University of Pittsburgh
School of Public Health
Department of Human Genetics

HUGEN 2071
Genomic Data Processing and Structures

Fall 2023

Wednesdays 1:30PM-2:50PM
122 Victoria Building

Fridays 1:30–2:50 PM
A719 Public Health-Crabtree

3 Credits

COURSE DESCRIPTION

Bioinformatics involves an in-depth understanding of data and a substantial amount of data processing. This course focuses on the manipulation and management of human genetic and genomic data via two platforms: the R statistical computing environment and the Unix operating system. The course will also cover the major data formats and structures used to store human genetic and genomic data. A key component of the course will be hands-on analyses of example data sets in a variety of formats.

COURSE GOALS

Upon completion of this course, the student will be able to:

- Write efficient R code to manipulate data.
- Write efficient Unix scripts to manipulate data.
- Describe common data structures used for the storage of human genetics data.
- Create data files in common data structures used for the storage of human genetics data.

COURSE PREREQUISITES

HUGEN 2022 · Human Population Genetics
BIOST 2041 · Introduction to Statistical Methods 1

or approval of the instructor

Students should also have basic computing and programming skills.

FACULTY

Daniel E. Weeks, Ph.D.
3117 Public Health
412-624-5388
weeks@pitt.edu
Office Hours available upon request

Jon Chernus, Ph.D.
3124 Public Health
Phone: 412-648-2562
jmc108@pitt.edu
Office hours available upon request

CANVAS AND GITHUB CLASSROOM INSTRUCTION

This course will extensively use the University's Canvas site (canvas.pitt.edu) and GitHub Classroom (classroom.github.com). To login to Canvas, you must have a Pitt account. Each lecture will be accompanied by supporting material and further reading, all of which will be made available around the time of the lecture. It is the student's responsibility to check for, and read, this material. Discussion topics related to the course may also be posted on Canvas, and, for the purpose of determining a student's grade, participation in these discussions will be considered as equivalent to participation in class discussion. The instructors will use Canvas as the primary means of communicating with the students, who are expected to check the site on a regular basis throughout the semester. GitHub Classroom will complement Canvas by providing a mechanism for distributing and submitting coding assignments and coding projects, as well as enabling collaborative team coding assignments using state-of-the-art source code control systems. Using GitHub Classroom will also enable students to learn to use versioning in their coding and to code within a reproducible research paradigm.

Class attendance is expected. Though we plan to record each class via Zoom, recordings will be made available on an individual basis only, when class is missed due to illness or another valid reason. Of course, if you are required to isolate or quarantine, become sick, or are unable to come to class, contact us as soon as possible to discuss arrangements.

Accessibility

Ensuring an accessible and pleasant experience to all users, regardless of disability, is a key focus of Canvas. The Canvas platform was built using the most modern HTML and CSS technologies, and is committed to W3C's Web Accessibility Initiative and §508 guidelines. GitHub Classroom's compliance with §508 guidelines can be found at government.github.com/accessibility/.

HEALTH AND SAFETY

It is important that you abide by Pitt's Health Rules. These rules have been developed to protect the health and safety of all of us. If you do not comply, you will be asked to leave class. It is your responsibility have the required face covering when entering a university building or classroom if Pitt's rules require it. For the most up-to-date information and guidance, please visit the Power of Pitt site.

If you are required to isolate or quarantine, become sick, or are unable to come to class, contact us as soon as possible to discuss arrangements. Do not come to class if you are sick, may be sick, or should isolate/quarantine.

EVALUATION AND GRADING

Evaluation will be based on the following components:

Online Syllabus Review

There will be one (1) online syllabus review. It will consist of a series of questions posed online through a Google Form about the syllabus.

The online syllabus review is passed by correctly answering 80% of its questions. Retaking the syllabus review does not use mulligans (described in the late policy below).

Feedback Questions and Comments

At the end of each class this term, we will ask you to submit one question or comment on the class content. These questions or comments demonstrate engagement with and aid you in keeping up with the course material.

Please submit only questions that you wish to have answered. If you do not have such a question, please instead submit a comment about what you learned or how we might better help you learn.

Feedback questions and comments will be due by 10 PM the evening after class. Credit will be awarded for those questions and comments that show good faith engagement with the class material. Comments can include describing what you learned that was new or summaries of what you think was most valuable or useful from the class or associated active learning activities.

To earn an A in the class you must submit one question or comment for 90% or more of the times required. To earn a B, ≥ 80% of the times; to earn a C, ≥ 70% of the times; to earn a D, ≥ 60% of the times.

Homework Assignments

There will be approximately twelve (12) homework assignments during the term. Each homework assignment will ask you to complete coding tasks that reflect what is being learned in class.

Homework assignments will typically be due twenty-four (24) hours before a class. A late assignment receives 0 points, unless you use a late policy mulligan, described in the assignment late policy below, for one (1) twenty-four (24) hour extension. An assignment submitted after the extension receives 0 points.

There are limited opportunities to make up a poorly completed assignment, described in the make-up assignment policy below.

We will be discussing solutions to the homework assignments in class, so it is essential that they are submitted on time.

To earn an A in the class you must pass 90% or more of the tasks across all of the homework assignments. To earn a B, ≥ 80% of the tasks; to earn a C, ≥ 70% of the tasks; to earn a D, ≥ 60% of the tasks.

Projects

A common but not unchallenging task in working with large genetic data sets is cleaning and preparing them for deposition in the National Institutes of Health Database of Genotypes and Phenotypes (dbGaP). All NIH-funded projects that generate large-scale data must place such data into dbGaP once the data are cleaned.

Midterm Project · Processing and Structures for Phenotypes

The midterm project of this course asks the students to prepare mock phenotype data for deposition into dbGaP. This includes merging, cleaning, and validating the data, as well as creating a data dictionary that describes the data.

Final Project · Processing and Structures for Genotypes

The final project of this course asks the students to prepare mock genotype data for deposition into dbGaP. This includes merging, cleaning, and validating genotypes and placing them in the format required for submission to dbGaP.

The specifications for achieving A, B, C, and D level work on the projects will be provided when the projects go live later in the term.

Both projects must be turned in on time. Think of these hard deadlines as being similar to inflexible grant deadlines that are frequently encountered in research. If exigent circumstances arise that require extensions or exceptions beyond this policy, please contact Drs. Weeks or Chernus at your earliest opportunity.

Late Policy Mulligans

We expect homework assignments to be handed in on time, so that we can freely discuss the solutions to the homework assignments in class.

Each student begins the term with three (3) late policy "mulligans." A late policy mulligan can be used for a single 24-hour extension on a homework assignment (extending the due-date to immediately before class rather than 24 hours before).

No late policy mulligans are necessary for the syllabus review.

No late policy mulligans can be used for the engagement questions and comments.

If exigent circumstances arise that require extensions or exceptions beyond this policy, please contact Drs. Weeks or Chernus at your earliest opportunity.

Make-Up Assignment Policy Mulligans

Each student begins the term with two (2) make-up assignment "mulligans." A mulligan can be used to submit a make-up assignment to replace a poorly completed homework assignment.

The make-up assignments will consist of reviewing missed assignment questions, what the objective behind the question was, writing a description of what went wrong when trying to answer the question, and the creation of a new homework question and answer that could achieve the same objective. It is due within 72 hours after the graded homework has been returned to you.

No make-up assignment policy mulligans are necessary for the syllabus review.

No make-up assignment policy mulligans can be used for the engagement questions and comments.

If exigent circumstances arise that require extensions or exceptions beyond this policy, please contact Drs. Weeks or Chernus at your earliest opportunity.

Grading

The grade for the class is determined by meeting all of the requirement for that particular grade given below:

Assessment	Earn D	Earn C	Earn B	Earn A
Syllabus quiz	✓	✓	✓	✓
Engagement questions	60%	70%	80%	90%
Homework assignments	60%	70%	80%	90%
Midterm project	*	**	***	****
Final project	*	**	***	****

* D-level work on the project
** C-level work on the project
*** B-level work on the project
**** A-level work on the project

To be specified upon assignment of the projects.

You must meet all thresholds to earn a grade, i.e., to earn an A you must achieve A-level work on every assessment across the class. Passing the syllabus quiz, submitting engagement questions for 85% of the required times, achieving 91% across the homework assignments, and completing A-level work on both projects will earn only a B, because the engagement in the course was at the B level.

SCHEDULE

Note that class-specific learning objectives, active learning links, and suggested readings are available in our online HuGen 2017 book at:

https://danieleweeks.github.io/HuGen2071/Readings.html

As these may be adjusted/updated throughout the term, you should consult it on a regular basis.

Class 1: 8/30/2023 Wed: Instructor: Jon Chernus	**Introduction and Overview**
Class 2: 9/1/2023 Fri: Instructor: Dan Weeks	**GitHub**
Class 3: 9/6/2023 Wed: Instructor: Dan Weeks	**R: Basics**
Class 4: 9/8/2023 Fri: Instructor: Dan Weeks	**R: Factors, Dates, Subscripting**
Class 5: 9/13/2023 Wed: Instructor: Dan Weeks	**R: Character Manipulation**
Class 6: 9/15/2023 Fri: Instructor: Dan Weeks	**R: Loops and Flow Control**
Class 7: 9/20/2023 Wed: Instructor: Dan Weeks	**R: Functions and Packages, Debugging R**
Class 8: 9/22/2023 Fri: Instructor: Dan Weeks	**R: Tidyverse**

Class 9: 9/27/2023 Wed: Instructor: Dan Weeks	**R: Recoding and Reshaping Data**
Class 10: 9/29/2023 Fri: Instructor: Dan Weeks	**R: Merging Data**
Class 11: 10/4/2023 Wed: Instructor: Dan Weeks	**R: Traditional Graphics & Advanced Graphics**
10/6/2023 Fri:	**No class - Fall Break**
Class 12: 10/11/2023 Wed: Instructor: Dan Weeks	**R: Exploratory Data Analysis**
Class 13: 10/13/2023 Fri: Instructor: Dan Weeks	**R: Interactive and Dynamic Graphics**
Class 14: 10/18/2023 Wed: Instructor: Dan Weeks	**Data Quality Checking and Filters**
Class 15: 10/20/2023 Fri: Instructor: Jon Chernus	**Unix: Basics, Streams, Redirection, & Pipe**
Class 16: 10/25/2023 Wed: Instructor: Jon Chernus	**Unix: Interacting with Processes, Cluster Jobs, Shell Scripting**
Class 17: 10/27/2023 Fri: Instructor: Ryan Minster	**Genetic Data Structures**
Class 18: 11/1/2023 Wed: Instructor: Ryan Minster	**PLINK I**
11/3/2023 Fri:	**No class - ASHG**
Class 19: 11/8/2023 Wed: Instructor: Ryan Minster	**PLINK II**
Class 20: 11/10/2023 Fri: Instructor: Jon Chernus and Dan Weeks	**PLINK Computer Lab**
Class 21: 11/15/2023 Wed: Instructor: Jon Chernus	**Unix: Data Manipulation**

Class 22: 11/17/2023 Fri: Instructor: Jon Chernus	**Unix: Pipes & Parallelization**
11/22/2023 Wed:	**No class - Thanksgiving**
11/24/2023 Fri:	**No class - Thanksgiving**
Class 23: 11/29/2023 Wed: Instructor: Jon Chernus	**Unix: Scripting, Control Structures and Variables**
Class 24: 12/1/2023 Fri: Instructor: Ryan Minster	**VCF, bcftools, vcftools**
Class 25: 12/6/2023 Wed: Instructor: Ryan Minster	**SAM & samtools**
Class 26: 12/8/2023 Fri: Instructor: Jon Chernus	**Genetic Data in R, GDS**
Class 27: 12/13/2023 Wed: Instructor: Jon Chernus	**Help with Final Project**
12/15/2023 Fri:	**No class**

COURSE MATERIALS

Online HuGen 2071 Course Book

This is a work in progress:

https://danieleweeks.github.io/HuGen2071/

Required Computer

As in many classes we will be doing interactive computer exercises, please bring a laptop capable of running RStudio and R to class.

Required Software (All available free online)

Web Browser

R *r-project.org*	R Studio *rstudio.com*
Global Protect *link*	GitHub Classroom *classroom.github.com*

Suggested Readings (All available free online thru the University of Pittsburgh)

To access materials, go to hsls.pitt.edu/remote and follow the instructions under "Remote access tip for Pitt users." The bookmarklet there is one of the easiest ways to quickly access materials for which Pitt has current subscriptions and that are available to you as a member of the University of Pittsburgh.

Bioinformatics for Geneticists
Editor: Michael R. Barnes
DOI: 10.1002/9780470059180
Web access: *onlinelibrary.wiley.com/book/10.1002/9780470059180*

Bioinformatics Data Skills
Author: Vince Buffalo
Publisher: O'Reilly 2015
Web access: *www.oreilly.com/library/view/bioinformatics-data-skills/9781449367480/?ar*

Data Manipulation with R
Author: Spector, Phil.
Publisher: New York: Springer, c. 2008.
Web access: *ebookcentral.proquest.com/lib/pitt-ebooks/detail.action?docID=371639#*

ggplot2: Elegant Graphics for Data Analysis
Author: Wickham, Hadley
Publisher: New York: Springer Aug. 2009
Web access: *doi.org/10.1007/978-0-387-98141-3*

Supplemental Readings/Bibliography
(Optional, All available free online thru the University of Pittsburgh)

Introductory Statistics with R
Author: Dalgaard, Peter.
Publisher: New York: Springer, c. 2002.
Web access: *ebookcentral.proquest.com/lib/pitt-ebooks/detail.action?docID=3035502#*

Current Protocols in Bioinformatics

Editor: Baxevanis AD, Stein LD, Stormo GD, Yates JR
Publisher: John Wiley and Sons, Inc., c. 2017
DOI: 10.1002/0471250953
Web access: *onlinelibrary.wiley.com/book/10.1002/0471250953*

R Programming for Bioinformatics
Author: Robert Gentleman
Publisher: Boca Raton : CRC Press, c2009.
Web access: https://learning.oreilly.com/library/view/r-programming-for/9781420063677/

Bioinformatics and Computational Biology Solutions Using R and Bioconductor
Editors: Robert Gentleman *et al.*
Publisher: New York: Springer Science+Business Media, c. 2005.
Web access: *link.springer.com/book/10.1007%2F0-387-29362-0*

ACADEMIC POLICIES

Disability Services

If you have a disability for which you are or may be requesting an accommodation, you are encouraged to contact both your instructor and Disability Resources and Services (DRS), 140 William Pitt Union, (412) 648-7890, drsrecep@pitt.edu, (412) 228-5347 for P3 ASL users, as early as possible in the term. DRS will verify your disability and determine reasonable accommodations for this course.

Academic Integrity

Students in this course will be expected to comply with the University of Pittsburgh's Policy on Academic Integrity. Any student suspected of violating this obligation for any reason during the semester will be required to participate in the procedural process, initiated at the instructor level, as outlined in the University Guidelines on Academic Integrity. This may include, but is not limited to, the confiscation of the examination of any individual suspected of violating University Policy. Furthermore, no student may bring any unauthorized materials to an exam, including dictionaries and programmable calculators.

To learn more about Academic Integrity, visit the Academic Integrity Guide for an overview of the topic. For hands- on practice, complete the Academic Integrity Modules.

Plagiarism

University policy:

> Integrity of the academic process requires that credit be given where credit is due. Accordingly, it is unethical to present as one's own work the ideas, representations, words of another, or to permit another to present one's own work without customary and proper acknowledgement of sources.
>
> A student has an obligation to exhibit honesty and to respect the ethical standards of the profession in carrying out his or her academic assignments. Without limiting the application of this principle, a student may be found to have violated this obligation if he or she:

10. Presents as one's own, for academic evaluation, the ideas, representations, or words of another person or persons without customary and proper acknowledgment of sources.
11. Submits the work of another person in a manner which represents the work to be one's own.

Source

To avoid plagiarism, you must give "customary and proper acknowledgment of sources" by appropriately and clearly identifying which thoughts are yours and which are others, and appropriately citing your sources.

Sophisticated plagiarism detection software will be used in this course. If plagiarism is detected, you will automatically receive a grade of zero for that assignment and the incident will be reported, as required, to your Dean.

Use of Artificial Intelligence Tools

Large language models (LLMs) like ChatGPT or GitHub Copilot can help an experienced programmer write, explain, and debug code more efficiently. But they are not a substitute for learning basic programming skills, especially as their output often requires its own debugging. As it will be in the settings for which this class seeks to prepare you, here you are permitted to use LLMs in your work (*with citation*), but you are of course responsible for any and all of their errors. Note that you must acknowledge/cite usage of these tools when used.

Course Recording

This class or portions of this class will be recorded by the instructors for educational purposes. These recordings will be shared only with students enrolled in the course via Canvas. These recordings will reside in the cloud and should not be redistributed.

To ensure the free and open discussion of ideas, students may not record classroom lectures, discussions and/or activities without the advance written permission of the instructor, and any such recording properly approved in advance can be used solely for the student's own private use.

Copyright Notice

These materials may be protected by copyright. United States copyright law, 17 USC § 101, *et seq*., in addition to University policy and procedures, prohibit unauthorized duplication or retransmission of course materials. See Library of Congress Copyright Office and the University Copyright Policy.

Sexual Misconduct, Required Reporting, & Title IX

If you are experiencing sexual assault, sexual harassment, domestic violence, and stalking, please report it to me and I will connect you to University resources to support you.

University faculty and staff members are required to report all instances of sexual misconduct, including harassment and sexual violence to the Office of Civil Rights and Title IX. When a report is made, individuals can expect to be contacted by the Title IX Office with information about support resources and options related to safety, accommodations, process, and policy. I encourage you to use the services and resources that may be most helpful to you.

As your professor, I am required to report any incidents of sexual misconduct that are directly reported to me. You can also report directly to Office of Civil Rights and Title IX: 412-648-7860 (M-F; 8:30am-5:00pm) or via the Pitt Concern Connection at: Make A Report

An important exception to the reporting requirement exists for academic work. Disclosures about sexual misconduct that are shared as a relevant part of an academic project, classroom discussion, or course assignment, are not required to be disclosed to the University's Title IX office.

If you wish to make a confidential report, Pitt encourages you to reach out to these resources:

The University Counseling Center: 412-648-7930 (8:30 A.M. TO 5 P.M. M-F) and 412-648-7856 (AFTER BUSINESS HOURS)

Pittsburgh Action Against Rape (community resource): 1-866-363-7273 (24/7)

If you have an immediate safety concern, please contact the University of Pittsburgh Police, 412-624-2121

Any form of sexual harassment or violence will not be excused or tolerated at the University of Pittsburgh.

For additional information, please visit the full syllabus statement on the Office of Diversity, Equity, and Inclusion webpage.

From the Office of Diversity, Equity, and Inclusion

Equity, Diversity, and Inclusion

The University of Pittsburgh does not tolerate any form of discrimination, harassment, or retaliation based on disability, race, color, religion, national origin, ancestry, genetic information, marital status, familial status, sex, age, sexual orientation, veteran status or gender identity or other factors as stated in the University's Title IX policy. The University is committed to taking prompt action to end a hostile environment that interferes with the University's mission. For more information about policies, procedures, and practices, visit the Civil Rights & Title IX Compliance web page.

I ask that everyone in the class strive to help ensure that other members of this class can learn in a supportive and respectful environment. If there are instances of the aforementioned issues, please contact the Title IX Coordinator, by calling 412-648-7860, or e-mailing titleixcoordinator@pitt.edu. Reports can also be filed online. You may also choose to report this to a faculty/staff member; they are required to communicate this to the University's Office of Diversity and Inclusion. If you wish to maintain complete confidentiality, you may also contact the University Counseling Center (412-648-7930).