Syllabus

Logistics

Graduate School of Public Health
Department of Human Genetics
HuGen 2070
Bioinformatics for Human Genetics

Tuesdays 1:30 PM–2:55 PM · 3121C Public Health
Fridays 1:30 PM–2:55 PM · 3121C Public Health
3 credit hours
Fall 2019

Contact Information: Please see the 'Faculty Information' section in CourseWeb.

University Course Catalog Description

This course focuses on manipulation and management of human genetic data. The course will cover programming approaches to the processing of phenotype and genotype data for genome-wide association analysis, sequence data, and integrated analyses in the R statistical computing language and Unix scripting language. A key component of the course will be hands-on analyses of example data sets.

Course Description

As technology makes it easier to generate large amounts of genetic data, the well-trained human geneticist needs to learn how to efficiently manipulate, manage, and annotate such data. This course will provide training in the use of the R statistical computing language and environment, in the use of genetic databases and tools, and Unix scripting as applied to data for genetic association studies. We will cover the programming and data structures that underlie the bioinformatics for genome-wide association analysis, sequence data, and integrated analyses. We will complement lectures with exercises in the use of publically-available software and intensive hands-on experience in the analyses of example data sets.

Course Goals

Upon completion of this course, the student will:

- Be able to write efficient R code to reformat, summarize, and visualize data.
- Be able to write efficient Unix scripts to reformat and summarize data.
- Have gained programming skills useful for managing and manipulating human genetic data.
- Describe and create files containing phenotypic and genetic data in a variety of common formats.
- Use data management tools written specifically for common formats to merge, filter and edit genetic data.

Faculty Availability

We welcome your questions. Please feel free to drop by our offices, or to set up an appointment. E-mail is also an excellent way to reach us. However, since we get so many e-mails, please use an informative subject line, starting with "**HuGen 2070:** "

Schedule 2019

Attached Files:
2019_Bioinf_HuGen_Schedule_26Aug2019.pdf (57.008 KB)

Num	Date	Day	Lecture Topic	Reading	Objective 1	Objective 2	Objective 3	Notes	HW
1	8/27/19	Tue	Introduction and Overview	Barnes (2007) Chapter 1	Review the syllabus	Describe bioinformatics	List various type of data used in genetics		Assign HW 1
2	8/30/19	Fri	Genetic Information & dbGaP		Describe what genetic information entails	Describe how genome-wide SNP data are generated	Enumerate the different technologies used in		HW 1 Due
3	9/3/19	Tue	R: Basics	Spector (2008) Chapters 1 & 2; Buffalo (2015) Chapter 8 'R Language Basics', p. 175-193.	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		Assign HW 2
	9/6/19	Fri	No Class					Department Retreat	
4	9/10/19	Tue	R: Factors, Dates, Subscripting	Spector (2008) Chapters 4, 5, 6	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		HW 2 Due Assign HW 3
5	9/13/19	Fri	R: Character Manipulation	Spector (2008) Chapter 7	To learn how to handle character data in R	To learn how to use regular expressions in R			
6	9/17/19	Tue	R: Reproducible Research	Gentleman. Reproducible research: a bioinformatics case study. Statistical applications in genetics and molecular biology (2005) vol. 4 pp. Article2	To understand the concepts of reproducible research	To learn to use R Markdown			HW 3 Due, Assign HW 4
7	9/20/19	Fri	R: Functions and Packages, Debugging R	Buffalo (2015) Chapter 8 'Debugging R Code', p. 236-237.	To learn how to write R functions and packages	To learn how to debug R code			
8	9/24/19	Tue	R: Tidyverse	Chapter 4 of "Statistical Inference via Data Science, A moderndive into R and the tidyverse": https://moderndive.com/4-wrangling.html	To learn how to use the pipe operator				HW 4 Due Assign HW 5
9	9/27/19	Fri	R: Recoding and Reshaping Data	Spector (2008) Chapters 8 & 9	To learn how to reformat and reshape data in R				
10	10/1/19	Tue	Merging Data	Spector (2008) Chapter 9; Buffalo (2015) Chapter 8 'Merging and Combining Data', p. 219-224.	To learn how to merge data using Unix and Python	To learn how to use the R 'merge' command			HW 5 Due Assign HW 6
11	10/4/19	Fri	R: Traditional Graphics & Advanced Graphics	Wickham (2009) Chapters 2 & 3	To learn the basic graphics commands of R	To learn the R graphing package ggplot2			
12	10/8/19	Tue	R: Exploratory Data Analysis	Buffalo (2015) Chapter 8 'Exploring Data Visually with ggplot2', p. 207-219.	To learn how to summarize data frames	To learn how to visualize missing data patterns	To learn how to visualize covariation		HW 6 Due Assign HW 7
13	10/11/19	Fri	R: Interactive and Dynamic Graphics	Wickham (2009) Chapters 2 & 3	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		
14	10/15/19	Tue	PLINK & PLINK Format		Describe PLINK formats	Create PLINK datafiles	Use PLINK to perform genetic association testing		HW 7 Due Work on Midterm Project
	10/18/19	Fri	No Class					ASHG	
15	10/22/19	Tue	Data Quality Checking and Filters		To learn how to check genotype data for quality				Midterm Project Due
16	10/25/19	Fri	R: GRanges	Buffalo (2015) Chapter 9.	To learn how to work with GenomicRanges.	To understand Genomic Ranges and strand issues.			Assign HW 9
17	10/29/19	Tue	Unix: Basics, Streams, Redirection, & Pipe	Buffalo (2015) Chapter 3	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		
18	11/1/19	Fri	Unix: Interacting with Processes, Cluster Jobs, Shell Scripting	Buffalo (2015) Chapter 7	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		HW 9 Due Assign HW 10
19	11/5/19	Tue	Unix: Data Manipulation	Buffalo (2015) Chapter 12	To learn Unix tools like sed and awk that can be used to manipulate data				
20	11/8/19	Fri	Unix: Pipes & Parallelization	Buffalo (2015) Chapter 12	To learn to string programs together to process data	To learn how to parallelize functions in Unix			HW 10 Due Assign HW 11

Num	Date	Day	Lecture Topic	Reading	Objective 1	Objective 2	Objective 3	Notes	HW
21	11/12/19	Tue	Unix: Scripting, Control Structures and Variables		To learn how to use control structures in Unix scripting	To learning how to use variables in Unix			
22	11/15/19	Fri	Genetic Data Structures		To learn about what genetic data is stored and principles for storing it				HW 11 Due Assign HW 12
23	11/19/19	Tue	PLINK Advanced		To learn how to use PLINK to manipulate data files				
24	11/22/19	Fri	SAM & samtools		To learn about SAM data format for sequence data	To learn about samtools to manipulate SAM data files			HW 12 Due Assign HW 13
	11/26/19	Tue	No Class					Thanksgiving	
	11/29/19	Fri	No Class					Thanksgiving	
25	12/3/19	Tue	VCF, bcftools, vcftools		To learn about VCF data format	To learn about bcftools and vcftools for manipulating VCF files			Assign HW 14
26	12/6/19	Fri	Genetic Data in R, GDS		To learn about data structures in R for storing genetic data				HW 13 Due
27	12/10/19	Tue	Mega2		To learn about programs for convert data from one format to another				HW 14 Due
	12/13/19	Fri	No Class						Final Project Due

Course Requirements and Grading

Student Performance Evaluation (Factors and Weights)

Evaluation will be based on the following components:

Homework & Quizzes (70% *of final grade)*

The homework assignments will comprise problems that extend the in class activities and build the skills needed to complete each of the projects.

Midterm Project (*15% of final grade*)

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.

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 describe the data.

Final Project (*15% of final grade*)

The final project of this course asks the students to prepare mock genotype data for deposition into dbGaP. This includes merging, cleaning, and validating the data.

Late Assignment Policy

We expect homework to be handed in on time. If you are not going to be able to make the deadline, please let us know the reason why. For valid reasons, we may be able to extend the deadline for you.

Even if your homework is late, we'd prefer that you hand your homework in, as we believe you'll learn something from attempting to do it. Reduced points may be awarded for late homework if it is handed in before we have handed out or discussed the answers and before we've completed correcting the homework.

Course Prerequisites

Hugen 2022 · Population Genetics

Biostat 2041 · Biostats Methods I

or approval of the instructor

Students should also have basic computing and programming skills.

Grading Scale

Grades scored between	will equal
97% and 100%	A+
94% and less than 97%	A
90% and less than 94%	A–
87% and less than 90%	B+
84% and less than 87%	B
80% and less than 84%	B–

77% and less than 80%	C+
74% and less than 77%	C
70% and less than 74%	C–
67% and less than 70%	D+
64% and less than 67%	D
60% and less than 64%	D–
0% and less than 60%	F

CourseWeb/BlackBoard Instruction

This course will extensively use the University's CourseWeb site http://courseweb.pitt.edu/ [also known as BlackBoard]. To login, you must have a Pitt account. Your login ID is the same as your login ID for your Pitt account and your password is the same as for your 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 CourseWeb, 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 the CourseWeb site as the primary means of communicating with the students, who are expected to check the site on a regular basis throughout the semester.

Course Materials

Readings (Available online):

Data manipulation with R
Author: Spector, Phil.
Publisher: New York: Springer, c2008.
Web access: http://site.ebrary.com/lib/pitt/Doc?id=10223421
Permanent URL: http://pittcatplus.pitt.edu/?itemid=|library/marc/voyager|5689739

ggplot2: Elegant Graphics for Data Analysis
Author: Wickham, Hadley Author
Publisher: New York: Springer Aug. 2009
Web access: http://dx.doi.org/10.1007/978-0-387-98141-3
Permanent URL: http://pittcatplus.pitt.edu/?itemid=|library/marc/voyager|6140635

Bioinformatics Data Skills
Editor: Vince Buffalo
Publisher: O'Reilly
Web access: http://proquest.safaribooksonline.com/9781449367480

Bioinformatics for Geneticists (Second Edition)
Editor: Michael R. Barnes
DOI: 10.1002/9780470059180
Web access: http://www3.interscience.wiley.com/cgi-bin/bookhome/114171153/

Supplemental Readings/Bibliography (Optional)

Introductory statistics with R
Author: Dalgaard, Peter.
Publisher: New York: Springer, c2002.
Web access: http://site.ebrary.com/lib/pitt/Doc?id=10047812

Current Protocols in Bioinformatics

Editor: Baxevanis AD, Stein LD, Stormo GD, Yates JR
Publisher: John Wiley and Sons, Inc., c2010
Web
access: http://www.mrw.interscience.wiley.com/emrw/9780471250951/cp/cpbi/toc

Single Nucleotide Polymorphisms: Methods and Protocols
Editor: Anton A. Komar
Publisher: New York: Springer, c2008.
Series: Methods in Molecular Biology
Volume: 578
Print ISBN: 978-1-60327-410-4
Web access: http://www.springerprotocols.com/Full/doi/10.1007/978-1-60327-411-1_3

Searching by using the MeSH Database
Source: NCBI Handbook (PubMed help page)
Web access: http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helppubmed&part=pubmedhelp#pubmedhelp.Searching_by_using_t

GoPubMed: exploring PubMed with the Gene Ontology
Authors: Doms A, Schroeder M.
Ref Nucleic Acids Res. 2005 Jul 1; 33 (Web Server issue): W783-6.
Web access: http://www.ncbi.nlm.nih.gov/pubmed/15980585

A quick guide for developing effective bioinformatics programming skills
Authors: Dudley and Butte.
PLoS Comput Biol (2009) vol. 5 (12) pp. e1000589
Web
access: http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1003789

R programming for bioinformatics
Author: Robert Gentleman
Publisher: Boca Raton : CRC Press, c2009.

Interactive and Dynamic Graphics for Data Analysis: with R and GGobi
Authors: Dianne Cook and Deborah F. Swayne
Publisher New York: Springer Verlag 2007.

Bioinformatics and computational biology solutions using R and Bioconductor
Editors: Robert Gentleman et al.
Publisher: New York: Springer Science+Business Media, c2005.

Academic Policies

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.

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: http://www.bc.pitt.edu/policies/policy/02/02-03-02.html

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.

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.

Accessibility

Blackboard is ADA Compliant and has fully implemented the final accessibility standards for electronic and information technology covered by Section 508 of the Rehabilitation Act Amendments of 1998. Please note that, due to the flexibility provided in this product, it is possible for some material to inadvertently fall outside of these guidelines.

Copyright Notice

Statement on Classroom Recording

To ensure the free and open discussion of ideas, students may not record classroom lectures, discussion 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.

Diversity

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 for Diversity at 412-624-3506 or nam137@pitt.edu;
- the University's Office of Diversity and Inclusion at 412-648-7860 or via this anonymous reporting form.

Sexual Misconduct, Required Reporting and Title IX

The University is committed to combating 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:

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.

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
- The University Counseling Center: 412-648-7930 (8:30 A.M. TO 5 P.M. Monday–Friday) 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.