ENVIRONMENTAL BIOINFORMATICS

Instructors: Harriet Alexander (halexander@whoi.edu)
Maria Pachiadaki (mpachiadaki@whoi.edu)
Carolyn Tepolt (ctepolt@whoi.edu)
TA: David Geller-McGrath (mcgrath1@mit.edu)

COURSE INFORMATION

An intensive, hands-on introduction to computational skills and a survey of modern computational theory and approaches for the manipulation and analysis of genomic data in non-model systems. This course is designed to synthesize theory (both biological and computational) with hands-on programming to equip students with the ability to understand and carry out hypothesis testing with genomic data.

Logistics:

- Tuesdays & Thursdays, 4:00PM - 5:30PM
- WHOI: Clark 271
- MIT: 54-823 (via videolink)

Office hours:

- Harriet: Thursdays 10:00 AM – 12:00 PM (Watson 109)
- Maria: Mondays 2:00 PM-4:00 PM (Redfield 322)
- Carolyn: Wednesdays 1:00 PM – 3:00 PM (Redfield 230)
- David: TBD
- MIT: you are not forgotten! While we wish we could come to MIT this semester, due to COVID limitations it is not likely to happen. However, we will have a TA (David) who will attend most lectures at MIT and will hold office hours at MIT. Additionally, we will also be available via Zoom during our office hours. Slack us if you would like to meet over zoom!

***Class resources:* we'll go over how to use these in class**

- GitHub Classroom: This will be the primary means of assigning and turning in homework. We hope that this resource will increase your comfort with Git while streamlining homework and project submission.
- Poseidon: All computation for this course will be done on WHOI's HPC. While we appreciate that you may be able to run some of these analyses on your own computers, the time and effort required to troubleshoot everyone's individual configuration is prohibitive. Feel free to set up / test anything you like on your own machine, but please run in-class exercises, homework, and project computation on Poseidon. We have designated class space for this, and non-JP students will be given temporary access to the cluster as guest students.
- Slack: We have set up a Slack channel as a general communication hub for the class. We hope that this platform will serve as a resource for everyone in the class to converse, troubleshoot, and help each other.

Assessment: This course will have six main homework assignments that align with the six course sections, plus an initial 0th homework to help build a foundation of working with a terminal. All homework will be assigned and submitted through GitHub Classroom. The 0th homework will count as 5% of the final grade. The five highest-scored ***completed*** main homework assignments will each count as 10% of the final grade. In addition, there will be one final project worth 35% of the final grade. The remaining 10% will be determined by active and prepared participation in class.

Final Project: Working in groups of two or three, students will select a paper of interest containing some variety of -omic analysis and will attempt to replicate the data analyses using the study's publicly archived data. Detailed instructions on the final project will be given out in the first class. This project should be carried out throughout much of the course, and there will be a number of milestone dates for project progress throughout the semester. The final product will be a Jupyter notebook annotating each step of the process and a comparison of project results to the original published results. All groups will give a 15-minute presentation on their reanalysis in the final week of classes.

Accommodations: If you need disability-related accommodations, please contact us as early in the semester as possible. If you will miss more than one class in a row, please let us know. We understand that life happens, and we're happy to work with you, but we can't help if we don't know.

COURSE STRUCTURE

Section 1: Computational science & introduction to programming
09 September: Class introduction, computer setup, command line introduction
14 September: HPC setup & login, continue with UNIX shell
16 September: Introduction to GitHub, how to organize biology projects
21 September: Introduction to Python, introduction to Jupyter notebooks
23 September: Python continued
28 September: History of sequencing, Python continued

Section 2: Introduction to biological algorithms and sequence data
30 September: Sequence alignment and comparing sequences
05 October: Sequence data and quality control
07 October: kmers, de bruijn graphs, and assembly
12 October: Assembly in practice
14 October: Functional and taxonomic annotation

19 October: PRESENTATIONS

Section 3: Environmental metagenomics

21 October: introduction to environmental metagenomics
26 October: Qiime2, targeted metagenomics
28 October: shotgun sequencing, metagenome assembly & binning

Section 4: Differential Expression Analysis
02 November: gene expression analysis, theory and experimental design
04 November: R and ggplot intro with transcriptomic data

Section 5: Intraspecific diversity
09 November: introduction to population genomics; identifying & working with SNPs
11 November: No class (Veterans Day)
16 November: selection, drift, & intraspecific adaptation
18 November: selection analyses in non-model species

23 November: Open lab (help with final project issues)

25 November: No class (Thanksgiving)

Section 6: Putting it all together: automation and best practices
30 November: pipelines, workflows, & reproducibility
03 December: Snakemake

07 December: Final project presentations
09 December: Final project presentations

DUE DATES FOR HOMEWORK AND PROJECTS

Completed homework must be on GitHub by **11:59 PM on the due date**. Typically, homework will be due the day before a class, rather than on a class day itself.

LATE POLICY: Talk to us BEFORE homework is late and we will work something out. Otherwise, late homework loses 10% per day.

For the final project, there are a series of milestone dates periodically throughout the class, as you choose a project and work through it. More details will be given in the final project handout, which you'll get on the first day of class. For planning purposes, key project dates are also included in this list. Note that project milestones must be **completed** by these dates. Please don't leave them until the last minute, especially the check-in meetings since these may take a little while to schedule.

13 September: HW#0A – UNIX shell part 1, assigned 9 September
20 September: HW#0B – UNIX shell part 2, assigned 14 September
21 September: Project: identify teams & papers, post these to the class Slack

27 September: HW#1A - python part 1, assigned 21 September
30 September: Project: 1st check-in meeting with faculty mentor

04 October: HW#1B - python part 2, assigned 28 September
05 October: Project: create Gantt chart outlining project tasks & who will do them
07 October: Project: in HPC, create project file structure and populate with source data
11 October: HW#2A – sequence quality, assigned 5 October
19 October: Project: in class, present project introduction & plan of work
20 October: HW#2B – genome assembly and assessment, assigned 12 October

01 November: HW#3 – metagenomics, assigned 21 October
08 November: HW#4A - gene expression part 1, assigned 2 November
09 November: Project: 2nd check-in meeting with faculty mentor
15 November: HW#4B - gene expression part 2, assigned 9 November
22 November: HW#5 – intraspecific variation, assigned 16 November

06 December: Project: ALL materials due by midnight (inc. final project & presentation)
07 December: Project: final project presentations, Part 1
08 December: HW#6 – pipelines & integration, assigned 30 November
09 December: Project: final project presentations, Part 2