MCB 5430: Analysis of Eukaryotic Functional Genomic Data

Mon, Wed. 1-3pm
Beach Hall room 202
Fall 2016 Syllabus

Instructor Information: Leighton Core
Email: leighton.core@uconn.edu

Office hours: by appointment

Text references*:

Practical Computing for Biologists. Steven H. D. Haddock & Casey Dunn (2011).

Getting started with R: an Introduction for Biologists. Andrew P. Beckerman & Owen L. Petchey (2012)

R in Action: Data Analysis and Graphics with R. Robert I. Kabacoff (2011).

* All books can be purchased on Amazon.com

Course Description:
This course will cover the creation of workflows for the processing and analysis of data from next generation sequencing experiments. The focus will be on eukaryotic functional genomics datasets such as ChIP-seq, RNA-seq. Students will learn basic programming skills necessary to complete these tasks including: commands for navigating and operating in the terminal environment, basic shell scripting for creating pipelines, parsing and analyzing data. Students will also be introduced to the R programming language for analysis and display of processed data. Use of available 'off the shelf' analysis tools will be covered and incorporated into workflows.

Learning Outcomes:
Upon completing MCB 5429, students will be able to:

- Use the terminal to navigate their computer and perform everyday operations from the command line.
- Use shell scripting to automate processing and analysis of next-generation sequencing data.
- Properly design ChIP-seq and RNA-seq experiments
- Perform QC analysis on NGS data.
- Map ChIP-seq and RNA-seq data to genomes
- Use public genome browsers to display their own data and retrieve publicly available data.
- Process and perform preliminary analyses of ChIP-seq and RNA-seq data
- Perform peak calling on ChIP-seq data and search for underlying DNA motifs

- Perform differential gene expression analysis on RNA-seq data.
- Perform Gene Ontology analysis on sets of genes identified through genomic analysis

Course Format: Each class will have a short lecture on the assigned topic. The intent is for students to spend the majority of class time learning computer skills necessary for analyzing genomic data.

Course Materials: Lecture slides, notes and in-class demonstrations will be posted to HuskyCT blackboard site. If the HuskyCT site goes down, materials will be emailed.

BBC server access: Some programs will be run on the UConn bioinformatics server. To obtain access to the server fill out the request form at the following link:
http://bioinformatics.uconn.edu/contact-us/
Choose 'account for course' from the first drop down menu, and 'MCB5429' from the course drop down menu.

Useful links from UConn Bioinformatics Core:

Understanding the BBC cluster:
http://bioinformatics.uconn.edu/understanding-the-bbc-cluster-and-sge/#our-cluster

Unix basics:
http://bioinformatics.uconn.edu/unix-basics

Other BBC tutorials:
http://bioinformatics.uconn.edu/resources-and-events/tutorials/

Course Schedule*: *** subject to change**

Week 1:

8/29: Overview of Next Generation Sequencing (NGS), Functional Genomics, and course goals.

8/31: Introduction to Linux terminal.
- Navigating the terminal environment.
- Basic command line utilities.

Week 2:

9/5: No Class – Labor Day

9/7: Introduction to Linux (continued).
- Dealing with text files

- Creating / running shell scripts.

(Homework 1 assigned)

Week 3:

9/12: Introduction to Linux (continued).

- More useful Linux commands
- Scripting with loops.

9/14: The UConn BBC server

- Connecting remotely to the UConn BBC server.
- Using the Nano editor.
- Submitting jobs to the server.

(Homework 1 due)

Week 4:

9/19: Installing programs and editing search path.
Downloading data from public sources

9/21: QC and preprocessing of NGS data.

- fastX and fastQC tools for data filtering, read trimming, adapter clipping.
- Creating pipelines to automate processing and mapping of data.

(Homework 2 assigned)

Week 5:

9/26: Mapping NGS data:

- Overview of read alignment methods.
- Bowtie alignment of ChIP-seq data

2/28: Post-processing of mapped data

- Making bed and bedgraph files.
- Organizing and prioritizing pipeline output.

(Homework 2 due)

Week 6:

10/3: Using UCSC genome browser to view genome annotation tracks and your data.

10/5: Common ChIP-seq analyses:

- Calling ChIP-seq peaks with MACS.

(MIDTERM ASSIGNED)

Week 7:

10/10: Common ChIP-seq analyses:

- Determining reads in peaks, peak location relative to genes.

10/12: Common ChIP-seq analyses:

- MEME: Identification of motifs under discreet peaks.
- MAST: mapping motifs back to genomes.
- FIMO: determine occurrences of motif in selected sequence.

Week 8:

10/17: Introduction to R:

- Reading, writing, viewing and manipulating tables.

10/19: R basic plotting:

- Plotting of distributions

(MIDTERM DUE)

Week 9:

10/24: R basic plotting:

- Plotting of groups of data

10/26: R: Making heatmaps

(Homework 3 assigned)

Week 10:

10/31: Writing R functions

11/2: Introduction to RNA-seq

- Alignment considerations
- Alignment of RNA-seq data

(Homework 3 due)

Week 11:

11/7: RNA-seq (continued)

11/9: Differential gene expression analysis

- EdgeR

(Homework 4 assigned)

Week 12:

11/14: Differential gene expression analysis (continued)

- DEseq

11/16: Gene ontology and gene set enrichment analysis
(Homework 5 due)
(FINAL PROJECT ASSIGNED)

Week 13:
11/21: No Class -Thanksgiving Recess
11/23: No Class -Thanksgiving Recess

Week 14:
11/28: Student projects: independent analysis of public data
11/30: Student projects: independent analysis of public data

Week 15:
12/5: Student projects: independent analysis of public data
12/7: Student projects: independent analysis of public data

***Changes in the syllabus:** Every effort will be made to follow the course outline for classroom lectures and assignments. However, given that this is still a new format for this course, some changes in the syllabus may be unavoidable. Students are responsible for being aware of these changes. If you miss a class you should check the blackboard site for lecture and class notes, as well as potential changes to the syllabus.

Homework: Homework assignments will be announced in class and are due the following week. All assignments will be posted on the blackboard site for the course. Homework will be submitted via the HuskyCT site or via email to the instructor. **Assignments should be named with the NetID and assignment number (e.g. xyx15002_HW1).** Assignments are due by 5pm on the scheduled due date. Late assignments will lose 5% of total points per day, including weekends.

Course Grades:

Final Grade: Based on a 200-point scale:
- In class exercises: 20 points (10%)
- 4 homework assignments: 20 points each; 80 points total (40%)
- Midterm Project: 40 points (20%)
- Final Project: 60 points (30%)

Useful and potentially useful links: (more will be distributed during classes)

Terminal/Linux:
http://lifehacker.com/5633909/who-needs-a-mouse-learn-to-use-the-command-line-for-almost-anything
http://ryanstutorials.net/linuxtutorial/

Free Linux for Dummies: http://it-ebooks.info/book/784/

Python:
https://www.python.org/
http://www.codecademy.com/
http://www.learnpython.org/

R:
http://manuals.bioinformatics.ucr.edu/home/ht-seq
http://www.r-bloggers.com/using-apply-sapply-lapply-in-r/
https://cran.r-project.org/doc/contrib/Torfs+Brauer-Short-R-Intro.pdf

Publishing guidelines for data analysis:
http://melissagymrek.com/science/2014/01/09/show-me-the-data.html