INTRODUCTION TO GENETIC EPIDEMIOLOGY

(EPID0754)

Prof. Dr. Dr. K. Van Steen

CHAPTER 2: INTRODUCTION TO GENETICS

1 Basics of molecular genetics

1.a Where is the genetic information located?

The structure of cells, chromosomes, DNA and RNA

1.b What does the genetic information mean?

Reading the information, reading frames

1.c How is the genetic information translated?

The central dogma of molecular biology

2 Overview of human genetics

2.a How is the genetic information transmitted from generation to generation?

Review of mitosis and meiosis, recombination and cross-over

2.b Variability is the key to "information"

Polymorphisms and mutations, trait variation

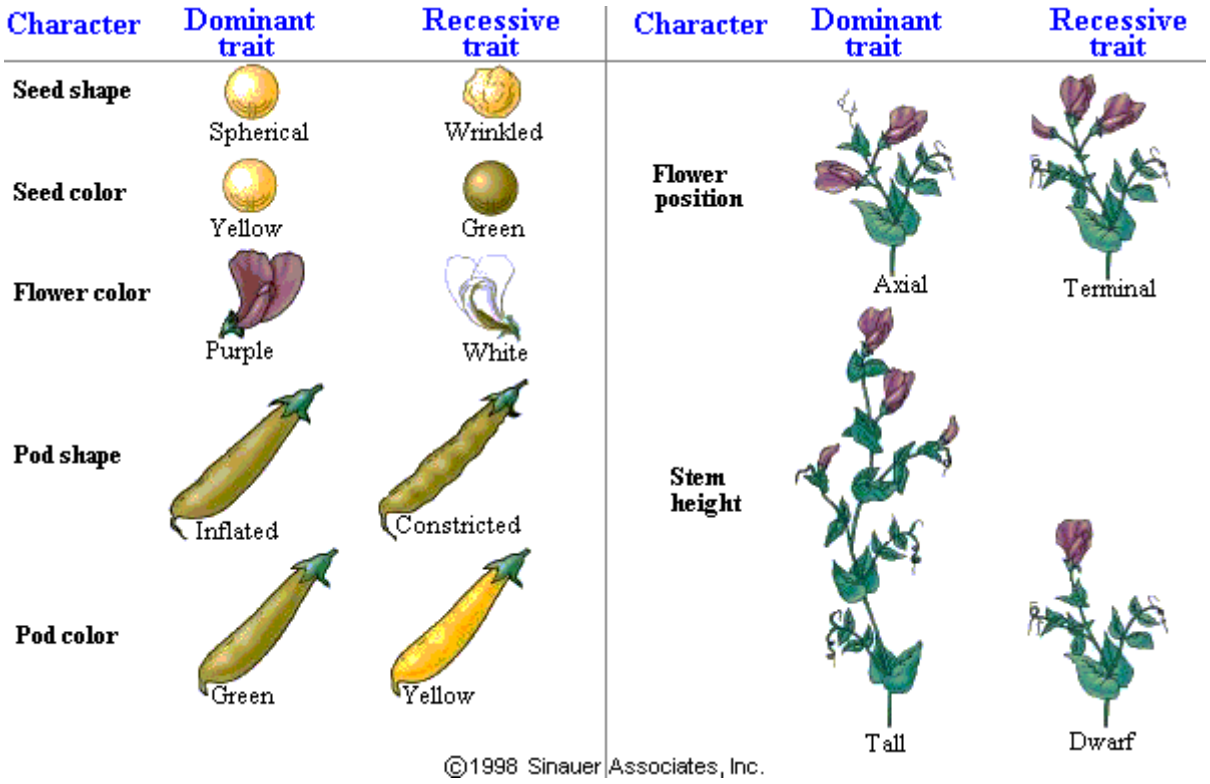
1 Basics of molecular genetics

1.a Where is the genetic information located?

Mendel

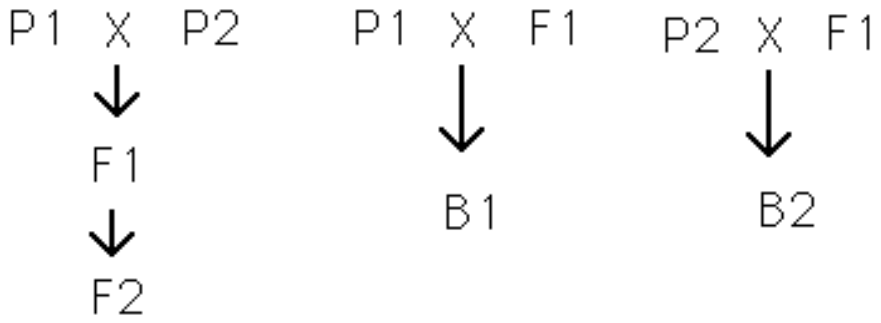
- Many traits in plants and animals are heritable; genetics is the study of these heritable factors
- Initially it was believed that the mechanism of inheritance was a masking of parental characteristics
- Mendel developed the theory that the mechanism involves random transmission of discrete “units” of information, called genes. He asserted that,
    - when a parent passes one of two copies of a gene to offspring, these are transmitted with probability 1/2, and different genes are inherited independently of one another (is this true?)

Mendel's pea traits

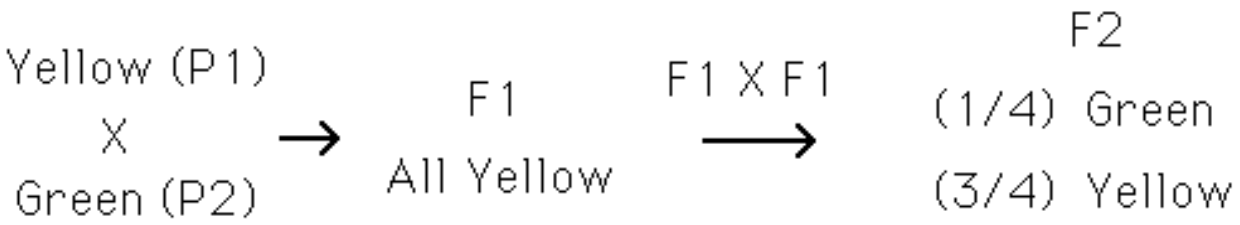


Some notations for line crosses

- Parental Generations (P_1 and P_2)
- First Filial Generation $F_1 = P_1 \times P_2$
- Second Filial Generation $F_2 = F_1 \times F_1$
- Backcross one, $B_1 = F_1 \times P_1$
- Backcross two, $B_2 = F_1 \times P_2$



What Mendel observed



- The F_1 were all Yellow
- Strong evidence for discrete units of heredity , as "green" unit obviously present in F_1, appears in F_2
- There is a 3:1 ratio of Yellow : Green in F2

What Mendel observed (continued)

- Parental, F_1 and F_2 yellow peas behave quite differently

P1 Yellow
X
Green (P2)
→ All Yellow

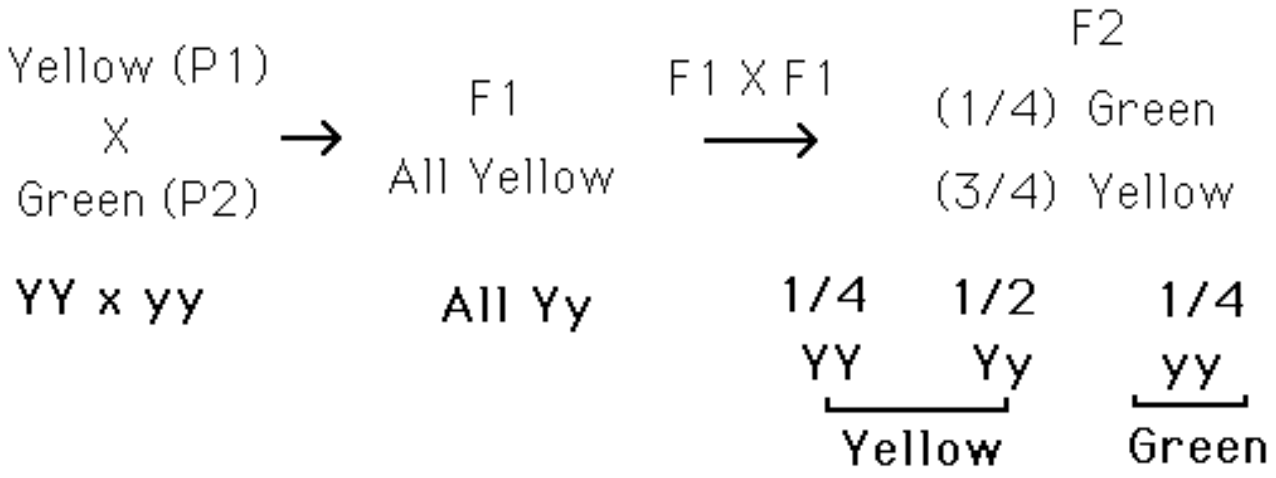
F1 Yellow
X
Green (P2)
→ 1/2 Yellow
1/2 Green

F2 Yellow
X
Green (P2)
→ 2/3 Yellow
1/3 Green

Mendel's conclusions

- **Mendel's first law** (law of segregation of characteristics)

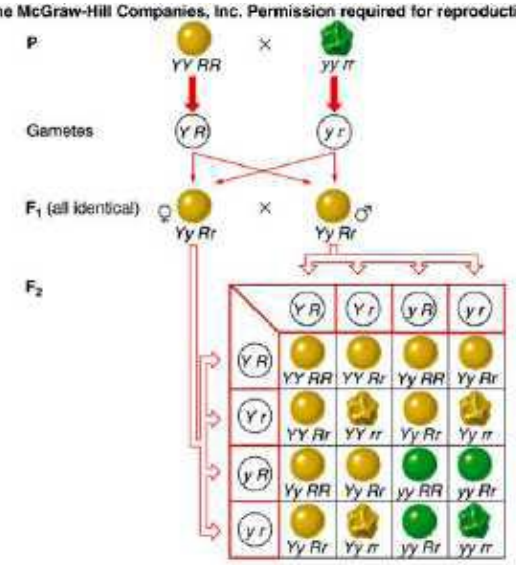
  This says that of a pair of characteristics (e.g. blue and brown eye colour) only one can be represented in a gamete. What he meant was that for any pair of characteristics there is only one gene in a gamete even though there are two genes in ordinary cells.



Mendel's conclusions (continued)

- **Mendel's second law** (law of independent assortment)

  This says that for two characteristics the genes are inherited independently.



Type	Genotype	Phenotype	Number	Phenotypic ratio
Parental	Y– R–	yellow round	315	9/16
Recombinant	yy R–	green round	108	3/16
Recombinant	Y– rr	yellow wrinkled	101	3/16
Parental	yy rr	green wrinkled	32	1/16

Ratio of yellow (dominant) to green (recessive) = 12:4 or 3:1
Ratio of round (dominant) to wrinkled (recessive) = 12:4 or 3:1

Mendelian transmission in simple words

- One copy of each gene is inherited from the mother and one from the father. These copies are not necessarily identical
- Mendel postulated that mother and father each pass one of their two copies of each gene independently and at random
- At a given locus, the father carries “alleles” a and b and the mother carries “alleles” c and d, the offspring may be a/c, a/d, b/c or b/d, each with probability 1/4 → note the notation with the “/” … (notational variations exist)
- Transmission of genes at two different positions, or loci, on the same *chromosome* (see later) may actually NOT be independent. If dependent, they are said to be *linked.* → related to physical proximity.

The cell as the basic unit of biological functioning

- Let us take it a few levels up …
- Although the tiniest bacterial cells are incredibly small, weighing less than 10-12 grams, each is in effect a veritable micro-miniaturized factory containing thousands of exquisitely designed pieces of intricate molecular machinery, made up altogether of one hundred thousand million atoms, far more complicated than any machinery built by man and absolutely without parallel in the non-living world.
- Each microscopic cell is as functionally complex as a small city. When magnified 50,000 times through electron micrographs, we see that a cell is made up of multiple complex structures, each with a different role in the cell's operation.

(http://www.allaboutthejourney.org/cell-structure.htm)

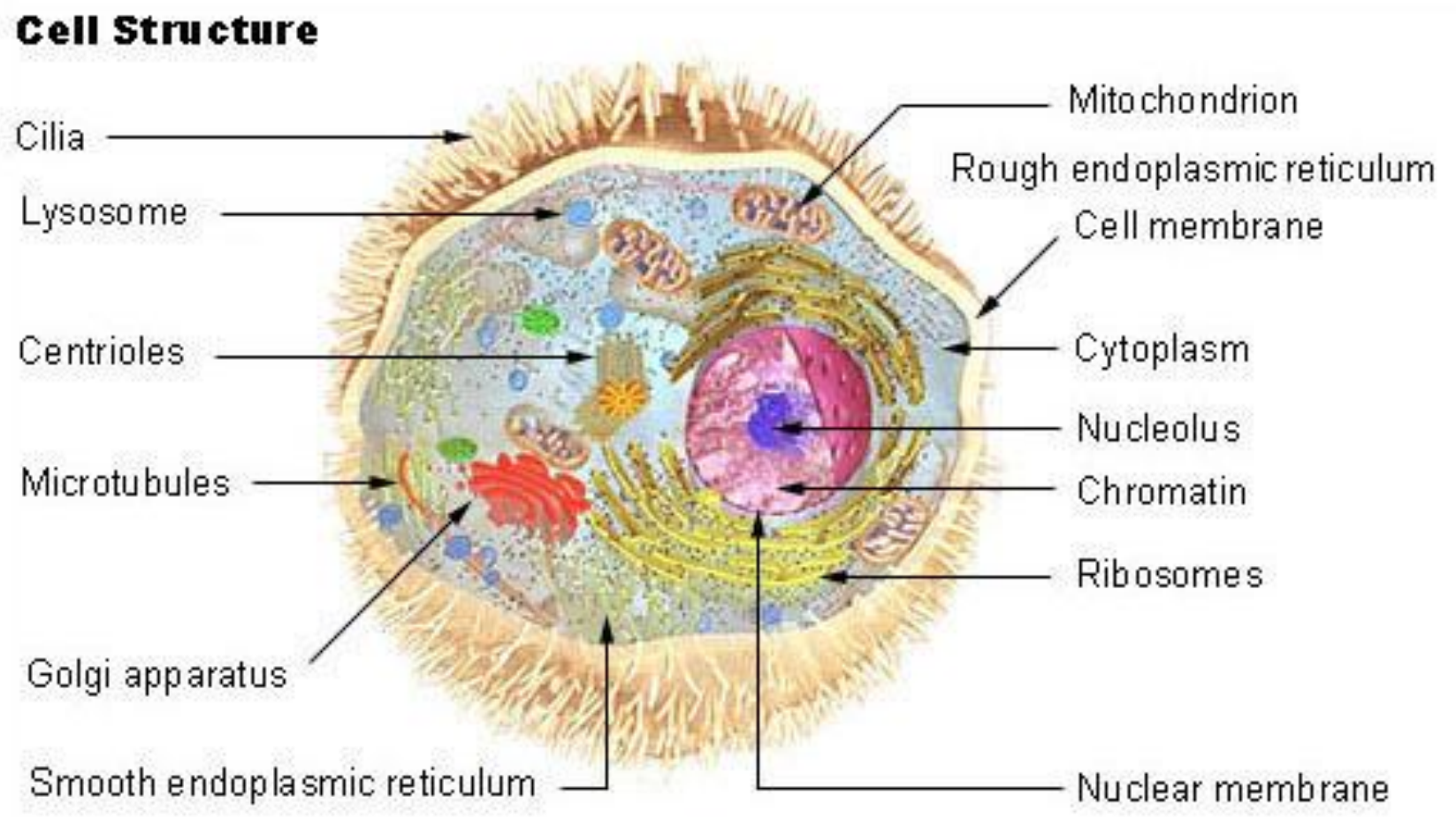
The cell as the basic unit of biological functioning

- Using the city comparison, here's a simple chart that reveals the design of a typical human cell:

City	Cell
Workers	Proteins
Power plant	Mitochondria
Roads	Actin fibers, Microtubules
Trucks	Kinesin, Dinein
Factories	Ribosomes
Library	Genome
Recycling center	Lysosomes
Police	Chaperones
Post office	Golgi Apparatus

(http://www.allaboutthejourney.org/cell-structure.htm)

The cell as the basic unit of biological functioning



(http://training.seer.cancer.gov/anatomy/cells_tissues_membranes/cells/structure.html)

- *Eukaryotes*: organisms with a rather complex cellular structure. In their cells we find organelles, clearly discernable compartments with a particular function and structure.
  - The organelles are surrounded by semi-permeable membranes that compartmentalize them further in the cytoplasm.
  - The Golgi apparatus is an example of an organelle that is involved in the transport and secretion of proteins in the cell.
  - Mitochondria are other examples of organelles, and are involved in respiration and energy production

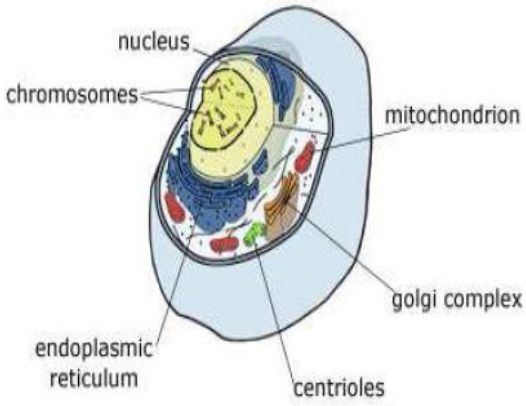
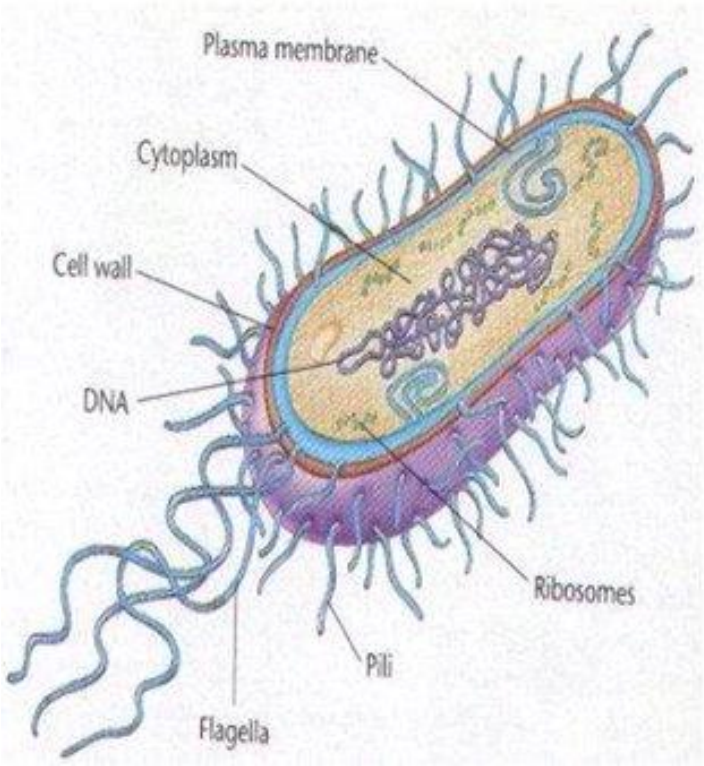


Image adapted from: National Human Genome Research Institute.

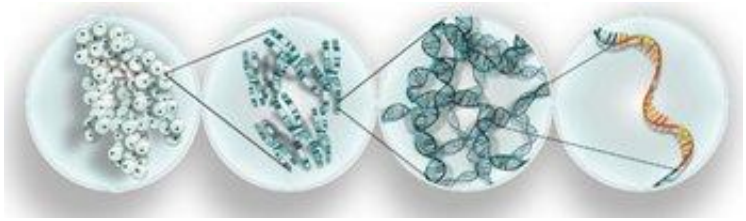
A typical eukaryotic cell.

- *Prokaryotes*: cells without organelles where the genetic information floats freely in the cytoplasm



The miracle of life

- There are three main platforms to explain the miracle of human life:



(VIB, Biotechnology)

  - The cells of the living organism. The cells are thus the basic unit of all biological functions
  - The genetic instructions that are responsible for the properties of the cell
  - The biological mechanisms that are used by the cells to carry out the instructions.
- The genetic instructions are stored in code in the DNA. The collection of all possible genetic instructions in a cell is called the *genome.*

History revealed that genes involved DNA

Geneticists already knew that DNA held the primary role in determining the structure and function of each cell in the body, but they did not understand the mechanism for this or that the structure of DNA was directly involved in the genetic process.

British biophysicist Francis Crick and American geneticist **James Watson** undertook a joint inquiry into the structure of DNA in 1951.

(http://www.pbs.org/wgbh/nova/genome)

Watson and Crick

"We wish to suggest a structure for the salt of deoxyribose nucleic acid (D.N.A). This structure has novel features which are of considerable biological interest."

(Watson JD and Crick FHC. A Structure for DNA, *Nature*, 1953)

MOLECULAR STRUCTURE OF NUCLEIC ACIDS

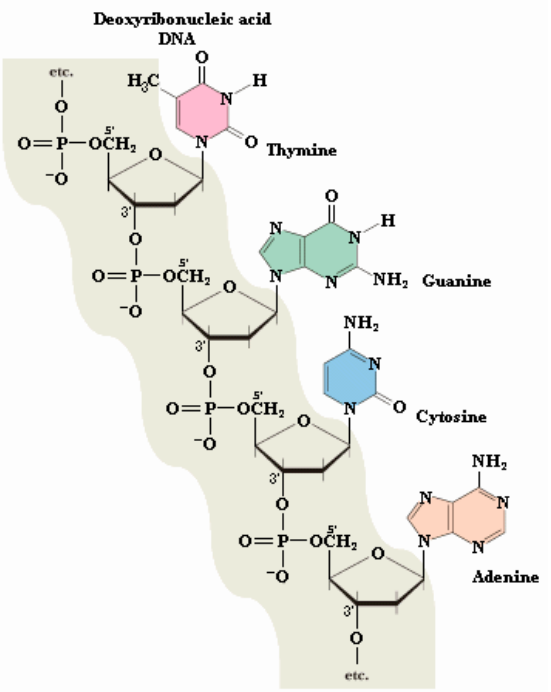
A Structure for Deoxyribose Nucleic Acid

What does “DNA” stand for?

- Deoxyribonucleic acid (DNA) IS the genetic information of most living organisms. In contrast, some viruses (called retroviruses) use ribonucleic acid as genetic information. “Genes” correspond to sequences of DNA
- DNA is a polymere (i.e., necklace of many alike units), made of units called nucleotides.
- Some interesting features of DNA include:
  - DNA can be copied over generations of cells*: DNA replication*
  - DNA can be translated into proteins: *DNA transcription* into RNA, further translated into proteins
  - DNA can be repaired when needed: *DNA repair*.

What does "DNA" stand for?

- There are 4 nucleotide *bases*, denoted A (adenine), T (thymine), G (guanine) and C (cytosine)
- A and G are called purines, T and C are called pyrimidines (smaller molecules than purines)
- The two strands of DNA in the double helix structure are complementary (sense and anti-sense strands); A binds with T and G binds with C



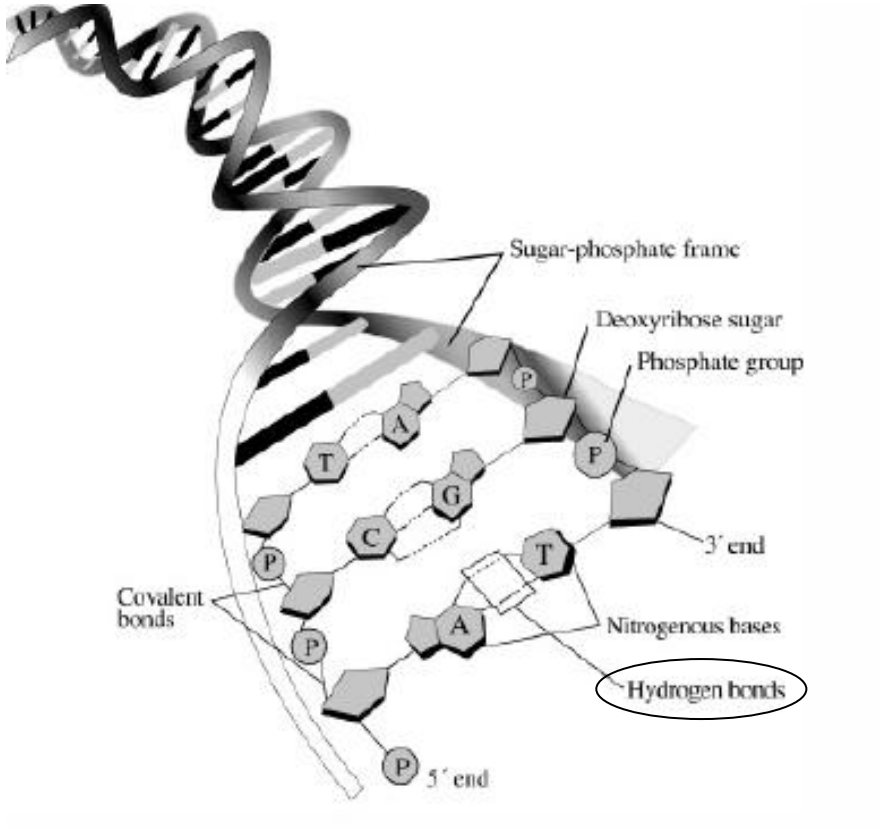
(Biochemistry 2nd Ed. by Garrett & Grisham)

Primary structure of DNA

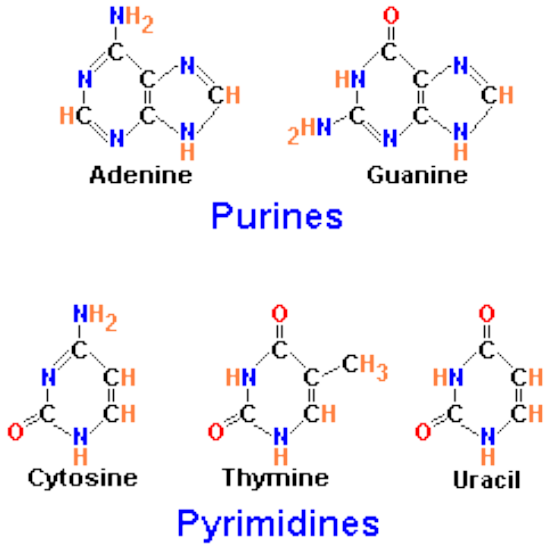
The 3 dimensional structure of DNA can be described in terms of primary, secondary, tertiary, and quaternary structure.

- The primary structure of DNA is the sequence itself - the order of nucleotides in the deoxyribonucleic acid polymer.
- A *nucleotide* consists of
  - a phosphate group,
  - a deoxyribose sugar and
  - a nitrogenous base.
- Nucleotides can also have other functions such as carrying energy: ATP
- Note: Nucleo **s** ides are made of a sugar and a nitrogenous base…

Nucleotides



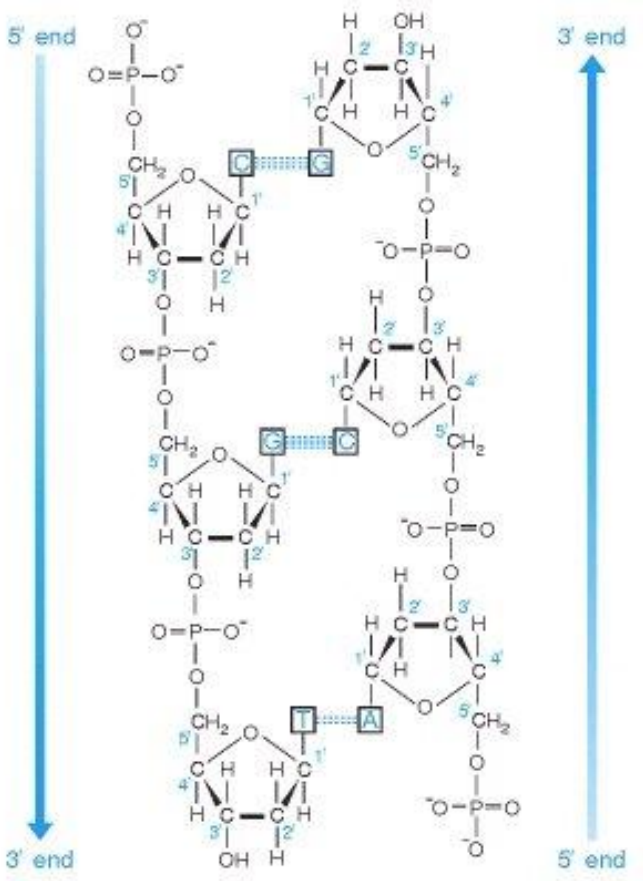
Nitrogenous bases



(http://www.sparknotes.com/101/index.php/biology)

Secondary structure of DNA

- The secondary structure of DNA is relatively straightforward - it is a double helix.
- It is related to the hydrogen bonding
- The two strands are anti-parallel.
  - *The 5' end* is composed of a phosphate group that has not bonded with a sugar unit.
  - *The 3' end* is composed of a sugar unit whose hydroxyl group has not bonded with a phosphate group.



Major groove and minor groove

- The double helix presents a major groove and a minor groove (Figure 1).
  - The major groove is deep and wide (backbones far apart)
  - The minor groove is narrow and shallow (backbones close to each other)

- The chemical groups on the edges of GC and AT base pairs that are available for interaction with proteins in the major and minor grooves are color-coded for different types of interactions (Figure 2)

Figure 1

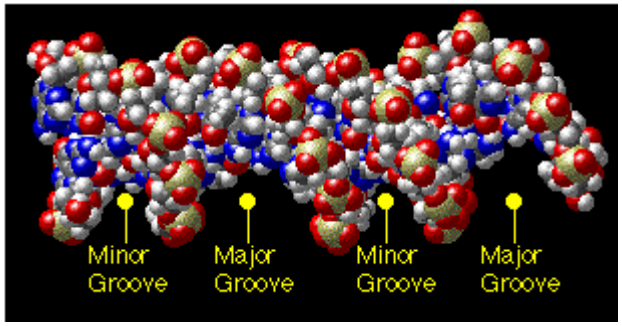
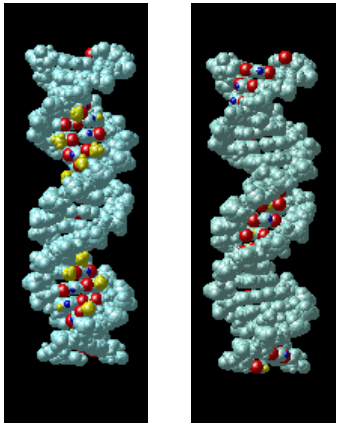


Figure 2

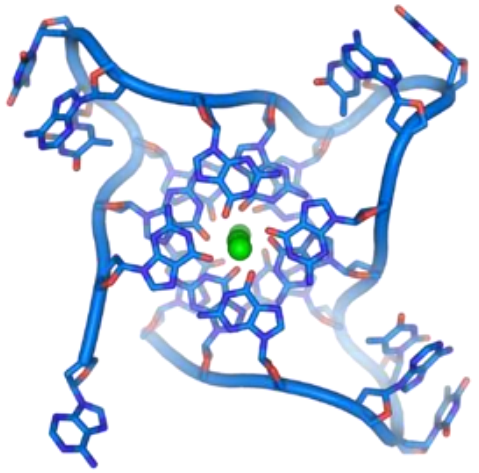


Tertiary structure of DNA

- This structure refers to how DNA is stored in a confined space to form the chromosomes.
- It varies depending on whether the organisms prokaryotes and eukaryotes:
  - In prokaryotes the DNA is folded like a super-helix, usually in circular shape and associated with a small amount of protein. The same happens in cellular organelles such as mitochondria .
  - In eukaryotes, since the amount of DNA from each chromosome is very large, the packing must be more complex and compact, this requires the presence of proteins such as histones and other proteins of non-histone nature
- Hence, in humans, the double helix is itself super-coiled and is wrapped around so-called histones (see later).

Quaternary structure of DNA

- At the ends of linear chromosomes are specialized regions of DNA called telomeres.
- The main function of these regions is to allow the cell to replicate chromosome ends using the enzyme telomerase, since other enzymes that replicate DNA cannot copy the 3 'ends of chromosomes.
- In human cells, telomeres are long areas of single-stranded DNA containing several thousand repetitions of a single sequence TTAGGG.

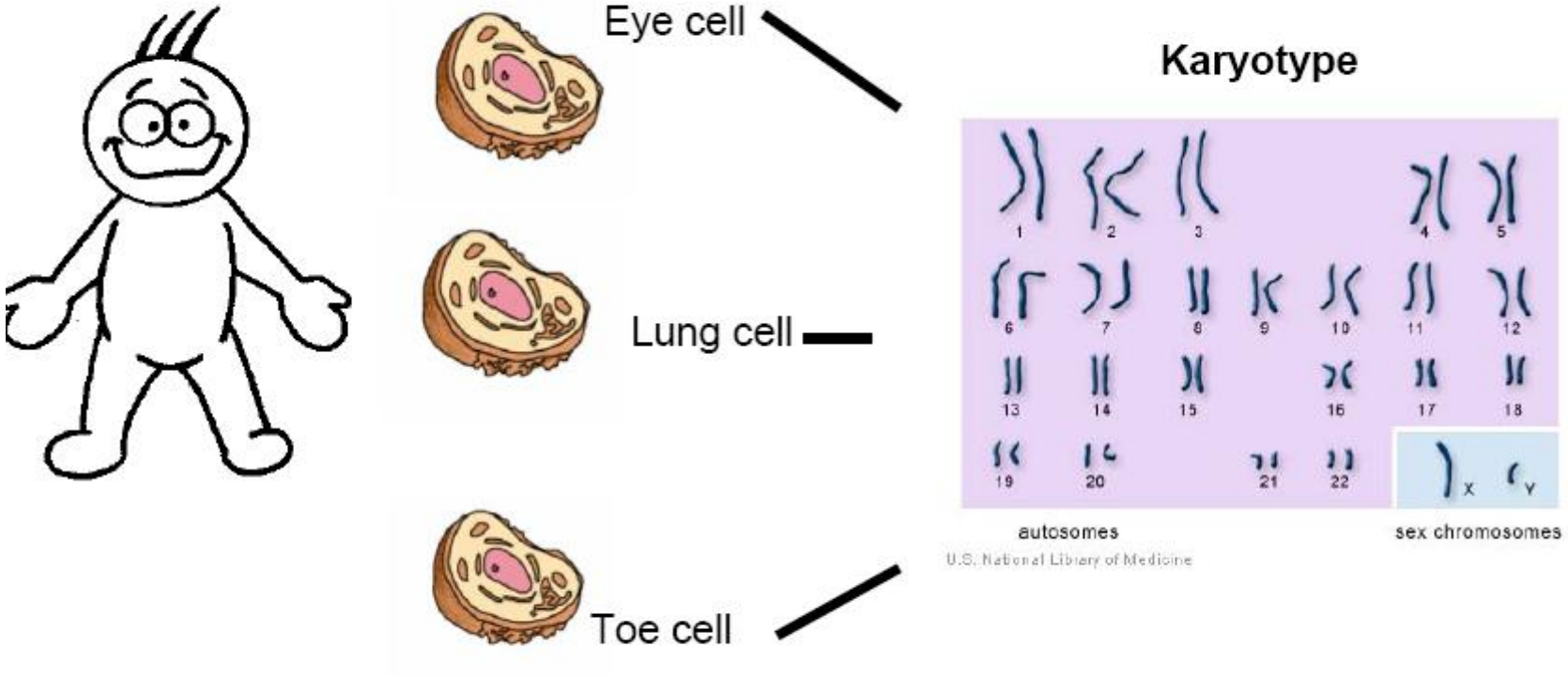


(http://www.boddunan.com/miscellaneous)

The structure of DNA

- A wide variety of proteins form complexes with DNA in order to replicate it, transcribe it into RNA, and regulate the transcriptional process (central dogma of molecular biology).
  - P*roteins* are long chains of amino acids
  - An *amino acids* being an organic compound containing amongst others an amino group (NH_2) and a carboxylic acid group (COOH))
  - Think of aminco acids as 3-letter words of nucleotide building blocks (letters).

Every cell in the body has the same DNA

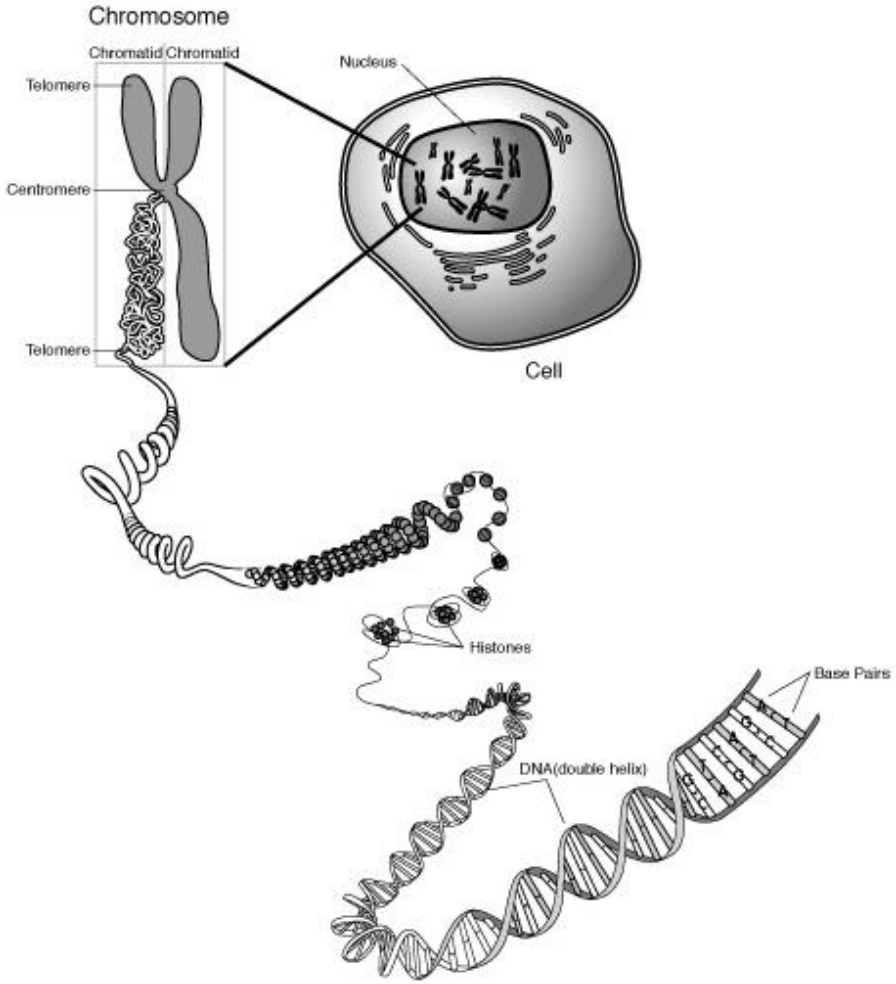


- One base pair is 0.00000000034 meters
- DNA sequence in any two people is 99.9% identical – only 0.1% is unique!

Chromosomes

- In the nucleus of each cell, the DNA molecule is packaged into thread-like structures called chromosomes. Each chromosome is made up of DNA tightly coiled many times around proteins called histones (see later) that support its structure.
- Chromosomes are not visible in the cell's nucleus—not even under a microscope—when the cell is not dividing.
- However, the DNA that makes up chromosomes becomes more tightly packed during cell division and is then visible under a microscope. Most of what researchers know about chromosomes was learned by observing chromosomes during cell division.

Histones: packaging of DNA in the nucleus

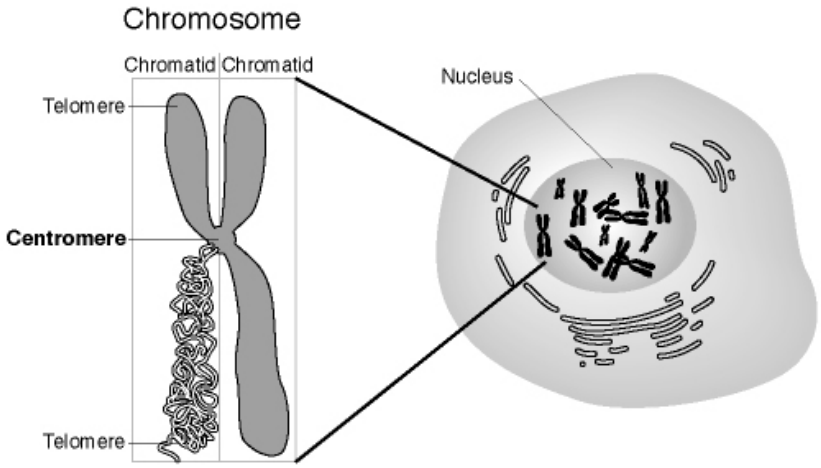


http://www.accessexcellence.org/AB/GG/chromosome.html

- *Histones* are proteins rich in lysine and arginine residues and thus positively-charged.
- For this reason they bind tightly to the negatively-charged phosphates in DNA.

Chromosomes

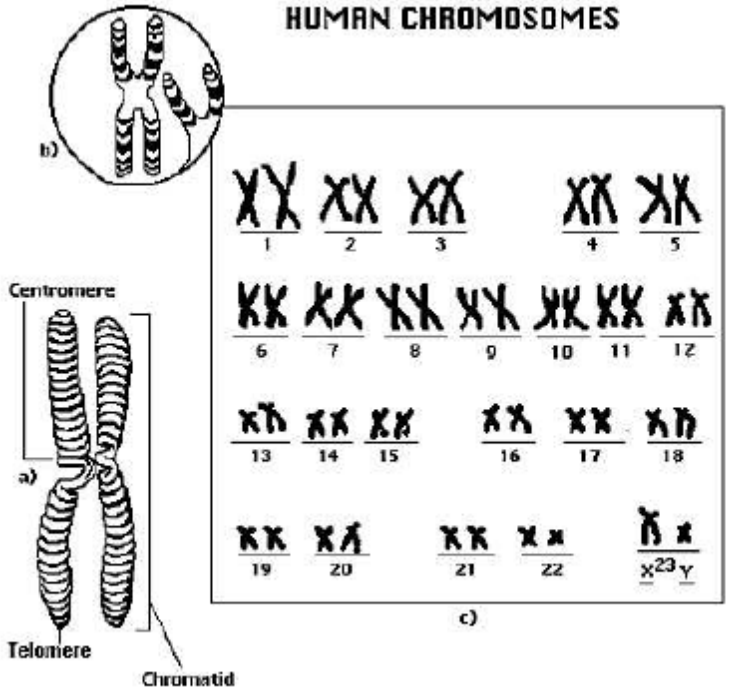
- All chromosomes have a stretch of repetitive DNA called the centromere. This plays an important role in chromosomal duplication before cell division.
- If the centromere is located at the extreme end of the chromosome, that chromosome is called acrocentric.
- If the centromere is in the middle of the chromosome, it is termed metacentric
- The ends of the chromosomes (that are not centromeric) are called telomeres. They play an important role in aging.



(www.genome.gov)

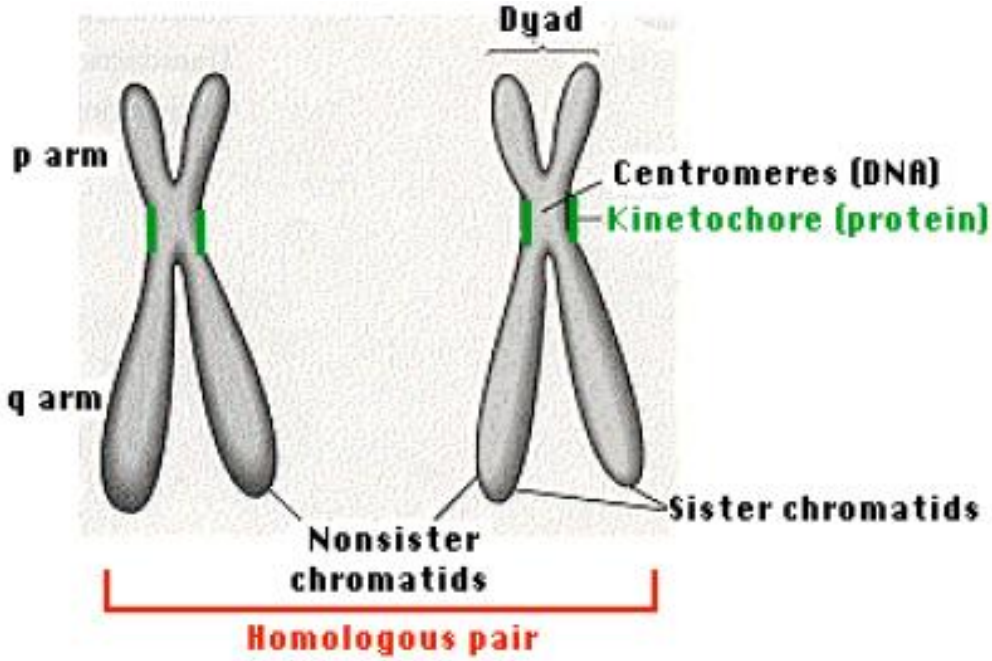
Chromosomes

- The short arm of the chromosome is usually termed *p* for petit (small), the long arm, *q*, for queue (tall).
- The telomeres are correspondingly referred to as *pter* and *qter*.



Chromatids

- A chromatid is one among the two identical copies of DNA making up a replicated chromosome, which are joined at their centromeres, for the process of cell division (mitosis or meiosis – see later).



Sex chromosomes

- Homogametic sex :

  that sex containing two like sex chromosomes

  - In most animals species these are females (XX)
  - Butterflies and Birds, ZZ males

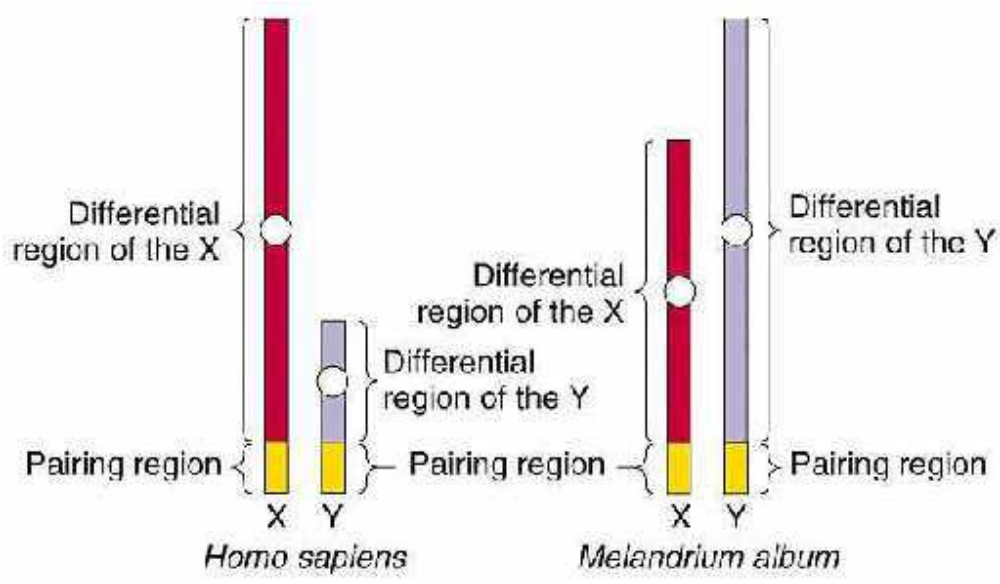
- Heterogametic sex:

  that sex containing two different sex chromosomes

  - In most animal species these are XY males
  - Butterflies and birds, ZW females
  - Grasshopers have XO males

Pairing of sex chromosomes

- In the homogametic sex: pairing happens like normal autosomal chromosomes
- In the heterogametic sex: The two sex chromosomes are very different, and have special pairing regions to insure proper pairing at meiosis



X-inactivation

- X-inactivation (also called lyonization) is a process by which one of the two copies of the X chromosome present in female mammals is inactivated
- X-inactivation occurs so that the female, with two X chromosomes, does not have twice as many X chromosome gene products as the male, which only possess a single copy of the X chromosome

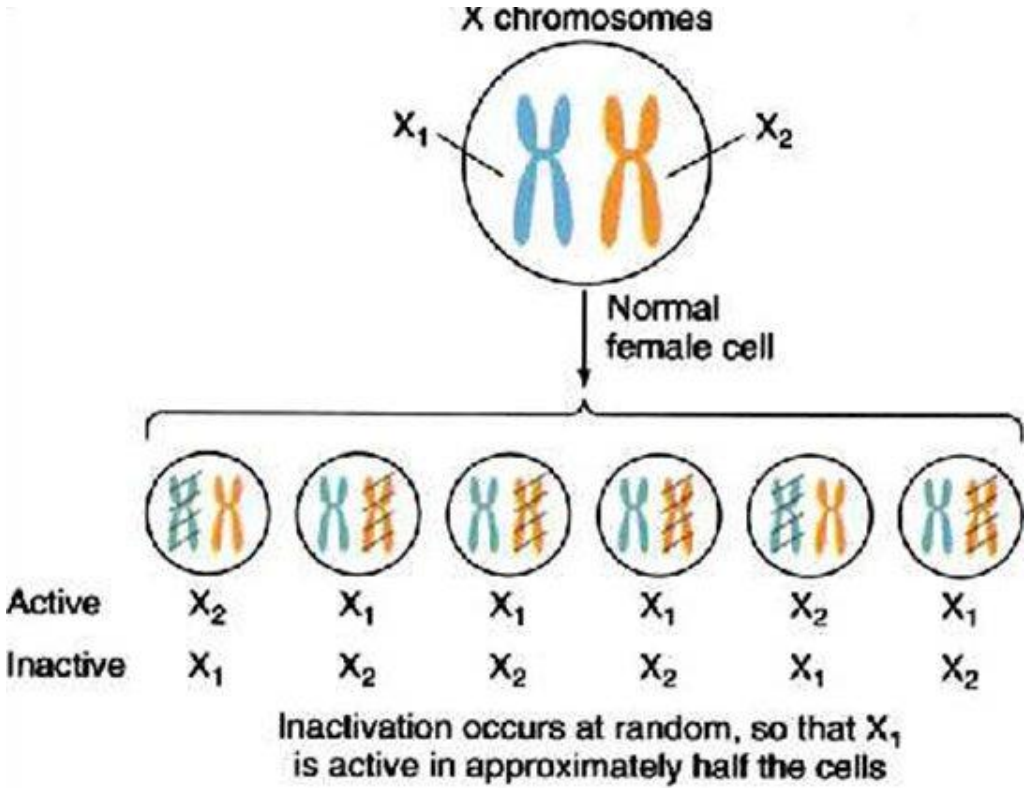
The ginger colour of cats (known as "yellow", "orange" or "red" to cat breeders) is caused by the "O" gene. The O gene changes black pigment into a reddish pigment. The O gene is carried on the X chromosome. A normal male cat has XY genetic makeup; he only needs to inherit one O gene for him to be a ginger cat. A normal female is XX genetic makeup. She must inherit two O genes to be a ginger cat. The O gene is called a sex-linked gene because it is carried on a sex chromosome.

If the female cat inherits only one O gene, she will be tortoiseshell (heterozygous for red colour). The formation of red and black patches in a female cat with only one O gene is through a process known as X-chromosome inactivation. Some cells randomly activate the O gene while others activate the gene in the equivalent place on the other X chromosome. → **epigenetic inheritance**

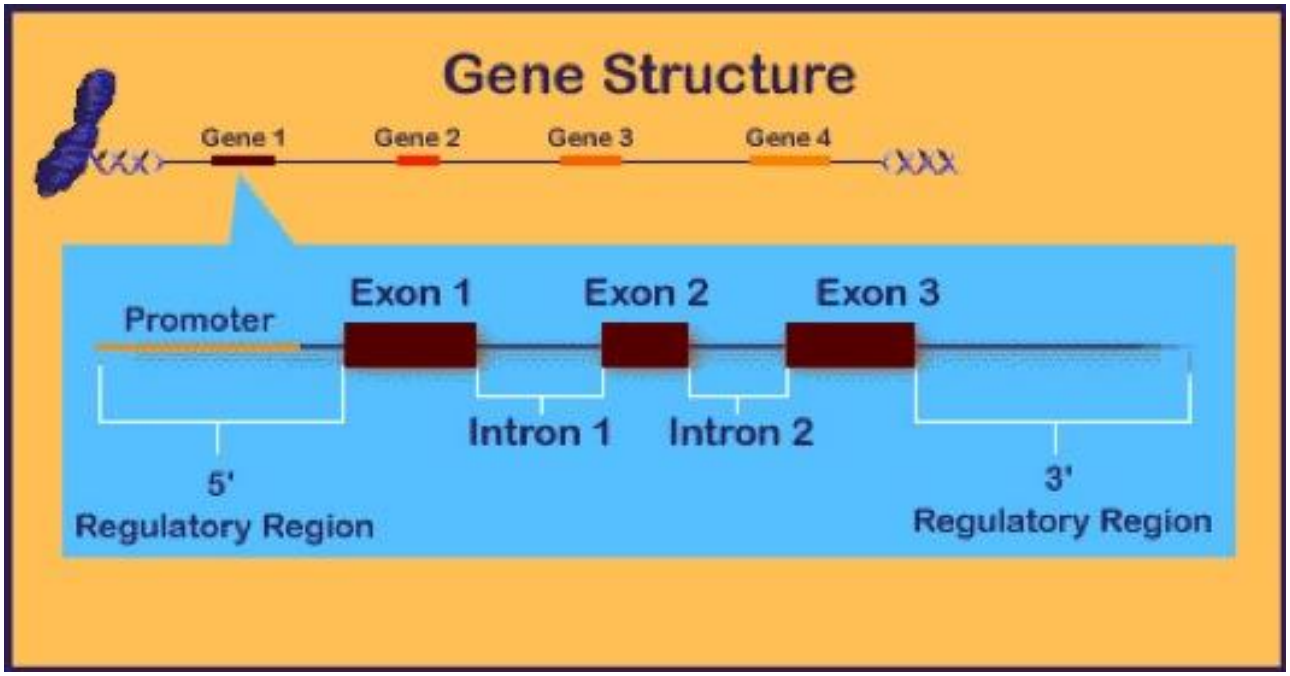
(wikipedia)

X-inactivation

- The choice of which X chromosome will be inactivated is random in placental mammals such as mice and humans, but once an X chromosome is inactivated it will remain inactive throughout the lifetime of the cell.



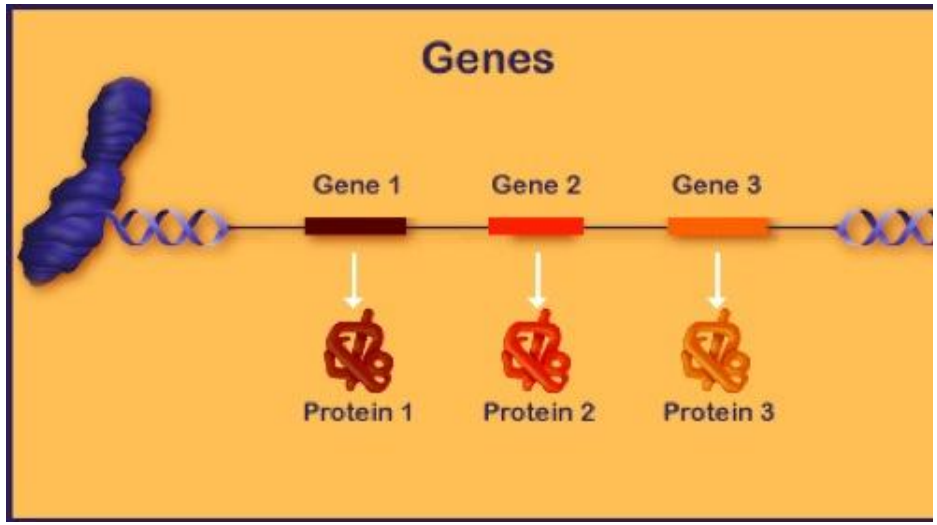
1.b What does the genetic information mean?

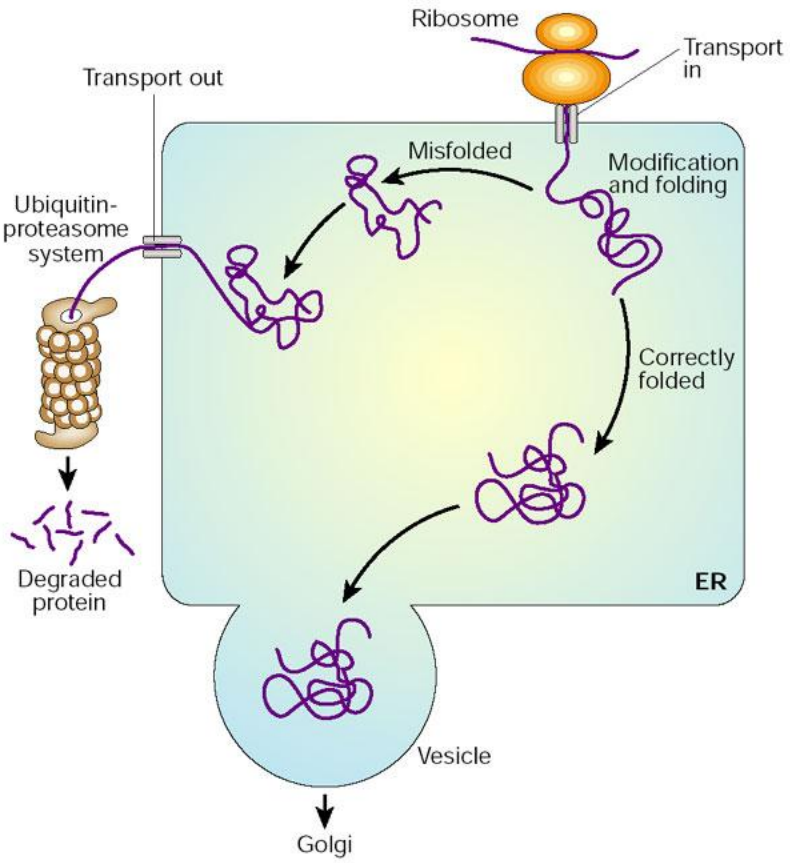


(Roche Genetics)

- *Promoter*: Initial binding site for RNA polymerase in the process of gene expression. First transcription factors bind to the promoter which is located 5' to the transcription initiation site in a gene.

Genes and Proteins

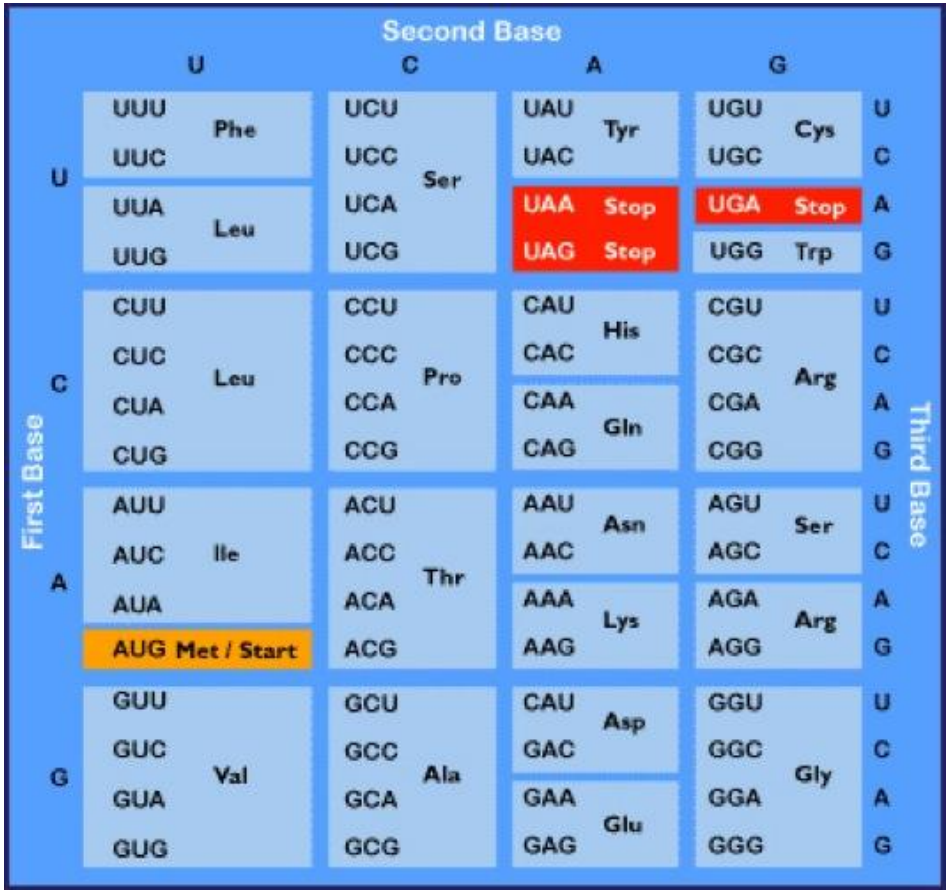




(Roche Genetics)

(http://www.nature.com/nature/journal/v426/n6968/images/nature02261-f2.2.jpg)

Translation table from DNA building stones to protein building stones



(Roche Genetics)

- Where does the U come from?

Comparison between DNA and RNA

- Pieces of coding material that the cells needs at a particular moment, is transcribed from the DNA in RNA for use outside the cell nucleus.

Char	DNA	RNA
Major cellular site	nucleus	cytoplasm (cell area outside nucleus)
Major function	genetic material;	carries out instructions
	directs protein synthesis;	for protein synthesis
	replicates itself before cell div.	
Sugar	deoxyribose	ribose
Bases	A, C, T, G	A, C, U(racil), G
Structure	double strand coiled	single straight or
	into a double helix	folded strand

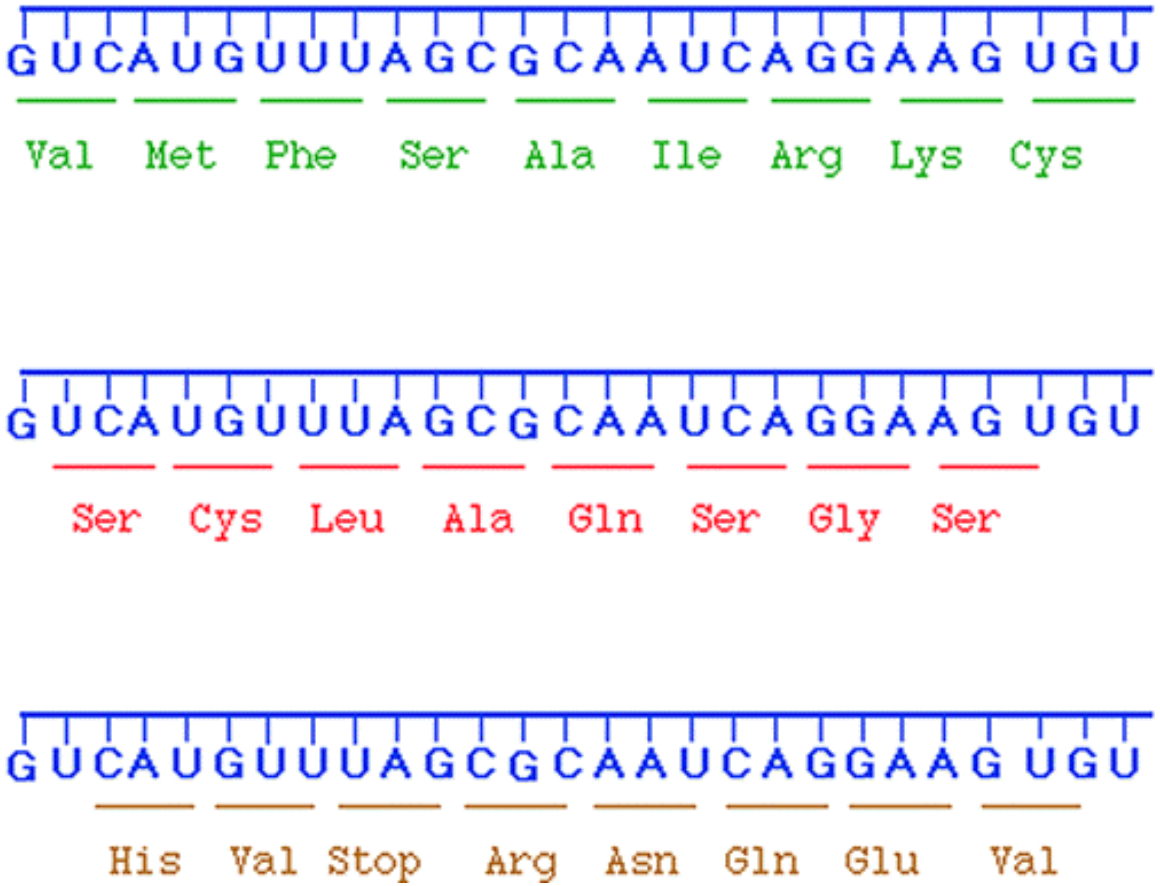
(Human Anatomy & Physiology - Addison-Wesley 4th ed)

- Note that in RNA U(racil), another pyrimidine, replaces T in DNA

Reading the code

- Because there are only 20 amino acids that need to be coded (using A, C, U or G), the genetic code can be said to be degenerate, with the third position often being redundant
- The code is read in triplets of bases.
- Depending on the starting point of reading, there are three possible variants to translate a given base sequence into an amino acid sequence. These variants are called *reading frames*

Reading the code

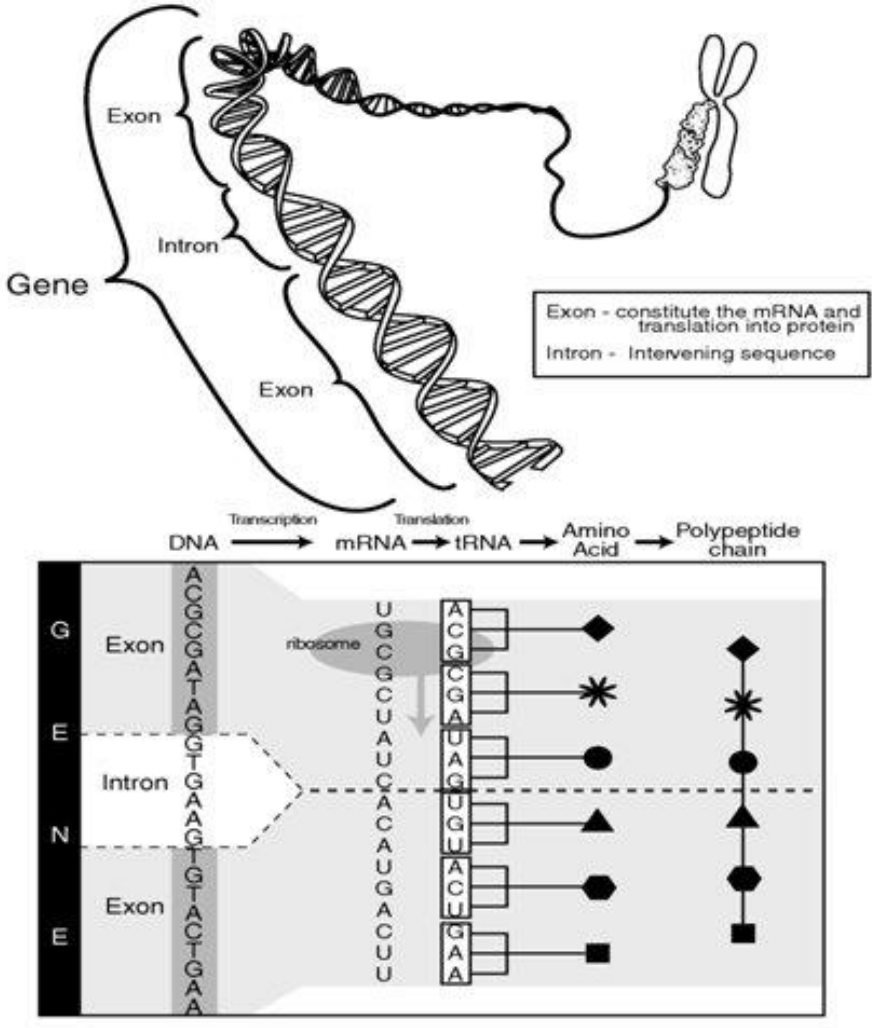


1.c How is the genetic information translated?

The link between genes and proteins: nucleotide bases

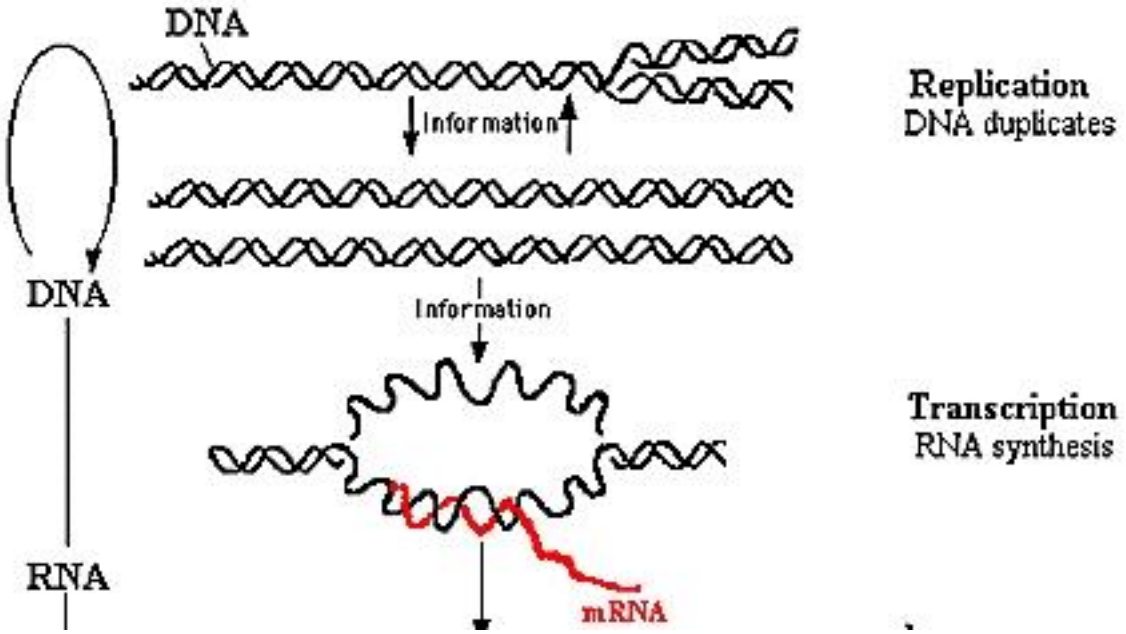
- A gene codes for a protein, but also has sections concerned with gene expression and regulation (E.g., promoter region)
- The translation of bases into amino acids uses RNA and not DNA; it is initiated by a START codon and terminated by a STOP codon.
- Hence, it are the three-base sequences (codons) that code for amino acids and sequences of amino acids in turn form proteins

DNA makes RNA, RNA makes proteins, proteins make us

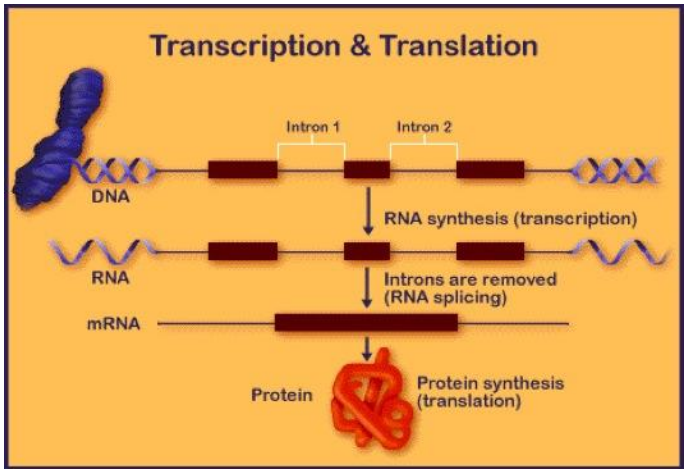


Central dogma of molecular biology

- Stage 1: DNA replicates its information in a process that involves many enzymes. This stage is called the ***replication*** stage.



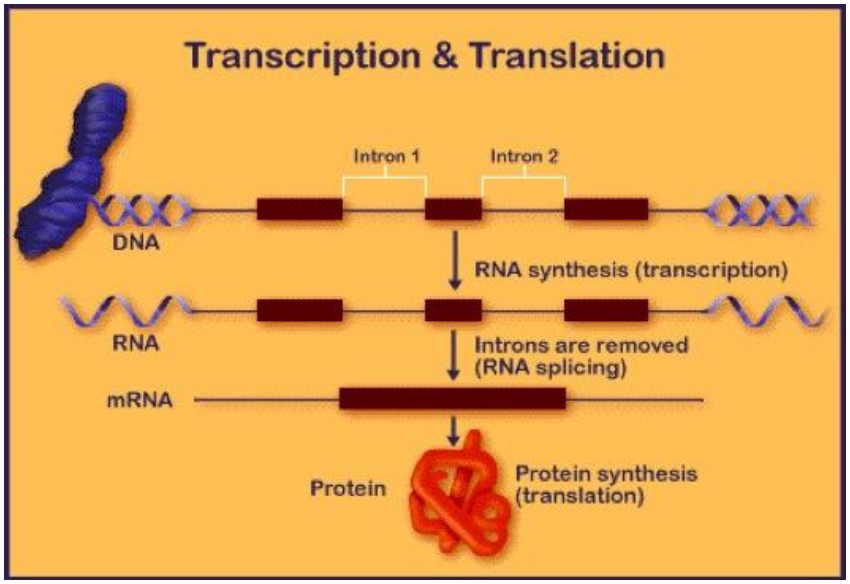
- Stage 2: The DNA codes for the production of messenger RNA (mRNA) during ***transcription*** of the sense strand (coding or non-template strand)



(Roche Genetics)

So the *coding strand* is the DNA strand which has the same base sequence as the RNA transcript produced (with thymine replaced by uracil). It is this strand *which contains codons*, while the non-coding strand (or anti-sense strand) contains anti-codons.

- Stage 3: In eukaryotic cells, the mRNA is ***processed*** (essentially by splicing) and migrates from the nucleus to the cytoplasm

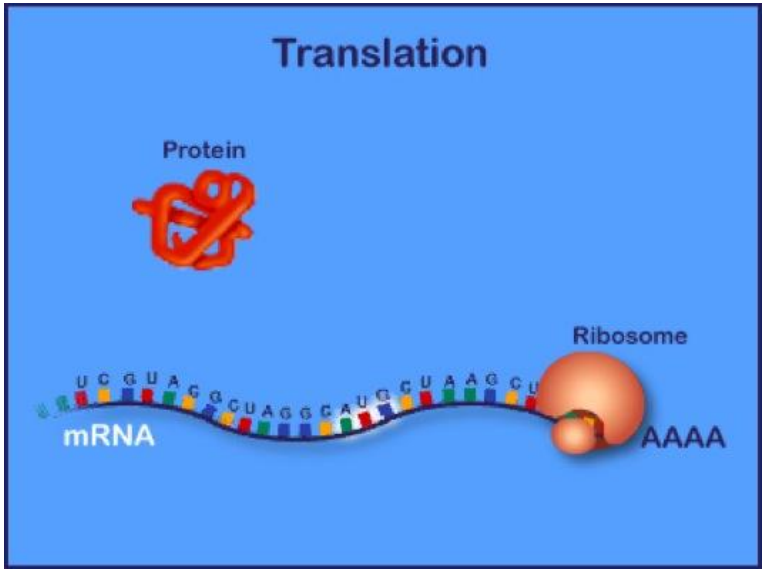


(Roche Genetics)

- Stage 4: mRNA carries coded information to ribosomes. The ribosomes "read" this information and use it for protein synthesis. This stage is called the ***translation*** stage.

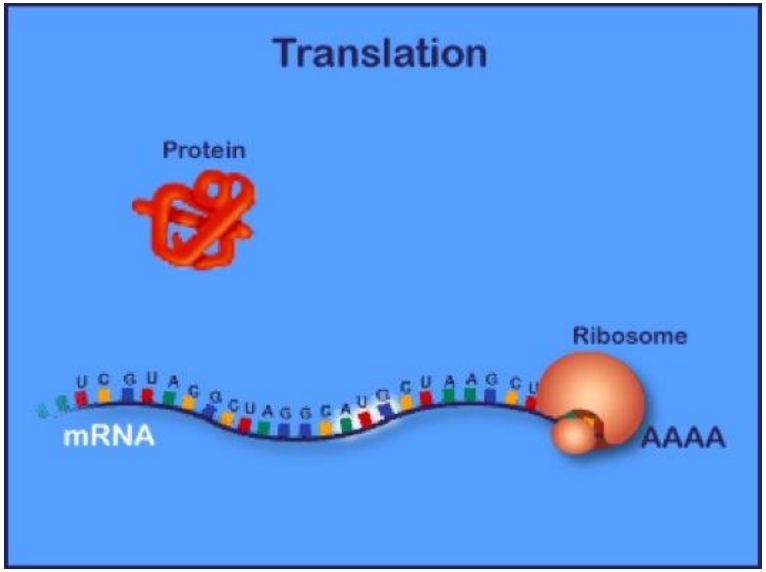
  The direction of reading mRNA is 5' to 3'. tRNA (reading 3' to 5') has anticodons complementary to the codons in mRNA

Translation is facilitated by two key molecules



- *Transfer RNA* (tRNA) molecules transport amino acids to the growing protein chain. Each tRNA carries an amino acid at one end and a three-base pair region, called the anti-codon, at the other end. The anti-codon binds with the codon on the protein chain via base pair matching.

Translation is facilitated by two key molecules (continued)

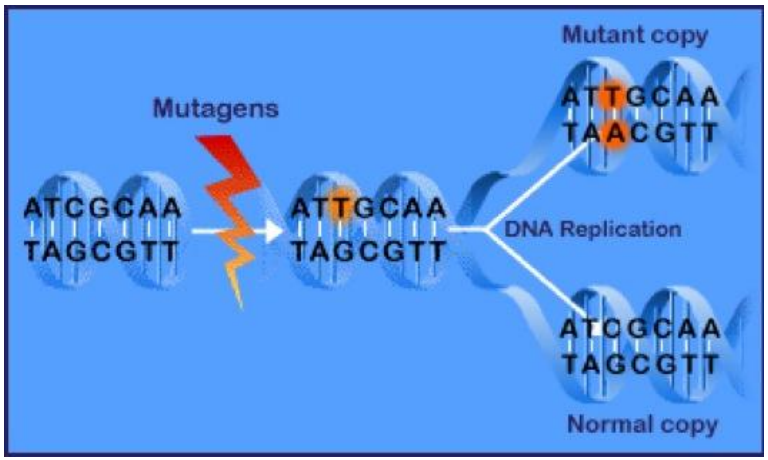


(Roche Genetics)

- *Ribosomes* bind to the mRNA and facilitate protein synthesis by acting as docking sites for tRNA. Each ribosome is composed of a large and small subunit, both made of ribosomal RNA (rRNA) and proteins. The ribosome has three docking sites for tRNA

DNA repair mechanisms

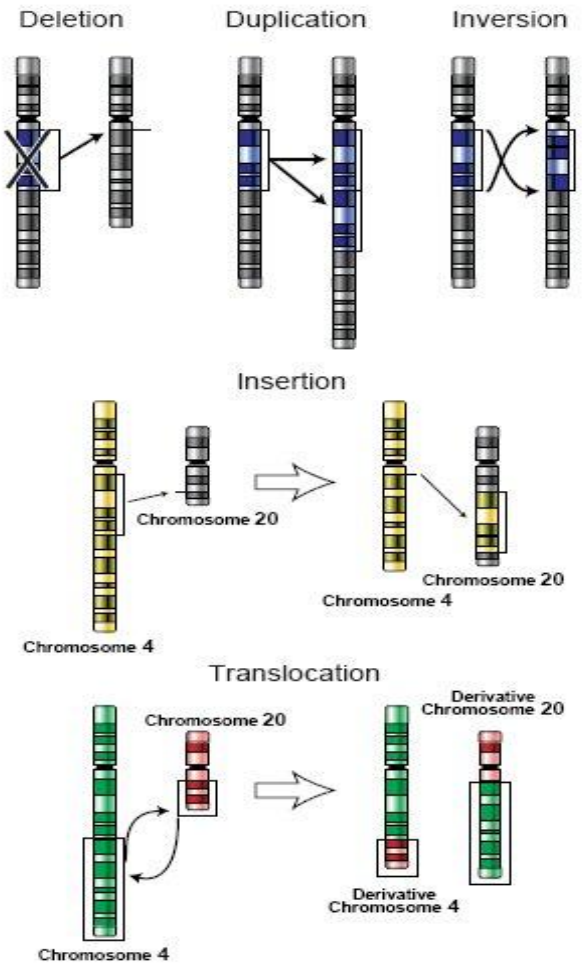
- In biology, a mutagen (Latin, literally origin of change) is a physical or chemical agent that changes the genetic material (usually DNA) of an organism and thus increases the frequency of mutations above the natural background level.
- As many mutations cause cancer, mutagens are typically also carcinogens.
- Not all mutations are caused by mutagens: so-called "spontaneous mutations" occur due to errors in DNA replication, repair and recombination.



(Roche genetics)

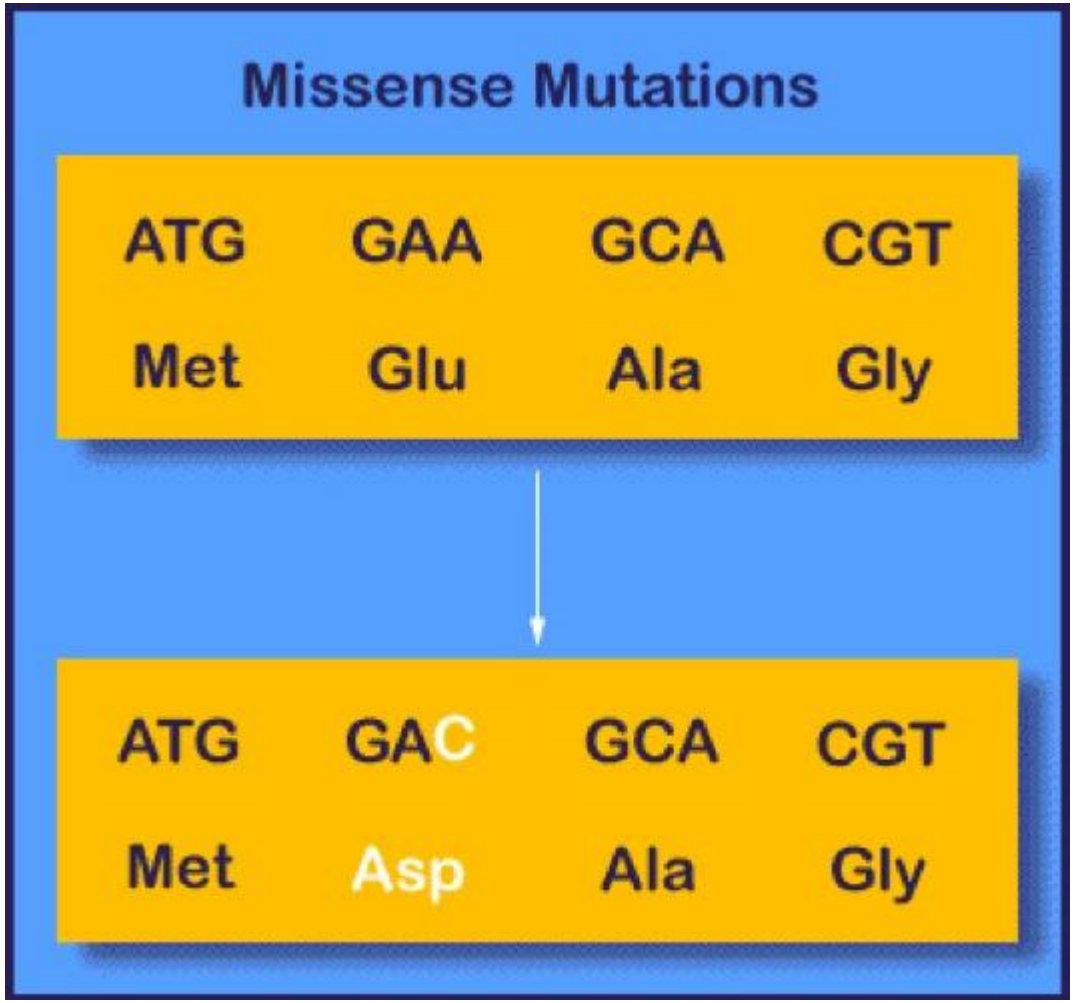
Types of mutations

- Deletion
- Duplication
- Inversion
- Insertion
- Translocation

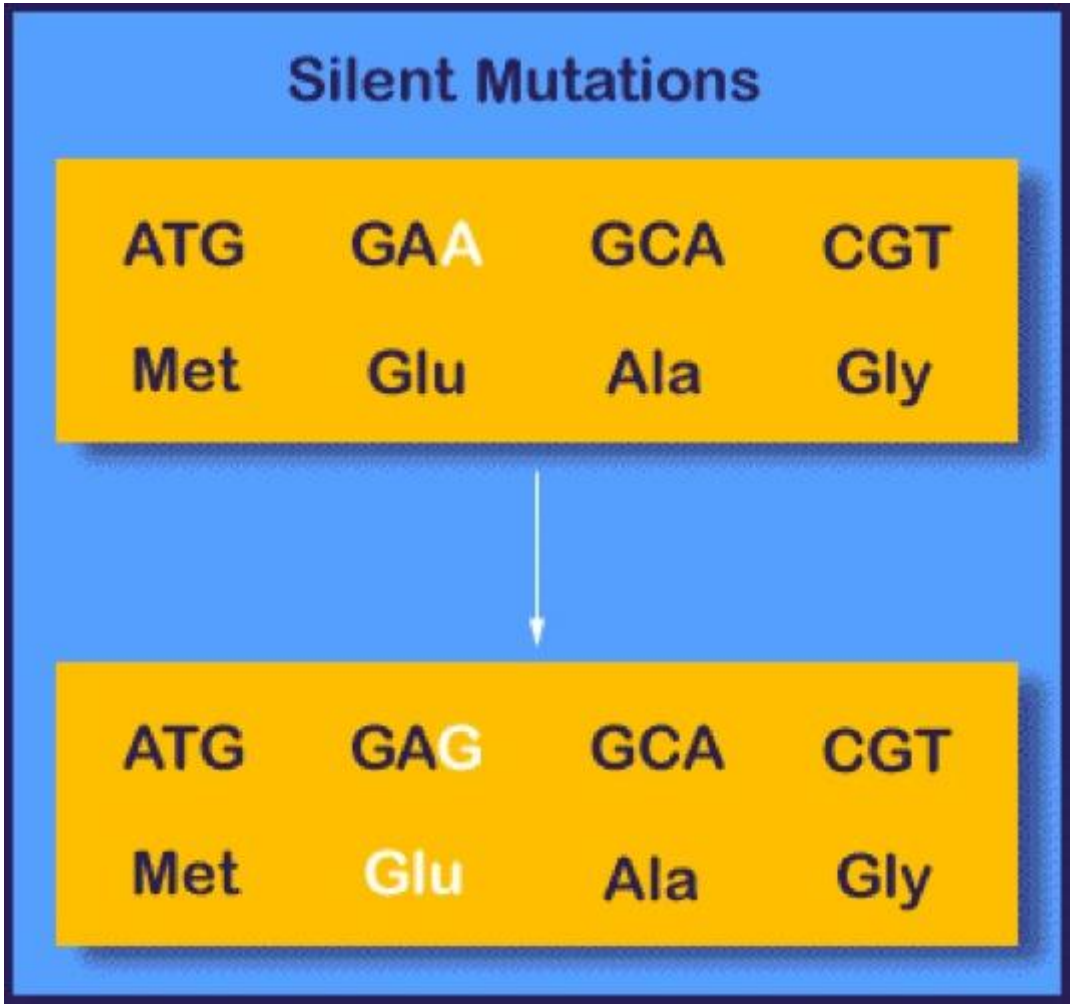


(National Human Genome Research Institute)

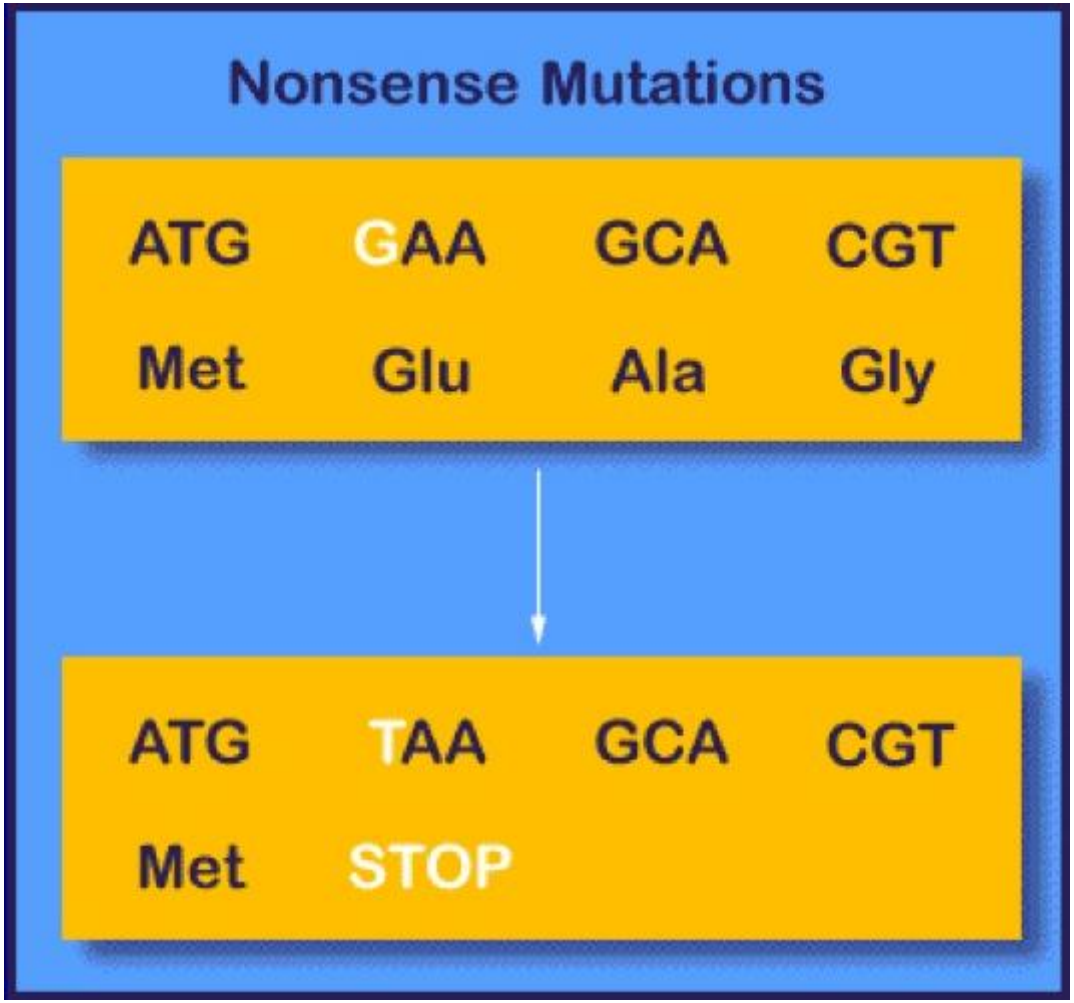
Types of mutations (continued)



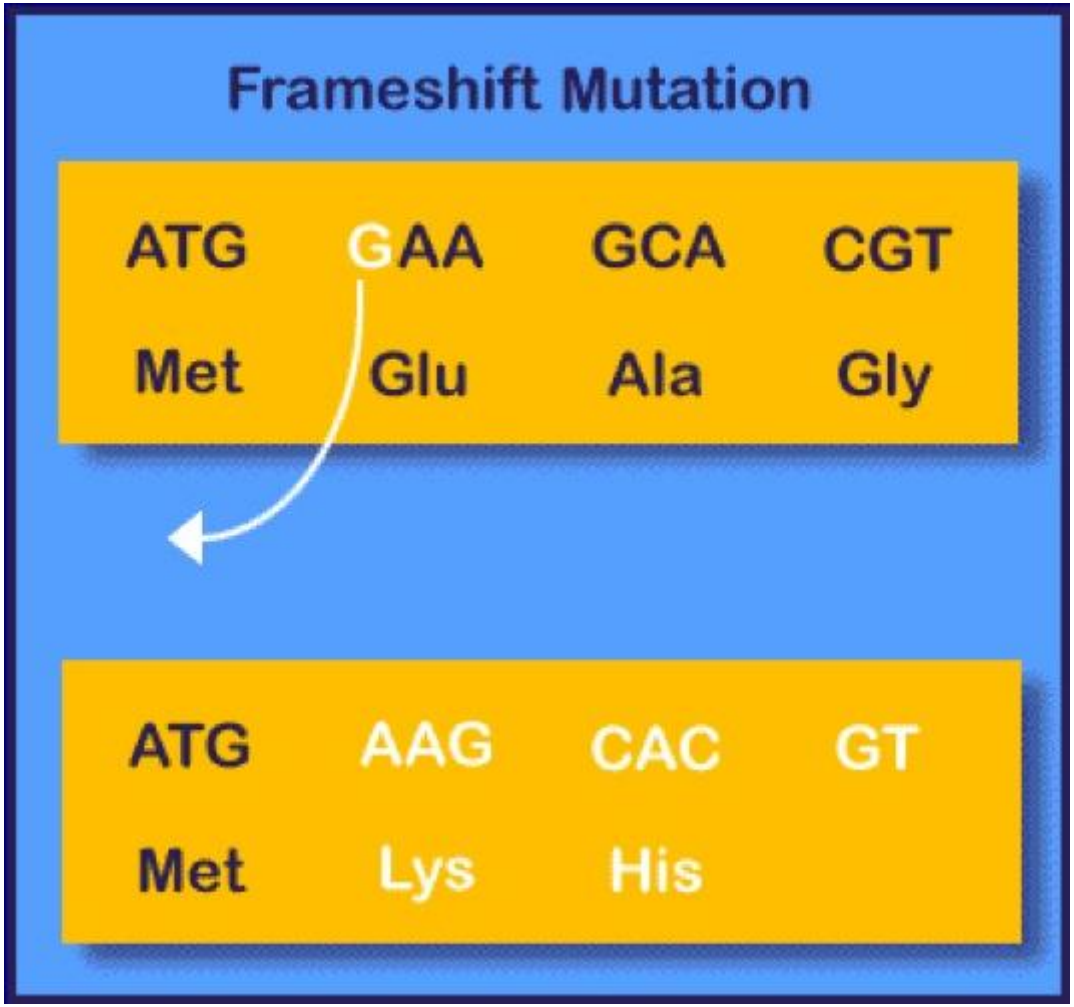
Types of mutations (continued)



Types of mutations (continued)

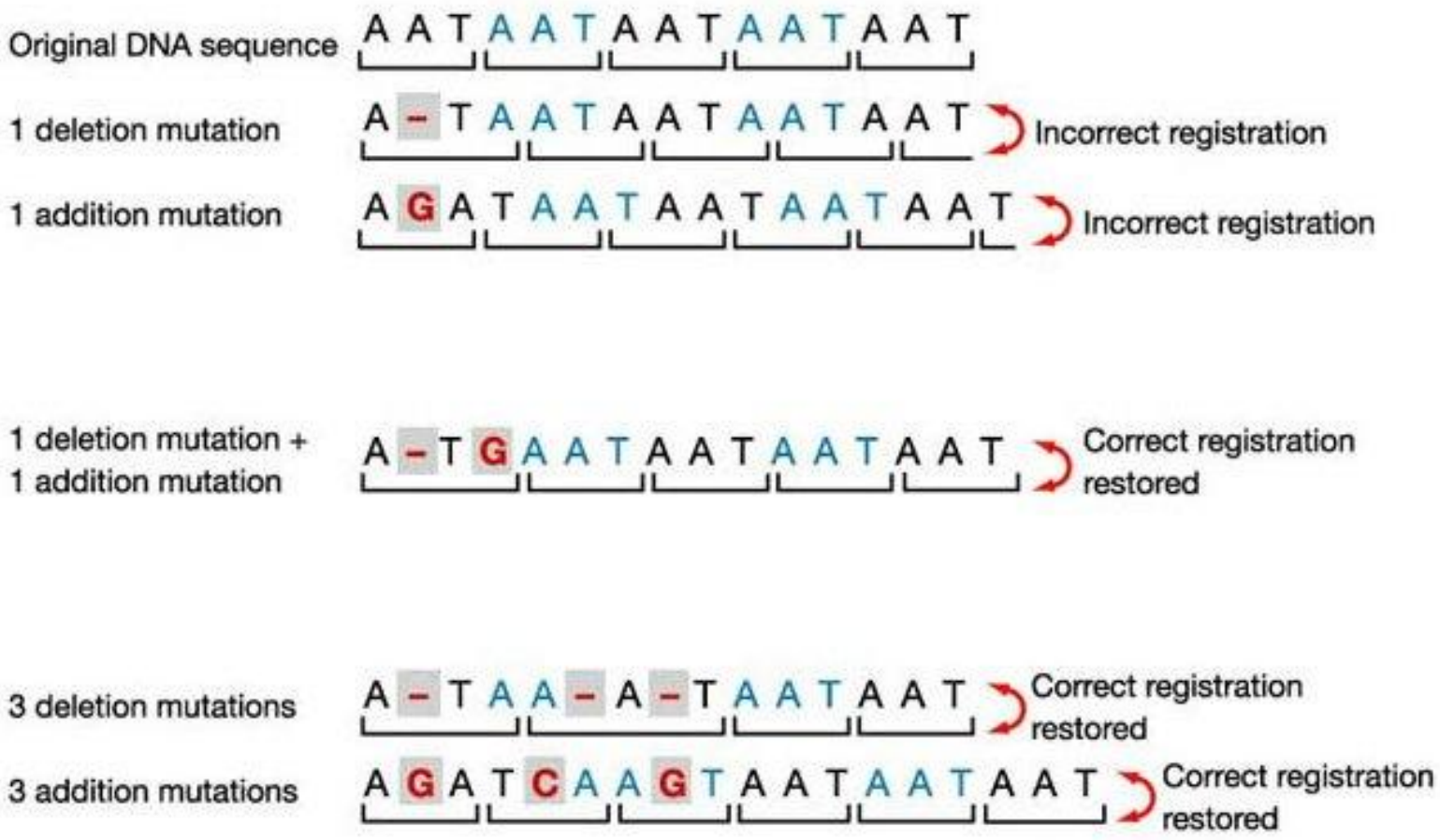


Types of mutations (continued)



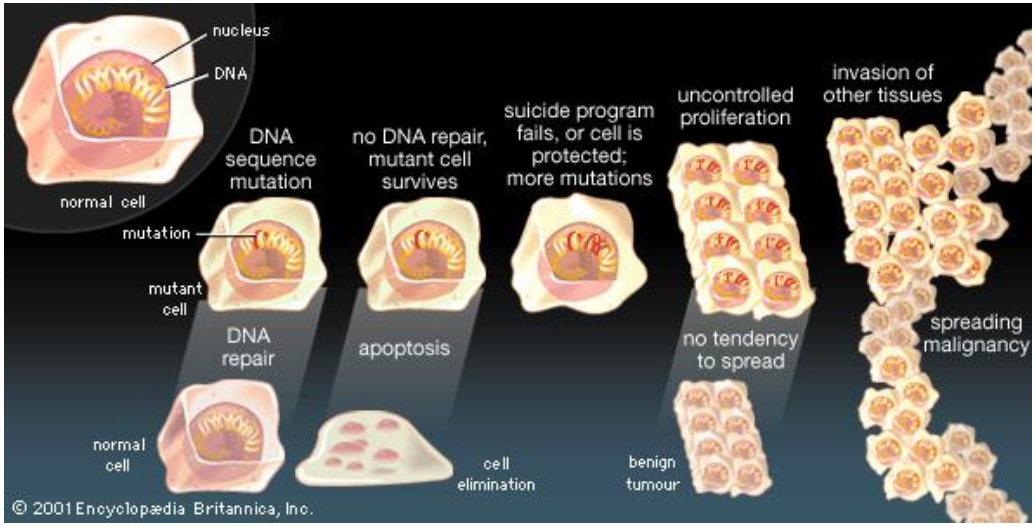
DNA repair mechanisms

- Where it can go wrong when reading the code …



DNA repair mechanisms

- damage reversal: simplest; enzymatic action restores normal structure without breaking backbone
- damage removal: involves cutting out and replacing a damaged or inappropriate base or section of nucleotides
- damage tolerance: not truly repair but a way of coping with damage so that life can go on



DATABASES

Human Mutation

OFFICIAL JOURNAL HGVS HUMAN GENOME VARIATION SOCIETY www.hgvs.org

The Roche Cancer Genome Database (RCGDB)

Jan Küntzer,[1]* Daniela Eggle,[1] Hans-Peter Lenhof,[2] Helmut Burtscher,[3] and Stefan Klostermann[1]

[1]Roche Diagnostics GmbH, Pharma Research Scientific Informatics, Penzberg, Germany; [2]Saarland University, Center for Bioinformatics Saar, Saarbrücken, Germany; [3]Roche Diagnostics GmbH, Discovery Oncology, Penzberg, Germany

Communicated by Rolf Sijmons

Received 9 October 2009; accepted revised manuscript 7 January 2010.
Published online 2 February 2010 in Wiley InterScience (www.interscience.wiley.com). DOI 10.1002/humu.21207

ABSTRACT: Sequence variations are being studied for a better understanding of the mechanism and development of cancer as a mutation-driven disease. The systematic

cells that have accumulated the for the tumor most favorable set of genetic aberrations [Pinkel and Albertson, 2005].

The estimate of only ~350 "cancer genes" [Futreal et al., 2004; Haber and Settleman, 2007] conflicts with the finding that many

(http://onlinelibrary.wiley.com/doi/10.1002/humu.21207/pdf)

Distinguish between polymorphisms and mutations

- With have already introduced the concept of a genetic marker. In general they can also be seen as “flagposts” to capture genetic variation.
- The verb mutation describes the process by which new variants of a gene arise. As a noun it is used to describe a rare variant of a gene.
- Polymorphisms are more common variants (more than 1%).
- Most mutations will disappear but some will achieve higher frequencies due either to random genetic drift or to selective pressure
- The most common forms of variants are:
  - repeated sequences of 2, 3 or 4 nucleotides (microsatellites)
  - single nucleotide polymorphisms (SNPs) in which one letter of the code is altered

Non-synonymous SNP

- A SNP that alters the DNA sequence in a coding region such that the amino acid coding is changed.
- The new code specifies an alternative amino acid or changes the code for an amino acid to that for a stop translation signal or vice versa.
- Non-synonymous SNPs are sometimes referred to as coding SNPs.

Synonymous SNP

- Synonymous SNPs alter the DNA sequence but do not change the protein coding sequence as interpreted at translation, because of redundancy in the genetic code.
- Exonic SNPs may or may not cause an amino acid change

2 Overview of human genetics

2.a How is the genetic information transmitted from generation to generation

Understanding heredity

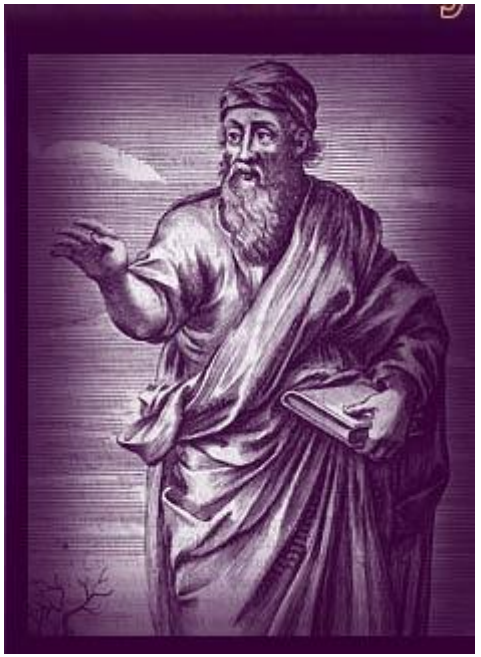
- Pythagoras
- Empedocles
- Aristotle
- Harvey
- Leeuwenhoek
- de Maupertuis
- Darwin
- Mendel
- Morgan
- Crick & Watson
- McClintock

(http://www.pbs.org/wgbh/nova/genome)

Pythagoras (580-500 BC)

Pythagoras surmised that all hereditary material came from a child's father. The mother provided only the location and nourishment for the fetus.

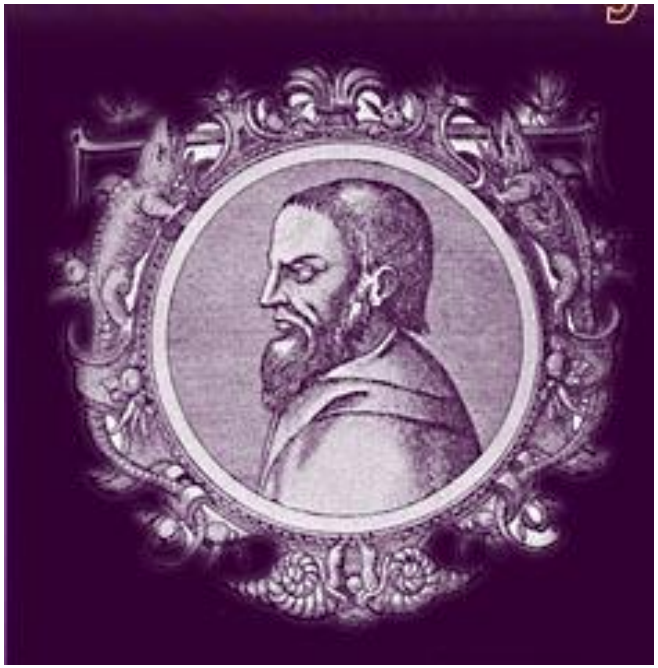
Semen was a cocktail of hereditary information, coursing through a man's body and collecting fluids from every organ in its travels. This male fluid became the formative material of a child once a man deposited it inside a woman.



Aristotle (384-322 BC)

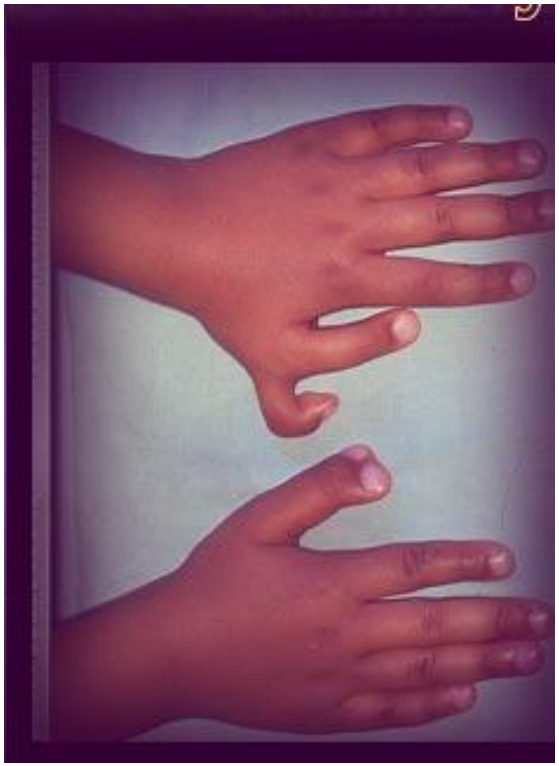
Aristotle's understanding of heredity, clearly following from Pythagorean thought, held wide currency for almost 2,000 years.

The Greek philosopher correctly believed that both mother and father contribute biological material toward the creation of offspring, but he was mistakenly convinced that a child is the product of his or her parents' commingled blood.



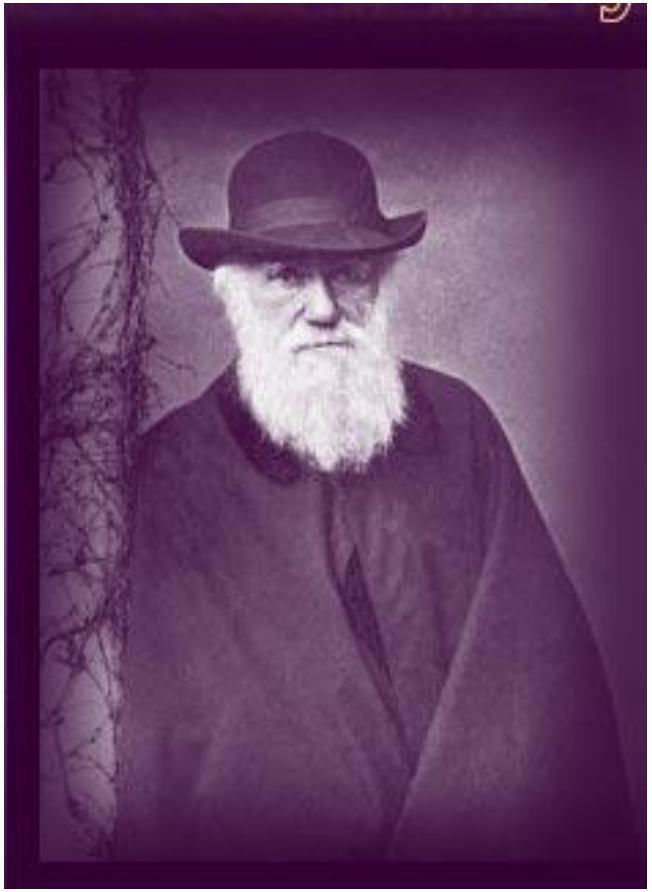
De Maupertuis (1698-1759)

In his 1751 book, Système de la nature (System of Nature), French mathematician, biologist, and astronomer Pierre-Louis Moreau de Maupertuis initiated the first speculations into the modern idea of dominant and recessive genes. De Maupertuis studied the occurrences of polydactyly (extra fingers) among several generations of one family and showed how this trait could be passed through both its male and female members.



Darwin (1809-1882)

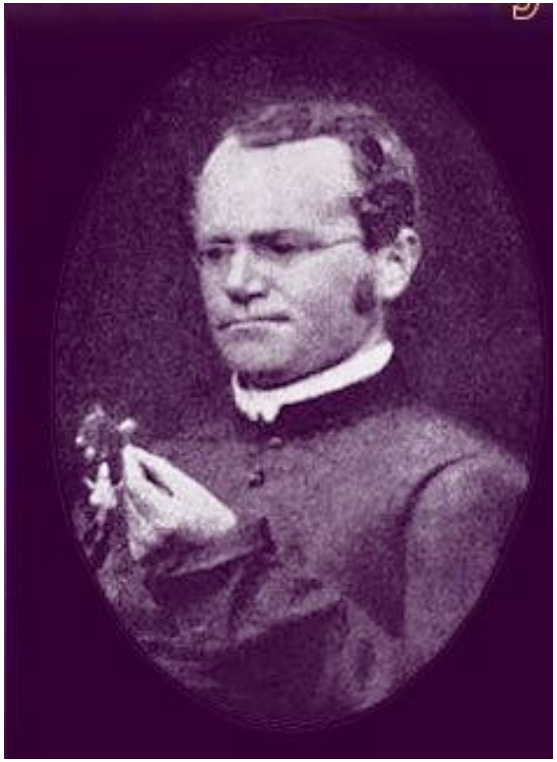
Darwin's ideas of heredity revolved around his concept of "pangenesis." In pangenesis, small particles called pangenes, or gemmules, are produced in every organ and tissue of the body and flow through the bloodstream. The reproductive material of each individual formed from these pangenes was therefore passed on to one's offspring.



Here we meet again … our friend Mendel (1822-1884)

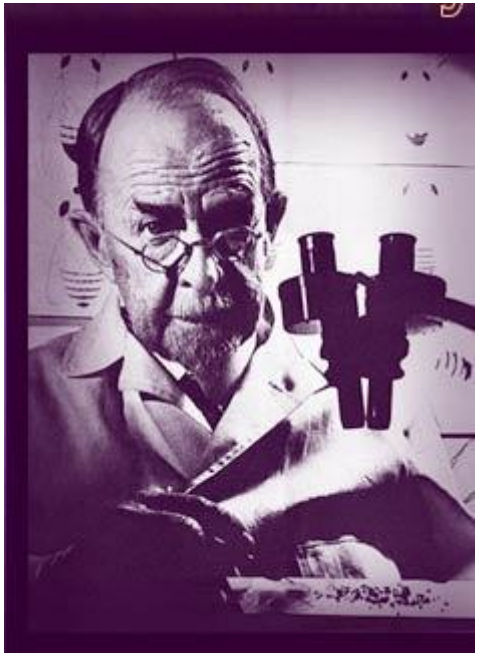
Gregor Mendel, an Austrian scientist who lived and conducted much of his most important research in a Czechoslovakian monastery, stablished the basis of modern genetic science. He experimented on pea plants in an effort to understand how a parent passed physical traits to its offspring. In one experiment, Mendel crossbred a pea plant with wrinkled seeds and a pea plant with smooth seeds.

All of the hybrid plants produced by this union had smooth seeds...



Morgan (1866-1945)

Thomas Hunt Morgan began experimenting with Drosophilia, the fruit fly, in 1908. He bred a single white-eyed male fly with a red-eyed female. All the offspring produced by this union, both male and female, had red eyes. From these and other results, Morgan established a theory of heredity that was based on the idea that genes, arranged on the chromosomes, carry hereditary factors that are expressed in different combinations when coupled with the genes of a mate.



Crick (1916-2004) and

Watson (1928-)

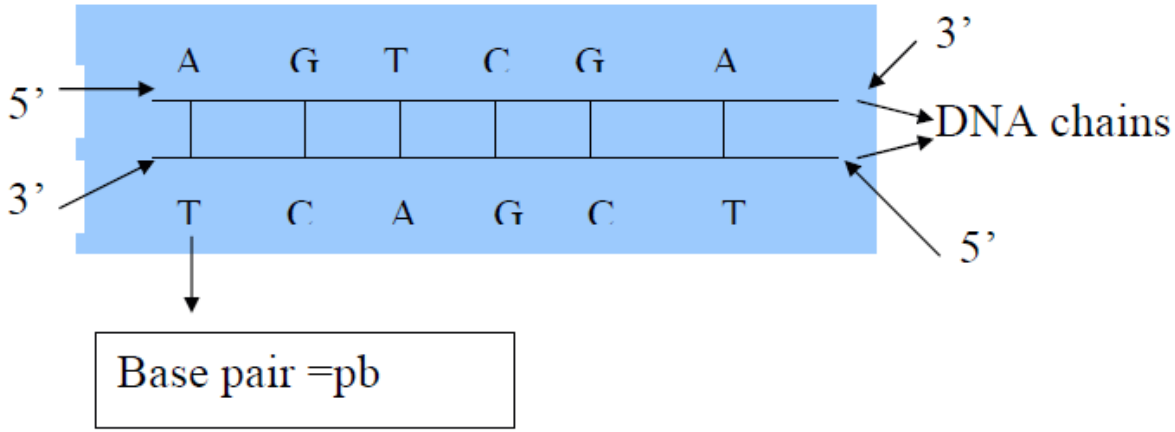
Employing X-rays and molecular models, Watson and **Crick** discovered the double helix structure of DNA. Suddenly they could explain how the DNA molecule duplicates itself by forming a sister strand to complement each single, ladder-like DNA template.

The MODERN human genome summarized

- The human genome consists of about 3×10^9 base pairs and contains about 22,000 genes

There are two DNA chains in one chromosome

— DNA has four bases A, G, T and C. A combined with T and G combined with C



The MODERN human genome (continued)

— bp is also used as length unit of chromosome or DNA sequence

— DNA sequence has direction. There are two sides (ends) called 5' side and 3' side.

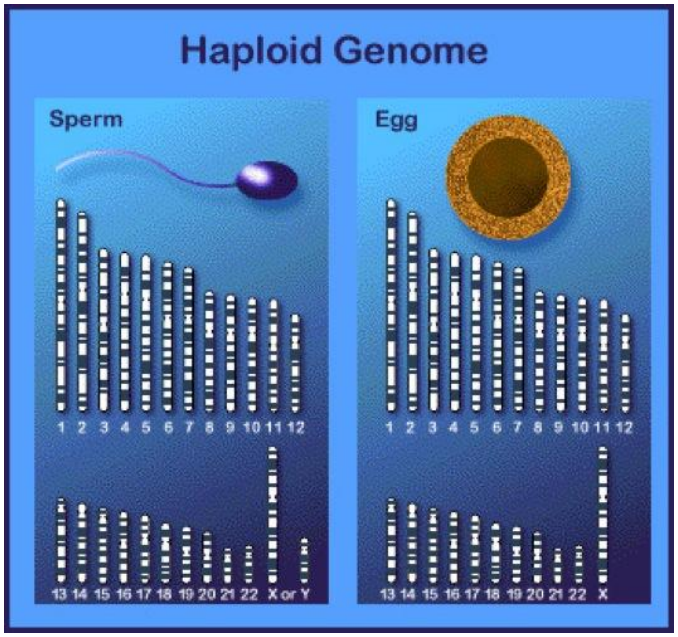
— The homologous chromosomes (chromosome 1, for example) have exactly same length for every individual.

The MODERN human genome (continued)

- Cells containing 2 copies of each chromosome are called diploid (most human cells). Cells that contain a single copy are called haploid.
- Humans have 23 pairs of chromosomes: 22 autosomal pairs (i.e., homologous pairs) and one pair of sex chromosomes
- Females have two copies of the X chromosome, and males have one X and one Y chromosome
- Much of the DNA is either in introns or in intragenic regions … which brings us to study the transmission or exploitation of genetic information in more detail.

Genetic information is inherited via meiosis

- Paternal genes (via sperm) and maternal genes (via egg) are donated to offspring
- Yet, parents won't lose genetic information, nor offspring will have too much genetic information



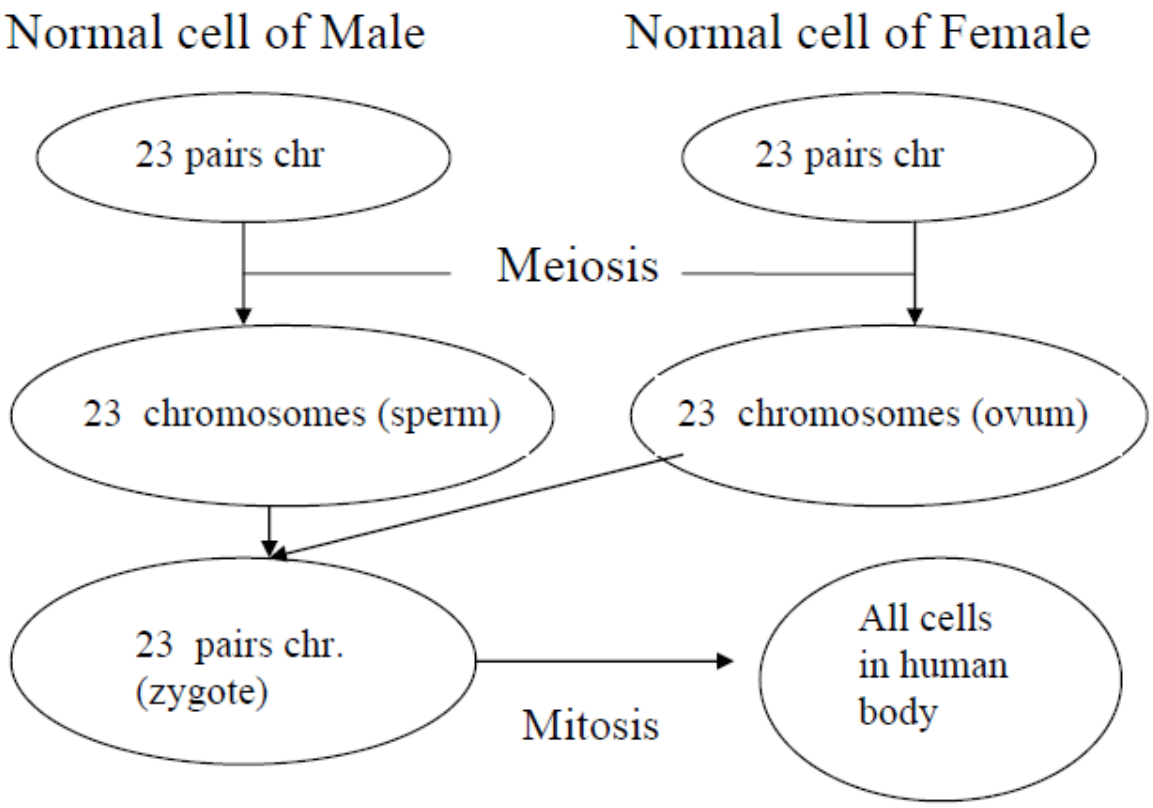
(Roche Genetics)

Meiosis in detail

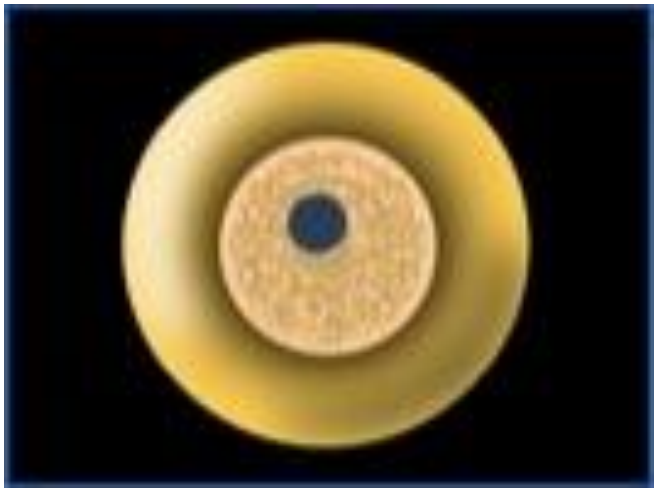
- Meiosis is a process to convert a diploid cell to a haploid gamete, and causes a change in the genetic information to increase diversity in the offspring.
- In particular, meiosis refers to the processes of cell division with two phases resulting in four haploid cells (gametes) from a diploid cell. In meiosis I, the already doubled chromosome number reduces to half to create two diploid cells each containing one set of replicated chromosomes. Genetic recombination between homologous chromosome pairs occurs during meiosis I. In meiosis II, each diploid cell creates two haploid cells resulting in four gametes from one diploid cell (mitosis).

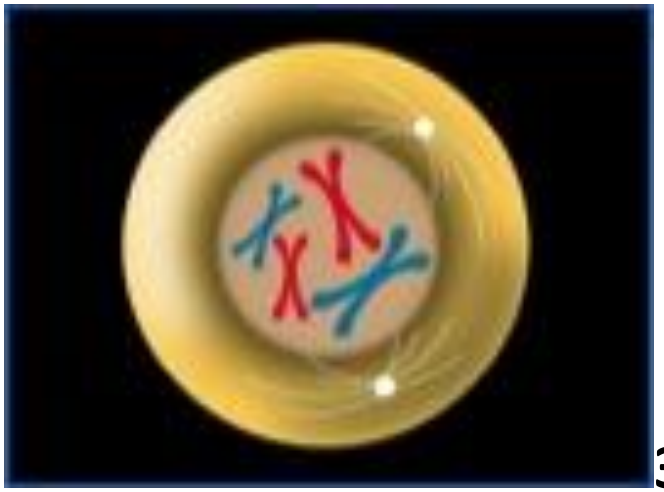
- Check out a nice demo to differentiate meiosis from mitosis: http://www.pbs.org/wgbh/nova/miracle/divide.html

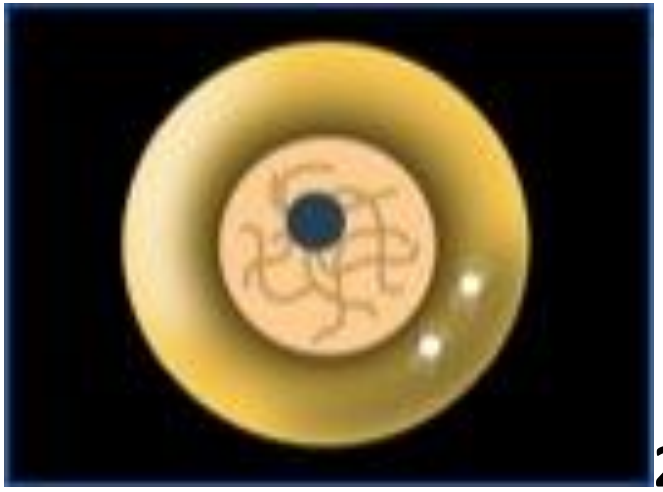
Meiosis in detail

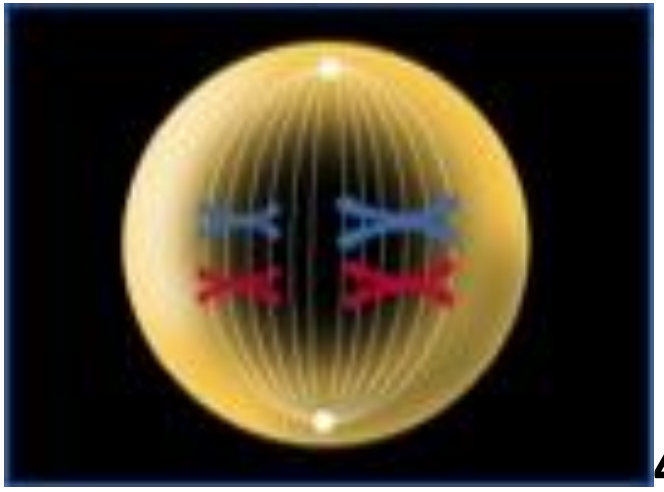


Meiosis in detail

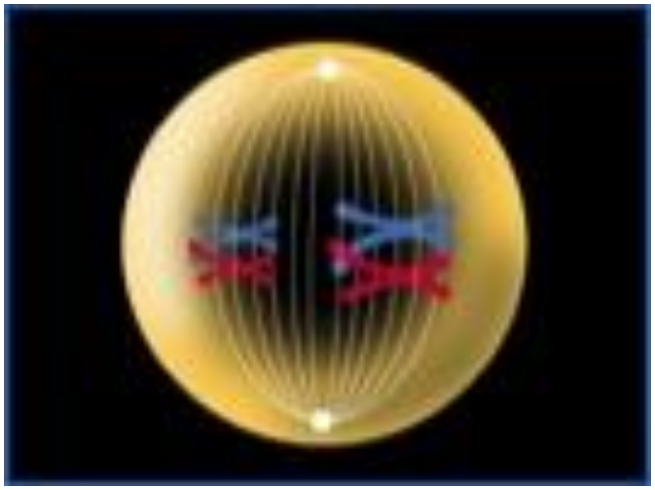

1


3

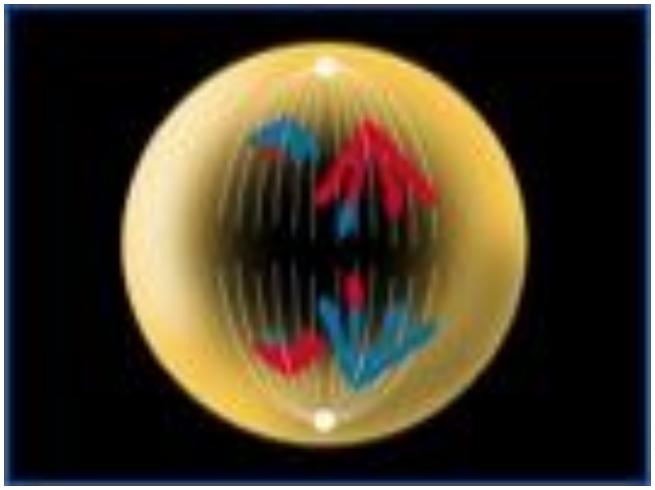

2


4

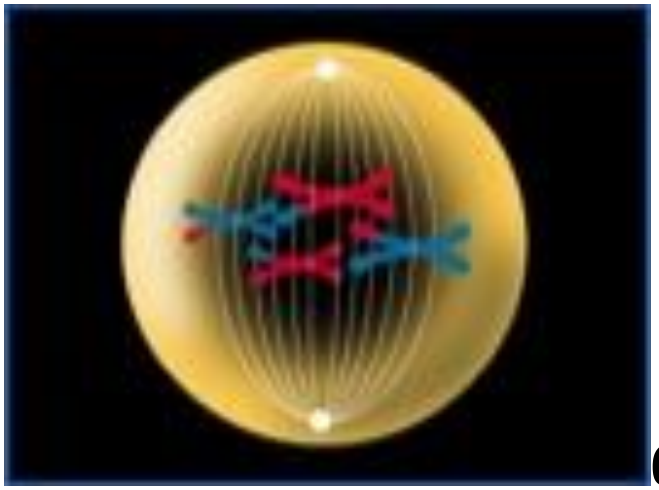
Meiosis in detail



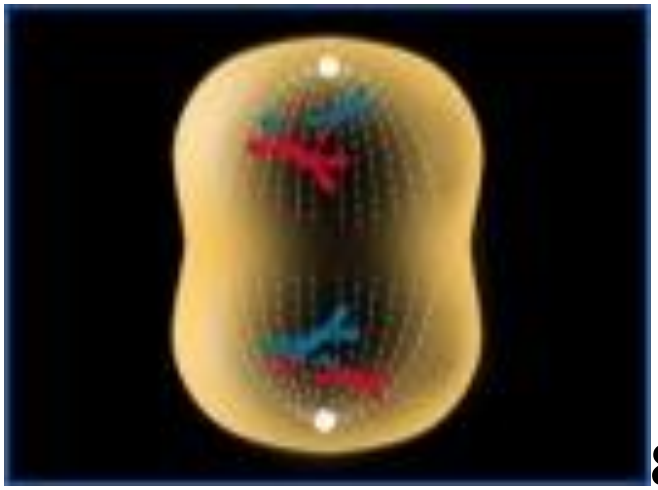
5



7

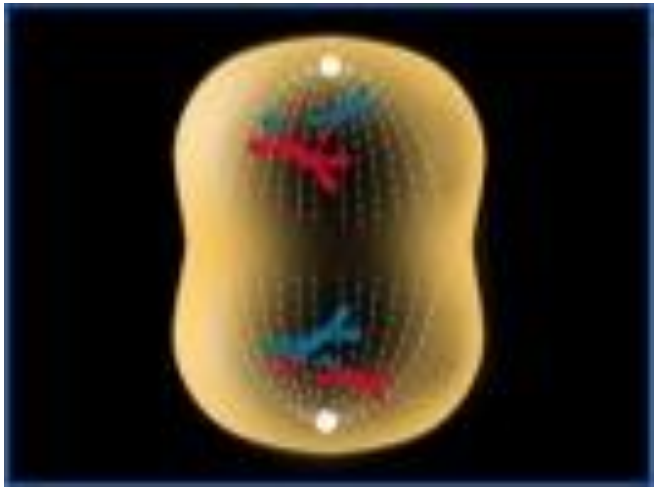


6



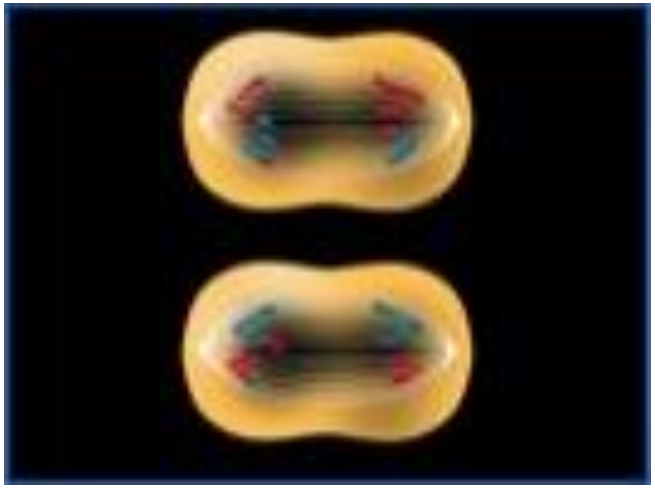
8

Meiosis in detail

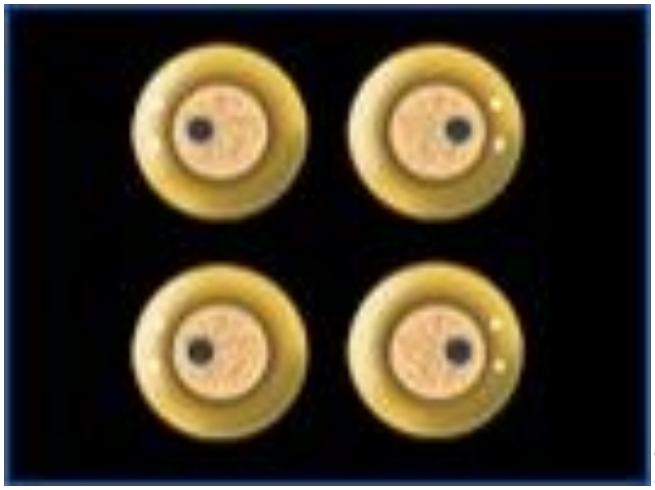


8

9



10



11

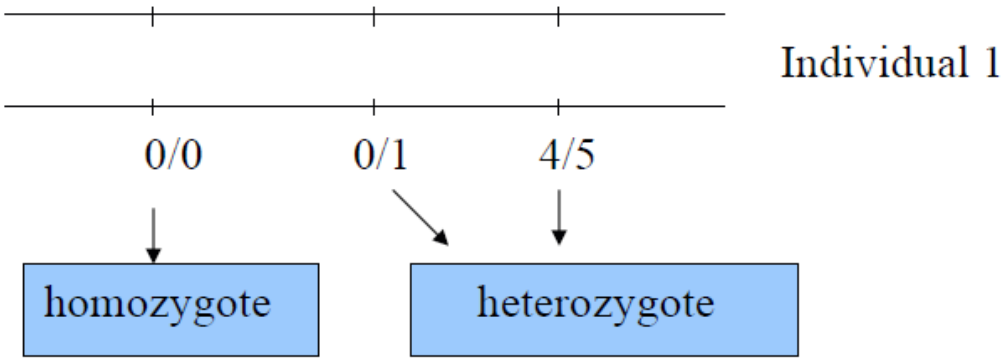
Genetic Terminology

- *Gene*:
  - A segment of DNA within a chromosome which has a specific genetic function
  - Length from several bps to several million bps
  - Gene is not the smallest unit of genetic material
  - Before the discovery of DNA, people believe that gene is the smallest unit
- *Locus*:
  - A specific position in chromosome. It may be 1bp or several bps in length
- Gene, marker and locus are sometimes used interchangeably in the literature!

- *Alleles*: DNA sequences within a locus (flavors of a gene or variant)
- *Genetic marker*:
  - Flagpost to capture genetic variation:
  - Two main kinds
    - SNPs (single nucleotide polymorphisms): 1bp in length; usually has 2 possible alleles.
    - Macrosatellite markers: length from several bps to several hundreds of bps; many possible alleles
    - The *heterozygosity* of a marker is defined as the probability that two alleles chosen at random are different. If π is the (relative) frequency of the *i*-th allele, then heterozygosity can be expressed as:

$$\text{Heterozygosity} = 1 - \sum_i \pi_i^2$$

- *Genotype*: At each locus there is an allele in each chromosome of the homologous chromosome pair. The two alleles together (no specific order) are called genotype



For one individual, the genotype at one marker contains two alleles.

If the two alleles are the same, the genotype is called **homozygote.**

Otherwise, it is called **heterozygote**.

- The *phenotype* is the characteristic (e.g. hair color) that results from having a specific genotype ;
- The *trait* is a coded (e.g. for actual statistical analysis) of the phenotype.

Example: Blood type (ABO locus, three allele A, B and O)

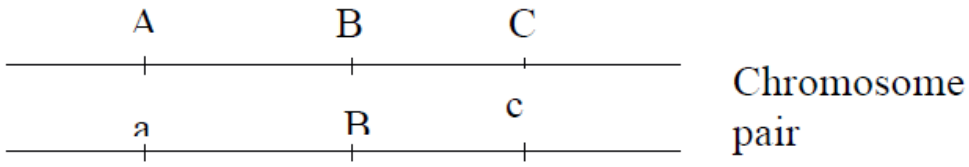
Phenotype	Genotype
A	A/A, A/O
B	B/B, B/O
AB	A/B
O	O/O

Here, A and B both mask the presence of the O allele.
— A and B are said to be **dominant** to O;
— O is **recessive** to A and B.
— A and B are **co-dominant.**

- There are two main different measures for heredity:
  - *Broad heritability*:
    proportion of total phenotypic variance accounted for by all genetic components (coefficient of genetic determination)
  - *Narrow heritability*:
    proportion of phenotypic variance accounted for by the additive genetic component

Recombination introduces extra variation

- A collection of linked loci (loci that tend to be inherited together) is called a *haplotype*



Genotype A/a B/B C/c

The two haplotypes are ABC and aBc.

- In practice (for codominant alleles), we can only observe multilocus genotype {A/a B/b C/c}. So the possible haplotype pairs are {ABC, aBc} and {ABc, aBC}

Typical data set as follow:

Ind. ID	Marker 1	Marker 2	Marker 3
1	0/0	0/1	4/5
2	1/0	0/2	8/7
3	1/1	2/1	9/6
4	0/1	3/2	8/8
5	0/0	0/3	6/5
6	1/0	2/2	7/7
	(two alleles)	(4 alleles)	(6 alleles)

For individual 6, two haplotypes are
{1,2,7; 0,2,7}.

For individual 1, we do not know the two haplotypes.
Tow possibilities are
{0,0,4; 0,1,5} or {0,1,4; 0,0,5}

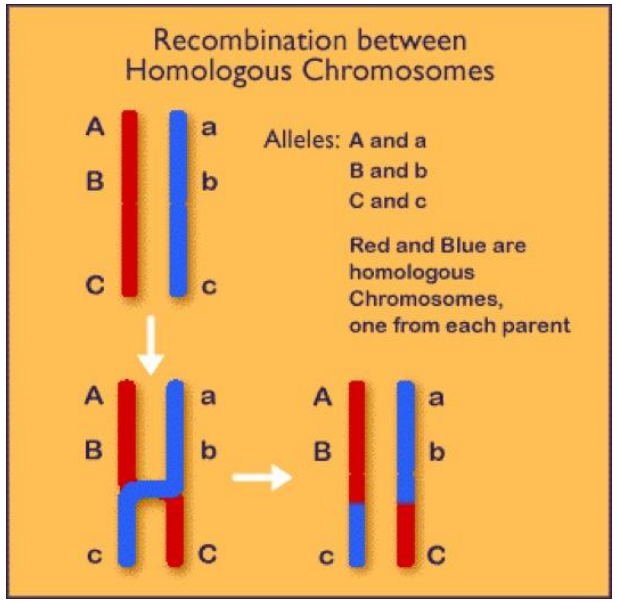
Recombination

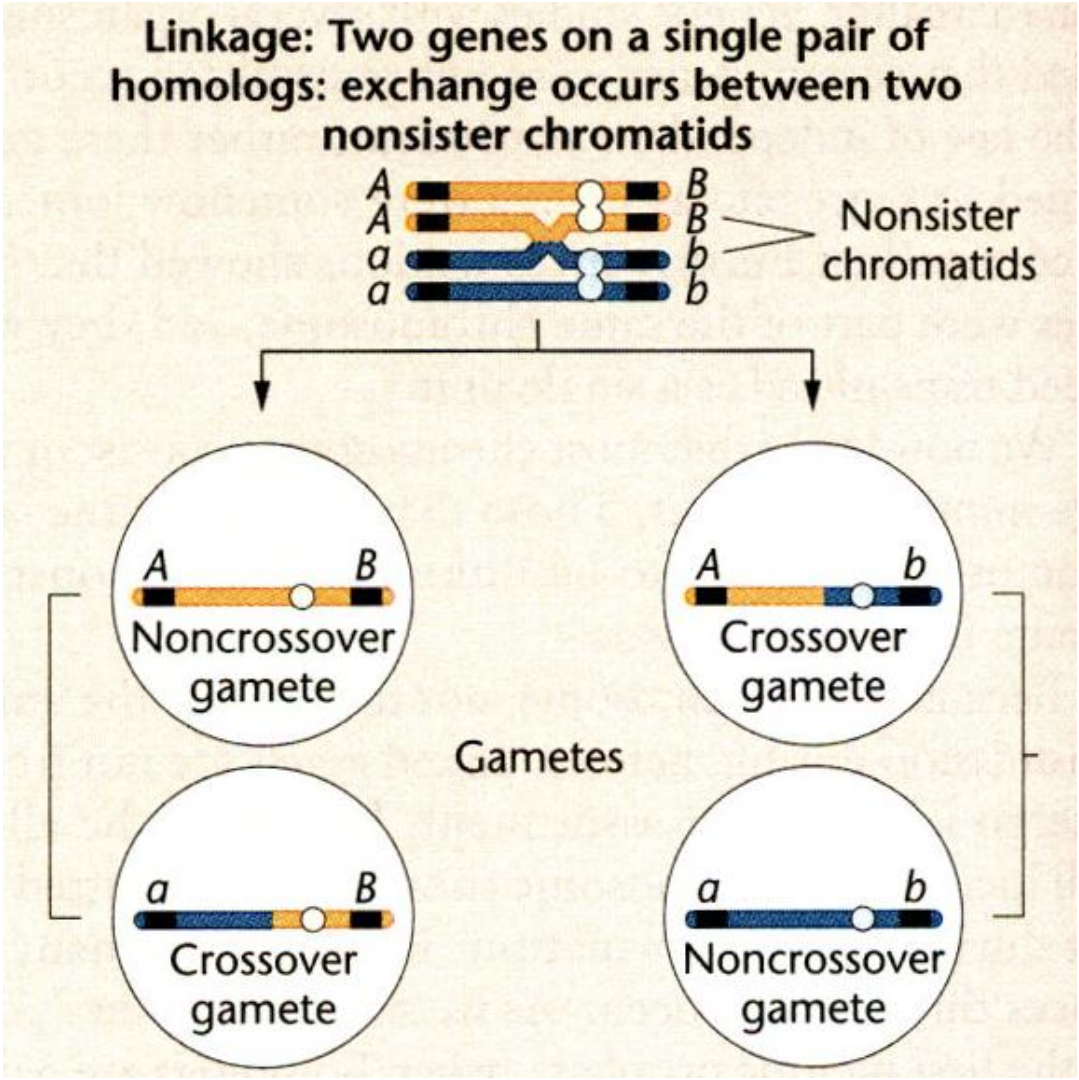
- Immediately before the cell division that leads to gametes, parts of the homologous chromosomes may be exchanged

  An individual with haplotypes A-B and a-b may produce gametes

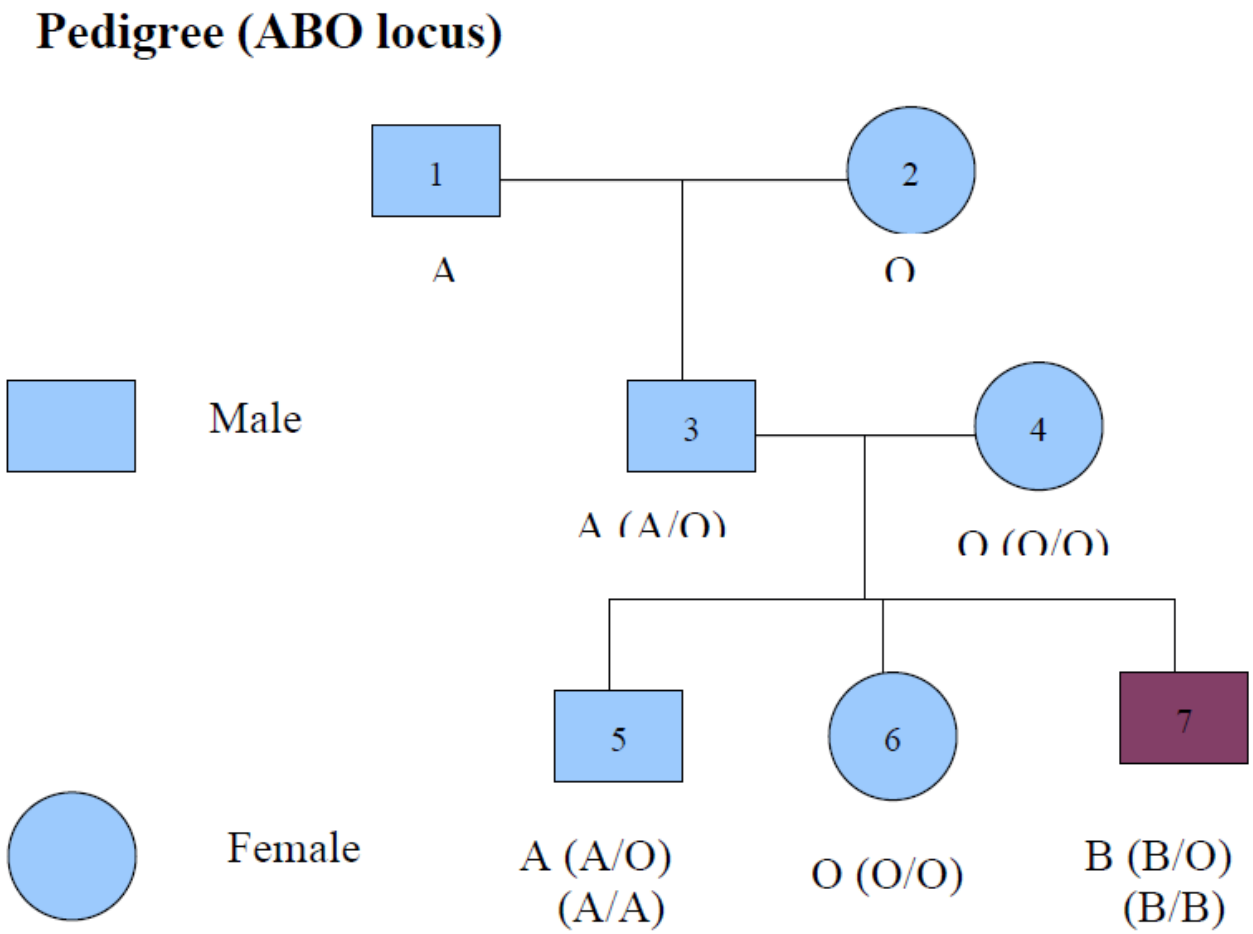
  A-B and a-b or A-b and a-B.

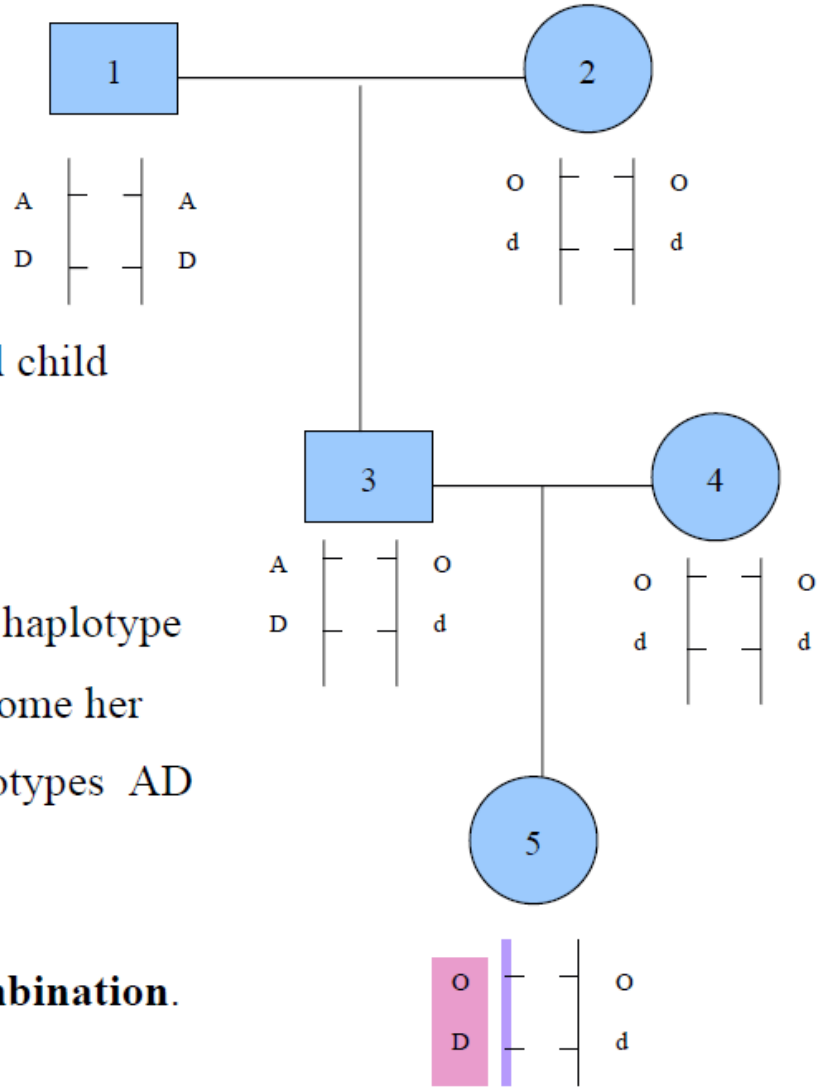
- The last two examples are indicative for a process called cross-over (i.e. the process by which two chromosomes pair up and exchange sections of their DNA). Recombination refers to the result of such a process, namely genetic recombination.




Linkage: Two genes on a single pair of homologs: exchange occurs between two nonsister chromatids
A
A
a
a
B
B
b
b
Nonsister chromatids
A
B
Noncrossover gamete
A
b
Crossover gamete
Gametes
a
B
Crossover gamete
a
b
Noncrossover gamete

Example: Is child 7 the biological child of 3 and 4?



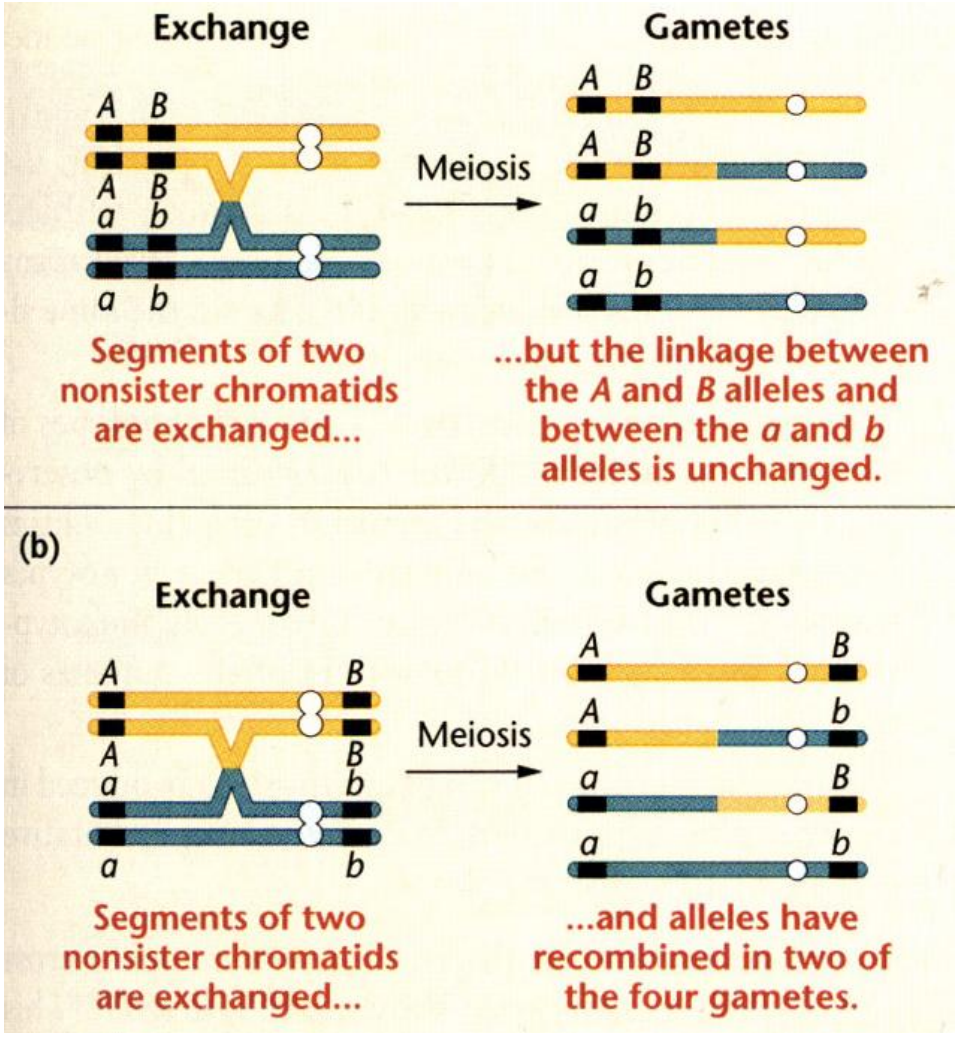


- Is 5 the biological child of 3 and 4? Yes !

- Why 5's paternal haplotype OD is different frome her father's two haplotypes AD and Od.

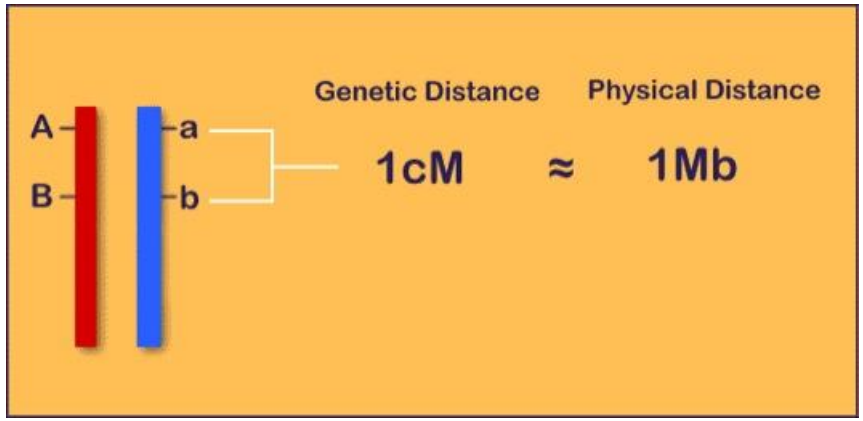
This is called **recombination**.

Recombination can change allele arrangements



Recombination is related to genetic distance

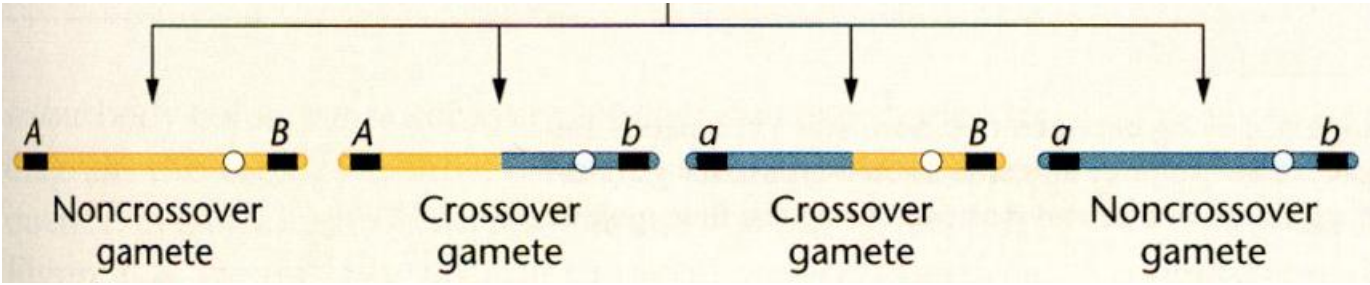
- The greater the physical distance between two loci, the more likely it is that there will be recombination.
- This forms the basis of mapping strategies such as “linkage” and “association”.
- So recombination is related to “distance” D. In a way, it forms a bridge between “physical distance” and “genetic distance”



(Roche Genetics)

Genetic distance (continued)

- Genetic distance uses recombination rate: 1cM ≈ 1% recombination rate
- The probability of recombination between two markers during meiosis is termed the *recombination fraction (or recombination rate)* [the proportion of gametes that are recombinant between the two loci], and is usually denoted by ϑ.



- What are the extreme values of the recombination fraction?

Genetic distance (continued)

- Unlinked genes (on different chromosomes) cosegregate 50% of the time. Defines maximum
- Recombination frequency measures recombinant gametes NOT number of crossovers
- Some recombination events are NOT observed:
    - between sister chromatids
    - double recombination (double-crossover)

Example:

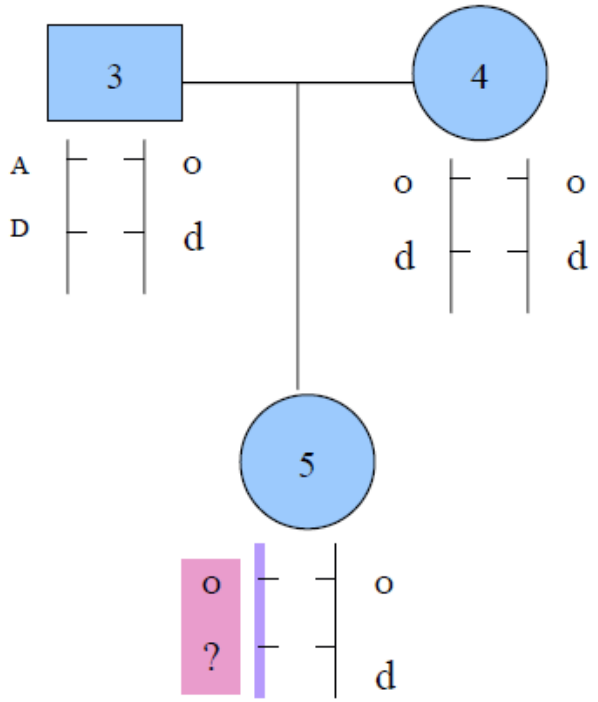
If θ=0, then ?=d,

If θ<1/2, then the

?=d with probability >1/2.

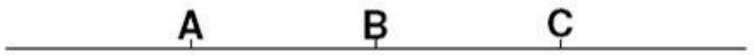
If θ=1/2, ?=d and ?=D with

Equal probability.



Genetic distance (continued)

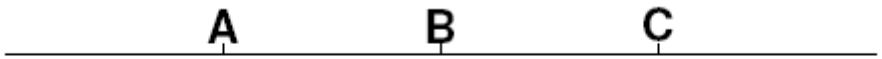
- Since a recombination event is only observed if there are an odd number of crossovers between the two loci, recombination fractions are not additive.
- In general, a genetic map function M(D) = ϑ provides a mapping from the additive genetic distance D to the non-additive recombination fraction ϑ between a given pair of loci.
- Several models exist for recombination rates, but the "constant recombination rate" model is the simplest:
  - A simplified model is that loci can be arranged along a line in such a way that, with each meiosis, recombinations occur at a constant rate.



Genetic distance (continued)

- In the simplest setting, the relationship between the recombination frequency and the genetic distance D_{AB} between loci A and B is then given by Haldane's map function as follows:

$$D_{AB} = -\frac{1}{2}\log_e(1 - 2\theta_{AB})$$



$$D_{AC} = D_{AB} + D_{BC}$$

- In practice, real-life is more complicated, due to settings for which the model of constant rate or independence of recombinations does not fit
  - Under the *Kosambi map function*, complete *interference* is assumed for small map distances and a decreasing amount of interference accompanies increasing distances. *Hot spots* cause uneven relationship between physical and genetic distances

Genetic distance (continued)

- An extra real-life complication is that recombination appears to be more frequent in females than in males:
  - Total female map length: 44 Morgans
  - Total male map length: 27 Morgans
  - Total sex-averaged map length: 33 Morgans
- On average, 1 cM corresponds to about 10^6 bases (i.e. 1000kb or 1Mb).
  - The total length of the human genome is “on average” 33 Morgans (≈ 3×10^9 bases)

2.b Variability is the key to “information”

- Variation at genetic loci (see before); variation at genetic markers
- Trait variation may have genetic and/or non-genetic explanations
  - In many cases, the same phenotype can result from a variety of different genotypes (sometimes termed *phenocopies*)
  - Equally, the same gene may have several different phenotypic manifestations. This phenomenon is called *pleiotropy*.
  - “Association studies” between markers and trait may reveal this
- Variation in chromosome numbers between species
  - All animals have a characteristic number of chromosomes in their body cells called the diploid (or 2n) number.
  - The gametes contain the haploid number (n) of chromosomes.

Diploid numbers of commonly studied organisms

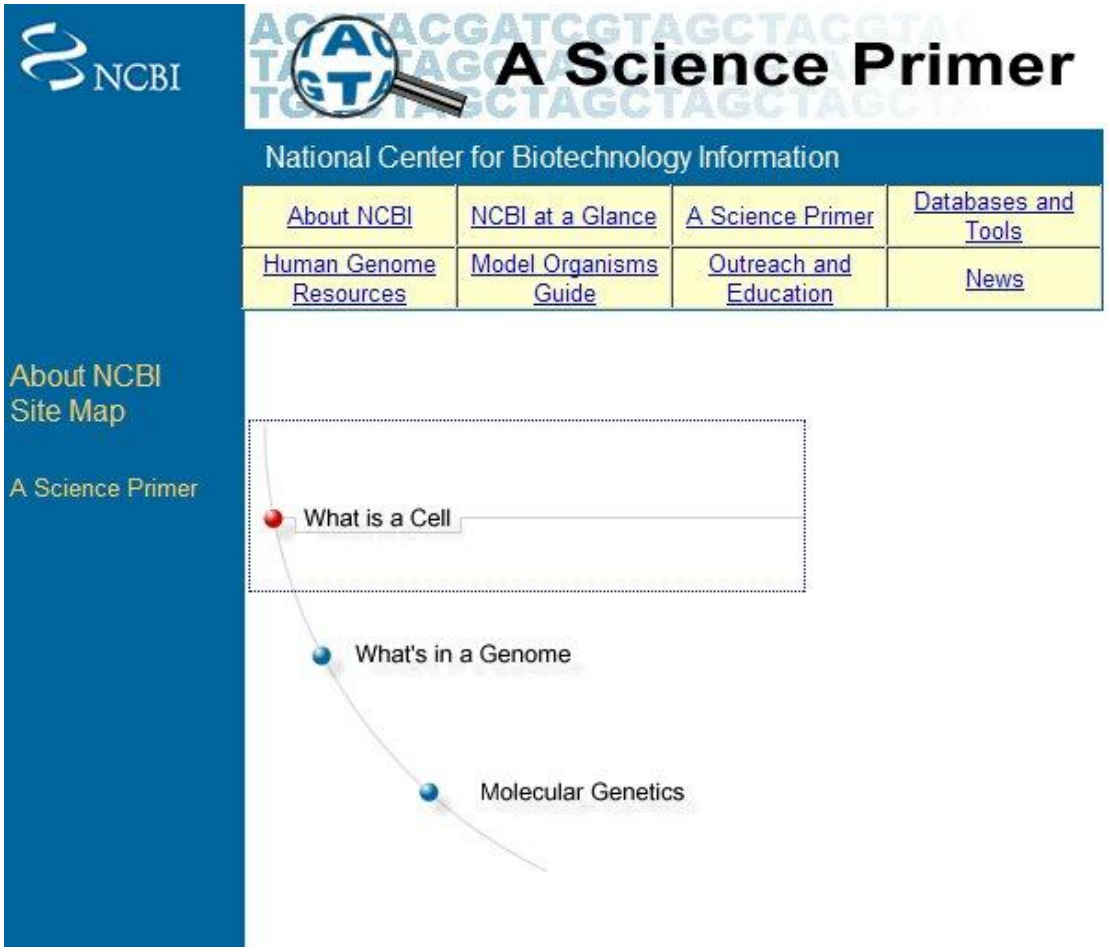
References:

- Ziegler A and König I. *A Statistical approach to genetic epidemiology*, 2006, Wiley. (Chapter 1, Sections 2.3.1; 3.1, 3.2.2; 5.1, 5.2.1-5.2.3)
- Burton P, Tobin M and Hopper J. Key concepts in genetic epidemiology. *The Lancet*, 2005
- Clayton D. Introduction to genetics (course slides Bristol 2003)
- URLs:
  - http://www.rothamsted.ac.uk/notebook/courses/guide/
  - http://www.cellbio.com/courses.html
  - http://www.genome.gov/Education/
  - http://www.roche.com/research_and_development/r_d_overview/education.htm
  - http://nitro.biosci.arizona.edu/courses/EEB320-2005/
  - http://atlasgeneticsoncology.org/GeneticFr.html
  - http://www.worthpublishers.com/lehninger3d/index2.html
  - http://www.dorak.info/evolution/glossary.html

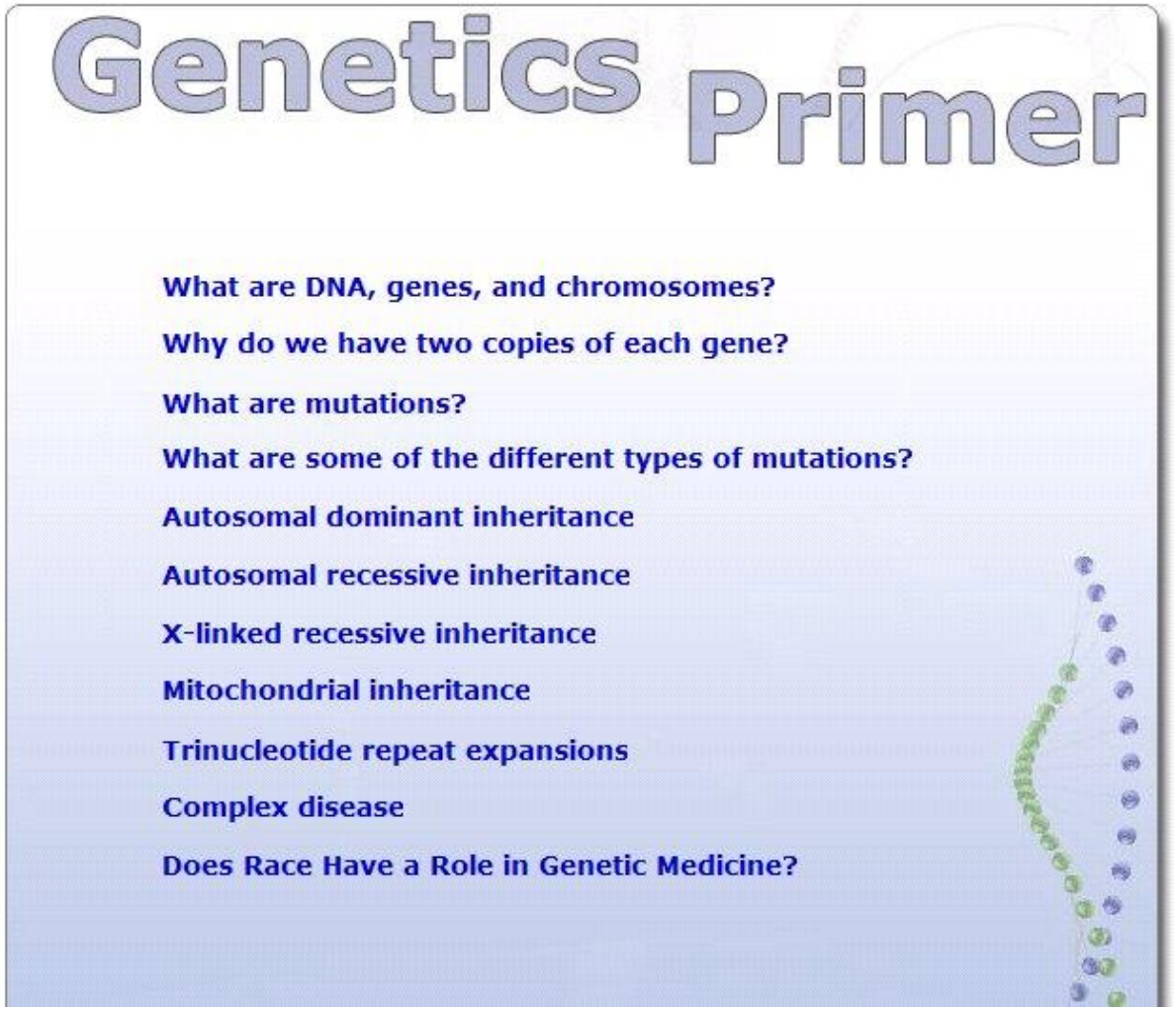
For a primer on the Human Genome Project

  - http://www.sciencemag.org/content/vol291/issue5507/

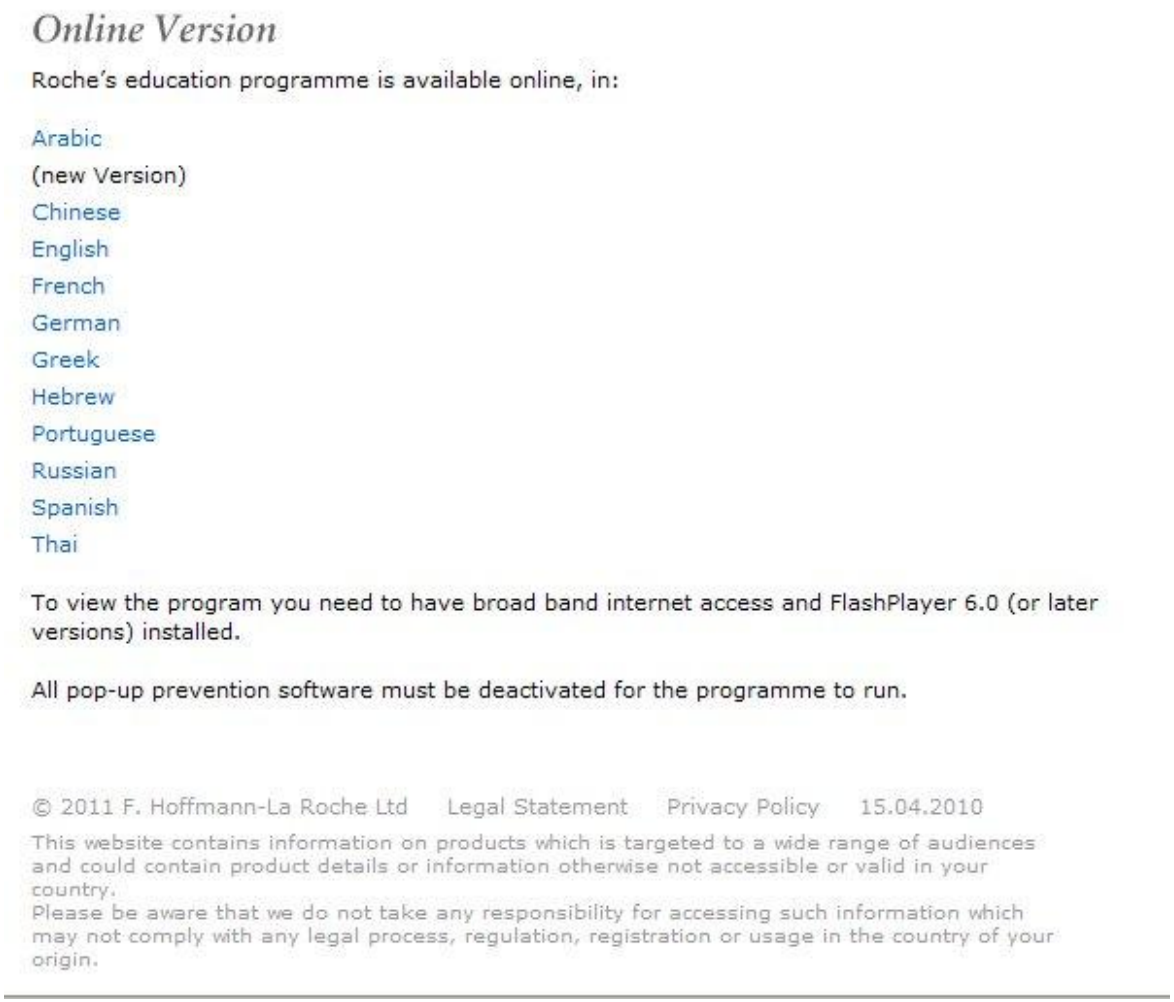
For additional info on concepts:



(http://www.ncbi.nlm.nih.gov/About/primer/genetics.html)



(http://www.nchpeg.org/pa/index.php?option=com_content&view=article&id=56&Itemid=56)



Online Version

Roche's education programme is available online, in:

Arabic
(new Version)
Chinese
English
French
German
Greek
Hebrew
Portuguese
Russian
Spanish
Thai

To view the program you need to have broad band internet access and FlashPlayer 6.0 (or later versions) installed.

All pop-up prevention software must be deactivated for the programme to run.

(http://www.roche.com/education)