

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/303309837>

Medical genetics and genomics 2016

Book · May 2016

CITATIONS

0

READS

49,839

6 authors, including:



Csaba Szalai

Semmelweis University

275 PUBLICATIONS 4,833 CITATIONS

[SEE PROFILE](#)



Sára Tóth

Semmelweis University

90 PUBLICATIONS 2,240 CITATIONS

[SEE PROFILE](#)



Erna Pap

Semmelweis University

37 PUBLICATIONS 2,334 CITATIONS

[SEE PROFILE](#)



Andras Falus

Semmelweis University

357 PUBLICATIONS 6,396 CITATIONS

[SEE PROFILE](#)

MEDICAL GENETICS AND GENOMICS

Editor: Csaba Szalai, PhD, university professor

Authors:

Chapter 1, 9, 10, 11, 12, 13, 14, 15: Csaba Szalai

Chapter 2: Valéria László

Chapter 3, 4, 5, 7, 8: Sára Tóth

Chapter 6: Erna Pap

Chapter 16: András Falus and Ferenc Oberfrank,

Summary

The book contains the substance of the lectures and partly of the practices of the subject of 'Genetics and Genomics' held in Semmelweis University for medical, pharmacological and dental students. The book updated in 2016 starts with a short introduction to basic genetics and molecular biology and then topics from human genetics mainly from medical point of views. Some of the 16 chapters deal with medical genetics, but the chapters also introduce to the basic knowledge of cell division, cytogenetics, epigenetics, developmental genetics, stem cell biology, oncogenetics, immunogenetics, population genetics, evolution genetics, nutrigenetics, and to a relative new subject, the human genomics and its applications for the study of the genomic background of complex diseases, pharmacogenomics and for the investigation of the genome environmental interactions. As genomics belongs to systems biology, a chapter introduces to basic terms of systems biology, and concentrating on diseases, some examples of the application and utilization of this scientific field are also be shown. The modern human genetics can also be associated with several ethical, social and legal issues. The last chapter of this book deals with these issues. At the end of each chapter there are questions, with which the readers can ascertain whether they understood and/or learned the chapter. Because it is an e-book, some terms and definitions has a hyperlink for more detailed explanations in the World Wide Web. Besides university students, the book is also recommended to all those who are interested in modern medical genetics and genomics and want to be up-to date in these subjects.

Keywords: Mitosis, meiosis, mutations, cytogenetics, epigenetics, Mendelian inheritance, genetics of sex, developmental genetics, stem cell biology, oncogenetics, immunogenetics, human genomics, genomics of complex diseases, genomic methods, population genetics, evolution genetics, pharmacogenomics, nutrigenetics, gene environmental interaction, systems biology, bioethics.



Typotex Kiadó

2013

Updated in 2016

COPYRIGHT: András Falus, Valéria László, Ferenc Oberfrank, Erna Pap, Csaba Szalai, Sára Tóth, Budapest University of Technology and Economics

Creative Commons NonCommercial-NoDerivs 3.0 (CC BY-NC-ND 3.0)

This work can be reproduced, circulated, published and performed for non-commercial purposes without restriction by indicating the author's name, but it cannot be modified.

Lector: Viktor Molnár, MD, head of Csertex Research Laboratory

ISBN 978 963 279 187 6

Prepared under the editorship of Typotex Kiadó

Responsible manager: Votisky Zsuzsa

Made within the framework of the project Nr. TÁMOP-4.1.2/A/1-11/1-2011-0079, entitled „Konzorcium a biotechnológia és bioinformatika aktív tanulásáért”.

Nemzeti Fejlesztési Ügynökség

www.ujszechenyiterv.gov.hu

06 40 638 638



MAGYARORSZÁG MEGÚJUL



A projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Content

1.	Basic genetics and molecular biology	9
1.1.	Basic genetics	9
1.2.	Mendelian inheritance	9
1.2.1.1.	Dominant-recessive inheritance	10
1.2.2.	Polygenic inheritance	10
1.2.3.	Genetic pleiotropy	11
1.2.4.	Sex-linked inheritance	11
1.2.5.	Linked inheritance, recombination	12
1.2.6.	Mutation	13
1.3.	Basics of molecular biology	13
1.3.1.	Some characteristics of the human DNA	14
1.3.2.	Replication	17
1.3.3.	Transcription	18
1.3.4.	Structure of the genes	18
1.3.5.	The genetic code	19
1.3.6.	Translation	20
2.	Transmission of genetic information	22
2.1.	Cell cycle and regulation of cell cycle	22
2.1.1.	G ₀ - G ₁ transition	23
2.1.2.	G ₁ – S transition, S-phase	25
2.1.3.	G ₂ – M transition	26
2.1.4.	M-phase	27
2.1.4.1.	Chromosome structure	29
2.1.4.2.	Structure and role of mitotic spindle	32
2.1.4.3.	Metaphase – anaphase transition	35
2.1.5.	Cytokinesis	37
2.1.6.	Operation of cell cycle checkpoints	37
2.2.	Chromosome territories	39
2.3.	Meiosis	41
2.3.1.	Phases of meiosis	42
2.3.2.	Oogenesis	46
2.3.3.	Spermatogenesis	47
2.3.4.	Regulation of meiosis	50
3.1.	Mutation and polymorphism	52
3.2.	The classification of mutations	53
3.3.	Gene mutations	55
3.4.	DNA repair	61
3.5.	Mutagenicity tests	64
3.6.	Nomenclature of the genetic variants	65
3.6.1.	Levels of variations	65
3.6.2.	Positions of the variations	66
3.6.3.	Specific changes	67
3.7.	Useful web-sites:	67
3.8.	Questions	68
4.	Cytogenetics	69
4.1.	Structural chromosome aberrations	70
4.1.1.	Deletions	71
4.1.2.	Duplications	71
4.1.3.	Translocations	72

4.1.3.1.	Reciprocal translocations.....	72
4.1.4.	Inversions	74
4.1.5.	Ring (ring) chromosome	75
4.1.6.	Isochromosome	75
4.1.7.	Dicentric chromosome	79
4.1.8.	Acentric fragment	80
4.2.	Numerical chromosome aberrations	80
4.2.1.	Euploid chromosome mutations	80
4.2.2.	Aneuploid chromosomal aberrations	81
4.2.3.	The most common numerical chromosomal abnormalities.....	83
4.2.3.1.	Trisomy 21	84
4.2.3.2.	Trisomy 13	85
4.2.3.3.	Trisomy 18	85
4.2.4.	Numerical sex chromosome aberrations	85
4.2.4.1.	Turner syndrome	85
4.2.4.2.	Klinefelter syndrome	86
4.2.4.3.	Triple X syndrome	86
4.2.4.4.	Double-Y syndrome, "superman" or Jacobs syndrome.....	86
4.3.	Uniparental disomy (UPD).....	87
4.4.	Mixoploid mutations	87
4.4.1.	Mosaicism.....	88
4.4.2.	Chimerism	88
4.5.	Useful web-sites:.....	89
4.6.	Questions	89
5.	Epigenetics	91
5.1.	Epigenetic changes - molecular modifications.....	91
5.1.1.	DNA methylation.....	92
5.1.2.	CpG as mutation hot spot	92
5.1.3.	Histone modifications	93
5.2.	Non-coding RNAs	94
5.3.	Epigenetic phenomena	94
5.3.1.	X-chromosome inactivation	94
5.3.2.	Genomic imprinting	96
5.3.2.1.	Imprinting related diseases	97
5.3.2.2.	Evolutionary theories of imprinting	98
5.4.	The significance of epigenetic effects.....	98
5.5.	Useful web-sites.....	101
5.6.	Questions	101
6.	Mendelian Inheritance: autosomal inheritance.....	102
6.1.	Introduction	102
6.2.	Interpretation of some basic genetic terms.....	104
6.3.	Phenomena that fine-tune classical monogenic inheritance.....	106
6.4.	Autosomal dominant inheritance	110
6.4.1.	General characteristics of autosomal dominant (AD) inheritance	110
6.4.2.	Diseases due to the mutation of structural genes	112
6.4.2.1.	Marfan syndrome	112
6.4.2.2.	Osteogenesis imperfecta.....	112
6.4.3.	Diseases due to mutations of receptor genes	112
6.4.3.1.	Achondroplasia	112
6.4.4.	Mutations of the gene of a protein with a yet unknown function	113

6.4.5.	Mutation of Protooncogenes.....	114
6.4.6.	Pharmacogenetic diseases	114
6.5.	Autosomal recessive inheritance	115
6.5.1.	General characteristics of autosomal recessive (AR) inheritance	115
6.5.2.	Enzymopathies.....	115
6.5.3.	Cystic fibrosis	117
6.5.4.	Haemoglobinopathies.....	118
6.6.	Genes and Tumors	118
6.7.	Genes and Drugs.....	120
6.8.	Conclusion.....	120
6.9.	Questions	121
7.	The role of sex in heredity	122
7.1.	X-linked inheritance.....	122
7.1.1.	X-linked dominant (XD) Inheritance	122
7.1.2.	X-linked recessive (XR) Inheritance.....	124
7.2.	Y-linked (holandric) Inheritance	126
7.3.	Sex influenced inheritance	126
7.4.	Sex limited inheritance.....	127
7.5.	Genomic imprinting.....	127
7.6.	Cytoplasmic inheritance	127
7.6.1.	Maternal genetic effect.....	127
7.6.2.	Mitochondrial inheritance	128
7.7.	The X chromosome inactivation.....	129
7.8.	Questions	130
8.	Genetics of biological processes	132
8.1.	Developmental genetics	132
8.1.1.	Morphogens	133
8.1.2.	Homeobox genes.....	133
8.2.	The genetics of sex.....	134
8.2.1.	Male sex determination in mammals	134
8.2.2.	Development of female sex in mammals	137
8.3.	Stem cell biology	138
8.4.	Oncogenetics	139
8.4.1.	Oncogenes	140
8.4.2.	Tumor suppressor genes.....	141
8.4.3.	Anti-apoptotic genes	142
8.4.4.	Telomerase.....	142
8.5.	Immunogenetics.....	143
8.6.	Useful web-sites:.....	147
8.7.	Questions	147
9.	Introduction to genomics	148
9.1.	Genomics.....	148
9.2.	Human Genome Project	149
9.3.	DNA sequencing.....	151
9.4.	Participants in the Human Genome Project	155
9.5.	Some results of the HGP	156
9.6.	Variations in the human genome	161
9.7.	Junk DNA in the human genome.....	165
9.8.	Comparative genomics.....	168
9.9.	Literature	170

9.10.	Questions	171
10.	Genomic approach to complex diseases	173
10.1.	General features of the complex diseases.....	173
10.2.	Environmental factors	174
10.3.	Why is it important to study the genomic background of the complex diseases? ..	174
10.4.	Heritability of the complex diseases	175
10.5.	Calculating heritability	176
10.6.	Difficulties in the studies of the genomic background of complex diseases.....	177
10.7.	Development of genomic methods, problems	180
10.8.	Problems of rare variants.....	182
10.9.	Epigenetics of the complex disease	183
10.10.	The random behavior of the genome	183
10.11.	Statistical problems	183
10.12.	Possible solutions.....	184
10.13.	Why are the complex diseases more frequent in our days?.....	186
10.13.1.	Thrifty gene hypothesis	186
10.13.2.	Hygiene hypothesis	187
10.13.3.	Additional theories	188
10.14.	Literature	190
10.15.	Questions	191
11.	Genomic methods for complex diseases	193
11.1.	Genetic markers	193
11.2.	Methods for the genomic backgrounds of diseases	194
11.2.1.	Study of genetic variants	194
11.2.2.	GWAS	196
11.2.3.	Evaluation of GWAS results	197
11.2.4.	Partial genome screenings	198
11.2.5.	Positional cloning.....	198
11.2.6.	Personal genomics.....	199
11.2.7.	New generation sequencing (NGS).....	200
11.2.8.	Measurement of gene expression.....	200
11.2.9.	Determination of the methylation of the genome	201
11.2.10.	Additional microarray-based methods	202
11.3.	Animal models.....	202
11.3.1.	The advantages of the animal models	202
11.3.2.	Shortcomings of animal models	204
11.3.3.	Experimental disease models.....	205
11.4.	Literature	206
11.5.	Questions	207
12.	Population and evolution genetics	209
12.1.	Population genetics	209
12.1.1.	Types of sample collection	209
12.1.2.	Selection of populations for genetic studies	211
12.1.3.	Hardy Weinberg equilibrium.....	211
12.1.4.	Linkage and haplotype	213
12.1.5.	Founder populations	215
12.1.6.	Association studies	216
12.1.7.	Risk calculation.....	218
12.2.	Evolutionary genetics	219
12.2.1.	Gene environmental interactions and the human genome	219

12.2.1.1.	Natural selection	219
12.2.1.2.	Role of infections in formation of the genome	220
12.2.1.3.	Genetic drift	220
12.2.2.	Why are some lethal mutations frequent?	221
12.2.3.	Examples for effects forming the genome	224
12.3.	Literature	228
12.4.	Questions	229
13.	Gene environmental interaction	231
13.1.	Penetrance of the genetic variants	231
13.2.	Interactions between highly penetrant variations and the environment	232
13.3.	Examples for interactions between low penetrant variations and environment	233
13.4.	Smoking-genome interaction	234
13.4.1.	Genomic background of smoking	235
13.4.2.	Smoking-gene interaction in disease susceptibilities	236
13.4.3.	Smoking-gene interactions in complex diseases	237
13.5.	Examples for gene-environmental interactions	239
13.6.	Genomic investigations of the gene-environmental interaction	244
13.7.	Nutrigenetics and nutrigenomics	246
13.8.	The future of gene-environmental interaction	248
13.9.	Literature	249
13.10.	Questions	252
14.	Pharmacogenomics	255
14.1.	Goals of pharmacogenomics	255
14.1.1.	Drug development	255
14.1.2.	Adverse drug response	256
14.2.	Genomic background of adverse effects	258
14.3.	Difficulties of the pharmacogenomic researches	259
14.4.	Genetic variants influencing pharmacokinetics	260
14.5.	Genes influencing pharmacodynamics	262
14.6.	Examples of pharmacogenetic studies	263
14.6.1.	Pharmacogenetics in oncology	263
14.6.2.	Pharmacogenetics of statins	264
14.6.3.	Clopidogrel	265
14.6.4.	MODY	266
14.6.5.	Pharmacotherapy of asthma	267
14.6.6.	Interaction between genetic variations and β_2 -agonists	268
14.6.7.	Interaction between genetic variations and leukotriene antagonists	268
14.7.	The future of pharmacogenomics	270
14.8.	Literature	271
14.9.	Questions	274
15.	Systems biologic approach of diseases	276
15.1.	Introduction	276
15.2.	Displaying interactions	276
15.3.	Human interactome	277
15.4.	Disease genes in the networks	278
15.5.	Nodes and edges in diseases	282
15.6.	Human Diseasome	283
15.7.	Shared gene hypothesis	284
15.8.	Shared metabolic pathway hypothesis	285
15.9.	Shared microRNA hypothesis	286

15.10.	Phenotypic Disease Network (PDNs)	286
15.11.	Application of systems biological approaches.....	287
15.12.	Literature	292
15.13.	Questions	294
16.	Bioethical and research ethical issues in genetic research	295
16.1.	Background.....	295
16.2.	The ethically challenging areas and of genetic research, the "border" issues ...	296
16.3.	The biobanks.....	300
16.4.	Some general ethics-related issues.....	301
16.5.	The specific genetic research bioethics and research ethics	302
16.6.	The ethics issues of commercialization of genetic information	302
16.7.	The genetic research, biobanks, data management and ethics legislation.....	303
16.8.	Conclusion	305
16.9.	Bibliography	305

1. Basic genetics and molecular biology

Csaba Szalai

1.1. Basic genetics

The laws of inheritance are investigated by genetics. The different nucleic acids (DNA and RNA) in the living organism play a central role in the inheritance of the different features. The information in the DNA molecule is inherited from one generation to the next generation through **reproduction**. It means that the hereditary material is the DNA (in some viruses the RNA), more exactly the **genes** which are the functional units which determine the nature of the features.

Gene definition: Genes are the units of inheritance. Genes are pieces of DNA that contain information for synthesis of ribonucleic acids (RNAs) or polypeptides. Earlier only those units were regarded genes, which coded proteins. Nowadays, genes are also those, which code functional RNAs, which are not transcribed to proteins. These are called **non-coding RNAs**. In the so-called RNA-viruses (e.g. influenza, HIV1) genes are coded only in the form of RNA. The appearance of an organism which results from the expression of an organism's genes as well as the influence of environmental factors and the interactions between the two is called **phenotype**. The genetic background of an organism is called **genotype**.

The majority of the DNA content of the cells is packaged in **chromosomes** and DNA can be also found in **mitochondria**. In diploid cells a couple of **homologous chromosomes** are a set of one maternal chromosome and one paternal chromosome that pair up with each other inside a cell during meiosis. These copies have the same genes in the same locations, or **loci**. In the nature a given gene can have different variations, these are called **alleles**. In a given population the most frequent allele of a gene is called **wild type**. If in a diploid cell the same alleles occur in a given locus of the homologous chromosomes then the organism is **homozygous**, if the alleles are different, it is **heterozygous** at this locus.

1.2. Mendelian inheritance

The founder of the modern science of genetics was Johann Gregor Mendel. Mendel's pea plant experiments established many of the rules of heredity, now referred to as the laws of Mendelian inheritance. Below Mendel's Laws are summarized (see more details in [Wikipedia](#)).

Law of Segregation (the "First Law") states that the two alleles for a heritable character segregate (separate from each other) during gamete formation and end up in different gametes.

Law of Independent Assortment (the "Second Law"), also known as "Inheritance Law", states that separate genes for separate traits are passed independently of one another from parents to offspring.

Law of Dominance (the "Third Law") states that recessive alleles will always be masked by dominant alleles. Therefore, a cross between a homozygous dominant and a homozygous recessive will always express the dominant phenotype, while still having a heterozygous genotype.

A **Mendelian trait** is one that is controlled by a single locus in an inheritance pattern. In such cases, a mutation in a single gene can cause a disease that is inherited according to Mendel's laws. These diseases are called **monogenic diseases**.

1.2.1.1. *Dominant-recessive inheritance*

In **dominantly inherited diseases** only one faulty gene is enough for the manifestation of the disease. Such disease is e.g. familial hypercholesterolemia or Huntington disease. In cases of **recessive diseases** the faulty gene product is compensated by the normal variant. In this case two mutated homologous genes are required for the manifestation of the disease. Such diseases are e.g. cystic fibrosis or albinism.

The **codominant inheritance** is a variation of the dominant-recessive inheritance. In case of codominant inheritance two different alleles of a gene can be expressed, and each version makes a slightly different protein. Both alleles influence the genetic trait or determine the characteristics of the genetic condition. E.g. blood type AB is inherited in a codominant pattern. Here the A and B blood group is dominant over 0 blood group and show codominant inheritance to each other. It means that if a person has one gene for A blood group, one for 0 blood group then his/her blood group will be A, in the case of one A and one B, the blood group will be AB.

1.2.2. *Polygenic inheritance*

In most cases, however, a trait or feature is determined more than one allele pair. Often the genes are large in quantity but small in effect. Examples of human polygenic inheritance are

height, skin color, eye color, weight and diseases like diabetes mellitus, high blood pressure, asthma, allergy or atherosclerosis.

1.2.3. Genetic pleiotropy

It can also occur that an allele pair is responsible for more than one trait. Here, the product of the gene participates in several metabolic pathways, which have effects on different organs or tissues. In this case mutations in this gene can have different consequences. It is called genetic pleiotropy. A classic example of pleiotropy is the human disease phenylketonuria (PKU). This disease can cause mental retardation and reduced hair and skin pigmentation, and can be caused by any of a large number of mutations in a single gene that codes for the enzyme phenylalanine hydroxylase, which converts the amino acid phenylalanine to tyrosine, another amino acid.

1.2.4. Sex-linked inheritance

Humans have altogether 46 chromosomes, which consist of 22 pairs of autosomes in both females and males and two sex chromosomes. There are two copies of the X-chromosome in females (homogametic), but males have a single X-chromosome and a Y-chromosome (heterogametic). Genes on the X or Y chromosome are called sex-linked. Since humans have many more genes on the X than the Y, there are many more X-linked traits than Y-linked traits. The gender of the offspring is determined by the sperm.

Cytogenetics is a field of genetics dealing with species or cell specific number of chromosomes, and their structure and characteristic segments, their functional roles, and all the differences - namely the chromosomal mutations - related to them. With cytogenetic methods (e.g. with chromosome staining) the chromosome X and Y can be easily differentiated. Chromosome X is significantly larger than the Y. Both chromosomes contain homologous and non-homologous regions. The non-homologous regions contain genes which do not have pairs in the other chromosome. In males these genes are in **hemizygotic state**.

Females possessing one **X-linked recessive mutation** are considered **carriers** and will generally not manifest clinical symptoms of the disorder. All males possessing an X-linked recessive mutation will be affected, since males have only a single X-chromosome and therefore have only one copy of X-linked genes. All offspring of a carrier female have a 50% chance of inheriting the mutation if the father does not carry the recessive allele. All female

children of an affected father will be carriers (assuming the mother is not affected or a carrier), as daughters possess their father's X-chromosome. No male children of an affected father will be affected, as males only inherit their father's Y-chromosome.

Because sex chromosomes contain different numbers of genes, different species of organisms have developed different mechanisms to cope with this inequality. **Dosage compensation** is the equalization of gene expression in males and females of a species. In humans, the females (XX) silence the transcription of one X chromosome, and transcribe all information only from the one expressed X chromosome. Thus females have the same amount of expressed X-linked genes as the human males (XY) who have just the one X chromosome to express from which to transcribe and express genes. The X inactivation, also called lyonization happens early in embryonic development at random. The inactive X chromosome in a female somatic cell is called a **Barr body**.

1.2.5. Linked inheritance, recombination

If two genes are close to each other on a chromosome, then the associated traits inherited together. In this case the Mendel's first and second laws are not valid and this phenomenon is called **linked inheritance** and the **two genes are in linkage**.

In meiosis, DNA replication is followed by two rounds of cell division to produce four daughter cells with half the number of chromosomes as the original parent cell. The two meiotic divisions are known as meiosis I and meiosis II. Before meiosis begins, during S phase of the cell cycle, the DNA of each chromosome is replicated so that it consists of two identical sister chromatids attached at a centromere (Figure 1). In meiosis I, homologous chromosomes pair with each other and can exchange genetic material in a process called chromosomal crossover or **crossing over or homologous recombination**. During this process the alleles of the two homologous chromosomes can exchange with each other. Recombination results in a new arrangement of maternal and paternal alleles on the same chromosome. Although the same genes appear in the same order, but some alleles are different (Figure 1.1)

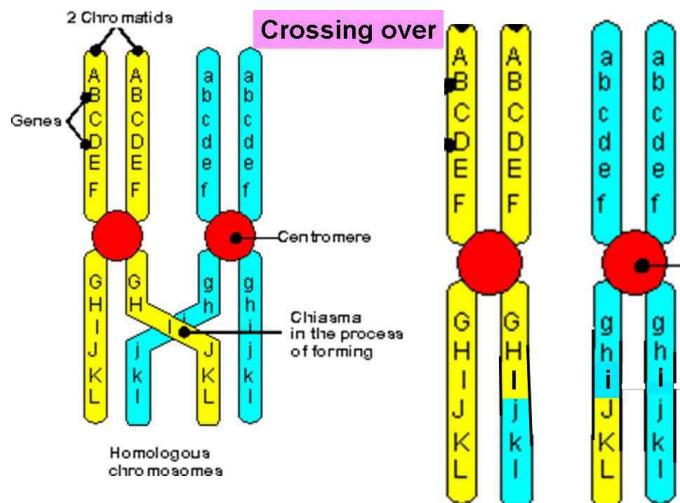


Figure 1. 1. During meiosis crossing over occurs between two homologous chromosomes. On two strands the order of the alleles changed. (<http://www.yourarticlelibrary.com/biology/notes-on-the-process-and-mechanism-of-crossing-over/12069/> 22/06/2015).

1.2.6. Mutation

Mutation is a permanent alteration in the DNA sequence, such that the sequence differs from what is found in most people. Mutations range in size; they can affect anywhere from a single DNA building block (base pair) to a large segment of a chromosome that includes multiple genes. The exact cause of a given mutation is generally unknown. Mutations can occur through the effect of a mutagenic agent or during the replication of the DNA. During replication, on average, one mutation occurs in every 100 millionth nucleotides. The effect of the majority of mutations does not appear phenotypically, but if it occurs on a functionally important position, then the consequence can even be serious illnesses, like cancer or inherited diseases. If the mutation occurs in a gene, then it is called **gene mutation**.

The chromosomes are fragile structures, if they are mutated it is called **chromosome mutation**. Chromosome mutation can be: duplication, deletion, translocation, inversion and insertion.

1.3. Basics of molecular biology

The **central dogma** in molecular biology can be described as "DNA makes RNA and RNA makes protein," a positive statement which was originally termed the sequence hypothesis by Crick (Figure 1.2). However, this simplification does not make it clear that the central dogma

as stated by Crick does not preclude the reverse flow of information from RNA to DNA, only ruling out the flow from protein to RNA or DNA.

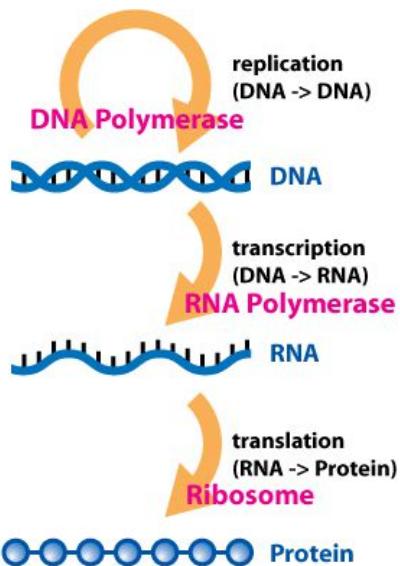


Figure 1.2. The central dogma of molecular biology.

https://en.wikipedia.org/wiki/Central_dogma_of_molecular_biology#/media/File:Central_Dogma_of_Molecular_Biochemistry_with_Enzymes.jpg 26/02/2016.

1.3.1. Some characteristics of the human DNA

The proteins coded by the DNA in our cells determine the structures and functions of the cells. If there is a mutation in the DNA, it can change the structure and function of the protein, which can have consequences on the function of the cell and can lead to diseases. Let's see the structure of the DNA in our cells.

The backbone of the DNA strand is made from alternating **phosphate** and **sugar residues** (Figure 1.3). The sugar in DNA is **2-deoxyribose**, which is a **pentose** (five-carbon) sugar. The sugars are joined together by phosphate groups that form phosphodiester bonds between the third and fifth carbon atoms of adjacent sugar rings. These asymmetric bonds mean a strand of DNA has a direction. In a **double helix** the direction of the nucleotides in one strand is opposite to their direction in the other strand: the strands are antiparallel. The asymmetric ends of DNA strands are called the 5' (five prime) and 3' (three prime) ends, with the 5' end having a terminal phosphate group and the 3' end a terminal hydroxyl group. One major difference between DNA and RNA is the sugar, with the 2-deoxyribose in DNA being replaced by the alternative pentose sugar **ribose in RNA**. The four bases found in DNA

are **adenine** (abbreviated A), **cytosine(C)**, **guanine (G)** and **thymine (T)**. These four bases are attached to the sugar/phosphate to form the complete **nucleotide**, as shown for adenosine monophosphate. The nucleobases are classified into two types: the **purines**, A and G, being fused five- and six-membered heterocyclic compounds, and the **pyrimidines**, the six-membered rings C and T.^[10] A fifth pyrimidine nucleobase, **uracil (U)**, usually takes the place of thymine in RNA and differs from thymine by lacking a methyl group on its ring. Uracil is not usually found in DNA, occurring only as a breakdown product of cytosine.

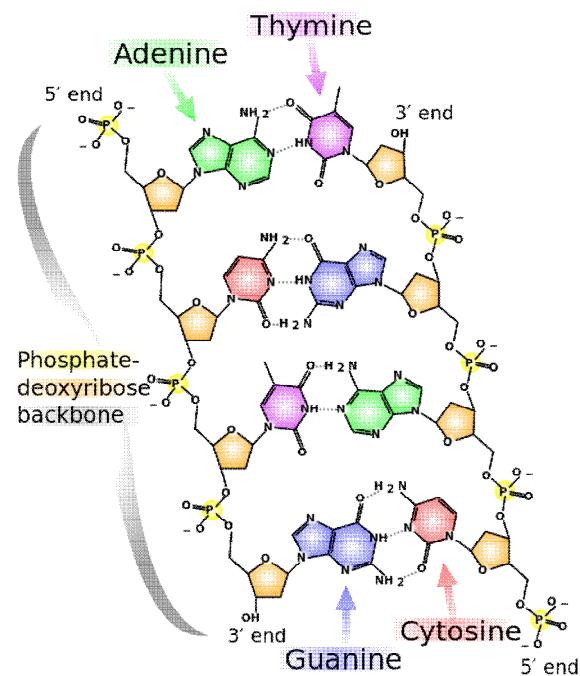


Figure 1.3. Structure of the DNA. https://en.wikipedia.org/wiki/DNA#/media/File:DNA_chemical_structure.svg

26/02/2016.

In a DNA double helix, each type of nucleobase on one strand bonds with just one type of nucleobase on the other strand. This is called **complementary base pairing**. Here, **purines form hydrogen bonds to pyrimidines**, with adenine bonding only to thymine in two hydrogen bonds, and cytosine bonding only to guanine in three hydrogen bonds. This arrangement of two nucleotides binding together across the double helix is called a **base pair**. As hydrogen bonds are not covalent, they can be broken and rejoined relatively easily. The two strands of DNA in a double helix can therefore be pulled apart like a zipper, either by a mechanical force or high temperature. As a result of this complementarity, all the information in the double-stranded sequence of a DNA helix is duplicated on each strand, which is vital in

DNA replication. Indeed, this reversible and specific interaction between complementary base pairs is critical for all the functions of DNA in living organisms.

A DNA sequence is called "**sense**" if its sequence is the same as that of a messenger RNA copy that is translated into protein. The sequence on the opposite strand is called the "**antisense**" sequence. Both sense and antisense sequences can exist on different parts of the same strand of DNA (i.e. both strands can contain both sense and antisense sequences).

In human cells DNA is in two compartments. **Nuclear DNA**, or nuclear deoxyribonucleic acid (nDNA), is DNA contained within a nucleus of the cell. Nuclear DNA encodes for the majority of the genome, with DNA located in mitochondria coding for the rest. Nuclear DNA adheres to Mendelian inheritance, with information coming from two parents, one male and one female. The other DNA containing compartment is the mitochondria. Mitochondria are cellular organelles within eukaryotic cells that convert chemical energy from food into a form that cells can use, adenosine triphosphate (ATP). In most multicellular organisms, including humans the **mitochondrial DNA (mtDNA) is inherited from the mother (maternally inherited)**.

Nuclear DNA and mitochondrial DNA differ in many ways. The structure of nuclear **DNA chromosomes is linear** with open ends and includes 46 chromosomes containing more than 3 billion nucleotides ($3.38 * 10^9$). **Mitochondrial DNA** chromosomes have closed, **circular** structures, and contain 16,569 nucleotides. Nuclear DNA is located within the nucleus of eukaryote cells and usually has two copies per cell while mitochondrial DNA is located in the mitochondria and contains 100-1,000 copies per cell. Nuclear DNA contains more than 20 thousands protein coding and more than 23 thousands non-coding genes. The mitochondrial DNA contains 37 genes. Of the 37 genes 13 are protein coding, 2 rRNA and 22 tRNA coding genes. The mutation rate for nuclear DNA is less than 0.3% while that of mitochondrial DNA is generally higher.

As mitochondria is the “powerhouse of the cell”, mutations of its DNA will effect on the power production processes of the cell, and will have serious consequences especially in tissues with large power need, like liver, neurons and muscle. As the mutation rate in the mitochondrial DNA higher, the mitochondrial diseases usually deteriorate with age, and can play also a role in the aging processes.

1.3.2. Replication

DNA replication is the process of producing two identical replicas from one original DNA molecule. This biological process occurs in all living organisms and is the basis for biological inheritance. DNA is made up of two strands and each strand of the original DNA molecule serves as a template for the production of the complementary strand, a process referred to as semiconservative replication. Cellular proofreading and error-checking mechanisms ensure near perfect fidelity for DNA replication.

DNA polymerases are a family of enzymes that carry out all forms of DNA replication (Figure 1.4).

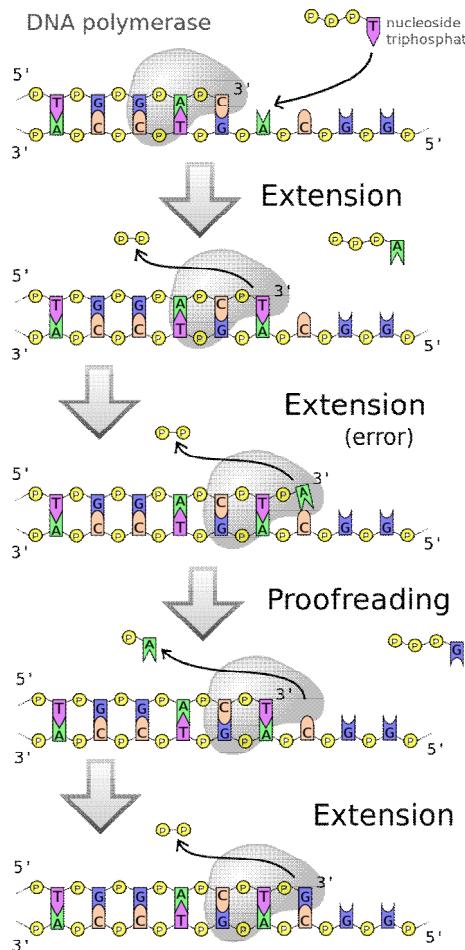


Figure 1.4. DNA polymerases add nucleotides to the 3' end of a strand of DNA. If a mismatch is accidentally incorporated, the polymerase is inhibited from further extension. Proofreading removes the mismatched nucleotide and extension continues.

https://en.wikipedia.org/wiki/DNA_replication#/media/File:DNA_polymerase.svg 26/02/2016.

1.3.3. Transcription

Transcription is the first step of **gene expression**, in which a particular segment of DNA is copied into RNA (messenger RNA or mRNA) by the enzyme RNA polymerase. RNA polymerase, and therefore the initiation of transcription, requires the presence of a core **promoter sequence** in the DNA. Promoters are regions of DNA that promote transcription and, in eukaryotes, are found at -30, -75, and -90 base pairs upstream from the transcription start site. **Transcription factors** are proteins that bind to these promoter sequences and facilitate the binding of RNA polymerase.

One strand of the DNA, the *template strand* (also noncoding or antisense strand), is used as a template for RNA synthesis. As transcription proceeds, RNA polymerase traverses the template strand and uses base pairing complementarity with the DNA template to create an RNA copy. Although RNA polymerase traverses the template strand from $3' \rightarrow 5'$, the coding (non-template or sense) strand and newly formed RNA can also be used as reference points, so transcription can be described as occurring $5' \rightarrow 3'$. This produces an RNA molecule from $5' \rightarrow 3'$, an exact copy of the coding strand (except that thymines are replaced with uracils, and the nucleotides are composed of a ribose (5-carbon) sugar).

1.3.4. Structure of the genes

The [structure of a protein coding gene](#) consists of many elements of which the actual protein coding sequence is often only a small part. These include DNA regions that are not transcribed as well as untranslated regions of the RNA.

Firstly, flanking the **open reading frame**, all genes contain **regulatory sequence** that is required for their expression. In order to be expressed, genes require a **promoter sequence**. The promoter is recognized and bound by **transcription factors** and RNA polymerase to initiate transcription.

Additionally, genes can have regulatory regions many kilobases upstream or downstream of the open reading frame. These act by binding to transcription factors which then cause the DNA to loop so that the regulatory sequence (and bound transcription factor) become close to the RNA polymerase binding site. For example, enhancers increase transcription by binding an activator protein which then helps to recruit the RNA polymerase to the promoter; conversely [silencers](#) bind repressor proteins and make the DNA less available for RNA polymerase.

The **transcribed pre-mRNA** contains untranslated regions at both ends which contain a ribosome binding site, terminator and start and stop codons. In addition, most eukaryotic open reading frames contain untranslated **introns** which are removed before the **exons** are translated. The sequences at the ends of the introns, dictate the splice sites to generate the final mature mRNA which encodes the protein or RNA product. **Splicing** is a modification of the nascent pre-mRNA transcript in which introns are removed and exons are joined. For nuclear encoded genes, splicing takes place within the nucleus after or concurrently with transcription.

Alternative splicing is a regulated process during gene expression that results in a single gene coding for multiple proteins. In this process, particular exons of a gene may be included within or excluded from the final, processed messenger RNA (mRNA) produced from that gene. Consequently the proteins translated from alternatively spliced mRNAs will contain differences in their amino acid sequence and, often, in their biological functions (see Figure 1.5). Notably, alternative splicing allows the human genome to direct the synthesis of many more proteins than would be expected from its 20,000 protein-coding genes.

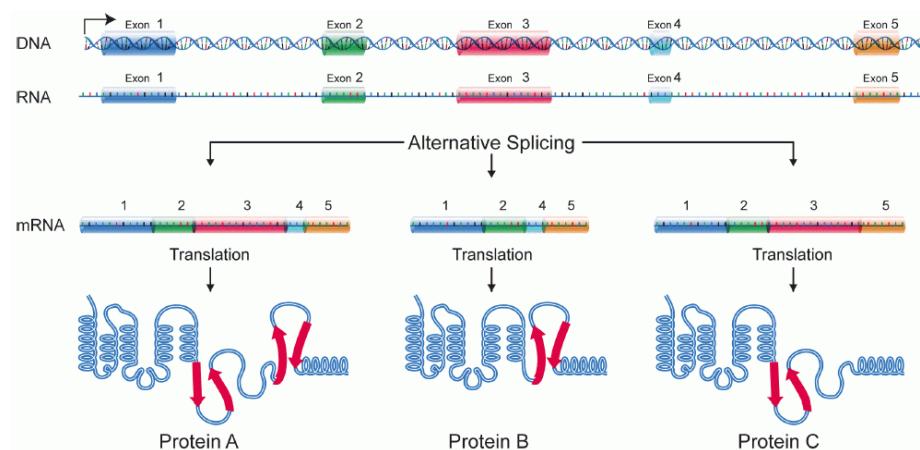


Figure 1.5. Alternative splicing. Because of this process, one gene can code for several proteins. https://en.wikipedia.org/wiki/Alternative_splicing#/media/File:DNA_alternative_splicing.gif 26/02/2016..

1.3.5. The genetic code

The **genetic code** is the set of rules by which information encoded within genetic material (DNA or mRNA sequences) is translated into proteins by living cells. Biological decoding is accomplished by the ribosome, which links amino acids in an order specified by mRNA, using transfer RNA (tRNA) molecules to carry amino acids and to read the mRNA

three nucleotides at a time. The genetic code is highly similar among all organisms and can be expressed in a simple table with 64 entries.

The code defines how sequences of these nucleotide triplets, called **codons**, specify which amino acid will be added next during protein synthesis. With some exceptions, **a three-nucleotide codon in a nucleic acid sequence specifies a single amino acid**. Because the vast majority of genes are encoded with exactly the same code, this particular code is often referred to as the canonical or standard genetic code, or simply *the* genetic code, though in fact some variant codes have evolved. For example, protein synthesis in human mitochondria relies on a genetic code that differs from the standard genetic code.

Translation starts with a chain initiation codon or **start codon**. The most common start codon is AUG, which is read as methionine.

The **three stop codons** have been given names: UAG is *amber*, UGA is *opal* (sometimes also called *umber*), and UAA is *ochre*. Stop codons are also called "**termination**" or "**nonsense**" **codons**. They signal release of the nascent polypeptide from the ribosome because there is no cognate tRNA that has anticodons complementary to these stop signals, and so a release factor binds to the ribosome instead.

Degeneracy is the redundancy of the genetic code. The genetic code has redundancy but no ambiguity. For example, although codons GAA and GAG both specify glutamic acid (redundancy), neither of them specifies any other amino acid (no ambiguity).

1.3.6. Translation

In [translation](#), mRNA is decoded by a ribosome to produce a specific amino acid chain, or polypeptide. The polypeptide later folds into an active protein and performs its functions in the cell. The ribosome facilitates decoding by inducing the binding of complementary transfer RNA (tRNA) anticodon sequences to mRNA codons. The tRNAs carry specific amino acids that are chained together into a polypeptide as the mRNA passes through and is "read" by the ribosome (Figure 1.6). The entire process is a part of gene expression.

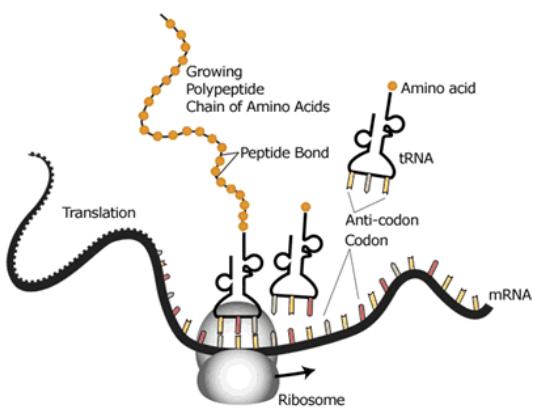


Image adapted from: National Human Genome Research Institute.

Figure 1.6. Translation.

After the translation some proteins are modified through a process called **post-translational modification**. Post-translational modification refers to the covalent and generally enzymatic modification of proteins during or after protein biosynthesis. Post-translational modifications can occur on the amino acid side chains or at the protein's C- or N-termini. They can extend the chemical repertoire of the 20 standard amino acids by introducing **new functional groups** such as phosphate, acetate, amide groups, or methyl groups.

Other forms of post-translational modification consist of cleaving peptide bonds, as in processing a propeptide to a mature form or removing the initiator methionine residue. The formation of disulfide bonds from cysteine residues may also be referred to as a post-translational modification. For instance, the peptide hormone insulin is cut twice after disulfide bonds are formed, and a propeptide is removed from the middle of the chain; the resulting protein consists of two polypeptide chains connected by disulfide bonds.

2. Transmission of genetic information

Valéria László

2.1. Cell cycle and regulation of cell cycle

In a given organism the genetic information (DNA) is transferred from cell to cell during the cell cycle. In the cell cycle, the cellular content is duplicated then it is halved. However, a distinction must be drawn between the nuclear and cytoplasmic events. DNA duplication (in chromatin form of DNA) and halving (in chromosome form of DNA) are very precisely regulated processes, resulting **two genetically identical cells**. At the same time the growing of the cytoplasm followed by division in two are less strictly regulated events of cell cycle.

The duplication of cellular ingredients occurs in **interphase**, that is divided into **G₁** (preduplication or preceding DNA duplication), **S** (DNA synthesis) and **G₂** (postduplication) phases. In **M-phase** the previously duplicated cellular content is separated, in **mitosis** the chromosomes, followed by **cytokinesis**, the division of cytoplasm.

Cell proliferation rate in an adult multicellular organism is variable. Moreover most of the cells are in so-called **G₀** phase, where there is **no cell division**, sometimes not even growth. The cells need extracellular stimuli, e.g. growth factors and / or adhesion to other cells or extracellular matrix in order to reenter G₁ phase.

In the cell cycle a very sophisticated control system (cell cycle control system) functions, whose essential components are the **cyclin-dependent protein kinases**, the **Cdk-s**. Cdk-s are activated by another protein family, by cyclins, the amount of which cyclically varies during the cell cycle. Beside cyclins, the activity of cyclin-dependent kinases is regulated by other factors, too. These factors include activating and inhibiting Cdk kinases which phosphorylate Cdk-s, resulting Cdk activation and inhibition respectively. Phosphate residues are removed by phosphatases, modifying Cdk activity. According to their names, cyclin-dependent kinase inhibitors inhibit Cdk activity. The amount of all the proteins mentioned before may be regulated via transcriptional and translational level and by proteasomal degradation, followed by ubiquitination. All these

together allow a highly organized, complex but gentle control of the cell cycle. The cyclin-dependent kinases, the main actors of cell cycle control system, operate the cell cycle through phosphorylation of many different target proteins. Recently in addition to cyclin-dependent kinases the role of some other kinases (e.g. Polo, Aurora etc.) was found.

The phases of cell cycle are not interchangeable, they have to follow each other in a strict order. Operation of **checkpoints** in the cell cycle ensures to give rise to genetically identical cells by cell cycle (Figure 2.1).

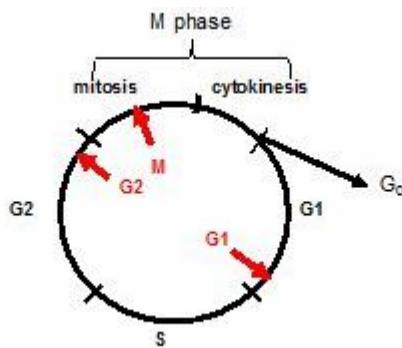


Figure 2.1. Phases (G₁, S, G₂, M) and checkpoints (G₁, G₂, M) of cell cycle. Cell cycle control system allows to overstep checkpoints if the conditions are suitable for the cell to proceed to the next phase.

The main checkpoints are the following: **G₁ checkpoint** (in higher eukaryotes it is referred to as **restriction point**), where first of all the integrity of DNA is checked, operates at the end of the G₁ phase. The second checkpoint is at the end of G₂ phase, it is the **G₂ checkpoint**, where the accuracy and integrity of DNA is monitored. Finally, the function of **M checkpoint**, in the metaphase of mitosis is to ensure the appropriate attachment of all chromosomes to the microtubules of the mitotic spindle before the duplicated chromosomes are separated. And now let us see a brief summary of multicellular (mammalian) cell cycle and the regulation.

2.1.1. G₀ - G₁ transition

In an adult multicellular organism most cells do not divide, they are found in a special phase, G₀ phase. G₀ phase cells lack functional cyclins and cyclin-dependent kinases, the main cell cycle regulators. If proliferation is necessary, these G₀ phase cells have to return into the cell cycle, essentially have to pass G₁ checkpoint or restriction point. It is induced by growth

factors or extracellular matrix components initiating transcription and translation of D cyclin and reduction of Cdk inhibitors by stimulating their proteasomal degradation. These Cdk inhibitors: p16, p15, p18 and p19 specifically inhibit Cdk4 and Cdk6 by preventing the binding of activating D cyclin, and also the activity of Cdk-cyclin complex. The main target of active Cdk4/6-D cyclin complex is pRb (Rb stands for retinoblastoma, a malignant disease of the retina caused by the mutation of pRb encoding gene), p107 and p130 proteins. The phosphorylation of these proteins causes conformational changes and they release E2F transcription factors. And it is the turning point in G0-G1 transition, because E2F transcription factors induce the transcription of several S-phase specific genes, such as E cyclin, A cyclin, thymidine kinase, DNA polymerase etc. E cyclin activates Cdk2 whose main target, similarly to the Cdk4/6-D-cyclin is Rb protein, the phosphorylation of which is enhanced (positive feedback). Cdk2 has another activator, A cyclin, their complex is essential in S phase initiation (Figure 2.2).

Disadvantageous environmental effects, e.g. hypoxia (excessive proliferation of cells may result not sufficient blood flow) or DNA damages activate G₁ checkpoint machinery and it will stop the cell cycle. The amount and activity of **p53** is increased which in turn induces the transcription of a Cdk inhibitor protein, p21. p21 is a general Cdk inhibitor, hence it inhibits all Cdk-cyclin complexes: Cdk4 / 6 - D cyclin, Cdk2- E cyclin and Cdk2- A cyclin, so the cell cycle is halted and the cell may not enter S phase. This general Cdk inhibitor family has two other members, p27 and p57. These proteins prevent the duplication of damaged DNA, suspend the cell cycle, allowing error correction. Briefly, their activity prevents the cell cycle resulting genetically different cells (Figure 2.2).

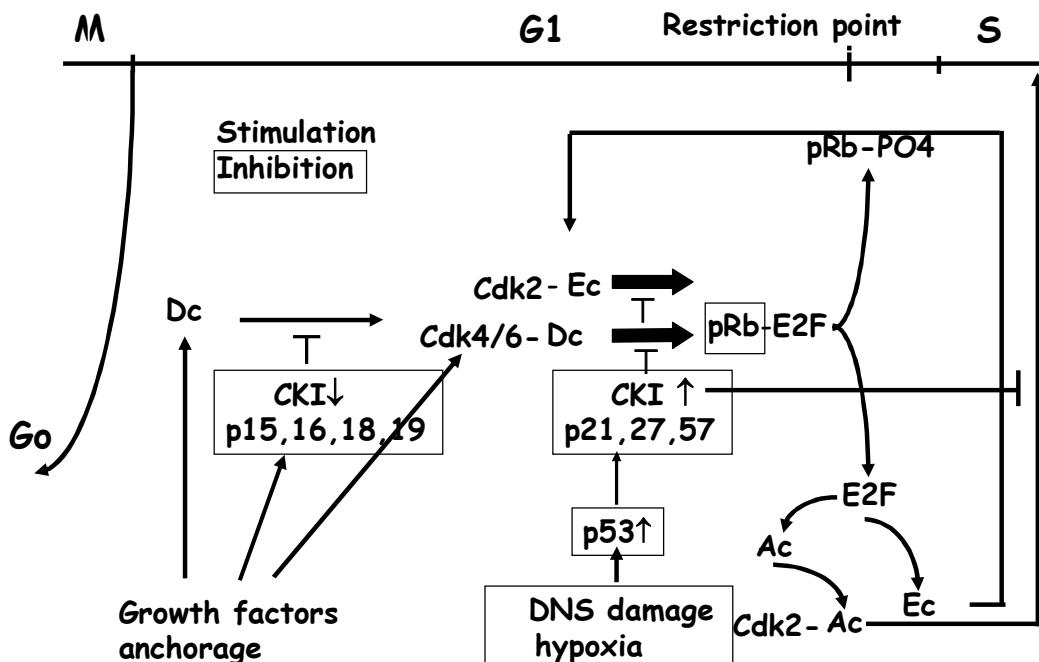


Figure 2.2. Summary of G₀–G₁ transition

The Cdk inhibitor encoding genes are **tumor suppressor genes** whose mutations in homozygote form (recessive) are the main contributors of tumor development. The most well-known tumor suppressor gene species are p53 and pRb encoding genes. About half of the tumors lack functional p53. The genes encoding cell cycle stimulating proteins (Cdk-s, cyclins, growth factors and many others) are **protooncogenes**. Their mutation in heterozygote form (dominant) is also involved in tumor development.

2.1.2. G₁ – S transition, S-phase

The main S phase event is the **DNA duplication**, the replication. Since eukaryotic cell DNA is much higher than prokaryotic, the replication starts simultaneously at several sites, called **origos**, and occurs in both directions. Initiation proteins associate with origos where DNA unwinds, followed by the attachment of further components of replication complex. As throughout the whole cell cycle cyclin-dependent kinases play major role in G₁-S transition, too. **Cdk2-E cyclin complex** activation requires the degradation of the Cdk inhibitor p27, which step is initiated by an ubiquitin ligase, **SCF** (Skp-Cullin-F-box protein). Finally the activated Cdk2 (and another protein kinase, Cdc7) phosphorylate some, not exactly known members of the replication complex. This effect fulfills another role, too, namely it prevents

the formation or the binding of new initiation complexes, and hence the DNA is replicated only once.

In DNA replication both strands of DNA serve as template, and appropriate nucleotides built in according to the complementarity of the nucleotide chain. As it was demonstrated by the experiments applying radioactive isotope labelled monomers, **DNA synthesis is semi-conservative**, since after the replication both DNA molecules have one old and a newly synthesised strand. Since the sequence of DNA strand unambiguously determines the sequence of complementary DNA strand, the arising two DNA molecules are equal.

2.1.3. *G₂ – M transition*

The regulation of G₂ - M transition is better known than that of G₁ - S transition. The M-phase is triggered by **MPF (M-phase or Mitosis Promoting Factor)**, that is a complex of B cyclin and Cdk1. After the binding of these proteins post-translational modifications are required for the final activation. Cdk1 component of the complex is the substrate of two kinases, one is an activating kinase which adds a phosphate group to a tyrosine, the other is an inactivating kinase which phosphorylates a threonine residue of the protein. The latter is removed by a phosphatase (product of a gene belonging to Cdc25 gene family), and this is the last step in MPF activation (Figure 2.3). But all these events will only happen if G₂ checkpoint machinery finds DNA undamaged and correctly replicated.

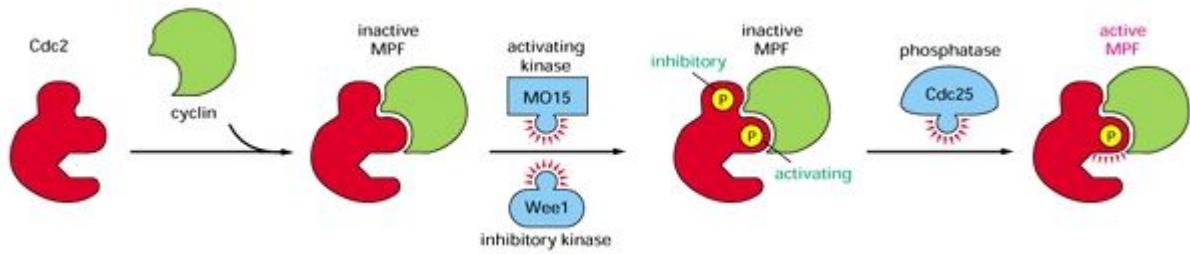


Figure 2.3. MPF activation. B cyclin binds to Cdk1 which is phosphorylated by an activating and an inactivating kinase. Inactivating phosphate group is cleaved by a phosphatase resulting an active MPF.

Source:

<http://www.ncbi.nlm.nih.gov/books/NBK28366/figure/A4636/?report=objectonly>;

29/07/2013.

MPF has numerous substrates, first of all it activates **Cdc25 protein**, thus by a positive feedback control more and more MPF is activated. In mammalian cells there are three phosphatases: Cdc25A, B and C, at this point of cell cycle regulation, the C type operates.

Then, MPF triggers M-phase through the phosphorylation of further target proteins, like **lamin A, B and C**, components of nuclear lamina, a structure attached to the inner nuclear membrane. It results disintegration of nuclear membrane.

MPF indirectly inhibits **actomyosin ATP-ase** activity causing rearrangement of microfilaments and consequently rounding of the cell and also inhibiting premature cytokinesis.

One of the major events, the chromosome condensation is triggered also by MPF, through the phosphorylation of **condensins, H1 and H3 histones**.

Phosphorylation of **MAP-s** (microtubules associated proteins) changes the arrangement of microtubule system and induces mitotic spindle formation needed for chromosome separation.

In G₂-M phase transition the APC (Anaphase Promoting Complex) having role later in metaphase-anaphase transition is indirectly activated by MPF.

2.1.4. *M-phase*

The M-phase is a complex process of successive steps, a series of events, used to be divided into **mitosis** and **cytokinesis**. In the first half of the M-phase, in mitosis the

doubled DNA divides in two, followed by the separation of cytoplasm, by the phase of cytokinesis.

In mitosis the following phases are distinguished:

Prophase. In the nucleus the nuclear chromatin gradually changes to chromosomes by the maximal condensation of DNA. Since before the M-phase the DNA has been replicated, each chromosome comprises two chromatids (sister chromatids). In the cytoplasm the centrosome, which also has been doubled in interphase, splits into two and move to opposite poles of the cell, and organize the mitotic spindle composed of microtubules.

Prometaphase. Nucleolus disappears, the chromosome development continues. The nuclear membrane disintegrates, too. Kinetochore microtubules binding kinetochore protein complex associate to the centromere region of each chromatid.

Metaphase. The chromosomes are arranged in the equatorial plane of the cell by the help of kinetochore microtubules. Kinetochore regions face the two poles of the cell and the kinetochore microtubules bind to sister chromatids of a chromosome from opposite direction.

Anaphase. Sister chromatids of chromosomes split and move toward the poles of the cell. In the first half of anaphase (anaphase A) the kinetochore, later in the second half of anaphase (anaphase B) the polar microtubules operate. It is the shortest phase of mitosis.

Telophase. Kinetochore microtubules disappear, nuclear membrane is reorganized around the chromatids at the cell poles. Chromosomes decondense, they become chromatin. Nucleoli are reformed. Polar microtubules lengthen further the cell.

The mitosis, the division of nuclear content is followed by

Cytokinesis. The separation of the cytoplasm begins in the late anaphase and is completed after the telophase. In the middle of the cell, perpendicular to the axis of the mitotic spindle cleavage furrow appears which gradually deepens and thus the connection between the two half cells narrows. The overlapping region of polar microtubules makes so-called midbody. Finally, the cytoplasm completely splits.

Let us see in more detail the processes listed above.

2.1.4.1. Chromosome structure

In M-phase the long eukaryotic DNA molecules have to be packed in small chromosomes to be able to accurately halve without breaks. Meanwhile, the original length of the DNA (several cm) is reduced by ten thousands fold (few μm). The molecular mechanism of this packaging is still not known in detail. The major points of a widely accepted model are described below (Figure 2.4).

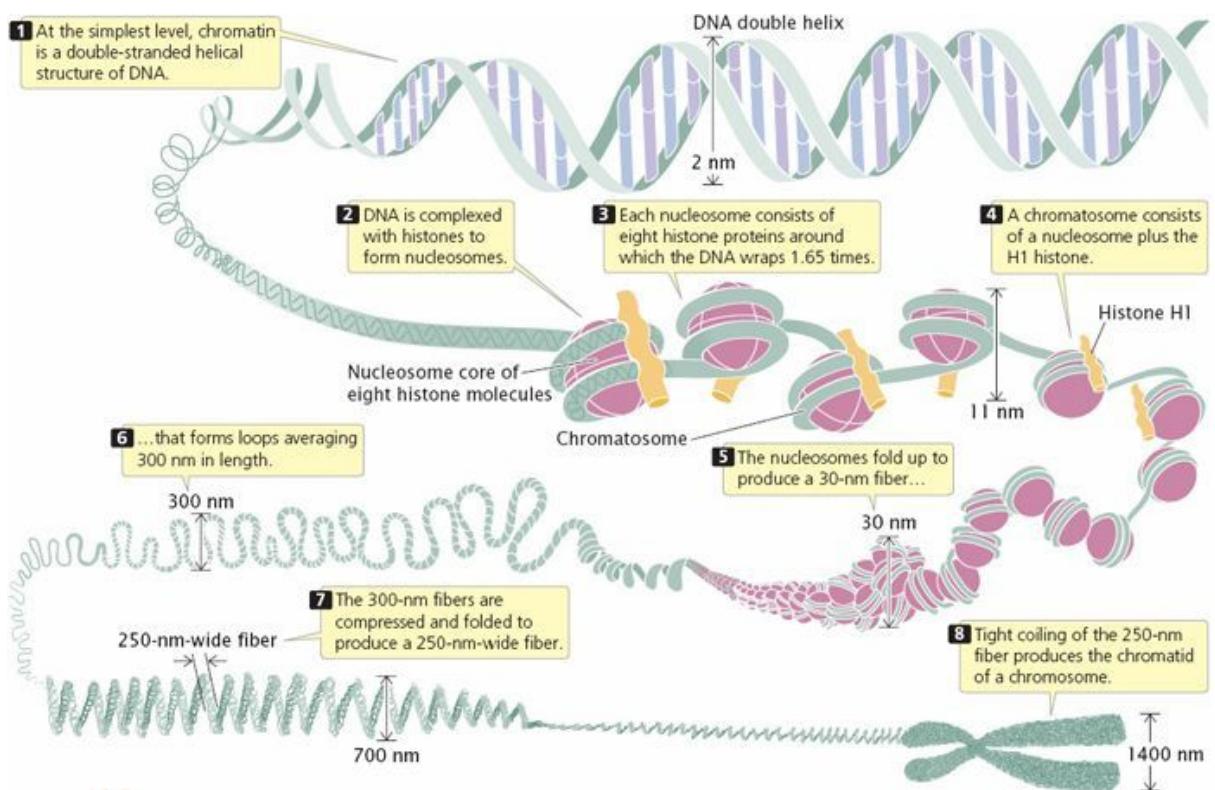
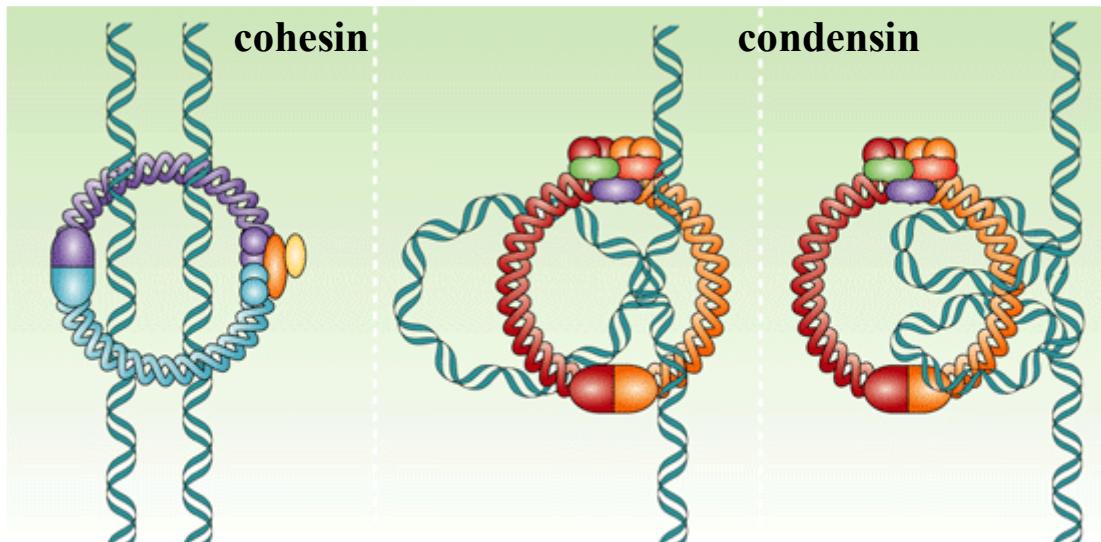


Figure 2.4. From the DNA to the chromosome

Source: <http://www.nature.com/scitable/topicpage/eukaryotic-genome-complexity-437>; 20/02/2013.

Two nm wide DNA double helix wraps the octamers of histones (2 of each H2A, H2B, H3 and H4 histone molecules) forming nucleosomes, disc-like structures connected by the continuous DNA molecule. It is called **nucleosomal structure** having a diameter of 11 nm. H1 histone folds six nucleosomes in one plane to give a diameter of 30 nm fiber called chromatin or solenoid. The chromatin fiber is attached to a protein scaffold and forms loops. These loops are the basic unit of replication and transcription, and this structure is 300 nm wide. Finally, it is further compressed and folded to produce the chromatids of 1400 nm wide metaphase chromosome (Figure 2.4). The final step of chromosome

condensation is induced by the MPF activated condensins. There are two protein complexes of similar structure influencing different DNA functions: the **condensins** and the **cohesins**. They are composed of different SMC (structural maintenance of chromosomes) proteins having ATPase activity and regulatory functions, all associate in a ring-like structure (Figure 2.5).



Nature Reviews | Genetics

Figure 2.5. Structure of cohesin and condensin

Source: http://www.nature.com/nrg/journal/v4/n7/box/nrg1110_BX3.html; 19/02/2013.

Metaphase chromosome has very characteristic morphological structure. As the DNA is doubled in the S phase, chromosome comprises two **sister chromatids**. After DNA synthesis, the DNA molecules are held together by the ring-like cohesin complexes. Much of this cohesins detaches during the prophase, and at the end of metaphase it is found only at the primary constriction of chromosomes specified as **centromere** region. This pericentromeric cohesin is cleaved in early anaphase allowing the separation of chromatids. Chromosomes are usually classified according to the location of the centromere region (see Chapter 4, Cytogenetics). During prophase and prometaphase a special three-layer plate of protein structure called **kinetochore** associates to the centromeres of chromosomes. Beside many other proteins kinetochore contains both dynein and kinesin-type motor proteins, and the role of it is to bind kinetochore microtubules (about 30–40/sister chromatids). In scleroderma which is an autoimmune disease, patients produce antibodies against some of the kinetochore proteins.

The centromere divides the sister chromatids into two **arms**, the ends of the arms are called **telomeres**. Loss of telomeres makes the chromosomes unstable (Figure 2.6).

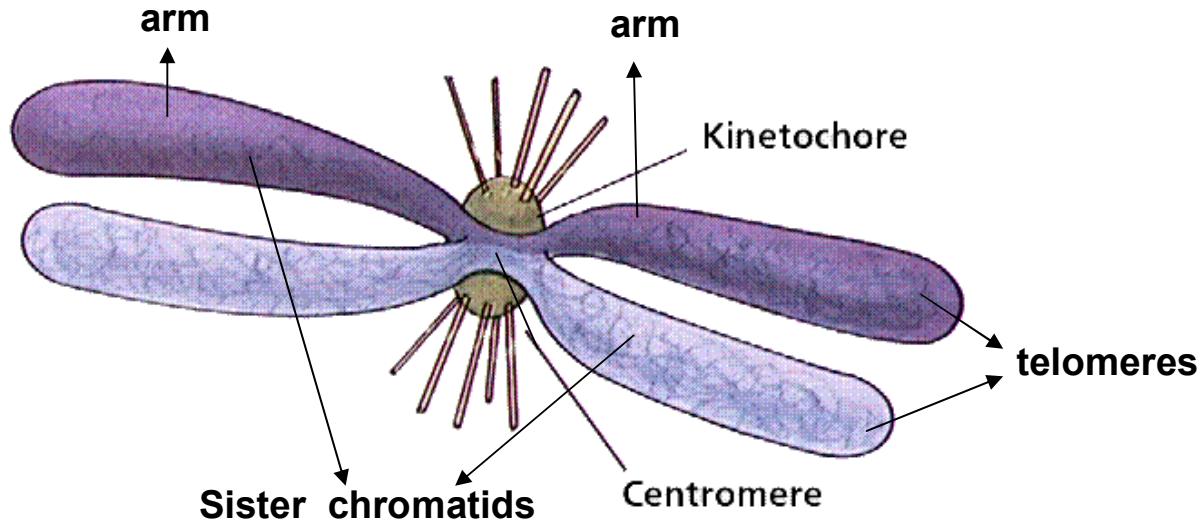


Figure 2.6. Eukaryotic chromosome

Source: <http://www.emc.maricopa.edu/faculty/farabee/biobk/biobookmito.html>;
20/02/2013.

There are 5 pairs of human chromosomes having not only primary, but secondary constriction or **NOR** (nucleolar organizer region) too, which contains a high number of copies of the large (45S) rRNA gene.

1.1.4.2. Disappearance and re-formation of nuclear envelope

As it was mentioned before, the lamins of nuclear lamina attached to the inner surface of nuclear envelope are phosphorylated by MPF causing the dissociation of nuclear membrane into vesicles. Lamin B remains in the membrane of vesicles, but lamin A and C are found in soluble form in the cytoplasm. The highly organized nuclear pores are also decomposed. At the end of mitosis, in the telophase, the phosphatases are activated and dephosphorylate the lamins. The reformation of nuclear envelope begins on the surface of chromosomes. They move closer to each other, and the membranes fuse and the pores are also reorganized. Finally the chromosomes decondense to chromatin.

2.1.4.2. Structure and role of mitotic spindle

The components of mitotic spindle are the **centrosomes** and the **microtubules**. In human cells, the major microtubule organizer center (MTOC) is the centrosome; in interphase cells it is located usually near to the nucleus. The structure of centrosome is the following: in the center there are two perpendicular cylindrical bodies (the **centrioles**), which are connected by proteins at their bases. The centrioles are made of 9×3 microtubules in windmill-like arrangement. Around them an amorphous, unstructured material, the **pericentriolar matrix** is located, in which numerous different proteins are found. The microtubules grow out from the pericentriolar matrix in star-like manner, this region is called **aster** (Figure 2.7). Minus ends of microtubules face the centrioles, their plus ends face outward. The microtubules are organized, the tubulin heterodimer polymerization is induced by a special subtype of tubulin, found in **γ -tubulin rings** in the pericentriolar matrix.

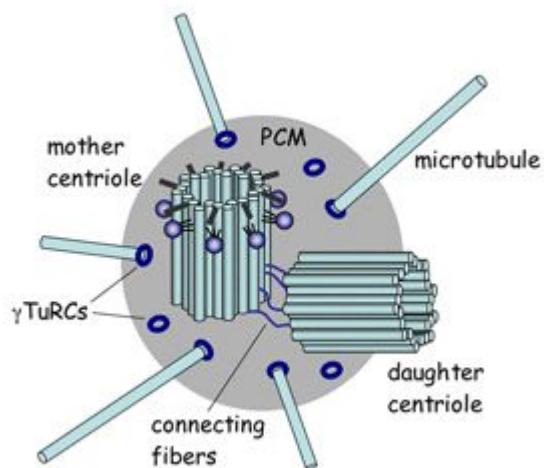


Figure 2.7. Schematic figure of centrosome. There are two centrioles in the center, surrounded by pericentriolar matrix (PCM) which is the nucleation site of microtubules (γ TuRC = γ tubulin ring complex).

Source: <http://www.irbbarcelona.org/index.php/es/research/programmes/cell-and-developmental-biology/microtubule-organization>; 19/02/2013.

To produce genetically identical cells through the cell cycle not only the accurate replication and separation of DNA, but the precise **duplication and division of centrosome** are needed. If there is no centrosome duplication, there is no bipolar mitotic spindle, no division and the chromosomes are not able to separate in two. However, if

centrosome is repeatedly duplicated, more poles are made in the cells and the chromosomes are unevenly distributed between the daughter cells (see causes of atypical divisions). In late G₁ phase the centrioles slightly move away from each other, and in the S-phase perpendicularly to the original one the development of new centrioles begin. In late G₂ and early mitosis the two pairs of centrioles are separated from each other, migrate to the two poles of the cell by the help of microtubule system and motor proteins. At the poles the new centrosomes nucleate the microtubules of mitotic spindle (Figure 2.8).

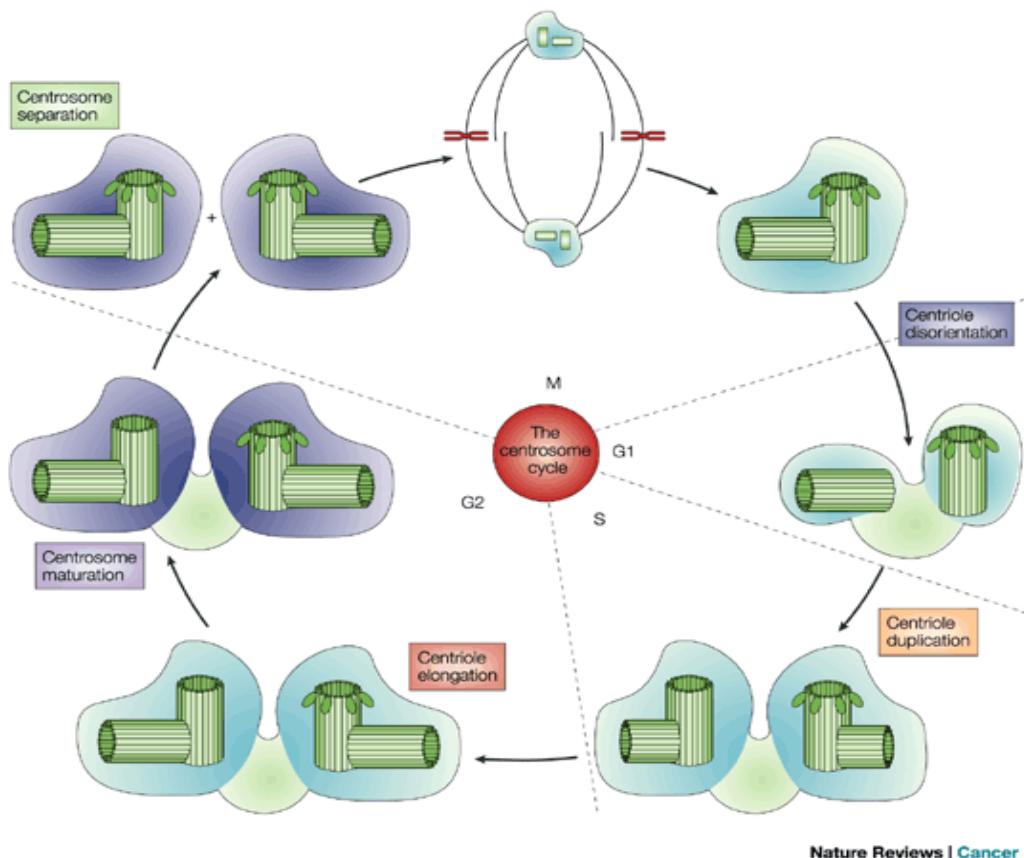


Figure 2.8. Duplication and separation of centrosome

Source: http://www.nature.com/nrc/journal/v2/n11/box/nrc924_BX3.html; 19/02/2013.

Mitotic spindle organization also needs the activation of MPF. At the beginning of division the MAP-s, microtubule associated proteins are phosphorylated by MPF, which in turn changes the characteristic interphase microtubule arrangement and induces the development of mitotic spindle. In interphase there are few, long and relatively stable microtubules. Oppositely the mitotic spindle is characterized by many, short and highly dynamic microtubules.

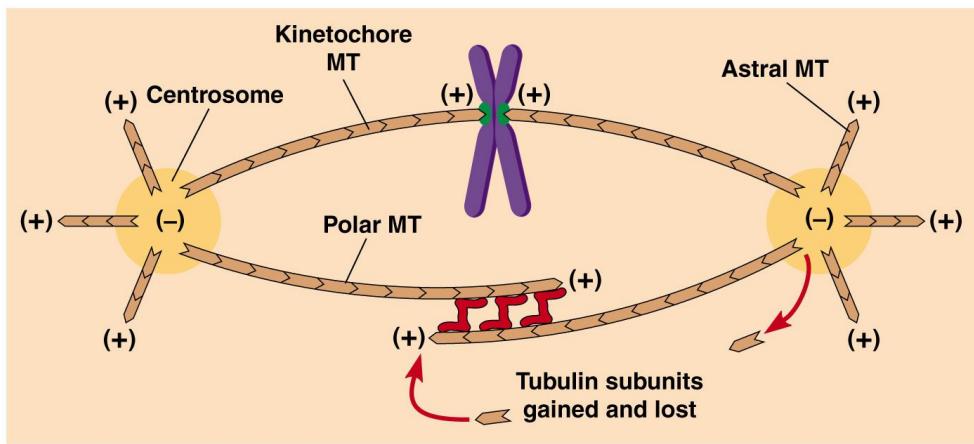
In prophase many, dynamic microtubules grow in all directions back away from the centrosomes. The attachment to any structure by their + end stabilizes the microtubules. The microtubules growing from different poles may bind to each other giving rise to the partly overlapping **polar microtubules**. In the overlapping region + end motor proteins are found, which stabilize the polar microtubules and are also needed to push apart the two poles in anaphase B.

In prometaphase after the disruption of nuclear envelope the microtubules randomly may bind not only to each other, but also to the chromosomes. It has been demonstrated that the chromosomes by their kinetochore region move on the microtubules. This movement is a kind of sliding mediated by dynein motor protein.

Around the centrosome the effect of a special unidentified force, referred as polar wind is noticed. It means that from the poles of the cell all bigger particles are excluded, probably due to the mechanical effect of intensively growing microtubules. As they grow, they may push any particles, e.g. chromosomes. At the same time the + ends of microtubules randomly growing from the two opposite poles may bind to the two different kinetochores of chromosomes. These microtubules are the **kinetochore microtubules**.

In the metaphase the chromosomes are arranged in the equatorial plane by the help of kinetochore microtubules. This arrangement is not static, the chromosomes are oscillating according to the dynamism of microtubules. At the same time the length of microtubules is constant, because the rate of polymerization on both ends is the same.

The third type of microtubules are the **astral** ones, they grow from the centrosome toward the plasmamembrane. The role of them is not clear (Figure 2.9).



© 2012 Pearson Education, Inc.

Figure 2.9. Structure of mitotic spindle

Source: http://www.mun.ca/biology/desmid/brian/BIOL2060/BIOL2060-19/19_25.jpg; 19/02/2013.

2.1.4.3. *Metaphase – anaphase transition*

In metaphase–anaphase transition the **anaphase promoting complex (APC)** has to be activated. APC is a specific **ubiquitin ligase**, an enzyme which binds ubiquitin to its substrate proteins, targeting them to proteasomal degradation. One of the main substrates is the separase inhibitor, the **securin**. Destruction of securin activates **separase**, which in turn cleaves the sister chromatid binding cohesin from the chromosomes, allowing the pulling of sister chromatids to the poles by kinetochore microtubules. The other substrate is the **B cyclin** and its degradation inactivates MPF. Inactivation of MPF triggers the completion of M-phase; disappearance of mitotic spindle, chromosome decondensation, reorganization of nuclear envelope and the cytokinesis, too.

In the metaphase **M (spindle)-checkpoint machinery** operates; its function will be discussed below. Briefly its significance is to ensure the precise halving of chromosomes, pulling one sister chromatid to one pole and the other one to the other pole.

In the following figure (Figure 2.10) the different microtubule-kinetochore attachments are shown. The accurate segregation of sister chromatids is ensured by amphitelic attachment. If the attachment is not correct, the kinetochore microtubules are not bound to the sister chromatids of a chromosome from opposite poles, there are free kinetochores, the APC remains inactive and the mitosis is stopped in metaphase until the attachment is corrected. The colchicine causing disruption of microtubules stops the division in this way. If the chromosomes are arranged correctly in the metaphase, APC is activated and the cell may step the anaphase.

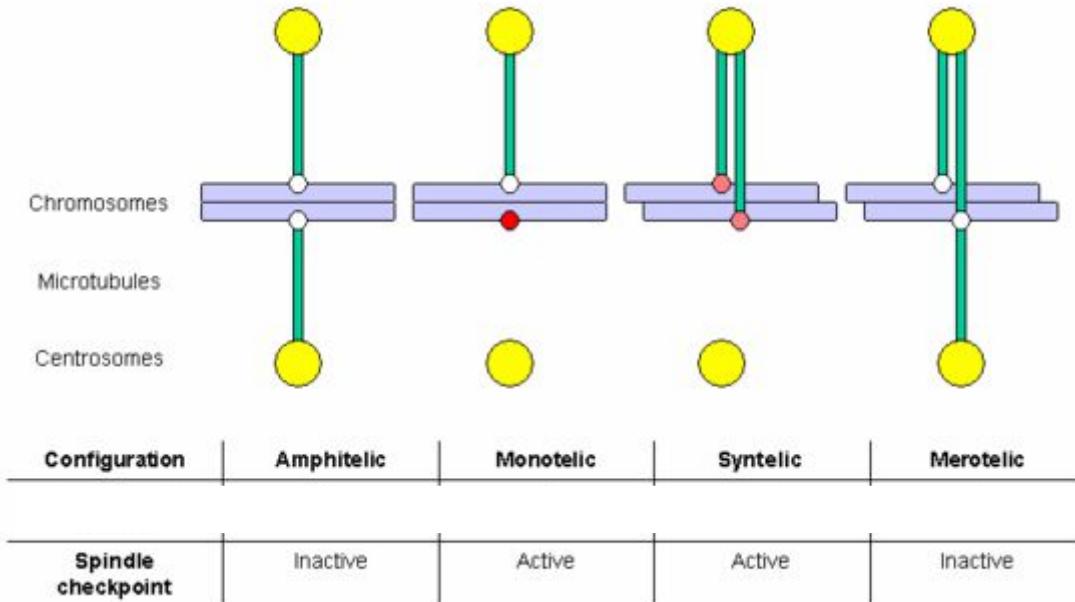


Figure 2.10. Attachment of kinetochore microtubules and kinetochor

Source: http://en.wikipedia.org/wiki/File:MT_attachment_configuration-en.png; 03/07/2013.

In anaphase two subphases are distinguished according to the type of microtubules operating. In **anaphase A** the kinetochore microtubules shorten by depolymerization on both ends. The depolymerization is coupled to the movement of the chromosome, the sister chromatids are pulled to the opposite poles of the cell. The poles are further separated in **anaphase B** by the help of polar microtubules which are growing, but the overlapping region remains constant. Consequently the sister chromatids are further separated (Figure 2.11).

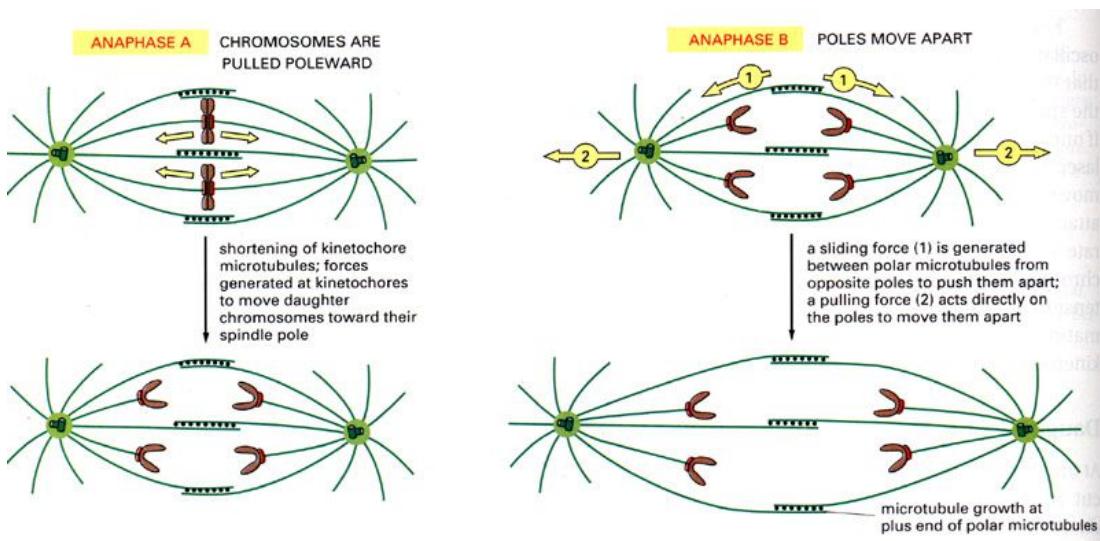


Figure 2.11. Anaphase A and B

Source : <http://greatcourse.cnu.edu.cn/xbfzswx/wlkc/kcxx/11English.htm> ; 20/02/2013.

2.1.5. Cytokinesis

The division of cytoplasm is carried out by other components of the cytoskeleton than the separation of chromosomes, but the two cytoskeletal systems are not independent. The site of cytoplasmic cleavage is denoted by the mitotic spindle. Asymmetric positioned mitotic spindle results in asymmetric cytokinesis, in different sized cells. In late anaphase (anaphase B), after the migration of two sister chromatids to the poles beneath the plasma membrane, perpendicularly to the mitotic spindle axis, a **contractile ring** composed of **actin** and **myosin II filaments** is formed. The regulation of contractile ring development is not exactly known, but the role of kinases and monomeric G-proteins is suspected. The sliding of actin and myosin filaments on each other eventually leads to the progressive cleavage of the cell. Finally, below the contractile ring, the two cells are only connected by the so-called midbody. The new cell membranes develop by the fusion of vesicles, which transport probably takes place along the microtubules of midbody.

2.1.6. Operation of cell cycle checkpoints

The checkpoints and their significance have been mentioned in the introduction of the chapter. Here the operation of the main checkpoints, G₁, G₂ and M checkpoints is briefly discussed. The **checkpoint machinery** is composed of three main components which make up a cascade. The **sensor** detects the errors occurring in DNA molecules, this signal is transmitted to a **transducer**, which finally triggers **effector** proteins.

In G₁ and G₂ checkpoints the DNA damages, for example the single-stranded DNA, or double strand breaks are recognized by certain proteins which activate transducers. The activated transducers, which are protein kinases (not cyclin dependent) phosphorylate, e.g. in G₁ checkpoint the p53 protein. This phosphorylation stabilizes p53 and stops the cell cycle. So in G₁ checkpoint the effector protein is p53. In G₂ checkpoint Cdc25 is the effector molecule. In the case of DNA damage the transducer inactivates Cdc25 by phosphorylation, that is why the inactive Cdc25 is not able to activate MPF by the cleavage of inactivation phosphate group, so the cell cycle is stopped before M-phase.

In the M checkpoint the sensor proteins bind to free kinetochores of chromosomes. However, these proteins recruit a protein which is required for the APC function, so in the case of free kinetochores APC is not functional; the cell is retained in metaphase.

Obviously the precise operation of checkpoint machinery is more complex than it has been described before; the details are being discovered nowadays. Anyway its accurate function is essential to give rise to genetically identical cell by cell cycle.

The failure of sufficient regulation or checkpoint machinery may result in **atypical divisions**. Although some of these atypical divisions, depending on the species and cell type is not necessarily abnormal, but the majority of them is characteristic of tumor cells. Of course, from each of atypical proliferation genetically diverse cells are originated.

In **endomitosis** the nuclear envelope remains intact, therefore the amount of cellular DNA content increases. In parallel with the size of the nucleus and the whole cell enlarges, too, so **giant cells** are formed. The sister chromatid separation inside the intact nuclear causes the increase in chromosome number of the cell, these are referred to as **polyploid cells**. If sister chromatids remain together, it leads to the formation of **giant chromosomes** (polytene) composed of many sister chromatids instead of two.

Failure of cytokinesis results in giant cells, too, but these cells have more nuclei.

Many division abnormalities are caused by the mitotic spindle defects. The normal division is bipolar due to the precise duplication and separation of centrosomes. Abnormalities either in the duplication or the separation of centrosomes may cause so-called **multipolar divisions, depending on the number of poles**: tri-, tetra-, etc. polar divisions.

The consequence of **non-disjunction** of sister chromatids of a chromosome can be easily calculated, it leads to the change of chromosome number (aneuploidy) in both cells,

in one of them one more and in the other one less chromosome is found. The reason for such a defect is the syntelic or monotelic kinetochore-microtubule attachment (Figure 2.10).

However, the merotelic attachment may result in **bridge formation** (or anaphase bridge, because it becomes visible in anaphase). First the sister chromatid to which microtubules bind from both poles make a kind of bridge but later it more probably breaks. The breakage of chromosomes leads to structural chromosomal abnormalities. The chromosome fragment without centromere is excluded from the nucleus and makes a so-called micronucleus in the cytoplasm. This phenomenon is used in mutagenicity assays to detect compounds which cause/increase chromosomal breakages.

2.2. Chromosome territories

While chromosomes appear as condensed, elongated structures during the process of cell division, for most of the lifetime of a cell, chromosomes do not look anything like this. So, the question is what chromosomes look like in the nucleus of a cell between cell divisions?

It was the FISH technique that has revealed how chromosomes look like during the interphase of the cell cycle. It turned out that among the several models suggested previously, the **chromosome territory model** proved to be the right one. Although there are still a lot of open questions a few topics have already been clarified:

- Chromosome territories are **irregular in shape** but typically about 1 to 2 micrometers in diameter, and they consist of smaller subdomains.
- Chromosome territories border each other closely; the neighboring **chromosomes can invade the territories of each other** and intermingle at their peripheries.
- Chromosome territories are known to be **arranged radially** around the nucleus. Chromosome territories are **semiconserved from parent to daughter cell** during cell division, with locations in the daughter cell similar to those in the parent cell.
- Patterns of chromosome arrangement are **specific to both cell type and tissue type**.
- The organization of interphase chromosome territories changes during differentiation, quiescence and senescence.
- The **gene-rich chromosomes are present in the nuclear interior** while **gene-poor chromosomes are located at the periphery**.
- Regions of DNA that are within the nuclear interior show higher transcriptional activity.

- Group of genes with common tasks are in the vicinity of each other and are under coordinated regulation. Loops may carry genes to even very remote sites in the nuclear space for coregulation in an **expression hub** (Figure 2.12).
- Chromosome territories are also dynamic structures, with genes able to relocate from the periphery towards the interior once they have been switched on. In other cases, genes may move in the opposite direction, or simply maintain their position.
- Chromosome territories **can reposition in diseases**, which might provide novel insights into disease mechanisms and why genes are incorrectly expressed in diseases.

In summary, this functional compartmentalization of the chromosomes makes possible:

- Coordinated functioning of the DNA packed in the nucleus; synchronized switching off and on of hundreds/thousands of genes
- Coordinated functioning of genes “working for common aims”
- Fine-tuning of the regulation of gene-gene interactions
- Organized transportation of gene products
- Separation of active genes from the inactive ones

See more details (source): http://www.mechanobio.info/topics/synthesis/go-0006323/03_go-0006323

<http://www.nature.com/scitable/topicpage/chromosome-territories-the-arrangement-of-chromosomes-in-3025>

Inter- and intrachromosomal loops

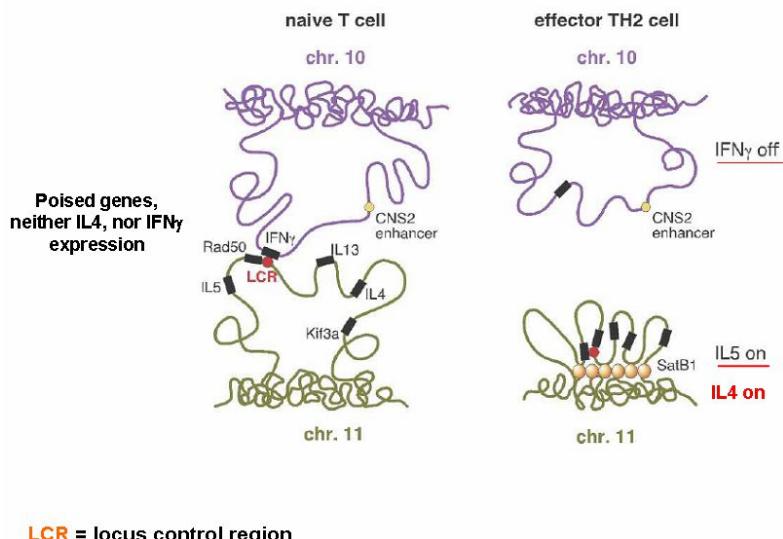


Figure 2.12. Inter- and intrachromosomal loops in the regulation of gene expression. Loops may carry genes to remote sites in the nuclear space for coregulation in an expression hub.

2.3. Meiosis

There are two forms of genetic information transmission from one generation to the next one. Firstly asexual reproduction, which is typical for the lower organisms evolved. It is a simple process, the offspring develop from the somatic cells of a single parent, thus they are genetically identical to the parent organism.

In sexual reproduction the offspring have mixed genome of two parents, so they are genetically different from both parents and from each other. Sexual reproduction has a great evolutionary advantage for the species, because the individuals gain high genetic variability allowing the adaptation to the unexpected circumstances. Sexual reproduction is crucial for the survival of species. In sexually reproducing organisms, there are two successive generations of cells: the diploid somatic cells give rise to haploid cell by meiosis and the haploid cells, which are reduced to gametes in animals. The species-specific chromosome number is restored by the fusion of gametes resulting diploid zygote, and the life of a new individual starts.

How these haploid cells are formed in meiosis? The essence of the process is double: on one hand the chromosome number is halved, secondly the parental genetic information is mixed.

2.3.1. *Phases of meiosis*

In meiosis there are two successive divisions: **meiosis I** and **meiosis II**.

Similarly to mitosis, in meiosis I prophase the chromosomes are condensed, the nucleolus and the nuclear envelope disappear. The main event of this phase is the **homologous recombination**, the exchange of sequences of paired homologous chromosomes (maternal and paternal chromosomes of the same size and shape, and having the same genes).

It is the longest phase of meiosis which is divided into five substages: **leptotene**, **zygotene**, **pachytene**, **diplotene** and **diakinesis**. In **leptotene** the two sister chromatids (having identical DNA due to duplication in S-phase) containing chromosomes are very thin fiber-like structures, which randomly bind by their both ends to the nuclear envelope. Later they move to a distinct point of nuclear envelope, close to centrosome, forming a bouquet-like structure. Thus the homologous chromosomes are close to each other, which is necessary for the next stage process. In **zygotene** the pairing, also known as synapse, of homologous chromosomes begins. Recent studies have shown that even before the pairing, probably in early leptotene the double stranded DNA-s break at several hundred sites. The pairing of homologous chromosomes is helped by a ladder-like protein structure, by the **synaptonemal complex**. It has lateral, transversal filaments, the overlapping transversal filaments form the central region of the structure (Figure 2.13).

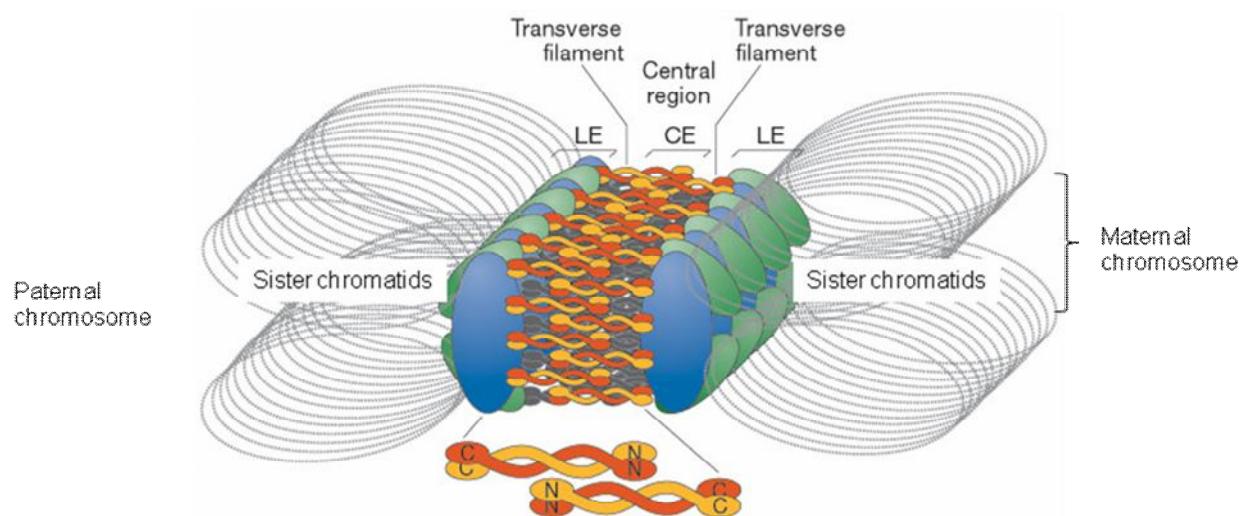


Figure 2.13. Structure of synaptonemal complex. It keeps together the homologous

chromosomes (bivalent, tetrads) like a zip.

Source : http://drugline.org/img/term/synaptonemal-complex-14373_1.jpg ; 20/02/2013.

Due to DNA condensation, chromosomes become thicker and more visible and the synapses are completed in **pachytene**. After pairing they form structures composed of two chromosomes, maternal and paternal one (**bivalent**), both having two sister chromatids (**tetrad**). The tight binding between the homologous chromosomes leads to apparent decrease in number of chromosomes (pseudoreduction). The majority of double-stranded DNA breaks are repaired, but at some of them homologous recombination (**crossing-over**), exchange of corresponding chromatids occur. This process is mediated by the recombination nodules, large 100 nm sized multi-enzyme complexes, which appear on the synaptonemal complex. The detailed molecular mechanism of crossing over is not discussed here. The crossing over may occur between any chromatids, but it results new combination of genes if it happens between non-sisters. The number of crossing overs between non-sister chromatids of a chromosome pair is 1-3. There is compulsory recombination even between the basically not homologous X and Y chromosomes at their pseudoautosomal regions (PAR). Checkpoint machinery controls the appearance and the process of crossing over, underlining the significance of homologous recombination.

In **diplobose** stage the synaptonemal complex largely detaches, thus the members of homologous pairs may slightly move away from each other, so the chromosomes are linked only at the sites of crossing overs, referred as chiasmata. Finally, in diakinesis the homologous separation continues, but the bivalents are still connected at chiasmata, found between sister chromatids of homologous chromosomes, and also by aploid s which held together sister chromatids of a chromosome. Later the aploid s dissociate from the arms and keep the chromatids together only at centromeric regions. During the prophase kinetochore region develops on chromosomes, but in contrast to mitosis, both kinetochores of a chromosome face one pole, while the kinetochore of the homologous face opposite poles (Figure 2.14).

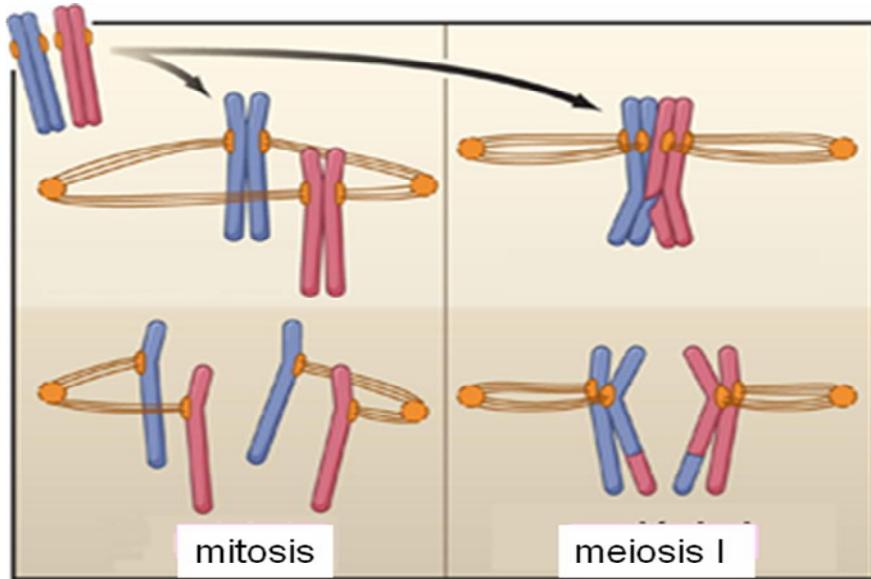


Figure 2.14. Kinetochore orientation in mitosis and meiosis I

Source : <http://www.sciencedirect.com/science/article/pii/S0092867406011524> ; 20/02/2013.

- In **first division metaphase** not the single chromosomes, but the chromosome pairs are arranged in the equatorial plane, whereas the chiasmata still connect the homologs. Chiasmata only disappear at the end of metaphase.
- In the **anaphase** the kinetochore microtubules pull the homologous chromosomes and not the chromatids toward the poles, since the kinetochores of a chromosome face the same pole. Thus the synapses not only allow the cross-over, but also needed to halve the number of chromosomes. The separation of homologous, which member of a pair is pulled to a given pole is a random process. It increases further the genetic variation. In human it is 2^{23} . In **telophase** the nuclear membrane is reorganized, and the cytoplasm splits. Arising cells are haploid, that is why the first division of meiosis is called **reduction division**. The chromosomes are still composed of two sister chromatids, which will separate in the following part, in meiosis II.

The first division is followed by a short interphase, in which there is **no DNA replication**.

Second division of meiosis is also divided into pro-, meta-, ana- and telophase, but these phases are essentially very similar to the phases of mitosis. Thus, in metaphase the single chromosomes are arranged in the equatorial plane, and in the anaphase the sister chromatids of the chromosomes are separated. The orientation of kinetochores is also similar to mitosis.

In the telophase the nuclear envelopes are reorganized, two nuclei are formed and then the cytoplasms are also halved.

Finally the **meiosis results from a diploid cell four haploid cells, the gametes**. After the fusion of two haploid cells, in the zygote chromosome number of the species is reconstituted. At the same time the **genetic information of the gametes is different** caused by the **homologous recombination in meiosis I prophase** and the **random assortment of homologous in meiosis I anaphase**. These processes provide high genetic variability needed for the survival of the species.

The most frequent abnormality of meiosis is the **non-disjunction** (Figure 2.15) either in meiosis I or II. Obviously non-disjunction of both the homologous chromosomes (in the first division), and the sister chromatids (in the second division) alters the chromosome number of resulting gametes. Involvement of such gametes in fertilization may lead to so-called **aneuploid genome mutation**.

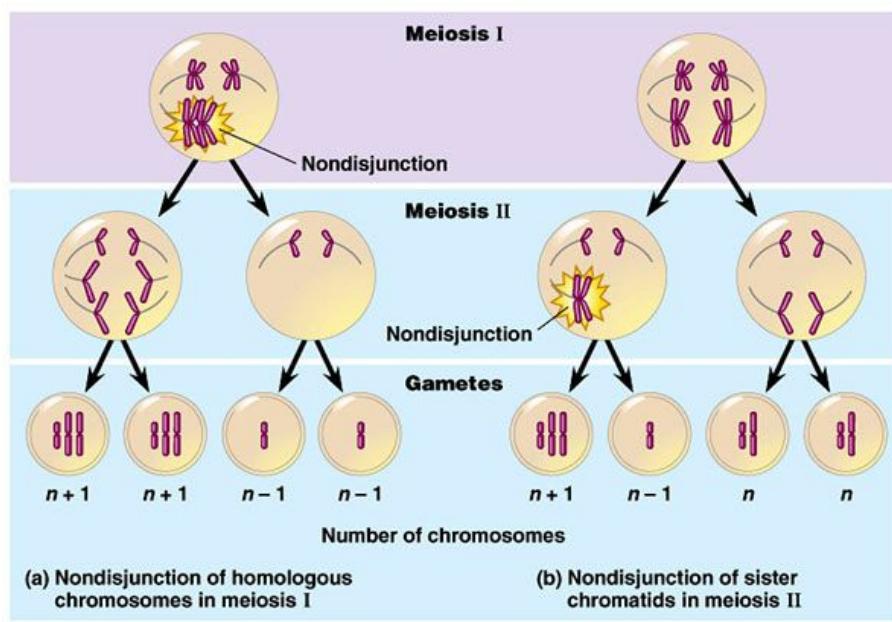


Figure 2.15. Meiotic non-disjunction

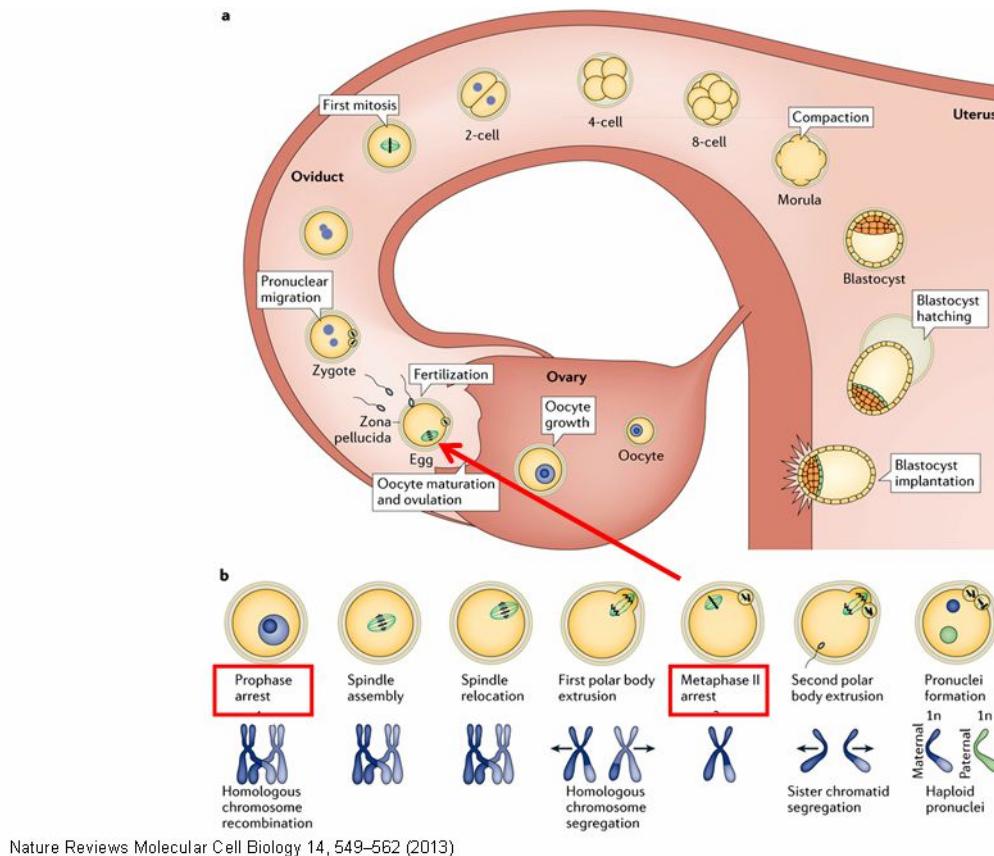
Source : http://drugline.org/img/term/meiotic-nondisjunction-9351_1.jpg ; 20/02/2013.

In vertebrates the gamete formation is a complex process, the meiosis is only a part of it. At the very beginning of ontogeny primordial germ cells migrate to the developing gonads. Several mitotic divisions are followed by meiosis, and finally in male gametogenesis a differentiation step gives rise to mature gametes.

2.3.2. Oogenesis

In most animals the female gamete (egg) is very large compared to somatic cells. Eggs contain yolk: different nutrients (lipids, proteins, carbohydrates) sufficient for the early development of embryo, until self-feeding will be able. Although the egg of mammals contains small amount of yolk (oligolecithal), it is much bigger than the body cells of the organism. The size of a human egg is about 100 µm.

In developing gonads of embryo the **primordial germ cells** (46 chromosomes) develop to **oogonia** (46 chromosomes) which divide by mitosis. The cell entering the first meiotic division is **primary oocyte** (46 chromosomes). Meiosis I is halted, the cells may remain in prophase diplotene stage for decades. Meanwhile, a coat, the zona pellucida develops around them, and in the cytoplasm cortical granules accumulate, which content is released after the sperm penetration, preventing the penetration of further sperms. From the puberty due to hormonal effects, cyclically one cell resumes meiosis I. Division of the cytoplasm is asymmetric, the larger cell is the **secondary oocyte** (23 chromosomes), whereas the smaller cell is **polocyte** (or polar cell, also has 23 chromosomes). The unequal cytokinesis is likely provided by the asymmetric mitotic spindle position. Secondary oocyte continues meiosis, enters meiosis II but it is halted in the metaphase. Secondary oocyte is ovulated in this stage. The completion of meiosis II is triggered by the penetration of sperm, the fertilization. The meiosis II results a fertilized ovum which contains two pronuclei (23-23 chromosomes) and a polocyte (23 chromosomes). The polocyte derived from the first division may divide, too, but the studies usually demonstrate the presence of only two polocytes (Figure 2.16). Polocyte derived from the first division is termed primary, the other derived from the second division is termed secondary polocyte. (in Figure 2.17 there are three polocytes!)



Nature Reviews Molecular Cell Biology 14, 549–562 (2013)

Figure 2.16. From oocyte to embryo

2.3.3. Spermatogenesis

While in most species the egg is the largest cell, not capable of independent movement, the other gamete, the sperm is the smallest cell and is able to move. In male organism **primordial germ cells** (46 chromosomes) migrate into developing testis where they become **spermatogonia** (46 chromosomes) in the external wall of testis. From the puberty, spermatogonia divide continuously by mitosis. A group of them enters meiosis I; these are the **primary spermatocytes** (46 chromosomes). Meiosis I gives rise to two haploid cells, called **secondary spermatocytes** (23 chromosomes). The second meiotic division makes for haploid round, immobilized cells, called **spermatids** (23 chromosomes). In both divisions the cytokinesis is incomplete, the secondary spermatocytes and spermatids, too, are connected to each other by cytoplasmic bridge.

After meiosis differentiation process, cytological morphogenesis begins, which results in actively motile sperms. This step is called **spermiohistogenesis**, which happens embedded in Sertoli cells. From Sertoli cells, the sperms are placed in the lumen of testis (Figure 2.17).

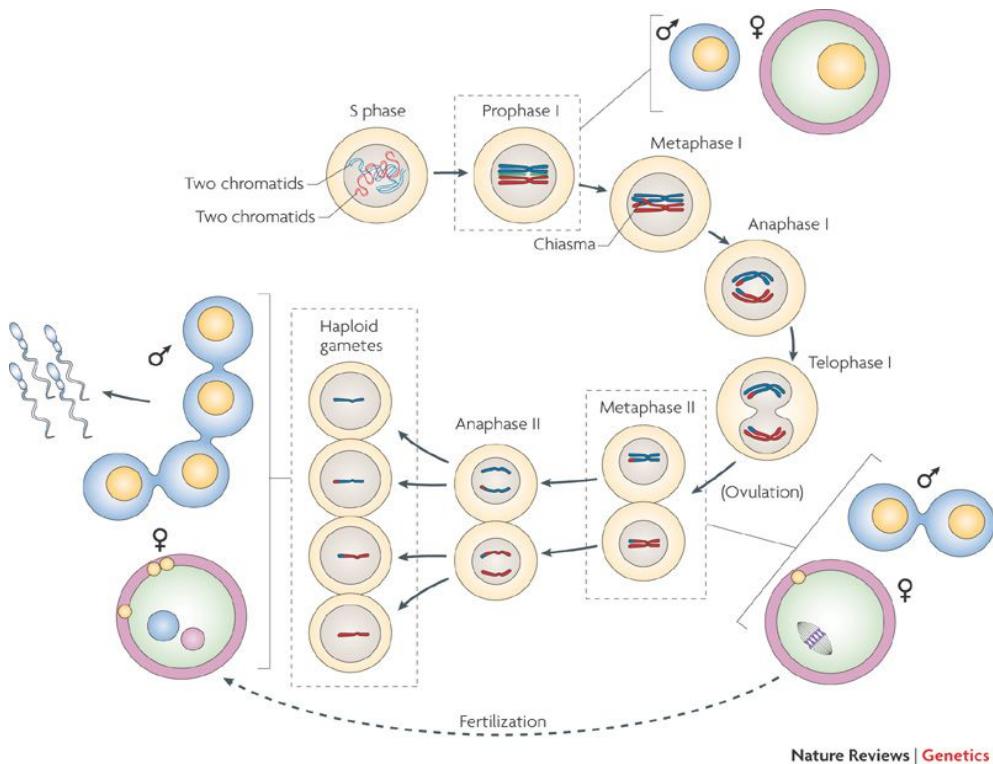


Figure 2.17. Comparison of oogenesis (pink) and spermatogenesis (blue).

Details are in the text.

Source : http://www.nature.com/nrg/journal/v11/n2/fig_tab/nrg2723_F1.html#figure-title ; 20/02/2013.

The typical structure of differentiated sperms (head, midpiece and tail part) serves only one purpose, to safely convey DNA content to the egg.

The head contains the nucleus where DNA is in completely heterochromatic form to occupy the smallest space. Protamines being more positively charged proteins than histones are needed for DNA to be packed so condense. In front of the nucleus, a giant secretory vesicle, acrosome is located. The acrosome vesicle containing hydrolytic enzymes is responsible for dissolving the different coats of eggs during fertilization. In the midpiece fused mitochondria are located, forming so-called mitochondrial sheath, where ATP necessary for the movements of the sperm is produced. The tail essentially is a flagellum composed of $9 \times 2 + 2$ microtubule system, additionally at the periphery there are nine dense keratin containing fibrils whose function is still not clear.

Recent studies have demonstrated the differences of spermatogenesis or oogenesis regarding the meiosis I prophase. These are summarized in Table 1. Differences between the whole two processes can be seen in Table 2.

Primary spermatocyte **primary oocyte**

Sites of chromosomes where synapsis starts

ends inside

Synaptonemal complex

thicker less thicker
shorter longer

Sites of chiasmata in chromosomes

ends inside

Number of chiasmata

less more

Table 1. Differences between spermatogenesis and oogenesis during meiosis I prophase

	Spermatogenesis	Oogenesis
Initialization of the process during life	Puberty	2 nd month of the embryonic life – halted in the diplotene stage of meiosis I in the 5 th month
Rhythm of the process	Continuous, without interruption	Cyclic from the puberty: monthly one oocyte pro ovum continues meiosis, enters meiosis II but it is halted in the metaphase, completion of meiosis II is triggered by the fertilization
Time required for gamete formation	60-65 days	Even 50 years. From the beginning of the 5 th month of the embryonic life the homologues in the primary oocytes are held together at the chiasmata
Effectiveness of the regulatory mechanism	Elimination of faulty spermatocytes by effective meiotic checkpoint control system	The checkpoint control system of the meiotic spindle formation is less effective leading to an increased risk of nondisjunction and aneuploidies
Result of gametogenesis Cell number at fertilization	4 equivalent sperms 15-150 million sperms/ ejaculation	One oocyte + 2 polocytes One (two) oocyte(s)/month
Termination of the process during life	Continues for a lifetime	Terminates at menopause

Table 2. Differences between male and female gametogenesis

2.3.4. *Regulation of meiosis*

The regulation of meiotic division has primarily been studied in amphibian and fish oogenesis. Firstly **MPF** itself has been found to be the regulator of amphibian oogenesis, and only later was discovered that it was equal to M-phase or mitosis triggering factor (MPF), which is composed of Cdk1 and cyclin B.

There are crucial differences in the process of oogenesis and spermatogenesis, consequently in the regulation, too. In the following the major differences, regarding the regulation are summarized.

The first difference is the start time of meiosis. **The oogenesis begins in early embryonic life, the spermatogenesis in puberty.** What makes this difference? It has been demonstrated that **retinoic acid** is responsible for meiotic trigger. Retinoic acid that is

metabolized by CYP26B1 is produced in both sexes embryo. Since the amount and the activity of CYP26B1 is higher in male than in female embryo, retinoic acid is degraded in males, but in females as a signal induces the expression and effect of STRA8 transcriptional factor. In oogonia STRA8 triggers the meiosis.

The meiosis I starts, but stops in prophase diplotene stage. The elevation of **cAMP** has been detected in primary oocyte, which in turn inactivates MPF via protein kinase A. From puberty cAMP decreases cyclically in a primary oocyte, indirectly activating MPF, needed to continue meiosis I.

The second arrest of the meiosis is in the metaphase of meiosis II. The role of a not precisely identified protein, called cytostatic factor (**CFS**) is assumed. It is under the regulation of Mos protooncogene (serine/threonine kinase) and inhibits APC activity via MAPK pathway. In the case fertilization the Ca ion level is elevated that is needed for the reactivation of APC. The activated APC induces the degradation of B cyclin and securin in the proteasome. The not inhibited separase may separate cohesin from the centromeres of chromosomes, allowing the sister chromatid separation in anaphase. Lack of B cyclin inactivates MPF and the cell, which is already the zygote, may complete the division.

Then, the zygote starts segmentation (cleavage) which is characterized by a special, so-called embryonic cell cycle. In embryonic cell cycle there is no G1 and G2, consequently cell growth, the cell size gradually decreases. Embryonic cell cycle is regulated by the same MPF, which controlled the meiosis.

3. Genetic variations

Sára Tóth

3.1. Mutation and polymorphism

Szalai Csaba

According to the classic definition **mutations** are sudden heritable changes in the DNA. The process (change) itself is still called mutation, but due to the fast development of genetics and genomics, two terms related to the variations in the sequence had to be modified. Next to the above mentioned definition the term mutation is also used to indicate a disease-causing change or sometimes rare change. Similarly, the term **polymorphism** is used both to indicate a non disease-causing change or a change found at a frequency of 1% or higher in the population. In the era of advanced DNA sequencing tools and personal genomics, these earlier definitions of mutation and polymorphism are antiquated. The 1 % or higher frequency associated with a polymorphism is an arbitrary number and in addition, there are examples that variations with >1% frequencies can cause diseases. For instance, sickle-cell anemia is caused by a nucleotide change whose frequency is >1% in a gene coding for the beta chain of the hemoglobin protein. The disease manifests in people who have two copies of the mutated gene. To prevent this confusion a new usage of these terms is suggested. It is proposed that in most cases, instead of mutation and polymorphism neutral terms like "**sequence variant**", "**alteration**" and "**allelic variant**" should be used. The term "mutation" may be used to indicate the result of a recent mutation event which has been detected using as a reference the germline DNA of the same individual. Therefore, a mutation would be a "DNA variant" acquired over the lifetime of an organism, i.e. a **somatic mutation**. In this sense, mutations are the principal causes of many diseases like cancer but are typically not inherited by their offspring. Alterations in the DNA of germ cells – sperms and eggs – can be inherited by offspring and are currently called **germline mutations**. In this case, the term mutation should be used only if the germline "variant" has been detected using as a reference the germline DNA of the same individual.

In the case a sequencing project did not include as a reference the germ-line DNA of an individual, the term "mutation" could not be used and should be replaced by the neutral term "variant". Therefore, in the sequencing report the alternative use of the term "mutation" or "variant" will also clarify which kind of reference was adopted. Importantly, the term "**polymorphism**" should only be used in the context of a population. Accordingly, this term cannot be approved to classify variants in personal genomics. (See:

<http://bmcmedgenomics.biomedcentral.com/articles/10.1186/s12920-015-0115-z> and
<http://www.hgvs.org/mutnomen/recs.html> for more details.)

3.2. The classification of mutations

To evaluate mutations it is crucial to know where, in what cell types they occur. **Somatic mutation** is formed in a given somatic cell, and when this cell divides successively the mutation is transmitted to the offspring, so a mutation carrying **cell clone** is formed. Depending on the time of the mutation, whether it is formed earlier or later in ontogeny, the number of cells involved, and the size of mutant clones will differ.

A classic example of somatic mutations when one of the eyes is blue, the other is brown. Then the mutation occurred after the separation of the eye primordium. The other situation is when brown spots can be seen in a blue eye; in this case the mutation took place after the development of the two separate eyes. Somatic mutations in natural conditions - except the vegetative propagation of plants - are not passed on to offspring.

However, the medical significance is not negligible, since somatic mutations may also play a role in tumorigenicity. According to **Knudson's hypothesis** (two hit theory) for the development of certain cancers affecting tumor suppressor genes (see in chapter 8, Genetics of biological processes), two successive mutations are required. The tumor suppressor gene mutations are recessive, so two mutant copies are necessary for the complete loss of function, and then the tumor formation. One is usually inherited, while other is formed only in one or certain organs such the previous heterozygous state is lost and the homozygosity of the mutant tumor suppressor gene leads to tumor formation. The phenomenon is called **loss of heterozygosity = LOH**, and by using modern molecular biological methods, it may be suitable for the detection of pre-cancerous conditions.

Germline mutations occurring in primordial germ cells or in germ cells during gametogenesis are inherited to the offspring, and therefore they have crucial importance in medicine.

According to their origin there are **spontaneous**, due to defective DNA replication, and **induced mutations** caused by various environmental effects (radiation, chemicals, etc.).

The mutation frequency depends on the evolutionary level, since in prokaryotes in the absence of DNA repair the mutation rate is much higher than eukaryotes. Accordingly, high

mutation frequency was observed in mitochondria with its prokaryotic-like DNA. This is about ten times (!) higher than the mutation rate of the nuclear DNA, which is about 10^{-5} per gene per generation.

The most frequent spontaneous mutations are: 1 / deaminations or 2 / depurinations
1 / during deamination cytosine is converted to uracil, and adenine to hypoxanthine;
2 / during depurination the sugar-phosphate backbone of DNA remains intact, but the purine base, e.g. guanine is lost, so it will be a ‘toothless’ DNA, so a gap is formed in the DNA strand.

The frequency of spontaneous mutations is influenced by the cell or tissue type, rapidly proliferating cells and tissues carry more spontaneous mutations, as the higher the rate of DNA duplication, the greater the chance of erroneous incorporation of nucleotides.

The best known induced mutation is the UV radiation-induced one (Figure 3.1). The UV radiation induces dimerization of thymine, this distorts the DNA, interferes with DNA replication and transcription. Similarly, incorporation of certain base analogues as 5-bromodeoxyuridine (BrdU) during DNA replication and the subsequent incorrect DNA repair, results in a change in the original sequence.

Alkylating compounds (EMS = ethyl-methane-sulfonate, MNU = methyl-nitroso-urea) attaches ethyl or methyl groups to certain DNA bases like guanine, and thus O6-methylguanine arises. This causes mutation because O6-methylguanine pairs thymine instead of cytosine, therefore following the methylation of DNA bases in the subsequent replication another base will be incorporated into the newly synthesized strand, so the mutation is fixed. Intercalating compounds (proflavin, acridine orange) either merge into the DNA strand between bases, or create DNA loops which in future replication and repair processes will lead to shorter DNA (deletion) or to DNA lengthening (duplication). Some carcinogens (e.g. benzo(a)pyrene) connects larger molecules to the DNA, so-called **adducts** are created.

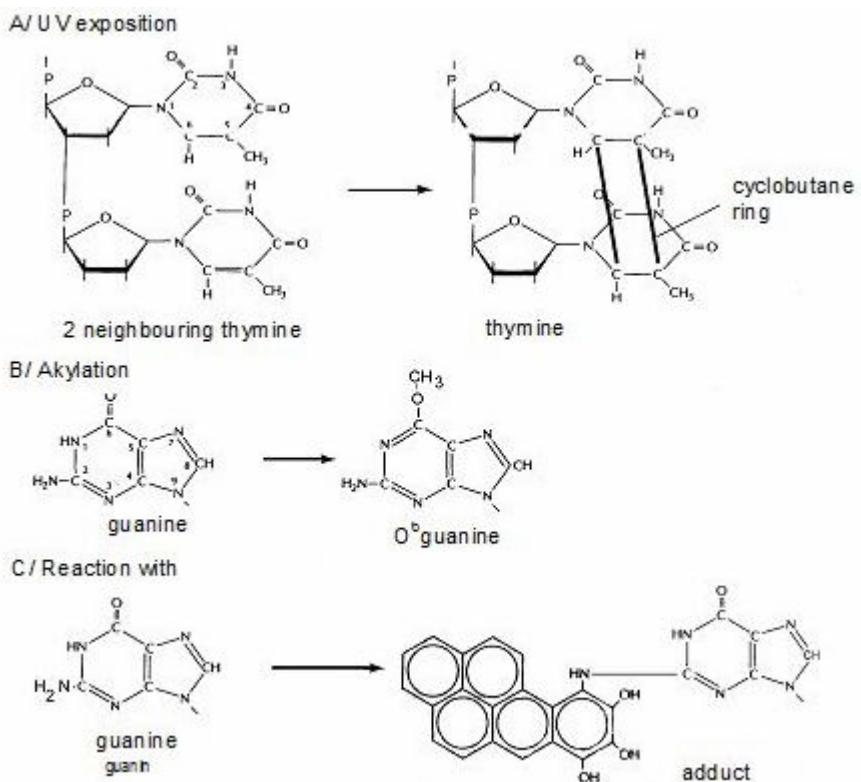


Figure 3.1. Examples of induced mutations

The mutations were also classified according to the degree of change caused in the DNA. Thus, we can talk about gene mutation - this is sometimes called point mutation, chromosome mutation when the lesion involves several genes, and genome mutation, which can affect the entire genetic material.

3.3. Gene mutations

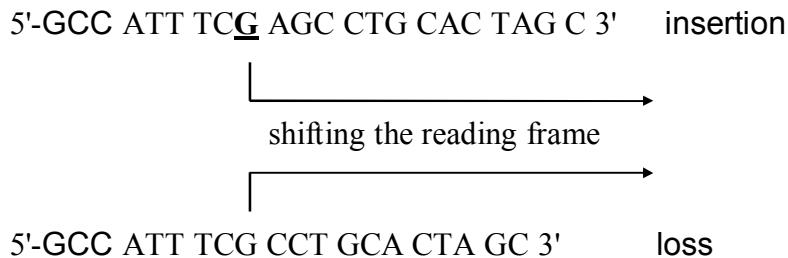
Gene mutations can affect a single base of a gene - this is called point mutation in narrow sense - and may affect a greater or lesser portion of the gene. A single base limited mutation can be addition, deletion, or base substitution. If the number of bases added or deleted is not equal to three or a whole number multiple of it (6, 9, 12 etc.), a so-called **frame-shift mutation** is created. This means that downstream from the mutation site the transcriptional reading of the coded information is changed (see example below).

Original sequence:

5'-GCC ATT TCA ACT GCC TGC AGC 3'



MUTATION



If 3 or 6, 9, etc. is the number of bases inserted or eliminated, so ***in frame mutation*** occurs, only the affected section of information content changes the rest of the gene's does not.

There are two additional opportunities in base substitution, ***the transition*** and ***the transversion***. In the first case a purine is replaced by another purine base or a pyrimidine base by another pyrimidine e.g. A → G or C → T. In the latter case, a purine base is exchanged by a pyrimidine base or vice versa a pyrimidine by a purine.

The effects of substitution mutations are more varied. There are missense, nonsense and silent or sense mutations.

Following a ***missense mutation*** the codon changes, and thus a different amino acid is incorporated into the protein, e.g. the mutation causing sickle cell anemia, wherein base exchange results in the incorporation of valine in place of glutamine as 6th amino acid in the β-globin chain of hemoglobin.

Due to a ***nonsense mutation*** the original codon changes to stop codon, so early termination of the protein chain, and a shorter protein molecule will be the result. In the case of ***silent or sense mutation*** due to degenerated genetic code, although the codon changes, the same amino acid is incorporated into a protein, so the mutation has no consequences. This mainly occurs when the 3rd and 2nd base of the codon is replaced by another one.

Whereas, during a cell's life DNA replicates several times and may be repeatedly exposed to mutagenic agents several - even different types - mutations may occur. Some of the recurrent mutations affect the part is already mutated, restoring the original sequence. Then a so called ***back or reverse mutation*** takes pace, and thus the potentially harmful consequences of the first mutation are eliminated.

There is another possibility to apparently eliminate the consequence of a mutation. In prokaryotes it has been observed that although the DNA mutated, different amino acid was not built into the protein. It turned out that in this case tRNA also mutated and yet it delivered the original amino acid to the altered mRNA codon. Such tRNA is called *suppressor tRNA*.

If the mutation affected an extended sequence (any number of bases) in the gene, then gene deletion, addition, or when the given sequence is reversed, gene inversion takes place. Longer deletions may also affect not only one gene, but either gene families, (where genes with similar but not identical function, derived from an ancestral gene, are located directly one after the other in the DNA). Such is the case of deletions affecting globin gene families in various hemoglobin disorders (hemoglobinopathies), for example in thalassemias (its mutation mechanism is discussed at mutation hot spots). Of course, the longer the lesion, the more severe the consequences, that is the more wrong, altered or functionless the protein product as well.

A special case of gene mutations, namely the additions is, when Alu sequences or LINE elements are integrated into the coding region of a gene by transposition or retrotransposition. Then the addition of the jumping element (transposon or retroposon) breaks the original exon sequence, and thus leads to RNA and protein formation of altered information content. In the case of hemophilia A addition of an Alu sequence causes the disease.

Similarly, gene duplication due to recombination results in mutation as well. *It occurs either in meiosis, when unequal crossing over between non-sister chromatids leads to gene duplication, or in mitosis, when rarely mitotic recombination (crossing over) takes place between sister chromatids.* In the latter case, it is a somatic mutation, and may lead to tumorigenesis by creating such daughter cells, where in one of them the heterozygosity is lost. (In the other daughter cell three copies of the gene are present, where gene duplication is in one of the homologous chromosomes while the other homologue is with a single locus of the given gene).

The so-called ***mutational hot spots*** should be mentioned here. Individual DNA sequences, genes more likely mutate, where repetitive sequences are found. These repeats may interfere with the replication and meiotic pairing of the homologous chromosomes. The replication abnormality has physical causes: symmetrical or repetitive sequences located on the same strand of unwound DNA can pair on the basis of complementarity or can form loops, and thus disturb the function of enzymes involved in replication and repair. For example, in hemophilia B, where large direct sequences with CG repeats occur in the factor IX. coding gene, there are

10 to 100 times more mutations. This higher mutation frequency may be attributable to epigenetic causes as well. Methylated cytosine easily deaminates to thymine, thus leading to a C → T transition on one strand and a G → A transition on the other.

The above mentioned ***uneven crossing*** (see Chapter 4, Cytogenetics) can explain the repetition of larger sequences, sometimes of entire genes. For this phenomenon a good example is the formation of α-thalassemia. Normally, both homologs of chromosome 16 contain two α-globin genes one after the other. The uneven crossing over can result in gametes, which for example contain 1, or 3 α-globin genes. Following the fertilization of such zygotes may be formed in which one more or one less α-globin genes are present. The person's health depends on the number of α-globin genes: 0 copy - intrauterine lethality, 1 copy - severe anemia, 2 copies - mild anemia, three copies - asymptomatic carrier can be the consequence. Today, more than 30 diseases are known, which are caused by uneven crossing over (e.g. red-green color blindness).

Not only the direct repeats, but ***palindromic sequences*** (sequences of which base sequence in 5'-3' direction is the same on both strands) can frequently lead to additions and deletions.

A special type of gene mutations are the ***repeat mutations*** where different numbers of nucleotides are repeated forming so called repetitive units. Beside the best known ***trinucleotide repeat mutations*** there are other length mutations involving up to 24 nucleotides long units, thus resulting in the accumulation of e.g. octapeptide units in the protein (Creutzfeldt-Jakob disease).

Several types of trinucleotide repeat mutations are known.

1. ***Polyglutamine diseases*** associated with the expansion of CAG triplets (CAG is the code of glutamine). The currently known CAG trinucleotide repeat mutations results in severe neurological diseases, so called neurodegenerative disorders, but numbers of disease causing repeats are different. In the case of Huntington's chorea and Kennedy disease repeats affect the protein coding region of the gene.
2. Another large group of known trinucleotide repeat mutations are ***the polyalanine diseases*** in which GCN triplets (N can be any nucleotide) are accumulated, causing alanine accumulation in the protein. They are mainly transcription factor affecting mutations which generally lead to developmental malformation syndromes such as synpolydactylia or hand-foot-genital syndrome.

3. While the aforementioned trinucleotide repeats are in the ***coding regions*** of the genes, in the case of myotonic muscular dystrophy and fragile X syndrome the repeats are in the ***untranslated region = UTR***, and generally the number of repeats is larger as well.

SOME TRINUCLEOTIDE REPEAT DISORDERS

Disease	Frequency	Trinucleotide	Normal allele	Mutant allele	Repeat numbers
Huntington	1:10000	(CAG) _n	11 - 34	42 - 100	
Fragilis X	1: 2000	(CGG) _n	10 - 50	52 - 500	
Myotonic d.	1: 8000	(CTG) _n	5 - 35	50 - 200	
Kennedy	1: 50000	(CAG) _n	11 - 31	40 - 65	

It is typical for the repeat mutations that the diseases are caused only above a certain number of repeats, so there is a so-called ***premutation state***, and that the ***repeat expansion*** (the increase in the repeat number) takes place during meiosis. The increase in the number of repeats (e.g. CAG) is the result of a process (already mentioned at the mutation hot spots), when during the replication one of the DNA strands is looped due to the repetitive sequences. If this looping affect the newly synthesized strand, the replication apparatus can detect it as if the replication had not been completely carried out from the template strand, so more repeat units are added to the new strand. This phenomenon is called ***replication skipping***.

With this the new strand contains several new repeat units as well. The old and the new strands are different in length, and then the repair mechanism corrects it by adding a sufficient number of new repeats to the old strand. The opposite phenomenon also exists - that may decrease the number of repeats. In this case looping affects the template strand, i.e. the newly synthesized strand will be shorter than the original was. However, in this case the repair will correct the error, too, i.e. the unnecessary number of repeats will be cut out from the old strand, so eventually there will be a DNA molecule containing fewer repeat units.

Since DNA replication takes place before mitosis and meiosis and, in principle, the repeat number can change in both cases. In contrast, in the case of GCN trinucleotide repeats the change in the repeat number is rather explained by a meiotic event namely by uneven crossing over.

Since the number of repeats varies from generation to generation, the repeat mutations are not stable and therefore they are recently also called ***dynamic mutations***. In the case of prokaryotes, this dynamism has an important role in counteracting with the effects of the host's immune system; in eukaryotes it may play a role in tumorigenicity.

The role of repeat mutations that cause disease is easy to see, since with the addition of increased number of repeats into the coding sequence, with this expansion, the structure of the gene involved will be more and more distorted, therefore the protein is coded by it becomes more and more altered and non-functioning.

There is a phenomenon called ***anticipation*** - long been known in human genetics but cannot be explained for a long time. **Anticipation means that a hereditary disease transmitted from generation to generation will appear in younger and younger age and in more and more severe form.** Since the repeat expansions of medical importance takes place especially in meiosis (or in the previous S phase), and the gene in question will be increasingly damaged by this process, the above mentioned phenomenon is well explained.

In case of gene mutations not only the size of the mutation, the length of the DNA sequence in question is important, but in eukaryotes, including humans the place of mutation is important as well. It is not all the same whether a mutation is in a coding or non-coding region or in an untranslated region, in an exon or an intron, or even in the border between them. In the latter case, ***so-called splicing mutations*** can be formed, because the exon-intron boundary sequences play an important role in the intron looping, in the lariat formation thereby in spliceosome function. As a consequence of splicing mutations an exon is lost, or an intron can be translated, and a definitely defective protein will be the end result.

Even mutations of a single gene, which occur at various places, or base exchanges or splicing mutations can cause symptoms of completely different diseases, or symptoms of different severity as it is known for a large number of mutations in cystic fibrosis.

The role of the ***UTR mutations*** had not been understood till the past few years, since at first glance we might think that a mutation affecting the non-coding DNA sequences only, and therefore no defective protein had been produced, could not cause symptoms or disease. In contrast, we now know that the 5'UTR region of the mRNA is required for ribosome binding, and normal protein synthesis. Thus, it was also understood that some of the trinucleotide repeat mutations in which the expansion affects the UTR region why cause diseases. Moreover, the methylation of cytosines in a number of repeats induces epigenetic alterations (attachment of methyl-binding proteins and / or non-coding RNAs, chromatin remodeling), which also explains the role of the UTR mutations in disease.

Although as a result of the Human Genome Project the human DNA sequence is almost completely known, but knowing the sequence does not imply the identification of the gene and knowing the gene does not automatically mean the understanding of its function.

This is particularly problematic in those cases where the mutation results in a new protein with different functions, but neither the original protein nor the mutation itself is known. These are the so-called ***gain of function mutations***. It makes the genetic analysis more difficult. It is typical for Huntington disease, where the exact original function of the finally identified ***huntingtin protein*** is still not well understood.

The situation is simpler when a mutation changes the structure or the function of a previously known protein. Then ***loss-of-function mutations*** are present like in the case of phenylketonuria or sickle cell anemia. These loss-of-function mutations are usually recessive, so they manifest in the phenotype only in homozygous form. This may be so because for most of the gene products the exact quantity is not crucial and the system works correctly with half dose as well. However, there are ***dose-sensitive genes***, where 50% of the product is not enough. This is called ***haploinsufficiency*** (haplo = single; insufficient = failure). So the phenotype is abnormal even in heterozygotes, therefore dominant inheritance is typical for loss-of-function mutations.

It also occurs that, due to the mutation not only the original function of the gene product is lost, but the mutant product prevents normal function of the normal product, e.g. they cannot dimerize. This is called ***dominant-negative effect***, when the mutation is associated with dominant phenotype and is also expressed in heterozygous form.

3.4. DNA repair

If the DNA or the genes themselves can mutate in so many ways, it is not surprising that a number of mechanisms developed during evolution to ensure the integrity of the genetic material that protect against mutations and to correct the resulting errors. The name for these mechanisms ***is DNA repair***.

These mechanisms are grouped according to

- 1 / whether it reverses the chemical reaction that causes mutation; it is the ***direct repair*** or
- 2 / it cuts out the incorrect bases and replaces by correct ones; it is known as ***excision repair***.

1 / The best example of the direct repair is the removal UV-induced thymine dimers. Photoreactivation is a process typical mainly for prokaryotes and for some eukaryotes (e.g.

yeast). In this process the cyclobutane ring formed between the pyrimidine bases are slit by the use of visible light energy, and the bases remain in their original location, and the previous structure is restored.

Although UV radiation is one of the most mutagenic effects (just think about the growing ozone hole and the intensive UV radiation exposure over Earth's surface), unfortunately, many species, including humans are not capable of this photoreactivation repair. Will it be explained by the later evolution of man which followed the formation of the protective ozone layer, and by the small amounts of UV rays reaching the earth's surface?

Another direct repair mechanism is the removal of alkylated bases. The O⁶-methylguanine methyltransferase enzyme removes the methyl group of guanine by linking it to a cysteine base in its active site. Such enzymes are found both in pro- and eukaryotes.

2 / The excision repair is more common than the direct repair. There are three types:

- a / base excision
- b / nucleotide excision
- c / mismatch repair

a / During the **base excision repair** the single incorrectly incorporated base is spliced out and the DNA polymerase fills the gap by using the intact complementary strand as a template.

b / In **nucleotide excision repair** not only the mutated part e.g. thymine dimer is cleaved off, but also the preceding and following few other nucleotides, i.e. a shorter oligonucleotide sequence. Then, the DNA polymerase fills the gap based on the undamaged complementary strand, and DNA ligase connects the old and the repaired section. For this process in humans seven different genes are required and mutations in any of them are associated with uncorrected UV radiation-induced mutations. It is typical for rare inherited disorders, such as the Cockayne syndrome or xeroderma pigmentosum. The latter disease is a good example of genetic heterogeneity, since the errors of different excision repair enzymes result in the same symptoms.

c/ During **mismatch repair** not exactly complementary bases that not exactly fit to the double helix are recognized and removed. During DNA replication - in parallel with the synthesis - mismatched bases are detected and removed by the proof-reading (3' → 5' exonuclease) activity of the DNA polymerase enzyme. Those that escape this process are corrected by the enzymes of the mismatch repair complex.

While bacterial mismatch repair is relatively well, human one is less known. However, we

know that the common hereditary colon cancer is caused by mutations in some genes of the mismatch repair protein complex. **So not only mutations in a specific protein coding gene, but any molecular defects in the repair mechanism can also lead to disease.**

Both the direct and the excision repair take place before DNA replication, ensuring that only the correct DNA molecule's replication is possible. However, the cells have "multiple insurance", so there are two additional, alternative, ***post-replication repair mechanisms*** in case of the failure of the first two repair processes. One of them is the ***recombination repair***, the other is the so called error-prone or ***SOS repair***. During the recombination repair the mutation remains uncorrected - for example a thymine dimer inhibits DNA synthesis, so there will be a gap in the right place in the new strand.

(The synthesis is not interrupted completely because the DNA polymerase - as in the case of Okazaki fragments can synthesize a new strand in pieces.) The gap is filled after recombining with the original strand, while the gap of the original strand derived from the previous recombination is filled by DNA polymerase and ligase enzymes in cooperation. This late repair mechanism makes the correction of a mutation possible before the next DNA replication.

The ***SOS or error-prone repair*** is known only in prokaryotes (although similar mechanisms are assumed in eukaryotes as well), and it works only in extreme cases where cell survival is at stake. When much of the DNA is damaged by strong radiation or other mutagenic agents, there is no time for the precise yet time-consuming repair mechanisms mentioned above, but to correct the DNA damage quickly and inaccurately, thus avoiding the immediate cell death. It is obvious that such a mechanism in multicellular eukaryotes is not necessary, since the death of a single cell does not lead to the death of the whole organism, the other cells will take over the function of the lost cell.

The most difficult is to correct those mutations which are usually caused by ionizing radiation or oxidative damage resulting ***double-stranded DNA breaks***, since then - in contrast to the previously mentioned repair mechanisms - there is no template strand serving as the basis of correction. The double-stranded breaks - given that free ends are generated - increase chromosome instability and thus can lead to structural chromosomal abnormalities. There are two mechanisms to correct them:

- a. the so-called ***non-homologous end joining (NHEJ)***

b. the ***homologous recombination***

In **NHEJ** a special DNA ligase with a cofactor brings the broken ends together. If the double-stranded break creates fragments with overhanging strands and microhomologous sequences then the repair is most likely correct. If the fragments have blunt ends, then there is a high chance to unite not related pieces and generate structural chromosome abnormality as well.

For error correction ***homologous recombination repair*** uses either the correct sequence of the homologous chromosome, or in G2 phase of the cell cycle the sister chromatid already formed as template by an enzyme system similar to that used in the crossing over.

Of course, not only various mechanisms for repair are available to protect genome integrity, but inactivation systems as well which can neutralize or inactivate mutagenic agents. Such as the peroxisomal system in which the oxidative and thus mutagenic superoxides are eliminated by superoxide dismutase that converts peroxides H_2O_2 , then catalase cleaves and thus neutralizes it.

3.5. Mutagenicity tests

To avoid the adverse effect of mutations, however, we cannot rely exclusively on cells with evolutionarily integrated repair mechanisms, but everything possible must be done to prevent the production and the marketing materials of any mutagenic effect. This is the aim of the ***mutagenicity tests***. According to international regulations, all prospective medicines and chemicals are subjects to the so-called pharmacological safety studies including different types of the mutagenicity tests. It is important that the widest possible spectrum of - from prokaryotes to eukaryotes, including mammals and human - in vivo or in vitro tests be carried out to eliminate the mutagenicity, the mutation-inducing effects.

Generally, to detect point mutations bacterial direct tests such as the Ames test are the first, when the test material is directly administered to the appropriate culture of bacterial strains.

However, it is also possible that not the test substance itself, but one of its metabolite is mutagenic, in this case mammalian liver microsome fraction containing enzymes necessary for the metabolism is also added to the experimental system.

Mutagenicity tests not only for the detection of point mutations but for the detection of mutations affecting DNA repair and causing numerical and structural chromosome aberrations are also available.

One of the most widely used method is the in vitro ***sister chromatid exchange (SCE) technique***, by which the exchange between sister chromatids can be detected. Although it is still not known why sister chromatids exchange in normal somatic cells - as their genetic material is 100% identical (incorrect replication has negligible effect) - but DNA-damaging, mutagenic substances can multiply the normal 4-5 SCEs/ cells (mitosis) exchange frequency.

However, this technique has medical and diagnostic significance, since there are fortunately rare, inherited diseases, such as Bloom's syndrome with increased chromosome fragility and instability, which characterized by ≈ 60 SCE / cell in the SCE test and these values serve as a basis for diagnosis.

Categories of ***beneficial, neutral and harmful mutations*** are used for the assessment of consequences of mutations in population and evolution genetics. Then the mutation is evaluated not from the individual's point of view but from the survival of the species. However, we must not forget that in this case the mutation is not alone, but in relation to the environment investigated. The best, now classic example is the case of white and black pigmented versions (morphs) of peppered moth (*Biston betularia*) in England (see http://en.wikipedia.org/wiki/Peppered_moth). It also warns that not only the genetic material, but its environment is changing, and that was once beneficial or neutral environment for one of them, it was detrimental to the other, or vice versa. It should be noted also that the ***wild type / mutant allele*** discrimination applies to a particular environment, population state and ***the wild type allele always means the mutant, which is the most common under those circumstances.***

3.6. Nomenclature of the genetic variants

Csaba Szalai

In genetics the uniform and unequivocal description of sequence variants in DNA and protein sequences is very important. The Human Genome Variation Society has been established that created the uniform rules of the nomenclature of sequent variants. Below some details can be found about the most recent recommendations. See for more details in:

<http://www.hgvs.org/mutnomen/recs.html>.

3.6.1. Levels of variations

The most important rule is that all variants should be described at the most basic level, i.e. the DNA level. Descriptions should always be in relation to a ***reference sequence***, either

a **genomic** or a **coding DNA** reference sequence. Discussions on which type of reference sequence to prefer, genomic or coding DNA, have been lively. Although theoretically a genomic reference sequence seems best, in practice a coding DNA reference sequence is preferred.

- "c." for a **coding DNA** sequence (like c.76A>T)
- "g." for a **genomic** sequence (like g.476A>T)
- "m." for a **mitochondrial** sequence (like m.8993T>C,
- "n" for a **non-coding RNA** reference sequence (gene producing an RNA transcript but not a protein)
- "r." for an **RNA** sequence (like r.76a>u)
- "p." for a **protein** sequence (like p.Lys76Asn)

For a clear distinction, descriptions at DNA, RNA and protein level are unique;

- **DNA-level**
in capitals, starting with a number referring to the first nucleotide affected (like c.76A>T or g.476A>T)
- **RNA-level**
in lower-case, starting with a number referring to the first nucleotide affected (like r.76a>u)
- **protein level**
in capitals, starting with a letter referring to first the amino acid affected (like p.Lys76Asn)

3.6.2. Positions of the variations

Coding DNA as the Reference Sequence

- there is no nucleotide 0
- nucleotide 1 is the A of the ATG-translation initiation codon
- the nucleotide 5' of the ATG-translation initiation codon is -1, the previous -2, etc.
- the nucleotide 3' of the translation stop codon is *1, the next *2, etc.
- intronic nucleotides:

- beginning of the intron; the number of the last nucleotide of the preceding exon, a plus sign and the position in the intron, like c.77+1G, c.77+2T, etc.
- end of the intron; the number of the first nucleotide of the following exon, a minus sign and the position upstream in the intron, like c.78-1G.
- in the middle of the intron, numbering changes from "c.77+.." to "c.78-.."; for introns with an uneven number of nucleotides the central nucleotide is the last described with a "+"

3.6.3. Specific changes

- ">" indicates a **substitution** at DNA level (like c.76A>T)
- "del" indicates a **deletion** (like c.76delA)
- "dup" indicates a **duplication** (like c.76dupA); duplicating insertions are described as duplications, not as insertions; ACTTTGTGCC to ACTTTGTGGCC is described as c.8dupG (not as c.8_9insG)
- "ins" indicates a **insertion** (like c.76_77insG)
- "inv" indicates an **inversion** (like c.76_83inv)
- "[]" indicates an **allele** (like c.[76A>T])

Two sequence variants in one individual:

- two sequence changes in different alleles (e.g. for recessive diseases) are listed between square brackets, separated by a ";"-character;
c.[76A>C];[87delG]
- two sequence variants in one allele are listed between square brackets, separated by a ";"-character; c.[76A>C; 83G>C]

3.7. Useful web-sites:

www.genomic.unimelb.edu.au mdi/

www.hvgs.org/mutnomen/

<http://www.hgmd.cf.ac.uk/ac/hahaha.php>

<http://decipher.sanger.ac.uk/>

www.ncbi.nlm.nih.gov

3.8. Questions

1. When the terms mutation and polymorphism may be used?
2. What kind of mutations do you know according to their origin?
3. Give examples of some physical and chemical mutagens!
4. Why could a double-stranded DNA break lead to structural chromosomal abnormality?
5. What is the difference between the causes leading to polyalanine and polyglutamine diseases?
6. What is the explanation for the existence of mutational hot spots?
7. What is the connection between the anticipation and nucleotide repeat mutations?
8. Why is SOS repair not found in multicellular organisms?
9. When does mutation repair take place?
10. Give examples of some mutagenicity tests!
11. What could be the consequence of splicing mutations?

4. Cytogenetics

Chromosome mutations

Sára Tóth

Cytogenetics is a field of genetics dealing with species or cell specific number of chromosomes, and their structure and characteristic segments, their functional roles, and all the differences - namely the chromosomal mutations - related to them. Chromosome mutations are changes in the structure or in the number of chromosomes, and since they are relatively rare in this respect they differ from normally occurring common, harmless chromosome polymorphisms. Since both types of chromosome aberrations affecting many genes, and since the size of chromosomes or their affected segments are within the limits of microscopic resolution therefore they can be examined by light microscope, as opposed to gene mutations only be identified by molecular biological techniques. However, the application of modern hybridization based (FISH and CGH) techniques allow the identification of small structural changes (e.g. microdeletions or CNVs) previously unrecognized by light microscope.

Two aspects of the chromosomal abnormalities are regarded crucial: when and where they happen. While chromosome mutations may be formed during both mitosis and meiosis, those may occur in meiosis, lead to defective gamete formation, and to the birth of affected offspring. Thus their medical significance is greater than that of mitotic chromosome aberrations. From the point of mitotic chromosomal abnormalities it is also important when during development and in what kind of cell they are formed. Mutations occurred during the early cleavage divisions may have serious consequences for the entire organism, while aberrations occurred in a continuously proliferating cell type (e.g. epithelial cells) in adulthood may have negligible role. However, certain chromosomal mutations may have a role in the formation and subsequent rapid proliferation of tumor cells.

Two chromosomal regions have special importance in the formation of chromosome aberrations: the centromeres and the telomeres.

The **centromere** is primary constriction of chromosomes where sister chromatids are connected, situated in strictly imposed position of the chromosome. The kinetochore and through that kinetochore microtubules bind to it. Its significance is in the segregation of sister chromatids or chromosomes during anaphase, therefore it plays role mainly in the formation

of numerical chromosome mutations. Chromosome pieces without centromere (acentric fragments) do not reach the right pole of the cell, but lost in successive divisions.

There are many repetitive GC-rich sequences around the centromere, and the centromere itself includes the latest replicating DNA.

The chromosome ends, the *telomeres* are rich in TTAGGG repetitive sequences, and ensure the integrity and stability of the chromosome structure and play a role in cell aging, in tumorigenicity and the formation of structural chromosome aberrations, since in the absence of a telomere the chromosome structure becomes unstable and fragments without telomeres easily adhere, opening the way for a wide variety of disorders.

4.1. Structural chromosome aberrations

The prerequisite of structural chromosome aberrations is breakage of chromosome/s which can be spontaneous or induced. The classification of structural aberrations is based on the number and the location of breaks within chromosomes (Figure 4.1).

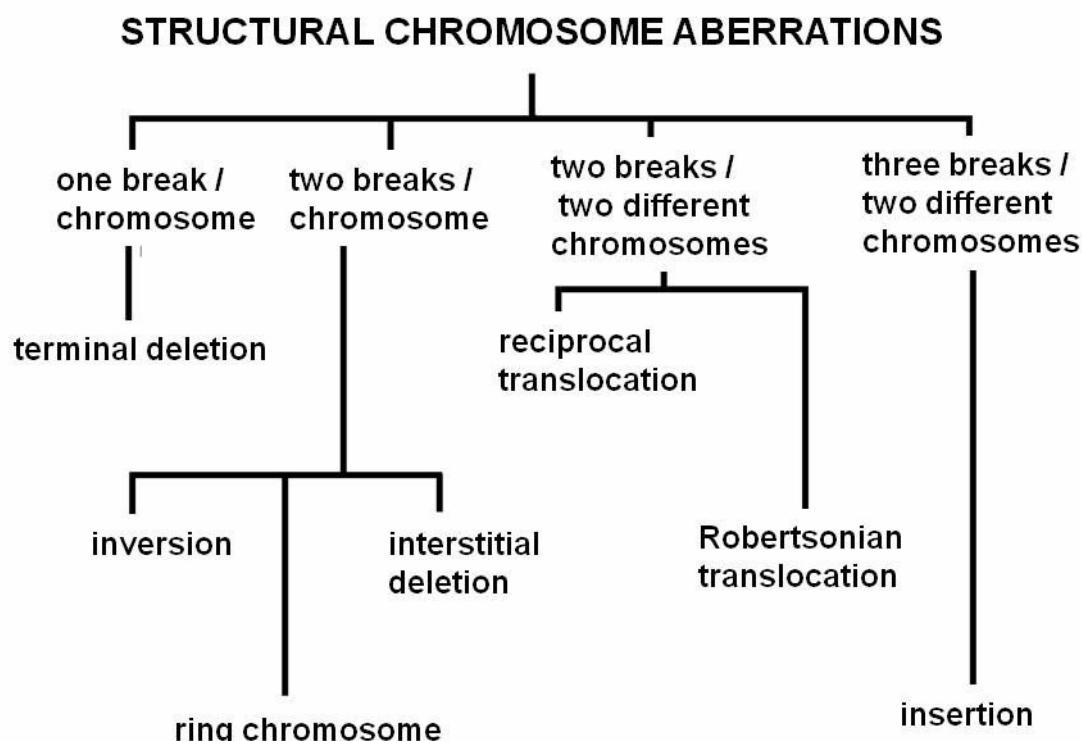


Figure 4.1. The classification of structural chromosome aberrations

4.1.1. Deletions

If a chromosome is broken, and the broken piece lost, we are talking about deletion. Then the genetic information carried by the broken piece will be absent from the cell involved, whereupon the cell does not function normally or die. Since the deletions eliminate certain functions therefore certain proteins for example enzymes are not produced. By the help of deletions the location of the gene eliminated can be mapped - it was one of the earliest methods of gene mapping, the ***deletion mapping***.

If the break is close to the end of the chromosome, a ***terminal deletion*** is generated. In this case, in addition to other genes telomere is lost, too and this also contributes to the severity of symptoms, to early lethality. The best known example of a terminal deletion is the ***cat cry (cri du chat) syndrome***: the short arm of chromosome 5 is deleted (5p-). The disease is named after the affected newborns characteristic mewing cry.

There are two breaks within one chromosome in the case of ***interstitial deletion***, and the intermediate piece is lost. Such lesions usually may cause severe physical and mental disabilities, spontaneous abortion, premature death depending on the chromosome involved. The best known interstitial deletion affects the long arm of chromosome 15: del15 (q11-13). This is one of the causes of Prader-Willi or Angelman syndrome (see Chapter 5, Epigenetics and genomic imprinting). In the former case paternal deletion, in the latter one maternal deletion is found.

Also interstitial, but small, so-called ***microdeletions*** are in background of Williams and DiGeorge syndromes (del7q11.23 and del22q11.2) as well.

4.1.2. Duplications

During duplication a chromosomal segment is duplicated. It's either a replication error or due to meiotic unequal crossing over. In both cases the repetitive sequences occurring in the affected region may explain the "slipping" of the replication apparatus or the non-exact pairing of the non-homologous chromosomes (skipping). Like deletions, duplications are also used to identify the chromosomal location of a gene or group of genes, so to map a gene.

4.1.3. Translocations

For the formation of translocations more than one, usually 2 or 3 breaks are needed. The broken part / s are transferred to another chromosome. Depending on the origin of the broken piece or on the number of fragments translocated there are different sub-groups of the translocations.

4.1.3.1. Reciprocal translocations

At least two breakpoints are expected in the reciprocal translocations, which may be in two homologous chromosomes or in two completely different non-homologous ones. The broken fragments of chromosomes are exchanged then join to a new location.

As a result, two chromosomes of altered structure are created. However, this does not cause phenotypic changes, i.e. symptoms or disease in most cases. This is the case of ***balanced translocation***. This phenomenon can be explained by just changing the position of the affected genes, not the genes themselves. Breakpoints are usually in non-coding regions, as the proportion of the coding regions of the human genome is <3%.

In cases where the breakpoint is within a gene, following the translocation the gene itself is affected, so the abnormal product - with different function, activity or amount, or perhaps unable to function - is responsible for the appearance of pathological traits, e.g. tumor formation.

The best example of reciprocal translocations leading to the formation of the ***Philadelphia chromosome (Ph₁)***, is between 9 and 22 chromosomes, its cytogenetic abbreviation is t(9;22)(q34;11). This translocation occurs in ***chronic myeloid (CML)*** or ***acute lymphocytic leukemia (ALL)***. The breakpoint in chromosome 22 is in the BCR (breakpoint cluster region) gene, while the breakpoint of chromosome 9 affects in the ABL (Abelson murine leukemia) proto-oncogene. Since the ABL gene encodes a tyrosine kinase as the result of the translocation a bcr / abl fusion protein is produced which not only has a greater molecular weight than the original enzyme, but also a higher activity. In fact, during this translocation the well-regulated promoter of ABL gene is lost, and the gene permanently overexpressed. Finally this leads to uncontrolled cell proliferation, i.e. the development of the tumor.

Another medically important example is the Burkitt's lymphoma caused mostly by Epstein-Barr virus. In this disease the c-myc proto-oncogene coded by chromosome 8 is translocated to chromosome 14 or 2 or 22 [t(8; 14) or t (8;2) and t(8;22)], where either the immunoglobulin heavy-chain (on chromosome 14) or Ig light chain genes -κ chain on

chromosome 2 and λ chain on chromosome 22 are located. As genes coding the Ig chains are continuously transcribed, and therefore the translocated c-myc - which encodes a transcription factor acting in heterodimeric form – is constantly overexpressed, too and leads to the increase of cell proliferation, ultimately tumorigenesis. These two cases are examples of the relationship between translocations and proto-oncogenes, where the overexpression of a normal protein (Burkitt's lymphoma), or a fusion protein - although of normal function - regulation of independent production (CML) is responsible for tumor formation.

If three instead of two breakpoints occur, *insertional translocation* or *insertion* takes place. Then one piece of the broken chromosome (2 breaks) is incorporated, inserted to the other chromosome (one break).

Since human chromosome set consists of 46 chromosomes, and any piece of any of these chromosomes may change places, so it is obvious that the number of translocations is almost infinite.

The special case of translocations is the *Robertson's translocation* or *centric fusion* (Figure 4.2). In these structural chromosomal abnormalities only acrocentric chromosomes can be involved so in humans only one of the 13, the 14, the 15, the 21 and the 22 chromosomes. Not only the types of chromosomes involved, but at the breakpoints are strictly determined: a break is always in the centromere or near to centromere.

In this way abnormal - fusion - chromosomes can be formed, one of which initially can contain two centromeres (dicentric), but ultimately only one centromere remains active, the other is without a centromere, therefore it is lost during the subsequent divisions and thus the number of chromosomes is reduced .

If the break is exactly in the centromere, a two centromeric, although rearranged chromosome is formed. As the result of the rearrangement from the initial acrocentrics two different sized, a larger and a smaller metacentric or submetacentric chromosome is created. The small submetacentric containing NOR (nucleolar organizer) regions on both arms is lost during successive divisions. However, since there are 10 NOR regions in the human genome, on the short arms of the acrocentric chromosomes therefore any loss of two does not lead to phenotypic change, that is, the centric fusion is balanced.

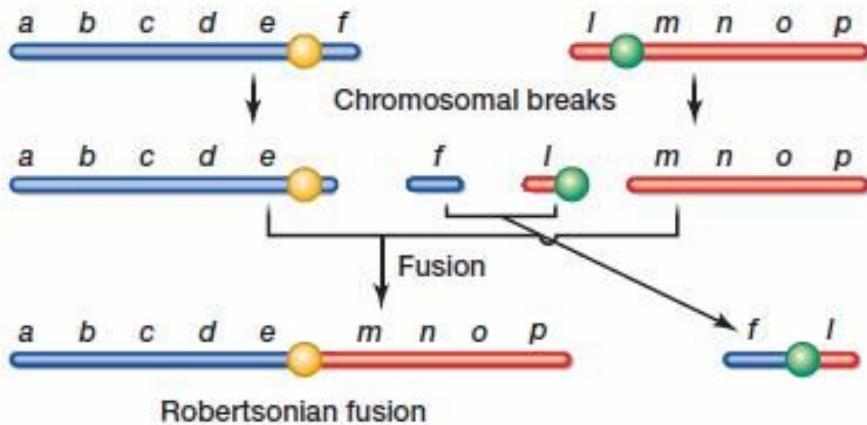


Figure 4.2. Robertsonian translocation or centric fusion

The centric fusion not only results in an abnormal chromosome structure, but the **chromosome number is reduced from 45 to 46**.

Based on current cytogenetic evidence, chromosome number reduction occurred in hominid evolution can be explained by two consecutive centric fusions. While the great apes the gorillas, chimpanzees and orangutans have 48 chromosomes, humans have 46. This means that the centric fusion has had to occur after the line of apes and humans separated during evolution.

4.1.4. Inversions

The inversion is a structural chromosome aberration in which the same chromosome breaks twice and the fragment between the breakpoints turns 180 degrees. There are two types:

- 1 / pericentric
- 2 / paracentric inversion

1 / In **pericentric inversion** the chromosome breakages are on both arms, that is on both sides of the centromere. The pericentric inversion of chromosome 9 is relatively common and found frequently in couples with recurrent abortions.

2 / In paracentric inversion breakpoints are on the same arm of the chromosome, thus in the turn of the fragment the centromere is not involved. For both the para- as well as the pericentric inversion is true that the breakpoints are normally in non-coding regions, so the carriers have normal phenotype. According to the present-day knowledge inversions also played a role in the evolution of human chromosomes, as some human chromosomes derived from the chromosomes of apes and old-world (Catharrinae) monkeys. Moreover, on the basis of presence or absence of such rearrangements (e.g. inv2) the Sumatran and Bornean orangutans are divided into two separate sub-species, and even today they are considered separate species.

4.1.5. *Ring (ring) chromosome*

In this case, there are breaks on both arms of the chromosome - usually near the telomeres - then broken ends fold and a ring chromosome is formed. The fragments broken are lost during successive divisions, so the information encoded by their genes as well. The carriers depending on the chromosome involved and the size of the region lost are more or less severely affected. The somatic retardation - a physical developmental retardation - can be explained by the fact that the DNA replication of the ring chromosome is often erroneous: devil ring, giant ring (due to duplication), rearranged (recombinant) chromosome, or even two ring chromosomes within a cell, i.e. change in chromosome number will be the end result.

4.1.6. *Isochromosome*

The isochromosome is an abnormal chromosome containing the same genes on both arms. Upon formation the sister chromatids are not separated parallel to the long axis of the chromosome, and migrate towards the poles of the cell, but the plane of their separation is perpendicular to the longitudinal axis.

Thus aberrant chromosomes ultimately cells containing them are formed which contain either the short arm or the long arm specific information only on both arms and the information of the other arm is lost.

Since a chromosome arm is rich in many genes, so the surplus or the lack of these lead to severe, often lethal consequences. X and Y chromosomes seem to be exceptional since in most of the known viable isochromosome cases these chromosomes are involved. This is because the Y chromosome is relatively gene poor and the X chromosome inactivation has a compensating effect.

The above mentioned structural abnormalities are usually formed, except duplication, prior to DNA duplication (G1 phase), and therefore replication of damaged DNA leads to identical sister chromatids. Their separation at the end of mitosis results in identical chromosomal aberration carrying daughter cells. Although these types of aberrations can be formed in both mitosis and meiosis, in a strictly medical point of view the latter one is more important because gametes with chromosome mutations can lead to the birth of affected /mutant offspring.

From amongst structural aberrations the translocations and inversions exist not only in balanced, asymptomatic forms. In the case of carriers however, the birth risk of ***chromosomally unbalanced***, physically and mentally retarded child, with severe developmental abnormalities is very high. The symptoms often lead to intrauterine death so to spontaneous abortion or stillbirth. This is due to the difficult pairing of structurally deficient and normal homologous chromosomes in the first meiotic division (Figure 4.3).

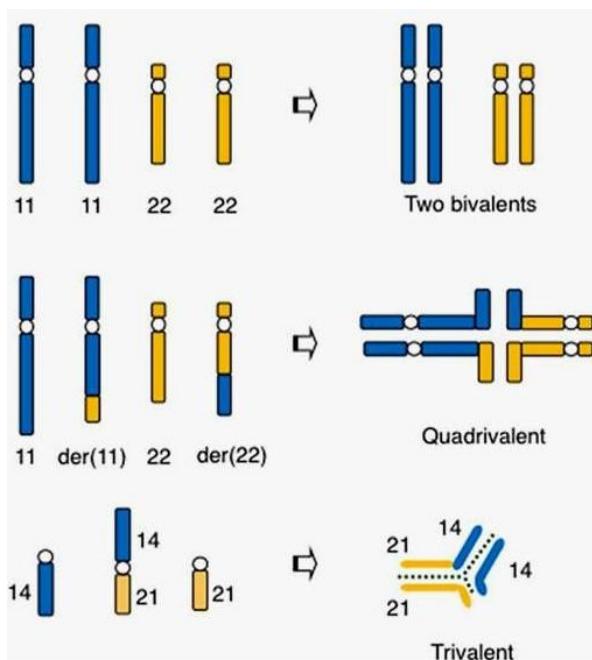


Figure 4.3. Meiotic consequences of some translocations

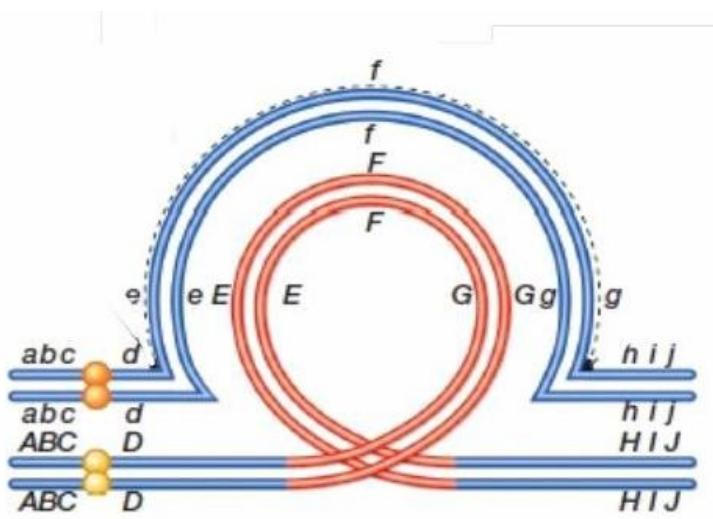


Figure 4.4. Meiotic consequence of an inversion (paracentric)

In meiosis, in reciprocal translocation, instead of the usual bivalent, a quadrivalent and in centric fusion a trivalent, in inversion a loop is formed (Figure 4.4), which makes difficult or sometimes prevents the normal segregation especially if we take into consideration the recombination/crossing over between the homologous chromosomes as well, which likely lead to unbalanced gametes carrying abnormal chromosomes. From a single reciprocal translocation more than 10 different of unbalanced gametes can be derived (Figure 4.5). Some of these are no longer limited to the gain or loss of a chromosome segment - partial trisomy or monosomy - but also have numerical chromosomal abnormalities. Then segregation itself is abnormal, not 2:2, but 3:1 or 4:0, meaning that not 2-2 chromosomes segregate into the daughter cells, but 3 or 4 are present in one of them, while in the other there is 0 or 1.

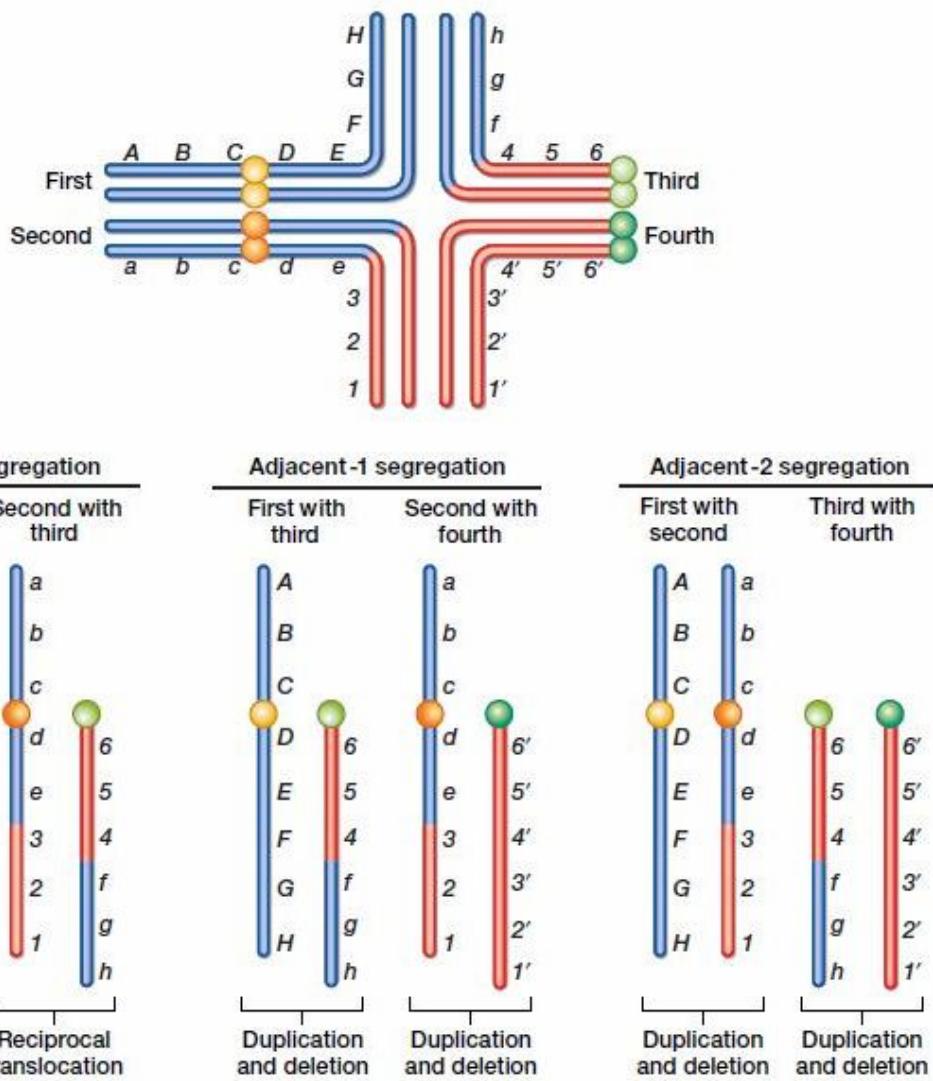


Figure 4.5. Some types of meiotic segregation in a reciprocal translocation

The most severe segregational abnormalities may also inhibit gametogenesis and thus cause **infertility, sterility**. The best example is the centric fusion between homologous acrocentric chromosomes. For example, from t(15;15) or t(14;14) Robertson's translocations viable offspring cannot be born, from t(21;21) centric fusion either unviable or Down syndromic offspring can be born (Figure 4.6).

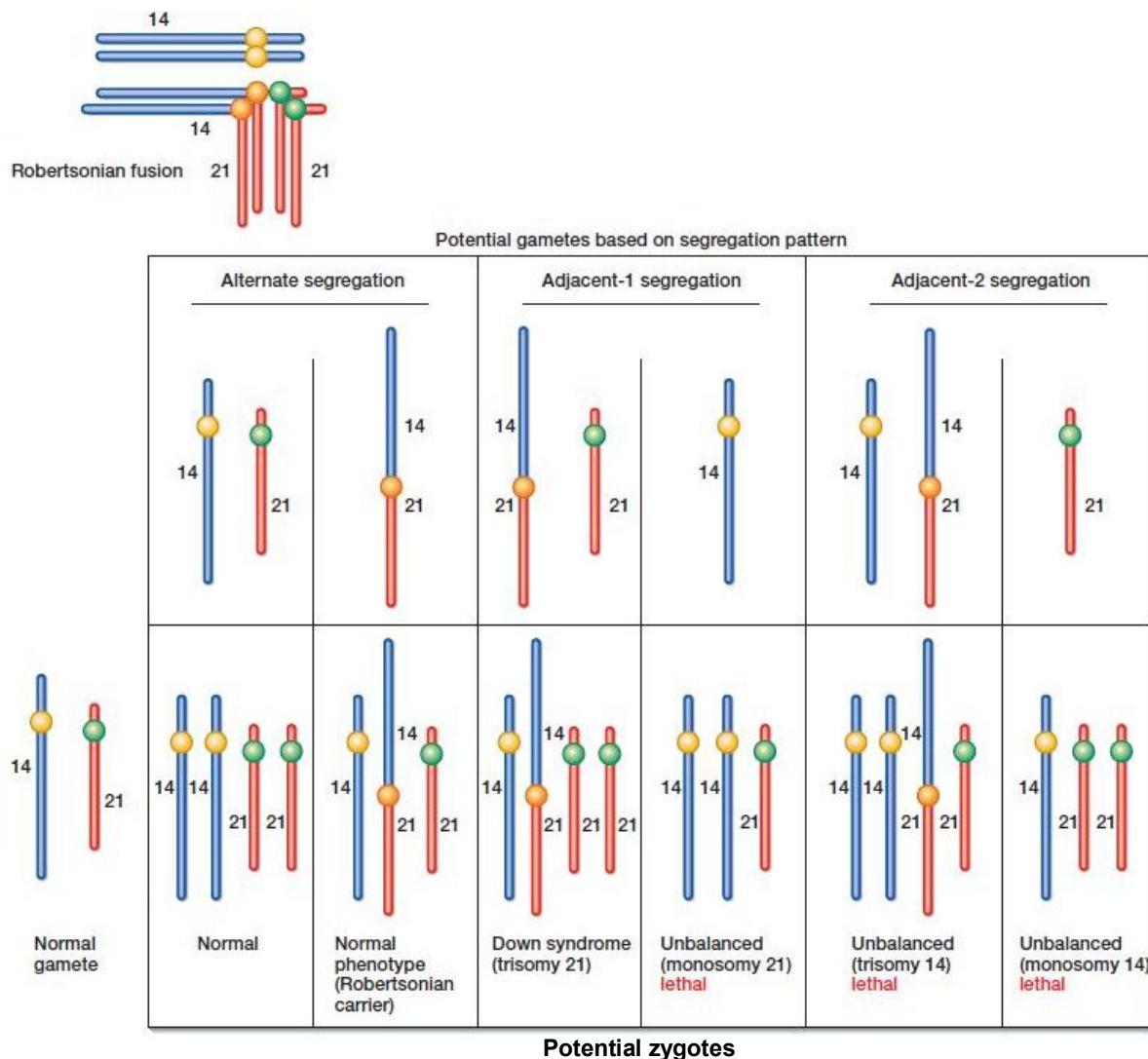


Figure 4.6. Gametes and offspring derived from a t(14;21) centric fusion

4.1.7. Dicentric chromosome

A dicentric chromosome with two centromeres can be created not only by centric fusion - as mentioned above -, but between the short arms of two non-homologous acrocentric chromosomes by non-homologous recombination.

Since dicentric chromosomes cannot migrate to the appropriate poles at time of cell divisions, and through a still unclear mechanism one of the two centromeres is inactivated,

and thus an abnormal chromosome having one centromere is maintained.

4.1.8. Acentric fragment

More rarely broken fragment(s) without a centromere remain in the cytoplasm as small fragments. Due to the absence of centromere such pieces cannot migrate to cell poles and either a so called micronucleus is formed or they are during the subsequent cell divisions, and finally only the deleted chromosomes has retained within the cell. Since these acentric fragments are most commonly induced by some chromosome breakage causing mutagenic agents such as radiation, therefore they can be used for testing the mutagenic effects (micronucleus test).

4.2. Numerical chromosome aberrations

The numerical anomalies, when one or more chromosomes are in excess or missing, ultimately modify the entire genome size, so they can be considered genome mutations as well.

There are three types of numerical chromosome aberrations:

- 1 / euploid
- 2 / aneuploid
- 3 / mixoploid mutations

4.2.1. Euploid chromosome mutations

In the case of euploidy each chromosome is present in the same number, i.e. in a haploid cell everything is present only once, twice in diploids, three times in triploids and so on.

The haploid chromosome number - i.e. typical of the gametes - is n , its exact multiples, that is, $2n$, $3n$, etc. found in the euploid somatic cells. Polyploidy means, if we find a multiple of n , either in gametes or in somatic cells. However, mutations only have arisen if the individuals or just the cell have chromosomes in a number different from the species specific (haploid or diploid) set. **The multiplication of the chromosome set occurs in the M phase of the cell**

cycle due to the defects of microtubules and / or the abnormal organization of mitotic spindle.

In plants, polyploidy is compatible with normal life moreover it is economically quite advantageous, since the multiplication of chromosomes leads not only to the multiplication of genes, but also to the multiplication of their products. Thus the protein content and the crop yield grow, such as banana, wheat etc. In plant cells, the polyploidy found is either the result of that species' evolution or the result of conscious plant breeding work. For the induction of polyploidy spindle poisons - Colcemid, Colchicine, Vincristine, Vinblastine - alkaloids inhibiting the polymerization of spindle microtubules can also be used. In this case ***autopolyploidy*** is created, since every chromosome is of the same species. In contrast, hybrids (e.g. wheat) established by crossing of species or related species are ***allopolyploids***.

Unlike plants polyploidy in animals or in humans is lethal, leads to death in utero. With the exception of certain cells/tissues, for example bone marrow megakaryocytes and a part of the regenerating liver cells which are also polyploids. In these cases it is fixed during the millions of years of evolution what type of cell has the multiplication of chromosome number during ontogeny.

In 10% of the spontaneously aborted fetuses triploidy occurs. Interestingly, 90% of them is of paternal origin, derived either from fertilization by a diploid sperm or from double fertilization. Only a minority comes from fertilization of a diploid egg.

4.2.2. Aneuploid chromosomal aberrations

Aneuploidy is a chromosomal abnormality when only a certain chromosome is in excess or missing. If there is only one chromosome instead of the normal two homologues we are talking about ***monosomy***, if there are three copies ***trisomy*** occurs. If a particular chromosome is not found at all in a cell / organism ***nullisomy*** is present. The latter is lethal both in humans and animals, but in plants is not. Generally speaking, in humans / animals the excess of chromosomes is tolerated better than the chromosomal deficiency. Several somatic and especially sex chromosome aneuploidies - trisomies - occur in live-born, but only one - the X chromosomal monosomy (Turner syndrome) occurs in live-born. The aneuploid mutations are due to ***mitotic or meiotic non-disjunctions***, when the sister chromatids or the chromosomes do not separate in the anaphase – because of the abnormality of the kinetochore, the centromere or both. Less frequently (uniparental disomy) a chromatid / chromosome lagging

behind the others in the anaphase - ***anaphase lag*** - do not get to the right pole, and therefore not to the daughter cell.

Due to this one of the daughter cells is with an extra chromosome, while there is a deficiency in the other. Of course, from medical point of view meiotic non-disjunctions are more important as these lead to defective gametes, and finally to affected offspring.

In the case of ***mitotic non-disjunction***, it is crucial, when and in which cell type's division occurs. The early non-disjunction, eventually involving many cells / tissues leads to severe consequences (mosaicism).

The ***meiotic non-disjunctions*** are grouped according to when they occur – in the first or in the second meiotic division. In the first meiotic non-disjunction, some pairs of homologous chromosomes are not segregated, whereas in the second meiotic non-disjunction - as in mitotic non-disjunctions the sister chromatids are not separated. These have different consequences accordingly.

Following ***the first meiotic non-disjunction*** all four progeny cells - in spermatogenesis the four sperms - will have an abnormal chromosome set – will be aneuploid. Two is with an additional chromosome (n+1); two is without one (n-1). In ***the second meiotic non-disjunction only the half of the daughter cells*** are affected. They will also be with an extra or an absent chromosome.

The fusion of such abnormal gamete with a normal one results in trisomic or monosomic zygote.

In the case of trisomies there is difference in the origin of the three homologues depending on in which meiotic division the mutation took place. ***Trisomies derived from the first meiotic division all three homologues are of different origin*** (e.g. one is from the maternal grandmother, the other is from the maternal grandfather, and the third is inherited from the father). However, ***in trisomies from the second meiotic non-disjunction two homologues are identical*** (e.g. either from the maternal grandmother or from the maternal grandfather) and ***only the third comes from the other parent***, from the father. ***70% of the human aneuploid chromosome mutations are derived from the first and 30% from the second meiotic non-disjunction.***

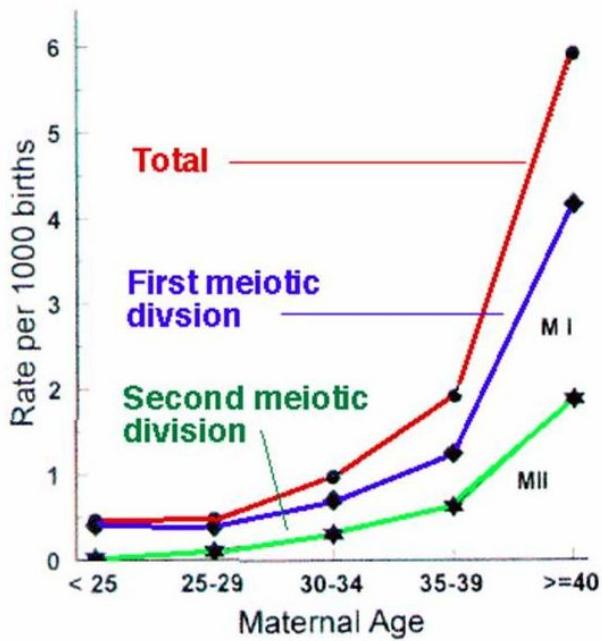


Figure 4.7. Rate of meiotic non-disjunctions in function of maternal age

So most of the meiotic non-disjunctions occur during the first meiotic division and are of maternal origin. The frequency of maternal non-disjunctions and the aneuploid offspring (like Down syndrome) - increases with maternal age (Figure 4.7). The reason of this lies in the characteristics of female gametogenesis: probably the aging of the synaptonemal complex, which reduces the chance of co-segregation of homologues leads to the formation of gametes with abnormal chromosome number.

This is why above a certain maternal age (35-40) prenatal tests are recommended or required, to determine whether a fetus carries a numerical chromosome aberration or not.

4.2.3. The most common numerical chromosomal abnormalities

All chromosome trisomies except the one of the largest human chromosome (chromosome 1) were found in spontaneously aborted fetuses, among live born three autosomal trisomies and some involving sex chromosomes occur (Figure 4.8). Two things are suggested by this: first, the fate of trisomies is strongly dependent on the number and type and function of genes present in the chromosome, on the other hand there is a strong intrauterine selection, so the most severely affected fetuses die in utero. These are confirmed by the fact that in spontaneously aborted fetuses the most common abnormality is the trisomy 16, which although affects a relatively small chromosome, is never found in live born! All the monosomies, with the exception of the X chromosomal monosomy are incompatible with life.

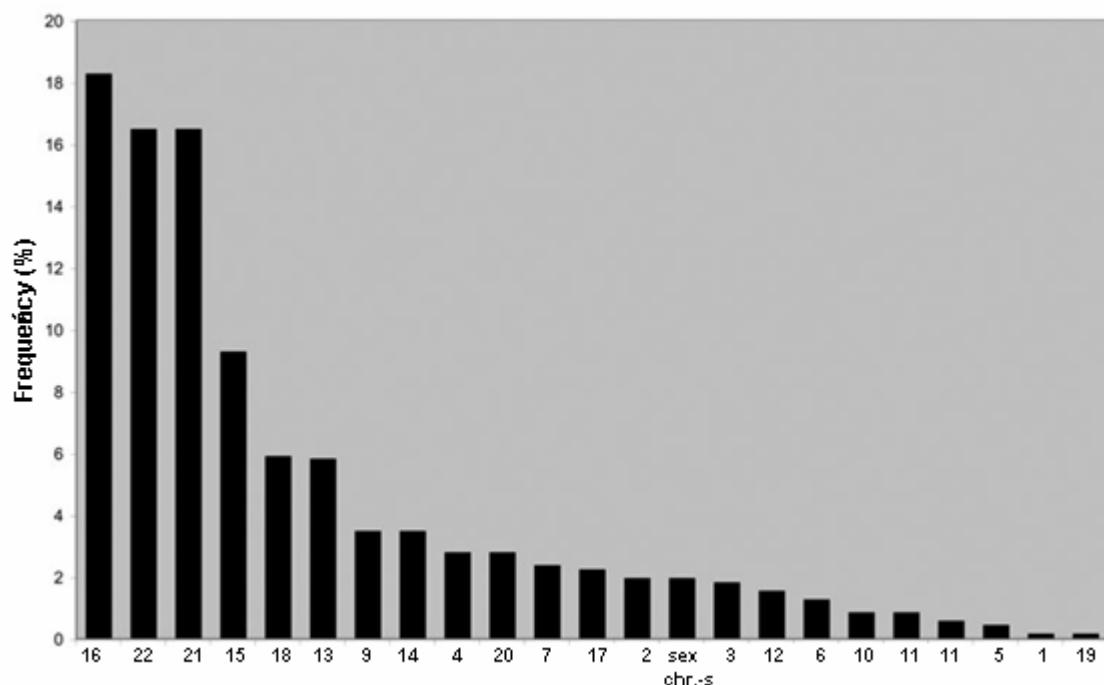


Figure 4.8. Autosomal numerical chromosome aberrations

4.2.3.1. Trisomy 21

Trisomy 21 is the cause of **Down syndrome**. Although the non-disjunction of chromosome 21 is not the only cause of Down syndrome -a smaller proportion of the cases is due to either centric fusion or translocation - it is the most common type. Despite the fact that trisomy 21 fetuses die in utero the average population frequency of Down syndrome is 1:650, but this value increases dramatically with maternal age, at 45 years of age it is more than 1:100!

Although today live-born trisomy 21 patients have more or less the same life expectancy than healthy individuals, but the leukemia and some other disease prevalence is higher among them than in the general population. In recent decades, there is a significant change in the social status of Down syndromic individuals, whereas before they were excommunicated, teaching them was thought to be impossible, now increasing efforts have been made to facilitate their social integration (e.g. special kindergartens, in many countries they are taught together with healthy children in public school classes, sporting events etc.).

4.2.3.2. Trisomy 13

Trisomy 13 is the **Patau syndrome**. Similar to Down syndrome it is most commonly derived from maternal non-disjunction. 65% of such non-disjunctions derived from the first meiotic division. Frequency of birth is 1:12 500 - 1:21 700. Only <5% of these infants survive the first year of life.

4.2.3.3. Trisomy 18

Trisomy 18 is the **Edwards syndrome**. It is primarily due to maternal non-disjunction. 95% of the cases are due to non-disjunction in the first meiotic division. The frequency is 1:6000 - 1:10000 live-born but the frequency at the time of conception can be much higher, since approx. 95% of the fetuses die within the womb. 30% of the Edwards syndromic abnormal newborns die within one month, > 95% of them die within a year.

4.2.4. Numerical sex chromosome aberrations

4.2.4.1. Turner syndrome

The **Turner syndrome** is characterized by **45,X0 karyotype. This is the only viable monosomy**. The explanation for this lies in the fact that while both homologues of the autosomes are necessary to the normal phenotype - so their monosomy is lethal – by contrast in females only one X chromosome is active (see the dose compensation in X inactivation), so a functional monosomy and Barr body negativity can be maintained. However, for the normal development of female sex characteristics both X chromosome is needed, as indicated by the symptoms of Turner syndrome.

Although the frequency is 1:5000 in newborn female infants, the conception rate is much higher, but 99% of them spontaneously abort. This is in good agreement with the concepts of the viability of monosomies. 80% of these cases are due to paternal meiotic non-disjunctions, therefore in these patients only one maternal X chromosome is present.

While it is easy to understand the sexual development related characteristics of the syndrome, the low height is still not fully explained. It is assumed that a gene coding a protein of the small ribosome subunit (*RPS4X*) may also play a role. Because this gene has a Y chromosomal counterpart (*RPS4Y*) as well, both in normal females and males double dose of this ribosomal protein is produced. In Turner syndromic individuals less than sufficient

amount is produced, and if the ribosome number is less than normal it will largely influence the production of other proteins, and thus indirectly the body height, too.

Although Turner syndrome is often characterized by normal intelligence there is a difference in verbal skills, social integration between patients inherited their X chromosome from father or the mother. Maternal X carriers, according to surveys are weaker in these features than the patients inherited paternal X. The phenomenon is explained by the different methylation of the two types X chromosome and the genomic imprinting.

4.2.4.2. *Klinefelter syndrome*

Klinefelter syndrome is characterised by 47,XXY karyotype and male phenotype. The frequency is 1:1000. Nearly it is derived with the same probability from maternal (56%) and paternal (44%) non-disjunction. 36% of the maternal non-disjunctions take place in the first meiotic division. Since there are two X chromosomes, thus they are Barr body positive. Their sterility can also be attributed to presence of 2 X chromosomes, since certain X chromosomal gene products are in a higher dose than in normal fertile males.

4.2.4.3. *Triple X syndrome*

Feminine phenotype and 47, XXX karyotype are present. 89% is of maternal, 8% is of paternal origin, and the remaining 3% is due to post-fertilization mitotic non-disjunction. Neonatal frequency is 1:1000. Two Barr bodies are typical.

4.2.4.4. *Double-Y syndrome, "superman" or Jacobs syndrome*

In this case ***normal, slightly taller than the average males have 47,XYY karyotype.*** The birth rate is 1:1000. They derived only from paternal second meiotic non-disjunction. In contrast to all meiotic non-disjunctions, ***the formation is not affected by age,*** as paternal gametogenesis is continuous from puberty, there are no aged sperms.

They are also featured by poorly tolerated frustration and stronger aggressivity; perhaps that is why this chromosome abnormality is found in greater numbers amongst imprisoned men. The aggressiveness and the possible criminal tendency are strongly debated, and it would only be 100% decided if the entire male population would be karyotyped and comparative data about their aggressivity would have been available as well. Today many

different aggressiveness associated genes and gene mutations are known, that is why the role of the Y chromosome in aggressiveness is questioned.

Knowing the characteristics of meiotic division we could ask that the two aneuploidies (47,XXX and 47,XYY) with normal fertility are characterized by greater prevalence of similar disorders among offspring or not. For example, in the case of double Y syndrome the following karyotypes offspring are expected in the offspring: 2 XXY, 2 XY, 1 XX and 1 XYY. In contrast, birth of only normal offspring was reported so far, however its exact explanation is still not known.

4.3. Uniparental disomy (UPD)

This abnormality which is not or hardly identifiable by cytogenetic methods were recognized - due to molecular biological techniques - in the past decades. The UPD means that the person concerned has a normal chromosome number, but *the homologues of a certain chromosome – in contrast to normal – are from the same parent, either from the father or from the mother*. As for the formation two consecutive numerical aberrations are in the background: a meiotic non-disjunction and an *anaphase lag* occurring during the early cleavage divisions. So in fact a trisomic zygote is formed first, and subsequently the 3 homologue is lost. Depending on whether first or second meiotic non-disjunction occurred, *uniparental heterodisomy or uniparental isodisomy* is present. The first case is when the child inherits two different homologues from the parent (one maternal and one paternal), that is non-disjunction occurred in the first meiosis. The latter is when the two homologues inherited are the same (either both are maternal or paternal) suggesting a second meiotic non-disjunction.

In UPD depending on the parental origin of the homologues, and due to genomic imprinting, different symptoms may be seen. The different symptoms in some of the Prader-Willi and Angelman syndrome cases are not due to the 15q deletion, but the UDP.

4.4. Mixoploid mutations

In mixoploidy or in mutations associated with mixed ploidy usually two (sometimes more) cell lines with different chromosome numbers are found within an organism. There are two forms: *mosaicism and chimerism*.

4.4.1. Mosaicism

In genetics a mosaic is a living creature, where two cell lines of different chromosome numbers, but of the same origin are present in the body. They are either ***aneuploid or polyploid mosaics***.

The former occurs as a result of mitotic non-disjunction or anaphase lag during cleavage, two cell lines of different chromosome number are formed, when one is normal and the other is aneuploid, generally trisomic. For example, assuming a two-cell embryo, if one cell is divided normally and the other is abnormally, then finally 2 normal and 1 trisomic and 1 monosomic cells are present. Since the monosomic cells are not viable eventually the ratio of the normal and the trisomic cells will be 2:1.

In the case of ***polyploid mosaicism*** a normal and a polyploid (generally triploid / tetraploid) cell line are present. In this case, however mitotic spindle error leads to the formation of the aberration. Again, assuming a two-cell embryo, if one divides normally, the other not, ultimately there will be 3 cells instead of four, and two are normal and one is tetraploid.

Depending on the time the aberration occurs (during cleavage or in organogenesis or even later in development), the symptoms become more or less severe. So the proportion of normal and defective cells is crucial. Mosaisms involving sex chromosomes are relatively common.

In the case of ***gonadal mosaicism*** only the cells in the germ line have abnormal chromosome number, thus the risk of numerical aberrations in the offspring is high. Unfortunately, the detection of such defects is still not possible routinely, but the birth of an abnormal offspring of the patient can indicate this. Mosaicism in a broader sense is a somatic mutation, when different mutants (alleles) of a given gene are located in different organs or in different cells of the same organ (for example eyes with different colors: one is blue and the other is brown or a blue eye with brown spots).

4.4.2. Chimerism

After the lion-headed, bird-legged, snake-tailed monster of Greek mythology the creature that has two cell lines of different origin - derived from different zygotes - is called ***chimera***. A chimera is derived either from fusion of fraternal twins, or from double fertilization of an egg and a polar body (polocyte), or from transplacental haematopoietic stem cells exchange between fraternal twins (blood group chimerism).

Recently, the chimera referred to as transgenic animals / plants, which contain cells of different origin, derived from either the fusion of few-cell-embryos, or via the microinjection of foreign genes into fertilized oocytes.

Recently several publications dealt with the phenomenon of ***microchimerism***. It has been known for 25 years that in maternal body after being pregnant with a male fetus - after giving birth, and even after abortion - Y chromosome carrying or Y body-positive cells can be detected in the bloodstream. It is now found that these foreign cells in maternal body detected many years (decades!) later, not only survived, but probably also proliferated. This means that stem cells of the male fetus were transferred by the bloodstream to the mother's body, where they reached and adhered to certain organs and formed cell clones. Therefore a hypothesis is also suggested that some putative autoimmune diseases are actually not autoimmune but against the cells - 50 % foreign to female body (immunologically incompatible) immune reaction are awoken. This also serves as an explanation why autoimmune diseases are more common in women. However, transplacental cell migration in the opposite direction (from mother to fetus) - cannot be excluded, and may play a role in the tolerance against alloantigenes, although its mechanism and consequences are not well known.

4.5. Useful web-sites:

- [www.nlm.nih.gov / MedlinePlus / geneticsbirthdefects.html](http://www.nlm.nih.gov/MedlinePlus/geneticsbirthdefects.html)
- [www.mayoclinic.com / health-information](http://www.mayoclinic.com/health-information)
- [www.rarechromo.org / html / ChromosomesAndDisorders.asp # ANAL](http://www.rarechromo.org/html/ChromosomesAndDisorders.asp#ANAL)
- www.trisomy18support.org
- www.t21online.com
- www.williams.org
- www.turnersyndrome.org
- www.klinefeltersyndrome.org
- www.pwsausa.org

4.6. Questions

1. What are the causes of aneuploidy and polyploidy?
2. What are the main regions of chromosomes?
3. Explain the low incidence of monosomies!

4. What is microchimerism, and what is its biological significance?
5. In what diseases has UPD an etiologic role?
6. What are the different positions of chromosomal breakpoints?
7. What techniques are used for the detection of chromosomal aberrations?
8. What are chimerism and mosaicism?
9. What are the possible consequences of centric fusions?
10. What is the explanation of the higher frequency of first meiotic non-disjunctions?

5. Epigenetics

Sára Tóth

In recent years, epigenetics has become one of the fastest growing areas of genetics. In this subject - according to PubMed database - last year only, over 10 000 scientific papers have been published.

The term epigenetics is connected to Conrad Waddington, who in the early '50s when studying the processes of ontogeny talked about a so-called epigenetic landscape when he tried to explain how an extraordinary variety of cells can develop from a single cell, the zygote. Although they are the same genetically, but morphologically, functionally different, due to what point of the scenery they reach (mountain, valley or slope), so how the gene could be regulated during development. Today, those mitotically and / or meiotically transmissible processes are called epigenetic phenomena that alter the function, so the expression of each gene, without affecting the DNA sequence itself, that the changes in gene expression are not due to mutations.

The range of these phenomena and the known enzymes and regulatory proteins involved in these processes is expanding, and epigenetic changes related to almost all aspects of life have been reported. Parallel with the increasing knowledge of epigenetic processes, many previously unexplained observations, phenomena became understandable.

5.1. *Epigenetic changes - molecular modifications*

In the toolbar of epigenetics it has an essential role to alter the building blocks - DNA and histones - of chromatin. The resulting epigenetically modified DNA, and the variously modified histones attached depending on the modifications also attract different non-histone proteins thus ***chromatin is fundamentally affected and remodelled***. The chromatin has two functional stages: heterochromatin, representing a closed, nontranscribed state and euchromatin which is a loose, open structure accessible to components involved in transcription. Epigenetic modifications make a further, more refined control possible.

5.1.1. DNA methylation

The epigenetic modification of DNA is methylation of cytosine when 5-methyl-cytosine (5MeC) is created. In this case methylated cytosines are almost exclusively located in so-called **CpG dinucleotides**. CpG dinucleotides are covalently linked cytosine and guanine bases in one of the DNA strands. Methylated cytosine pair guanine as the unmethylated and therefore the information coded by the DNA remains unchanged. However, the methyl group of the 5MeC facing the major groove of DNA, and therefore it is accessible to the different DNA-binding proteins. CpG dinucleotides are found mainly in the promoter region of the genes, but depending on the gene, "inside" the gene in the exons and introns can also be found. Not all CpGs are methylated it also depends on the cell type, its metabolic condition how these CpG dinucleotides are methylated, i.e. what is the **methylation pattern**. The methylation of CpGs of promoter provides a basic regulation of gene expression: the methylation usually (but there are exceptions) leads to inhibition of gene expression. Since epigenetic marks are transmitted from cell division to cell division but from generation to generation are usually not, this means that the DNA methylating enzyme system is specialized accordingly. There are two main methylating enzymes known: **the maintenance DNA methyltransferase (DNMT1) and de novo DNA methyltransferase (DNMT3)**. During DNA replication **DNMT1** methylates cytosines of the CpGs in the new complementary DNA strand - in accordance with the old strand, thereby maintaining the original pattern of DNA methylation. The **DNMT3** can methylate cytosines which had not previously been methylated. This is important in gametogenesis when the original pattern inherited from the parents erased and a new methylation pattern - appropriate for the sex of the organism - is built up. **DNA demethylases** are involved in the removal of methylation patterns.

5.1.2. CpG as mutation hot spot

Cytosine spontaneously deaminates to uracil. This instability can also apply to methylated cytosine but in this case not uracil but thymine will be the result. Thus **from a CpG dinucleotide a TpG dinucleotide is formed, and it has been a change in the DNA sequence, i.e. a mutation**. Analysis of mutational databases revealed that for a number of diseases CpG dinucleotides are mutational hot spots.

The chemical lability of cytosine, i.e. its mutability is shown by the fact that, although methylation is a general feature of DNA the frequency of methylated cytosines is much lower than expected. In humans, only 3% of cytosines are methylated. It seems that in a longer

evolutionary interval CpG frequency slowly but gradually decreases due to the constant CpG → TpG transformation. Despite the small CpG frequency in vertebrate genomes, there are short non-methylated DNA sequences, which CpG value correspond to the expected frequency. These so called ***CpG islands*** are CG rich and often found in several hundred nucleotides long stretches at the 5' end of the genes. Scattered in the human genome around 27-30 thousand CpG islands are found. For these areas, the CpG → TpG transformation is not typical. The CpG stability is either due to that there can be no methylation, or that these islands are functionally so important that natural selection prevents their loss. **There are CpG islands in the promoter of approx. 50% of the human genes, which are generally unmethylated. Their abnormal methylation is pathological and can lead to the formation of tumors by changing the regulation of gene expression** (see Genetics of biological processes).

5.1.3. Histone modifications

In addition to DNA methylation the role of histone modifications is also essential for epigenetic processes. Histones are evolutionarily highly conserved, DNA binding, basic - lysine and arginine rich - proteins. The H2A, H2B, H3 and H4 histones in 2-2 copies are involved nucleosomal octamer structure, while the H1 histones bind to the so-called linker DNA connecting the nucleosomes to each other. The N-terminal tail of the nucleosomal histones protrudes, and this is the area where histone modifications occur. The main targets of histone modifications are primarily the ***lysine*** residues with suitable positions in histone H3 and H4 tails which can be methylated, acetylated, phosphorylated, and ubiquitinilated etc. These modifications form a ***histone pattern***, depending on the cell type, its state of development and physiological function and of the gene or gene sequence in question. **This pattern is the so called histone code, which fundamentally affect the expression of the area concerned.** Both methylated DNA and modified histones attract several methylated DNA - or histone-binding proteins and non-coding RNAs as well and the interacting members of the complex formed determine the epigenetic pattern characteristic for developmental stages, cells and genes. Any mutation in any element of this complex can lead to abnormal epigenetic signals and thus malfunction or disorder. **An example of this is the Rett syndrome, where mutation of the *MECP2* (methylated cytosine binding protein) gene is in the background.**

5.2. Non-coding RNAs

Following the success of the human genome project it has become apparent that only <2% of the human genome is protein-coding so rightly ask the question: what is the function of the others?

Today we know that an important part the genome is to determine the smaller or longer regulatory RNAs (see also Chapter 9). These RNAs through RNA-RNA, RNA-DNA, RNA-protein interactions can modify the expression of genes, without modifying their DNA sequence. Such RNAs can affect transcription by inhibiting it, e.g. Xist RNA (see below), but may act post-transcriptionally e.g. microRNAs, which inhibit translation of mRNA. The non-coding RNA can be derived either from the same chromosome where the gene is regulated (*cis-action*) or from another chromosome (*trans-action*).

5.3. Epigenetic phenomena

The most important epigenetic events are: ***genomic imprinting and X chromosome inactivation***. Furthermore, **carcinogenesis, aging, some psychiatric disorders, and even conduct disorders can be associated with epigenetic processes**. Some observations indicate that these changes - not altering DNA sequences, but affecting its function - can be transmitted not only somatically so from mitosis to mitosis, but instead through meiosis during gametogenesis to gametes, and thus after the fertilization they will be typical of the progeny as well so there are signs of the so called ***transgenerational epigenesis***.

5.3.1. X-chromosome inactivation

Mammals are diploid organisms and consequently both alleles of an autosomal locus responsible for a particular trait are functioning, i.e. biallelic gene expression is characteristic for them. If one of the autosomes is eliminated due to chromosome mutations so only one allele remains at a single locus, this generally results in severe or lethal symptoms. In contrast, the sex chromosomes form homologous pairs (XX) only in females, whereas in males the Y chromosome is not a functional homolog of the X chromosome. While the Y chromosome contains only a few genes mainly responsible for sex determination (*SRY*) and gametogenesis (e.g. *AZF*), the X chromosome has genes determining a large number of somatic traits. If genes coded by the single X chromosome of males are sufficient for the normal development and normal physiological processes of the individual, it is obvious that one X chromosome has to be enough for the female body. This means that evolutionarily became necessary **to**

equalize the different X chromosomal gene doses of the two sexes, so to compensate the dose differences. This *dose compensation* is also called *Lyonisation* after Mary Lyon, the scientist who described this phenomenon. In mammals, including humans, the dose compensation in females is achieved through the inactivation of one of the X chromosomes. However, it should be emphasized that other mechanisms also exist for dose compensation in other organisms with heteromorphic sex chromosomes. X chromosome inactivation takes place at the beginning of mammalian embryonic development, in the blastocyst stage. As a first step of dose compensation **chromosome counting** is carried out by a mechanism not yet fully understood in details. This means that the cell acquires the information about the quantity of X chromosomes. Where two or more X chromosomes are in the cell, only one is active and the other (or others) is (are) inactivated. This *inactivation is random*, i.e. either the maternal or the paternal X is inactivated. However, once the selected X is inactivated, the given status of the cell is maintained lifelong in each of its descendant daughter cells.

Due to the random X inactivation in females there are cells in which the maternal, and there are in which the paternal X chromosome becomes inactive. Thus a so-called ***functional mosaicism*** is typical for women. The inactive X chromosome is intact, most of its genes are not expressed, except the genes of the pseudoautosomal regions found near to the telomere of both arms of the X chromosome (PAR1 and PAR2 regions) and the ***few genes that escapes X inactivation***. These remain active on the inactive X chromosome. To find out why these genes remained and how, is the subject of intensive research today. Although the inactivation in somatic cells is passed from daughter cells to their progeny cells, it does not mean that it is the case in germ cells, too. During oogenesis the inactivated X chromosome is reactivated, and it remains active in the mature gamete. In X inactivation the so-called ***XIC = X inactivation center*** plays a crucial role. Here, in the Xq13 region the ***XIST*** gene is found that is transcribed only from the inactive X chromosome. The ***product is a large non-coding RNA*** which is covering the would-be inactive X by a not yet fully known mechanism. The *XIST* expression is followed by several other epigenetic events such as DNA methylation, histone methylation, a change in the histone composition as shown by the macroH2A histone variant, increased chromosome condensation, and ultimately late DNA replication will be characteristic to the inactive X chromosome, i.e. the DNA of the inactive X chromosome starts replication after the replication of the other chromosomes' DNA. The ***increased chromosome condensation leads to heterochromatinization, and then the transcription is inhibited, thus the chromosome becomes inactive.***

5.3.2. Genomic imprinting

Based on the classical genetic experiments, it appeared that it does not matter even in heterozygosity which allele comes from which parent. Since both alleles are expressed, the origin is not important. However, some animal studies or rare human diseases, suggested that it is at least not true for each gene. During mouse embryo manipulations it was found that when the nucleus of a mouse oocyte was injected into another oocyte of the same mouse, then diploid cell created, a ***gynogenote*** just started the embryonic development, but soon died because the fetal membranes were not formed. When the experiment was repeated in a way that an enucleated oocyte got two sperm nuclei, although the embryonic development was also not normal, it was different from the former phenomenon. There was no embryo in such ***androgenote*** only hyperproliferated fetal membranes. In other words, on the basis of mouse experiments it is concluded that maternal and paternal halves of the genome are not functionally equivalent. Rare human diseases such as ***complete hydatidiform mole*** also suggest this. In this case only paternal chromosomes are found in the otherwise diploid sample. That is because an empty egg is fertilized by either two sperms or by a diploid one, which is also supported by the fact that the sample proved to be homozygous for all loci by further tests. The reverse of this was found when ***teratocarcinomas*** were analyzed the abnormal tissue had only maternal chromosomes. On the basis of these initial observations it is considered that each chromosome carries a marker which refers to the parental origin. This signal is fixed at some point during gametogenesis, which somehow imprinted in the genetic material. The parental origin specific marking of the genome is called ***genomic imprinting***.

To identify the mechanism of imprinting further attempts were made. When the mouse experiments were repeated in a way that only one pair of chromosomes or a distal or a proximal part of the chromosome was purely of paternal or maternal origin, it is found that not the total genome, but only certain chromosome segments, certain genes carry markers of parental origin. Such a phenomenon is the **uniparental disomy (UPD)** (see Chapter 4, Cytogenetics), where after rare chromosome segregation anomalies, regarding the chromosome number, normal diploid organisms are created with both homologues of certain chromosomes / chromosome segments derived from the same parent, and who have severe symptoms, depending on the chromosomes involved. From the point of DNA base sequence it is irrelevant, which comes from which parent, therefore the labeling should be epigenetic. **If a father transmits an imprinted gene to his child, which he inherited from his mother, then the gene carries maternal imprinting in the father, but the kid will inherit paternal imprinting of this gene.** This means that ***imprinting is reversible***. That is similar to X

inactivation, where an epigenetic mark or pattern resulting in imprinting is inherited without further changes in somatic cells, but **in germ cells the original inherited pattern is erased**, and in the individual -appropriate to the gender - a new female or male epigenetic pattern, imprint is built up. To our knowledge, there are about 100 imprinted genes in humans, which generally play a role in ontogeny - especially around the implantation period – in growth and behavior. These genes are not completely dispersed in the genome, but form groups, so-called differentially imprinted regions (clusters). In mice, chromosome 7, in humans 11 and 15 are particularly rich in imprinted chromosomal regions.

5.3.2.1. *Imprinting related diseases*

The research of the imprinting related diseases is still in its infancy, as many very finely tuned mechanisms may lead to the development of these. The best known disorders due to imprinting are **Prader-Willi and Angelman syndromes**, where the 15q11-q13 region is affected. While in Prader-Willi syndrome the maternal UPD or paternal deletion of the above mentioned region is the cause of the disease, Angelman syndrome may be caused by maternal deletion or paternal UPD of this region or by mutation of the ***UBE3A*** (ubiquitin ligase) also located in this region. Moreover, in both cases, mutations of ***the center responsible for imprinting (IC)*** occur. Prader-Willi syndrome is characterized by obesity, small hands and feet, underdeveloped genitalia, mild mental retardation. Symptoms observed in Angelman syndrome are quite different. Developmental retardation, compulsive movements, laughter (that is why the disease was formerly called happy puppet syndrome), poor speech ability or complete inability of the speech are the characteristic features.

More known rare diseases related to imprinting:

- a. Beckwith-Wiedemann syndrome: in which two differentially imprinted regions (clusters) can be found in the 11p15.5 region with the *H19*, *IGF2* and the *KCNQ10T* genes. The former cluster is associated with childhood kidney tumor disorders (Wilms' tumor, Beckwith-Wiedemann syndrome). During tumor formation loss of heterozygosity occurs in renal tissue (LOH; see chapter 3), but then almost always (over 90%) the maternal allele is lost.
- b. Silver-Russell syndrome (7p11.2 or 11p15.5);
- c. the pseudo-hypoparathyroidism (20q13.2)
- d. transient neonatal diabetes mellitus (6q24).

5.3.2.2. Evolutionary theories of imprinting

There are several theories to explain the evolutionary origin of imprinting. One of the most well-known is the so-called ***conflict of parental interests*** theory. Thus, fathers will be able to spread their genes best when there is a lot of offspring. If, the maternal body is exhausted due to the numerous births, or the mother dies, the father can produce more offspring with another partner, and further transmits his genes. In contrast, the mothers interests are to save resources, i.e. that not one child will use all maternal resources they also can survive additional reproductive cycles and eventually successfully transmit their genes. That is, the paternal genes stimulate - even at the expense of the mother - fetal growth, while the maternal ones restrict the fetal access to nutrient resources. This concept is well suited to the hydatidiform mole and the observations in the case of ovarian teratomas and by the fact that the *IGF2* (insulin-like growth factor 2 = insulin-like growth factor 2) and its receptor (*IGFR*) genes are imprinted as well. The imprinting of these two genes is specific: paternal *IGF2* gene is weakly, and the *IGFR* gene is highly methylated, while in the mother the opposite is observed. The significance of this is that the effects of the two parents equalized in this way more growth factor is in vain when the amount of the receptor is reduced. As the Prader-Willi and Angelman syndromes do not confirm the above theory, therefore, there should be other unknown reasons for imprinting. According to one of these new theories upright position and balance shift during pregnancy may have a role. The maternal imprinting restricts fetal growth and thus shifting the center of gravity during pregnancy, thus making the upright posture and walking more stable, which could be crucial and life-saving for early human ancestors.

5.4. The significance of epigenetic effects

DNA methylation and subsequent changes in histone code and in chromatin structure are essential for gene regulation. We can assume that these mechanisms are crucial in establishing and maintaining cell and tissue identity. A good example was given by examining the tumor cells, where hypermethylation of the tumor suppressor genes and the consequent inhibition of their expression or hypomethylation of oncogenes and the increase of their activity were observed. Both changes may play an important role in oncogenesis. In addition to the carcinogenesis assisted reproductive technologies also highlighted the role of epigenetic

changes. The accumulated data since the birth of Dolly, the first cloned mammalian suggest that the cloned mammals are not normal e.g. often larger than the normally conceived animals, and their neonatal mortality is more frequent as well. Since the genetic material of the nucleus from adult organism used for cloning via nuclear transfer to the enucleated oocyte previously undergone a series of epigenetic changes to function normally these changes should be reversed after implantation. However, in this *epigenetic reprogramming* the oocyte cytoplasm is involved as well, it seems that under these artificial circumstances it does not work perfectly: the reprogramming is usually incomplete and imperfect.

The importance of epigenetic reprogramming is shown by the higher frequency of Beckwith-Wiedemann syndrome of offspring conceived through IVF procedures. Although the frequency of 1:5000 is not too high, it's higher than the value observed in the naturally conceived offspring. Probably artificial conditions of IVF techniques are not favourable to epigenetic reprogramming. However, the epigenetic changes play a crucial role in adaptation to the environment. The importance of environment in epigenetics is proven by *twin studies*. **The epigenetic similarity of identical twins, (e.g. their DNA methylation patterns and histone modifications) is very high, but it decreases as they are getting older, due to increasing epigenetic differences induced by their different environment, lifestyle and diet.**

A controversial theory based on epidemiological studies is the *transgenerational epigenesis*. Swedish studies have also associated the nutrient supply of the father and paternal grandparents in childhood and the proband's life-span or mortality due to diabetes or cardiovascular diseases. (Recent animal studies showed a correlation between the parental high-fat diet and the obesity and diabetes of the offspring). Others described a relationship between the age when fathers started smoking and body mass index (BMI) of their 9-year-old offspring. These observations are very difficult to explain at present, especially on the maternal side, where metabolic signals transferred via the placenta should be taken into consideration. However, on the father's side the sperm-mediated epigenetic transmission is easier to interpret.

From the point of view of transgenerational epigenetic processes the role and the delivery of modifications created by such environmental effects as diet and environmental pollutants (e.g. pesticides) to the offspring are particularly interesting and thought-provoking. Since folates play a key role in the synthesis of methyl donors required for DNA methylation, so it is understandable that the content of dietary folate has an epigenetic importance as well. It could be justified by a murine experiment later become famous. The wild-type mice' fur color is the

so-called ***agouti*** (a peculiar brownish-gray color), besides there is an A^{vy} (**viable yellow**) mutation causes a yellow coat color. **This allele is metastable, as it leads to yellow fur only in unmethylated state, in methylated state an unchanged agouti coat color develops.** In addition, in heterozygotes ($^uA^{vy}A$) the non-methylated mutant allele is dominant, whereas the methylated form is recessive concerning the wild-type allele. The $^uA^{vy}$ unmethylated allele is dominant against the methylated $^mA^{vy}$ allele, so the coat color in the animals homozygous for of A^{vy} is in function of the methylation of the alleles. In a conclusive experiment homozygote (AA) mothers were crossed with heterozygous ($A^{vy}A$) fathers. During pregnancy in one group the maternal diet was rich in methyl donors, while the others got normal diet. In the first group the majority of heterozygous offspring was agouti-colored or smaller and less yellow spots were seen on them (referring to the limited expression of the non-methylated mutant allele). A normal diet resulted in exactly the opposite effect, there were more completely yellow or large yellow spotted in the heterozygous offspring. In this case, the effect was observed in the of F_2 generation as well, but not in later generations. This metastable mutation has other consequences e.g. the yellow furred animals are generally fat and have higher frequency of tumors, so it further confirms **the relationship between epigenesis and oncogenesis.**

In another series of experiments, pregnant mothers were treated with ***vinclozolin***, an anti-androgen pesticide. Here methylation differences (25 different DNA sequences were tested) even in the F_4 generation were observed. Then offspring of both sexes were affected by the maternal treatment, but these effects were transmitted further only patrilinearily. This suggests that transgenerational epigenetic effects can be gender and organ specific, and their consequences may occur later in adulthood. So they can cause some late-onset diseases of adulthood.

Based on the above and other similar experiments, and human data, it is now known that there are two ways of maternal transmission of epigenetic effects. **One is the direct transgenerational epigenesis where (as in fathers) epigenetic changes (epimutations) effect on germ-line cells and the other is the epigenetic reprogramming of development (by maternal intrauterine or perinatal parental care behavior).** Although the nature of epigenetic changes and their transgenerational transmission is not yet known and understood in every detail, but we can say that they play an important role in the fine regulation of gene expression of both normal and abnormal developmental processes.

5.5. Useful web-sites

www.geneimprint.com

<http://atlasgeneticsoncology.org/Deep/GenomImprintID20032.html>

<http://teach.genetics.utah.edu/content/epigenetics/>

<http://epigenie.com/>

5.6. Questions

1. What is the purpose of dosage compensation?
2. What could be an evolutionary explanation for imprinting?
3. What is a differentially methylated cluster?
4. What molecular alterations are in the background of epigenetic changes?
5. Why CpG dinucleotides can be mutation hot spots?
6. What is the role of non-coding RNAs in X inactivation?
7. What mechanisms can cause Angelman syndrome?
8. What is the histone code?
9. What are the CpG islands and what is their epigenetic significance?
10. What is chromatin remodeling?

6. Mendelian Inheritance: autosomal inheritance

Erna Pap

6.1. Introduction

Mendelian inheritance is the basis of classical genetics. Although our knowledge about classical genetics has significantly expanded lately, the understanding of the heredity of the human diseases / traits can still be related to Mendelian inheritance.

Those patterns of inheritance are considered Mendelian in a simplified way, which fulfill two criteria: on the one hand Mendel's principles can be applied, on the other hand the environment has no influence on them. The classical classification of the hereditary patterns is the following: autosomal dominant and recessive, codominant, X-linked dominant and recessive, and Y-linked.

The rigid validity and interpretation of these patterns have been questioned in the last few decades. Numerous phenomena can be observed at the expression of monogenic diseases whose understanding raises difficulties upon the classical rules: either the principle of dominance does not fully appear, or the severeness of the disease is variable, furthermore the environmental factors may provide some influence on the manifestation of the same mutated gene. It has become clear by now, that the „one gene – one locus” genotype determines only partially the clinical symptoms of a disease, and that the combination of several secondary genetic effects with environmental factors together contributes to the manifestation of the disease. These phenomena are among others penetrance, variable expressivity, oligogenic influence and as an epigenetic factor X -inactivation in women. Furthermore, a large number of already clarified and still unclarified epigenetic components can alter the onset, the severeness and the course of a disease.

The genetics of human traits is difficult to study, as it is not possible to make back-crosses, and the time lag between the generations is much longer than in the case of classical models, such as bacteria, yeast cells, drosophila, mice and rats. Additionally, the size of the families is significantly smaller than in classical models.

Up till now more than 6000 monogenic traits/diseases have been revealed. It is to be underlined, that less than 6000 genes lie behind these monogenic diseases. Some diseases are caused by microdeletions or other chromosomal structural variations, but as they can follow

the rules of Mendelian heredity, they are classified as monogenic diseases in human genetics. This also shows that the hereditary pattern is not applicable to genes, rather to characteristics. Human diseases following Mendelian inheritance pattern are much less frequent in a population than those of human polygenic (multifactorial) inheritance (Table 1). It may happen that some physicians will only rarely face a patient suffering from some kind of monogenic diseases during their practice. This topic is still of crucial importance for future doctors, as the genetic properties of mankind, can be „deduced” this way. Several genetic processes, intergenic interactions, even the whole genetics can be relatively more easily studied in monogenic systems. The obtained knowledge can contribute then to the understanding of the development, the clinical course and the curability of the polygenic, multifactorial diseases. Often in these diseases the genes themselves have not even been identified yet. Numerous genes play a role in their development, therefore the simplified connections as „gene – mRNA – protein - symptoms” cannot be set up either.

Listing and discussing all monogenic diseases is not the goal of this present chapter. Besides other study books about the subject “Clinical Genetics” and a huge online reference source is available: the OMIM (online Mendelian inheritance in men <http://omim.org/>). We intend to widen the view: to draw attention to some new aspects besides the general laws, thereby to reveal the complexity of heredity and demonstrate that there are no absolute truths in nature – presently in genetics – and that the mutual effects of genes with other genes and with the environment offer unlimited possible variations in the individuum. All of these may lead us to get closer to the desired personalised medicine, hopefully not too far in the future. At the same time the field shifts from genetics to the world of genomics.

Monogenic		
Huntington's disease	AD	1:10,000
Osteogenesis imperfecta	AD	1:10,000
Familial hypercholesterolemia	AD	1:500
Polycystic kidney	AD	1: 800
Cystic fibrosis	AR	1:3,600
Phenylketonuria	AR	1:12,000
Albinism	AR	1:1000-10,000
Duchenne muscular dystrophy	XR	1: 4,500 (in boys)
Mitochondrial		

Leber optic neuropathy		1:50,000
Chromosomal		
Prader-Willi Syndrome	Deletion	1:30,000
Down Syndrome	Trisomy	1:500 (not hereditary in general)
Multifactorial (Complex)		
Schizophrenia		1: 100
Diabetes mellitus II.*		7: 100
Diabetes mellitus I		4: 1000
Maniac depression		1: 10
Breast carcinoma*		1: 10 (in women)
Allergy		2: 10
Asthma		1: 10
Hypertonia*		3: 10
Obesity*		1: 10

* Significantly more frequent in adulthood or in elders.

Table 1.
The prevalence of some diseases

6.2. Interpretation of some basic genetic terms

Gene: a DNA sequence coding for one or more functional product(s). These can be proteins via mRNA or can be any other kind of RNA, like t tRNA, rRNA, snRNA, miRNA, siRNA, etc. In the so-called RNA-viruses (e.g. influenza, HIV1) genes are coded only in the form of RNA.

Allele: it is an alternative form of a gene (one member of a pair) that is located at a specific position (locus) on a specific chromosome. A diploid cell contains two alleles at a time. Their relationship determines whether the inheritance of a gene-coded trait is dominant or recessive.

Multiple allelism: Certain genes have more than 2 different alleles that result in different phenotypes. It means that there are more than two phenotypes available depending on the mode of inheritance of the alleles. An example for it in humans is the AB0 blood groups for which 3 different alleles in one gene are responsible.

Complex heterozygotes: those persons, whose „disease gene” contains two mutated alleles but both carry a different mutation (e.g. the mutations of *CFTR* gene in Cystic Fibrosis). Although the disease is manifested because both alleles are affected, from the point of view of the presence of different mutations one cannot speak about homozygosity. This formation is

said to be complex heterozygosity as the nucleotide sequence of the two alleles is not identical in respect to the mutations.

Genetic heterogeneity: a single phenotype or genetic disorder may be caused by multiple numbers of alleles or locus variants. Genetic heterogeneity can be classified as either allelic or locus.

- **Locus heterogeneity:** the mutations of several different genes result in the same or in very similar phenotype (e.g. Retinitis Pigmentosa, deafness).

In the case of Retinitis Pigmentosa about 100 genes have been described whose mutations create the same symptoms: loss of vision or blindness due to the death of rods and cones.

In the case of deafness more than 50 genes have been recognized to result each in the same symptom.

- **Allelic heterogeneity:** different mutant alleles of the same gene result in similar symptoms /diseases. (e. g. mutations of *FGFR3* gene. See later!).

Dominance / recessivity: the expression „dominance / recessivity” refers to phenotype, not to genotype. A trait is said to be dominant when it is visible even in heterozygous form. In case it is not, it is said to be recessive. These expressions show the relationship between the two alleles of the same gene. If the expression of one of the alleles represses the expression (the phenotypic form) of the other one, the inheritance pattern is dominant.

What makes a trait dominant or recessive?

Missense mutations and the mutations of regulatory proteins often result in gain of function, as a new, dangerous protein may be coded. Due to the appearance of the new effect these mutations usually cause dominant phenotype, the synthesized proteins function independently from the fact that they should stay inactive. In contrast, nonsense or frameshift mutations cause loss of function, as the protein cannot fulfill its previous function. If a 50% gene activity is still sufficient for the maintenance of the normal vital functions, the inheritance is recessive. If not, **haploinsufficiency** occurs.

Co-dominance: both alleles are manifested phenotypically in heterozygotes. AB0 blood group system is the classical example. Earlier incomplete dominance was also called co-dominance. Incomplete dominance occurs when the phenotype of the heterozygous phenotype is distinct from and often intermediate to the phenotypes of the homozygous phenotypes. E.g. *LDLR* mutations in familial hypercholesterolemia.

6.3. Phenomena that fine-tune classical monogenic inheritance

Most but not all of the following phenomena alter the „one gene – one trait” pure relationship, at the same time practically all of them pose difficulties for the geneticist or for the physician in pedigree analysis.

Expressivity: it determines the strength of the gene, i.e. it shows how strong the manifestation of a given trait/disease is. In autosomal dominantly inherited diseases the severity of the illness can often vary in the affected persons even suffering from the very same variant. This phenomenon is called variable expressivity.

Penetrance: In medical genetics it is the frequency with which a heritable disease is manifested in a given population (family) by individuals carrying the disease-causing allele. Complete penetrance means that the disease is manifested in all members of the population who carry that allele. In autosomal dominant inheritance all heterozygotes are supposed to show the trait, but in the case of incomplete penetrance it does not always appear. In this case a heterozygous parent may not show the trait, but inheriting the disease-causing dominant allele, his/her offspring may. The penetrance of the phenotype (the percent of the manifestation in the carriers) can be calculated in a family

$$P \% = 100 \times \text{number of affected persons} / \text{number of obligate carriers}.$$

An individual is considered to be an obligate carrier who has minimum one affected parent and he(she) himself(herself) has minimum one affected offspring or he himself is affected. For instance if 3 persons manifest the disease out of 4 obligate carriers (heterozygotes) in the same family, the disease shows a 75 % penetrance, meaning that the penetrance is incomplete.

Neither incomplete penetrance nor variable expressivity can be interpreted by Mendel's principles. These two phenomena modulate to the greatest extent the rules of classical Mendelian inheritance, interfering with the pedigree analysis as well. In these cases supposedly more than one gene is expressed, suggesting that interplay between the products of other genes and their mutual effect can influence the pure expression of the „driver” gene. (See later modifier genes!) Besides, a whole arsenal of epigenetic regulatory mechanisms (see Chapter 4.) may alter the expression of a gene, either at transcriptional or at translational level. These fine regulatory mechanisms are orchestrated not only by intracellular, but also by environmental factors. In this way the two most typical characteristics of Mendelian monogenic inheritance – the effect of only one gene and the lack of any environmental effect – seem to be untrue. The effect of additional genes mentioned above shift the monogenic

inheritance pattern toward an oligogenic one, furthermore, the influence of the environment cannot be excluded either.

Anticipation: in each generation the disease gets manifested either earlier and/or in a more severe form. Anticipation could be considered as a specific case of the variable expressivity; still, it is somewhat different. Variable expressivity is explained by the modifying effect of the environment and of other genes; meanwhile different genetic reasons lie in the background of anticipation. It is well known by now, that the earlier manifestation of the disease from generation to generation is caused by a trinucleotide repeat expansion and has probably no connection with effects of the environmental or modifier genes. If the person, carrying the mutant allele, dies earlier than the disease could have been manifested, the case can be considered as incomplete penetrance in respect to the examined family. Obviously, anticipation complicates the exact analysis of a pedigree as well.

Complex or compound heterozygotes: the offspring often inherits two different disease-causing alleles of the same gene from the heterozygous parents in recessive inheritance, which means that the „aa” genotype should rather be described as „ a_1a_2 ”. The different disease-causing alleles can result in differences in the severity of the manifestation of the disease. At some extent, theoretically, the effect of the two alleles could even quench each other.

Pleiotropy: a single gene controlling or influencing multiple (and possibly unrelated) phenotypic traits. Variants in this type of gene will simultaneously affect more than one trait. The explanation is that the gene is expressed in multiple organs; furthermore, fulfilling perhaps completely different functions (see Chapter 12, the example of *EDAR* gene). The protein coded by the gene may be an intermediate molecule of further metabolic cascades or the protein can serve as regulatory molecule, participating in the regulation of several different processes. As for structural proteins, they can be practically present in all tissues.

Heterogeneity: it seems to be the opposite of pleiotropy as variants of multiple genes, independently from each other, results in the same phenotype. In this case two affected parents can have phenotypically healthy, unaffected children if the disease is recessive. Multiple gene abnormalities are seen in disorders such as deafness.

Phenocopy: environmental factors induce the manifestation of a genetic disease despite the presence of healthy alleles (for example deafness caused by Rubella virus infection). The opposite situation is when an inherited disease, due to environmental influences, will be less or will not be manifested at all despite the presence of the mutant alleles. This can occur by

the effect of medicaments or for instance in phenylketonuria by proper diet. The phenocopy is called pathological when the phenotype is pathological despite a healthy genetic background, and it is called normal when the phenotype is healthy despite a mutated genotype.

„de novo”, new mutation: a genetic mutation that individuals neither possessed nor transmitted in a population through many generations, but the disease appears in one of the offsprings unexpectedly. In this case the new spontaneous mutation must have occurred in the germ line of one of the parents of the affected child. Certain genes undergo mutations more frequently, this explains why certain diseases never disappear from a population, as mutations of their genes can be „recreated” in some individuals. For instance Rett syndrome occurs in 95%, achondroplasia in 80% due to new mutation.

Influence of the age: after having had healthy children, originally healthy parents can have an affected child suffering from autosomal dominant disease due to the elderly age of one of the parents. The mutation appears in the germline of the parent as a new mutation. Interestingly, in monogenic diseases, the mutation shows a stronger correlation with the age of the father (~ above 50). The explanation is that gametogenesis is sustained throughout their life in men, spermatogonia undergo multiple divisions, the regulatory mechanisms fail to function properly, and the mutations during replication become stable.

Lethal/sublethal genes: the genetic mutation causes 100% or less than 100 % of the death of the affected individual. Sublethal genes cause the death of only some of the individuals. In special cases the dominant allele can cause death before the affected person would have offsprings. (e.g.. Hutchinson-Guilford – progeria <http://ghr.nlm.nih.gov/condition/hutchinson-gilford-progeria-syndrome>). This could lead to the elimination of the disease-causing allele from a population, but in some genes lethal variants can be formed relatively frequently. In the case of Huntington’s disease the situation is different: the onset of the lethal disease appears relatively late in life, so the affected person will have had children by the time of the manifestation, therefore the lethal disease-causing allele will be transmitted.

„Modifier genes”: genes that influence the expression of another gene. These are interactions between two or more genes of different loci. When an originally monogenic disease shows weaker or stronger symptoms, the reason is that the expression of the mutated gene is often altered by other gene effects. It has been already proven that the manifestation of some diseases is slightly controlled by the mutated forms of specific modifier genes. This offers an explanation to the variable course of the disease in different persons (see above expression, penetrance). **Epistasis** has been considered for long as a separate phenomenon, by today it has become clear though that it is about an interaction between certain main genes and known or

still unknown modifier genes. Until now relatively few modifier genes have been identified, but it is supposed that in most cases not only one gene but a set of modifier genes is involved in the manifestation of the disease. (See Chapter 15 – Systems biology). The picture is further complicated by the fact that the modifier genes themselves follow some kind of hereditary pattern and that they can also be polymorphic, therefore they can differently modify the main gene. The inheritance is called ***oligogenic*** in those hereditary diseases, whose development and manifestation have been proven to be influenced by such modifier genes, like in the case of cystic fibrosis and polycystic kidney. Until recently both have been outlined as classical monogenic diseases. As the methods of full genome sequencing or exome sequencing (see Chapter 11) have become cheaper, in accordance with the systems biology theory it has become possible to demonstrate that every monogenic disease is caused practically not only by the disease-causing variants of the „driver gene” but that parallel the variants of many other genes also contribute to the development of the symptoms.

Heterozygote advantage: the heterozygous genotype has a higher relative fitness to either homozygous genotype. The key may be that a particular allele may have advantages under given conditions, although a different allele may be favored when conditions change. In the case of certain autosomal recessive diseases heterozygotes have reproductive advantage due to environmental factors. This alters significantly the frequency of the disease in these populations. Independently of the environmental factors, modifier genes are supposedly also involved in this phenomenon.

Influence of the sex: the manifestation or severity of certain diseases is different in men and women. (See details in Chapter 7.) In the case of sex influenced traits the autosomal genes are expressed more in one of the sexes (for e.g.. boldness). In congenital adrenal hyperplasia (CAH) altered phenotypes develop in both sexes. In the so-called sex restricted diseases the phenotype is manifested only in one of the sexes, although the inheritance is autosomal. Due to the fact that specific hormones are needed for the expression of the disease, it will be manifested in one gender only. In pubertas praecox for instance the level and effect of sex hormones play the main regulatory role in the development of the disease.

The influence of the environment: some monogenic diseases - despite the disease-causing genotype - are manifested only when particular inducing environmental effects hit the organism. The inducing factors are usually medicament or food. Earlier these diseases used to be called ecogenetic (porfiria, malignus hypertermia, glucose-6-phosphate-dehydrogenase deficiency) (see 6.4.4). Either an altered function of the modifier genes or some epigenetic event lies in the background of the inducing effect.

Table 2 shows a short summary of the occurrence of the above mentioned terms/phenomena and some autosomal diseases.

	Allele heterogeneity	Locus heterogeneity	Variable expressivity	Incomplete penetrance	Pleiotropy	Influence of paternal age	Anticipation	Heterozygote advantage	Phenocopy
Achondroplasia	X					X			
Marfan Syndrome	X		X		X	X			
Osteogenesis imperfecta	X		X	X	X				
Familial hypercholesterolemia	X	X		X	X				X
Polydactyly	X	X	X	X					
Huntington's disease				Depending of the repeat number			X		
Deafness	X	X							X
Cystic fibrosis	X		X		X				X
Phenylketonuria	X							X	X
Albinism (albino phenotype)	X	X			X				
CAH	X				X			X	X
Xeroderma pigmentosum	X				X				
Sickle cell anemia					X			X	

Table 2.

Summary of the genetic characteristics and phenomena in connection with some AD and AR diseases.

6.4. Autosomal dominant inheritance

6.4.1. General characteristics of autosomal dominant (AD) inheritance

The number of known AD diseases is about 4500 out of more than 6000 up till today known monogenic traits/diseases. As we have already described in the introduction, there are less than 6000 genes behind these monogenic traits/diseases: sometimes microdeletions, structural variations cause the disease which follow the Mendelian inheritance pattern, therefore in human genetics they are classified as monogenic diseases. The prevalence of AD diseases is between 1/ 1000 and 1/ 10 000 live births. This number can of course vary, in

familial hypercholesterolemia the prevalence is for instance 1/ 500, in achondroplasia 1/20,000.

The disease causing genes are located on body chromosomes (autosomes). Severity and prevalence is equal in both sexes. The phenotype is manifested in the Aa heterozygous genotype already, which means that one disease-causing allele is sufficient to elicit the symptoms.

In certain cases there is no difference in the severeness of the symptoms between homozygotes and heterozygotes. For instance in Huntington's disease one mutated allele of huntingtin gene already codes for such a deeply malfunctioning protein, - as a negative dominant mutation- that it elicits the lethal outcome. (Not to be mistaken: the onset and the slower or faster development of the disease does not depend on the homozygous or heterozygous genotype, but on the trinucleotide repeat number in the mutated allele. See 6.3.2.: Anticipation.) On the other hand the homozygous state can result significantly more severe outcome of certain diseases than the heterozygous one. (Examples: osteogenesis imperfecta, Marfan syndrome, Waardenburg syndrome or familial hypercholesterolemia.) Homozygosity is quite rare in AD diseases, since both parents need to be affected. (When new mutations appear, they practically never occur at the same time on both alleles of an individual.) The disease is sometimes accompanied by so severe symptoms that the disease itself is the obstacle of starting a family. As the disease can be recognized due to the dominant phenotype, sometimes affected parents themselves refuse to have children, they do not risk it, or finally they decide to visit a genetic counselling. As it has been already discussed in Chapter 6.3 new mutations can arise, therefore the „disease gene” cannot be eliminated from a population.

Possible causes of dominant inheritance can be haploinsufficiency, dominant negative effect and loss of function or gain of function mutations.

The AD inherited diseases are caused mostly by mutations in genes coding for structural proteins, regulatory proteins, receptors and proto-oncogenes. Pleiotropy, variable expressivity, incomplete penetrance, arousal of a new mutation and the influence of paternal age are particular characteristics of this kind of inheritance. These will be discussed in the following paragraphs in connection with a few diseases as examples.

6.4.2. Diseases due to the mutation of structural genes

6.4.2.1. Marfan syndrome

Fibrillin gene is mutated. Pleiotropy in this case is clearly understandable, since fibrillin is one of the most important extracellular proteins present in the elastic and non-elastic connective tissues. Multiple organs can be affected: lungs, skin, kidneys, skeletal system, vascular system, cornea, etc. The type of the affected organs and the severity of the symptoms may show individual differences, implicating variable expressivity. The arousal of the mutation shows strong correlation with paternal age.

6.4.2.2. Osteogenesis imperfecta

The different mutations of collagen genes produce differently severe symptoms. The collagen gene family comprises 45 genes, distributed on different chromosomes, coding for proteins of somewhat different nature. Considering that collagen is the most abundant protein in the extracellular matrix, it is not surprising that the mutations of these genes have pleiotropic effect as well. Also expressivity is variable; the fragility of the bones in homozygotes can be so severe, that it causes death already during the process of birth. On the other hand it is possible that either deafness - as a result of the abnormal development of the ear bones, or only breakable bones develop.

Sometimes the inheritance of osteogenesis imperfecta shows incomplete penetrance. It has not been revealed exactly yet, why some heterozygous persons do not show the symptoms in AD diseases, but it is suggested that modifier genes are involved in the process.

6.4.3. Diseases due to mutations of receptor genes

6.4.3.1. Achondroplasia

The disease is caused by the mutation of *FGFR3* (fibroblast growth factor receptor type 3) gene. Depending on the locus of the mutation in the gene, three different diseases are distinguished: - achondroplasia, hypochondroplasia and tanatophoroplasia - .

The mutation is a „gain of function” mutation; the receptor stays active without ligand as well. The mutation arises usually as new mutation and it shows strong correlation with the age of the father.

6.4.3.2. Familial hypercholesterolemia: it is a relatively frequent disease with a prevalence of 1:500. The severity of the symptoms depends on whether the genotype is homozygous or heterozygous. More than 100 mutations have been identified in the gene of LDL receptor- this is a case of allelic heterogeneity. But variants in other genes (e.g. *APOB*, *PCSK9*) can cause similar symptoms, i.e. the same disease, which is the case of locus heterogeneity. In addition the disease also has recessive types, like mutations in the *ARH* or *CYP7A1* genes.

6.4.3.3. Polycystic kidney disease: with a prevalence of 1:800, it is a quite frequent AD disease. The disease is connected to the malfunction of a receptor-ion channel complex that consists of two polycystin proteins, regulating G-protein signalization and a membrane bound Ca^{++} channel. *PKD1* gene codes for the protein polycystin1 and *PKD2* gene for polycystin2. 80% of the mutations hit the *PKD1* gene. The onset of the disease is in early or late adulthood, it shows variability even in the same family regarding both the time of the onset and the severity of the symptoms. The differences depend on environmental factors, as infections can promote the development of the symptoms. As for genetic influence, the manifestation may depend on the site of the mutation in the affected allele. Additionally, the outcome of the disease is strongly influenced by modifier genes whose variants further complicates the picture. Some variants of nitrogen-monoxide synthase gene (*NOS3*) induce an early and more severe development of the disease.

6.4.4. *Mutations of the gene of a protein with a yet unknown function*

6.4.4.1. Huntington's disease: it is caused by a trinucleotide repeat expansion. The exact role of the huntingtin protein, coded by the healthy allele is unknown. Within cells, this protein may be involved in chemical signaling, transporting materials, attaching (binding) to proteins and other structures, and protecting the cell from self-destruction (apoptosis). A gain of malfunction appears with the increase of the number of CAG repeats. Normally, the CAG segment is repeated 10 to 35 times within the gene. People with Huntington's disease have 36 to more than 120 CAG repeats. People with 36 to 39 CAG repeats may or may not develop the signs and symptoms of Huntington disease, while people with 40 or more repeats almost always develop the disorder. The increase of the repeat number occurs mostly in the paternal germline (see 6.3 and 6.4.1).

6.4.5. Mutation of Protooncogenes

See 6.6.

6.4.6. Pharmacogenetic diseases

There are some monogenic diseases which - despite their mutated genotype - get manifested only in the presence of particular inducing environmental effects. An older classification ranged pharmacogenetic diseases into the group of ecogenetic diseases. The term „pharmacogenetic” is not exact either, as the inducing factors can also be others than only medicaments.

6.4.6.1. Porfiria

The affected genes code for the enzymes of the steps of haeme and porphyrin synthesis. Different types of porfiriias are distinguished, depending on the enzyme whose gene had been mutated in the cascade. The inheritance pattern is mostly autosomal dominant, rarely recessive. Although the disease-causing alleles are present in the patients, they do not lead to phenotypic appearance unless certain specific environmental factors induce the manifestation of the disease. These factors are from different origins: drugs, alcohol, steroids (for example contraceptives) stress, starvation, light, etc. The classification of porfiriias in internal medicine and their biochemical background will not be discussed in this chapter. Instead, once more, we intend to draw attention to the fact that genes *per se* are not „omnipotent” and although the monogenic background is clarified, the inheritance pattern of this disease differs significantly from the classical Mendelian schema.

6.4.6.2. Malignant hyperthermia

The disease can be caused by the mutation of at least six different genes. Mutations of *CACNA1S* and *RYR1* (ryanodine receptor) genes are the most frequent. The product of *CACNA1S* gene regulates the function of ryanodine receptor. As ryanodine receptor regulates the function of Ca⁺⁺ ion channels, the mutation of either *CACNA1* or *RYR1* genes results in the efflux of large amounts of Ca⁺⁺ ions from the sarcoplasmic reticulum to the cytosol due to the faster opening and slower closing of the ion channels. The increased Ca⁺⁺ ion concentration causes increased muscle contraction and increased heat production, resulting in unquenchable high fever and even death. This is indeed a pharmacogenetic disease as it is

exclusively triggered by drugs, namely by those that are commonly used as general anesthetics.

6.5. Autosomal recessive inheritance

6.5.1. General characteristics of autosomal recessive (AR) inheritance

The disease / trait is manifested in homozygotes only (aa). The prevalence and severeness of the disease is the same in both sexes. The significant spread and presence of some diseases in some populations may be explained evolutionary by founder effect and heterozygous advantage. Complex heterozygous state is characteristic for multiple recessive diseases. The type of the pedigree is horizontal, sometimes sporadic. Consanguineous marriages significantly promote the increase of affected persons. The frequency of the disease will be higher in these families than in the surrounding population. Consanguinity makes „hidden genes” visible, as heterozygotes are more „concentrated” in these families than in the rest of the population.

The AR inherited diseases are caused mostly by mutation of an enzyme, haemoglobin or tumorsuppressor genes. Cystic fibrosis also belongs here; it does not fit in any of the above mentioned groups though. The frequency of AR diseases is in an average of 2.5:10,000 live births in Europe, although this number can vary. The prevalence of cystic fibrosis is 1:1600, of phenylketonuria is 1-10,000, of mucopolysaccharidosis is 1:50,000.

6.5.2. Enzymopathies

The disease-causing alleles cause decreased enzymatic activity already in heterozygotes. Its value can fall exactly between the enzymatic activity values of homozygous affected and homozygous healthy persons. It is also possible that the level of the defeated metabolic product significantly increases in the organism; still, there are no signs of the manifestation of the disease.

The phenomenon of pleiotropy can be easily interpreted in the case of enzymopathies, as the affected enzymes often catalyze steps of cascade reactions. If the enzyme is missing from the beginning of the cascade or from a branching point, more than one metabolic process will be probably damaged.

6.5.2.1. **Phenylketonuria** (PKU) is caused by the lack of phenylalanine hydroxylase enzyme, resulting in toxic phenylpyruvate production instead of tyrosine. The enzyme deficiency

causes problems in the cascade of tyrosine conversion. Although tyrosine is supplemented in the organism with diet, it remains below the normal needs of tyrosine, therefore DOPA and melanin synthesis suffers damages. PKU can be treated with phenylalanine free or phenylalanine poor diet. While the manifestation of the toxic product can be avoided, light skin color, blue eyes, light hair color remain as characteristic phenotype in the affected persons. It has been recently discovered that the expression of the several hundred types of phenylalanine hydroxylase mutations is influenced by modifier genes.

6.5.2.2. The so-called **classical albinism** is caused by mutation of tyrosine kinase gene, therefore melanin synthesis fails. This enzyme is a component of a cascade reaction series as well, which explains the pleiotropic effects in albinism. It is to be underlined that the mutations of multiple other genes also result in a similar albino phenotype. The causes of these diseases are deficiencies in intracellular melanin transport. Mutations of several different genes have already been identified, also the diseases have been named differently. (See Table 3.)

6.5.2.3. Pleiotropic effects in **congenital adrenal hyperplasia** (CAH) can be also explained by failures in the enzymatic cascade of steroid synthesis. 21-hydroxylase deficiency results from a unique mutation with two highly homologous near-copies in series consisting of an active gene (*CYP21A*) and an inactive pseudogene (*CYP21P*). Mutant alleles result from recombination between the active and pseudo genes.
<http://pediatrics.aappublications.org/content/106/6/1511.full> The frequency of CAH among Yupi eskimos is 1:300, while in other populations it is about 1:10-15,000. This high prevalence can be explained by heterozygote advantage developed against Haemophilus influenzae B.

6.5.2.4. In **Xeroderma pigmentosum** one (or more) of the genes of the nucleotide excision repair (NER) of DNA is mutated, this is an example for locus heterogeneity. (See 6.2.).
<http://ghr.nlm.nih.gov/condition/xeroderma-pigmentosum>

Mutated gene	Affected protein	Damaged function	Result
Tyrosinase 11q14-3	Tyrosinase enzyme	DOPA and melanin synthesis <i>(Classical and oculocutan albinism)</i>	
P gene 15q12	Transmembrane protein of melanosome membrane	unknown	CLASSICAL ALBINISM
TyRP1 gene 9p23	Tyrosinase-related protein1	Iontransporter or chaperone or melanosome proteincomplex stabiliser	
LYST gene 1q42-43	Unknown protein	Suppository role in the transfer of molecules from Golgi to target place	SYNDROMES WITH ALBINO PHENOTYPE
HPS gene 10q23.1-23.3	Complex proteins	Molecule „trafficking” <i>(Hermansky-Pudlak syndrome)</i>	
Rab27a gene 15q21	GTP-binding protein Rab 27a	Melanosom transport on microtubuli <i>(Grisicelli syndrome)</i>	
MYO5a gene 15q21	Myosin 5a	Melanosomes can not leave melanocytes <i>(Grisicelli syndrome)</i>	

Table 3.

Classical albinism and syndromes with albino phenotype. Their genetic background and the altered / lost cell biological functions.

6.5.3. Cystic fibrosis

The gene of the disease, the *CFTR* gene does not belong to any of the gene groups showing „classical” AR inheritance, as the disease is not caused by enzymatic or haemoglobin mutations. The mutation hits the gene of a chloride ion channel regulatory protein. It is the most frequent AR disease in the European populations. There is no evidence of increased mutation rate in this disease; instead heterozygotes might have had increased fitness against cholera over both homozygous genotypes in the middle age. The phenomenon is the so-called heterozygote advantage. (See 6.3 and Chapter 12.)

Logically, the phenomenon of pleiotropy in the case of Cystic fibrosis is not a consequence of metabolic cascade discrepancy. It is the consequence of the increased dense mucus secretion in several organs, whose ducts become plugged this way. As pathogens can easily invade the mucus, chronic inflammation develops and the pancreas, the lungs, the seminiferous tubules

get into severe conditions. The disease cannot be cured presently, but its severity can significantly differ among the patients, suggesting the effect of modifier genes and demonstrating that specific mutations in the same gene can lead to differently damaged protein functions. Out of the more than 800 different mutations that have been identified in the *CFTR* gene so far, deltaF508 is the most frequent one. This is the deletion of a triplet from exon 10, coding for the 508th aminoacid. Due to the large number of mutations, complex heterozygosity can be assigned to the genotype of the disease (see 6.3).

6.5.4. *Haemoglobinopathies*

6.5.4.1. The cause of **sickle cell anemia** is one of the most well-known mutations. The 6th aminoacid glutamine is substituted onto valine in the betaglobin chain of the haemoglobin molecule, due to a transversion substitution in the gene. The mutated haemoglobin is called haemoglobin S. It causes the sickle shape of the red blood cells and gives rise to multiple additional pleiotropic effects. The development of the heterozygote advantage against malaria can also be attributed to this mutation. (See Chapter 12.)

6.5.4.2. **Thalassemia** can be caused by several types of mutation. Deletions, frameshift and splicing mutations can arise in both alpha and beta chains of the haemoglobin. This explains at the same time the differences in the severity and the geographical spreading of the types of thalassemias. Heterozygotes are partially protected against malaria in the case of this disease as well.

6.6. Genes and Tumors

The field of oncogenetics will be discussed in more details in Chapter 8. Still, it is important to mention this field by monogenic inheritance, as many of the tumor causing genes have already been identified and the relation „one gene – one damaged protein product” can be traced. In spite of this fact tumors are considered to belong to the group of multifactorial hereditary diseases: 1. their development is dramatically influenced by environmental factors; 2. the occurrence or loss of the products of multiple mutated genes together result in the manifestation of the disease. This so-called „**multiple hit theory**” can be attributed to the fact that many different proteins can take over the role of a damaged function for relatively

long time in the life of the affected cell, so the mutation of one gene alone will not induce the manifestation of the tumor.

More or less all cells carry some kind of genetic failure. One or two such mutations will not transform the cell into a tumor cell. The genetic damage – mutation - produced in the original cell will be preserved and the offspring cells will inherit it. Proto-oncogenes are a group of genes that cause normal cells to become cancerous when they are mutated. Mutations in proto-oncogenes are typically dominant in nature, and the mutated version of a proto-oncogene is called an oncogene. The inheritance of tumor suppressor genes is recessive. The transformation results in the altered activity of the protein. The products of the mutated genes are involved in signalling pathways regulating cell growth, division and differentiation.

If the mutation occurs in the germ line, the mutated allele may be transmitted into the offsprings. Genes endowing the predisposition for a given tumor have already been identified in a large number of cancer types. It is to be underlined that although tumor suppressor genes exert their negative effect in homozygous recessive form, their inheritance pattern is said to be dominant because of their manifestation, that is, the disease appears often in every generation in a family. (As described above, vertical family tree is a characteristic of the dominant inheritance pattern.) *RB1*, *BRCA1- 2*, *APC* genes, etc., could be picked out at random. When in the carrier person the mutant allele is located in cells that proliferate intensively in certain periods of life, a second, new mutation can relatively easily arise on the originally healthy allele, as well („two hit theory”). Therefore albeit the mutation of both alleles of the gene is required for the malfunction of the coded protein, because of the increased probability of the occurrence of the second hit, already one parent is sufficient to dispose a higher risk for the development of the disease in the offspring. Additionally, it is not to be forgotten, that more than one (usually several) recessive homozygous or heterozygous dominant allele pairs must be present in the cell in order to elicit the development of the malignus tumor itself.
<http://themedicalbiochemistrypage.org/oncogene.php>

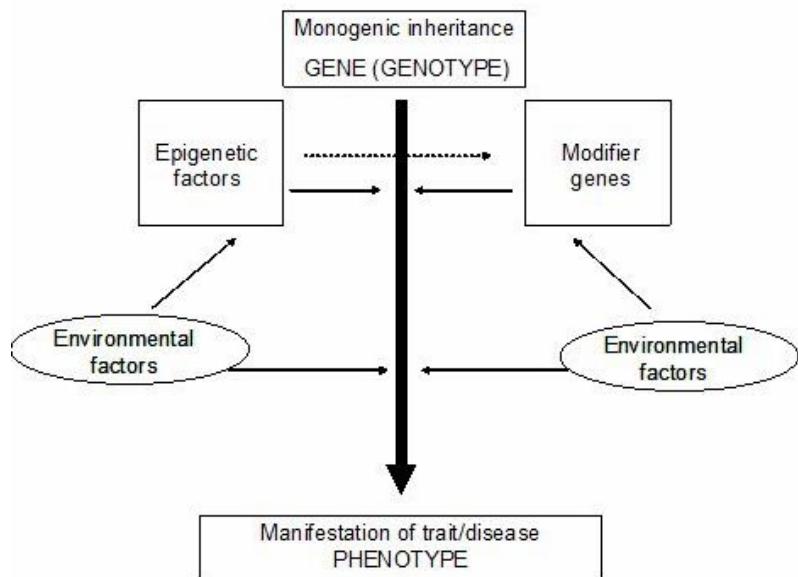


Figure 6.1.

Factors influencing monogenic inheritance

6.7. Genes and Drugs

It is important to keep always in mind in medical practice that every single person may react differently to the very same drug. This can mainly be attributed to the genetic variations in the individuals. The topic will be discussed in more details in the chapter of Pharmacogenomics (Chapter 13.) It is to be mentioned here as well, that there are mutations inherited in a monogenic AD, AR or X-linked way, which are responsible for the adverse reactions caused by certain drugs. These diseases are very rare and they are not to be confused with the pharmacogenetic diseases discussed in 6.4.6. (See http://ebooks.thieme.com/pdfreader/color-atlas-genetics41792_p.273.) It is not to be confused either with the fact that people can react very differently even to the most common medicament, which can be explained by the large number of genomic variants, as well.

6.8. Conclusion

A new interpretation of the classical monogenic inheritance and of the application of Mendelian principles has arisen in the last few decades. It seems undoubted by today that environmental factors, epigenetic effects and the products of the so-called modifier genes all influence the phenotypic manifestation of the allele pair (gene) which is responsible for a given trait/disease. Even monogenic inheritance that was believed relatively simple earlier,

seems to be more complicated and complex. (See Systems biology, Chapter 15.) Although the complexity was assigned to the polygenic, multifactorial inheritance, our view about monogenic diseases widens due to the new discoveries of genetics / genomics (see Fig. 6.1).

6.9. Questions

1. Describe Mendel's principles!
2. Define the following terms! --- gene, allele, multiple allelism, complex or compound heterozygotes, locus heterogeneity, allele heterogeneity, dominance, recessivity, codominance.
3. Which phenomena interfere with the classical application of Mendel's principles in the case of monogenic diseases?
4. Define the following terms! Give examples of the diseases for each term!---- pleiotropy, expressivity, penetrance, anticipation, phenocopy, complex or compound heterozygotes, heterogeneity, sublethal/lethal gene, new mutation, modifier gene.
5. Describe the meaning and give examples: the age and the sex influence the manifestation of some diseases.
6. Which are the classical monogenic inheritance patterns?
7. How has the discovery of oligogenic inheritance pattern affected our view of the monogenic inheritance? Give examples!
8. What types of genes are usually mutated in the case of AD and AR diseases? Give examples for each type!
9. Does the environment influence the manifestation of diseases following monogenic inheritance patterns?
10. Describe the inheritance and the manifestation of tumors and pharmacogenetic diseases with respect to the environmental effects!

Recommended readings

<http://ebooks.thieme.com>

7. The role of sex in heredity

Sára Tóth

Our gender on the one hand acts directly through the sex chromosomes and the genes encoded by them, on the other hand through the characteristics of gametogenesis and fertilization it influences indirectly the appearance of our characteristics.

7.1. X-linked inheritance

The fact that in humans **females are homogametic and males are heterogametic** makes the interpretation of both types of X-linked inheritance patterns difficult. Because women have two X chromosomes, they may be homo- or heterozygous for an X-linked trait / disease. Those women whose heterozygosity is proven by pedigree analysis and / or genotyping, are called ***obligate heterozygotes*** (conductors), while those who are only presumptive heterozygotes based on the family tree (i.e. they have no affected offspring, but have affected brothers) are called ***facultative heterozygotes***.

In contrast, men have only one X chromosome therefore they are ***hemizygous***, so they are either affected or carriers when the X chromosome is mutated, or healthy, if a normal (non-mutant) X chromosome is present. **As for X-linked disorders 1/3 of the affected males are new mutation carriers!** After all, the hemizygous males have reproductive disadvantage since the trait or disease is always manifested in them so the mutant gene is selected out from the population. In women, X chromosome inactivation further complicates the picture: depending on the X inactivation the phenotype can be quite varied - mild or severe - in heterozygotes.

7.1.1. ***X-linked dominant (XD) Inheritance***

In this case, the pedigree pattern is similar to the autosomal dominant, but the two sexes are affected differently.

The main features of this type of inheritance are:

- 1 / vertical family tree
- 2 / twice as many women affected as men, 2 : 1 female : male ratio

3 / 50% of the offspring of an affected women - regardless of their sex – are sick

4 / **all daughters of an affected man are affected while all sons are healthy** (the father always gives his X chromosome to his daughters, the Y to his sons!)

5 / symptoms of the affected women are often milder and more variable than that of the affected men

While the symptoms of homozygous dominant X^AX^A females are alleviated only by the X inactivation, whereas in heterozygous X^AX^a women the product (protein) coded by the normal allele X^a can do the same it as well.

Traits / diseases determined by genes on ***the X chromosomal PAR1 region*** e.g. the **Xg blood group antigen and amelogenesis imperfecta** (incomplete teeth enamel production) have such inheritance. In the latter one the enamel layer of the teeth is missing and such teeth grow carious more easily.

The most known X-linked dominant disorder is the **hypophosphataemia** (formerly called vitamin D-resistant rickets, coded on the long arm of the X chromosome), which is characterized by growth retardation in childhood, rickets and low serum phosphate level. It is a treatable disease by large doses of vitamin D and phosphate! The **fragile X syndrome**, a trinucleotide (CGG) repeat mutation caused disease is also X-linked dominant. This is the most common cause of male mental retardations. While the normal repeat number is <30, this number is between 200 and 2000 in the affected individuals. Between about 50 and 200 repeats there is the so-called **premutation or gray zone**. The adult affected males are characterized with a long face, protruding ears, large jaws and large testes. In addition to mental retardation, behavioral problems and mood swings are part of the symptoms. The protein encoded by the **FMR1 gene** probably causes the symptoms by binding the mRNAs of other genes involved in the functions of the nervous system.

The assessment of the X-linked dominant pedigrees is complicated by the so-called **X-linked male lethality**. Since there is no normal allele the hemizygous, male embryos already die in utero. In this case, there are usually not as many offspring in the family to realize the 2:1 female : male sex ratio characteristic of such inheritance. **Incontinentia pigmenti** associated with hemizygous lethality is a disorder of pigmentation characterized by blistering of the skin in early childhood and with partial hair loss that manifests only in women. **Rett syndrome**, which is basically a neural developmental disorder, is also characterized by male lethality but moreover epigenetic phenomena are involved as well. In girls the typical progressive symptoms of loss of speech and acquired motor functions, the compulsive hand-

wringing, ataxia and seizures are due to the mutation of the methyl-cytosine binding protein coding ***MECP2 gene***.

7.1.2. X-linked recessive (XR) Inheritance

To date, more than 400 traits with such inheritance pattern are identified. This value is much greater than that would be estimated on the basis of the number of human genes per chromosome and this fact is due to the easier detection and identification of such traits because of the specific male inheritance pattern derived from heterogametic sex. Amongst such traits / diseases there are relatively harmless, with mild symptoms such as **red-green color blindness**, others with severe symptoms such as **hemophilia**, and lethal as **Duchenne muscular dystrophy**.

The characteristics of X-linked recessive inheritance are:

- 1 / zigzag or knight's move pattern: the disease is transferred from mother to son
and from son to his daughter
- 2 / there are many more men affected than women
- 3 / sick women are born to affected father and obligate heterozygote mother
- 4 / affected man usually comes from healthy parents where the mother is
obligate carrier
- 5 / there is no man-to-man transmission

Although **hemophilia** is known for at least 4,000 years - as already mentioned in the Talmud that in families where one of the sons of the matrilineal relatives died due to bleeding out at circumcision as a result, their newborn sons were not circumcised - the first point mutation was described only in 1986. The X-linked recessive hemophilia has two forms: **Hemophilia A**, which is due to the failure of blood clotting **factor VIII**, and **hemophilia B**, which is due to the failure of blood clotting **factor IX**.

In 40% of hemophilia A cases a specific mutation of the factor VIII gene occurs. The intron 22 of the gene contains two small genes of unknown function, the F8A and F8B. About 400 kb away there are more copies of F8A of as well. **Among these copies intrachromosomal crossing over takes place during meiosis, causing the inversion of the corresponding chromosome piece and thus factor VIII gene falls apart in two distant pieces.**

This is the cause of the lack of clotting factors and hemophilia. The most common mutation causing hemophilia occurs in the **paternal germ line** during meiosis. The large

number of divisions and the concomitant increased spontaneous mutation rate typical to paternal gametogenesis explain among other things that mutations occur with higher probability in the offspring of aged fathers.

One of the best known and most studied **cytoskeletal diseases** is **Duchenne muscular dystrophy**. This X-linked recessive disease, which was described in the second half of the last century, begins with difficulties of standing up in the 2nd-3rd years of life - **Gower's sign** - and associated with increasing muscle weakness.

The boys around the age of 10 are wheelchair-bound then die around 20 years of age. Because the disease is relatively common (incidence of 1:3500), and to this day is incurable, it is clear that it is intensively investigated. Thus came to light that the cause of the disease is a gene mutation affecting a cytoskeletal protein called **dystrophin**. The dystrophin, a muscle cell specific protein whose C-terminal end is bound to the sarcolemma through a glycoprotein complex of six components and the N-terminus linked to the actin cytoskeleton. **The dystrophin is the product of the largest currently known gene, which is 2400 kb in length, and thus its transcription takes more than 16 hours.** The function of dystrophin in muscle is the cell membrane stabilization. The mutation is often a frame-shift causing deletion, and thus the cell does not produce dystrophin, or a protein with completely altered structure and function is synthesized. If only an in-frame mutation occurs in the dystrophin gene, that is only a small part is deleted, then the so-called **Becker muscular dystrophy** with milder symptoms is formed. **The Duchenne and Becker muscular dystrophies are due to different mutant alleles of the same gene, so they are examples of allelic heterogeneity as well. As many other mutations (for example, point mutations, and duplications) occur in the dystrophin gene, multiplex allelism is also typical for it.**

Since the affected men generally do not reach reproductive age, they can not transmit their mutant gene to the offspring, so this **sub-lethal mutant gene** should gradually disappear from the population. However, the incidence of the disease is fairly constant; it is just possible as the rate of new mutations is high, that the mutant gene is repeatedly produced. **According to new observations deletion mutations involving the dystrophin gene take place typically in the maternal germ line while the other types of mutations are rather common in the paternal germ line, but the reason has not been known yet.**

The X chromosome inactivation further complicates the pedigree analysis also in X-linked recessive inheritance. The phenotype of heterozygous females varies depending on the ratio of healthy X^A and mutant X^a bearing cells. **If the gene product is a soluble protein, such as the clotting factors in hemophilia, the effect is “averaged”. In other words, these women**

are asymptomatic but biochemically will be different from normal. However, where the product is localized to a given cell type, there the symptoms appear in a mosaic form. Such as the **hypohidrotic ectodermal dysplasia**, where the mutation causes the absence of sweat glands and the abnormal development or deficiency of dentition.

7.2. Y-linked (holandric) Inheritance

Currently, only the male sex determining and the male gametogenesis related genes: the ***SRY*** and ***AZY*** Y-linked inheritance are proven. **There are not known Y-specific and not male infertility related hereditary diseases!**

The only somatic characteristics transmitted from man to man, the hairy ear is not Y-linked, but the exact gene and its locus is not yet known.

The Y-linked inheritance features are:

- 1 / only males are affected
- 2 / the affected men's father is affected
- 3 / all the sons of affected men are affected

7.3. Sex influenced inheritance

In the case of some traits the gene is differently expressed in the two genders. It could be either the consequence of male lethality mentioned earlier in connection with X-linked inheritance or it could be due to the influence of other genes which means that the gene is expressed differently or not expressed at all therefore the manifestation of the gene depends on the sex of the affected individual.

The best known such trait is **baldness**, which is autosomal dominant in men, so it is expressed both in homo- and heterozygotes. On the contrary **it is autosomal recessive in women where it is expressed only in homozygotes with high testosterone level**. One type of **pubertas praecox (precocious puberty)** having the same genetic background is manifested mainly in men, too. In this disease ***the luteinizing hormone receptor (LHR) is mutated***, so it induces increased testosterone synthesis and those somatic changes characteristic for the premature puberty even in the absence of the ligand.

7.4. Sex limited inheritance

In the case of traits **when the gene is strictly expressed only in one sex, the inheritance is sex limited.** Milk production is a classic example of it, since there are mammary glands in both sexes, but milk secretion is characteristic only for females. The fact is long been recognized by cattle breeders that the milk yield depends on the bull, too! In fact, not only the amount of secreted milk, but the composition (e.g. its lipid content) is also dependent on the paternal genes. The development of **pre-eclampsia**, which takes more than 50 000 women's lives a year, can be affected by paternal genes. In this case, the paternal genome half of the developing embryo affects the development of the placenta in a way, that it may cause a sudden increase of maternal blood pressure towards the end of the pregnancy.

7.5. Genomic imprinting

The parental origin dependent gene expression is the genomic imprinting, which because of the characteristics of the process (DNA methylation, histone modifications, chromatin remodeling) is discussed in Chapter 5 (Epigenetics).

7.6. Cytoplasmic inheritance

7.6.1. *Maternal genetic effect*

The role of sex is also proven for other specific forms of heredity. For example, in cytoplasmic inheritance with maternal genetic effect molecules (mRNAs, non-coding RNAs, or proteins) stored in the oocyte modify the development of the offspring by influencing the gene expression after fertilization. Thus, **the expression of genes in the offspring can be different without mutation had occurred in the genome, so this kind of inheritance is a consequence of epigenetic changes.** For Drosophila and other lower ranked eukaryotes there are several evidences e.g. the formation of dextral or sinistral shells. Then during oocyte maturation factors coded by the dominant or the recessive maternal alleles, and produced by the nurse cells, are passed to the egg and to the zygote and subsequently, i.e. during the early cleavage divisions can modify the orientation of the axis of mitotic spindle leading to dextral or sinistral shell formation. Likewise, there are transgenic mouse model examples for the manifestation of traits not coded in the genome of the offspring but induced by paternal sperm RNAs. However, the role of this type of cytoplasmic inheritance in humans has yet to be verified.

7.6.2. Mitochondrial inheritance

The role sex is undisputed in the inheritance of mitochondria since in this case cell organelles found in the oocyte cytoplasm are exclusively transmitted by the mother to the offspring. Several theories exist which explain how this happens. According to one of them during fertilization of the sperm neck – the midpiece - cannot get to the egg, so the zygote, and later the developing organism contain only maternal mitochondria. According to another theory, the sperm's mitochondria enter the egg, but do not contribute genetic information to the embryo. Instead, they are eliminated somehow. It is suggested that paternal mitochondria are marked with ubiquitin to select them for later destruction inside the embryo. **This means that the mother transmits her mitochondria to all of her offspring - sons and daughters alike -, but in the next generation her daughters, the sons cannot pass them further. This maternal inheritance does not follow the Mendelian rules, so it can be considered as one type of non-Mendelian inheritance as well.**

Many diseases are due to mutations in mitochondrial DNA. There are 59 known mitochondrial disorders, but fortunately all of them are rare. Since the main function of the mitochondria is the oxidative phosphorylation, therefore mitochondrial diseases mainly affect organs with the highest energy need (muscles, nervous system). One of the most well-known mitochondrial diseases is **Leber's optic neuropathy** due to a point mutation, and which is usually characterized by adolescent or young adult-onset central vision loss.

It must be mentioned here that there are mitochondrial diseases caused by nuclear DNA mutations, since the vast majority of mitochondrial proteins are coded in the nucleus! They show Mendelian inheritance, such as the autosomal dominant progressive external ophtalmoplegia (external eye muscle paralysis) coded by the long arm of chromosome 10.

In the case of mitochondrial diseases two types of mutations - point mutations and deletions - are known to occur. The severity of the symptoms depends on the type of mutation, the number of mutant mitochondria and naturally the tissue type. The majority of mitochondrial mutations do not take place in the germ line, they are generally somatic but, in addition, the amount of the mutant mitochondria may vary from cell to cell during successive cell divisions even within a tissue. Therefore the cytoplasm of the cells will be different. When the cell cytoplasm contains the same normal or the same mutant mitochondria **homoplasmy**, when two types of mutant or both normal and mutant mitochondria are found simultaneously **heteroplasmy** is present. Understandably the severity of symptoms can be

variable. In the offspring of heteroplasmic mothers the severity of symptoms may be different depending on how many mutated mitochondria were passed to the egg.

The analysis of mitochondrial DNA and homo-and heteroplasmy was used recently for the identification of the remains of the Russian royal family executed in 1918. Since mitochondria are maternally inherited, so the living matrilineal relatives of the imperial family and their descendants were tested and compared to the mitochondrial DNA extracted from the remains. Not only the identification of the remains was successful, but it has also been shown that Tsar Nicholas II was heteroplasmic.

Also, mitochondrial DNA testing can help to decide an old debate in human evolution: does the ancestors of modern man (*Homo sapiens*) and Neanderthal populations interbreed (see chapter 9)? Comparative analysis of mitochondrial mutations is often used in population genetics for the detection of ethnic origin, ancestry and relationships of certain populations. It turned out that American Indians have mutant, deleted mitochondria, so if someone has such deletion, he or she must have at least one Indian female ancestor. Similarly, mitochondrial DNA analysis has shown that the Hungarians and Finns are genetically not, only through their language related!

7.7. The X chromosome inactivation

In somatic cells, the paternal and the maternal copy of all autosomal genes are expressed. So these are present in a double dose. The only exceptions are the so called imprinted genes (see Chapter 5). However, the expression of genes encoded by the X sex chromosome is influenced whether the X chromosome is of the male or the female, and the fact that the X and Y chromosomes are not homologous except the genes of PAR. Thus the X chromosomal genes in females can be transcribed in twice as large doses than in men. This is prevented by a phenomenon called **dose compensation**. Due to the X chromosome inactivation, described in chapter 5, there is **functional mosaicism** in women. The best known examples are the calico (or tortoiseshell) cats. Only the female cats are black and red mottled. The size and distribution of patches depends on where and in how many cells the black or red color coding X chromosomes are inactive. While somatic cells are characterized by random X inactivation, the extra-embryonic membranes (placenta) have imprinted parental origin dependent X chromosome inactivation. The placenta always has the paternal X in inactive form.

The inactive X chromosome can be detected in interphase. Adhering to the nuclear membrane, a heavily stained **sex chromatin**, the so-called **Barr body** is seen in the epithelial cell nuclei A

drumstick-shaped appendix of the segmented nucleus of neutrophils is a particular manifestation of the inactive X, so the Barr body. The rapid detection and microscopic examination of Barr bodies are simple, in the past it was used for quick sex determination in connection with sports competitions.

Initially, it was thought that the whole X chromosome is inactive, but we now know that ***the PAR regions are never inactivated!*** Moreover, non-inactivated X chromosomal genes outside the PAR were also found, a part of them has a functional therefore transcribed homologue in the Y chromosome, while the other part has only non-functional pseudogene on Y (such as the steroid sulfatase (*STS*) gene and the anosmin gene responsible for Kallman syndrome).

In other species, where heteromorphic sex chromosomes also occur other mechanisms exist for dose compensation. The X chromosome in male *Drosophila* is twice as active, than in females. A 1:1 ratio instead of 2:2 is formed this way. It is also possible that both female X-s are only half as active as the male one, thus $1/2+1/2 : 1 = 1:1$ is the final ratio.

An interesting possibility of the X chromosome inactivation is the so-called ***skewed X inactivation***. This means that in certain tissues always one – let's say - always the paternal X chromosome is inactivated. This may have far-reaching consequences. It is attempted to explain by this the higher frequency of certain autoimmune diseases (e.g. SLE) observed in females. In the thymus maturing T lymphocytes can only tolerate those antigens which are encoded by the active X-chromosome, and not the antigens coded by the other, the inactive one. Thus, all the cells / tissues where the other X chromosome is active are considered non-self, and immune response is generated against them, resulting in autoimmune disease symptoms. Of course, this cannot be the sole cause of autoimmune diseases, since it cannot be explained by this why the disease manifests in different ages.

7.8. Questions

1. What is the role of RNAs in cytoplasmic inheritance?
2. What kinds of dose compensation mechanisms are known?
3. What is the supposed role of skewed X inactivation in the development of autoimmune diseases?
4. Based on pedigree analysis how can we distinguish the X linked dominant inheritance from the autosomal dominant one?

5. What are homo- and heteroplasmy?
6. What can be the consequences of maternal heteroplasmy?
7. What do you know about the genetics of pre-eclampsia?
8. What are the characteristics of the inheritance of precocious puberty?
9. Which genes can escape the X inactivation?
10. What are the differences amongst the symptoms of a carrier woman, if the X-linked gene encodes a soluble or a cell-bound product?

8. Genetics of biological processes

Sára Tóth

In this section we give a brief insight into the genetics of 3 biological processes:

- a. Developmental genetics
- b. Oncogenetics
- c. Immunogenetics

8.1. Developmental genetics

The fact, how the several different cell types characteristic for our body can derive from a single fertilized egg, remained unexplained by biologist and geneticist for a long time. There are more than 200 different cell types in an adult body whose genetic material and DNA are identical, but functionally differ significantly. Just the epigenetic approach made / makes a more accurate interpretation of developmental changes possible.

During ontogenesis the developmental potentials are gradually narrowing, and thus from the ***totipotent zygote***, through the ***pluripotent inner cell mass (ICM) of the blastocyst***, and then as differentiation goes ahead, from ***the multipotent cells of the germ layers*** finally ***specific unipotent cells*** are formed, which are able to generate cells similar to themselves only. The ***differentiation*** starts with a naïve but versatile cell, which is genetically determined for a differentiation pathway and subsequently at first reversible and then irreversible changes result in the ***terminally differentiated state***, when the cell is showing all the features that are realized from the original genetic program.

However, it soon became clear that **ontogeny is actually none other than cell fate determination**. In the course of cell differentiation, information is needed both ***about lineage and position***, that is, ***cellular identity***. While in the former primarily ***intrinsic*** / internal (for example, the symmetric or asymmetric cell divisions), in the latter outer / ***extrinsic factors*** - cell-cell and cell-matrix interactions or soluble morphogens – play a role.

There are good examples for division asymmetry in neurogenesis (neuroblast → neuron and glia) and in male and female gametogenesis. Amongst the gametogenetic processes in the former, A and B spermatogonia, in the latter polocyte and egg are resulted in from asymmetrical divisions. **The asymmetry is not necessarily a morphological difference, often only functional differences are present in the cells derived (then in their**

descendants as well). Although the cause of asymmetry is not entirely clear in each case, it is assumed that certain gene products are not evenly distributed in the initial cell, and through influencing the spindle axis they lead to asymmetry. Another possible explanation of the asymmetry is based on the differences of the stability of mitotic spindle microtubules (kinetochore, astral, polar) and of the traction forces generated by them.

8.1.1. Morphogens

Morphogens involved in cell differentiation are soluble molecules, whose effects depend on their concentration gradient. Such a morphogen is the *activin*, which is able to determine different cell types depending on its concentration (e.g. in vitro ~ 0.1 ng / ml concentration mesenchymal, while in 1.0 ng / ml skeletal muscle differentiation is induced by).

Another *morphogen is the sonic hedgehog (SHH)*, which has a role in the differentiation of the neural tube, and in the separation of the eyes. The sonic hedgehog produced by ventral, central cells of the neural tube gradually diffuse to dorsal cells, where in almost negligible concentration sensory neurons are generated, on the other hand, from the ventral and lateral cells *due to its large(r) concentration motor neurons differentiate*.

8.1.2. Homeobox genes

At the very beginning of ontogeny maternal gene products act (see The role of sex in heredity chapter), later from amongst the embryo's own genes segmentation genes are expressed, and this is followed by the expression of positional identity and the axes (cranio-caudal and proximo-distal in the limb) determining *HOX (homeobox)* genes. The homeobox genes form a large gene family that is highly conserved evolutionarily, and the order of spatial expression of family members from Drosophila to mammals is the same. All members of the gene family code a transcription factor. As they regulate steps of the ontogeny through a transcription factor cascade, they are considered *master regulators* of development. All members of the family have a sequence of 60 codons, a *homeobox* by that the DNA-binding protein motive, the *homeodomain* of these transcription factors are determined. In humans, there are four *HOX* gene families: the *HOXA*, *HOXB*, *HOXC* and *HOXD* which include a variable number of genes.

The relationship between morphogens and *Hox* genes was elicited by experiments carried out on differentiating limbs of chick embryos. The so-called *zone of polarizing activity (ZPA)* of the limb expresses the sonic hedgehog morphogen, and as a result of its concentration

gradient, the HoxD family members - expressed in a specific order - create fingers. The morphogen - HOX gene relationship is shown by the fact that in humans both *SHH* and *HOXD* gene family mutations cause the same abnormality, e.g. the *holoprosencephaly*.

Of course, for the normal differentiation not only these early key genes are necessary, but a number of growth factor induced specific transcription factor cascades and the proper regulation of cell proliferation and apoptosis, too, but their detailed discussion far exceed the limits of this section.

8.2. The genetics of sex

A branch of developmental genetics is dealing with the sex determination and sex differentiation, and with all the genetic process by which the male and female gender-specific phenotype develops.

8.2.1. Male sex determination in mammals

Many years of research was needed to identify the gene, the molecular "master switch" that triggers the sexual differentiation process. This is the *SRY gene* localized to the short arm of the Y chromosome, near the pseudoautosomal (PAR1) region. **If the Y chromosome is present, and there is normal SRY, the male sexual differentiation takes place in any case.** The role of SRY was demonstrated by a *transgenic mouse* experiment when a DNA fragment corresponding to the mouse *Sry* was microinjected to the inner cell mass of a female blastocyst stage embryo with XX chromosomes. Such embryos were implanted to surrogate mothers and then mice were born with male genitalia, external sex characteristics and even with male sexual behaviors. So *Sry* was enough to change the gender; this is indeed the gene responsible for the male sex differentiation. However, these mice were not fertile, attributable to the presence of the two X chromosomes; a situation which is similar to human Klinefelter syndrome.

The SRY encodes a protein - previously has been named TDF, testis determining factor - which is in contrast to earlier concepts, not a conventional transcription factor, but a protein that after binding the DNA bends it, thus allowing the classical transcriptional factors to access neighbouring DNA sequences, genes.

The next step in **the differentiation cascade initiated by SRY** is **the anti-Mullerian hormone (AMH) or MIS (Mullerian inhibiting substance) production by the developing testicular Sertoli cells**. Thus, the development towards the female sex differentiation, that is,

the development of Mullerian duct is inhibited. Shortly afterwards the production of testosterone in Leydig cells starts, and this leads to the development of male gonads and external genitalia.

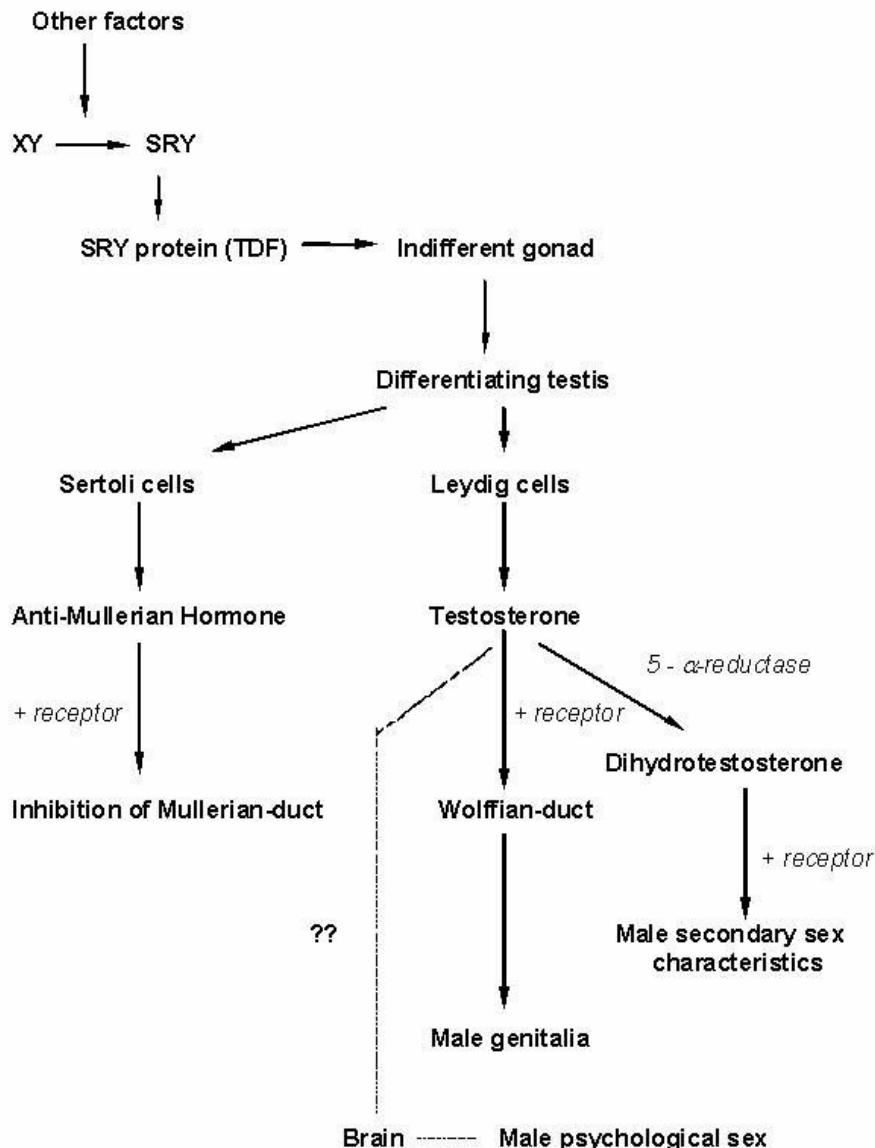


Figure 8.1. Male sex determination

Beside the experiments mentioned above, the role of SRY was suggested by human diseases associated with the abnormalities in sex development. Such is the **sex reversal** where in the presence of XX sex chromosomes male phenotype or at XY genotype female phenotype develops. The possible explanation is that in paternal meiosis, the obligate crossing over is not in PAR1, but it is shifted proximally towards the centromere. Thus, the SRY gene is transferred to the X chromosome, and thereby a recombinant, aberrant X and microdeleted Y is formed.

There are also sex revertants, when female phenotype is formed because of a ***mutated SRY***. In these cases, the HMG (high mobility group) part, the DNA binding domain of the protein is wrong, and in the absence of DNA binding the differentiation cascade cannot start.

Although the SRY alone is sufficient for male sex determination, i.e. to induce the differentiation, however, many other autosomal (e.g. chromosome 17 localized ***SOX9*** [SRY HMG box related genes] a transcription factor encoding gene), and ***X chromosome localized genes are necessary to switch on SRY and to the whole process of sexual differentiation***.

For the normal sexual differentiation not only the sufficient quality and quantity of the inducers, but their ***adequate receptors*** are necessary, too. Their mutations also cause disturbed sexual development.

The androgen insensitivity syndrome (AIS), formerly known as testicular feminization (X-linked recessive hereditary disease) drew attention to a gene localized on chromosome X, which is involved in male sexual differentiation. In this disease beside XY genotype and normal serum testosterone level female external sexual characteristics develop, although there are testes in the abdominal cavity! Since neither ovaries nor a uterus develop, these patients are sterile. It was concluded from the symptoms that the problem could be after testosterone induction in the differentiation cascade (either its receptor, or its signalization or the target genes may be incorrect). Finally, ***the testosterone receptor mutation(s)*** were verified as the cause of the syndrome.

The role of pituitary-derived hormones in the sex differentiation disorders was demonstrated by the symptoms of ***Kallman's syndrome*** patients. The most common symptoms are anosmia (lack of sensing smell) and the complete absence of testicular functions, although XY sex chromosomes are present. The disease is caused by a deletion of gene located proximally from the PAR1 region of the X chromosome. ***The gene encodes a cell adhesion protein which has a role in neuronal migration***. A part of these stem cells migrate to the olfactory nerve, another part to the hypothalamus during development. In the latter area they produce gonadotropin-releasing hormone (GHRH) and thus indirectly - through the gonadotropin synthesis of the pituitary – they effect on gonadal differentiation. This will be apparent in the unusual symptoms: hypogonadotrop-hypogonadism and anosmia. The ***KAL1*** gene has a Y chromosomal homologue, too but it is an inactive ***pseudogene***.

8.2.2. Development of female sex in mammals

At the time of the discovery of the *SRY* gene, in 1990, it was thought that female sex determination is a passive process, i.e. that in the absence of Y chromosome, so of *SRY*, definitely female gender would develop. Using the computer language it was told that this is the "default pathway". Later examining rare woman to man sex reversal families it turned out that there is a *female sex determining gene*, too, which is the *R-spondin1 (RSPO1)*. So if it is mutated male phenotype with 46, XX karyotype is formed. Unlike *SRY*, it defines a *soluble ligand* which competes with *WNT4* factor for a membrane receptor (frizzled) and triggers the β -catenin pathway, and leads to the target gene activation, and thereby the female sex determination and differentiation. Of course, like men, also a variety of other transcription and growth factors are needed to reach the terminally differentiated state, and perhaps they are even less known than the ones in male sex determination and differentiation. But it is certain that the components of the two systems mutually inhibit each other.

According to our current knowledge, in the bipotential gonads male and female determinants are in balance, and only later, at the time of the expression of *SRY* and *RSPO1* the balance is shifted to one way or the other.

Mutations in the steroid metabolism – which plays an important role in female sexual differentiation as well – cause the autosomal recessive *congenital adrenal hyperplasia or adrenogenital syndrome*. Then female infants of XX genotype are born with not obvious external genitalia, generally with enlarged clitoris. Other symptoms are adrenal enlargement, salt loss and lack of cortisone. The disease is due to *21 - α -hydroxylase enzyme mutation*. Due to this mutation the progesterone cannot be converted to deoxycortisone, but to 17-OH-progesterone. The latter one has an androgen-like effect, and it is responsible for the masculinisation of the external genitals. Although the incidence of the disease is 1:8.000-25.000 in the Caucasian population, among the Yupik Eskimos it is very common, 1:500. This is probably due to the fact that *heterozygotes have selection advantage* over bacterial infection caused by the *Haemophilus influenzae* type B strain, which causes not only simple cold but meningitis in the normal AA genotype Eskimos.

As a summary, the sex differentiation abnormalities can be primarily caused by the following inherited disorders:

- a. mutations of *SRY*, rarely of *RSPO1* or structural abnormalities affecting these genes
- b. disorders of steroid (androgen / estrogen) biosynthesis

- c. mutations of the androgen receptor
- d. defects of the AMH gene
- e. X0/XY mosaicism
- f. mutations in genes involved in the differentiation of mesoderm or the nephrogonotome
(for example *SF1*, *WT-1*)

Numerical aberrations involving the sex chromosomes can cause abnormalities in the sexual development because they push the balance of the regulatory elements of this process, while the listed disorders either influence the ***chromosomal sex*** or result in the abnormal development of genitals and sex specific phenotype, so that of the - ***gonadal and genital sex***.

However, there are also known disorders when both gonadal as well as the genital sex are normal, but abnormalities still exist in gametogenesis. For some cases autosomal mutations are responsible, for example the gonadotropin-releasing hormone or sex steroid biosyntheses are wrong, but Y-chromosome-linked hereditary infertilities are now also known, which are caused by Y-chromosomal long arm localized genes (e.g. *AZF* gene = **azoospermy factor**).

8.3. Stem cell biology

The development of biological knowledge has made the culture of mouse embryonic stem cells and then the creation of transgenic mice possible at the beginning of 1980s. Since then, the two areas interact, and the improving knowledge about stem cells has contributed to the expansion of genetic knowledge related to development and differentiation, which result in the first human embryonic stem cell line generation in 1996. From there the road led to the ***induced pluripotent stem (iPS) cell***. The potentials of pluripotent embryonic stem cells derived from the inner cell mass of the blastocyst were recognized early, and by the appropriate selection of culture conditions to differentiate them into a particular cell type – i.e. insulin-producing β -cells or dopaminergic neurons – their medical use is within reach. However, the use of human embryonic stem cells has always been a ***serious ethical problem***. Where should the used embryos come from? Is it justifiable to create human embryos only for experimental purposes? What would be the fate of the so called supernumerary embryos derived from in vitro fertilization? Answers to these questions were different in different countries, which were codified by the legislation of the country. (In Hungary, a new human

stem cell line cannot be created, but with permission, you can work with a few lines previously established abroad.)

Therefore the pioneering achievements of Shinya Yamanaka were accepted with great appreciation - eventually Nobel Prize awarded in 2012 - in the international scientific community. His method helped ***the adult differentiated (unipotent) cells to de-differentiate into the pluripotent cells of ICM or epiblast.*** The cell type created in this way is called ***induced pluripotent stem (iPS) cell.*** For re-programming pluripotency factors with known roles in developmental genetics were used in combinations that plasmids containing the genes of these factors transfected the target cells. Since among these (*LIF*, *Sox2*, *KLF4*, *cMYC*) there was an oncogene (*cMYC*) as well, the therapeutic use of cells thus created would not have been safe, so today there are many other ways for reprogramming, e.g. oncogene-free combinations, or only the mixtures of proteins responsible for pluripotency are used. Although the cells created can be directionally differentiated similarly to embryonic stem cells, and in the case of their therapeutic use immunological incompatibilities are not expected, since the cell donor and the recipient were the same person, unfortunately, ***the epigenetic changes taking place during differentiation and of course reprogramming are still not known in all details. We do not know the epigenetic pattern formation and erase in every detail, and even more we cannot directionally influence it.***

With all these iPS cells are of great importance in developmental genetics and in therapeutic research, as this would enable the treatment of many diseases and would overcome difficulties due to donor shortages.

8.4. Oncogenetics

Although the process and causes of carcinogenesis have already been discussed in several other subjects, this section will cover the major genetic events of the development of tumors, because at cellular level tumors may also be considered genetic disorders.

The cancers affect 1 in 3 people worldwide; a man has ~ 40% chance of the cancer. Even this high frequency indicates that **tumors are usually not of monogenic origin**, with the exception of rare monogenic tumors such as *retinoblastoma*, or *Li-Fraumeni syndrome*. **There are a number of underlying genetic susceptibility factors (mutations) and environmental effects.**

The cancer can be described as a group of diseases characterized by unlimited proliferation and spread of mutant cells in the body.

The following steps are the hallmarks of carcinogenesis:

- a. growth signal autonomy
- b. unlimited replicative potential
- c. evasion of growth inhibitory signals
- d. evasion of apoptosis
- e. angiogenesis
- f. invasion and metastasis

Mutations can be spontaneous as well as induced by some environmental factor, as it is discussed in the chapter of Mutations and polymorphisms. **Most mutagens are also carcinogens. The mutations of three large gene families play key roles in carcinogenesis.** These are the **oncogenes, tumor suppressor genes, and the so-called mutator genes.** The mutator genes involved in DNA repair (see there), so through their malfunctions mutations are fixed in the genome leading to tumorigenesis.

8.4.1. Oncogenes

Oncogenes are actually genes (proto-oncogenes) of changed normal function, which are essentially involved in cell cycle regulation. Such genes include genes encoding growth factors (such as EGF) and their receptors (such as EGFR), the components involved in their signal transduction (such as Ras, Raf) and transcription factors. Mutations of these lead to the growth factor independent unlimited cell proliferation – e.g. this can be the result of the constitutive activation of mutant receptor tyrosine kinases. **Oncogenes are activated not only by point mutations in the above players, but by gene amplification or chromosome translocations** (e.g. the t(9; 22) translocation leading to Ph1 chromosome described in chronic myeloid leukemia which results in a fusion protein with increased tyrosine kinase activity) as well.

In addition to classic genetic alterations, epigenetic changes - epimutations - also can cause oncogene activation. It is known that increased genome hypomethylation during aging often affects oncogenes. This not only explains the higher activity of oncogenes, but the known phenomenon that certain cancers' incidence increases with age. **A specific example for the relationship between oncogenes and epigenetics is given by the**

imprinted IGF2 (insulin-like growth factor 2). Normal colonic epithelial cells express only the maternal allele, but in colon tumors the imprinting is lost (LOI = Loss Of Imprinting), the paternal allele is expressed, and the tumor develops.

8.4.2. Tumor suppressor genes

The evasion of growth inhibitory signals is due to mutations of the tumor suppressor genes. The normal tumor suppressors together with protooncogenes regulate cell cycle and control the integrity of the genome. If damage is detected in the genome, the cell cycle is stopped and the cell starts error correction. On the bases of these two sub-functions there are **gate keepers and care takers** mentioned. The former includes the classic tumor suppressors, e.g. **RB** and **TP53** genes, the latter **the DNA repair genes - also known as mutator genes** - (e.g. **MLH1** and **MSH2 mismatch repair genes**).

While a single mutant allele of protooncogenes is sufficient for oncogenesis, so there must be dominant mutation, in the case of tumor suppressors both alleles should be mutated for the loss of the growth inhibiting function. Here, then, the mutant is recessive. In the care taker or mutator genes the **haploinsufficiency phenomenon** may play a role in oncogenesis, as in the case of mutation of one allele, the remaining normal allele has only reduced ability to function, and in many cases even this is sufficient to tumor induction due to the large number of uncorrected mutations.

Knudson set up the **so-called two-hit hypothesis** after investigating the tumor suppressors (RB). Thus, the development of certain cancers requires two successive mutational events affecting tumor suppressor genes. It is usually already inherently present (**familial retinoblastoma**), while the other is formed only in one or certain organs, and as the previously heterozygous state is lost, the homozygous mutant tumor suppressor gene leads to tumor formation. In **sporadic cases**, both mutations take place in the same person. The phenomenon is called **loss of heterozygosity** = **LOH**, and after being identified by the modern molecular biological experimental methods, it may be suitable for the detection of pre-cancerous condition.

Similarly to oncogenes and tumor suppressor genes, epigenetics and epimutations may play a role as well. While **CpG dinucleotides** in the promoters of normal tumor suppressors are not methylated, thereby ensuring gene expression, in tumors they are often hypermethylated so the transcription of the gene is inhibited, and the protection against excessive cell proliferation is lost. Another epigenetic relationship is the formation

of tumor suppressor protein and ***HDAC*** (*histone deacetylase*) complex. The normal suppressor proteins interact with HDAC, thereby triggering the chromatin remodelling, the heterochromatinization which limits the functioning of genes in the affected area, thereby inhibiting cell proliferation. The mutant suppressors are unable to do so, therefore the euchromatic structure remains and proliferation continues.

8.4.3. Anti-apoptotic genes

The ***TP53*** tumor suppressor gene as a guardian of the genome plays a role not only in cell cycle arrest and DNA repair stimulation just after the DNA damage, but also in the induction of apoptosis, when for large scale irreparable damages are present. That is because of the mutation not only the cell cycle may proceed invariably, but also severely damaged mutant cells will survive, so ***TP53*** mutations bilaterally contribute to tumorigenesis. This can explain why ***Li-Fraumeni syndrome associated mutations of this gene cause a wide variety of tumors affecting many different organs simultaneously.***

Malfunctions of the intrinsic apoptotic pathway, due to the involvement of ***p53 and /or mitochondrial Bcl-2*** may cause not only the lack of apoptosis, but the resistance of tumor cells to chemotherapy.

8.4.4. Telomerase

It is known that eukaryotic DNA is shortened in somatic cells from division to division because of the characteristics of replication. This occurs in the subtelomeric and telomeric repetitive sequences of chromosomes, and following approx. 50-70 divisions it leads to cell senescence, arrest of cell division and aging. In germ line cells **the telomerase enzyme, which comprises a reverse transcriptase, and a telomeric DNA complementary RNA can restore the length of the telomere.** It's crucial in the transmission of the same sized genome from generation to generation. However, ***telomerase activity is also linked to cancer cells. They can restore the telomeres either by up-regulating telomerase enzyme or by recombination based alternative telomere lengthening.***

If a cell - due to different mutations - avoids cell death caused by the extreme short telomeres, its genome becomes unstable, leading to the oncogenic transformation of the cell through the aforementioned mutations (amplifications, translocations). This can be further strengthened by mutated genes induced telomerase (e.g. c-MYC via binding to the promoter of telomerase can activate it).

8.5. Immunogenetics

In teaching immunology a number of genetic processes - critical in the function of the immune system - were also discussed. Of these, perhaps the most specific is the process, which leads to the enormous diversity of immunoglobulins (B cell receptors) and T-cell receptors. **All people are capable to produce approx. 10^{11} different antibodies, although the human haploid genome size is "only" 3×10^9 bp.**

Somatic gene rearrangement and somatic mutations make the elimination of this discrepancy. There are only 3 immunoglobulin (Ig) / B cell receptor (BCR) loci (*IGH*, *IGK*, and *IGL*) in the human genome which determine the heavy and the two light chains (κ and λ), and four T-cell receptor (TCR) loci (*TRA*, *TRB*, *TRG* and *TRD*) which make the four (α , β , γ and δ) TCR chain synthesis possible.

Despite the enormous diversity, each individual B-and T-cell is **monospecific**, i.e. only one type of Ig or TCR heterodimer with a unique antigen binding site can be produced, which are specific for one antigen only. Different B- and T-cells express different Ig and TCR heterodimers specific to different antigens, so this enables the population of billions of cells of the immune system to recognize virtually any antigen. Diversity is due to the specific organization and expression of the corresponding genes.

Take, for example, a reminder of the immunoglobulins. Each Ig consists of 4 chains, 2 heavy (H) and two light (L) chains. Each chain contains a variable (V) and a constant (C) region. We cannot find complete genes for heavy and light Ig chain determination in the human genome, but each H and L chain are defined by a number of separate genes. The heavy chain variable region has three domains: the V (variable), D (diversity) and J (joining). (As for the light chain, the D segment is missing.) **At H locus about 200 V, approx. 30 D and J gene 9 (including 3 pseudogenes) are found.** These genes are inactive in the cells of non-immune organs and become active only during the T-and B-cell maturation. In the primary immune organs one from these genes (a V, a D and a J) are randomly combined and brought together that a new fusion exon is formed, which determines the H chain variable region.

This process is called **somatic gene rearrangement or somatic recombination** (Figure 8. 2), which is achieved through **DNA splicing**. The **RAG1 and -2 enzymes encoded by the recombination activating gene 1 and 2 are involved V-D-J recombination.**

The D-J and the V-DJ rearrangements of immunoglobulin heavy chains and the V-JC recombinations of the λ - and V-J rearrangements of the κ -light chains are carried out in this way. The further rearrangements of both heavy (VDJ-C) and light chains (VJ-C) are due **to mRNA splicing**.

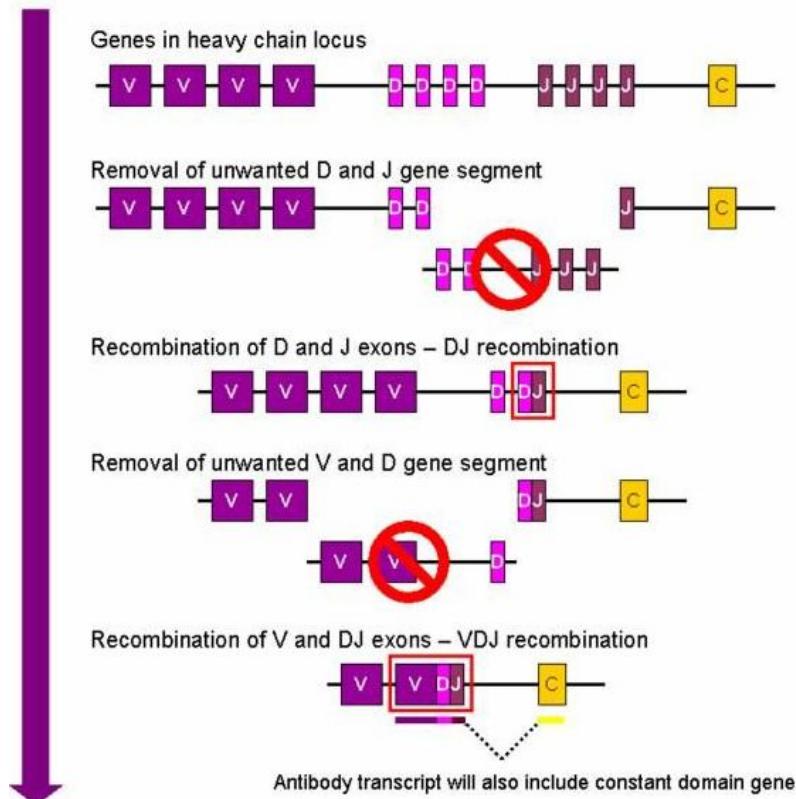


Figure 8.2. Somatic gene rearrangement

RNA splicing also takes place during the switch from the membrane-bound IgM C domain to the soluble IgM C domain.

T-cell receptor somatic gene rearrangement takes place in similar manner that is observed in the case of immunoglobulins.

The diversity is further increased by the facts that recombination can be shifted a few bases in 5' to 3' direction and that the RAG recombinases can cause double strand DNA breaks, whose repair is inaccurate, and therefore the specificity of antibodies can be even more different.

On the other hand, there is **somatic hypermutation** by which random base-exchange mutations occur in the V region of B cells. This mechanism does not function in other cells, and other genes are not affected, only a ca. 1.5 kb region. This only takes place during the activation of B cells: when it starts to divide after interacting with an antigen, **the resulting somatic hypermutation alter the antigen-binding region**. The cells binding the antigen best

survive and divide more than the other B cells. This process is called ***affinity maturation***. This is triggered by the ***activation induced the cytidine deaminase (AID)***, which deaminates cytidine to uracil. This base mismatch - not exactly repaired by a variety of mechanisms - can result in a number of different mutations.

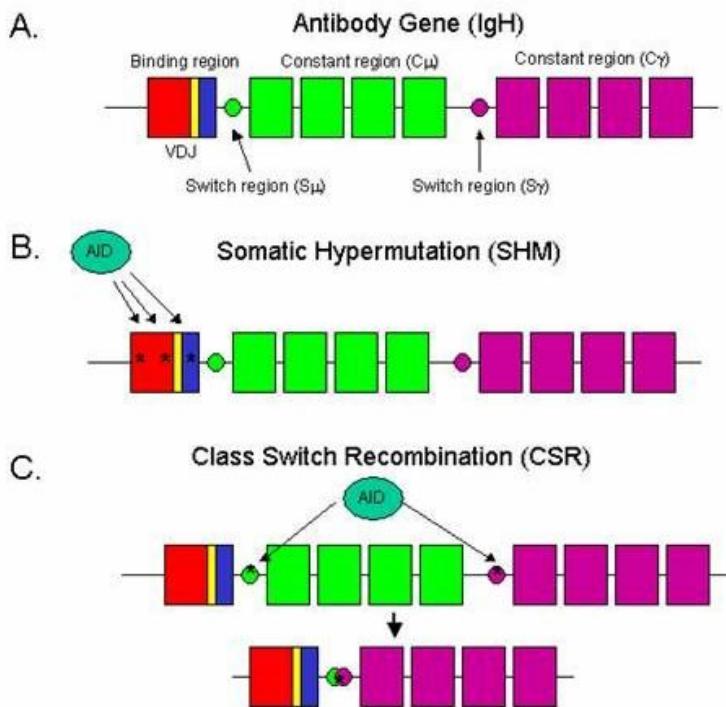


Figure 8. 3. Somatic hypermutation and class switch

Another case of somatic gene rearrangements is the ***immunoglobulin class switching***. The heavy chains on the bases of their C region can be divided into five classes. Each H gene contains the C regions of all 5 Ig classes in a chromosomal organization that the C region of IgM is closest to the V region. There is a number of different C regions of the IgG and the IgA classes which are further divided into subclasses. IgM is the first type of antibodies that is produced by each B cell. However, after a while, B cell switches to another class of antibodies. It is ***a third DNA splicing***, in which the DNA is spliced out between VDJ and constant regions, thereby creating a new class of Ig. **Due to the fact that the variable region will remain unchanged, the antigen specificity will not change either. Its affinity towards the antigen remains unchanged, only interacts with other effector molecules.**

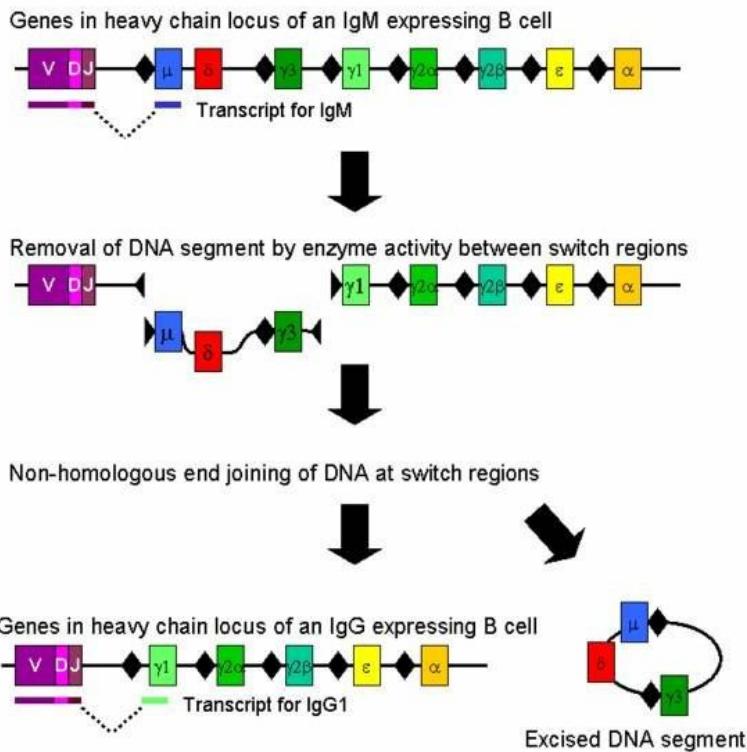


Figure 8. 4. Imunoglobulin class switch: IgM → IgG

Such diversification mechanisms often result in non-functional Ig genes: they contain either stop codons or the reading frame is shifted.

In developing B cells a so-called ***allelic exclusion*** mechanism is used, in which each B cell only produces one active L chain and one active H chain. The cell tries each copy of the L genes and each copy of H genes: when an active chain is created, no further DNA splicing is needed. However, if a non-functional Ig is created, the cell then tries the next H or L gene. This process continues until the active H and L chains are completed, or until all of the genes have been tried (in this case, however, the cell dies).

In addition to the above mechanisms, certain epigenetic mechanisms also play role in diversification, for example histone modifications and subsequent chromatin remodelling create a so-called recombination center, thus the Ig or TCR regions become accessible by RAG recombinases.

The immunological role of epigenetics is enlightened by a previously not mentioned monogenic disease, the ***ICF syndrome***. This *Immunodeficiency (agammaglobulinaemia), Centromeric instability of chromosomes 1, 9, 16 and Facial dysmorphia* associated disease is due to the mutation of ***DNMT3B***, a DNA de novo methyltransferase, an autosomal gene (chromosome 20). Then, although the mutation affects only one gene, the varied symptoms

are the consequences of defective methylation of many other genes, i.e. they are due to the lack of correct spatial and temporal methylation.

8.6. Useful web-sites:

www.imgt.org

www.p53.iarc.fr6index.html

www.methylgene.com

www.cancer.org

8.7. Questions

1. What kind of oncogene activation mechanisms do you know?
2. What is LOI?
3. What is the function of care taker and gatekeeper genes?
4. What do you know about the iPS cells?
5. What is sonic hedgehog and what is its effect based on?
6. What is the role of HOX genes?
7. What is the role of SRY and RSPO1?
8. Give an example of epigenetic changes related to carcinogenesis!
9. Explain the Knudson's hypothesis!
10. Why is somatic recombination not regarded as an epigenetic mechanism?

9. Introduction to genomics

Csaba Szalai

9.1. Genomics

Although the science of genomics has a past of several decades, it became well-known only in the last 20 years even for the majority of natural scientists. It is one of the most rapidly developing scientific areas, but still for most people, even for physicians or pharmacists graduated before the 90s, it covers mainly unknown concepts. Because of this, here I will define some terms.

First of all: What is the genome? The **genome** is the entirety of an organism's hereditary information. It is encoded either in DNA or, for many types of viruses, in RNA. The genome includes both the genes and the non-coding sequences of the DNA/RNA. The genome in humans is the total haploid DNA content of a diploid cell, plus the mitochondrial DNA. Because there is a difference between the female and male genomes, since males have two types of sex chromosomes (X and Y), in this definition we have to take this also into consideration.

The next important question is: what is genomics? There are possibilities for a number of definitions, but perhaps the simplest is: **Genomics** is the study of the function, structure and interactions of the genome. The term genomics also involves the special genomic methods. Besides studying the DNA, genomics among others can also involve the study of RNA (transcriptomics) and proteins (proteomics), and bioinformatics. Because in English a lot of terms in this area end with “omics”, **omics** has become a new synthetic term, and been used widespread in biology and related sciences (<http://en.wikipedia.org/wiki/Omics>; <http://www.omicsworld.com/>).

In some definitions genomics is defined as the synonym of molecular systems biology, in which the life is studied on the level of genomic organizations. But, genomics is **rather part of the systems biology**. According to the subject of the genomics, there are several subtypes of genomics, like structural genomics, comparative genomics, plant genomics, human genomics, pharmacogenomics or medical genomics, etc. Here we deal mainly with the last three of them.

There is still one important question, which can often cause problems for a lot of people. What is the difference between genetics and genomics? In reality, there is no sharp difference between the two terms, but in general, if a gene or a genetic variation is studied, or the

heredity of some traits is investigated, then it is called genetics. If several genes or the genome as a system are studied, then it is genomics. Because genomics is part of the systems biology, it requires more complex and sophisticated methods. However, even in the scientific usage, genetics and genomics are often interchangeable.

9.2. Human Genome Project

The big leap in our knowledge in genomics can be attributed to the Human Genome Project (HGP). Here, I introduce some important aims and results. More details can be found in <http://www.genome.gov/10001772>.

The HGP was developed in collaboration between the National Institute of Health (NIH) and the United States Department of Energy (DOE), and begun on October 1st, 1990 to map the human genome. The participation of the DOE in the initiation of HGP can be explained that after atomic bombs were dropped during World War II, Congress told DOE to conduct studies to understand the biological and health effects of radiation and chemical by-products of all energy production. DOE experts thought that the best way to study these effects was at the DNA level. Some aims of the HGP were:

- Mapping and Sequencing the Human Genome
- Mapping and Sequencing the Genomes of Model Organisms (like mouse, drosophila (fruit fly), the first plant *Arabidopsis thaliana* (thale cress), chimpanzee, *Caenorhabditis elegans* (round worm), pathogens, domestic animals, etc.)
- Data collection and analysis and storage the direct product of the human genome
- Ethical and Legal Considerations
- Research Training
- Technology Development
- Technology Transfer (into the private sector)

More details: <http://www.genome.gov/10001477>. The overall budget was \$3 billion, and it was planned for 15 years.

The project started with some difficulties, since in 1998, about at the half time of the whole project, only 5% of the sequence of the human genome was known, and it seemed hopeless to carry out the project successfully. **Craig Venter**, one of the leaders of the project suggested a new method, called “shotgun sequencing” (see: e.g. http://en.wikipedia.org/wiki/Shotgun_sequencing: “In

shotgun sequencing DNA is broken up randomly into numerous small segments, which are sequenced using the chain termination method to obtain reads. Multiple overlapping reads for the target DNA are obtained by performing several rounds of this fragmentation and sequencing. Computer programs then use the overlapping ends of different reads to assemble them into a continuous sequence.”) The other leaders, including **Francis Collins**, director of the [National Human Genome Research Institute](#) (NHGRI) (https://en.wikipedia.org/wiki/National_Human_Genome_Research_Institute) opposed this suggestion arguing that although the method worked for small bacterial genomes, but the human genome is too large for such a method. Because of this, Venter left the HGP and sought funding from the private sector to fund Celera Genomics (https://en.wikipedia.org/wiki/Celera_Corporation). The goal of the company was to sequence the entire human genome and release it into the public domain for non-commercial use in much less time and for much less cost than the public human genome project. The goal consequently put pressure on the public HGP and spurred several groups to redouble their efforts to produce the full sequence. DNA from five demographically different individuals was used by Celera to generate the sequence of the human genome; one of the individuals was Venter himself. In 2000, Venter and Francis Collins jointly made the announcement of the mapping of the human genome, a full three years ahead of the expected end of the HGP. On the 15 February 2001, the Human Genome Project consortium published the first Human Genome in the journal *Nature*, and was followed, one day later, by a Celera publication in *Science*. Despite some claims that shotgun sequencing was in some ways less accurate than the clone-by-clone method chosen by the HGP, the technique became widely accepted by the scientific community and is still the de facto standard used today.

The published genome in both cases contained the so-called draft sequence of human genome, which means that it contained several gaps, and sequencing mistakes. On average, the whole genome was read at 4-5 fold coverage. Later most of the gaps were filled, and the mistakes corrected at higher coverage, but this project is still in progress even today. The official end of the HGP was announced in April 2003 with fewer gaps, and at 8-9 average coverage. More details: http://en.wikipedia.org/wiki/Whole_genome_sequencing.

9.3. DNA sequencing

DNA sequencing is the process of reading the nucleotide bases in a DNA molecule. Since the beginning of the HGP it has been developing continuously. In HGP the DNA was sequenced with Sanger method, i.e. with dideoxy or chain termination sequencing (http://en.wikipedia.org/wiki/Sanger_sequencing). The whole budget of the HGP was \$3 billion. The first man, whose genome was sequenced was Craig Venter for \$70 million. In 2001 the sequencing of one human genome took a minimum of 1 year. It is obvious that both the price and the time are not appropriate for routine investigations, or even for sequencing several human genomes. It became clear that the Sanger method could not be developed much further to become much cheaper and faster. But it was also obvious that much cheaper and faster sequencing would have an immense leap in pharmaceutical research, personal medicine, but it could be used for countless aims. To accelerate the development, the **Archon Genomics X PRIZE** was established. The J. Craig Venter Science Foundation offered the \$500,000 (US) Innovation in Genomics Science and Technology Prize in September 2003 aimed at stimulating development of less expensive and faster sequencing technology. The announcement was: US\$ 10 million prize is to be awarded to "the first Team that can build a whole human genome sequencing device and use it to sequence 100 human genomes within 30 days or less, with an accuracy of no more than one error in every 1,000,000 bases sequenced, with an accuracy rate of at least 98% of the genome, and at a recurring cost of no more than \$1,000 (US) per genome (http://en.wikipedia.org/wiki/Archon_X_Prize; Figure 9.1).



Figure 9.1. Archon Genomics X PRIZE

Source: (([5](#); 15/02/2013.)

This was announced in 2006, and at that time it seemed to be utopian, but the high demand for low-cost sequencing has driven the development of high-throughput sequencing (also called as next-generation sequencing, or **NGS**) technologies that parallelize the sequencing process, producing thousands or millions of sequences at once. Some example for the new techniques:

- A parallelized version of **pyrosequencing** was developed by **454** Life Sciences, which has since been acquired by **Roche Diagnostics**.
- Solexa, now part of **Illumina**, developed a sequencing technology based on **reversible dye-terminators**.
- Applied Biosystems' **SOLiD** technology employs **sequencing by ligation**.
- **Ion semiconductor sequencing** developed by Ion Torrent Systems Inc. (now owned by Life Technologies). This method of sequencing is based on the detection of hydrogen ions that are released during the polymerization of DNA, as opposed to the optical methods used in other sequencing systems.

More details: http://en.wikipedia.org/wiki/DNA_sequencing

The methods were so successful that in 2007 the new generation sequencing (NGS) became the method of the year (<http://www.nature.com/nmeth/journal/v5/n1/full/nmeth1157.html>) in *Nature Methods* magazine. In 2007 the genome of James Watson was sequenced with the 454 technology in 2 months and for \$1 million. It was still far away from the aim, but it was a big step ahead. Since then, the price has been lower and lower, and the time shorter and shorter (Figure 9.2). E.g. in June 2009, Illumina announced that they were launching their own Personal Full Genome Sequencing Service at a depth of 30× for US\$48,000 per genome.

In November 2009, Complete Genomics published a peer-reviewed paper in *Science* demonstrating its ability to sequence a complete human genome for US\$1,700. If true, this would mean the cost of full genome sequencing has come down exponentially within just a single year from around US\$100,000 to US\$50,000 and now to US\$1,700.

In 2011 Complete Genomics charges approximately US\$10,000 to sequence a complete human genome (less for large orders).

In May 2011, Illumina lowered its Full Genome Sequencing service to US\$5,000 per human genome, or US\$4,000 if ordering 50 or more.

In January 2012, Life Technologies introduced a sequencer to decode a human genome in one day for \$1,000.

In October 2011 the Xprice foundation revised the rules. The objective was to sequence the genomes of 100 centenarians with high accuracy and 98% completeness within 30 days for \$1000 or less. Interest was tepid, however, and only two of the eight contenders in the original contest registered by the 31 May deadline—the company Ion Torrent, and a lab at Harvard University. In 2013 XPrize announced that the foundation was calling off the contest because it “was not incentivizing the technological changes” laid out by the XPrize board and the genomics prize’s chair, genome sequencer J. Craig Venter. XPrize Senior Director Grant Campany said the fact that only two competitors signed up suggested that the \$10 million prize wasn’t a sufficient incentive for sequencing companies, which are already making hundreds of millions of dollars, to invest their R&D funds in the challenge. At the same time, he says, no company is sequencing whole genomes to the accuracy the contest required—the focus is on less accurate sequences and the subset of the genome that codes for proteins.

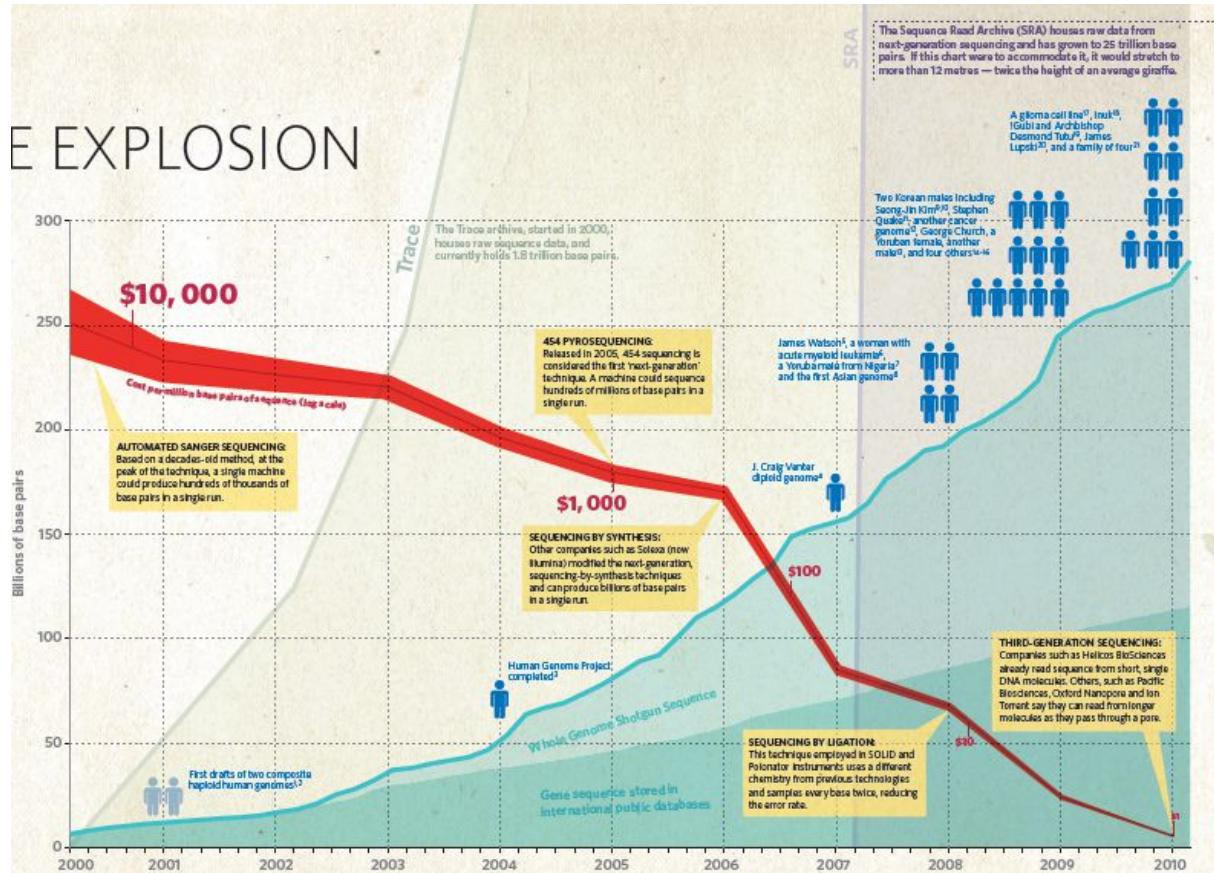


Figure 9.2. Changing of the price of DNA sequencing (red line) and the amount of DNA sequence data between 2000 and 2010 in logarithmic scale. In 2000, the price of sequencing 1 million DNA was \$10 thousand, which reduced in 2010 to \$1. The finished DNA sequence in 2000 started with 8 million base pair (bp) and doubled in every 18 months. By 2010 it increased to 270 billion bp. But this number is dwarfed comparing to the raw data that has been created and stored by researchers around the world in *Trace archive and Sequence Read Archive* (SRA). Here, the amount of data was 25 trillion bp in 2010, which in this scale would be 12 m high, twice the height of a giraffe.

Source: <http://www.nature.com/news/2009/091021/full/464670a.html>; 15/02/2013.

Taking advantage of the development in sequencing, several genome projects have been launched, like the 1000 Genomes project. The **1000 Genomes Project**, launched in January 2008, is an international research effort to establish by far the most detailed catalogue of human genetic variation. Its aim was to sequence the genomes of at least one thousand anonymous participants from a number of different ethnic groups. In 2010, the project finished its pilot phase, which was described in detail in a publication in a *Nature* paper.

Because of the rapid development of the NGS, now they are planning to sequence the genome of 2500 persons (<http://www.1000genomes.org/>).

Similar projects are the **Genome 10k project**. It aims to assemble a genomic zoo—a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately one for every vertebrate genus (<http://genome10k.soe.ucsc.edu/>), and the **i5k project**, which plans to sequence the genomes of 5,000 insect and related arthropod species over 5 years <http://www.arthropodgenomes.org/wiki/i5K>).

As the technology developed larger and larger genome projects were initiated. In 2015 the **100,000 genome project** was launched. The project plans to sequence 100,000 genomes from patients with rare diseases, plus their families, and patients with cancer. Combining genomic sequence data with medical records is also planned. It is also studied how to use genomics in healthcare and how best to interpret the data to help patients.

(<https://www.genomicsengland.co.uk/the-100000-genomes-project/>).

In 2015 a still larger project was proposed by the administration of president Obama in the USA. It was proposed to spend \$215 million on a “precision medicine” initiative the centerpiece of which will be a national study involving the health records and **DNA of one million volunteers**. (<https://www.technologyreview.com/s/534591/us-to-develop-dna-study-of-one-million-people/>)

9.4. Participants in the Human Genome Project

First, it is interesting, whose genomes have been sequenced first? In *Science*, where Venter et al. published one of the two papers about the sequence of the human genome (<http://www.sciencemag.org/content/291/5507/1304.long>), the sample selection was detailed as follows: “...the initial version of a completed human genome should be a composite derived from multiple donors of **diverse ethnic backgrounds**. Prospective donors were asked, on a voluntary basis, to self-designate an ethnogeographic category (e.g., African-American, Chinese, Hispanic, Caucasian, etc.). We enrolled 21 donors. Three basic items of information from each donor were recorded and linked by confidential code to the donated sample: **age, sex, and self-designated ethnogeographic group**. From females, □130 ml of whole, heparinized blood was collected. From males, □130 ml of whole, heparinized blood was collected, as well as five specimens of semen, collected over a 6-week period.” ... “**DNA from five subjects** was selected for genomic DNA sequencing: **two males and three**

females—one African-American, one Asian-Chinese, one Hispanic-Mexican, and two Caucasians.”... “The decision of whose DNA to sequence was based on a complex mix of factors, including the goal of achieving diversity as well as technical issues such as the quality of the DNA libraries and availability of immortalized cell lines.”

The official HGP collected the samples in two centers, with similar criteria (<http://www.nature.com/nature/journal/v409/n6822/full/409860a0.html>). It must be added, however, that later it turned out that Celera sequenced mainly the genome of Craig Venter, thus he became the first named person, whose genome was sequenced.

In the HGP, altogether 18 countries participated, but the USA had the main role. There are several web pages about the results of the HGP: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml), http://en.wikipedia.org/wiki/Human_Genome_Project, etc.

9.5. Some results of the HGP

The sequencing of the human genome has brought several interesting, sometimes unexpected results. Our knowledge about it has been continuously expanding since then, and will be expanding for decades. Perhaps one of the most unexpected results was that the human genome contains **hardly more than 20 thousand genes**. Originally, most experts estimated the number of genes to be around 100 thousand. This story shows how unexpected was this result for the experts: “ Between 2000 and 2003, a light-hearted betting pool known as “GeneSweep” was run in which genome researchers could guess at the number of genes in the human genome. A bet placed in 2000 cost \$1, but this rose to \$5 in 2001 and \$20 in 2002 as information about the human genome sequence increased. One had to physically enter the bet in a ledger at Cold Spring Harbor, and all told 165 bets were registered. Bets ranged from 25,497 to 153,438 genes, with a mean of 61,710...” (Source: <http://www.genomicron.evolverzone.com/2007/05/human-gene-number-surprising-at-first/>). Thus, the lowest bet for the gene number was 25,497, which won ultimately, although it was still higher than the actual gene number, which were around 21 thousand. In Table 1 there are some statistical data about the human genome, updated in February 2016.

Size of the genome

Golden Path Length (bp): 3,096,649,726

Gene counts

Coding genes	20,313
Non-Coding Genes	25,180
Small non coding genes	7,703
Long non coding genes	14,896
Misc non coding genes	2,307
Pseudogenes (noncoding sequences similar to an active protein)	14,453

Sequence variations

Short Variants (SNPs, indels, somatic mutations): 149,490,457

Structural variants: 4,149,389

Table 1. Some statistical data about the human genome. Source:
http://www.ensembl.org/Homo_sapiens/Info/Annotation

Some interesting results of the human genome, corrected with new data, from e.g. the 1,000 genome project:

- Largest gene: *DMD*, which codes for **dystrophin**; size: 2,224,919 bases; location: Xp21.2
- Longest coding sequence: *TTN*, codes for **titin**; coding sequence: 104,076 bp; 34,692 amino acid
- Longest exon: *TTN*: 17,106 bp
- Most exon: *TTN*; 351
- 20% of the genome is gene desert (a region >500 kbp without a gene)
- Gene rich chromosomes: 17, 19, 22 (the **richest is the 19**, with 1,458 coding and 980 non-coding genes)
- Gene-poor chromosomes: Y, 4, 13, 18, and X; the **poorest is the Y** with 72 coding and 137 non-coding genes and < 1.0 gene/Mb
- The 5' end of the 98.12% of the **introns are GT bases and AG at the 3' end**; 0.76% is GC-AG

- The **recombination is higher in females** than in males, but the number of **mutations is higher in male meiosis**, which means that the majority of the mutations originates from males.
- Every new-born receives about **60 mutations from the parents**.
- Every individual has **250-300 loss-of-function mutations** in the annotated genes, among which **50-100 genes are involved in Mendelian diseases**. It shows, among others, why it is so dangerous when the parents are relatives. The closer is the kinship, the higher is the probability that the child receives two mutations from the same gene, resulting in recessive diseases, or even multigenetic syndromes.
- 46% of the human genome consists of **repeats**. A lot of them are transposons, i.e. jumping genes, inactivated about 40 million years ago. The most frequent repeats are called **Alu**, which occupy of the 10.6% of the genome.
- 145 human genes originate from bacteria, through horizontal gene transfer.
- There are long repeated regions in the **pericentromeric and subtelomeric regions**.
- At present 156 imprinted genes are known. **Imprinting** is a genetic phenomenon by which certain genes are expressed in a parent-of-origin-specific manner. Appropriate expression of imprinted genes is important for normal development, with numerous genetic diseases associated with imprinting defects including Beckwith–Wiedemann syndrome, Silver–Russell syndrome, Angelman syndrome and Prader–Willi syndrome. 56% of these genes are maternally, 44% are paternally imprinted (<http://www.geneimprint.com/site/genes-by-species>; http://en.wikipedia.org/wiki/Genomic_imprinting).
- There are 27-29,000 CpG islands. **CpG islands** or **CG islands** are genomic regions that contain a high frequency of CpG sites (http://en.wikipedia.org/wiki/CpG_island). The "p" in CpG refers to the phosphodiester bond between the cytosine and the guanine, which indicates that the C and the G are next to each other in sequence. 99% of the **methylation** occurs at CG dinucleotides, which influence the transcription of the nearby genes, and play important roles in genetic regulation, imprinting and cell differentiation. The methylation occurs on the cytosine. About 70% of human promoters have a high CpG content. In the **ENCODE** project it was found that 96% of CpGs exhibited differential methylation in at least one cell type or tissue assayed, and levels of DNA methylation **correlated with chromatin accessibility**. **Methylation in the promoter reduces, in the gene bodies increases the expression of the genes**. In stem cells 25% of the methylation occurs in CA, instead of CG.

- Besides methylation of the CpG islands, **modifications** (methylation, acetylation, etc.) of the **histone proteins** around the chromosomes also play an important role in the **regulation of gene expression**. To study these phenomena the **Human Epigenome Consortium** was founded and the **Human Epigenome Project** was launched (<http://www.epigenome.org/>). From these a new scientific area has been formed, called **epigenomics, or epigenetics**.
- There are two different genome region types, which participate in the regulation of gene expression. **Promoter** regions located near the genes they transcribe, on the same strand and upstream, towards the 5' region of the sense strand; and the **enhancer regions** that regulate expression of distant genes. Beyond the linear organization of genes and transcripts on chromosomes lies a more complex (and still poorly understood) network of **chromosome loops and twists** through which promoters and more distal elements, such as enhancers, can communicate their regulatory information to each other. In the ENCODE project more than 70.000 promoter and nearly 400.000 enhancer regions were detected. **Enhancers** are often **cell-type specific**.
- Several **paralogous genes** have been detected. According to the definition, paralogs are two genes or clusters of genes at different chromosomal locations in the same organism that have structural similarities indicating that they derived from a common ancestral gene, and have since diverged from the parent copy by mutation and selection or drift. By contrast, **orthologous genes** are ones which code for proteins with similar functions, but exist in different species, and are created from a speciation event.
- Altogether 14,453 pseudogenes have been detected so far. In contrast to paralogs, **pseudogenes** are dysfunctional relatives of genes that have lost their protein-coding ability or are otherwise no longer expressed in the cell. **Duplicated pseudogenes** have intron-exon-like genomic structures and may still maintain the upstream regulatory sequences of their parents. In contrast, **processed pseudogenes**, having lost their introns, contain only exonic sequence and do not retain the upstream regulatory regions. In the human genome, processed pseudogenes are the most abundant type due to a burst of retrotranspositional activity in the ancestral primates 40 million years ago. Originally thought as functionless, pseudogenes have been suggested to exhibit different types of activity. Firstly, they can **regulate the expression** of their parent gene by decreasing the mRNA stability of the functional gene through their over-

expression. A good example is the *MYLKP1* pseudogene, which is up-regulated in cancer cells. The transcription of *MYLKP1* creates a non-coding RNA (ncRNA) that inhibits the mRNA expression of its functional parent, *MYLK*. Moreover, studies in *Drosophila* and mouse have shown that small interfering RNA (siRNA) derived from processed pseudogenes can regulate gene expression by means of the RNA-interference pathway, thus **acting as endogenous siRNAs**. In addition, it has also been hypothesized that pseudogenes with high sequence homology to their parent genes can regulate their expression through the generation of **anti-sense transcripts**. Finally, pseudogenes can compete with their parent genes for **microRNA** (miRNA) binding, thereby modulating the repression of the functional gene by its cognate miRNA. According to predictions, at least 9% of the pseudogenes present in the human genome are actively transcribed.

Human knockouts. On average, every person carries mutations that inactivate at least one copy of 200 genes and **both copies of around 20 genes**. These ‘loss of function’ mutations have long been implicated in monogenic diseases. Most, however, seem to be harmless — and some are even beneficial to the persons carrying them. E.g. lack of the *CCR5* genes renders individuals to be protected against HIV-1 infection. Individuals who are homozygous for null mutations of the fucosyltransferase 2 (*FUT2*) gene do not secrete ABO antigens and are protected against some strains of Norovirus. Individuals lacking a gene called *LPA* that codes for a blood lipoprotein have a significantly lower risk of heart attacks and stroke.

There are several web pages containing information about the genomes of human and other organisms (e.g.: <http://genome.ucsc.edu/>; <http://www.ensembl.org/>; <http://www.ncbi.nlm.nih.gov/>). There is still an important topic, not detailed above, which is about the variations in the genome. We consider it, however, so important that there is a special subchapter for this topic (see below).

The mapping of the human genome has not been finished after the completion of the HGP. The **Genome Reference Consortium** has been founded, whose main task is to map the missing gaps. These are located in difficult-to-sequence regions, usually in repeat-rich regions. At the completion of the HGP about 350 gaps were in the genome. These regions are not small; they represent about 5% of the genome. To fill these gaps are far from easy, which is shown by the fact that 6 years after the initiation of this project, in 2009, only 50 such gaps were completed.

9.6. Variations in the human genome

One of the main tasks in the HGP was to reveal the variations in the human genome. The most frequent variation in the genome is the **single nucleotide polymorphism (SNP)**. In general, or in the original sense, we speak about an SNP if the population frequency of a variation is more than 1%. The less frequent variations are usually called mutations. But in the recent years the definition and the usage of the term SNP has been changed. See the definition in chapter 3.1. Often, most single nucleotide variations are called SNP or recently rather **SNV (single nucleotide variation)**, regardless of their frequency. **Minor allele frequency (MAF)** refers to the frequency at which the less common allele occurs in a given population. If the MAF of an SNP is >5%, than it is called frequent, if between 0.5-5%, it is low, if it is below 0.5%, it is rare.

Relatively, most SNPs are in the intron, then in the intragenic regions, and the rarest are in the exons. At present, about 150 million SNPs and short variants are known (Table 1), but only about 0.1% of these change amino acid codons, and only about 40% of these are non-conservative. At the first glance most SNPs are neutral, but the ENCODE project (see below) revealed that a lot of them are involved in regulation. The majority of **SNPs** found associated with **disease** are **outside the coding region**.

Large structural variations were detected already in the HGP, but at the population level they were regarded as less significant, comparing to the SNPs. Later, as the genomic methods improved, it turned out that there is a huge number of smaller and larger copy variations in the human genome. **Copy-number variations (CNVs)**—a form of structural variation—are alterations of the DNA of a genome that results in the cell having different number of copies of one or more sections of the DNA. CNVs correspond to relatively large regions of the genome that have been deleted (fewer than the normal number) or duplicated (more than the normal number) on certain chromosomes. For example, the chromosome that normally has sections in order as A-B-C-D might instead have sections A-B-C-C-D (a duplication of "C") or A-B-D (a deletion of "C") (see Figure 9.3). This variation accounts for roughly 12% of human genomic DNA and each variation may range from about one kilobase to several megabases in size. Until now, 2900 genes have been found (13% of the genes) effected by the CNVs. This means that the individuals differ from each other in this respect as well, i.e. certain genes can miss in some individuals, or there can be more copies of a gene in some people. Usually, it causes no large phenotypic differences, but there are **several diseases, where CNVs can play a role**, like Crohn's disease, Alzheimer disease, autism, obesity, AIDS, etc.

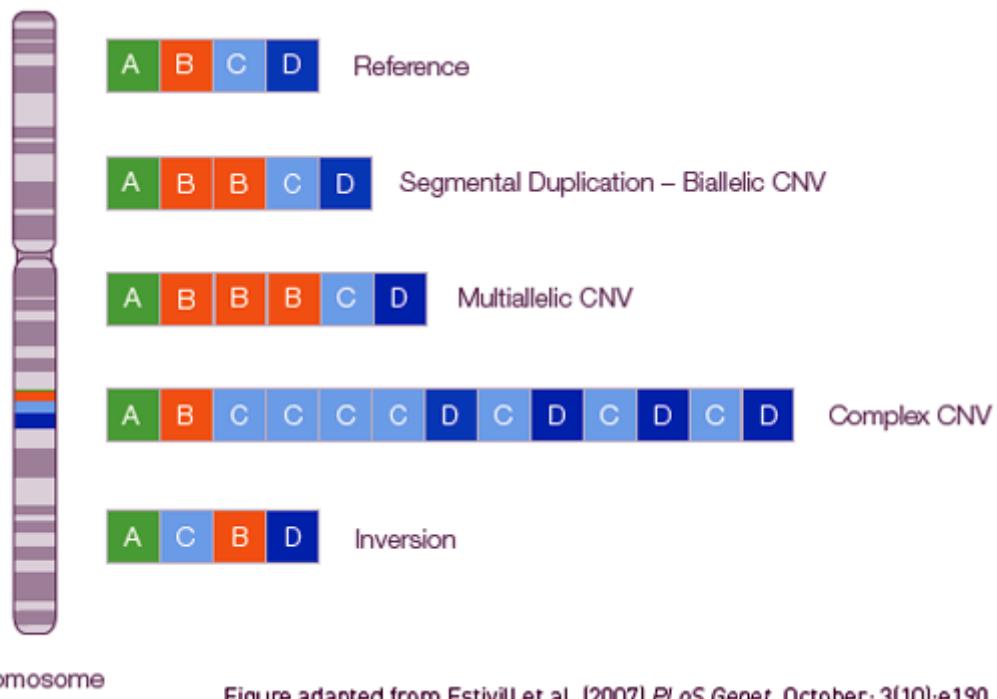


Figure adapted from Estivill et al. (2007) PLoS Genet. October; 3(10):e190.

Figure 9.3. Structural variants in the genome. Source: <http://www.thermofisher.com/hu/en/home/life-science/pcr/real-time-pcr/qpcr-education/what-can-you-do-with-qpcr/genotyping.html> (24/02/2016)

CNVs can play a role in **transplantation**. If in the organ acceptor, owing to a CNV, a gene is missing, and the gene is present in the donor, a **graft-versus-host disease** can develop in spite of MHC identity, i.e. an immune response could develop against the gene product.

Structural differences were also observed in **concordant twins**. This observation questions the long standing notion that monozygotic twins are essentially genetically identical, and also shows that structural variation might also originate during somatic development.

The discovery of the abundance of the CNVs also changed our view of the genomic differences among individuals. In the first two papers about the sequence of the human genome it was stated that the genomic difference between two individuals is 0.1%, i.e. any two persons are in 99.9% identical. At that time it caused a large media coverage. The differences were attributed in these papers mainly to SNPs. Later the diploid sequences of both Craig Venter and James Watson have been published. Analysis of diploid sequences has shown that non-SNP variation, i.e. **CNVs accounts for much more human genetic variation** than single nucleotide diversity. It is estimated that approximately 0.4% of the genomes of unrelated people typically differ with respect to copy number. When copy number variation is included, human to **human genetic variation is estimated to be at least 0.5%** (99.5% similarity). In Table 2 the number of variations can be seen in some sequenced

genomes. But, according to more recent studies, between individuals, separated historically long ago from each other, the difference can be as high as 2-3%. For this difference large genomic rearrangements can be responsible. Populations separated by distance tend to drift apart genetically over time, and roughly 95% of variability between populations is a result of this random drift. For some differences the natural selection is responsible (see Chapter 12).

Number of SNPs (SNVs)		
Genome of J. Craig Venter	3,213,401	
Genome of James Watson	3,322,093	
Asian genome	3,074,097	
Yoruban (African) genome	4,139,196	
Structural variations in Venter's genome		
n	Long (bp)	
CNV	62	8,855–1,925,949
Insertion/deletion	851,575	1–82,711
Block substitution	53,823	2–206
Inversion	90	7–670,345

Table 2.

Number of variations in some sequenced genomes

There was a large change in our view regarding the development of modern human genome. In a *Science* paper published in May 2010, **Scante Pääbo**'s international team found that a small amount—1% to 4%—of the nuclear DNA of Europeans and Asians, but not of Africans, can be traced to **Neanderthals** (<http://www.sciencemag.org/content/328/5979/680.full>). The most likely model to explain this was that early modern humans arose in Africa but **interbred with Neanderthals** in the Middle East or Arabia before spreading into Asia and Europe, about 50,000 to 80,000 years ago. Seven months later, on 23 December, the team published in Nature the complete nuclear genome of a girl's pinky finger from **Denisova Cave** in the Altai Mountains of southern Siberia. To their surprise, the genome was neither a Neanderthal's nor a modern human's, yet the girl was alive at the same time, dating to at least 30,000 years ago

and probably older than 50,000 years. Her DNA was most like a Neanderthal's, but her people were a distinct group that had long been separated from Neanderthals. By comparing parts of the Denisovan genome directly with the same segments of DNA in 53 populations of living people, the team found that the Denisovans shared 4% to 6% of their DNA with Melanesians from Papua New Guinea and the Bougainville Islands. Those segments were not found in Neanderthals or other living humans. The most likely scenario for how all this happened is that after Neanderthal and Denisovan populations split about 200,000 years ago, modern humans interbred with Neanderthals as they left Africa in the past 100,000 years. Thus Neanderthals left their mark in the genomes of living Asians and Europeans. Later, a subset of this group of moderns—who carried some Neanderthal DNA—headed east toward Melanesia and interbred with the Denisovans in Asia on the way. As a result, Melanesians inherited DNA from both Neanderthals and Denisovans, with as much as 8% of their DNA coming from archaic people (http://en.wikipedia.org/wiki/Denisova_hominin).

Later it was shown that archaic people contributed more than half of the alleles that code for proteins made by the human leukocyte antigen system (HLA), which helps the immune system to recognize pathogens. Pääbo's team published the complete genome of the Denisovan cave girl. She didn't carry B*73—and it hasn't been found in Siberia—but she carried two other linked HLA-C variants, which occur on the same stretch of chromosome 6. If living people have any of these variants, they almost always carry at least two of the three variants—as did the cave girl. So even though she lacked B*73, the researchers propose that all three variants were inherited, often in pairs, from archaic humans in Asia. From immunology it is known that the more heterogeneous a population in certain HLA genes is, the more successful it is in defending against pathogen challenges. Thus, it seems that archaic genome contributed to modern human HLA variations and selection fitness through horizontal gene transfer.

After the HGP, several projects contributed to the SNP databases. Such projects were the **HapMap projects** (<http://hapmap.ncbi.nlm.nih.gov/>; http://en.wikipedia.org/wiki/International_HapMap_Project) and the **1000 genome project**. E.g. in the pilot paper of the 1000 genome project 15 million SNPs, 1 million short structural variations (insertion, deletions) and 20 thousand large structural variations were published (<http://www.sciencemag.org/content/330/6004/574>). Most of them were new variations.

The MHC (HLA) region (6p21.3) is quite special, regarding the density of the variations. In this 7.6 Mb long region are located the MHC genes playing important roles in immune response and transplantations. Here, in the MHC class III region can be found the highest gene density (59 expressed genes), and the highest genetic diversity. In a study, in a 4 Mb region 37 thousand SNP and 7 thousand structural variations have been detected, which correspond to a genetic diversity one order of magnitude higher than in other parts of the genome.

9.7. Junk DNA in the human genome

The low number of genes surprised the experts. E.g. it was difficult to explain, how it was possible that the very simple worm *Caenorhabditis elegans* which is hardly 1 mm, contains about the same number of genes as humans. The proportion of the gene coding region in the human genome is a mere 1.2%. Earlier it was thought that the main task of the genome is to store protein coding genes, thus the non-coding region was named junk. But, the majority of experts were sure that the human genome must not contain 98.8% junk. To clarify this contradiction, a project was launched called Encyclopedia of DNA elements (**ENCODE**) (<http://genome.ucsc.edu/ENCODE/>, <http://www.genome.gov/10005107>). The project was initiated with a \$12 million pilot phase to evaluate a variety of different methods for use in later stages. A number of then-existing techniques were used to analyse a portion of the genome equal to about 1% (30 million base-pairs). The results of these analyses were evaluated based on their ability to identify regions of DNA, which were known or suspected to contain functional elements. The pilot phase was successfully finished and the results were published in June 2007 in *Nature* and in a special issue of *Genome Research*.

In September 2007, NHGRI began funding the production phase of the ENCODE project. In this phase, the goal was to analyse the entire genome and to conduct additional studies. The project utilized the huge technical development in genomic methods, like sequencing (NGS). The aim was to determine which regions are transcribed into RNA, which regions are likely to control the genes that are used in a particular type of cell, and which regions are associated with a wide variety of proteins. The primary assays used in ENCODE are ChIP-seq, DNase I Hypersensitivity, RNA-seq, and assays of DNA methylation (see methods in Chapter 11).

In September 2012, the project released results in 30 papers published simultaneously in several journals, including *Nature* (<http://www.nature.com/encode/#/threads>). The most striking finding was that the fraction of human DNA that is biologically active is considerably higher than even the most optimistic previous estimates. In an overview paper, the ENCODE

Consortium reported that its members were able to assign biochemical functions to over 80% of the genome. Later, however, it was argued the notion that the majority of eukaryotic noncoding DNA is functional is very difficult to reconcile with the massive diversity in genome size observed among species, including among some closely related taxa. E.g. the size of the onion genome is 16 Gbp. The so-called **onion test** simply asks: if most eukaryotic DNA is functional at the organism level, be it for gene regulation, protection against mutations, maintenance of chromosome structure, or any other such role, then why does an onion require five times more of it than a human? Several analyses of sequence conservation between humans and other mammals have found that about 5% of the genome is conserved and an additional 4% of the human genome is under lineage-specific selection pressure. Altogether, **9% of the human genome shows signs of functionality** is actually consistent with the results of ENCODE and other large-scale genome analyses.

In short, the following functional elements could be distinguished:

- Protein coding regions
- RNA coding regions (not transcribed to protein)
- Promoter regions (binding transcription factors; more than 70,000)
- Enhancer regions (distant regulatory regions, about 400,000)
- Long range chromatin interactions
- DNA methylation sites
- Histone modification

In the 30 papers there are a lot of interesting results, let us see some of them:

It was found that about **75% of the genome is transcribed at some point in some cells**, and that genes are highly interlaced with **overlapping transcripts that are synthesized from both DNA strands**. These findings force a rethink of the definition of a gene and of the minimum unit of heredity.

Some studies were based on the **DNase I hypersensitivity (DHS) assay**. DHSs are genomic regions that are accessible to enzymatic cleavage as a result of the displacement of nucleosomes (the basic units of chromatin) by DNA-binding proteins. The authors identified **cell-specific patterns of DNase I hypersensitive sites** that show remarkable concordance with experimentally determined and computationally predicted binding sites of transcription factors. DNA binding of a few high-affinity transcription factors displaces nucleosomes and creates a DHS, which in turn facilitates the binding of further, lower-affinity factors. The

results also support the idea that transcription-factor binding can block DNA methylation, rather than the other way around — which is highly relevant to the interpretation of disease-associated sites of altered DNA methylation.

Beyond the linear organization of genes and transcripts on chromosomes lies a more complex (and still poorly understood) network of **chromosome loops and twists** through which promoters and more distal elements, such as enhancers, **can communicate their regulatory information** to each other.

ENCODE defined 8800 small **RNA molecules** and 9600 long noncoding RNA molecules, each of which is at least 200 bases long. It was found that various ones **home in on different cell compartments**, as if they have fixed addresses where they operate. Some go to the nucleus, some to the nucleolus, and some to the cytoplasm, for example.

Earlier investigations aimed for finding genomic background of complex diseases or traits (e.g. height) found that the majority (~93%) of disease- and trait-associated variants lay within noncoding sequence, complicating their functional evaluation. The map created by ENCODE reveals that many of these **disease-linked regions include enhancers or other functional sequences**. And **cell type is important**. In one study, several variants were studied that were earlier found strongly associated with systemic lupus erythematosus, a disease in which the immune system attacks the body's own tissues. It was noticed that the variants identified in genomic studies tended to be in regulatory regions of the genome that were active in an immune-cell line, but not necessarily in other types of cell.

In one cell line (K562) 127,417 promoter-centered chromatin interactions were detected, 98% of which were intra-chromosomal (which is called **gene kissing**). Multi-genic interactions could be detected in 90% of the genes including several promoter-promoter and promoter-enhancer interactions. Most of these interactions are cell-specific. It means that the majority of **functional elements are cell-specific**, they may be functional in one cell, and have no function in others.

Scientists could eventually accumulate enough data **to predict the function of unexplored sequences**. This process, called **imputation**, has long been a goal for genome annotation. According to some opinions there may be a phase transition where imputation is going to be more powerful and more accurate than actually doing the experiments.

Another study showed that **part of the genome is conserved to maintain its 3D structure**. Mutation in these could result in diseases.

According to these findings the majority of the genome contains logistic (regulatory) information, which contradicts to previous views that the main task of the genome is for

coding building blocks (proteins). It can also explain why the evolution of multicellular organisms lasted for 3.5 billion years. It is namely known that the life on Earth developed 4 billion years ago, but multi-cellular organisms existed only in the last 500 million years. Until then only single-cellular organisms existed.

The other explanation which could resolve the low gene number - complex organism contradiction is that about 94% of the genes code for multiple proteins. Through **alternative splicing**, several proteins with different size and functions can be transcribed from one gene in the different tissues. In simpler organisms the extent of this is more limited, or does not exist at all. In addition, there are several possibilities for **post-transcriptional modifications** of the proteins. According to the estimations the **number of proteins** in humans can be as high as **2 million**.

9.8. Comparative genomics

According to the definition, **comparative genomics** is the study of the relationship of genome structure and function across different biological species or strains (http://en.wikipedia.org/wiki/Comparative_genomics). It searches answers for such questions, like which are the **human specific genes** (e.g. comparing the human and chimpanzee genomes); which genes are **essential for life** (conserved genes in all organism); for multi-cellular organisms (single-cellular vs. multi-cellular), for the eukaryotic evolution (eukaryotic vs. prokaryotic), for the mammals (e.g. mouse vs. drosophila), etc.

Comparative genomics exploits both **similarities and differences** in the genomic sequences of different organisms to infer how selection has acted upon the sequences. Those sequences that are responsible for similarities between different species should be **conserved** through time (**stabilizing selection**), while those responsible for differences among species should be divergent (**positive selection**). Those sequences that are unimportant to the evolutionary success of the organism will be **unconserved (selection is neutral)**.

Comparative genomic studies suggest that **owing to mammalian conservation ~5% of the human genome is conserved** due to non-coding and regulatory roles. According to a recent study combining population genomic information from the 1000 Genomes Project and biochemical data of the ENCODE Project, **an additional 4% of the human genome** is subject to **lineage-specific constraint**. Regulatory elements under human constraint in non-conserved regions were found near **color vision and nerve-growth genes**, consistent with **purifying selection** for recently evolved functions (<http://www.sciencemag.org/content/337/6102/1675.full>).

Comparing the genome of mouse and human, it turned out that 99% of the protein coding genes in the mouse have human homologs. It means that at gene level, the two species differ from each other in about 300 genes. Thus, the mouse is suitable for model organism, for investigation of gene functions, for disease models in genetic diseases, etc.

The closest relative of the human species is the chimpanzee. Applying molecular dating it is known that human and chimpanzee speciation occurred less than 6.3 million years ago. But, because of the reduced amount of variation on the X chromosome, humans and chimpanzees were still exchanging X chromosomes 1.2 million years after the species split. The similarities of **human and chimpanzee** protein sequences are remarkable. About 50,000 amino acid differences separate us and chimpanzees. The identification of non-alignable sequences in the two genomes that were due to small- and large-scale segmental deletions and duplications, showed that the **overall difference between the two genomes is actually ~4%**.

In contrast, the common chimpanzee (*Pan troglodytes*) and human **Y chromosomes** are very **different** from each other. Many of the differences between the chimpanzee and human Y chromosomes are due to gene loss in the chimpanzee and gene gain in the human. It was found that the chimp Y chromosome has only two-thirds as many distinct genes or gene families as the human Y chromosome, and only 47% as many protein-coding elements as humans.

Another interesting finding was that difference was found in the conserved ***FOXP2*** gene between humans and chimpanzee. Because in humans mutations in this gene cause a severe speech and language disorder, this was named (wrongly) the **language gene** (http://news.nationalgeographic.com/news/2001/10/1004_TVlanguagegene.html). In contrast, Pääbo found that the *FOXP2* gene is the same in modern humans and in Neanderthals, raising the possibility that the Neanderthals could speak.

Comparison of human and Neanderthal genome indicated that there are only 1000 to 2000 amino acid differences between the two species. The researchers found 78 protein-altering sequence changes that seem to have arisen since the divergence from Neanderthals several hundred thousand years ago, plus a handful of other genomic regions that show **signs of positive selection in modern humans**. These are linked to sperm motility, wound healing, skin function, genetic transcription control and cognitive development.

9.9. Literature

1. http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml 2009.
2. International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. *Nature* 2001;409:860-921.
3. Venter JC et al. The sequence of the Human Genome. *Science* 2001;291:1304-51.
4. International Human Genome Sequencing Consortium: Finishing the euchromatic sequence of the human genome [Nature 431, 931 - 945 \(21 October 2004\)](http://www.nature.com/nature/journal/v431/n6091/full/nature03190.html)
5. <http://genomics.xprize.org/>
6. Rusk N, Kiermer V. Primer: Sequencing—the next generation. *Nature Methods* 2008;5:15.
7. <http://www.genome.gov/10005107>; 2009.
8. Pennisi E. 1000 Genomes Project Gives New Map Of Genetic Diversity. *Science* 2010; 330: 574-5.)
9. <http://www.epigenome.org/>; 2009.
10. Redon R. et al.: Global variation in copy number in the human genome. *Nature* 2006; 444: 444-454.
11. Armour JA. Copy number variation and antigenic repertoire. *Nat Genet*. 2009;41(12):1263-4.
12. Bruder CE et al.: Phenotypically concordant and discordant monozygotic twins display different DNA copy-number-variation profiles. *Am J Hum Genet*. 2008;82:763-71.
13. Ng PC et al. Genetic variation in an individual human exome. *PLoS Genet*. 2008 Aug 15;4(8):e1000160.
14. Reich D et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*. 2010 Dec 23;468(7327):1053-60.
15. Green RE et al. A draft sequence of the Neandertal genome. *Science*. 2010 May 7;328(5979):710-22.
16. Reich D et al. Denisova admixture and the first modern human dispersals into southeast Asia and oceania. *Am J Hum Genet*. 2011 Oct 7;89(4):516-28.
17. Burbano HA et al. Targeted investigation of the Neandertal genome by array-based sequence capture. *Science*. 2010 May 7;328(5979):723-5.
18. Gibbs W.W. (2003) "The unseen genome: gems among the junk", [Scientific American](http://www.scientificamerican.com/article.cfm?articleID=0001940883750400), 289(5): 46-53.

19. The ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007; 447:799-816.
20. Parker SC, Hansen L, Abaan HO, Tullius TD, Margulies EH. Local DNA Topography Correlates with Functional Noncoding Regions of the Human Genome. *Science*. 2009; 324: 389 – 392.
21. Fire A, Xu S, Montgomery M, Kostas S, Driver S, Mello C (1998). "Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*". *Nature* 391 (6669): 806–11.
22. Swami M. RNA world: A new class of small RNAs *Nature Reviews Genetics* 2009;10, 425.
23. Waterston RH. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002; 420 (6915) 520 - 562.
24. Kirkness EF et al. The Dog Genome: Survey Sequencing and Comparative Analysis. *Science*. 2003; 301:1898-1903
25. Krause J et al. The Derived FOXP2 Variant of Modern Humans Was Shared with Neandertals. *Current Biology* 2007; 17: 1908-1912
26. ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012 Sep 6;489(7414):57-74.

9.10. Questions

1. What is genome?
2. What is genomics?
3. What is the difference between genetics and genomics?
4. When did the Human Genome Project start?
5. Which organizations were involved in the initialization of the HGP?
6. Give some examples about the main aims of the HGP!
7. What was the name of the sequencing method which was proposed by Craig Venter, and what was the name of the company funded by him?
8. What is the Archon Genomics X PRIZE?
9. Give some examples for larger genome projects!
10. What is the size of the human genome?
11. Approximately, how many protein coding genes are in the human genome?
12. How many percent is the protein coding part of the human genome?
13. Which one is the largest human gene?

14. Which gene has the longest coding sequence?
15. Which chromosome is the gene richest?
16. Which is the gene poorest chromosome?
17. When are the most inherited mutations produced?
18. What are the CpG islands?
19. What is the genetic imprinting?
20. What is the name for the most frequent repeats?
21. How many percent is the repeat part of the human genome?
22. On average, how many genes with loss-of-function mutations does an individual carry?
23. What is the SNP?
24. What is the MAF?
25. What do Copy Number Variations mean?
26. What is the CNP?
27. On average, what is the difference on genomic level between two individuals?
28. Did the homo sapiens interbred with other species or subspecies?
29. Which is the most variable part of the human genome?
30. What are the pseudogenes and what can be the role of them?
31. What are the paralogs?
32. Relatively, where are the SNPs the most frequent?
33. What is the junk DNA, and what is the importance of it?
34. What is the aim of the ENCODE project?
35. Give some examples for the results of the ENCODE projects!
36. What results did the DNase I hypersensitivity (DHS) assays in the ENCODE project give?
37. What is the connection between the transcription factors and the DNA methylation?
38. What is the gene kissing?
39. What can be the roles for the RNA in the genome?
40. What is the comparative genomics?
41. How many percent of the human genome is conserved?
42. What does the comparison of the human genome with the Neanderthal and chimpanzee genome show?

10. Genomic approach to complex diseases

Csaba Szalai

10.1. General features of the complex diseases

The complex or multifactorial diseases are those which **develop through interactions** of a few (oligogenic) or several (polygenic) genes and the environmental factors. The complex diseases, in contrast to the monogenic diseases, which affect only a small fraction of the population, are often very frequent and more or less everybody is affected by them. Complex diseases are the **endemic, non-communicable diseases**, or **NCD**, which are non-infectious and non-transmissible between persons, like cancer, asthma, hypertension, diabetes mellitus, cardiovascular diseases, Alzheimer disease, etc. It must be added, however, that the different traits (inner or outer properties, like e.g. height, eye color or predisposition to drug abuse) are also multifactorial, and the methods for revealing their genomic backgrounds are the same as those for diseases, and the two even can be connected, thus there is no sharp distinction between them.

Some features of the complex diseases:

- They are often very **frequent** (vs. monogenic diseases). E.g. the prevalence of asthma is between 6-10% in different countries, and there are ~300 million asthmatics in the world. Cardiovascular diseases are responsible for 39% of all death.
- **Familial aggregation** but no Mendelian inheritance can be observed. It means that certain diseases in certain families occur more frequently than in the general population, but by studying the family tree no Mendelian inheritance can be detected (dominant, recessive, etc.).
- They are usually **more frequent in post-reproductive age**. It means that in contrast to the majority of monogenic diseases, which develop in early ages, the complex diseases develop in age, when the affected individuals already have children, thus the genetic variations that caused the susceptibility to the disease could be passed over to the next generation.
- They have huge **economic significance**. E.g. European Union spends for the cardiovascular diseases €170 billion/year, the USA \$300 billion/year; for asthma \$18 billion/year, etc.

- The **prevalence** of a number of complex diseases has been **increasing** in the last decades (e.g. allergy, obesity, asthma, type 2 diabetes mellitus (T2DM), melanoma, etc.).
- Often **comorbidity** can be observed (see Chapter 15, Systems biology).

It must be added that from certain point of view all the diseases can be considered multifactorial. E.g. mono-causative conditions like infectious diseases can also be complex. The genomic background and additional environmental factors influence whether an individual becomes ill in an infectious environment or what symptoms he/she will develop. Even the symptoms of monogenic conditions are influenced by genomic and environmental factors.

10.2. Environmental factors

Environmental factors have decisive roles in the development of complex diseases. In most cases, the genomic background only makes susceptible to certain diseases, and the environmental factors are needed for the manifestation. Environmental factors are everything which is not genetic. Some examples: food, smoking, intrauterine factors (through epigenetic mechanisms their effects can last for the whole life), infections, the way of life, upbringing, weather, etc.

10.3. Why is it important to study the genomic background of the complex diseases?

- Perhaps the most important is that it helps to **explore the molecular pathomechanism**. In contrast to the traditional methods, the genomic methods are often hypothesis free, i.e. they do not require any knowledge about the pathogenesis. In this way novel pathways and mechanisms can be detected, which can offer new drug targets or new therapies.
- The genomic studies can reveal the **genetic differences between people**, offering novel possibilities for personal therapies, and connections can be found between the success of the therapy and the genetic background (see Pharmacogenomics, Chapter 14).
- **Genetic risks** to the different diseases can be detected. In this way, right after the birth the genomic background and the risk to different diseases of a new-born can be

determined, which offers the possibilities to change from “diagnose and treat” to “predict and prevent”. Earlier it was regarded as the most important task of the medical genomics, but later it turned out that in most cases the sum risk to a multifactorial disease is so complex that it is usually impossible to give a clinically relevant estimation.

It must be noted, however, that the genomic results in medicine infiltrate only very slowly to the clinical practice, and many things are different than was expected in the 90s. Besides, genomics obeys the **First Law of Technology**: we invariably overestimate the short-term impacts of new technologies and underestimate their longer-term effects. Here are some reasons that can be the cause of the above mentioned problems:

1. The individuals are **genetically too heterogeneous**, which makes the personal therapy very difficult, although there are some positive examples, especially in cancer therapy.
2. Even if an increased risk to a certain disease is recognized, which can be prevented by changing of lifestyle, people usually do not incline to change very easily, e.g. everybody knows the increased risk associated with smoking, alcoholism, drugs, or sedentary lifestyle, but most people do not care about it.
3. The connection between genomic background and phenotype is much more complex than previously thought (see later).

10.4. Heritability of the complex diseases

If we consider that the prevalence of many diseases has increased in the last decades, although the genomic background of the population surely did not change dramatically, how can we prove that a disease has a heritable fraction?

The simplest method is the calculation of the **λ_R value**. Here R is the abbreviation of the “relative”. If siblings are studied then “s” is for siblings (λ_s). For the calculation of this value the family aggregation of a disease is compared to the population frequency. E.g. if the concordance of a disease in twin pairs is 0.8, and the population frequency of this disease is 0.2, then the λ_s is $0.8/0.2 = 4$. From these two things can be seen: (1) If $\lambda_s = 1$, then the disease has no genetic background, i.e. the risk to it is not heritable and if it is higher than 1 the disease may have a heritable component; (2) diseases whose population prevalence is high (common diseases), the λ_s cannot be high. If the denominator is low and the counter is high like in the cases of the monogenic diseases, the λ_s is very high. E.g. for the monogenic cystic

fibrosis $\lambda_s = 500$, while for the multifactorial and frequent type 2 diabetes mellitus (T2DM) $\lambda_s = 3.5$, for asthma $\lambda_s = 2$, for the less frequent schizophrenia $\lambda_s = 8.6$ and for type 1 diabetes mellitus (T1DM) $\lambda_s = 15$.

What are the possible **problems with the λ values**, which cause that it is not as easy to calculate, as it should be? First of all, siblings, especially twins grow up in the **same environment**, eat similar food, are exposed to similar effects (infections, weather, psychical effects), all of which can cause biases, because if they have the same disease, it can be that not the genetic, but the common environment (e.g. virus, or diet) was the cause of the disease (phenocopy, Table 1). This can be corrected if twins who were grown up separately are involved in the studies. The problem with the latter is that it is much more difficult to recruit such twins, and in addition, the very strong (mainly common) intrauterine effects cannot be excluded. Several studies show namely that perhaps the intrauterine effects are the strongest environmental factor. It is well-known that preterm labor makes susceptible to several adult diseases (e.g. T2DM), the high weight at birth to T1DM, and it was also shown that IQ is strongly influenced by intrauterine effects (smoking, alcohol or diet). But, λ_R is still used and can give important information about the genetic fraction of the disease. In addition, after the World War II, when a lot of twins remained alone in Denmark, it was a general practice there that the twin pairs were adopted separately, thus it was possible later to recruit such population for genetic studies.

Another good solution is **twin studies** where the concordance rate of monozygotic twins is compared to that of fraternal twins. Because monozygotic twins are genetically identical, and fraternal twins are not, but the environmental factors, especially in early childhood are as similar as possible, it is a possible method to estimate the genetic and environmental contributions to a trait.

10.5. Calculating heritability

The **heritability** of a trait within a population is the proportion of observable differences in a trait between individuals within a population that is due to genetic differences. Factors including genetics, environment and random chance can all contribute to the variation between individuals in their observable characteristics (in their "phenotypes"). Heritability can change without any genetic change occurring (e.g. when the environment starts contributing to more variation).

Any particular phenotype can be modeled as the sum of genetic and environmental effects:

Phenotype (P) = Genotype (G) + Environment (E).

Likewise the variance in the trait – $\text{Var}(P)$ – is the sum of genetic effects as follows:

$$\text{Var}(P) = \text{Var}(G) + \text{Var}(E) + 2 \text{Cov}(G,E).$$

In a planned experiment $\text{Cov}(G,E)$ can be controlled and held at 0. In this case, heritability is defined as:

$$H^2 = \frac{\text{Var}(G)}{\text{Var}(P)}$$

H^2 is the broad-sense heritability. See for more details here: <https://en.wikipedia.org/wiki/Heritability>

In the scientific literature heritability is often given **in percent**. E.g. the heritability of height is 80%. Heritability cannot be interpreted at an individual level; it is specific to a particular population in a particular environment.

A prerequisite for heritability analyses is that there is some population variation to account for. In practice, all traits vary and almost all traits show some heritability. For example, in a population with no diversity in hair color, "heritability" of hair color would be undefined ($\text{Var}(P) = 0$). In populations with varying values of a trait, variance could be due to environment (hair dye for instance) or genetic differences, and heritability could vary from 0-100%.

This last point highlights the fact that heritability cannot take into account the effect of factors which are invariant in the population. Factors may be invariant if they are absent and don't exist in the population (e.g. no one has access to a particular antibiotic), or because they are omni-present (e.g. if everyone is drinking coffee).

10.6. Difficulties in the studies of the genomic background of complex diseases

At the beginning of the genomic era, even right after the completion of the HGP, it was generally thought that genomic would revolutionize the medicine, and in a few years the era of personal therapy would come. But now, in 2016 we know that there are only some scattered examples of the personal therapy, and it is not expected to be widespread even in the next years. What can be the reason for this?

According to the general opinion, one of the main reasons for this failure is due to the very complex regulation of the genome (see Chapter 9), and the multifactorial nature of the diseases and traits. Figure 10.1 shows that if a QT is determined by genetic factors, what population distribution is expected for this QT. First of all, let us define the QT and some related terms:

- **QT means quantitative trait**, which is a value that can be quantified by numbers, like IgE or LDL-C levels, height or IQ. QT can vary in degree and can be attributed to polygenic effects, i.e., product of two or more genes, and their environment.
- **Discontinuous trait** is either/or traits that do not have any range. E.g. cleft lip. It can also be used in QT studies.
- **QTL or quantitative trait loci** are stretches of DNA containing or linked to the genes that underlie a quantitative trait, or in family studies the loci segregate together with the QT. Such loci e.g. are those which segregate together with elevated cholesterol or fasting insulin level.

Let us go back to Figure 10.1. If a locus has two alleles with equal frequencies, one of which reduces the value of the trait, the other increases it, then, as depicted in Figure 10.1A, in respect of the QT, the population can be divided into three groups. In Figure 10.1B and C those cases can be seen when there are two or three loci influencing the QT. In case of 3 loci there can be as much as 7 different genotypes associated with a QT value in the population. In cases of multifactorial traits, the QT is usually influenced by several hundreds of loci, plus the environmental factors. In these cases the distribution will be continuous, and if we determine the QT, there is a huge number of possible genotype and environmental factor combinations which can be responsible for the given value. It means that the determination of the QT will give very little information regarding the genotype.

In Table 1 there are some characteristics which make the determination of the genetic background of the multifactorial diseases difficult.

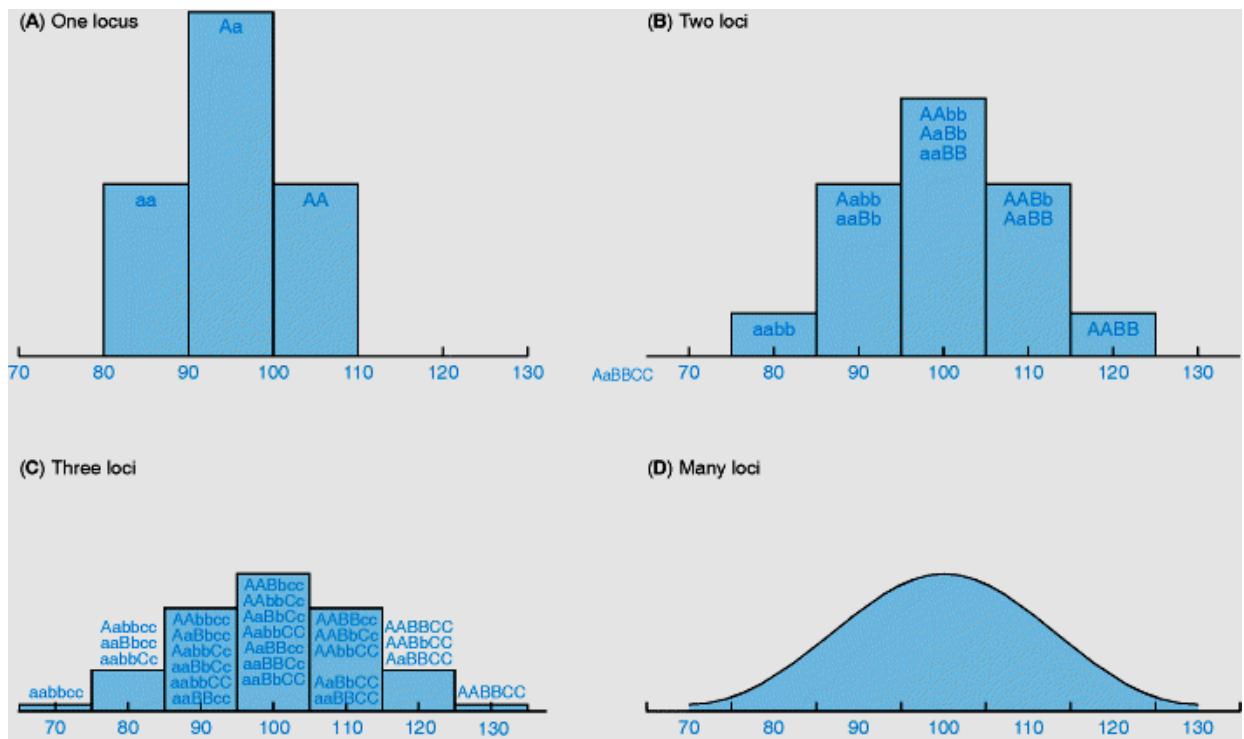


Figure 10.1.

- A. If a locus has two alleles with equal frequency, one of which reduces the value of the trait, the other increases it, then, in respect of the QT the population can be divided into three groups.
- B, C. There are two and three loci influencing the QT, respectively. In Figure 10.1B, in case of 3 loci there are already 7 genotypes associated with the QT value of.
- D. In cases of multifactorial traits, the QT is usually influenced by several hundreds of loci, plus the environmental factors. In these cases the distribution will be continuous.

Source:

<http://www.ncbi.nlm.nih.gov/books/NBK7564/>;

<http://www.ncbi.nlm.nih.gov/books/NBK7564/figure/A2478/?report=objectonly>; 31/05/2013)

Problems	Explanation
Genetic heterogeneity	Different allelic combinations lead to similar phenotypes.
Phenocopy	Environmental factors lead to the same clinical phenotype as do the genetic factors. In other words, the environmental condition mimics the phenotype produced by a gene.
Pleiotropy	The same genetic variations or genotype combinations can lead to different phenotypes.
Incomplete penetrance	Some individuals fail to express the trait, even though they carry the trait associated alleles.
The exact diagnosis is difficult	Often in complex diseases there are no standard diagnoses. There are subtypes of the diseases that cannot be differentiated with standard methods. The symptoms can change with the time, or manifest in episodes. Different diseases with similar symptoms. Concordance of different diseases.

Table 1.

Factors, which make the determination of the genetic backgrounds of the complex diseases difficult

10.7. Development of genomic methods, problems

As for both researchers and the whole society the significance of genomic results are widespread appreciated, this has led to a large-scale effort for the development of genomic methods and huge breakthroughs have been achieved (see Chapter 11).

But there is no reason for the total satisfaction, since most of the aims have not been achieved. In 2009, Manolio et al. published a widespread cited table in a paper (Figure 10.2 was prepared according to the table in the paper), which summarizes the results of studies aiming at determining the genomic background of multifactorial diseases and traits (<http://www.wired.com/wiredscience/2009/10/beyond-the-genome/>). These results show that the GWAS (genome

wide association studies, see Chapter 11), which were thought to be the very method for determining the genomic background of complex traits, could determine only a small fraction of the heritability proportion of the majority of the traits. It means that most variants identified until then conferred relatively small increments in risk, and explained only a small proportion of familial clustering, leading many to question how the remaining, '**missing**' heritability can be explained. And the situation has not improved considerably since then. E.g. height is one of the QTs which is easy to determine, and it is known that the heritability of it is about 80%. In several studies, large populations were collected and several GWAS were carried out. In one study, 44 loci were determined, which were responsible only for 5% of the heritability. Later, 180 loci could be determined, but they were still responsible only for 10% of the heritability. This is true for the majority of the complex diseases. E.g. this value for T2DM is 6%, for fasting glucose level is 1.5%, for early myocardial infarction is 2.8%. The exceptions are diseases, where there are only a couple of mutations with strong impact, like in the case of macular degeneration. In contrast, the determination of the genetic background of monogenic diseases is a great success; it has been clarified for about 4000 such diseases so far.

What can be the reason for this situation, which is often called the **dark matter of heritability**? Previously, some explanations have already been mentioned and below some additional ones will be given.

The Case of the Missing Heritability

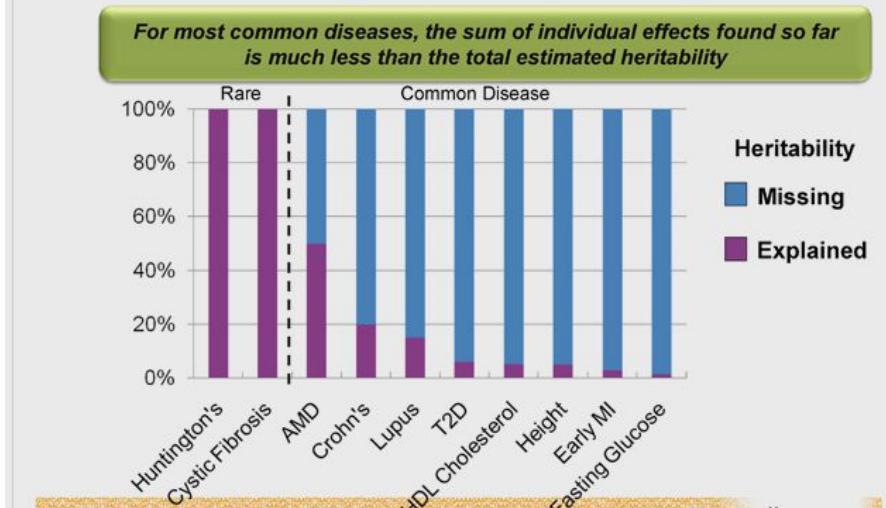


Figure 10.2. Adapted from: Manolio TA et al. Finding the missing heritability of complex diseases. *Nature*. 2009 Oct 8;461(7265):747-53 (<http://www.ncbi.nlm.nih.gov/pubmed/19812666>).

10.8. Problems of rare variants

GWAS work with pre-made chips, which could determine known variations with a population frequency of **>5%** (MAF = minor allele frequency). There is a theory named **common disease - common variants or CD/CV**, which says that common diseases are caused by several common (frequent) variants with weak effects. The weak effects of these variants are accumulated causing higher susceptibility to a disease. If the environmental factors are unfavourable, then the disease can develop. It proved to be true for a lot of traits, like Alzheimer disease, where the roles of the common apoE4 variants or the obesity where the roles of variations in the FTO genes were verified. But, there are also proofs for the so-called **common disease rare variants hypothesis (CD/RV)**, which states that the common diseases are caused by rare variants with strong effects. Example is the breast cancer where thousands of rare variants with strong effects have been found. The rare variants cannot be determined with GWAS, and the traditional statistical methods are not suitable for their detection. It is suggested that even in diseases, where common variations are known, there are also rare variations with strong effect.

The rare variants can also cause another statistical problem called **synthetic associations**. In this case rare variants at the locus create multiple independent association signals captured by

common tagging SNPs causing that variants, which do not participate in the given phenotype, will be falsely named.

10.9. Epigenetics of the complex disease

One of the most significant discoveries of the last decades was that the modifications (e.g. methylation) of the nucleotides can be inherited as well. As it is detailed in Chapter 5, this is studied by **epigenetics**.

Several lines of evidence prove that one of the main reasons for the increasing prevalence of the complex diseases in the developed countries is the effect of environmental factors associated with modern lifestyle. And it is through epigenetic marks that environmental factors like diet, stress and prenatal nutrition can make an imprint on genes that is passed from one generation to the next.

Unfortunately, the traditional and relatively simple genomic methods, like GWAS are not capable of studying the changes of epigenetic factors. With the development of NGS, however, it is becoming easier to study epigenetics as well (see Chapter 11), but as epigenetic marks are different from tissue to tissue (even from cell to cell), and can change throughout the whole life, this will never be an easy task, and it is very difficult to interpret these results. In addition, the epigenetic effects must be interpreted together with the other levels of genomic regulations, making systems biological methods necessary (Chapter 15).

10.10. The random behavior of the genome

In September 2010 researchers published in Nature that genetic circuits that regulate cellular functions are subject to **stochastic fluctuations, or ‘noise’**, in the levels of their components (<http://nicorg.pbworks.com/w/file/fetch/53677630/Functional%20Roles%20of%20Noise%20in%20Genetic%20Circuits.pdf>). It means that the behavior of the genome is sometimes random and thus **cannot be predicted in 100%**. It means that it is theoretically impossible even with more developed genomic and informatic methods to exactly forecast the future traits (phenotypes) of a newborn.

10.11. Statistical problems

The next problem originates from the evaluation methods, i.e. from the statistics. The most variations associated with increased risk to complex diseases, increase the risk with only $\leq 10\%$. It means that the chance in the carriers for the development of the disease is only ≤ 1.1 times higher than in non-carriers. Detecting variations with such weak effects is very difficult. In addition, as the **population is genetically heterogeneous**, and interactions between these

variants are needed, the possible number of genetic backgrounds associated with increased risk is practically infinite. In statistical point of view it is advantageous if the population is larger, but the larger population is genetically more heterogeneous, thus the effect of each genetic variant is diluted, becoming less significant and may be lost.

Another problem is called the **multiple testing problem**.

If in a GWAS 100 thousand genetic variations are measured, in a statistical point of view it means that 100 thousand independent measurements are carried out. In this case the probabilities of the false results are summed up. In statistics, $p < 0.05$ is used as a significance threshold. It means that the probability of the false statement is 5% (we can make a false statement 5 times in 100 independent investigations). One of the methods to correct this is called **Bonferroni correction** (see http://en.wikipedia.org/wiki/Bonferroni_correction). In this case, 0.05 is divided by the number of the measurements (in this case with 100 thousand; $p = 0.05/100.000 = 5 \times 10^{-7}$). But the number of the independent investigations depends not only on the number of the measurements, but on several other factors, like the number of the samples, the clinical parameters and the type of tests, etc. But the Bonferroni correction is too conservative, i.e. if the correction is applied, only the strongest effects can be detected. In contrast, according to the CD/CV hypothesis the complex diseases develop through interactions between multiple genetic variants with weak effects and the environment. In addition, as the genetic factors interact with each other, if we want to calculate this interaction as well, it would increase the number of independent questions to a very large number. It means that the Bonferroni corrections and similar other methods are not capable of detecting the variants of weak effects, i.e. other methods are needed.

10.12. Possible solutions

There are several developments which try to cope with the above mentioned problems. E.g. utilizing the results of the 1000 Genome Project, new chips are under development, which can measure rarer ($MAF < 0.05$) variants as well (e.g. Illumina 5M chip). Furthermore, next to genotyping based methods, the new generation sequencing (NGS) may be soon suitable for population based studies. With the NGS, **all type of variations can be detected**. It must be added, however, that **the statistical problems are even larger** with this method, since it can give terabit size of data and millions of variations, many of which can be sequencing mistakes, or unknown variations whose functional characterizations are immensely difficult.

As the sequencing methods are developing, it is easier to get **epigenetic data**, evaluation of which, however, is also challenging (see above).

There are a couple of **new solutions for the statistical problems** as well. E.g. to overcome several of the limitations, **probabilistic graphical models** (PGMs) were proposed. Thanks to their ability to efficiently and accurately represent complex networks, PGMs represent powerful tools to dissect the genetic susceptibility of complex diseases. **Bayesian networks** are a popular class of PGMs, its graphical representation presents a crucial advantage and is able to efficiently deal with SNP–SNP interactions impacting the phenotype, a situation that is called **epistasis**. As Bayes statistics can evaluate networks, it is a suitable evaluation method for systems biology (<http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0033573>).

It is assumed that with better statistics significantly **more information can be extracted** even from the present results. E.g. in a paper it has been stated that from the old results but with better statistics they could explain 67% of the heritability of height, in contrast to the 5% in the original paper. In this paper rather than considering SNPs one by one, the new statistical analysis considers what effect all the SNPs together have on height (<http://www.nature.com/news/2010/100620/full/465998a.html>).

In another paper the genetic background of hypertension was studied. They reevaluated the results of a meta-analysis of several GWAS, which did not find any associated variants (owing to the too conservative Bonferroni correction, and the heterogeneous nature of this disease). In the new statistics the authors did not consider individual SNPs, but examined whether there are **pathways where the distribution of the variations are statistically different** in the hypertensive population relative to the controls. In this paper several pathways were found associated with the disease.

It is also a great challenge that the majority (~93%) of disease- and trait-associated variants emerging from these studies **lie within non-coding sequence**. It is therefore very difficult to explain how these variants influence the trait. In a study of the **ENCODE** (<http://www.nature.com/nature/journal/v489/n7414/full/nature11247.html>) project it was found that in a given cell line, **76.6% of all non-coding GWAS SNPs** either **lie within a DNase I hypersensitive site** (DHS) (57.1% or 2931 SNPs), or are in complete linkage disequilibrium (LD) with SNPs in a nearby DHS (see Chapter 9 for DHS and 11 for LD). DHSs show remarkable concordance with experimentally determined and computationally predicted binding sites of transcription factors and enhancers. With the help of the results of the ENCODE and similar other projects it will be much easier to determine the function of a variant lying in non-coding region of the genome.

10.13. Why are the complex diseases more frequent in our days?

It is well-known that a lot of complex diseases are significantly more frequent in the last decades, especially in the developed countries. They are often called diseases of civilization. These are diseases that appear to increase in frequency as countries become more industrialized.

For this, one of the best examples is obesity. It is associated with several other diseases, like hypertension, cardiovascular diseases and T2DM. On the web page of <http://www.cdc.gov/obesity/data/trends.html> we can follow how in the USA the prevalence of obesity has increased in the last decades. And the tendency is similar in the allergic and other diseases as well.

What can be the reasons for these trends? The genomic background of the humans cannot be changed in these years! It means that the environment has changed in a very disadvantageous manner in this respect. As the complex diseases develop through interaction between genetic and environmental factors, it is necessary to know the environmental factors, which interact with the genome.

Below some popular hypotheses are described trying to explain these trends.

10.13.1. *Thrifty gene hypothesis*

The **thrifty gene hypothesis** was proposed by geneticist James V. Neel in 1962 to resolve some of the above mentioned problems. Thrifty genes are genes which enable individuals to efficiently collect and process food to deposit fat during periods of food abundance. According to the hypothesis, the 'thrifty' genotype would have been advantageous for hunter-gatherer populations, especially child-bearing women, because it would allow them to fatten more quickly during times of abundance. Fatter individuals carrying the thrifty genes would thus better survive times of food scarcity. However, in modern societies with a constant abundance of food, this genotype efficiently prepares individuals for a famine that never comes. The result is widespread chronic obesity and related health problems like diabetes (http://en.wikipedia.org/wiki/Thrifty_gene_hypothesis).

The theory regarding the salt-conserving phenotype also belongs to the thrifty phenotype hypothesis, and it hypothesizes, why the frequency of hypertension is elevated especially in the USA black population. Earlier it was advantageous if salt-losing was kept minimal, especially for populations living in warm climate. In those times it was not as easy to get salt as it is today. As normal level of sodium chloride is essential for life, the ancestral sodium-conserving genotypes gave a selection advantage. In our days when salt is abundant, and a lot

of black people now live in cooler climate, this genetic background leads to a higher susceptibility to salt sensitive hypertension. But it is also true for non-black-population. The selection pressure made the humans salt craving, and now most people eat much more salt than would be needed.

There are several spectacular examples for the thrifty gene hypothesis. Pima Indians live in the USA and in Mexico. Earlier these people lived a hunter-gatherer existence and enjoyed both prosperity and good health. Shortly after the Pima Indians encountered the Anglos and Mexicans, however, they suffered a great famine. Over the course of just a decade or two, the Pima went from being rich and healthy to being poor and unhealthy. In our days, however, Pima Indians have an unusually high rate of obesity and T2DM. The prevalence of these diseases is above 50%. In contrast, the rate of obesity in the Indians living in Mexico is about 8%, although the genetic background of the two populations must be very similar.

According to this hypothesis, we can say that these diseases are caused by normal genes, which got to an unfavorable environment.

10.13.2. *Hygiene hypothesis*

The increasing prevalence of allergy is often explained by the hygiene hypothesis. This says that the lack of early childhood exposure to infectious agents and parasites increases the susceptibility to allergic diseases by suppressing natural development of the immune system. The immune system of a newborn is polarized toward a Th2 immune response. Because of the lack of infections we fail to induce a Th1 polarized response early in life, so as we grow up we are more prone to developing Th2 induced disease, like allergy or asthma. The rise of autoimmune diseases in the developed world has also been linked to the hygiene hypothesis (http://en.wikipedia.org/wiki/Hygiene_hypothesis). This hypothesis is also called Th1 trigger defect.

A lot of examples confirm this hypothesis. E.g. Karelians live in Russia and Finland. The Finnish Karelians live in much higher hygiene and they are significantly more allergic.

NOD (non-obese diabetes) mice are good animal models for the autoimmune T1DM. If they are kept in germ-free environment, the autoimmune diabetes develops much more frequently in them.

It was also shown that early antibiotic usage causes altered gut flora, and is associated with obesity and several other diseases.

It must be added, however, that life expectancy is significantly higher in populations with higher hygiene, thus the hygiene hypothesis does not mean that we must live in a dirty environment, but we have to recognize the mechanism how infections can prevent the disease. The other associated hypothesis is the **Th1 maturation hypothesis**. It says that children with an inborn **Th1 maturation defect** might survive by better health care and antibiotic use at the cost of higher asthma and allergy rates.

10.13.3. Additional theories

Similar to the above mentioned hypotheses is the theory, which says that there are **genetic variants**, which **gave earlier a selection advantage**, but in our days they **make us susceptible to certain diseases**. Such can be the variants associated with inflammation. Earlier it could be advantageous if a variant was associated with a stronger inflammatory response, because it could give a higher chance for survival of the different bacterial and viral infections or parasites. But today, it can make the carriers more susceptible to different chronic inflammatory diseases, like asthma, atherosclerosis or autoimmune diseases.

There are also variants, which are advantageous in younger age, but harmful in older. It is called **antagonistic pleiotropy**. The stronger inflammatory response can be advantageous in younger age, but harmful in older, because chronic diseases in old age are often inflammatory, like atherosclerosis, rheumatoid arthritis or Alzheimer disease.

It was found, e.g. that a variant in the toll like receptor gene (**TLR4** gene, Asp299Gly (D299G) variant), which is associated with a weaker immune response to Gram negative bacteria, and higher susceptibility to sepsis, is more frequent in centenarians.

Another example is the e4 allelic variant of the **APOE gene**. A strong detrimental effect of the e4 allele on survival was demonstrated which was mostly attributed to women with moderate lifespans of 70 to 95 years. One potential mechanism could be associated with inflammation which may be involved in aging through two main pathways associated with immunosenescence and synergies with chronic diseases that have inflammatory components. In contrast, several studies provided support for a beneficial role of the e4 allele in early life. For example, it was shown that the proportion of the e4 allele was significantly smaller in spontaneously aborted embryos than in adults. The proportion of the e4 allele was also found to be significantly larger in healthy liveborn infants compared with stillborn infants and with adults. These findings suggest that the e4 allele can benefit early survival. Studies also show that ApoE4 may protect against early life infectious diseases such as, e.g., diarrhea and liver damage caused by the hepatitis C virus infection. A putative protective mechanism may be

associated with an enhanced function of the immune system in early life with a role of ApoE as an immunomodulator (14).

The **old friends hypothesis** is connected with the hygiene hypothesis. This hypothesis says that the rise in allergies and inflammatory diseases is at least partly due to gradually losing contact with the range of microbes our immune systems evolved with. During evolution humans have evolved dependency on microbial exposure, such that our immune systems cannot now function properly without it. It is hypothesised that the "old friends" include commensal organisms (the normal microbiota of the skin, gut and respiratory tract of humans and animals) and some potentially pathogenic organisms such as helminths (worms), which establish chronic infections or carrier states. Since the 1800s, when allergies began to be more noticed, the mix of microbes we have lived with, and eaten, drunk and breathed in has been steadily changing. These changes include clean drinking water, safe food, sanitation and sewers, and maybe overuse of antibiotics. Whilst vital for protecting us from infectious diseases, these will also have inadvertently altered exposure to the 'microbial friends' which inhabit the same environments (<http://www.sciencedaily.com/releases/2012/10/121003082734.htm>).

There is a new method or rather a discipline developed in the genomic era that is capable to study hundreds of microbes at the same time. It is called **metagenomics**. Metagenomics utilize the power of next generation sequencing and the associated bioinformatics. From any samples (like tools or any environmental samples) the isolated genetic materials are sequenced resulting in sequencing data of a plenty of microbes at the same time. Then the sequencing data of all the members of the sampled microbial communities are put together with the help of bioinformatics. Because of its ability to reveal the previously hidden diversity of microscopic life, metagenomics offers a powerful lens for viewing the microbial world that has the potential to revolutionize understanding of the entire living world.

Another theory says that **diesel oil and some pollen** can react with each other, making the pollen more aggressive and more allergenic. That can be the explanation why the prevalence of allergy is higher in big cities.

Another theory says that there are a lot of **people who live elsewhere than their ancestors**, who were selected for a certain environment. To the new environment they are not able to adapt, and thus develop diseases. E.g. black people whose ancestors lived in Africa and now are living in Canada or in North Europe do not receive enough sunlight, and in this way they develop D-vitamin deficiency and autoimmune diseases. See additional examples in Chapter 13 (Gene environmental interaction).

It is also possible that as different **populations are mixed**, certain alleles from different populations can get next to each other, can interact, and this can lead to certain diseases.

10.14. Literature

1. Sørensen TI et al. Childhood body mass index--genetic and familial environmental influences assessed in a longitudinal adoption study. *Int J Obes Relat Metab Disord*. 1992 Sep;16(9):705-14.
2. Manolio TA et al. Finding the missing heritability of complex diseases. *Nature*. 2009 Oct 8;461(7265):747-53.
3. Kaati G et al. Cardiovascular and diabetes mortality determined by nutrition during parents' and grandparents' slow growth period. *Eur J Hum Genet*. 2002 Nov;10(11):682-8.
4. Eldar A, Elowitz MB. Functional roles for noise in genetic circuits. *Nature*. 2010 Sep 9;467(7312):167-73.
5. Moore JH, Asselbergs FW, Williams SM (2010) Bioinformatics challenges for genome-wide association studies. *Bioinformatics* 26: 445-455.
6. Yang J et al. Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet*. 2011;19(7):807-12.
7. Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*. 2008 Nov;92(5):265-72.
8. Johnson RJ, Andrews P, Benner SA, Oliver W. Theodore E. Woodward award. The evolution of obesity: insights from the mid-Miocene. *Trans Am Clin Climatol Assoc*. 2010;121:295-305;
9. Lev-Ran A, Porta M. Salt and hypertension: a phylogenetic perspective. *Diabetes Metab Res Rev*. 2005 Mar-Apr;21(2):118-31.
10. Franceschi C et al. Inflamm-aging. An evolutionary perspective on immunosenescence. *Ann N Y Acad Sci*. 2000 Jun;908:244-54.
11. Capri M et al. Human longevity within an evolutionary perspective: the peculiar paradigm of a post-reproductive genetics. *Exp Gerontol*. 2008 Feb;43(2):53-60.
12. Candore G et al. Inflammation, longevity, and cardiovascular diseases: role of polymorphisms of TLR4. *Ann N Y Acad Sci*. 2006 May;1067:282-7.
13. ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012 Sep 6;489(7414):57-74.
14. Kulminski AM, et al. Age, gender, and cancer but not neurodegenerative and cardiovascular diseases strongly modulate systemic effect of the Apolipoprotein E4 allele

on lifespan. PLoS Genet. 2014 Jan 30;10(1):e1004141.
<http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1004141#pgen.1004141-Fabris1>

10.15. Questions

1. What are the complex diseases?
2. What features have the complex diseases?
3. Why is it important to study the genomic background of complex diseases?
4. What are the difficulties which cause that genomic results infiltrate only slowly to the practice?
5. How can we prove that a disease has a heritable fraction?
6. What are the problems with the λ values?
7. How can the bias because of the environmental factors be mitigated?
8. What is the heritability of a trait?
9. What is the QT?
10. What are the discontinuous traits?
11. What is the QTL?
12. What are the factors, which make the determination of the genetic backgrounds of the complex diseases difficult?
13. What is genetic heterogeneity?
14. What is phenocopy?
15. What is genetic pleiotropy?
16. Give some examples why it is difficult to determine the phenotype of a complex disease!
17. What does it mean: missing heritability in genomics?
18. What are the CD/CV and CD/RV hypothesis?
19. What is synthetic association?
20. What is the consequence of the random behavior of the genome?
21. Why has it been unsuccessful so far to determine the genetic background of complex diseases?
22. What is the Bonferroni correction?
23. What results were provided by the ENCODE projects regarding the GWAS results?

24. What is the thrifty gene hypothesis?
25. What is the hygiene hypothesis?
26. What is the Old Friends hypothesis?
27. What is metagenomics?
28. What is the antagonistic pleiotropy?
29. What can be the consequence of the migration?

11. Genomic methods for complex diseases

Csaba Szalai

In this chapter the main methods for the investigation of the genomic backgrounds of the complex diseases and some related theoretical considerations will be summarized. The basic genetic methods will not be described here.

11.1. Genetic markers

A **genetic marker** is usually a sequence variation with a known location on a chromosome that can be used to identify individuals, with a relative high chance to differentiate between different alleles on homologous chromosomes. Genetic markers can be used to study the relationship between an inherited disease and its genetic cause (for example, a particular mutation of a gene that results in a defective protein). It is known that pieces of DNA that lie near each other on a chromosome tend to be inherited together (they are linked). This property enables the use of a marker, which can then be used to determine the precise inheritance pattern of the gene that has not yet been exactly localized. Genetic markers have to be easily identifiable, associated with a specific locus, and highly polymorphic, because homozygotes do not provide any information.

One of the most popular markers are the **microsatellites**, or **simple sequence repeats (SSRs)** or **short tandem repeats (STRs)**, are repeating sequences of 2-6 base pairs of DNA. Often they are very polymorphic, meaning that individuals are often heterozygotic to them, which means that they differ in the number of repeats.

They are widely used in mapping disease genes or differentiate between individuals. The human genome is now mapped by approximately 30,000 highly polymorphic microsatellites. The average length of **linkage disequilibrium (LD)** for microsatellites is ~100 kb, which is considerably higher than that of SNPs. Therefore, a single microsatellite captures a larger genomic region than does a single SNP. Microsatellites also provide **several other advantages**, such as a higher information content (6–10 alleles as compared with 2 alleles for SNPs), and a smaller interpopulation variability. Most existing forensic DNA databases are STR-based. It has been demonstrated that 20–50 ascertained autosomal SNPs could reach match probabilities similar to those obtained with 10–15 forensically used STRs.

But the **disadvantages of the STRs** are that the detection methods are quite complicated relative to those of the SNPs, they are much rarer than SNPs, and their mutation rates are 100,000 times higher.

Nowadays, the **advantages of the SNPs are much more significant**, and mainly because of their number and simple detection techniques, they will replace STRs in most areas. E.g. forty-five unlinked autosomal SNPs were ascertained by screening more than 500 candidate SNPs in 44 worldwide populations. These 45 ascertained SNPs have high levels of heterozygosity and low levels of population differentiation and are therefore suitable for universal human identification purposes. Multiplex genotyping assays for these SNPs have been developed.

11.2. Methods for the genomic backgrounds of diseases

11.2.1. Study of genetic variants

There are two main groups of approaches in the search for disease associated genetic variants:

- **Hypothesis-driven approaches**, like candidate gene association studies, single gene sequencing, etc. These are usually genetic methods.
- **Hypothesis-free approaches**, like whole genome screening, genome wide association studies (GWAS), whole genome sequencing, exome sequencing, and microarray measurements. These are usually genomic methods.

Genetic variations play important roles in disease susceptibilities, differences between individuals or in responses to drugs, and the study of them is important in discovery of novel drug targets, personal therapies or pharmacogenetics, etc. The HGP and the subsequent different genome projects (Human Variome Project, HapMap, 1000 Genome project, etc.) detected millions of genetic variants. Presently, there are about 150 million known short variants, and more than 4 million structural variants in the databases (http://www.ensembl.org/Homo_sapiens/Info/Annotation).

The simplest method for the study of the genetic background of a disease is the candidate gene association study. In these studies genes are selected, which are thought to play a role in the disease. Then, genetic variations are searched in these genes. Earlier the genes were sequenced in several individuals, now the databases contain practically all the common variants. The first one is often called **wet laboratory method**, the latter one ***in silico* method**.

Then, the selected variants are genotyped, and their frequencies are compared in the population with and without the studied trait (disease). If the frequencies of the variants differ in a statistically significant way between the two populations, then they are suspected to play a role in the disease susceptibility. Several 10 thousand such investigations have been carried out in the last decades in different diseases. But, there were a lot of problems with these studies. One of the problems is the multiple testing problems, but in a different way than discussed in the previous chapter. Because here, the same variants have been tested in different laboratories, and naturally only the positive results have been published; the negative ones have been discarded. And, if 100 laboratories study the same variants, there is a chance that one of them gets a positive association purely by chance. This is called **publication bias**. Because of this, hundreds of false positive results (and genes) have been published.

The other problem is that with this methods only those genes can be studied whose role was already known in the disease, and in this way no new mechanism could be detected.

The hypothesis-free genomic methods theoretically could solve this last problem. First, **whole genome screenings** were developed and carried out in several diseases. In this method families were screened with microsatellites. Those families were recruited where there were at least two affected siblings. These studies are also called **affected sib pair (ASP) studies**, or **linkage studies**. Here LOD scores were calculated. The **LOD score** (logarithm (base 10) of odds) is a statistical test often used for linkage analysis. The LOD score compares the likelihood of obtaining the test data if the two loci, or the disease phenotype and a locus are indeed linked, to the likelihood of observing the same data purely by chance. Positive LOD scores favor the presence of linkage, whereas negative LOD scores indicate that linkage is less likely. A LOD score greater than 3.0 is considered evidence for linkage. A LOD score of +3 indicates 1000 to 1 odds that the linkage being observed did not occur by chance. On the other hand, a LOD score of less than -2.0 is considered evidence to exclude linkage (http://en.wikipedia.org/wiki/Genetic_linkage).

The method has given a lot of interesting results, but there have been several problems with it. First, it is difficult to collect families with two affected siblings, second, the genotyping of the microsatellites are very cumbersome and expensive. Because of this latter, the number of microsatellites in the studies was limited (usually not more than 400), thus the resolution was very low. This means that it was a great chance that disease associated loci, which were not in linkage with any of the microsatellites were lost. In addition, these studies could determine only genomic regions (because of the limited number of markers), and not genes. And often,

these regions are large, several megabase long and contain several hundreds of genes. In this way, additional methods are needed for the determination of the genes.

11.2.2. GWAS

Presently, the most popular method for the study of the genomic background of complex diseases and traits is called **GWAS (genome-wide association study)**, also known as whole genome association study (WGA study or WGAS). The method has become possible, when arrays and chips have been developed with which first 100 thousand, then several million SNP could be genotyped in one measurement, and the price of one chip has become relatively cheap, i.e. about \$100. First, only SNPs were determined, later, when the significance of CNVs became apparent, they were involved as well. The CNVs were determined through their known linkage with SNPs. In 2007 this method was selected for the breakthrough of the year (<http://science.sciencemag.org/content/318/5858/1842.full>).

There are two main companies in the markets, Affymetrix and Illumina. The Affymetrix Genome-Wide Human SNP Array 6.0 features 1.8 million genetic markers, including more than 906,600 SNPs and more than 946,000 probes for the detection of CNVs.

The Illumina HumanOmni5-Quad (Omni5) BeadChip can detect 4.3 million tagSNPs selected from the International HapMap and 1000 Genomes Projects that target genetic variation down to 1% minor allele frequency (MAF).

In GWAS the distribution (frequencies) of the variants is compared in the different populations; usually one of them is affected with the trait, the other is not. But, with the development of the statistical methods GWAS has become capable of studying the genomic background of continuous traits (like fasting glucose levels or blood pressure) as well. In this latter case there are no different groups.

GWAS has been offering a great chance for the investigation of the genomic background of the diseases, which have been utilized by a lot of research groups and consortia. Because of the strict statistical conditions and the large investigated populations, the results of GWAS may contain only few false results; and because this is a hypothesis-free method, there is a possibility that it reveals new aspects of the disease. To make these important results public, a web page was established on 25 November 2008 which later moved to the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) at <http://www.ebi.ac.uk/gwas>. The Catalog is a quality controlled, manually curated, literature-derived collection of all published GWAS assaying at least 100,000 SNPs and all SNP-trait associations with p-values $< 1.0 \times 10^{-5}$. The Catalog also publishes a GWAS

diagram of all SNP-trait associations, with genome-wide significance of p-values $\leq 5.0 \times 10^{-8}$, mapped onto the human genome by chromosomal locations and displayed on the human karyotype (Figure 11.1).

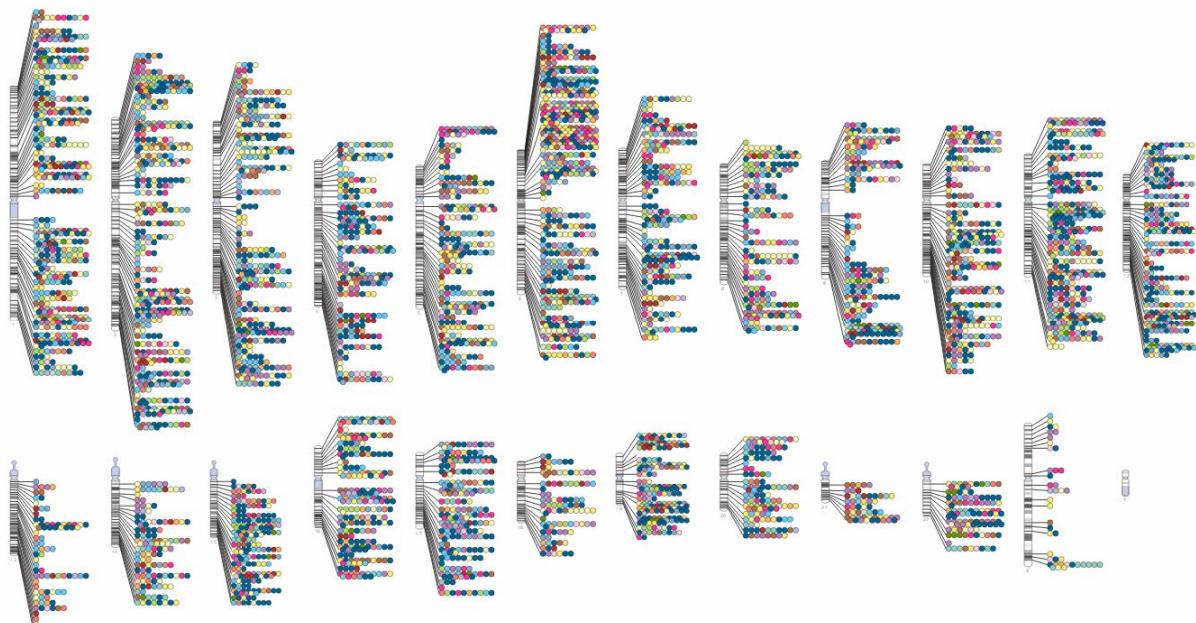


Figure 11.1. The diagram shows all SNP-trait associations with $p\text{-value} \leq 5.0 \times 10^{-8}$, published in the GWAS Catalog. Each colored plot represents a given trait. Source: <http://www.ebi.ac.uk/gwas/diagram> 19/02/2016.

11.2.3. Evaluation of GWAS results

The evaluation and handling of GWAS data are a great challenge for the bioinformaticians. As we discussed in Chapter 10, one of the main problems is the multiple testing problem. If the p value of a SNP corresponds to the Bonferroni corrected value, then it is said that it reached the level of **genome wide significance**. It is, e.g. in case of 1 million SNPs 5×10^{-8} . As the main characteristics of the complex diseases are **variants with weak effects**, this low p value often can only be achieved through involving large populations. Often the number of participants must be $>100,000$, which is very difficult and expensive to collect, and which is in case of rarer diseases even impossible. Because of this, GWAS are often carried out by large international consortia (e.g. WTCCC, Chapter 12).

A method to attenuate this problem can be, if several smaller populations are investigated independently. In this way the p values in the independent studies for each SNP are multiplied, and it is easier to achieve the low values (e.g. $10^{-3} \times 10^{-3} = 10^{-6}$). Usually, a discovery GWAS

is carried out in a smaller population (**discovery cohort**). Then, SNPs are selected with a not so strict p value (e.g. cut off value $< 5 \times 10^{-2}$), then several independent populations are collected (**replication cohorts**), and only the selected SNPs are studied. The SNPs which are confirmed in the replication cohorts can be those which are associated with the disease.

As it is discussed in Chapter 10 (10.11. Possible solutions), new statistical methods are also under development, such as Bayesian statistics and **pathway analysis**. For this latter, several databases are available like **Gene Ontology (GO)**; (<http://www.geneontology.org>) or KEGG (Kyoto Encyclopedia of Genes and Genomes (<http://www.genome.jp/kegg/>)).

Gene Set Enrichment Analysis (GSEA) is a computational method, which was originally developed for gene expression studies and can be applied in GWAS as well. This determines whether different a priori defined sets of genes show statistically significant, concordant differences between two biological states (e.g. phenotypes). Then the sets of genes are ranked according to their associations.

With these methods several new disease associated pathways have been detected.

11.2.4. Partial genome screenings

Partial genome association studies (PGAS) are simpler variations of the above described genome screenings, and are usually preceded by other studies. In these studies selected genome regions (found previously by whole genome screening or GWAS) are screened with SNPs more densely located than in the previous studies, or tagSNPs in genes or regulatory regions are genotyped, which were detected by other studies (e.g. by GWAS or gene expression measurements).

In these studies a lot of the statistical difficulties are solved, and it is a greater chance to find the responsible genes or variants.

11.2.5. Positional cloning

Positional cloning is the method, when genes are identified and verified with *in vitro* and *in vivo* experiments after locating the region by linkage studies (See more details: http://en.wikipedia.org/wiki/Genetic_screen#Positional_cloning).

11.2.6. Personal genomics

Next to the strict scientific studies, the results and methods in genomics are utilized by profit oriented companies. It is called **direct to consumer (DTC) service**. The different companies offer different services with different techniques. Here are some quotations from the Wikipedia 10/10/2012 (http://en.wikipedia.org/wiki/Personal_genomics):

Companies which offer genome-wide personal genomics services have already gone to market and are selling their services direct to the consumer.

There are some controversial ethical and legal issues connected with these services, e.g.: “Genetic discrimination is discriminating on the basis of information obtained from an individual’s genome. Genetic non-discrimination laws have been enacted in some US states and, at the federal level, by the Genetic Information Nondiscrimination Act (**GINA**). The GINA legislation prevents discrimination by health insurers and employers, but does not apply to life insurance or long-term care insurance. Patients will need to be educated on interpreting their results and what they should be rationally taking from the experience. It is not only the average person who needs to be educated in the dimensions of their own genomic sequence but also professionals, including physicians and science journalists, who must be provided with the knowledge required to inform and educate their patients and the public.”

Some companies use the collected data also for scientific studies, naturally, after the informed consent has been signed by the participants. The most successful personal genomic company is 23andme, which genotyped by 2015 over 1 million individuals (<https://en.wikipedia.org/wiki/23andMe>). It has already published several papers about the genomic background of different diseases including e.g. Parkinson disease (<http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1002141>).

The experiences are usually positive. Earlier it has been generally believed that if an individual receives data that he/she has a greater risk to a certain disease, she/he will be depressive or it will influence his/her psyche adversely. But according to surveys no such things have been experienced. Possibly, it is similar to the case when somebody is a smoker or obese, and has a greater chance for the development of several serious illnesses. In contrast, most participants changed their lifestyle to avoid the disease. It must be added, however, that these participants must be more health-conscious than the average people.

11.2.7. New generation sequencing (NGS)

More details can be seen in Chapter 9.

Presently NGS is too expensive, and results in too many and uncertain data, and in this way it is not suitable for population studies. Often its reduced variations are used. Such application is the **exome sequencing** (http://en.wikipedia.org/wiki/Exome_sequencing), where only the protein coding regions are sequenced, which is about 30 Mb, 1% of the whole genome. According to the estimations, 85% of the mutations causing monogenic diseases are in this region. Its disadvantage is that in the complex diseases, the majority of variations are outside this region. But possibly, the rarer variations with strong effects are rather here.

With the development of NGS several methods have been worked out utilizing the technique. The central method of the ENCODE project was the **DNase-seq** (<http://www.nature.com/encode/#/threads>). The DNase I enzyme will preferentially cut live chromatin preparations at sites where nearby there are specific (non-histone) proteins. The resulting cut points are then sequenced using high-throughput sequencing to determine those sites ‘hypersensitive’ to DNase I, corresponding to open chromatin. The cell-specific patterns of DNase I hypersensitive sites show remarkable concordance with experimentally determined and computationally predicted binding sites of transcription factors and enhancers.

ChIP-seq: Chromatin immunoprecipitation followed by sequencing. Specific regions of crosslinked chromatin, which is genomic DNA in complex with its bound proteins, are selected by using an antibody to a specific epitope. The enriched sample is then subjected to high-throughput sequencing to determine the **regions in the genome most often bound by the protein to which the antibody was directed**. Most often used are antibodies to any chromatin-associated epitope, including transcription factors, chromatin binding proteins and specific chemical modifications on histone proteins.

11.2.8. Measurement of gene expression

The third basic method for the investigation of genomic background of diseases is the measurement of gene expression. The theory is that if the expression level of a gene is different in a pathologic tissue relative to the healthy tissue, then this gene may participate in the pathomechanism of the disease. Similarly, external effects change the level of gene expression, thus their (e.g. drugs) effects can be studied. The problem is that it is often very difficult to get the tissue concerned. E.g. for the study of atherosclerosis the best tissue is the atherosclerotic plaque from large arteries, in asthma the lung, in Alzheimer disease the brain, in T1DM the pancreas, etc. In this way these measurements are carried out in most diseases

from other sources, like tissues from model-animals, or from cell or tissue cultures. The most significant exceptions in this respect are the tumors, which are usually removed, and can be analysed *ex vivo*. It is the reason, why the gene expression measurement gets into the clinical practice only in oncology.

This technique developed in similar speed as the sequencing or the genotyping. First, the expression of only one gene could be measured and was quantified by comparing its expression level to an internal control, or housekeeping gene (**Housekeeping genes** are typically constitutive genes that are required for the maintenance of basic cellular function, and are expressed in all cells of an organism; and their expression level is similar under normal and pathophysiological conditions. Such a gene is e.g. β -actin.) Later specific equipment was developed (real time PCR) for the quantification of the gene expression, and then came the microarrays, with which the expression of all genes could be measured on one chip or array. First, the accuracy and reproducibility of the measurement were not very good and it was very expensive, but later the methods improved considerably and became significantly cheaper, thus it could get into the standard diagnosis methods in some areas.

Nowadays, there has been approaching a novel method, which will probably replace the traditional (but not too old) method. This method is the **RNA-sequencing, or RNA-seq**, which means isolation of RNA sequences, often with different purification techniques to isolate different fractions of RNA followed by high-throughput sequencing (NGS). The advantage of this method relative to the microarray that it provides information on differential expression of genes, including gene alleles and differently spliced transcripts; non-coding RNAs; post-transcriptional mutations or editing; and gene fusions (<http://en.wikipedia.org/wiki/RNA-Seq>). Sometimes gene expression is influenced by different variant loci. **Expression quantitative trait loci (eQTLs)** are genomic loci that influence the expression levels of mRNAs.

11.2.9. Determination of the methylation of the genome

The development of NGS also accelerated and improved the **epigenetic studies**. The methylation of the DNA most often occurs at CG dinucleotides on cytosine. Usually, it is detected by methylation-specific PCR and comparative sequencing. The method is based on a chemical reaction of **sodium bisulfite (NaHSO₃)** with DNA that **converts unmethylated cytosines to uracil**, followed by PCR. However, methylated cytosines will not be converted in this process, and primers are designed to overlap the CpG site of interest, which allows one to determine methylation status as methylated or unmethylated. The samples can be sequenced also on NGS platform. The sequences obtained are then re-aligned to the reference

genome to determine methylation states of CpG dinucleotides based on mismatches resulting from the conversion of unmethylated cytosines into uracil.

11.2.10. Additional microarray-based methods

Several new microarray-based methods have been developed in the recent years. After discovering of the **miRNA**, products have appeared in the market, with which these could be measured.

Array-comparative genomic hybridization (array CGH) is a technique to detect CNVs. It can be used to create a virtual karyotype, and is capable of detecting new structural variations in tumor tissues, and can be used in prenatal or preimplantation diagnosis, or diagnosing the genomic background of birth abnormalities (http://en.wikipedia.org/wiki/Comparative_genomic_hybridization).

ChIP-on-chip array is a technique that combines chromatin immunoprecipitation ("ChIP") with microarray technology ("chip"). It is used to investigate interactions between proteins and DNA *in vivo*. The ChIP-seq is the same method using NGS.

11.3. Animal models

11.3.1. The advantages of the animal models

Animal models and experiments have an important role in the study of the genomic background of complex diseases. There are diseases, where most of our knowledge about their molecular pathogenesis originates from animal models. Let us see what advantages the use of animals has.

- The experimental animals can often be crossed freely. In contrast to humans, there is a greater possibility to investigate the connection between genotypes and phenotypes through several generations. It is much easier to carry out **QTL studies**. The generation time of a mouse is 2 months (in humans it is 20-30 years), thus in a year several mouse generations can be studied. With directed crosses and backcrosses the segregation of the genetic markers and symptoms, or QTs can be tracked during several generations.
- There are different mouse strains, which differ from each other in disease susceptibilities or other phenotypes. These can be used as animal-models, or through crosses we can study the connection between segregation of genetic markers and

phenotypes. These animals can be kept in strictly controlled environments, thus the effects of these can be easier studied.

- At gene level human is not so different from the rest of the animals. The essential genes are usually the same; we differ from the mouse only in 300 genes. But, species like *Drosophila melanogaster* (fruit fly) are also widely used, and a lot of pathways (like Hippo pathway, which is conserved and plays an important role in organ size control) were first discovered in this animal. Already two Nobel prizes have been given because of the studies of this species (http://en.wikipedia.org/wiki/Drosophila_melanogaster).
- There are a lot of experiments which in humans may not be carried out from ethical reasons, only in animals.
- There is much easier to get tissues from animals (lung, brain etc.) and measure gene expression, etc. The diagnoses of diseases are much more accurate.
- There are a lot of animal-models for different human diseases.
- There is a possibility to develop genetically modified animals.

Let us see the last two points in more detail. There are two basic types of genetically modified animals used in these studies. One of them is the **knock out or KO animals**. In KO animals researchers inactivate, or "knock out," an existing gene by replacing it or disrupting it with an artificial piece of DNA. Among them the mice are the most significant for studying the role of genes which have a known sequence, but whose functions have not yet been determined (http://en.wikipedia.org/wiki/Knockout_mouse). By causing a specific gene to be inactive in the mouse, and observing any differences from normal behavior or physiology, researchers can infer its probable function.

In 2006, 33 research centers in 9 countries founded the **International Knockout Mouse Consortium**, then in July 2011 the **International Mouse Phenotyping Consortium** aiming to build a huge, shared resource for biomedical research (<http://www.mousephenotype.org/>). Mouse embryonic stem cells have been produced, in which researchers have "knocked out" each of the more than 20,000 specific mouse genes that code for proteins. By growing mice from these cells, researchers can gain insight into the role that the missing genes play in health and disease. The phenotyping effort will aim to probe the anatomy, development, physiology, behavior, and disease traits of 5000 of these mouse lines by the end of 2016 (<http://news.sciencemag.org/scienceinsider/2011/09/the-consortium-that-will-launch-.html>).

Knocking out genes can also be used for animal-models for different diseases (Table 1). In these animals the molecular pathomechanisms or different therapies can be studied.

The other types of the genetically modified animals are called **transgenic**. Here, the **genes are over-expressed**, or new genetic information is inserted into the mouse genome. These animals can be used for the same aims as the KO animals.

The over-expressed genes are usually under the regulation of promoters with strong activity. It is also possible that the promoters are only active in certain organ or tissue. In this way the gene will be over-active only in this organ. E.g. *SCGB1A1* is expressed only in the lung, but here it is highly expressed. IL5 gene was introduced after the promoter region of this gene in a mouse strain. The mouse over-expressed the IL5 in its lung. As IL-5 is an eosinophil colony-stimulating factor, it is a major regulator of eosinophil accumulation in tissues, and can modulate eosinophil behavior at every stage from maturation to survival, IL5 over-expression caused eosinophilia in the lung of this mouse, and asthmatic symptoms developed.

Earlier it was difficult to study **essential genes** in animals, because lack or over-expression of these genes are often lethal in embryonic development. To avoid these problems **conditional transgenic animal models** were developed. In these animals the gene will be inactivated or induced in vivo. One of the best-known methods for this is the Cre-Lox recombination. Cre-Lox recombination is a site-specific recombinase technology which allows the DNA modification to be targeted to a specific cell type or be triggered by a specific external stimulus (https://en.wikipedia.org/wiki/Cre-Lox_recombination).

RNA interference can also be used for transient in vivo gene inactivation. In *Caenorhabditis elegans* all the genes were studied by inactivating them with RNAi. This method can also be used in more developed animals, even in humans.

11.3.2. Shortcomings of animal models

It is true that the difference between animals and human at gene level is relative small. But as it turned out (see ENCODE project), about 80% of the genome have certain functions. And at genomic levels the differences are larger. Comparing human and mouse it is also true that some processes are different, or homolog genes can have different functions. Because of these, all the results from animal models have to be confirmed in humans or human systems.

Similarly, the similar diseases can have different pathomechanisms, and there are human diseases for which no good animal models have been developed so far.

E.g. the gene resistin have been discovered in mouse white adipose tissue, which expressed it, and this expression caused insulin resistance, one of the main characteristics of T2DM. Later it turned out that in humans resistin was expressed in macrophages.

11.3.3. Experimental disease models

Because different diseases can be developed by gene manipulations, this can be utilized for developing experimental disease models. It can also be carried out by inducing random mutations by mutagenic agents, then by phenotyping the animals, different strains can be developed by different crosses. Afterwards by genomic screening the genomic background of the diseases can be found out. This method is also capable of studying complex diseases. Such mice are produced e.g. in the Jackson Laboratory (<http://jaxmice.jax.org/>), from where mice strains with a given phenotype can be ordered. One of the widespread methods for this is ENU mutagenesis. **ENU**, also known as *N*-ethyl-*N*-nitrosourea, is a highly potent mutagen. For a given gene in mice, ENU can induce 1 new mutation in every 700 loci. Mutation is usually induced in male mice, which are then crossed with a wild type female. The G₁ progeny can be screened to identify dominant mutations. However, if the mutation is recessive, then G₃ individuals homozygous for the mutation can be recovered from the G₁ males (see: <http://en.wikipedia.org/wiki/ENU>).

There are databases with mouse phenotypes, like **Mouse Phenome Database** which characterizes strains of laboratory mice to facilitate translational discoveries and to assist in selection of mouse strains for experimental studies (<http://phenome.jax.org/db/q?rtn=docs/aboutmpd>).

There are some widespread animal models for polygenic diseases, like Non-obese diabetic (NOD) mouse, spontaneously hypertensive rat (SHR), Dahl salt sensitive rat, New Zealand Obes (NZO) mouse, etc.

Gene	Modification	Animal	Diseases
LDLR	KO	mouse	atherosclerosis
APOE	KO	mouse	atherosclerosis
Ob (LEP in humans: Leptin)	KO (ob/ob)	mouse	obesity
LPR (Leptin receptor)	KO (db/db)	mouse	obesity
TBX21	KO	mouse	asthma
ANP	KO	mouse	hypertension
SNCA	over-expression	drosophila	Parkinson disease
Mutant APP	over-expression	mouse	Alzheimer disease

Table 1.

Some animal models for human diseases developed by manipulating a gene

11.4. Literature

1. International HapMap 3 Consortium, Altshuler DM et al. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010;467(7311):52-8.
2. International HapMap Consortium, Frazer KA et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007;449(7164):851-61.
3. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;7;447:661-78.
4. Pennisi E. 1000 Genomes Project Gives New Map Of Genetic Diversity. *Science* 2010; 330: 574-5.)
5. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet*. 2007 Dec;81(6):1278-83.
6. <https://www.23andMe.com/>

7. Kaye J. The regulation of direct-to-consumer genetic tests. *Hum Mol Genet*. 2008;17:180-3.
8. Allayee H, Ghazalpour A, Lusis AJ. Using mice to dissect genetic factors in atherosclerosis. *Arterioscler Thromb Vasc Biol*. 2003 Sep 1;23(9):1501-9.
9. Mehrabian M et al. Identification of 5-lipoxygenase as a major gene contributing to atherosclerosis susceptibility in mice. *Circ Res*. 2002 Jul 26;91(2):120-6.
10. Rapp JP. Genetic analysis of inherited hypertension in the rat. *Physiol Rev*. 2000 Jan;80(1):135-72.
11. ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012 Sep 6;489(7414):57-74.

11.5. Questions

1. What are the genetic markers? Give examples!
2. What are the advantages and disadvantages of the STRs relative to the SNPs?
3. What basic methods are known for the study of the genomic background of the complex diseases?
4. What are in genomics the in silico vs. wet laboratory methods?
5. Give examples for the hypothesis-driven and hypothesis-free genetic/genomic methods!
6. What is the candidate gene association study?
7. In what genomic studies are the microsatellites used?
8. What is the disadvantage of the linkage studies?
9. What is the LOD score?
10. What is the GWAS?
11. What are the main difficulties of GWAS?
12. What does genome wide significance mean?
13. What is the pathway analysis?
14. What is the gene set enrichment analysis?
15. What is the PGAS?
16. What is the positional cloning?
17. What does personal genomics mean?
18. What can be the problems with the DTC genomics companies?
19. What does exome sequencing mean?

20. What does DN-ase-seq mean?
21. What is the ChIP-seq?
22. What are the advantages of the microarray gene expression measurements?
23. What is the RNA-seq method?
24. What is the CGH?
25. How is it possible to determine the methylation pattern?
26. What is the advantage of using animal models for studying human diseases?
27. What genomic modified animal models do you know and what are they good for?
28. What are conditional transgenic animals?
29. What are the shortcomings of the animal models?
30. Give examples for the experimental disease models!

12. Population and evolution genetics

Csaba Szalai

12.1. Population genetics

According to the definition population genetics is the study of the frequency and interaction of alleles and genes in populations. This is one of the most frequently used methods for studying the genetic and genomic background of diseases in humans. In the first part of this chapter the basic terms and methods of human population genetic studies are described.

12.1.1. Types of sample collection

In human population genetic studies there are two main basic methods for sample collection. The first is the retrospective sample collection, which is used for **retrospective studies**. In this method, for investigating the genomic background of a disease, participants are recruited and clinical data and biological samples are collected according to their disease-status. Here, there are usually two groups (or cohorts), which can be compared from different points of view. Collection of the patients (participants with the disease) is relatively simple, especially for physicians specialized for the disease concerned. Collection of controls is more problematic, because they usually have to match the patients in age and gender distribution, ethnicity (see population stratification below), and they should not have any diseases which can cause bias in the results, thus they usually appear more seldom at the doctor. The study design is very important. It is essential for the successful study that all the vital data are collected from the participants including laboratory data, clinical parameters and environmental factors. This latter can be collected e.g. by filling pre-made questionnaires. All the patients must be informed about the study, and informed consent must be signed by all patients or parents (in case of minors). The study must be conducted according to the principles expressed in the Declaration of Helsinki (<http://www.wma.net/en/30publications/10policies/b3/>), and approvals by Ethics Committees are also necessary. Most population genetic studies are carried out in this way, because it is relatively simple, cheaper and easier to be carried out than the other types of study. These studies are called **case-control studies**. Next to many small studies, there are several large studies for the genetic backgrounds of the diseases. The

Wellcome Trust Case-Control Consortium (**WTCCC**) started in 2005 (<http://www.wtccc.org.uk/>). This is a consortium of 50 research groups across the UK. The WTCCC aims were to exploit progress in understanding of patterns of human genome sequence variation along with advances in high-throughput genotyping technologies, and to explore the utility, design and analyses of genome-wide association studies (GWAS). The WTCCC has substantially increased the number of genes known to play a role in the development of some of our most common diseases and identified approximately 90 new variants across all of the diseases analysed. The WTCCC also carried out major experiments to address the genome-wide measurement of copy number variation (CNV) within 19,000 samples tested in phase one (16,000 disease samples and 3,000 common controls) and additional samples from breast cancer.

The WTCCC became a great success, and in 2008 the **WTCCC2** was established. Here 120,000 samples were collected and several further diseases were investigated with GWAS. In 2009 WTCCC3 was also launched.

In **prospective studies** the samples and data are usually collected from healthy participants, and then, their life has been tracked for decades by regular visits, data and sample collections. The organization of these studies are much more complex, they are more expensive and last much longer than the retrospective studies, but usually the results are more valuable and less biased. In retrospective studies there can be several biases. E.g. persons who died of the disease, or the mildest, who remained unrecognized can be underrepresented.

There are several famous prospective studies. The first such large study started in 1948 in Framingham, USA, and collected samples and data from 5209 participants. Since then the **Framingham Heart Study** has been in progress and already the 3rd generation has also been involved. Most of our knowledge about the risk factors of the cardiovascular diseases has come from this study.

The **UK Biobank project** is even larger. It started in 2007 and aimed for collecting samples and data about 500,000 volunteers in the UK, at ages from 40 to 69. The volunteers will be followed for at least 25 years thereafter. Its main aim is to investigate the respective contributions of genetic predisposition and environmental exposure (including nutrition, lifestyle, medications, etc.) to the development of disease.

The **Avon Longitudinal Study of Parents and Children (ALSPAC)** - which is also known as Children of the 90s - is a long-term health research project. More than 14,000 mothers enrolled during pregnancy in 1991 and 1992, and the health and development of their children have been followed in great detail ever since. The ALSPAC families have provided a vast

amount of genetic and environmental information over the years. Around 2012 the first grandchildren were born and also included in the study. On the web page of the study several interesting results can be read (<http://www.bristol.ac.uk/alspac/>).

12.1.2. Selection of populations for genetic studies

When patients are selected for a retrospective study aiming at investigating the genomic background of a disease, there are two main strategies. We can apply **strict conditions**, i.e. the symptoms of all patients must be exactly the same. But very often it is very difficult, even impossible. E.g. in case of asthma there can be a lot of phenotypes, all of them are associated with asthma, like rhinitis, conjunctivitis, dermatitis, high IgE level, eosinophilia, all of which have partly different genetic background, but none of them is necessary for the development of the disease. There are asthmatics that have all of these symptoms, some only parts of them, and some none of them. The more similar the phenotypes of the patients are, the larger is the chance that their genetic background is the same, and the easier is to clarify it. But, if we want to include only those patients in a group who have exactly the same phenotype, it would be very difficult to collect enough patients, and we will have difficulties in the statistical analysis. If we apply **loose conditions**, e.g. all the asthmatics are included, then we could have larger populations, but their genetic background will be heterogeneous, and the effects of the individual genetic variations will be diluted.

One of the possible solutions is that we use **intermediate phenotypes**, also called **endophenotypes**. E.g. in case of allergy we can collect participants with high IgE levels, or in hypertension patients with low rennin levels, or in obesity patients with high waist to hip ratio, or high leptin levels, in atherosclerosis patients with high LDL-C, or low HDL-C, etc. In these cases we do not determine the genetic background of the diseases, but the QTL for these QTs, and these are associated with the diseases, thus we can determine some parts of the genetic background.

12.1.3. Hardy Weinberg equilibrium

For the reliable results in population genetic studies the deviation from the Hardy Weinberg equilibrium (HWE) must always be controlled. **The HWE gives the expected distribution of the genotypes in a random population.**

In cases of two alleles, the relative frequencies of the alleles can be given with two equations:

$$p + q = 1 \text{ and } p^2 + 2pq + q^2 = 1,$$

where:

- p is the frequency of the major allele,
- q is the frequency of the minor allele,
- p^2 and q^2 are the frequencies of the homozygotes,
- $2pq$ is the frequency of the heterozygotes.

We can calculate the **expected number of patients** in each genotype group by multiplying the number of patients in the population with the frequencies. E.g. if the minor allele frequency is 0.2 ($q = 0.2$; $p = 0.8$), and the number of patients is 100, then the expected number of minor allele homozygotes is $0.2^2 \times 100 = 4$; the heterozygotes $2 \times 0.2 \times 0.8 \times 100 = 32$; the major allele homozygotes $0.8^2 \times 100 = 64$.

In population genetic studies it **must be controlled, whether the real numbers in each genotype group deviate from those of the theoretically expected values**. For this, several methods and online programs are available, like in web page of <http://ihg2.helmholtz-muenchen.de/cgi-bin/hw/hwa1.pl>, where a lot of alleles can be investigated at the same time, and if we compare two populations, the program carries out association studies as well.

About HWE more can be read in the Wikipedia (https://en.wikipedia.org/wiki/Hardy%20Weinberg_principle). Here we show the possible causes and consequences if the results statistically deviate from the equilibrium. In population genetic studies investigating the genetic background of a disease, the deviation from HWE can be originated from different factors:

- Wrong genotyping.
- The sampling was not random. E.g. there are a lot of relatives in the collected population.
- The population is inbred. In both last cases, the rate of homozygotes is often increased.
- The studied allele is in a CNV (repeat) region. In this case the rate of heterozygotes can be elevated.
- The studied genotype plays a role in the disease. E.g. it triggers the disease in homozygote form (recessive diseases). In that case the actual number of homozygotes is larger than the expected.

If there is a deviation from the HWE in controls, then the allele must usually be excluded from the study, because it can result in wrong conclusions. In cases (case-control studies) at the first four points the situation is the same, but at the fifth, because the real causes are unknown, the results must be further investigated. In both controls and cases the genotyping can be repeated with an alternative method, and the family connections must be studied, if it is possible. At genomic studies (e.g. in case of GWAS) when several thousands of genotypes are investigated parallel, the investigation of family relations is possible; there are even programs for it.

The last point is the most interesting one in cases. If a genotype is significantly more frequent in cases, then it can mean that it increases the risk to the disease; if it is significantly rarer, then it can protect from it. Both are valuable pieces of information regarding the genetic background of the disease, but must be confirmed with alternative methods. In both cases it can be said that the genotype is associated with the disease (see more in 11.1.6).

12.1.4. Linkage and haplotype

It is well known that in the genetic life cycle, during meiosis crossing over or homologous recombination takes place between homologous chromosomes. It means that part of the genetic material exchange between the chromosomes resulting in new combinations of DNA within chromosomes. E.g. in male meiosis the average number of crossing over is 49. From our point of view it means that if a marker locus was next to a mutation, they can be separated in the next generation. Because in linkage studies genetic markers are used for tracking the disease-associated variants, this has serious consequences. For the calculation of the chance of co-inheritance of two alleles, the **linkage disequilibrium (LD)** was introduced. In population genetics, **linkage disequilibrium is the non-random association of alleles at two or more loci that may or may not be on the same chromosome**. In other words, linkage disequilibrium is the occurrence of some combinations of alleles or genetic markers in a population more often or less often than would be expected from a random formation of haplotypes from alleles based on their frequencies.

Example:

In a population genetic study two loci were genotyped with alleles of A and G, respectively; both have a population frequency of 50%. If there is no linkage between them, i.e. they are inherited independently, than the occurrence of AG combination is $50\% \times 50\% = 25\%$ in the

population. If we measure 40% occurrence, it means that they are not inherited independently, then there is LD between them. But the situation is the same if they occur significantly rarer together (e.g. 10%) than expected.

Many measures of LD have been proposed, though all are ultimately related to the frequency difference between a two-marker haplotype and the frequency expected assuming that the two markers are independent. The two commonly **used measures of linkage disequilibrium are D' and r^2 .** D' is a population genetics measure that is related to recombination events between markers and is scaled between 0 and 1. A D' value of 0 indicates complete linkage equilibrium, which implies frequent recombination between the two markers and statistical independence under principles of Hardy-Weinberg equilibrium. A D' of 1 indicates complete linkage disequilibrium, indicating no recombination between the two markers. Alternatively, r^2 is the square of the correlation coefficient, and is a more statistical measure of shared information between two markers. The r^2 measure is commonly used to determine how well one SNP can act as a surrogate for another. There are multiple dependencies between these two statistics, but most notably r^2 is sensitive to the allele frequencies of the two markers, and can only be high in regions of high D' (<http://biodatamining.biomedcentral.com/articles/10.1186/1756-0381-2-7>).

If two or more alleles are next to each other and are in LD, then we say that they are on the same **haplotype**. In population genetics haplotype frequencies are frequently determined in a population, and it can also be associated with altered risk to a disease. The haplotypes or LD maps of a population are very important, which is demonstrated by the success of the **International HapMap project**. The first project started in 2002, and already 3 phases have taken place (<http://hapmap.ncbi.nlm.nih.gov/>) and have been of great importance on population genetic studies.

In population genetic studies, when several adjacent SNPs are genotyped, their LD map can be depicted in a form of a triangle (**Figure 12.1**) or heat maps. For calculation of haplotypes there are several online software products, like [Haploview 4.1](#).

Let us go back to the linkage. In genomic or genetic studies linkage can have two meanings. The first one corresponds to the previously described meaning, i.e. the alleles are often inherited together, and thus there is linkage between them. But linkage can be between a phenotype (QT) and a marker. Here can be hypothesized that there is a locus (genetic variation) linked to the marker, which influences the QT.

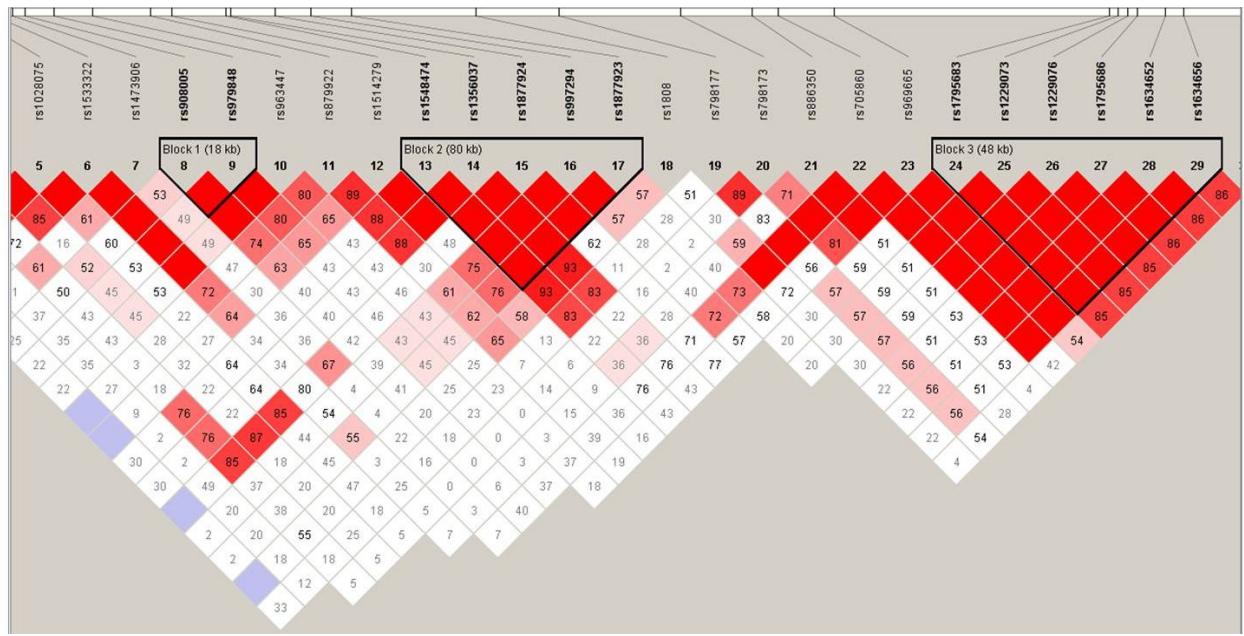


Figure 12.1. LD (or heat) map.

There are 3 regions of triangles grouping red blocks inferring 3 haplotype blocks. The numbers above the map show the marker numbers and names of the alleles. The numbers in the squares are r^2 between markers (SNPs).

Source: <http://woratanti.files.wordpress.com/2009/11/picture6.jpg>; 18/02/2013.

12.1.5. Founder populations

As the probability of a crossing over between two loci is proportional to their physical distance, this is used for the estimation of the so called genetic distance, and the **unit centiMorgan (cM)** was introduced. It is defined as that **distance between loci for which the expected average number of intervening chromosomal crossovers in a single generation is 1%**. The *centimorgan* was named in honor of the Nobel-laureate geneticist [Thomas Hunt Morgan](#).¹ 1 cM corresponds to about 1 million base pairs in humans on average; it is a little bit larger in females than in males.

If a mutation arises in an individual, and is transmitted to the next generation, the loci around it can change because of crossing over. The farther a locus is, the larger is the possibility that in a few generations the two will not occur on the same chromosome. In linkage studies it has a great importance. The closer the marker to the mutation is, the larger is the possibility that the mutation is detected with the help of the marker. It means that as densely located markers

must be used as possible. But the human genome is enormous, and the STR markers are very labour-intensive to genotype. In most linkage studies about 3-400 STR were used, which on average are about 8-10 cM distance from each other. As the average haplotype blocks in the human genome are between 1-100 kb, there is a great chance in these studies that the markers miss the mutations. To overcome some of these limitations, the so-called **founder populations** were often used in these studies. Founder populations have been founded by small number of individuals, and because of geographical or cultural isolation, they have married within the population, and thus they are genetically relative close to each other (inbred vs. outbred populations). This has resulted that the individuals are separated by only a few meiosis (see founder effect: https://en.wikipedia.org/wiki/Founder_effect), and these populations have larger haplotype blocks, and in genetic studies the chances are significantly higher that the markers are in LD with the disease-causing variation. The French Canadians of Quebec, the Finnish, the Icelandic or populations in small islands are classical examples of founder populations. The founder effects can arise from cultural isolation. For example, the Amish or Hutterite populations in the United States, the gipsy populations in Hungary or the Ashkenazi Jews exhibit founder effects.

With the development of the genomic methods using densely located markers (like GWAS or NGS), the founder populations have become less significant in these studies.

12.1.6. Association studies

Association studies are the most popular population genetic methods to clarify the genetic background of complex traits. In these studies usually two populations are compared, one with the QT (e.g. disease), and one without it. These are the **case-control studies**. Genetic markers are genotyped and calculated, whether these markers are associated with the disease or not.

The association is a statistical term. In frequency-based methods, if a marker is associated with the disease, it means that it occurs significantly more frequently or less frequently in the cases. The association can have different meanings:

1. Direct effect: the marker allele has direct effect on the disease.
2. Natural selection: the marker allele increases the chance of survival of the disease.
3. False association due to population stratification (see below).
4. Statistical error (type I error, see in Chapter 9).

5. The marker allele is in LD with the variation, which directly influences the phenotype.

When the results are evaluated, all these points must be considered for a reliable conclusion.

In association studies **population stratification** (https://en.wikipedia.org/wiki/Population_stratification) is the presence of a systematic difference in allele frequencies between cases and controls possibly due to different ancestry. Population stratification can be a problem because the association found could be due to the underlying structure of the population and not a disease associated locus. To take a classic example, a GWAS for skill with chopsticks carried out in San Francisco might identify human leukocyte antigen A1 (*HLA-A1*) as an allele associated with chopstick skill, simply because this allele is more common in people of East Asian origin. Also the real disease causing locus might not be found in the study if the locus is less prevalent in the population, where the case subjects are chosen (http://www.nature.com/nrg/journal/v14/n1/full/nrg3382.html?WT.ec_id=NRG-201301).

For this reason, it was common in the 1990s to use family-based data where the effect of population stratification can easily be controlled for using methods such as the **transmission disequilibrium test**. In this test family trios are used (two parents and the affected child), and the over-transmission of an allele from heterozygous parents to affected offspring is investigated. If the allele has a role in the disease, it is transmitted with greater probability to the offspring. For an allele only those parents may be involved in the analysis who are heterozygotes to the allele.

In our globalized world several populations can live next to each other (like e.g. in the USA), and it is not so easy to collect case-control populations with balanced ethnicity, and the unbalanced populations can lead to both type I and type II errors. But if the structure is known or a putative structure is found, there are a number of possible ways to implement this structure in the association studies and thus compensate for any population bias. Most contemporary genome-wide association studies take the view that the problem of population stratification is manageable, and that the logistic advantages of using unrelated cases and controls make these studies preferable to family-based association studies. These methods can also handle the situation, when the populations mixed at genomic level (like black Americans with the Caucasians). This is called **population admixture**. E.g. the average African-American genome, for example, is 73.2% African, 24% European, and 0.8% Native American and 4% of European Americans carry African ancestry. Latinos, meanwhile, carry an average of 18% Native American ancestry, 65.1% European ancestry (mostly from the Iberian

Peninsula), and 6.2% African ancestry (<http://www.sciencemag.org/news/2014/12/genetic-study-reveals-surprising-ancestry-many-americans>).

This phenomenon can also be utilized by a genomic method, called **admixture mapping**.

In association studies, especially in GWAS, **the associated allele is frequently in LD** with the disease causing locus and not a causative variant. Here, the next task is to identify the responsible locus. The first step here is usually sequencing, but in complex diseases, where the responsible allele is frequent, **in silico methods** can be used. For this, databases like [dbSNP](#) can be used, and the linkage can be analyzed by the [Haploview](#) software. If there is a suspected allele, its possible function in the disease must be established with *in vitro* and *in vivo* methods. Usually it is far from simple; often it is the most difficult task. If it is not in the coding region of a gene, or does not change an amino acid code, then the task is still more difficult. Next to the wet labor experiments, the results of the [ENCODE](#) project and several predicting software products can also help in this job.

12.1.7. Risk calculation

In association studies there are values which show the robustness of the association. One of these values is the *p*-value, which is the probability that the null hypothesis is true. In association studies null hypothesis is that there is no association. The null hypothesis is rejected when the *p*-value is less than the significance level α , which is usually 0.05. When the null hypothesis is rejected ($p < 0.05$), the result is said to be statistically significant.

In retrospective studies the **odds ratio (OR)** is used for the estimation of the risk. This value has a direct connection with the *p*-value. **The odds ratio is the ratio of the odds of the association of an allele with the trait (disease) in cases to the odds of it in the control group.**

In prospective studies and in clinical trials the **relative risk (RR)** is used. **Relative risk is a ratio of the probability of the association in the case group versus a control group.**

If these values are greater than 1, then the risk is elevated, if they are <1, it is lower. Next to these values, their 95% confidence intervals (**95%CI**) must be given, which depends mainly on the population size. The larger the population, the narrower the 95%CI is. For statistically significant association both values of the 95%CI must be above 1, if the OR, or RR is >1, and below 1, if these values <1. E.g. OR = 2.2 (95%CI 1.3-3.9) is significant, OR = 2.2 (95%CI 0.9-4.5) is not significant. If the p-value < 0.05, then the OR is significant.

12.2. Evolutionary genetics

The modern evolutionary genetics originates from the synthesis of Darwinian evolutionary theories, the genetics and the modern molecular biology. Here we discuss it only from human and medical points of view and mainly concentrate on factors playing a role during the evolution of the modern human genome (https://en.wikipedia.org/wiki/Modern_evolutionary_synthesis).

12.2.1. Gene environmental interactions and the human genome

12.2.1.1. Natural selection

The modern human genome was developed through its constant interaction with the environment. The following selection mechanisms have contributed to this process (see https://en.wikipedia.org/wiki/Natural_selection):

- **Directional selection** occurs when a certain allele has a greater fitness than others, resulting in an increase of its frequency.
- **Stabilizing selection** lowers the frequency of alleles that have a deleterious effect on the phenotype – that is, produce organisms of lower fitness.
- **Purifying selection** results in functional genetic features, such as protein-coding genes or regulatory sequences, being conserved over time due to selective pressure against deleterious variants.
- **Balancing selection** does not result in fixation, but maintains an allele at intermediate frequencies in a population. E.g. **heterozygote advantage** when the heterozygote state can provide an advantage in a certain environment.

The human species has a special situation in the animal world. According to a theory of Eva Jablonka the human genome is able to tolerate potentially toxic variants thanks to clothing, tools, agriculture and other cultural innovations that allow individuals with these variants to survive. The human genome has accumulated more than its fair share of potentially harmful genetic changes — in protein coding regions, promoters and even the loss of entire genes. The relaxed selection created by human culture allowed the evolution of more diversity and complexity, but it has also made humans more reliant on the innovations that freed them from selection.

12.2.1.2. Role of infections in formation of the genome

The different microorganisms and **infections** have one of the largest roles in the formation of human genome. This selection factor played even important role in the history of the modern human (see e.g. https://en.wikipedia.org/wiki/Population_history_of_indigenous_peoples_of_the_Americas), and its effect can be felt even today. Think about the large epidemics like cholera, pest, influenza, pox, TBC, etc., which frequently decimated the population. Individuals who contracted one of these infections often died and were not able to pass their genome to the next generation. In contrast, there were individuals who were resistant or survived and could pass their genomes which gave them greater fitness. The today population is the descendant of those who survived all the infections, and were able to pass their genome to the next generation. Naturally, not only the genome influences how individuals respond to an infection, but several other factors as well, like the actual physical state, other infections, age, epigenetic states, pure chance, etc.

During the last years several traces of these microorganism–human genome interactions could be detected in the human genome. E.g.:

- 145 human genes originate from bacteria, through **horizontal gene transfer**.
- 8% of our genome originates from retroviruses. These two points are examples of **gene flow**.
- There are human genes specifically against bacterial or viral infections, like genes for pattern recognition receptors like *TLRs*, *MBL*, *CD14*, *NOD2* etc., or for antiviral proteins like *APOBC3G*, *TRIM5*, *BST2*.

The microorganisms render selection pressure on human genome even today. In a GWAS carried out in 52 populations, 441 variants mapped to 139 human genes were identified significantly associated with virus-diversity. Analysis of functional relationships among genes subjected to virus-driven selective pressure identified a complex network enriched in viral products-interacting proteins (<http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000849>).

12.2.1.3. Genetic drift

The genetic composition of a population is also influenced by random effects. Such an effect is e.g. when a small group in a population splinters off from the original population and forms a new one. The random sample of alleles in the just formed new colony is expected to grossly misrepresent the original population in at least some respects. It is even possible that the

number of alleles for some genes in the original population is larger than the number of gene copies in the founders, making complete representation impossible. When a newly formed colony is small, its founders can strongly affect the population's genetic make-up far into the future. This is a type of the **bottleneck effect**.

Another random event can be, when an individual (e.g. a king) has disproportionately more descendants than others in a small population, and his alleles become more frequent, or a war, or a natural disaster kill the majority of the population. This is called genetic drift and is different from the natural selection.

The genetic drift can have medical consequences, when an allele which became widespread in an environment, but can make susceptibility to a disease in another.

From evolutionary genetic point of view **population genetics** is the study of allele frequency distribution and change under the influence of the evolutionary processes: natural selection, genetic drift, mutation and gene flow.

12.2.2. *Why are some lethal mutations frequent?*

According to the Darwinian Theory and logic, the frequencies of deleterious mutations or mutations which produce organisms of lower fitness should decrease from generation to generation. But some deleterious mutations causing serious diseases are very frequent in certain populations. How is it possible?

The explanation is the **balancing selection, or heterozygote advantage**, which describes the case in which the heterozygote genotype has a higher relative fitness than either the homozygote dominant or homozygote recessive genotype. Polymorphism can be maintained by selection favoring the heterozygote, and this mechanism is used to explain the occurrence of some kinds of genetic variability. A common example is the case where the heterozygote conveys both advantages and disadvantages, while both homozygotes convey a disadvantage.

One of the most well-known examples in the Caucasian population is the [cystic fibrosis](#). This is the most common autosomal recessive genetic disorder. Its population frequency is 1/2500-3000, and the prevalence of mutation carriers is 1/28. CF is caused by a mutation in the gene for the protein cystic fibrosis transmembrane conductance regulator (CFTR). This protein is required to regulate the components of sweat, digestive juices, and mucus. CFTR regulates the movement of chloride and sodium ions across epithelial membranes, such as the alveolar epithelia located in the lungs. Although most people without CF have two working copies of

the CFTR gene, only one is needed to prevent cystic fibrosis due to the disorder's recessive nature.

The ΔF508 mutation (deletion of 3 bases from the gene, causing deletion of the phenylalanine 508 from the protein), is the most frequent cause of the disease, the majority of the affected carries this mutation. The symptoms are very serious already in early childhood. In 1959, the median age of survival of children with cystic fibrosis in the USA was six months. The survival time improved significantly, but males are often infertile due to congenital absence of the vas deferens. Thus, it is exactly the disease type, in which the mutation is lethal and according to the natural selection the frequency of the mutation should be lowered from generation to generation.

The ΔF508 mutation is estimated to be up to 52,000 years old. Numerous hypotheses have been advanced as to why such a lethal mutation has persisted and spread in the human population. The following hypotheses have been proposed as possible sources of heterozygote advantage (https://en.wikipedia.org/wiki/Cystic_fibrosis):

- Cholera: With the discovery that cholera toxin requires normal host CFTR proteins to function properly, it was hypothesized that carriers of mutant CFTR genes benefited from resistance to cholera and other causes of diarrhea.
- Typhoid: Normal CFTR proteins are also essential for the entry of *Salmonella Typhi* into cells, suggesting that carriers of mutant CFTR genes might be resistant to typhoid fever. No *in vivo* study has yet confirmed this. In both cases, the low level of cystic fibrosis outside of Europe, in places where both cholera and typhoid fever are endemic, is not immediately explicable.
- Diarrhea: It has also been hypothesized that the prevalence of CF in Europe might be connected with the development of cattle domestication. In this hypothesis, carriers of a single mutant CFTR chromosome had some protection from diarrhea caused by lactose intolerance, prior to the appearance of the mutations that created lactose tolerance.
- Tuberculosis: Another possible explanation is that carriers of the gene could have some resistance to TB.

The heterozygote advantage is easier to be explained in **sickle cell anemia**. This is also an autosomal recessive disease, which is very frequent in individuals of African origin, especially in people (or their descendants) from parts of tropical and sub-tropical regions where malaria is or was common. In the USA the prevalence of the disease is 1/600 in black,

and the frequency of carriers is 1/12. There are certain symptoms, however, even in heterozygotes. Five percent of carriers have blood in their urine, and have symptoms if they are deprived of oxygen (for example, while climbing a mountain). The main cause of the disease is a mutation in the gene that produces beta-globin, a protein needed to produce normal haemoglobin. This is a Glu6Val mutation and the gene product is called **haemoglobin S**.

Haemoglobin S ([sickle cell trait](#)) provides a survival advantage over people with normal haemoglobin in regions where malaria is endemic. The trait is known to cause significantly fewer deaths due to malaria, especially when *Plasmodium falciparum* is the causative organism. This is a prime example of natural selection, evident by the fact that the geographical distribution of the gene (for haemoglobin S) and the distribution of malaria in Africa virtually overlap. Because of the unique survival advantage, people with the trait increase in number as more people infected with malaria and having the normal haemoglobin tend to succumb to the complications.

Although the precise mechanism for this phenomenon is not known, several factors are believed to be responsible.

- Infected erythrocytes tend to have lower oxygen tension, because it is significantly reduced by the parasite. This causes sickling of that particular erythrocyte, signalling the phagocytes to get rid of the cell and hence the parasite within.
- Since the sickling of parasite infected cells is higher, these selectively get removed by the reticulo-endothelial system, thus sparing the normal erythrocytes.
- Excessive vacuole formation occurs in those parasites infecting sickle cells.
- Sickle trait erythrocytes produce higher levels of the superoxide anion and hydrogen peroxide than do normal erythrocytes; both are toxic to malarial parasites.

The sickle cell trait was found to be 50% protective against mild clinical malaria, 75% protective against admission to the hospital for malaria, and almost 90% protective against severe or complicated malaria.

Another example of the selection advantage is the frequent **CCR5Δ32 mutation** in the Caucasian population. The CCR5 is a chemokine receptor, and the 32 bp deletion does not cause any significant symptoms in humans, although the resulted protein product is

functionless. But, as CCR5 is an essential co-receptor for the HIV-1 to enter the cells, the lack of it protects individuals from the infection. In the European population the frequency of the Δ32 mutation is very high, 1/100 individuals homozygote for it, and thus protected from HIV-1 and AIDS. Even heterozygotes have some advantage. Although they can be infected by the virus, but the AIDS disease developed significantly slower in them (2-4 vs. 6-8 years, in untreated people).

Interestingly, the mutation occurs only in people with European ancestry. According to researches the mutation appeared 7000 (2900-15750) years ago. Possibly there was an epidemic at that time in this population, in which the pathogen used the same receptor for the infection. So far, this infection has not yet been identified, but there are some suspects like pest and pox. In 2012 it was identified the CCR5 as a cellular determinant required for cytotoxic targeting of subsets of myeloid cells and T lymphocytes by the *Staphylococcus aureus* leukotoxin ED (LukED). CCR5-deficient mice are largely resistant to lethal *S. aureus* infection thus this finding put forth the possibility that resistance to *S. aureus* leukotoxins may have influenced the selection of the Δ32 allele (<http://www.ncbi.nlm.nih.gov/pubmed/23235831>).

12.2.3. Examples for effects forming the genome

In the last years more and more people have been sequenced in different populations. It gives the possibility to compare their genomes and detect population-specific variations, some of them developed through population-specific gene-environmental interactions. Let us see some examples!

One of the most marked differences between populations may be the **skin color**. The environmental factor which induced the differences was the **sunlight**. Individuals from populations lived in sunny environment have darker skin, those who experienced less sunlight have lighter skin. The main cause of the selection pressure in the sunlight is the UV radiation. It can cause DNA mutations and cancer in the skin (melanoma). The best defense against it is the melanin produced by melanocytes, which gives the skin dark color. The other selection pressure is the **vitamin D** produced in the skin when exposed to sunlight. But high melanin content decreases this process. Two genes *SLC24A5* and *SLC45A2* were identified as major determinants of pigmentation in humans and in other vertebrates. The allele A111T in the former gene and the allele L374F in the latter gene are both nearly fixed in light-skinned Europeans, and can therefore be considered ancestry informative marker. An L374F substitution in *SLC45A2* was found at 100% frequency in a European sample, but was absent

in Asian and African samples. An association study has shown that the Phenylalanine-encoding allele is correlated with fair skin and non-black hair in Europeans. The modern humans who came out of Africa to originally settle Europe about 40,000 years ago are presumed to have had dark skin, which is advantageous in sunny latitudes. And the new data confirm that about 8500 years ago, early hunter-gatherers in Spain, Luxembourg, and Hungary also had darker skin: They lacked versions of two genes—*SLC24A5* and *SLC45A2*—that lead to depigmentation and, therefore, pale skin in Europeans today. But in the far north—where low light levels would favor pale skin—the team found a different picture in hunter-gatherers: Seven people from a 7700-year-old archaeological site in southern Sweden had both light skin gene variants, *SLC24A5* and *SLC45A2*. They also had a third gene, *HERC2/OCA2*, which causes blue eyes and may also contribute to light skin and blond hair.

There are examples even today for the selection pressure of the sunlight. The prevalence of melanoma has been increasing in light-skinned populations due to excess exposure to sunlight (sunbathe), and 100% of black-skinned individuals of African origin, 93% of Indians and 85% of East-Asians living in Canada suffer from vitamin D deficiency. The health consequences can hit pregnant women particularly hard. In a large analysis of vitamin D levels in 50,000 pregnant women in the United States in the 1960s, it was found that low vitamin D was common in blacks—but not whites—and was associated with preterm birth.

Available food can also be a selection factor. One example is the *AMY1* gene coding for salivary amylase, which digests starch. *AMY1* is one of the few genes in the human genome that show extensive copy-number variation between individuals. Extra *AMY1* copies endow the individuals carrying them with the capacity to produce more salivary amylase. In a study two groups were investigated: one consisted of four populations with a low-starch diet and the other of three populations from agricultural societies and hunter-gatherers in arid environments, who traditionally eat high-starch food. Strikingly, twice as many members of the high-starch-diet group had at least six copies of *AMY1*. This difference could not be explained by geographical factors, because both groups contained people of Asian and African origin. Instead, the authors propose that variations in *AMY1* copy number are more likely to have been influenced by positive natural selection. So what is the advantage of having more salivary amylase? Significant digestion of starch occurs during chewing. This is crucial, and probably vital, in people likely to suffer from diarrhoeal diseases. Moreover, after being swallowed, salivary amylase is carried to the stomach and intestines, where it aids other digestive enzymes. Of the three copies of the *AMY1* gene registered in the reference sequence

of the human genome, variations in nucleotide sequences are small. This suggests that the duplication of these genes may have occurred relatively recently, possibly even since the evolution of modern humans about 200,000 years ago (<http://www.nature.com/nature/journal/v449/n7159/full/449155a.html>).

Siberia, where local temperatures occasionally drop below -70°C in the winter and only animals are available for consumption for much of the year, is one of the most extreme habitats human populations have adapted to since their dispersal out of Africa. With agriculture being unsustainable in this part of the world as a result of its extremely cold environment, these coastal populations mostly fed on marine mammals for **a high-fat diet rich** in n-3 polyenoic fatty acids. Such a diet **would have led the populations to be in a permanent state of ketosis**, and where metabolism is mainly “lipocentric” (ketone bodies, fatty acids) rather than “glucocentric” (glucose), as found in a high-carbohydrate diet. In a genome-wide SNP genotype study (<http://www.sciencedirect.com/science/article/pii/S0002929714004224>) of 200 Siberian individuals, the strongest signals of positive selection detected by tests for haplotype homozygosity and allele differentiation mapped to a 3 Mb region containing 79 protein-coding genes at chr11: 66–69 Mb in Northeast Siberian populations. Then the genomes of 25 unrelated individuals from the Chukchi, Eskimo, and Koryak populations were sequenced. A causative variant as a nonsynonymous G>A transition (rs80356779; c.1436C>T [p.Pro479Leu] on the reverse strand) was identified **in CPT1A gene, a key regulator of mitochondrial long-chain fatty-acid oxidation**. The derived allele is associated with hypoketotic hypoglycemia and high infant mortality in people leaving outside the arctic area or eating “more healthy diet”, yet occurs at high frequency in Canadian and Greenland Inuits and was also found at 68% frequency in our Northeast Siberian sample. The c.1436C>T mutation might have conferred a metabolic advantage for the Northeast Siberian populations in dealing with their traditional high-fat diet. The deleterious effect of the mutation might be explained by a change from the traditional diet to a more carbohydrate-based one or by recent cultural shifts and environmental stressors such as fasting and pathogens. This study also proves that sometimes it can have serious health consequences if people live at a different location or eat different food than their ancestors and useful genetic variations in a certain environment can be harmful in another.

Prehistoric and contemporary human populations living at altitudes of at least 2,500 meters above sea level may provide unique insights into human evolution (http://news.nationalgeographic.com/news/2004/02/0224_040225_evolution.html). Indigenous highlanders living at **high altitude have evolved different biological adaptations for surviving** in the oxygen-thin air.

The Andeans adapted to the thin air by developing an ability to carry more oxygen in each red blood cell by **having higher haemoglobin concentrations** in their blood. Tibetans compensate for low oxygen content much differently. They increase their oxygen intake by taking **more breaths** per minute than people who live at sea level. In addition, Tibetans may have a second biological adaptation, which expands their blood vessels, allowing them to deliver oxygen throughout their bodies more effectively than sea-level people do. Tibetans' lungs synthesize larger amounts of nitric oxide from the air they breathe. One effect of **nitric oxide** is to increase the diameter of blood vessels, which suggests that Tibetans may offset low oxygen content in their blood with increased blood flow.

To pinpoint the genetic variants underlying Tibetans' relatively low haemoglobin levels, researchers collected blood samples from nearly 200 Tibetan villagers living in three regions high in the Himalayas. When they compared the Tibetans' DNA with their lowland counterparts in China, their results pointed to the same culprit — a gene on chromosome 2, called ***EPAS1*, involved in red blood cell production and haemoglobin concentration** in the blood. While all humans have the *EPAS1* gene, Tibetans carry a **special version of the gene**. Over evolutionary time individuals who inherited this variant were better able to survive and passed it on to their children, until eventually it became more common in the population as a whole (<http://www.livescience.com/6543-revealed-tibetans-survive-thin-air.html>).

Next to the Tibetan population, adaptation to high altitude have also been reported for highland Andean and Ethiopian populations. One of the important findings that has emerged from this combined work is that although the biological pathway implicated in each continental region is consistent, the specific loci that have been identified in high-altitude populations are often not consistent, which indicates convergent adaptation among different geographical regions.

Lactose intolerance is the inability to digest lactose, a sugar found in milk and to a lesser extent milk-derived dairy products. Most mammals normally become lactose intolerant after weaning, but some human populations have developed lactase persistence, in which lactase production continues into adulthood. This means that the **lactose intolerance may be regarded as the ancient “wild type” phenotype**. It is estimated that 75% of adults worldwide show some decrease in lactase activity during adulthood. The frequency of decreased lactase activity ranges from 5% in northern Europe through 71% for Sicily to more than 90% in some African and Asian countries. This distribution is now thought to have been caused by **recent natural selection favoring lactase persistent** individuals in cultures that rely on dairy products. While it was first thought that this would mean that populations in

Europe, India, and Africa had high frequencies of lactase persistence because of a particular mutation, it has now been shown that lactase persistence is caused by **several independently occurring mutations**. These last two examples are examples for **convergent evolution**, which means that different processes in different population lead to similar phenotypes.

Often, different traits can be developed in individuals, which are **only side-effects** of the changes induced by natural selection. One of the reasons of this is that **most of these genes are pleiotropic**: that is, they are individually involved in several different traits. For example, ***EDAR*** regulates hair follicle density and the development of sweat glands and teeth. In humans, selective pressures on *EDAR* favoring changes in body temperature regulation and hair follicle density in response to colder climates may have influenced tooth shape, although this trait probably does not affect population fitness.

An additional possible mechanism of a trait that has no apparent reason is that allele causing a phenotypic feature can be in **LD with the selected variant**, can be on the same haplotype and thus inherited together with it.

Bacteria can acquire mutations or genes which are advantageous for their survival through horizontal gene transfer, e.g. genes for antibiotic resistance. Recent data provided evidence that 1%–6% of modern Eurasian genomes were inherited from ancient hominins, such as Neandertals or Denisovans, with specific genomic regions presenting up to 64% of Neandertal ancestry. In the context of immunity, there is increasing evidence to suggest that modern humans have acquired advantageous variations through admixture with ancient hominins for *HLA* class I genes, *STAT2*, or the *OAS* gene cluster. In modern humans it was shown that archaic people contributed more than half of the alleles that code for proteins made by the human leukocyte antigen system (HLA), which helps the immune system to recognize pathogens. Thus, it seems that **archaic genome contributed to modern human genome and selection fitness through horizontal gene transfer**.

12.3. Literature

1. International HapMap 3 Consortium, Altshuler DM et al. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010;467(7311):52-8.
2. International HapMap Consortium, Frazer KA et al. A second generation human haplotype map of over 3.1 million SNPs. *Nature*. 2007;449(7164):851-61.

3. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;7;447:661-78.
4. Pennisi E. 1000 Genomes Project Gives New Map Of Genetic Diversity. *Science* 2010; 330: 574-5.)
5. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet*. 2007 Dec;81(6):1278-83.
6. <https://www.23andMe.com/>
7. Kaye J. The regulation of direct-to-consumer genetic tests. *Hum Mol Genet*. 2008;17:180-3.
8. Allayee H, Ghazalpour A, Lusis AJ. Using mice to dissect genetic factors in atherosclerosis. *Arterioscler Thromb Vasc Biol*. 2003 Sep 1;23(9):1501-9.
9. Mehrabian M et al. Identification of 5-lipoxygenase as a major gene contributing to atherosclerosis susceptibility in mice. *Circ Res*. 2002 Jul 26;91(2):120-6.
10. Rapp JP. Genetic analysis of inherited hypertension in the rat. *Physiol Rev*. 2000 Jan;80(1):135-72.
11. ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012 Sep 6;489(7414):57-74.

12.4. Questions

1. What is population genetics?
2. What types of sample collection methods do you know? Give some examples!
3. What strategies do you know for the selection of patients in retrospective studies?
What are the advantages and disadvantages?
4. What does endophenotypes mean?
5. What is HWE?
6. What can be the causes for the deviation from the HWE?
7. What does linkage disequilibrium mean?
8. What measures for LD do you know?
9. What does haplotype mean?
10. What is cM?
11. What can be the causes for genetic linkage?
12. What are the founder populations and how can they be used in genetic studies?

13. In population genetic studies what can be the cause of an association?
14. What does population stratification mean?
15. What methods do you know for the control of the problem of population stratification?
16. What does the population admixture mean in population genetics?
17. What values can be used for the estimation of risk in association studies?
18. What is the evolutionary genetics?
19. What selection processes contributed to the development of the human genome?
20. Give examples for the microorganism–human genome interactions!
21. What is the genetic drift?
22. What is the bottleneck effect?
23. Why are some lethal mutations frequent?
24. Why is $\Delta F508$ mutation frequent?
25. Why is the sickle cell anemia frequent in certain populations?
26. What mutation is protective against AIDS?
27. Give examples for effects forming the genome!
28. Give examples for the selection pressure of the sunlight today!
29. Give an example for the selection pressure of the available food!
30. Give an example for the selection pressure of the life in high altitude!
31. What do you know about the genetic background of lactose intolerance?
32. What does convergent evolution mean?
33. What can be the reason that sometimes traits may appear which have nothing to do with the selection pressure?
34. What role did horizontal gene transfer play in the development of the human genome?

13. Gene environmental interaction

Csaba Szalai

In the previous chapter some examples were shown how environmental factors have played roles in the formation of the human genome. In this chapter examples will be shown how genetic variations can influence the effect of environmental factors on their carriers.

13.1. Penetrance of the genetic variants

According to the definition **penetrance** in genetics is the proportion of individuals carrying a particular variant of a gene (allele or genotype) that also express an associated trait (phenotype). In medical genetics, the penetrance of a disease-causing variation is the proportion of individuals with the variation who exhibit clinical symptoms. For example, if a variation in the gene responsible for a particular autosomal dominant disorder has 95% penetrance, then 95% of those with the variation will develop the disease, while 5% will not. From gene-environmental point of view genetic variants can be:

- **Highly penetrant.** A variant is highly penetrant, when the trait it produces will almost always be apparent in an individual carrying the allele. The variant significantly influences the function of a gene or its products. These are e.g. the monogenic disease causing variants.
- **Low penetrant.** A variant has low penetrance, when only sometimes produce perceptible phenotype. The variant does not influence significantly the function of the gene or its products, or the intact gene is not essential for the normal phenotype. But in interactions with other variants or environmental factors its effect may be apparent. E.g. frequent variants can belong to this class.

Naturally, the transition between the two types of variants is continuous. The genomic background of an individual strongly influences his/her respond to the environmental stimuli. Individuals, who are hypersensitive to an environmental stimulus, usually have monogenic diseases, because normal amount of environmental factors can cause a disease. Like people with cutaneous **porphyria**, who are extreme light-sensitive, and develop symptoms like blisters, necrosis of the skin and gums, itching and swelling in response to normal sunlight. In

these people deficiency in the enzymes of the porphyrin pathway leads to insufficient production of haeme. But among people without deleterious, highly penetrant genetic variations there are light-sensitive (light-skinned), and light resistant (dark-skinned), and their population distribution is usually continuous and normal (Gaussian) (Figure 13.1).

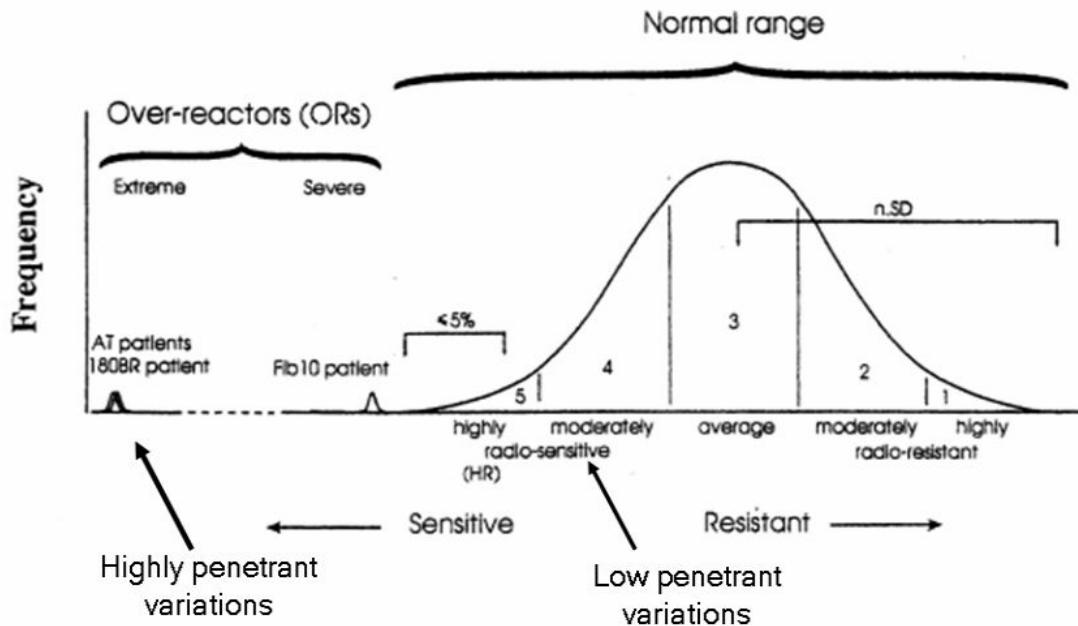


Figure 13.1. Population distribution of certain genotypes influencing the respond to environmental stimuli (here radiation). AT: ataxia telangiectasia (people with A-T have enhanced sensitivity to radiation, disposition to lymphoma and leukaemia)

Source: <https://books.google.hu/books?id=eiL4CAAAQBAJ&pg=PA66&lpg=PA66&dq#v=onepage&q&f=false>

13.2. Interactions between highly penetrant variations and the environment

Besides porphyria there are several examples where individuals are hypersensitive to certain environmental stimuli, but if they could avoid these factors they can prevent, or reduce the more serious symptoms. E.g. xeroderma pigmentosum is an autosomal recessive genetic disorder of DNA repair, in which the ability to repair damage caused by UV light is deficient. In extreme cases, all exposure to sunlight must be forbidden, no matter how small; as such, individuals with the disease are often colloquially referred to as Children of the Night. Multiple basal cell carcinomas and other skin malignancies frequently occur at a young age in those with the disease.

Much more frequent is the phenylketonuria (PKU) characterized by a mutation in the gene for the hepatic enzyme phenylalanine hydroxylase (PAH), rendering it nonfunctional. This enzyme is necessary to metabolize the amino acid phenylalanine to the amino acid tyrosine.

When PAH activity is reduced, phenylalanine accumulates and is converted into phenylpyruvate. Untreated PKU can lead to mental retardation, seizures, and other serious medical problems. The mainstream treatment for classic PKU patients is a strict phenylalanine-restricted diet supplemented by a medical formula containing amino acids and other nutrients. The current recommendation is that the PKU diet should be maintained for life. Patients who are diagnosed early and maintain a strict diet can have a normal life span with normal mental development. Because it is quite frequent (1/10,000 births) it is commonly included in the newborn screening panel of most countries.

Familial hypercholesterolemia (FH) is one of the most frequent autosomal diseases with co-dominant inheritance, occurring in 1:500 people in most countries; homozygous FH is much rarer, occurring in 1 in a million births. It is caused by mutation in the low density lipoprotein receptor (**LDLR**). In heterozygote individuals the symptoms of the disease (high LDL-cholesterol serum level, early atherosclerosis, myocardial infarction) can be reduced or prevented with specific low cholesterol diets and drugs (e.g. statins). In homozygotes the symptoms are much more serious, and the treatment is more difficult.

13.3. Examples for interactions between low penetrant variations and environment

Alpha 1-antitrypsin deficiency is caused by variation in the *SERPINA1* gene (earlier A1AT) that causes defective production of alpha 1-antitrypsin. It is a co-dominant trait caused most often by a common polymorphism called Z mutation (Glu342Lys; rs28929474). Alpha-1 antitrypsin deficiency occurs worldwide, but its prevalence varies by population. This disorder affects about 1 in 1,500 to 3,500 individuals with European ancestry. The carriers (2-5%) are sensitive to cigarette smoke and other air-borne particles leading to lung emphysema, COPD, or asthma.

Factor V Leiden, caused by the Leiden mutation in F5 gene (R506Q). In this disorder, the Leiden variant of factor V of the coagulation system cannot be inactivated by activated protein C. Factor V Leiden is the most common hereditary hypercoagulability disorder amongst Eurasians. It is named after the city Leiden (Netherlands), where it was first identified in 1994. It is very common; its carrier frequency is 6.5%.

The variation prevents efficient inactivation of factor V. When factor V remains active, it facilitates overproduction of thrombin leading to generation of excess fibrin and excess clotting. The carriers have usually no symptoms, but oral contraceptives, long physical inactivity (air travel) or pregnancy may increase the risk for development of deep vein thrombosis.

Prothrombin G20210A mutation is in F2 gene of the coagulation system. Its frequency is 1-3%. It is on the same pathway as the previous one, and participates in the same gene-environmental interactions. The two variations have synergistic effect. If they occur together, the chance for deep vein thrombosis is 2.6 times greater than in people with Leiden mutation alone.

Hereditary haemochromatosis. It is the most common cause of iron overload. Excess iron accumulates in tissues and organs disrupting their normal function. The most susceptible organs include the liver, adrenal glands, heart, skin, gonads, joints, and the pancreas; patients can present with cirrhosis, polyarthropathy, adrenal insufficiency, heart failure or diabetes. The hereditary form of the disease is most common among those of Northern European ancestry, in particular those of Celtic descent. Its cause is variations in the HFE gene. 80-100% of patients are homozygote to C282Y variation, and the H63D variation is also a risk factor in C282Y carriers. In population with Celtic ancestry the homozygote frequency of C282Y is very high (1/100-1/300), and it is also quite high in other Caucasian population. But the frequency of the disease is significantly lower, which shows that other factors are necessary for the development of the disease. These factors include alcoholism (causing damage to the liver), diet with high C vitamin content (increases the absorption of iron), red meat (high iron content). There are risk reducing factors, like substances that inhibit iron absorption, such as high-tannin tea, calcium, and foods containing oxalic and phytic acids (such as collard greens, which must be consumed at the same time as the iron-containing foods in order to be effective).

13.4. Smoking-genome interaction

The smoking is one of the strongest and well-measurable environmental factors, and there are several genetic and genomic results about its interaction with the genome. This interaction can be studied in two different aspects. First, it may sound surprising, but smoking can be regarded as a **complex, multifactorial disease**, similarly to drug abuse and alcoholism. On

the other hand, it is well-known that **it increases the risk of several diseases**, like lung cancer, asthma, COPD, atherosclerosis or Alzheimer-disease. These diseases do not develop in all smokers, meaning that other factors are necessary for the development of these diseases. The strongest among these factors is the genomic background of the smokers.

13.4.1. Genomic background of smoking

The heritability of smoking is higher than that of several other polygenic diseases, it is 60%. It is well known that in smokers stress induces cigarette-craving. In a study it was investigated, whether this had a genetic background. Significantly stronger stress-induced cigarette craving was found for individuals carrying either the ***DRD2* (D2 dopamine receptor gene)** A1, or the ***SLC6A3* (dopamine transporter gene)** nine-repeat allelic variants. Stress-induced craving was markedly higher for those carrying both alleles, compared to those with neither, consistent with the separate biological pathways involved (receptor, transporter). These findings provide strong support for the possibility that the dopamine system is involved in stress-induced craving and suggest a potential genetic risk factor for persistent smoking behavior. This pathway plays also a role in drug abuse or alcoholism. Both allelic variants are associated with lower brain dopaminergic function, and these basal deficits, in turn, are thought to increase the incentive salience of drug use in the presence of triggers (e.g. stress) that might be related to acute phase increases in dopamine levels (<http://www.nature.com/tpi/journal/v4/n2/full/6500227a.html>).

Another important pathway that plays a role in the addiction to smoking is the **nicotine pathway**. The ***CYP2A6*** gene (19q13.2) codes for an enzyme responsible for the degradation of nicotine. **Deficiency of this enzyme is quite common** (10-17.6%), and causes reduced degradation of nicotine, and people with this deficiency have reduced possibility of smoking addiction. If they smoke cigarette, they smoke less, and have reduced risk to cancer and emphysema. The nicotine is accumulated in their organisms, their craving will be reduced more quickly, and thus less toxic substance from the smoke will get to their bodies.

Several GWAS were carried out to study the genomic background of smoking. In 2008 three independent GWAS identified a SNP (rs1051730) in a nicotinic receptor subunit gene, which associated with both smoking and lung cancer. There was a discussion, which one is the real association. Then, with the help of association studies it was verified that the genomic region (***CHRNA5-CHRNA3-CHRN B4, 15q24; CHRNA = neuronal acetylcholine receptor subunit***

alpha), where there are several nicotinic receptors was responsible for the association with strong smoking and the link with lung cancer is primarily mediated through the smoking-related phenotypes. When patients with lung cancer were taken out of the population, the association remained (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2614129/>).

The significance of the nicotinic receptors was shown by further studies, where additional **nicotinic receptor gene cluster (CHRNA6–CHRN B3)** on chromosome 8p11 was found to be **associated with smoking** and also with the quantity of cigarette smoked in a day.

In a Hungarian study a highly significant association between ever smoking (past + current smokers) and a specific MHC haplotype was observed. The 8.1 ancestral haplotype occurred more frequently in the ever smokers than in the never smokers [odds ratio: 4.97 (1.96-12.62); P = 0.001], and such associations were stronger in women (odds ratio = 13.6) than in men (odds ratio = 2.79). An independent study in Icelandic subjects (n = 351) yielded similar and confirmative results. Considering the documented link between olfactory stimuli and smoking in females, and the presence of a cluster of odorant receptor genes close to the MHC class I region, the findings implicate a potential role of the MHC-linked olfactory receptor genes in the initiation of smoking (<http://www.ncbi.nlm.nih.gov/pubmed/15339882>).

13.4.2. Smoking-gene interaction in disease susceptibilities

Previously we have mentioned that smoking in people with **alpha 1-antitrypsin deficiency** can trigger lung emphysema, COPD and asthma.

The product of the gene **GSTM1** glutathione S-transferase plays a role in the detoxification of electrophilic compounds, including carcinogens, therapeutic drugs, environmental toxins and products of oxidative stress, by conjugation with glutathione. Its null variant is very frequent (39%). This deficiency **in smokers** is associated with **increased risk to asthma and lung cancer**. Vitamin C and E can be protective.

Eighty-ninety percent of patients with **rheumatoid arthritis (RA)** have certain subtypes of HLA-DRB1: DRB1*0401, *0404, *0405, *0408, *0101, *0102, which are called shared epitopes (**HLA-(DRB1) SE**). Carriers of HLA-SE have an increased susceptibility to RA and this also has a prognostic significance. In RA, **smoking is the most important environmental risk factor**. In the last years it has been discovered that in RA patients **anti-CCP (anti-cyclic citrullinated peptide) auto-antibodies** can be detected. In **HLA-DRB1 SE** carriers smoking can lead to appearance of anti-CCP antibodies. It starts in the lung and years afterwards RA develops. Smoking activates the enzyme peptidylarginine deiminase which

catalyzes the conversion of arginine to citrulline in the proteins in the lung. The cigarette smoke functions as local adjuvant, which leads to the production of anti-CCP. The HLA-DRB1-SE variants bind and present citrullinated proteins especially well. Months or years later a mild inflammation in the joints can trigger appearances of citrullinated proteins. In individuals, who have high anti-CCP level it can lead to development of chronic RA. If a smoker is a HLA-SE allele carrier, his/her relative risk is 6.5, in two allele-carriers it is 21.

A **gene-gene-environmental interaction** can be observed in those who have null variant in the **GSTM1 gene, HLA-SE carriers, and smokers**. They have 58-fold risk to development of RA.

The **expected life time** of smokers is significantly lower than that of non-smokers. Individuals who are carriers of the *C4B*Q0*, an inactive variants of the complement *C4B* gene in the HLA region (6p21.3), have reduced life expectancy. The population frequency of this variant is 16% in young age (below 45), and reduces to 6% in people of 70-79 years of age. These findings were detected in Hungarian populations and confirmed in Icelandic; and showed that carriers of the *C4B*Q0* had a substantially increased risk to suffer from myocardial infarction or stroke, and were sorted out from the healthy elderly population. This was associated strongly with smoking both in Iceland and Hungary. The findings indicated that the *C4B*Q0* genotype could be considered as a major covariate of smoking in precipitating the risk for acute myocardial infarction and associated deaths (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1942025/>).

The ***CYP1A1*** gene belongs to the cytochrome P450 superfamily (CYP). Enzymes in this group catalyze the oxidation of organic substances and they are the **main detoxifying agents** in the organism. ***CYP1A1 degrades the toxins in the cigarette smoke***. The most frequent cancer in children is the acute lymphoid leukemia (ALL). Children whose parents are smokers have a significantly higher risk. In a study it was found that if the father smokes at home, then the risk is 1.8-fold. The variations in the *CYP1A1* gene alone do not influence the risk to ALL, but if the children are carriers of certain haplotypes of *CYP1A1* and their fathers smoke at home, then the risk is 2.8-fold, if the father is a strong smoker, the risk is 4.9-fold (<http://www.ncbi.nlm.nih.gov/pubmed/18691756>).

13.4.3. Smoking-gene interactions in complex diseases

Smoking increases the risk of several complex diseases. Let us see some examples, what genomic backgrounds could influence this risk.

Smoking is a risk factor in asthma. In a whole genome screening a hypothesis was tested, whether inclusion of exposure to environmental tobacco smoke (ETS) would improve the ability to map genes for **asthma** (<http://www.atsjournals.org/doi/full/10.1164/ajrccm.163.6.2001101#VsbffnhAdU>).

144 white families from the Collaborative Study for the Genetics of Asthma were screened by 323 microsatellites as genetic markers, and environmental information about exposure to ETS during infancy was incorporated in the study. Three regions showed a significant increase from the baseline LOD score (chromosome 1p between D1S1669-D1S1665 markers; 5q at D5S1505-D5S816; and 9q at D9S910). The highest LOD score was found on chromosome 5q. In this genomic region 3 candidate genes were found between the markers of D5S1505-D5S816: *ADRB2*, *IL4* and *IL13*. Among these the strongest candidate is the ***ADRB2*** which codes for the **adrenergic β2 receptor**, because it is expressed in the lung and binds substances from the cigarette smoke. The receptor has a common variant: Arg16Gly, which influences the amount of expressed receptors, and has pharmacogenetic significance (see in Chapter 14). In another study it was found that compared with never-smoking Gly-16 homozygotes, those **ever-smokers who are Arg-16 homozygotes had a significantly increased risk of asthma** (odds ratio = 7.81; 95% confidence interval [CI]: 2.07 to 29.5). This association showed a clear dose-response relationship with the number of cigarettes smoked.

The smoking increases the risk of **atherosclerosis and T2DM** (type 2 diabetes mellitus) as well. The ***CYP1A1*** gene has a polymorphism called MsPI (T6235C). The C allele is associated with a better inducible gene, its frequency is 10%, and it increases the risk to atherosclerosis and T2DM and higher rate of complications only in mild smokers. In heavy smokers the risk of these diseases are so high that the weak effect of this polymorphism could not be detected. This observation suggests that the presence of the rare C allele of the ***CYP1A1*** gene in smokers may enhance predisposition to severe CAD and T2DM (<http://www.sciencedirect.com/science/article/pii/S0021915001007237>).

The variants of the gene ***APOE*** (E2, E3, E4; figure 13.2) influence the susceptibility to several diseases, like Alzheimer disease, or atherosclerosis. The variants are quite frequent and differ from each other in their reduction potential, and affinity to lipoprotein receptors. The APOE4 has the lowest reduction potential, meaning that it can reduce least effectively the oxidative stress induced by smoking. In a study the highest levels of oxLDL and risk to atherosclerosis were measured in APOE4 smokers.

Carrying APOE4 is associated with high risk to Alzheimer disease, the same is true for smoking (OR = 4.93) but this risk is the highest in APOE4 smokers (OR = 6.56).

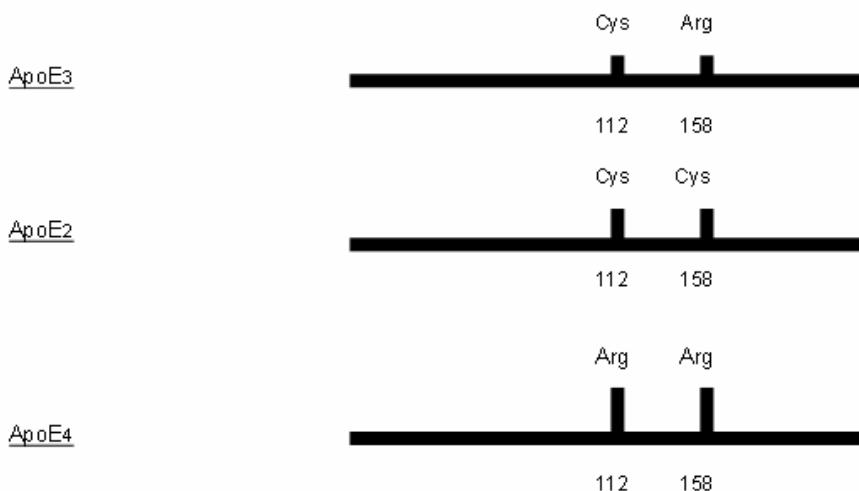


Figure 13.2. The three frequent variants of the apoE.

https://commons.wikimedia.org/wiki/File:ApoE_i_variants_alpha%2B71%C3%A8liques.png

13.5. Examples for gene-environmental interactions

The high frequency of the APOE4 allele (16%) can be explained by selection advantage (see also in chapter 11.13.3.). The E4 allele is associated with higher level of LDL-C and binds to LDL receptor with higher affinity. In populations where plant-based nutrition was the dominant (e.g. by gatherers, or in some islands), the individuals ate less cholesterol, and thus the APOE4 gave a selection advantage. And indeed, descendants of these populations have a higher frequency of this allele.

The *APOE* alleles influence the life expectancy as well. APOE2 is associated with the highest life expectancy, and the APOE4 with the lowest. Probably, APOE4 makes more sensitive to harmful environmental factors. It is often claimed that genes affecting health in old age, such as cardiovascular and Alzheimer diseases, are beyond the reach of natural selection. But in a population study a gradual increase with each generation of the E2 and E3 alleles of the gene at the expense of the E4 allele has been found. The E2 allele frequency was found to increase slightly more rapidly than that for E3 (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2848859/>).

It is well-known that high serum cholesterol levels can often be reduced through low-cholesterol diet, but this does not function in everybody. In a study it was found that among others it was influenced by the *APOE* genotypes. The carrier status of the E2 allele was associated with poor response, while that of the E4 with good response. This implies among others that certain alleles can have both positive and negative effects.

Leukotrienes are inflammatory mediators generated from arachidonic acid by the enzyme 5-lipoxygenase coded by the *ALOX5* gene on 10q11.2. Since atherosclerosis involves arterial inflammation, in a study it was investigated whether a repeat polymorphism in the *ALOX5* gene promoter could relate to atherosclerosis and that this effect could interact with the dietary intake of competing 5-lipoxygenase substrates. The *ALOX5* genotypes, carotid-artery intima-media thickness, and markers of inflammation were determined in a randomly sampled cohort of 470 healthy, middle-aged women and men from the Los Angeles Atherosclerosis Study. Dietary **arachidonic acid** and **marine n-3 fatty acids** were measured with the use of six 24-hour recalls of food intake. Variant *ALOX5* homozygotes (lacking the common allele) were found in 6.0 percent of the cohort. Mean intima-media thickness adjusted for age, sex, height, and racial or ethnic group was increased among carriers of two variant alleles, as compared with carriers of the common (wild-type) allele. Increased dietary **arachidonic acid significantly enhanced the apparent atherogenic effect of genotype, whereas increased dietary intake of n-3 fatty acids blunted the effect.** According to this study variant *ALOX5* genotypes identify a subpopulation with increased atherosclerosis. The observed **diet-gene interactions** further suggest that dietary arachidonic acids promote, whereas marine n-3 fatty acids inhibit the leukotriene-mediated inflammation that leads to atherosclerosis in this subpopulation (<http://www.ncbi.nlm.nih.gov/pubmed/14702425>).

High level of homocysteine is associated with CAD (coronary atherosclerotic disease). It contributes to damage of the endothelial wall, proliferation of smooth muscle in the blood vessel, and to the development of atherosclerotic plaques. The enzyme ***MTHFR*** (methylenetetrahydrofolate reductase) and the vitamin folic acid are important players in the homocysteine metabolic pathway. A common **thermolabile variant** of the *MTHFR* gene, C677T (Ala222Val) was associated with high homocysteine level and increased CAD risk in people with low folate intake. The 677TT genotype was associated with the highest risk, but it **could be reduced with folate intake.**

These two last examples showed that knowing certain genotypes can be advantageous if the environmental factors (here food) can be easily changed to blunt their harmful effects. It is utilized by most personal genomic DTC companies, where on the basis of the genotypes of the customers personal advice is given.

CD14 is part of the innate immunity and codes for the lipopolysaccharide receptor with a ligand found on the surface of Gram negative bacteria. With the help of toll like receptor 4

(*TLR4*), it induces a Th1 immune response against the pathogens. In the SNP -159C/T of the *CD14* gene, the rarer T allele increases the level of transcription, and the soluble CD14, and decreases the IgE level. In a French study it was studied whether different environments influenced the effect of this allelic variant on allergic rhinitis. The *CD14* -159TT genotype was associated with 2-fold reduced risk to atopy and rhinitis. Exposure to a farming environment in early life was associated with a similar reduced risk of nasal allergies. When farm exposure and *CD14* -159C/T were considered together, the risk of nasal allergies and atopy was the most reduced in the subjects who combined both an early-life exposure to a farming environment and the -159TT genotype (OR = 0.21 meaning ~5-fold risk reduction). This study showed that a gene-by-environment interaction between *CD14* -159C/T and environmental exposure in childhood may modify the development of atopy (<http://www.sciencedirect.com/science/article/pii/S0091674906013662>). This polymorphism could be considered in interventions studies that use microbial stimuli to reduce sensitization. In a similar study *TLR2* and *CD14* SNPs were associated with asthma and atopic asthma, respectively. In addition, *CD14*, *TLR2*, *TLR4*, and *TLR9* SNPs modified associations between country living and asthma (<http://www.atsjournals.org/journal/ajrccm>).

The ***CD14*** gene is a good example for the observation that the effects of a genotype can be even opposite depending on the environmental factors. An interesting finding was published about Karelian ethnic groups living both in Finland and Russia (<http://www.ncbi.nlm.nih.gov/pubmed/19222419>).

Considering the prevalence of asthma/allergic diseases, an East-West gradient has been consistently confirmed between Western affluent countries and Eastern developing countries, with, e.g. atopic diseases being more prevalent in Western than Eastern Europe. Finnish Karelans (Western environment) have previously been shown to have a higher prevalence of allergic disease than Russian Karelans (Eastern environment). These two areas are geographically adjacent and are expected to have similar outdoor air pollution. The Karelans were one ethnic group until they were artificially divided by a new Finnish/Russian border during the Second World War, after which changes have occurred mainly on the Russian side of the border with an influx of white Russians. However, the genetic make-up of the two populations should still be similar and, indeed, any differences can be readily detected as allele distribution differences between the populations on each side of the border. The major differences between Finnish and Russian Karelans therefore are likely to be in the cultural, economic and lifestyle conditions, with which they lived since their separation at the time of

the Second World War. The study analysed two asthma/atopy-related genes, *CD14* and *CC16*, which were chosen due to strong evidence of gene by environment interactions.

Opposite effects on asthma-related phenotypes were found for specific alleles of both *CD14* -159C/T and *CC16* A38G in the Karelian women. Of particular interest was the finding of **paradoxical gene by environment responses** for several allergy phenotypes. For some of these, itchy rash in particular, the ***CD14* TT genotype conferred the greatest risk among those in Finland, but the TT genotype was associated with the least risk in Russians**. A paradox was also found for *CC16* as the AA genotype was associated with the greatest risk of rhinitis and allergic eye symptoms in Finnish subjects, but the least risk for these phenotypes in Russians. Gene by environment interactions have been suggested to explain the inconsistencies in the associations of *CD14* -159C/T with atopic phenotypes and an **endotoxin switch model** has been postulated, in which the *CD14* promoter polymorphism changes the threshold, at which environmental endotoxin stimulation leads to a TH2 immune response. Several population-based and family-based studies showed that the *CD14* -159C/T polymorphism had an interactive effect with endotoxin exposure on atopic phenotypes. The T allele of the *CD14* -159C/T SNP, with possibly higher expression of *CD14*, exhibited protective effects on atopy with low exposure to endotoxin, in contrast, being a risk factor with high exposure to endotoxin. Russian Karelians in the Eastern environment may have had higher levels of endotoxin exposure, relative to Finnish Karelians in the Western environment. Indeed, recent studies found that Finnish Karelian children had a lower prevalence of microbial antibodies and less exposure to microbial loads in drinking water, compared with Russian Karelian children. With regard to the endotoxin switch model, Russian women with the T allele of *CD14* -159C/T, potentially exposed to higher levels of endotoxin relative to Finnish women, should have an increased risk for atopic phenotypes. However, the study found that in Russian women the T allele was protective against atopic conditions, which is inconsistent with the switch model and indicates the complexity of interactions between environmental exposure to endotoxin and genotypes of *CD14* (see in Chapter 10, Hygiene hypothesis).

Besides smoking, alcohol consumption is another strong and measurable environmental factor. The first step of the degradation of alcohol is catalyzed by alcohol dehydrogenases (ADH). Among white populations, variant alleles are common at the *ADH3* locus (present in 40 to 50 percent). At the *ADH3* locus, the $\gamma 1$ allele differs from the $\gamma 2$ allele by two amino acids at positions 271 and 349. Pharmacokinetic studies show a 2.5-fold difference in the maximal

velocity of ethanol oxidation between the homodimeric $\gamma 1$ isoenzyme (associated with a fast rate) and the homodimeric $\gamma 2$ isoenzyme (associated with a slow rate). This difference is thought to affect the rate of oxidation of blood ethanol, although the *ADH3* polymorphism had no apparent effect on blood alcohol levels in a short-term study of high-dose alcohol consumption in humans. Epidemiologic studies have associated the *ADH3* polymorphisms with alcohol-associated diseases, such as alcoholism ($\gamma 2\gamma 2$), alcohol-related end-organ damage ($\gamma 1\gamma 1$), and oropharyngeal cancer ($\gamma 1\gamma 1$). In a large study 14,916 USA male physicians between 40-84 years of age were followed up for 12 years (<http://www.nejm.org/doi/full/10.1056/NEJM200102223440802>). During this follow-up 396 had myocardial infarction (MI). In those who had *ADH3* $\gamma 2\gamma 2$ genotype and consumed at least 14 gram alcohol/day had 0.14 risk of MI (7.1-fold risk reduction) comparing to those who had $\gamma 1\gamma 1$ genotype and consumed no alcohol. It corresponds to the J-shaped curve found when the relationship between alcohol use and total mortality were studied (<http://circ.ahajournals.org/content/116/11/1306.long>). The nadir of the curves based on recent meta-analysis suggested optimal benefit at approximately half a drink per day. Fewer than 4 drinks per day in men and fewer than 2 per day in women appeared to confer benefit. Reductions in cardiovascular death and nonfatal myocardial infarction were also associated with light to moderate alcohol intake. Although some studies suggested that wine had an advantage over other types of alcoholic beverages, other studies suggested that the type of drink was not important. Heavy drinking was associated with an increase in mortality, hypertension, alcoholic cardiomyopathy, cancer, and cerebrovascular events, including cerebrovascular haemorrhage. Paradoxically, light-to-moderate alcohol use actually reduced the development of heart failure and did not appear to exacerbate it in most patients who had underlying heart failure. Numerous mechanisms have been proposed to explain the benefit that light-to-moderate alcohol intake has on the heart, including an increase of high-density lipoprotein cholesterol (HDL-C), reduction in plasma viscosity and fibrinogen concentration, increase in fibrinolysis, decrease in platelet aggregation, improvement in endothelial function, reduction of inflammation, and promotion of antioxidant effects. According to the previous findings the effect of alcohol consumption is influenced by genetic background, and *ADH3* genotypes have a strong role in it. In this study, participants who drank alcohol, and had $\gamma 2\gamma 2$ genotype, had higher mean HDL-C level, which is known to be protective against atherosclerosis and related diseases like MI (<http://www.ncbi.nlm.nih.gov/pubmed/11207350>).

The **antagonistic pleiotropy** described in Chapter 10 is also a good example for the gene-environmental interaction.

13.6. Genomic investigations of the gene-environmental interaction

In the last years with the development of genomic methods and large biobanks with detailed clinical and environmental data, the possibilities for more thorough studies of the gene-environmental interactions have increased considerably (see e.g. [UK Biobank](#) or [ALSPAC](#)).

Below, there are some examples for the new types of studies using genomic methods and large biobanks.

In a large study involving several populations gene-environmental interactions were investigated in MI. Earlier one of the most robust genetic associations for **cardiovascular disease (CVD) have been found with the chromosome 9p21 region**. In this study it was investigated whether environmental factors (nutrition) influence the effect of variants in 9p21 on MI risk. All four SNP risk variants increased the risk of MI by about a fifth. However, the effect of the SNPs on MI was influenced by the “prudent” diet pattern score of the [INTERHEART](#) participants (multiethnic population with 8,114 individuals (3,820 cases and 4,294 controls) from five ethnicities—European, South Asian, Chinese, Latin American, and Arab), a score that includes fresh fruit and vegetable intake as recorded in food frequency questionnaires. That is, the risk of MI in people carrying SNP risk variants was influenced by their diet. The strongest interaction was seen with an SNP called rs2383206, but although rs2383206 carriers who ate a diet poor in fruits and vegetables had a higher risk of MI than people with a similar diet who did not carry this SNP, rs2383206 carriers and non-carriers who ate a fruit- and vegetable-rich diet had a comparable MI risk. Overall, the combination of the least “prudent” diet and two copies of the risk variant were associated with a two-fold increase in risk for MI in the INTERHEART study. Additionally, data collected in the [FINRISK](#) study, which characterized healthy individuals living in Finland at baseline and then followed them to see whether they developed CVD (19,129 Finnish individuals with 1,014 incident cases of CVD), revealed a similar interaction between diet and 9p21 SNPs. These findings suggest that the risk of CVD conferred by chromosome 9p21 SNPs may be influenced by diet in multiple ethnic groups. Importantly, they suggest that **the deleterious**

effect of 9p21 SNPs on CVD might be mitigated by consuming a diet rich in fresh fruits and vegetables (<http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001106>).

Similar study has been carried out in connection with **obesity**. More than 20,000 individuals with European ancestry were involved. Altogether 12 SNPs were selected and genotyped previously showed strong association with increased body mass index (BMI). A genetic predisposition score for each individual was calculated and their occupational and leisure-time physical activities were assessed by using a validated self-administered questionnaire. Then, the researchers used modeling techniques to examine the main effects of the genetic predisposition score and its interaction with physical activity on BMI/obesity risk and BMI change over time. The researchers found that each additional BMI-increasing allele was associated with an increase in BMI equivalent to 445 g in body weight for a person 1.70 m tall and that the size of this effect was greater in inactive people than in active people. In individuals who have a physically active lifestyle, this increase was only 379 g/allele, or 36% lower than in physically inactive individuals, in whom the increase was 592 g/allele. Furthermore, in the total sample each additional obesity-susceptibility allele increased the odds of obesity by 1.116-fold. However, the increased odds per allele for obesity risk were 40% lower in physically active individuals (1.095 odds/allele) compared to physically inactive individuals (1.158 odds/allele). The findings of this study indicate that the genetic predisposition to obesity can be reduced by approximately 40% by having a physically active lifestyle. The findings of this study suggest that, while **the whole population benefits from increased physical activity levels, individuals who are genetically predisposed to obesity would benefit more than genetically protected individuals** (<http://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000332>).

Several studies showed that **regular consumption of coffee reduced the risk of Parkinson disease (PD)**. In the next example the results of a genomic study will be described, in which it was investigated with the help of GWAS, what genomic background influenced the positive effect of coffee-drinking (<http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002237>).

In this study genome-wide genotype data and lifetime caffeinated-coffee-consumption data on 1,458 persons with PD and 931 without PD were involved. A **genome-wide association and interaction study (GWAIS)** was performed, testing each SNP's main-effect plus its interaction with coffee, adjusting for sex, age, and two principal components. Subjects were stratified as heavy or light coffee-drinkers and a GWAS was carried out in each group. The rs4998386 SNP and the neighbouring SNPs in *GRIN2A* gene were associated with PD via heavy coffee consumption. *GRIN2A* encodes an NMDA-glutamate-receptor subunit and

regulates excitatory neurotransmission in the brain. In stratified GWAS, the ***GRIN2A* signal was present in heavy coffee-drinkers (OR = 0.43; P = 6×10⁻⁷) but not in light coffee-drinkers** (Figure 13.4). This study was a proof of concept that inclusion of environmental factors can help identify genes that are missed in GWAS. Both adenosine antagonists (caffeine-like) and glutamate antagonists (*GRIN2A*-related) are being tested in clinical trials for treatment of PD. *GRIN2A* may be a useful pharmacogenetic marker for subdividing individuals in clinical trials to determine which medications might work best for which patients.

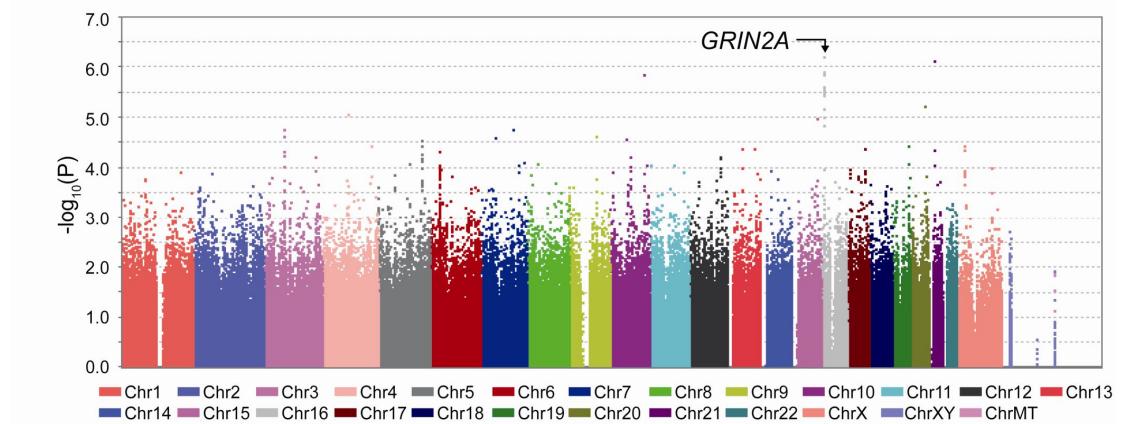


Figure 13.3. Manhattan plot depicts the results of a genome-wide association and interaction study (GWAIS) in heavy coffee drinkers with *GRIN2A* achieving the lowest P values. Source: <http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1002237>

13.7. Nutrigenetics and nutrigenomics

Nutrition plays an important role among environmental factors, because this heterogeneous factor has an effect on everybody and significantly influences the phenotype of each individual. Naturally, the effect of the nutrition is influenced by the genomic background of the individuals, which is studied by the nutrigenetics or nutrigenomics. There are usually two types of such studies. One of them investigates how the genetic variations influence the effect of the nutrition; and the other one how the nutrition influences the expression of genes. Previously in this chapter, several examples have been shown about the first one, like results in connection with the *APOE* and *ALOX5* genotypes, alcohol consumption and *ADH3*, coffee-drinking and *GRIN2A*, prudent diet and chromosome 9p21 variants, and folic acid and *MTHFR*.

There are a few examples in connection with the second one, like the study which investigated the effect of virgin olive oil on gene expression. The consumption of **virgin olive**

oil with a high phenolic content significantly affected the expression of 98 proinflammatory genes, many of which are known to be involved in inflammatory processes controlled by cytokines and transcription factors such as NF-κB, and pathways such as mitogen-activated protein kinase. The authors suggest that phenols may downregulate their gene expression thus reducing inflammation. These results are consistent with previous studies showing beneficial effects of olive oil on plasma lipid levels and reduced inflammation (<http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-11-253>).

Although nutrition is one of the strongest environmental factors, and nutrigenetic results would be very helpful in personal nutrition, it will surely not become part of the routine investigations in the near future, but a lot of personal genomic companies with direct-to-consumer services have included these in their services.

What advantages can present the nutrigenomics?

- It can offer information for personal nutrition.
- Personal nutrition may be associated with better physical and mental health.
- Symptoms of several diseases can be attenuated or prevented.
- These can lead to lower health care costs.

There are individuals who, because of their genetic background, are extremely sensitive to certain foods. E.g. individuals with **glucose-6-phosphate dehydrogenase (G6PD) deficiency** can have fatal haemolytic anaemia after consumption of broad bean (*vicia faba*). It is an X-linked hereditary condition, thus it occurs mainly in males. The disease is also called **favism** after the other name of the bean, fava bean. G6PD deficiency is the most common human enzyme defect, being present in more than 400 million people worldwide. African, Middle Eastern and South Asian people are affected the most along with those who are mixed with any of the above. A side effect of this disease is that it confers protection against malaria, in particular the form of malaria caused by *Plasmodium falciparum*, the most deadly form of malaria (see more <http://en.wikipedia.org/wiki/Favism>). Many pharmacological substances are potentially harmful to people with G6PD deficiency. Henna has been known to cause haemolytic crisis in G6PD-deficient infants. In these cases knowing the genetic background of the carriers can be life saving.

Several loci have been shown to modulate the **relation between saturated fats and body weight**, but the ***APOA2* –265T > C variant is the most convincingly replicated**. The protein product of the *APOA2* gene is a major part of the HDL (Figure 13.4). Mediterranean individuals with *APOA2* –265 CC genotype and high saturated fatty acid (SFA) intake had 6.8% greater BMI and higher prevalence of obesity. CC subjects with low SFA intake displayed lower plasma ghrelin than CC subjects with high SFA intake (all $P < 0.05$). ***APOA2* –265 CC subjects were more likely to exhibit behaviours that impede weight loss** and less likely to exhibit the protective behaviour. The –265C allele is associated with food consumption (higher energy intake) total fat and proteins and obesity (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3827635/>).

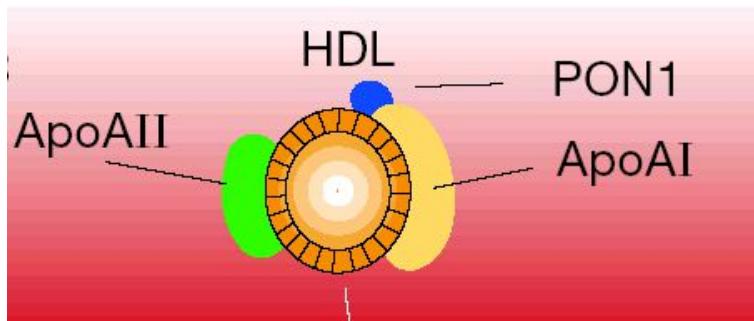


Figure 13.4. Main proteins in high density lipoprotein (HDL)

Naturally, the nutrigenetic investigations are quite expensive, but the main problem is not this. The population genetic studies give probabilities, and population averages. A given result may be valid and significant for the whole population, but not for every individual. It can occur that an individual carries the “good genotype”, but because of other factors (genetic and other environmental), the effect of an environmental factor has an opposite or even harmful effect than in the population average. E.g. somebody has the advantageous SNP in the *GRIN2A* gene in connection with PD, but must not drink coffee, because his/her stomach is sensitive for it and ulcer can develop. But the more factors are considered in the analysis, the more reliable the prediction at individual level can be.

13.8. The future of gene-environmental interaction

There is a huge potential in the results of gene-environmental interactions. Beyond the scientific significances, these studies can give results that can be utilized in the everyday life. The main problem with the environmental factors is that they are very difficult to measure,

often are unpredictable, random and even invisible. E.g. air pollution is a serious risk factor to a lot of diseases (asthma, COPD, atherosclerosis), but its effect on the different individuals is difficult to quantify. The effects of the environmental factors are influenced by the age, physical and mental states of the individuals. They are even influenced by the effects of earlier factors.

The development of genomic and evaluation methods and biobanks with high quality data will contribute to better and better understanding of the effects of the environmental data. The study of genome-environment interaction can yield additional data about the pathomechanism of the diseases, the optimal individual lifestyle or the optimal personal therapy. But there are networks of interactions in everybody, which requires **systems biologic methods**, and experts to evaluate them. It can be imagined that in the future **decision support systems** will help the physicians or dietitians who can involve hundreds or thousands or even millions of data about each individual and can predict the effect of the nutrition or other environmental factors on a personal level. Presently, however, it is not known, whether such reliable systems can be developed at all, and if yes, then when?

A special part of the gene-environmental interaction is the interaction between genes and the drugs, which are called pharmacogenetics or pharmacogenomics. Because from medical point of view it is especially important; thus the whole next chapter is about this theme.

13.9. Literature

1. Laland KN, Odling-Smee J, Myles S. How culture shaped the human genome: bringing genetics and the human sciences together. *Nat Rev Genet.* 2010 Feb;11(2):137-48.
2. International Human Genome Sequencing Consortium: Initial sequencing and analysis of the human genome. *Nature* 2001;409:860-921.
3. Venter JC et al. The sequence of the Human Genome. *Science* 2001;291:1304-51.
4. Barreiro LB, Quintana-Murci L. From evolutionary genetics to human immunology: how selection shapes host defence genes. *Nat Rev Genet.* 2010 Jan;11(1):17-30.
5. Fumagalli M et al. Genome-wide identification of susceptibility alleles for viral infections through a population genetics approach. *PLoS Genet.* 2010 Feb 19;6(2):e1000849.
6. Szalai Cs, Czinner A, Császár A, Szabó T, Falus A: Frequency of the HIV-1 resistance CCR5 deletion allele in Hungarian newborns. *Eur J Pediatr* 1998; 157:/9:782.

7. Hütter G, Ganepola S. The CCR5-delta32 polymorphism as a model to study host adaptation against infectious diseases and to develop new treatment strategies. *Exp Biol Med (Maywood)*. 2011 Aug 1;236(8):938-43.
8. Tishkoff SA et al. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet*. 2007 Jan;39(1):31-40.
9. Tully G. Genotype versus phenotype: human pigmentation. *Forensic Sci Int Genet*. 2007 Jun;1(2):105-10.
10. Reich D et al. Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *Am J Hum Genet*. 2011 Oct 7;89(4):516-28.
11. Chambers V et al. Haemochromatosis-associated HFE genotypes in English blood donors: age-related frequency and biochemical expression. *J Hepatol*. 2003 Dec;39(6):925-31.
12. Erblich J et al. Stress-induced cigarette craving: effects of the DRD2 TaqI RFLP and SLC6A3 VNTR polymorphisms. *Pharmacogenomics J*. 2004;4(2):102-9.
13. Minematsu N et al. Association of CYP2A6 deletion polymorphism with smoking habit and development of pulmonary emphysema. *Thorax*. 2003 Jul;58(7):623-8.
14. Stevens VL et al. Nicotinic receptor gene variants influence susceptibility to heavy smoking. *Cancer Epidemiol Biomarkers Prev*. 2008 Dec;17(12):3517-25.
15. Füst G, Arason GJ, Kramer J, Szalai C et al. Genetic basis of tobacco smoking: strong association of a specific major histocompatibility complex haplotype on chromosome 6 with smoking behavior. *Int Immunol*. 2004 Oct;16(10):1507-14.
16. Lundström E et al. Gene-environment interaction between the DRB1 shared epitope and smoking in the risk of anti-citrullinated protein antibody-positive rheumatoid arthritis: all alleles are important. *Arthritis Rheum*. 2009 Jun;60(6):1597-603.
17. Criswell LA et al. Smoking interacts with genetic risk factors in the development of rheumatoid arthritis among older Caucasian women. *Ann Rheum Dis*. 2006 Sep;65(9):1163-7.
18. Blaskó B et al. Low complement C4B gene copy number predicts short-term mortality after acute myocardial infarction. *Int Immunol*. 2008 Jan;20(1):31-7.
19. Füst György, Kramer Judit, Kiszel Petra, Blaskó Bernadette, Szalai Csaba, Guðmundur Johann Arason, Chack Yung Yu. C4BQ0, egy génvariáns, amely jelentősen csökkenti az esélyt az egészséges öregkor megélésére. *Magyar Tudomány*, 2006/3 266. o.

20. Lee KM et al. Paternal smoking, genetic polymorphisms in CYP1A1 and childhood leukemia risk. *Leuk Res.* 2009 Feb;33(2):250-8.
21. Susan Colilla et al. Evidence for gene-environment interactions in a linkage study of asthma and smoking exposure. *J Allergy Clin Immunol* 2003;111:840-6.
22. Wang Z et al. Association of asthma with beta(2)-adrenergic receptor gene polymorphism and cigarette smoking. *Am J Respir Crit Care Med.* 2001 May;163(6):1404-9.
23. Wang XL et al. Effect of CYP1A1 MspI polymorphism on cigarette smoking related coronary artery disease and diabetes. *Atherosclerosis.* 2002 Jun;162(2):391-7.
24. Talmud PJ, Hawe E, Miller GJ. Analysis of gene-environment interaction in coronary artery disease: lipoprotein lipase and smoking as examples. *Ital Heart J.* 2002 Jan;3(1):6-9.
25. Kivipelto M et al. Apolipoprotein E epsilon4 magnifies lifestyle risks for dementia: a population-based study. *J Cell Mol Med.* 2008 Dec;12(6B):2762-71.
26. Rusanen M et al. Midlife smoking, apolipoprotein E and risk of dementia and Alzheimer's disease: a population-based cardiovascular risk factors, aging and dementia study. *Dement Geriatr Cogn Disord.* 2010;30(3):277-84.
27. Drenos F, Kirkwood TB. Selection on alleles affecting human longevity and late-life disease: the example of apolipoprotein E. *PLoS One.* 2010 Apr 2;5(4):e10022.
28. Dwyer JH et al. Arachidonate 5-lipoxygenase promoter genotype, dietary arachidonic acid, and atherosclerosis. *N Engl J Med.* 2004 Jan 1;350(1):29-37.
29. Zhang G et al. Opposite gene by environment interactions in Karelia for CD14 and CC16 single nucleotide polymorphisms and allergy. *Allergy.* 2009 Sep;64(9):1333-41.
30. Alam MA et al. Association of polymorphism in the thermolabile 5, 10-methylene tetrahydrofolate reductase gene and hyperhomocysteinemia with coronary artery disease. *Mol Cell Biochem.* 2008 Mar;310(1-2):111-7.
31. Bufalino A,. Maternal polymorphisms in folic acid metabolic genes are associated with nonsyndromic cleft lip and/or palate in the Brazilian population. *Birth Defects Res A Clin Mol Teratol.* 2010 Nov;88(11):980-6.
32. Chen L et al. Alcohol intake and blood pressure: a systematic review implementing a Mendelian randomization approach. *PLoS Med.* 2008 Mar 4;5(3):e52.
33. Hines LM et al. Genetic variation in alcohol dehydrogenase and the beneficial effect of moderate alcohol consumption on myocardial infarction. *N Engl J Med.* 2001 Feb 22;344(8):549-55.

34. Capri M et al. Human longevity within an evolutionary perspective: the peculiar paradigm of a post-reproductive genetics. *Exp Gerontol.* 2008 Feb;43(2):53-60.
35. Candore G et al. Inflammation, longevity, and cardiovascular diseases: role of polymorphisms of TLR4. *Ann N Y Acad Sci.* 2006 May;1067:282-7.
36. Do R et al. The Effect of Chromosome 9p21 Variants on Cardiovascular Disease May Be Modified by Dietary Intake: Evidence from a Case/Control and a Prospective Study. *PLoS Medicine* 2011;9 (10)
37. Li S et al. Physical activity attenuates the genetic predisposition to obesity in 20,000 men and women from EPIC-Norfolk prospective population study. *PLoS Med.* 2010 Aug 31;7(8). pii: e1000332. PubMed PMID: 20824172; PubMed Central PMCID: PMC2930873.
38. Lu Y, Feskens EJ, Dolle ME et al. Dietary n-3 and n-6 polyunsaturated fatty acid intake interacts with FADS1 genetic variation to affect total and HDLcholesterol concentrations in the Doetinchem Cohort Study. *Am J Clin Nutr* 2010; 92:258–265.
39. Ordovás JM, Robertson R, Cléirigh EN. Gene-gene and gene-environment interactions defining lipid-related traits. *Curr Opin Lipidol.* 2011 Apr;22(2):129-36.
40. Hamza TH et al. Genome-wide gene-environment study identifies glutamate receptor gene GRIN2A as a Parkinson's disease modifier gene via interaction with coffee. *PLoS Genet.* 2011 Aug;7(8):e1002237.

13.10. Questions

1. From a gene environmental point of view what does it mean that a genetic variant has high or low penetrance?
2. What is the distribution of the population regarding responses to environmental stimuli?
3. Give examples for the interactions between highly penetrant genetic variants and the environment!
4. Give examples for the interactions between low penetrant genetic variants and the environment!
5. What aspects can be investigated studying the smoking-genome interactions?
6. Give examples for genes playing roles in the addiction to smoking!
7. With which disease did genetic variants in the nicotinic receptors associate in different GWAS?

8. What roles do the variants of the *CYP2A6* genes play in smoking?
9. What does the association between the 8.1 ancestral haplotype of the MHC region and smoking initiation implicate?
10. Give examples for genes in which variations can influence the health of the smokers!
11. What kind of environmental factor interacts with HLA-DRB1 SE, and what can be the consequences?
12. What consequences have been found for carriers of the C4B*Q0 variants?
13. What gene has an important role in the degradation of the toxins in the smoke? What can be the consequences of the variations in this gene?
14. What gene has variations which influenced the risk to asthma in smokers?
15. Give some examples for the gene environmental interactions regarding the *APOE* gene!
16. Which gene variations can influence the effect of consuming food rich in arachidonic acid on intima-media thickness?
17. What food supplements would you recommend for men carrying promoter polymorphisms in the *ALOX5* gene?
18. What food supplements would you recommend for individuals carrying the thermolabile variant of the *MTHFR* gene?
19. What environmental factors and how can influence the effect of variations in the *CD14* gene?
20. Is it possible that a genetic variation can have opposite effects in different populations? Explain it!
21. What environmental factor can interact with the variations in the *ADH3* gene, and how?
22. With what diseases did the 9p21 chromosome region associate, and what and how influenced this association?
23. What non-genetic factor influenced the effect of polymorphism associated with risk to obesity?
24. What is GWAIS and what did it find in Parkinson disease?
25. What is nutrigenomics or nutrigenetics and what is its significance?
26. How was it proved that virgin olive oil had anti-inflammatory effect?
27. What is favism and what are its consequences?

28. What are the difficulties in the gene–environmental interaction studies and what is its significance?

14. Pharmacogenomics

Csaba Szalai

14.1. Goals of pharmacogenomics

14.1.1. Drug development

Pharmacogenomics has two main goals. One of them is to **search for new drugs and drug targets** with the help of genomic methods. It has great significance, because current existing therapies only hit about 400 different drug targets compared to the 20-22 thousand protein coding genes coding for about 2 million different proteins (because of e.g. posttranslational modifications, splice variants). In addition, DNA or RNA sequences can also be regarded as potential targets, and we know from the ENCODE project (<http://www.nature.com/encode/#/threads>) and the amount of conserved regions that about 10% of the human genome has some functions, and the number of cell specific enhancer regions is about 400,000 (see Chapter 9.7). Naturally, only a fraction of these are really drug targets, but according to estimations there are at least 10 times more drug targets than presently exists.

In contrast, however, drug development presently has been in crisis. In the last years the FDA has only approved between 18 and 30 new drugs annually, while between 1998 and 2002 on average this number was 68 (<http://www.fda.gov/Drugs/informationondrugs/ucm079436.htm>). The identification of new drug targets can significantly be **accelerated by the new high throughput genomic methods**, and in addition, the price of the drug development can be considerably reduced. E.g. with the help of GWAS, several million genetic variants can be studied in a short time and at a relatively low price, and if a variant is associated with a disease or its symptoms or endophenotypes (QT), then it means that there should be a sequence near to it, which plays a role in the associated phenotype, and the sequence itself or the coded protein or RNA or the influenced pathway can be regarded as potential drug targets. An additional advantage of the genomic methods is that they can be **hypothesis-free** (i.e. it is not necessary to know the pathomechanism), and thus new pathways can be discovered. Unfortunately, long time is needed (up to 20 years) from target identification till bringing a drug candidate to market, but genomic and other modern methods can shorten this time considerably.

14.1.2. Adverse drug response

The main topic of this chapter is about the other main goal of pharmacogenomics. This deals with **the influence of genetic variations on drug response in patients by correlating gene expression or single-nucleotide polymorphisms with a drug's efficacy or toxicity**. By doing so, pharmacogenomics aims to develop rational means to optimize drug therapy, with respect to the patients' genotype, to ensure maximum efficacy with minimal adverse effects. Pharmacogenomics is the whole genome application of pharmacogenetics, which examines the single gene interactions with drugs.

These types of studies are at least as significant as the discovery of new drug targets. Genomic differences between people can result in significant differences in their responses to the drugs. E.g. in 30% of people the β-blockers used against hypertension are **ineffective**, while the antidepressants are ineffective in 50% of the treated persons. The situation is similar in the cases of most drugs.

The **adverse drug response** can cause even larger problems. According to statistical data several hundred thousand people die in the USA annually because of the adverse effects of drugs. In a study researchers examined data from approximately 1,000 patients who had been admitted to a large Liverpool hospital. Among the 290 patients who were readmitted within one year and for whom data were available, 21 percent had been readmitted at least partly because of an adverse drug reaction (http://www.naturalnews.com/027866_drugs_side_effects.html).

The differences of people in responses to drugs have strong genetic background. According to estimations, **60-80% of these differences are due to genetic differences between people**.

There are more and more pharmacogenetic information about different drugs, but this only slowly goes into the practice. In a retrospective survey where 53 thousand patients and 99 drugs with FDA-approved pharmacogenetic information were studied, it was found that in 300-600 cases the severe adverse effects could have been avoided if the pharmacogenetic test had been carried out.

The two branches of pharmacogenomics overlap with each other. If a drug candidate passes the phase I clinical trial, a whole genome SNP screening could be carried out in phase II. An **abbreviated SNP profile to predict efficacy** could be identified in this phase II by detecting those SNPs along the genome that are in linkage disequilibrium, when patients with efficacy are compared with patients who did not respond to the drug candidate. An abbreviated profile of these small regions of linkage disequilibrium that differentiate efficacy can then be used to select patients for larger phase III studies. This could make many of these phase III studies smaller and therefore more efficient. Pharmacogenetics could also be used during the initial

post-marketing surveillance period (phase IV) to identify **SNP markers associated with serious but rare adverse events** (Fig. 14.1). These markers could be added to the SNP markers for efficacy and common adverse events identified during development to produce a comprehensive medicine response profile, and to identify which patients respond to the drug and which patients will be at high risk for an adverse event (<http://www.ncbi.nlm.nih.gov/pubmed/10866212>). These could be done for existing drugs, and subpopulations could be selected, in which the drugs could be efficient and the risk for an adverse effect is minimal. There is an interesting example of this: the case of BiDil or NitroMed suggested in the treatment of congestive heart failure. The clinical trials, which investigated an ethnically mixed population, did not prove the efficiency of the drug, and thus it was not approved by the FDA. But later, when the trial was repeated on an African-American population, the trial had to be stopped, because the difference was so significant between the placebo and the treated populations. It became the first **ethnic-specific drug** (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1687161/>). Because the **ethnicity, or the color of the skin are rather subjective, and do not influence directly the drug-response**, genetic markers could be determined which could predict the real connection.

There are a lot of examples when an approved drug must be withdrawn from the market, because of serious, but rare adverse effects. This causes \$billion loss for the pharmaceutical companies, but this is also harmful for those sick people, for whom the drug was efficient. If the cause of the serious adverse effects could be determined, which could be genetic, then the individuals who have a high risk for the adverse effects, could be treated with alternative therapy.

Theoretically it is possible that in the future, everybody will have a genomic profile, available for the physicians, who with the help of a decision-support system would be able to select the optimal drugs or therapy for the patients. This could be the ideal case for the personal therapy.

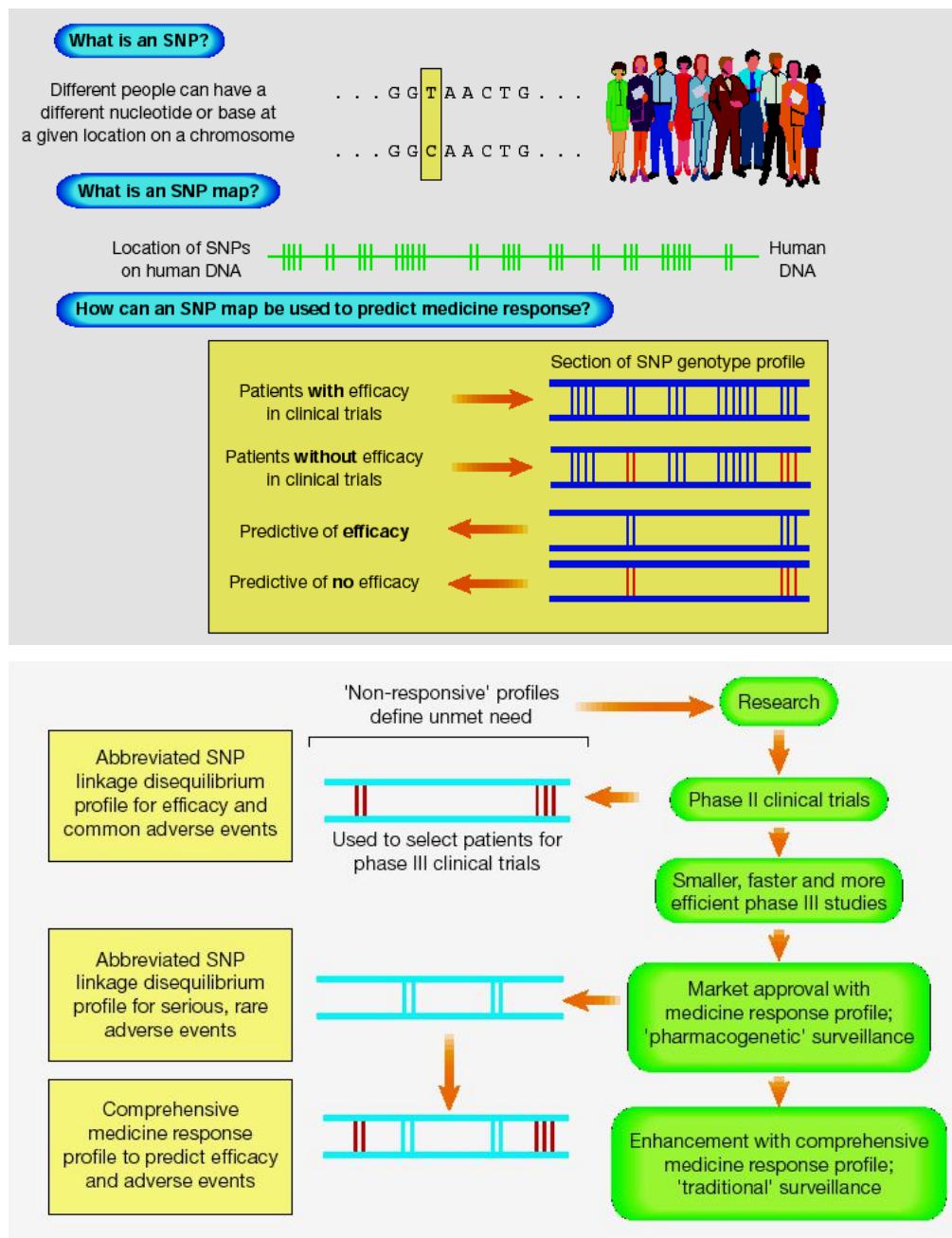


Figure 14.1. How can genotyping help for selecting the right patients for a given drug in clinical trials and later in the real life? Source: <http://www.ncbi.nlm.nih.gov/pubmed/10866212>

14.2. Genomic background of adverse effects

One of the main questions of pharmacogenomics is that, what the mechanism is, with which the genetic variants influence the drug-response. There are three main possible mechanisms:

Pharmacokinetic: Genetic variations, which influence the mechanisms of absorption and distribution of the administered drug, the chemical changes of the substance in the body, and the effects and routes of excretion of the metabolites of the drug.

Pharmacodynamic: Genetic variants, which are in the genes of the drug targets or in their associated pathways. Pharmacodynamics is often summarized as the study of what a drug does to the body, whereas pharmacokinetics is the study of what the body does to a drug.

Idiosyncratic: Genetic variations in genes coding for proteins, which are not in the drug target or pharmacokinetic pathways, but could influence the drug response. The adverse effects could be caused by e.g. an enzymopathy, so that the triggering substance cannot be processed properly in the organism and causes symptoms by accumulating or blocking other substances to be processed.

14.3. Difficulties of the pharmacogenomic researches

It can be asked that if the significance of the pharmacogenomics and the interest of the pharmaceutical industry are so great, then why there are so few perceived results? One of the explanations is that the main development of the high throughput genomic, bioinformatic and other methods have been carried out only in the last few years, and there was not enough time (10-15 years) for the marketing of the drugs developed by the new methods.

But there are other characteristics which can cause difficulties. Often environmental factors can cause similar effects as the genetic variants, which is called **phenocopy**. From a statistical point of view it can cause great difficulties in the evaluation.

Another disturbing factor is **gene-gene interactions**. As was detailed in Chapters 10 and 11, it is very difficult to detect and evaluate them. The variants can occur in the same or different genes, and strengthen or weaken the effects of each other. And as the distributions of the genetic variants can be different between different populations, the perceived effects of individual variants can also differ.

There is also a less ethical explanation to why the pharmacogenetics results are so few today. In some cases the pharmaceutical companies are not interested that their drugs may be used only on people where the drugs are really effective, because it can result in less user and less profit.

Perhaps the Food and Drug Administration (FDA) in the USA is the most famous authorities which decides on drug approval, direction of use or drug labeling. On the web page of FDA a table can be found with the list of **FDA-approved pharmacogenomic biomarkers in drug labels** (<http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm>). In November 2012 there were 118 rows with drug names in this table, in May 2016, 164. It shows that the

number is increasing, although with not a very high speed. **The most rows are connected to oncology**, then psychiatry. Among the genes the most frequent ones belong to the CYP gene family. The most frequent gene is the ***CYP2D6*** and then the ***CYP2C19***. Because the **CYP gene family** plays an important role in the drug metabolism, it shows that the **pharmacokinetic variants are overrepresented** in this list.

In the followings we show only a few examples of the above mentioned list, and will concentrate rather on the researches which are carried out in this topic. We use only a few types of diseases and drugs as examples. These examples show the promises but also the difficulties of pharmacogenomics in clinical use.

It must be noted that most results are genetic and not genomic, but in this area the terms of pharmacogenomics and pharmacogenetics are often used as synonyms, and we used them in a similar way.

14.4. Genetic variants influencing pharmacokinetics

According to estimations, the effects of about 20% of the drugs on the market are influenced by variants in genes coding for enzymes responsible for the degradation of the drugs. If a variant increases the activity of the enzyme (fast metabolism), then the drug is excreted too fast, and may not be able to exert its total effect. If a variant has an opposite effect (slow metabolism), the drug can accumulate and become toxic and may have more adverse effects.

The **CYP (cytochrom P-450) gene family** has 57 members and they play a role in oxidizing endogenous compounds and xenobiotics. Variants in this gene are responsible for 80% of all adverse drug responses. ***CYP2D6*** as the most frequently occurring gene on the previously mentioned FDA list acts on one-fourth of all prescribed drugs. Approximately 10% of the population has a slow acting form of this enzyme and 7% a super-fast acting form, while 35% are carriers of a non-functional 2D6 allele, which elevates considerably the risk of adverse drug reactions, when the individuals are taking multiple drugs.

In the case of ***CYP2C9*** (cytochrome P450 2C9), approximately 10% of the population are carriers of at least one allele for the slow-metabolizing form and may be treatable with 50% of the dose at which normal metabolizers are treated.

Warfarin is an anticoagulant normally used in the prevention of thrombosis and thromboembolism, the formation of blood clots in the blood vessels and their migration elsewhere in the body respectively. Warfarin is the most widely prescribed oral anticoagulant drug.

Warfarin activity is determined partially by genetic factors. The FDA offers to use genetic tests to improve their initial estimate of what is a reasonable warfarin dose for individual patients. Polymorphisms in two genes (*VKORC1* and *CYP2C9*) are particularly important.

- *CYP2C9* polymorphisms explain 10% of the dose variation between patients, mainly among Caucasian patients, as these variants are rare in African American and most Asian populations. These *CYP2C9* polymorphisms do not influence time to effective INR as opposed to *VKORC1*, but does shorten the time to INR >4 (International Normalized Ratio: https://en.wikipedia.org/wiki/Prothrombin_time#International_normalized_ratio).
- *VKORC1* polymorphisms explain 30% of the dose variation between patients: particular mutations make *VKORC1* less susceptible to suppression by warfarin. There are two main haplotypes that explain 25% of variation: low-dose haplotype group (A) and a high-dose haplotype group (B). *VKORC1* polymorphisms explain why African Americans are on average relatively resistant to warfarin (higher proportion of group B haplotypes), while Asian Americans are generally more sensitive (higher proportion of group A haplotypes). Group A *VKORC1* polymorphisms lead to a more rapid achievement of a therapeutic INR, but also a shorter time to reach an INR over 4, which is associated with bleeding.

Despite the promise of pharmacogenomic testing in warfarin dosing, its use in clinical practice is controversial. In August 2009 the Centers for Medicare and Medicaid Services concluded that "the available evidence does not demonstrate that pharmacogenomic testing of CYP2C9 or VKORC1 alleles to predict warfarin responsiveness improves health outcomes in Medicare beneficiaries." Two randomized controlled trials have found that even though genetic testing may predict stable doses of warfarin, the testing does not increase time in therapeutic range (i.e., time in the desired level of anticoagulation). A recent study found that prospective genotyping reduced hospitalization rates for patients just starting warfarin therapy (<https://en.wikipedia.org/wiki/Warfarin>). In another study INR measurements did not entirely capture the information provided by *CYP2C9* and, especially *VKORC1* genotypes, the latter remained the most informative predictor of stable warfarin dose requirements in the studied Brazilian cohort (<http://www.bloodjournal.org/content/113/17/4125.full?ssq-checked=true>).

Because of the known clinical significance of CYP polymorphisms, there are **CYP chips** available for the determination of the known predictor genotypes.

Suxamethonium chloride, also known as **suxamethonium** or **succinylcholine**, is a nicotinic acetylcholine receptor agonist, used to induce muscle relaxation and short-term paralysis, usually to facilitate tracheal intubation. In 1/2500 individuals the enzyme **butyrylcholinesterase** (also known as pseudocholinesterase, plasma cholinesterase and is encoded by the *BCHE* gene) that hydrolyses many different choline esters and also this compound, has no activity due to mutations in both genes, which can cause serious adverse reactions like apnea.

Mercaptopurine (its brand name Purinethol) is an immunosuppressive drug used to treat e.g. leukemia, pediatric non-Hodgkin's lymphoma, and inflammatory bowel disease (such as Crohn's disease and ulcerative colitis). Its metabolizing enzyme is thiopurine methyltransferase (**TPMT**). Its gene has three known SNPs causing enzyme deficiency. Patients with TPMT deficiency are much more likely to develop dangerous myelosuppression.

The gene for **multidrug resistance-1 (MDR-1)** belongs to the **ABC-transporter** family and the name of its gene is *ABCB1*. It is expressed at the apical membrane of the mucosal epithelium all along the gastrointestinal tract, at the biliary canalicular membrane of hepatocytes and on the apical surface of cells in the proximal kidney tubules protecting our cells against toxic compounds, including some drugs. In the *ABCB1* 3435T/C SNP, the T allele is associated with lower expression level of the gene, and higher rate of adverse drug reactions (<http://www.ncbi.nlm.nih.gov/pubmed/17434155>).

Anthracyclines are potent cytostatic drugs, the correct dosage of which is critical to avoid possible cardiac side effects. **ABCC1** (MRP1) is expressed in the heart and takes part in the detoxification and protection of the cells from toxic effects of xenobiotics, including anthracyclines. Polymorphisms in this gene influence the cardiac side effects of the anthracyclines (<http://www.ncbi.nlm.nih.gov/pubmed/21929509>).

14.5. Genes influencing pharmacodynamics

There are significantly fewer results regarding genetic polymorphisms influencing pharmacodynamics. The previously mentioned **VKORC1** is an example of it (see above in details).

14.6. Examples of pharmacogenetic studies

14.6.1. Pharmacogenetics in oncology

As we could see earlier the most drugs with validated pharmacogenetic tests are in oncology. The most important **driver mutation of chronic myeloid leukemia** is **t(9;22)(q34;q11)** translocation which leads to the **Philadelphia chromosome**. This gives rise to a **fusion gene, BCR-ABL1**, that juxtaposes the *ABL1* gene on chromosome 9 (region q34) to a part of the BCR ("breakpoint cluster region") gene on chromosome 22 (region q11). The disease can be treated with **tyrosine kinase inhibitors** (e.g. imatinib, nilotinib, dasatinib). But, there can be several mutations in the fusion gene which can lead to **therapy resistance** for certain drugs. There are a lot of data about which variants can cause resistance against which drugs. With **sequencing** the fusion gene it can be predicted the optimal kinase inhibitor. E.g. the T315I mutation is a quite frequent variation in the fusion gene and cause resistance to most drugs. But there are newly developed tyrosine kinase inhibitors (ponatinib, axitinib) to which the fusion gene with the mutation is sensitive. The **effectiveness of the therapy can be followed up with quantitative PCR worked out for the fusion gene**. After the initiation of the therapy the level of the *BCR-ABL1* fusion gene is measured in every third month. If the level of the fusion gene is not decreased in a given rate an alternative therapy must be chosen <http://www.ncbi.nlm.nih.gov/pubmed/23803709>.

In **non small cell lung cancer** the **mutations of the EGFR and KRAS** are frequently interrogated. The mutations in the gene influence the effectiveness of the drugs. EGFR tyrosine kinase inhibitors are usually very effective in cases where the *EGFR* gene is mutated in the tumor cells.

In some cancer the driver mutations can be the **amplification of a gene**, like in certain breast and stomach cancers. The amplification of the *HER2* gene, located on chromosome 17, can be connected to a lot of cancer types. There are approved drugs on the market which specifically inhibits the *HER2* proteins and are used in *HER2*-positive breast and stomach cancers. There can also be point mutations in the tyrosine kinase domain of the *HER2* gene. It occurs in 4 % of lung cancers and 3% of colon cancers. There is a specific small molecule tyrosine kinase inhibitor which inhibits the faulty *HER2* protein. This drug inhibits both the EGFR and the *HER2* tyrosine kinase enzymes.

Next to genetic variations, gene expression levels in the tumor can also give useful information about the required patient (tumor) specific therapy. In case of breast cancer, after removing the tumor, chemotherapy must be applied in some patients. Chemotherapy causes

unpleasant side effects, but low risk patients do not need it. It was shown that gene expression profile of the tumorous tissue can give useful information for stratifying patients according to their risks. The FDA approved MammaPrint measures the expression profile of 70 genes. In a prospective clinical study for a breast cancer recurrence assay, the utility of the MammaPrint 70-gene signature was confirmed to identify those breast cancer patients that may safely forgo chemotherapy. (<http://www.agendia.com/healthcare-professionals/breast-cancer/mammaprint/>).

14.6.2. Pharmacogenetics of statins

Statins (or **HMG-CoA reductase inhibitors**) are a class of drugs used to lower cholesterol levels by inhibiting the enzyme HMG-CoA reductase, which plays a central role in the production of cholesterol in the liver. Increased cholesterol levels have been associated with cardiovascular diseases, and statins are therefore used in the prevention of these diseases. Statins **have rare but severe adverse effects, particularly muscle damage**, and some doctors believe they are overprescribed. The best-selling statin is atorvastatin, marketed as Lipitor (manufactured by Pfizer) and Torvast. By 2003 atorvastatin became the best-selling pharmaceutical in history.

Because of their widespread use and rare but severe adverse effects, a lot of pharmacogenetic studies have been carried out. The **CYP enzymes** play an important role in the metabolism of statins. CYP3A4 metabolize the lovastatin, simvastatin and the atorvastatin. The level of **CYP3A4** can show a 10-fold difference between people indicating genetic variants. In one study, a -290A/G promoter SNP influenced significantly the LDL-C level after atorvastatin treatment, while the M445T variant influenced the LDL-C level both before and after treatment (<http://www.ncbi.nlm.nih.gov/pubmed/14697480>). Individuals carrying rs35599367, C>T SNP in intron 6 of the gene needed 0.2-0.6-fold simvastatin dose reduction for the optimal lipid level.

CYP3A5 contribute to biotransformation of some statin. There is a 6986 G/A SNP in intron 3 of the gene, which influence significantly the expression of the gene. Only 10% of the European populations show high CYP3A5 expression, and in these people the lovastatin, simvastatin and the atorvastatin treatment is significantly less effective (<http://www.ncbi.nlm.nih.gov/pubmed/15284534>).

The **multidrug resistance-1 (MDR-1, ABCB1)** significantly influences the transport and localization of the statins. In a study the previously mentioned *ABCB1* C3435 SNP influenced the LDL-C level in atorvastatin therapy (<http://www.ncbi.nlm.nih.gov/pubmed/19891551>).

Polymorphisms in the **HMGCR** gene encoding HMG-CoA reductase, the main target of statins, also influence the response to the drugs.

Mutations in the **LDLR** gene cause familiar hypercholesterolemia (FH). The responds to statins of the mutation carriers depend on the types of the mutations. People with null mutations respond worse than people with mutations influencing only the functions of the receptor.

GWAS were also carried out in this topic. The Study of the Effectiveness of Additional Reductions in Cholesterol and Homocysteine (SEARCH) identified a SNP in the *SLC01B1* gene (*SLCO1B1*5*) which is associated with statin-induced myopathy in simvastatin (Zocor) treated patients with cardiovascular diseases (<https://clinicaltrials.gov/ct2/show/NCT00124072>).

It must be noted, however, that the pharmacogenetic results in connection with the statins are rather controversial, and thus in the FDA-approved list there are only two items of statins with pharmacogenomic drug labels.

14.6.3. Clopidogrel

Clopidogrel is an oral, thienopyridine class antiplatelet agent used to inhibit blood clots in coronary artery disease, peripheral vascular disease, and cerebrovascular disease. It is marketed by Bristol-Myers Squibb and Sanofi under the trade name **Plavix**. The drug works by irreversibly inhibiting a receptor called P2Y₁₂, an adenosine diphosphate (ADP) chemoreceptor on platelet cell membranes. Adverse effects include haemorrhage, severe neutropenia, and thrombotic thrombocytopenic purpura (TTP). It is prescribed for 40 million patients annually.

Clopidogrel is a pro-drug activated in the liver by cytochrome P450 enzymes, including **CYP2C19**. Three-four percent of the Caucasian population homozygote, while 24% heterozygote for the inactive variants of the gene associating with higher rate of cardiovascular complications.

GWAS was carried out in an Amish population, and a SNP in the **CYP2C19** gene was identified, which was associated with reduced drug response, and this was responsible for

12% of the drug response variations (<https://www.genomeweb.com/dxpgx/gwas-implicates-cytochrome-p450-gene-plavix-response>) . The traditional factors (BMI, age, cholesterol level) were responsible for only 10% of the variations. This was later confirmed in another study and in a 12-year follow-up study the CYP2C19 status was the only independent risk factor, when cardiovascular death, non-fatal myocardial infarction or coronary revascularization were applied as target values. In another study two variants of the *ABCB1* were shown to be associated with adverse drug response. The product of this gene plays a role in the absorption of the drug. *CYP2C19* has a gain of function allele (CYP2C19*17) which codes for an ultra-fast metabolizing form of the enzyme. Carriers of this allele respond better to the drug (<http://www.ncbi.nlm.nih.gov/pubmed/22123178>). Presently, FDA recommends alternative therapies for poor responders, and in March 2010 the warnings about *CYP2C19* genotypes were put into the drug label.

14.6.4. MODY

Maturity onset diabetes of the young or MODY is a monogenic form of diabetes with an autosomal dominant inheritance. It means that the children inherit the disease (diabetes) from their diseased parents with 50% chance. All the genes mutated in MODY influence the insulin secretion. According to estimations about 5% of patients diagnosed with type 1 diabetes mellitus and also an unknown part of patients with type 2 diabetes have in reality MODY.

Until now 14 genes have been detected in which mutations can cause MODY, but 3 of them are relative frequent and detection of mutations in them has a therapeutic significance. In **MODY2** the disease is caused by mutations in the glucokinase (*GCK*) gene. The patients have high blood glucose levels, but it turned out that it does not associate with any secondary disease or worsening symptoms, thus it usually **requires no drug treatment**.

Mutations in ***HNF1A*** and ***HNF4A*** genes (**MODY3** and **MODY1**, respectively) have much more serious consequences and require treatment. But it turned out that they react quite well to **sulfonylureas** and not so well to metformin and usually do not require intravenous insulin treatment for years (Figure 14.2).

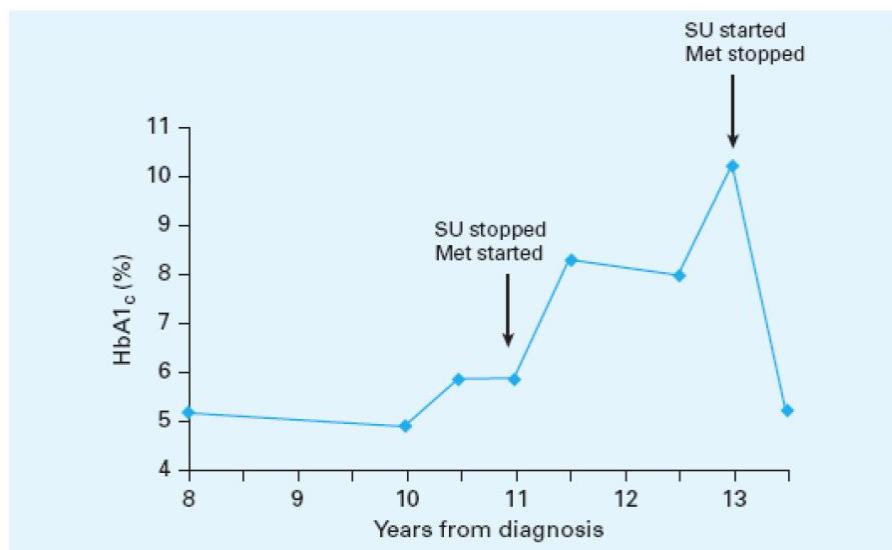


Figure 14.2. Patients with MODY1 and MODY3 react quite well to sulfonylurea (SU). Metformin (Met) which is the most frequent drug used in type 2 diabetes is much less effective in these forms of diabetes as can be seen in the elevated HbA1c blood level in the patient. HbA1c or glycated hemoglobin serves as a marker for average blood glucose levels over the previous 3 months before the measurement as this is the lifespan of red blood cells.

14.6.5. Pharmacotherapy of asthma

There are four major classes of asthma pharmacotherapy currently in widespread use: (1) β_2 -agonists used by inhalation for the relief of airway obstruction (e.g. albuterol, salmeterol, fenoterol); (2) glucocorticosteroids for both inhaled and systemic use (e.g., beclomethasone, triamcinolone, prednisone); (3) theophylline and its derivatives, used for both the relief of bronchospasm and the control of inflammation; and (4) inhibitors and receptor antagonists of the cysteinyl-leukotriene pathway (e.g. montelukast, pranlukast, zafirlukast, zileuton).

Variability in individual asthma treatment response may be due to many factors, including the severity and type of disease, treatment compliance, intercurrent illness, other medication taken (drug–drug interaction), environmental exposures, and age. However, there is reason to believe that genetic factors underlie much of the observed treatment response variance. A study of treatment response to glucocorticosteroids, a beta-2 adrenergic agonist, and an experimental leukotriene inhibitor has found that up to 60–80% of the variance in drug response may be due to differences between individuals. This value corresponds to the maximum limit of genetic variance, and indicates that a clinically relevant part of the response

to the main classes of asthma drugs may be due to genetic determinants (<http://www.ncbi.nlm.nih.gov/pubmed/18311188>).

To date, investigations in the field of asthma pharmacogenomics have focused on three classes of asthma therapies: β 2-agonists, leukotriene antagonists and glucocorticosteroids. Below examples are shown of pharmacogenetic studies about β 2-agonists and leukotriene antagonists in asthma.

14.6.6. Interaction between genetic variations and β 2-agonists

The 5q31-33 is an important pharmacogenomic region for asthma. β 2-agonists are used widely by inhalation for the relief of airway obstruction. These drugs act via binding to the **β 2 adrenergic receptor (ADRB2)**, a cell surface G protein-coupled receptor located on 5q32. Responses to this drug are currently the most investigated pharmacogenomic pathway in asthma. Two coding variants (at positions 16 and 27) within the *ADRB2* gene have been shown in vitro to be functionally important (<http://www.ncbi.nlm.nih.gov/pubmed/15090197>). The Gly16 receptor exhibits enhanced downregulation in vitro after agonist exposure. In contrast, Arg16 receptors are more resistant to downregulation. Because of linkage disequilibrium, individuals who are Arg/Arg at position 16 are much more likely to be Glu/Glu at position 27; individuals who are Gly/Gly at position 16 are much more likely to be Gln/Gln at position 27. The position 27 genotypes influence but do not abolish the effect of the position 16 polymorphisms with regard to downregulation of phenotypes in vitro. Retrospective studies and prospective clinical trials have suggested that adverse effects occur in patients homozygous for arginine (Arg/Arg), rather than glycine (Gly/Gly), at position 16. Bronchodilator treatments avoiding β 2-agonist may be appropriate for patients with the Arg/Arg genotype (<http://www.ncbi.nlm.nih.gov/pubmed/15500895>).

14.6.7. Interaction between genetic variations and leukotriene antagonists

Leukotrienes, released by eosinophils, mast cells and alveolar macrophages, are among the main mediators in asthma, inducing airway obstruction, migration of eosinophils and proliferation of smooth muscle. Of the three enzymes exclusively involved in the formation of the leukotrienes (5-lipoxygenase (*ALOX5*), leudotriene C4 (**LTC4**) synthase, and **LTA4** epoxide hydrolase), ALOX5 is the enzyme required for the production of both the cysteinyl-leukotrienes (LTC4 , LTD4 , and LTE4) and LTB4. ALOX5 activity in part determines the

level of bronchoconstrictor leukotrienes present in the airways, and pharmacological inhibition of the action of ALOX5 or antagonism of the action of the cysteinyl-leukotrienes at their receptor is associated with an amelioration of asthma. A polymorphism located in the promoter of the *ALOX5* gene decreases gene transcription, and less enzyme is produced when the number of repeats of an Sp1 binding motif GGGCGG, which acts as a transcription modulating site, is different from the usual number of 5 (<http://www.ncbi.nlm.nih.gov/pubmed/16364163>). In a study in the United States approximately 6% of asthma patients did not carry a wild-type allele at the *ALOX5* core promoter locus ([Drazen et al., 1999](#)). It was hypothesized that patients possessing the altered promoter might be less responsive to a leukotriene modifier. In randomized, double-blind, placebo-controlled trials of ABT-761, an ALOX5 inhibitor, which is a derivative of the antileukotriene drug Zileuton this hypothesis was investigated. The primary outcome of the clinical study was improvement in FEV1 (forced expiratory volume in 1 second). In the unstratified population, the inhibitor produced a 12% to 14% improvement in FEV1. Patients homozygous for the wild-type promoter had a 15% improvement in FEV1. In contrast, those patients homozygous for the mutant version of the promoter had a significantly decreased FEV1 response. Otherwise the *ALOX5* core promoter locus does not account for all patients who did not respond to ALOX5 inhibition, which suggests that there may be other gene defects in the pathway leading to a lack of response to this form of treatment. It was suggested that patients who fail to respond to ALOX5 inhibition are those in whom other mechanisms are responsible for asthmatic airway obstruction.

LTC4 synthase is a membrane-bound glutathione transferase expressed only by cells of hematopoietic origin and is a key enzyme in the synthesis of cys-LTs, converting LTA4 to LTC4. The gene encoding LTC4 synthase is located on 5q35. An adenine to cytosine transversion has been found 444 bp upstream (-444) of the translation start site of the LTC4 synthase gene and reported that the polymorphic C -444 allele occurred more commonly in patients with aspirin intolerant asthma (AIA) ([Sanak et al., 1997](#) and [2000](#)). A 5-fold greater expression of LTC4 synthase has been demonstrated in individuals with AIA when compared with patients with aspirin-tolerant asthma; furthermore, the expression of LTC4 synthase mRNA has also been shown to be higher in blood eosinophils from asthmatic subjects compared with control subjects and was particularly increased in eosinophils from patients with AIA. In addition, it was found that, among subjects with asthma treated with zafirlukast (a leukotriene receptor antagonist), those homozygous for the A allele at the -444 locus had a lower FEV1 response than those with the C/C or C/A genotype ([Palmer et al., 2002](#)).

It must be noted that these examples are researches, and the results have not got into clinical practice yet.

14.7. The future of pharmacogenomics

The responses to drugs are influenced by interactions between many genetic and environmental factors. As was discussed in earlier chapters and will be shown in the next chapter, complex network of interactions can be drawn from these, evaluation of which needs systems biologic approaches. But a lot of exact genomic (and epigenomic), clinical and environmental data are also needed for the proper networks, which require large prospective studies, and a lot of basic researches for better understanding the behaviour of our genome and the whole organism from molecular to whole body levels. In addition, better evaluation methods are also needed for extracting as much information from the available data as possible.

In the last decades an immense development has been achieved in these areas, but we are only at the beginning of the progress, and it is not even sure, which of the pharmacogenetic goals are achievable? Presently, for the majority of drugs, there are no reliable pharmacogenetic tests, and also there are very few personal therapies or drugs.

The question is, when it will be a reality that everybody will have genomic data, from which every physician with the help of a user-friendly decision-support system can decide, which therapy or drug will be the most effective and to which the patients have no adverse drug response.

In the beginning of the 90s, even serious experts predicted that in a few years these goals would come true. But as we have learned the immense complexity of the genome and the whole organism, it turned out that, presently it is not even known whether it will ever be a reality? In 2016 there are very few genomic results that have gone into the practice. Mainly variations in the protein coding regions with strong effects can give clinically relevant information; the effects of common variants are usually unpredictable and clinically unusable. But we are only at the beginning of this process, and regarding the huge development of the last decades, we can be sure that the number of the usable pharmacogenomic tests or personal therapies will be expanded in the future.

14.8. Literature

1. <http://www.fda.gov/>
2. <http://www.fda.gov/downloads/AboutFDA/Transparency/Basics/UCM247465.pdf>
3. <http://www.fda.gov/Drugs/ScienceResearch/ResearchAreas/Pharmacogenetics/ucm083378.htm>.
4. ENCODE Project Consortium et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012 Sep 6;489(7414):57-74.
5. Erdelyi DJ, Kamory E, Zalka A, Semsei AF, Csokay B, Andrikovics H, Tordai A, Borgulya G, Magyarosy E, Galantai I, Fekete G, Falus A, Szalai C, Kovacs GT. The role of ABC-transporter gene polymorphisms in chemotherapy induced immunosuppression, a retrospective study in childhood acute lymphoblastic leukaemia. *Cell Immunol*. 2006 Dec;244(2):121-4.
6. Erdélyi DJ, Kámory E, Csókay B, Andrikovics H, Tordai A, Kiss C, Félné-Semsei Á, Janszky I, Zalka A, Fekete G, Falus A, Kovács GT, Szalai C. Synergistic interaction of ABCB1 and ABCG2 polymorphisms predicts the prevalence of toxic encephalopathy during anticancer chemotherapy. *Pharmacogenomics J*. 2008 8: 321-327.
7. Semsei AF, Erdelyi DJ, Ungvari I, Csagoly E, Hegyi MZ, Kiszel PS, Lautner-Csorba O, Szabolcs J, Masat P, Fekete G, Falus A, Szalai C, Kovacs GT. ABCC1 polymorphisms in anthracycline induced cardiotoxicity in childhood acute lymphoblastic leukemia. *Cell Biol Int*. 2011 Sep 20. [Epub ahead of print] PubMed PMID: 21929509.
8. Tan GM, Wu E, Lam YY, Yan BP. Role of warfarin pharmacogenetic testing in clinical practice. *Pharmacogenomics*. 2010 Mar;11(3):439-48.
9. Gasche Y et al. Codeine intoxication associated with ultrarapid CYP2D6 metabolism. *N Engl J Med*. 2004 Dec 30;351(27):2827-31.
10. <http://en.wikipedia.org/wiki/Statin>
11. Kajinami K, Brousseau ME, Ordovas JM, Schaefer EJ. CYP3A4 genotypes and plasma lipoprotein levels before and after treatment with atorvastatin in primary hypercholesterolemia. *Am J Cardiol*. 2004 Jan 1;93(1):104-7.
12. Kivistö KT et al. Lipid-lowering response to statins is affected by CYP3A5 polymorphism. *Pharmacogenetics*. 2004 Aug;14(8):523-5.

13. Mangavite LM, Wilke RA, Zhang J, Krauss RM. Pharmacogenomics of statin response. *Curr Opin Mol Ther.* 2008 Dec;10(6):555-61.
14. Mangavite LM, et al.. Clinical implications of pharmacogenomics of statin treatment. *The Pharmacogenomics Journal* (2006) 6, 360–374.
15. <http://en.wikipedia.org/wiki/Clopidogrel>
16. Myburgh R, Hochfeld WE, Dodgen TM, Ker J, Pepper MS. Cardiovascular pharmacogenetics. *Pharmacol Ther.* 2012 Mar;133(3):280-90.
17. Rosenson RS. A treasure of pharmacogenomic insights into postprandial lipoproteinemia and therapeutic responses to fibrate therapy: lessons from GOLDN. *Curr Atheroscler Rep.* 2009 May;11(3):161-4.
18. Wojczynski MK et al. Apolipoprotein B genetic variants modify the response to fenofibrate: a GOLDN study. *J Lipid Res.* 2010 Nov;51(11):3316-23.
19. Liggett, S.B.: Assay Drug Dev Technol. Polymorphisms of adrenergic receptors: variations on a theme. 2003; 1: 317-326.
20. Liggett, S.B.: Pharmacogenetics of beta-1- and beta-2-adrenergic receptors. *Pharmacology.* 2000;61:167-173.
21. Martinez, F.D. et al. Association between genetic polymorphisms of the beta2-adrenoceptor and response to albuterol in children with and without a history of wheezing. *J Clin Invest.* 1997;100,3184-3188.
22. McGraw, D.W., Forbes, S.L., Kramer, L.A., Liggett, S.B.: Polymorphisms of the 5' leader cistron of the human beta2-adrenergic receptor regulate receptor expression. *J Clin Invest.* 1998;102,1927-1932.
23. Israel, E., Drazen, J.M., Liggett, S.B. et al. Effect of polymorphism of the beta(2)-adrenergic receptor on response to regular use of albuterol in asthma. *Int Arch Allergy Immunol.* 2001;124,183-186.
24. Lazarus, S.C. et al. Long-acting beta2-agonist monotherapy vs continued therapy with inhaled corticosteroids in patients with persistent asthma: a randomized controlled trial. *JAMA.* 2001;285,2583-2593.
25. Israel, E. et al. Use of regularly scheduled albuterols treatment in asthma: genotype-stratified, randomised, placebo-controlled cross-over trial. *Lancet.* 2004; 364,1505-1512.
26. Kalayci O, Birben E, Sackesen C, Keskin O, Tahan F, Wechsler ME, Civelek E, Soyer OU, Adalioglu G, Tuncer A, Israel E, Lilly C. ALOX5 promoter genotype, asthma severity and LTC production by eosinophils. *Allergy.* 2006 Jan;61(1):97-103.

27. Drazen, J.M. et al. Treatment of asthma with drugs modifying the leukotriene pathway. *N Engl J Med.* 1999, 340,197-206.
28. Drazen, J.M. et al.: Pharmacogenetic association between ALOX5 promoter genotype and the response to anti-asthma treatment. *Nat Genet.* 1999,;22,168-170.
29. Sampson, A.P. et al. Variant LTC(4) synthase allele modifies cysteinyl leukotriene synthesis in eosinophils and predicts clinical response to zafirlukast. *Thorax.* 2000,55, Suppl 2:S28-31.
30. Sanak M et al. Enhanced expression of the leukotriene C(4) synthase due to overactive transcription of an allelic variant associated with aspirin-intolerant asthma. *Am J Respir Cell Mol Biol.* 2000, 23,290-296.
31. Sanak, M. et al. Leukotriene C4 synthase promoter polymorphism and risk of aspirin-induced asthma. *Lancet.* 1997,350,1599-1600.
32. Whelan, G.J. et al. Effect of montelukast on time-course of exhaled nitric oxide in asthma: influence of LTC4 synthase A(-444)C polymorphism. *Pediatr Pulmonol.* 2003,36,413-420.
33. Hawkins GA et al. The glucocorticoid receptor heterocomplex gene STIP1 is associated with improved lung function in asthmatic subjects treated with inhaled corticosteroids. *J Allergy Clin Immunol.* 2009 Jun;123(6):1376-83.e7.
34. Tantisira, K.G. et al. Molecular properties and pharmacogenetics of a polymorphism of adenylyl cyclase type 9 in asthma: interaction between beta-agonist and corticosteroid pathways. *Hum Mol Genet.* 2005; 14: 1671-1677.
35. Tantisira, K.G. et al. TBX21: a functional variant predicts improvement in asthma with the use of inhaled corticosteroids. *Proc Natl Acad Sci U S A.* 2004;101:18099-18104.
36. Tantisira, K.G. et al. Molecular properties and pharmacogenetics of a polymorphism of adenylyl cyclase type 9 in asthma: interaction between beta-agonist and corticosteroid pathways. *Hum Mol Genet.* 2005, 14, 1671-1677.
37. Tantisira KG et al. Genomewide association between GLCCI1 and response to glucocorticoid therapy in asthma. *N Engl J Med.* 2011 Sep 29;365(13):1173-83.
38. Palmer, L.J. et al. Pharmacogenetics of asthma. *Am J Respir Crit Care Med..* 2002,15, 861-866.
39. Distefano JK, Watanabe RM. Pharmacogenetics of Anti-Diabetes Drugs. *Pharmaceuticals (Basel).* 2010 Aug 1;3(8):2610-2646.

40. Konoshita T; Genomic Disease Outcome Consortium (G-DOC) Study Investigators. Do genetic variants of the Renin-Angiotensin system predict blood pressure response to Renin-Angiotensin system-blocking drugs?: a systematic review of pharmacogenomics in the Renin-Angiotensin system. *Curr Hypertens Rep.* 2011 Oct;13(5):356-61.
41. Manunta P et al. Physiological interaction between alpha-adducin and WNK1-NEDD4L pathways on sodium-related blood pressure regulation. *Hypertension.* 2008 Aug;52(2):366-72.
42. Turner ST et al. Genomic association analysis suggests chromosome 12 locus influencing antihypertensive response to thiazide diuretic. *Hypertension.* 2008 Aug;52(2):359-65.
43. Chung CM et al. A genome-wide association study identifies new loci for ACE activity: potential implications for response to ACE inhibitor. *Pharmacogenomics J.* 2010 Dec;10(6):537-44.
44. Corvol JC et al. The COMT Val158Met polymorphism affects the response to entacapone in Parkinson's disease: a randomized crossover clinical trial. *Ann Neurol.* 2011 Jan;69(1):111-8.
45. Arbouw ME et al. Novel insights in pharmacogenetics of drug response in Parkinson's disease. *Pharmacogenomics.* 2010 Feb;11(2):127-9.

14.9. Questions

1. What main goals has pharmacogenomics?
2. What is the significance of pharmacogenomics?
3. How can genetic variations be used in clinical trials?
4. With what mechanisms can genetic variations influence the drug-response?
5. What are the difficulties of pharmacogenomic researches?
6. What diseases and what gene family are overrepresented in the FDA table with approved pharmacogenomic biomarkers in drug labels?
7. Give examples for genes influencing pharmacokinetics!
8. What and how can genetic variations of *CYP2C9* and *VKORC1* influence?
9. How can CYP chips be used?
10. What can genetic variations in butyrylcholinesterase influence?

11. What gene can influence the adverse effect of mercaptopurine?
12. What roles can ABC-transporters have in pharmacology?
13. To what gene family does the gene whose genetic variations can influence the cardiac side effects of the anthracyclines belong?
14. Give examples of a gene influencing the pharmacodynamics of warfarin!
15. What is the main effect of the statins?
16. What gene family plays an important role in the metabolism of statins?
17. What gene family plays a role in the transport of statins?
18. What is the main effect of clopidogrel and what gene family is responsible for the activation of the pro-drug?
19. What kind of study has been carried out for the investigation of the pharmacogenomics of clopidogrel in an Amish population? What has it found?
20. What gene variation influences the effect of β 2-agonists?
21. Give an example for gene influencing the effect of anti-leukotrienes!
22. What can be the possible future of the pharmacogenomics?

15. Systems biologic approach of diseases

Csaba Szalai

15.1. Introduction

In the previous chapters it has been pointed out that with the development of genomic methods, computers and bioinformatics there are new possibilities for better understanding and modeling of living organisms as complex systems, which are more similar to the reality. With the spreading of high throughput methods (microarray measurements, new generation sequencing, etc.) we can get immense amount of data, and it is well-known that these data points are not independent, but in connection and interaction with each other. If e.g. a SNP locates in the regulatory region of a gene, it influences not only the expression of this gene, but also the operation of those proteins which are in interaction with the product of the gene. Furthermore, another SNP can influence the effect of this SNP in both positive and negative ways. In a living organism these interactions are on several levels, and now it is clear that if we want to interpret the effect of a mutation or an environmental factor, we must consider these interactions. In biology, the scientific field that focuses on complex networks of interactions within biological systems and tries to map and interpret these is called [systems biology](#).

According to the definition, systems biology is the study of the interactions between the components of biological systems, and how these interactions give rise to the function and behavior of that system.

In the last years, due to the above mentioned progresses, systems biology has been developed considerably. Below, concentrating on diseases, basic terms of systems biology will be introduced, and some examples of the application and utilization of this scientific field will be shown.

15.2. Displaying interactions

In systems biology interactions are displayed in the form of networks which are often called **graphs**. These can also be called interactome networks (<http://www.ncbi.nlm.nih.gov/pubmed/21164525>). The network consists of interacting components, which are called **nodes**, and interactions are depicted as lines called **edges** (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3102045/>).

In this simplified approach, the functional richness of each node is lost. Despite or even perhaps because of such simplifications, useful discoveries can be made. As regards cellular systems, the nodes are metabolites and macromolecules such as proteins, RNA molecules and gene sequences, while the edges are physical, biochemical and functional interactions that can be identified with a plethora of technologies. One challenge of network biology is to provide maps of such interactions using systematic and standardized approaches and assays that are as unbiased as possible. The resulting networks of interactions between cellular components, can serve as scaffold information to extract global or local graph theory properties. Once shown to be statistically different from randomized networks, such properties can then be related back to a better understanding of biological processes.

Some properties of these networks were first described and published by Albert László Barabási, a physicist of Hungarian origin in [Science](#) (1999) and [Nature](#) (2000). The earliest network models assumed that complex networks are wired randomly, such that any two nodes are connected by a link with the same probability p . This Erdős–Rényi model generates a network with a Poisson degree distribution, which implies that most nodes have approximately the same degree, that is, the same number of links, while nodes that have significantly more or fewer links than any average node are exceedingly rare or altogether absent. In contrast, many real networks, from the World Wide Web to social networks, are scale-free, which means that their degree distribution follows a power law rather than the expected Poisson distribution. In a **scale-free network** most nodes have only a few interactions, and these coexist with a few highly connected nodes, the **hubs**, that hold the whole network together. This scale-free property has been found in all organisms for which protein-protein interaction and metabolic network maps exist, from yeast to human (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3102045/>).

15.3. Human interactome

Owing to the conservation of biochemical and molecular functions across species, much of our current understanding of cellular networks is derived from model organisms. Yet, in the past decade we witnessed an exceptional growth in human-specific molecular interaction data, helping us understand the interlocking networks that play a key role in human disease (<http://www.ncbi.nlm.nih.gov/pubmed/21164525>).

Most attention is focused on molecular networks, including: **protein interaction networks**, whose nodes are proteins linked to each other via physical (binding) interactions; **metabolic**

networks, whose nodes are metabolites linked if they participate in the same biochemical reactions; **regulatory networks**, whose directed links represent regulatory relationships between a transcription factor and a gene, or post-translational modifications, such as those between a kinase and its substrates; and **RNA networks**, capturing the role of RNA-DNA interactions such as small non-coding microRNAs and siRNAs in regulating gene expression. In parallel, an increasing number of studies rely on **phenotypic networks** that include: **co-expression networks**, in which genes with similar co-expression patterns are linked; and **genetic networks**, in which two genes are linked if the phenotype of a double mutant differs from the expected phenotype of two single mutants.

15.4. Disease genes in the networks

Above, the term of hubs has been mentioned, which are nodes with disproportionately many connections suggesting that in biological networks hub proteins must play a special biological role. Indeed, evidence from model organisms indicates that hub proteins tend to be encoded by essential genes, and that genes encoding hubs are older and evolve more slowly than genes encoding non-hub proteins.

The deletion of genes encoding hubs also leads to a larger number of phenotypic outcomes than the deletion of genes encoding less connected proteins. While the strength of evidence for some of these effects is still debated, by virtue of the many interactions they have, one expects that the absence of a hub would affect the function of an exceptional number of other proteins. This assumption has led to the hypothesis that, in humans, hubs should typically be associated with disease genes. Indeed, one study found that **disease proteins in the OMIM Morbid Map have more protein-protein interactions than non-disease proteins** in literature-curated protein-protein interaction databases.

Note, however, that the essential gene concept in simple organisms does not map uniquely into disease genes in humans. Indeed, some human genes are essential in early development, so functional changes in them often lead to first-trimester spontaneous abortions (embryonic lethality). Mutations in such ‘essential’ genes cannot propagate in the population, as individuals carrying them cannot reproduce. In contrast, individuals can tolerate for a long time the disease-causing mutations, often past their reproductive age. The question is, are both (disease and essential) genes associated with hubs? Goh et al found that **essential genes show a strong tendency to be associated with hubs** and expressed in multiple tissues, i.e., they tend to be located at the functional center of the interactome (Fig. 15.1)

(<http://www.pnas.org/content/104/21/8685.long>). Yet, in contrast with our initial hypothesis, **non-essential disease genes do not show a tendency to encode hubs and tend to be tissue-specific**. That is, from a network perspective, **these genes segregate at the functional periphery of the interactome** (Fig. 15.2). In summary, in human cells it is the **essential genes, and not the disease genes that are encoding hubs**. This difference can be understood from an evolutionary perspective: mutations that disrupt hubs have difficulty propagating in the population, as the absence of hubs create so many disruptions that the host may not survive long enough to reproduce. Thus, only mutations that impair functionally or topologically peripheral genes can persist, accounting for the family of heritable diseases, especially those that appear in adulthood (<http://www.ncbi.nlm.nih.gov/pubmed/21164525>).

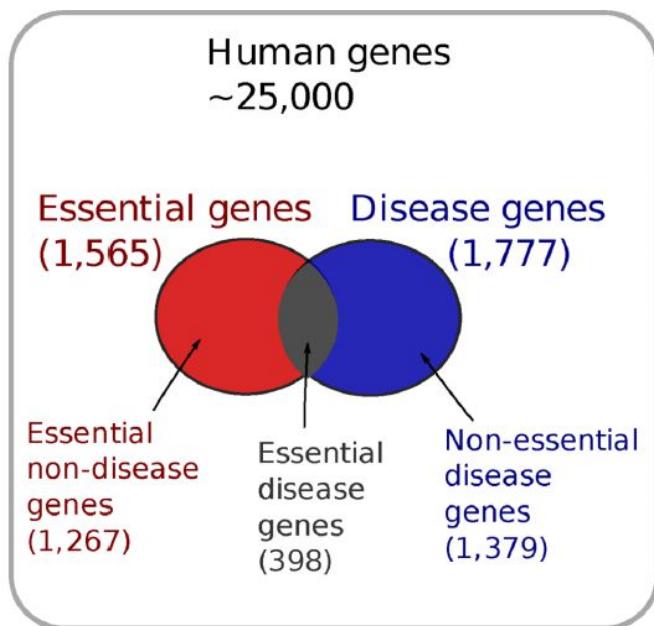


Figure 15.1. Disease and essential genes in the interactome

Of the approximately 25,000 genes, only about 1,700 have been associated with specific diseases. In addition, about 1,600 genes are known to be *in utero* essential, i.e., their absence is associated with embryonic lethality.

Source: <http://www.ncbi.nlm.nih.gov/pubmed/21164525>. 18/02/2013.

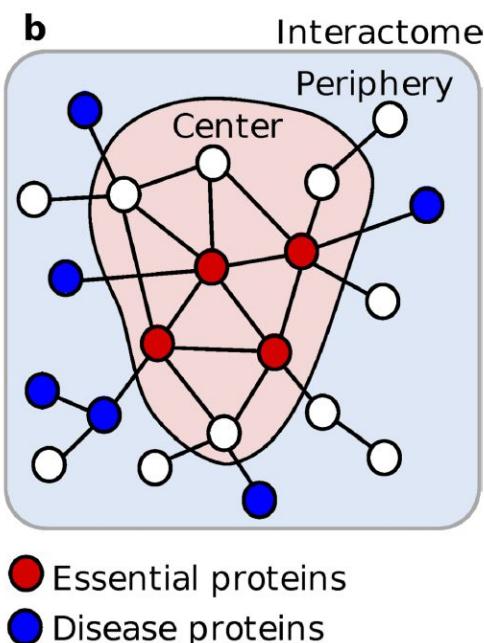


Figure 15.2. Schematic illustration of the differences between essential and non-essential disease genes. Non-essential disease genes (illustrated as blue nodes) are found to segregate at the network periphery whereas *in utero* essential genes (illustrated as red nodes) tend to be at the functional center (encode hubs, expressed in many tissues) of the interactome.

Source: <http://www.ncbi.nlm.nih.gov/pubmed/21164525>; 18/02/2013.

It must be added, however, that the above mentioned findings are referred mainly to monogenic diseases, and highly penetrant mutations (Chapter 13). Considering **low penetrant mutations, common SNPs and complex diseases, there are several examples of hub proteins** in disease networks associating with many diseases. E.g. common SNPs in the **TNF** gene are associated with asthma, atherosclerosis, obesity, T1DM, T2DM and Alzheimer disease. Similarly, β_2 adrenerg receptor (**ADRB2**) is also a hub protein. Its variations are associated with asthma, responses to drugs, obesity and hypertension. **PPARG** codes for a typical hub protein, since mutations in it can cause hypertension, obesity, T2DM, and atherosclerosis.

If a gene or molecule is involved in a specific biochemical process or disease, its direct interacting partners might also be suspected to play some role in the same biochemical process. In line with this hypothesis, proteins involved in the same disease show a high propensity to interact with each other. For example, Goh et al. observed 290 physical interactions between the products of genes associated with the same disorder, representing a

10-fold increase relative to random expectation (<http://www.pnas.org/content/104/21/8685.long>). Furthermore, it was found that genes linked to diseases with similar phenotypes have a significantly increased tendency to interact directly with each other. These observations indicate that if we identify a few disease components, the other disease-related components will likely be in their network-based vicinity. That is, we expect that each disease can be linked to a well-defined neighbourhood of the interactome, often referred to as a **disease module** (Fig. 15.3). Thus **a disease module represents a group of network components that together contribute to a cellular function whose disruption results in a particular disease phenotype.**

These disease modules can be identified by several biochemical and genomic methods, even *in silico* on the basis of currently available data using bioinformatics approaches. E.g. Chen et al., relied on co-expression networks constructed from liver and adipose tissues, facilitating the identification of sub-networks associated with genetic loci linked to obesity- and diabetes-related DNA variations (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2841398/>). The results confirmed the connection between obesity and a macrophage-enriched metabolic subnetwork, validating three previously unknown genes, *LPL*, *LACTB*, and *PPM1L*, as obesity genes in transgenic mice.

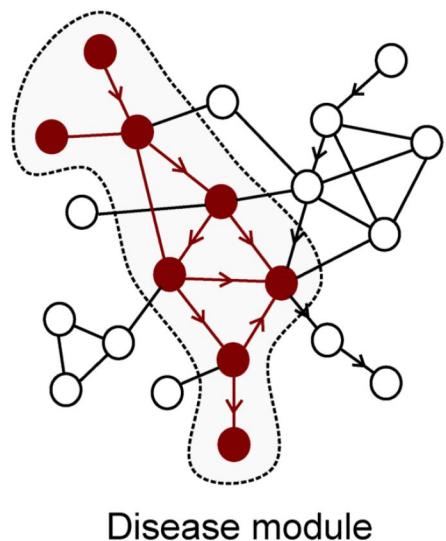


Figure 15.3. A disease module represents a group of nodes whose perturbation (mutations, deletions, copy number variations, or expression changes) can be linked to a particular disease phenotype, shown as red nodes.

Source: <http://www.ncbi.nlm.nih.gov/pubmed/21164525>; 18/02/2013.

15.5. Nodes and edges in diseases

From the interactome network theory it can be deduced that the whole human interacting network can be drawn as a complex, large network. From systems biological, network-based approach, diseases can be explained as results of the **perturbation** of this network (Fig. 15.4). Networks can be perturbed in two ways: **removing nodes** (e.g. protein deletion due to null mutation in a gene) or **modifying edges** (e.g. through a mutation in the ligand binding domain of a receptor).

The consequences on network structure and function are expected to be radically dissimilar for node removal versus edgetic perturbation. **Node removal** not only disables the function of a node, but also **disables all the interactions of that node with other nodes, disrupting in some way the function of all of the neighbouring nodes**. An **edgetic disruption**, removing one or a few interactions, but leaving the rest intact and functioning, **has subtler effects** on the network, though not necessarily on the resulting phenotype. The distinction between node removal and edgetic perturbation models can provide new clues on mechanisms underlying human disease, such as the different classes of mutations that lead to **dominant versus recessive modes of inheritance**.

The idea that the disruption of specific protein interactions can lead to human disease complements canonical gene loss/perturbation models, and is poised to explain confounding genetic phenomena such as **genetic heterogeneity**. Matching the edgetic hypothesis to inherited human diseases, approximately **half of 50,000 Mendelian** alleles available in the human gene mutation database can be modeled as **potentially edgetic** if one considers deletions and truncating mutations as node removal, and in-frame point mutations leading to single amino-acid changes and small insertions and deletions as edgetic perturbations. This number is probably a good approximation, since thus far disease-associated genes predicted to bear edgetic alleles using this model have been experimentally confirmed. For genes associated with multiple disorders and for which predicted protein interaction domains are available, it was shown that putative edgetic alleles responsible for different disorders tend to be located in different interaction domains, consistent with **different edgetic perturbations conferring strikingly different phenotypes** (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3102045/>).

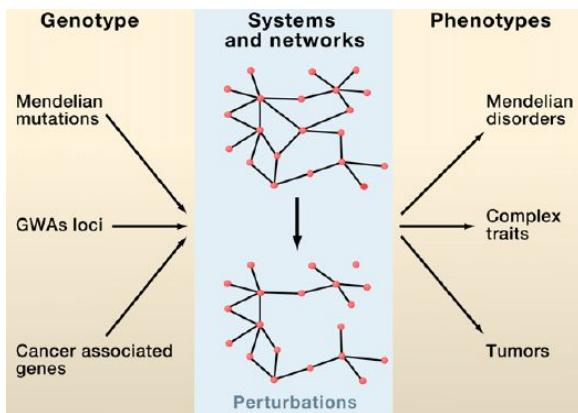


Figure 15.4. Perturbations in biological systems and cellular networks may underlie genotype-phenotype relationships.

By interacting with each other, genes and their products form complex cellular networks. The link between perturbations in network and systems properties and phenotypes, such as Mendelian disorders, complex traits, and cancer, might be as important as that between genotypes and phenotypes. There are examples of node removal as well as edge modification.

Source: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3102045/>; 18/02/2013.

15.6. Human Diseasome

The highly interconnected nature of the interactome means that at the molecular level, it is difficult, if not counter-intuitive, to consider diseases as being invariably independent of one another. Indeed, different disease modules can overlap, so that perturbations caused by one disease can affect other disease modules. The systematic mapping of such network-based dependencies between the pathophenotypes and their disease modules has culminated in the concept of the **diseasome**, representing **disease maps whose nodes are diseases and whose links represent various molecular relationships between the disease-associated cellular components**. Uncovering such links between diseases not only helps us understand how different phenotypes, often addressed by different medical sub-disciplines, are linked at the molecular level, but can also help us comprehend why certain groups of diseases arise together. The co-morbidity of conditions culled from the diseasome offers insights that may yield novel approaches to disease prevention, diagnosis, and treatment. Diseasome-based approaches could also aid drug discovery, in particular when it comes to the use of approved drugs to treat molecularly linked diseases (<http://www.ncbi.nlm.nih.gov/pubmed/21164525>).

15.7. Shared gene hypothesis

If the same gene is linked to two different disease pathophenotypes, this linkage is often an indication that the two diseases have a common genetic origin. Goh et al. used the gene-disease associations collected in the OMIM database to build a network of diseases that are linked if they share one or several genes (Fig. 15.5). <http://www.pnas.org/content/104/21/8685.long>

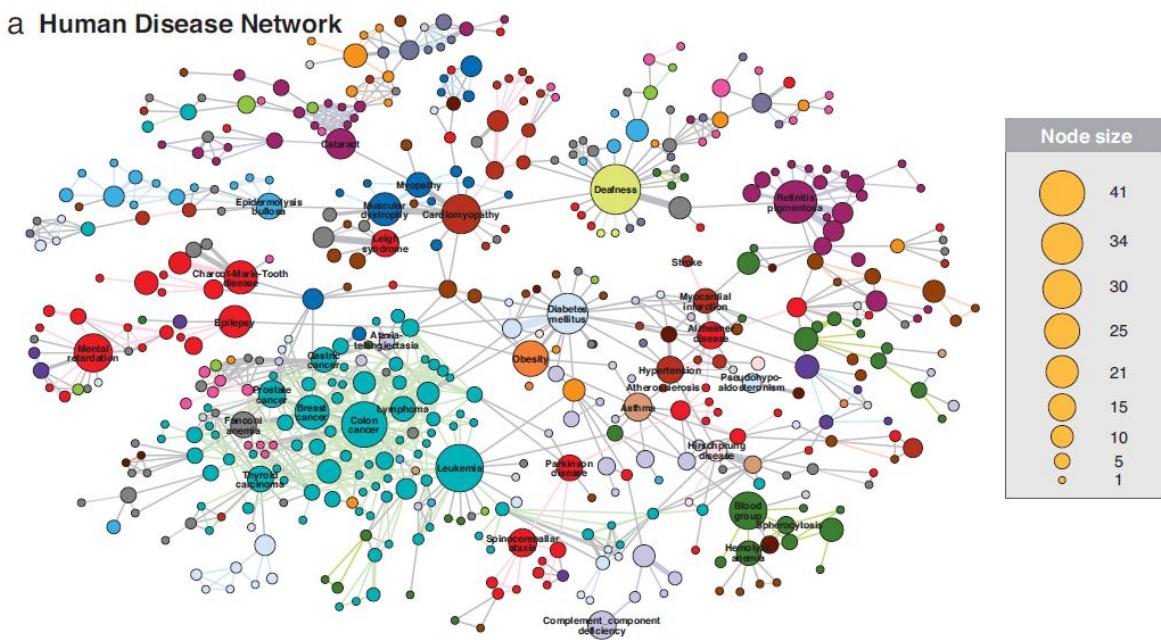


Figure 15.5. Human disease network (HDN)

In the HDN, each node corresponds to a distinct disorder, colored based on the disorder class to which it belongs. A link between disorders in the same disorder class is colored with the corresponding dimmer color; and links connecting different disorder classes are gray. The size of each node is proportional to the number of genes participating in the corresponding disorder, and the link thickness is proportional to the number of genes shared by the disorders it connects. The name of disorders with >10 associated genes are indicated.

Source: <http://www.pnas.org/content/104/21/8685.long>, 18/02/2013.

It was shown that a patient is **twice as likely to develop a (comorbid) disease if that disease shares a gene with the primary disease**, than if that disease does not share a gene with the primary disease. E.g. obese people often develop T2DM and hypertension. People with

T2DM and/or hypertension develop often atherosclerosis, and atherosclerotic people often develop Alzheimer disease. All of the diseases have co-morbidities and have shared genes.

Yet, many disease pairs that share genes do not show significant comorbidity. This lack of comorbidity may occur, in part, because different mutations on the same gene can have different effects on the function of the gene product and on its organ-based expression, therefore, different pathological consequences that are context-dependent. Such ‘edgetic’ alleles affect a specific subset of links in the interactome, and individuals who harbor different mutations in the same gene can develop different disorders. Consistent with this view, disease pairs associated with mutations that affect the same functional domain of a protein show higher comorbidity than disease pairs whose mutations occur in different functional domains (Fig. 15.6) (<http://www.ncbi.nlm.nih.gov/pubmed/21164525>).

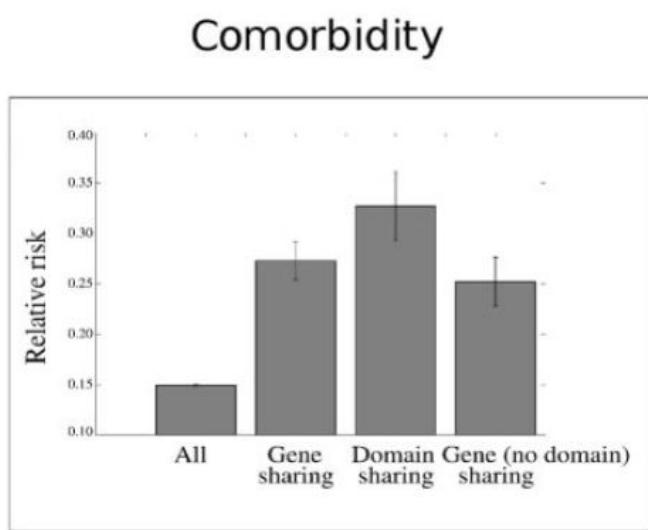


Figure 15.6.

Comorbidity between diseases linked in the HDN measured by the logarithm of relative risk, indicating that if the disease-causing mutations affect the same gene (2nd column), then the comorbidity is 2-times higher. If it affects the same domain of the shared disease protein, then the comorbidity is even higher.

Source: <http://www.ncbi.nlm.nih.gov/pubmed/21164525>; 18/02/2013.

15.8. Shared metabolic pathway hypothesis

An enzymatic defect that affects the flux of one reaction may potentially affect the fluxes of all downstream reactions in the same pathway, leading to disease phenotypes that are normally associated with these downstream reactions. Thus, for metabolic diseases, links

induced by shared metabolic pathways are expected to be more relevant than the links based on shared genes. In support of this hypothesis, [Lee et al](#) constructed a metabolic disease network, in which two disorders are connected if the enzymes associated with them catalyze adjacent reactions. Comorbidity analysis confirms the functional relevance of metabolic coupling: disease pairs linked in this network have a 1.8-fold increased comorbidity, compared to disease pairs that are not linked metabolically (<http://www.ncbi.nlm.nih.gov/pubmed/21164525>).

15.9. Shared microRNA hypothesis

Prompted by the increasing evidence of the role of miRNAs in human disease, Lu et al. connected disease pairs whose associated genes are targeted by at least one common miRNA molecule. The obtained network displays a disease class-based segregation: for example, cancers share similar associations at the miRNA level, leading to a distinct cancer cluster, which, for example, differs from the cluster associated with cardiovascular diseases, in the miRNA-based disease network (<http://www.ncbi.nlm.nih.gov/pubmed/21164525>).

15.10. Phenotypic Disease Network (PDNs)

One can also link disease pairs based on the directly observed comorbidity between them, obtaining a phenotypic disease network. For example, [Rzhetsky et al.](#) inferred the comorbidity links between 161 disorders from the disease history of 1.5 million patients at the Columbia University Medical Center, and [Hidalgo et al.](#) built a network involving 657 diseases from the disease history of over 30 million Medicare patients. In these maps, two diseases are connected if their comorbidity exceeds a predefined threshold. The PDN is blind to the mechanism underlying the observed comorbidity, which may be rooted in molecular-level dependencies, or in environmental or treatment-related perturbations of the network. Yet, PDN captures disease progression, as patients tend to develop diseases in the network vicinity of diseases they have already had. Furthermore, patients who are diagnosed with diseases with more links in the PDN, show a higher mortality than those diagnosed with less connected diseases.

Another use of phenotypic information to build a disease network was suggested by [Van Driel et al.](#), who employed text mining to assign to over 5,000 human phenotypes in the OMIM database a string of phenotypic features from the medical subject heading vocabulary. The overlap of their phenotypic descriptions was used to link various diseases, finding that phenotypic similarity positively correlates with the molecular signatures of two linked

diseases, from relatedness at the level of protein sequence to protein motifs and direct protein–protein interactions between the disease-associated proteins.

These studies indicate that the molecular-level links between the known disease components have direct epidemiological consequences, leading to observable comorbidity patterns.

While most efforts focused on the role of single molecular or phenotypic measure to capture disease-disease relationships (such as shared genes or metabolites), a comprehensive understanding requires us to inspect multiple sources of evidence, from shared genes to protein-protein interaction based relationships, shared environmental factors, common treatments, affected tissues and organs, and phenotypic manifestations (<http://www.ncbi.nlm.nih.gov/pubmed/21164525>).

15.11. Application of systems biological approaches

In Chapter 10 it was discussed that in complex diseases most variants identified until now conferred relatively small increments in risk, and explained only a small proportion of familial clustering. Gao et al (<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2818071/>) tried to alleviate this problem in a systems biological study, in which efforts have been made to prioritize positional candidate genes for complex diseases utilize the **protein-protein interaction (PPI) information**. 266 known disease genes, and 983 positional candidate genes from the 18 established linkage loci of T1DM, were compiled from the **T1Dbase** (<http://t1dbase.org>). It was found that the PPI network of known **T1DM genes has distinct topological features from others, with significantly higher number of interactions among themselves**. Then those positional candidates were defined to be new candidate disease genes that were first degree PPI neighbours of the 266 known disease genes. This led to a list of 68 genes.

Then it was investigated whether the characteristics of these correspond to those of disease genes. Disease genes have more interactions, and are cited more often in scientific papers. For the predicted genes, one may argue that their appearance in T1DM publications could be a result of their interactions with the known disease genes, as interacting genes often appear in the same publications. To address this issue, all PubMed records were excluded from the analysis of predicted genes that have cited the known T1DM genes. Out of the 68 new candidates 13 (~20%) **were cited significantly more often** than random in T1DM publications, as compared to only ~6.9% of the Human Protein Reference Database genes. This was a ~3-fold enrichment. As a group, members of the 68 list were significantly ($p < 10^{-7}$) more likely to appear in T1DM-related publications than members of a random set of 68

genes. It shows that there is a high possibility that these genes play a role in T1DM. Out of the 68 novel candidates, more than a third (24) interact with at least two known disease genes, and about a sixth (12) interacts with at least three. It shows the **connection between disease modules** and provides further proof that they are really disease genes. This is intuitive, as subsets of genes having much more interactions with each other than with others are likely to be from a same functional network module, and consequently to be involved in the same physiological processes and disease phenotypes.

The number of independent baits (known T1DM genes) for each gene was also determined. Figure 15.7 shows the PPI network of the top 5 candidates in terms of number of baits. On the top are ***ESR1*** and ***VIL2***, each with 6 baits. Interestingly, they are also among the top in terms of independent citations in T1D-related publications and network degrees. *ESR1*, or estrogen receptor 1, has been cited in 139 (124, after removing co-citation with known disease genes) T1DM-related publications, which ranked number 1 (1) out of the 68 candidates; the number for *VIL2* is 30 (29), ranked number 8. The odds ratios to random genes are all greater than 1, at 9.6 for *VIL2* and 6.2 for *ESR1*. Both have abundant interactions with other proteins, with $k=163$, #1 of the 68 for *ESR1*; and $k=43$, #11 for *VIL2*. These are within the top 2% of all genes, and both can be considered hubs. It must be added that in this case the disease is not caused by mutations in these genes, and thus here the term ‘disease gene’ has a little different meaning as previously used. But they have hub function in protein interaction networks, and part of the disease network. From all of these top 5 genes in Figure 15.7 can be shown that they have significantly more connections compared to the average and all are in pathways connected to T1DM.

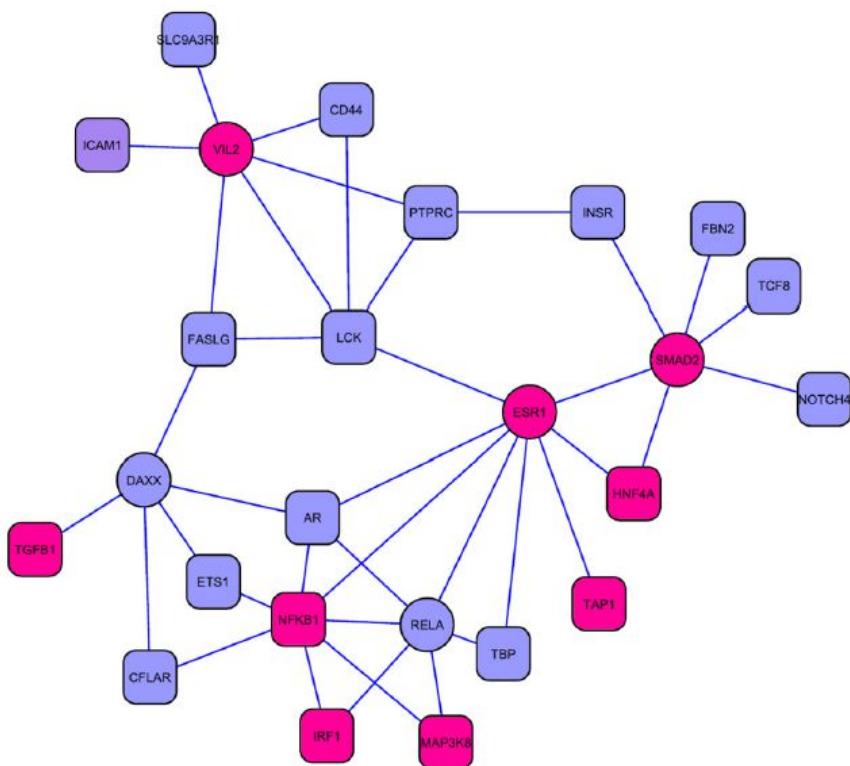


Figure 15.7. Protein-protein interaction network of the top 5 predictions (ellipse) in T1DM among the 68 proteins and their corresponding baits (round rectangle; interacting known T1DM genes). Bright magenta nodes represent genes with significant citation in T1DM-related publications ($p < 0.01$).

Source: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2818071/figure/F7/>; 18/02/2013.

New results were gained in a study, where **interaction network was deduced from GWAS results of 5 complex diseases** (<http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018660>). Five neurodegenerative and/or autoimmune complex human diseases (Parkinson's disease-Park, Alzheimer's disease-Alz, multiple sclerosis-MS, rheumatoid arthritis-RA and Type 1 diabetes-T1DM) were included. **Pathway enrichment analyses** were performed on each disease interactome independently. Several issues related to immune function and growth factor signalling pathways appeared in all autoimmune diseases, and, surprisingly, in Alzheimer's disease. Furthermore, the paired analyses of disease interactomes revealed significant molecular and functional relatedness among autoimmune diseases, and, unexpectedly, between T1DM and Alz. T1DM-Alz pair had the highest rank, followed by the autoimmune disease pairs MS-RA and T1DM-RA, and then by MS-Alz and RA-Alz. All Park pairs scored very low. These results are shown in [Figure 15.8](#) displaying a network summary of

relationships among the five diseases. This systems biological approach revealed some new and interesting results, from which some are shown below.

Numerous immune related pathways were enriched in autoimmune interactomes. This result was expected as many of the susceptibility genes in RA, T1D and MS were immune related. Notably, the pathways B-cell activation and T-cell activation appeared in all autoimmune diseases. It is well known that both arms of adaptive immunity greatly contribute to autoimmunity. Surprisingly, the same pathways appeared in Alzheimer's disease. The **role of adaptive immunity in Alz** so far has remained under-explored, however some studies suggest altered T cell phenotypes and responses in such patients. Interestingly, regular use of anti-inflammatory drugs reduces the odds of developing Alz. This observation that the Alz genetic framework may have an impact on immune function, questions the classical distinction between inflammatory and non-inflammatory diseases, and supports the hypothesis that, even though the primary insult is not inflammation but neurodegeneration, immunological pathways play a role in the etiopathogenesis of Alzheimer's disease.

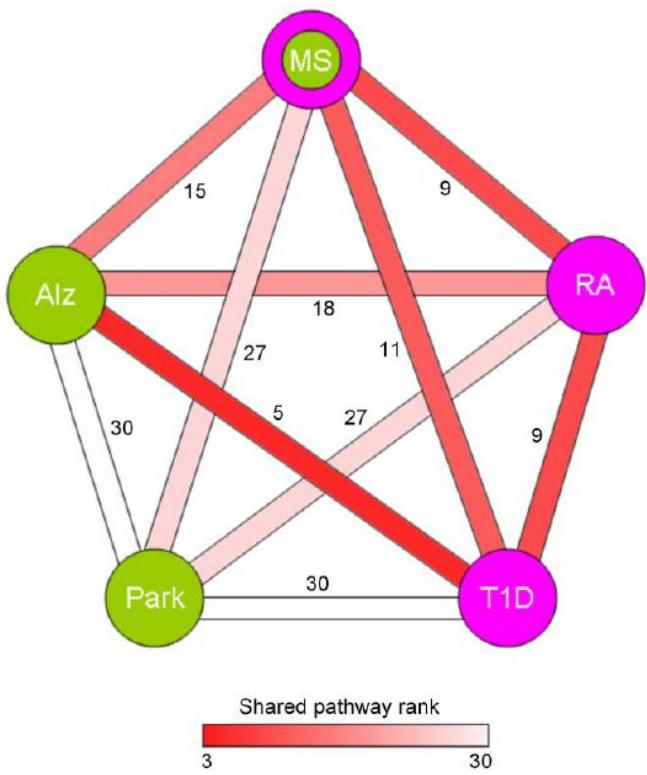


Figure 15.8. Overall disease relatedness based on shared pathways in the [Panther](#), [KEGG](#) and [CGAP-BioCarta](#) databases. Green nodes indicate the neurodegenerative disorders, whereas pink nodes highlight the autoimmune diseases. The color of the edges connecting the nodes reflects the shared pathway rank ranging from 3 (highest relatedness) to 30 (lowest relatedness).

Park: Parkinson's disease, Alz: Alzheimer's disease, MS: multiple sclerosis, RA: rheumatoid arthritis and T1D: Type 1 diabetes.

Source: <http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0018660>; 18/02/2013.

Owing to the often unknown interactions between drug targets and other cellular components, drugs whose efficacy was predicted by specific target-binding experiments may not have the same effect in different clinical settings, in which that target is of modified contextual importance (e.g. tissue-specific isoform compensates for the loss of function of the inhibited protein). Furthermore, **single-target drugs may, perhaps, correct some dysfunctional aspects of the disease module, but could alter the activity of other network neighbourhoods, leading to detectable side effects.** This network-based view of drug action implies that most disease phenotypes are difficult to reverse through the use of a single 'magic bullet,' i.e., an intervention that affects a single node in the network. While network-based approaches represent a relatively recent trend in drug discovery, given the intricate network

effects drug development must face, the nascent field of **network pharmacology**, at the intersection of network medicine and polypharmacology, is poised to become an essential component of drug development strategies. The efficacy of this approach has been demonstrated by **combinatorial therapies** of AIDS, cancer, or depression, raising an important question: can one systematically identify multiple drug targets with optimal impact on the disease phenotype? This is an archetypical network problem, leading to methods to identify optimal drug combinations starting either from the metabolic network, or from the bipartite network linking compounds to their drug-response phenotypes. Research in this direction has led to potentially safer multi-target combinations for inflammatory conditions, or to the identification of 14 optimal anti-cancer drug combinations.

Equally important, drug-target networks that link approved or experimental drugs to their protein targets have helped organize the considerable knowledge base encoding the interplay between diseases and drugs. Its analysis demonstrated the **preponderance of palliative drugs**, i.e., **drugs that do not target the actual source of the disease (i.e., the disease-associated proteins) but proteins in the network neighbourhood of the disease proteins**.

The first step of **rational drug design** is an understanding of the cellular dysfunction caused by a disease. By definition, this **dysfunction is limited to the disease module**, which means that **one can reduce the search for therapeutic agents** to those that induce detectable changes in the particular disease module. This represents a significant reduction of the search space, also aiding the development of biomarkers for disease detection, as changes in the activity of the disease module components are expected to show the strongest correlations with disease progression (<http://www.ncbi.nlm.nih.gov/pubmed/21164525>).

15.12. Literature

1. Barabási AL, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. *Nat Rev Genet.* 2011 Jan;12(1):56-68.
2. Vidal M, Cusick ME, Barabási AL. Interactome networks and human disease. *Cell.* 2011 Mar 18;144(6):986-98.
3. Barabasi AL, Albert R. Emergence of scaling in random networks. *Science.* 1999 Oct 15;286(5439):509-12. PubMed PMID: 10521342.
4. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. The large-scale organization of metabolic networks. *Nature.* 2000 Oct 5;407(6804):651-4.

5. Albert R, Jeong H, Barabasi AL. Error and attack tolerance of complex networks. *Nature*. 2000 Jul 27;406(6794):378-82.
6. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. *Proc Natl Acad Sci U S A*. 2007 May 22;104(21):8685-90. <http://www.pnas.org/content/104/21/8685.long>
7. Duarte NC et al. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *PNAS*. 2007; 104:1777–1782.;
8. Ma H et al. The Edinburgh human metabolic network reconstruction and its functional analysis. *Molecular Systems Biology*. 2007; 3:135.
9. Chen Y et al. Variations in DNA elucidate molecular networks that cause disease. *Nature*. 2008 Mar 27;452(7186):429-35. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2841398/>
10. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabási AL. The human disease network. *Proc Natl Acad Sci U S A*. 2007 May 22;104(21):8685-90.
11. Lee D-S et al. The implications of human metabolic network topology for disease comorbidity. *PNAS*. 2008; 105:9880–9885.
12. Lu M et al. An Analysis of Human MicroRNA and Disease Associations. *Plos ONE*. 2008; 3:e3420.
13. Hidalgo C et al. A Dynamic Network Approach for the Study of Human Phenotypes. *Plos Computational Biology*. 2009; 5 e1000353.
14. van Driel MA et al. A text-mining analysis of the human phenome. *European Journal of Human Genetics*. 2006; 14:535–542.
15. Edwards YJ et al. Identifying consensus disease pathways in Parkinson's disease using an integrative systems biology approach. *PLoS One*. 2011 Feb 22;6(2):e16917.
16. Gao S, Wang X. Predicting Type 1 Diabetes Candidate Genes using Human Protein-Protein Interaction Networks. *J Comput Sci Syst Biol*. 2009 Apr 1;2:133.
17. Menon R, Farina C. Shared molecular and functional frameworks among five complex human disorders: a comparative study on interactomes linked to susceptibility genes. *PLoS One*. 2011 Apr 21;6(4):e18660.
18. Binder CJ et al. (2004) IL-5 links adaptive and natural immunity specific for epitopes of oxidized LDL and protects from atherosclerosis. *J Clin Invest* 114: 427–437.
19. Taleb S, Tedgui A, Mallat Z (2010) Adaptive T cell immune responses and atherogenesis. *Curr Opin Pharmacol* 10: 197–202.)

15.13. Questions

1. What is systems biology?
2. How are the interactions displayed in systems biology?
3. What properties have the interaction networks in different organisms?
4. Give examples for the interaction networks in biologic systems!
5. Who described first the properties of the modern biologic network?
6. What are the hubs in the networks?
7. What is the link between essential genes, disease genes and hubs?
8. What is a disease module in biologic network?
9. How can diseases be explained from systems biological point of view?
10. What is the concept of diseasome and how can it be used?
11. What does the shared gene hypothesis say and what are its consequences?
12. What are the shared metabolic pathway and microRNA hypotheses?
13. What are the phenotypic disease networks and what can be their significance?
14. With systems biological methods how have new T1DM genes been detected?
15. How was it investigated whether the detected new T1DM genes corresponded to the characteristics of the disease genes?
16. What can be deduced from the interaction network of the GWAS results of 5 diseases regarding Alzheimer disease?
17. What is the significance of the network pharmacology?
18. What are palliative drugs?
19. What can be the first step of the rational drug design?

16. Bioethical and research ethical issues in genetic research

Ferenc Oberfrank and András Falus

Genetic studies in recent decades have undergone a fundamental transformation. This transformation implies consequential change in the application area as well. Genetic research is not purely a cognitive activity of certain sectors of science and technology, but also represents a kind of "cultural" function altering the society itself.

Efficiency, effectiveness, risks and perspectives are related to the values, norms and communication skills of the social environment. This pragmatic approach makes it necessary to establish and maintain a supportive social and political environment for genetic research and the practical application of the results. Due to comprehensive, multifaceted, long-term and unforeseen effects, however, in addition to the pragmatic approach, further analysis, social dialogue and consistency are also needed.

16.1. Background

In recent decades, our genetic knowledge has expanded rapidly. This is primarily owing to the development of modern biomedicine, however, it was greatly accelerated by the U.S. and British governments supporting the "Human Genome Project" and related programs. The program has fulfilled the main objective of the first stage, that is, the molecular description of the human genome, much faster than originally planned. This was due to several factors:

1. The program leaders managed to maintain a public policy and administrative support of the program, and also the social support with the assistance of the media. In spite of disputes flaring up occasionally, the majority has not cast doubts on giving priority to this program, because in the long run, benefits could be expected for most people, with acceptable and manageable risks.
2. Exploiting the funding properly, the leaders of the program involved excellent and motivated scholars, research groups and development teams, appropriate institutional arrangements, well-organized and gradually expanding international cooperation.
3. The unforeseen division of the program and the appearance of a more profit-oriented "provocative" action did not ultimately jeopardize the achievement of the original objectives; on the contrary, the competing program clearly had an accelerating effect.

This has introduced the concept of private interest, followed by a significant involvement of private funds and further actors, goals, prospects.

4. The current state of human civilization is particularly favourable for the interconnection of biomedical and technological approaches. The synergic operation of biotechnology, information technology and other new technologies (e.g. nanotechnology) has launched and maintains a huge development of biomedical research methodology and tools.

Genetic research has now been extended to multiple applications in most civilized countries, which are attempting to ensure an acceptable quality of life for their population, and maintain a sustainable economic and cultural sphere. However, it is a tall order to eliminate the undesirable implications of these practical applications, in order to ensure their ethical acceptability. The possible means to avoid these problems include both general and professional training, ethical, regulatory and compliance enforcement, proactive attention of ethics committees, regulatory activities (licensing, inspection), the autonomous engagement of professional organizations, and social dialogue.

The current situation is expected to remain valid for some time. In addition to scientific development, this is also predicted by social needs and expectations, the dominant geopolitical, economic, political and cultural trends.

One of the main fields of application of genetic research is diagnosis and treatment. The major changes in paradigm of clinical practice and medical approaches have already started and will continue based on the rapidly changing developments. Its center of gravity will gradually shift from the symptoms and treatment of patients towards the preventive studies in asymptomatic status, and personalized medical interventions are going to be repositioned.

16.2. The ethically challenging areas and of genetic research, the "border" issues

In science it is very important to be aware of the limits. Moreover, knowing the limits makes a science a real science. In other words, one does not want to draw conclusions which are not approved. The question of existence and non-existence is treated differently in philosophy, theology and biology. Based on today's scientific knowledge, the miracle of genesis has not yet been satisfactorily answered. There are hypotheses, assumptions, but no clear evidence is

available. The experts in genetics cannot really answer the question "why" concerning life. No matter how deep is their professional knowledge, one cannot claim e.g. that the DNA rules the world, and all accidental events are determined by the DNA. This is a kind of vulgar materialist "violation of boundaries". (The other "illegal border crossing" is the "vulgar theology" (e.g. the Bible is interpreted as a fundamental „word-to-word” document).

Systems biology approach of human genetics means genomics, proteomics, metabolomics and bioinformatics together. This includes Mendelian inheritance, disease threads, common non-communicable disease risk factors, personalized diagnostics and pharmacogenomics for therapy, studies on very large populations and mapping of genetic variations.

Parallel development of genetic research, biotechnology and (bio)informatics has generated the collection, recording, storage, processing and analysis of a huge amount of data. The availability, quality and access of different sources of genetic information carry a great value for potential usage. A wide range of genetic databases, biobanks were created in the recent decades. Both the research and the society are deeply concerned to preserve and disseminate this information, in order to optimally exploit these potentials in research and development of applications, health promotion, clinical and economic fields.

One may already witness an immense progress, with the development of relatively inexpensive, „easy to use,” laboratory tests for the detection of inherited monogenic diseases, and for the prediction of associated morbidity and mortality. However, in spite of the expectations, there has been no major breakthrough in the therapy of these diseases. This situation raises serious ethical issues, further complicated by the fact that anyone (without professional skill) may have an easy access to gene diagnostics tests on the internet. In this context, the most important task is to improve the medical and genetic knowledge of the social strata and ensure the availability of genetic counselling.

Some genetic conditions are inherited not only in the family, but also have ethnic, racial and social aspects. This may be associated with social stigmatization, discrimination and exclusion phenomena as well.

Genetic research of common non-communicable chronic diseases – high blood pressure, cancer, diabetes, neuropsychiatric disorders, etc. – has received attention. This is a very complex task, since many genetic factors, as well as environmental, social, economic, cultural and other factors are combined, resulting in a complex clinical picture. Serious ethical issues are raised in these studies, as interdisciplinary collaboration is needed to move forward, which means having to cross traditional biomedical research paradigm frameworks, while respecting the principles of human research ethics and standards.

The aim of pharmacogenomic studies is to clarify the role of the genetic variants (polymorphisms) in single-drug treatment responses (e.g. toxicity, efficacy, dosing), and to uncover individual differences in genetic diversity by exploring various epidemiological studies in various populations.

The promising population genetic studies carry several ethical challenges. Part of the population tracking studies are not aimed to get specific biomedical data, but to make use of the data obtained from coded or anonymous DNA samples.

Modern molecular biology, genetics and genomics science provided breakthrough in three areas of biology (and medicine as well), these are biotechnology, gene therapy and gene diagnostics. All three aspects of molecular biology raise major ethical questions.

Biotechnology generates novel compounds, drugs, pharmaceuticals with molecular biology tools. With this approach, very expensive and not sufficiently effective drugs will be cheaper and more efficient. However, it has to be noted that - because of the very similar technology - unfortunately the production of narcotic drugs will also be easier and more accessible. The plant and animal biotechnology is suitable for the creation of genetically modified organisms (GMO = Genetically Modified Organisms). On one hand, this may enhance food production at improved quantity and quality; on the other hand, its irresponsible, uncontrolled use, without international monitoring based on consensus and publicity, might cause damage to health and the environment (or even have an adverse effect on the biosphere). It is sad that in our business-oriented world benefits on both sides are exploited. It is clear that biotechnology cannot escape the economic laws, so this issue requires a complex approach.

The development of gene diagnostics is impressive. In today's gene amplification techniques, from a single hair (with a hair follicle containing a few hundred cells) full genetic identification could be performed. Increasingly sophisticated techniques (gene chips, micro beads and an automated DNA sequencer) have the ability to quickly and accurately answer genetic questions. These tests are applicable in genetic diseases, infections (e.g. at blood transfusion) and personal identification. The forensic sciences and other branches of justice (e.g. paternity issues) will also benefit from these scientific procedures. Today, in the era of genomics, gene expression patterns may reveal more variations and the application of gene diagnostics is more accurate and precise. The new science of bioinformatics is also significant. Accessing the internet, "in silico" work can be done by biologists: searching in DNA databases, they can perform advanced, useful research on the computer screen. It can be said that besides "single instrumental parts" (i.e. the individual genes), also whole "orchestral harmonies" (= patterns of thousands of genes and information content of biological pathways)

will be assessed. It is more and more feasible to make progressive genetic "predictions", to see the outcome of certain diseases (e.g. the possibility of tumor metastases in cancer), and anticipate the side effects of drugs. This latter option has huge benefits (not only financially, but also in terms of health, in the management of unnecessary delays in treatments). New personalized vaccines are under development with the computational tools of the new science of immune-genomics. It is clear, however, that with faster, more complete genetic diagnosis, professionals and the subjects of gene diagnostics have to face various new laws (labor law, insurance), and ethics ("prejudices").

The situation has become extremely difficult for doctors as to when and what to say to the patient. Even if the doctor emphasizes the eventuality and limitations of our knowledge, the patient or his relatives might insist on facing the results, and want to know the hazards and chances according to the actual state of science. An increase in the mass of accessible data is due to the availability of the exponentially improving international data bank networks. In addition to the advantages, one should see the danger of the misused interpretations, as well. The most important tool is education, the modernization of teaching biological sciences and a sober, honest, sincere dissemination of knowledge, even if the market-oriented information industry leaves less and less space for science. Anyway, there are some very encouraging trends (e.g. "University of All Knowledge"). The basic principle of genetics is still the probability of the inheritance (and not the definite faith). In spite of knowing and emphasizing this, every day there are newly occurring, sometimes ethically difficult situations. Perhaps even more problems emerge in connection with gene therapy (manipulation of genes). Gene therapy means gene transfer or modification in human cells (DNA) with some genetic dysfunctions which cause a disease. Although there are much more failures in this area than successes, yet seductive promises of healing tend to overshadow the legitimate scientific skepticism warning us to take caution and moderation. While it is true that more and more successful techniques (e.g. gene silencing) are available for genetic improvement (the more accurate knowledge of genomics of the human genome also helps), but we are still far from real successes in gene therapy. The sensation-hype trends in the commercial mass media are worsening the public reaction, by drawing an unrealistic picture for the public. Hopefully, attractive and meaningful new scientific dissemination organs will gain ground in this area as well.

Today it is a more or less universally accepted agreement that, as long as there are no technical barriers, diseases can be cured by genetic tools (probably including disease prevention), but skills, mental abilities are not permitted to be improved. It should be noted,

however, that the boundaries between the two concepts are not clear enough, which raises several additional ethical problems. In any case, perhaps fortunately, in spite of the scientific reductionism excesses, now it is quite clear that those processes of brain-psychic-emotional intelligence cannot be interpreted or modified by genetic methods (at least not more than by chemical, pharmacological effect).

16.3. The biobanks

The biobanks are collections of the materials removed from the biological samples (organs, tissues, cells, DNA, etc.). Data banks collect, store (under appropriate conditions), preserve and protect the samples and the information. The word „biobank” carries a general meaning. Each biobank is distinguished from each other according to the living source (human, dog, wheat, yeast biobank, etc.) or the living tissues stored (human blood, human kidney tumors, human DNA biobank, etc.).

The DNA biobanks (genetic biobank) are collections where not only organs, tissues, DNA, but also the genetic information of organisms (or genomic DNA, etc.) are stored.

However, the biobanks are more than simple repositories of the samples. Each biological sample is stored together with a data set that characterizes the living creature. Separate legislations control the collections of different organisms. Obviously, the human biobanks are the most regulated by law.

In human biobanks it is known of each biological sample whom it belongs to. Human biobanks frequently contain detailed clinical (medical) data, if a person belongs or belonged to a disease group. The person providing the biological sample should sign an information statement (informing about the nature of the biobank, biological sampling method, and the possible side effects of sampling information) and an informed consent form. The personal data of the person with the given biological sample are securely stored in biobanks in written and/or electronic form, regulated by data protection as codified by law.

In human biobanks the biological samples should be kept strictly anonymously. Each biological sample receives a registration number. The identity of biological sample has to be kept (name, address or date of birth). The registration numbers of biological samples, as well as the personal and clinical data are stored in electronic form. The electronic data storage is hidden by appropriate security bars. In some cases, the sample receives a final anonymous code (e.g. genetic characterization of populations), so it has never been possible to connect a sample to an individual. More often, however, the pseudomisation is the appropriate

procedure where the identity of the donor may be protected by a code or multiple codes, and only the doctor (under the pledge of oath of Hippocrates) is entitled to learn the identity, and solely for medical reasons.

A large number of biological samples is essential for the scientific study of certain diseases and other biological processes (e.g. when testing rare, but highly effective alleles). The traditional procedure is very slow (starts with collection of the biological sample by inviting patients, building up biobank structure *de novo*, etc.), it impairs the research because months or years are required for the appropriate number (hundreds or thousands) of sample collection. The carefully created international biobanks accelerate the research, because samples and clinical data are available prior to the research.

In scientific research, samples can only derive from biobanks with an explicit ethical permission (as these studies are not carried out for medical reasons).

Blood banks are specific types of biobanks, where the volunteers (donors) give blood. These biobanks are specialized to collect donated blood, which are used in surgery; tissue banks are used for transplantation (e.g., corneal transplantation for corneal banks). These biobanks store samples for a relatively short period of time because the viability of the cells is limited. Samples for laboratory diagnostic tests should be discarded after performing the test (after diagnosis). The procedure is also strictly regulated by law.

16.4. Some general ethics-related issues

As usual, the scientists involved in genetic research also claim the freedom of research, which can only be restricted by transparent and predictable rules. The scientific community generally accepts the limitations that relate to the value of human life and the dignity of the human research subject. Sensitive fields are, however, those where social consensus cannot be achieved, such as when the human life begins, the ethical status of the embryo and the fetus, the resulting research opportunities, such as embryo research, obtaining stem cells for research purposes, but there may be sharp debates about animal experimentation as well. Both researchers and society are divided on these issues, which are very difficult to overcome. In this case the necessary means to ensure public trust is a renewable social dialogue, and researchers should undertake a public burden.

Another common ethics-related problem in the area of the research is the selection of topics. It is well known that the primary sources of funding are going to the problem-solving research of developed countries. Unfortunately, the overwhelming majority of the population of

underdeveloped countries with much less resources receives much less attention. However, sometimes it happens that, when the research in the developed countries cannot be executed due to ethical constraints, it is continued in undeveloped countries.

The various areas of genetic research, the methods used in many traditional researches raise ethical questions. Based on the research, ethics gradually developed after the 2nd World War (e.g. Nuremberg Code, Helsinki Declaration, Belmont Report, CIOMS, UN, UNESCO, Council of Europe documents, national and international law), these issues are handled well by the institutions (research institutes, ethics committees, scientific councils, national and international professional and political organizations, authorities and agencies).

16.5. The specific genetic research bioethics and research ethics

Genetic research generates very new ethical issues, appearing in contexts previously not known, which means they cannot be answered in the "traditional" way.

One such key issue is the question of the informed consent. This raises a series of theoretical and practical problems for biobanks, bio-libraries, sample collections for genetic information in connection with exploiting the opportunities offered for scientific investigation.

Another important issue is the ownership of genetic information and the right of participation in commercialization benefits.

16.6. The ethics issues of commercialization of genetic information

The practice of bio-innovation and intellectual property management in the non-human biotechnology discoveries has been previously established. The benefits of patent protection, the economic and commercial potential, have well developed in the market. This raised very serious ethical issues and generated controversy. These issues are particularly challenging in the context of human genome.

It is a widely accepted principle, also confirmed by international law that the human genome is the common heritage and property of mankind, and the results of genetic research provide scientific evidence that humans living today share a common origin. On this basis, only the common interest and charitable purpose could be acceptable to exploit the research results, and unrestricted access to them has to be ensured. It is very difficult, however, to put these noble principles into practice. A system should be developed, which is based on these common values and common interests, but also provides the advantage that personal

motivations (scientific knowledge, ambition) and economic efficiency (value for money, efficiency) could be achieved. It is also very important that the system be fair: all who contribute should benefit from the results.

The bio-innovation system is first and foremost determined by the standards that can be derived from the values using the principles. All of these are deeply rooted in the society and culture, the political and professional institutions of which create them.

16.7. The genetic research, biobanks, data management and ethics legislation

The genetic research and applications described above brought about the development of the ethical rules. Various standards, declarations, guidelines, rules, and soon national and international legal documents, contracts have been created. Over the past twenty years, many of these standards have been published. Serious problems are the variety of standards, their owners, as well as the heterogeneity and the fact that their nomenclature is not unified. The goal in each case, taking into account the cultural diversity, is the development of a globally uniform and consistent professional and ethical legislation, its common maintenance and enforcement. It is a discouraging experience that the nascent consensus rules are too general to be applicable. The practical norms, however, in many important questions remain in the draft, because the different nations cannot agree on them.

Significant professional support was given by the ELSI section of Human Genome Project Program, which is a funded program for serious ethical, legal and social aspects to be tested. Later, the European Union has also launched similar programs.

The ethical principles of genetic research can be partly derived from the Nuremberg Code, the Helsinki Declaration (1964) and repeated amendments, the Council for International Organization of Medical Sciences (CIOMS) guidelines. A number of aspects of health and life science issues derived from relevant French National Ethics Advisory Council and the U.S. National Bioethics Advisory Commission declarations. Their positions give important information, but sometimes generate controversy. An important forum for negotiating standards is the national bioethics committee's global summits.

A specific example of the international ethical and legal norms was released as the European Convention on Human Rights and Biomedicine (Oviedo Convention, 1997), which, despite all the refraining, embodies a European consensus on genetic research and the practical application of the principles. Additional Protocol prohibits reproductive cloning of human beings. Politics has also manifested, when the European Parliament made its decision on

September 7, 2007 about the prohibition of therapeutic cloning of human embryos. In many countries, including Hungary, the prohibition law had already been established.

The UNESCO Universal Declaration on the Human Genome and Human Rights (1997) is ceremonial, but gives only a little practical help; still it could lead to a global consensus that research ethical rules be respected around the world. The UNESCO released an International Declaration on Human Genetic Data on the use and protection of genetic data as well. The World Health Organization (WHO) also released a number of decisions, draft international guidelines, reports and recommendations since 1998. Of these, perhaps the most important was the one about the genetic databases (2003).

Ethical standards in relation to the utilization of human genetic innovation results are included in the European Directive on the Legal Protection of Biotechnological Inventions (1998) and the European Patent Convention as well.

The declaration of HUGO on DNA samples in 1998 was a ground-breaking document, which was followed by the recommendations on DNA Banking by the Ethics Committee Recommendations of the Royal College of Physicians, United Kingdom (2000). The declaration called Opinion on biobanks for research by the German National Ethics Council (2004) and the joint resolution of the governing biobanks previously published by the French and German National Ethics Committee (2003) were milestones. An important document was released by the European Society of Human Genetics, which is a recommendation of the DNA data banking (2001) and the proposal of the Council of Europe on the regulation of the biomedical archiving of human biological materials (2003).

The pioneer national legislation of genetic research, biobanks and data protection issues was published in Australia, Singapore, the U.S., France and Canada (Quebec). The Hungarian Human Genetic Law, based on a serious professional work, was born in 2008. One must mention the activities of the Australian and Canadian legal reform committees, as the effects of their activities in these areas exceeded the national framework.

In the future, one can expect specific professional and ethical problems regarding national genomics programs, such as the first Icelandic medical database (in the cooperation of the Icelandic government and a company called deCODE), the Estonian Genome Project and the residential database of Tonga.

Genetics has an epoch-making significance, which was recognized and reflected in world affairs: the United Nations Millennium Declaration (2000) dealt with it and the G8 Summit has also repeatedly taken sides on the issue.

16.8. Conclusion

Obviously, human knowledge cannot be separated from the mental and physical events of the society, which we all must be aware of in the era of genetics and genomics. Seeking renewal and strengthening, our nation has to adapt to the consequences of the development of our knowledge in many spheres (economic, legal, ethical, religious life). We must proceed towards the light against the unscientific darkness. This may serve clear priorities, values and commitment, and the accurate designation of personal responsibility.

16.9. Bibliography

Bernice Elger, Nikola Biller-Andorno, Alexandre Mauron and Alexander M. Capron (ed.): Ethical Issues in Goberning Biobanks – Global Perspectives; Ashgate (2008)

Ferencz Antal, Kosztolányi György, Falus András, Kellermayer Miklós, Somfai Béla, Jelenits István, Hámori Antal: Biogenetika és etika (Sapientia füzetek 4.); Vigília Kiadó (2005)

The Advisory Committee on Health Research: Genomics and World Health; World Health Organization (2002); Jan Helge Solbakk, Soren Holm, Bjorn Hofmann: The Ethics of Research Biobanking; Springer (2009)