BIOM 610 Data Analysis for Life Science
Spring 2021

Course Co-Directors:

Daniel Beiting, Ph.D. Assistant Professor of Pathobiology School of Veterinary Medicine Rm 314 Hill Pavilion 380 S. University Avenue Phone: 215-898- 9247 Email: beiting@upenn.edu

Russell Takeshi Shinohara, Ph.D. Associate Professor of Biostatistics in Biostatistics and Epidemiology Center for Clinical Epidemiology and Biostatistics Department of Biostatistics, Epidemiology and Informatics Perelman School of Medicine 217 Blockley Hall 423 Guardian Drive Office: 215-746-1090 Email: russell.shinohara@pennmedicine.upenn.edu

Course Description:

Technological advances have transformed fields that rely on data by providing a wealth of information ready to be analyzed. From working with single genes to comparing entire genomes, biomedical research groups around the world are producing more data than they can handle and the ability to interpret this information is a key skill for any practitioner. The skills necessary to work with these massive datasets are in high demand, and this course will help you learn those skills. Using the open-source R programming language, you'll gain a nuanced understanding of the tools required to work with complex life sciences and genomics data. You'll learn the mathematical concepts — and the data analytics techniques — that you need to drive data- driven research. From a strong foundation in statistics to specialized R programming skills, this course will lead you through the data analytics landscape step-by-step. Taught by Rafael Irizarry from the Harvard T.H. Chan School of Public Health, and offered through the Harvard partnership with EdX.com, this four part course will enable new discoveries and will help you improve individual and population health. If you're working in the life sciences and want to learn how to analyze data, enroll now to take your research to the next level.
https://www.edx.org/professional-certificate/harvardx-data-analysis-for-life-sciences

The course will be 100% virtual and asynchronous (self-paced). All lecture videos will be available at the start of the course, and organizers will provide students with a recommended schedule for navigating the course content.

Laptop required. Students are evaluated upon completion of video lectures.

Enrollment expected to be 10 to 60 students. This course should be taken by 1st year PhD students from CAMB, IGG, NGG and PGG and MD/PhD and VMD/PhD candidates to fulfill the statistics rrequirement.

Course Structure:

Module 1: Statistics and R Random variables, Distributions, Inference (p-values and confidence intervals), Exploratory Data Analysis, and Non- parametric statistics
Module 2: Introduction to Linear Models and Matrix Algebra Learning objectives: Matrix algebra notation, Matrix algebra operations, Application of matrix algebra to data analysis, Linear models, and Brief introduction to the QR decomposition
Module 3: Statistical Inference and Modeling for High-throughput Experiments Learning objectives: Organizing high throughput data, Multiple comparison problem, Family Wide Error Rates, False Discovery Rate, Error Rate Control procedures, Bonferroni Correction, q-values, Statistical Modeling Hierarchical Models and the basics of Bayesian Statistics, and Exploratory Data Analysis for High throughput data
Module 4: High-Dimensional Data Analysis Learning objectives: Mathematical Distance, Dimension Reduction, Singular Value Decomposition and Principal Component Analysis, Multiple Dimensional Scaling Plots, Factor Analysis, Dealing with Batch Effects, Clustering, Heatmaps, and Basic Machine Learning Concepts

The course emphasizes the following core competencies: knowledge within program area (biostatistical method, statistical analysis in R, and high- dimensional data analysis and interpretation); computational methodologies (data analysis; programming and computing)

Through the use of scripting in an open-source programming language for statistics and plotting, students will learn how to create transparent and reproducible analyses.

After completion of the course, students should be able to develop a robust statistical plan to explore and analyze high-dimensional datasets. Specifically, the students should be able to produce statistical summaries from large tabular data, select the appropriate statistical methods for analysis of the data in the R programming language, and summarize, plot and interpret the results. In addition, students will learn how to read, interpret, and critically evaluate statistical concepts in the literature. Students will gain experience in designing and analyzing a research study, which will enhance several key competencies that are an important part of their PhD training.

The course will be 1-credit and will be self-directed, with students spending approximately 3 hours per week, for a total of 14 weeks, following the lecture and course materials for the Harvard/EdX certificate in Data Analysis for Life Sciences. This 'certificate' course consists of four separate modules, and will be offered every spring semester. Weekly opt-in sessions (held by Zoom) will be hosted by Graduate Student Teaching Assistants (TAs) from Epidemiology and Biostatistics and will provide a setting for course students to ask questions, get feedback, and discuss course videos in more detail. A course message board will be set-up via Slack and will be used as a platform for additional Q&A and interactions between students, TAs and Course Director