ENVIRONMENTAL BIOINFORMATICS

Instructors: H. Alexander, M. Pachiadaki, C. Tepolt

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, 2:30PM – 4:00PM
- WHOI: Clark 271
- MIT: 54-823 (via videolink)

Office hours:

- Harriet: Thursdays 9:00 AM – 11:00 AM (Watson 109)
- Maria: Wednesdays 9:00 AM-11:00 AM (Redfield 322)
- Carolyn: Wednesdays 1:00 PM – 3:00 PM (Redfield 230)
- MIT: you are not forgotten! We will periodically come up to Cambridge on class days to assist with in-class exercises and to hold office hours. While we don't have a set schedule for this, we will email all MIT-based students in advance to let you know when an instructor will be in Cambridge. Our goal is to have one instructor physically at MIT for one class per week, and we will take turns so we all get to meet you. We are also available via Skype during our office hours by prior arrangement.

***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 homework assignments that align with the six course sections. All homework will be assigned and submitted through GitHub Classroom. The five highest-scored *completed* 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 15% will be determined by active and prepared participation in class.

Final Project: Working in pairs, 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.

COURSE STRUCTURE

Section 1: Computational science & introduction to programming
5 September: class introduction, computer setup, command line introduction
10 September: HPC setup & login, introduction to GitHub & UNIX
12 September: UNIX continued, introduction to Python
17 September: Python continued, introduction to Jupyter notebooks
19 September: Pandas data frames, plotting in Python

Section 2: Introduction to biological algorithms
24 September: sequence alignment
26 September: BLAST, introduction to phylogenetics

Section 3: Genome and transcriptome data
1 October: introduction to genome assembly
3 October: working with HTC data (fastq files)
8 October: transcriptome assembly & applications
10 October: gene expression analysis
15 October: NO CLASS (Indigenous Peoples' Day)

17 October: Class presentations: brief introduction to final projects

Section 4: Environmental metagenomics
22 October: introduction to environmental metagenomics
24 October: Qiime2, targeted metagenomics
29 October: shotgun sequencing, single-cell genomics
31 October: metagenome assembly & binning

5 November: Open lab (bring your computational questions & answers)

Section 5: Intraspecific diversity
7 November: introduction to population genomics
12 November: identifying & working with single-nucleotide polymorphisms (SNPs)
14 November: selection, drift, & intraspecific adaptation
19 November: selection analyses in non-model species

21 November: Open lab (bring your computational questions & answers)

Section 6: Putting it all together: automation and best practices
26 November: pipelines, workflows, & reproducibility
28 November: NO CLASS (Thanksgiving)
3 December: Snakemake

5 December: Final project presentations
10 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.

16 September: HW#1A – UNIX, assigned 10 September
17 September: Project: identify teams & papers, post these to the class Slack
20 September: Project: 1st check-in meeting with faculty mentor
23 September: HW#1B – python, assigned 17 September
26 September: Project: create Gantt chart outlining project tasks & who will do them

3 October: Project: in HPC, create project file structure and populate with source data
7 October: HW#2 – more python & phylogenetics, assigned 24 & 26 September
17 October: Project: in class, present project introduction & plan of work
23 October: HW#3 – differential gene expression, assigned 8 October

4 November: HW#4 – metagenomics, assigned 24 October
5 November: Project: 2nd check-in meeting with faculty mentor
20 November: HW#5 – intraspecific variation, assigned 7 November

4 December: Project: ALL materials due by midnight (inc. final project & presentation)
5 December: Project: final project presentations, Part 1
9 December: HW#6 – pipelines & integration, assigned 26 November
10 December: Project: final project presentations, Part 2