ETH zürich

551-1119-00L Microbial Community Genomics

Introduction to Comparative Genomics

Sequencing technologies: an historical perspective

- 1953: Discovery of the structure of DNA
- 1965: “Sequencing” of the first tRNA
- 1972: Sequencing of first complete gene (coat protein of bacteriophage MS2)
- 1977: Release of “**chain termination method**” → **FIRST GENERATION SEQUENCING**
- 1996: Beginning of **SECOND or NEXT-GENERATION SEQUENCING**
- 2005: Implementation of pyrosequencing in automated system
- 2007: Illumina acquires Solexa
- 2010: Beginning of **THIRD-GENERATION SEQUENCING**

- 1953: Discovery of the structure of DNA

James Watson

Francis Crick

Maurice Wilkins

Rosalind Franklin

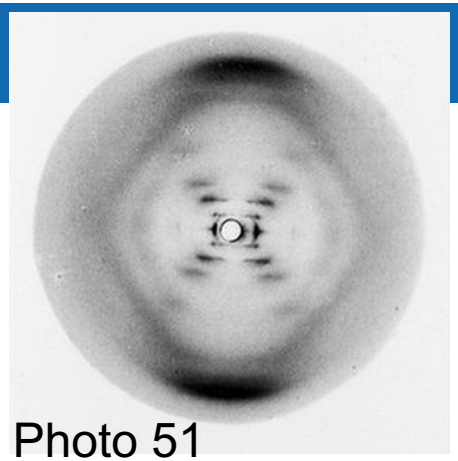
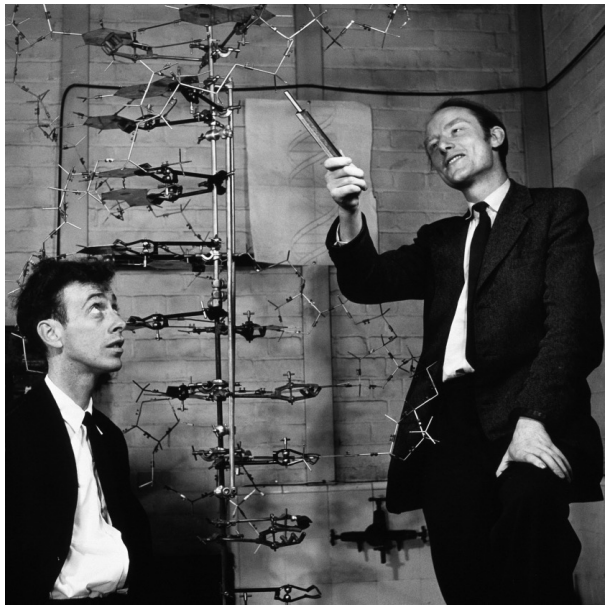
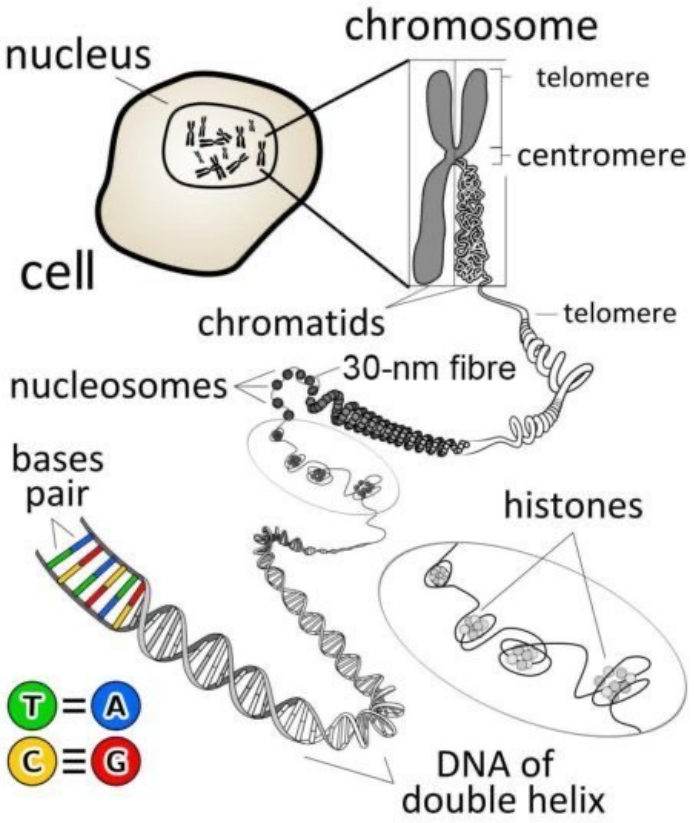


Photo 51



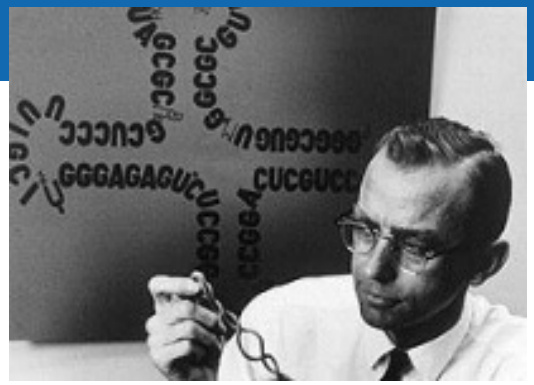


https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/

- 1953: Discovery of the structure of DNA
- 1965: “Sequencing” of the first tRNA

→ use of ribonucleases with cleaving sites at specific nucleotides

→ reconstruction of the original nucleotide sequence by determining the order in which small fragments occurred in the tRNA molecule



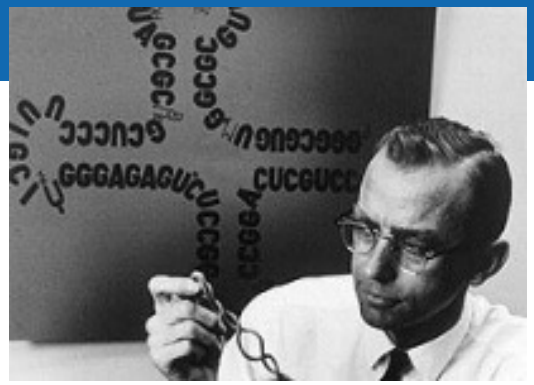
Robert W. Holley

https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/

- 1953: Discovery of the structure of DNA
- 1965: “Sequencing” of the first tRNA

  → use of ribonucleases with cleaving sites at specific nucleotides

  → reconstruction of the original nucleotide sequence by determining the order in which small fragments occurred in the tRNA molecule
- 1972: Sequencing of first complete gene (coat protein of bacteriophage MS2) via RNAse digestion and isolation of oligonucleotides



Robert W. Holley

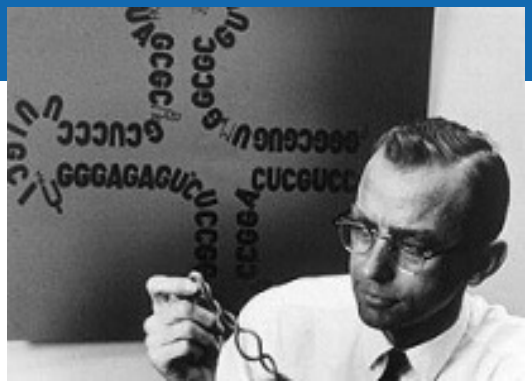
Walter Fiers

https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/

- 1953: Discovery of the structure of DNA
- 1965: "Sequencing" of the first tRNA

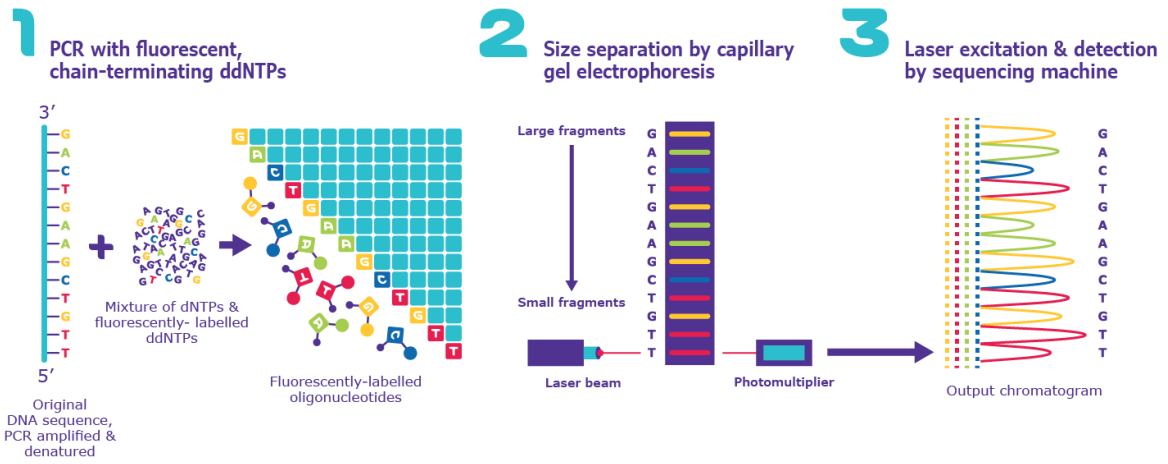
  → use of ribonucleases with cleaving sites at specific nucleotides

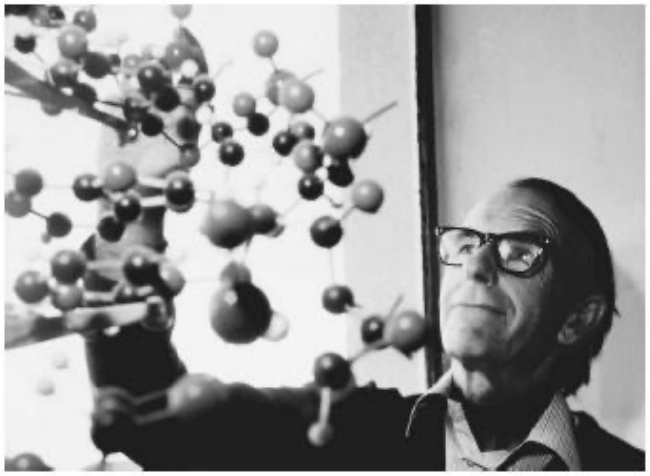
  → reconstruction of the original nucleotide sequence by determining the order in which small fragments occurred in the tRNA molecule
- 1972: Sequencing of first complete gene (coat protein of bacteriophage MS2) via RNAse digestion and isolation of oligonucleotides
- 1977: Release of "**chain termination method**" utilizing radiolabeled partially digested fragments → **FIRST GENERATION SEQUENCING**



Robert W. Holley

Walter Fiers



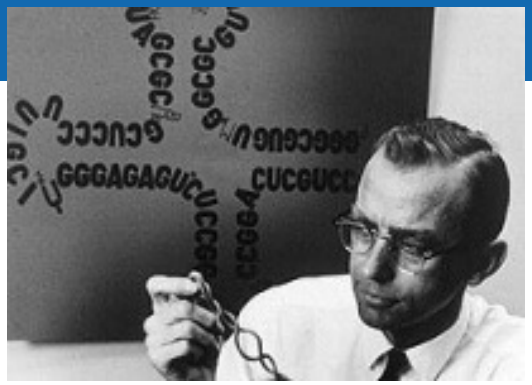


Frederick Sanger

- 1953: Discovery of the structure of DNA
- 1965: "Sequencing" of the first tRNA

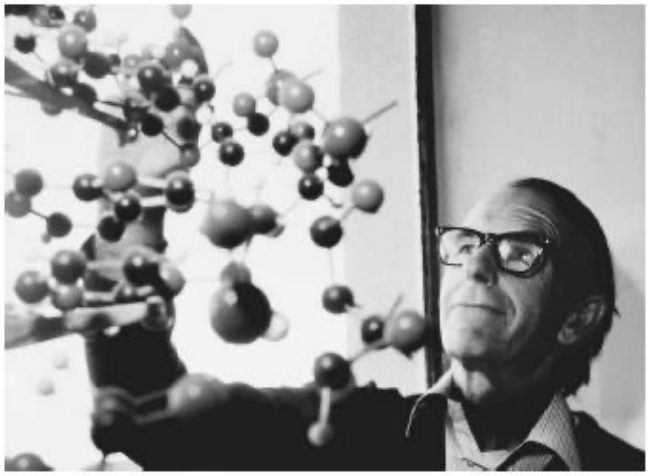
  → use of ribonucleases with cleaving sites at specific nucleotides

  → reconstruction of the original nucleotide sequence by determining the order in which small fragments occurred in the tRNA molecule
- 1972: Sequencing of first complete gene (coat protein of bacteriophage MS2) via RNAse digestion and isolation of oligonucleotides
- 1977: Release of "**chain termination method**" utilizing radiolabeled partially digested fragments → **FIRST GENERATION SEQUENCING**



Robert W. Holley

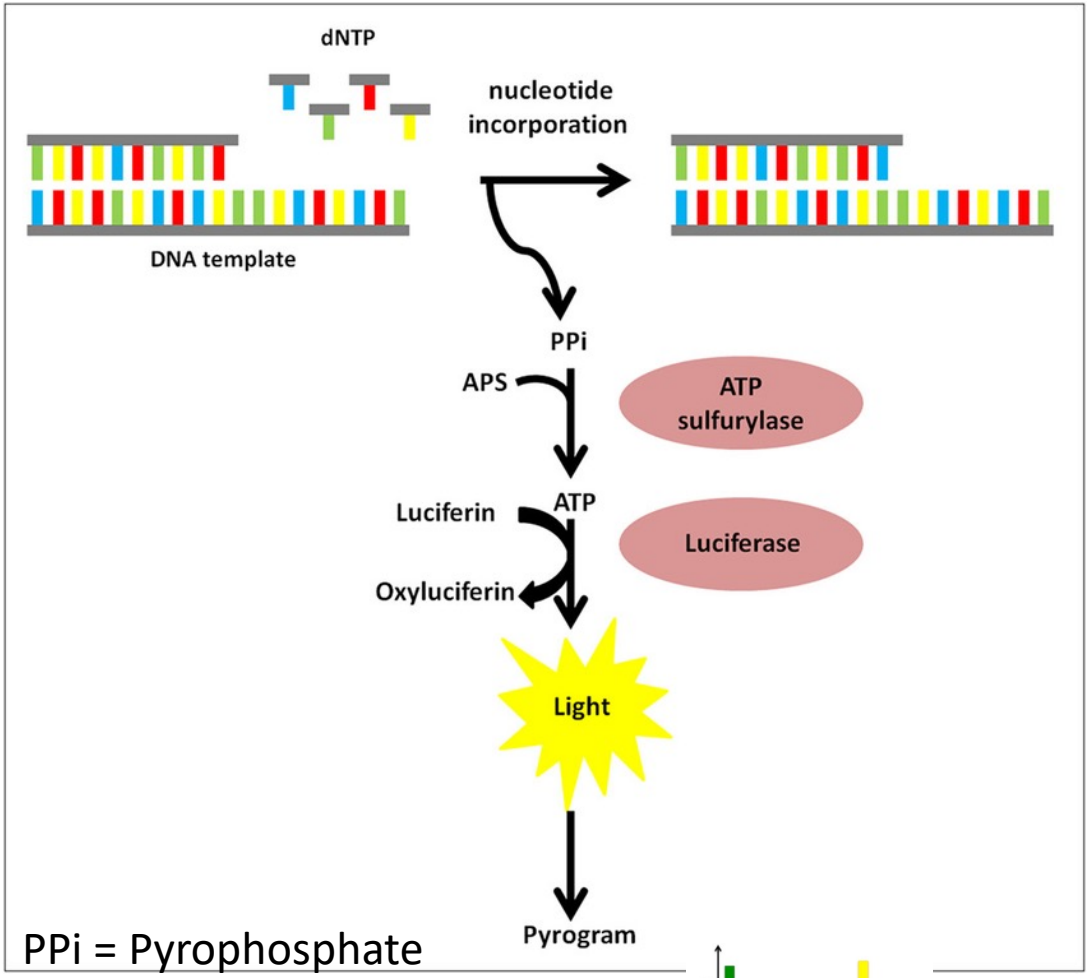
Walter Fiers

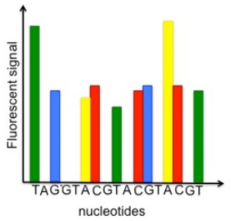


Frederick Sanger

→ Main sequencing technology for next 25 years
→ Key innovations mainly in automation of wet-lab and data analysis pipelines

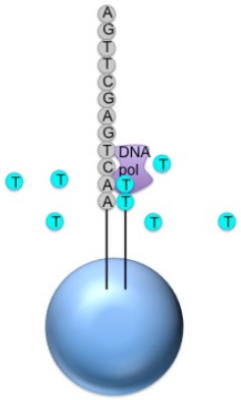
- 1996: Beginning of **NEXT-GENERATION SEQUENCING**
  → Pyrosequencing

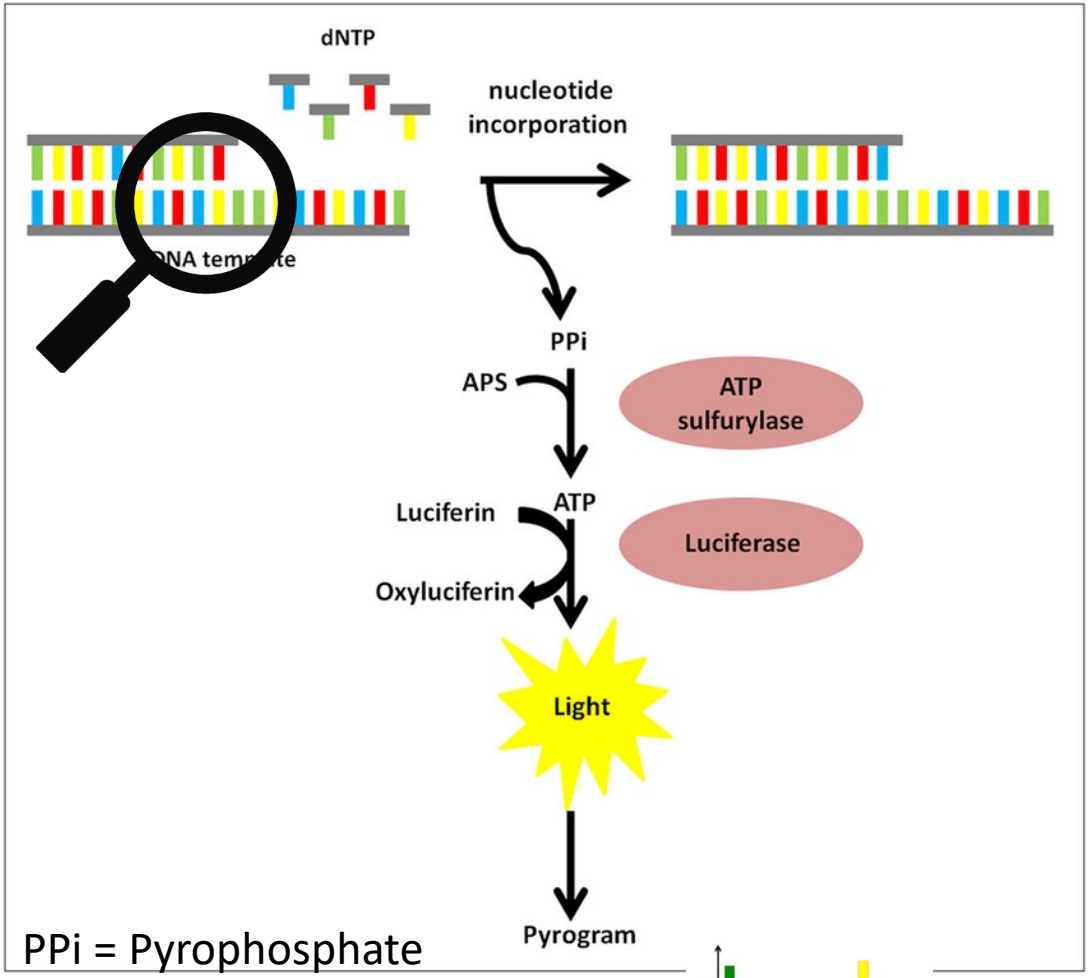




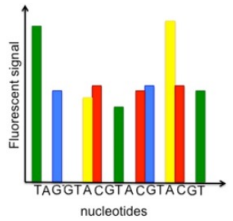
https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/

- 1996: Beginning of **NEXT-GENERATION SEQUENCING**
  → Pyrosequencing



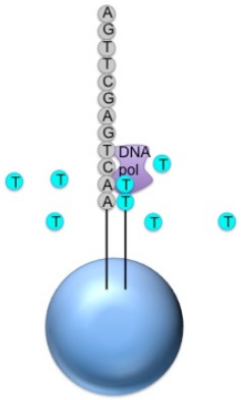


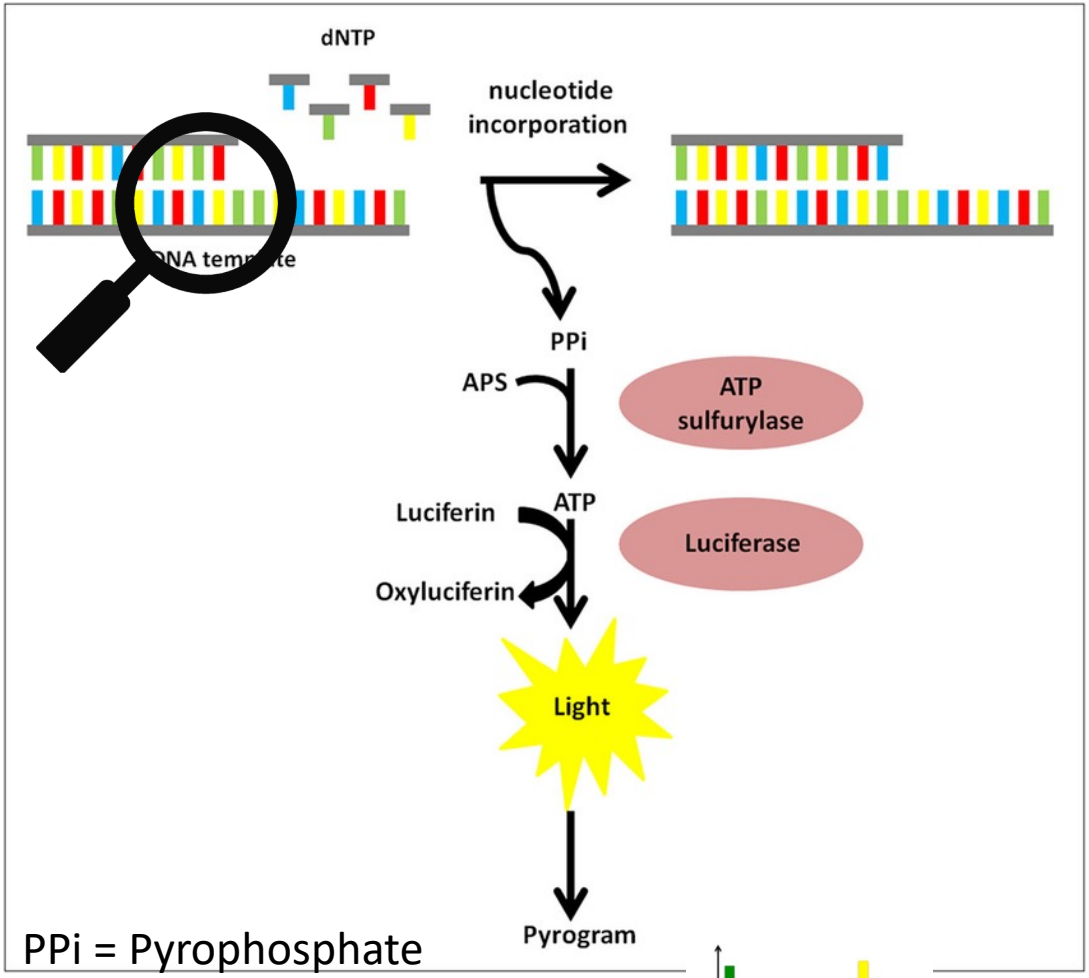
PPi = Pyrophosphate



https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/
https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/next-generation-sequencing/454-sequencing/

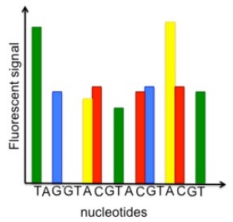
- 1996: Beginning of **NEXT-GENERATION SEQUENCING**
  → Pyrosequencing



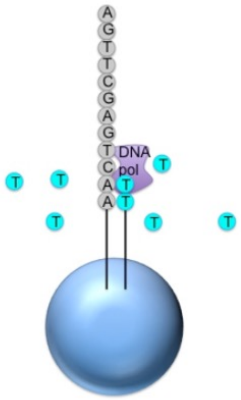


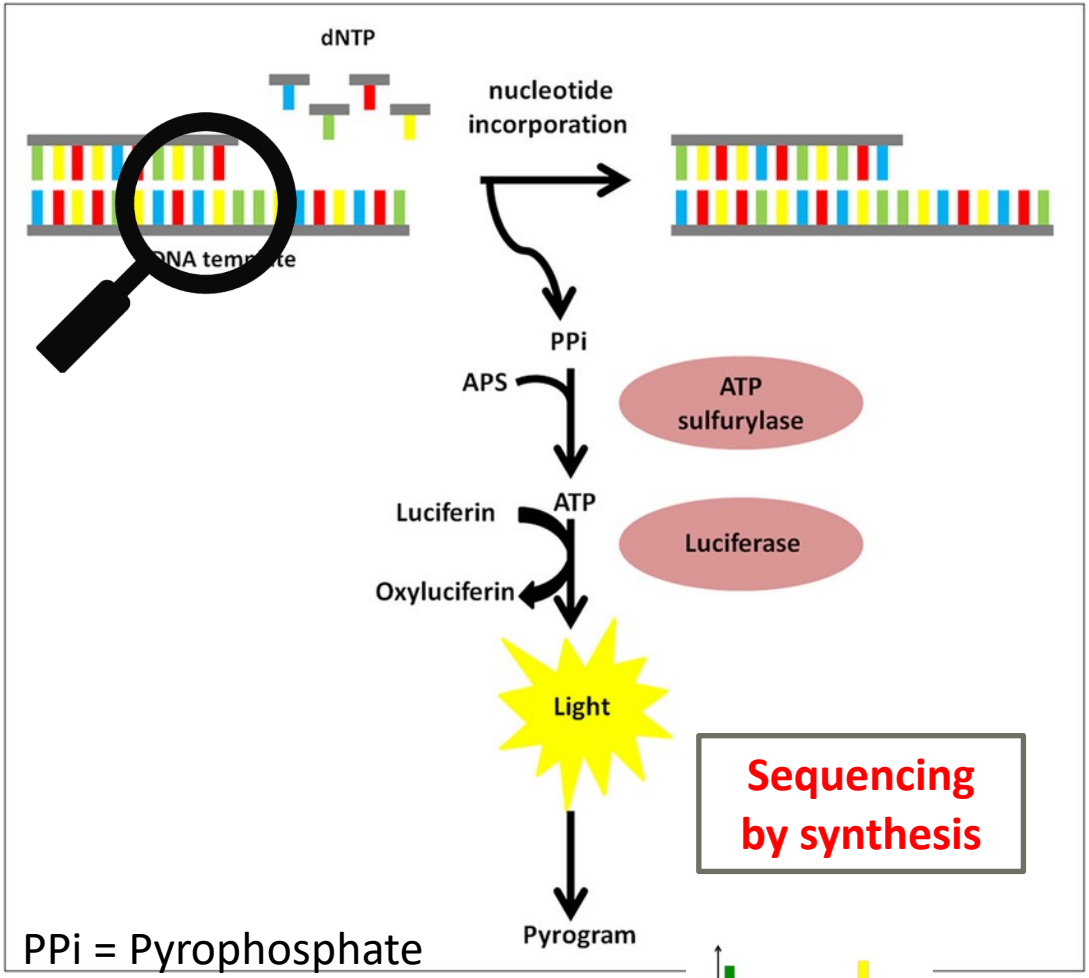
→Each read will have a different length because different numbers of nucleotides will be added during each wash

PPi = Pyrophosphate

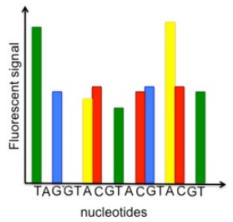


- 1996: Beginning of **NEXT-GENERATION SEQUENCING**
  → Pyrosequencing





→Each read will have a different length because different numbers of nucleotides will be added during each wash



- 1996: Beginning of **NEXT-GENERATION SEQUENCING**
  → Pyrosequencing
- 2005: Implementation of pyrosequencing in automated system
  → 454 sequencing platform

Roche 454 Sequencing System

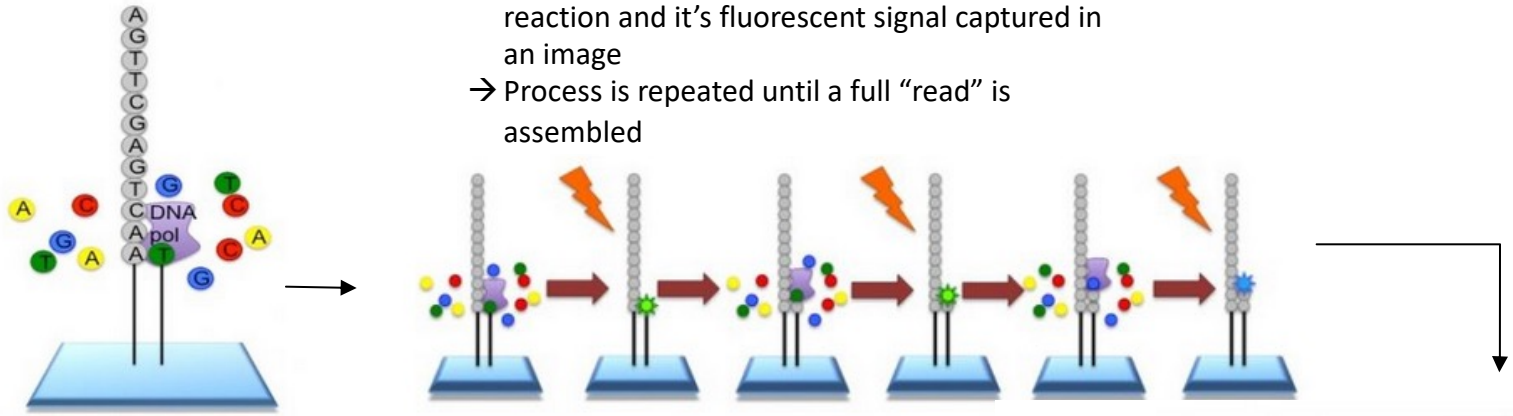
https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/

- 1996: Beginning of **NEXT-GENERATION SEQUENCING**
  - → Pyrosequencing
- 2005: Implementation of pyrosequencing in automated system
  - → 454 sequencing platform
- 2007: Illumina acquires Solexa
  - → Advanced sequencing technology
  - → Improved throughput

Illumina MiSeq Sequencing platform

→ In each cycle, one dNTP is incorporated into the reaction and it's fluorescent signal captured in an image

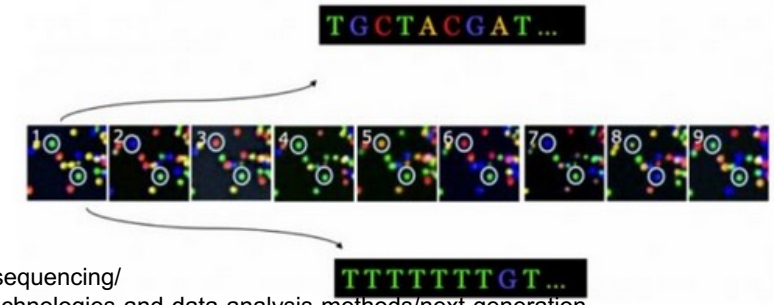
→ Process is repeated until a full "read" is assembled



→ Single-stranded short DNA sequences are bound to a glass flow cell

→ Flow cell is flooded with fluorescently labelled dNTPs and polymerase

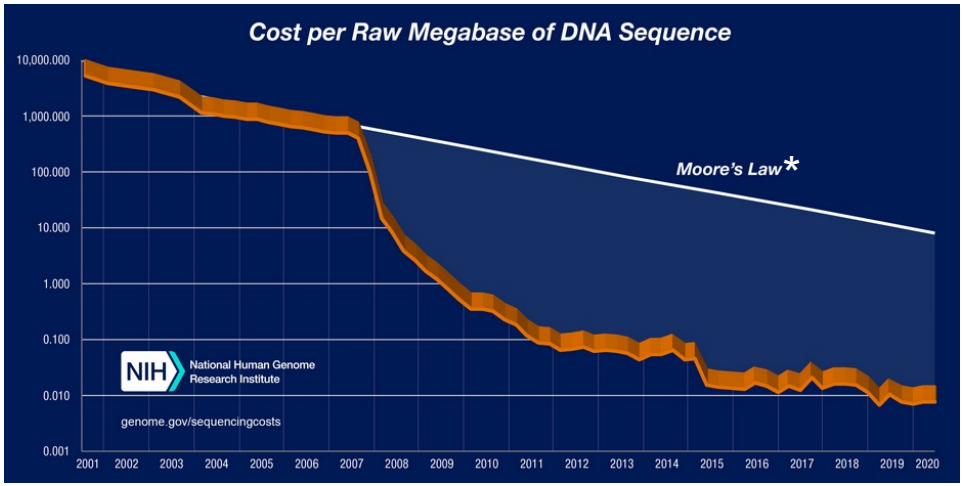
→ Translation of thousands of fluorescent images into a sequence read

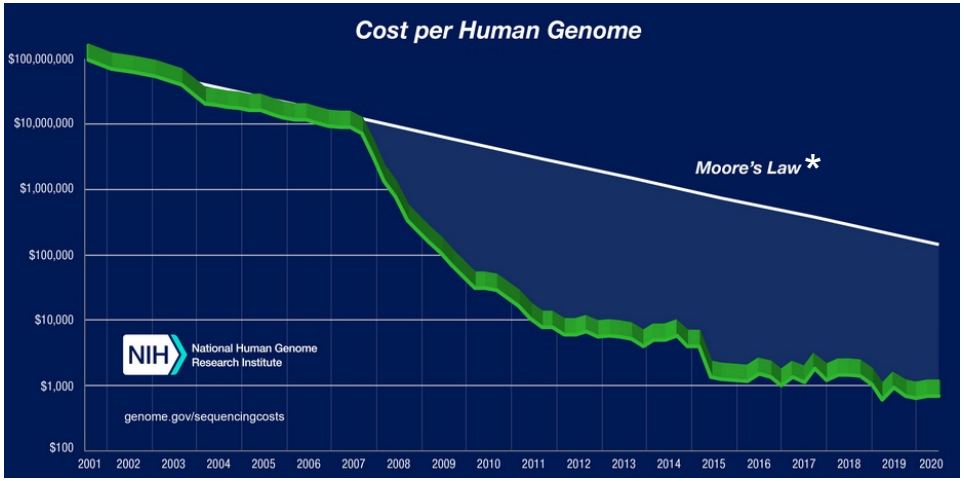


D BIOL

https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/
https://www.ebi.ac.uk/training/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/next-generation-sequencing/454-sequencing/

Improvements in DNA sequencing: Some numbers…



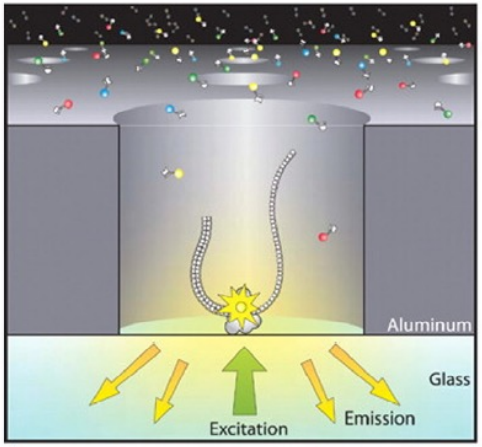


*Moore's law is an observation and projection of a historical trend. Rather than a law of physics, it is an empirical relationship linked to gains from experience in production

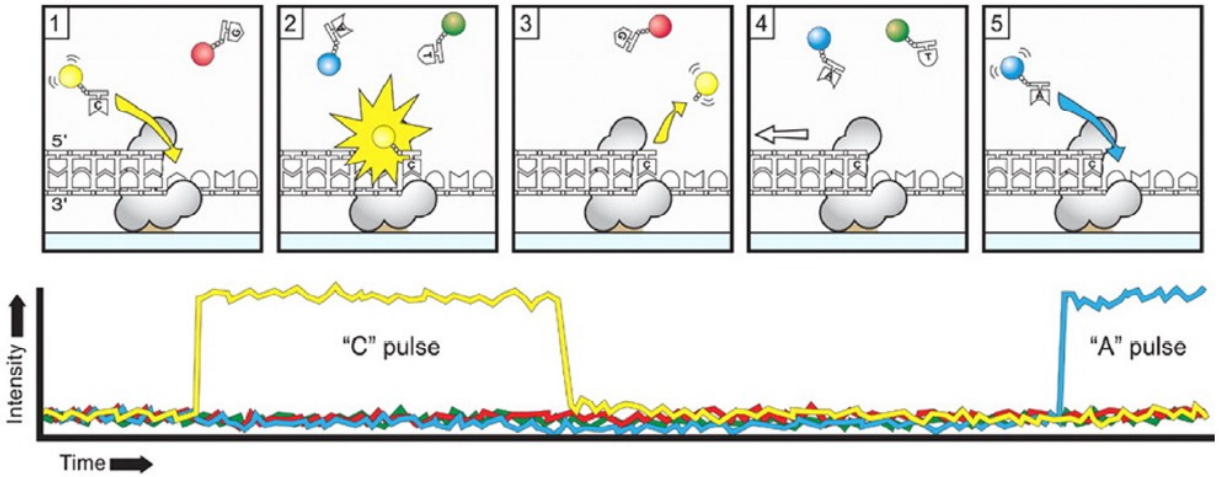
https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

- 2010: Beginning of **THIRD-GENERATION SEQUENCING**
  → PacBio sequencing (Pacific Biosciences, Inc.)

PacBio RSII sequencer



→ polymerase immobilized at the bottom of a "well" (zero-mode waveguide ZMW) in a SMRTcell



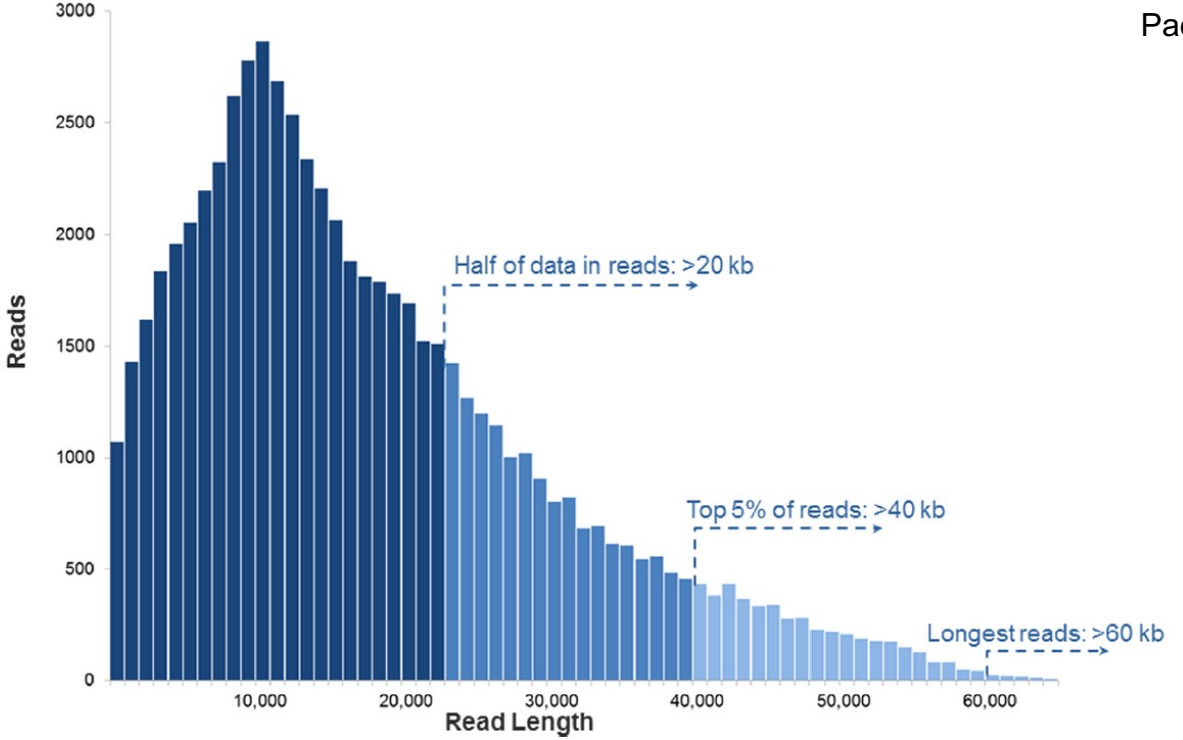
→ Incorporation of fluorescent dNTPs produces a base-specific light pulse
→ Replication process in all ZMWs is recorded as a "movie" in real-time

https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/
Rhoads & Au (2015) PacBio sequencing and its applications *Genomics, Proteomics & Bioinformatics* 13(5): 278-289

- 2010: Beginning of **THIRD-GENERATION SEQUENCING**
  → PacBio sequencing (Pacific Biosciences, Inc.)

PacBio RSII sequencer

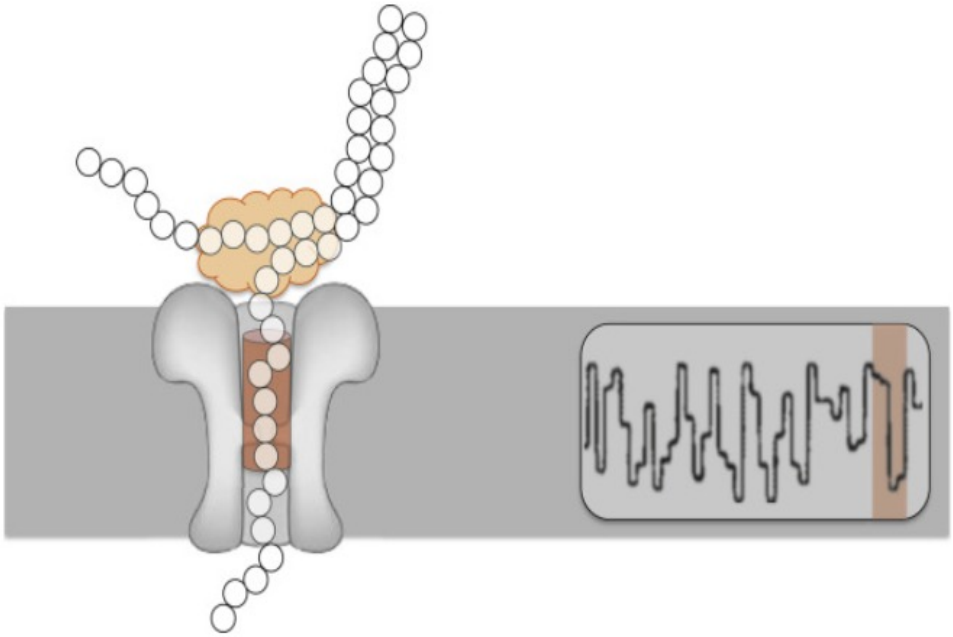
→ **Generation of long-reads!!**



https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/
Rhoads & Au (2015) PacBio sequencing and its applications *Genomics, Proteomics & Bioinformatics* 13(5): 278-289

- 2010: Beginning of **THIRD-GENERATION SEQUENCING**
  - → PacBio sequencing (Pacific Biosciences, Inc.)
  - → Nanopore sequencing (Oxford Nanopore Technologies)

Nanopore MinION

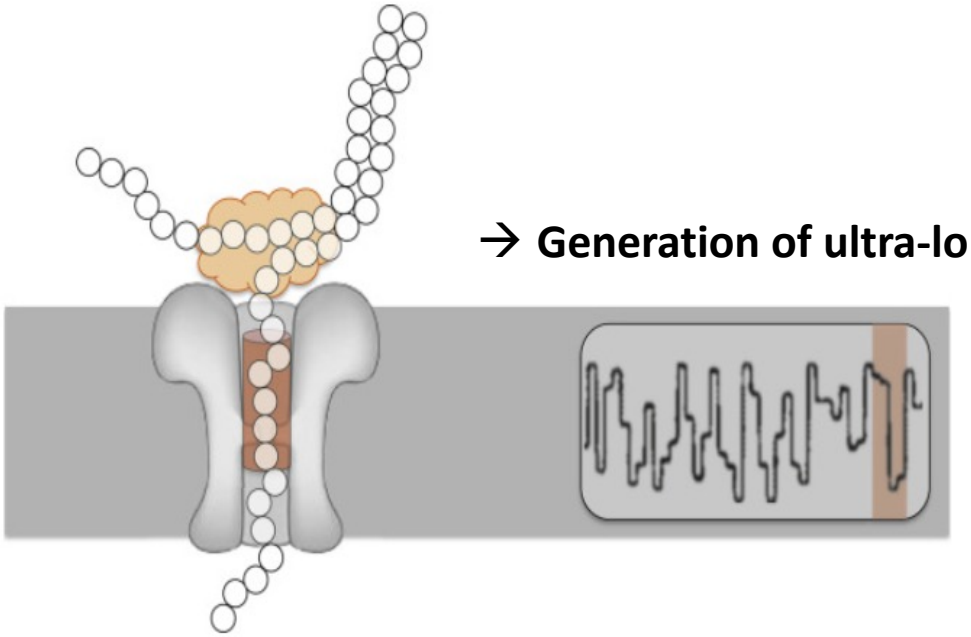


- → single-stranded DNA/RNA molecules pass through protein nanopore
- → Each nucleotide that passes the pore leads to a different change in electrical current across pore
- → Resulting signal is decoded to provide sequence information

https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/
https://www.sciencedirect.com/topics/neuroscience/nanopore-sequencing
https://nanoporetech.com/applications/dna-nanopore-sequencing

- 2010: Beginning of **THIRD-GENERATION SEQUENCING**
  → PacBio sequencing (Pacific Biosciences, Inc.)
  → Nanopore sequencing (Oxford Nanopore Technologies)

Nanopore MinION

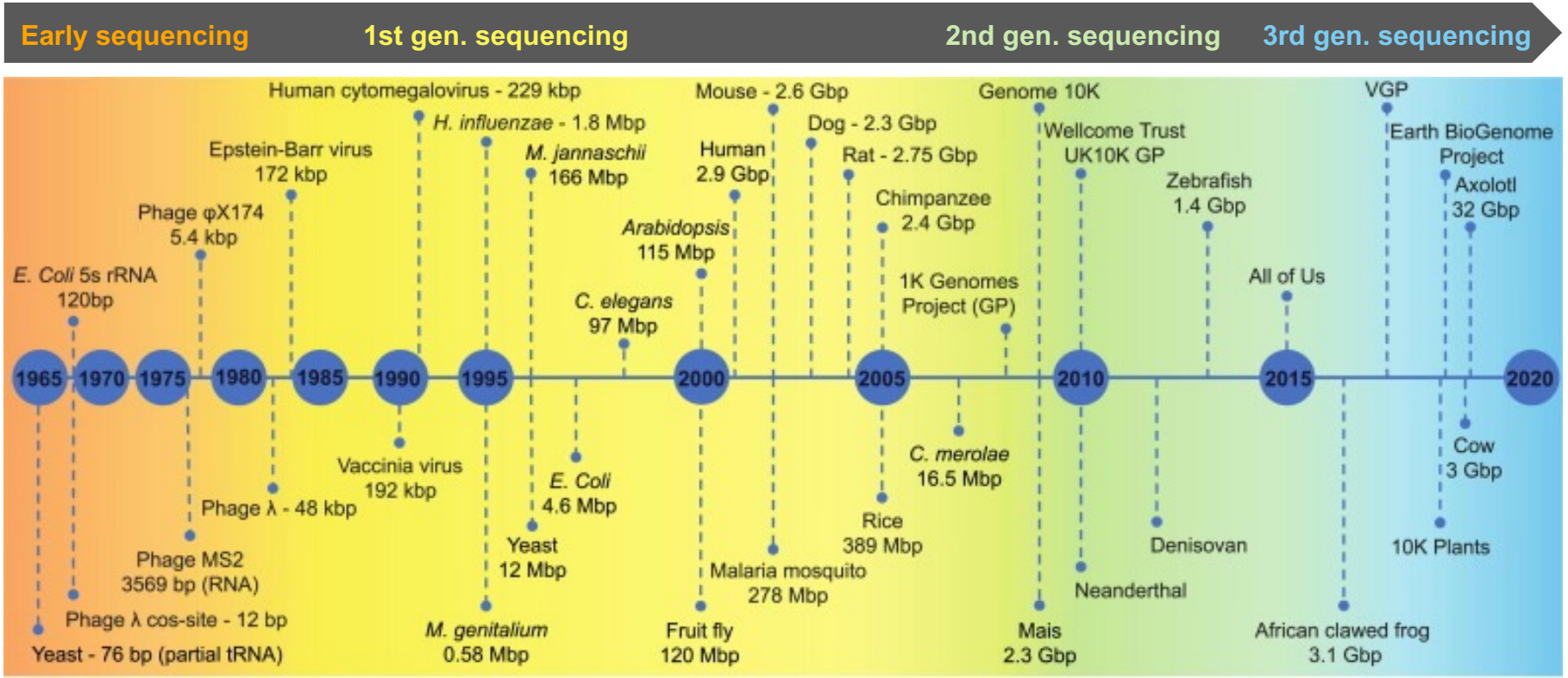


→ **Generation of ultra-long reads (2Mb)!!**

→ single-stranded DNA/RNA molecules pass through protein nanopore
→ Each nucleotide that passes the pore leads to a different change in electrical current across pore
→ Resulting signal is decoded to provide sequence information

https://the-dna-universe.com/2020/11/02/a-journey-through-the-history-of-dna-sequencing/
https://www.sciencedirect.com/topics/neuroscience/nanopore-sequencing
https://nanoporetech.com/applications/dna-nanopore-sequencing

Genome sequencing: an historical perspective



https://doi.org/10.1016/j.csbj.2019.11.002

See also: https://www.nature.com/immersive/d42859-020-00099-0/index.html for milestones of genome sequencing.

The great plate count anomaly

- The "**great plate count anomaly**" is the term we use to describe the observation that microscopic cell counts are significantly higher than corresponding counts of "*colony forming units*" on agar plates.

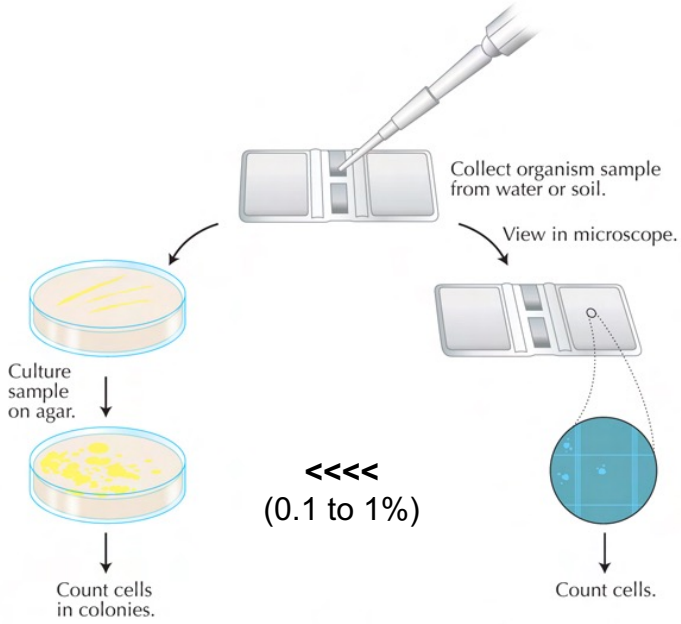
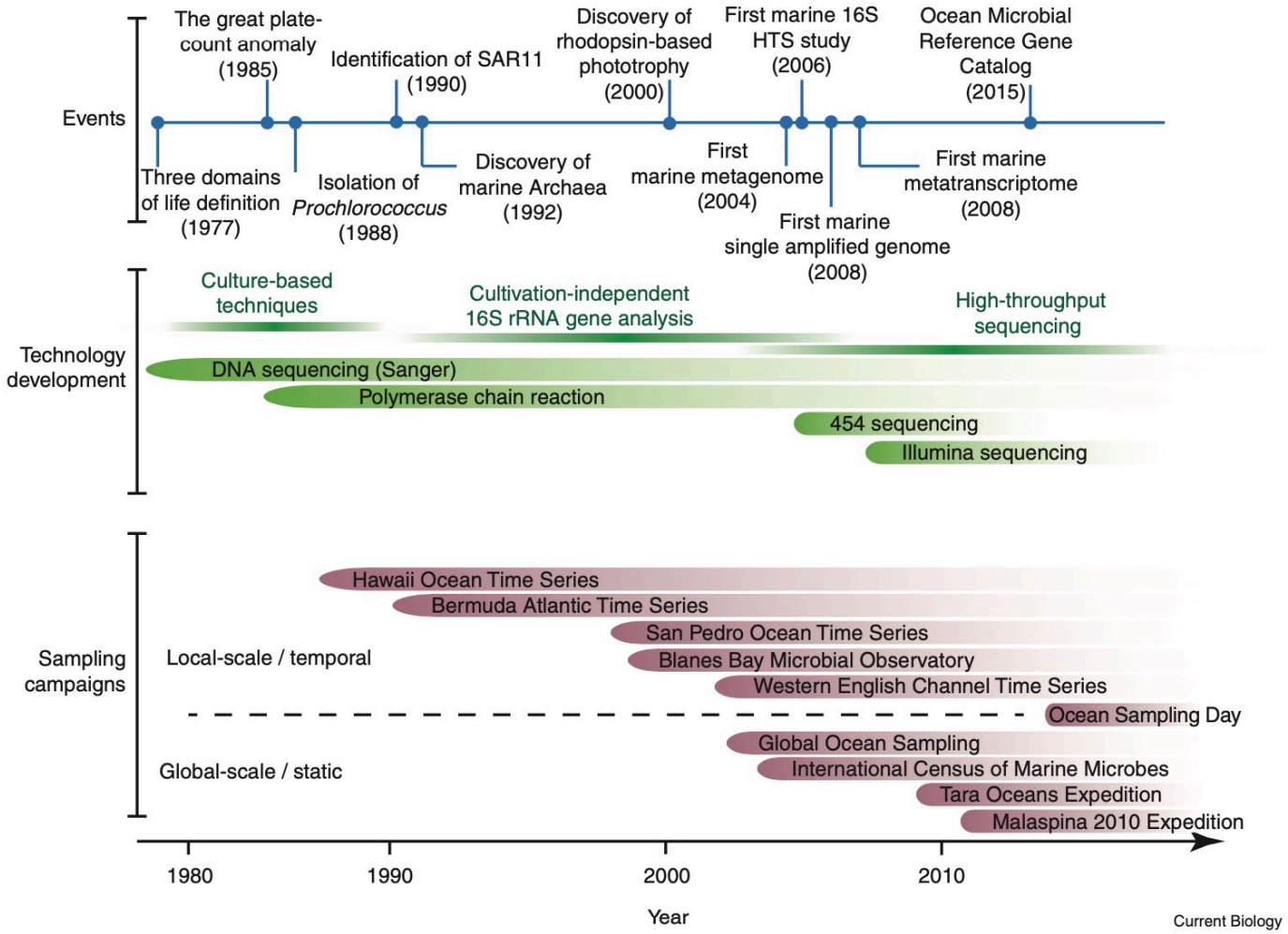


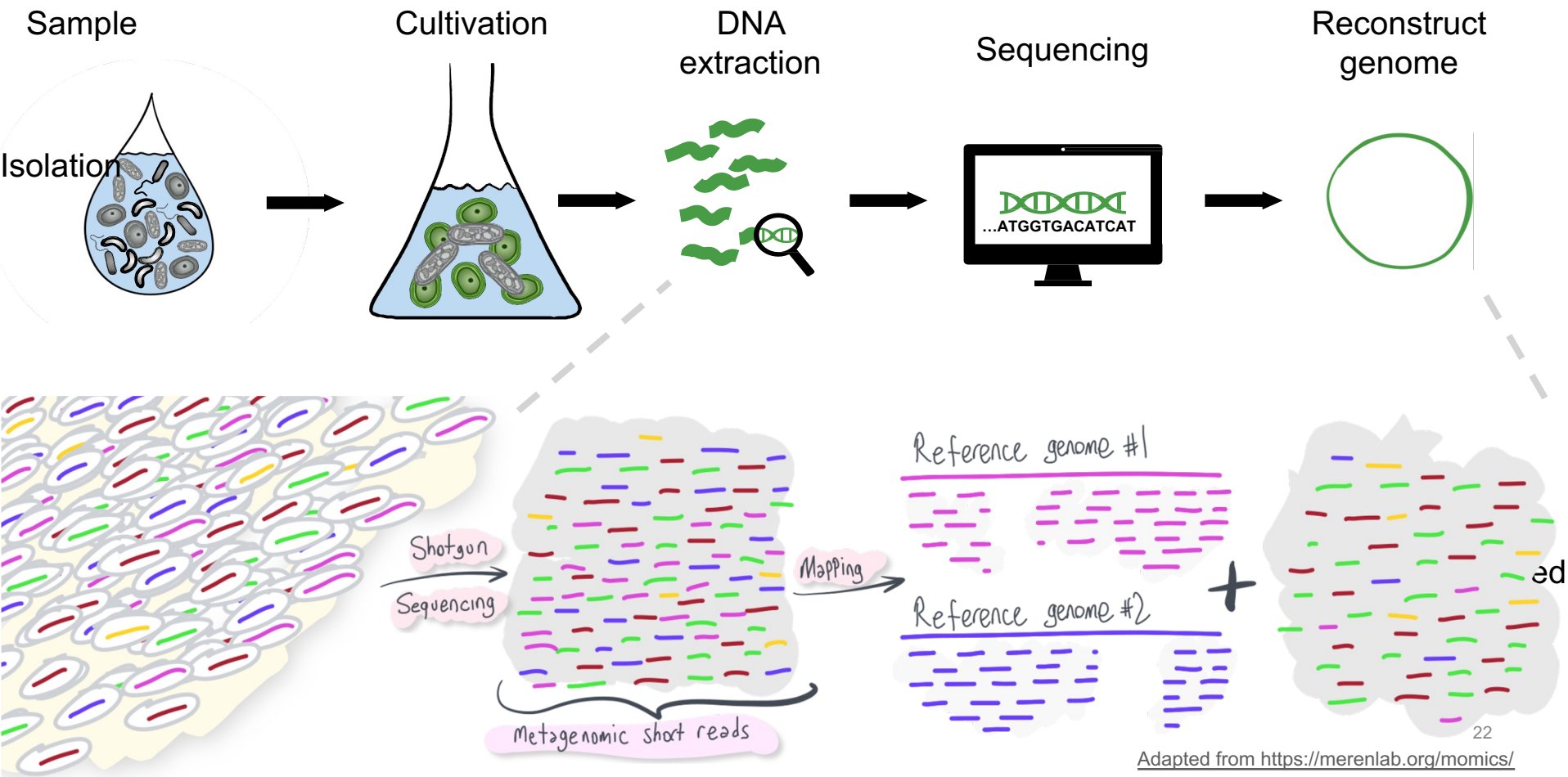
FIGURE 6.12. The great plate count anomaly. Plate counts of cells obtained by cultivation are usually much lower, sometimes by orders of magnitude, than those from direct cell counts under a microscope. Possible reasons are (1) the differing nutritional requirements of the organism, (2) the organism may enter a noncultivatable resting state, or (3) the organism may rely on other organisms and thus cannot be cultivated in isolation.

Marine microbial diversity: an historical perspective



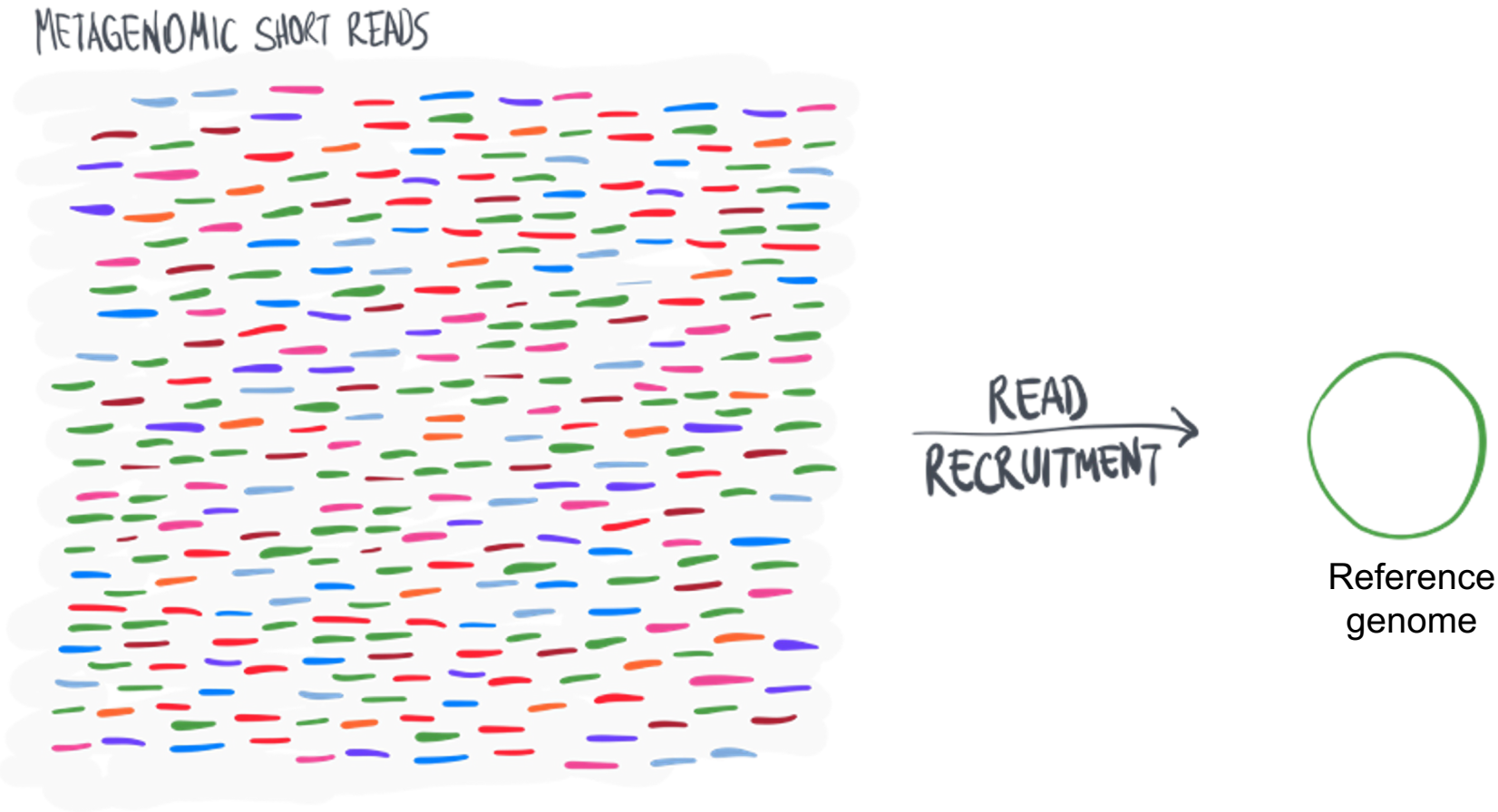
Salazar & Sunagawa, 2017, Current Biology

Traditional microbiology | | Culture-based microbiology

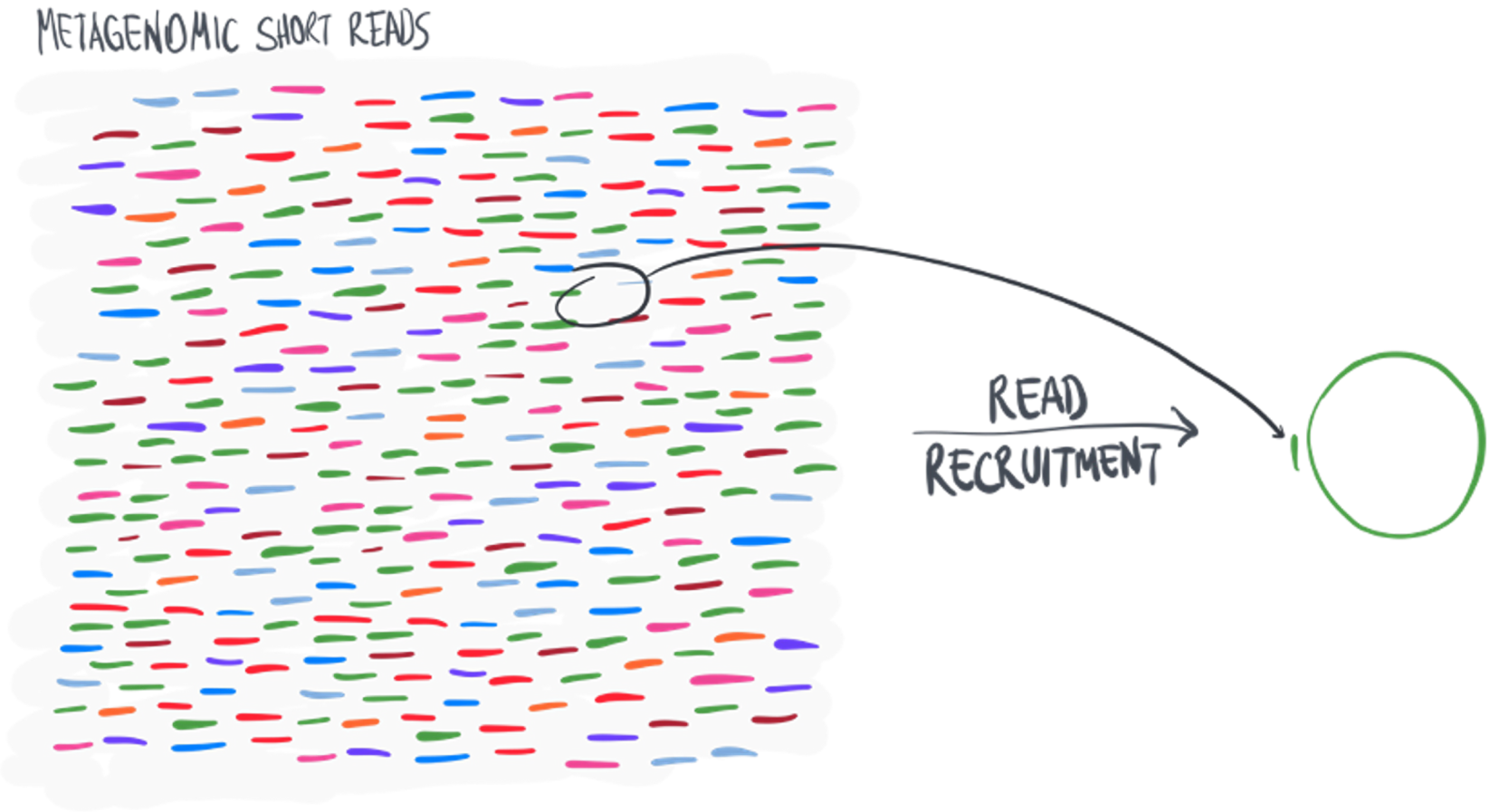


Adapted from https://merenlab.org/momics/

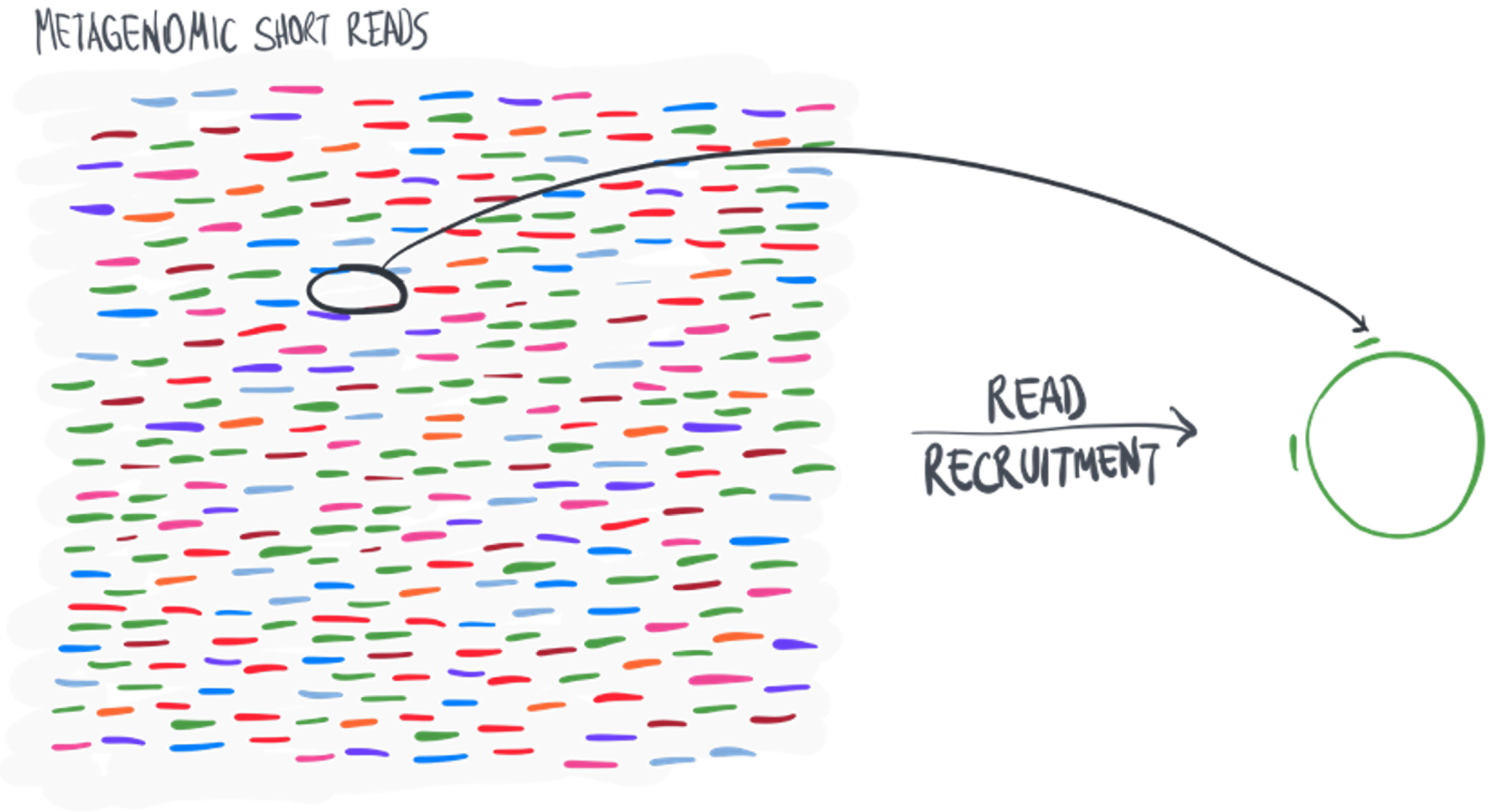
Metagenomics | | Mapping to a reference



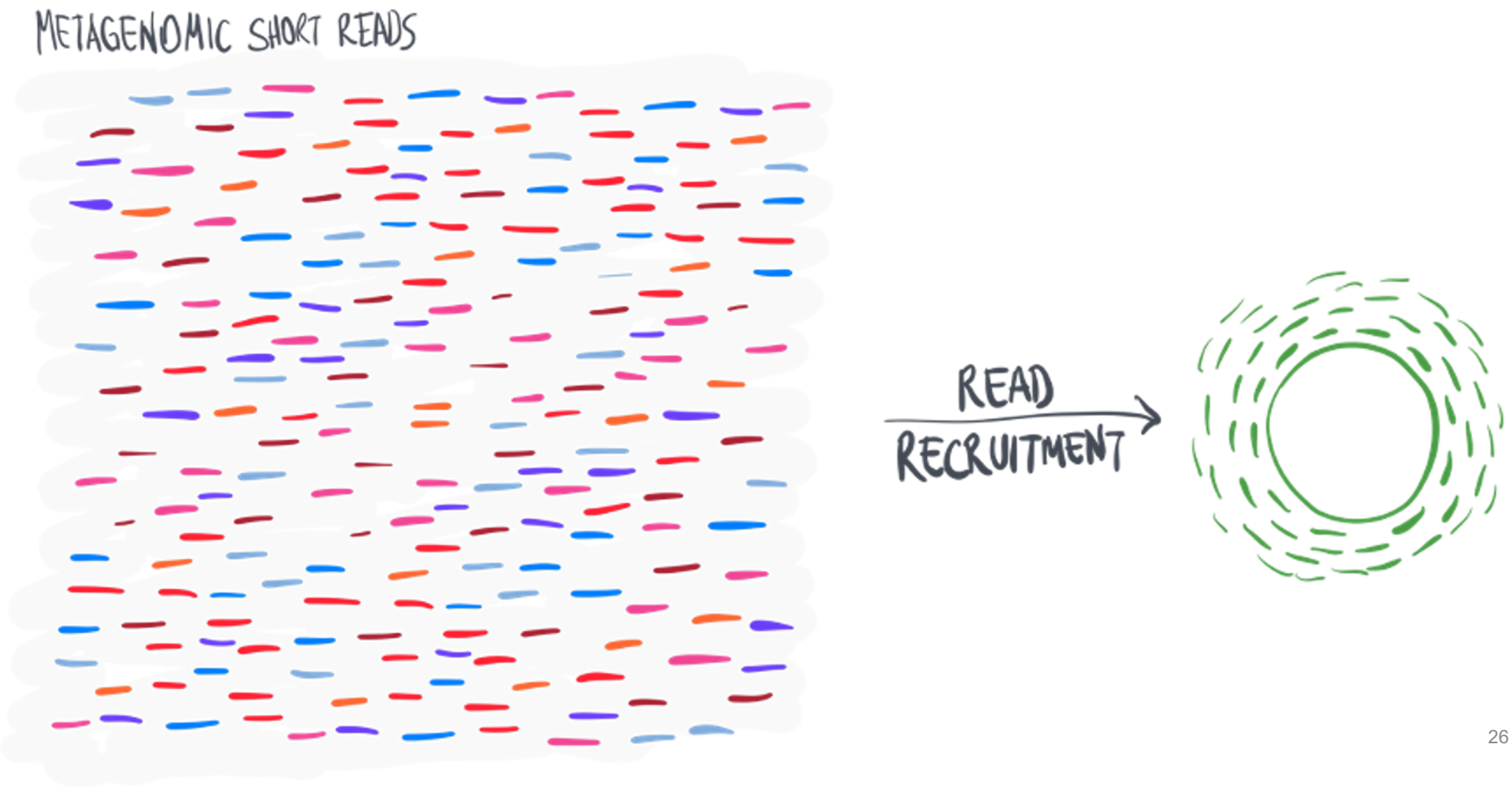
Metagenomics | | Mapping to a reference



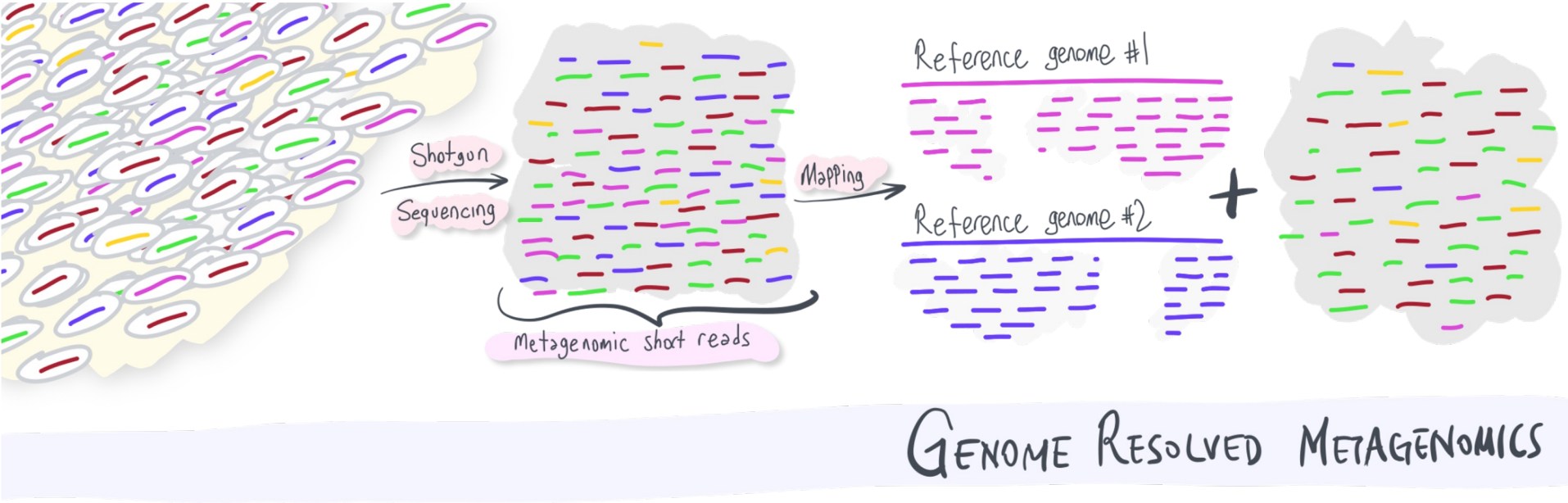
Metagenomics | | Mapping to a reference

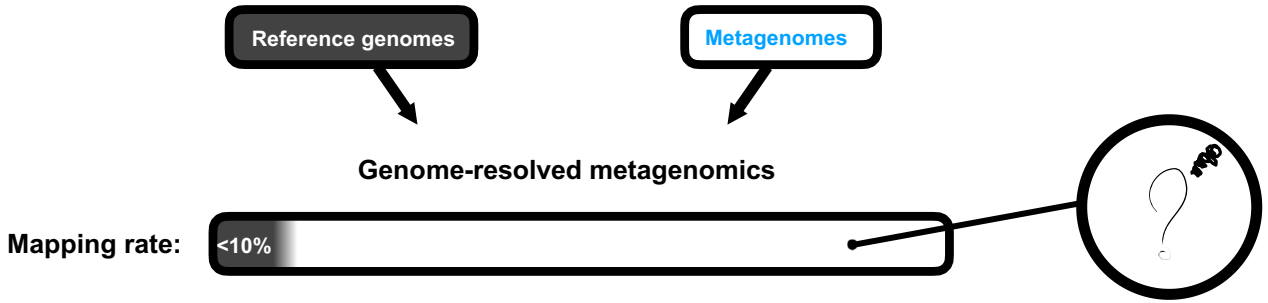


Metagenomics | | Mapping to a reference



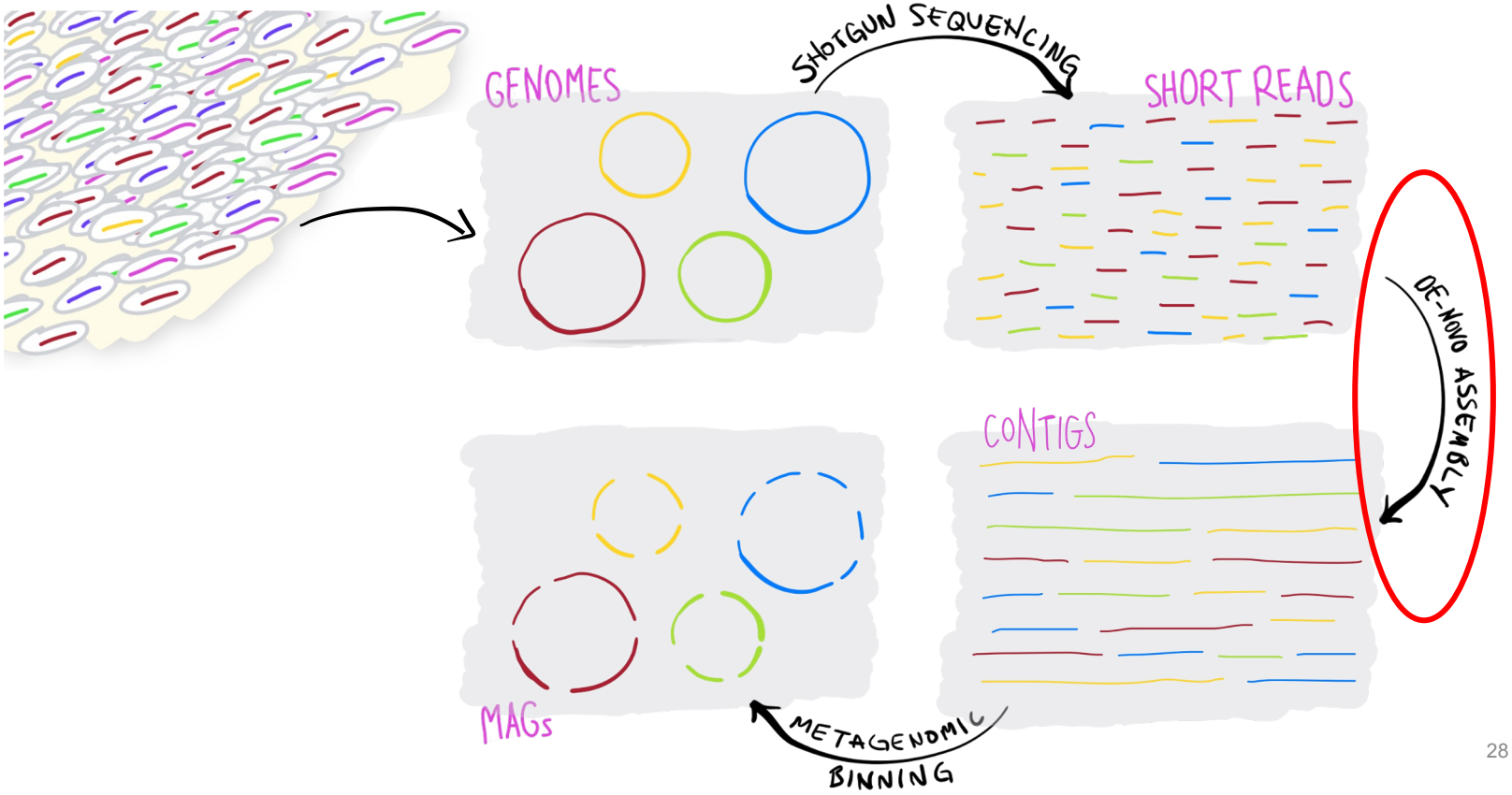
Metagenomics | | Mapping rates





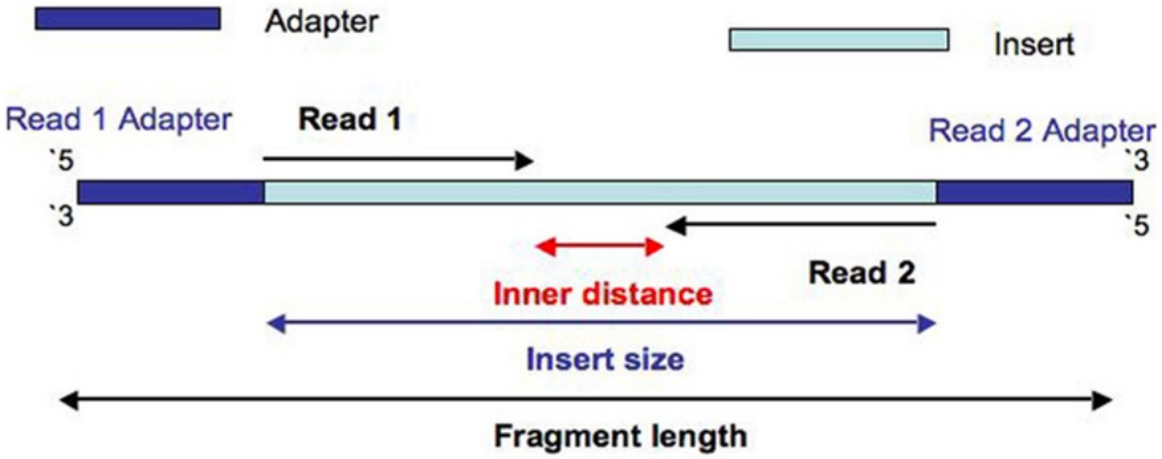
Current state | | Culture-independent microbiology

Current approaches DO NOT require to isolate organisms or reference genomes:



Background: DNA sequencing libraries

- DNA extracted from a metagenomic sample is randomly sheared into inserts of known size distribution (i.e., min, max, mean)
- Adapters are added to facilitate the sequencing of these inserts



Note: Paired end reads may be overlapping providing the possibility to “merge” reads into one or not. In the latter case, the sequence between the paired reads remains unknown, while the length can be estimated due to the known insert size distribution

Step 1: Data quality control - sources of errors

1. Low base calling quality scores

original sequence

A C T G A A C T A A G T A

sequenced read

A C T G A A C T C A A N A T T T A G C T G C A

40 40 40 40 30 20 40 40 10 30 20 5 30 adapter

1. Same base
2. Different base
3. Additional base
4. Undefined base

Other sources of error

2. Residual adapter sequences
3. Residual control DNA sequences (e.g., “PhiX spike-ins”)
4. Contamination from non-target organisms

Base calling quality (*phred*) scores

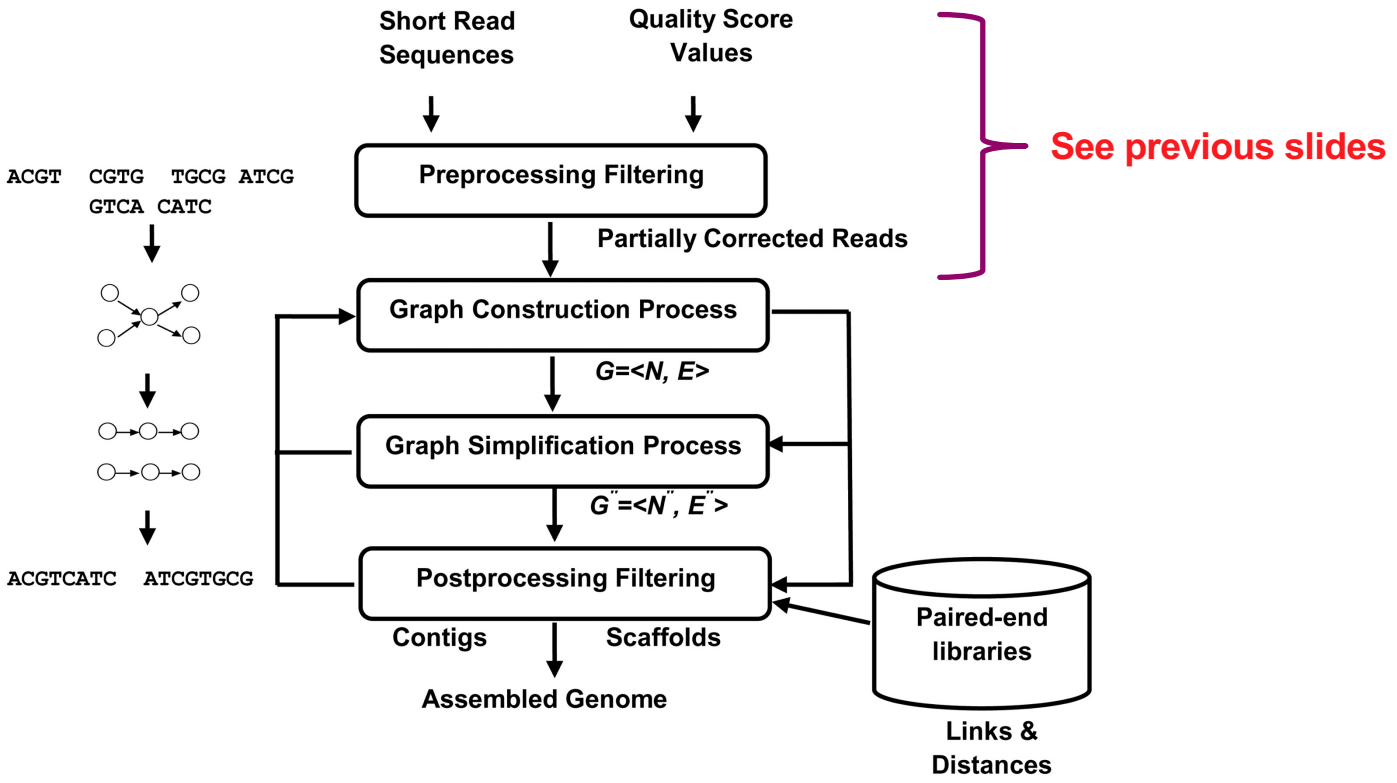
$Q = -10 \log_{10} P$

Probability of error: $P = 10^{-Q/10}$

Probability of truth: $1 - P$

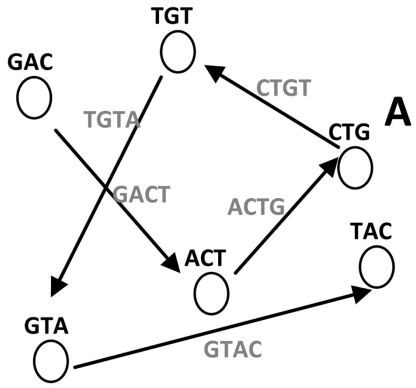
Quality score	% Correct Base
40	99.99
30	99.9
20	99
10	90

Step 2: Assembly and scaffolding: overview



El-Metwally et al., PLoS Comp Biol, 2013

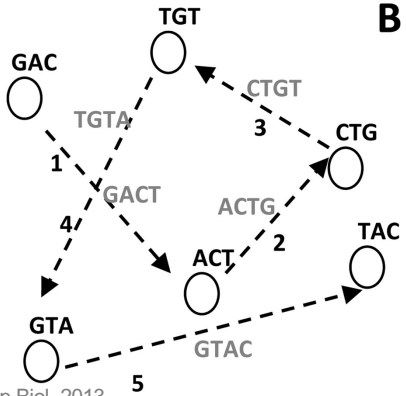
Graph construction: k-mer based assembly



R_1= GACTGTA R_2= ACTGTAC

Set of 3-Kmers of R_1= GAC, ACT,CTG,TGT,GTA

Set of 3-Kmers of R_2= ACT,CTG,TGT,GTA ,TAC



Example of an Eulerian path :

GACT
ACTG
CTGT
TGTA
GTAC

C Assembled Reads :

GACTGTAC

A) <u>k-mer-based</u> graph

Nodes = k-mers

Edges = k-1 overlaps

B) <u>Layout</u> shortest Eulerian path

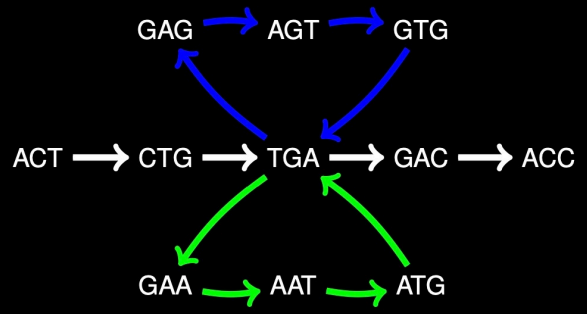
Visit each <u>edge</u> once

C) Combine into <u>consensus</u>

El-Metwally et al., PLoS Comp Biol, 2013

Graph construction: k-mer based assembly

An ambiguous assembly graph



Because of ambiguities and low-coverage regions, a single path is almost never found in theory, and is really never found in practice.

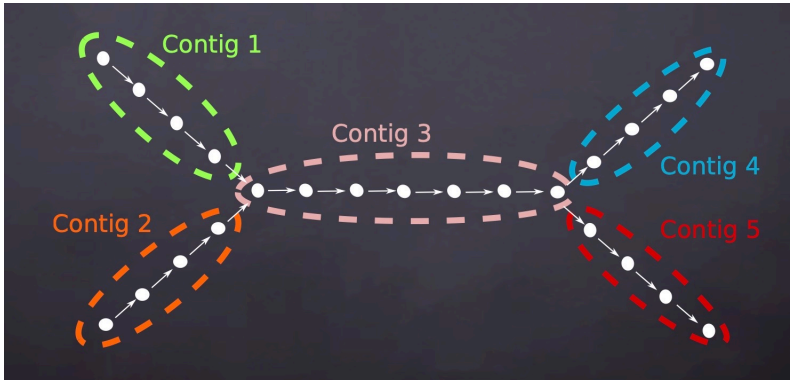
Assembly in practice

Return a **set of paths** covering the graph, such that *all possible assemblies* contain these paths.

Assembly of the above graph

An assembly is the following set of paths:

$$\{\texttt{ACTGA}, \texttt{GACC}, \texttt{GAGTG}, \texttt{GAATG}\}$$



Post processing: scaffolding

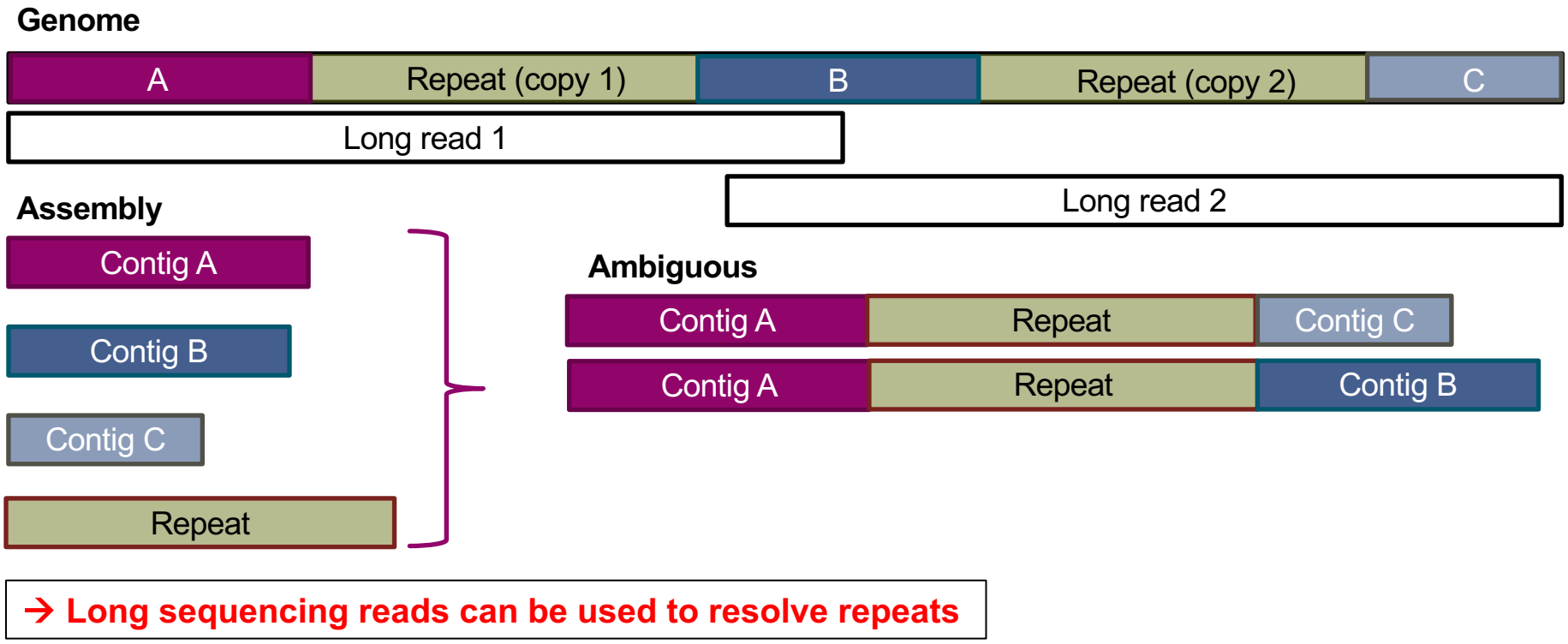
A) Align paired-end reads

B) Orientation of contigs
→ according to read orientation

C) Scaffolding
→ use information on insert size distribution and fill 'gaps' with 'Ns'

El-Metwally et al., PLoS Comp Biol, 2013

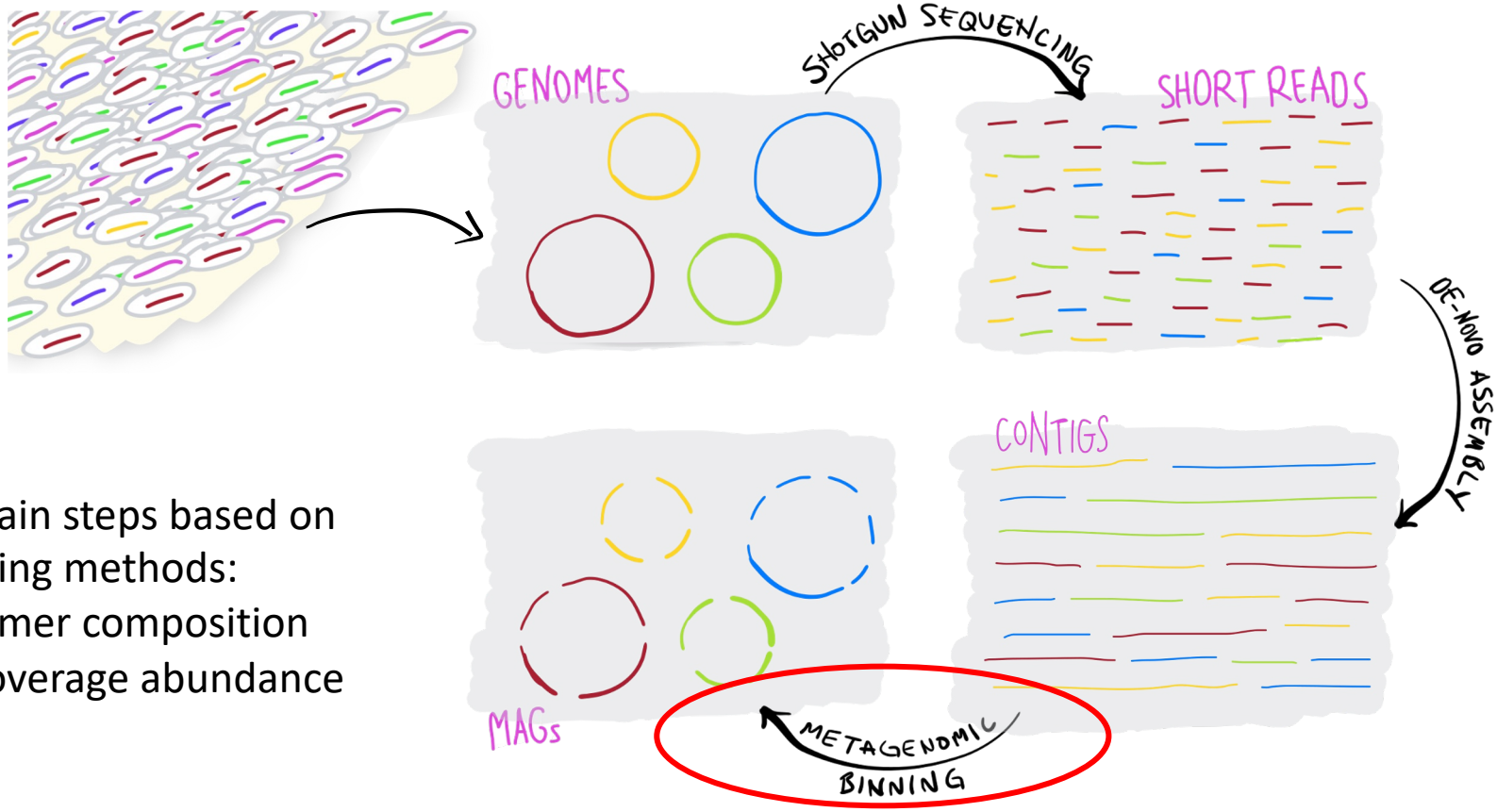
Repetitive sequences in genomes prevent full assembly



→ **Long sequencing reads can be used to resolve repeats**

Current state | | Culture-independent microbiology

Current approaches DO NOT require to isolate organisms or reference genomes:



Two main steps based on clustering methods:

- k-mer composition
- coverage abundance

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1

k=2

GTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	3

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	3

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	3

k=2

GTTTTGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	0	1	0	0	0	0	1	1	1	0	0	1	3

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	1	1	0	0	0	0	1	1	1	0	0	1	3

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	1	1	0	0	0	0	1	1	1	0	0	2	3

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	1	1	0	0	0	1	1	1	1	0	0	2	3

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	2	1	0	0	0	1	1	1	1	0	0	2	3

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
0	0	0	2	1	0	0	0	1	1	1	1	0	0	2	4

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
10	3	2	2	4	2	0	2	2	2	6	0	1	2	1	1

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
10	3	2	2	4	2	0	2	2	2	6	0	1	2	1	1

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC
GAAGCAGAAAAGAAACTCCTTAATCATGCCAAAAC

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
10	3	2	2	4	2	0	2	2	2	6	0	1	2	1	1

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC
GAAGCAGAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	GA	CA	CC	CG	GC	AT	TA

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
1	0	2	2	1	0	0	2	2	2	2	3	1	2	4	10

GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
10	3	2	2	4	2	0	2	2	2	6	0	1	2	1	1

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC +
GAAGCAGAAAAGAAACTCCTTAATCATGCCAAAAC

AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
11	3	4	4	5	2	0	2	2	1

k=2

GTTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y										
Z										
L										
K										
M										

k=2

ACTTCCGCAGTCGGGCATTACGCGTTGTGGAATGA

	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y	4	5	2	4	5	4	4	3	2	1
Z										
L										
K										
M										

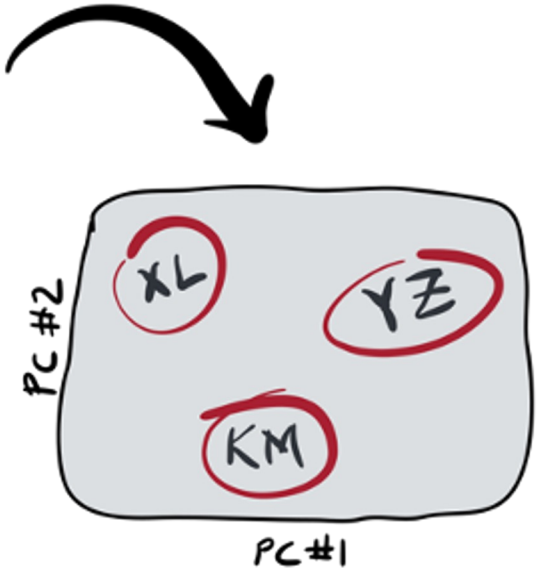
k=2

GGGCCTGCGCCGGTCCAGTCACCCGGCTGCGACCT

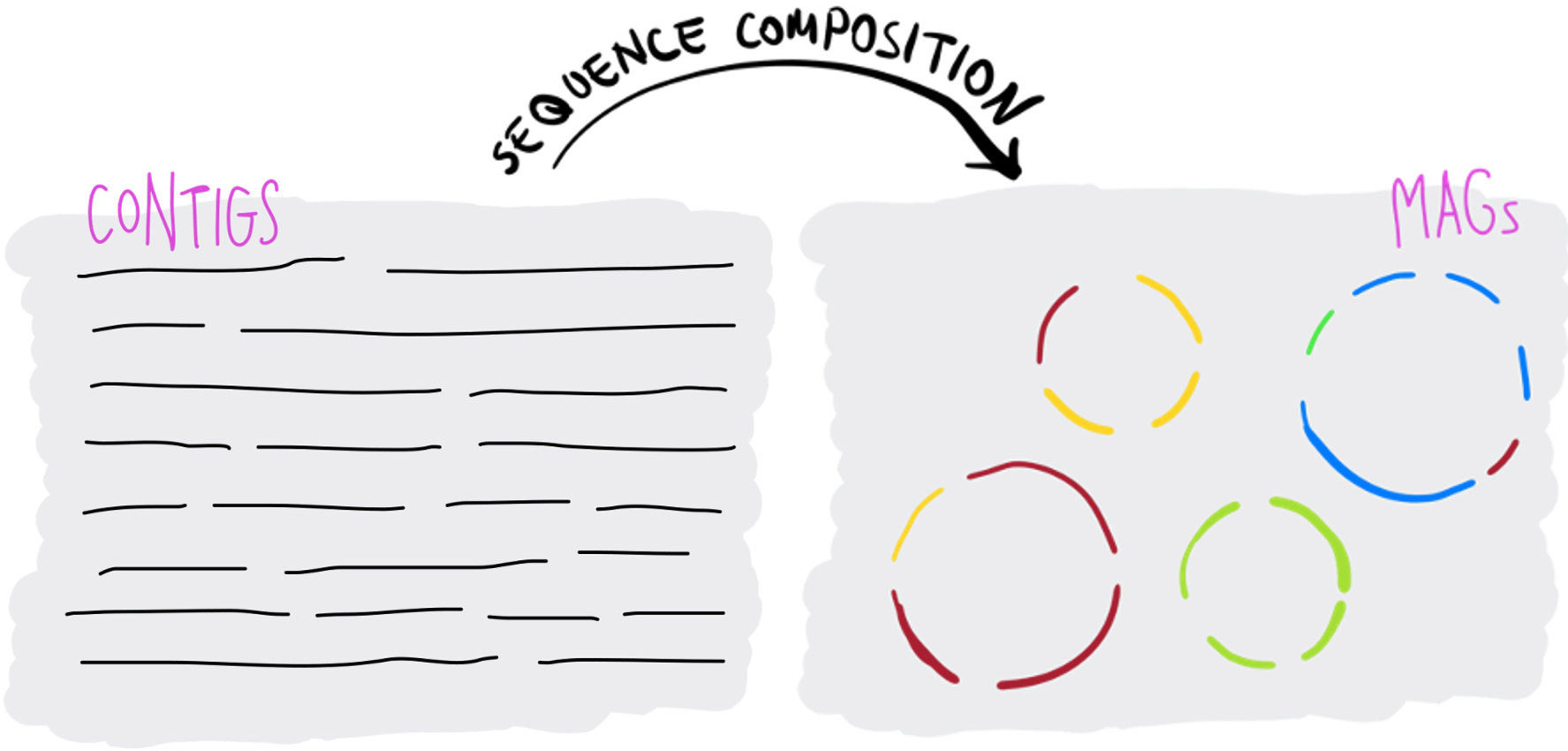
	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y	4	5	2	4	5	4	4	3	2	1
Z	4	5	3	2	4	1	5	5	2	3
L	11	6	3	2	2	3	2	1	1	4
K	1	1	2	2	1	8	9	10	0	0
M	0	4	4	3	4	10	4	5	0	0

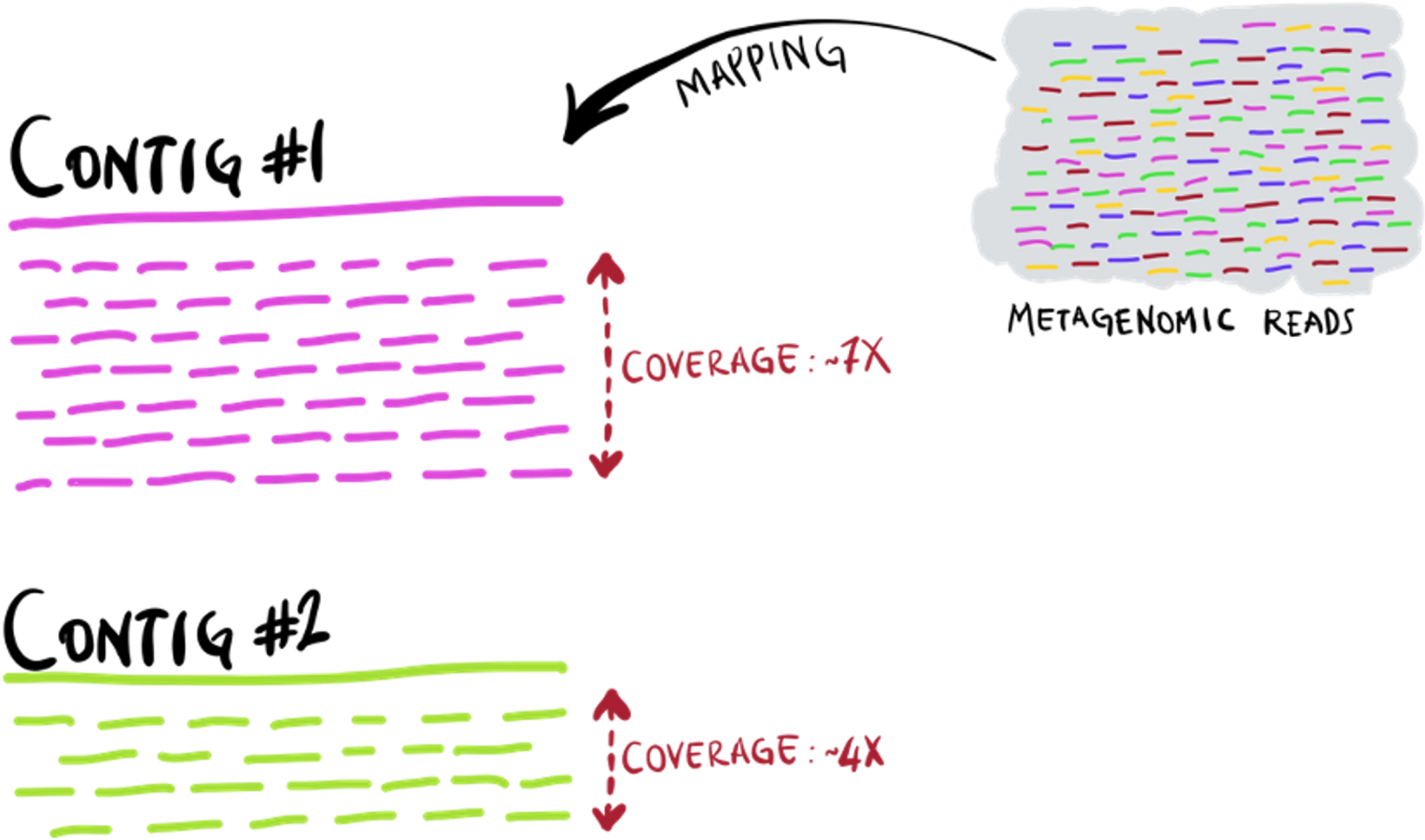
k=2

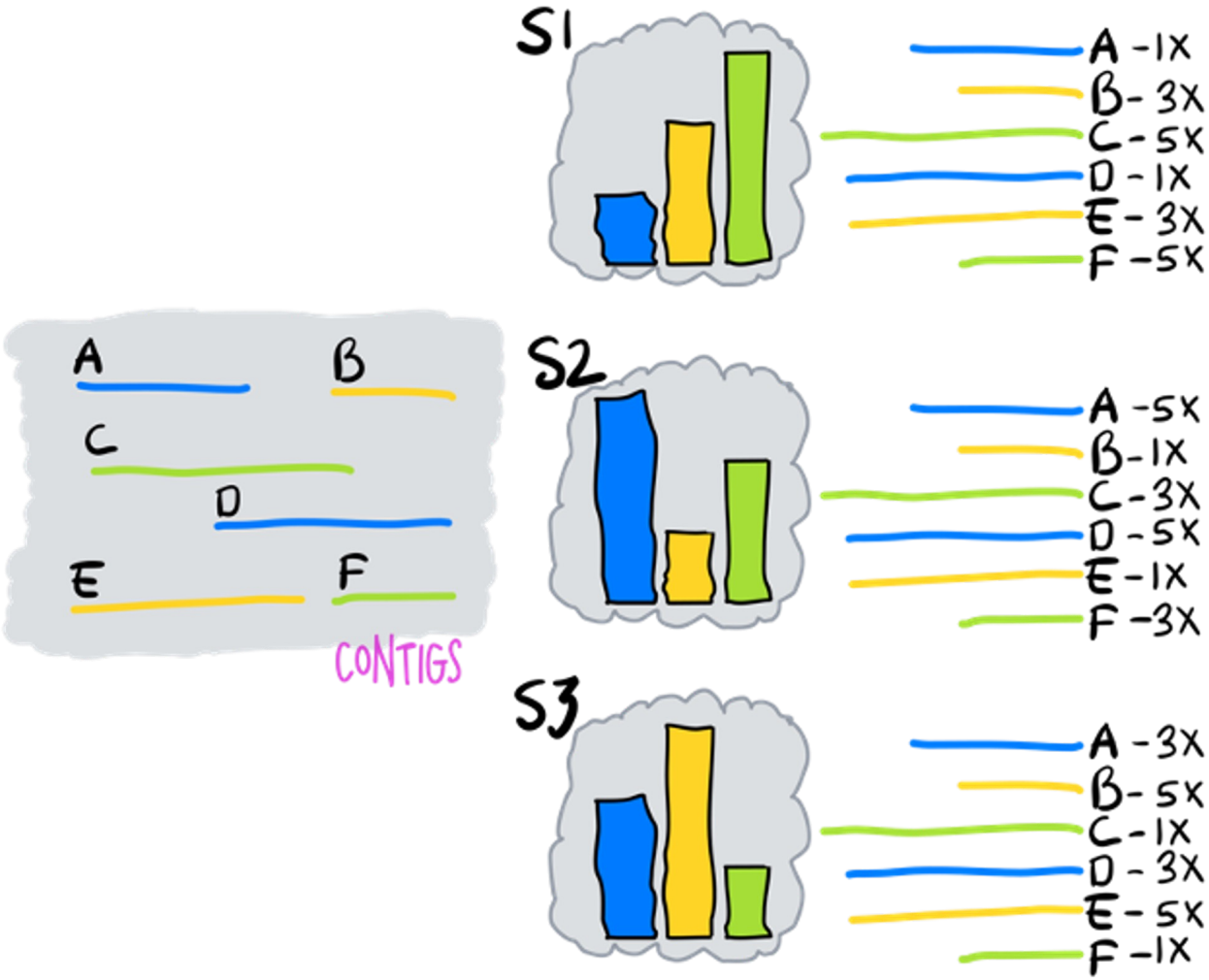
	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y	4	5	2	4	5	4	4	3	2	1
Z	4	5	3	2	4	1	5	5	2	3
L	11	6	3	2	2	3	2	1	1	4
K	1	1	2	2	1	8	9	10	0	0
M	0	4	4	3	4	10	4	5	0	0

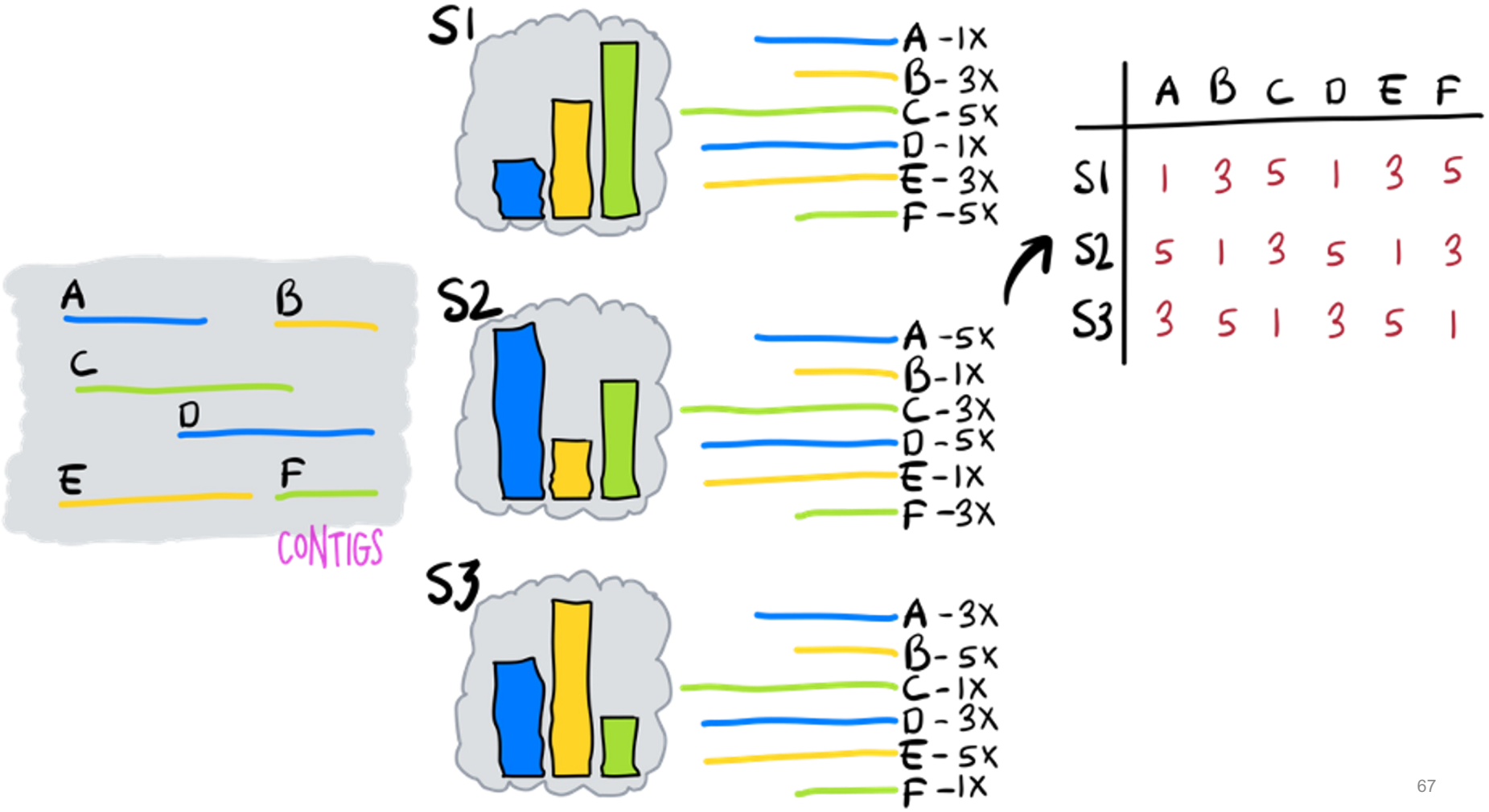


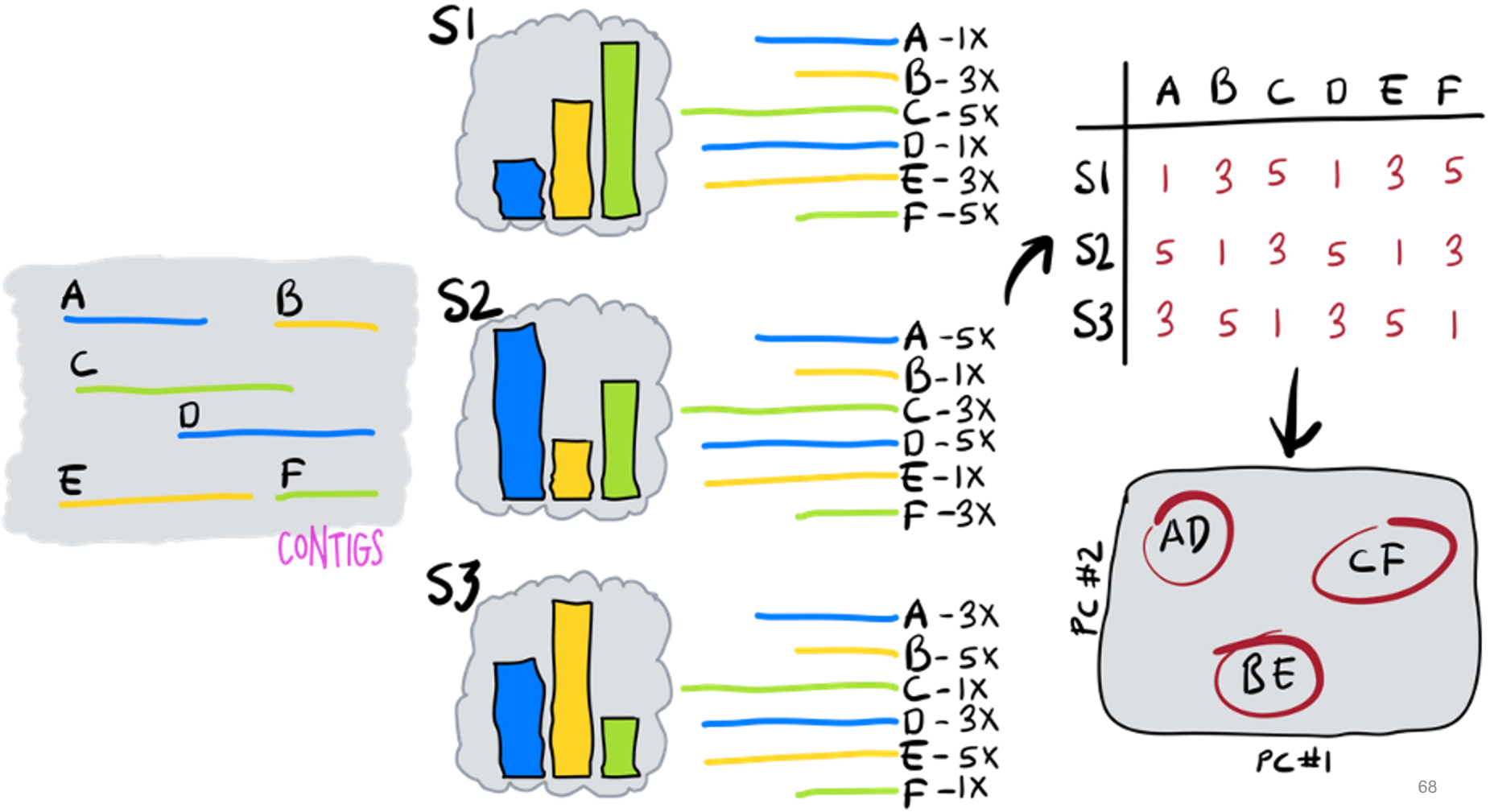
k=2


SEQUENCE COMPOSITION
CONTIGS
MAGs

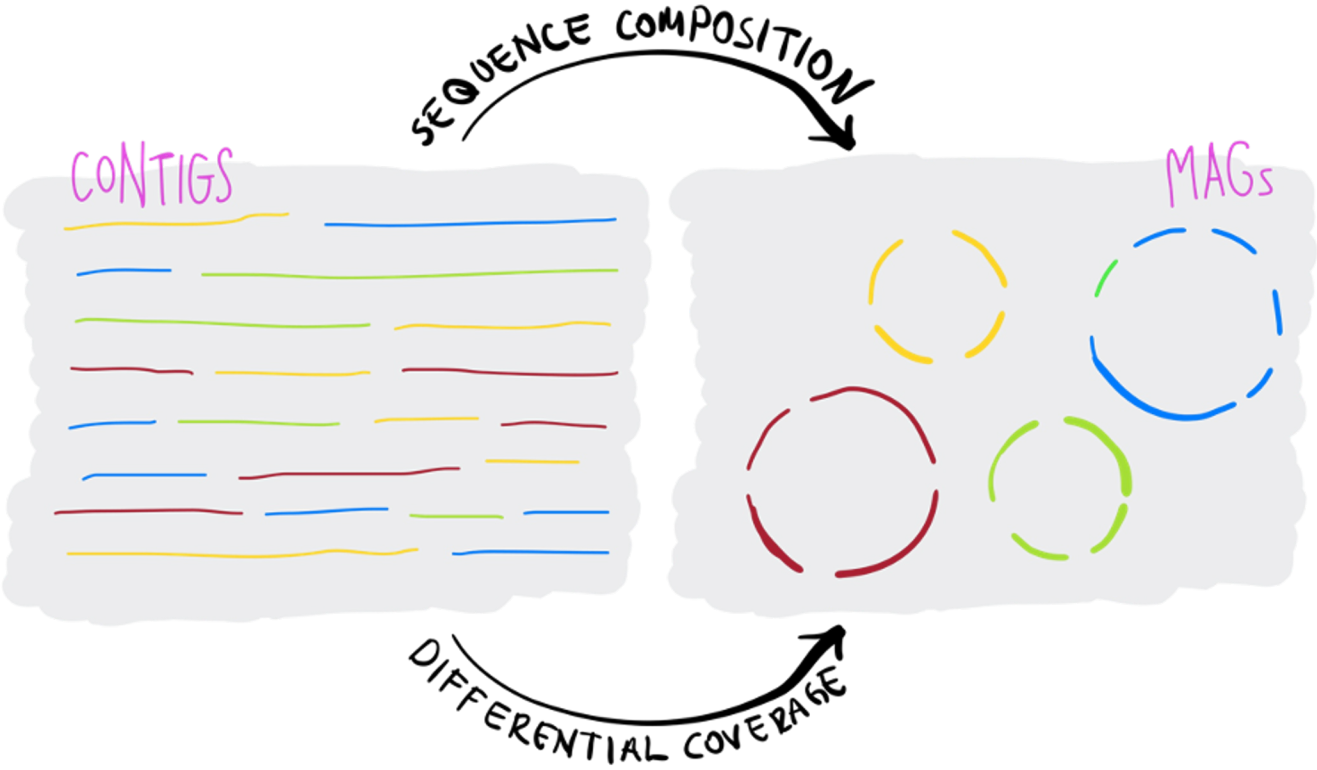

MAPPING
METAGENOMIC READS
CONTIG #1
COVERAGE : ~7X
CONTIG #2
COVERAGE : ~4X


S1
A -1X
B- 3X
C- 5X
D -1X
E- 3X
F -5X
A
B
C
D
E
F
CONTIGS
S2
A -5X
B-1X
C-3X
D -5X
E-1X
F -3X
S3
A -3X
B- 5X
C-1X
D- 3X
E- 5X
F -1X

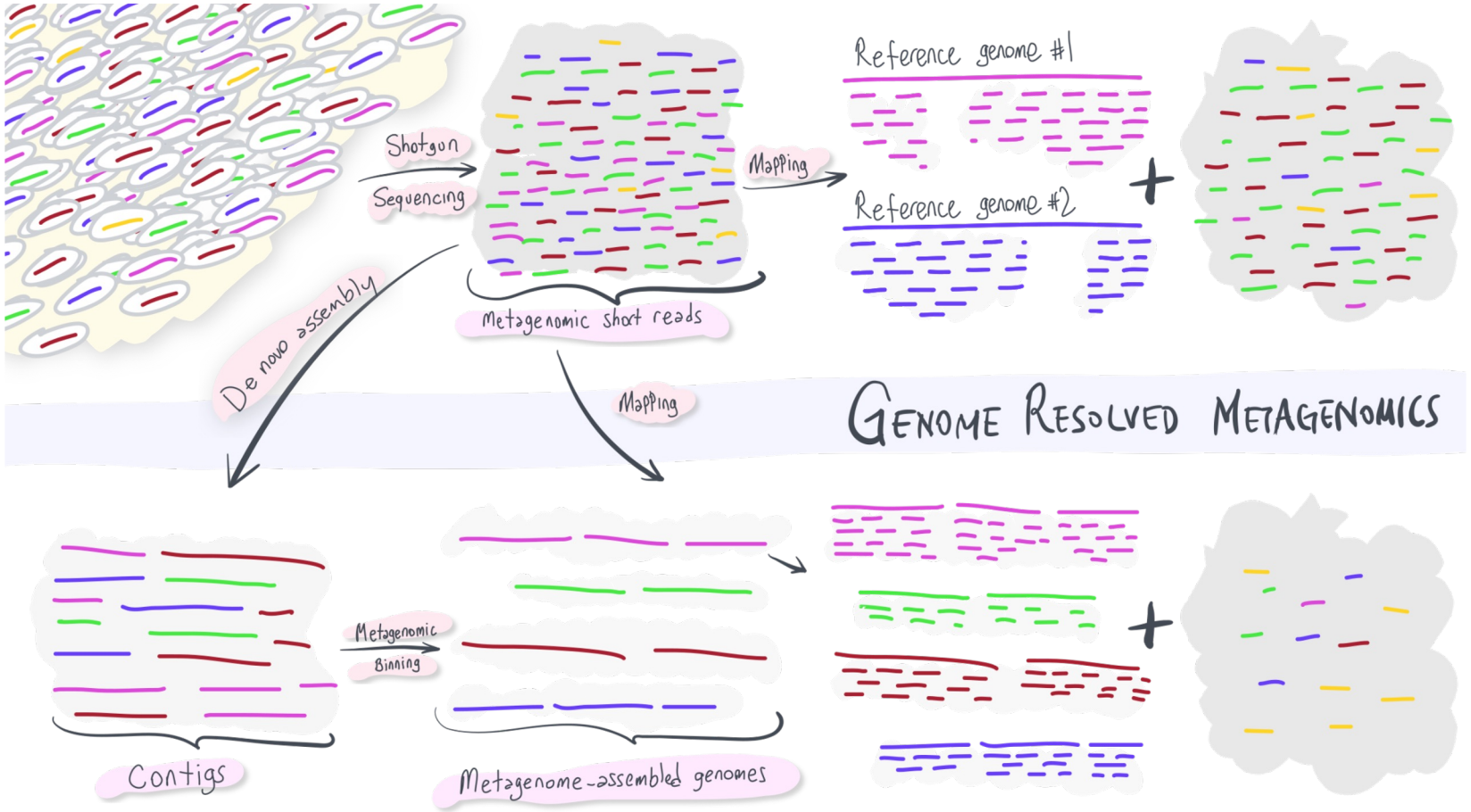

S1
A -1X
B- 3X
C- 5X
D -1X
E- 3X
F -5X
S2
A -5X
B- 1X
C- 3X
D -5X
E -1X
F -3X
S3
A -3X
B- 5X
C- 1X
D- 3X
E- 5X
F -1X
A
B
C
D
E
F
CONTIGS
A B C D E F
S1 1 3 5 1 3 5
S2 5 1 3 5 1 3
S3 3 5 1 3 5 1


A
B
C
D
E
F
CONTIGS
S1
A -1X
B- 3X
C- 5X
D -1X
E- 3X
F -5X
S2
A -5X
B-1X
C-3X
D -5X
E -1X
F -3X
S3
A -3X
B- 5X
C-1X
D- 3X
E- 5X
F -1X
A B C D E F
S1 1 3 5 1 3 5
S2 5 1 3 5 1 3
S3 3 5 1 3 5 1
AD
CF
BE
PC #2
PC #1

MAG reconstruction|| Sequence composition & diff. coverage



Metagenomics | | Summary

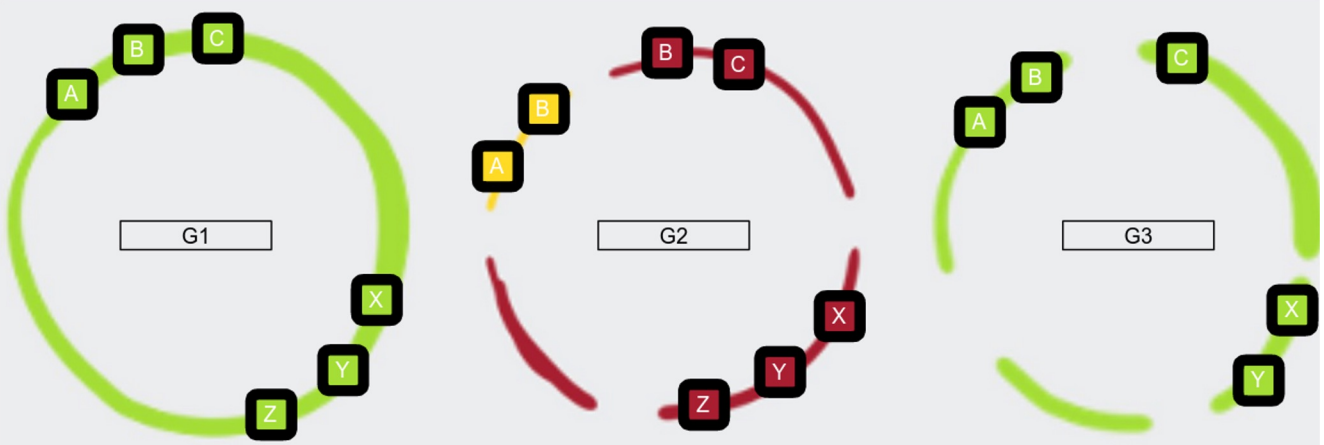


Evaluation of the reconstruction || How complete are our results?

Evaluation of the reconstruction | | How complete & clean are our results?

Universal single-copy marker genes:

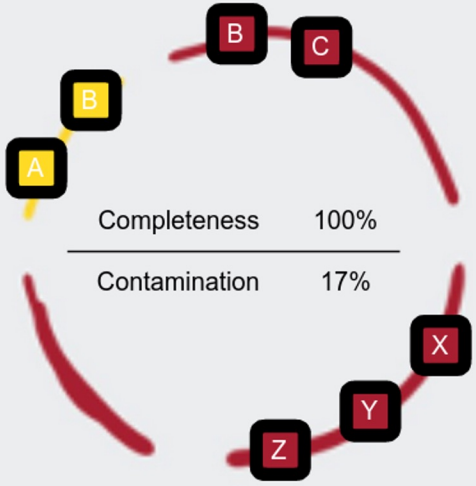
- genes present in every genome
- Between 40 and 120 genes for Bacteria/Archaea depending on cutoffs



	G1	G2	G3
A	1	1	1
B	1	2	1
C	1	1	1
X	1	1	1
Y	1	1	1
Z	1	1	0

Evaluation of the reconstruction | | How complete & clean are our results?

Completeness: % of single-copy marker genes found in the genome
Contamination: % of single-copy marker that are found >1



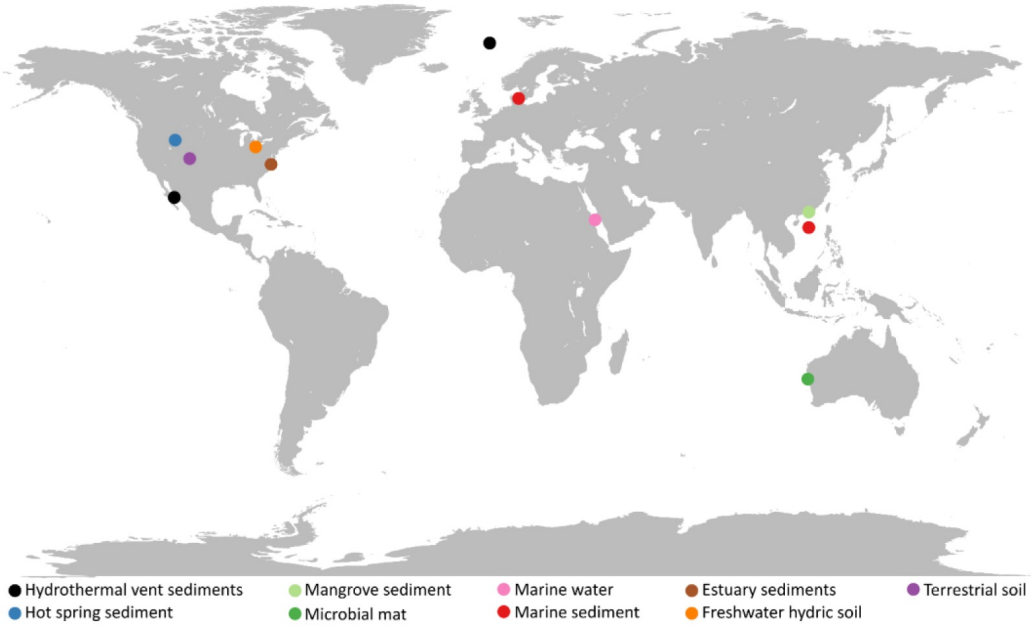
	G1	G2	G3
A	1	1	1
B	1	2	1
C	1	1	1
X	1	1	1
Y	1	1	1
Z	1	1	0

Recent findings in marine microbial genomics

- Discovery of hidden clades
  - Asgard Archaea
  - Candidate phyla radiation (CPR)
  - DPANN
- Discovery of new metabolisms: COMAMMOX
- Re-definition of known metabolisms: the case of nitrogen fixation

Discovery of hidden clades: Asgard archaea

Asgard archaea or Asgardarchaeota is a proposed superphylum consisting of a group of archaea that includes Lokiarchaeota, Thorarchaeota, Odinarchaeota, and Heimdallarchaeota. It appears the eukaryotes have emerged within the Asgard, which supports the two-domain system of classification over the three-domain system.



Global distribution of metagenomic-assembled sequences of Asgard archaea. Asgard metagenomic-assembled genomes from NCBI Assembly and MG-RAST databases were recorded for information related to location and environmental context of sampling (November 2018)

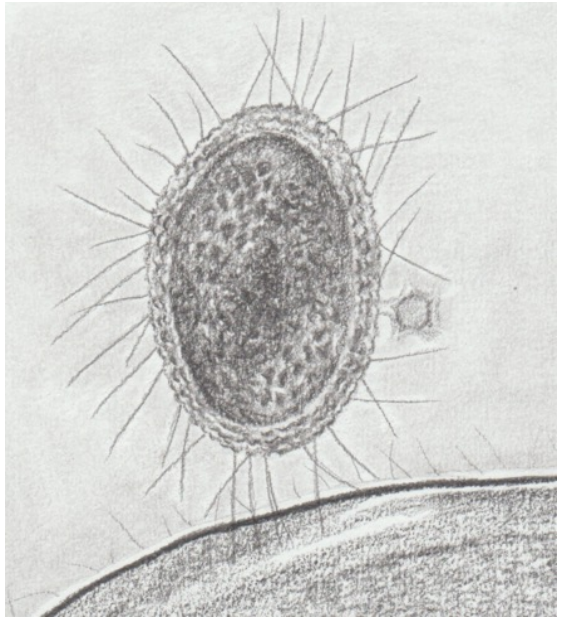
Discovery of hidden clades: Candidate phyla radiation

The **candidate phyla radiation (CPR group)** is a large evolutionary radiation of bacterial lineages whose members are mostly uncultivated and only known from metagenomics and single cell sequencing.

CPR lineages are generally characterized as:

- having **small genomes** and
- **lacking several biosynthetic pathways and ribosomal proteins**.

This has led to the speculation that they are likely obligate symbionts.



Discovery of hidden clades: DPANN

DPANN is a superphylum of Archaea first proposed in 2013. They are known as nanoarchaea or ultra-small archaea due to their smaller size. They exhibit limited metabolic capacities reflected in the fact that many lack central biosynthetic pathways for nucleotides, aminoacids, and lipids. They are mostly anaerobic and cannot be cultivated.

RESEARCH ARTICLE | MICROBIOLOGY |

Insight into the symbiotic lifestyle of DPANN archaea revealed by cultivation and genome analyses

Hiroyuki D. Sakai, Naswandi Nur, Shingo Kato, +5, and Norio Kurosawa Authors Info & Affiliations

Edited by Edward DeLong, Daniel K. Inouye Center for Microbial Oceanography: Research and Education, University of Hawaii at Manoa, Honolulu, HI; received August 26, 2021; accepted November 12, 2021

January 12, 2022 | 119 (3) e2115449119 | https://doi.org/10.1073/pnas.2115449119

Significance

The DPANN superphylum is a grouping of symbiotic microorganisms categorized based on their genomic contents and a few examples of cultivation experiments. Although the genome information of DPANN archaea is increasing year by year, most of them have remained uncultivated, limiting our knowledge of these organisms. Herein, a thermoacidophilic symbiotic archaeon (ARM-1) from the DPANN superphylum was successfully cultivated and characterized. We determined its physiological, morphological, and genomic characteristics in detail and obtained experimental evidence of the symbiotic lifestyle of this archaeon. Notably, ARM-1 is a symbiotic archaeal strain that showed dependence on a range of host species in a laboratory culture. The results significantly contribute to the true understanding of the physiology and ecology of DPANN archaea.

Discovery of new metabolisms: COMAMMOX

Comammox (COMplete AMMonia OXidation) is the name attributed to an organism that can convert ammonia into nitrite and then into nitrate through the process of nitrification.
Nitrification has traditionally thought to be a two-step process, where ammonia-oxidizing bacteria and archaea oxidize ammonia to nitrite and then nitrite-oxidizing bacteria convert to nitrate.
Complete conversion of ammonia into nitrate by a single microorganism was first predicted in 2006. In 2015 the presence of microorganisms that could carry out both conversion processes was discovered within the genus Nitrospira, and the nitrogen cycle was updated.

Published: 26 November 2015

Complete nitrification by *Nitrospira* bacteria

Holger Daims, Elena V. Lebedeva, Petra Pjevac, Ping Han, Craig Herbold, Mads Albertsen, Nico Jehmlich, Marton Palatinszky, Julia Vierheilig, Alexandr Bulaev, Rasmus H. Kirkegaard, Martin von Bergen, Thomas Rattei, Bernd Bendinger, Per H. Nielsen & Michael Wagner

Nature **528**, 504–509 (2015) | Cite this article

Published: 26 November 2015

Complete nitrification by a single microorganism

Maartje A. H. J. van Kessel, Daan R. Speth, Mads Albertsen, Per H. Nielsen, Huub J. M. Op den Camp, Boran Kartal, Mike S. M. Jetten & Sebastian Lücker

Nature **528**, 555–559 (2015) | Cite this article

Re-definition of known metabolisms: the case of nitrogen fixation

Nitrogen fixation is the conversion of molecular nitrogen into ammonia or related nitrogenous compounds, typically in soil or aquatic systems. Biological nitrogen fixation or diazotrophy is an important microbials mediated process that converts dinitrogen gas to ammonia using the nitrogenase protein complex.

Two very recent findings, both based on the reconstruction of marine MAGs, have re-defined our knowledge of diazotrophy:

- The existence and relevance of marine **Heterotrophic Bacterial Diazotrophs** (HBDs)
- The existence of **non-diazotrophic Trichodesmium** (a genus previously though to be strictly diazotrophic).

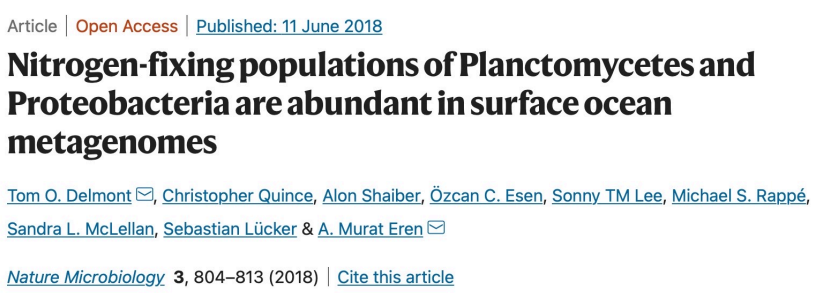
RESEARCH ARTICLE | MICROBIOLOGY

Discovery of nondiazotrophic *Trichodesmium* species abundant and widespread in the open ocean

Tom O. Delmont | Authors Info & Affiliations

Edited by Paul G. Falkowski, Rutgers, The State University of New Jersey, New Brunswick, NJ, and approved September 13, 2021 (received for review July 8, 2021)

November 8, 2021 | 118 (46) e2112355118 | https://doi.org/10.1073/pnas.2112355118



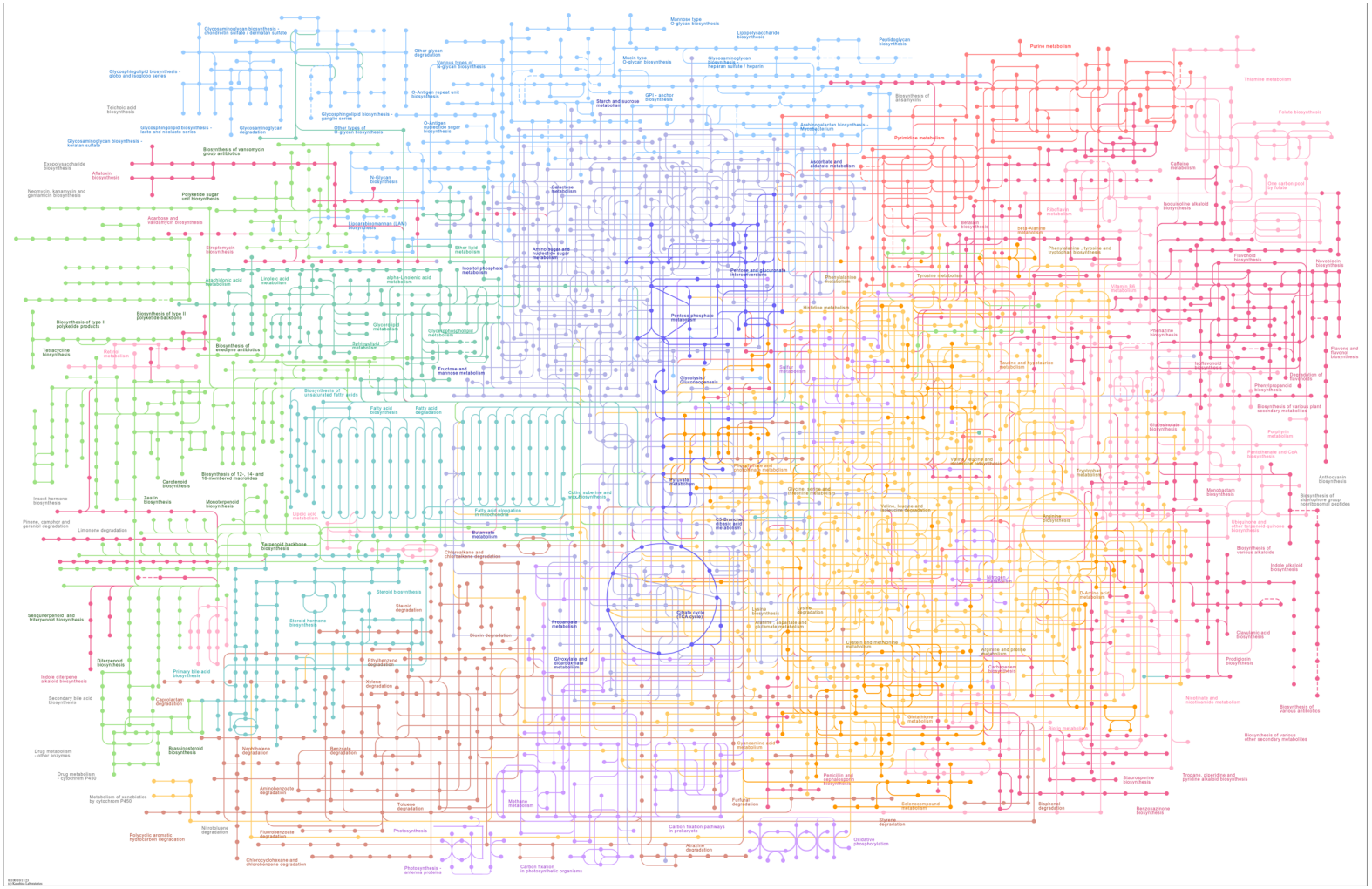
Article | Open Access | Published: 11 June 2018

Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes

Tom O. Delmont, Christopher Quince, Alon Shaiber, Özcan C. Esen, Sonny TM Lee, Michael S. Rappé, Sandra L. McLellan, Sebastian Lücker & A. Murat Eren

Nature Microbiology **3**, 804–813 (2018) | Cite this article

Making sense out of the metabolic complexity



Genome-scale metabolic models

Genome-scale metabolic models (GEMs) provide a representation of an organism's metabolism, encompassing all known metabolic reactions and associated genes.

- A systematic way of describing the metabolism of organisms.
- A way of translating genomic information into functional and physiological information.
- A way to infer the metabolism of organisms that cannot be isolated (and thus subject to experimentation).