STAT 540: Statistical Methods for High-Dimensional Biology

2021-2022 Winter Term 2

Overview

Credits: STAT 540 is a 3 credit graudate course with a mandatory computing seminar. It is cross-listed as STAT 540, BIOF 540, GSAT 540.

Course website: stat540-ubc.github.io/

Course-level learning objectives

- Perform exploratory data analysis and visualize genomics data
- Apply tailored statistical methods to answer questions using high dimensional biological data
- Make your work reproducible, reusable, and shareable
- Work with real data in a collaborative model

Selected topics

- Basics of molecular genetics/genomic and data collection assays (methods)
- Basic probability and math foundations
- Exploratory data analysis and data quality control
- Normalization, batch correction
- Causal inference and confounding effects
- Basic statistical inference ("one gene at a time") – linear models
- Large-scale inference ("genome-wide") – multiple testing
- Analysis of microarray, RNASeq, and epigenetics data
- Principal Component Analysis and clustering (unsupervised machine learning)
- Classification and cross validation (supervised machine learning)
- Gene set analysis and gene networks
- Genome-wide association analysis (GWAS)
- Single-cell genomics

Teaching Team

Instructors

- Keegan Korthauer (She/Her/Hers)
  Email: keegan@stat.ubc.ca
  Virtual office hours: Thursday 9-10am (Pacific time)
- Yongjin Park, PhD (He/Him/His)
  Email: ypp@stat.ubc.ca
  Virtual office hours: Wednesday 1-2pm (Pacific time)

Teaching Assistants

- Asfar Lathif (He/Him/His)
  Email: asfarlathifbt@gmail.com
  Virtual office hours: Tuesday 10-11am (Pacific time)
- Marco Tello (He/Him/His)
  Email: Marco.TelloPalencia@bcchr.ca
  Virtual office hours: Thursday 11am-12pm TBA (Pacific time)

Schedule

Note that as part of UBC's response to the COVID-19 omicron variant, lecture and seminar will be held online for the first two weeks of the term (synchronously via Zoom; meeting links posted in Canvas). It is expected that lecture and seminar will be held in person starting January 24. In case of any changes to this plan, stay tuned to UBC Broadcast communications.

Lectures held in person will also be live-streamed whenever possible (links shared in Canvas) to minimize disruption for those students who are unable to attend class due to illness or isolation. Seminar sessions will not be recorded, as these are primarily interactive.

Lectures (Sec 201)

- Time : Mon Wed 9:30-11am
- Location:
  - **Before January 24**: online (Zoom meeting links posted in Canvas)
  - **On January 24 and after**: in person in ESB 2012
- See course website for lecture materials and schedule

Seminars (Sec S2B)

- Time : Mon 12pm-1pm
- Location:
  - **Before January 24**: online (Zoom meeting links posted in Canvas)
  - **On January 24 and after**: in person in ESB 1012
- See course website for schedule and seminar materials
- We strongly recommend reading the seminar materials prior to attending each seminar.

Course communication

As we begin the term with web-oriented learning, it is vital that we have a dedicated plan for how we will communicate throughout the semester.

Announcements

Course announcements will be posted in the course Github Discussion repository (you will be granted access after we collect your Github user IDs). You are responsible for checking it regularly (subscribe to email notifications for this repo to make this easier).

General questions

Please also use the course Github Discussion repository for posting questions (e.g. topics discussed in class, questions about course organization, assignment clarifications, if you are stuck on an assignment and need help). This ensures the message can be seen by the entire teaching team, and that others in the class who might have the same question can learn from responses. You are also welcome to share your input on questions posted by others using comments.

Private matters

For private matters (e.g. requesting an extension or concerns about your grades), please contact the Teaching team by email (listed above). Please refrain from using email to ask general questions described above.

Group work

In your final project groups, we encourage you to make use of (1) the discussion feature in your GitHub Teams group, and (2) the issues feature in your group's project repo. In addition, you are encouraged to meet regularly via an online platform of your choice (or in person if conditions permit). Please reach out if you have any questions or challenges in this space.

Prerequisites and Resources

This interdisciplinary course requires a broad range of skills at the interface of statistics, molecular biology / genomics, and computing. While you may have some background in one or more of the following areas, you are not expected to be an expert in all. That said, to be successful in the course, you may need to spend a bit more time in the areas that you have less experience in. Although there are no official prerequisites for the course, here is a list of useful skills to bring into the course and/or learn along the way.

Statistics:

- You should have already taken a university level introductory statistics course.
- This free online book ”Modern Statistics for Modern Biology” by Susan Holmes and Wolfgang Huber is a great reference for introduction or review of many of the statistical concepts that are relevant for this course.
- This free online book ”Data Analysis for the Life Sciences” by Rafael Irizarry and Michael Love is another great resource for introduction or review of many of the statistical concepts relevant in this course, with an emphasis on implementation in R.

Biology:

- No requirements, but you are expected to learn things like, e.g. the difference between DNA and RNA, and the difference between a gene and a genome.
- This free online book "Concepts of Biology" by Fowler, Roush Wise is a great resource for biological concepts, in particular chapters 6 and 9
- This free online book "Biology" by Clark, Douglas Choi goes more in-depth, see Chapters 14, 15, and 16 for material on genetics that is particularly relevant for this course.
- No matter your prior experience, when you come across a new biological concept during the course or in your research, you might need to spend a bit of time 'learning as you go' - this is expected and something I still do regularly in my day-to-day research!

R:

- No experience required but be prepared to do a lot of self-guided learning if you haven't taken other courses on R or used it in your research.
- Start now by installing R and the HIGHLY RECOMMENDED "integrated development environment" (IDE) RStudio - both are free and open source.
- You should be able to run R on your own computer during each seminar session.
- If you are new to R, check out this blog post on getting started with R.
- This free online book "Introduction to Data Science" by Rafael Irizarry is also a great resource for getting more in-depth with R, programming basics, and the tidyverse. In particular see Chapters 1-5.

Other computing tools:

- In this course we'll be using the version control software Git and its web-based hosting and collaborative platform GitHub.
- The online resource "Happy Git and GitHub for the useR" from Jenny Bryan is a great reference for these tools as we learn them.
- Another helpful git resource is Hadley Wickham's webinar "Collaboration and time travel - version control with git, github and RStudio"
- We'll learn about using R markdown to generate readable and reproducible reports with code and text, and you'll be using that a lot in this course - see Chapter 18 of the 'Happy Git' resource mentioned above: "Test drive R markdown".

Evaluation

You will have three individual assignments, six seminar submissions (one divided into two parts), and one group project. Deadlines are all by 11:59 pm (Pacific time) on the due date, and late assignments are penalized 10% per day/partial day.

For more detail on each of these assignments and detailed instructions on how to turn in your work, see the course website

Intro Assignment (5%)

An introductory assignment designed to assess basic knowledge of GitHub, R and Rmarkdown

Seminar completion (10%)

- You will submit short ”deliverables” for seminars 1, 2 (split into two parts
- Seminar 2a and 2b together count as one deliverable), 3, 4, 5, and 7
- Each of these six Seminar session deliverables is weighted equally, but the lowest score will be dropped (so that the 5 deliverables with highest score will each count for 2% of the final grade).
- These deliverables give practical experience applying the knowledge that will be helpful on the homework assignment, final project, and (hopefully) your future research.
- Each deliverable is due on the Friday following the TA-led session for that seminar

Paper critique (5%)

Practice your ability to go through a paper, identify the biological problem that the authors want to address, and critique how they chose to approach this question. Read, summarize and critique this paper

Analysis assignment (30%)

- Involves detailed analysis of real data using R
- This assignment will assess your ability to understand and apply the methods learned in class

Group project (50%)

- A semester-long data analysis group project that will allow you to apply the techniques covered in class to a research question of your choosing
- Groups of target size of 4 students will be formed at the beginning of the course
- Important checkpoints during the term (with deliverables):
  - Initial project proposal (one-paragraph)
  - Finalize the one page project proposal
  - Progress report
  - 15 minute oral presentation
  - GitHub repository
  - Individual report deadline