QCB455/COS551 – Introduction to Genomics & Computational Biology

Fall 2021

INSTRUCTORS

Joshua Akey – jakey@princeton.edu
Mona Singh – mona@princeton.edu
Claire McWhite - cmcwhite@princeton.edu

TEACHING ASSISTANTS

Micah Fletcher – micahf@princeton.edu
Danny Simpson – ds65@princeton.edu
Riley Skeen-Gaar – rileyrs@princeton.edu

We will use Ed on canvas to host all communication. Please note that all questions about problem sets should be asked on Ed.
There are many advantages to using a discussion forum, including if someone else has had the same question as you, you get your answer immediately. Students are free to answer questions that other students ask, but please do not answer questions if you are just guessing. Please note that no code or solutions should be posted to Ed. Note: discussion questions will be posted on Canvas

Please allow up to 24 hours for the TAs to respond to questions.

LECTURES

Tuesday and Thursday 11-12:30 in Carl Icahn Lab 101. In case we have to convert to remote teaching, pre-recorded lectures will be made available on canvas by 11am on the date designated below in the syllabus. Students should watch the videos and go through the lecture notes within 48 hours of their posting.

PRECEPTS (optional)

Day/time: Tuesdays, 7:30 – 8:20 pm EST
Location: CIL 101

Precepts will give an introduction to R, and will work through some problems relevant for homework and the material covered in lectures. Precepts are optional.

OFFICE HOURS

Faculty office hours to answer questions about the material presented in lecture:
Dr. Singh: Tuesday 1pm EST in CIL 250 on the Tuesdays she lectures.
Dr. Akey: Tuesday 1pm EST in CIL 146 on the Tuesdays he lectures
Dr. McWhite: Tuesday 1pm EST in CIL (room TBA) on the Tuesdays she lectures.

TAs’ office hour to answer questions about precept and problem sets: **Wednesdays, 2-3 PM and Thursdays 2-3 PM.**
Location: TA office hours will be held via Zoom, there is a permanent link also available on Ed in the Welcome message.
https://princeton.zoom.us/j/92645020034?pwd=cXBLVHRoaGdFaUU3QVFZMWp6RTV6Zz09

COURSE DESCRIPTION

The course provides an overview of experimental and computational approaches for deciphering genomes and studying molecular systems. We focus on methods for analyzing large-scale "omics" data, such as genome and protein sequences, gene expression, and molecular interaction networks. We also cover the basic biology of the genome, statistical concepts relevant to genomics (data evaluation, estimation of true and false positives, significance testing, multiple testing correction), machine learning methods (e.g., hidden Markov models, clustering, classification techniques) as applied to problems arising in computational biology, and algorithms relevant for genomics.

Note: This is a cross-listed, highly interdisciplinary course. The material covered will include basic biology, computational biology, and statistics, and will provide a broad background that students can then use in future coursework and/or research in quantitative and computational biology. Students who want an in-depth view of any one of these areas and are less interested in the other two may wish to take another course. Given the interdisciplinary nature of this course, we understand that students have different technical strengths. Some of the material may be out of your "comfort zone." Our goal is to have an engaged group of students who are excited about and committed to learning the material we cover. Historically, students who make a good faith effort in the class obtain good grades.

PREREQUISITES

Some programming experience. COS126 or equivalent preferred.

HOMEWORK

Homework assignments can be completed in R or the language of your choice. TAs will cover the basics of programming in R during the first two optional help sessions and will answer any question on R code during office hours or help sessions. Due to time limitations, TAs will only provide feedback and give partial grading on R code. Assignments submitted in R can be formatted using the R markdown language, or can otherwise submitted as a pdf file along with the R code. Homework to be graded must be submitted through canvas.

Homework assignments are due at **5:00 pm EST** on the scheduled due date. All students will have 5 total 'late days' which can be used throughout the semester to turn in homework or other assignments past their due date (5 for the entire semester, not per assignment). Once an individual's late days are used up, late assignments will be subject to grade penalties: –10% of total points for up to 24 hours delay; –25% for up to 48 hours delay; –50% for up to 72 hours delay; no credit for delays greater than 72 hours. Please allow up to 24 hours for the TAs to respond to questions about the problem sets.

FINAL PROJECT

Final projects may be done individually or in groups of up to three. Group projects are expected to be more ambitious and deeper in scope, so that the amount of work a person would do is the same whether doing an individual or pair project. There are two options for the final project:

(1) You can reproduce results from an "omics" publication of your choice. Each student is expected to choose a paper and repeat the analysis (or parts of the analysis) of the published data by writing original code and generating new figures. The students are expected to reproduce or refute the results of the original publication, validating or rejecting the authors' conclusions.
(2) You can perform independent research, involving computational or statistical analysis, that builds on existing work and is on a topic of your choosing.

The evaluation of the final project will be done in four steps:

1. Proposal: The students will write a project proposal describing their plan proposed for the final project. The proposal must be 1-2 pages long. For option (1), this proposal should include a summary of the chosen publication, why it is important, a description of the data you have already obtained, and which parts of the analysis you plan to reproduce. For option (2), this proposal should include why the problem is important, a summary of prior computational research on the topic, a description of the data you have already gathered for analysis, what analysis you plan to do, and how you will judge the success of your work.
2. Progress report: Please give a 1-2 page update on the data analysis you have performed. Describe (and preferably include) what code you've already written, any difficulties you've encountered, and what remains to be done for your final project.
3. Final paper: Each student will **individually** write a short paper with an introduction, description of the methods, results, and conclusions, and including their code. Details will be forthcoming. For pair or group projects, a student's paper should focus on his or her contributions.

The final project write up is due at 5 pm EST on Dean's date (Tuesday **December 14, 2021**). Please note that university regulations do not allow us to make extensions beyond this date.

GRADING POLICY

Participation	10% (answering or asking questions during lecture, discussions on canvas, answering online discussion questions posted after lectures.)
Homework	45% (5% for assignment 0, 10% assignments 1-4)
Final project	45% (10% for proposal (part 1), 10% for progress update (part 2), 25% for final report)

Re-grading: assignments submitted for regrading will be re-checked as a whole and additional points, on any part of the assignment, may be added or deducted.

COLLABORATION POLICY

Students are allowed to work in groups on homework. However, each student must write up his/her own homework report (code and answers), and their turned in assignment should mention each person they talked to about the homework (including TAs and other students).

CALENDAR

Monday	Tuesday	Wednesday	Thursday	Friday
Aug. 30	**Aug. 31**		**Sept. 2** **Josh Akey** Introduction to Computational Biology Homework 0 assigned	**Sept. 3**
Sept. 6 **Labor Day** (no classes)	**Sept. 7** **Josh Akey** Data science concepts I: Study design, Exploratory Data analysis, and Probability Distributions **Precept 1** Introduction to programming in R - part I	**Sept. 8**	**Sept. 9** **Josh Akey** Data science concepts II: Hypothesis tests, multiple testing correction, and Evaluating classification tests	**Sept. 10** Homework 0 due Homework 1 assigned
Sept. 13	**Sept. 14** **Josh Akey** DNA Sequencing I: Sequencing technologies and data files **Precept 2** Introduction to programming in R - part II	**Sept. 15**	**Sept. 16** **Josh Akey** DNA Sequencing II: Analysis of next-generation sequencing data	**Sept. 17**

Sept. 20	**Sept. 21** **Josh Akey** Population Genomics I: Genetic variation, the coalescent, and linkage disequilibrium **Precept 3** p-values, multiple hypothesis correction	**Sept. 22**	**Sept. 23** **Josh Akey** Population Genomics II: Population structure and polygenic risk scores	**Sept. 24** Homework 1 due Homework 2 assigned
Sept. 27	**Sept. 28** **Josh Akey** RNA-seq **Precept 4** RNA-seq and negative binomial modeling	**Sept. 29**	**Sept. 30** **Josh Akey** eQTLs and allele-specific expression	**Oct. 1**
Oct. 4	**Oct. 5** **Josh Akey** CRISPR-based functional genomics **Precept 5** Linear modeling	**Oct. 6**	**Oct. 7** **Mona Singh** Gene expression clustering, gene ontology and other functional standards	**Oct. 8** Homework 2 due Homework 3 assigned
Oct. 11	**Oct. 12** **Mona Singh** Regulatory genomics: Chip-seq **Precept 6** Hypergeometric distribution, Poisson distribution	**Oct. 13**	**Oct. 14** **Mona Singh** Regulatory genomics: Motif finding	**Oct. 15**
Oct. 18 Fall recess	**Oct. 19** Fall recess	**Oct. 20** Fall recess	**Oct. 21** Fall recess	**Oct. 22** Fall recess

Oct. 25	**Oct. 26** **Mona Singh** Sequence analysis: BLAST and Multiple sequence alignment **Precept 7** E-values, extreme value distribution	**Oct. 27**	**Oct. 28** **Mona Singh** Sequence analysis: HMMs and domains	**Oct. 29** Homework 3 due Project proposal assigned
Nov. 1	**Nov. 2** **Mona Singh** Predicting protein features: neural network introduction **Precept 8** Project proposal discussion	**Nov. 3**	**Nov. 4** **Mona Singh** Cancer genomics I	**Nov. 5** Project proposal due Project update assigned
Nov. 8	**Nov. 9** **Mona Singh** Cancer genomics II **Precept 9** Supervised machine learning, classification	**Nov. 10**	**Nov. 11** Networks I: Introduction	**Nov. 12** Homework 4 assigned
Nov. 15	**Nov. 16** **Mona Singh** Networks II: Clustering and comparison **Precept 10** Network measures	**Nov. 17**	**Nov. 18** **Claire McWhite** Networks III: Technologies, Visualizations, Genetic Networks	**Nov. 19** Project update due
Nov. 22	**Nov. 23** **Claire McWhite**	**Nov. 24** Thanksgiving Break	**Nov. 25** Thanksgiving Break	**Nov. 26** Thanksgiving Break

	Networks IV: Functional networks **NO precept**			
Nov. 29	**Nov. 30** **Claire McWhite** Proteomics I **Precept 11** Final project discussion	**Dec. 1**	**Dec. 2** **Claire McWhite** Proteomics II	**Dec. 3**
Dec. 6 Homework 4 due	**Dec. 7** Reading period begins	**Dec. 8**	**Dec. 9**	**Dec. 10**
Dec. 13	**Dec. 14** Project final paper due	**Dec. 15**	**Dec. 16**	**Dec. 17**