# 582314 Genetics for Computer Scientists, 2 CU

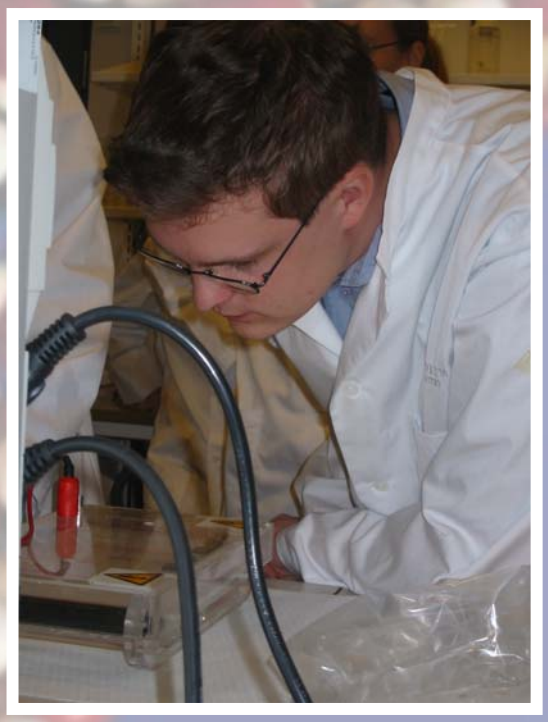
**15.3.-19.3.2004**

**Cum laude approbatur, optional courses**

**Department of computer science,**
**University of Helsinki**

## Course assistants:

Ranja Eklund
Inkeri Majuri
Vesa Ollikainen
Päivi Onkamo
Anna Rautanen
Jarno Tuimala



## Organizers:

- Helsinki Institute for Information Technology, Basic Research Unit, Department of Computer Science, University of Helsinki
- The Multifactorial diseases research group in Biomedicum, University of Helsinki
- The Finnish IT center for Science, CSC, Espoo

# Contents

# DNA and Genes

## From a Gene to a Functional Protein

At first, the biological university appears to be amazingly diverse from huge trees to tiny flowers, from a single cell bacterium, visible only under microscope, to human and other multicellular animals of all kind. Behind this variation overlies a powerful uniformity: all biological systems are composed of the same types of chemical molecules and they all use similar principles of organization at the cellular level.

**Figure 1.** Diversity of life

**Cells** are the basic components of all the living things. Despite of the wide variety in the function and appearance of cells, the basic structure of all cells is very similar. The biological universe consists of two types of cells: **prokaryotic cells**, which lack a defined nucleus and have a simplified internal organization, and **eukaryotic cells**, which have a more complicated internal structure including a defined, membrane-limited **nucleus**. Prokaryotic organisms consist of two distinct types: bacteria (often called "true" bacteria) and archaea. Eukaryotes include all the other forms of life, for example all animals, plants and fungi.

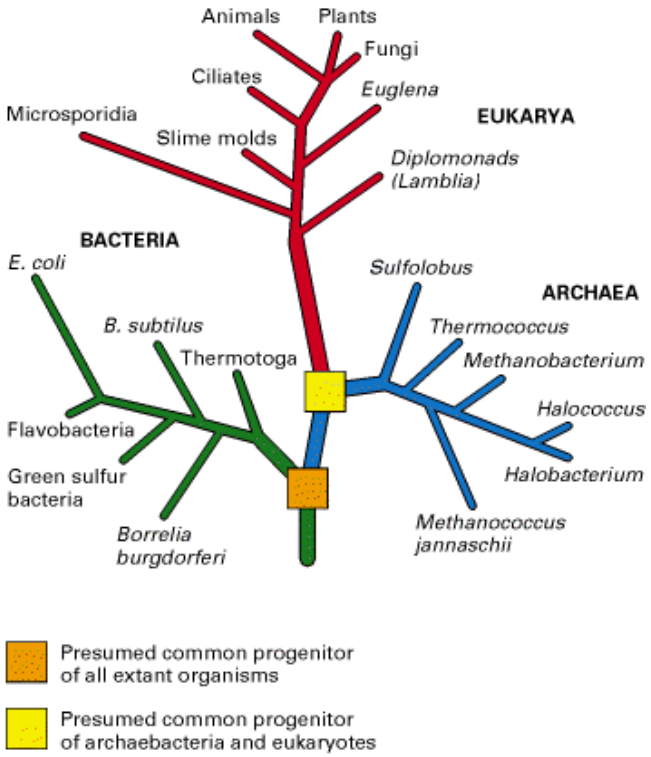



**Figure 2:** The three kingdoms of life.
The overall structure of a cell is very similar between all eukaryotes, and only a few differences are found when comparing animal and plant cells.

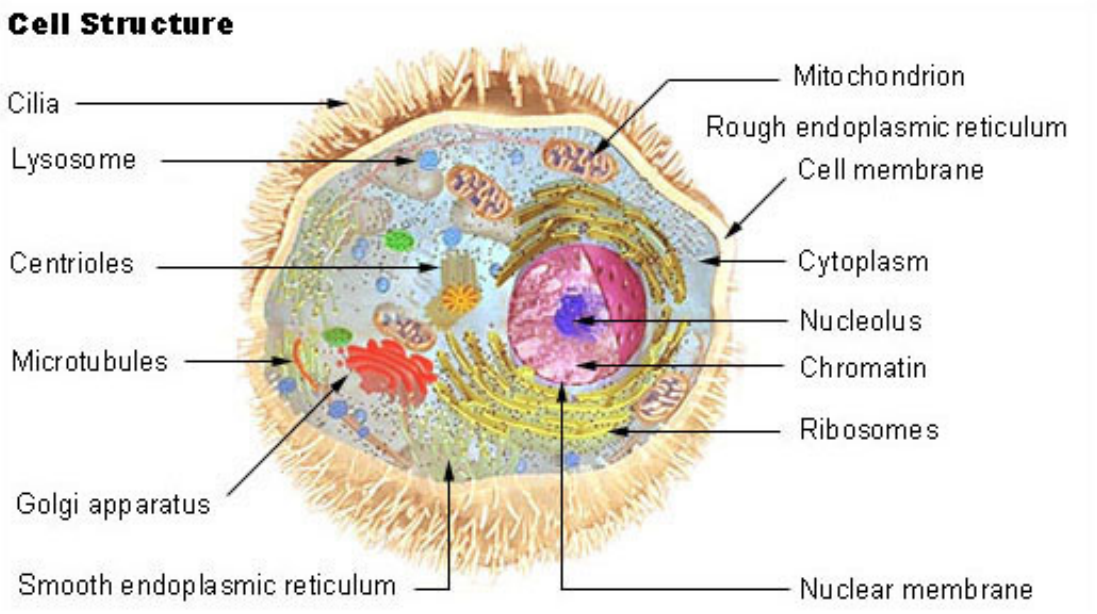



**Figure 3:** Cell Structure of an animal cell

Genetic material of cells is located to **the nucleus** of the cell. The genetic information is stored in **Deoxyribonucleic acid (DNA).** DNA is the storehouse, or a cellular library, that contains all the information required to build the cells and tissues of an organism. Usage of the information stored in DNA is what makes us at the same time unique and functional organisms.

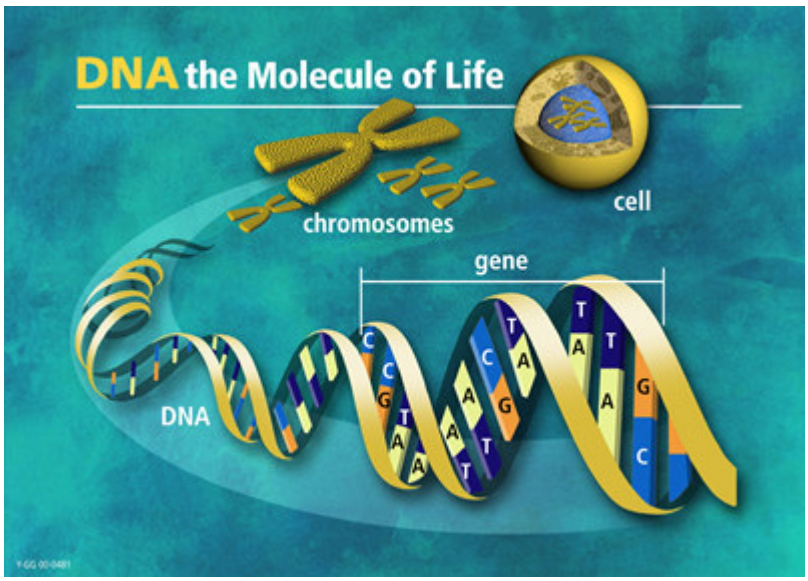



**Figure 4:** DNA is located in the nucleus of the cell

The genetic information stored in DNA is made available to the cell by the process called **gene expression**. The use of the genetic information is described to be the central dogma, which states that the **information of a gene** is transferred from the DNA and converted to a **protein**. **RNA** molecules work as messengers in this process. During gene expression, DNA molecules copy their information by directing the synthesis of a RNA molecule coding the complementary sequence in a process called **transcription**. The RNA then directs the synthesis of a protein whose amino acid sequence is determined by the base sequence of the RNA. This process is known as **translation**. The amino acid sequence of the protein determines its 3D structure, which in turn defines the protein's function. The central dogma states that the transfer of information can occur only in one direction from DNA to RNA to protein, and it can not happen in a reverse way. The function of the cells, and in turn of living organisms, is dependent on the coordinated activity of many different proteins. The biological information contained in the genes acts as a set of instructions for synthesizing functional proteins at the right time in the right place.

## DNA

DNA is a polymer containing a chain of nucleotide monomers. Each nucleotide consists of a sugar, a base and a phosphate group. The sugar in the DNA is 2´-deoxyribose and it consists of five carbon atoms. There are four types of bases: adenine (A) and guanine (G), which have two carbon rings, **purines** and thymine (T) and cytosine (C), which have a single carbon ring and are called **pyrimidines**. The bases are connected to 1´carbon of the deoxyribose-sugar. Together a sugar and a base are termed a nucleoside. Nucleosides have from one to three phosphate groups attached to the sugar and they are building blocks of DNA.

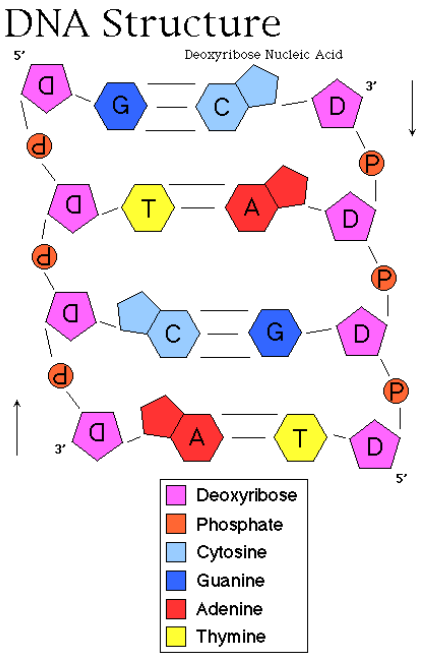



**Figure 5:** The DNA structure

**DNA-chains** are formed when nucleosides with three phosphates are joined to form **DNA polynucleotide chain**. Two phosphates are lost during polymerization and the nucleotides are joined together by the remaining phosphate. The bond is called **phosphodiester bond** and it is formed always in same chemical direction. The order of the bases varies and the **sequence of bases** encodes the genetic information. Polynucleotides can be extremely long.

**DNA molecules** are composed of two polynucleotide strands wrapped around each other to form a double helix. The sugar-phosphate part of the molecule forms a backbone. The bases face inwards and are stacked on the top of each other. The two nucleotide chains run in opposite directions. The double helix is usually twisted in right handed way and it rotates a full circle every 10 bases.

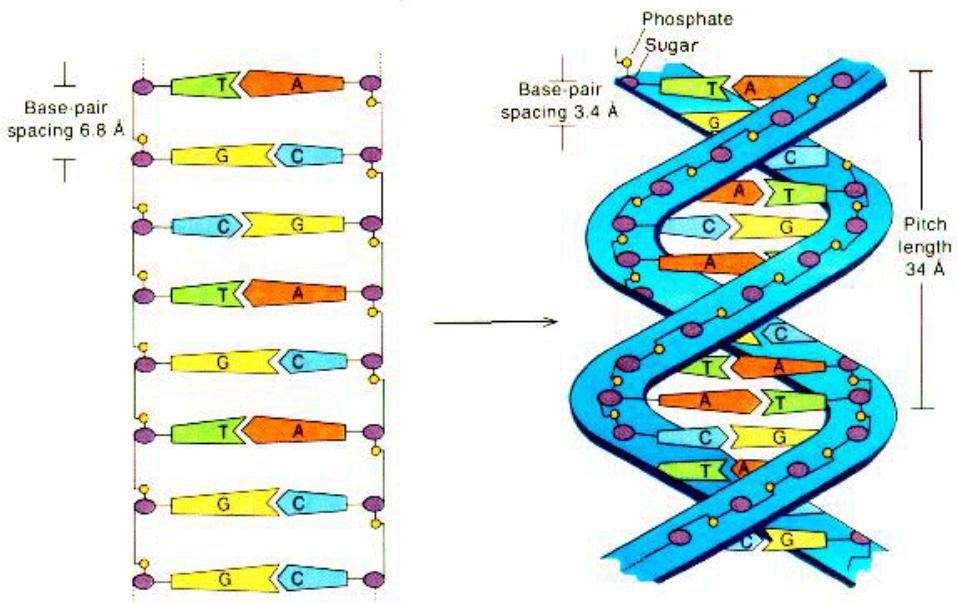



**Figure 6:** DNA molecules are composed of two polynucloetide strands wrapped around each other to form a double helix.

**Complementary Pairing**

The bases of the two polynucleotide chains interact with each other. The space between the polynucleotides is such that two-ring purine interacts with a single-ring pyrimidine. This means that thymine always interacts with adenine and guanine interacts only with cytosine. Interaction between the two bases is stabilized by hydrogen bonds, which form between the bases. Two hydrogen bonds are formed between A and T and three between G and C. This means that the G-C bonds are stronger than A-T-bonds. The way that the bases form pairs between the two DNA strands is known as the **complementary base pairing**. Other combinations than G-C and A-T do not work because they are too large or small to fit inside of the helix and the incorrect pairing also prevents correct forming of hydrogen bonds needed to stabilize the double helix structure. Because G must always bind to C and A to T, the sequence of the two strands are related to each other and are said to be **complementary** with each other. The sequence of one strand can be used to predict the sequence of the other one and it also determines it. This means that one strand can be used as a model while making the other one. This is a vital mechanism for retaining genetic information and passing it on to other cells following cell division. Complementary base pairing is central to the way that DNA sequences are later transcribed into mRNA.

**The Genetic Code**

The genetic code describes how base sequences are converted into amino acid sequences during protein synthesis. The DNA sequence of a gene is divided into a series of units of three bases. Each set of three bases is called a **codon** and it specifies a particular amino acid. The four bases in DNA can combine as total $4^3 = 64$ codons, which specify the 20 amino acids found in proteins. Because the number of codons is greater than the number of amino acids, almost all amino acids are encoded by more than one codon. Usually the identity of the last letter of the codon is insignificant for the amino acid determination. Some specific codons have also role in directing the protein synthesis as they function as an initiation and stop codons for protein synthesis. In addition to identifying the start of protein synthesis, the initiation codon determines the **reading frame** of DNA sequence. Normally only one reading frame contains useful information and the two other reading frames usually contain several stop codons which prevent them from being used to direct protein synthesis.

Initially the genetic code was believed to be universal, meaning that all organisms would recognize individual codons as the same amino acids. However, it has been shown that there are some rare exceptions in the code.

| First Base | Second Base: U | Second Base: C | Second Base: A | Second Base: G | Third Base |
|---|---|---|---|---|---|
| U | UUU Phe | UCU Ser | UAU Tyr | UGU Cys | U |
| | UUC Phe | UCC Ser | UAC Tyr | UGC Cys | C |
| | UUA Leu | UCA Ser | UAA Stop | UGA Stop | A |
| | UUG Leu | UCG Ser | UAG Stop | UGG Trp | G |
| C | CUU Leu | CCU Pro | CAU His | CGU Arg | U |
| | CUC Leu | CCC Pro | CAC His | CGC Arg | C |
| | CUA Leu | CCA Pro | CAA Gln | CGA Arg | A |
| | CUG Leu | CCG Pro | CAG Gln | CGG Arg | G |
| A | AUU Ile | ACU Thr | AAU Asn | AGU Ser | U |
| | AUC Ile | ACC Thr | AAC Asn | AGC Ser | C |
| | AUA Ile | ACA Thr | AAA Lys | AGA Arg | A |
| | AUG Met | ACG Thr | AAG Lys | AGG Arg | G |
| G | GUU Val | GCU Ala | GAU Asp | GGU Gly | U |
| | GUC Val | GCC Ala | GAC Asp | GGC Gly | C |
| | GUA Val | GCA Ala | GAA Glu | GGA Gly | A |
| | GUG Val | GCG Ala | GAG Glu | GGG Gly | G |

**Figure 7**. The genetic code describes how bases are converted into amino acid sequences during protein synthesis. DNA's code is first translated to a mRNA molecule. In mRNA synthesis thymine bases (T) are replaced by uracile bases (U). mRNA is used as a instruction for protein synthesis.

## Genes

The biological information needed by an organism to reproduce itself is contained in its DNA. The genetic information is encoded in the base sequence of the DNA, and it is organized as a large number of genes. Genes contain the instructions for the synthesis of a protein. The genetic information is expressed as different kind of proteins taking care of proper function of the cells and organisms.

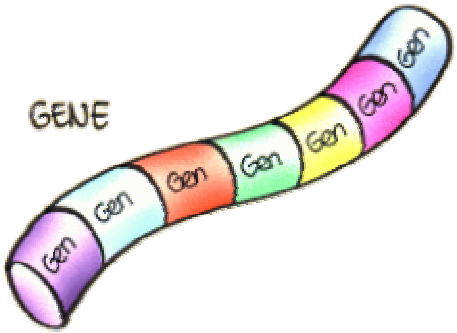



**Figure 8**: Simplified figure of a group of genes.

In physical terms, **a gene** can be described as a discrete segment of DNA with a base sequence that encodes the amino acid sequence of a polypeptide. In molecular terms, a gene is the entire DNA sequence required for synthesis of a functional protein. Besides the area coding the actual protein (coding region), a gene includes control elements and areas, which function is not known. Genes vary a lot in size, from 100 base pairs to several million base pairs. In higher animals genes are present on a series of extremely long DNA molecules called chromosomes. Genes do

not appear tightly one after the other. Vice versa, the genes are very dispersed and they are separated from each other by sequences that do not appear to contain useful information. Only one of the two strands of the DNA double helix carries the biological information and is called the **template strand**. It is used to produce a RNA-copy of complementary sequence, which directs the synthesis of a protein. The other strand is called non-template strand. Both strands of the double helix have the potential to act as the template strand: different genes are encoded on different strands.

The capacity of DNA molecules to store information is enormous. For a DNA molecule that is *n* bases long, the number of different combinations of the four bases is $4^n$. Even for a very short DNA molecule the number of different sequences possible is very high.

**Organizing Cellular DNA**

Because the total length of cellular DNA in cells is up to hundred of thousand times the cell's length, the packing of DNA must be very efficient. DNA is organized into chromosomes and the packing is crucial in order to fit them into cells. In eukaryotic cells, DNA is associated with about equal mass of special proteins called **histones**. DNA and histones form together a highly condensed structure, called **chromatin**. The building block of chromatin is the **nucleosome**, consisting of eight histone proteins, around which is wrapped about 146 base pairs of DNA. Each eukaryotic chromosome contains a single linear DNA molecule. Chromatin is further organized into large units of hundreds to thousands of kilobases in length called chromosomes. Each chromosome is composed of a single DNA molecule packed into nucleosomes and folded into a 30 nm fiber, which is attached to a scaffold at specific sites. Additional folding of the scaffold further compacts the structure into the highly condensed form of chromosomes.

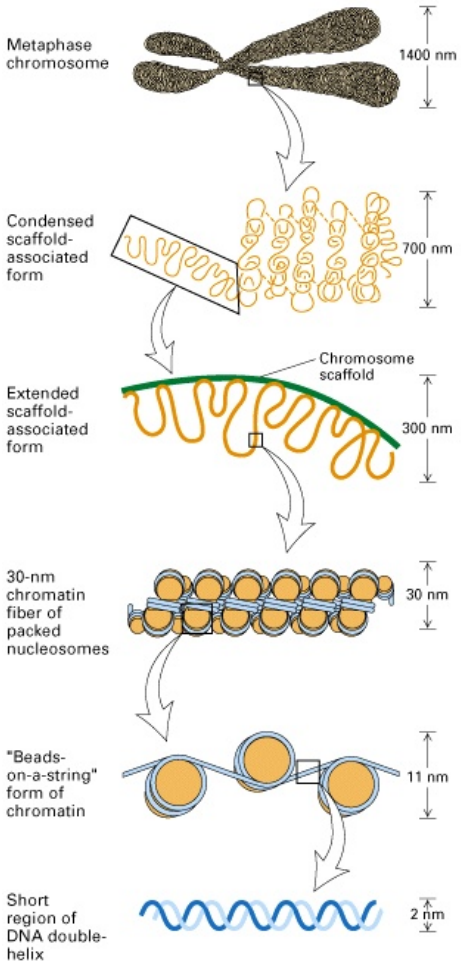



**Figure 9**: Packing of chromatin and the chromosome scaffold in chromosomes.

**The Human Genome**

The term "human genome" is used to describe the different types of sequence that together make up the total DNA of a human cell. The DNA in the human genome is about 3 billion base pairs long and is estimated to contain about 30 000 genes. The DNA is arranged as a set of 23 chromosomes, each of which is a single, double stranded DNA molecule 55 - 250 million base pairs long. The genes and gene related sequences account for about 25 % of the DNA and only 5 % of the genomic DNA in humans encodes proteins. The reminder is called extragenic DNA and has no known function.

The coding information in a gene is present as a series of segments of DNA sequences called **exons** separated from each other by intervening non-coding sequences called **introns**. Genes vary in length and also in respect to the number and sizes of the introns. Additional sequences are also present and associated with genes. Leader and trailer sequences occur at the ends of a gene and they are needed in protein synthesis as signals for protein synthesis machinery. Promoter sequence regulates protein synthesis and is needed to activate it. Some genes exist as numbers of copies with identical or related sequences that can be grouped into gene-families.

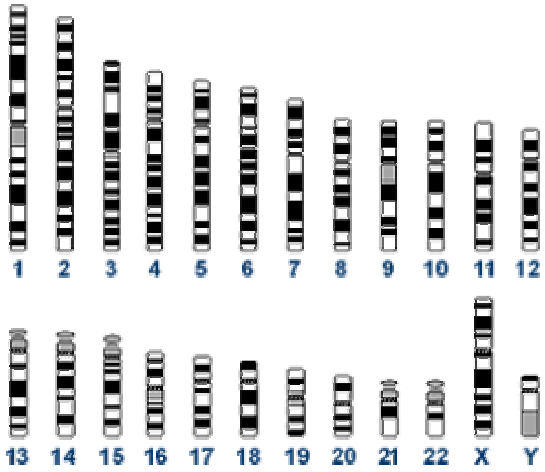



**Figure 18**: Human DNA is arranged as a set of 23 chromosomes.

**Genotype and phenotype**

Phenotype is the **outlook**, the physic appearance, of an organism. It is the sum of the atoms, molecules, macromolecules, cells, structures, metabolism, energy utilization, tissues, organs, reflexes and behaviors. Therefore phenotype is anything part of the observable structure, function or behavior of a living organism.

Genotype is the **genetic information** coded in to the DNA of an organism. The genetic information is used as a set of instructions for building and maintaining an organism. The genetic information is copied at the time of cell division or reproduction and is passed from one generation to the next. The genotype gives the instructions for everything happening in a cell and an organism.

**Passing on the genetic information**

Genetic information is passed on to offspring in the **sexual reproduction**. The sexual reproduction is the only way that new characteristics can develop. The genetic information is located to our chromosomes. All **somatic** human cells have **46 chromosomes**, from which we have inherited 23 from both parents. Somatic cells include all the cells found from the body, except those needed to produce offspring. Therefore the **diploid** chromosome count of a human cell (2n) is 46, which includes of two versions of one chromosome, one from the mother and one from the father.

Cells needed to carry the genetic information for the next generations, **gametocytes**, differ from the normal cells. They have only **23 chromosomes** and are called **haploid** (n). Production of gametocytes includes special phase, **meiosis**, which is needed to reduce the amount of chromosomes from 46 to 23. Meiosis is also necessary in the context of enabling the **rearrangement** of the genetic material and so the evolution of the new characteristics. Gametocytes only have one single copy of each gene.

In the fertilization, an egg (n, 23 chromosomes) from the mother and a sperm (n, 23 chromosomes) from the father merge and undergo a fusion to form a **zygote** (2n, 46 chromosomes). The fertilized egg undergoes millions of **mitosis** and the cells differentiate to form everything needed to construct a new individual.

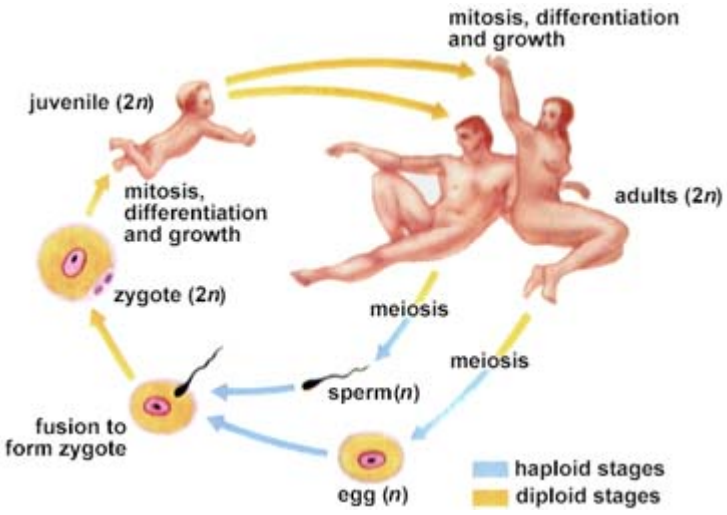
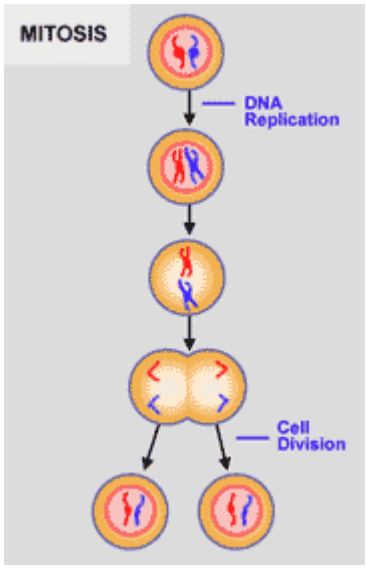



**Figure 10:** Sexual reproduction includes haploid and diploid stages

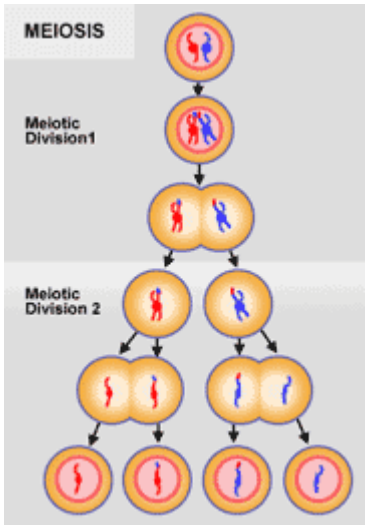
## Mitosis

Mitosis takes place every time when normal, asexual cell divides. During mitosis each chromosome divides into two daughter cells. The number of chromosomes per cell remains unchanged because before a cell enters mitosis, each chromosome **has been duplicated** in DNA replication and consists of two identical strands known as **chromatids**. Mitosis is the process whereby each of the chromatid pairs separate and disperse into a separate daughter cell. Mitosis is a continuous process, which usually lasts 1-2 hours. Mitosis produces **two similar copies** from the mother cell. Mitoses are happening all the time, everywhere in an organism. Mitosis is needed for example for the hairs and nails to grow, skin to renew and so on.



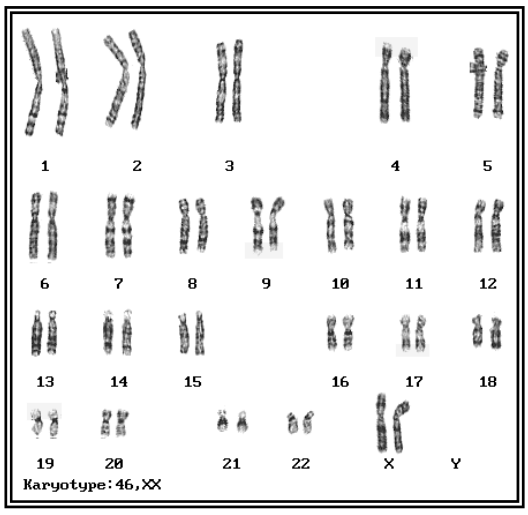

## Meiosis

Meiosis is the process of nuclear division, which occurs during the final stage of **gamete formation**. After mitosis the daughter cells have a diploid chromosome count (2n, 46 chromosomes). In the meiosis the diploid count is **halved** so that each mature gamete receives a haploid chromosome count (N, 23 chromosomes). Meiosis takes place only at the final stage of gamete maturation. Meiosis can be considered as two divisions known as meiosis I and meiosis II.




## Inheritance

Human genome includes 46 chromosomes, from which 44 are called **autosomes**. Autosomes are uniform in female and male. Remaining two chromosomes are **sex chromosomes**. In human sex chromosomes are type X or Y. Females have two X chromosomes and male have one X and one Y chromosome.



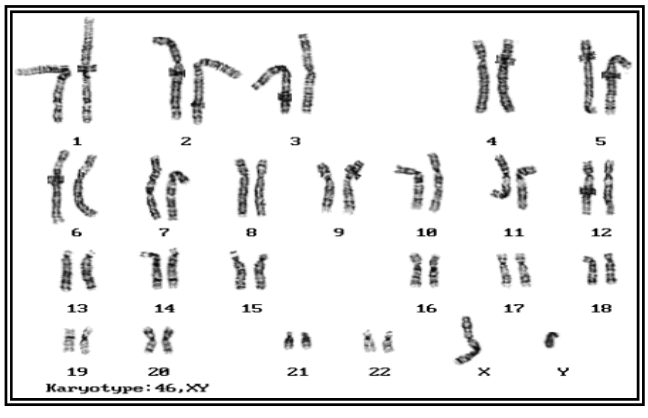



**Figure 11**: Normal female and male chromosomes.

We have two versions from each of our autosomal chromosomes, the **homologous chromosomes**, one from the mother and one from the father. Consequently we have also two copies from each gene. The location of a gene in a chromosome is called a **locus**. **Allele** is a mutually exclusive form of the same gene, occupying the same locus on the homologous chromosomes and governing the same function. Person can be heterozygous or homozygous according to one allele. **Heterozygous** means that two different forms of one allele are found from the homologous loci. If the homologous loci are **homozygous** they are occupied by similar alleles. A character is said to be **dominant** if it is manifest in the heterozygote and **recessive** if not. The dominance and recessiveness are properties of characters, not genes.

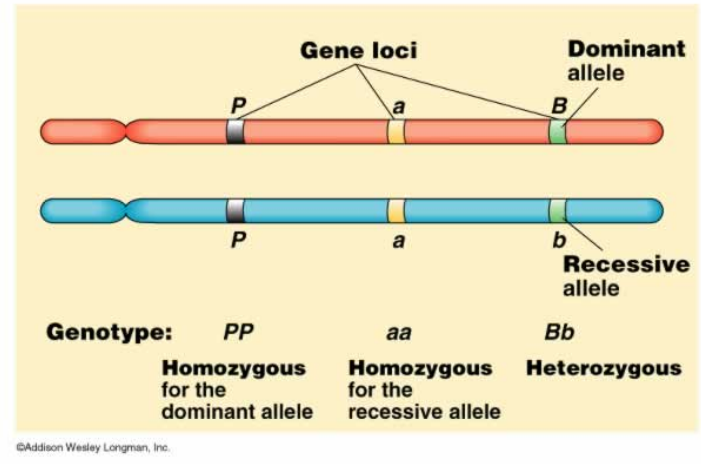



**Figure 12:** Homologous chromosomes.

**Mendelian inheritance**

The simplest genetic characters are those whose presence or absence depends on the genotype at the single locus. That does not mean that this characteristic is coded by one gene, because the expressions of all the human characters require a large number of genes and environmental factors. Sometimes a particular genotype at one locus is needed for the character to be expressed. These are called Mendelian characters: There are over 10,000 Mendelian characters nowadays known for humans. Mendelian pedigree patterns can be shown with drawn pedigrees. Special symbols have been chosen to represent family members and their characteristics in the genetic pedigrees.

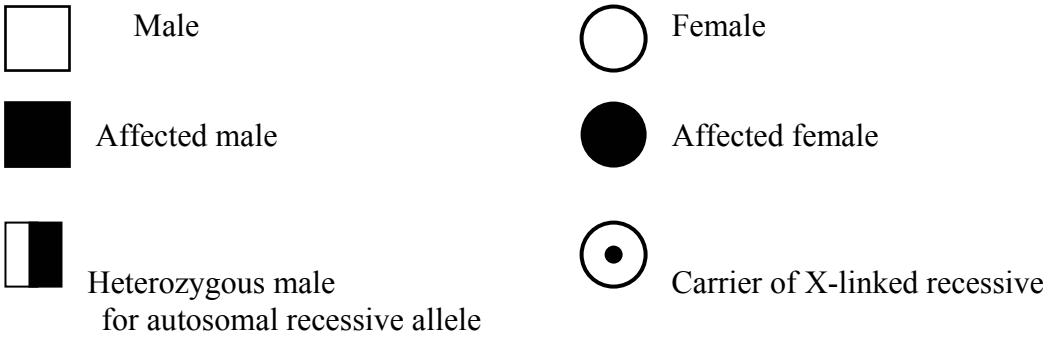



**Figure 13:** Samples of the symbols commonly used in genetic pedigrees

**a. Autosomal dominant inheritance**

An affected person usually has at least one affected parent. Affected individuals can be of either sex, and the characteristic can be transmitted by either sex.

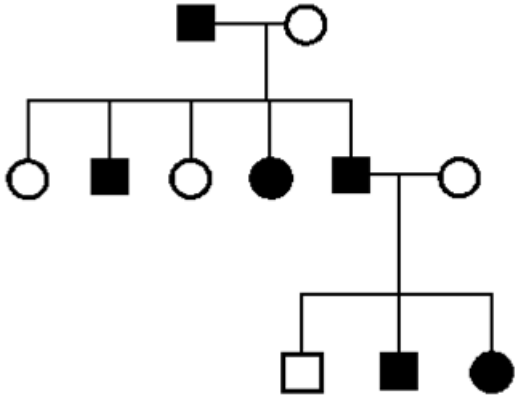


**Figure 14**: Typical pedigree showing autosomal dominant inheritance.

**b. Autosomal recessive inheritance**

Affected offspring are usually born to unaffected parents, so the parents are usually carriers. The trait affects both sexes alike.

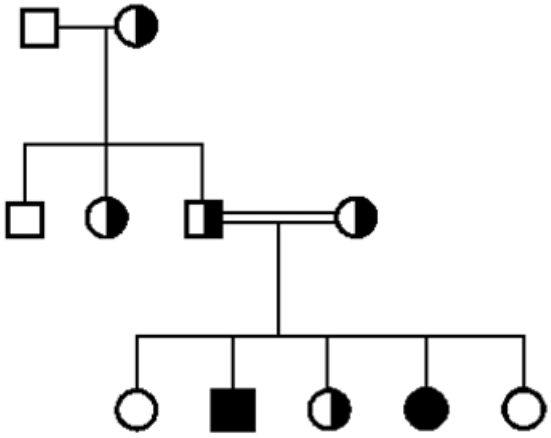


**Figure 15**: Typical pedigree showing autosomal recessive inheritance

c. **X-linked recessive inheritance**

Affects mainly males. Affected males are usually born to unaffected parents, when the mother is usually an asymptomatic carrier and may have affected male relatives. Females can be affected if the father is affected and the mother is a carrier. Female affection can also be due to some other rare coincidence.

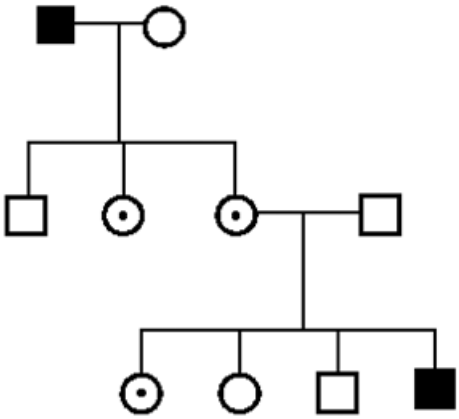


**Figure 16**: Typical pedigree of X-linked recessive inheritance.

d. **X-linked dominant inheritance**
Affects either sex, but more often females than males. Females are often mildly and more variably affected than males.

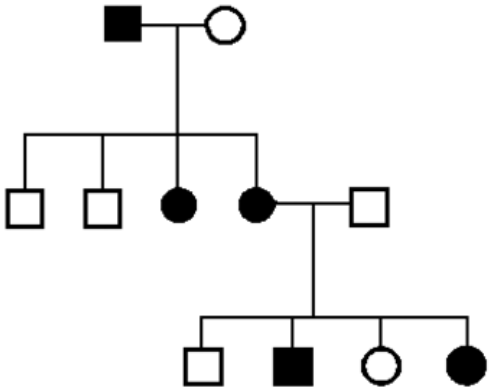


**Figure 17**: Typical pedigree of X-linked dominant inheritance.

e. **Y-linked inheritance**
Affect only males. Affected male always have an affected father.

**Non-Mendelian characteristics**
Non-Mendelian characteristics may depend on two, three or **many genetic loci**. The characteristic is also affected by greater or smaller contributions from **environmental factors**. Non-Mendelian inheritance can also be termed as **multifactorial**. The genetic determination may involve a small number of genetic loci (oligogenic) or greater amount of loci (polygenic).

**Mutations**

The DNA sequence of a gene determines the amino acid sequence of the protein that it encodes. It is important that the DNA sequence is conserved because alterations in the protein's amino acid sequence may affect the ability of the protein to function, which in turn may have serious effects on the organism. Alterations in the DNA are caused by many chemical and physiological agents and they are also due to rare errors in DNA replication. These changes are known as **mutations**. Once introduced, the DNA sequence changes are made permanent by DNA replication and are passed on to daughter cells in following cell division. Since DNA mutations occur constantly, cells have developed different kind of systems to maintain the correctness of the genome i.e. correcting mutations. These methods are crucial for the cell to overcome deleterious mutations and survive.

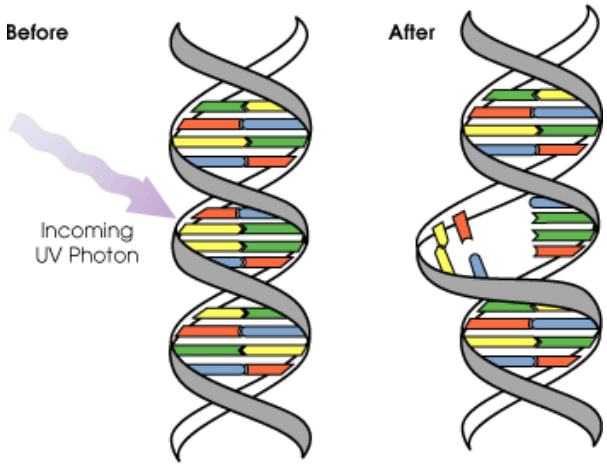



**Figure 19:** DNA mutations can be caused by environmental factors, for example UV-light.

Two important terms that describe an organism carrying a mutation are **genotype** and **phenotype**. Genotype is used to describe the mutation and the gene it occurs in. Phenotype describes the effect of mutation to the organism. An organism that has the usual phenotype is called a **wild** type and an organism whose phenotype has changed as a result of a mutation is called a **mutant**. Mutations occur in two forms: **point mutations** which involve a change in the base present at any position in a gene and **gross mutation** which involve alterations of a longer stretches of DNA.

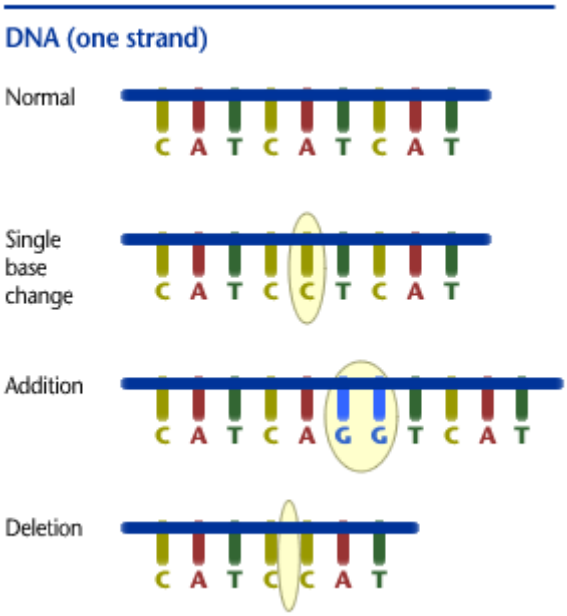



**Figure 20:** Different kinds of point mutations.

Point mutations fall into a number of categories, each with different consequence for the protein encoded by the gene. **Missense** mutations involve the alteration of a single base, which change a codon in the way that the amino acid that it codes is altered. **Nonsense mutations** change the codon in a way that it is altered to a termination codon and this way has an effect to the protein's size, form and functionality. **Frame shift** mutation result from the insertion of extra bases or deletion of existing bases from the DNA sequence of a gene. This leads to alteration of a reading frame and building of a mutant protein. Mutation is called **silent** if it occurs at the third base of a codon and the amino acid encoded by that codon is not changed. Silent mutations do not have effect on the encoded protein and do not result a mutant phenotype. They tend to accumulate in the DNA of organism where they are known as **polymorphisms.**

Gross mutations involve major alterations in the DNA and have serious effects on the encoded proteins and thus, a mutant phenotype. Gross mutations fall into three subcategories. **Deletions** involve a loss of portion of the DNA sequence. **Insertions** are additions of extra bases, usually from another part of chromosome. **Rearrangements** involve mutations caused by DNA sequences changing positions with each other.

Among humans and other higher animals the DNA mutations play significant role in development of diseases. Inherited diseases are caused by mutations that are passed on from parents to offspring. Mutations can occur at random in any cell, but normally, a mutation in a single cell has no effect on the organism because the cells are continually being replaced. However, a mutation in a germ cell (sperm or ovum) can be passed after the conception and will be present in every cell the resulting offspring and therefore cause a disease. Most genetic diseases are caused by mutations in single gene, which may vary a lot. If the disease is caused by mutation in one gene, disease is said to be **monogenic**. In contrary to the monogenic diseases, **multifactorial** diseases are caused by co-operative action of different mutations in different genes and environmental factors.

**Fifty Years of DNA and the Hugo Project**

Little more than fifty years ago, in April 1953, Francis Crick and James Watson wrote a letter to the *Nature* magazine, in which they proposed that they had invented a structure, "which has novel features which are of considerable biological interest". They were absolutely right, because the structure they were representing was the molecular structure of DNA. This momentous discovery, which was culmination of many researchers, was one of the most significant landmarks of 20$^{th}$ century science.

The genetic code was cracked at 1960, when a group of researchers figured out that three letter of DNA encodes particular amino acid. A three letter word made out of four letters could have more than enough permutations to encode the 20 amino acids. The final breakthrough for biotechnology happened in the 1970s due to inventing some ingenious tricks for manipulating DNA. Means for cutting and pasting of DNA were invented and the cloning was discovered. Cloning means piecing together fragments of DNA from different species and moving them to another animal, for example to bacteria, where they could be copied as a part of bacteria's own genetic material. In 1983 PCR-technique (polymerase chain reaction) was invented, which enabled making millions of copies of specific DNA segments in test tubes. These DNA techniques are most essential for the modern genetic engineering.

The developing techniques have been coupled with fast growing amount of information. One of the major goals of the biology has been the sequencing the human genome. The rate of this process has been incredible. In the 1970s scientists were figuring out how the sequencing should be done and only 30 years later the first draft version of the human genome was revealed. In August 2003 the international Human Genome Sequencing Consortium reached their summit, unveiling a complete respectable version of the human genome. By doing this they finished the

HUGO project, which has been compared to the man's first landing to the moon. Finishing the human genome project has also been seen as a beginning for a new century of biology.

Even though the whole sequence of human and many other organisms are known, there still remains a lot more to find out. The next goal is to find out which genes produce which proteins with the tools of proteomics. Also much more needs to be found out about the regulation of gene expression. The results of the human genome project will also have a lot to offer to the future of medicine.

Figures:
1. http://research.amnh.org/biodiversity/center/images/biod.gif
2. http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowSection&rid=mcb.figgrp.204
3. http://training.seer.cancer.gov/module_anatomy/images/illu_cell_structure.jpg
4. http://www.writing.ucsb.edu/faculty/samuels/dna.jpg
5. http://jan.ucc.nau.edu/~lrm22/lessons/dna_notes/dna_structure.gif
6. http://user.uniserve.com/~ghatton/dna2.gif
7. http://www.irn.pdx.edu/~newmanl/GeneticCode.GIF
8. http://www.gensuisse.ch/school/images/gene/gene.gif
9. http://ww2.mcgill.ca/biology/undergra/c200a/f09-35.jpg
18. http://anatomy.med.unsw.edu.au/cbl/embryo/DNA/images/hIdiotGram.gif
19. http://earthobservatory.nasa.gov/Library/UVB/Images/dna_mutation.gif
20. http://press2.nci.nih.gov/sciencebehind/cancer/images/cancer42.gif
http://www.rri.kyoto-u.ac.jp/ReDi/rls_2.jpg
http://www.csiro.au/helix/dna/index.html (inheritance)
http://genetics.gsk.com/chromosomes.htm (mitosis and meiosis)
http://www.biology.iupui.edu/biocourses/N100/2k2humancsomaldisorders.html (karyotypes)
http://www.anselm.edu/homepage/jpitocch/genbio/geneticsnot.html (homologous chromosomes)
http://129.128.91.75/de/genetics/70gen-inherit.html (inheritance patterns)

## Polymerase Chain Reaction

Polymerase chain reaction (PCR) is a method that is used for the amplification of a target DNA-fragment. PCR utilizes similar molecules that nature uses for copying DNA prior to cell division: an enzyme called DNA polymerase, which catalyzes the reaction, deoxynucleoside triphosphates (dNTPs; nucleotides), which are the building blocks of DNA and two oligonucleotide primers, short synthetic strands of DNA that define the beginning and end of the DNA stretch to be copied. A DNA polymerase enzyme walks along the segment of DNA, reading its code and creating a copy by assembling nucleotides in the sequence defined order.

There are three major steps involved in PCR, denaturation, annealing, and extension, each of them taking place at a different temperature. The cycles are done on an automated cycler, which rapidly heats and cools the test tubes containing the reaction mixture. The cycles are repeated for 25 to 40 times. During **denaturation** step at 94°C, the double-stranded DNA melts and opens into single-stranded (template) DNA. In **annealing**, which is performed at the range of 45-60°C, primers anneal to specific, nucleotide-sequence-defined region of the single-stranded template DNA. During the **extension** at 72°C, the DNA polymerase enzyme catalyzes the reaction, in which the annealed primers are extended by the incorporation of nucleotides in sequence-defined order. As a result, a double-stranded DNA copy of the original target is produced. Since the product of one PCR cycle becomes additional template for the next cycle, there is an exponential increase of the number of copies. The PCR process can therefore give rise to millions of copies of the target sequence. See the schematic picture of PCR process on the next page.

PCR technique has revolutionized the world of molecular biology. Kary Mullis invented PCR in 1983 and was 10 years later awarded the Nobel Prize in Chemistry for his work. PCR is commonly used in medical and biological research labs for a variety of tasks, such as the detection of hereditary diseases, DNA sequencing, forensic science, evolutionary studies and population genetics. Due to its simplicity, sensitivity, specificity, and reliability, PCR is actually an integral part of almost every genome study.

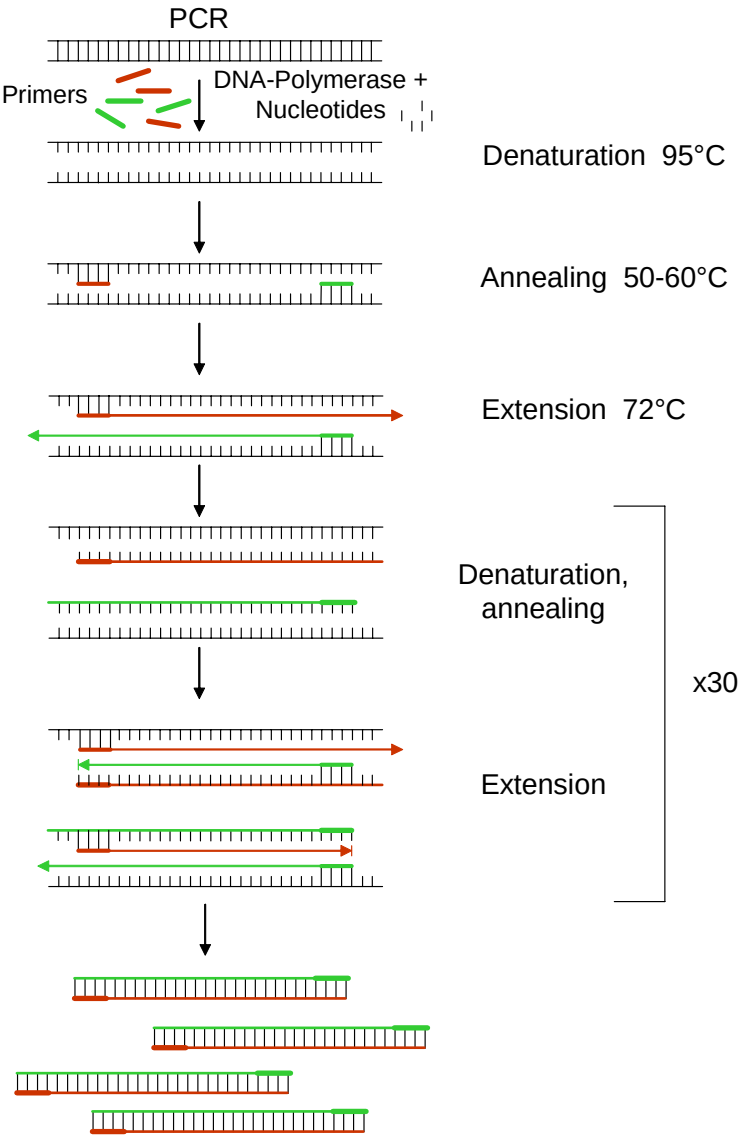



Figure 1. The PCR cycles.

## Restriction enzymes

Restriction Enzymes, also called restriction endonucleases, are enzymes that cut DNA. They are found and purified from bacteria. To protect themselves, many types of bacteria have developed a method to destroy any foreign or mutated DNA. These bacteria have built a special restriction-modification system, which consists of DNA cutting restriction enzymes and modifying methylase enzymes. Bacteria's own DNA is protected from digestion by methylations. There are three types of restriction endonucleases. With types I and III there is no strict control over the sequence that is recognized by the enzyme. The type II enzymes are more useful for recombinant DNA technology, since they recognize specific, usually palindrome sequences of DNA and cut the sugar-phosphate backbone of the DNA in the region of the recognized sequence. This is called enzyme digestion. Some type II restriction enzymes make staggered, also called "sticky", cuts in the opposite strands creating complementary, single stranded ends; other restriction enzymes make a cut across both strands creating DNA fragments with "blunt" ends.

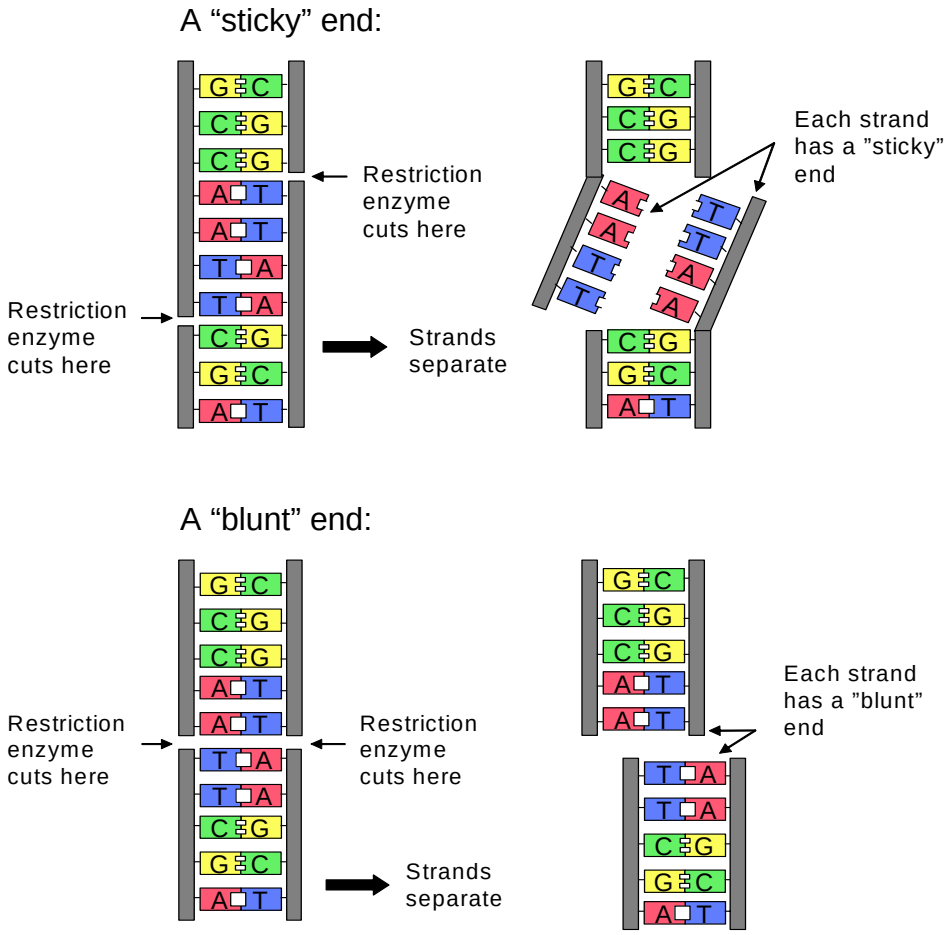



Figure 2. Cuts provided by restriction enzymes.

Today there are over 500 different restriction enzymes commercially available. They have become essential tools for recombinant DNA technology.

Birren, B, Green ED, Klapholz S, Myers, RM and Roskams J.Genome Analysis, A laboratory Manual, Volume 1, Analyzing DNA. Cold Spring Harbor Laboratory Press, 1997

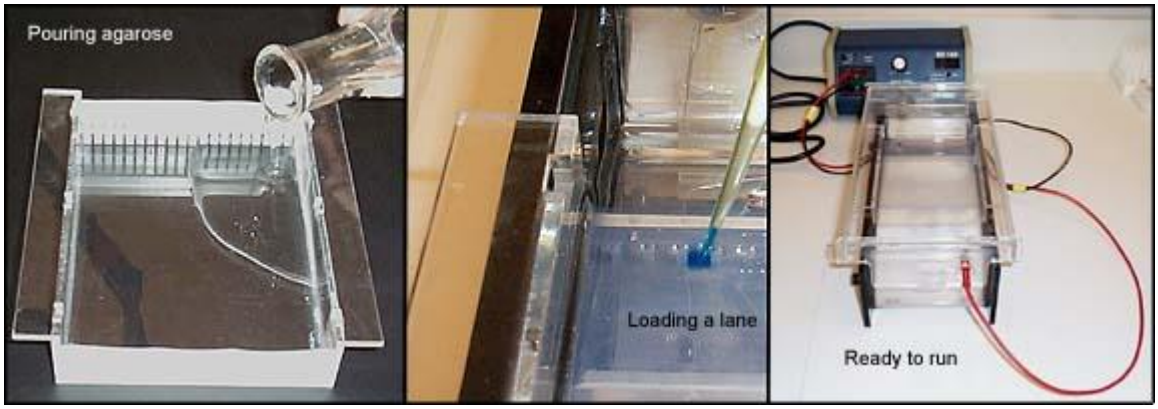
Newton, CR. PCR Essential data. John Wiley & Sons. 1995

## Gel electrophoresis

Gel electrophoresis is a method that separates macromolecules-either nucleic acids or proteins-on the basis of size, electric charge, and other physical properties.
A gel is a colloid in a solid form. The term electrophoresis describes the migration of charged particle under the influence of an electric field. *Electro* refers to the energy of electricity. *Phoresis*, from the Greek verb *phoros*, means "to carry across." Thus, gel electrophoresis refers to the technique in which molecules are forced across a span of gel, motivated by an electrical current. Activated electrodes at either end of the gel provide the driving force. A molecule's properties determine how rapidly an electric field can move the molecule through a gelatinous medium.
Many important biological molecules such as amino acids, peptides, proteins, nucleotides, and nucleic acids, posses ionisable groups and, therefore, at any given pH, exist in solution as electically charged species either as cations (+) or anions (-). Depending on the nature of the net charge, the charged particles will migrate either to the cathode or to the anode.

**How does this technique work?**
Gel electrophoresis is a technique used for the separation of nucleic acids and proteins. Separation of large (macro) molecules depends upon two forces: charge and mass. When a biological sample, such as proteins or DNA, is mixed in a buffer solution and applied to a gel, these two forces act together. The electrical current from one electrode repels the molecules while the other electrode simultaneously attracts the molecules. The frictional force of the gel material acts as a "molecular sieve," separating the molecules by size. During electrophoresis, macromolecules are forced to move through the pores when the electrical current is applied. Their rate of migration through the electric field depends on the strength of the field, size and shape of the molecules, relative hydrophobicity of the samples, and on the ionic strength and temperature of the buffer in which the molecules are moving. After staining, the separated macromolecules in each lane can be seen in a series of bands spread from one end of the gel to the other.

**Electrophoresis of Nucleic Acids**
Gel electrophoresis is the process by which scientists can sort pieces of DNA cut with restriction enzymes by size. An agarose or polyacrylamide gel is loaded with the DNA fragments and current is passed through the gel. Since DNA is negatively charged, it will migrate towards the positive pole. The DNA will not migrate at the same rate, however. Larger pieces of DNA collide with the gel matrix more often and are slowed down, while smaller pieces of DNA move through more quickly. Since different genes have different nucleotide sequences, restriction enzymes will cut them at different places, generating different size DNA fragments. By using gel electrophoresis, biologists can tell which gene is which based upon the sizes of the fragments generated when a gene is treated with a restriction enzyme.



## Gene Expression

The term **gene expression** commonly refers to the entire process whereby the information encoded in a particular gene is decided into a particular protein. Gene expression is the process by which the genetic information of the genes is made available to the cell. During gene expression, DNA molecules copy their information by directing the synthesis of a **messenger RNA** molecule (mRNA), coding the complementary sequence in a process called **transcription**. The mRNA then directs the synthesis of a protein whose amino acid sequence is determined by the base sequence of the RNA. This process is known as **translation**.

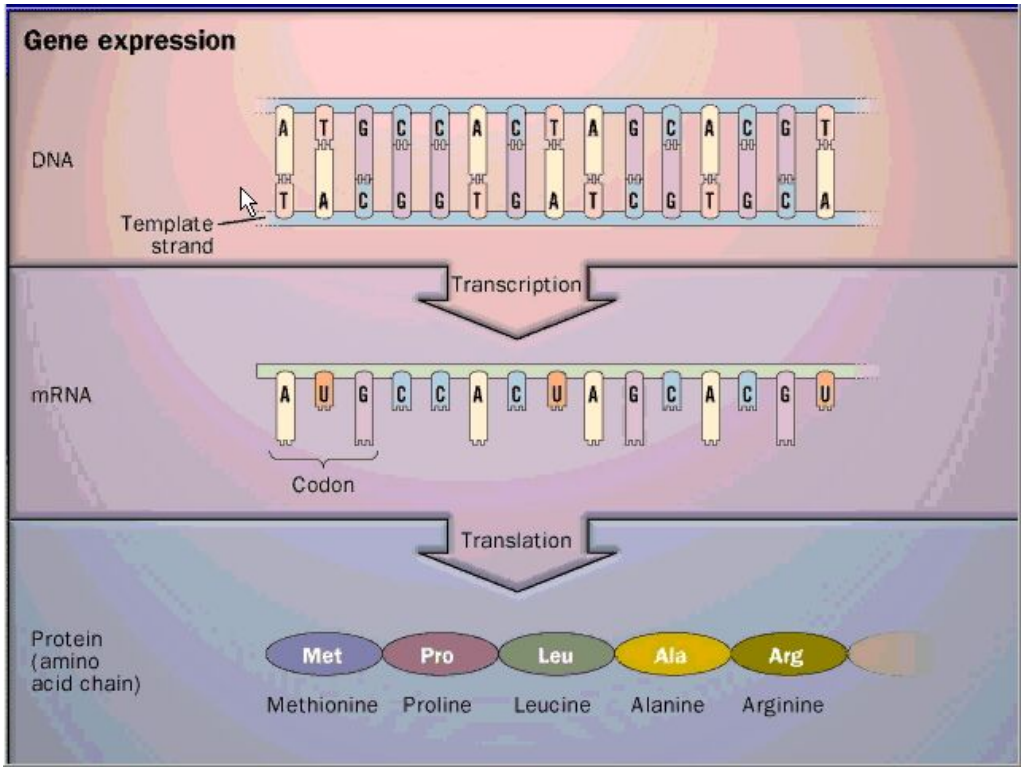



**Figure 1:** The DNA sequence is converted to protein sequence in a multi-phased reaction.

In the transcription, an mRNA copy of the DNA sequence is produced. The RNA is synthesized by special cellular enzymes (protein molecules that catalyses chemical reactions of other substances without being destroyed in the reaction) called **RNA polymerases**. mRNA is produced using the template strand of the DNA as a template and the molecule synthesized is called a **transcript**. Transcript's sequence is complementary to the DNA sequence and the thymine bases (T) of the DNA are replaced with urasile bases (U). In eukaryotes, the initial mRNA transcript is processed by addition of some extra adenosines (A) to the tail part of mRNA molecule, and by splicing, which removes non-coding DNA sequences, introns, from the preliminary mRNA. These processes are needed to produce a functional mRNA molecule. Processed mRNA may undergo translation to produce a protein or it may also be used as a RNA molecule. If the mRNA is being used in protein production, the mRNA is transported from its site of synthesis in the nucleus to the cytoplasm where translation occurs.

Decoding of the nucleotide sequence in mRNA into the amino acid sequence of protein depends on **transfer RNAs** (tRNA). All tRNAs have a similar structure, that include acceptor

part that it used for attachment of a specific amino acid in the other end and a part with a three-base anticodon sequence in the other end of the tRNA molecule. The anticodon part of tRNA can form a base-pair structure with its corresponding codon in mRNA. So there is a corresponding tRNA molecule for every mRNA codon. tRNAs are used to carry amino acids to the place of the protein synthesis and they also have important role in recognizing the codons from mRNA.

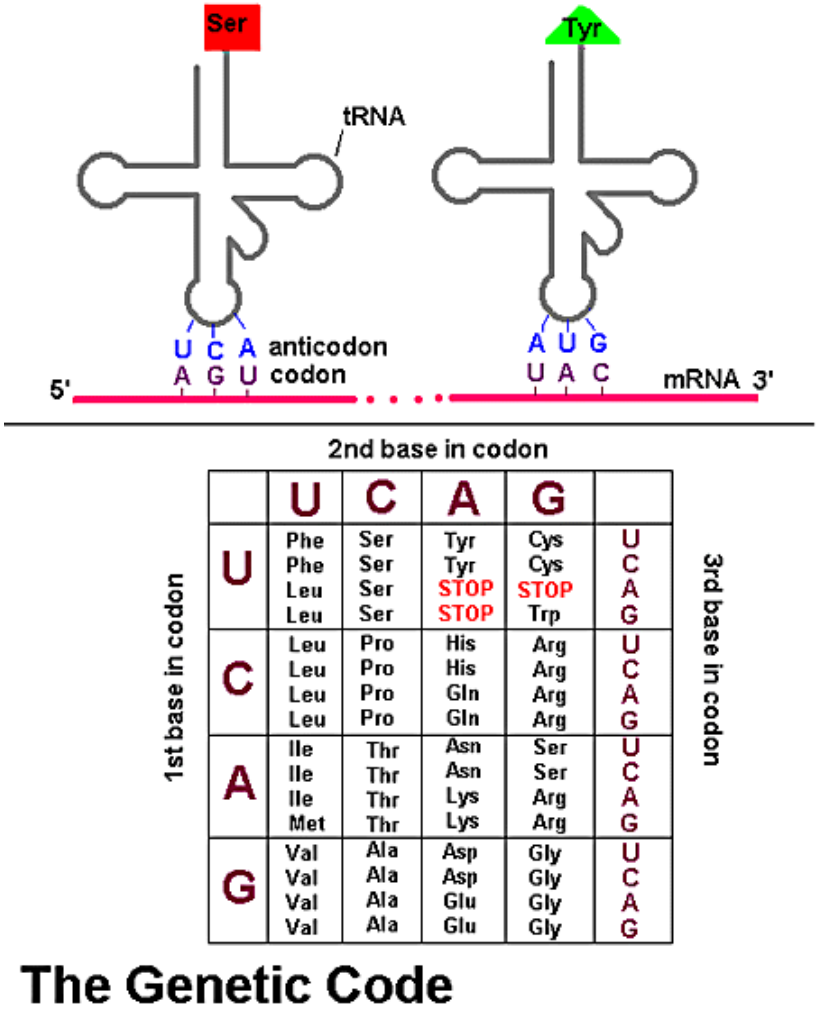



**Figure 2**: tRNA molecules are needed to convert the sequence carried by an mRNA to a protein. All tRNA molecules have similar structure that includes acceptor part with specific amino acid in one end and part with anticodon sequence in other end of the molecule.

Protein synthesis takes place on the ribosomes, which are small particles found in the cytoplasm of the cell. The ribosome recognizes the initiation codon from the mRNA with the help of some specific initiation factors. Elongation of a protein chain entails three steps, which are repeated over and over again. In the first step, the tRNA with corresponding amino acid forms a pair with the corresponding codon in the mRNA. In the second stage a bond is formed between the new amino acid and the growing protein chain. Formation of the bond is accompanied by movement of the tRNA in order to depart from the ribosome. Finally the ribosome translocates to the next codon, and the old tRNA is discharged from the ribosome. These steps are repeated until the ribosome arrives to the site of stop codon. The termination is carried out by the help of termination factors, which recognize the stop codon and prevent the ribosome to go on. Following the termination, the nascent protein is released from the ribosome. The ribosome dissociates and the mRNA is released.

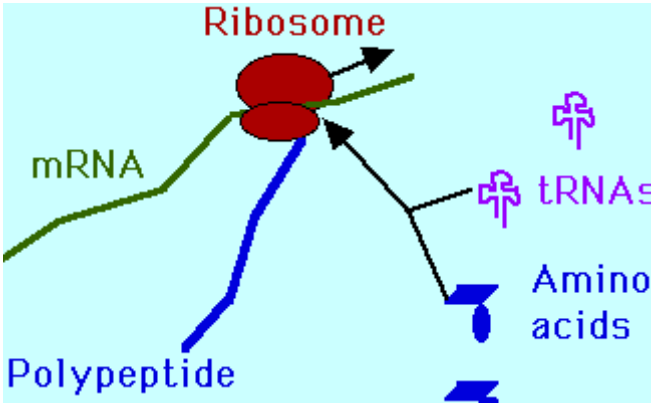



**Figure 3:** Protein synthesis takes place on ribosomes where amino acids carried by tRNAs are joined together to form a polypeptide chain according to the instructions written in mRNA.

Following translation, proteins may be modified by the addition of chemical groups or by cleavage of the proteins. Modification may be required for full functional activity of the protein. Before the protein is activated it needs to be folded correctly to achieve the correct three-dimensional structure. Correct 3D structure is crucial for protein's convenient activity.

**Figure 4:** Correct folding of proteins is needed to achieve correct functional properties. Functional proteins are composed of different kinds of structural elements.

## Expression Control

One of the underlying principles of molecular cell biology is that the actions and properties of each cell type are determined by the proteins it contains. But what determines the types and amounts of the various proteins that characterize a particular cell type? The determining factors are the **concentration** of each protein's corresponding RNA, the **frequency** at which the mRNA is translated to the protein, and the **stability** of the protein itself. The concentration of various mRNAs is determined largely by which genes are transcribed and their rate of transcription in a particular cell type. Thus the differential transcription of different genes largely determines the action and properties of cells. Not all the genes that are present in a cell are expressed and different genes are expressed during different stages of cells life. Different genes are also expressed in different kind of cells. The overall combination of genes that are active in a cell

determines the characteristics of a cell and its function within the organism. For example, different sets of genes are active in muscle cells when compared, for example, to nerve cells.

Theoretically, regulation at any one of the various steps in transcription and translation could lead to differential gene expression in different cell types or developmental stages or in response to external conditions. Although examples of regulation at each of the steps in gene expression have been found, control of the **transcription initiation** – the first step – is the most important mechanism for determining whether or not most genes are expressed and how much of the encoded mRNAs, and consequently proteins, are produced. In eucaryotic genomes the control elements controlling the initiation of transcription can be located right next to the gene but they can also be located to many kilobases away from the start site. Most important control element is the **promoter**, which determines the site of transcription initiation and directs binding of the enzyme needed to produce mRNA. Beside the actual control elements, other control elements, enhancers, have also been found. All control elements are needed for efficient activation of transcription.

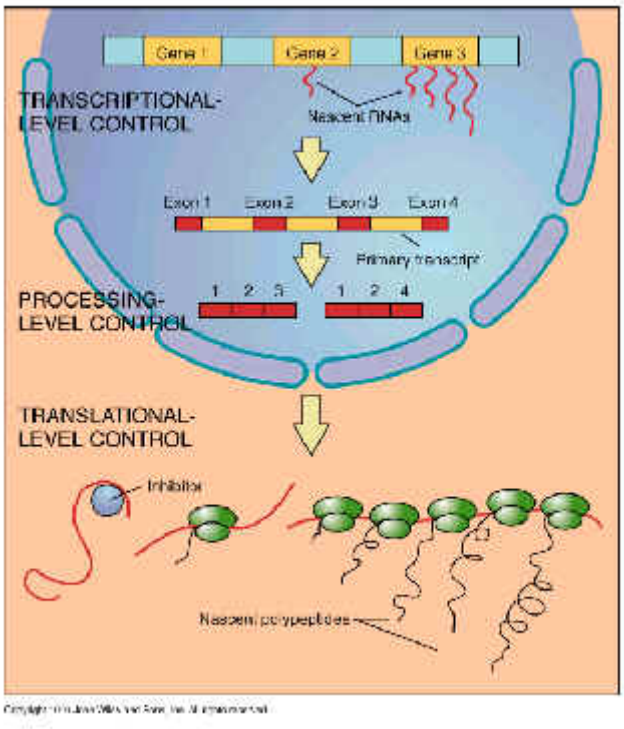



**Figure 5:** Gene expression is carefully controlled during all the steps of the expression.

## Studying the Gene Expression

It is widely known that the products of thousands of genes in a living organism cooperate in a complicated way creating a functional system. Large number of genes has made the study of gene expression difficult. Traditional methods in molecular biology generally work on a "one gene in one experiment" basis, which means that the throughput is very limited and the "whole figure" of a gene expression is hard to obtain. In the past years, a new technology, called **DNA microarrays**, has been introduced. It makes possible to monitor simultaneously the activities of thousands of genes. This is a great advantage, because monitoring gene expression lies at the heart of a wide variety of medical and biological research projects, including classifying diseases, understanding basic biological processes and identifying new drug targets.

An array is an orderly arrangement of samples. It provides a medium for matching known and unknown DNA samples based on **base-pairing rules** and automating the process of identifying

the unknowns. In general, arrays are described as macroarrays or microarrays, the difference being the size of the samples spots. Microarrays contain usually thousands of spots for different genes. DNA microarrays are fabricated by high-speed robots, generally on glass or nylon base, on which **probes with known identities** are attached. There are two variants of the DNA microarray technology in terms of probes. Probes, which are attached to the membrane can be **cDNA-probes**, composed of 500-5000 bases each, or **oligonucleotides**, consisting of 10-80 nucleotides. The probes may be an assortment of all the known genes of a the study organism, e.g. mouse, or all known genes expressed by special tumor cells.

The basic concept behind the use of DNA arrays for gene expression study is simple. First appropriate gene chip for the study is chosen. Then RNA is isolated from the cells under study and converted to cDNA, which is appropriate form of nucleic acid for the study. Samples are labeled with a special chemical label and they are let to form a hybrid with nucleic acid probes attached to the solid support. By monitoring the amount of labeled sample associated with each DNA location, it is possible to calculate the abundance of each mRNA species represented in the sample. The identities of the DNAs attached to the solid support are known, so the identification of the sample mRNAs attached to the probes is possible. Due to the huge amount of information gained with microarrays efficient bioinformatics is needed to be able to understand the meaning of the results.

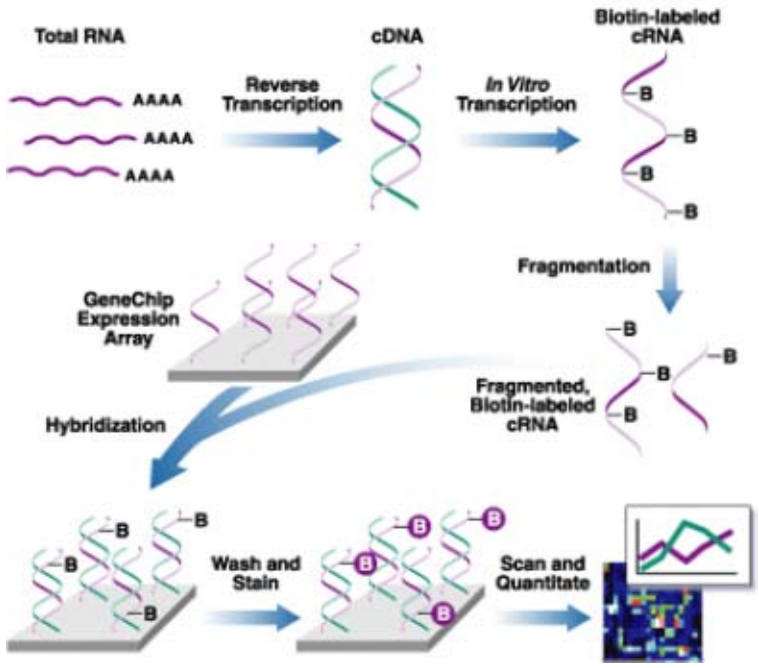



**Figure 6:** Schematic figure of affymetrix eukaryotic gene expression assay.

1. http://www.mdx.ac.uk/www/lifesciences/alex/images/dna_prot.jpg
2. http://www-stat.stanford.edu/~susan/courses/s166/genetic.gif
3. http://www.mrothery.co.uk/images/translation.gif
4. http://pubs.acs.org/cen/coverstory/7839/7839sci13xa.ce.gif
5. http://www.usd.edu/~jfallan/risticb413sp2001/lec13.1.jpg
6. http://www.affymetrix.com/technology/ge_analysis/index.affx

# Overview of DNA microarray data analysis

The DNA microarray data analysis starts from the results of scanned images. At this point, images have been evaluated, bad spots have been investigated, and the spots have preferably been scored with flags indicating whether the spot was good, bad, or borderline. This is crucial, because in the later stages of the analysis the visual inspection of individual spots is not possible. A typical result file from image analysis is depicted in Figure I.

**Figure I.** A result file for one scanned DNA microarray. Every row contains information for one gene, which is unambiquously defined by Probe Set Name. For every gene, amount af labeled sample bound to the oligonucleotide probe is recorded in the column Signal. Every signal value is associated with a quality value (Detection) and quality p-value (detection p-value).

| | Probe Set Name | Signal | Detection | Detection p-value |
|---|---|---|---|---|
| 1 | Pae_16SrRNA_s_at | 32.4 | A | 0.749276 |
| 2 | Pae_23SrRNA_s_at | 4.6 | A | 0.660442 |
| 3 | PA1178_oprH_at | 78.4 | M | 0.05447 |
| 4 | PA1816_dnaQ_at | 3.6 | A | 0.846089 |
| 5 | PA3183_zwf_at | 7.7 | A | 0.908831 |
| 6 | PA3640_dnaE_at | 2.3 | A | 0.990699 |
| 7 | PA4407_ftsZ_at | 26.2 | A | 0.562335 |
| 8 | Pae_16SrRNA_s_st | 6.6 | A | 0.824011 |
| 9 | Pae_23SrRNA_s_st | 64.2 | A | 0.204022 |
| 10 | PA1178_oprH_st | 49.8 | P | 0.011455 |

Next steps in the analysis are preprocessing, normalization, and quality control. The goal of these analyses is to organize results in a meaningful fashion, and to remove variation which is attributable to systematic errors. Normalization also makes the signal values from different chips directly comparable. This is a minimal prerequisite for successful data analysis.

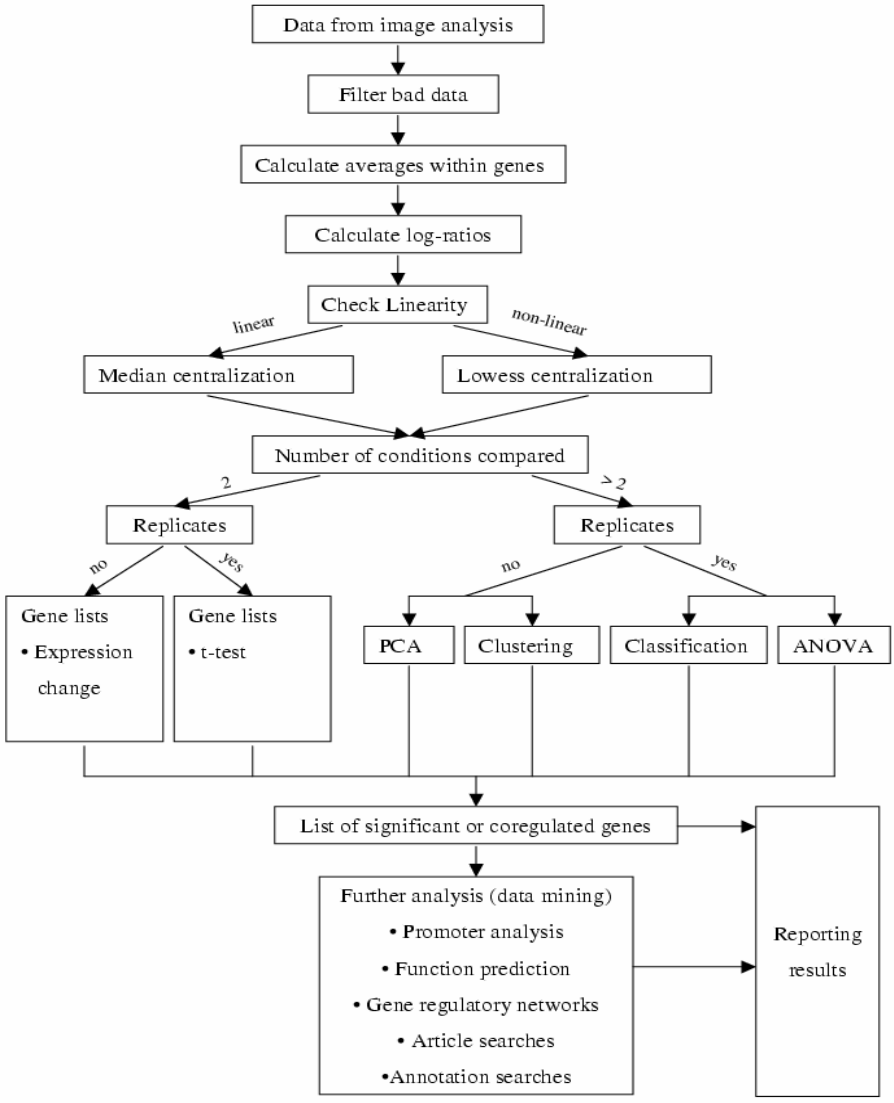
In the further analysis steps, statistically significant, quality-checked data is separated from not interesting and not-trustworthy data. The next step is to find the differentially expressed genes using statistical tools, or to group the good quality data (usually only a small fraction of the original raw data) into meaningful clusters by, e.g., clustering. The goal of clustering is to find similarly behaving genes or patterns related to time scale, time point, developmental phase or treatment of the sample.

After clustering, we may be only at the very beginning of the challenging part of the data analysis, of course, depending on what we are looking for. Next we need to link the observations to biological data, to regulation of genes, and to annotations of functions and biological processes.

With an enormous amount of data, we need standardized systems and tools for data management in order to publish the results in a proper and sound way, as well as to be able to benefit from other publicly available gene expression data. One established standard for publishing DNA microarray results is MIAME standard. The European database (using MIAME) for DNA microarray results is Array Express.

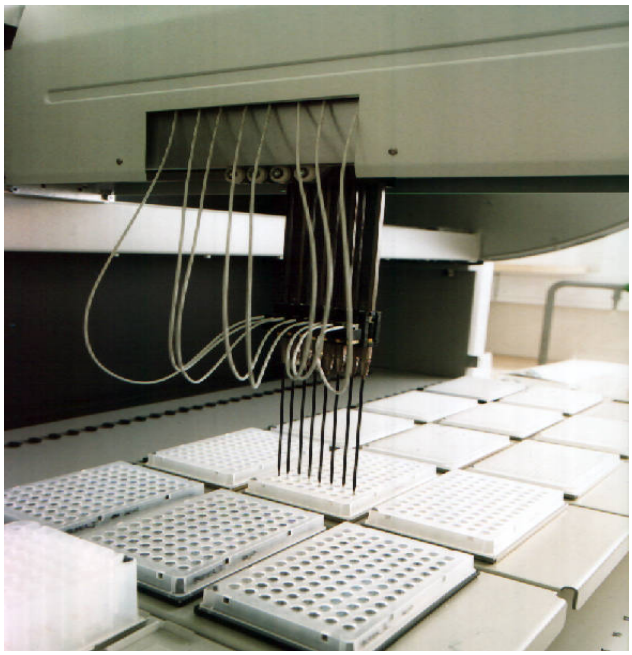
To get an overview of the exploratory data analysis pipeline, consult Figure II. The flow chart also helps to choose the right method for the situation. The flow chart should be read flexibly, though. For example, there can be several filtering steps instead of just one shown on the chart. Or, when there are more than two conditions in the experiment, the data can be analyzed using two conditions route. Note that all possible orders of analysis have not been shown for clarity, and the schema is only meant to be of help in the analysis process, not to be taken as an absolute truth.

**Figure II.** Overview of data analysis methods. Note that all possible orders of analysis are not shown.




This is a modified excerpt from the book DNA Microarray Data Analysis published by CSC. The complete book is freely available at http://www.csc.fi/oppaat/siru/.

# High-throughput genotyping



Genetic differences between individuals have a major role in identification of susceptibility genes for diseases as well as determining individual drug response. Therefore technologies that enable the identification and genotyping a vast number of genetic polymorphisms in large populations have increased rapidly in recent years. Automated, high-throughput methods have been developed mainly to score **microsatellite** markers and single nucleotide polymorphisms (**SNP**s).

Microsatellites are highly polymorphic, short repetitive sequences of two-, three or four bases (di-, tri- and tetra nucleotides, respectively). The number of repeats varies among normal individuals, from a few up to 30 or so. Microsatellites are distributed evenly throughout the genome, which makes them useful markers in search of disease-predisposing genes. Microsatellites are also called as STR (simple tandem repeat), SSR (simple sequence repeat) and SSLP (simple sequence length polymorphism). These alleles are easy to detect by polymerase chain reaction (PCR). PCR products are run on an electrophoretic gel or capillaries and identified based on the length of the fragment. Even fragments differing by only one repetitive unit (CA for example) can be distinguished from each other. To increase the throughput of microsatellite genotyping, automated instrumentation for PCR setup, liquid handling, electrophoresis and data processing and storage is needed.

SNPs (single nucleotide polymorphisms) are single base variations (e.g. AAGC**C**TAA → AAGC**T**TAA), which occur approximately every 1000 basepairs along the human genome.
SNPs are the most simple and common form of genetic polymorphism, comprising 90% of all human DNA polymorphism. Most SNPs do not have effect on cell function, but others are believed to predispose people to disease, or to influence the individual's response to a drug. Even SNPs that are located in non-coding regions of a genome are highly important for genetic studies, since they can be used as genetic markers. There is enormous interest in SNPs at the moment; highly because it is believed that particularly SNPs could be useful markers in association studies for multifactorial diseases. Compared to microsatellites they are highly abundant and genetically more stable.

Various methodologies exist for detecting the alternative alleles of SNPs, each having its draws and backs. The ideal SNP genotyping system would offer high throughput capacity, ease of assay design, robustness, affordable price, automated genotype calling, and most importantly, accurate and reliable results. Common to nearly all SNP genotyping methods is that they include a PCR step to achieve sensitivity and specificity needed in genotyping. Further detection to discriminate between alleles is based on e.g. allele-specific hybridization, allele-specific primer extension, allele-specific oligonucleotide ligation, or allele-specific enzymatic cleavage. There are also several options to detect the allelic discrimination obtained with above mentioned chemistries; monitoring the light emitted by the products (chemiluminescence, fluorescence, fluorescence resonance energy transfer, fluorescence polarization), by measuring the mass of the products (mass spectrometry), or by detecting a change in the electrical property when the products are formed. The requirement of a PCR amplification step is the principal factor that limits the throughput of SNP genotyping assays today. Multiplex PCR amplification or using pooled DNA samples as a template is one answer to this problem. Also methods that avoids PCR amplification step have been recently developed.

## Finnish Genome Center as an example

Finnish Genome Center (FGC) is a national facility for the genetic research of multifactorial diseases, hosted by the University of Helsinki. FGC collaborates internationally and nationwide with research groups interested in understanding causes of common diseases and provides genotyping service. FGC is a good example of the medium sized genotyping lab.
FGC is a facility for high-throughput genotyping, aiming at a throughput of 10,000 polymerase chain reactions per day. PCR assays are set up on 384-well plates, using semiautomatic pipetting stations and a liquid handling robotic workstation. Either gel or capillary electrophoresis can be used to separate and size the PCR fragments. From the beginning FGC has focused mainly on the genome wide scans using 400 evenly spaced polymorphic microsatellite markers. Today the focus is more and more on microsatellite fine-mapping using custom oligonucleotides, and setting up a capacity to genotype biallelic single nucleotide polymorphisms (SNPs).

FGC uses three different SNP genotyping methods, which are based on RFLP (restriction fragment length polymorphism), single nucleotide primer extension (SNuPe) and primer extension detected with MALDI-TOF mass spectrometry. Each method utilizes PCR amplification of a DNA fragment containing a SNP site. The reaction is carried out in 5 µl volume on 384-well microtiter plates. In the RFLP method, the PCR products are digested with specific restriction enzyme. This traditional method is modified by including a natural or an artificially created additional restriction site in the PCR product to serve as an internal control site for the digestion reaction. Approximately ten different markers are pooled together and the fluorescently labelled fragments are separated based on their sizes with an ABI 377 –DNA Sequencer.

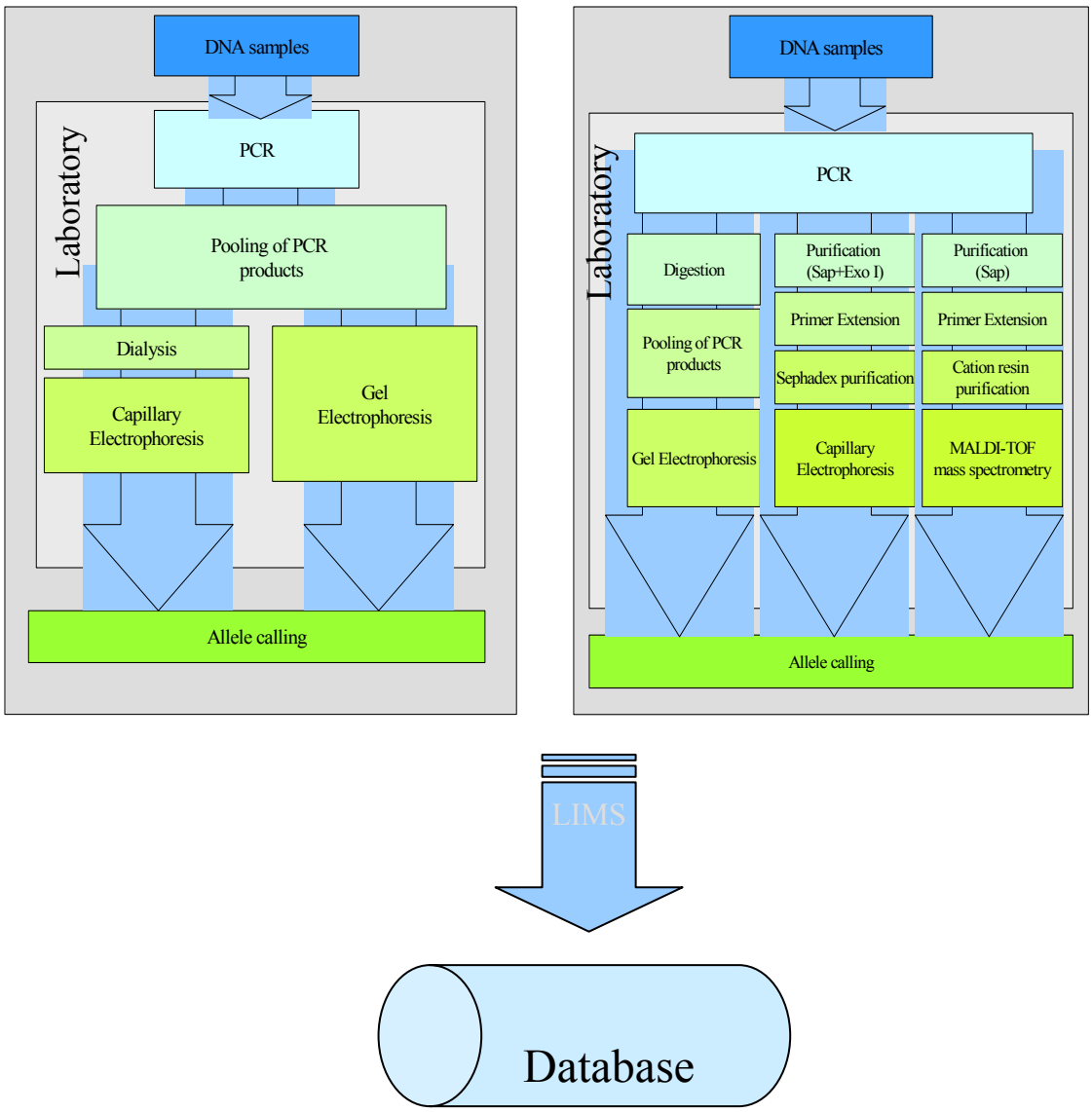
In the SNuPe method an enzymatically purified PCR product is used as a template for the single nucleotide primer extension reaction. The primer extension reaction mixture contains nucleotide terminators (ddATP, ddGTP, ddCTP, ddTTP), each of them labelled with a different fluorescent dye. The specific detection primer is annealed to the target sequence and extended by a single nucleotide terminator (ddNTP) complementary to the polymorphic nucleotide. Unincorporated nucleotides and additional salts are removed by using Sephadex filtration. The extension products are analyzed on a MegaBace 1000 capillary electrophoresis instrument based on the colour of the dye.

As in the previous method, the primer extension method utilizing a MALDI-TOF mass spectrometry uses an enzymatically purified PCR product as a template for the primer extension reaction. During the reaction, the primer is extended by a specific number of nucleotides

depending on the allele and design of the assay. In the reaction mixture, all four nucleotides A, T, C, and G are present as either dNTPs or ddNTPs (for regular SNP assays, usually three nucleotides are present as ddNTPs and one as dNTP). The incorporation of a ddNTP terminates the extension of the primer. Using a DNA polymerase that incorporates both ddNTPs and dNTPs at the same rate, the primer extension reaction produces allele-specific extension products of different masses depending on the sequence analyzed. The purified primer extension products are then spotted on microarray chips and the masses are measured by using the ionization-technique known as **m**atrix-**a**ssisted **l**aser **d**esorption/**i**onization (MALDI) and the time-**o**f-**f**light (TOF) mass spectrometer. The ions are accelerated by high voltage and separated based on the time it takes them to travel at the detector.

A schematic picture of the genotyping workflow at Finnish Genome Center:

**Microsatellite genotyping workflow** **SNP genotyping workflow**




REFERENCES

Kwok, P-Y (2001) Methods for genotyping single nucleotide polymorphisms. *Annu. Rev. Genomics Hum*. Genet. 2:235-258

Syvänen, A-C (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Rev. Genet.* 2: 930-942.

## Genetic association analysis

Search for significant correlations between gene variants and phenotypes is called *(genetic) association analysis*. For example, studying variation of ApoE –gene in a group of healthy people and group of patients affected with Alzheimer's disease would most probably show strong statistical correlation between allele 4 and the disease phenotype. It is a well-known fact that this allele confers increased risk for the disease. Usually, the researcher has obtained genotype information on one or several earlierly specified candidate genes, and tries find candidate gene variant(s) overrepresented in affected individuals. The correlation may be searched with any statistical approaches, but among the genetics community a popular approach is the very simple, traditional $\chi^2$–homogeneity test.

**Example 1:** Locus A is candidate gene for *Systemic Lupus Erythematosus* (SLE). There are two alleles at the locus. In a sample of 200 individuals (100 cases + 100 controls) there were

| | Affected | Unaffected |
|---|---|---|
| Allele 1 | 79 | 46 |
| Allele 2 | 21 | 54 |

- Allele 1 seems associated, based on the differences in distributions in affected and healthy individuals
- A statistical test is carried out to quantify the significance of the finding:

$\chi^2$- homogeneity test

| **Observed** | Affected | Healthy | Σ |
|---|---|---|---|
| Allele 1 | 79 | 46 | 125 |
| Allele 2 | 21 | 54 | 75 |
| Σ | 100 | 100 | 200 |

| **Expected** | Affected | Healthy | Σ |
|---|---|---|---|
| Allele 1 | 62.5 | 62.5 | 125 |
| Allele 2 | 37.5 | 37.5 | 75 |
| Σ | 100 | 100 | 200 |

The idea is to compare the observed frequencies (in the left table) to frequencies expected under **null hypothesis of no association between alleles and the occurrence of the disease (independency).** Thus, the expected frequencies are simply calculated as product of marginal frequencies form the Observed table.

The test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Where $o_{ij}$ is observed class frequency in the cell at ith row and jth column, $e_{ij}$ expected (under null hypothesis of no association), and $k$ is the number of classes in the table. Degrees of freedom for the test is: df=(r-1)(s-1), r=number of rows in the table, s=number of columns.

Here,

$$\chi^2 = \sum_{i,j} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \frac{(79-62.5)^2}{62.5} + \frac{(46-62.5)^2}{62.5}$$

$$+ \frac{(21-37.5)^2}{37.5} + \frac{(54-37.5)^2}{37.5} = 23.23$$

df=1 $p$<<0,001

where $p$-value is obtained from the $\chi^2$–distribution with given numbers of degrees of freedom (here, df=1). **The *p*-value is low enough that the null hypothesis can be rejected (the probability that the observed frequencies would differ this much from null hypothesis frequencies just by coincidence is less than 0.001)**.

In the example, the gene is not a direct cause for the disease, but it affects the *probability* of the disease. The situation can be illustrated like this:

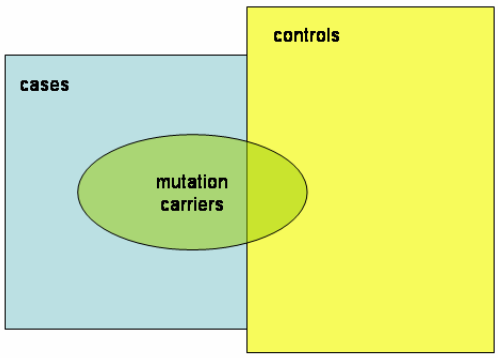



**Figure 1.** The relationship between a set of cases (affected) and a set of controls (unaffected) and disease mutation carrier proportions.

This is the most typical situation in the human disease genetics nowadays, e.g. in the study of common complex diseases such as diabetes or asthma. The goal is to find genes conferring increased risk to disease based on a hypothesis that the risk mutation(s) are relatively common.

Quite often, there are no candidate genes known for a disease, so one has to look first for the genome region of interest. This is carried out by saturating the genome with marker genes (like microsatellites and SNPs), genotyping them in a sample of affected and unaffected individuals, and analyzing the association (or the genetic *linkage*; co-segregation of marker genes and disease phenotype in pedigrees, see the next chapter). The disease gene may lie anywhere in the area or sometimes, does not exist there at all. An example of a first few lines of an SNP data:

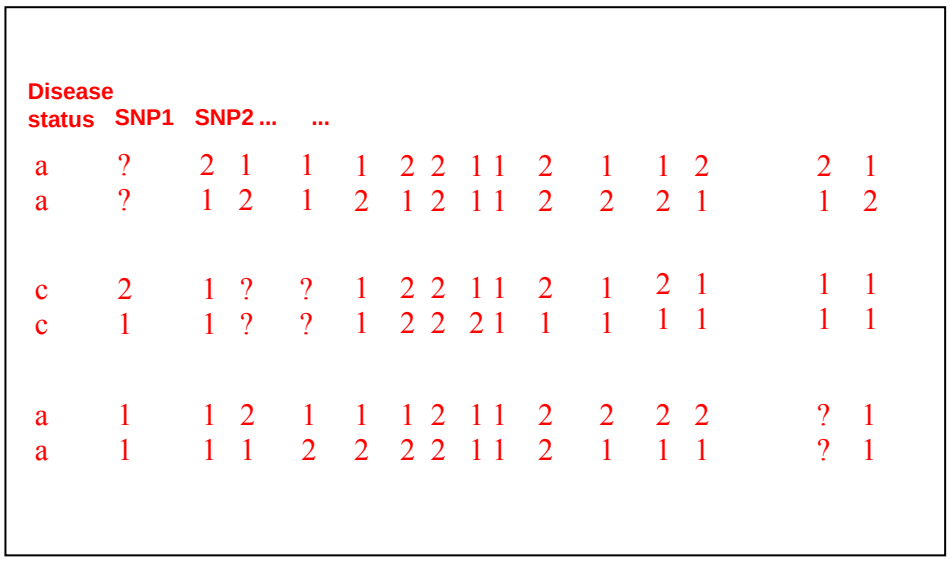



**Figure 2.** First six rows of an SNP data. Question mark denotes for missing alleles.

Now, the task would be to find either an allele or *allele string* (haplotype) which is overrepresented in chromosomes of the affected individuals, and to prove that the difference in frequencies is statistically significant. The data has been ascertained based on phenotype (affected vs unaffected, or for example, from upper and lower ends of quantitative distribution), in order to enrich the number of disease risk conferring mutations in the sample.

- Markers may be microsatellites, SNPs or any varying sites in DNA
- Test with which to measure the association
- *Multiple testing* problem: making one test per each allele in each locus or all 1-, 2- and 3-marker long haplotypes means that a great number of tests will be conducted. Thus, the probability that by chance some of them show "significant" association is high. How to overcome this difficulty? → Permutation testing, empirical p-values

**What does a phenotype-marker correlation then imply?**

- Either, the associating allele / site is itself the disease causing mutation (if we are really lucky), or it is in *linkage disequilibrium (LD)* with it.

## LD process

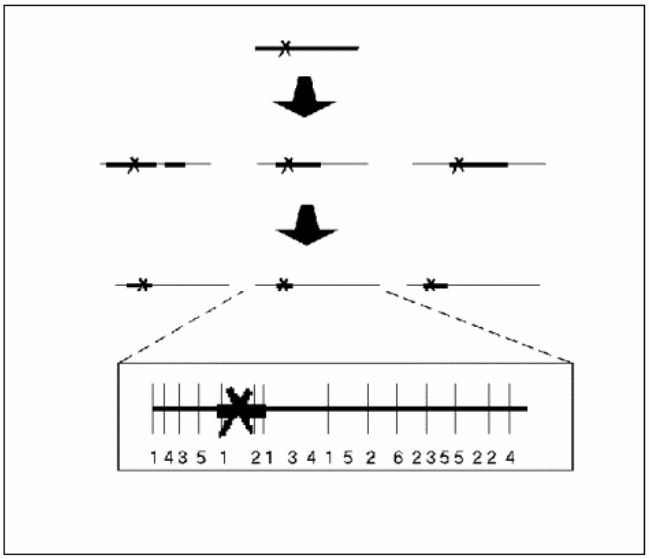
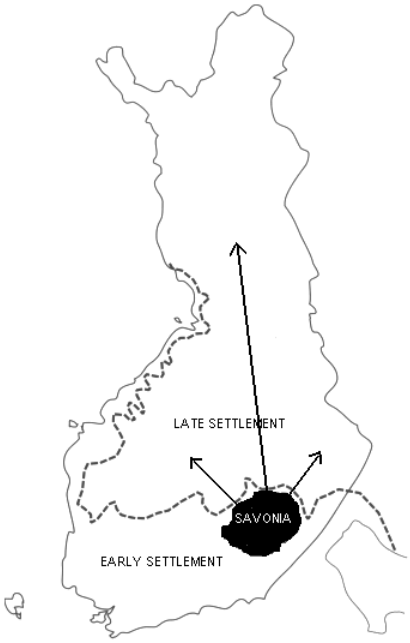



**Figure 3.** A disease mutation born *n* generations ago in one (founder) chromosome has been transmitted to the next generations. Recurrent *recombinations* narrow down the area of the ancestral chromosome hitch-hiking along with the disease mutation from generation to generation. In the present generation, picking up affected individuals carrying the disease gene and genotyping the area with a dense marker map, one would observe short stretch of shared haplotype in several affected individuals (haplotype 1-2-1 in the figure). This phenomenon is also called "founder effect".

If we have found a marker allele or haplotype in LD with the disease, we know we are "reasonably" near to the actual disease locus– and may continue the search with even more densely spaced markers around the marker(s) found and eventually show a functional gene in the area with effect on the molecular pathway affecting the trait in question.



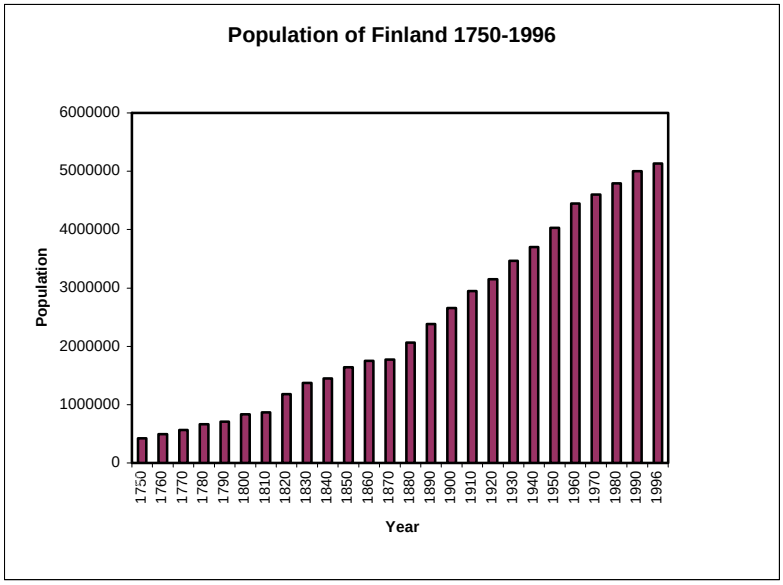



**Figure 5** (previous page and above)**.** Finnish population history. The southern and coastal areas have been inhabited for at least several thousands years ago, whereas the central and northern parts only in 16$^{th}$ century by small emigrant groups from a district in South-East of Finland, Savonia. Adapted from Norio (2000): Suomi-neidon geenit. Population size in Finland has grown from less than half a million in 1750 to 5,1 million in 1996 (adapted from Kere 2001). Thus, for most genetic diseases, the expected number of different founder mutations is small compared to larger and older populations.

## Limitations of the LD mapping

The relationship of the distance between markers with respect to the strength of LD: Theoretically, the amount of LD follows exponential curve as a function of physical distance (**Figure 6**).

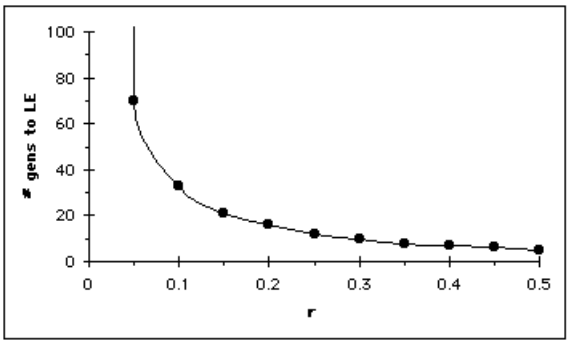



**Figure 6.** Theoretical decay of LD as function of physical distance.

However, the situation often looks more like this:

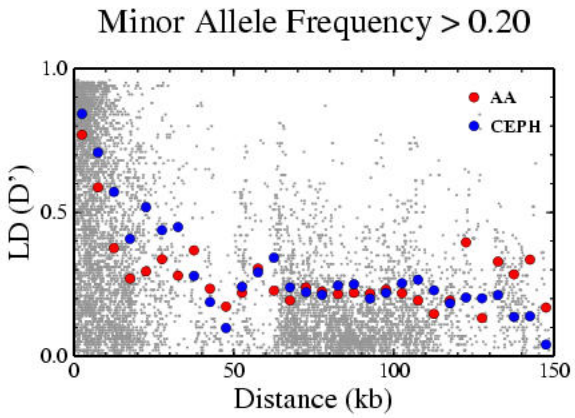



**Figure 7:** Linkage disequilibrium (D') for the African American (red, AA) and European (blue, CEPH) populations binned in 5 kb bins after removing all SNPs with minor allele frequencies less than 20%. 3429 SNPs (Source http://www.fhcrc.org/labs/kruglyak/PGA/pga.html)

This is because the amount of LD is affected by several possible reasons

- Random drift
- Natural selection
- Population structure
- New mutations
- Founder effect

LD is continuous process →
Depending on the relative strength of these forces, the overall level and the distances to which the LD extends will vary a great deal between genetic areas and populations. Thus, there is a limit in what we can observe when searching for associations using LD!

Despite these shortcomings, interesting challenges exist:

- Looking for haplotypes
- Better methods for measurement of association
- Taking into consideration the *disease models*

## Haplotype Pattern Mining (HPM)

Haplotype Pattern Mining (HPM, Toivonen et al, AJHG 67: 133-145, 2000), is a data mining based method, in which trait-associated haplotype patterns, possibly containing gaps, are searched for and the location of a disease gene predicted based on the locations of the associated haplotype patterns. The method was originally designed for LD-based gene mapping in genome-wide level.

**Algorithm:**

- Scan for all recurrent patterns in case/control data (using depth-first search) ⇒ set of patterns which have high enough frequency in the data set (figure 8):

| Marker/ Pattern | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | Frequency |
|---|---|---|---|---|---|---|---|---|---|
| P1 | 4 | * | * | 6 | 3 | * | * | * | 24 |
| P2 | 4 | 1 | * | 6 | 3 | * | * | * | 21 |
| P3 | 4 | 1 | 6 | 6 | 3 | * | * | * | 17 |
| P4 | 4 | 1 | 6 | 6 | 3 | * | 1 | * | 16 |
| P5 | 4 | * | * | 6 | 3 | 6 | 1 | * | 16 |
| P6 | * | * | 6 | 6 | 3 | 6 | 1 | * | 13 |
| P7 | * | 5 | 4 | 3 | * | * | * | * | 12 |
| P8 | 3 | 5 | 4 | 3 | * | * | * | * | 12 |
| P9 | 4 | 1 | 4 | * | * | * | * | * | 11 |

- The pattern *P1* might be observed, for example, in genotype ({4,1}, {2,2}, {5,3},{6,7}, {4,3}, {1,8}, {5,5}, {3,2}).
- The haplotype patterns are evaluated by their strength of association to the phenotype, and markerwise score for each marker is calculated. This is done by measuring the strength of association of each of the patterns by simple $\chi 2$–test, and then scoring, for each marker in turn, all haplotype patterns which overlap the marker and the value of $\chi 2$–test exceeds a pre-specified threshold

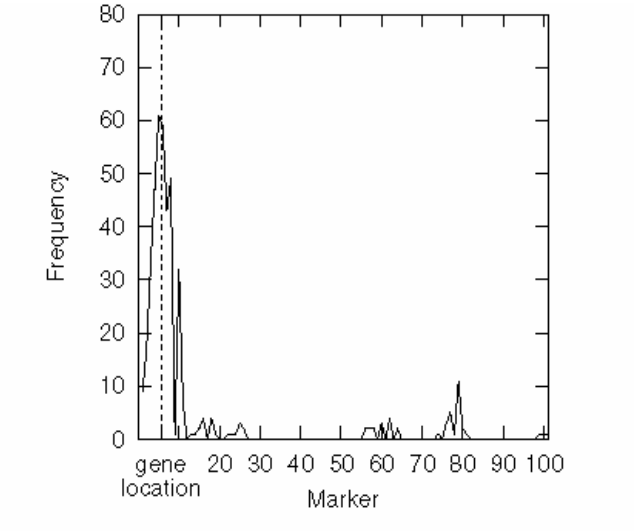



**Figure 9**. An example of a correct localization by HPM. The disease mutation lies at the position shown by vertical line.

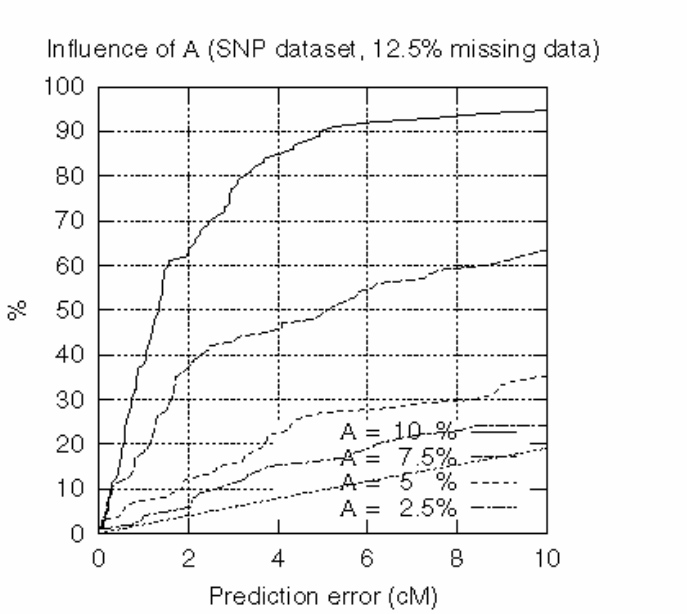



**Figure 10.** Localization power with simulated SNPdata (density 3 SNPs per 1 cM). Isolated population with a 500-year history was simulated. Disease model was monogenic with disease allele frequency varying from 2.5-10 % in the affecteds. 12.5 % of data was missing. Sample size 100 cases and 100 controls.

# Basics of Linkage Analysis

## Idea of linkage analysis

Besides association analysis, *linkage analysis* is one of the two main approaches to disease gene mapping. Linkage analysis utilizes pedigree data, and it often constitutes the first phase of a genome mapping study.

Two loci are said to be *linked* if they are located close to each other in the same chromosome. The purpose of linkage analysis is to find a marker or markers that are linked to the hypothetical disease locus. (From the previous lectures we know that markers are loci with known variation between individuals.)

Recall that crossovers occur in the chromosomes during meiosis (Figure 1). Computational modelling of human meiosis constitutes the very basis for linkage analysis: as a result of the events in meiosis, the chromosomal material transmitted to the offspring can be a combination of the two parental chromosomes. It is a common assumption that the number of crossover events follows Poisson distribution and their locations in the chromosome are random and independent from each other. The probability that two loci are separated by (an odd number of) crossovers, yielding in an observable recombination, is called *recombination factor* and is denoted by $\theta$. If the two loci are wide apart, or in different chromosomes, the expected value of $\theta$ is close to or equal than 0.5. The closer the loci are in the same chromosome, the smaller the expected value of $\theta$.

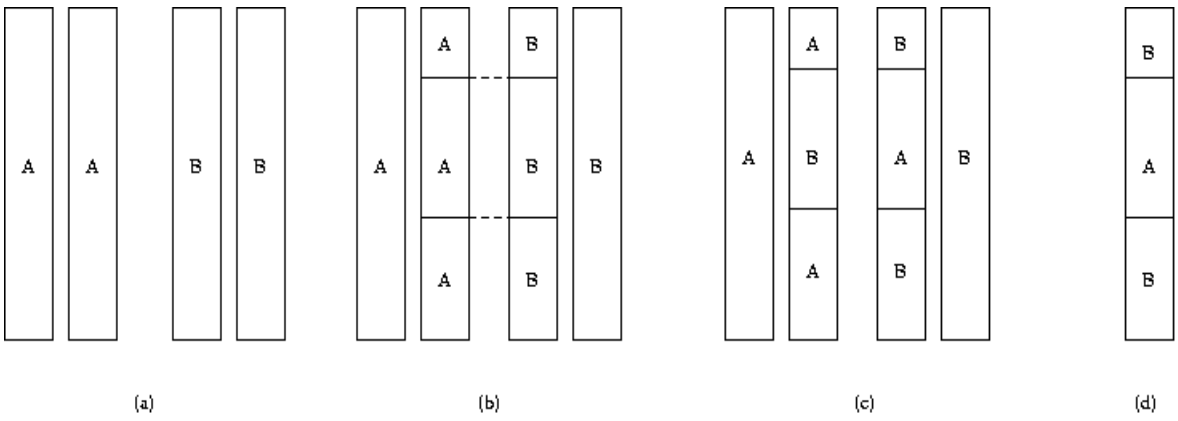



**Figure 1.** A highly simplified figure of meiosis. The two chromatids A and B of a parent duplicate (a) and exchange genetic material between crossover points (b). The resulting chromatids (c) may be recombinants, and one of them may eventually end in a gamete (d)

In linkage analysis at its simplest, the analyst tries to estimate the value of recombination fraction $\theta$ between a marker and a hypothetical disease locus. This is based on the observable meioses in the pedigrees (deduced from the genetic markers). The underlying idea is that common inheritance in the two loci indicates that the loci are probably tightly linked in the genetic map. Detecting strong linkage between a marker and the hypothetical disease locus is thus equivalent to locating the disease gene.

A simple approach for linkage analysis is thus to estimate the recombination fraction between the hypothetical disease locus and each marker in turn. The gene is then expected to be close to the marker or markers that is tightly linked to the gene (low value of $\theta$) and gives strong evidence for linkage. The degree of evidence is often expressed by LOD-score, which is a derivative of likelihood-ration test, familiar from statistics.

Linkage analysis requires pedigree data. The pedigrees can be as simple as two-generation nuclear pedigrees, or they can contain several generations and tens or even hundreds of individuals. Assuming that the underlying disease locus is the same, the pedigrees themselves need not be collected from the same genetic isolate. The alleles associated with the disease locus may vary between families.

**Markers and information**

Before going into details of linkage computations, let us consider the degree of information that can be obtained from genetic markers. Since the information content of an individual marker is limited, markers do not provide full information about inheritance. This is illustrated by concepts of sharing an allele IBS and IBD: We say that two individuals share an allele **IBS** (*identical-by-state*), if the alleles have a similar label. They share an allele **IBD** (*identical-by-descent*) if they, in addition, have inherited the same copy of allele from a common ancestor.

Consider the three example pedigrees in Figure 2. A single marker is genotyped for all individuals in the pedigrees.

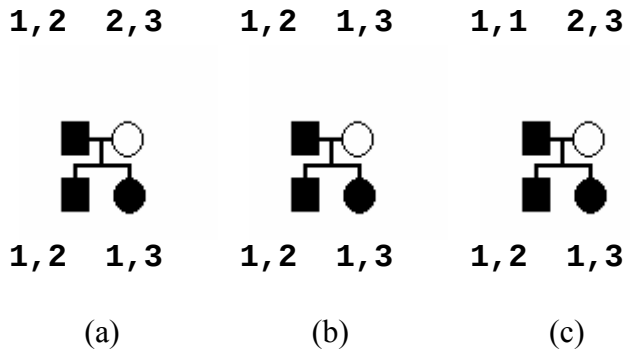



**Figure 2.** Three nuclear pedigrees with different degrees of IBD sharing of the offspring in spite of similar observed alleles.

In example (a) the children share allele 1 identical-by-state (IBS). They also share allele 1 identical-by-descent (IBD) since they must have inherited it from the same parental chromosome, namely that of the father.

In example (b), allele 1 of the son is necessarily inherited from the mother. Allele 1 of the daughter is inherited from the father. The children share allele 1 identical-by-state (IBS), but they do not share any alleles identical-by-descent (IBD).

In example (c), the inherited alleles labeled as 1 can be of same or different origin (stemming either from father's father or father's mother). Even though the children share allele 1 identical-by-state (IBS) with certainty, they share allele 1 identical-by-descent (IBD) with probability of 0.5.

Generating statistics for linkage based solely on alleles shared among affected relatives would be rather straightforward. As the previous examples illustrate, looking at allelic states is, however,

not enough to obtain full information about inheritance in marker loci. To overcome this limitation, most linkage analysis methods try to deduce the exact inheritance patterns, or their probability distributions, based on the observed marker data. Even though favorable in terms of results, this also creates the main computational burden of the algorithms.

### Building blocks of linkage analysis

In a gene mapping study, the input data for linkage analysis contains at least the following components:

1. Pedigree structures
2. Affection statuses of the involved individuals
3. Marker locations
4. Marker allele frequencies
5. Genotypes of the individuals at marker loci.

In addition, there may be information about environmental factors as well as the assumed disease model (in terms of penetrances and disease allele frequencies in the population).

## Types of linkage analysis

Linkage analysis methods and approaches can be classified by several criteria. The main division is often made between *parametric* and *non-parametric* methods. Parametric linkage methods require an assumption of disease model (expressed in terms of penetrances, i.e. conditional probabilities for phenotypes given genotypes, and disease allele frequency). So-called nonparametric methods do not require such assumptions, even though they may perform better with certain true disease models.

The next division can be made by use of phenotype information: the disease status can be either *dichotomous* (affected or healthy, as in the previous examples) or *continuous* (such as blood pressure). Most disease gene mapping studies use dichotomous response variables.

The linkage methods generally use one of the two algorithms with different ways of traversing the search space. Methods based on *Elston-Stewart algorithm* are exponential in relation to the number of markers, but linear to size of pedigree. On the other hand, the time and memory consumption of methods based on *Lander-Green algorithm* grow exponentially as a function of pedigree size, whereas the growth is only linear in respect to the number of markers. Furthermore, there are heuristic methods which often allow analyzing a large number of markers in large pedigrees in reasonable time (with a possibility of finding a sub-optimal result).

A further division is often made between *two-point* and *multipoint* linkage analysis. The former means that one marker is compared to the disease status at a time, whereas multipoint methods take into account several adjacent markers simultaneously.

The study may focus either in the entire genome or a collection of candidate regions. In a full *genome scan*, no a priori assumptions of the disease locus; all loci are considered equally likely. As a result, one is often interested in finding prominent regions for subsequent analysis, instead of narrow regions with confirmed statistical significance. In a *candidate gene* approach, the analyst focuses in a set of previously suspected candidate regions (often based on previously reported work). The task can be either to confirm linkage, finding additional evidence for a result reported earlier, or exclude linkage, ruling out genes or regions from subsequent analysis.

## Parametric linkage analysis

In parametric linkage analysis, the disease model is assumed known, or at least fixed. In the analysis, the putative location of the disease gene is the free parameter, and the task is to find such a value for this parameter that maximizes the likelihood of the data.

In two-point analysis, the parameter for the gene location, denoted by $\theta$, gives the genetic distance between the marker and the disease locus. Consider a single nuclear family (Figure 3) with two parents and two offspring:

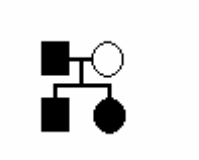


**Figure 3. A nuclear family.**

For a given pedigree structure, a *likelihood function* gives the probability of the data given the parameter values. The likelihood function for the pedigree of Figure 3 can be written as:

$$L_j = \Sigma_{gF} P(g_F)\, P(y_F \mid g_F)$$
$$\Sigma_{gM} P(g_M) P(y_M \mid g_M)$$
$$\Sigma_{gOi}\, P(g_{Oi} \mid g_F, g_M)\, P(y_{Oi} \mid g_O)$$

where $P(g_F)$ and $P(g_M)$ are the two-loci genotype probabilities for the mother and father, and $P(O_i|g_F, g_M)$ is the probability of a genotype for child $i$ given the parental genotypes. Here, the last expression depends on the distance between marker and disease loci ($\theta$). In the expression, $P(y_F|g_F)$, $P(y_M|g_M)$ and $P(y_{Oi}|g_{Oi})$ are the penetrances given the corresponding genotypes.

In linkage analysis, this expression is maximized. The estimate for the recombination fraction ($\theta$ ') indicates whether the disease locus is likely to reside close or distant to the marker. The goodness of the estimate is measured by *LOD score* which is defined as:

$$LOD(\theta') = \log_{10} \frac{L(\theta = \theta')}{L(\theta = 0.5)} = \log_{10} L(\theta = \theta') - \log_{10} L(\theta = 0.5)$$

The nominator in the expression corresponds to the null hypothesis situation (no linkage, i.e. $\theta$ =0.5).

By convention, a LOD score higher than three is taken as evidence of significant linkage. This is based on considerations of genome-wide significance in the presence of multiple testing inherent in full-genome scans.

The expression of the likelihood can be complex. Even in this simple case above, the summation needs to be done over all possible two-loci genotypes for the father, mother, and each of the offspring. As the pedigree structure becomes more complex, so does the expression of the likelihood. Thus, the values of the likelihood function are in practice constructed and evaluated *in silico*.

**Example of parametric linkage analysis**

Figure 3 shows a numerical example of parametric linkage analysis. The pedigree depicted in the top left corner of the figure contains parents and seven offspring, four of which are affected. We make an assumption of a dominant, fully penetrant disease, which means that the disease allele (denoted by D) must be inherited from the father to each of the affected offspring.

Because of the assumptions, the – rather complex – expression of the likelihood function is now reduced into the following form:

$$LOD(\theta') = \log_{10} \frac{L(\theta = \theta')}{L(\theta = ½)} = \frac{\sum_{gF} (P(g_F) \sum_{gOi} P(g_{Oi} \mid g_F, \theta = \theta'))}{\sum_{gF} (P(g_F) \sum_{gOi} P(g_{Oi} \mid g_F, \theta = ½))}$$

There is uncertainty is in the two-locus genotypes of the father: the disease gene can be in the same chromosome as either allele 1 or allele 3. The two main branches in the tree diagram of Figure 3 represent this uncertainty: each branch is equally likely *a priori*. In each of these main branches, there are four alternatives how the genetic material can be transmitted from the father to the offspring, the disease allele *D* can go together with allele 1 or allele 3, or the normal allele (+) can go together with either of the two marker alleles. Out of these four possibilities, two require a recombination between the disease and marker loci, occurring at probability $\theta$. Evidently, a recombination does not occur in the corresponding meiosis with probability 1-$\theta$.

For each of the eight branches, observed frequencies can be computed from offspring genotypes and phenotypes, and incorporated into the expression of the likelihood.

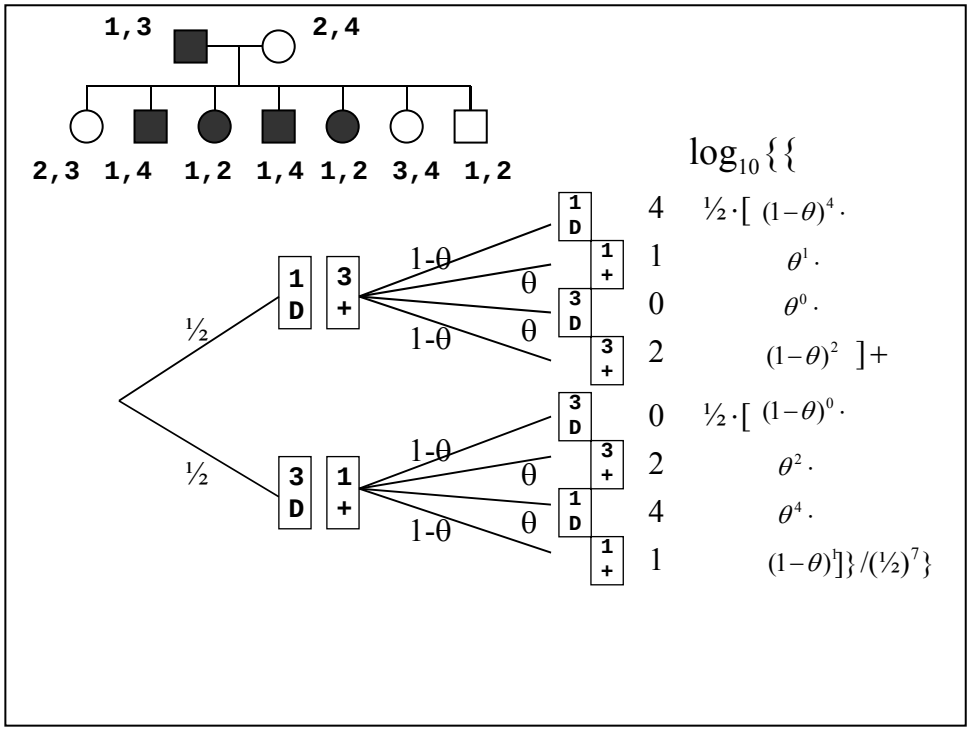



**Figure 4. An example of LOD score calculation for given pedigree and marker data.**

Maximizing the likelihood gives a point estimate of 0.14 for the recombination fraction $\theta$. The corresponding LOD score is 0.56. This is not significant, but in case we had six similar pedigrees, the overall LOD score would have exceeded the significance threshold of LOD=3.

**Extensions and implementations**

In the previous example, we conducted parametric two-point linkage analysis for a single marker. The algorithm can be extended to take into account multiple markers simultaneously, and to cope with more complex (general) pedigree structures as well as missing genotype and phenotype information.

Among the most widely-used implementations of parametric linkage analysis are Linkage and Genehunter software packages.

## Non-parametric linkage analysis

In non-parametric linkage analysis, no assumptions about the disease model are made. Thus, these methods are more suitable for the analysis of complex diseases, where the true disease model is generally unknown. The methods rely on evaluating the IBD sharing among the affected individuals in the pedigrees.

**NPL at its simplest: Affected sib-pair test**

The *affected sib pair test* (ASP) is the simplest test for non-parametric linkage. Even though rarely used in its basic form, it illustrates the basic intuition behind the nonparametric family of linkage tests. To get an idea of the test, the following example shows the computation of the test statistic for a sample of nine families, each having two affected offspring:

| Father | Mother | Child | Child | IBD=0 | IBD=1 | IBD=2 |
|---|---|---|---|---|---|---|
| 1,4 | 2,3 | 1,3 | 1,3 | | | 1 |
| 1,4 | 2,4 | 4,4 | 4,4 | | | 1 |
| 2,3 | 4,4 | 2,4 | 3,4 | | | |
| 2,5 | 1,3 | 2,3 | 2,3 | | | 1 |
| 2,5 | 1,4 | 4,5 | 1,2 | 1 | | |
| 3,4 | 2,4 | 2,4 | 2,4 | | | 1 |
| 3,5 | 2,4 | 2,3 | 2,5 | | 1 | |
| 4,5 | 2,5 | 5,5 | 5,5 | | | 1 |
| 2,2 | 4,5 | 2,4 | 2,5 | | | |
| 3,3 | 2,5 | 3,5 | 3,5 | | | |
| 3,5 | 4,4 | 4,5 | 3,4 | | | |
| 1,2 | 2,4 | 1,4 | 1,4 | | | 1 |
| 1,5 | 2,2 | 2,5 | 2,5 | | | |
| 2,5 | 1,3 | 1,5 | 1,5 | | | 1 |
| | | | | 1 | 1 | 7 |

From these families, it is calculated how often the offspring share 0, 1, and 2 alleles IBD. This can be done unequivocally only for families with two heterozygous parents. Finally, the observed counts are compared to the expected distribution, where 50% cases correspond to IBD=1, and 25% to each of IBD=0 and IBD=2 (This can easily be verified by enumerating the alternatives with a pen and a paper). The fit to the expected distribution is evaluated with a chi-squared test with 2 degrees of freedom. The general form of the test statistic is

$$\sum_{i=1}^{n} \frac{(o_i - e_i)^2}{e_i}$$

Applied to this case, the computation of the statistic goes as follows:

$$\chi^2 = \frac{(1-2{,}25)^2}{2{,}25} + \frac{(1-4{,}5)^2}{4{,}5} + \frac{(7-2{,}25)^2}{2{,}25} = 13{,}45$$

In this case we would get a significant result, since the corresponding p-value is approximately $p$=0.0012. However, it should be noted that the sample is far too small for reliable usage of a test statistic that is based on an asymptotic distribution.

**Extensions and implementations**

The ASP test has been generalized to take into account multiple markers, unknown genotypes, arbitrary pedigree sizes, and arbitrary pairs and groups of affected relatives. The state-of-art nonparametric linkage analysis methods are implemented in Merlin and Genehunter software packages. In these implementations, the computation proceeds as follows:

1. Find out the possible inheritance patterns (inheritance vectors) that are compatible with the observable genotype data in each pedigree, and compute their probabilities at each marker locus.
2. Propagate the results of the marker loci both to both directions on the map using a hidden Markov model.
3. For each point in the chromosome, compute the value of the NPL statistic based on the estimated IBD sharing between affected relatives.
4. Normalize the score statistic using the mean and variance of the NPL statistic under null hypothesis. These were obtained as side products of the calculation.
5. Combine the normalized statistics over the different pedigrees to obtain an overall NPL statistic for a given point in the chromosome.

## Conclusions

Linkage analysis utilizes pedigree data and constitutes one of the two main approaches to disease gene mapping. It often serves as a starting point for a full genome scan. The idea is to find regions where alleles are inherited together with the hypothetical disease allele in the pedigrees. The inheritance patterns are deduced based on the observed marker data, and the alleles may vary between families.

Linkage methods can be classified by many criteria. For instance, methods that require an assumption of a disease model are called parametric methods whereas non-parametric methods do not require such assumptions. Nowadays, non-parametric methods are more popular, since the main interest is in complex diseases. Analysis is restricted by the drawback that exact methods and programs for linkage analysis have exponential memory and computing time requirement either in respect to the size of the pedigree (Merlin and Genehunter) or to the number of markers (Linkage).

## Sources and further reading


- E. Lander, P. Green, Construction of multilocus genetic linkage maps in humans. Proc. Nat. Acad. Sci, USA, 84 (1987), 2363-2367.
- L. Kruglyak, M. Daly, M. Reeve-Daly, E. Lander, Parametric and nonparametric linkage analysis: a unified multipoint approach.
- G. Abecasis, S. Cherny, W. Cookson, L. Cardon, Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. Nature Genetics, 30 (2002), 97-101.
- A. Ingólfsdóttir, A. Christensen, J. Hansen, J. Johnsen, et al., Formalization of Linkage Analysis. BRICS Report Series. RS-02-7 (2002).

# Work 1. DNA Extraction

In this exercise, DNA is extracted from white blood cells. First the cells are broken up with low salt buffer (TKM1) and nuclear pellets are centrifuged. The nuclear pellet is resuspended in high salt buffer and detergent (SDS) is added. This reagent is similar to soap and breaks up the nuclei. Afterwards, proteins are precipitated with high salt solution (NaCl) and DNA is precipitated with 100% ethanol.

Buffers

**TKM-1**
10 mM Tris-HCl pH 7.6
10 mM KCl
10 mM MgCl
2 mM EDTA pH 7.6
(2.5% Nonidet P-40)

**TKM2**
10 mM Tris-HCl pH 7.6
10 mM KCl
10 mM MgCl
2 mM EDTA pH 7.6
0.4 M NaCl

20% SDS (Sodium dodecyl sulfate)

**5M NaCl**

## Preparation of DNA from blood

1. Divide blood into two centrifuge tubes and add 5 ml of TKM1+Nonidet P-40 – buffer. Invert several times and let the tubes stand for 10 minutes at room temperature (RT).

2. Centrifuge at 3000 RPM for 10 minutes at RT in a tabletop centrifuge.

3. Slowly pour off supernatant (liquid) and save nuclear pellet (the small pellet at the very bottom of the tube) and wash pellet in 10 ml of TKM1-buffer (without Nonidet P-40), shake well. Centrifuge as before and pour off supernatant.

4. Add 800 µl of TKM2-buffer into tube and resuspend pellet with pasteur-pipette. Transfer suspension into 2 ml Eppendorf-tube.

5. Add 50 µl of 20% SDS and then mix whole suspension thoroughly by pipetting back and forth several times.

6. Incubate tubes at 55°C overnight.

---

7. Add 360 µl of 5M NaCl into tube and mix well.

8. Centrifuge at 12000 RPM for 10 min in microcentrifuge.

9. Pipet supernatant (containing DNA) carefully into 10 ml tube. Add 3.5 ml of 100% ethanol (-20°C). Invert tube several times until DNA precipitates.

10. Remove the precipitated DNA strands with melted pasteur-pipette and put DNA in 10 ml of 70% ethanol to another tube. Let the tubes stand at room temperature overnight.

---

11. Collect DNA strands with melted pasteur-pipette and let them dry on the top of pipette. Dissolve strands in 300 µl of sterile water in 1.5 ml Eppendorf-tube.

12. Measure the concentration of DNA: take 2 µl of sample into 98 µl of water (dilution 1:50). Measure sample in spectrophotometer (Gene Quant II RNA/DNA calculator).

The concentration of DNA is calculated by measuring the DNA-absorbance (wavelength 260). At this wavelength, a solution containing a DNA concentration of 50 ng/µl has an absorbance value of 1. DNA is free of proteins if the ratio of A260/A280 is equal to 1.8.

13. DNA is stored at -20°C.

# Work 2. PCR-RFLP

## IDENTIFICATION OF MTHFR (C677T) POLYMORPHISM BY USING PCR-RFLP TECHNIQUE

A common polymorphism, C→T substitution at nucleotide 677 in methylenetetrahydrofolate reductase (MTHFR) gene leads to the conversion of an alanine to a valine at amino acid residue 223. This amino acid substitution generates a thermolabile enzyme with reduced enzyme activity. Carriers of the mutation have elevated plasma homocysteine levels. Therefore this mutation has been speculated to represent a genetic risk factor for vascular diseases. The polymorphism has been shown to be weakly associated to increased risk of venous thromboembolism in some studies, but the results are controversial. The allele frequency in the Finnish population is approximately 20 %.

The aim of this study is to determine the genotype of MTHFR (C677T) polymorphism of 30 individuals selected randomly among Finnish blood donors. A fragment containing the polymorphism of interest is first PCR amplified. The PCR product is then digested with a specific restriction enzyme *HinfI*. *HinfI* recognizes the sequence shown below, and cuts the double-stranded DNA leaving so called "staggered" ends.

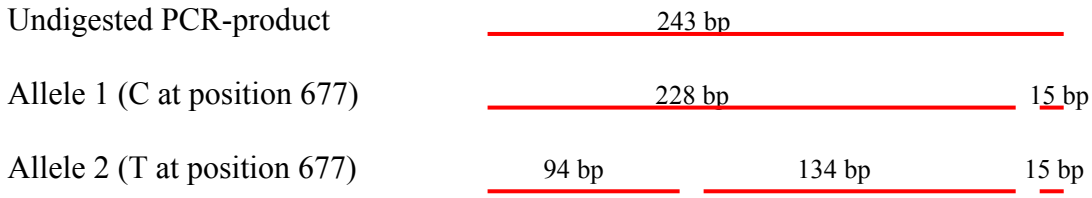
Recognition sequence of *HinfI*:

5' …..G | ANTC….. 3'

3' …..C TNA | G…..5'

Conversion of nucleotide C to T at the nucleotide position 677 of MTHFR gene alters the sequence, and creates a new digestion site for *HinfI* enzyme. However, to make sure that digestion reaction is performing as expected, the amplified region includes also a second recognition site for *HinfI*. This sequence functions as an internal control site for digestion reaction.

Fragments of following sizes are created:




The digested fragments are separated by agarose gel electrophoresis with ethidium bromide staining. The genotype is interpreted based on the sizes of the fragments.

**Materials**

- 6 randomly selected DNA samples (10 ng/µl) per each group
- 10x PCR buffer
- $MgCl_2$ (25 mM)
- dNTP mixture (2,5 mM each)
- Ampli Taq Gold DNA polymerase
- *HinfI* Restriction enzyme
- Agarose
- Ethidium bromide
- 1xTBE
- 50 bp DNA ladder
- 6xLoading dye

**Methods**

## PCR

1. Prepare one master-mix for all six PCR reactions by combining the following reagents in an Eppendorf tube:

| | 1x | Master-mix (8x) |
|---|---|---|
| 10x PCR buffer | 2,5 µl | 20,0 µl |
| $MgCl_2$ (25 mM) | 3,0 µl | 24,0 µl |
| dNTP mixture (2,5 mM each) | 2,5 µl | 20,0 µl |
| Forward primer (5 µM) | 1,0 µl | 8,0 µl |
| Reverse primer (5 µM) | 1,0 µl | 8,0 µl |
| Sterile $H_2O$ | 9,75 µl | 78,0 µl |
| Ampli Taq Gold DNA polymerase (5U/µl) | 0,25 µl | 2,0 µl |

2. For each PCR reaction, combine the template DNA and PCR mastermix in a thin-walled PCR tube:

| | |
|---|---|
| Template DNA (10 ng/µl) | 5 µl |
| PCR Mastermix | 20 µl |

3. Transfer the samples into a thermal cycler and perform thermal cycling under the conditions given below:

1. 95°C for 10 min
2. 95°C for 30 sec
3. 58°C for 30 sec
4. 72°C for 30 sec
5. Repeat steps 2.-4. for 29 times
6. 72°C for 7 min
7. 4°C forever

## Digestion

1. Prepare a digestion master-mix by combining the following reagents in an Eppendorf tube. Keep the enzyme on ice! :

| | 1x | MasterMix (8x) |
|---|---|---|
| Sterile $H_2O$ | 1,5 µl | 12 µl |
| *HinfI* (10 U/µl) | 0,5 µl | 4 µl |

2. Combine 20 µl of PCR product and 2 µl of digestion mastermix in a thin-walled PCR-tube. Mix the reagents by pipetting up and down.

Incubate at 37°C for 4 hours, followed by heat inactivation at 80°C for 20 minutes. Incubation is done in a thermal cycler (PCR machine).

## Agarose gel electrophoresis

**To prepare a 2,5 % agarose gel**, 5 g of agarose is mixed with 200 ml of 1xTBE. The mixture is heated in a microwave oven until agarose is completely dissolved and solution is clear. After cooling solution to about 60°C, four drops of Ethidium bromide is added. The gel solution is then poured into a casting tray containing a sample comb. The gel is allowed to solidify at room temperature.

1. **To prepare samples for the electrophoresis**, add 4 µl of 6x loading dye into each digested PCR product and mix carefully with pipette.

2. Each group will combine the rest of undigested PCR products together by pipeting 5 µl of each to the new eppendorf tube. Add 4 µl of 6x loading dye into combined undigested PCR –sample.

The comb is removed carefully. The tray containing the gel is inserted in the electrophoresis chamber which is filled with electrophoresis buffer (1xTBE).

3. Load the gel by pipeting 24 µl of the sample containing the loading buffer into a sample well. Each group will have 6 digested and 1 undigested samples.

Electrophoresis is performed at 100 volts until loading dye has migrated an appropriate distance (approximately 2 hours). DNA, which is negatively charged in solution, will migrate towards the positive pole. To visualize DNA, the gel is placed on an ultraviolet transilluminator. A Polaroid picture is taken of the gel.

## Interpretation of genotyping results

The sizes of separated fragments can be estimated based on the known fragment sizes of used size standard (50 bp DNA ladder). Small fragments under 50 bp are really difficult to visualize on the agarose gel. For this reason the 15 bp fragment obtained after digestion with *HinfI* can not be seen.

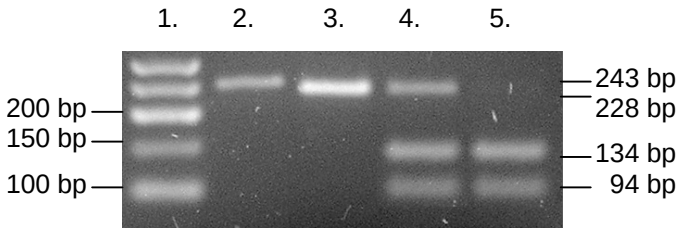



Figure 1. MTHFR (C677T) fragments after *HinfI* digestion separated on 2,5% agarose gel.

1. 50 bp DNA ladder
2. Undigested PCR-product
3. Homozygote C (1 1)
4. Heterozygote CT (1 2)
5. Homozygote T (2 2)

Calculate the allele frequencies based on the obtained genotypes. Also *Hardy-Weinberg equilibrium* is calculated to confirm the correctness of results.

**Calculation of allele frequencies by using the observed genotypes:**

Genotypes:
$N_{AA}$ = number of homozygotes for the allele A
$N_{Aa}$ = number of heterozygotes
$N_{aa}$ = number of homozygotes for the allele a

Allele frequency for the allele A: $$p = \frac{2N_{AA} + N_{Aa}}{2(N_{AA} + N_{Aa} + N_{aa})}$$

Allele frequency for the allele a: $$q = \frac{2N_{aa} + N_{Aa}}{2(N_{AA} + N_{Aa} + N_{aa})}$$

**Hardy-Weinberg equilibrium**
The frequency of homozygotes and heterozygotes stay in constant proportions in population, if certain conditions, such as random mating, are fulfilled. The proportions are given by Hardy-Weinberg equation:

$$p^2+2pq+q^2=1$$

where

- $\boldsymbol{p}$ is the allele frequency for the allele **A**
- $\boldsymbol{q}=\mathbf{1}-\boldsymbol{p}$ is the allele frequency for the allele **a**
- genotype frequency for the **AA** genotype is $\boldsymbol{p}^2$
- genotype frequency for the **Aa** genotype is $\mathbf{2}\boldsymbol{pq}$
- genotype frequency for the **aa** genotype is $\boldsymbol{q}^2$

H-W equilibrium means that formulation of genotypes is independent: the probability that an individual receives allele A (with probability *p*) does not affect the probability with which the individual receives the second allele (A with probability *p*, a with *q*)

In genotyping studies, the H-W equilibrium is routinely tested: On the basis of the observed genotype distribution, the expected occurrence rates of the different genotypes can be calculated according to the Hardy-Weinberg principle and compared with the observed occurrence rates by $\chi^2$ -test. A *p*-value of $<0.05$ is considered statistically significant.

1. Calculate the allele frequencies based on observed genotypes as shown above.
2. Calculate the expected genotype frequencies according to Hardy-Weinberg equilibrium.
3. Calculate the expected number of genotypes based on expected genotype frequencies.
4. Compare the number of expected and observed genotypes by using the $\chi^2$ -test.

## Genetics glossary

| | |
|---|---|
| Allele | An alternative form of a gene or a marker |
| Base pair (bp) | A pair of complementary nitrogenous bases (adenine and thymine or guanine and cytosine) in a DNA molecule. Also, the unit of measurement for DNA sequences (e.g., 200 bp) |
| Chromosome | A single DNA molecule containing genes (and markers) in linear order. Inhumans, 23 pairs of chromosomes, each pair containing one chromosome from each parent, carry the entire genetic code |
| Crossing over | The interchange of sections between pairing homologous chromosomes during meiosis |
| Disease model | Number of genes, their effects, environmental factors, and interactions which affect the disease susceptibility for a certain disease. Disease with genetic contribution may be monogenic (Mendelian one-gene disease), oligogenic, where just a few genes are involved, or polygenic with several genes with weak effects each, for example |
| Gene | Basic element of heredity that determines traits, coding for proteins |
| Genetic association | Correlation of presence of a disease or a trait with presence of certain marker allele(s) (or alleles at genes), observed at the population level |
| Genotype | The particular alleles at specified locus present in an individual |
| Haplotype | A string of alleles from genes or markers which are located closely together on the same chromosome and which tend to be inherited together |
| IBD | Identity By Descent, where two copies of an identical allele have been inherited from a common ancestor (see IBS) |
| IBS | Identity By State. Any two copies of an allele which are chemically identical. Need not to be inherited from same source (see IBD) |
| LD | Linkage disequilibrium. Alleles of separate loci occur together at population level more often than can be accounted for by chance. Usually indicates that the loci are physically close to each other on the chromosome |
| Linkage | The tendency of genes in proximity of each other to be inherited together. The closer the loci, the greater the probability that they will be inherited together |
| Locus (plural loci) | The specific site of a particular gene or marker on its chromosome |
| Marker | A gene or a stretch of non-coding DNA sequence, the alternative forms (alleles) of which can be reliably detected by genotyping technologies |
| Marker map | The positions of a set of marker genes chosen for some particular mapping study |
| Pedigree | A family tree diagram which shows the genetic history of a particular (often multigenerational) family |
| Phase | Parental origin of a haplotype or chromosome |
| Phenotype | The observable and measurable characteristics of an organism, e.g. presence of a disease, which may or may not be genetic |
| Population | A group of organisms of the same species relatively isolated from other groups of the same species |
| Recombination | The process by which offspring derive a combination of genes (or markers) different from that of either parent. Occurs by crossing over |
| SNP | Single nucletide polymorphism differing in a single base pair. |
| Trio, triplet | An offspring and the parents (family trio) |