Introduction

- Some of the objectives for genetic studies include:
  - Identify the genetic causes of phenotypic variation
  - Have better understanding of human evolution
  - Drug development: finding genes responsible for a disease provides valuable insight into how pathways could be targeted
- Recent decades have produced major advances in the science of genetics
- The amount of data available for use in genetic studies has increased astronomically
- In the past few years we have seen the release of the first drafts of the entire human genome and the genomes of model organisms.

Challenges of Human Genetics

- The most notable experiments have unequivocal interpretation:
  - Unequivocal interpretation is rare in human genetics
  - Generally can not design the perfect experiment: have to work with data we have at our disposal
  - Interpretation is of the greatest importance
- How do our data and results inform us with respect to the fundamental questions we are trying to address?
- What are the alternative interpretations of our data?
- Is it possible to distinguish among these alternatives?
- With so much data and so many options, there is a pressing need for well-designed studies and accurate and efficient statistical methods.
- Relative to experimental methods, analysis if fast and inexpensive

The Need for Experimental Design and Statistics

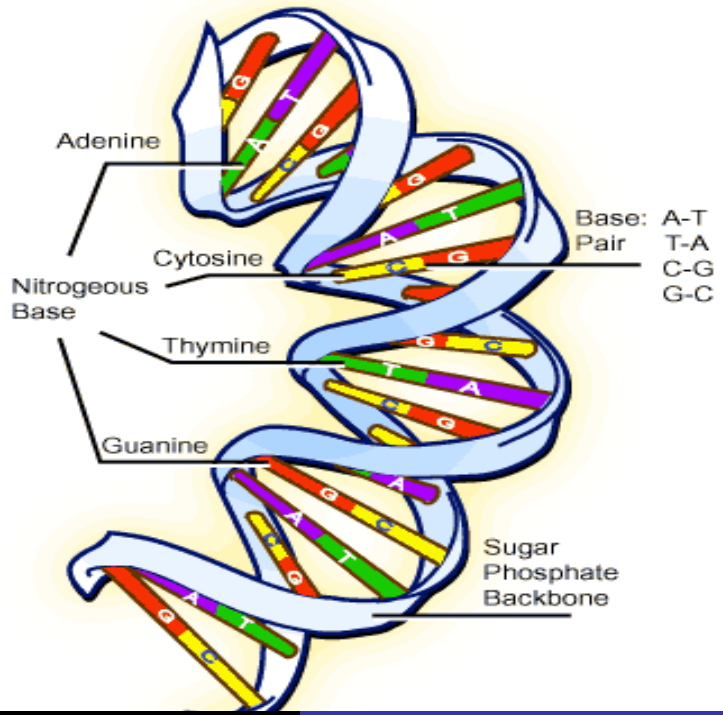
- Considering the cost of collecting family information and conducting molecular genetic experiments, we are obligated to get everything we can out of all of the data that we have at our disposal.
- Our goal for the quarter will be
  - study potential designs that incorporate genetic data
  - learn the corresponding methods for analyzing data from these designs
- Our goals in these tasks will be to:
  - understand the basic idea of each type of study
  - know the assumptions each type of analysis depends on for validity
  - understand the limitations of different types of studies
  - learn how to correctly interpret study results

The basic structure of a gene

- It is well established that human characteristics are inherited from parents to offspring in discrete unites called genes.
- Vast amount of info regarding the precise molecular mechanisms of genetic transmission from parent to offspring.
- A **gene** is the most fundamental unit of heredity that controls the transmission and expression of one or more traits.
- The chemical structure of a gene is deoxyribonucleic acid (DNA).

The basic structure of a gene

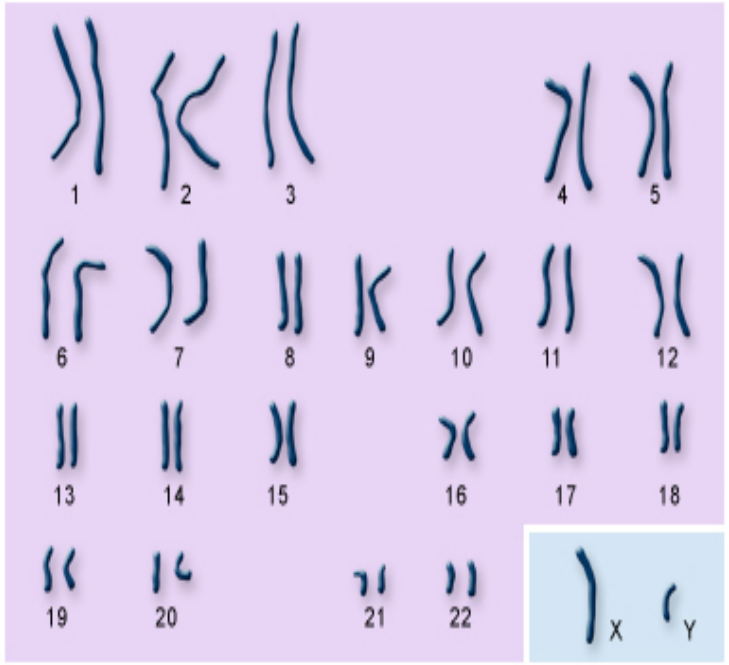
- A gene can be viewed as a two long strands of DNA which are normally bound to each other lengthwise by hydrogen bonds and are twisted around each other as a double helix.
- The subunits are called nucleotides which contain the nitrogenous bases
- There are four different nitrogenous bases, called adenine (A), guanine (G), cytosine (C), and thymine (T).



Chromosomes

- We can think of DNA as a sequence of the four letters A, G, C, and T.
- An important feature of DNA is that A is always paired with T, and G is always paired with C.
- Genes are found on chromosomes in the nucleus of cells.
- **Chromosomes** are very long strands of DNA.
- Every species has its own characteristic number of different chromosomes.
- Humans have 23 pairs of chromosomes, 22 autosomes and 1 pair of sex chromosomes.

Chromosomes



U.S. National Library of Medicine

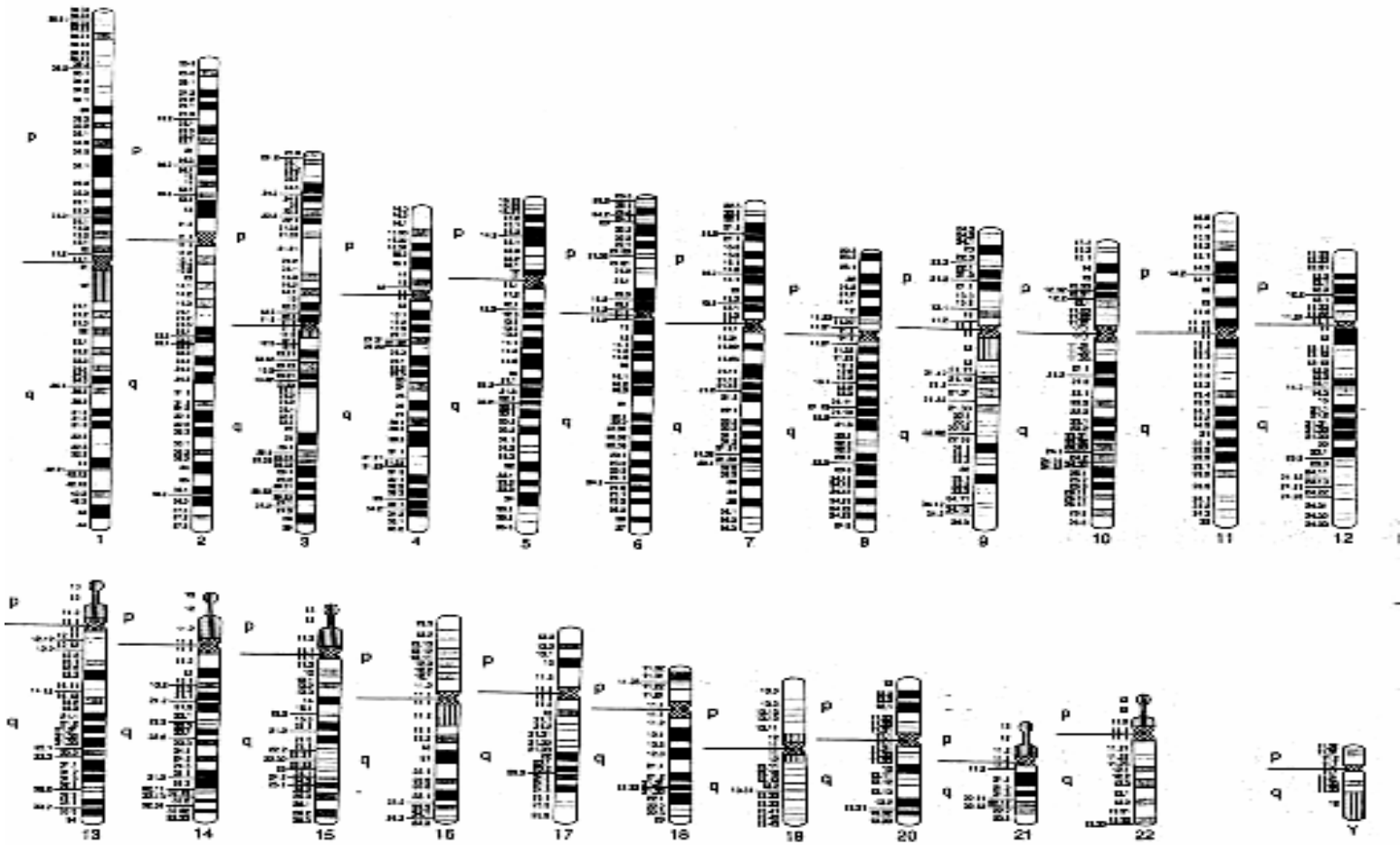
- The 22 autosomal chromosomes are numbered in order of decreasing length from 1 to 22.
- For every pair of chromosomes, one is inherited from the mother of an individual and one is inherited from the father of an individual.

Chromosomes

- Chromosomes that are of the same pair and carry the same set of genes and are called **homologous**. (e.g. both chromosome 21)
- **Mitosis** is cell division that yields two identical diploid cells, both of which have two pairs of each chromosome.
- **Meiosis** is a special type of cell division that happens in reproductive tissue yielding haploid cells (which have one of each chromosome) called **gametes**. In females, the gametes are the egg cells and in males the gametes are the sperm cells.
- The **centromere** is a region of the chromosome that is the attachment site for the spindle fiber that moves the chromosome during cell division.
- The centromere defines two arms of the chromosome, the short arm **p** and the long arm **q**.
- When treated with special stains, each arm appears to be divided into a number of bands, which are numbered from the centromere.

Chromosomes

- The approximate location of a gene is often specified by the chromosome number (i.e. 1,2,...,22,X,Y), the arm (p,q), and the band (1,2,3,...).

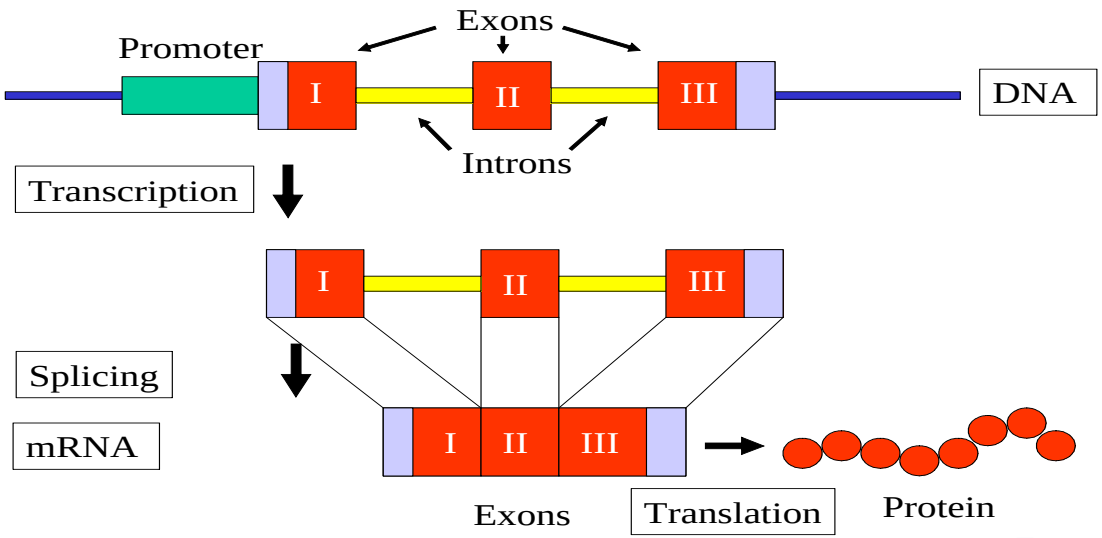


The Human Genome

- The entire DNA characteristics of a species is called its genome.
- The human genome has about 3 billion base pairs per haploid.
- Approximately 2% of the human genome is coding and 98% of the human genome is non-coding.
- A gene is a sequence of DNA that is transcribed into mRNA (messenger RNA), which, in turn, is translated into protein.
- For RNA, uracil (U) is substituted for thymine in DNA.
- In a recent build of the human genome, annotation data are available for approximately 32,000 genes with around 18,000 confirmed genes.
- Genes vary enormously in length from less than a thousand base pairs to over a million base pairs (megabases).

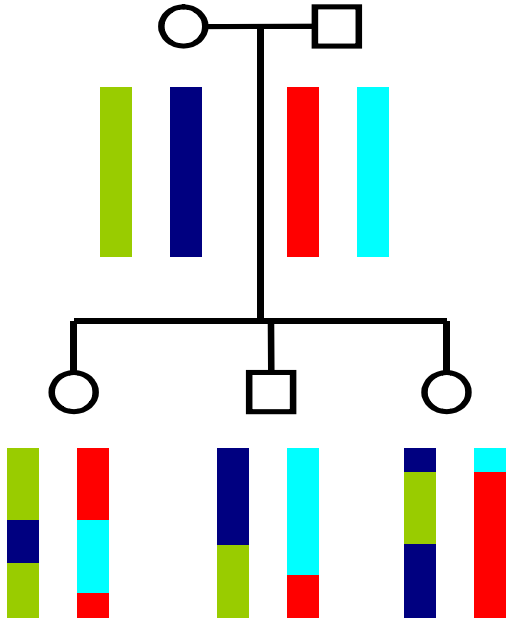
Coding Sequences

- Genes do not form a continuous sequence but consists of several coding segments called **exons** that are separated by non-coding segments called **introns**.
- Non-coding regions and introns are sometimes called "junk" DNA.
- This term can be misleading because non-coding regions may indeed have a function.
- Some non-coding regions are known to be involved in the regulation of nearby coding sequences.



Recombination

- A chromosome inherited by an offspring from a parent is actually a mosaic of the parent's two chromosomes.
- **Genetic Recombination** $\longrightarrow$ genetic material is exchanged between a chromosome of paternal origin and the corresponding chromosome of maternal origin.



Genetic and Physical Maps

- **Crossovers** are the points of exchange
- **Recombination fraction** between two loci on a chromosome is the probability that the two loci end up on regions of different origin $\longrightarrow$ occurs when the two loci are separated by an odd number of crossovers
- **Genetic Maps** $\longrightarrow$ give the order and distances (recombination fraction) between genes or genetic markers.
- **Physical Maps** $\longrightarrow$ sets of ordered markers and the physical distance (base pairs) between them

More Genetic Terminology

- More than 99 percent of loci of the DNA sequences on the 23 chromosome pairs are identical in all humans
- A **genetic marker** is a strand of DNA that is polymorphic: has some variation in the human population.
- A genetic marker can have two or more different states and we an **allele** is the state at a marker.
- Single Nucleotide Polymorphism (SNP) has two allelic types: highly abundant (1 per 1000 base-pairs)
- Short Tandem Repeats (microsattelites) GTAGTAGTAGTAGTA...
- For a chromosome pair, the two alleles at a single genetic marker are called an individual's **genotype**.
- **Homozygous** genotypes have alleles that are identical.
- Genotypes that have two different alleles are said to be **Heterozygous** .
- A **haplotype** is a sequence of alleles allong a chromosome.

Introduction to Segregation Analysis

- In the mid 1800's, Gregor Mendel demonstrated the existence of genes based on the regular occurrence of certain characteristic ratios of dichotomous characters (or traits) among the offspring of crosses between parents of various characteristics and lineages.
- These ratios are known as segregation ratios
- The analysis of segregation ratios remains an important research tool in human genetics.
- The demonstration of such ratios for a discrete trait among the offspring of certain types of families constitutes strong evidence that the trait has a simple genetic basis.

Mendelian Genetics

- Simple Mendelian disorders or traits can be adequately modeled using Mendels laws. Generally, these traits are close to completely penetrant.
- Mendel's Laws
  - **Law of Segregation (The "First Law")**: The alleles at a gene segregate (separate from each other) into different gametes during meiosis. An individual receives with equal probability one of the two alleles at gene from the mother and one of two alleles at a gene from the father.
  - **Law of Independent Assortment (The "Second Law")**: The segregation of the genes for one trait is independent of the segregation of genes for another trait, i.e., when genes segregate, they do so independently

Mendelian Genetics

- **Mode of Inheritance** is the manner in which a particular genetic trait or disorder is passed from one generation to the next.
- **Classical Mendelian experiments** use inbred strains of animals or self-fertilized plants so that individuals in each of the starting generation parental groups are homozygous at every locus and are genetically identical

Example 1: Rabbits

grey $\times$ albino

$\downarrow$

grey F1

$\downarrow$

grey : black : albino F2

9 : 3 : 4

- Proposed genetic model: color is controlled by two genes. Gene 1 controls the presence of color: alleles C and c. Gene 2 controls whether color is grey or black: alleles G and g

IBD Sharing Probabilites for Outbreds

Gene 1	Gene 2	Phenotype
CC or Cc	GG or Gg	grey rabbit
	gg	black rabbit
cc	any genotype	albino rabbit

- Parental generation groups $\longrightarrow$ homozygous for both genes
- Show how this proposed genetic model explains the observed segregation ratios of the phenotypes.

Example 1: Rabbits

$$CCGG \times ccbb$$
$$\downarrow$$
$$CcGg \qquad \text{F1}$$
$$\downarrow$$

F2	CG	Cg	cG	cg
CG	grey	grey	grey	grey
Cg	grey	black	grey	black
cG	grey	grey	albino	albino
cg	grey	black	ablino	albino

So expected phenotype relative frequencies for grey: black: albino are $9 : 3 : 4$

Example 2: Mice

grey $\times$ chocolate

$\downarrow$

grey F1

$\downarrow$

grey : black : chocolate F2

- Proposed genetic model: color is controlled by three genes. Gene 1 controls the presence of color: alleles C and c. Gene 2, with alleles G and g, and Gene 3, with alleles B and b, interact to produce the color of mice that have alleles to produce color.
- Parental generation $\longrightarrow$ homozygous CC at Gene 1

Gene 2	Gene 3	Phenotype
GG or Gg	Any genotype	grey mouse
gg	BB or Bb	black mouse
	bb	chocolate mouse

- If this model where the correct model for color, what segregation ratios of the phenotypes in the F2 generation would we expect for grey : black : chocolate?

Example 2: Mice

$$GGBB \times ggbb$$
$$\downarrow$$
$$GgBb \qquad \text{F1}$$
$$\downarrow$$

F2	GB	Gb	gB	gb
GB	grey	grey	grey	grey
Gb	grey	grey	grey	grey
gB	grey	grey	black	black
gb	grey	grey	black	chocolate

So expected phenotype relative frequencies for grey: black: albino are $12 : 3 : 1$

Example 3: Bean Flower Color

Flowers come in shades from white to purple. Quantify color: white (0) to purple (10)

$$10 \times 0$$
$$(\text{purple} \times \text{white})$$
$$\downarrow$$
$$5$$
$$\downarrow$$

color	10	9	8	7	6	5	4	3	2	1	0
relative counts											

- Proposed genetic model: color is controlled by two genes with additive effects.
- Gene 1: A=3, a=0
- Gene 2: B=2, b=0
- If this model where the correct model for color, what relative counts would we expect?

Example 3: Bean Flower Color

	AB	Ab	aB	ab
AB	10	8	7	5
Ab	8	6	5	3
aB	7	5	4	2
ab	5	3	2	0

color	10	9	8	7	6	5	4	3	2	1	0
relative counts	1	0	2	2	1	4	1	2	2	0	1

Aggregation and Segregation Analysis in Human Genetics Studies

- Aggregation and segregation studies are generally the first step when studying the genetics of a human trait.
- Aggregation studies evaluate the evidence for whether there is a genetic component to a study.
- They do this by examining whether there is familial aggregation of the trait.
- Questions of interest include
  - Are relatives of diseased individuals more likely to be diseased than the general population?
  - Is the clustering of disease in families different from what you'd expect based on the prevalence in the general population?

Aggregation and Segregation Studies

- Aggregation Study Example: Alzheimer's Disease -
  Studies based on twins have found differences in concordance rates between monozygotic and dizygotic twins. In particular, 80% of monozygotic twin pairs were concordant whereas only 35% of dizygotic twins were concordant. In a separate study, first-degree relatives of individuals (parents, offspring, siblings) with Alzheimer's disease were studied. First degree relatives of patients had a 3.5 fold increase in risk for developing Alzheimer's disease as compared to the general population. This was age-dependent with the risk decreasing with age-of-onset.

Reference: Bishop T, Sham P (2000) Analysis of multifactorial disease. Academic Press, San Diego

Aggregation and Segregation Studies

- Segregation analysis moves beyond aggregation of disease and seeks to more precisely identify the factors responsible for familial aggregation. For instance,
  - Is the aggregation due to environmental, cultural or genetic factors?
  - What proportion of the trait is due to genetic factors?
  - What mode of inheritance best represents the genetic factors?
  - Does there appear to be genetic heterogeneity?

Segregation analysis for autosomal dominant disease

- Consider a disease that is believed to by the caused by a fully penetrant rare mutant allele at an autosomal locus.
- Let D be the allele causing the disorder and let d represent be the normal allele.
- There are 9 possible mating types (can collapse to six mating types due to symmetry)
- Each of these mating types will produce offspring with a characteristic distribution of genotypes and therefore a distribution of phenotypes.
- The proportions of the different genotypes and phenotypes in the offspring of the six mating types are know as the segregation ratios of the mating types.

Segregation analysis for autosomal dominant disease

- These specific values of the segregation ratios can be used to test whether a disease is caused by a single autosomal dominant gene.
- Suppose that a random sample of matings between two parents where one is affected and one is unaffected is obtained
- Out of a total of n offspring, r are affected.
- Since autosomal dominant genes are usually rare, it is reasonable to assume that the frequency of allele D is quite low and that most affected individuals are expected to have genotype of Dd instead of DD.
- What are the matings in the sample under this assumption?
- How can we test if the observed segregation ratios in the offspring are what is expected if the disease were indeed caused by an autosomal dominant allele?
- The Binomial distribution can be used to model this data.

The Binomial Distribution

The binomial distribution is a very common discrete probability distribution that arises in the following situation:

- A fixed number, n, of trials
- The n trials are independent of each other
- Each trial has exactly two outcomes: "success" and "failure"
- The probability of a success, p, is the same for each trial

If X is the total number of successes in a binomial setting, then we say that the probability distribution of X is a **binomial distribution** with parameters n and p: $X \sim B(n, p)$

$$P(X = x) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

Segregation analysis for autosomal dominant disease

- Let X be the number of offspring that are affected.
- Under the null hypothesis, X will have a binomial distribution

$$P(X = x) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

  where p is the probability that an offspring is affected.
- We are interested in testing
  - H_0: $p = \frac{1}{2}$ vs. H_a: $p \neq \frac{1}{2}$
- Out of a total of n offspring, r are affected. The p-value is the probability of observing a value at least as extreme as r. If $r < \frac{n}{2}$, the p-value is

$$\sum_{x=0}^{r} \binom{n}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{(n-x)} + \sum_{x=n-r}^{n} \binom{n}{x} \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{(n-x)}$$

$$= \left(\frac{1}{2}\right)^{n-1} \sum_{x=0}^{r} \binom{n}{x}$$

Autosomal dominant disease example

- Marfan syndrome, a connective tissue disorder, is a rare disease that is believed to be autosomal dominant (and actually is!).
- 112 offspring of an affected parent and an unaffected parent are sample
- 52 of the offspring are affected and 60 are unaffected
- Are these observations consistent with an autosomal dominant disease.
- The p-value is

$$= \left(\frac{1}{2}\right)^{112-1} \sum_{x=0}^{52} \binom{112}{x} = 0.5085$$

- What if only 42 of the offspring are affected?

$$= \left(\frac{1}{2}\right)^{112-1} \sum_{x=0}^{42} \binom{112}{x} = 0.0104$$

Normal Approximation to Binomial

- If $X \sim B(n, p)$, and n is large enough such that

$$np \geqslant 10 \qquad \text{and} \qquad n(1-p) \geqslant 10$$

- Then X is approximately $N\left(\mu_X = np, \sigma_X = \sqrt{np(1-p)}\right)$
- For the Marfan syndrome data with 52 offspring affected,

$$z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{52.5 - (112)(.5)}{\sqrt{112(.5)(.5)}} = -.661$$

  P-value is $2P(Z \geqslant |z|) = 2(0.2539) = .5079$, where Z follows a standard normal distribution
- For the Marfan syndrome data with 42 offspring affected, the p-value is .0107.

Segregation analysis for autosomal recessive disease

- How can you do segregation analysis to test if a disease that is fully penetrant autosomal recessive?
- For this model we know that affected individuals are DD, but unaffected individuals could be Dd or dd.
- One proposal is to look at the segregation ratios in families with at least one affected individual. What are some problems with this proposal?

Patterns of Inheritance

- Exercise: Characterize the pattern of inheritance one would expect to see in a pedigree for autosomal dominant and recessive genes. Do the same for x-linked inheritance. Assume full penetrance.
  - Dominant autosomal
  - Recessive autosomal
  - Dominant X-linked
  - Recessive X-linked

Genetic Models

Genetic Models

- Single major locus: Simple Traits
  - Dominant model
  - Recessive model
  - Additive
  - Multiplicative
- Multifactorial/polygenic: Complex Traits
  - Multifactorial (many factors)
  - polygenic (many genes)
  - Generally assumed that each of the factors and genes contribute a small amount to phenotypic variability
- Mixed model - single major locus with a polygenic background

Single Major Locus

- A single gene, usually assumed to have only 2 alleles, contributes to the phenotypic variability
- Let's consider a dichotomous trait (or binary trait) where an individual can be either affected or unaffected

Single Major Locus Parameters

- q_1 = frequency of allele increasing risk of disease, where $q_1 + q_2 = 1$
- Penetrance parameters
  - f_{11} = probability of being affected given 11 genotype
  - f_{12} = probability of being affected given 12 genotype
  - f_{22} = probability of being affected given 22 genotype
- K_p=population prevalence of the disease
- $K_p = q_1^2 f_{11} + 2 q_1 q_2 f_{12} + q_2^2 f_{22}$
- Genotype Relative Risk - It is common to represent the risk of a genetic variants relative to the average population
  - $R_{11} = \frac{P(\text{affected}|11)}{K_p} = \frac{f_{11}}{K_p}$
  - $R_{12} = \frac{f_{12}}{K_p}$
  - $R_{22} = \frac{f_{22}}{K_p}$

Penetrance Parameters

- The penetrance parameters determines the model type
- Consider the following parameterization
  - $f_{11} = k$
  - $f_{12} = k - c_{12}$
  - $f_{22} = k - c_{22}$

  where $k - 1 \leqslant c_{12} \leqslant k$ and $k - 1 \leqslant c_{22} \leqslant k$, with $0 \leqslant k \leqslant 1$, $c_{12} \geqslant 0$, and $c_{22} \geqslant 0$
- What is the relationship between c_{12} and c_{13} for an additive model?
- What are the parameter values for a fully penetrant dominant disease?
- Note that if both $c_{12} = 0$ and $c_{22} = 0$, then the locus is not involved with the phenotype, and k would be equal to K_p.

Multiplicative Model

- A multiplicative model is given below
  - $f_{11} = r^2 k$
  - $f_{12} = rk$
  - $f_{22} = k$

  where with $0 \leqslant k \leqslant 1$, $r \geqslant 1$, and $0 \leqslant r^2 k \leqslant 1$

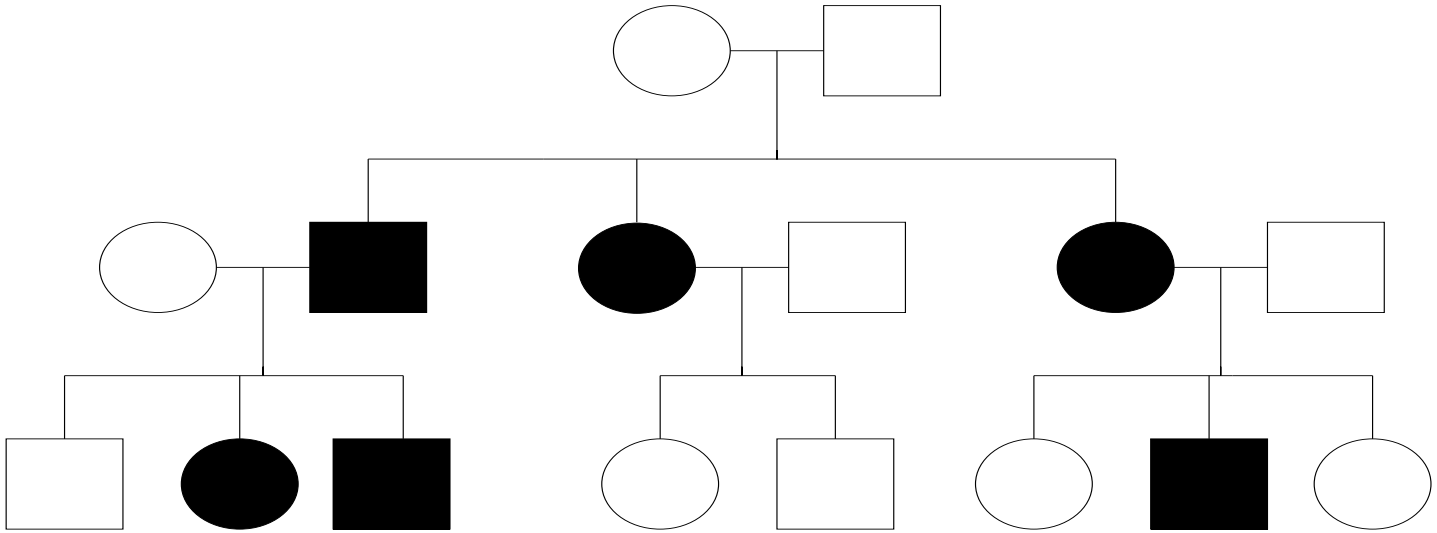
Genetic Model for Quantitative Trait

- For a dichotomous trait, a penetrance parameter is defined for each genotype as the $P(\text{trait}|\text{genotype})$.
- For a quantitative trait, Y, the penetrance function describes the distribution of the trait conditional on an individual's genotype, $f(Y|\text{genotype})$.
- Location of the heterozygote mean determines whether the allele increasing susceptibility to the disease or increasing the value of the phenotype is dominant, additive, recessive, or etc.
- Assume that the quantitative trait approximately follows a Normal distribution for each genotype group. If you compared the trait distributions for the genotype groups, what would you expect to see for the following models:
  - A quantitative trait controlled by a dominant gene:
  - A quantitative trait controlled by a recessive gene:
  - A quantitative trait controlled by an additive gene:

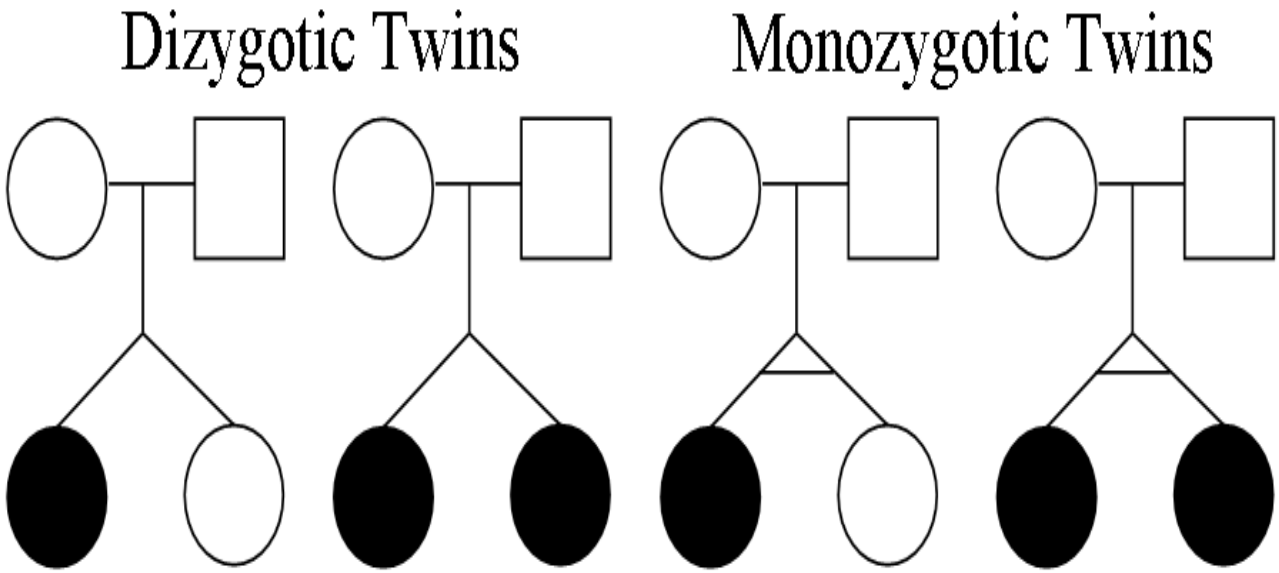
Genetic Heterogeneity

- Genetic Heterogeneity is common for complex traits,
- Genetic heterogeneity - The presence of apparently similar characters for which the genetic evidence indicates that different genes or different genetic mechanisms are involved in different pedigrees. In clinical settings genetic heterogeneity refers to the presence of a variety of genetic defects (that) cause the same disease, often due to mutations at different loci on the same gene, a finding common to many human diseases including alzheimer's disease, cystic fibrosis, and polycystic kidney disease
- Pedigree - A diagram of the genetic relationships and medical history of a family using standardized symbols and terminology
- Founder - Individuals in a pedigree whose parents are not part of the pedigree.

Extended Pedigree

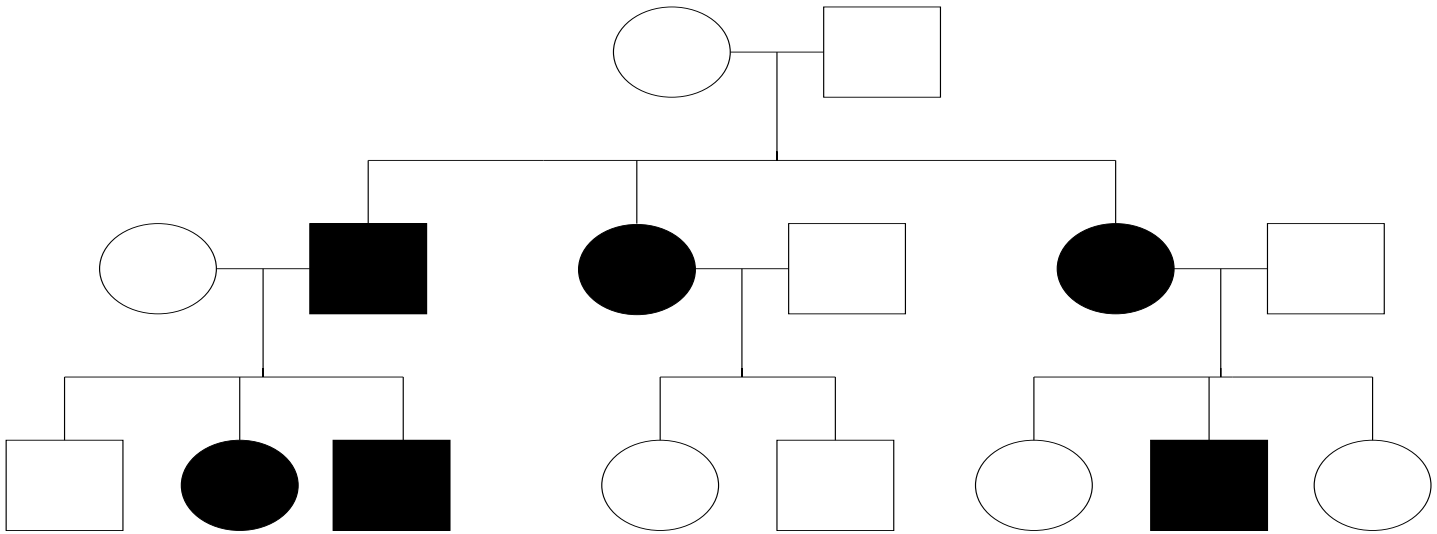


Pedigrees with Twins



Genetic Models II

Extended Pedigree with Multiple Affected Individuals



Risk Ratios

- The correlation patterns among relatives provide a simple yet powerful means of discriminating between genetic models for a trait.
- For a given genetic model, it is possible to calculate risk ratios for relatives.
- The risk ratio λ is the relative risk of individuals in a particular class of relatives to the risk of disease in the general population.
- λ is subscripted by class of relatives, e.g. λ_S for sibling risk ratio
- If i and j are siblings, then $K_S = P(i \text{ is affected} | j \text{ is affected})$ and $\lambda_S = \frac{K_S}{K_p}$
- K_p is the prevalence of the disease in the general population

Sibling Risk Example

- Consider a disease that is caused by a single mutant allele at an autosomal locus. Assume that the mutation has a dominant mode of inheritance and is fully penetrant.
- Let D be the allele causing the disorder and let d represent be the normal allele. Let the p be frequency of the D allele in the population. Assuming Hardy-Weinberg Equilibrium at the locus, what is K_S for this genetic model?
- We need to calculate the following

$$K_S = P(\text{individual is affected} | \text{ sibling is affected})$$

$$= \frac{P(\text{individual is affected } \textbf{and} \text{ sibling is affected})}{P(\text{ sibling is affected})}$$

Sibling Risk Example

- To calculate the denominator, we must figure out the probability of a diseased individual. Note that

$$P(\text{individual is affected}) = \sum_{\text{genotypes}} P(\text{affected}|\text{genotype})P(\text{genotype})$$

- What types of matings could produce a diseased child and how frequent is each mating in the population?
- For each mating type, what is the probability of producing a pair of diseased children? Note that given the parental mating type, transmissions to offspring are independent.
- Based on the answers to parts 2 and 3, we can calculate the numerator: $P(\text{both siblings affected}) =$

$$\sum_{\text{mating types}} P(\text{both siblings affected}|\text{mating type})P(\text{mating type})$$

Sibling Risk Example

Mating Type	P(Mating Type)	P(Affected Sib Pair|Mating Type)