for Pandas) then use the alias to call the package. For instance, `pd.read_csv` rather than `pandas.read_csv` when the `pd` alias is used for Pandas.

Take note of the way the csv import command is constructed. First the user specifies the name of package (ie. pandas) and then the function within the package (ie. read_csv). The package name and function name is separated by a period.

Next, use `type` to find out the data type or structure for hbr_uhr_chr22_counts.

```
type(hbr_uhr_chr22_counts)
```

```

```

```
pandas.core.frame.DataFrame
```

Take a look a the first few rows of hbr_uhr_chr22_counts.

```
hbr_uhr_chr22_counts.head()
```

	Geneid	HBR_1.bam	HBR_2.bam	HBR_3.bam	UHR_1.bam	UHR_2.bam	UHR_3.bam
0	U2	0	0	0	0	0	0
1	CU459211.1	0	0	0	0	0	0
2	CU104787.1	0	0	0	0	0	0
3	BAGE5	0	0	0	0	0	0
4	ACTR3BP6	0	0	0	0	0	0

Figure 1: Example of a data frame.

Because `hbr_uhr_chr22_counts` is a Pandas data frame, it is possible to append one of the many Pandas commands to it. For instance, the `head` function was appended to display the first five rows of hbr_uhr_chr22_counts. The data frame name and function is separated by a period. This is perhaps one of the most appealing aspects of Python syntax. Note that the `head` function was followed by `()`. If the parentheses is blank, then by default the first five lines will be shown. There will be more examples of the Pandas `head` function in a subsequent lesson.

Lists and Tuples

Lists and tuples are one dimensional collections of data. The tuple is an immutable list, in which the elements cannot be modified. However, lists are mutable.

To create a list, enclose the contents in square brackets.

```
sequencing_list=["whole genome", "rna", "whole exome"]
```

To create a tuple, enclose the contents in parentheses.

```
sequencing_tuple=("whole genome", "rna", "whole exome")
```

Lists and tuples are indexed and can contain duplicates. The first item in a list or tuple has an index of 0 (Python uses a 0 based indexing), the second item has an index of 1, and the last item has an index of n-1 where n is the number of items. Indices can be used to recall items in a list or tuple.

```
sequencing_list[1]
```

```
'rna'
```

What if users wanted to extract the first two items in sequencing list?

```
sequencing_list[0:2]
```

```
['whole genome', 'rna']
```

But will the following work?

```
sequencing_list[0,1]
```

No, there is an error. More on this in section that covers loops and iterators.

```
TypeError                                 Traceback (most recent cal
Cell In[61], line 1
----> 1 sequencing_list[0,1]

TypeError: list indices must be integers or slices, not tuple
```

List versus tuples (mutable versus immutable)

```
sequencing_list[1]="single cell RNA"
```

```
sequencing_list
```

```
['whole genome', 'single cell RNA', 'whole exome']
```

```
sequencing_tuple[1]="single cell RNA"
```

```
TypeError                                 Traceback (most recent cal
Cell In[48], line 1
```

```
----> 1 sequencing_tuple[1]="single cell RNA"

TypeError: 'tuple' object does not support item assignment
```

Adding and removing from a list

Suppose there is a list called `states`.

```
states=["florida", "alabama", "wisconsin", "tennessee"]
```

To add another state as the last entry in the list, used the `append` attribute for a list.

```
states.append("kentucky")
states
```

```
['florida', 'alabama', 'wisconsin', 'tennessee', 'kentucky']
```

To remove an item from a list, use `remove` attribute.

```
states.remove("wisconsin")
states
```

```
['florida', 'alabama', 'tennessee', 'kentucky']
```

The `remove` attributes removes the first occurrence of an item from a list. See https://www.w3schools.com/python/python_lists_remove.asp *(https://www.w3schools.com/python/python_lists_remove.asp)* for methods on removing items from a list by index.

To add multiple items at the end of a list, use the `extend` attribute.

```
states.extend(["texas", "missouri"])
states
```

```
['florida', 'alabama', 'tennessee', 'kentucky', 'missouri', 'texas',
```

Making a Copy of a List

Suppose there is a list called `list1` that contains the following numbers.

```
list1=[1,2,3,4,5]
list1
```

```
[1, 2, 3, 4, 5]
```

Next, create copy of list1 was made and assigned to variable list2.

```
list2=list1
list2
```

```
[1, 2, 3, 4, 5]
```

Then insert 0 as the first item in list2.

```
list2.insert(0,0)
list2
```

```
[0, 1, 2, 3, 4, 5]
```

When assigning list2 to list1 using =, Python will point list2 to the values stored in list1 (ie. list1 and list2 are referencing the same list). Because lists are mutable, the changes to list2 are reflected in list1 as well.

```
[0, 1, 2, 3, 4, 5]
```

Set list1 back to [1,2,3,4,5].

```
list1=[1,2,3,4,5]
```

Next, use the `deepcopy` module from the Python package `copy` to make a copy of list1 called list2. To call a module within a Python package follow this general syntax of `package.module`. For instance, to call `deepcopy` use `copy.deepcopy`.

```
import copy
list2=copy.deepcopy(list1)
list2
```

Set the first element of list2 to 0.

```
list2.insert(0,0)
list2
```

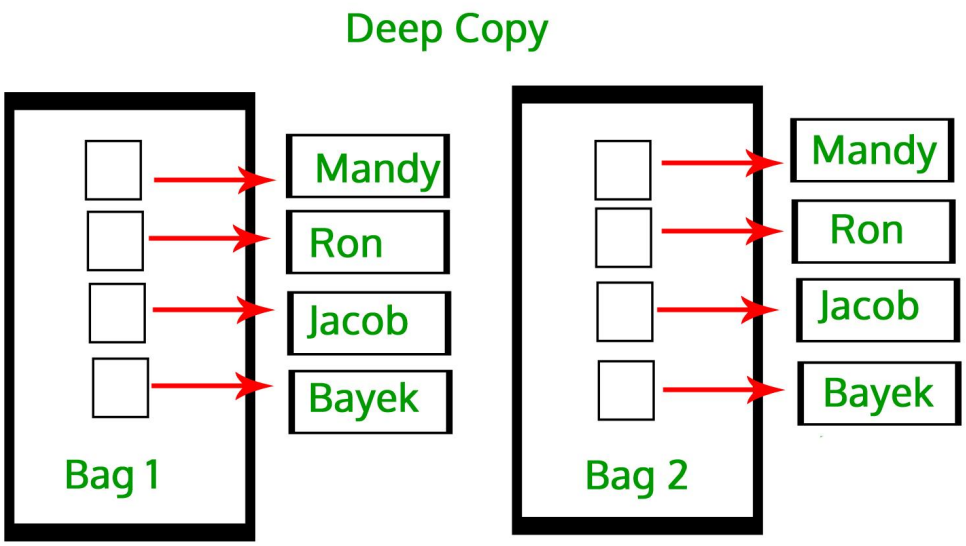
```
[0, 1, 2, 3, 4, 5]
```

Finally, recall list1.

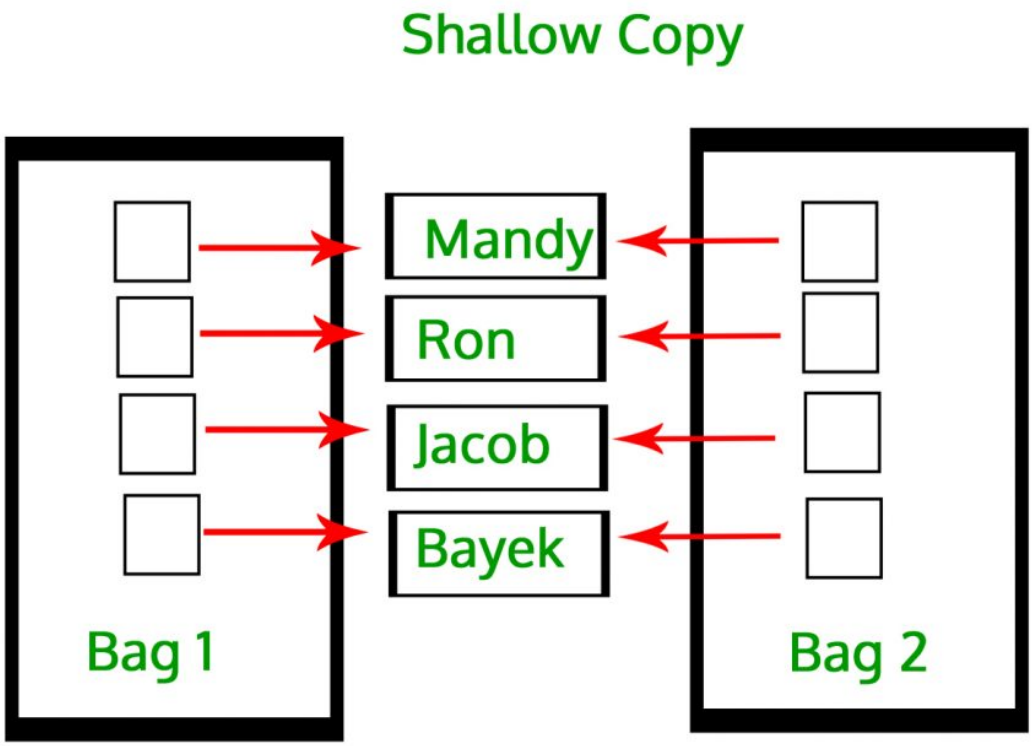
```
list1
```

```
[1, 2, 3, 4, 5]
```

There actually two types of copies in Python. One is called shallow copy and the other is deep copy. To create a shallow copy of list1 and store it as list2, just do `list2=list1.copy()`. However, caution still need to taken when shallow copying as this could also lead to unintended changes to the original variable. To create an independent copy of a variable, use deep copy. See https://www.geeksforgeeks.org/copy-python-deep-copy-shallow-copy/# *(https://www.geeksforgeeks.org/copy-python-deep-copy-shallow-copy/#)* to learn more.



Source: https://www.geeksforgeeks.org/copy-python-deep-copy-shallow-copy/# *(https://www.geeksforgeeks.org/copy-python-deep-copy-shallow-copy/#)*



Source: https://www.geeksforgeeks.org/copy-python-deep-copy-shallow-copy/# *(https://www.geeksforgeeks.org/copy-python-deep-copy-shallow-copy/#)*

Instructions for modifying Python lists can be found at the W3 school *(https://www.w3schools.com/python/python_lists.asp)*

Arrays

Given a list of numbers, it is difficult to perform mathematical operations. For instance

```
list_of_numbers=[1,2,3,4,5]
```

Multiplying list_of_numbers by 2 will duplicate this list. However, multiplying a list of numbers by two should double every number in that list. Thus, the expected result is [2,4,6,8,10]. To resolve this, convert the list to an array using the package numpy *(https://numpy.org)*.

```
list_of_numbers*2
```

```
[1, 2, 3, 4, 5, 1, 2, 3, 4, 5]
```

Use the `array` function of numpy to convert list_of_numbers to an array called array_of_numbers.

```
import numpy
```

```
array_of_numbers=numpy.array(list_of_numbers)
```

```
array_of_numbers*2
```

```
array([ 2,  4,  6,  8, 10])
```

The array of numbers shown here is a one dimensional array. A special case of arrays is the matrix, which is two dimensional. Like data frames, matrices store values in columns and rows. Matrices are encountered in computation and are used to store numeric values (see here for more on matrices *(https://youtu.be/lZcyZHomFQc)*).

Loops and Iterators

Loops and iterators are great for performing repeated tasks. In Python, users will see `for` and `while` loops. To learn about loops, first recreate `sequencing_list`.

```
sequencing_list=["whole genome", "rna", "whole exome"]
```

Then add a few more items the sequencing_list. To add multiple items to Python lists, just use the `.extend` attribute.

```
sequencing_list.extend(["chip", "atac"])
sequencing_list
```

```
['whole genome', 'rna', 'whole exome', 'chip', 'atac']
```

The following `for` loop will print elements with index 2, 3, and 4 from sequencing_list and can be explained as follows.

- `for` is a type of loop to iterate over repetitive tasks in Python. To use the `for` loop,
  - An index is needed to keep track of where in the repetitive task the loop is in. For instance, this index can inform the loop which item in a list that it is currently performing a task on. The index can be named anything. This example will use `i` as it is very common across computing.

- Next, the loop needs to know the starting and ending point for the repetitive task. The example below uses a range of 2 through 5. Thus, the index `i` will initially take on the value of 2, then increment by 1 in each pass of the loop and stop when `i` equals 5.
- A ":" follows `for` loop line. The action for the `for` loop is written in the next line but tab indented. In the example below, the action is the print the `i`th item in the sequencing_list.

```
for i in range(2,5):
    print(sequencing_list[i])
```

```
whole exome
chip
atac
```

The start and end in a `for` loop does not necessarily need to be numeric in Python. The following will loop through sequencing_list and print each element. In the loop below, `sequence_type` is set as the index and the loop will print each element of `sequencing_list`.

```
for sequence_type in sequencing_list:
    print(sequence_type)
```

```
whole genome
rna
whole exome
chip
atac
```

There is also the `while` loop. The example below will print the first four items in sequencing list using `while`. Just like the `for` loop, the `while` loop needs an index to help it keep track of where it is at in the task. Here, the index is `i` and it is initiated with the value 0 outside the `while` loop. Next, the `while` loop will proceed to print the `i`th item in `sequencing_list` as long as `i` is less than 4. The `index i` is incremented by 1 in the `while` loop.

```
i=0
while i < 4:
    print(sequencing_list[i])
    i=i+1
```

```
whole genome
rna
whole exome
chip
```

What would happen if `i` was initialized to 4 and the `while` loop would iterate until `i` is equal 0.

```
i=4
while i >= 0:
    print(sequencing_list[i])
    i=i-1
```

The above `while` loop will just print the items in sequencing_list in reverse order.

```
atac
chip
whole exome
rna
whole genome
```

A `for` loop can be used to solve the issue why `sequencing_list[0,1]` did not work to subset the first and second items in `sequencing_list`. In the command construct below, `to_subset` will hold a list containing 0 and 1, which correspond the indices for the first and second item in sequencing_list. In the following line, `sequencing_list[i]` will subset the `i`th item in sequencing_list but only those indices included in `to_subset`, which the `for` loop will iterate through.

```
to_subset=[0,1]
[sequencing_list[i] for i in to_subset]
```

```
['whole genome', 'rna']
```

To subset the first and second item in `sequencing_list`, the `map` command can also be used.

Definition

"The `map()` function is used to apply a given function to every item of an iterable, such as a list or tuple, and returns a map object (which is an iterator)." -- https://www.geeksforgeeks.org/python-map-function/?ref=lbp *(https://www.geeksforgeeks.org/python-map-function/?ref=lbp)*

```
list(map(sequencing_list.__getitem__, [0,1]))
```

```
['whole genome', 'rna']
```

What if the user wanted to add the word "sequencing" at the end of each sequencing type in `sequencing_list`? To do this, the `map` function can be used to iterate through `sequencing_list` and `lambda` can be used to execute the function that adds " sequencing" to the end of every item in sequencing_list.

Definition

"A lambda function is a small anonymous function. A lambda function can take any number of arguments, but can only have one expression." -- https://www.w3schools.com/python/python_lambda.asp *(https://www.w3schools.com/python/python_lambda.asp)*

In the example below, `lambda` is used to define a function that adds " sequencing" to whatever value is passed onto the variable `sl`. In this instance, sequencing_list, the last argument in the `map` function is passed to `sl`.

```
list(map(lambda sl: sl+" sequencing", sequencing_list))
```

```
['whole genome sequencing', 'rna sequencing', 'whole exome sequencing
 'atac sequencing']
```

Another example of combining `map` and `lambda` to iterate over a task is shown in the commands below where every entry in numbers_list will be square.

```
numbers_list1=[1,2,3,4,5,6]
list(map(lambda j: j**2, numbers_list1))
```

```
[1, 4, 9, 16, 25, 36]
```

An alternative for squaring every element in `numbers_list1` is to use list comprehension *(https://www.w3schools.com/python/python_lists_comprehension.asp)*, which will essentially allow the use of one liner `for` loop to complete the task.

```
numbers_list1=[1,2,3,4,5,6]
numbers_list1=list(j**2 for j in numbers_list1)
numbers_list1
```

```
[1, 4, 9, 16, 25, 36]
```

Dictionaries

Dictionaries are key-value pairs and these are encountered as ways to specify options in some Python packages.

```
my_dictionary={"apples":"red","oranges":"orange","bananas":"yellow"}
```

Subsetting a Dictionary

There are several methods for subsetting a dictionary. See https://www.geeksforgeeks.org/get-a-subset-of-dict-in-python/ *(https://www.geeksforgeeks.org/get-a-subset-of-dict-in-python/)*.

First, just enclosing one of the keys in square brackets will retrieve its associated value.

```
my_dictionary['bananas']
```

```
yellow
```

A `for` loop can be used to subset a dictionary as well. In the example below, a new dictionary called `apples_bananas` is created just to hold the key and value pairs for apples and bananas in `my_dictionary`. To do this, follow the steps below.

1. Create any variable with a list that contains dictionary keys to extract. In this example, the variable will be named `keys_to_extract` and the list will contain apples and bananas, which are keys in `my_dictionary`.
2. Next, create an empty dictionary called `apples_bananas` by setting to empty `{}`.
3. In the `for` loop, iterate through `keys_to_extract` using the variable `k` to keep track of progress. If `k` is in `my_dictionary`, then use the dictionary's `.update` attribute to write it into `apples_bananas`. `apples_bananas` can be written to because Python dictionaries are mutable.

```

```

```
keys_to_extract = ['apples', 'bananas']
apples_bananas={}
for k in keys_to_extract:
    if k in my_dictionary:
        apples_bananas.update({k: my_dictionary[k]})
```

```
apples_bananas
```

```
{'apples': 'red', 'bananas': 'yellow'}
```

The above `for` loop can be condensed to a one liner using dictionary comprehension *(https://www.geeksforgeeks.org/python-dictionary-comprehension/)*.

```
keys_to_extract = ['apples', 'bananas']
apples_bananas={k: my_dictionary[k] for k in keys_to_extract if k in
```

An alternative to using a `for` loop to extract the key and values for apple and bananas in `my_dictionary` is Python's `zip` and `map` commands.

Definition

"The zip() function in Python combines multiple iterables such as lists, tuples, strings, dict etc, into a single iterator of tuples. Each tuple contains elements from the input iterables that are at the same position." -- https://www.geeksforgeeks.org/zip-in-python/ *(https://www.geeksforgeeks.org/zip-in-python/)*

To demonstrate `zip`, consider the lists below.

```
a1=[1,2,3]
a2=[3,4,5]
list(zip(a1,a2))
```

A list where the first, second, and third items in `a1` and `a2` are paired together.

```
[(1, 3), (2, 4), (3, 5)]
```

Next, recall that the `map` command takes an iterable item like a list and performs a certain function with it.

```
keys_to_extract = ['apples', 'bananas']
list(map(my_dictionary.get,keys_to_extract))
```

The above commands will return a list with values for apples and bananas in `my_dictionary` where the `map` function will use the dictionary's `.get` attribute to retrieve values for keys list in `keys_to_extract`.

```
['red', 'yellow']
```

Given that `zip` will perform element-wise combination on iterable items such as list, it can be used to generate key and value pairs from `keys_to_extract` and `my_dictionary` using the command below where `dict` is used to specify creation of a dictionary.

```
dict(zip(keys_to_extract, map(my_dictionary.get, keys_to_extract)))
```

```
{'apples': 'red', 'bananas': 'yellow'}
```

Updating a dictionary

Use the a dictionary's `update` attribute to add values at the end.

```
my_dictionary.update({'pears': 'green'})
```

The code below will also append the `pear` and `green` value pair to `my_dictionary`.

```
my_dictionary['pears']='green'
```

```
{'apples': 'red', 'oranges': 'orange', 'bananas': 'yellow', 'pears':
```

To add multiple items to the end of a dictionary, use `.update`.

```
my_dictionary.update({'avocado': 'green', 'kiwis': 'brown'})
```

```
{'apples': 'red', 'oranges': 'orange', 'bananas': 'yellow', 'pears':
```

The dictionary's `.pop` attribute can be used to remove an item.

```
my_dictionary.pop('pears')
```

The result of `.pop` is that the value for the removed dictionary key is returned.

```
green
```

Recall `my_dictionary` to make sure pears was removed.

```
{'apples': 'red', 'oranges': 'orange', 'bananas': 'yellow', 'avocado
```

To delete multiple items, just create a list of keys to remove and assign this list to a variable. Below, `keys_to_remove` will be used to store avocado and kiwis, which are keys from `my_dictionary` to remove.

```
keys_to_remove=['avocado', 'kiwis']
list(map(my_dictionary.pop, keys_to_remove))
```

The output for the above command is the values for the keys in `my_dictionary` that were removed.

```
['green', 'brown']
```

Recall `my_dictionary` to make sure avocado and kiwi was removed.

```
{'apples': 'red', 'oranges': 'orange', 'bananas': 'yellow'}
```

> Write a `for` loop to print "today is a work day" if the day of the week is Monday, Tuesday, Wednesday, or Thursday". If it is Friday, print "It is almost the weekend". For Saturday, print "Happy weekend!" and Sunday, print "Monday will come whether you like it or not. Hint: use a list that contains the days of the week.

{{Sdet}}{{Ssum}}Answer{{Esum}}

```
for days in days_of_week:
    if days=="Friday":
        print(days, ":", "It is almost the weekend.")
    elif days=="Saturday":
        print(days, ":", "Happy weekend!")
    elif days=="Sunday":
        print(days, ":", "Monday will come whether you like it or no
```

```
    else:
        print(days, ":", "Today is a work day.")
```

{{Edet}}

Data Wrangling using Python

Learning Objectives

After this lesson, participants will

- Be able to import tabular data into Python using Pandas
- Be able to explore and modify tabular data through various data wrangling approaches, including
    - retrieving dimensions
    - subsetting
    - obtaining descriptive statistics
    - replacing column names
    - performing mathematical operations
    - filtering
    - removing and adding columns

Importing Tabular Data using Pandas

Pandas *(https://pandas.pydata.org)* is a popular Python package used to work with tabular data.

To work with Pandas, first activate it using the `import` command.

```
import pandas
```

Sometimes the name of a package is long, so users might want to shorten it by creating an alias. The alias `pd` is often used for Pandas. To add an alias, just append `as` followed by the user defined alias to the package import command. If importing a package using an alias, then the package needes to be called using the assigned alias. For instance, if `pd` was used to alias Pandas, then use `pd.read_csv` to import a csv file.

```
import pandas as pd
```

This exercise will use the `read_csv` function of Pandas to import a comma separated value (csv) file called `hbr_uhr_chr22_rna_seq_counts.csv`, which contains RNA sequencing gene expression counts from the Human Brain Reference (hbr) and Universal Human Reference (uhr) study *(https://rnabio.org/module-01-inputs/0001/05/01/RNAseq_Data/)*. This data will be stored as the variable `hbr_uhr_chr22_counts`.

```
hbr_uhr_chr22_counts=pandas.read_csv("./hbr_uhr_chr22_rna_seq_counts
```

Take a look at the first few rows of `hbr_uhr_chr22_counts` by appending the `head` attribute to `hbr_uhr_chr22_counts`.

```
hbr_uhr_chr22_counts.head()
```

	Geneid	HBR_1.bam	HBR_2.bam	HBR_3.bam	UHR_1.bam	UHR_2.bam	UHR_3.bam
0	U2	0	0	0	0	0	0
1	CU459211.1	0	0	0	0	0	0
2	CU104787.1	0	0	0	0	0	0
3	BAGE5	0	0	0	0	0	0
4	ACTR3BP6	0	0	0	0	0	0

Figure 1: The first five rows of `hbr_uhr_chr22_counts`. The first column contains genes and the subsequent columns contain gene expression counts for each of the samples. The left most column of this data frame contains the row indices.

Because `hbr_uhr_chr22_counts` is a Pandas data frame (`type(hbr_uhr_chr22_counts)`, see lesson 2), it is possible to append one of the many Pandas commands to it. For instance, the `head` function was appended to display the first five rows of `hbr_uhr_chr22_counts`. The data frame name and function is separated by a period. This is perhaps one of the most appealing aspects of Python syntax. Note that the `head` function was followed by `()`. If the parentheses are blank, then the default first five lines will be shown. To view the first 10 rows of hbr_uhr_chr22_counts do the following.

```
hbr_uhr_chr22_counts.head(10)
```

	Geneid	HBR_1.bam	HBR_2.bam	HBR_3.bam	UHR_1.bam	UHR_2.bam	UHR_3.bam
0	U2	0	0	0	0	0	0
1	CU459211.1	0	0	0	0	0	0
2	CU104787.1	0	0	0	0	0	0
3	BAGE5	0	0	0	0	0	0
4	ACTR3BP6	0	0	0	0	0	0
5	5_8S_rRNA	0	0	0	0	0	0
6	AC137488.1	0	0	0	0	0	0
7	AC137488.2	0	0	0	0	0	0
8	CU013544.1	0	0	0	0	0	0
9	CT867976.1	0	0	0	0	0	0

Figure 2: Include an integer inside the parentheses of `pandas.dataframe.head()` function to view a specific number of lines in a tabular dataset.

The function `tail` can be used to view by default the bottom five lines of a tabular dataset. Similar to `head`, the number of lines shown can be customized by specifying an integer inside the parentheses.

```
hbr_uhr_chr22_counts.tail()
```

Get Dimensions of a Data Frame

Pandas data frames have a function `shape` that informs of the number of rows and number of columns in a data frame (in other words the dimensions of a tabular dataset). To get the dimensions for `hbr_uhr_chr22_counts`, do the following

```
hbr_uhr_chr22_counts.shape
```

The `hbr_uhr_chr22_counts` data frame has 1335 rows and 7 columns according to the output below.

```
(1335, 7)
```

Note

The elements in tabular data can be referred to by their row and column positions.

The `size` function returns the number elements in a data frame. For instance, `hbr_uhr_chr22_counts` has 1335 rows and 7 columns, which means that it has 1335 times 7 elements (or 9345).

To view more information regarding a data frame including column headers and data type in each column use the following.

```
hbr_uhr_chr22_counts.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1335 entries, 0 to 1334
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Geneid     1335 non-null   object
 1   HBR_1.bam  1335 non-null   int64
 2   HBR_2.bam  1335 non-null   int64
 3   HBR_3.bam  1335 non-null   int64
 4   UHR_1.bam  1335 non-null   int64
 5   UHR_2.bam  1335 non-null   int64
 6   UHR_3.bam  1335 non-null   int64
dtypes: int64(6), object(1)
memory usage: 73.1+ KB
```

To get descriptive statistics for a data frame use the `describe` attribute.

```
hbr_uhr_chr22_counts.describe()
```

```
       HBR_1.bam   HBR_2.bam   HBR_3.bam   UHR_1.bam   UHR_2.bam   U
count  1335.000000 1335.000000 1335.000000 1335.000000 1335.000000 1
mean     29.530337   36.264419   32.084644   50.694382   33.419476   4
std      99.177874  120.617793  108.237694  197.575081  122.598310  1
min       0.000000    0.000000    0.000000    0.000000    0.000000    0
25%       0.000000    0.000000    0.000000    0.000000    0.000000    0
50%       0.000000    0.000000    0.000000    1.000000    1.000000    1
75%       8.000000   10.000000    9.000000   13.000000   12.000000    1
max    1532.000000 1797.000000 1637.000000 4027.000000 2406.000000 1
```

The above descriptive statistics table for `hbr_uhr_chr22_counts` has too many decimal places. Append the `round` attribute to include only one decimal.

```
hbr_uhr_chr22_counts.describe().round(decimals=1)
```

```
       HBR_1.bam   HBR_2.bam   HBR_3.bam   UHR_1.bam   UHR_2.bam   U
count  1335.0      1335.0      1335.0      1335.0      1335.0      1
mean   29.5        36.3        32.1        50.7        33.4        4
std    99.2        120.6       108.2       197.6       122.6       1
min    0.0         0.0         0.0         0.0         0.0         0
25%    0.0         0.0         0.0         0.0         0.0         0
50%    0.0         0.0         0.0         1.0         1.0         1
75%    8.0         10.0        9.0         13.0        12.0        1
max    1532.0      1797.0      1637.0      4027.0      2406.0      1
```

Row Indices/Names

Figure 3 shows the first 10 rows of `hbr_uhr_chr22_counts`. The left most column, which contains labels starting with "0" is referred to as the row indices or row names. Users can specify a column in the dataset as the row indices or row names using the `index_col` option in `read_csv`. For instance, the `hbr_uhr_chr22_rna_seq_counts.csv` dataset could be imported with gene names as the row indices. To do this, add the `index_col=0` option to `read_csv`. Gene names in `hbr_uhr_chr22_rna_seq_counts.csv` is the first column and is denoted as column "0" in Python. Thus, setting `index_col=0` ensures that the gene names will be set as the row indices or row names (see Figure 3).

```
hbr_uhr_chr22_counts_1=pandas.read_csv("./hbr_uhr_chr22_rna_seq_count
```

Geneid	HBR_1.bam	HBR_2.bam	HBR_3.bam	UHR_1.bam	UHR_2.bam	UHR_3.bam
U2	0	0	0	0	0	0
CU459211.1	0	0	0	0	0	0
CU104787.1	0	0	0	0	0	0
BAGE5	0	0	0	0	0	0
ACTR3BP6	0	0	0	0	0	0
...	...	...	...	...	...	...
ACR	0	0	0	0	2	0
AC002056.5	0	0	0	0	0	0
AC002056.3	0	0	0	0	0	0
RPL23AP82	41	59	54	32	23	34
RABL2B	74	62	54	68	50	47

Figure 3. The `index_col=0` option in `pandas.read_csv` sets the gene names as row names in the imported data frame.

Data Wrangling

Subsetting

To subset a column of a pandas dataframe, the bracket notation followed by the column name can be used. For instance, to extract the `Geneid` column in `hbr_uhr_chr22_counts` do the following.

```
hbr_uhr_chr22_counts["Geneid"]
```

To subset multiple columns using bracket notation, just inlcude a list of column names to subset.

```
hbr_uhr_chr22_counts[["Geneid", "UHR_3.bam"]]
```

The command below will subset the expression counts for the RABL2B gene.

```
hbr_uhr_chr22_counts[hbr_uhr_chr22_counts["Geneid"]=="RABL2B"]
```

```

```

```
        Geneid  HBR_1.bam   HBR_2.bam   HBR_3.bam   UHR_1.bam   UHR_
1334    RABL2B      74          62          54          68          
```

The "|" symbol can be used as the "or" operator so to also subset the counts for RPL23AP82 use the following.

```
hbr_uhr_chr22_counts[(hbr_uhr_chr22_counts["Geneid"]=="RABL2B") | (hb
```

```
        Geneid      HBR_1.bam   HBR_2.bam   HBR_3.bam   UHR_1.bam   U
1333    RPL23AP82       41          59          54          32      
1334    RABL2B          74          62          54          68      
```

Alternatively, use the `isin` function and provide a list of genes to retrieve.

```
hbr_uhr_chr22_counts[hbr_uhr_chr22_counts["Geneid"].isin(["RABL2B", '
```

Use "." to reference a column. For instance, the `Geneid` column in `hbr_uhr_chr22_counts`.

```
hbr_uhr_chr22_counts.Geneid
```

Applying the "." notation to subset the expression for a specific gene.

```
hbr_uhr_chr22_counts[hbr_uhr_chr22_counts.Geneid=="RABL2B"]
```

```
        Geneid  HBR_1.bam   HBR_2.bam   HBR_3.bam   UHR_1.bam   UHR_
1334    RABL2B  74          62          54          68          50  
```

Subsetting by Integer Positions

Given that the elements in a data frame can be referenced by its row and column positions, what would be the approach for extracting the element in row 60 and column 5? The solution is the command below, which returns a result of 2. The row and column numbers are enclosed in "[]" and separated by a comma.

```
hbr_uhr_chr22_counts.iloc[60,5]
```

```

```

```
2
```

The above method for subsetting the element in row 60 and column 5 of `hbr_uhr_chr22_counts` is great if the goal is to extract the value and do numeric operation on it. But what if the user wants to return the element along with the corresponding gene in data frame format?

To do this, enclose the row and column indices to extract in their own inner set of square brackets as shown below. Column 0, which contains the gene name is also included in the brackets containing the column indices of interest.

```
hbr_uhr_chr22_counts.iloc[[60],[0,5]]
```

```
    Geneid  UHR_2.bam
60  CCT8L2      2
```

Pandas offers different approaches for subsetting rectangular data tables. One method is `iloc`.

> `iloc` is a "purely integer-location based indexing for selection by position" -- https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.iloc.html# *(https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.iloc.html#)*. The row and column positions are enclosed in "[]".

`iloc` allows for retrieval of elements in multiple rows and columns. For instance, the following can be used to retrieve the elements in rows 60 and 65 and columns 0, 4, 5, and 6 in `hbr_uhr_chr22_counts`. Note that the row and column positions are enclosed in an outer set of "[]". Within the outer set of "[]" the first set of "[]" enclose a comma separated list of row positions while the second set of "[]" enclose a comma separated list of column positions.

```
hbr_uhr_chr22_counts.iloc[[60,65],[0,4,5,6]]
```

```
    Geneid      UHR_1.bam   UHR_2.bam   UHR_3.bam
60  CCT8L2          1           2           0
65  SLC25A15P5      2           2           4
```

Note

When working with tabular data, the covention is to specify the row first then the column when referring to or subsetting elements from the table.

To get the first three rows of `hbr_uhr_chr22_counts` do the following. Note that it retrieves the rows with indices 0, 1, and 2.

```
hbr_uhr_chr22_counts.iloc[:3]
```

```
    Geneid  HBR_1.bam   HBR_2.bam   HBR_3.bam   UHR_1.bam   UHR_2.bar
0   U2          0           0           0           0           0
1   CU459211.1  0           0           0           0           0
2   CU104787.1  0           0           0           0           0
```

What will be the output for `hbr_uhr_chr22_counts.iloc[[3],:]`?

{{Sdet}}{{Ssum}}Solution{{Esum}}

The row with an index of 3 will be retrieved.

```
    Geneid  HBR_1.bam   HBR_2.bam   HBR_3.bam   UHR_1.bam   UHR_2.bar
3   BAGE5       0           0           0           0           0
```

{{Edet}}

Subsetting using column names

Panda's `loc` function allows for subsetting by row or column names. For instance, to retrieve the `Geneid` column, do the following. The ":" denotes get every row.

```
hbr_uhr_chr22_counts.loc[:,['Geneid']]
```

```
        Geneid
0       U2
1       CU459211.1
2       CU104787.1
3       BAGE5
4       ACTR3BP6
...     ...
1330    ACR
1331    AC002056.5
1332    AC002056.3
1333    RPL23AP82
1334    RABL2B
```

To retrieve the counts for the gene SLC25A15P5, use the following where SLC25A15P5 is the subsetting criteria, where

- `hbr_uhr_chr22_counts.loc[:,'Geneid']` extracts the Geneid column.
- `=="SLC25A15P5"` will filter out the row with the SLC25A15P5 gene.

```
hbr_uhr_chr22_counts[hbr_uhr_chr22_counts.loc[:,'Geneid']=="SLC25A15P
```

```
    Geneid      HBR_1.bam    HBR_2.bam  HBR_3.bam   UHR_1.bam   UHR_2
65  SLC25A15P5      0           0           0           2           2
```

To retrieve counts for more than one gene, enclose the genes of interest in a list and use the `isin` function to filter out the rows containing the genes in the list.

```
hbr_uhr_chr22_counts[hbr_uhr_chr22_counts.loc[:,'Geneid'].isin(["SLC2
```

```
    Geneid      HBR_1.bam   HBR_2.bam   HBR_3.bam   UHR_1.bam   UHR_2
60  CCT8L2          0           0           0           1           2
65  SLC25A15P5      0           0           0           2           2
```

To find all of the SLC genes in `hbr_uhr_chr22_counts`, the following could be used where `str.startswith` searches for text that starts a pattern (ie. "SLC"). Other options for pattern matching include `str.endwith` and `str.contains`.

```
hbr_uhr_chr22_counts.loc[hbr_uhr_chr22_counts.loc[:,'Geneid'].str.sta
```

```
        Geneid      HBR_1.bam   HBR_2.bam   HBR_3.bam   UHR_1.bam   U
54      SLC9B1P4        0           0           0           0
65      SLC25A15P5      0           0           0           2
109     SLC25A18        100         111         74          6
181     SLC25A1         32          50          41          226
249     SLC9A3P2        0           0           0           0
268     SLC7A4          19          25          14          9
494     SLC2A11         54          63          46          28
726     SLC35E4         18          32          26          21
783     SLC5A1          0           0           0           0
795     SLC5A4          7           12          5           13
955     SLC16A8         9           13          11          11
```

```
1046    SLC25A17        39          39          40          119
1099    SLC25A5P1       0           0           1           0
```

Replacing Column Names

To view the column headings of a data frame use the `column` function. For instance,

```
hbr_uhr_chr22_counts.columns
```

```
HBR_1.bam
HBR_2.bam
HBR_3.bam
UHR_1.bam
UHR_2.bam
UHR_3.bam
```

The `str.replace` function can be used to replace a string with something else. Here, it used to remove ".bam" from the sample names in the column heading.

```
hbr_uhr_chr22_counts.columns=hbr_uhr_chr22_counts.columns.str.replace
```

Mathematical Operations on Data Frames and Filtering

Pandas enables mathematical operations on data frames. For instance, one might want to sum the total counts across all samples for each gene. The `sum` function can be used to do this. Setting `axis=1` will sum up the counts for each row or gene. Because the `Geneid` column is a string, it is necessary to first subset only the numeric columns.

```
hbr_uhr_chr22_counts.loc[:, ['HBR_1', 'HBR_2', 'HBR_3', 'UHR_1', 'UHR
```

Below, genes with zero counts across all samples are removed from `hbr_uhr_chr22_counts` and stored as `hbr_uhr_chr22_counts_filtered`. To accomplish this set `hbr_uhr_chr22_counts.loc[:, ['HBR_1', 'HBR_2', 'HBR_3', 'UHR_1', 'UHR_2', 'UHR_3']].sum(axis=1)!=0` and use as a filter criteria. Essentially, this will keep the genes whose expression across samples is not 0.

```
hbr_uhr_chr22_counts_filtered=hbr_uhr_chr22_counts.loc[hbr_uhr_chr22_
```

Removing and Adding Columns to a Data Frame

This exercise will use the differential gene expression analysis table from the hbr and uhr study.

```
hbr_uhr_deg_chr22=pandas.read_csv("./hbr_uhr_deg_chr22.csv")
```

The `info()` attribute will retrieve information regarding the `hbr_uhr_deg_chr22` data frame, which includes the column names.

```
hbr_uhr_deg_chr22.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1335 entries, 0 to 1334
Data columns (total 18 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   name            1335 non-null   object
 1   baseMean        1335 non-null   float64
 2   baseMeanA       1335 non-null   float64
 3   baseMeanB       1335 non-null   float64
 4   foldChange      971 non-null    float64
 5   log2FoldChange  971 non-null    float64
 6   lfcSE           971 non-null    float64
 7   stat            971 non-null    float64
 8   PValue          971 non-null    float64
 9   PAdj            971 non-null    float64
 10  FDR             639 non-null    float64
 11  falsePos        639 non-null    float64
 12  HBR_1.bam       1335 non-null   float64
 13  HBR_2.bam       1335 non-null   float64
 14  HBR_3.bam       1335 non-null   float64
 15  UHR_1.bam       1335 non-null   float64
 16  UHR_2.bam       1335 non-null   float64
 17  UHR_3.bam       1335 non-null   float64
dtypes: float64(17), object(1)
memory usage: 187.9+ KB
```

The `hbr_uhr_deg_chr22` table contains differential gene expression analysis results. Relevant columns include:

- name: gene names
- log2FoldChange: the gene expression change between the two treatment groups

- PAdj: the adjusted p-value associated with statistical confidence of the expression change
- The columns labeled with the sample names (ie. columns 12 through 17) are the normalized gene expression counts

Use `str.replace` to remove ".bam" from the sample names in columns 12 through 17.

```
hbr_uhr_deg_chr22.columns=hbr_uhr_deg_chr22.columns.str.replace(".bar
```

To drop columns in a Pandas data frame, use the `.drop` function and specify the name(s) of the column(s) to remove. The example below removes columns baseMean, baseMeanA,and baseMeanB, however it does not overwrite the original data frame. To ovewrite, include the `inplace` option and set it to `True`.

```
hbr_uhr_deg_chr22.drop(columns=["baseMean","baseMeanA", "baseMeanB"])
```

```
hbr_uhr_deg_chr22.drop(columns=["baseMean","baseMeanA", "baseMeanB"]
```

Subset the name, log2FoldChange, and PAdj columns in hbr_uhr_deg_chr22 and save to a new data frame `hbr_uhr_deg_chr22_1`.

```
hbr_uhr_deg_chr22_1=hbr_uhr_deg_chr22.loc[:,["name", "log2FoldChange"
```

```
hbr_uhr_deg_chr22_1.head()
```

```
    name    log2FoldChange      PAdj
0   SYNGR1      -4.6        5.200000e-217
1   SEPT3       -4.6        4.500000e-204
2   YWHAH       -2.5        4.700000e-191
3   RPL3         1.7        5.400000e-134
4   PI4KA       -2.0        2.900000e-118
```

Next, add a column called "-log10PAdj" to `hbr_uhr_deg_chr22_1`, which will contain the negative of log10 of the values in the PAdj column. "-log10PAdj" is used in volcano plots that depict gene expression change versus statistical confidence. To calculate -log10PAdj, the package numpy will be used. Numpy *(https://numpy.org)* enables scientific calculations.

```
import numpy
```

```
hbr_uhr_deg_chr22_1["-log10PAdj"]=numpy.negative(numpy.log10(hbr_uhr_
```

Take a look at the first several lines of hbr_uhr_deg_chr22_1

```
hbr_uhr_deg_chr22_1.head()
```

```
    name    log2FoldChange      PAdj            -log10PAdj
0   SYNGR1  -4.6                5.200000e-217   216.283997
1   SEPT3   -4.6                4.500000e-204   203.346787
2   YWHAH   -2.5                4.700000e-191   190.327902
3   RPL3    1.7                 5.400000e-134   133.267606
4   PI4KA   -2.0                2.900000e-118   117.537602
```

Other methods for adding new column to a Pandas data frame include `insert` and `assign`.

The final task for this lesson is to add a column that indicates whether a gene is up regulated, down regulated, or has no change based on the log2FoldChange and PAdj values. The criteria are as follows.

- PAdj >= 0.01: no change (marked as ns in the column)
- Absolute value of log2FoldChange <2: no change (marked as ns in the column)
- log2FoldChange >= 2 and PAdj < 0.01: (up regulated)
- log2FoldChange <=2 and PAdj < 0.01: (down regulated)

To code this in Python, the first step is to drop the NA values from the `hbr_uhr_deg_chr22_1` using `dropna` with the `axis` option set to 0 to remove rows that contain NA and `inplace` set to True to so that the original dataframe will be modified.

```
hbr_uhr_deg_chr22_1.dropna(axis=0, inplace=True)
```

Next, create a list called significance_criteria that contains the criteria shown above. In the criteria list below, "&" is the Boolean for "and". To calculate the absolute value of log2FoldChange, `numpy.absolute` is used.

```
significance_criteria=[(hbr_uhr_deg_chr22_1["PAdj"]>=0.01),
                       (numpy.absolute(hbr_uhr_deg_chr22_1["log2Fold(
                      (hbr_uhr_deg_chr22_1["log2FoldChange"]>=2) & (
                      (hbr_uhr_deg_chr22_1["log2FoldChange"]<=-2) &
```

Then, create a list called significance_status that indicates whether the criteria are ns (not significant), up, or down. These statuses have to correspond to the order in which the criteria were listed in significance_criteria.

```
significance_status=["ns","ns","up","down"]
```

Finally, `numpy.select` will be used to assign values to the significance column.

```
hbr_uhr_deg_chr22_1["significance"]=numpy.select(significance_criter
```

```
hbr_uhr_deg_chr22_1.head(4)
```

```
    name    log2FoldChange          PAdj        -log10PAdj  significa
0   SYNGR1      -4.6          5.200000e-217   216.283997  down
1   SEPT3       -4.6          4.500000e-204   203.346787  down
2   YWHAH       -2.5          4.700000e-191   190.327902  down
3   RPL3         1.7          5.400000e-134   133.267606  ns
```

Write this data frame to a csv file in the `/data/username/pies_data` folder, which should be the present working directory. Replace username with the user's Biowulf account ID. The `to_csv` command in Pandas is used to write data frames to csv files. Setting `index=False` ensures that the csv file will not have row names.

```
hbr_uhr_deg_chr22_1.to_csv("./hbr_uhr_deg_chr22_with_significance_le
```

This lesson has shown the participants various data wrangling approaches using the Python package Pandas. The capabability of Pandas expand to more than what is covered here, participants are encouraged to check out the Pandas documentations *(https://pandas.pydata.org/docs/)* to learn more.

Data Visualization using Python

Learning Objectives

This lesson will provide participants with enough knowledge to start using Python for data visualization. Specifically, participants should

- Be able to use the package Seaborn to
  - Construct plots that range from very basic to elegant as well as biologically relevant
  - Customize plots including altering font size and adding custom annotations

Python Data Visualization Tools

Seaborn *(https://seaborn.pydata.org)* is a popular Python plotting package, which is the tool that will be introduced in this lesson. Seaborn is an extension of and builds on Matplotlib *(https://matplotlib.org)* and is oriented towards statistical data visualization. However, there are other packages, including those that are domain specific, implement grammar of graphics, and are used for creating web-based visualization dashboards. A non-exhaustive list of Python plotting packages is shown below.

- Matplotlib *(https://matplotlib.org)*
- Plotnine: implements grammar of graphics for those familiar with R's ggplot2 *(https://plotnine.readthedocs.io/en/stable/)*
- bioinfokit: genomic data visualization *(https://github.com/reneshbedre/bioinfokit)*
- pygenomeviz: visuazlize comparative genomics data *(https://moshi4.github.io/pyGenomeViz/)*
- Dash bio: create interactive data visualizations and web dashboards *(https://dash.plotly.com/dash-bio)*

Visualization using Seaborn

Load Packages

```
import pandas
import numpy
import matplotlib.pyplot as plt
import seaborn
```

Modify the Basic Plot Elements with Seaborn.

To plot using Seaborn, start the command with `seaborn` followed by the plot type (where plot type can be any plot, for instance if the user wants a scatter plot then the command would be `seaborn.scatterplot`), separated by a period. A common alias for Seaborn is `sns`.

```
seaborn.plot_type
```

This section will use Seaborn's `scatterplot` to explore how to work with and modify basic elements of plotting. The foundations learned in this section form the basis for creating advanced and elegant plots.

The data that will be plotted is a point located at 5 on the x axis and 5 on the y axis. To generate `x` and `y`, `numpy.array` will be used. Here, x and y are single element arrays that store the number 5.

```
x=numpy.array([5])
y=numpy.array([5])
```

Plot `x` and `y` using Seaborn's `scatterplot` function (see Figure 1 for results), which takes data frames or Numpy arrays as input. Here, x will be plotted on the x axis, and y will be plotted on the y axis. The plot can be stored as a variable, which in this example is plot0.

```
plot0=seaborn.scatterplot(x=x, y=y)
plt.show()
```

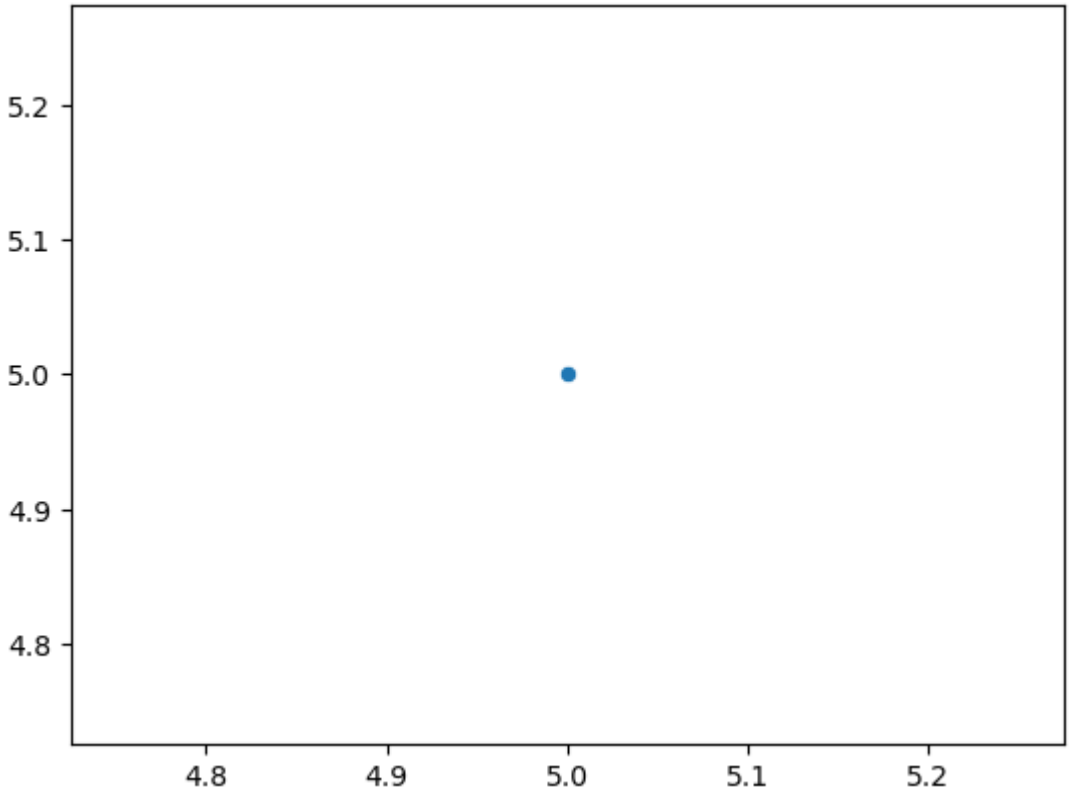


Figure 1

The plot in Figure 1 has no axes labels. Axes labels are an integral part of an informative data visualization. It might also be useful to include meaningful x and y limits. To do this, append the various `.set*` attributes to the plot. See Figure 2a for result.

- `set_xlabel`: specify x axis label (`size` is used to set the label font size)
- `set_ylabel`: specify y axis label
- `set_xlim`: sets the x axis limits
- `set_ylim`: sets the y axis limits
- `set_xticks`: sets the location of x axis tick marks
- `set_xticklabels`: sets the x axis tick mark labels, `size` is used to set the tick mark label font size
- `set_yticks`: sets the location of y axis tick marks
- `set_yticklabels`: sets the y axis tick mark labels, `size` is used to set the tick mark label font size

```
plot0=seaborn.scatterplot(x=x, y=y)
plot0.set_xlabel("x axis", size=14)
plot0.set_ylabel("y axis", size=14)
plot0.set_xlim(0,10)
plot0.set_ylim(0,10)
plot0.set_xticks([0,2,4,6,8,10])
plot0.set_xticklabels(labels=["0","2","4","6","8","10"], size=15)
```

```
plot0.set_yticks([0,2,4,6,8,10])
plot0.set_yticklabels(labels=["0","2","4","6","8","10"], size=15)
plt.show()
```

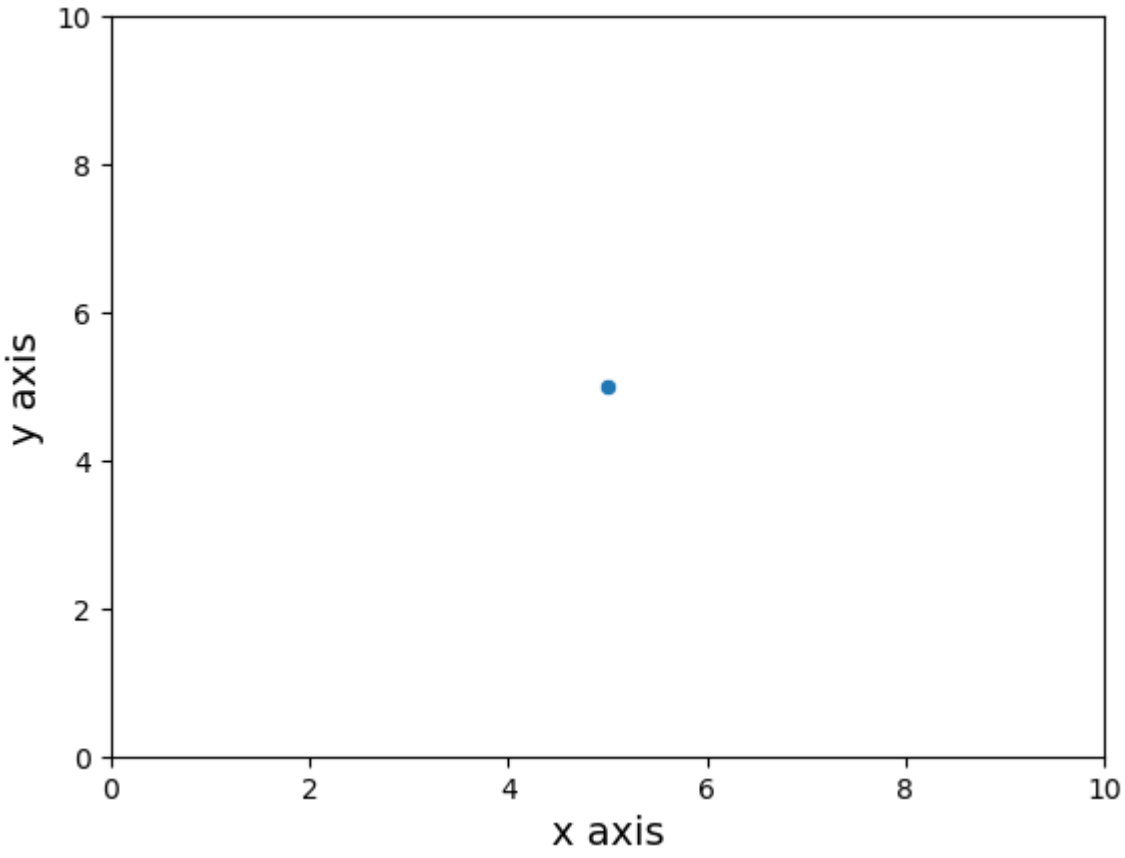


Figure 2

The `plotting_context` of a Seaborn plot contains parameters that determine scaling of plot elements (see https://seaborn.pydata.org/generated/seaborn.plotting_context.html *(https://seaborn.pydata.org/generated/seaborn.plotting_context.html)*). To view these parameters, do the following.

```
print(seaborn.plotting_context())
```

Essentially, the Seaborn plotting context is a dictionary containing key-value pairs that control the aesthetics of the data visualization.

```
{'font.size': 12.0, 'axes.labelsize': 12.0, 'axes.titlesize': 12.0,
```

These parameters can be changed using the `set_context` (see https://seaborn.pydata.org/generated/seaborn.set_context.html *(https://seaborn.pydata.org/generated/*

seaborn.set_context.html)) function by providing a customized dictionary and assigning it to the `rc` argument.

```
help(seaborn.set_context)
```

```
Help on function set_context in module seaborn.rcmod:

set_context(context=None, font_scale=1, rc=None)
    Set the parameters that control the scaling of plot elements.

    This affects things like the size of the labels, lines, and other
    of the plot, but not the overall style. This is accomplished usin
    matplotlib rcParams system.

    The base context is "notebook", and the other contexts are "paper
    and "poster", which are version of the notebook parameters scaled
    values. Font elements can also be scaled independently of (but re
    the other values.

    See :func:`plotting_context` to get the parameter values.

    Parameters
    ----------
    context : dict, or one of {paper, notebook, talk, poster}
        A dictionary of parameters or the name of a preconfigured set
    font_scale : float, optional
        Separate scaling factor to independently scale the size of th
        font elements.
    rc : dict, optional
        Parameter mappings to override the values in the preset seabo
        context dictionaries. This only updates parameters that are
        considered part of the context definition.
```

Tip

See https://seaborn.pydata.org/tutorial/aesthetics.html *(https://seaborn.pydata.org/tutorial/aesthetics.html)* to learn about adjust aesthetics for Seaborn plots.

To change the x and y axes tick label font size to 20, use `seaborn.set_context(rc={'xtick.labelsize': 20, 'ytick.labelsize': 20})` prior to constructing a Seaborn plot.

The code above can be modified to generate a more complex scatter plot. For instance, the inputs for x and y can be changed to numeric arrays of five 6 elements each.

```
x=numpy.array([0,1,2,3,4,5])
y=numpy.multiply(2,x)
print("x is a numeric array composed of: ", x)
print("y is a numeric array composed of: ", y)
```

```
x is a numeric array composed of:  [0 1 2 3 4 5]
y is a numeric array composed of:  [ 0  2  4  6  8 10]
```

The code used to generate Figure 2 can then be run again with modifications to the x and y axes limits to generate the plot shown in Figure 3. To produce a line plot representation of Figure 3, simply change the plot type to lineplot (`seaborn.lineplot`).

```
plot0=seaborn.scatterplot(x=x, y=y)
plot0.set_xlabel("x axis", size=14)
plot0.set_ylabel("y axis", size=14)
plot0.set_xlim(0,6)
plot0.set_ylim(0,12)
plot0.set_xticks([0,2,4,6])
plot0.set_xticklabels(labels=["0","2","4","6"], size=15)
plot0.set_yticks([0,2,4,6,8,10,12])
plot0.set_yticklabels(labels=["0","2","4","6","8","10","12"], size=1!
plt.show()
```

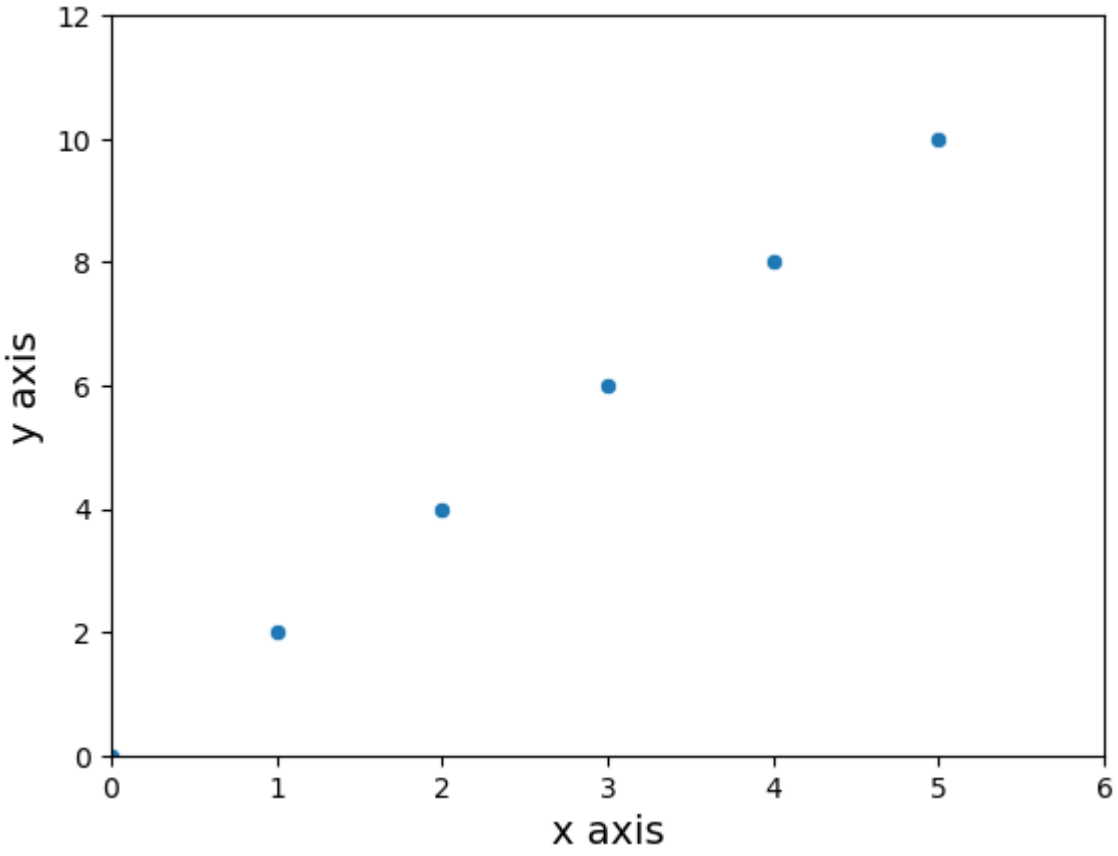


Figure 3

Constructing Biologically Relevant Plots

The next exercise is to practice creating a scatter plot on a biologically relevant dataset. Namely, the differential expression results from the hbr and uhr RNA sequencing study will be used to create a scatter plot depicting log2 fold change of gene expression on the x axis and negative log10 of the adjusted p-values on the y axis. This special case of scatter plot is called a volcano plot.

Step one is to import the data using Panda's `read.csv` command.

```
hbr_uhr_deg_chr22=pandas.read_csv("./hbr_uhr_deg_chr22_with_significa
```

Now, review the contents of this data table by doing the following.

```
hbr_uhr_deg_chr22.head(4)
```

```
    name    log2FoldChange      PAdj        -log10PAdj  significance
0   SYNGR1      -4.6        5.200000e-217  216.283997  down
1   SEPT3       -4.6        4.500000e-204  203.346787  down
```

```
2    YWHAH        -2.5          4.700000e-191   190.327902   down
3    RPL3         1.7           5.400000e-134   133.267606   down
```

To create the volcano plot, provide the following arguments. See Figure 4 for result.

- The data frame (ie. hbr_uhr_deg_chr22)
- What to plot on the x axis (ie. log2FoldChange)
- What to plot on the y axis (ie. "-log10PAdj")

```
plot1=seaborn.scatterplot(hbr_uhr_deg_chr22,x="log2FoldChange", y="-
```

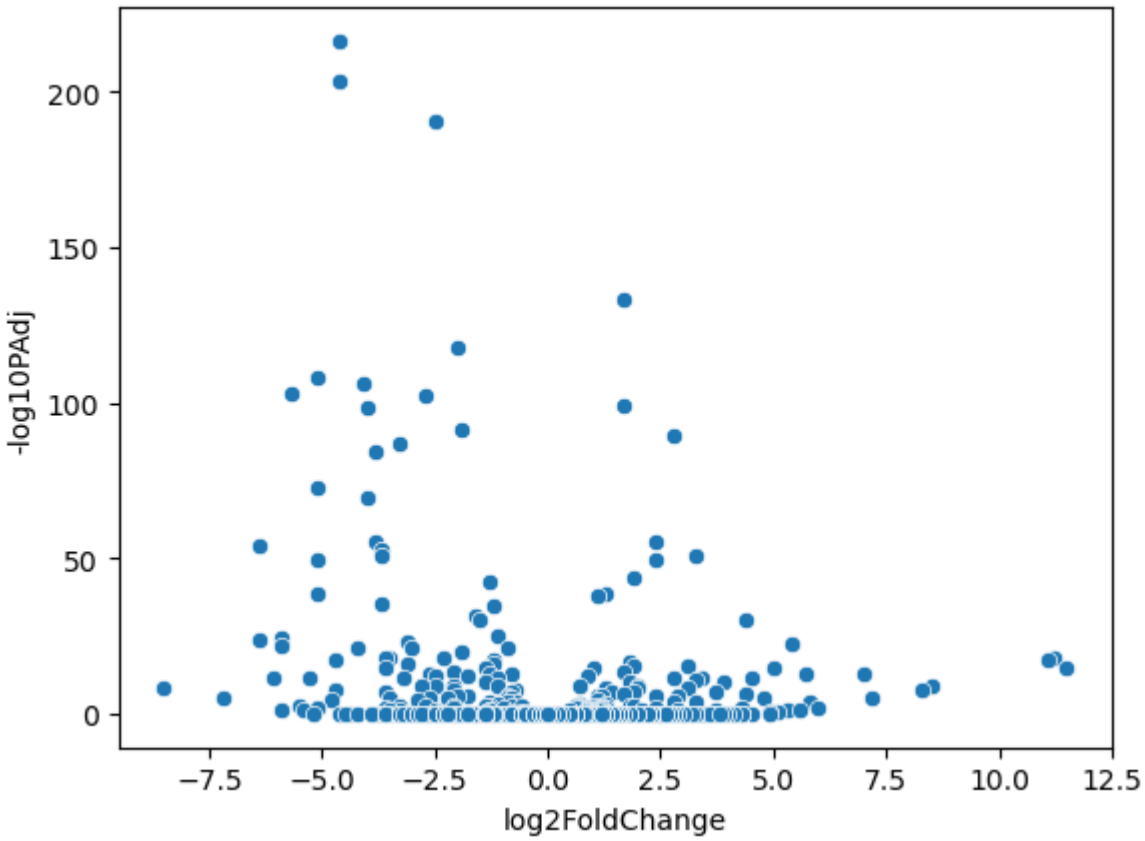


Figure 4

The volcano plot in Figure 4 does not help with visualizing the up, down, an non-significant genes. Fortunately, the `hue` option can be used to distinguish these. See Figure 5.

```
plot1=seaborn.scatterplot(hbr_uhr_deg_chr22,x="log2FoldChange", y="-
```

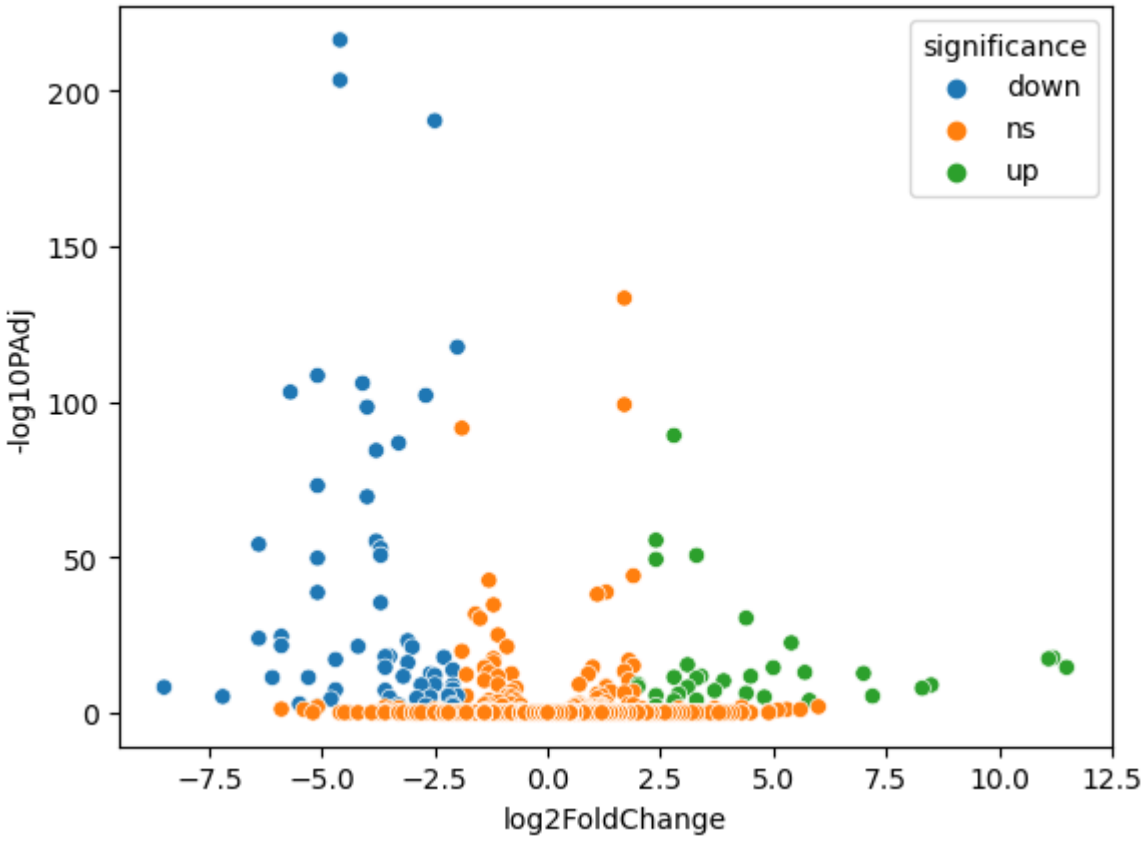


Figure 5

It would be informative to label some of the top significant differentially expressed genes in the volcano plot. To do this, import the file `hbr_uhr_deg_chr22_top_genes.csv` and assign it to the data frame `hbr_uhr_deg_chr22_top_genes`.

```
hbr_uhr_deg_chr22_top_genes=pandas.read_csv("./hbr_uhr_deg_chr22_top_
```

```
hbr_uhr_deg_chr22_top_genes
```

The table contains the top two differentially expressed genes according to the adjusted p-value (PAdj). The task to do is to label the points corresponding to these two genes on the volcano plot. The values for log2FoldChange and -log10PAdj will serve as the x and y coordinates for plotting the gene name.

```
     name    log2FoldChange       PAdj          -log10PAdj  significance
0    XBP1         2.8          7.300000e-90     89.136677   up
1    SYNGR1       -4.6         5.200000e-217    216.283997  down
```

To label the two top differentially expressed genes, start by constructing the volcano plot from Figure 5. Then, use a `for` loop to iterate through the name column in the data frame `hbr_uhr_deg_chr22_top_genes`. In the `for` loop

- `i`: the number that keeps track of the row number in the data frame `hbr_uhr_deg_chr22_top_genes` and is used to
  - reference the x coordinate or log2FoldChange value in that row
  - reference the y coordinate or -log10PAdj value in that row
- `enumerate`: iterate through the `name` column in `hbr_uhr_deg_chr22_top_genes` and stores the name to variable `gene_name`. `i` is incremented as it iterates through the name column within the `for` loop

```
plot1=seaborn.scatterplot(hbr_uhr_deg_chr22,x="log2FoldChange", y="-
for i, gene_name in enumerate(hbr_uhr_deg_chr22_top_genes["name"]):
    plot1.text(hbr_uhr_deg_chr22_top_genes["log2FoldChange"][i],
               hbr_uhr_deg_chr22_top_genes["-log10PAdj"][i],gene_name)
```

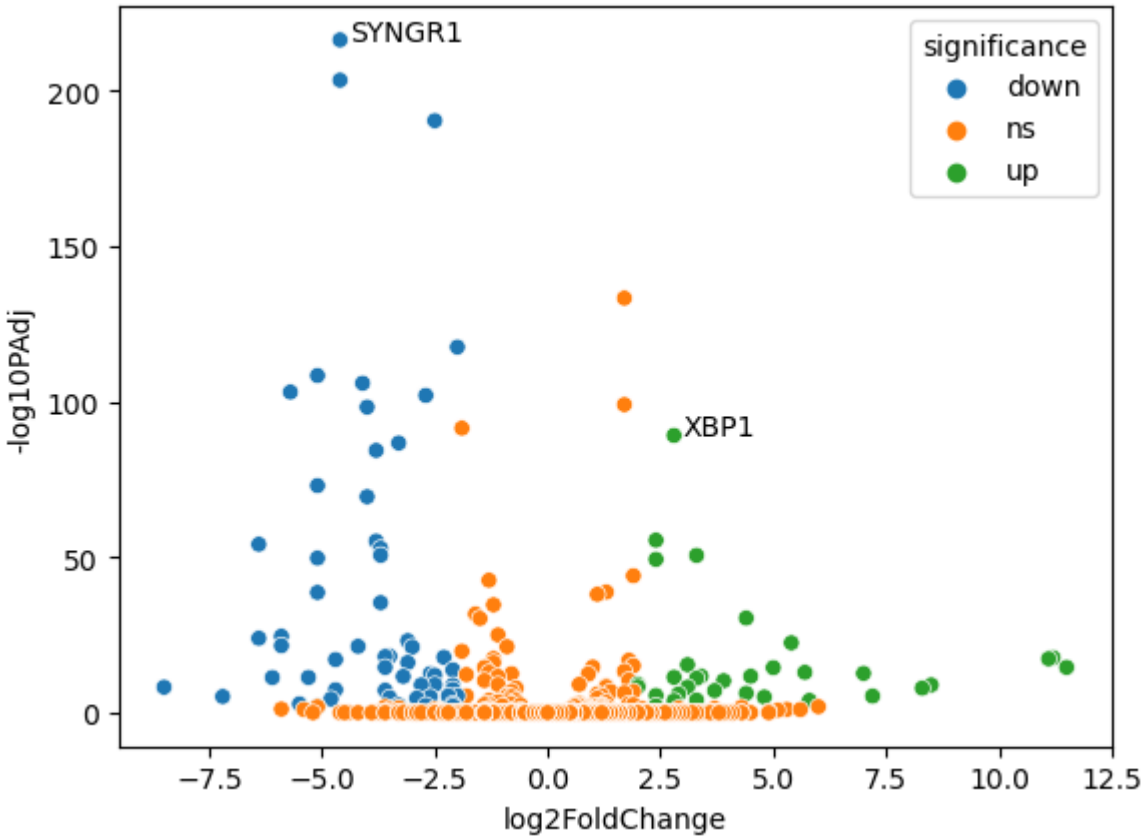


Figure 6

The next visualization is the heatmap and dendrogram combination, which helps with visualizing clusters and patterns. Heatmap and dendrogram can be used in RNA sequencing studies to inspect whether there are cluster of genes with similar expression patterns among study

groups. The normalized counts for the top differential expressed genes in the hbr and uhr study will be used to construct a heatmap/dendrogram using Seaborn's `clustermap`.

Import the data.

```
hbr_uhr_top_deg_normalized_counts=pandas.read_csv("./hbr_uhr_top_deg_
```

The `seaborn.clustermap` command below generates a clustermap of the top differential expressed genes in the hbr and uhr study. The arguments and options are as follows.

- Argument: The dataset (ie. `hbr_uhr_top_deg_normalized_counts`)
- Options:
  - `z_score=0`: scale the rows by z-score
  - `cmap`: specify color palette (ie. viridis)
  - `figsize`: specify figure size
  - `vmin`: minimum value on the color scale bar
  - `vmax`: maximum value on the color scale bar
  - `cbar_kws`: dictionary containing key value pair that specifies the title to the color scale bar
  - `cbar_pos`: coordinates for placement of the color scale bar

```
plot4=seaborn.clustermap(hbr_uhr_top_deg_normalized_counts,z_score=0
                         figsize=(8,8),vmin=-1.5, vmax=1.5,cbar_kws={
                         cbar_pos=(0.855,0.8,0.025,0.15))
```

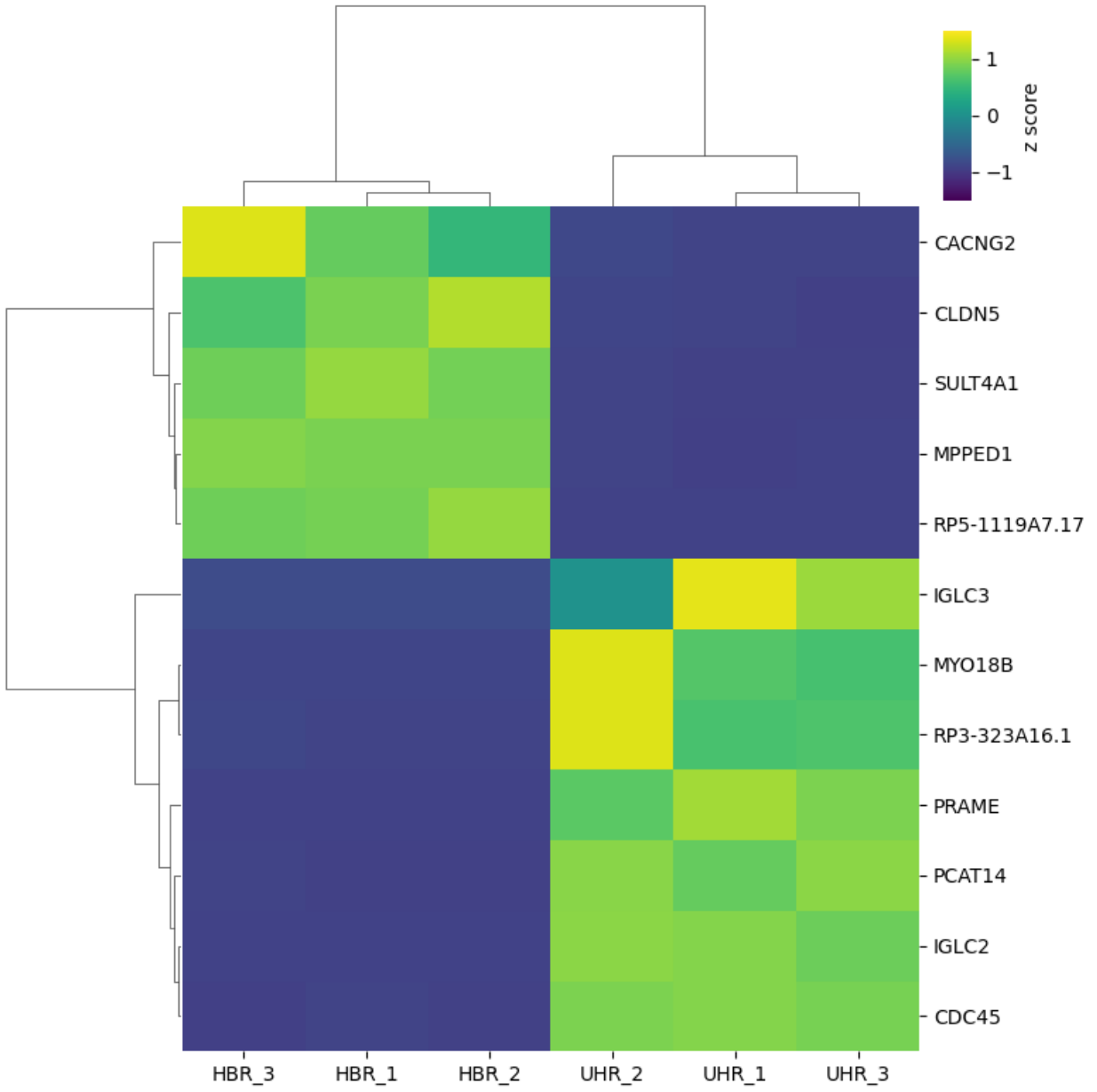


Figure 9: Expression heatmap of the top 12 differentially expressed genes in the HBR and UHR study

Below, a Pandas Series, called samples that contains a mapping of colors to study samples is created.

```
samples=pandas.Series({"HBR_1":"orangered", "HBR_2":"orangered", "HBI
```

Then a variable, column_colors is created that contains a mapping of the hbr_uhr_top_deg_normalized_counts column headings to the colors specified in samples. This is accomplished using the `map` command.

```
column_colors=hbr_uhr_top_deg_normalized_counts.columns.map(samples)
```

The option `col_colors`, which is then added to display a color bar on the top of the heatmap that helps to distinguish treatment groups (ie. hbr or uhr).

Other options added include

- `ax_heatmap.set_xticklabels`: allows for customizing the x axis labels' fontsize and rotation. This requires using `ax_heatmap.get_xmajorticklabels()` to get the x axis tick labels
- `ax_cbar.tick_params`: sets the size for the color scale bar labels
- `ax_col_colors.set_title`: sets the title and location bar displaying the treatment group to color mapping

```
plot4=seaborn.clustermap(hbr_uhr_top_deg_normalized_counts,z_score=0
                         figsize=(8,8),vmin=-1.5, vmax=1.5,cbar_kws={
                         col_colors=column_colors, cbar_pos=(0.855,0.8
plot4.ax_heatmap.set_xticklabels(plot4.ax_heatmap.get_xmajorticklabel
plot4.ax_cbar.tick_params(labelsize=12)
plot4.ax_col_colors.set_title("treatment",x=-0.1,y=0.01)
plt.show()
```

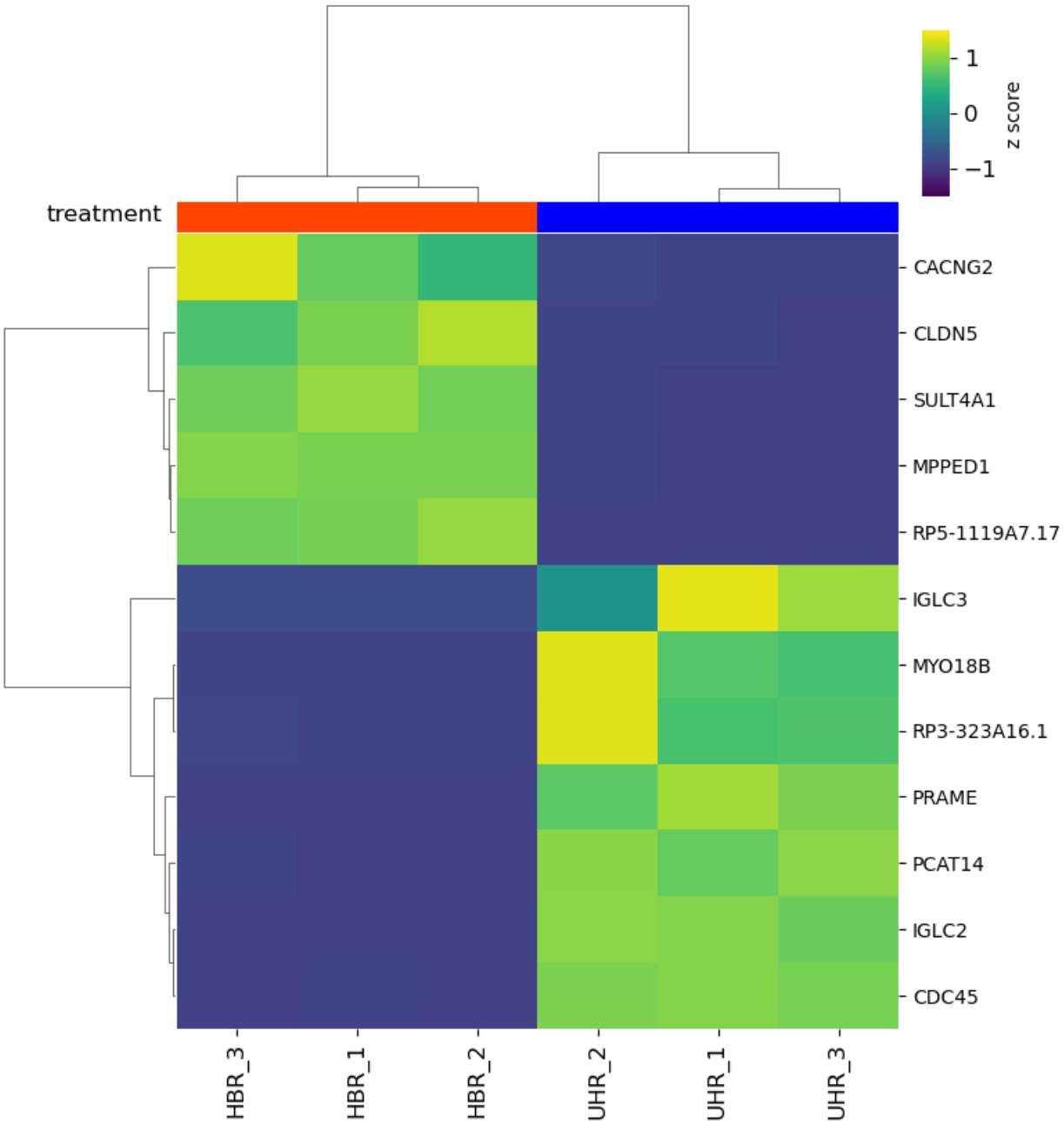


Figure 10: Expression heatmap of the top 12 differentially expressed genes in the HBR and UHR study with treatment group annotations.

Even though there is a color bar that separates the HBR and UHR groups in the heatmap, there is no legend showing which color corresponds to which group. This can be resolved by using the Patch module from Matplotlib.

Note

"A patch is a 2D artist with a face color and an edge color." -- https://matplotlib.org/stable/api/_as_gen/matplotlib.patches.Patch.html *(https://matplotlib.org/stable/api/_as_gen/matplotlib.patches.Patch.html)*

```
from matplotlib.patches import Patch
```

Then create a dictionary (here it will be called `treatment_groups`) to store the treatment and color value pairs.

```
treatment_groups={'HBR': 'orangered', 'UHR': 'blue'}
```

Next, create variable called `handles` to store in list, the colored tiles corresponding to each treatment group (ie. HBR or UHR) in the legend.

```
handles=[Patch(facecolor="orangered"), Patch(facecolor="blue")]
```

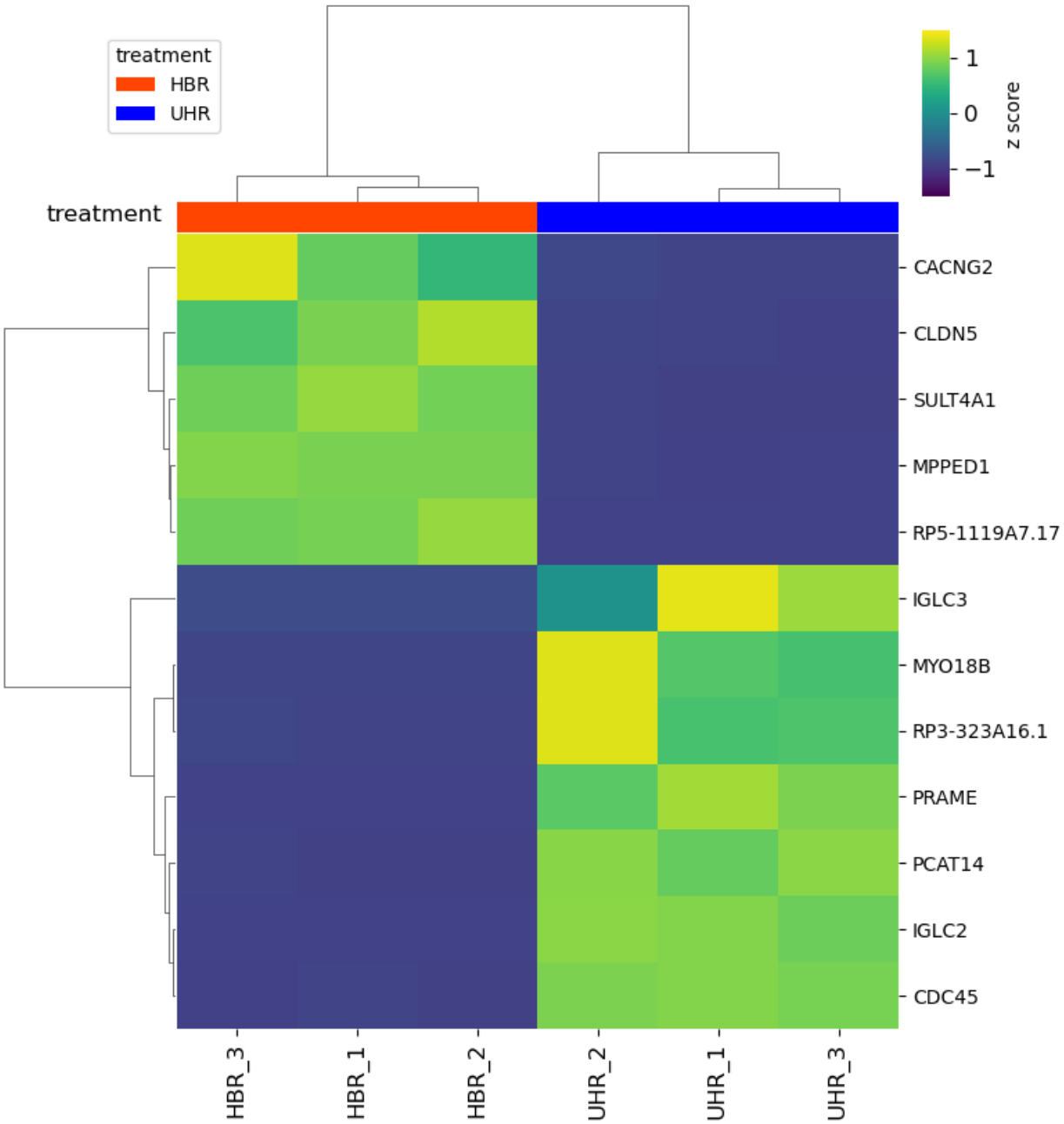
Finally, add `plt.legend(handles, treatment_groups, title='treatment', bbox_to_anchor=(-25, 1))` to the `clustermap` construct where:

- `handles` passes the color tiles set from Matplotlib's Patch module.
- `treatment_groups` will label each patch according the group that it belongs (orangered for HBR and blue for UHR).
- The title of the legend is set using the `title` option.
- `bbox_to_anchor` helps set the coordinate location in the plot in which the legend should appear.

Note

`plot4.savefig("./hbr_uhr_heatmap_with_legend.png")` is used to save the heatmap as a PNG. Users will need to replace `plot4` with the specific variable that was used to store the plot.

```
plot4=seaborn.clustermap(hbr_uhr_top_deg_normalized_counts,z_score=0
                         figsize=(8,8),vmin=-1.5, vmax=1.5,cbar_kws=(
                         col_colors=column_colors, cbar_pos=(0.855,0.8
plot4.ax_heatmap.set_xticklabels(plot4.ax_heatmap.get_xmajorticklabe
plot4.ax_cbar.tick_params(labelsize=12)
plot4.ax_col_colors.set_title("treatment",x=-0.1,y=0.01)
plt.legend(handles, treatment_groups, title='treatment', bbox_to_anch
plt.show()
plot4.savefig("./hbr_uhr_heatmap_with_legend.png")
```


treatment
HBR
UHR
1
0
−1
z score
treatment
CACNG2
CLDN5
SULT4A1
MPPED1
RP5-1119A7.17
IGLC3
MYO18B
RP3-323A16.1
PRAME
PCAT14
IGLC2
CDC45
HBR_3
HBR_1
HBR_2
UHR_2
UHR_1
UHR_3

Using Python through Command Line on Biowulf

Copy Example Data to User's Biowulf Data Directory

This class will use Jupyter Lab installed on Biowulf for interactions with Python. To get started, open a Terminal (if working on a Mac) or a Command Prompt (if working on Windows) and sign into the user's Biowulf accounts.

In the `ssh` command construct below, be sure to replace user with the participant's own Biowulf login ID.

```
ssh user@biowulf.nih.gov
```

Next, change into the participant's Biowulf data directory. Remember to replace user with the participant's own Biowulf login ID.

```
cd /data/user
```

Then, copy the `pies_data` directory in `/data/classes/BTEP` on Biowulf to the `pies_class_2025`.

```
cp -r /data/classes/BTEP/pies_data .
```

Finally, change into `pies_data`.

```
cd pies_data
```

Starting Jupyter Lab

Step 1 to starting Jupyter Lab on Biowulf via command is to request an interactive compute session using `sinteractive` with the following options. The local temporary/scratch storage space and RAM allocation can be modified base on user needs.

- `--gres=lscratch:5`: to allocate 5gb of local temporary/scratch storage space
- `--mem=2gb`: to request 2gb of memory or RAM

- `--tunnel`: to open up a channel of communication between local machine and Biowulf to allow interaction with applications like Jupyter Lab

```
sinteractive --gres=lscratch:5 --mem=2gb --tunnel
```

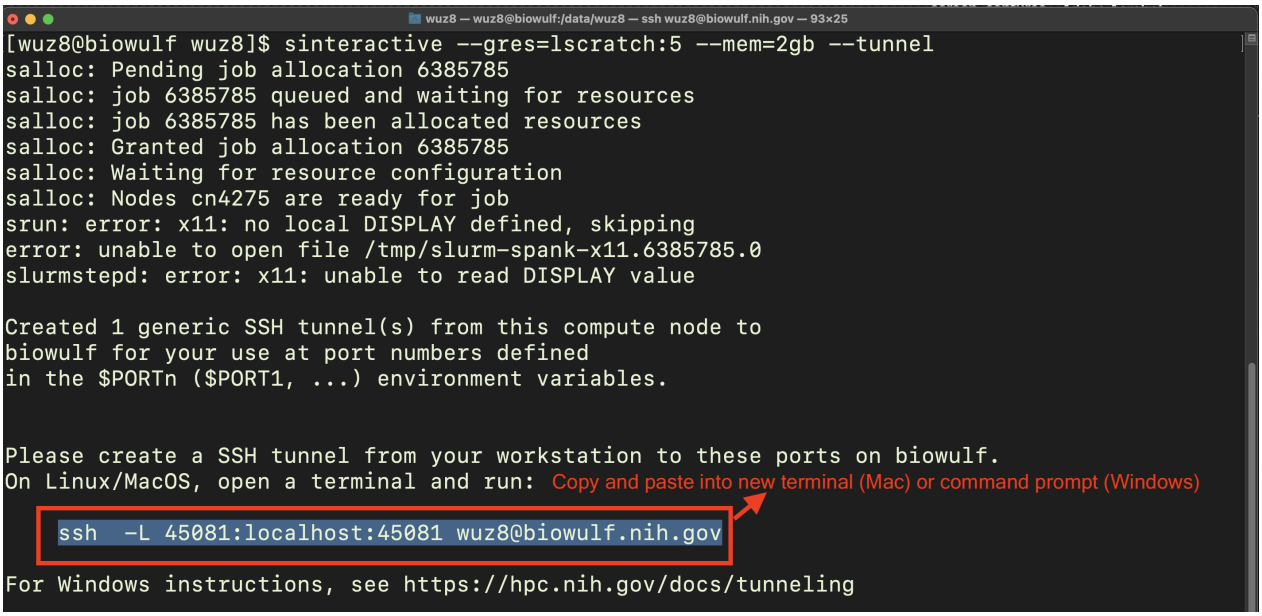


Figure 1: After interactive session resources have been allocated, users will see a `ssh` command that looks like that enclosed in the red rectangle in the figure below. Open a new terminal (if working on a Mac) or command prompt (if working on a Windows computer) and then copy and paste this `ssh` command into the new terminal.

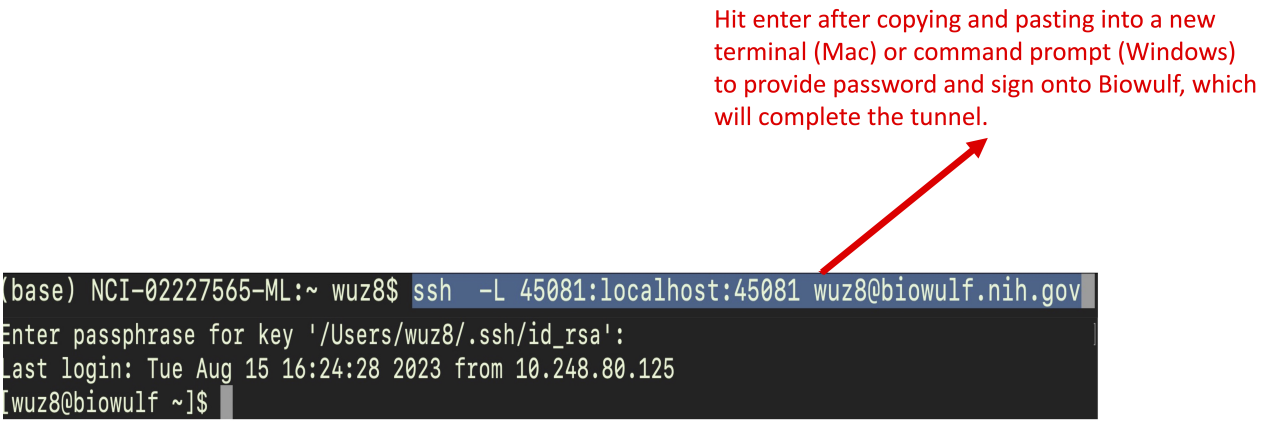


Figure 2: Hit enter after copying and pasting the `ssh` command to a new terminal to provide password and log into Biowulf. This will complete the tunnel.

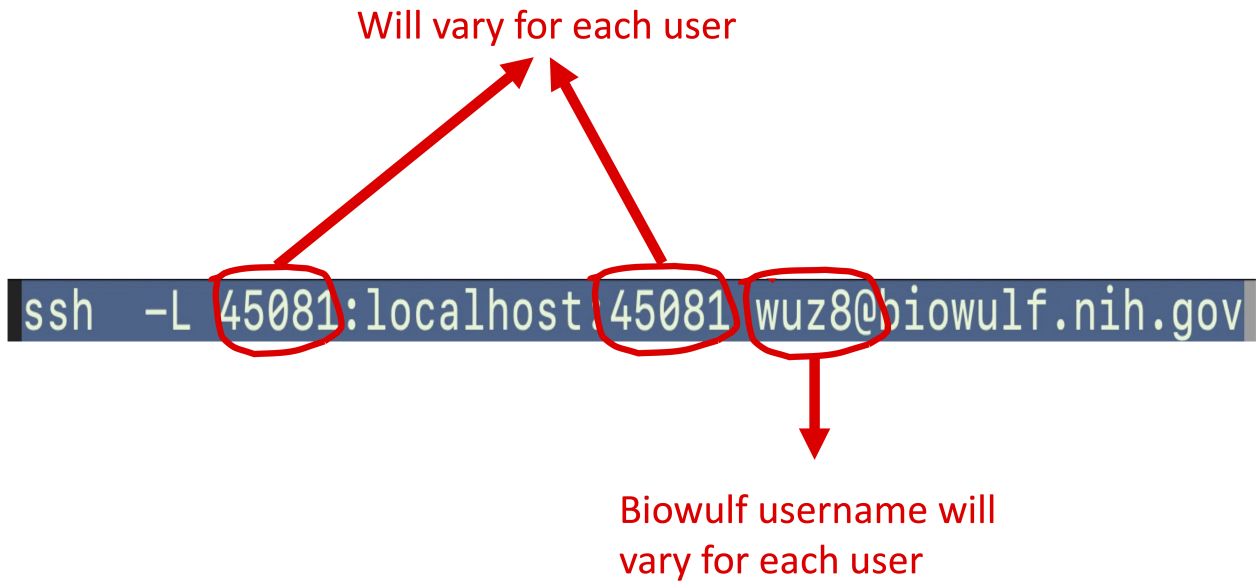


Figure 3: In the `ssh` command shown in Figure 1 and Figure 2, the numbers preceding and following "localhost" will differ depending on user. Also, the Biowulf username will differ for each user (wuz8 is the instructor's Biowulf username).

```
salloc: job 6385785 queued and waiting for resources
salloc: job 6385785 has been allocated resources
salloc: Granted job allocation 6385785
salloc: Waiting for resource configuration
salloc: Nodes cn4275 are ready for job
srun: error: x11: no local DISPLAY defined, skipping
error: unable to open file /tmp/slurm-spank-x11.6385785.0
slurmstepd: error: x11: unable to read DISPLAY value

Created 1 generic SSH tunnel(s) from this compute node to
biowulf for your use at port numbers defined
in the $PORTn ($PORT1, ...) environment variables.


Please create a SSH tunnel from your workstation to these ports on biowulf.
On Linux/MacOS, open a terminal and run:

    ssh  -L 45081:localhost:45081 wuz8@biowulf.nih.gov

For Windows instructions, see https://hpc.nih.gov/docs/tunneling

[wuz8@cn4275 wuz8]$ module load jupyter
[+] Loading git 2.39.2  ...
[+] Loading jupyter
[wuz8@cn4275 wuz8]$
```

Figure 4: Go back to the terminal (Mac) or command prompt (Windows) with the interactive session (look for cn#### at the prompt). Do `module load jupyter` from here.

```
[wuz8@cn4275 wuz8]$ jupyter lab --ip localhost --port $PORT1 --no-browser
    To access the server, open this file in a browser:
        file:///spin1/home/linux/wuz8/.local/share/jupyter/runtime/jpserver-363837-open.html
    Or copy and paste one of these URLs:
        http://localhost:45081/lab?token=ad4b828f83a0fd8ad468cadaed56590b8a34f7f0418e76f3
     or http://127.0.0.1:45081/lab?token=ad4b828f83a0fd8ad468cadaed56590b8a34f7f0418e76f3
```

Copy either of the http links to local browser

Figure 5: Start a Jupyter lab session using `jupyter lab --ip localhost --port $PORT1 --no-browser` and copy and paste either one of the http links to a local browser.

Practice questions

Lesson 2 practice questions

Question 1

Generate a list called `twelve` that contains numbers 1 through 12 and then afterwards, subset it to a list called `even_numbers` that contains only the even entries.

> **Hint**
>
> Google how to find the remainder of a division operation.

{{Sdet}}{{Ssum}}solution{{Esum}}

```
twelve=[1,2,3,4,5,6,7,8,9,10,11,12]
```

```
even_numbers=list()
for i in twelve:
    if i % 2 == 0:
        even_numbers.append(i)
```

OR

```
even_numbers=list()
even_numbers=[i for i in twelve if i % 2 == 0]
```

OR

```
even_numbers=list(filter(lambda i: i % 2 == 0, twelve))
```

{{Edet}}

Question 2

Create a list called `numeric` grades that contains 90, 75, 80, 95, and 100. Then loop through `numeric_grades` and print the student's letter grade using the following criteria.

- `>=90: A`
- `<90 but >=80: B`

- <80 but >=70: C
- <70 but >=60: D
- Below 60: Failed

Hint

Use Google to find out how to make multiple comparisons within Python's elif statement.

```
numeric_grades=[90,75,80,95,100]
student_name=['Yoda', 'Cat', 'Dog', 'Mouse', 'Spock']
```

{{Sdet}}{{Ssum}}solution{{Esum}}

```
for i in range(len(numeric_grades)):
    if numeric_grades[i]>=90:
        print(student_name[i], "got an A")
    elif (numeric_grades[i]<90) & (numeric_grades[i]>=80):
        print(student_name[i], "got a B")
    elif (numeric_grades[i]<80) & (numeric_grades[i]>=70):
        print(student_name[i], "got a C")
    elif (numeric_grades[i]<70) & (numeric_grades[i]>=60):
        print(student_name[i], "got a D")
    else:
        print(student_name[i], "Failed")
```

{{Edet}}

Lesson 3 practice questions

Question 1

Import hcc1395_chr22_rna_seq_counts.csv and store it as hcc1395_chr22_counts.

{{Sdet}}{{Ssum}}Solution{{Esum}}

```
import pandas
```

```
hcc1395_chr22_counts=pandas.read_csv("./hcc1395_chr22_rna_seq_counts
```

{{Edet}}

Question 2

How many rows and columns are in hcc1395_chr22_counts?

{{Sdet}}{{Ssum}}Solution{{Esum}}

```
hcc1395_chr22_counts.shape
```

```
(1335, 7)
```

{{Edet}}

Question 3

What are the column names in hcc1395_chr22_counts and how to view the first 10 rows of this data set?

{{Sdet}}{{Ssum}}Solution{{Esum}}

```
hcc1395_chr22_counts.head(10)
```

Alternatively, use `hcc1395_chr22_counts.columns` to get the column headings for this data frame.

{{Edet}}

Question 4

How many genes start with the letter "C" in hcc1395_chr22_counts?

{{Sdet}}{{Ssum}}Solution{{Esum}}

```
hcc1395_chr22_counts.loc[hcc1395_chr22_counts.loc[:,'Geneid'].str.sta
```

{{Edet}}

Question 5

Import hcc1395_deg_chr22.csv and store it as hcc1395_deg_chr22.

{{Sdet}}{{Ssum}}Solution{{Esum}}

```
hcc1395_deg_chr22=pandas.read_csv("./hcc1395_deg_chr22.csv")
```

{{Edet}}

Question 6

Remove ".bam" from the column headers of hcc1395_deg_chr22.

{{Sdet}}{{Ssum}}Solution{{Esum}}

```
hcc1395_deg_chr22.columns=hcc1395_deg_chr22.columns.str.replace(".bar
```

{{Edet}}

Question 7

Subset out the following columns from hcc1395_deg_chr22 and store it as hcc1395_deg_chr22_1.

- name
- log2FoldChange
- PAdj

{{Sdet}}{{Ssum}}Solution{{Esum}}

```
hcc1395_deg_chr22_1=hcc1395_deg_chr22.loc[:,["name", "log2FoldChange'
```

Use the `.head` function to check of the subsetting was done correctly.

```
hcc1395_deg_chr22_1.head()
```

{{Edet}}

Question 8

Add a column to hcc1395_deg_chr22_1 that contains the negative log10 of the PAdj value.

{{Sdet}}{{Ssum}}Solution{{Esum}}

```
import numpy
```

```
hcc1395_deg_chr22_1["-log10PAdj"]=numpy.negative(numpy.log10(hcc1395_
```

{{Edet}}

Lesson 4 practice questions

Question 1

Create a volcano plot for the differential expression analysis results for the hcc1395 data (hint: import hcc1395_deg_chr22_with_significance.csv)

{{Sdet}}{{Ssum}}Solution{{Esum}}

```
import pandas
import matplotlib.pyplot as plt
import seaborn
```

```
hcc1395_deg_chr22=pandas.read_csv("./hcc1395_deg_chr22_with_significa
```

```
plot1=seaborn.scatterplot(hcc1395_deg_chr22,x="log2FoldChange", y="-
plt.show()
```

{{Edet}}

Question 2

Label the two most differential expressed genes in the volcano plot. As a hint, first import hcc1395_deg_chr22_top_genes.csv. What message shows up upon running the code and suggest a plausible solution.

{{Sdet}}{{Ssum}}Solution{{Esum}}

```
hcc1395_deg_chr22_top_genes=pandas.read_csv("./hcc1395_deg_chr22_top_
```

```
plot1=seaborn.scatterplot(hcc1395_deg_chr22,x="log2FoldChange", y="-
for i, gene_name in enumerate(hcc1395_deg_chr22_top_genes["name"]):
    plot1.text(hcc1395_deg_chr22_top_genes["log2FoldChange"][i],
               hcc1395_deg_chr22_top_genes["-log10PAdj"][i],gene_name)
plt.show()
```

```
posx and posy should be finite values
```

{{Edet}}

Question 3

Import hcc1395_top_deg_normalized_counts.csv and create an expression heatmap. Use the Viridis color palette.

{{Sdet}}{{Ssum}}Solution{{Esum}}

```
hcc1395_top_deg_normalized_counts=pandas.read_csv("./hcc1395_top_deg_
```

```
plot2=seaborn.clustermap(hcc1395_top_deg_normalized_counts,z_score=0
                         figsize=(8,8),vmin=-1.5, vmax=1.5,cbar_kws=(
plt.show()
```

{{Edet}}

Question 4

Add a bar on the top of the heatmap that shows which treatment group the samples belong to.

{{Sdet}}{{Ssum}}Solution{{Esum}}

```
samples=pandas.Series({"hcc1395_normal_rep1":"orangered", "hcc1395_nc
column_colors = hcc1395_top_deg_normalized_counts.columns.map(sample
plot2=seaborn.clustermap(hcc1395_top_deg_normalized_counts,z_score=0
                         figsize=(8,8),vmin=-1.5, vmax=1.5,cbar_kws=(
                         col_colors=column_colors, cbar_pos=(0.05,0.8
plot2.ax_heatmap.set_xticklabels(plot2.ax_heatmap.get_xmajorticklabe
plot2.ax_cbar.tick_params(labelsize=12)
plot2.ax_col_colors.set_title("treatment",x=1.09,y=-0.3)
plt.show()
```

{{Edet}}

Finding help

The document provides useful links where participants can find help for the Python packages that were addressed during the course series.

Pandas - package for working with tabular data *(https://pandas.pydata.org)*

- Pandas API reference gives instructions for each command *(https://pandas.pydata.org/docs/reference/index.html)*. To get to the API reference, either
    - Navigate to the the Documentation section at the Pandas homepage and click on API reference (Figure 1).
    - OR, click on the the Documentation tab at the top of the Pandas homepage and click on the tile labeled API reference in the subsequent page (Figure 2).

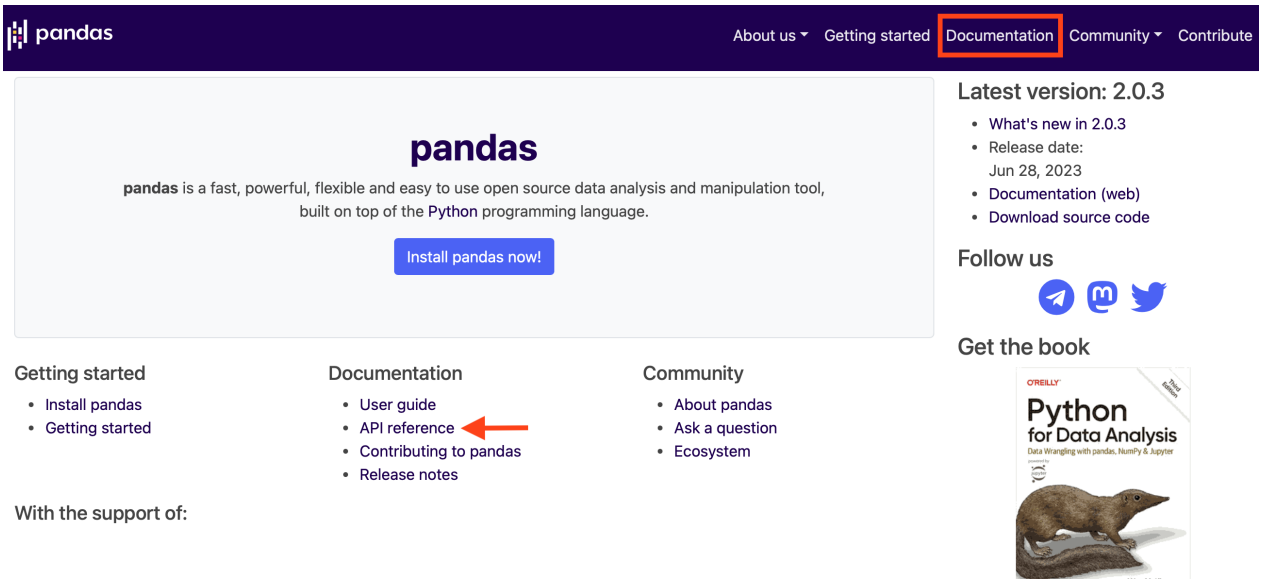


Figure 1

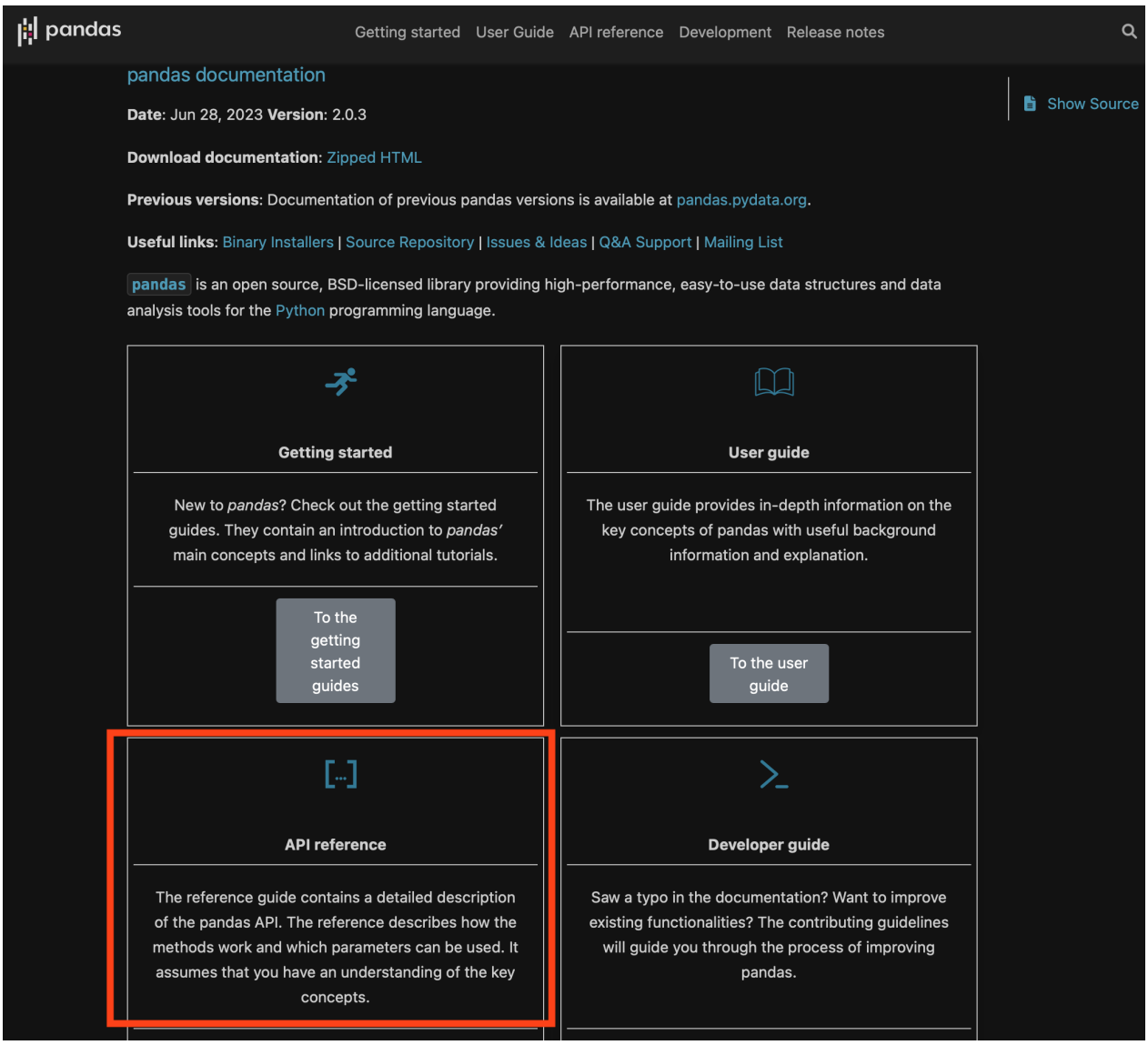


Figure 2

Seaborn for data visualization *(https://seaborn.pydata.org/index.html)*

- Seaborn API reference gives instructions for each command *(https://seaborn.pydata.org/api.html)*. To get to the Seaborn API reference, click on API at the top of the Seaborn website.

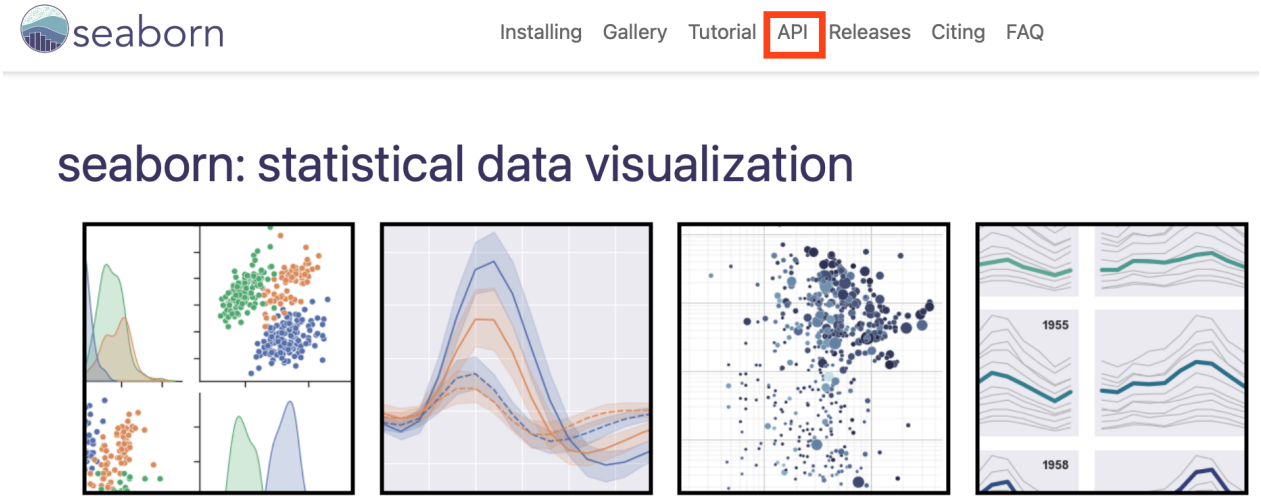


Figure 3

Numpy for scientific computing *(https://numpy.org/doc/stable/index.html)*

- Numpy API reference *(https://numpy.org/doc/stable/reference/index.html)*. To get to this, select Documentation at the top of the Numpy homepage (Figure 4) and then click on either of the links to the API reference (Figure 5).

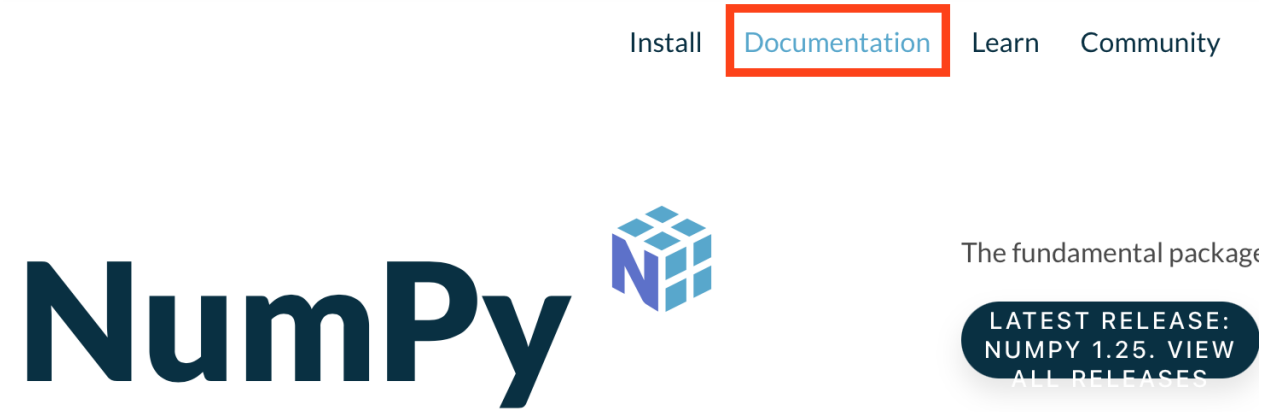


Figure 4

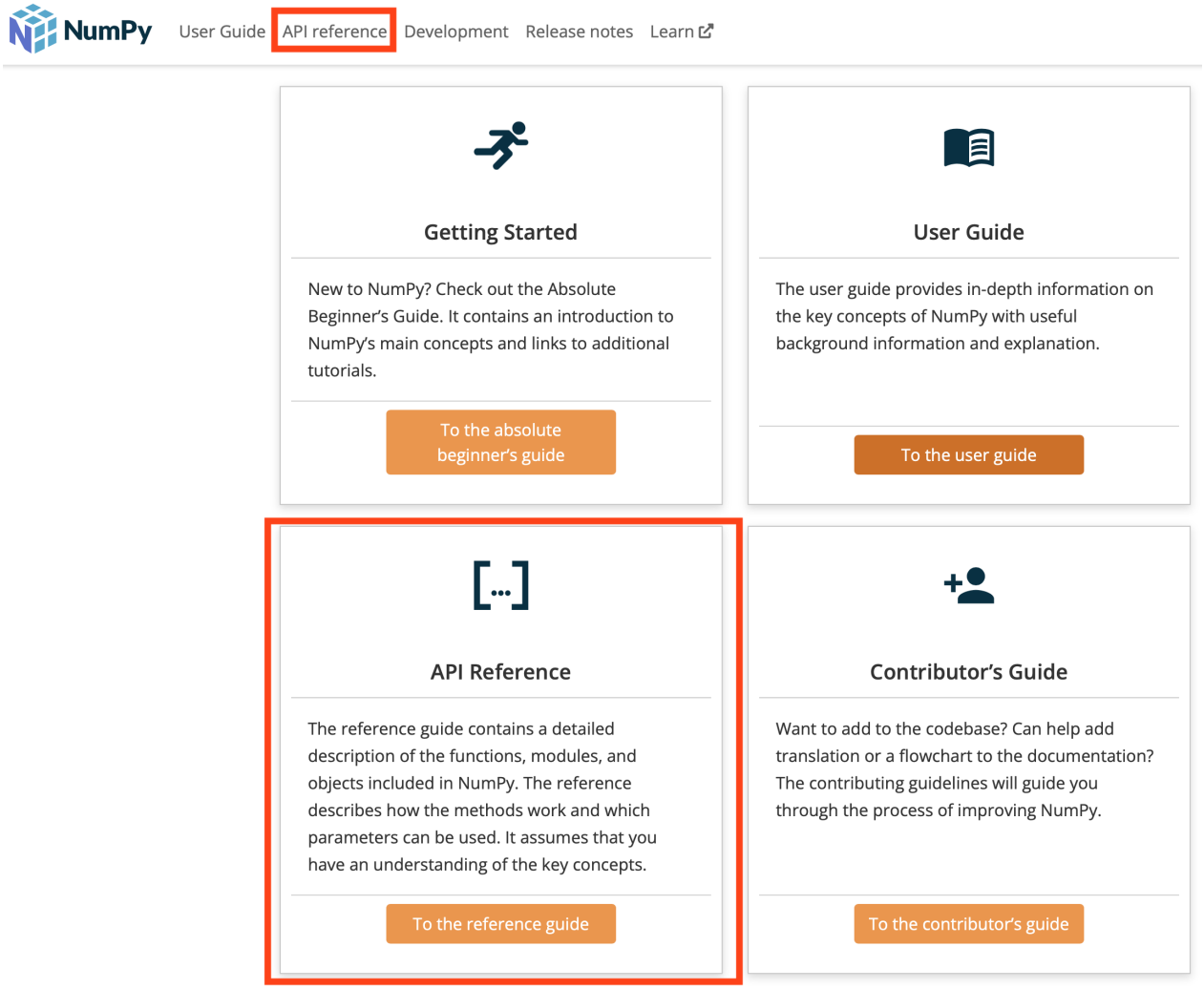


Figure 5

Matplotlib for data visualization *(https://matplotlib.org)*

- Matplotlib API reference *(https://matplotlib.org/stable/api/index)*. To get to this, click on reference at the top of the Matplotlib homepage (Figure 6).

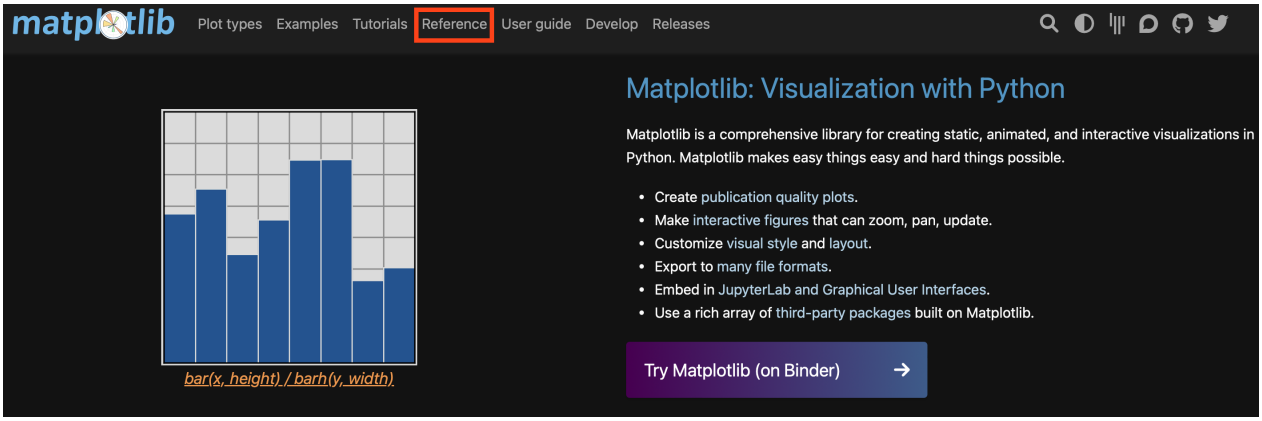


Figure 6