

BIOL 6397-02 & MATH 6397-04

Genomic Data Analysis

Spring 2019 Syllabus

This course provides an introduction to analyzing genomic data to answer biological questions. The class has a mixed enrollment of Biology and Biochemistry (BIOL) PhD students and students in the Masters in Statistics and Data Science (MSDS) program. The course will be a combination of lectures by the instructor, classroom discussions, demonstrations of data analysis, and hands-on exercises where you analyze publicly available data. You will also do homework assignments and a final project.

Instructor: Dr. Richard Meisel, rmeisel@uh.edu, SR2 421C

Co-instructor: Dr. Binod Manandhar, BINOD@math.uh.edu, PGH 214

Meeting times:

Monday & Wednesday, 2:30-4:00 PM in SR2 129

Every other Wednesday, MSDS students enrolled in MATH 6397 will meet in CBB 118

Readings:

There is no textbook for this course. Instead, we will read review articles and primary literature that describes how genomic data are used to answer biological questions. These readings will be from the perspectives of biologists using genomic approaches as well as statisticians and computer scientists who are developing tools for analyzing data. Readings will be assigned to correspond to lectures, discussions, and activities that we will be doing in class. All of these readings are linked in the schedule below and will be posted to blackboard as PDFs.

In addition to the assigned readings for each class period, there two references that are recommended to ensure students are familiar with the computational and biological material covered in this course (both are available as PDFs or links to online resources on blackboard):

1. [A Primer for Computational Biology](#) by Shawn O'Neil – introduction to Unix and R
2. [Genomes, 2nd edition](#) by Terence A Brown – introduction to genome biology

Schedule:

- The course is organized into seven blocks that each last ~2 weeks. First, we will spend 2 weeks on introductory material. There will then be five modules that will each last 2 weeks. In the seventh block you will work on your final project, present the project to the class, and produce a report on your final project.
- The introductory block and all five modules will include some class meetings that are held jointly with the MSDS students and BIOL students, and one class meeting that is held separately. In the modules, the first three class meetings are held together in SR2 129. The fourth class meeting of each module (every other Wednesday) is held separately, with the MSDS students meeting in CBB 118. The separate meetings are indicated in **Bold Blue** in the schedule below.

Date	Topic
	Introductory Material

Jan 14	Introduction to Genomics NHGRI – A Guide to Your Genome Del Giacco & Cattaneo – “ Introduction to Genomics ”
Jan 16	Introduction to Unix and R O’Neil – A Primer for Computation Biology
Jan 21	NO CLASS – MLK HOLIDAY
Jan 23	Working in Unix and R
Module 1: Using RNA-Seq to Study Gene Expression	
Jan 28	Read mapping Trapnell & Salzberg (2009) “ How to map billions of short reads onto genomes ” Li (2013) “ Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM ”
Jan 30	Quantifying gene expression Bray et al. (2016) “ Near-optimal probabilistic RNA-seq quantification ” Patro et al. (2017) “ Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference ”
Feb 4	Differential gene expression Love et al. (2014) “ Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 ” Pimentel et al. (2017) “ Differential analysis of RNA-seq incorporating quantification uncertainty ”
Feb 6	RNA-seq exercise Love et al. (2018) “ Analyzing RNA-seq data with DESeq2 ”
Module 2: DNA-Protein Interactions and Chromatin	
Feb 11	Introduction to DNA-protein interactions and chromatin ENCODE Project Consortium (2012) “ An integrated encyclopedia of DNA elements in the human genome ”
Feb 13	ChIP-seq, ATAC-seq, etc. Kharchenko et al. (2011) “ Comprehensive analysis of the chromatin landscape in Drosophila ” Buenrostro et al. (2013) “ Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position ”
Feb 18	ChIP-seq data analysis Zhang et al. (2008) “ Model-based analysis of ChIP-Seq (MACS) ”
Feb 20	ChIP-seq exercise Guertin (2015) “ Computational analysis of ChIP-seq and RNA-seq data ”
Module 3: Genetic Variation and Genome-Wide Association Studies	
Feb 25	Genetic variation in populations The 1000 Genomes Project Consortium (2012) “ An integrated map of genetic variation from 1,092 human genomes ”
Feb 27	Inferring sequence variation McKenna et al. (2010) “ The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data ” Poplin et al. (2018) “ Scaling accurate genetic variant discovery to tens of thousands of samples ”
Mar 4	Genome-wide association (GWA) King et al. (2012) “ Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource ”
Mar 6	GWA exercise

	GATK Best Practices DSPR Tutorial
Mar 11 & 13 NO CLASS – SPRING BREAK	
Module 4: Microbiomes and Metagenomics	
Mar 18	Microbiomes The Human Microbiome Project Consortium (2012) " Structure, function and diversity of the healthy human microbiome " Sunagawa et al. (2015) " Structure and function of the global ocean microbiome "
Mar 20	Microbial profiling and metagenomics Schloss et al. (2009) " Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities " Caporaso et al. (2010) " QIIME allows analysis of high-throughput community sequencing data "
Mar 25	Genomic data from microbes Falony et al. (2016) " Population-level analysis of gut microbiome variation " Zhernakova et al. (2016) " Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity "
Mar 27	Microbial genomics exercise Quince et al. (2017) " Shotgun metagenomics, from sampling to analysis " Kozich et al. (2013) " Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform "
Module 5: Long Read Sequencing, Genome Assembly, and Annotation	
Apr 1	Long read sequencing Miller et al. (2018) " Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing "
Apr 3	Genome assembly Bickhart et al. (2017) " Single-molecule sequencing and chromatin conformation capture enable de novo reference assembly of the domestic goat genome "
Apr 8	Transcriptomes and annotation Cantarel et al (2008) " MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes " Wang et al. (2016) " Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing "
Apr 10	Long read sequence exercise Koren et al. (2017) " Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation "
Final Project Preparation	
Apr 15-22	Final project discussions
Apr 24-29	Final Project presentations

The above schedule is subject to change. If this is necessary, reasonable notice will be given.

Grading:

- Your course grade will be out of 100 points determined by **attendance**, six **bi-weekly assignments**, and a **final project**. Attendance will be worth 10 points (10% of your class grade). Each of the six assignments will be worth 10 points (the sum of all six assignments will be worth 60% of your grade). The remaining 30% of your grade will be based on your final project.

- You are expected to attend every class meeting. You are allowed to miss one class meeting unexcused. Any additional missed classes without a pre-approved excuse will result in a 1-point deduction from your attendance grade (i.e., each class meeting is worth 1% of your final grade).
- The six **bi-weekly assignments** will be associated with every fourth class meeting where we do exercises (in **Bold Blue** in the schedule above). You will submit all assignments to blackboard. The assignments will be due one week after the associated class meeting. One point out of 10 will be deducted for each day after the due date that an assignment is turned in late.
- Your **final project** will consist of a proposal, data analysis, a report, and a presentation. Your grade on the final project will depend on all four components. More information about the **final project** will be provided later in the semester.
- Your letter grade in the course will be based on the following scale:

A	> 93	C	73 – 77
A–	90 – 93	C–	70 – 73
B+	87 – 90	D+	67 – 70
B	83 – 87	D	63 – 67
B–	80 – 83	D–	60 – 53
C+	77 – 80	F	< 60

This scale may be adjusted depending on the average performance of students in the course. Your letter grade will not be lower than the scale above, but you may receive a higher letter grade than the scale.