

Machine learning for genomic medicine

Outline:

Molecular biology basics
Comparative genomics
Next Generation Sequencing (NGS)
NGS experiment types
Computational analysis of NGS data
Worked example: variants call

Introduction to molecular biology

Molecular biology the branch of biology that study gene structure and function at the **molecular level**.

The **Molecular biology** field overlaps with other areas, particularly genetics and biochemistry.

Molecular biology allows the lab to be predictive in nature (events that occur in the future are strictly dependent from previous “molecular states”)

Introduction to molecular biology

Many different types of organisms on this planet...
They are classified into **three** main groups:

Eukaryotes

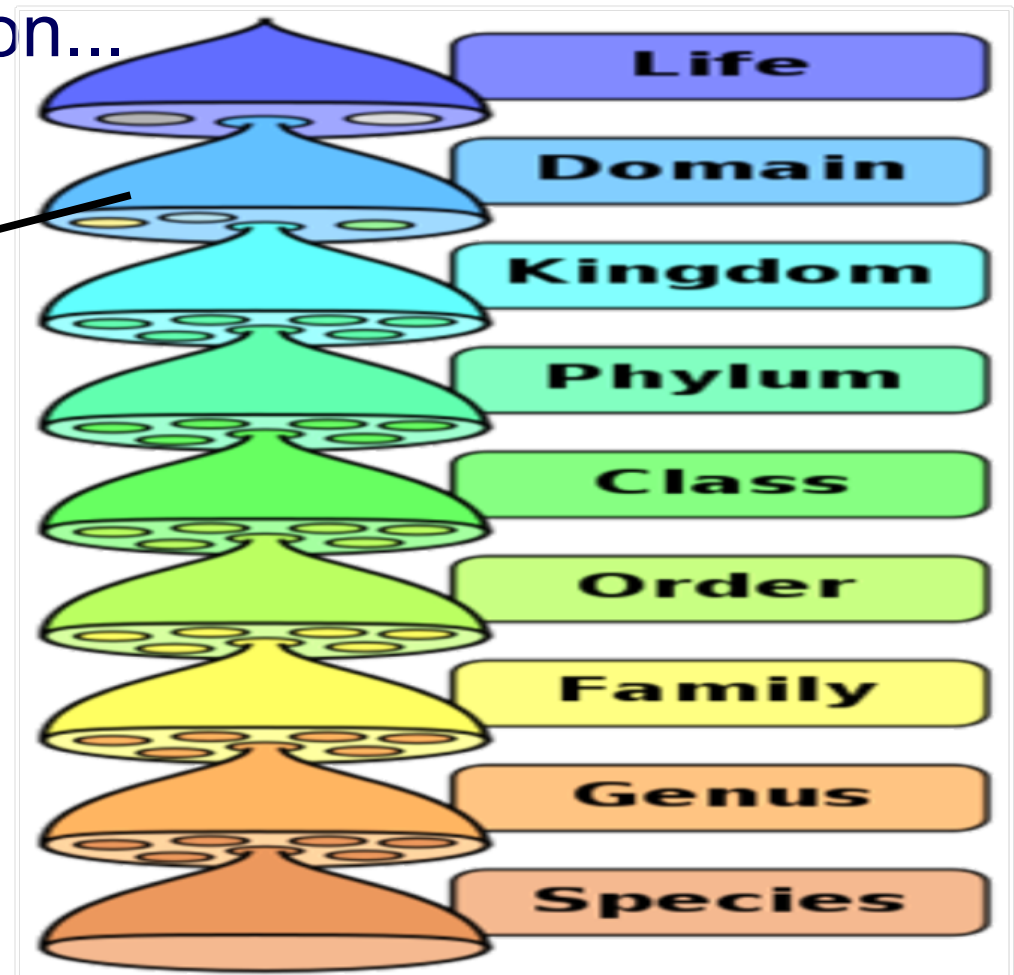
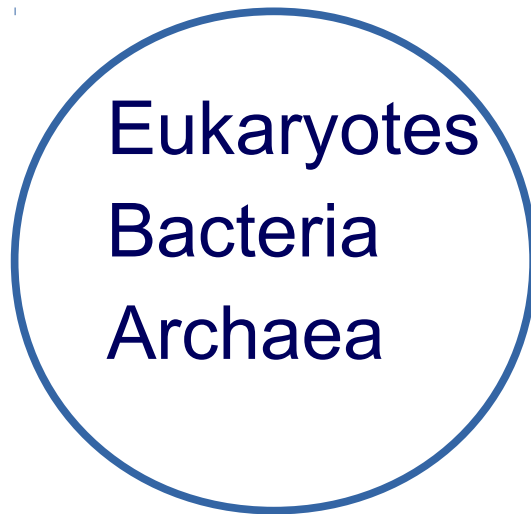
Bacteria

Archaea

Some organisms lies outside the main three
groups (i.e. viruses)

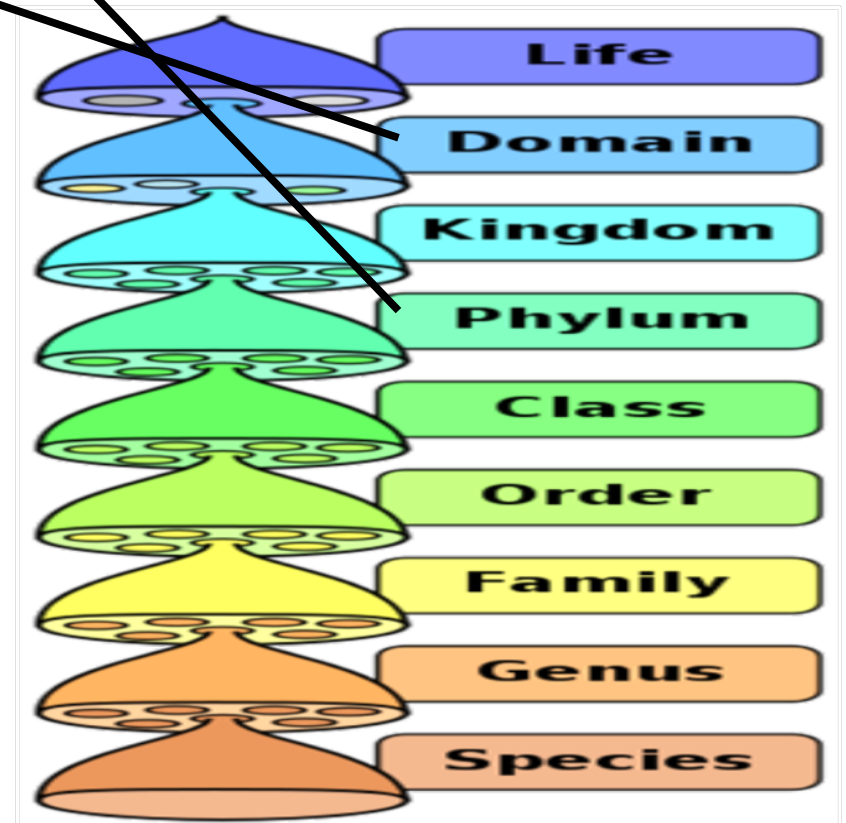
Introduction to molecular biology

Classification of organisms: beside the three main groups (**domains**) there are other and more specific levels of classification...

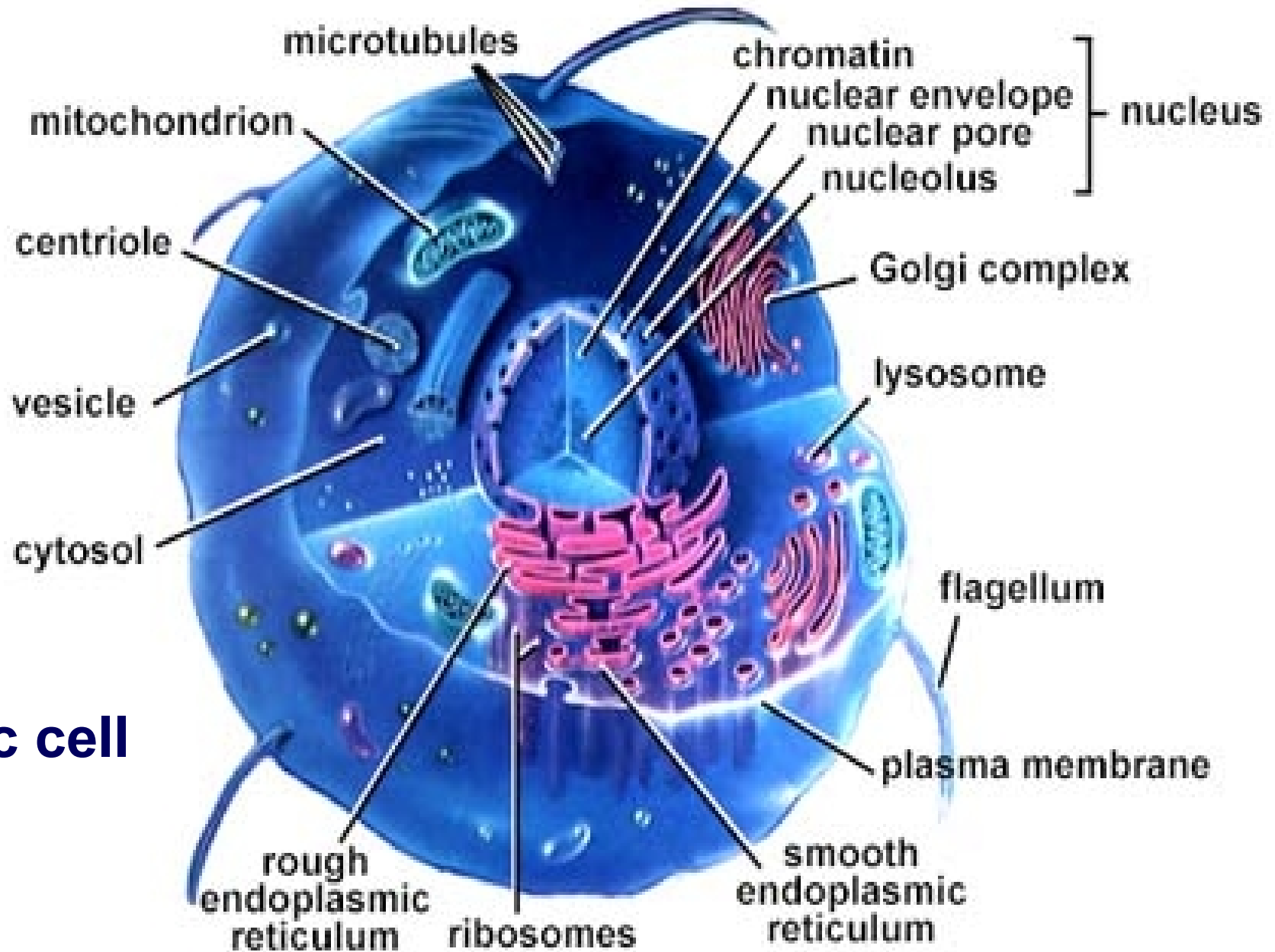


Introduction to molecular biology

Eukaryota; Metazoa; Chordata; Craniata;
Vertebrata; Euteleostomi; Mammalia; Eutheria;
Euarchontoglires; Primates; Haplorrhini; Catarrhini;
Hominidae; Homo; sapiens



Introduction to molecular biology



Eukaryotic cell

Introduction to molecular biology

Eukaryotic cell

Eukaryotic cells are found in animals, plants, fungi and protists cell;

Cell with a **true nucleus**, where the genetic material is surrounded by a membrane;

Eukaryotic genome (the whole genetic information contained into a single cell) is more **complex** than that of prokaryotes and distributed among multiple chromosomes;

Eukaryotic DNA is linear;

Eukaryotic DNA is complexed with proteins called **histones**;

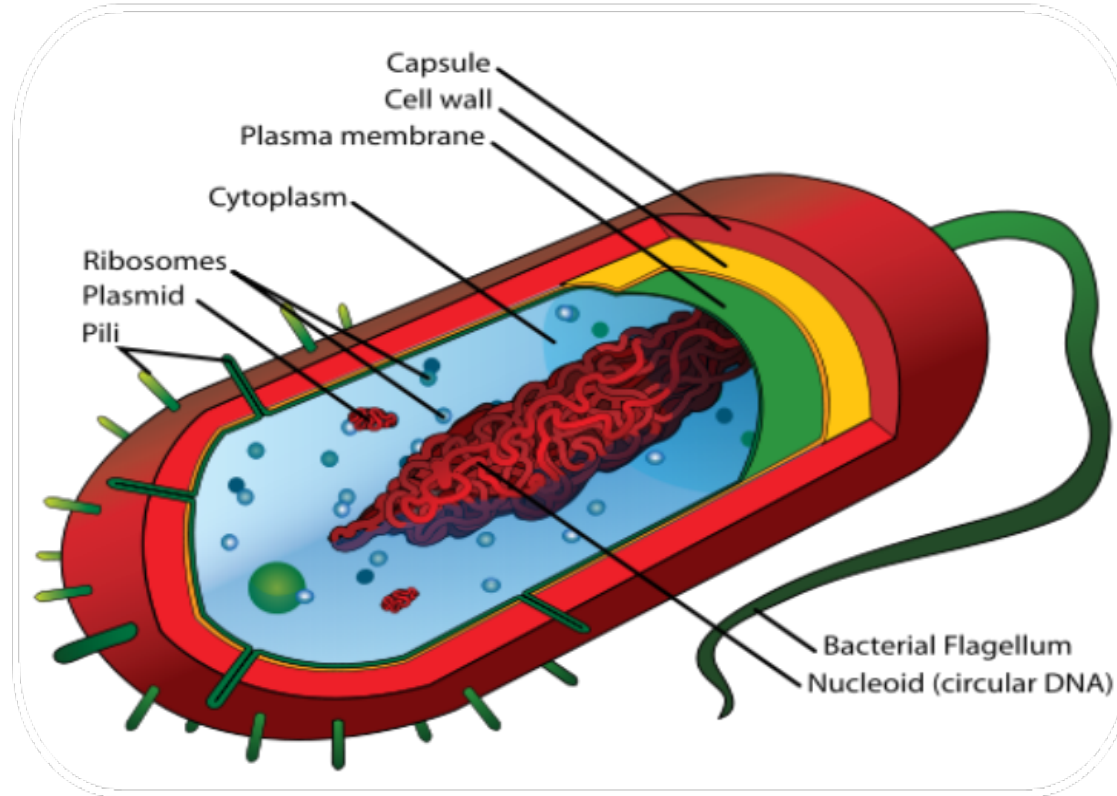
Numerous membrane-bound organelles;

Complex internal structure;

Cell division by **mitosis**.

Introduction to molecular biology

Prokaryotic cell



Introduction to molecular biology

Prokaryotic cell

Unicellular organisms, found in all environments. These include bacteria and archaea;

Without a nucleus; no nuclear membrane (genetic material dispersed throughout cytoplasm);

No membrane-bound organelles;

Cell contains only one circular DNA molecule contained in the cytoplasm;

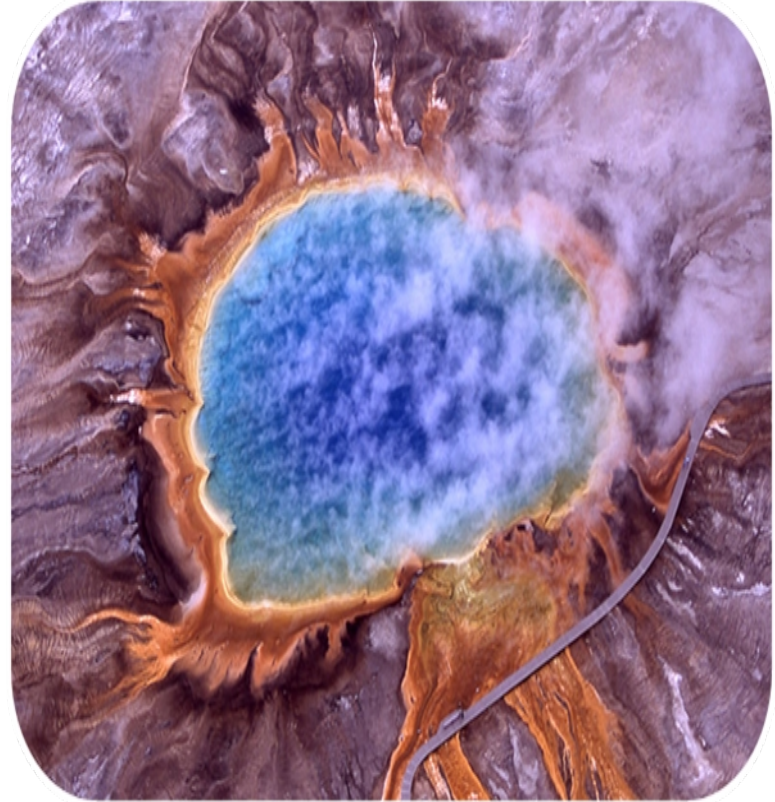
DNA is naked (no histone);

Simple internal structure and Cell division by simple binary fission.

Introduction to molecular biology

Archaea

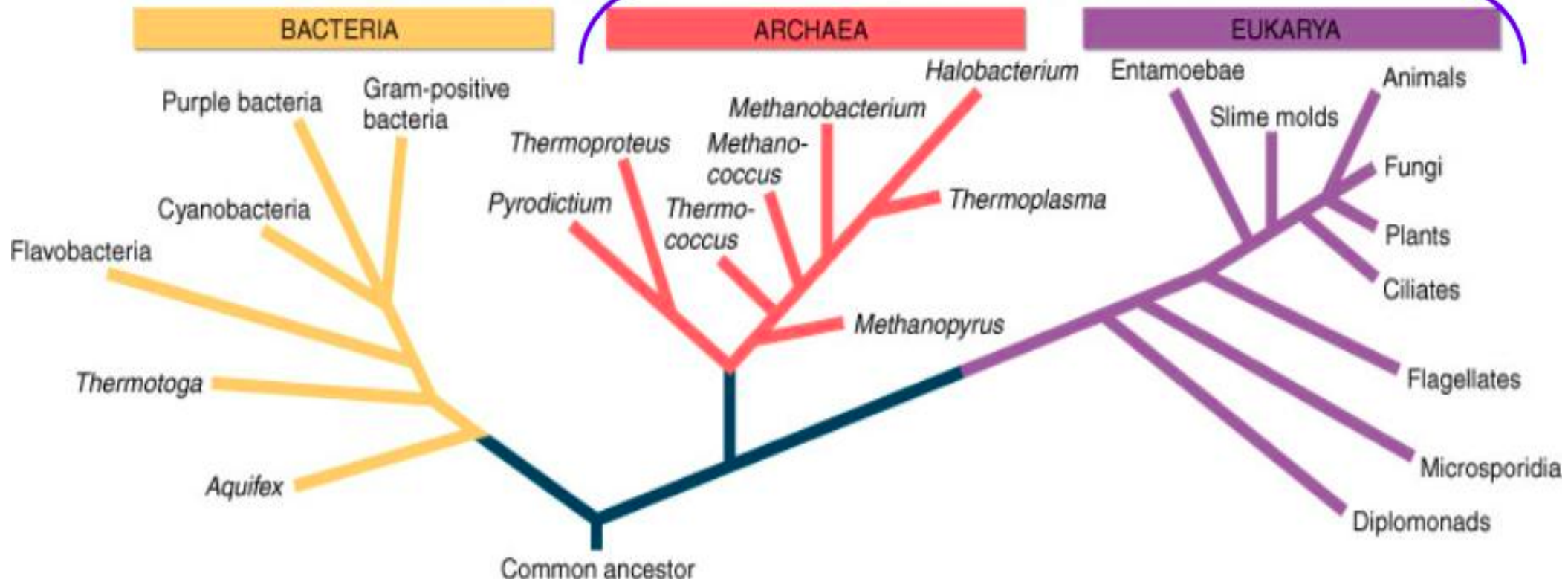
Are very similar to prokaryotes; organisms **without nucleus** but some aspects of their molecular biology are more similar to those of eukaryotes.



Introduction to molecular biology

The tree of life

More closely related to each other than either is to bacteria



Introduction to molecular biology

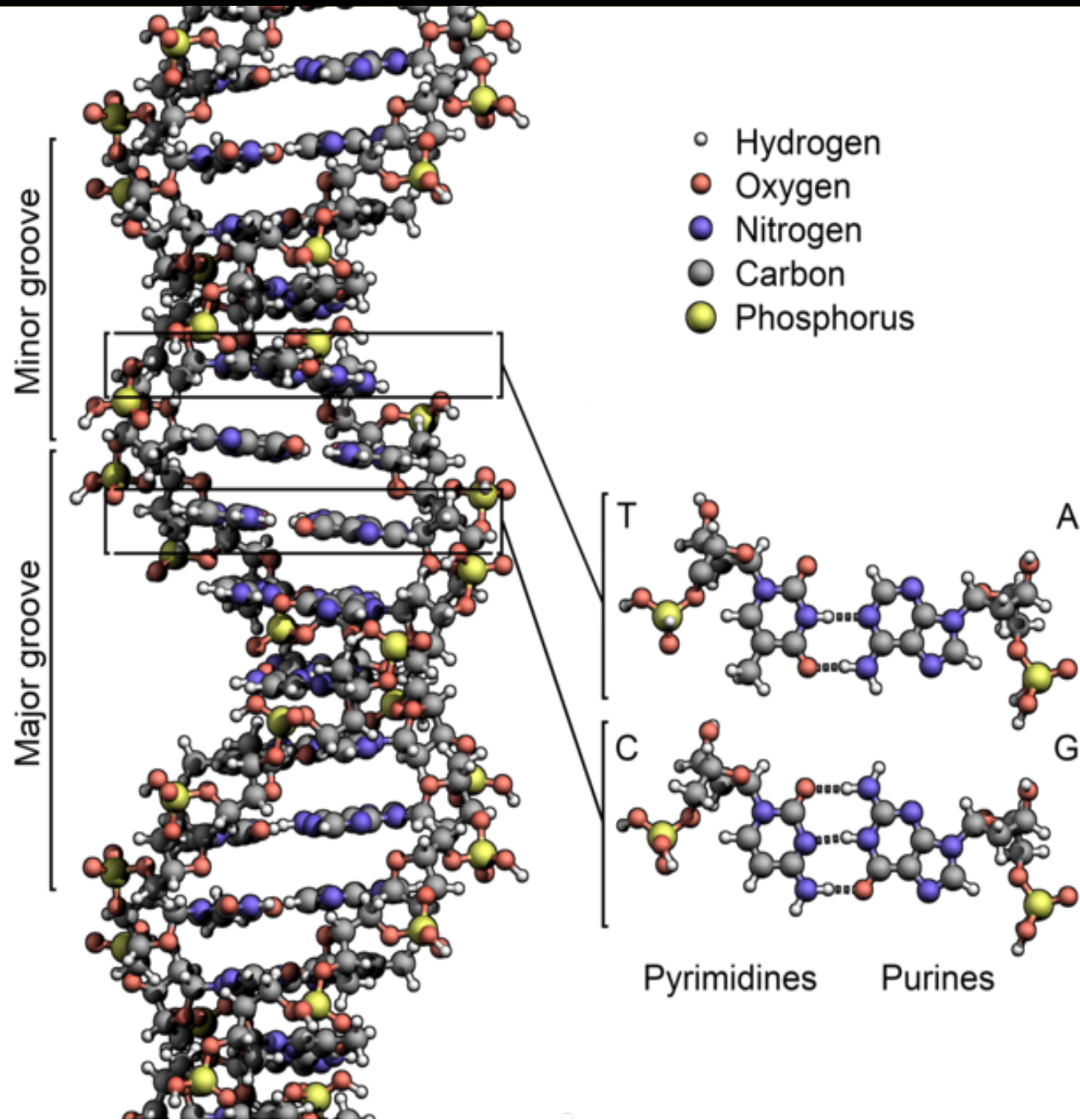
The genome

Totality of genetic information of an organism.

The information is stored into a nucleic acid, a very long macromolecule called DNA , **deoxyribonucleic acid** (for some viruses, the genetic information is stored into a similar macromolecule, the RNA **ribonucleic acid**)

Introduction to molecular biology

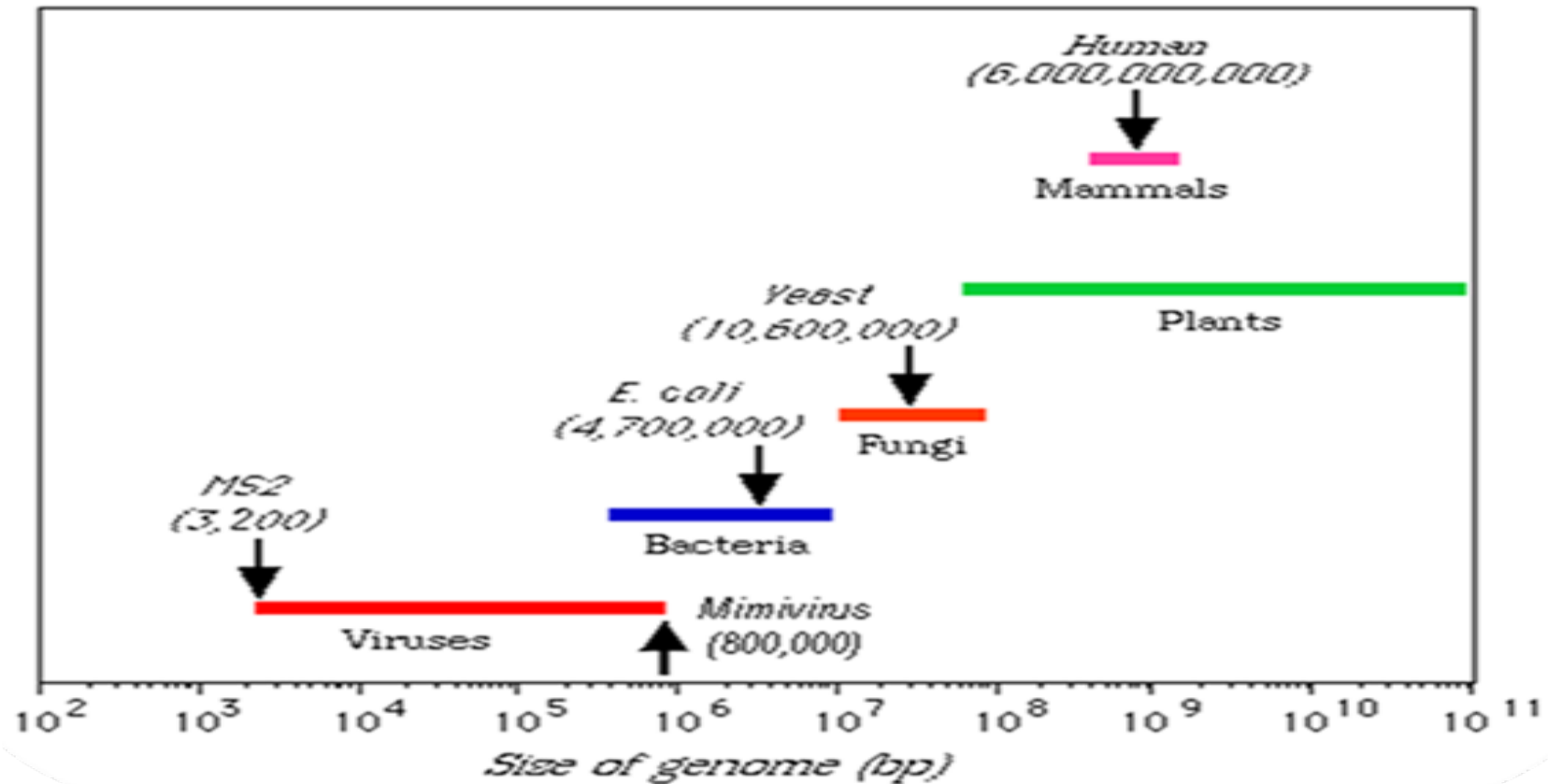
DNA structure :



Introduction to molecular biology

DNA and Genomes :

Comparison of Genome Size:



Introduction to molecular biology

The genetic information storage problem:

If you stretched the DNA in **one** human cell all the way out, it would be about **2m** long and all the DNA in all your cells put together would be about twice the diameter of the Solar System.

One of your cells has, on the average, a 10–100 μm diameter.

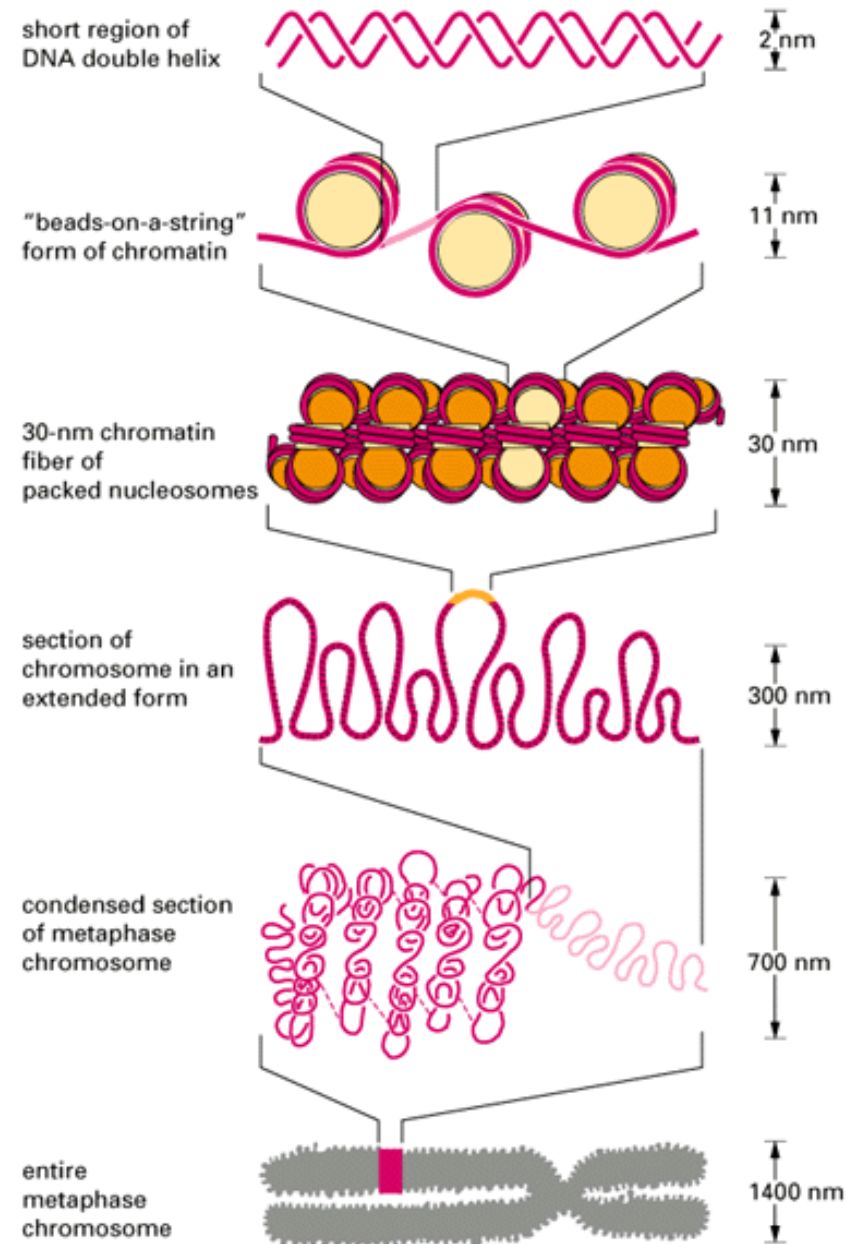
How can be a 2m long molecule into a cell with a diameter of 100 μm (at best)?

Introduction to molecular biology

The genetic information storage problem:

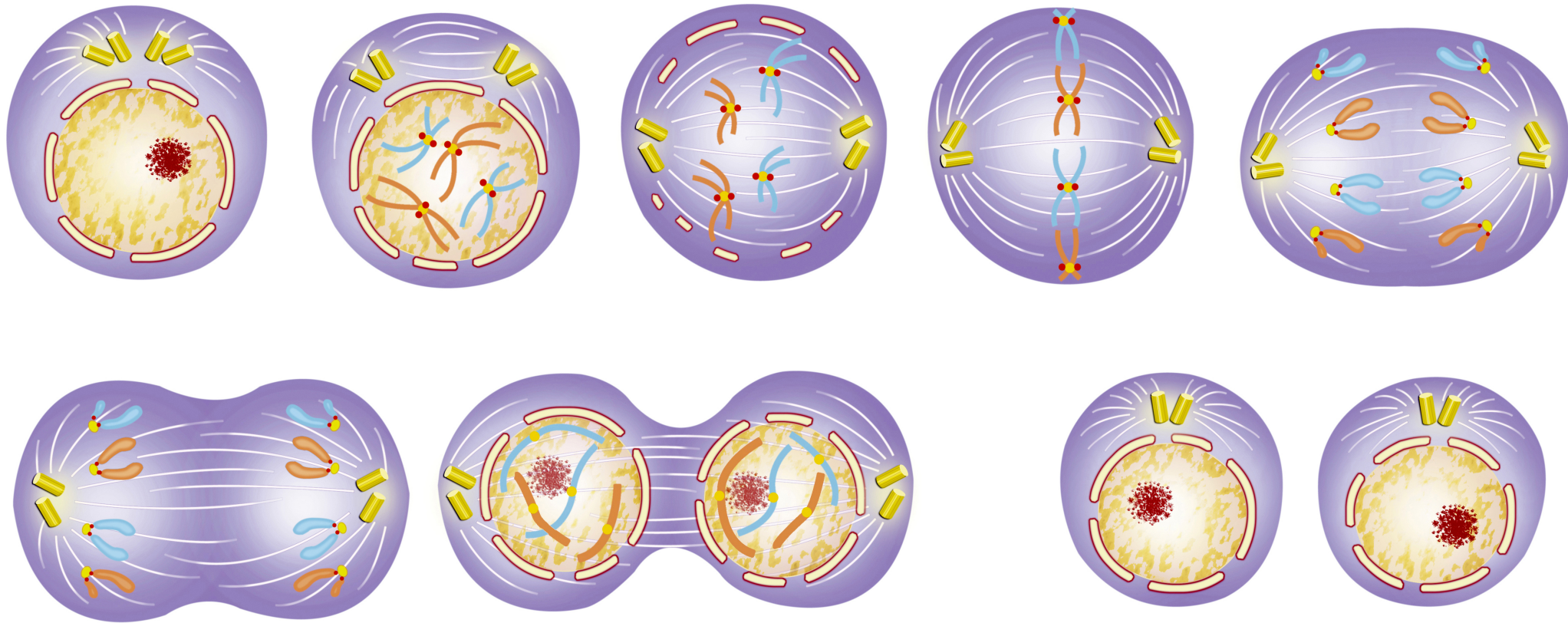
Chromatine

- Eukaryotic genomes are packaged with chromatin proteins
- Heterochromatin (highly condensed, untranscribed)
- Euchromatin (more accessible, transcribed)
- Each cell: unique pattern of heterochromatin and euchromatin



Introduction to molecular biology

DNA state is not constant :



(cell cycle overview)

Introduction to molecular biology

Overview. The key roles of cell division :

- The ability of organisms to produce more of their own kind best distinguishes living things from nonliving matter
- The continuity of life is based on the reproduction of cells, or **cell division**

Introduction to molecular biology

Overview. The key roles of cell division :

- In **unicellular** organisms, division of one cell reproduces the entire organism
- **Multicellular** organisms depend on cell division for
 - **Development** from a fertilized cell
 - **Growth**
 - **Repair**
- Cell division is an integral part of the **cell cycle**, the life of a cell from formation to its own division

Introduction to molecular biology

Most cell division results in genetically identical daughter cells

Most cell division results in daughter cells with **identical** genetic information, **DNA**

The exception is **meiosis**, a special type of division that can produce sperm and egg cells

Introduction to molecular biology

Cellular organization of the genetic material

All the DNA in a cell constitutes the cell's **genome**

A genome can consist of a single DNA molecule (common in prokaryotic cells) or a number of DNA molecules (common in eukaryotic cells)

DNA molecules in a cell are packaged into **chromosomes**

Introduction to molecular biology

Cellular organization of the genetic material

Eukaryotic chromosomes consist of **chromatin**, a complex of DNA and protein that **condenses** during cell division.

Every eukaryotic species has a characteristic number of chromosomes in each cell nucleus

Somatic cells (nonreproductive cells) have two sets of chromosomes

Gametes (reproductive cells: sperm and eggs) have half as many chromosomes as somatic cells

Introduction to molecular biology

Distribution of chromosomes during eukaryotic cell division

In preparation for cell division, **DNA is replicated** and the chromosomes condense

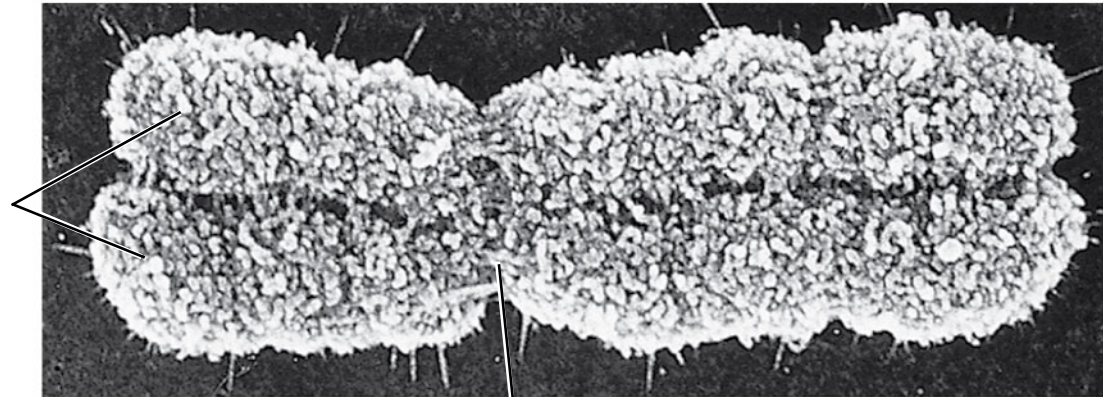
Each duplicated chromosome has two **sister chromatids** (joined copies of the original chromosome), which separate during cell division

The **centromere** is the narrow “waist” of the duplicated chromosome, where the two chromatids are most closely attached

Introduction to molecular biology

Distribution of chromosomes during eukaryotic cells division

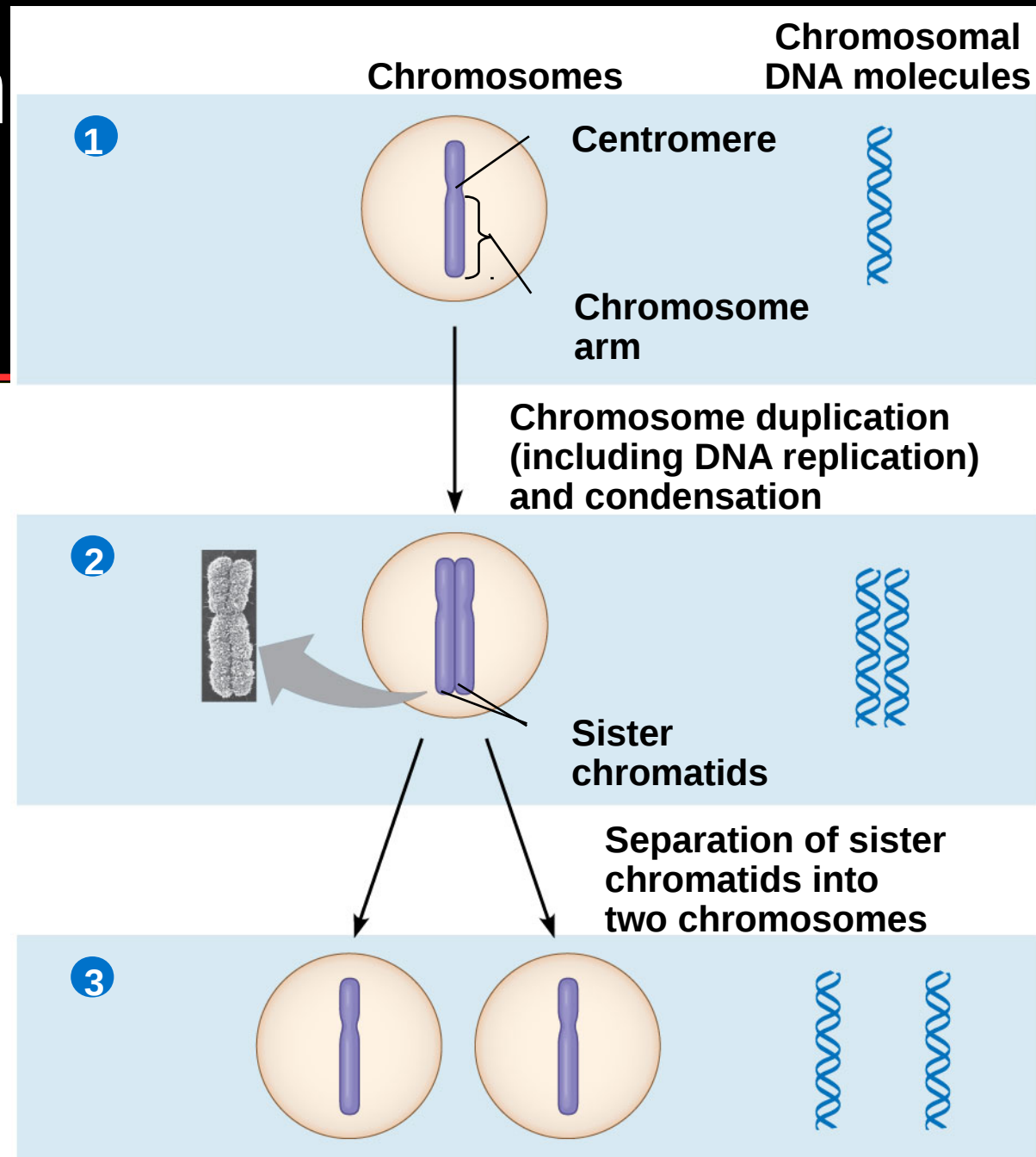
Sister chromatids



Centromere

0.5 μm

Introduction molecular biology



Introduction to molecular biology

Eukaryotic cell division

- Eukaryotic cell division consists of
 - **Mitosis**, the division of the genetic material in the nucleus
 - **Cytokinesis**, the division of the cytoplasm
- Gametes are produced by a variation of cell division called **meiosis**
- Meiosis yields nonidentical daughter cells that have only one set of chromosomes, half as many as the parent cell

Introduction to molecular biology

Phases of the cell cycle

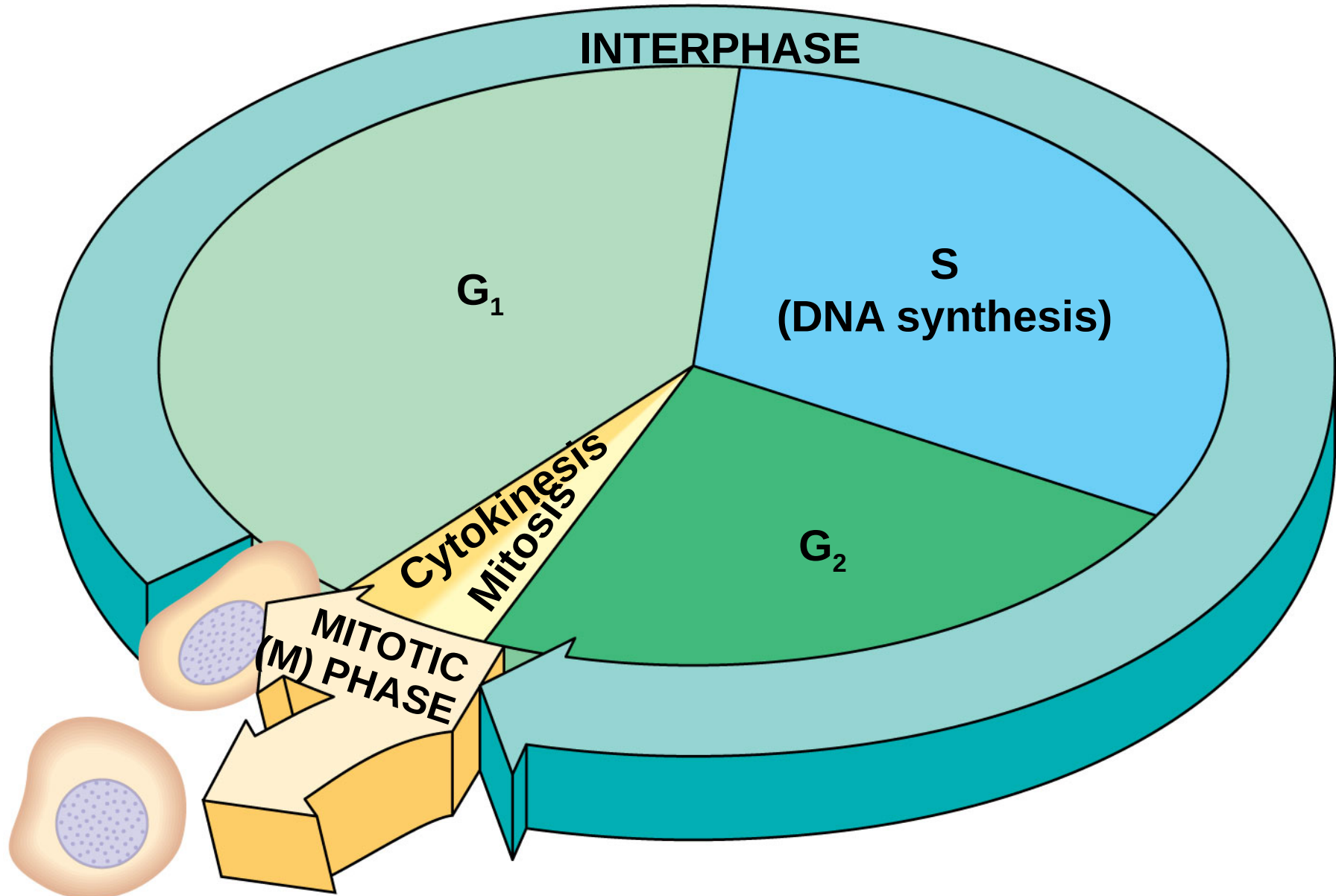
- The cell cycle consists of
 - **Mitotic (M) phase** (mitosis and cytokinesis)
 - **Interphase** (cell growth and copying of chromosomes in preparation for cell division)

Introduction to molecular biology

Phases of the cell cycle

- Interphase (about **90%** of the cell cycle) can be divided into subphases
 - **G₁ phase** (“first gap”)
 - **S phase** (“synthesis”)
 - **G₂ phase** (“second gap”)
- The cell grows during all three phases, but chromosomes are duplicated only during the S phase

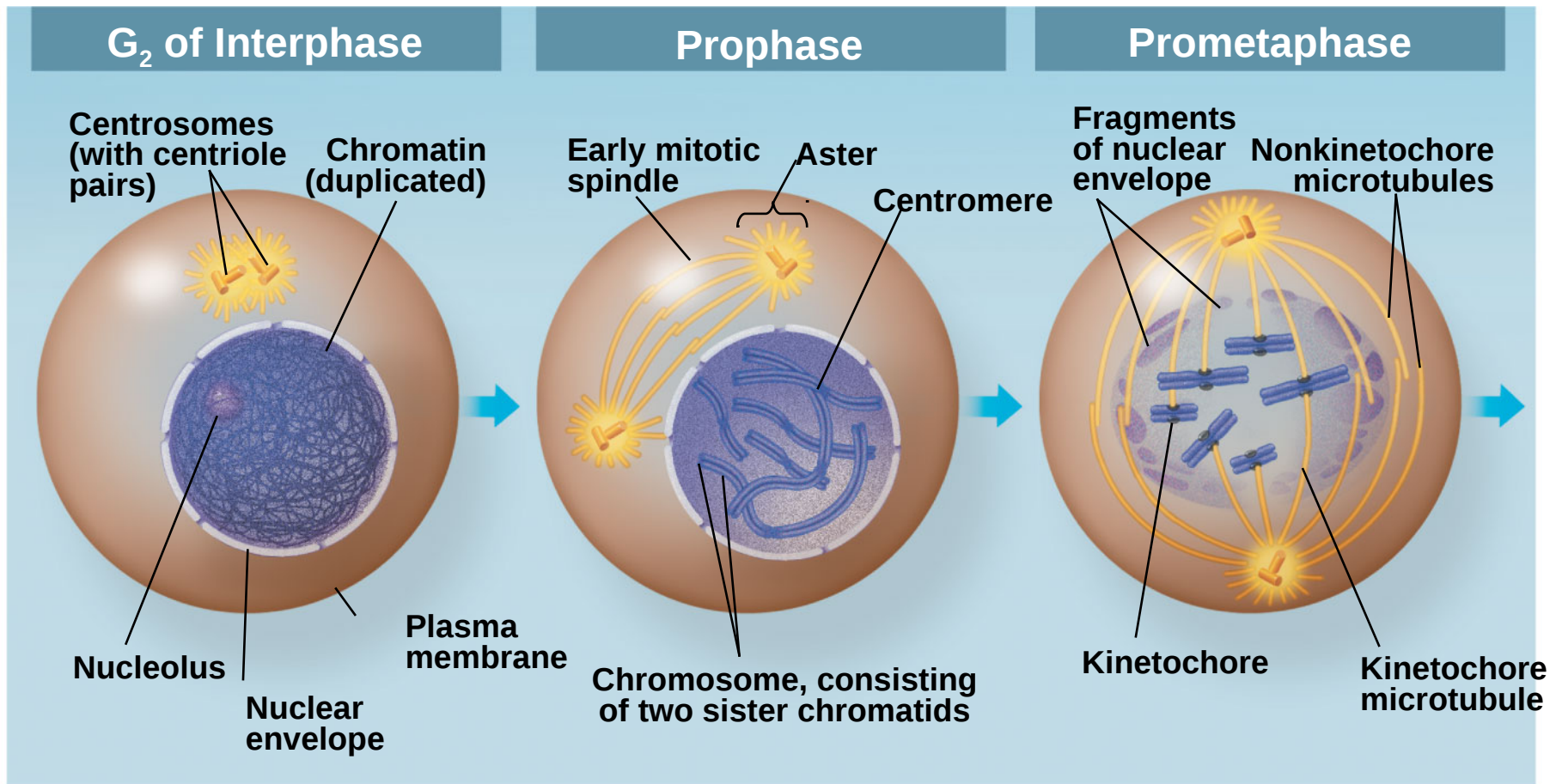
Introduction to molecular biology



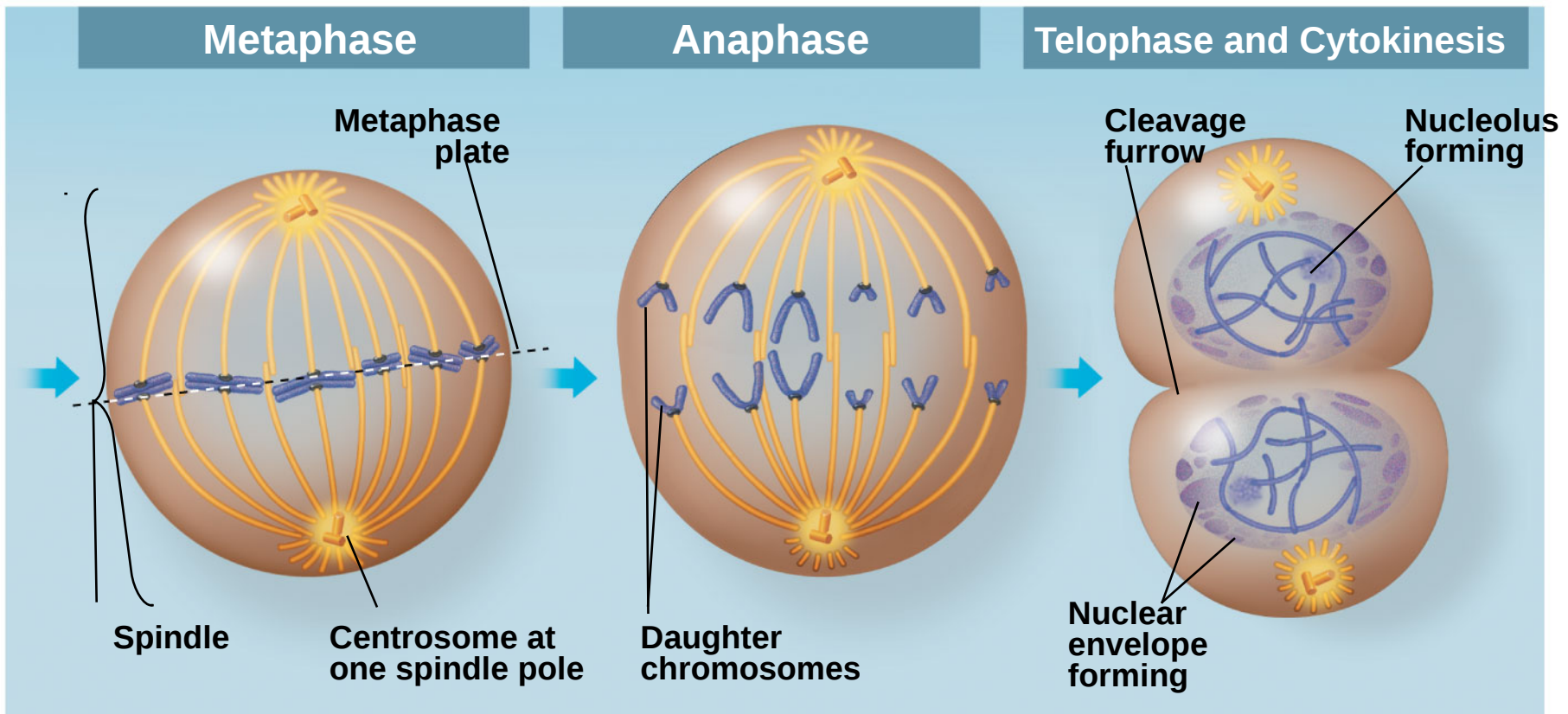
Introduction to molecular biology

- Mitosis is conventionally divided into five phases
 - **Prophase**
 - **Prometaphase**
 - **Metaphase**
 - **Anaphase**
 - **Telophase**
- Cytokinesis overlaps the latter stages of mitosis

Introduction to molecular biology



Introduction to molecular biology

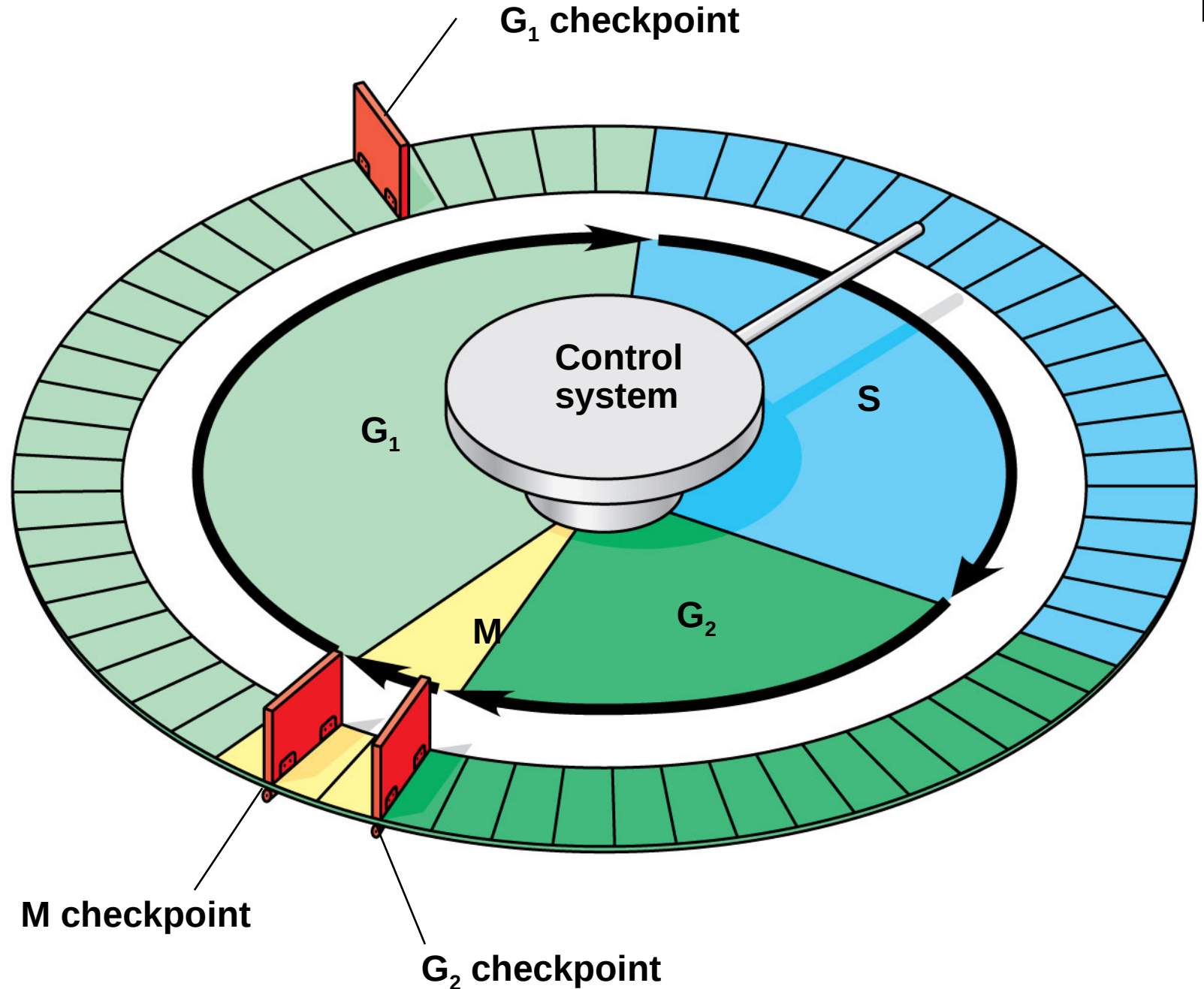


Introduction to molecular biology

The eukaryotic cell cycle is regulated by a molecular control system

- The **frequency** of cell division varies with the type of cell
- These differences result from **regulation at the molecular level**
- Cancer cells manage to escape the usual controls on the cell cycle

Introduction to molecular biology

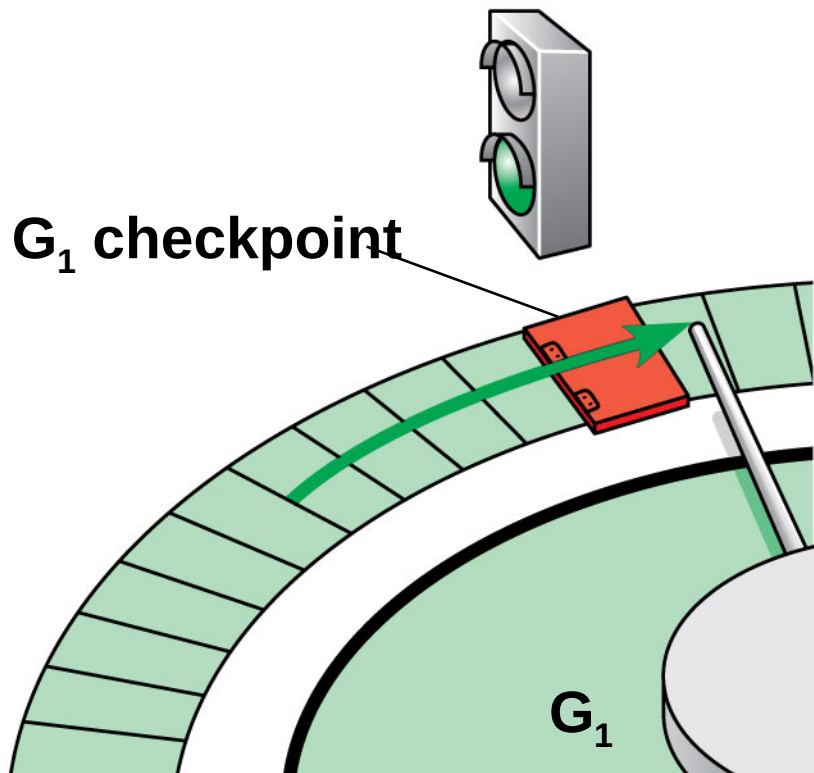


Introduction to molecular biology

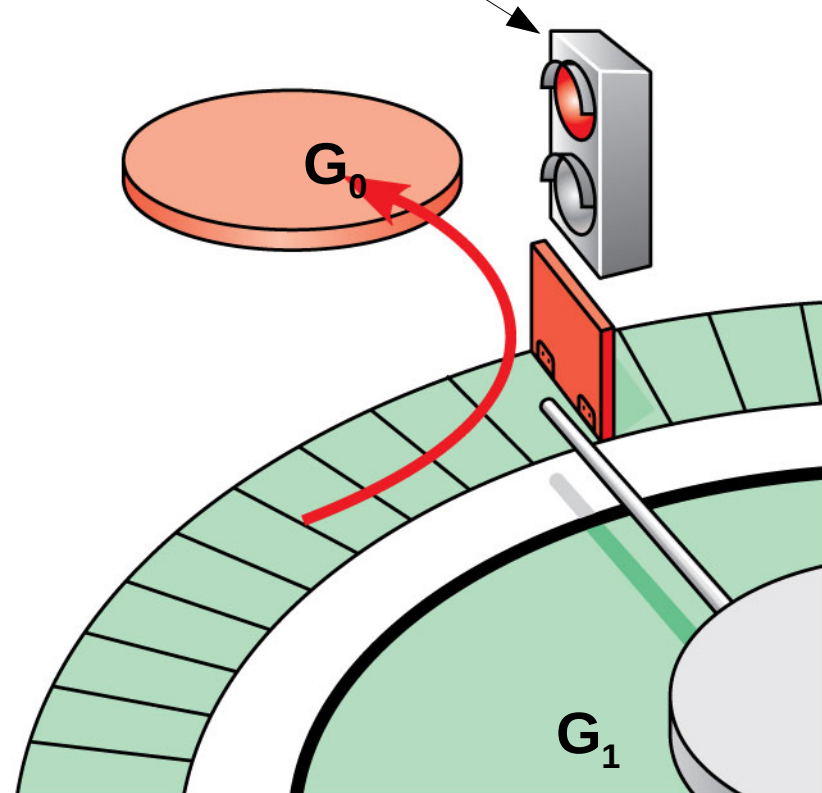
- For many cells, the G_1 checkpoint seems to be the most important
- If a cell receives a go-ahead signal at the G_1 checkpoint, it will usually complete the S, G_2 , and M phases and divide
- If the cell does not receive the go-ahead signal, it will exit the cycle, switching into a nondividing state called the **G_0 phase**

Introduction to molecular biology

Semaphore activated by DNA replication errors / lesions



(a) Cell receives a go-ahead signal.



(b) Cell does not receive a go-ahead signal.

Introduction to molecular biology

The genetic information flow:

A cell is composed by many types of molecules. Each of them can be roughly divided into two broad categories: molecules that must be taken from the environment and molecules that can be synthesized by the cell itself (i.e. proteins).

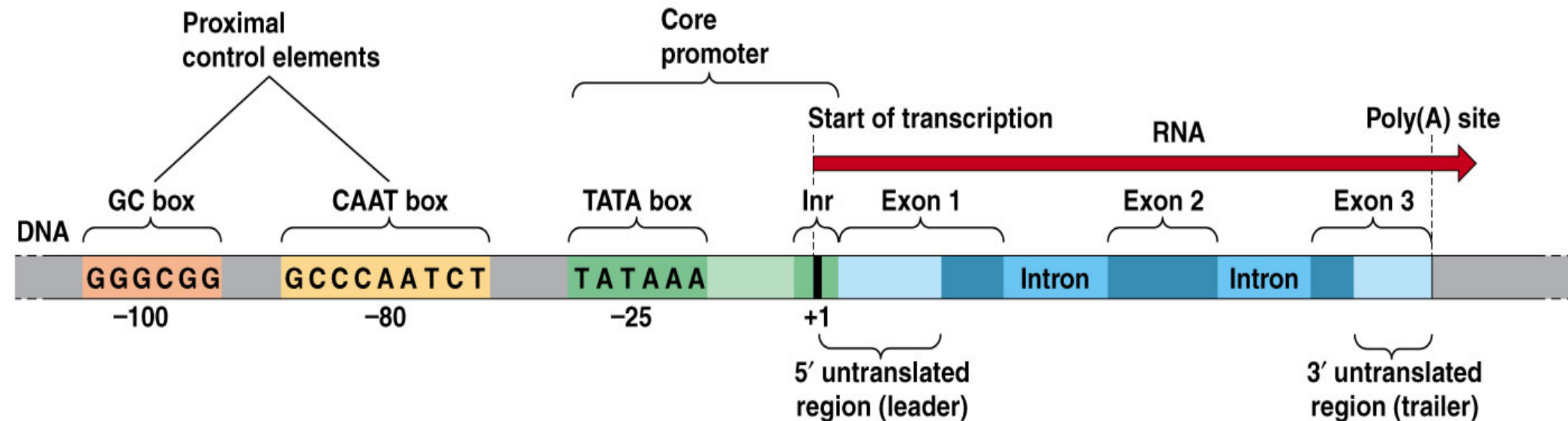
The cell is built using proteins. These molecules can have a structural role (components of walls able to divide the cell in compartments) or can be used to promote many chemical reactions if they belongs to the class of proteins called enzymes.

All the “source code” required to synthesize proteins (and other types of molecules) is stored in DNA.

How is this information read (and used)?

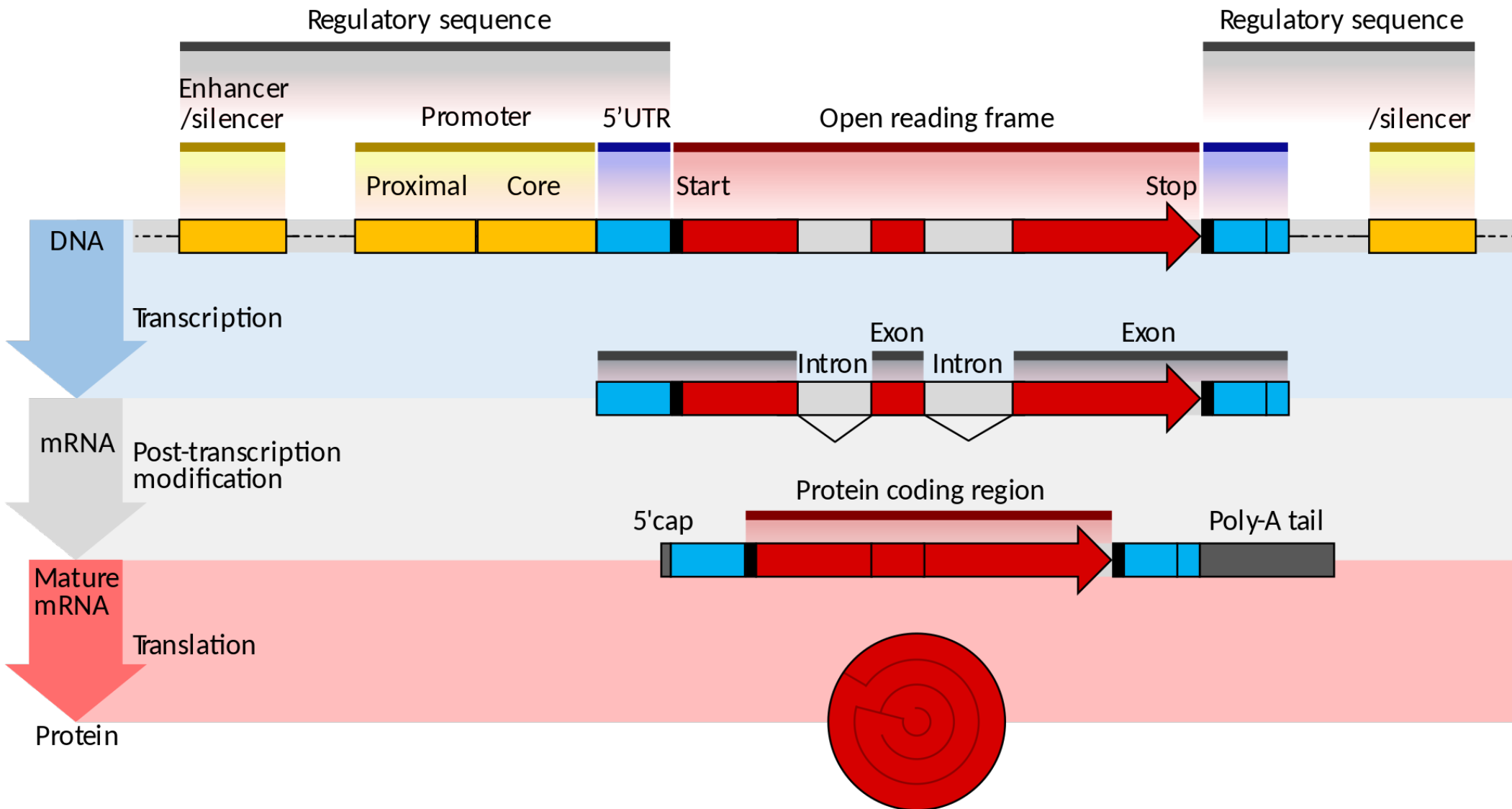
Introduction to molecular biology

The smallest information unit: the gene



Introduction to molecular biology

From DNA to proteins → information flow



Introduction to molecular biology

Gene expression (contents)

- Gene Expression
- The Gene Structure
- Transcription
- Genetic Code and Protein Synthesis
- Regulation of Gene Expression
- Prokaryotes Vs Eukaryotes
- Gene Expression Analysis

Introduction to molecular biology

Gene expression (definition)

- The process by which a gene's information is converted into the structures and functions of a cell by a process that produce a biologically functional protein or RNA molecule (gene products).
- Gene expression is assumed to be controlled at various points in the sequence of events leading to RNA/protein synthesis.

-

Introduction to molecular biology

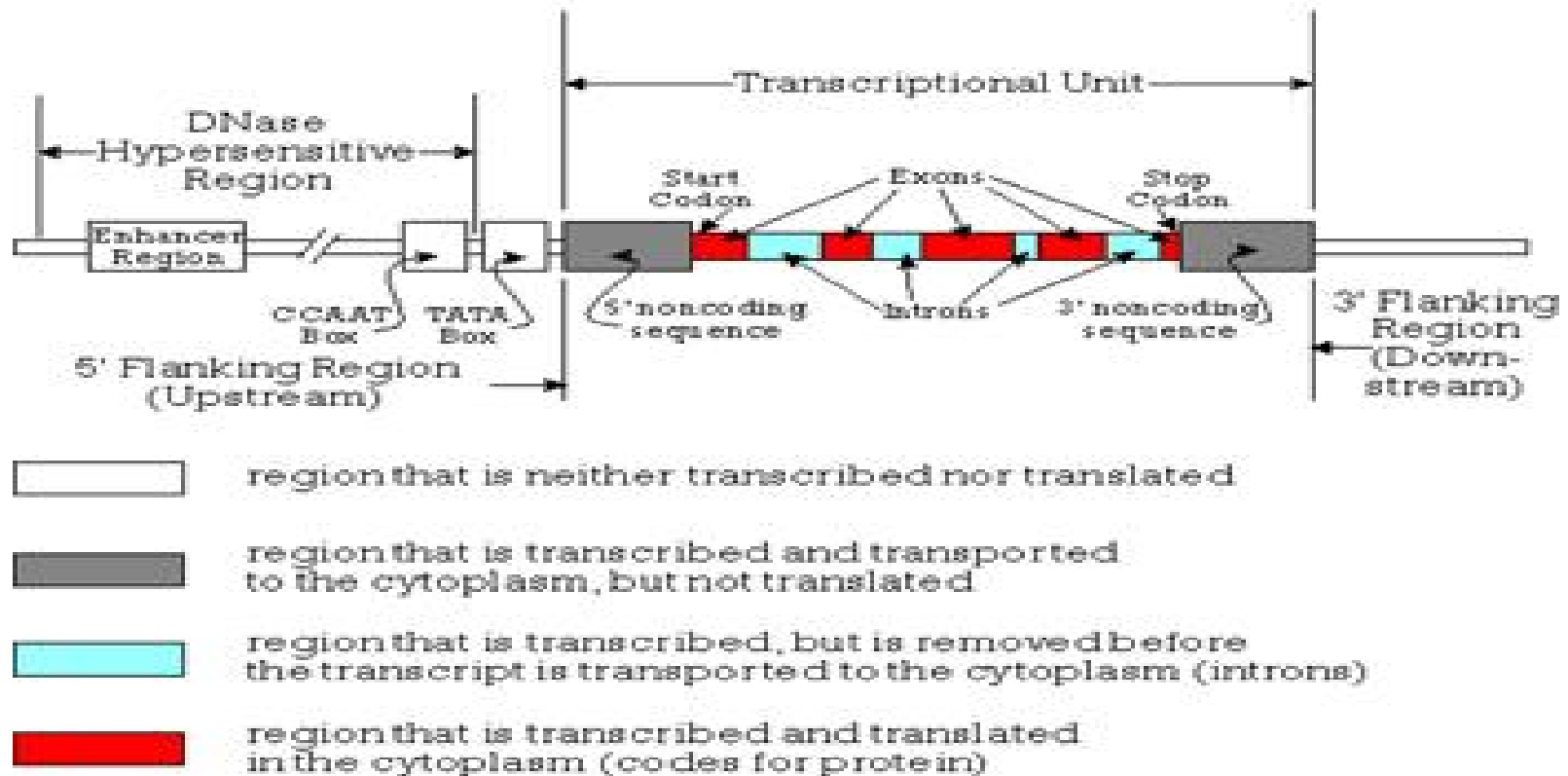
(eukaryotic) Gene structure

- Eukaryotic gene structure: in most eukaryotic genes, in contrast to typical bacterial genes, the coding sequences (**exons**) are interrupted by noncoding DNA (**introns**). The gene must have some functional regions (exon(s); start signals; stop signals; regulatory control elements).
- In the average eukaryotic gene 7-10 exons spread over 10-16kb of DNA.

Introduction to molecular biology

Gene structure model

Typical Model for a Eukaryotic Gene
(not all genes have all parts)



Introduction to molecular biology

From DNA to Protein synthesis: (at least) four stages

- Transcription
- RNA processing
- Translation
- Post-translation processing

Introduction to molecular biology

Gene expression

Transcription

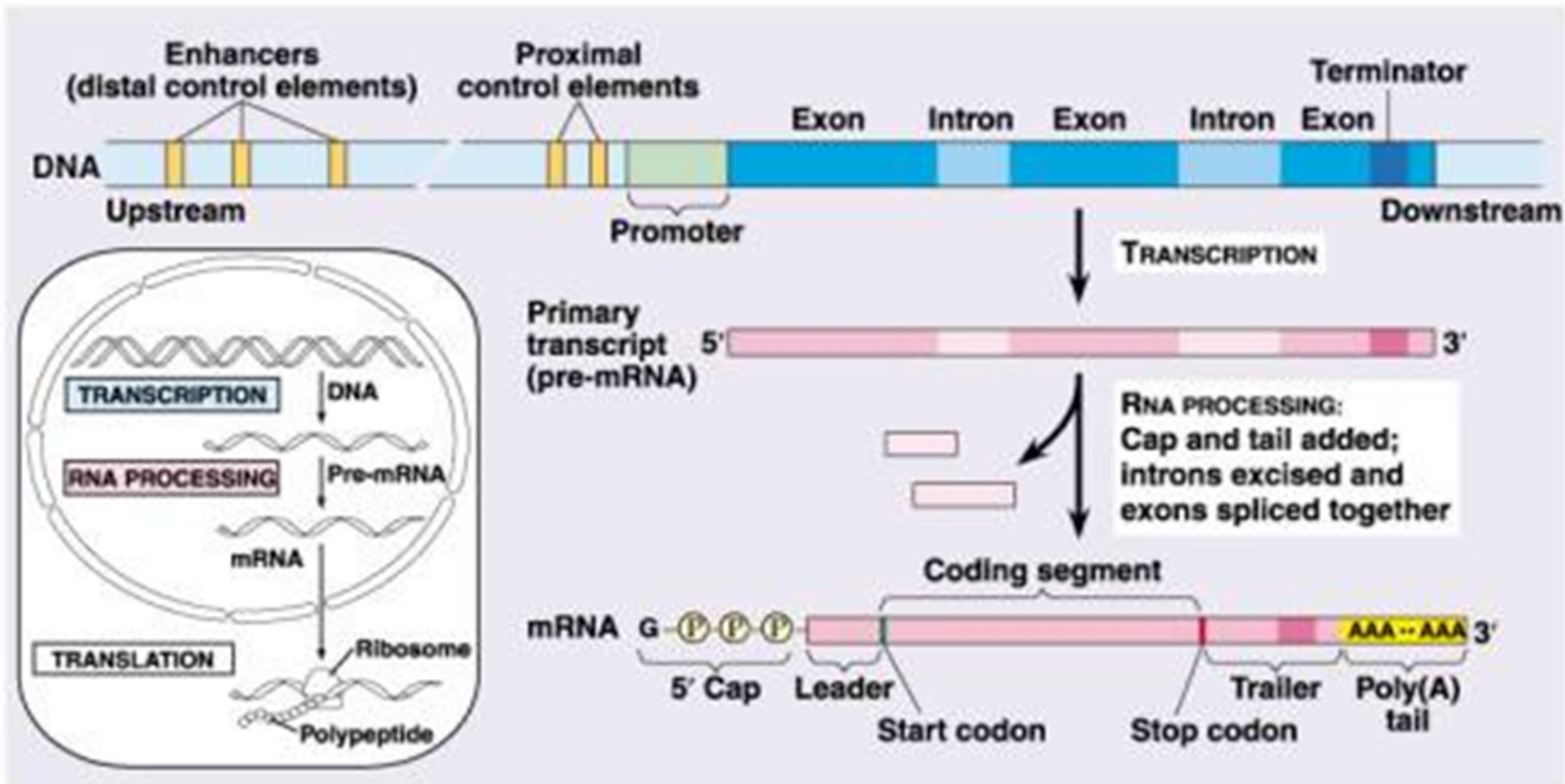
- Synthesis of an RNA that is complementary to one of the strands of DNA.

Translation

- Ribosomes read a messenger RNA (mRNA) and make protein according to its instruction.

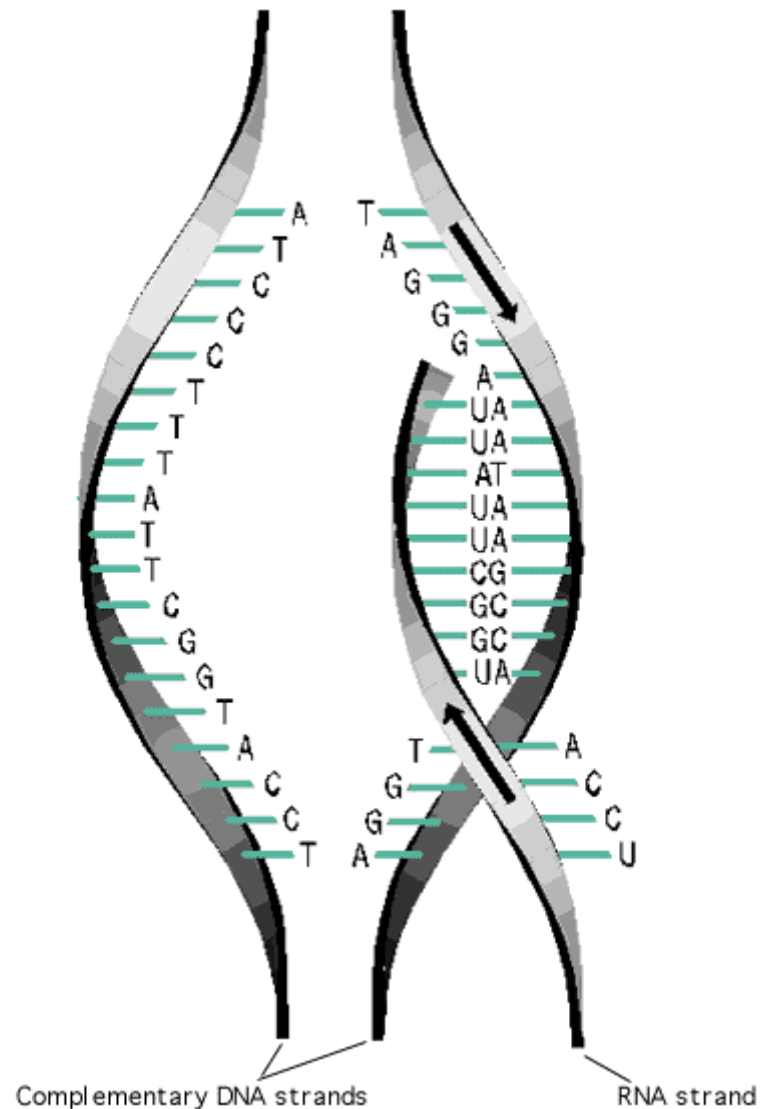
Introduction to molecular biology

Protein synthesis overview



Introduction to molecular biology

Step I : gene transcription



Introduction to molecular biology

Proteins (enzymes) involved in gene transcription

RNA polymerase: The enzyme that controls transcription and:

- Search DNA for initiation site,
- Unwinds a short stretch of double helical DNA to produce a single-stranded DNA template,
- Selects the correct ribonucleotide and catalyzes the formation of a phosphodiester bond,
- Detects termination signals where transcript ends.

Introduction to molecular biology

Eukaryotic RNA polymerases have different roles in transcription

<u>Polymerase I</u>	<u>nucleolus</u>	Makes a large precursor to the major rRNA (5.8S, 18S and 28S rRNA in vertebrates)
<u>Polymerase II</u>	<u>nucleoplasm</u>	Synthesizes hnRNAs, which are precursors to mRNAs. It also makes most small nuclear RNAs (snRNAs)
<u>Polymerase III</u>	<u>Nucleoplasm</u>	Makes the precursor to 5SrRNA, the tRNAs and several other small cellular and viral RNAs.

Introduction to molecular biology

Eukaryotic promoter

Eukaryotic Promoter lies upstream of the gene. There are several different types of promoter found in human genome, with different structure and different regulatory properties class/I/II/III.

Conserved eukaryotic promoter elements	Consensus sequence
CAAT box	GGCCAATCT
TATA box	TATAA
GC box	GGGCGG
CAP site	TAC

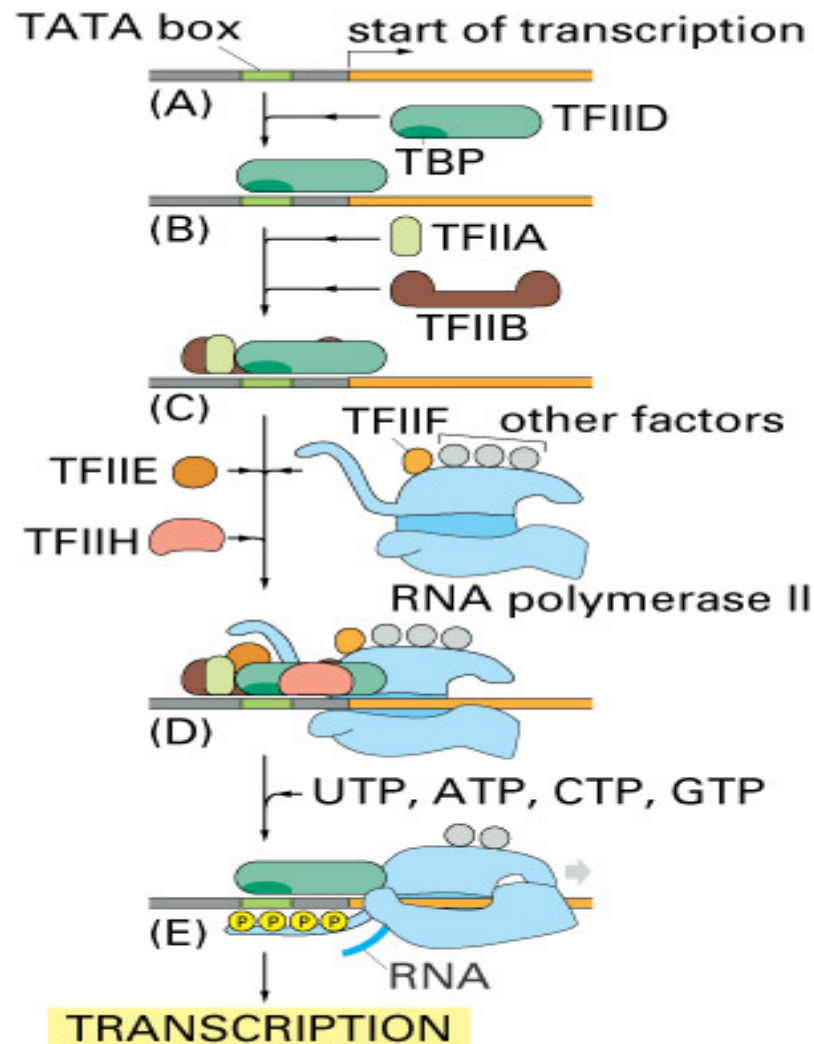
Introduction to molecular biology

Transcription factors

- **Transcription factors** are proteins that bind to DNA **near the start** of transcription of a gene.
- **Transcription factors** either inhibit or assist RNA polymerase in initiation and maintenance of transcription.

Introduction to molecular biology

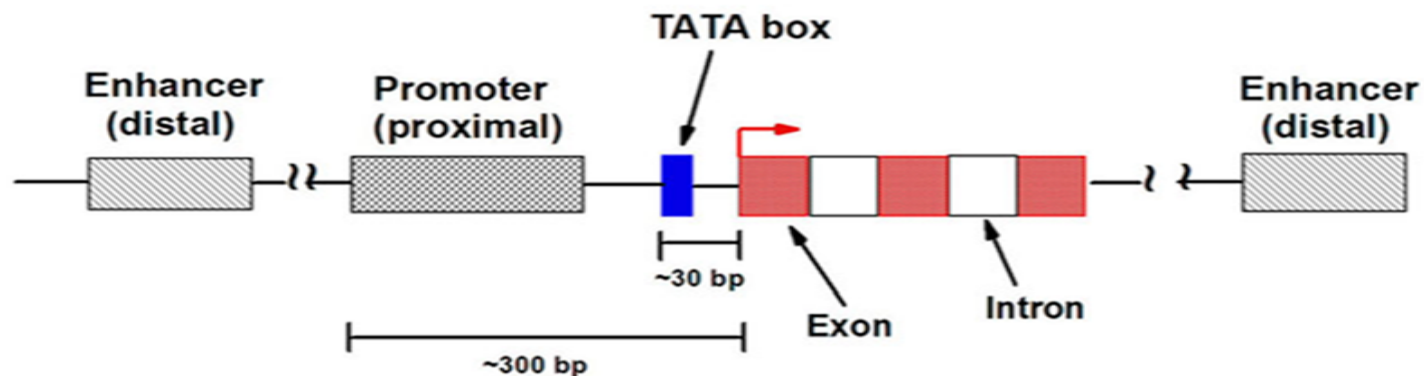
Start of transcription (a cascade of events)



Introduction to molecular biology

Enhancers

Enhancers are stretches of bases within DNA, about 50 to 150 base pairs in length; the activities of many promoters are greatly increased by **enhancers** which can exert their stimulatory actions over distances of several thousands base pairs.



Introduction to molecular biology

Preinitiation complex

- The general transcription factors combine with RNA polymerase to form a preinitiation complex that is competent to initiate transcription as soon as nucleotides are available.
- The assembly of the preinitiation complex on each kind of eukaryotic promoter (class II promoters recognized by RNA polymerase II) begins with the binding of an assembly factor to the promoter.

Introduction to molecular biology

Transcription is divided into three distinct phases:

Initiation

Elongation

Termination

Introduction to molecular biology

Transcription INITIATION

- The polymerase binding causes the unwinding of the DNA double helix which expose at least 12 bases on the template.
- This is followed by initiation of RNA synthesis at this starting point.

Introduction to molecular biology

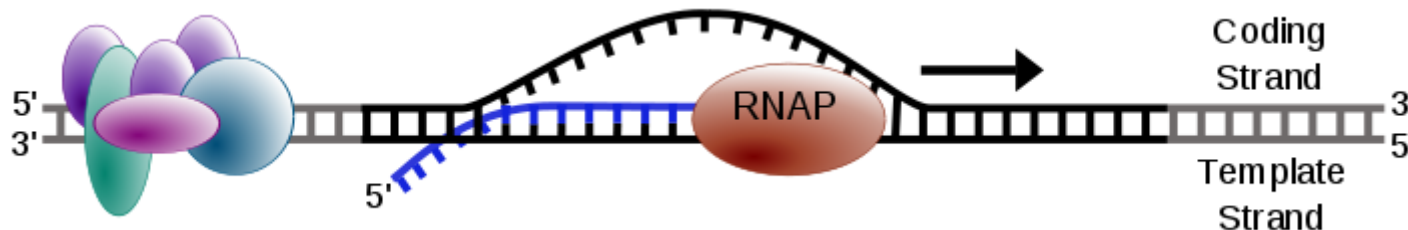
1. Initiation

- The RNA polymerase starts building the RNA chain; it assembles ribonucleotides triphosphates: ATP; GTP; CTP and UTP into a strand of RNA.
- After the **first** nucleotide is in place, the polymerase joins a second nucleotide to the first, forming the initial phosphodiester bond in the RNA chain.

Introduction to molecular biology

2. Elongation

- **RNA polymerase** directs the sequential binding of ribonucleotides to the growing RNA chain in the 5' - 3' direction.
- Each ribonucleotide is inserted into the growing RNA strand following the rules of base pairing. This process is repeated until the desired RNA length is synthesized.....



Introduction to molecular biology

3. Termination

- Terminators at the end of genes; signal termination. These work in conjunction with RNA polymerase to loosen the association between RNA product and DNA template. The result is that the RNA dissociate from RNA polymerase and DNA and so stop transcription.
- The product is **immature** RNA or pre mRNA (Primary transcript).

Introduction to molecular biology

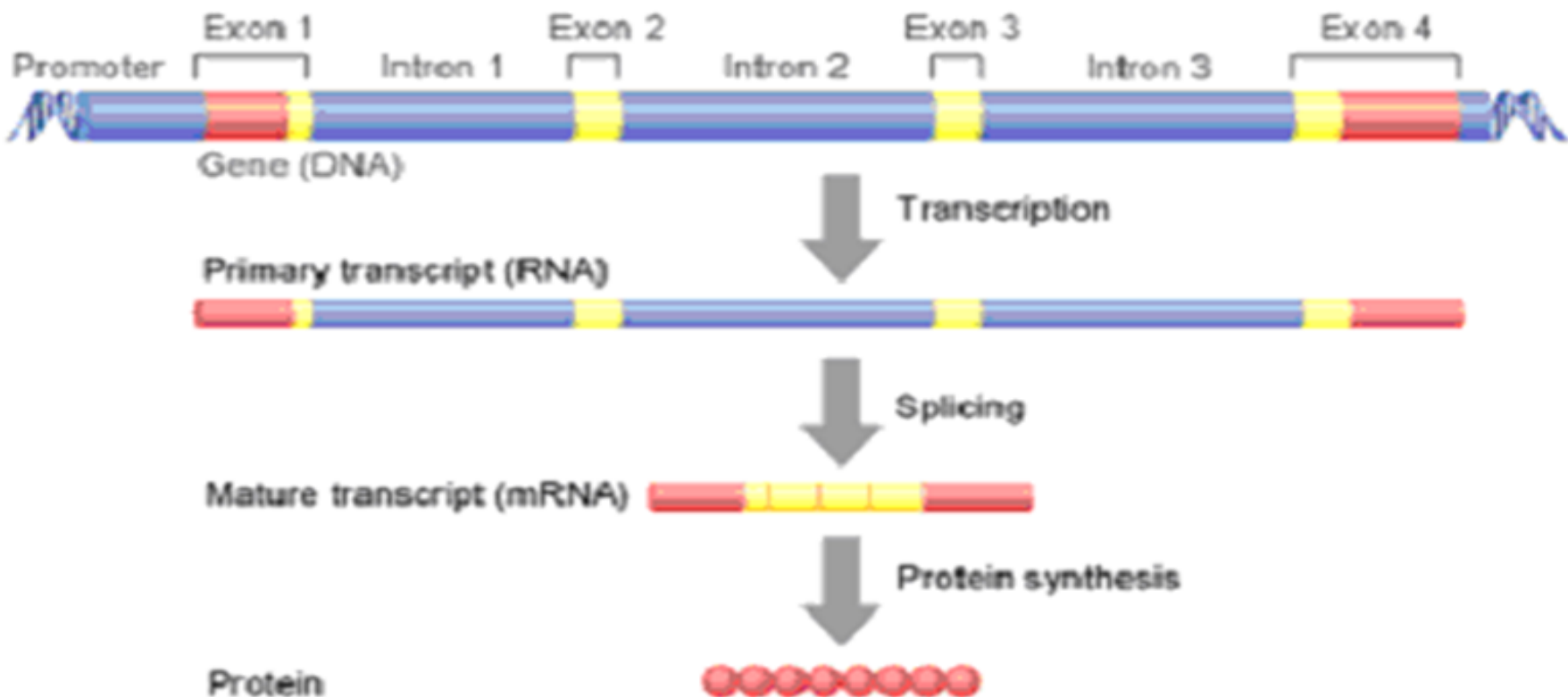
After transcription ends ...

- The primary product of RNA transcription: the hnRNAs contain both intronic and exonic sequences.
- These hnRNAs are **processed in the nucleus to give mature mRNAs** that are transported to the cytoplasm where they participate in protein synthesis.

Introduction to molecular biology

From DNA to protein

Structure of a Gene



Introduction to molecular biology

RNA processing : Pre-mRNA → mRNA (three steps)

- Capping
- Splicing
- Addition of poly A tail

Introduction to molecular biology

RNA processing

- **Capping**
 - The cap structure is added to the 5' of the newly transcribed mRNA precursor in the nucleus prior to processing and subsequent transport of the mRNA molecule to the cytoplasm.
- **Splicing:**
 - Step by step removal of pre mRNA introns and joining of remaining exons; it takes place on a special structure (a protein complex) called spliceosome.

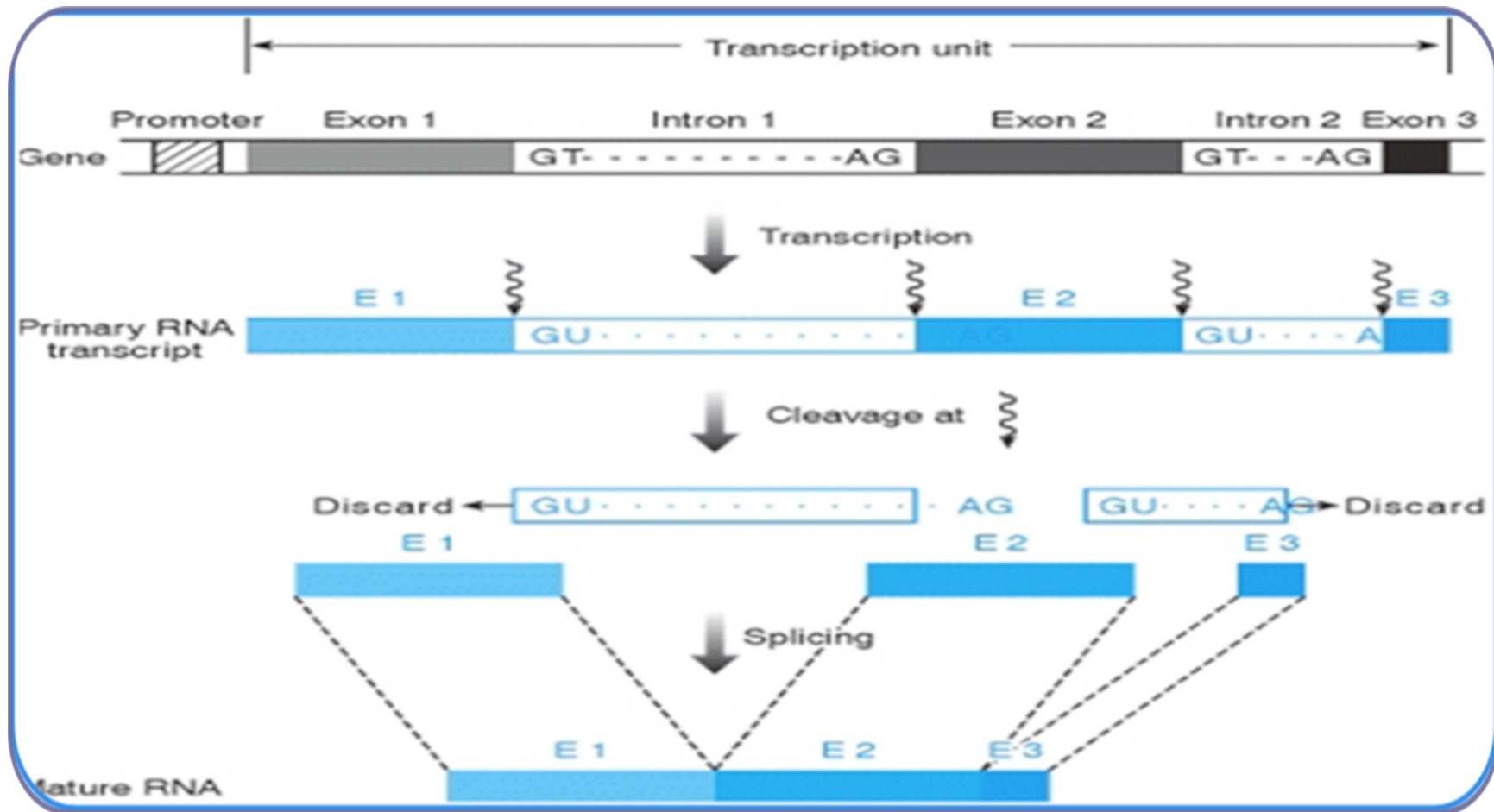
Introduction to molecular biology

RNA processing

- Addition of poly A tail:
 - Synthesis of the poly (A) tail involves cleavage of its 3' end and then the addition of about 40-200 adenine residues to form a poly (A) tail.
 - This is a **timer** ... throughout the life of the mRNA molecule the tail is constantly shortened. When the length of the tail reach a “critically short” length the molecule is destroyed.

Introduction to molecular biology

Alternative splicing

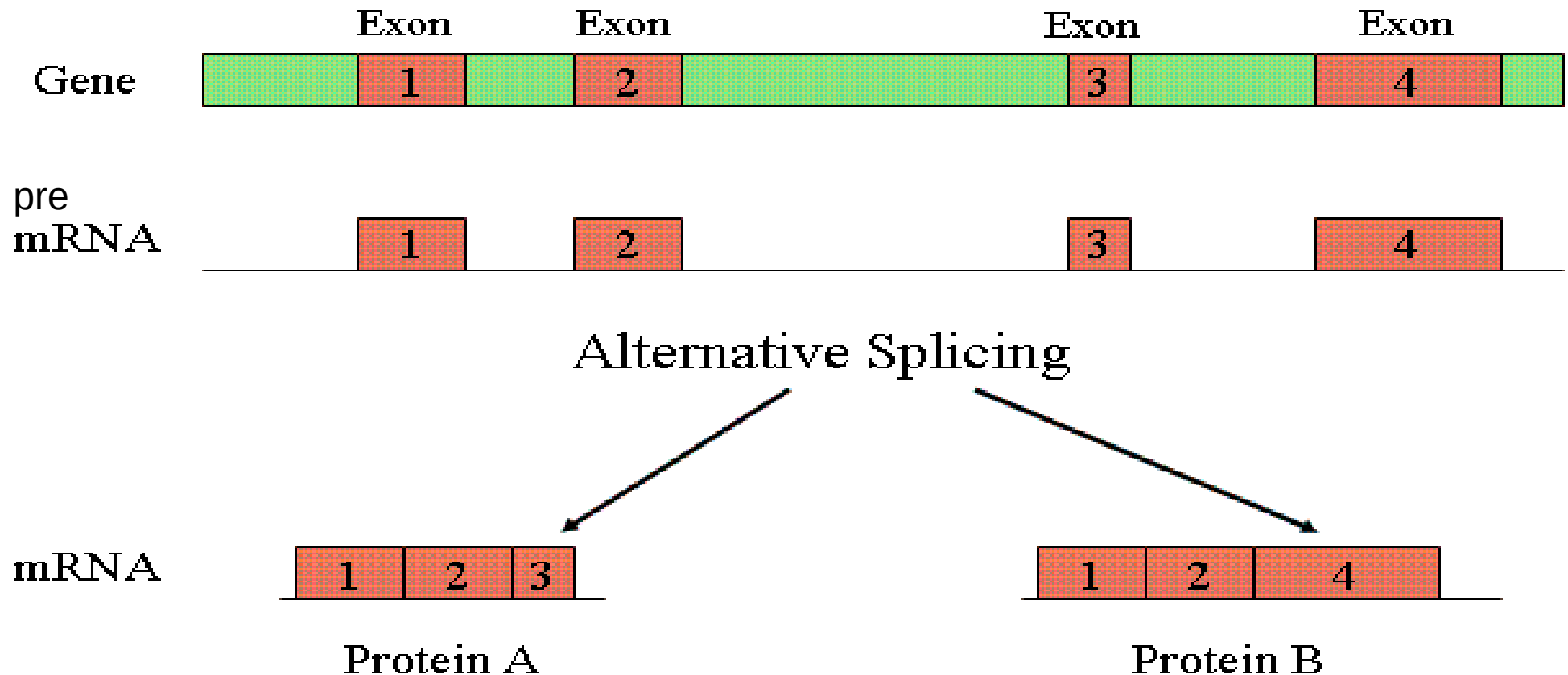


Introduction to molecular biology

Alternative splicing

- **Alternative splicing:** is a very common phenomenon in higher eukaryotes. It is a way to get more than one protein product out of the same gene and a way to control gene expression in cells.

Introduction to molecular biology

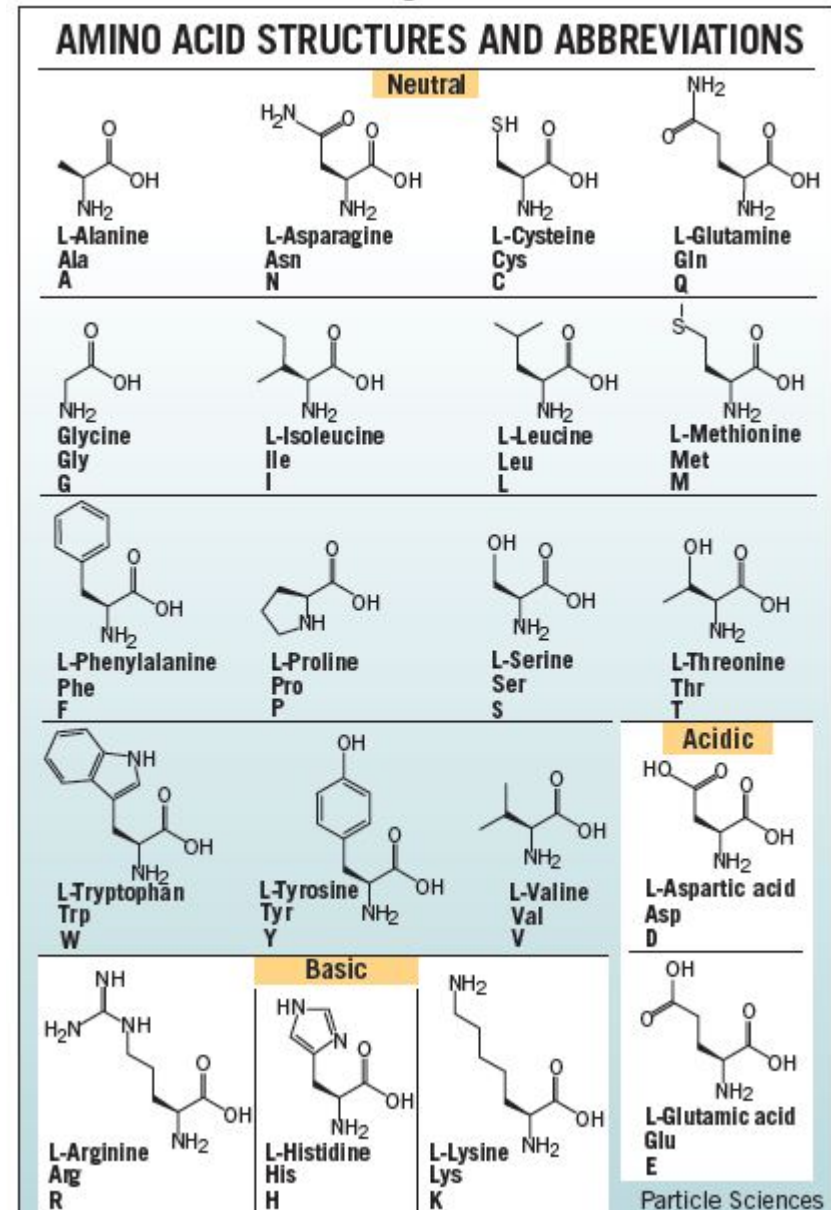


Introduction to molecular biology

Translation: from mRNA to proteins

- Proteins are linear molecules (as we seen for DNA/RNA) but they are not composed by nucleotides (a,c,t,g). Their sequence is generated from an alphabet of 20 amino acids.

Figure 1



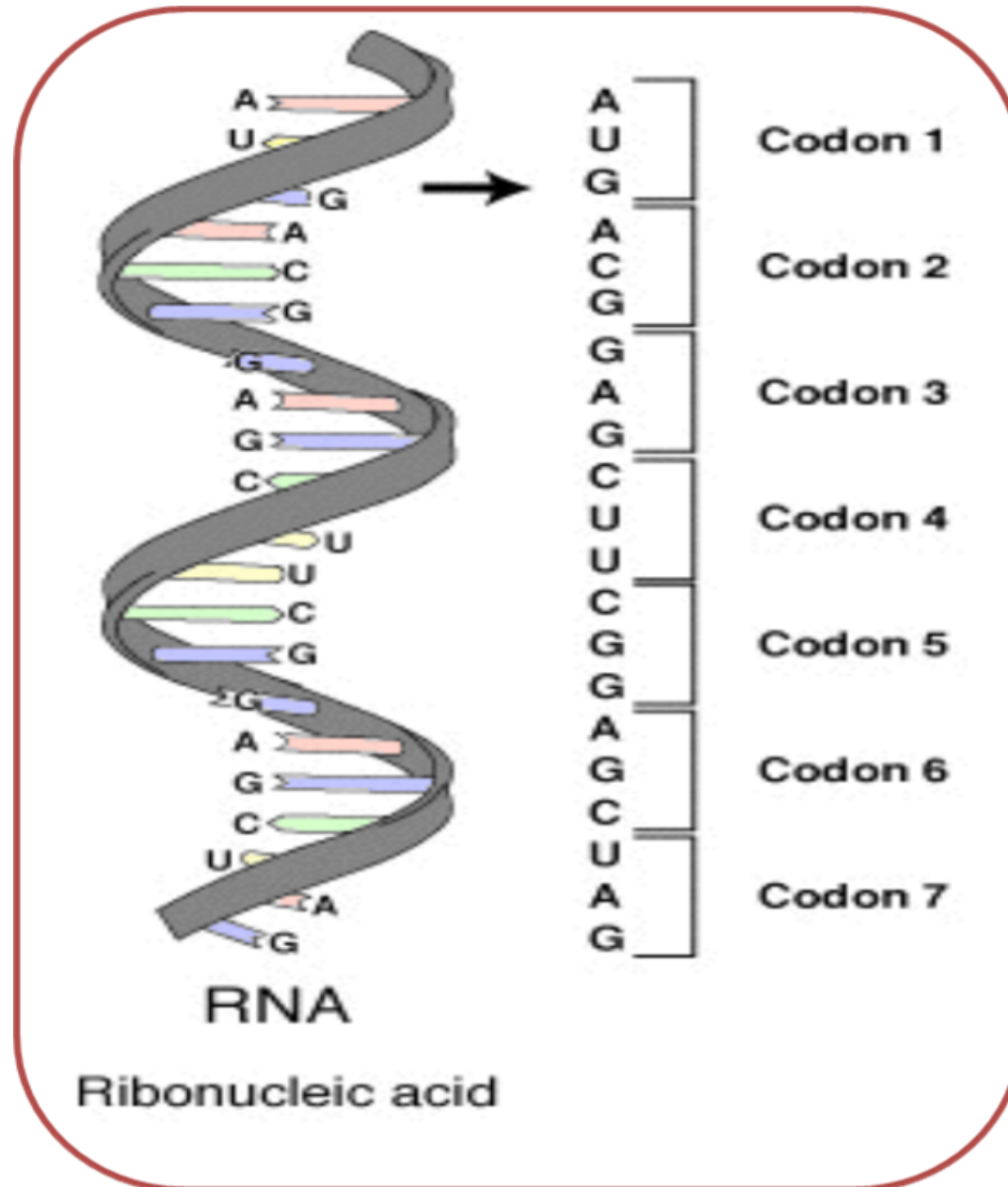
Introduction to molecular biology

The genetic code (how mRNA is read)

- The sequence of codons in the **mRNA** defines the primary structure (the sequence of amino acids) of the final protein.
- **Three** nucleotides in mRNA (a codon) specify **one** amino acid in a protein.

Introduction to molecular biology

NB.
In RNA we
have U instead
of T



Introduction to molecular biology

The genetic code

- **The triplet sequence of mRNA that specify certain amino acid.**
 - 64 different combination of bases; 61 of them code for 20 amino acids (AA); the last three codon (UAG, UGA, UAA) don not code for amino acids; they are **termination** codons.
- **Degenerate**
 - More than one triplet codon specify the same amino acid.

Introduction to molecular biology

The genetic code

■ **Unambiguous**

- Each codon specifies a particular amino acid, the codon ACG codes for the amino acid threonine, **and only** threonine.

■ **Non overlapping**

- This means that successive triplets are read in order. Each nucleotide is part **of only one triplet codon**.

Introduction to molecular biology

The genetic code

		Second Letter							
		T	C	A	G				
First Letter	T	TTT } Phe TTC } TTA } Leu TTG }	TCT } Ser TCC } TCA } TCG }	TAT } Tyr TAC } TAA } Stop TAG } Stop	TGT } Cys TGC } TGA } Stop TGG } Trp	T	C	A	G
	C	CTT } Leu CTC } CTA } CTG }	CCT } Pro CCC } CCA } CCG }	CAT } His CAC } CAA } Gln CAG }	CGT } Arg CGC } CGA } CGG }	T	C	A	G
	A	ATT } Ile ATC } ATA } ATG } Met	ACT } Thr ACC } ACA } ACG }	AAT } Asn AAC } AAA } Lys AAG }	AGT } Ser AGC } AGA } Arg AGG }	T	C	A	G
	G	GTT } Val GTC } GTA } GTG }	GCT } Ala GCC } GCA } GCG }	GAT } Asp GAC } GAA } Glu GAG }	GGT } Gly GGC } GGA } GGG }	T	C	A	G

DNA Codon

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

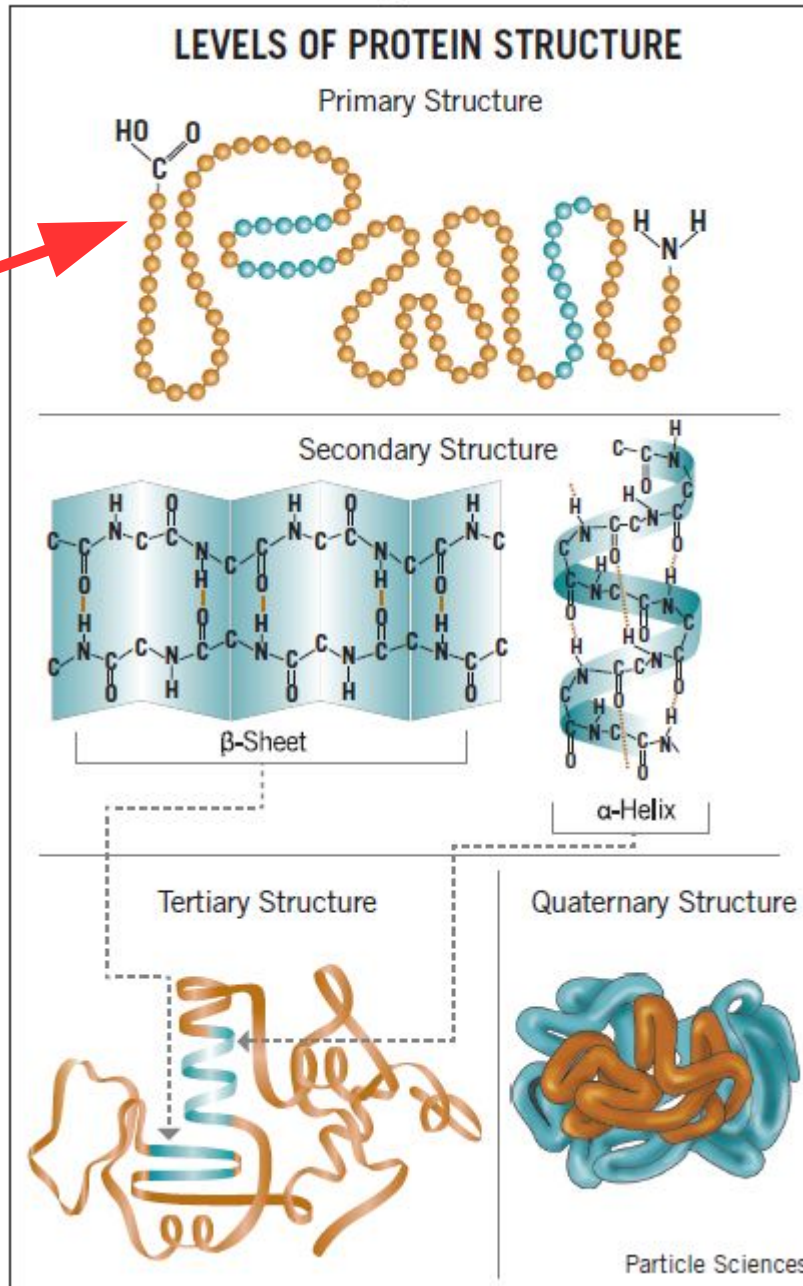
RNA Codon

Introduction to molecular biology

Protein structure levels:

Primary structure (sequence) is encoded directly by RNA codons which (in turn) are encoded directly by the gene DNA sequence

Figure 2



DNA



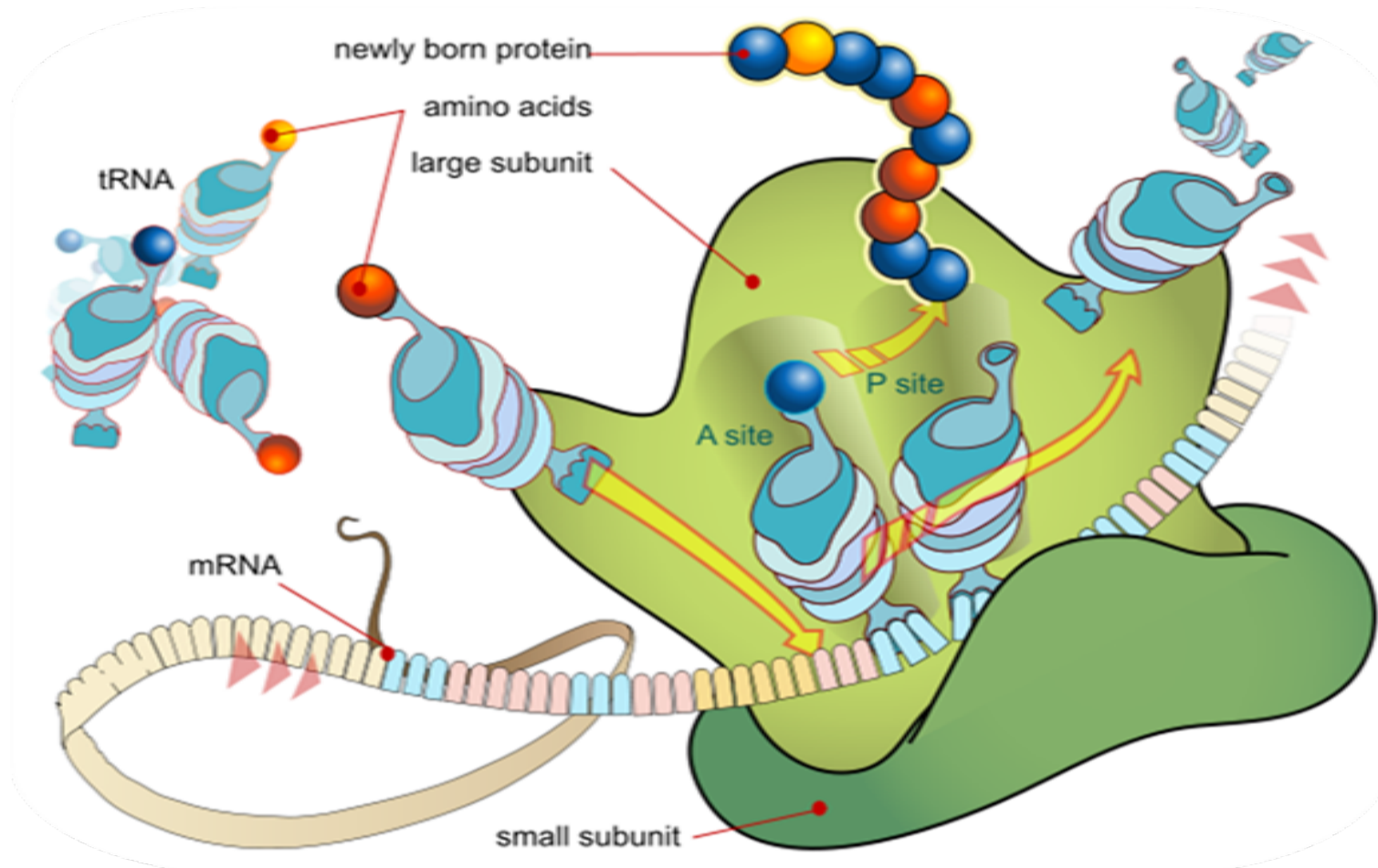
mRNA



Protein

Introduction to molecular biology

Translation



Introduction to molecular biology

Translation

- **Translation** is the process by which ribosomes read the genetic message in the mRNA and produce a protein product according to the message's instruction.

Introduction to molecular biology

Requirements for translation

- **Ribosomes**
- **tRNA**
- **mRNA**
- **Amino acids**
- **Initiation factors**
- **Elongation factors**
- **Termination factors**
- **Aminoacyl tRNA synthetase enzymes:**
- **Energy source:**

Introduction to molecular biology

Ribosomes

- Eukaryotic ribosomes are larger. They consist of **two subunits**, which come together to form an 80S particle;
 - 60S subunit holds (three rRNAs 5S, 5.8S, 28S and about 40 proteins).
 - 40S subunit contains (an 18S rRNA and about 30 proteins).

Introduction to molecular biology

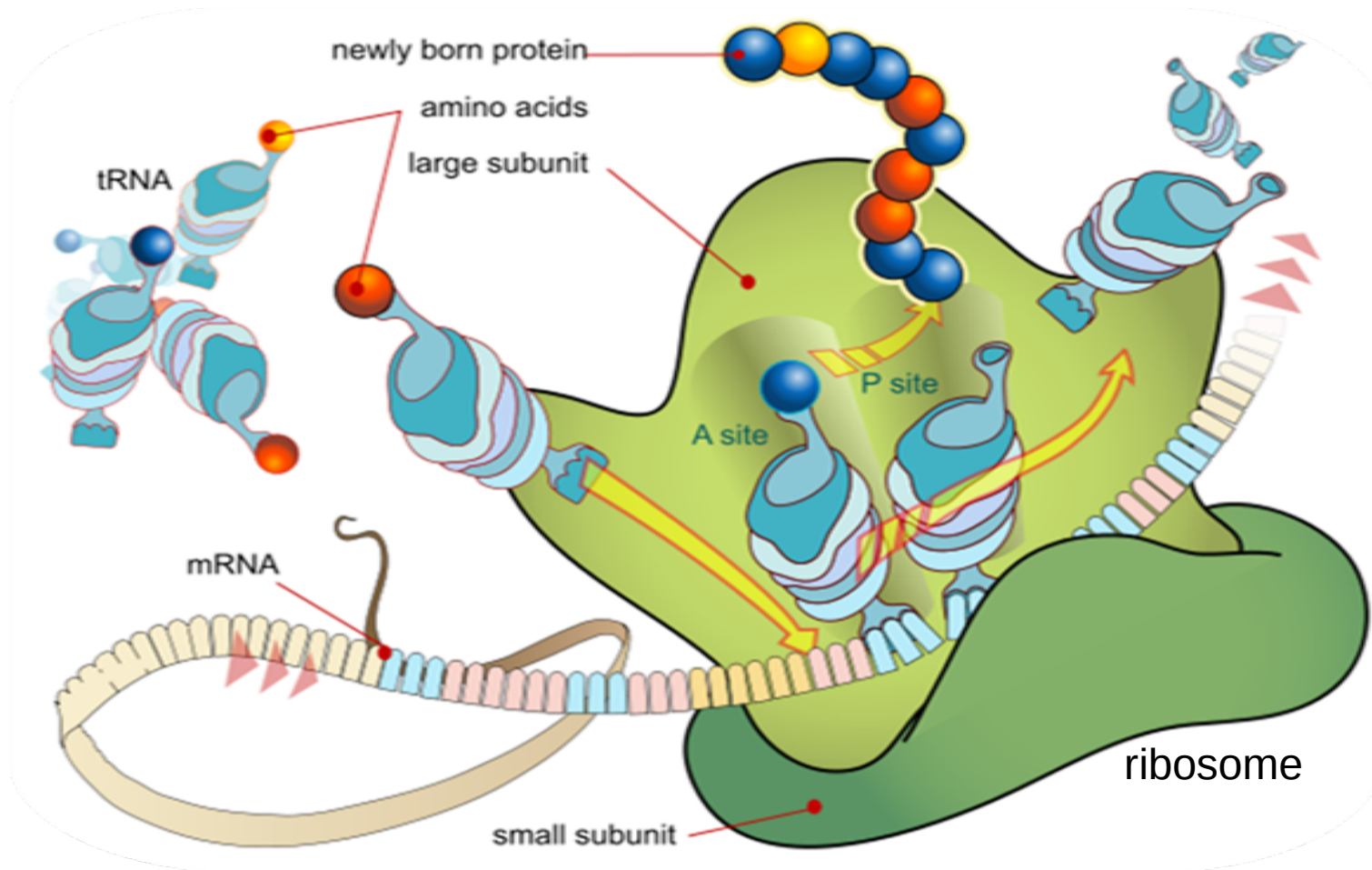
Ribosomes

The large ribosomal subunit contains three tRNA binding sites, designated A, P, and E.

- The A site binds an aminoacyl-tRNA (a tRNA bound to an amino acid);
- P site binds a peptidyl-tRNA (a tRNA bound to the peptide being synthesized).
- The E site binds a free tRNA before it exits the ribosome.

Introduction to molecular biology

Ribosomes



Introduction to molecular biology

Preparatory steps for protein synthesis

- First, aminoacyl tRNA synthetase joins amino acid to their specific tRNA.
- Second, ribosomes must dissociate into subunits at the end of each round of translation.

Introduction to molecular biology

Protein synthesis occur in three phases

- Accurate and efficient initiation occurs; the ribosomes binds to the mRNA, and the first amino acid attached to its tRNA.
- Chain elongation, the ribosomes adds one amino acid at a time to the growing polypeptide chain.
- Accurate and efficient termination, the ribosomes releases the mRNA and the polypeptide.

Introduction to molecular biology

Translation phase I : Initiation

- The initiation phase of protein synthesis requires over 10 eukaryotic Initiation Factors (eIFs): Factors are needed to recognize the cap at the 5' end of an mRNA and binding to the 40s ribosomal subunit.
- Binding the initiator Met-tRNA^{iMet} (methionyl-tRNA) to the 40S small subunit of the ribosome.

Introduction to molecular biology

Translation phase I : initiation

- **Scanning** to find the **start codon** by binding to the 5' cap of the mRNA and scanning downstream until they find the first **AUG** (initiation codon).
- The **start codon** must be located and **positioned** correctly in the **P site** of the ribosome, and the initiator tRNA must be positioned correctly in the same site.
- Once the mRNA and initiator tRNA are correctly bound, the 60S large subunit binds to form 80s initiation complex with a release of the eIF factors.

Introduction to molecular biology

Translation phase II : elongation

■ **Translocation;** translocation of the new peptidyl t-RNA with its mRNA codon in the A site into the free P site occurs. Now the A site is free for another cycle of aminoacyl t-RNA codon recognition and elongation. Each translocation event moves mRNA, one codon length through the ribosomes.

Introduction to molecular biology

Translation phase III : termination

- Translation termination requires specific protein factors identified as releasing factors, RFs in *E. coli* and eRFs in eukaryotes.
- The signals for termination are the same in both prokaryotes and eukaryotes. These signals are **termination codons present in the mRNA**. There are 3 termination codons, UAG, UAA and UGA.

Introduction to molecular biology

Translation phase III : termination

- After multiple cycles of elongation and polymerization of specific amino acids into protein molecules, a **nonsense codon = termination codon** of mRNA appears in the A site. This is recognized as a terminal signal by eukaryotic releasing factors (eRF) which cause the **release of the newly synthesized protein** from the ribosomal complex.

Introduction to molecular biology

Eukaryotic gene expression

- Essentially **all humans' genes contain introns**. A notable exception is the histone genes which are intronless.
- Eukaryote genes are not grouped in operons. Each eukaryote gene is transcribed separately, **with separate transcriptional controls on each gene**.
- Eukaryotic mRNA is modified through RNA splicing.
- Eukaryotic mRNA is generally monogenic (monocistronic); code for **only one polypeptide**.

Introduction to molecular biology

Glossary

- Alleles are forms of the same gene with small differences in their sequence of DNA bases.
- **Alternative splicing**: is a very common phenomenon in higher eukaryotes. It is a way to get more than one protein product out of the same gene and a way to control gene expression in cells.
- Exon: a segment of a gene that is represented in the mature RNA product. Individual exons may contain coding DNA and/or noncoding DNA (untranslated sequences).
- **Bioinformatics** is the application of computer science and information technology to the field of biology and medicine
- Introns (intervening sequence) (A noncoding DNA sequence): Intervening stretches of DNA that separate exons.
- Primary transcript: The initial production of gene transcription in the nucleus; an RNA containing copies of all exons and introns.
- RNA gene or non-coding RNA gene: RNA molecule that is not translated into a protein. Noncoding RNA genes produce transcripts that exert their function without ever producing proteins. Non-coding RNA genes include transfer RNA (tRNA) and ribosomal RNA (rRNA), small RNAs such as snoRNAs, microRNAs, siRNAs and piRNAs and lastly long ncRNAs.
- Enhancers and silencers: are DNA elements that stimulate or depress the transcription of associated genes; they rely on tissue specific binding proteins for their activities; sometimes a DNA elements can act either as an enhancer or silencer depending on what is bound to it.
- Activators: Additional gene-specific transcription factors that can bind to enhancer and help in transcription activation.
- Open reading frame (ORF): A reading frame that is uninterrupted by translation stop codon (reading frame that contains a start codon and the subsequent translated region, but no stop codon).
- Directionality: in molecular biology, refers to the end-to-end chemical orientation of a single strand of nucleic acid. The chemical convention of naming carbon atoms in the nucleotide sugar-ring numerically gives rise to a 5' end and a 3' end ("five prime end" and "three prime end"). The relative positions of structures along a strand of nucleic acid, including genes, transcription factors, and polymerases are usually noted as being either *upstream* (towards the 5' end) or *downstream* (towards the 3' end).
- Reverse Transcription: Some viruses (such as HIV, the cause of AIDS), have the ability to transcribe RNA into DNA.
- Pseudogenes. DNA sequences that closely resemble known genes but are nonfunctional.
- More: <http://www.ncbi.nlm.nih.gov/books/NBK7584/>

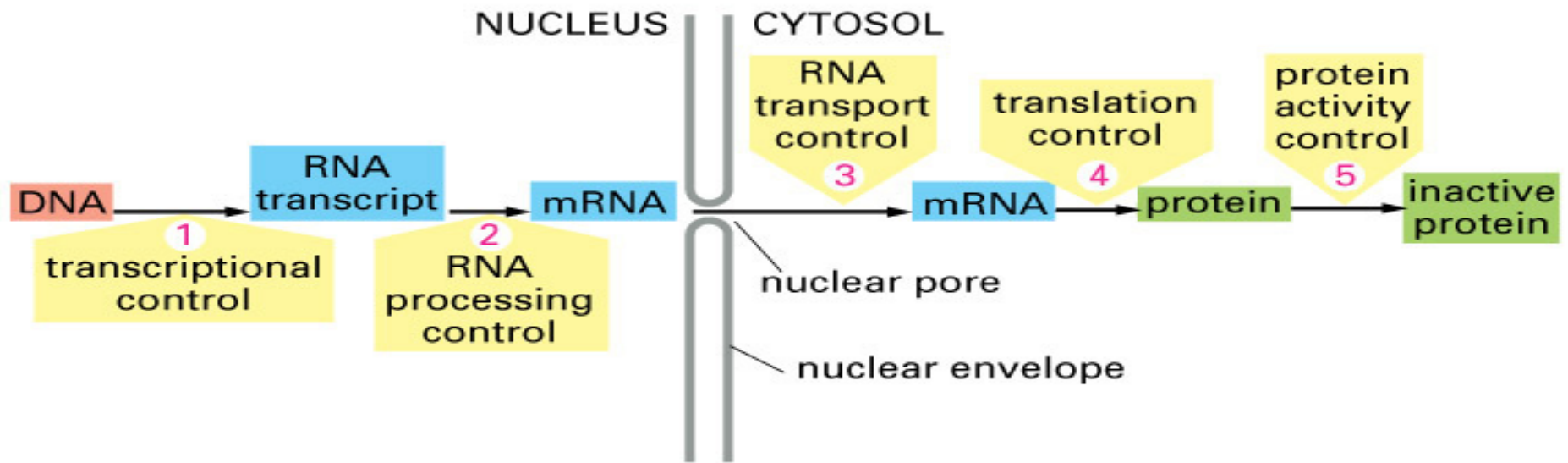
Introduction to molecular biology

Control of gene expression

- Transcriptional
- Posttranscriptional
- Translational
- Posttranslational

Introduction to molecular biology

Control of gene expression (overview)



Introduction to molecular biology

Control of gene expression depends on various factors including:

- **Chromosomal** activation or deactivation.
- Control of initiation of **transcription**.
- Processing of **RNA** (e.g. **splicing**).
- Control of RNA transport.
- Control of **mRNA** degradation.
- Control of initiation of **translation** (only in **eukaryotes**).
- **Post-translational** modifications.

Introduction to molecular biology

Trends in understanding gene regulation:

- Past focus has been on understanding transcription initiation.
- There is increasing elucidation of **posttranscriptional and translational regulation**.
- Mechanisms can be elaborate and interdependent, especially in development.
- Regulation relies on precise **protein-DNA and protein-protein contacts**.

Introduction to molecular biology

The vocabulary of gene regulation:

- **Housekeeping gene**
 - under **constitutive expression**
 - constantly expressed in approximately all cells
- **Regulated gene**
 - Levels of the gene product rise and fall with the needs of the organism.
 - Such genes are **inducible**.
 - able to be turned on
 - Such genes are also **repressible**.
 - able to be turned off

Introduction to molecular biology

RNA polymerase binding to promoters is a major target of regulation

- RNA polymerases bind to *promoter* sequences near the starting point of transcription initiation.
- The **RNA pol-promoter interaction** greatly influences the rate of transcription initiation.
- Regulatory proteins (transcription factors) work to **enhance or inhibit this interaction** between RNA pol and the promoter DNA.

Introduction to molecular biology

Activators (proteins) improve contacts between RNA polymerase and the promoter

- Binding sites in DNA for activators are called **enhancers**.
- In bacteria, enhancers are usually adjacent to the promoter.
 - often adjacent to promoters that are “weak” (bind RNA polymerase weakly), so the activator is necessary
- In eukaryotes, enhancers may be very distant from the promoter.

Introduction to molecular biology

Positive regulation

- **Positive regulation involves activators.**
- Enhance activity of RNA polymerase
 - Activator-binding sites are near promoters that weakly bind RNA Pol or do not bind at all.
 - It may remain bound until a molecule signals dissociation.
 - Alternatively, the activator may only bind when signaled.

Positive regulation
Molecular signal causes dissociation of activator from DNA, inhibiting transcription.

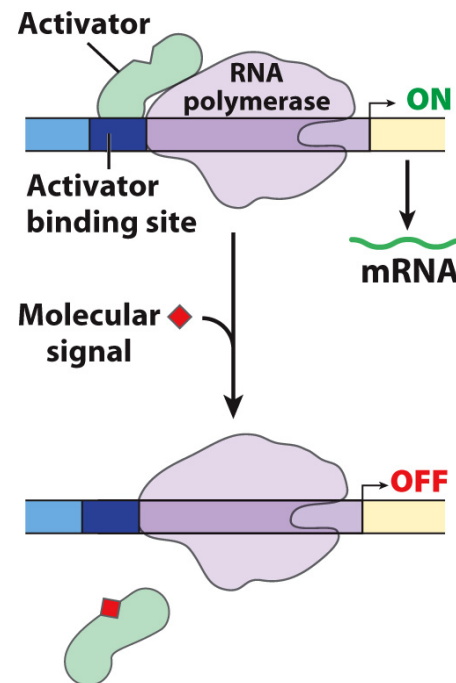


Figure 28-4c
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

Positive regulation
Molecular signal causes binding of activator to DNA, inducing transcription.

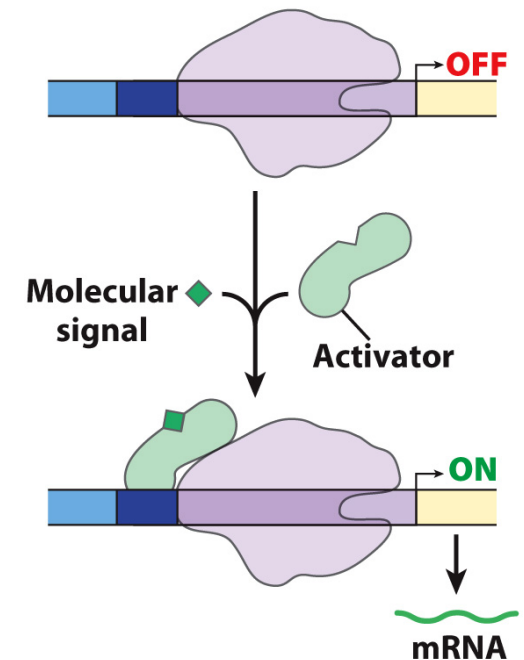


Figure 28-4d
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

Introduction to molecular biology

DNA looping allows eukaryotic enhancers to be far from promoter

- Activators can influence transcription at promoters thousands of bp away.
- How? Via **formation of DNA loops**
- Looping can be facilitated by **architectural regulator proteins**.
- **Co-activators** may mediate binding by binding to both activator and RNA polymerase.

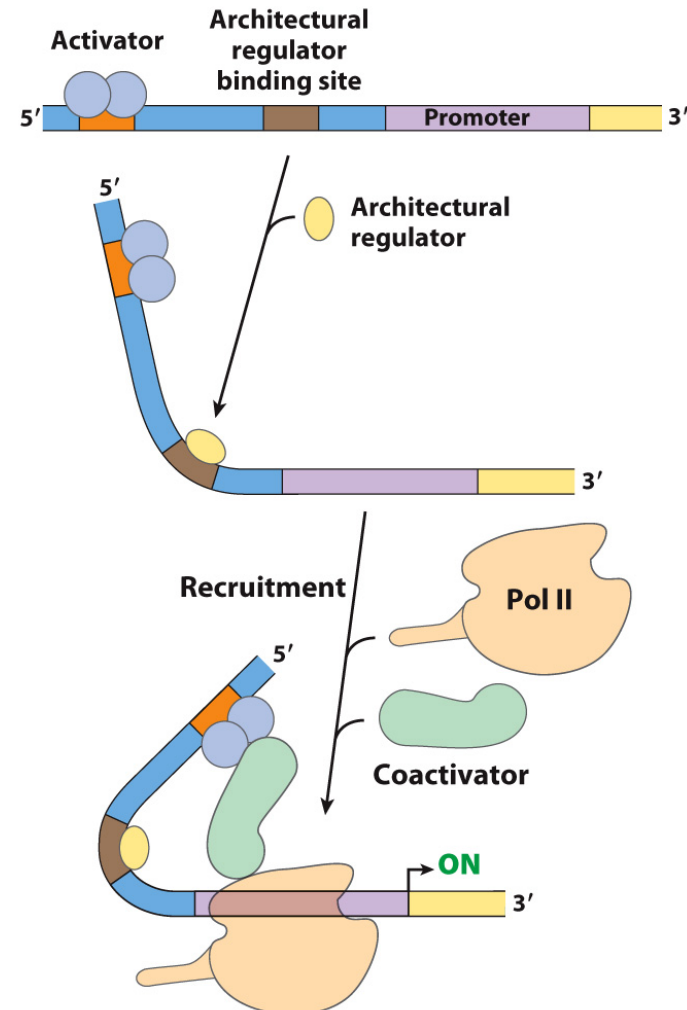


Figure 28-5
Lehninger Principles of Biochemistry, Seventh Edition
© 2017 W. H. Freeman and Company

Introduction to molecular biology

Eukaryotic gene regulation relies on combinatorial control

- In yeast, there are only **300** transcription factors for **thousands** of genes.
- Transcription factors mix and match.
- **Different combinations regulate different genes.**
- Eukaryotic gene regulation relies on protein-protein interactions.

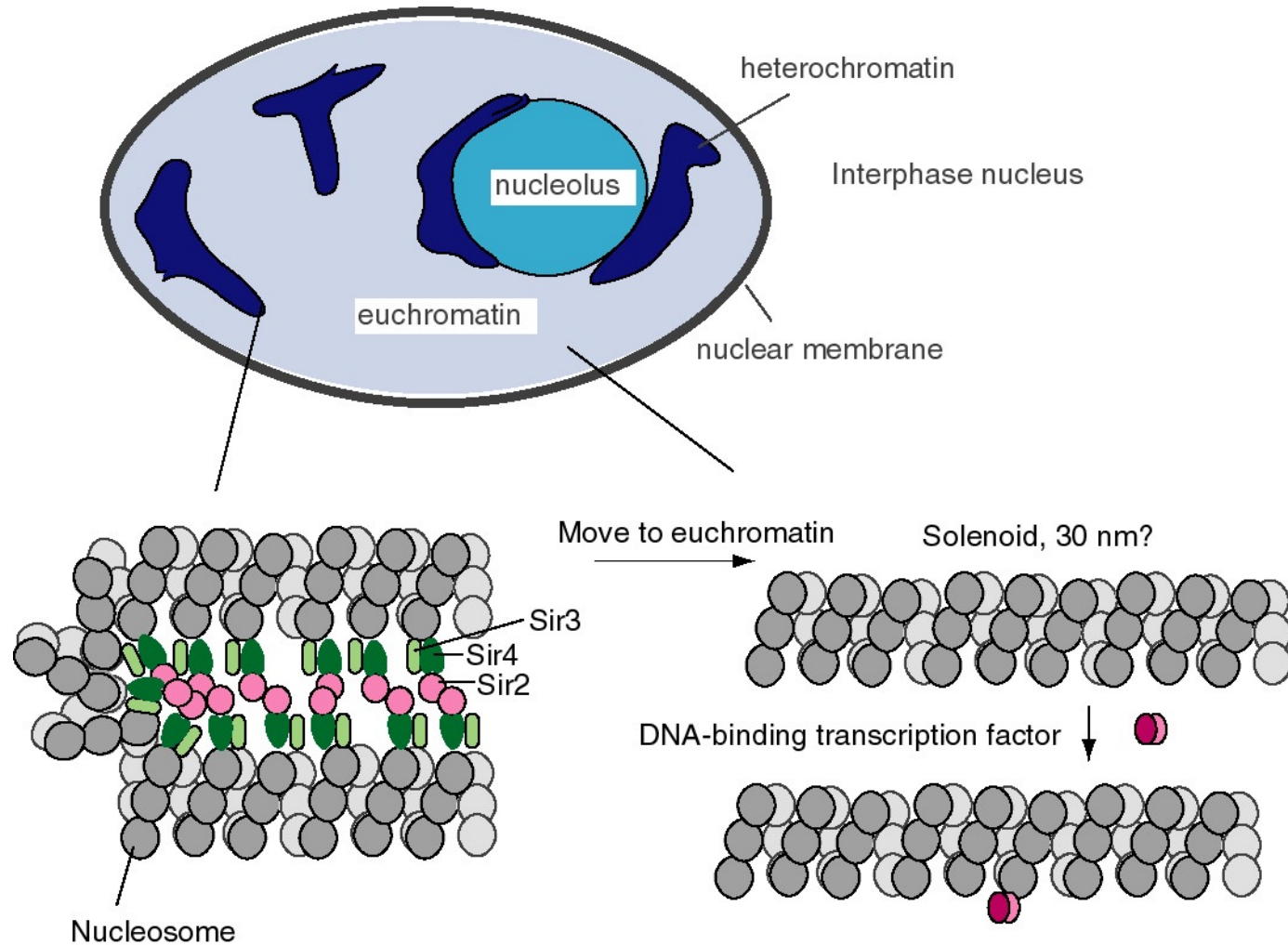
Introduction to molecular biology

Features of eukaryotic gene regulation

- Access of eukaryotic promoters to RNA polymerase is hindered by chromatin structure.
 - thus **requires remodeling chromatin**
- Positive regulation mechanisms predominate and are *required* for even a basal level of gene expression.
- Eukaryotic gene expression requires a complicated set of proteins.

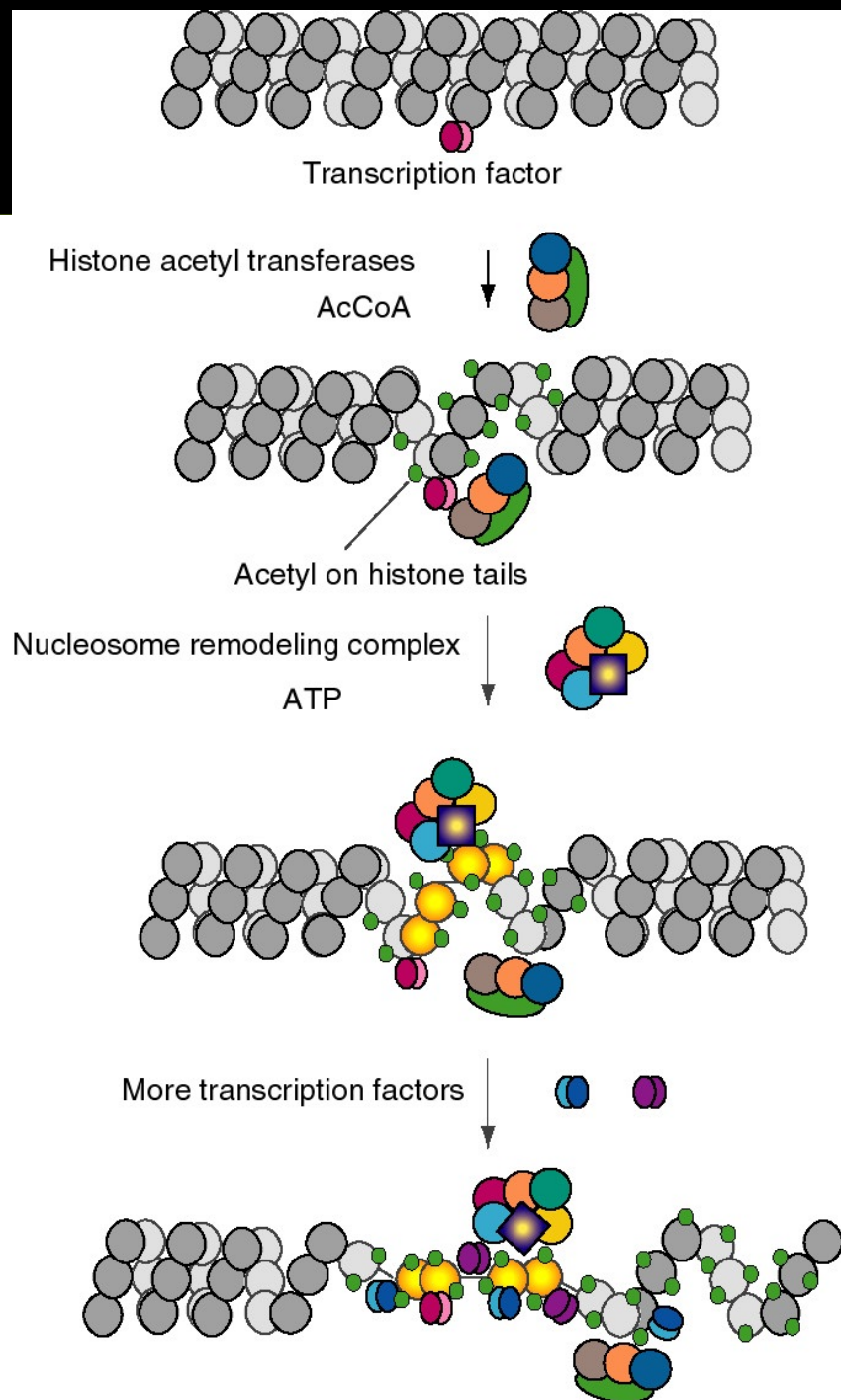
Introduction to molecular biology

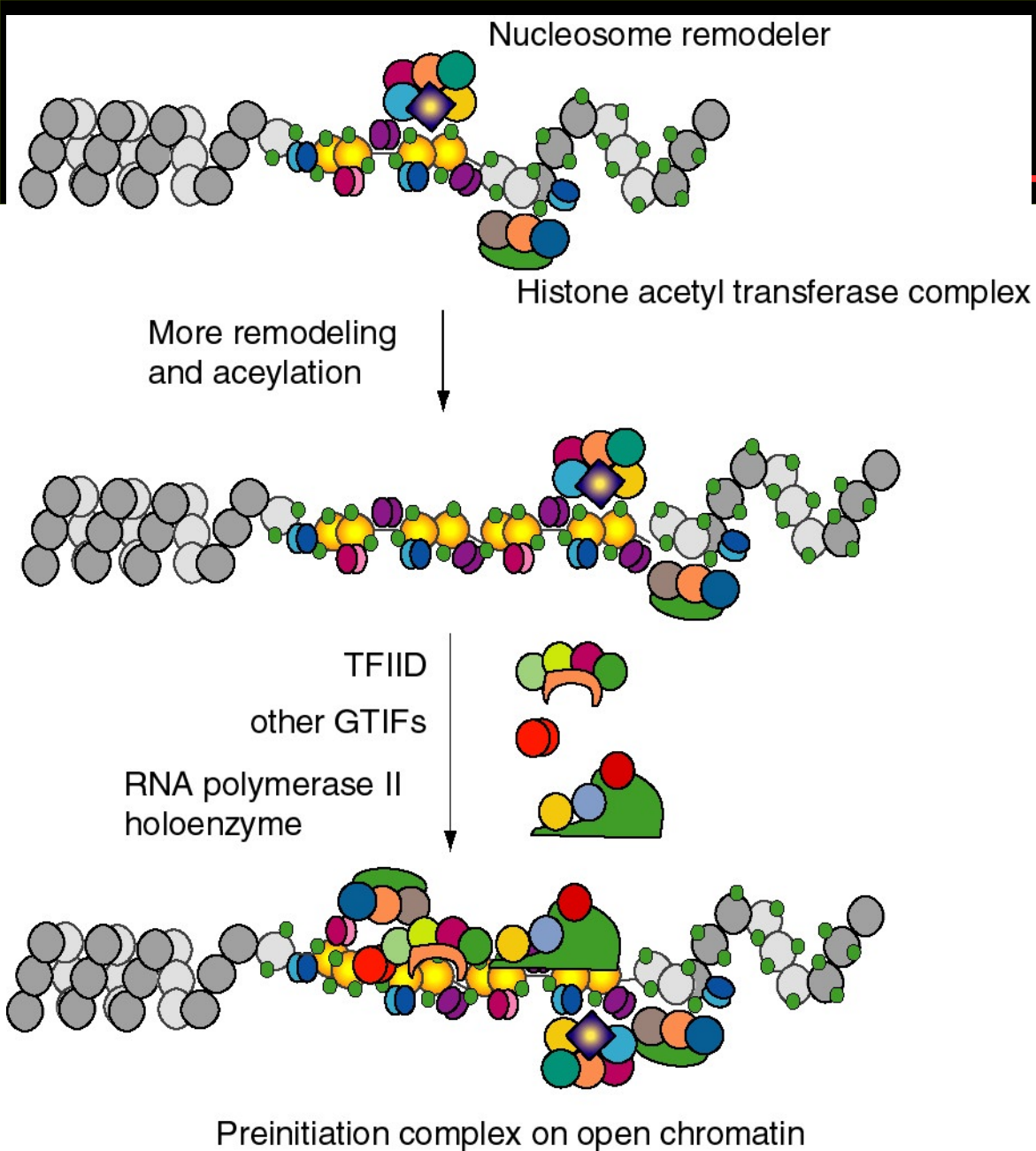
From silenced to open chromatin



molecular biology

From silenced to open chromatin





Assembly of transcription initiation complex on open chromatin

Once transcription starts the downstream nucleosomes are removed so that all the gene length can be read by RNA polymerase

Introduction to molecular biology

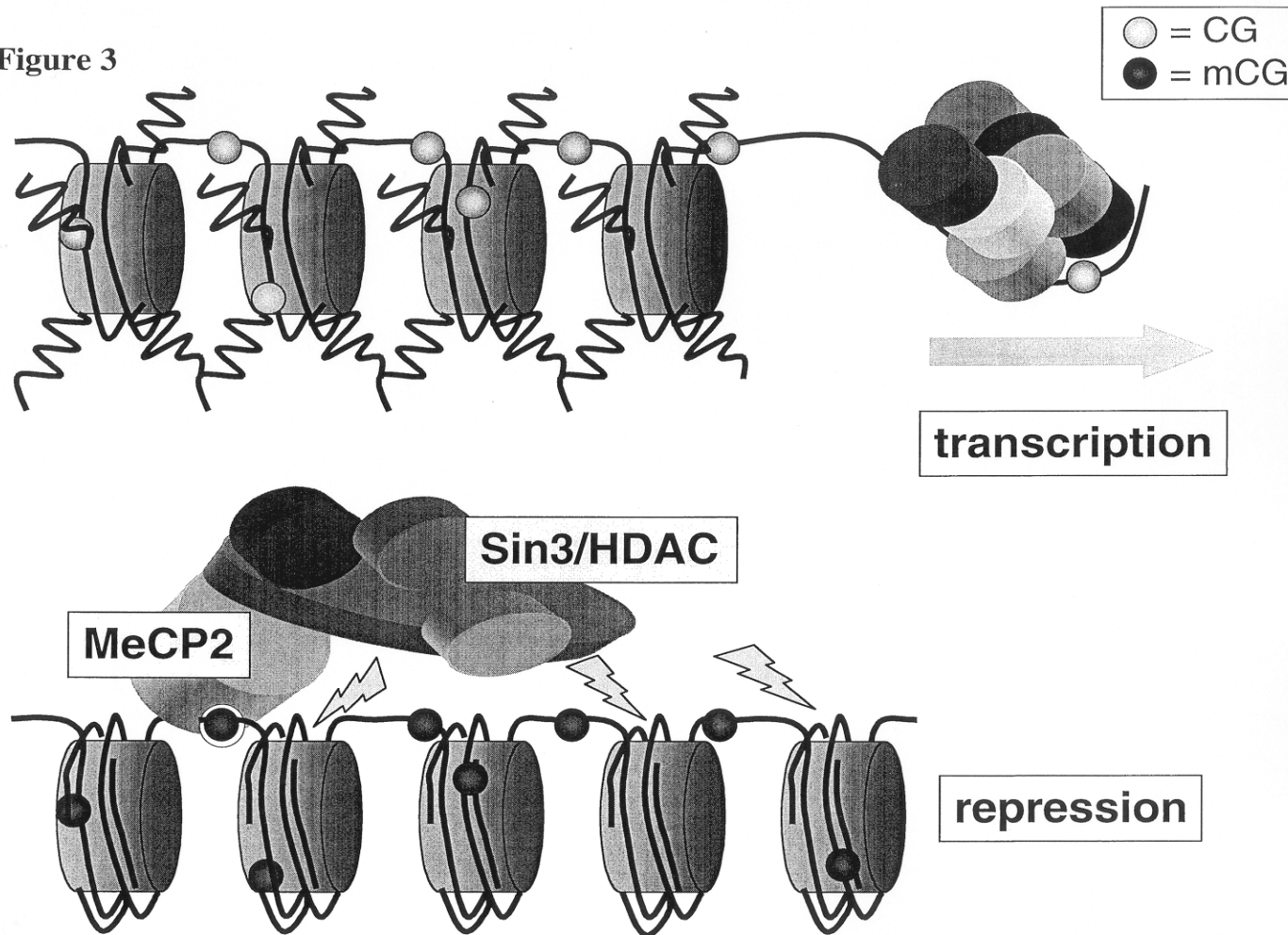
The initial binding of transcription factors can destabilize nucleosomes

- Destabilize histone/DNA interactions.
- Bound transcription factors can thus participate in nucleosome displacement and/or rearrangement.
- Provides **sequence specificity** to the formation of DNase hypersensitive sites (long stretches of open chromatin).
- **DNase hypersensitive sites** may be
 - nucleosome free regions or
 - factor bound, remodeled nucleosomes which have an increased accessibility to nucleases.

Introduction to molecular biology

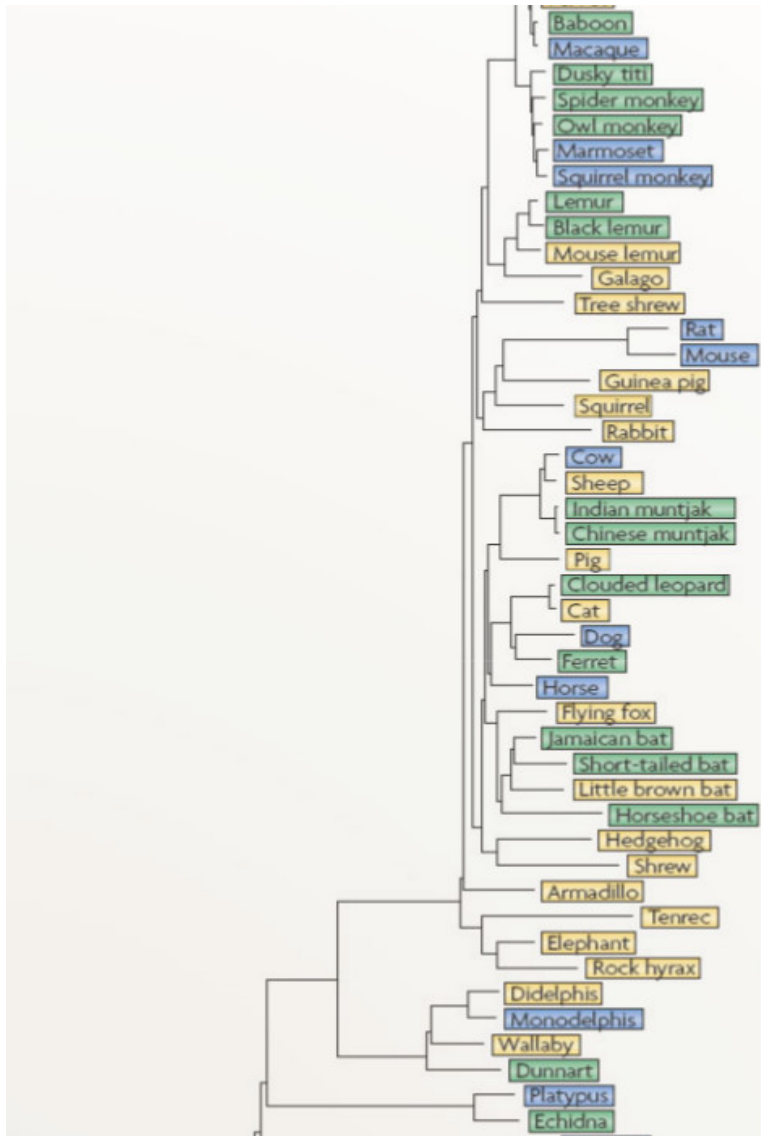
Modification of histones can promote transcript factor binding

Figure 3



Comparative genomics

Comparative genomics → Evolutionary pressures → functional inference



Next challenge: distinguish functional DNA and assign a role to it

-> genome-wide predictions by comparative genomics

-> experimental tests, also on a large scale, e.g., ENCODE

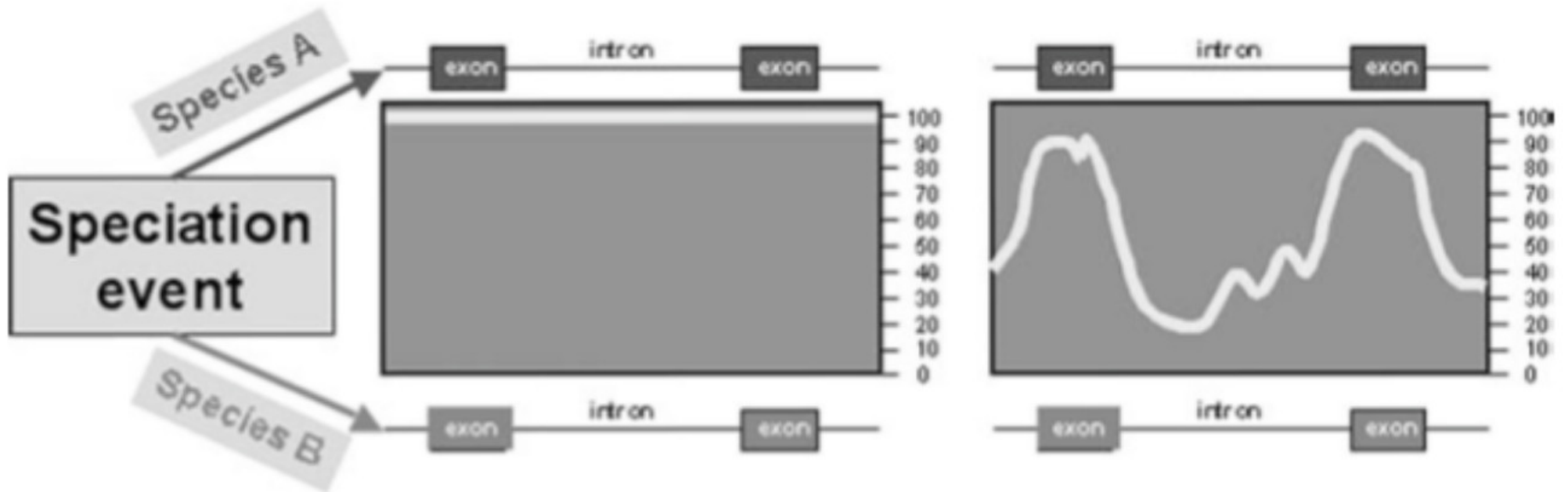
Comparative genomics

COMPARATIVE GENOMICS

- Goal of comparative genomics:
 - Finding DNA sequences that show significant signs of positive or negative selection (*and infer such sequences are functional*)
 - Positive (Darwinian) selection: fixation of advantageous alleles
 - Results in adaptive evolution
 - Sequence changes more rapidly than the bulk
 - Negative (purifying) selection: removal of disadvantageous alleles
 - Sequence is constrained by its function to remain similar to its ancestor
 - Sequence changes more slowly than the bulk
 - 'constrained' sequences: similarity level is greater than expected for neutral DNA
- Mutations of single bases; indels (insertions and deletions, resulting from replication errors or recombination); chromosomal rearrangements
- Conserved sequence: reliable alignment

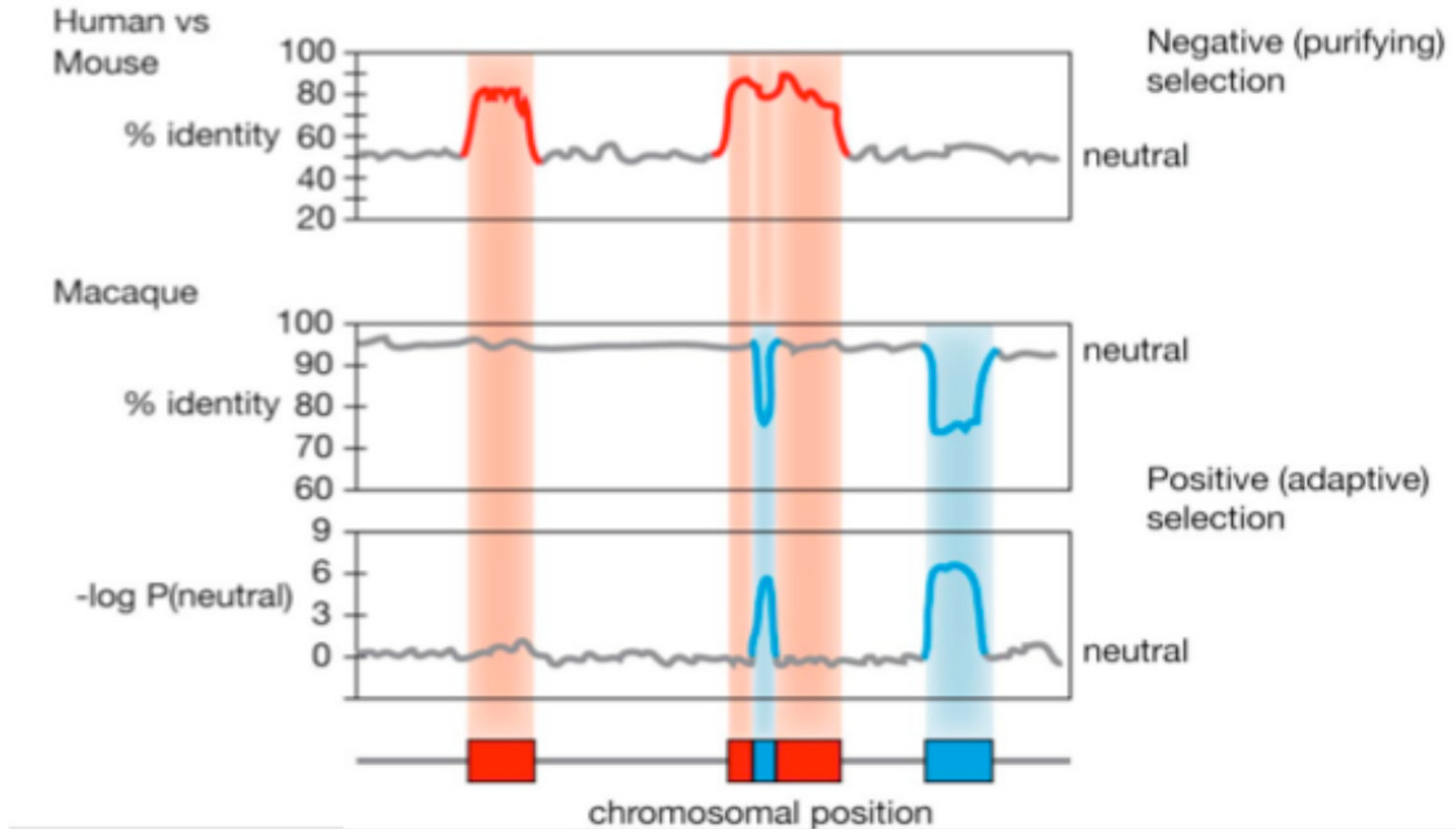
Comparative genomics

Evolution of functional regions over time



Comparative genomics

Evolution of functional regions over time



- Purifying selection: rate of sequence change is slower than that of neutral DNA
- Positive (Darwinian) selection: rate of sequence change is faster than that of neutral DNA

Comparative genomics

Neutral DNA

- Fourfold synonymous sites
- Pseudogenes (genes that are no longer active, i.e. genes that got their promoters inactivated during evolution)
- Ancestral repeats

Comparative genomics

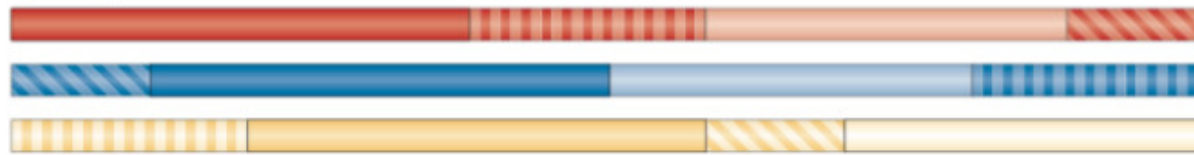
How to compare genome? Genomes sequence alignment

- Optimize a similarity score
 - Matching symbols favored
 - Mismatches not favored
 - Gaps penalized
 - Gap-open penalty + affine gap penalties (additional smaller penalty for each additional position included in the gap)
- Global alignment
 - Maps each symbol in one sequence onto a corresponding symbol in another sequence
 - Needleman-Wunsch 1970
 - ClustalW
 - AVID, LAGAN
- Local alignment
 - When a match of only a portion of two or more sequences is desired
 - Smith-Waterman 1981
 - Blast
 - blastZ (long genomic DNA sequences)

Comparative genomics

Aligning large genomic sequences

- Find reliable alignments within homology blocks and deduce how the various homology blocks are connected in genomes of compared species.
- Genes on same chromosome = syntenic
 - Groups of genes that are syntenic in humans are frequently also syntenic in mouse; i.e., *conserved synteny*
 - They frequently also maintain a similar order and orientation (indicates homology = similarity because of common ancestry)
 - Homologous segments rarely extend for entire chromosomes
 - One human chromosome will align with several homology blocks in mouse, many of which are on different chromosomes.

a Sequenced genomes**b Reconstruction of homologous collinearity relationships****c Base-pair alignment**

```

TGCCCCGTGATCACCCAAGTTGGCCAGAGACCCTGGGGTGGGGCTGATTCTGTCTGGATATACGGGGAGGGGTAAGCATGAGGAA
.....C.....C.....A.....CC.....
CAT...C..TGT..G.....CA.....C..CT..C.....G.GT..A.....G.....C.
G.....G..A.....ACA.....T.....C..C.A.A..GGGT..A.C.....G..TG...ACC
A..AGG-.....G.C..CTG...CA.....T.....C.C.....GAC..T..A.....G..AG.
A..AGG-.....G.C..CTG...CA.....CC.....G.GAA..T.....A.....AG.
CA...-C.....G.G..T.....TA.....C..C.CA..G.G..A.....C..T.....G.
CAT...-TG...CA...T.A..C-.....T.....C..A..T.GA..T...T.-G.GG..A--
CA.T--...TG.T.GA.A...GC..G.....CAC.A..C.CTC..CCAG.GT...A..C.CG.AG...AG.
CA.T.T-...GGT..G.....GA..GT..A...T...GA.CC..C...G..GT.....A.C..TG...AG.
CAT..T-...TG.G.G.....A.TG.....A..T...AG.C.AC.....G.GT.....A.C.GATG...AG.
CA..A-...TG..G.CCA...CC.CA.....CT..C...C...G.G.GT.....A..GGA.G...AG.
CA...T-..CTG...G.....T..CT...C-----T.....G.TG...T..
.....C.....A.G.....GT--T.T.....C..C...C.G.GT..A.....AC..TG..T.A..
CA...A.....A.G.....C..A.....C..C.C.G.G.GT..A...A.C..CTG...A..
....TC...GG..AGGT.....C..AT.....T.....C..C.A..GCG.GT..CA.....T-----

```

d Constraint detection

```

TGCCCCGTGATCACCCAAGTTGGCCAGAGACCCTGGGGTGGGGCTGATTCTGTCTGGATATACGGGGAGGGGTAAGCATGAGGAA
.....C.....C.....A.....CC.....
CAT...C..TGT..G.....CA.....C..CT..C.....G.GT..A.....G.....C.
G.....G..A.....ACA.....T.....C..C.A.A..GGGT..A.C.....G..TG...ACC
A..AGG-.....G.C..CTG...CA.....T.....C.C.....GAC..T..A.....G..AG.
A..AGG-.....G.C..CTG...CA.....CC.....G.GAA..T.....A.....AG.
CA...-C.....G.G..T.....TA.....C..C.CA..G.G..A.....C..T.....G.
CAT...-TG...CA...T.A..C-.....T.....C..A..T.GA..T...T.-G.GG..A--
CA.T--...TG.T.GA.A...GC..G.....CAC.A..C.CTC..CCAG.GT...A..C.CG.AG...AG.
CA.T.T-...GGT..G.....GA..GT..A...T...GA.CC..C...G..GT.....A.C..TG...AG.
CAT..T-...TG.G.G.....A.TG.....A..T...AG.C.AC.....G.GT.....A.C.GATG...AG.
CA..A-...TG..G.CCA...CC.CA.....CT..C...C...G.G.GT.....A..GGA.G...AG.
CA...T-..CTG...G.....T..CT...C-----T.....G.TG...T..
.....C.....A.G.....GT--T.T.....C..C...C.G.GT..A.....AC..TG..T.A..
CA...A.....A.G.....C..A.....C..C.C.G.G.GT..A...A.C..CTG...A..
....TC...GG..AGGT.....C..AT.....T.....C..C.A..GCG.GT..CA.....T-----

```

Figure 1 | **Overview of comparative sequence analysis.** **a** | Genomes of different species are sequenced by various strategies and assembled by computational algorithms. **b,c** | Homologous collinear segments are then identified and aligned. **d** | Finally, downstream analyses such as identifying constrained sequences can be carried out.

Comparative genomics

Various methods → results stored in dedicated web based (genomic) browsers

- Ensembl <https://www.ensembl.org>
 - *Mercator* for homologous collinearity
 - *PECAN* alignments
 - **GERP** constraints
- UCSC Genome Browser <https://genome.ucsc.edu/>
 - *Chains-and-nets* for homologous collinearity
 - *MultiZ* alignments
 - **Phastcons** constraints

Comparative genomics

Phylogenetic depth of the genome alignments

Table Q.2. Portions of the human genome conserved and constrained between various species.

Comparison species ^a	Distance from human		Fraction of human intervals aligning to comparison species ^d			
	Divergence time (Myr) ^b	Substitutions per synonymous site ^c	Total genome ^e	Coding exons ^f	Regulatory regions ^g	UCEs ^h
chimpanzee	5.40	0.015	0.95	0.96	0.97	0.99
macaque	25.0	0.081	0.87	0.96	0.96	0.99
dog	92.0	0.35	0.67	0.97	0.87	0.99
mouse	91.0	0.49	0.43	0.97	0.75	1.00
rat	91.0	0.51	0.41	0.95	0.70	1.00
opossum	173	0.86	0.10	0.82	0.32	0.95
chicken	310	1.2	0.037	0.67	0.06	0.95
zebrafish	450	1.6	0.023	0.65	0.03	0.76
Number			2.858x10 ⁹ nucleotides	250,607	1369	481

Alignments with distantly related species: indicator of constraint

Comparative genomics

Portion of human genome under constraints

- 5% under constraint
 - Lower bound estimate of the portion of the human genome that is functional
 - DNA sequences not included in this estimate:
 - Those diverged for new functions in different lineages
 - Those acquired new function recently through adaptive evolution
 - 1.2% protein-coding
 - 0.7% UTR of mature mRNA
 - Remaining 3% (larger than ~2% for mRNA!)
 - Noncoding RNAs
 - Regulatory sequences

Comparative genomics

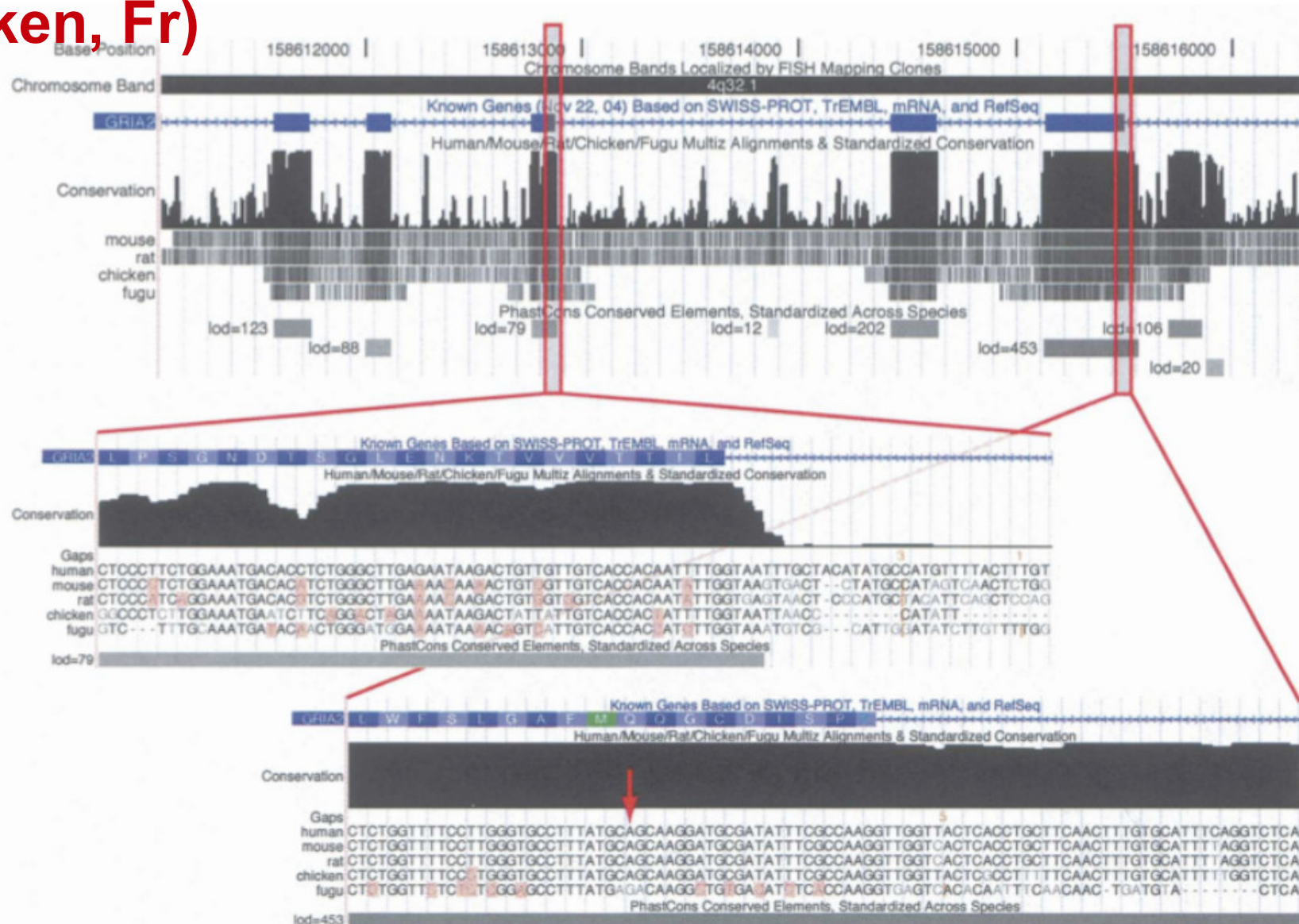
Specific sequences constraints level estimation

- phastCons
 - Phylogenetic Hidden Markov Model
 - Two states of conservation: one neutral and one constrained
 - Posterior probability that any aligned position came from the constrained state



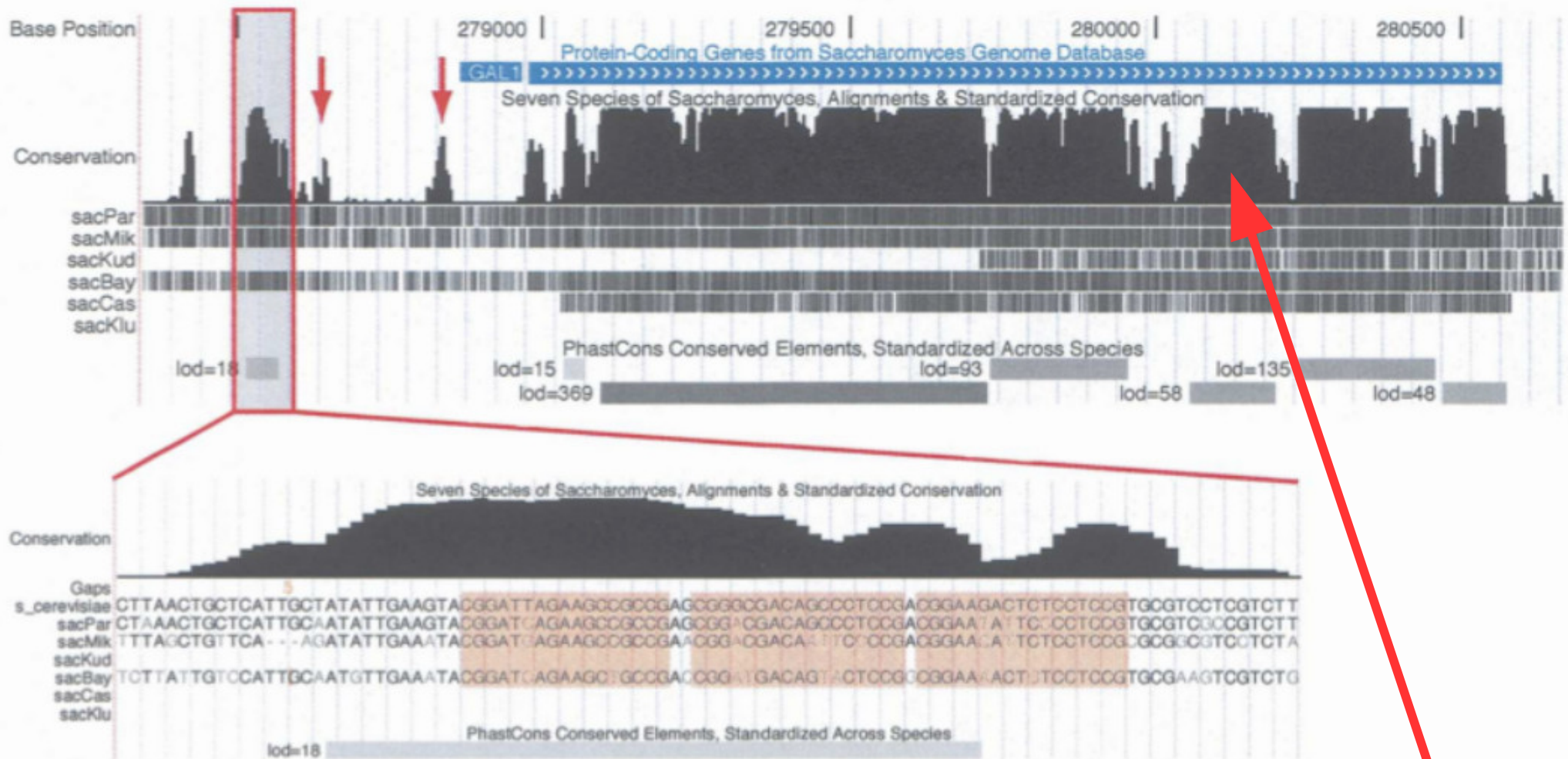
Comparative genomics

PhastCons → exon (3' terminal region) (Hs, Mm, Rn, chicken, Fr)



Comparative genomics

PhastCons → upstream regulatory region (7 yeasts)



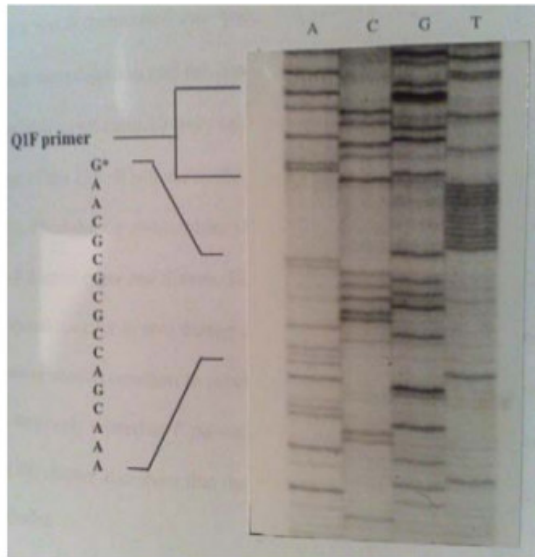
NB: PhastCons (as GERP,...) → 1 real score per base

NGS

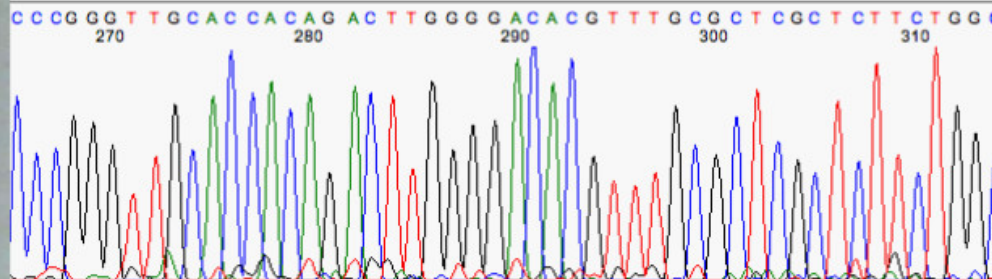
Gene Expression and Variation Analysis by Next Generation Sequencing

NGS

Previous sequencing technologies



Fluorescent dye-terminator sequencing
(90s – today)



© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

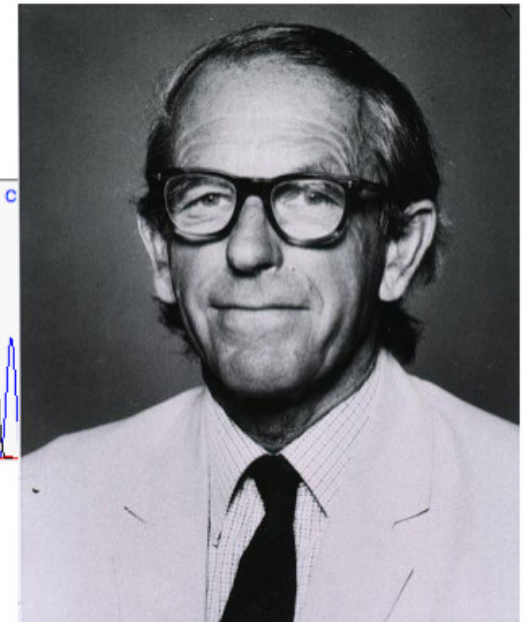


Image from the National Library of Medicine, public domain.

Fred Sanger
Nobel laureate 1958,1980

© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

Traditional sanger / chain termination methods
Radioactive slab gels.
(70s,80s,90s)

4 separate lanes, 1 per base

How does this work?

What is the data like? How long?

How much money (per base)?

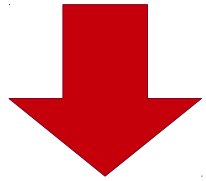
Length: 500-1000bp

Data is bad at beginning and end of read

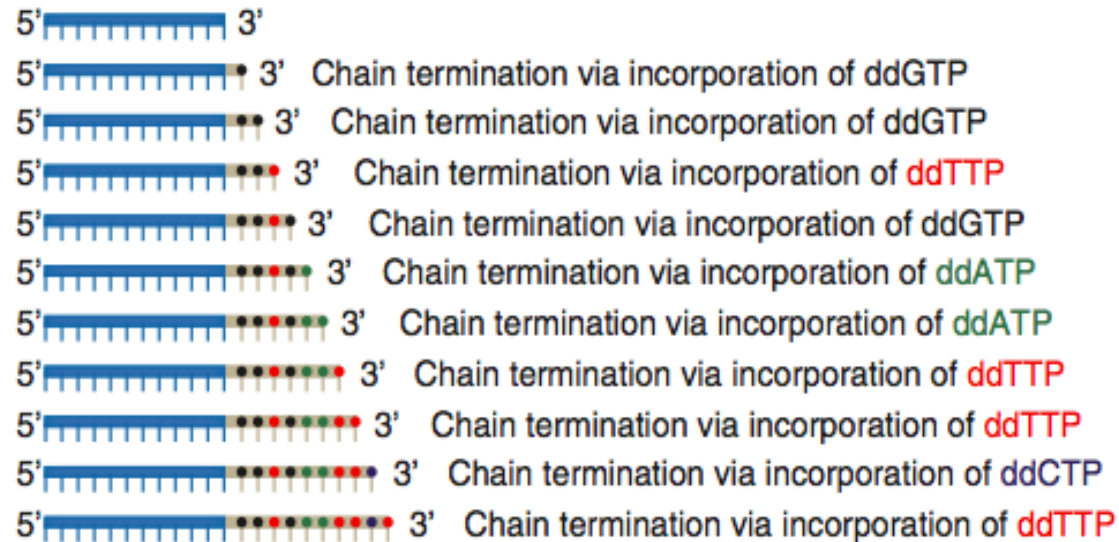
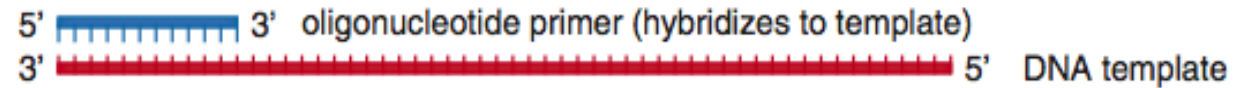
~cents per base

NGS

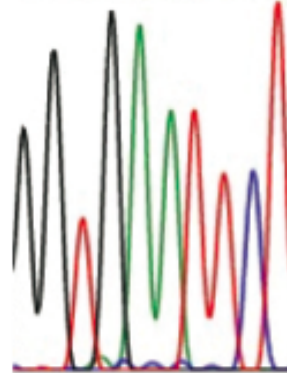
DNA sequencing by Sanger method



Primer elongation, chain termination upon incorporation of ddNTP, separation, detection



G G T G A A T T C T



Capillary gel electrophoresis to separate DNA fragments by size

Laser detection of labeled ddNTPs

Determination of DNA sequence inferred by pattern of chain termination

NGS

View genomic DNA (here from the beta globin locus) from the Trace Archive at NCBI: FASTA format

Show as **FASTA** in color

>gnl|ti|981051509 name:17000177953277 [Send to BLAST](#)

Quality score: not available >=0 - <20 >=20 - <40 >=40 - <60 >=60 - <80 >=80 - <100

```

TTTCGAATAATTTAAAATACATCATTGCAATGAAAAATAAATGTTTTTTATTAGGCAGAATCCAGATGCTCA
AGGCCCTTCATAATATCCCCCAGTTTAGTAGTTGGACTTAGGGAACAAAGGAAACCTTTAATAGAAATTGG
ACAGCAAGAAAGCGAGCIIAGIGAIACIIGIGGGCCAGGGCAIIAGCCACACCAGCCACCACIIICIGAI
AGGCAGCCTGCCTGGTGGGGTGAATTTCTTTGCCAAAGTGTATGGGCCAGCACACAGACCAGCACGTTGCC
CAGGAGCTGTGGGAGGAAGATAAAGAGGTATGAACATGATTAGCAAAAAGGGCCTAGCTTGGACTCAGAATA
ATCCAGCCTTATCCCCAACCATAAAATAAAGCAGAATGGTAGCTGGATTGTAGCTGCTATTAGCAATATG
AAACCTCTTACATCAGTTACAATTTATATGCAGAAATATTTATATGCAGAGATATTGCTATTGCCTTAAC
CCAGAAATTATCACTGTTATTCTTTAGAATGGTGCAAAAGAGGCATGATACATTGTATCATTATTGCCCTG
AAAGAAAGAGATTAGGGAAAGTATTAGAAATAAGATAAACAAAAAAGTATATTAAGGAAAGAAAGCATT
TTTTAAAATTACAAATGCAAAAATTACCCTGATTTGGTCAATTATGTGTACACATATTAACATTACT
TTTAAACCATAAATATGTATAATGGATTATGTATCAATTAAAAATAAAAGAAAATAAAGTAGGGAGATTA
TGAATATGCAAAAT

```

Show as **Quality** in color

>gnl|ti|981051509 name:17000177953277

Quality score: not available >=0 <20 >=20 <40 >=40 <60 >=60 <80 >=80 <100

```

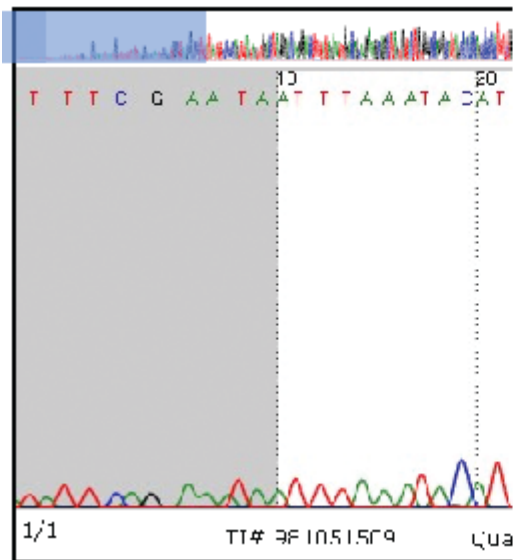
12 11 10 10 10 10 12 12 15 27 29 29 29 29 29 29 29 28 28 30 30 30 30 30 30 30 30 30 30 30 30 30 30
30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 30 32 32 32 32 32 32 30 30 30 30
30 31 30 30 32 32 31 31 31 30 30 31 31 30 31 31 32 32 31 31 31 30 30 31 30 34 34 34 34 34 40 34 31 31 31
30 31 31 31 34 34 32 32 32 32 35 35 35 32 32 35 32 35 33 33 33 33 30 34 33 33 33 33 33 33 34 34 34 33
30 34 34 34 33 35 34 33 30 33 30 33 33 33 31 34 34 34 34 31 34 34 31 33 35 34 34 34 34 34 34 34 35 34 34
32 34 34 34 41 41 34 34 34 34 33 33 33 33 33 33 33 33 34 34 34 34 34 34 33 34 41 41 41 30 30 30 33 32
36 36 38 41 41 38 34 37 36 37 32 32 41 37 41 41 41 41 41 41 41 38 41 38 41 41 45 45 45 45 45 45 37 37
36 36 37 36 36 45 45 45 36 36 36 37 37 36 45 45 45 37 36 36 43 43 43 43 45 45 45 45 45 45 45 45 45
43 43 43 43 43 43 43 45 45 45 45 45 45 45 43 43 43 43 37 37 36 36 37 37 36 36 45 45 45 45 45 45
37 37 45 45 45 45 45 45 37 36 36 36 37 37 37 38 41 38 41 41 38 38 33 36 36 31 33 36 33 36 36 32 32 41 34
41 41 34 34 41 41 41 36 33 36 34 34 36 34 33 33 33 33 33 33 32 34 38 38 38 38 38 34 34 33 34 34 34 34
32 34 41 41 35 36 34 34 34 34 31 31 34 34 41 36 34 34 34 35 34 34 37 40 40 37 40 40 37 40 34 34 34 34
34 34 34 34 34 35 33 34 31 30 30 30 33 30 35 34 34 37 37 34 34 34 34 34 34 34 35 34 35 34 31 31 34 34 34

```

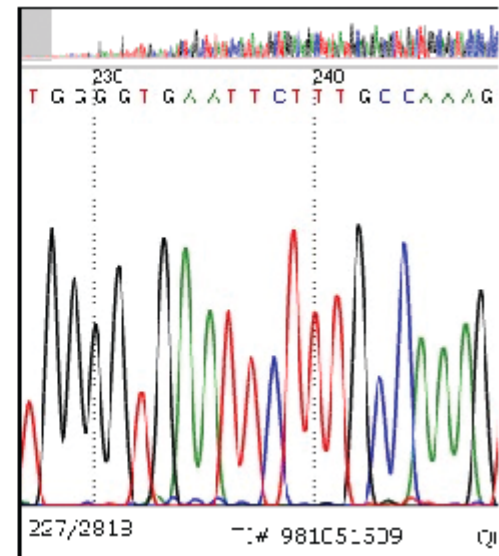
NGS

Examples of Sanger sequencing traces

Low quality reads



High quality reads



NGS

Comparison of NGS technologies

Technology	Read length (bp)	Reads per run	Time per run	Cost per megabase	Accuracy
Roche 454	700	1 million	1 day	\$10	99.9%
Illumina	50-250	<3 billion	1-10 days	~\$0.10	98%
SOLiD	50	~1.4 billion	7-14 days	\$0.13	99.9%
Ion Torrent	200	<5 million	2 hours	\$1	98%
Pacific Biosciences	2900	<75,000	<2 hours	\$2	99%
Sanger	400-900	N/A	<3 hours	\$2400	99.9%

NGS

Next Generation Sequencing

- All the sequences technologies since Sanger sequencing.
- Many sequencing technologies, but one is hugely dominant:

<https://www.illumina.com/>

<http://allseq.com/knowledgebank>

For modern-ish review of other sequencing platforms

NGS

How does Illumina Sequencing works?

- Massively parallel sequencing of **short reads** – 40bp-300bp

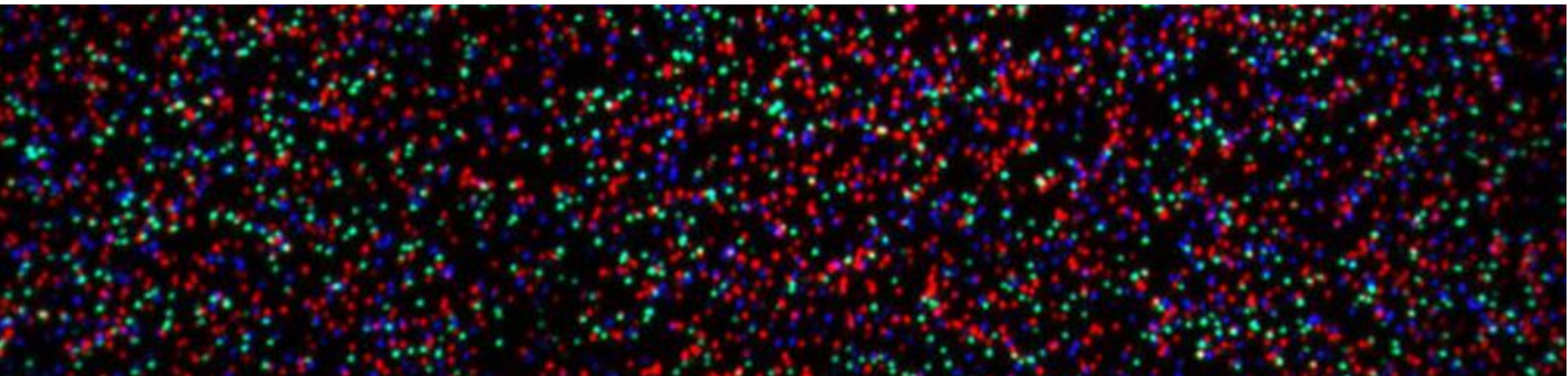
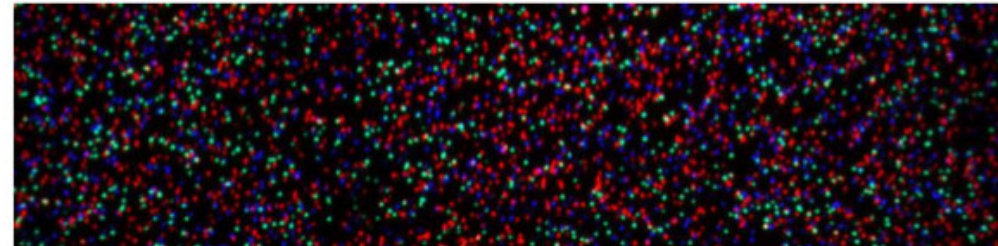
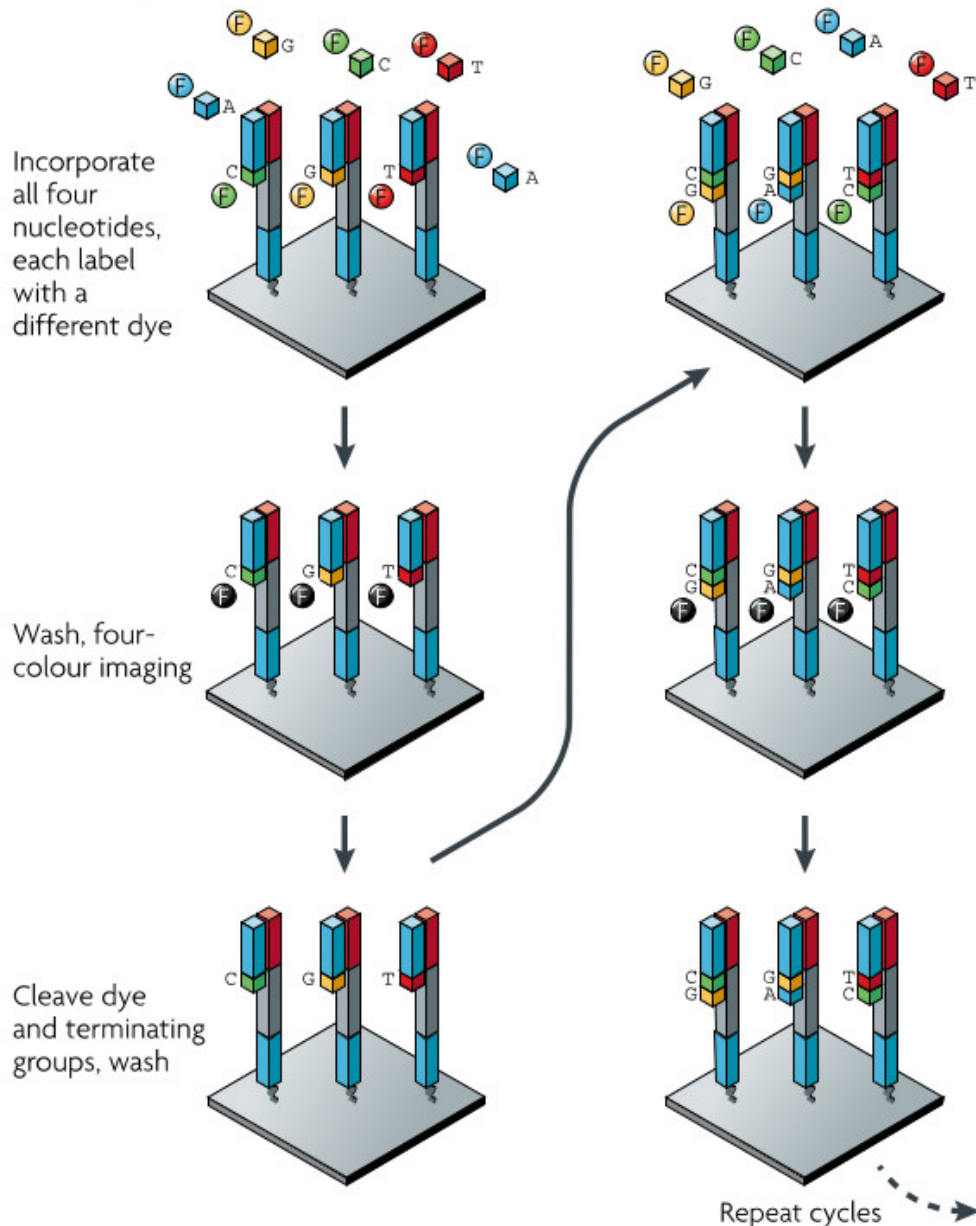


Image of Illumina HiSeq flowcell, every spot (cluster) on the flowcell is a **unique** sequencing reaction. Each spot is 1um or less. Sequencing happens on a flowcell, you buy sequencing capacity by lane. Each lane gives you 200M + reads, and costs upwards of \$1000 (Illumina HiSeq 2500)

Illumina / Solexa Sequencing

a Illumina/Solexa — Reversible terminators



© Massey University. All rights reserved. This content is excluded from our Creative Commons license. For more information, see <http://ocw.mit.edu/help/faq-fair-use/>.

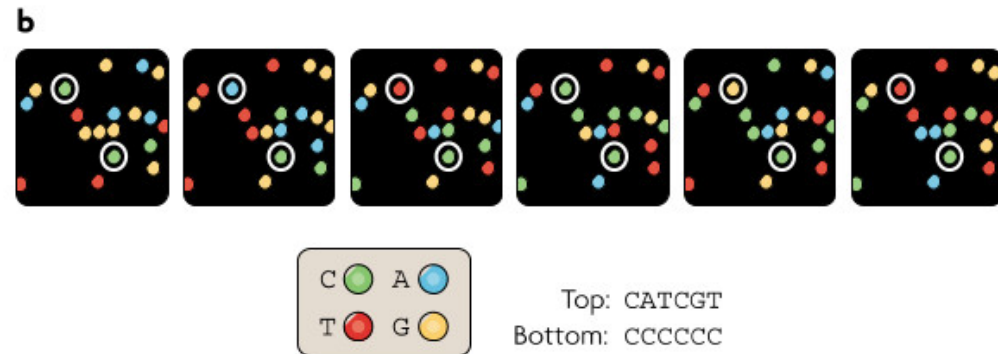
1 base per cycle.

~1 hour per cycle (20mins chemistry, 40mins imaging)

~1000 molecules per cluster

<1um per cluster

*varies somewhat depending on Illumina instrument

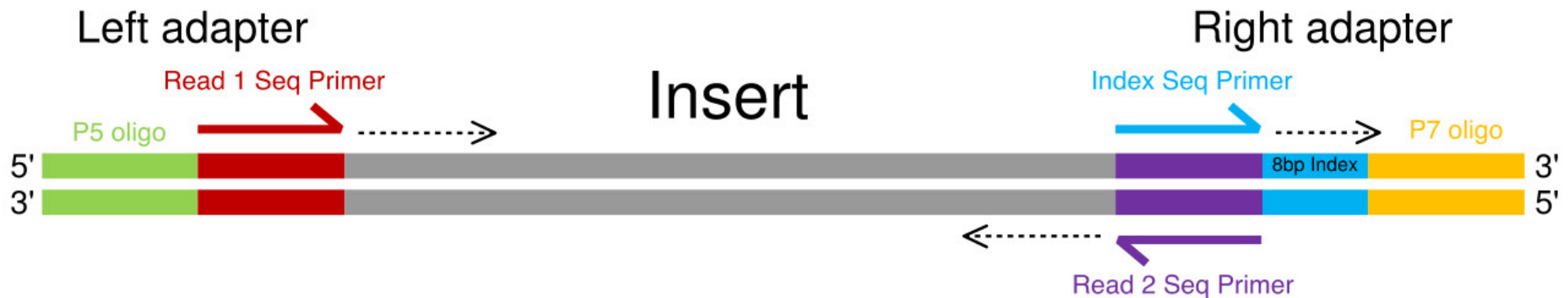


Reprinted by permission from Macmillan Publishers Ltd: *Nature Review Genetics*.

Source: Metzker, M. L. "Sequencing Technologies — the Next Generation." *Nature Reviews Genetics* 11 (2010): 31–46. © 2010.

NGS

Anatomy of an Illumina sequencing fragment

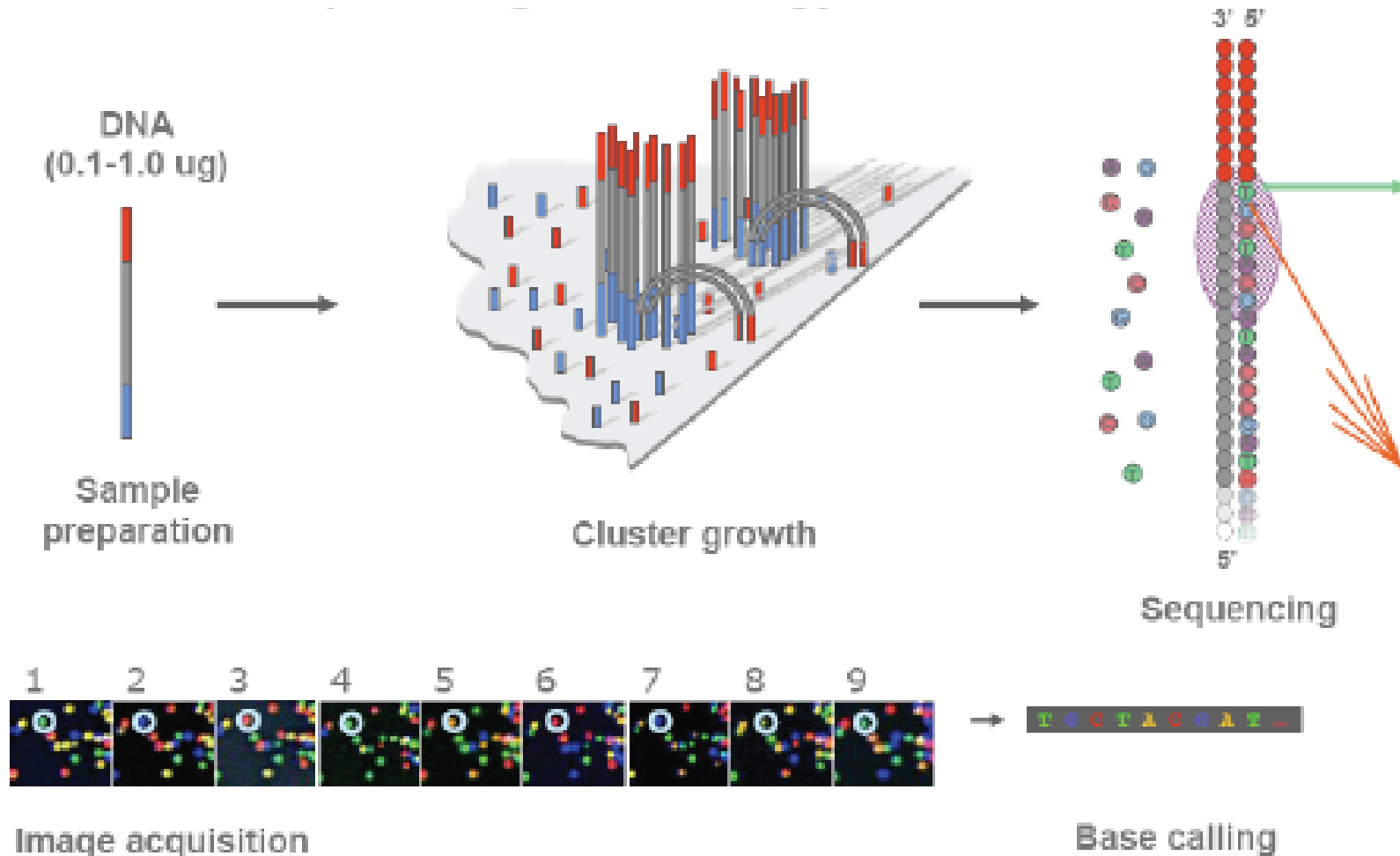


- P5 & P7 oligos bind fragment to flowcell
- Can have single ended sequencing. Only Read 1 (plus index read if multiplexing/barcoding)
- Paired end sequencing. Read 1&2 (plus index read)
- Index read gives you the multiplexing / barcoding that lets you put multiple samples onto the same "lane"

NGS

Next Generation Sequencing technology

Illumina



NGS

Movie on Illumina sequencing

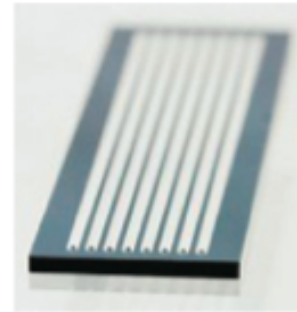
<https://www.youtube.com/watch?v=womKfikWlxM>

Next Generation Sequencing technology

Illumina

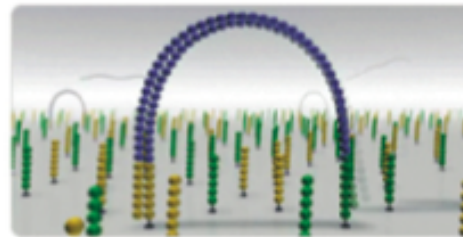
Randomly fragment genomic DNA

Library preparation



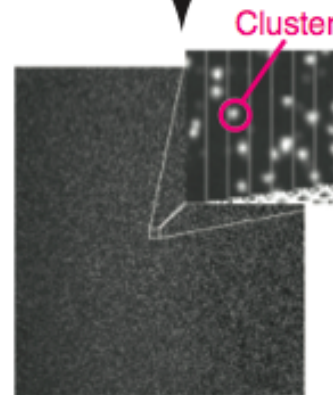
Samples immobilized on surface of a flow cell (8 lanes)

Solid phase amplification



- Bridge amplification (inverted U) generates clusters on surface of flow cell
- ~Ten million single-molecule clusters per square centimeter

Sequencing by synthesis



- Each cycle: add polymerase, one labeled deoxynucleoside triphosphate (dNTP) at a time (four labeled dNTPs per cycle)
- Image fluorescent dyes
- Call nucleotide
- Enzymatic cleavage to remove

NGS

Next Generation Sequencing technology **Illumina**

Disadvantage:

- Short read length (~150 bases)

Advantages:

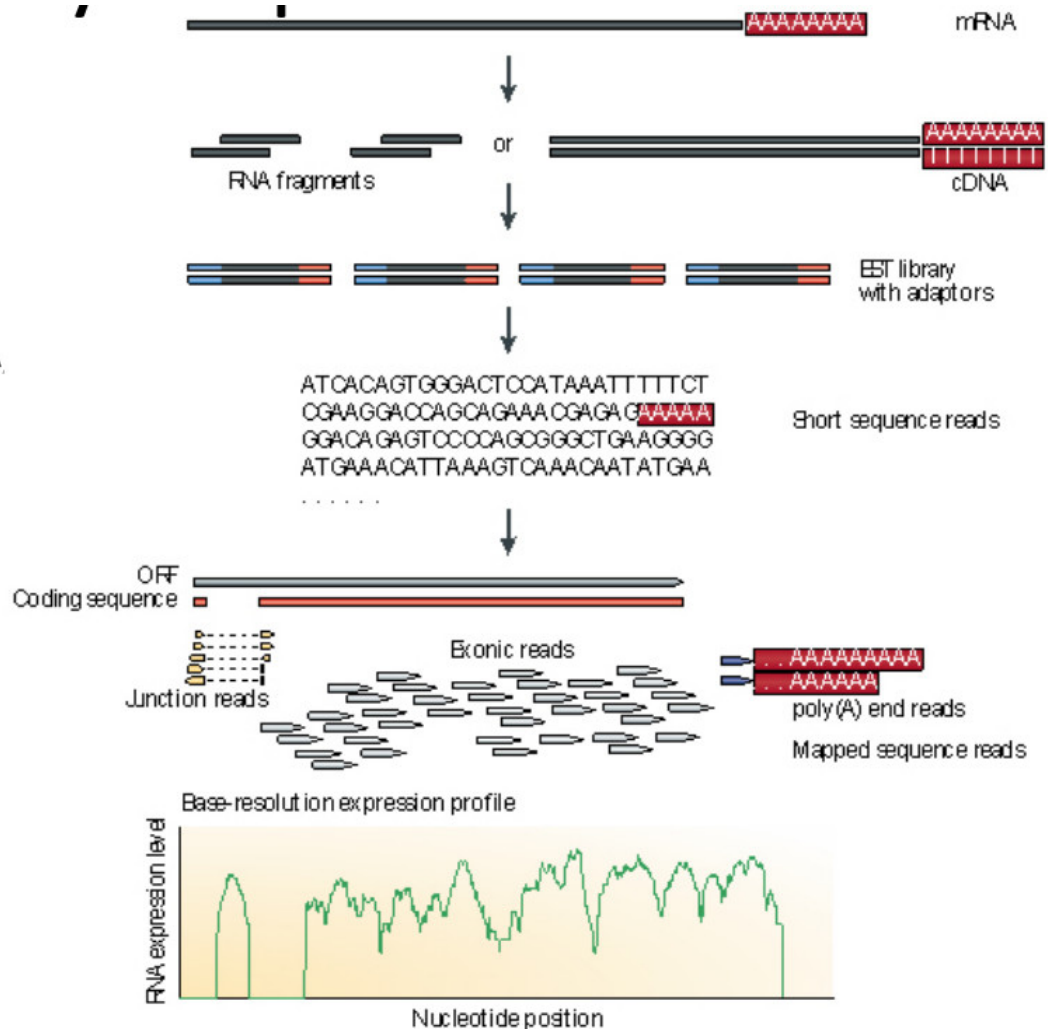
- Very fast
- Low cost per base
- Large throughput; up to 1 gigabase/experiment
- Short read length makes it appropriate for resequencing
- No need for gel electrophoresis
- High accuracy
- All four bases are present at each cycle, with sequential addition of dNTPs. This allows homopolymers to be accurately read.

NGS

How to get DNA suitable for sequencing? Library preparation

Depending on what type of library prep you do, can have totally different types of experiments (DNA, RNA, methylation, ribosome profiling)

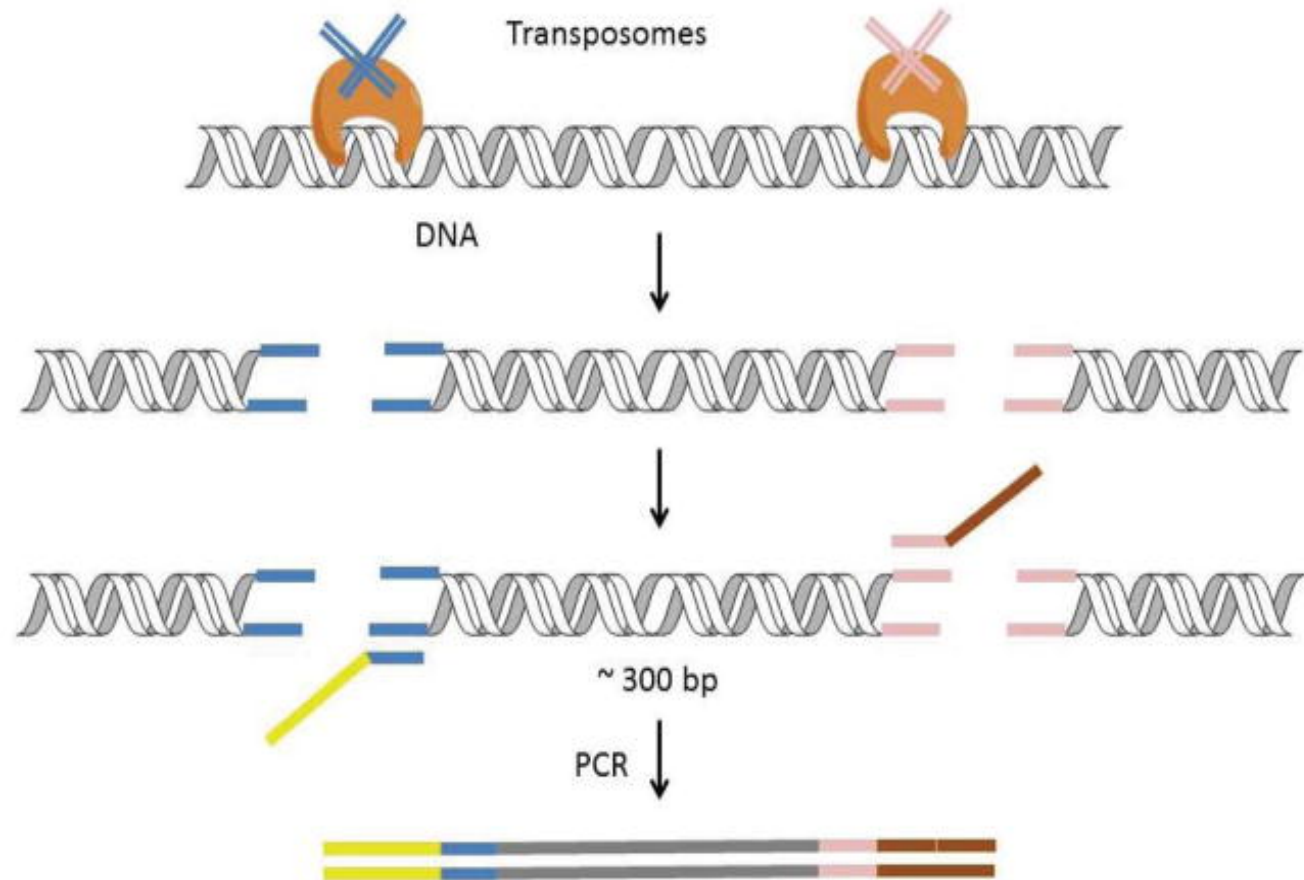
But ultimately, everything looks the same when it is converted to the final flowcell ready fragment: a dsDNA fragment with asymmetric adaptors



NGS

DNA library preparation using a transposase-based method (Nextera) developed by Illumina

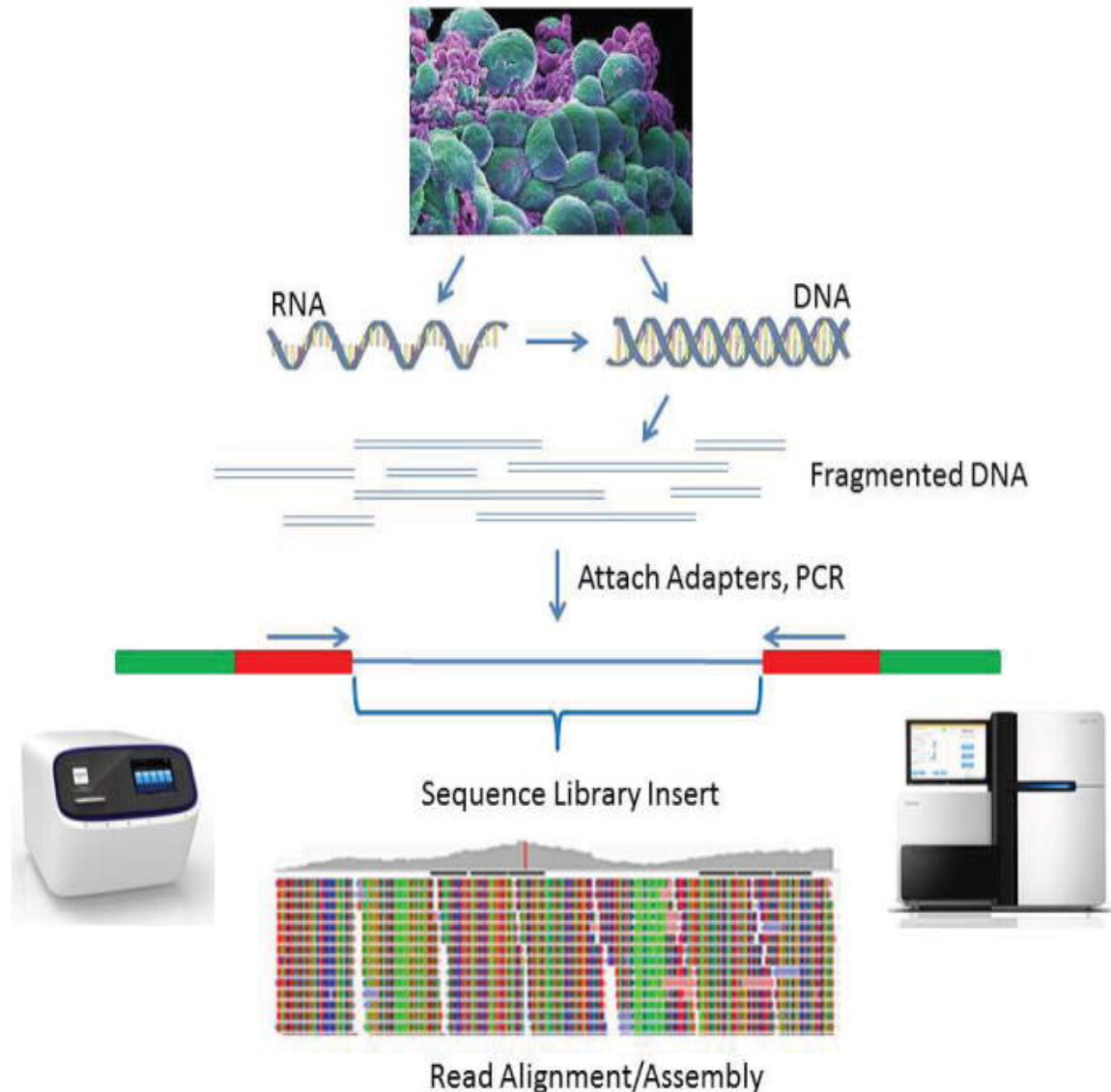
The transpososome complex comprises an engineered transposase pre-loaded with two double-stranded sequencing adapters. The transpososome simultaneously **fragments the DNA and inserts the adapters**. The full Illumina adapter sequences are completed during subsequent PCR cycling, after which the library is ready for quantitation and loading onto the flow cell.



NGS

Basic workflow for NGS library preparation

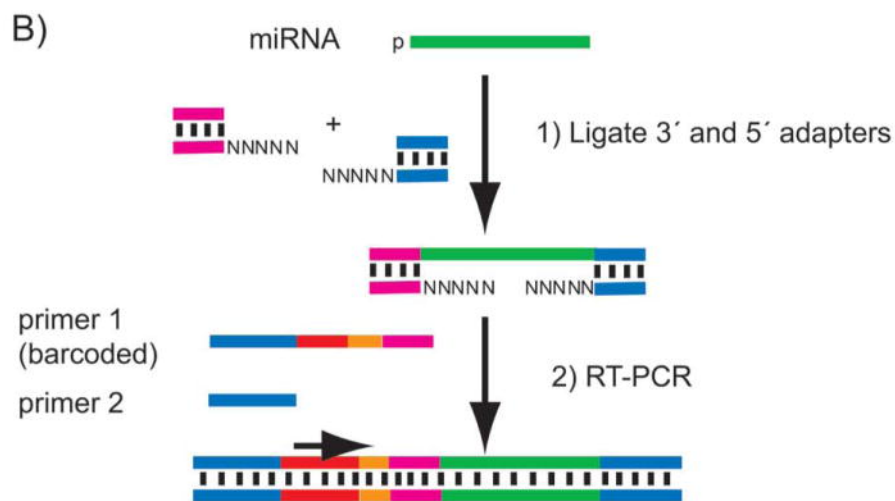
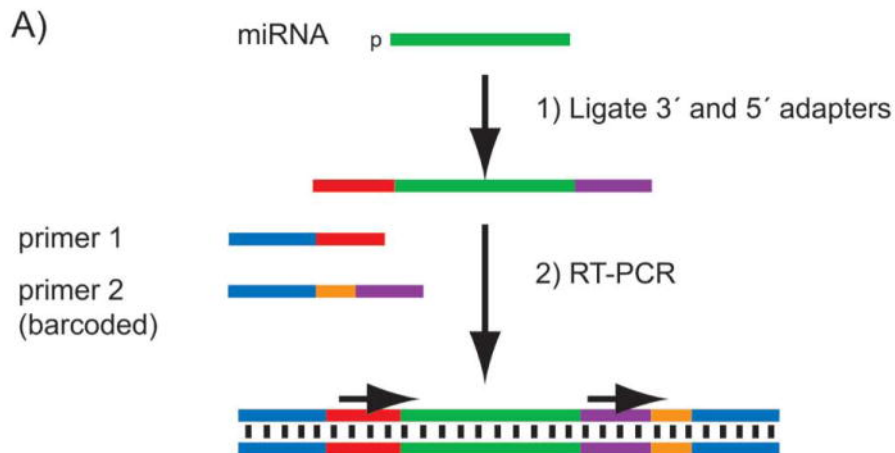
RNA or DNA is extracted from sample tissue/cells and **fragmented**. RNA is converted to cDNA by reverse transcription. DNA Fragments are converted into the library by ligation to sequencing adapters containing specific sequences designed to interact with the NGS platform, either the surface of the flow-cell (Illumina) or beads (Ion Torrent). The next step involves clonal amplification of the library, by either **cluster generation for Illumina** or microemulsion PCR for Ion Torrent. The final step generates the actual sequence via the chemistries for each technology. One difference between the two technologies is that **Illumina allows sequencing from both ends of the library insert** (i.e., paired end sequencing). Cell photograph courtesy of Annie Cavanagh, Wellcome Images.



NGS

Library preparation workflow for miRNA-seq

A) The Illumina workflow ligates a 3' adenylated DNA adapter to the 3' end of miRNA in a total RNA sample. Then, an RNA adapter is ligated to the 5' end of the miRNA. The doubled-ligated products are RT-PCR amplified to introduce barcodes for multiplex applications and generate sequencing libraries. The first read sequences the insert miRNA; a second and separate sequencing read is necessary to sequence the barcode. B) Ion Torrent's workflow uses an RNA ligase to attach 5' and 3' adapters composed of hybrid RNA-DNA duplexes. An RT-PCR reaction amplifies the sample and introduces the barcodes to the library construct. In this method, the barcode and the miRNA insert are sequenced in a single read.

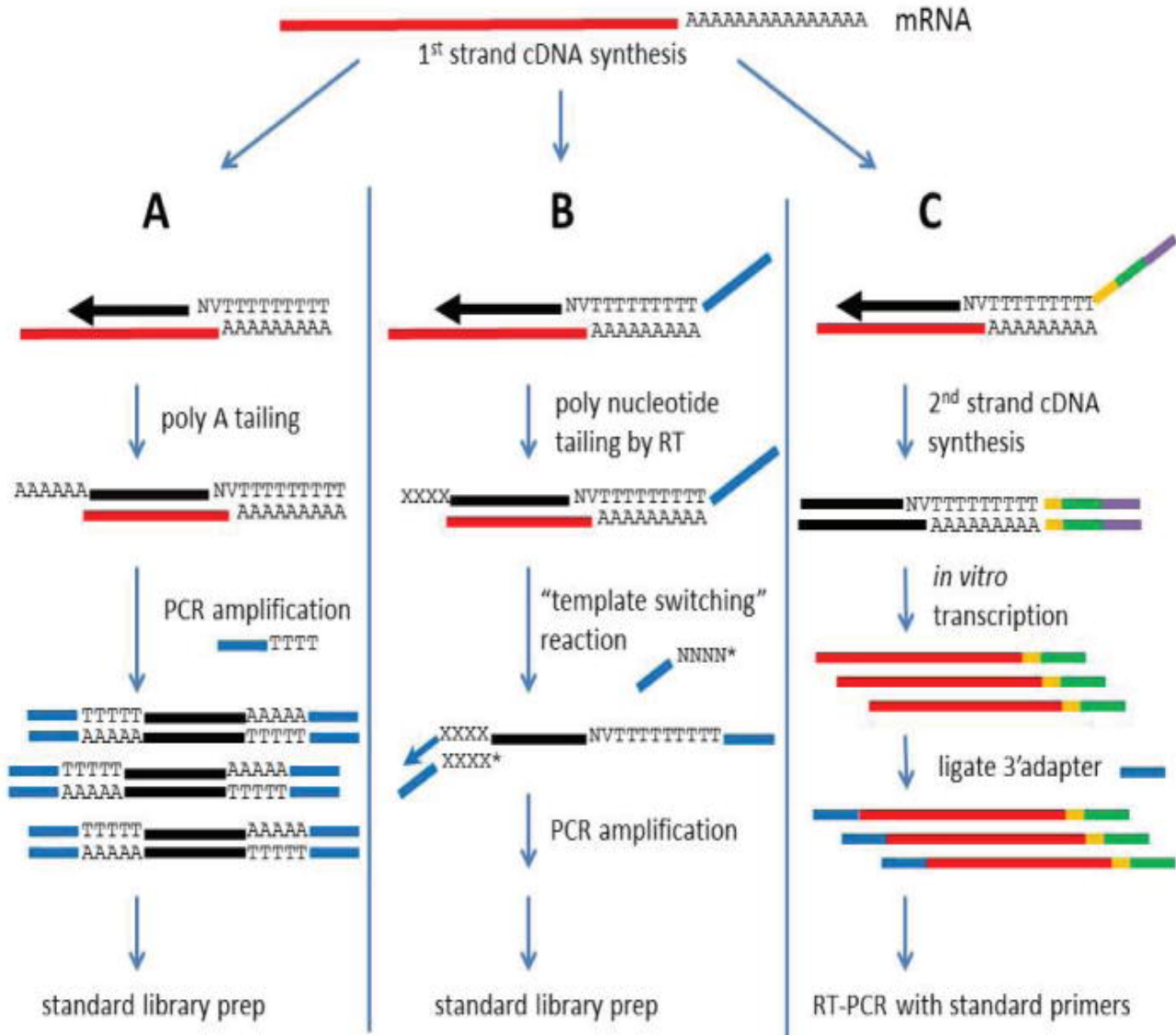


- = sequences required for amplification on flowcell or beads
- = sequence primer hybridization site
- = barcode sequencing primer hybridization site
- = barcode sequence
- = universal adapter sequence

NGS

RNA-seq libraries from single cells

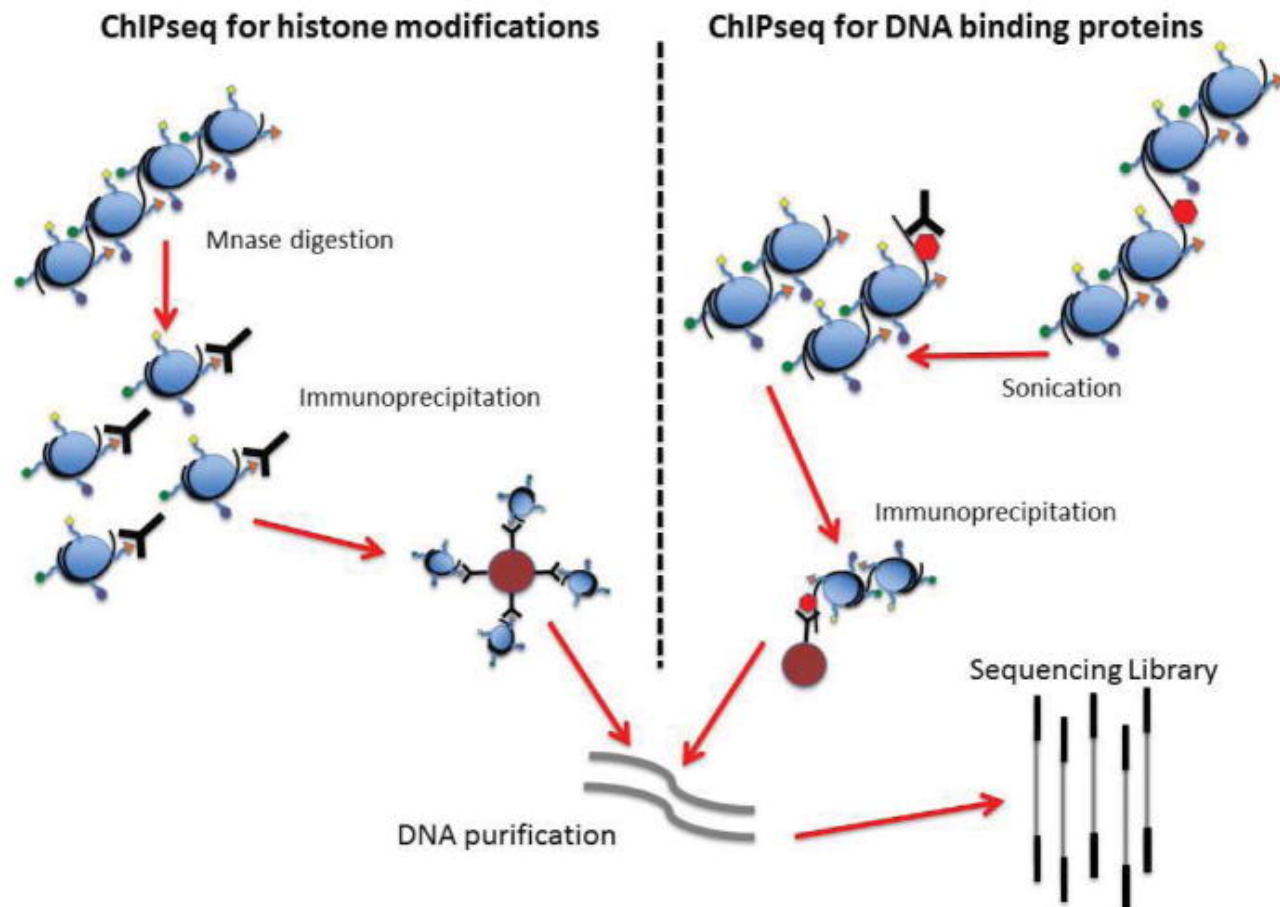
A) Poly-adenylated RNA is reverse transcribed with an anchored oligo-dT primer carrying a universal primer sequence at its 5' end. Next, poly-nucleotide tailing is used to add a poly(A) tail to the 3' end of the cDNA. This cDNA can now be amplified with universal PCR primers containing an oligo-dT sequence at the 3' end. Amplified cDNA can then be used in a standard DNA library construction protocol. B) An anchored oligo-dT primer initiates cDNA synthesis and adds a universal primer sequence. Next, the cDNA is polynucleotide tailed by the RT, producing a 3' overhanging tail. Template switching is initiated on the 3' end of the cDNA by hybridization of a second universal primer sequence containing complementary bases at its 3' end. The template switching oligonucleotide is 3' blocked (*) to prevent extension by the polymerase, whereas the 3' end of the cDNA is extended to copy the second universal primer sequence onto the end of the cDNA. The cDNA can now be amplified by PCR. The PCR products created are then taken into a standard library protocol. C) cDNA synthesis is initiated using a barcoded (orange) and anchored oligo-dT primer containing an Illumina adapter sequence (green) and T7 promoter sequence (purple) at the 5' end. After second strand cDNA synthesis, the fully duplex T7 promoter element is used to initiate *in vitro* transcription and generate cRNA copies of the cDNA with the 5' Illumina adapter and barcode. Finally, a second Illumina adapter is ligated to the 3' end of the cRNA. Doing a final RT-PCR amplification completes the construction of the library.



NGS

Chromatin immunoprecipitation-sequencing (**ChIP-seq**) procedure for detecting sequences at the sites of histone modifications or the recognition sequences of DNA binding proteins

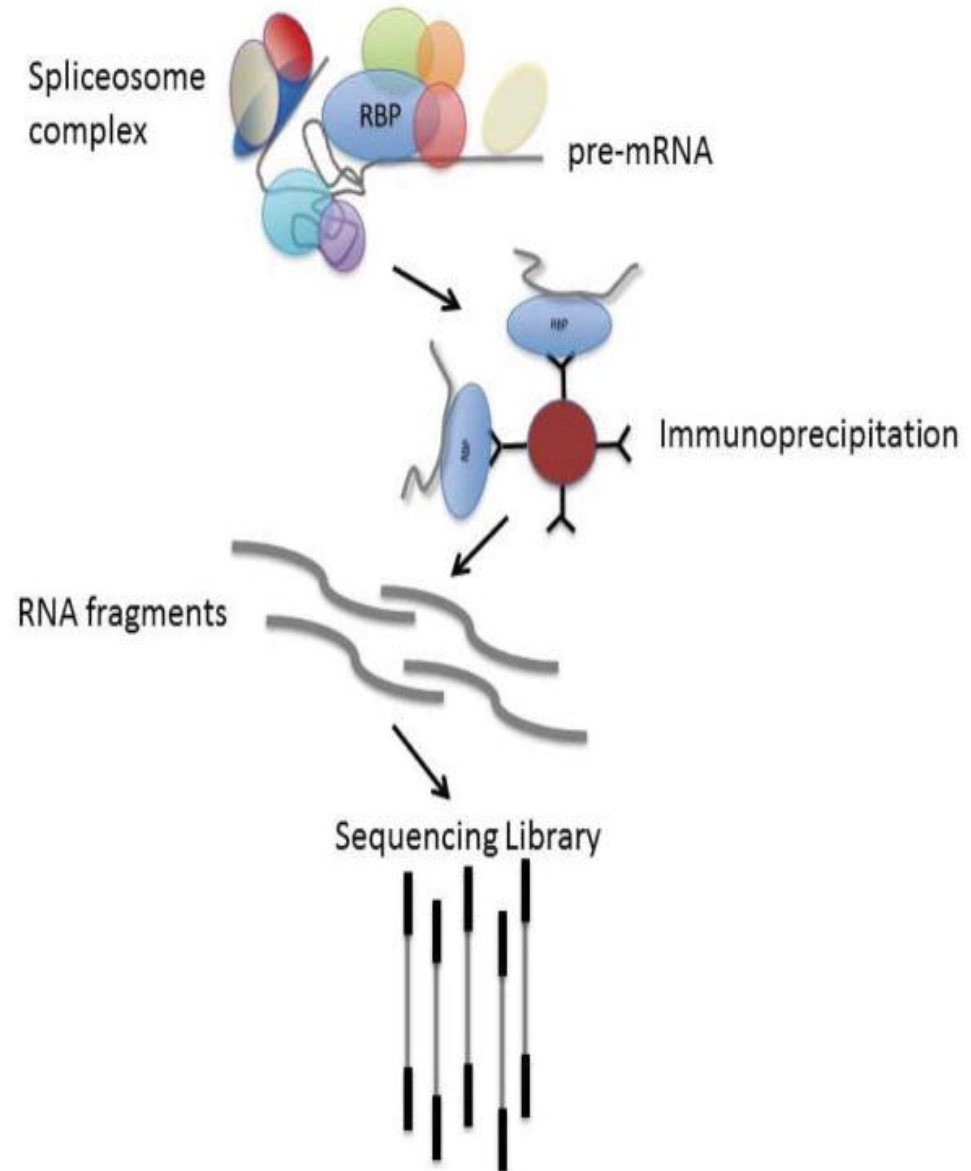
Chromatin is crosslinked, fragmented either by micrococcal nuclease digestion or by sonication, and then incubated with antibodies for either the histone modification or protein of interest. Immunoprecipitation is performed using either Protein A or Protein G beads. After washing, the DNA is uncrosslinked, eluted from the beads and purified, at which point the DNA can be taken into standard DNA library construction protocols.



NGS

RNA immunoprecipitation (**RIP-seq**) done by targeting RNA binding proteins (RBPs)

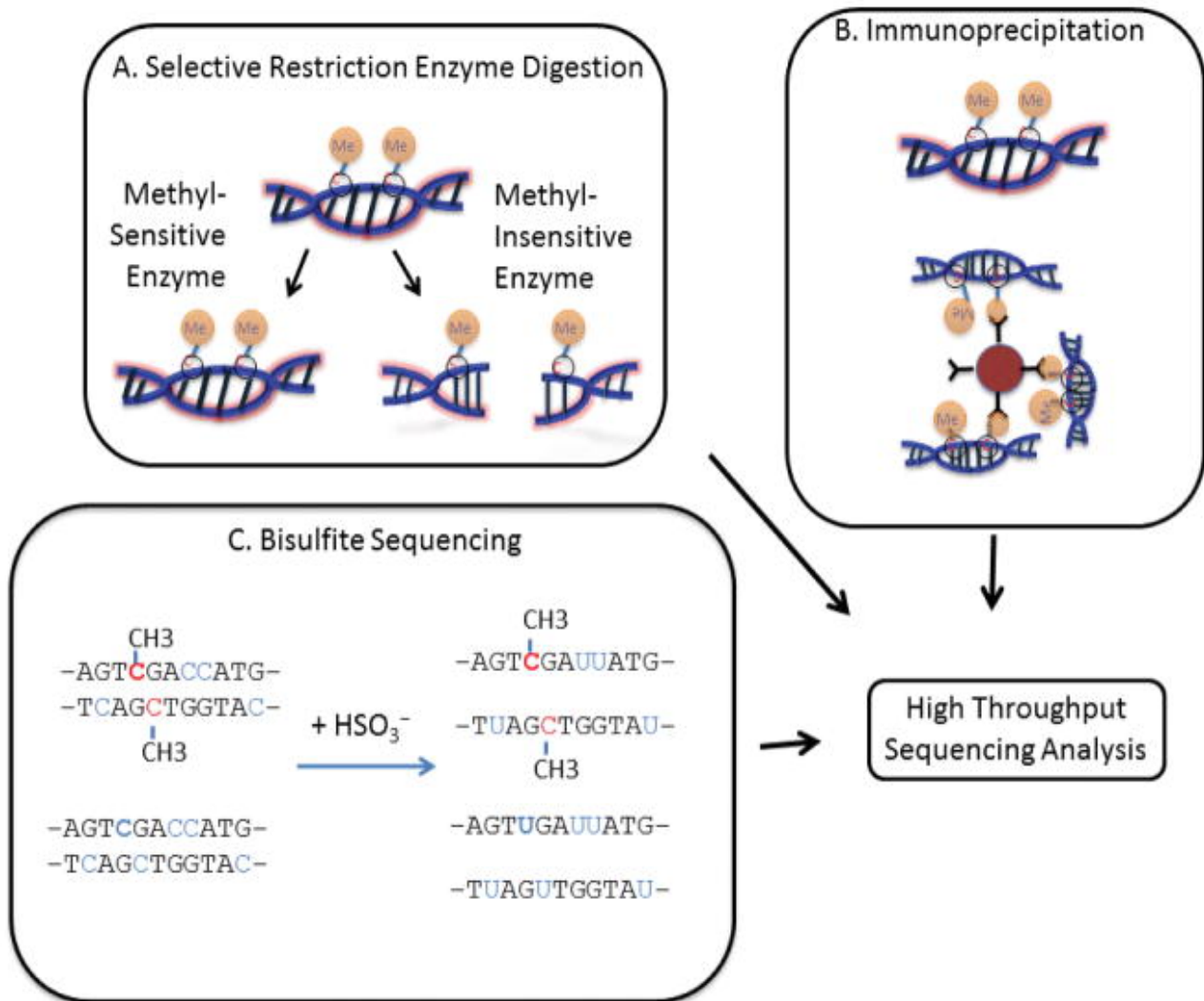
The basic principle of **RIP-seq** is immunoprecipitation of RBPs that are bound to target RNA molecules. The RNA molecules are then purified and a sequencing library is created. In some protocols, the RBP complex is chemically crosslinked to the target RNA; that crosslinking must be reversed after immunoprecipitation. We have found that crosslinking is not necessary for simple RIP-seq where the objective is to identify the RNA molecules bound by RBP, but it is required for CLIP-seq protocols that are used to identify the specific sequence motifs for RBP binding. The immunoprecipitation step can be done with antibodies directed at the specific RBP of interest, or the RBP can be tagged and expressed in the cells under study..



NGS

Approaches for the study of CpG methylation epigenetics (**Methylseq**)

A) A combination of methyl-sensitive and methyl-insensitive restriction enzymes can be used to selectively identify and compare the CpG methylation status of specific regions of sequence. **B)** Antibodies that specifically recognize methylated cytosines can be used to immunoprecipitate DNA fragments, followed by deep sequencing. **C)** Chemical treatment of DNA with sodium bisulfite results in the conversion of unmethylated cytosines to uracils. In contrast, methylated cytosines are protected. Subsequently, deep sequencing of these libraries reveals the methylation status of individual nucleotides.



NGS

At the end of the NGS experiment:

It is crucial to understand that **no matter what you did** during the library preparation step you will get a LARGE collection of (relatively) short DNA fragments sequences: the **reads**

DNA lib → **genome** reads

RNA lib → **RNA** reads (ready to be mapped on the genome)

ChIP-seq → genome fragments in **euchromatin (open state)**

Methylseq → genome reads in **methylated CpG regions**

RIP-seq → genome reads corresponding to **RNA binding sites**

MiRNA-seq → genome reads corresponding to **microRNAs**

...

NGS

Computational analysis of NGS reads. After alignment you can see that:

DNA-seq reads → a given person has (say) at position 1,250,370 of chromosome 18 a nucleotide C instead of a nucleotide A. This is a single nucleotide variant (SNV)

RNA-seq reads → the number of reads mapped for each gene is directly proportional to the activity state of a gene (more transcription => more reads). So you can **measure** the **expression level** of a gene for a given sample/person.

...

NGS

Whats the data like : the **FASTQ** file format

```
@WIGTC-HISEQ:4:1107:1232:1988#TTAGGC/1;0
TGAAACTATTTTCACCCAGACAGATGCCATATTTGAATTC
+WIGTC-HISEQ:4:1107:1232:1988#TTAGGC/1;0
]`Z`^`RS\_baas^__bPR_J^V\\[VbR[\[_aSI^V^B
@WIGTC-HISEQ:4:1107:1117:1992#TTAGGC/1;1
GTGGGGATGTTGCGACTGGATTCATGGCAACTCCTCTGACA
+WIGTC-HISEQ:4:1107:1117:1992#TTAGGC/1;1
___eeecgbeefghffhiffiiiiifhbbghhhhhfhfb
@WIGTC-HISEQ:4:1107:1647:1958#TTAGGC/1;1
CTGTAATTGGCTTCCGACGACTTGGGAATGATAGCATCGAA
+WIGTC-HISEQ:4:1107:1647:1958#TTAGGC/1;1
\_S`cdeffeggfghfihhihiifghbfffhifhfhfgh
@WIGTC-HISEQ:4:1107:1629:1991#TTAGGC/1;1
GGCAACAGCGGTCTTGGAGACGGCAGCAGCGGTACCTCCT
+WIGTC-HISEQ:4:1107:1629:1991#TTAGGC/1;1
__bJ`cdeffceghhhihiffdghgghihfdUedgibg]
@WIGTC-HISEQ:4:1107:1516:1994#TTAGGC/1;1
GTCCATCGAGCCATGGGGTCTTGACTGTGGTGATGAAGAA
+WIGTC-HISEQ:4:1107:1516:1994#TTAGGC/1;1
_abeeeeegfggiiiiicfhihihihhiiegbgffhhi
@WIGTC-HISEQ:4:1107:2130:1974#TTAGGC/1;1
GTCCGTCGTTTCCTGGTGCTCCTGGTTGTCCATCAGCTCC
+WIGTC-HISEQ:4:1107:2130:1974#TTAGGC/1;1
bb_eeeeegfggghiiiffgihhhhfihhhifhfiihhiii
@WIGTC-HISEQ:4:1107:2078:1977#TTAGGC/1;1
ATGGAGTTGTCTCAAACGTCTGCACGATCTCCTTCACGAT
+WIGTC-HISEQ:4:1107:2078:1977#TTAGGC/1;1
bbbeeedeggggghiiiiifgiiiiiihiiiihiiiihihh
```

Orange -> Sequence data

Line 1: Read identifying metadata

@WIGTC-HISEQ -> Instrument name

4 -> Flowcell lane #4

1107:1117:1992 -> X,Y and tile #

#TTAGGC -> Barcode

/1 -> Forward read (/2 is reverse read)

Line 2:

ATCG... The actual nucleotide sequence data

Blue -> Quality data

Line 3: Read identifying metadata (same as line 1)

Line 4: Quality data. 1 character per base.

http://en.wikipedia.org/wiki/FASTQ_format

Quality data can be in different encodings!

Check encoding of a given FASTQ file with the FastQC program (on galaxy and standalone)

FASTQ groomer on Galaxy can convert quality score encodings

NGS

Whats the data like : the **FASTQ** file format

```
@WIGTC-HISEQ:4:1107:1232:1988#TTAGGC/1;0
TGAAACTATTTTCACCCAGACAGATGCCATATTTGAATTC
+WIGTC-HISEQ:4:1107:1232:1988#TTAGGC/1;0
] \z ` `RS \_baaS^__bPR_J^V\ \[VbR[\[_aSI^V^B
```

Single read information:

. 4 lines

NGS

Whats the data like : the **FASTQ** file format

```
@WIGTC-HISEQ:4:1107:1232:1988#TTAGGC/1;0
TGAAACTATTTTCACCCAGACAGATGCCATATTTGAATTC
+WIGTC-HISEQ:4:1107:1232:1988#TTAGGC/1;0
] \Z ``RS\_baaS^__bPR_J^V\\[VbR[\[_aSI^V^B
```

Orange -> Sequence data

Line 1: Read identifying metadata

@WIGTC-HISEQ -> Instrument name

4 -> Flowcell lane #4

1107:1117:1992 -> X,Y and tile #

#TTAGGC -> Barcode

/1 -> Forward read (/2 is reverse read)

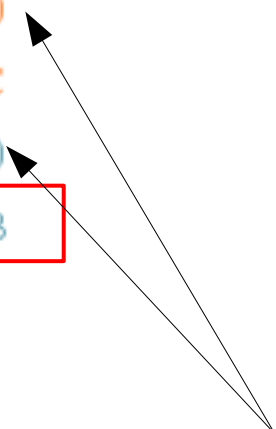
Line 2:

ATCG... The actual nucleotide sequence data

NGS

Whats the data like : the **FASTQ** file format

```
@WIGTC-HISEQ:4:1107:1232:1988#TTAGGC/1;0
TGAAACTATTTTCACCCAGACAGATGCCATATTTGAATTC
+WIGTC-HISEQ:4:1107:1232:1988#TTAGGC/1;0
] \Z ` `RS \_baaS^__bPR_J^V\[VbR[\[_aSI^V^B
```



Blue -> Quality data

Line 3: Read identifying metadata (same as line 1)

Line 4: Quality data. **1 character per base.**

http://en.wikipedia.org/wiki/FASTQ_format

Quality data can be in different encodings!

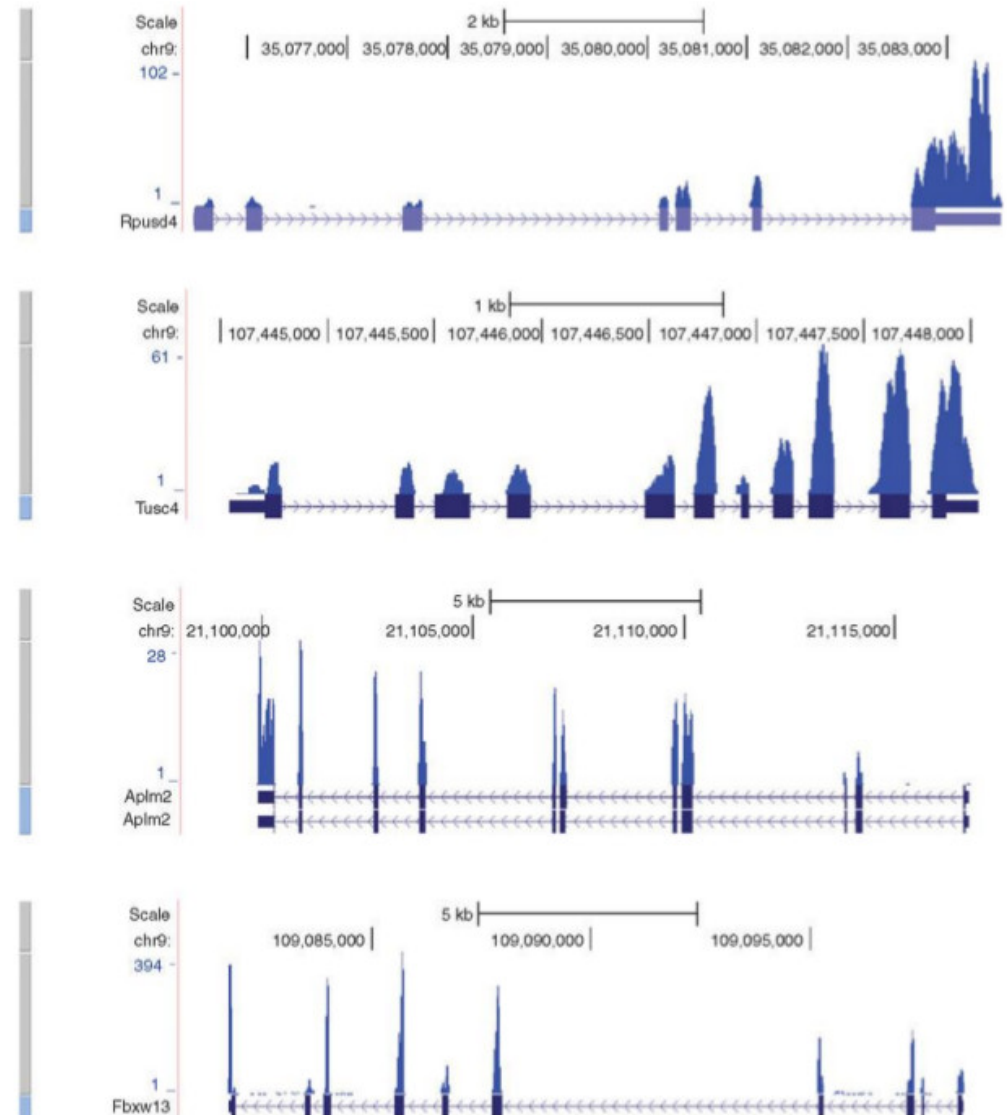
Check encoding of a given FASTQ file with the FastQC program (on galaxy and standalone)

FASTQ groomer on Galaxy can convert quality score encodings

NGS

FASTQ file → gene transcription quantification

1. Start with raw reads: FASTQ
2. Some quality filtering, dropping bad reads, removing or trimming reads include sequence from adaptors
3. “Map” to a reference genome using a splice-junction-aware mapper/sequence aligner. TopHat is commonly used. Produces .bam file.
4. Count how many reads map to a given gene. That count is proportional to the abundance of that transcript in the original sample. Cuffdiff of the Cufflinks suite can do this. Produces a spreadsheet.



NGS

NGS the problem of alignment

Program	Website	Open source?	Handles ABI color space?	Maximum read length
Bowtie	http://bowtie.cbcb.umd.edu	Yes	No	None
BWA	http://maq.sourceforge.net/bwa-man.shtml	Yes	Yes	None
Maq	http://maq.sourceforge.net	Yes	Yes	127
Mosaik	http://bioinformatics.bc.edu/marthlab/Mosaik	No	Yes	None
Novoalign	http://www.novocraft.com	No	No	None
SOAP2	http://soap.genomics.org.cn	No	No	60
ZOOM	http://www.bioinfor.com	No	Yes	240

From: [Nat Biotechnol. Author manuscript; available in PMC 2010 May 1.](#)

Published in final edited form as:

Nat Biotechnol. 2009 May; 27(5): 455–457.

doi: 10.1038/nbt0509-455.

Recent software tools allow the mapping (alignment) of **millions or billions** of short reads to a reference genome.

. For the human genome, this would take thousands of hours using BLAST.

. Reads may come from regions of repetitive DNA (exacerbated by sequencing errors)

NGS

Alignment to a reference genome: example of short-read alignment (Bowtie) results

References to which reads match

reads

quality scores

GA-CS_7_1_743_1919	-	241C3	9156	ATTTAAATCAAATTTTTCTCTATAAC	0;7III6IIII99C9;I;IIIIIII\$	0
GA-CS_7_1_208_1926	+	766H19	71940	GTATCATCGGCCATGGTCACTCATAT	\$I8IG@I@I9B=BCA5I'2/) .,)+0	0
GA-CS_7_1_176_1936	+	760L22	132731	GGGGGAAGTAATAGATTTACGGGTCA	\$IIIIIIIIIIII3I=III=?;II?=	0
GA-CS_7_1_157_1959	+	957L9	111040	GTTTCCTTATCTGTAGAAGGGGGTAA	\$IIIIIIIIIIIGIIIEIIII9II2I>,@	0
GA-CS_7_1_876_1939	+	760L22	126907	GCATTAGCAAACCTAAAAAAATGTTT	\$IIIIIIIIIIIIII@F:<9=3II:I	0
GA-CS_7_1_681_1981	+	760L22	102970	GATTGAATATCAGGTCTGGTACAAAA	\$IGIIIFIIIIICDBI4) II<8766&*	0
GA-CS_7_1_248_744	-	241C3	98493	TGTATCCATATACTTACAGTTTCAAC	&9,89087II+E5</4>+II4I8II\$	0
GA-CS_7_1_625_1953	-	205J11	7292	ACAAGCCTCTAGAAACAGATAGTTTC	+>:<0:34@>?II6IIIIIDIIII?EI\$	0
GA-CS_7_1_650_1988	-	100J8	117470	TTTGAAAAGAAGGTGGTGAAAAATTC	,19ICII8FIAGHAIIIIIIIII@II\$	1
GA-CS_7_1_206_1844	-	760L22	92090	TTAAAGTCTTTTGCAAGCTGTGTCAC	04)2) .8.31;;+>7+E:6I2IF2I\$	0

NGS

Alignment can be used to detect genetic variation

2660	A	37	@,,,,,,,,,T,,,,,,,,,,,,,
2661	G	31	@,,,,,,,,,,,,,
2662	G		←,,,,,,,,,,,,,
2663	A	31	@,,,,,g,,,,,,,,,,,,,
2664	A	30	@,,,,,,,,,,,,,
2665	G	28	@,,,,,,,,,,,,,
2666	G	28	@,,,,,,,,,,,,,
2667	G	28	@,,,,,,,,,,,,,
2668	A	28	←,,,,,,,,,,,,,
2669	C	25	@,,,,,,,,,,,,,
2670	A	27	@,,,,,,,,,,,,,
2671	A	27	@,,,,,,,,,,,,,
2672	T	29	@,,,,,,,,,,,,,
2673	G	28	@,,,,,,,,,,,,,
2674	A	29	@ggGGGGGGGggggGggggGgggGggGg
2675	G	28	@,,,,,,,,,,,,,
2676	G	27	@,,,,,,,,,,,,,
2677	G	27	@,,,,,,,,,,,,,
2678	A	26	@,,,,,T,,,,,,,,,,,,,
2679	A	28	@,,,,,,,,,,,,,
2680	G	28	@,,,,,,,,,,,,,
2681	C	25	@,,,,,,,,,,,,,
2682	A	27	@,,,,,,,,,,,,,
2683	A	27	@,,,,,,,,,,,,,
2684	G	24	@,,,,,,,,,,,,,
2685	G	24	@,,,,,,,,,,,,,
2686	A	24	@,,,,,T,,,,,,,,,,,,,
2687	G	24	@,,,,,,,,,,,,,
2688	A	23	@,,,,,,,,,,,,,
2689	G	24	@,,,,,,,,,,,,,
2690	C	25	@,,,,,,,,,,,,,
2691	A	27	@,,,,,,,,,,,,,
2692	G	27	@,,,,,,,,,,,,,
2693	C	27	@,,,,,,,,,,,,,
2694	T	27	@,,,,,,,,,,,,,
2695	A	27	@,,,,,,,,,,,,,
2696	G	28	@,,,,,,,,,,,,,

Reference sequence
(5 Mb, fasta format)

Read depth

Reference sequence A;
Sample has G 29 times

. and , denote
agreement with
reference on top,
bottom strands

MAQ analysis

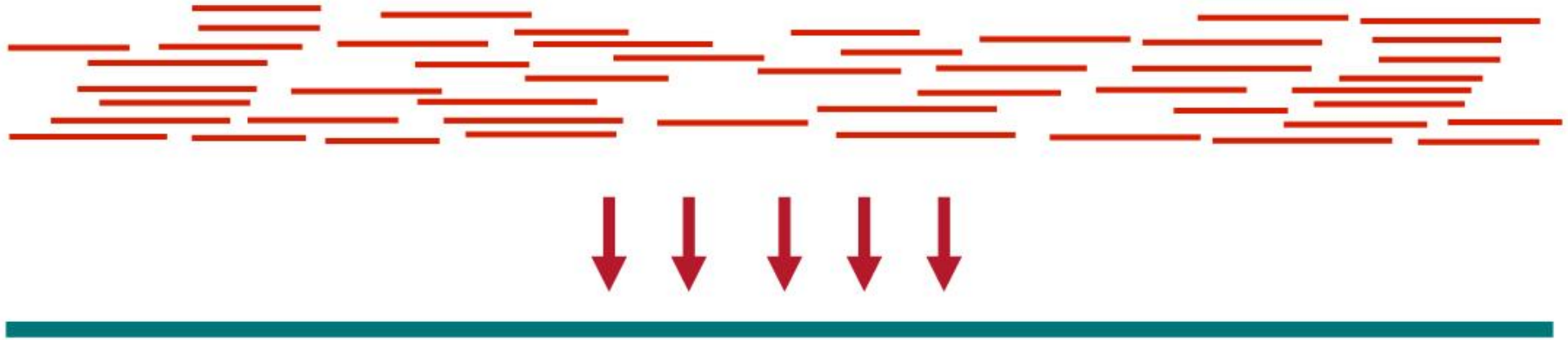
NGS

BWA: a popular short reads aligner

- Aligns short reads (<200 base pairs) to a reference genome
- Fast, accurate
- Learn more at <http://bio-bwa.sourceforge.net/>
- Command-line software for the Linux environment (like essentially all NGS tools)

Read mapping

Read mapping:



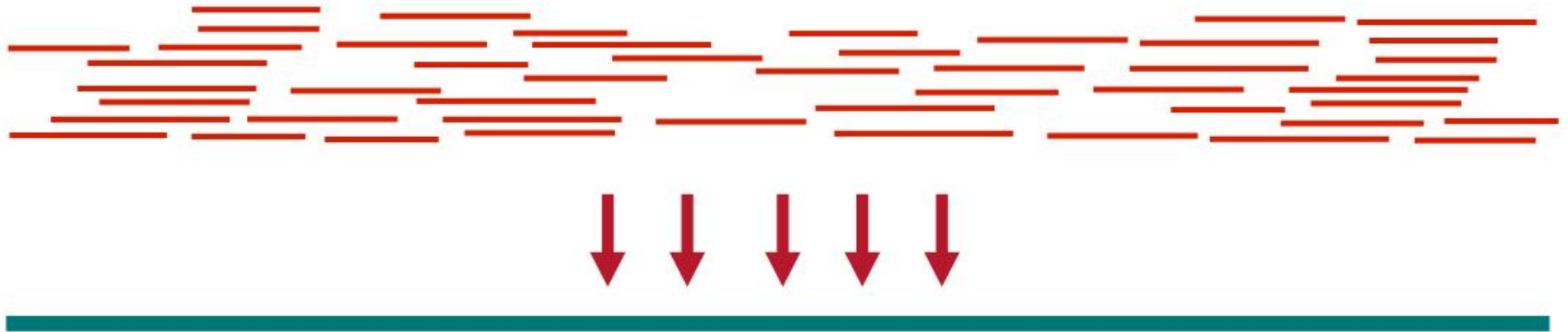
```
CATCGACCGAGCGCGATGCTAGCTAGGTGATCGT.....  
TGCCGCATCGACCGAGCGCGATGCTAGCTAGGTGATCGT...  
GCATGCCGCATCGACCGAGCGCGATGCTAGCTAGGTGATCGT  
GTGCATGCCGCATCGACCGAGCGCGATGCTAGCTAGGTGATC
```

```
.....AGGTGCATGCCGCATCGATCGAGCGCGATGCTAGCTAGCTGATCGT.....
```

- Want ultra fast, highly similar alignment
- Detection of genomic variation

Read mapping

Read mapping:

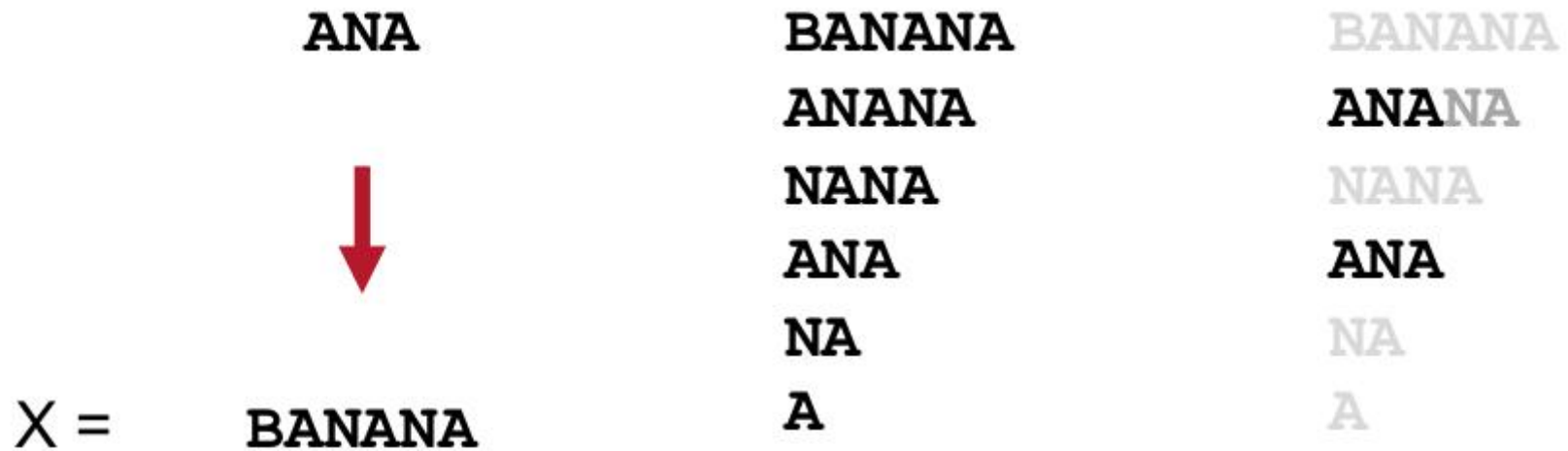


```
          CATCGACCGAGCGCGATGCTAGCTAGGTGATCGT.....
         TGCCGCATCGACCGAGCGCGATGCTAGCTAGGTGATCGT...
        GCATGCCGCATCGACCGAGCGCGATGCTAGCTAGGTGATCGT
       GTGCATGCCGCATCGACCGAGCGCGATGCTAGCTAGGTGATC
.....AGGTGCATGCCGCATCGATCGAGCGCGATGCTAGCTAGCTGATCGT.....
```

- Modern fast read aligners: BWT, Bowtie, SOAP
 - Based on *Burrows-Wheeler transform*

Read mapping

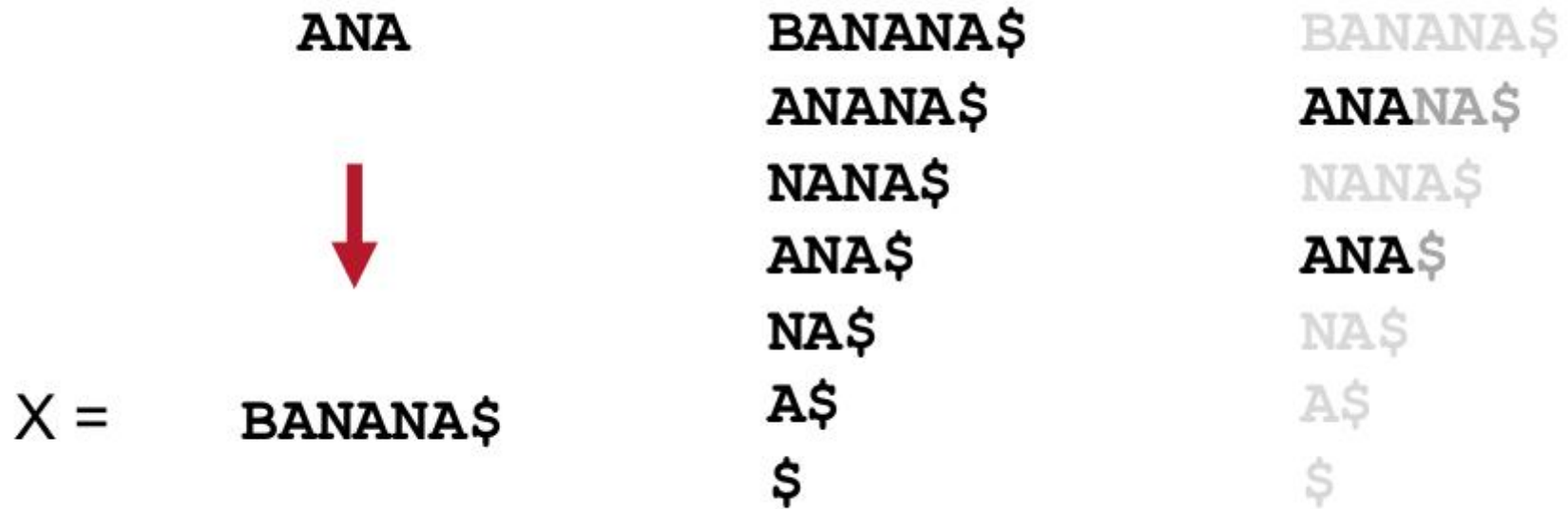
Burrows-Wheeler Transform:



suffixes of
BANANA

Read mapping

Burrows-Wheeler Transform:



Read mapping

Burrows-Wheeler Transform:

ANA



BANANA\$

X =

BANANA\$

ANANA\$B

NANA\$BA

ANA\$BAN

NA\$BANA

A\$BANAN

\$BANANA

BANANA\$

ANANA\$B

NANA\$BA

ANA\$BAN

NA\$BANA

A\$BANAN

\$BANANA

BANANA\$

ANANA\$B

NANA\$BA

ANA\$BAN

NA\$BANA

A\$BANAN

\$BANANA

Read mapping

Burrows-Wheeler Transform:

ANA



X = BANANA\$

BANANA\$

ANANA\$B

NANA\$BA

ANA\$BAN

NA\$BANA

A\$BANAN

\$BANANA

BANANA\$

ANANA\$B

NANA\$BA

ANA\$BAN

NA\$BANA

A\$BANAN

\$BANANA

\$BANANA**A**

A\$BANAN**N**

ANA\$BAN**N**

ANANA\$**B**

BANANA\$

NA\$BAN**A**

NANA\$**B****A**

Read mapping

Burrows-Wheeler Transform:

ANA



BANANA\$

X =

BANANA\$
ANANA\$B
NANA\$BA
ANA\$BAN
NA\$BANA
A\$BANAN
\$BANANA

BANANA\$
ANANA\$B
NANA\$BA
ANA\$BAN
NA\$BANA
A\$BANAN
\$BANANA

\$BANANA
A\$BANAN
ANA\$BAN
ANANA\$B
BANANA\$
NA\$BANA
NANA\$BA

Read mapping

Burrows-Wheeler Transform:

ANA
↓
X = BANANA\$

BANANA\$
ANANA\$B
NANA\$BA
ANA\$BAN
NA\$BANA
A\$BANAN

BANANA\$
ANANA\$B
NANA\$BA
ANA\$BAN
NA\$BANA
A\$BANAN

\$BANANA
A\$BANAN
ANA\$BAN
ANANA\$B
BANANA\$
NA\$BANA
NANA\$BA

BWT matrix of string 'BANANA'

$$\text{BWT}(\text{BANANA}) = \text{ANNB\$AA}$$

Read mapping

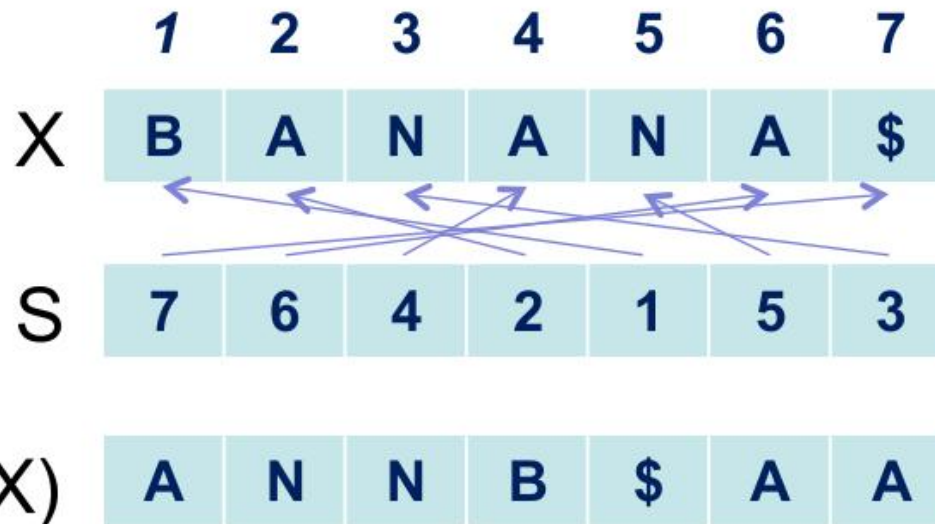
Suffix arrays :

\$BANANA	1	\$BANANA
A\$BANAN	2	A\$BANAN
ANA\$BAN	3	ANA\$BAN
ANANA\$B	4	ANANA\$B
BANANA\$	5	BANANA\$
NA\$BANA	6	NA\$BANA
NANA\$BA	7	NANA\$BA

Suffixes are sorted in the BWT matrix

Define suffix array S:

$S(i) = j$, where $X_j \dots X_n$ is the i -th suffix lexicographically



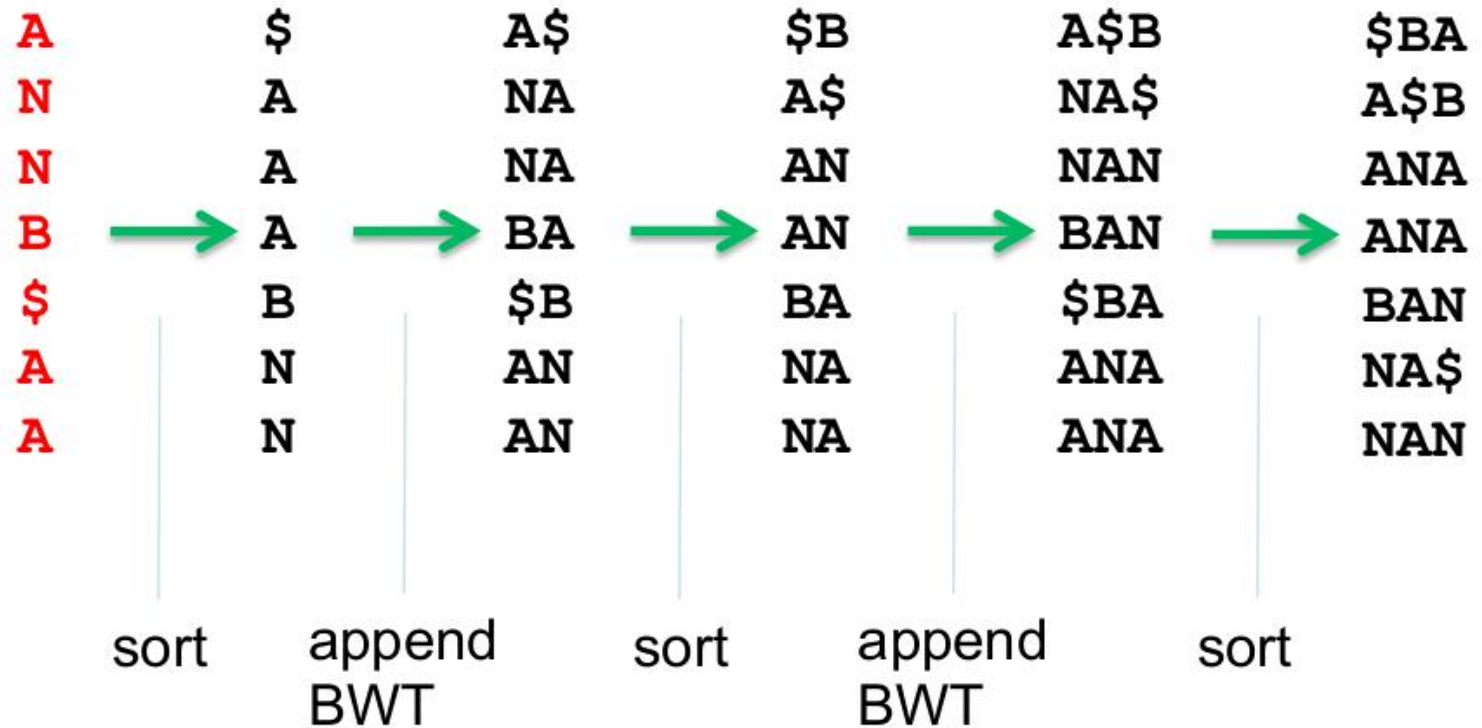
BWT(X) constructed from S:
At each position, take the letter to the left of the one pointed by S

Read mapping

Reconstructing BANANA :

\$BANANA
A\$BANAN
ANA\$BAN
ANANA\$B
BANANA\$
NA\$BANA
NANA\$BA

BWT matrix of string 'BANANA'



Read mapping

Reconstructing BANANA - faster:

```
$BANANAA  
A$BANANN  
ANA$BANN  
ANANA$B  
BANANA$  
NA$BANA  
NANA$BA
```

BWT matrix of
string 'BANANA'

Lemma. The i -th occurrence of character c in last column is the same text character as the i -th occurrence of c in the first column

```
$BANANAA  
A$BANANN  
ANA$BANN  
ANANA$B  
BANANA$  
NA$BANA  
NANA$BA
```

Read mapping

Reconstructing BANANA - faster:

Lemma. The i -th occurrence of character c in last column is the same text character as the i -th occurrence of c in the first column

```
$BANANA
A$BANAN
ANA$BAN
ANANA$B
BANANA$
NA$BANA
NANA$BA
```

```
A $BANAN
N A$BANA
N ANA$BA
B ANANA$
$ BANANA
A NA$BAN
A NANA$B
```

BWT matrix of
string 'BANANA'

Read mapping

Reconstructing BANANA - faster:

Lemma. The i -th occurrence of character c in last column is the same text character as the i -th occurrence of c in the first column

\$BANANA
A\$BANAN
ANA\$BAN
ANANA\$B
BANANA\$
NA\$BANAN
NANA\$BA

A	\$BANAN
N	A\$BANA
N	ANA\$BA
B	ANANA\$
\$	BANANA
A	NA\$BAN
A	NANA\$B

A	\$BANAN
A	NA\$BAN
A	NANA\$B

} Same words,
same sorted order

BWT matrix of
string 'BANANA'

Read mapping

Reconstructing BANANA - faster:

```
$BANANA
A$BANAN
ANA$BAN
ANANA$B
BANANA$
NA$BANA
NANA$BA
```

BWT matrix of
string 'BANANA'

Lemma. The i -th occurrence of character 'a' in last column is the same text character as the i -th occurrence of 'a' in the first column

LF(): Map the i -th occurrence of character 'a' in last column to the first column

LF(r): Let row r contain the i -th occurrence of 'a' in last column

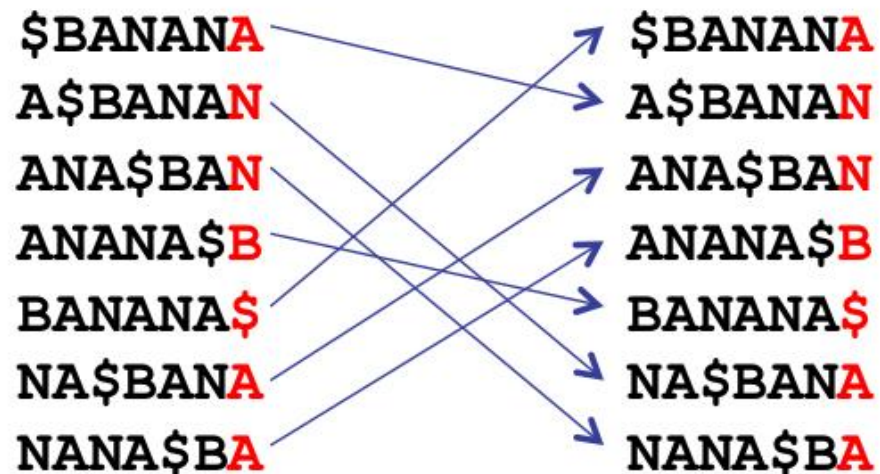
Then, $LF(r) = r'$; r' : i -th row starting with 'a'

Read mapping

Reconstructing BANANA - faster:

LF(r): Let row r be the i-th occurrence of 'a' in last column
Then, LF(r) = r'; r': i-th row starting with 'a'

\$BANANA
A\$BANAN
ANA\$BAN
ANANA\$B
BANANA\$
NA\$BANA
NANA\$BA



LF[] = [2, 6, 7, 5, 1, 3, 4]

Row LF(r) is obtained by rotating row r one position to the right

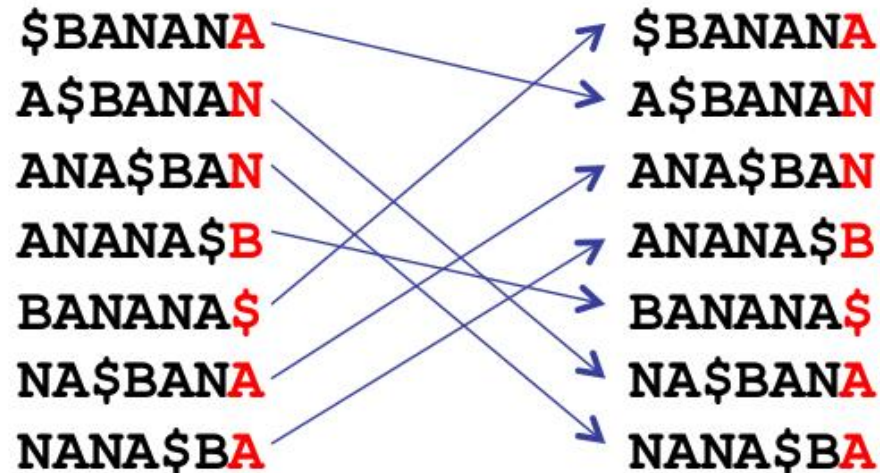
BWT matrix of string 'BANANA'

Read mapping

Reconstructing BANANA - faster:

LF(r): Let row r be the i -th occurrence of 'a' in last column
Then, $LF(r) = r'$; r' : i -th row starting with 'a'

\$BANANA
A\$BANAN
ANA\$BAN
ANANA\$B
BANANA\$
NA\$BANAN
NANA\$BA



LF[] = [2, 6, 7, 5, 1, 3, 4]

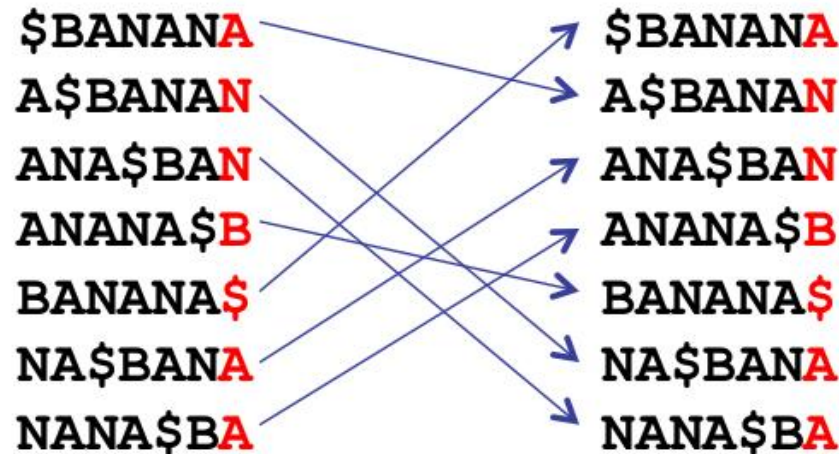
Therefore, the last character in row $LF(r)$ is the character before the last character in row r

BWT matrix of string 'BANANA'

Read mapping

Reconstructing BANANA - faster:

\$BANANA
A\$BANAN
ANA\$BAN
ANANA\$B
BANANA\$
NA\$BANA
NANA\$BA



$$LF[] = [2, 6, 7, 5, 1, 3, 4]$$

Computing LF() is easy:

Let $C(a)$: # of characters smaller than 'a'

Example: $C(\$) = 0$; $C(A) = 1$; $C(B) = 4$; $C(N) = 5$

Let row r end with the i -th occurrence of 'a' in last column

Then, $LF(r) = C(a) + i$

BWT matrix of
string 'BANANA'

Read mapping

Reconstructing BANANA - faster:

\$BANANA
A\$BANAN
ANA\$BAN
ANANA\$B
BANANA\$
NA\$BAN
NANA\$BA

BWT matrix of
string 'BANANA'

	A	N	N	B	\$	A	A	
C()	1	5	5	4	0	1	1	C() copied for convenience
index i	1	1	2	1	1	2	3	indicating this is i-th occurrence of 'c'
LF()	2	6	7	5	1	3	4	LF() = C() + i

Reconstruct BANANA:

```
S := "" ; r := 1 ; c := BWT[r] ;  
UNTIL c = '$' {  
    S := cS ;  
    r := LF(r) ;  
    c := BWT(r) ; }  
}
```

Read mapping

Searching for query “ANA”:

$L(W)$: lowest index in BWT matrix where W is prefix
 $U(W)$: highest index in BWT matrix where W is prefix

Example:

$$L(\text{“NA”}) = 6$$

$$U(\text{“NA”}) = 7$$

Lemma (prove as exercise)

$$L(aW) = C(a) + i + 1,$$

where $i = \#$ ‘a’s up to $L(W) - 1$ in $\text{BWT}(X)$

$$U(aW) = C(a) + j,$$

where $j = \#$ ‘a’s up to $U(W)$ in $\text{BWT}(X)$

Example:

$$\begin{aligned} L(\text{“ANA”}) &= C(\text{‘A’}) + \# \text{ ‘A’s up to } (L(\text{“NA”}) - 1) + 1 \\ &= 1 + (\# \text{ ‘A’s up to } 5) + 1 \\ &= 1 + 1 + 1 = 3 \end{aligned}$$

$$U(\text{“ANA”}) = 1 + \# \text{ ‘A’s up to } U(\text{“NA”}) = 1 + 3 = 4$$

\$BANANA
A\$BANAN
ANAS\$BAN
ANANAS\$B
BANANAS\$
NAS\$BANA
NANAS\$BA

BWT matrix of
string ‘BANANA’

Read mapping

Searching for query “ANA”:

```
$BANANA
A$BANAN
ANA$BAN
ANANA$B
BANANA$
NA$BANA
NANA$BA
```

BWT matrix of
string ‘BANANA’

Let

$LFC(r, a) = C(a) + i$, where $i = \#$ ’a’s up to r in BWT

```
ExactMatch(W[1...k]) {
```

```
    a := W[k];
```

```
    low := C(a) + 1;
```

```
    high := C(a+1); // a+1: lexicographically next char
```

```
    i := k - 1;
```

```
    while (low <= high && i >= 1) {
```

```
        a = W[i];
```

```
        low = LFC(low - 1, a) + 1;
```

```
        high = LFC(high, a);
```

```
        i := i - 1; }
```

```
    return (low, high);
```

```
}
```

Read mapping

Summary of BWT algorithm:

Suffix array of string X:

$S(i) = j$, where $X_j \dots X_n$ is the j -th suffix lexicographically

- BWT follows immediately from suffix array
 - Suffix array construction possible in $O(n)$, many good $O(n \log n)$ algorithms
- Reconstruct X from $BWT(X)$ in time $O(n)$
- Search for all exact occurrences of W in time $O(|W|)$
- $BWT(X)$ is easier to compress than X

NGS

BWA, BOWTIE and other aligners produce output in the SAM format

Column	Description
1	QNAME Query (pair) NAME
2	FLAG bitwise FLAG
3	RNAME Reference sequence NAME
4	POS 1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ MAPping Quality (Phred-scaled)
6	CIGAR extended CIGAR string
7	MRNM Mate Reference sequence NaMe ('=' if same as RNAME)
8	MPOS 1-based Mate POSition
9	ISIZE Inferred insert SIZE
10	SEQ query SEquence on the same strand as the reference
11	QUAL query QUALity (ASCII-33 gives the Phred base quality)
12	OPT variable OPTional fields in the format TAG:VTYPE:VALU

NGS

Sequence alignment/map (**SAM**) format and the **BAM** format

- **SAM** is a common format having sequence reads and their alignment to a reference genome.
- **BAM** is the binary form of a SAM file.
- Aligned BAM files are available at repositories (Sequence Read Archive at NCBI, ENA at Ensembl)
- SAMTools is a software package commonly used to analyze SAM/BAM files.
- Visit <http://samtools.sourceforge.net/>

NGS

SAM file format

(1) The query name of the read is given (M01121...)

(2) The flag value is 163 (this equals 1+2+32+128)

(3) The reference sequence name, chrM, refers to the mitochondrial genome

(4) Position 480 is the left-most coordinate position of this read

(5) The Phred-scaled mapping quality is 60 (an error rate of 1 in 10⁶)

(6) The CIGAR string (148M2S) shows 148 matches and 2 soft-clipped (unaligned) bases

```
home/bioinformatics$ samtools view 030c_S7.bam | less
M01121:5:000000000-A2DTN:1:2111:20172:15571      163      chrM
480      60      148M2S      =      524      195      AATCTCATCAAT
ACAACCCTCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATACCCCGAACC
AACCAAACCCCAAAGACACCCCCACAGTTTATGTAGCTTACCTCCTCAAAGCAATAACC
TGAAAATGTTTAGACGGG      BBBBFFFB5@FFGGGFGEggGEGAAACGHFHFEGGAGFFH
AEFDGG?E?EGGGFGHFGHF?FFCHFHO0E@EGFGGEEE1FFEEHGBGEFFFGGGG@</0
1BG212222>F21@F11FGFG1@1?GC<G11?1?FGDGGF=GHFFFHC.-
RG:Z:Sample7      XC:i:148      XT:A:U      NM:i:3      SM:i:37
AM:i:37      X0:i:1      X1:i:0      XM:i:3      XO:i:0      XG:i:0      MD:Z:19C109C0A17
```

(7) An = sign shows that the mate reference matches the reference name

(8) The 1-based left position is 524

(9) The insert size is 195 bases

(10) The sequence begins AATCT and ends ACGGG (its length is 150 bases)

(11) Each base is assigned a quality score (from BBBB ending FHC.-)

(12) This read has additional, optional fields at accompany the MiSeq analysis

NGS

SAMTools tview visualization of aligned reads from a BAM file



The image shows a terminal window displaying the SAMTools tview visualization of aligned reads. The window title is "matt@atlantis: ~/UNIMI/TEACHING/PhDCOURSE_1819/PRACTICALS/SAMTOOLS_inst/samtools_primer/samtools_prim...". The terminal shows a reference sequence at the top: "CGCTGATTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTGTAGAAAGCGCGTGGTCACAACGTTACCGTT". Below the reference sequence, there are multiple rows of aligned reads, each represented by a series of vertical bars (reads) and their corresponding nucleotide bases (A, C, G, T). The reads are aligned to the reference sequence, and some mismatches are visible. The terminal window also shows a menu bar with "File Edit View Search Terminal Tabs Help" and a status bar at the bottom.

There are many tools to view SAM/BAM files. A popular software package (SAMTools, used in Linux) includes **tview** visualization of reads from a BAM file

NGS

Variant Call Format (VCF) file summarizes variation

Column	Mandatory	Description
CHROM	Yes	Chromosome
POS	Yes	1-based position of the start of the variant
ID	Yes	Unique identifier of the variant; the dbSNP entry rs1413368 is given in our example
REF	Yes	Reference allele
ALT	Yes	A comma-separated list of alternate nonreference alleles
QUAL	Yes	Phred-scaled quality score
FILTER	Yes	Site filtering information; in our example it is PASS
INFO	Yes	A semicolon-separated list of additional information. These fields include the gene identifier GI (here the gene is NEGR1); the transcript identifier TI (here NM_173808); and the functional consequence FC (here a synonymous change, T296T).
FORMAT	No	Defines information in subsequent genotype columns; colon separated. For example, GT:AD:DP:GQ:PL:VF:GQX in our example refers to genotype (GT), allelic depths for the ref and alt alleles in the order listed (AD), approximate read depth (reads with MQ=255 or with bad mates are filtered) (DP), genotype quality (GQ), normalized, Phred-scaled likelihoods for genotypes as defined in the VCF specification (PL), variant frequency, the ratio of the sum of the called variant depth to the total depth (VF), and minimum of {genotype quality assuming variant position, genotype quality assuming nonvariant position} (GXQ).
Sample	No	Sample identifiers define the samples included in the VCF file

NGS

Variant Call Format (VCF) file summarizes variation

```
##fileformat=VCFv4.2
##FILTER=<ID=PASS,Description="All filters passed">
##samtoolsVersion=1.3.1+htslib-1.3.2
##samtoolsCommand=samtools mpileup -g -f genomes/NC_008253.fna alignments/sim_reads_aligned.sorted.bam
##reference=file://genomes/NC_008253.fna
##contig=<ID=gj|110640213|ref|NC_008253.1|,length=4938920>
##ALT=<ID=*,Description="Represents allele(s) other than observed.">
##INFO=<ID=INDEL,Number=0,Type=Flag,Description="Indicates that the variant is an INDEL.">
##INFO=<ID=IDV,Number=1,Type=Integer,Description="Maximum number of reads supporting an indel">
##INFO=<ID=IMF,Number=1,Type=Float,Description="Maximum fraction of reads supporting an indel">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Raw read depth">
##INFO=<ID=VDB,Number=1,Type=Float,Description="Variant Distance Bias for filtering splice-site artefacts in RNA-seq data (bigger is better)",Version="3">
##INFO=<ID=RPB,Number=1,Type=Float,Description="Mann-Whitney U test of Read Position Bias (bigger is better)">
##INFO=<ID=MQB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality Bias (bigger is better)">
##INFO=<ID=BQB,Number=1,Type=Float,Description="Mann-Whitney U test of Base Quality Bias (bigger is better)">
##INFO=<ID=MQSB,Number=1,Type=Float,Description="Mann-Whitney U test of Mapping Quality vs Strand Bias (bigger is better)">
##INFO=<ID=SGB,Number=1,Type=Float,Description="Segregation based metric.">
##INFO=<ID=MQ0F,Number=1,Type=Float,Description="Fraction of MQ0 reads (smaller is better)">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="List of Phred-scaled genotype likelihoods">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##INFO=<ID=AF1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele frequency (assuming HWE)">
##INFO=<ID=AF2,Number=1,Type=Float,Description="Max-likelihood estimate of the first and second group ALT allele frequency (assuming HWE)">
##INFO=<ID=AC1,Number=1,Type=Float,Description="Max-likelihood estimate of the first ALT allele count (no HWE assumption)">
##INFO=<ID=MQ,Number=1,Type=Integer,Description="Root-mean-square mapping quality of covering reads">
##INFO=<ID=FQ,Number=1,Type=Float,Description="Phred probability of all samples being the same">
##INFO=<ID=PV4,Number=4,Type=Float,Description="P-values for strand bias, baseQ bias, mapQ bias and tail distance bias">
##INFO=<ID=G3,Number=3,Type=Float,Description="ML estimate of genotype frequencies">
##INFO=<ID=HWE,Number=1,Type=Float,Description="Chi^2 based HWE test P-value based on G3">
##INFO=<ID=DP4,Number=4,Type=Integer,Description="Number of high-quality ref-forward , ref-reverse, alt-forward and alt-reverse bases">
##bcftools_callVersion=1.3.1+htslib-1.3.2
##bcftools_callCommand=call -c -v variants/sim_variants.bcf
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT alignments/sim_reads_aligned.sorted.bam
gj|110640213|ref|NC_008253.1| 736 . T G,C 31.0001. DP=68;VDB=0.827789;SGB=-
0.693147;MQ0F=0;AF1=1;AC1=2;DP4=0,0,0,61;MQ=41;FQ=-204.988 GT:PL 1/1:64,178,0,66,147,60
```

... just **one** SNP identified in this VCF file , it is at position 736

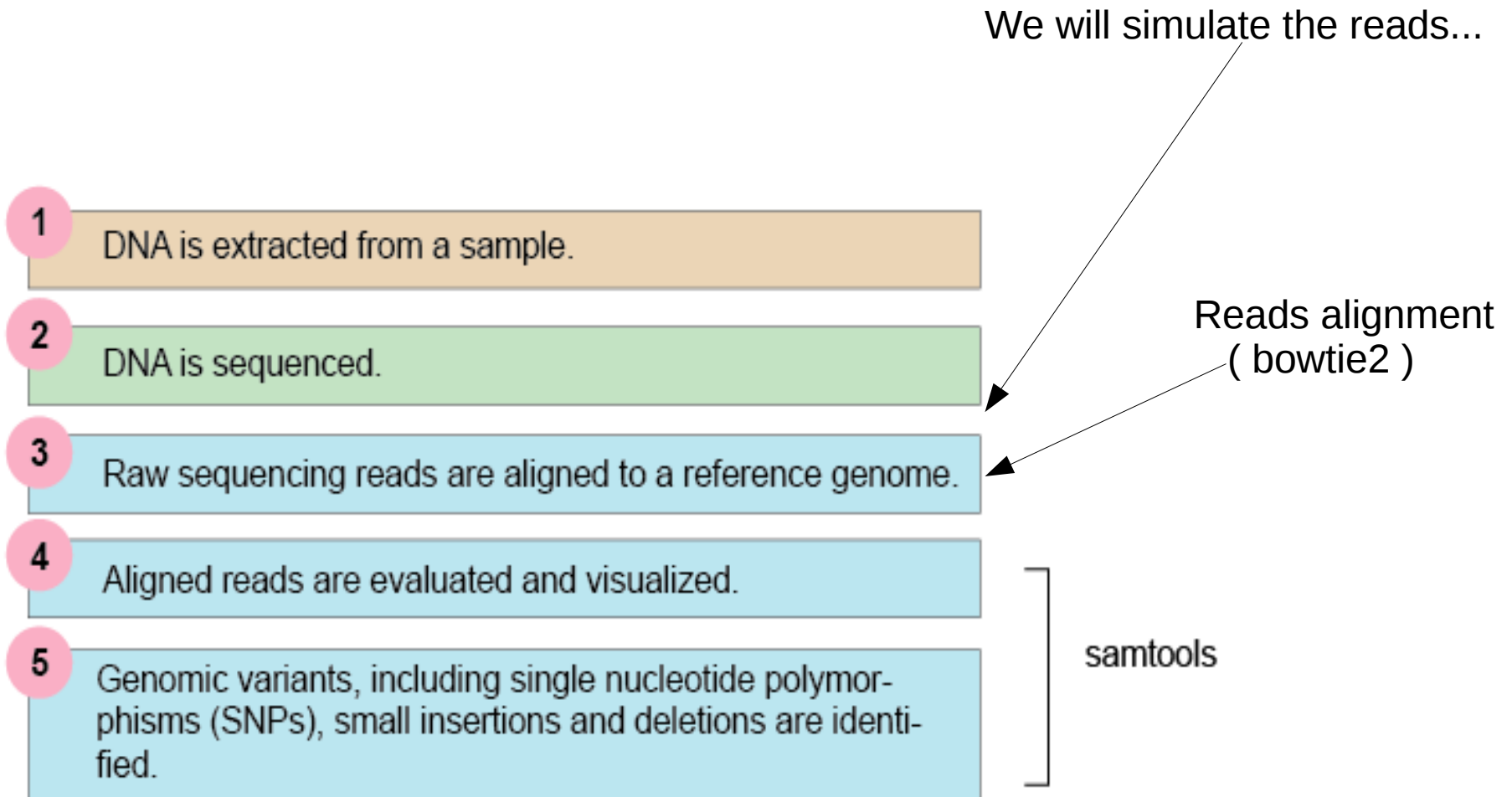
NGS

Variant Call practicals using:

- wgsim (whole genome simulator)**
- bowtie2**
- samtools**
- bcftools**

NGS

Variant Call practicals:



NGS

Variant Call practicals using:

STEP I : reads simulation, script:
generate_simulated_reads.sh

```
#!/bin/sh
```

```
# First, make sure wgsim is installed
```

```
if ! type wgsim > /dev/null 2>&1; then
```

```
    echo "Could not find wgsim. Please adjust your path or download from:
```

```
    https://github.com/lh3/wgsim."
```

```
fi
```

```
# Generate simulated reads
```

```
wgsim -N1000 -S1 genomes/NC_008253_1K.fna simulated_reads/sim_reads.fq /dev/null
```

NGS

Variant Call practicals using:

STEP II : reads alignment, script: align_to_genome.sh

```
#!/bin/sh
```

```
# First, make sure bowtie is installed
```

```
if ! type bowtie2 > /dev/null 2>&1; then
```

```
    echo "Could not find bowtie2. Please adjust your path or download from: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml"
```

```
fi
```

```
# Align the simulated reads against the reference genome
```

```
bowtie2 -x indexes/e_coli -U simulated_reads/sim_reads.fq -S  
alignments/sim_reads_aligned.sam
```

NGS

Variant Call practicals using:

STEP III : variants call, script: identify_variants.sh

```
# Convert SAM to BAM
# -b: output BAM
# -S: input is SAM
printf "\n>Converting SAM to BAM\n"
samtools view -b -S -o alignments/sim_reads_aligned.bam alignments/sim_reads_aligned.sam

# Sort and Index BAM
printf "\n>Sorting and indexing BAM\n"
# this is the way the old version of samtools did this step. Use instead the new sort interface
#samtools sort alignments/sim_reads_aligned.bam alignments/sim_reads_aligned.sorted
samtools sort -o alignments/sim_reads_aligned.sorted.bam alignments/sim_reads_aligned.bam
samtools index alignments/sim_reads_aligned.sorted.bam
```

Part 1/3

NGS

Variant Call practicals using:

STEP III : variants call, script: identify_variants.sh

```
# Identify SNPs: requires two distinct steps.  
# These steps can be piped together, but for clarify, they are issued  
# as two independent steps below.
```

Part 2/3

```
# First, run samtools mpileup to calculate likelihoods  
# -g: generate BCF output (genotype likelihoods)  
# -f: reference sequence file  
printf "\n>Running mpileup\n"  
samtools mpileup -g -f genomes/NC_008253.fna alignments/sim_reads_aligned.sorted.bam >  
variants/sim_variants.bcf
```

NGS

Variant Call practicals using:

STEP III : variants call, script: identify_variants.sh

```
# Second, run bcftools to actually call the SNPs
# -c: SNP calling (force -e)
# -v: output potential variant sites only
# -e: likelihood based analyses
printf "\n>Calling variants with bcftools\n"
# it was the old call variants convention (view) change to call
bcftools call -c -v variants/sim_variants.bcf > variants/sim_variants.vcf

printf "\nAll done.\n"
```

Part 3/3

```
# Then, you can do tview...
#samtools tview alignments/sim_reads_aligned.sorted.bam genomes/NC_008253.fna
```