Applied Genomics
BIOL-GA.1130

This course provides a comprehensive introduction to the analysis of next generation DNA sequence (NGS) data. Through a combination of lectures, hands-on computational training, discussions of scientific papers, and assignments using real data, students will learn the foundations of analytical methods, the computational skills to implement those methods, and the reasoning skills to critically assess the primary literature in genomics. The course will cover all commonly used NGS methods including genome sequence analysis, gene expression analysis and protein-nucleic acid interactions. To gain practical expertise in executing bioinformatic analyses, students will undertake a series of assignments using real data. Students will also complete an individual project that integrates skills and concepts covered during the class and that is tailored to meet their background and training.

The course is designed for students with a background in biology who have some experience with statistical analyses using the R programming language. The course is also appropriate for computer science students with some biology background who wish to improve their skills in translating biological problems into computational approaches. The course is based on the premise that biological and computational research is now inextricably intertwined.

The goal of the course is to provide students with the necessary skills to move seamlessly from acquisition of genomic data to its analysis using UNIX commands and programming in R. Students will read primary research from the genomics and bioinformatics literature and work with real large-scale datasets. The course will teach students to synthesize data from the literature and devise novel computational experiments to test new ideas or hypotheses.

For the final project, students will be required to read the primary literature, identify a biological problem and the available datasets to work with, and undertake a computational analysis to tackle the problem. Students will generate a written report of their study and present their projects to instructors and fellow students.

Prerequisites: Statistics in Biology, or equivalent background in statistics and R programming with instructor permission.

Grading Scheme

20% class participation
15% midterm exam
25% home work assignments (five total)
40% final project

Syllabus

Instructors: David Gresham and Manpreet Katari
Teaching Assistant: Kostya
Course Number: BIOL-GA.1130
Credits: 4

Lecture/Computational lab: Wednesday 12:30pm-3:15pm
Recitation: Monday 12:30-1:30 p.m

Location of Lecture and Recitation: 12 Waverly Place (CGSB) Room L111

Prior to class:
-apply for HPC account
-attend R and UNIX bootcamp
-January 22-23 1pm-5pm
-12 Waverly Place (CGSB) Room L111

Each class will be lead by either David Gresham (DG) or Manpreet Katari (MK)

Week 1: Introduction to next generation sequencing (MK, DG)

Reading Assignment 1: None
Lecture 1: What is next generation sequencing, file formats, R/Bioconductor, Unix, HPC
Computer Lab 1: Access HPC, execute unix and R commands, PBS scripts, visualization, getting data from databases
Assignment 1: Getting started with HPC, R and NGS file types.

Week 2: Genome Alignment (MK)

Reading Assignment 2:
Lecture 2: How to align a genome using short read sequences
Computer Lab 2: FASTQC, Bowtie and BWA examples
Assignment 2: Align sequenced genome using Bowtie and BWA

Week 3: Detecting variants with next gen sequencing (DG)

Reading Assignment 3: "*An integrated map of genetic variation from 1092 human genomes",* Nature. (2012)

Lecture 3: Identifying SNPs, CNVs, translocations, low abundance mutations in NGS data
Computer Lab 3: samtools, bcftools
Assignment 3: Paired End alignment, VCF generation and CNV detection

Week 4: RNA-seq I: Alignment and Quantification (MK)

Reading Assignment 4: Mortavzi et al.,
Lecture 4: Mapping RNA-seq reads, quantification, splice variants, RNA editing
Computer Lab 4: Tophat, Cufflinks
Assignment 4: Gene expression analysis using RNA-seq with TopHat

Week 5: RNA-seq II: Analysis (DG)

Reading Assignment 5: Rapaport et el.,
Lecture 5: Differential gene expression analysis, hierarchical clustering, GO term enrichment
Computer Lab 5: edgeR, cluster
Assignment 5: Differential gene expression analysis, clustering, and PCA

Week 6: ChIP-seq (DG)

Reading Assignment 6: Boyle et al.,
Lecture 6: Design and Analysis of ChiP seq experiments
Computer Lab 6: MACS, detecting peaks, PWM

Week 7: Mid-term

Week 8: De-novo genome assembly (MK)

Reading Assignment 8:
Lecture 8: How to assemble a genome sequence without a reference
Computer Lab 8: SOAP-de novo, Velvet

Week 9: Project proposals

Week 10: Meta-genomics (MK)

Reading Assignment 9:
Lecture 9: Amplicon-based, reference bases, de novo
Computer Lab 9:

Week 11: Network analysis (DG)

Reading Assignment 11: None

Lecture 12: Generating and analyzing networks of interactions
Computer Lab 12: Cytoscape and iGraph

Week 12: Analyzing Protein-RNA interactions using NGS (DG)

Reading Assignment 11:
Lecture 12: Mapping protein-RNA interactions using NGS, PAR-Clip, RIbo-seq
Computer Lab 12:

Week 13: Project Presentations I

Week 14: Project Presentations II