Exploring microbiomes with cultivation-independent genome-resolved metagenomics

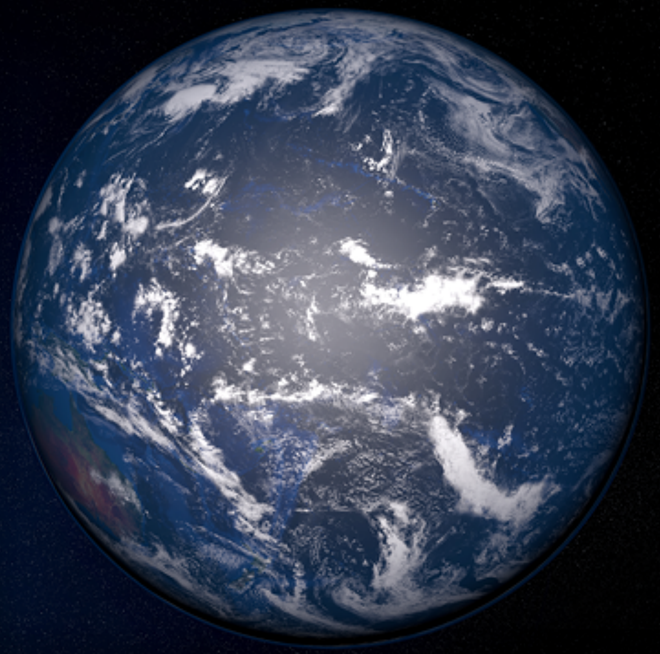
Lucas Paoli, paolil@ethz.ch

Block Course HS 2021

Why explore microbiomes?

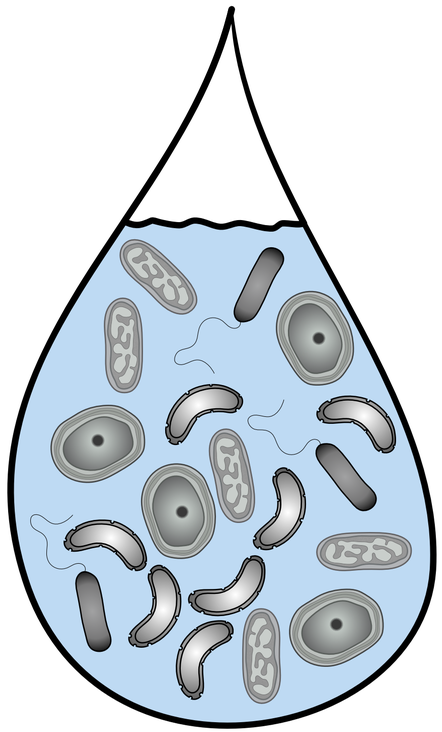
- Ubiquitous across earth’s ecosystems
- Support global food webs
- Underpin biogeochemical cycles
- Determine Host’s health and disease
- …
- Untapped metabolic diversity

Oceans cover >70% of the planet



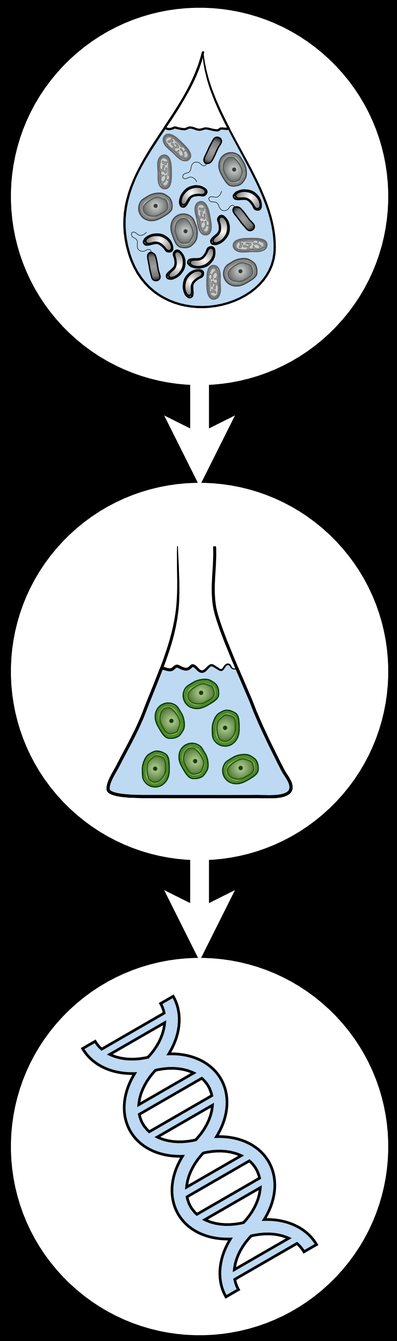
> 500,000 microbial cell per mL

> 50% of the oxygen production



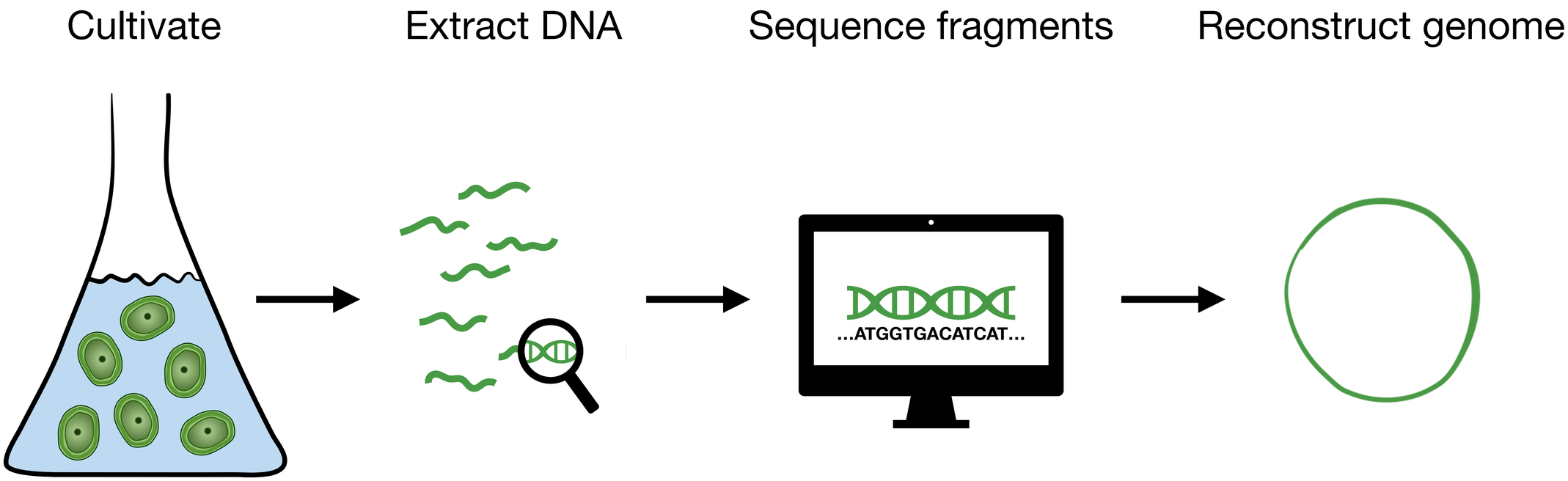
Traditional microbiology

Traditional microbiology
Cultivation-based analysis of microbiomes



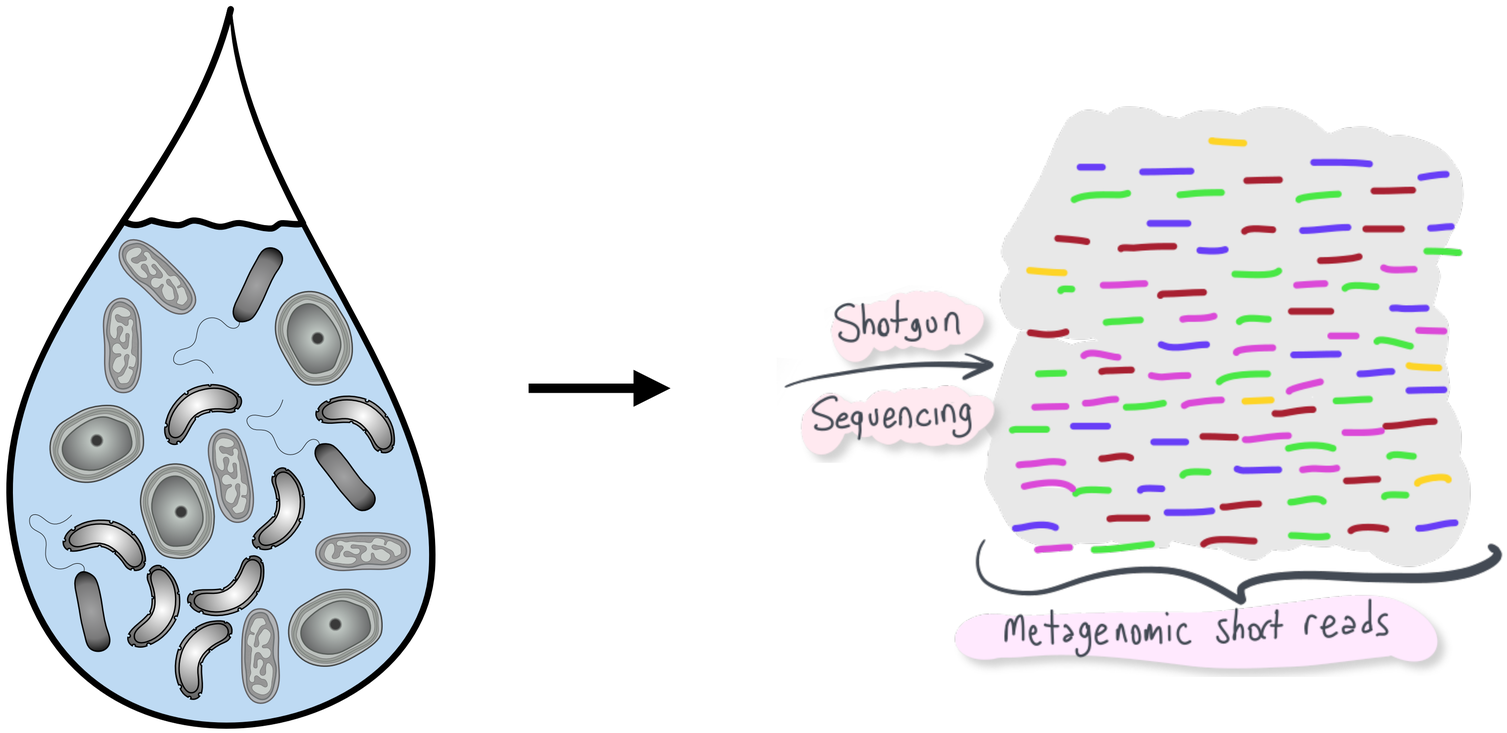
- Isolate
- Cultivate
- Sequence

Image: David Sauveur

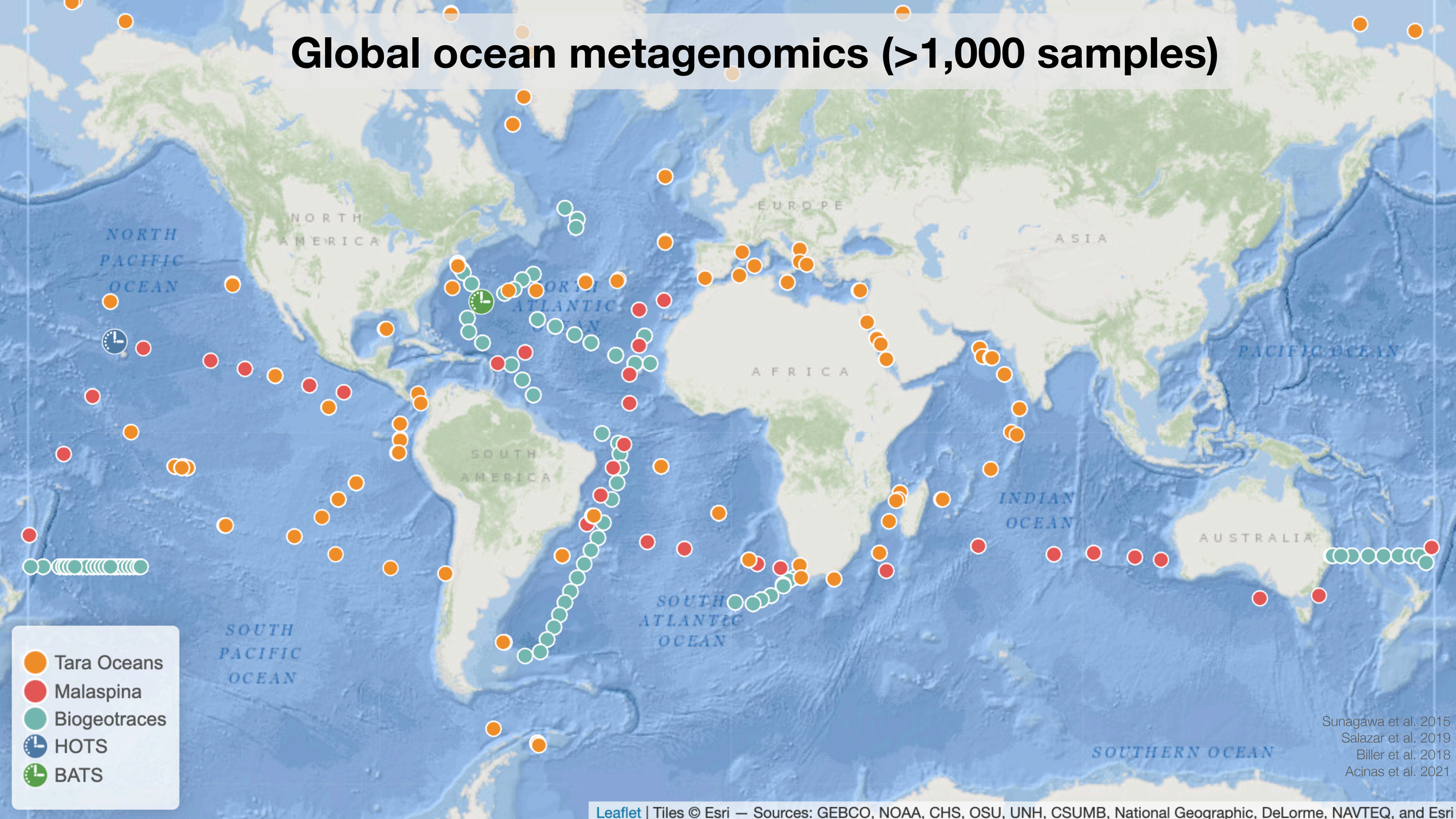

Cultivate
Extract DNA
Sequence fragments
Reconstruct genome
...ATGGTGACATCAT...

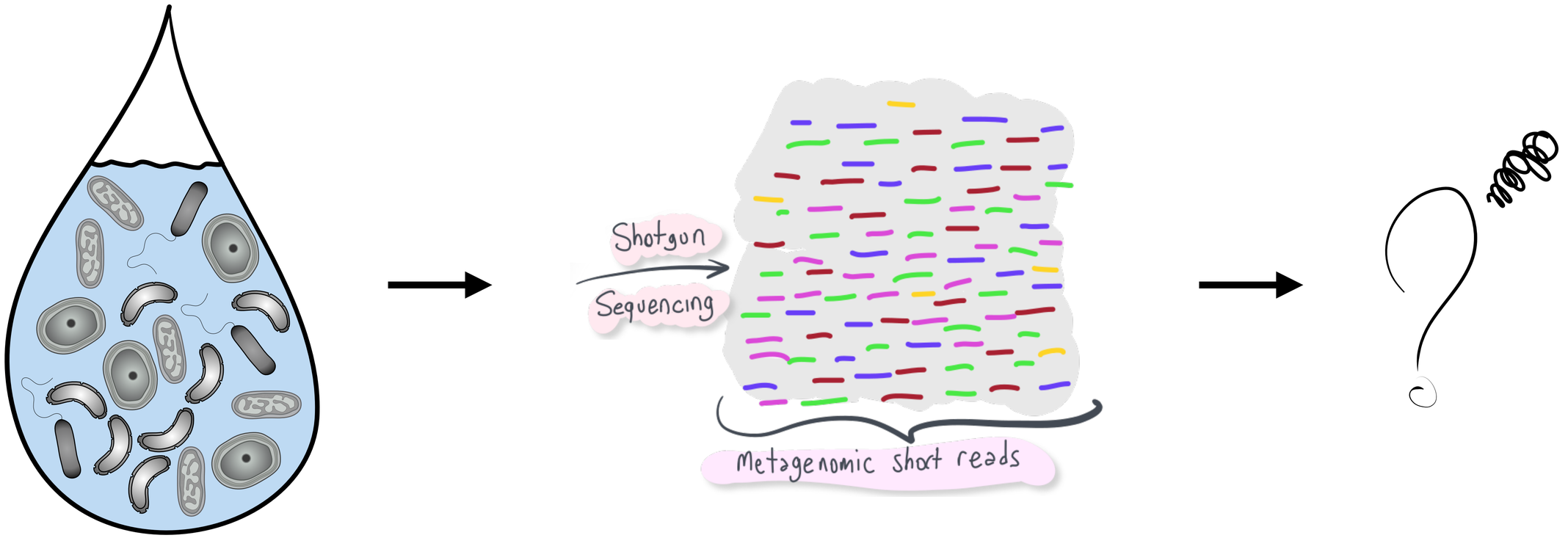
Metagenomics

Accessing the genomic content of microbiomes


Shotgun
Sequencing
Metagenomic short reads

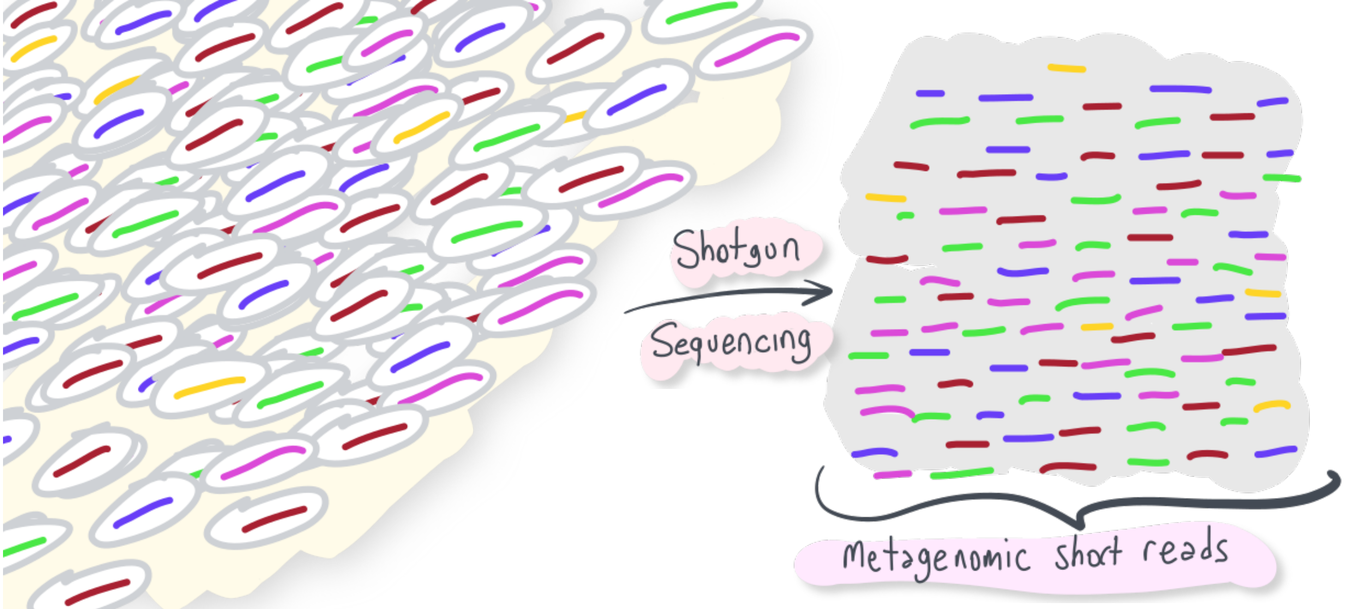
Global ocean metagenomics (>1,000 samples)



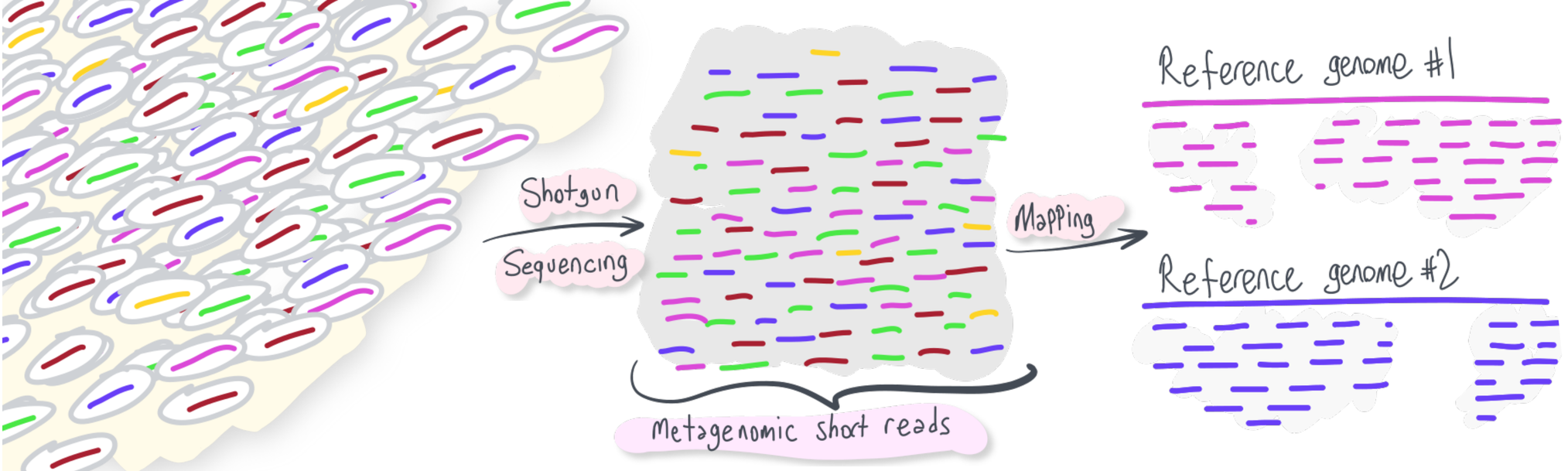

Shotgun
Sequencing
Metagenomic short reads

Genome-resolved Metagenomics

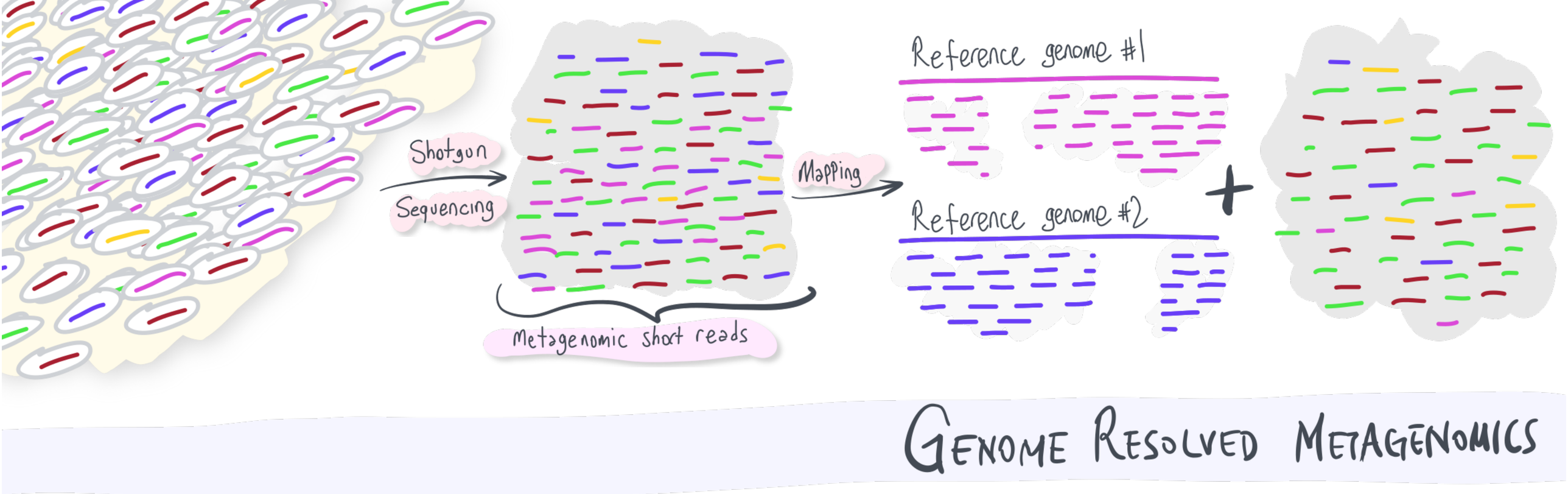
Contextualizing the genomic content of microbiomes



Genome Resolved Metagenomics



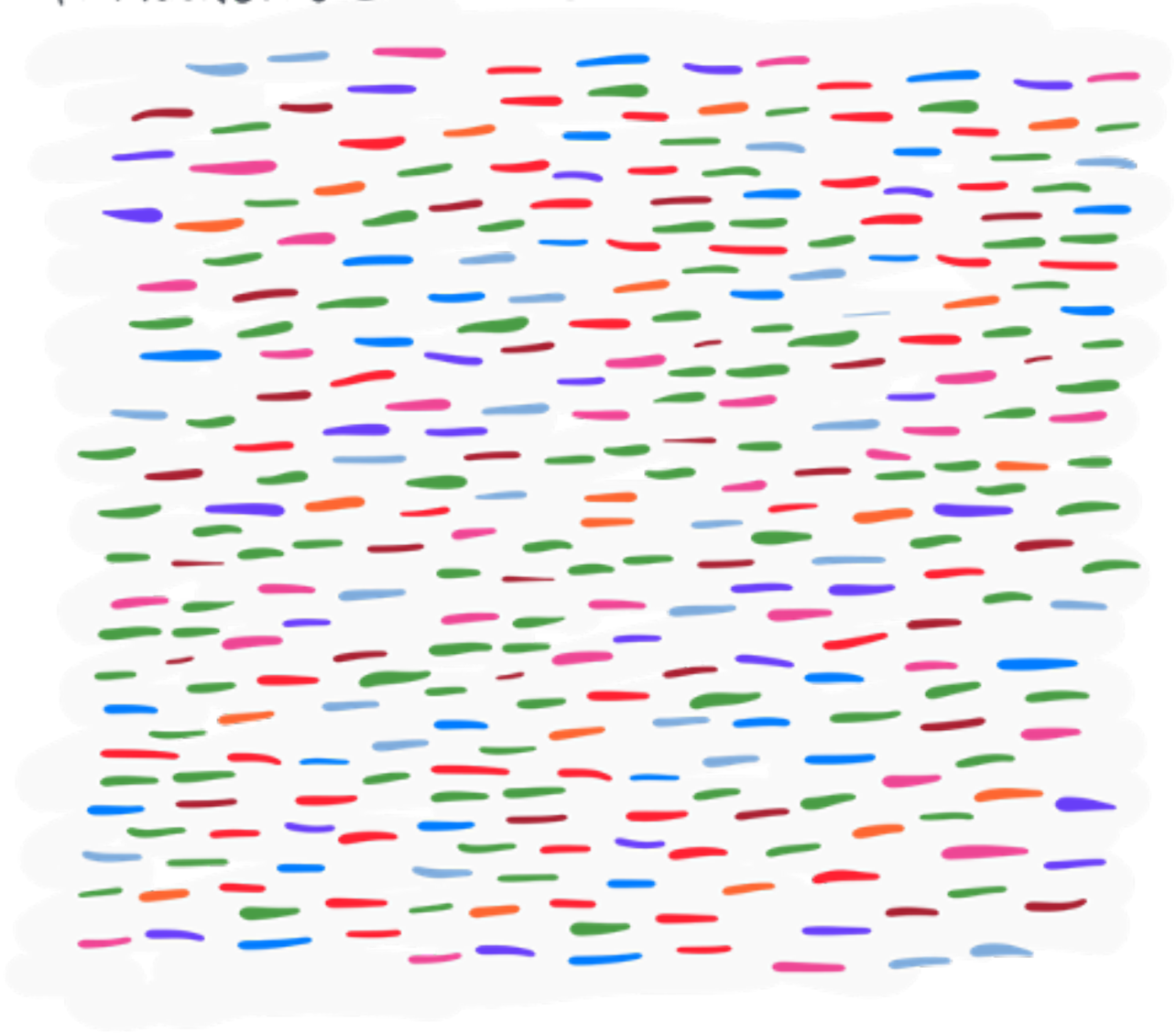
Genome Resolved Metagenomics



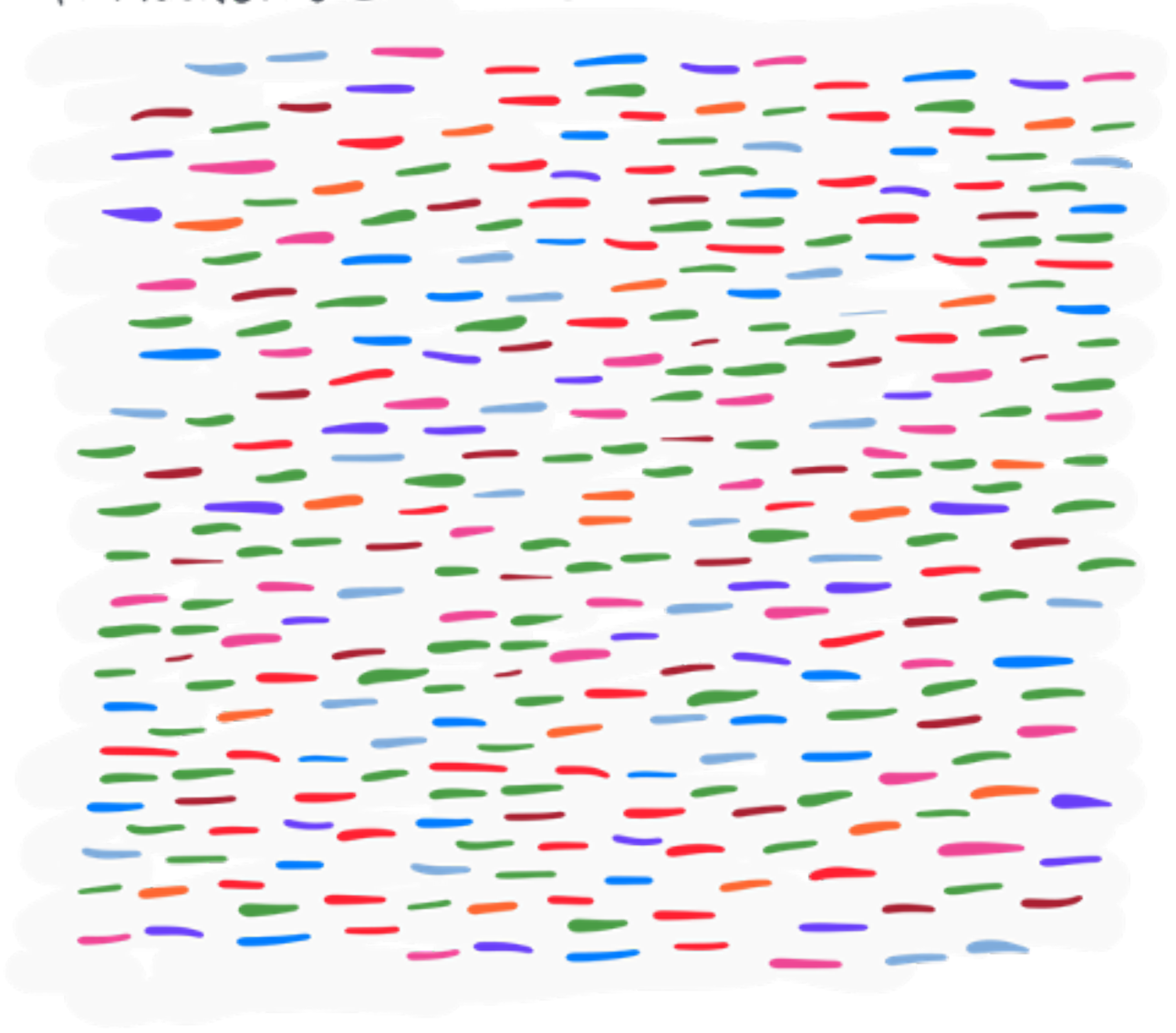
Genome Resolved Metagenomics

Mapping a metagenome
Alignment of reads to a reference

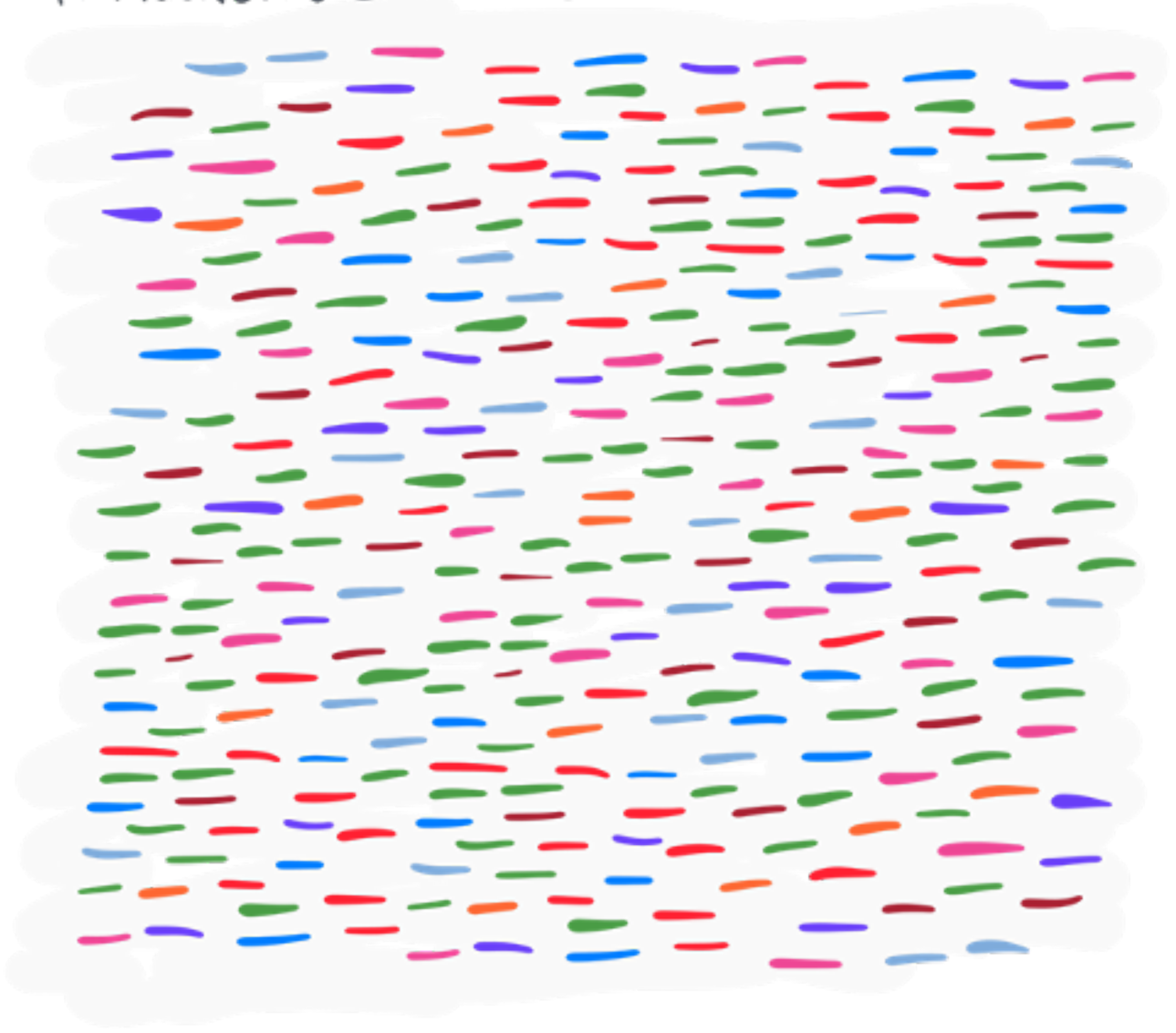
METAGENOMIC SHORT READS



METAGENOMIC SHORT READS

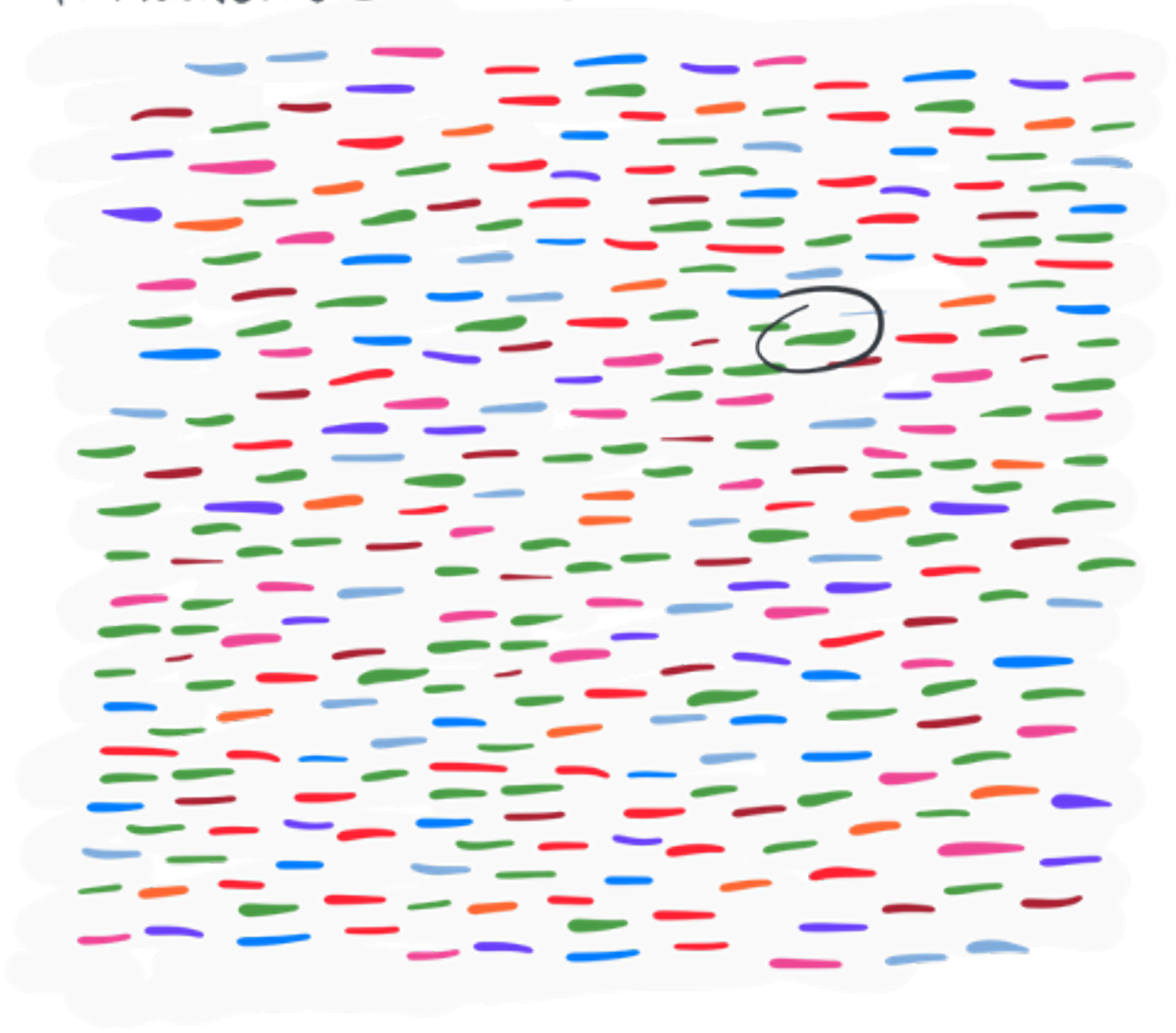


METAGENOMIC SHORT READS

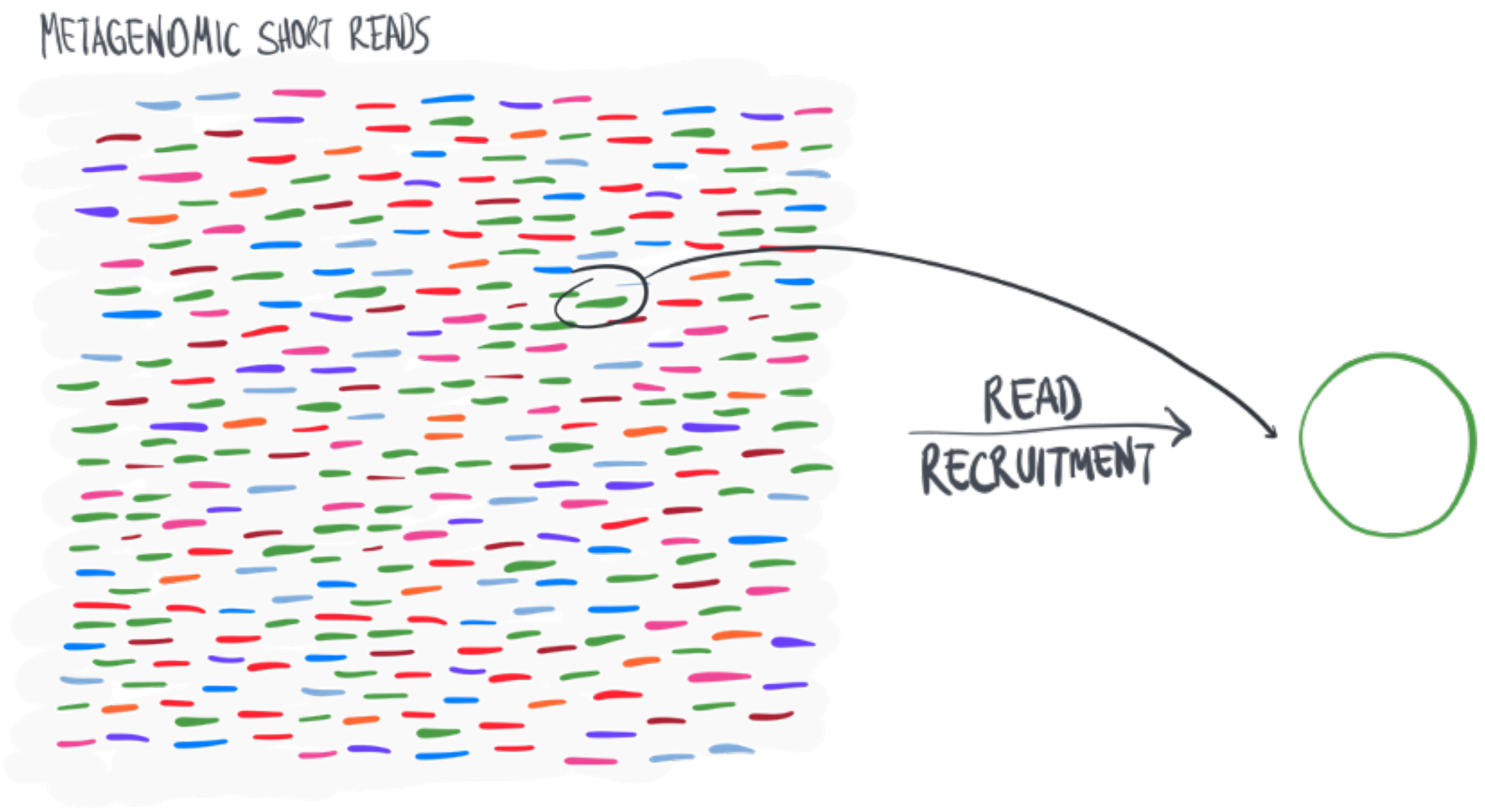


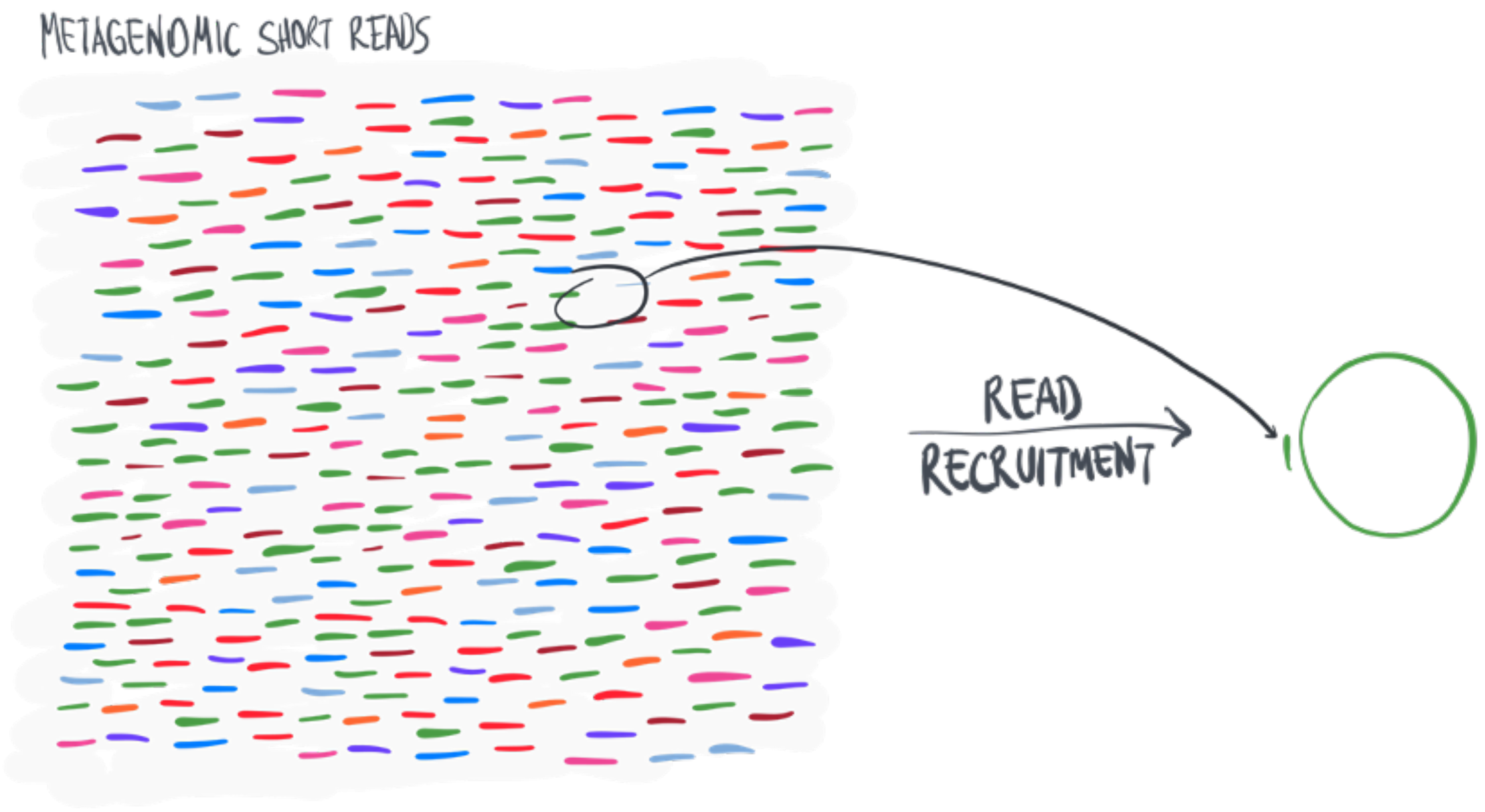
READ RECRUITMENT

METAGENOMIC SHORT READS

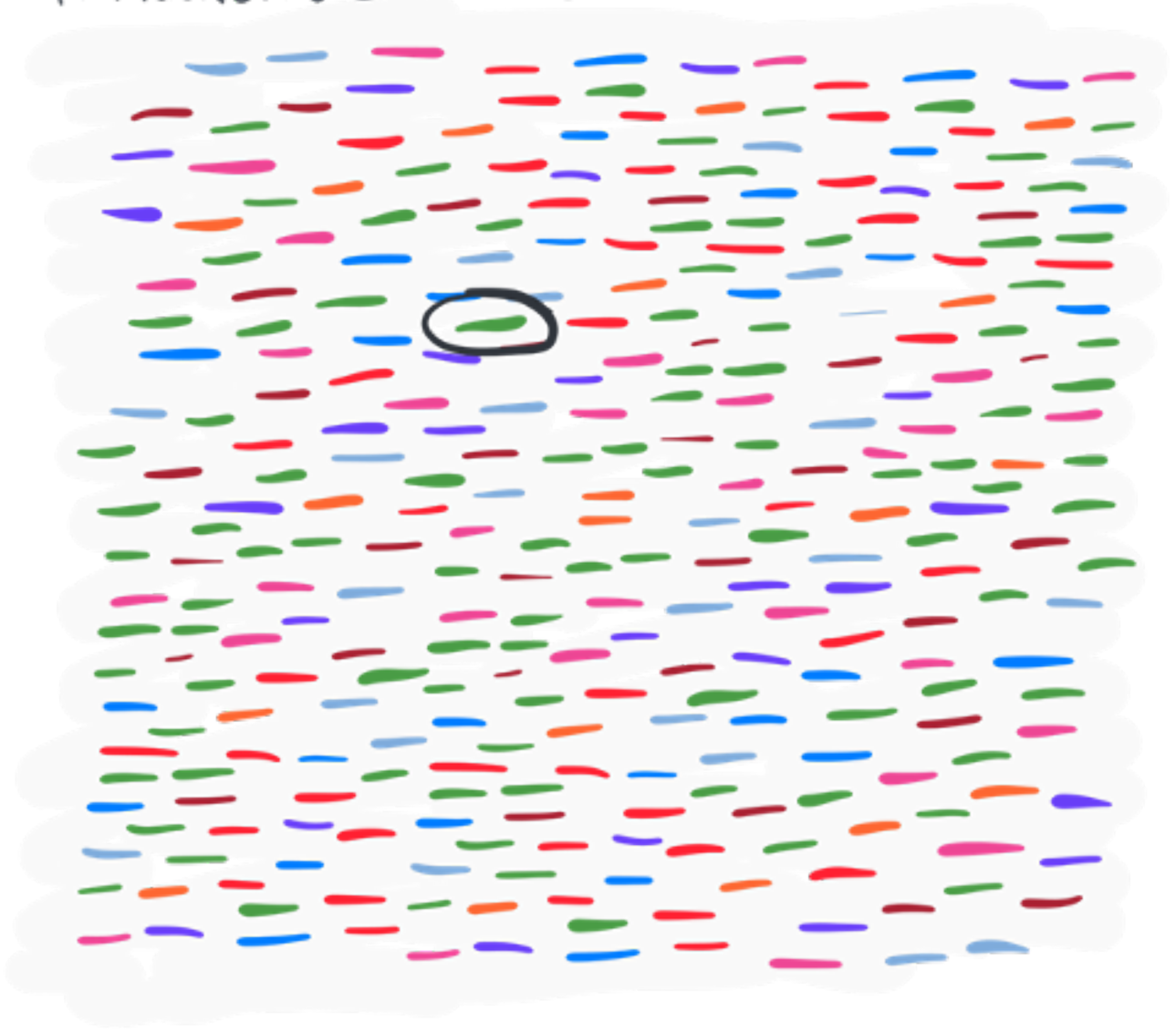


READ RECRUITMENT

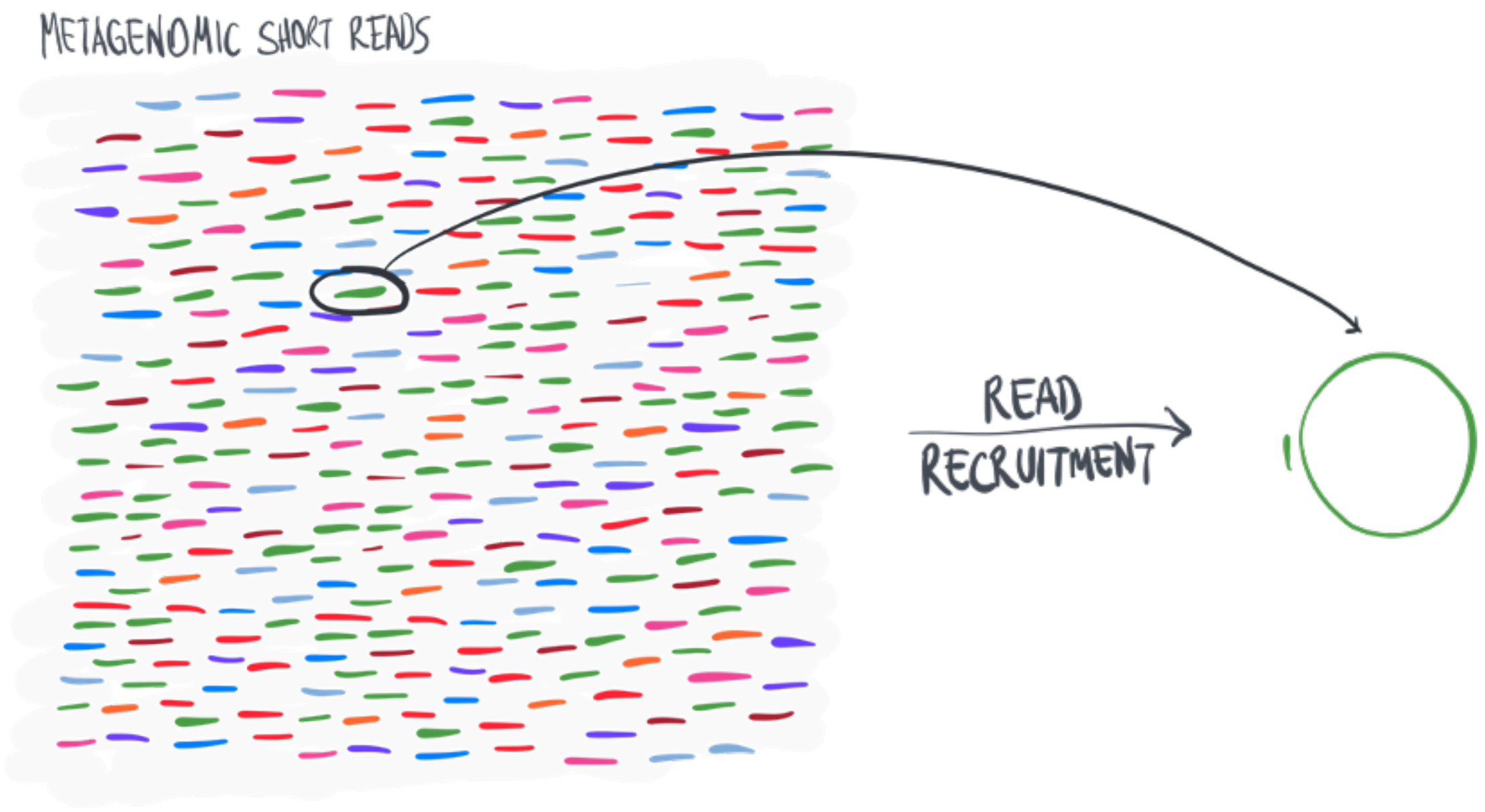

METAGENOMIC SHORT READS
READ
RECRUITMENT

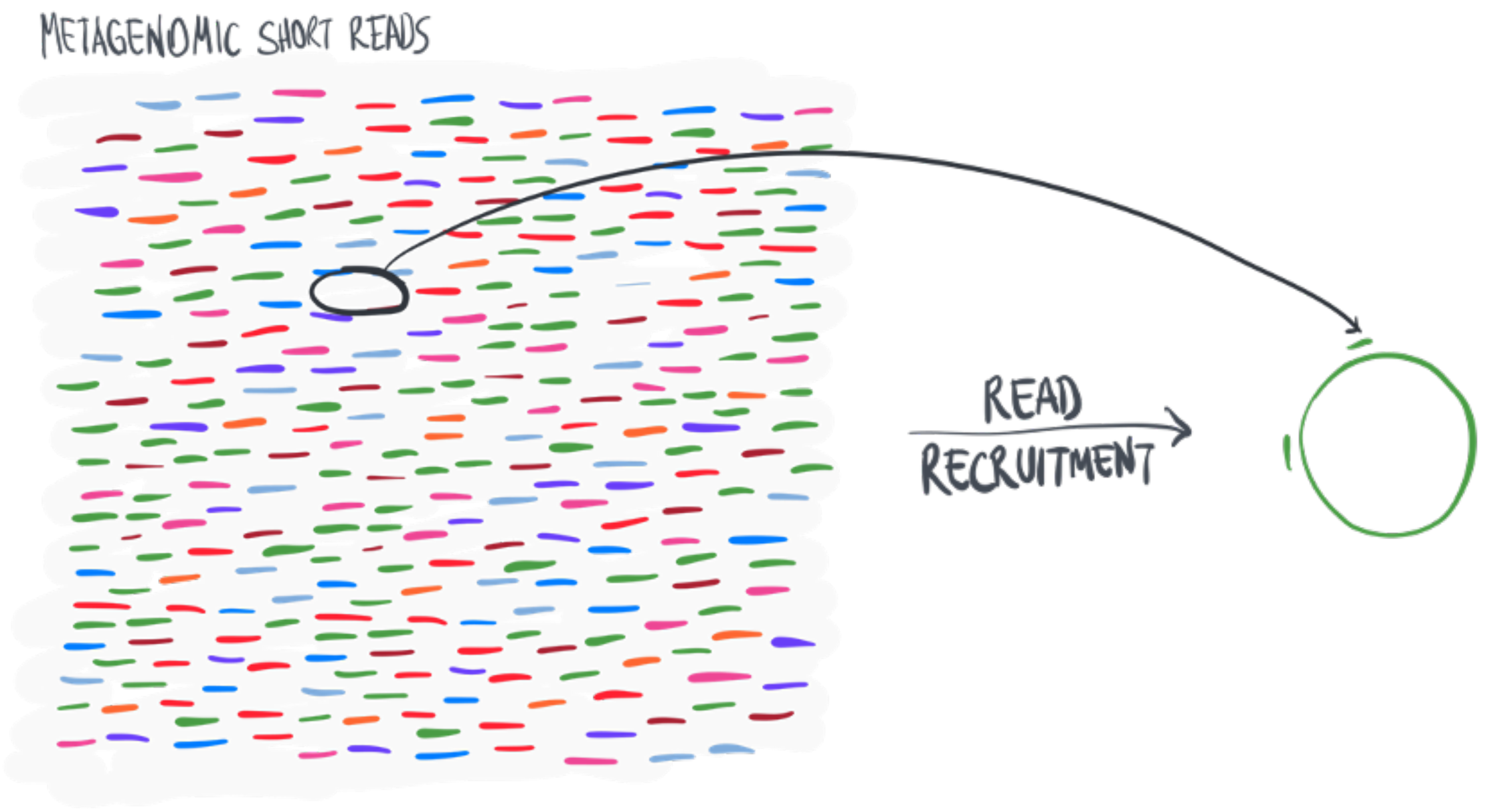

METAGENOMIC SHORT READS
READ
RECRUITMENT

METAGENOMIC SHORT READS

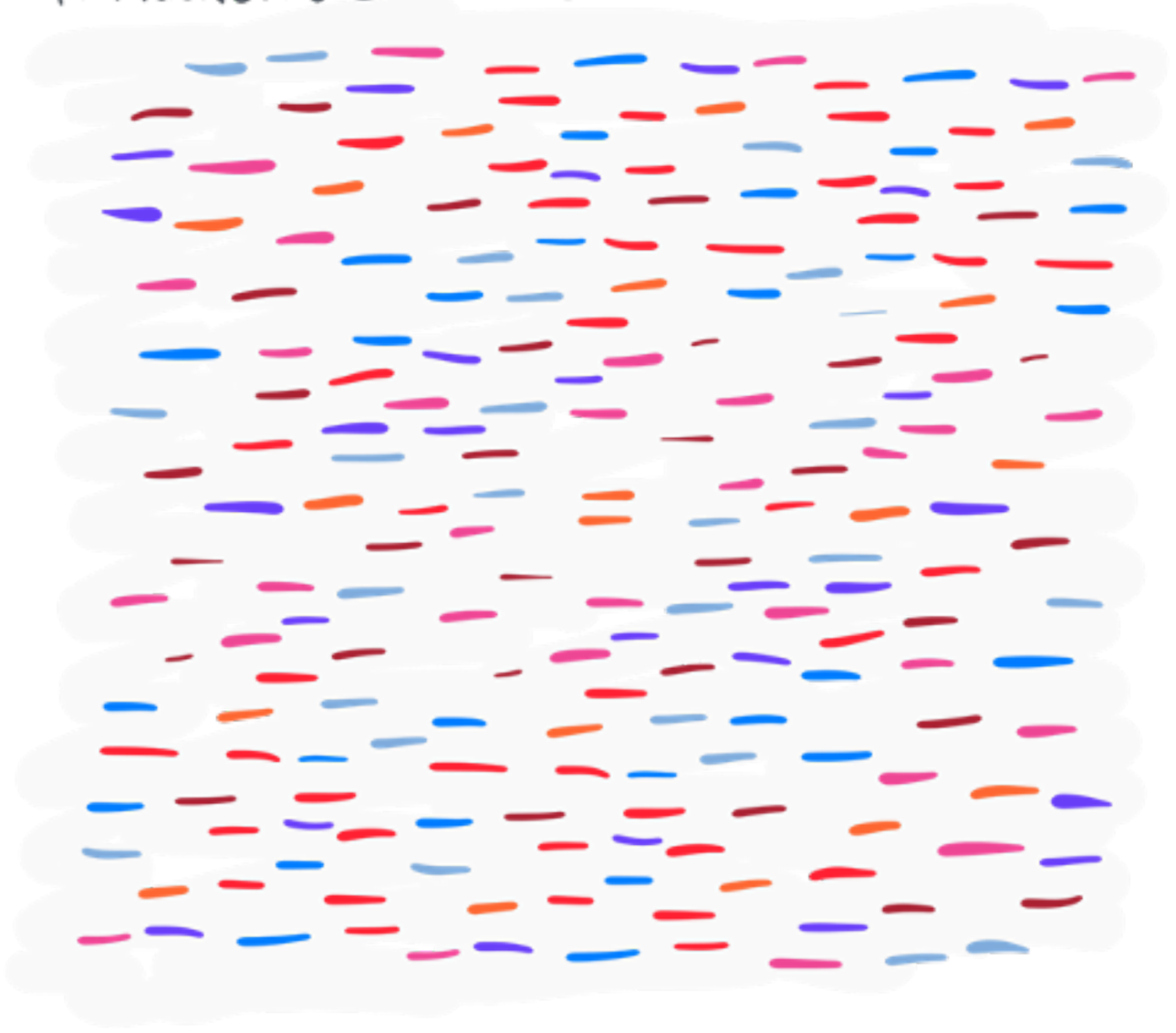


READ RECRUITMENT


METAGENOMIC SHORT READS
READ
RECRUITMENT


METAGENOMIC SHORT READS
READ
RECRUITMENT

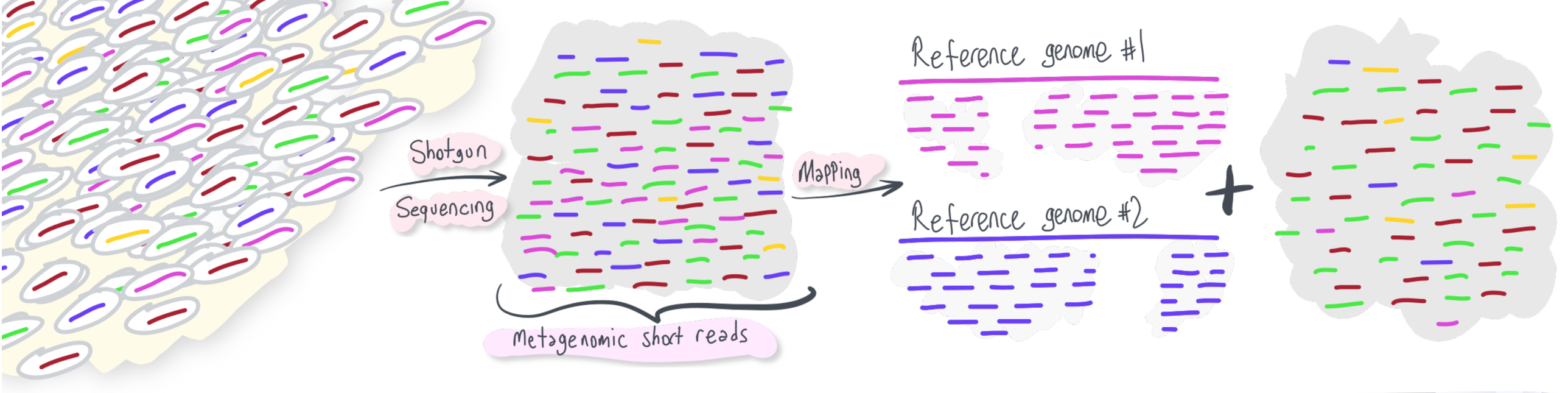
METAGENOMIC SHORT READS



READ RECRUITMENT

Mapping rates

How much can we put in the reference context?



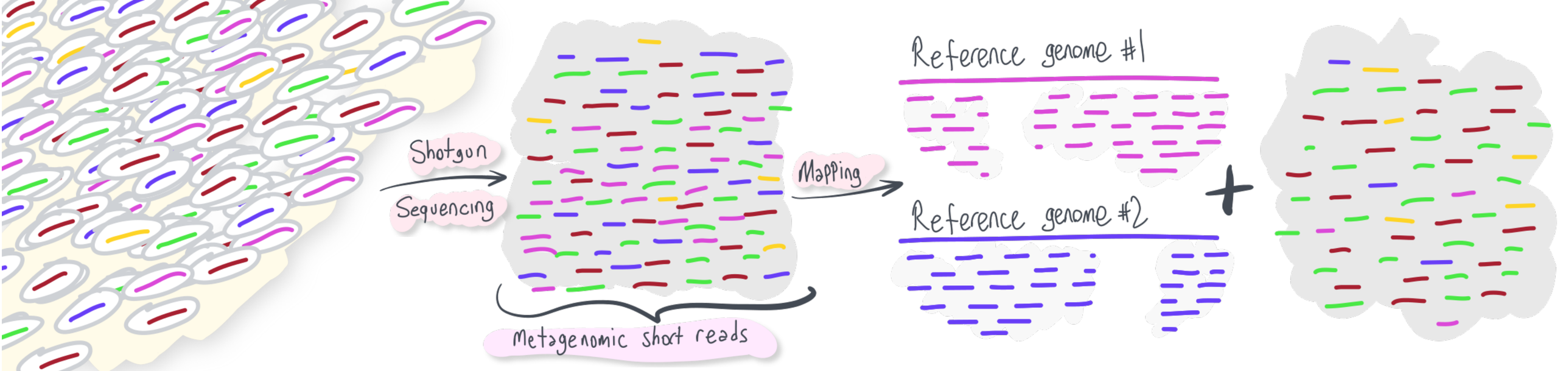
Genome Resolved Metagenomics

Reference genomes

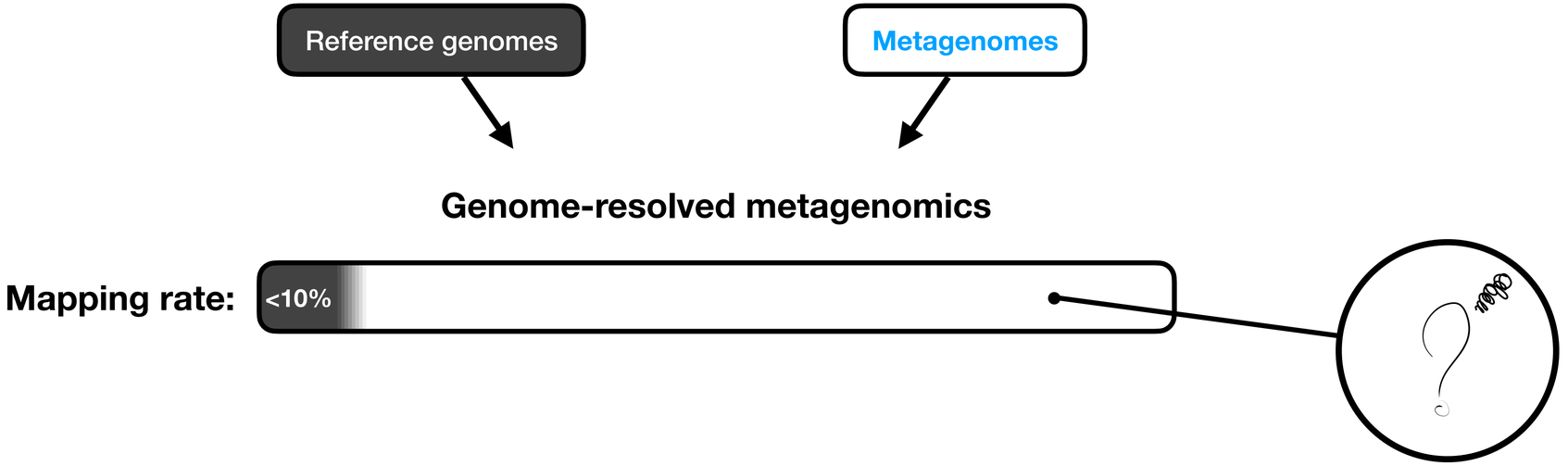
Metagenomes

Genome-resolved metagenomics

Mapping rate:

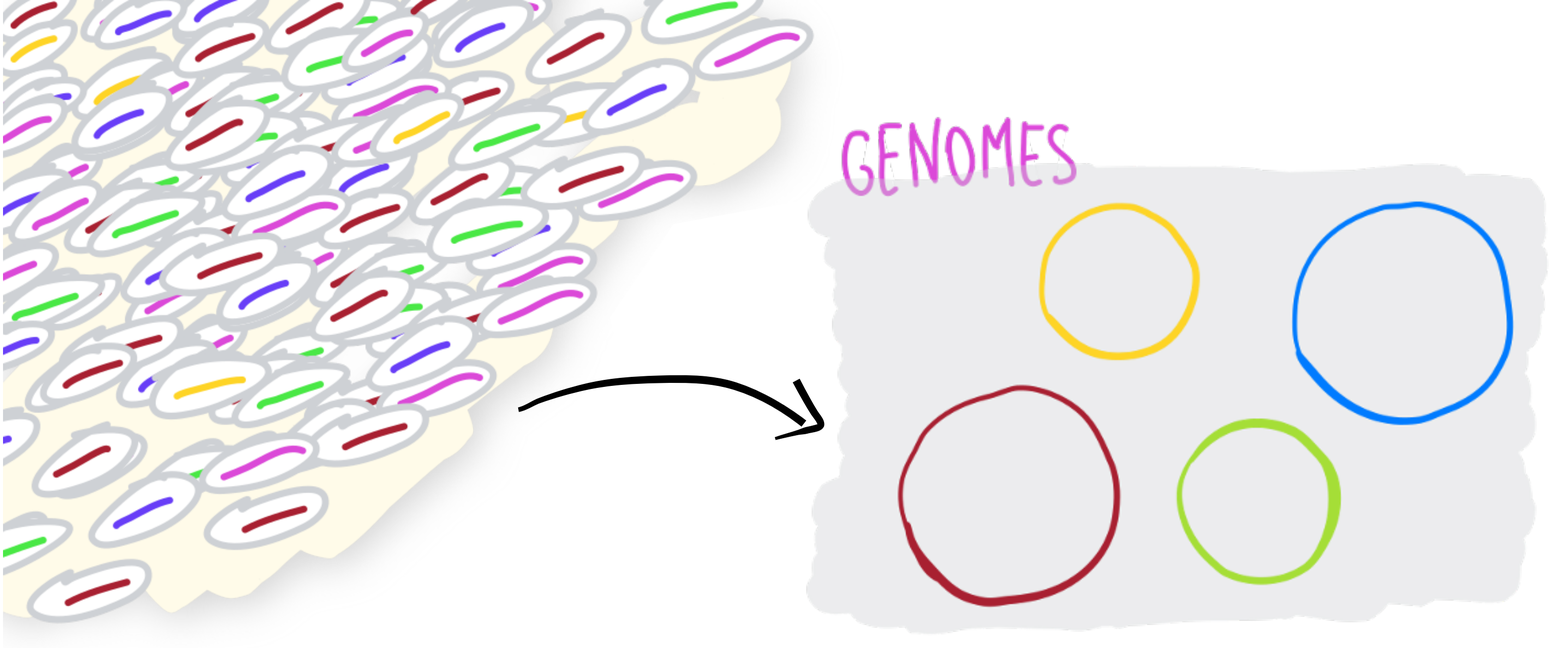


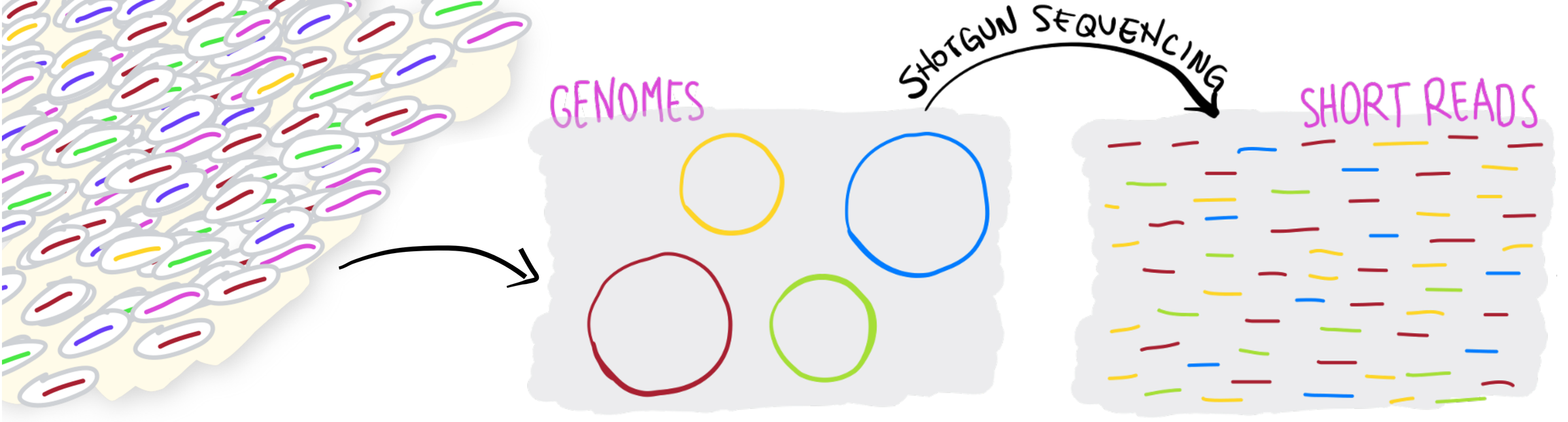
Genome Resolved Metagenomics

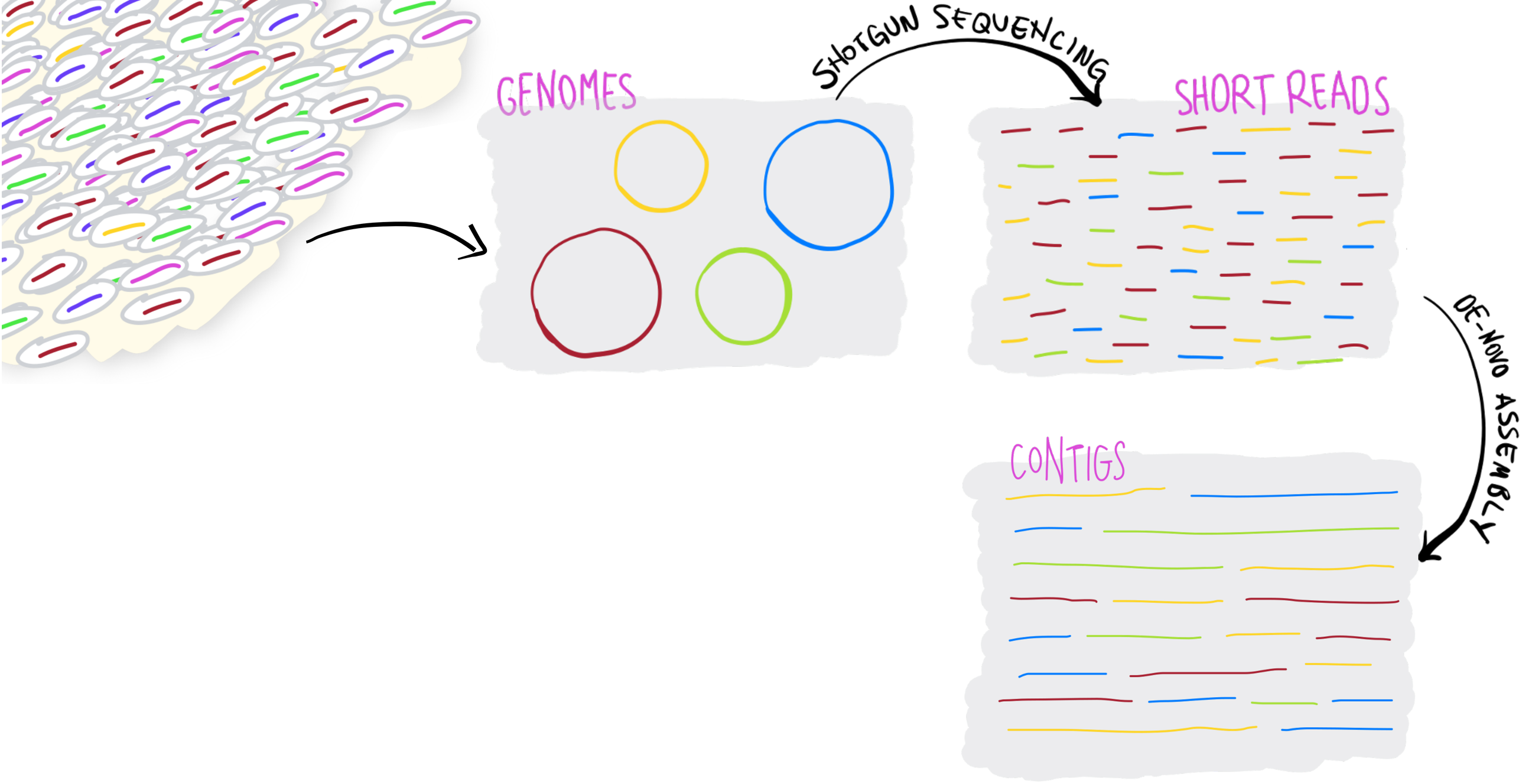


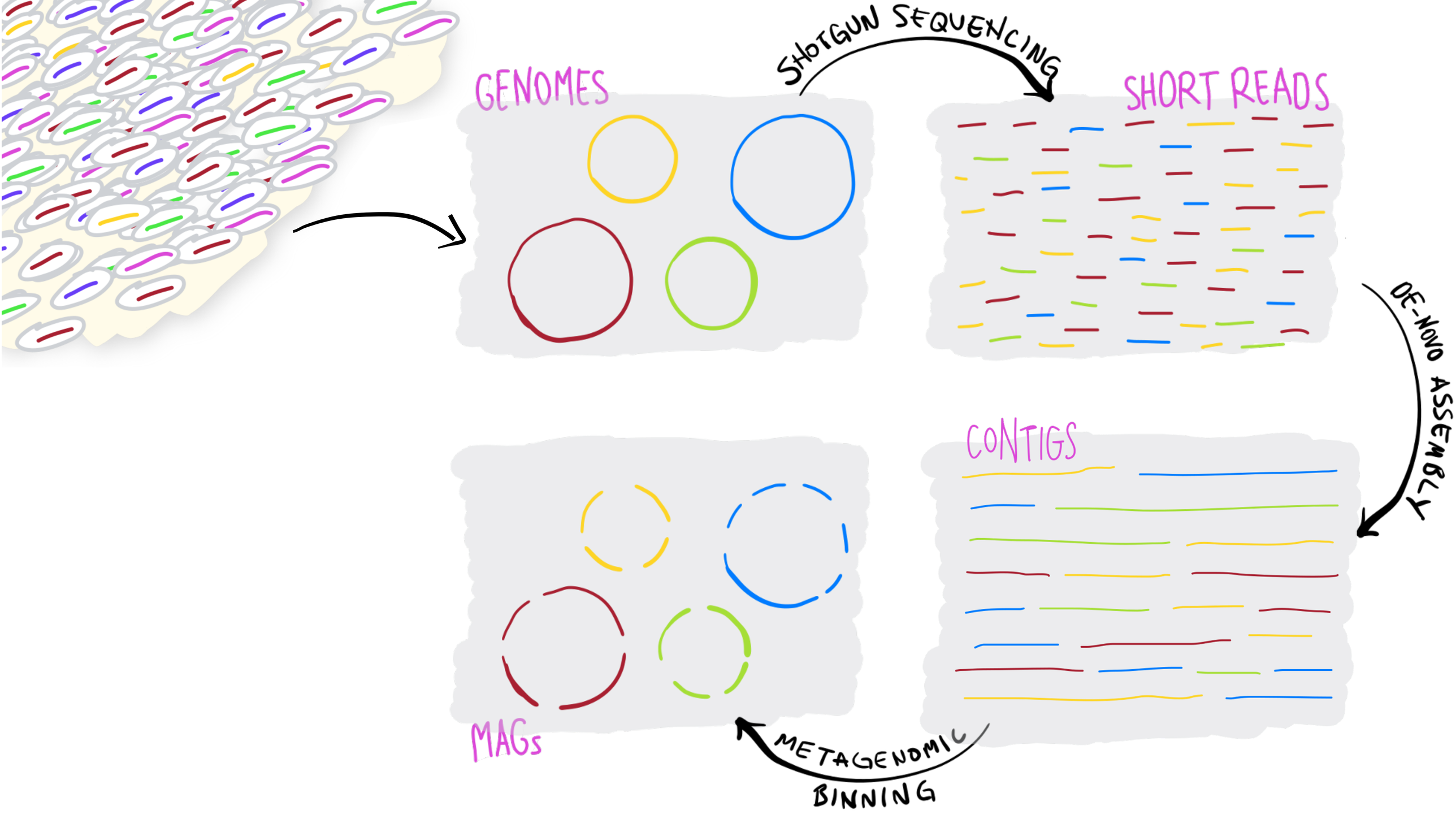
Accessing the missing genomic content

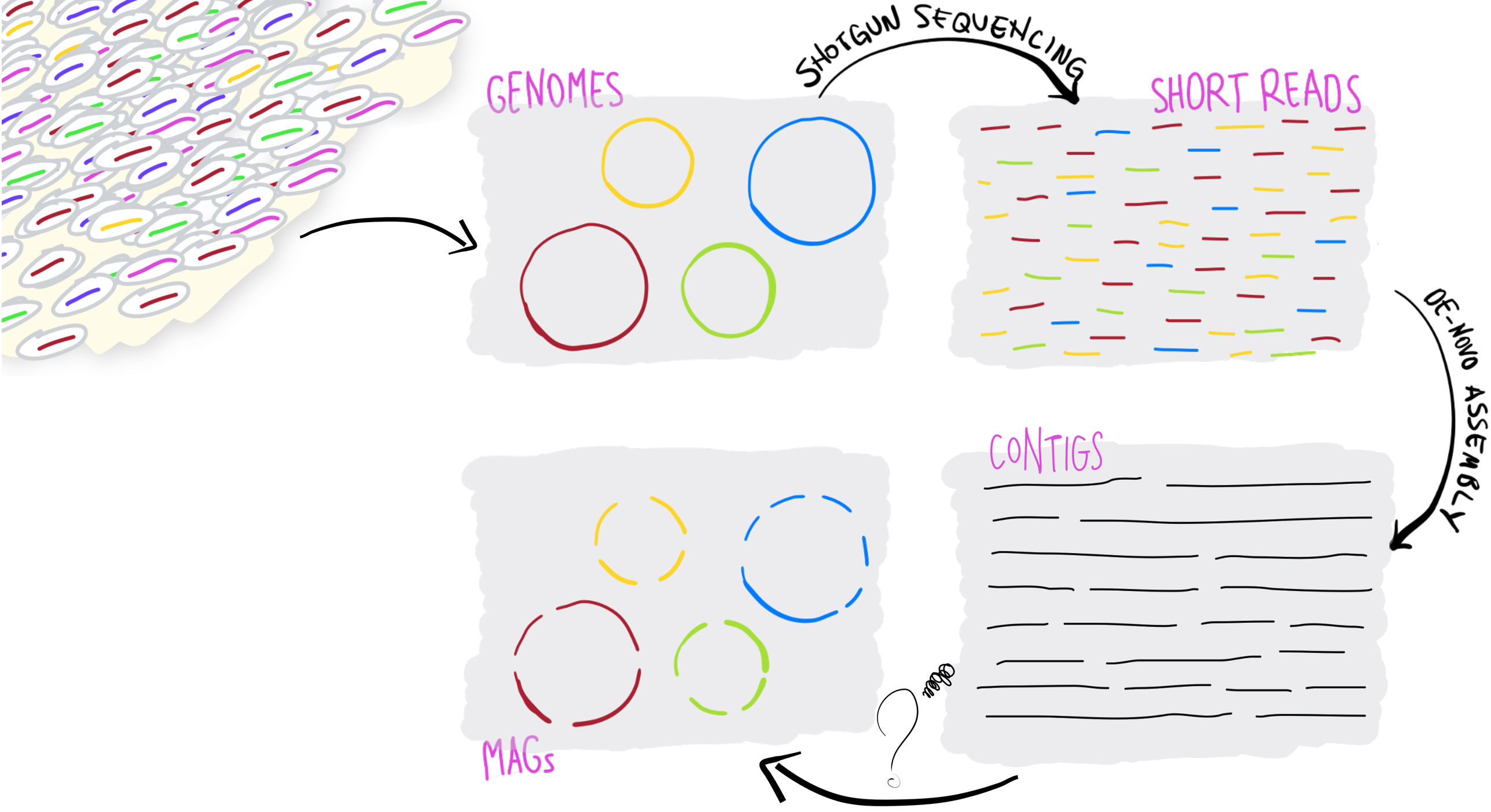
Cultivation-independent methods


GENOMES


GENOMES
SHOTGUN SEQUENCING
SHORT READS


GENOMES
SHOTGUN SEQUENCING
SHORT READS
DE-NOVO ASSEMBLY
CONTIGS


GENOMES
SHOTGUN SEQUENCING
SHORT READS
DE-NOVO ASSEMBLY
CONTIGS
METAGENOMIC
BINNING
MAGs


SHOTGUN SEQUENCING
GENOMES
SHORT READS
DE-NOVO ASSEMBLY
CONTIGS
MAGs
?
100%

Sequence composition
Computing k-mer frequencies

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	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

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

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

→ PALINDROMES :)

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										

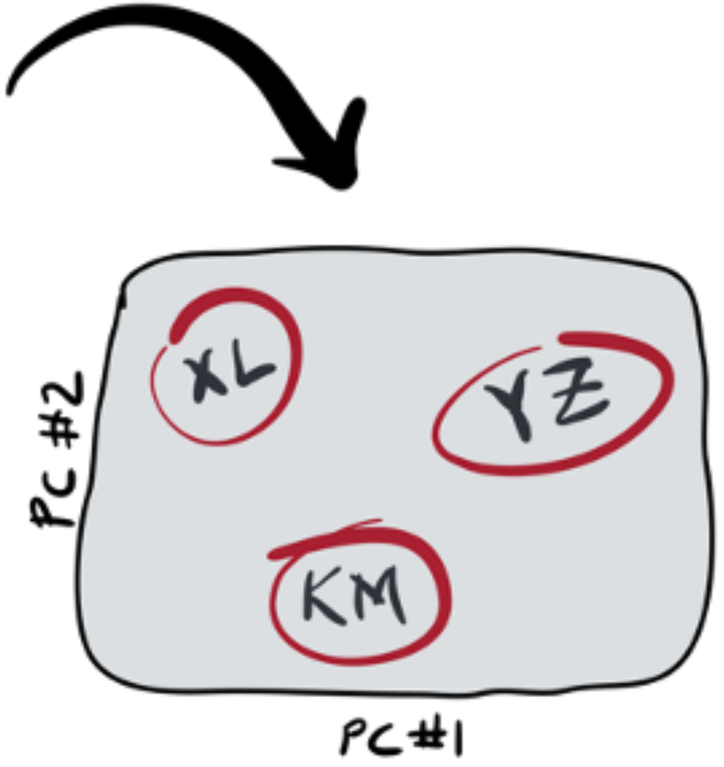
k=2

GGGCCTGCGCCGGTCCAGTCACCCGGCTGCGACCT

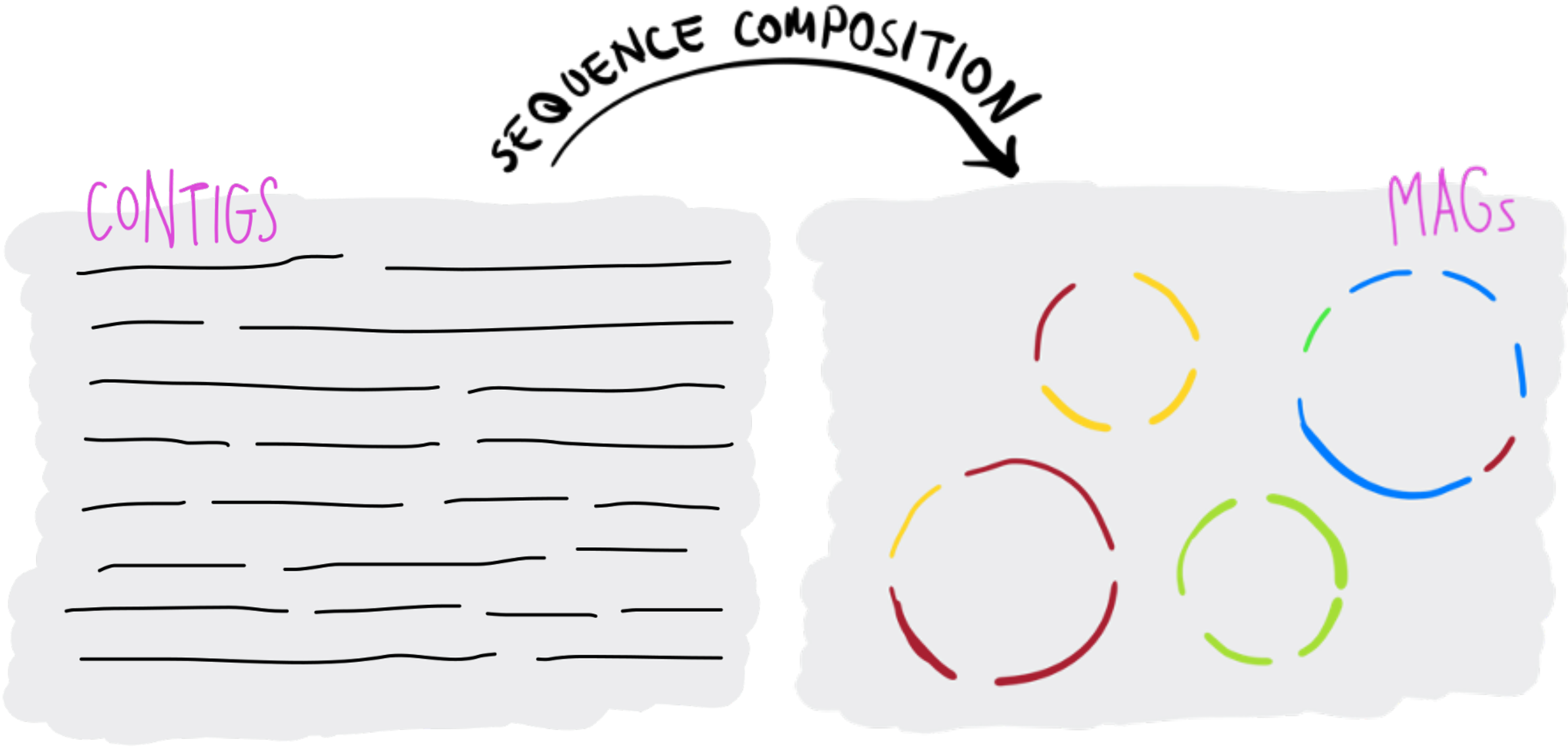
	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y	4	5	2	4	5	4	4	3	2	1
Z	4	5	3	2	4	1	5	5	2	3
L	11	6	3	2	2	3	2	1	1	4
K	1	1	2	2	1	8	9	10	0	0
M	0	4	4	3	4	10	4	5	0	0

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

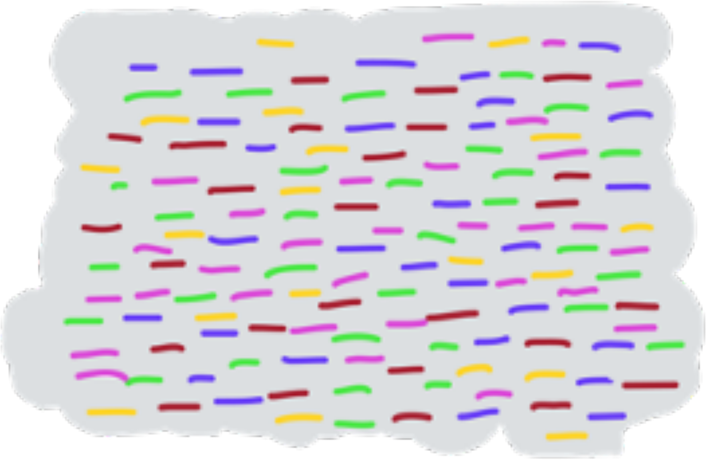
Abundance correlation

Counting stuff across samples

CONTIG #1

CONTIG #2

CONTIG #1

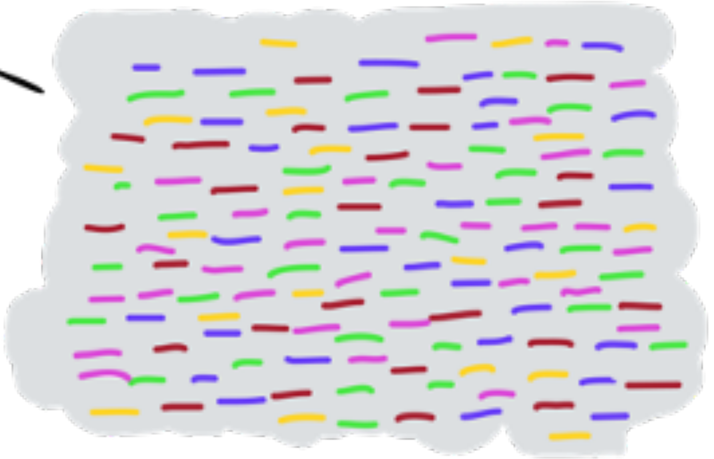


METAGENOMIC READS

CONTIG #2

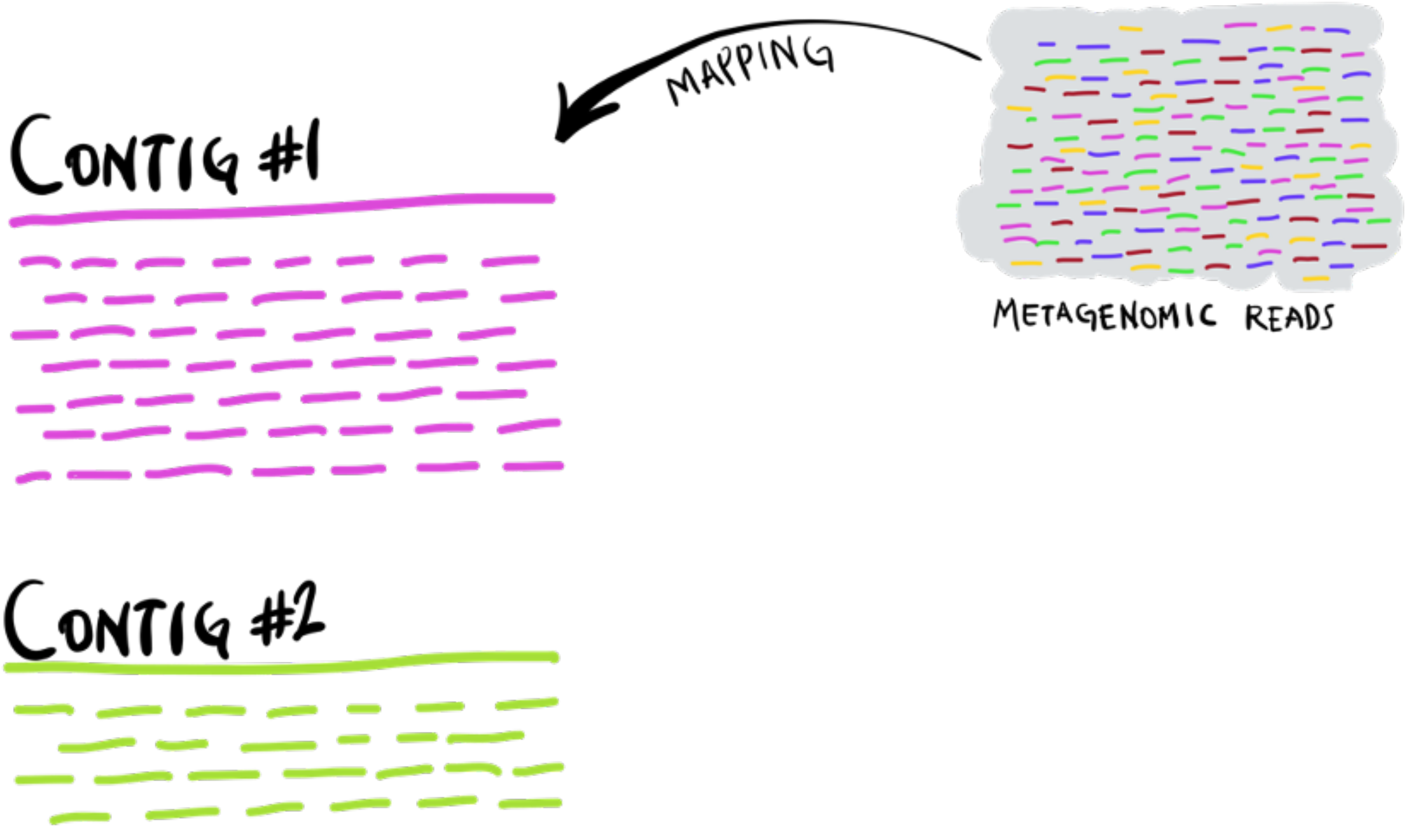
CONTIG #1

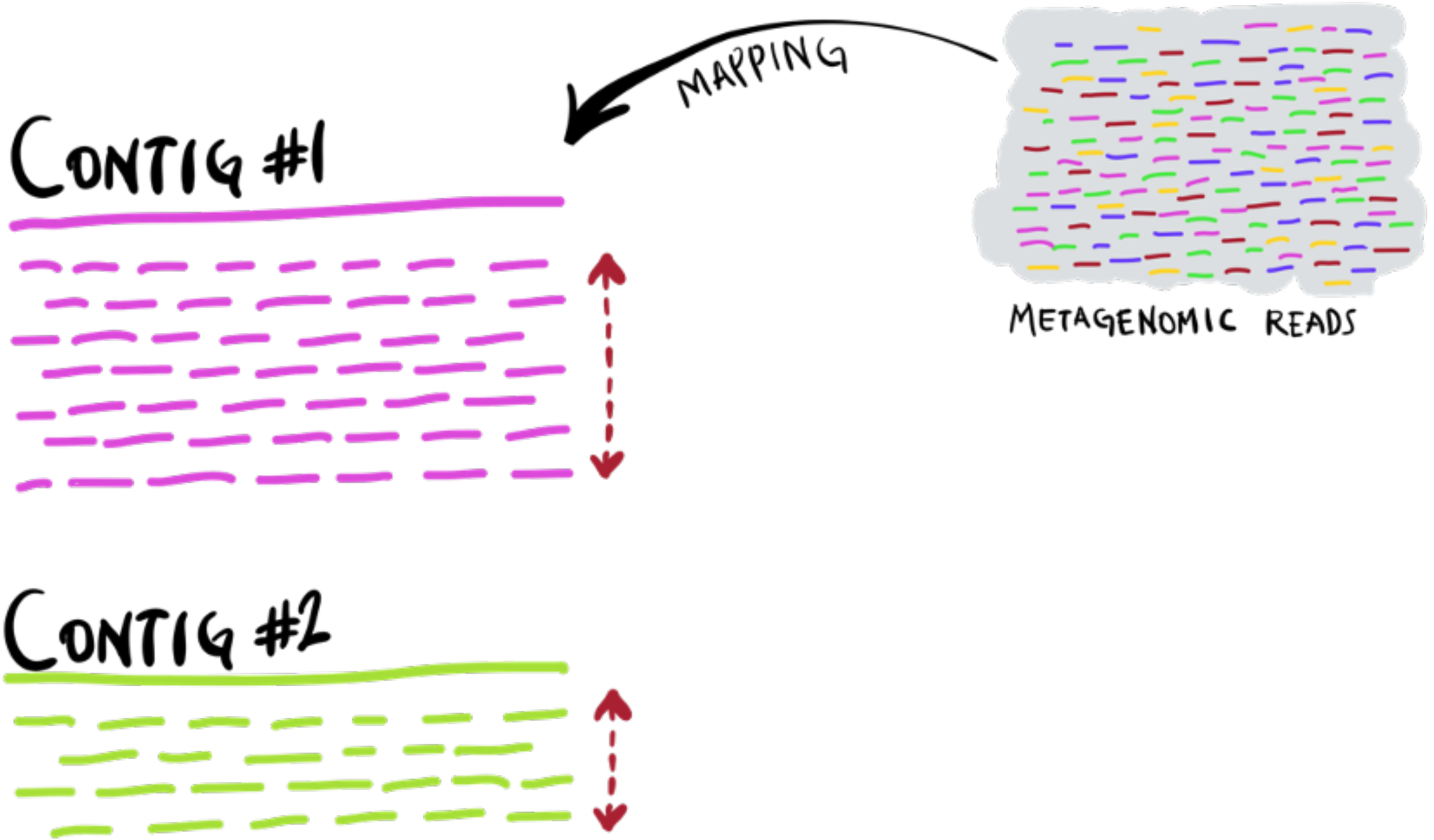
MAPPING

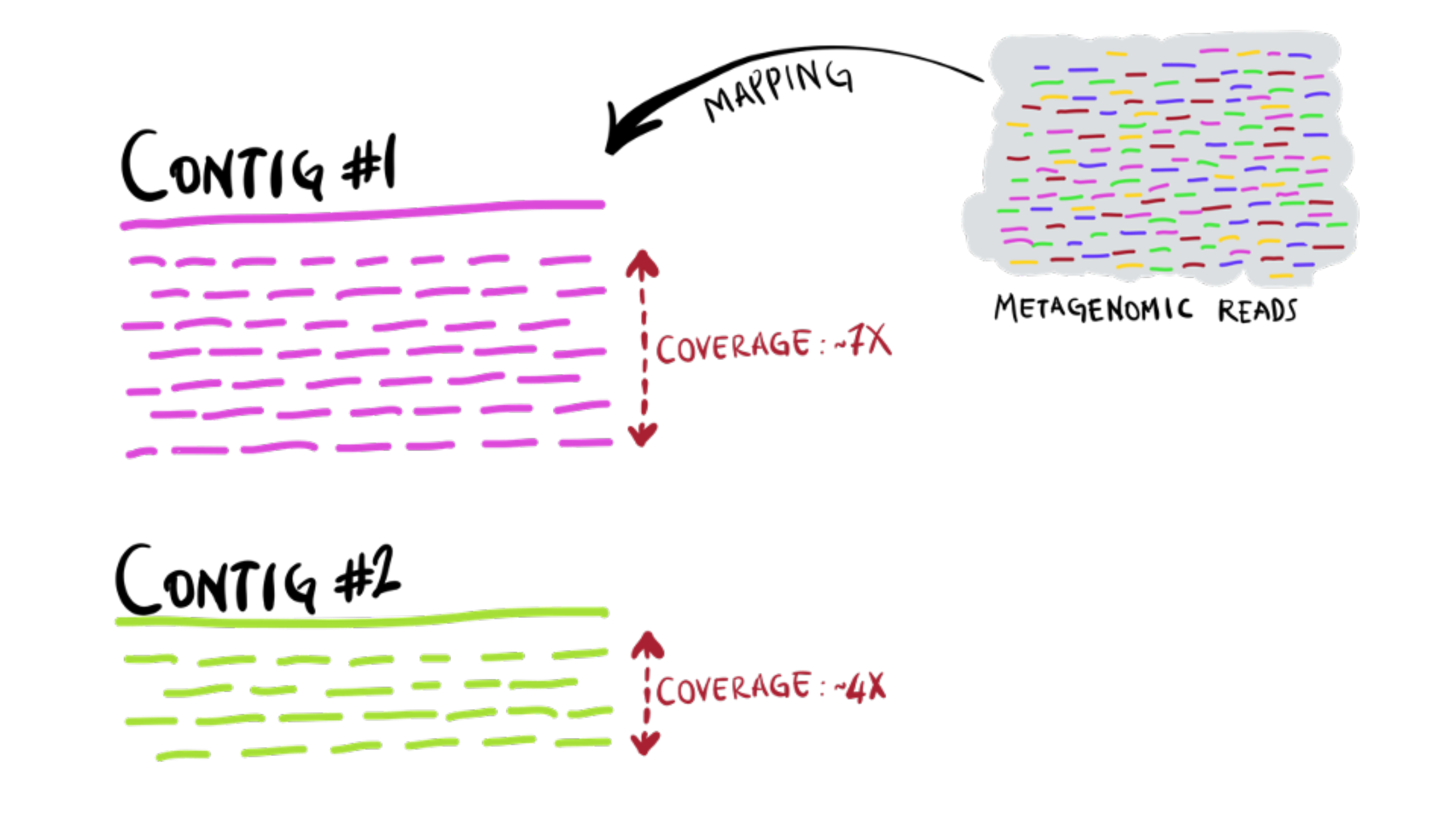


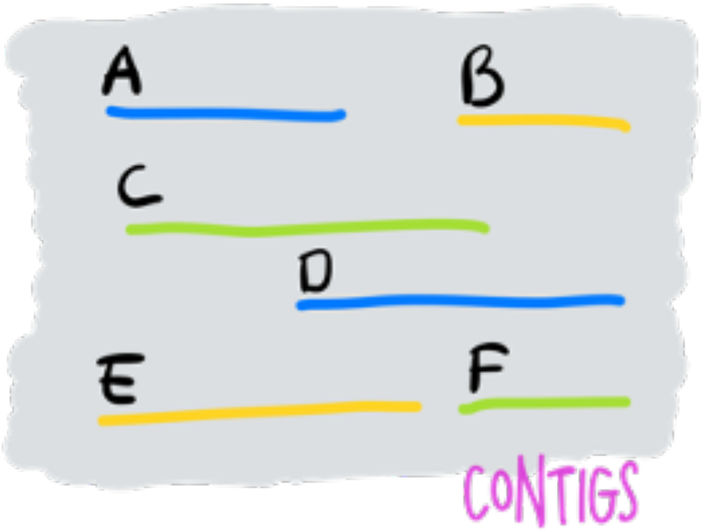
METAGENOMIC READS

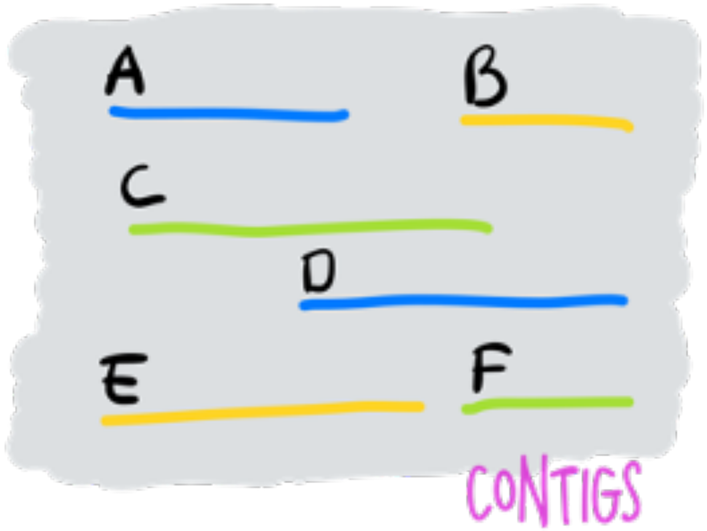
CONTIG #2


MAPPING
CONTIG #1
METAGENOMIC READS
CONTIG #2


MAPPING
METAGENOMIC READS
CONTIG #1
CONTIG #2


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

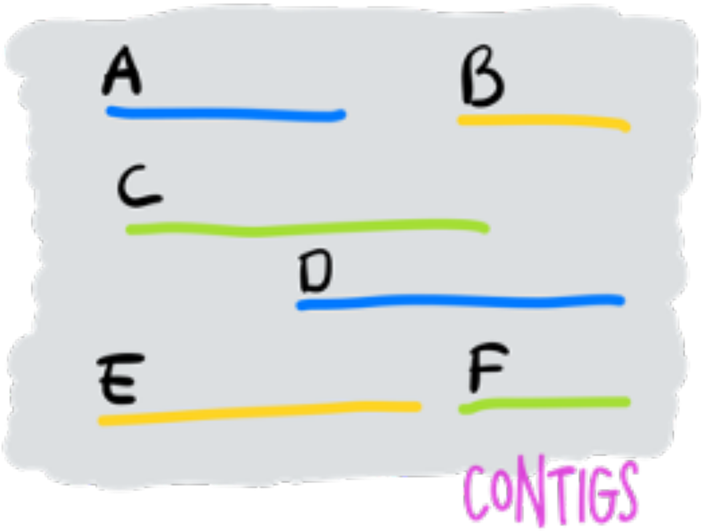

A
B
C
D
E
F
CONTIGS

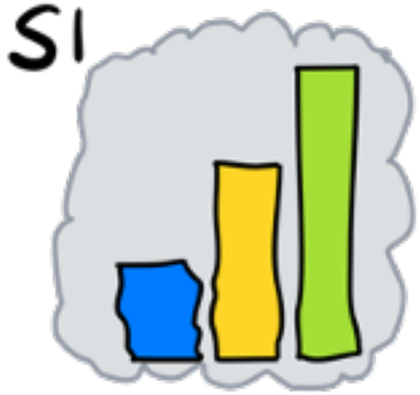


S1

S2

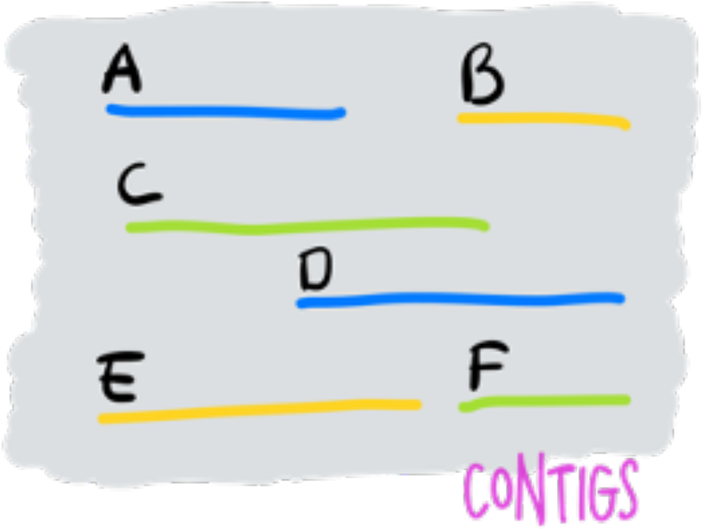
S3

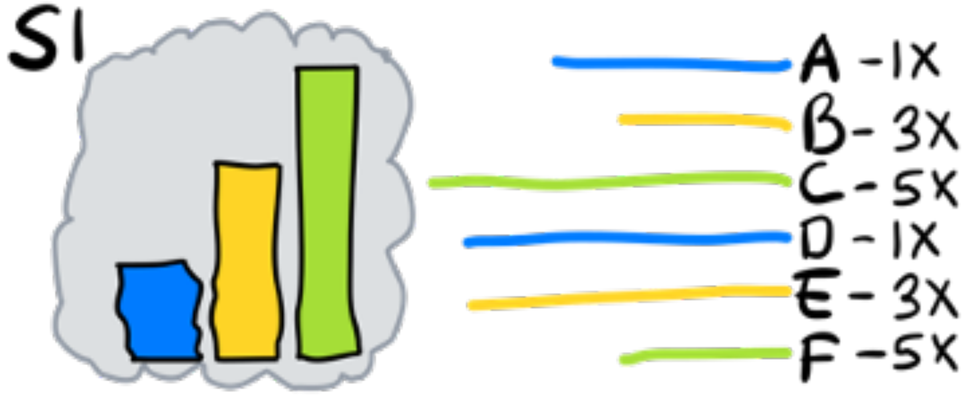

A
B
C
D
E
F
CONTIGS


S1

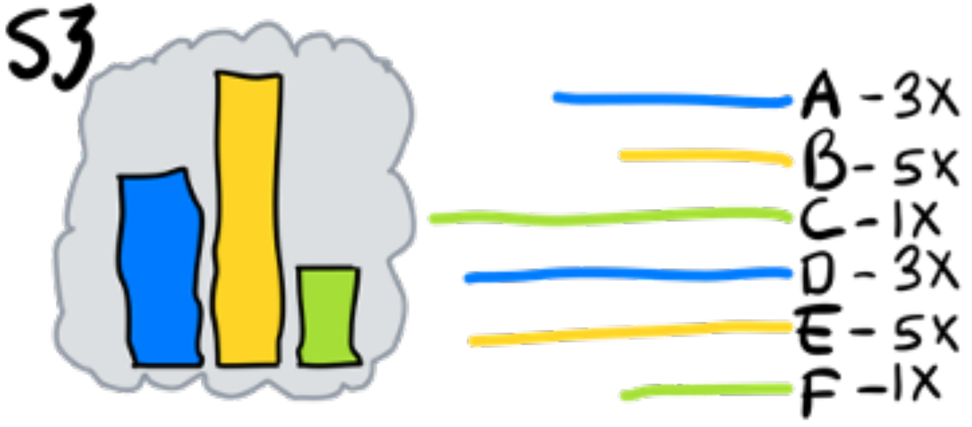

S2

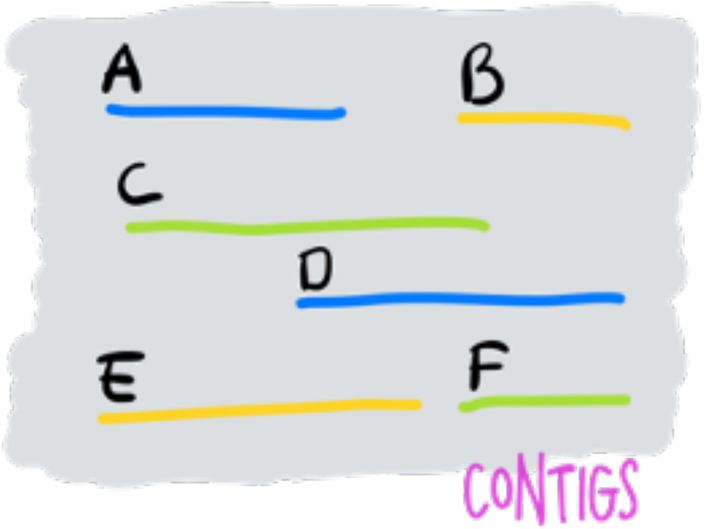

S3

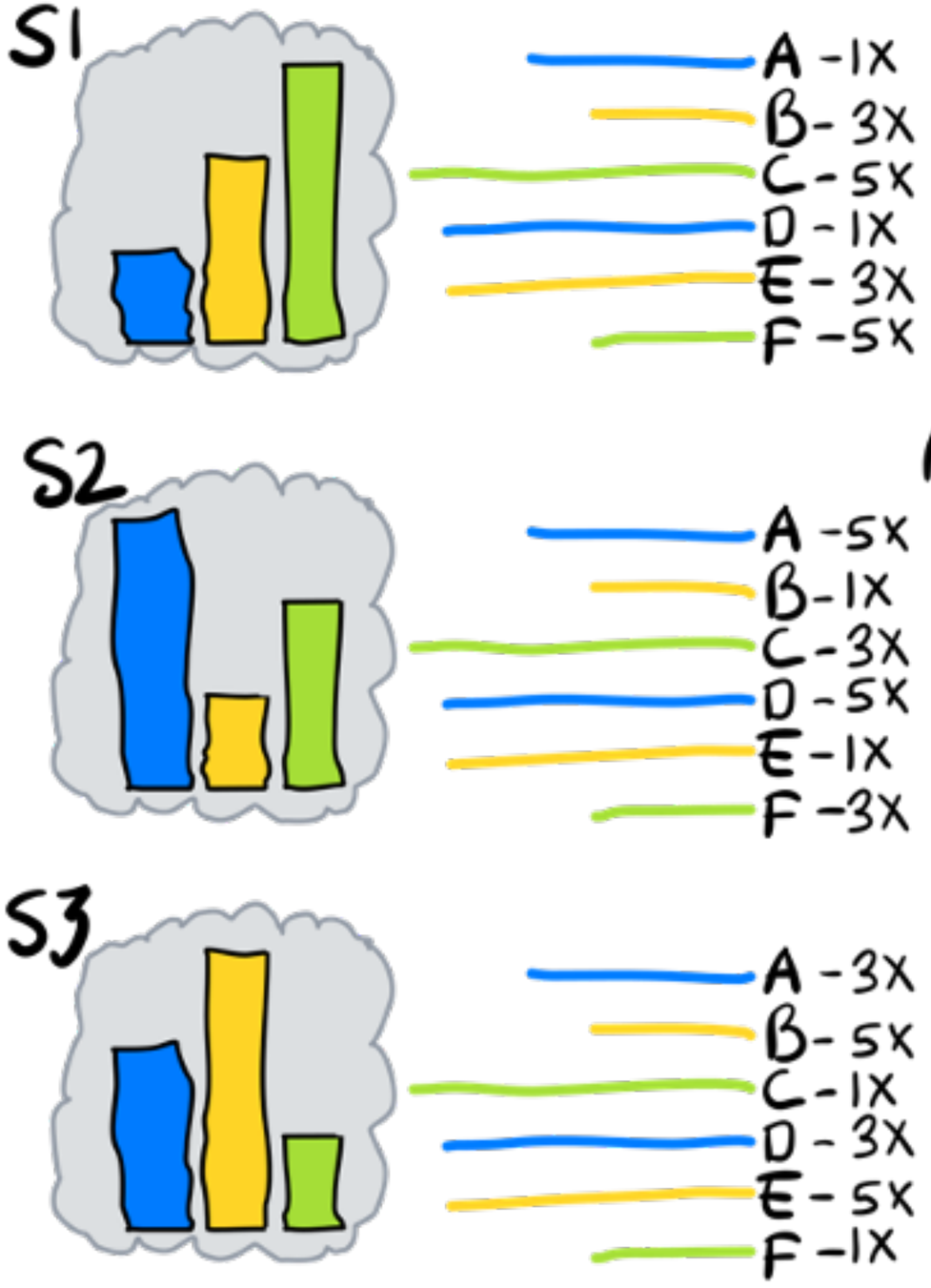

A
B
C
D
E
F
CONTIGS


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

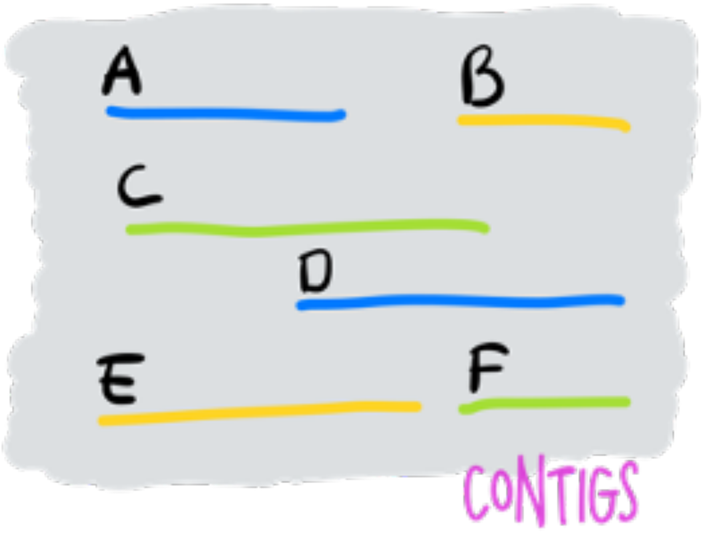

S2
A -5X
B-1X
C-3X
D-5X
E-1X
F -3X

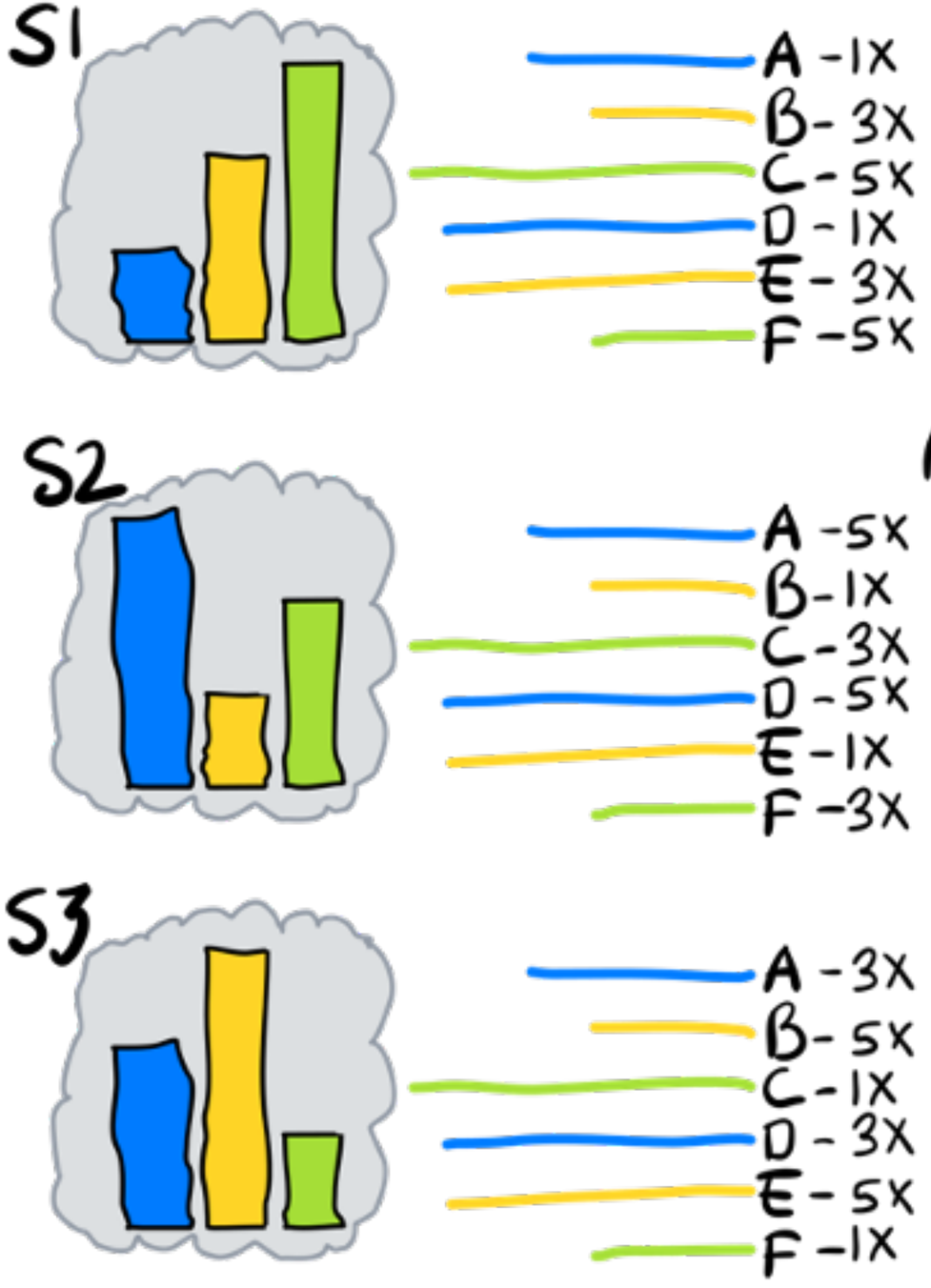

S3
A -3X
B- 5X
C-1X
D- 3X
E- 5X
F -1X



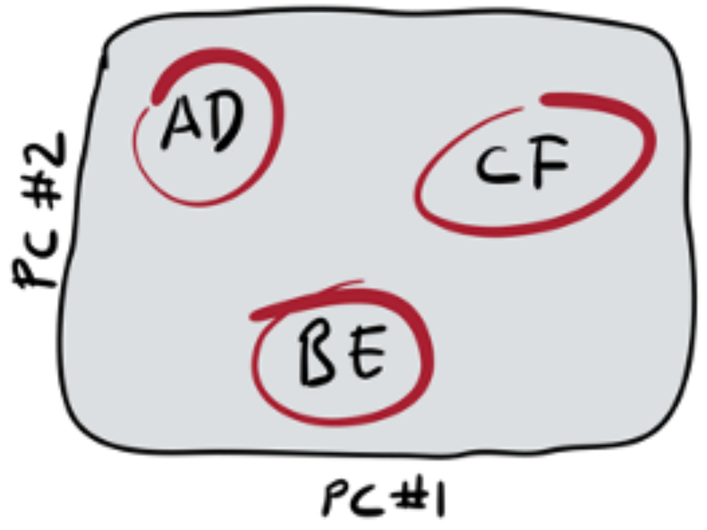


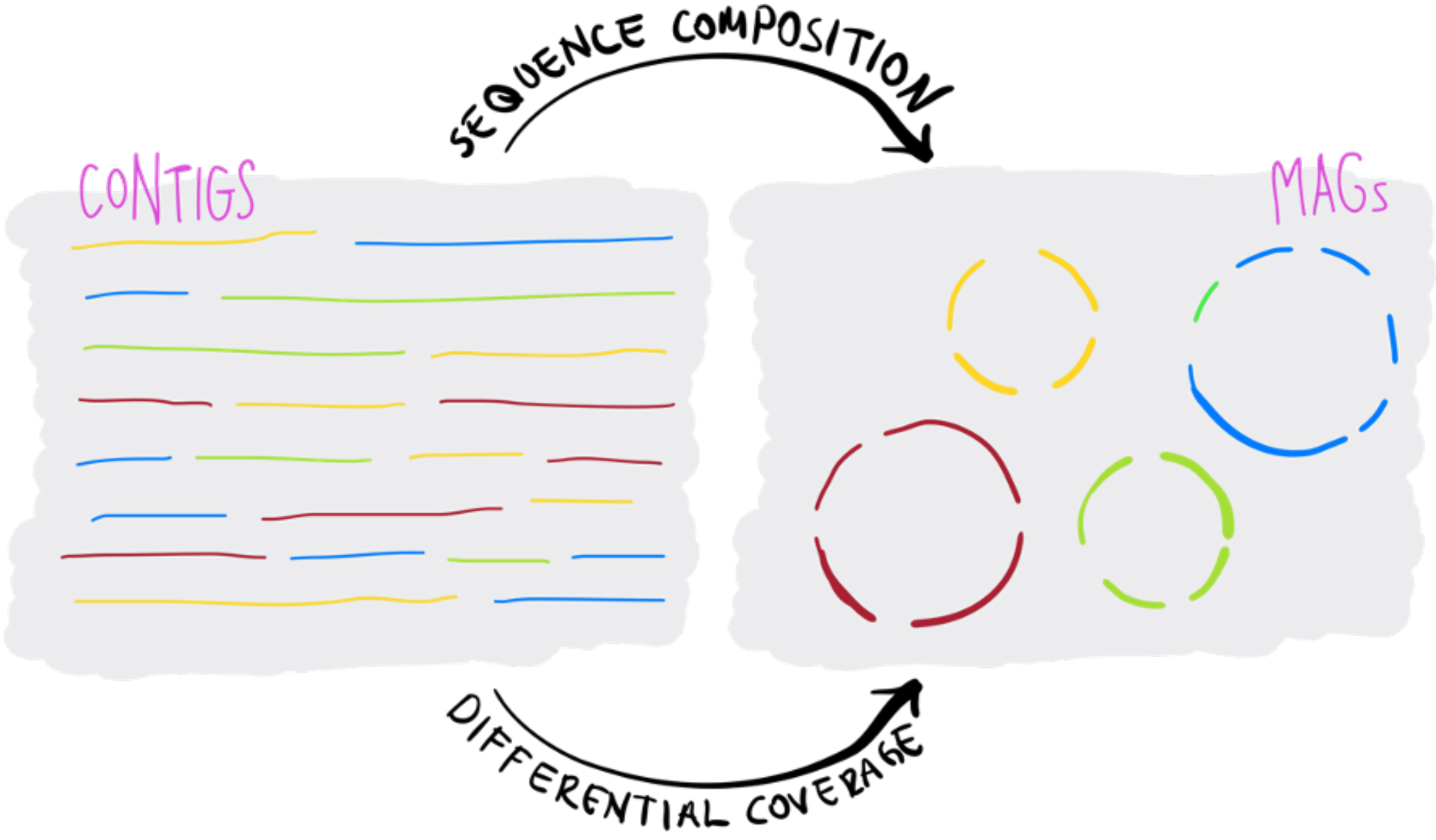
	A	B	C	D	E	F
S1	1	3	5	1	3	5
S2	5	1	3	5	1	3
S3	3	5	1	3	5	1





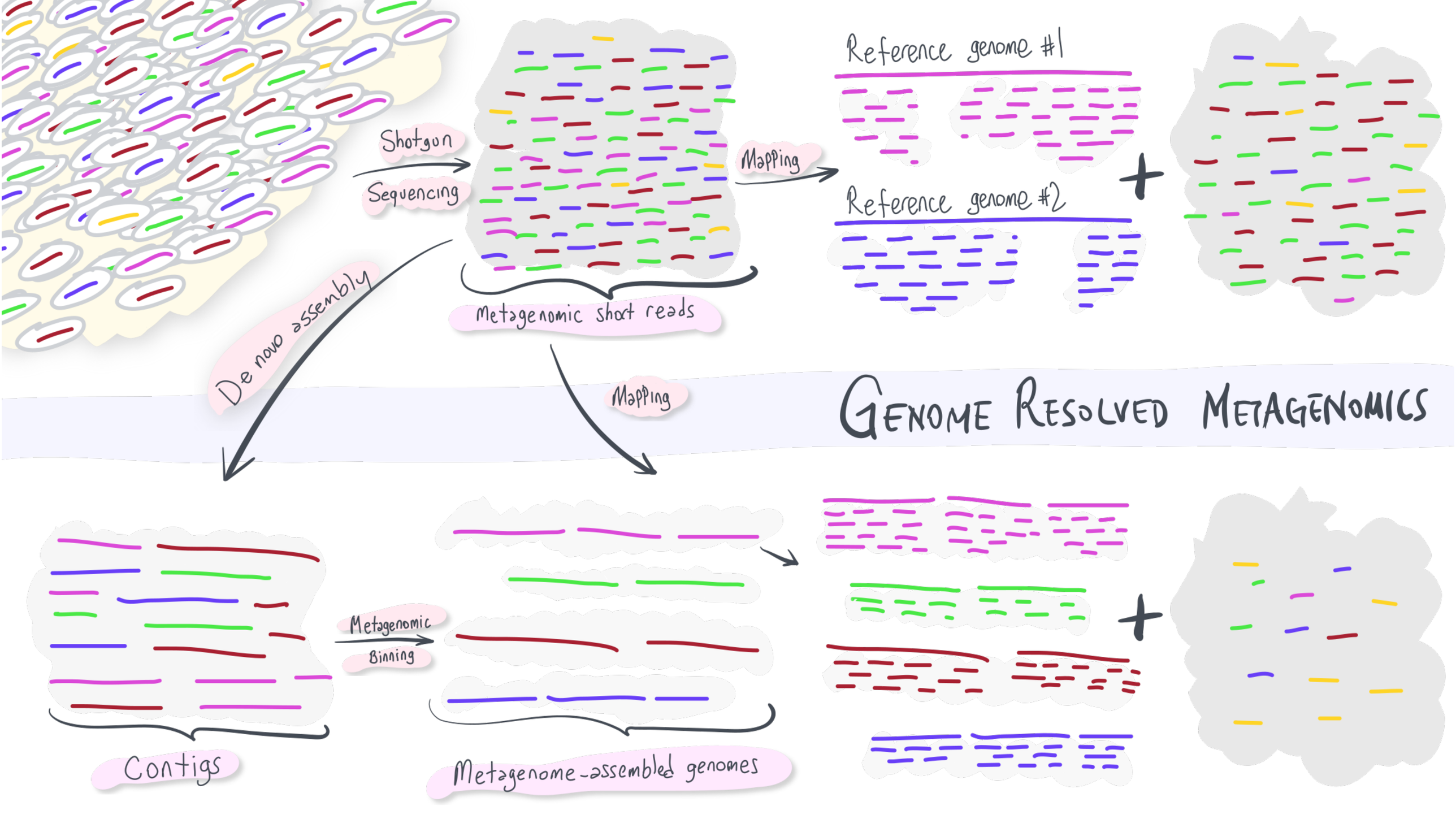
	A	B	C	D	E	F
S1	1	3	5	1	3	5
S2	5	1	3	5	1	3
S3	3	5	1	3	5	1




SEQUENCE COMPOSITION
CONTIGS
MAGs
DIFFERENTIAL COVERAGE

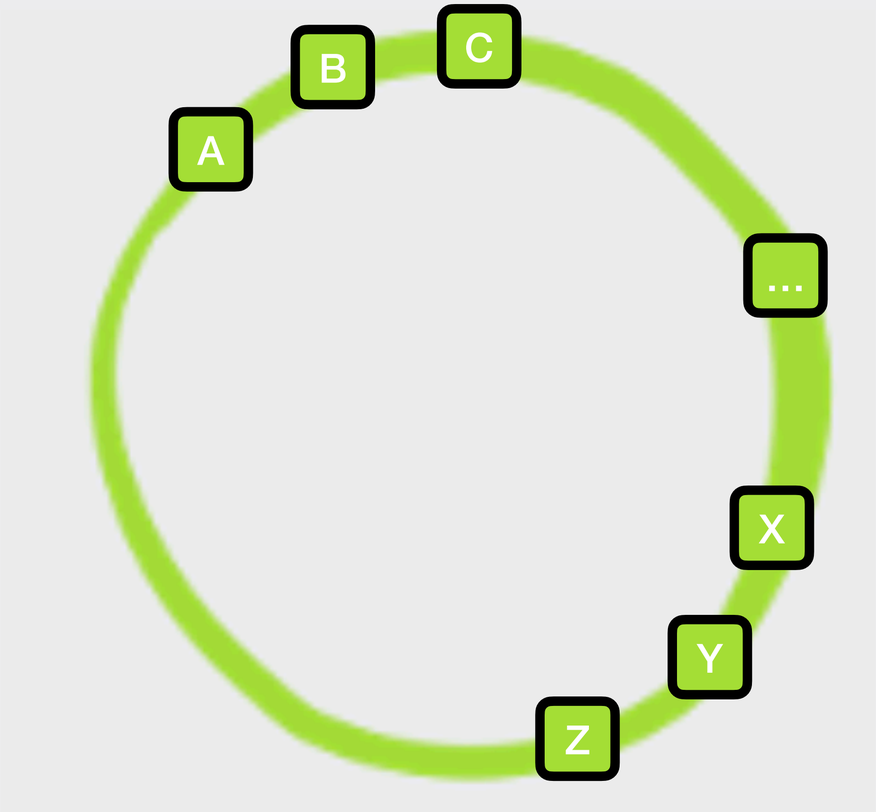
Cultivation-independent genome-resolved metagenomics

A summary

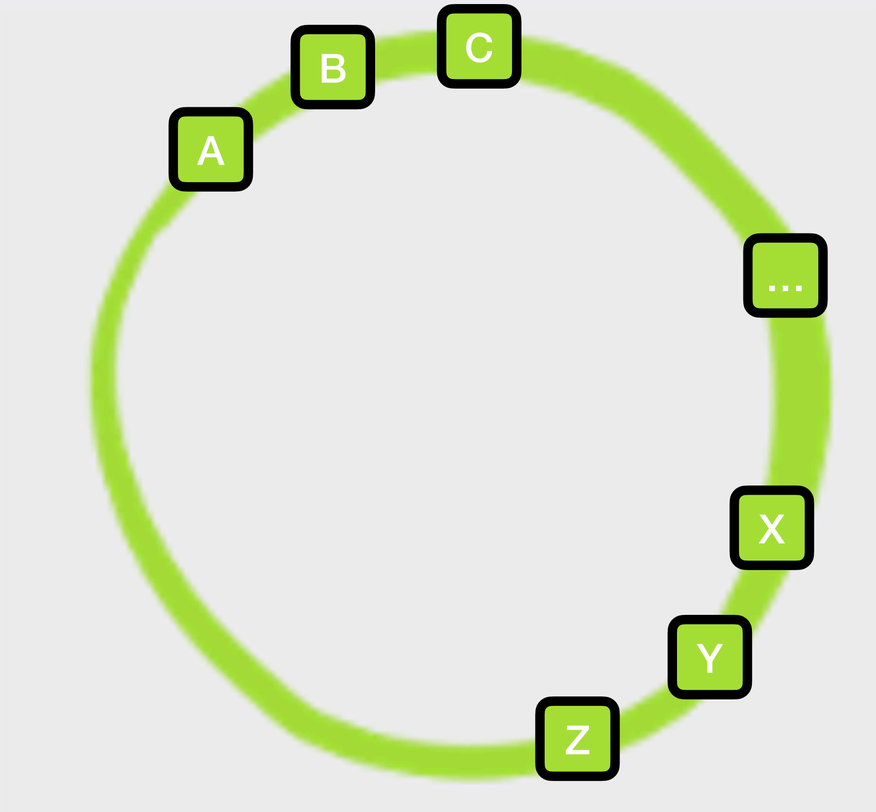

Shotgun
Sequencing
Metagenomic short reads
Mapping
Reference genome #1
Reference genome #2
De novo assembly
Mapping
GENOME RESOLVED METAGENOMICS
Contigs
Metagenomic
Binning
Metagenome-assembled genomes

Nothing’s perfect
Evaluation and limits of MAGs

Universal single-copy marker genes

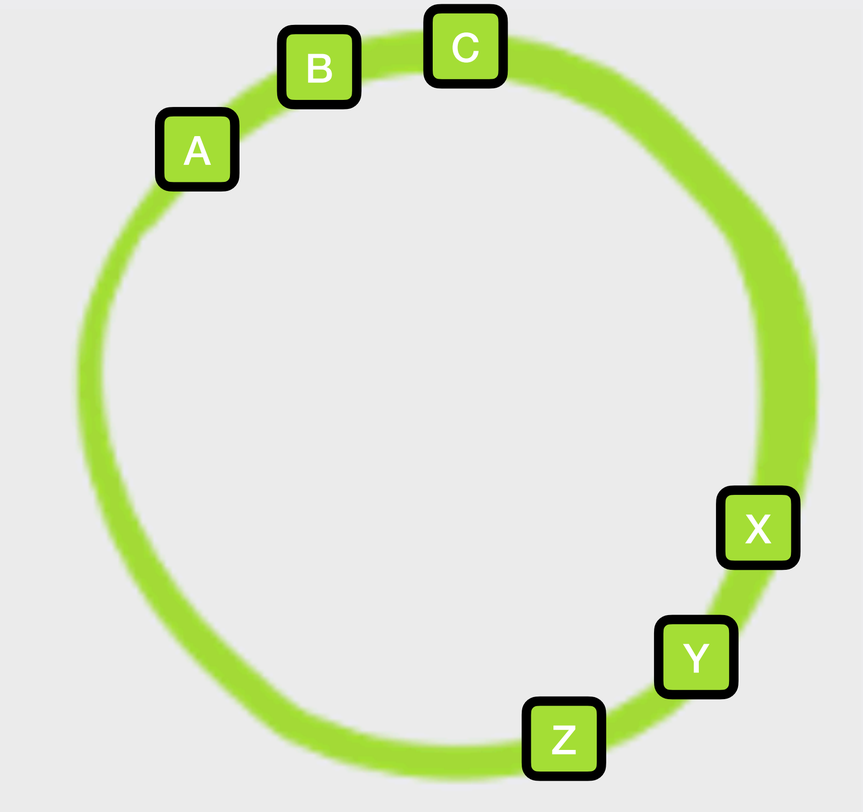


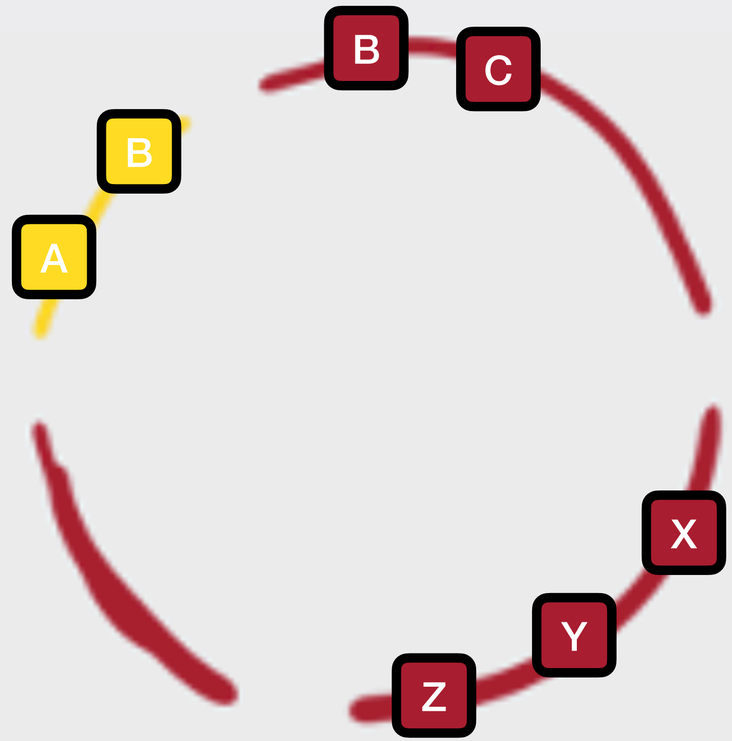
Universal single-copy marker genes

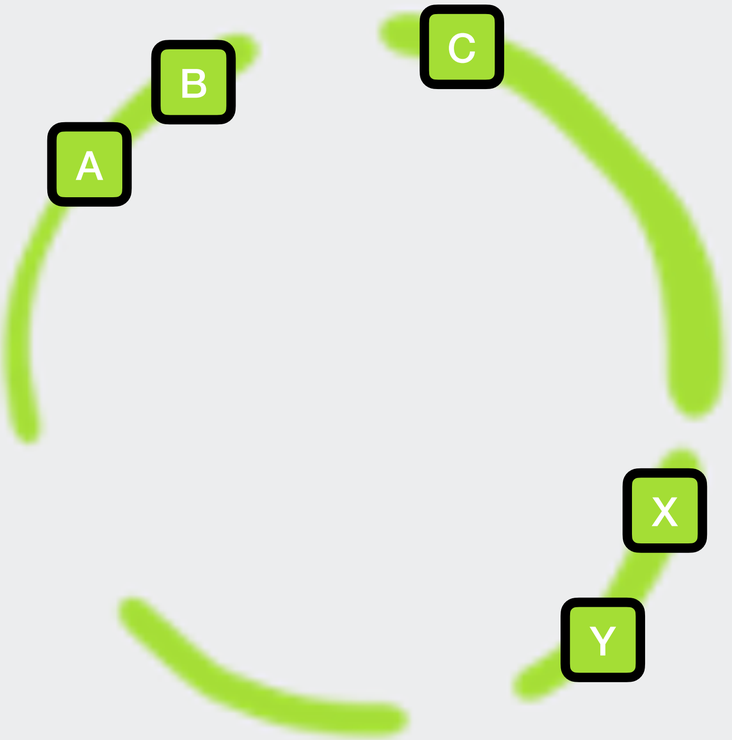


Between 40 and 120 for Bacteria/Archaea depending on cutoffs

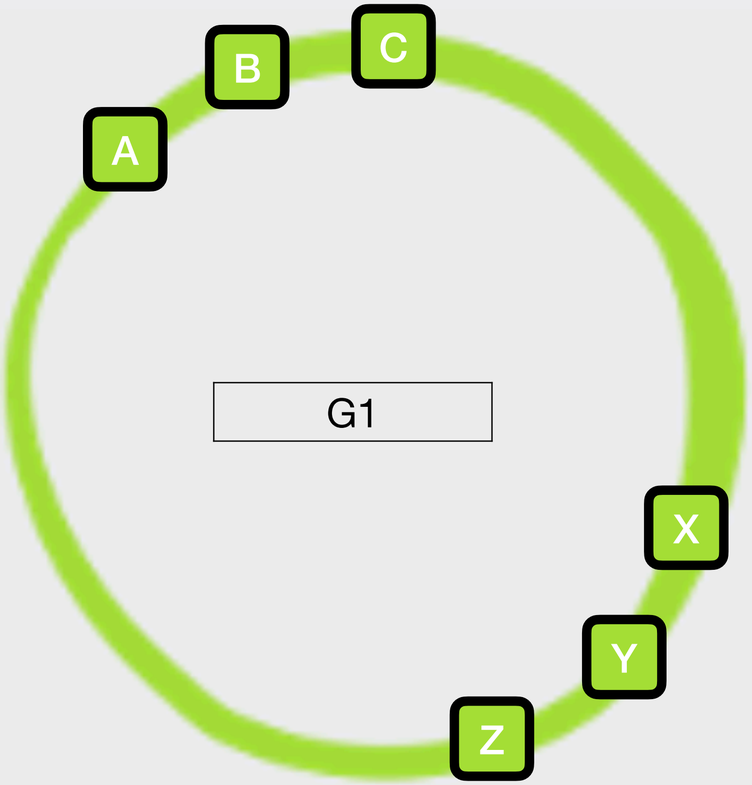
Universal single-copy marker genes

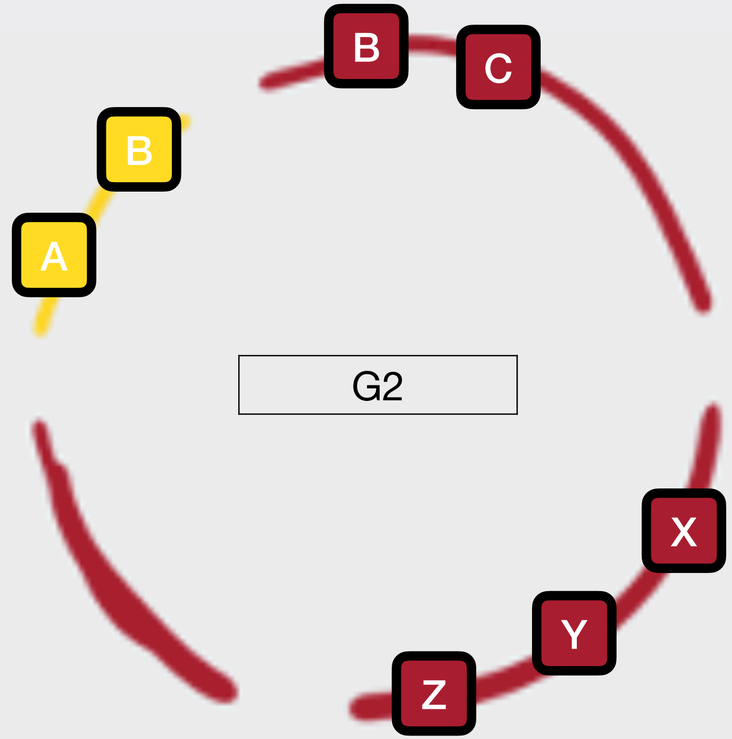


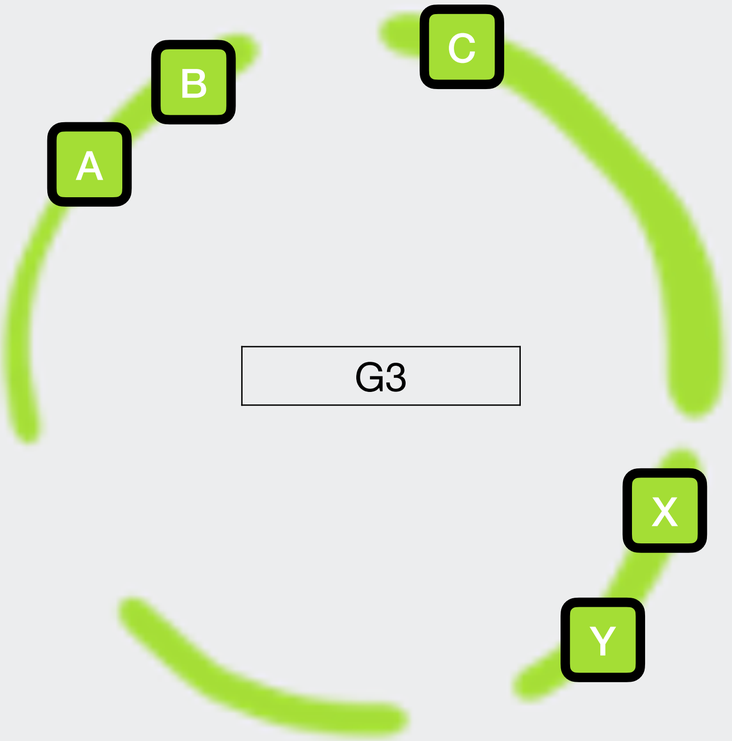




Universal single-copy marker genes



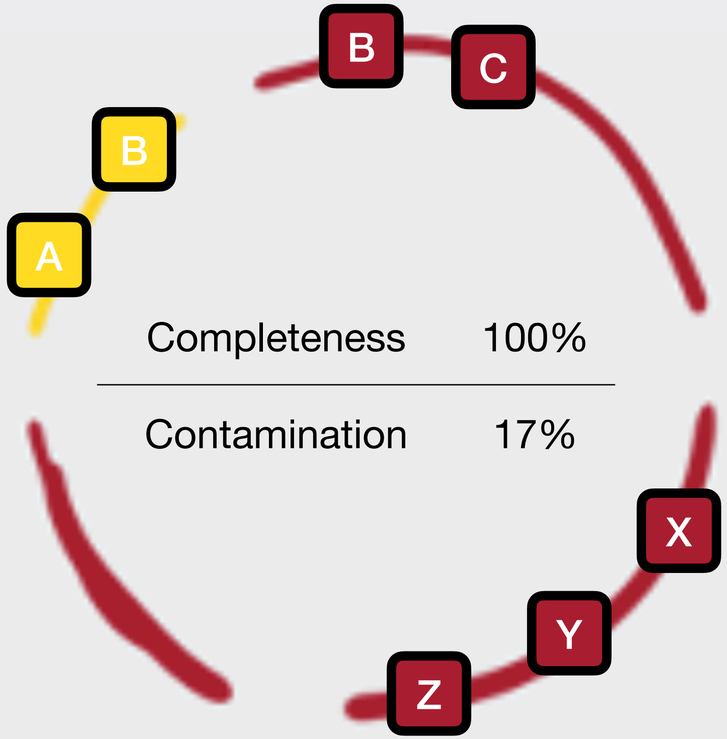




	G1	G2	G3
A	1	1	1
B	1	2	1
C	1	1	1
X	1	1	1
Y	1	1	1
Z	1	1	0

Universal single-copy marker genes

Completeness	100%
Contamination	0%



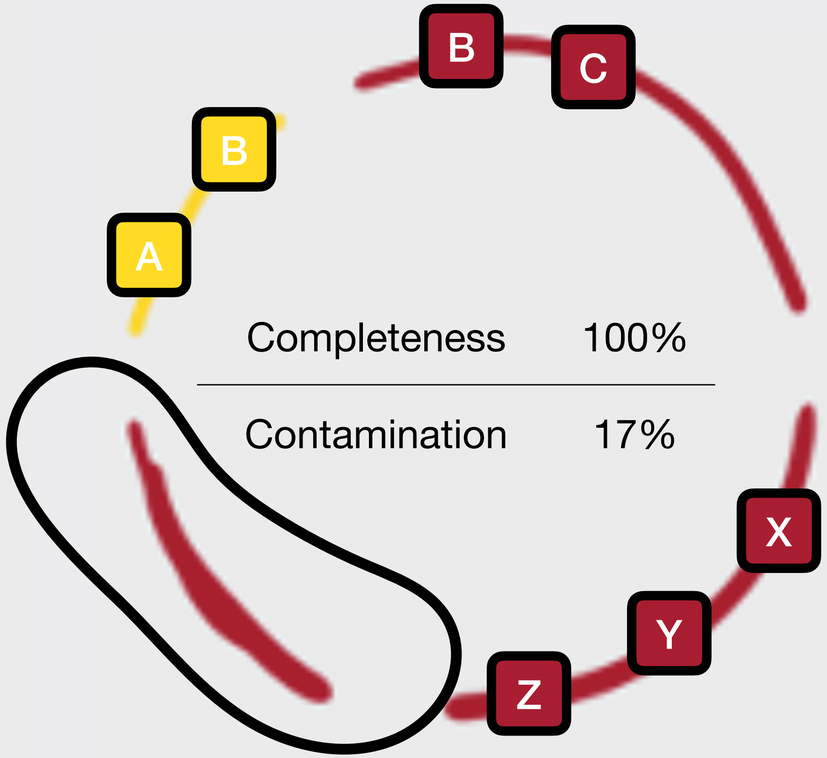
Completeness	100%
Contamination	17%

Completeness	83%
Contamination	0%

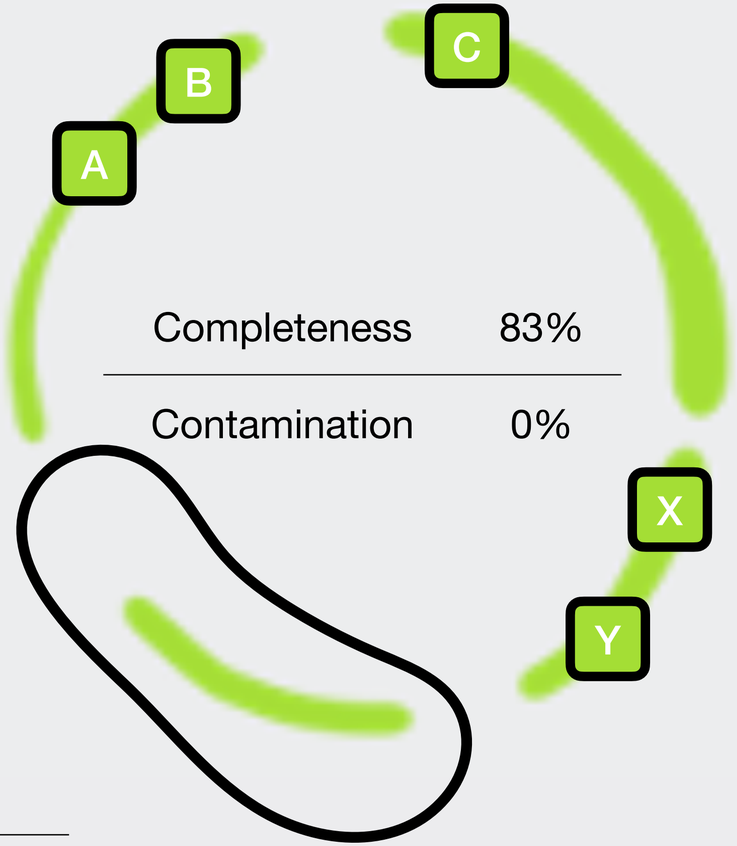
	G1	G2	G3
A	1	1	1
B	1	2	1
C	1	1	1
X	1	1	1
Y	1	1	1
Z	1	1	0

Universal single-copy marker genes

Completeness	100%
Contamination	0%



Completeness	100%
Contamination	17%



Completeness	83%
Contamination	0%

	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

Material

This course uses a lot of material from https://merenlab.org/momics/, I invite you to have a look.

If you want details on the bioinformatics behind you can start by having a look here: https://astrobiomike.github.io/genomics/metagen_anvio

A. Murat Eren (Meren) (PI)

☐ Web ☐ Email ☐ Twitter ☐ LinkedIn ☐ Github ☐ ORCiD

☐ Address: Knapp Center for Biomedical Discovery, 900 E. 57th St., MB 9, RM 9118, Chicago, IL 60637 USA

☐ Phone: +1-773-702-5935 ☐ Fax: +1-773-702-2281

I am a computer scientist with a deep appreciation for the complexity of life. I design algorithms and experiments to better understand microbes and their ecology. [photos: 1, 2, 3].

» MBL Fellow, Marine Biological Laboratory.
» Assistant Professor, The Department of Medicine at the University of Chicago.
» Committee on Microbiology, The Biomedical Sciences Cluster at the University of Chicago.

Mike Lee

☐ Web ☐ Email ☐ Twitter ☐ LinkedIn ☐ Github

» NASA Space Biology Fellow, NASA Ames Research Center.
» JCVI Research Fellow, J. Craig Venter Institute.

✍ Combining reference genome annotations with your own in pangenomes (Sat, Dec 01, 2018)
✍ Anvi'o 'views' demystified (Mon, May 08, 2017)
✍ Making anvi'o use your own HMM collection (Sat, May 21, 2016)

Application to the ocean microbiome

Uncharted biosynthetic potential

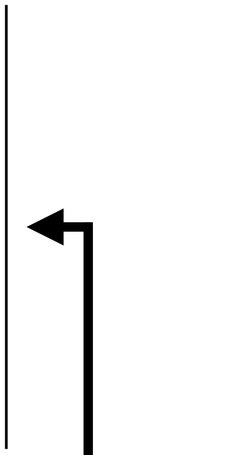
Why explore environmental microbiomes?

- Ubiquitous across earth's ecosystems
- Support global food webs
- Underpin biogeochemical cycles
- Determine Host's health and disease
- …
- **Untapped metabolic diversity**
  - **Biosynthetic potential**
    - **New Enzymes**
    - **New Natural products**

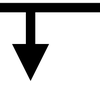
Applications

Why explore environmental microbiomes?

- Ubiquitous across earth's ecosystems
- Support global food webs
- Underpin biogeochemical cycles
- Determine Host's health and disease
- …
- **Untapped metabolic diversity**
  - **Biosynthetic potential**
    - **New Enzymes**
    - **New Natural products**

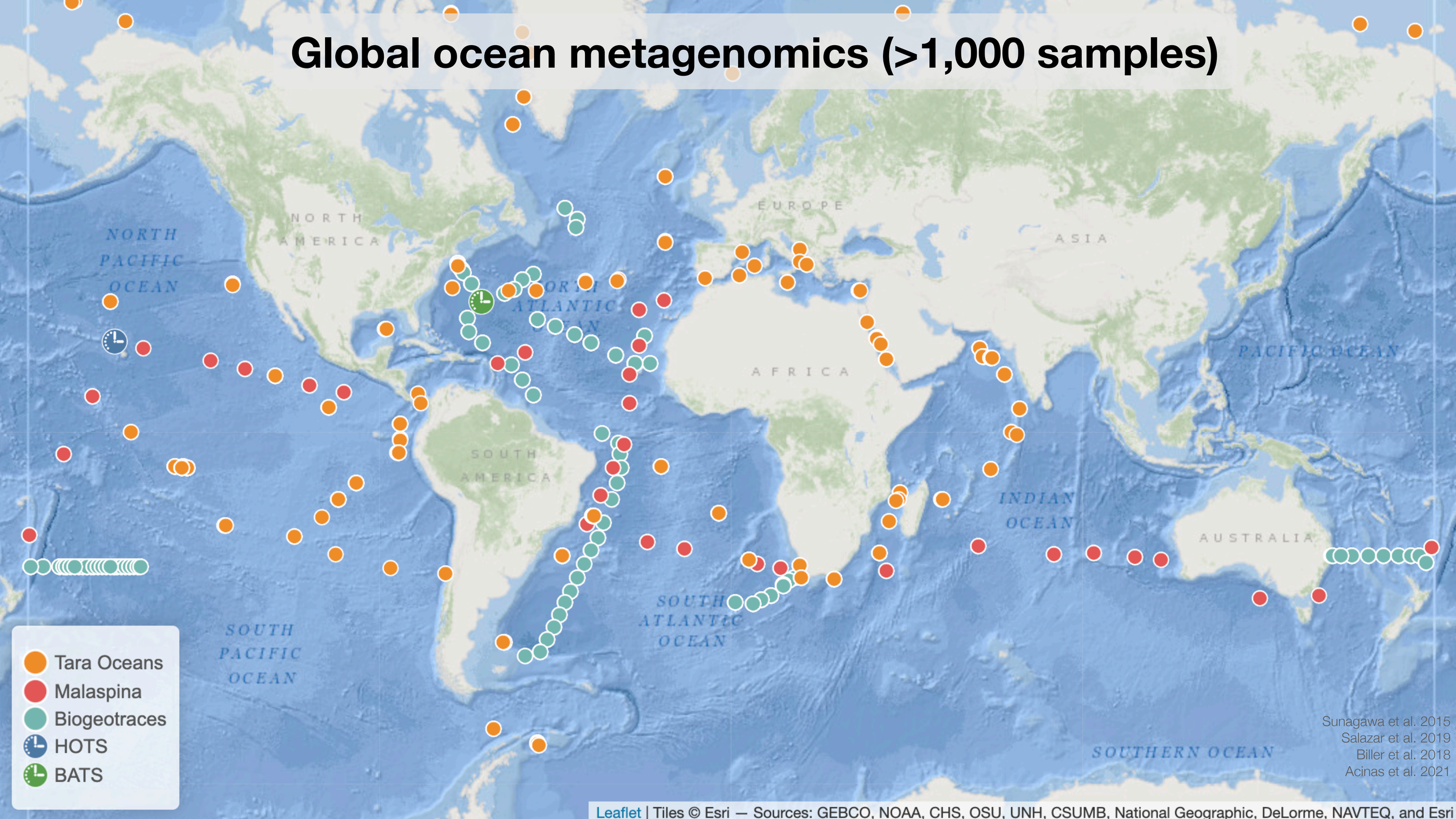


Microbial Interactions

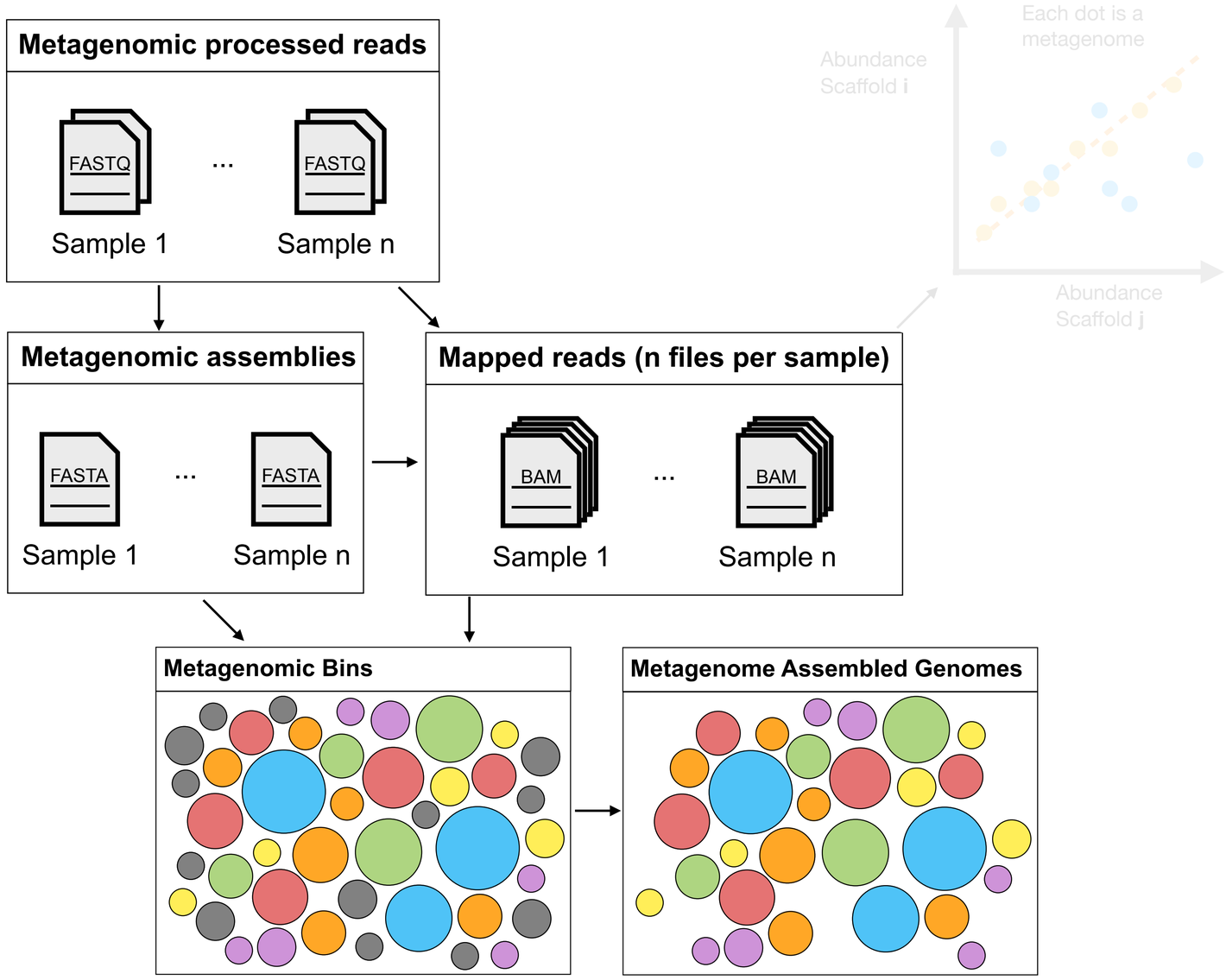


Applications

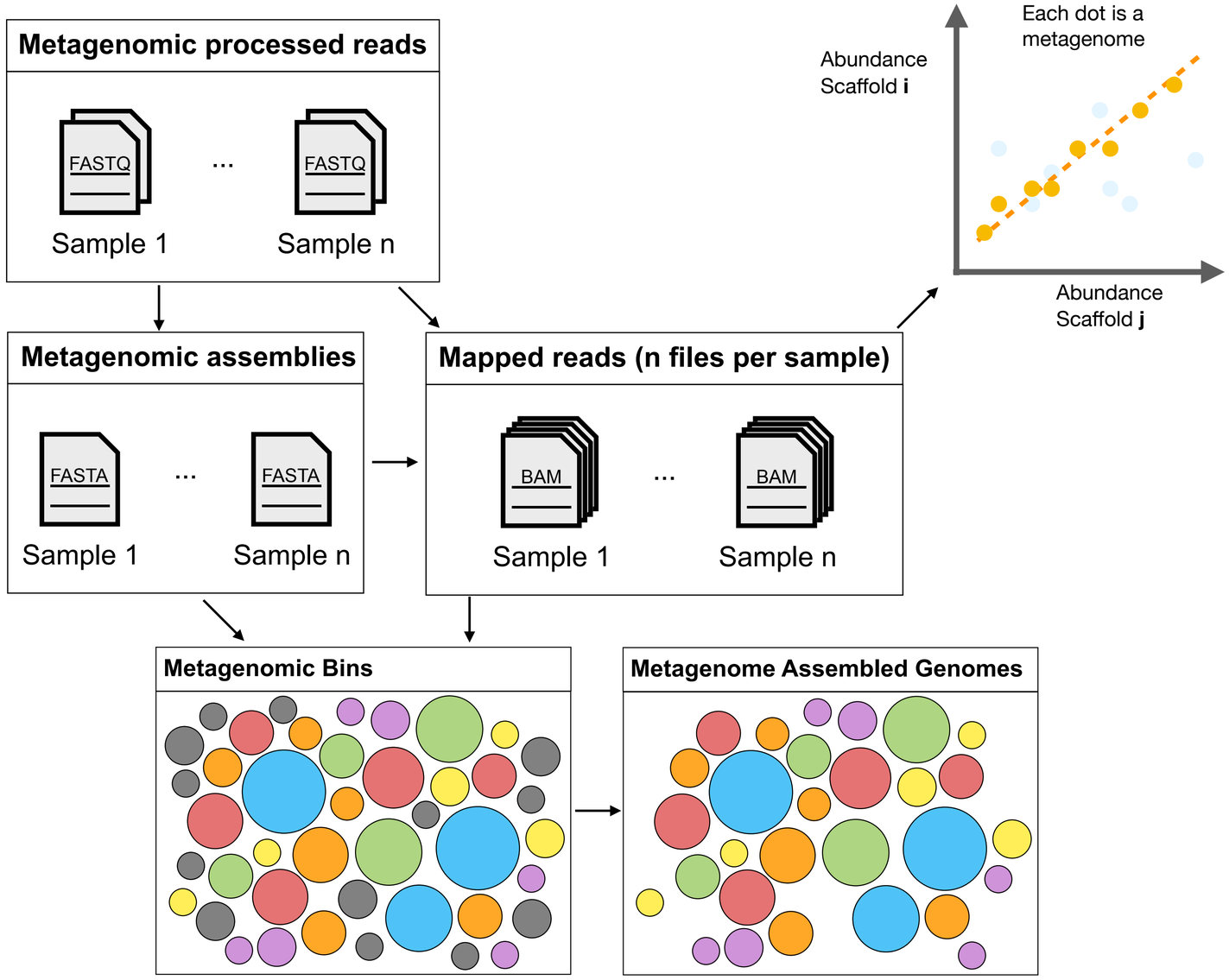
Global ocean metagenomics (>1,000 samples)



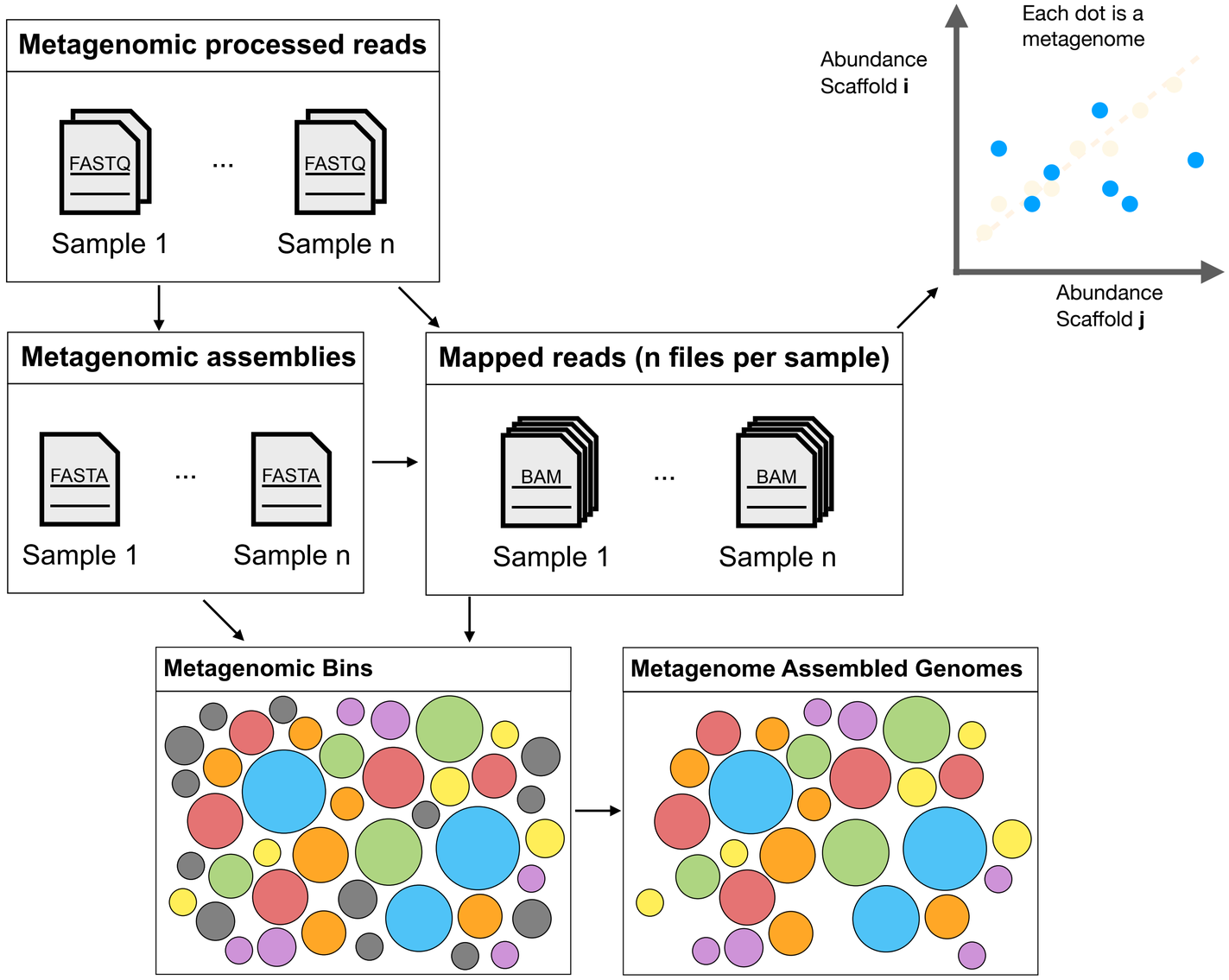
Reconstructing genomes from metagenomes (MAGs)



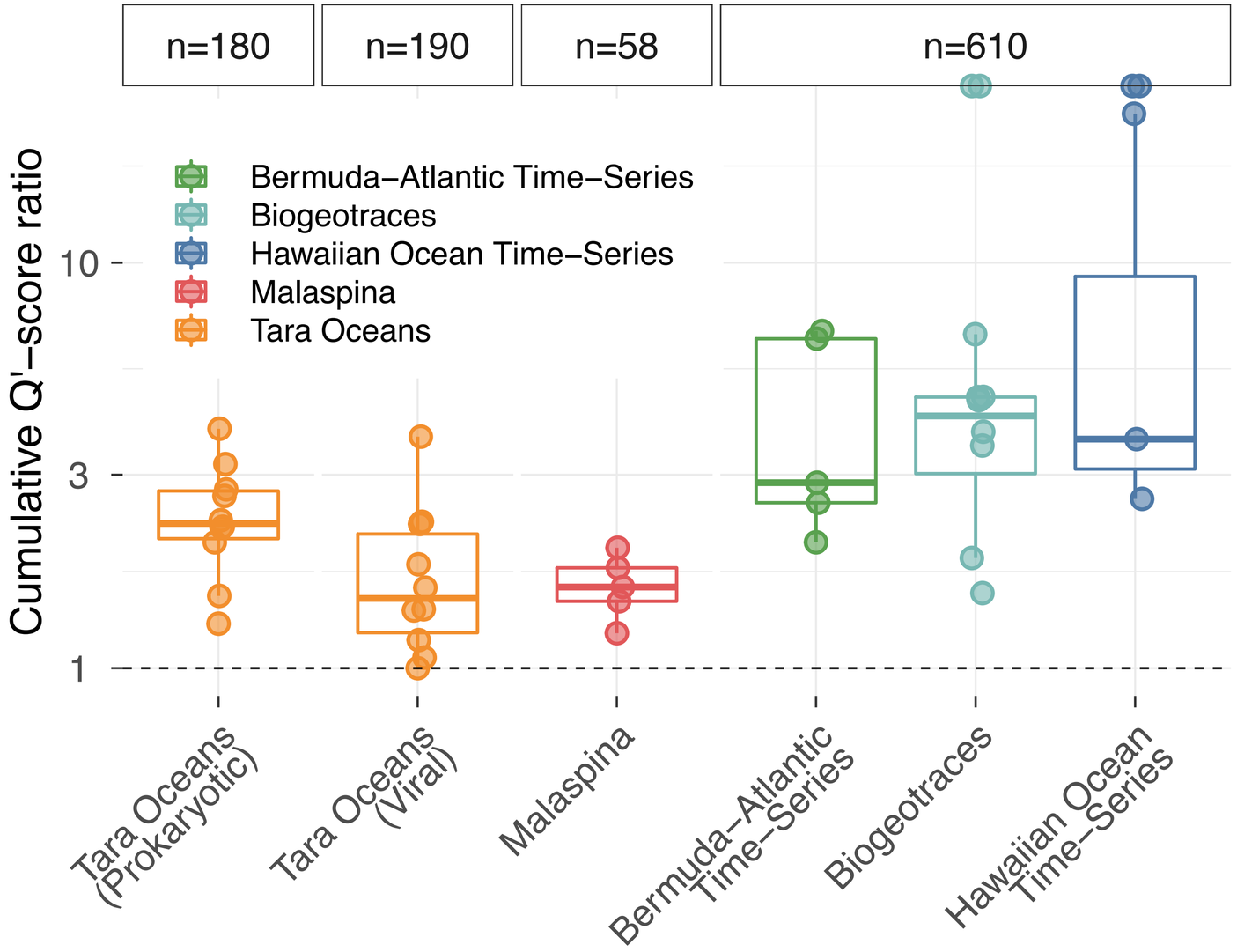
Reconstructing genomes from metagenomes (MAGs)



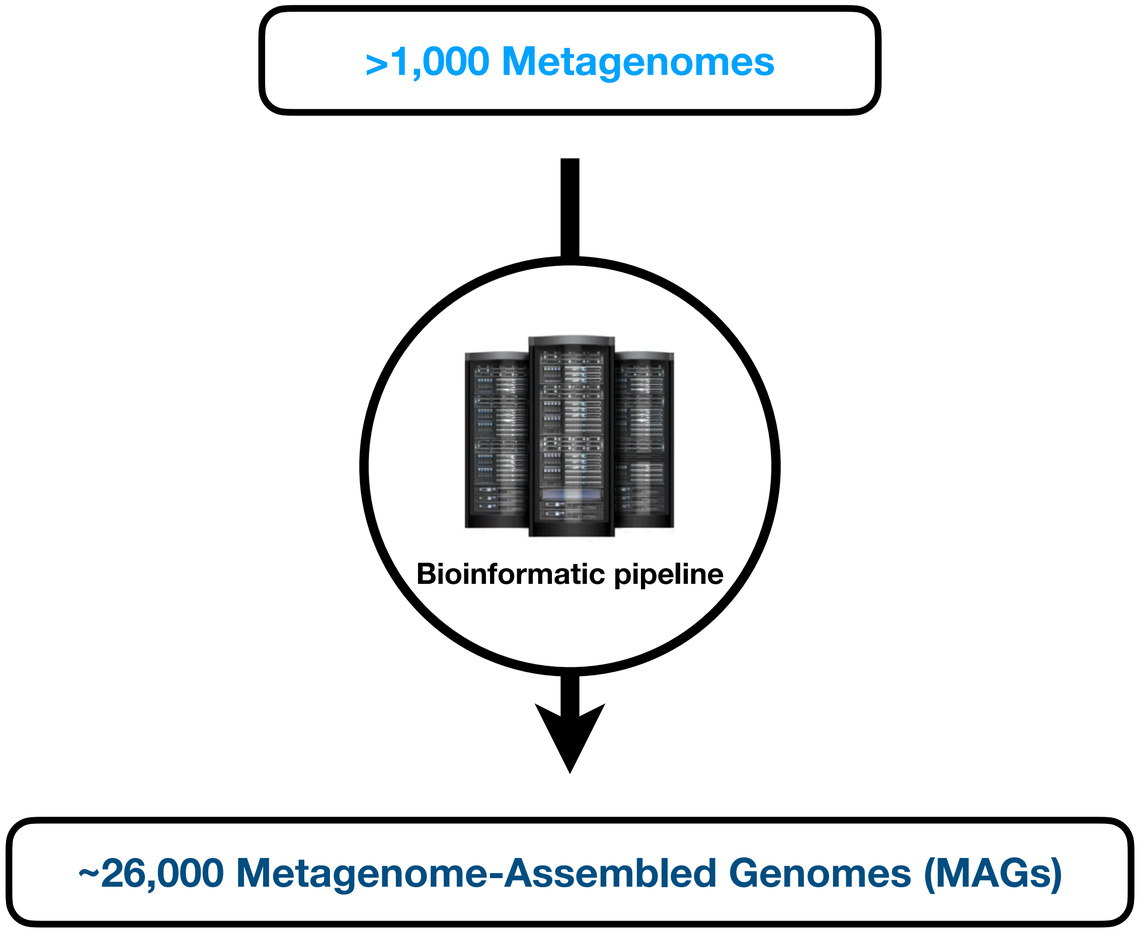
Reconstructing genomes from metagenomes (MAGs)



Abundance correlation improves binning results three-folds



Reconstructing microbial genomes from metagenomes



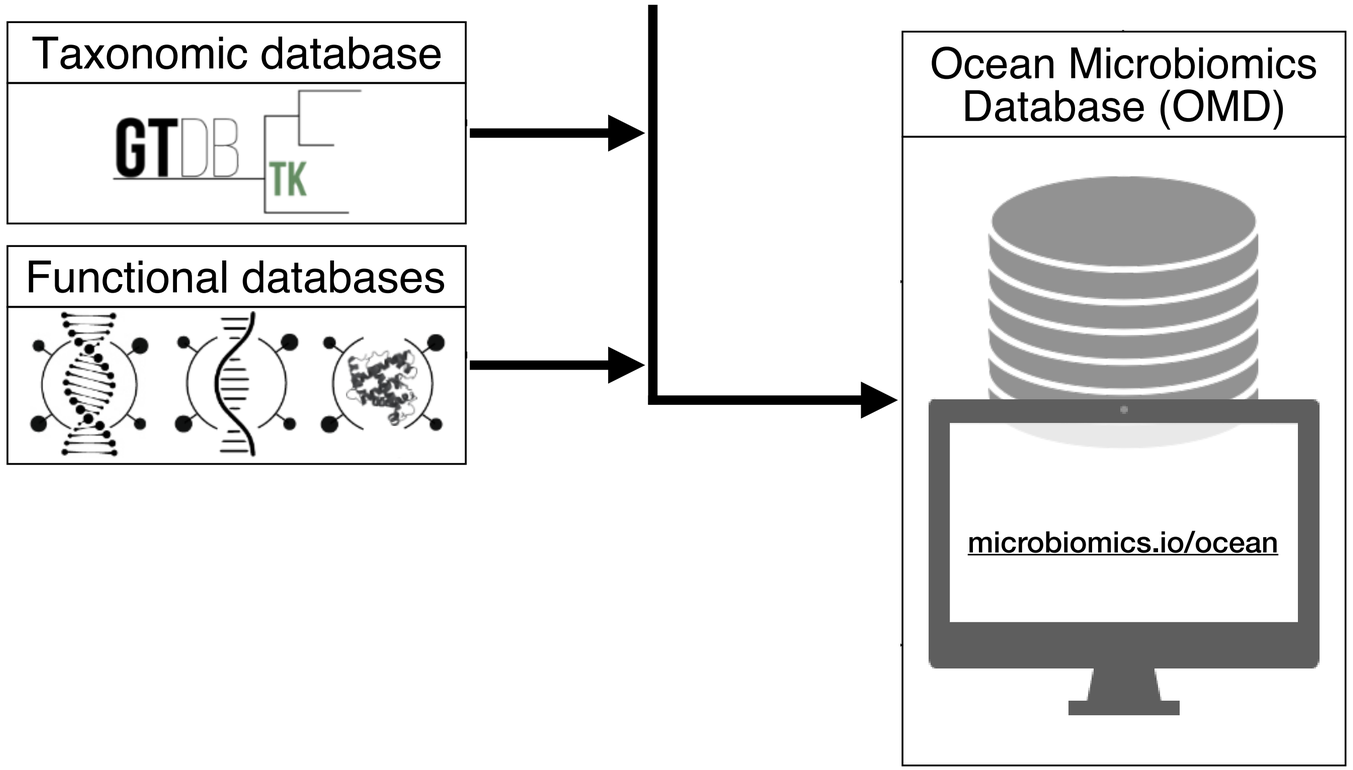
Integrating cultivation dependent and independent methods

- ➡ Newly reconstructed MAGs (~26,000
- ➡ Manually curated MAGs (~1,000)
- ➡ Single amplified genomes (SAGs) (~6,000)
- ➡ Reference genomes from isolates (~2,000)

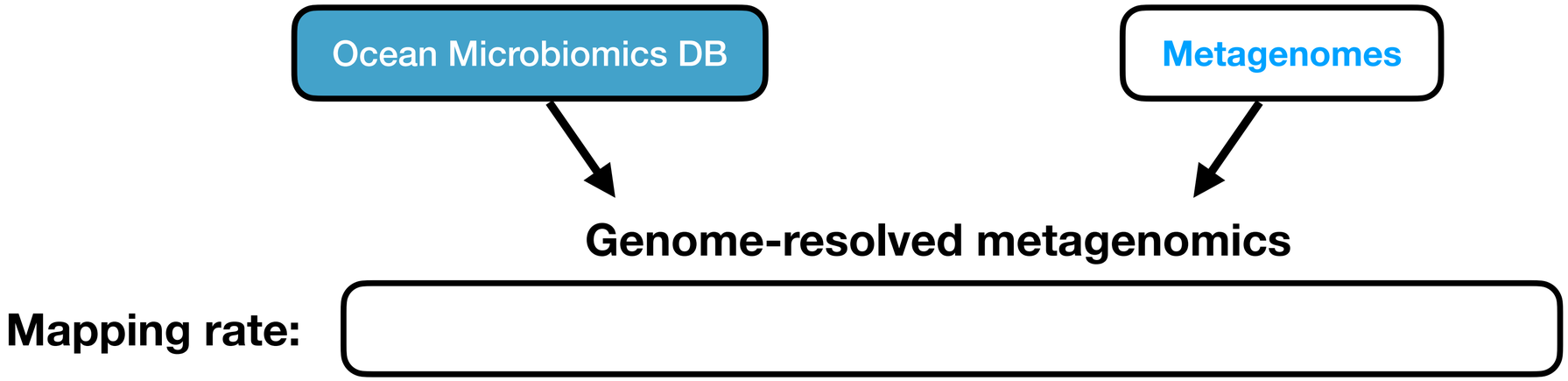
Delmont et al. 2018
Pachiadaki et al. 2019
Klementsen et al. 2018

Establishing a rich ocean microbiome resource

- ➡ Newly reconstructed MAGs (~26,000
- ➡ Manually curated MAGs (~1,000)
- ➡ Single amplified genomes (SAGs) (~6,000)
- ➡ Reference genomes from isolates (~2,000)



Improved representation of the ocean microbiome



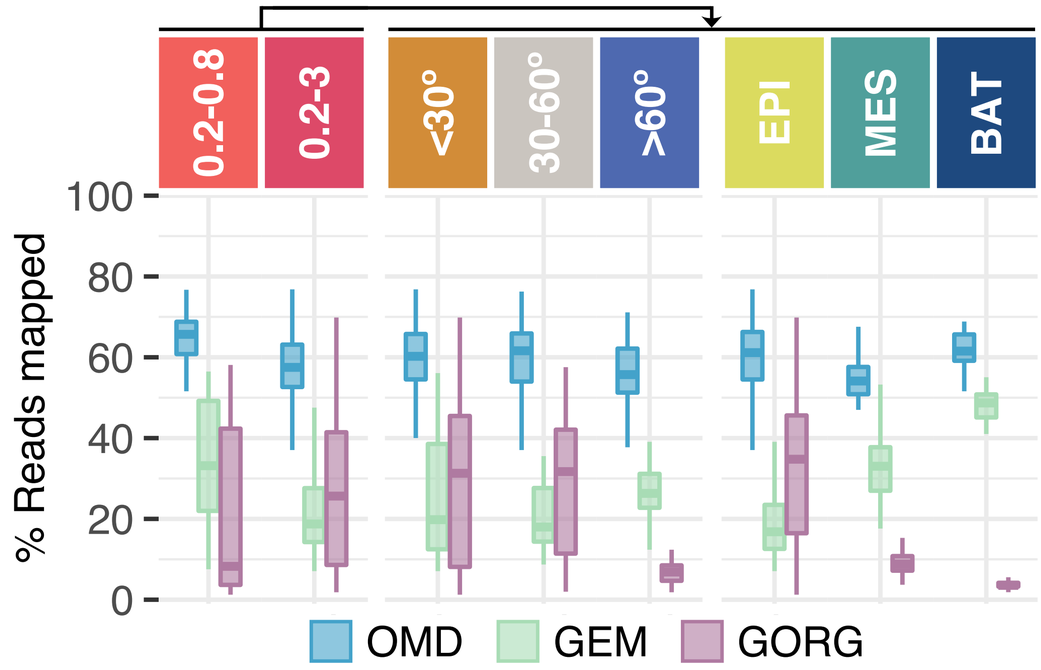
Improved representation of the ocean microbiome

Ocean Microbiomics DB

Metagenomes

Genome-resolved metagenomics

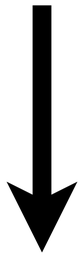
Mapping rate:



Pachiadaki et al. 2019
Nayfach et al. 2020

Providing access to its biosynthetic potential

~35k genomes



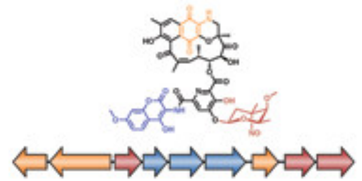
~40k Biosynthetic Gene Clusters (BGCs)

➡ **How unique is it compared to currently sequenced microbes?**

Compared to sequenced and characterised biosynthetic potential

NCBI

RefSeq
~200,000 genomes



Characterised
biosynthetic pathways

1.2 M BGCs

BIGFAM database

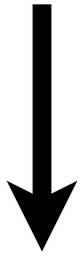
2,000 BGCs

MIBiG

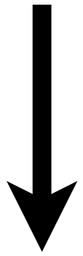
Kautsar et al. 2020
Kautsar et al. 2021

Grouping BGCs into relevant units

~35k genomes



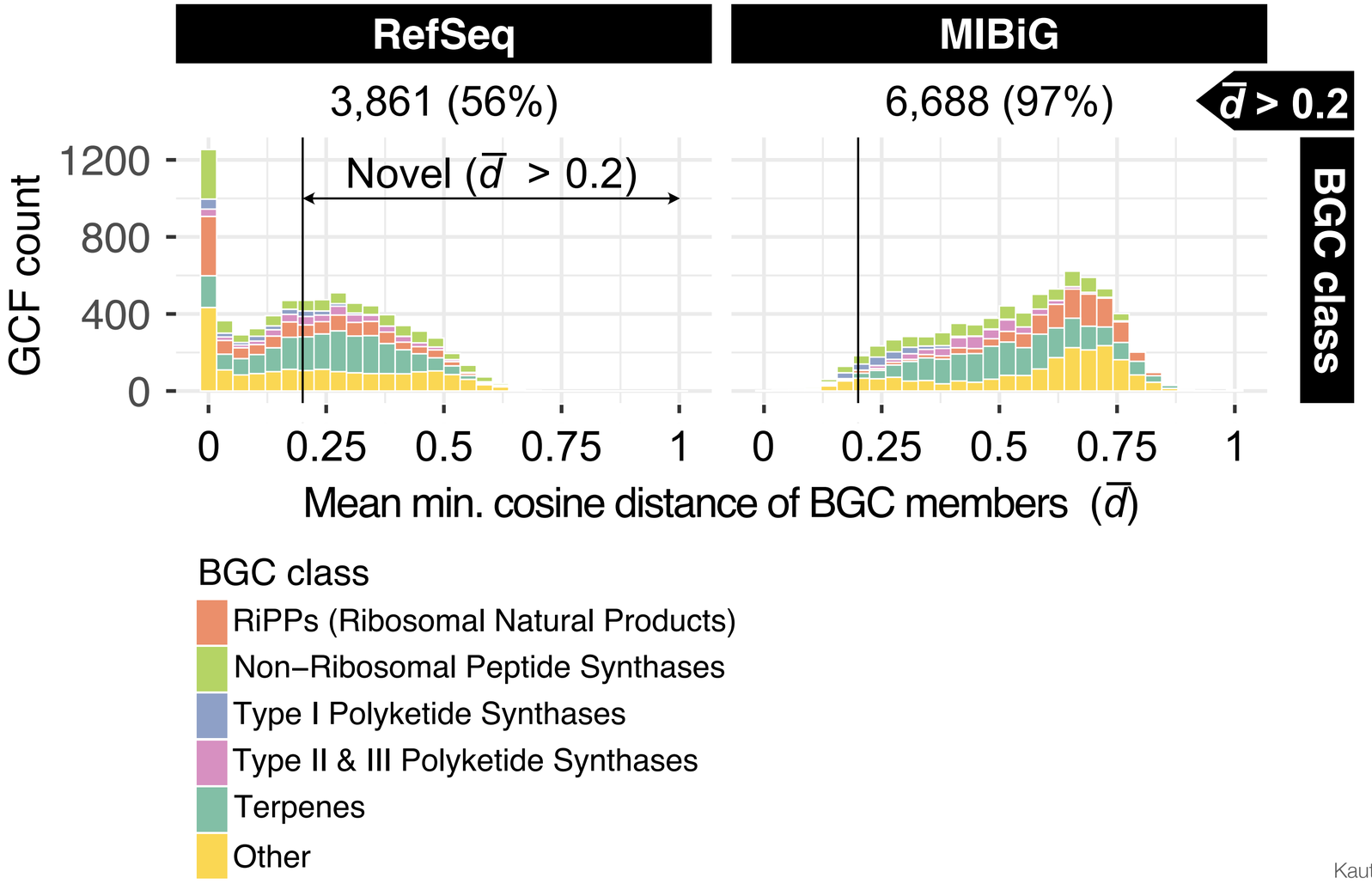
~40k Biosynthetic Gene Clusters (BGCs)



~7k Gene Cluster Families (GCFs)

Kautsar et al. 2021

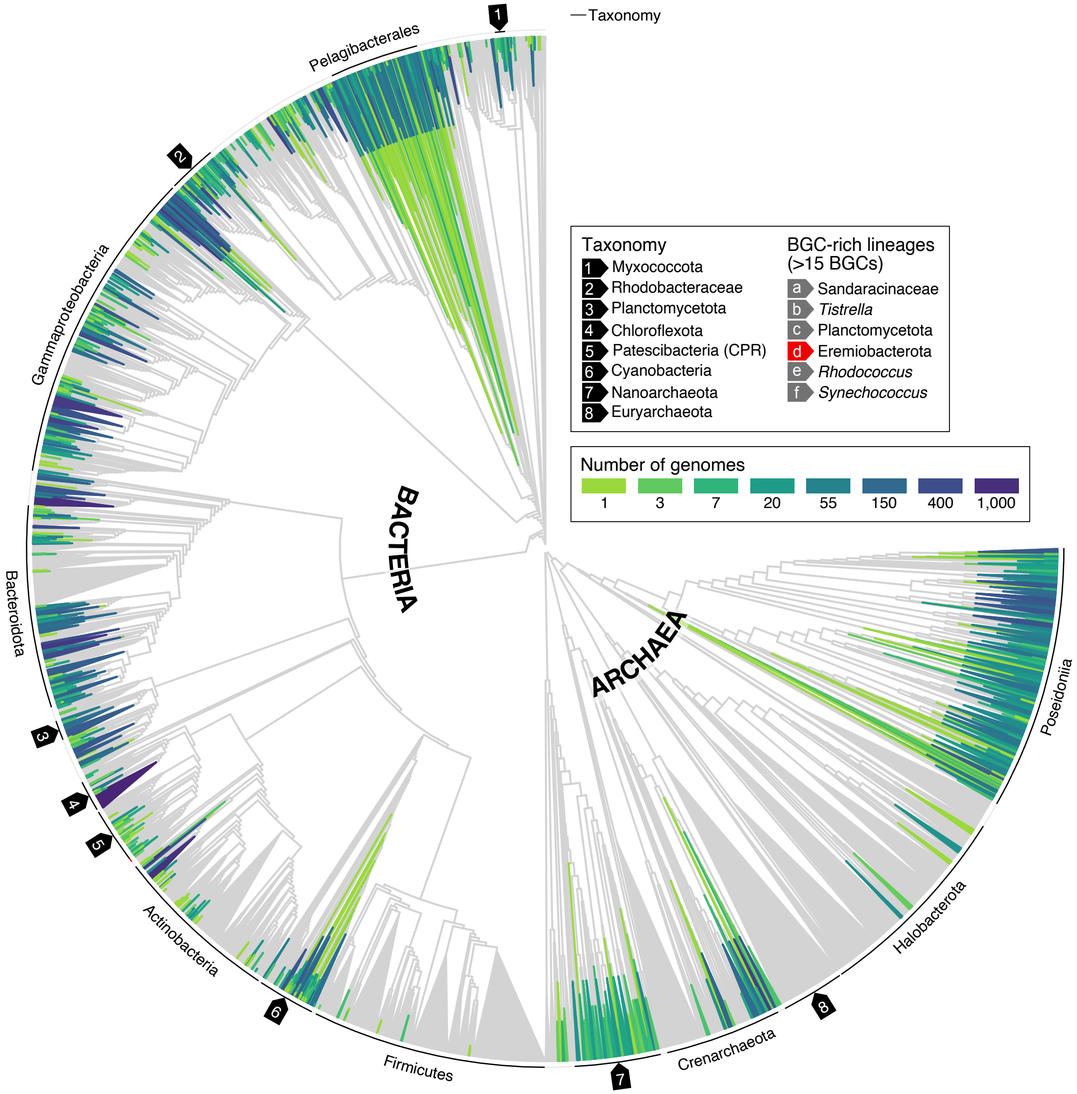
With large potential for new compounds



Kautsar et al. 2020
Kautsar et al. 2021

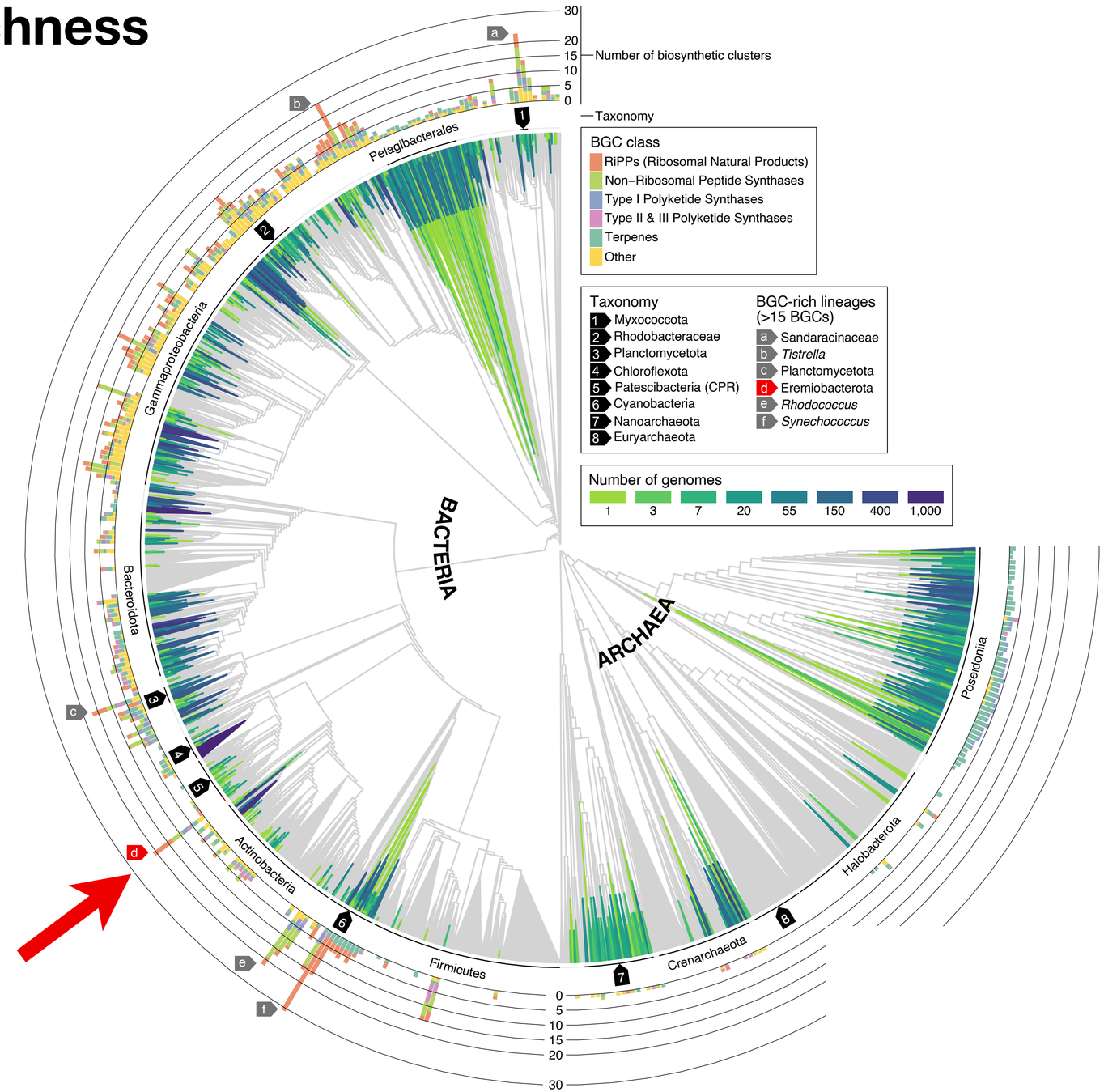
➡ **Are there BGC-rich microbial lineages to be discovered in the ocean?**

Phylogenomic distribution of the ocean biosynthetic potential



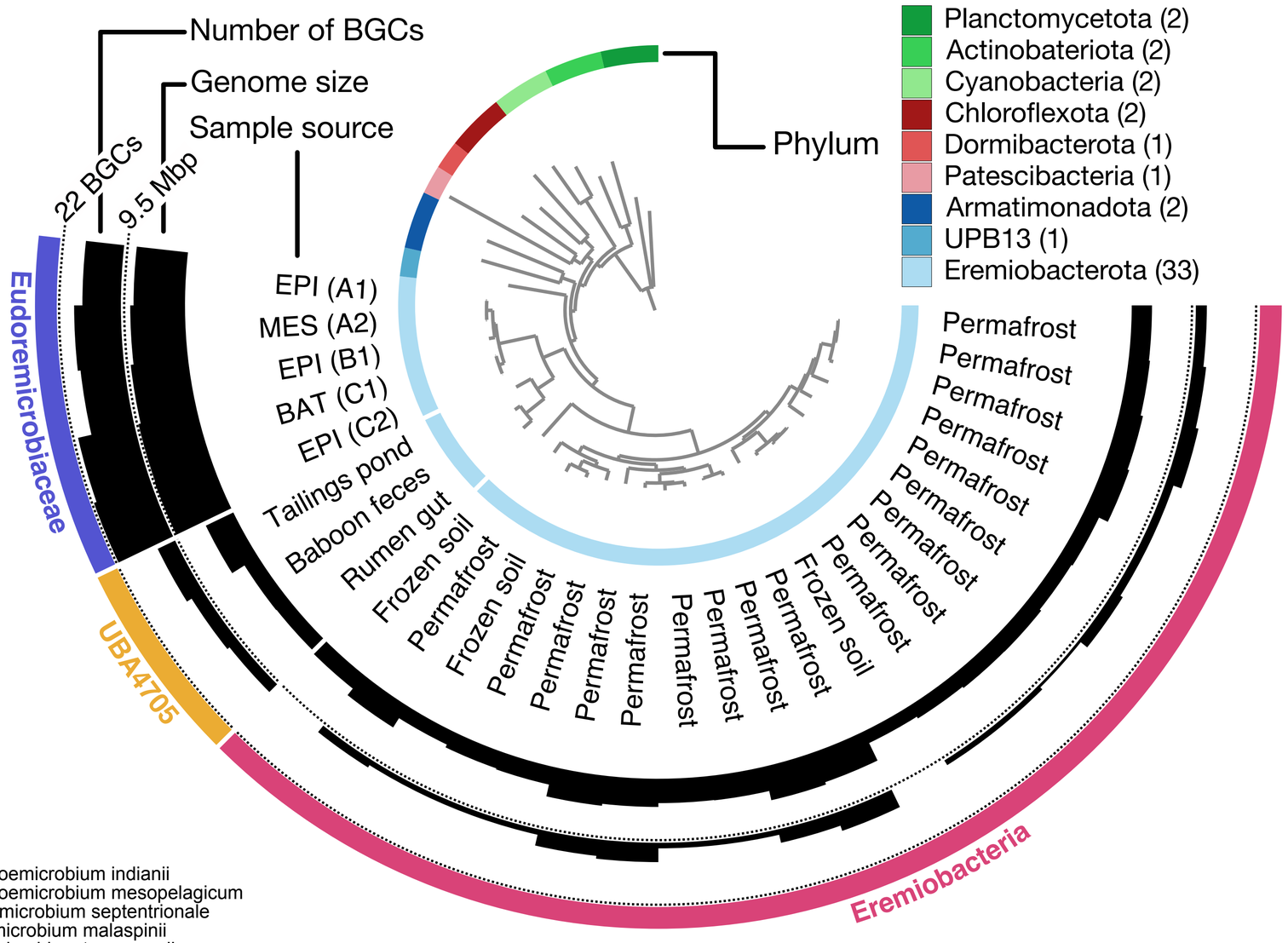
Parks et al. 2018

Eremiobacterota, uncultivated phylum with unsuspected BGC richness



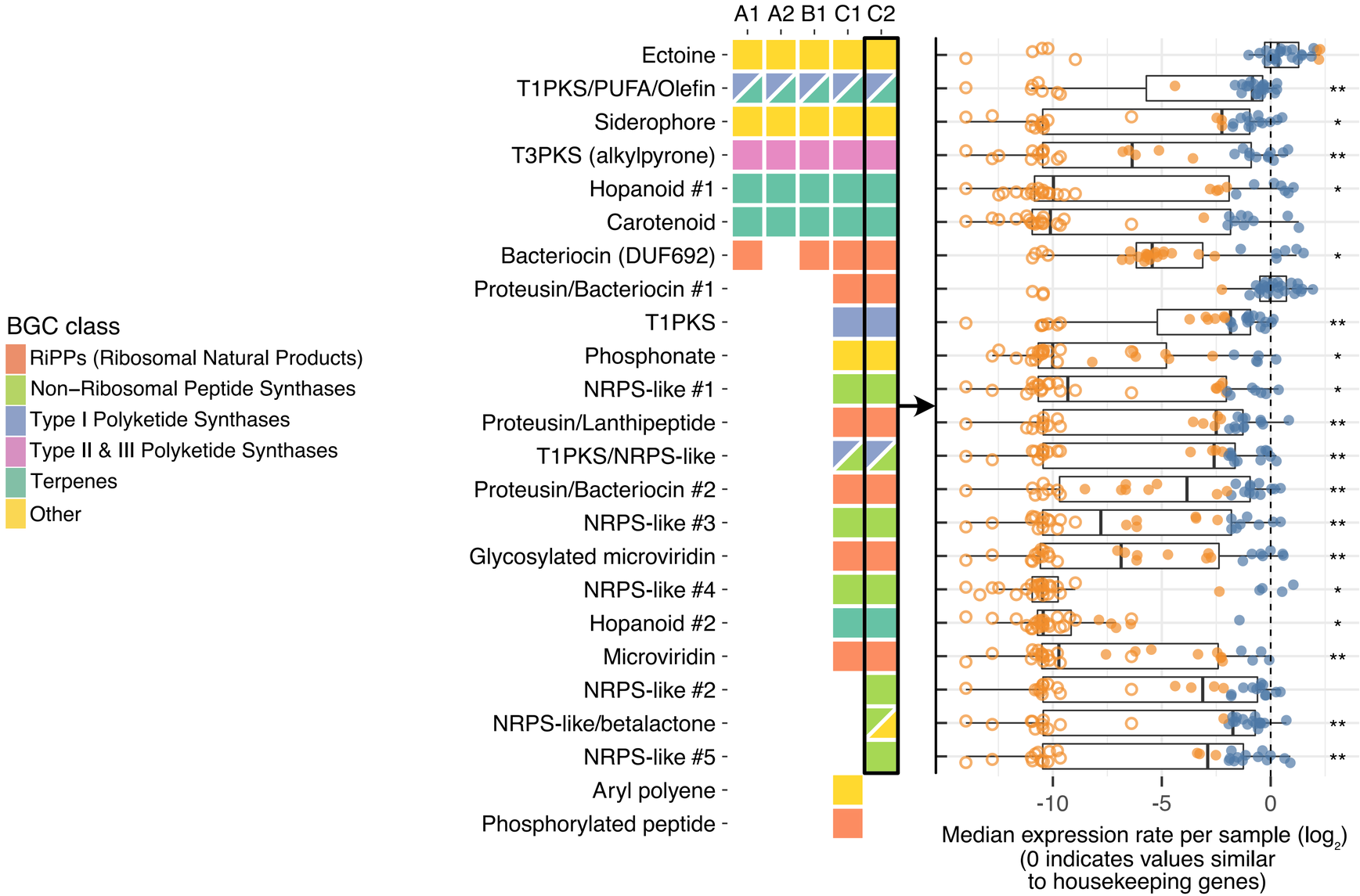
Parks et al. 2018

A novel marine Eremiobacterota lineage: The candidate family Eudoremicrobiaceae

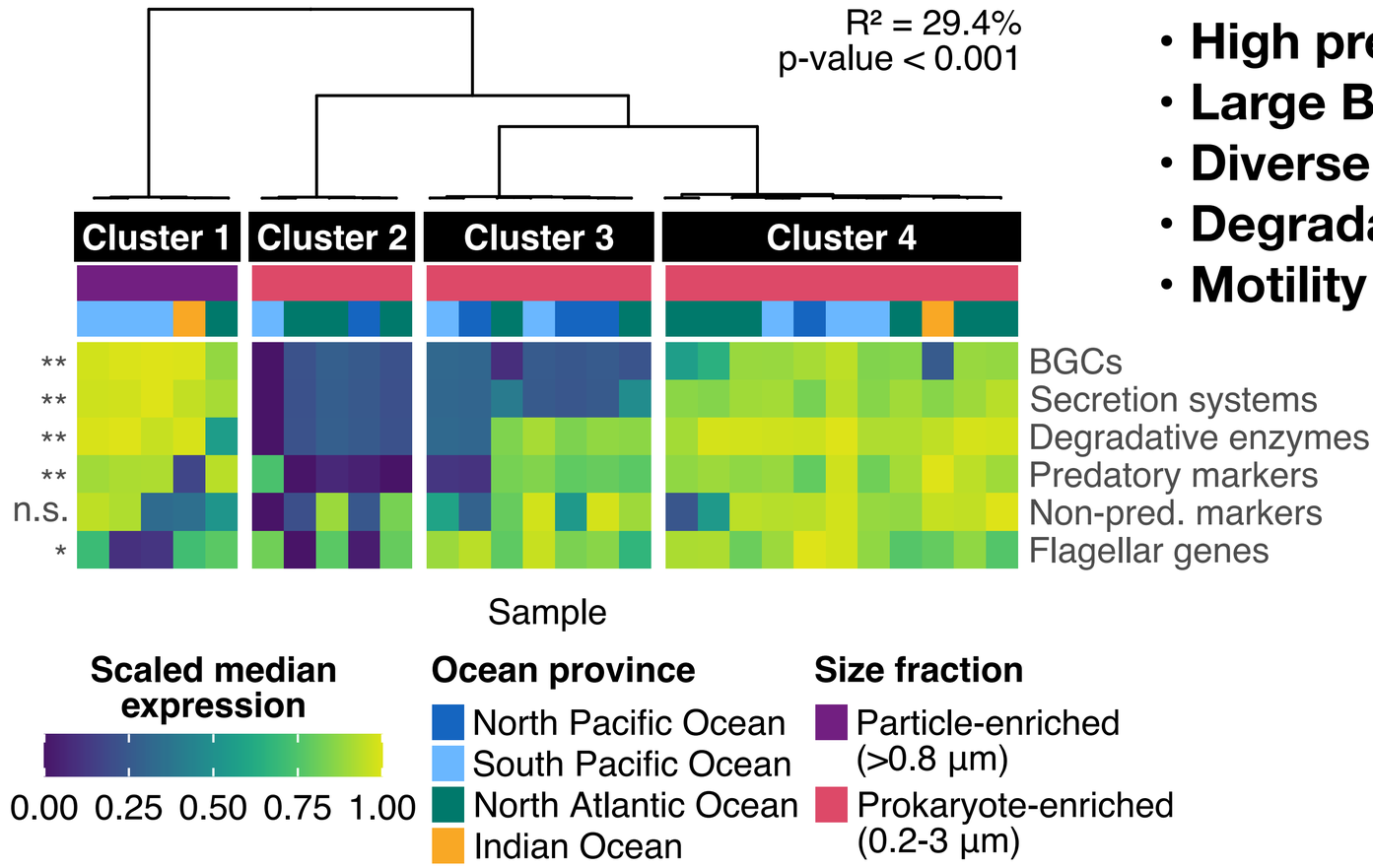


A1 *Ca.* Amphithoemicrobium indianii
A2 *Ca.* Amphithoemicrobium mesopelagicum
B1 *Ca.* Autonoemicrobium septentrionale
C1 *Ca.* Eudoremicrobium malaspinii
C2 *Ca.* Eudoremicrobium taraoceanii

The newly identified bacterial family has a diverse and actively expressed BGC repertoire



A biosynthetic potential that may supports a putative predatory lifestyle

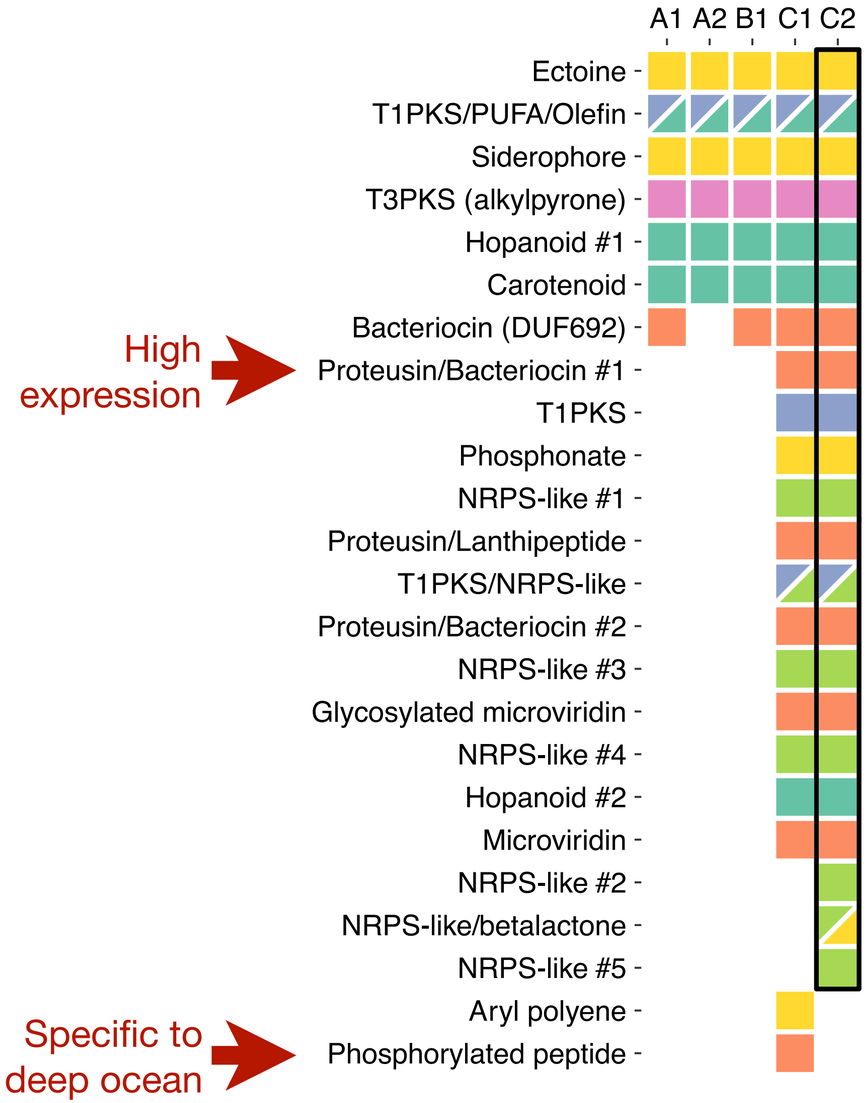


- **High predatory scores**
- **Large BGC diversity**
- **Diverse secretion systems**
- **Degradative enzymes**
- **Motility**

Pasternak et al. 2012

➡ **Is this computational approach sufficiently powerful to predict new enzymology and natural products?**

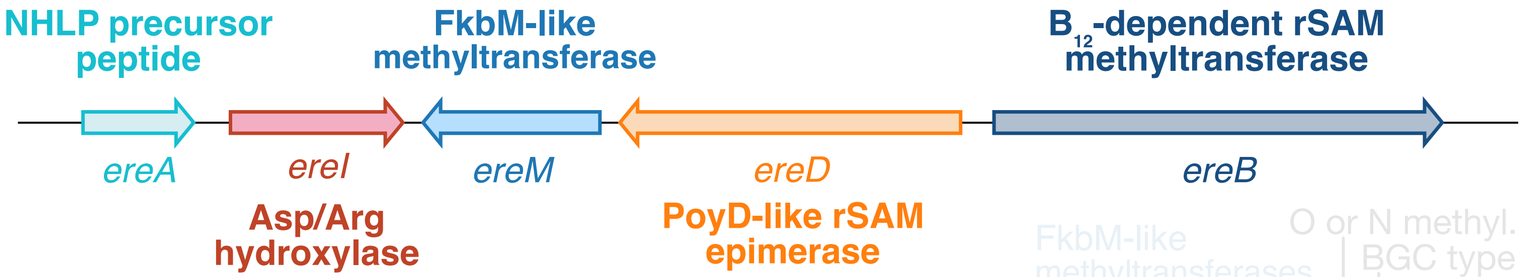
Probing Eremiobacterota's biosynthetic potential for new enzymes and natural products



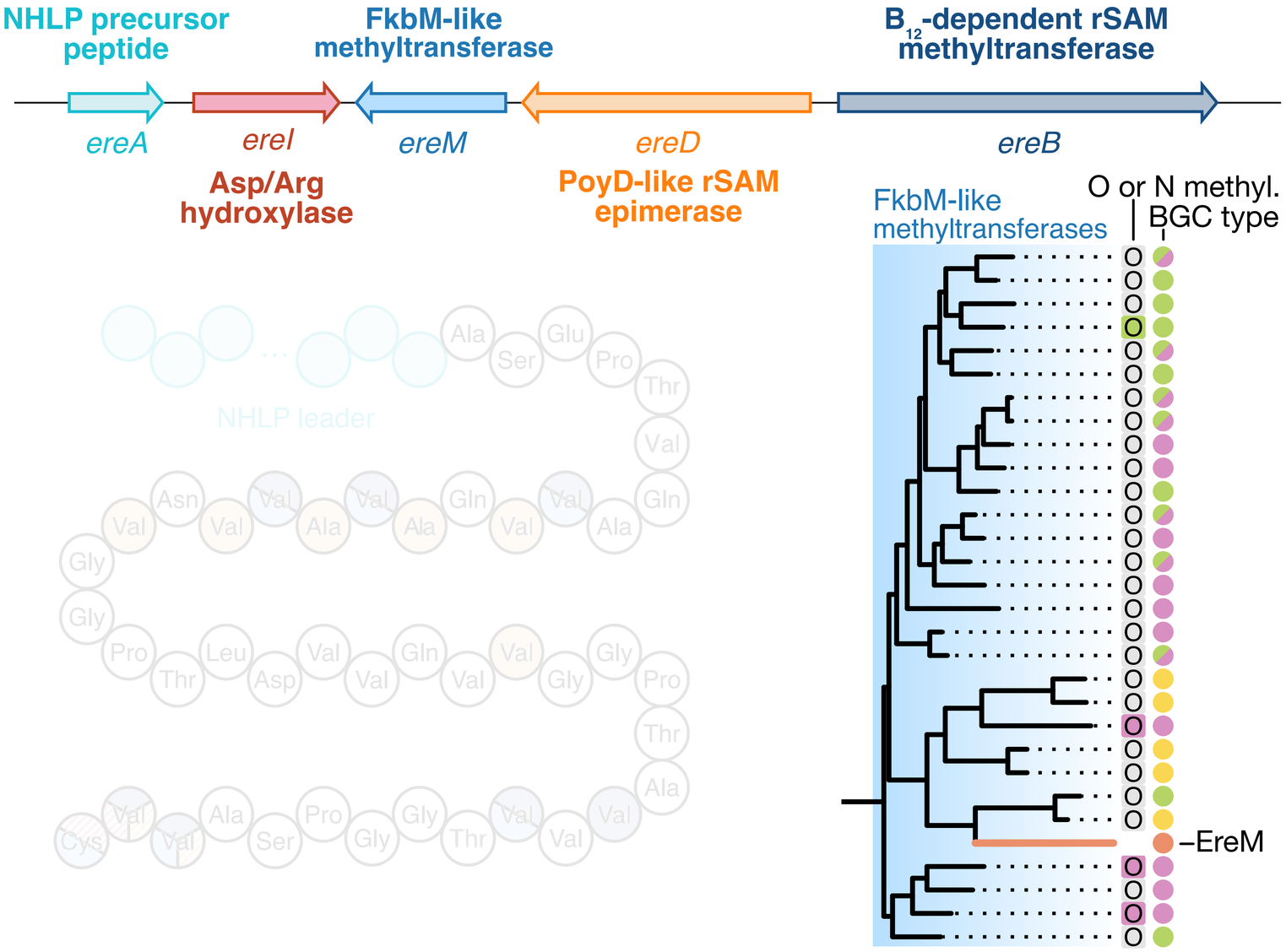
Characterising two **predicted novel** Ribosomal BGCs using:

- Non-standard heterologous expression
- Tandem Mass Spectrometry (MS/MS)
- Isotope labelling
- Nuclear Magnetic Resonance (NMR)

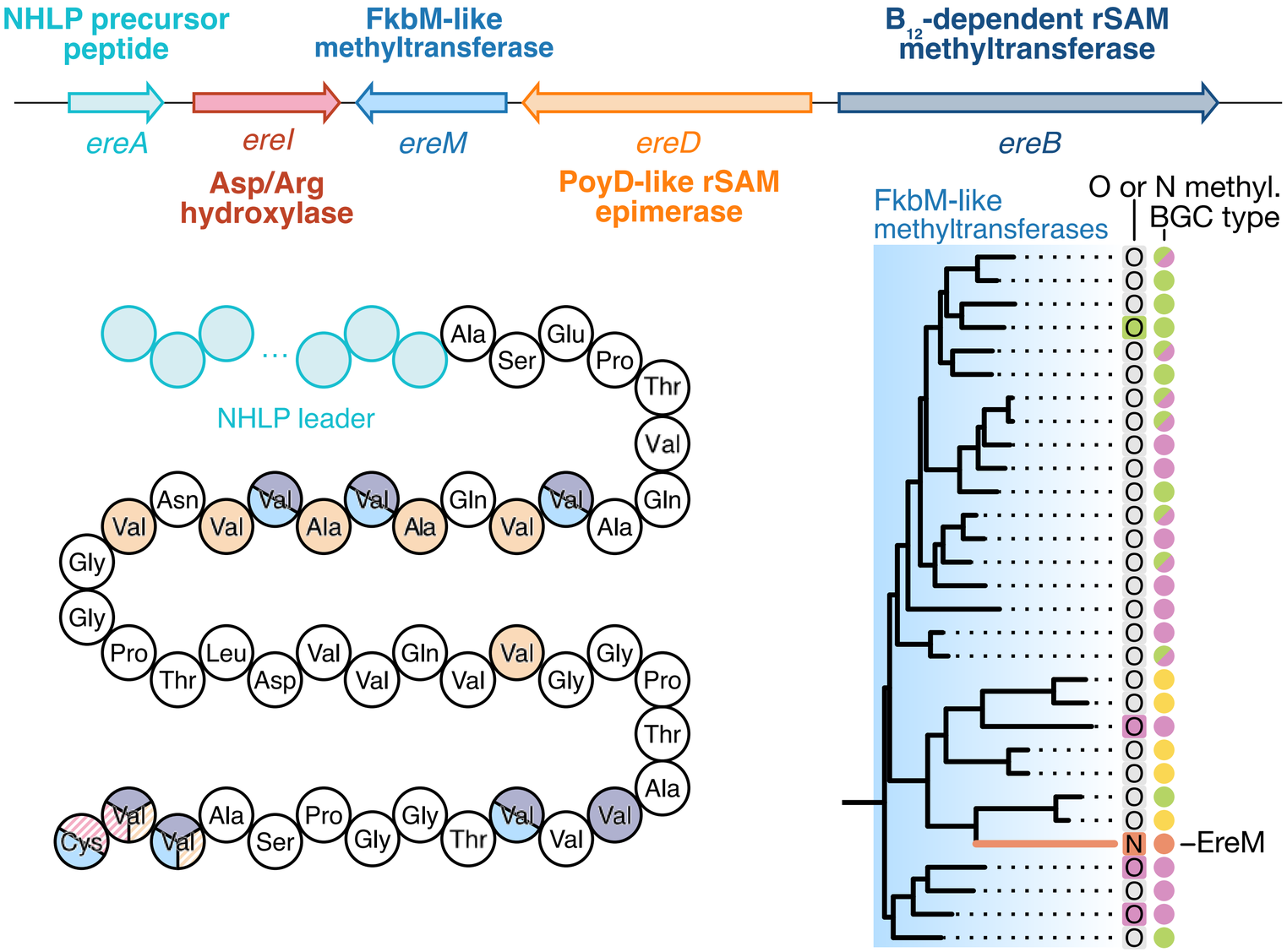
Intricate proteusin cluster reveals new enzymology



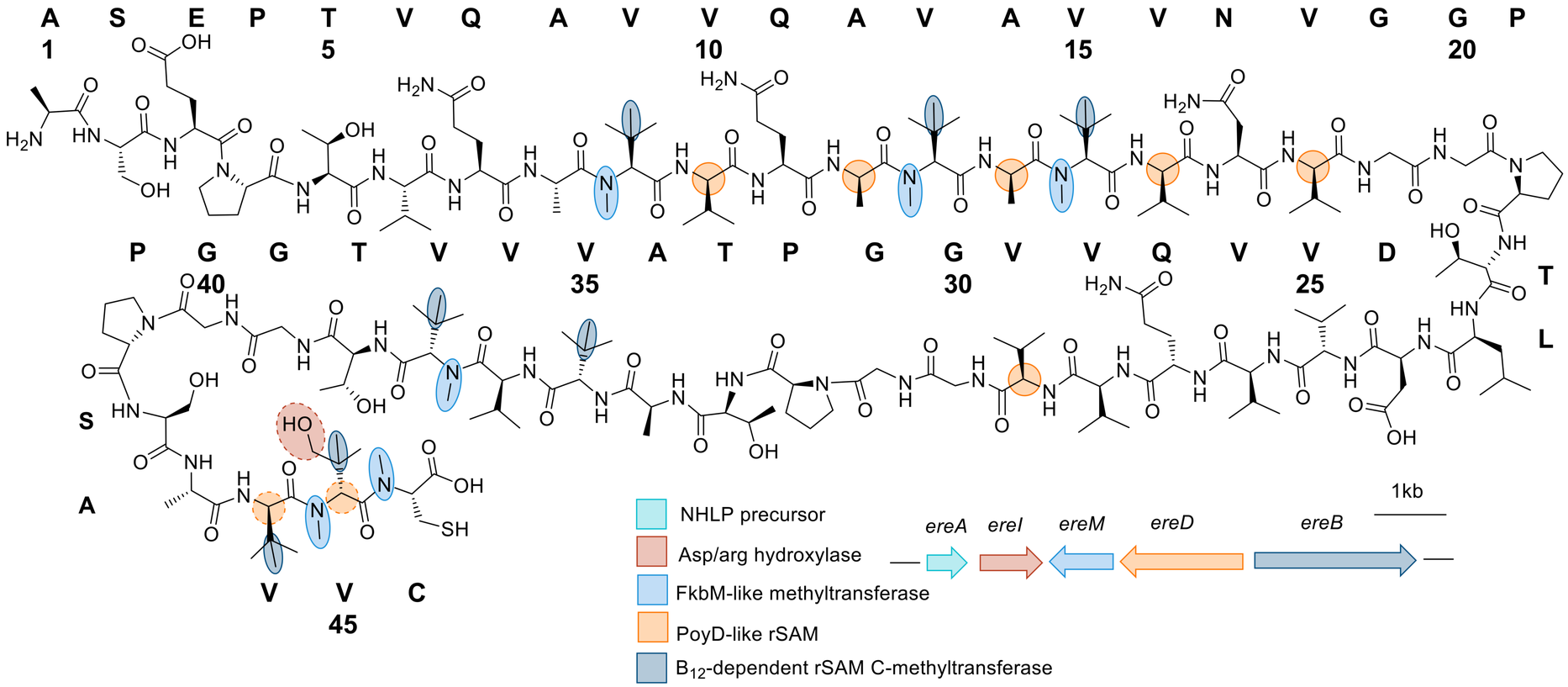
Intricate proteusin cluster reveals new enzymology



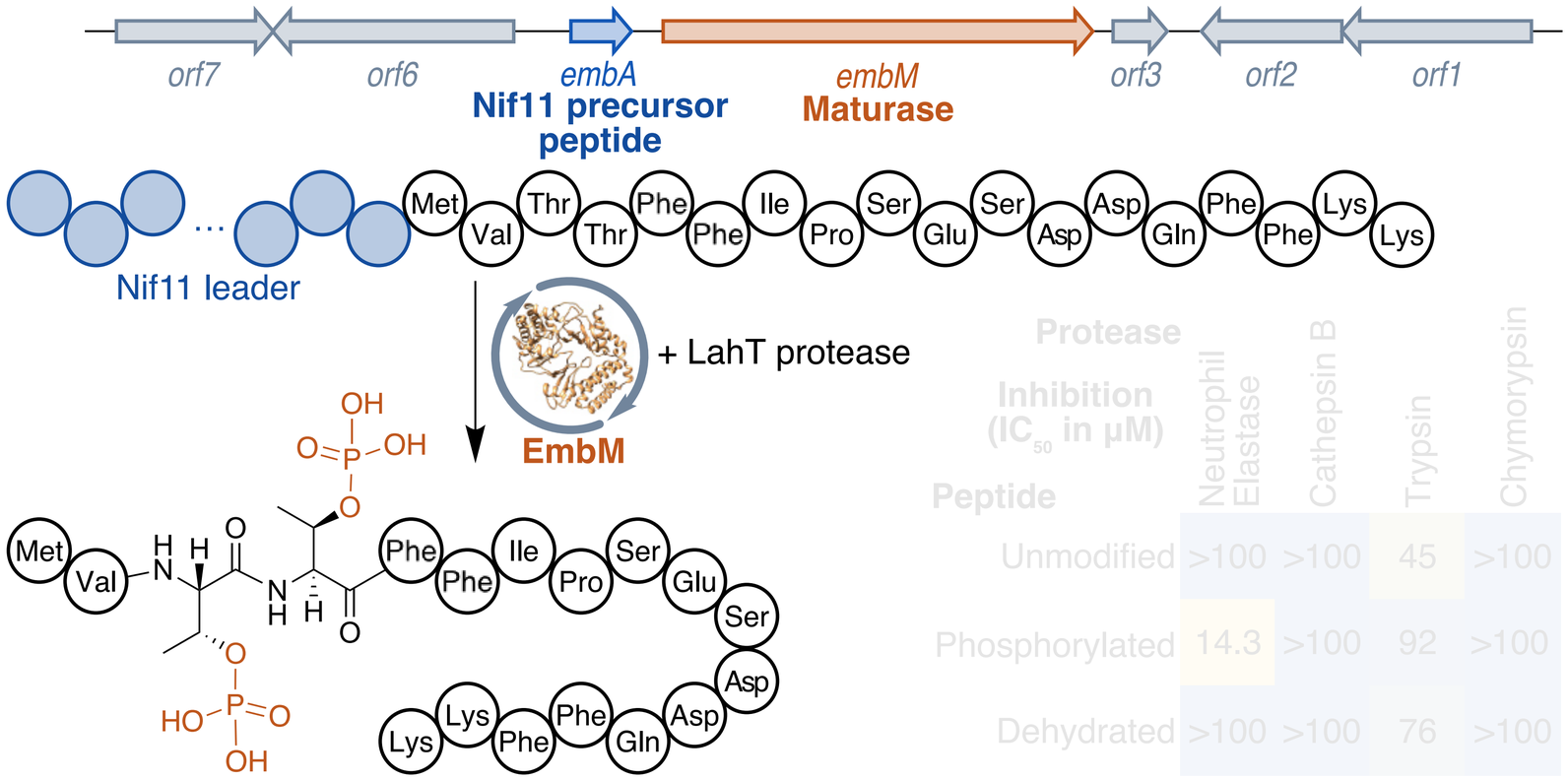
Intricate proteusin cluster reveals new enzymology



A predicted O-methyltransferase with amide-N-methylation activity

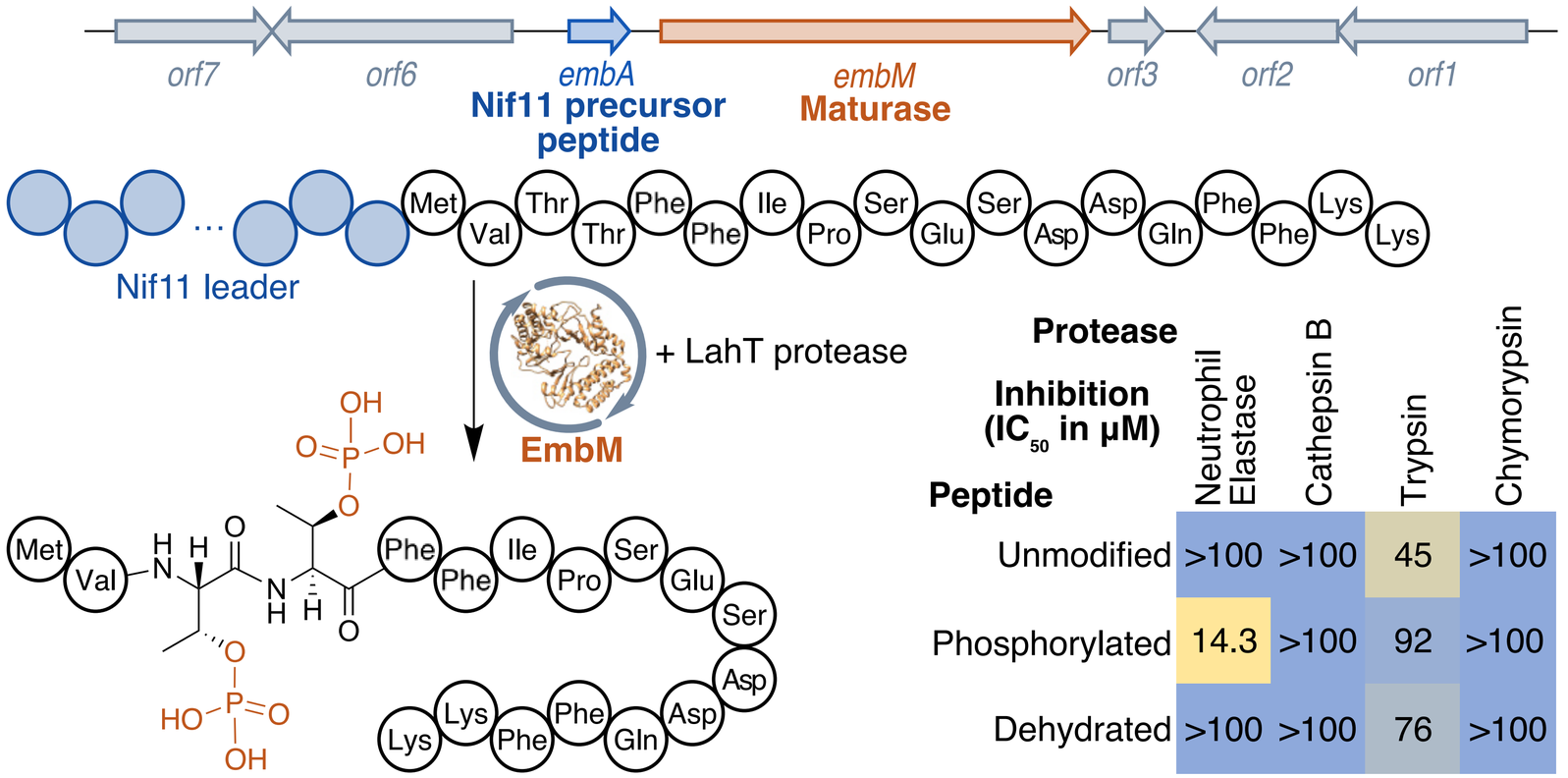


A new RiPP cluster with phosphorylation as sole modification



Protease Inhibition (IC_{50} in µM) Peptide	Neutrophil Elastase	Cathepsin B	Trypsin	Chymorypsin
Unmodified	>100	>100	45	>100
Phosphorylated	14.3	>100	92	>100
Dehydrated	>100	>100	76	>100

With Human Elastase inhibition



Protease / Inhibition (IC_{50} in μM) / Peptide	Neutrophil Elastase	Cathepsin B	Trypsin	Chymorypsin
Unmodified	>100	>100	45	>100
Phosphorylated	14.3	>100	92	>100
Dehydrated	>100	>100	76	>100

Conclusions

- **Genome-resolved microbiomics** as a mean to **explore environmental microbiomes** and **discover novel enzymology** and **natural products**
- This approach provides **evolutionary and ecological context** to the **biosynthetic potential**
- Bioinformatics-guided **experimental characterisation** is **necessary** and can still lead to **unpredicted discoveries**

Thank you for your attention

Sunagawa Lab

ETH zürich

SIB Swiss Institute of Bioinformatics

FNSNF
Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

Helmut Horten Stiftung

TARA OCEANS

Fondation tara océan
explorer et partager

ETHZ, Institute of Microbiology:

Jörn Piel

Serina Robinson

Clarissa Forneris

Florian Hubrich

Questions?

Image: François Aurat