

Quantitative Genomics and Genetics

BIOCB 4830/6830; PBSB.5201.03

Jason Mezey

Department of Computational Biology
Department of Genetic Medicine

TA: Beulah Animah Agyemang-
Barimah
baa95@cornell.edu

TA: Samuel (Sam) Terkper
Ahuno
sta4008@med.cornell.edu

Spring 2024: Jan 23 - May 9
T/Th: 8:40-9:55

Why you're here

Spring 2024 Course Announcement

Quantitative Genomics and Genetics

Professor: Jason Mezey
Computational Biology (Cornell)
Department of Genetic Medicine (Weill Cornell)

Dates: Jan 23 – May 9
Days: Tues. and Thurs.
Time: 8:40 am – 9:55 am

COURSE DESCRIPTION: A rigorous treatment of analysis techniques used to understand the genetics of complex phenotypes when using genomic data. This course will cover the fundamentals of statistical methodology with applications to the identification of genetic loci responsible for disease, agriculturally relevant, and evolutionarily important phenotypes. Data focus will be genome-wide data collected for association analysis, as well as for inbred and pedigree experimental designs. Analysis techniques will focus on the central importance of generalized linear models in quantitative genomics with an emphasis on both Frequentist and Bayesian computational approaches. Tools learned in class will be implemented in the computer lab, during which the language R will be taught from the ground up (no previous experience required or expected)

GRADING: S/U or Letter Grade.

CREDITS: 4 (lecture + computer lab).

SUGGESTED PREREQUISITES: At least one class in Genetics and one class in probability and / or statistics.

Today

- Logistics (time/locations, registering, syllabus, schedule, requirements, computer labs)
- Intuitive overview of the goals and the field of quantitative genomics
- The foundational connection between biology and probabilistic modeling
- Begin our introduction to modeling and probability

Times and Locations I

- Lectures are every Tues. / Thurs. 8:40-9:55AM - see class schedule (to be posted!)
- In-person lecture locations:
 - Ithaca: In-person lectures in Weill Hall 224 (Tues) and Weill Hall 226 (Thurs)
 - NYC: Except today (!!) in-person lectures will be in Belfer Building (69th between 1st and York) on the 2nd or 3rd floor (on Thurs Jan 25! Lecture will be in BB204-A,B) - I will post a schedule soon!
- For those on Zoom:
 - NYC students are joining by zoom now (please mute / unmute to ask questions)
 - We will discuss zoom access going forward in future lectures...
- Lectures will be recorded:
 - These will be posted along with slides / notes
 - In person lecture attendance is not required BUT I (strongly) encourage you to come to class...

Times and Locations II

- There is a REQUIRED computer lab
- **FIRST COMPUTER LAB WILL BE NEXT WEEK (Thurs. Feb 1 / Fri. Feb 2)**
- For those IN ITHACA (= Labs with Beulah):
 - Lab 1: 3:35-4:25PM on Thurs. (Mann Library B30A)
 - Lab 2: 9:05-9:55AM on Fri. (Mann Library B30A)
- For those IN NYC (= Labs with Sam!):
 - FRIDAYS (only!) 9-10AM in A-950 Auditorium, 1300 York Ave (9th floor)
 - If you have a problem with this lab time, please contact me (see following slides...)

Times and Locations III

- Again: the computer lab is REQUIRED (if you take the course for credit!)
 - We take attendance (= this will impact your grade)
 - We will teach you R from the ground up (= don't worry if you know nothing about R or programming in general)
- HOWEVER, if you are already an experienced R programmer:
 - You may skip the first **2 labs** without penalty
 - If you really feel you will get nothing out of the labs please contact me and we can discuss...

Times and Locations IV

- I (Jason) will hold office hours for both campuses by zoom at a time TBD (stay tuned for more information!)
- NO office hours this week - this will start next week
- You may also set up individual sessions with me (Jason) by appointment

Registering for the class I

- If you can register for this class, please do so (even if you plan to audit!!)
- If you register, you may take this class for a grade (letter in Ithaca, Honors / HP / etc. at Weill), P/F, S/U, or Audit
- If you cannot register for some reason, you are still welcome to take the class (e.g., sit-in) and, if you do the work, we will grade it as if you are registered (!!)
- If you audit or do not register officially, while not required, I strongly recommend that you do the work for the class, (i.e. homework / exams / project / computer lab)
- My observation is that you are likely to be wasting your time if you do not do the work but I leave this up to you...

Registering for the class II

- In Ithaca:
 - You must register for both the lecture (3 credits) and computer lab (1 credit) if you take the course for a letter grade
 - If you are an undergraduate, register for BIOCB 4830 (lecture and lab); graduate student, register for BIOCB 6830 (same)
- In NYC:
 - At Weill: the course (PBSB.5021.03) should be available in the Graduate School drop-down at learn.weill.cornell.edu
 - If Other: check with WCMC registrar for instructions
- Please contact me if there are any issues with registering (!!)

Grading

- We will grade undergraduates and graduates separately (!!)
- Grading: problem sets (20%), computer lab attendance (5%), project (25%), mid-term (20%), final (30%)
 - A short problem set ~6 total
 - Exams will be take-home (open book)
 - A single project (~1 month)
- Note that while computer lab attendance impacts your grade, lecture attendance does not (= again, lecture attendance is optional - although highly recommended...)

Class Resource I: CANVAS

Spring 2024

BIOCB4830/BIOCB6830

Recent Announcements

Recent Activity in BIOCB4830/BIOCB6830

No Recent Messages You don't have any messages to show in your stream yet. Once you begin participating in your courses you'll see this stream fill up with messages from discussions, grading updates, private messages between you and other users, etc.

[View Course Calendar](#)

[View Course Notifications](#)

To Do

Nothing for now

- Note: not populated with content yet... but will be by end of week (please check back!)

Getting on Canvas / emailing

- We will use CANVAS for this class (for everything: posting, announcements, emailing, homework / work uploads, discussion posts, labs, etc.:)
- Everyone in Ithaca should automatically be signed up (if you are registered for the class...)
- If you are at Weill Cornell please register ASAP (!!) using your CWID at: <https://request.canvas.cornell.edu>
- Please email me if you cannot sign up on the class Canvas for some reason (!!)
- ALL EMAIL for any aspect of the course must be sent through Canvas (we will stop answering direct emails after the first week of the course)
- PLEASE DON'T email Jason / Beulah / Sam's direct email after the first week (=we will ignore you - unless its an emergency...)

Canvas materials

- We will post information about the course and a schedule updated during the semester (check back often!!)
- I will post slides for all lectures and I will TRY to post slides before each lecture (no promises!)
- We will post videos of lectures (with a delay in most cases)
- We will post a “Partial” Textbook: *Quantitative Genetics 2022* (Shizhong Xu) - I will post individual relevant sections on Canvas to accompany lectures (stay tuned...)
- All homeworks, exams, keys, etc. will be posted on Canvas and you will upload your work to Canvas
- All computer labs and code will be posted on Canvas

Canvas posting / discussions

The image shows a screenshot of the Canvas LMS interface for a course. On the left is a vertical navigation menu with icons for Account, Dashboard, Ed Discussion (circled in red), Canvas, Inbox, History, and Help. The main content area is titled "Recent Activity in BIOCB4830/BIOCB6830" and contains a message box with an information icon and the text: "No Recent Messages You don't have any messages to show in your stream yet. Once you begin participating in your courses you'll see this stream fill up with messages from discussions, grading updates, private messages between you and other users, etc." A red arrow points from this message box to a detailed view of an Ed Discussion. The detailed view shows a post titled "Welcome! #1" by Jason Mezey (STAFF) in the "General" category, posted 3 days ago. The post content includes a welcome message and a list of tips for using Ed Discussion.

BIOCB4830/BIOCB6830

Spring 2024

Recent Announcements

View Course Calendar

View Course Notifications

To Do

Nothing for now

Home

Syllabus

Course Materials

Ed Discussion

No Recent Messages You don't have any messages to show in your stream yet. Once you begin participating in your courses you'll see this stream fill up with messages from discussions, grading updates, private messages between you and other users, etc.

ed BIOCB 4830 - Ed Discussion

New Thread

Search

Filter

WELCOME!

General Jason Mezey STAFF 3d

UNPIN STAR WATCHING VIEWS 86

Hi everyone,

We're using Ed Discussion for class Q&A.

This is the best place to ask questions about the course, whether curricular or administrative. You will get faster answers here from staff and peers than through email.

Here are some tips:

- Search before you post
- Heart questions and answers you find useful
- Answer questions you feel confident answering
- Share interesting course related content with staff and peers

For more information on Ed Discussion, you can refer to the [Quick Start Guide](#).

All the best this semester!

Jason

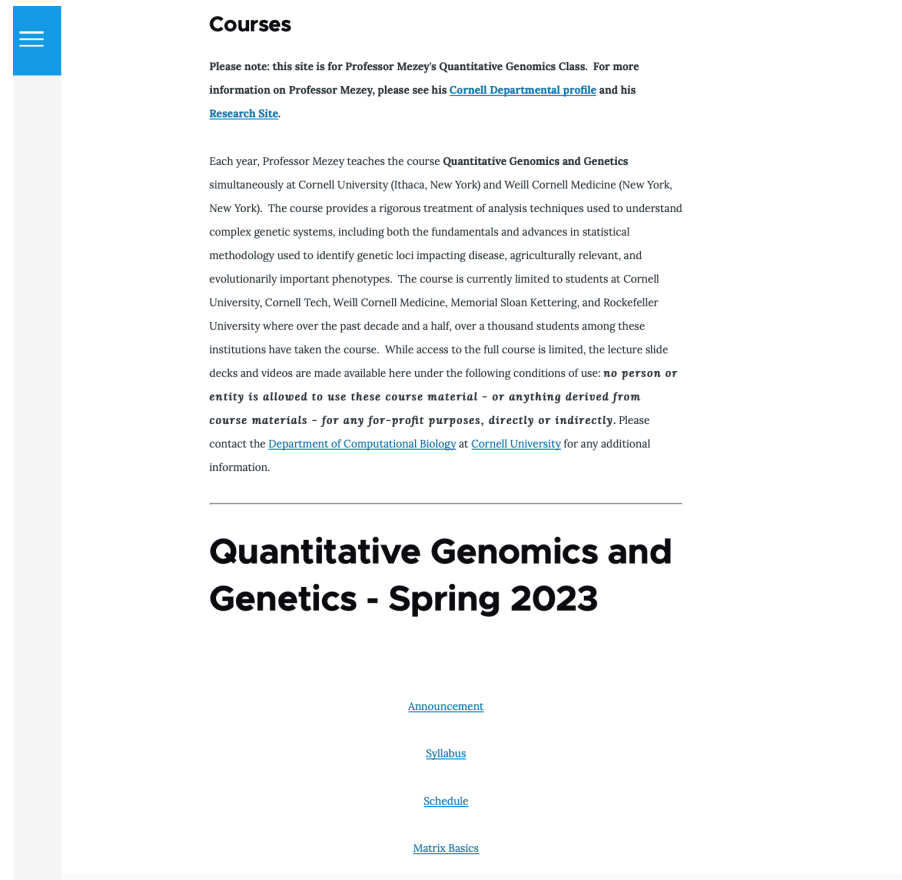
Comment Edit Delete ...

Canvas posting

- Posting Protocol:
 - Feel free to post questions and comments
 - Public posts (let the community of students and instructors help out)
 - Private questions: email Jason, Beulah, or Sam (using Canvas)
- Please note that expected response times to questions will be minimum >24hrs (sometimes longer..) depending on the availability of the instructors
- We encourage public posts so that your classmates can help you out as well (this worked great in previous years!)

Class Resource II: Website

- An additional class website: <https://mezeylab.biohpc.cornell.edu>



The screenshot shows a website page with a blue header containing a white hamburger menu icon. The main content area is white with a light gray vertical bar on the left. The page is titled "Courses" and includes a note about Professor Mezey's Quantitative Genomics Class, a paragraph of course details, and a section for "Quantitative Genomics and Genetics - Spring 2023" with links to "Announcement", "Syllabus", "Schedule", and "Matrix Basics".

Courses

Please note: this site is for Professor Mezey's Quantitative Genomics Class. For more information on Professor Mezey, please see his [Cornell Departmental profile](#) and his [Research Site](#).

Each year, Professor Mezey teaches the course **Quantitative Genomics and Genetics** simultaneously at Cornell University (Ithaca, New York) and Weill Cornell Medicine (New York, New York). The course provides a rigorous treatment of analysis techniques used to understand complex genetic systems, including both the fundamentals and advances in statistical methodology used to identify genetic loci impacting disease, agriculturally relevant, and evolutionarily important phenotypes. The course is currently limited to students at Cornell University, Cornell Tech, Weill Cornell Medicine, Memorial Sloan Kettering, and Rockefeller University where over the past decade and a half, over a thousand students among these institutions have taken the course. While access to the full course is limited, the lecture slide decks and videos are made available here under the following conditions of use: **no person or entity is allowed to use these course material - or anything derived from course materials - for any for-profit purposes, directly or indirectly.** Please contact the [Department of Computational Biology at Cornell University](#) for any additional information.

Quantitative Genomics and Genetics - Spring 2023

[Announcement](#)

[Syllabus](#)

[Schedule](#)

[Matrix Basics](#)

- This has not yet been updated but currently has videos from last year and lecture slide decks from last year (=same content) - take a look!

What you will learn in this class I

- A rigorous introduction to basics of probability and statistics that is intuition based (not proof based)
- Foundational concepts of how probability and statistics are at the core of genetics, which are complete enough to build additional / more advance understanding (i.e., enough to “get your hooks into the subject”)
- Exposure to many advanced probability / statistics / genetics / algorithmic concepts that will allow you to build additional understanding beyond this class
- Clear explanations for convincing yourself that the basics of mathematics and programing are not hard (i.e. anyone can do it if they devote the time)

What you will learn in this class II

- An intuitive and practical understanding of linear models and related concepts foundation to statistics, machine learning, and computational biology
- The computational approaches necessary to perform inference with these models (EM, MCMC, etc.)
- The statistical model and frameworks that allow us to identify specific genetic differences responsible for differences in organisms that we can measure
- You will be able to analyze a large data set for this problem, e.g. a Genome-Wide Association Study (GWAS)
- You will have a deep understanding of quantitative genomics that from the outside seems diffuse and confusing

Should I be in this class I

- No probability or statistics: not recommended
- Limited probability or statistics (high school, a long time ago, etc.): if you take the class be ready to work (!!)
- Prob / Stats (e.g. BTRY 4080+4090 or BTRY 6010+6020 in Ithaca, Quantitative understanding in biology at Weill, etc.): you'll be fine
- No or limited exposure to genetics: you'll be fine
- No or limited exposure to programming: you'll be fine (we will teach you "programming" in R from the ground up)
- Strong quantitative background (e.g. stats or CS graduate student): you may find the intuitive discussion of quantitative subjects and the applications interesting

Should I be in this class II

- Every year many students have concerns - please don't let the following dissuade you...
- (1) It's too late for me to learn this
 - = wrong, it does not matter when you start (e.g., the students in my lab learn it all once they join my lab)
- (2) I'm not smart enough to learn this (e.g., I've taken math classes and I couldn't follow them / when other students talk I don't know what they're taking about, etc.)
 - = wrong, if you've gotten this impression, you've been in inappropriate or badly taught classes and / or you've been talking to insecure students (or faculty) who think "knowing" math that has been figured out by others / explaining math concepts in an unclear way means they are intelligent (it does not...)
- (3) It's not worth my time
 - = this is a more personnel question but given the way the world is moving it probably worth your time if you can do it...

Should I be in this class III

- Final thoughts on this from a previous student:

“As I have mentioned before, I entered this course with limited background in R and GWAS. However, thanks to your course, I now feel much more comfortable with both. I am preparing the specific aims for my A exam proposal and I now feel confident that I will be able to succeed in this project. Of course, I still have a lot to learn, but I feel like I built a great foundation on these topics/skills in your class.

I know that students from different backgrounds take your class, so I wanted to share with you some of the things that helped me the most to get through the class and do well. You are welcome to share these points with future students like me that may be a little intimidated by the class at first.

1. Re-watching the lectures (KEY!)
2. Going over the solutions to the lab exercises as soon as the TAs released them
3. Going to office hours
4. Trusting the process!

Questions about
logistics?

Introduction to genetics and probability basics

- Today, we will provide a (brief and) broad introduction to the field of *quantitative genomics*, is a field concerned ***with the modeling of the relationship between genomes and phenotypes and using these models to discover and predict***
- In this class, we will be concerned with the most basic problem of quantitative genomics: ***how to identify genotypes where differences among individual genomes produce differences in individual phenotypes*** (i.e. genetic association studies) which is the foundation of all genetic analysis work
- The same analysis concepts and approaches also underly all work in ANY data science work, whether you are applying statistics, computational statistics, or machine learning approaches

Genotype and Phenotype

- We know that aspects of an organism (measurable attributes and states such as disease) are influenced by the genome (the entire DNA sequence) of an individual
- This means difference in genomes (genotype) can produce differences in a phenotype:
 - Genotype - any quantifiable genomic difference among individuals, e.g. Single Nucleotide Polymorphisms (SNPs). Other examples?

GAATTC
GAACTC

TCGCGAA-----TTCCCAT
TCGCGAAACGTTTCCCAT

- Phenotype - any measurable aspect of an organisms (that is not the genotype!). Examples?

An illustration continued...

- The problem: for any two people, there can be millions of differences their genomes...
- How do we figure out which differences are involved in producing differences and which ones are not?
- This course is concerned with how we do this
- Note that the problem (and methodology) applies to any measurable difference, for any type of organism!!

Why do we want to know this?

If you know which genome differences are responsible:

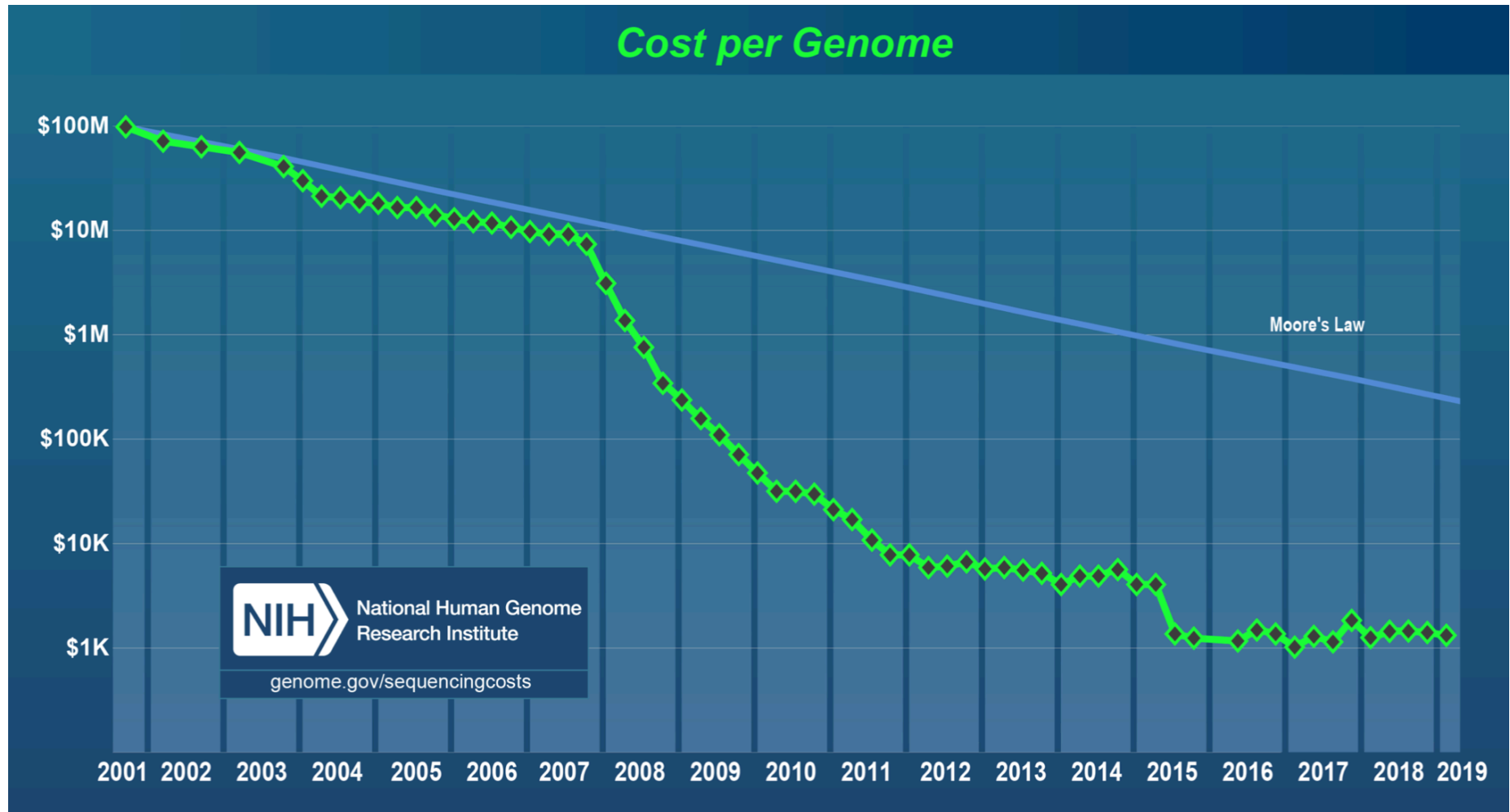
- From a child's genome we could predict adult features
- We target genomic differences responsible for genetic diseases for gene therapy
- We could predict an individual's risk for having a disease
- We can explain why a disease has a particular frequency in a population, why we see a particular set of differences
- We can manipulate genomes of agricultural crops to be disease resistant strains
- These differences provide a foundation for understanding how pathways, developmental processes, physiological processes work
- The list goes on...

History of genetics (relevant to Quantitative Genetics)

- Relevant history:
 - 1900-1980: statistical analysis of the patterns of inheritance (i.e. the resemblance between relatives).
 - 1980-2002: mapping (= identification) of the genetic loci responsible for most Mendelian diseases (e.g. diseases where alleles at a 'single' genetic locus determines disease).
 - 2002-present: 'age of genomics' first convincing mapping of genetic loci for complex traits (i.e. cases where genotype cannot be inferred directly from the phenotype).

In sum: during the last two decades, the greater availability of DNA sequence data has completely changed our ability to make connections between genome differences and phenotypes

Present / future: advances in next-generation sequencing driving the field



Connection of genomics-genetics

- Traditionally, studying the impact / relationship of the genome to phenotypes was the province of fields of “Genetics”
- Given this dependence on genomes, it is no surprise that modern genetic fields now incorporate genomics: the study of an organism’s entire genome (wikipedia definition)
- However, one can study genetics without genomics (i.e. without direct information concerning DNA) and the merging of genetics-genomics is quite recent

The impact of Genomic Data on genetic analysis

- Before the “Genomic Era” genetic analysis was part of three different fields that used different analysis techniques: **Medical Genetics**, **Agricultural Genetics**, and **Evolutionary Genetics**
- The reason was they were analyzing different systems / interested in different questions AND they did not have the data available to do what they really wanted to do: *identify which differences in a genome (genotypes) were responsible for differences in phenotypes of interest (!!)*
- Once genomic data (i.e., data on the entire genome) became available the starting analysis of all of these fields became the same (i.e., analyzing which differences impacted phenotypes) *and they started using the same set of methods (!!)* = effectively unifying these fields into modern “Quantitative Genetics / Genomics”
- This is the reason the Quantitative Genetics literature before the Genomic Era is so difficult to follow / seems so diffuse... but after this class you will understand how to go back and figure out this literature (!!)

Why this is a good time to be learning about this subject

- Mapping (identifying) genotypes (genetic loci) with effects on important phenotypes is perhaps the major use of genomic data and a major focus of genomics
- However, the data collection, experimental, and statistical analysis techniques for doing this are still being developed
- The current statistical approaches are the focus of this course (i.e., you will have a solid foundation by the end)
- The importance is just now starting to permeate broadly (i.e., we are now in the “internet generation” for genomics and the impact of genomics on biology)
- The basic statistical approaches are (=should be) applied in ANY analysis of ANY genomic data for ANY purpose

That's it for today

- Next lecture, we will begin our formal and technical introduction to probability where we will start by defining the concepts of a “system”, “experiments” and “experimental trials”, and “sample outcomes” and “sample spaces”