Quantitative Genomics and Genetics

BTRY 4830/6830; PBSB.52201.03

Jason Mezey
Department of Computational Biology
Department of Genetic Medicine

TA: Mitchell (Mitch) Lokey
mgl77@cornell.edu
(Cornell, Ithaca)

TA: Samuel (Sam) Terkper Ahuno
sta55@cornell.edu
(NYC - note: Ithaca email not med)

Spring 2023: Jan 24 - May 9
T/Th: 8:05-9:20

Why you're here

Spring 2023 Course Announcement

Quantitative Genomics and Genetics

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

Dates: Jan 24 – May 9
Days: Tues. and Thurs.
Time: 8:05 am – 9:20 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. 8:05-9:20AM - see class schedule (to be posted)
- In-person lecture locations:
  - Ithaca: All in-person lectures in Weill Hall 226
  - NYC: Many different locations (!!) - schedule to be posted (i.e., you will have to check every lecture!)
- Zoom option:
  - Remote (to both Ithaca / NYC) students are joining by zoom now (please mute / unmute to ask questions)
  - By next week, we will have a zoom option for everyone (we will discuss)
- Lectures will be recorded:
  - These will be posted along with slides / notes
  - I 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 2 / Fri. Feb 3)**
- PLEASE NOTE (!!): LAB TIMES ARE DIFFERENT THAN LISTED
- For those IN ITHACA (= Labs Mitch!):
  - Lab 1: 5:30-6:30PM on Thurs. (Weill Hall 226)
  - Lab 2: 8-9AM on Fri. (Weill Hall 226)
  - THIS WEEK: if you go to the lab (see next) that you registered for / please contact me if you need to switch
- For those IN NYC (= Labs taught by Sam!):
  - Lab 1: 4-5PM on Thurs. (In WCMC1300 Classroom; G [B215], H [B217])
  - Lab 2: 9-10AM on Fri. (By zoom - please stay tuned for invite)
  - If you go to the lab (see next) and you are in NYC please to the Thurs. section OR if you are remote (e.g. in Houston) attend Fri by zoom

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

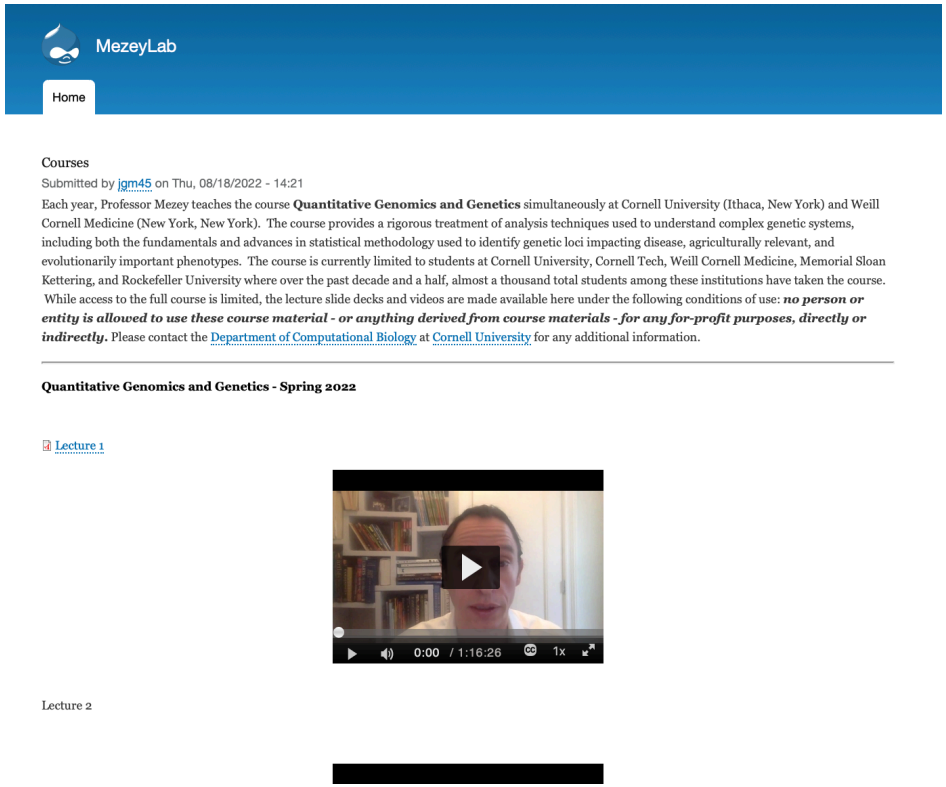
- In Ithaca:
  - You must register for both the lecture (3 credits) and computer lab (1 credit) if you take the course for a letter grade
  - If you are an undergraduate, register for BTRY 4830 (lecture and lab); graduate student, register for BTRY 6830 (same)
- In NYC:
  - At Weill: the course (PBSB.5021.03) 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 (= lecture attendance is optional - although highly recommended…)

Class Resource 1: Website

- The class website will be a under the "Classes" link on my site: https://mezeylab.biohpc.cornell.edu



- This has not yet been updated but will be by end of week (syllabus, schedule, etc.) and lecture decks and videos!

Website resources

- We will post information about the course and a schedule updated during the semester (check back often!!)
- We will post videos of lectures (delay in most cases)
- The website lecture videos from last year and will (eventually) include lecture slide decks from last year (=same content) - take a look!
- There is no textbook for the class but I will post slides for all lectures
- All homeworks, exams, keys, etc. will be posted elsewhere (see slides that follow)
- All computer labs and code will be posted elsewhere (see slides that follow)

Class Resources II: Piazza

- MAKE SURE YOU SIGN UP ON PIAZZA whether you officially register or not = all communication for the course (!!)
- Class: https://piazza.com/cornell/spring2023/btry4830btry6830
- If you are at Cornell (Ithaca) you can sign up right now!
- If you are at Weill you may not be able to sign up using the following instructions PLEASE EMAIL ME DIRECTLY at **jgm45@cornell.edu** and I will get you on
- To register:
  - Step 1: Sign up on Piazza (if you don't have an account already)!
  - Step 2: Enroll in the same course (regardless if you are grad or undergrad!)
  - Make sure you're signing up for Spring 2023 (!!)
- EVERYBODY PLEASE GET REGISTERED ASAP (!!)

Email and Posting

- ALL EMAIL for any aspect of the course must be sent through PIAZZA (we will stop answering direct emails after the first week of the course)
- PLEASE DON'T email Jason / Mitch / Sam's direct email after the first week (=we will ignore you - unless its an emergency…)
- Posting Protocol:
  - Post all questions and comments on Piazza.
  - Public posts (Let the community of students and instructors help out)
  - Private posts (To Jason and Mitch and Sam)
- 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 III: CMS

- Assignments and computer labs (!!) will be posted on Cornell CMS (as BTRY 4830)
- This is not yet setup (please stay tuned for information on how to register)
- Note: ALL submissions should be made through the CMS website (=please don't email submissions to Jason or Mitch or Sam)

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

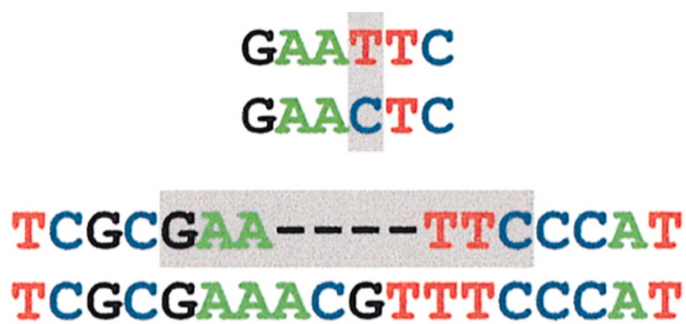
- Today, we will provide a (brief and) broad introduction to the field of *quantitative genomics*, is a field concerned ***with the modeling of the relationship between genomes and phenotypes and using these models to discover and predict***

- 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?



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

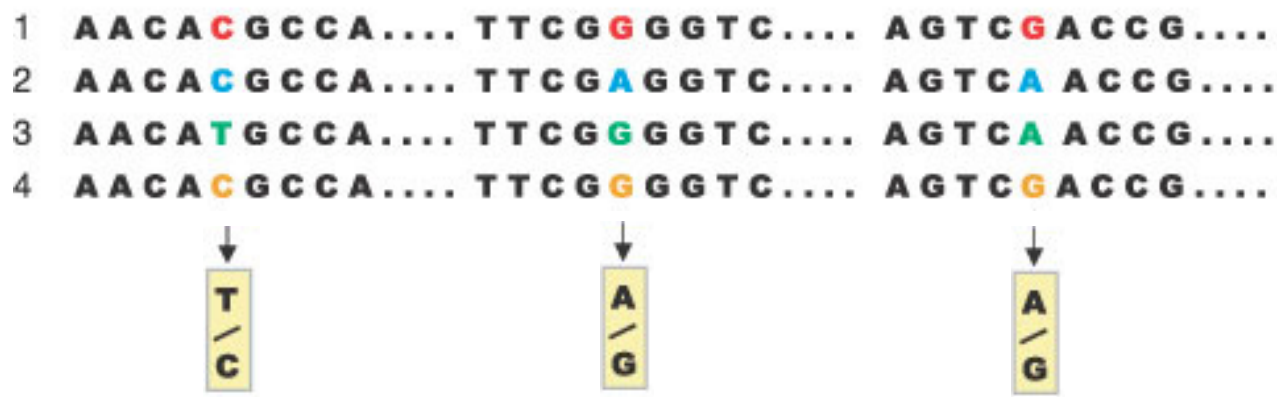
An illustration

Example: People are different...

Physical, metabolism, disease, countable ways.

We know that environment plays a role in these differences

...and for many, differences in the genome play a role



For any two people, there are millions of differences in their DNA, a subset of which are responsible for producing differences in a given measurable aspect.

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:

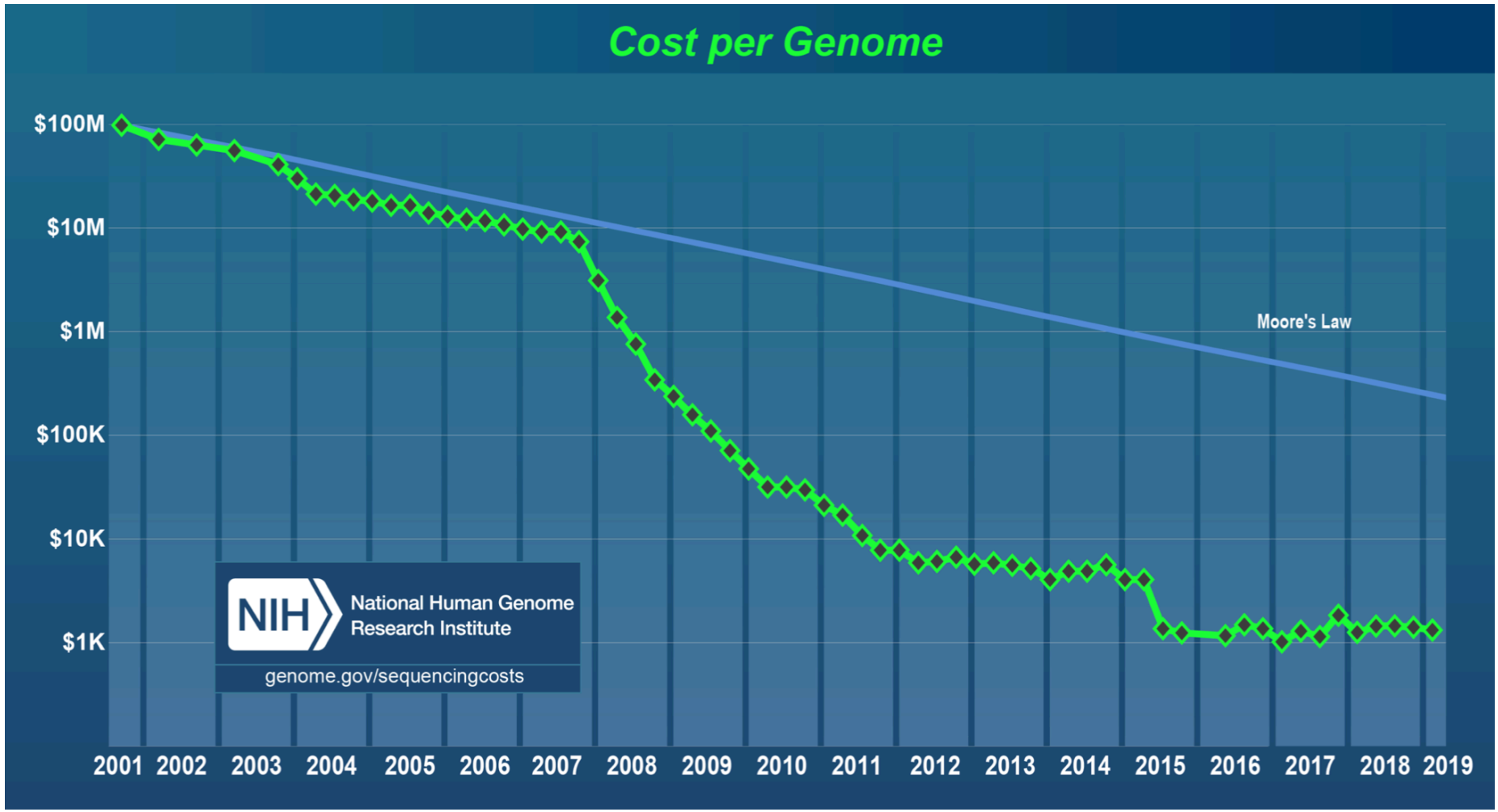
- From a child's genome we could predict adult features
- We could predict an individual's risk for having a disease
- We target genomic differences responsible for genetic diseases for gene therapy
- We can manipulate genomes of agricultural crops to be disease resistant strains
- We can explain why a disease has a particular frequency in a population, why we see a particular set of differences
- 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 inheritence (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

Motivating intro to prob & statistics: foundational biology concepts

- In this class, we will use ***statistical modeling*** to say something about *biology*, specifically the relationships between genotype (DNA) and phenotype
- Let's start with the biology by asking the following question: why DNA?
- The structure of DNA has properties that make it worthwhile to focus on...

It's the same in all cells

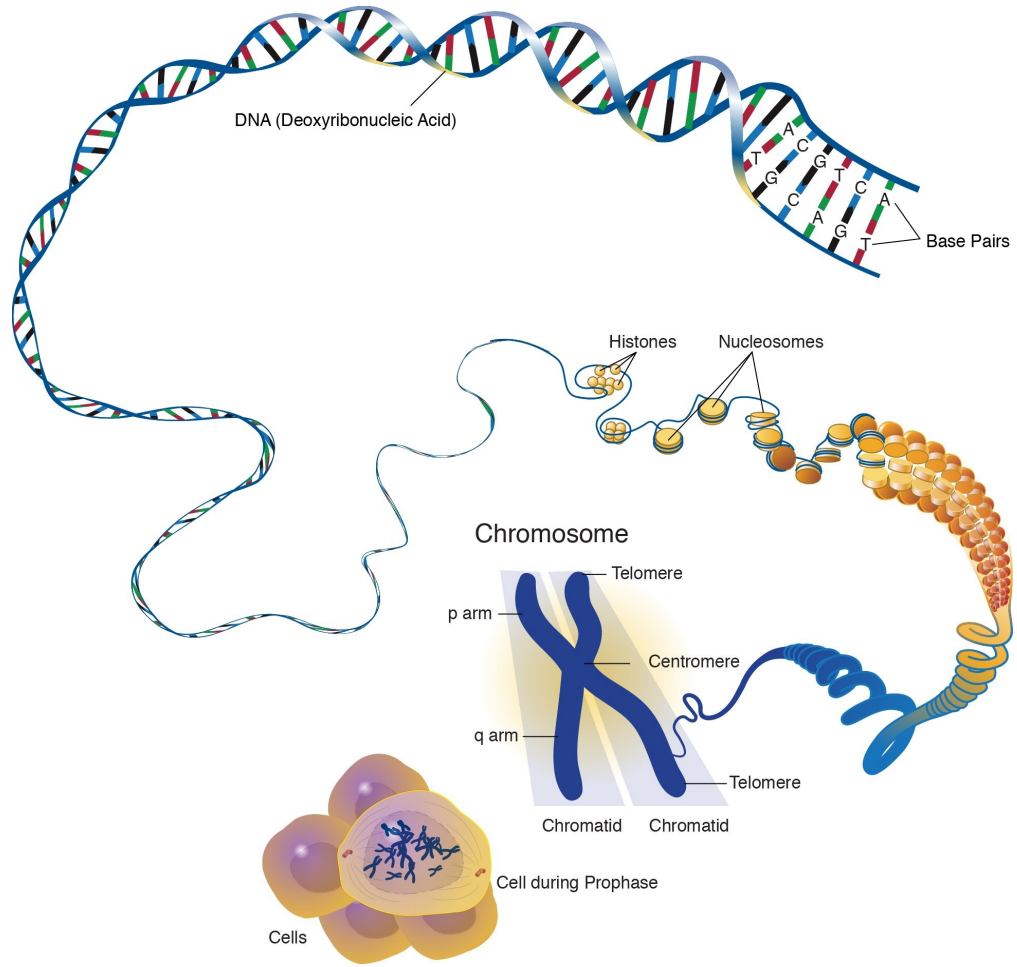
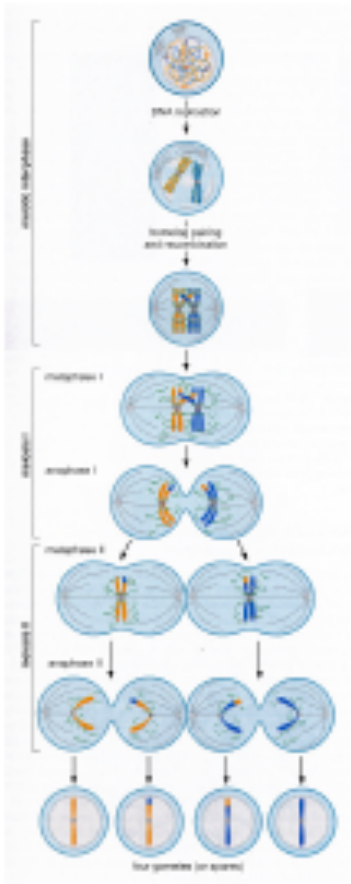
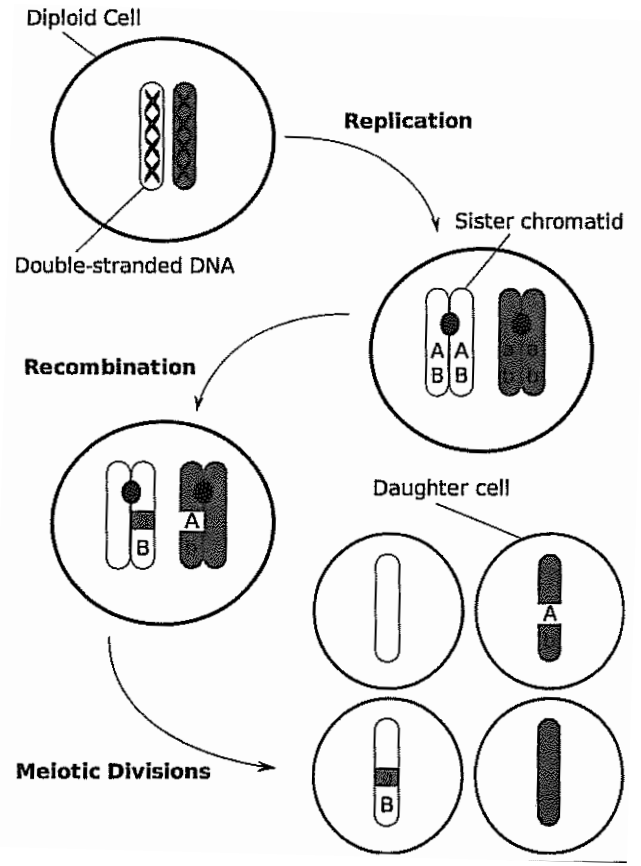


Figure 1: A simplified schematic showing genome organization in human cells. The DNA of a genome is located within the nucleus of a cell. The genome is organized in long strings that are tightly coiled around protein structures to form chromosomes. Each string is a double helix where the building blocks are A-T and G-C nucleotide pairs © *kintalk.org*.

with a few exceptions (e.g. cancer, immune system...)

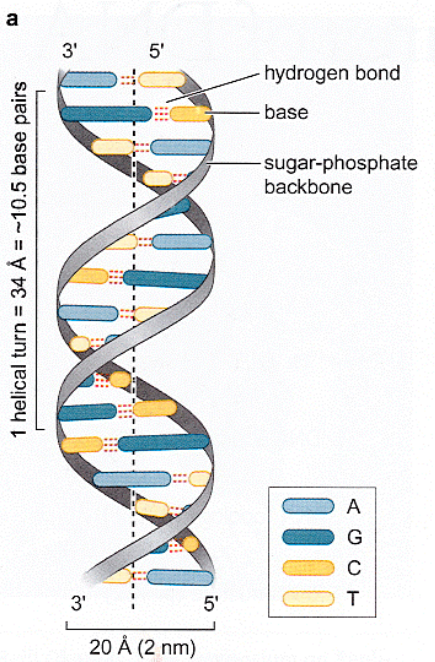
It's passed on to the next generation

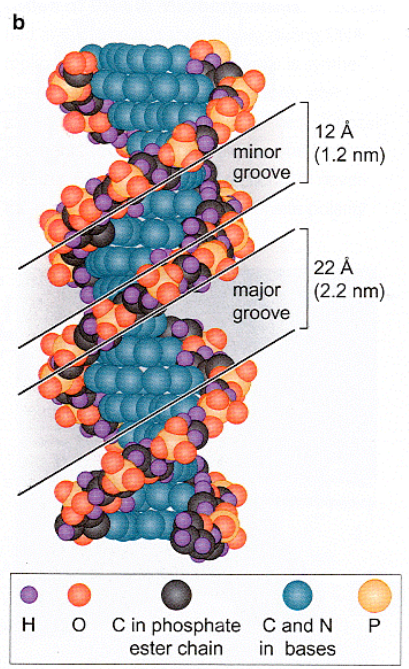


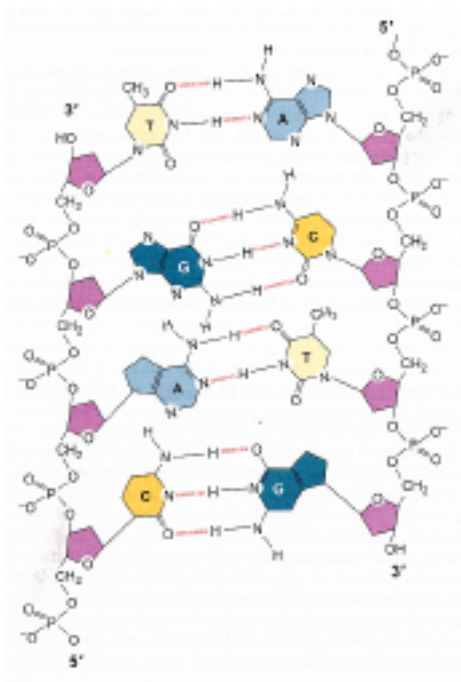


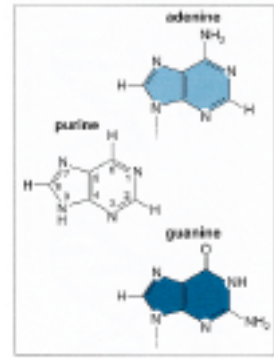
Credit: Watson et al., *Molecular Biology of the Gene*, CSHL Press, 2004

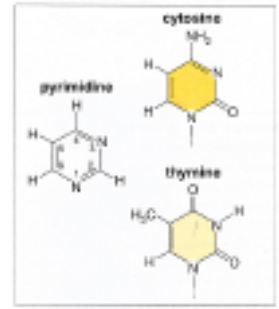
It has convenient structure for quantifying differences









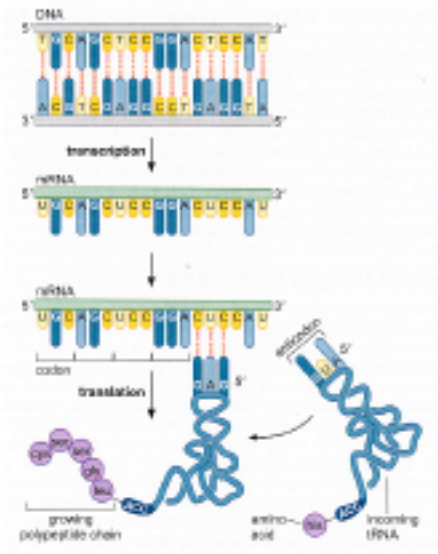


Credit: Watson et al., *Molecular Biology of the Gene*, CSHL Press, 2004

It's almost the same in each individual in a species

1 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....
2 AACACGCCA.... TTCGAGGTC.... AGTCA ACCG....
3 AACATGCCA.... TTCGGGGTC.... AGTCA ACCG....
4 AACACGCCA.... TTCGGGGTC.... AGTCGACCG....

It's responsible for the construction and maintenance of organisms



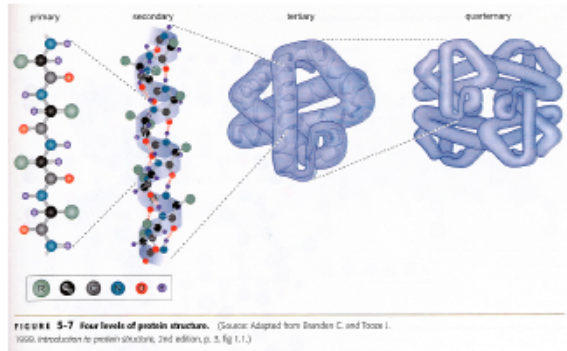


FIGURE 5-7 Four levels of protein structure. (Source: Adapted from Branden C. and Tooze J. 1999. Introduction to protein structure, 2nd edition, p. 5, fig 1.1.)

Credit: Watson et al., *Molecular Biology of the Gene*, CSHL Press, 2004

Note: other regions of genomes can impact phenotypes...

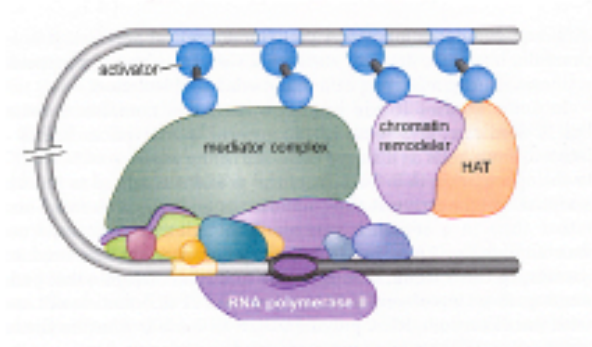


FIGURE 12-16 Assembly of the pre-initiation complex in presence of Mediator, nucleosome modifiers and remodelers, and transcriptional activators. In addition to the general transcription factors shown in Figure 12-15, transcriptional activators bound to sites near the gene recruit nucleosome modifying and remodeling complexes, and the Mediator Complex, which together help form the pre-initiation complex.

Statistics and probability I

- **Quantitative genomics** is a field concerned with the ***modeling*** of the relationship between *genomes* and *phenotypes* and using these models to ***discover*** *and* ***predict***
- We will use frameworks from the fields of probability and statistics for this purpose
- Note that this is not the only useful framework (!!) - and even more generally - mathematical based frameworks are not the only useful (or even necessarily "the best") frameworks for this purpose

Statistics and probability II

- A non-technical definition of probability:
  a mathematical framework for modeling under uncertainty
- Such a system is particularly useful for modeling systems where we don't know and / or cannot measure critical information for explaining the patterns we observe
- This is exactly the case we have in quantitative genomes when connecting differences in a genome to differences in phenotypes

Statistics and probability III

- We will therefore use a probability framework to model, but we are also interested in using this framework to discover and predict
- More specifically, we are interested in using a probability model to identify relationships between genomes and phenotypes using DNA sequences and phenotype measurements (=Data)
- For this purpose, we will use the framework of *statistics*, which we can (non-technically) define as a system for interpreting data for the purposes of prediction and decision making given uncertainty

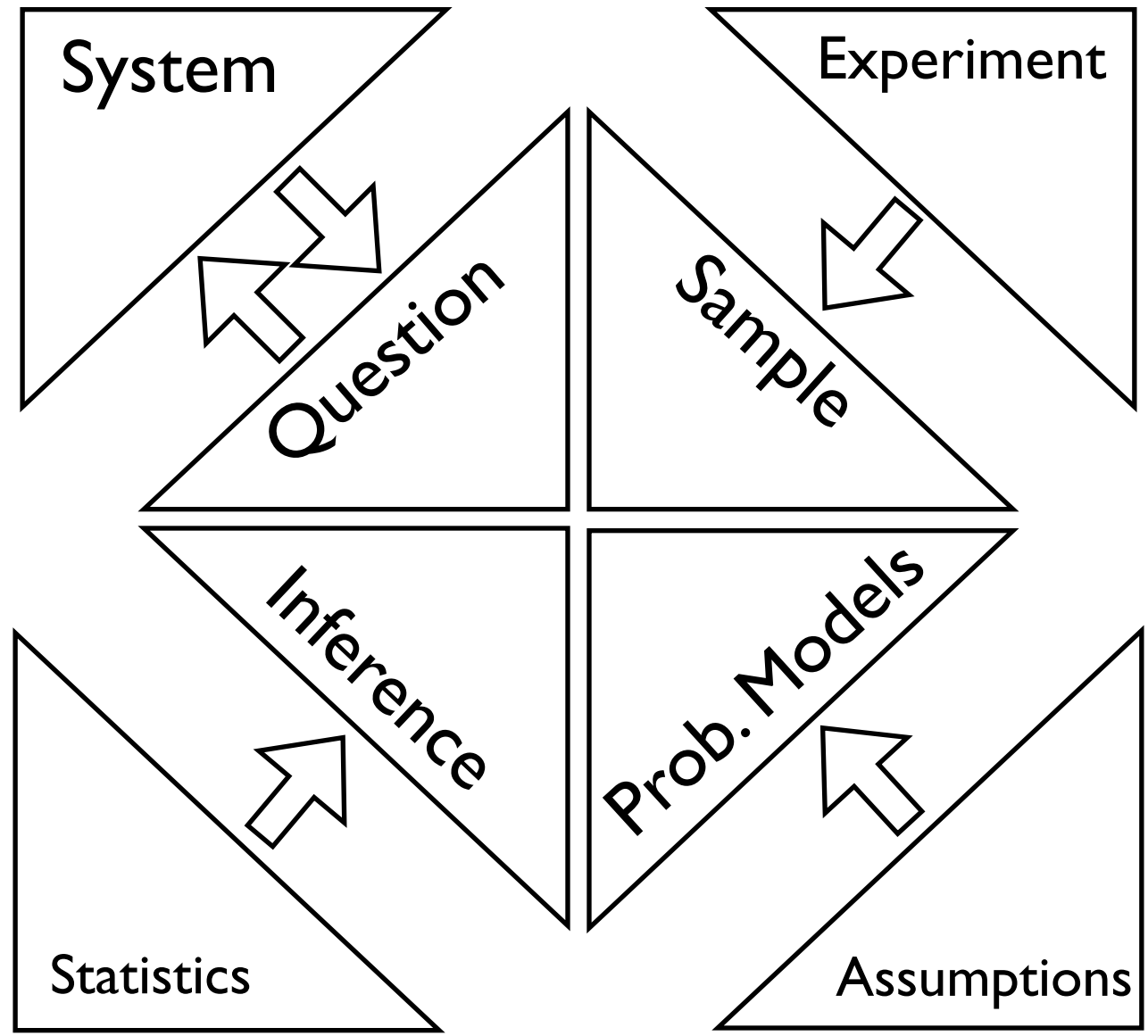
Definitions: Probability / Statistics

- **Probability** (non-technical def) - a mathematical framework for modeling under uncertainty

- **Statistics** (non-technical def) - a system for interpreting data for the purposes of prediction and decision making given uncertainty

These frameworks are particularly appropriate for modeling genetic systems, since we are missing information concerning the complete set of components and relationships among components that determine genome-phenotype relationships

Conceptual Overview



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"