American Museum of Natural History

American Museum of Natural History
Richard Gilder Graduate School

RGGS Comparative Genomics 2 – Computational Methods (Session 5)

Jose Barba

Gerstner Scholar in Bioinformatics & Computational Biology

AMNH Arachnology Lab
Institute for Comparative Genomics
jbarba@amnh.org
josebarbamontoya.github.io

September 26, 2024
ICG conference room

Session 5 outline

- Complete tutorial of version control with GitHub from Session 4
- Introduction to R for phylogenomics
- Introduction to Python for phylogenomics

Additional matters

- **Any question about your paper presentations for Session 6?**
  - Each student will choose an -omics paper that they find innovative, exciting, relevant to their work, or particularly interesting. On October 10, they will deliver a 10-minute presentation providing a concise overview of the research question addressed and a thorough explanation of the computational methods employed
- **The quiz scheduled for the next session will be an oral recapitulation of earlier topics**
- **Session 13 — November 28**
  - Reschedule the class to December 5 (double session)

Version control with GitHub

- **Basic Git commands:**
  - `git init`: Initialize a Git repository
  - `git clone <repo>`: Clone a repository to your local machine
  - `git add <file>`: Stage a file for a commit
  - `git add .`: Stage all changes for the next commit
  - `git commit -m "message”`: Commit changes with a message
  - `git push`: Push local changes to the GitHub repository
  - `git pull`: Fetch and merge changes from the remote repository to your local one
  - `git checkout -b branch-name`: Create and switch to a new branch
  - `git merge branch-name`: Merge another branch into the current one
  - `git log`: check the history of commits

Version control with GitHub

- **A tutorial for setting up and using GitHub with Git, particularly focused on version control through SSH and Git basics.**
  - It walks through essential steps, including generating SSH keys, setting up a repository, making commits, and handling branching, merging, and conflicts

- **Instructions to download the GitHub version control tutorial to the home directory:**
  1. Open the terminal
  2. Type `cd ~`
  3. Enter the following command `wget https://raw.githubusercontent.com/josebarbamontoya/rggs_comparative_genomics_2/main/session_04/github_version_control_tutorial.sh`, if `wget` is not available, use `curl -0` instead

Introduction to R for phylogenomics

- **What is R? A programming language designed specifically for statistical computing and graphics, as well as data manipulation**
  - Widely used scripting language in academia
  - Easily accessible —open source and intuitive
  - Good for data wrangling and and crunching
  - Super good for creating publication quality figures
  - Has a lot of packages and functions to help you solve your research questions
  - You can write your own functions and packages

Introduction to Python for phylogenomics

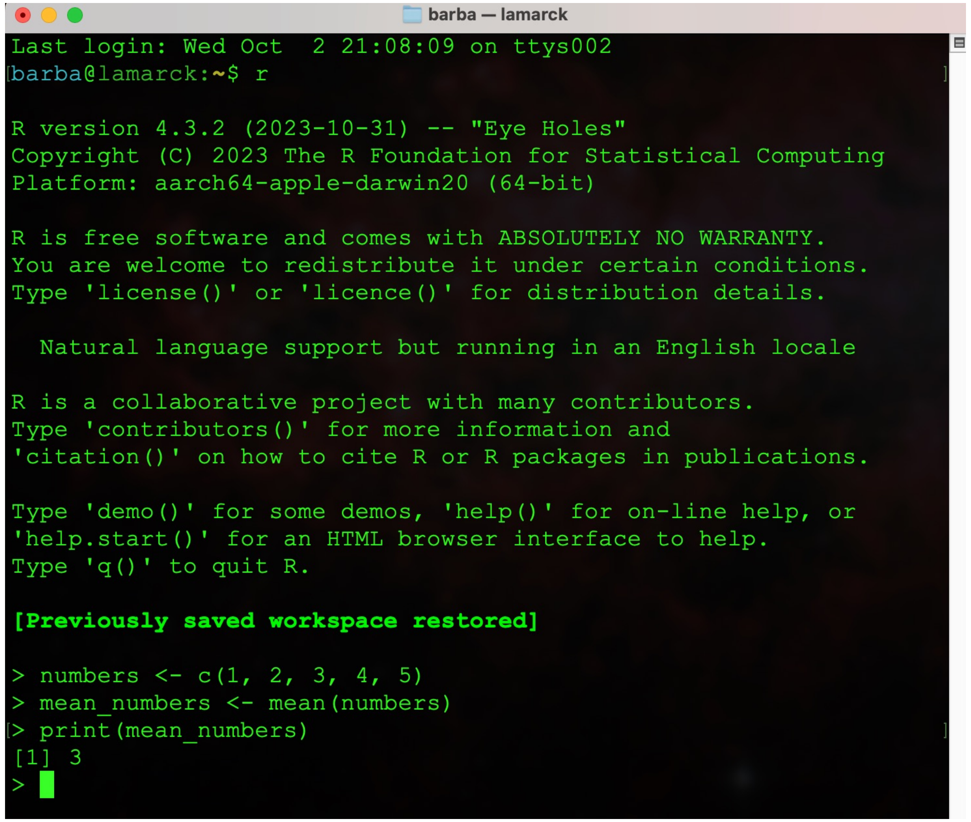
- **What is Python? A programming language designed for a wide range of applications, including statistical analysis and graphics, software and web development, as well as data manipulation**
  - Widely used scripting language in academia and industry
  - Easily accessible – open source and intuitive
  - Powerful for data wrangling and crunching
  - Super good for creating publication quality visualizations
  - Has a lot of libraries and frameworks to solve diverse research problems
  - You can write your own functions, packages, and modules

So, should I use R or Python for my analysis?

- R is a specialized programming language designed for statistical analysis and data visualization
- Python is a versatile and powerful programming language, suitable for a broader range of applications, including data science
- The choice between R and Python depends on the specific needs of the project and the user's background
- Both languages can be integrated using RStudio, JupyterLab, or the Conda and reticulate packages

Introduction to R for phylogenomics

- **How can I use R?**
  - Interactively in the command line
    > r
  - As a scripting language in the command line
    > rscript my_script.r
  - Interactively in RStudio
    https://posit.co/downloads/
  - \`<-\` instead of \`=\` to assign values
  - Generally, functions do not work with missing (NA) values. You may need to use the argument na.rm = TRUE to ignore them



```
Last login: Wed Oct  2 21:08:09 on ttys002
barba@lamarck:~$ r

R version 4.3.2 (2023-10-31) -- "Eye Holes"
Copyright (C) 2023 The R Foundation for Statistical Computing
Platform: aarch64-apple-darwin20 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

> numbers <- c(1, 2, 3, 4, 5)
> mean_numbers <- mean(numbers)
> print(mean_numbers)
[1] 3
> 
```

Introduction to R for phylogenomics

- **Installing functions and packages**
  - Install from CRAN (https://cran.r-project.org/)

    ```
    > install.packages(“example_package”)
    ```
  - Install from BioConductor (http://bioconductor.org/)

    ```
    > if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager") BiocManager::install(”example_package")
    ```
  - Install from source in R

    ```
    > install.packages(path_to_file, repos = NULL, type="source")
    ```
  - Install from source in the terminal

    ```
    > R CMD INSTALL example_package.tar.gz
    ```

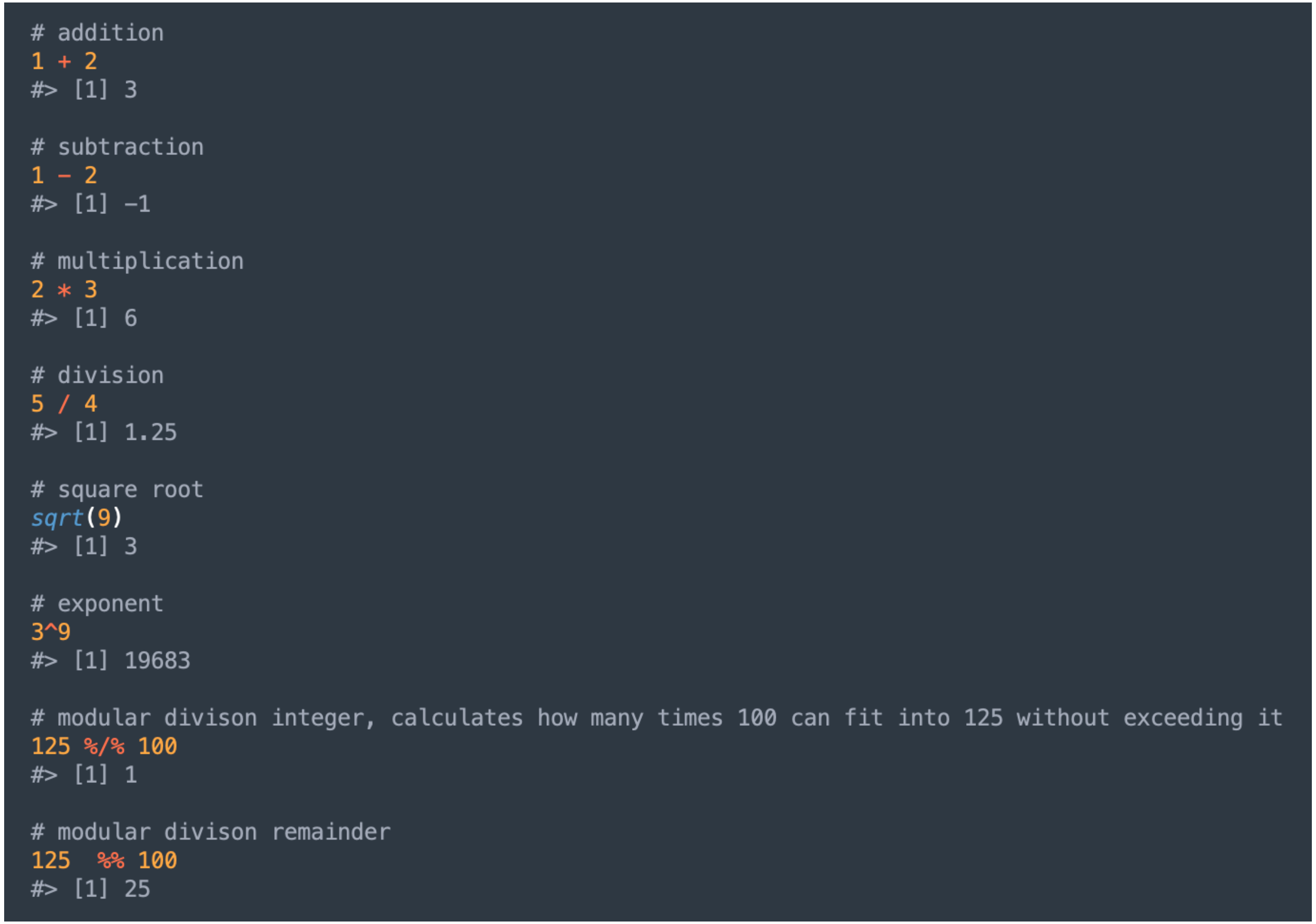
Introduction to R for phylogenomics

- **An introductory tutorial on R for phylogenomics**
  - It explores data manipulation, analysis, and visualization within the context of phylogenomics

- **Instructions to download the tutorial to the home directory:**
  1. Open the terminal
  2. Type `cd ~`
  3. Enter the following command `wget https://raw.githubusercontent.com/josebarbamontoya/rggs_comparative_genomics_2/main/session_05/r_tutorial.r`, if `wget` is not available, use `curl -0` instead

Introduction to R for phylogenomics

- R as a calculator



```
# addition
1 + 2
#> [1] 3

# subtraction
1 - 2
#> [1] -1

# multiplication
2 * 3
#> [1] 6

# division
5 / 4
#> [1] 1.25

# square root
sqrt(9)
#> [1] 3

# exponent
3^9
#> [1] 19683

# modular divison integer, calculates how many times 100 can fit into 125 without exceeding it
125 %/% 100
#> [1] 1

# modular divison remainder
125  %% 100
#> [1] 25
```

Introduction to R for phylogenomics

- **R data types and structures**
  - integer: whole numbers, like a chromosome position (e.g., 15739170)
  - numeric: decimal values, such as GC content (e.g., 0.4281)
  - factor: categorical variables, like nucleotide types (A, C, G, T)
  - logical: Boolean values (TRUE/FALSE) for conditions, such as whether a site is a CpG site
  - null: empty or non-existent values (e.g., NA)
  - character: text strings, such as gene names (e.g., "BRCA1")
  - complex: complex numbers with real and imaginary parts (e.g., `2 + 3i`)
  - list: an ordered collection of varied objects (e.g., `list(name = "Sample1", values = c(1, 2))`)
  - data frame: a table structure with columns of different types and rows as observations (e.g., genomic features)
  - matrix: a two-dimensional array with uniform element types (e.g., gene expression data).
  - array: a multi-dimensional extension of a matrix (e.g., a three-dimensional array for multiple measurements)
  - tibble: a modern data frame with better printing and subsetting (e.g., `tibble(gene = c("gene1", "gene2"), expression = c(5.3, 2.1))`)

Introduction to R for phylogenomics

- **Packages for phylogenomic data manipulation, analysis and visualization**
  - ape: tools for analysis of phylogenetics and evolution
  - bioconductor: a repository of r packages for bioinformatics
  - dendextend: enhancements for dendrogram functionality
  - dplyr: tools for data manipulation
  - geiger: tools for analyzing evolutionary rates
  - ggplot2: comprehensive data visualization capabilities
  - ggtree: visualization and manipulation of phylogenetic trees
  - msa: multiple sequence alignment package
  - phangorn: phylogenetic reconstruction and analysis tools
  - phytools: tools for phylogenetic comparative biology
  - phyclust: phylogenetic clustering methods
  - treeio: input and output functionalities for tree data
  - tidyverse: a collection of r packages designed for data science
  - ips: interface to various phylogenetic software

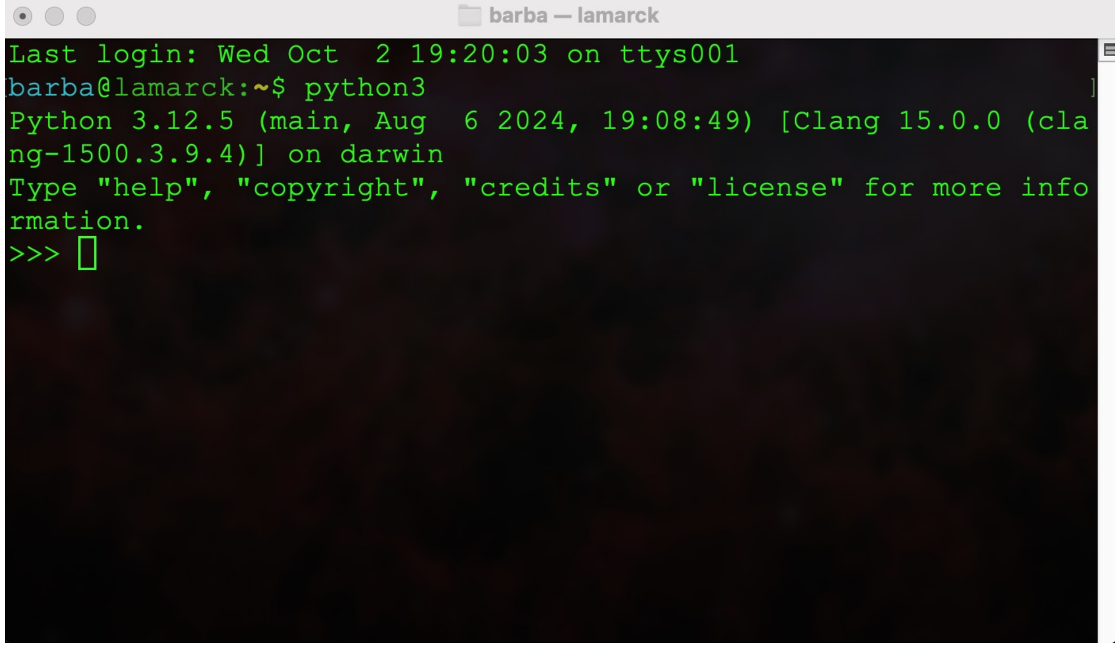
Introduction to Python for phylogenomics

- **How can I use Python?**
  - Interactively in the command line

    > python
  - As a scripting language in the command line

    > python my_script.py
  - Interactively in JupyterLab

    https://jupyterlab.readthedocs.io/en/latest/



```
Last login: Wed Oct  2 19:20:03 on ttys001
barba@lamarck:~$ python3
Python 3.12.5 (main, Aug  6 2024, 19:08:49) [Clang 15.0.0 (cla
ng-1500.3.9.4)] on darwin
Type "help", "copyright", "credits" or "license" for more info
rmation.
>>> 
```

Introduction to Python for phylogenomics

- **Installing functions, modules, and packages**
  - Install from PyPI (https://pypi.org/)

    ```
    > pip install “example_package”
    ```
  - Install from Conda (https://docs.conda.io/projects/conda/en/latest/user-guide/install/index.html)

    ```
    > conda install “example_package”
    ```
  - Install from Homebrew (https://brew.sh/)

    ```
    > brew install “example_package”
    ```
  - Install from a requirements file

    ```
    > pip install -r requirements.txt
    ```
  - Install from source in Python

    ```
    > pip install path_to_package_directory
    ```
  - Install from source in the terminal

    ```
    > python setup.py install
    ```

Introduction to Python for phylogenomics

- **Python data types and structures**
  - int: whole numbers, like a chromosome position (e.g., 15739170)
  - float: decimal values, such as GC content (e.g., 0.4281)
  - str: text strings, such as gene names (e.g., "BRCA1")
  - bool: Boolean values (True/False) used for conditions, such as whether a gene is expressed or not
  - none: empty or non-existent values (e.g., None)
  - list: an ordered collection of varied objects (e.g., `[1, 2, "Sample1"]`)
  - tuple: an immutable ordered collection of varied objects (e.g., `(1, 2, "Sample1")`)
  - dict: a collection of key-value pairs, like genomic features (e.g., `{"gene": "BRCA1", "expression": 5.3}`)
  - set: an unordered collection of unique elements (e.g., `{1, 2, 3}`)
  - array: a multi-dimensional array from the NumPy library for numerical computations (e.g., `numpy.array([[1, 2], [3, 4]])`)
  - dataframe: a table structure from the pandas library with columns of different types and rows as observations (e.g., `pd.DataFrame({"gene": ["gene1", "gene2"], "expression": [5.3, 2.1]})`)

Introduction to Python for phylogenomics

- **Modules and packages for phylogenomic data manipulation, analysis and visualization**
  - biopython: a set of tools for biological computation, including sequence analysis and phylogenetics
  - pandas: data manipulation and analysis library for structured data
  - scikit-bio: tools for bioinformatics, including sequence alignment and phylogenetic analysis
  - dendropy: library for phylogenetic computing, including manipulation of phylogenetic trees
  - matplotlib: plotting library for creating static, animated, and interactive visualizations
  - phylo: a module in Biopython for working with phylogenetic trees
  - ete3: toolkit for the analysis and visualization of trees
  - pyscaffold: tool for constructing and analyzing phylogenetic trees and their relationships
  - numpy: library for numerical computations and support for large multi-dimensional arrays and matrices
  - csv: module for reading and writing csv files, useful for data input and output
  - sys: module for accessing system-specific parameters and functions, useful for interacting with the Python runtime
  - rpy2: interface to R from Python, allowing the use of R packages and functions within Python code
  - os: module for interacting with the operating system, providing functionalities for file and directory management
  - io: deal with input and output (I/O) operations

Introduction to Python for phylogenomics

- **An introductory tutorial on python for phylogenomics**
  – It explores data manipulation, analysis, and visualization within the context of phylogenomics

- **Instructions to download the tutorial to the home directory:**
  1. Open the terminal
  2. Type `cd ~`
  3. Enter the following command `wget https://raw.githubusercontent.com/josebarbamontoya/rggs_comparative_genomics_2/main/session_05/python_tutorial.py`, if `wget` is not available, use `curl -0` instead