

# MEDICAL GENOMICS

**Василий Евгеньевич Раменский**

[ramensky@gmail.com](mailto:ramensky@gmail.com)

**Анастасия Александровна Жарикова**

[azharikova89@gmail.com](mailto:azharikova89@gmail.com)

НМИЦ Терапии и профилактической медицины  
Факультет биоинженерии и биоинформатики МГУ

# MEDICAL GENOMICS

- Big genomic data enable inference without intervention (aka reverse genetics)
- Data are just around the corner



МОСКВА, 1868.

Из фото. Стрел. Митрофан. Г. в. Моск.

ЕКАТЕРИНИНСКАЯ БОЛЬНИЦА  
У ПЕТРОВСКИХ ВОРУТЬ



У ПЕТРОВСКИХ ВОРУТЬ

# MEDICAL GENOMICS


## **Part I. Кто виноват?**

1. Mutations: origins and rates
2. Mutations: transmission
3. Mutations in time: some basics of population genetics
4. Mutations in space: genes and consequences
5. Mutations in individuals and populations

## **Part II. Что делать?**

6. Mendelian diseases: gene discovery and diagnostics
7. Some basics of quantitative genetics
8. Complex diseases: gene discovery and allelic architecture

# Remarks

- Important info: kodomo
- English
- Molecular genetics + population genetics + medical genetics + statistical genetics + genetic epidemiology + bioinformatics  $\Rightarrow$  no single textbook
- Many topics not covered: immunology, pathogens, microbiome, therapy, genome editing
  
- Definitions
- Questions and exercises
- Homework slides 
- Summary, concepts, further reading



# Textbooks

1. T.Strachan, A.Read. Human Molecular Genetics. 2011. ISBN 0815345895.
2. J. Gillespie. Population genetics. A concise guide 1998 ISBN 0-8018-5764-6
3. S. Szalai, et al. Medical genetics and genomics. 2016. [https://www.researchgate.net/publication/303309837\\_Medical\\_genetics\\_and\\_genomics\\_2016](https://www.researchgate.net/publication/303309837_Medical_genetics_and_genomics_2016)
4. A.Griffiths et al. An Introduction to Genetic Analysis. Freeman/Worth, 11 ed. 2015 ISBN 1464109486
5. Бочков Н.П., Пузырев В.П., Смирнихина С.А. Клиническая генетика. Учебник. Под ред. Н.П. Бочкова. ГЭОТАР-Медиа, 4-е издание, 2018. ISBN 978-5-9704-4628-7



**MUTATIONS:**

ORIGINS AND RATES

# Lecture plan

- Human karyotype
- Mitosis and DNA replication
- Replication fidelity and mutation rate
- Exogenous and endogenous DNA damage
- DNA repair mechanisms
- De novo mutations: single nucleotide variants
- Structural variants and CNVs
- Aneuploidy

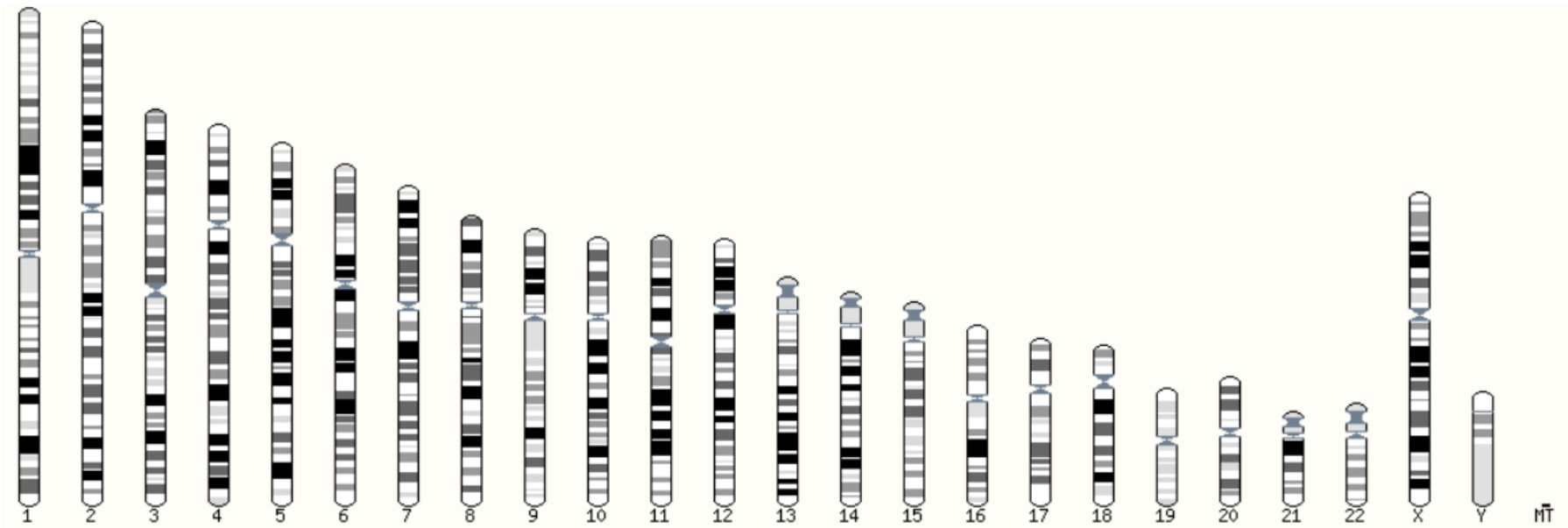
# Human karyotype



22 pairs of **autosomal** chromosomes + 2 **sex** chromosomes

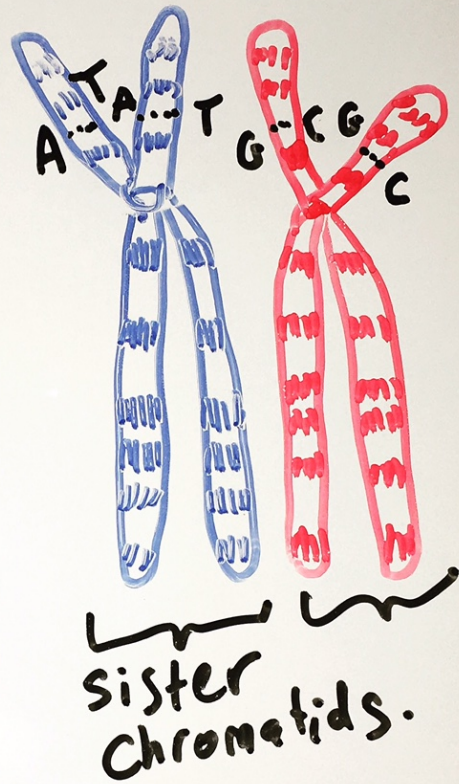
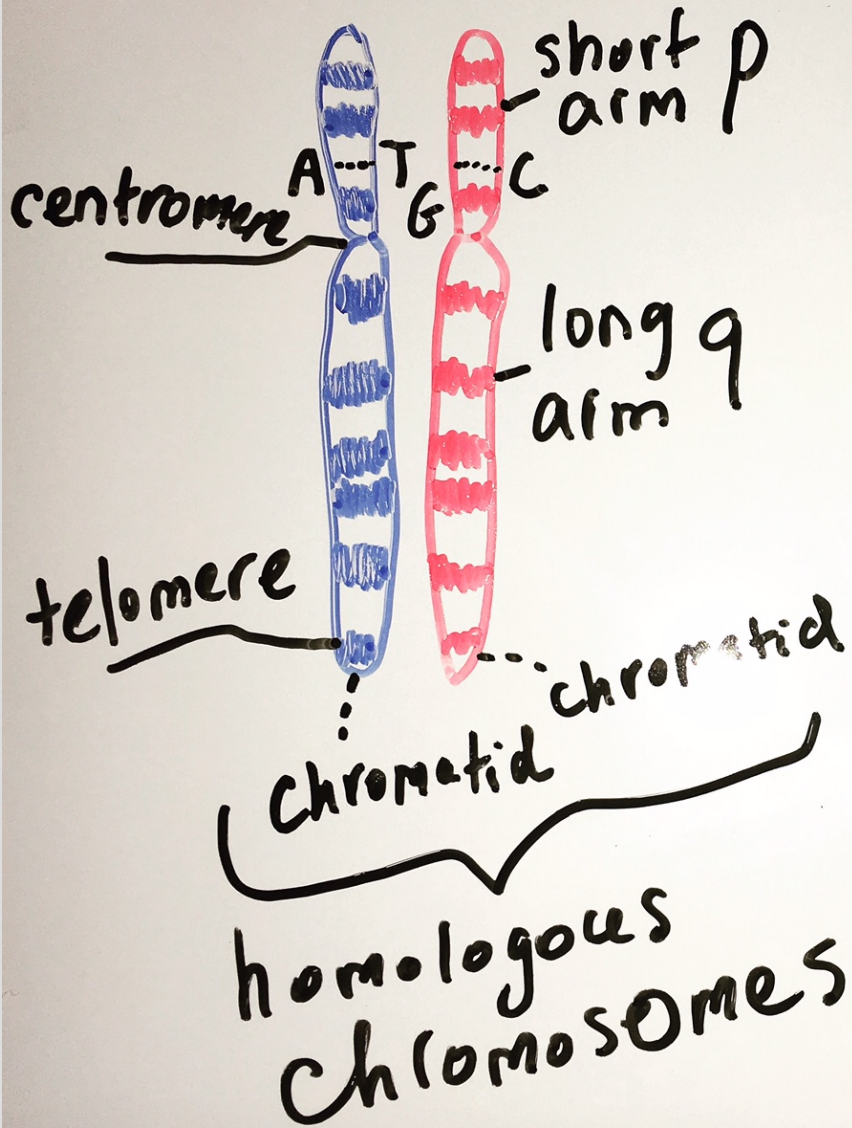
# Human karyotype

- **Euchromatin** (2.9 Gbp): the gene-rich, transcriptionally active regions of the nuclear genome
- **Heterochromatin** (0.2 Gbp): tightly packed (condensed), transcriptionally inactive, highly repetitive DNA. Location: centromeres, telomeres.
- **Metacentric chromosomes** have the centromere in the center, such that both arms are of nearly equal length.
- **Acrocentric chromosomes** (13, 15, 21, 22) have unequal arms.

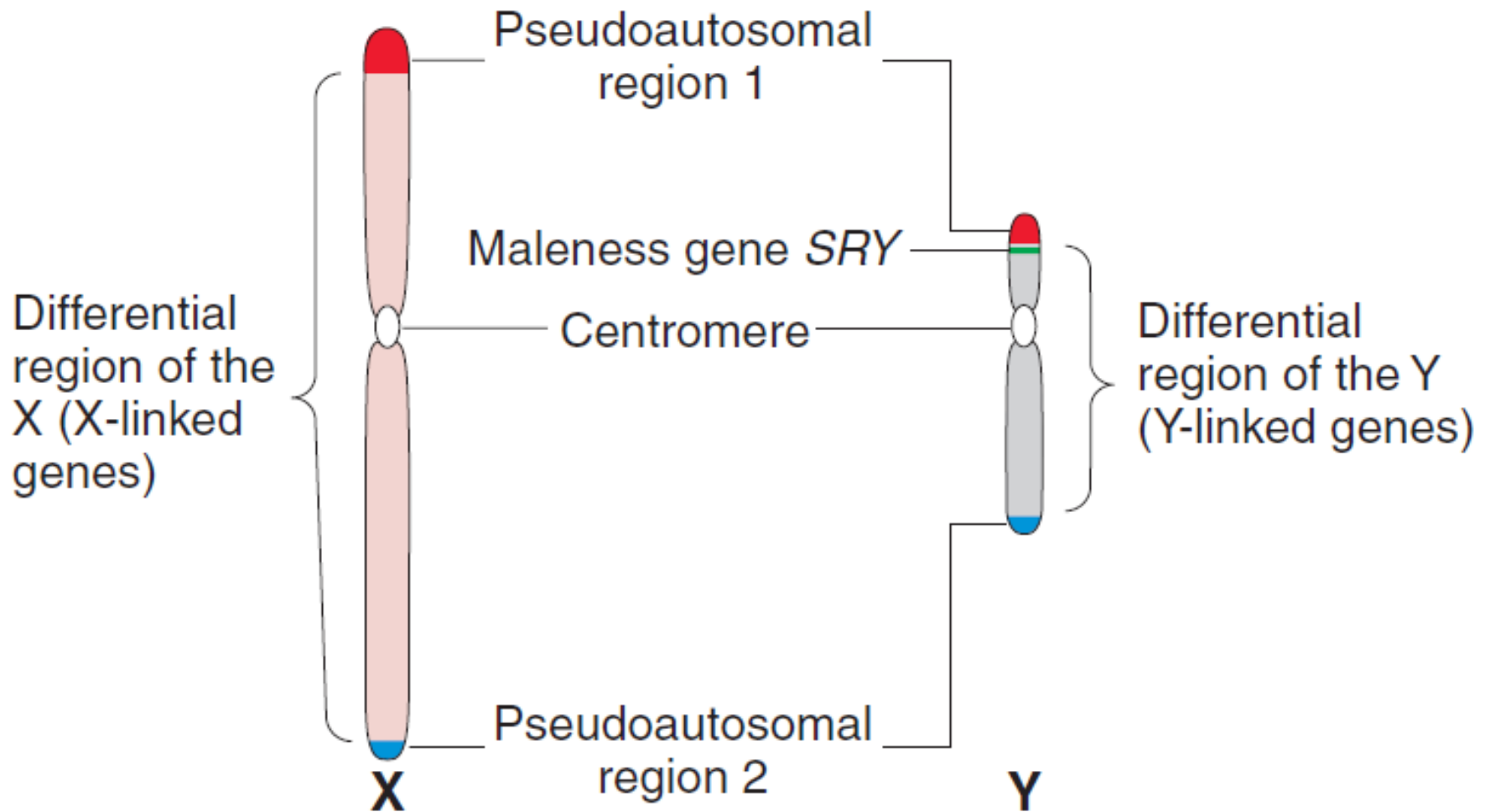




A:T G:C alleles



# Sex chromosomes



Women: XX, men: XY

Q: transmission of Y chromosome

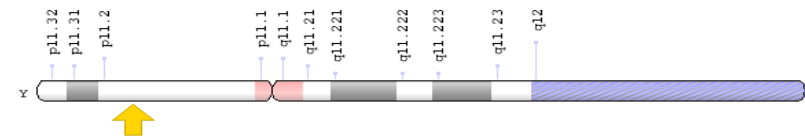
# How to check sex of an NGS sample?

# How to check sex of an NGS sample?

- **Heterozygous / Homozygous ratio on chromosome X**

Het/Hom < 0.8 => M, Het/Hom > 0.8 => F; suspicious : 0.5-1.0

- ***SRY* gene** (Sex-determining Region Y):  
intronless sex-TF protein, responsible for  
the initiation of male sex determination  
in mammals

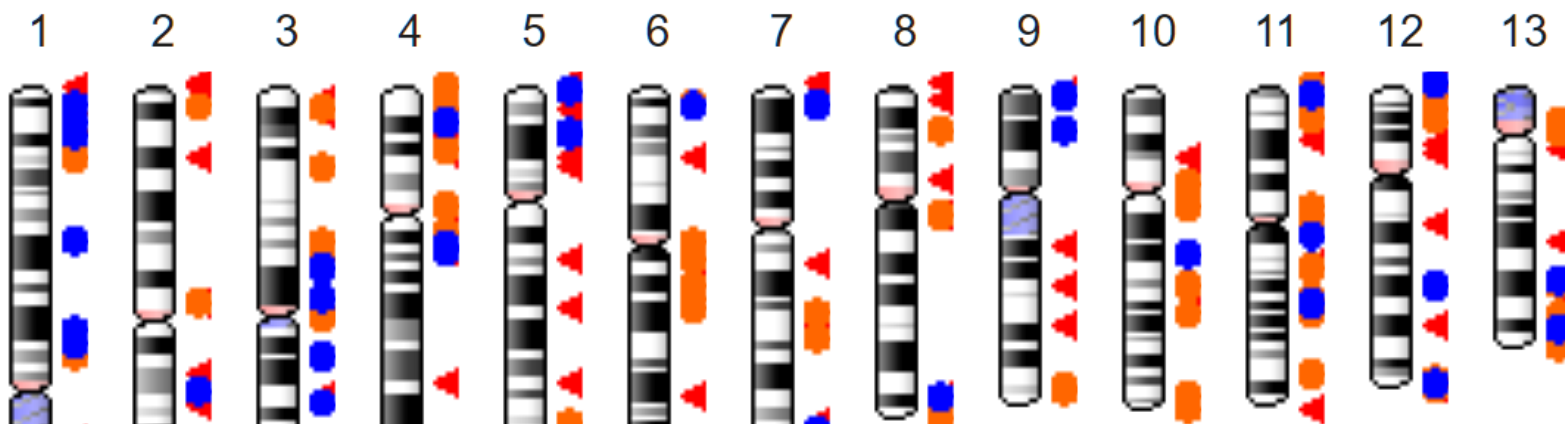


- **The human amelogenin genes: *AMELX* and *AMELY***

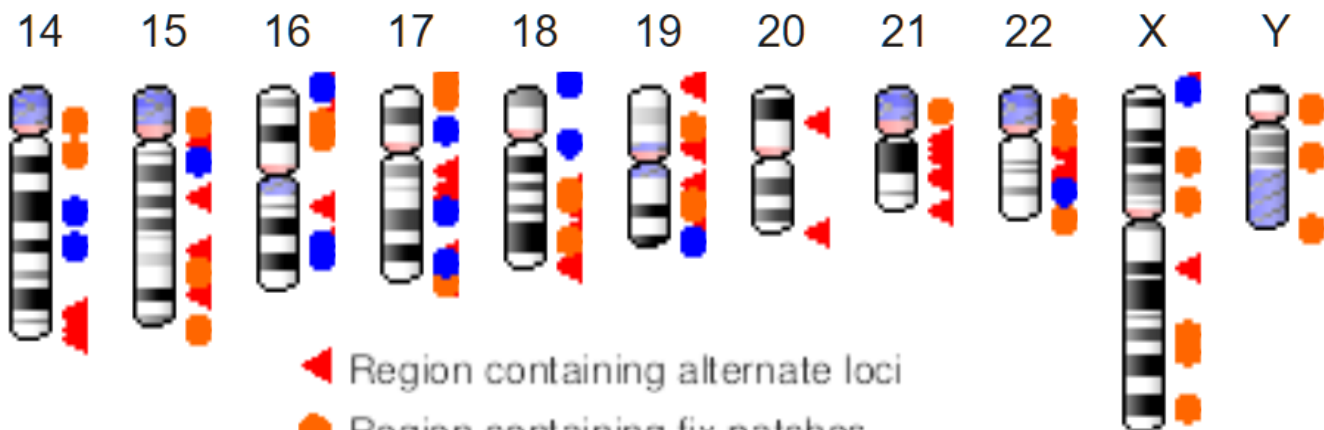
Short arms of X and Y sex chromosomes, share 84% sequence identity. A 6 bp insertion/deletion difference in the first intron of the *AMELY* and *AMELX* genes is typically targeted in forensic sex identification (Tzvetkov 2010 *Pharmacogenomics*)



# Genome Reference Consortium GRCh38.p13



Chromosomes	Mbp
Nuclear	3088
Mitochondrial	0.017
Unknown, alt	121.0



- ◄ Region containing alternate loci
- ◼ Region containing fix patches
- Region containing novel patches

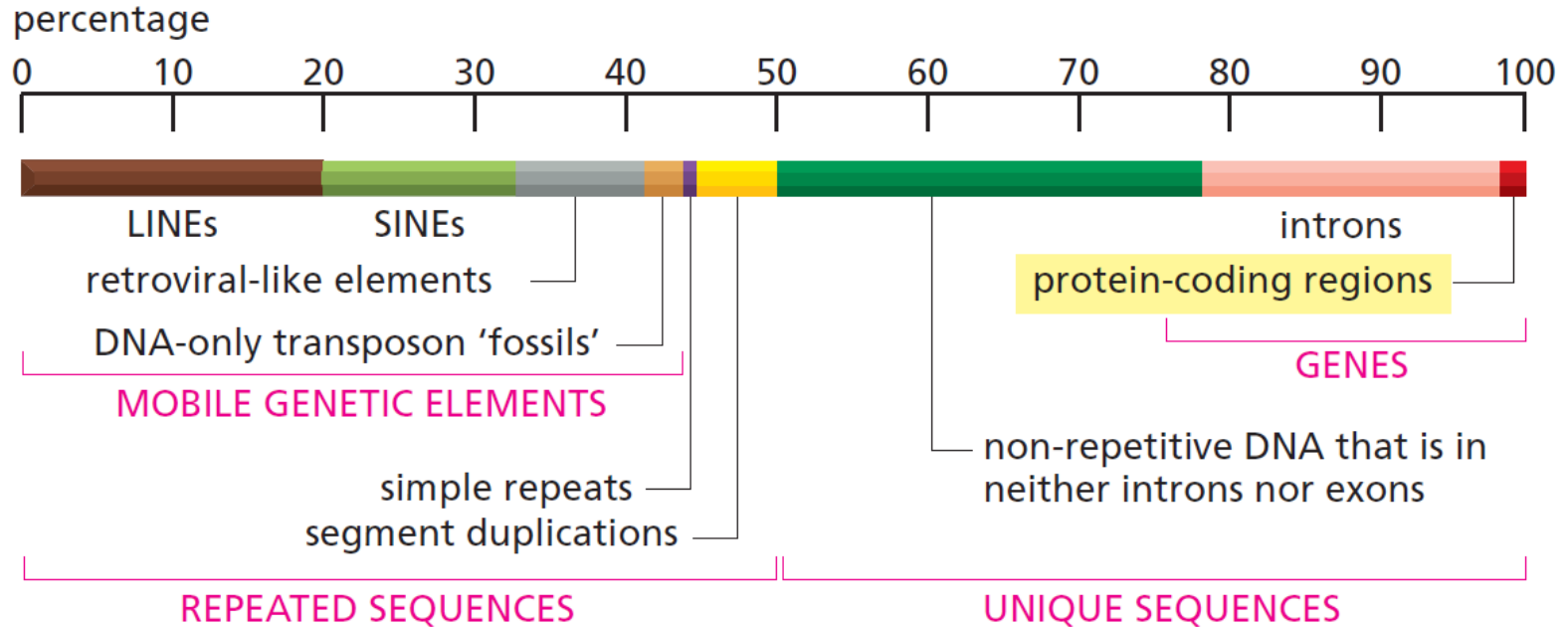


# Human genome contents

Regions	Length, Mbp	%	Description
<b>Genes</b>	1,200	37.5	Genomic locus where transcription occurs
Exons	48	1.5	Transcribed genomic region that remain in the RNA after splicing
Other (introns, UTRs)	1,152	36.0	Regions of a coding cDNA which are not translated
<b>Unique and regulatory sequences</b>	510	15.9	
<b>Interspersed repeats</b>	1,400	43.8	
LINES	640	20.0	~850,000 Long Interspersed Elements (~7,000 bp). Retrotransposed elements containing open reading frames encoding (often inactive) reverse transcription machinery
SINEs, Alu repeats	420	13.1	~1,500,000 Short Interspersed Elements. Retrotransposed elements <500 bp that contain tRNA, snRNA and rRNA, which require other mobile elements to be transposed.
LTR retrotransposons	250	7.8	Transposable elements characterized by the presence of Long Terminal Repeats (LTRs) directly flanking an internal coding region
DNA transposons	90	2.8	Class II transposable elements that move through a DNA intermediate
<b>Microsatellites</b>	90	2.8	A region in the genomic sequence containing short tandem repeats of 2-10bp
<b>Total</b>	3,200	100.0	



# Human genome contents



Alberts - *Essential Cell Biology*

Element	Transposition	Structure	Length	Copy number	Fraction of genome
LINES	Autonomous		1–5 kb	20,000–40,000	21%
SINEs	Nonautonomous		100–300 bp	1,500,000	13%
DNA transposons	Autonomous		2–3 kb	300,000	3%
	Nonautonomous		80–3000 bp		

# ENSEMBL gene annotation GRCh38 v.99

Gene biotype	Genes (Transcripts)	%	Description
Protein coding	19,968 (153,197)	32.9	Genes that contain an open reading frame (ORF)
Pseudogenes	15,263	25.2	Genes that have homology to known protein-coding genes but contain a frameshift and/or stop codon(s) which disrupts the ORF
To be confirmed	1,060	1.7	Require experimental validation
T-cell receptors, immunoglobulins	408	0.7	Undergo somatic recombination before transcription
RNA genes	23,977	39.5	
lncRNA	16,880		A non-coding gene >200bp in length
snRNA	1,910		Processing of pre-messenger RNA
miRNA	1,879		A small RNA (~22bp) that silences the expression of target mRNA
snoRNA	942		Post-transcriptional modification of other RNAs
Other	2,366		rRNA, sRNA, scRNA, scaRNA, miscRNA
<b>Total</b>	<b>60,676 (227,818)</b>	<b>100</b>	

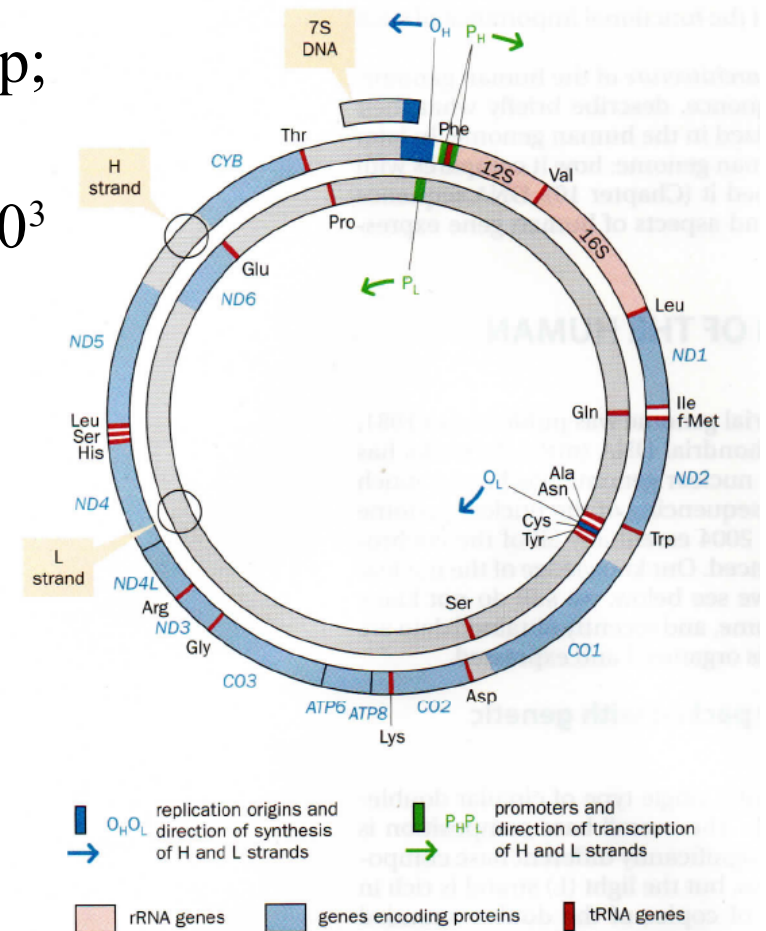
# ENSEMBL gene annotation GRCh38 v.99

Chromosome	Approximate length (bp)	Protein-coding genes	Non-protein coding genes	Pseudogenes
1	248956422	2047	1964	1233
2	242193529	1303	1605	1033
3	198295559	1075	1160	768
4	190214555	753	984	732
5	181538259	881	1200	710
6	170805979	1041	989	803
7	159345973	989	977	893
8	145138636	670	1041	629
9	138394717	778	786	678
10	133797422	728	880	568
11	135086622	1312	1053	815
12	133275309	1036	1197	627
13	114364328	321	586	378
14	107043718	820	857	519
15	101991189	613	986	513
16	90338345	867	1033	467
17	83257441	1185	1198	531
18	80373285	269	608	246
19	58617616	1474	895	514
20	64444167	543	594	250
21	46709983	231	403	183
22	50818468	492	513	332
X	156040895	843	640	872
Y	57227415	63	108	392
Mitochondrial	16569	13	24	



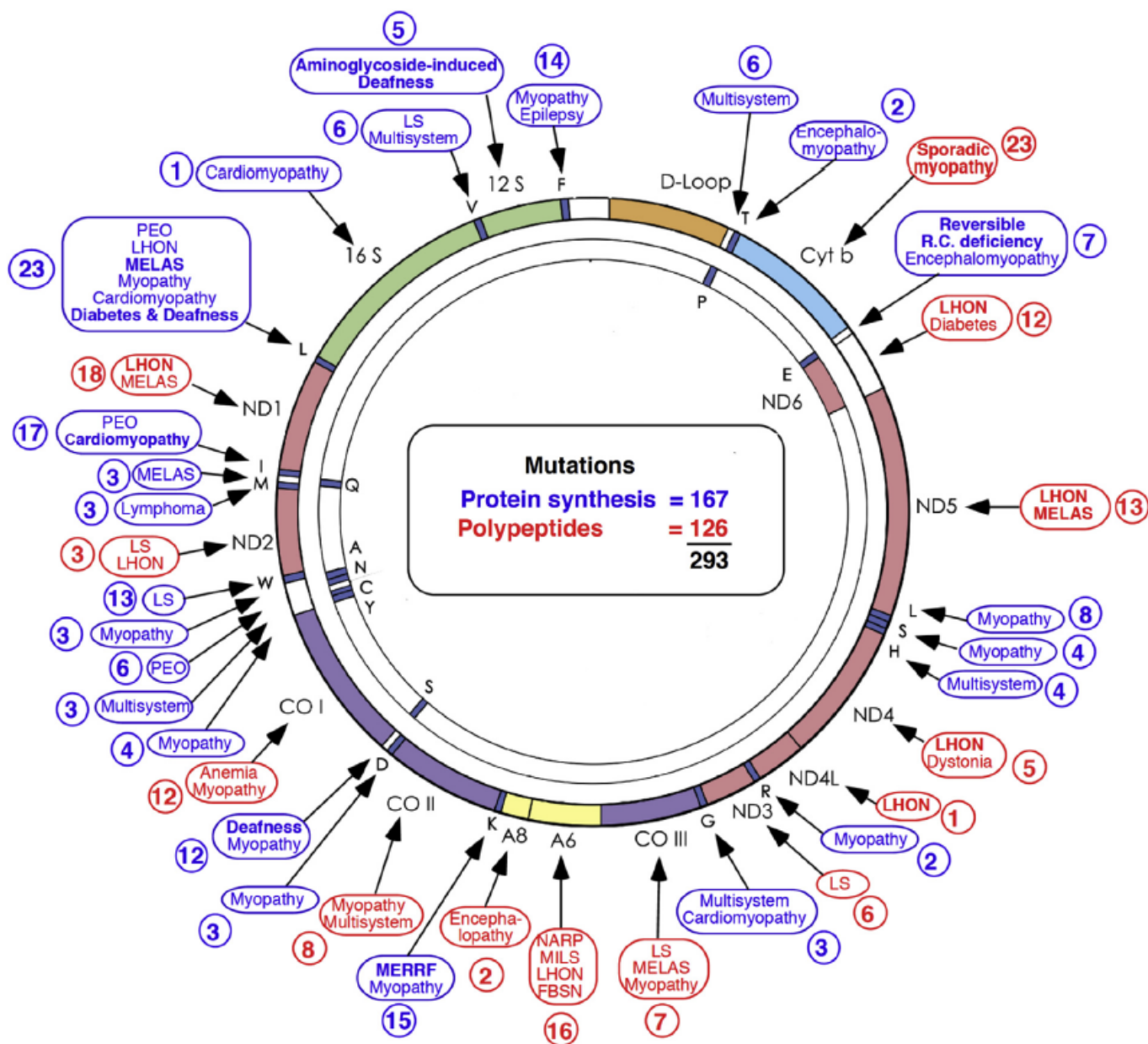
# Mitochondrial genome

- **mtDNA**: circular, double-stranded, 16,569 bp; H and L chains; similar to bacteria
- Egg only, maternally inherited; each cell:  $\sim 10^3$  copies; highly heterogeneous
- 37 genes: 22 tRNA + 2 rRNA + 13 coding
- 13 polypeptides are part of **mitochondrial respiratory complex** (Sugars  $\rightarrow$  ATP), together with multiple nuclear genes
- mtDNA is to some extent autonomous, with its own genetic code
- Stop codons: TAA, TAG, AGA, AGG
- **Mitochondrial diseases**: a heterogeneous group of inherited anomalies in oxidative phosphorylation due to mutations in the mitochondrial (70%) or nuclear DNA (30%)
- $\sim 300$  disease-causing point mutations known in mtDNA





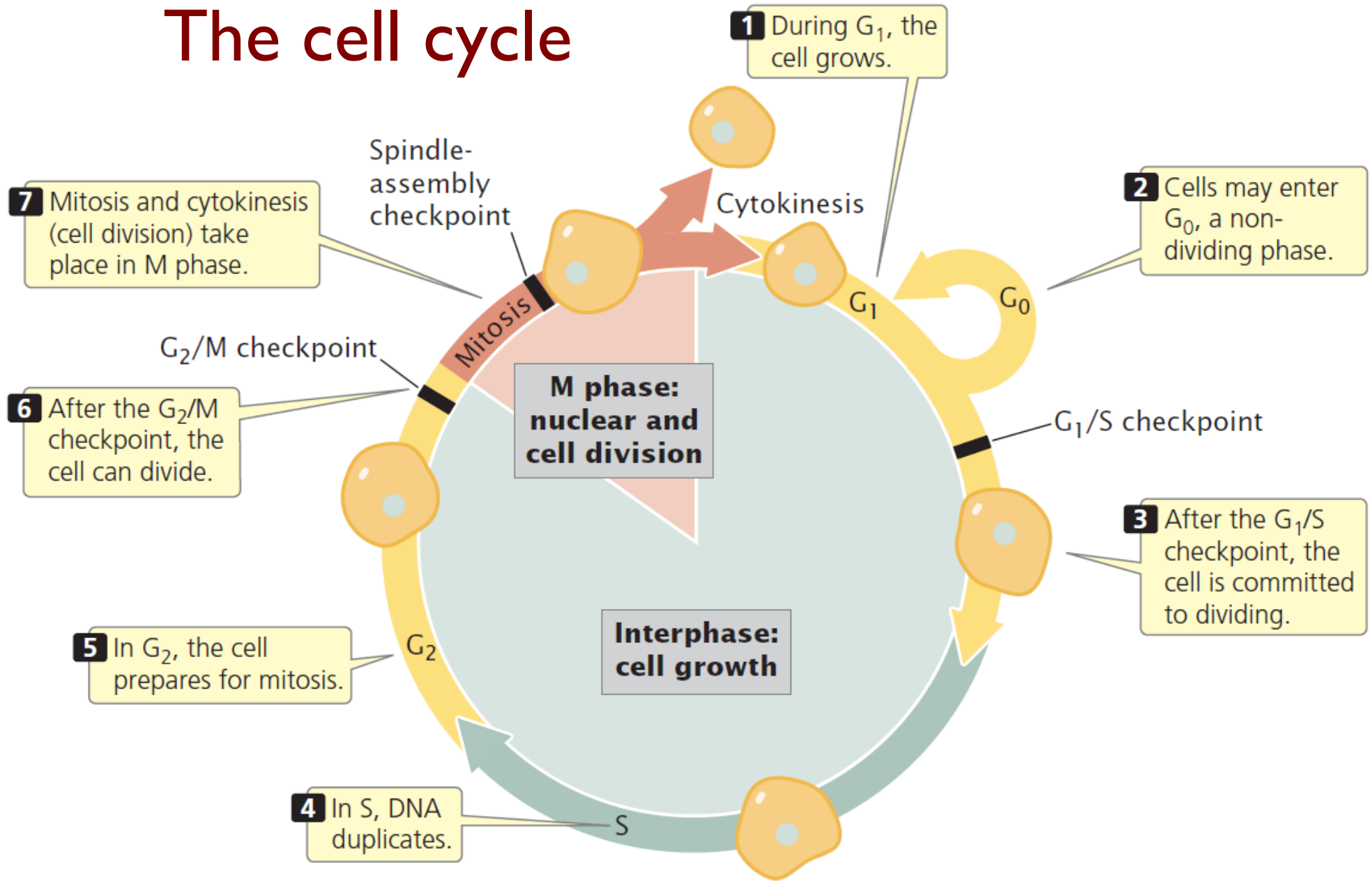
# Pathogenic mutations in mtDNA





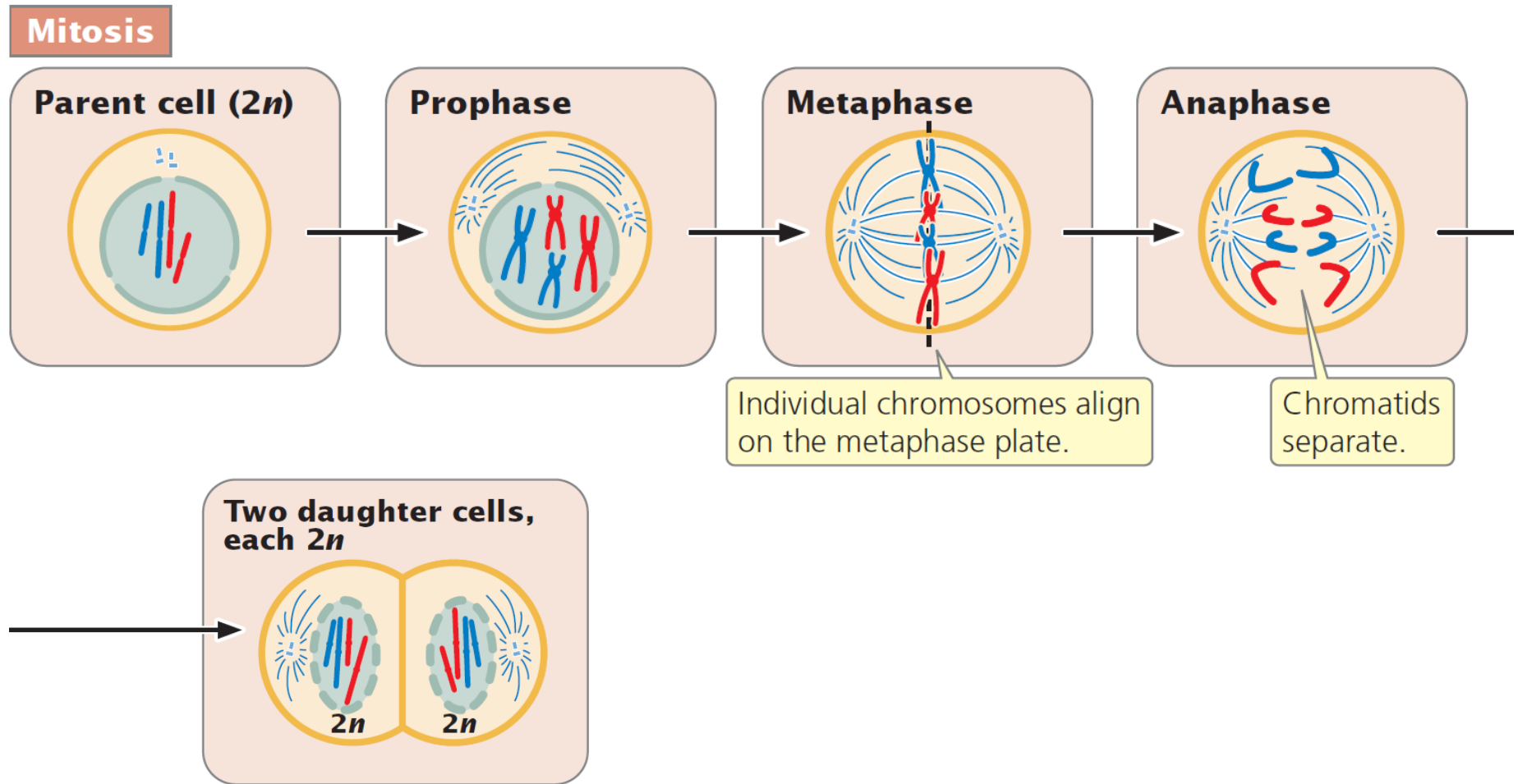


# The cell cycle



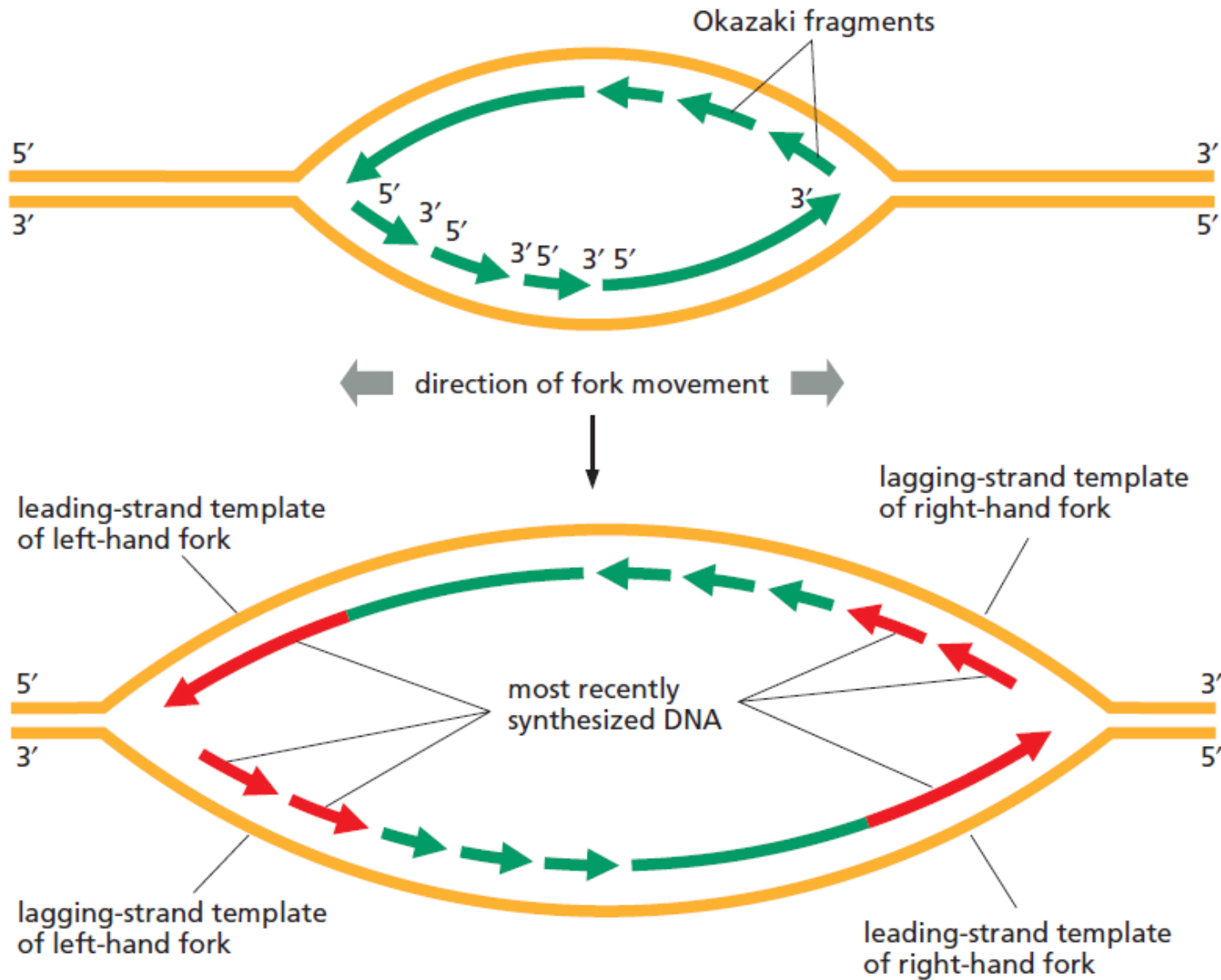
## 2.7 The cell cycle consists of interphase and M phase.

# Mitosis

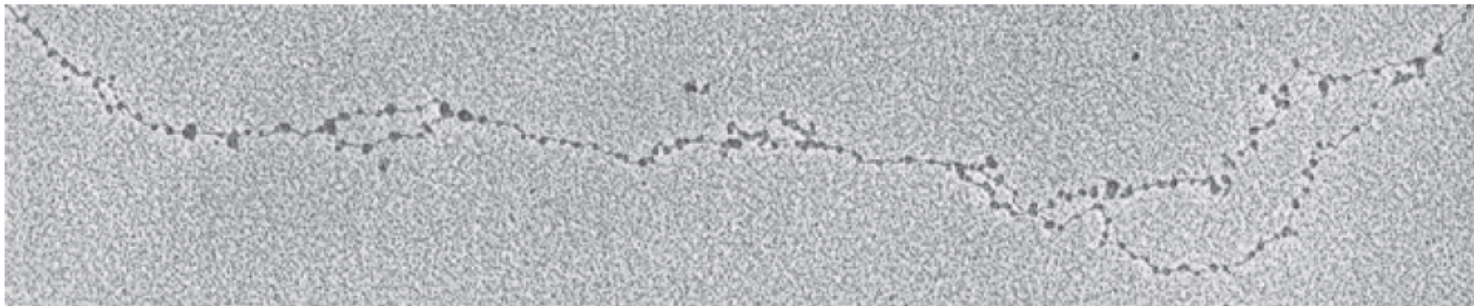
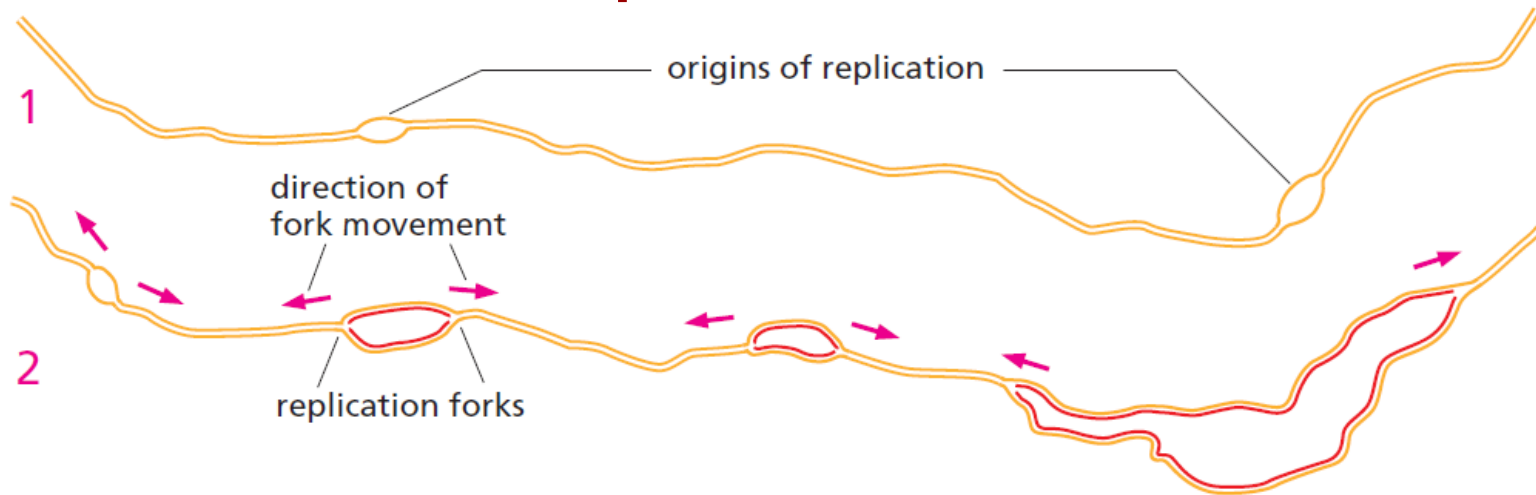


**Mitosis:** a type of cell division that results in two daughter cells with the set of chromosomes as the parent nucleus, typical of ordinary tissue growth

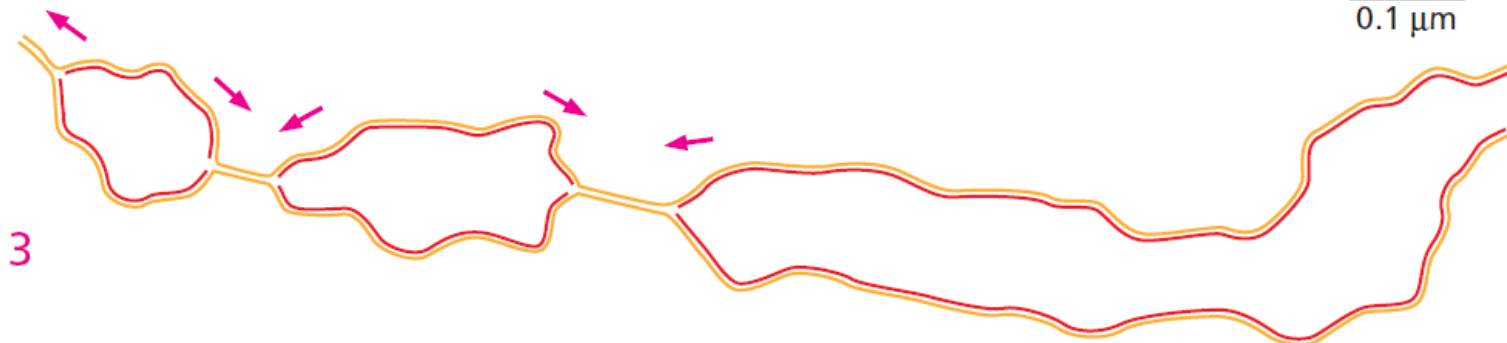
# DNA replication forks



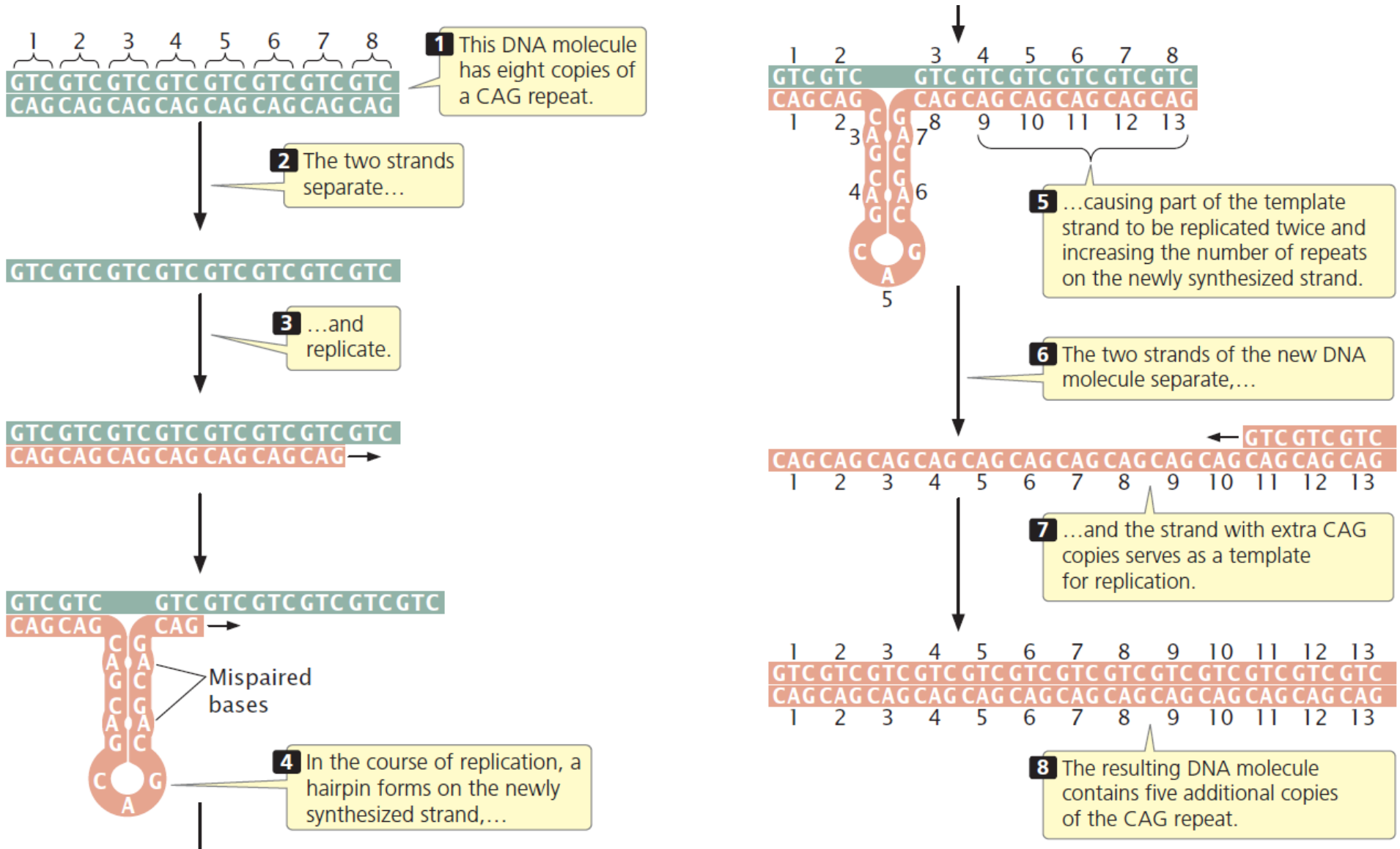
# DNA replication forks



0.1  $\mu\text{m}$

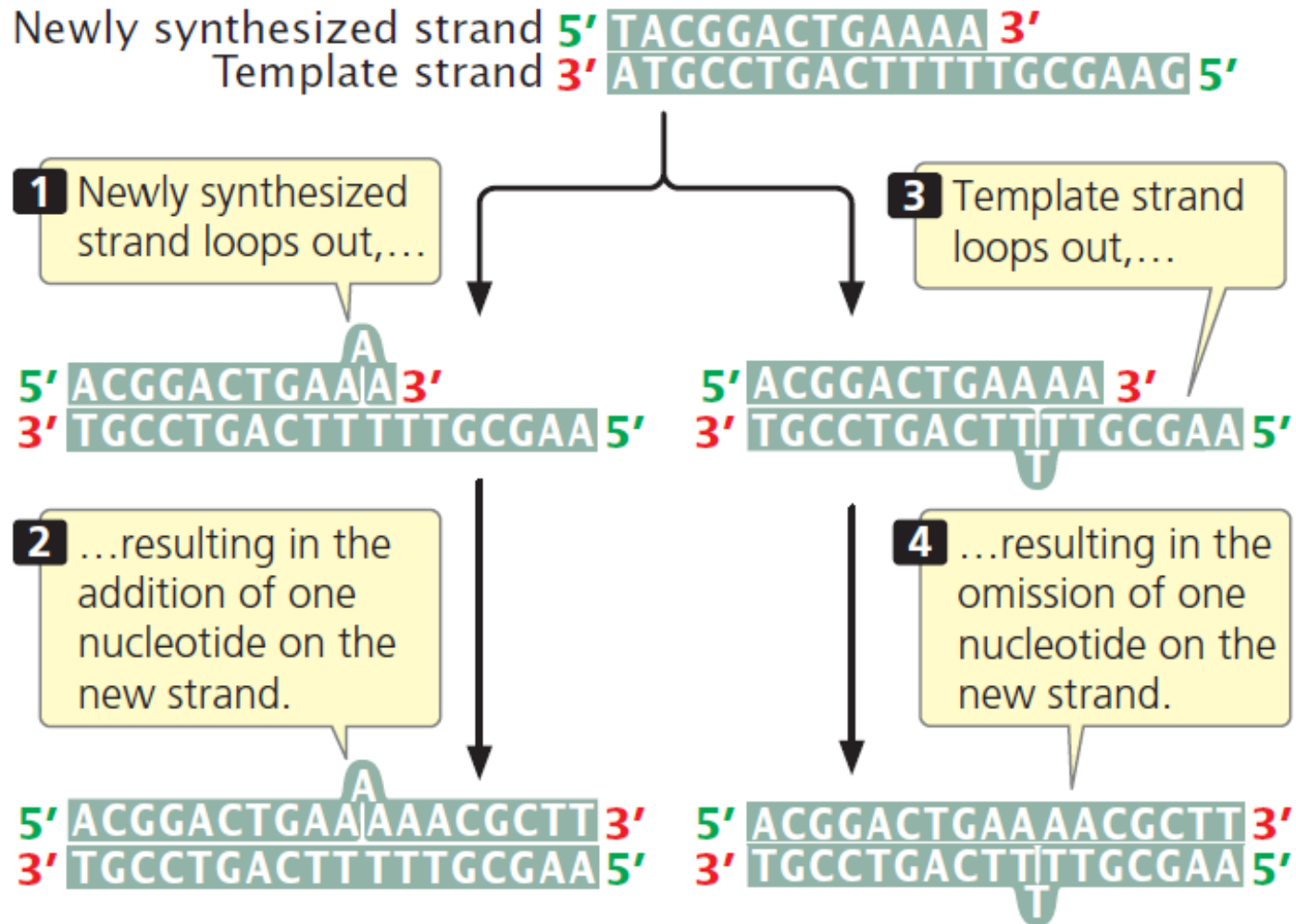


# Repeat expansion during replication





# Repeat expansion during replication



13.13 Insertions and deletions may result from strand slippage.

# Repeat expansion and disease

**Table 13.1** Examples of genetic diseases caused by expanding trinucleotide repeats

Disease	Repeated Sequence	Number of Copies of Repeat	
		Normal Range	Disease Range
Spinal and bulbar muscular atrophy	CAG	11–33	40–62
Fragile-X syndrome	CGG	6–54	50–1500
Jacobsen syndrome	CGG	11	100–1000
Spinocerebellar ataxia (several types)	CAG	4–44	21–130
Autosomal dominant cerebellar ataxia	CAG	7–19	37–220
Myotonic dystrophy	CTG	5–37	44–3000
Huntington disease	CAG	9–37	37–121
Friedreich ataxia	GAA	6–29	200–900
Dentatorubral-pallidoluysian atrophy	CAG	7–25	49–75
Myoclonus epilepsy of the Unverricht–Lundborg type*	CCCCGCCCGCG	2–3	12–13

*Exercise:* find related genes in OMIM database





# OMIM<sup>®</sup>

## Online Mendelian Inheritance in Man<sup>®</sup>

An Online Catalog of Human Genes and Genetic Disorders

Updated February 12, 2021



Dissected OMIM Morbid Map Scorecard (Updated February 12th, 2021) :

Class of phenotype	Phenotype	Gene *
Single gene disorders and traits	5,740	4,006
Susceptibility to complex disease or infection	694	499
"Nondiseases"	151	119
Somatic cell genetic disease	231	130

\*Some genes may be counted more than once because mutations in a gene may cause more than one phenotype and the phenotypes may be of different classes (e.g., activating somatic BRAF mutation underlying cancer, [164757.0001](#). and germline BRAF mutation in Noonan syndrome, [164757.0022](#).)

# Mutations

**Mutations** are random changes in DNA sequences

Mutations are the cause of all genetic variation and genetic disease.

Mechanisms of mutation:

- Spontaneous replication errors
- Endogenous (spontaneous) DNA damage: deamination, depurination
- Exogenous (induced) DNA damage: chemical agents, radiation

**Variants** = mutations (recent changes), polymorphisms (segregating in a population), engineered (non-random) changes

# Mutations

**Single nucleotide variant:** change of the base of a single DNA nucleotide (90%)

- Transition (G>A, C>T)
- Transversion (C>G, etc)

**Short deletion:** removal of few (<50bp?) nucleotides (6%)

- Deletion of a unique sequence
- Contraction of a short repeat

**Short insertion:** addition of few (<50bp?) nucleotides (2%),

- Insertion of a unique sequence
- Expansion of a short repeat

**Structural variant (2%):** sequence change ~1 kbp and larger in size

- Balanced

Inversion or translocation

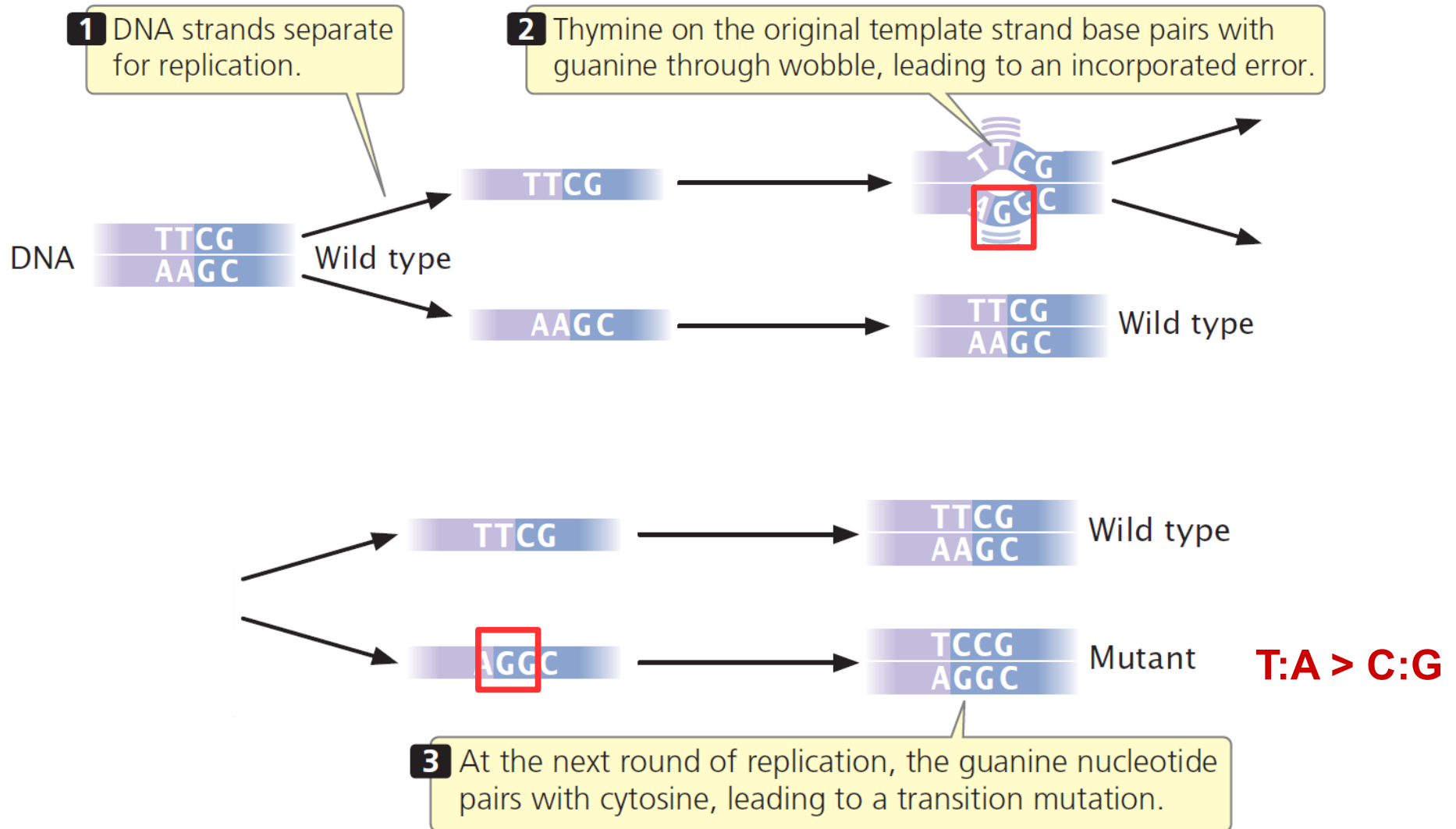
- Unbalanced (aka **CNV, copy number variant**)

Tandem or dispersed duplication, deletion, insertion

**Aneuploidy:** wrong number of whole chromosomes: nullisomy, monosomy, trisomy



# Replication errors become mutations





# Standard and non-standard base pairing

## Standard base-pairing arrangements



Thymine (common form)

Adenine (common form)



Cytosine (common form)

Guanine (common form)

## Anomalous base-pairing arrangements



Cytosine (rare form)

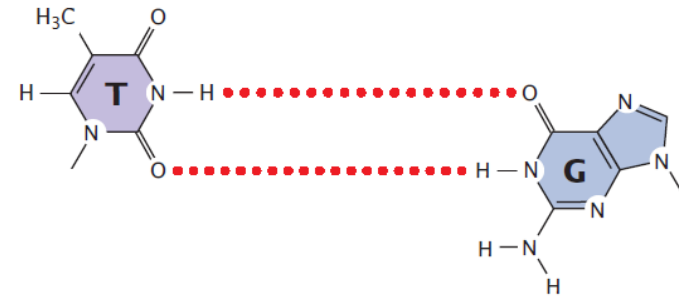
Adenine (common form)



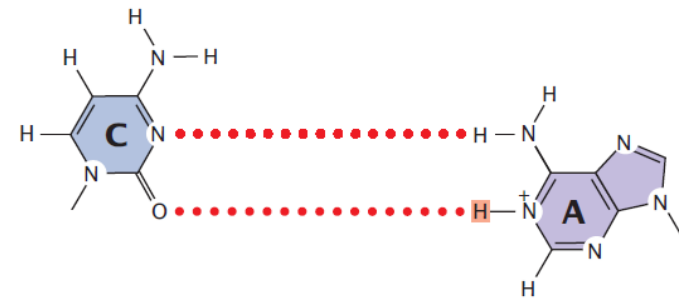
Thymine (common form)

Guanine (rare form)

## Non-Watson-and-Crick base pairing



Thymine-guanine wobble

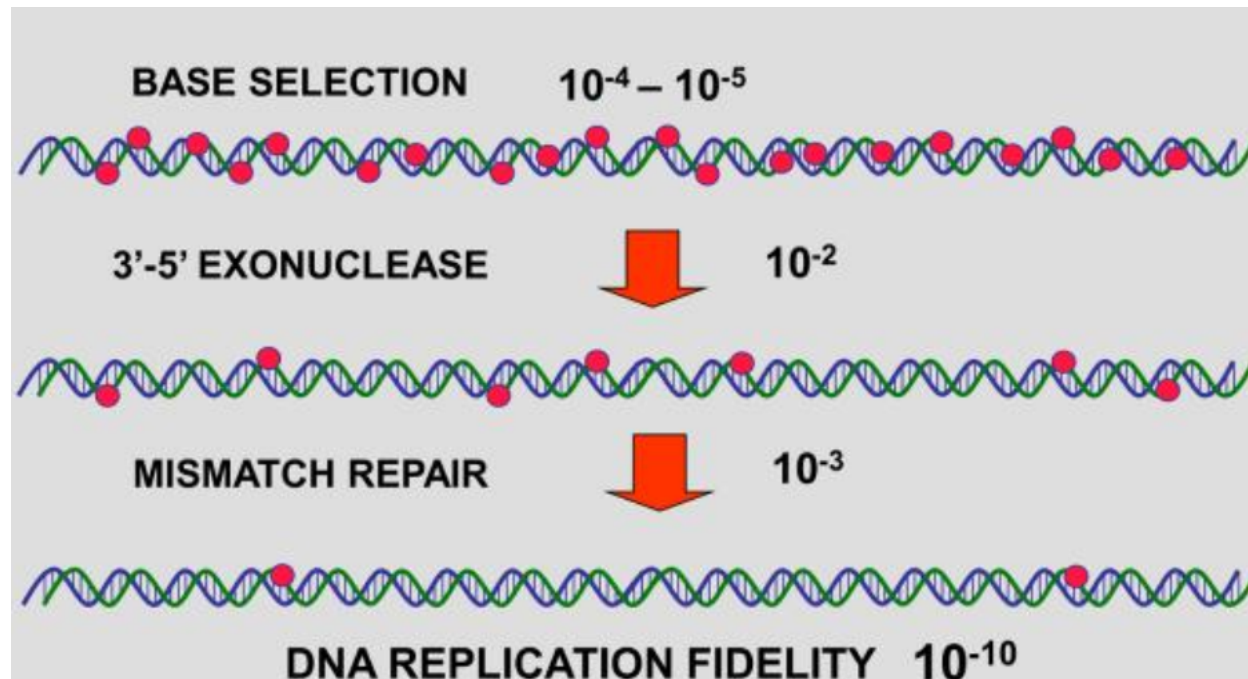


Cytosine-adenine protonated wobble

# Mechanisms of replication fidelity

**Overall mutation rate:**  $10^{-10}$  per nucleotide per replication

1. DNA polymerase:  $\sim 10^{-5}$  error rate
2. Proofreading 3' $\rightarrow$ 5' exonuclease removes 99% mispairings:  $\sim 10^{-2}$
3. Mismatch repair (MMR) machinery removes and restores DNA fragment around the mismatch:  $\sim 10^{-3}$



# Mechanisms of replication fidelity

**Overall mutation rate:**  $10^{-10}$  per nucleotide per replication

**TABLE 6–1 ERROR RATES**

US Postal Service on-time delivery of local first-class mail	13 late deliveries per 100 parcels
Airline luggage system	1 lost bag per 200
A professional typist typing at 120 words per minute	1 mistake per 250 characters
Driving a car in the United States	1 death per $10^4$ people per year
DNA replication (without mismatch repair)	1 mistake per $10^7$ nucleotides copied
DNA replication (including mismatch repair)	1 mistake per $10^9$ nucleotides copied

# Mutation rate and its consequences

$S$ : mutation rate per nucleotide per cell division

$K$ : the average number of germline cell divisions per generation, from zygote to zygote ( $\sim 30$  in females,  $\sim 60\text{--}500$  in males)

$N$ : genome size

**Mutation rate per genome:**  $S \times K \times N$

$\sim 10^{-10}$  per nucleotide per cell division (or  $\sim 10^{-8}$  per generation, because there are  $\sim 100$  cell divisions and rounds of DNA replication per human generation  $\Rightarrow \sim 100$  *de novo* mutations in a newborn

1)  $\sim 1\%$  of all newborns being affected by a serious disease due to a *de novo* mutation. If the mutation rate were 100 times higher,  $10^{-8}$  per cell division, we would immediately **go extinct**.

2)  $10^{14}$  cells in human body  $\Rightarrow$  total number of somatic mutations in each person ?



# Mutation rate and its consequences

Genes Genet. Syst. (2019) 94, p. 13–22

## Spontaneous *de novo* germline mutations in humans and mice: rates, spectra, causes and consequences

Mizuki Ohno\*

The human body consists of approximately  $10^{14}$  cells and undergoes approximately  $10^{16}$  cell divisions in a lifetime, resulting in **over  $10^{15}$  cumulative mutations per individual** (Frank, 2014).

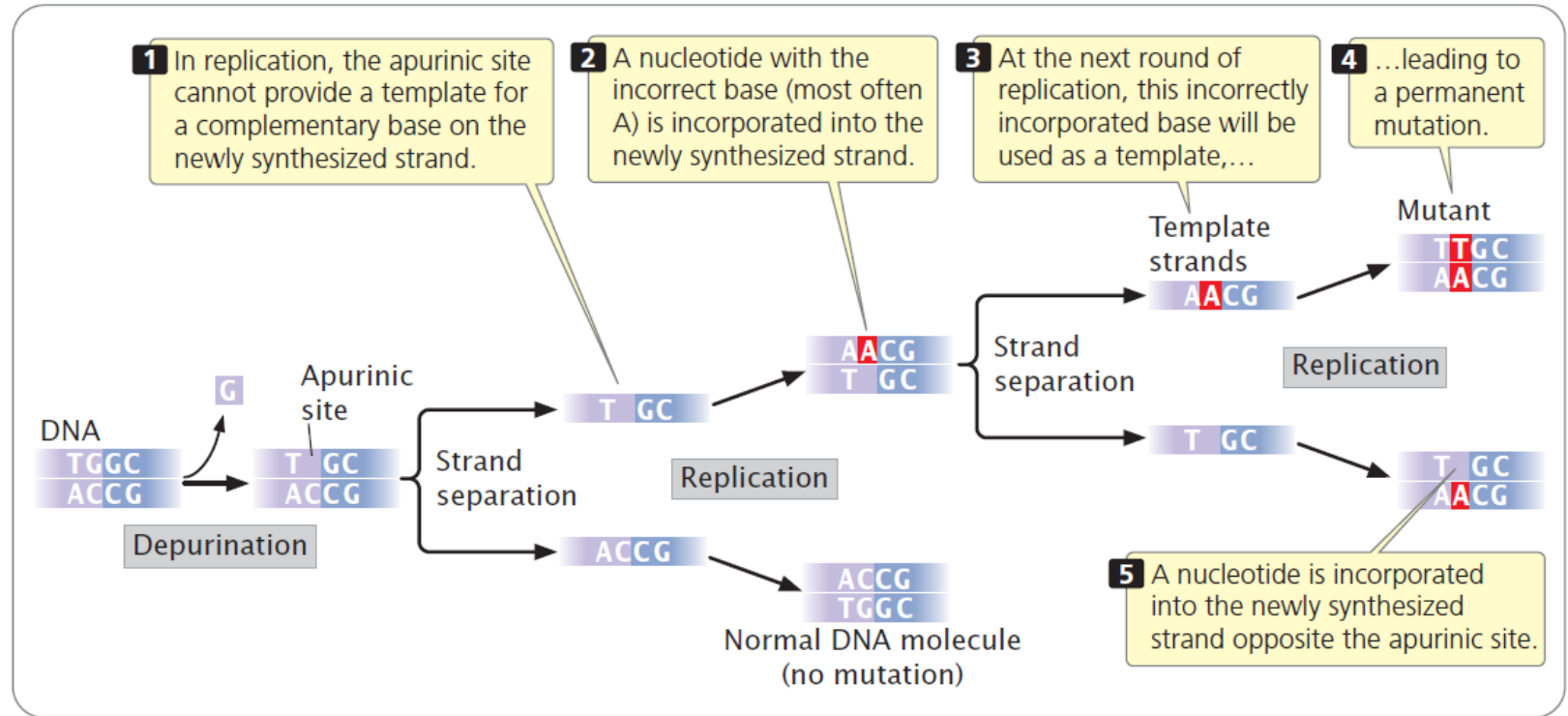
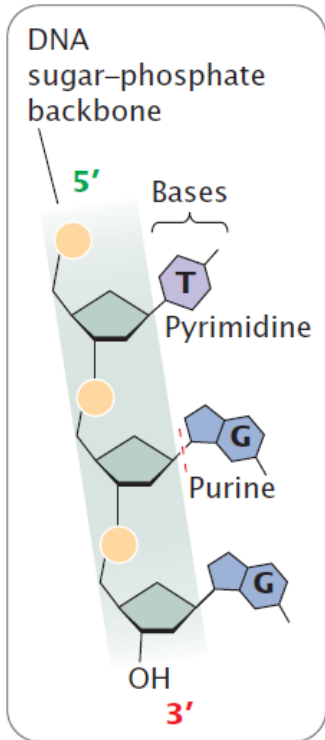
If  $10^6$  stem cells in intestinal tissue generate transient daughter cells once a week with a mutation rate of approximately  $10^{-9}$  per nucleotide per cell division, the intestinal epithelium of a 60-year-old human would have accumulated more than  $10^9$  independent mutations. Thus, **nearly every genomic site is likely to be mutated in at least one cell in this organ** (Lynch, 2010a, 2010b).





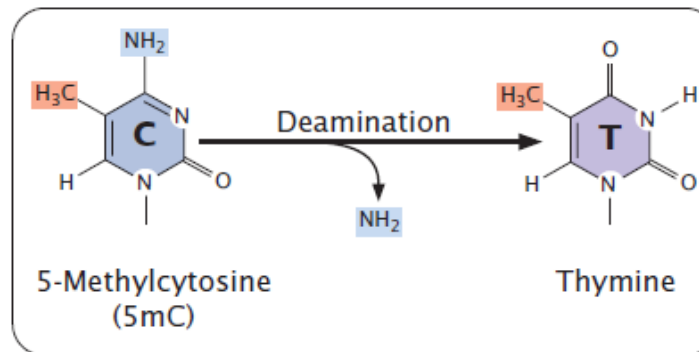
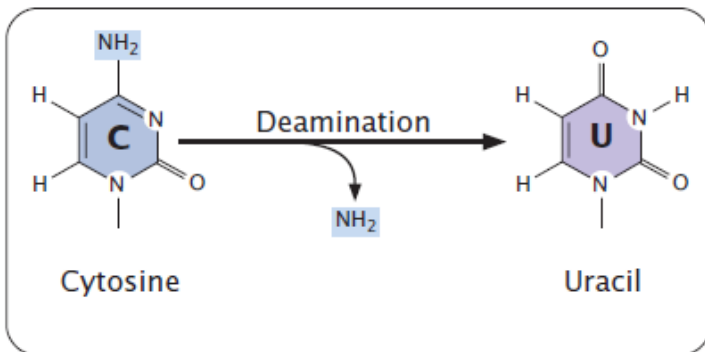
# Endogenous DNA damage

Depurination G:C → A:T



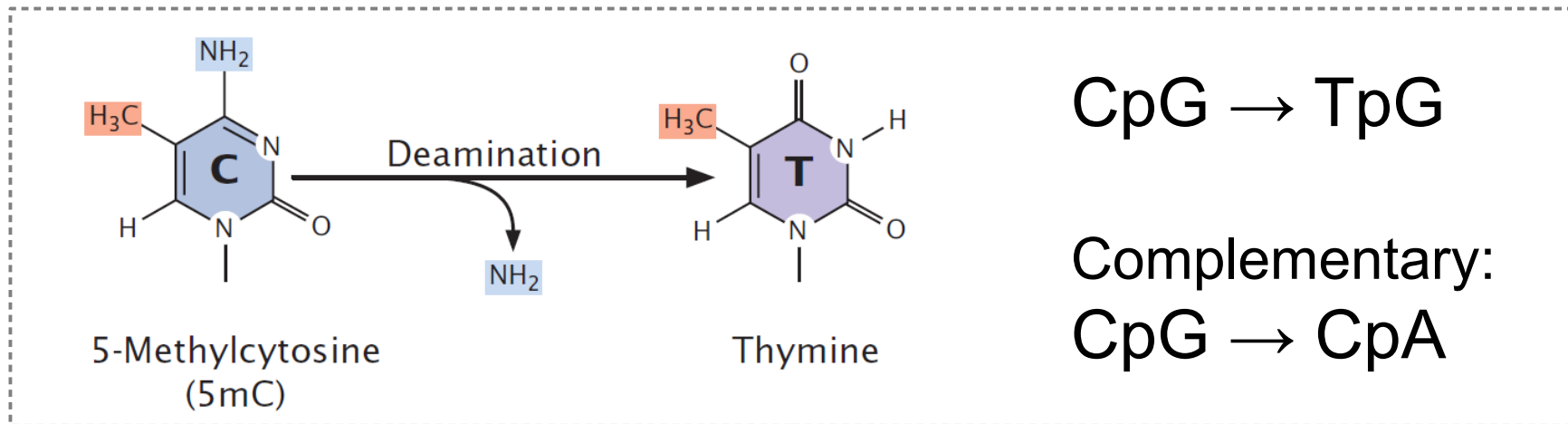
Deamination C:G → U:A → T:A

C:G → 5mC:G → T:G → T:A





# Deamination of 5'-methylcytosine



- Cannot be detected by DNA repair system, because it produces a normal base
- Most mutations occur in male germ cells (M/F = 7:1), because of heavy methylation of sperm DNA and high number of cell divisions
- Example: 46% of point mutations in coagulation factor VIII (*F8*) in unrelated hemophilia A patients
- 23% of all mutations in Human Gene Mutation Database (1998)

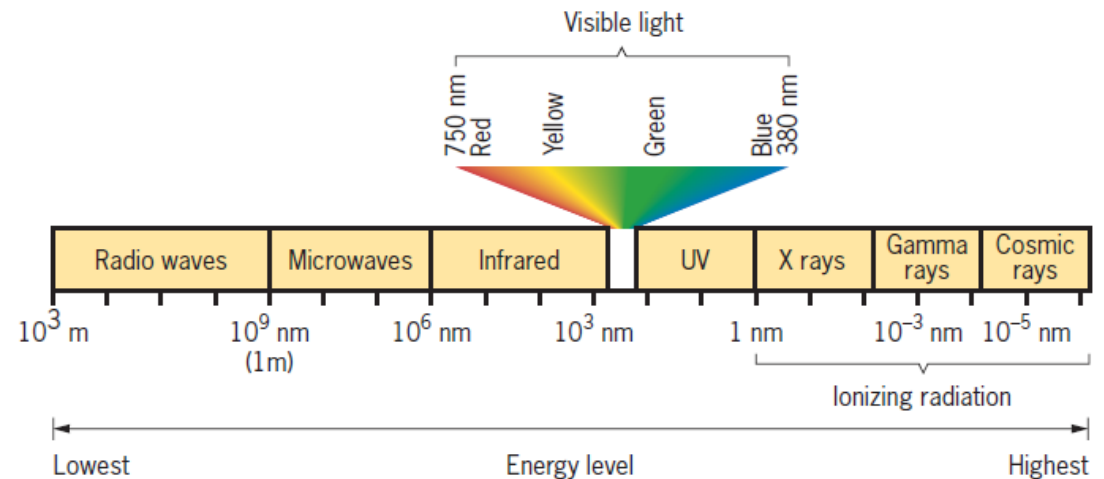
# Exogenous DNA damage

## Chemical mutagens

- **Base analogs:** 5-bromouracil, 2-aminopurine
- **Alkylating agents:** methyl ( $-\text{CH}_3$ ) and ethyl ( $-\text{CH}_3-\text{CH}_2$ ) groups added to nucleotide bases
- **Deamination:** nitrous acid deaminates cytosine, creating uracil
- **Hydroxylamine:** adds a hydroxyl group ( $-\text{OH}$ ) to cytosine
- **Intercalating agents:** proflavin, acridine orange, ethidium bromide, dioxin

## Radiation

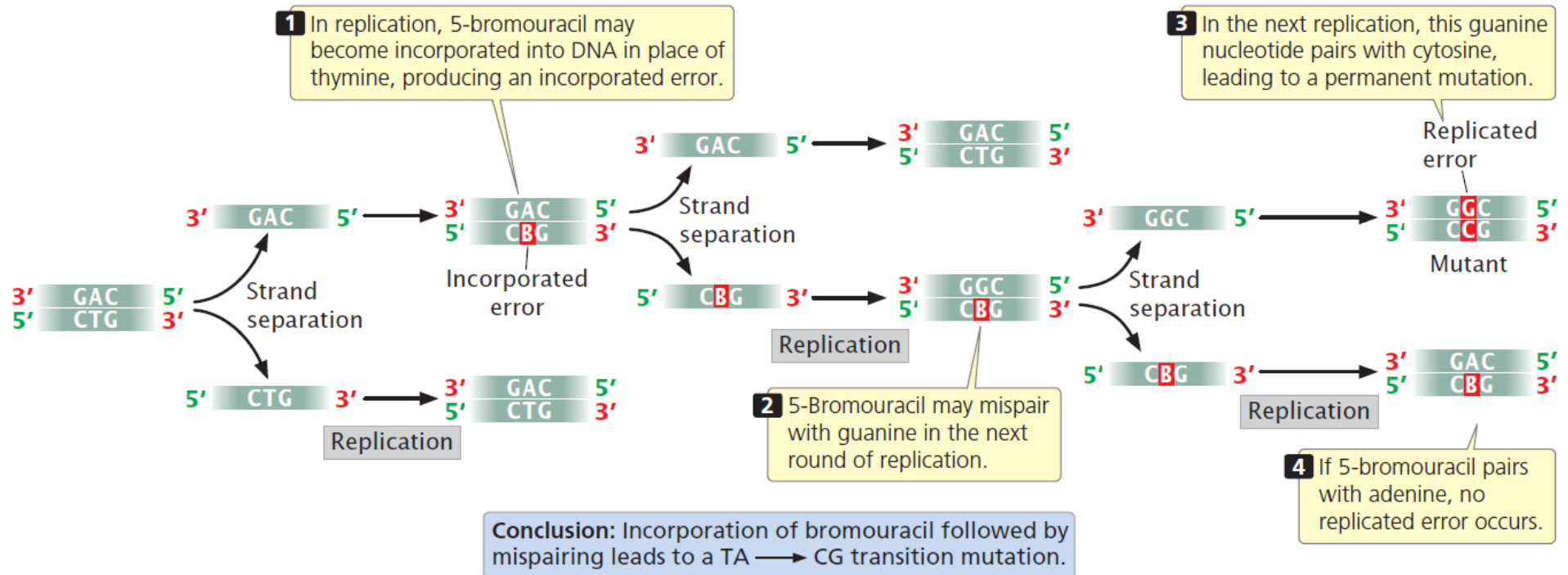
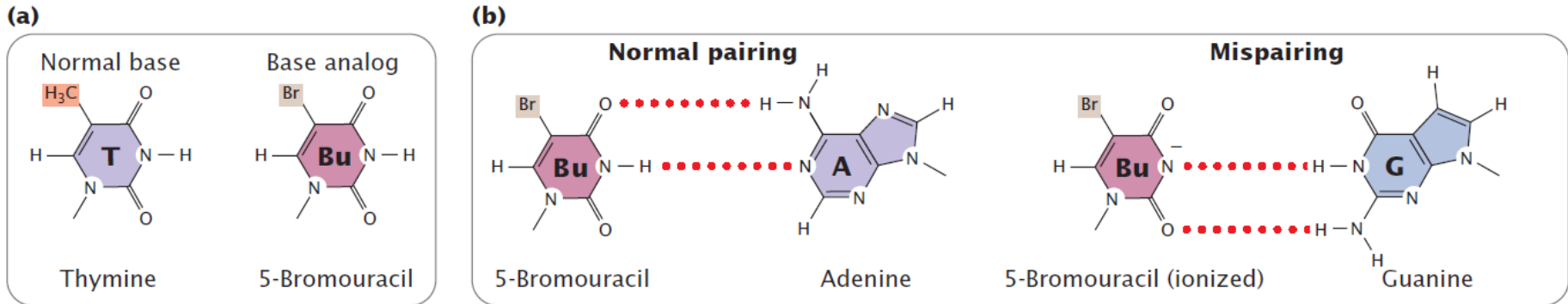
- **Ionizing:**  $\sim 10^{-5} - 1 \text{ nm}$
- **Ultra-violet:**  $\sim 1 - 380 \text{ nm}$





# Exogenous DNA damage

## Chemical mutagens: 5-bromouracil



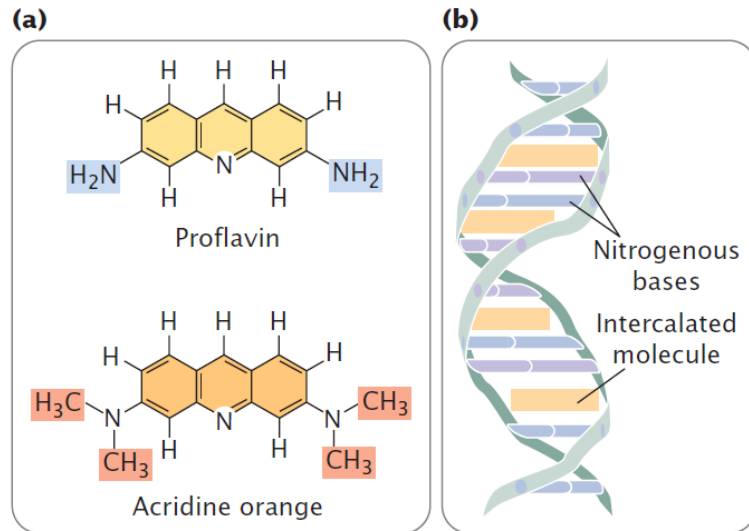


# Exogenous DNA damage

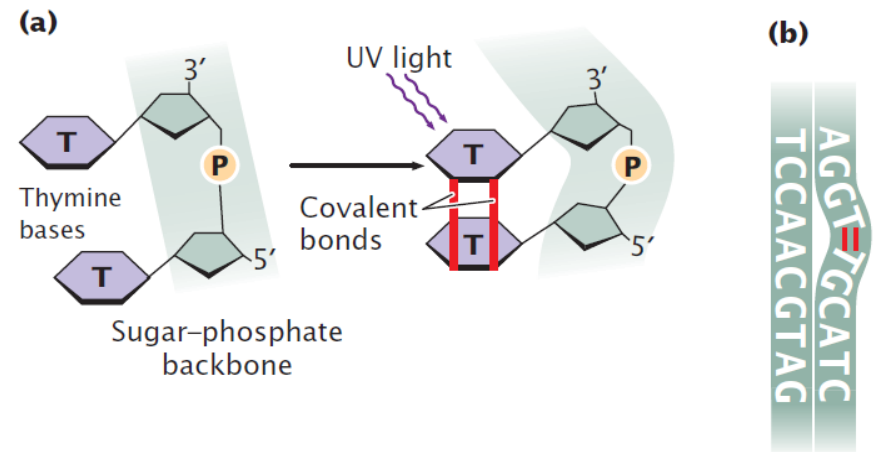
## Chemical mutagens

	Original base	Mutagen	Modified base	Pairing partner	Type of mutation
(a)	 Guanine	EMS Alkylation	 $O^6$ -Ethylguanine	 Thymine	CG → TA
(b)	 Cytosine	Nitrous acid ( $HNO_2$ ) Deamination	 Uracil	 Adenine	CG → TA
(c)	 Cytosine	Hydroxylamine ( $NH_2OH$ ) Hydroxylation	 Hydroxylamino-cytosine	 Adenine	CG → TA

# Exogenous DNA damage



13.20 Intercalating agents such as proflavin and acridine orange insert themselves between adjacent bases in DNA, distorting the three-dimensional structure of the helix and causing single-nucleotide insertions and deletions in replication.



13.21 Pyrimidine dimers result from ultraviolet light.  
(a) Formation of thymine dimer. (b) Distorted DNA.

**Intercalating agents:** distorted DNA  $\Rightarrow$  insertions and deletions

**Ionizing radiation:**

- Free radicals, reactive ions  $\Rightarrow$  altered bases
- Double-strand breaks

**UV light:** Pyrimidine dimers (TpT, CpC, CpT)  $\Rightarrow$  distorted DNA  $\Rightarrow$  replication blocked  $\Rightarrow$  apoptosis or continued error-prone replication



# Endogenous DNA damage

**Depurination:** about 5000 adenine or guanine bases are lost every day from each nucleated human cell by spontaneous hydrolysis of the base-sugar link

**Deamination:** at least 100 cytosines each day in each nucleated human cell are spontaneously deaminated to produce uracil.

**Attack by reactive oxygen species:** highly reactive superoxide anions and related molecules are generated as a by-product of oxidative metabolism in mitochondria. They can also be produced by the impact of ionizing radiation on cellular constituents. These reactive oxygen species attack purine and pyrimidine rings.

**Nonenzymatic methylation:** accidental nonenzymatic DNA methylation by S-adenosyl methionine produces about 300 molecules per cell per day of the cytotoxic base 3-methyl adenine, plus a quantity of the less harmful 7-methyl guanine.

Strachan, Read. *Human Molecular Genetics*, Chapter 13



# Exogenous DNA damage

**Ionizing radiation:** gamma- and X-rays can cause single-strand or double-strand breaks in the sugar-phosphate backbone.

**Ultraviolet radiation:** UV-C rays (with a wavelength of about 260 nm) are especially damaging, but the major source of UV damage in humans is from the UV-B rays (260-315 nm) in sunlight that can penetrate the ozone layer. UV radiation causes cross-linking between adjacent pyrimidines on a DNA strand to form cyclobutane pyrimidine dimers and other abnormal photoproducts.

**Environmental chemicals:** these include hydrocarbons (for example, in cigarette smoke), some plant and microbial products such as the aflatoxins found on moldy peanuts, and chemicals used in cancer chemotherapy. Alkylating agents can transfer a methyl or other alkyl group onto DNA bases and can cause cross-linking between bases within a strand or between different DNA strands.

Strachan, Read. *Human Molecular Genetics*, Chapter 13



# Sources of point mutations

