HARVARD T.H. CHAN | SCHOOL OF PUBLIC HEALTH

BST 281: Genomic Data Manipulation (Spring 2024)

Lecture: Mon & Weds, 3:45 – 5:15pm, FXB G12
Lab (optional): Weds, 1:00 – 1:50pm, FXB G11
Version: 2024-01-29

Instructional Staff

Instructors

Dr. Eric A. Franzosa
Senior Research Scientist, Dept. of Biostatistics
Office: HSPH, Building 1, Room 413 or Suite 412
E-mail: franzosa@hsph.harvard.edu (*include "BST 281" in the subject when emailing staff*)

Dr. Kelsey N. Thompson
Research Associate, Dept. of Biostatistics
Office: HSPH, Building 1, Room 413 or Suite 412
E-mail: kthompson@hsph.harvard.edu

Instructors' office hour: Fri, 10:00am, via Canvas Zoom
Additional in-person office hours are available by appointment (email to setup)

Teaching Fellows

Corriene (Corri) Sept (corriene_sept@g.harvard.edu)
Office hour: Thurs, 10:00am, via Canvas Zoom

Minye Zhou (minyezhou@hsph.harvard.edu)
Office hour: Thurs, 1:00pm, via Canvas Zoom

Credits

5.0 credits

Course Description

Genomic Data Manipulation (BST 281) provides a practical introduction to the tools and techniques used to obtain, analyze, and interpret a wide variety of modern genome-scale data types. The course requires and assumes prior familiarity with command-line environments and Python and/or R scripting at an introductory level. Students will apply those skills to achieve a better understanding of principles underlying the algorithms and software methods of modern genome-scale data manipulation, and will combine those approaches to aid in the interpretation of data from their individual research and/or public repositories. We will introduce the types of experimental results commonly encountered in genomic data analysis (high-throughput sequencing, gene expression, protein-protein interaction networks, etc.) and freely available online sources for these data.

BST 281 consists of two required 90-minute meetings per week, divided into lecture/discussion sections and hands-on activities, and an optional lab/recitation section (to be scheduled during Lecture 01). Students will complete biweekly homework assignments to reinforce concepts from lecture and gain further hands-on experience with genomic data manipulation. Working in groups, students will present a modern research paper in the field of genomic data science in journal-club style as the course's equivalent of a midterm exam. As a final project, students will (again working in groups) design and execute a series of novel transformations on a biological dataset and collectively document their methods, findings, and experience.

Prerequisites

- BST272, BST273, or equivalent computing/programming experience authorized by the primary instructor (typically in Python or R).
- Prior experience with the tenets of molecular and cell biology (e.g. the central dogma, principles of protein structure, and cellular organization) is important for having the necessary vocabulary to communicate in the course.
- Prior experience with undergraduate-level statistics is useful but not strictly required.

Learning Objectives

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

- Process and manipulate an array of 'omics data types (sequence, expression, structure, etc.) and understand how they may be generated or obtained from public repositories.
- Apply basic computational and statistical tools to analyze these data types.
- Demonstrate the use of general computational tools (including Python/R scripting) for manipulating genomic/experimental data.
- Critically analyze broad areas of current research in quantitative biology.
- Pursue more advanced/specialized coursework in computational biology.

For Ph.D. in BPH students: This course covers the following Ph.D. in BPH competency: "*Apply computational and statistical tools and/or techniques to obtain, analyze, and interpret a variety of modern genome-scale data types.*"

Course Readings

Course readings will be provided in 1-2 page review documents associated with each lecture. In recent years we've been transitioning away from required textbook readings toward online articles and literature reviews. We will continue to provide historical suggested readings from the following two textbooks, which you may look into if interested/helpful:

- Introduction to Genomics, Lesk (3rd edition)
- Bioinformatics and Functional Genomics, Pevsner (3rd edition)

Similarly, a small number of suggested readings on topics related to statistical methods will be provided from the following textbook:

- Principles of Biostatistics, Pagano and Gauvreau (2nd edition)

Course Structure

Website (Canvas): https://canvas.harvard.edu/courses/133856. Lecture materials, recordings, assignments, and announcements will be organized via Canvas (typically using the "modules" tab). The Canvas Discussion Board is available for non-private interaction with the instructional staff and Zoom meetings (where applicable) can be launched from the Zoom tab.

Lecture Meetings: There will be two required 90-minute in-person lecture meetings per week. Lectures are focused on structured presentation of new course material by the instructors interspersed with discussion questions. Roughly every-other lecture will conclude with a hands-on demo of the day/week's topic, many of which will continue in the course's optional Lab (see below). Lectures will be recorded for optional review and to accommodate unexpected absences or pre-excused academic travel. Lecture recordings are NOT an alternative to in-person attendance (see "Participation" under the assessments section below).

Lab Session: The TFs will host a Lab session once per week for up to one hour. Participation in the Lab is optional: it will not affect your Participation grade for the course and is not counted toward the course's credit hours. Lab provides an environment to 1) review expectations for the current homework assignment, 2) review the solution to the previous homework assignment, 3) review material covered in recent lectures, and 4) finish hands-on activities from lecture. Lab will *not* introduce new scientific content, but *may* be used for supporting skills development (e.g. instruction on how to deliver an effective journal club presentation). Extra Lab time can be used for general Q&A with the TFs (similar to office hours). The Lab day and time will be determined in the first week of the course to maximize students' ability to participate. Lab is an in-person event and will not be recorded.

Office Hours: There will be one joint instructor office hour and two TF office hours per week. Office hours are intended to provide feedback on the current homework assignment (or final project in the final weeks of the course) and clarify material from the lectures. The dates, times, and formats (in-person vs. Zoom) of the office hours will be set during the first week to maximize students' ability to participate. Office hours will not be recorded.

Grading, Progress, and Assessment

The final grade for this course will be based on:

- Five, biweekly problem sets (10% each, 50% total)
- Midterm journal club presentation (15%)
- Final project (25%)
- Class participation (10%)

Homework assignments (50%)

Students will complete biweekly homework assignments to reinforce their understanding of lecture material and practice genomic data manipulation methods. Assignments will be posted by class time on Mondays and will be due via Canvas hand-in by 11:59pm on the Friday of the *following* week. Five assignments will be completed during the first 12 weeks of the semester; students will transition to working on their Final Projects for the final 4 weeks. Assignments will be provided in literate programming frameworks (chiefly Jupyter Notebooks) and will consist of a combination of 1) paragraph-style writing about analysis methods and biological interpretation, 2) reporting on output data and processes conducted external to the notebook, 3) data visualization, and 4) code-based data analysis conducted in the notebook itself. Students will submit both their executed notebook file and a print-to-PDF copy of the notebook for grading.

Midterm journal club presentation (15%)

Working in groups of ~4, students will select, digest, and present to the class a recent, high-quality paper with a considerable genomic data/methods component. Presentations will last 25 minutes including 5-7 minutes for Q & A. Presentations will replace instructor-led lectures during 2 weeks of the course spaced to divide the semester roughly into thirds (6 presentations per week). Students will be given the opportunity to form groups and select presentation days early in the course. Groups must submit their chosen paper to the instructors for approval prior to their presentation date. Group members will receive a shared grade based on the clarity of their presentation/critique of a paper's background, methods, and results; handling of questions from the audience; balance of speaking time; and slide quality. Bonus points will be awarded for presenting papers that directly relate to the preceding unit in the course. Additional details of the presentation format will be posted to Canvas and reviewed in Lab.

Final Project (25%)

Working in groups of ~4*, students will propose, refine, and execute a research project in genomic data analysis over the final 4 weeks of the course. Projects will be framed as a series of transformations/analyses of genome-scale data (~1/group member), starting from raw or lightly preprocessed (typically public) input data and ending with high-quality statistical inferences, data visualizations, and written discussion. Prior to beginning non-trivial work on their project, groups will submit short proposals detailing the project's motivation, input data, and proposed methods to the instructors. Proposals typically undergo 1-2 rounds of rapid iteration to arrive at a body of work that is both sufficient and feasible for a group of a given size and the 4 weeks of Final Project work. Students are allowed to propose new work using data from the individual research of one or more group members; proposing ongoing or completed work from individual research or other courses is prohibited.

Groups will submit a single write-up describing the Final Project work and its findings alongside individualized data + code packets. The write-up is expected to contain (short) introduction and discussion sections written jointly by the group members and per-member descriptions and visualizations of individual contributions. Each student's Final Project grade will consist of a majority individual and minority shared component (with the latter based on the jointly written sections and overall project cohesion). Grading will focus on the quality of the student/group's analysis methods and their descriptions of data, methods, and results; there will be no penalties if a reasonable analysis failed to produce "exciting" results. Additional details of the Final Project write-up will be posted to Canvas and reviewed in Lab.

Groups will additionally present their Final Projects in "lightning talk" format (i.e. ~10 minutes, including one audience question) in the final week of the course.

(*A student's Final Project group can be different from their Journal Club group, though most students opt to work with the same group for both assignments*.)

Participation (10%)

Participation in BST 281 is quantified based on in-class surveys assigned and discussed during lecture meetings. The further goal of these surveys is to recruit real-time feedback about the course and its contents, including comments on pacing/difficulty, feedback on student presentations, and comprehension questions covering material from the preceding lectures. Surveys will be graded based on the completeness and thoughtfulness of responses rather than objective correctness (where applicable). These surveys are NOT intended to function as "pop quizzes" in the traditional sense. Surveys will be conducted via the Canvas "Quiz" system or on paper (as such, please make sure to bring a pen or pencil to class). As with all assignments, survey responses should represent a student's individual thinking and should not be copied from other sources (see the "Original Work Policy" below). Students are permitted to miss 0-2 surveys without penalty, after which the Participation grade drops off exponentially. Students should alert the instructors to planned absences (e.g. academic travel) as soon as possible and to unplanned absences (e.g. due to illness or emergency) when able.

Additional Policies

Original Work Policy

Students are encouraged to discuss assignments (including non-group assignments) liberally with their peers in BST 281. However, final analyses (including any requested computer code) are expected to be completed individually. Likewise, written responses are expected to represent a student's individual thinking, and should not be copied from or co-developed with another human intelligence. It is generally dangerous to directly review another student's assignment (completed or in-progress) due to the potential for over-converged solutions.

The policy above extends straightforwardly to interactions with generative artificial intelligence (GAI) systems, such as Alphabet's *Bard* or OpenAI's *ChatGPT*. These systems are excellent tools for seeking clarification or additional examples of complex concepts, including computer code. However, they can also be abused to compose complete solutions (text or code) for you. Hence, from the standpoint of Academic Integrity, we consider submitting a solution shaped by GAI to be equivalent to submitting a solution shaped by another human intelligence.

Whether seeking human or GAI assistance, posting assignment questions online is prohibited EXCEPT when posting to this course's Canvas Discussion Board.

Students should consult with the instructors when pursuing outside help if there is any question as to its appropriateness. We consider seeking/accepting inappropriate help to be a violation of the School's Academic Integrity policy, and it typically results in substantial loss of credit for the assignment. Please be aware that text and code submissions are automatically compared within and across class years to identify statistically improbable similarities.

Getting Help

As stressed above, the one exception to the "no online posts" rule is the Canvas Discussion Board (CDB), where students are permitted to post *unsolved* assignment questions in order to get clarification from the instructional team. Solutions, including partial solutions (e.g. non-functional code), should not be posted to the CDB. The CDB is also a resource for finding peers for study groups and for Journal Club and Final Project work. Questions that are not appropriate for the

CDB can be brought to Lab, TF office hours, or the instructors' office hour. If emailing the instructional team with questions, please put "BST 281" in the subject line. "Regular" email is preferred over Canvas messaging. Please do not email the same question to each staff member individually; if you have an urgent question, it is preferable to CC the entire instructional team. Using the CDB is preferred over email where comfortable/appropriate as it allows all students to benefit from the clarifications provided by the instructional staff.

Late Work Policy

Aim to hand in assignments on time (typically Friday 11:59pm of the week *after* the assignment was first posted). Students can submit work up to two days late and have it scored out of a maximum of 90% of the potential points as a lateness penalty (this second deadline will typically be Sunday 11:59pm after the initial Friday due date). Work cannot be submitted for credit beyond two days of lateness. Non-emergency extensions may be granted if requested with justification >24 hours in advance of the assignment deadline. The Final Project cannot be turned in late unless special arrangements were made with the instructors in advance (e.g. taking a temporary incomplete grade, INC, for the course). Similarly, there is no late completion option for missed in-class surveys (see "Participation").

Final Letter Grades

The following table indicates the default lower limits for each letter grade:

0	70	73	77	80	83	87	90	93
F	C-	C	C+	B-	B	B+	A-	A

Final weighted percentage scores are compared to this table to determine final letter grades without any additional rounding (for example, a score of 89.9 maps to a B+). Depending on class performance, grades may be curved upward, or these limits lowered, to shift the overall grade distribution to students' benefit. Note that, while Canvas should accurately summarize the grades you've received on assignments and the surveys you've completed, its final grade calculations are not official.

Please see the next page for the lecture and assignment schedule.

Lecture & Assignment Schedule

Week	Lecture	Date	Day	Topic	Assignment
01	01	2024-01-22	M	Special: Logistics & Computational Experiments	--
01	02	2024-01-24	W	Quantitative fundamentals 1: Descriptive stats	--
02	03	2024-01-29	M	Quantitative fundamentals 2: Inference	HW1
02	04	2024-01-31	W	Sequences 1: Molecules to data	HW1
03	05	2024-02-05	M	Sequences 2: Alignment + mapping	HW1
03	06	2024-02-07	W	Genomics 1: Sequence QC + assembly	HW1
04	07	2024-02-12	M	Genomics 2: Comparative genomics + phylogenetics	HW2
04	08	2024-02-14	W	Genomics 3: Genome annotation + mapping redux	HW2
05	--	2024-02-19	M	NO CLASS: President's Day	HW2
05	09	2024-02-21	W	Special: Scientific data visualization	HW2
06	10	2024-02-26	M	Metagenomics 1: Concepts + amplicon methods	HW3
06	11	2024-02-28	W	Metagenomics 2: Shotgun methods	HW3
07	--	2024-03-04	M	**Journal Club: Series 1A**	HW3
07	--	2024-03-06	W	**Journal Club: Series 1B**	HW3
08	--	2024-03-11	M	NO CLASS: Spring Break	--
08	--	2024-03-13	W	NO CLASS: Spring Break	--
09	12	2024-03-18	M	Gene expression 1: Arrays + bulk RNAseq	HW4
09	13	2024-03-20	W	Gene expression 2: Single-cell methods + clustering	HW4
10	14	2024-03-25	M	Gene expression 3: Classification + catch-up	HW4
10	15	2024-03-27	W	Networks 1: Networks intro + physical interactions	HW4
11	16	2024-04-01	M	Networks 2: Regulatory motifs + interactions	HW5
11	17	2024-04-03	W	Networks 3: Genetic perturbations + interaction	HW5
12	--	2024-04-08	M	**Journal Club: Series 2A**	HW5
12	--	2024-04-10	W	**Journal Club: Series 2B**	HW5
13	--	2024-04-15	M	BST 281 HOLIDAY: Marathon Monday OR Buffer	FP
13	18	2024-04-17	W	Special: Protein families and structures	FP
14	19	2024-04-22	M	Special: Systems biology and dynamics	FP
14	20	2024-04-24	W	Special: Quantitative genetics / GWAS	FP
15	21	2024-04-29	M	Special: Epigenomics	FP
15	22	2024-05-01	W	Special: Metabolomics and proteomics	FP
16	--	2024-05-06	M	**Final Project Lightning Talks A + in-class help**	FP
16	--	2024-05-08	W	**Final Project Lightning Talks B + in-class help**	FP