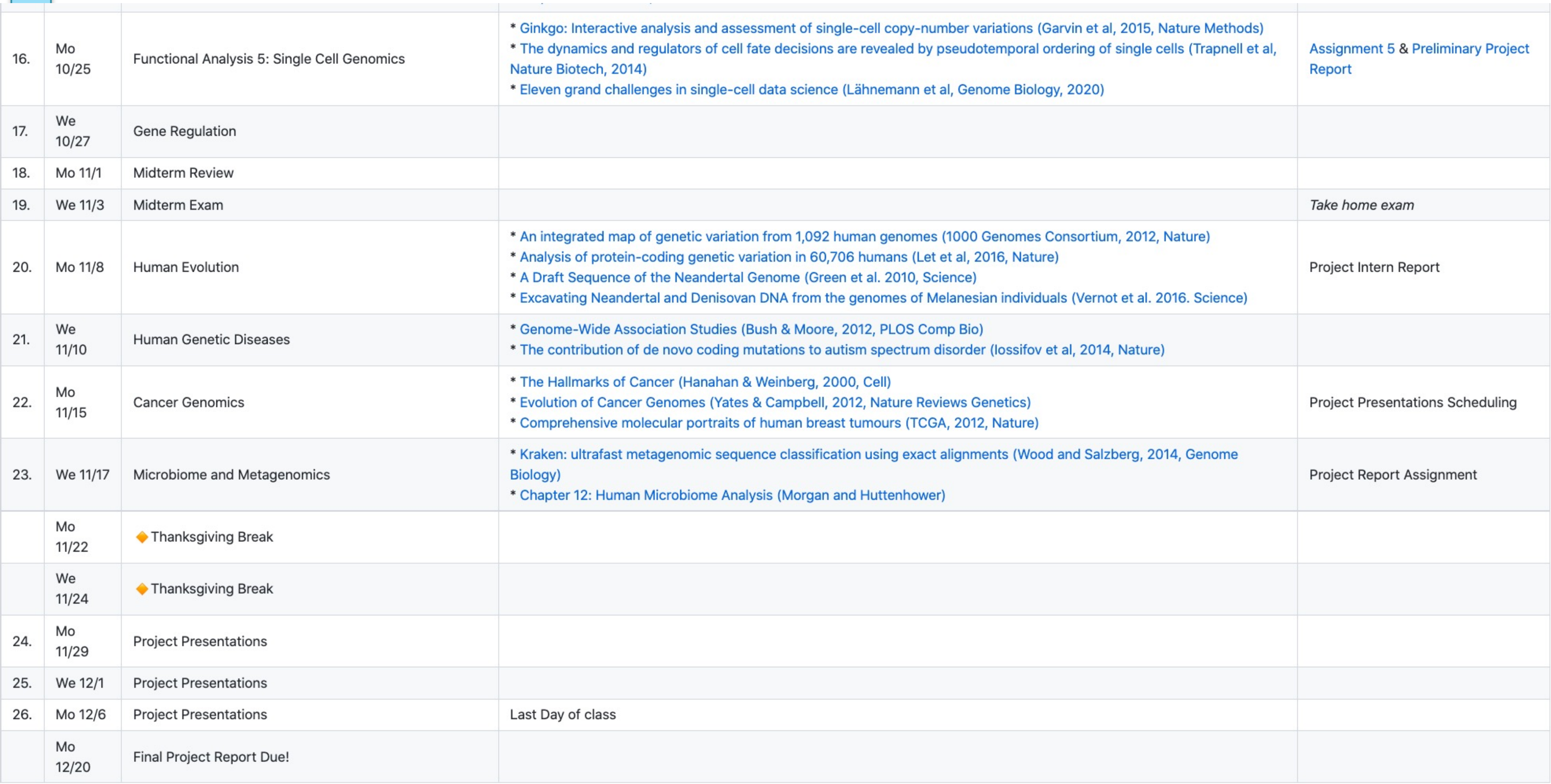
Lecture 18. Midterm review

Michael Schatz

November 1, 2021

Advanced Biomedical Research

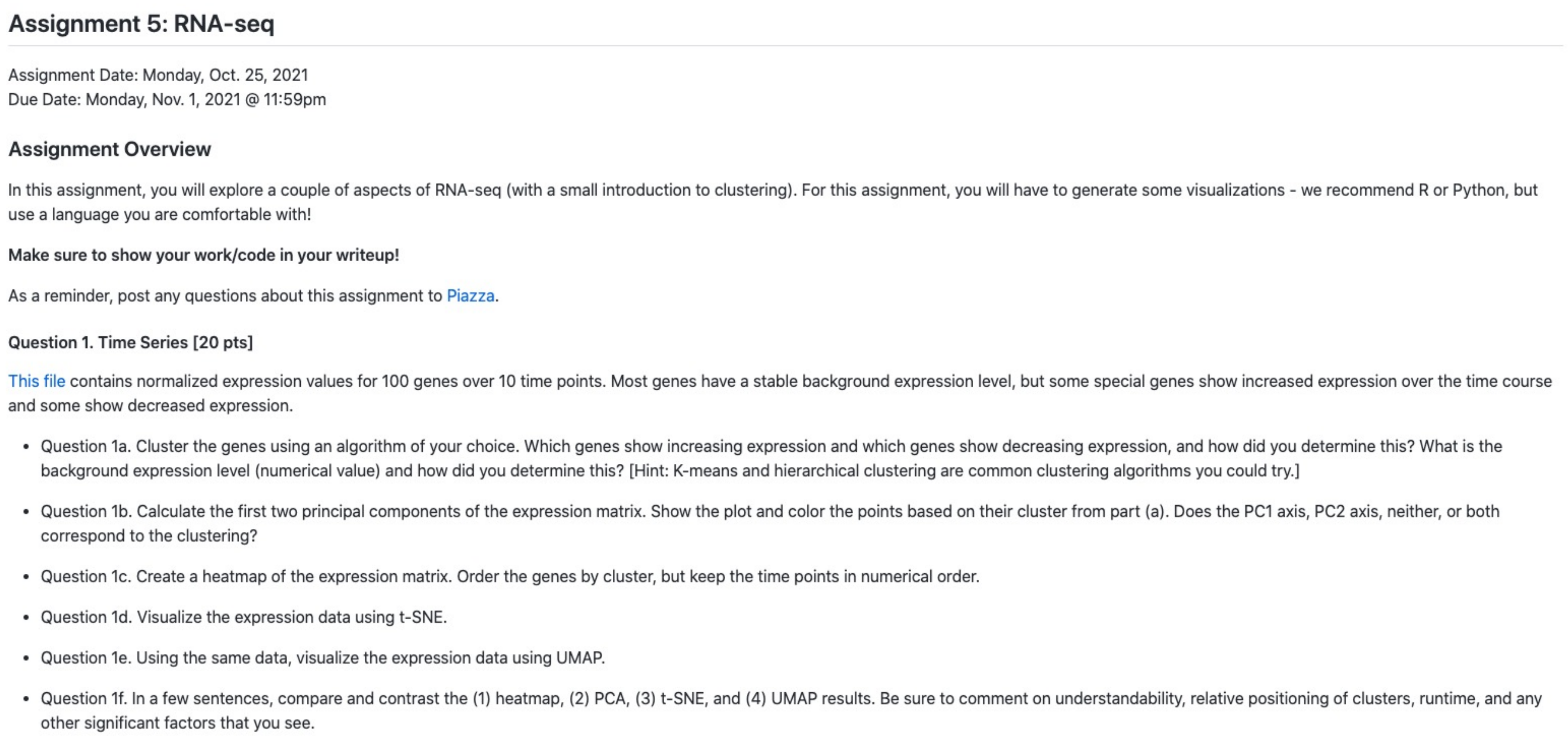
Updated schedule!



16.	Mo 10/25	Functional Analysis 5: Single Cell Genomics	* Ginkgo: Interactive analysis and assessment of single-cell copy-number variations (Garvin et al, 2015, Nature Methods) * The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells (Trapnell et al, Nature Biotech, 2014) * Eleven grand challenges in single-cell data science (Lähnemann et al, Genome Biology, 2020)	Assignment 5 & Preliminary Project Report
17.	We 10/27	Gene Regulation		
18.	Mo 11/1	Midterm Review		
19.	We 11/3	Midterm Exam		*Take home exam*
20.	Mo 11/8	Human Evolution	* An integrated map of genetic variation from 1,092 human genomes (1000 Genomes Consortium, 2012, Nature) * Analysis of protein-coding genetic variation in 60,706 humans (Let et al, 2016, Nature) * A Draft Sequence of the Neandertal Genome (Green et al. 2010, Science) * Excavating Neandertal and Denisovan DNA from the genomes of Melanesian individuals (Vernot et al. 2016. Science)	Project Intern Report
21.	We 11/10	Human Genetic Diseases	* Genome-Wide Association Studies (Bush & Moore, 2012, PLOS Comp Bio) * The contribution of de novo coding mutations to autism spectrum disorder (Iossifov et al, 2014, Nature)	
22.	Mo 11/15	Cancer Genomics	* The Hallmarks of Cancer (Hanahan & Weinberg, 2000, Cell) * Evolution of Cancer Genomes (Yates & Campbell, 2012, Nature Reviews Genetics) * Comprehensive molecular portraits of human breast tumours (TCGA, 2012, Nature)	Project Presentations Scheduling
23.	We 11/17	Microbiome and Metagenomics	* Kraken: ultrafast metagenomic sequence classification using exact alignments (Wood and Salzberg, 2014, Genome Biology) * Chapter 12: Human Microbiome Analysis (Morgan and Huttenhower)	Project Report Assignment
	Mo 11/22	🔶Thanksgiving Break		
	We 11/24	🔶Thanksgiving Break		
24.	Mo 11/29	Project Presentations		
25.	We 12/1	Project Presentations		
26.	Mo 12/6	Project Presentations	Last Day of class	
	Mo 12/20	Final Project Report Due!		

https://github.com/schatzlab/biomedicalresearch2021

Assignment 5: RNAseq
Due Nov 1 @ 11:59pm



Assignment 5: RNA-seq

Assignment Date: Monday, Oct. 25, 2021
Due Date: Monday, Nov. 1, 2021 @ 11:59pm

Assignment Overview

In this assignment, you will explore a couple of aspects of RNA-seq (with a small introduction to clustering). For this assignment, you will have to generate some visualizations - we recommend R or Python, but use a language you are comfortable with!

Make sure to show your work/code in your writeup!

As a reminder, post any questions about this assignment to Piazza.

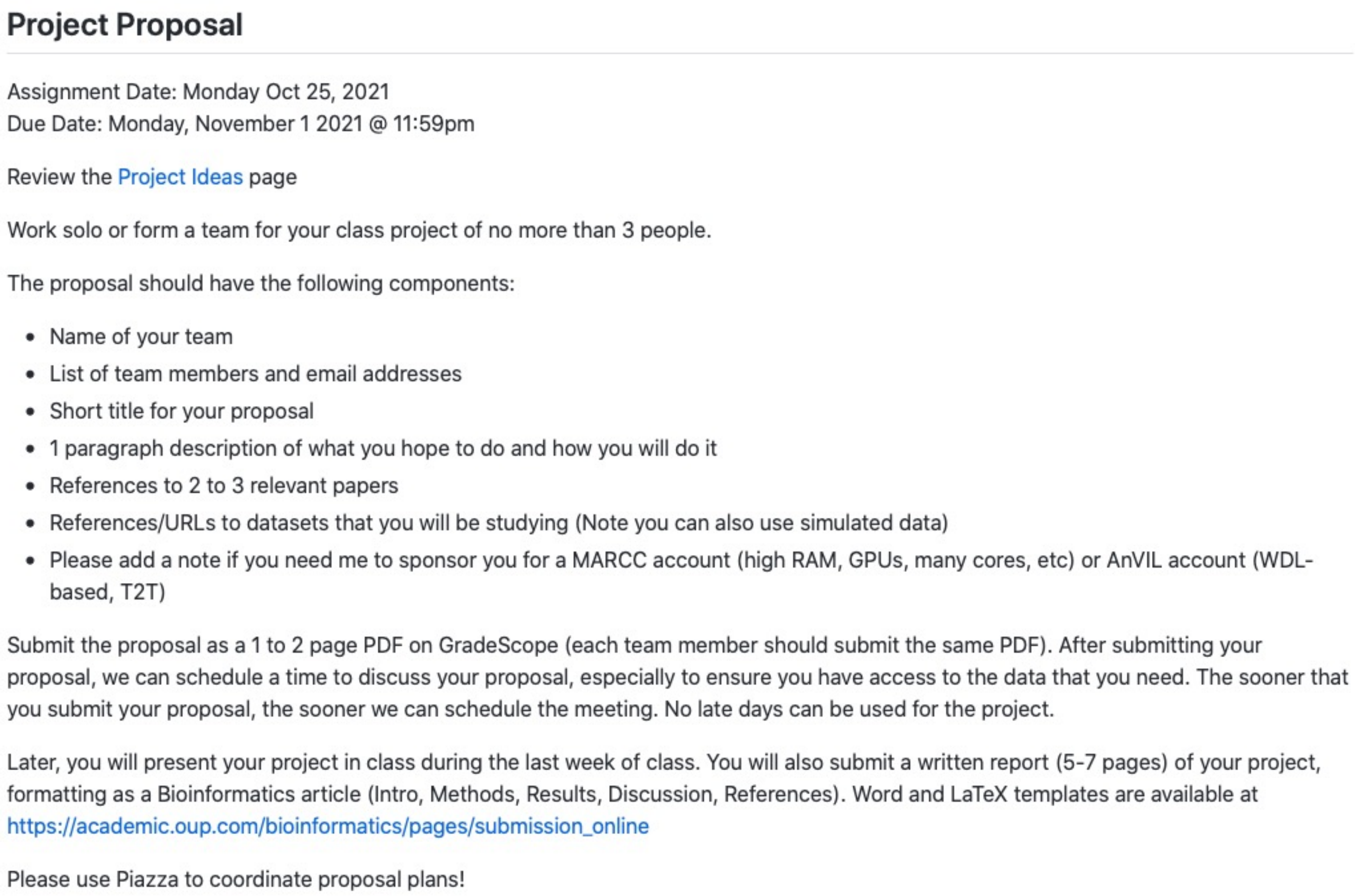
Question 1. Time Series [20 pts]

This file contains normalized expression values for 100 genes over 10 time points. Most genes have a stable background expression level, but some special genes show increased expression over the time course and some show decreased expression.

- Question 1a. Cluster the genes using an algorithm of your choice. Which genes show increasing expression and which genes show decreasing expression, and how did you determine this? What is the background expression level (numerical value) and how did you determine this? [Hint: K-means and hierarchical clustering are common clustering algorithms you could try.]
- Question 1b. Calculate the first two principal components of the expression matrix. Show the plot and color the points based on their cluster from part (a). Does the PC1 axis, PC2 axis, neither, or both correspond to the clustering?
- Question 1c. Create a heatmap of the expression matrix. Order the genes by cluster, but keep the time points in numerical order.
- Question 1d. Visualize the expression data using t-SNE.
- Question 1e. Using the same data, visualize the expression data using UMAP.
- Question 1f. In a few sentences, compare and contrast the (1) heatmap, (2) PCA, (3) t-SNE, and (4) UMAP results. Be sure to comment on understandability, relative positioning of clusters, runtime, and any other significant factors that you see.

https://github.com/schatzlab/biomedicalresearch2021

Class Project!
Proposal Due Nov 1



Project Proposal

Assignment Date: Monday Oct 25, 2021
Due Date: Monday, November 1 2021 @ 11:59pm

Review the Project Ideas page

Work solo or form a team for your class project of no more than 3 people.

The proposal should have the following components:

- Name of your team
- List of team members and email addresses
- Short title for your proposal
- 1 paragraph description of what you hope to do and how you will do it
- References to 2 to 3 relevant papers
- References/URLs to datasets that you will be studying (Note you can also use simulated data)
- Please add a note if you need me to sponsor you for a MARCC account (high RAM, GPUs, many cores, etc) or AnVIL account (WDL-based, T2T)

Submit the proposal as a 1 to 2 page PDF on GradeScope (each team member should submit the same PDF). After submitting your proposal, we can schedule a time to discuss your proposal, especially to ensure you have access to the data that you need. The sooner that you submit your proposal, the sooner we can schedule the meeting. No late days can be used for the project.

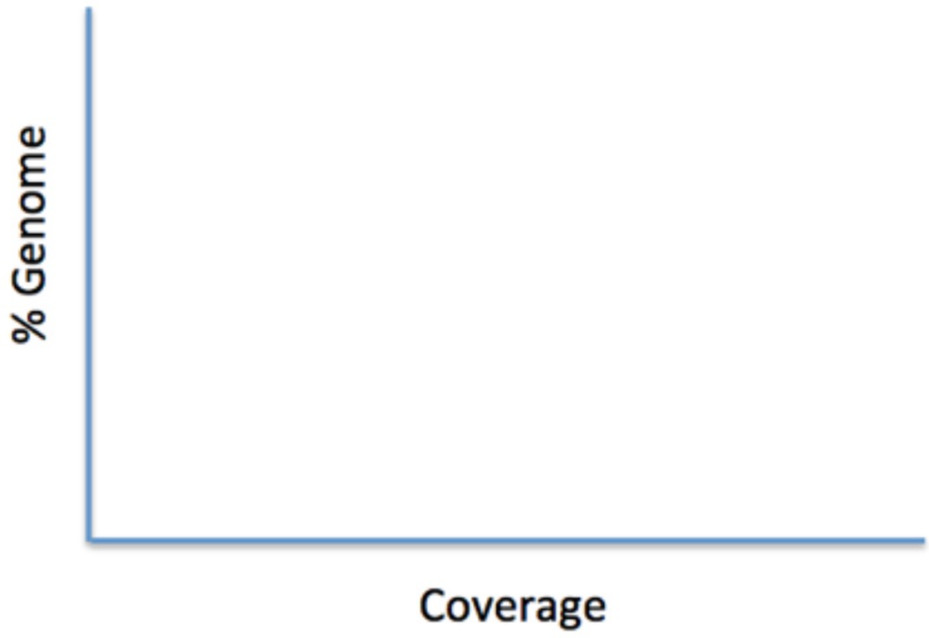
Later, you will present your project in class during the last week of class. You will also submit a written report (5-7 pages) of your project, formatting as a Bioinformatics article (Intro, Methods, Results, Discussion, References). Word and LaTeX templates are available at https://academic.oup.com/bioinformatics/pages/submission_online

Please use Piazza to coordinate proposal plans!

https://github.com/schatzlab/biomedicalresearch2021/blob/main/project/proposal.md

Sample Question

Q3. The Maryland blue crab genome is 1 Gbp in size. Approximately how many 100bp reads should we sequence so that we expect at least 99.85% of the genome will be sequenced at least 40 times? Sketch the expected coverage distribution for this number of reads; be sure to clearly label the mean coverage, and how 40 fold coverage relates to the mean. (Hint: In a normal distribution, 68.2% of the data is within 1 standard deviation of the mean, 95.4% within 2, 99.7% within 3, and 99.9% within 4)



Exam Topics

Genomics

- Genomics Technologies
  - Illumina, PacBio, Nanopore
- Genome Assembly
- Human Genome & T2T
- Whole Genome Alignment
- Read mapping
- Variant Identification
- Gene Finding
- RNA-seq
- Methyl-seq, Chip-Seq, Hi-C
- Genome Annotation

Quantitative Techniques

- Normal, Poisson, Binomial, P-value
- de Bruijn and overlap graphs
- kmers
- Dot plots
- Quality Values (Phred Scale)
- Full text indexing & suffix arrays
- Seed & Extend
- Hidden Markov Models
- PCA / t-SNE / UMAP
- Differential Expression
- Expectation Maximization

What is the goal? What is the approach? What are the key challenges?

How did we explore these topics in the homeworks and lectures?