INTRODUCTION TO GENETIC EPIDEMIOLOGY (EPID0754)

Prof. Dr. Dr. K. Van Steen

CHAPTER 2: INTRODUCTION TO GENETICS

1 Basics of molecular genetics

Where is the genetic information located?

The structure of cells, chromosomes, DNA and RNA

2 Human genetics

The *human* genome

How is genetic information transmitted from generation to generation?

Variation is key to information: mutations and polymorphisms

1 Basics of molecular genetics

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 decade we have seen the release of the first drafts of the entire human genome and the genomes of model organisms.

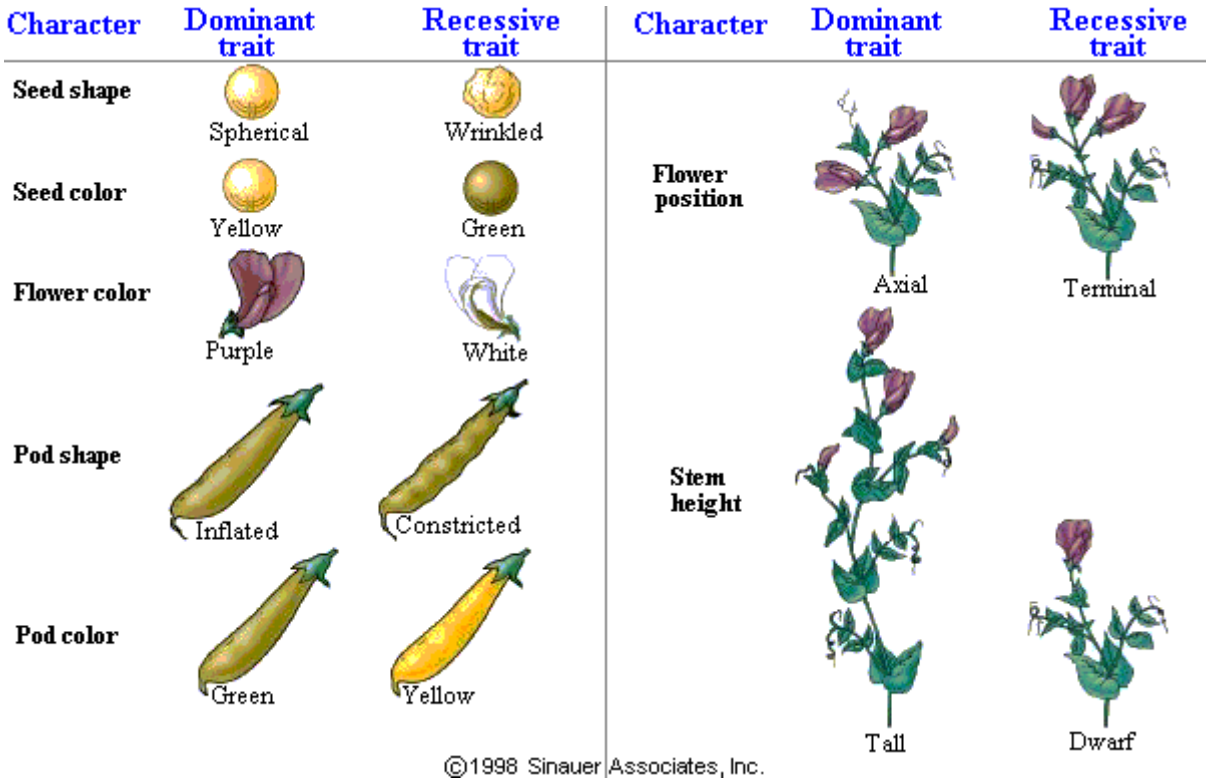
- The most notable experiments have unequivocal interpretation:
  - Unequivocal interpretation is rare in human genetics
  - Generally cannot design the perfect experiment: have to work with data we have at our disposal
  - Interpretation is of the greatest importance
- How do our data and results inform 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

Where is the genetic information located?

Mendel

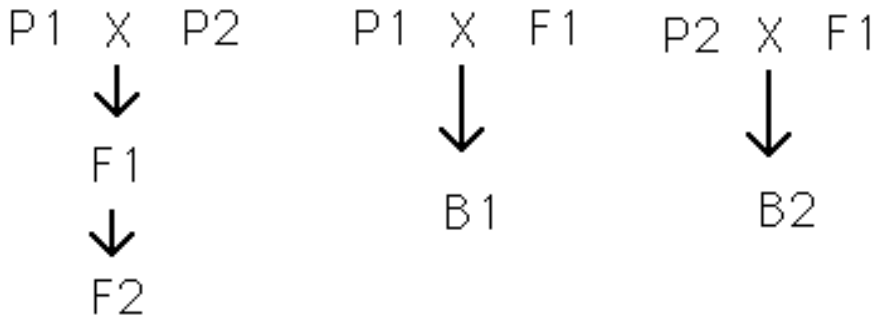
- Many traits in plants and animals are heritable; genetics is the study of these heritable factors
- Initially it was believed that the mechanism of inheritance was a masking of parental characteristics
- Mendel developed the theory that the mechanism involves random transmission of discrete **"units" 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

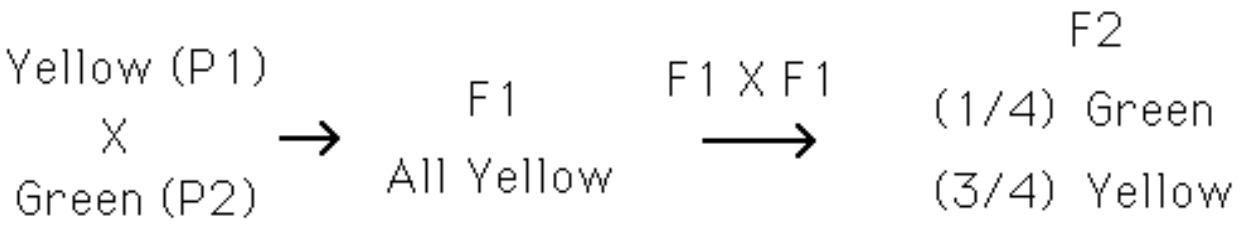


Some notations for line crosses

- 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

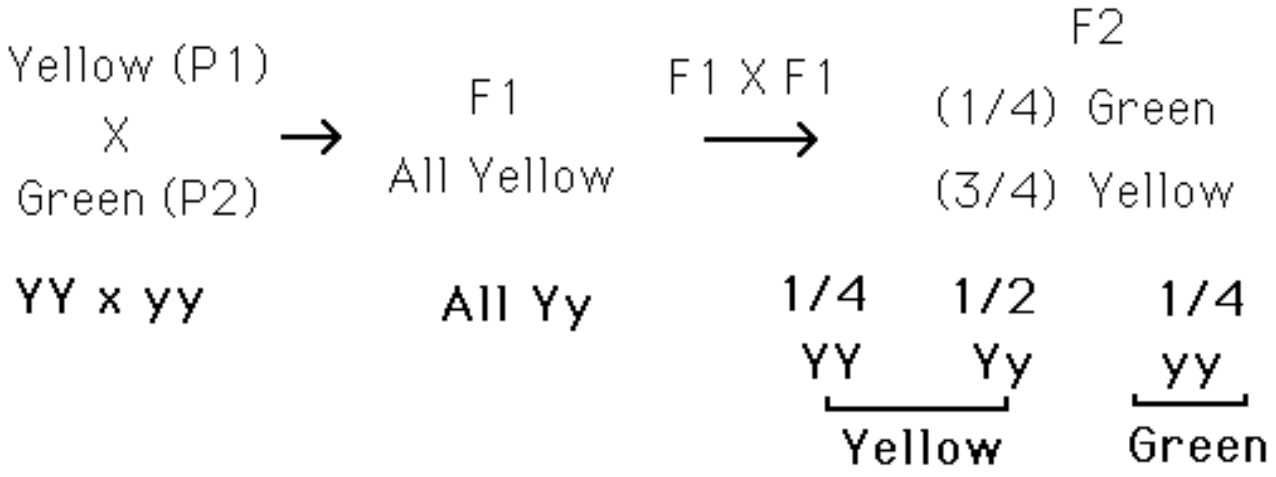
F1 Yellow
X
Green (P2)
→ 1/2 Yellow
1/2 Green

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.

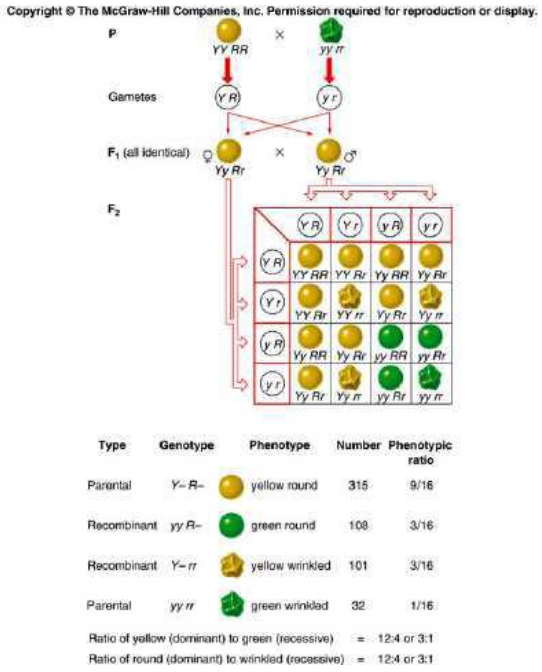


Mendel's conclusions (continued)

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

  This says that for two characteristics the genes are inherited independently

  ???????

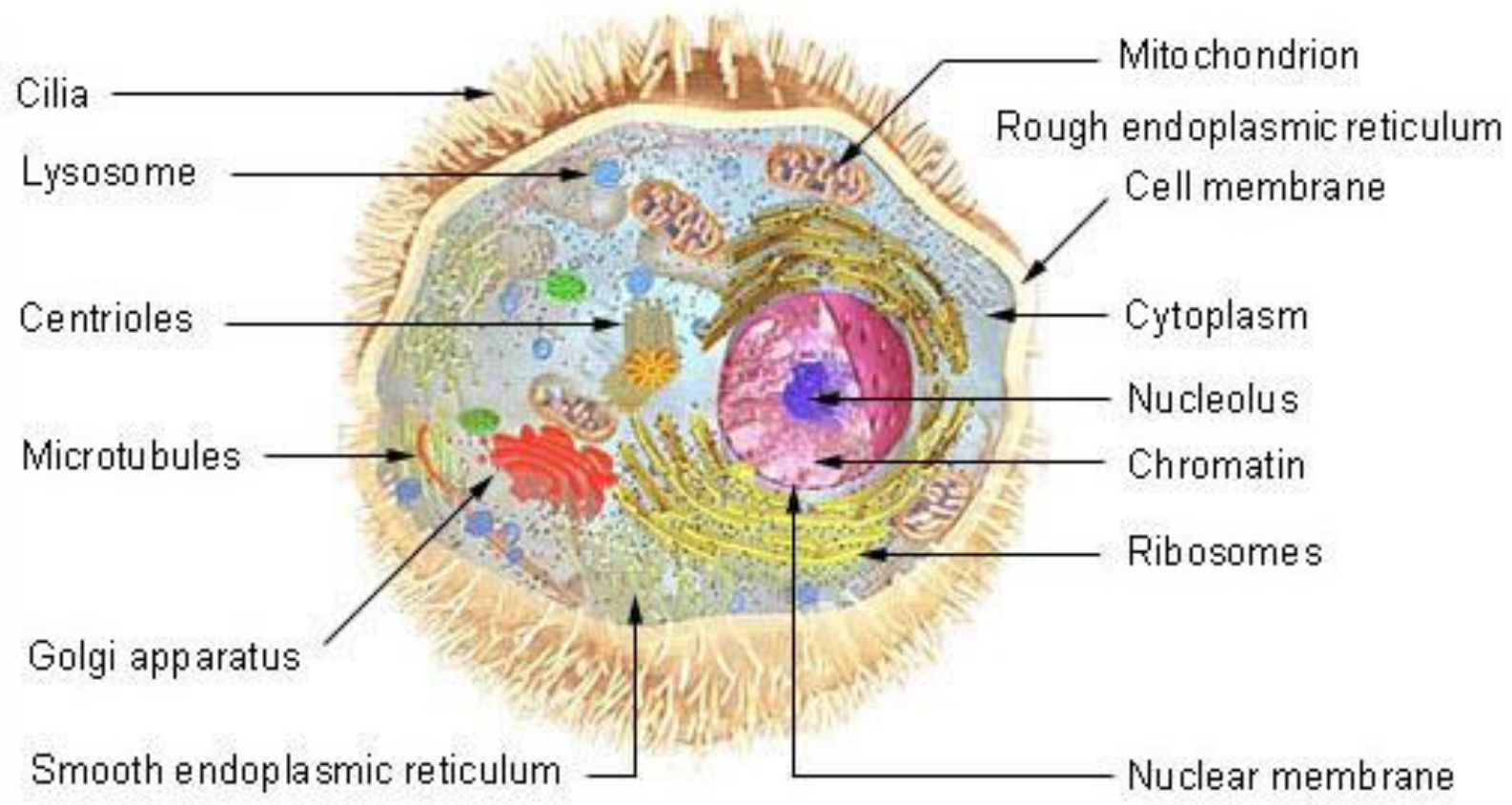


Type	Genotype	Phenotype	Number	Phenotypic ratio
Parental	Y– R–	yellow round	315	9/16
Recombinant	yy R–	green round	108	3/16
Recombinant	Y– rr	yellow wrinkled	101	3/16
Parental	yy rr	green wrinkled	32	1/16

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

Question: How do we have to interpret the by Mendel identified "units of information"? Look at cell biology

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

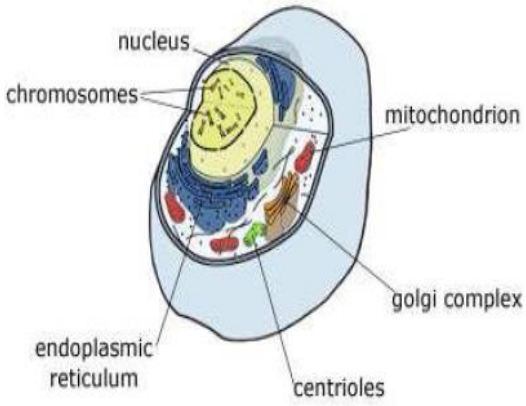
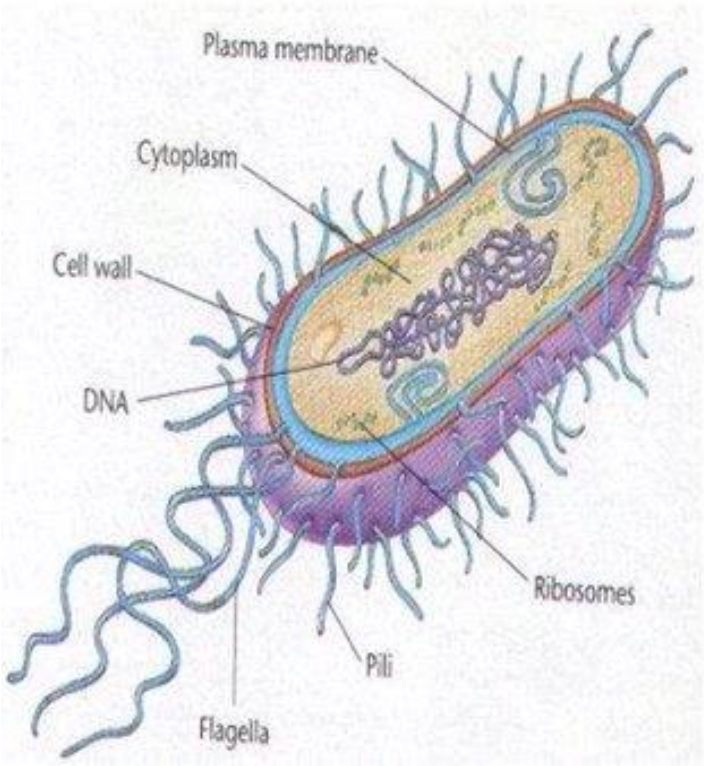


Image adapted from: National Human Genome Research Institute.

A typical eukaryotic cell.

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



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.

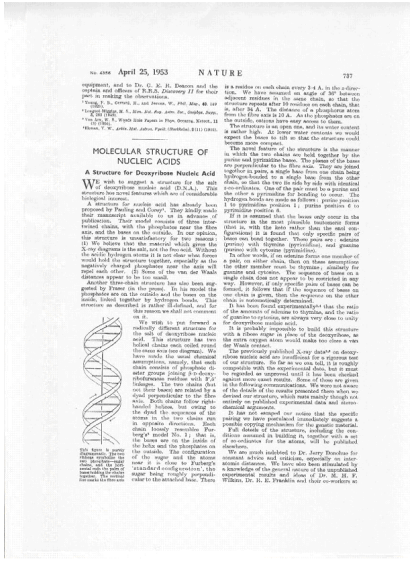
British biophysicist Francis Crick and American geneticist **James Watson** undertook a joint inquiry into the structure of DNA in 1951.

(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

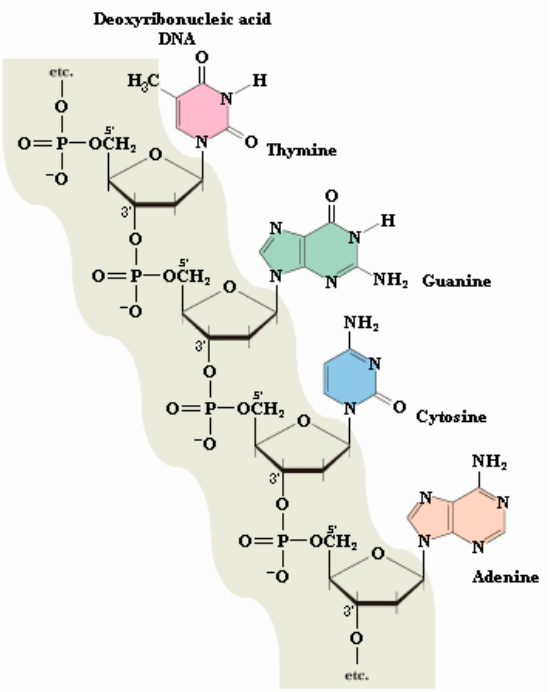
A Structure for Deoxyribose Nucleic Acid

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)

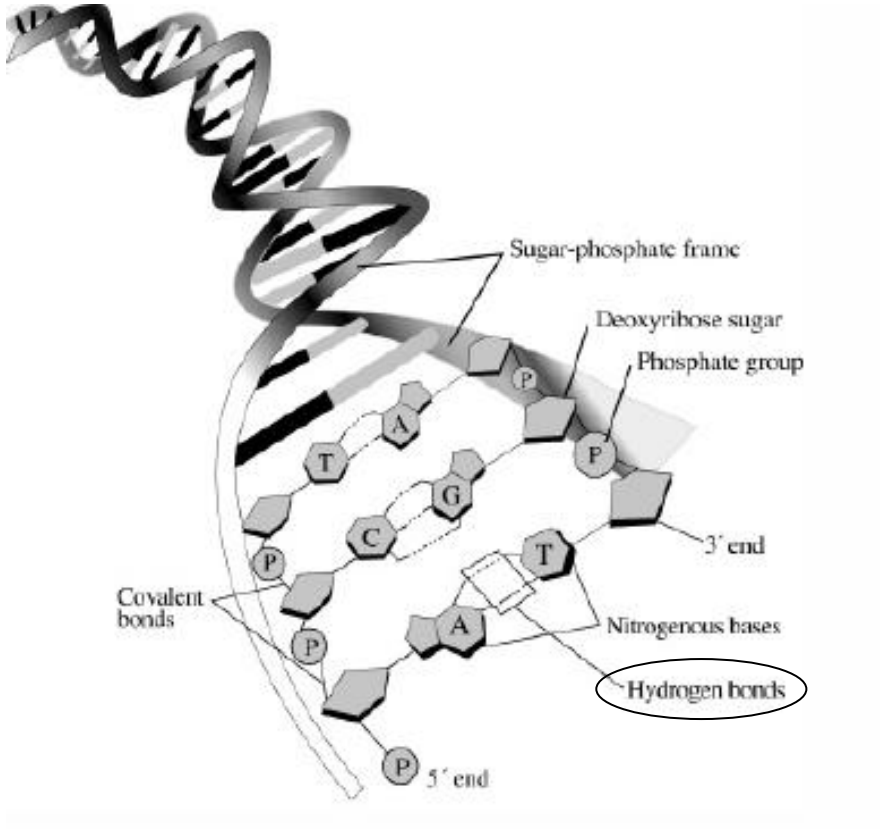
- In fact, there are 4 structures that are relevant when talking about DNA…

Primary structure of DNA

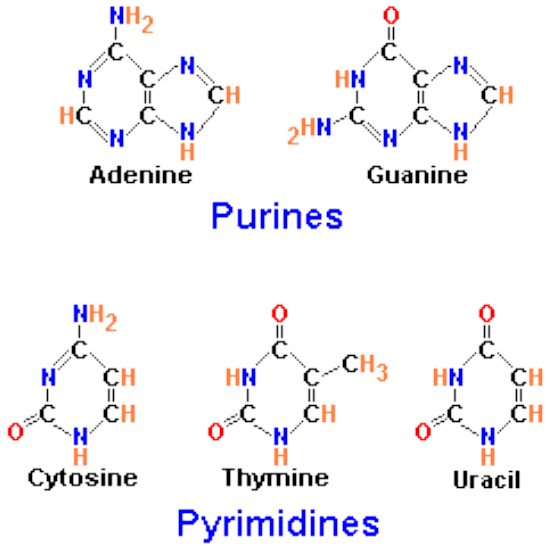
The 3 dimensional structure of DNA can be described in terms of primary, secondary, tertiary, and quaternary structure.

- 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



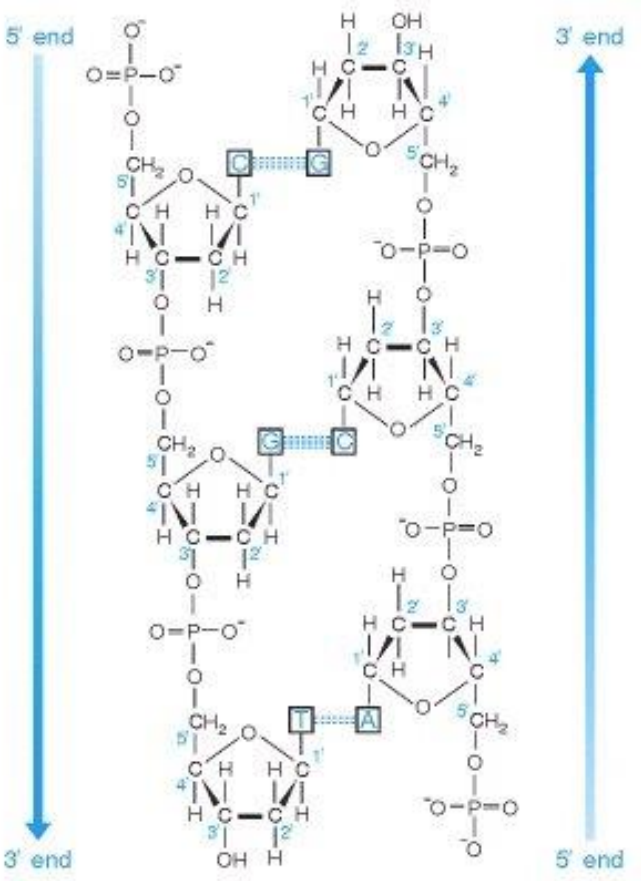
Nitrogenous bases



(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

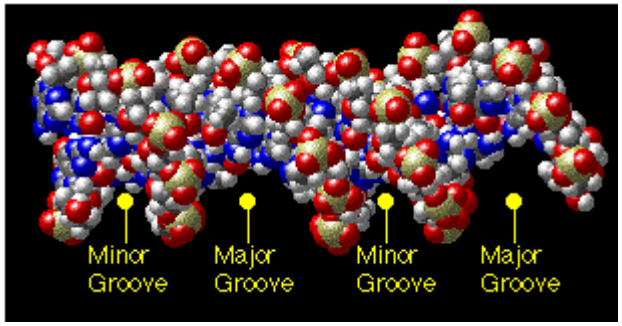
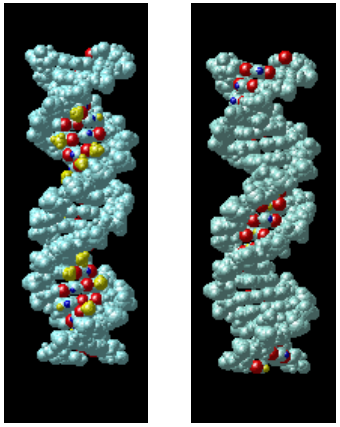


Figure 2

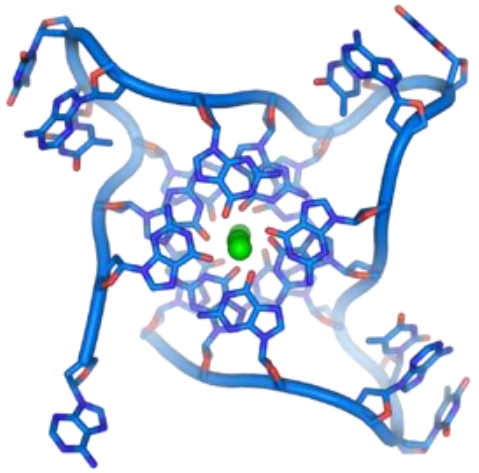


Tertiary structure of DNA

- 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).

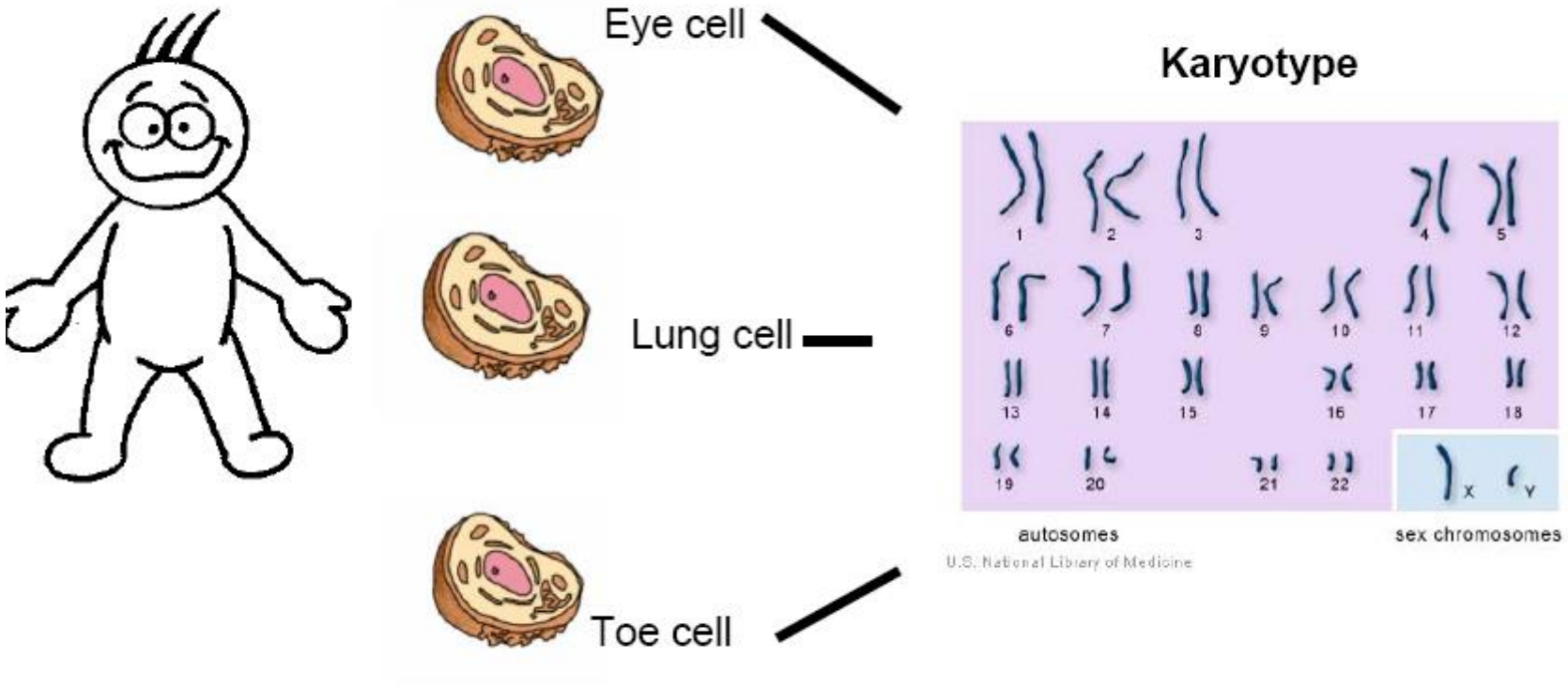
Quaternary structure of DNA

- 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)

Every cell in the body has the same DNA



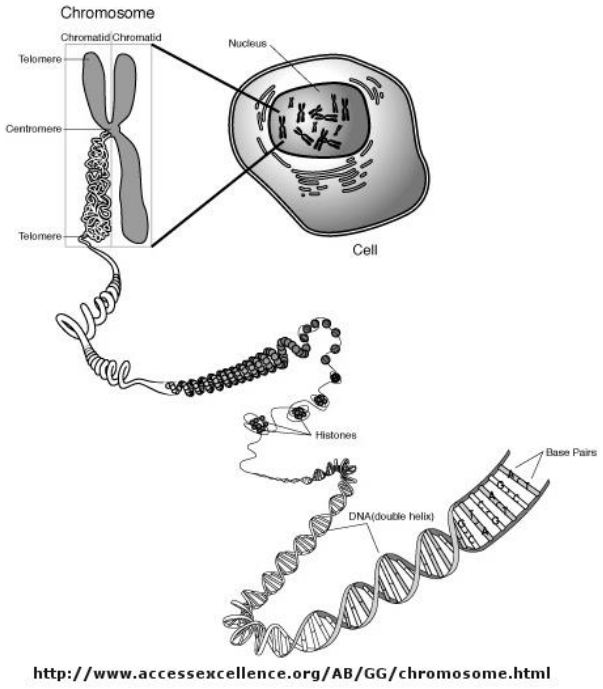
- One base pair is 0.00000000034 meters
- DNA sequence in any two people is 99.9% identical – only 0.1% is unique!

DNA makes RNA, RNA makes proteins, proteins make us

- 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 – see later).
- 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]

Chromosomes

- For convenience, 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** that support its structure.

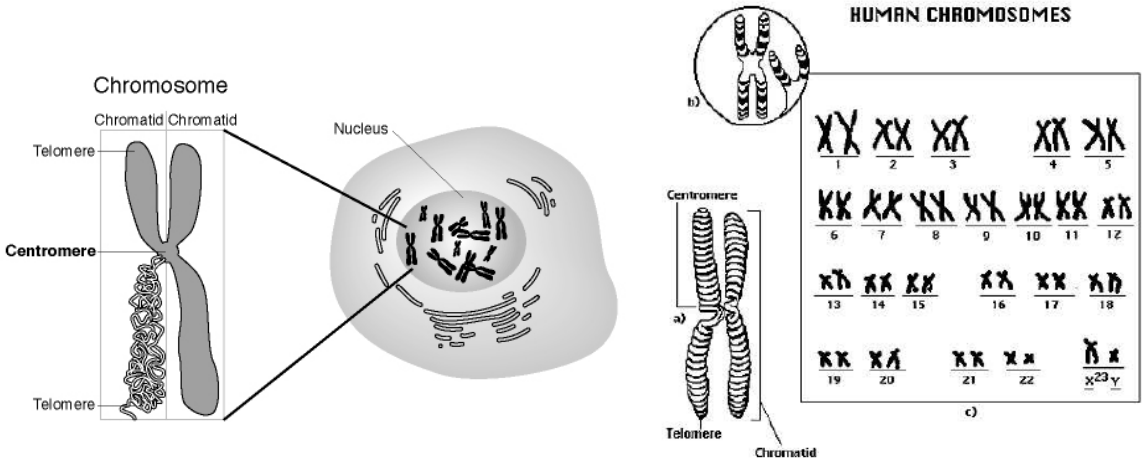


- *Histones* are proteins rich in lysine and arginine residues and thus positively-charged.
- For this reason they bind tightly to the negatively-charged phosphates in DNA.

Chromosomes

- 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.
- **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.

- Chromosomes that are of the same pair and carry the same set of genes and are called **homologous**. (e.g. both chromosome 21)
- 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.



2 Human genetics

The human genome

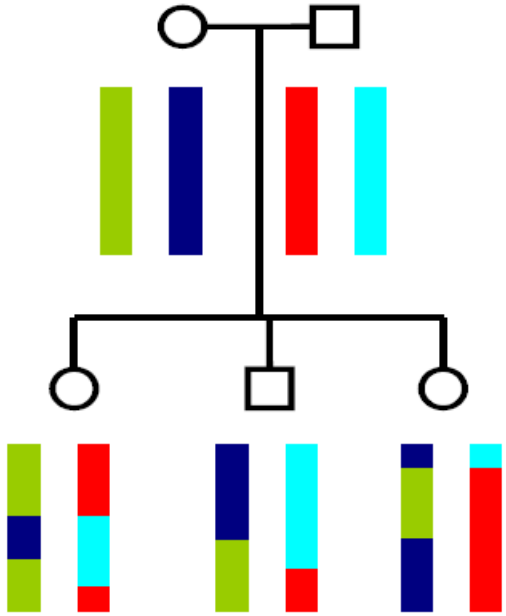
- The entire DNA characteristics of a species is called its **genome**.
- The human genome has about 3 billion base pairs per haploid.
- Approximately 2% of the human genome is coding and 98% of the human genome is non-coding.
- A gene is a sequence of DNA that is transcribed into mRNA (messenger RNA), which, in turn, is translated into protein (see later for more details).
- For RNA, uracil (U) is substituted for thymine in DNA.
- There are about 20,000 genes for humans
- Genes vary enormously in length from less than a thousand base (Kb) pairs to over a million base pairs (Mb)

The human genome

- 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
- 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**

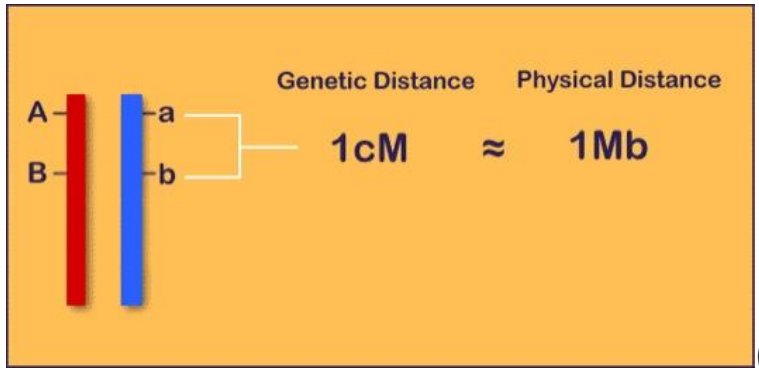
Recombination

- A chromosome inherited by an offspring from a parent is actually a mosaic of the parent's two chromosomes.
- **Genetic Recombination**: 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. This occurs when two loci are separated by an odd number of crossovers
- **Genetic maps** give the order and distances (recombination fraction) between genes and genetic markers
- **Physical maps** refer to sets of ordered markers and the physical distance (base pairs) between them

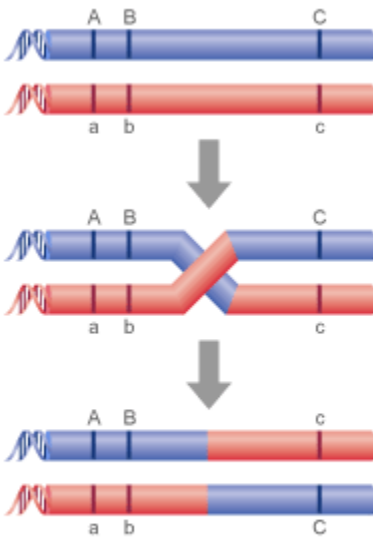


(Roche Genetics)

Linkage and linkage analysis

- Non-independent transmission between two loci implies linkage. Linkage is related to physical proximity.
- **Linkage analysis** aims at finding out the rough location of the gene relative to another DNA sequence called a genetic marker, which has its position already known. It is has long been the traditional way to search for disease genes
- In this course, the emphasis lies on genetic association analysis, another way to find genetic disease (complex trait) predisposing loci

Linkage analysis

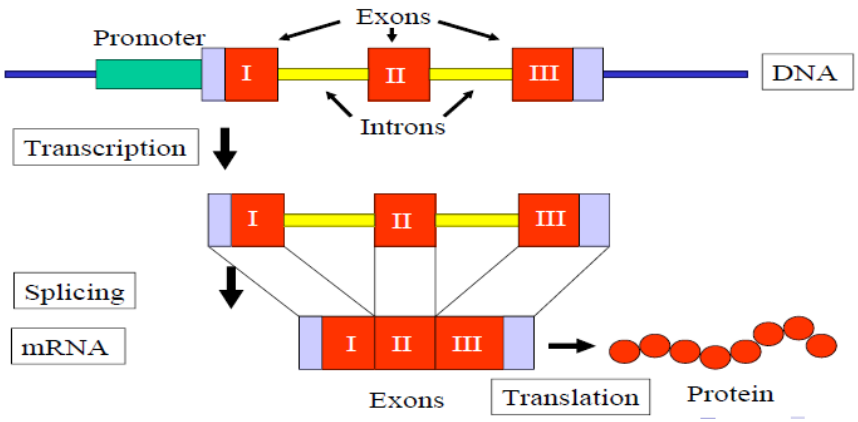


The top diagram shows paternal (blue) and maternal (red) chromosomes

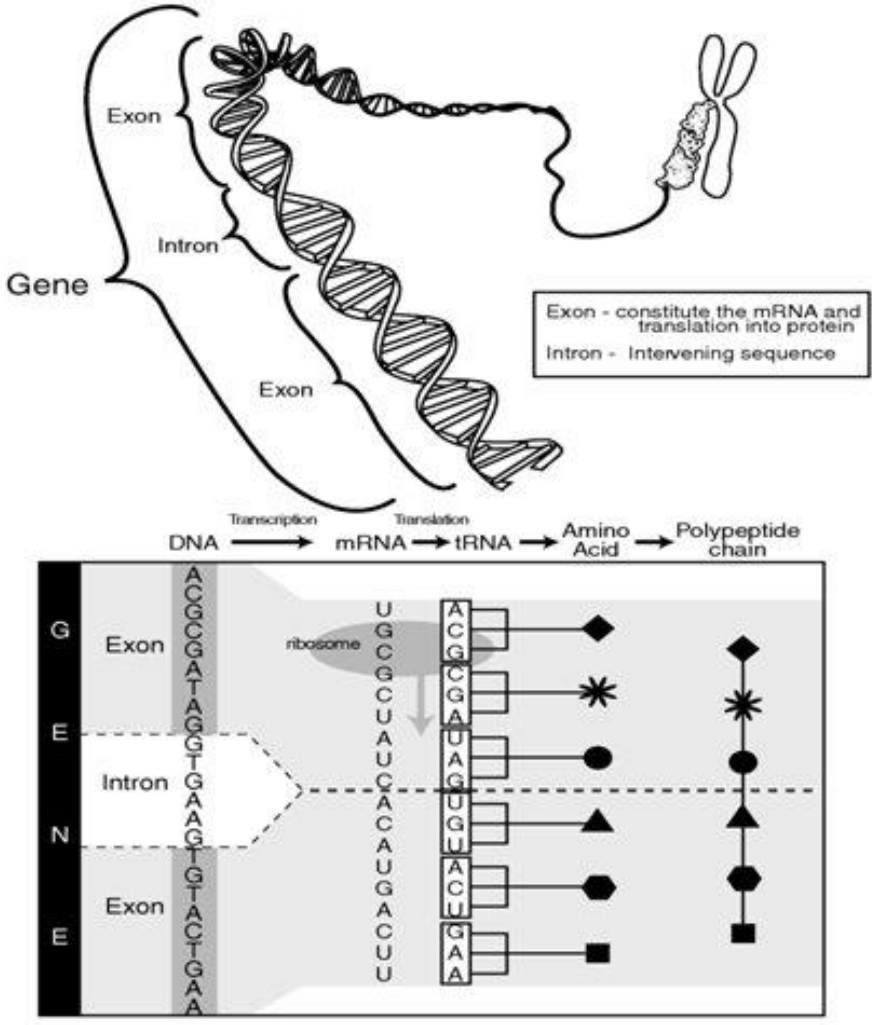
aligned in a germ cell, a cell that gives rise to eggs or sperm. Three DNA sequences are shown, labelled A, B and C. The capital letters represent the paternal alleles and the lower case letters represent the maternal alleles. The middle panel shows the physical process of recombination, which involves crossing over of DNA strands between the paired chromosomes. The bottom panel shows what happens when the crossover is resolved. The maternal and paternal alleles are mixed (recombined) and these mixed chromosomes are passed to the sperms or eggs. If A is the disease gene and B and C are genetic markers, recombination is likely to occur much more frequently between A and C than it is between A and B. This allows the disease gene to be mapped relative to the markers B and C.
(http://genome.wellcome.ac.uk/doc_WTD020778.html)

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

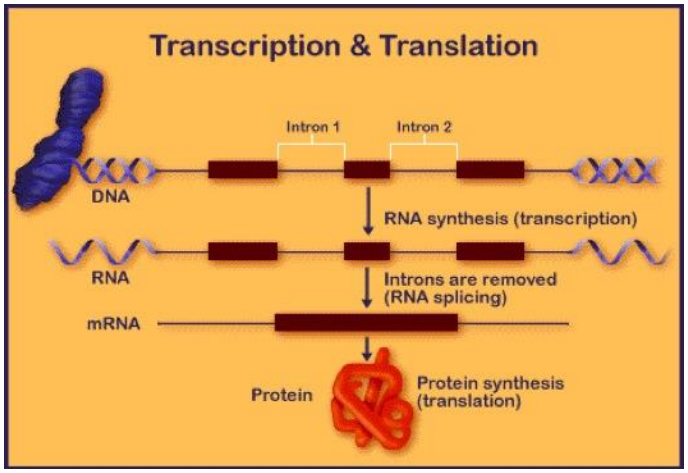


DNA makes RNA, RNA makes proteins, proteins make us



Central dogma of molecular biology

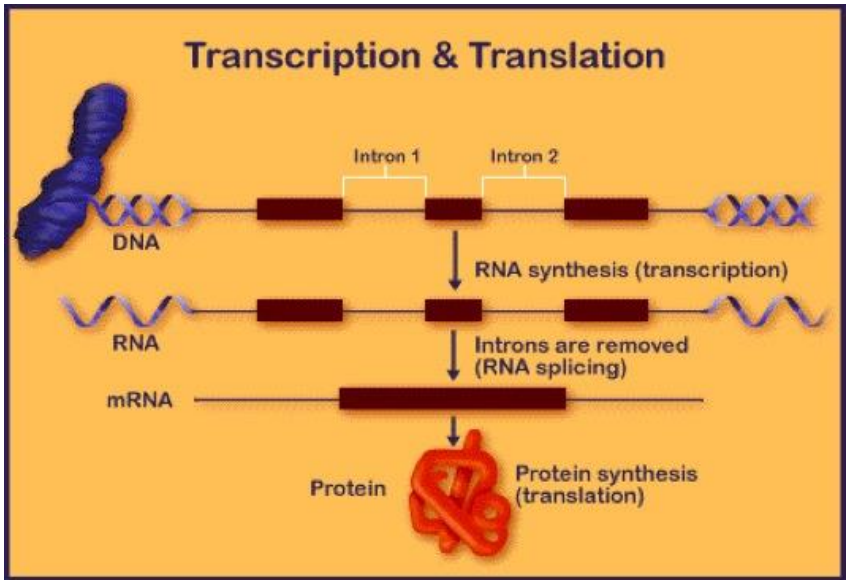
- Stage 1: DNA replicates its information in a process that involves many enzymes. This stage is called the ***replication*** stage.
- Stage 2: The DNA codes for the production of messenger RNA (mRNA) during ***transcription*** of the sense strand (coding or non-template strand)



(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.

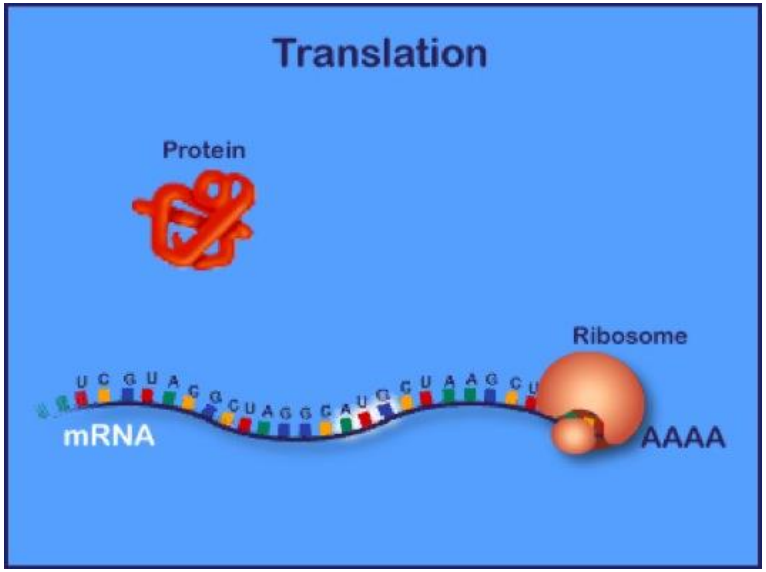
- Stage 3: In eukaryotic cells, the mRNA is ***processed*** (essentially by splicing) and migrates from the nucleus to the cytoplasm



(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.

Translation is facilitated by two key molecules: transfer RNA and ribosomes

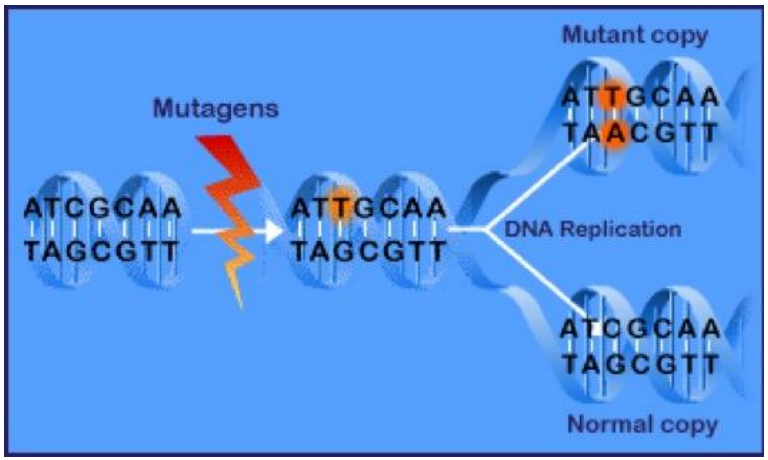


- *Transfer RNA* (tRNA) molecules transport amino acids to the growing protein chain. Each tRNA carries an amino acid at one end and a three-base pair region, called the anti-codon, at the other end. The anti-codon binds with the codon on the protein chain via base pair matching.

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

DNA repair mechanisms

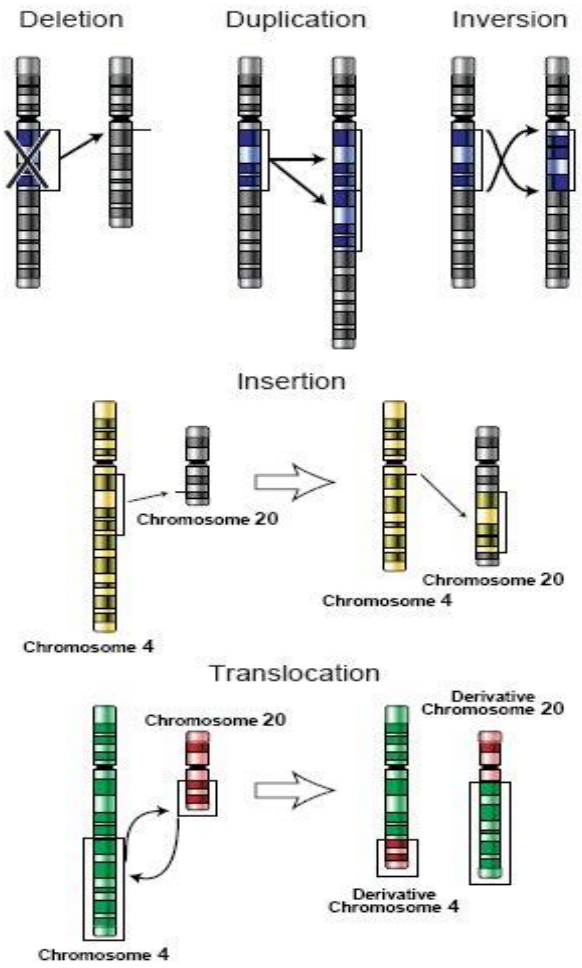
- In biology, a mutagen (Latin, literally origin of change) is a physical or chemical agent that changes the genetic material (usually DNA) of an organism and thus increases the frequency of mutations above the natural background level.
- As many mutations cause cancer, mutagens are typically also carcinogens.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



(Roche genetics)

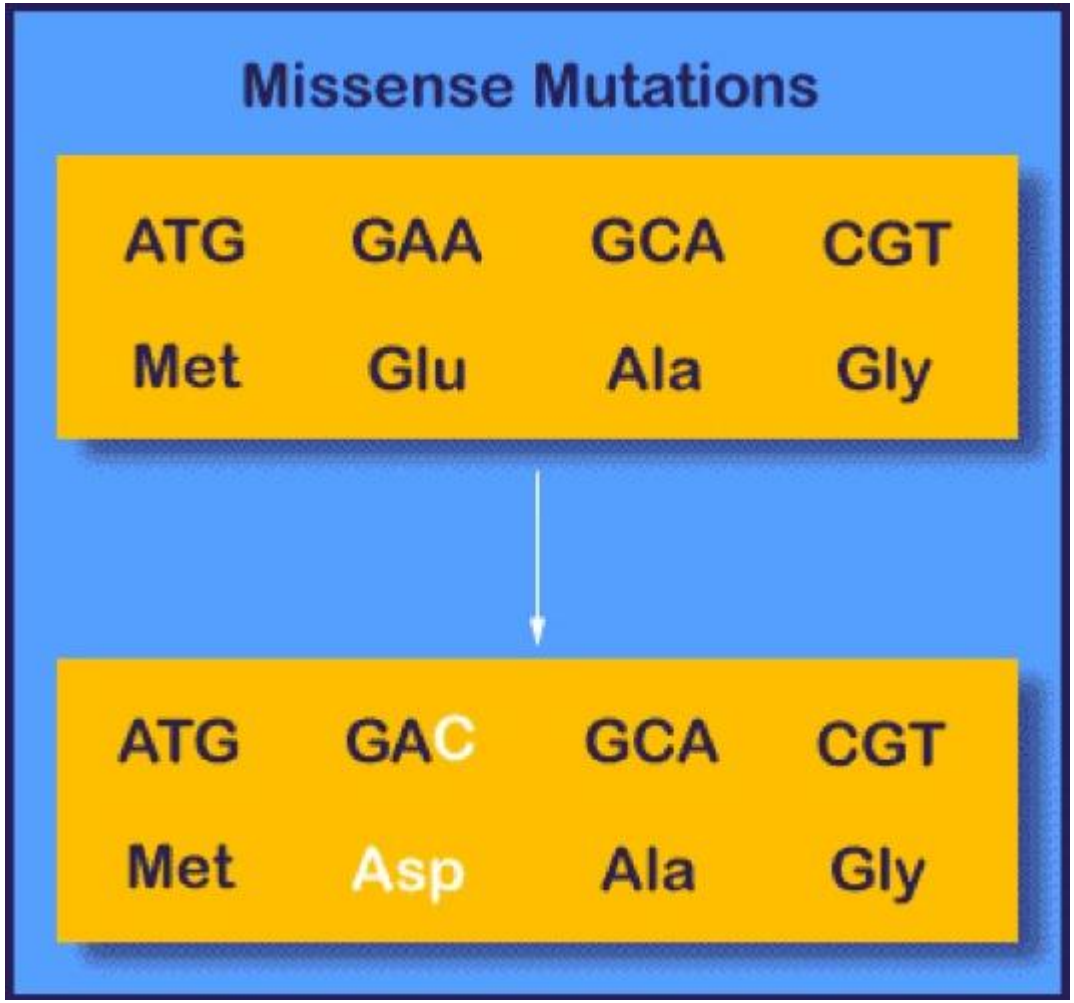
Types of mutations

- Deletion
- Duplication
- Inversion
- Insertion
- Translocation

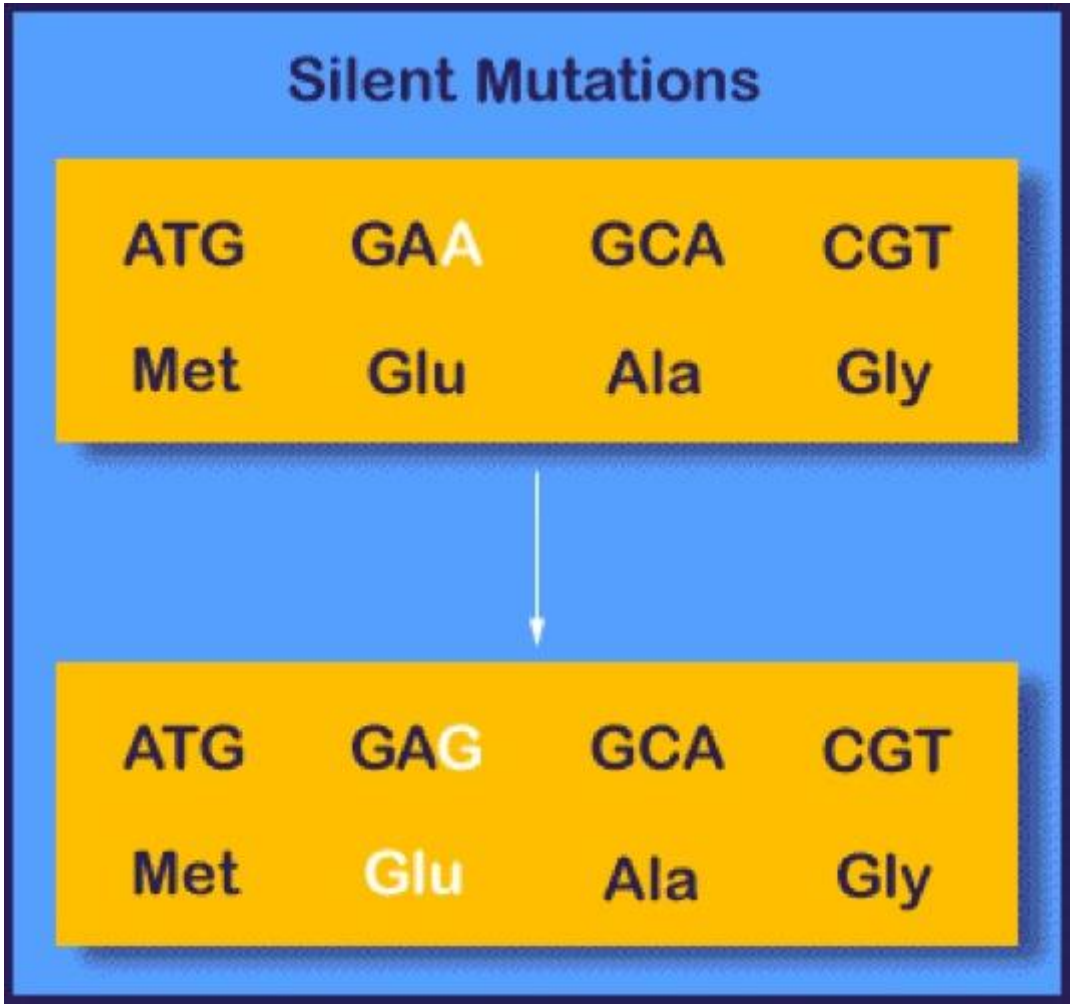


(National Human Genome Research Institute)

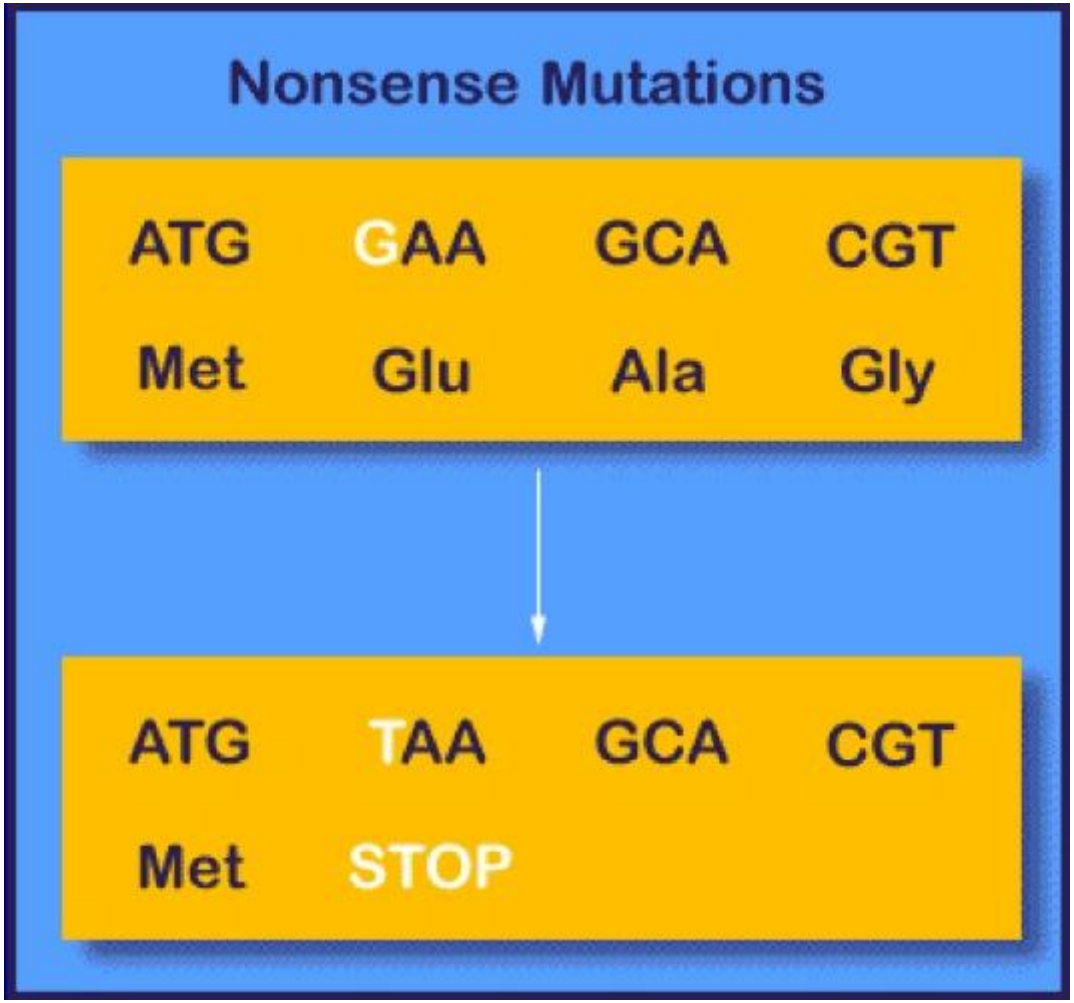
Types of mutations (continued)



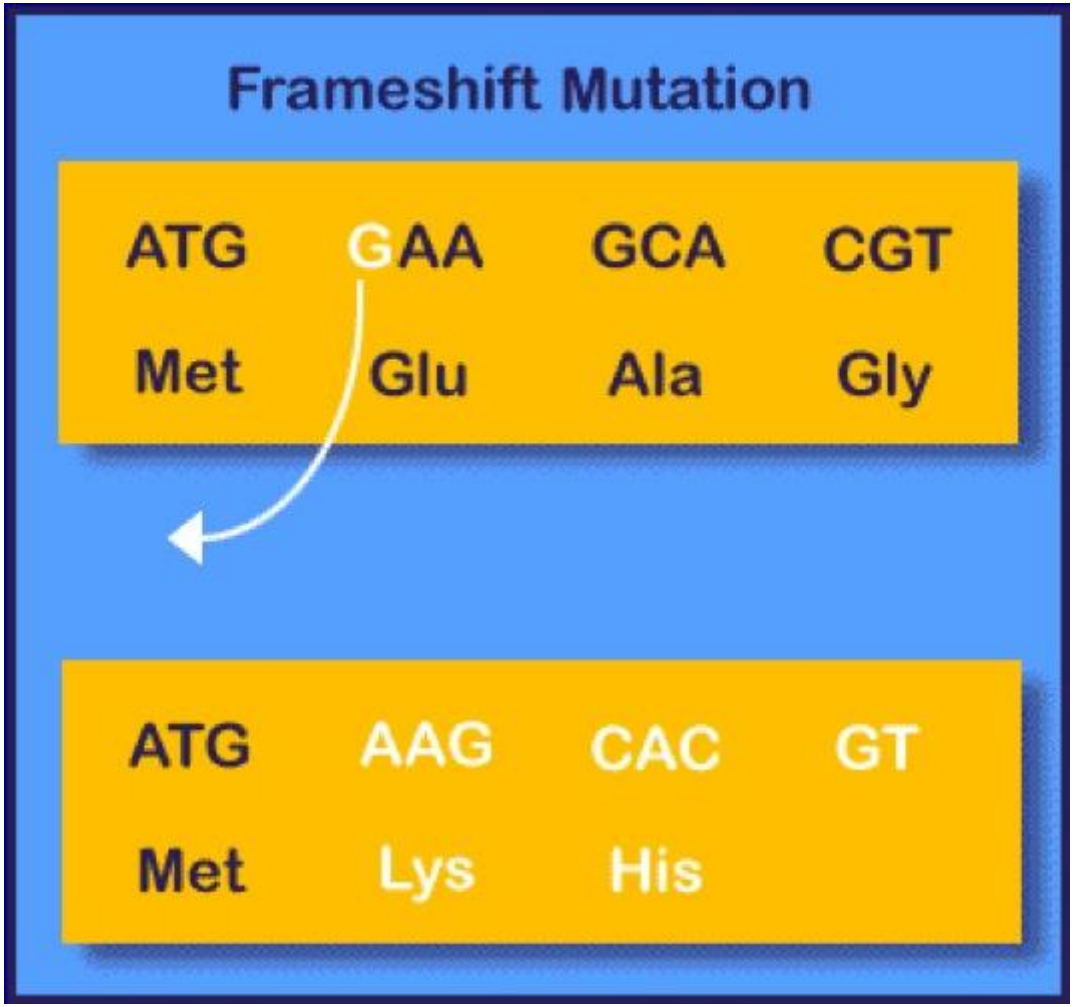
Types of mutations (continued)



Types of mutations (continued)

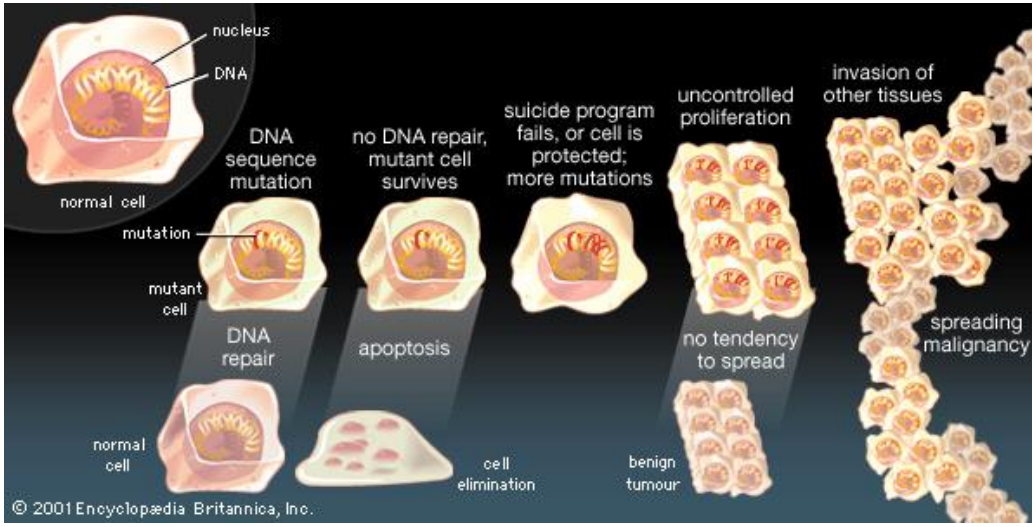


Types of mutations (continued)



DNA repair mechanisms

- **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



Variation is key

- 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 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...

Variation is key

- For a chromosome pair, the two alleles at a single genetic marker are called an individual's **genotype.**
- **Homozygous genotype**s have alleles that are identical (otherwise we talk about **heterozygous genotypes**).
- A **haplotype** is a sequence of alleles along the same chromosome.

Important references

- Ziegler A and König I. *A Statistical approach to genetic epidemiology*, 2006, Wiley
- URLs :
  - http://courses.washington.edu/b516/lectures_2009/?C=M;O=A
  - http://www.roche.com/education