University of Pittsburgh
Graduate School of Public Health
Department of Human Genetics

HUGEN 2071
Genomic Data Processing and Structures

Fall 2021

Tuesdays 12:30–1:55 PM, A425 Public Health-Crabtree
Fridays 1:30–2:55 PM, 2121C Public Health

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.
Office: TBD
Phone: TBD
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.

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

During this pandemic, it is extremely important that you abide by the public health regulations , the University of Pittsburgh's health standards and guidelines, and Pitt's Health Rules. These rules have been developed to protect the health and safety of all of us. Universal face covering is required in all classrooms and in every building on campus, without exceptions, regardless of vaccination status. This means you must wear a face covering that properly covers your nose and mouth when you are in the classroom (and therefore eating and drinking in class are not allowed). 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. For the most up-to-date information and guidance, please visit coronavirus.pitt.edu and check your Pitt email for updates before each class. If you are required to isolate or quarantine, become sick, or are unable to come to class, contact the instructors as soon as possible to discuss arrangements.

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.

The initial deadline for submission of the online syllabus review is Friday Sept 3 at 1:30 PM. A failed syllabus review can be retaken as many times as is necessary to pass it until Wed Sep 8 at 11:00 AM. Retaking the syllabus review does not use mulligans (described in the late policy below).

Engagement Questions and Comments

For some classes this term, we will ask you to submit at least two questions or comments on the readings or videos prior to class. These questions or comments demonstrate engagement with and aid you in keeping up with the course material.

Engagement questions and comments will be due one (1) hour before class. Credit will be awarded for those questions and comments that show good faith engagement with the material. Comments can include describing what you learned that was new or summaries of what you think was most valuable or useful from the reading or video.

To earn an A in the class you must submit two questions or comments 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 ten (10) 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 be due twenty-four (24) hours before 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 Minster 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 Minster 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 Minster 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 assigned readings may be updated throughout the term.

8/27/2021 Fri: **Introduction and Overview**
Instructor: Jon Chernus

Required Reading:
Barnes (2007) Chapter 1

Carey MA, Papin JA. Ten simple rules for biologists learning to program. *PLoS Comput Biol*. 2018;14(1): e1005871.
doi.org/10.1371/journal.pcbi.1005871

Dudley JT, Butte AJ. A quick guide for developing effective bioinformatics programming skills. *PLoS Comput Biol*. 2009;5(12):e1000589.
doi.org/10.1371/journal.pcbi.1000589

Learning objectives:

Review the syllabus
Describe bioinformatics
List various type of data used in genetics

8/31/2021 Tue: **Genetic Information & dbGaP**
Instructor: Jon Chernus

Required Reading:

Mailman MD, Feolo M, Jin Y, *et al*. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*. 2007;39(10):1181-1186.
doi.org/10.1038/ng1007-1181

Tryka KA, Hao L, Sturcke A, *et al*. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res*. 2014;42(Database issue):D975-D979.
doi.org/10.1093/nar/gkt1211

Wong KM, Langlais K, Tobias GS, *et al*. The dbGaP data browser: a new tool for browsing dbGaP controlled-access genomic data. *Nucleic Acids Res*. 2017;45(D1):D819-D826.
doi.org/10.1093/nar/gkw1139

Learning objectives:

Describe what genetic information entails
Describe how genome-wide SNP data are generated
Enumerate the different technologies used in genotyping

9/3/2021 Fri: **GitHub**
Instructor: Dan Weeks

Required Reading:

Happy Git and GitHub for the useR. https://happygitwithr.com/

Perez-Riverol Y, Gatto L, Wang R, *et al*. Ten Simple Rules for Taking Advantage of Git and GitHub. *PLoS Comput Biol*. 2016;12(7):e1004947. doi.org/10.1371/journal.pcbi.1004947

Learning objectives:

To learn how to use GitHub
To learn how to use GitHub Classroom
To learn how to use GitHub within RStudio

9/7/2021 Tue: **R: Basics**
Instructor: Dan Weeks

Required Reading:
Spector (2008) Chapters 1 & 2; Buffalo (2015) Chapter 8 'R Language Basics', p. 175-193.

Learning objectives:

To become familiar with the R language and concepts
To learn how to read and write data with R
To learn control flow: choices and loops

9/10/2021 Fri: **R: Factors, Dates, Subscripting**
Instructor: Dan Weeks

Required Reading:
Spector (2008) Chapters 4, 5, 6

Learning objectives:

To learn how to handle factors and dates with R
To learn how to subset data with R
To learn how to manipulate characters with R

9/14/2021 Tue: **R: Character Manipulation**
Instructor: Dan Weeks

Required Reading:
Spector (2008) Chapter 7

Learning objectives:
To learn how to handle character data in R
To learn how to use regular expressions in R

9/17/2021 Fri: **R: Reproducible Research**
Instructor: Dan Weeks

Required Reading:
Gentleman. Reproducible research: a bioinformatics case study. Statistical applications in genetics and molecular biology (2005) vol. 4 pp. Article2
https://doi.org/10.2202/1544-6115.1034

Learning objectives:
To understand the concepts of reproducible research
To learn to use R Markdown

9/21/2021 Tue: **R: Functions and Packages, Debugging R**
Instructor: Dan Weeks

Required Reading:
Buffalo (2015) Chapter 8 'Debugging R Code', p. 236-237.

Learning objectives:
To learn how to write R functions and packages
To learn how to debug R code

9/24/2021 Fri: **R: Tidyverse**
Instructor: Dan Weeks

Required Reading:
Chapter 4 of "Statistical Inference via Data Science,
A moderndive into R and the tidyverse": https://moderndive.com/4-wrangling.html

Learning objectives:
To learn how to use the pipe operator
To learn how to use Tidyverse functions

9/28/2021 Tue: **R: Recoding and Reshaping Data**
Instructor: Dan Weeks

Required Reading:
Spector (2008) Chapters 8 & 9

Learning objectives:
To learn how to reformat and reshape data in R

10/1/2021 Fri: **Merging Data**
Instructor: Dan Weeks

Required Reading:
Spector (2008) Chapter 9; Buffalo (2015) Chapter 8 'Merging and Combining Data', p. 219-224.

Learning objectives:
To learn how to merge data using Unix and Python
To learn how to use the R 'merge' command

10/5/2021 Tue: **R: Traditional Graphics & Advanced Graphics**
Instructor: Dan Weeks

Required Reading:
Wickham (2009) Chapters 2 & 3

Learning objectives:
To learn the basic graphics commands of R
To learn the R graphing package ggplot2

10/8/2021 Fri: **R: Exploratory Data Analysis**
Instructor: Dan Weeks

Required Reading:
Buffalo (2015) Chapter 8 'Exploring Data Visually with ggplot2', p. 207-219.

Learning objectives:
To learn how to summarize data frames
To learn how to visualize missing data patterns
To learn how to visualize covariation

10/12/2021 Tue: **R: Interactive and Dynamic Graphics**
Instructor: Dan Weeks

Required Reading:
Wickham (2009) Chapters 2 & 3

Learning objectives:
To learn how to use interactive and dynamic graphics to explore your data more thoroughly
To learn to use iPlots and Ggobi
To learn to use plotly

10/15/2021 Fri: **No class - Fall Break**

10/19/2021 Tue: **No class -ASHG**

10/22/2021 Fri: **Data Quality Checking and Filters**
Instructor: Dan Weeks

Learning objectives:
To learn how to check genotype data for quality

10/26/2021 Tue: **PLINK & PLINK Format**
Instructor: Ryan Minster

Required Reading:
Midterm Project Due

Learning objectives:
Describe PLINK formats
Create PLINK datafiles
Use PLINK to perform genetic association testing

10/29/2021 Fri: **Unix: Basics, Streams, Redirection, & Pipe**
Instructor: Jon Chernus

Required Reading:
Buffalo (2015) Chapter 3

Learning objectives:
To learn basic Unix commands
To learn how streams operate in Unix
To learn out to pass streamed data from program to program in Unix

---

11/2/2021 Tue: **Unix: Interacting with Processes, Cluster Jobs, Shell Scripting**
Instructor: Jon Chernus

Required Reading:
Buffalo (2015) Chapter 7

Learning objectives:
To learn how to interact with running processes
To learn about the cluster and how to submit jobs there
To learn how to write a script that can run in Unix

11/5/2021 Fri: **Unix: Data Manipulation**
Instructor: Jon Chernus

Required Reading:
Buffalo (2015) Chapter 12

Learning objectives:
To learn Unix tools like sed and awk that can be used to manipulate data

---

11/9/2021 Tue: **Unix: Pipes & Parallelization**
Instructor: Jon Chernus

Required Reading:
Buffalo (2015) Chapter 12

Learning objectives:
To learn to string programs together to process data
To learn how to parallelize functions in Unix

11/12/2021 Fri: **Unix: Scripting, Control Structures and Variables**
Instructor: Jon Chernus

Learning objectives:
To learn how to use control structures in Unix scripting
To learning how to use variables in Unix

11/16/2021 Tue: **Genetic Data Structures**
Instructor: Ryan Minster

Learning objectives:
To learn about what genetic data is stored and principles for storing it

---

11/19/2021 Fri: **PLINK Advanced**
Instructor: Ryan Minster

Learning objectives:
To learn how to use PLINK to manipulate data files

11/23/2021 Tue: **No class - Thanksgiving**

Thanksgiving

11/26/2021 Fri: **No class - Thanksgiving**

Thanksgiving

11/30/2021 Tue: **SAM & samtools**
Instructor: Ryan Minster

Learning objectives:
To learn about SAM data format for sequence data
To learn about samtools to manipulate SAM data files

12/3/2021 Fri: **VCF, bcftools, vcftools**
Instructor: Ryan Minster

Learning objectives:
To learn about VCF data format
To learn about bcftools and vcftools for manipulating VCF files

---

12/7/2021 Tue: **Genetic Data in R, GDS**
Instructor: Jon Chernus

Learning objectives:
To learn about data structures in R for storing genetic data
To learn about the GDS format

12/10/2021 Fri: **Mega2**
Instructor: Dan Weeks

Learning objectives:
To learn about programs for convert data from one format to another

---

12/14/2021 Tue: **Help with Final Project**
Instructor: Jon Chernus

12/16/2021 Thursday:

Final Project Due at 12 noon

12/17/2021 Fri: **No Class**

COURSE MATERIALS

Required Software (All available free online)

Web Browser

R
r-project.org

R Studio
rstudio.com

Pulse Secure
pulsesecure.net

GitHub Classroom
classroom.github.com

Required 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

Academic Integrity

All students are expected to adhere to the school's standards of academic honesty. Cheating/plagiarism will not be tolerated. The Graduate School of Public Health's policy on academic integrity, which is based on the University policy, is available online in the Pitt Public Health Academic Handbook. The policy includes obligations for faculty and students, procedures for adjudicating violations, and other critical information. Please take the time to read this policy.

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.

COVID-19 & Public Health

In the midst of this pandemic, it is extremely important that you abide by public health regulations and University of Pittsburgh health standards and guidelines. **While in class, at a minimum, this means you must wear a face covering** and comply with physical distancing requirements; other requirements may be added by the University during the semester. These rules have been developed to protect the health and safety of all community members. Failure to comply with these requirements will result in you not being permitted to attend class in person and could result in a Student Conduct violation. For the most up-to-date information and guidance, please visit coronavirus.pitt.edu and check your Pitt email for updates before each class.

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.

Accommodation for Students with Disabilities

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, 140 William Pitt Union, 412-648-7890 as early as possible in the term.

Sexual Misconduct, Required Reporting, & Title IX

The University is committed to combatting sexual misconduct. As a result, you should know that University faculty and staff members are required to report any instances of sexual misconduct, including harassment and sexual violence, to the University's Title IX office so that the victim may be provided appropriate resources and support options. What this means is that as your professor, I am required to report any incidents of sexual misconduct that are directly reported to me, or of which I am somehow made aware.

There are two important exceptions to this requirement about which you should be aware:

1. A list of the designated University employees who, as counselors and medical professionals, do not have this reporting responsibility and can maintain confidentiality, can be found here: https://www.diversity.pitt.edu/civil-rights-title-ix/make-report/responsible-employee-program-and-reporting
2. An important exception to the reporting requirement exists for academic work. Disclosures about sexual misconduct that are shared as 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 are the victim of sexual misconduct, Pitt encourages you to reach out to these resources:

- Title IX Office: 412-648-7860
- SHARE @ the University Counseling Center: 412-648-7930 (8:30 AM–5:00 PM Mon–Fri) and 412-648-7856 (after business hours)

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

Other reporting information is available here*: https://www.diversity.pitt.edu/civil-rights-title-ix/make-report*

Diversity Statement

The University of Pittsburgh Graduate School of Public Health considers the diversity of its students, faculty, and staff to be a strength and critical to its educational mission. Pitt Public Health is committed to creating and fostering inclusive learning environments that value human dignity and equity. Every member of our community is expected to be respectful of the individual perspectives, experiences, behaviors, worldviews, and backgrounds of others. While intellectual disagreement may be constructive, no derogatory statements, or demeaning or discriminatory behavior will be permitted.

If you feel uncomfortable or would like to discuss a situation, please contact any of the following:

- the course instructor;
- the Pitt Public Health Associate Dean responsible for diversity and inclusion;
- the University's Office of Diversity and Inclusion at 412-648-7860 or at https://www.diversity.pitt.edu/civil-rights-title-ix/make-report/report-form (anonymous reporting form)