**Bioinformatics Coding Crash Course**

**Dates**: Every Monday, Wednesday, and Friday for 6 weeks, time TBD (1 hr each)

June 1st 2020 – July 8th 2020

**Purpose**: While we are all spending an increasing amount of time working from home during the COVID-19 pandemic, this time provides an opportunity to learn coding and bioinformatics skills for analyzing bioinformatics data from home. This course will introduce you to the basics of programming, particularly for bioinformatics data processing, analysis, and visualization. Completing this course will provide you with the foundation needed to use programming skills in your research and allow you to ask new questions of your data.

Classes will be conceptual and while you may try to follow along with the code in class, class time is not for debugging individual cases. The bulk of the hands-on time will come from completing the out-of-class assignments and can be discussed in more detail in during the Friday review sessions. When you are reviewing the lectures and completing the homework tasks, you should approach troubleshooting in the following manner:

1. Perform an internet search to find a question on Stack Overflow or BioStars that is similar to your problem and try to solve the problem by adapting the provided solutions to the questions.
2. Post the question to the Slack page. The purpose of the Slack page is to serve as a platform for group sharing and troubleshooting of problems. Questions should only pertain to one specific question (“How to I declare a variable in Python?”) and not be a general blanket question (“How do I write the script from yesterday’s homework?”). Provide your classmates with the challenge, the code you tried, and any related errors from your attempt.
3. Each Friday, we will hold an informal office hours session where class members can bring questions to discuss, review, and troubleshoot.

Lecture recordings will be available @ [YouTube](https://www.youtube.com/playlist?list=PLF1C3_ICktkLRFzEKmf-WzXzyjZZmaN94) the day after the lecture. Slides will be available @ [GitHub](https://github.com/j-berg/bioinformatics-bootcamp).

**Outcomes**: By completing this course, you will become familiar and comfortable with basic coding skills so that you have the foundation needed to continue learning and using programming skills in your research. You will become familiar with bioinformatics tools, how to download them, use them, and access their documentation for custom use. You will become aware of the types of questions you can address using coding and bioinformatics.

**Prerequisites**:

* Fill out a Utah CHPC [account application](https://www.chpc.utah.edu/userservices/accounts.php)
* Sign-up for the Coding Bootcamp Slack Channel: *bioinfo-bootcamp.slack.com*
* Download [R](https://ftp.osuosl.org/pub/cran/) and [RStudio](https://rstudio.com/products/rstudio/download/#download)
* Complete the [Introduction to Command Line](http://linuxcommand.org/index.php) lessons
* Windows Users: Download [PuTTY](https://www.chiark.greenend.org.uk/~sgtatham/putty/latest.html) to access CHPC resources

**Schedule**:

Week 1: Command Line and Navigating Files

Mon: Navigating folders and files in the command line

* Navigating the file system
* Copying, moving, and renaming files
* Transferring files
* Truncating, sampling, and outputting files

Wed: Bioinformatics file types and their manipulation

* Downloading and installing Conda and introduction to package management
* GTF file format
* FASTA/FASTQ
* BAM/SAM and samtools
* Data tables
* Downloading external files
* Some data table arithmetic
* Other*.*

Week 2: Command Line and Processing Sequence Data

Mon: Processing an RNA-seq file

* SRAtoolkit
* Read pre-processing and quality control
* Genome and transcriptome alignment
* Read post-processing
* Read and isoform quantification
* Data compilation and formatting
* Automating processing to a set of files

*HW: Download a series of FASTQ sequencing files from a* [*GEO*](https://www.ncbi.nlm.nih.gov/geo/) *dataset related to your own research project and align and quantify the files. Choose a dataset with biological replicates.*

Wed: Command line arithmetic for RNA-seq alignments

* Review sequencing alignment in command line
* IGV
* Download and use other tools to perform follow-up analyses of the data

*HW: Choose 10 genes from your samples before and determine whether there is differential splicing occurring between samples.*

Week 3: Python and Introduction to Variable Types and Programming

Mon: Data types in programming

* Introduction to your computer and how it speaks
* Integers, floats, strings, arrays/lists, objects/dictionaries, data tables
* Importing functions from internal and external libraries

*HW: Create a dictionary of a mock classroom of students and their test scores. Create an array of students’ names, include names that were not in the dictionary. Access each student in the array, determine if they are in the dictionary programmatically, and print out a sentence that states “[student] got a [score] on their test.”*

Wed: Loops, conditionals, and functions

* If, else if, else and flow control
* For and while loops
* Modularizing a task using a function

*HW: 1) Create a list of strings. For each string, if the number of characters is greater than 5, print out the string as is. If it is less than or equal to 5, print out the string plus “(this is a small word)”. 2) Create a list of numbers. If the number is less than 100, divide by 2 and append to a new list. If the number falls between 100 and 200, divide by 3 and append to the new list. If the number is greater than 200, divide by 4 and append to the new list. Important: For each task, create a function that takes as input the array to be processed and outputs the new array of modified values.*

Week 4: Python and Coding a Bioinformatics Tool

Mon: Writing a gene dictionary to rename gene IDs in a data table

* Import a read count table
* Import a GTF file as a data table
* Extract gene IDs and their corresponding gene names
* Remap gene IDs from read count table to their appropriate gene names
* Use user inputs to allow for input of data table and GTF to output the remapped data table

*HW: Take your RNA-seq data table that was mapped to gene IDs from before and remap to gene names*

Wed: Algorithm design in Python and optimizing performance

* Designing more complex algorithms and pseudo-code
* Optimizing performance and why and when it matters

*HW: Review concepts from Python modules*

Week 5: R and Bioinformatics Software

Mon: Open-source bioinformatics software

* Introduction to R Studio
* How to use variables, loops, etc. in R
* Using DESeq2 to perform differential expression analysis of your read count data
* Accessing the documentation of tools
* Outputting files

*HW: Perform differential expression analysis of your read count table from before. Identify the 25 most up-regulated and down-regulated genes from your dataset. Do these make sense in the context of the model?*

Wed: Other statistical software and functions in R

* Normalization of data
* P-value corrections
* Linear regressions
* PCA
* T-tests, ANOVAs, etc.

*HW: Determine the correlation coefficients of the read counts from the biological replicates of your dataset. Perform PCA of the datasets and determine if the experimental groups cluster together.*

Week 6: R and Visualization

Mon: Visualization in R

* Boxplots, violin plots, and swarm plots
* Scatter plots
* Heatmaps
* Outputting plots

*HW: Plot the biological replicates as scatter plots and overlay the correlation coefficient lines. Do this for at least two replicates from each experimental type. Take a gene of interest from the differential expression data and create a boxplot overlaid with a swarm plot of the data for that gene. Create a gene-clustered heatmap of the count data to identify gene expression clusters. Are there any clusters that stand out to you?*

Wed: Visualization in R and conclusions

* Histograms and density curves
* Kaplan-Meier survival curves
* KS tests
* Continued education and troubleshooting in programming and bioinformatics

*HW: Review concepts from R modules*

**Continued Learning Resources:**

* <https://github.com/harvardinformatics/learning-bioinformatics-at-home>
* <https://github.com/griffithlab/rnaseq_tutorial_v1>
* <http://compgenomr.github.io/book/>
* <https://github.com/quinlan-lab/sllobs-biostats>
* <https://github.com/quinlan-lab/applied-computational-genomics>
* <http://web.stanford.edu/class/bios221/book/>
* <https://www.chpc.utah.edu/presentations/Summer2020CHPCPresentationSchedule.php>