

Introduction to Microbial Genomics and DNA sequencing: Basic Principles

GastroPak Microbial Bioinformatics
Workshop

DR CHRISTOPHER QUINCE

Earlham/Quadram Group Leader



Plan for the week

https://github.com/Sebastien-Raguideau/GastroPak_Workshop

Overview

- 1) Microbial world
- 2) Microbial cell structure
- 3) Microbial genomics
- 4) DNA sequencing and 'omics

GLOBAL
EDITION



BROCK BIOLOGY OF **MICROORGANISMS**

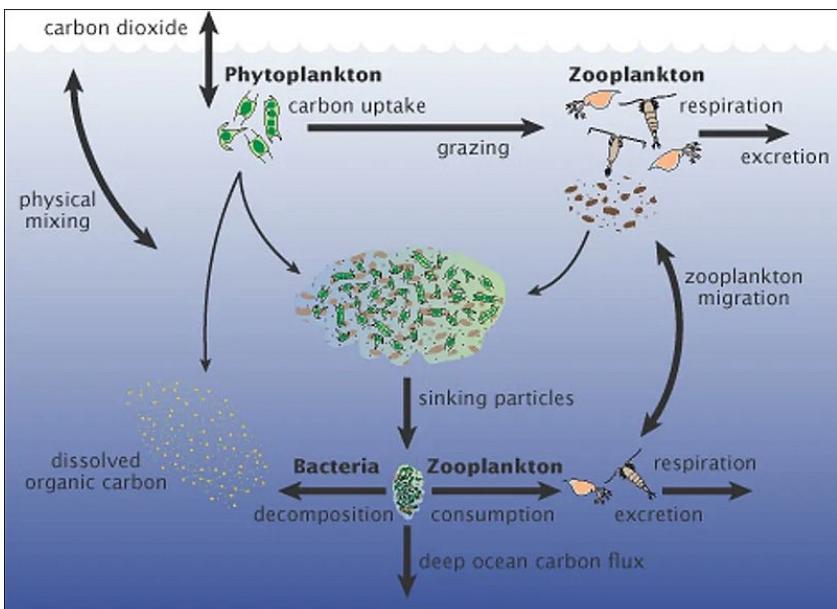
SIXTEENTH EDITION

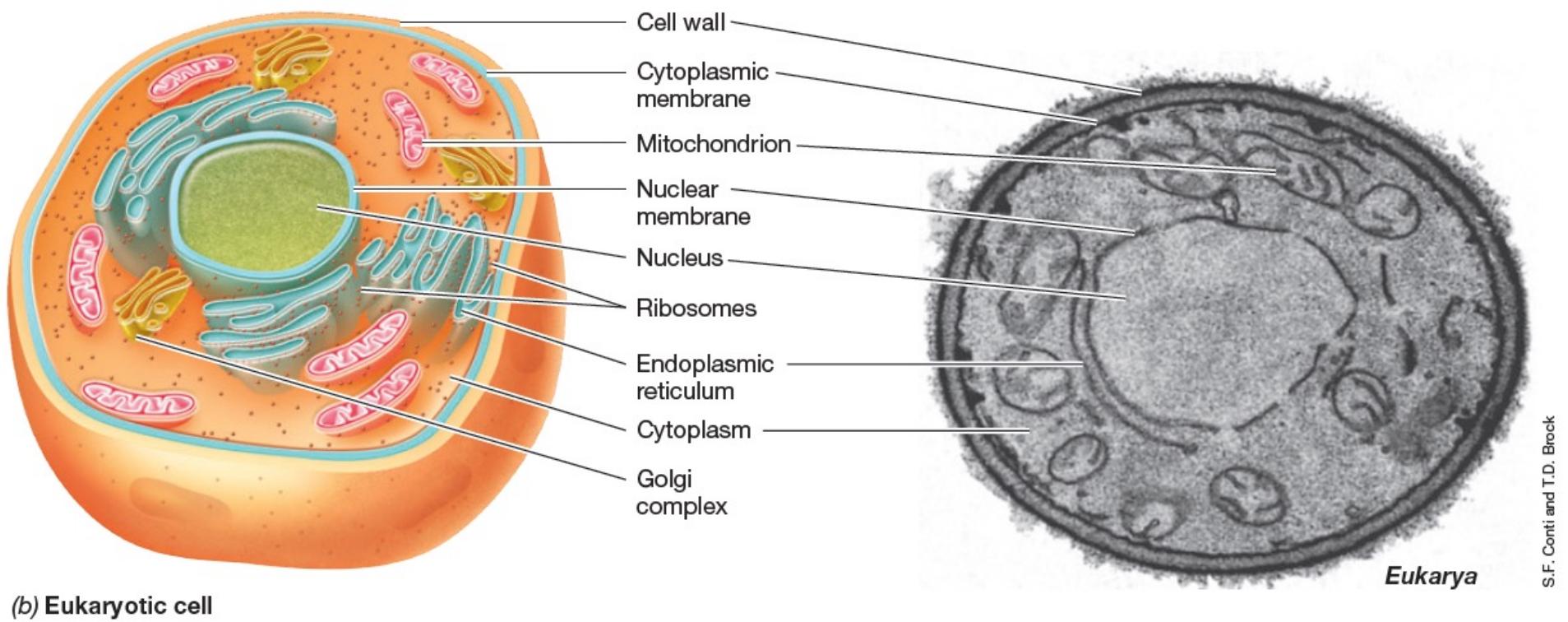
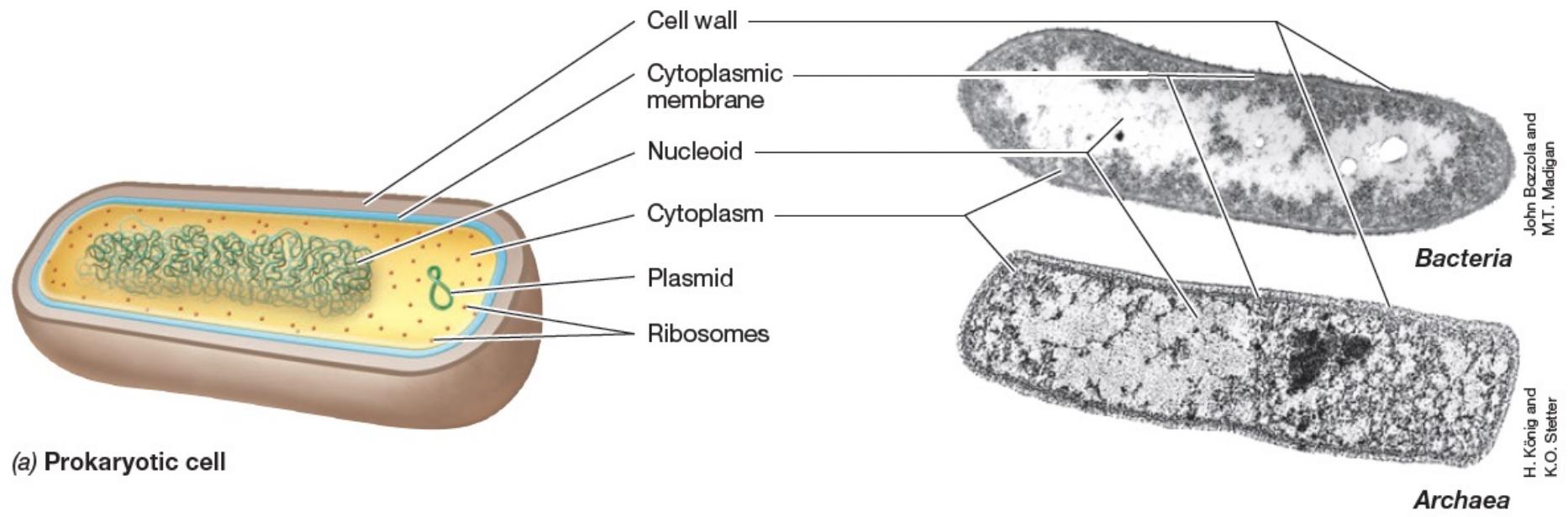
Madigan • Bender • Buckley • Sattley • Stahl



Why are they important?

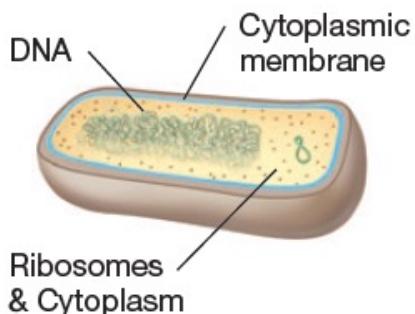
- What is a microbe:
 - Anything **microscopic**
 - Bacteria, archaea, eukaryote or **virus**
- Bacterial pathogens e.g. *Salmonella enterica*
- Commensal human microbiota in the gut plays a vital role in the immune system, gut-brain axis etc.
- Environmental microbiota key in major biogeochemical cycles e.g. soil, plankton





Structure

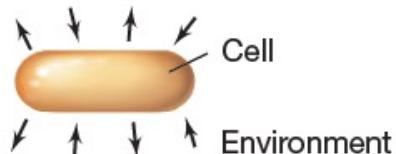
All cells have a cytoplasmic membrane, cytoplasm, a genome made of DNA, and ribosomes.



Metabolism

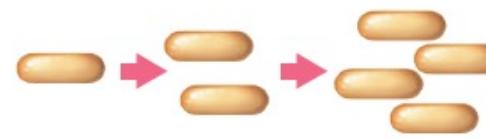
All cells use information encoded in DNA to make RNA and protein. All cells take up nutrients, transform them, conserve energy, and expel wastes.

1. Catabolism (transforming molecules to produce energy and building blocks)
2. Anabolism (synthesizing macromolecules)



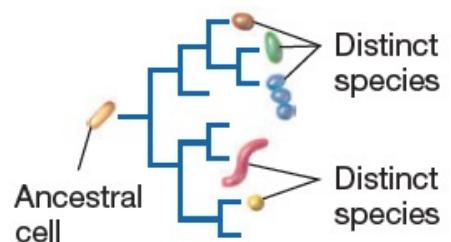
Growth

Information from DNA is converted into proteins, which do work. Proteins are used to convert nutrients from the environment into new cells.



Evolution

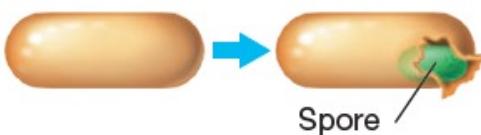
Chance mutations in DNA cause new cells to have new properties, thereby promoting evolution. Phylogenetic trees built from DNA sequences capture evolutionary relationships between species.



Properties of some cells:

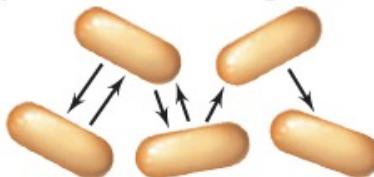
Differentiation

Some cells can form new cell structures such as a spore.



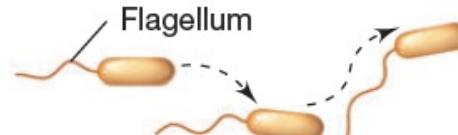
Communication

Cells interact with each other by chemical messengers.



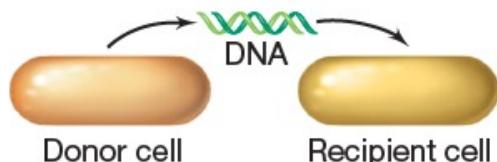
Motility

Some cells are capable of self-propulsion.



Horizontal gene transfer

Cells can exchange genes by several mechanisms.

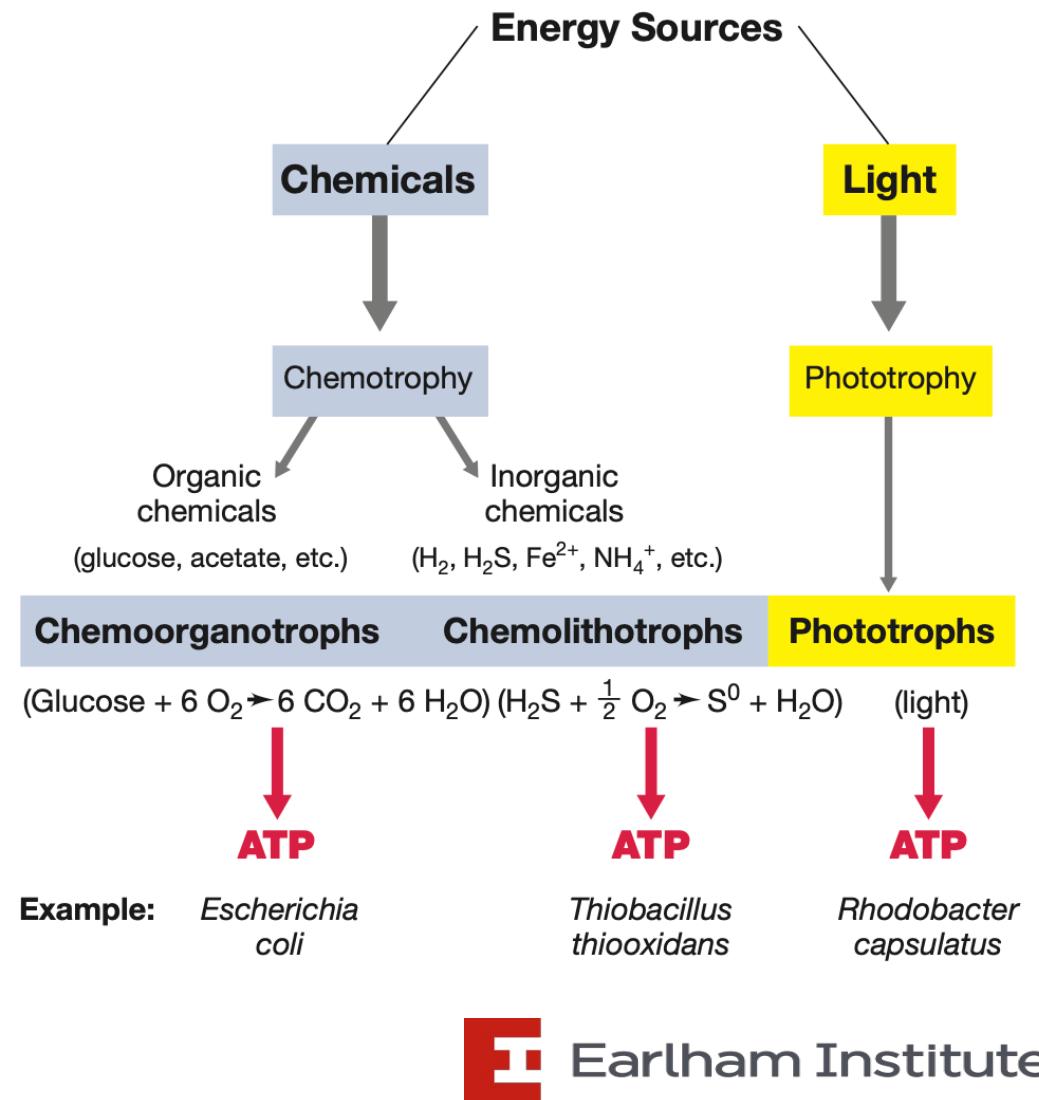


Prokaryotic microbes

- Can be aerobic or anaerobic
- Utilise a wide range of energy sources

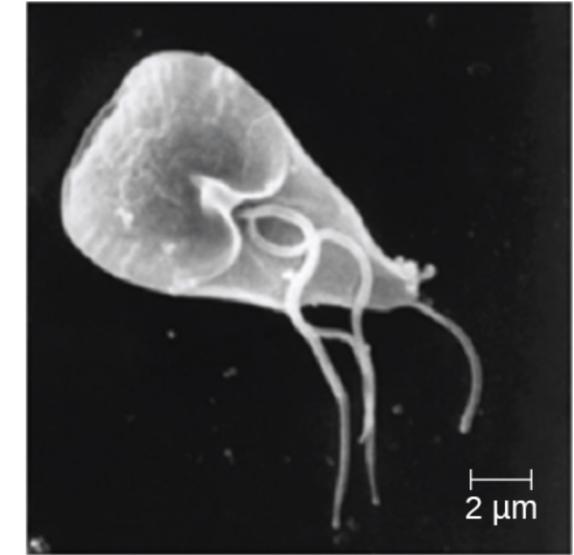
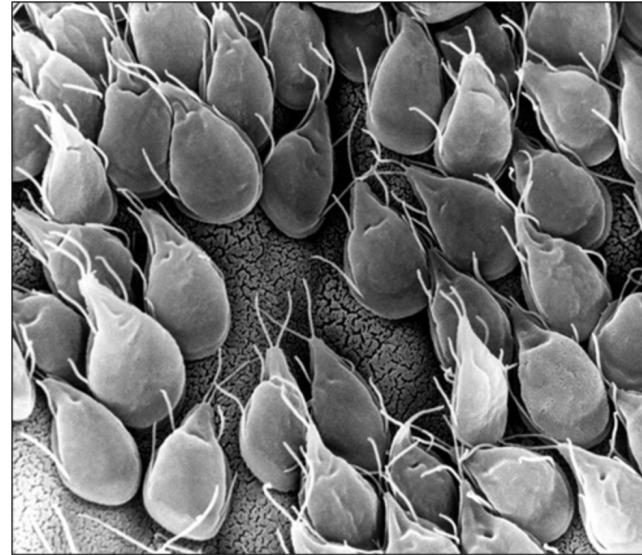


(b) Bloom of purple bacteria in a salt marsh

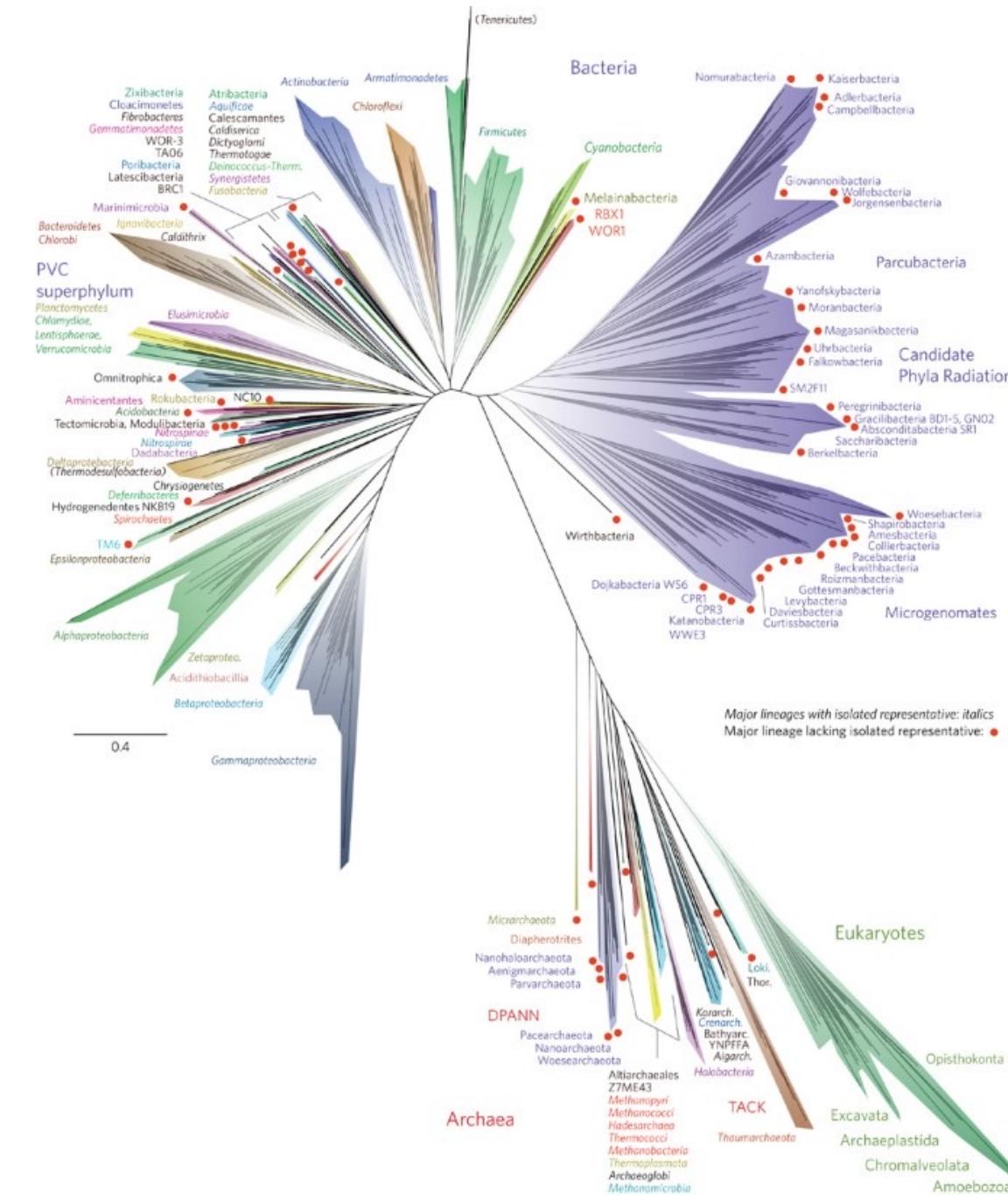


Eukaryotic microbes

- Includes important gastroenteritic pathogens e.g. *Giardia lamblia* and *Cryptosporidium sp.*
- Also yeasts
- Picoeukaryotic plankton

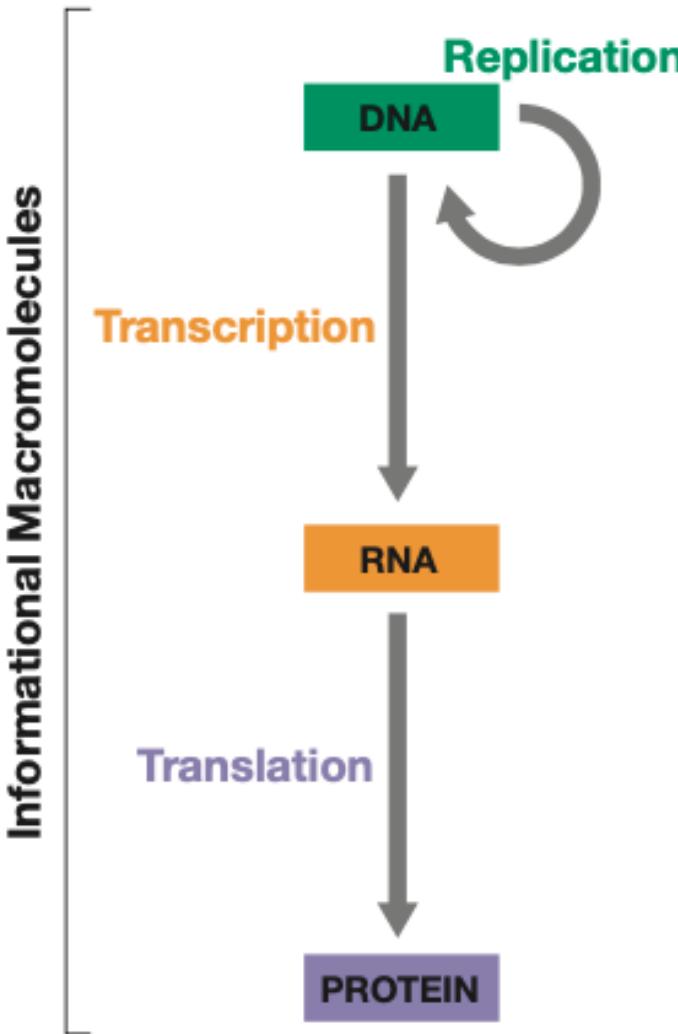


Tree of life



Hug et al. Nature Microbiology (2016)

Central dogma of molecular biology



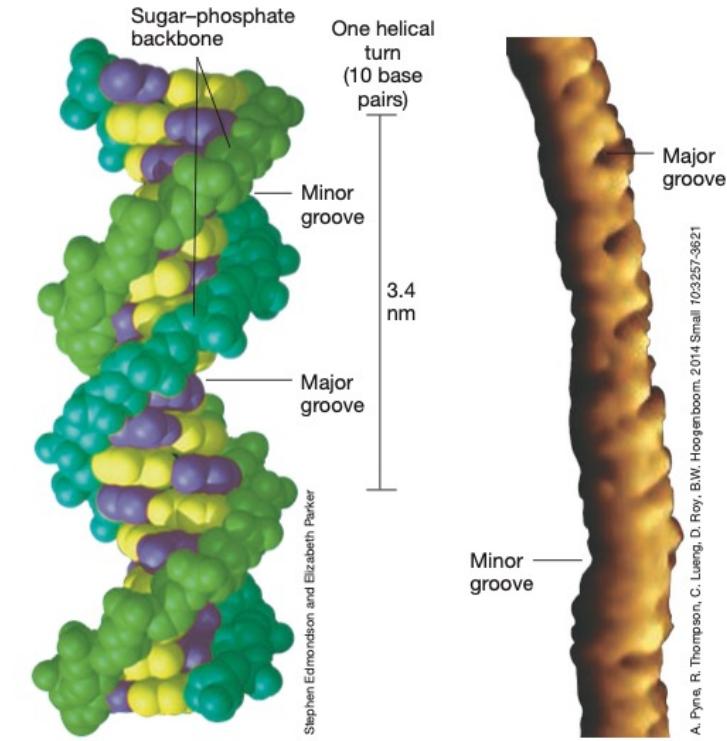
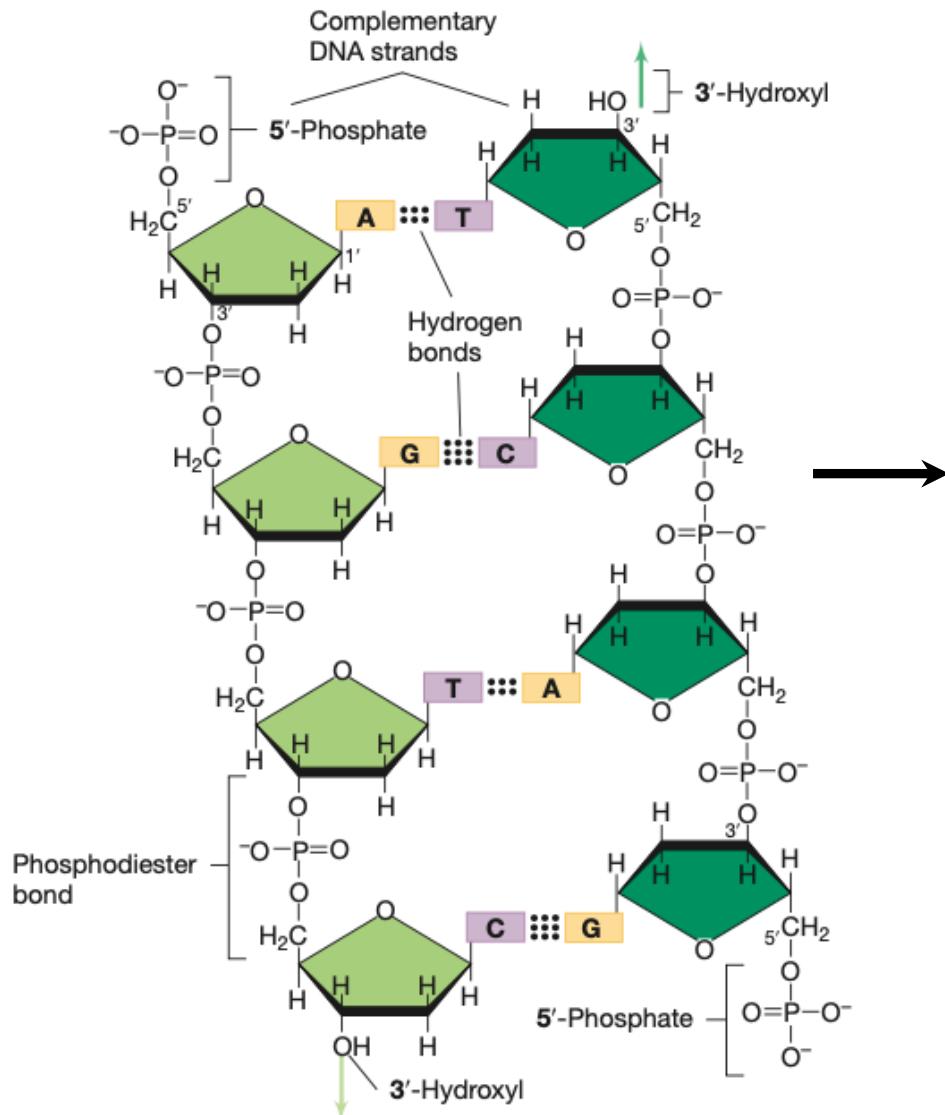
- Informational Macromolecules
- 1) **Replication.** During replication, the DNA double helix is duplicated. Replication is catalyzed by the enzyme *DNA polymerase*.
 - 2) **Transcription.** The transfer of genetic information from DNA to RNA is called transcription. Transcription is catalyzed by the enzyme *RNA polymerase*.
 - 3) **Translation.** The formation of a polypeptide using the genetic information transferred to mRNA by DNA is a process that occurs on the ribosome.



Earlham Institute

Properties of DNA and replication

- DNA comprises four bases with complementary binding:
 - Purine (A or G) to pyrimidine (C or T)
- Replication occurs during cell division by DNA polymerase
- Forms double helix structure



- Genes are transcribed to mRNA - complementary to 5' to 3' strand
- Limited post-transcription modification
- Ribosomes translate mRNA to proteins
- These are comprised of 22 naturally occurring amino acids
- Perform most enzymatic and some structural roles in cell

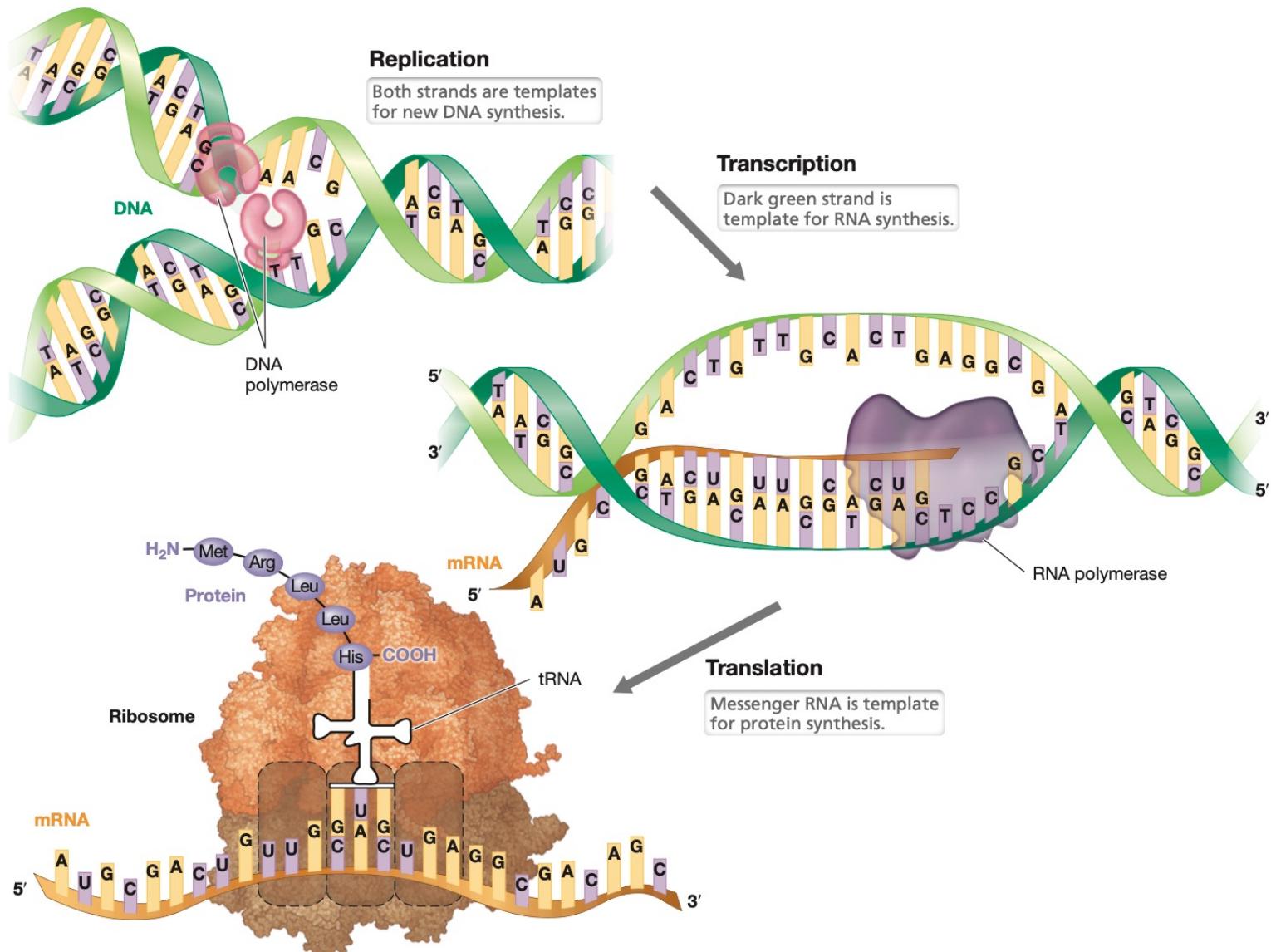


Figure 6.5 Synthesis of the three types of informational macromolecules in the processes of replication (DNA → DNA), transcription (DNA → RNA), and translation (RNA → protein). Note that for any particular gene only one of the two strands of the DNA double helix is transcribed.

Transcription in prokaryotes

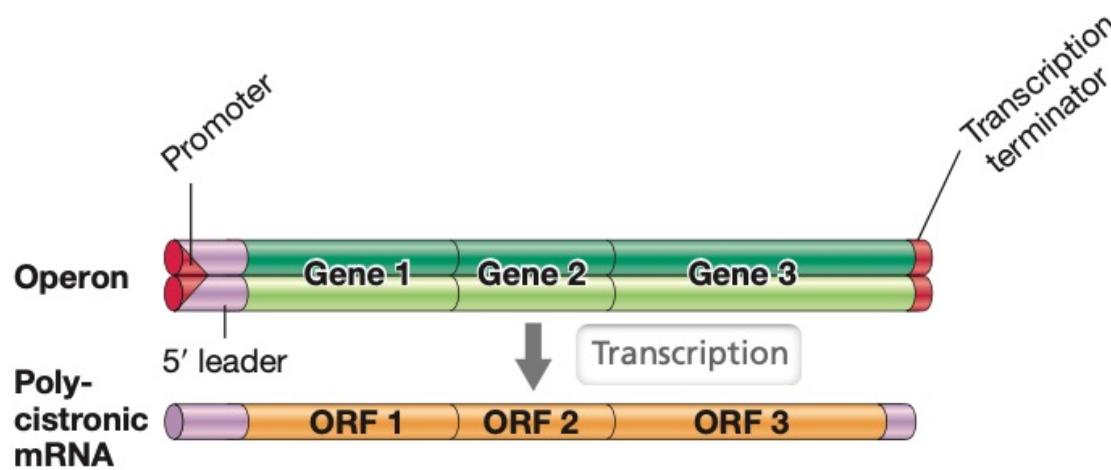
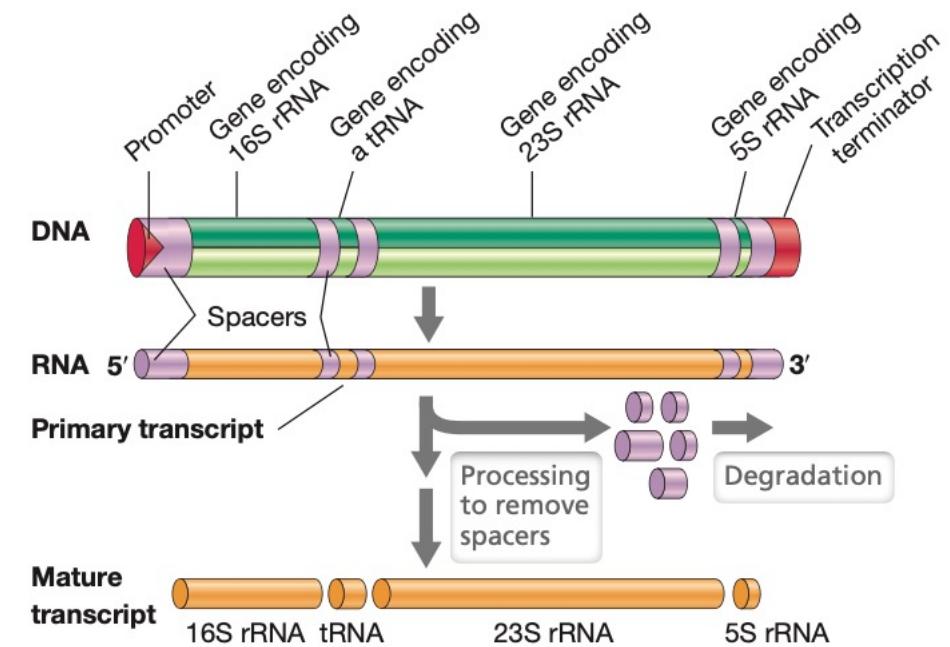


Figure 6.22 Operon and polycistronic mRNA structure. Note that a single promoter controls the three genes within the operon and that the polycistronic mRNA molecule contains an open reading frame (ORF) corresponding to each gene.



- A ribosomal rRNA transcription unit from *Bacteria* and its subsequent processing.
- In *Bacteria*, all rRNA transcription units have the genes in the order 16S rRNA, 23S rRNA, and 5S rRNA
- Spacer between the 16S and 23S rRNA genes contains a tRNA gene.
- *Escherichia coli* contains seven rRNA transcription units.

Translation and the genetic code

TABLE 6.4 The genetic code as expressed by triplet base sequences of mRNA

Codon	Amino acid	Codon	Amino acid	Codon	Amino acid	Codon	Amino acid
UUU	Phenylalanine	UCU	Serine	UAU	Tyrosine	UGU	Cysteine
UUC	Phenylalanine	UCC	Serine	UAC	Tyrosine	UGC	Cysteine
UUA	Leucine	UCA	Serine	UAA	None (stop signal)	UGA	None (stop signal)
UUG	Leucine	UCG	Serine	UAG	None (stop signal)	UGG	Tryptophan
CUU	Leucine	CCU	Proline	CAU	Histidine	CGU	Arginine
CUC	Leucine	CCC	Proline	CAC	Histidine	CGC	Arginine
CUA	Leucine	CCA	Proline	CAA	Glutamine	CGA	Arginine
CUG	Leucine	CCG	Proline	CAG	Glutamine	CGG	Arginine
AUU	Isoleucine	ACU	Threonine	AAU	Asparagine	AGU	Serine
AUC	Isoleucine	ACC	Threonine	AAC	Asparagine	AGC	Serine
AUA	Isoleucine	ACA	Threonine	AAA	Lysine	AGA	Arginine
AUG (start) ^a	Methionine	ACG	Threonine	AAG	Lysine	AGG	Arginine
GUU	Valine	GCU	Alanine	GAU	Aspartic acid	GGU	Glycine
GUC	Valine	GCC	Alanine	GAC	Aspartic acid	GGC	Glycine
GUA	Valine	GCA	Alanine	GAA	Glutamic acid	GGA	Glycine
GUG	Valine	GCG	Alanine	GAG	Glutamic acid	GGG	Glycine

^aAUG encodes N-formylmethionine at the beginning of polypeptide chains of *Bacteria*.

64 codons map to 22 amino acids
T → U in RNA

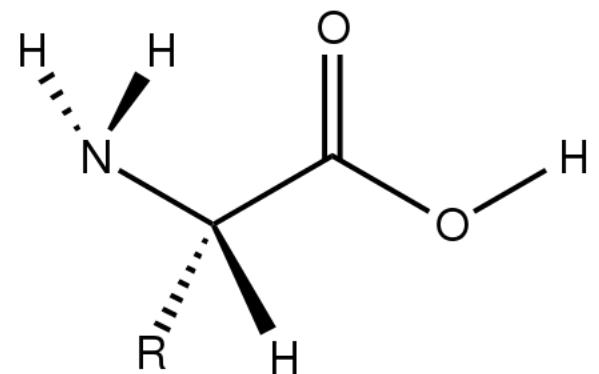
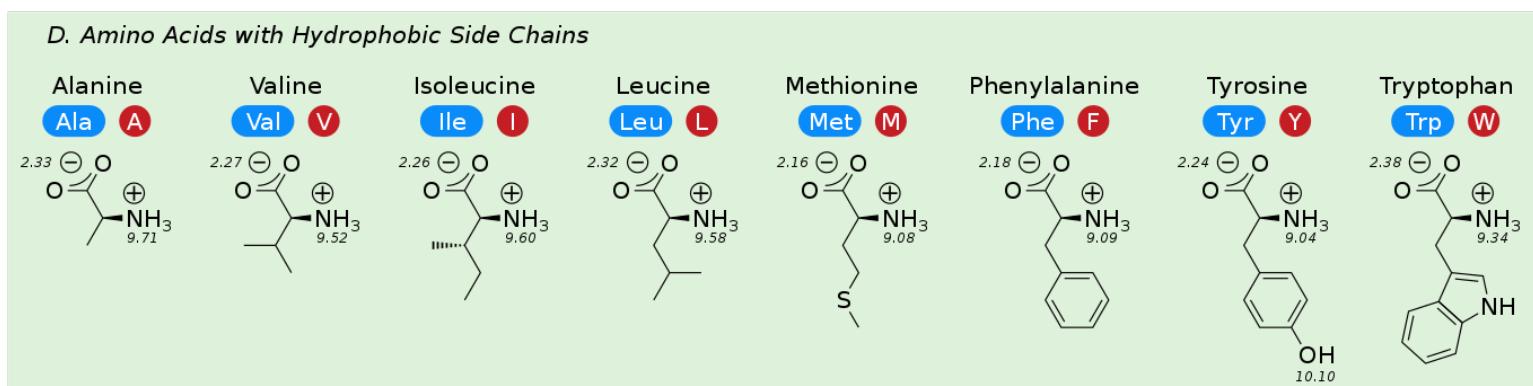
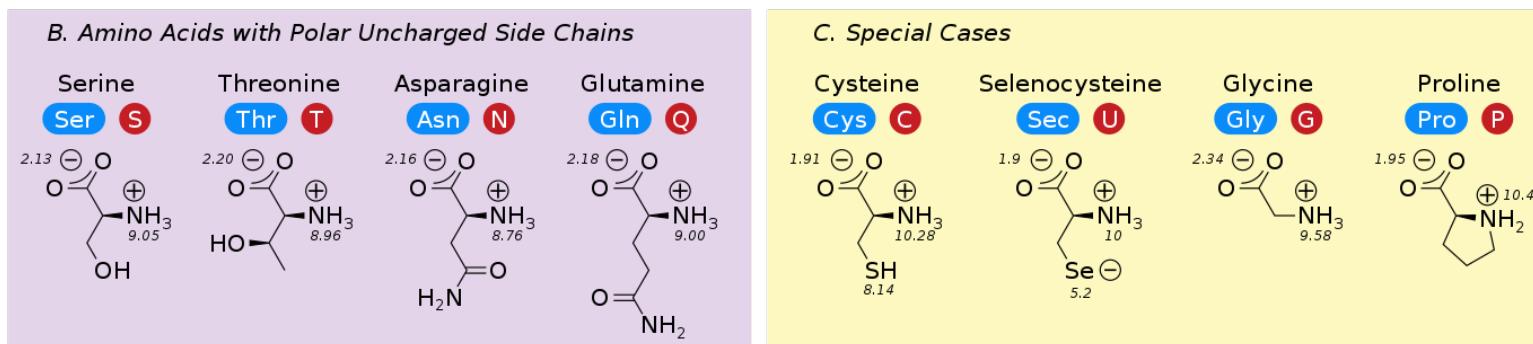
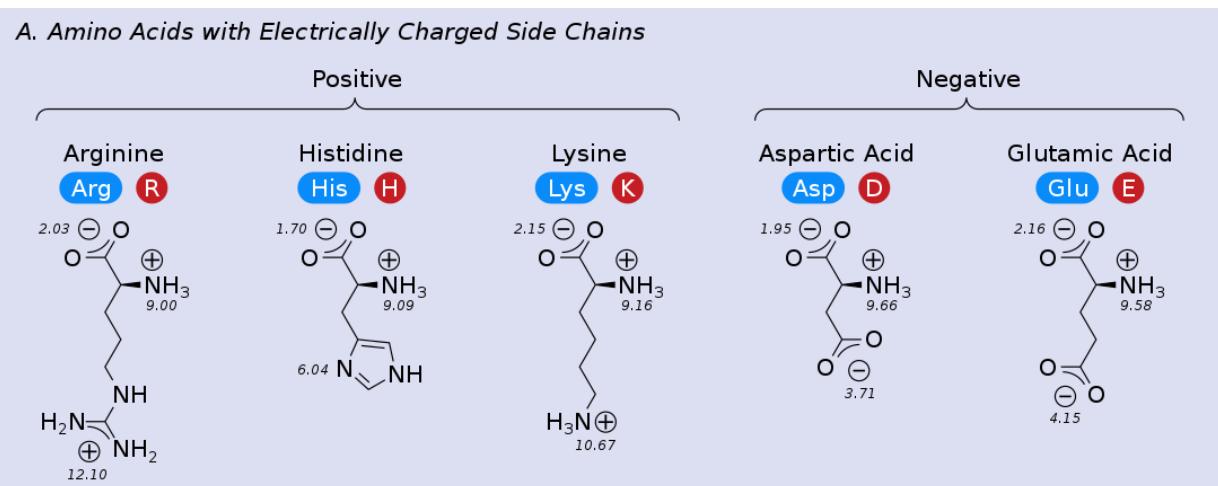
Properties of amino acids

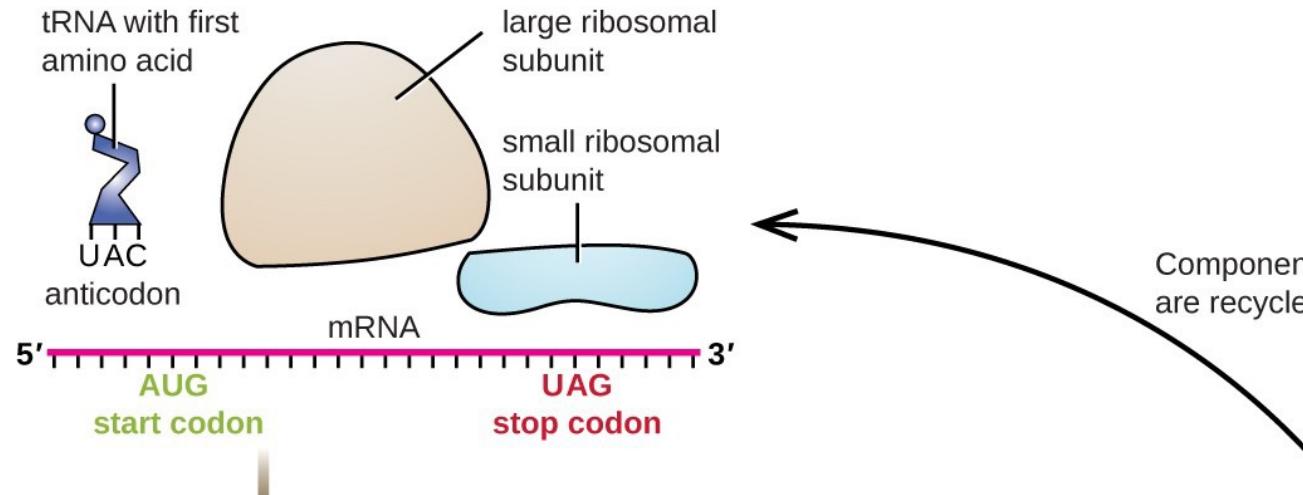
TWENTY-ONE PROTEINOGENIC α -AMINO ACIDS

Side chain charge at physiological pH 7.4

pK_a values shown italicized

⊕ Positive ⊖ Negative



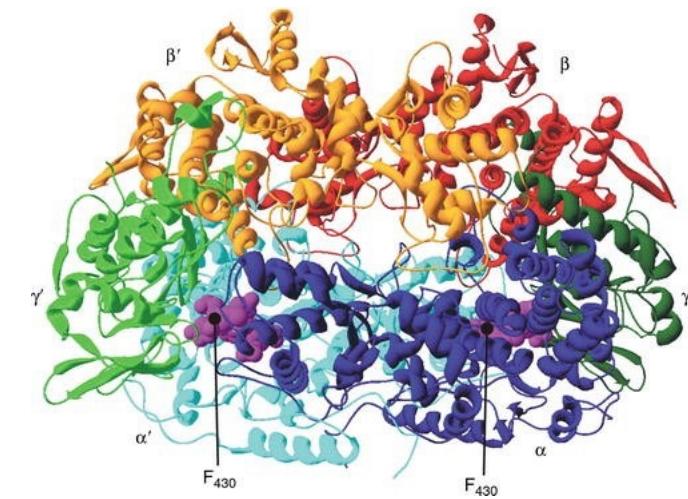
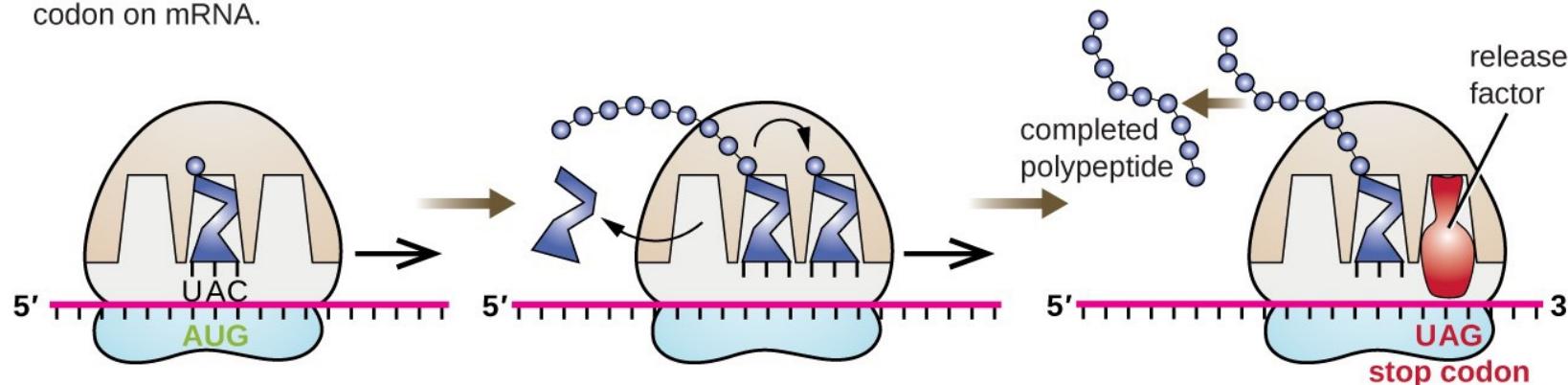


Components are recycled.

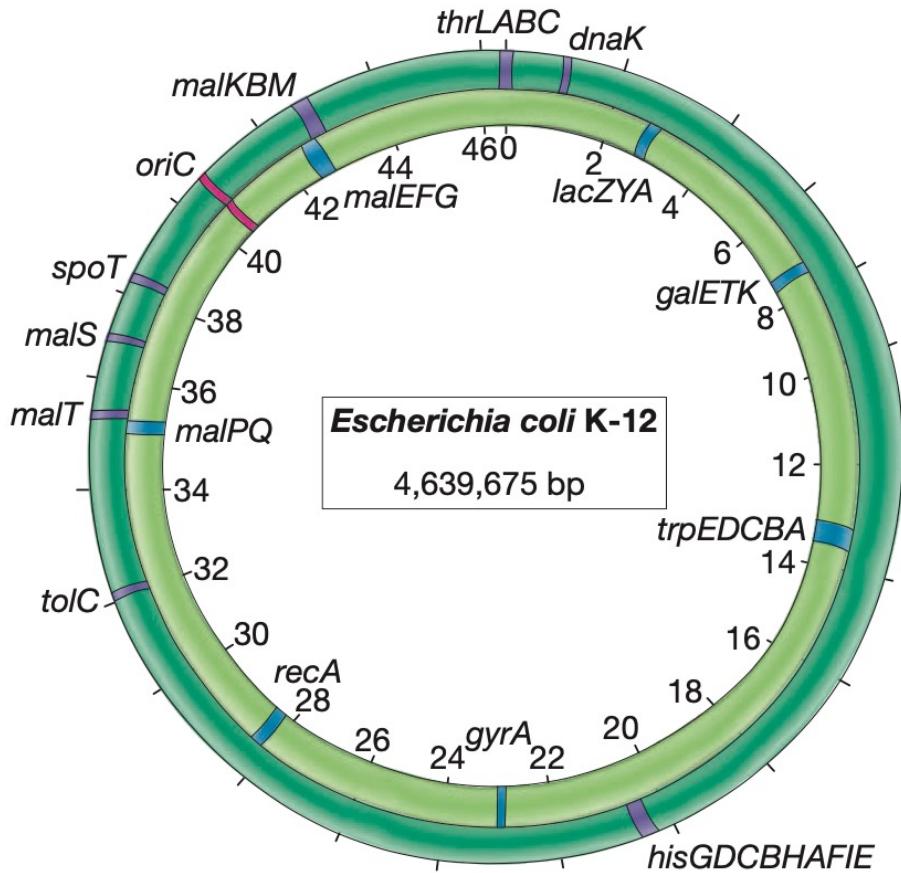
INITIATION
Transitional complex forms, and tRNA brings first amino acid in polypeptide chain to bind to start codon on mRNA.

ELONGATION
tRNAs bring amino acids one by one to add to polypeptide chain.

TERMINATION
Release factor recognizes stop codon, translational complex dissociates, and completed polypeptide is released.



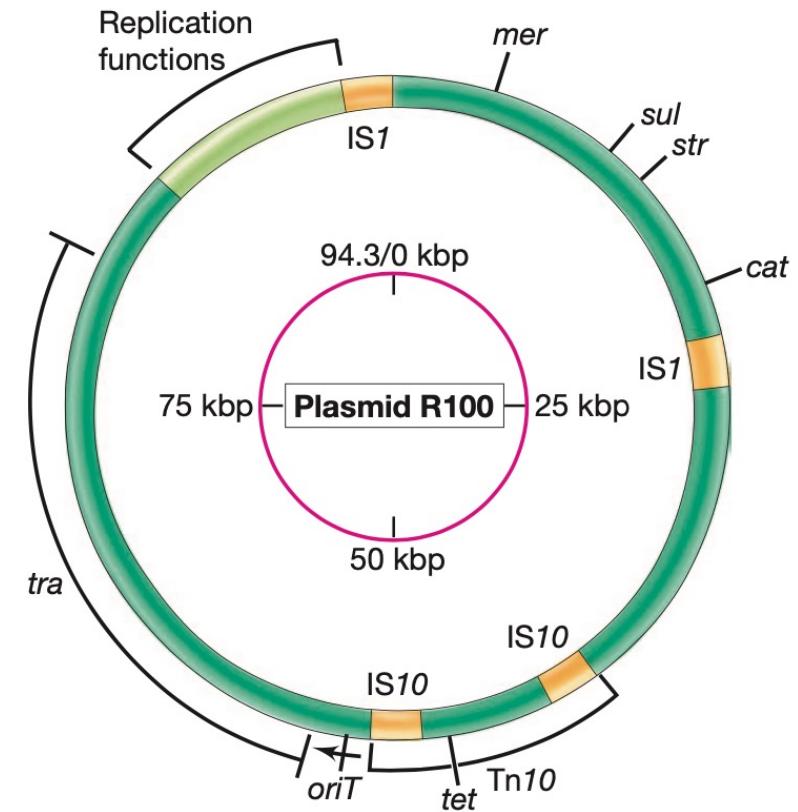
Typical genomic structure in prokaryotes



The chromosome of *Escherichiacoli* strain K-12

- Distances are given in 100 kilobases of DNA.
- The chromosome contains 4,639,675 base pairs and 4288 open reading frames (genes).
- Genes are on either strand.
- Depending on the DNA strand, the locations of a few genes and operons are indicated.
- Replication proceeds in both directions from the origin of DNA replication, *oriC*, indicated in red.

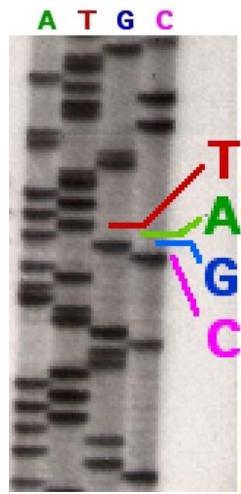
Plasmids



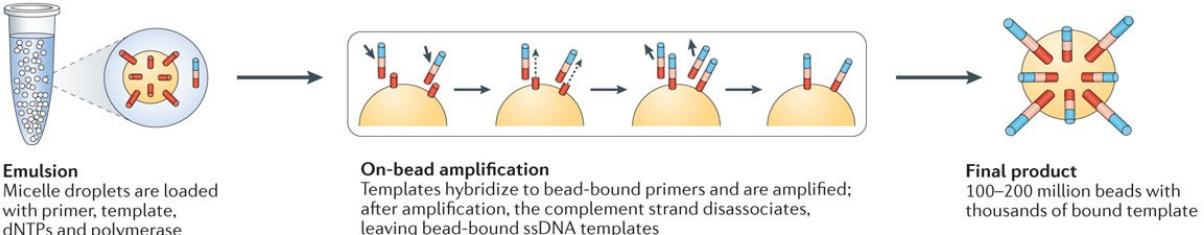
- **Genetic map of the resistance plasmid R100.** The inner circle shows the size in kilobase pairs. The outer circle shows the location of major antibiotic resistance genes and other key functions: *mer*, mercuric ion resistance; *sul*, sulfonamide resistance; *str*, streptomycin resistance; *cat*, chloramphenicol resistance; *tet*, tetracycline resistance; *oriT*, origin of conjugative transfer; *tra*, transfer functions. The locations of insertion sequences (IS) and the transposon Tn10 are also shown. Genes for plasmid replication are found in the region from 88 to 92 kbp.

DNA sequencing

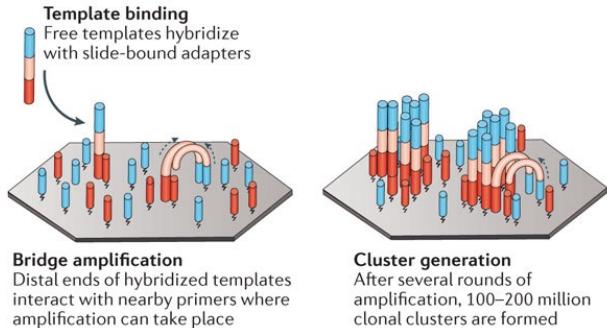
- First machines for sequencing DNA slow and costly e.g. electrophoresis based Sanger
- Followed by next generation machines:
 - Use multiplexed amplification in wells (454) or clusters on plates (Illumina) or chips (IonTorrent)
 - Sequencing by synthesis
- Third generation single molecule sequencing
 - Nanopore
 - Pacific Biosciences



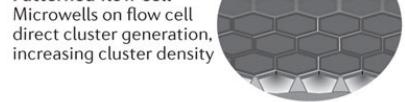
a Emulsion PCR
(454 (Roche), SOLiD (Thermo Fisher), GeneReader (Qiagen), Ion Torrent (Thermo Fisher))



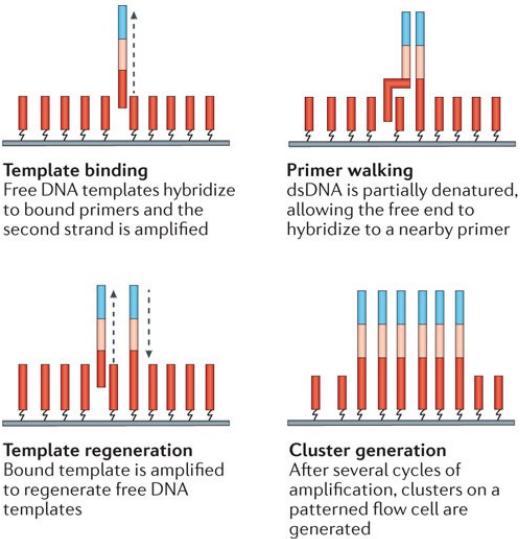
b Solid-phase bridge amplification
(Illumina)



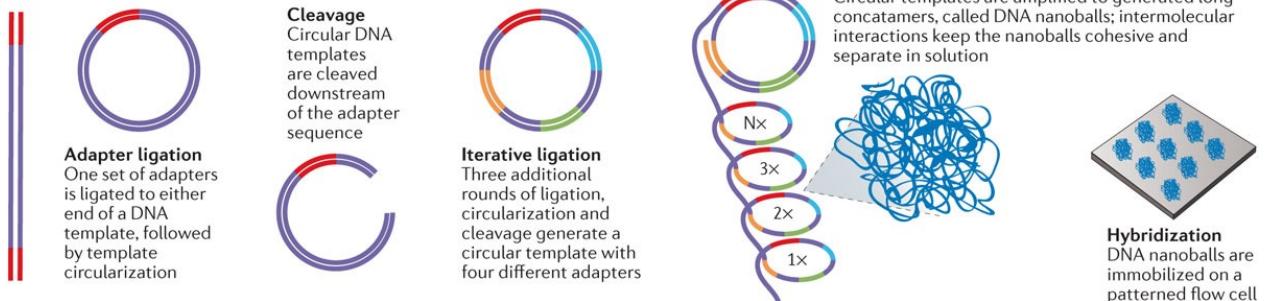
Patterned flow cell



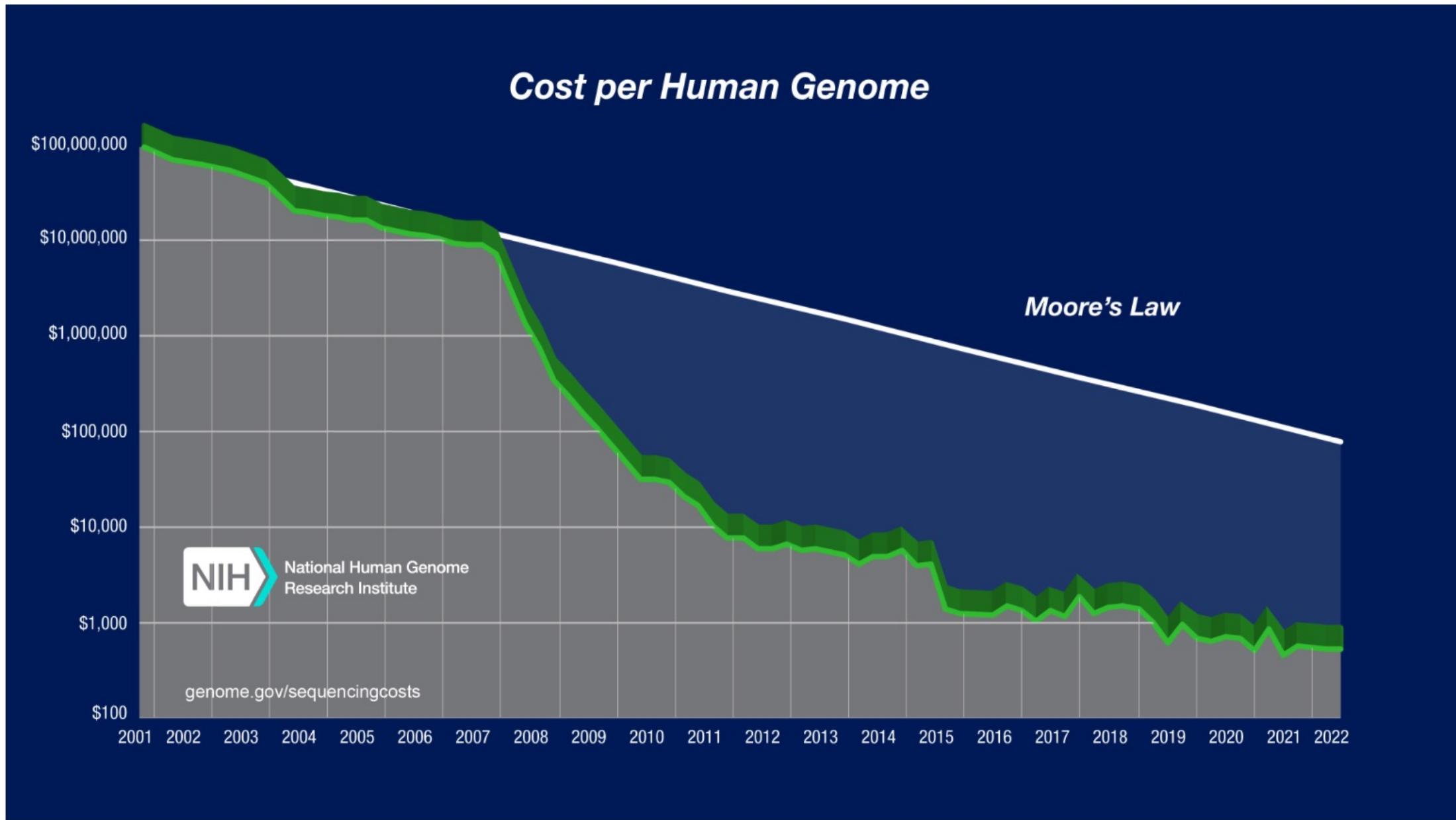
c Solid-phase template walking
(SOLiD Wildfire (Thermo Fisher))



d In-solution DNA nanoball generation
(Complete Genomics (BGI))



Cost per Human Genome



Cost per genome data - 2022

Nanopore sequencing

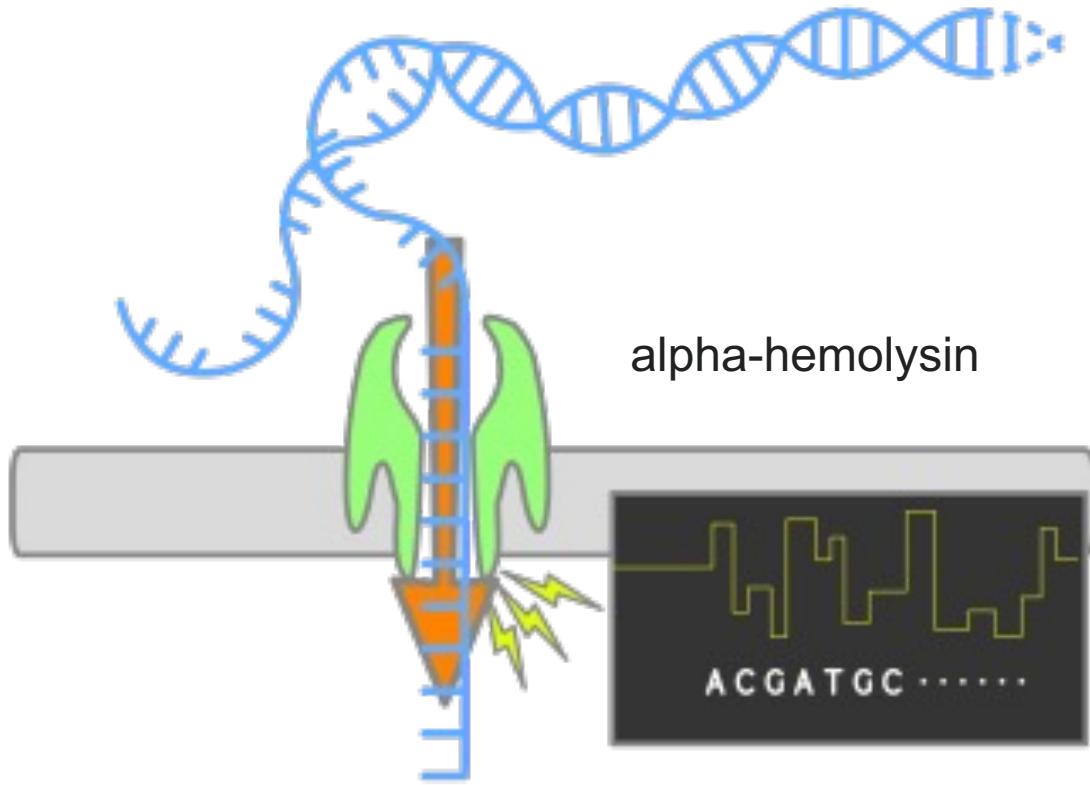


Table 7

Comparison of various high-performing sequencing instruments *.

Manufacturer	Read length	Data output	Max. run time (hours)	Chemistry	Key applications **
Illumina (NovaSeq 6000)	300 PE	6 Tb (6000 Gb)	44	Sequencing by synthesis	SS-WGS and TGS, TGEP, 16sMGS, WES, SCP, LS-WGS, CA, MS, MGP, CFS, LBA
Thermo Fisher Scientific Ion Torrent (Ion GeneStudio S5 Prime)	600 SE	50 Gb	12	Sequencing by synthesis	WGS, WES, TGS
GenapSys (16 chips)	150 SE	2 Gb	24	Sequencing by synthesis	TS, SS-WGS, GEV, 16S rRNA sequencing, sRNA sequencing, TSCAS
QIAGEN (GeneReader)	100 SE	Not available	Not available	Sequencing by synthesis	Cancer research and identifying mutations
BGI/Complete Genomics	400 SE	6 Tb (6000 Gb)	40	DNA nanoball	Small and large WGS, WES and TGS
PacBio (HiFi Reads)	25 Kb	66.5 Gb	30	Real-time sequencing	DN sequencing, FT, identifying ASI, mutations, and EPM
Nanopore (PromethION)	4 Mb	14 Tb (14000 Gb)	72	Real-time sequencing	SV, GS, phasing, DNA and RNA base modifications, FT, and isoform detection

*Performance comparison is given as per manufacturer's description. ** Applications by all sequencers of the respective manufacturer are listed. **Full names are given in Abbreviations.

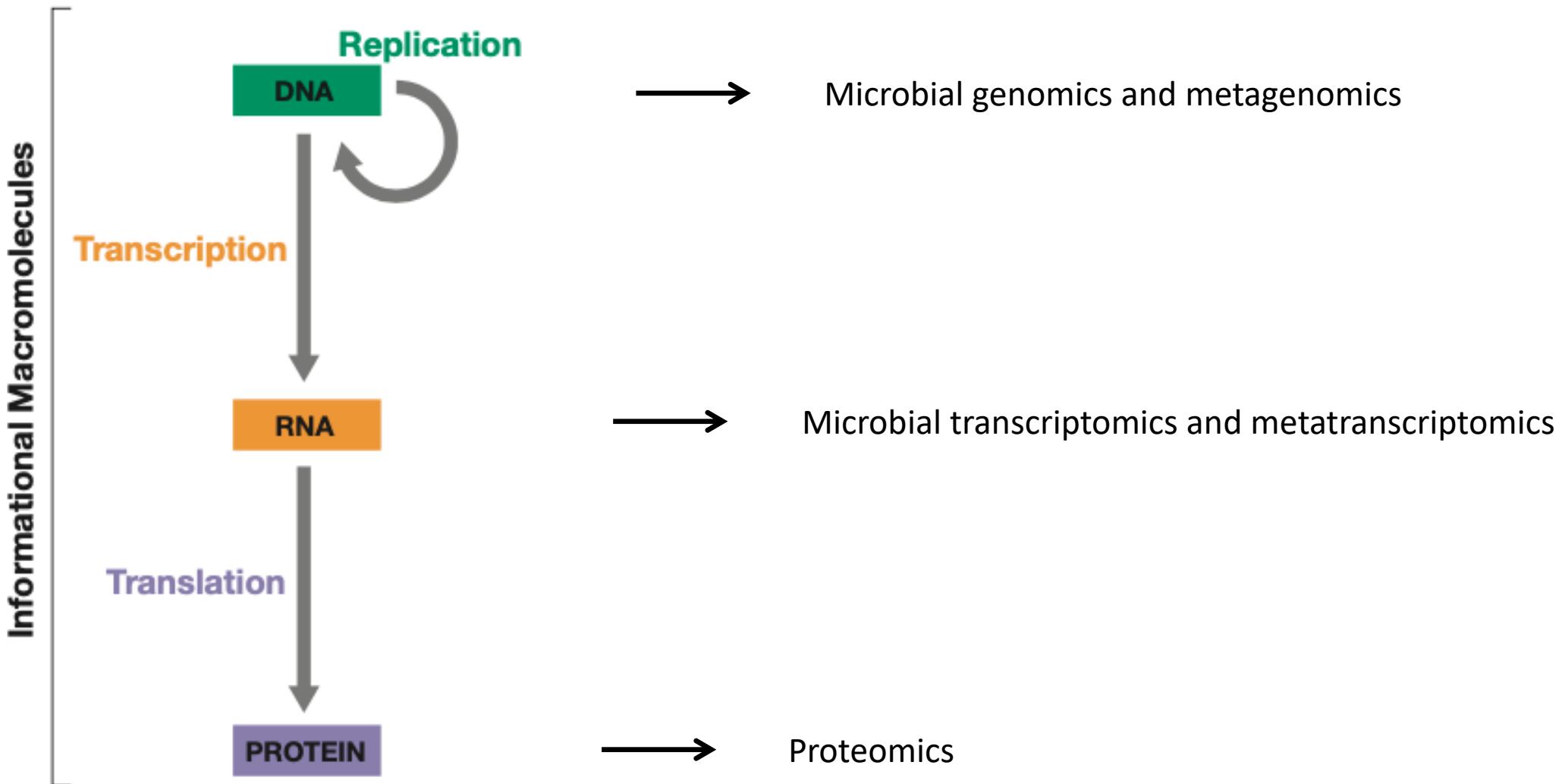
Basically:**Illumina cheap short reads****Nanopore cheap long noisy reads****HiFi PacBio expensive long accurate reads**

High-Performance Sequencing Platform

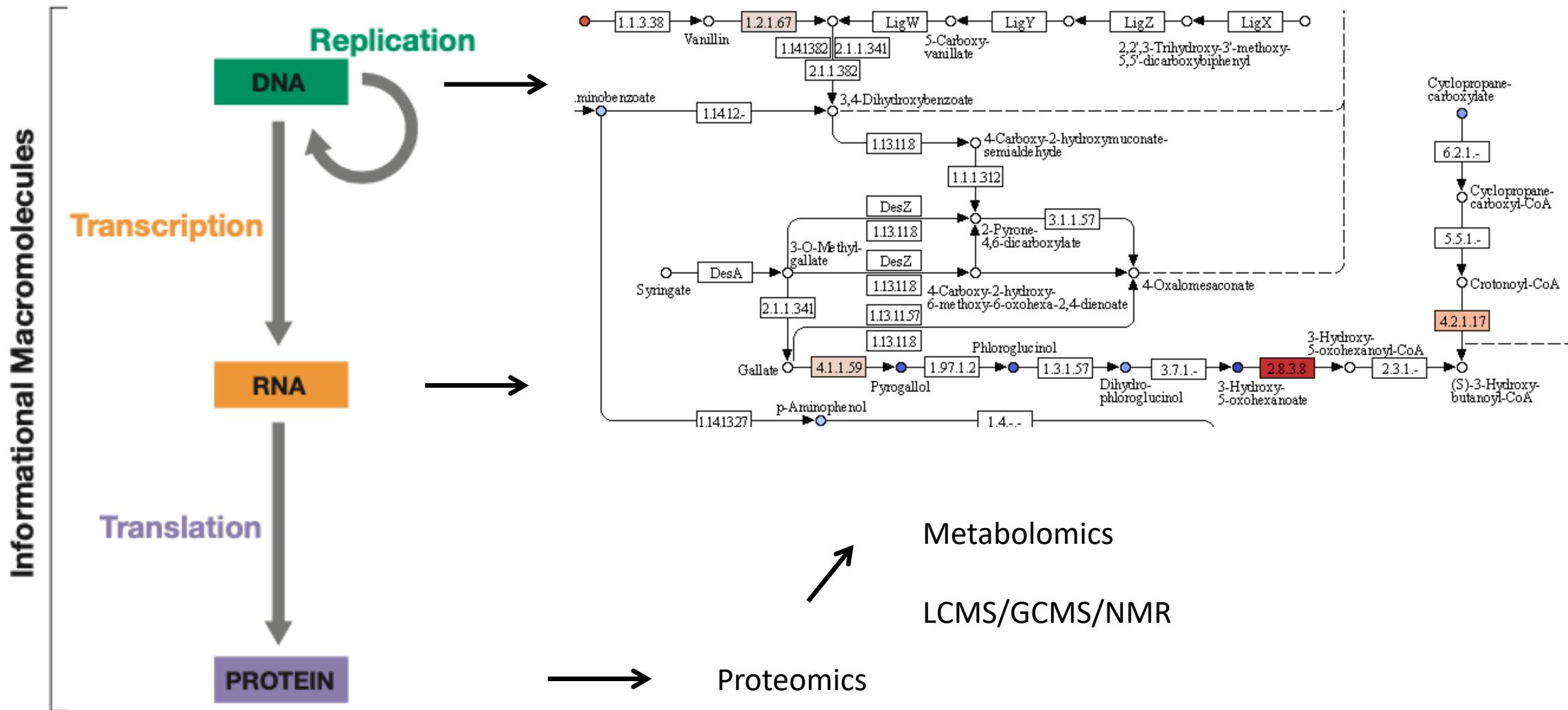
2023 upgrade

Short reads	Illumina NovaSeq X Plus	Illumina NextSeq 1000	Illumina iSeq100
	<p>Illumina NovaSeq X Plus</p> <p>2 independent stages (2.5B) 10B (25B) 100-300 cycles 2 8 lanes 2 channels Dragen built-in XLEAP SBS In production ✓</p>	 <p>Illumina NextSeq 1000</p> <p>100M 300/400M 100-600 cycles 1 lane 2 channels Dragen built-in Standard SBS In production ✓</p>	 <p>Illumina iSeq100</p> <p>4M clusters 300 cycles 1 lane 1 channel Standard SBS Self-service Coming soon</p>
	<p>Pacific Biosciences Revio</p> <p>4 independent stages 25M ZMWs NVIDIA GPU built-in In production ✓</p>	 <p>Pacific Biosciences Sequel IIe</p> <p>4 stages 8M ZMWs IsoSeq Small amplicons In production ✓</p>	 <p>Oxford Nanopore P2 Solo</p> <p>2 independent stages 2,675 channels GPU workstation Low maintenance cost Validation in progress</p>

Other 'omics in the study of the microbiome



Other 'omics in the study of the microbiome



Summary

- Microbes key roles both as individual players and in a community context
- Technology has driven the ‘omics revolution of microbiology
- Understand what aspects of cellular dynamics ‘omics data relates too