Pragmatic Bioinformatics for Bench Scientists Taught by Bench Scientists
GS04 1781; July 1 – August 5, 2020

Instructors: Jichao Chen, Ph.D., M.H.S.  jchen16@mdanderson.org (course director)
Sidney Wang, Ph.D.  Hsi.Ming.S.Wang@uth.tmc.edu
Kunal Rai, Ph.D. KRAi@mdanderson.org
Ed Ostrin, M.D., Ph.D.  EJOstrin@mdanderson.org
George Eisenhoffer, Ph.D.  GTEisenhoffer@mdanderson.org

Course description: Bioinformatics is becoming essential in the genomic era. Witnessing both the power and the complexity of bioinformatics, bench scientists, despite providing most of the biological insights, often feel left out as simply data generators, and frustrated when collaborating with data analyzers. This course, taught by bench scientists who have published on specific bioinformatics topics, aims to empower bench scientists with valid statistical and computational methods to be able to explore data and communicate with computational scientists. It is pragmatic because it covers as-needed theoretical background and teaches usable, instead of efficient, programming in the format of a dry-lab protocol that generates publication-quality figures. It consists of 6 modules covering principles, RNA, DNA, protein, images, and freeware. Each module takes 4 hours of class time, which consists of one 30-minute background and three 1-hour sessions with significant hands-on time. Grades are based on homework.

Objectives:
1. Understanding of the basic experimental design and statistical analysis in wet-lab research;
2. Develop working knowledge of bioinformatics analysis of RNA-seq, ChIP-seq, proteomics data;
3. Understand the basics and applications of image quantification;
4. Choose and use existing software;
5. Build necessary vocabulary to communicate with computational scientists.

Outline followed by individual modules’ content:
Module 1 (Principles): Experimental design and statistical testing – Dr. Sidney Wang (July 1st)
Module 2 (RNA): Single-cell RNA-seq using Seurat and Monocle – Dr. Jichao Chen (July 8th)
Module 3 (DNA): ChIP-seq and (sc)ATAC-seq – Dr. Kunal Rai (July 15th)
Module 4 (Protein): Mass spectrometry and protein structure – Dr. Ed Ostrin (July 22nd)
Module 5 (Image): ImageJ and Imaris – Dr. George Eisenhoffer (July 29th)
Module 6 (Freeware): GSEA and HOMER – Dr. Jichao Chen (August 5th)

Format: 1-5 pm on Wednesday afternoons
30 min: background
1st hr session: topic 1
10 min break
2nd hr session: topic 2
10 min break
3rd hr session: topic 3
Module 1 (Principles): Experimental design and statistical testing

30 min background: Why (how) is study design important? Learning from common mistakes in the published literature. Use examples that students are familiar with (e.g. gel image, qPCR) to illustrate the concept of sampling (mean, variance and underlying distribution). Further use these examples to illustrate how sampling bias and researcher bias could lead to erroneous conclusions. Similar principles applies to the broader context of publication bias.

1st hr session: First look at the data/low-level analysis/data transformation. Use example dataset to illustrate why we need normalization/standardization to help cultivate intuition on how it works. Also discuss the assumptions and scenario where the assumptions fail. Common approaches to visualize and evaluate data: PCA, hierarchical clustering.

2nd hr session: Modeling data. Explain how linear models work in the context of differential expression analysis (e.g. formulate a t-test in a linear model framework). Start with using limma to analyze example array dataset. Discuss assumptions of linear modeling. Transition to sequencing data. Introduce popular method for modeling RNA-seq data such as edgeR and DESeq (negative binomial) and voom-limma (normal with weight adjustment). Compare results to evaluate the differences and the consequences of not properly modeling the data.

3rd hr session: Statistical tests and concepts in genomics: What are we testing? What is the null hypothesis (under the linear model framework). Different ways of formulating the test (e.g. test modeling coefficient, nested linear models using likelihood ratio test to compare model fit). Effect size vs. significance. Multiple testing, p value, FDR, FWER (Bonferroni). Using John Storey q value method to estimate FDR and illustrate the idea behind multiple testing.

Conclude: Coming back to study design (Power, False positive, and False negative). Recap and emphasize why it is important to design the study beforehand, and explain how we can stride for publishing robust conclusions in the face of publication bias.


Homework: 40 points: Identify bias in an example dataset and adjust for it in DE tests
30 points: Use simulated data to illustrate how sample size affects the outcome of a study
30 points: Use simulated data to illustrate multiple testing issues and make a case for the necessity for adjustment
Module 2 (RNA): Single-cell RNA-seq using Seurat and Monocle

30 min background: scRNA-seq vs bulk RNA-seq; install Loupe Cell Browser; install R libraries (Seurat, Monocle, UMAP);

1st hr session: Seurat object; pre-processing (QC, normalization, scaling); cell type identification (doublet removal, marker genes)

2nd hr session: Plots (feature plot, dot plot, violin plot, volcano plot); differential gene expression

3rd hr session: Data transfer between Seurat and Monocle; Monocle trajectory (trajectory plot, branch heatmap, BEAM plot);

Course materials/online reading: scRNA-seq CellRanger output (control and Nkx2-1 mutant in PNAS; PMID: 31548395); unannotated code

Homework: 50 points: publication-quality figure (10 points each)
40 points: code annotation
10 points: customization (2 points each)
Module 3 (DNA): ChIP-seq and (sc)ATAC-seq

30 min background: ChIP RNA-seq vs RNA-seq; antibodies in immunostaining, FACS, co-IP, and ChIP (is an isotype control all you need?).

1st hr session: Sequence processing: trimming (what’s sequenced by Illumina: primers, barcodes); alignment (reference genome build); peak calling (MACS); merge replicates

2nd hr session: Visualization using deepTools or EA-seq; plots (coverage plot, profile plot, heatmaps); differential binding analysis

3rd hr session: ATAC-seq vs ChIP-seq; scATAC-seq with Signac (comparing to Seurat; cell type, plots, differential, motif).

Course materials/online reading: unannotated code

Homework: 50 points: publication-quality figure
            40 points: code annotation
            10 points: custom beautification (2 points each)
Module 4 (Protein): Mass spectrometry and protein structure

30 min background: Protein vs nucleic acid (no PCR; similarity vs identity; 3D structure); install PDB viewer.

1st hr session: Sequence alignment (score; global vs local; BLAST; phylogenetic tree); protein ID in mass spectrometry (enzyme; fingerprint)

2nd hr session: Application: crypto-peptides

3rd hr session: Identify and visualize protein motifs (helix-loop-helix, zinc finger, beta barrel, etc.)

Course materials/online reading:

Homework: 50 points: identify crypto-peptides (10 points each)
50 points: identify protein motifs (10 points each)
Module 5 (Image): ImageJ and Imaris

30 min background: Pixels; 8/16/32-bit; color; 3D; time-lapse. Install ImageJ/FIJI and Imaris viewer

1st hr session: FIJI basics: import, channel, color, level, selection, stack, plugin, macro, custom hot keys


3rd hr session: Imaris viewer: rendering, slicer, movie

Course materials/online reading:

Homework: 80 points: write a user’s instruction manual on how to count nuclei manually and automatically in FIJI, with screenshots for illustration

20 points: an Imaris movie
Module 6 (Freeware): GSEA and HOMER

**30 min background:** available freeware: GSEA, HOMER, Galaxy, Bioconductor etc.; criteria: function (input/output), interface, support.

**1st hr session:** GSEA: underlying statistics; validity of the gene sets (GO/IPA; cellular context); metagene analysis.

**2nd hr session:** Homer clustering for RNA-seq data.

**3rd hr session:** Homer motif analysis for ChIP-seq data.

**Course materials/online reading:**

**Homework:**
- 30 points: write a user’s instruction manual on how to use GSEA, with screenshots for illustration
- 35 points: write a user’s instruction manual on how to use HOMER clustering, with screenshots for illustration
- 35 points: write a user’s instruction manual on how to use HOMER motif analysis, with screenshots for illustration