Review of Genetic and Genomic Concepts

JSM 2018

Thanks to Eric Sobel for many of the slides

Disclaimer

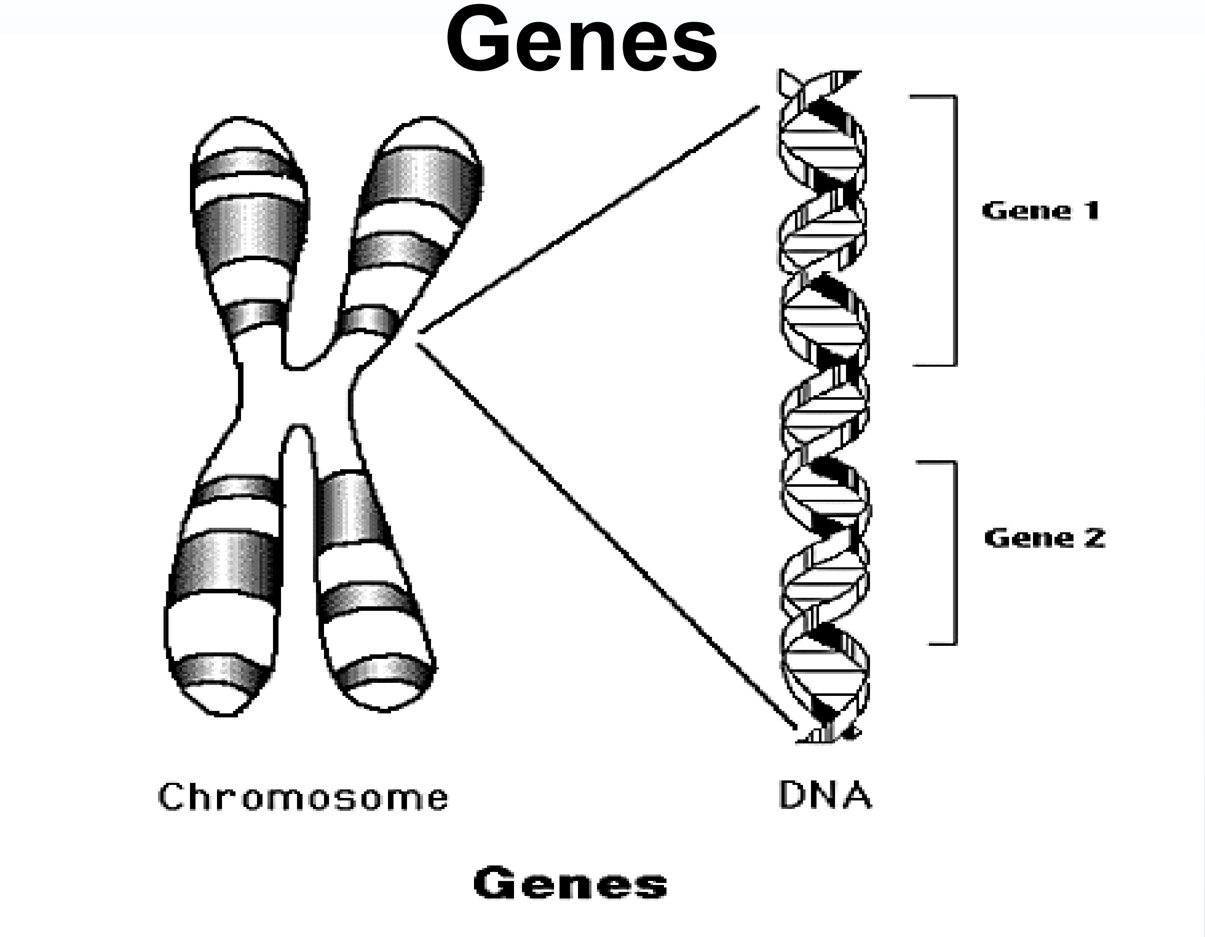
- This review is designed to make the upcoming Julia application lectures easier to understand.
- It is **NOT** a full overview of genetics and genomics.

Genetic Data

REVIEW OF DNA

- Cells are the working unit of almost every living system.
- DNA provides the instructions for these cells.
- DNA is made up of strands of sugars and phosphates where nucleotide bases have been attached (A, G, C, T).

DNA bases are the Building Blocks of Genes



A gene is composed of strings of bases (A,G, C, T) held together by a sugar phosphate backbone.

More about DNA

- DNA sequence can be denoted in a short handed way. P = phosphate A = adenine, G = guanine, T = thymine, C = cytosine
- Example:

```
5'… ATTGC … 3'
3'… TAACG … 5'
```

- The base order provides the first order instructions to the cells on how to create and maintain an organism.
- The human genome is packaged into 46 chromosomes, 23 homologous pairs. The DNA of one member of each pair is maternally derived and the DNA of the other is paternally derived. Chromosomes also contain proteins, such as histones, that provide structure and allow regulation of expression (RNA production, protein production).

Sequence Variation can lead to Phenotypic Variation

- The sequence of the two members of a chromosomal pair are not identical. Places where they differ are called variants.
- For 22 of the pairs, the sequence of the two members are nearly identical everywhere for all humans, but the members of the 23rd pair are quite different in males.
- The 23rd pair is the sex chromosomes. Females are XX and males are XY. Y is the smallest human chromosome.
- A phenotype is an observable trait or disease status.
- Variation in DNA leads to variation in phenotype.

Forms of DNA variation

Variation in the genome occurs because of mistakes in copying the DNA.

Examples:

A**A**GCT in mom A**G**GCT in child (substitution)

AAGAAGAAG in mom AAG**AAGAAG**AAGAGG (insertion)

AAG**AA**GAAG in mom AAGGAAG in child (deletion)

AAG**AAGCCT**TTA in mom AAG**TCCGAA**TTA in child (inversion)

These different forms are called variants

The position of a gene on a chromosome is called a locus

Variants at a locus are called alleles

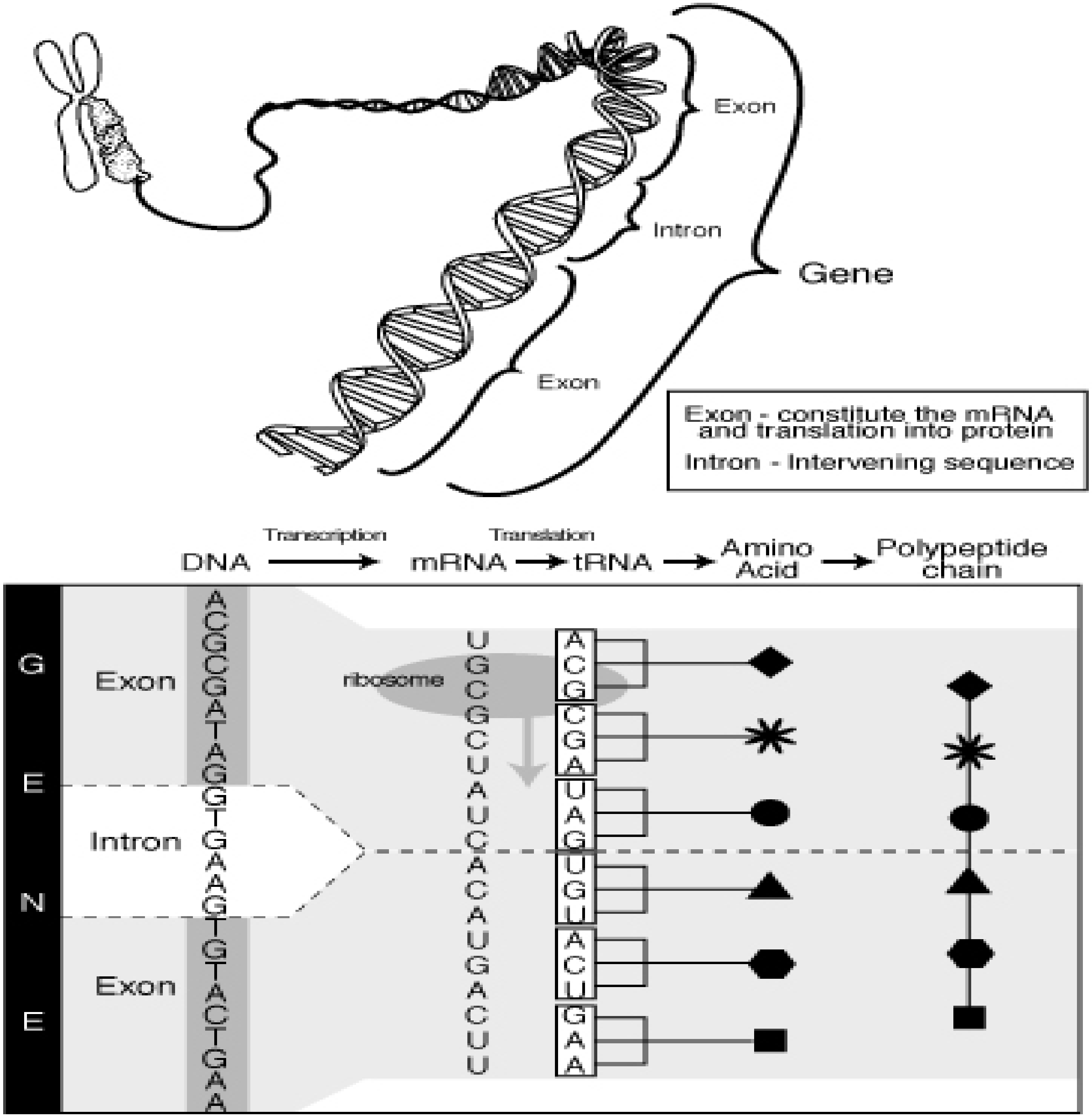
Summary of Terminology

Position	…12345678901234567890…
Maternally inherited chromosome	…GCAGGCCAGTTCATCCTCGA…
Paternally inherited chromosome	…GCCGGCCAGCTCATGCTCGA…

Locus: Any defined location or region of the genome.
Allele: The value (DNA seq.) inherited from a parent at a locus.
Genotype: The two alleles of an individual at a locus;
Position 3: A/C Heterozygote (A|C = ordered genotype)
Position 4: G/G Homozygote
Haplotype: set of alleles from one parent; for positions 3, 4, 10
AGT (maternally inherited) and CGC (paternally inherited)
Multilocus Genotype for positions 3, 4, 10:
AC.GG.TC (AGT/CGC)

The Blueprint and its Products: DNA –> RNA –> Proteins

- DNA is duplicated as cells divide
- DNA is transcribed into RNA.
- Expression: RNA production is regulated Once made, RNA is processed (portions spliced out etc)
- Expression: RNA is translated into protein (also a regulated process)



Exon - constitute the mRNA and translation into protein

Intron - Intervening sequence

How are Genes Passed on? Mendel's First Law

- An individual carries two copies of each gene (one on each copy of the chromosome) , one inherited from each parent. The gene at any given location is transmitted randomly and with probability ½.

  P(A/G -> G) = ½

  P(G/G -> G) = 1 and

  P(A/A ->G) = 0

Mode of Inheritance

- Mode of Inheritance = how the genotype effects the phenotype.
- Codominant Genotypes – when the phenotypic results of both alleles can be observed.
- Examples:
  - MN blood group (phenotypes M, MN, N)
  - hair texture (curly, wavy, straight hair)
  - Rose flower color (red, pink, white).
- Dominant and Recessive alleles – One allele masks the presence of the other when looking at the phenotype.
  - Example: ABO antigens of the red blood cell. Allele A is dominant to O if the genotype A/O results in the same phenotype as A/A. O is recessive to A.
- In genome wide association studies we usually assume that the markers act additively so we can count the number of minor alleles (0,1,or 2) and use as a continuous covariate.

Recoding of Alleles

- The actual nucleotides or repeated sequences of a polymorphism are often recoded to a set of numbers or letters.
- Letters or numbers are used to denote alleles at a theoretical locus.
- The alleles at a single nucleotide polymorphism (SNP) might be recoded to “A” & “a” or to “1” & “2”.
- A missing allele is often coded as 0.

Recoding of Genotypes

- The genotype is defined by the alleles at the locus.
- For SNPs, genotypes can be identified by a single integer that represents the number of minor alleles in the genotype, taking values of 0, 1 or 2.

Mendel’s Second Law of Independent Assortment of Loci

Suppose there are 2 loci with two alleles each. If a parent has genotypes A|a and B|b, then she will transmit A and B to her child with probability ¼.

Often violated! Holds when the loci are on different chromosomes or very far away on the same chromosome (the two loci are unlinked), but violated when the two loci are close together (the two loci are linked). The probability of co transmission of A and B will be much greater than ¼.

Violation of Mendel’s second law allows us to map the location of genes.

Violation of Mendel’s second Law: Gene Mapping and Chromosomal Recombination

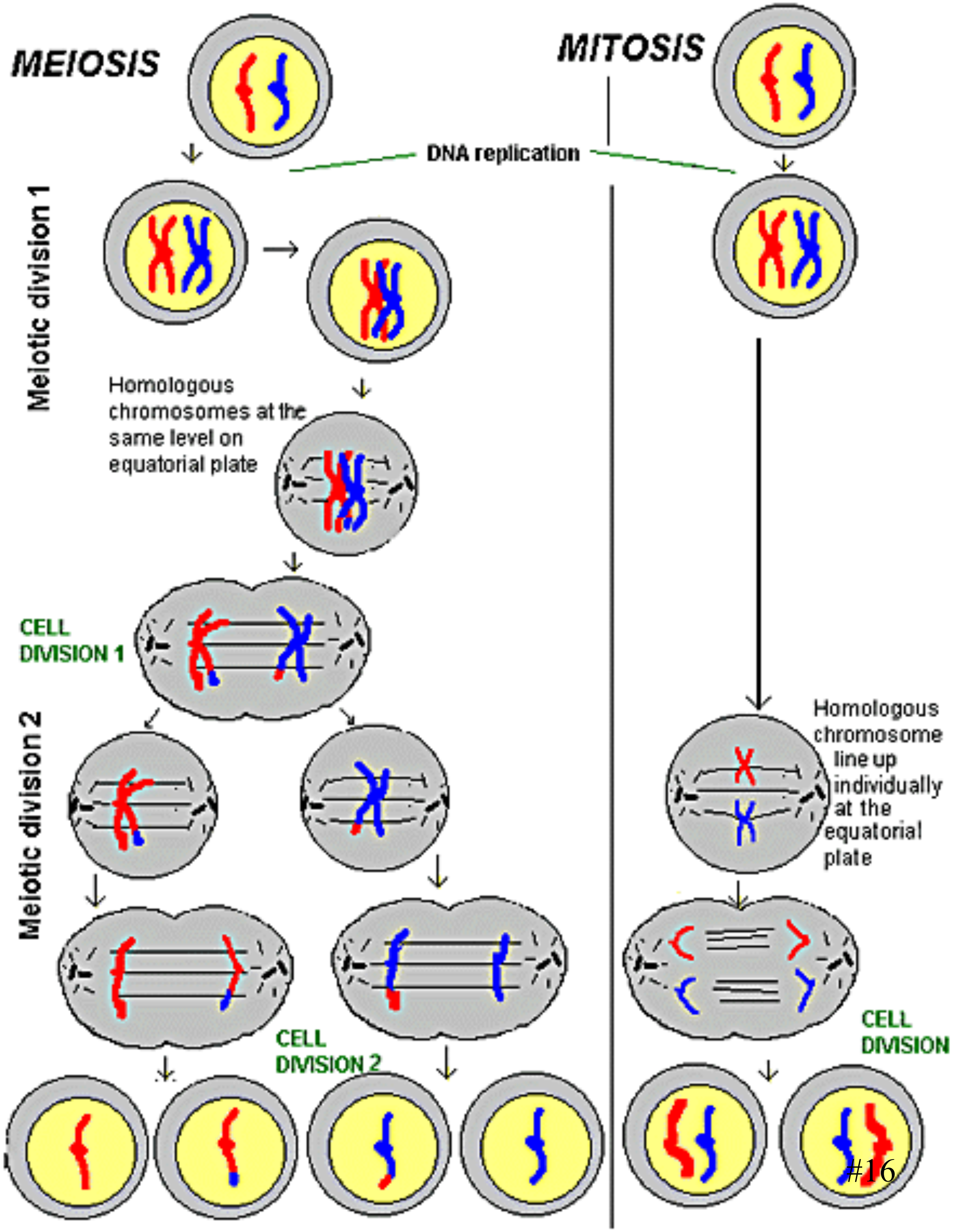
Gene mapping relies on the recombination between markers and trait genes.

- In the formation of the egg and sperm (the gametes), the chromosomes duplicate and pair up with their homologous chromosome.
- During this process, they form physical connections called chiasmata (singly chiasma). Chiasmata are randomly found on the chromosomes.
- There can be a break and reformation at a chiasma so that genetic material among the two chromosomes in a pair are exchanged.
- If the alleles are different at two loci (A/a and B/b), then an individual’s genotypes are called doubly heterozygous. To observe recombination, we need some individuals in the family to be doubly heterozygous.
- The probability of observing recombination ranges from 0 to ½ and is called the recombination fraction.

Meiosis and Mitosis

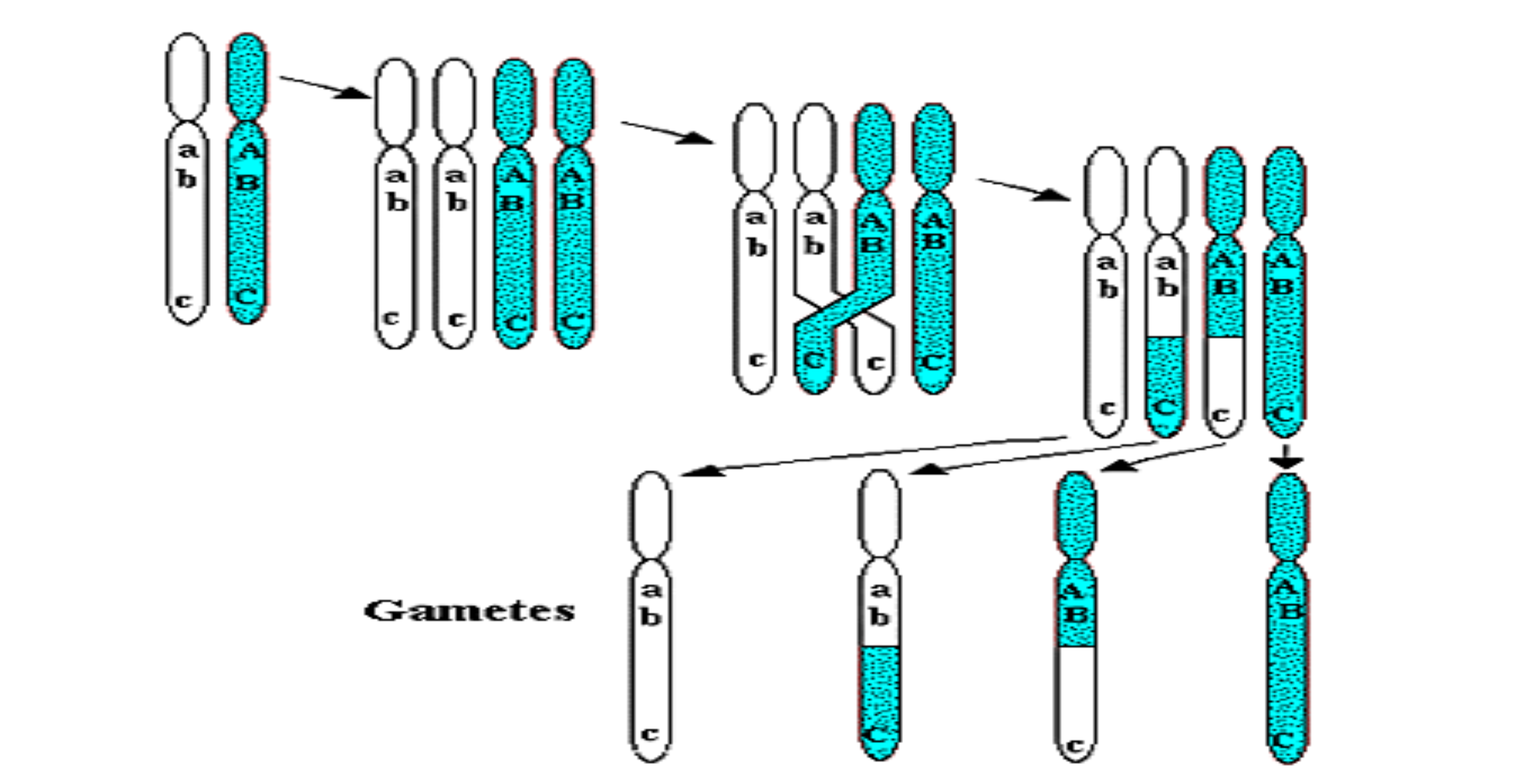
Errors occurring during DNA copying, called mutations, give rise to variation in DNA.
(Other sources?)

Somatic mutation versus germ line mutation.


MEIOSIS
MITOSIS
DNA replication
Meiotic division 1
Homologous chromosomes at the same level on equatorial plate
CELL DIVISION 1
Meiotic division 2
Homologous chromosome line up individually at the equatorial plate
CELL DIVISION 2
CELL DIVISION

gametes

Pairs of Homologous Chromosomes



Crossing-over and recombination during meiosis

Linkage Disequilibrium

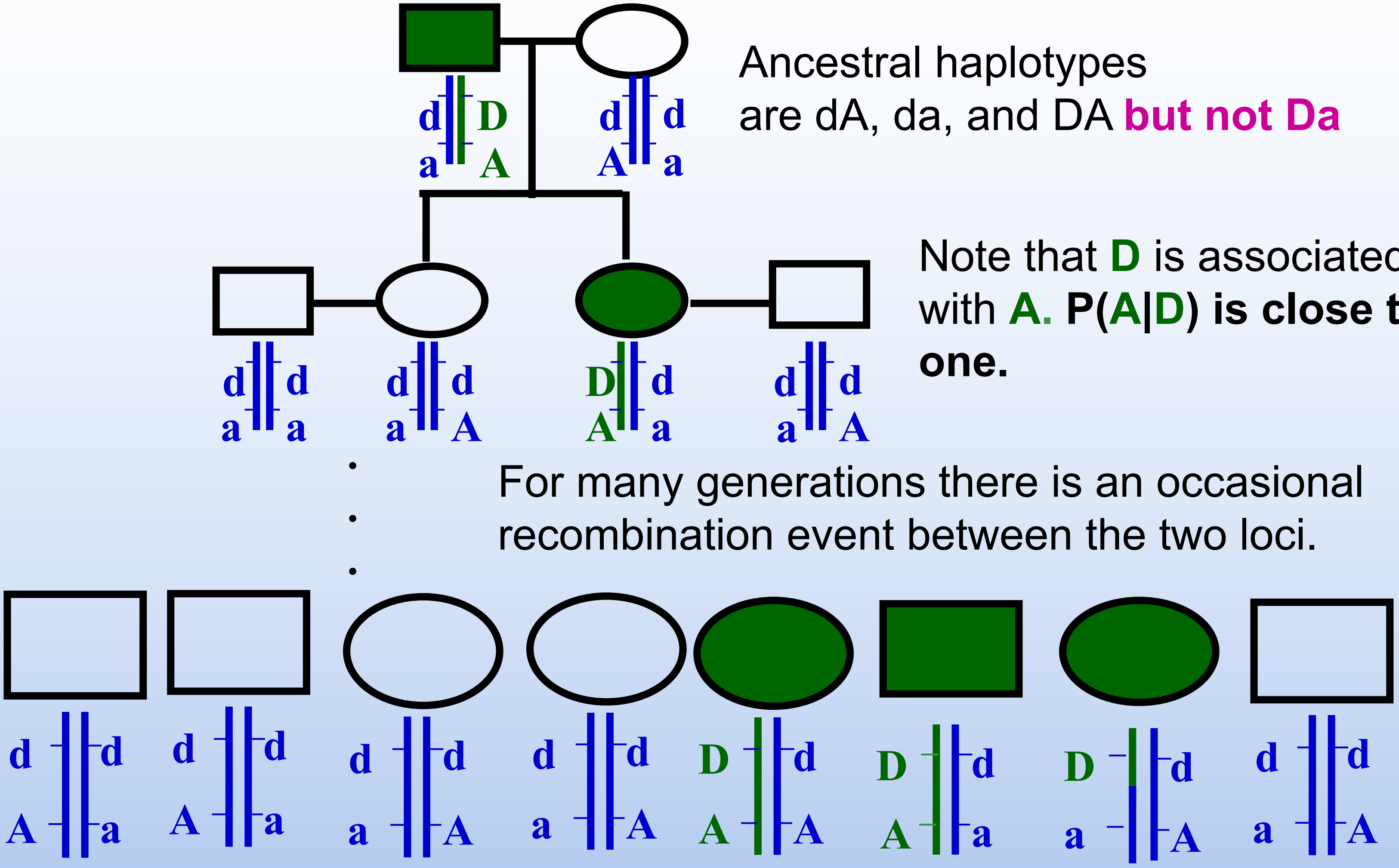
- Linkage disequilibrium is a measure that expresses the extent to which alleles at two loci are non-randomly associated within a population.
- Linkage disequilibrium implies a shared ancestry between alleles at two loci that has not been eroded by recombination.

Factors Influencing LD:

- LD throughout the genome reflects the population history, breeding patterns and geographic structure.
- LD in specific genomic regions reflects the evolutionary forces such as selection, gene conversion and mutation.
- How these local and global factors affect LD between a specific pair of loci will depend on the recombination rate.

Linkage Disequilibrium (LD)

One of the population founders carries an allelic variant that increases risk of a disease. The disease gene is very close to a marker.



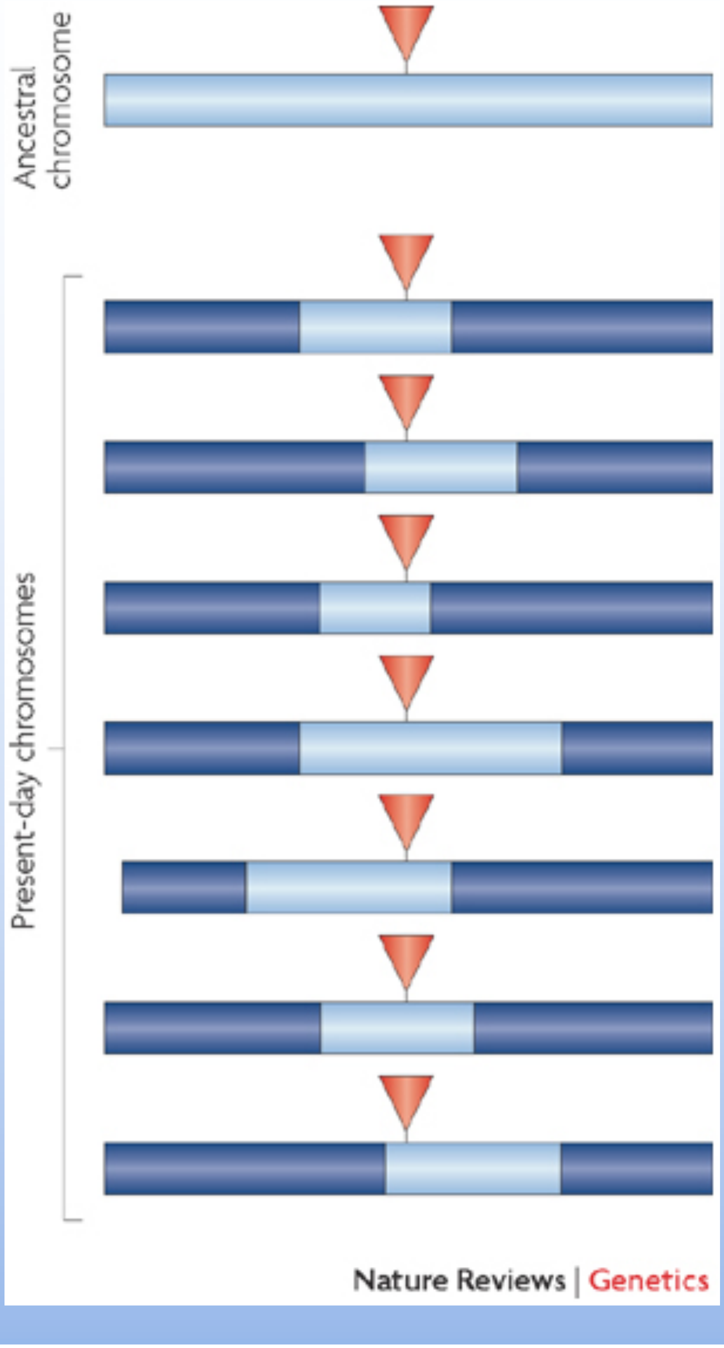
Ancestral haplotypes are dA, da, and DA **but not Da**

Note that **D** is associated with **A. P(A|D) is close to one.**

For many generations there is an occasional recombination event between the two loci.

The P(A|D) has decreased, as has the degree of association between **D** and **A**, but still P(A|D) > P(A) and P(haplotype: A D) > P(A) x P(D)

Another view of LD



Nature Reviews | Genetics

A new mutation is in complete LD with its background chromosome

Many generations later the region of complete LD is much smaller, but partial LD can still be observed around the mutation site.

Linkage Disequilibrium (LD)

Two loci, A and B, have alleles A, a and B, b all with non-zero frequency.

The two alleles at the two loci, can result in as few as two, and as many as four haplotypes. {AB, Ab, aB, ab}

	A	a	
B	p_{11}	p_{21}	p_{+1}
b	p_{12}	p_{22}	p_{+2}
	p_{1+}	p_{2+}	

The frequencies of the four possible haplotypes can be denoted as p_{ij}, where *i* refers to the allele at the A locus and *j* for the B locus.

Allele frequencies are given by the marginal totals.

The linkage disequilibrium coefficient expresses the non-independence of the alleles at the two loci.

Linkage Disequilibrium coefficient: $\mathrm{D} = p_{11} - p_{+1} \times p_{1+}$

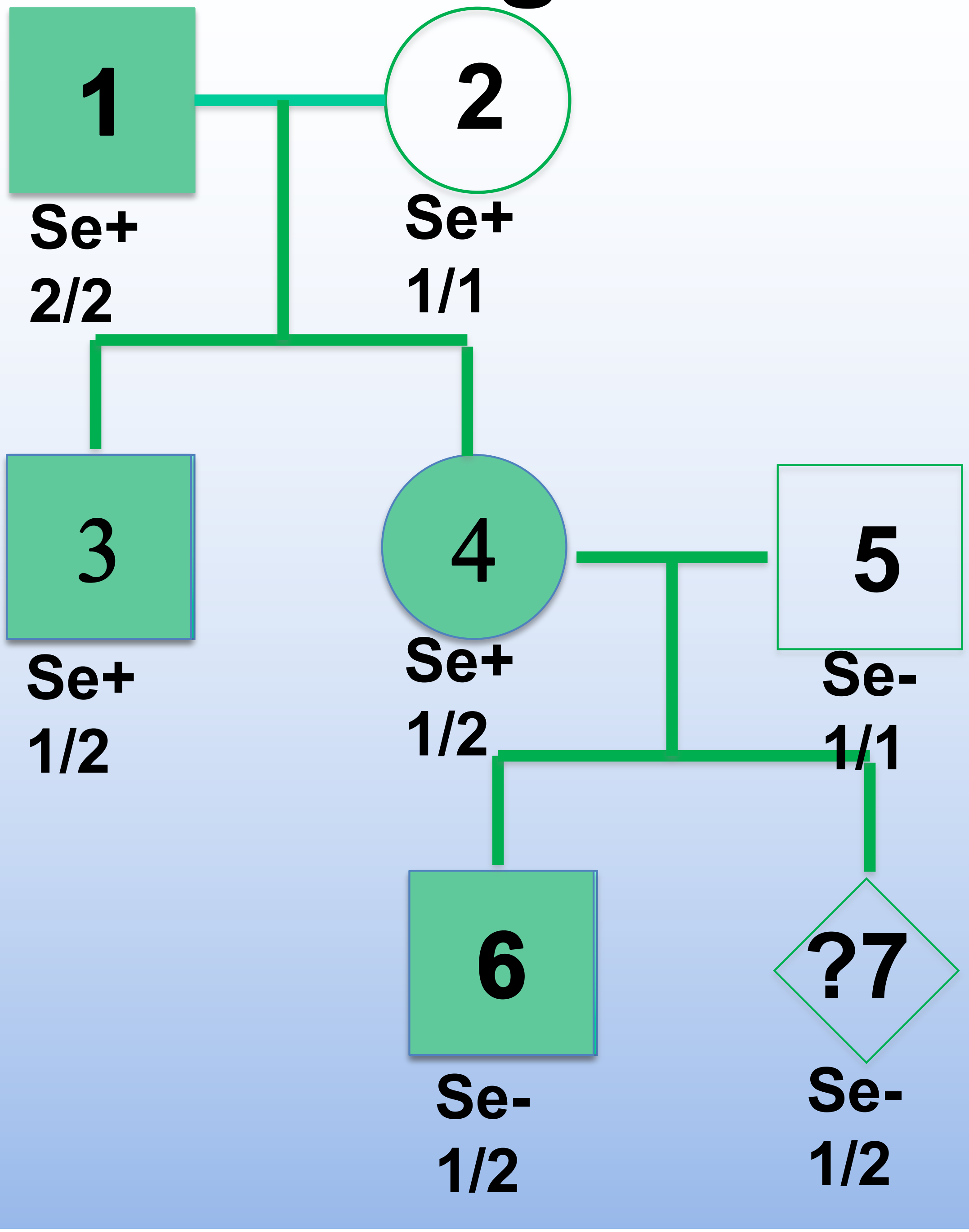
Linkage Equilibrium: $p_{11} = p_{+1} \times p_{1+}$

Linkage Disequilibrium: $p_{11} \neq p_{+1} \times p_{1+}$

How are the Family Data Displayed?

- Pedigrees: Genetics implicitly relies on families.
- We depict with stylized drawings and set conventions.
  - Males are displayed as squares, females as circles. Deceased individuals are denoted with a line through their symbol.
  - Mating pairs are denoted by a horizontal line that directly connects them
  - Offspring are denoted by a vertical line from a mated pair. Either both parents are included (even if no information is available about one or both of them) or neither of them.
  - Those individuals without parents in the pedigree are called founders.
  - Siblings are denoted by their common relationship to their parents. (forms a fork like structure).

Pedigree



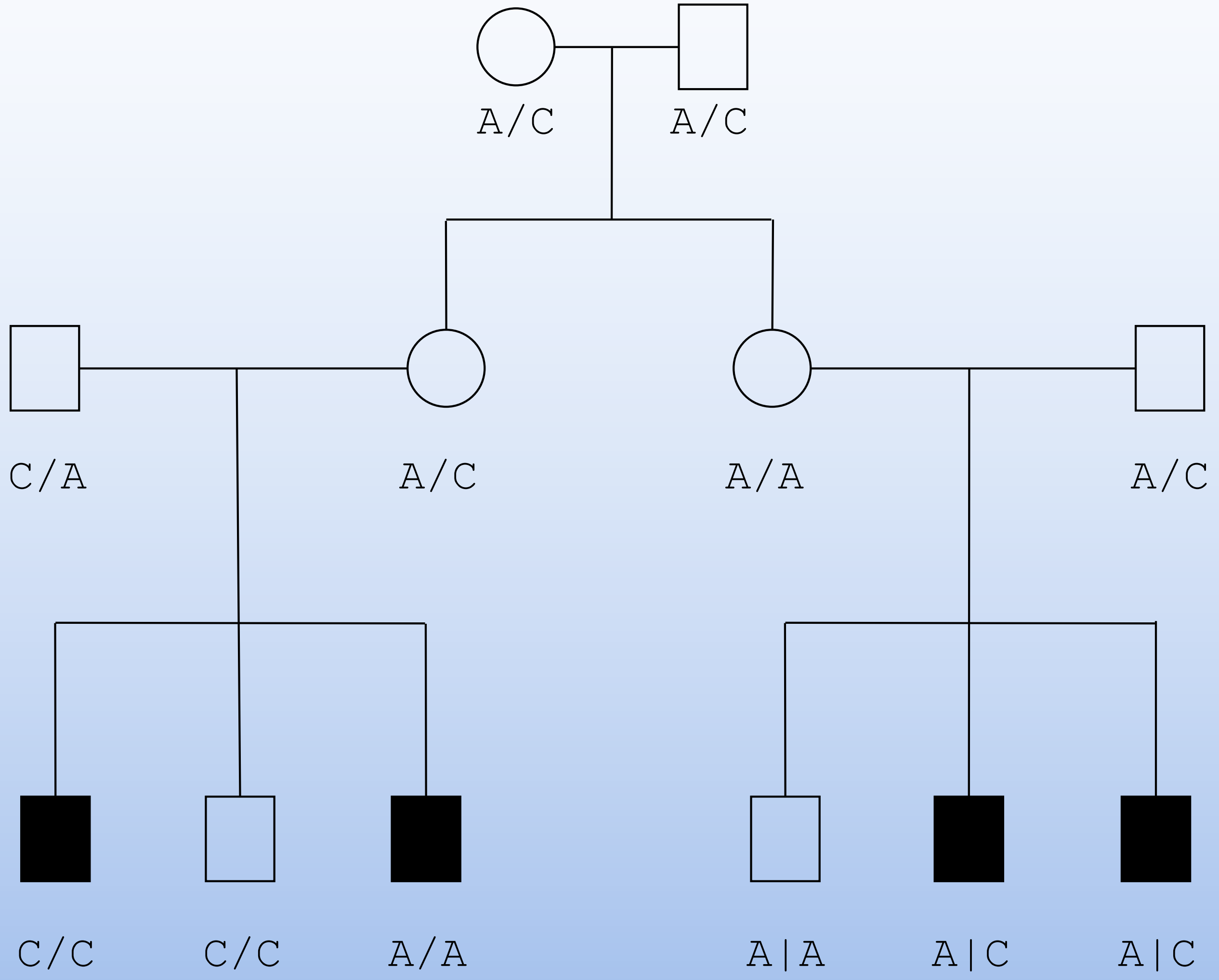
Pedigree as a Data Frame

```
Pedigree,Person,Mother,Father,Sex,MD,Se,M2
Bottom,1,,,male,Affected,+,2
Bottom,2,,,female,Normal,+,0
Bottom,5,,,male,Normal,-,1
Bottom,3,2,1,male,Affected,+,0
Bottom,4,2,1,female,Affected,+,1
Bottom,6,4,5,male,Affected,-,1
Bottom,7,4,5,male,,-,1
```

IBD vs IBS

- Two *alleles* are called **identical-by-descent** (IBD) if they are copies of the same ancestral allele.
- Two *alleles* are called **identical-by-state** (IBS) if they have the same name/value, i.e., are in the same state.
- If two alleles are IBD they will be IBS.

IBD vs IBS: Examples



Gene Mapping Overview

- Within candidate region or genome-wide, genotype at loci of known position.
- Search for correlation between locus genotype and trait phenotype, that is, for regions of DNA that co-segregate with phenotype.
- Positive result provides localization of susceptibility gene to chromosomal region.
- As with all statistical tests, the crucial task is to determine how significantly different is the result from what might happen by chance.

Association Overview

- Association is simply a statistical statement about the co-occurrence of alleles and phenotypes.

- Allele ***A*** at a locus is associated with disease allele ***D*** at a trait locus if people who have ***D*** (cases) also have ***A*** more often than would be predicted from the population frequencies of ***D*** and ***A***:
  p(**A** & **D**) > p(**A**)p(**D**).

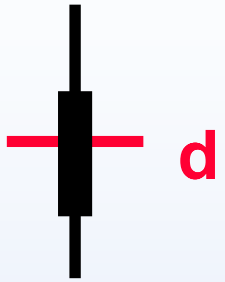
- If affected individuals are significantly less likely to have allele ***A***, then ***A*** is still associated but protective.

What are the Consequences of LD for Association Analysis?

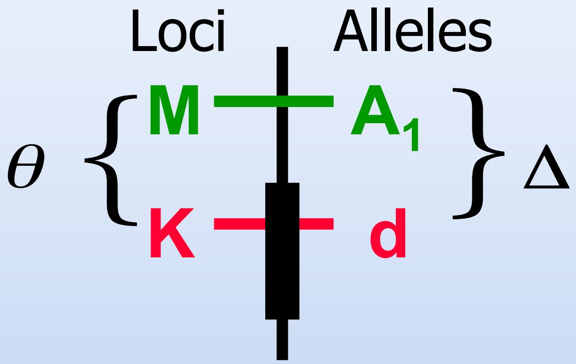
- Violation of Mendel Law #2 means that we can expect SNPs that are close in distance will be highly correlated.
- As the distance increases we expect the correlation to decrease.
- This property is both good and bad.
  - Good – we don't need to examine association at every single location along the chromosome.
  - Bad – there is a resolution limit. That is we can't pinpoint the most likely locus to be the causal one simply by using an association test.

Three Causes of Allelic Association

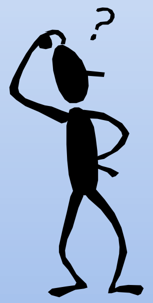
- best: allele increases disease susceptibility
  - **candidate gene studies**



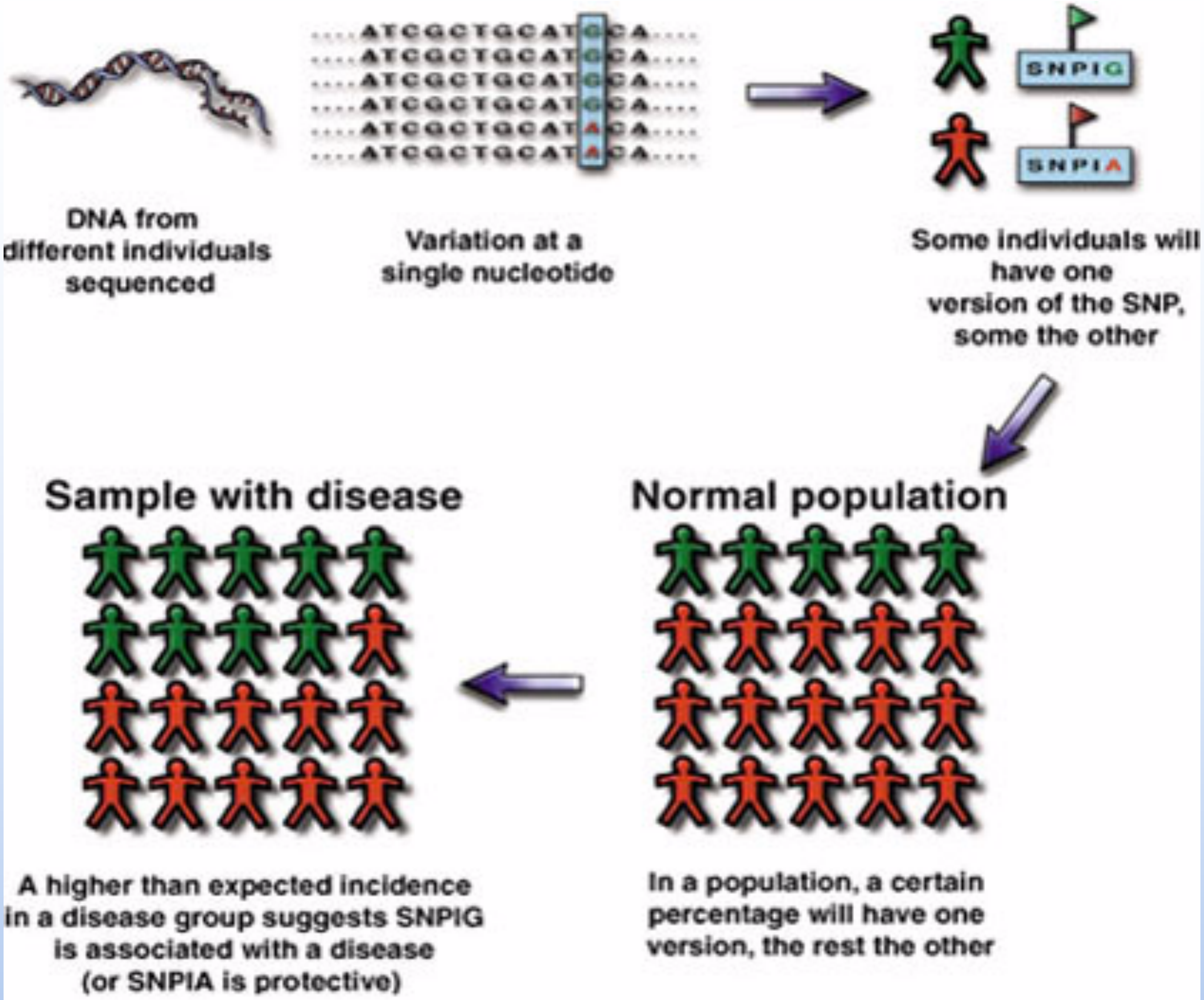
- good: some subjects share common ancestor
  - **linkage disequilibrium studies**



- bad: association due to population stratification
  - **family-based studies offer protection**



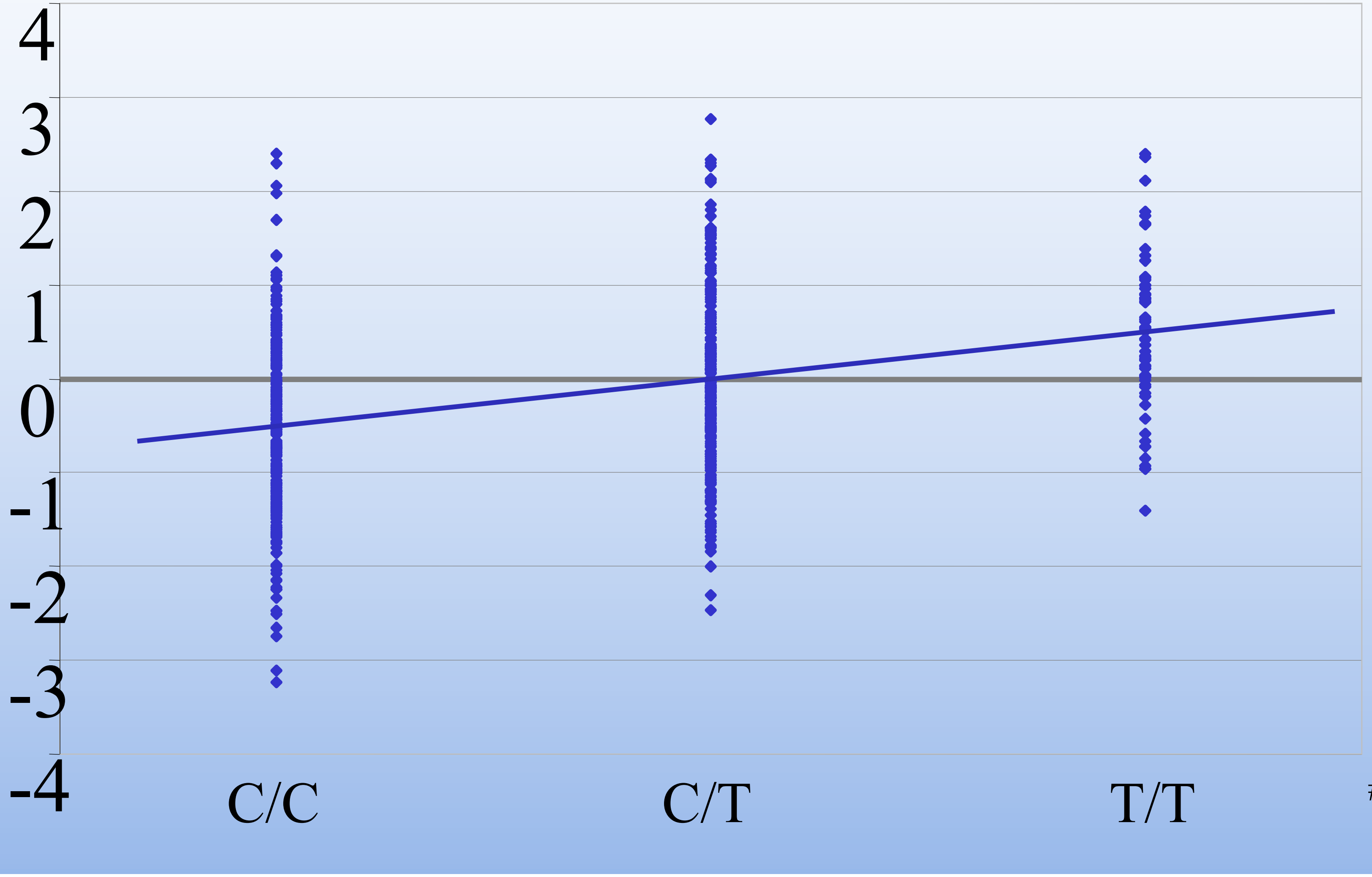
Case-Control association using a SNP


....ATCGCTGCATGCA....
....ATCGCTGCATGCA....
....ATCGCTGCATGCA....
....ATCGCTGCATGCA....
....ATCGCTGCATACA....
....ATCGCTGCATACA....
SNPIG
SNPIA
DNA from different individuals sequenced
Variation at a single nucleotide
Some individuals will have one version of the SNP, some the other
Sample with disease
Normal population
A higher than expected incidence in a disease group suggests SNPIG is associated with a disease (or SNPIA is protective)
In a population, a certain percentage will have one version, the rest the other

Because marker allele causes the disease or because of linkage disequilibrium, disease status and a marker allele are associated.

Wellcome News, Issue 20 Q3 1999

Association between a Locus and a Quantitative Trait in a Population-based Sample



Research Opportunity: Going beyond Genotype-Phenotype Association

- Genetic association studies are fraught with false positives. Replication using another population is necessary but it's not sufficient to be confident in the results.
- Need evidence of the biological pathway leading from the DNA to the phenotype.
- Genomics – genome-wide expression, epigenomic, microbiome and proteomic data, etc,
  - Massive data sets that reflect whole genome pathways. Creates computational and inferential problems to be addressed.
  - Integrating these data to make inferences regarding causal pathways is a very active area of research.

Genetics and Genomics

- A huge field.
- We just shratched the surface.
- More statisticians needed in the field.

Medical/Molecular Genetics References

- Human Molecular Genetics, by Strachan and Read.
- Medical Genetics, by Jorde, Carey, Bamshad
- Thompson & Thompson Genetics in Medicine by Nussbaum, McInnes, and Willard.

Statistical Genetics References

- Mathematical and Statistical Methods for Genetic Analysis, 2nd edition, by Ken Lange.
- The Fundamentals of Modern Statistical Genetics by Nan Laird and Christoph Lange
- Statistical Human Genetics Methods and Protocols 2nd Editor: Robert Elston.