

Introduction to NGS Data

Matt Gitzendanner magitz@ufl.edu

Oleksandr Moskalenko om@hpc.ufl.edu

Getting Data to HPC

SFTP client to move files
to / from your computer



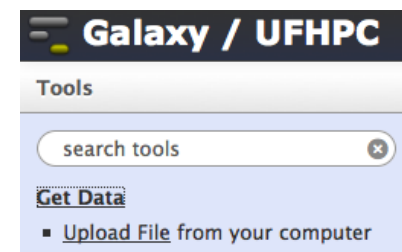
e.g.: Cyberduck, FileZilla

ssh client to connect to
submit.hpc.ufl.edu



e.g.: Terminal, PuTTY

- ▶ SFTP, ssh, bbcp
- ▶ Have ICBR do it for you!
- ▶ We can assist



Large data import into Galaxy

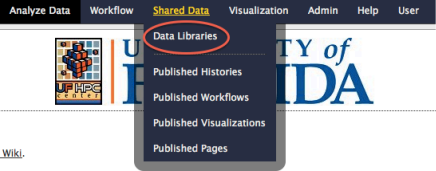
- ▶ Wiki page has step-by-step directions
 - http://wiki.hpc.ufl.edu/doc/Galaxy_Data_Import

Inside the Galaxy

Create your personal incoming folder

Please note that at the moment the Galaxy admins must enable shared library access for each new user. If you don't see the options listed below please email galaxy@hpc.ufl.edu and request access.

- Move the mouse pointer to the "Shared Data" menu. A menu will open. Click on the "Data Libraries" menu item.



Wiki:

- The list of shared data libraries will be shown in the main window. Click on the "Incoming" data library.

Data Libraries

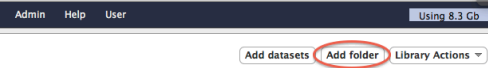
search dataset name, info, message, dbke

Advanced Search

Data library name ↓

- C_neo_MutX
- Ensembl Regulatory Build 63
- GMS 6001 MACS Exercise
- Incoming**
- Test Phylogenetic Data
- Training datasets

- Click on the "Add folder" button in the top right corner.



Uploaded By	Date	File Size
-------------	------	-----------

- Fill out the form to create a folder named after your galaxy user name (for example `jdco@ufl.edu`).

Galaxy / UF HPC

Quality Assessment

- ▶ Poor quality reads lead to problems in assembly and mapping
 - Need to remove adapters
 - Trim low quality sequence
 - Remove identical sequences
 - PCR duplicates
- ▶ Quality assessment
 - FastQC
 - Galaxy and CLI

- Take care with:
 - Quality encoding
 - Paired-end data

Quality Filter

▶ Sickle

- Especially for filtering paired-reads
 - Creates a file of single reads for pairs where one read is discarded and the other is good

FastQ paired records kept: 3584484 (1792242 pairs)

FastQ single records kept: 216860 (from PE1: 25453, from PE2: 191407)

FastQ paired records discarded: 44980 (22490 pairs)

FastQ single records discarded: 216860 (from PE1: 191407, from PE2: 25453)

▶ Fastxtoolkit

- Has many great utilities
- Does not handle filtering of paired-reads

Compressed data formats

▶ Gzip

- A standard file compression tool
- Files end in .gz
- Many NGS applications can natively process gzipped data files
 - Use when possible: reduces storage needs, but also disk and network I/O in analyzing your data

Things to watch out for

- ▶ Converting quality formats unintentionally
 - Keep track of input/output formats of data manipulation tools
 - Some will convert for you
 - Can be helpful, but need to know what tool is doing
- ▶ Don't need to transfer/keep all data provided by ICBR to HPC
 - Definitely keep the data
 - Can reprocess raw data or view some run stats
 - Rarely used though
 - Full folder from 1 lane: 54G
 - BaseCalls folder from 1 lane: 17GB

Training Schedule

- ✓ Jan 14: Intro to UFHPC, getting started
- ✓ Jan 28: The Linux/Unix Shell - An Introduction
- ✓ Feb 4: Running Jobs, Submission Scripts, Modules
- ✓ Feb 11: Dr. Dhruva Chakravorty: Amber
- ✓ Feb 18: Galaxy Overview, The Basics
- ✓ Feb 25: Dr. David Ostrov: Molecular Docking
- ✓ Mar 11: NGS Data Techniques: General Methods and Tools
- ▶ Mar 18: NGS: Reference Based Mapping & de Novo Assembly
- ▶ Mar 25: Phylogenetic Analyses
- ▶ Apr 1: Multiprocessing at the HPC Center
- ▶ Apr 8: Introduction to GPU nodes
- ▶ Apr 15: Tentative: Overview of the new cluster and storage
- ▶ Apr 22:
- ▶ May 2: Spring 2013 Research Computing Day (noon-4pm)

UF Research Computing

- ▶ Help and Support (Continued)
 - <http://wiki.hpc.ufl.edu>
 - Documents on hardware and software resources
 - Various user guides
 - Many sample submission scripts
 - <http://hpc.ufl.edu/support>
 - Frequently Asked Questions
 - Account set up and maintenance

