Course Description
This course will introduce graduate students to fundamental concepts surrounding the collection, analysis and interpretation of large biological datasets, and the challenges big data poses to both individual scientists and the scientific community. Using a combination of lectures, computational workshops, and discussions of current scientific papers, students will build the computational and analytical skills to examine large biological datasets, present their results, and critically evaluate the work of others. Lectures and analysis will mostly focus on experiments that use next-generation sequencing such as genome sequencing, RNA-seq, and bar-seq, but will also highlight other major sources of big data such as high-throughput phenotyping. Methods covered will include sequencing analysis pipelines, significance tests, multiple hypothesis testing, and clustering. Work for the class will consist of weekly programming and literature assignments and a self-designed and self-directed final data analysis project. The course is mainly directed toward biology graduate students with a limited background in programming, statistics or data analysis. However, students with quantitative backgrounds and some biological knowledge may interested if the their aim is to gain a better understanding of the methods and concepts surrounding big biological data.

Aims of the Course. Our ability to collect biological data is growing at an exponential pace. Because of this, modern biologists must build a skill set to analyze and critically evaluate large data sets. The primary aim of this course is to lay a practical foundation to perform these critical tasks. Students will learn how to handle files and run programs from the unix command line, to perform data analysis and programming using the high level programming language R, and to perform statistics appropriate for large biological datasets. These skills will be applied in hands-on workshops and weekly assignments throughout the course, and in individual work towards a final project. A second aim is to introduce students to the various ways in which big data is being generated, their advantages and their limitations. Included will be a discussions surrounding large multi-lab data collection projects. Third, this class aims to prepare students to understand and critically evaluate primary literature that includes big data. This will be accomplished through a mix of journal club discussions and re-analysis of primary literature. Lastly, this course aims is to improve presentation skills by having students lead discussions of journal articles and present final projects.

Credits 2

Grading Scheme
30% class participation / article presentation
30% class projects / assignments
40% final project