Amy Stonelake

March 3, 2021

Bioinformatics for Beginners, Post-bac Edition

3 sets of 2 week-long, 2x/week, one-hour sessions, that makes 4 hours per topic

I just got my (RNA-Seq) data back from the sequencing center, now what?

Teaching will be done using the GOLD learning environment

Session 1, April – Unix command line, getting a Biowulf account, decompressing data

Session 2, May – RNA-Seq Quality control, alignment to genome, differential gene expression, pathway analysis (great StatQuest Video, https://www.youtube.com/watch?v=tlf6wYJrwKY)

Session 3, June – R for data analysis and visualization ??? OR transitioning to Biowulf

Stand-alone one-time webinars on NCI prop software (make sure software license NCI wide, not just CCR, check with DG and MM)

Schedule THESE NOW! They will happen outside of regular class times.

Partek Flow (NCI) – need/will benefit from basic Unix skills to work with this.

Snapgene

Qiagen IPA

One-Page Syllabus for “Bioinformatics for Beginners, Post-bac Edition” (3 courses)

(URL)

Instructors: Amy Stonelake

Co-Instructors: Peter FitzGerald, Carl McIntosh, Des Tillo

To participate in this course, you need your NCI-issued government computer and a reliable internet connection.

All classes and help sessions will be held in Amy Stonelake’s Personal Room:

<https://cbiit.webex.com/meet/stonelakeak>

This class will be taught with the GOLD learning environment on the DNAnexus platform. Every learner will get their own login and password.

This is the first of three courses we have developed for beginning bioinformatics learners with the scenario “I’ve just gotten my sequence data back from the Sequencing Center, how do I understand and analyze it?”.

Getting help in this course:

1. Breakout room during class if you are lost or confused
2. Course Question and Answer Forum (URL here) will be used to answer questions about the course.
3. Scheduled help sessions.

(A class listserv for announcements – this is admin)

Session One – Why learn Bioinformatics? And Beginner Unix.

1. Tues, 3/30
2. Thursday, 4/1
3. Tues, 4/6
4. Thursday, 4/8

Session Two Intro to bulk RNA-Seq

1. Tues, 5/4
2. Thursday, 5/6
3. Tues, 5/11
4. Thurs, 5/13

Session Three – continue with RNA-Seq, Intro to Biowulf and Globus

1. Tues, 6/8
2. Thurs, 6/10
3. Tues, 6/15
4. Thurs, 6/17

**Why learn Bioinformatics? And Beginner Unix:**

1. Who should take this course:

Learners who want to work with Next Gen Sequence data

1. Pre-requisites: None, this class if for beginner level bioinformatics learners.
2. Learning Objectives

In the class learners will be able to:

1. Understand why every bench scientist should learn some bioinformatics
2. Log into and utilize the GOLD learning environment for class content and lessons
3. Work with Unix files and directories to manage Next Gen Sequencing data and associated files
4. Understand data formats (FASTA, FASTQ) and learn how to work with them at the Unix command line

**Beginner RNA-Seq: (will we need more than 4 classes for this topic? Probably yes.)**

1. Who should take this course?
2. Pre-requisites: Learners should be (1) familiar with the GOLD learning environment and have (2) beginner level knowledge of working in a Unix environment
3. Learning Objectives:

Learners will:

1. Understand the basics of bulk RNA-Seq, how the experiments should be set up for best results and problems to avoid (batch effects).
2. Interpret the information returned by the sequencing centers and determine next steps of analysis (Quality control, etc.)
3. Perform the beginning steps of bulk RNA-Seq analysis, including assaying sequence quality, sequence data trimming/ removing adapters and alignment to genome
4. Identify resources available to them within NCI to help them analyze their data such as proprietary software (Partek Flow, NIDAP, etc.).

Prerequisites: Learners should be (1) familiar with the GOLD learning environment, (2) have beginner level Unix command line skills and (3) understand the steps of bulk RNA-Seq analysis including data formats, inputs and outputs

Learning Objectives in the **Analyzing bulk RNA-Seq Data** class

After taking this class, learners will

1. Understand all the steps of a complete bulk RNA-Seq analysis, from FASTQ files to biological pathway analysis
2. Visualize sequence data graphically with the Integrative Genome Viewer (IGV) tool
3. Log in to the NIH High Performance Cluster Biowulf, understand the directory structure and how to work with files
4. Use the “Globus” service to transfer and share data files

Learning Objectives in the **Data Visualization with R** class

Learners will:

1. Become familiar with the R and RStudio programs in different environments, including GOLD, Biowulf, and their own computer
2. Import data into the R environment and manipulate it
3. Create visualizations of bulk RNA-Seq data including volcano plots…

Will need to reserve Biowulf student accounts on some days…

Meeting Dates:

Help sessions (including pre-sessions):

Class Discussion Board:

I just got my data back from the sequencing center. Now what? (moving big data with Globus, Biowulf accounts, requesting enough space, interpreting sequence quality data, good habits for naming files and directories).

NCI and CCR also have purchased software for various types of analysis. These platforms offer a Graphical User Interface (GUI) for point-and-click analyses. It is recommended you attend a training if you are interested in any of these software tools:

Partek Flow for bulk and single cell RNA-Seq

SnapGene and/or Geneious Prime

Qiagen Ingenuity Pathway Analysis (IPA)

For a complete list of software available to CCR scientists, please see: <https://ostr.ccr.cancer.gov/bioinformatics/software/>

Beginner Unix for Scientists, topics to be covered:

Intro Class One, Tuesday, 3/30

1. Bioinformatics: What is it good for?
2. GitHub resource for class content
3. Introduction to the GOLD online learning platform (everyone login)
4. How to post to the Class Question and Answer Forum

Class Two, Thursday, 4/1

1. First Unix command (ls)
2. Where am I? (pwd)
3. Creating, and removing files (touch, nano editor, rm) and directories (mkdir, rmdir).
4. Good names for files\_and\_directories and bad names for files and directories. (unix does not like spaces in file names and certain characters are forbidden).
5. Moving and renaming files, all with one command! (mv)
6. Moving around the directory structure (cd commands, absolute and relative paths, root directory /, take me home!)
7. Flags and command options - making programs do what they do
8. Working with large sequence data files (decompressing files and working with compressed files).
9. Using tab complete and your Unix history to make life easier
10. Help, help I need some help! (man pages, help pages and the Internet is your friend)

Class Three, Tuesday, 4/6

1. Quick review of class one Unix commands
2. Working with file content (input and output, append)
3. Working with files within the directory tree.
4. Less is more and more is less.
5. A very useful function, wc.
6. What's that echo? Am I on the right $PATH?
7. Combining commands with pipe (|). Where the heck is pipe anyway?
8. Finding information in files (grep, head, tail, sort and stuff like that)
9. Permissions - when all else fails check the permissions

Final Class Four, Thursday, 4/8

1. Intro to NIH High Performance Cluster “Biowulf”, home and data dir, requesting enough space for your data and analysis, sinteractive, swarm and batch, don’t work on the login node
2. Getting a Globus Account
3. Login to Biowulf with Student ID and execute some Unix commands (review)

Incorporate "quick quizzes" live polls during the class to reinforce concepts.

What will class look like?

One hour long

Webex

Breakout room for extra help/ I am lost or confused.

Polls to break the ice and quick quizzes to determine uptake of material.

~~Pre-sessions will be held to get everyone on the GOLD system.~~

Create login/password for everyone plus a couple of student1, student2 for late arrivals.

1. Login to GOLD
2. Execute a Unix command? (pwd or ls)
3. ~~Email~~ [~~ncibtep@nih.gov~~](mailto:ncibtep@nih.gov) ~~with names of files or something to show you’ve logged in.~~

Create GitHub pages for class, use unix review and other materials, maybe even from Data Carpentry.

Schedule: Bioinformatics for Beginners, Post-Bac Edition

Classes will be held at 3 PM via WebEx (use Amy Stonelake Room)

Polls/ Quick Quizzes: 3-4 per one hour session

Create Q and A forum for Class

Everyone please introduce themselves and post what types of data they will be working with to Q and A forum

Create Class listserv

Session Dates: All classes are held from 3 – 4 PM

Help Sessions:

Vendor Software Dates:

Session One

1. Tuesday, 3/30
2. Thursday, 4/1
3. Tuesday, 4/6
4. Thursday, 4/8

Session Two:

1. Tuesday, 5/4
2. Thursday, 5/6
3. Tuesday, 5/11
4. Thursday, 5/13

Session Three

1. Tuesday, 6/8
2. Thursday, 6/10
3. Tuesday, 6/15
4. Thursday, 6/17

FAQ:

1. What is the GOLD learning environment and how do I use it?
2. Do I have to know any bioinformatics to take these classes? No – these classes are designed for beginners with little to no experience in bioinformatics analyses.
3. Can I upload my RNA-Seq data to GOLD and analyze it using the programs that are already installed there? No - the GOLD system is a teaching environment only - it is not for analyzing real data. As part of the course, you will learn to access the NIH high-performance cluster "Biowulf". You will learn how to upload your data to Biowulf and work on it there.

We have developed a series of three courses for beginning bioinformatics learners with the scenario “I’ve just gotten my RNA-Seq data back from the Sequencing Center, how do I understand and analyze it?”.

These three courses will introduce learners to (1) working within the Unix environment, (2) analyzing the sequence data step-by-step with commonly used bioinformatics programs, and (3) creating visualizations of the data in the R/RStudio environment.

As part of each class, we should point them to some resources to continue learning.

Set up class Q and A forum. (have them post here before class starts to test?)

Notes on Bioinformatics

Notes on Unix

Graphical User Interface (GUI) vs working at the command line

Which is better?

Well….it depends.

Bash is the default shell on unix. A shell is a program for reading commands and running programs. Tasks can be automated with the shell.

Typically the “$” is used as the prompt, but can be other characters.

Type a command, and then press “Enter” or “return”.

Options/flags for commands.