COMPUTATIONAL GENOMICS – BIOL 7210 A – Spring 2024

Instructor: ***Christopher Gulvik****, GTA:* ***Haojun Song****, GTA:* ***Jianshu Zhao****, and*
Professor: ***I. King Jordan***
cgulvik@gatech.edu *hsong343@gatech.edu* *jzhao399@gatech.edu*

Course Summary

The science of genomics involves the intersection of experimentation and computation. Computers are required to handle the massive amount of data produced by genome sequencing projects. More importantly however, genome sequencing efforts yield 'information' alone, which can only be converted into 'knowledge' via the application of computational genomics techniques.

In this class, the students will convert raw genomic information (i.e., nucleotide sequence reads) into knowledge through the use of computational genomics tools and applications. The class will be provided with unassembled (raw) genome sequence data and will proceed through four distinct stages of analysis and interpretation of those data:

1 - **read sequence cleaning and genome assembly ("Group 1")**,
2 - **gene prediction and annotation ("Group 2")**,
3 - **genotyping and taxonomic and quality assessments ("Group 3")**, and
4 - **comparative genomics ("Group 4")**.

This course will be entirely practical in nature. Students will learn to do the actual work of computational genomics. Expert guest lecturers will be brought in to provide information on state-of-the-art computational genomics tools. Based on this information, other class lectures and their own research, students will be solely responsible for choosing which tools (e.g., algorithms, programs, or databases) to use, how to implement them, and for generating and thoroughly documenting their final results. All results will be shared and documented.

Class lecture sessions will be held on Tuesdays and Thursdays from 8:00am to 9:15am in the Clough UG Learning Commons room 102. There is no textbook. Required and recommended readings will be made available on GATech's internal Canvas learning management system. Students are required to use online databases, public open source repositories, and the scientific literature to inform their choice of computational tools to be used. Since there is no textbook and many of the sessions involve class discussion and lab activities rather than formal lecture, attendance and class participation are absolutely mandatory.This course has several different websites that students will use for different purposes.

Canvas Learning Management System

- All lectures will be posted here in the Files → "Lectures" subsection
- Homework exercises must be submitted by students in the "Assignments" section here

- Class-wide discussion opportunities for each topic are available in the "Discussion" section here for students, GTAs, and instructors to ask for clarification, help, or share with others on the subject
- Any virtual guest lectures, done through Zoom, will have URLs posted for everyone to join in the "Zoom" section here. Video cameras are required to be on during the lecture for attendance and participation to count
- All grades will be posted in the "Grades" section here

Internal Course GitHub

- You must be on-site on a GATech network or off-site connect to GATech's VPN service to have access to this internal webpage
- Group projects will have their code and primary output files stored here within their own respective repositories
- Exercise homework assignments will be posted in MarkDown file format for viewing instructions and specific grade distributions for each assignment
- General and specific guidance after students present their initial group project plans will be posted in their respective repositories

Course Grading

Attendance and participation	**[class][individual]**	**10%**
Background & strategy presentations (2x 7.5%)	**[class][group]**	**15%**
Exercise homework assignments (4 x 5%)	**[home][individual]**	**20%**
Final presentations & results (2 x 20%)	**[home][group]**	**40%**
GitHub homework assignment	**[home][individual]**	**4%**
Software management homework assignment	**[home][individual]**	**5%**
Workflow homework assignments (2 x 3%)	**[home][group]**	**6%**

Attendance and Participation

Classroom attendance and participation are mandatory. Participation will be judged by the degree to which each student participates in class lectures and discussions (by asking questions, answering questions, offering ideas and opinions), during group presentations (by asking questions during others' presentations, by engaging the audience during their own presentation, by connecting their presentation to previous class discussions, by working successfully in a small group), and during computer laboratory activities (by performing analyses and working with other students). Students who show up late or miss class will lose 10% of their class participation grade each time.

Exercise Homework Assignments

Students will have four in-class exercise activities throughout the semester. Each exercise will have the topic introduced and a hands-on in-class series of exercises will be performed by each student independently. GTAs will provide some technical assistance to students, however part of student participation is for student peer helping as well. Students will leave the class exercises with a general understanding of the topic and experience with basic and intermediate level application handling. The exercise homework assignments will follow on this, by having students expand their understanding of each topic through independent reading to accomplish an advanced level of application handling outside of class. For all four exercises, the results as well as documentation of how the results were generated are required to be submitted to Canvas.

Teams and Groups

The class will be split up into large *Teams*, where each individual team must accomplish four main tasks throughout the semester on a specific outbreak dataset. Each team will receive a different dataset from an ongoing outbreak investigation. Intra-team collaboration is a core component to this course and essential, but students are also encouraged to share ideas and discuss with other teams as well. Each team is required to create and agree upon a contract document, and it must include: data management practices, communication methods, meeting schedules, decision-making processes, internal roles, conflict handling, internal deadlines, and how grading should be distributed. Teams will be referenced by numbers (1-6).

Within each team, four individual *Groups* will be formed to accomplish a specific task. Each group will give a series presentations and laboratories/demos. **Group presentations and labs/demos** will be judged by the depth of analysis presented, the clarity of presentation, the utility of the exercises, the appropriateness and justification of the choices made, the validity and robustness of the results and the thoroughness of the documentation. All student code and analysis contributions must be shared and documented on GitHub – https://github.gatech.edu/comgenomics2024. Specific requirements for the presentations will be provided during class sessions. Contributions of each individual student to the overall group effort must be meticulously detailed and documented. Groups will be referenced by letters (A-D).

Computing Environment

This class requires the Linux command line. You will be required to install the Linux command line. There are a few ways you can do this:

- MacOS: already has the command line
- Ubuntu Linux: already has the command line
- Windows: WSL2 must be installed and configured from the Microsoft app store

COVID-19 Accommodations

As per University System of Georgia policies, instruction for this course will be face-to-face. In-person attendance is mandatory. We will accommodate students who need to quarantine due to COVID-19 or students who become ill for any reason. Lecture slides will be distributed via Canvas.

Additional Resources

Computational Genomics is a cutting edge scientific field and accordingly, it is difficult to find a single centralized resource or book. Therefore, we elect to provide you with a several different resources in various media formats, and some might be more valuable to you than others based on your preferred learning style (e.g., textbooks, video recordings, webpages). Other readings may be provided week by week.

Bioinformatics

- Bioinformatics and Functional Genomics, third edition by Jonathan Pevsner, 2015. Wiley-Blackwell. ISBN-13: 978-1118581780
- Rob's Computational Genomics Manual website and accompanying YouTube recorded lectures by Dr. Rob Edwards.
- Conner's Pathogen Genomics Course website by Dr. Conner Meehan.
- Ben's Computational Genomics website and accompanying YouTube recorded lectures by Dr. Ben Langmead.

Clear Communication

- Smart Brevity: The Power of Saying More with Less by Jim VandeHei, Mike Allen, and Roy Schwartz, 2022. New York, NY: Workman Publishing Co, Inc. ISBN-13: 978-1523516971
- Even a Geek Can Speak by Joey Asher, 2006. Persuasive Speaker Press. ISBN-13: 978-0978577605.
- The CDC Clear Communication Index (CCI) tool – CCI is a field-tested, validated tool that helps you assess use of clear communication criteria in your materials
- Making Data Talk by the National Cancer Institute

General Career and Learning

- A PhD Is Not Enough: A Guide To Survival In Science by Peter J. Feibelman, 1993. New York, NY: Basic Books. ISBN-13: 978-0465022229
- The Making of an Expert by K. Anders Ericsson, Michael J. Prietula, and Edward T. Cokely. 2007. Harvard Business Review 85(7-8):114-121.
- Active learning increases student performance in science, engineering, and mathematics by Scott Freeman, Sarah L. Eddy, Miles McDonough, Michelle K. Smith, Nnadozie Okoroafor, Hannah Jordt, and Mary Pat Wenderoth. 2014. Proc Natl Acad Sci U S A 111(23):8410-8415. DOI: 10.1073/pnas.1319030111.

Health Equity

- Health Equity Guiding Principles for Inclusive Communication website
- Global Public Health Equity Guiding Principles for Communication website

Jargon Reduction Tools

- Everyday Words for Public Health Communication – website database of public health terms and their more familiar, everyday alternatives
- Environmental Health Thesaurus website

Class	Week	Date	Day	Time (EST)	Topic or *Task Deadline*	People
1	1	01/09/2024	Tue	8:00 - 9:15	Introduction, Logistics & Teams	Christopher Gulvik
2	1	01/11/2024	Thu	8:00 - 9:15	Code Management with GitHub & Group Assignments	Christopher Gulvik
3	2	01/16/2022	Tue	8:00 - 9:15	Software Management with Conda	Christopher Gulvik
	2	*01/16/2022*	*Tue*	*23:55*	*Due: GitHub Exercise*	*All Students*
4	2	01/18/2024	Thu	8:00 - 9:15	Software Management with Docker	Frank Ambrosio & Curtis Kapsak
	2	*01/18/2022*	*Thu*	*9:15*	*Due: Group Contracts Exercise*	*All Students*
5	3	01/23/2024	Tue	8:00 - 9:15	Read Sequence Cleaning & Genome Assembly Concepts	Christopher Gulvik
	3	*01/23/2024*	*Tue*	*23:55*	*Due: Software Management Exercise*	*All Students*
	3	*01/23/2024*	*Tue*	*23:55*	*Grades: GitHub Exercise*	*Haojun Song & Jianshu Zhao*
6	3	01/25/2024	Thu	8:00 - 9:15	Read Sequence Cleaning & Genome Assembly Exercises	Haojun Song & Jianshu Zhao
7	**4**	**01/30/2024**	**Tue**	**8:00 - 9:15**	**Read Sequence Cleaning & Genome Assembly Background & Strategy**	**Students – (1) Groups**
	4	*01/30/2024*	*Tue*	*23:55*	*Grades: Software Management Exercise*	*Haojun Song & Jianshu Zhao*
	4	*01/30/2024*	*Tue*	*23:55*	*Due: Read Sequence Cleaning & Genome Assembly Exercise*	*All Students*
8	4	02/01/2024	Thu	8:00 - 9:15	Workflow Big Data Processing in Nextflow	Christopher Gulvik (virtual)
	4	*02/01/2024*	*Thu*	*23:55*	*Grades: Read Sequence Cleaning & Genome Assembly Exercise*	*Haojun Song & Jianshu Zhao*
9	5	02/06/2024	Tue	8:00 - 9:15	Workflows with Nextflow (continued)	Christopher Gulvik (virtual)
10	5	02/08/2024	Thu	8:00 - 9:15	CDC's Advanced Molecular Detection (AMD) Program and Public Health	Dhwani Batra & Kristen Knipe (virtual)
11	6	02/13/2024	Tue	8:00 - 9:15	Gene Prediction & Annotation Concepts	Christopher Gulvik
12	6	02/15/2024	Thu	8:00 - 9:15	Gene Prediction & Annotation Exercises	Haojun Song & Jianshu Zhao
13	**7**	**02/20/2024**	**Tue**	**8:00 - 9:15**	**Read Sequence Cleaning & Genome Assembly, Results**	**Students – (1) Groups**
	7	*02/20/2024*	*Tue*	*23:55*	*Due: Workflow #1 Exercise*	*All Students*
14	**7**	**02/22/2024**	**Thu**	**8:00 - 9:15**	**Gene Prediction & Annotation, Background & Strategy**	**Students – (2) Groups**
	7	*02/22/2024*	*Thu*	*23:55*	*Due: Gene Prediction & Annotation Exercise*	*All Students*
15	8	02/27/2024	Tue	8:00 - 9:15	Genotyping and Taxonomic and Quality Assessments Concepts	Christopher Gulvik
	8	*02/27/2024*	*Tue*	*23:55*	*Grades: Workflow #1 Exercise*	*Christopher Gulvik*
16	8	02/29/2024	Thu	8:00 - 9:15	Genotyping and Taxonomic and Quality Assessments Exercises	Haojun Song & Jianshu Zhao
	8	*02/29/2024*	*Thu*	*23:55*	*Grades: Gene Prediction & Annotation Exercise*	*Haojun Song & Jianshu Zhao*
17	9	03/05/2024	Tue	8:00 - 9:15	Sequences in the Wind: How to use genomics to mitigate a viral pandemic	Benjamin Rambo-Martin & Kristine Lacek
18	9	03/07/2024	Thu	8:00 - 9:15	Open Lab Session	*All Students*
	9	*03/07/2024*	*Thu*	*23:55*	*Due: Genotyping and Taxonomic and Quality Assessments Exercise*	*All Students*
19	**10**	**03/12/2024**	**Tue**	**8:00 - 9:15**	**Genotyping and Taxonomic and Quality Assessments, Background & Strategy**	**Students – (3) Groups**
20	**10**	**03/14/2024**	**Thu**	**8:00 - 9:15**	**Gene Prediction & Annotation, Results**	**Students – (2) Groups**
	10	*03/14/2024*	*Thu*	*23:55*	*Grades: Genotyping and Taxonomic and Quality Assessments Exercise*	*Haojun Song & Jianshu Zhao*
	11	03/19/2024	Tue		Spring Break	
	11	03/21/2024	Thu		Spring Break	

21	12	03/26/2024	Tue	8:00 - 9:15	Modern approaches to genomic epidemiology & Investigating foodborne outbreaks with genomic epidemiology	Lee Katz (virtual)
22	12	03/28/2024	Thu	8:00 - 9:15	Phylogenetics and Advanced Molecular Evolution Techniques and Applications	Jessica Chen (virtual)
23	**13**	**04/02/2024**	**Tue**	**8:00 - 9:15**	**Genotyping and Taxonomic and Quality Assessments, Results**	**Students – (3) Groups**
24	13	04/04/2024	Thu	8:00 - 9:15	Comparative Genomics Concept	Christopher Gulvik
25	14	04/09/2024	Tue	8:00 - 9:15	Comparative Genomics Exercise	Haojun Song & Jianshu Zhao
26	14	04/11/2024	Thu	8:00 - 9:15	Data wrangling, summarizing, and visualizing big data for data-driven conclusions	TBD
	14	*04/11/2024*	*Thu*	*23:55*	*Due: Comparative Genomics Exercise*	*All Students*
27	**15**	**04/16/2024**	**Tue**	**8:00 - 9:15**	**Comparative Genomics, Background & Strategy**	**Students – (4) Groups**
28	15	04/18/2024	Thu	8:00 - 9:15	Investigating foodborne outbreaks with genomic epidemiology	Sung Im (virtual)
	15	*04/18/2024*	*Thu*	*23:55*	*Due: Workflow #2 Exercise*	*All Students*
	15	*04/18/2024*	*Thu*	*23:55*	*Grades: Comparative Genomics Exercise*	*Haojun Song & Jianshu Zhao*
29	**16**	**04/23/2024**	**Tue**	**8:00 - 9:15**	**Comparative Genomics Results**	**Students – (4) Groups**
30	16	04/25/2024	Thu	8:00 - 9:15	Open Lab Session	All Students
	16	*04/25/2024*	*Thu*	*23:55*	*Grades: Workflow #2 Exercise*	*Christopher Gulvik*
	17	04/30/2024	Tue		No Class	
	17	05/02/2024	Thu		No Class; End of Term	
	18	*05/06/2024*	*Mon*		*Due: Grade Submission Deadline*	*Christopher Gulvik*
	18	*05/07/2024*	*Tue*		*Due: Grade Submission Deadline*	*All Students*

Content = In-person, student learning opportunity

Content = In-person, student-guided presentation

Content = deadline

Content = online only virtual Zoom meeting