Computation for Human Biology
Fall 2017
Tuesdays and Thursdays 12:00-1:20PM
Location: TBD

Course Description

Biology and medicine are becoming increasingly data-intensive fields. This course is designed to introduce students interested in human biology and related fields to methods for working with large biological datasets. There will be in-class activities analyzing real data that have revealed insights about the role of the genome and epigenome in health and disease. For example, we will explore data from large-scale gene expression and chromatin state studies. The course will provide an introduction to the relevant topics in biology and to fundamental computational skills such as editing text files, formatting and storing data, visualizing data and writing data analysis scripts. Students will become familiar with both UNIX and Python. This course is designed at the introductory level. Previous university-level courses in biology and programming experience are not required.

Course Objectives

Students will be able to:

1. Use Unix and/or Python to view, sort and parse large data sets such as those from genome-wide gene expression studies.
2. Use computational methods to interpret if a variant in the genome is likely to exert its effects via a protein coding region or through regulatory elements including promoters or enhancers.
3. Analyze datasets from cases and controls to identify sites in the genome that are likely to be relevant to a disease.
4. Query a large data set and visualize the data by making a scatter plot, histogram or heatmap.
5. Conduct a collaborative programming project applying best practices for generating reproducible data analysis scripts.

Instructors

Anshul Kundaje, PhD.
Assistant Professor, Dept. of Genetics, Dept. of Computer Science
Office: Lane Building, L301
Office Hours: TBA
e-mail: akundaje@stanford.edu

Annette Salmeen, D.Phil.
Core Course Coordinator, Program in Human Biology
Office: Human Biology Building 20 Room 21F
Office hours: TBA
e-mail: asalmeen@stanford.edu phone: 650-723-7842

Class Notes and Readings

There are no textbooks required for this class. Class notes, articles or videos will be posted on Canvas. In-class activities will be posted as Jupyter notebooks (http://jupyter.org/). The notebooks will include background information as well as opportunities to run and edit blocks of Unix or Python code. Instructions for accessing the Jupyter notebooks will be provided in class.

There are various on-line resources for writing code that may be helpful for this course, these references are listed by topic in the class schedule and are indicated as "For Reference".

Computing

The in-class activities will require a laptop computer or other mobile device that can connect to the internet. Please bring your laptops to class. If you have any concerns about bringing a laptop to class please contact one of the instructors to let them know.

Grading

Class participation	10%
Pre-class assignments	10%
In-class Programming or Data Analysis activities	10%
Weekly Programming or Data Analysis assignments	40%
Collaborative Computational Biology Programing Project	30%

Pre-class assignments (due Tues. or Thurs. at 9AM)
The pre-class assignments will consist of ~3-5 multiple choice, short answer or reflection questions to help students prepare for the class discussions and in-class activities. The assignments will be based on short reading assignments, videos or class notes that will be posted on Canvas. Reading assignments may provide background for the biology topics or computational methods that will be used in class or may introduce students to social implications or ethical debates surrounding the topics that are being covered. Students should work independently on these assignments.

In-class Programming or Data Analysis activities (submit at the end of class or by 8PM the day of class)

Most class sessions will include in-class Python programming or data analysis activities. Students will complete these assignments by logging into Jupyter notebooks which will be hosted in the cloud. Students may work collaboratively on these assignments.

Weekly Programming or Data Analysis Assignments (due Mon. at 12PM)
The weekly programming or data analysis assignments are an opportunity to practice and build upon the methods that are presented in class. Students will write their own code for these assignments. The assignments will be posted on Canvas on Wednesdays by 5:00PM and will be due on Mondays by 12:00PM. Students should work independently on these assignments.

Collaborative Computational Biology Programing Project

The goal of this project is to help students experience different roles in a programming project and to apply the computational methods that they will learn in the course in the context of a new dataset. Working in teams, students will build Jupyter notebooks showing their analysis and visualization of contributions of genetic variability to one of three diseases. Instructors will help students access datasets from GWAS studies on Alzheimer's disease, obesity or Crohn's disease and will provide an assignment sheet with instructions for the steps of the project. Teams may select the dataset they wish to work on.

Assignment Objectives

1) Participate in a collaborative team-based programming project to gain insights into the workflow for computational projects
2) Experience different roles in a computational project including the role of project manager, code implementer and documentation provider.
3) Apply data analysis methods from the course to new problems
4) Create visualizations to help interpret and represent data.

Honor Code

Students may work together and share ideas for all of the In-class Programming or Data Analysis activities as well as for the Collaborative Computational Biology Programming Project. Students should complete and submit the weekly pre-class assignments and weekly data analysis assignments on their own. If students consult any sources besides the class notes for the weekly data analysis assignments, citations should be provided.

Students with Documented Disabilities

Students with Documented Disabilities: Students who may need an academic accommodation based on the impact of a disability must initiate the request with the Office of Accessible Education (OAE). Professional staff will evaluate the request with required documentation, recommend reasonable accommodations, and prepare an Accommodation Letter for faculty. For students who have disabilities that don't typically change appreciably over time, the letter from the OAE will be for the entire academic year; other letters will be for the current quarter only. Students should contact the OAE as soon as possible since timely notice is needed to coordinate accommodations. The OAE is located at 563 Salvatierra Walk (phone: 723-1066, URL: http://oae.stanford.edu

Affordability of Course Materials

Stanford University and its instructors are committed to ensuring that all courses are financially accessible to all students. If you are an undergraduate who needs assistance with the cost of course textbooks, supplies, materials and/or fees, you are welcome to approach me directly. If would prefer not to approach me directly, please note that you can ask the Diversity & First-Gen Office for assistance by completing their questionnaire on course textbooks & supplies: http://tinyurl.com/jpqbarn or by contacting Joseph Brown, the Associate Director of the Diversity and First-Gen Office (jlbrown@stanford.edu; Old Union Room 207). Dr. Brown is available to connect you with resources and support while ensuring your privacy.

Class Schedule and Assignments

	Topics	Assignments and Resources
UNIT 1: Introduction to Molecular Genetics: What are genes, DNA, RNA and proteins? Getting Started with Python		
Class 1	Course Introduction: What is Computational Biology and why does it matter? What is a gene and how can we read DNA sequences into a computer?	
In-class Activity 1	Get started with Python Working with command line Setting up a Working Directory Getting the Path of a File Downloading a Gene Sequence Reading a Gene sequence	For Reference: UNIX Command-line bootcamp http://rik.smith-unna.com/command_line_bootcamp/?id=9xnbkx6eaof
Class 2	How can we predict the protein product of a gene?	Watch: Cracking your Genetic Code (2012) Nova Documentary https://www.youtube.com/watch?v=3OidRzGlhS8 Watch:"Part 1- What is a gene?" https://www.23andme.com/gen101/
In-class Activity 2	Working with String Variables Write the complementary strand for a DNA sequence Predict the mRNA transcript for a gene Find an open reading frame Predict the protein product of a gene	For Reference: Jones, M., (2017) Python for Biologists. http://pythonforbiologists.com/index.php/introduction-to-python-for-biologists/2-printing-and-manipulating-text/
Class 3	How can we compare two or more DNA sequences or compare a DNA sequence to a reference genome?	Read: Touchman, J. (2010) Comparative Genomics. *Nature Education Knowledge* 3(10):13. http://www.nature.com/scitable/knowle

		dge/library/comparative-genomics-13239404
In-class Activity 3	Align gene sequences from two different organisms Align gene sequences from the same organism Use BLAST to find similar gene sequences in a database	For Reference: Jones, M., (2017) Python for Biologists. http://pythonforbiologists.com/index.php/applied-python-for-biologists/applied-python-1/
UNIT 2: Introduction to Functional Genomics: How are cell types different? Using Python to work with biological datasets		
Class 4	What causes variation in gene expression levels? What are non-coding regions of the genome doing?	Read: Kolta, G. (Sept. 5th 2012) "Bits of Mystery DNA, Far from Junk, Play Crucial Role", *New York Times* http://www.nytimes.com/2012/09/06/science/far-from-junk-dna-dark-matter-proves-crucial-to-health.html
In-class Activity 4	Finding potential regulators for a gene using BEDTools (closestBed).	For Reference: Quinlan, A., BED tools http://bedtools.readthedocs.io/en/latest/content/tools/closest.html
Class 5	How do you identify regulatory regions in the genome and measure gene expression levels?	Read: Diep, F. (April 12th, 2013), "Friction over Function: Scientists Clash on the Meaning of ENCODE's Genetic Data", *Scientific American* https://www.scientificamerican.com/article/friction-over-function-encode/
In-class	Looking at sequencing reads and	

Activity 5	sequencing files, an introduction to FASTQ files.	
Class 6	Finding enhancer and promoter regions in DNA sequences	Read: Chi, K. (Oct. 13th 2016), "The dark side of the human genome" http://www.nature.com/nature/journal/v538/n7624/full/538275a.html
In-class Activity 6	Extracting sequences corresponding to promoter or enhancer regions for a gene from CHIP-seq data (GetFastaBED, intersectBED). Datasets will be from the ENCODE project.	For Reference: Bailey et al., (2013) "Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data", *Plos Computational Biology.* http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003326 Quinlan, A., BED tools http://bedtools.readthedocs.io/en/latest/content/tools/intersect.html http://bedtools.readthedocs.io/en/latest/content/tools/getfasta.html
Class 7	How do you compare regulatory regions across cell types?	Read: Kolta, G. (Feb. 18th 2015) "Project Sheds Light on What Drives Genes", New York Times http://www.nytimes.com/2015/02/19/health/scientists-shed-light-on-circuits-that-control-genes.html
In-class Activity 7	Analysis of CHIP-seq data in the Epigenome Roadmap Browser	For Reference: Roadmap Epigenomic Tutorial, WashU, EpiGenome Browser

		http://epigenomegateway.wustl.edu/support/Browser_tutorial_v3_wREMC.pdf
Class 8	How do you compare gene expression levels across cell types?	
In-class Activity 8	Analysis of processed RNA-seq data Making heatmaps of gene expression data	For Reference: Griffith et al., (2015) " Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud", *Plos Computational Biology.* http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004393
Class 9	Visualizing gene expression differences across cell types	Watch: Ng, Andrew, K Means Clustering Algorithm https://www.youtube.com/watch?v=xnWFIgr34Lk
In-class Activity 9	Principle Component Analysis K-means clustering of cell types Correlation distance	
Class 10	How can you learn about the function of a gene product by studying other organisms?	Read: Ashburner M., et al.,(2000) 'Gene ontology: tool for the unification of biology." *Nature Genetics* 25, 25-29. http://www.nature.com/ng/journal/v25/n1/full/ng0500_25.html
In-class Activity 10	Clustering of genes GO-term enrichment	For Reference: http://www.geneontology.org/page/documentation
UNIT 3: Introduction to Population Genetics: How do genomes vary across populations? Using Python to analyze genomes		
Class 11	How do gene sequences vary across humans?	Watch: Part 2 What are SNPs? https://www.23andme.com/gen101/

		Read:Explainer "Structural variation" https://www.broadinstitute.org/explainer-structural-variation Read: Welcome Trust Sanger Institute, "1000 genomes for mankind" http://www.sanger.ac.uk/news/view/1000-genomes-mankind
In-class Activity 11	The 1000 Genomes Project UK10K	
Class 12	Visualizing genetic variation in humans: A look at data from 23 and me	Read: *TBD*
In-class Activity 12	Principle Component Analysis of data from 23 and me	For Reference: Genetic Ancestry by analysing 23AndMe Data using Python http://online.cambridgecoding.com/notebooks/cca_admin/genetic-ancestry-analysis-python
Class 13	Using principle component analysis to find artifacts	
In-class Activity 13	Principle Component Analysis of data demonstrating lab to lab variation	
UNIT 4 : Introduction to Genetics and Disease Participating in collaborative Computational Biology Programing Project		
Class 14	Introduction to Collaborative Programing Project	Read: Noble, W, "A Quick Guide to Organizing Computational Biology Projects" (2009) http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1000424
In-class	Getting started with the collaborative	

Activity 14	programing project	
Class 15	What is a genome wide association study and what can it tell us about the genetic basis for disease?	Read: Bush W. and Moore J., (Dec. 27th, 2012) 'Genome-Wide Association Studies", *PLOS Computational Biology.* Section 1-3. http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002822#pcbi.1002822.s001
In-class activity 15	Linkage Disequilibrium Haplotypes Correlation	
Class 16	Replication of GWAS studies in different populations.	Read: Bush W. and Moore J., (Dec. 27th, 2012) 'Genome-Wide Association Studies", *PLOS Computational Biology.* Section 4-5. http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002822#pcbi.1002822.s001
In-class activity 16	Comparing GWAS results in different populations	
Class 17	Analyzing GWAS Studies	Read: Bush W. and Moore J., (Dec. 27th, 2012) 'Genome-Wide Association Studies", *PLOS Computational Biology.* Section 6. http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002822#pcbi.1002822.s001 Submit: Draft of Collaborative Computational Biology Programming Project.
In-class	Analyzing GWAS Studies	

activity 17	Group work on Collaborative Computational Biology Programming Project	
Class 18	Rare variation Somatic mutations cancer	Read: Kennedy, P. (Nov. 25th, 2016) "The Thin Gene", New York Times http://www.nytimes.com/2016/11/25/opinion/sunday/the-thin-gene.html
In-class Activity 18	ExAC browser	
Class 19	Genome Reading to Genome Writing: CRISPR-CAS9	
In-class activity 19		
Class 20	Collaborative Programing Projects	Submit: Final version of Collaborative Computational Biology Programming Project.
In-class activity 20	Presentation of Collaborative Programing Projects	

Collaborative Computational Biology Programing Project

Overview

Working in teams, students will build Jupyter notebooks showing their analysis and visualization of contributions of genetic variability to one of three diseases. Instructors will help students access RNA Seq and CHIP-Seq data from studies on Alzheimer's disease, obesity or Crohn's disease. Teams may select the dataset they wish to work on.

Objectives

5) Participate in a collaborative team-based programming project to gain insights into the workflow for computational projects
6) Experience different roles in a computational project including the role of project manager, code implementer and documentation provider.
7) Apply data analysis methods from the course to new problems
8) Create visualizations to help interpret and represent data.

Timeline

Week 7 Tuesday Groups select team based project
Week 9 Tuesday Drafts of projects submitted on Canvas
Week 8 Monday Review of projects with feedback on rubrics submitted on Canvas
Week 9 Monday Instructor feedback on projects submitted on Canvas

Logistics

Each team will have three to four members and the final projects will have three or four parts accordingly. Students will alternate participating in roles as the Project Manager, Code Implementer and Documentation Provider.

The three roles are as follows:

Project manager: Establishes an overview for the project goals and objectives. Creates a list of the analysis that will be completed and the visualizations that will be produced.

Code implementer: Writes the code to conduct the data analysis.

Documentation Provider: Creates the final iPython notebook implementing the group plans, annotating code when necessary and adding features such as captions for any visualizations.

Groups of 3

	Project Manager	Code Implementer	Documentation
Project 1	Student 1	Student 2	Student 3
Project 2	Student 2	Student 3	Student 1
Project 3	Student 3	Student 1	Student 2

Groups of 4

	Project Manager	Code Implementer	Documentation
Project 1	Student 1	Student 2	Student 3
Project 2	Student 2	Student 3	Student 4
Project 3	Student 3	Student 4	Student 1
Project 4	Student 4	Student 1	Student 2

Collaborative Computational Biology Project Road Map