American Museum of Natural History

American Museum of Natural History
Richard Gilder Graduate School

RGGS Comparative Genomics 2 – Computational Methods (Session 4)

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 4 outline

- **Are there any questions about session 3?**
- **Working on remote servers: HPC architecture, using a scheduler, submitting jobs**
- **Scripting**
- **Version control with GitHub**

Working on remote servers: HPC architecture, using a scheduler, submitting jobs

- **Connecting to remote servers using the terminal and a package manager (Cyberduck) — covered in Session 2:**
  - E.g., connect to the AMNH-Huxley server: **`ssh amnh_username@huxley-master.pcc.amnh.org`**
  - E.g., transport a directory from your computer to the AMNH-Huxley server: **`scp -r your directory amnh_username@huxley-master.pcc.amnh.org:/home/amnh_username`**

Working on remote servers: HPC architecture, using a scheduler, submitting jobs

- **High performance computing (HPC) architecture:**
  - We well work on a workshop on HPC at AMNH by Apurva Narechania, available at: https://amnh.sharepoint.com/sites/Bioinformatics/SitePages/HPCs-at-AMNH.aspx
  - The AMNH-HPC workshop notes are also available at: https://github.com/josebarbamontoya/rggs_comparative_genomics_2/blob/main/session_04/HPC_workshop_10292018_Apurva_Narechania.pdf

- **PBS scheduler:**
  - A tutorial of PBS by Sajesh Singh is available at: https://github.com/josebarbamontoya/rggs_comparative_genomics_2/blob/main/session_04/PBS_Tutorial_Sajesh_Singh.pdf

Working on remote servers: HPC architecture, using a scheduler, submitting jobs

- **A tutorial for HPC and scripting**
  - This is a comprehensive shell (bash) script tutorial for setting up an HPC environment and executing various bioinformatics tasks. It covers the creation of scripts, running software like BLAST and RAxML, and using PBS (Portable Batch System) for job scheduling in Huxley

- **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/hpc_and_scripting_tutorial`, if `wget` is not available, use `curl -0` instead

Scripting

- **In computing, scripting refers to the use of scripts—small programs written in programming languages—to automate, facilitate, or enhance computational tasks**
- **Common scripting languages include Bash (Unix shell), Python, R, and MATLAB. Each of these languages has specific libraries and frameworks that can be leveraged for different types of computational tasks**

Scripting

- **These scripts can perform a wide range of functions, including:**
  - Data analysis: Automating the processing and analysis of large datasets, such as statistical analyses, simulations, or data visualization
  - Workflow automation: Streamlining repetitive tasks in computational workflows, such as data collection, transformation, and reporting
  - Algorithm implementation: Coding algorithms for specific computational tasks, such as genome assembly and analysis, phylogenetic reconstruction, or machine learning implementation
  - Integration of tools: Facilitating communication between different software tools or systems, allowing for a more cohesive analysis pipeline
  - Reproducibility: Ensuring that computational analyses can be easily reproduced by documenting the steps taken in a script

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

Version control with GitHub

- **Resources for further learning:**
  - GitHub Docs: https://docs.github.com/en/get-started
  - Git Documentation: https://git-scm.com/doc
  - Git Cheat Sheet: https://education.github.com/git-cheat-sheet-education.pdf