**BIOS 6660: Analysis of Genomic Data using R and Bioconductor**
**Spring 2019 (3 CREDITS)**

**Instructors:** Weiming Zhang, PhD Research Instructor
Pamela Russell, MA Research Instructor
Lauren Vanderlinden, MS Research Instructor
**E-mail:** weiming.zhang@ucdenver.edu
pamela.russell@ucdenver.edu
lauren.vanderlinden@ucdenver.edu
**Office:**
Lauren: Building 406 room 202
Pam: Building 406 room 107
Weiming: Building 500 room C3010
**Lecture Time:** Tue & Thurs, 10:30 - 11:50 am
**Location:** Ed 2 South L28-1308
**Office Hours:**
Pam: Fridays 10:00-12:00 (Jan 25 - Feb 22)
Lauren: TBD
Weiming: TBD

**Prerequisites:** BIOS 6602 or **Corequisite** BIOS 6612, or equivalent graduate level statistics course with consent of instructor

**Course Description:** This course provides students with hands on experience in processing and analyzing RNA-seq, genetic and Epigenomic data using R and command line tools. The course will emphasize reproducible research principles, effective data management, and effective coding principles.

**Course Objectives:** At the completion of BIOS6660 students will be able to:
1. Navigate the Unix/Linux command line and utilize command line programs
2. Use Git and GitHub for version control of scripts and code
3. Use the statistical software R to perform basic programming, data manipulation, and statistical tests
4. Organize a software or data analysis project and enforce quality through software testing
5. Manage the components of a data analysis workflow including data, code, and scripts
6. Apply reproducible analysis practices to a complete data analysis workflow, including producing end-to-end analysis reports with R Markdown
7. Understand the basic organization of the Bioconductor consortium and able to find appropriate packages for the analysis purposes
8. Navigate the S4 object oriented data structure used in Bioconductor in order to perform basic modification for custom analysis
9. Perform a complete high-throughput data analysis routine; from problem formulation, data processing, data filtering, statistical analysis and result interpretation and presentation
10. Perform clustering, comparative, and predictive algorithm for high-throughput data
11. Understand the basic concept of sequencing technology (both RNA-Seq and ChIP-Seq) and develop analysis solutions for these large datasets
12. Be familiarized with publicly available databases such as GEO and SRA
13. Validate Omics studies via publicly available data

14. Understand the basic concept of DNA Methylation and develop analysis solutions for these large datasets
15. Understand the concept and terminology of genetic association study and the concept of Mendelian randomization.
16. Comprehend real life biological problems and create analysis pipelines to fulfill the requirement of the project
17. Communicate and interpret the analysis findings to biological scientists using non-statistical terms

**Course Goals:** After completion of the course, students will be able to apply various statistical tools in a Linux environment to handle large Omics datasets and conduct reproducible research; students will be able to explore the nature of the dataset, understand the challenge of each type of Omics data, and effectively perform data preprocessing, normalization, batch correction and data analysis. More importantly, students will see both the uniqueness and the common problems these different types of Omics data have. The course will prepare them to face future new data types coming out from the advancement of biotechnology. Students will be able to perform a complete data analysis project from start to finish, including communicating and understanding the important questions to be answered, exploring the nature of the dataset, hypothesis generation, interpreting the biological meaning of the results, and learning to work with participating scientists to fulfill the research requirement. More importantly, students will learn to work with real researchers to perform biological discovery. These objectives will be accomplished using the free open-source statistical software R and Bioconductor. Throughout the course, an emphasis will be placed on effective data analysis practices, code and data management, and reproducibility.

**Competencies mapped to this course**

This course partially or fully addresses the following MS core knowledge and competencies, and is used for assessing achievement:

| **Identifier** | **NEW MS-BIOS Competencies** |
|---|---|
| MS-BIOS 2 | Apply statistical concepts of basic study designs including bias, confounding and efficiency, and identify strengths and weaknesses of experimental and observational designs. |
| MS-BIOS 3 | Carry out exploratory and descriptive analyses of complex data using standard statistical software and methods of data summary and visualization. |
| MS-BIOS 4 | Carry out valid and efficient modeling, estimation, model checking and inference using standard statistical methods and software. |

| | |
|---|---|
| MS-BIOS 5 | Demonstrate statistical programming proficiency, good coding style and use of reproducible research principles using leading statistical software. |
| MS-BIOS 6 | Demonstrate basic skills necessary for collaborating with non-biostatistical scientists, including mapping study aims to testable hypotheses, carrying out basic power and sample size estimation and evaluation, and identifying appropriate design, modeling and analysis methods to address study hypotheses. |
| MS-BIOS 7 | Communicate, orally and in writing, simple and complex statistical ideas, methods and results in non-technical terms appropriate for collaborator needs (e.g. preparation of analysis section of grant proposals and methods and results sections of manuscripts). |

**Required Text:** No text required.

**Software Use:** We will be using the open-source R software (www.r-project.org). The software is platform independent and will run on most major computer operating systems. Students will have access to a linux server where most analyses will take place. Various sequencing, processing and analysis packages which will be run on linux include: Trim Galore!, RSEM, hisat2, stringtie, and MACS2.

**Course Website:** Updated syllabus, homeworks, readings and other documents will be posted on the Canvas class website: https://ucdenver.instructure.com/login

**Required Work:**
- Homework assignments: 12 reports due throughout the semester

**Evaluation:**
- Report 1: 6.7%
- Report 2: 6.7%
- Report 3: 6.7%
- Report 4: 6.7%
- Report 5: 6.7%
- Report 6: 8.4%
- Report 7: 8.4%
- Report 8: 8.4%
- Report 9: 8.4%
- Report 10: 11%
- Report 11: 11%
- Report 12: 11%

**Academic Honor and Conduct Code:** All students are expected to abide to the honor code of the Colorado School of Public Health. Unless otherwise instructed, all of your work in this

course should represent completely independent work. Students are expected to familiarize themselves with the Student Honor Code that can be found at http://www.ucdenver.edu/academics/colleges/PublicHealth/resourcesfor/currentstudents/academics/Documents/PoliciesHandbooks/CSPH_Honor_Code.pdf
Any student found to have committed acts of misconduct (including, but not limited to cheating, plagiarism, misconduct of research, and breach of confidentiality) will be subject to the procedures outlined in the Honor Code.

**Student Code of Conduct:** Adherence to the Student Code of Conduct is expected: http://www.ucdenver.edu/life/services/standards/Documents/CUDenver-CodeofConduct.pdf

**Disability Accessibility Statement:** The University of Colorado Denver is committed to providing reasonable accommodation and access to programs and services for students with disabilities. For students requesting accommodations, you will need to contact the Office of Disability Resources and Services (DRS). Their staff will assist in determining reasonable accommodations as well as coordinating the approved accommodations. The office is B500, Room Q20-EG305 and open Mon-Fri 8am-5pm (phone 303-724-5640, email sherry.holden@ucdenver.edu).

**Other Class Policies:** Please turn off (or set to vibrate) mobile phones during class. Attendance will not be taken. Participation is not required but is recommended.

## Course Schedule *(Subject to Revision)*

| Date | Homework | Topic | Instructor |
|---|---|---|---|
| Jan 22 | R1 out | Analysis horror stories<br>Course overview | PR |
| Jan 24 | | Intro to the command line<br>Version control with Git and GitHub | PR |
| Jan 29 | R1 due; R2 out | Intro to R | PR |
| Jan 31 | | R continued | PR |
| Feb 5 | R2 due; R3 out | R ecosystem<br>R Markdown | PR |
| Feb 7 | | Intermediate R: Tidyverse<br>Code organization | PR |
| Feb 12 | R3 due; R4 out | Code quality | PR |
| Feb 14 | | Data management | PR |
| Feb 19 | R4 due; R5 out | Reproducibility 1 | PR |
| Feb 21 | | Reproducibility 2 | PR |
| Feb 26 | R5 due; R6 out | RNA-Seq Preprocessing 1 | LV |
| Feb 28 | | RNA-Seq Preprocessing 2 | LV |
| March 5 | | RNA-Seq Data Analysis 1 | LV |
| March 7 | R6 due | RNA-Seq Data Analysis 2 | LV |
| March 12 | R7 out | RNA-Seq Data Analysis 3 | LV |
| March 14 | | ChIP-Seq Data Analysis 1 | LV |
| March 19 | | SPRING BREAK | |
| March 21 | | SPRING BREAK | |
| March 26 | R7 due, R8 out | ChIP-Seq Data Analysis 2 | LV |
| March 28 | | Public Databases and Validation 1 | LV |
| April 2 | R8 due, R9 out | Public Databases and Validation 2 | LV |
| April 4 | | Introduction to Data  Integration | LV |

| | | | |
|---|---|---|---|
| April 9 | R9 due, R10 out | Introduction to DNA Methylation | WZ |
| April 11 | | DNA Methylation Data Cleaning, Normalization, and Batch Correction 1 | WZ |
| April 16 | | DNA Methylation Data Cleaning, Normalization, and Batch Correction 2, and Troubleshooting | WZ |
| April 18 | R10 due, R11 out | DNA Methylation Data Analysis 1 | WZ |
| April 23 | | DNA Methylation Data Analysis 2 | WZ |
| April 25 | | DNA Methylation Data Analysis 3, and Troubleshooting | WZ |
| April 30 | R 11 due, R12 out | Basic of Genetic Data: Concept and Terminology | WZ |
| May 2 | | Genetic Association Study Data Analysis: Common Variants | WZ |
| May 7 | | Genetic Association Study Data Analysis: Rare Variants | WZ |
| May 9 | R 12 due | Overview of Mendelian Randomization | WZ |