Quantitative Genomics and Genetics

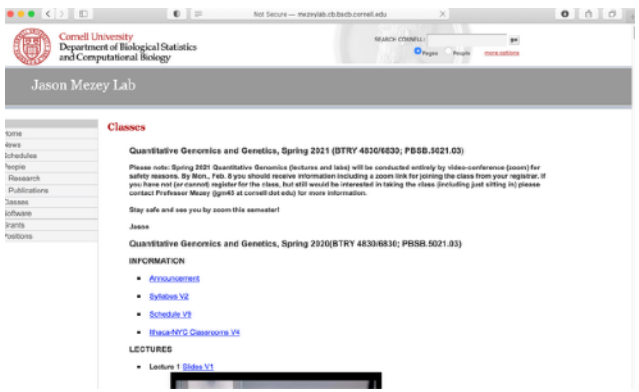
BTRY 4830/6830; PBSB.5201.03

Lecture 2: Introduction genetic concepts & probability basics

Jason Mezey
jgm45@cornell.edu
Feb. 11, 2021 (Th) 8:05-9:20

Announcements 1

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



- The first lecture (from Tues.) is currently only available for download for Cornell Ithaca students - we are aware of the problem and are working to fix…

- In the meantime, please pay attention to logistic requests (!!) = see next slide and see videos from Spring 2020 (same material!)

Announcements II

- PLEASE SIGN UP ON PIAZZA TODAY (!!!): piazza.com/cornell/spring2021/btry4830btry6830
- If you cannot sign up yourself, email me (jgm45@cornell.edu), Beulah (baa95@cornell.edu), or Scott (sdk2004@cornell.edu) and we will get you on it
- Sign up even if you are just "Sitting-In" or auditing the class (=all key communication will be done through PIAZZA)!

Announcements III

- CMS (!!) is how we will distribute / grade work for the class - if you are registered (or just sitting in and want to see the problem sets, exams, etc even if you want to do them) you must get on CMS now
- Everyone in Cornell, Ithaca who is officially registered for the course, should be able to get on the class CMS now (please email me if you cannot)
- Everyone else MESSAGE ME FROM PIAZZA (=do not email me directly) that you need to get up on CMS and PLEASE DO THIS TODAY (!!) - and please make sure I can see your name / include your email

Announcements V

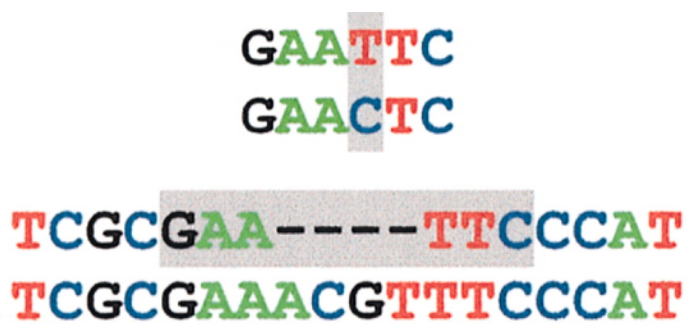
- Please officially REGISTER for this course IF YOU CAN (= if you just want to sit-in and not do the work please register for an Audit - you will automatically get Audit credit for the course)
- In general, my office hours will (most likely) be on Mon. 2-4PM but I will NOT hold office hours this Mon. (Feb. 15)!
- Other questions? Message me on Piazza…

Summary of lecture 2: 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)
- We will also begin our discussion of probability…

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?

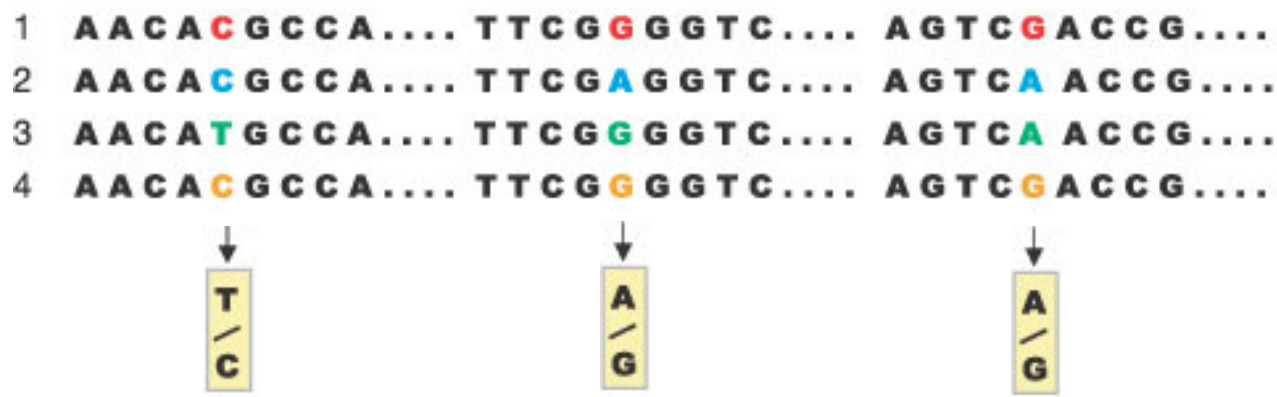
An illustration

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...

Quantitative genetics and connection to other disciplines

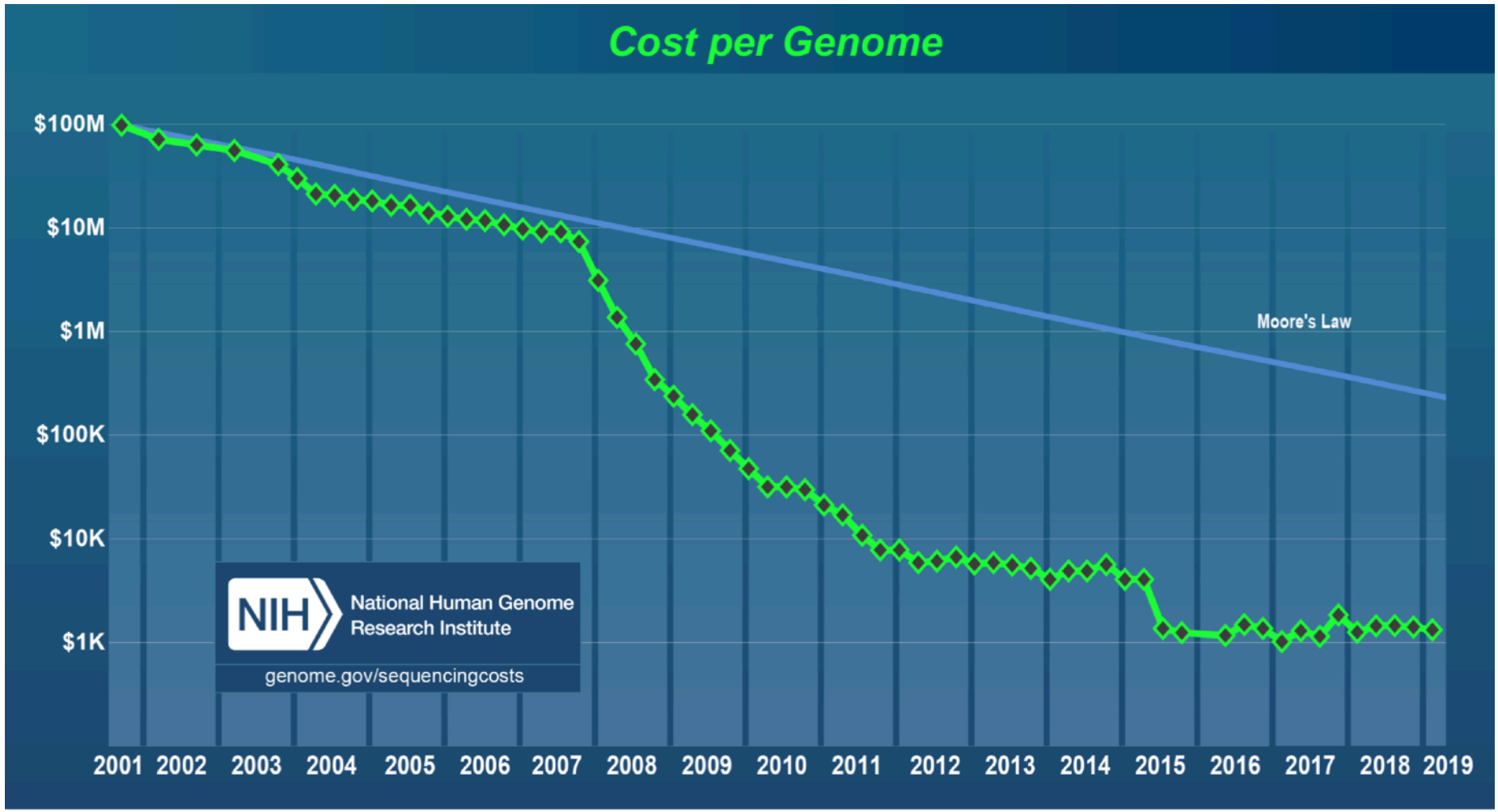
- Broad Classification of Fields of Genetics:
  - *Modeling Genetic Fields:* quantitative genetics; system genetics; population genetics; etc.
  - *Mechanism Genetic Fields:* Molecular Genetics; Cellular Genetics; etc.
  - *Model System Genetic Fields:* Human Genetics; Yeast Genetics; etc.
  - *Subject Genetic Fields:* Medical genetics; Developmental Genetics; Evolutionary Genetics; Agricultural Genetics, etc.
- **Quantitative genomics** is a field concerned with the *modeling* of the relationship between genomes and phenotypes and using these models to *discover and predict*

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

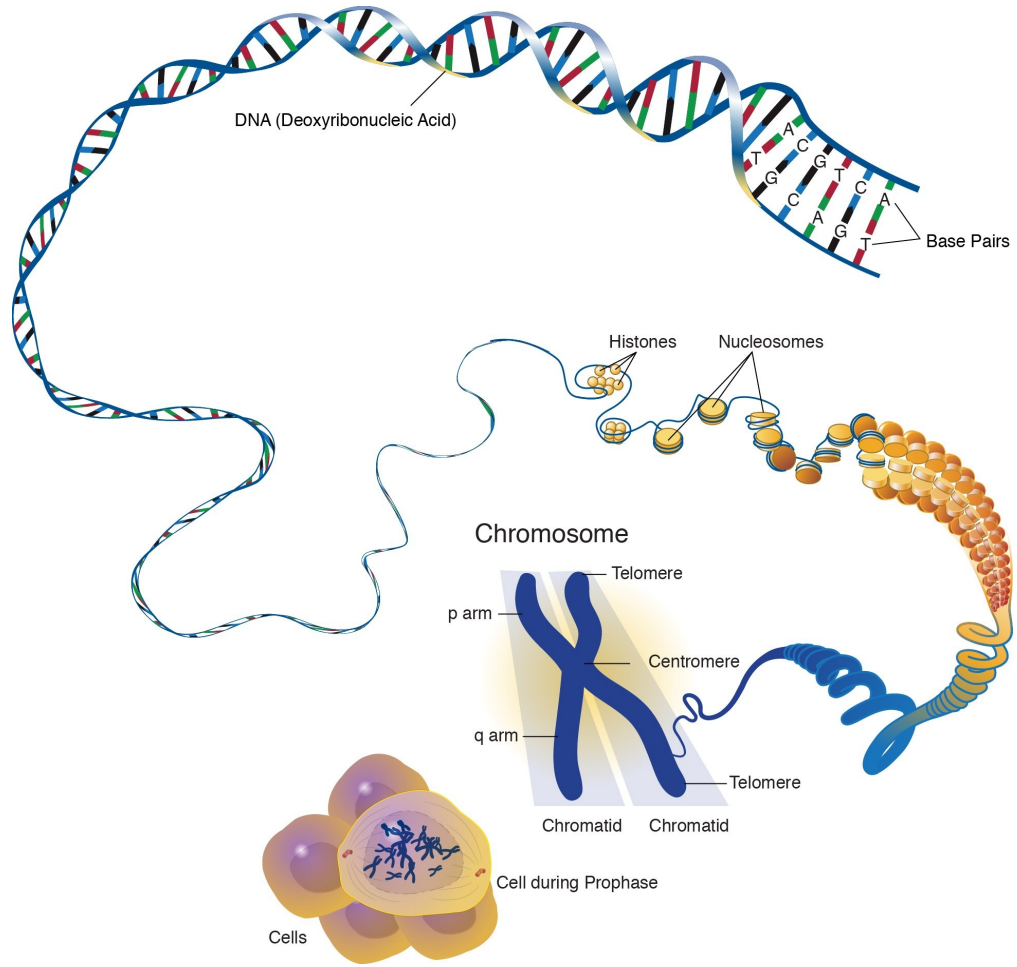
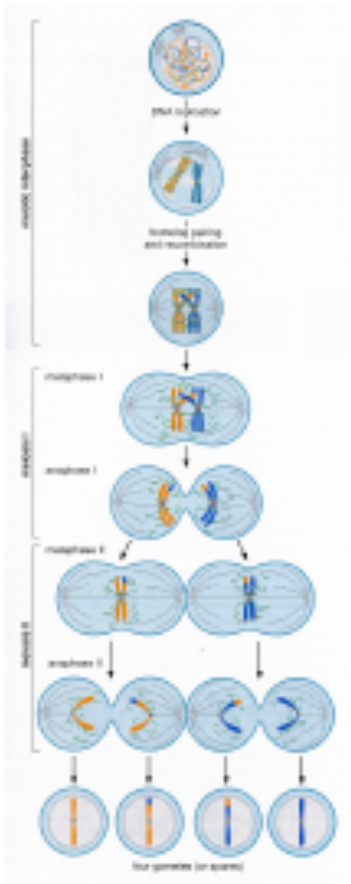
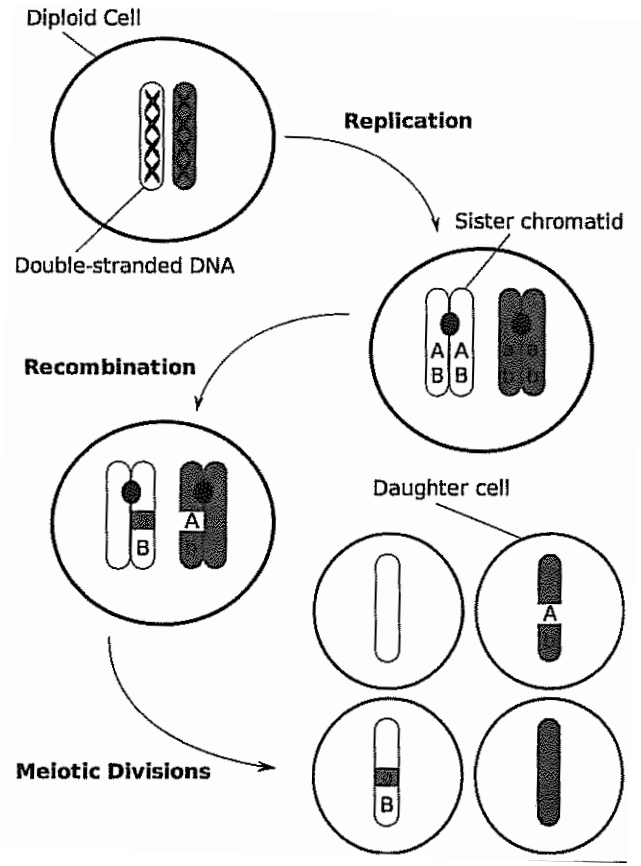


Figure 1: A simplified schematic showing genome organization in human cells. The DNA of a genome is located within the nucleus of a cell. The genome is organized in long strings that are tightly coiled around protein structures to form chromosomes. Each string is a double helix where the building blocks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

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

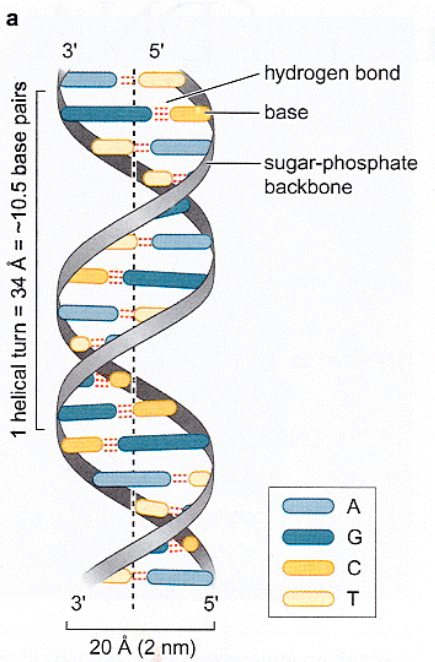
It's passed on to the next generation

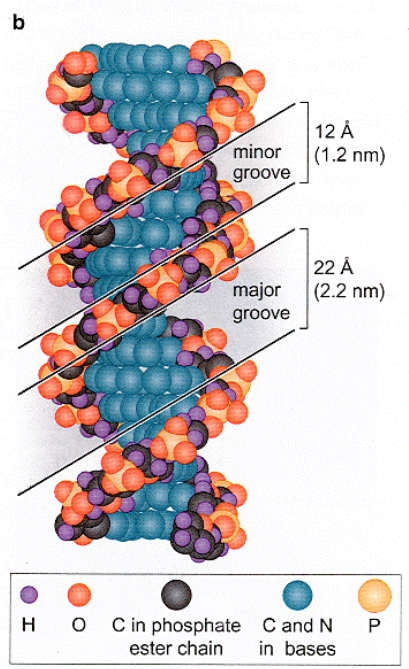


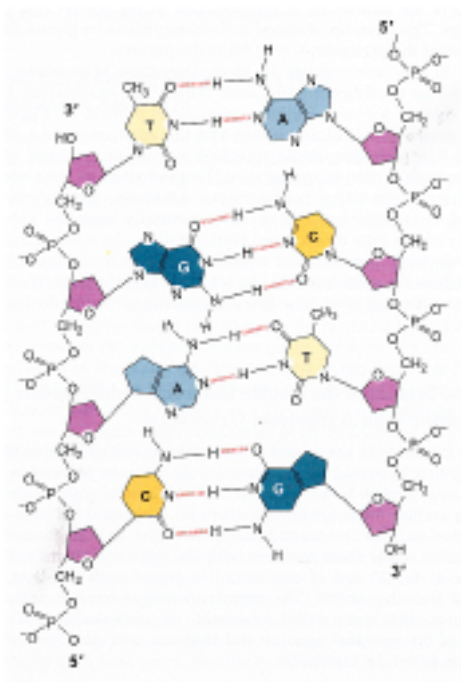


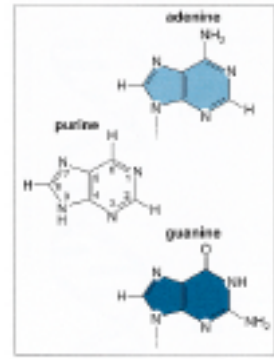
Credit: Watson et al., *Molecular Biology of the Gene*, CSHL Press, 2004

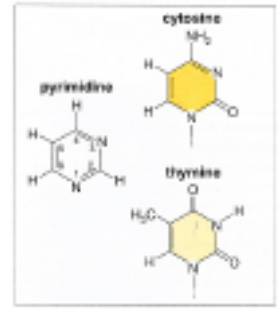
It has convenient structure for quantifying differences









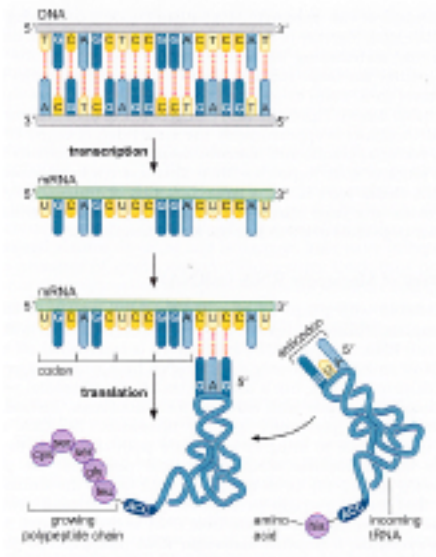


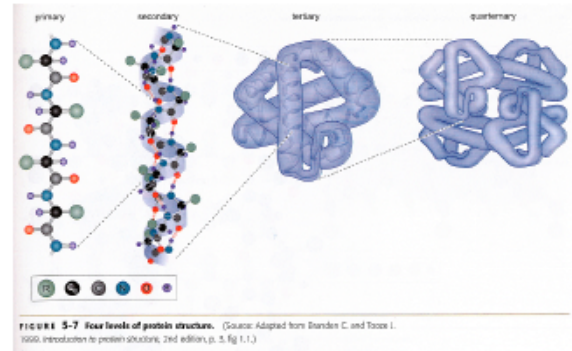
Credit: Watson et al., *Molecular Biology of the Gene*, CSHL Press, 2004

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





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

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

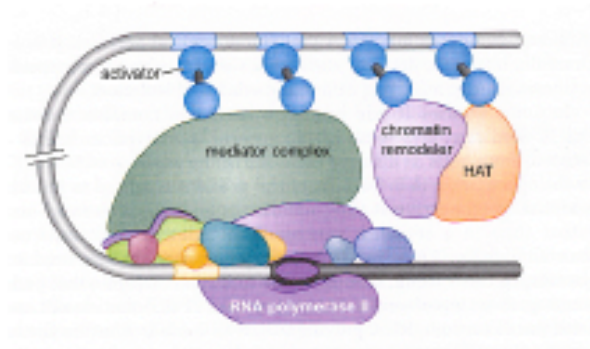


FIGURE 12-16 Assembly of the pre-initiation complex in presence of Mediator, nucleosome modifiers and remodelers, and transcriptional activators. In addition to the general transcription factors shown in figure 12-15, transcriptional activators bound to sites near the gene recruit nucleosomes modifying 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

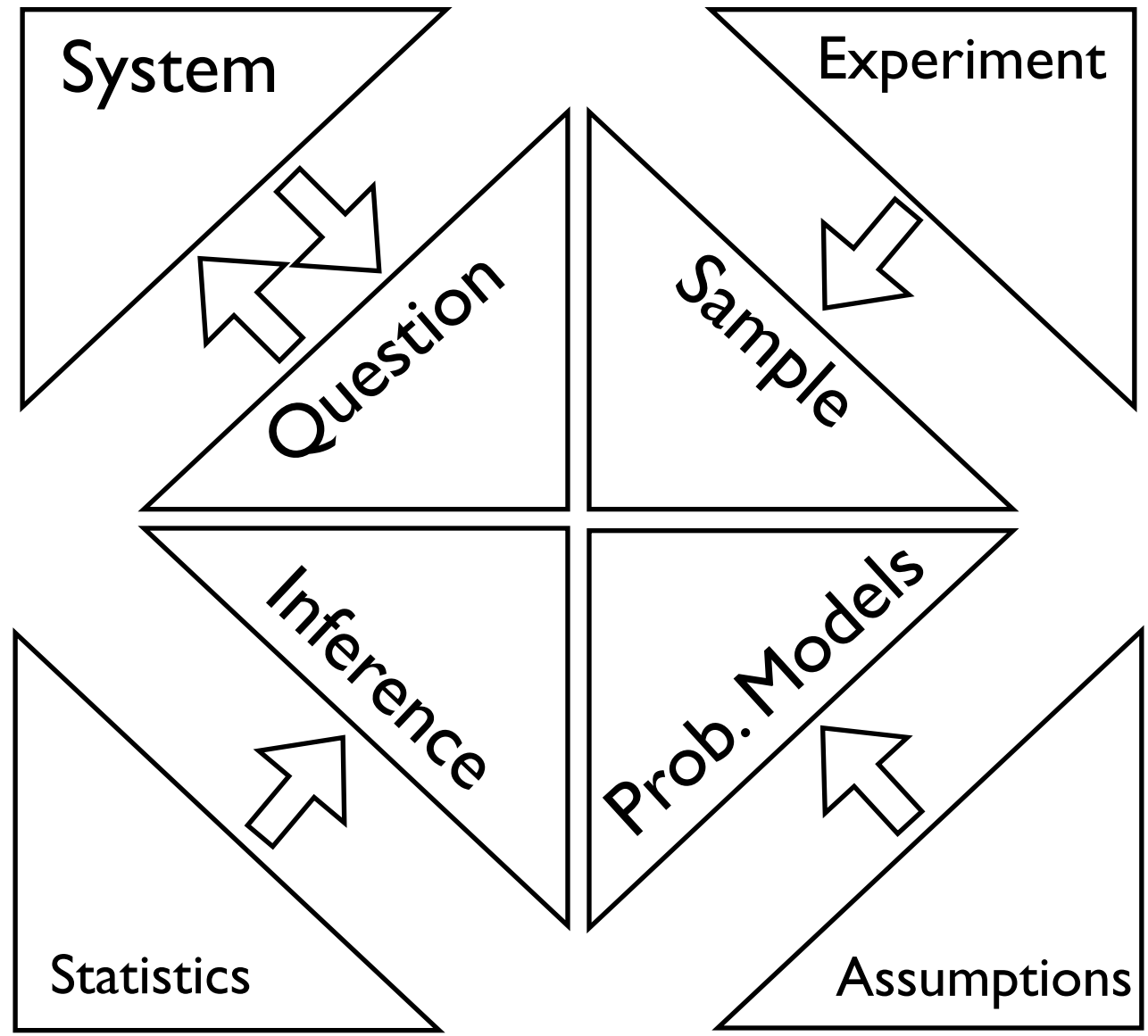
Definitions: Probability / Statistics

- **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

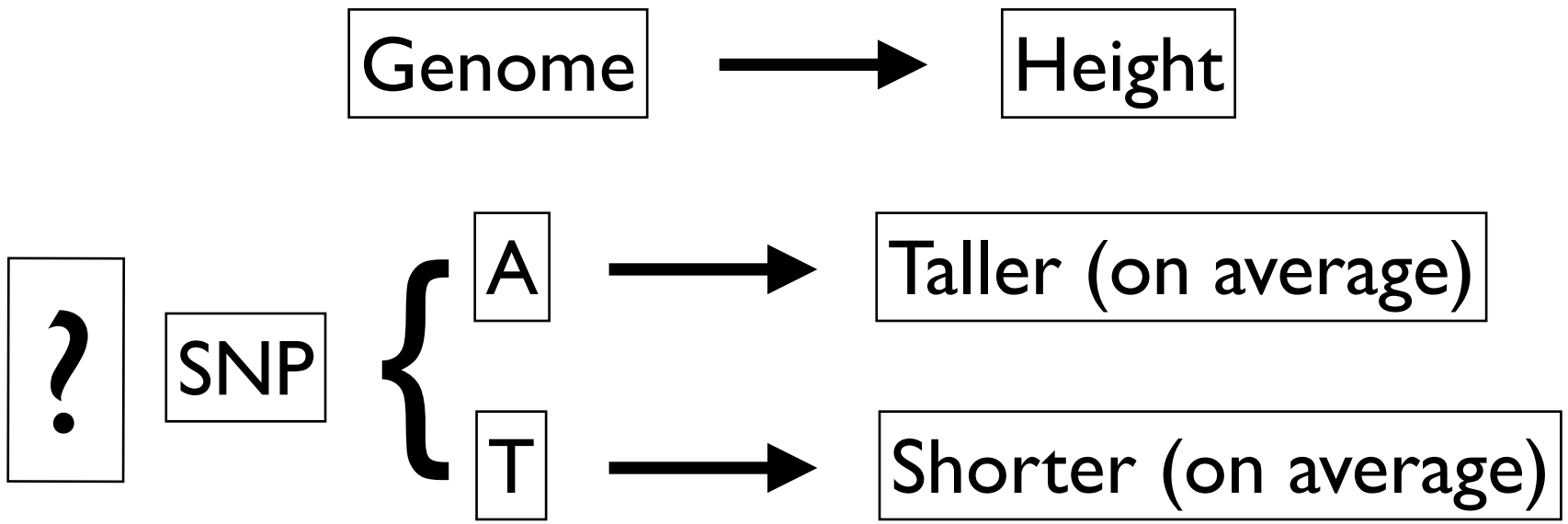
Conceptual Overview



Starting point: a system

- **System** - a process, an object, etc. which we would like to know something about

- Example: Genetic contribution to height



Starting point: a system

- **System** - a process, an object, etc. which we would like to know something about

- Examples: (1) coin, (2) heights in a population

Coin - same amount of metal on both sides?

Heights - what is the average height in the US?

Experiments (general)

- To learn about a system, we generally pose a specific question that suggests an experiment, where we can extrapolate a property of the system from the results of the experiment
- Examples of "ideal" experiments (System / Experiment):
  - SNP contribution to height / directly manipulate A -> T keeping all other genetic, environmental, etc. components the same and observe result on height
  - Coin / cut coin in half, melt and measure the volume of each half
  - Height / measure the height of every person in the US

Experiments (general)

- To learn about a system, we generally pose a specific question that suggests an experiment, where we can extrapolate a property of the system from the results of the experiment
- Examples of "non-ideal" experiments (System / Experiment):
  - SNP contribution to height / measure heights of individuals that have an A and individuals that have a T
  - Coin / flip the coin and observe "Heads" and "Tails"
  - Height / measure some people in the US

Experiments and samples

- **Experiment** - a manipulation or measurement of a system that produces an outcome we can observe
- **Experimental trial** - one instance of an experiment
- **Sample outcome** - a possible outcome of the experiment
- (Note: **Sample** - results of one or more experimental trials)
- Example (Experiment / Sample outcomes):
  - Coin flip / "Heads" or "Tails"
  - Two coin flips / HH, HT, TH, TT
  - Measure heights in this class / 1.5m, 1.71m, 1.85m, ...

That's it for today

- Next lecture, we will continue our discussion of probability!