Making Everything Easier!™

2nd Edition

Biology

FOR ~~DUMMIES~~®

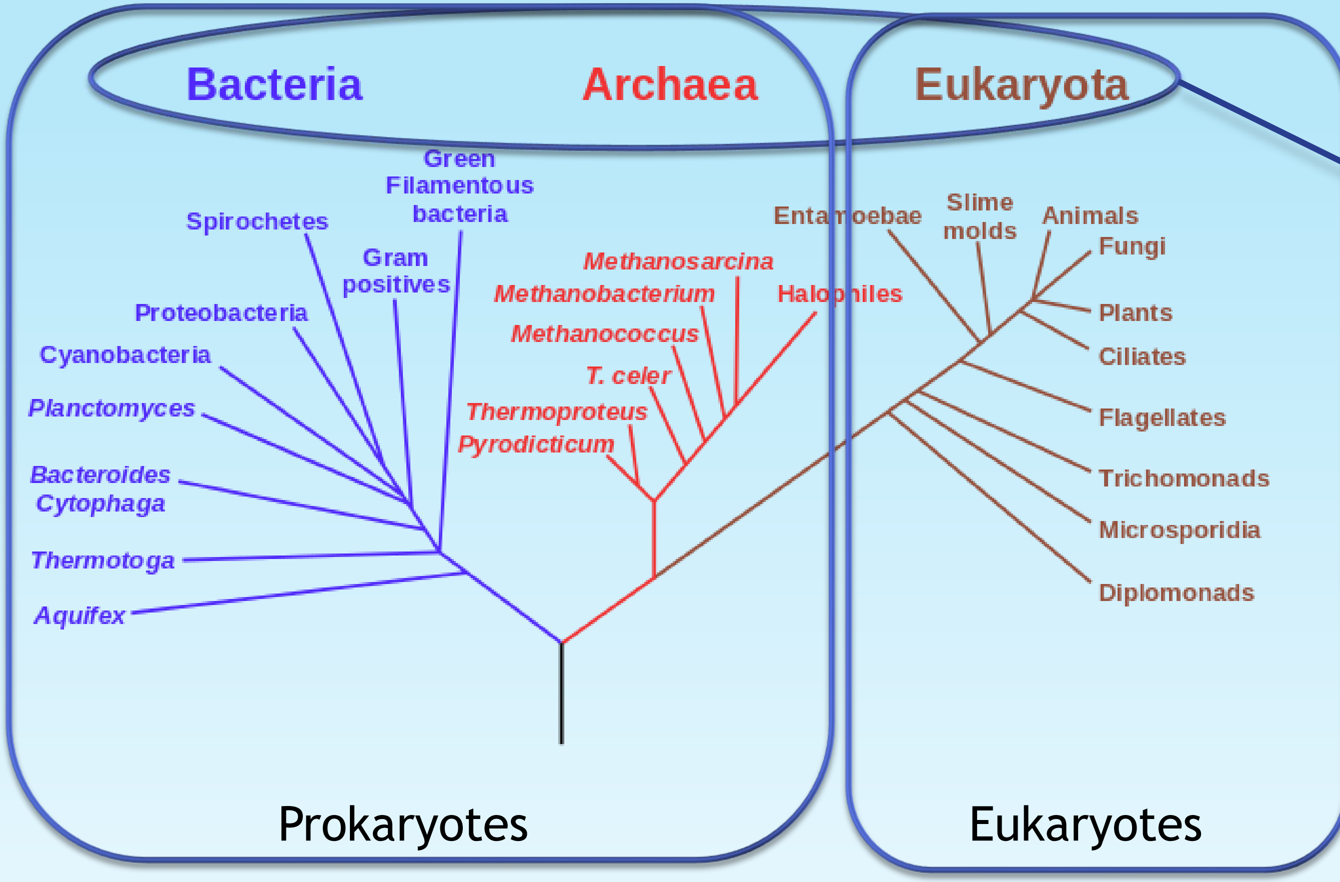
Bioinformaticians

Gregor Gilfillan

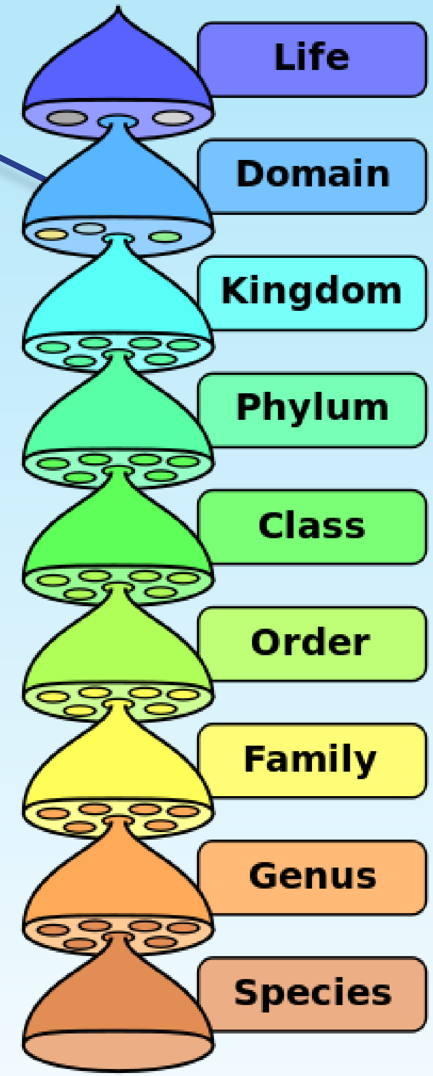
Topics:

1. Tree of life
2. Building blocks of life
3. Structure (and differences) of DNA and RNA
4. DNA makes RNA makes protein
5. Genomes and genomic features
6. Genetics
7. Epigenetics

1. Phylogenetic tree of life



Hierarchical classification of life



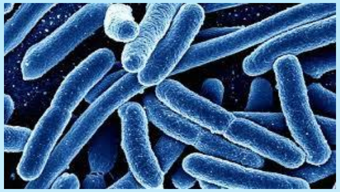
Tree assembled by comparison of rDNA sequences.

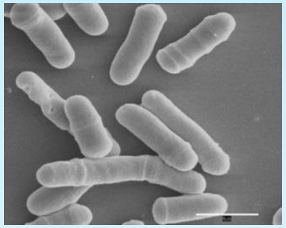
Viruses?

wikipedia

Prokaryotes vs Eukaryotes

Prokaryote





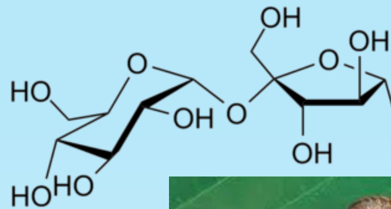
Eukaryote

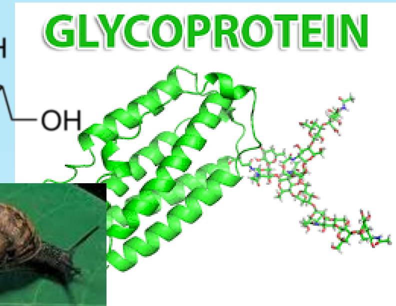
Size
Ribosome type
Chromosomes
Histone proteins
Membranes with steroid content

Defining differences = Eukaryotes:
- Contain DNA in a nucleus
- Contain organelles
(mitochondria, also chloroplasts in some)

2. Building blocks of life

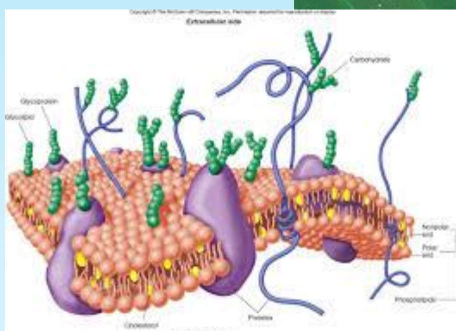
Carbohydrates





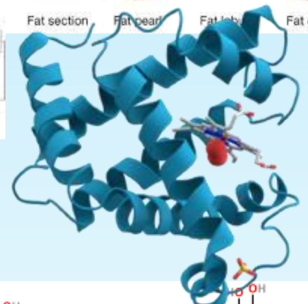
Energy / Structure / Identification / Lubrication / co-metabolites

Lipids (fats)

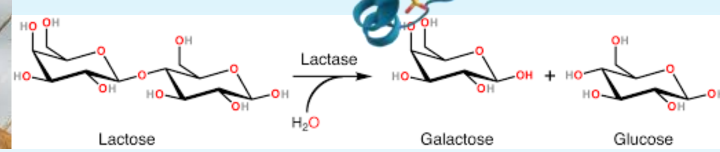


Energy storage / membranes / signaling / insulation

Proteins



Structures / Enzymes / signaling

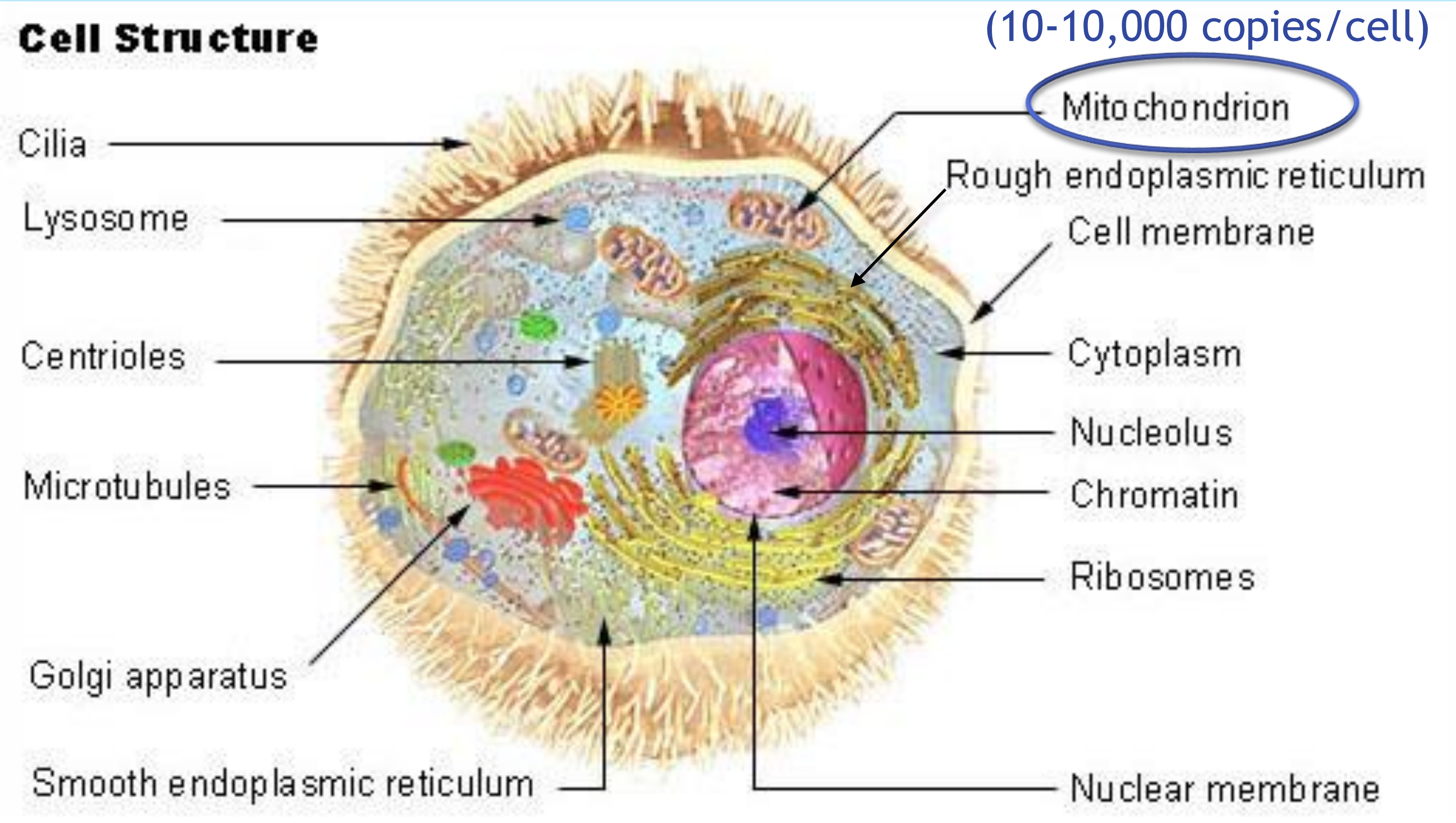


Nucleic acids

formation storage and transfer / ribozymes

Basic unit of life: the cell

Mitochondrial genome
(10-10,000 copies/cell)



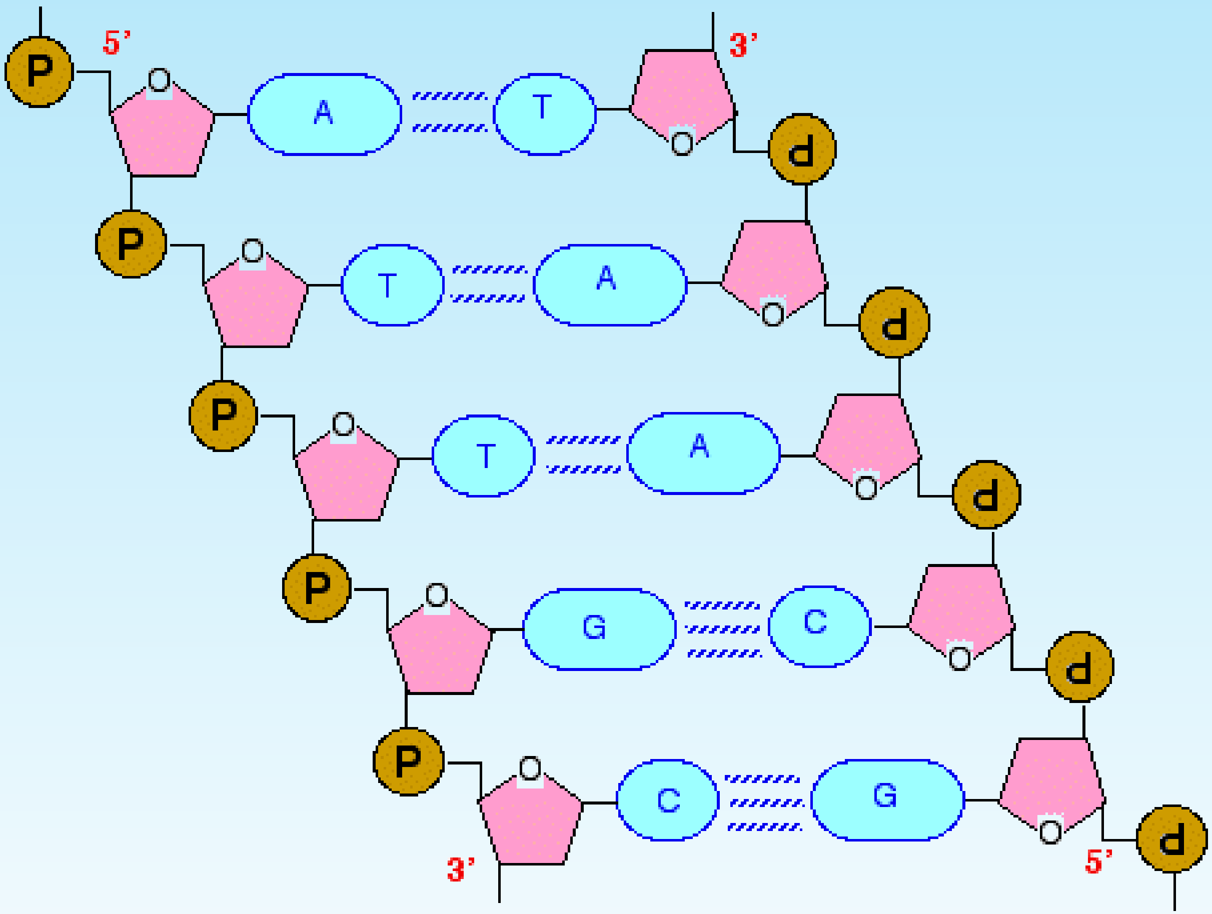
wikipedia

Average human = 5×10^{13} cells
Typical eukaryotic cell ca. 25 µm diameter / Typical bacterium ca. 1 µm
Smallest known cell (mycobacteria) = 0.1 µm
Largest (ostrich egg) = 20 cm

Topics:

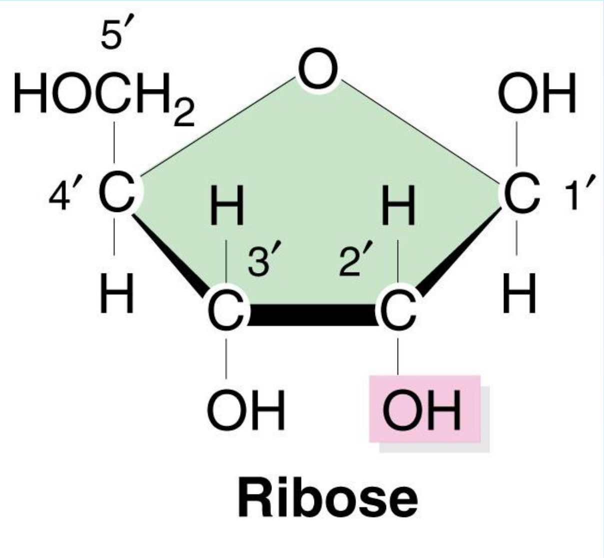
1. Tree of life
2. Building blocks of life
3. **Structure (and differences) of DNA and RNA**
4. DNA makes RNA makes protein
5. Genomes and genomic features
6. Genetics
7. Epigenetics

3. Structure and features of DNA and RNA

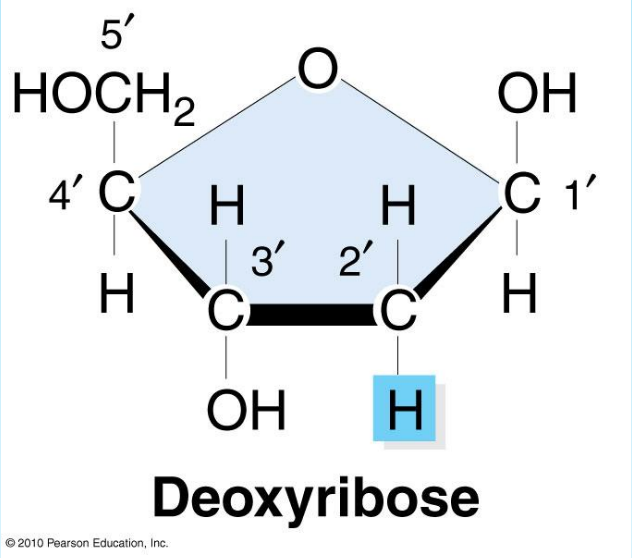


chemguide.co.uk

Sugar-phosphate backbone, bases and nucleotides



Ribose (found in RNA)

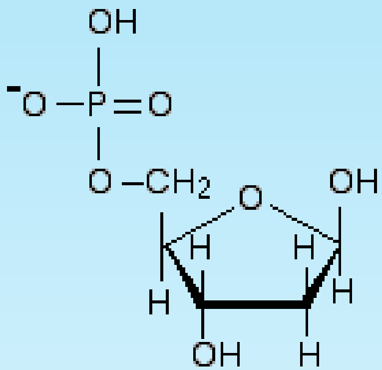


<ins>Deoxy</ins>ribose (found in DNA)

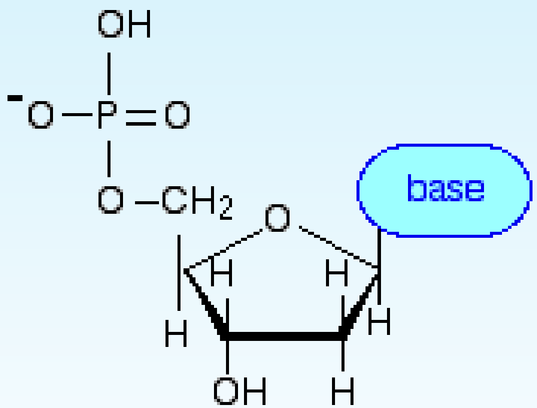
- Planar sugar molecules
- Carbon atoms 1-5 (numbered clockwise, from Oxygen atom)
- Written 1´(you say 1-prime) etc to distinguish these carbon atoms from those in the bases AGCT (which you don´t need to worry about)

(Deoxy)ribose-phosphate

Sugar phosphorylated at C5 = sugar-phosphate backbone of DNA / RNA

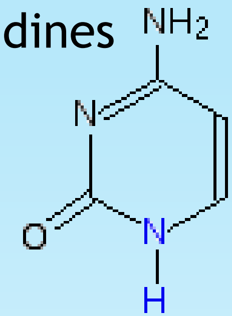


(Deoxy)ribose-phosphate + base = Nucleotide

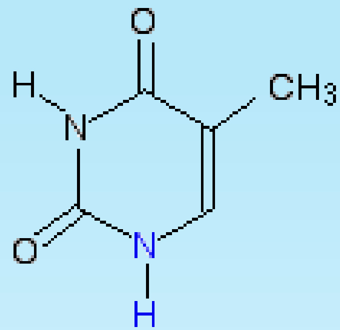


The bases:

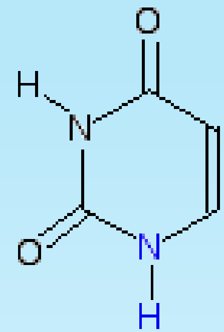
Pyrimidines



cytosine (C)

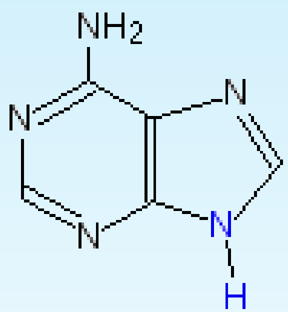


thymine (T)

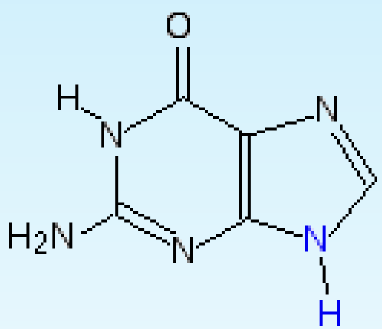


uracil (U)
= replaces T
in RNA

Purines



adenine (A)



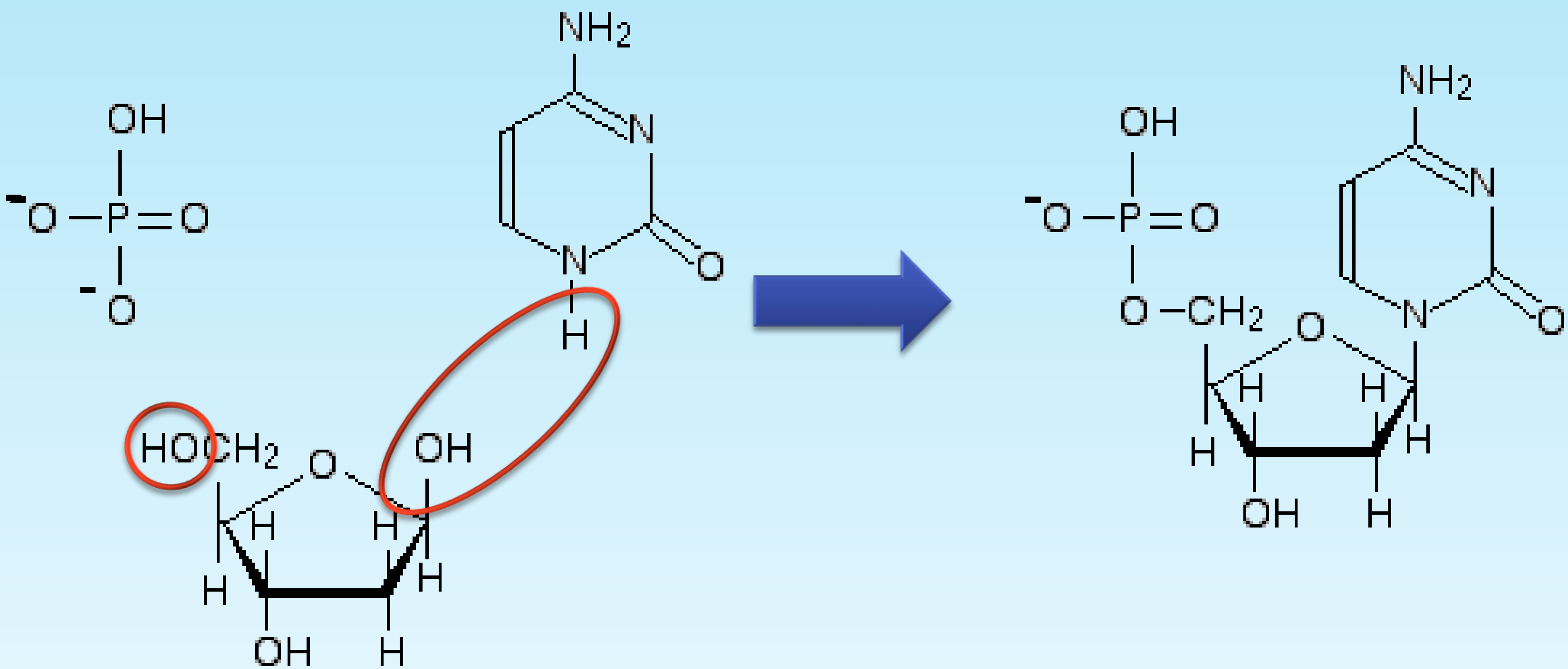
guanine (G)

Called bases because they are alkaline in chemical terms. But this is not of importance to their function in DNA. **Just call them bases!**

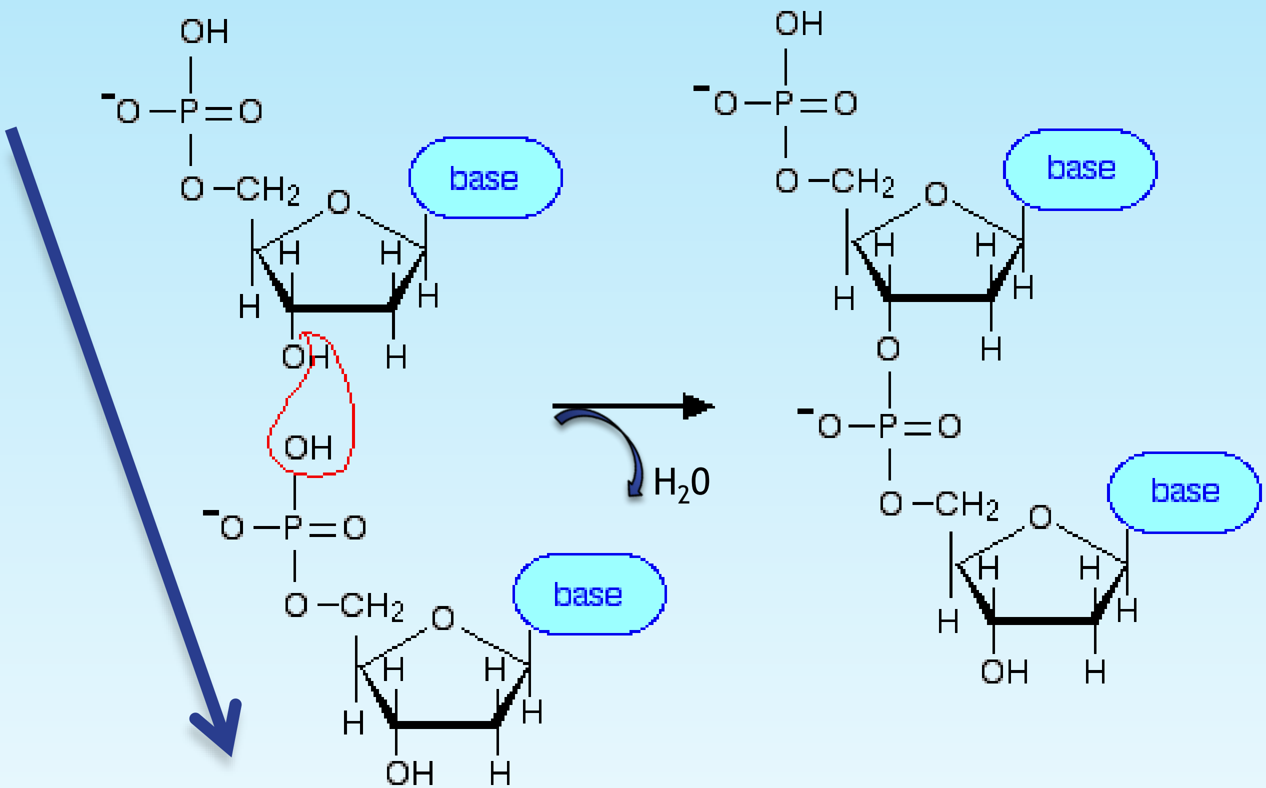
(Nitrogen atoms in blue = attach to C1´ of ribose in a nucleotide)

Phosphoric acid + sugar + base = nucleotide

(The basic repeating unit of nucleic acid)



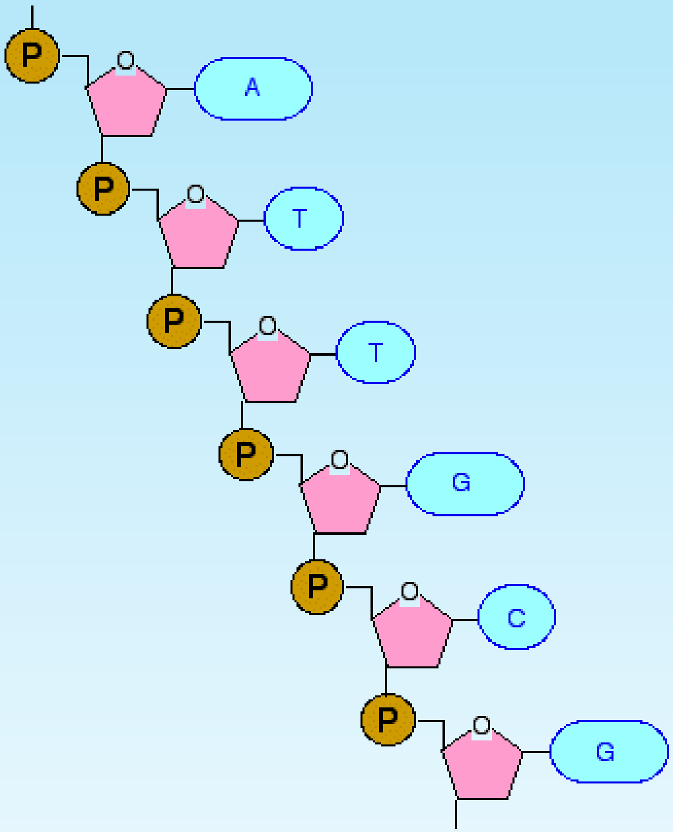
Joining nucleotides into a polynucleotide strand



Nucleophilic attack of 3´OH group onto 5´phosphate group of next nucleotide
=
Phosphodiester bond formation

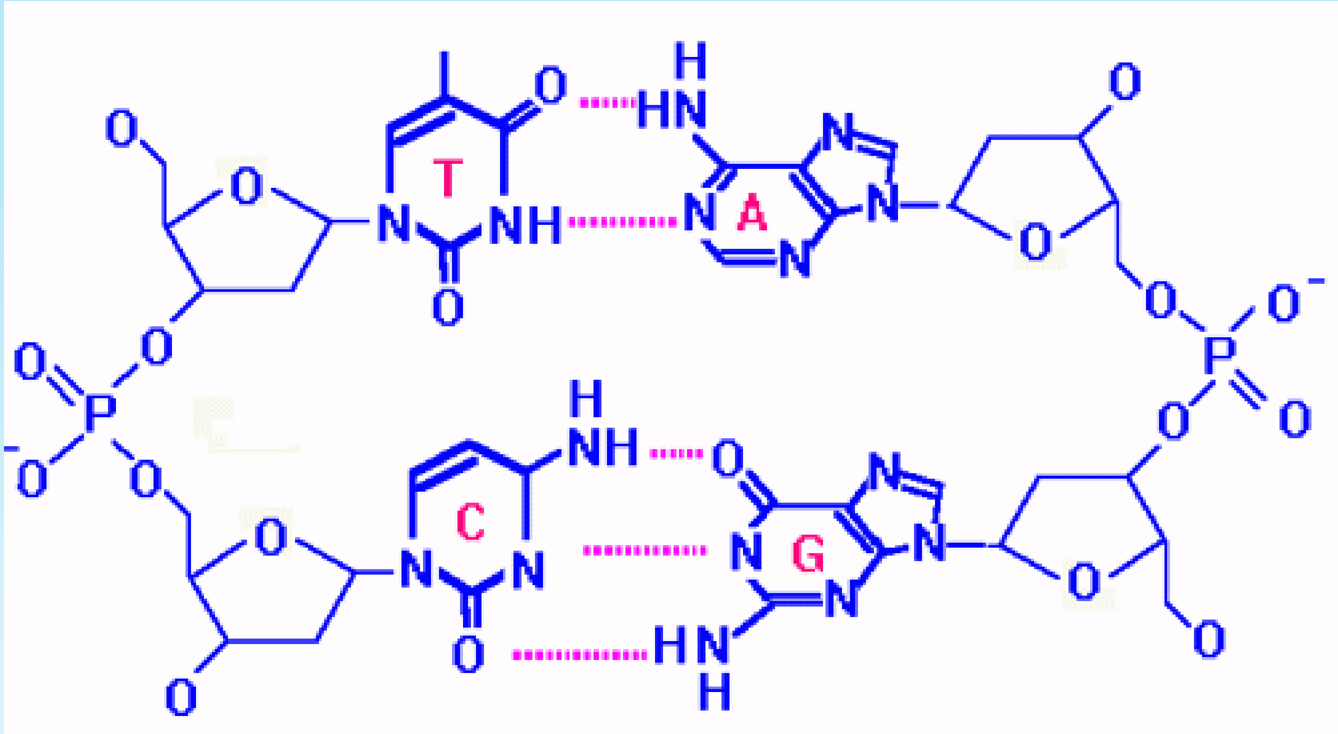
IMPORTANT: Biological reaction can ONLY proceed in 5´to 3´direction!

Simplified view of 1 strand….



…but DNA is double stranded…

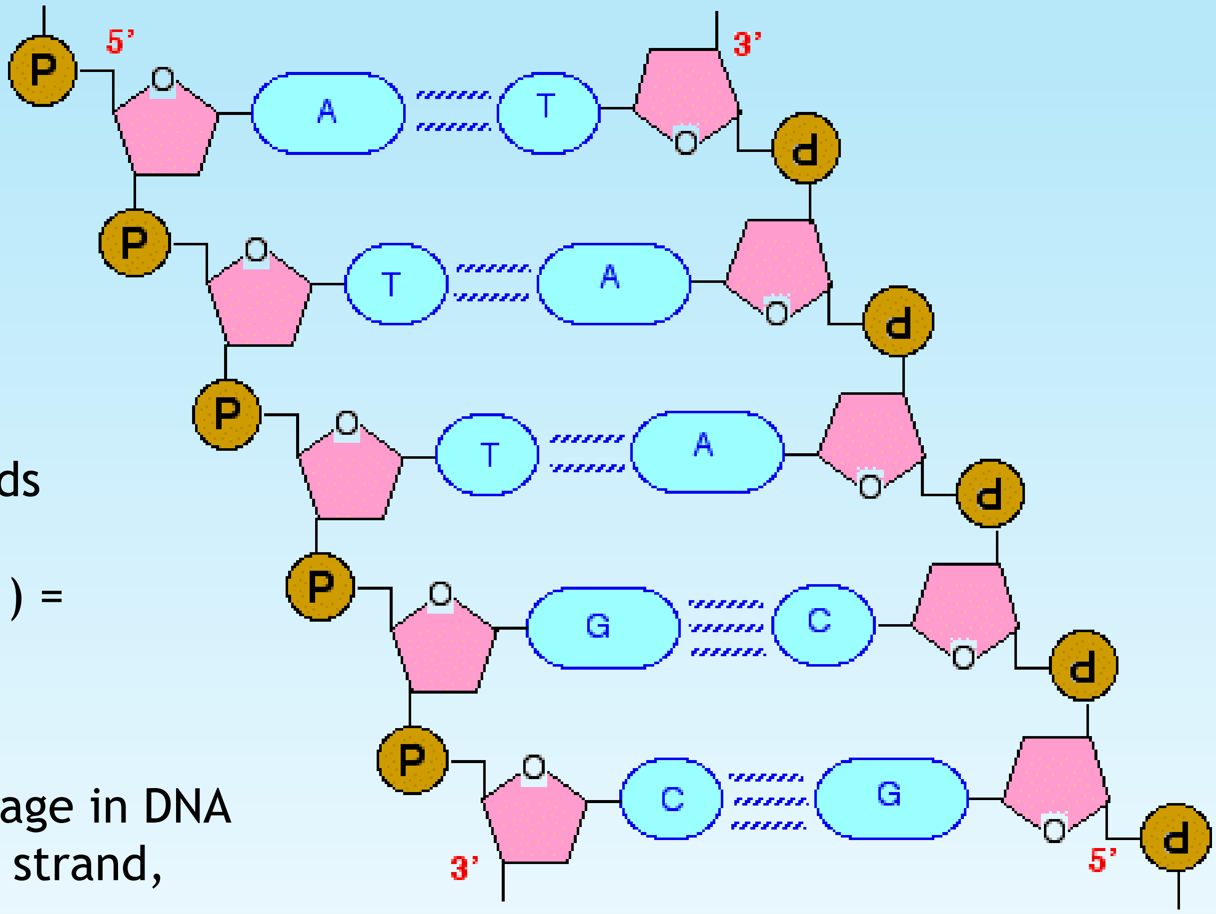
Base-pairing



- Based on hydrogen bonds (weak attractions, rather than strong chemical bonds)
- G always binds to C, A always binds to T
- **G:C = 3 hydrogen bonds / A:T = only 2**
- Always a purine and a pyrimidine in a pair

DNA with high GC content is more difficult to "melt" (separate the strands)

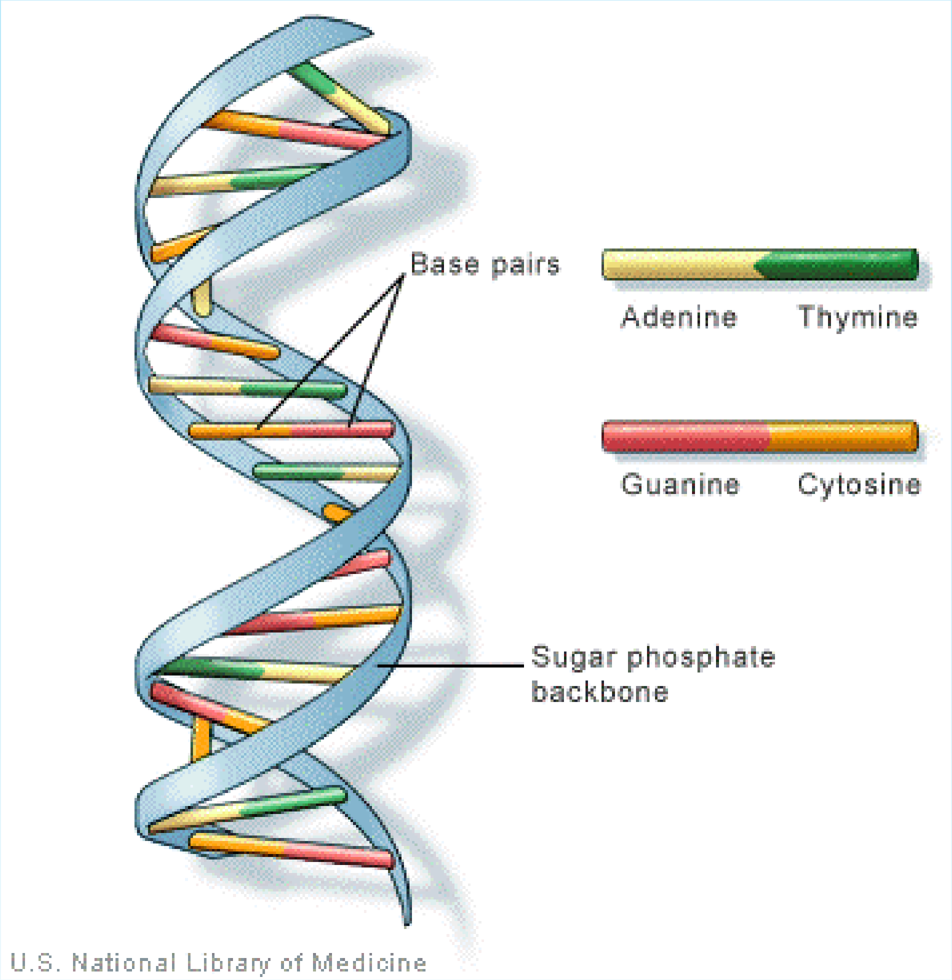
Structure of DNA

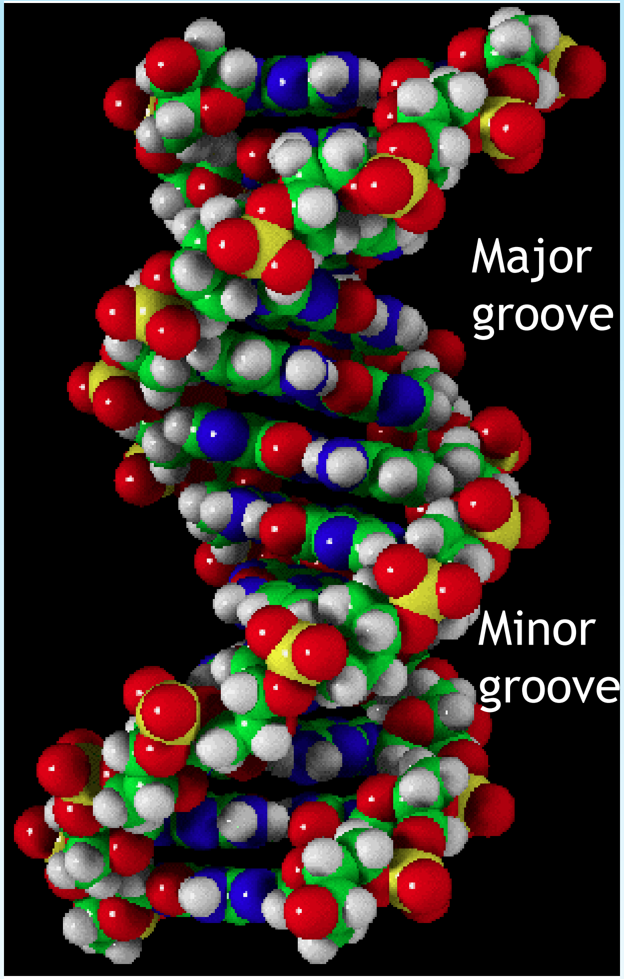


- Notice that strands go opposite directions (5´- 3´) = antiparallel.

- Basis of data storage in DNA = if you have one strand, you have all the information, and can copy it to double-stranded form.

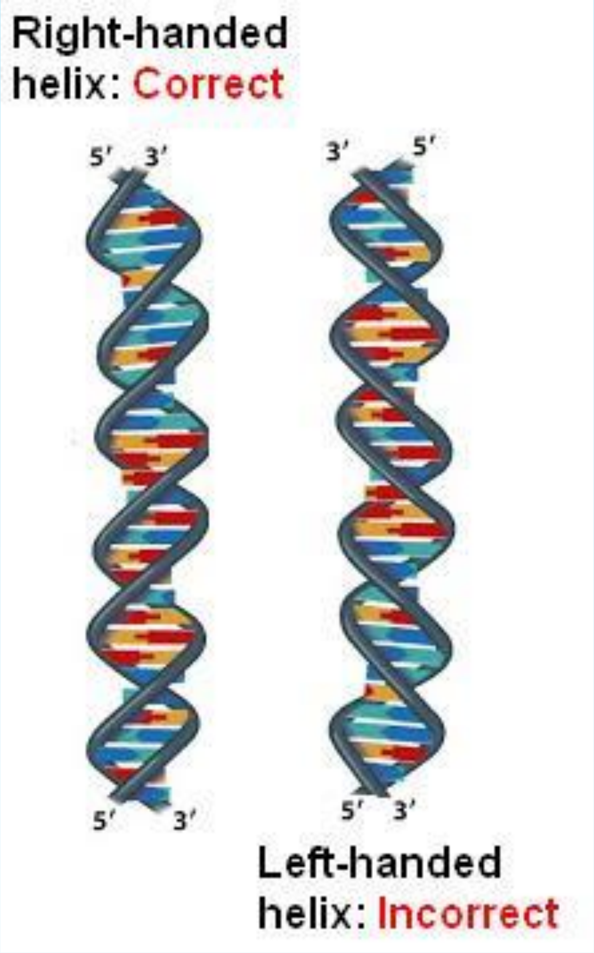
3D Structure of DNA = double helix





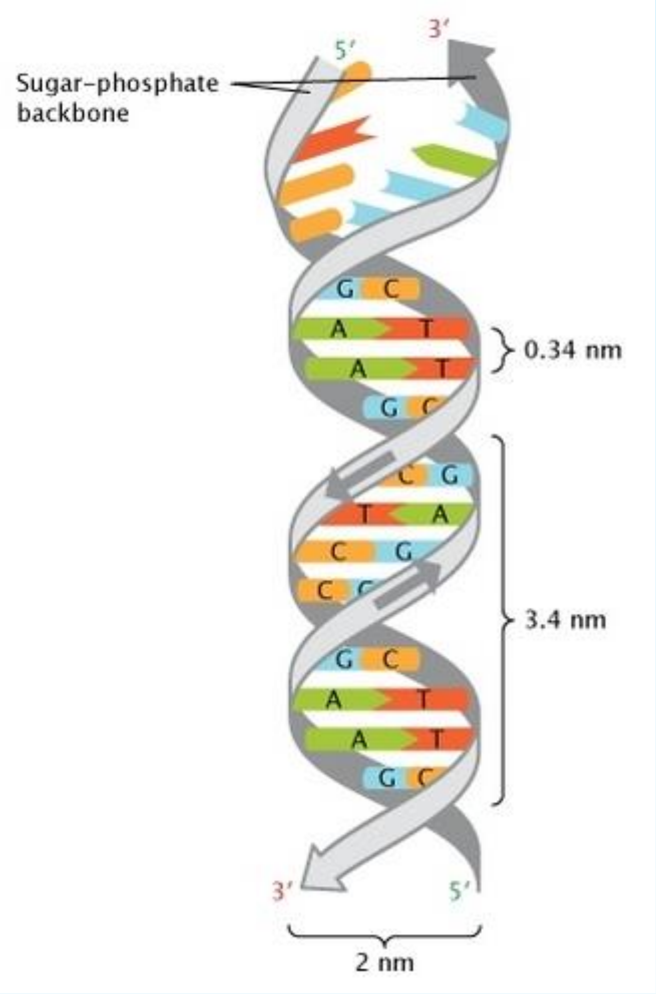
Some final details on DNA 3D structure:

It forms a right-handed helix…



Laurence Moran

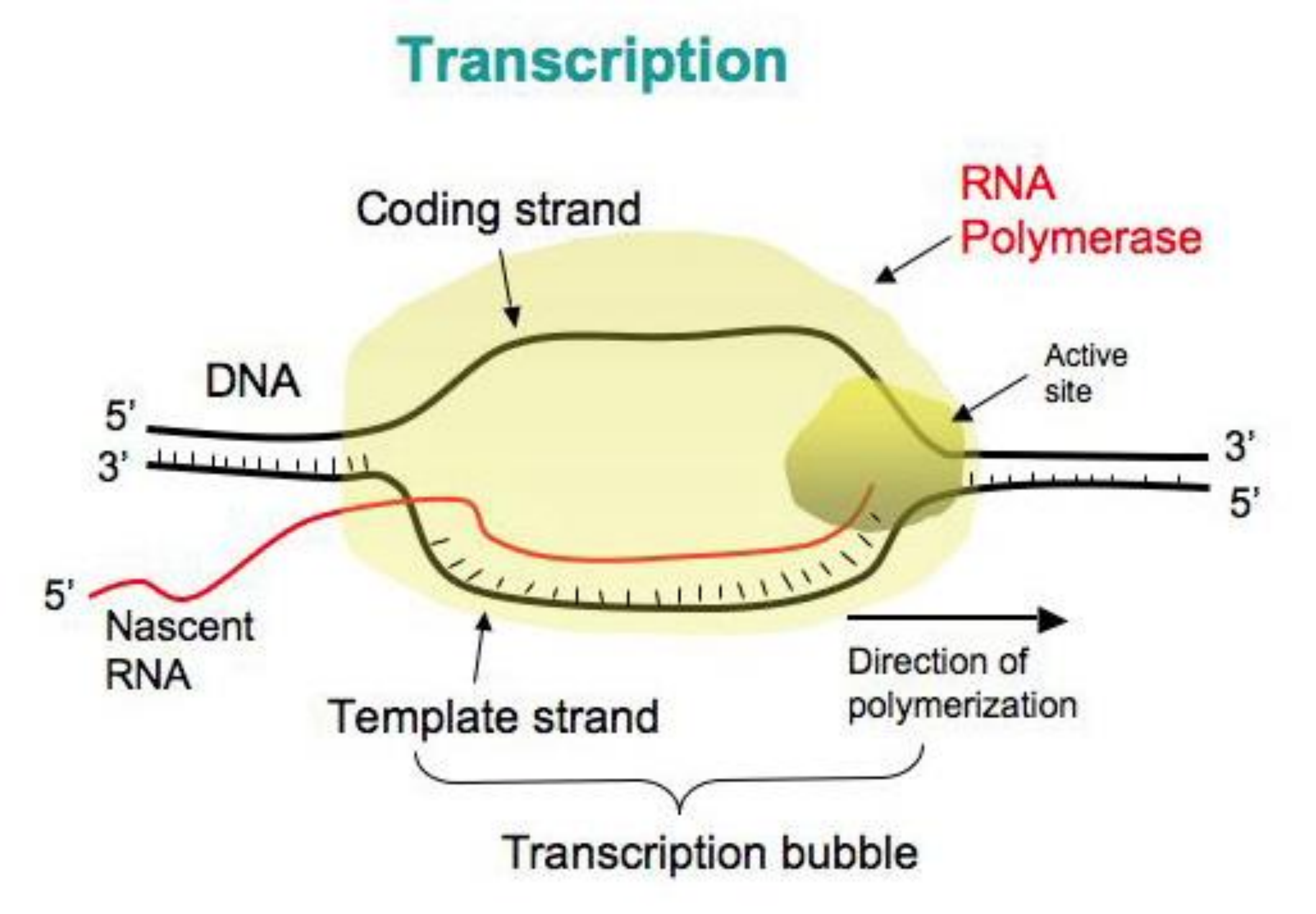
of (fairly) regular dimensions



10-10.5 bases per turn

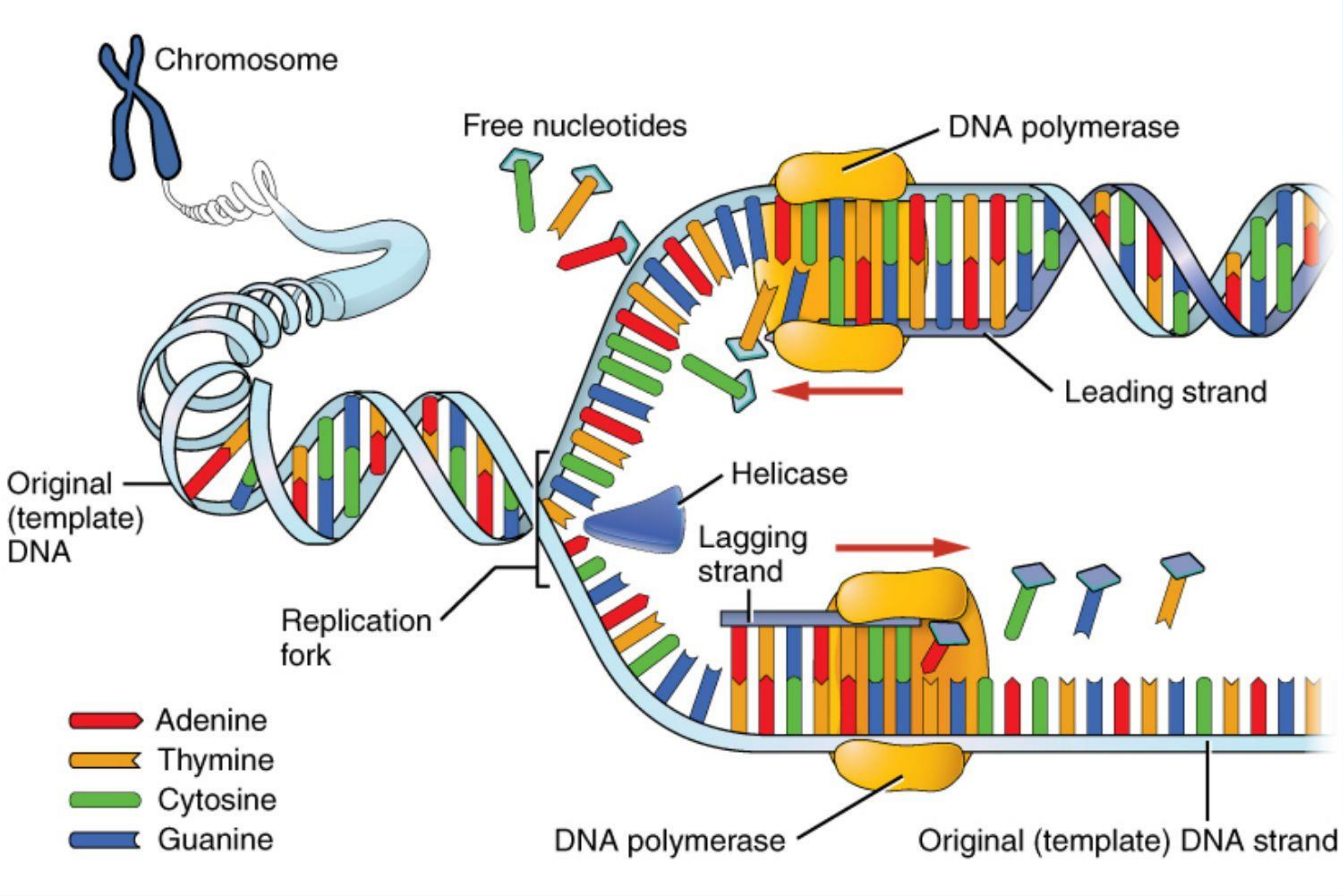
Transcription of RNA

By DNA-dependent RNA polymerase



Replication of DNA

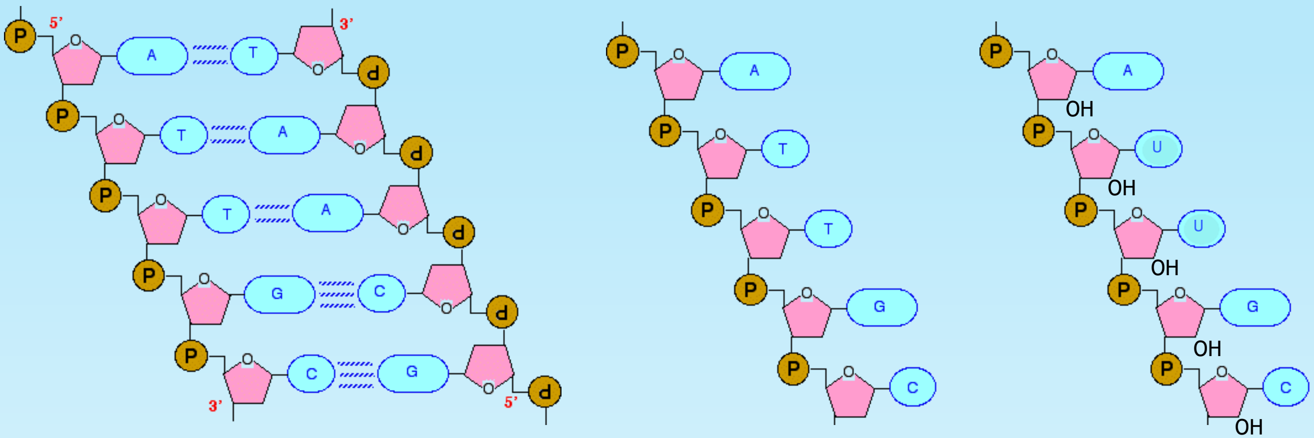
By DNA-dependent DNA polymerases



wikipedia

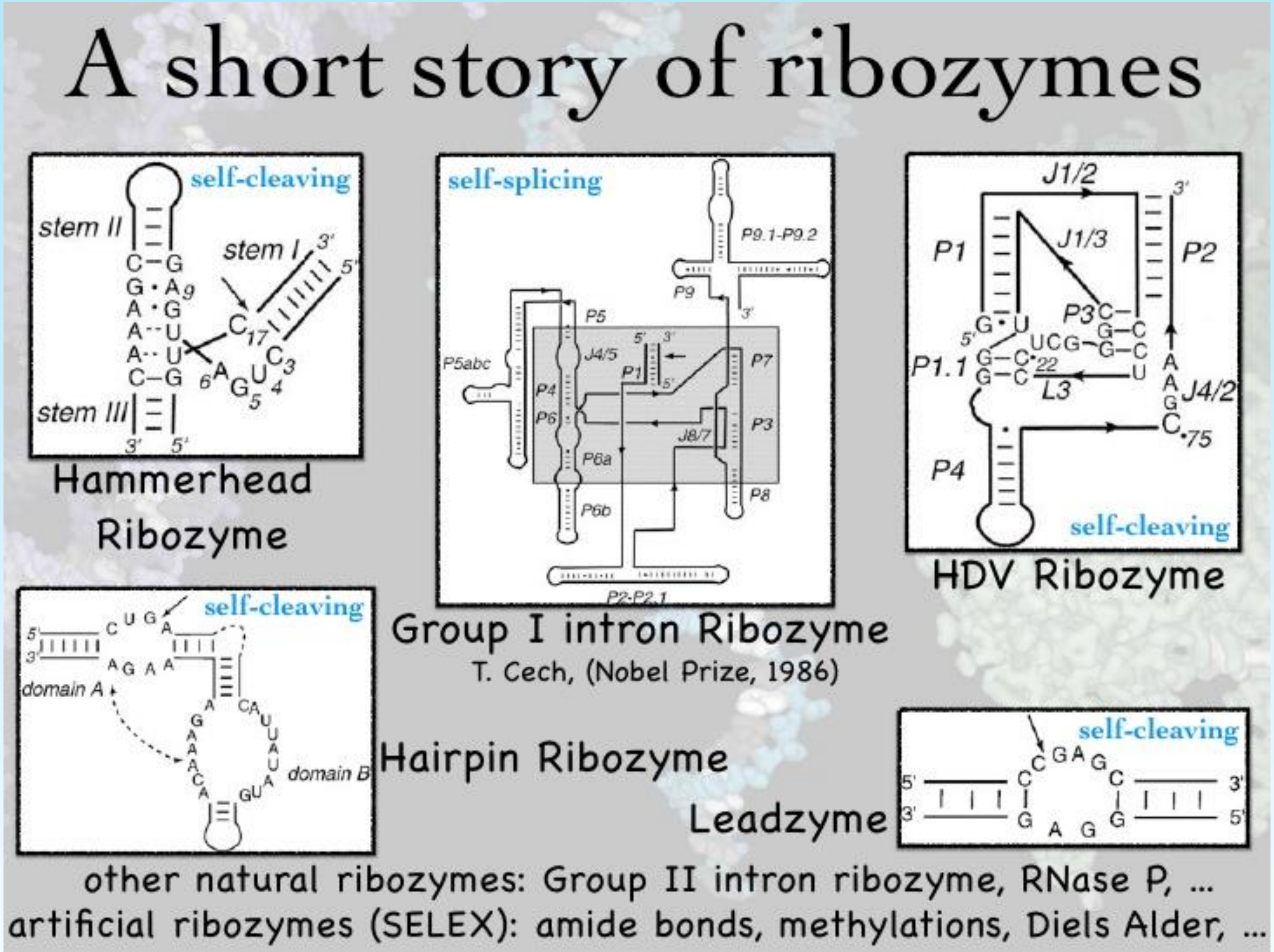
Stability / reactivity of DNA and RNA

dsDNA ssDNA ssRNA



Stability	+++++	+++	+
Reactivity	+	+	+++++

Reactive RNA: ribozymes (i.e. RNA enzymes)



Build some model DNA!

Pause

Video summary of DNA structure

https://www.youtube.com/watch?v=o_-6JXLYS-k

Topics:

1. Tree of life
2. Building blocks of life
3. Structure (and differences) of DNA and RNA
4. **DNA makes RNA makes protein**
5. Genomes and genomic features
6. Genetics
7. Epigenetics

4. DNA makes RNA makes protein

"The central dogma of molecular biology"
(Attributed to Francis Crick, although not the words he used)

replication DNA

transcription

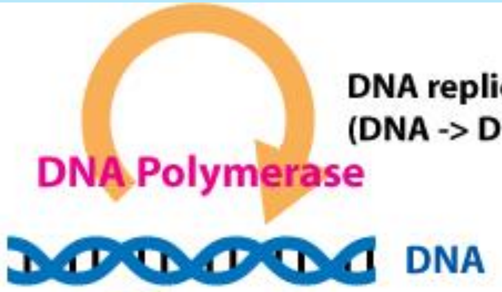
RNA

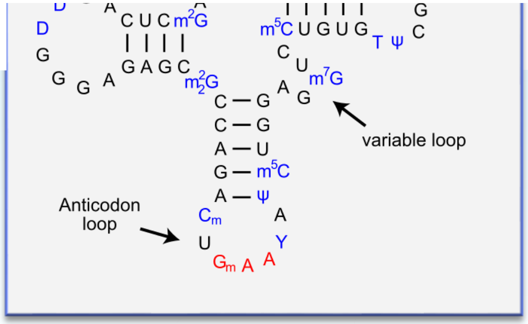
translation

protein

4. DNA makes RNA makes protein (cont.)

The central dogma in some more detail





wikipedia

The genetic code

- Genes use a 3-letter word (called a codon) to direct protein synthesis
- There are 20 amino acids that make up all natural proteins
- 3-letter code with an alphabet of 4 letters (AGCT) = 4^3 = 64 possible codons
- In RNA, letters become AGCU

Amino acids biochemical properties	nonpolar	polar	basic	acidic	Termination: stop codon

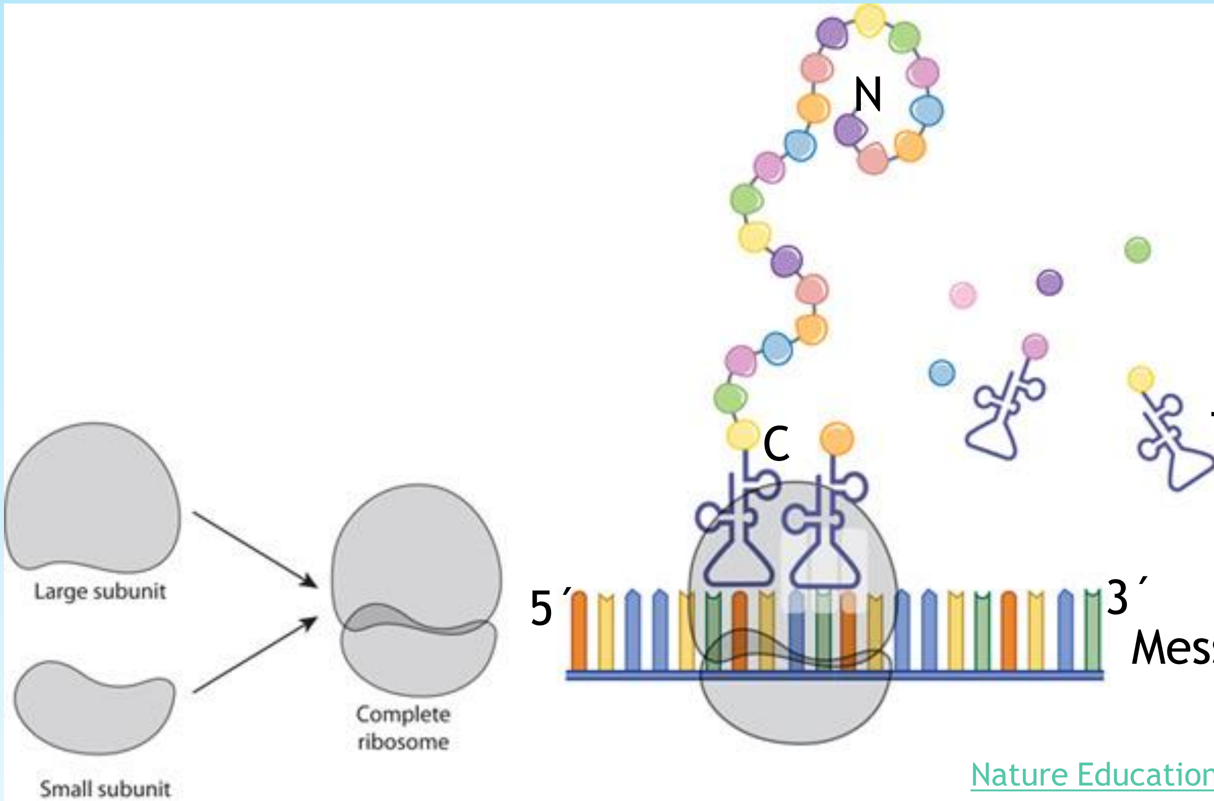
Standard genetic code

1st base	2nd base: U		C		A		G		3rd base
U	UUU	(Phe/F) Phenylalanine	UCU	(Ser/S) Serine	UAU	(Tyr/Y) Tyrosine	UGU	(Cys/C) Cysteine	U
	UUC		UCC		UAC		UGC		C
	UUA	(Leu/L) Leucine	UCA		UAA[B]	Stop (*Ochre*)	UGA[B]	Stop (*Opal*)	A
	UUG		UCG		UAG[B]	Stop (*Amber*)	UGG	(Trp/W) Tryptophan	G
C	CUU		CCU	(Pro/P) Proline	CAU	(His/H) Histidine	CGU	(Arg/R) Arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	(Gln/Q) Glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	(Ile/I) Isoleucine	ACU	(Thr/T) Threonine	AAU	(Asn/N) Asparagine	AGU	(Ser/S) Serine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	(Lys/K) Lysine	AGA	(Arg/R) Arginine	A
	AUG[A]	(Met/M) Methionine	ACG		AAG		AGG		G
G	GUU	(Val/V) Valine	GCU	(Ala/A) Alanine	GAU	(Asp/D) Aspartic acid	GGU	(Gly/G) Glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	(Glu/E) Glutamic acid	GGA		A
	GUG		GCG		GAG		GGG		G

Start codon

Translation

Occurs on the ribosome. All three major types of cell RNA involved.



Transfer rRNAs - tRNA

Messenger RNA - mRNA

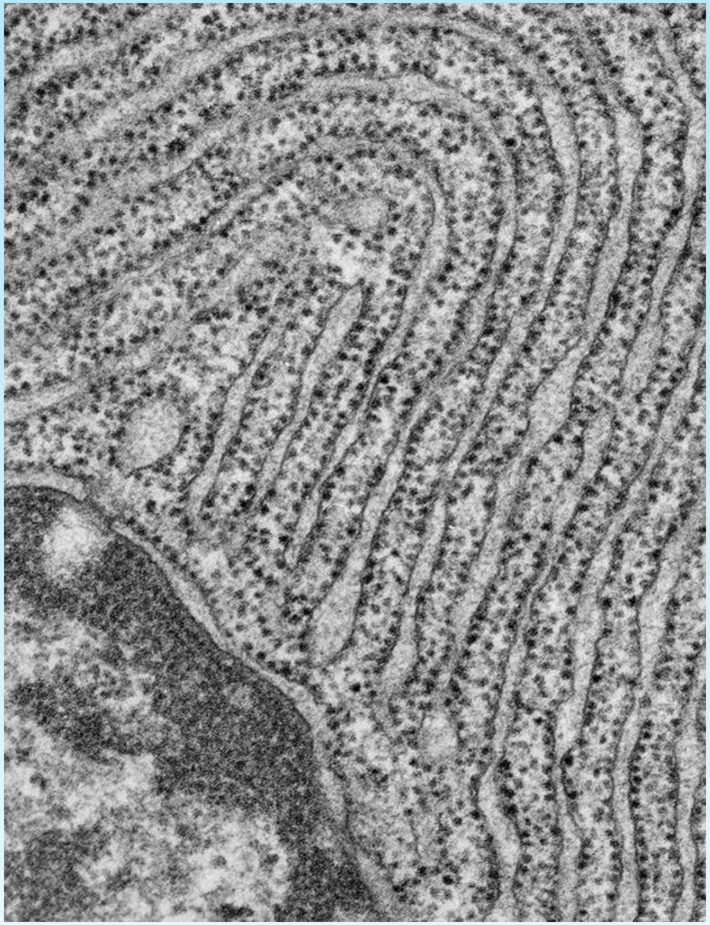
Contains ribosomal RNA
- rRNA (28S, 18S and 5S)

In a typical cell, >90% of all
RNA is rRNA!

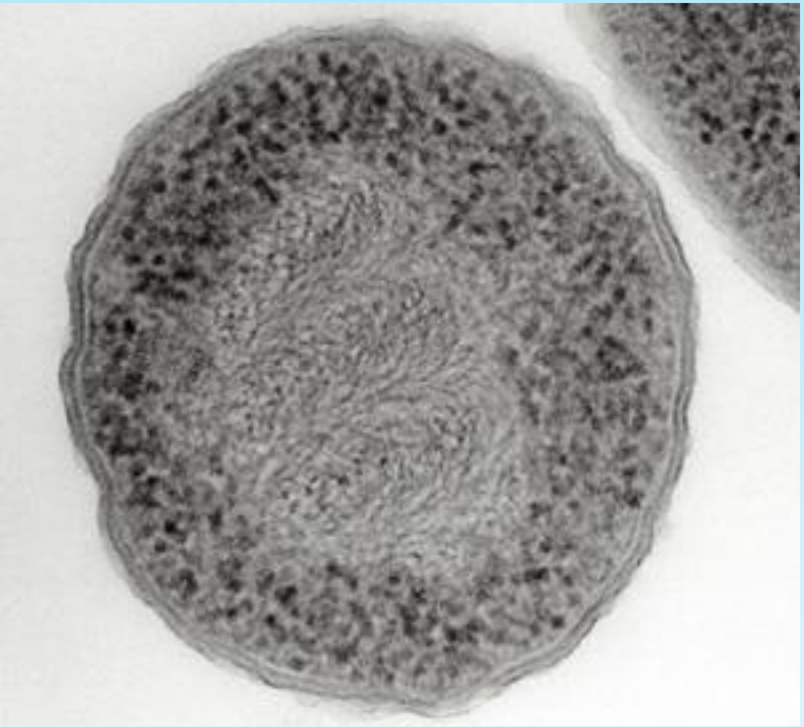
mRNA is read 5´to 3´
Protein is synthesis N-terminus to C-terminus

Ribosomes

The ribosome is a ribozyme!



Electron micrograph of eukaryotic cell showing endoplasmic reticulum studded with ribosomes.

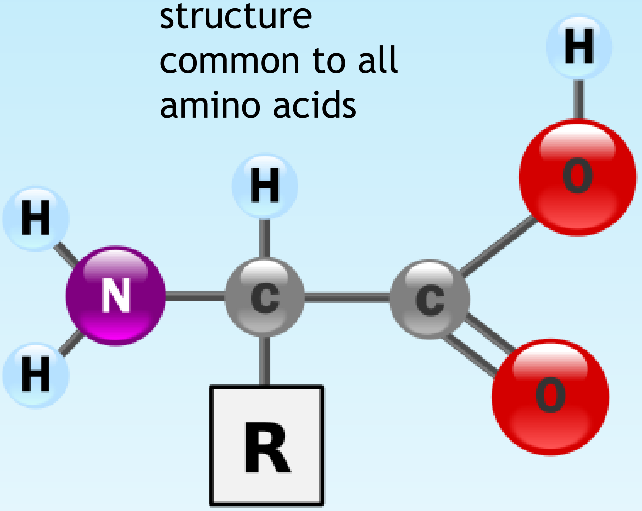


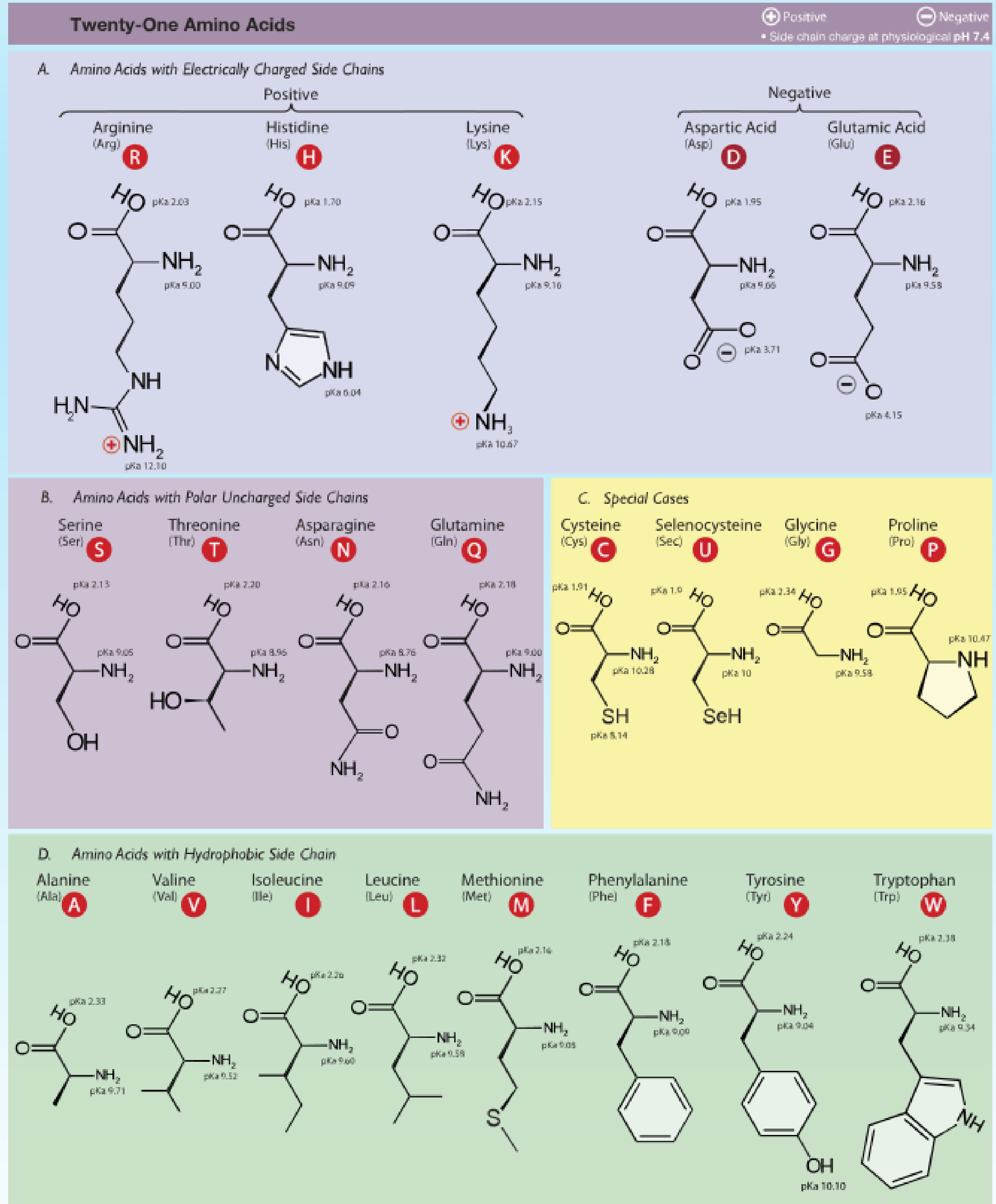
Electron micrograph of prokaryotic cell showing DNA in center and ribosomes on periphery.

Nature education

Amino acids and their properties

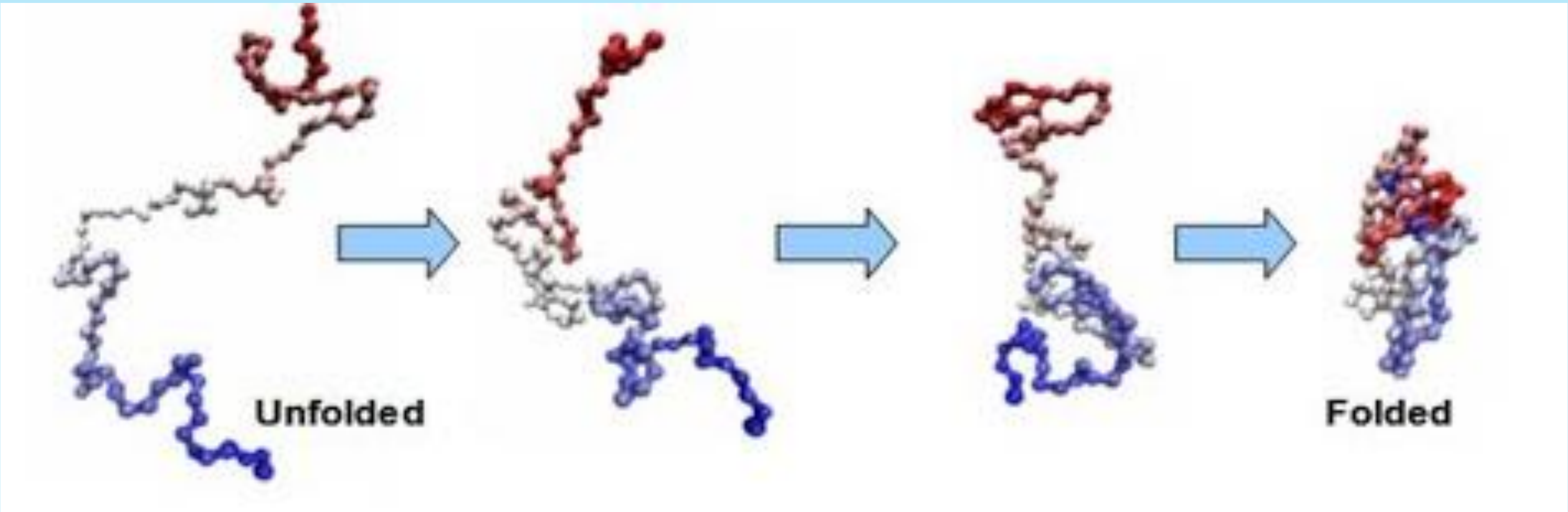
Chemical structure common to all amino acids





wikipedia

Protein folding



Factors driving / affecting protein folding:
- Hydrophobic interactions
- Ionic interactions (+ and – charges)
- Hydrogen bonds
- Metal ion binding
- Chaperone protein interactions
- Effects of salt and pH

Topics:

1. Tree of life
2. Building blocks of life
3. Structure (and differences) of DNA and RNA
4. DNA makes RNA makes protein
5. **Genomes and genomic features**
6. Genetics
7. Epigenetics

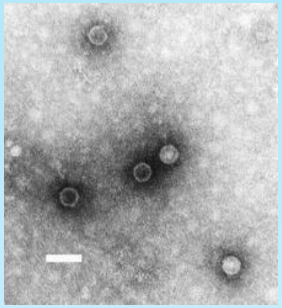
5. Genomes and genomic features

Not all genomes are of double-stranded DNA:

ssRNA viruses (e.g. Poliovirus)

dsRNA viruses (e.g. Rotavirus)

ssDNA (e.g. bacteriophage PhiX 174)



Poliovirus. Scale bar = 50 nM (credit: US EPA)

Viruses tend to have high mutation rates because:

- RNA polymerases do not have proofreading (spell checking) activity
- DNA is more stable in double-stranded form. Rearrangement and mutations more likely in ssDNA.

Genomes of dsDNA:

Ploidy (the number of sets / copies of chromosomes in a cell or organism):

Haploid = 1 copy of dsDNA genome
e.g. bacteria (typically circular genomes)*
e.g. yeast (16 linear chromosomes)
e.g. male bees, ants, wasps

Diploid = 2 copies of genome (one from mum, one from dad)
e.g. us (note that 3.2 Gb genome = haploid size)
e.g. yeast

Triploid = 3 copies
e.g. wheat

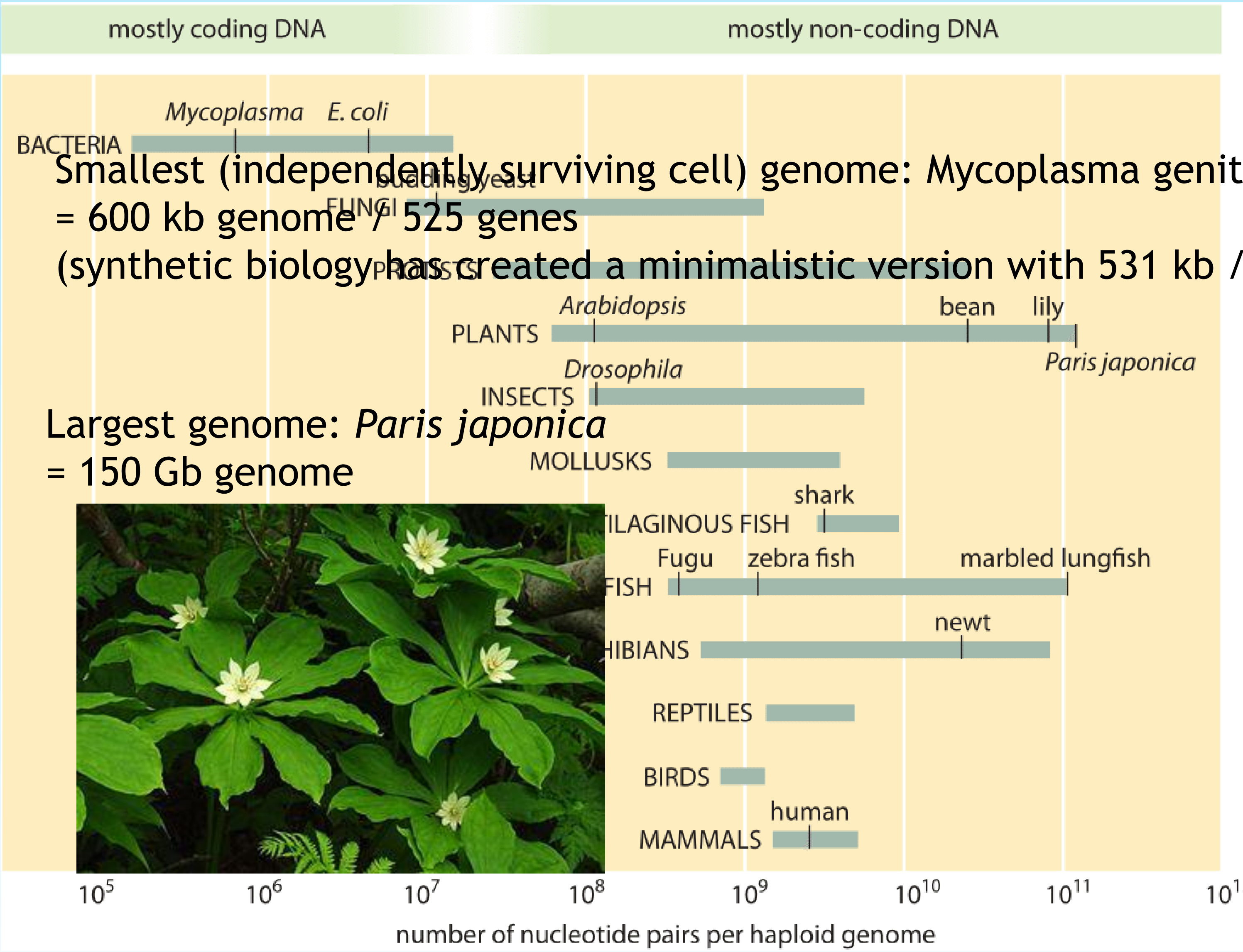
Tetraploid = 4 copies
e.g. potato, Salmon

1048576-ploidy has been measured in silk gland of silkworm!

Polyploidy

* Commonly also contain extra circles of DNA called plasmids, that can be transferred between bacteria (also between different species). Common mechanism for the spread of antibiotic resistance. Some plasmids can be present in multiple copies.

Genome Sizes



Smallest (independently surviving cell) genome: Mycoplasma genitalium
= 600 kb genome / 525 genes
(synthetic biology has created a minimalistic version with 531 kb / 473 genes)

Largest genome: *Paris japonica*
= 150 Gb genome

http://book.bionumbers.org/how-big-are-genomes/)

Genome Sizes

For organisms that have not had their genome sequenced / have poor assemblies / lots of repetitive DNA, genome size estimated using “C number” = amount of DNA (in pg) in a haploid genome (eg a sperm cell).

1 pg = 978 Mb (ca. 1 Gb)

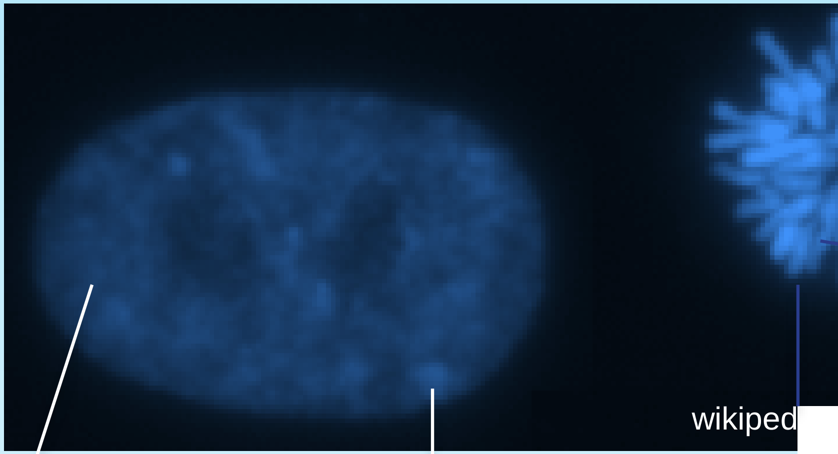
Database of eukaryotic genome sizes:
http://www.genomesize.com/

NOTE:
Large Genome ≠ many genes!

(generally means lots of repetitive and non-coding DNA)

Genomic structure and features

Cells stained with DNA-dye DAPI



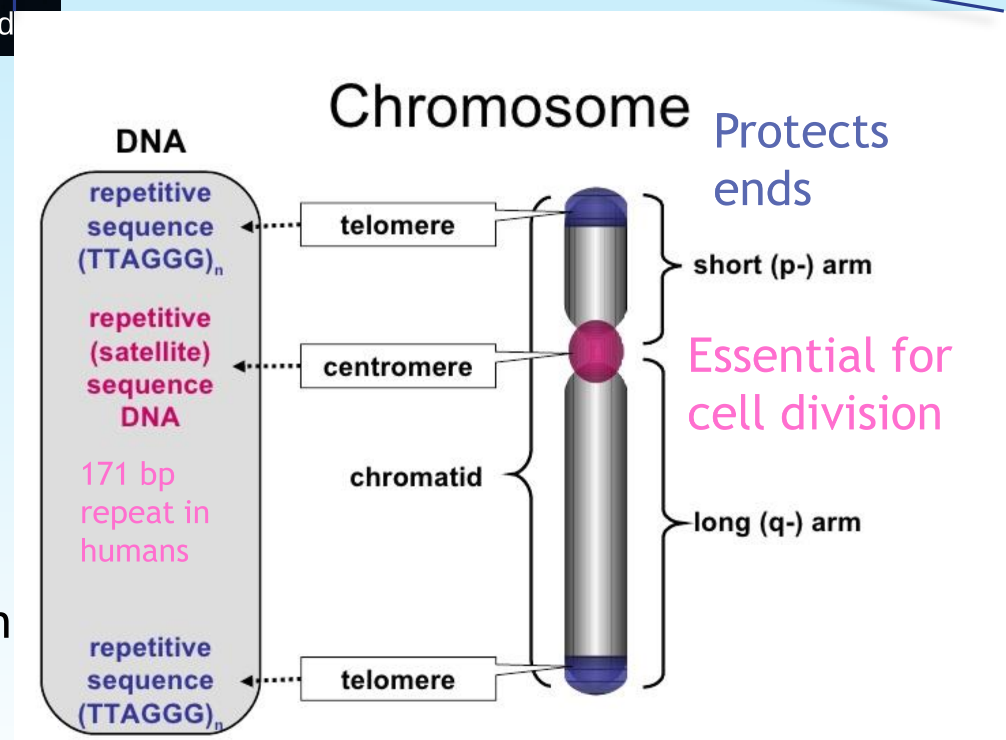
Nucleus undergoing mitosis (cell division)

Interphase nucleus
(how nuclei look most of the time)

Euchromatin
(open, expressed)

Heterochromatin
(dense, repressed)

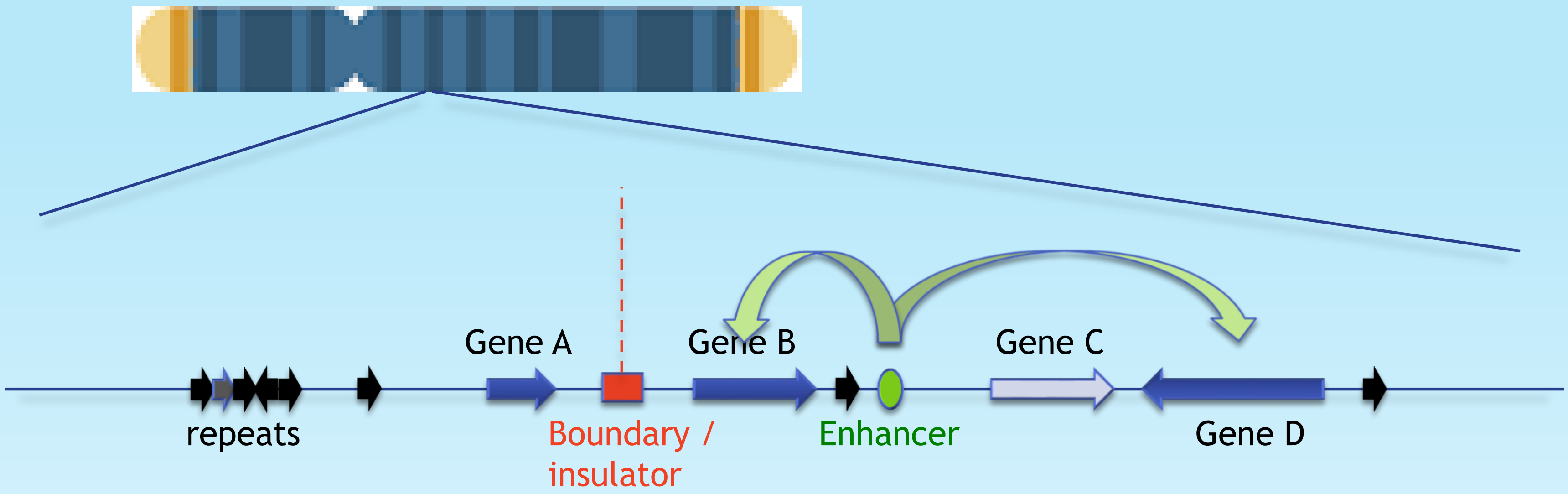
DNA in cells = protein bound = Chromatin



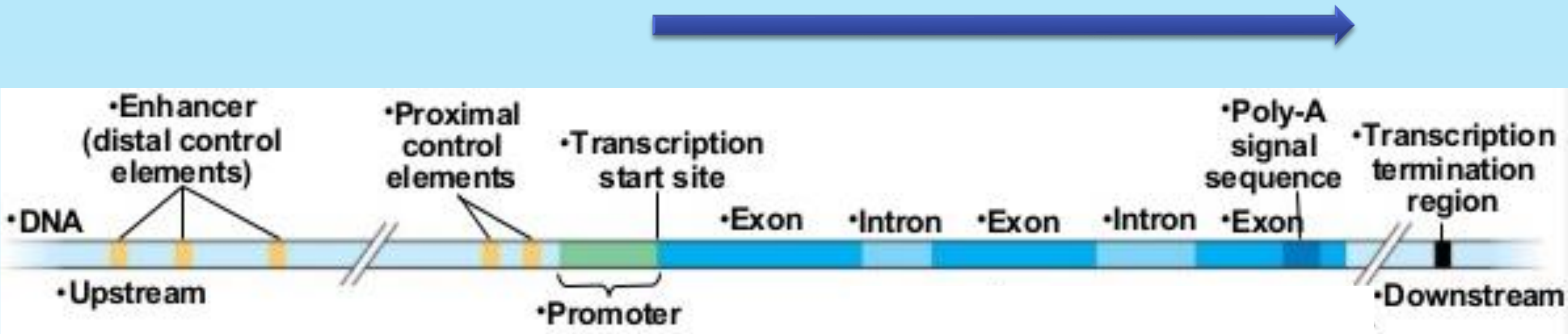
Protects ends

Essential for cell division

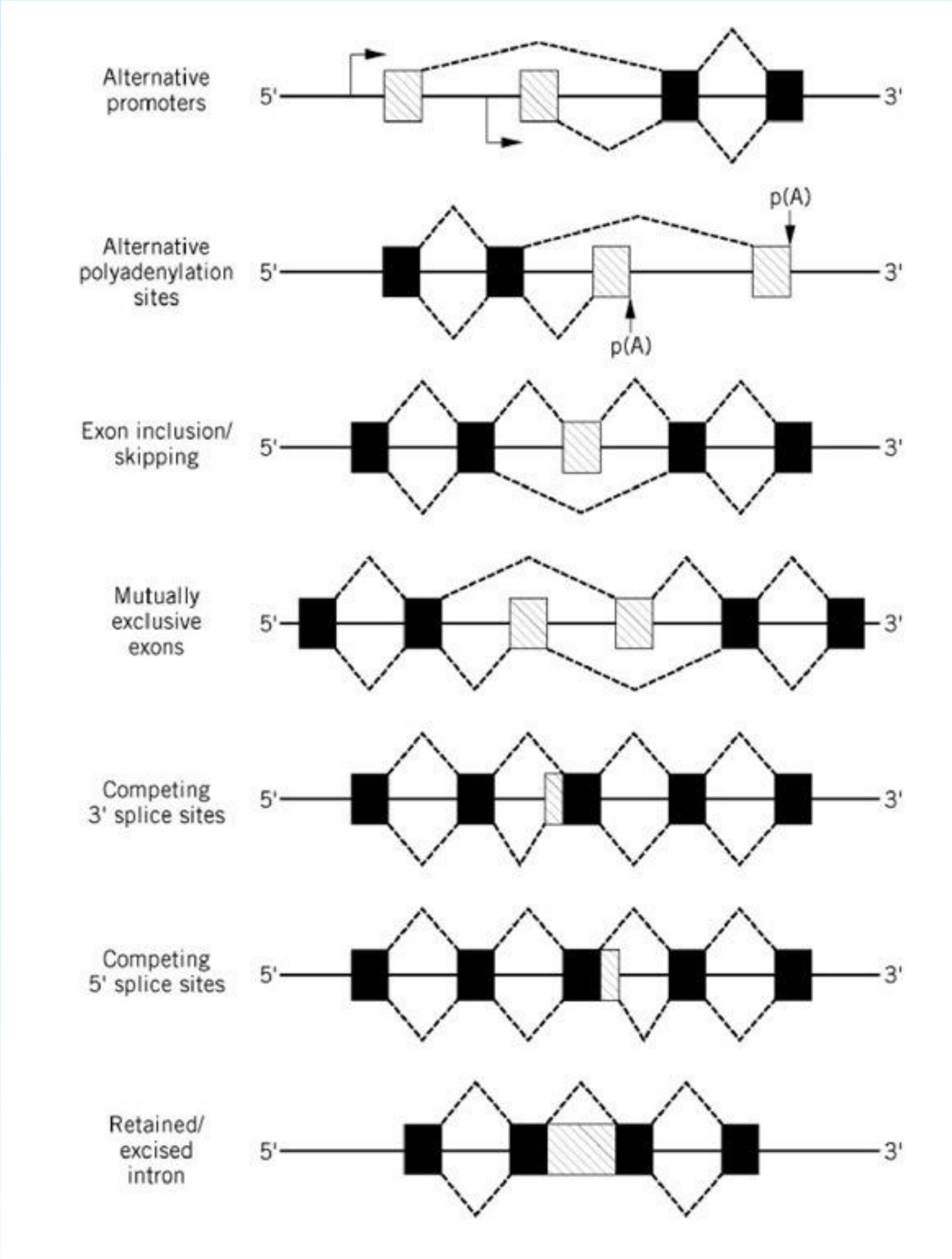
Genomic structure and features



Genes in more detail

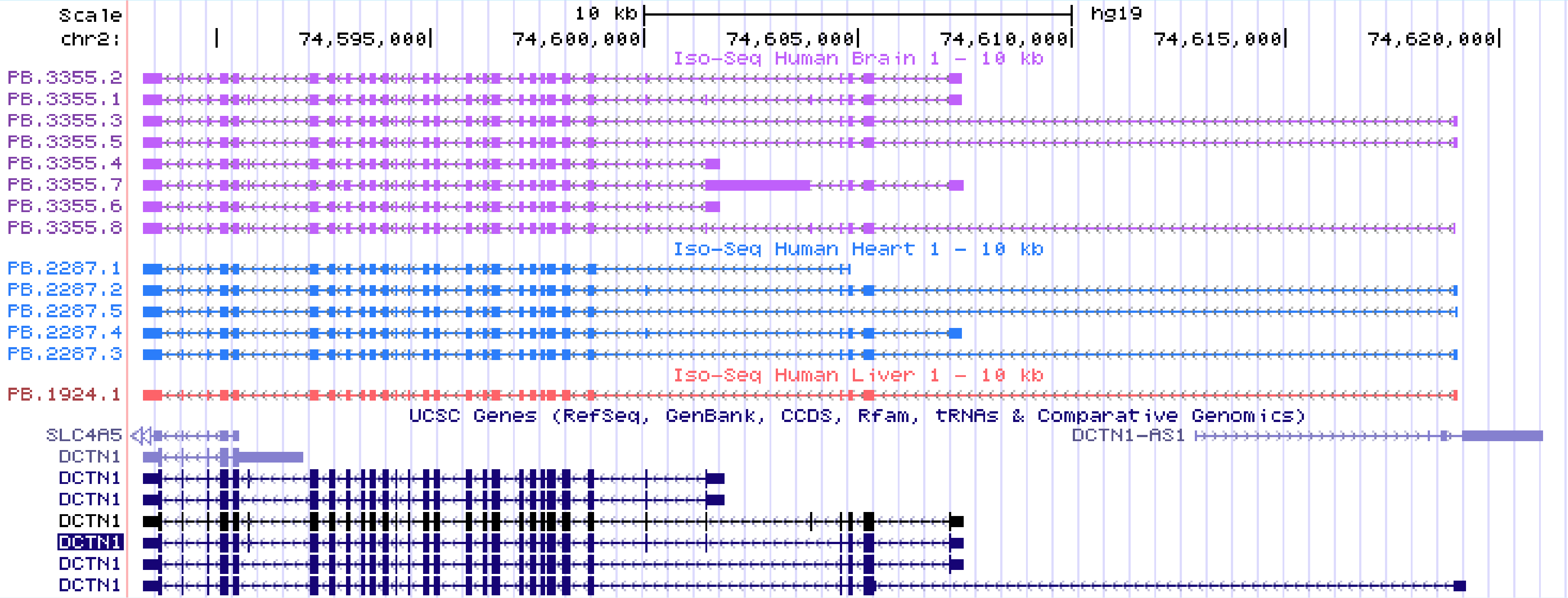


Splicing in more detail



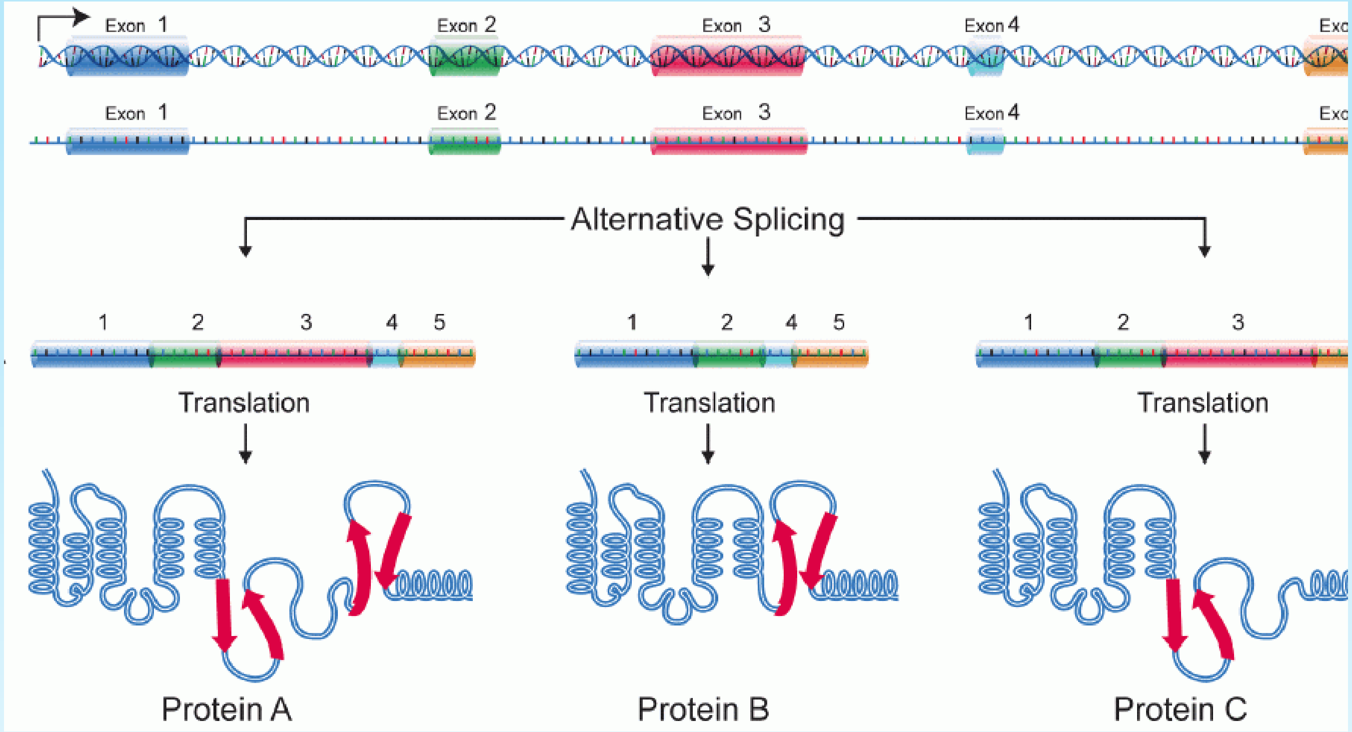
Splicing – one example

https://genome.ucsc.edu/



Note: Only ca. 1.5% of the human genome codes for protein!

Splicing: one gene = many proteins

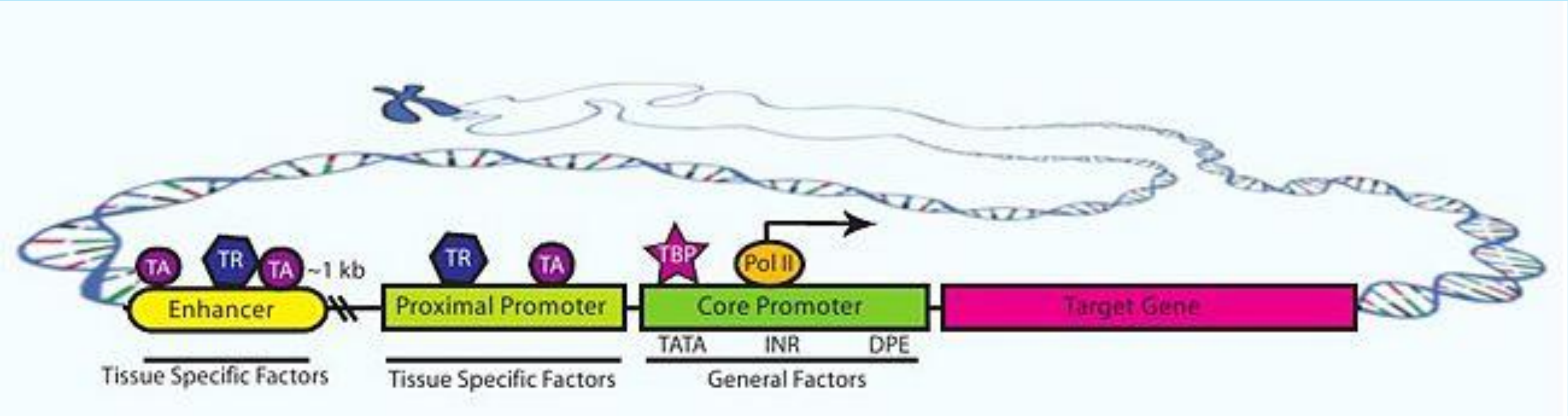


Average human gene = 8.8 exons
Average exon size ≈160 bp

To add to the complexity of the proteome: Most proteins can be post-translationally modified (acetylated, phosphorylated etc.).

Promoters

(of protein-coding genes = Pol II transcribed)*



ca. 1000 bp ca. 200 bp

Promoter

TATA = sequence typically found in ca. 50% promoters
(binds TBP; TATA binding protein)

INR = initiator element, where Polymerase binds

DPE = downstream promoter element

*
Pol I = transcribes the two large rRNAs (18S and 28S)
Pol III = transcribes small rRNA (5S), tRNAs, and other small RNAs.

Repeats

1. “Microsatellites”. Very short, < 10 bp,
   (particularly dinucleotide, trinucleotide and telomeric (6-8 bp)

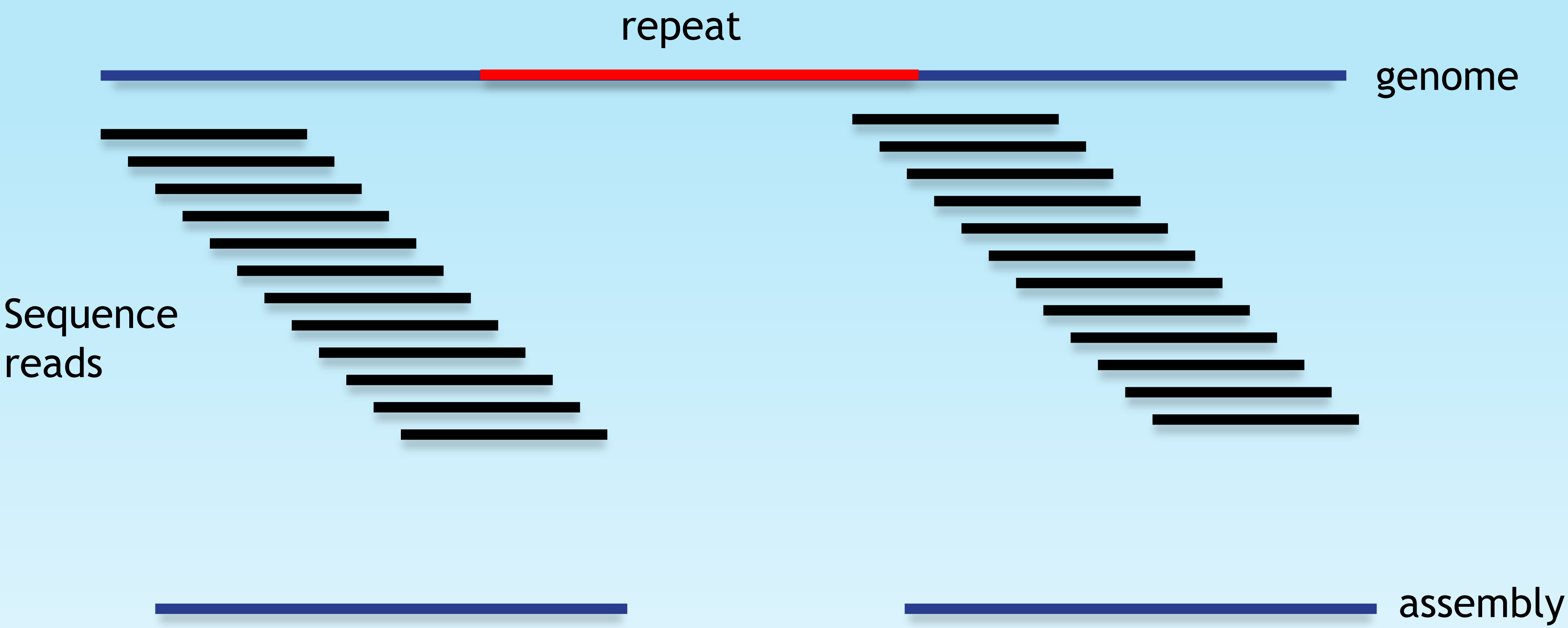
2. “Minisatellites”. 10-50 bp (eg centromeric repeats)

Also called SSRs, STRs and VNTRs. Used in forensics.

3. Transposons and retrotransposons.
- “Copy and paste” genetic elements
- Retro = Transcribed to RNA -> copied back into DNA at new location
- May encode enzymes to do so (many copies are dysfunctional remnants)

e.g. SINE and LINE elements
- 42% of the human genome is derived from retrotransposons!

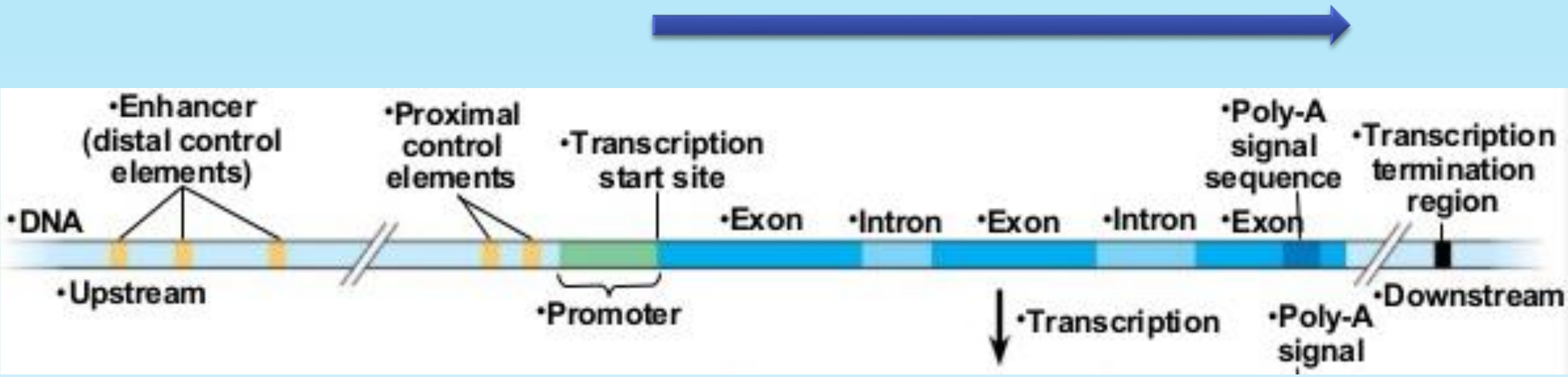
e.g. problem of repeats in genome assembly



Topics:

1. Tree of life
2. Building blocks of life
3. Structure (and differences) of DNA and RNA
4. DNA makes RNA makes protein
5. Genomes and genomic features
6. **Genetics**
7. Epigenetics

6. Genetics: What is a gene?



- You can also have non-coding genes (RNAs that do not contain a reading frame encoding protein) – eg. rRNA, antisense transcripts, Xist
- miRNAs sequences are often transcribed in UTR sequences
- Poly-cistronic RNAs (common in bacteria) = multiple gene sequences in one mRNA.

What is a locus?

Locus = place in the genome where a particular genetic feature resides.

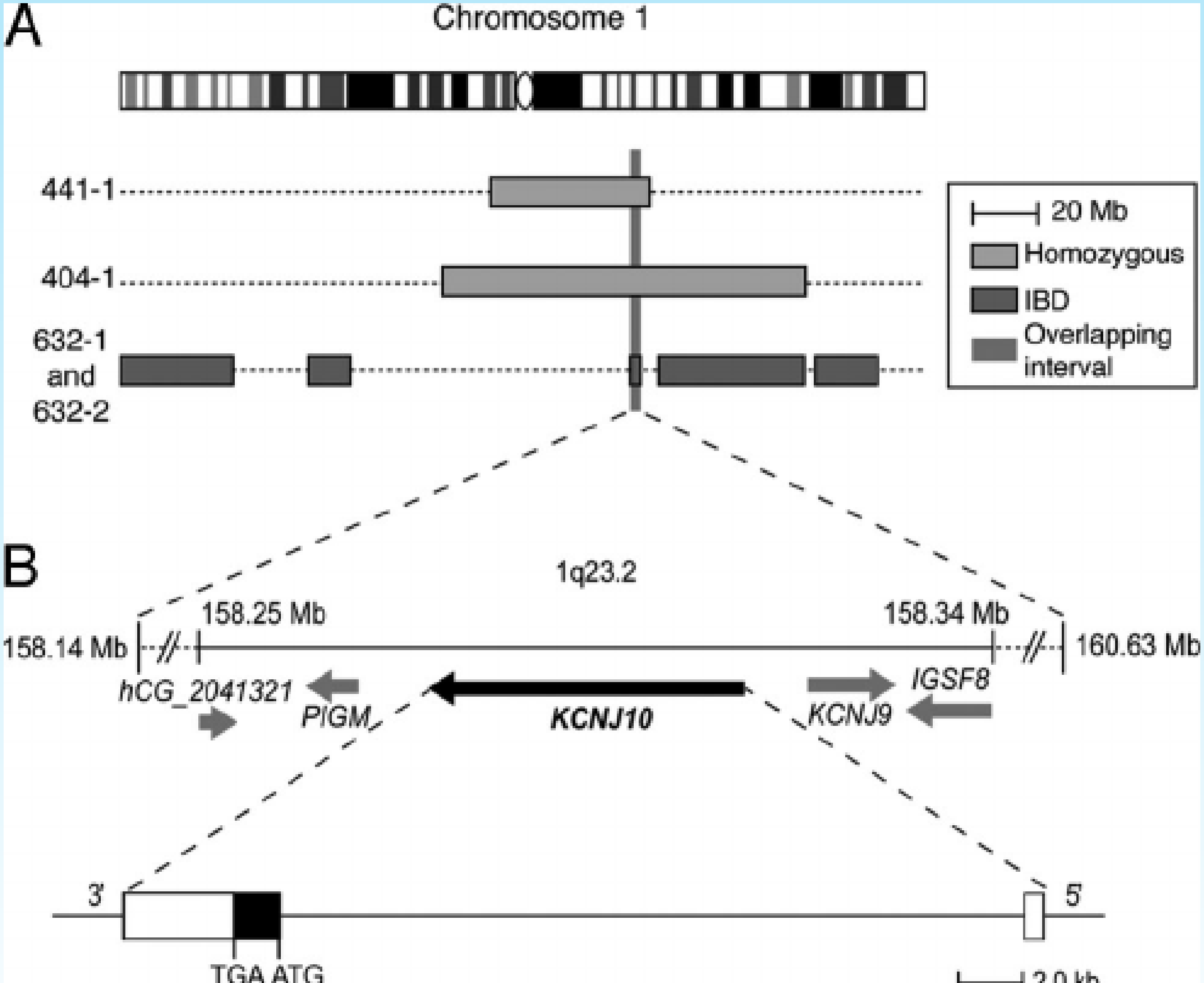
If the gene responsible for a particular trait (eg a disease) has not been identified, we may still know the locus resides on a particular area of a chromosome.

A locus can contain more than one gene, and there can be more than one locus for a kind of gene in the genome. e.g.

- the rDNA loci = in humans, 5 chromosomes carry the rDNA – each contains many copies of the rRNA gene (about 200 copies in total).
- HLA locus (important in immunology) = on chromosome 6, >200 genes
- The locus for Down´s syndrome = entire chromsome 21

E.g. SESAME* syndrome locus eventually mapped to *KCNJ10* gene

*seizures, sensorineural deafness, ataxia, mental retardation, and electrolyte imbalance

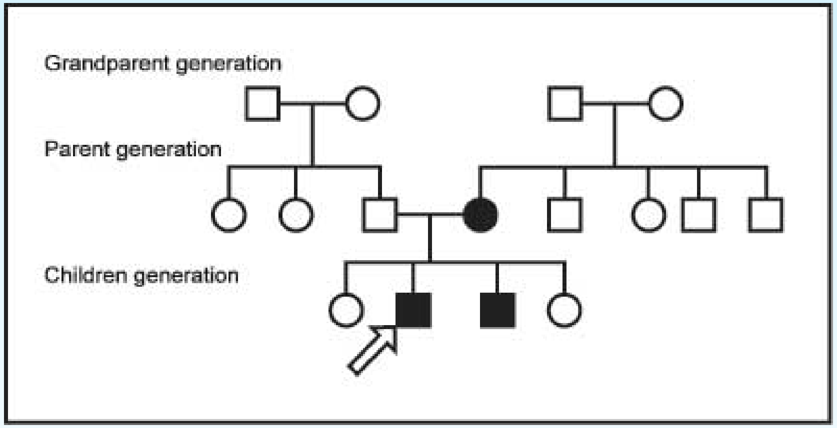


What is an allele?

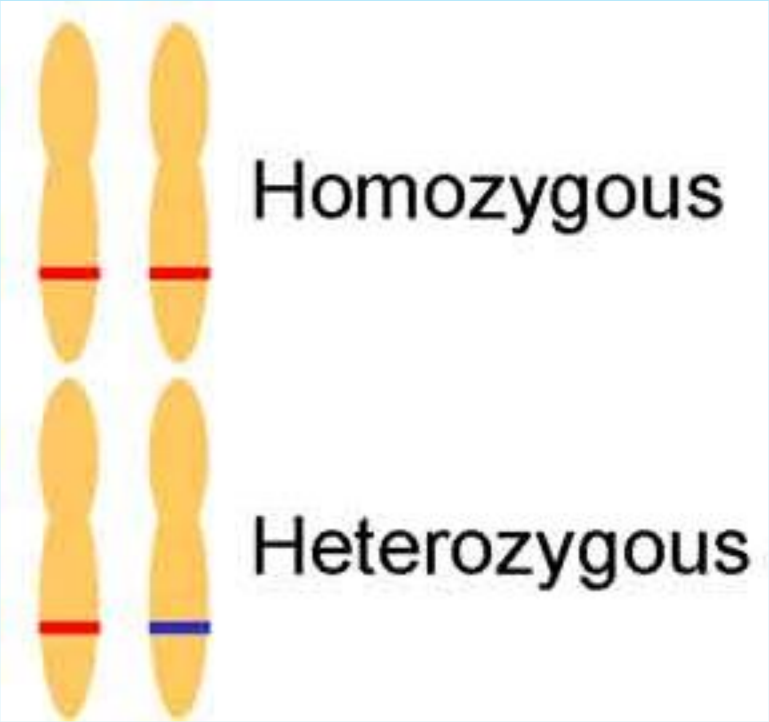
= A variant of any one gene.
Eg a gene where one variant encodes blue eyes, the second variant encodes brown eyes, has two alleles.

The variants can be single nucleotide polymorphisms (**SNPs)**, larger differences in sequence, copy number variants (**CNVs**), splice variants, insertions / deletions (**indels**).

Often discussed in terms of heredity

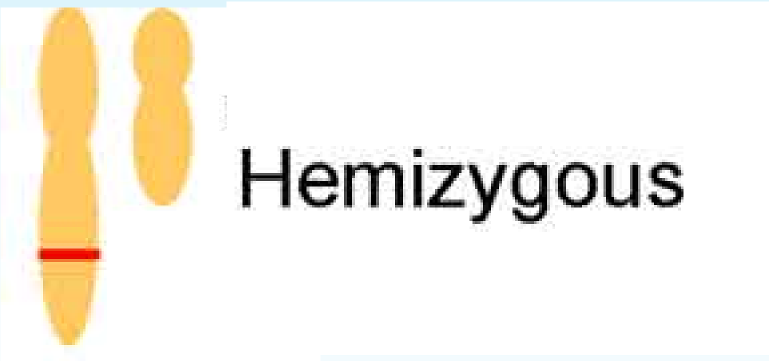


Zygosity



Both alleles the same

Two different alleles

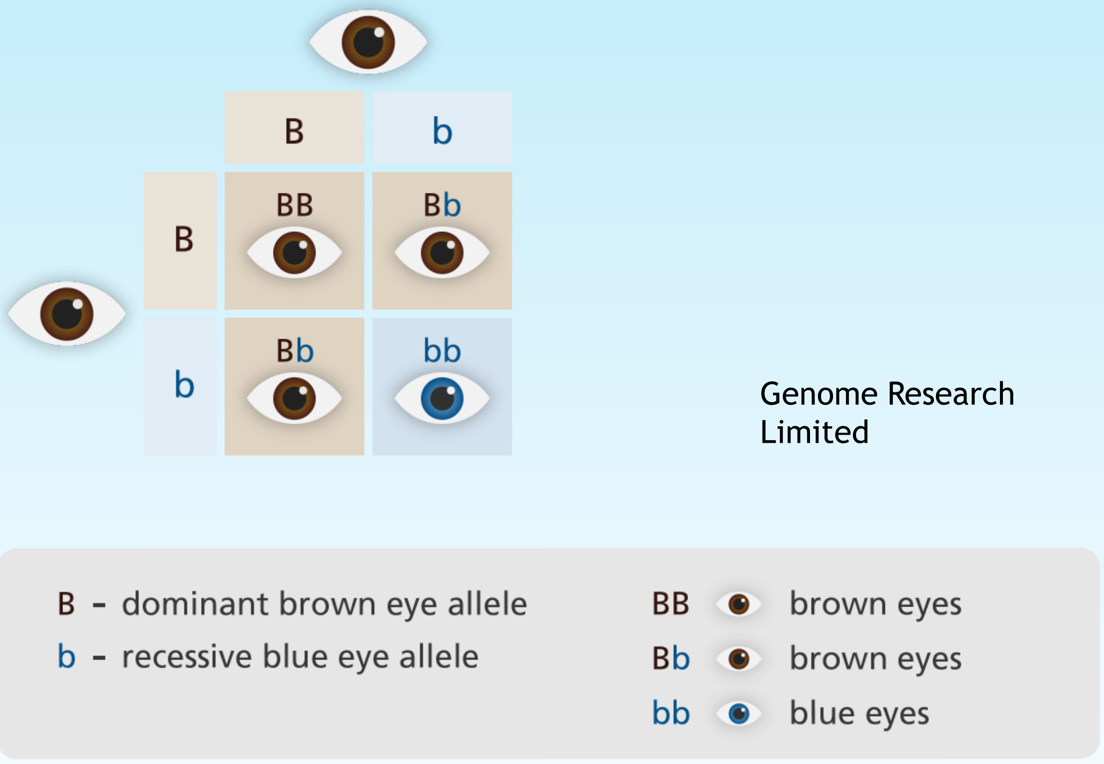


Only one allele possible

Dominant and recessive alleles

Since we have 2 copies of our genes (diploid):

- If both alleles at a given locus contribute to **phenotype** = co-dominance
- If only one allele is displayed in phenotype = dominant allele
- For a recessive allele to display phenotype, both copies must be the recessive allele.



Genome Research Limited

Mutations, variants and polymorphisms

Molecular genetics

AGCTTGCATTGAATCG

Process by which variation arises = **mutation**

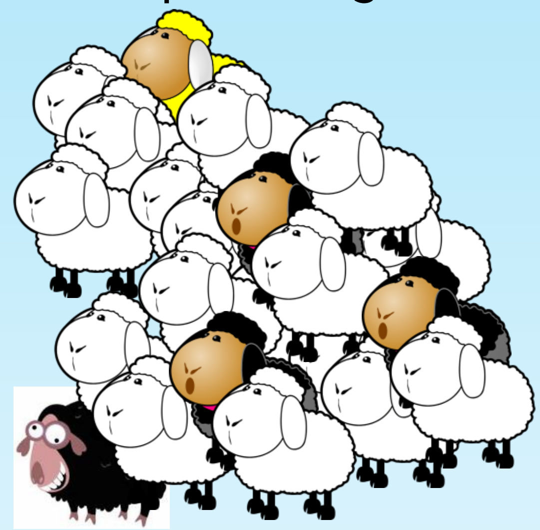
AGCTTGCAATGAATCG

Benign, no disease = variant

AGCTTGCAGTGAATCG

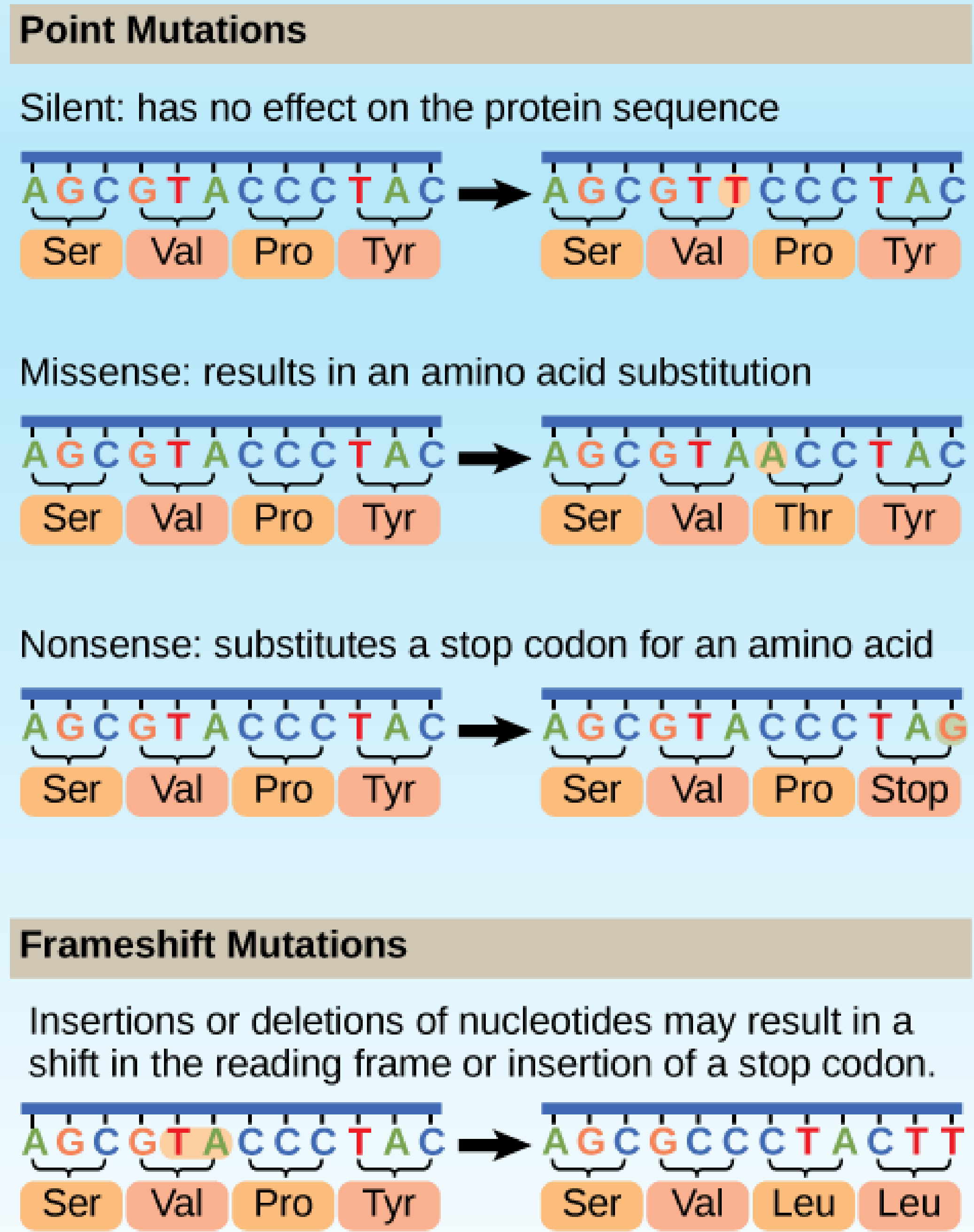
Disease-causing = mutation

Population genetics



Black coat variant (arose by mutation several generations ago) is a polymorphism **inherited** by a minority of individuals. This flock also contains two mutants that have arisen spontaneously.Only one looks damaging.

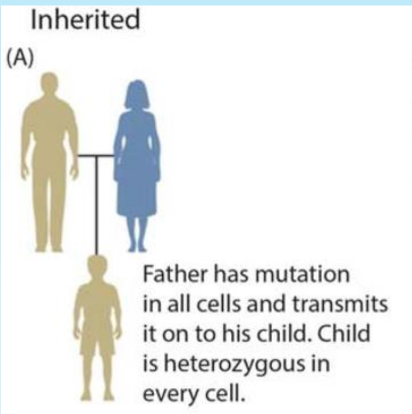
Silent, missense and nonsense mutations

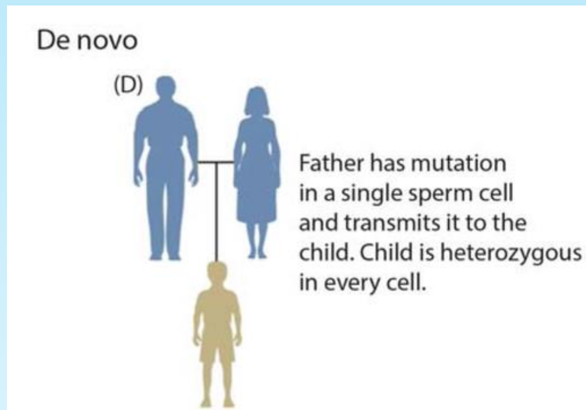


libretexts

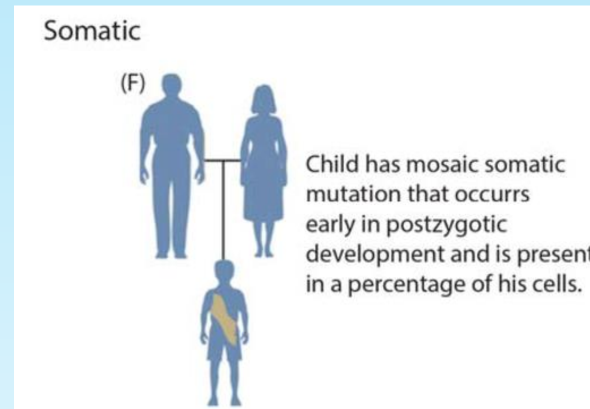
Somatic vs Germline mutations, and mosaicism

Germline = testes, ovaries (sperm and eggs)
Soma = all other tissues

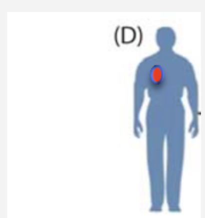




Spontaneous germline mutation



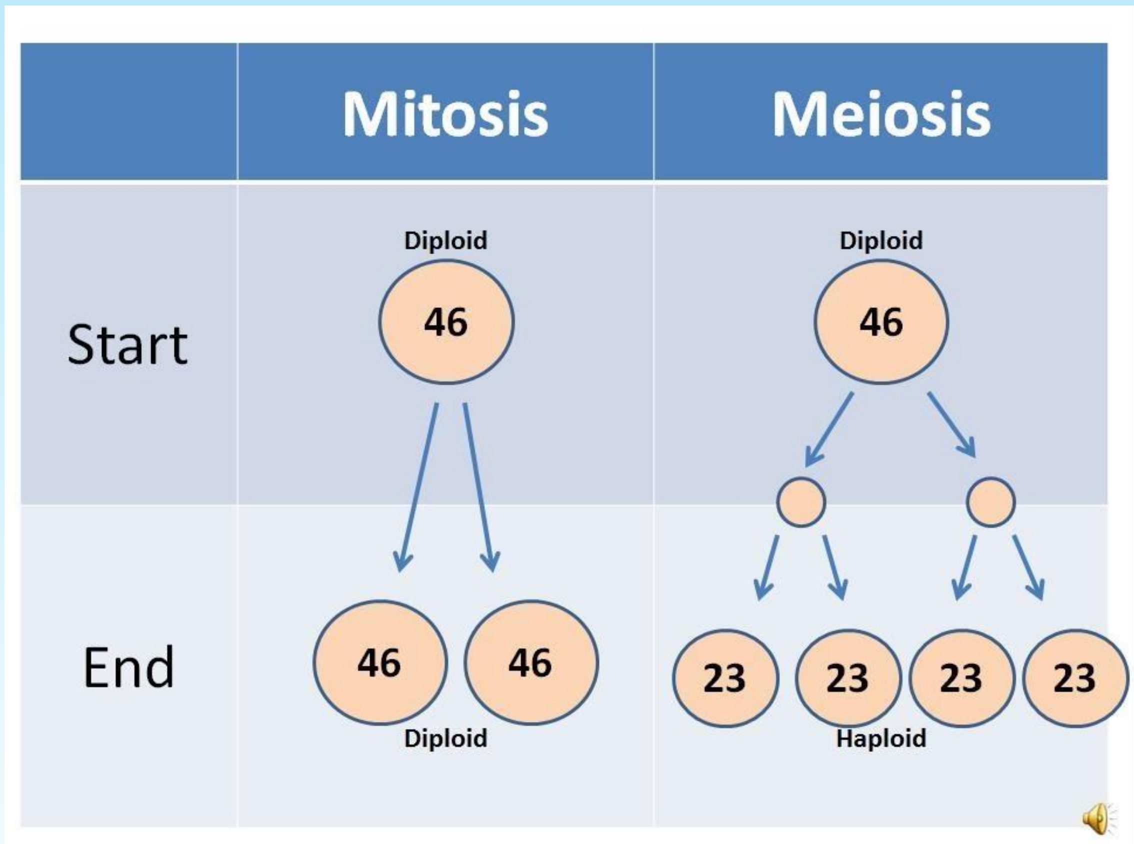
Mosaic individual.
If mutation present in germ cells, can be inherited by next generation



Man develops cancer

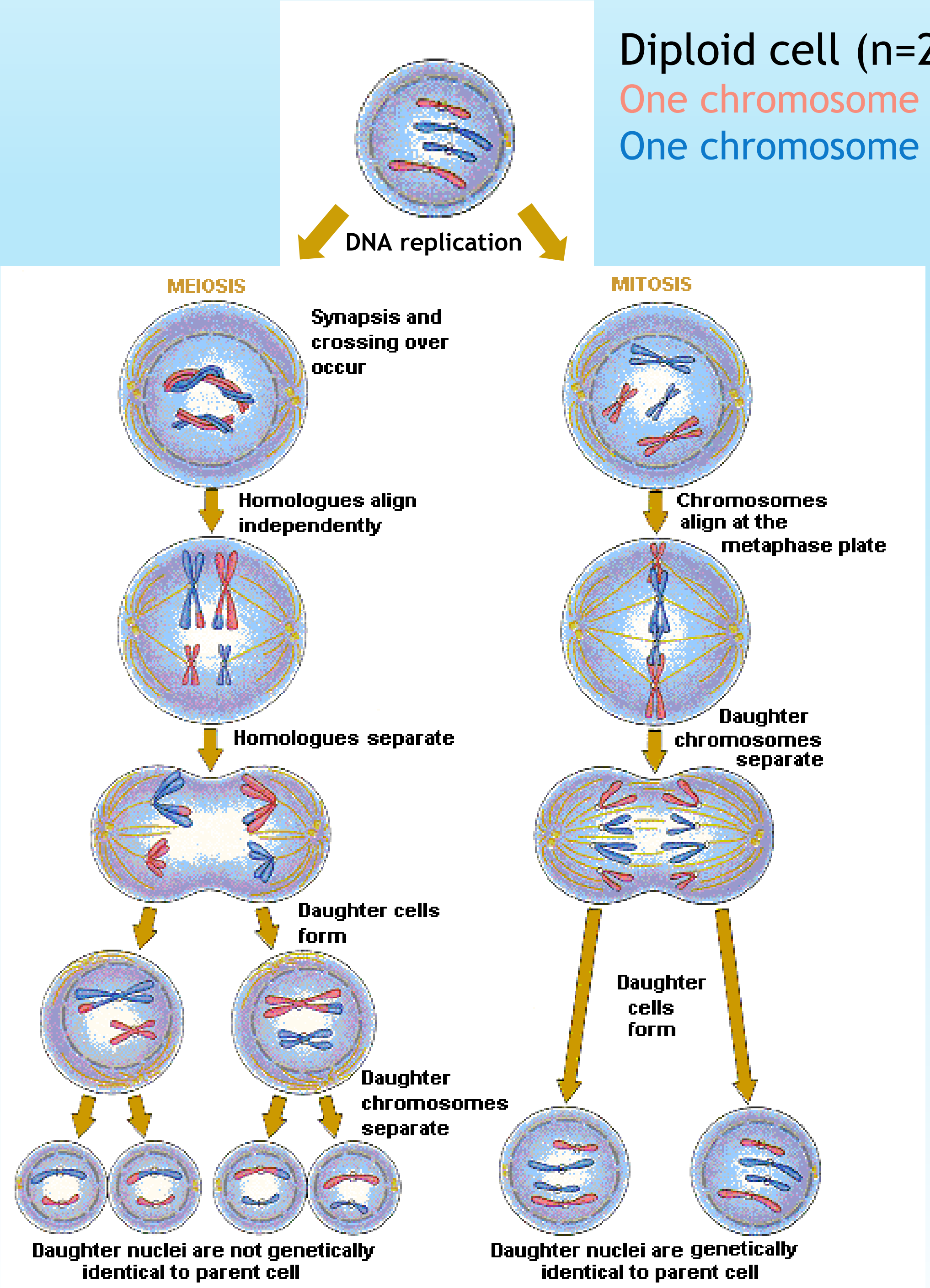
Spontaneous somatic mutation

Difference between somatic cell division and DNA replication (mitosis) and germ-line (meiosis)



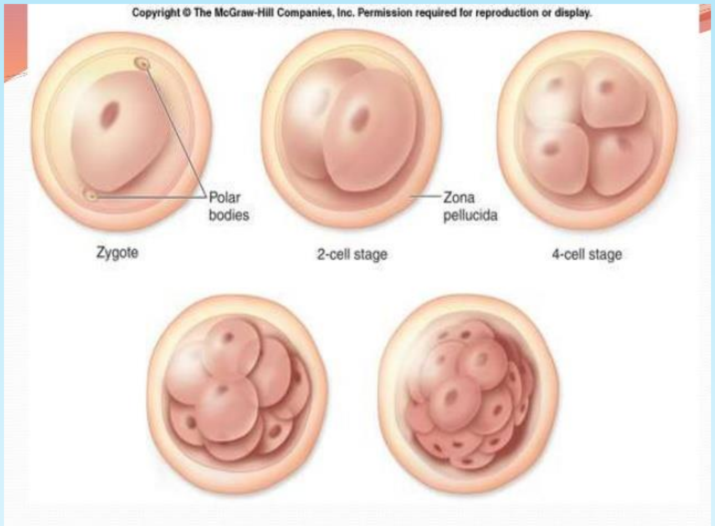
Difference between somatic cell division and DNA replication (mitosis) and germ-line (meiosis)

Diploid cell (n=2)
One chromosome set from mum
One chromosome set from dad



apbio

Every cell in your body contains the same DNA (almost)



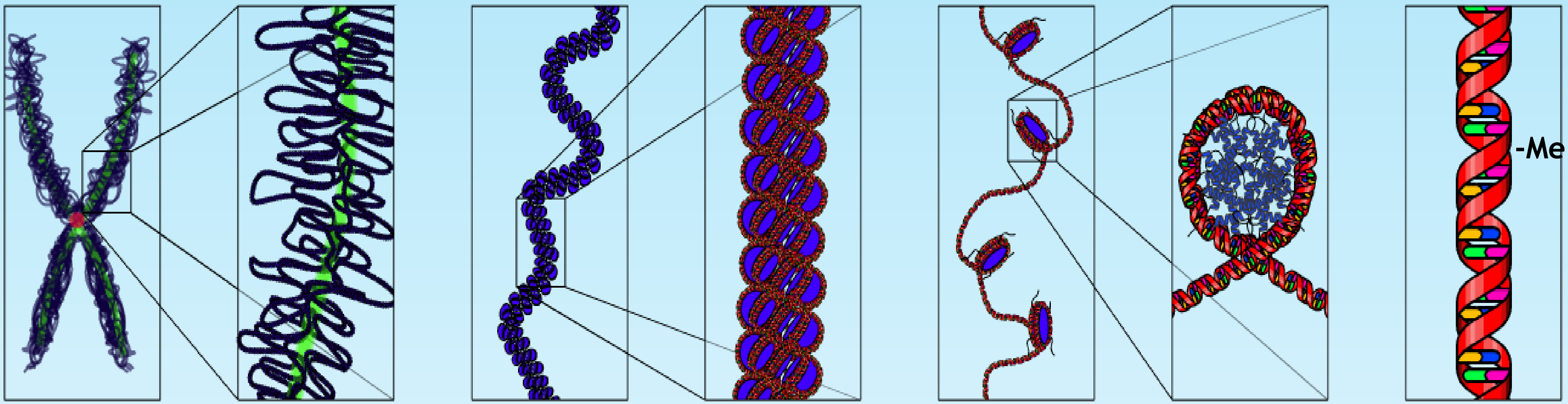
Exceptions:
- Red blood cells (lose their nuclei)
- B-cells (undergo genomic rearrangement to produce different antibodies)
- Sperm / egg and precursors = undergo meiotic recombination
- Cancer cells
- Somatic mutations that accumulate as you develop and age.

Which genes are turned on / off at a given time or in a tissue determine cell identity.

Topics:

1. Tree of life
2. Building blocks of life
3. Structure (and differences) of DNA and RNA
4. DNA makes RNA makes protein
5. Genomes and genomic features
6. Genetics
7. **Epigenetics**

7. Epigenetics: Heritable traits not encoded in the DNA sequence.



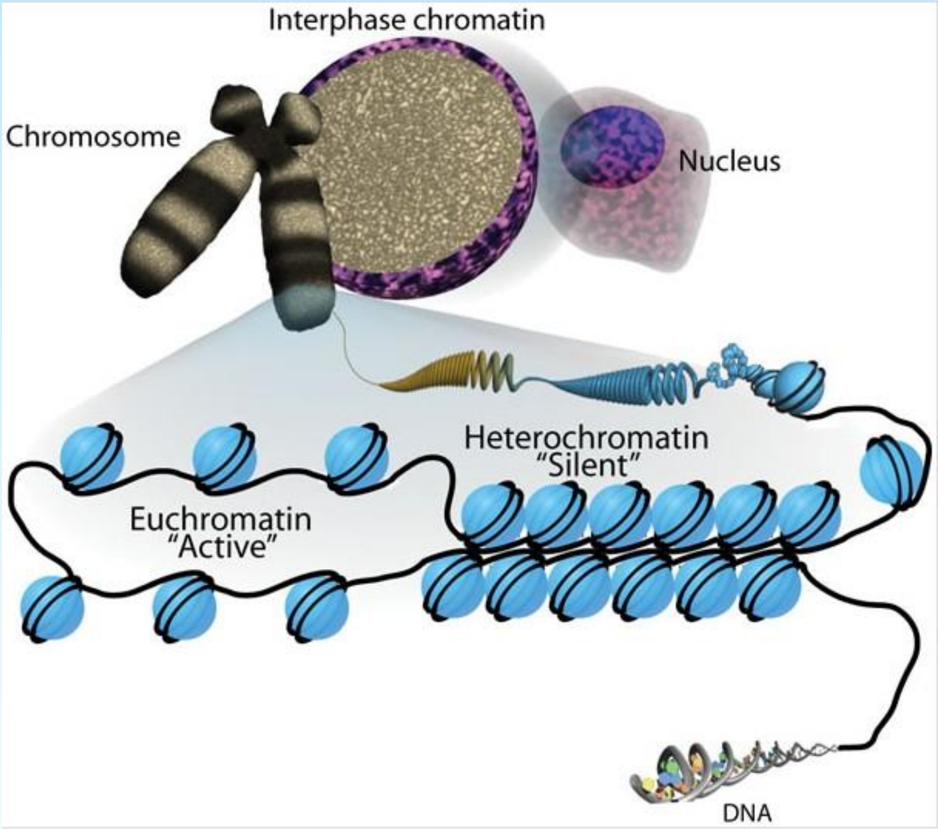
wikipedia

Mediated by:

-DNA methylation
-Histone variants
-Histone covalent modification
-Non-coding RNAs (miRNAs, long non-coding RNAs)
-Nucleosome remodelling (positioning) and chromatin compaction
-Nuclear position

DNA packaging (DNA + packaging proteins = chromatin)

-Each cell in your body contains about 2m DNA (in 46 pieces i.e. chromosomes)
- All that has to be packed into a nucleus that of approx 5 μm diameter – which is why you need chromatin!



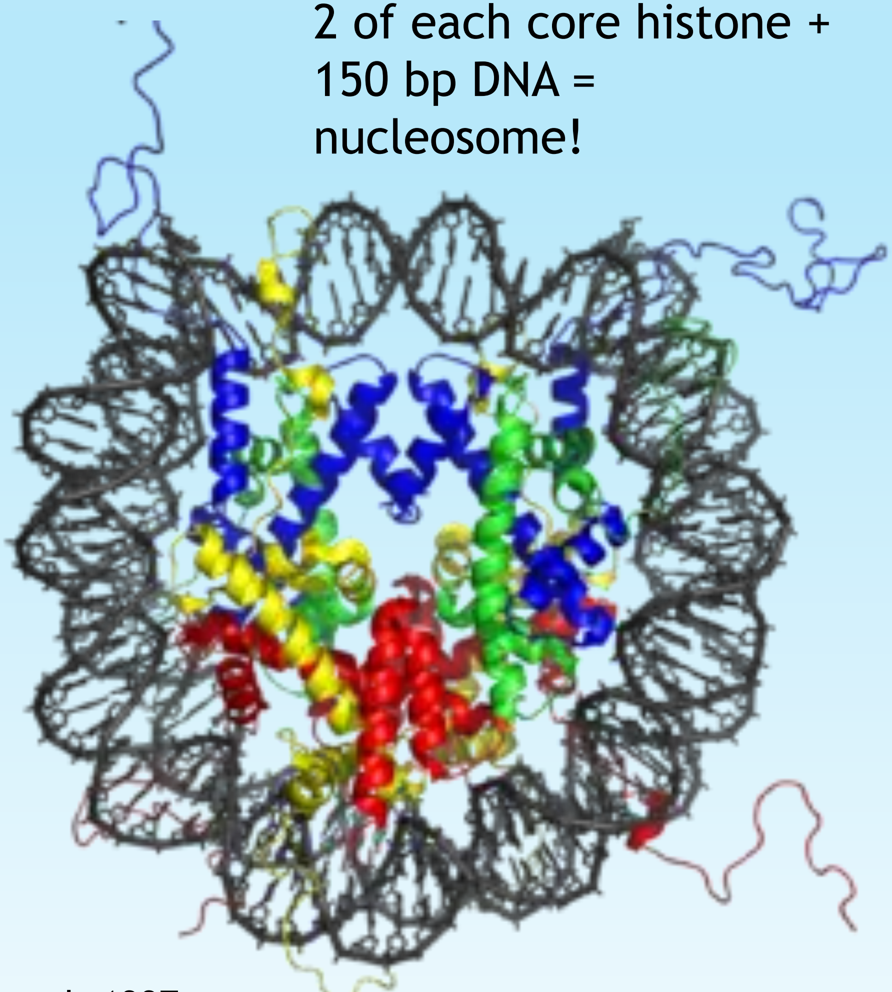
All the DNA in your body (10^{13}-10^{14} cells), if aligned end to end could stretch 5-50 times from the earth to the sun…

Histones + nucleosomes

Histone proteins:
- Small (ca. 100-140 aa)
- Highly conserved
- Package DNA

4 core histones:
H2A
H2B
H3
H4

2 of each core histone + 150 bp DNA = nucleosome!



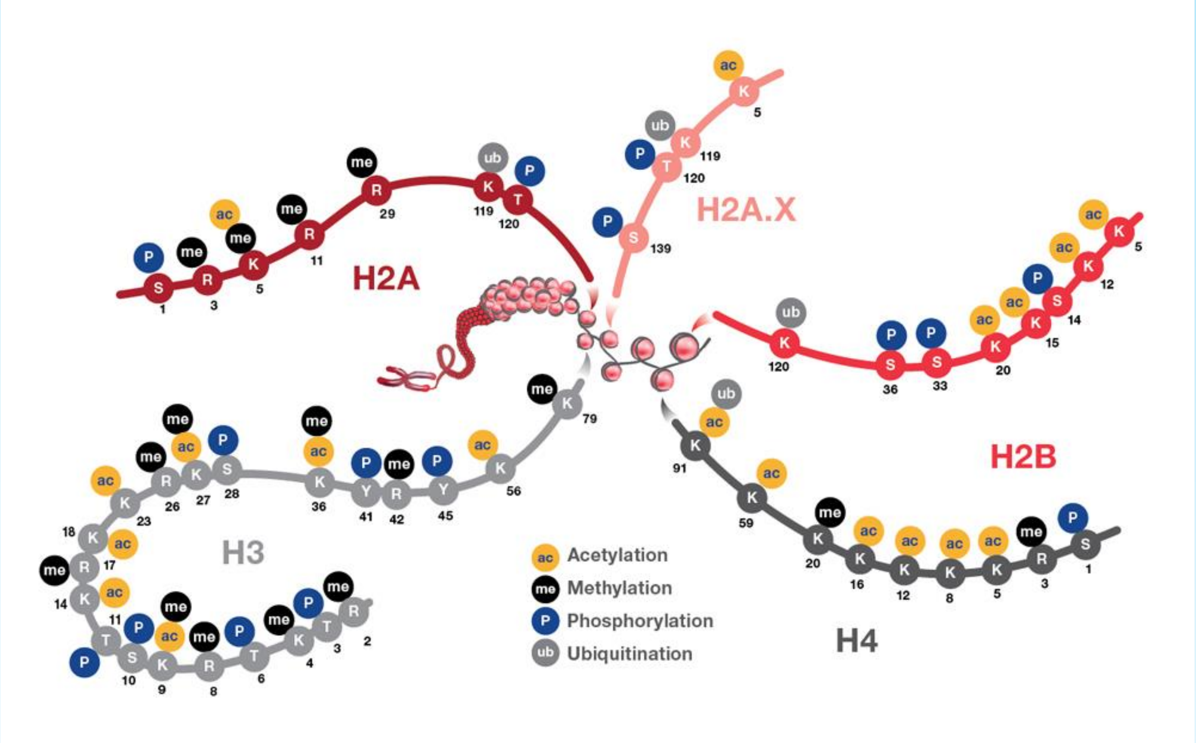
Luger et al, 1997

N-terminal TAILS

Histone tails are targets for modifications

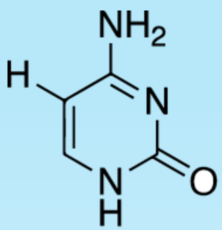
Post-translational modifications

-acetylation
-methylation
-phosphorylation
-Ubiquitylation
-SUMOylation
-Citrullination / deimination
-Proline isomerisation
-ADP-ribosylation
-Palmitoylation
-GlcNac

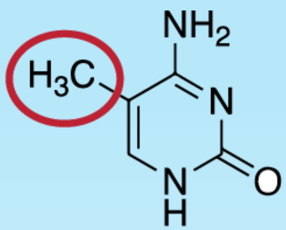


These modifications control how tightly the chromatin is packed (eg condenses to heterochromatin), attract or repulse transcription factors, signal damage etc.

DNA methylation



Cytosine



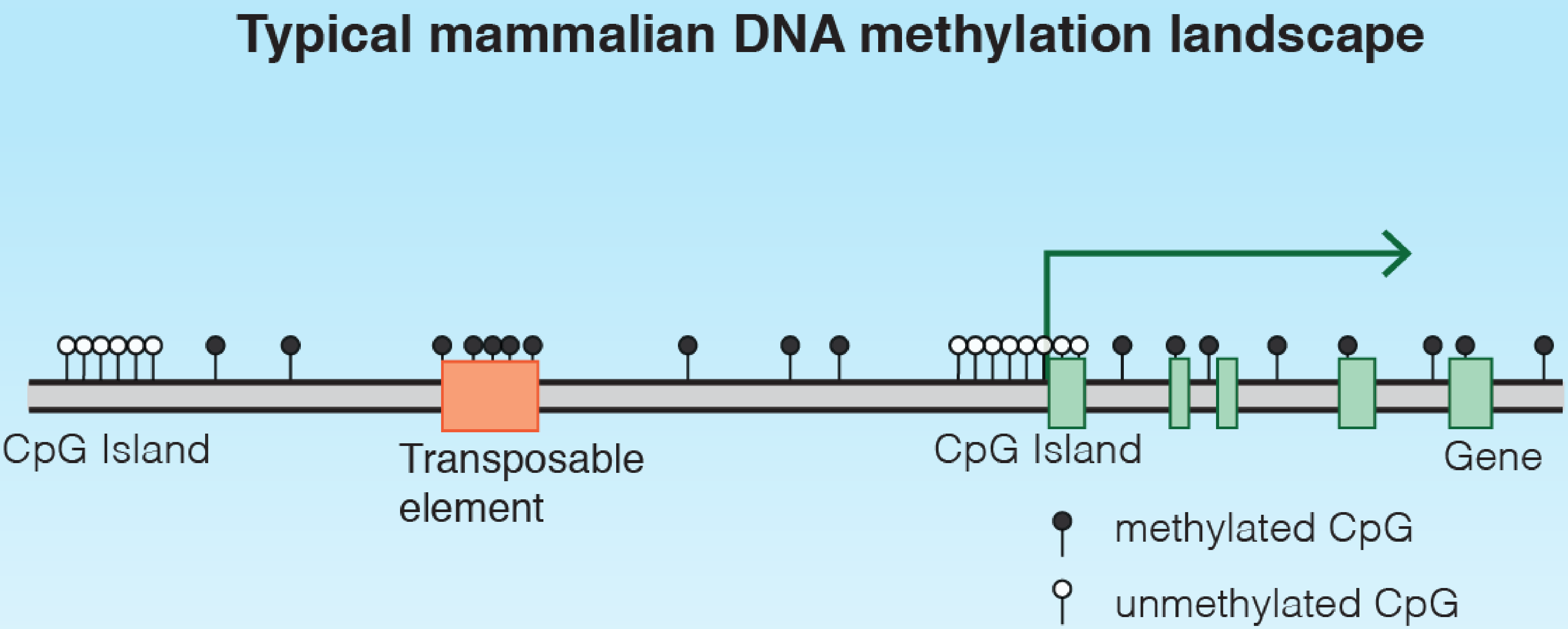
methylated Cytosine

(aka 5mC)

Also get adenine methylation

- Occurs most in CG (also known as CpG) context in mammals.
- Also common in CHH and CHG in plants (H= A, C or T)
- Adenine and cytosine methylation common in selected sequence motifs in bacteria as an antiviral defense. Express an enzyme that cuts a DNA sequence (restriction enzyme): non-methylated viral DNA gets cut, methylated bacterial DNA protected.

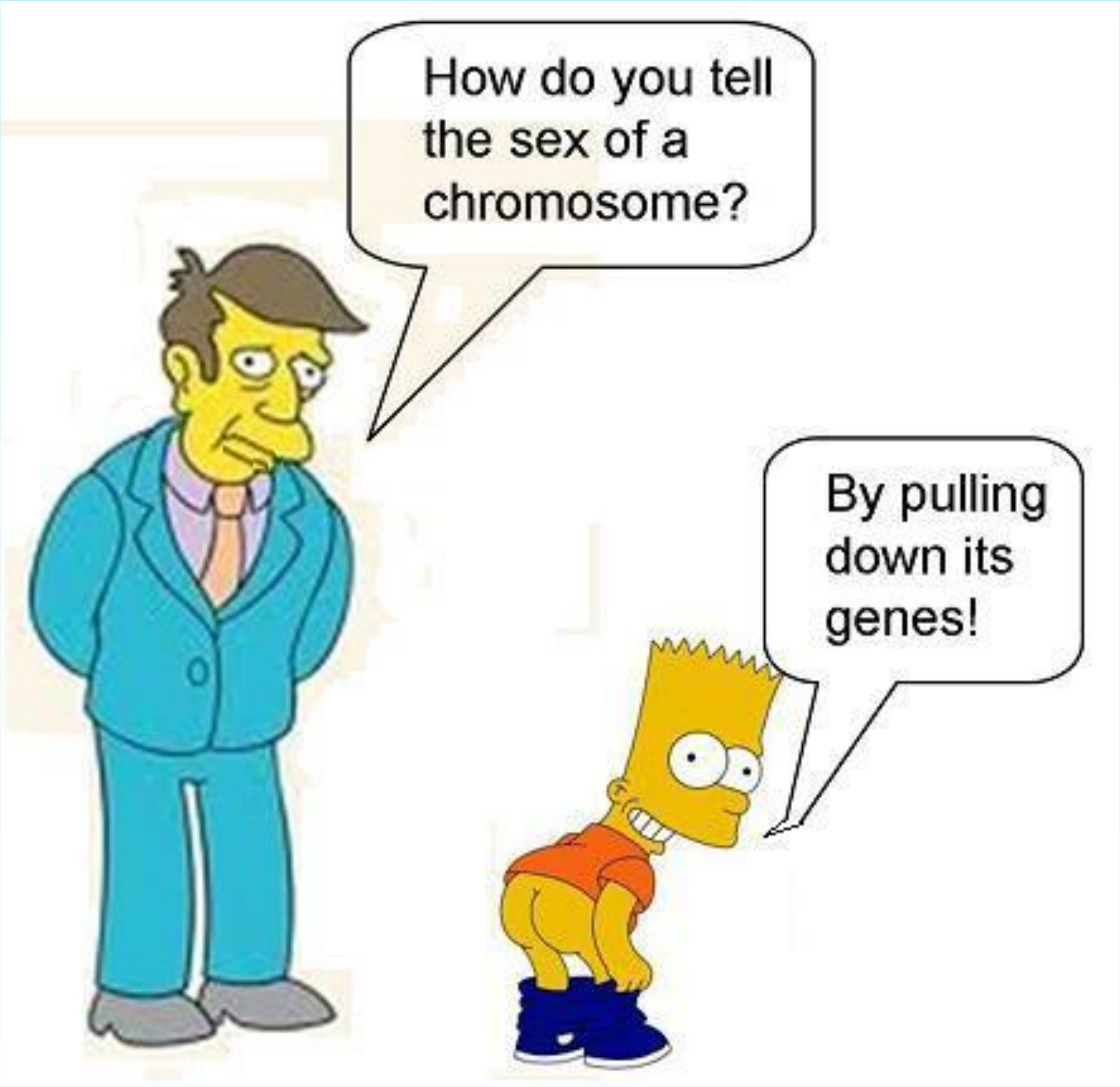
DNA methylation in vertebrates



- 5mC promotes transcriptional repression (can block TF binding directly, and can recruit proteins that cause chromatin condensation)
- Clusters of non-methylated C found at ca. 65 % of vertebrate promoters
- 5mC mutates more frequently to T, therefore gradual conversion of CG to TG in genome, except in promoters that are actively demethylated = gives rise to “CpG islands”.

Topics:

1. Tree of life
2. Building blocks of life
3. Structure (and differences) of DNA and RNA
4. DNA makes RNA makes protein
5. Genomes and genomic features
6. Genetics
7. Epigenetics


How do you tell the sex of a chromosome?
By pulling down its genes!

Thanks for your attention

www.sequencing.uio.no
post@sequencing.uio.no

NSC
NORWEGIAN SEQUENCING CENTRE

gregorg@medisin.uio.no