BME 1132
Probability and Biostatistics

Instructor: Ali AJDER, *Ph.D.*

Week-8

- Genetic terms…
- Sample space,
- Tree diagram,
- Probability measures and Events
- General rules for probability
- Conditional Probability
- Summary

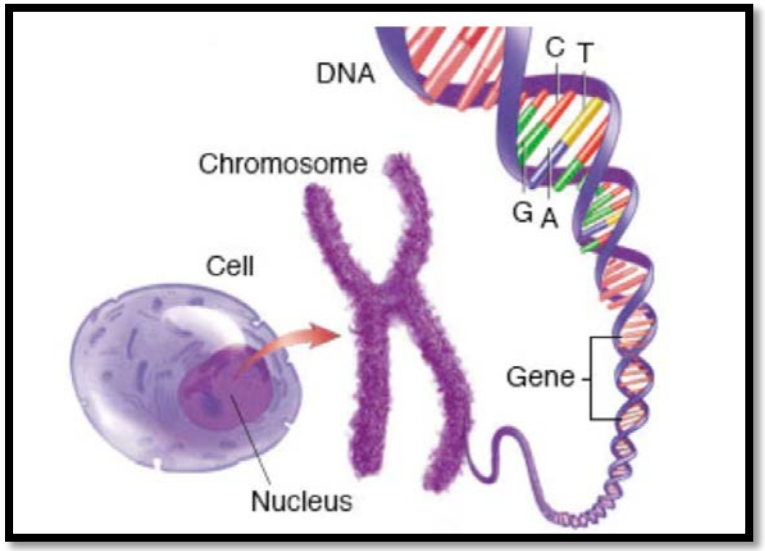
Example: Common Genetic Terms-1

A **gene** is a segment of double-stranded DNA, which itself is made of a sequence of four different nucleotides: adenine (A), guanine (G), thymine (T), or cytosine (C).

A **DNA segment** can be specified as a sequence of these four letters; for example {TAGCAAT}.

The alternate forms of a gene are called **alleles**.
{TA***G***CAAT} {TA***T***CAAT}
In this example, the alleles could be denoted as T and G.

Alleles are responsible for variation in **phenotypes**.



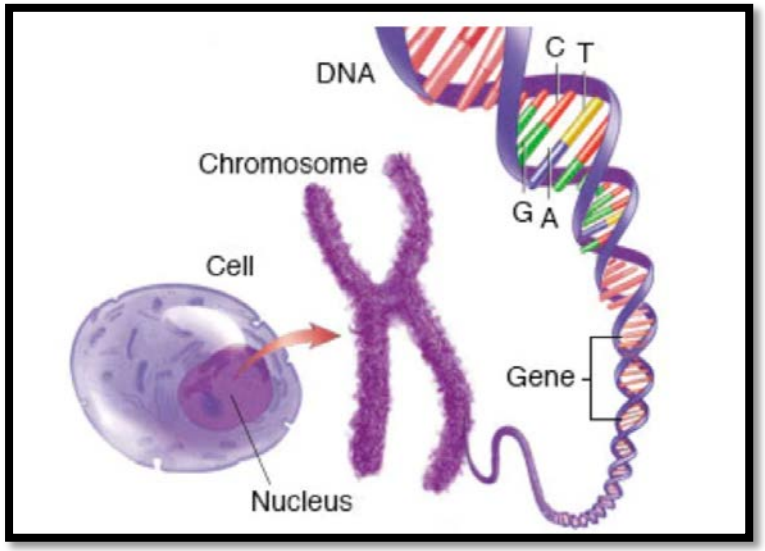
Phenotypes, in general, are observable traits, such as eye color, disease status, and blood pressure, due to genetic factors and/or environmental factors (e.g., diet, smoking, sun exposure).

Example: Common Genetic Terms-2

Genetic materials are stored on **chromosomes**.

Human somatic cells have two copies of each chromosome (one inherited from each parent); hence, they are called **diploid**.

Each pair of similar chromosomes are called **homologous chromosomes**.



The genotype (i.e., genetic makeup) of an individual for the bi-allelic gene **A** can take one of the three possible forms:

1-) *AA* **2-)** *aa* **3-)** *Aa*

The first two genotypes, AA and aa, are called ***homozygous***, which means the same version of the allele was inherited from both parents. That is, both homologous chromosomes have the same allele.

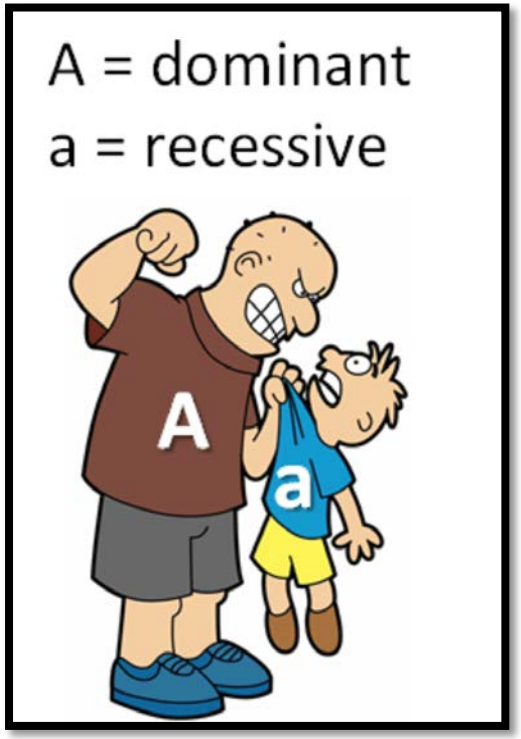
The last genotype, Aa, is called ***heterozygous***, which means different alleles were inherited.

Example: Common Genetic Terms-3

The presence of a specific allele does not always result in its corresponding trait (a characteristic such as eye color).

Some alleles are **recessive**, producing their trait only when both homologous chromosomes carry that specific variant.

On the other hand, some alleles are **dominant**, producing their traits when they appear on at least one of the homologous chromosomes.



For example, suppose that the allele a for gene **A** is responsible for a specific **disease**.
Furthermore, assume that a is a recessive allele.
Then, only a person with genotype aa will be affected by the **disease**.

Individuals with genotype AA or Aa will not have the **disease**.

Sample space

For a bi-allelic gene **A**,

The possible alleles are A and a,
and the possible corresponding genotypes are AA, Aa, and aa.

The collection of all possible outcomes is denoted S (Ω) and is called the sample space.

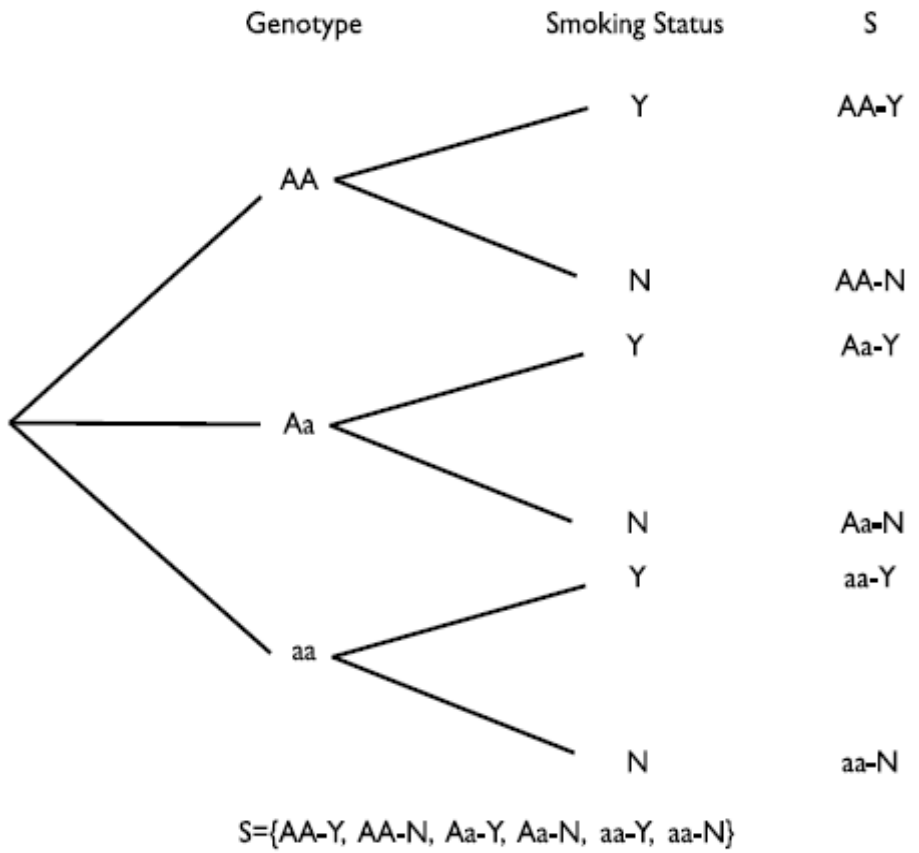
Bi-allelic gene: $S = \{A, a\}$
Genotype: $S = \{AA, Aa, aa\}$

Tree diagram

For a complex random phenomenon that is a combination of two or more other random phenomena, it might be easier to view the sample space with **tree diagrams**.
For example, suppose that we suspect that gene **A** is related to a specific disease, but genetic variation alone does not determine the **disease** status. Rather, it affects the risk of the disease. Further, we suspect that smoking (an environmental factor) is also related to the **disease**.

In this case, the random phenomenon we are interested in is the **combination** of genotype and smoking status ("Y" for smoking and "N" for not smoking).

All possible combinations (i.e., sample space) are identified using the tree diagram.



Probability measure and Events-1

The total probability of all outcomes in the sample space is always 1.

Bi-allelic gene: $P(A) + P(a) = 1$
Genotype: $P(AA) + P(Aa) + P(aa) = 1$

If we believe the two alleles are equally probable, we have

$$P(A) = P(a) = 1/2$$

We do **NOT** however assume equal probabilities for the genotypes.
To find the probability of each genotype, we first need to define **events**.

An **event** is a subset of the sample space S.

For the genotype example, $E = \{AA, aa\}$ is the event that a person is *homozygous*.

Probability measure and Events-2

In the tree diagram, the first set of branches represents the possible alleles for chromosome 1 (Ch1).
Likewise, the second set represents the possible alleles for chromosome 2 (Ch2).

By following the branches from the root to the tip, we can obtain the possible genotypes. Now, if we assume that all outcomes in the sample space are equally probable, then

$$P(A_1A_2) = P(A_1a_2) = P(a_1A_2) = P(a_1a_2) = 1/4$$

The labels for homologous chromosomes are arbitrary; we do not distinguish between genotypes A_1a_2 and a_1A_2.

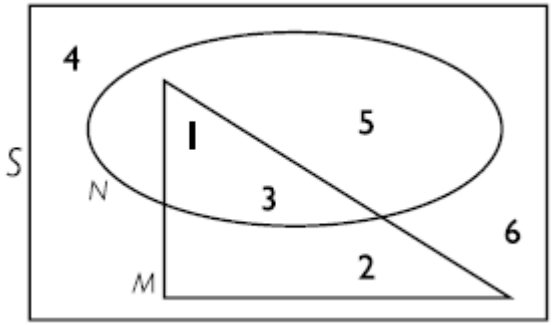
Therefore, we can create three new events: $AA = \{A_1A_2\}, Aa = \{A_1a_2, a_1A_2\}, aa = \{a_1a_2\}$.

The probabilities for these events are then

$$P(AA) = 1/4\,,$$
$$P(Aa) = 1/4 + 1/4 = 1/2$$
$$P(aa) = 1/2$$

General Rules for Probability: Example-1

Consider the die rolling example presented in the form of a Venn diagram in the Figure.



All the possible outcomes are contained inside the sample space S, which is represented by the rectangle.

We define two **events**;

The event M (shown as a triangle) occurs when the outcome is less than 4.
The event N (shown as an oval) occurs when the outcome is an odd number.

In this example, $P(M) = 1/2\,, P(N) = 1/2$

General Rules for Probability: Example-2

Now consider a bi-allelic gene **A** with two alleles A and a.
We assume that allele a is **recessive** and causes a specific **disease**. Then only people with the genotype aa have the **disease**.
We can define four events as follows:

The homozygous event: $HM = \{AA, aa\}$
The heterozygous event: $HT = \{Aa\}$
The no-disease event: $ND = \{AA, Aa\}$
The disease event: $D = \{aa\}$

$P(HM) =, ?\, P(HT) =?$
$P(ND) =?, P(D) =?$

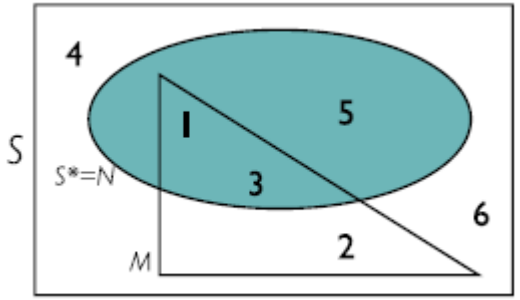
The shaded area shows the disease event (D);
whereas the unshaded area shows the no-disease event (ND).
The area with shaded border lines shows the homozygous event (HM).
The remaining part of the sample space, which includes the outcome Aa only, corresponds to the heterozygous event (HT).
Assume that the probabilities for different genotypes are;

$$P(AA) = 0.49, P(Aa) = 0.42, P(aa) = 0.09$$

Conditional Probabilities-1

The conditional probability, denoted $\boldsymbol{P(E_1|E_2)}$, is the probability of event E_1 given that another event E_2 has occurred.

Consider the die rolling example.
Recall that $P(M) = 1/2$.
Now suppose that we are told that N has occurred, that is, the outcome is in fact an odd number.
Then, the set of possible outcomes reduces to $S^* = N = \{1,3,5\}$
This new sample space is shaded in Figure.



Since the three possible outcomes, 1, 2, and 3, are still equally probable, the probability of each of them is now 1/3. Within this smaller space, the event M occurs if the outcome is either 1 or 3. (The outcome of 2 is no longer a possibility.)

These are two out of three equally probable outcomes. Therefore, the probability of M given that N has occurred (i.e., the conditional probability of M given N) is P(M|N) = 2/3.
In this case, knowing that the outcome is an odd number increased the probability of E_1 from 1/2 to 2/3.

Conditional Probabilities-2

The conditional probability of event E_1 given event E_2 can be calculated as follows: (assuming $P(E_2) \neq 0$)

$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}$$

In the die rolling example, the intersection of the two events is

$$M \cap N = \{1,3\}$$

with probability $P(E_1 \cap E_2) = 2/6 = 1/3$.
Therefore, the conditional probability of an outcome less than 4, given that the outcome is an odd number, is

$$P(M|N) = \frac{P(M \cap N)}{P(N)} = \frac{1/3}{1/2} = \frac{2}{3}$$

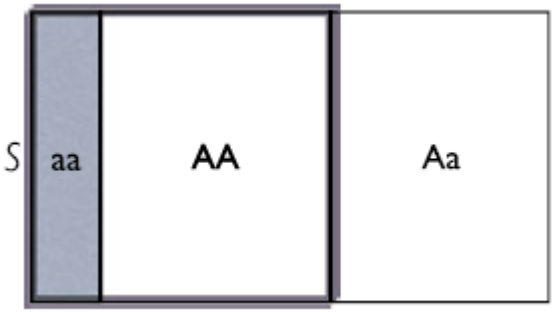
Conditional Probabilities-3

Now consider the gene-disease example.
Suppose we know that a person is homozygous and are interested in the probability that this person has the **disease**, $P(D|HM)$.

The probability of the intersection of D and HM is $P(D \cap HM) = P(\{aa\}) = 0.09$

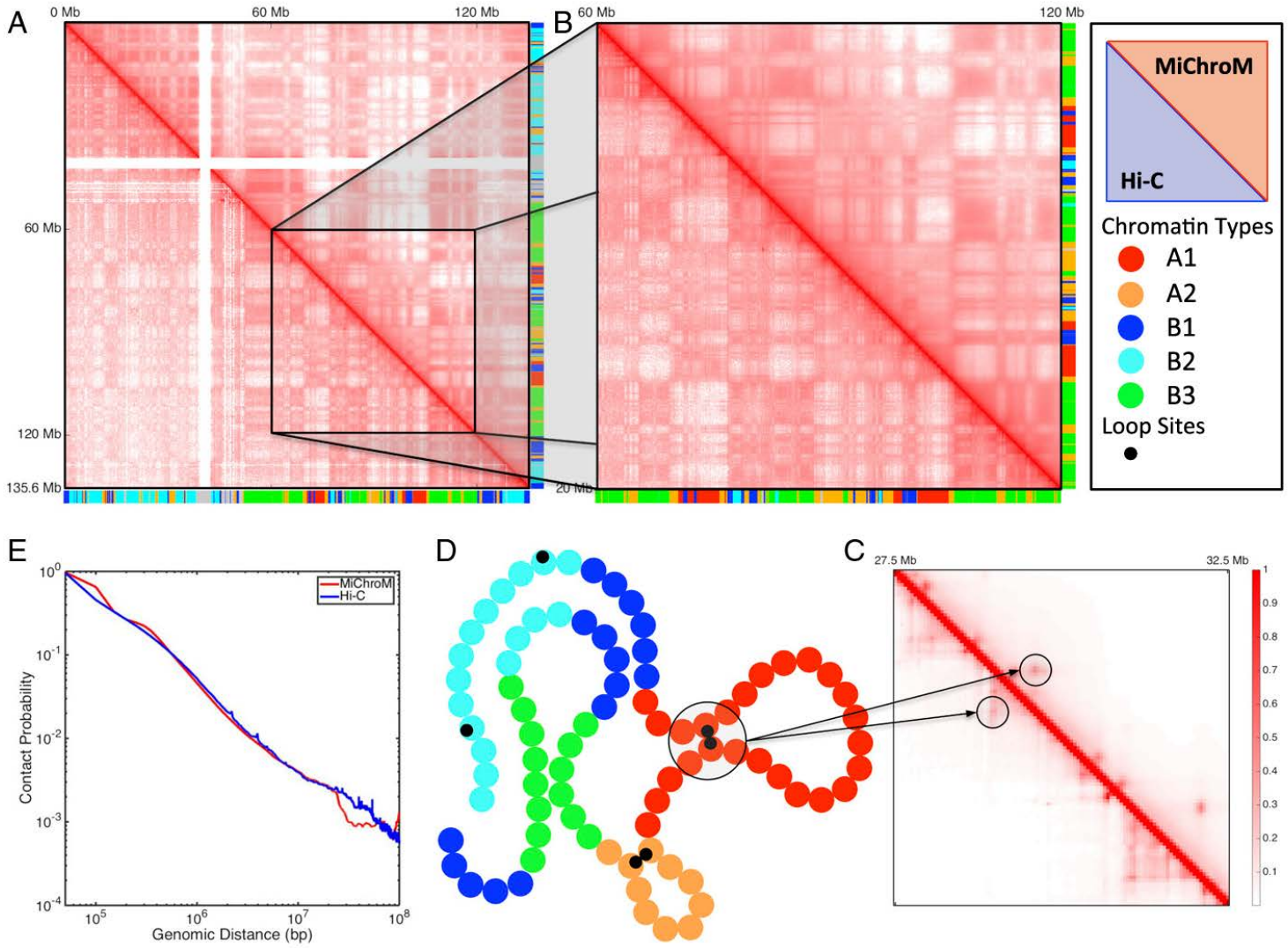
Using the formula of the conditional probability of having the disease knowing that the genotype is homozygous can be obtained as follows:

$$P(D|HM) = \frac{P(D \cap HM)}{P(HM)} = \frac{0.09}{0.58} = 0.16$$



In this case, the probability of the disease has increased from P(D) = 0.09 (the unconditional probability) to P(D|HM) = 0.16 (the conditional probability).

Questions?


A
0 Mb
60 Mb
120 Mb
60 Mb
120 Mb
135.6 Mb
B
60 Mb
120 Mb
20 Mb
MiChroM
Hi-C
Chromatin Types
A1
A2
B1
B2
B3
Loop Sites
E
MiChroM
Hi-C
Contact Probability
Genomic Distance (bp)
D
C
27.5 Mb
32.5 Mb