MMG 3320: Applications in Bioinformatics, 3 Credit Hours

Course Instructor: Dr. Princess Rodriguez
Office: Given Courtyard S453
Email: Princess.Rodriguez@med.uvm.edu
Phone: 802-656-1718
Tuesday/Thursday – 10:05-11:20, L/L CM 216

Course Description:
This course is intended for students in the biological sciences who have already completed the introductory course, Survey to Bioinformatic Databases. The advent of next-generation sequencing (NGS) platforms has made sequencing more accessible, faster, and less expensive, resulting in over 150,000 publicly available datasets that can be utilized to support research findings or testing hypotheses. In this course, students will learn and execute key steps in the bioinformatic workflow by processing a publicly available genomics dataset. By the end of the course, students will have accessed, processed, analyzed, visualized, and interpreted an NGS dataset of their choosing. The course covers several topics, including an introduction to UNIX, data processing, R programming basics (including data frames, cleaning, and fundamentals with ggplot2), as well as an in-depth overview of selected genomics analysis. Throughout the course, best practices for reproducible data and data management will be emphasized. The course uses a direct, hands-on approach, as such most classes are interactive and require student participation.

Course Learning Objectives:
By the end of the course, students will:

1. Develop an understanding of advanced genomics-based bioinformatic techniques and their applications in biological research,
2. Independently access and download publicly available genomic data from NCBI databases,
3. Gain proficiency in programming languages such as UNIX and R to automate bioinformatic workflows and manipulate data,
4. Explore the principals and techniques of data visualization and analysis,
5. Utilize new features and versions of computational programs to analyze genomics datasets as they become available,
6. Use online bioinformatics resources and programs to solve technical errors and determine data quality,
7. Interpret bioinformatic data generated and communicate results effectively to draw sound biological conclusions.

Prerequisites:
MMG 3310: Survey to Bioinformatic Databases

Course Resources:
There is no textbook required for this course. This course will be coordinated through Brightspace https://brightspace.uvm.edu/d2l/login. Brightspace will provide students with the course syllabus, assignments, grades, and links to supplemental web sites required for the course.

The course website(s) contain scripts, homework prompts, and lectures. Please note that materials will become available week-by-week.

- https://prodriguez19.github.io/Intro-to-shell/

A laptop computer will be required to participate in this course. In addition, each student will be provided with an account on the Vermont Advanced Computing Cluster (VACC). Students will use VACC - Open OnDemand (OOD) https://vacc-ondemand.uvm.edu/pun/sys/dashboard to access their VACC account, command line, and R/RStudio.

Course Outline:
This course is structured into four units which can broadly be described as:

- Unit 1 (Weeks 1 – 3): Introduction to UNIX
- Unit 2 (Weeks 4 – 7): Alignment to genome, file conversion, pre- and post-processed QC
- Unit 3 (Weeks 9): Variant analysis with GATK
- Unit 4: Week 10 – 12) : Advanced analysis and visualization with R
- Unit 5 (Weeks 13): Metagenomics analysis

Week 1: Jan 16th and 18th
Introduction to command line
Students will gain an understanding of the basic skills required to use the command line interface, including how to log into to the Vermont Advanced Computing Core (VACC), a high-performance computing cluster.

Week 2: Jan 23rd and 25th
Navigating and working with files
In a file system, files are organized into directories or folders which can be further divided into subdirectories. This week students will be introduced to the file system, how to navigate the file system, as well as, manipulating files and directories will be covered.

Week 3: Jan 30th and Feb 1st
Writing, searching, and creating shell scripts
An overview of text editors (`vim, nano`), searching and working with text files, pipe operators, environment variables, and creating shell scripts will be covered.

Week 4: Feb 6th and 8th
Considerations when selecting an NGS dataset to analyze from GEO
Experimental planning and considerations when selecting an NGS dataset to analyze will be covered. In addition, how to access a publicly available dataset from NCBI Gene Expression Omnibus (GEO) and job submissions using SLURM, an open-source workload manager designed for LINUX clusters will be described.

In addition, an overview of chromatin structure and function, histones and their modifications, and epigenetic regulation of gene expression will be provided. An emphasis will be on understanding techniques used to study these data (ChIP-Seq and RNA-Seq). **Paper discussion**

Week 5: Feb 13th and 15th

NGS Data Management & Preprocessing and Quality Control of sequence reads

Overview of best practices with NGS data management and popular tools used for transferring data (Filezilla) will be taught. Description of file formats commonly used in NGS (FASTQ, SAM/BAM) as well as FASTQC, a popular open-source tool for quality control of high-throughput sequencing data will be covered.

Genome Indexing and Alignment

Introduction to genome indexing and alignment algorithms. Provide an overview of tools (STAR, HiSAT2, BWA-mem) commonly used. Hands-on exercises in genome indexing and alignment will be performed during class. Post-alignment processing to remove low-quality reads, adapter sequences, and contaminants will also be covered.

Week 6: Feb 20th and 22nd

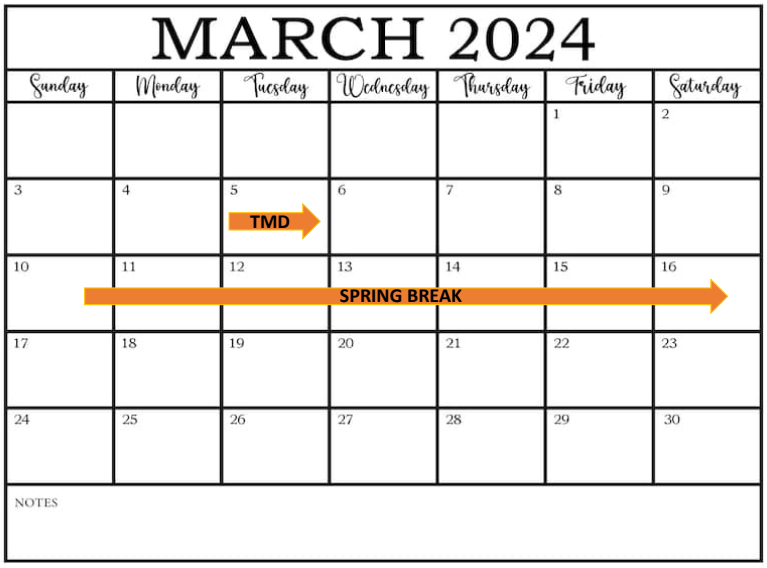
RNA-Seq: BAM to counts files

This week we will process BAM files using `HTSeq`. `HTSeq` is a python package that can be used to count the number of reads that overlap with features such as genes or exons.

Week 7: Feb 27th and 29th

ChIP-Seq: BAM to narrowPeaks files

This week we will process BAM files using `MACS2`. `MACS2` is a widely used tool for the identification of binding sites of transcription factors and histone modifications.



Week 8: Thursday, March 7th

The R basics: A Review

Students would have previously been exposed to R programming in the prerequisite course, MMG3310. This week will be a review of R syntax (comments, variables, functions, arguments), and data wrangling (data structures, vectors, data frames).

Week 9: March 19th and 21st

Variant analysis with GATK

This week are going to run GATK (Genome Analysis Toolkit). GATK was created by the Broad Institute for variant analysis and genotyping primarily for the human genome. However, it certainly can be used for any genome. **Paper discussion** followed by analysis of dataset to follow.

Week 10: March 26th and 28th

Differential Analysis and ggplot2

Includes normalization, identification of differentially expressed genes, interpretation, and batch effect removal with `DESeq2`. Overview of the grammar and how `ggplot2` is implemented, the building blocks of `ggplot2` plot.

Week 11: April 2nd and 4th

ggplot2 continued

How to create plots including scatterplots, bar charts, and box plots will be covered. Students will come to understand aesthetics, scales, and coordinate systems of `ggplot2`.

Week 12: April 9th and 11th

Pathway analysis with R

Introduction to functional analysis and its importance in interpreting biological data. Overview of Gene Ontology (GO) database and gene set enrichment analysis (GSEA) and R packages `clusterprofiler, pheatmap, biomaRt, fgsea, enrichplot`, and `DOSE` will be covered. We will use practical examples with real-world datasets to discuss best practices and common pitfalls of functional analysis with R.

Week 13: April 16th and 18th

Metagenomics analysis

This week we will be carrying out a common metagenomics workflow - identifying Operational Taxonomic Unities (OTUs) among samples taken from two metagenomes within a location. We will be starting with a set of sequenced reads (.fastq files), perform some quality control steps, assemble those reads, and finish by identifying and visualizing the OTUs among these samples. **Paper discussion** overview led by graduate student team will be followed by analysis of dataset.

Week 14: April 23rd and 25th

No class this week: Virtual 1-on-1

We will use this week to "catch-up" on any analysis required for presentation. Students will be required to check-in with Dr. Rodriguez for a 30-min 1-on-1 on Microsoft Teams. Sign-up sheet with times will be made available one week prior.

Week 15: April 30th and May 2nd

Final Project presentations

Students will present their findings in the form of a 15-min PowerPoint presentation on a publicly available dataset processed. Only students presenting are required to attend class and participate.

Grading Criteria:
Below is a breakdown of the overall grading. Collectively, homework assignments will account for 40% of the course grade. A total of 25% of the course grade will be assigned based on the final project. The purpose of this final project is for the student to access, analyze, and explore a publicly available genomics dataset. Students will also be graded on primary paper reflections, class participation and attendance for select classes. There is no final exam for this course.

Homework assignments	40%
Primary Literature Summaries and participation	35%
Final Project	25%
TOTAL	100%

Assessments (Graded Work):
Below are descriptions of the assessments that will be used to measure the students understanding of course concepts.

Homework assignments (40% of grade):
There will be a total of eleven homework assignments. These homework assignments can be categorized as ***mini*** and *take-home based* assignments.

- The ***mini*** homework assignments are developed to take students anywhere from 10 - 45 minutes to complete. Students will be granted time during class to start the homework but are still given an additional students 48 hours after assigned to complete and submit the homework. No late work will be accepted when it comes to mini-homework assignments. Specific guidelines in the form of a 20 to 50-point grade sheet will be provided to students.
- The *take-home based* homework assignments are developed to take students anywhere from 1 - 4 hours to complete. As a result, students are granted one week to complete and turn in the assignment. Each homework is designed to aid students prepare for the final project. Late homework be accepted but will be docked 10% of the overall grade for every day that the assignment is late. An assignment is considered late if it is not submitted by the time and due date specified. Three days past the due date (weekend included), the assignment will no longer be accepted, and the student will receive a ZERO. Specific guidelines in the form of a 100 to 150-point grade sheet will be provided to students one week before the assignment is due.

Primary Literature Summary and Participation (35%):
Students will be asked to summarize and evaluate up to four primary research articles. In the prerequisite course MMG3310, students were provided a comprehensive overview on how to perform a search on NCBI PUBMED, as well as strategies on how to summarize peer-reviewed literature. Building on these skills, students will select and comprehensively summarize an article that features an NGS dataset for their final project for one primary literature assignment. For the other paper summaries, students will only summarize select portions of the article selected by the instructor. Grading rubric in the form of a 50 to 100-point grade sheet will be provided to students.

In addition, a skill that requires training is scientific communication and evaluation of oral

presentations to express knowledge and understanding of a given topic. To ensure students are practicing these skills, students will be required to provide constructive feedback to their peers anonymously. These feedback forms will be available online for each student presenter. Students will also be asked to actively participate during all *paper discussion* sessions by asking questions, either at the end of the presentation or as the question arises during the presentation.

Final Project (25%):
One final project will be used to demonstrate an understanding of bioinformatics approaches acquired throughout the semester. The final project is an opportunity to apply tools using a real-world dataset, and then practice presenting this information to professors and peers in a formal scientific way. The self-selected trail will guide the downstream analysis performed by the student on their reanalysis of the publicly dataset selected.

Ski Trails: Ski trails are categorized by their level of difficulty, which are indicated by a color code system. This is used to signify to the skier the level of difficulty associated with the trail selected. Green trails are designed for beginners while blue trails are made for intermediate skiers with more technical skills, and finally black trails are designed for advanced skiers who are comfortable skiing at high speeds on challenging terrain. They often have the most obstacles and require a high level of technical skill to navigate.

Similarly, the final project for this course will mirror these ratings with the understanding that within this class there are varying levels of bioinformatic competency. Students will self-select the trail they would like to undertake during the first few weeks of the course. Detailed rubrics for each trail will be provided in the form of 100-point grade sheet one month prior presentations are scheduled. Below are the overall objectives for each trail.

- Green Trail: Replicating figures from a primary research article is a common exercise used to practice data generation. Here, the student will identify and then replicate one figure from the primary research article. Overall, the student will gain a better understanding of the data, identify potential errors or inconsistencies of the original analysis using software used by the authors to process and visualize the data, compare their results to the original figures, and draw biological conclusions.
- Blue Trail: A barrier to processing bioinformatic data is that there are many programs that can be used for a similar outcome in the pipeline. There are important considerations when selecting a short-read aligner, a program for read counting, or peak caller. Students who select the blue trail will be asked to modify select parameters within the bioinformatic pipeline and then compare and contrast how this impacts their overall results. Students who select this trail will be asked to comment on the overall usability, accuracy, speed, and complexity when testing these bioinformatic programs to understand which are best suited for the task at hand.
- Black Trail: When working with an NGS dataset, one common goal is to test an original hypothesis by analyzing the data and generating new insights. Students who select this trail will process and download an NGS dataset with this overall aim in mind. As such, the student will be asked to generate new and original insights that were not previously reported in the published body of work. The figures generated during the final

presentation should present new information that advances the understanding of the research question, rather than simply replicating existing figures.

Note: Regardless of the trail selected all students will download and process an NGS dataset, perform quality control, downstream analysis, and create outputs that will be presented to the class in the form of an oral presentation format using Microsoft PowerPoint. Each student will be evaluated on the overview of the topic given, experimental design and quality control performed, major findings and interpretation of plots generated with R/RStudio, grammar, format, overall delivery of topic.

Grading scale:
Your letter grade earned in the course will be based on the numerical ranges given below.

<60 = F	60 - 63 = D-	70 - 73 = C-	80 - 83 = B-	90 - 93 = A-
	64 - 66 = D	74 - 76 = C	84 - 86 = B	**94 - 96 = A**
	67 - 69 = D+	77 - 79 = C+	87 - 89 = B+	97 - 100 =A+

Decimals will be rounded up or down based on universal math conventions (.4 and below rounds down, .5 and above rounds up).

Citing your Sources:
You are required to appropriately cite your sources for your assignments. If you do not cite your sources, your assignments will automatically be marked 50% off. Plagiarism in any form is not acceptable. If assignments are plagiarized, the responsible student will be reported to the academic integrity office. Please see section titled Academic Integrity in this syllabus for more details about the University of Vermont's policies on cheating and plagiarism.

Office Hours:
Office hours are by appointment. I am happy to meet in-person or via zoom with 48-hour notice. Please feel free to email and ask questions or request meeting times if you need help with any homework assignment or project.

Email:
Please use proper etiquette when addressing your instructor. In the email subject please indicate the course in which you are enrolled and include all necessary information required to appropriately answer your question. My goal is to return emails within 48 hours but that is not always possible. If you send an email the night before an assignment is due, I cannot guarantee that we will be able to respond to your email promptly.

Student Learning Accommodations:
In keeping with University policy, any student with a documented disability interested in utilizing ADA accommodations should contact Student Accessibility Services (SAS), the office of Disability Services on campus for students. SAS works with students and faculty in an interactive process to explore reasonable and appropriate accommodations, which are communicated to

faculty in an accommodation letter. All students are strongly recommended to discuss with their faculty the accommodations they plan to use in each course. Faculty who receives Letters of Accommodation with Disability Related Flexible accommodations will need to fill out the Disability Related Flexibility Agreement. Any questions from faculty or students on the agreement should be directed to the SAS specialist who is indicated on the letter.

Contact SAS:
A170 Living/Learning Center;
802-656-7753
access@uvm.edu
www.uvm.edu/access

Important UVM Policies

Academic Integrity:
The Academic Integrity policy addresses plagiarism, fabrication, collusion, and cheating.

Code of Student Conduct:
UVM's Code of Student Conduct outlines conduct expectations as well as students' rights and responsibilities.

FERPA Rights Disclosure:
The purpose of UVM's FERPA Rights Disclosure is to communicate the rights of students regarding access to, and privacy of their student educational records as provided for in the Family Educational Rights and Privacy Act (FERPA) of 1974.

Grade Appeals:
If you would like to contest a grade, please follow the procedures outlined in this policy.

Grading:

This link offers information on grading and GPA calculation.

Religious Holidays:
Students have the right to practice the religion of their choice. If you need to miss class to observe a religious holiday, please submit the dates of your absence to me in writing by the end of the second full week of classes. You will be permitted to make up work within a mutually agreed-upon time. The complete policy is here.

Promoting Health & Safety:
The University of Vermont's number one priority is to support a healthy and safe community:
Center for Health and Wellbeing
Counseling & Psychiatry Services (CAPS) Direct Phone Line: (802) 656-3340
C.A.R.E. If you are concerned about a UVM community member or are concerned about a specific event, we encourage you to contact the Dean of Students Office (802-656-3380). If you would like to remain anonymous, you can report your concerns online by visiting the C.A.R.E. Team website.

Statement on Alcohol and Cannabis in the Academic Environment

As a faculty member, I want you to get the most you can out of this course. You play a crucial role in your education and in your readiness to learn and fully engage with the course material. It is important to note that alcohol and cannabis have no place in an academic environment. They can seriously impair your ability to learn and retain information not only in the moment you may be using, but up to 48 hours or more afterwards. In addition, alcohol and cannabis can:

- Cause issues with attention, memory and concentration
- Negatively impact the quality of how information is processed and ultimately stored
- Affect sleep patterns, which interferes with long-term memory formation

It is my expectation that you will do everything you can to optimize your learning and to fully participate in this course.