Human Genetics & Genomics: Bioinformatics Module 2021

Course Abstract

The goal of this course is to take students on an in-depth tour of a single common bioinformatics analysis pipeline for bulk RNA-sequencing analysis (RNA-seq). We will deconstruct the RNA-seq analysis pipeline starting with the easiest, highest level visualizations and working our way back through successively more advanced techniques. We will introduce common visualizations and exploratory analyses. We will start at the endpoint of the analysis, with an existing, finalized dataset, going step-by-step through the most common pipelines until we reach the very beginning, with alignment and quality control steps. We will emphasize practical utility of the methods, plus exposure to important ideas and concepts for students to explore on their own at a later date.

Course Format

The course takes place over 5 weeks on Tuesdays and Thursdays with 1.5 hour lectures. Each lecture will feature some review of the homework assignments, question sets, and technical walkthrough, followed by slides and discussion of newer material. The goal is that students should gain an understanding of fundamental concepts and be able to use their notes to recapitulate it with a future dataset if needed.

Special considerations for COVID-19

Due to social distancing protocols the course will continue online for a second year via password-locked Zoom, Slack.com or MS teams sessions. Details to be distributed via email.

Expectations: Attendance, Homework & Grading

The course is worth 100 points. Homework assignments, worth 10 points each, are expected to be turned in on time (*i.e.* time-stamped email *with a PDF attachment* no later than 8:59 AM PST the morning of the lecture listed as the due date). All assignments are graded for partial credit, and late assignments will be given a maximum of 50% of full credit *up to one lecture* after the original due date. Recommended readings are indicated on the syllabus. Assignment 1 is to choose one or more transcription factors from the ENCODE Transcription factor CRISPRi dataset and follow the lecture series in completing the steps of an analysis pipeline from data acquisition through through hypothesis testing and visualization using the GENAVi tool. Each stage of analysis and results is completed in successive homework assignments ("Labs") or question sets. Each question set is worth 10 points and will be graded based on completeness. Each student will also be required to present their homework at least once during the course prior to lecture or during a workshop. For this informal presentation, the student will be graded on their ability to discuss the figures they generated and any challenges encountered (10 points). A final exam worth 30 points will be given at the final lecture and will be based on the homework assignments. Students who have mastered the homework should find the exam to be straightforward.

Attendance is required no exceptions. Unexcused absences result in an incomplete grade. Advance permission may be obtained ahead of time or with extenuating circumstances after the fact from the graduate school (email to Emma Yates Kassler). Homework assignments are required on time (see schedule below) regardless of attendance or for half credit **one lecture late**.

Grades will be posted initially within 1 week after the final exam. However, it should be noted that Bioinformatics points are scaled proportionately to the number of lectures in the whole course, and averaged into the final biostats course grade with a different (and typically more lenient) grading scale.

Lecture & assignment schedule:

Tuesday, March 23 *Introduction* **Lecture topics:** Course structure and expectations. Install GENAVI. Availability of course materials. **Homework Assignment I:** (Lab) install GENAVi locally.

Thursday, March 25 *DGE* **Lecture topics:** ENCODE datasource. TCGA data and TCGAbiolinks-GUI. Preparing data for use in GENAVi. Differential gene expression (DGE) analysis using GENAVi. **Homework Asssignment II:** (Lab) Perform DGE on a TF of your choice.

Tuesday, March 30 *Review* Workshop / Homework Review

Thursday, April 1 *Exploratory Analysis* **Lecture topics:** Unsupervised clustering. Principle Components Analysis (PCA). Correlation. **Assignment III:** (Lab) Exploratory analysis of your chosen TFs in the CRISPRi dataset.

Tuesday, April 6 *GO Analysis* **Lecture topics:** Gene Ontology (GO) & Pathway Analysis, David, GoRilla, GENAVI. Gene Set Enrichment Analysis (GSEA) and MSigDB. Installing software. **Assignment IV:** (Lab) GO analysis of your TF DGE set.

Thursday, April 8 *Review* Workshop / Homework Review

Tuesday, April 13 *NGS & File Formats* **(Coetzee)** **Lecture topics:** Log files from alignment and QC analyses. Structure and interpretation of NGS file formats (*e.g.* FASTQ, .sam *etc.*) **Assignment V:** (Question set)

Thursday, April 15 *MultiQC* **(Coetzee)** **Lecture topics:** Performing Quality control with FASTQC and MultiQC. MegaQC for aggregating and databasing MultiQC reports. **Assignment VI:** (Question Set)

Tuesday, April 20 *Overview* **Lecture topics:** Question Set review. Course Overview / Review. Course Feedback and Evaluation.

Thursday, April 22 *Final Exam*

Oral presentations

On the date that each assignment is due, two students will present and discuss their answers to the assignments or question sets to discuss the results of a homework assignment and any challenges encountered (worth 10 points). We will schedule at the first lecture, and each pair of students is strongly encouraged to finish their assignments early and coordinate on their presentations. The goal of this exercise is to foster collaborative scientific storytelling and discussions. Students will not be penalized twice for wrong answers on their submitted work, but the presentation should exhibit strong communication of logic and reasoning behind the answers given.

Schedule and Due Dates:

These schedules are subject to change

day	date	lecture	hmwk (pts)	due
Tue	03/23	Intro	L1 (10)	
Thu	03/25	DGE	L2 (10)	L1
Tue	03/30	Wkshp	-	L2
Thu	04/01	Expl Anls	L3 (10)	
Tue	04/06	GO Anls	L4 (10)	L3
Thu	04/08	Wkshp	-	L4
Tue	04/13	Formats	QS5 (10)	
Thu	04/15	MultiQC	QS6 (10)	QS5
Tue	04/20	Review	-	QS6
Thu	04/22	Exam	E1 (30)	E1

Course schedule and outline. Numbered assignments are preceded by 'L' for labs, 'QS' for question sets and 'E' for exams. Optional times for oral presentation are during workshops.

Course Materials

All course materials, including this syllabus, lab workflows, question sets, and lecture slides will be made available in a Cedars account box.com folder.