THE OHIO STATE UNIVERSITY

PHR 8194

Introduction to the Structure, Analyses and Interpretation of Genomic data studies

Background:

There is a need to train biomedical researchers to undertake cutting-edge cellular experiments that apply genomic-based approaches coupled with computational and statistical analsyes to reveal novel biological understanding.

Meeting this training need has several hurdles. Biomedical training frequently focuses on the historical competencies of molecular biology and biochemistry, whereas genomic analyses requires bioinformatic approaches that are often outside of training programs. Bioinformatic approaches leverages biological understanding, statistical insight and computational skills.

PHR 8194 was designed to develop bioinformatic skills and abilities in graduate students who have either little or no previous experience in computational science and statistics.

Goals and Design of the course:

The goal of this *2 credit course* is to form the first step in a student's development towards becoming genomically-literate. That is, the course assumes no prior knowledge and has no pre-requisites beyond those of the College of Pharmacy/College of Medicine graduate programs. Over the course, students will be introduced to concepts of graphical user interfaces (GUIs) and to the R platform for statistical computing. Each module will focus on specific concepts and contain a lecture and practical computer time. *There are no computational requirements, other than to bring a laptop*.

The modules build to a student-driven individual analyses in the final weeks where each student presents the analyses of a separate publicly available genomic data set. The analyses is an opportunity to apply and integrate everything learnt in the previous modules.

Finally, having undertaken this course it is envisaged that the students will be able to then attend more specialist and advanced courses, either in the Dept of Biomedical Informatics at OSU, or nationally, for example at Cold Spring Harbor.

Importantly, this course is appropriate for students studying all aspects of biology, botany, zoology, microbiology, physiology, biochemistry, veterinary biosciences and related subjects.

<u>The Instructors</u>:

Dr. Lara E. Sucheston-Campbell. Dr. Sucheston-Campbell is an Associate Professor in the College of Pharmacy and Veterinary Medicine.

She graduated with her PhD in Genetic and Molecular Epidemiology from Case Western Reserve University and was recruited to the faculty at Roswell Park Comprehensive Cancer Center in in 2007, and then to OSU in 2016.

As a genetic epidemiologist, her research is focused on identifying, quantifying and characterizing the relationships between exposures and genomic variation that modifies susceptibility to complex phenotypes. For the last several years she has explored gene exposure relationships at a candidate gene and genome wide level in hematological malignancies, ovarian, prostate and breast cancer as well as body composition phenotypes. She has successfully proposed to assess these relationships in large well-defined cohorts and acts as both a PI and Co-I on NIH funded grants. She has authored and co-authored over 100 scientific papers and serves on peer-review journal editorial and statistical advisory boards, as well as NIH and international grant review committees.

sucheston-campbell.1@osu.edu

Dr. Moray J. Campbell. Dr. Campbell is a Research Associate Professor in the College of Pharmacy at OSU.

He received his PhD in microbiology from the University of Kent, UK, and undertook post-doctoral training at Cedars-Sinai Medical Center/UCLA, Los Angeles, before joining the faculty of the University of Birmingham, UK in 2000. In 2007 he was recruited to Roswell Park Comprehensive Cancer Center where in 2013 he also became an Associate Dean to lead the re-design of the graduate curricula, before moving in 2016 to OSU. He undertook training in computer science and statistics through the Bioinformatics MS degree program at Johns Hopkins University, graduating in 2016.

His research addresses cancer-driver epigenetic mechanisms in prostate and breast cancer, for which he is funded by the DoD and NIH. He has published over 100 papers, reviews and book chapters, and serves on peer-review panels for the Department of Defense Congressional Directed Medical Research Program, and the NIH and international grant review committees.

LESC and ***MJC*** will oversee and lead teaching on the course, and be present for all lectures. Further support will be provided from the graduate students and post-docs working in their groups, as required.

campbell.1933@OSU.edu

<u>**Core competencies and learning outcomes in PHR 8194:**</u>
(abridged from Wilson Sayresa et al. PMID: 29870542)

1. Explain the role of computation and data mining in hypothesis-driven and hypothesis-generating questions within the life sciences.

- Compare and contrast computer-based research with wet-lab research.
- Exploit public databases to identify gene targets in disease and therapy.

2. Summarize key computational concepts, such as algorithms and relational databases, and their applications in the life sciences.

- Explain how data are organized in relational databases (e.g., NCBI databases).

3. Apply statistical concepts used in bioinformatics.

- Understand that there is a probability of finding altered gene expression by chance (the P value) and that the size of the target database affects the probability (the E-value).
- Explain the statistical modeling used in common genomics approaches including differential gene expression and genetic variation.

4. Use bioinformatics tools to examine important areas of biology.

- Using multiple lines of evidence to annotate a gene.
- Appreciate how gene mutations alter function and relate to disease using a genome browser and tools such as OMIM and KEGG to place it in the context of a function and pathway important to the disease.

C5. Find, retrieve, and organize various types of biological data.

- Navigate and retrieve data from a genome browser.
- Retrieve data from protein and genome databases (e.g., GEO, PDB, NCBI).

C6. Explore and model biological networks, and data integration using bioinformatics.

- Appreciate techniques used to generate analyses and interpret the outputs to generate novel hypotheses.
- Analyze gene expression data to build an expression network.

C7. Develop awareness and expertise of command-line bioinformatics tools.

- Build and run statistical analyses using environments such as R.

C8. Describe and manage biological data types, structure, and reproducibility.

- Develop an awareness of, and ability to, manipulate different data types given the versioning of formats.
- Appreciate the various data formats used to store genomic data (e.g., FASTA)
- Compare reproducibility of biological and technical replicate data (e.g., transcriptomic data) using statistical tests (Spearman rank test and false discovery calculations).

C9. Interpret the ethical, legal, medical, and social implications of biological data. The increasing scale and penetration of human genomic data has greatly enhanced the ability to identify disease-related loci, druggable targets, etc. and to identify potential genes for

replacement with developing techniques. However, with this information also comes many ethical, legal, and social questions; suggested resolutions are often outpaced by the technological advances. Students should debate the medicinal, societal, and ethical implications of these information sets and techniques.

- Explain the implications, good and bad, of being able to walk into a doctor's office and have your genome sequenced and analyzed or of being able to obtain information from direct-to-consumer testing services.
- Be able to discuss different perspectives about who should have access to this data and how it should be protected.
- Describe how the scientific community protects against the falsification or manipulation of large datasets.

Course Overview

Module 1: Introduction to R statistical computing
Module 2: Analyses of gene expression; candidate and global
Module 3: Browsing the genome
Module 4: Genetic variation and genomic sequencing
Module 5: Final project presentations