**BST272: Computing Environments for Biology**

**Winter 2020, Jan. 7, 9, 14, 16**

**Time: TR 1:00-4:00**

**Location: Kresge 201 / LL6 (Jan. 14 only)**

**Instructor Information**

**Faculty**

Curtis Huttenhower, Professor, Biostatistics and Immunology and Infectious Diseases

SPH1 413, chuttenh@hsph.harvard.edu, 617-432-4912

Office Hours: SPH1 413, 9th 11:30-12:30, 14th 4:00-5:00

**Teaching** **Assistants**

Sagun Maharjan, Software Developer, Biostatistics

SPH1 412, smaharjan@hsph.harvard.edu, 617-432-5065

Office hours: SPH1-2 atrium, 13th and 21st 1:00-2:00

Sharifa Sahai, Ph.D. Student, Systems Biology

sharifasahai@g.harvard.edu

**Credits**

1.25 credits

**Course Description**

This course provides a high-level introduction to general computing environments appropriate for biological data analysis, as preparation for more advanced computational biology and bioinformatics courses. It is intended for biologists, clinician-researchers, other bench or translational scientists, or mathematicians with little to no computational or applied quantitative experience. It provides a compressed, highly interactive, hands-on introduction to basic command line, Python, and R environments for biological data analysis and visualization. It covers basic quantitative methods that can be carried out for 'omics data analysis in these environments and ensures that students have access to local and online (i.e. grid, cloud) resources for using these tools in the future. Finally, it thoroughly introduces freely available documentation and strategies for self-learning when using computational methods for biology research.

* **Pre-Requisites**

None

**Learning Objectives**

Upon successful completion of this course, you should be able to:

* Recognize and carry out the essentials of navigation of command line environments for biological data handling.
* Basic familiarity with the capabilities of general computing environments (Python, R) for quantitative biology and high-throughput molecular data analysis and visualization.
* Comfort with terminology and concepts for basic quantitative data structures: vectors, lists, matrices, sets, permutations, probabilities and probability distributions, conditional probabilities, hypothesis tests, modeling, and prediction.
* Ability to discover and understand documentation and resources available for further self-training as needed.

**Course Readings**: (Connect with the [Curriculum Center](https://www.hsph.harvard.edu/office-of-education/curriculum-center/) for best practices for course materials that includes securing copyright permissions and information on course materials fee categories)

* [required and recommended materials]

**Course Structure**

This is a small, highly hands-on course in which discussion is encouraged during lecture material and individual and small group participation is necessary during laboratory / tutorial sessions. Each session is organized into roughly 1/3 introduction of material via lecture and discussion, and 2/3 application of the material using interactive computing problem sets (individually and in small groups). Particularly due to the highly compact nature of the course schedule, students are expected to attend all sessions, be active participants in discussions and tutorials, and ask questions of the instructional staff and of each other to build comfort and understanding of quantitative computing environments for biology.

All course materials will be provided on the Canvas site, which will also be used to communicate scheduling, announcements, and to submit assignments. Students should bring laptops to class during all sessions if possible, and required software will be set up interactively during the relevant lab sessions.

**Grading, Progress and Assessment**

The final grade for this course will be based on:

* In-class laboratory submissions (4x, 10% each, 40% total)
* Homework problem sets (2x, 25% each, 50% total)
* Participation, based on attendance and engagement in discussion and tutorial activities (10%)

In-class laboratory activities are generally carried out collaboratively in groups, with worksheets submitted individually at the end of each session. Homework problem sets are to be completed individually and submitted on time. The maximum score for late work will fall exponentially: 90% if one day late, 75% if two, 40% if three, and all credit lost if four or more days late. Extensions may be granted with reason if requested at least 24 hours in advance of the assignment deadline. Final letter grades will be curved based on the percentiles of total scores received by students in the class.

**In-class laboratory submissions (4x, 10% each, 40% total)**

Laboratory tutorial activities include Jupyter notebooks, knitr documents, and brief written question-and-answer or descriptions of online activities. The activities themselves are carried out during class, typically in small groups, with the written results submitted individually at the end of each session.

**Homework problem sets (2x, 25% each, 50% total)**

Problem sets will be assigned during the course to be completed individually outside of class. These are a combination of written questions with hands-on analysis activities. Late submissions without prior notice are penalized as described above.

**Participation (10%)**

In a small, intensively, highly collaborative course, participation and discussion is especially important. Class participation will be included in the final grade based on a combination of active participation during lectures (questions posed by students, and answers provided by students to those posed during lecture), attendance, and engagement with small groups and instructional staff during laboratory tutorials. Students should arrive to class prepared to ask and answer questions, share their viewpoints in constructive and respectful ways, and otherwise actively engage with other students and the course instructor. Notification of class absences should be provided by email to the instructional team at least 24 hours in advance. Otherwise, class participation will be graded as follows:

* 10%: Always contributes to discussions by asking thoughtful questions; relates diverse topics from lecture sessions; is active and inquisitive during laboratory tutorials; discusses challenges with small group and instructors; attends all sessions.
* 5-9%: Sometimes contributes to discussions as above; interaction during only a subset of topic areas; completes laboratory activities but is not engaged with collaborators; more than one lecture or lab absences without notice.
* 0-4%: Rarely contributes to discussions or activities as above; more than one lecture absence without notice.

**Harvard Chan Policies and Expectations**

**Inclusivity Statement**

Diversity and inclusiveness are fundamental to public health education and practice. Students are encouraged to have an open mind and respect differences of all kinds. I share responsibility with you for creating a learning climate that is hospitable to all perspectives and cultures; please contact me if you have any concerns or suggestions.

**Bias Related Incident Reporting**

The Harvard Chan School believes all members of our community should be able to study and work in an environment where they feel safe and respected. As a mechanism to promote an inclusive community, we have created an anonymous bias-related incident reporting system. If you have experienced bias, please submit a report [here](https://reportinghotline.harvard.edu/) so that the administration can track and address concerns as they arise and to better support members of the Harvard Chan community.

**Title IX**

The following policy applies to all Harvard University students, faculty, staff, appointees, or third parties: [Harvard University Sexual and Gender-Based Harassment Policy.](http://hwpi.harvard.edu/files/title-ix/files/harvard_sexual_harassment_policy.pdf) Procedures [For Complaints Against a Faculty Member](http://hwpi.harvard.edu/title-ix/complaints-against-faculty-member-hsph)

Procedures[For Complaints Against Non-Faculty Academic Appointees](http://hwpi.harvard.edu/title-ix/complaints-against-non-faculty-academic-appointees-hsph)

**Academic Integrity**

Each student in this course is expected to abide by the Harvard University and the Harvard T.H. Chan School of Public Health School’s standards of Academic Integrity. All work submitted to meet course requirements is expected to be a student’s own work. In the preparation of work submitted to meet course requirements, students should always take great care to distinguish their own ideas and knowledge from information derived from sources.

Students must assume that collaboration in the completion of assignments is prohibited unless explicitly specified. Students must acknowledge any collaboration and its extent in all submitted work. This requirement applies to collaboration on editing as well as collaboration on substance.

Should academic misconduct occur, the student(s) may be subject to disciplinary action as outlined in the Student Handbook. See the [Student Handbook](https://www.hsph.harvard.edu/student-handbook/academic-support/academic-integrity/) for additional policies related to academic integrity and disciplinary actions.

**Accommodations for Students with Disabilities**

Harvard University provides academic accommodations to students with disabilities. Any requests for academic accommodations should ideally be made before the first week of the semester, except for unusual circumstances, so arrangements can be made. Students must register with the Local Disability Coordinator in the Office for Student Affairs to verify their eligibility for appropriate accommodations. Contact Colleen Cronin ccronin@hsph.harvard.edu in all cases, including temporary disabilities.

**Religious Holidays, Absence Due to**

According to Chapter 151c, Section 2B, of the General Laws of Massachusetts, any student in an educational or vocational training institution, other than a religious or denominational training institution, who is unable, because of his or her religious beliefs, to attend classes or to participate in any examination, study, or work requirement on a particular day shall be excused from any such examination or requirement which he or she may have missed because of such absence on any particular day, provided that such makeup examination or work shall not create an unreasonable burden upon the School. See the [student handbook](https://www.hsph.harvard.edu/student-handbook/student-life-policies/religious-holidays/) for more information.

**Grade of Absence from Examination**
A student who cannot attend a regularly scheduled examination must request permission for an alternate examination from the instructor in advance of the examination. See the [student handbook](https://www.hsph.harvard.edu/student-handbook/academic-support/hsph-grading-system/#absence) for more information.

**Final Examination Policy**

No student should be required to take more than two examinations during any one day of finals week. Students who have more than two examinations scheduled during a particular day during the final examination period may take their class schedules to the director for student affairs for assistance in arranging for an alternate time for all exams in excess of two. Please refer to the [student handbook](https://www.hsph.harvard.edu/student-handbook/academic-support/final-examination-policy/) for the policy.

**Course Evaluations**

Constructive feedback from students is a valuable resource for improving teaching. The feedback should be specific, focused and respectful. It should also address aspects of the course and teaching that are positive as well as those which need improvement.

Completion of the evaluation is a requirement for each course. Your grade will not be available until you submit the evaluation. In addition, registration for future terms will be blocked until you have completed evaluations for courses in prior terms.

**Course Schedule & Assessment of Student Learning**

|  |  |  |  |
| --- | --- | --- | --- |
| **Session topics** | **Objectives** | **Readings** | **Activities/****Assignments** |
| **Week of: Jan. 7-9** |
| **Module 1: Research computing for biological data analysis**\* Command line computing environments\* General command line data manipulation principles: files, redirects, pipes, and more\* Built-in command line data manipulation tools\* Accessing grid and cloud research computing environments**Module 2: Introductory biological data analysis in Python**\* Basics of Python as a general data manipulation environment\* Biological data types in Python\* Text, numerical, and structured data types\* Data input and output\* Data visualization | **Upon Successful completion of this week, you should be able to:** 1. Understand the capabilities of and navigate local command line environments2. Access basic grid and cloud environments for biological computing3. Carry out basic biological data manipulation in Python | **Required** Review syllabus in advance of class**Recommended** A Quick Guide for Developing Effective Bioinformatics Programming Skills (<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000589>)A Quick Guide to Organizing Computational Biology Projects (<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000424>)An Introduction to Programming for Bioscientists: A Python-Based Primer (<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004867>) | **Activity #1:** system-specific command line setup (Windows/Mac/Linux), local command line playground, Linux Journey interactive command line tutorial, Cannon-Odyssey/AWS/GCP**Homework #1:** DNA sequence and transcriptomic data analysis using built-in command line tools and Python (on the cloud if needed)**Activity #2:** local Python/Jupyter setup, Data Camp Python basics, FASTA/FASTQ genomic data handling, RNA-seq transcriptional data handling, plotting and visualization |

|  |
| --- |
| **Week of: Jan. 14-16** |
| **Module 3: Quantitative data analysis in Python and R**\* Python libraries for scientific data analysis: NumPy, SciPy, Pandas, Matplotlib, Seaborn\* Basics of R and RStudio\* Statistical methods and visualization in R\* Capabilities, pros, and cons of different biological data manipulations in different environments**Module 4: Quantitative data structures and finding additional resources**\* Data structures and how to use them: vectors, lists, matrices, sets, dictionaries, etc.\* Statistical methods and terminology: probabilities, probability distributions, and hypothesis tests\* Generative model fitting, discriminative predictors\* Online resources for quantitative biology and how to use them | **Upon Successful completion of this week, you should be able to:** 1. Understand quantitative data handling and visualization capabilities of Python and R2. Recognize and use terminology for general quantitative data structures, models, and methods3. Know how to find more help and documentation for computational biology tools | **Recommended** A Quick Guide to Teaching R Programming to Computational Biology Students (<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000482>)Ten quick tips for getting the most scientific value out of numerical data (<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006141>)Ten simple rules for biologists learning to program (<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005871>)Ten Simple Rules for Effective Computational Research (<https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003506>) | **Activity #3:** NumPy/SciPy/Pandas/Seaborn in Jupyter, RStudio setup, Data Camp R basics**Homework #2:** 'omics data statistics and visualization using Jupyter and knitr**Activity #4:** RStudio knitr for basic model fitting, prediction, hypothesis testing, and permutation tests; documentation scavenger hunts for Python, Python libraries, R/CRAN, Bioconductor, Stack Overflow, and general Googling |

* Please note, session topics and activities may be subject to change during the course