

Biology 4511/5511: Biological Data Analysis for High Throughput Sequencing Data

Instructor: Dr. Susheel Sagar Bhat (ssbhat@sas.upenn.edu), Lynch 104.
Office hours: scheduled via email

Time: 3:30-5:00pm Tuesdays and Thursdays

Location: LLAB109

Course description and structure:

The objective of this course is to make analysis of high-throughput sequencing data accessible and approachable. We will focus on two widely used methods: Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) and RNA sequencing (RNA-seq) to learn tools and methods available for analyzing data thus obtained. Students will mostly use user friendly GUI based publicly available tools to perform the analysis while interest in learning and using scripts in R will be encouraged. Students will learn the underlying principles of various tools that are used in such data analysis and why/when such tools are useful. An emphasis on hands-on experience is the priority in this course. A substantial amount of time is allocated to allow students to practice and explore the tools available to them aimed at helping students get comfortable with such analysis.

This is not a bio-informatics class. A comprehensive dive into biostatistics/programming is beyond the scope of this course.

As mentioned earlier, this course relies heavily on in-class exercises, only a subset of resources can be made available online. Students are encouraged to attend classes to gain the most out of this course.

Grading:

Assessment will take into consideration participation in in-class exercises and the group (groups will be pre-assigned) projects. Two progress assessment slots have been delineated in the course. A small project (assessment 1) and a standard publication-style project (final project) carried out by students in groups will be assessed. Students will be assessed on their understanding of the principles that underly high-throughput data analysis. On the final day of classes students will “peer review” all presentations and discuss possible alternatives to the approaches presented in the preceding days.

Grade breakdown will look like this: 10% class participation, 30% assessment 1, 40% final project and 20% peer-review score.

Contributions of individuals within the group will be assessed by surveying other group members and will affect overall scores of “assessment 1” and “final project” proportionally.

Timelines:

As indicated in the schedule, the ideas for the assessment projects will be discussed in class. Students will have ample time to think through and submit (marked SUBMISSION in the schedule) a rough outline of the projects that they will undertake on specified dates. Adherence to those dates will be ideal, however, delays in submission or presentation will be accommodated under exceptional circumstances.

Resources:

1. Galaxy web server: <https://usegalaxy.org>
2. R and R-studio: <https://rstudio-education.github.io/hopr/starting.html>;
<https://swirlstats.com/>
3. Sequencing standards and pipelines:
<https://www.encodeproject.org/data-standards/>

Academic integrity:

Students are expected to be familiar with and comply with Penn's Code of Academic Integrity (<https://catalog.upenn.edu/pennbook/code-of-academic-integrity>). Cheating, plagiarism, and other forms of academic dishonesty will be dealt with strictly and all violations will be penalized with zero points for the assignment or quiz in question. Students suspected of violations may be referred to the Office of Student Conduct (<https://www.osc.upenn.edu/>) for further disciplinary action.

Use of AI tools like ChatGPT is allowed, provided such tools are cited properly.

		Additional info
Jan 16	Introduction	Discussion of what will be expected in the assessment projects
Jan 21	Sequencing data: quality control and trimming	
Jan 23	Sequencing data: quality control and trimming ctd.	
Jan 28	In class exercise	
Jan 30	Aligning reads to genome	
Feb 4	Aligning reads to genome ctd.	
Feb 6	In class exercise	
Feb 11	ChIP-seq data analysis: overview	Project for assessment-1 (outline and idea) SUBMISSION
Feb 13	Peak calling and differentially enriched peaks	Project for assessment-1(outline and idea) SUBMISSION
Feb 18	Function and annotation analysis	
Feb 20	Visualizing ChIP-seq data	

Feb 25	In class exercise	
Feb 27	Practice and QA	
Mar 4	<u>Progress assessment-1</u>	
Mar 6	<u>Progress assessment-1 ctd.</u>	
Mar 18	RNA-seq data analysis: overview	Final project (outline and idea) discussion
Mar 20	Counting reads	Final project (outline and idea) discussion
Mar 25	Differential gene expression	
Mar 27	Gene ontology analysis	
Apr 1	In class exercise	
Apr 3	Combining RNA and ChIP-seq	Final project outline and proposed analysis. SUBMISSION
Apr 8	Presenting data: plots and plots	
Apr 10	Miscellaneous sequencing techniques and their use cases	
Apr 15	In class exercise	
Apr 17	Practice and QA	
Apr 22	<u>Final project presentation</u>	
Apr 24	<u>Final project presentation</u>	
Apr 29	<u>Discussion and group evaluation</u>	