




Zaawansowane Metody Analizy Danych w Biologii Molekularnej

Semest Letni 2018



Źródła zmienności genetycznej - stany chorobowe

Semest Letni 2018

Naturalne zróżnicowanie

GENES AND PHENOTYPES

Gene: a functional unit of inheritance, usually corresponding to the segment of DNA coding for a single protein.

Genome: all of an organism's DNA sequences.

locus: the site of the gene in the genome

alleles: alternative forms of a gene



Wild-type: the normal, naturally occurring type

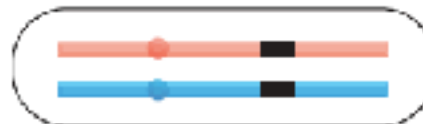


Mutant: differing from the wild-type because of a genetic change (a mutation)

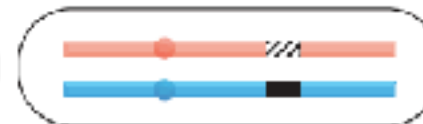
GENOTYPE: the specific set of alleles forming the genome of an individual

PHENOTYPE: the visible character of the individual

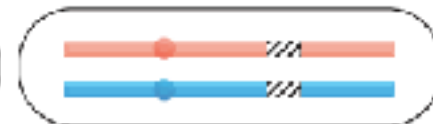
homozygous A/A



heterozygous a/A



homozygous a/a

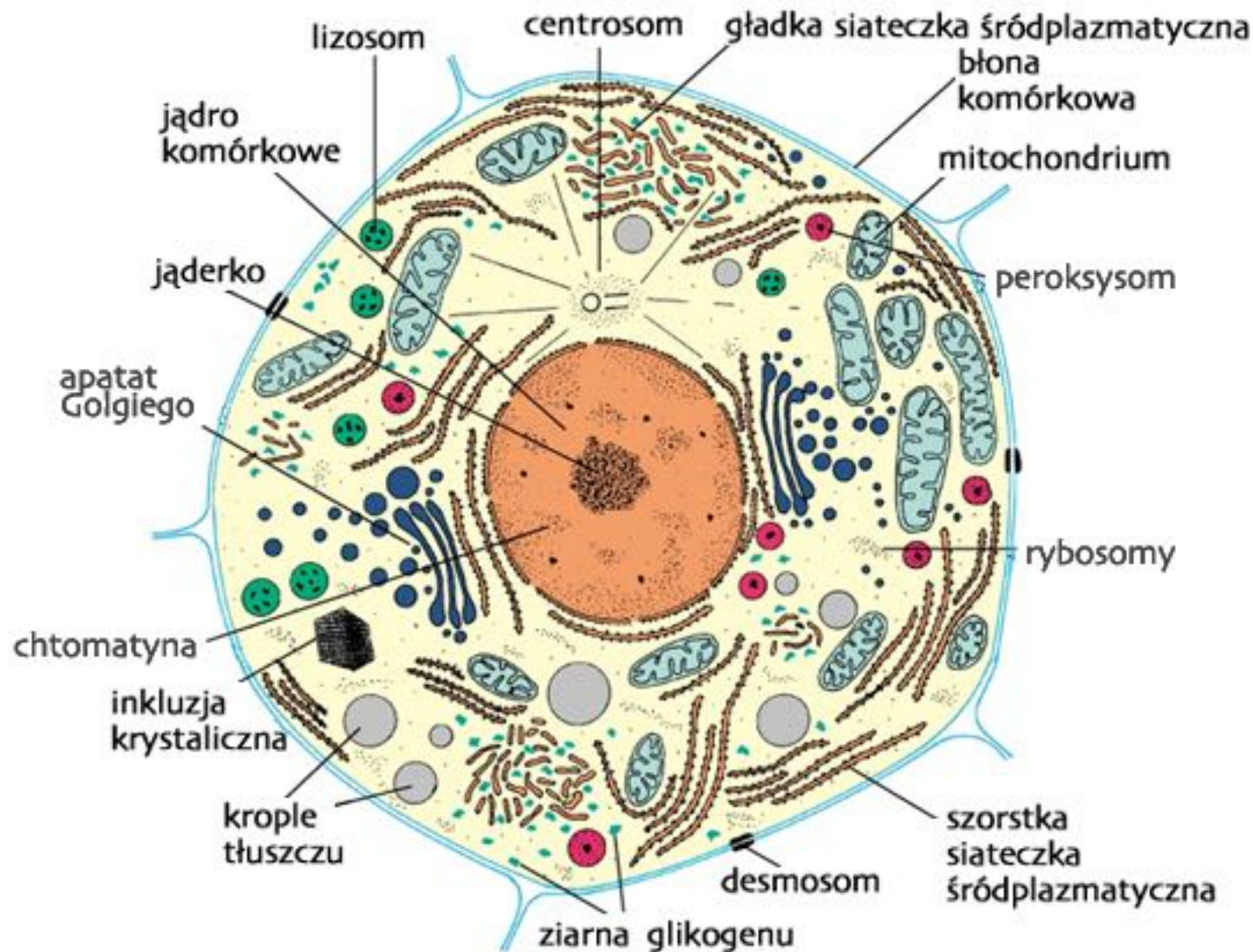


allele A is **dominant** (relative to a); allele a is **recessive** (relative to A)

In the example above, the phenotype of the heterozygote is the same as that of one of the homozygotes; in cases where it is different from both, the two alleles are said to be co-dominant.



Komórka eukariotyczna



Cykl komórkowy

Prosty fakt stojący u podstawy życia organicznego, aby otrzymać jedną żywą komórkę, to powielamy już istniejącą komórkę.

Każda żywa komórka posiada zestaw cech pozwalający na zachowanie ciągłości życia.

Zdolność do przekazania materiału genetycznego komórkom potomnym.



System kontroli cyklu komórkowego

Kaskada reakcji chemicznych uruchamiająca kolejne procesy

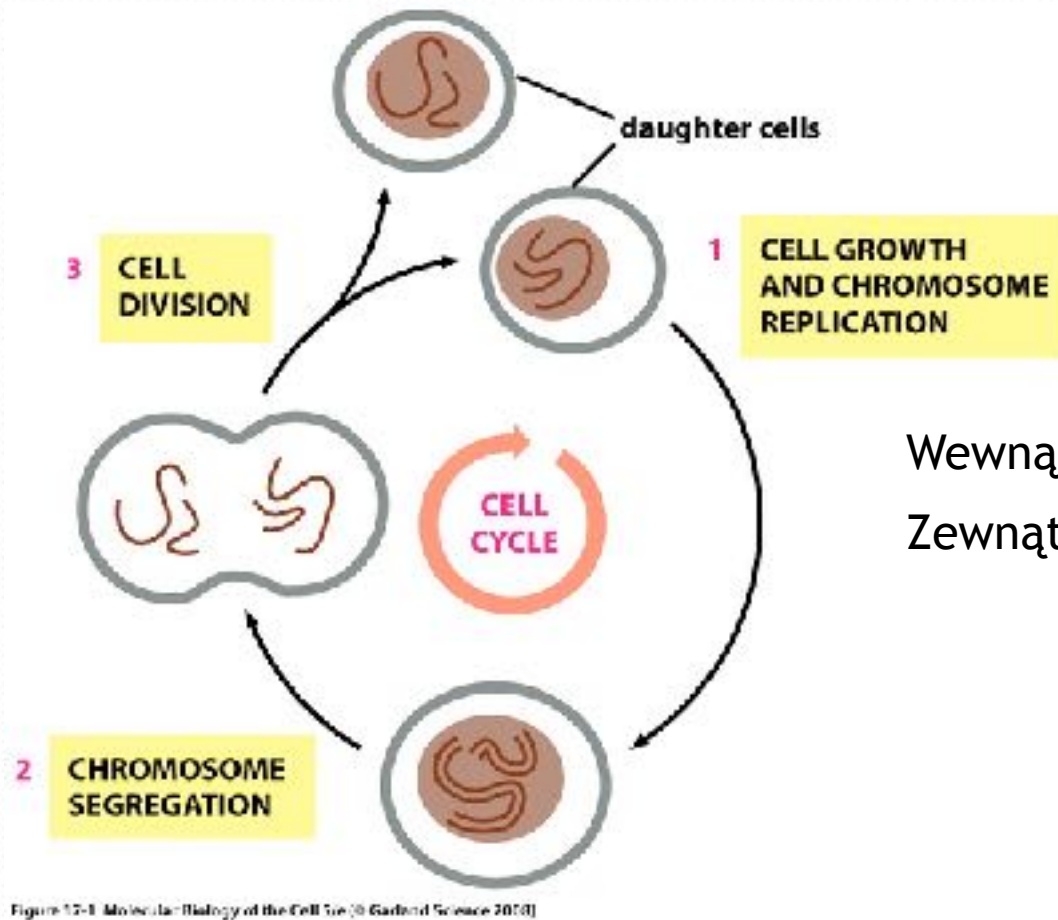


Figure 17-1 Molecular Biology of the Cell 5e (© Garland Science 2010)

Duplikacja chromosomów
Segregacja chromosomów

Wewnątrzkomórkowy system kontrolny
Zewnątrzkomórkowy system sygnałowy



Główne etapy cyklu komórkowego

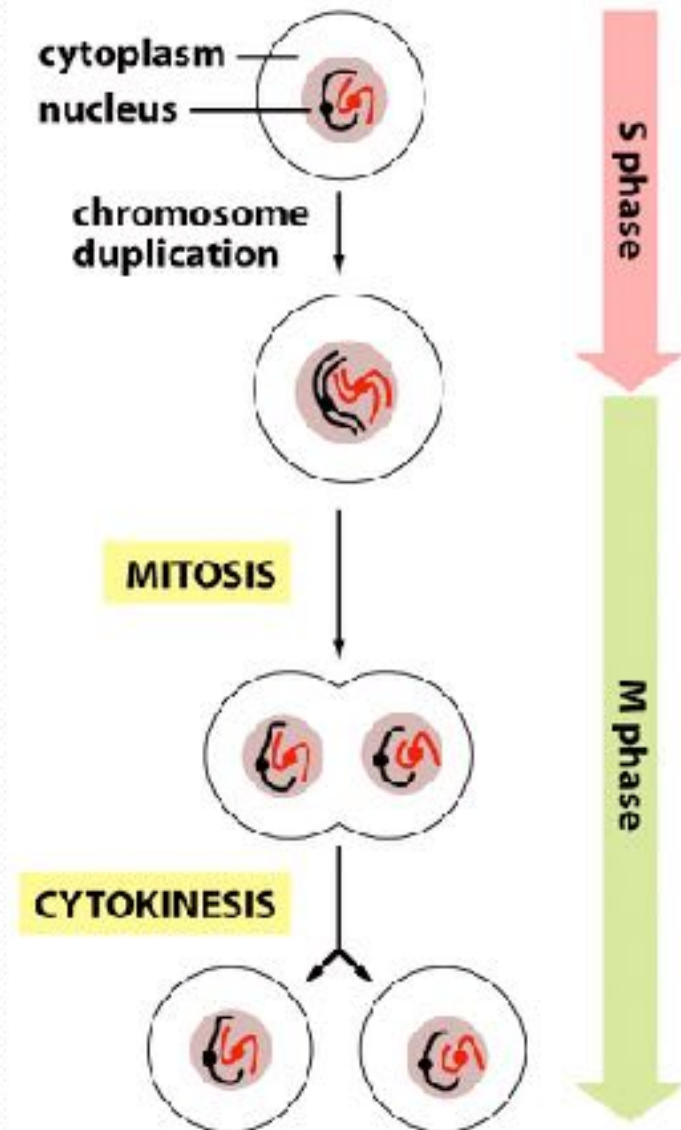


Figure 17-2 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Duplikacja chromosomów

S phase - od DNA synthesis

10-12 godzin

M phase - od mitosis

Mniej niż 1h w komórkach ssaków



Główne etapy cyklu komórkowego

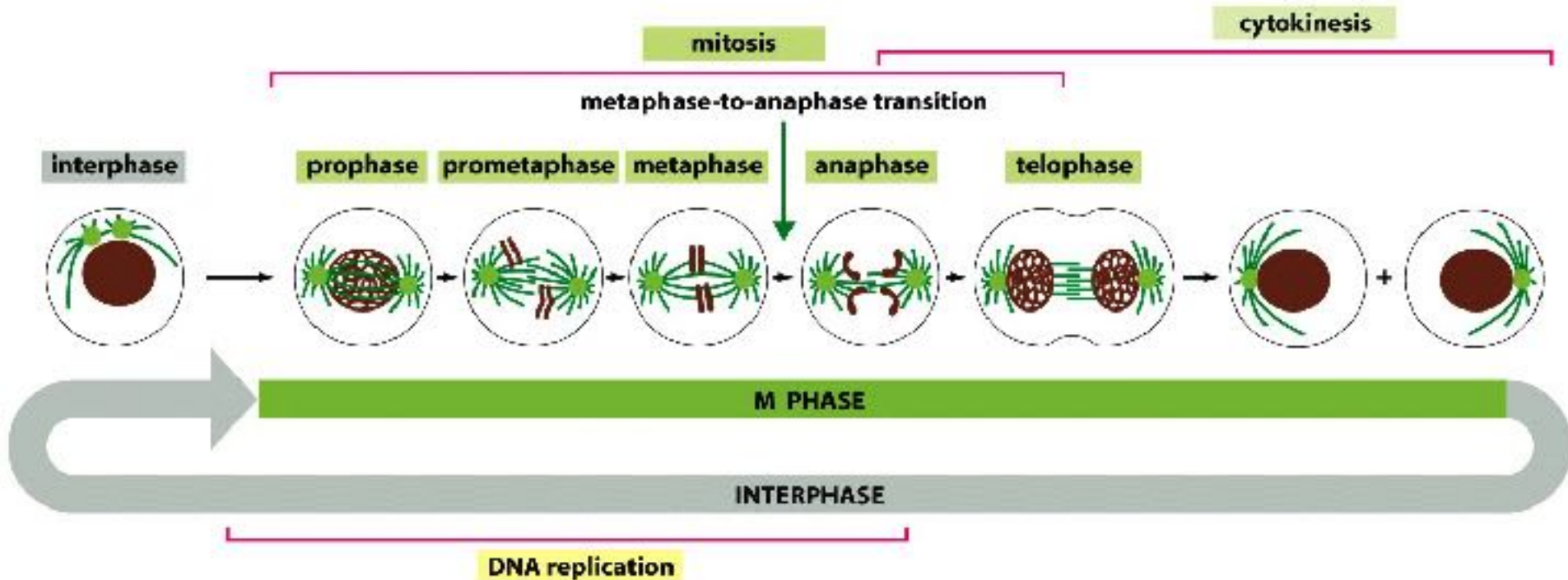


Figure 17-3 Molecular Biology of the Cell 5/e (© Garland Science 2008)



Rozmnażanie organizmów haploidalnych i diploidalnych

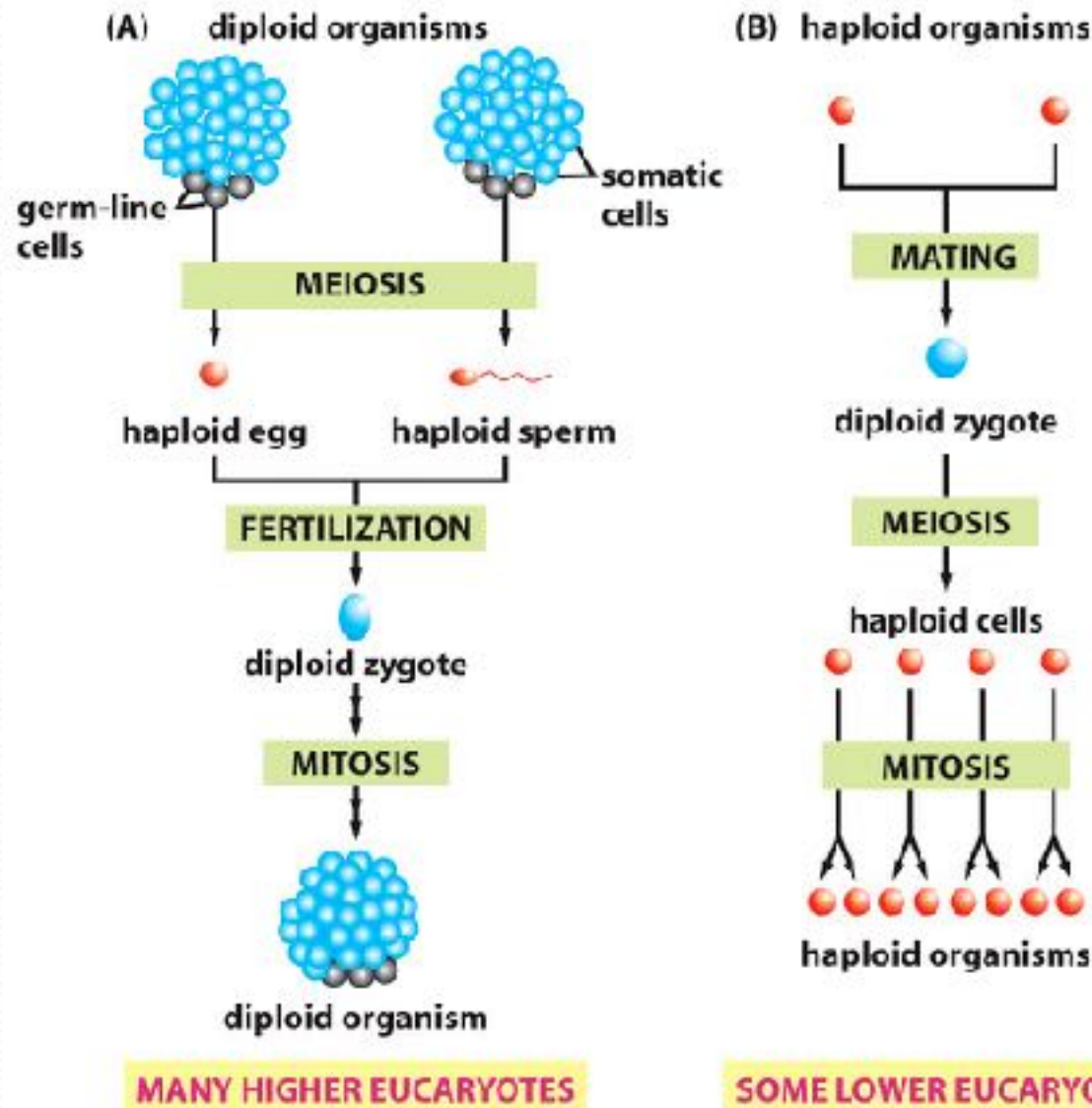
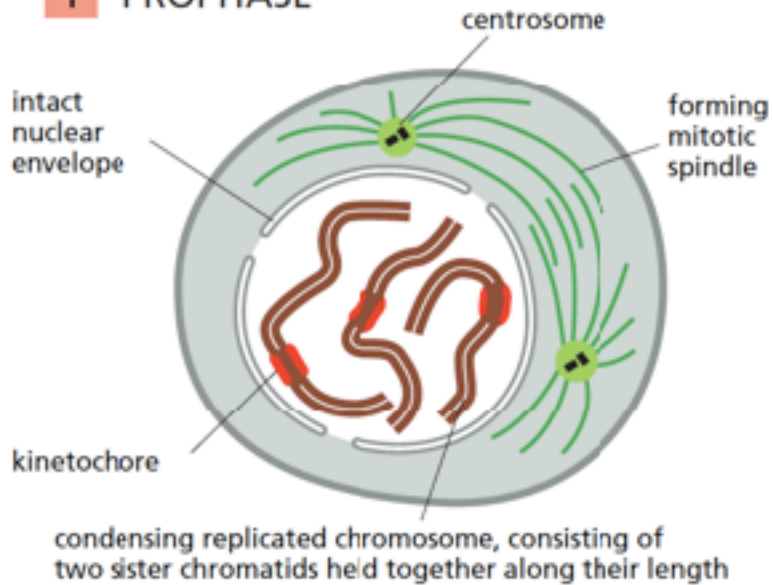


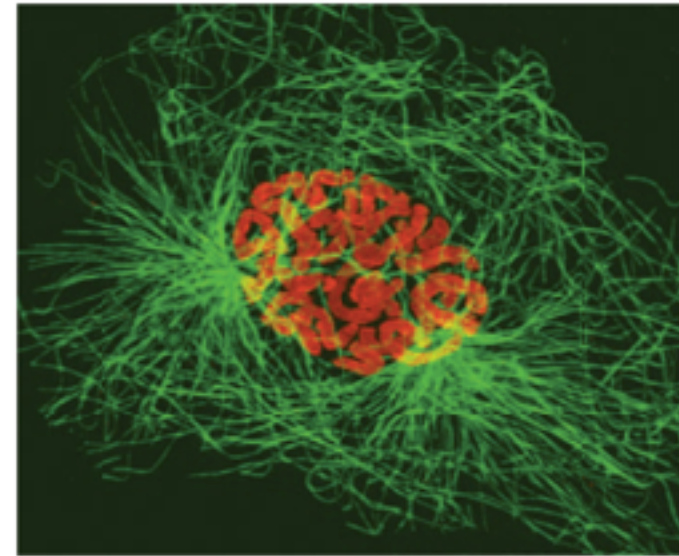
Figure 21-3 Molecular Biology of the Cell 5/e (© Garland Science 2008)



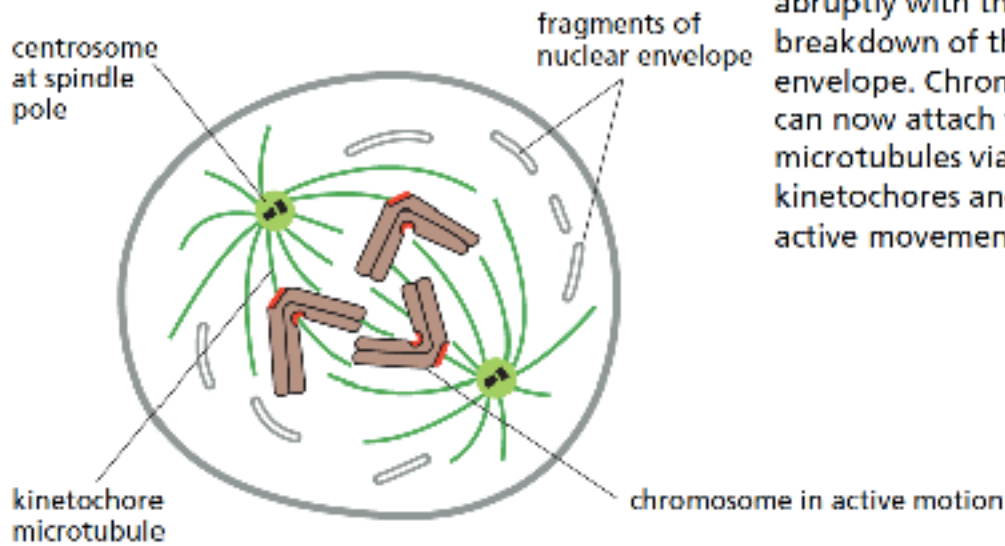
1 PROPHASE



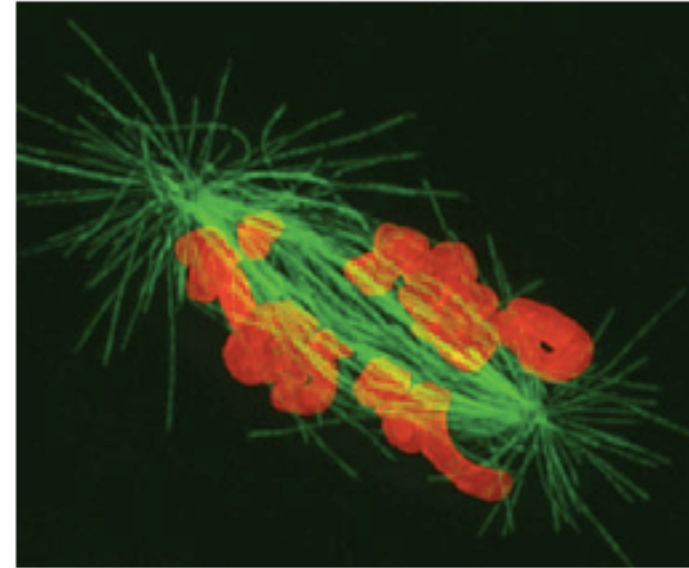
At **prophase**, the replicated chromosomes, each consisting of two closely associated sister chromatids, condense. Outside the nucleus, the mitotic spindle assembles between the two centrosomes, which have replicated and moved apart. For simplicity, only three chromosomes are shown. In diploid cells, there would be two copies of each chromosome present. In the photomicrograph, chromosomes are stained *orange* and microtubules are *green*.



2 PROMETAPHASE



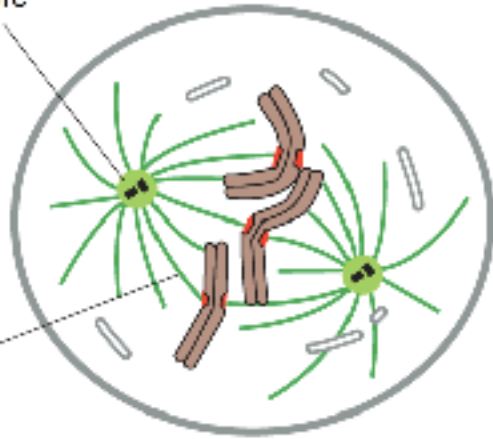
Prometaphase starts abruptly with the breakdown of the nuclear envelope. Chromosomes can now attach to spindle microtubules via their kinetochores and undergo active movement.



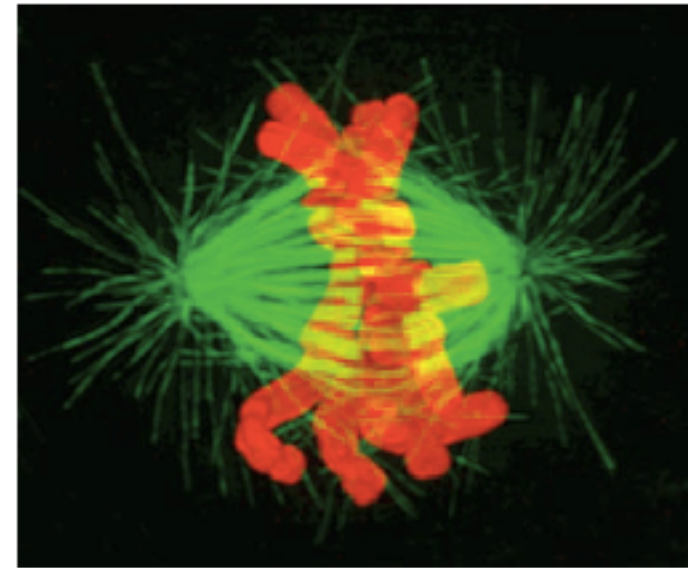
3 METAPHASE

centrosome at
spindle pole

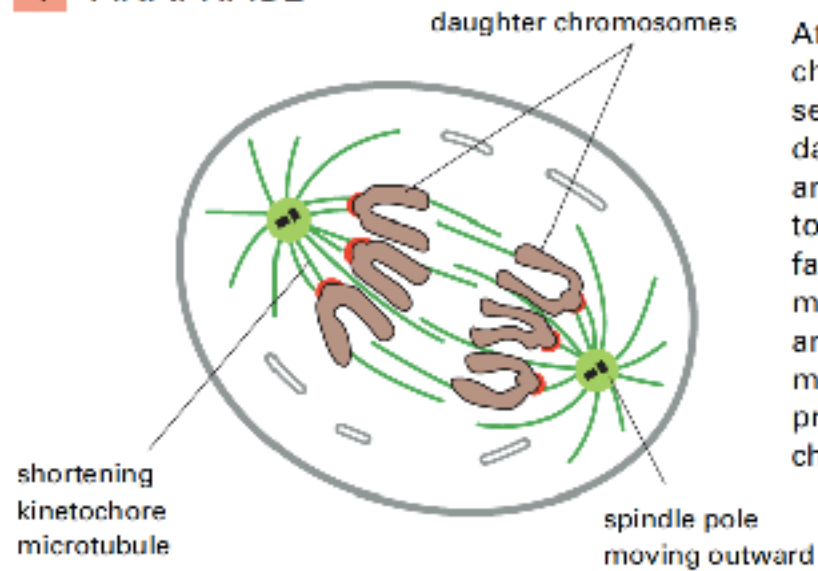
kinetochore
microtubule



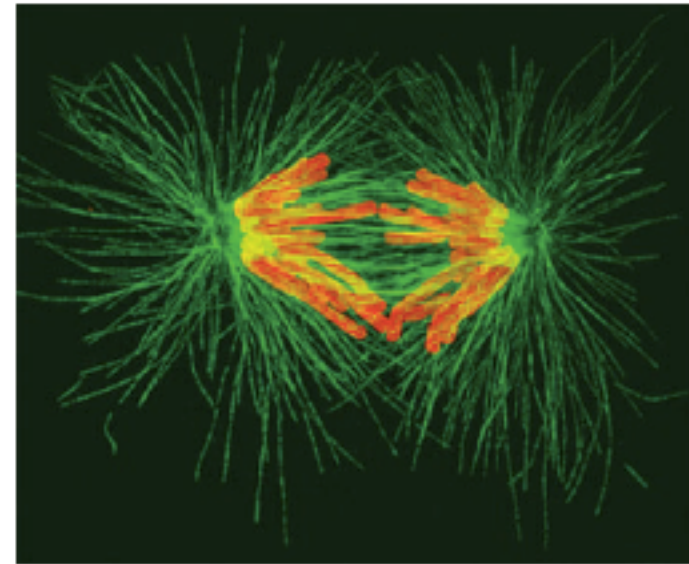
At **metaphase**, the chromosomes are aligned at the equator of the spindle, midway between the spindle poles. The kinetochore microtubules attach sister chromatids to opposite poles of the spindle.



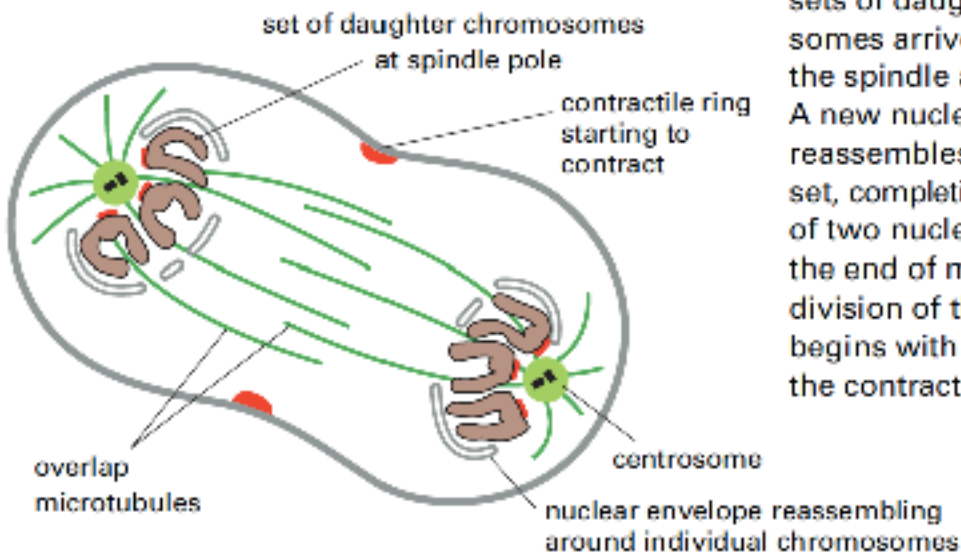
4 ANAPHASE



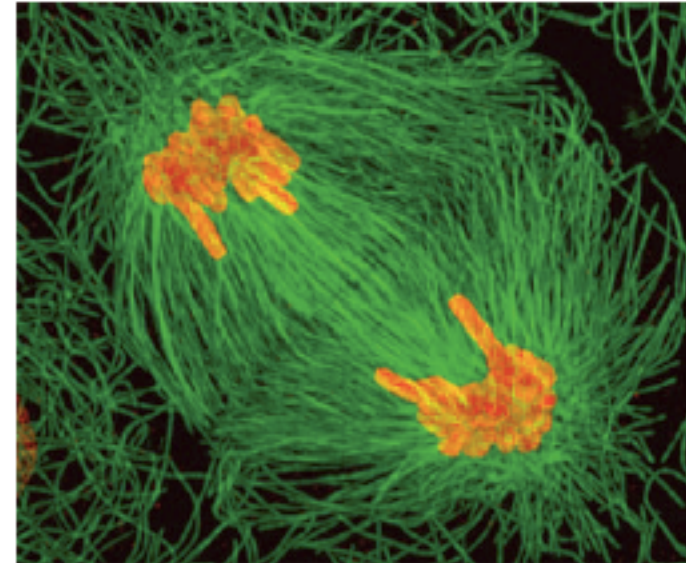
At **anaphase**, the sister chromatids synchronously separate to form two daughter chromosomes, and each is pulled slowly toward the spindle pole it faces. The kinetochore microtubules get shorter, and the spindle poles also move apart; both processes contribute to chromosome segregation.



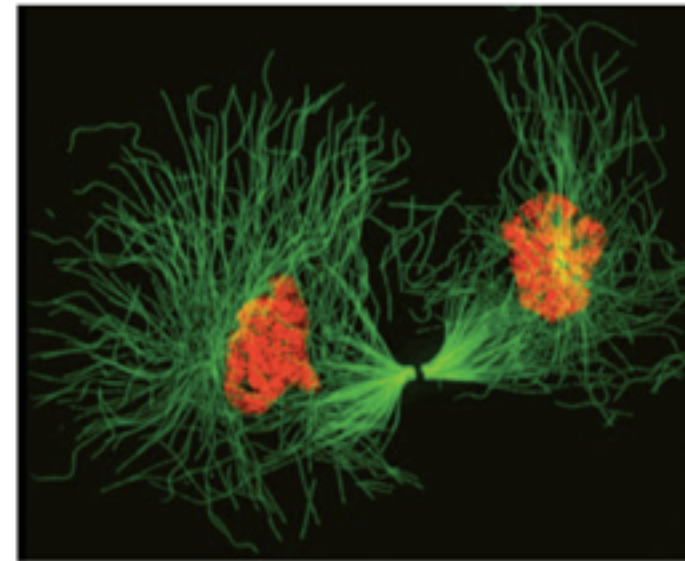
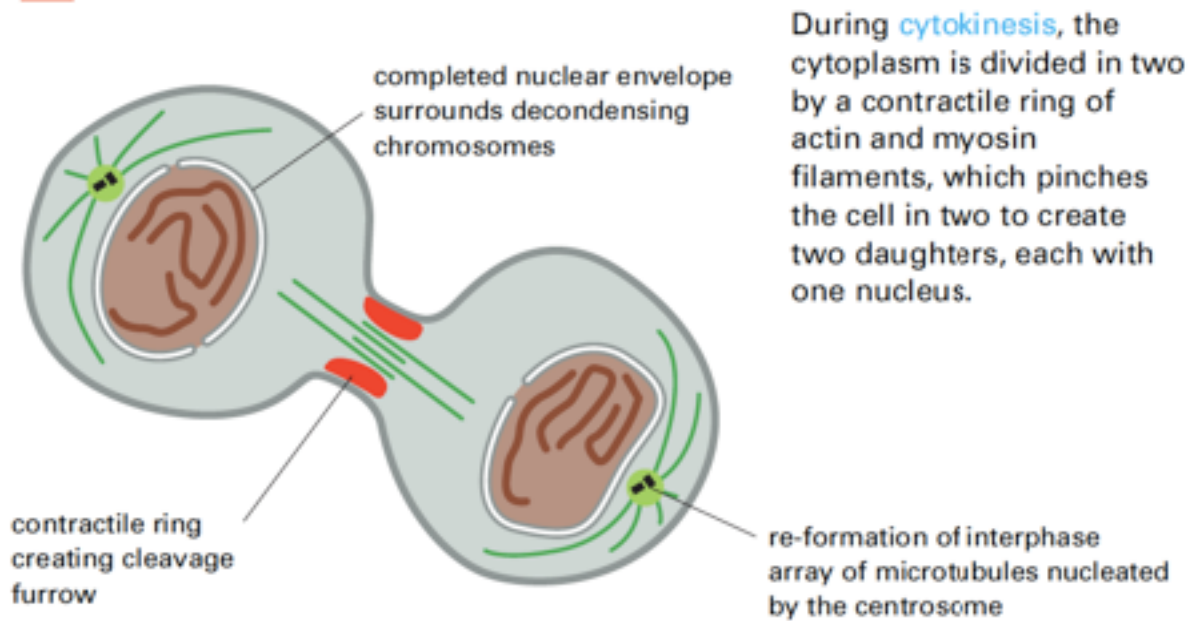
5 TELOPHASE



During **telophase**, the two sets of daughter chromosomes arrive at the poles of the spindle and decondense. A new nuclear envelope reassembles around each set, completing the formation of two nuclei and marking the end of mitosis. The division of the cytoplasm begins with contraction of the contractile ring.



6 CYTOKINESIS



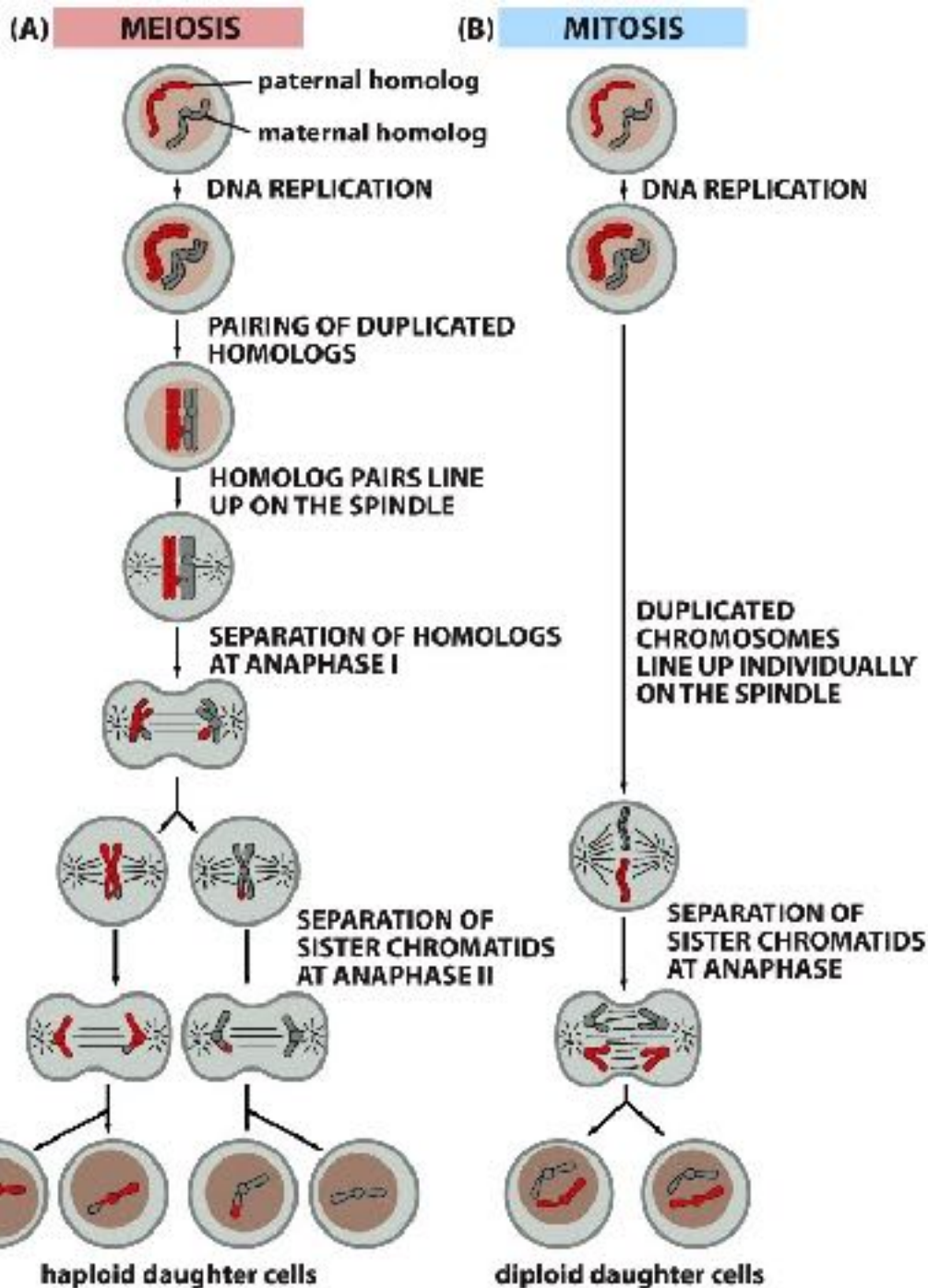
(Micrographs courtesy of Julie Canman and Ted Salmon.)



Podział redukcyjny - mejoza

Zapłodniona komórka jajowa dzieli się mitotycznie. Aby liczba chromosomów nie ulegała podwojeniu, komórki rozrodcze muszą mieć o połowę mniej chromosomów niż komórki wegetatywne.





autosomes,
sex chromosomes,
homologs,
homologous partner,
genetic recombination



replicated
paternal
chromosome 1

replicated
maternal
chromosome 1

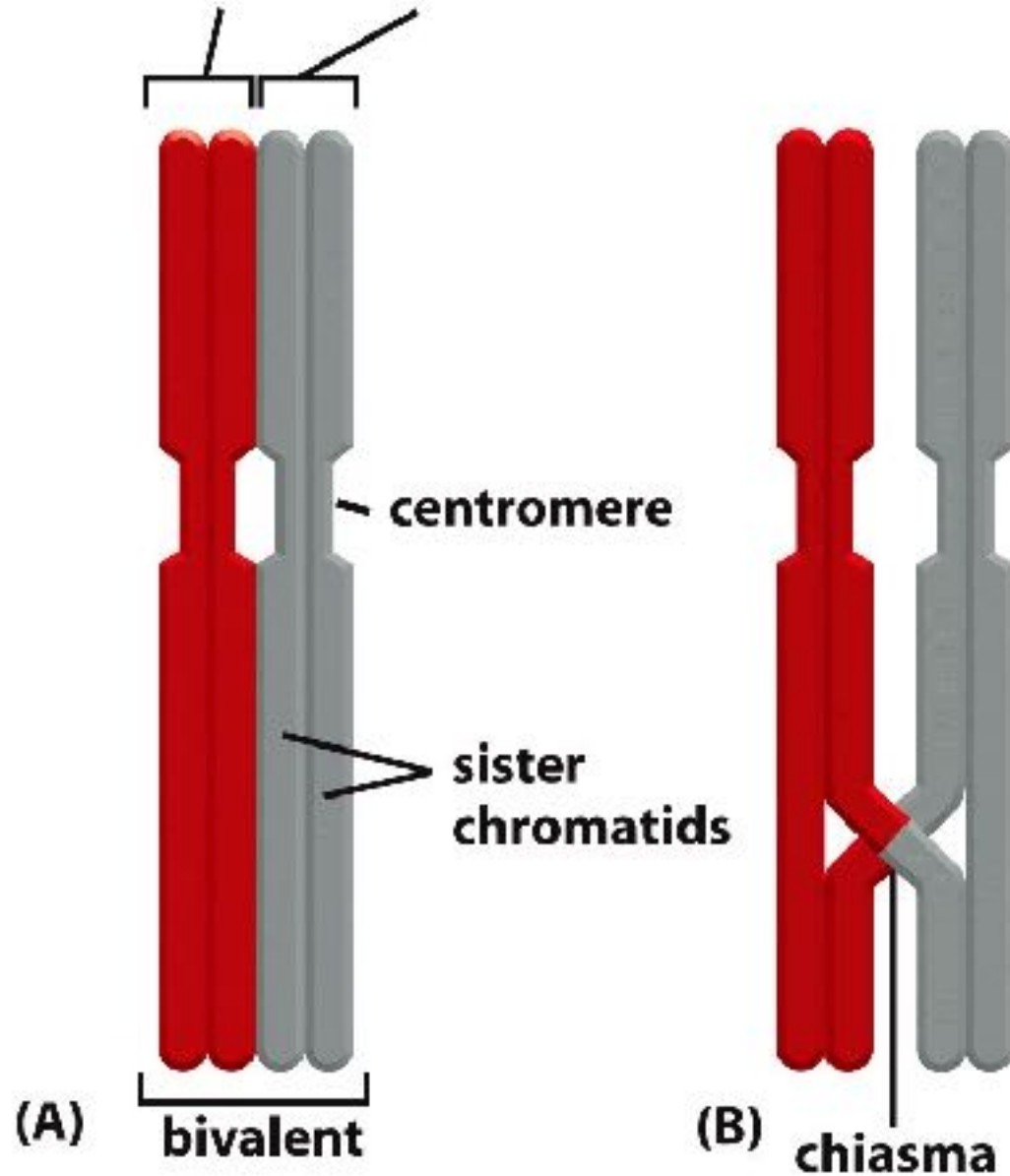


Figure 21-6 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Obserwacja postępu replikacji

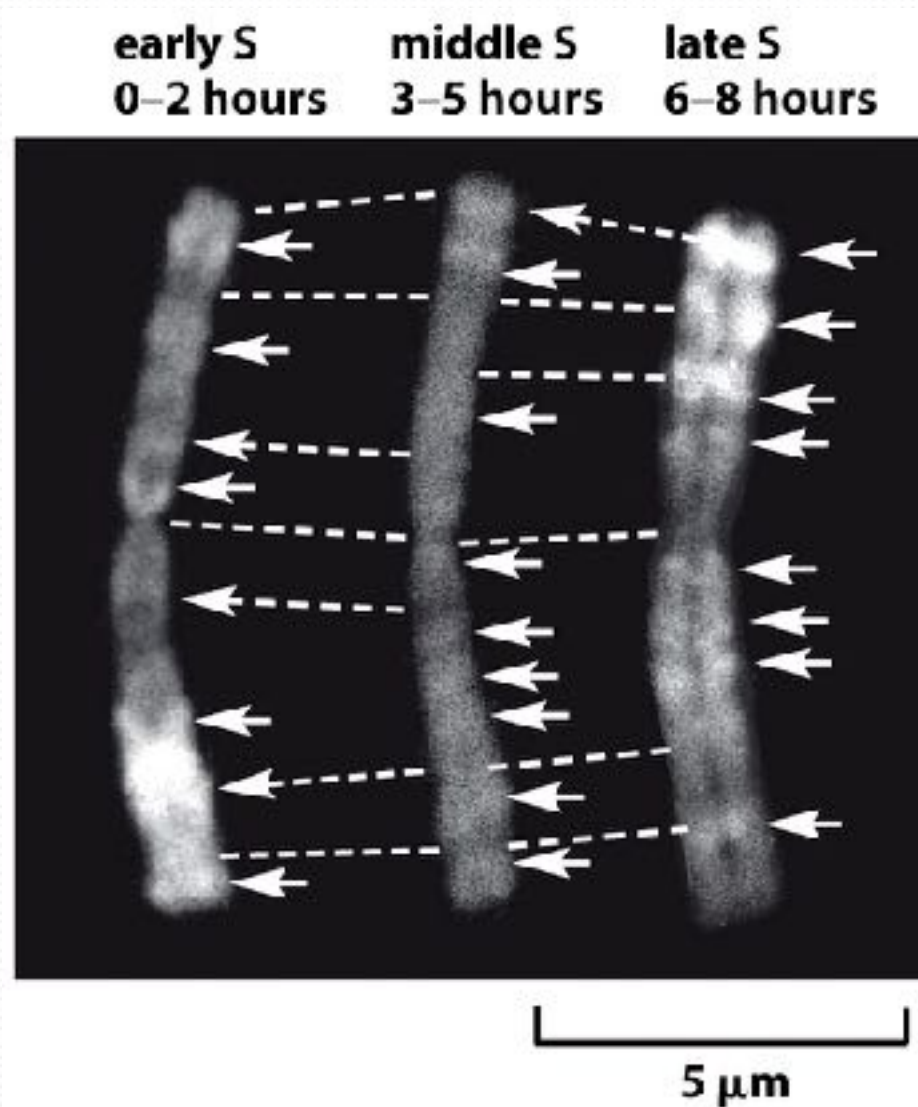


Figure 5-31 Molecular Biology of the Cell 5/e (© Garland Science 2008)



Obserwacja postępu replikacji

culture of yeast cells
arrested before DNA
replication begins

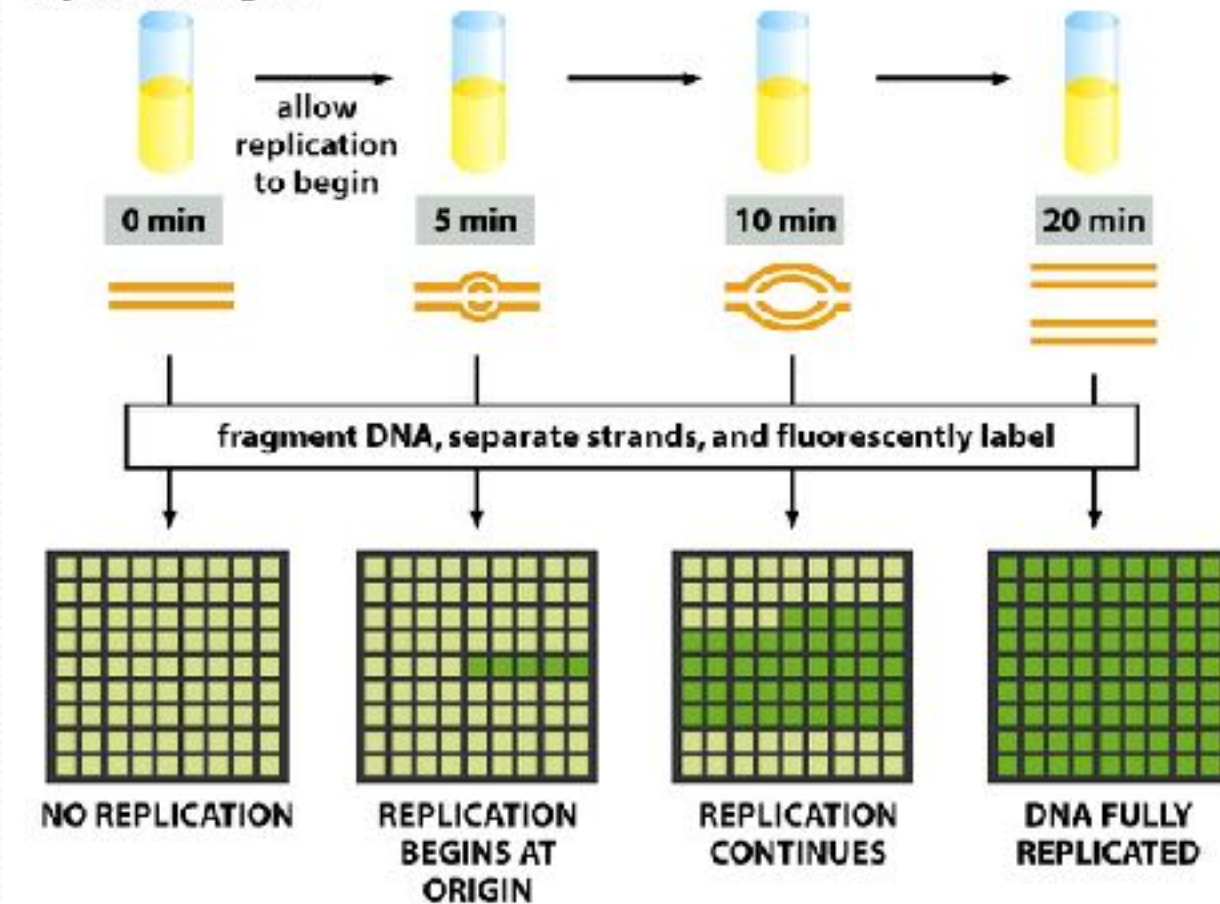


Figure 5-32 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Mutacje

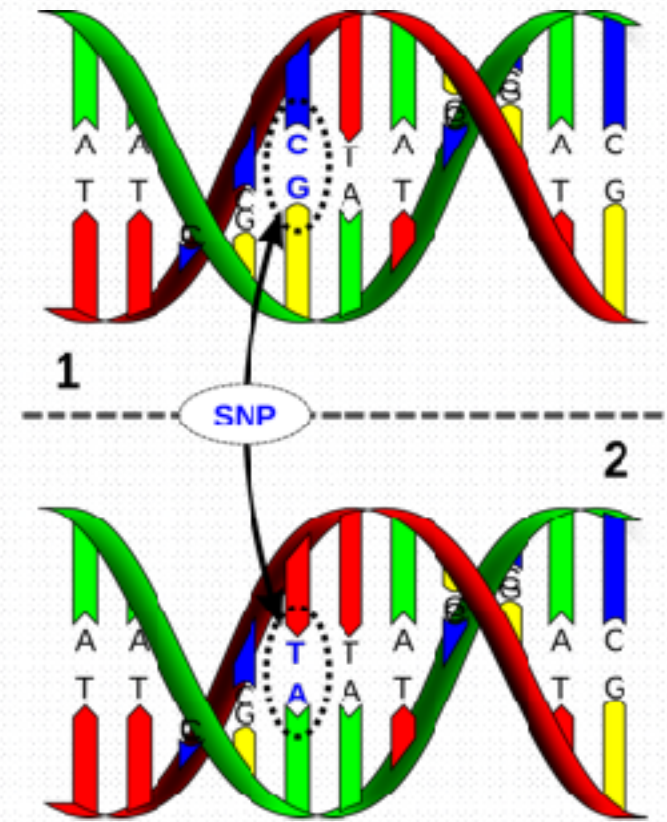
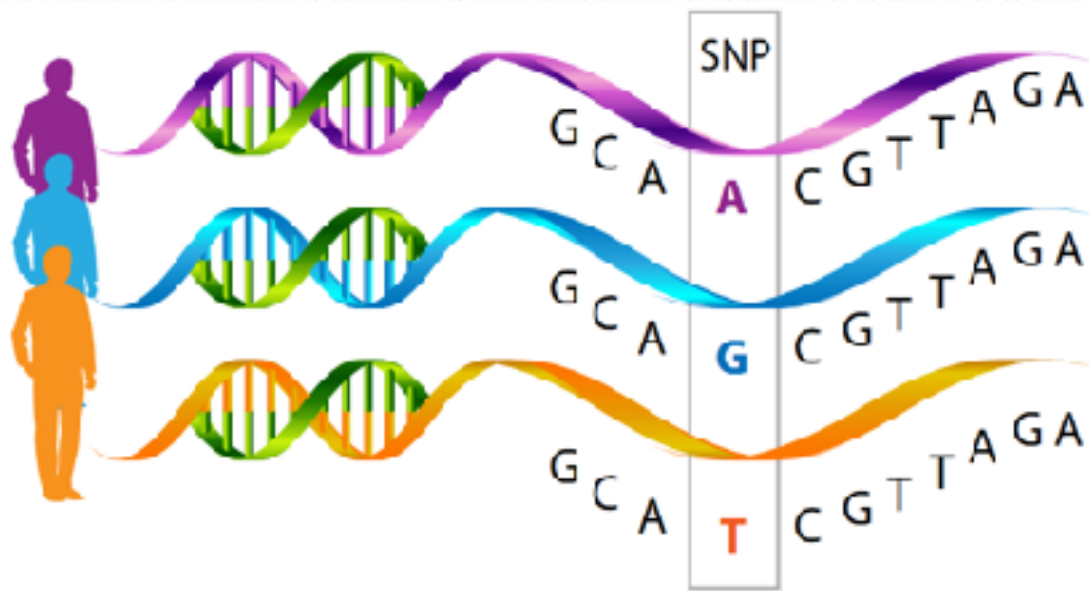
Losowe, trwałe zmiany w informacji genetycznej

Odnosząc się do organizacji przestrzennej, jak można poklasyfikować mutacje?

- Odnoszące się najczęściej do jednego genu - mutacje punktowe; para nukleotydów zmieniana jest na inną
- zmiany dotyczące fragmentów chromosomów - aberracje chromosomowe
- zmiany liczby chromosomów



Mutacje punktowe

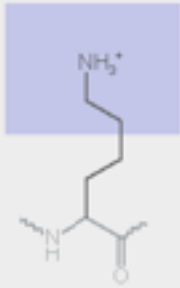
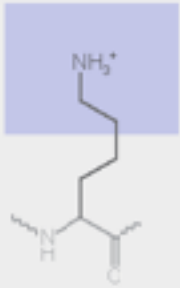
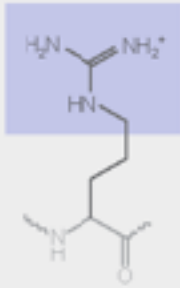
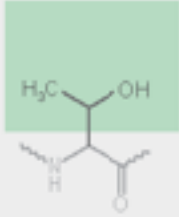


Co może spowodować mutacje? Ze względu na przyczyny jak podzielić mutacje?

- Substytucje,
- Insercje,
- Delecje.

Mutacje punktowe

Podział ze względu na efekt

	Mutacje punktowe				
	bez mutacji	cicha	nonsensowna	zmiany sensu	
				konserwatywna	niekonserwatywna
poziom DNA	TTC	TTT	ATC	TCC	TGC
poziom mRNA	AAG	AAA	UAG	AGG	ACG
poziom białek	Lys	Lys	STOP	Arg	Thr
					
				basic	polar

Mutacje letalne

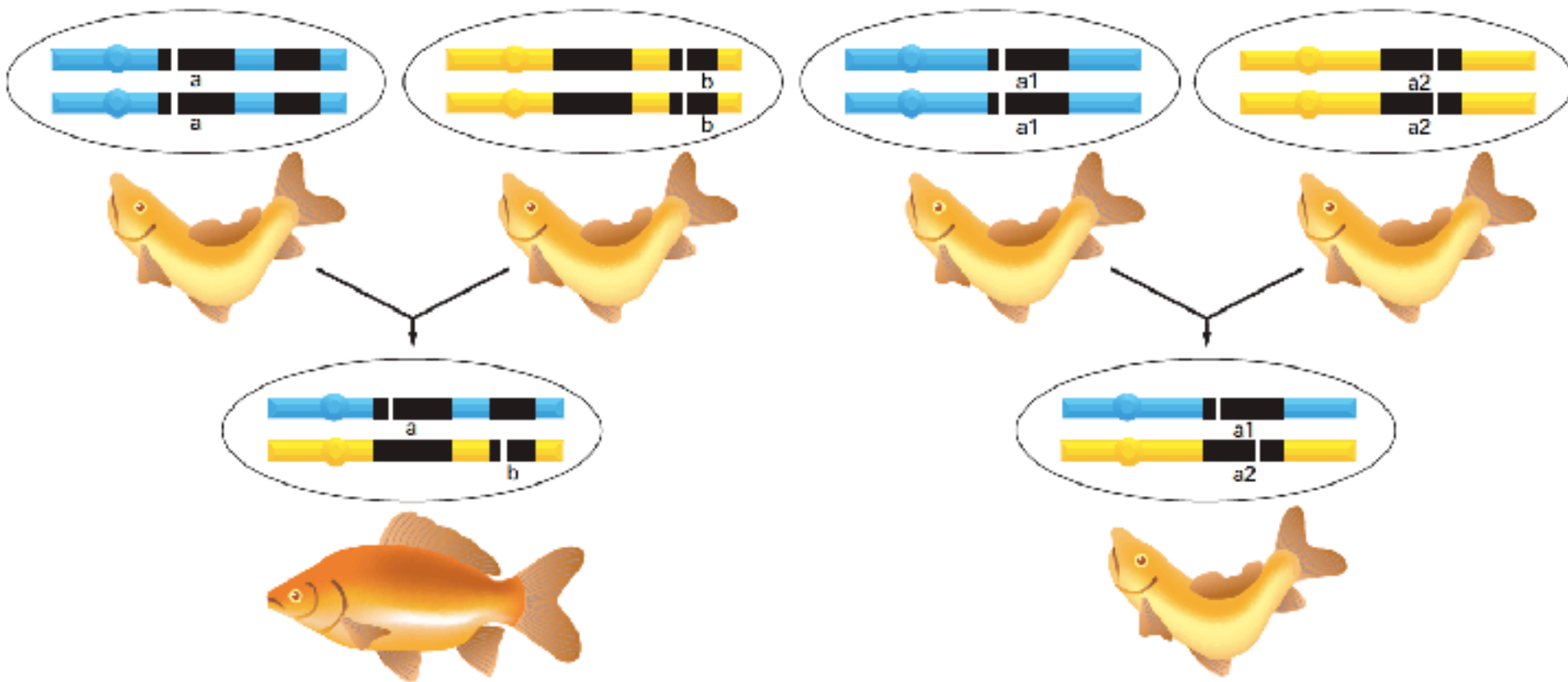
Konsekwencje mutacji

homozygous mutant mother

homozygous mutant father

homozygous mutant mother

homozygous mutant father

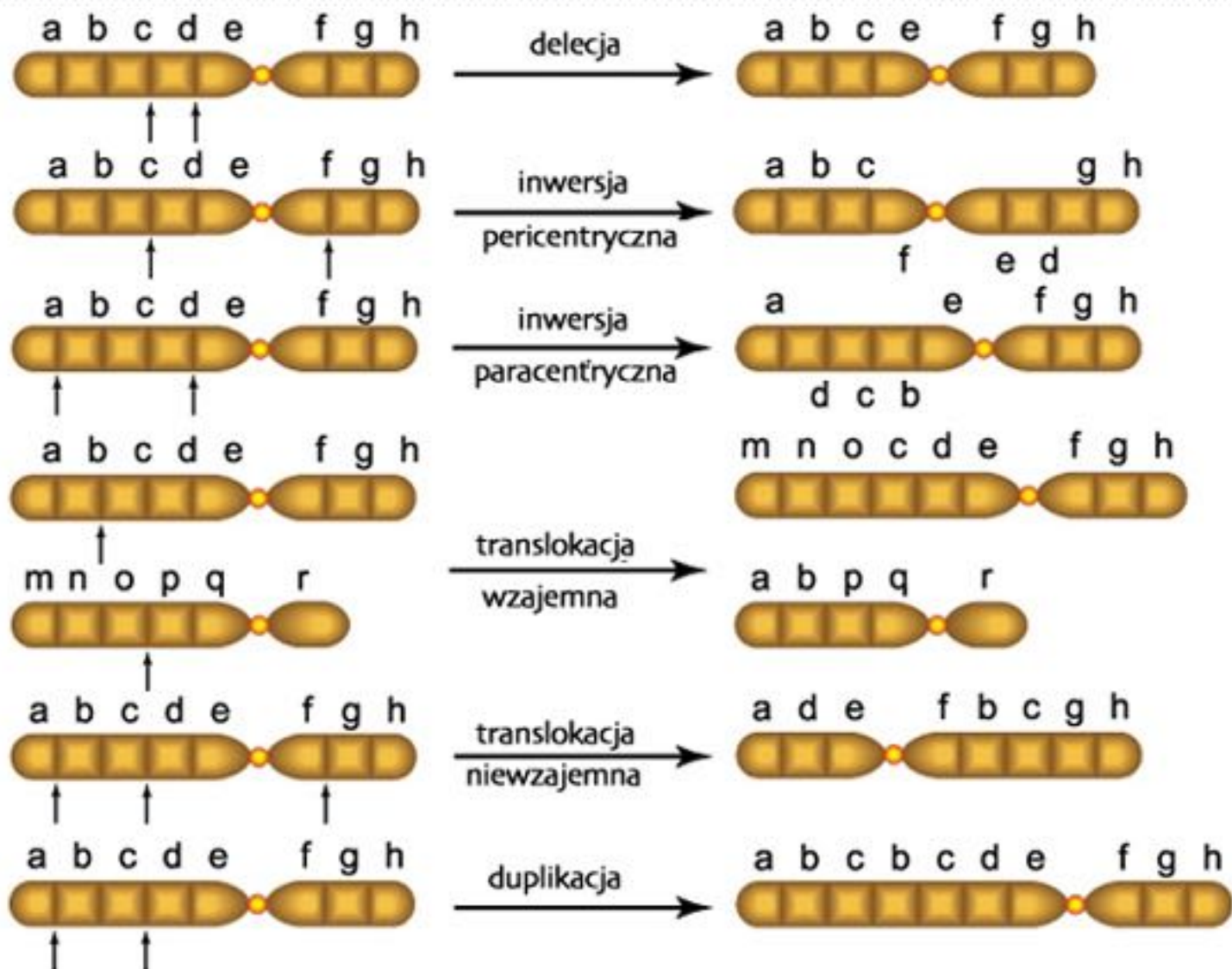


hybrid offspring shows normal phenotype:
one normal copy of each gene is present

hybrid offspring shows mutant phenotype:
no normal copies of the mutated gene are present



Aberracje chromosomowe



Zmiana liczby chromosomów

liczba chromosomów

haploidalny organizm	n
diploidalny organizm	$2n$ (euploidalna)
zmiana liczby chromosomów	aneuploidalna
jedna kopia danego chromosomu	$2n-1$ (monosomia)
diploidalny organizm	$2n + 1$ trisomia
poliploidalny organizm	$3n, 4n, 5n...$



Przykłady mutacji

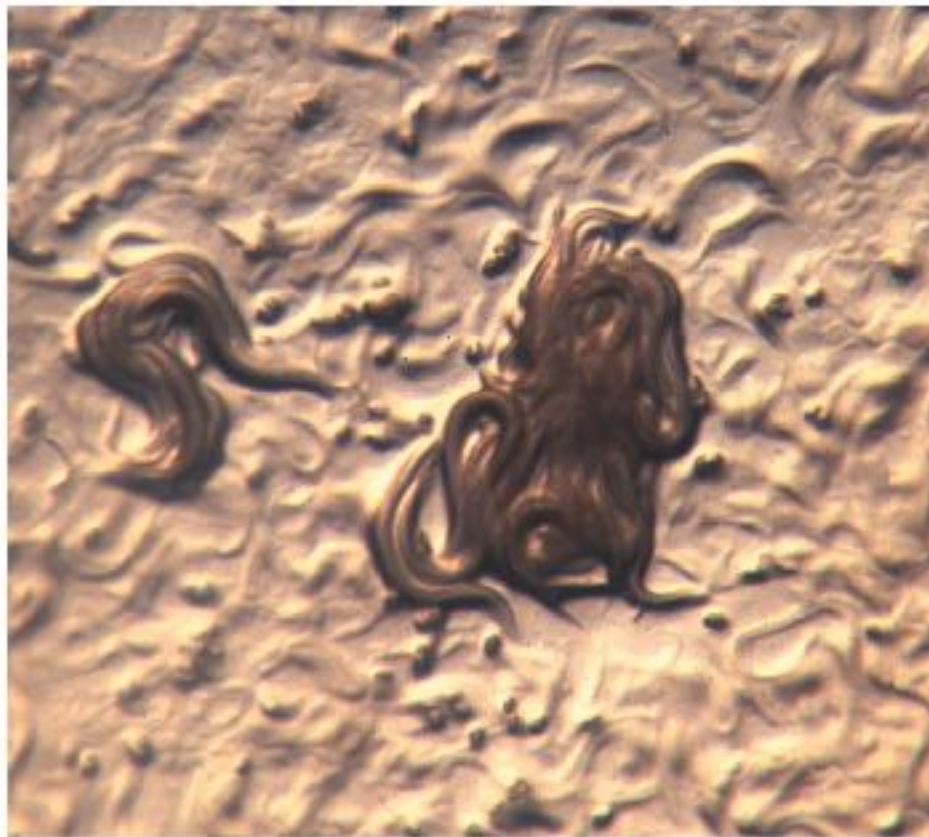


Figure 8-53 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Wyżlin



Przykłady mutacji



1 mm

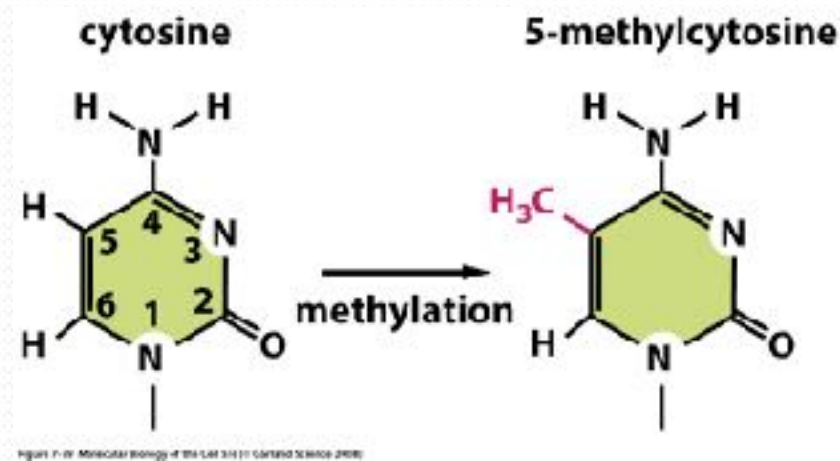
Figure 8-54 Molecular Biology of the Cell 5/e (© Garland Science 2008)

C. elegans



Zmiany epigenetyczne: metylacja DNA

- Metylacja DNA u kręgowców ogranicza się do cytozyny w sekwencji CG



Metylacja DNA

- Ewolucja - deaminacja zmetylowanej cytozyny prowadzi do powstania tyminy. Wiele obszarów CG przestało istnieć na skutek tego przekształcenia

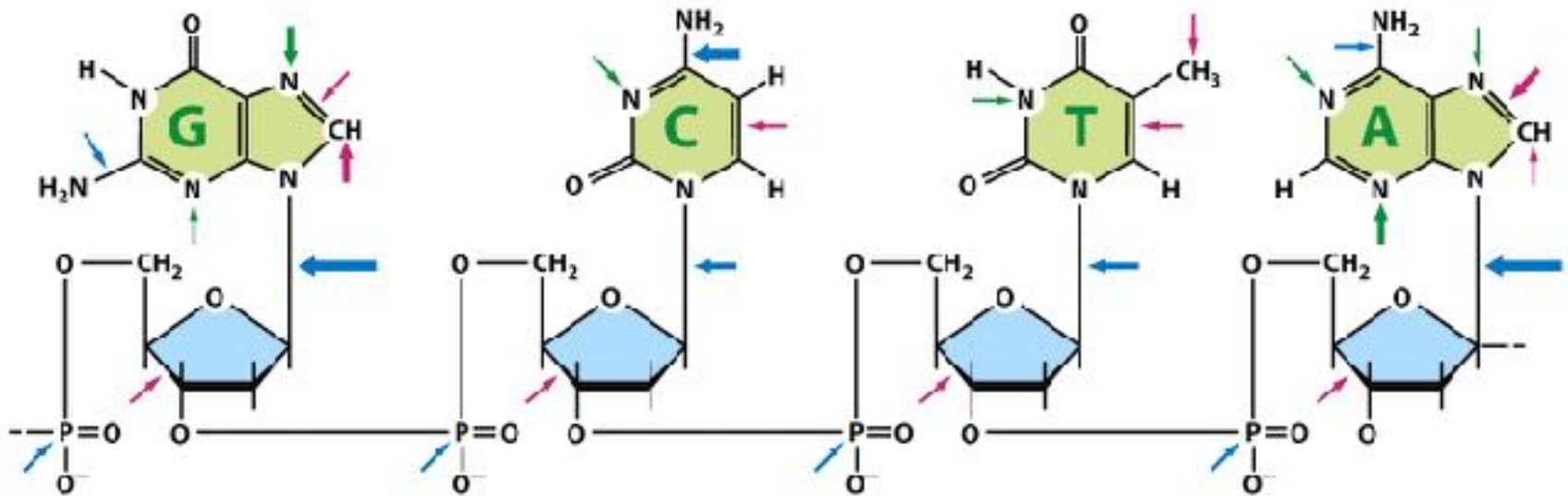


Figure 5-44 Molecular Biology of the Cell 5/e (© Garland Science 2008)

- Regiony wysokiej koncentracji CG nazywane są wyspami CG/CpG

Metylacja DNA

- Metylacja jest przekazywana nici potomnej

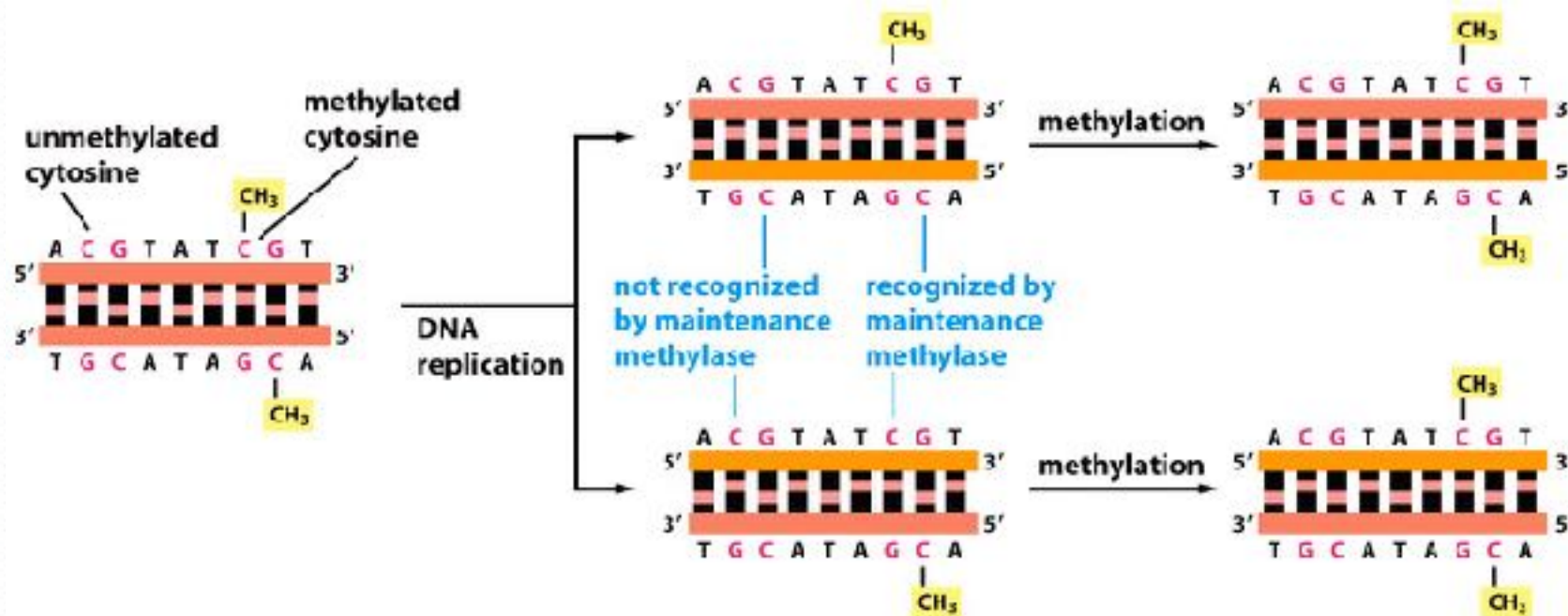
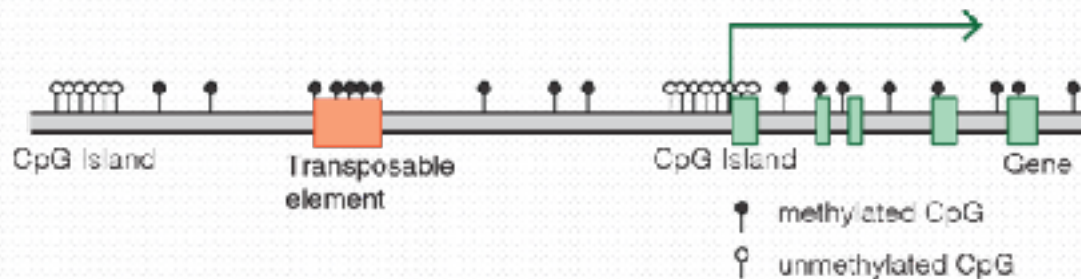


Figure 7-80 Molecular Biology of the Cell 5/e (© Garland Science 2008)

W momentach gwałtownego rozwoju organizmu obserwuje się demetylację wielu regionów DNA. Jaka może być tego przyczyna?

DNA methylation targets

Typical mammalian DNA methylation landscape

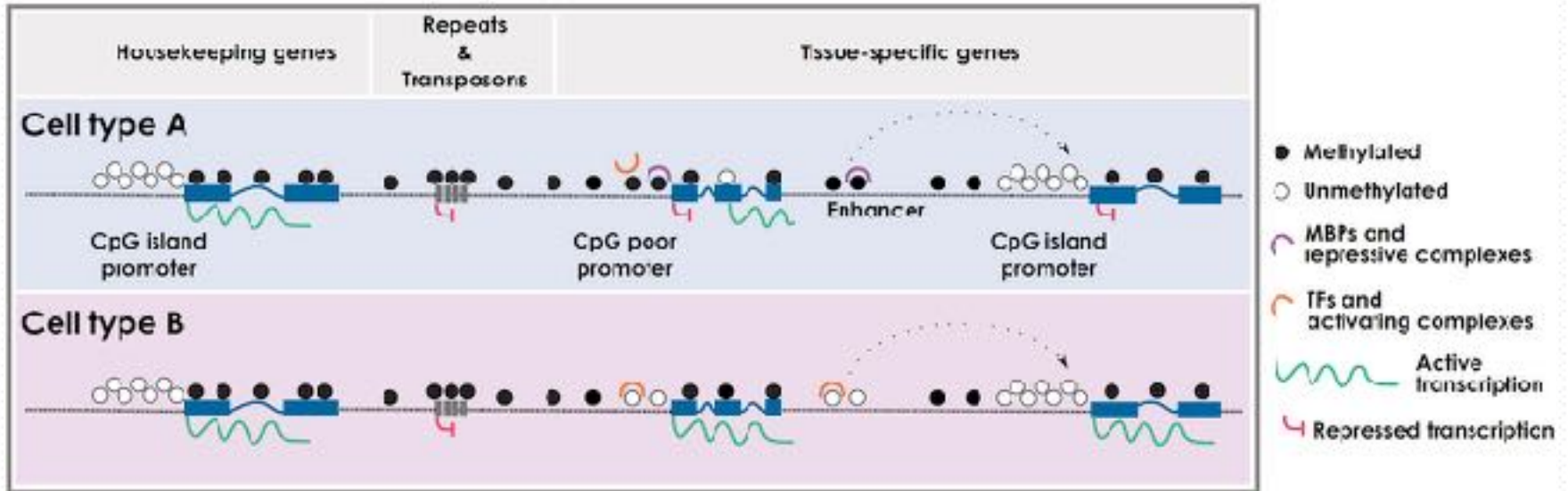


CpG islands:

- Length greater than 200bp
- G+C content above 50%
- ~ 25,000 of CpG islands
- ~75% being less than 850bp long
- ~50% being located in promoter regions
- ~25% being located within gene bodies
- 60% to 70% of genes have an island in the promoter
- Unmethylated and enriched in chromatin modification H3K4

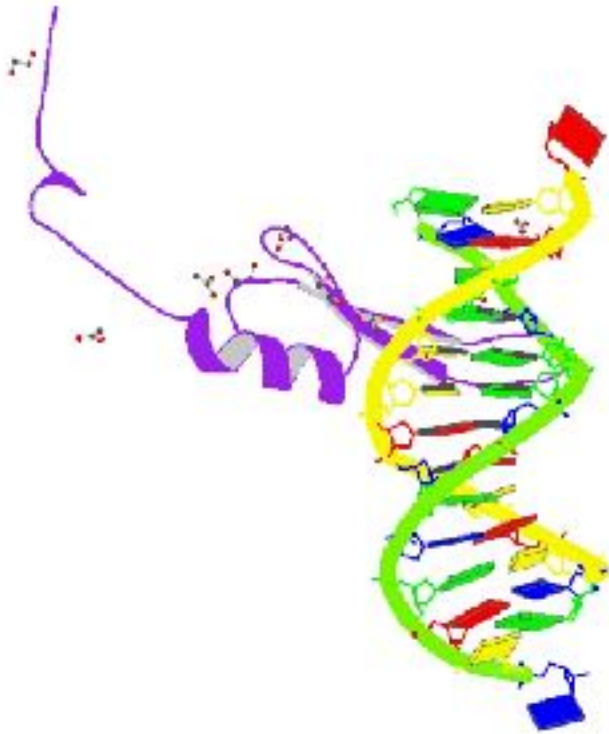


Repression of CpG-dense promoters



Methyl-CpG-binding domain

- Effects of DNA methylation are mediated
- Specific domain of ~70 residues
- Linked to other domains that recruit:
 - histone deacetylase complexes
 - chromatin remodeling factors



Crystal structure of methyl CpG Binding Domain of MBD4 in complex with the 5mCG/TG sequence

Journal List | Nucleic Acids Res. | v.21(21); 1993 Oct 25 | PMID: 11491

Nucleic Acids Research

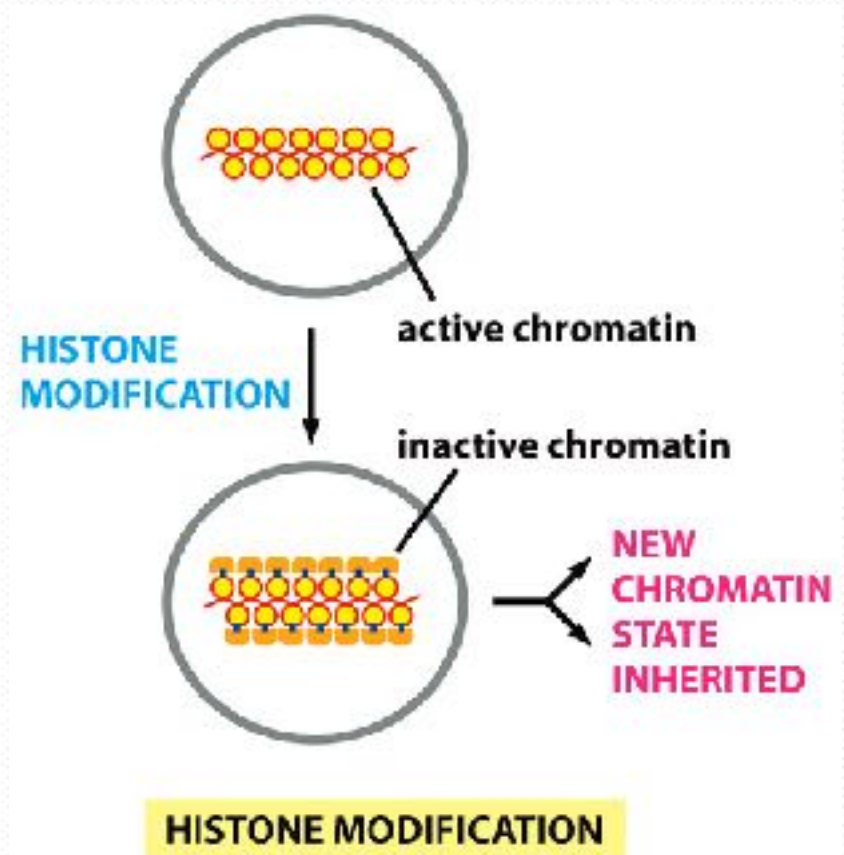
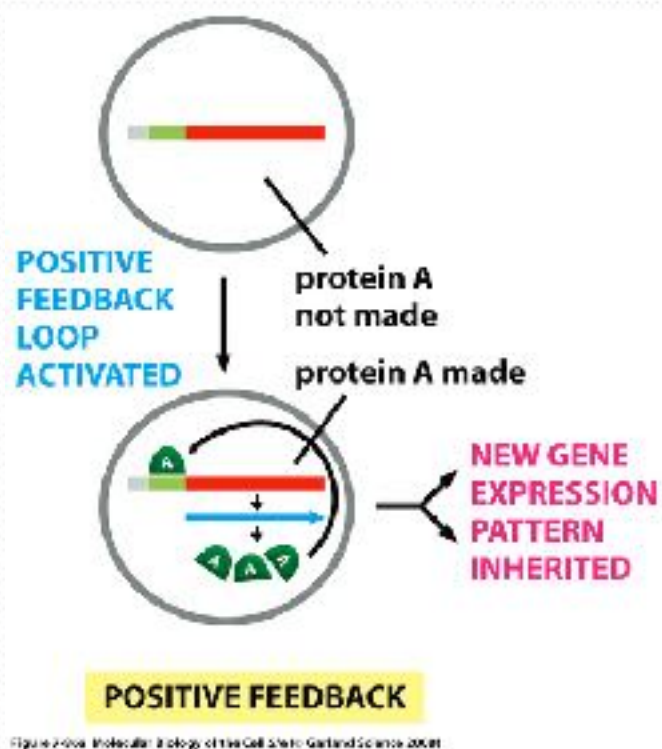
[Nucleic Acids Res.](#) 1993 Oct 25; 21(21): 4006-4012.

PMID: 11491

Dissection of the methyl-CpG binding domain from the chromosomal protein MeCP2.

[X. Nan](#), [R. R. Mowbray](#), and [A. Bird](#)

Dziedziczenie epigenetyczne



Dziedziczenie epigenetyczne

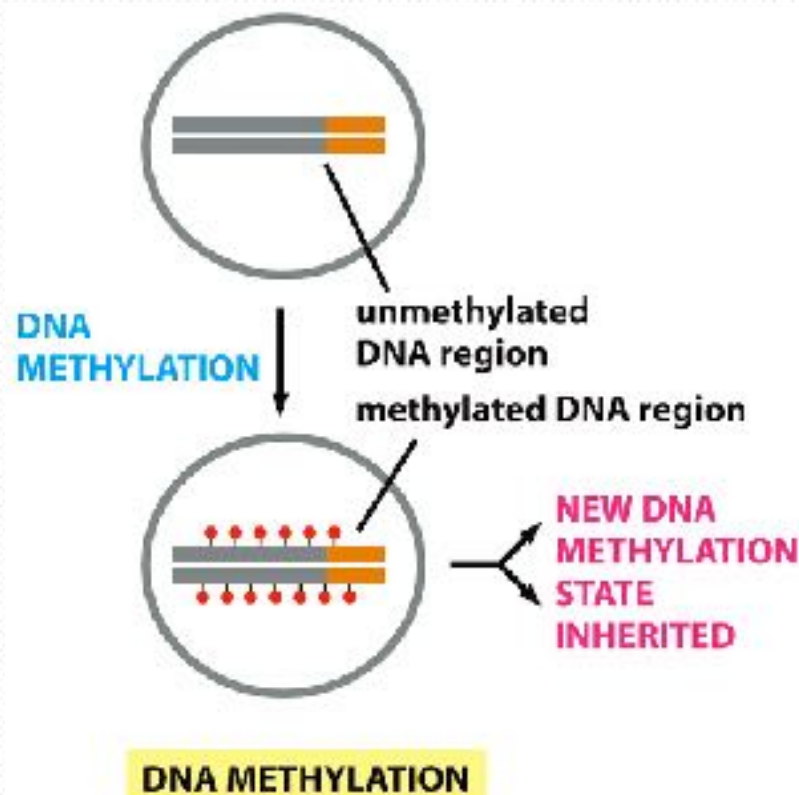


Figure 7-86a Molecular Biology of the Cell 5/e © Garland Science 2008

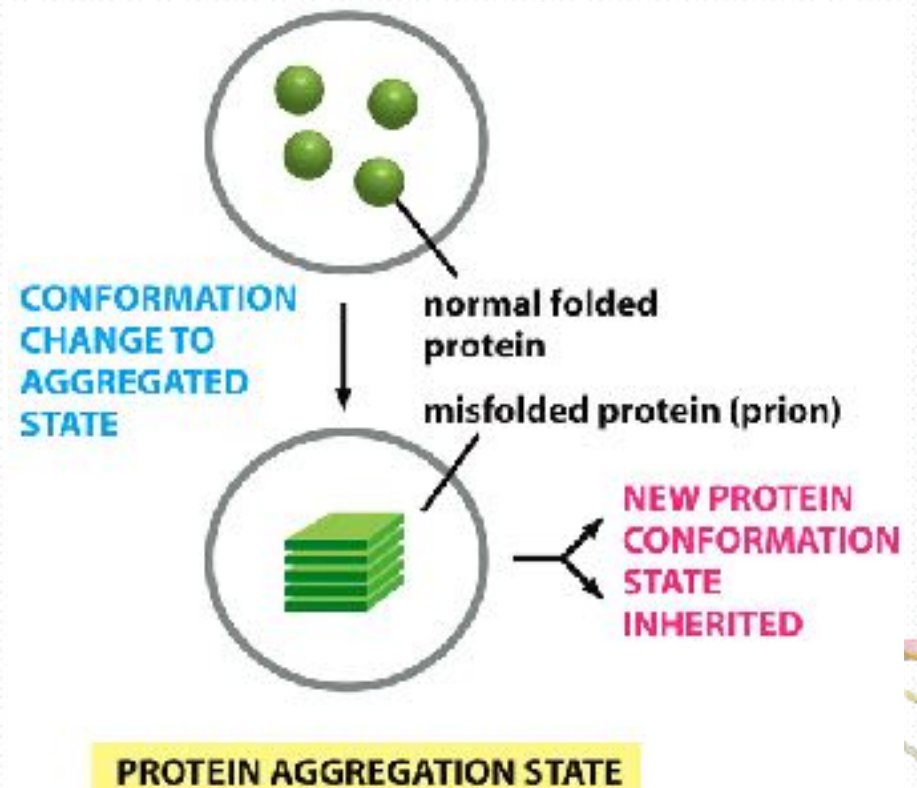


Figure 7-86a Molecular Biology of the Cell 5/e © Garland Science 2008

Nie tylko DNA ale i chromosomy zawierają wiele informacji, które są przekazywane komórkom potomnym i mają kluczowe znaczenie w regulacji ekspresji genów.

Powstawanie nowych genów

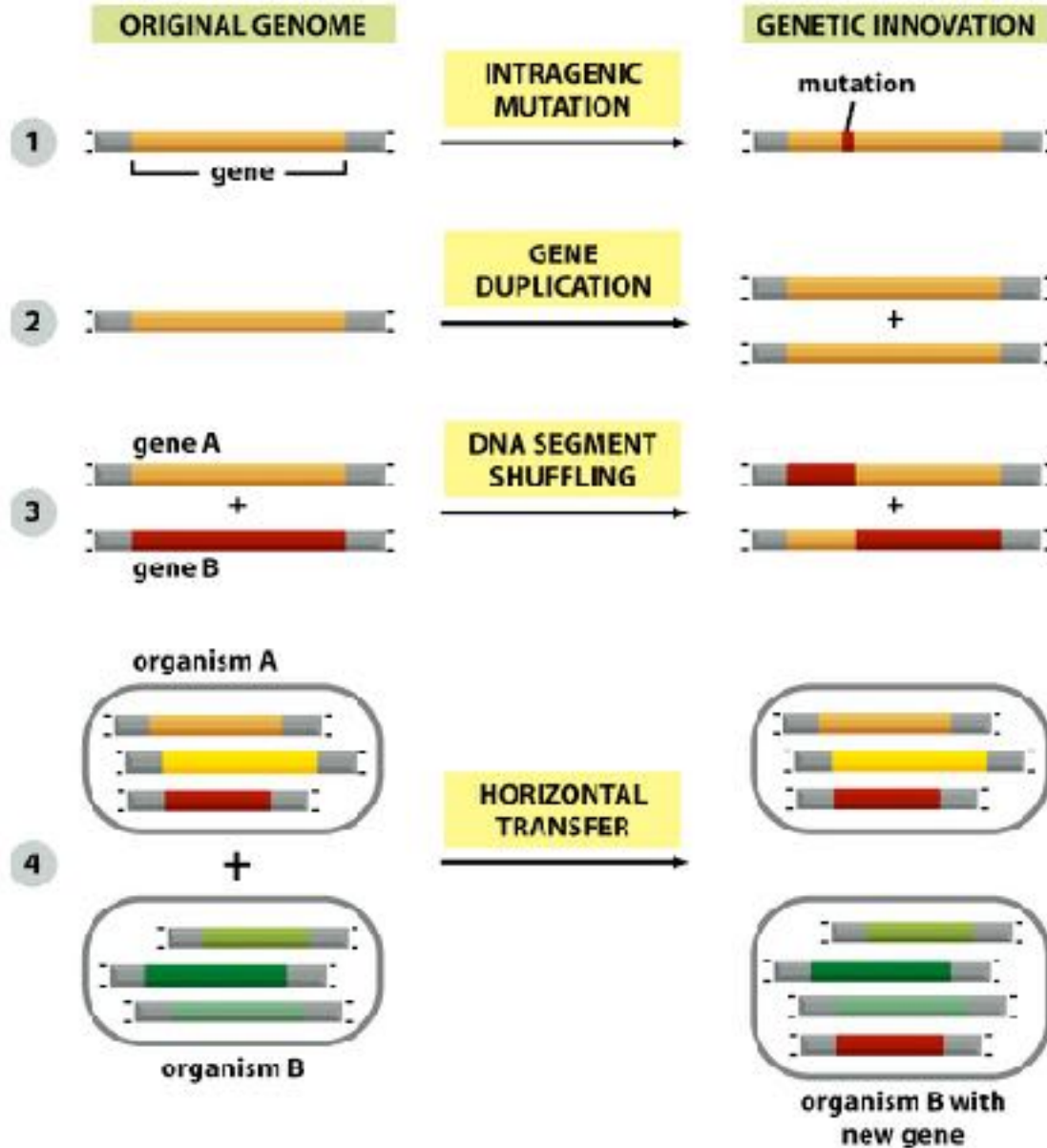


Figure 1-23 Molecular Biology of the Cell 5/e (© Garland Science 2008)



How genomes evolve

- view of the process of evolution
- information about family relationships among organisms
- molecular mechanisms by which evolution has proceeded

Homologous genes—that is, genes that are **similar** in both their nucleotide **sequence** and **function** because of a common ancestry—can often be recognised across vast phylogenetic distances.

- Nematode worms, fruit flies, yeasts, and even bacteria
- The protein-coding portion of a yeast gene can be substituted with its human homolog—even though humans and yeast are separated by more than a billion years of evolutionary history



Humans and mice

- have a common ancestor 80×10^6 years ago
- most of nucleotides could be changed by mutations
- functional sequences are much more conserved
- conserved regions vs. non conserved regions (less likely to be critical for function)

Roughly **5%** of the human genome consists of “**multi-species conserved sequences**”



Geny paralogiczne i ortologiczne

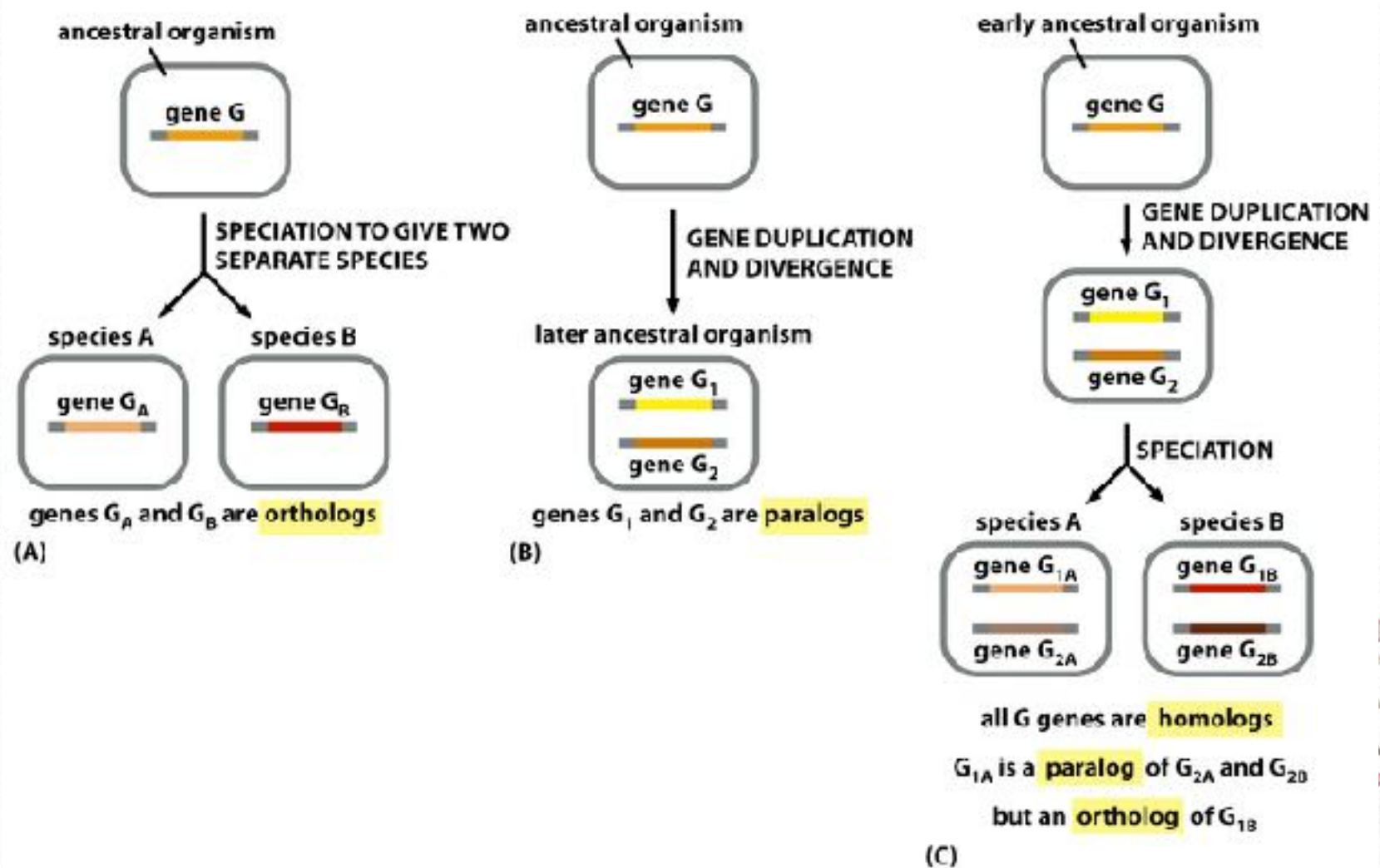


Figure 1-25 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Rodzina genów homologicznych

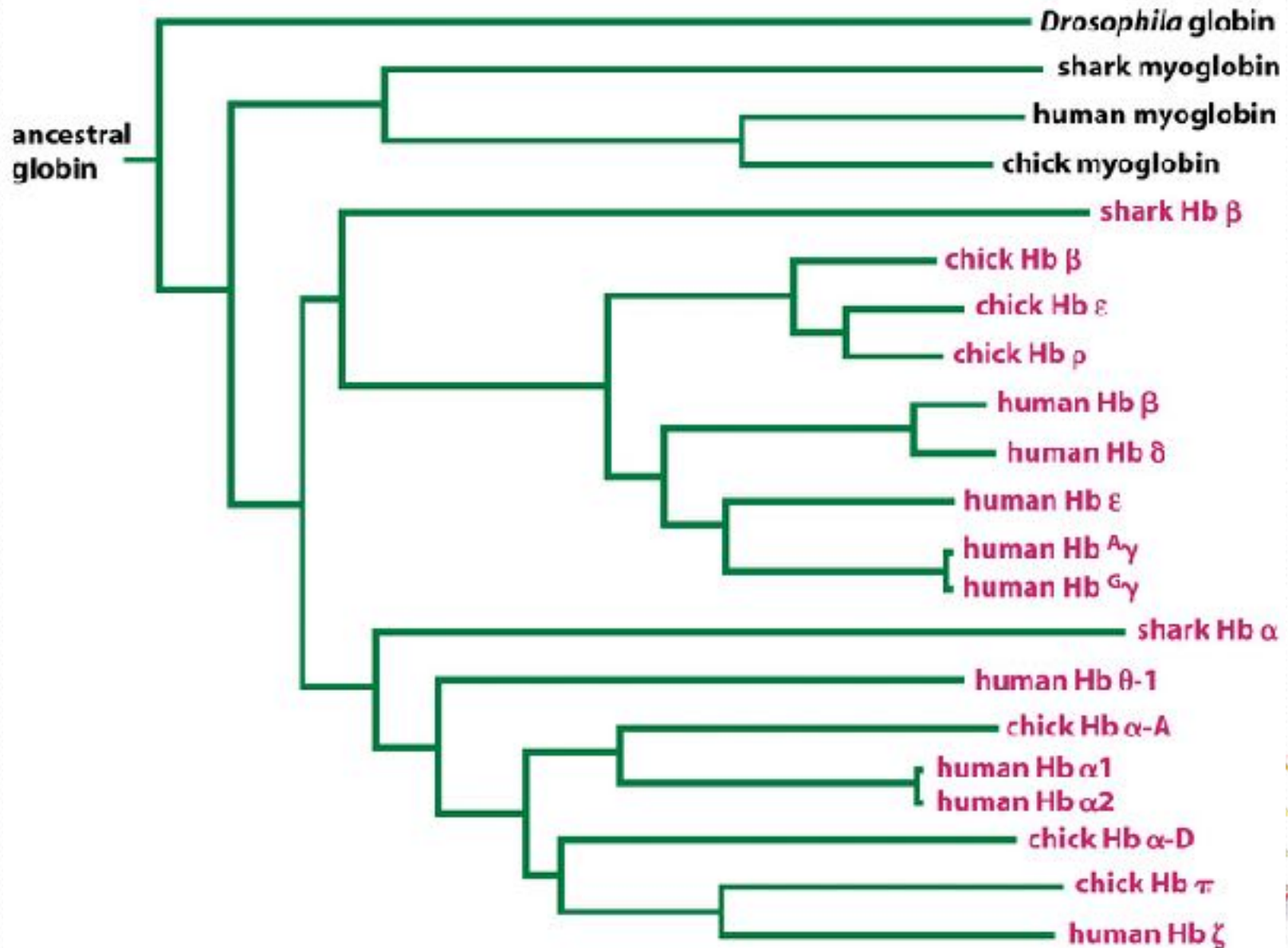


Figure 1-26 Molecular Biology of the Cell 5/e (© Garland Science 2008)

Wielkość genomów

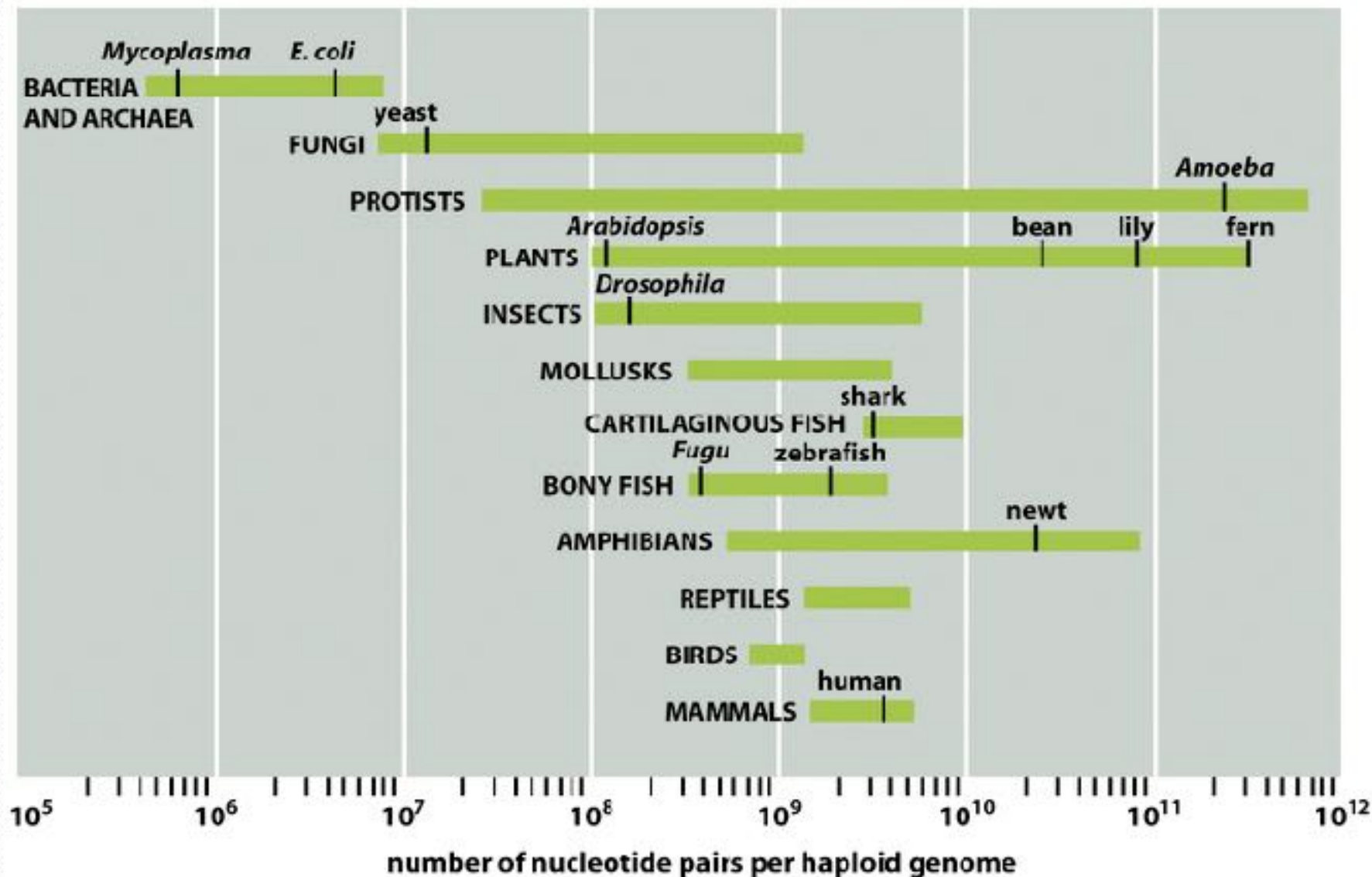


Figure 1-37 Molecular Biology of the Cell 5/e (© Garland Science 2008)

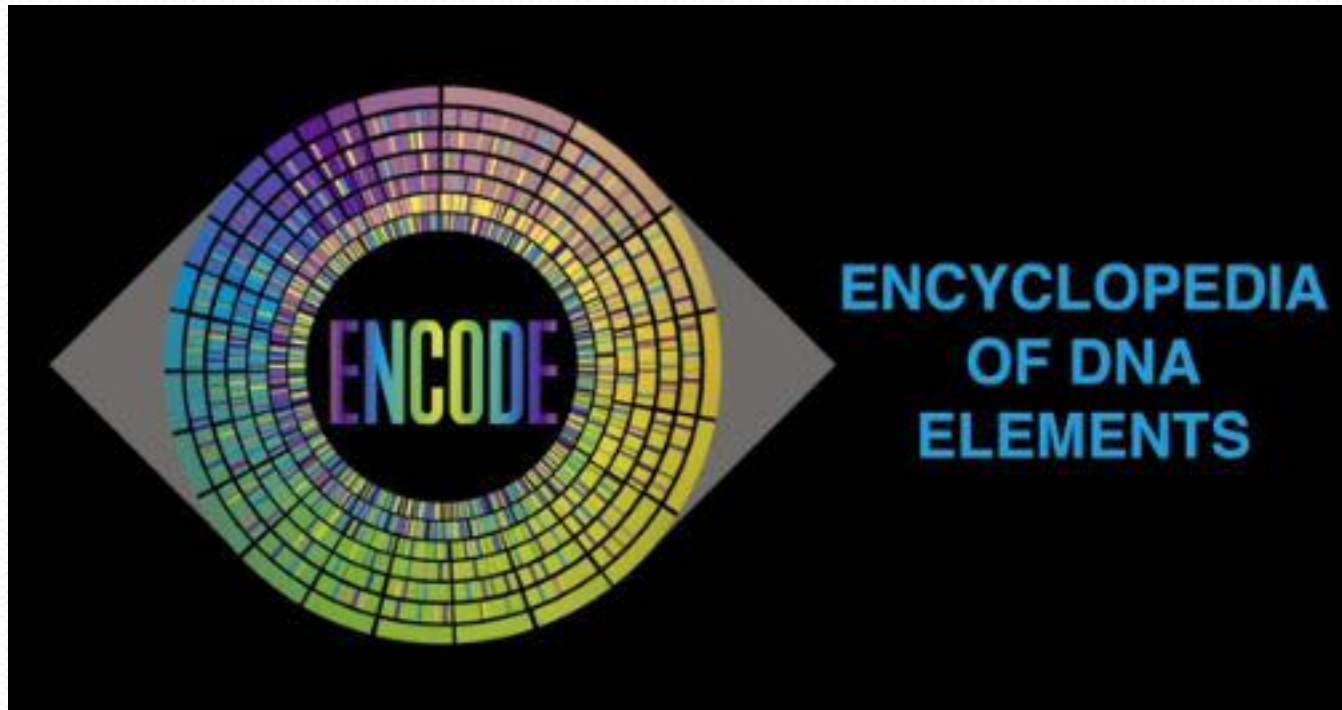
Genome alterations - failures of the normal mechanisms

- copying DNA
- maintaining DNA
- transposable elements

Evolution depends on accidents and mistakes followed by nonrandom survival.



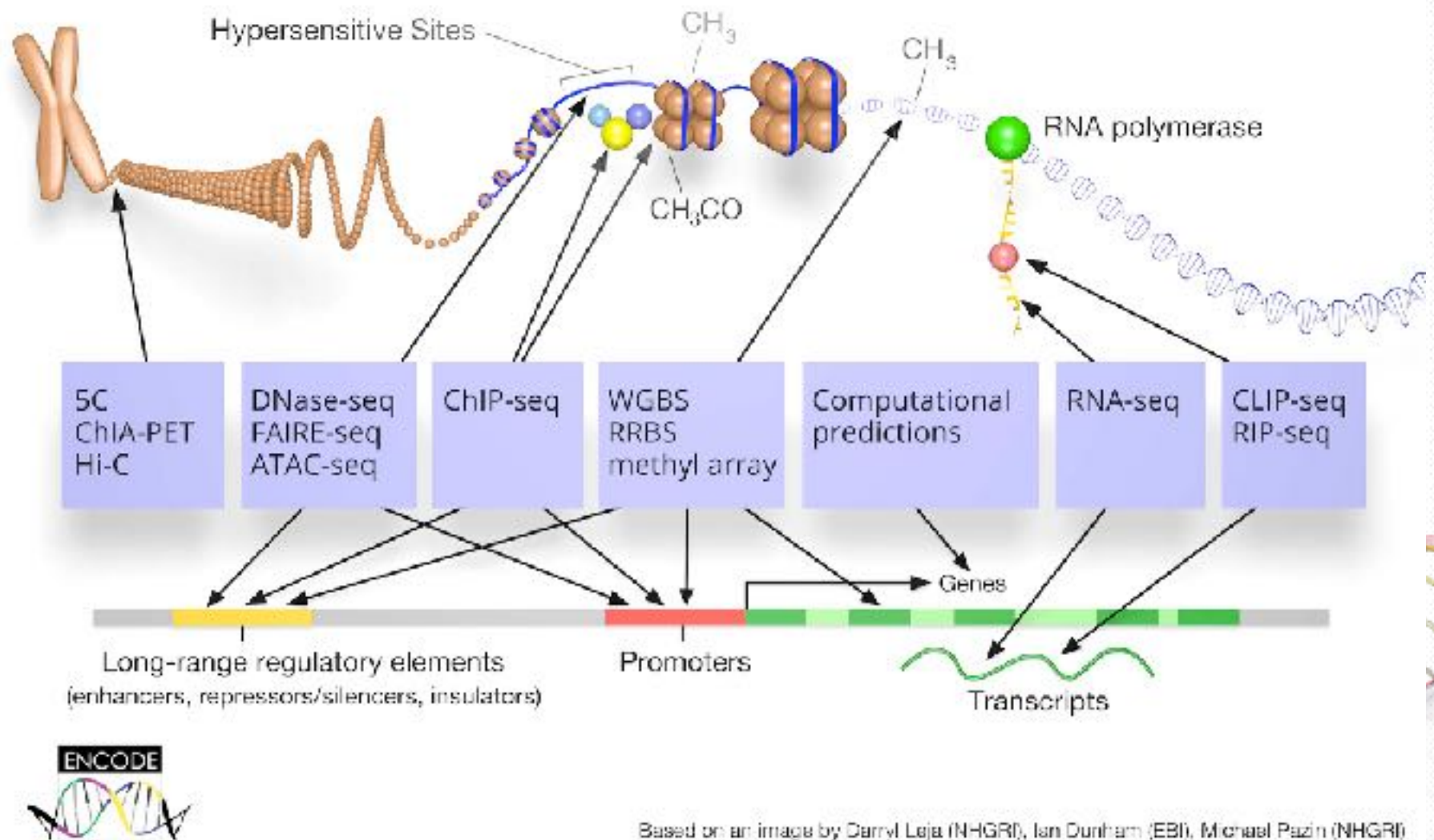
Encyclopedia of DNA Elements



Konsorcjum międzynarodowe, którego celem jest opracowanie obszernej listy elementów funkcjonalnych w ludzkim genomie, w tym elementów działających na poziomie białka, RNA i elementów regulujących komórki i aktywność genów



Encyclopedia of DNA Elements



ENCODE

<https://www.encodeproject.org>

ENCODE: Encyclopedia of DNA Elements

The ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

[Get Started](#)

[View Assay Matrix](#)

Project Biosample Type Assay Categories

ENCODE

<https://genome.ucsc.edu>



ENCODE at UCSC

UCSC Genome Browser

Genomes Genome Browser Tools Mirrors Downloads My Data Help About Us

Our tools

- **Genome Browser**
interactively visualize genomic data
- **BLAT**
rapidly align sequences to the genome
- **Table Browser**
download data from the Genome Browser database
- **Variant Annotation Integrator**
get functional effect predictions for variant calls
- **Data Integrator**
combine data sources from the Genome browser database
- **Gene Sorter**
find genes that are similar by expression and other metrics
- **Genome Browser in a Box (GBIB)**
run the Genome Browser on your laptop or server
- **In-Silico PCR**
rapidly align PCR primer pairs to the genome
- **LiftOver**
convert genome coordinates between assemblies
- **VistCseq**
interactively view in situ images of mouse and frog

More tools...

Our story

On June 22, 2000, UCSC and the other members of the International Human Genome Project consortium completed the first working draft of the human genome assembly, forever ensuring free public access to the genome and the information it contains. A few weeks later, on July 7, 2000, the newly assembled genome was released on the web at <http://genome.ucsc.edu>, along with the initial prototype of a graphical viewing tool, the UCSC Genome Browser. In the ensuing years, the website has grown to include a broad

What's new

Mar. 20, 2017 - 20 Species Conservation Track now available on RGSC 5.0/fin5

Mar. 15, 2017 - New Genome Browsers available for golden snub-nosed monkey, proboscis monkey, and turkey!

Mar. 3, 2017 - New NCBI RefSeq tracks released!

ENCODE

<https://genome.ucsc.edu/encode/>

genome.ucsc.edu

ENCODE at UCSC

ENCOD - Scutaj v Google

UCSC Genome Browser Home

Encyclopedia of DNA Elements at UCSC 2003 - 2012

Human Data at UCSC

- Downloads
- Experiment Matrix
- Search
- Genome Browser (hg19)
- Experiment List
- Cell Types

Mouse Data at UCSC

- Downloads
- Experiment Matrix
- Search
- Genome Browser (mm9)
- Experiment List
- Cell Types

Metadata Terms

- Registered Variables
- Antibodies
- Other Resources
- News Archive

First Production (2007-2012)

Pilot (2003-2007)

Contacts

About

The [Encyclopedia of DNA Elements \(ENCODE\)](#) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI). The goal of ENCODE is to build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

ENCODE results from 2007 and later are available from the ENCODE Project Portal, [encodeproject.org](#). This covers data generated during the two production phases 2007-2012 and 2013-present. The ENCODE Project Portal also hosts additional ENCODE access tools, and ENCODE project pages including up-to-date information about data releases, publications, and upcoming tutorials.

UCSC coordinated data for the ENCODE Consortium from its inception in 2003 (Pilot phase) to the end of the first 5 year phase of whole-genome data production in 2012. All data produced by ENCODE investigators and the results of ENCODE analysis projects from this period are hosted in the UCSC Genome browser and database. Explore ENCODE data using the image links below or via the left menu bar. **All ENCODE data at UCSC are freely available for download and analysis.**

Explore ENCODE data (2003 - 2012) at UCSC

View ENCODE data (2003 - 2012) in the UCSC Genome Browser

Search for data (current) at the ENCODE Portal

Search for ENCODE tracks (2003 - 2012) in the UCSC Browser

ENCODE


<https://genome.ucsc.edu/encode/downloads.html>

genome.ucsc.edu

ENCODE - szukaj w Google

UCSC Genome Browser Home

ENCODE at UCSC (2003-2012) - ENCODE Portal (Data 2007-present) - Downloads

 **ENCODE Downloads** 2007 - 2012

Overview

This page contains links to directories containing raw and processed data released as part of the ENCODE production phase (September 2007 - December 2012). Formats are described on the [File Format FAQ](#). For bulk download, retrieval by [FTP](#) is recommended along with rsync. Using rsync, for example "rsync -a -P rsync://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeDir/wgEncodeFile_...", has the advantage of starting up where it left off after a failure, when ran again. Production data files are covered by the [ENCODE data release policy](#). Links in the blue sections below show descriptions of the data as tracks in the UCSC Genome Browser. For further helpful information, such as links to a matrix to access files and answers to common user questions, please see the [ENCODE FAQ and Resources page](#).

The ENCODE Analysis Working Group (AWG) is a cross-consortium effort to perform integrated analysis of all ENCODE data types based on uniform processing. In September 2012 the results of AWG analysis of data produced from September 2007 through January 2011 were published in a series of coordinated [publications](#). Resulting files from the uniform processing reported in these publications are hosted on the ENCODE Analysis Data Hub for [download](#) and [visualization in the UCSC Browser](#). Visit the ENCODE [Integrative Analysis](#) and [Quality Metrics](#) pages for more information.

Human Genome Build 37 (hg19) ENCODE Analysis Data at UCSC

File Search

Integrated Regulation from UCSC	
DNase Clusters	Digital DNase-seq Hypersensitivity Clusters from ENCODE
Tss Factor ChIP	Transcription Factor ChIP-seq from ENCODE
Other	
Broad ChromHMM	Chromatin State Segmentation by HMM from ENCODE/Broad
Genome Segmentations	Genome Segmentations from DNase, FAIRE, Histone and TFBS Signals
Uniform DNase-seq	DNase-seq Hypersensitivity Uniform Peaks from ENCODE/Analysis
Uniform TFBS	Transcription Factor ChIP-seq Uniform Peaks from ENCODE/Analysis

Human Genome Build 37 (hg19) ENCODE Analysis Hub at the European Bioinformatics Institute

Metadata	
files.txt	Metadata terms for all Analysis Hub files
cvma	Controlled vocabulary used in ENCODE metadata
Uniform Peaks for DNase, FAIRE, Histone, and TFBS	
DNase-seq Peaks	DNase-seq Peaks of Open Chromatin
ENCODE Peaks	ENCODE Peaks of Open Chromatin

ENCODE

<https://genome.ucsc.edu/encode/downloads.html>

ENCODE at UCSC (2003-2012) - ENCODE Portal (Data 2007-present) - Downloads

ENCODE Downloads 2007 - 2012

Overview

This page links to directories of data generated by ENCODE project as part of the ENCODE project (September 2007 - December 2012). Formats and descriptions of the data are available in the [File Format FAQ](#). For bulk download, retrieval by [FTP](#) is recommended along with rsync. Using rsync, for example "rsync -a -P rsync://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeDir/wgEncodeFile_1", has the advantage of starting up where it left off after a failure, when ran again. Production data files are covered by the [ENCODE data release policy](#). Links in the blue sections below show descriptions of the data as tracks in the UCSC Genome Browser. For further helpful information, such as links to a matrix to access files and answers to common user questions, please see the [ENCODE FAQ](#) and [Resources page](#).

The ENCODE Analysis Working Group (AWG) is a consortium that perform integrated analysis of ENCODE data. In September 2012 the results of AWG analysis of data produced from October 2007 through January 2012 were published in a series of [hard-coded publications](#). Resulting files from the analysis processing reported in these publications are hosted on the ENCODE Analysis Data Hub for download and visualization in the UCSC Browser. See the ENCODE [Integrative Analysis and Quality Metrics](#) pages for more information.

Human Genome Build 37 (hg19): ENCODE Analysis Data at UCSC

File Search

Integrated Regulation from UCSC	
DNase Clusters	Digital DNase-seq Hypersensitive Clusters from ENCODE
Tn Factor ChIP	Transcription Factor ChIP-seq Uniform Peaks from ENCODE
Other	
Broad ChromHMM	Chromatin State Segmentation by HMM from ENCODE
Genome Segmentations	Genome Segmentations from DNase, FAIRE, Histone and TFBS Signals
Uniform DNase-seq	DNase-seq Uniform Peaks from ENCODE/Analysis
Uniform TFBS	Transcription Factor ChIP-seq Uniform Peaks from ENCODE/Analysis

Human Genome Build 37 (hg19): ENCODE Analysis Hub at the European Bioinformatics Institute

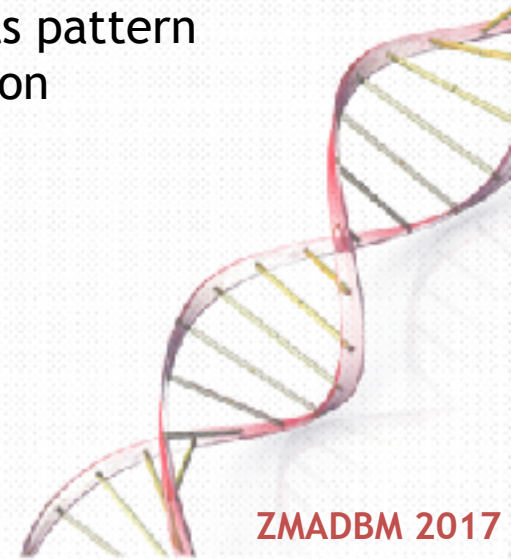
Metadata	
files.txt	Metadata terms for all Analysis Hub files
vocabulary	Controlled vocabulary used in ENCODE metadata
Uniform Peaks for DNase , FAIRE , Histone , and TFBS	
DNase-seq Peaks	DNase-seq Peaks of Open Chromatin
ENCODE Peaks	ENCODE Peaks of Open Chromatin

ENCODE

DNA Methylation	
HAIB Methyl RRBS	DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha
HAIB Methyl450	CpG Methylation by Methyl 450K Bead Arrays from ENCODE/HAIB

Reduced representation bisulfites sequencing (RRBS)

- high-throughput technique
- genome-wide methylation
- profiles on a single nucleotide level
- combines restriction enzymes and bisulfites sequencing
 - enzyme that cuts DNA at or near specific regions called restriction sites
 - is the use of bisulfite treatment of DNA to determine its pattern of methylation



ENCODE

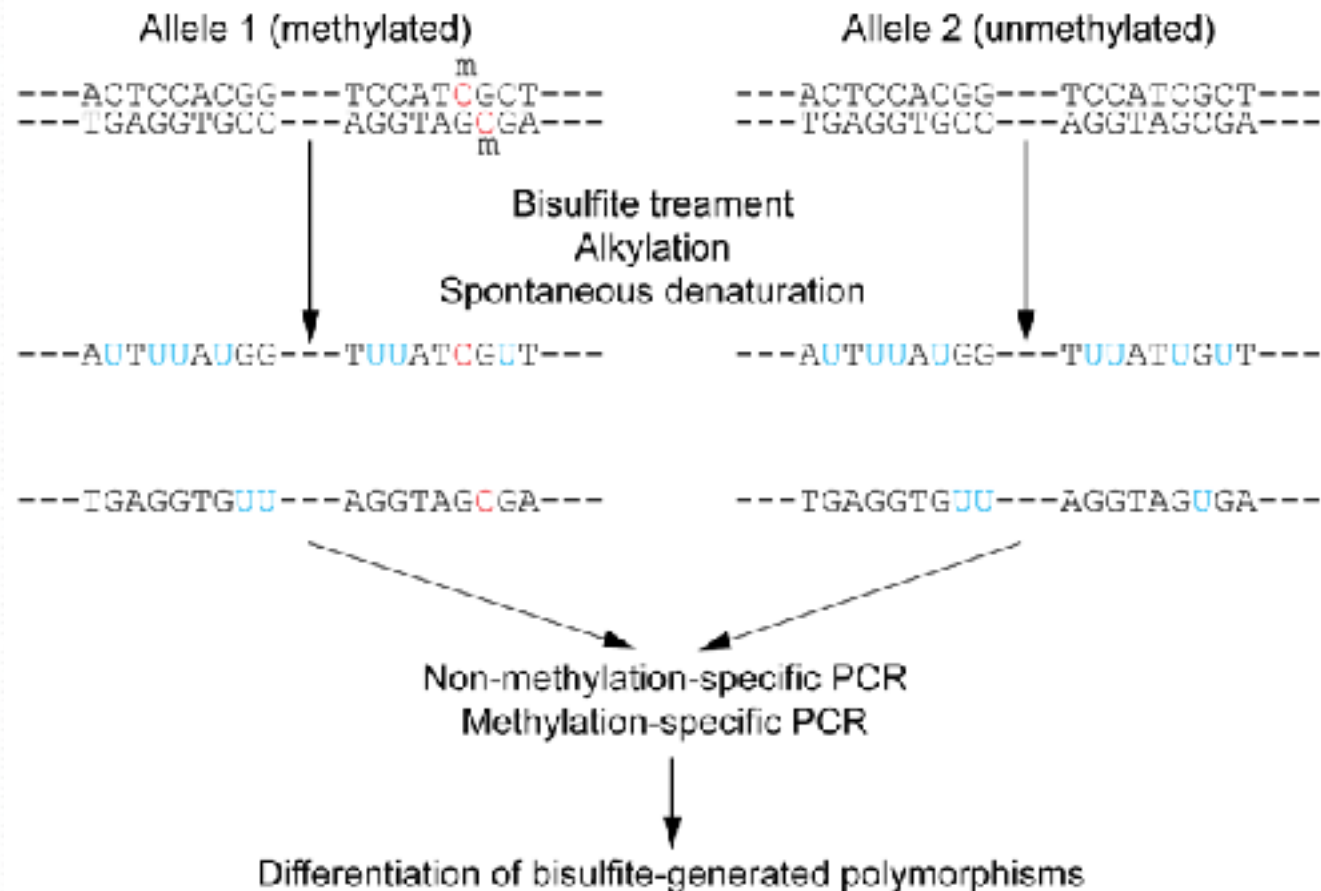
DNA Methylation

[HAIB Methyl RRBS](#)

DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha

[HAIB Methyl450](#)

CpG Methylation by Methyl 450K Bead Arrays from ENCODE/HAIB



ENCODE

DNA Methylation	
HAIB Methyl RRBS	DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha
HAIB Methyl450	CpG Methylation by Methyl 450K Bead Arrays from ENCODE/HAIB

RRBS protocol:

- Trawienie genomu enzymem restrykcyjnym niewrażliwym na metylację
- Enzymy nie są zależne od statusu metylacji sekwencji CpG
- MspI - enzym celuje w sekwencje 5'CGG3' i rozcina wiązania przed dinukleotydem CpG
- Naprawiane są lepkie końce 3'

Inny enzym restrykcyjny - PstI - rozpoznaje sekwencję 5'-CTGCAG-3' i przecina ją w następujący sposób:

5' C T G C A | G 3'
3' G | A C G T C 5'

co daje końce lepkie z wystającym 3':

5' C T G C A 3' 5' G 3'
3' G 5' 3' A C G T C 5'



ENCODE

DNA Methylation

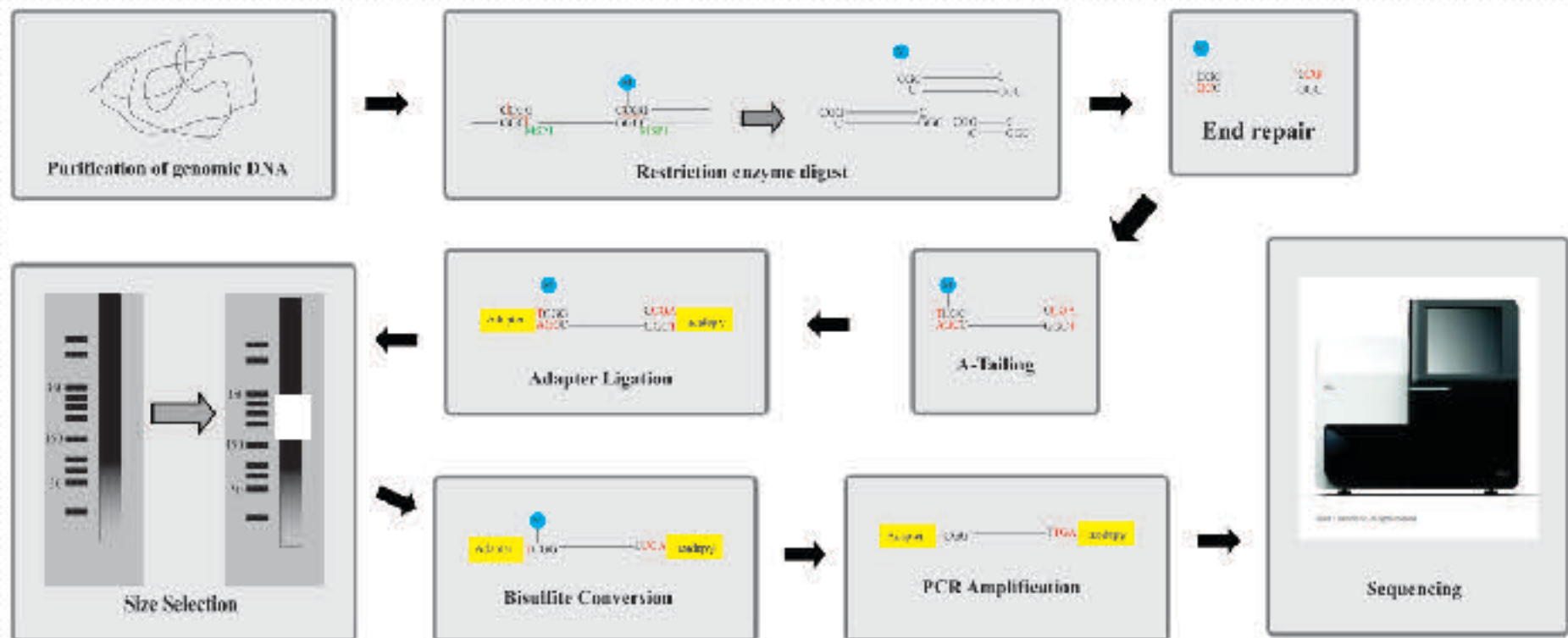
[HAIB Methyl RRBS](#)

DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha

[HAIB Methyl450](#)

CpG Methylation by Methyl 450K Bead Arrays from ENCODE/HAIB

RRBS protocol:



ENCODE

DNA Methylation

HAIB Methyl RRBS

DNA Methylation by Reduced Representation Bisulfite Seq from ENCODE/HudsonAlpha

HAIB Methyl450

CpG Methylation by Methyl 450K Bead Arrays from ENCODE/HAIB

http://rohsdb.cmb.usc.edu/GBshape/cgi-bin/hgTables?db=hg19&hgta_group=regulation&hgta_track=wgEncodeHaibMethyl450&hgta_table=wgEncodeHaibMethyl450Caco2SitesRep1&hgta_doSchema=describe+table+schema

Genomes Genome Browser Table Browser Tools Downloads Help on Table Browser

Schema for HAIB Methyl450 - CpG Methylation by Methyl 450K Bead Arrays from ENCODE/HAIB

Database: hg19 Primary Table: wgEncodeHaibMethyl450Caco2SitesRep1 Row Count: 482,421
Format description: Browser extensible data

Field	example	SQL type	info	description
bin	585	smallint(5) unsigned	range	Indexing field to speed chromosome range queries.
chrom	chr1	varchar(255)	values	Reference sequence chromosome or scaffold
chromStart	15885	int(11) unsigned	range	Start position in chromosome
chromEnd	15885	int(11) unsigned	range	End position in chromosome
name	cg1336834	varchar(255)	values	Name of item
score	805	int(11) unsigned	range	Score from 0-1000
strand	+	char(1)	values	+, or -
thickStart	15885	int(11) unsigned	range	Start of where display should be thick (start codon)
thickEnd	15885	int(11) unsigned	range	End of where display should be thick (stop codon)
itemRgb	16744152	int(11) unsigned	range	Used as ItemRgb as of 2004-11-22

Sample Rows

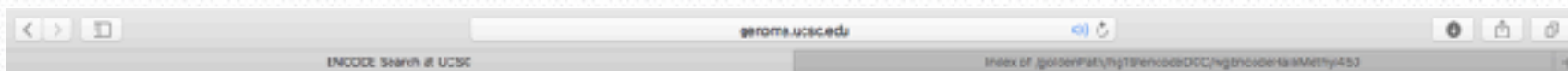
bin	chrom	chromStart	chromEnd	name	score	strand	thickStart	thickEnd	itemRgb
535	chr1	15884	15885	cg13069041	835	+	15884	15885	255,127,0
535	chr1	15826	15827	cg14091030	739	+	15826	15827	255,127,0
535	chr1	29405	29407	cg12045430	122	+	29406	29407	0,0,255
535	chr1	29424	29425	cg20825752	152	+	29424	29425	0,0,255
535	chr1	29434	29435	cg20391604	23	+	29434	29435	0,0,255
535	chr1	68848	68849	cg20253340	395	+	68848	68849	128,0,128
535	chr1	68901	68902	cg21070274	337	-	68901	68902	128,0,128
535	chr1	91560	91561	cg20130801	67	-	91560	91561	0,0,255
535	chr1	135252	135253	cg24335620	775	-	135252	135253	255,127,0
535	chr1	449075	449076	cg16162858	757	+	449075	449076	255,127,0

Note: all start coordinates in our database are 0-based, not 1-based. See explanation [here](#).

HAIB Methyl450 (wgEncodeHaibMethyl450) Track Description

ENCODE

<http://genome.ucsc.edu/encode/cellTypes.html>



ENCODE Cell Types 2007 - 2012

About ENCODE Cell Types

To facilitate integration of data between the contributing research groups, the ENCODE Consortium identifies common cell types for use by ENCODE contributors. During the first production phase of ENCODE (September 2007 to July 2012) the common cell types were divided into two **Tiers**. **Tier1** cells are of higher priority, and should be used within experiments before **Tier2** cells. Rationale for the selection is described on the [NHGRI ENCODE Common Cell Types](#) page. Additional cell types beyond the designated Tier1 and Tier2 could be used for ENCODE production; these are selected at the discretion of individual data production groups, and are designated Tier3.

Click the link in the **Documents** column of the table below to access the cell culture protocol document.

Common Cell Types: Tier 1 and Tier 2

Cell, tissue or DNA sample: Cell line or tissue used as the source of experimental material.

cell ¹	Tier ²	Description ³	Lineage ⁴	Tissue ⁵	Karyotype	Sex	Documents	Vendor ID	Term ID	Label
GM12878	1	B-lymphocyte, lymphoblastoid, International HapMap Project - CEPH/Utah - European Caucasian, Epstein-Barr Virus	mesoderm	blood	normal	F	ENCODE	Coriell GM12878	BTQ:0002042	GM12878
H1-hESC	1	embryonic stem cells	inner cell mass	embryonic stem cell	normal	M	ENCODE	WCell Research Institute WA01	CL:000007	H1-hESC
K562	1	leukemia, "The continuous cell line K-562 was established by Lozzio and Lozzio from the pleural effusion of a 53-year-old female with chronic myelogenous leukemia in terminal blast crises." - ATCC	mesoderm	blood	cancer	F	ENCODE	ATCC CCL-243	BTQ:0008064	K562

Total = 3

Cell, tissue or DNA sample: Cell line or tissue used as the source of experimental material.

cell ¹	Tier ²	Description ³	Lineage ⁴	Tissue ⁵	Karyotype	Sex	Documents	Vendor ID	Term ID	Label
A549	2	epithelial cell line derived from a lung carcinoma tissue. (PMID: 175022), "This line was initiated in 1972 by D.J. Gard, et al. through explant culture of lung carcinomatous tissue from a 50-year-old caucasian male."	endoderm	epithelium	cancer	M	Myers Crawford	ATCC CCL-185	BTQ:0008118	A549

ENCODE

<https://genome.ucsc.edu/cgi-bin/hgFileSearch?db=hg19>

The screenshot shows the UCSC Genome Browser's ENCODE Files Search page. The browser's address bar displays 'genome.ucsc.edu'. The page title is 'Search for Downloadable ENCODE Files in the Human Feb. 2009 (GRCh37/hg19) Assembly'. The search interface includes several input fields and dropdown menus:

- Track Name:** contains [text input]
- and Description:** contains [text input]
- and Group:** is [Any] [dropdown]
- and Data Format:** is [Any] [dropdown]

Below these fields, there are two additional search criteria:

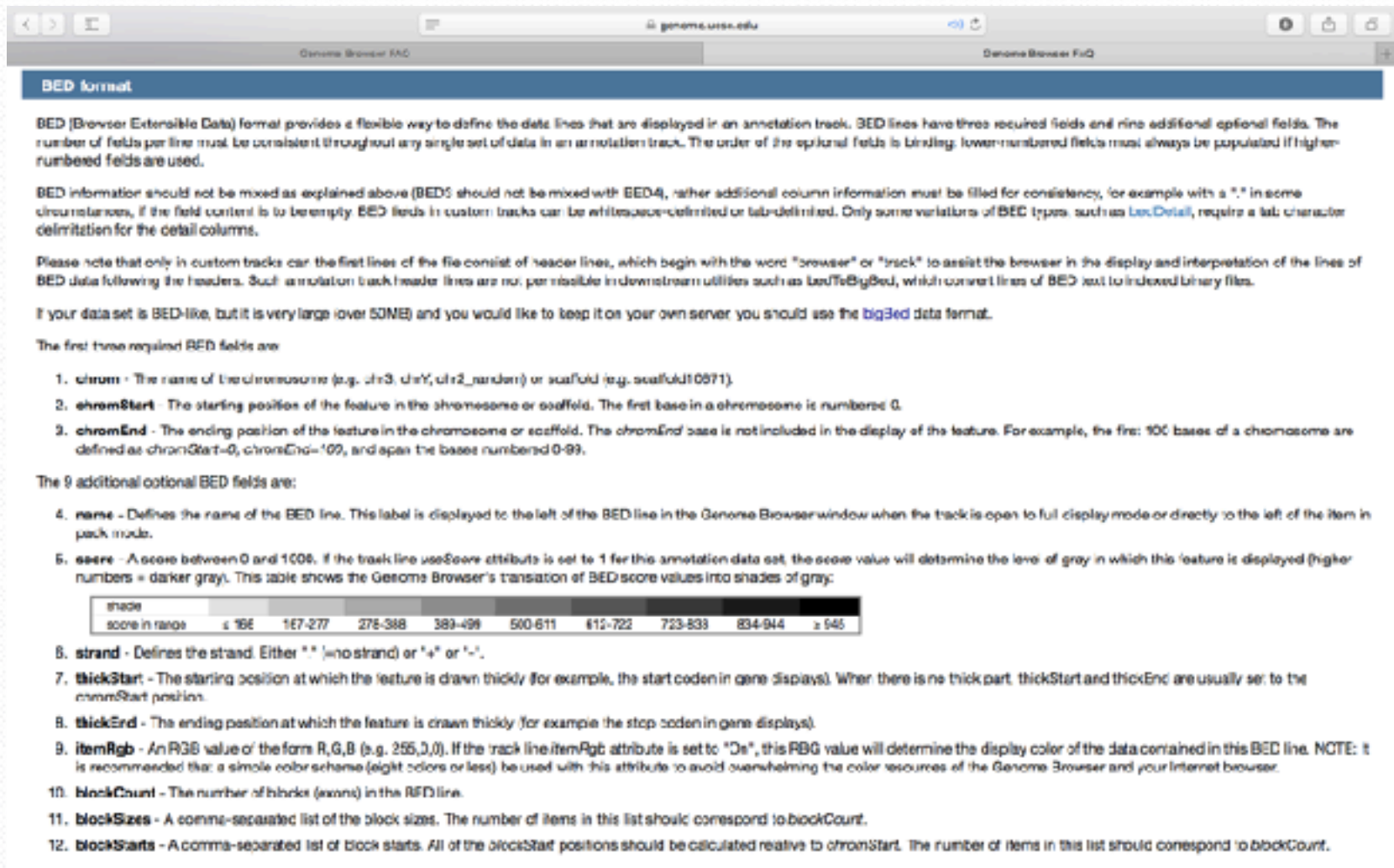
- Cell, tissue or DNA sample:** is among [Any] [dropdown] [Cell, tissue or DNA sample](#)
- Antibody or target protein:** is among [Any] [dropdown] [Antibody or target protein](#)

At the bottom of the search section, there are three buttons: 'search', 'clear', and 'cancel'. Below the search section, there is a section titled 'About Downloadable ENCODE Files Search' which provides instructions on how to use the search interface. The background of the page is a light yellow color.

ENCODE

<https://genome.ucsc.edu/FAQ/FAQformat.html>

<https://genome.ucsc.edu/FAQ/FAQformat.html#format1>



The screenshot shows the UCSC Genome Browser FAQ page for the BED format. The page title is "BED format". The content explains the BED (Browser Extensible Data) format, which provides a flexible way to define the data lines that are displayed in an annotation track. It states that BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track. The order of the optional fields is binding; lower-numbered fields must always be populated if higher-numbered fields are used.

BED information should not be mixed as explained above (BED3 should not be mixed with BED4; rather additional column information must be filled for consistency, for example with a "." in some circumstances, if the field content is to be empty). BED fields in custom tracks can be whitespace-delimited or tab-delimited. Only some variations of BED types, such as [bedDetail](#), require a tab character delimitation for the detail columns.

Please note that only in custom tracks can the first lines of the file consist of header lines, which begin with the word "browser" or "track" to assist the browser in the display and interpretation of the lines of BED data following the headers. Such annotation track header lines are not permissible in downstream utilities such as [bedToBigBed](#), which convert lines of BED text to indexed binary files.

If your dataset is BED-like, but it is very large (over 50MB) and you would like to keep it on your own server, you should use the [bigBed](#) data format.

The first three required BED fields are:

1. **chrom** - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10571).
2. **chromStart** - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.
3. **chromEnd** - The ending position of the feature in the chromosome or scaffold. The **chromEnd** base is not included in the display of the feature. For example, the first 100 bases of a chromosome are defined as **chromStart**=0, **chromEnd**=100, and span the bases numbered 0-99.

The 9 additional optional BED fields are:

4. **name** - Defines the name of the BED line. This label is displayed to the left of the BED line in the Genome Browser window when the track is open to full display mode or directly to the left of the item in pack mode.
5. **score** - A score between 0 and 1000. If the track line useScore attribute is set to 1 for this annotation data set, the score value will determine the level of gray in which this feature is displayed (higher numbers = darker gray). This table shows the Genome Browser's translation of BED score values into shades of gray:

shade
score in range
≤ 166
167-277
278-388
389-499
500-611
612-722
723-833
834-944
≥ 945
6. **strand** - Defines the strand. Either "." (no strand) or "+" or "-".
7. **thickStart** - The starting position at which the feature is drawn thickly (for example, the start codon in gene displays). When there is no thick part, **thickStart** and **thickEnd** are usually set to the **chromStart** position.
8. **thickEnd** - The ending position at which the feature is drawn thickly (for example the stop codon in gene displays).
9. **itemRgb** - An RGB value of the form R,G,B (e.g. 255,0,0). If the track line itemRgb attribute is set to "On", this RGB value will determine the display color of the data contained in this BED line. NOTE: it is recommended that a simple colorscheme (eight colors or less) be used with this attribute to avoid overwhelming the color resources of the Genome Browser and your Internet browser.
10. **blockCount** - The number of blocks (exons) in the BED line.
11. **blockSizes** - A comma-separated list of the block sizes. The number of items in this list should correspond to **blockCount**.
12. **blockStarts** - A comma-separated list of block starts. All of the **blockStart** positions should be calculated relative to **chromStart**. The number of items in this list should correspond to **blockCount**.

ENCODE - metylacje

1	chr16	53468112	53468162	cg00000029	601 +	53468112	53468162	255,127,0
2	chr3	37459206	37459256	cg00000108	952 +	37459206	37459256	255,127,0
3	chr3	171916037	171916087	cg00000109	918 +	171916037	171916087	255,127,0
4	chr1	91194674	91194724	cg00000165	335 -	91194674	91194724	128,0,128
5	chr8	42263294	42263344	cg00000236	699 -	42263294	42263344	255,127,0
6	chr14	69341139	69341189	cg00000289	727 +	69341139	69341189	255,127,0
7	chr16	28890100	28890150	cg00000292	550 +	28890100	28890150	128,0,128
8	chr8	41167802	41167852	cg00000321	845 -	41167802	41167852	255,127,0
9	chr1	230560793	230560843	cg00000363	645 +	230560793	230560843	255,127,0
10	chr15	23034447	23034497	cg00000622	14 +	23034447	23034497	0,0,205
11	chr9	139997924	139997974	cg00000658	829 -	139997924	139997974	255,127,0
12	chr19	54695678	54695728	cg00000714	279 +	54695678	54695728	128,0,128



ENCODE - geny

[illegible]

Pytania

- Wyznacz 100 genów które istotnie różnią się ekspresją pomiędzy IDH wild type i IDH mutant (df z poprzednich ćwiczeń na temat glejaków i TCGA).
- Wyznacz dla tych genów promotory (2000 bp upstream i 500 bp downstream) - plik w formacie .bed (skorzystaj z plików rds i bed z tych zajęć).
- Wyznacz dla tych genów zakresy eksonów - format .bed
- Wyznacz dla tych genów zakresy intronów - format .bed

