BIMS 8382: Introduction to Biomedical Data Science

*Stephen D. Turner**

Spring 2018

Contents

BIMS8382 Overview

About BIMS8382

This course introduces methods, tools, and software for reproducibly managing, manipulating, analyzing, and visualizing large-scale biological data. Specifically, the course introduces the R statistical computing environment and packages for manipulating and visualizing high-dimensional data, covers strategies for reproducible research, and culminates with analyses of real experimental high-dimensional biological data.

What BIMS8382 is *not*

BIMS8382 *is not a statistics class.* BIMS8382 is a *data science* course. There is a session devoted to essential statistical analysis but this 3-hour lesson offers neither a comprehensive background on underlying theory nor in-depth coverage of implementation strategies using R. Some general knowledge of statistics and study design is helpful, but isn't required.

BIMS8382 is not a *"Tool X"* or *"Software Y"* class. Students should take away from this series the ability to use an extremely powerful scientific computing environment (R) to do many of the things that they will do *across study designs and disciplines* – managing, manipulating, visualizing, and analyzing large, sometimes high-dimensional data. This data might be gene expression data, microbial genomics data, public health/demographic data, RNA-seq data, or something not biologically related at all such as movie preference trends from Netflix, or truck routing data from FedEx. Regardless, the same computational know-how and data literacy to do the same kinds of basic tasks in each. BIMS8382 may feature specific tools here and there (DESeq2 for RNA-seq analysis, ggtree for drawing phylogenetic trees, etc.), but these are not important – the same specific software or methods will likely not be used 10 years from now, but underlying data and computational foundation will be. **That** is the goal of this course – to arm students with a basic foundation, and more importantly, to **enable students to figure out how to use *this tool* or *that tool* on their own**, when they need to.

*sdt5z@virginia.edu

Course Schedule

Week 1: Intro to R

This novice-level introduction is directed toward life scientists with little to no experience with statistical computing or bioinformatics. This interactive introduction will introduce the R statistical computing environment. The first part of this workshop will demonstrate very basic functionality in R, including functions, functions, vectors, creating variables, getting help, filtering, data frames, plotting, and reading/writing files.

Week 2: Advanced Data Manipulation with R

Data analysis involves a large amount of janitor work – munging and cleaning data to facilitate downstream data analysis. This session assumes a basic familiarity with R and covers tools and techniques for advanced data manipulation. It will cover data cleaning and "tidy data," and will introduce R packages that enable data manipulation, analysis, and visualization using split-apply-combine strategies. Upon completing this lesson, students will be able to use the *dplyr* package in R to effectively manipulate and conditionally compute summary statistics over subsets of a "big" dataset containing many observations.

Week 3: Advanced Data Visualization with R and ggplot2

This session will cover fundamental concepts for creating effective data visualization and will introduce tools and techniques for visualizing large, high-dimensional data using R. We will review fundamental concepts for visually displaying quantitative information, such as using series of small multiples, avoiding "chart-junk," and maximizing the data-ink ratio. After briefly covering data visualization using base R graphics, we will introduce the *ggplot2* package for advanced high-dimensional visualization. We will cover the grammar of graphics (geoms, aesthetics, stats, and faceting), and using ggplot2 to create plots layer-by-layer. Upon completing this lesson, students will be able to use R to explore a high-dimensional dataset by faceting and scaling arbitrarily complex plots in small multiples.

Week 4: Reproducible Research & Dynamic Documents

Contemporary life sciences research is plagued by reproducibility issues. This session covers some of the barriers to reproducible research and how to start to address some of those problems during the data management and analysis phases of the research life cycle. In this session we will cover using R and dynamic document generation with RMarkdown and RStudio to weave together reporting text with executable R code to automatically generate reports in the form of PDF, Word, or HTML documents.

Week 5: Essential Statistics

This session will provide hands-on instruction and exercises covering basic statistical analysis in R. This will cover descriptive statistics, t-tests, linear models, chi-square, clustering, dimensionality reduction, and resampling strategies. We will also cover methods for "tidying" model results for downstream visualization and summarization.

Week 6: Survival Analysis

This session will provide hands-on instruction and exercises covering survival analysis using R. The data for parts of this session will come from The Cancer Genome Atlas (TCGA), where we will also cover programmatic access to TCGA through Bioconductor.

Week 7: Introduction to RNA-seq Data Analysis

This session focuses on analyzing real data from a biological application - analyzing RNA-seq data for differentially expressed genes. This session provides an introduction to RNA-seq data analysis, involving reading in count data from an RNA-seq experiment, exploring the data using base R functions and then analysis with the DESeq2 Bioconductor package. The session will conclude with downstream pathway analysis and exploring the biological and functional context of the results.

Statistics topics

Topics covered in *Week 5: Essential Statistics*

- Descriptive statistics
  - Missing data
  - Exploratory data analysis (histograms, scatterplots, etc).
- Linear models
  - T-tests
  - Paired vs unpaired
  - One-tailed vs two-tailed
  - Wilcoxon / Mann-Whitney U-tests
  - Linear models
  - ANOVA
  - Linear regression
  - Multiple regression
- Categorical data analysis
  - Contingency tables
  - Chi-square tests
  - Fisher exact tests
  - Logistic regression
- Power and sample size analysis
  - Power/sample size for t-test
  - Power/sample size for proportion/chi-square test

Added statistics topics / discussion for Spring 2018

Batch effects

Batch effects are sources of technical variation introduced during an experiment, such as processing with different reagents, handling by a different technician, sequencing on a different flow cell, or processing samples in groups on different days. If these *batch effects* are strongly confounded with the study variable of interest, they can call into question the validity of your results, and in some cases, render collected data completely useless. The papers below discuss batch effects and how they can be mitigated.

1. **Chapter 5** of Scherer, Andreas. *Batch effects and noise in microarray experiments: sources and solutions.*
2. Leek, Jeffrey T., et al. "Tackling the widespread and critical impact of batch effects in high-throughput data." *Nature Reviews Genetics* 11.10 (2010): 733-739. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3880143/.

What's my n?

"What's my n" isn't always a straightforward question to answer, especially when it comes to cell culture expriments. The post and article below go into some of these details.

1. Statistics for Experimental Biologists: "What is 'n' in cell culture experiments?" Available at http://labstats.net/articles/cell_culture_n.html.
2. Vaux, David L., Fiona Fidler, and Geoff Cumming. "Replicates and repeats—what is the difference and is it significant?." *EMBO reports* 13.4 (2012): 291-296. Available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3321166/.

Technical versus biological replicates

Technical replicates involve taking multiple measurements on the same sample. Biological replicates are different samples each with separate measurements/assays. While technical replicates can help calibrate the precision of an instrument or assay, biological replicates are necessary for statistical analysis to make inferences about a condition or treatment. Read the paper and note below for more information on technical vs biological replication.

1. Blainey, P et al. "Points of significance: replication." *Nature methods* 11.9 (2014): 879-880.
2. Illumina Technical Note: "The Power of Replicates." Available at https://www.illumina.com/Documents/products/technotes/technote_power_replicates.pdf.

Topics covered in *Week 6: Survival Analysis with R*

- General overview of topics / definitions
  - Survival analysis
  - Hazard
  - Survival functions
  - Kaplan-Meier curves
  - Censoring / missing data
  - Proportional hazards
  - Cox regression
  - Hazard ratios
- Producing life tables
- Creating Kaplan-Meier Curves
- Fitting cox proportional hazards regression models
- Cagetorizing continuous variables for K-M curves
- Analyses of lung cancer datasets, colon cancer datasets
- Application / analysis of TCGA data

Week 7 / potential changes?

The final week switches gears completely and focuses on a very specific type of analysis with RNA-seq data. In the past there is very high variability in students' interest in this topic. There are several students in prior years' classes that have no interest in gene expression analysis, as they don't foresee it ever becoming a part of their research or lab work.

This week could potentially be replaced with a class on advanced ***predictive modeling***. Topics may include:

- Exploration of different classification / regression strategies (e.g., Naive Bayes, k-means clustering)
- Exploration of different data mining / machine learning algorithms (e.g., Random Forests, neural networks)
- Creating training and testing sets for model creation and validation
- Cross-validation
- Bootstrapping
- Permutation testing
- Feature extraction
- Others. . . ?