Electronics and Computer Science

UNIVERSITY OF Southampton

COMP3212 Computational Biology

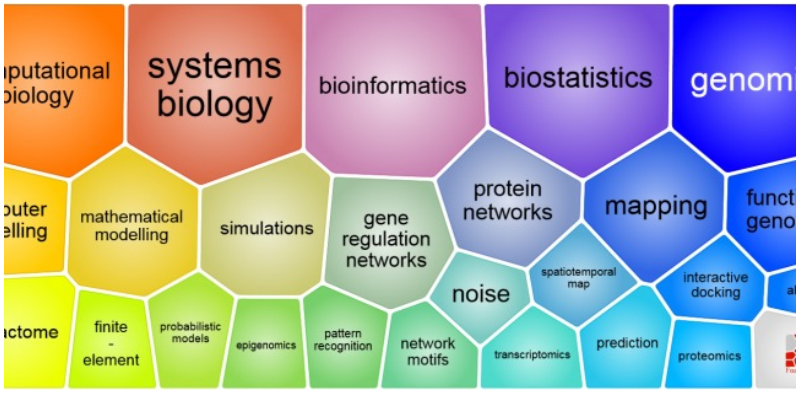
- Big Data Problem
- Computational tools
- New and growing datasets
- Dynamic Systems
- Networks
- Complexity

Björn Larsson, www.bjornlarsson.se

Tracy Melvin (tm@ecs.soton.ac.uk)

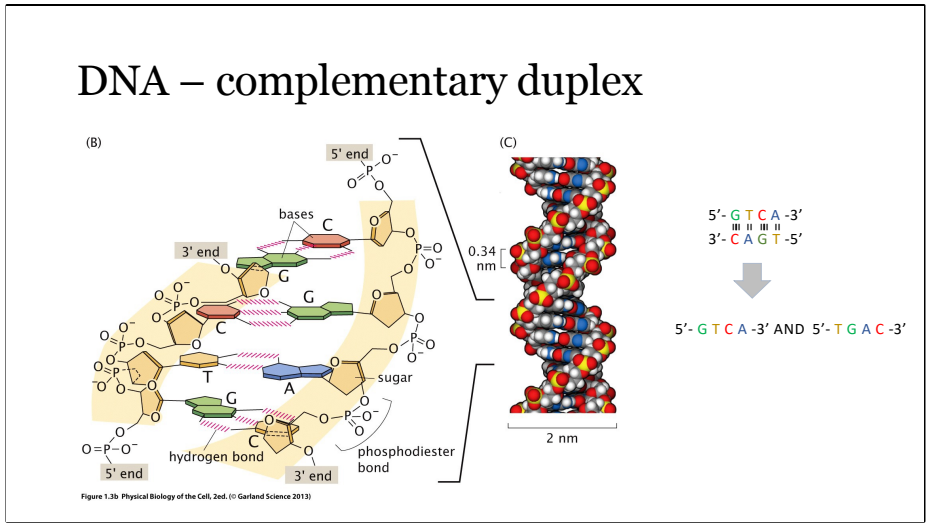
This module will provide an introduction to Computational Biology. Understanding 'Life' is currently a big data problem.
The role of computational tools so far in understanding the biology of humans for instance cannot be under estimated.
Time does not stand still, on a daily basis research is generating new data for many different organisms.
The existing data is supplemented with additional data, as we will learn, in most cases the data for many organisms is unique and has discrete variations.
All biological systems are dynamic; different molecules and structures inside the organisms vary in concentration. These are highly regulated and controlled by networks of interacting components.
Biology is highly complex, and the understanding of the complexity requires new computational approaches to be created.

COMP3212



https://erc.europa.eu/projects-figures/stories/computational-biology-spotlight-erc-projects

Computational Tools and approaches are pivotal to developing an understanding of the large datasets that have been accumulated for biological systems.
Indeed it is believed that biological systems are reliant upon ‘generated noise’ that is then regulated at the biomolecule, cell, tissue, organ level.

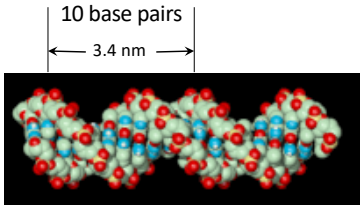


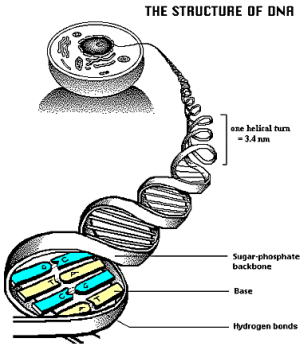
Nucleic acids are composed of *nucleotides*, made up of a 5-membered sugar, phosphate group and one of five primary nucleic acid bases, *adenine*, *guanine*, *cytosine*, *thymine* and *uracil*. There are two classes of nucleic acids (i) deoxyribonucleic acid (DNA), (ii) ribonucleic acid (RNA). For DNA the 5-membered sugar is deoxyribose and the primary bases are adenine, guanine, cytosine and *thymine*, and for RNA the sugar is ribose and the primary bases are adenine, guanine, cytosine and *uracil*.

Deoxyribonucleic Acid (DNA) is composed of two complementary polydeoxynucleotide chains that are associated *anti-parallel* to each other by hydrogen bonds between the base pairs (bp). adenine-thymine (by two hydrogen bonds), and guanine-cytosine (by three hydrogen bonds)

The duplex forms a double helix. As noted the two complementary DNA strands are associated anti-parallel to each other, the orientation is defined by the 5-membered deoxyribose with the phosphate group located at the 3' and 5' sites on the sugar. The convention is for the nucleic acid bases in a DNA sequence to be ordered from 5' to 3' with only one of the sequences in the duplex to be listed.

DNA – understanding the genomic data

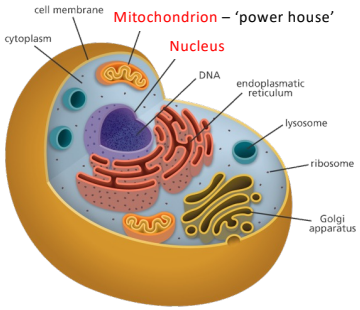




- DNA is a double helix
- The helical pitch is ~ 3.4nm
- DNA is ~ 0.2 nm wide
- The majority of the genomic DNA is present in the nucleus of the cell

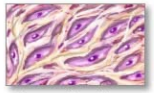
A human genome is made up of 64 billion nucleic acid base pairs. If the nuclear DNA in one single human cell was assembled end to end this would be 2 meters long. Although the majority of the genome is present in the nucleus, but further DNA is present in the mitochondrion (16.4k base pairs). We will learn about the mitochondrial DNA in a later lecture. As you see from this cartoon there are 10 base pairs (shown in blue) for one helical pitch. This is double helix with a wide (called major) and a narrow (called minor) groove in the double helical structure. The DNA helix is held together with the hydrogen bonds and the polymer is assembled with a sugar-phosphate backbone, shown on the left coloured red/yellow/grey and on the right as a ribbon schematic.

Organisation of the genome

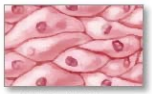


https://www.yourgenome.org/facts/what-is-a-cell

- A human cell contains **6.4 billion nucleotide pairs**.
- There is one copy of the genome inherited from the father and one from the mother, each contributing
- 3.2 billion nucleotide pairs
- The mitochondrial DNA is inherited from the mother
- A human body has ~$3x10^{12}$ cells containing nuclear DNA
- The DNA sequence in every nucleus is the same.

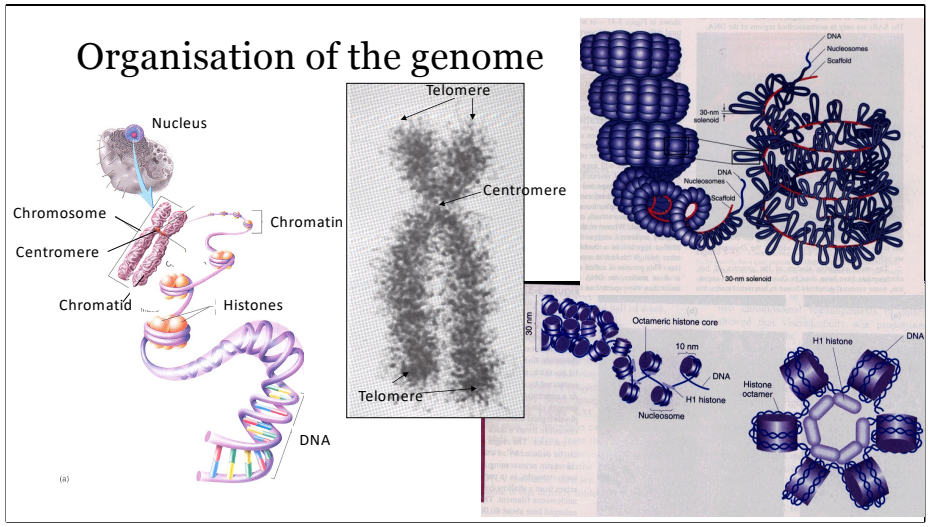

Connective tissue


Muscle tissue


Epithelial tissue

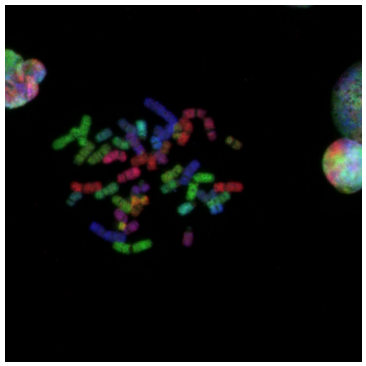

Nervous tissue

The human body is made up of ~200 different cell types. A human body contains ~$3x10^{13}$ cells, however 90% of the cells are enucleated blood cells (i.e. red blood cells), so there are ~$3x10^{12}$ cells containing nuclear DNA in a human body. The DNA is packaged within the nucleus and to a small extent in the mitochondria. The DNA is packaged as *chromosomes* in the nucleus of the cell. A small amount of DNA is also present in the mitochondria (16,500 base pairs); this contains 37 genes, 13 of which are key to the function of the mitochondria. Also present in a human is bacterial DNA (in bacteria), the quantity is debated, but is estimated to be ~$4x10^{13}$ bacteria (about 30% is lost during defecation).

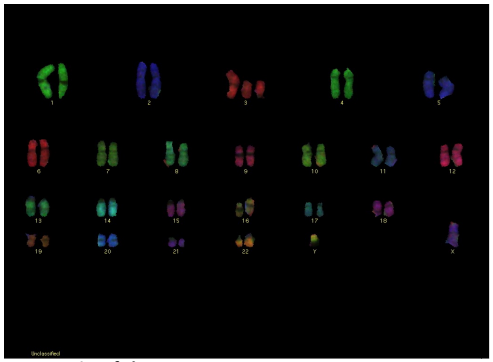


For a human there are 23 pairs of chromosomes, 22 are called autosomes and the 23rd pair are the sex chromosomes. The chromosomes are held within the nucleus of the cell. The autosomes are numbered by decreasing size (1-22). The largest Chromosome (Chr1) of 249 million base pairs (bps). However due to a historical error, the smallest chromosome was found to be Chromosome 21 (Chr21) and is 48 million bps. Chromosome 22 is the second smallest and has 51 million bps. As shown in the figure the chromosomes are often shown schematically as a X shaped structure. This is the structure that is formed during cell replication (i.e mitosis) to be described in a later slide. The Chromosome has a 'Centromere' holding the Chromatids together and 'telomeres' at the end. The figure in the middle is a low resolution representation of a micrograph obtained by electron microscopy. It is possible to see all the main loops. The DNA is highly organised and is held together with proteins called histones. These histones are bead like structures and the DNA is wrapped around these structures and then stacks of these histones are assembled on scaffolds for loops the assembled histone-DNA structures form 'solenoids' of 30nm cross-section. I like to consider the chromosome as a highly organised 'woolly jumper' like structure.

Organisation of the genome

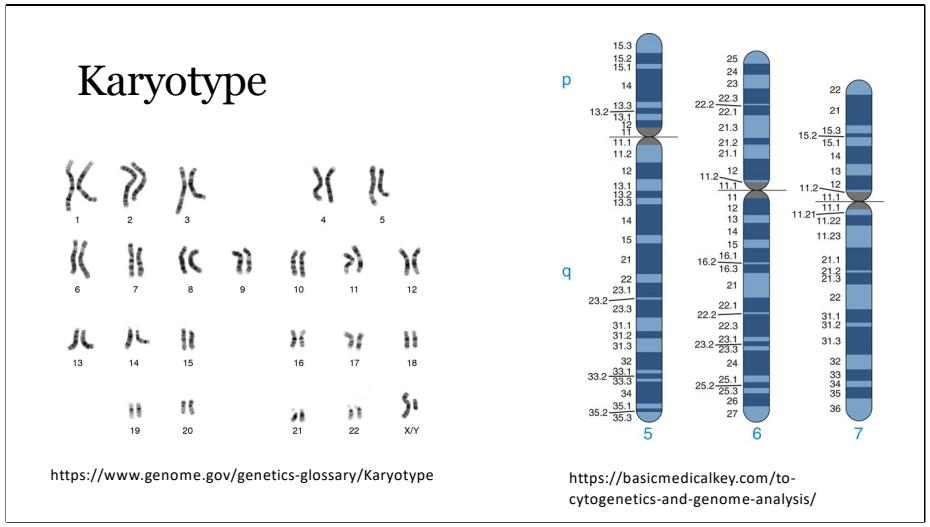


Nuclear DNA stained with different fluorophores



23 pairs of chromosomes
Chr1-22 from largest to smallest then XY – Male, XX - Female

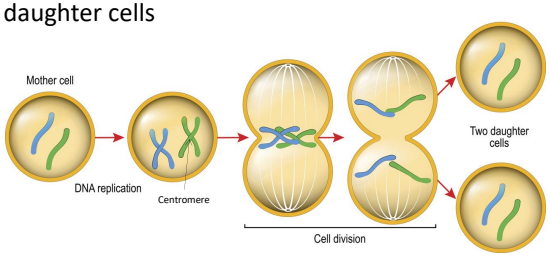
In these images there are samples of the human chromosomes stained with fluorescent dyes. The fluorescence emission wavelength (coloured here) are defined by the chromosome number (created by staining). In the right hand figure the data obtained on the right is 'sorted' using simple computational tools into the various sized/coloured pieces. The sample labelled 3 is shown as three pieces as one of the chromosomes was perhaps on top of another. For a human there are 23 pairs of chromosomes, 22 are called autosomes and the 23rd pair are the sex chromosomes. Females have two copies of the X chromosome and males have one X and one Y chromosome. The chromosomes are held within the nucleus of the cell. The autosomes are numbered by decreasing size (1-22). The largest Chromosome (Chr1) of 249 million bps. However due to a historical error, the smallest chromosome was found to be Chr21 and is 48 million bps. Chromosome 22 is the second smallest and has 51 million bps. You can see there are two chromosomes at the end, labelled X and Y. This is from a male, sex determinate chromosomes in females are XX.



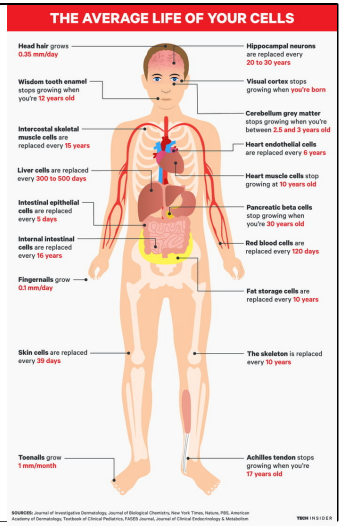
The chromosomes are always displayed with the large 'arms' below, the arms are labelled *p* (smallest) and *q* (largest). The term *karyotype* is the individual's collection of chromosomes, but is also used to define the technique that produces the image of an individual's chromosomes. The 'banding' is a result of a historical staining protocol with a chemical -Giemsa and is called G-banding. The cytogenetic location of genes are typically described by the Chromosome number, the *p* or *q* arm, and the banding number. The banding number starts at 11 closest to the centromere. The cytogenetic location of a DNA sequence is written according to the chromosome (1-22) and X and Y, arm and banding number. i.e. 17q12 (Chromosome 17, q arm, band 12). For example a sequence present in Chromosome 5 in the large arm in region 14 is labelled Chr5q14 Whilst the Giemsa staining (G banding) is no longer done routinely, the historical identification of sequences within the genomes still includes the location of the sequence in the genome 'loci'. This is found in the databases that we will talk about in a future lecture.

Cell Replication

Mitosis – a process where a single diploid cell divides into two identical daughter cells



Note: each of the chromosomes is duplicated by DNA replication, but the duplicate copies are held together by the centromere



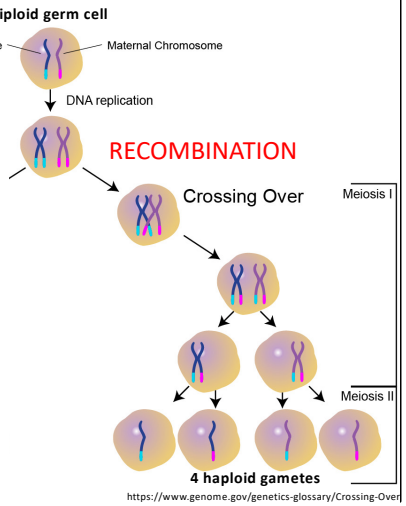
Mitosis (cell replication) occurs from within the embryonic stage (from Zygote (the fertilised egg)) through the whole adult life of a human and results in two genetically identical cells in which the total number of chromosomes is maintained. First DNA replication occurs to yield duplicated chromosomes which are condensed and are then one copy of each chromosome is pulled to opposite sides of the cells by spindle fibres. The cell then fully divides to *yield two identical diploid cells*.

It is important to remember that every cell prior and after replication contain two chromosomes and that the DNA sequences are similar – but different!

Inheritance

Meiosis – cell division in the **germ cells** in sexually-reproducing organisms to form the **gametes** (i.e. sperm and egg)

- Paternal chromosomes undergo a 'cross-over' – called **recombination**.
- There are an average of **26.4** recombination events for male and **41.1** for female germ cells.
- During fertilization cell fusion between the gametes (sperm and egg) results in a **diploid zygote containing** one copy each of the genomic DNA from the **Paternal** and the **Maternal** lineages.
- this results in **heterogeneity** in the inherited genomes between siblings (apart from twins).



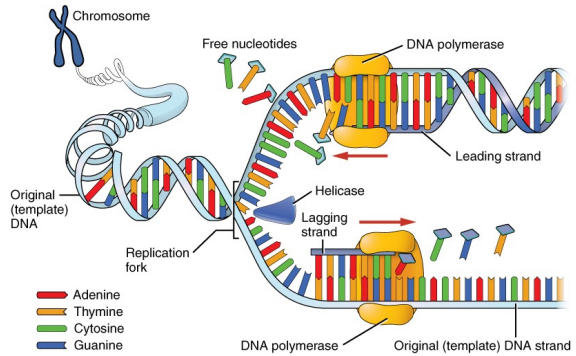
Now we will consider Meiosis (not mitosis as in the previous slide. Gametes are formed by *meiosis* in the gonads and are the reproductive cells of an organism. Human female gametes are called ova or egg cells and male gametes are sperm. Gametes are *haploid* cells, meaning that there is one copy of each chromosome (23) in each cell. At the start of meiosis in the germ cells, the homologous chromosomes pairs align in close proximity and each of the two chromosomes 'break' at the same location, the breaks combine with the opposite chromosome and form a connection, known as a *chiasma*. This is called homologous recombination. Following multiple homologous *recombinations* in all the autosomal chromosomes (autosome pair has the same morphology) and in the pseudoautosomal regions of the X/Y chromosomes, the cell divides twice *to yield four unique haploid* cells called gametes with one complete complement of chromosomes called *sister chromatids*.

Upon cell fusion of a male and female gamete a *zygote* is formed (not shown in the schematic). The zygote is *diploid*, meaning it contains two copies of the genome, one from each parent meaning that there will be heterogeneity between the two inherited genome sequences. In addition, the inherited sequences will be representative of different ancestral genes due to the multiple recombination events that have occurred during

meiosis. This results in heredity diversity within the offspring of the same parents. In addition, each individual's genome will be unique (unless the progeny arise from cell division of the zygote, *i.e.* twins). Although one copy of the genome is passed on from the father and mother to the offspring, the mitochondrial DNA is passed on from the mother only to all offspring.

The Zygote will undergo replication (mitosis) to generate the embryo.

DNA Replication



See the following animation: https://www.youtube.com/watch?v=TNKWgcFPHqw

DNA replication is shown in detail in the youtube video. In brief the helicase enzyme is responsible for 'breaking the hydrogen bonds that are formed between the nucleic acid bases' or what is called *melting* of the duplex to form two separate but complementary stands. Two double stranded DNA molecules are created by the enzyme, DNA polymerase together with free nucleotides (AGCT). Note the direction that this enzyme moves along the strand and how this is different for the two strands. The u-tube video gives a lot more detail. The key message is that two identical copies of DNA are made in creating the chromosome. This chromosome is then rapidly divided to yield two daughter cells in mitosis or four haploid gametes by meiosis. The difference between mitosis and meiosis is that following cell division the DNA sequence in each cell following mitosis is identical, whereas following meiosis the DNA sequence in each haploid gamete is different and is made up of different sequences originating from the mother and father. Note following fertilization of a haploid gamete the zygote will inherit different sequences from both sets of grandparents.

Mutations

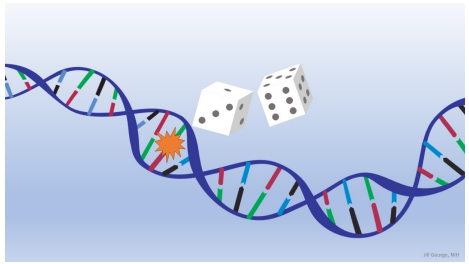
A mutation is an alteration in the nucleotide sequence of the genome of an organism.

Causes:

(1) DNA copying mistakes occur during replication.

(2) Exposure to environmental mutagens (i.e. radiation, chemical).

(3) Infection by retroviruses

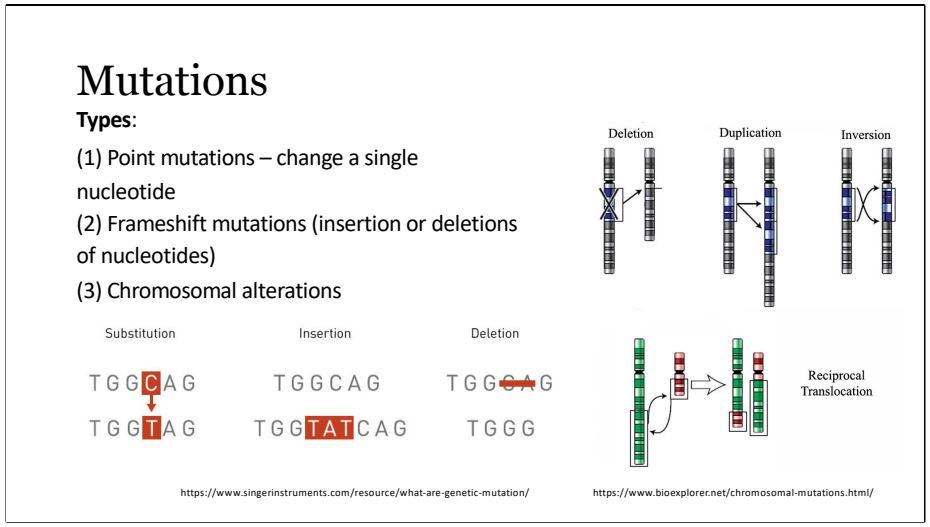


https://directorsblog.nih.gov/2017/04/04/random-mutations-play-major-role-in-cancer/

A *mutation* is the alteration of the nucleotide sequence of the genome of the organism, virus and extrachromosomal DNA. Genes can acquire mutations in their sequence during DNA or viral replication, mitosis, meiosis or as a result of other types of damage including environmental exposure (i.e. radiation, chemical), or by viruses.

Most mutations are either neutral in the phenotypical effect or deleterious – meaning they are removed by *negative selection*. Many mutations are corrected by the repair enzyme machinery of the cell. Very occasionally a mutation can convey an advantage in respect to survival or reproduction advantage to offspring providing *positive selection*.

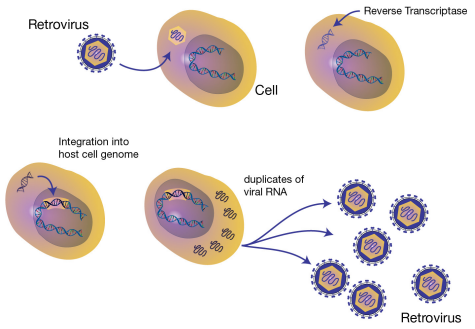
Mutations do not occur randomly in the genome, these occur at particular *hotspots.* These hotspots are typically defined by particular DNA sequence motifs/regions.



Mutations can include an exchange in a single nucleic acid base (the most common being G-T and is present in every 10 to 100kbps), an insertion or a deletion or a more profound change in the **DNA sequence**.
Other mutation types include deletions, duplication and inversion – these are when a sequence is lacking, a sequence is duplicated and a sequence is inverted within a **chromosome**.
A final type of mutation is reciprocal translocation and this is when there is an exchange (called translocation) between **two different chromosomes**.

Mutations

- Retro-viruses integrate into the host cell genome.
- Viruses can cause mutations within the genome.
- Ancient viruses are ~10% of the human genome.



https://www.genome.gov/genetics-glossary/Retrovirus

The human genome contains DNA where the sequences are derived from viruses. In many cases these were inserted into the genomes of pre-human ancestors. This DNA makes up about 8% of the human genome. In addition to these ancient viral sequences, cell infection by some viruses (i.e. HHV-6 a human DNA herpes virus) can result in inheritance of the viral DNA.

Transposons

- A **transposon** or a **transposable element** often called 'jumping *genes*' - change position within the genome, usually during replication.
- These can create or reverse mutations
- Transposons often result in duplication of the genetic material.
- Repeat sequences are formed each side of a transposon
- Approximately 44% of the human genome is occupied by transposons, <0.05% remain active.

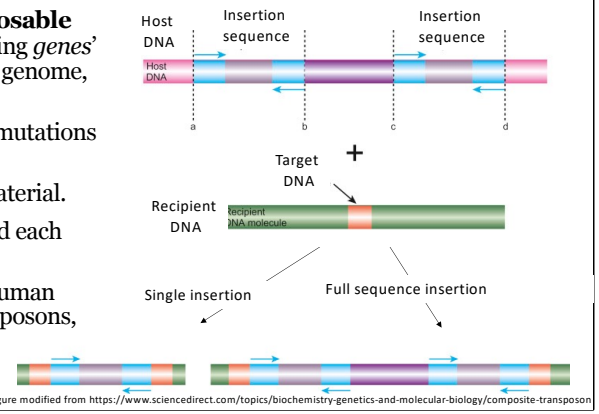


Figure modified from https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/composite-transposon

Transposons are often known as 'jumping genes'; they are transposable elements that are mobile repetitive sequences that make up ~ 45% of the human genome. Transposons can also act as mutagens as they can insert into a functional gene sequence, either in the intron, exon or promoter regions and can thus destroy or alter the gene's activity.

These are considered to have originated during evolution from viruses.

Mutations

Impact:

- Approximately 1 million 'DNA damages' occur in a human body in 1 day.
- The majority of these damages are repaired by elaborate repair enzymes.
- A small fraction of the DNA damage results in an alteration in the nucleotide sequence.
- The majority of these mutations are not deleterious, a small fraction lead to mutation.
- One in five 'healthy' adults may carry disease-related genetic mutations.
- 100-200 new mutations are passed on to the next generation.

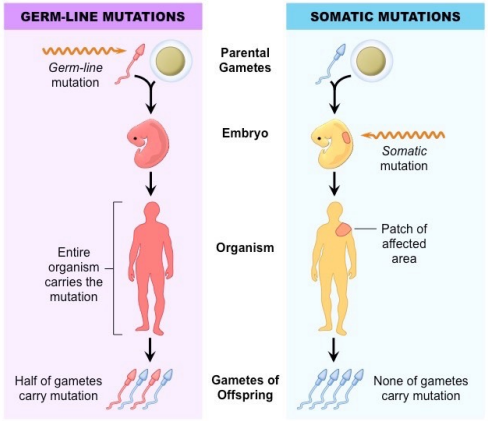
There are considered to be 1 million DNA damages in a human body every day. The majority of these are repaired. Although the progeny of parents inherit copies of the genome from each, they also inherit 100-200 new mutations. If we think carefully, it is only the fertilization of an egg that is going to yield an offspring - thus mutations that are inherited occur within the gonads of both parents only.

Mutations

Impact:

- Mutation is already present or occurs in the germ cells – Germline mutation
- Mutation occurs during the life of the individual – Somatic mutation.

https://ib.bioninja.com.au/standard-level/topic-3-genetics/33-meiosis/somatic-vs-germline-mutatio.html



Origins of mutations. Mutations in the germ-lines provide a whole organism that carries the new mutation. This organism can thus generate offspring there is a chance that they will inherit the mutation. These mutations are called Germline mutations.

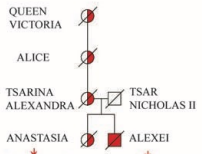
If mutations occur at any point after fertilization of the parental gametes then there are localized mutations. These mutations are carried through the cell lineage by mitosis. However none of the offspring carry the mutation.

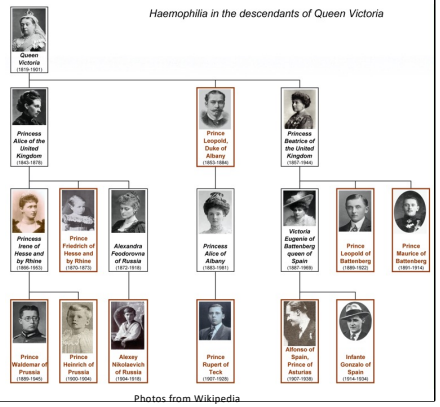
Germline mutation

Haemophillia is a sex-linked X chromosome disorder.

Manifests in males (**X**Y)

In females (**X**X) is **recessive**.

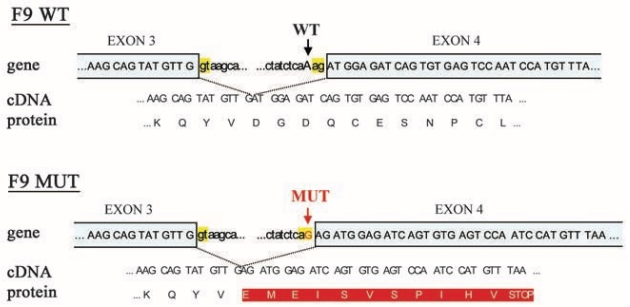




Single point Mutation present in X chromosome, thus the genetic disease manifests in males. Queen Victoria (with two X chromsomes) was recessive for the disorder. This means that she had no symptoms for Haemophillia, but some of her male descendants (male with XY chromosomes) were haemophilliacs and as a result died young. Recently the remains of the Romanov family were discovered and the DNA from all family members sequenced. It was found that Alexei (known to be Haemophilliac) and Anastasia was recessive for Haemophillia. The fact that the DNA was sequenced enabled the mutation to be detected.

Heritable mutation in X chromosome.

Haemophilia B in Romanovs caused by single point mutation A → G



WT = wild type (most common, forms functional protein)

F9 gene, present in X chromosome

cDNA – chromosomal DNA

EXON – coding region

Rogaev et al. Genotype Analysis Identifies the Cause of the "Royal Disease" *Science*, 2009, **326**, 817

The mutation present in Haemophilia B is caused by a single point mutation.
In the wild type (the sequence present in the majority of the population the mutation is A->G
In the next lecture we will revisit the Romanov example and in doing so will learn about genes, the structure of coding genes
Understanding the protein sequence
Impact of gene mutations on protein formation
The role of proteins and impact of protein folding and misfolding

Key ‘take home’ messages from the lecture

- DNA is made of two antiparallel strands of DNA of adenine + thymine and guanine + cytosine base pairs
- Each copy of the human genome is 3.2 billion nucleotide pairs, there are two copies in every cell apart from the germ cells. One copy is inherited from the mother and one from the father. Thus each human cell contains 6.4 billion nucleotide pairs.
- There are 23 pairs of chromosomes, one of these pairs is the sex chromosomes which are XY for male and XX for female
- The chromosomes are labelled with p (smallest) and q (largest) arms from the centromere starting from 11, locations along the chromosomes are identified from the G-stain banding
- Mitosis is the process that occurs during cell division leading to two identical cells with two sets of chromosomes – diploid cells.
- Meiosis is the process that occurs starting from a diploid germ cell to create four haploid gamete cells (sperm or egg).
- ‘Cross-overs’ called recombinations occur during meiosis, this results in hereditary diversity
- A mutation is an alteration in the nucleotide sequence of the genome of an organism (single point mutations, or more significant alterations from transposons, chromosome cross over.

This will not be described again, but it is important that all these points are understood – so any Questions?