ESAM 472/372:
An Introduction to the Analysis of RNA Sequencing Data

Fall Quarter 2024: Tuesdays & Thursdays 2:00 – 3:20 p.m., Evanston campus

Introduction

- Course overview
  - Learning how to work with sequencing data
  - Aligning sequencing reads to the genome
  - Performing the differential expression analysis
  - Interpreting and presenting the results
- Course goals
  - Basic knowledge of mathematical/statistical assumptions and methods
  - Facility with the computational tools used to analyze the data
  - Aplomb from performing your own differential gene expression analysis
  - An idea about where to look if you want to go beyond this course
- Why do RNA sequencing?
  - The central dogma of molecular biology
  - Simple and more complicated views of RNA
  - Illumina sequencing — high throughput!
  - Sequencing repositories: NCBI, ENA, DDBJ

Module 1: Working with the data on a compute cluster (Quest)

- Working remotely on Quest with secure shell (ssh)
- The unix command line and how to use it
- Navigating directories/folders from the command line
- The class project folder on Quest
- Getting data onto Quest and working with it
  - Manually with wget
  - Using awk to automate wget downloads
  - Using screen to keep a transfer running
  - Faster transfers with ascp
  - Transfers with the Globus web interface
  - Command-line Globus transfers
- The format of a fastq file
- Quality checking with FastQC
- Using the quest job queueing system
- Summarizing multiple quality checks with MultiQC

Module 2: Sequence alignment and pseudoalignment

- Substring matching
- Suffix arrays
- Suffix trees
- Burroughs-Wheeler Transformation
- Approximate matching
  - Edit distance
  - Levenshtein algorithm and dynamic programming
  - Alignment scoring
- Some alignment software history
- Spliced read alignment
- Single-ended vs. paired-end reads
- Non-stranded vs. stranded reads
- Obtaining a reference genome
- Obtaining a gene annotation
- Using STAR
  - Building an index
  - STAR options
  - Output files produced by STAR
- SAM and BAM files
- Manipulating SAM and BAM files with samtools
- Viewing aligned reads with IGV
- How different STAR options affect aligned reads
- RSeQC: post-alignment quality checking
- Gene-body coverage plots
- Post-alignment transcript integrity number (TIN) checking
- Automating the alignment of an entire experiment on Quest
- Aligning to the transcriptome with RSEM
- RSEM's expectation-maximization method of assigning reads
- Counts vs. transcripts per million (TPM)
- Pseudoalignment with kallisto
- Comparing counts and pseudo-alignment expected counts
- Dealing with primer/adapter contamination
- Paired-end read-through

Module 3: Read counts and differential expression analysis

- Read counting with STAR
- Different read counting modes
- Counting strand-specific reads
- Automating the building of an experiment count table
- htseq-count, featurecounts
- Multi-mapped reads
- Working with the read count table in R
- Simple analyses with read counts: correlation plots, histograms
- Poisson vs. negative binomial count distributions
- Fitting a negative binomial to read counts
- A biophysical model giving a negative binomial distribution
- Normalizing counts for sequencing depth
- Differential expression with DESeQ2
  - Generalized linear model for read counts
  - Maximum likelihood estimates
  - Empirical-Bayes variance shrinkage
  - Gene-wise dispersion estimates
  - Wald test
- Hypothesis testing; type I vs. type II errors
- Multiple hypothesis testing
  - Bonferroni correction
  - Simes' procedure
  - Benjamini-Hochberg false discovery rate
- How the number of replicates affects the differential expression analysis
- ERCC spike-ins
- Adding gene names to gene ids

Module 4: Processing and visualizing the results

- MA plots
- Variance stabilizing transformation
- rlog transformation
- Heatmaps
- Clustering
- Principal components analysis (PCA)
- Non-negative matrix factorization

Module 5: Single cell sequencing

- History of single-cell methods
- Droplet-based sequencing
- What are barcodes and unique molecular identifiers (UMIs)?
- Read format for 10X Genomics data
- Aligning and counting reads with Cellranger
- The feature-barcode matrix
- Visualizing data with the Cellranger Loupe Browswer
- Aligning and counting reads with STARsolo
- Cell/barcode filtering
- Processing single-cell data with Seurat
  - Obtaining basic statistics
  - Methods for normalizing data
  - Displaying a PCA plot
  - Choosing the number of principal components
  - Generating cell clusters with Seurat's nearest neighbor graph
  - Using t-SNE to do dimension reduction
  - How does t-SNE work?
  - Choosing parameters for the dimension reduction
  - Using UMAP to do dimension reduction
  - Finding cluster markers