Master of Science in Computational Biology

Module: Analysis of Next Generation Sequencing Data (ANGSD) [CMPB 5004 03]

Luce Skrabanek: las2017@med.cornell.edu
Mervin Fansler: mef3005@med.cornell.edu

Course Description:

Next generation DNA sequencing technology has revolutionized our ability to ask almost any question of our genome, epigenome or transcriptome. In Part I of this Module, we focus on the principles of the dominant technology: the Illumina short read sequencing by synthesis platform. The complete analysis pipeline is examined in detail, from the generation of raw reads, through alignment to the genome (Part II), and up to gene-centric analyses in Part III. At each step, there will be a strong emphasis on quality control, highlighting limitations and common pitfalls of the most commonly used tools, as well as ways to deal with them. In Part IV, alternate DNA sequencing technologies are surveyed, showcasing their applications.

Students will use the knowledge gained throughout this Module to apply to a practical project which will focus on the analysis of one or more NGS data types to address a biomedically relevant question.

Course Objectives:

After completing this Module, students will be able to:

- Have a deep appreciation of current DNA sequencing technologies, and an awareness of pitfalls, caveats, and confounding factors;
- Understand which technologies are appropriate for which use cases;
- Be aware of the details in deriving insights from raw data;
- Be able to critically assess next generation sequencing data and analyses, and be aware of common biases.

Semester:

Spring

Credits:
4

Pre-requisites:
None

Time/Day:
Tuesdays and Thursday, 10-11:30pm
January 10, 2023 - April 28, 2023

Assignments/Assessment:
60% of the grade will be assessed by an individual project, using techniques learned in class to explore a meaningful biological question. The project will be developed throughout the course, with opportunities every week to refine and get feedback. 40% of the grade will be assessed via weekly short programming exercises.

Grading:
Students in Master's programs will receive a letter grade (A-F); students in PhD programs will be graded using the Weill Cornell Graduate School grading system (Honors, High Pass, Low Pass, Fail).

Calendar:

Weeks 1-3; Part I: Design

- Introduction to High-Throughput DNA Sequencing
- Library preparation
- Illumina sequencing technology
- Experimental design

Weeks 4-6; Part II: Reads

Raw Data

- Public data access (GEO etc)
- Data formats
- Raw data QC

Read Alignment

- Reference genome
- Annotations [RNA-seq, ChIP-seq, methylation]
- Short read alignment, including filtering

Exploring Aligned Reads

- QC of mapped reads [RNA-seq, ChIP-seq, methylation, DNA]: sequencing depth, GC bias, genome coverage, contamination, correlation of replicate samples
- RNA-seq specific: 3'/5' bias, exon/intron coverage, gene numbers
- ChIP-seq specific: enrichment scores
- WGBS/eRRBS: CpG coverage

Weeks 7-12; Part III: Genes

Analysis of RNA-seq data

- Read counting; normalization; differential gene expression analysis

Analysis of ChIP-seq data

- Peak calling; differential peak analysis; motif finding; annotation

Analysis of DNA sequencing data

- Variant calling, tying into GWAS in Quantitative Genomics and Genetics module

Week 13; Part IV: Other technologies

- PacBio
- IonTorrent
- Nanopore
- Single cell DNA sequencing technologies
- **Week 14; Part V: Student presentations**