Genome structure and Organization

By

Aparna Patil Kose

Dept of Bioinformatics

Basics

- The information encoded in genetic material is responsible for establishing and maintaining the cellular and biochemical functions of an organism

- **Genome**: is all genetic material of an organism, which consists of DNA (or RNA in RNA viruses).
  - Includes both the genes (the coding regions) and the noncoding DNA, as well as mitochondrial DNA and chloroplast DNA.

- **Gene**: The basic unit of inheritance, segment of nucleic acid that encodes a functional protein or RNA.

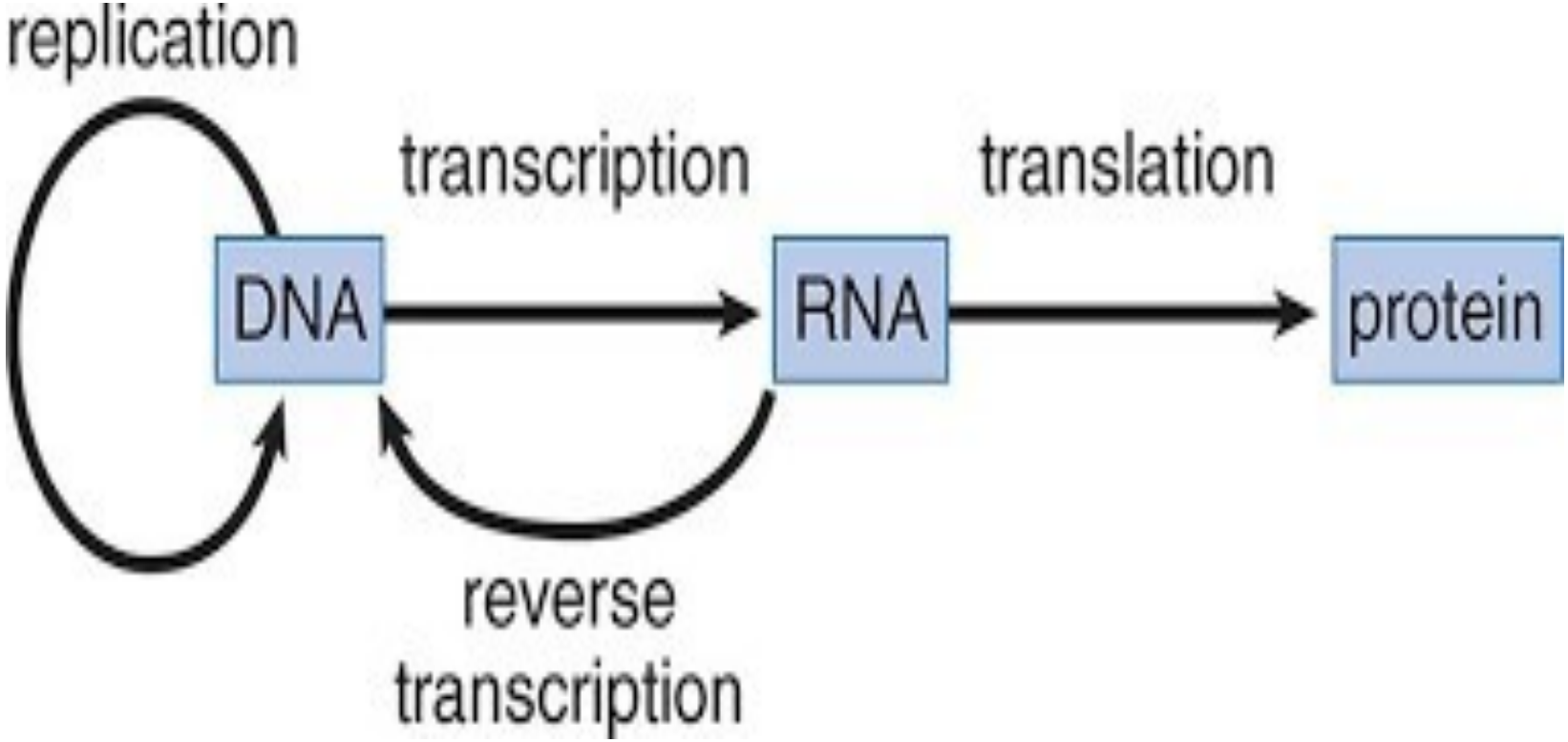
- **Gene expression**: The process by which gene products are made.

- **Chromosome**: Organized package of DNA found in the nucleus of the cell.
  - Different organisms have different numbers of chromosomes. Humans have 23 pairs of chromosomes-22 pairs of autosomes, and one pair of sex chromosomes, X and Y.

- **Allele**: An alternative form of a gene

- **Genotype**: The allelic constitution of a given individual.
  - The genotypes at locus A in a diploid individual may be AA, Aa, or aa

- **Gene cluster**: A group of related genes grouped together on a eukaryotic chromosome.



Central Dogma

Why is genome Organization important?

Genomes are organized into complex higher-order structures by folding of the DNA into chromatin fibers, chromosome domains, and ultimately chromosomes.

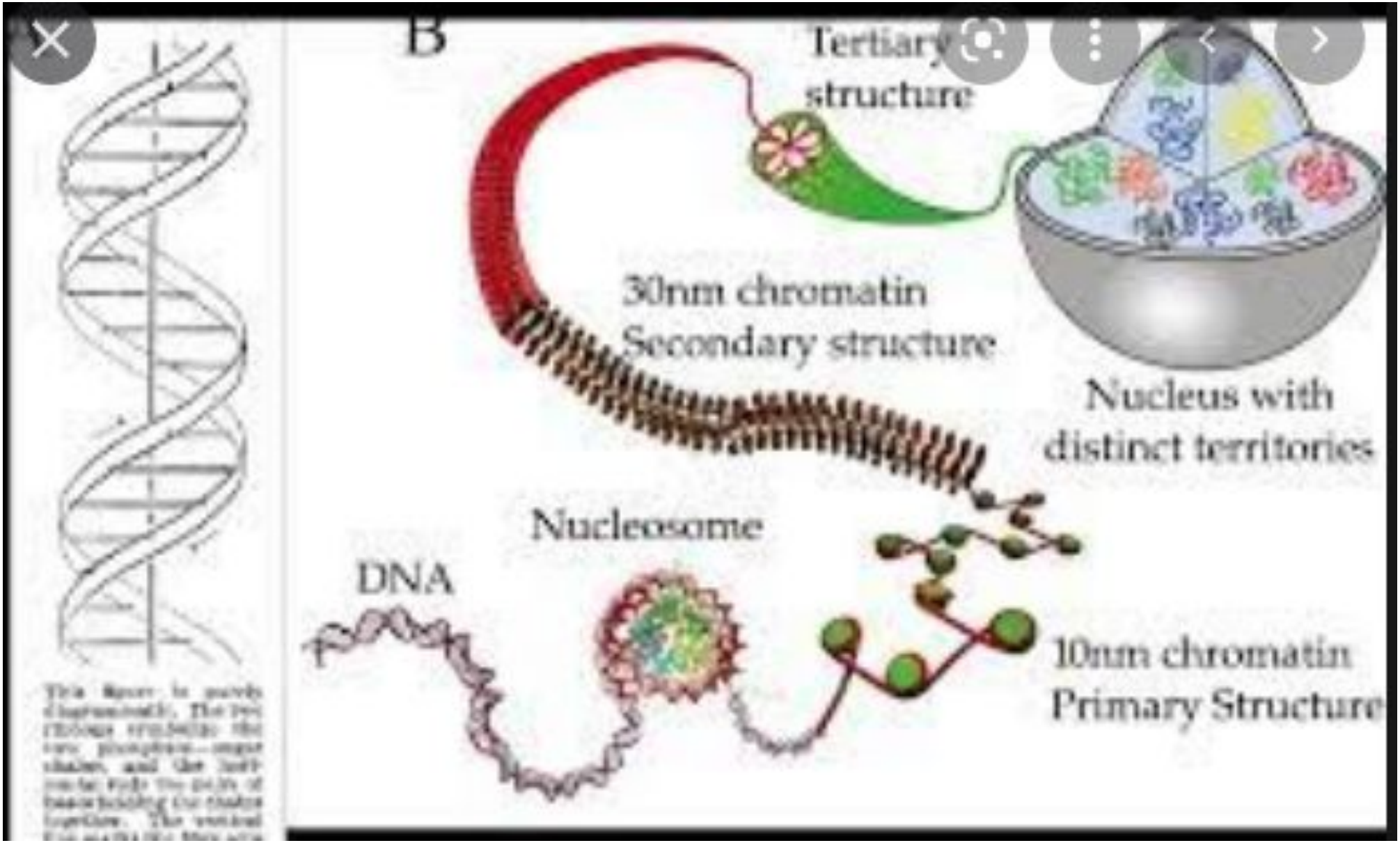
The higher-order organization of genomes is functionally important **for gene regulation and control of gene expression programs**.

Genomic organization

The **hereditary material** i.e. DNA (deoxyribonucleic acid) of an organism is composed of a sequence of four nucleotides in a specific pattern, which encode information as a function of their order.

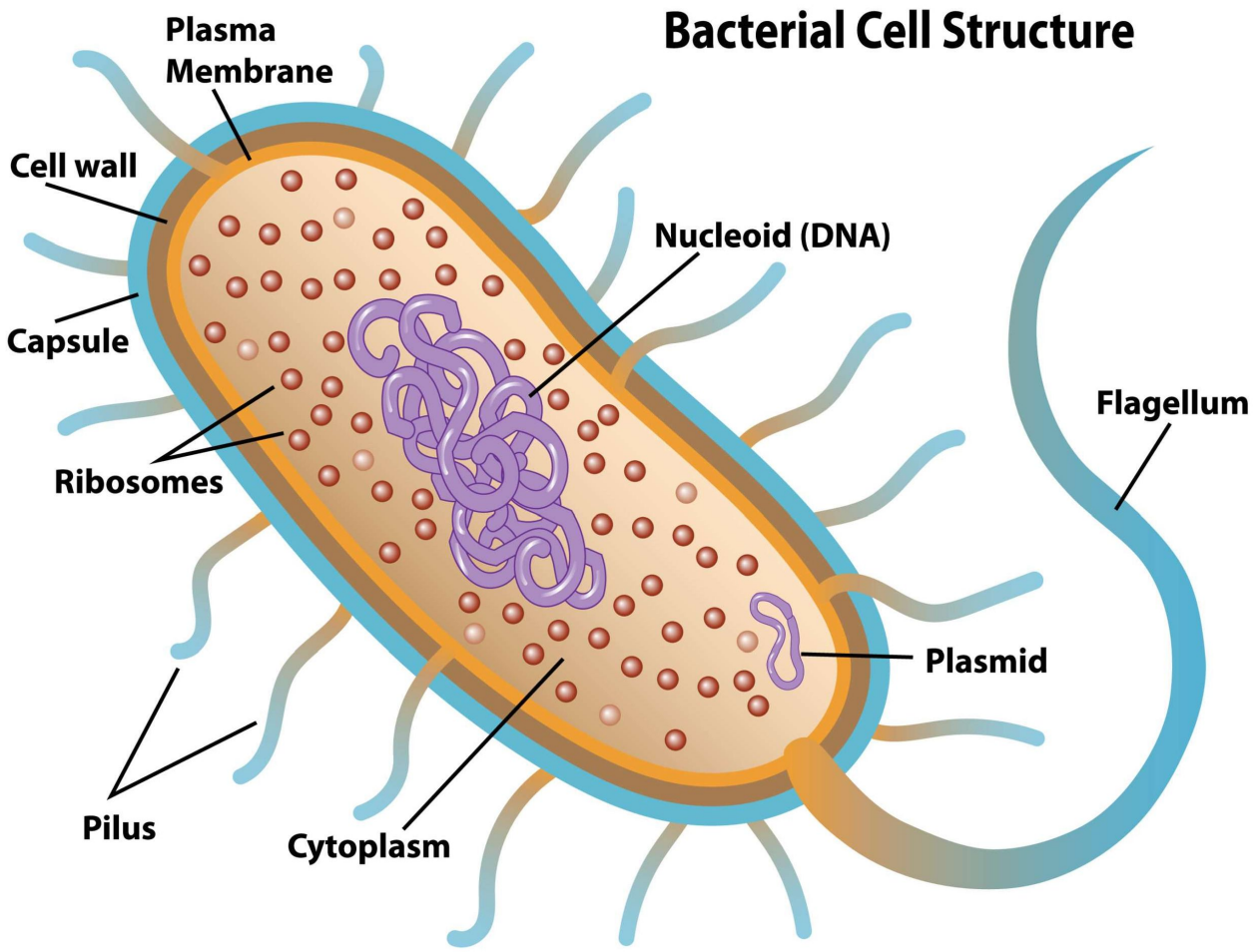
Genomic organization refers to the linear order of DNA elements and their division into chromosomes.

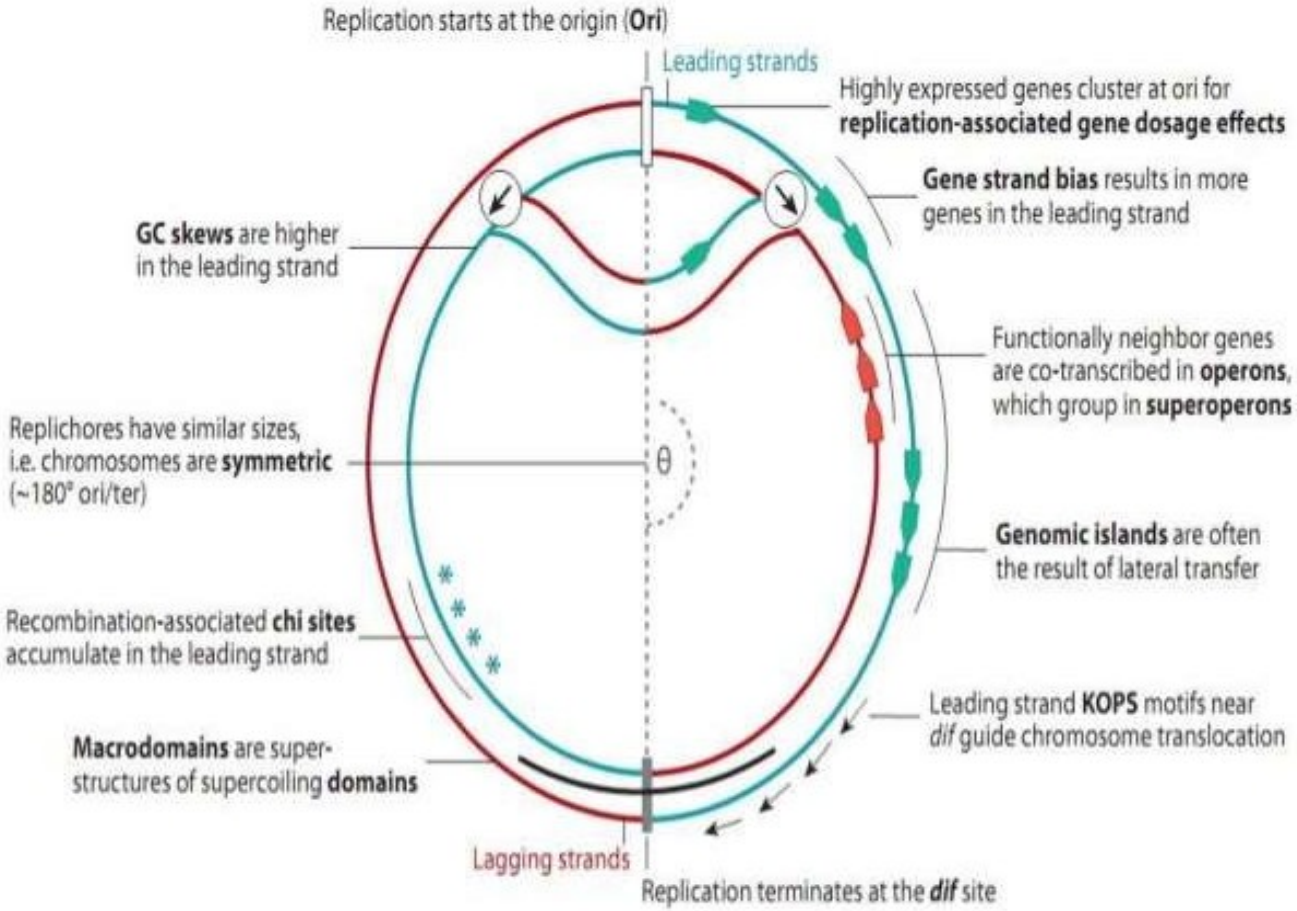
"Genome organization" can also **refer to the 3D structure of chromosomes** and the positioning of DNA sequences within the nucleus.


B
Tertiary structure
30nm chromatin
Secondary structure
Nucleus with distinct territories
Nucleosome
DNA
10nm chromatin
Primary Structure

Prokaryotic Genome

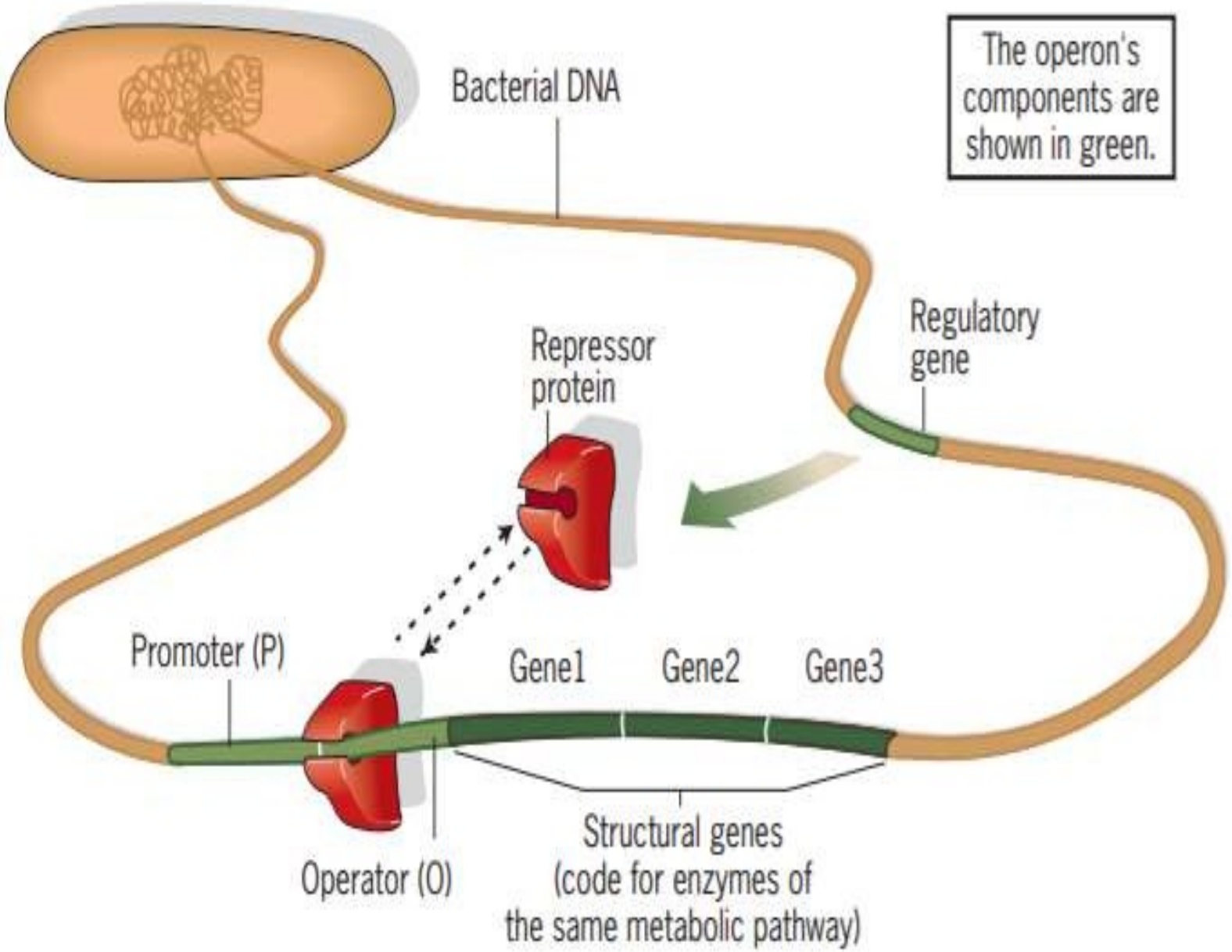
- Pro(before)-karyon (kernel) Include bacteria and Archaea
- The **nucleoid** (nucleus-like) - irregularly shaped region within the prokaryotic cell that contains all or most of the genetic material
- Relatively small genome with sizes ranging from 0.5 to 10Mbp (1 Mega base pairs=106 bp).
- Circular, double-stranded piece of DNA, multiple copies of which may exist at any time.
- The **gene density** in the genomes is **high**, with more than 90% of a genome sequence containing coding sequence and very few repetitive sequences.


Bacterial Cell Structure
Plasma Membrane
Cell wall
Capsule
Ribosomes
Nucleoid (DNA)
Flagellum
Plasmid
Pilus
Cytoplasm


Replication starts at the origin (Ori)
Leading strands
Highly expressed genes cluster at ori for replication-associated gene dosage effects
Gene strand bias results in more genes in the leading strand
GC skews are higher in the leading strand
Functionally neighbor genes are co-transcribed in operons, which group in superoperons
Replichores have similar sizes, i.e. chromosomes are symmetric (~180° ori/ter)
θ
Genomic islands are often the result of lateral transfer
Recombination-associated chi sites accumulate in the leading strand
Leading strand KOPS motifs near dif guide chromosome translocation
Macrodomains are super-structures of supercoiling domains
Lagging strands
Replication terminates at the dif site

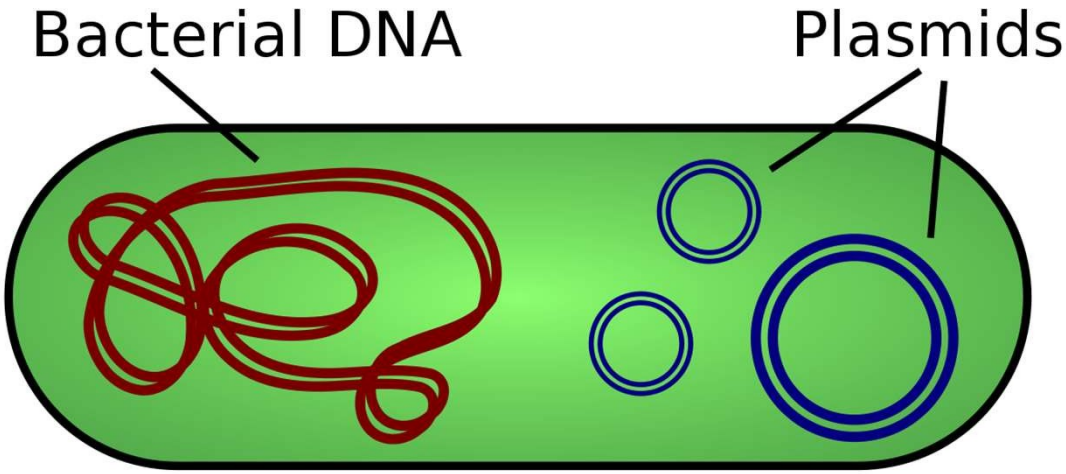
Operon

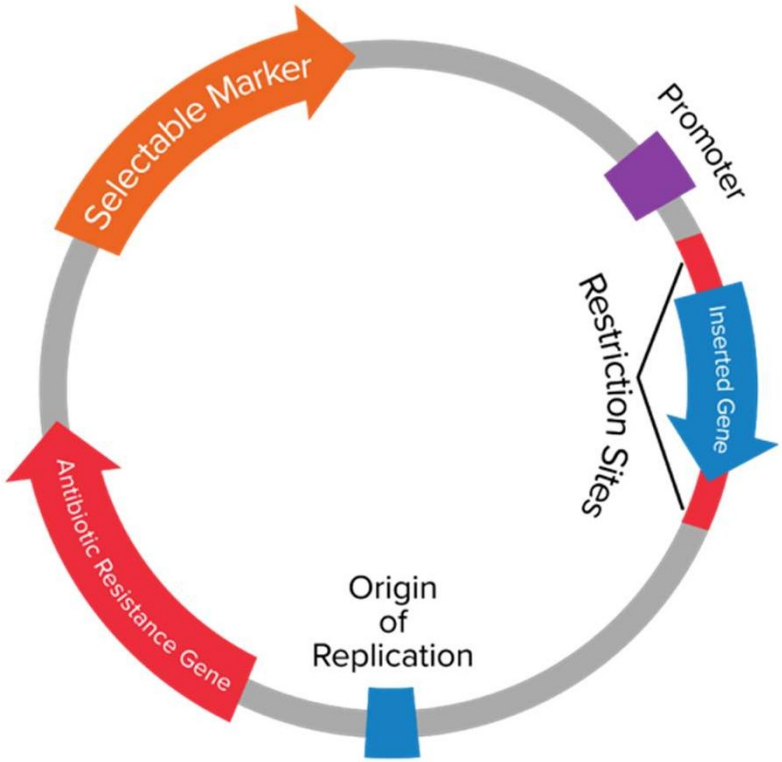
- Gene expression is an expensive process - takes a lot of energy to produce RNA and protein.
- If all the genes are turned on all the time, production of RNAs and proteins would drain the cell of energy control of gene expression is essential
- In Bacteria regulation is done by grouping functionally related genes together which can be regulated together easily.
- Group of contiguous, coordinately controlled genes is called an operon.
- First operon discovered - Lac operon - operon that encodes enzymes that permit a cell to metabolize the milk sugar lactose.


The operon's components are shown in green.
Bacterial DNA
Regulatory gene
Repressor protein
Promoter (P)
Gene1
Gene2
Gene3
Operator (O)
Structural genes (code for enzymes of the same metabolic pathway)

Plasmid

- **Self-replicating, double-stranded, circular** DNA molecules that are maintained in bacteria as independent **extrachromosomal** entities.
- Commonly used for vectors in genetic engineering
- Each plasmid has an origin of DNA replication - without this site, it cannot replicate in a host cell
- Copy number: average or expected number of copies per host cell
  - Low copy number (15–20)
  - Medium copy number(20–100 copies)
  - High Copy number(more than 100)

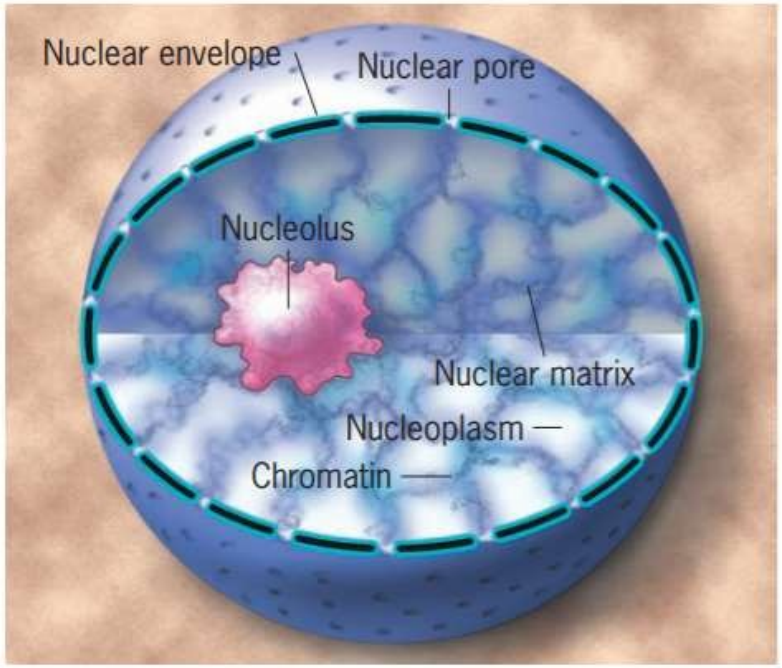


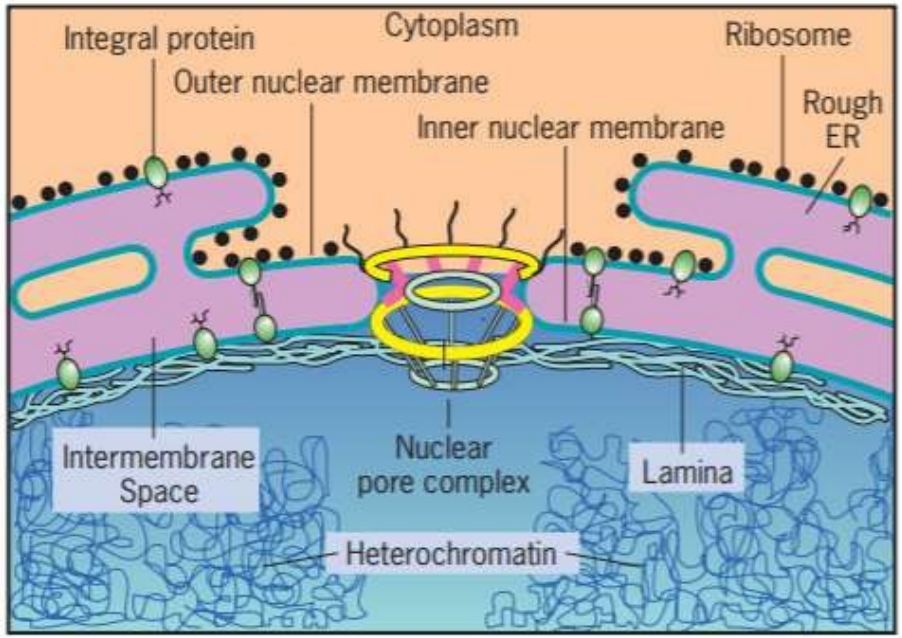


- Carry information for their own transfer from one cell to another (e.g., F plasmids),
- Encode resistance to antibiotics (R plasmids)
- carry specific sets of genes for the utilization of unusual metabolites (degradative plasmids),
- Some have no apparent functional coding genes (cryptic plasmids).

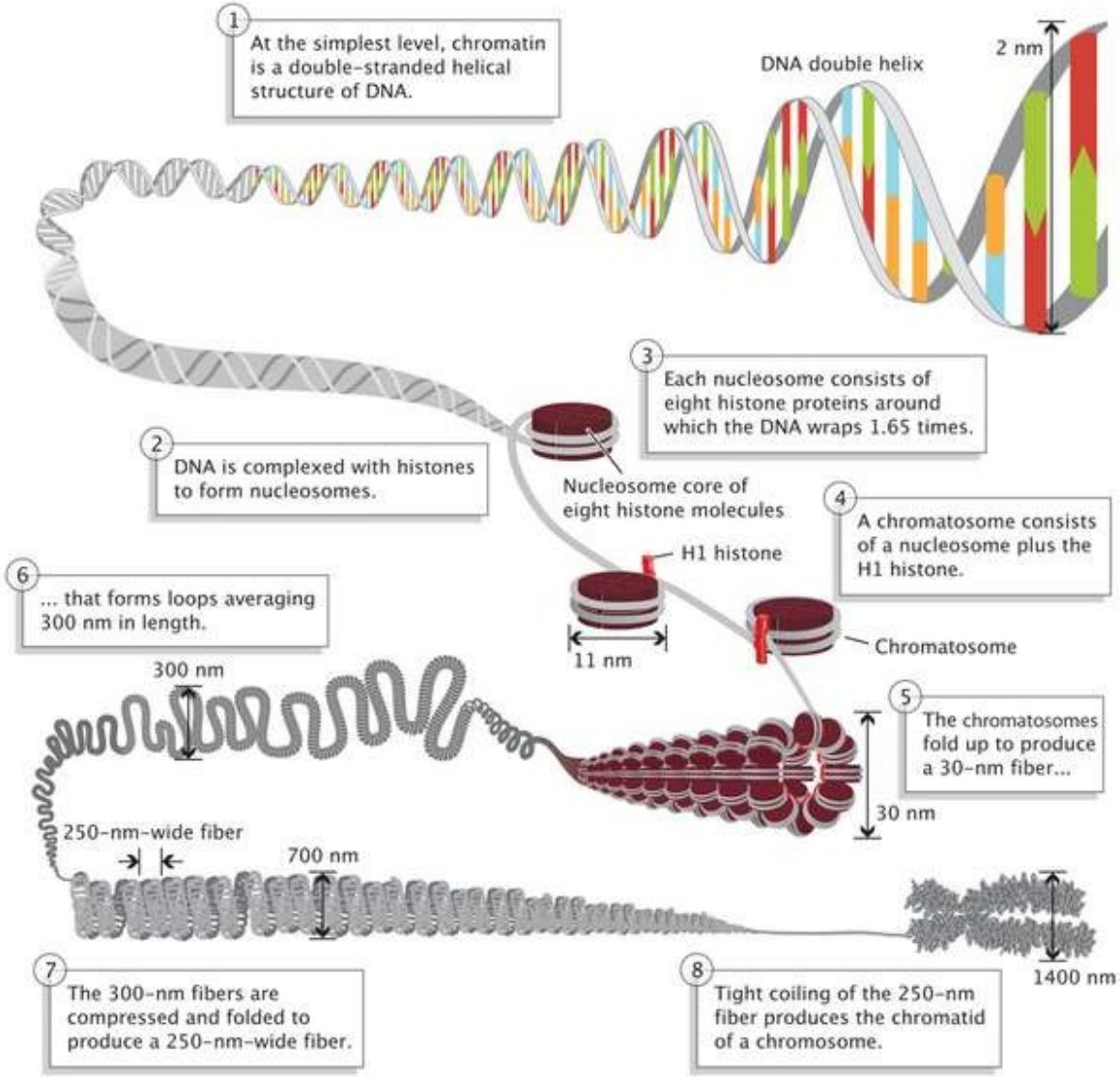
Eukaryotic Genome

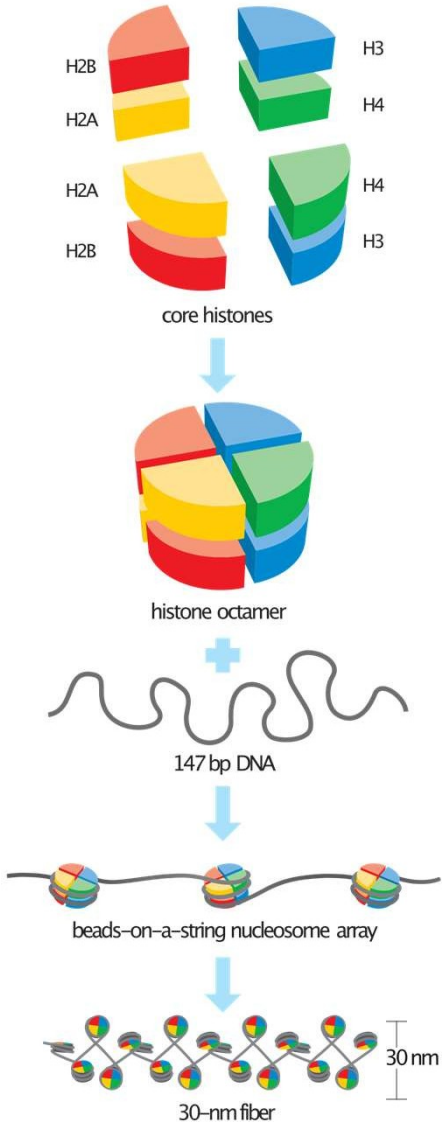
- Much larger than prokaryotic ones, with sizes ranging from 10 Mbp to 670 Gbp (1 Gbp = 109 bp)
- Linear DNA chromosomes nuclear genomes
- Very low gene density – In Humans, only 2-3% of the genome codes for genes, with about 1 gene per 100 kbp on average.
- Eukaryotic genomes are characterized by a mosaic organization of coding (Exons) and non coding regions (include introns, sequences for non-coding RNAs, regulatory regions, and repetitive DNA)

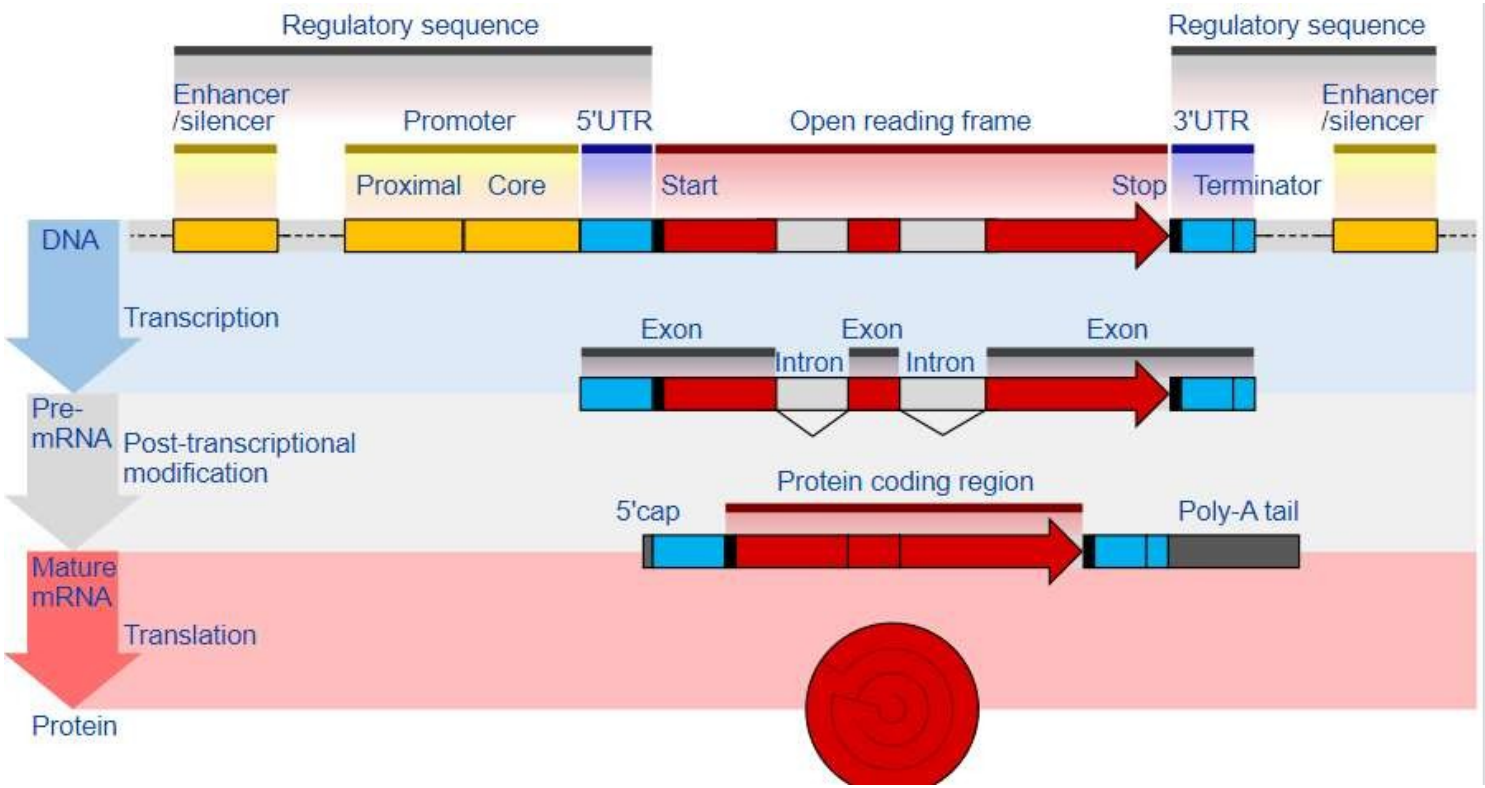




Eukaryotic Nucleus


1
At the simplest level, chromatin is a double-stranded helical structure of DNA.
DNA double helix
2 nm
2
DNA is complexed with histones to form nucleosomes.
3
Each nucleosome consists of eight histone proteins around which the DNA wraps 1.65 times.
Nucleosome core of eight histone molecules
4
A chromatosome consists of a nucleosome plus the H1 histone.
H1 histone
11 nm
Chromatosome
6
... that forms loops averaging 300 nm in length.
300 nm
5
The chromatosomes fold up to produce a 30-nm fiber...
30 nm
250-nm-wide fiber
700 nm
1400 nm
7
The 300-nm fibers are compressed and folded to produce a 250-nm-wide fiber.
8
Tight coiling of the 250-nm fiber produces the chromatid of a chromosome.


H2B
H2A
H3
H4
H2A
H2B
H4
H3
core histones
histone octamer
147 bp DNA
beads-on-a-string nucleosome array
30 nm
30-nm fiber



Eukaryotic Genome

Genomics of Microbes Genome

Microbial genomics, fuelling a rise in the number of available microbial genomes from microbial evolution to microbial diversity, host–pathogen interactions to disease-causing genetic variation, genomics has provided transformative insights into microbiology.

Moreover, genomic technologies show great potential for clinical diagnostics or the real-time detection and surveillance of epidemics.

Microbial Genomics: https://www.youtube.com/watch?v=fPNFk_BHi3Q

What sequencing technology can be used to detect structural variants?

Especially, long-read **sequencing** is powerful to **detect structural variants** and repetitive sequences.

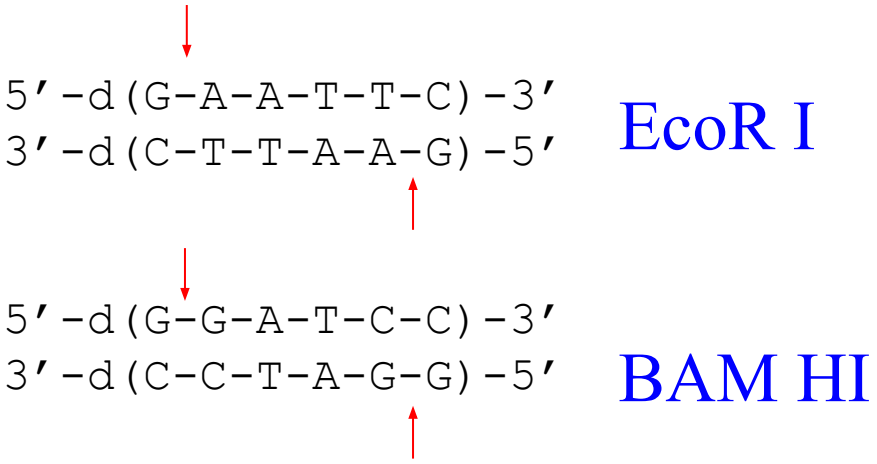
Strand-seq is the most suitable **detection** method for chromosomal inversions, a particularly challenging group of **structural variants**.

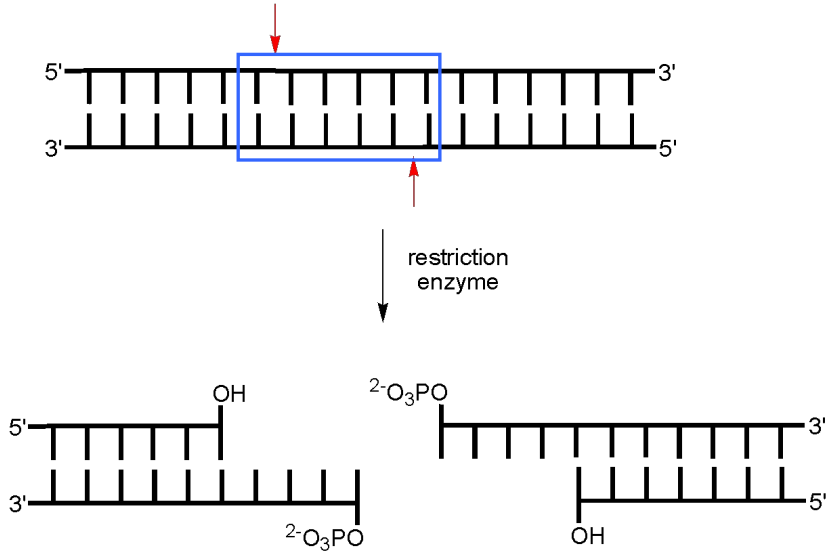
28.14: DNA Sequencing.

Maxam-Gilbert: relys on reagents that react with a specific DNA base that can subsequent give rise to a sequence specific cleavage of DNA

Sanger: Enzymatic replication of the DNA fragment to be sequenced with a DNA polymerase, Mg^{+2}, and dideoxynucleotides triphosphate (ddNTP) that truncates DNA replication

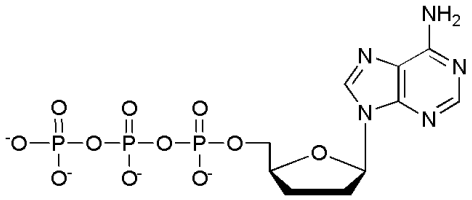
Restriction endonucleases: Bacterial enzymes that cleave DNA at specific sequences

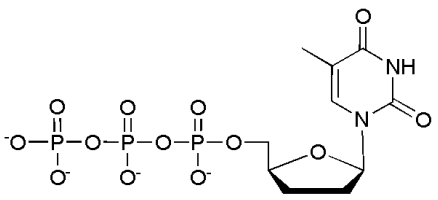


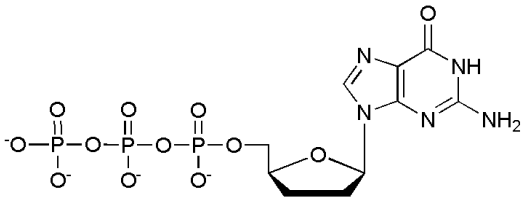


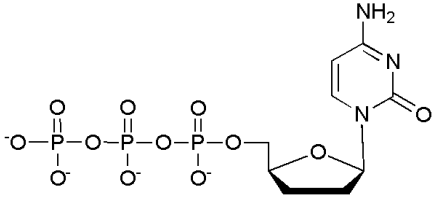
Sanger Sequencing:

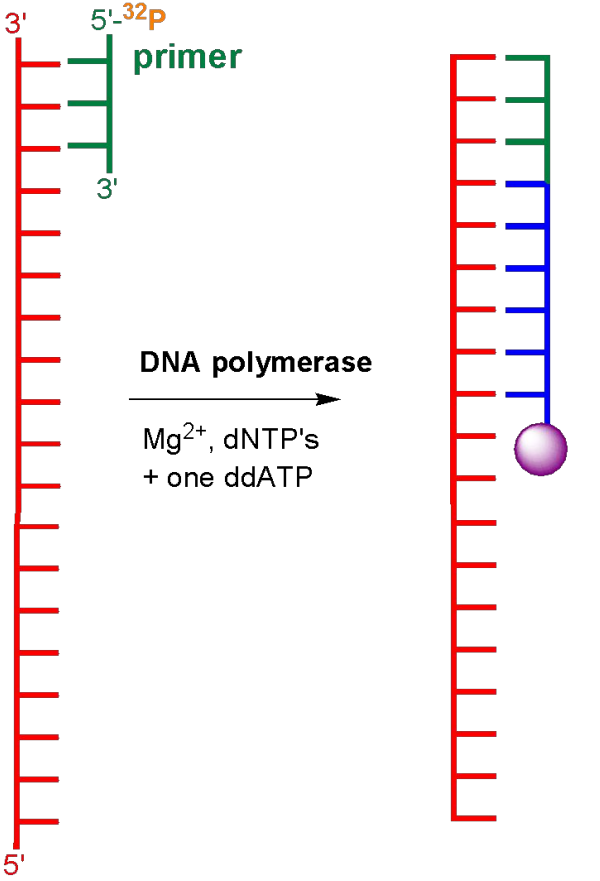
key reagent: dideoxynucleotides triphosphates (ddNTP)







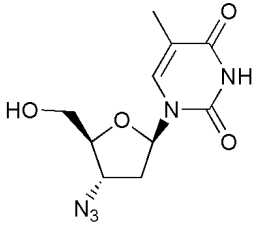


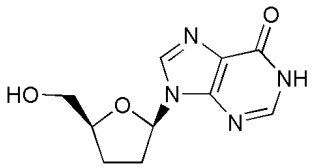


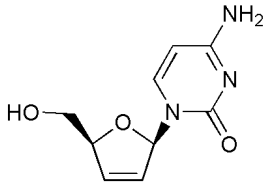
When a ddNTP is incorporated elongation of the primer is terminated

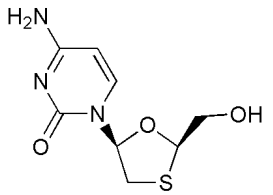
The ddNTP is specifically incorporated opposite its complementary nucleotide base

Anti-Viral Nucleosides

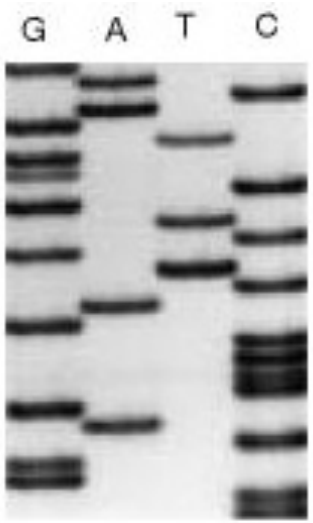


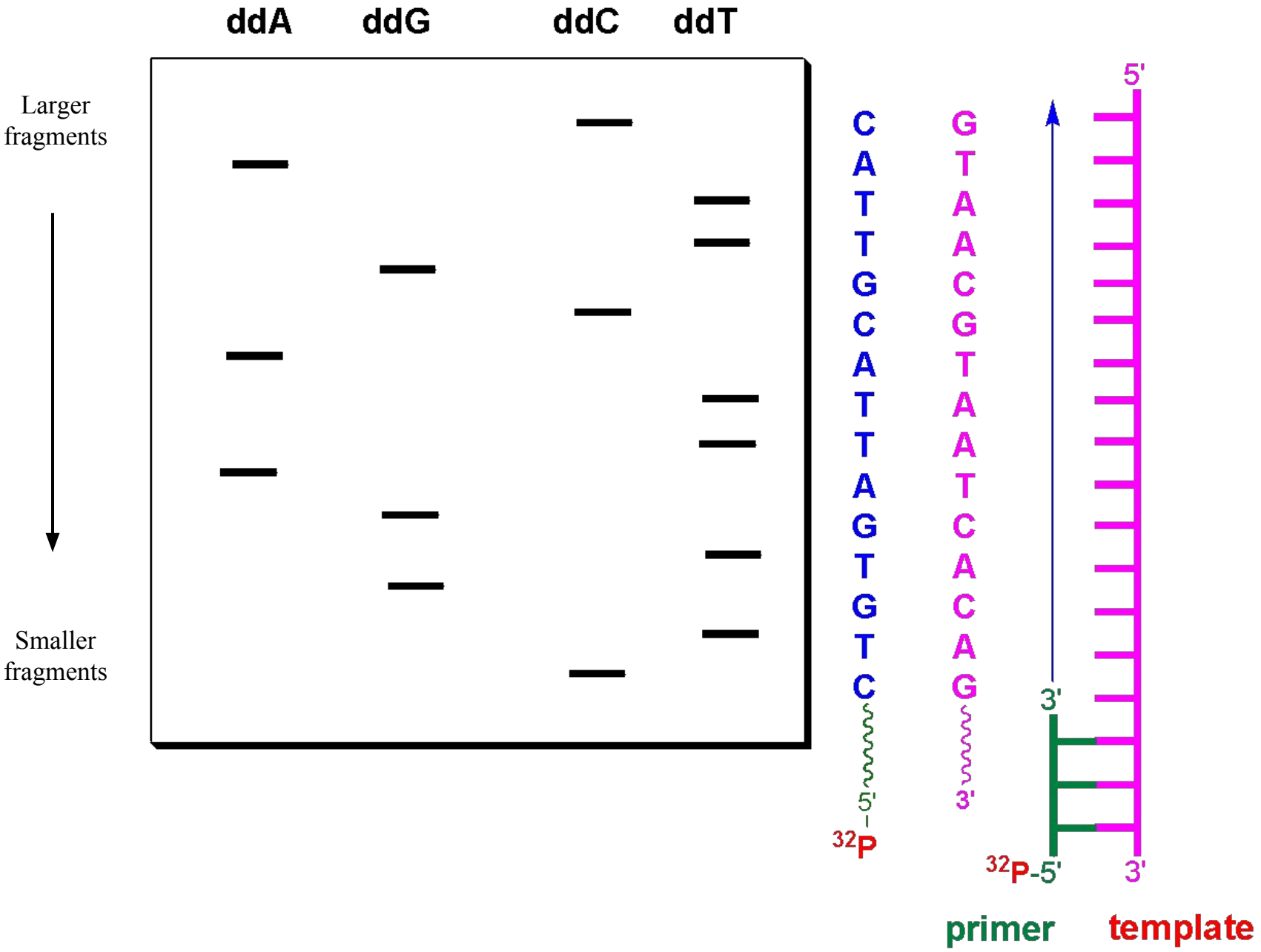






Sanger Sequencing





Bioinformatics Sequencing Method

Structural Variation Detection from

Next Generation Sequencing

(Ref: Structural Variation Detection from Next Generation Sequencing, Kai Ye 1*, George Hall2,3 and Zemin Ning2, Journal of Next Generation Sequencing & Applications, DOI: 10.4172/2469-9853.S1-007)

SEQUENCING DEVELOPMENT- NGS

Video:

https://sapac.illumina.com/science/technology/next-generation-sequencing.html
https://www.abmgood.com/marketing/knowledge_base/next_generation_sequencing_introduction.php

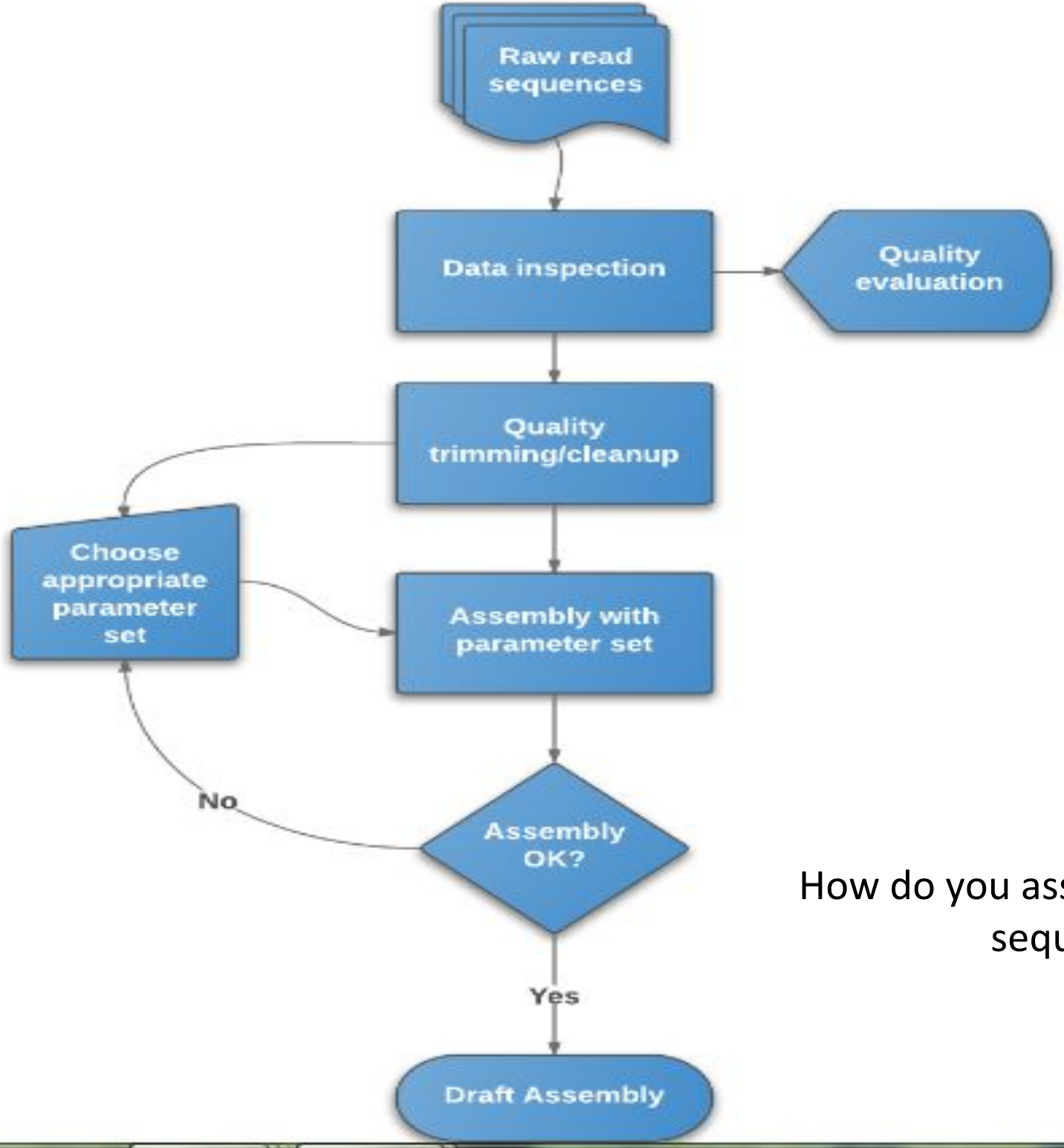
DNA, contains the blueprints of life.
Within its structures are the **codes required for the assembly of proteins and non-coding RNA –** these molecular machineries affect all the biological systems that create and maintain life**.**
By understanding the **sequence of DNA**, researchers have been able to **elucidate the structure and function of proteins as well as RNA** and have gained an understanding of the **underlying causes of disease.**
Next Generation Sequencing (NGS) is a powerful platform that has enabled the sequencing of **thousands to millions of DNA molecules simultaneously**.
This powerful tool is revolutionizing fields such as **personalized medicine, genetic diseases, and clinical diagnostics** by offering a high throughput option with the capability to **sequence multiple individuals at the same time.**

Assembly of data from genome sequencing

Genome assembly refers to the process of taking a large number of short DNA **sequences** and putting them back together to create a representation of the original chromosomes from which the DNA originated.

OR

In bioinformatics, **sequence assembly** refers to aligning and merging fragments from a longer DNA **sequence** in order to reconstruct the original **sequence**. ... Typically the short fragments, called reads, result from shotgun **sequencing** genomic DNA, or gene transcript (ESTs). De novo **genome assemblies** assume no prior knowledge of the source DNA **sequence** length, layout or composition.



How do you assemble a genome sequence?

How do you assemble a genome sequence? STEPS

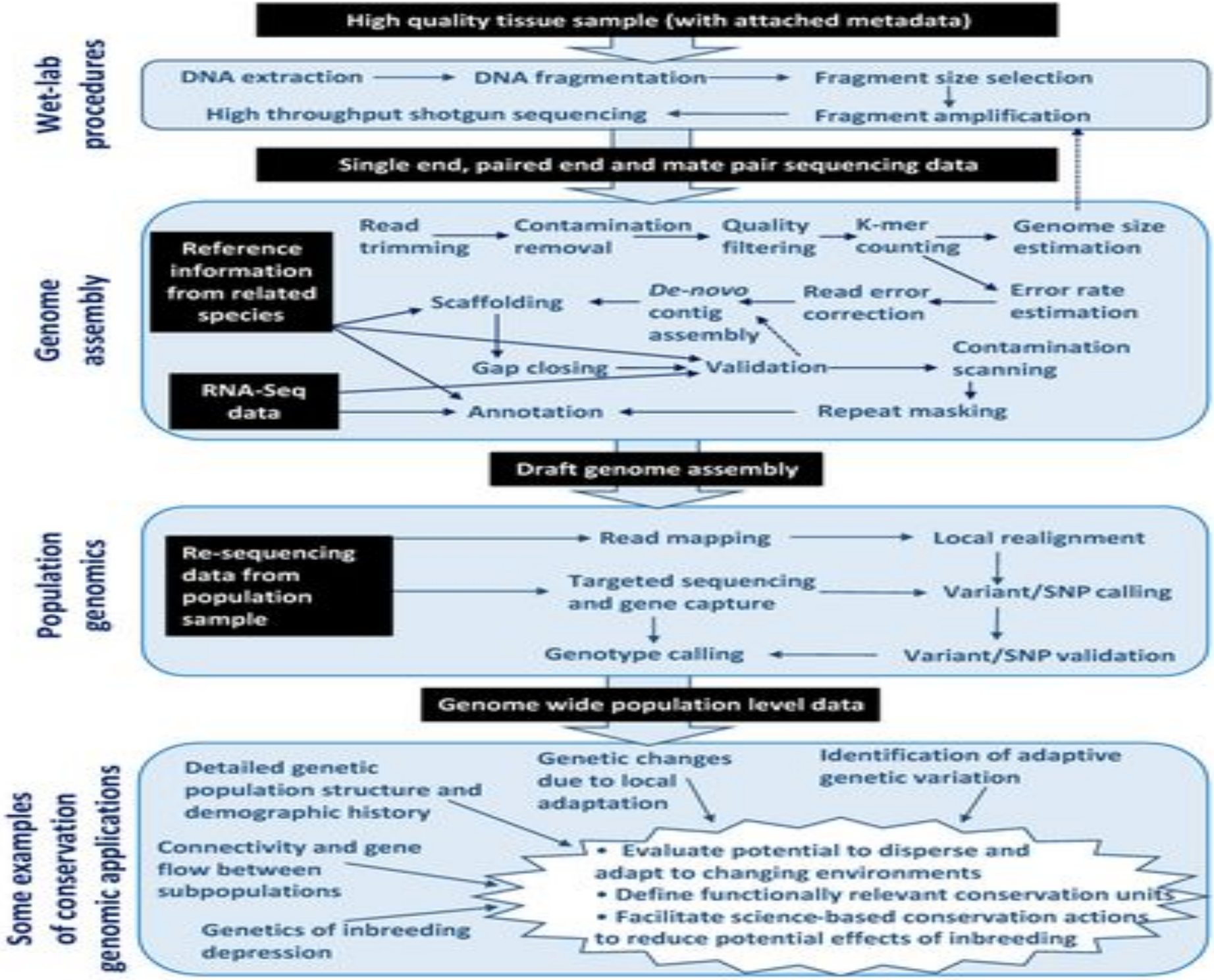
Step 1: Build a wide community for the project if possible
Step 2: Gather information about the target genome
Step 3: Design the best experimental workflow
Step 4: Choose the best sequencing platforms and library preparations
Step 5: Select the best possible DNA source and DNA extraction method
Step 6: Check the computational resources and requirements

Step 7: Choose the best computational design and pipeline
Step 8: Assemble the genome
Step 9: Check the assembly quality before annotation
Step 10: Genome annotation
Step 11: Build a searchable and sharable output format
Step 12: Reach out to the community to refine the assembly and annotation

Conclusions

There are no gold standards for genome assembly and annotation.
However, the availability of NGS data and their analytical tools has enabled the sequencing of several high-quality genomes of species of importance in aquaculture in recent years.
Beginners and small research groups still face challenges, because genome assembly and annotation are usually complex analytical procedures (or pipelines) requiring interdisciplinary collaborations (from biology to computer science) and hefty costs for refining/maintaining the genome.

● Availability of appropriate computational resources
● Collaboration with sequencing facility and bioinformatics groups
● Plan for amount and type of sequencing data needed
● Does funding allow to produce sufficient sequence coverage? If not, alternative approaches should be considered rather than producing a poor, low coverage, assembly
● Familiarization with data handling pipelines and file formats (see below)
● High-quality DNA sample (with individual metadata)
● Plan for analyses and publication


High quality tissue sample (with attached metadata)
Wet-lab procedures
DNA extraction
DNA fragmentation
Fragment size selection
Fragment amplification
High throughput shotgun sequencing
Single end, paired end and mate pair sequencing data
Genome assembly
Read trimming
Contamination removal
Quality filtering
K-mer counting
Genome size estimation
Reference information from related species
Scaffolding
De-novo contig assembly
Read error correction
Error rate estimation
Gap closing
Validation
Contamination scanning
RNA-Seq data
Annotation
Repeat masking
Draft genome assembly
Population genomics
Re-sequencing data from population sample
Read mapping
Local realignment
Targeted sequencing and gene capture
Variant/SNP calling
Genotype calling
Variant/SNP validation
Genome wide population level data
Some examples of conservation genomic applications
Detailed genetic population structure and demographic history
Genetic changes due to local adaptation
Identification of adaptive genetic variation
Connectivity and gene flow between subpopulations
Genetics of inbreeding depression
• Evaluate potential to disperse and adapt to changing environments
• Define functionally relevant conservation units
• Facilitate science-based conservation actions to reduce potential effects of inbreeding