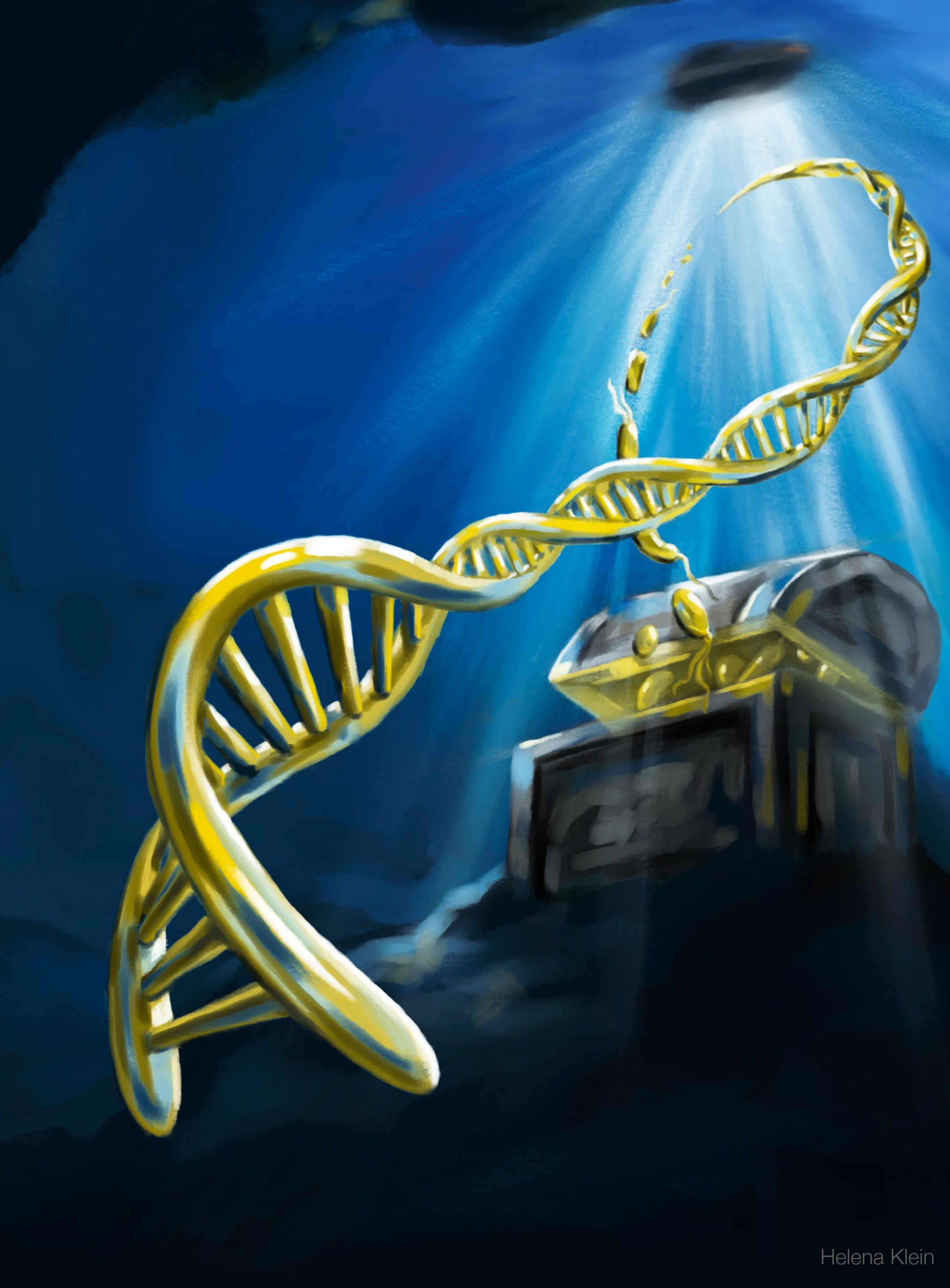


Mining microbiomes for future therapeutics

Dr. Lucas Paoli

Group Leader
Microbiome Immunity and
Ecology Lab



1. Why do we want to mine microbiomes?

Motivation

2. How do we mine microbiomes?

Methods

3. Mining the global ocean microbiome

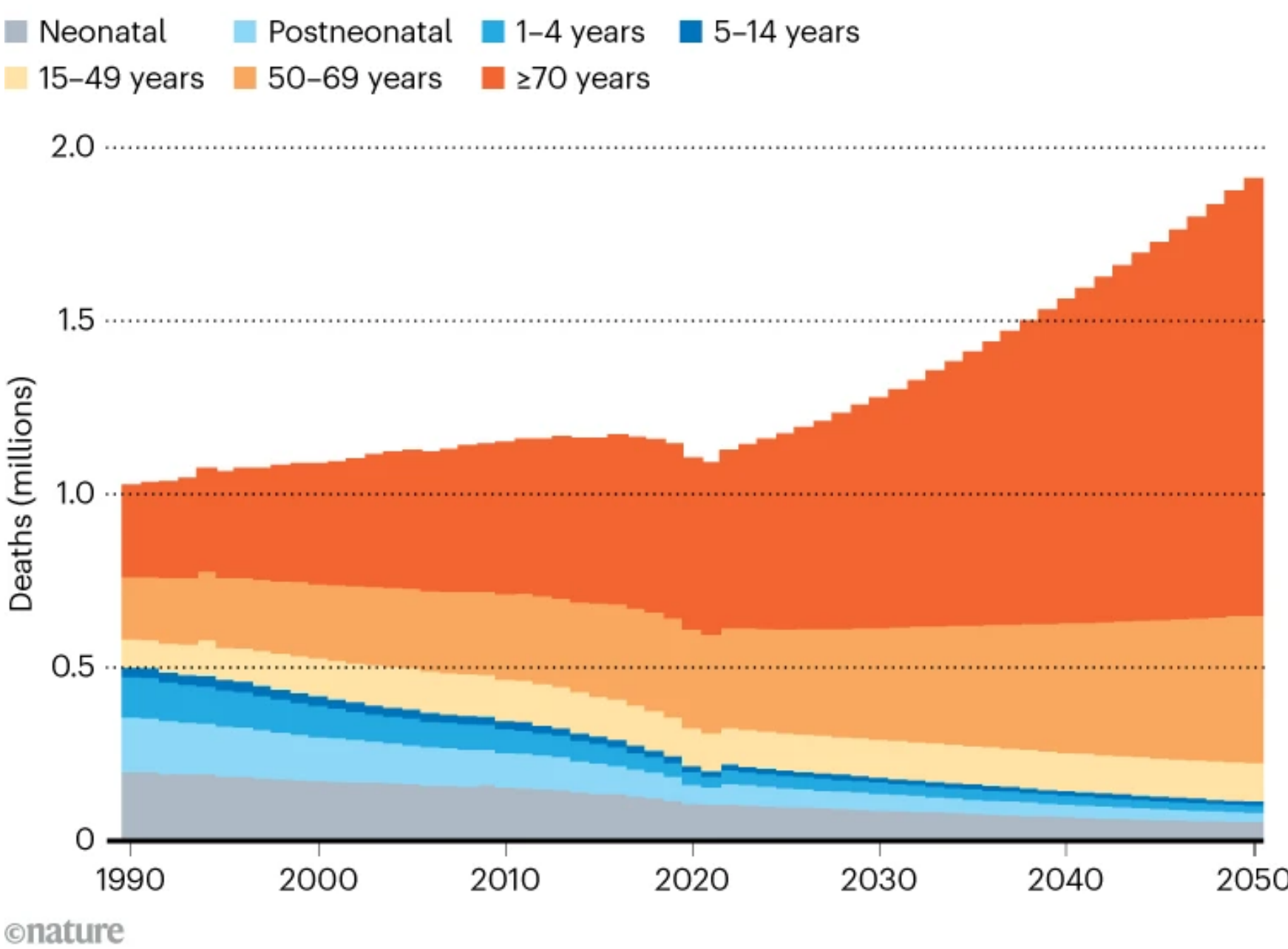
Results

Why do we want to mine microbiome?

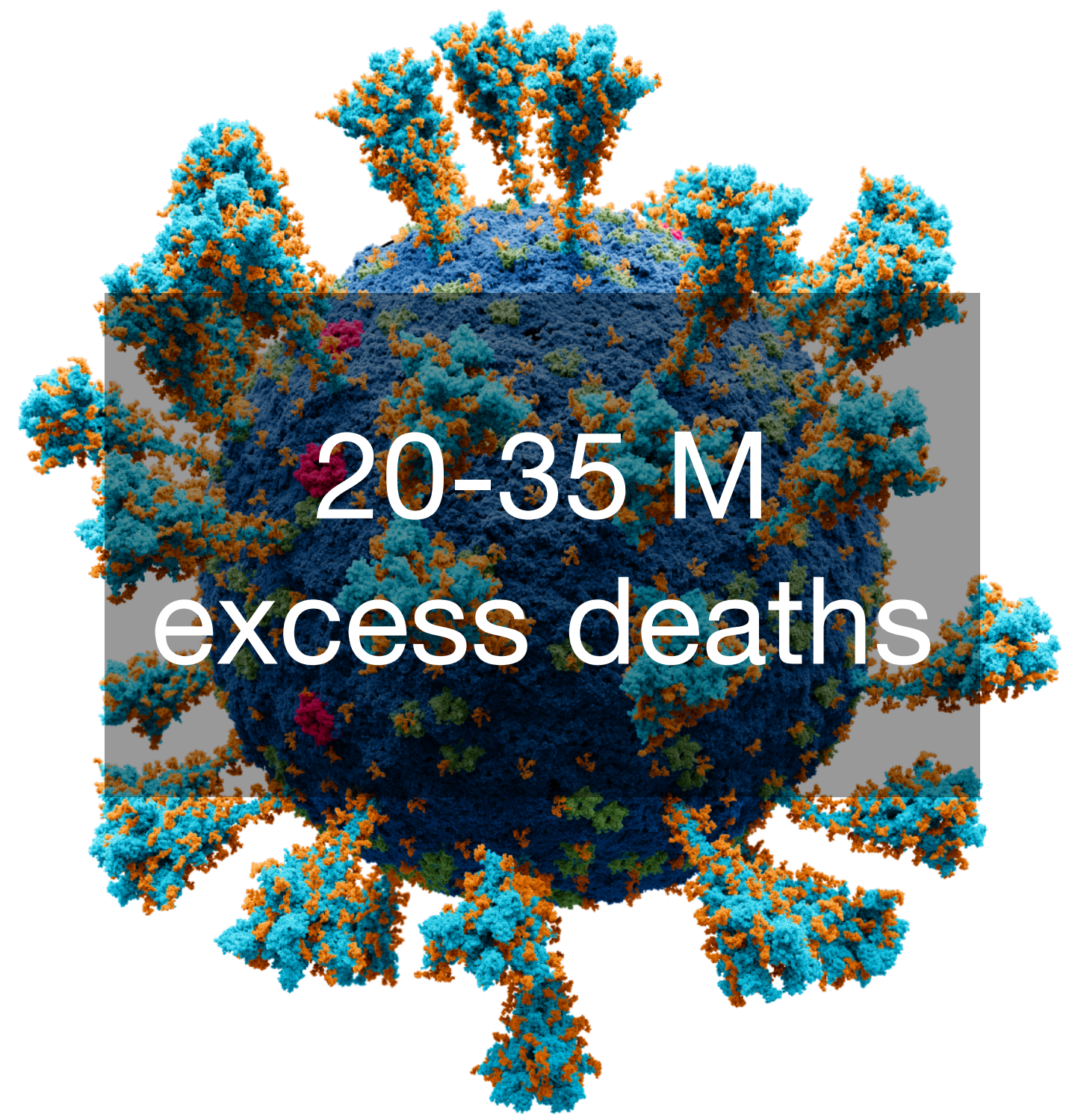
Therapeutic needs

RESISTANCE CRISIS

By 2050, antimicrobial resistance could be responsible for 1.91 million deaths per year. Mortality is projected to rise by around 70% among people aged 70 and older, but will continue to fall in young children and babies.

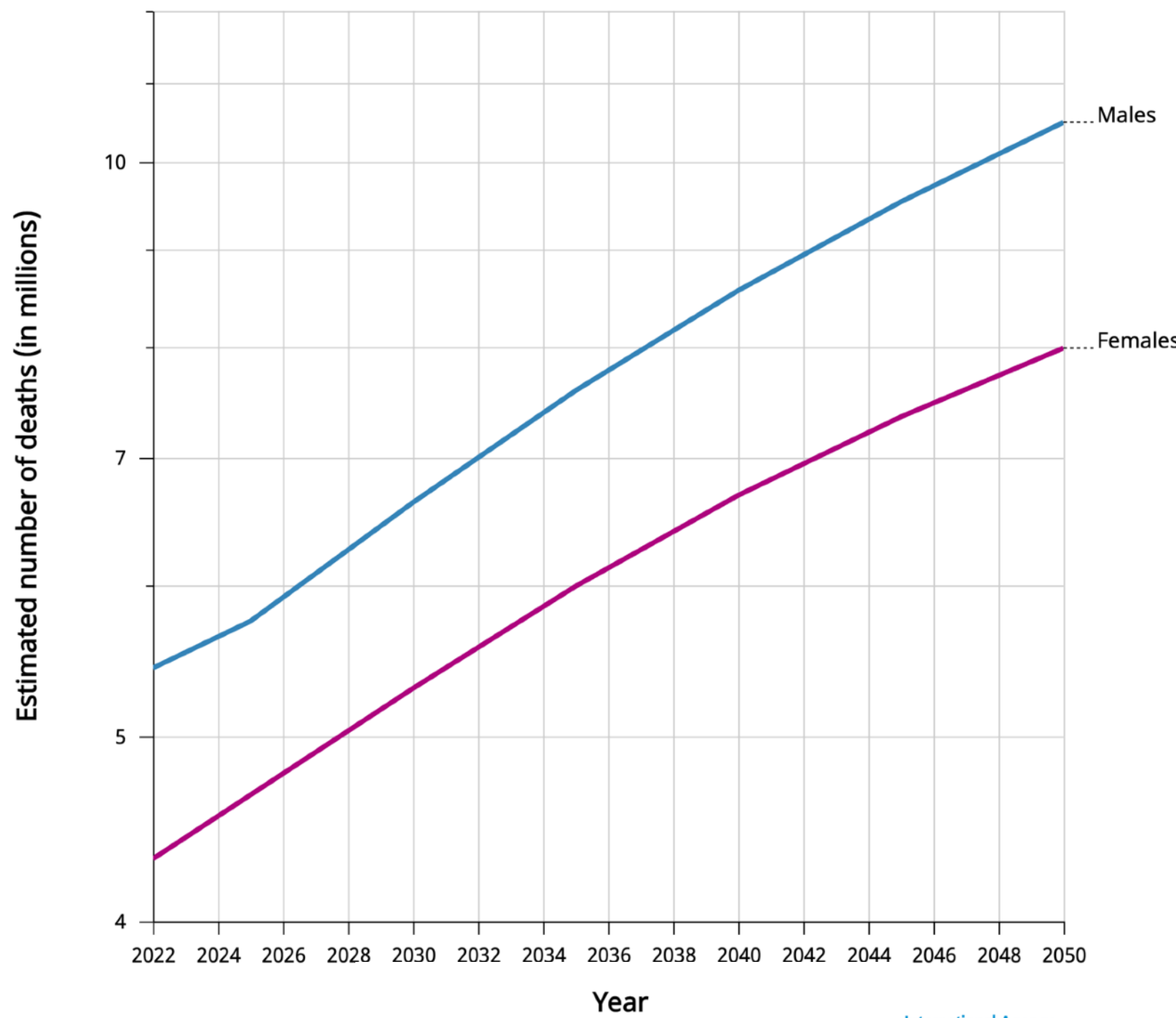


Antibiotics



Antivirals

Estimated number of deaths from 2022 to 2050, Males and Females, age [0-85+]
All cancers



Cancer Tomorrow | IARC - <https://gco.iarc.who.int/tomorrow>
Data version : Globocan 2022 (version 1.1)
© All Rights Reserved 2025

International Agency for Research on Cancer
World Health Organization

Cytotoxins

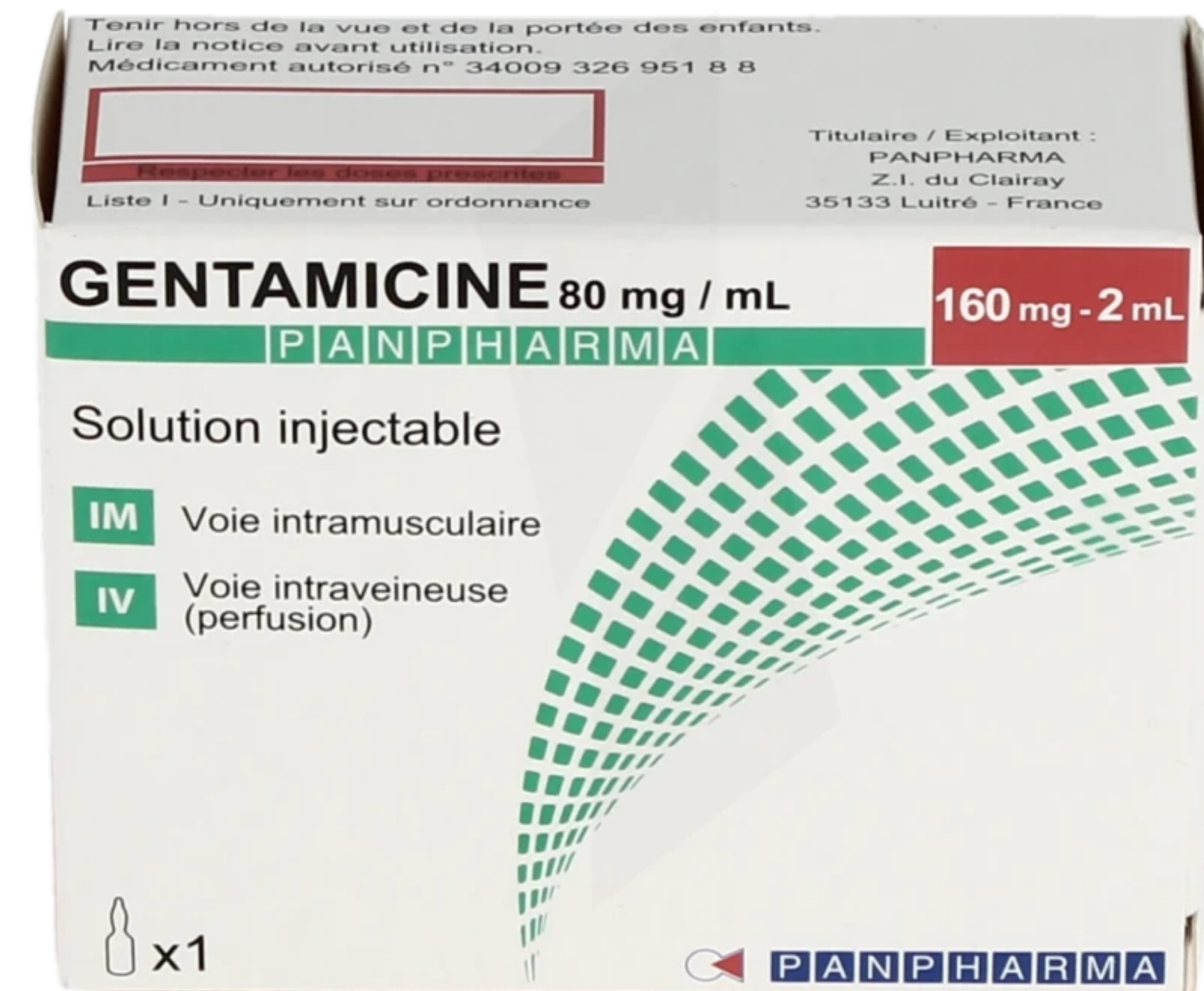
What's the first antibiotic you can think of?

Do you know where it's coming from?

Two common examples



Amoxicillin



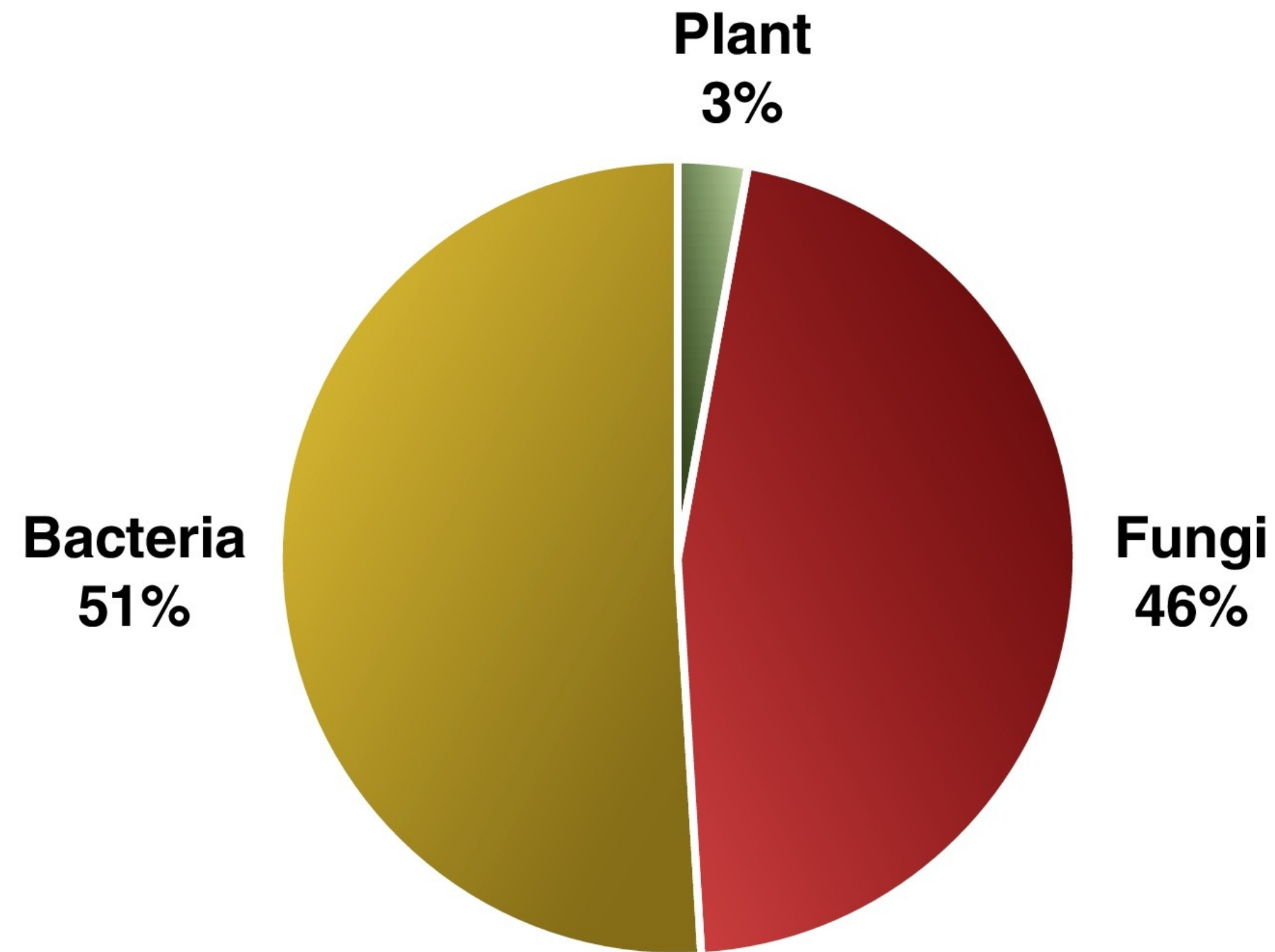
Gentamicin

Discovered/derived from microbes

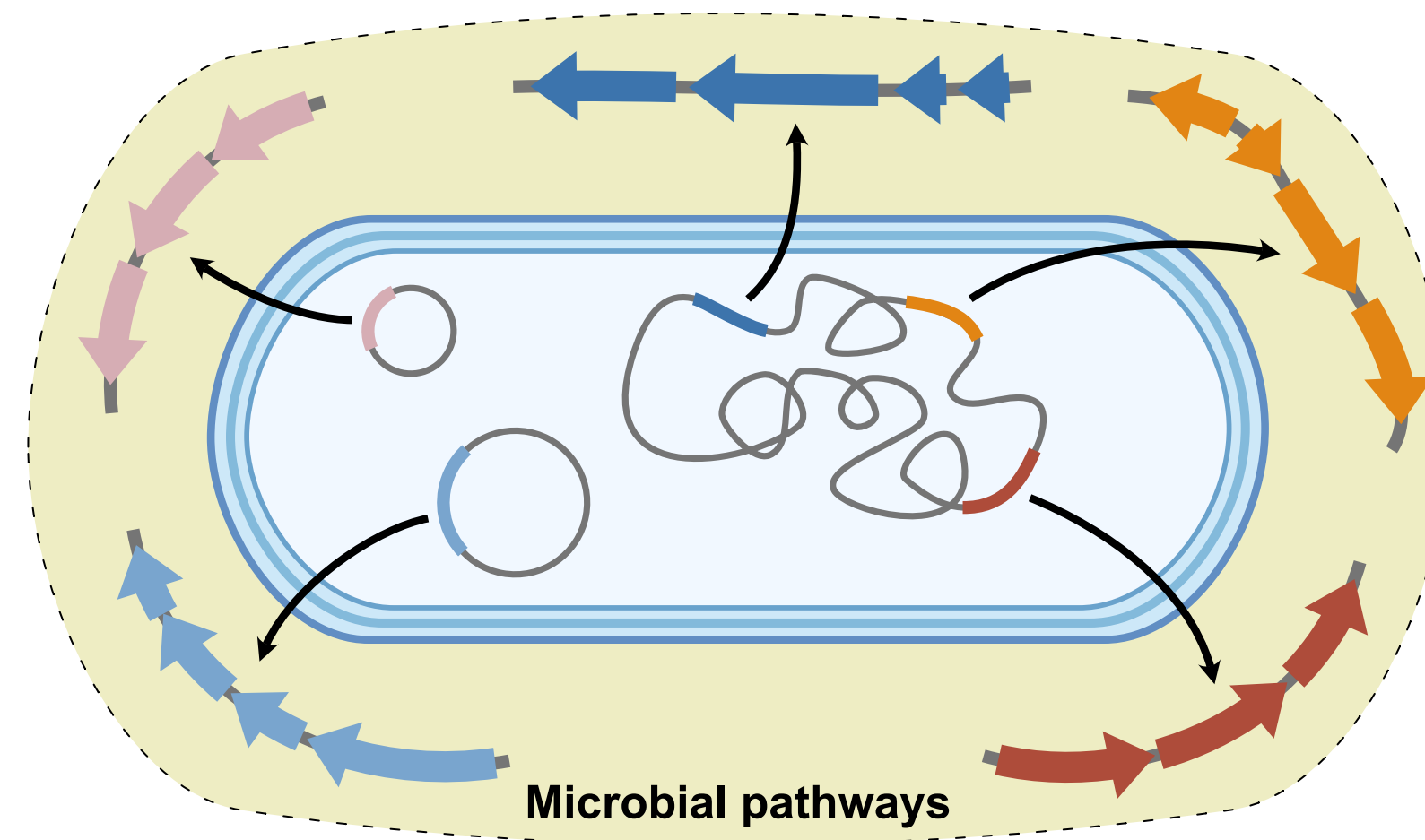


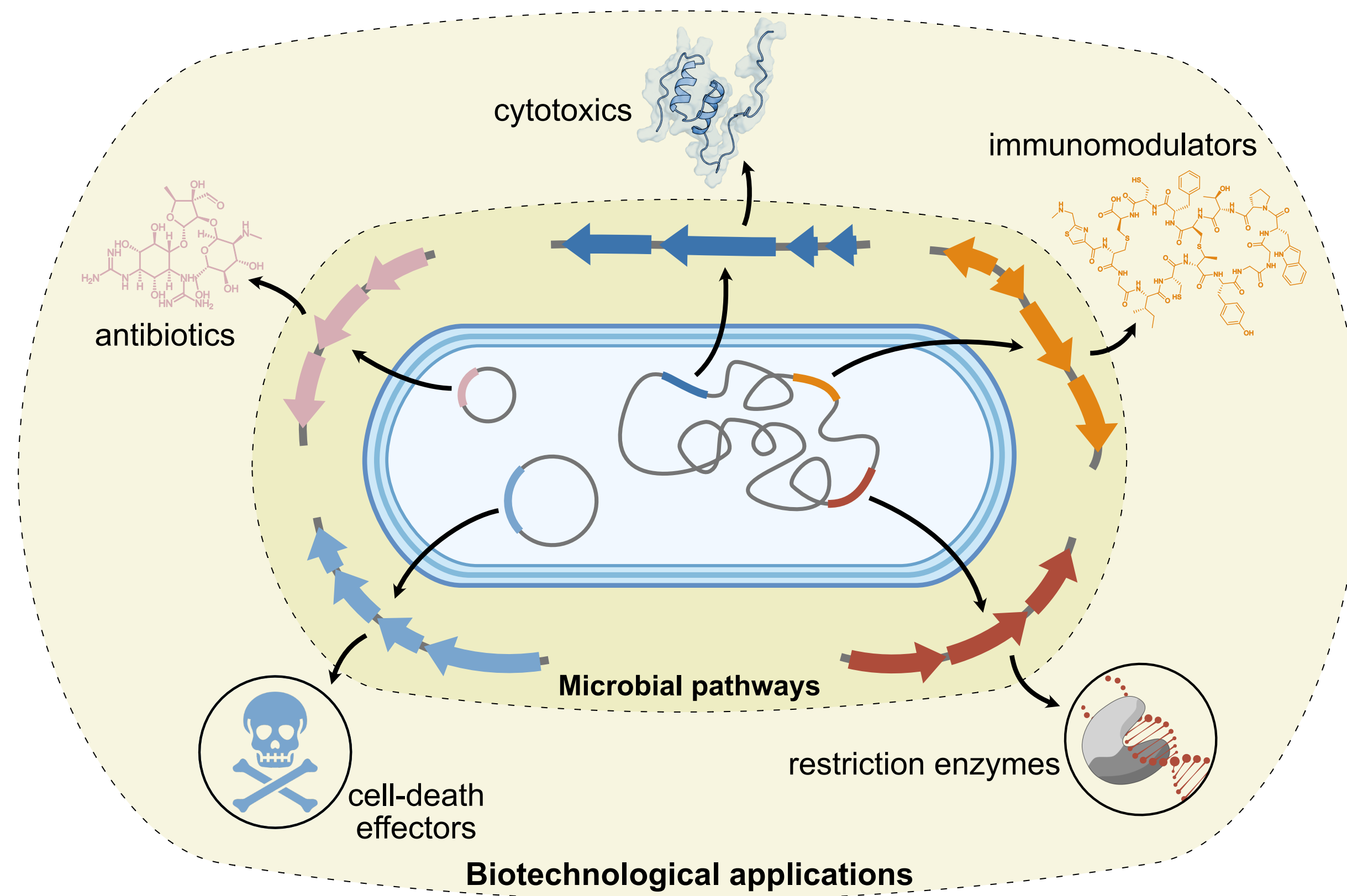
Microbes as a source of natural product drugs

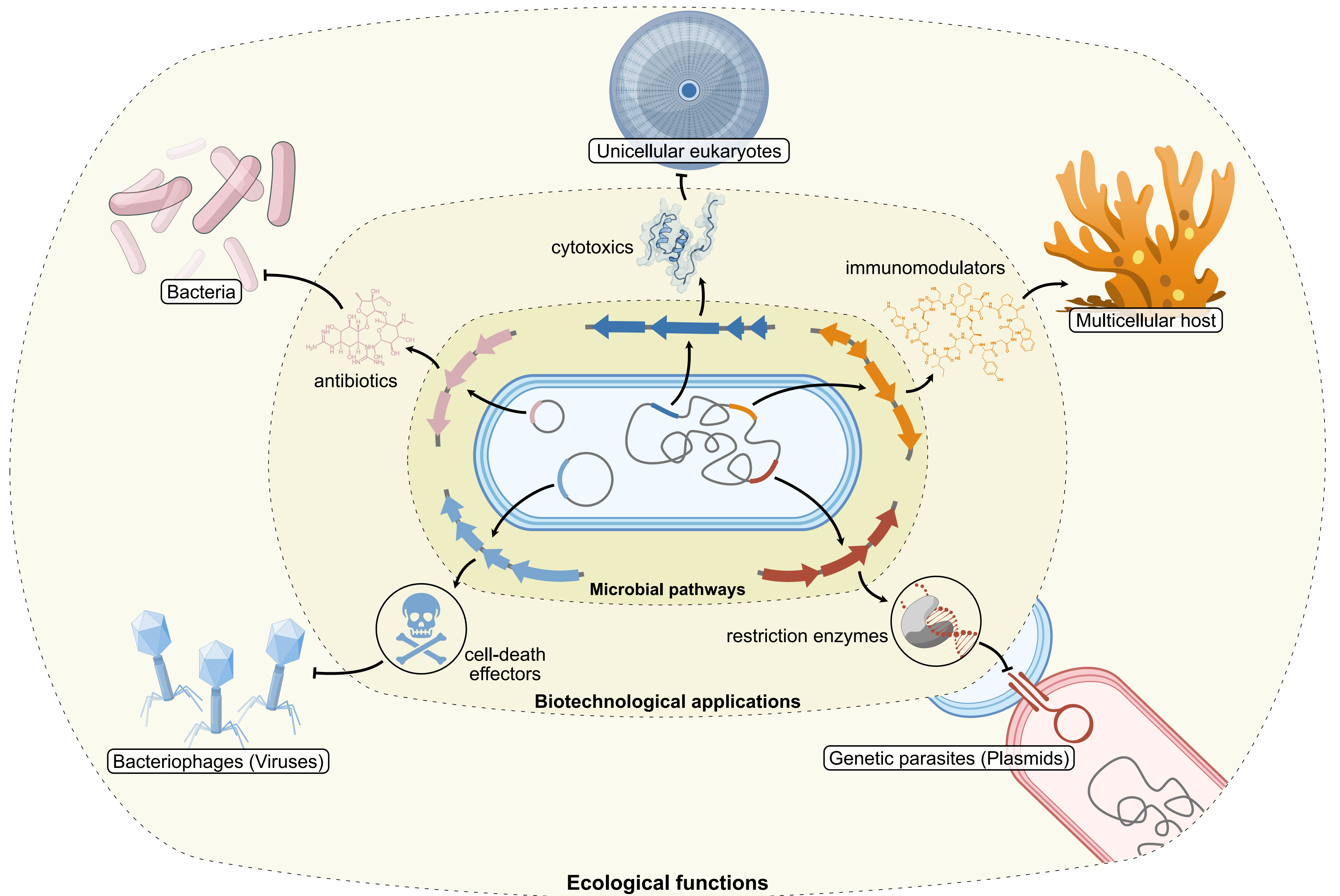
Source of FDA-approved antibacterial natural products

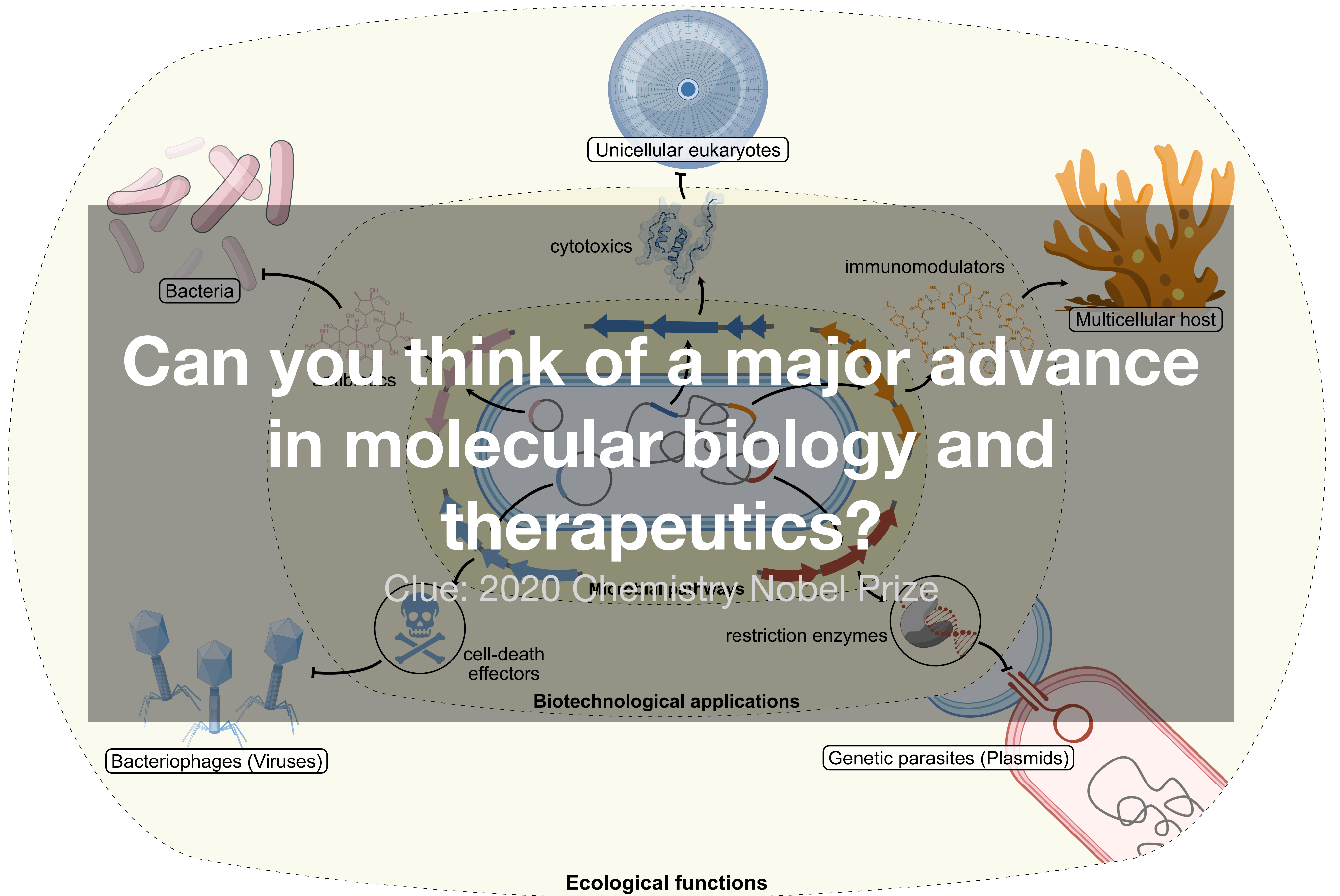


Microbial metabolism as a source of biotechnological and therapeutic innovation





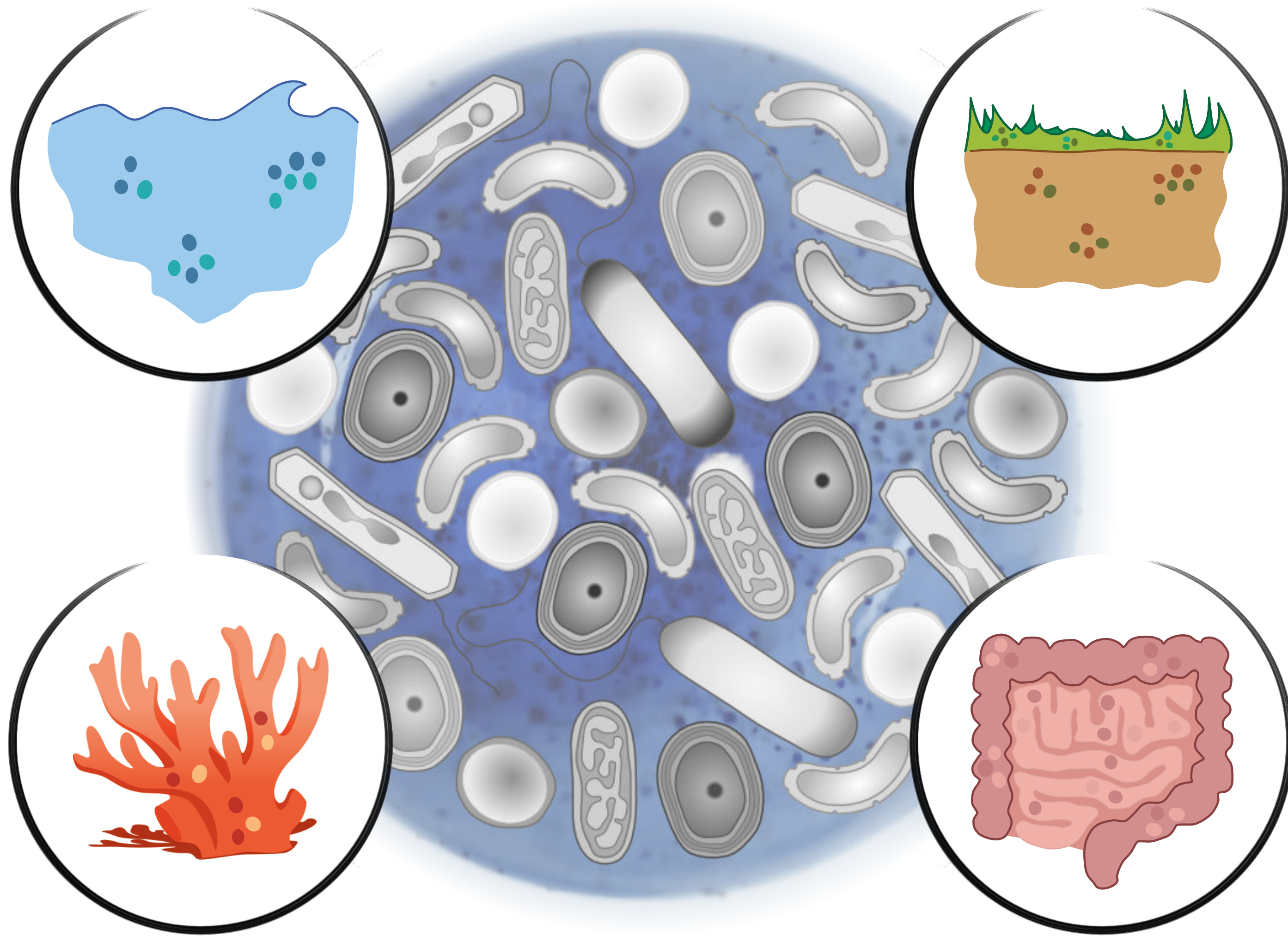




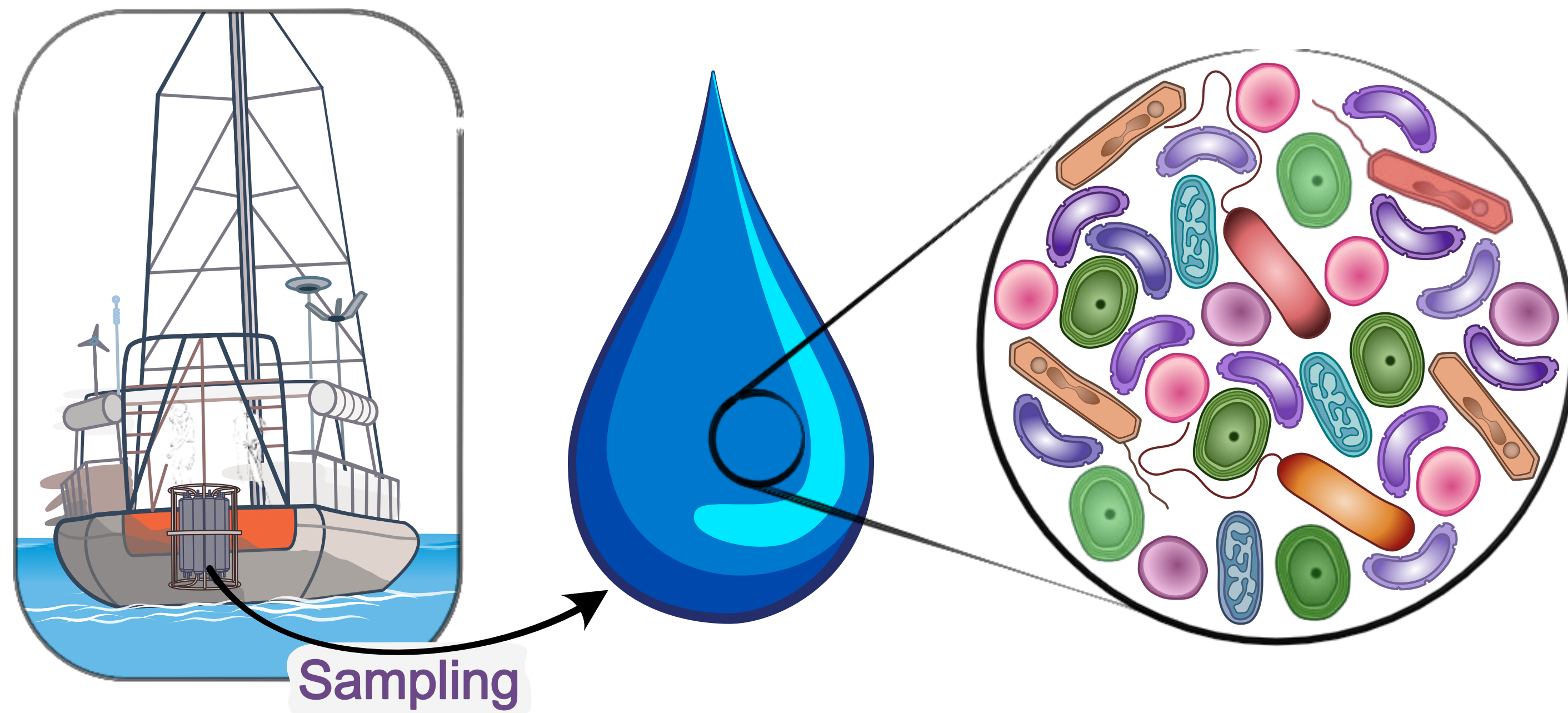
Microbes are ubiquitous on our planet



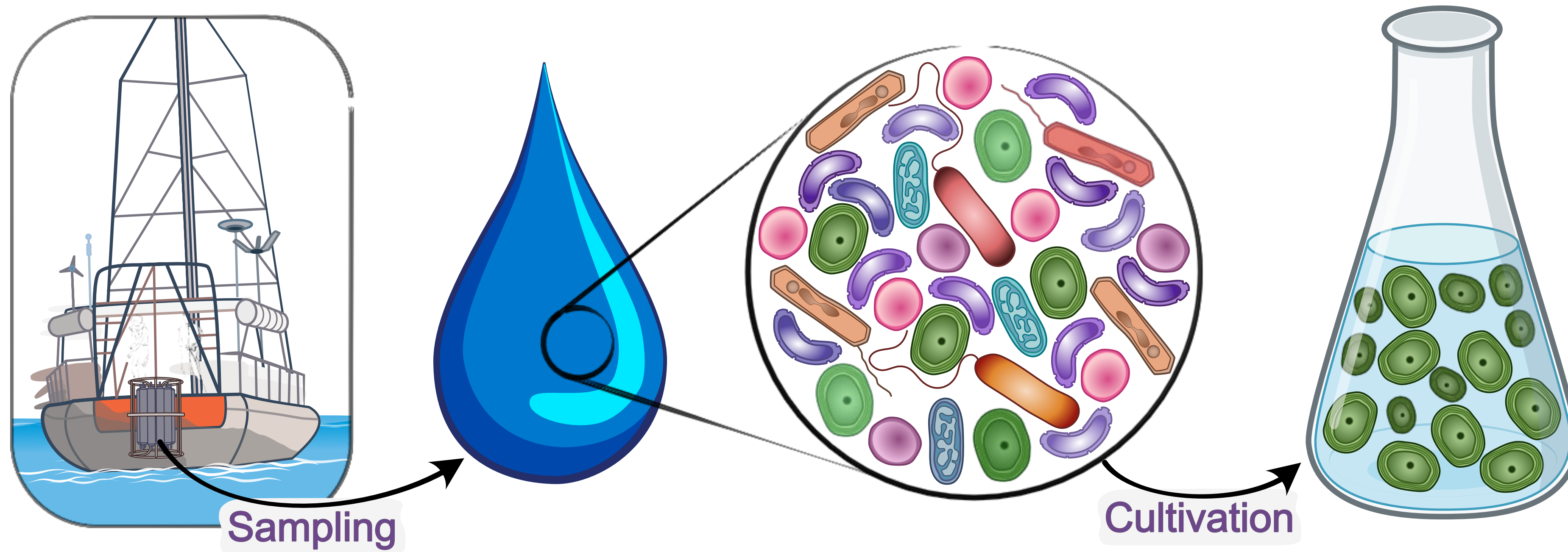
Microbiomes as a source for new applications



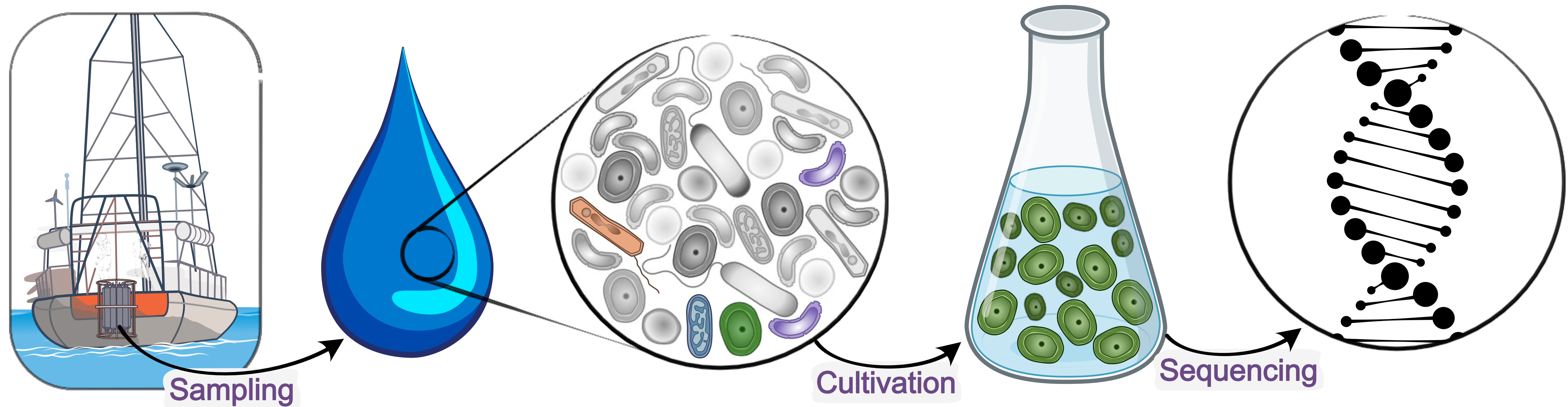
Cultivation- and genomics-based discovery



Cultivation- and genomics-based discovery



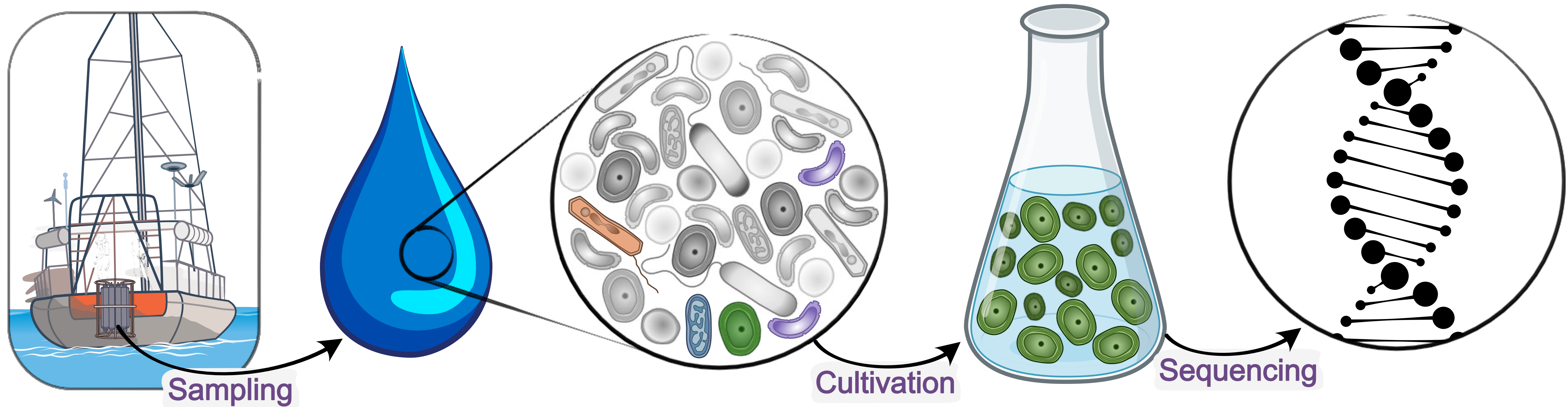
Cultivation- and genomics-based discovery



Access the genome:
- Functional blueprint

Bioinformatic mining

Uncultivated majority

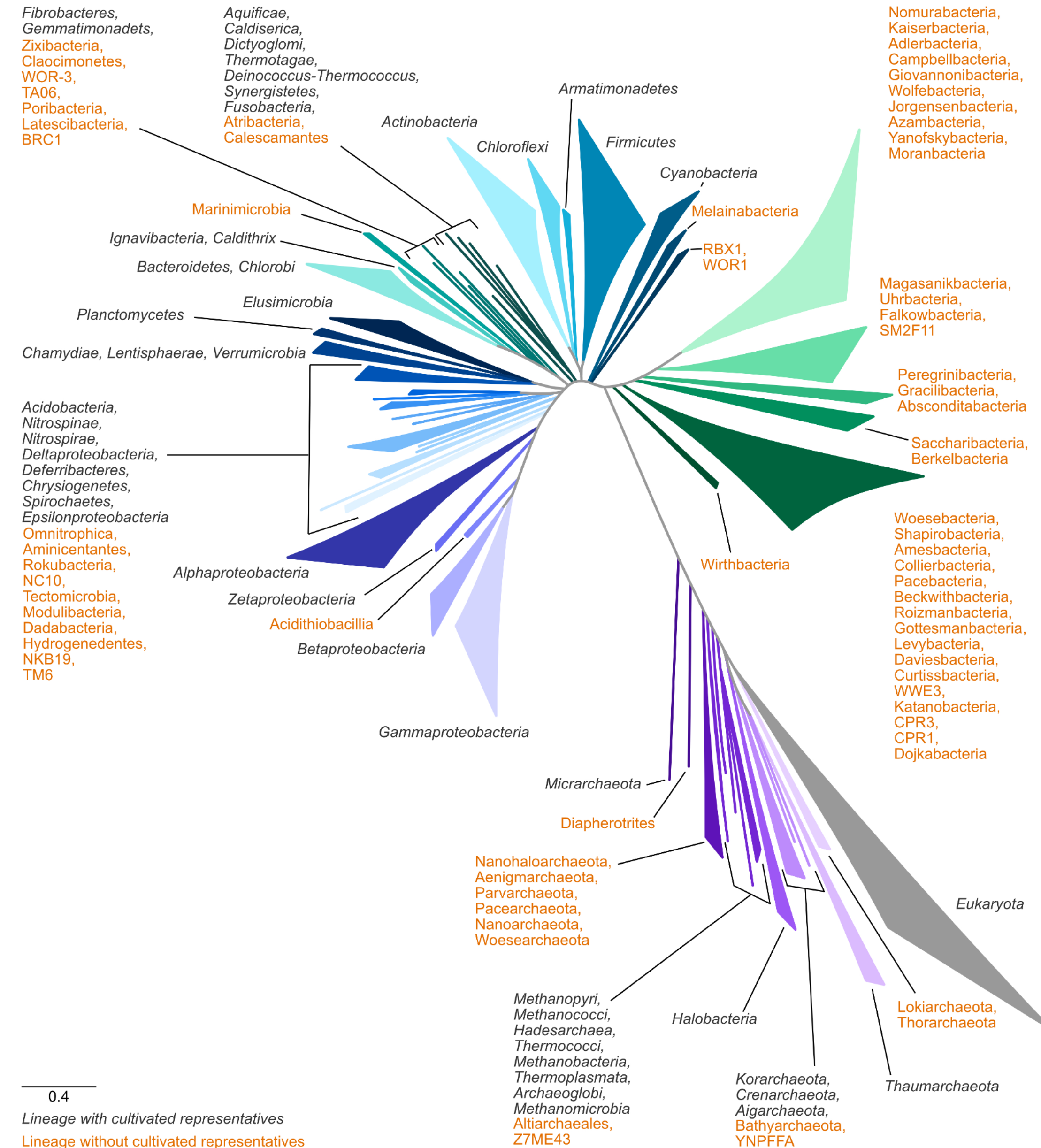


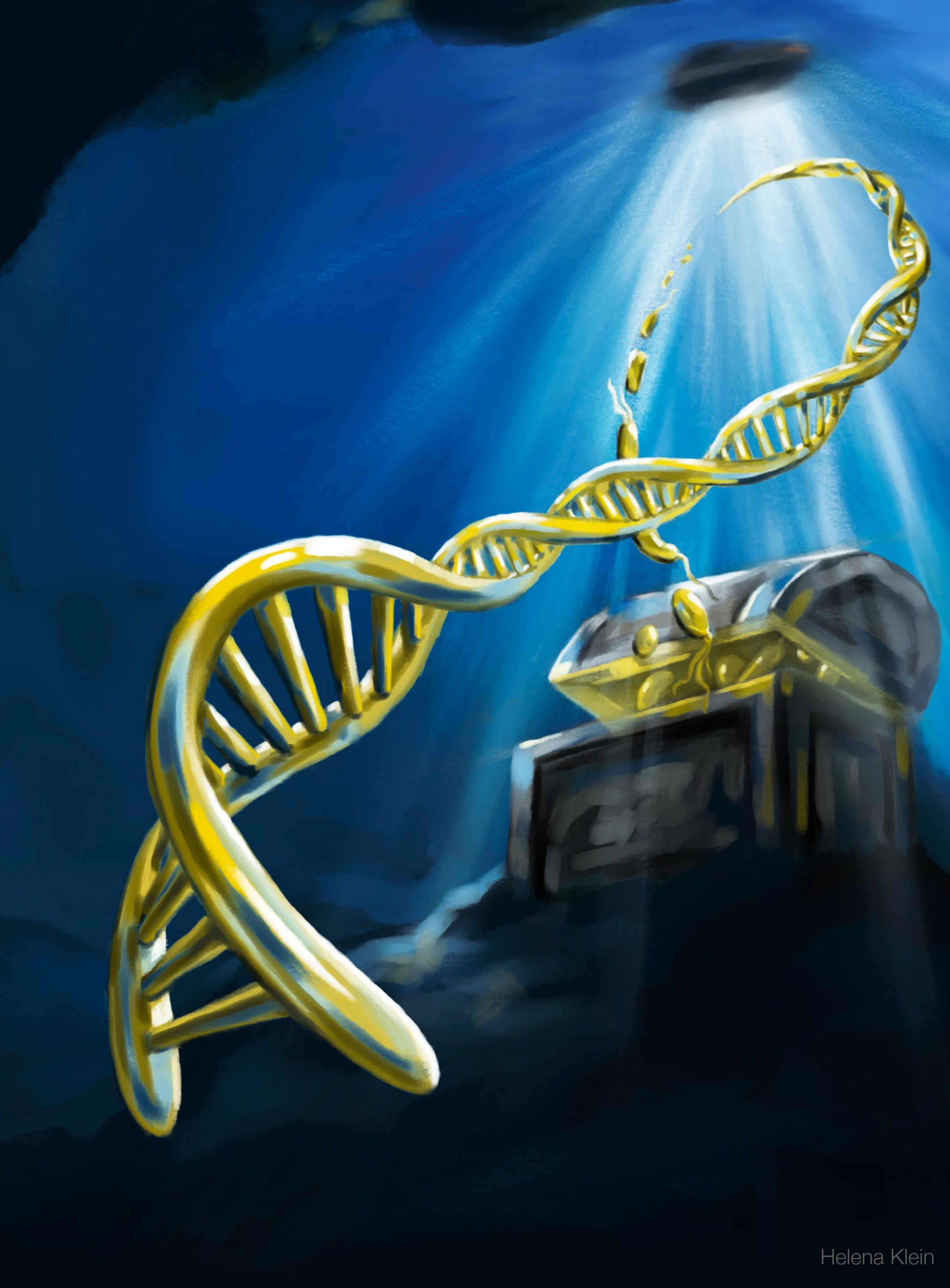
Access the genome:

- Functional blueprint

Bioinformatic mining

Uncultivated majority

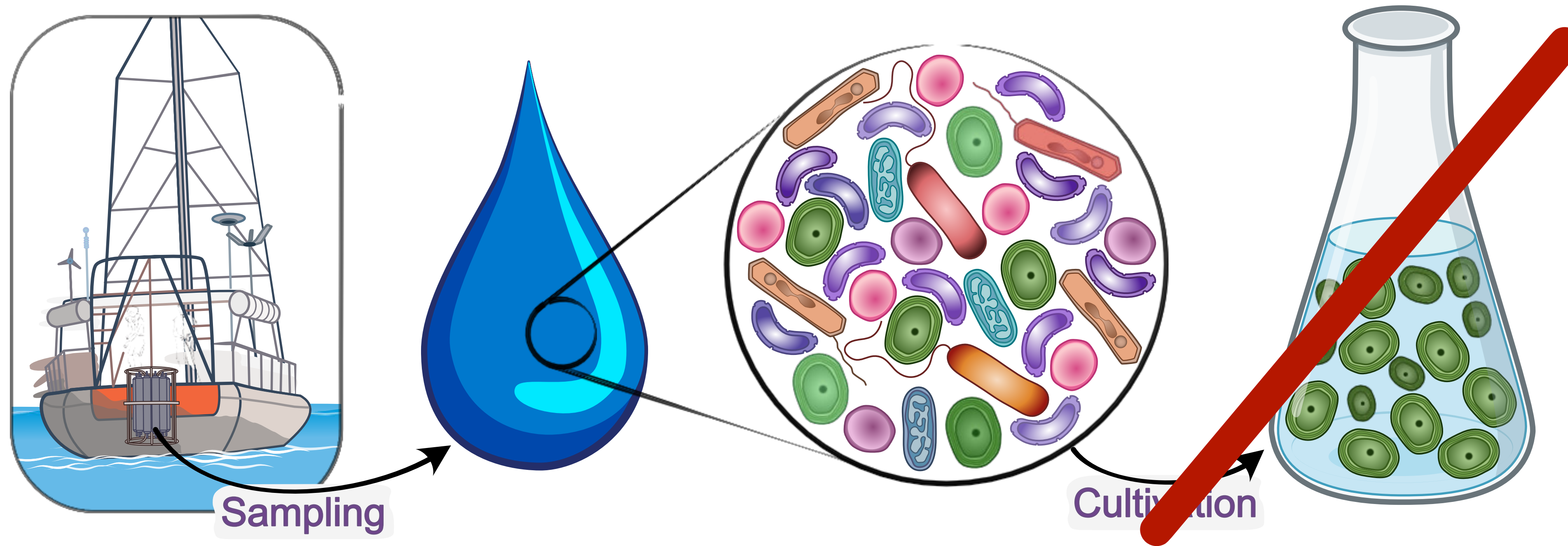




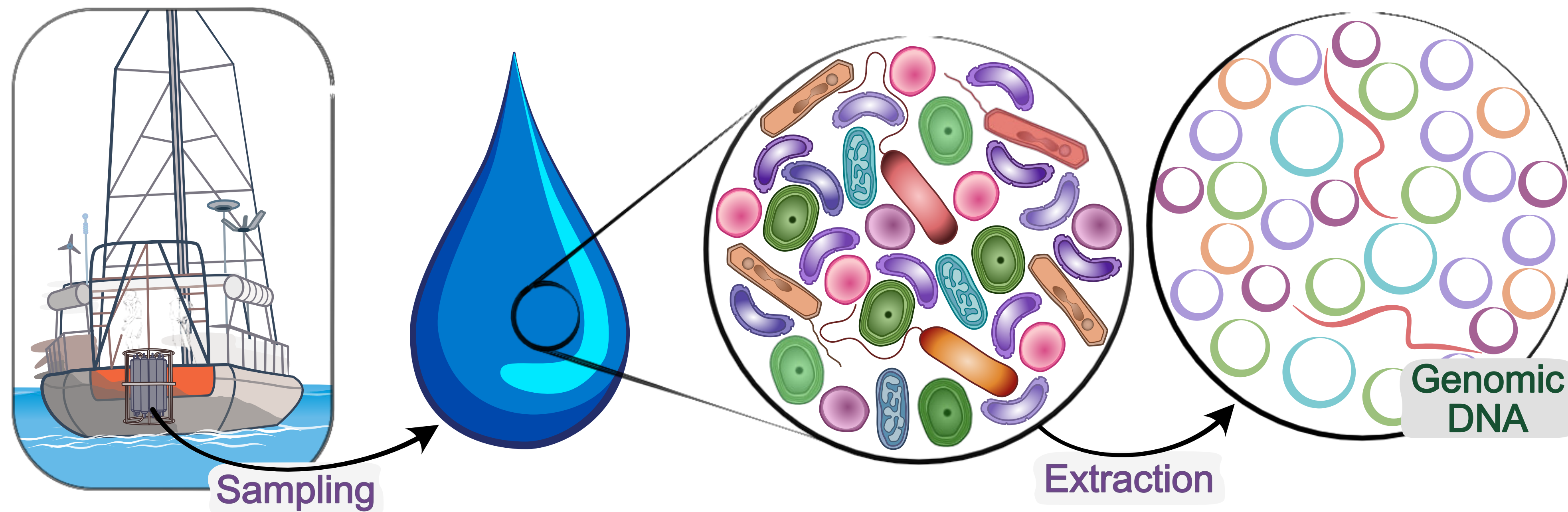
**Microbiomes as untapped
treasure troves for
biotechnological and
therapeutic solutions**

How do we mine microbiomes?

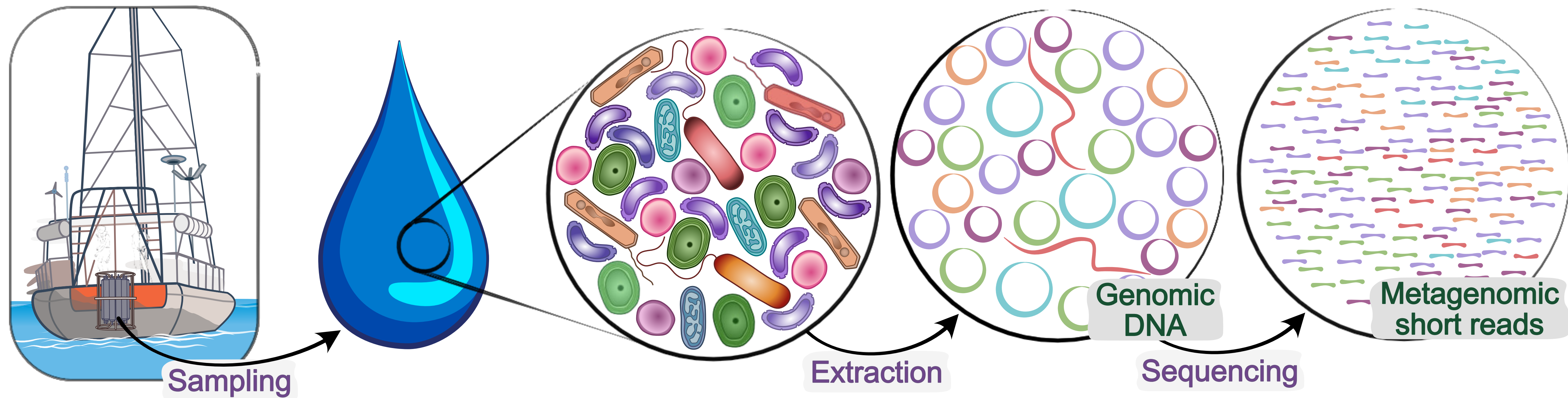
Cultivation-independent approach



Metagenomics: community-level genomics



Metagenomics: community-level genomics

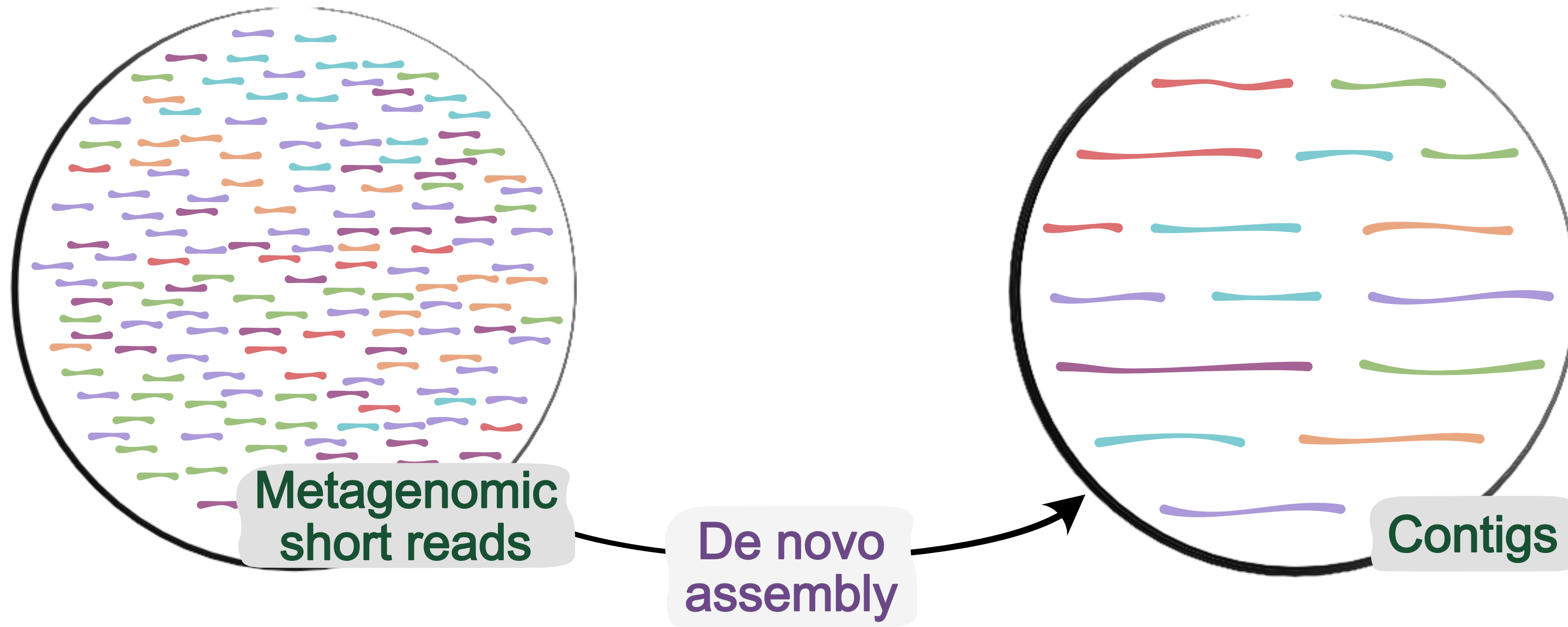


Small DNA fragments
(a tenth of a gene in
length)

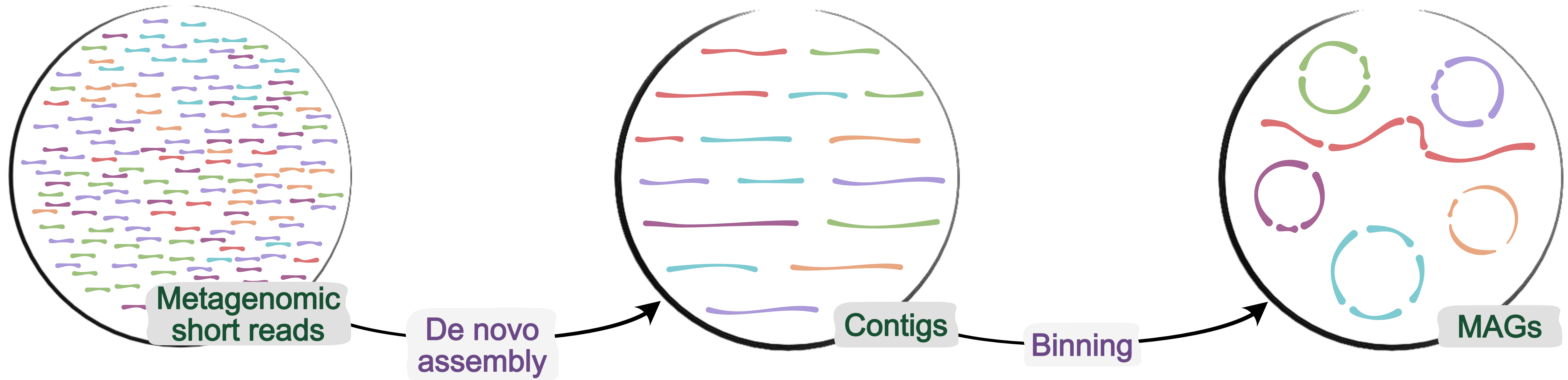
Cultivation-independent genome reconstruction



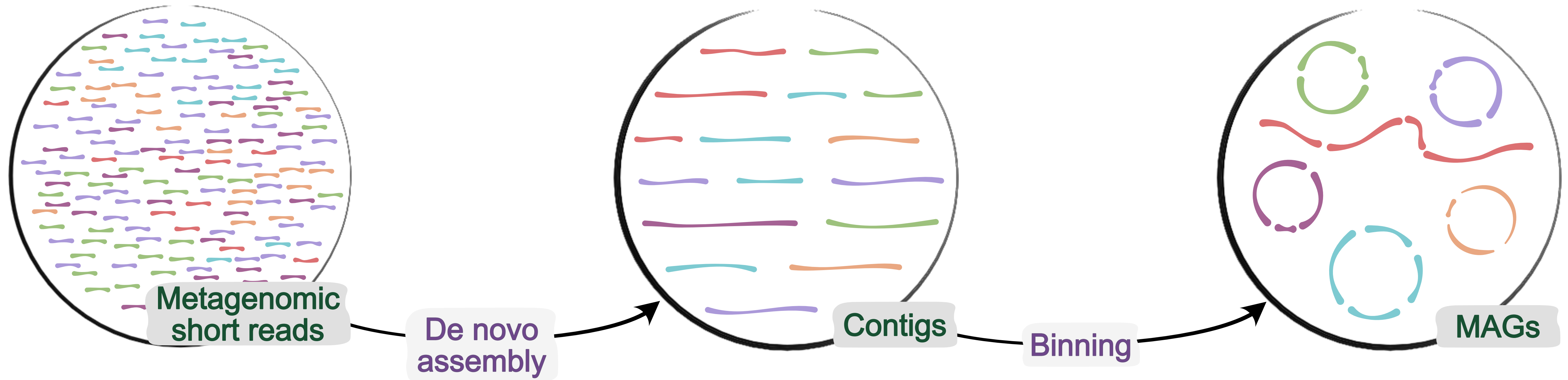
Cultivation-independent genome reconstruction



Cultivation-independent genome reconstruction

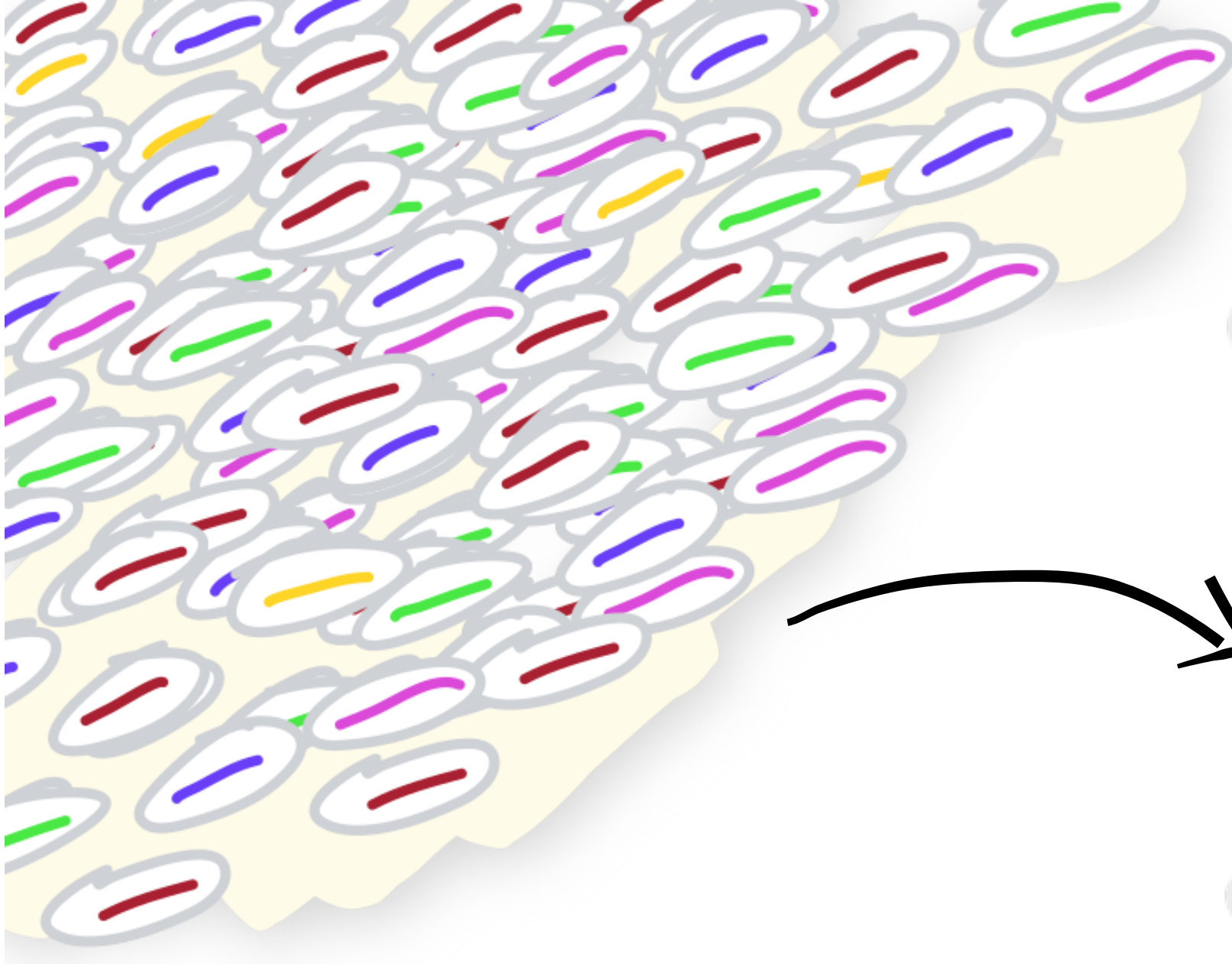


Cultivation-independent genome reconstruction

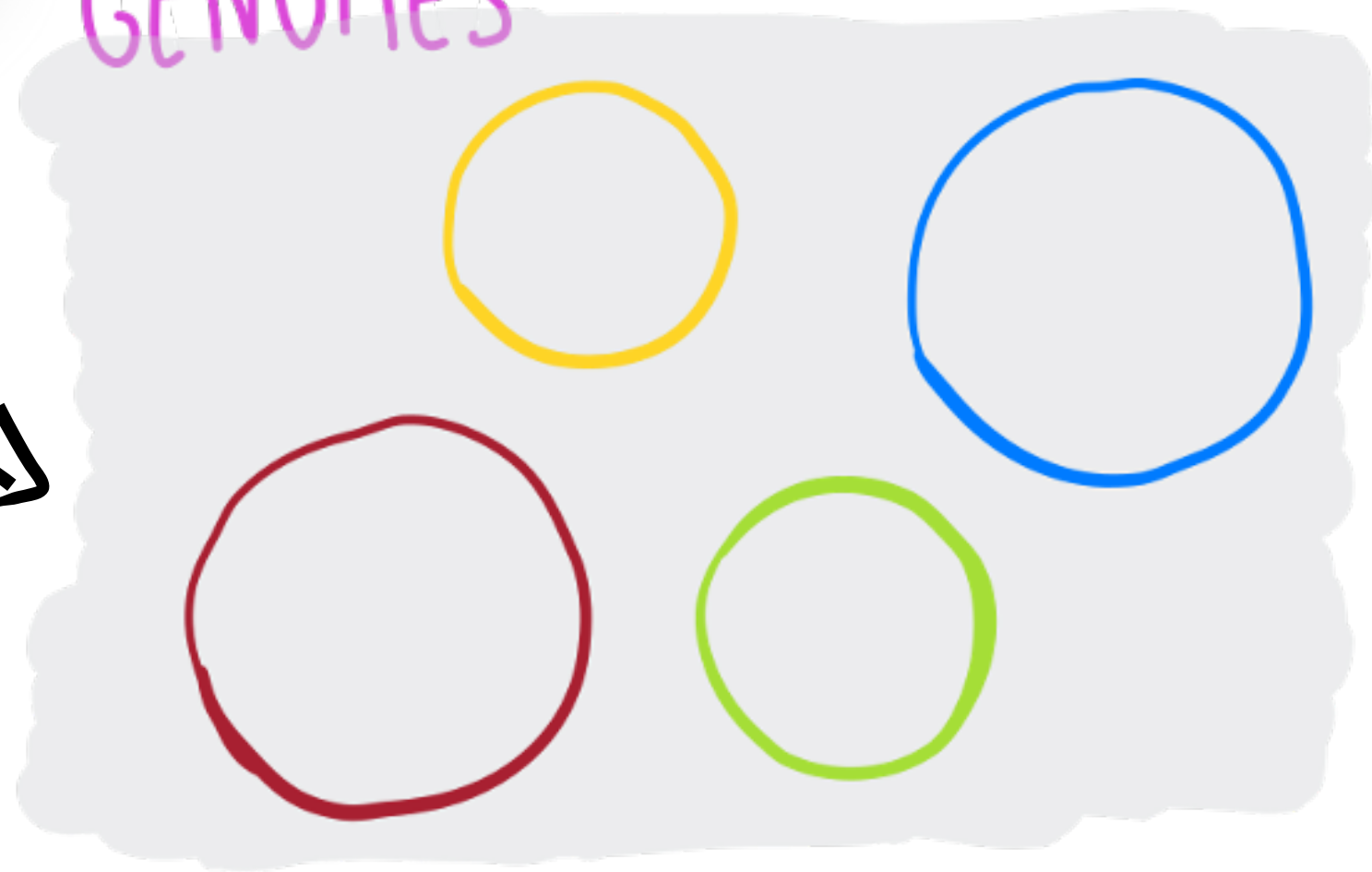


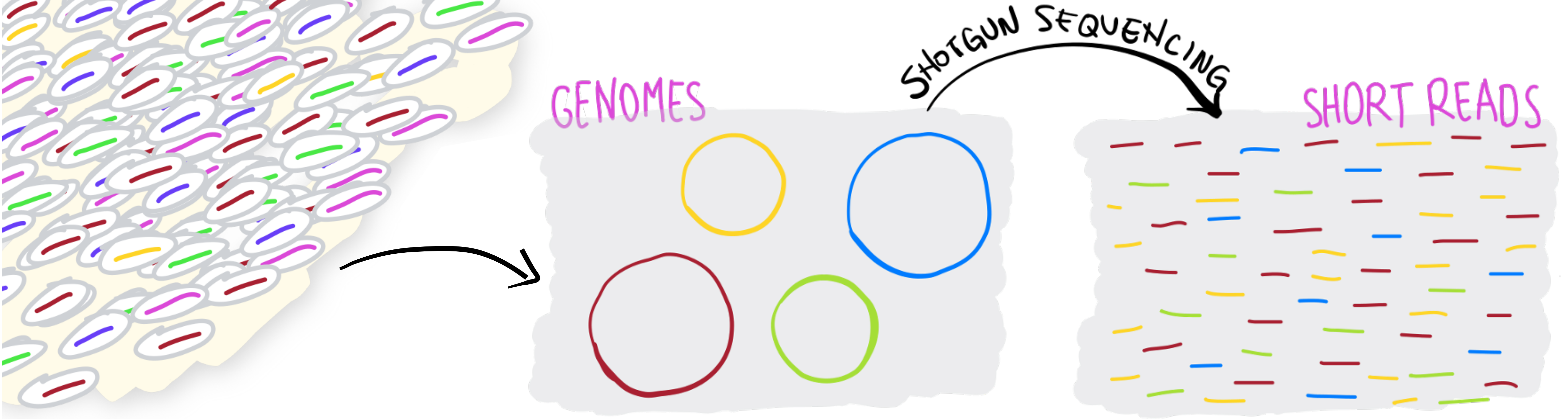
How does it work?

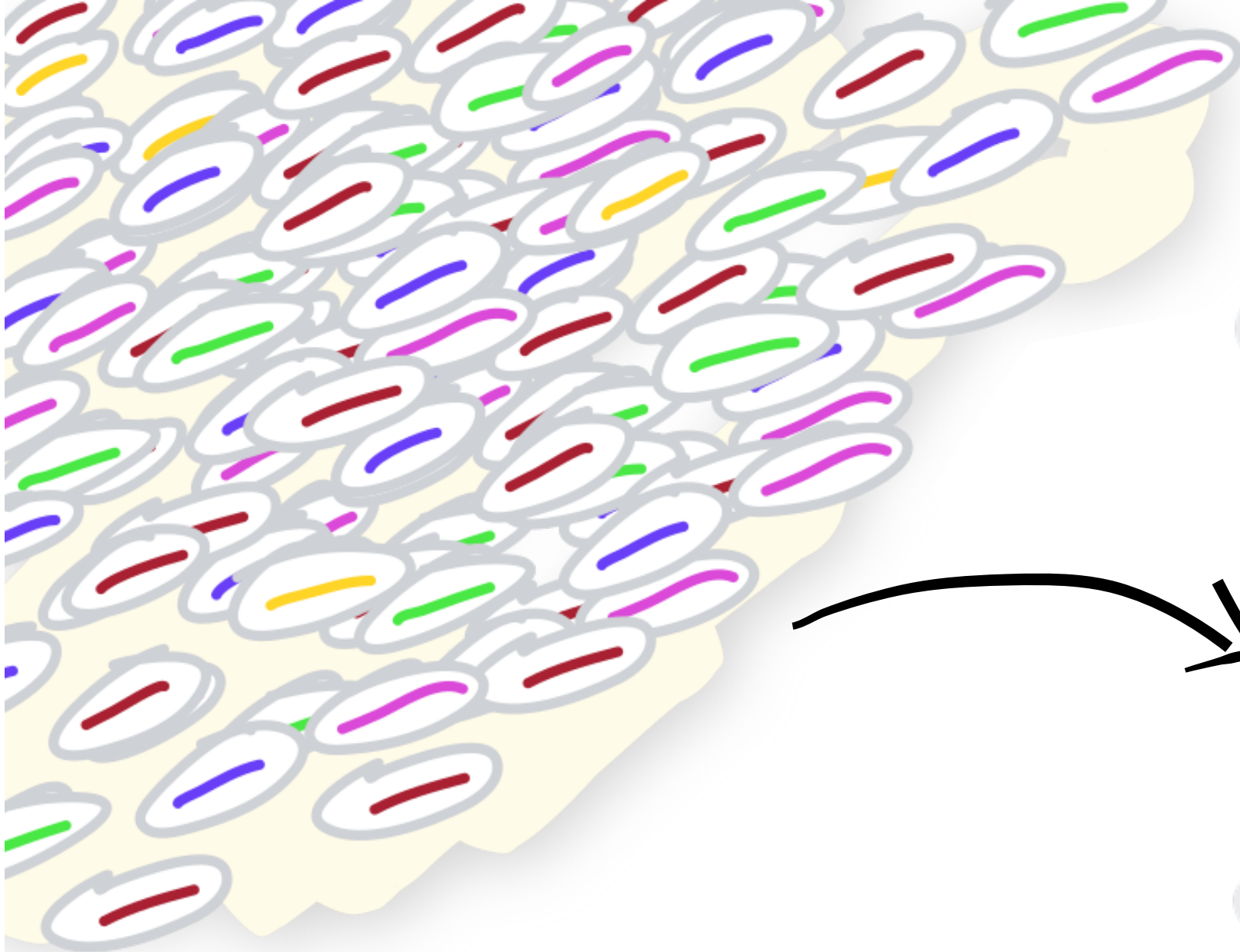
Metagenomic binning



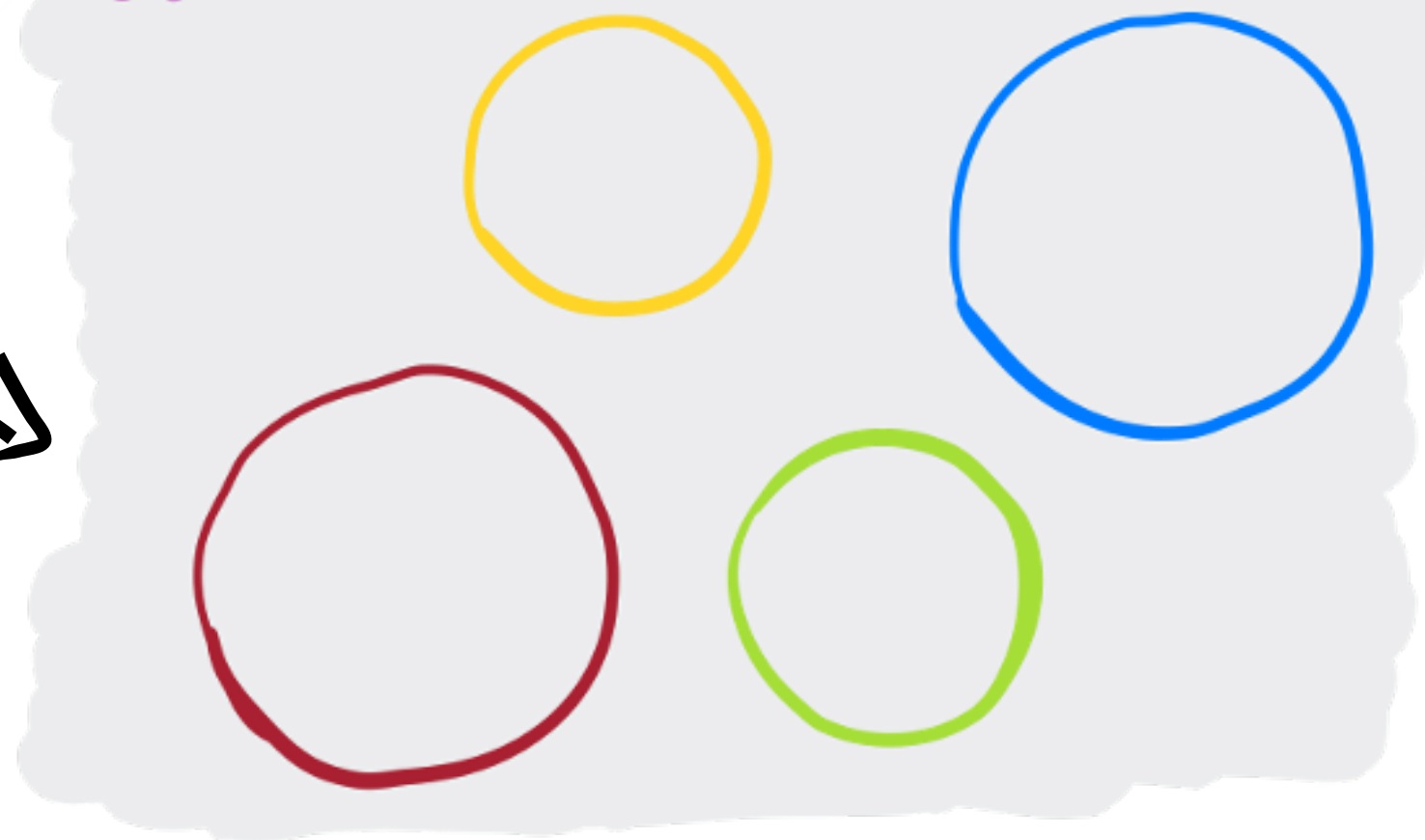
GENOMES







GENOMES

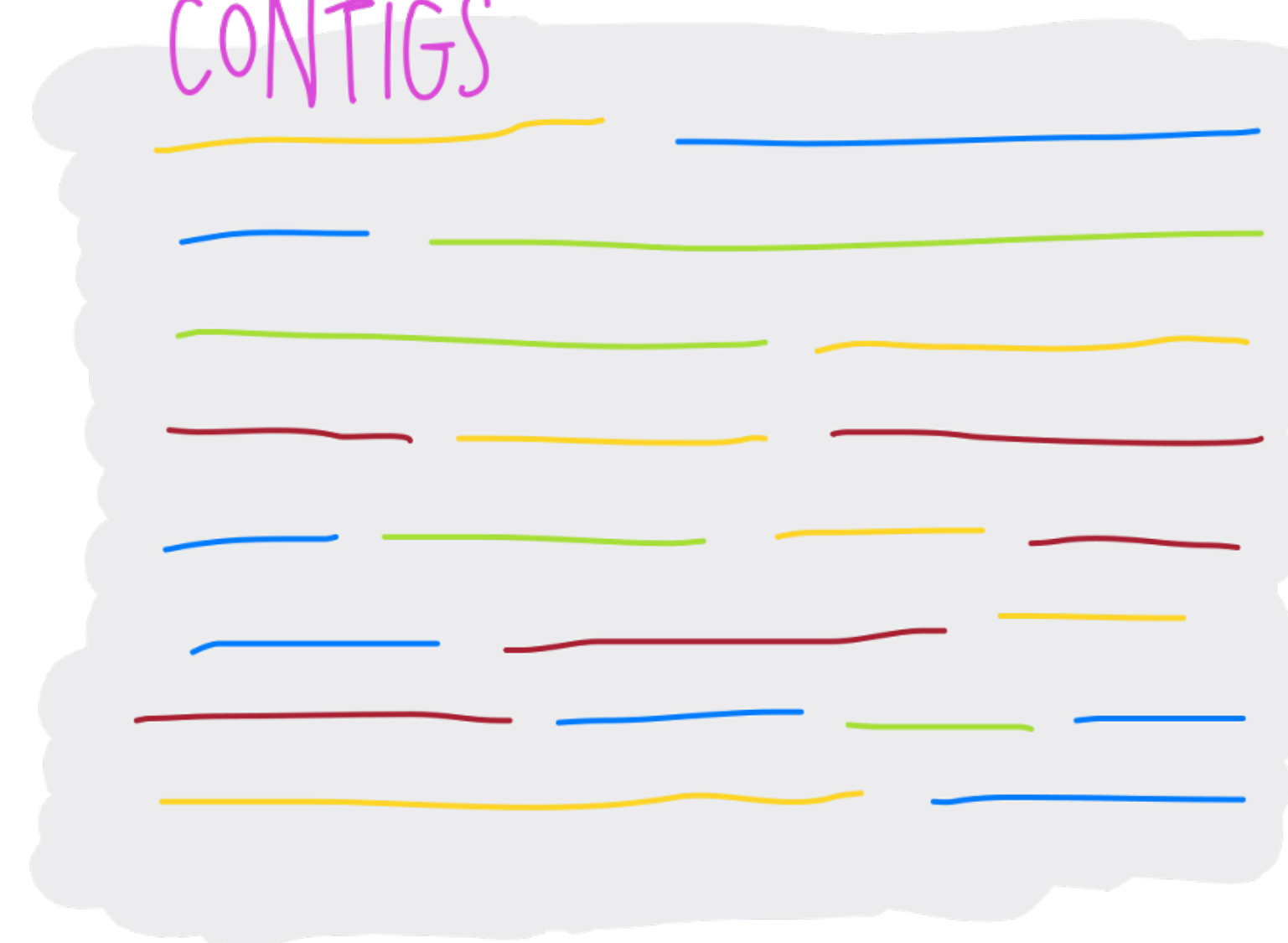


SHOTGUN SEQUENCING

SHORT READS



CONTIGS



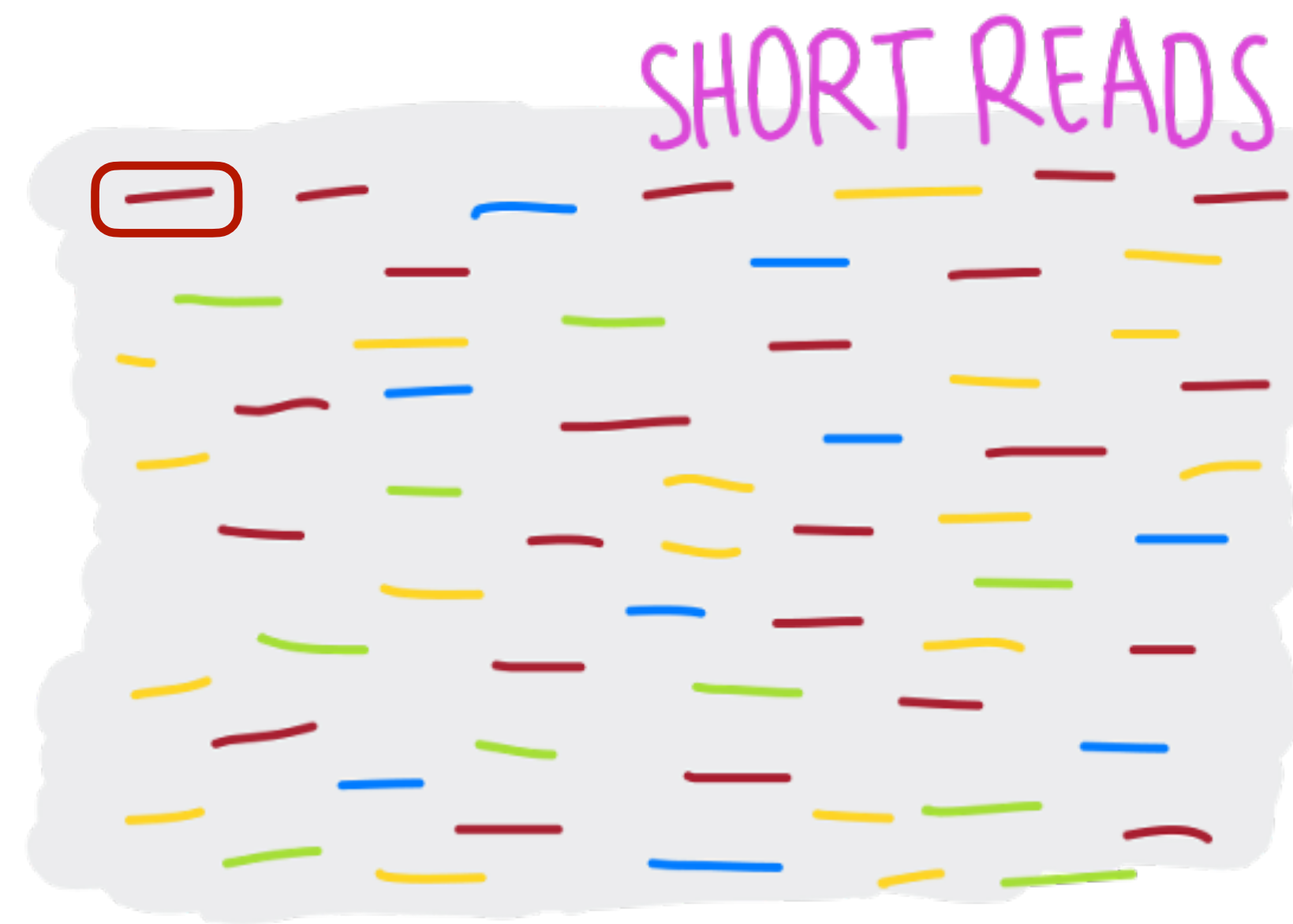
DE-NOVO ASSEMBLY

Metagenomic assembly

Basically a puzzle with bits of DNA

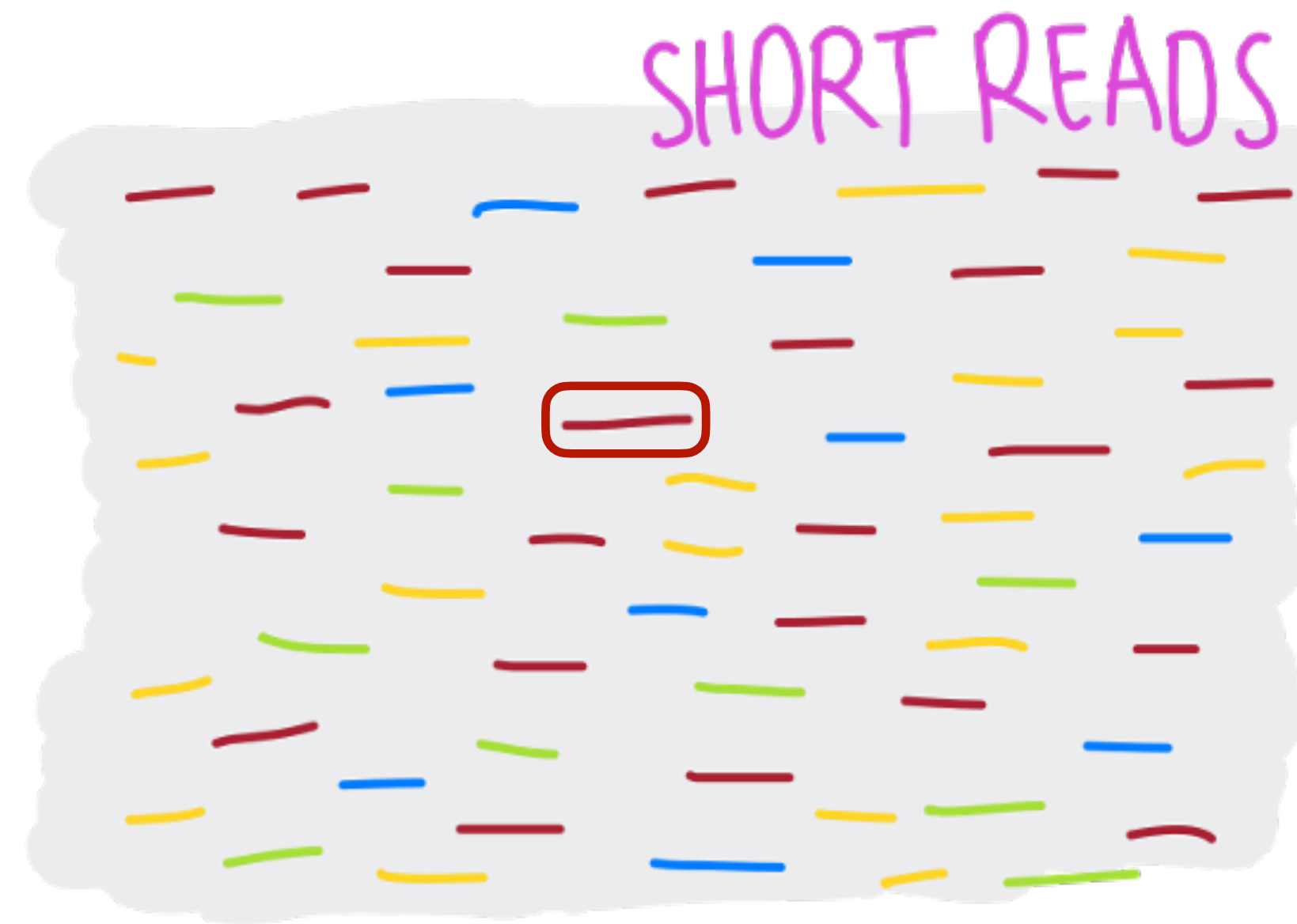
CONTIG #1

CONTIG #1
GTTTTGGCAT



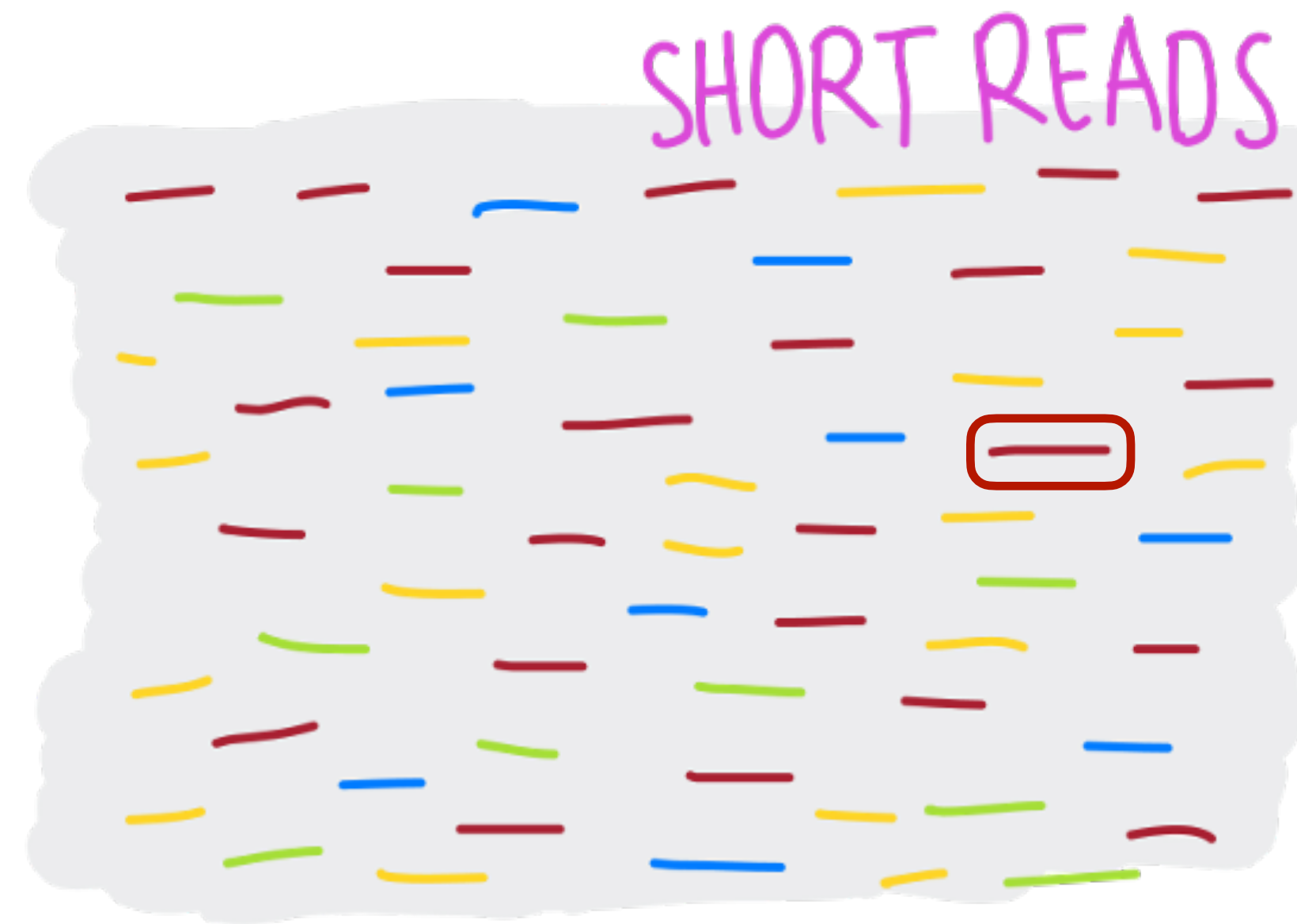
CONTIG #1

GTTTTGGCAT
TTTTGGCATG



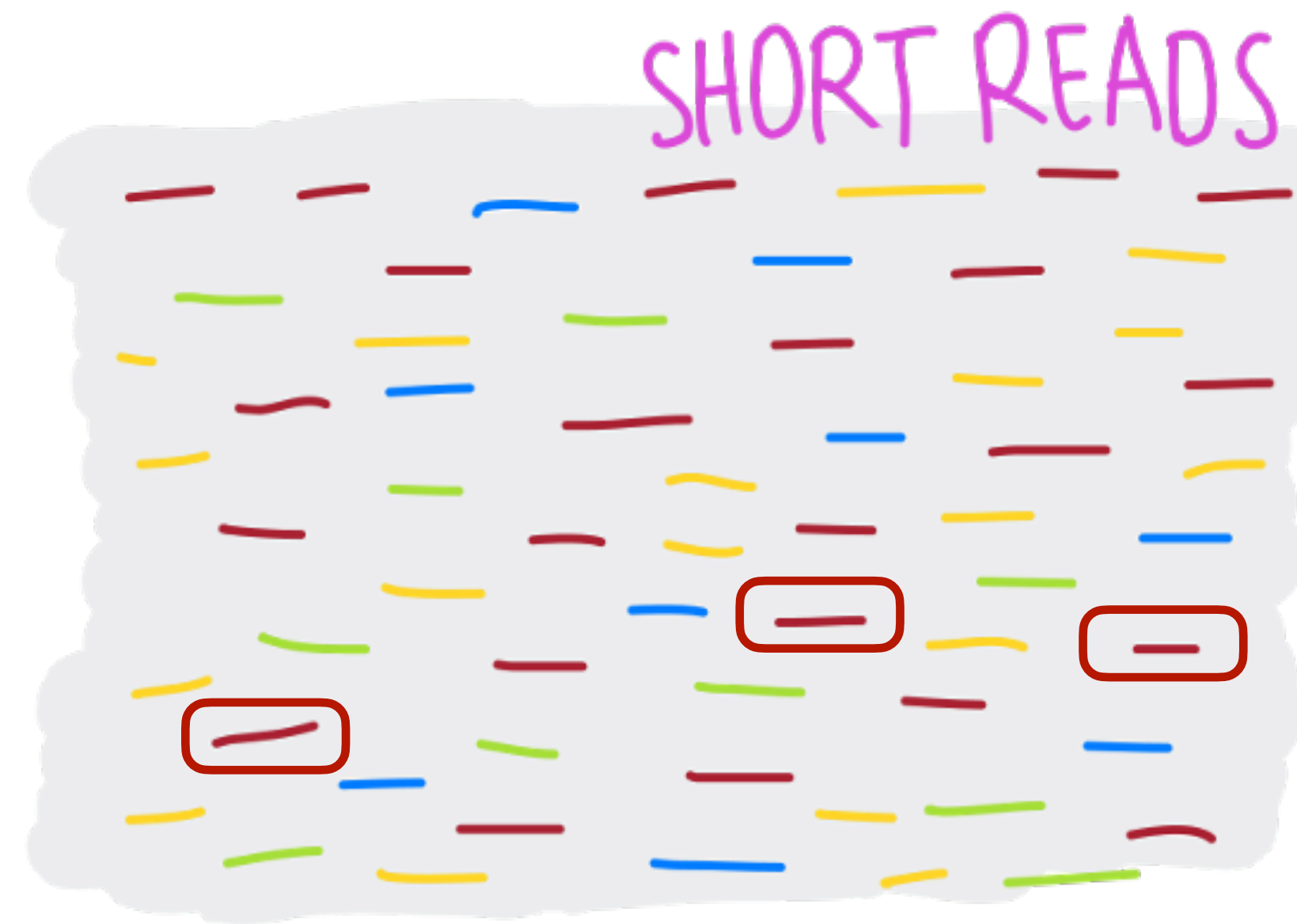
Contig #1

GTTTTGGCAT
TTTTGGCATG
TTTGGCATGA



Contig #1

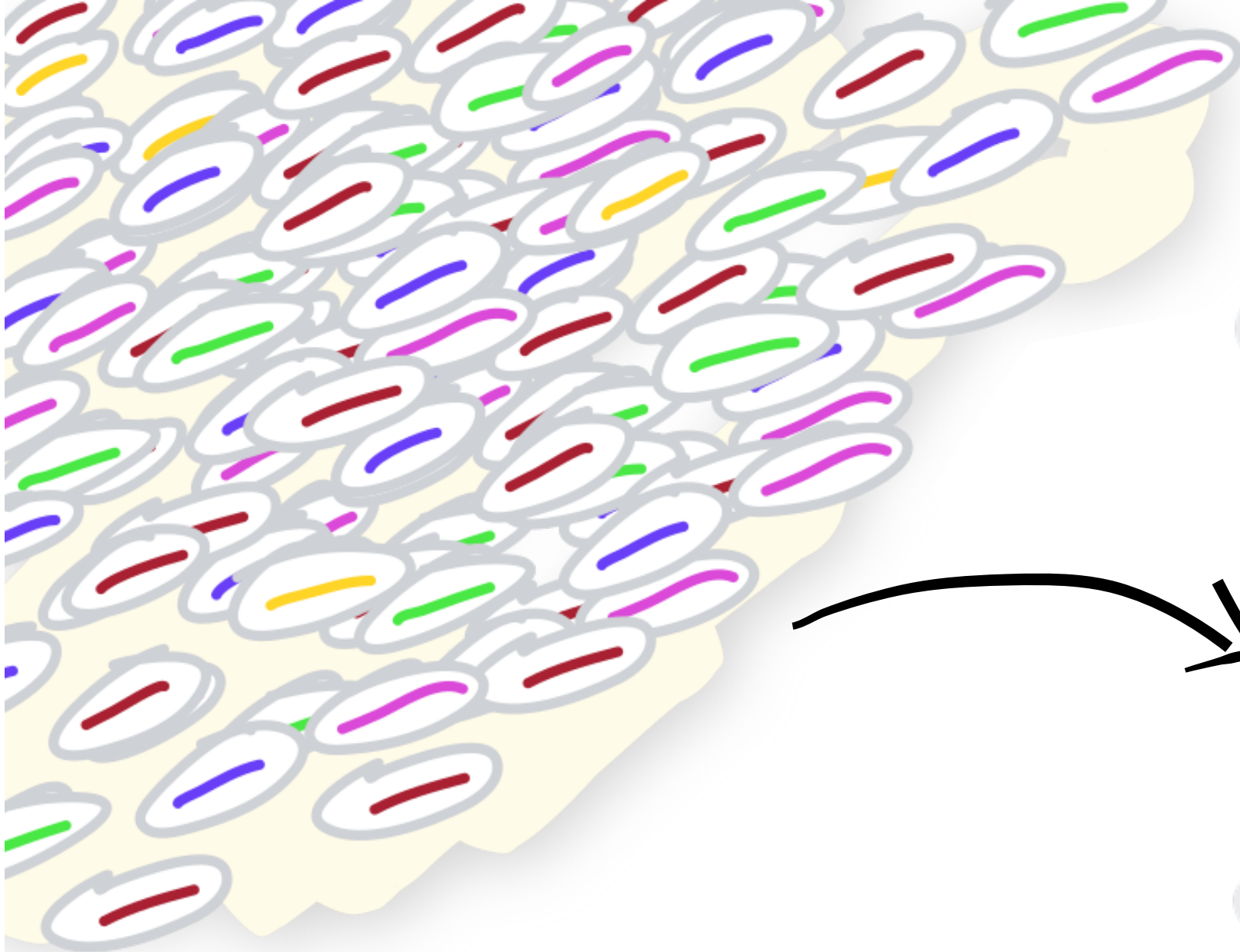
GTTTTGGCAT
TTTTGGCATG
TTTGGCATGA
TTGGCATGAT
TGGCATGATT
GGCATGATTA



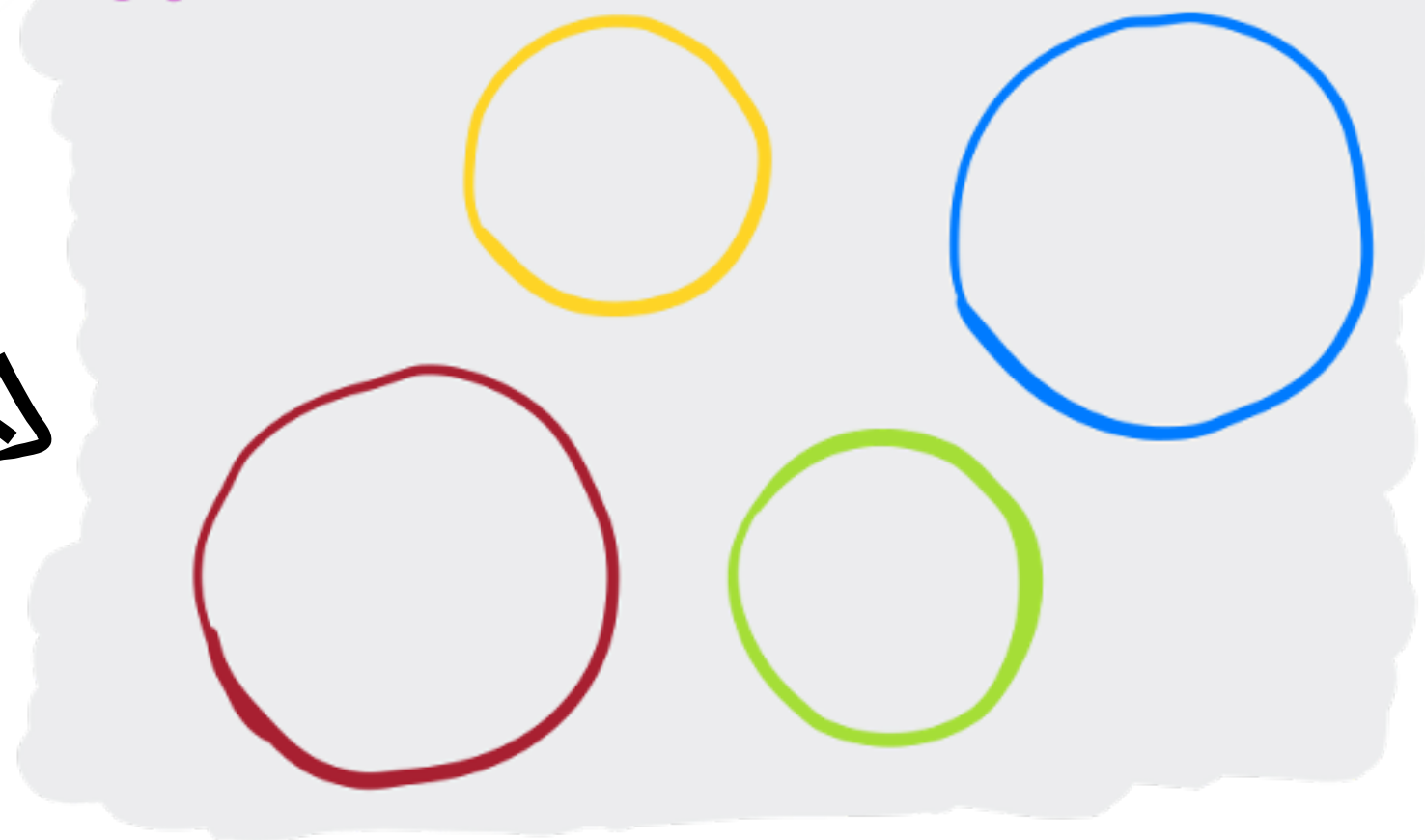
Contig #1

GTTTTGGCAT
TTTTGGCATG
TTTGGCATGA
TTGGCATGAT
TGGCATGATT
GGCATGATTA

GTTTTGGCATGATTA



GENOMES



SHOTGUN SEQUENCING



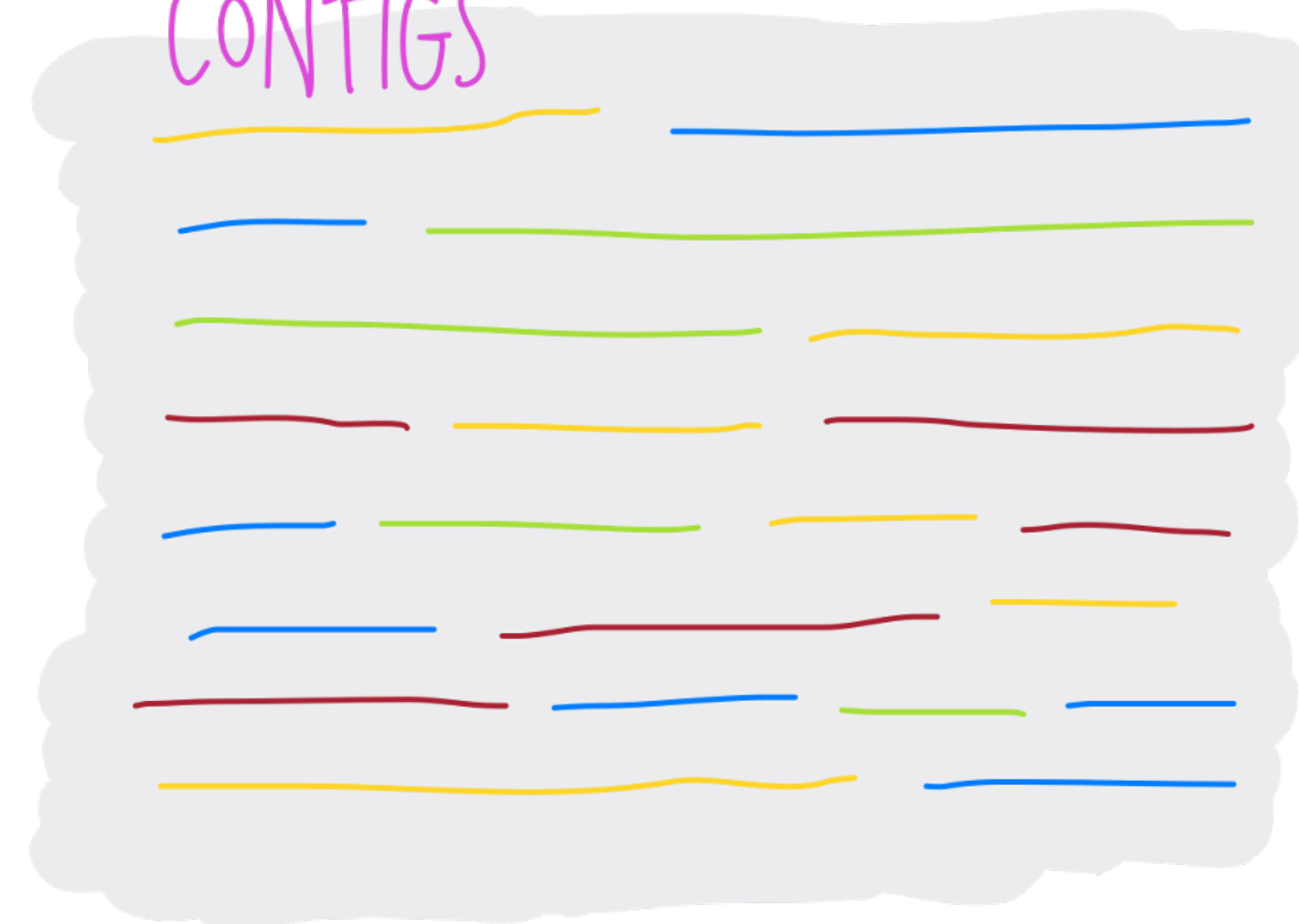
SHORT READS

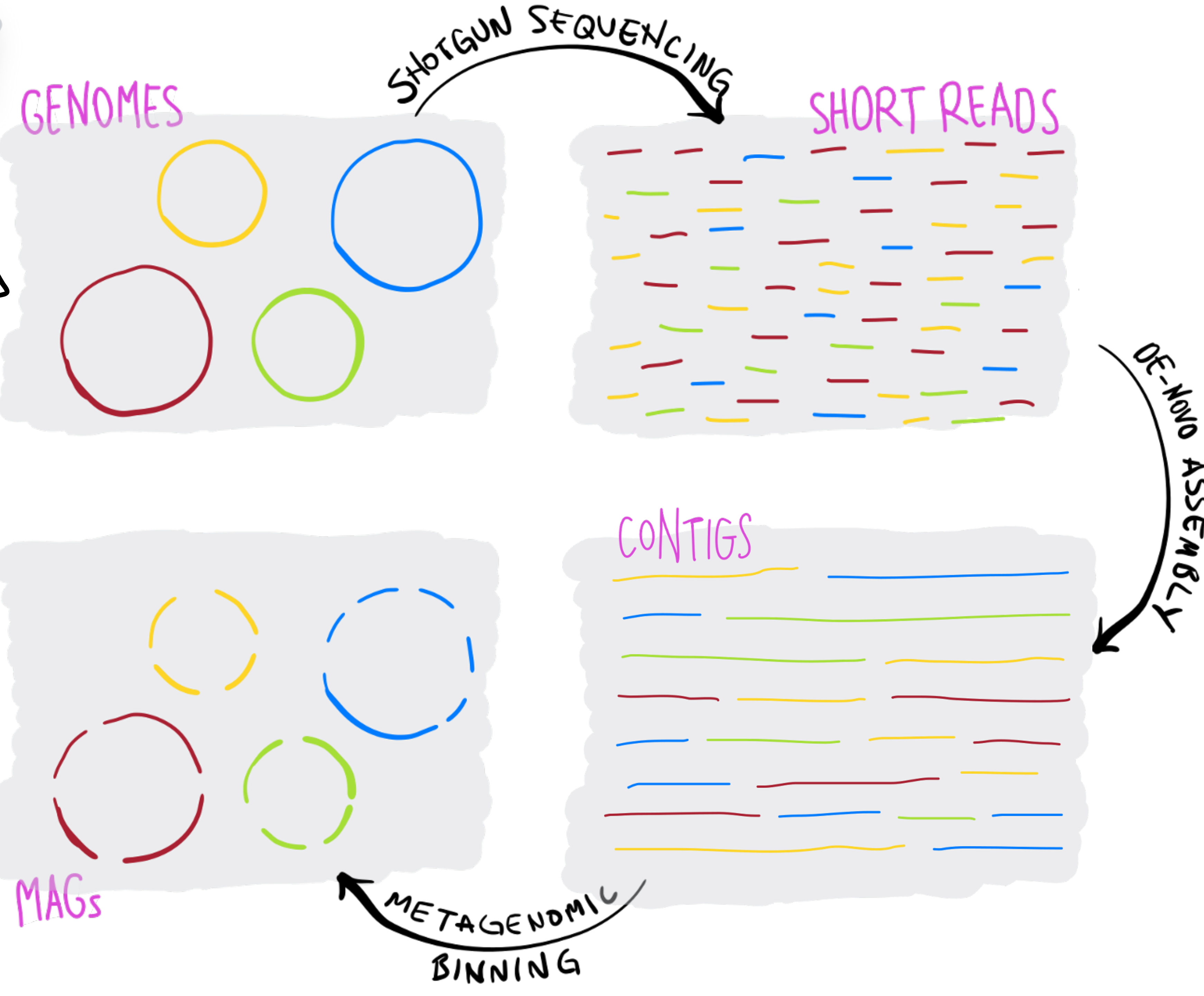
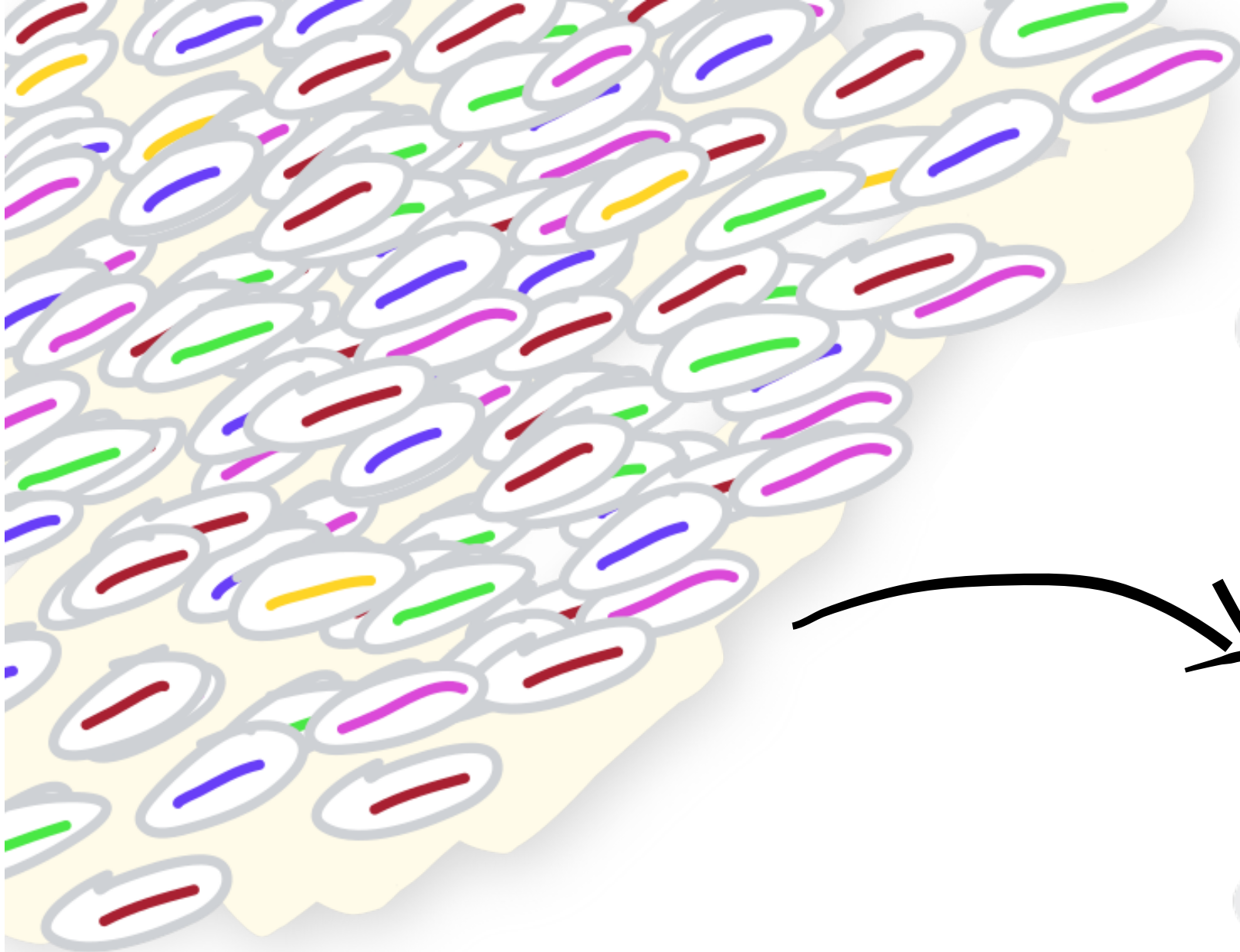


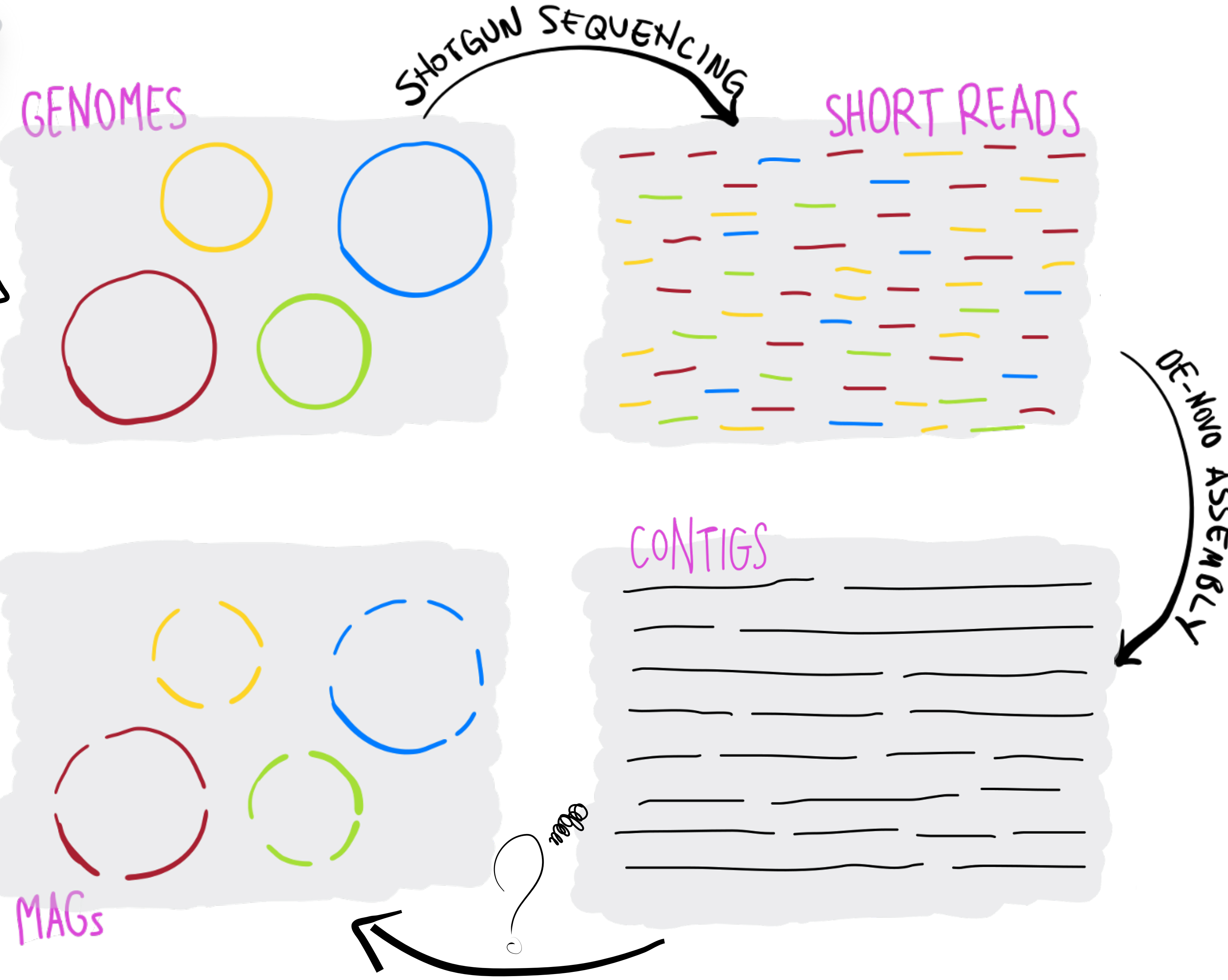
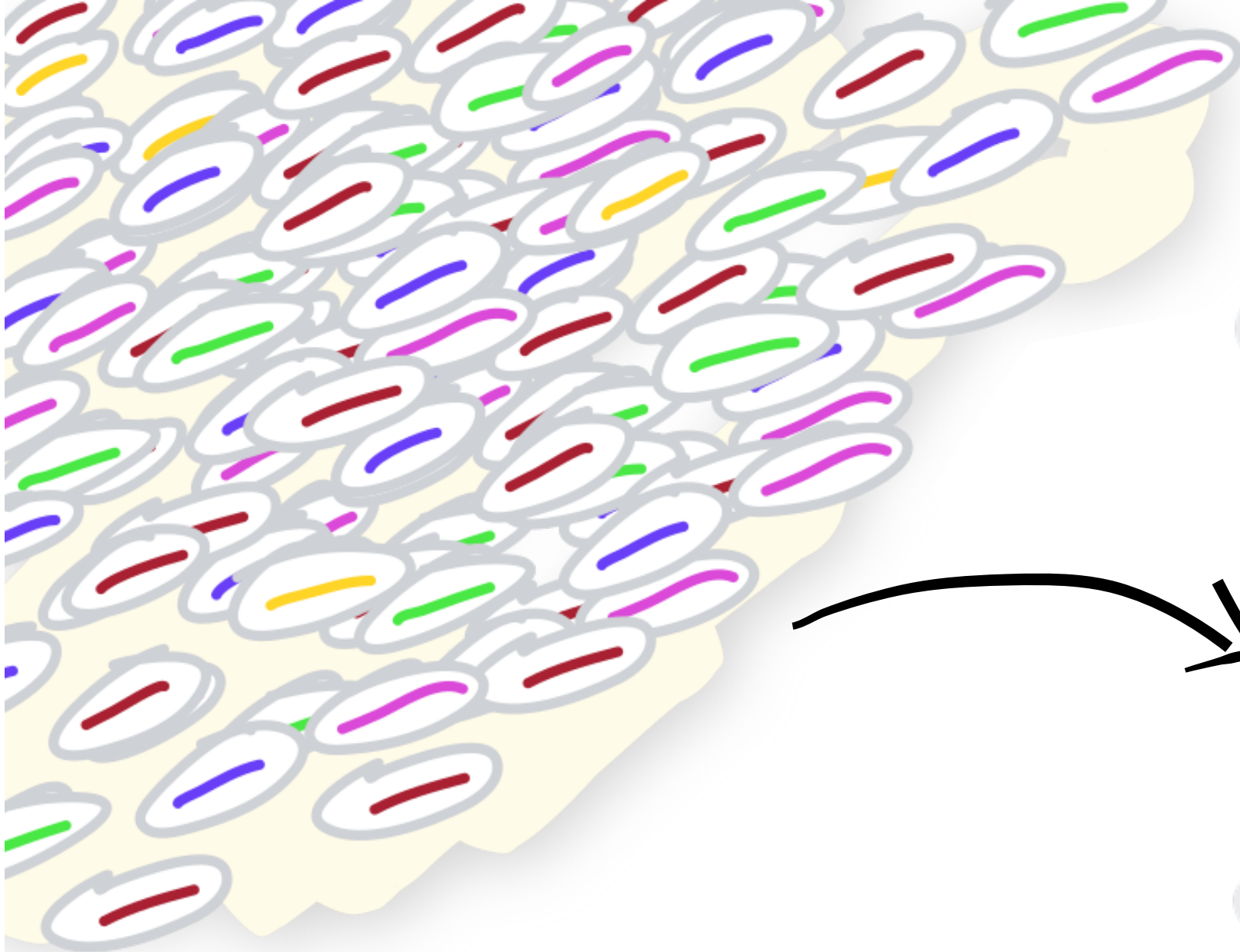
DE-NOVO ASSEMBLY



CONTIGS







Sequence composition

Computing k-mer frequencies

GTTTGGCATGATTAAGGAGTTTCTTTGTGCTTC

k=2

GTTTGGCATGATTAAAGGAGTTTCTTTGTGCTTC

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

k=2

GTTTTGGCATGATTAAAGGAGTTTCTTTTGTGCTTC

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

CTT TGGCATGATTAAAGGAGTTTCTTTTGTGCTTC

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

GTTTGGCATGATTAAAGGAGTTTCTTTGTGCTTC

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

GTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC

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	0	0	1	2	1	1

GTTTGGCATGATTAAGGAGTTTCTTTTGTGCTTC
GAAGCACAAAAGAAACTCCTTAATCATGCCAAAAC

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

→ PALINDROMES :)

k=2

GTTTGGCATGATTAAAGGAGTTTCTTTGTGCTTC

	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										

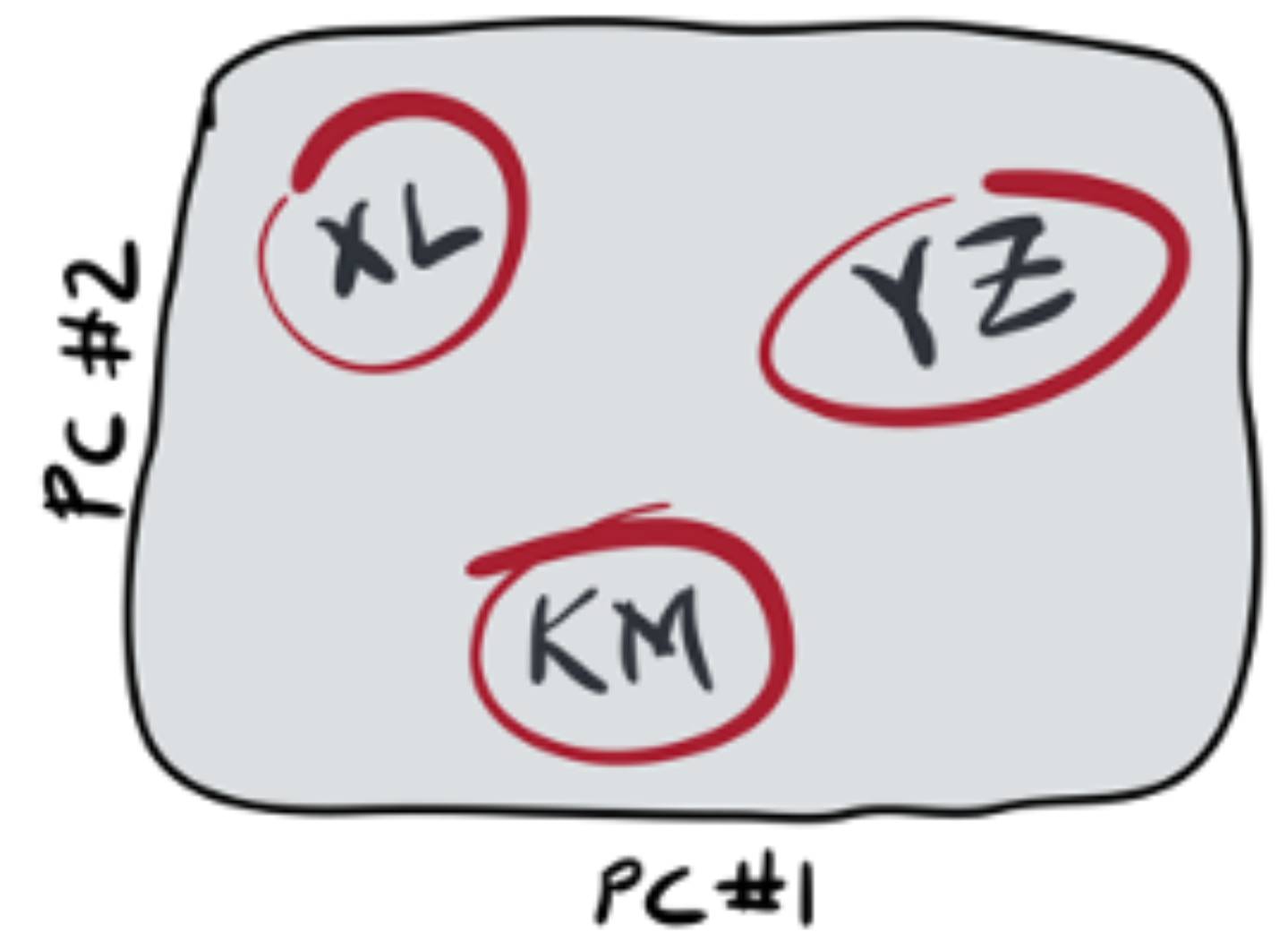
k=2

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

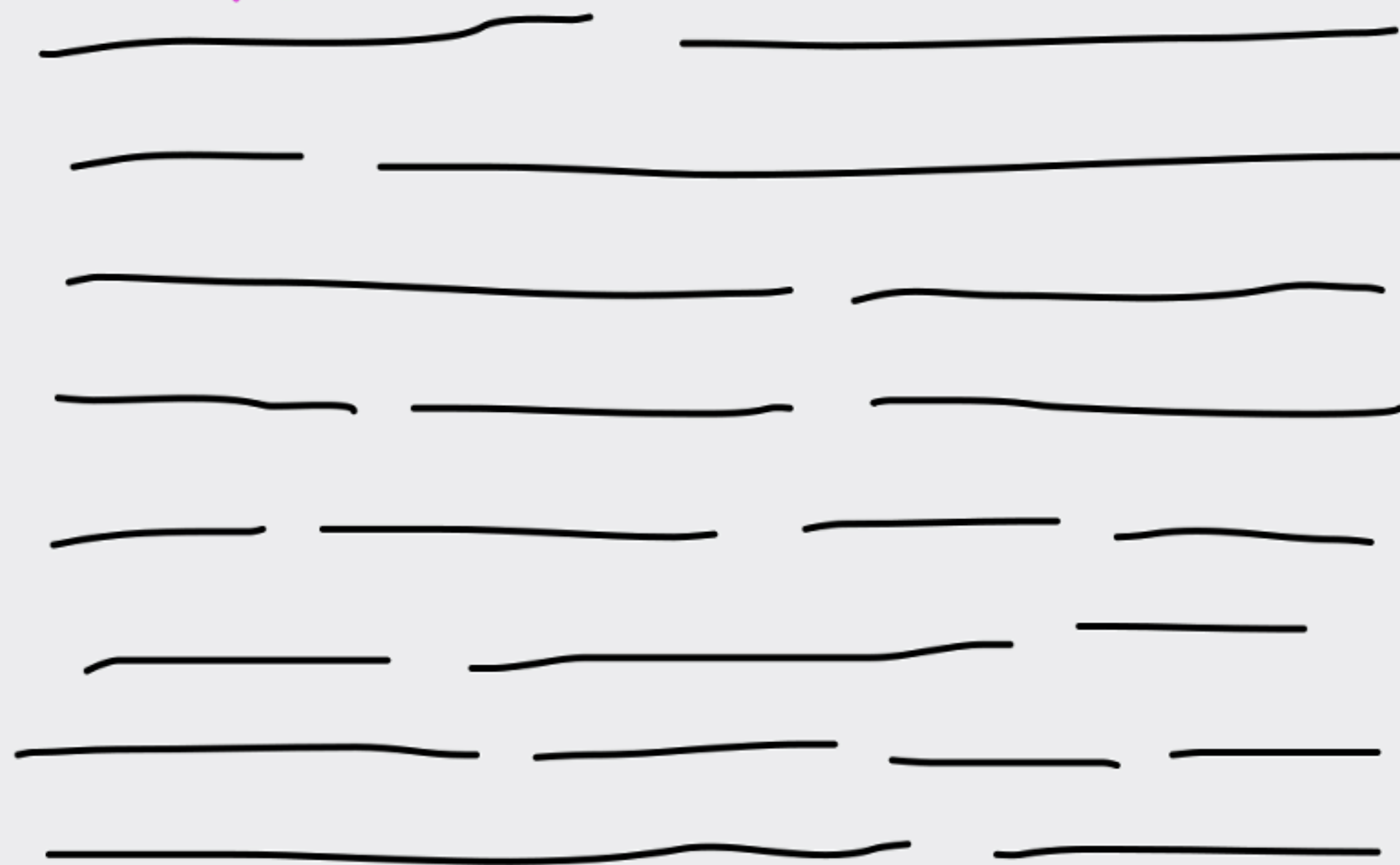


k=2

SEQUENCE COMPOSITION



CONTIGS



MAGs



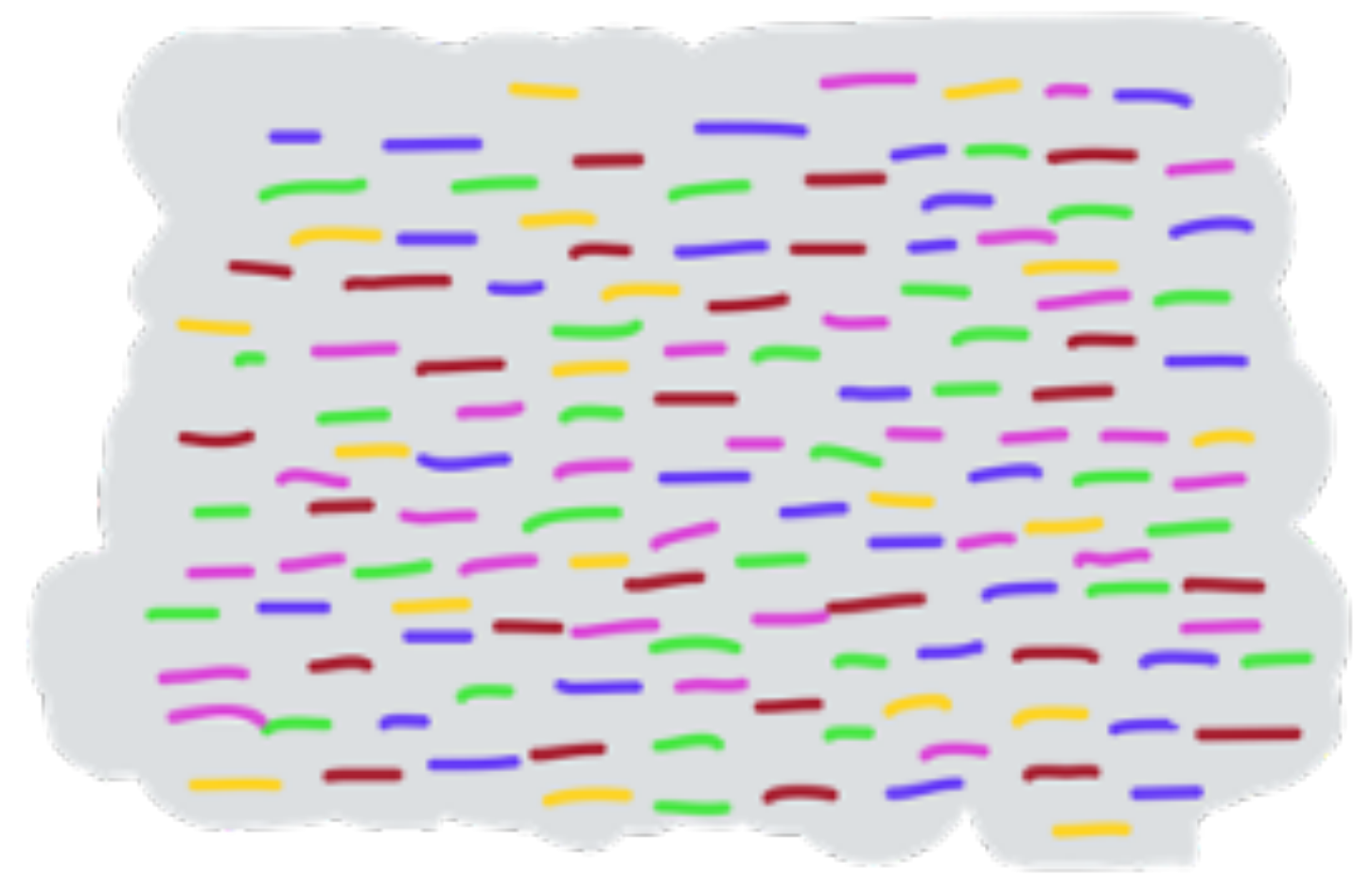
Abundance correlation

Counting stuff across samples

CONTIG #1

CONTIG #2

CONTIG #1

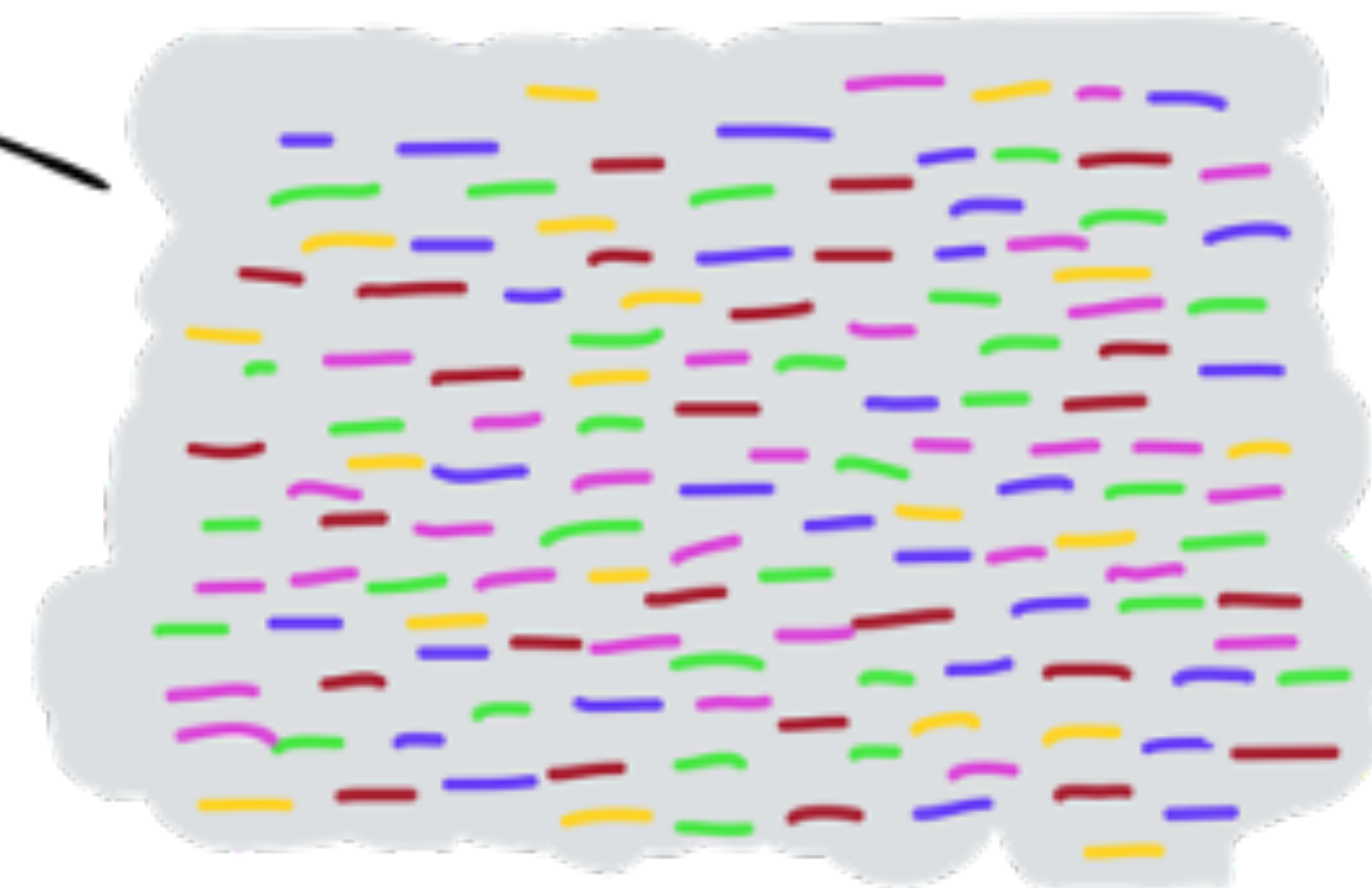


METAGENOMIC READS

CONTIG #2

CONTIG #1

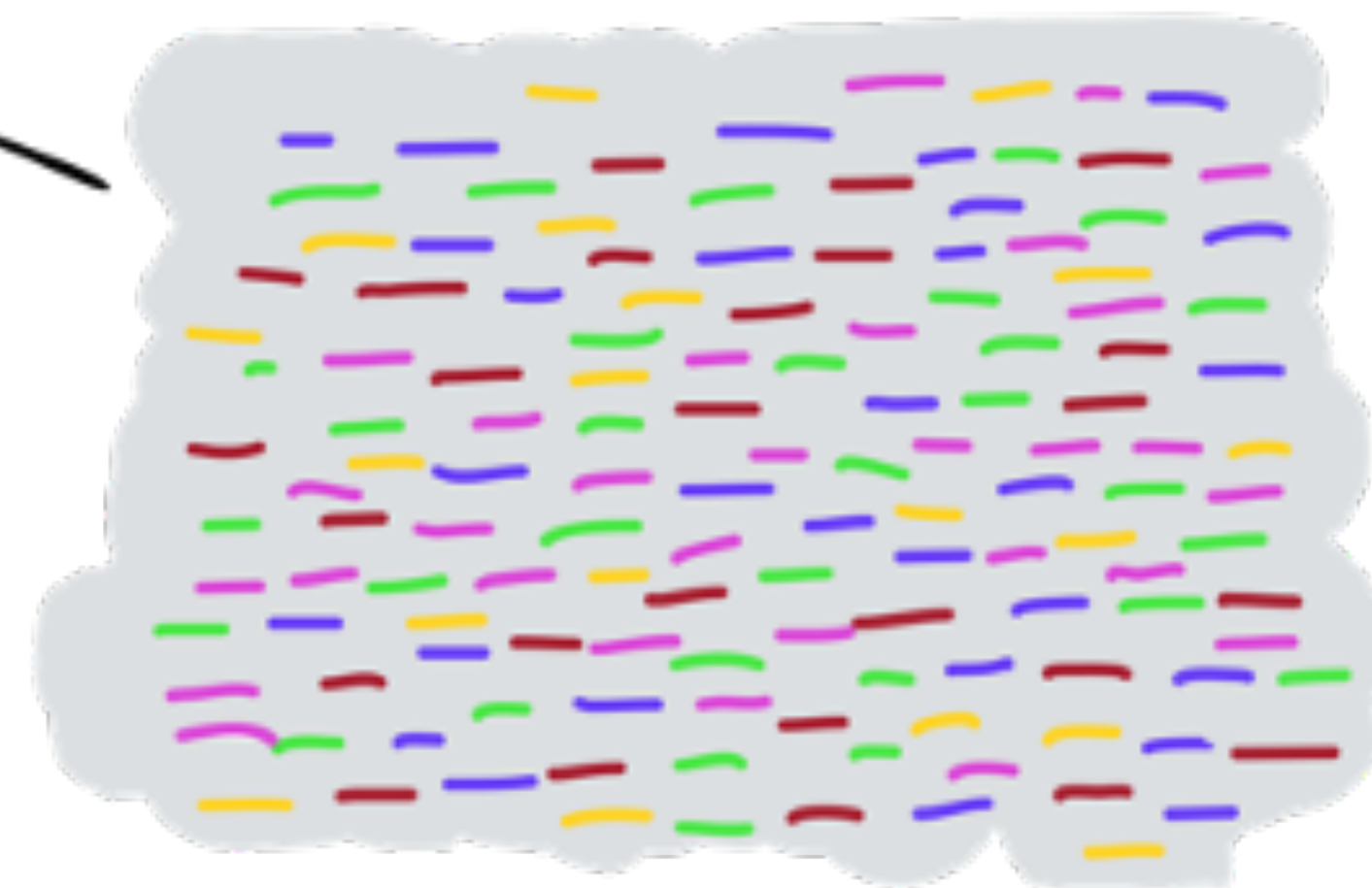
MAPPING



METAGENOMIC READS

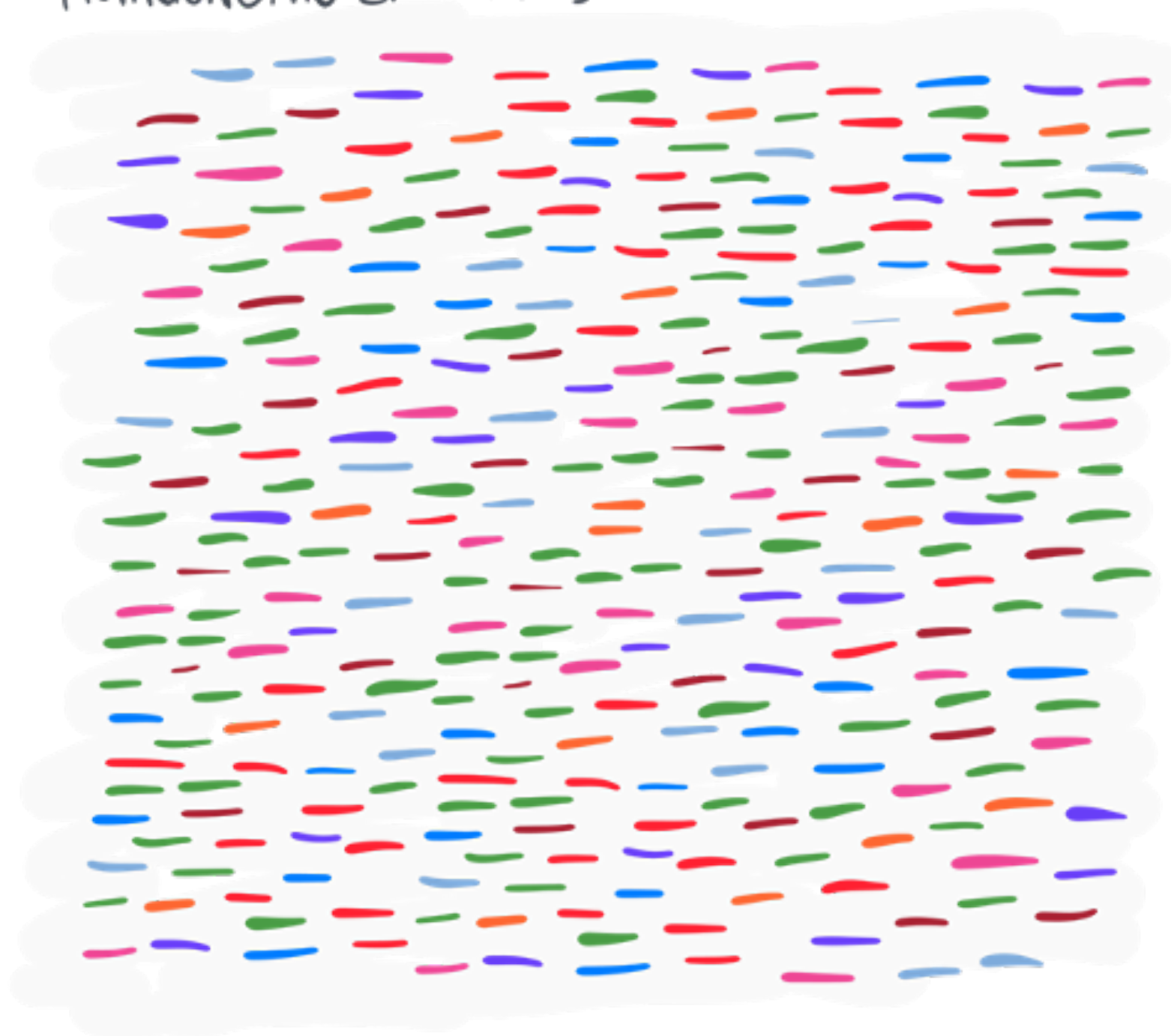
CONTIG #2

CONTIG #1



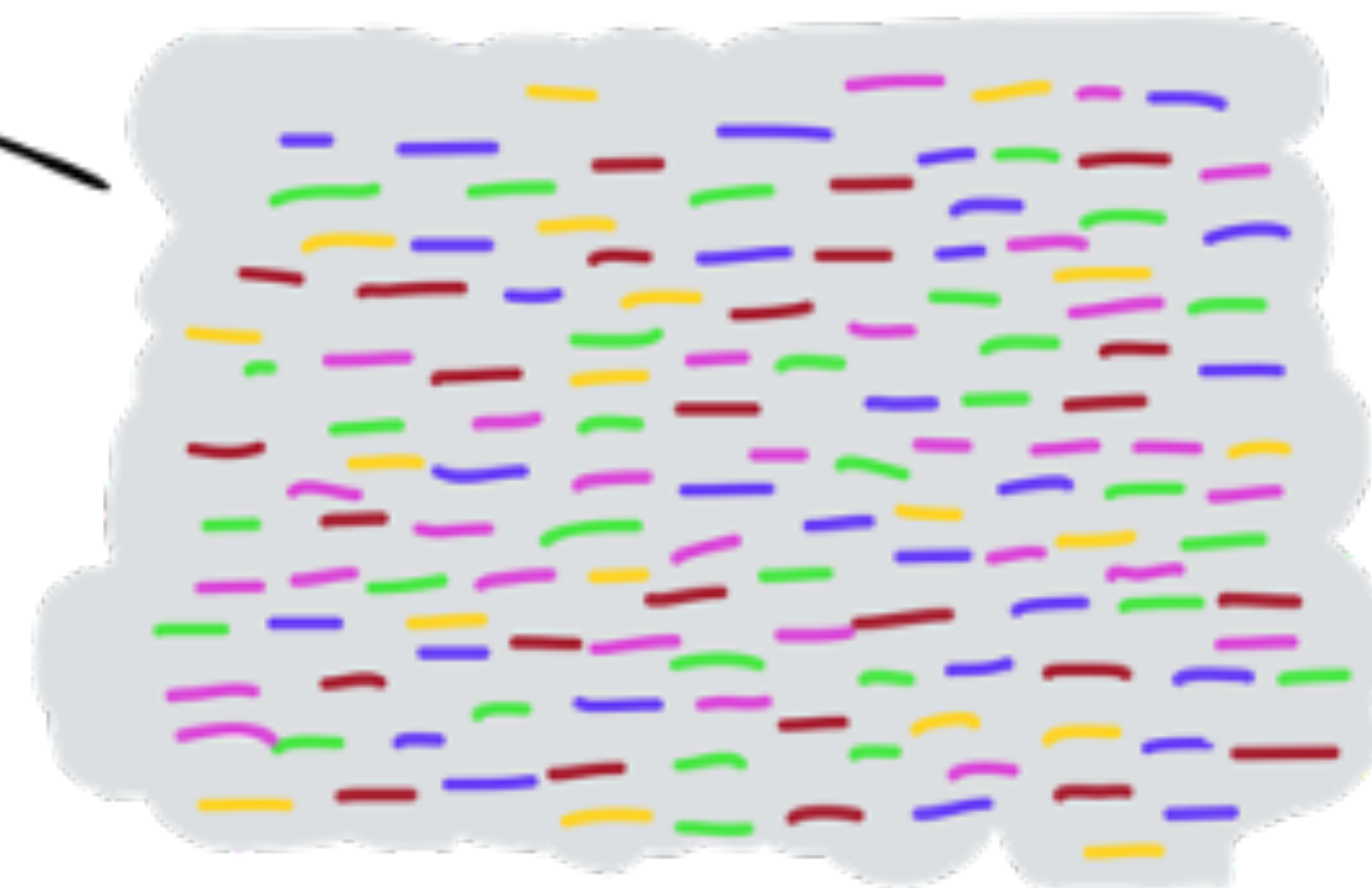
METAGENOMIC READS

METAGENOMIC SHORT READS



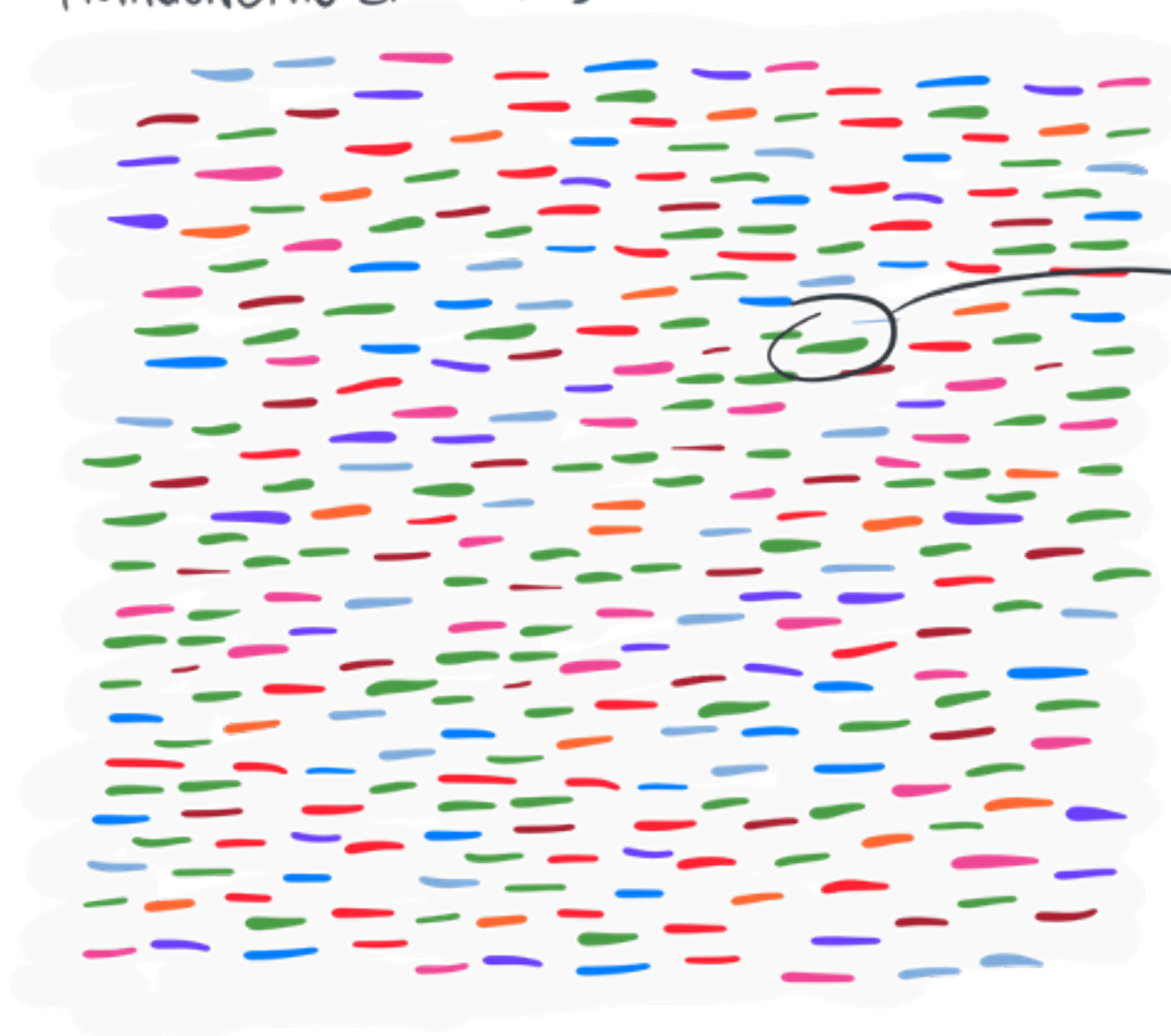
CONTIG #1

MAPPING



METAGENOMIC READS

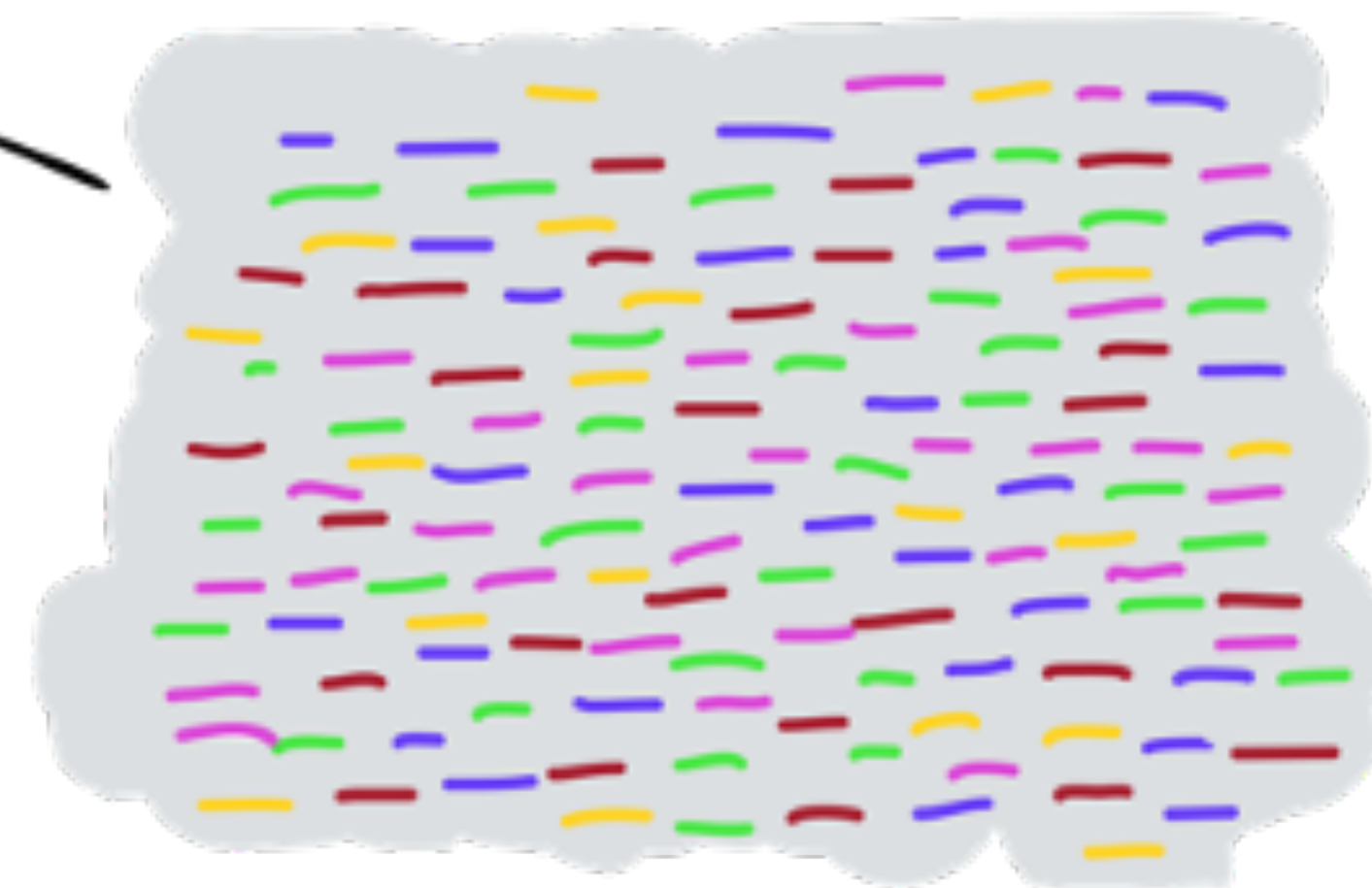
METAGENOMIC SHORT READS



READ
RECRUITMENT

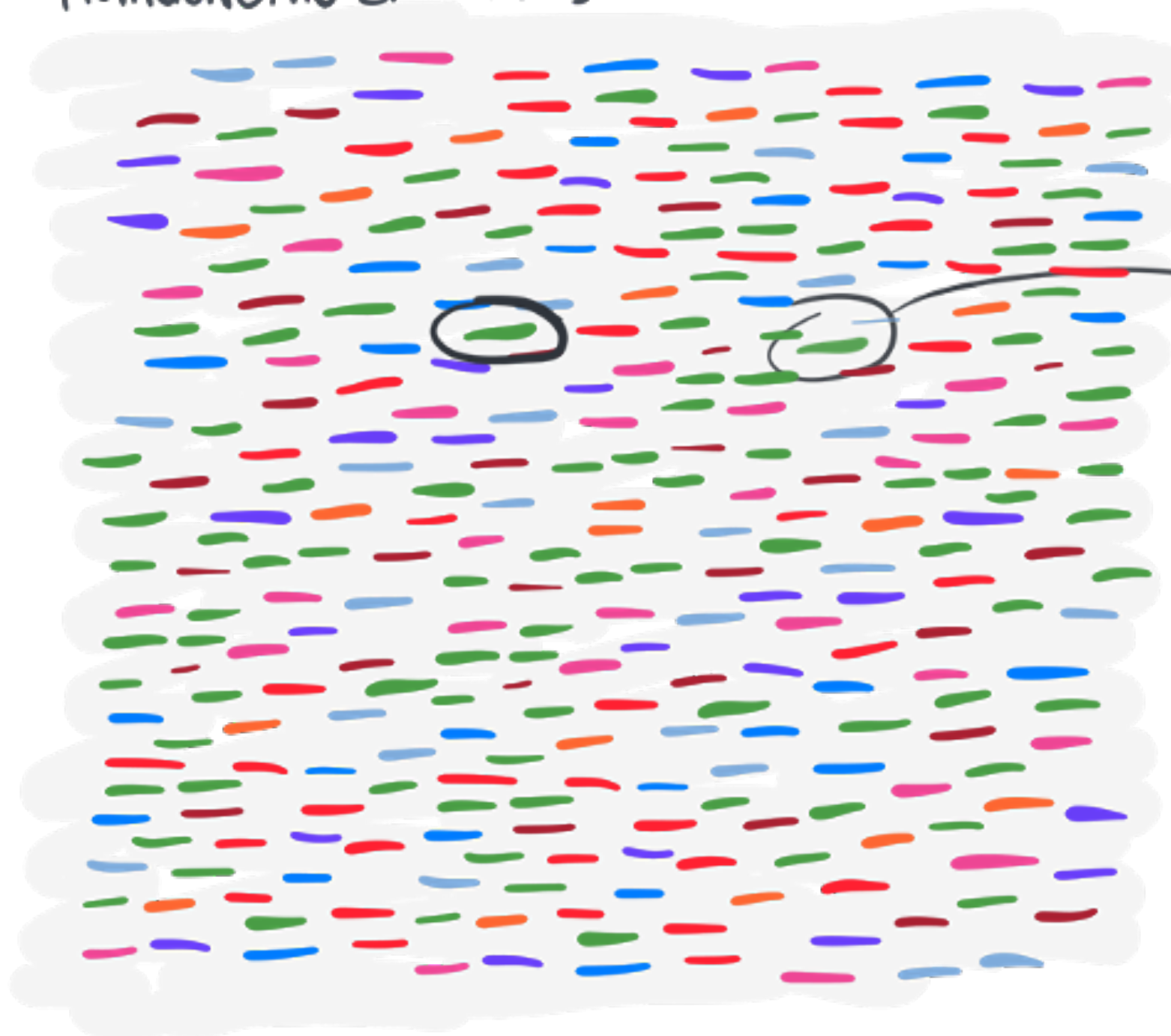


CONTIG #1

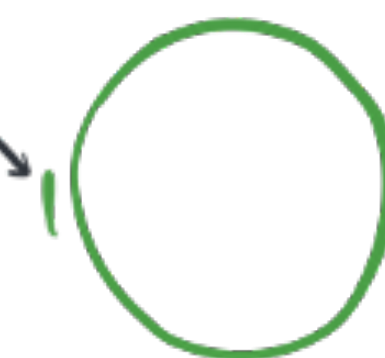


METAGENOMIC READS

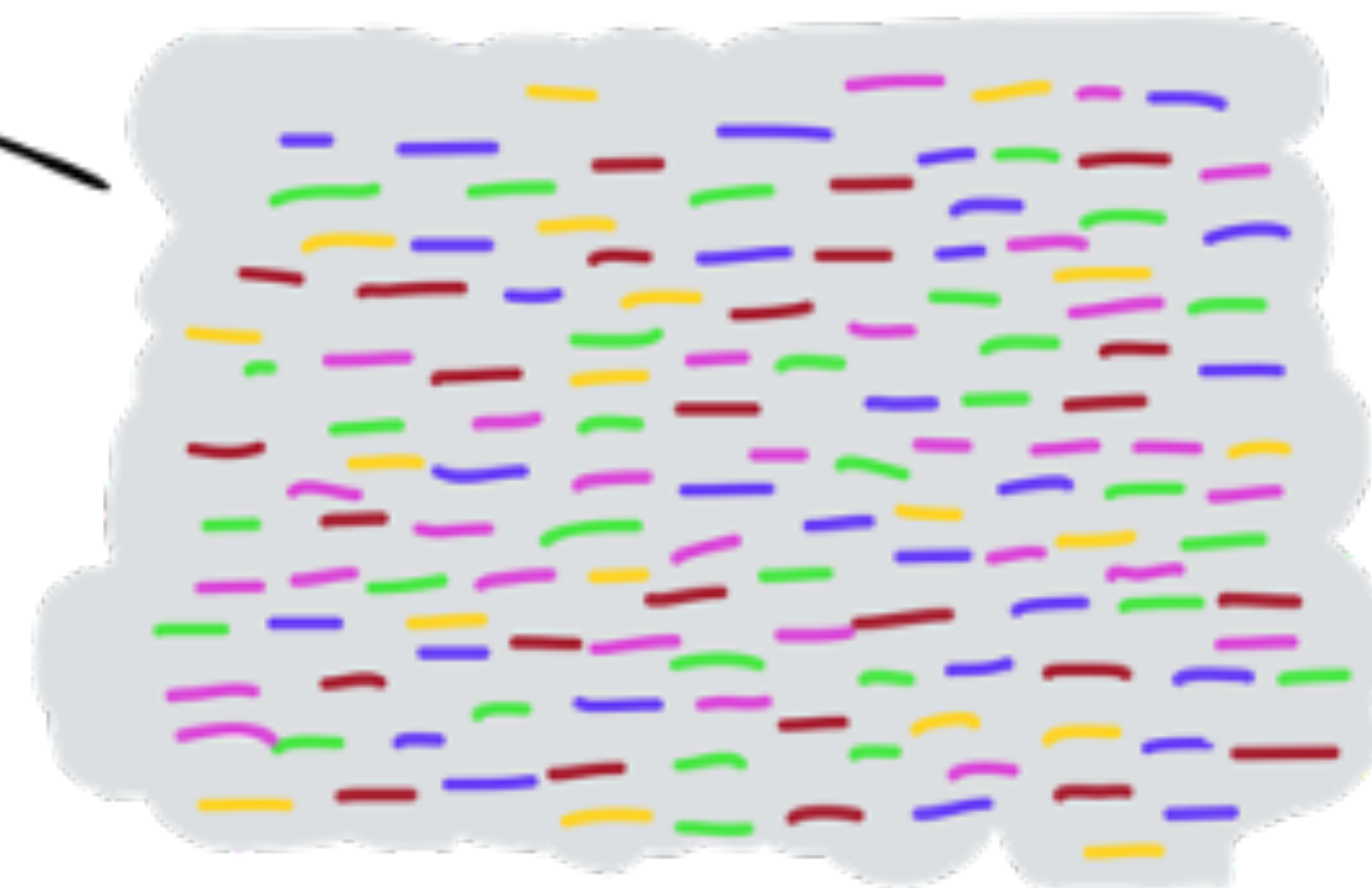
METAGENOMIC SHORT READS



READ
RECRUITMENT

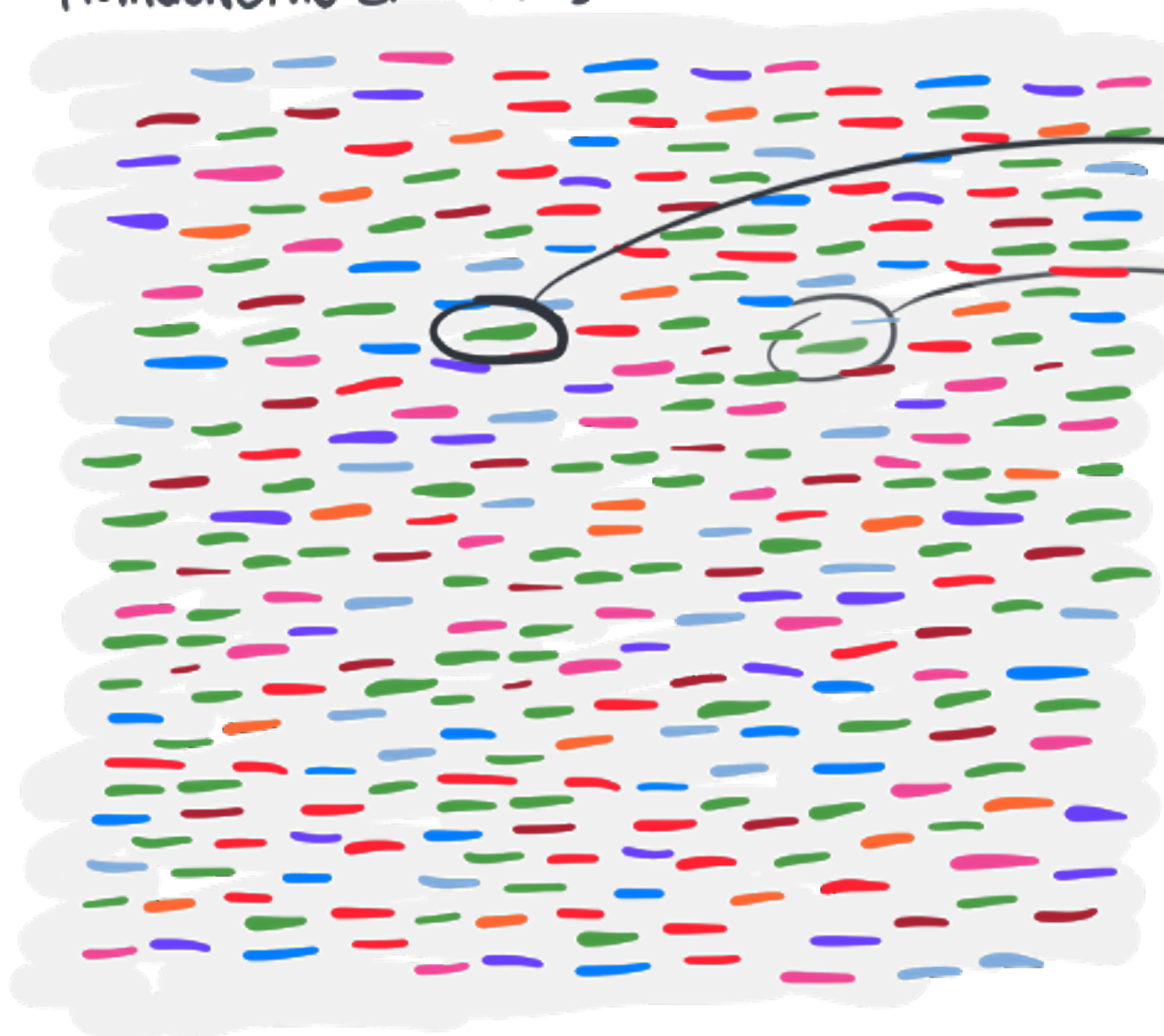


CONTIG #1

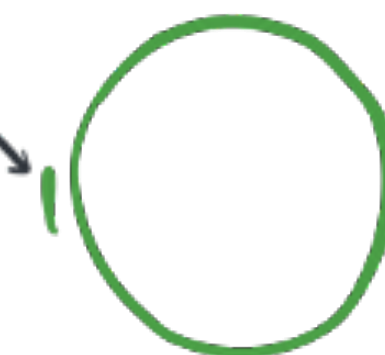


METAGENOMIC READS

METAGENOMIC SHORT READS

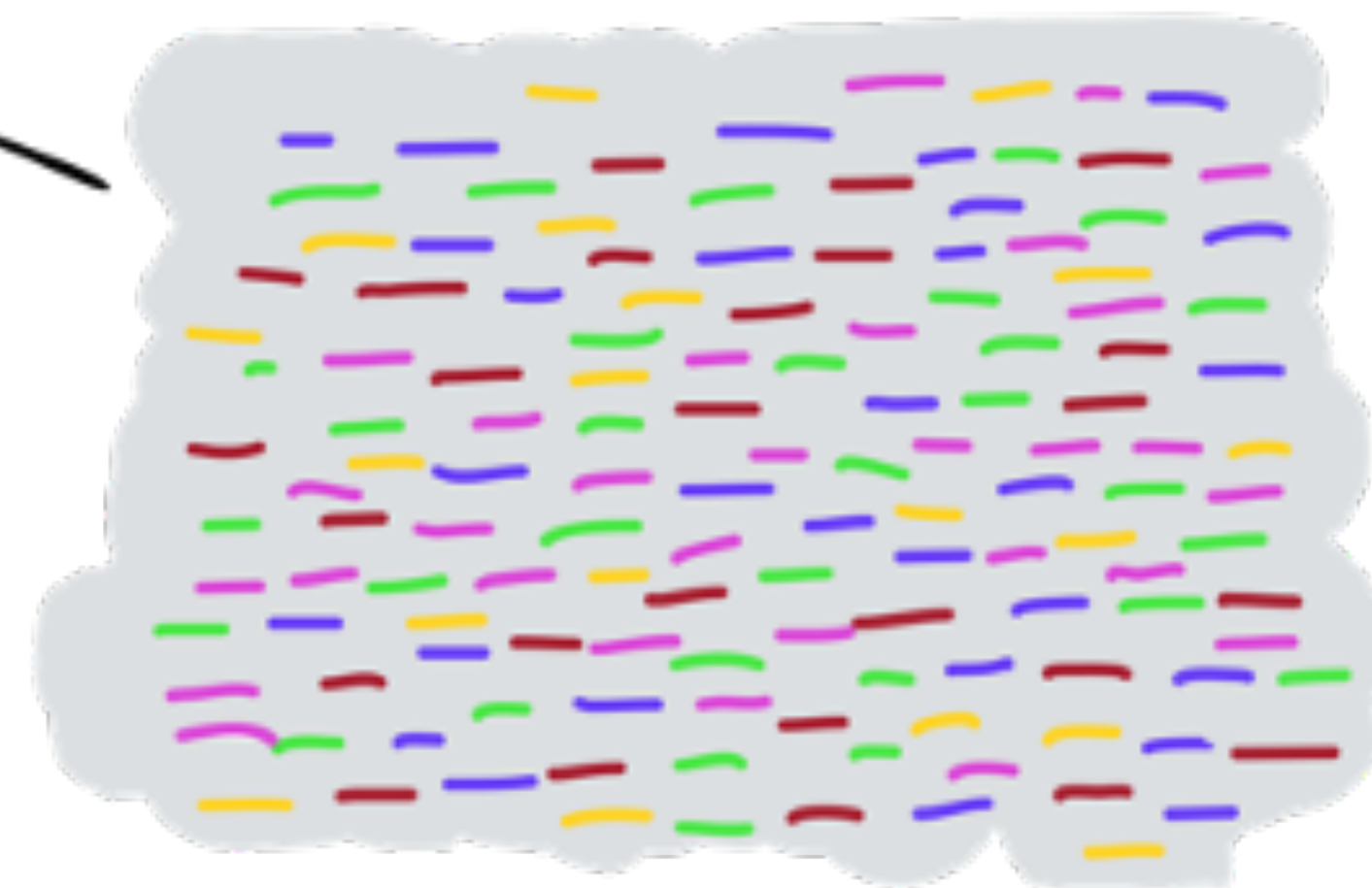


READ
RECRUITMENT



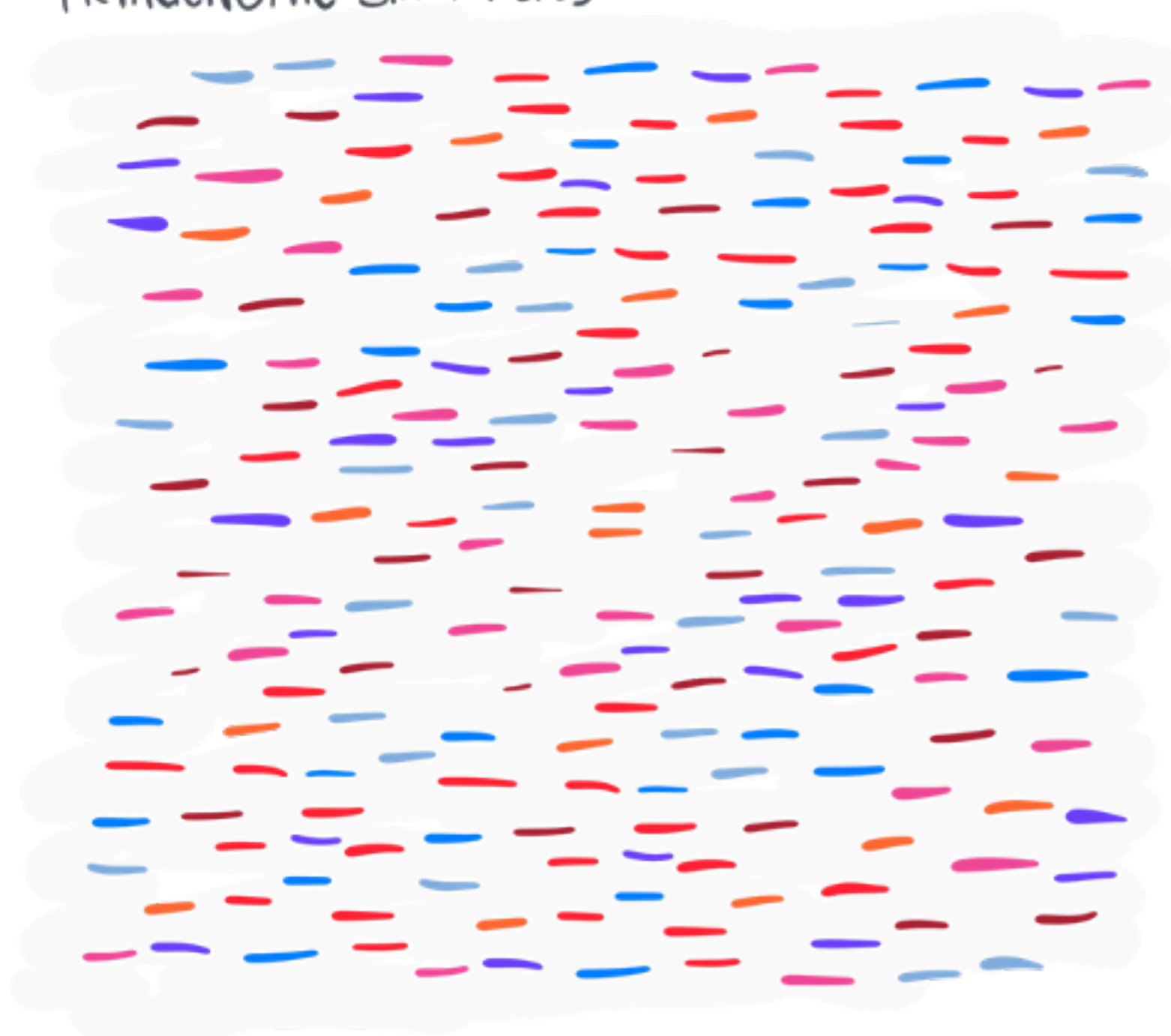
CONTIG #1

MAPPING



METAGENOMIC READS

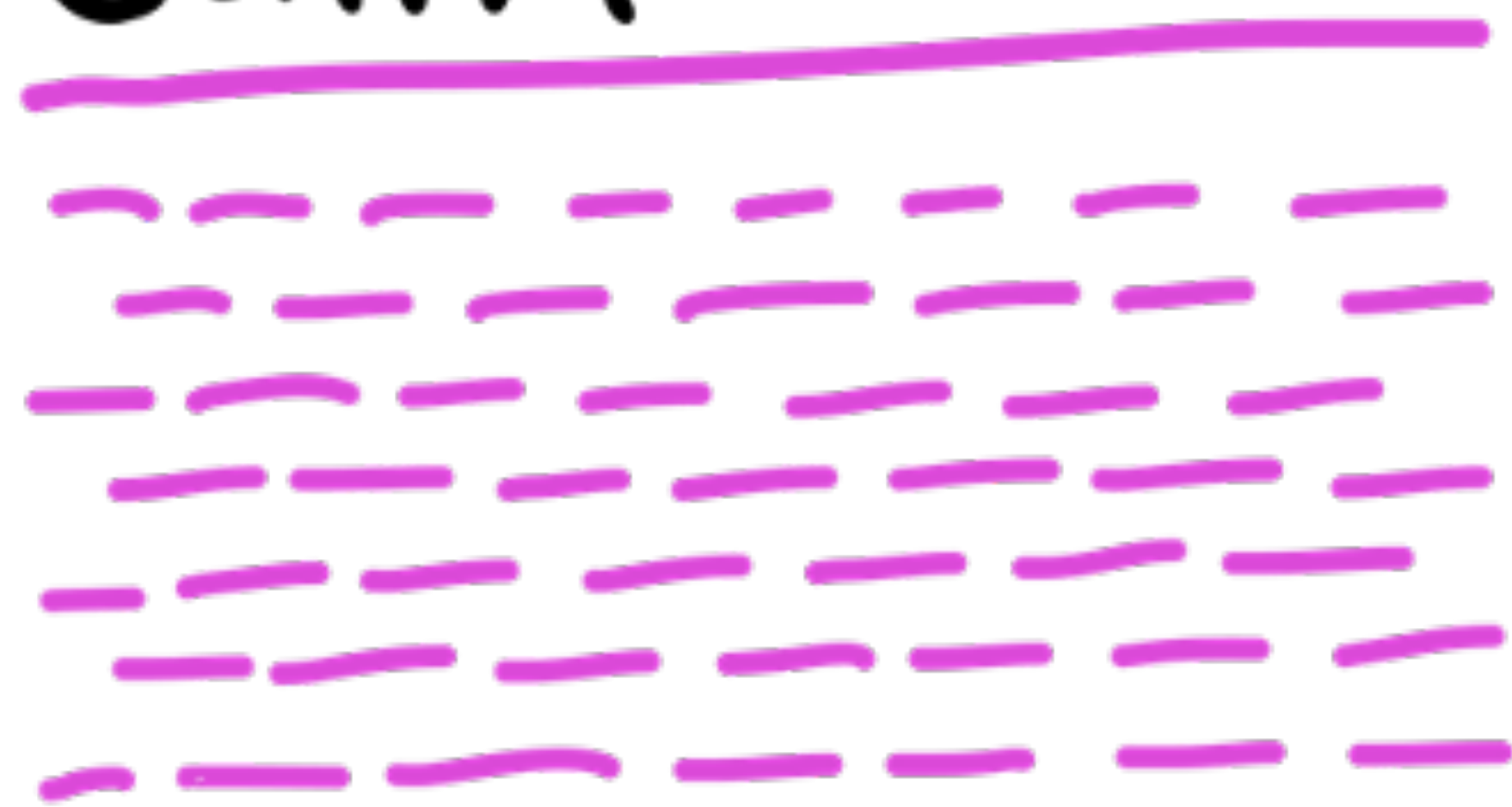
METAGENOMIC SHORT READS



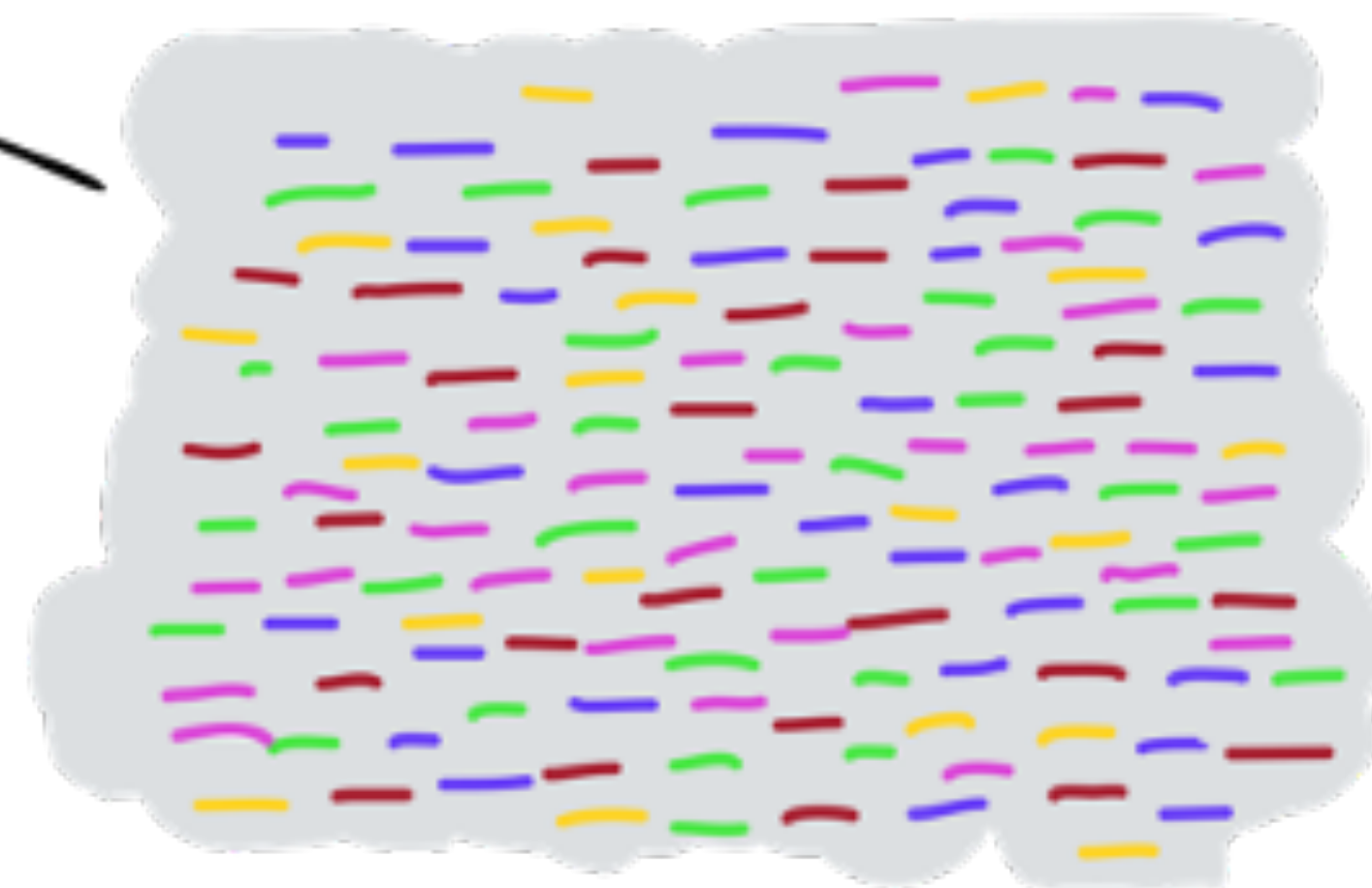
READ
RECRUITMENT



CONTIG #1

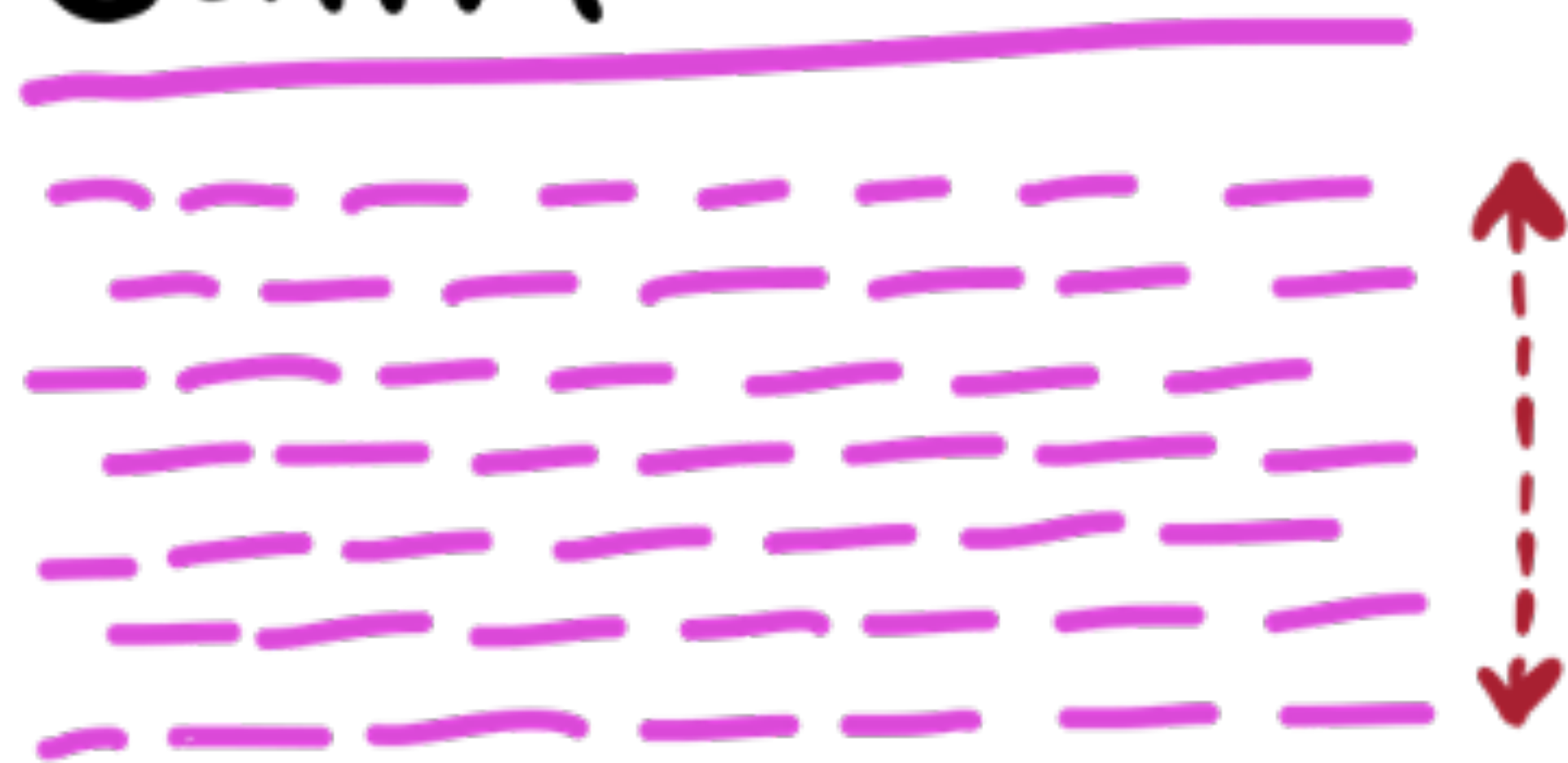


CONTIG #2



METAGENOMIC READS

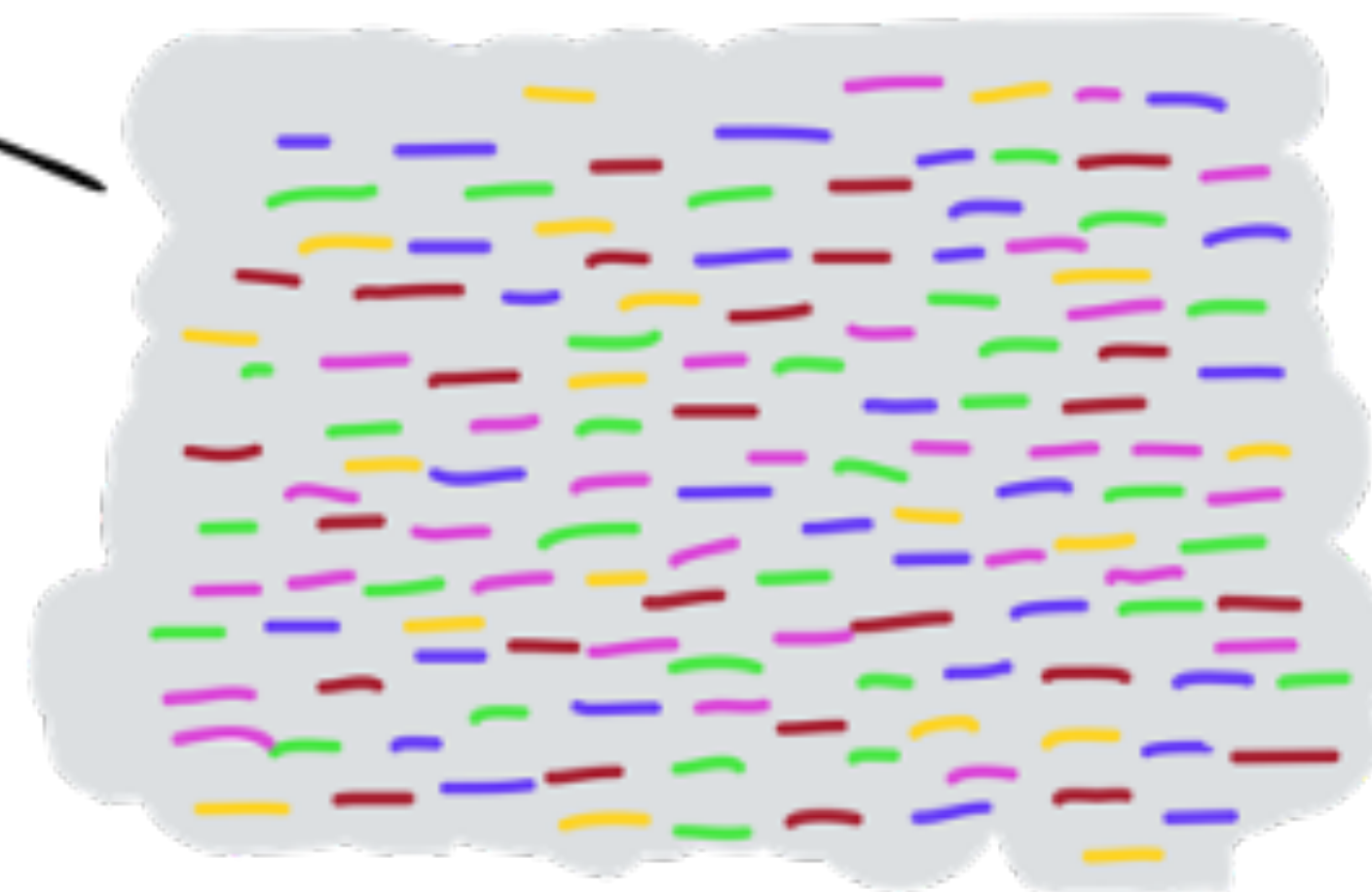
CONTIG #1



CONTIG #2

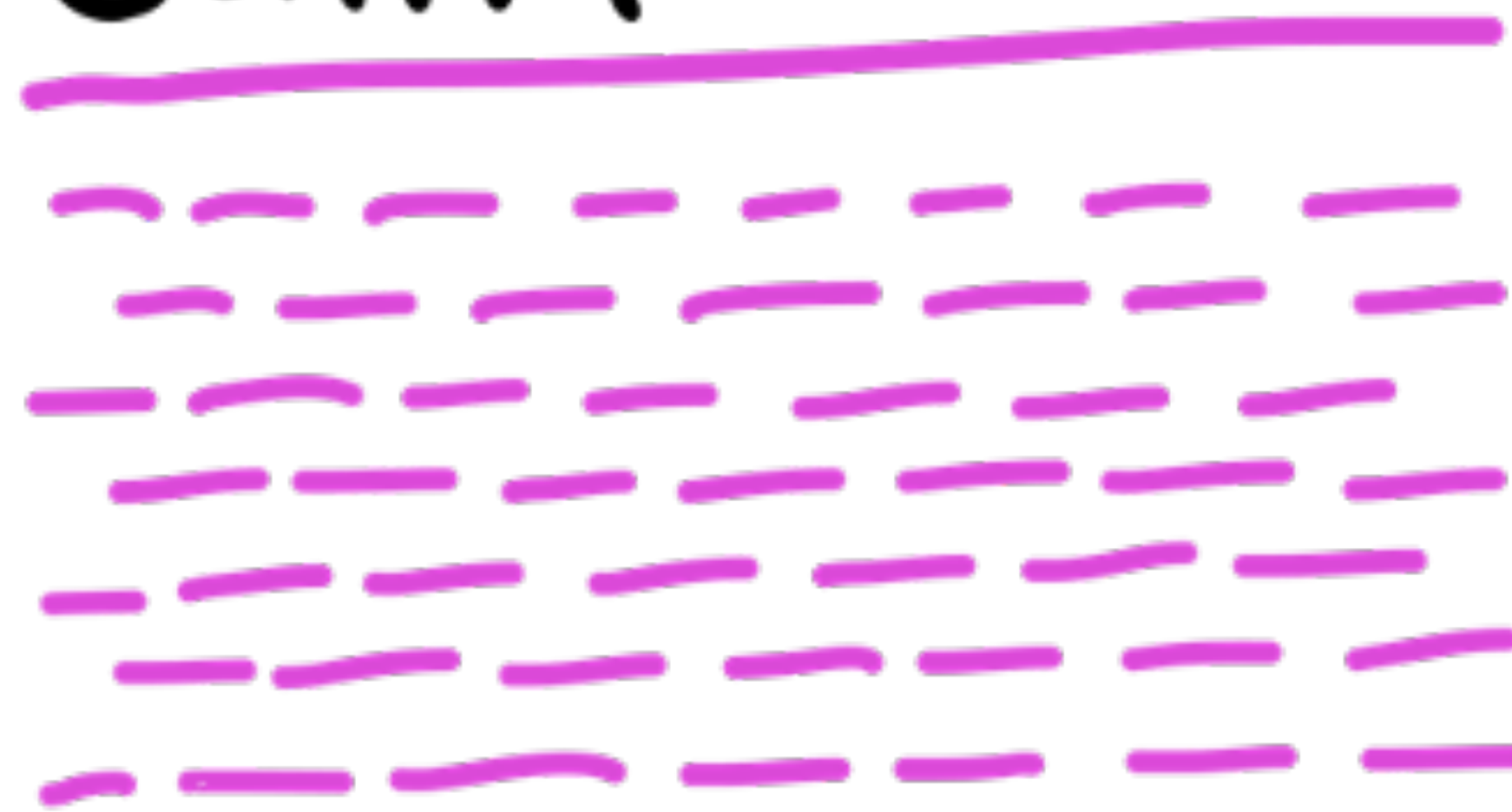


MAPPING



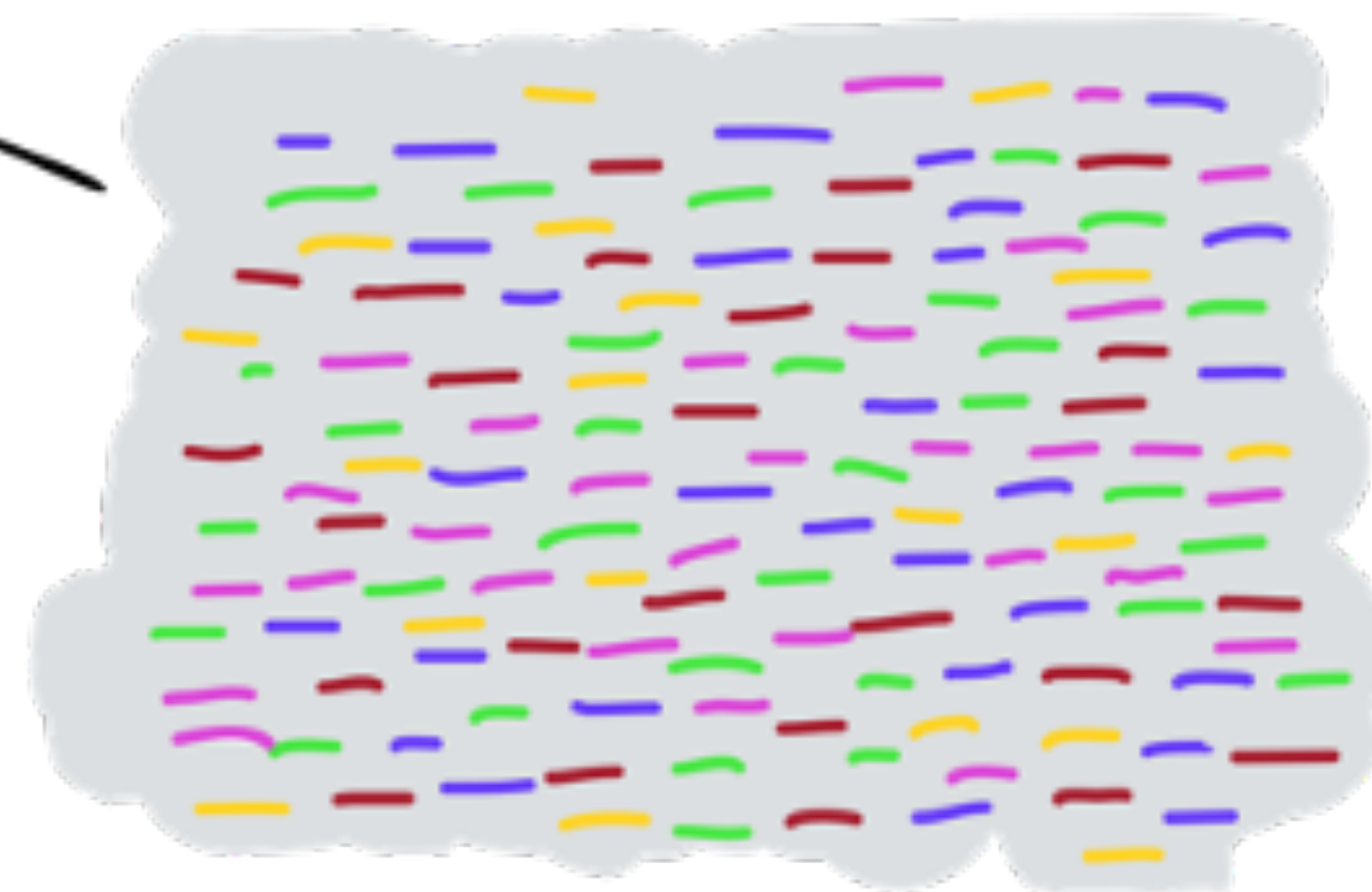
METAGENOMIC READS

CONTIG #1



COVERAGE: ~7X

MAPPING

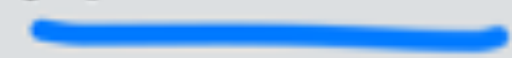
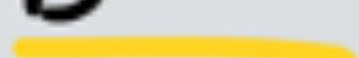


METAGENOMIC READS

CONTIG #2

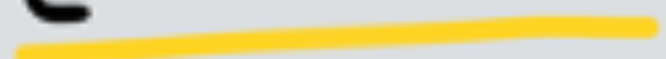


COVERAGE: ~4X

A  B 

C 

D 

E  F 

CONTIGS

S1



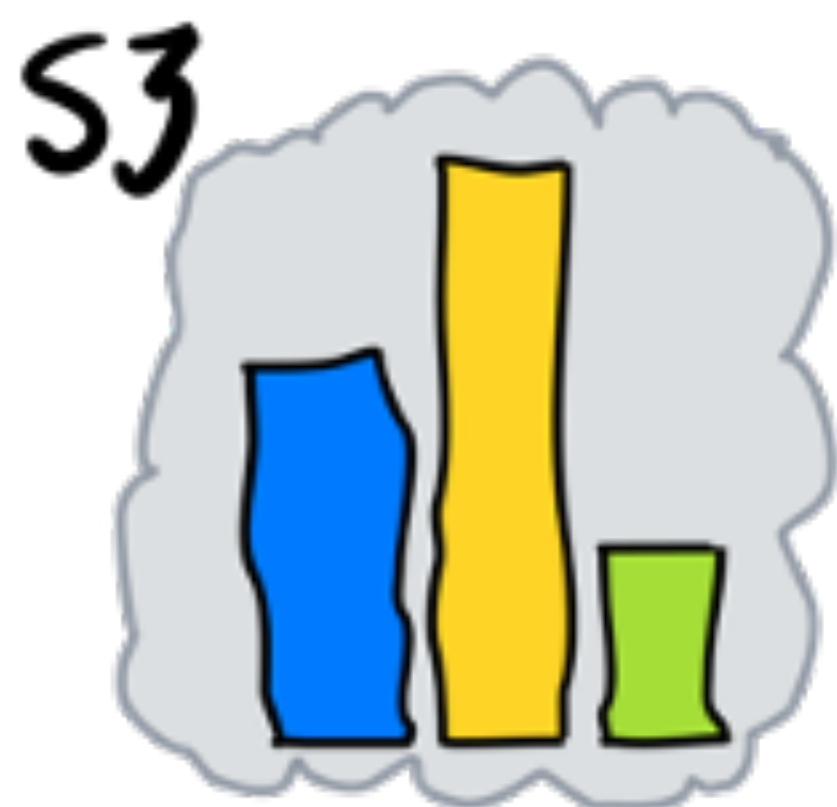
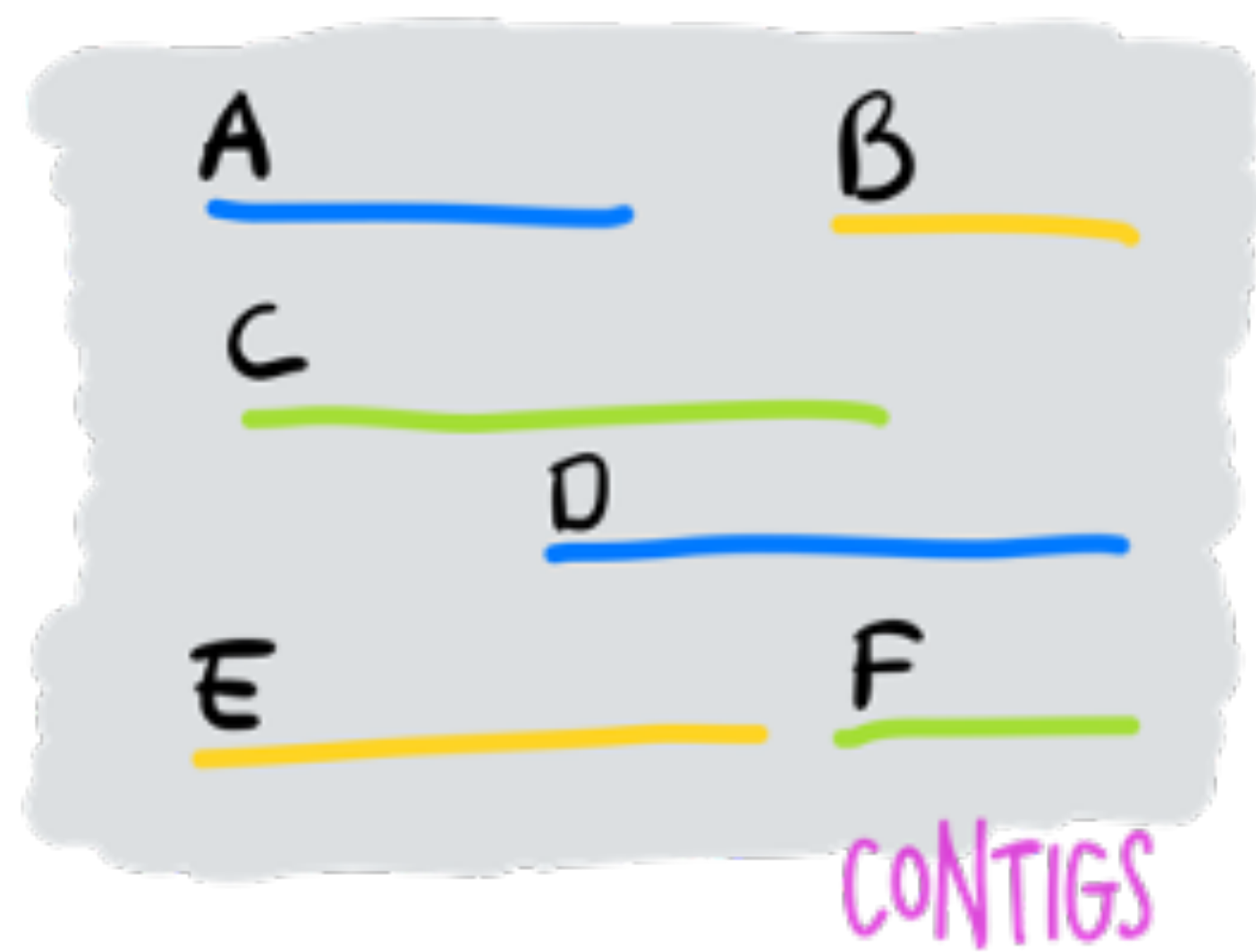
S2

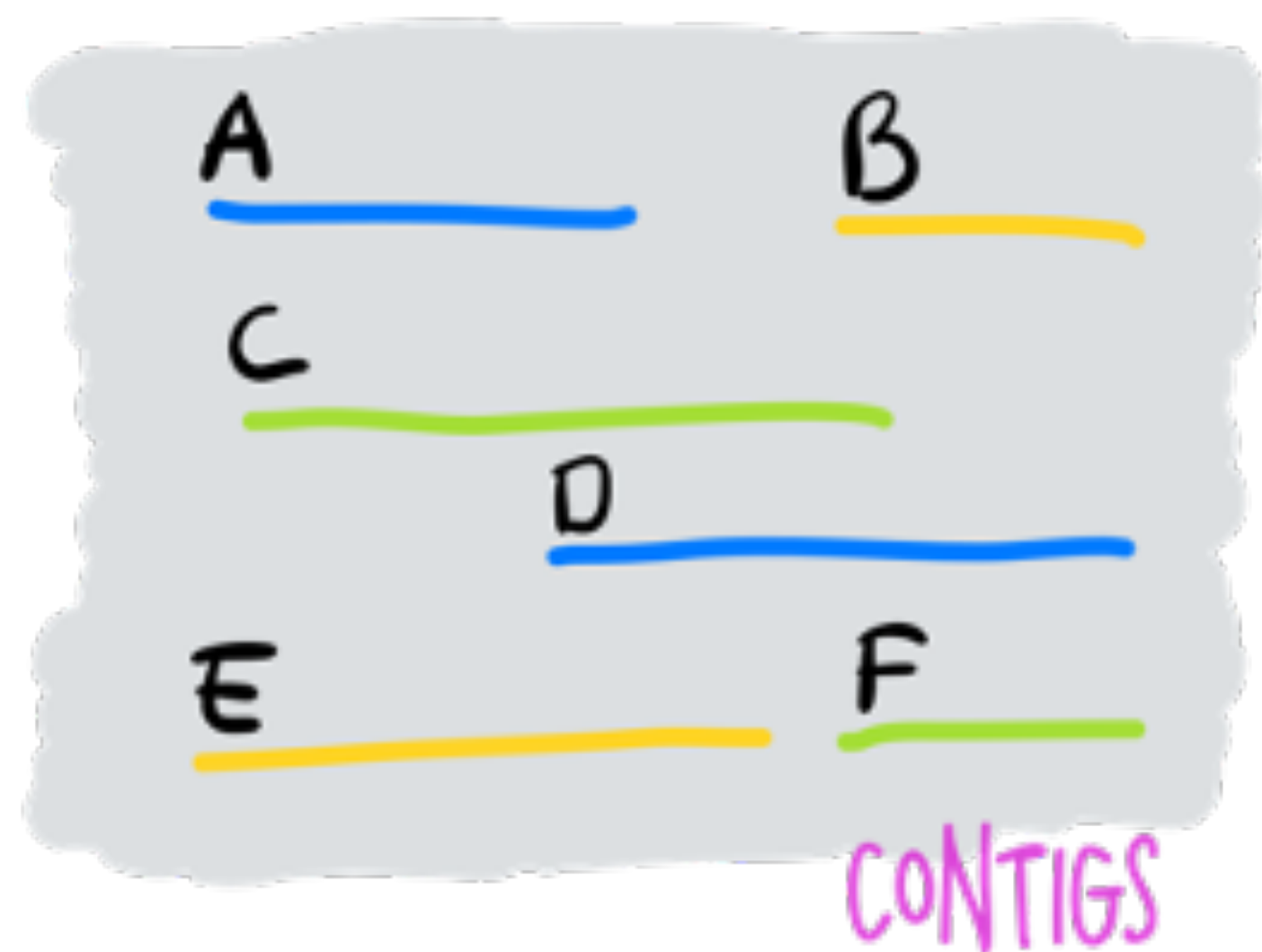


S3



CONTIGS





S1



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

S2

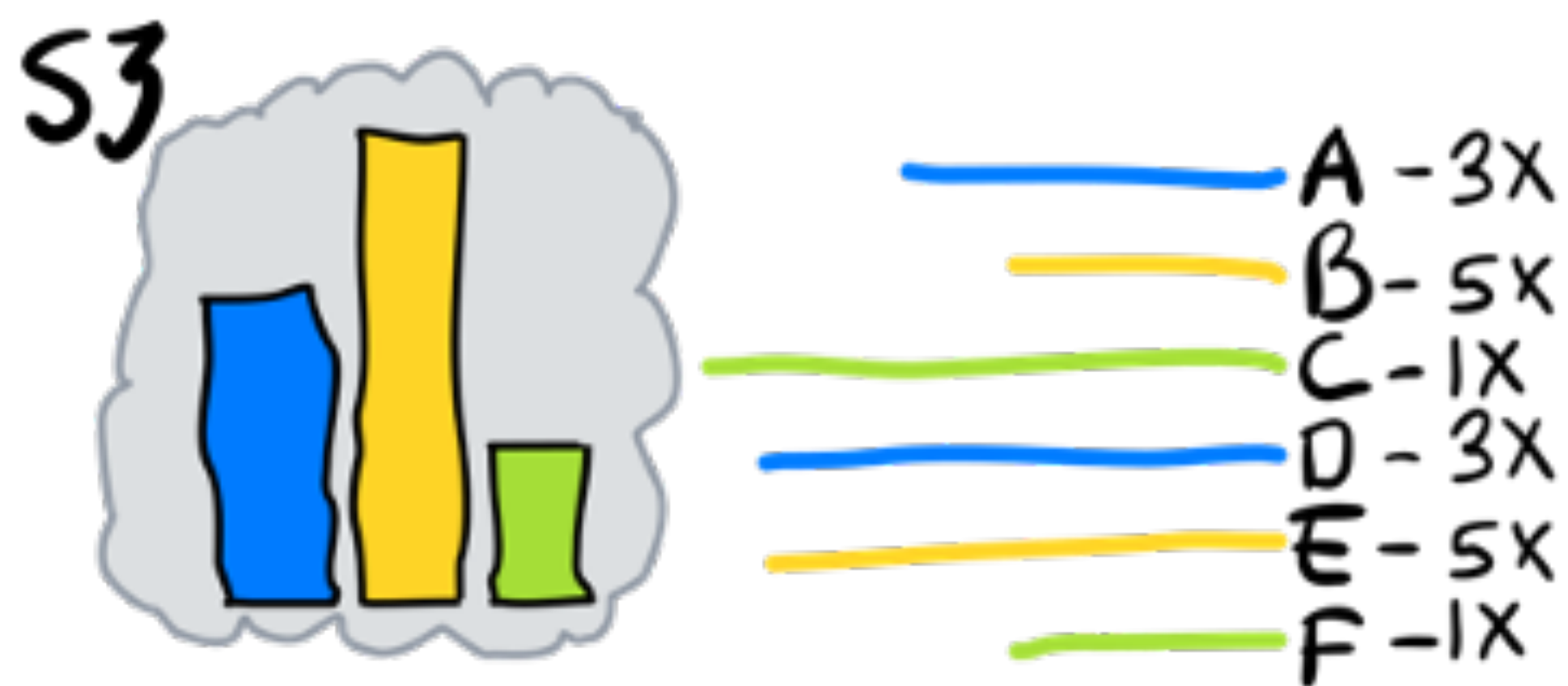
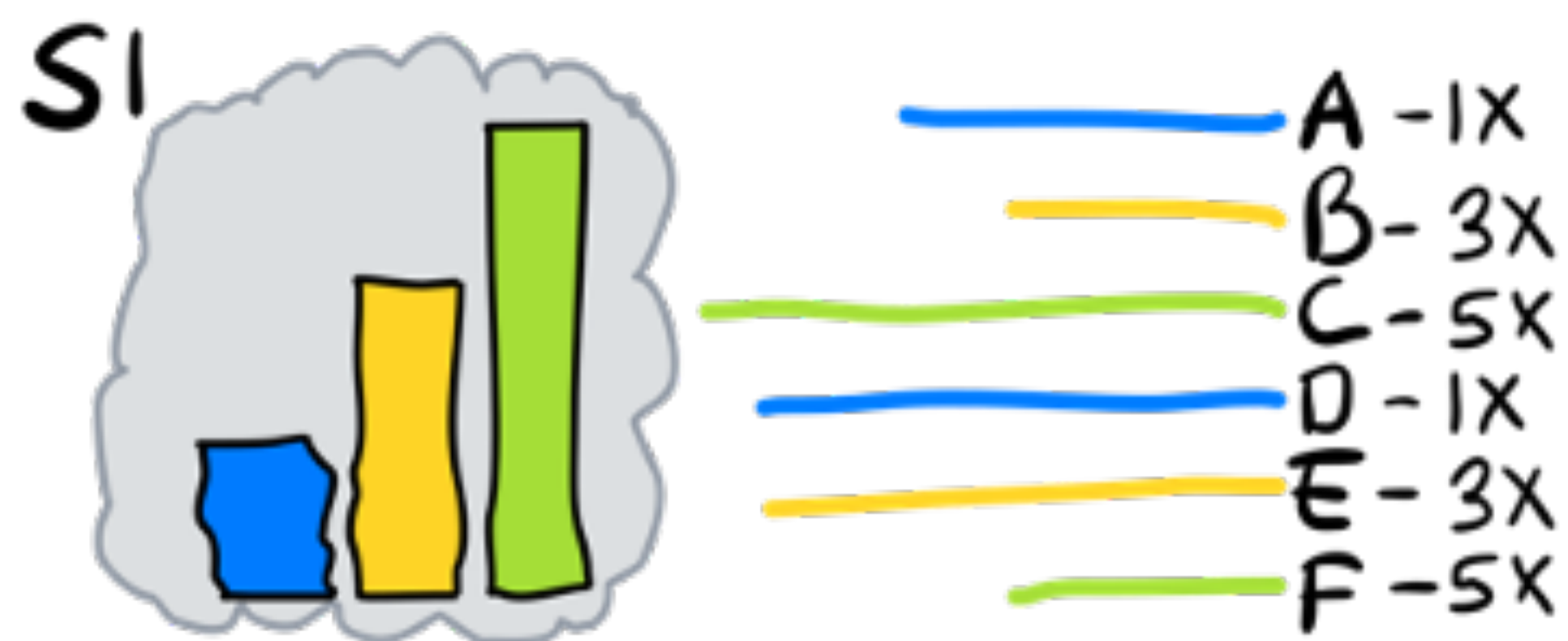
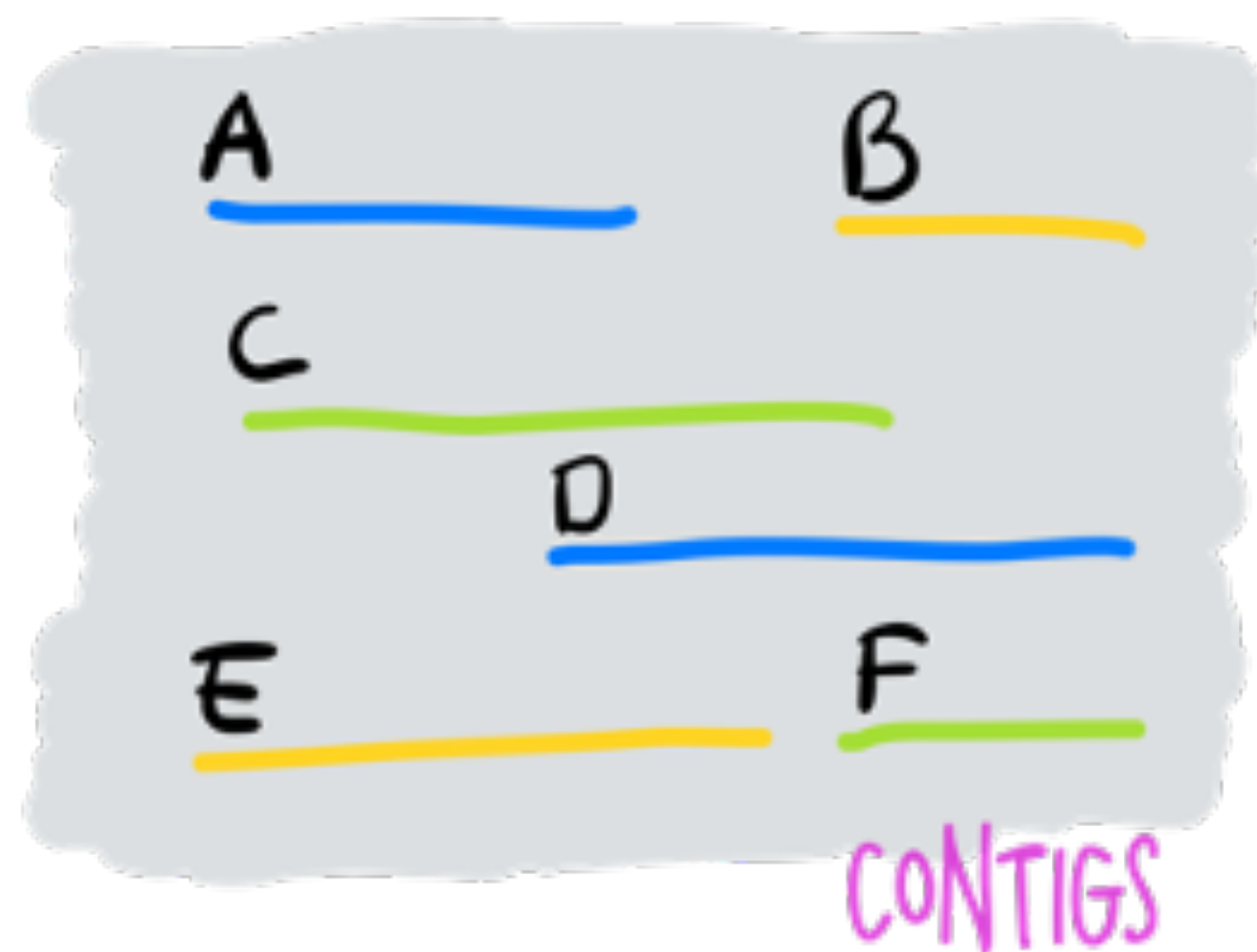


A - 5X
B - 1X
C - 3X
D - 5X
E - 1X
F - 3X

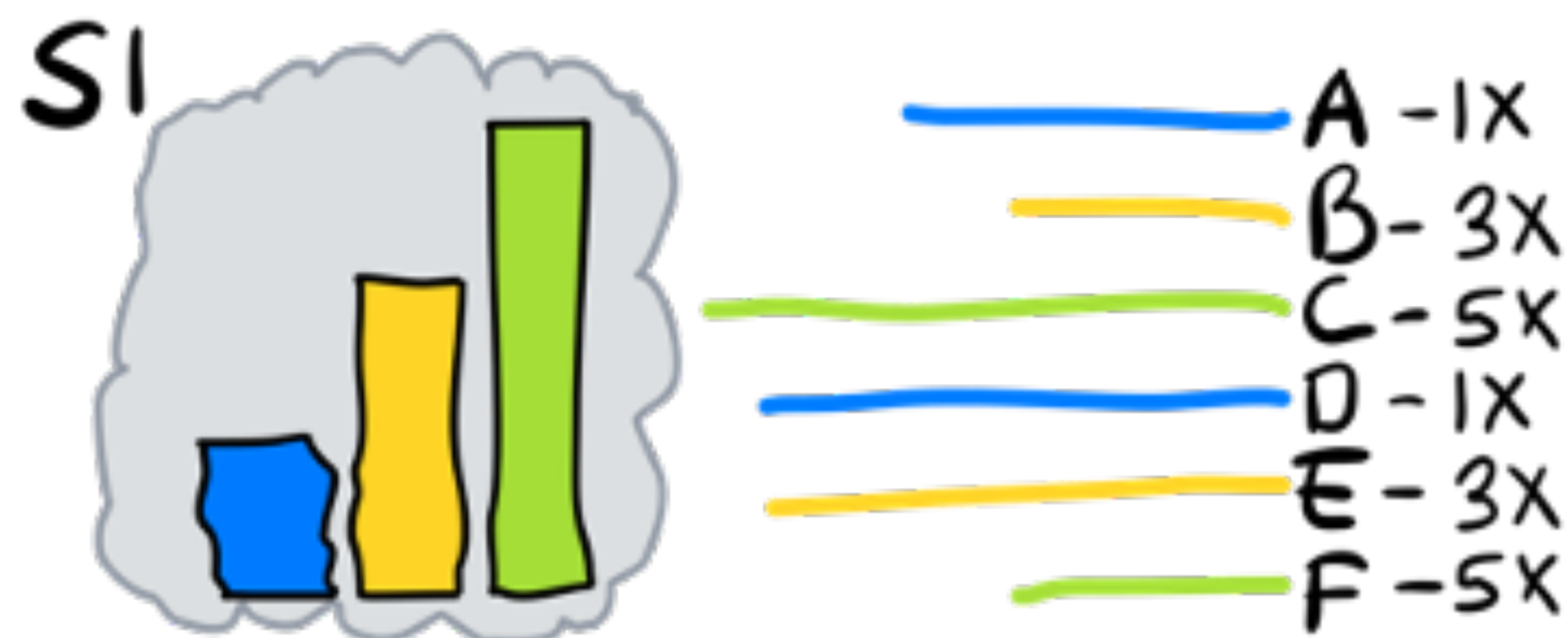
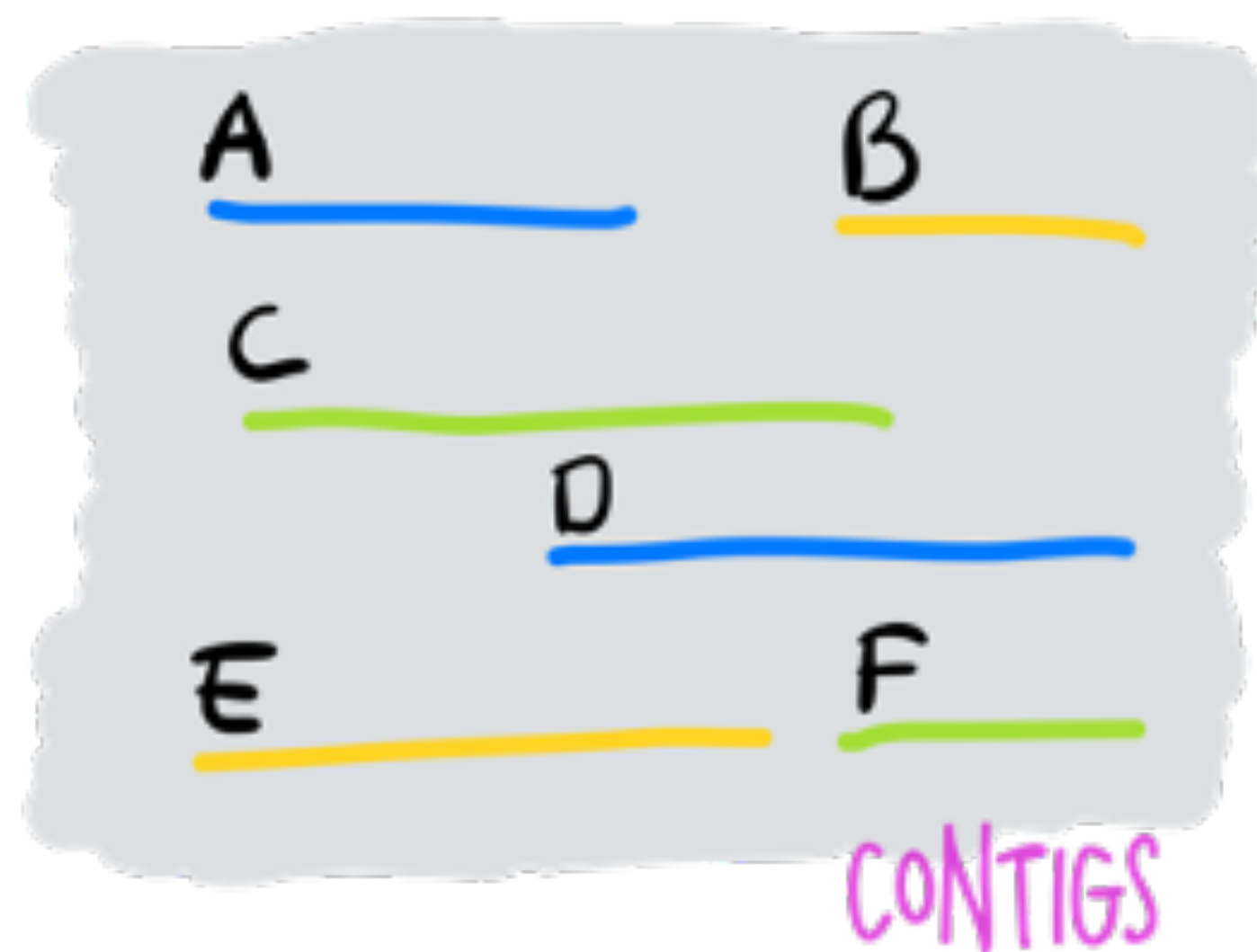
S3



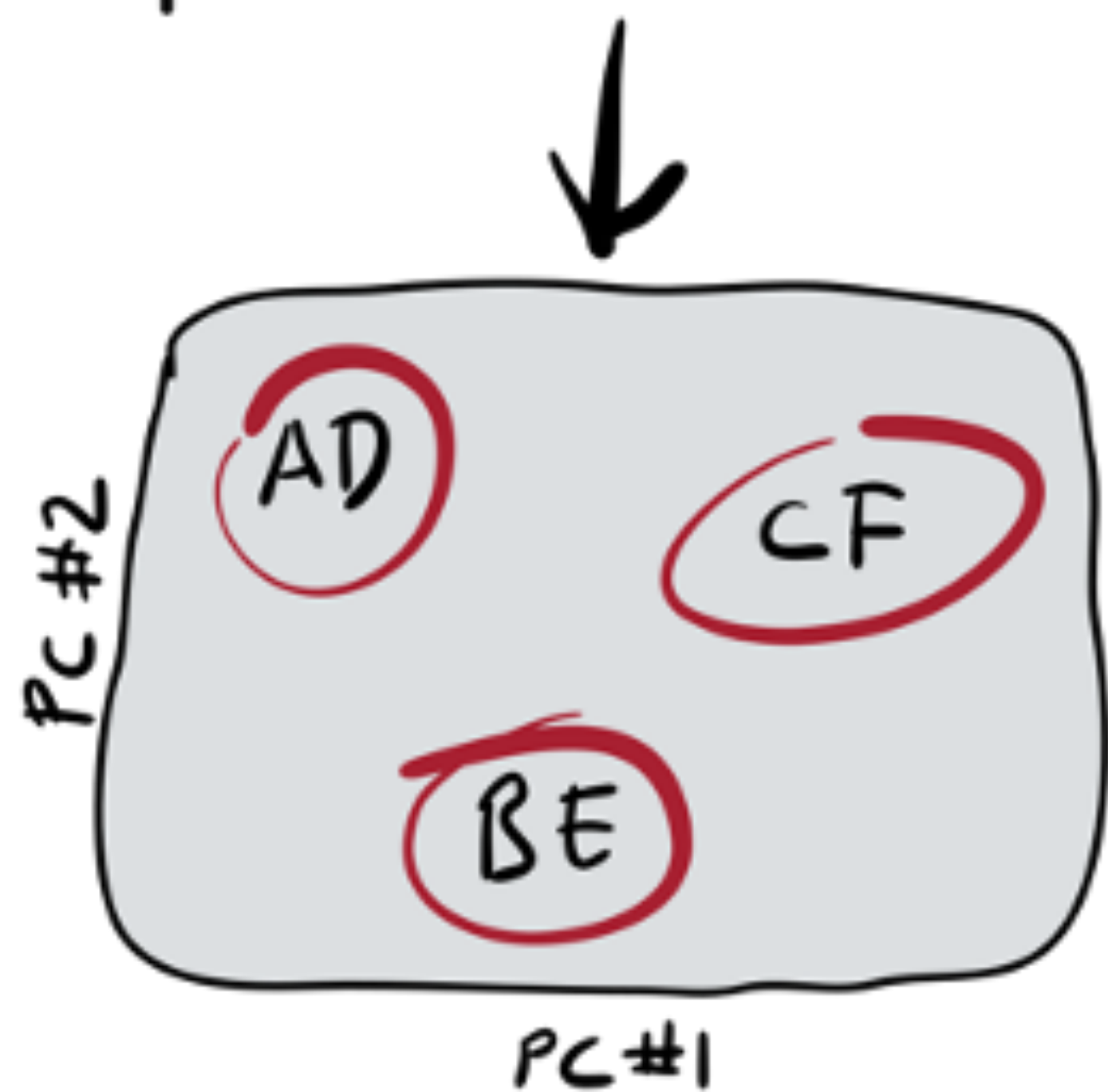
A - 3X
B - 5X
C - 1X
D - 3X
E - 5X
F - 1X



	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

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](#).



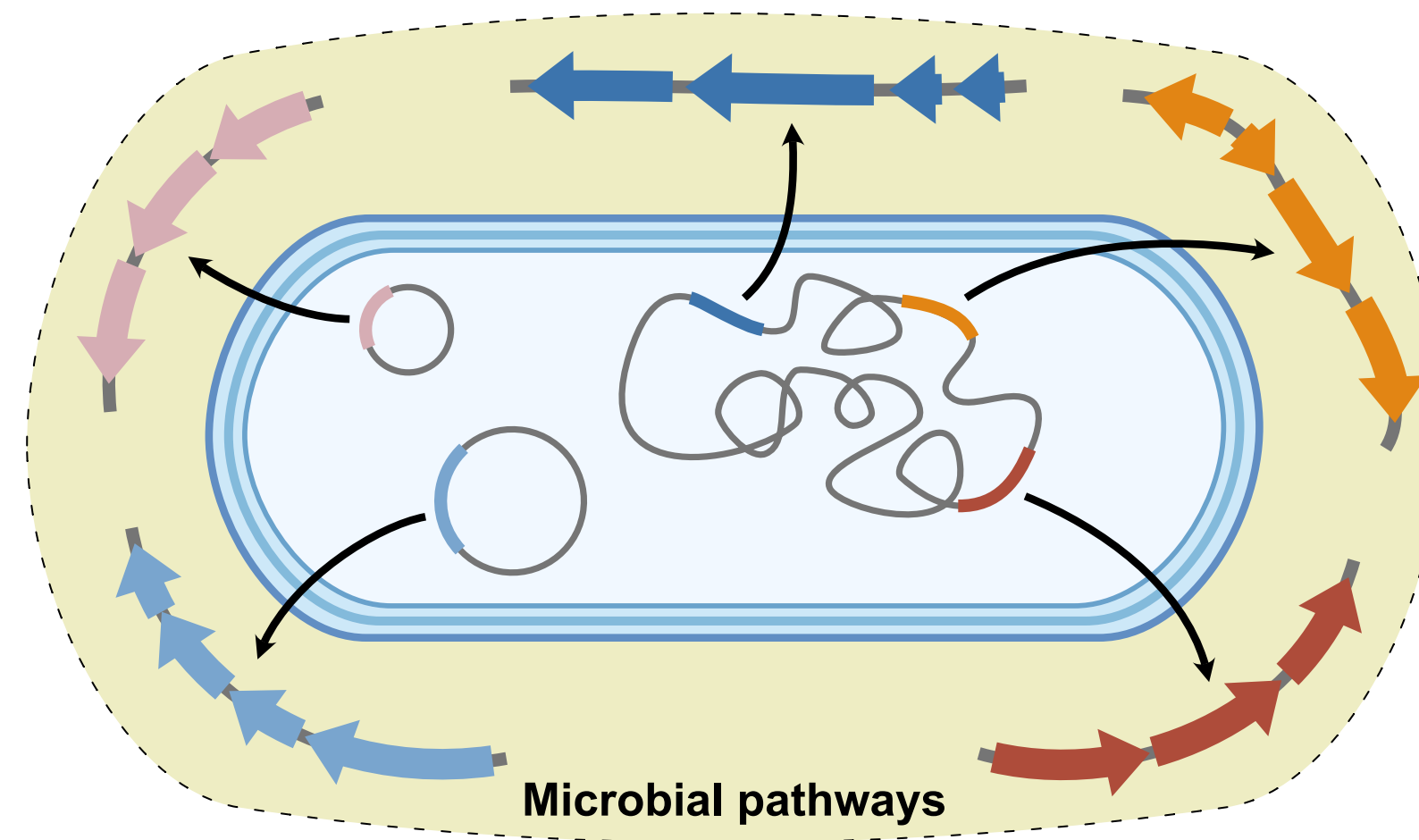
Mike Lee

□ Web □ Email □ Twitter □ LinkedIn □ Github

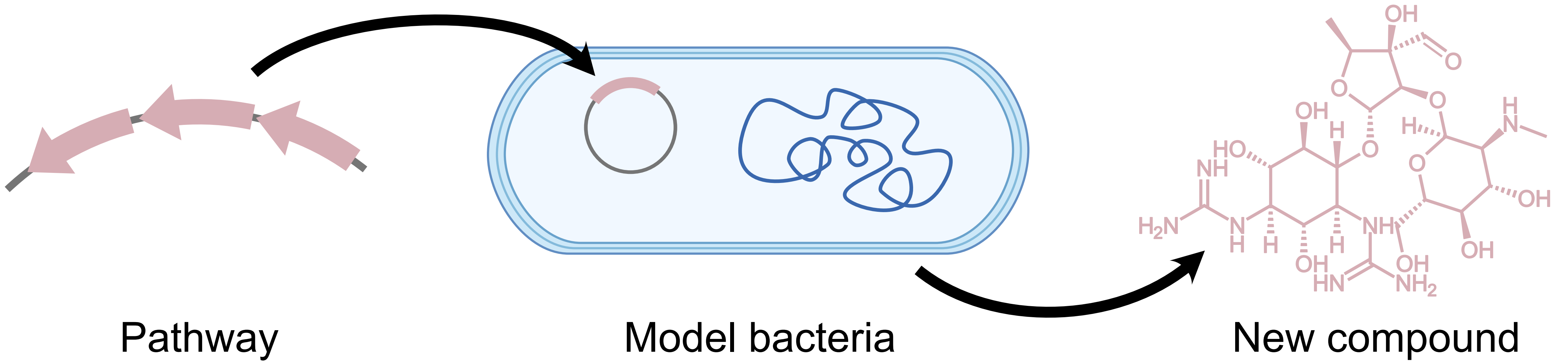
- » NASA Space Biology Fellow, [NASA Ames Research Center](#).
- » JCVI Research Fellow, [J. Craig Venter Institute](#).

- 👉 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)

Mining genomes for biosynthetic pathways

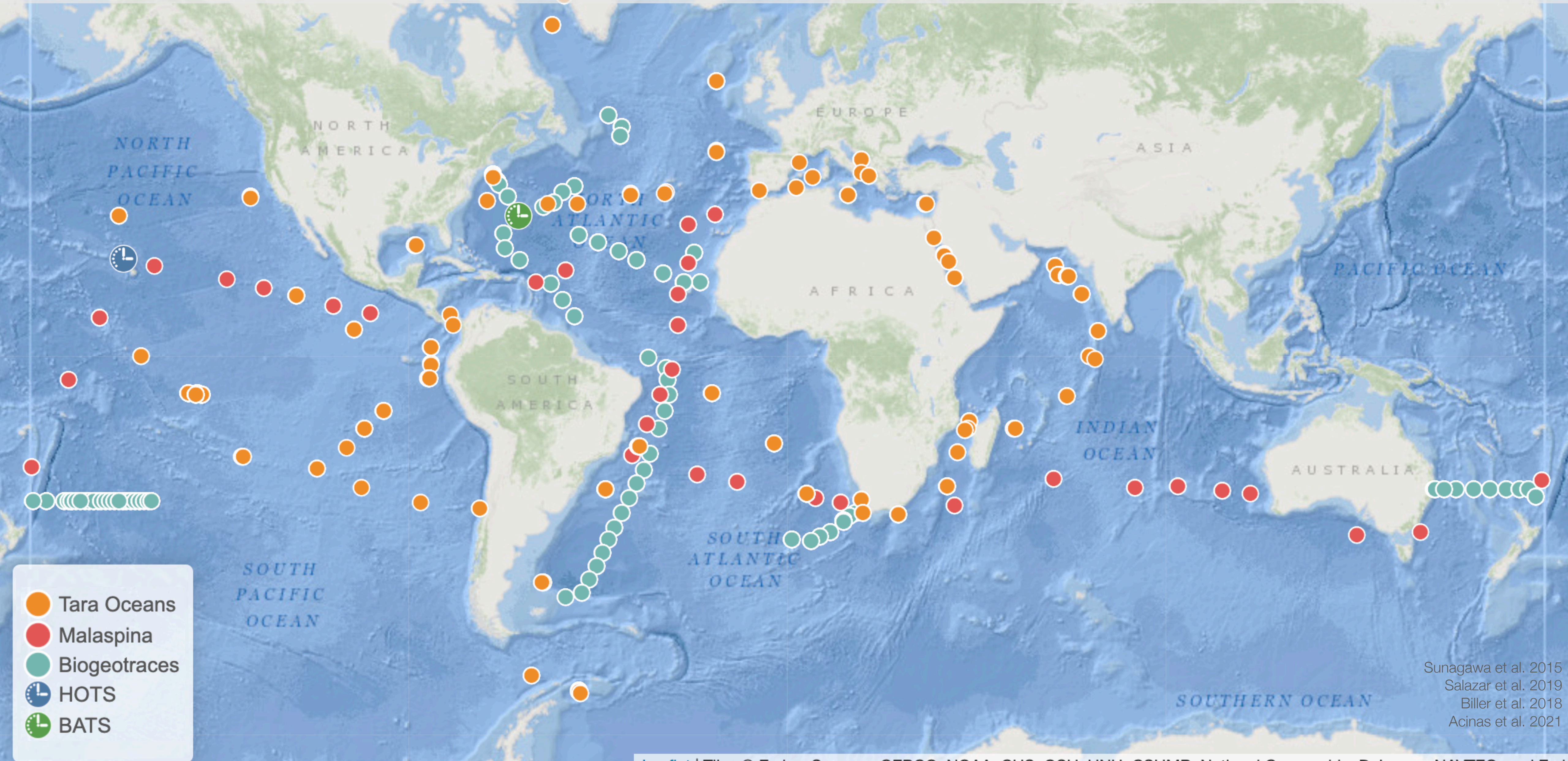


Synthetic biology: from sequences to compounds



Application to the global ocean microbiome

Global ocean metagenomes

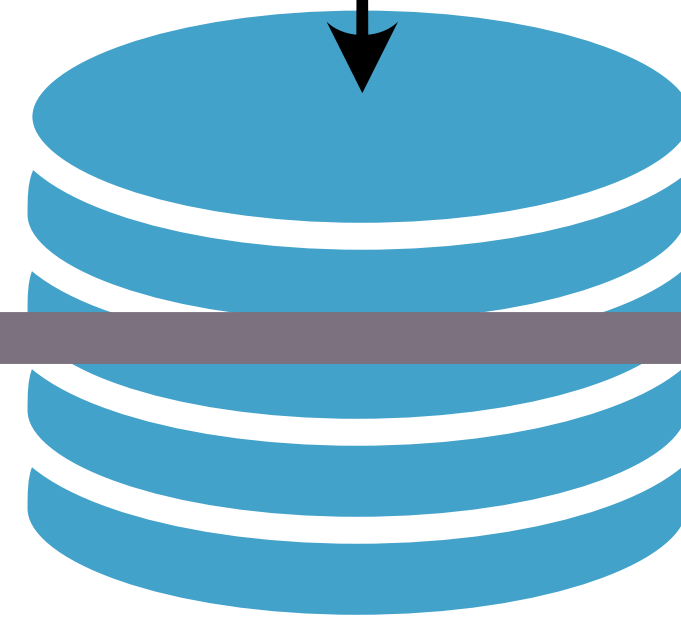


A genomic resource for the global ocean

>1,000 metagenomes



Bioinformatic
pipeline

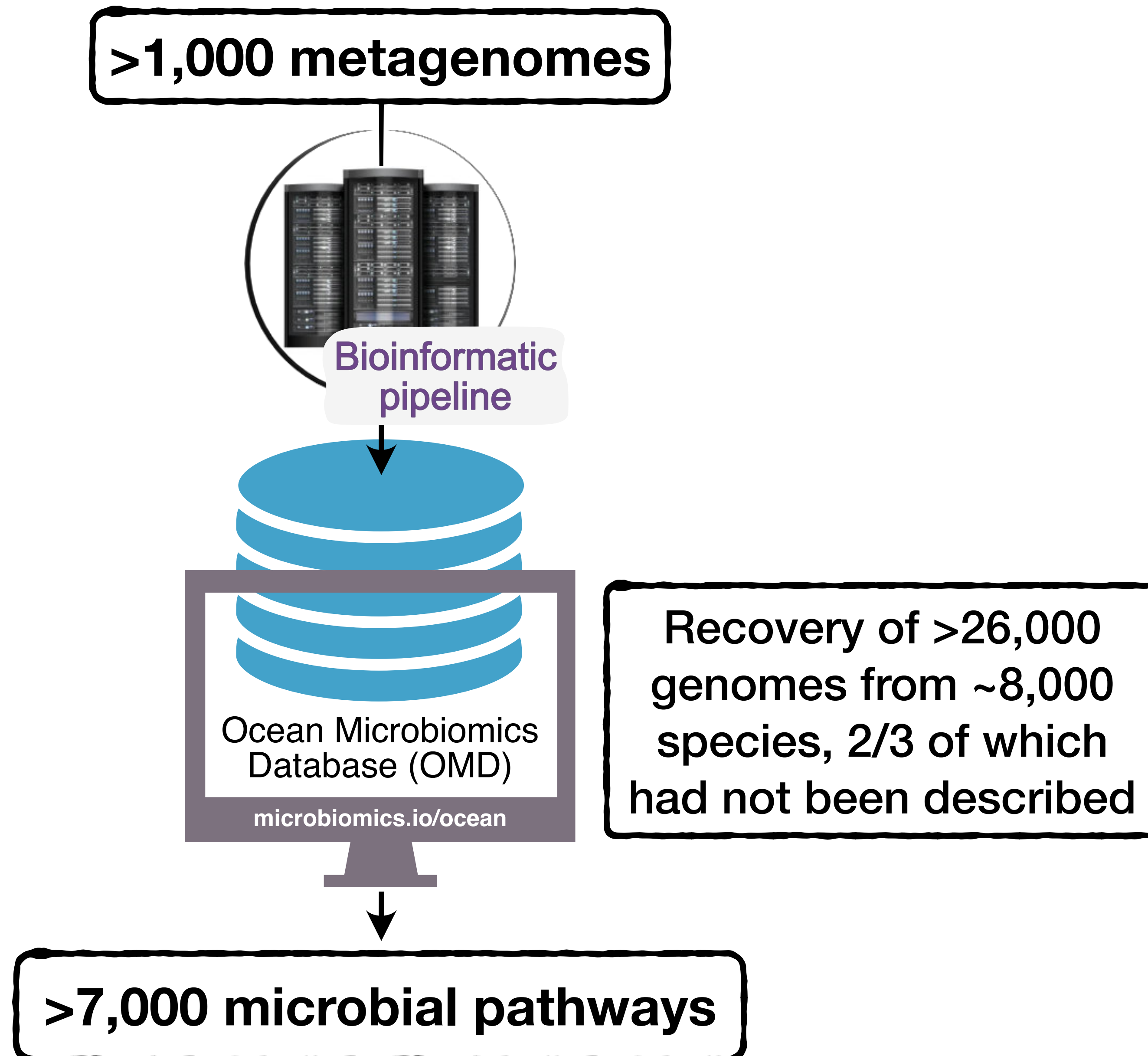


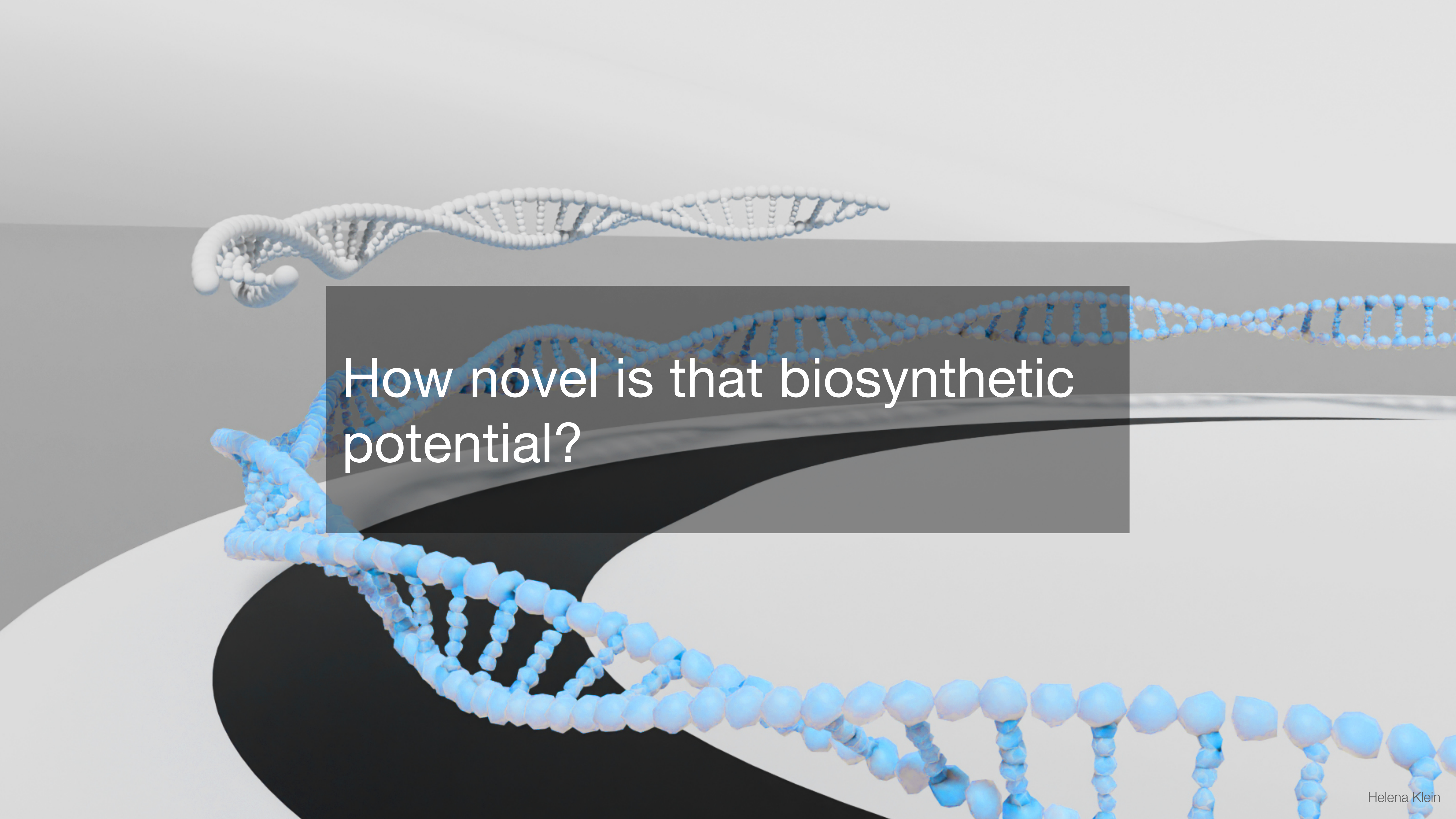
Ocean Microbiomics
Database (OMD)

microbiomics.io/ocean

Recovery of >26,000
genomes from ~8,000
species, 2/3 of which
had not been described

A genomic resource for the global ocean



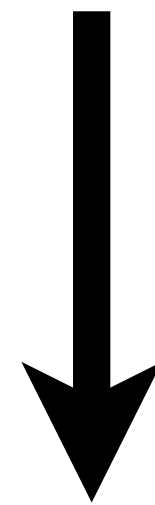


How novel is that biosynthetic potential?

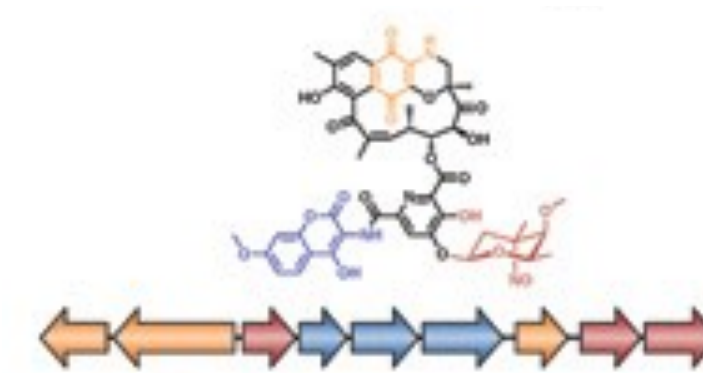
Comparison to sequenced or characterised BGCs



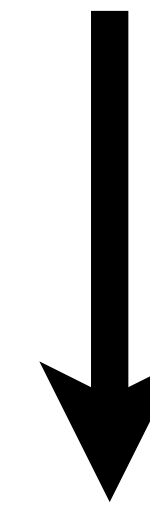
RefSeq
~200,000 genomes



1.2 M BGCs



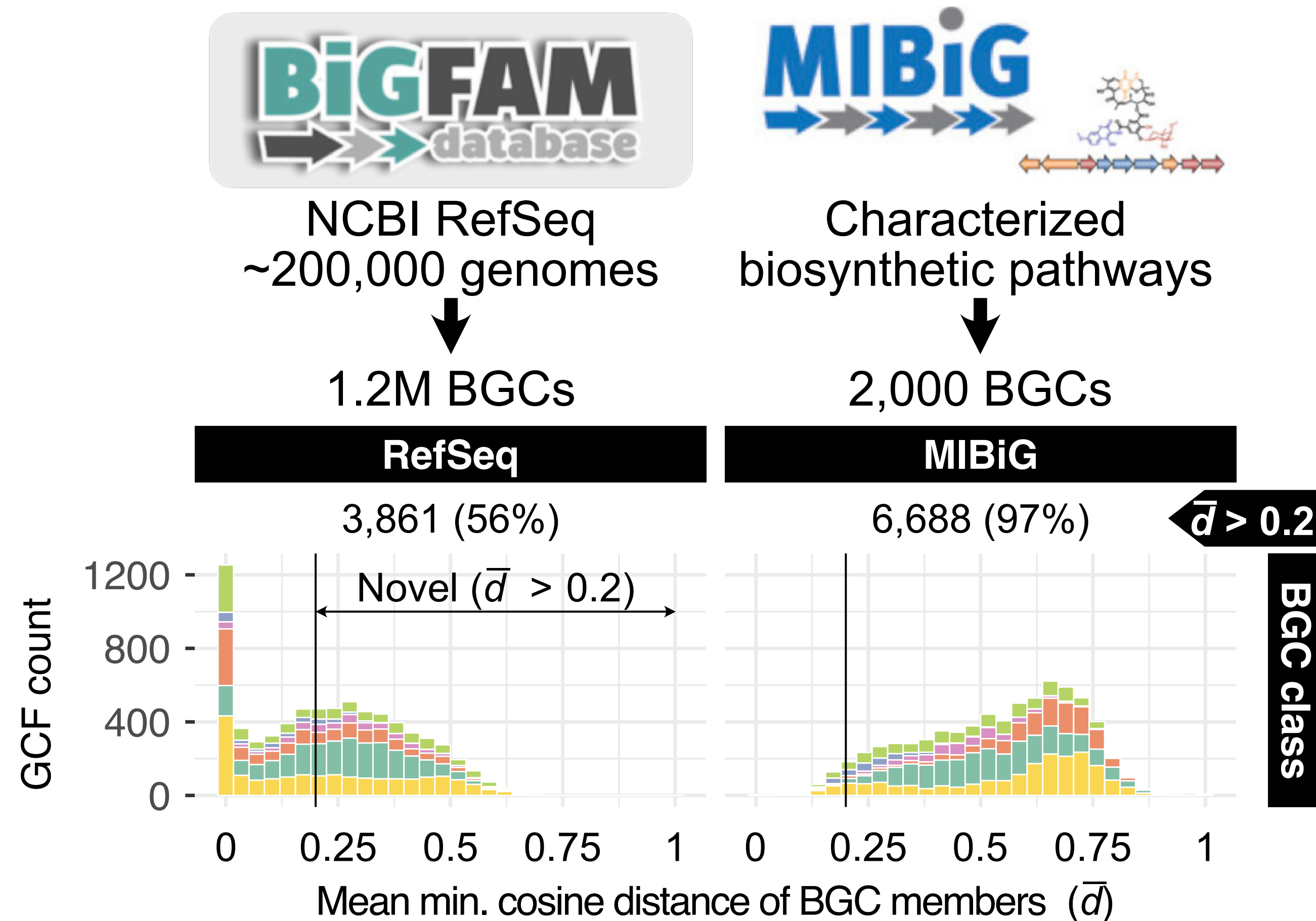
**Characterised
biosynthetic pathways**

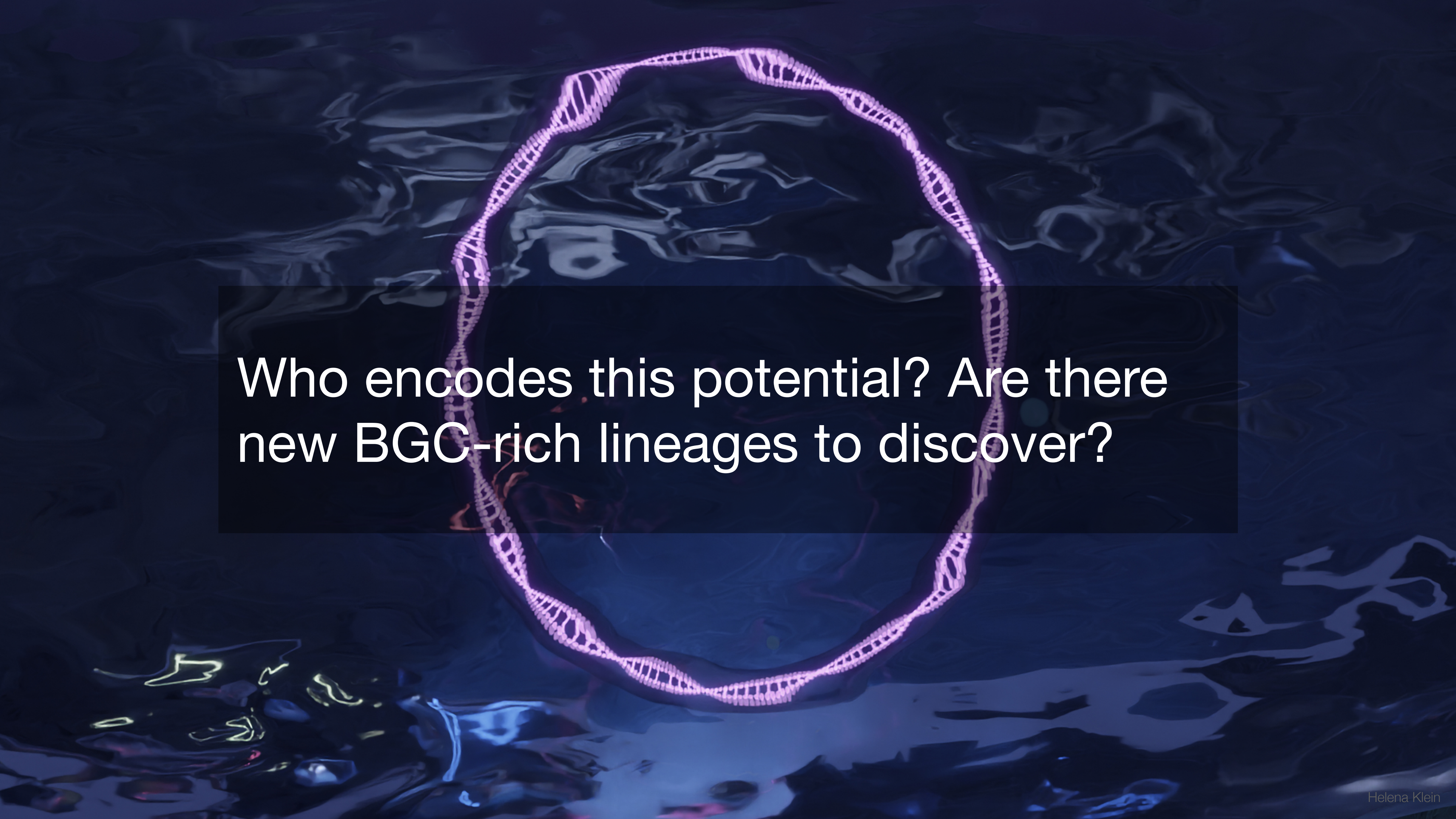


2,000 BGCs



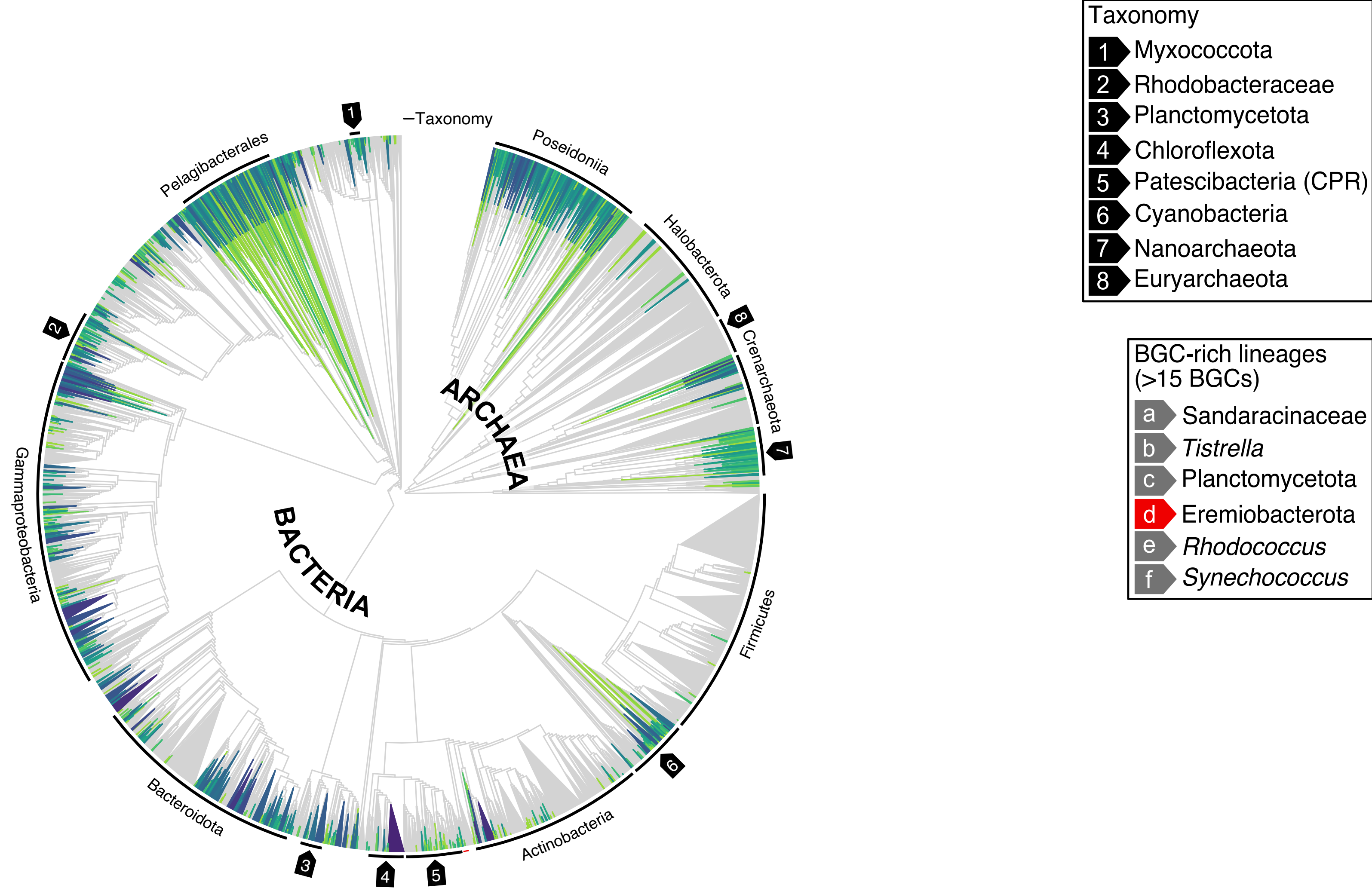
Reveals large potential for new compounds



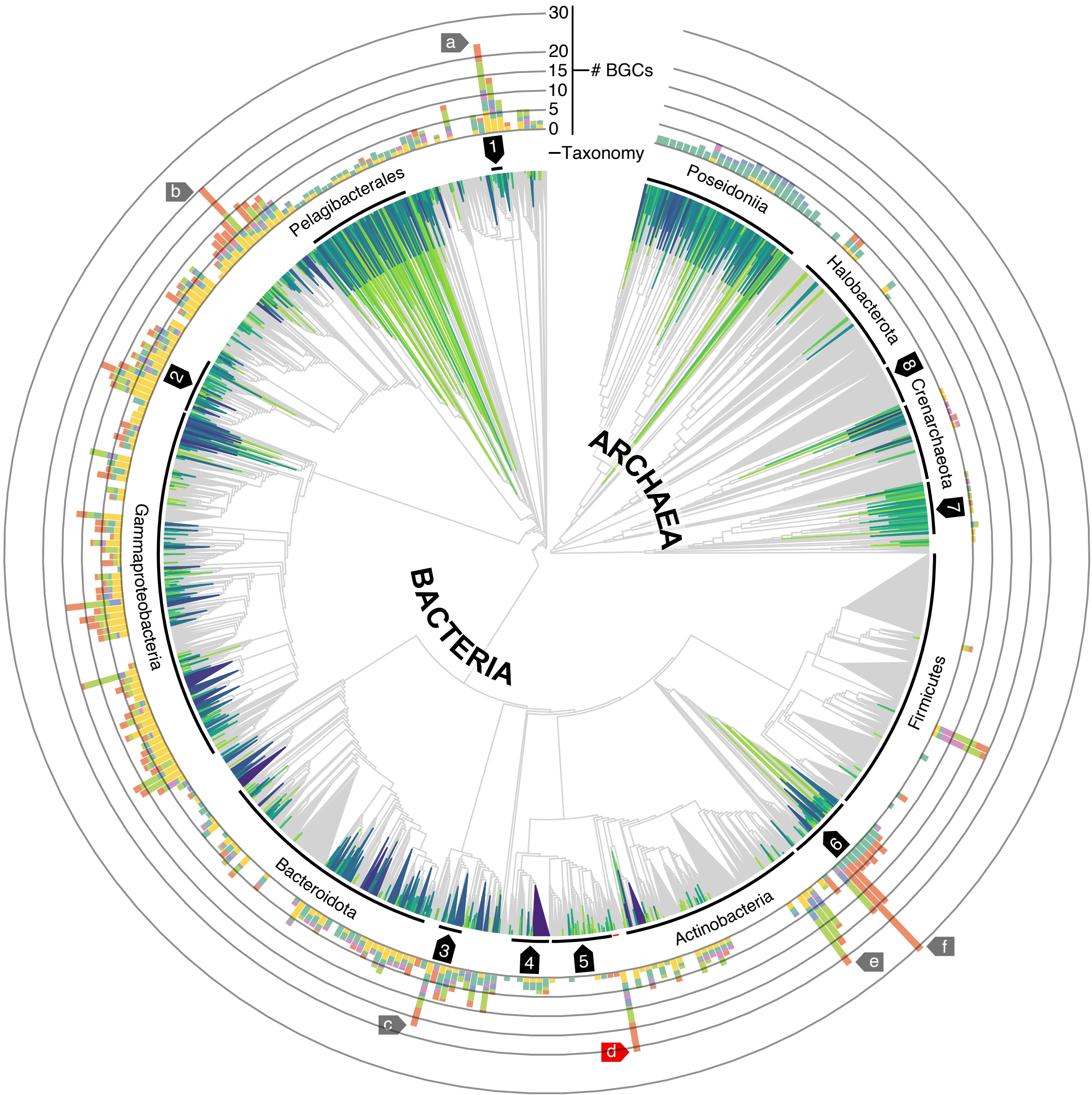


Who encodes this potential? Are there
new BGC-rich lineages to discover?

Distribution of the ocean's biosynthetic potential



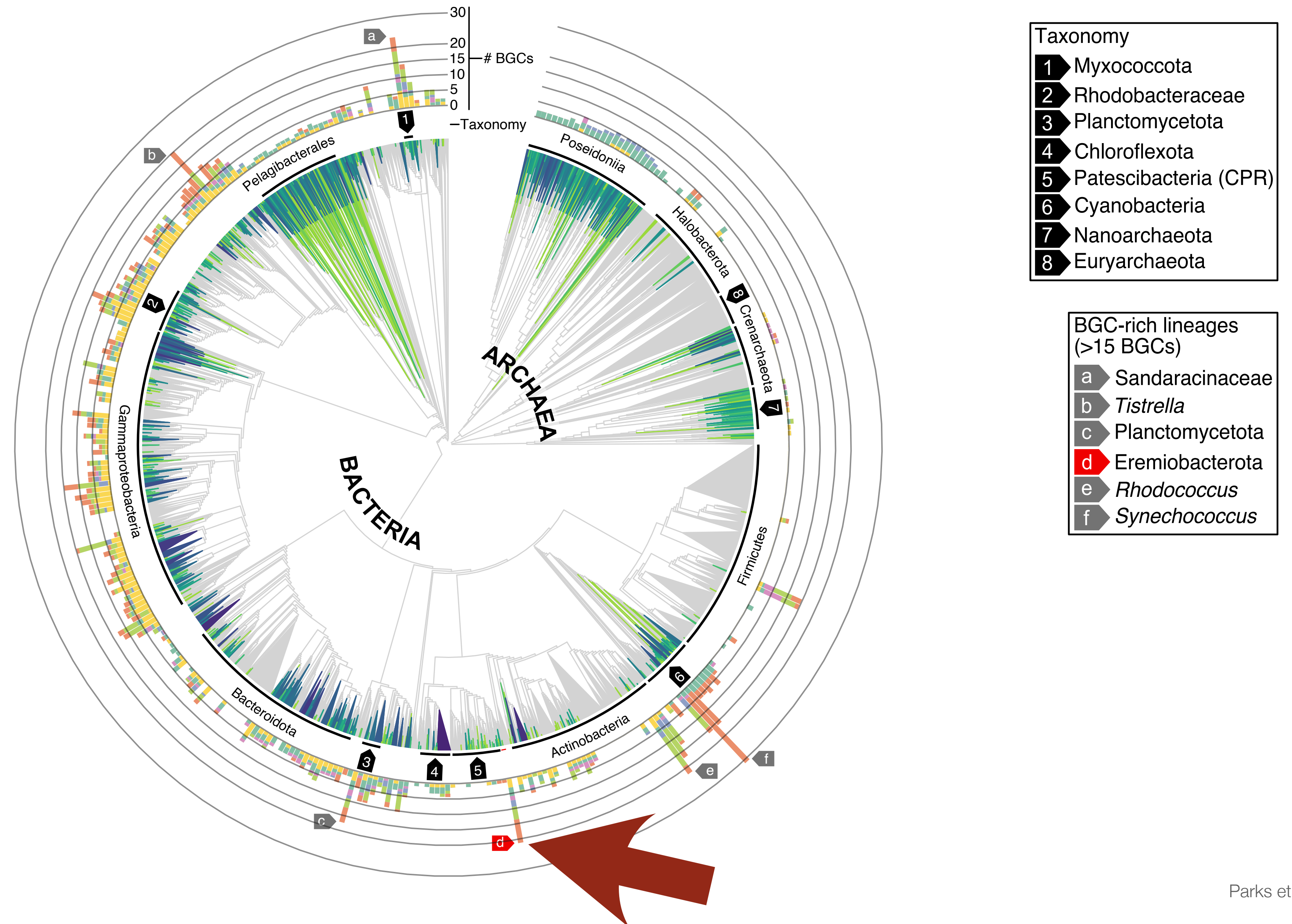
Distribution of the ocean's biosynthetic potential



- Taxonomy
- 1 Myxococcota
 - 2 Rhodobacteraceae
 - 3 Planctomycetota
 - 4 Chloroflexota
 - 5 Patescibacteria (CPR)
 - 6 Cyanobacteria
 - 7 Nanoarchaeota
 - 8 Euryarchaeota

- BGC-rich lineages (>15 BGCs)
- a Sandaracinaceae
 - b *Tistrella*
 - c Planctomycetota
 - d Eremiobacterota
 - e *Rhodococcus*
 - f *Synechococcus*

Eremiobacterota: unexpected BGC-richness



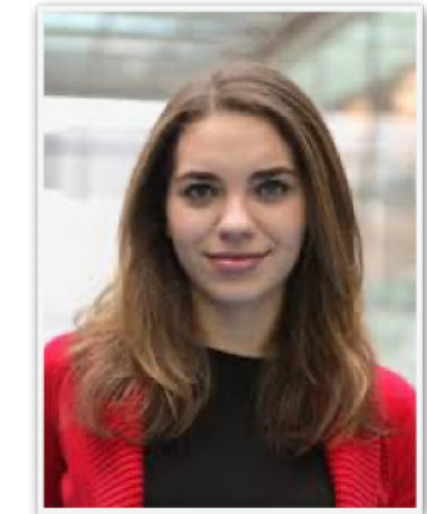
The background features several molecular models. On the left, there are structures with red and blue spheres. On the right, there are structures with blue and white spheres. In the center, there are structures with yellow and blue spheres. The text is overlaid on a dark blue rectangular area.

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

Experimental characterisation of this potential

Eudoremicrobium spp.

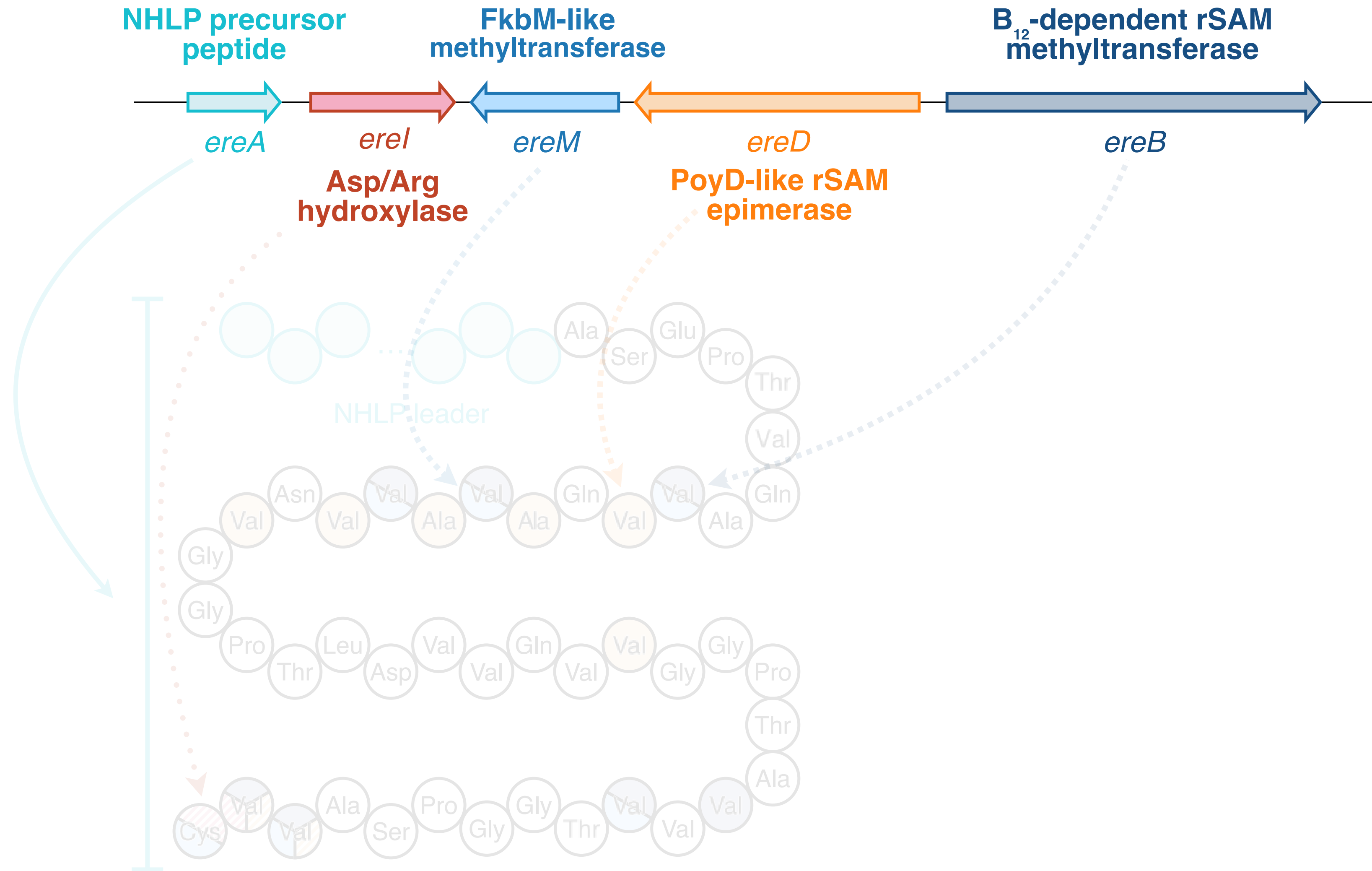
Ectoine -	■	■
T1PKS/PUFA/Olefin -	■	■
Siderophore -	■	■
T3PKS (alkylpyrone) -	■	■
Hopanoid #1 -	■	■
Carotenoid -	■	■
Bacteriocin (DUF692) -	■	■
Proteusin/Bacteriocin #1 -	■	■
T1PKS -	■	■
Phosphonate -	■	■
NRPS-like #1 -	■	■
Proteusin/Lanthipeptide -	■	■
T1PKS/NRPS-like -	■	■
Proteusin/Bacteriocin #2 -	■	■
NRPS-like #3 -	■	■
Glycosylated microviridin -	■	■
NRPS-like #4 -	■	■
Hopanoid #2 -	■	■
Microviridin -	■	■
NRPS-like #2 -	■	■
NRPS-like/betalactone -	■	■
NRPS-like #5 -	■	■
Aryl polyene -	■	■
Phosphorylated peptide -	■	■



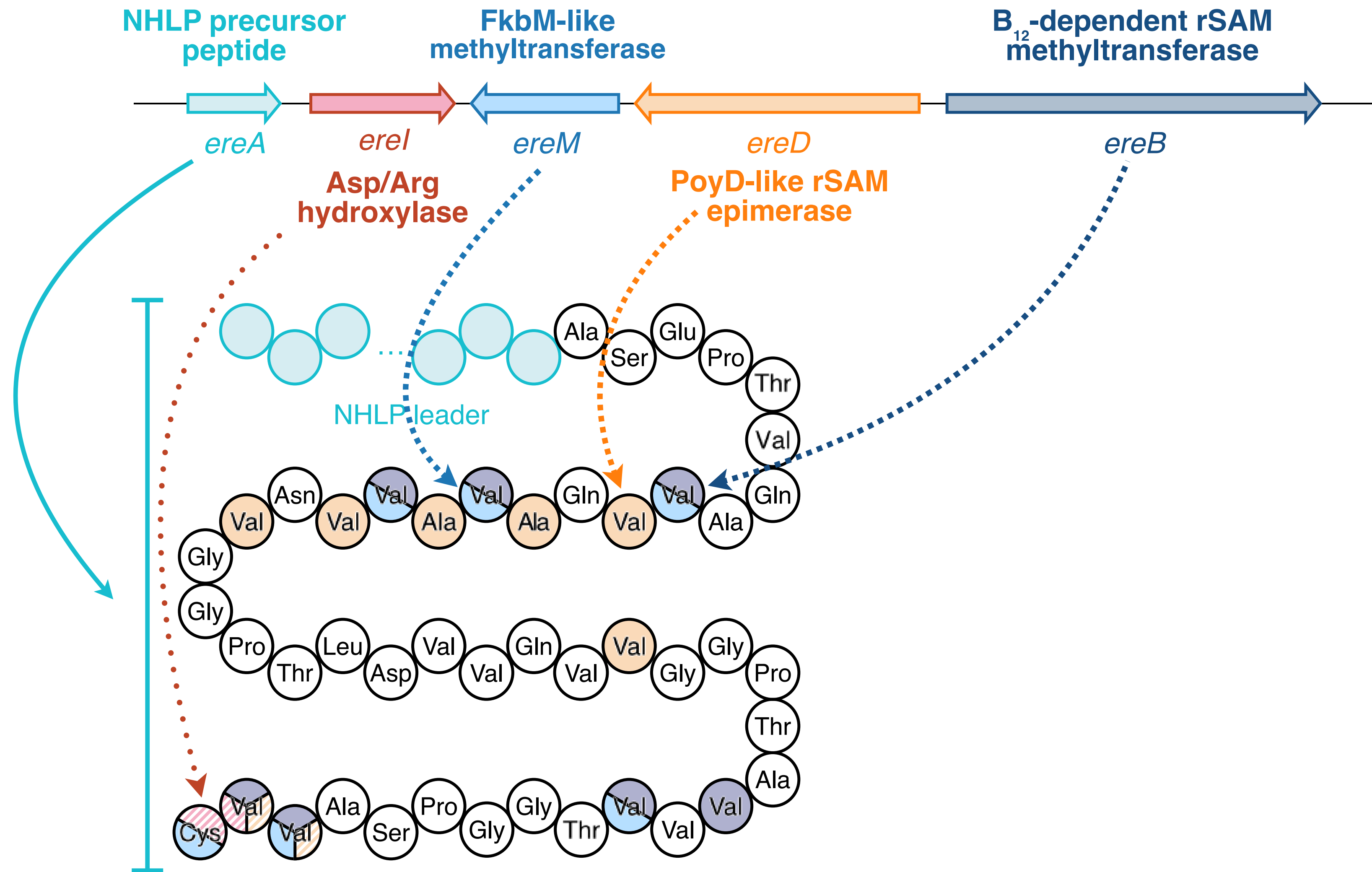
Characterising **two predicted** novel Ribosomal BGCs using:

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

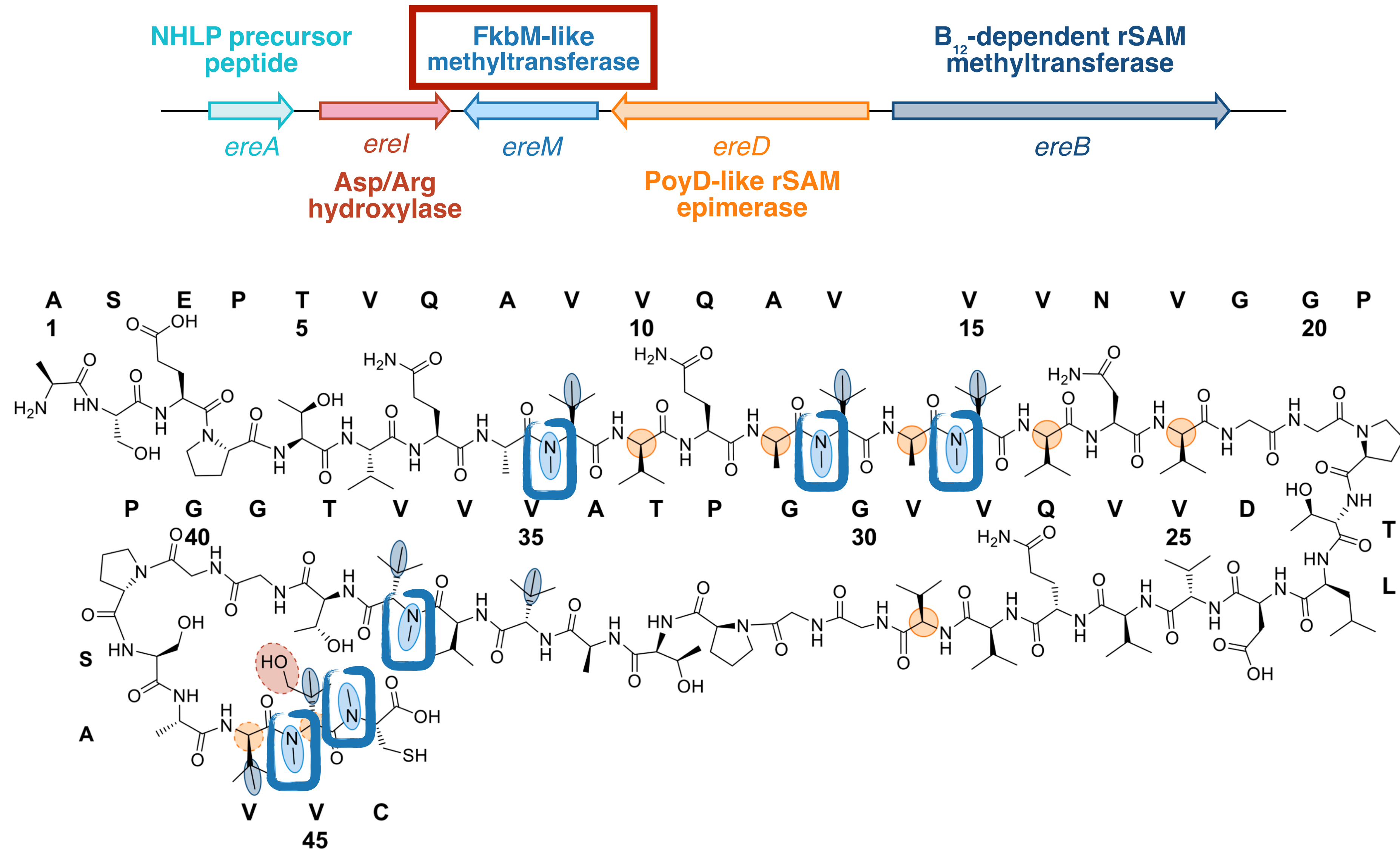
The logic of ribosomal natural products synthesis



Intricate BGC reveals new enzymology



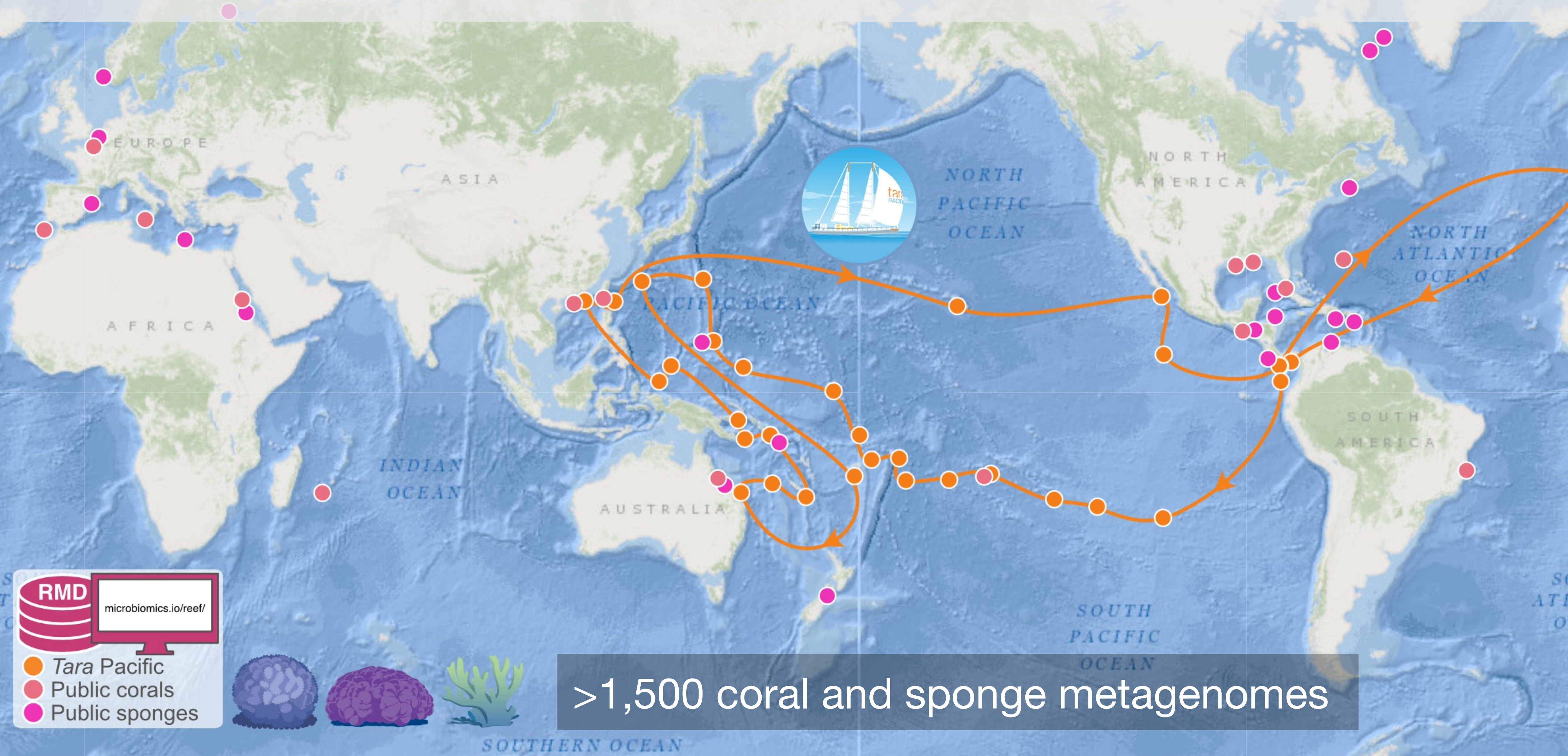
With unique trans amide-N-methylation



Application to the coral microbiome: an endangered treasure trove

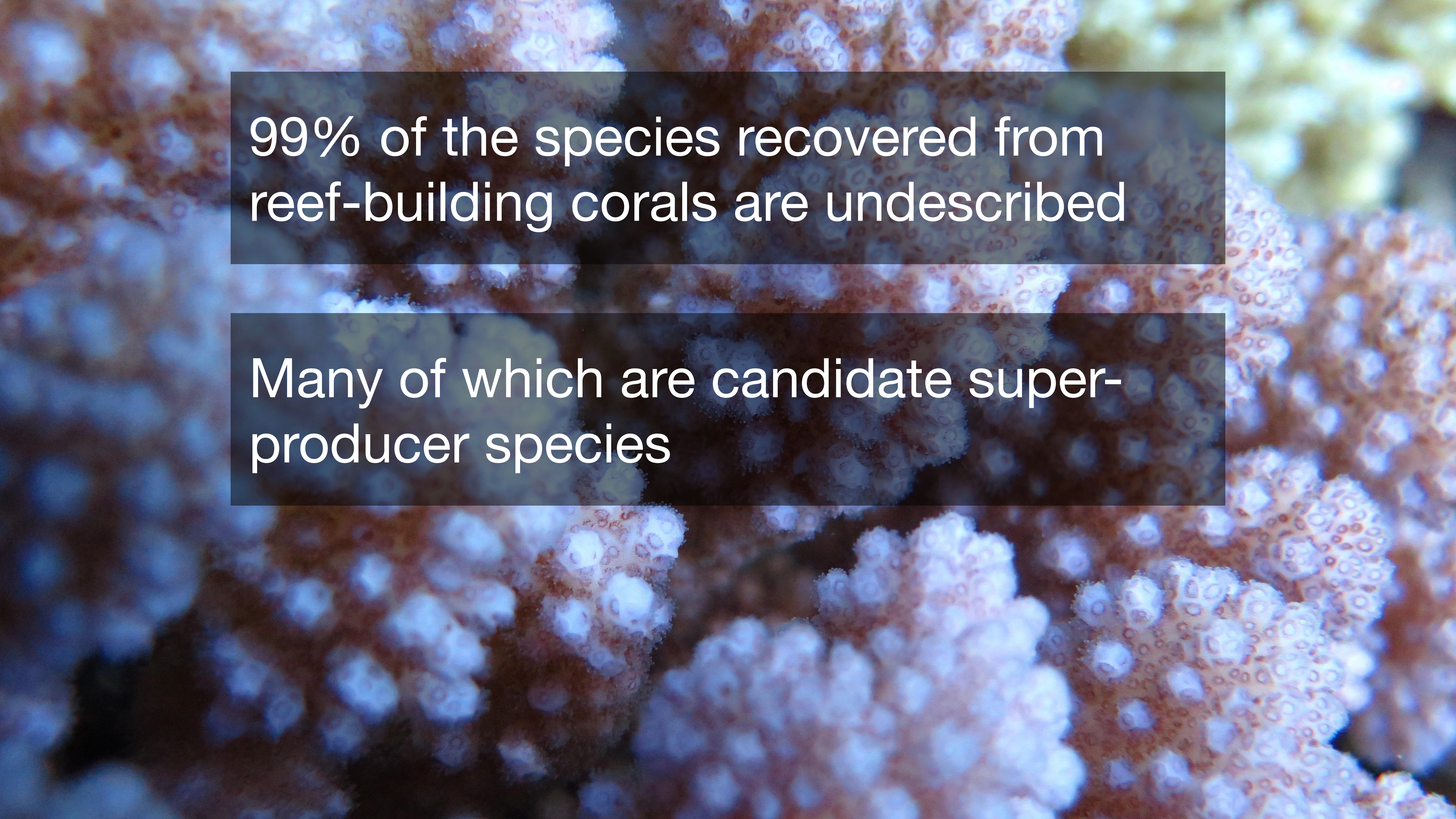


The Reef Microbiome Database



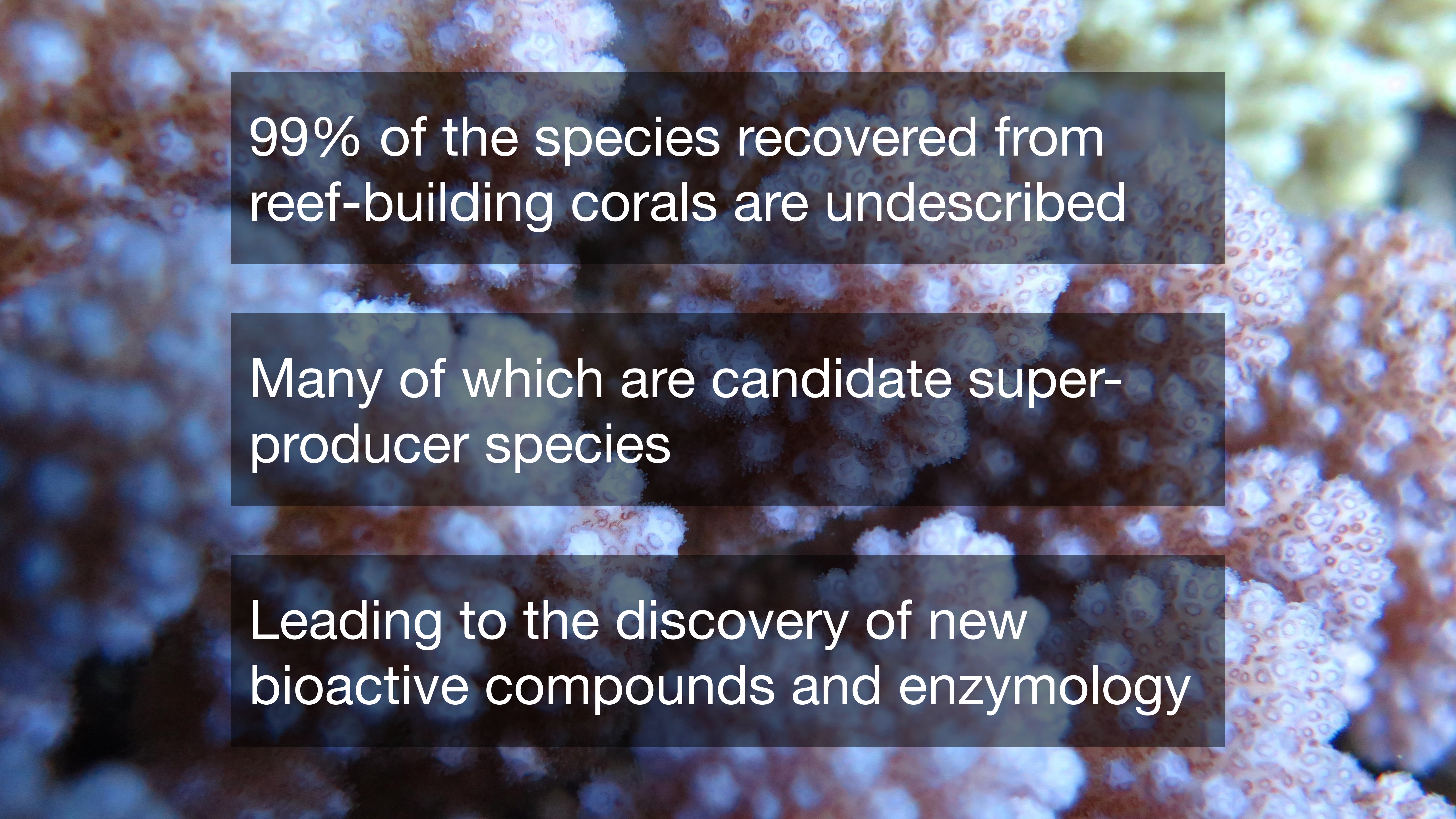
A close-up photograph of a coral reef. The coral has a complex, porous structure with many small, circular openings. The colors are primarily shades of brown, tan, and white, with some darker, almost black, areas in the shadows. The lighting is bright, highlighting the texture of the coral. A semi-transparent dark rectangle is overlaid on the upper left portion of the image, containing white text.

99% of the species recovered from
reef-building corals are undescribed

A close-up photograph of coral polyps, showing their characteristic circular openings and textured surface in shades of brown, tan, and white. The image is slightly out of focus, emphasizing the intricate patterns of the coral.

99% of the species recovered from
reef-building corals are undescribed

Many of which are candidate super-
producer species

A close-up photograph of a coral reef, showing the intricate, porous structure of the coral polyps. The colors are primarily light blue and white, with some darker, brownish areas. The background is slightly blurred, emphasizing the texture of the coral in the foreground.

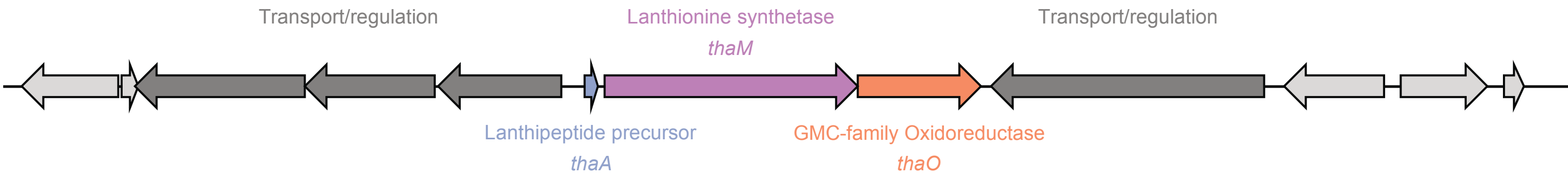
99% of the species recovered from
reef-building corals are undescribed

Many of which are candidate super-
producer species

Leading to the discovery of new
bioactive compounds and enzymology

New biosynthetic potential

- Acidobacteriota (phylum)
 - Subgroup 21 (class)
 - MAG13 (species - abundance-based match)
 - Vicinamibacteraceae (family)
 - Thermoanaerobaculaceae/Subgroup 10 (genus)
 - MAG48 (species - 98.9% id.)
 - Acanthopleuribacteraceae (family)
 - MAG20 (species - 100% id.)
 - Sulfidibacter corallicola (species - 100% id.)
 - Acanthopleuribacter pedis (species - 99.5% id.)



Reveals new enzymology

Acidobacteriota (phylum)

Subgroup 21 (class)

- └─ MAG13 (species - abundance-based match)

Vicinamibacteraceae (family)

Thermoanaerobaculaceae/Subgroup 10 (genus)

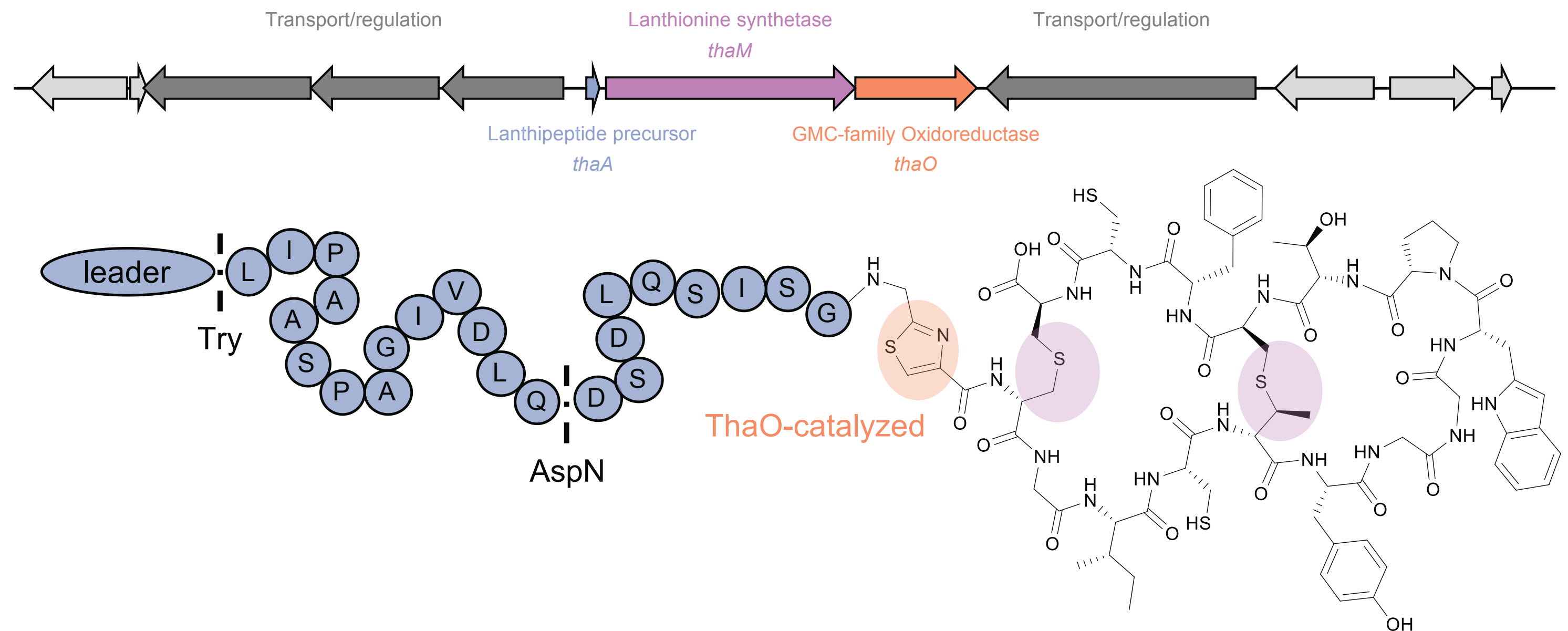
- └─ **MAG48 (species - 98.9% id.)**

Acanthopleuribacteraceae (family)

- └─ MAG20 (species - 100% id.)

- └─ Sulfidibacter coralicola (species - 100% id.)

- └─ Acanthopleuribacter pedis (species - 99.5% id.)



Reveals new enzymology

Acidobacteriota (phylum)

Subgroup 21 (class)

- MAG13 (species - abundance-based match)

- Vicinamibacteraceae (family)

- Thermoanaerobaculaceae/Subgroup 10 (genus)

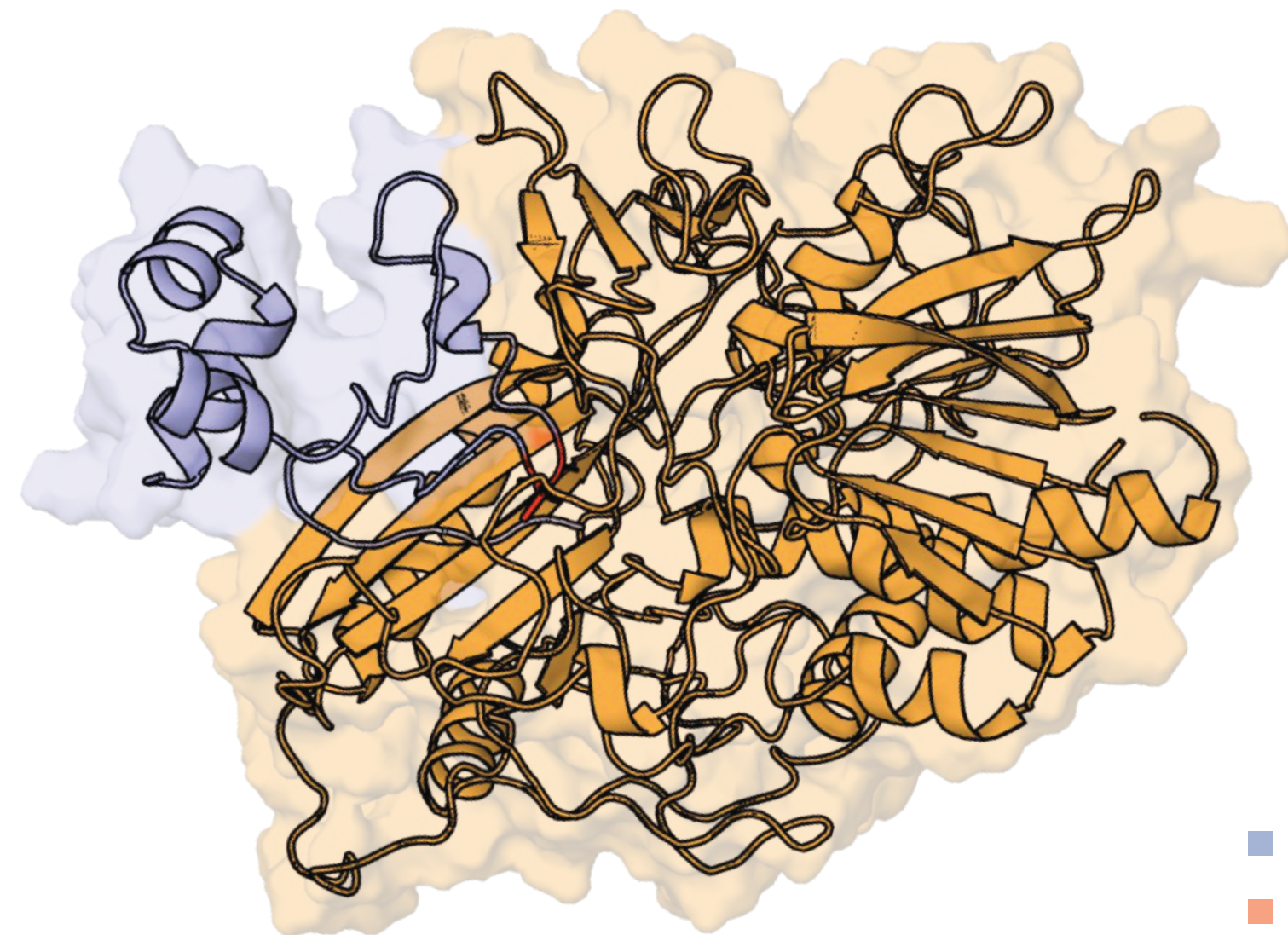
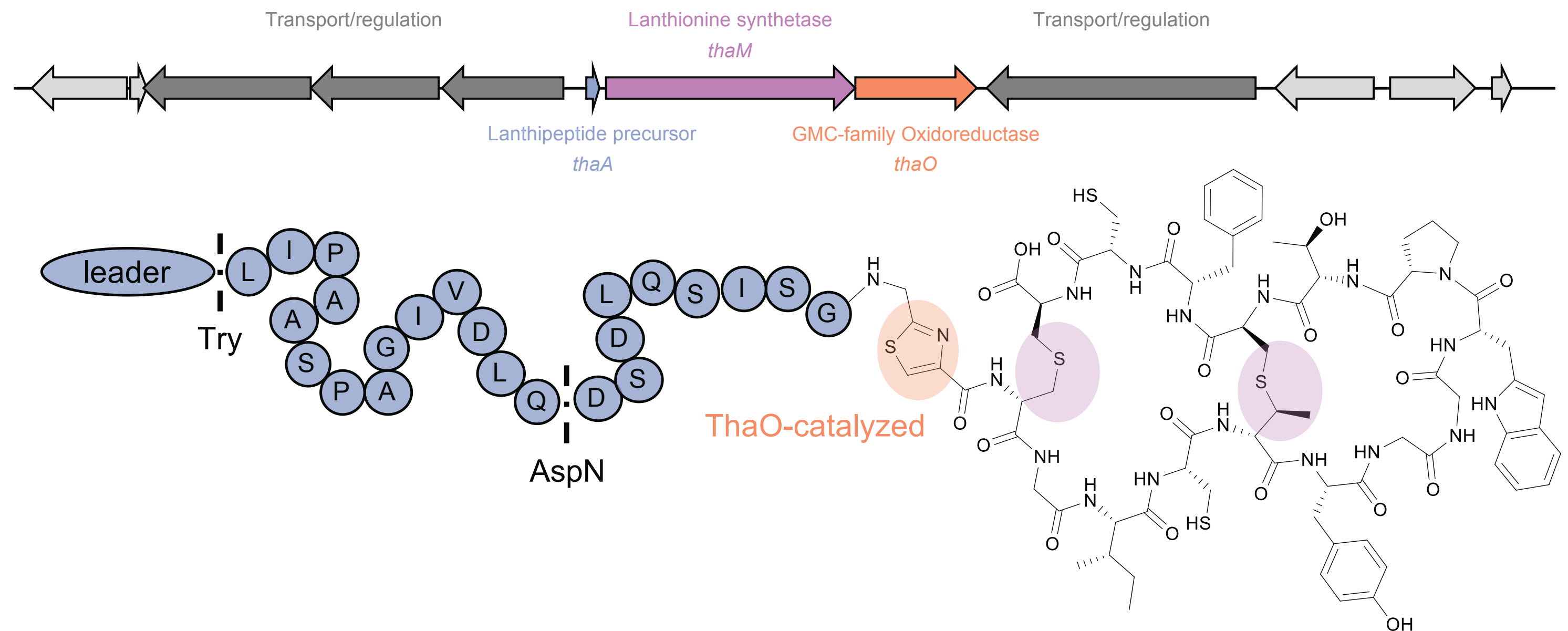
- **MAG48 (species - 98.9% id.)**

- Acanthopleuribacteraceae (family)

- MAG20 (species - 100% id.)

- Sulfidibacter corallicola (species - 100% id.)

- Acanthopleuribacter pedis (species - 99.5% id.)



ATP-independent
thiazole formation

■ ThaA
■ ThaO

As well as bioactivity

Acidobacteriota (phylum)

Subgroup 21 (class)

└─MAG13 (species - abundance-based match)

└─Vicinamibacteraceae (family)

└─Thermoanaerobaculaceae/Subgroup 10 (genus)

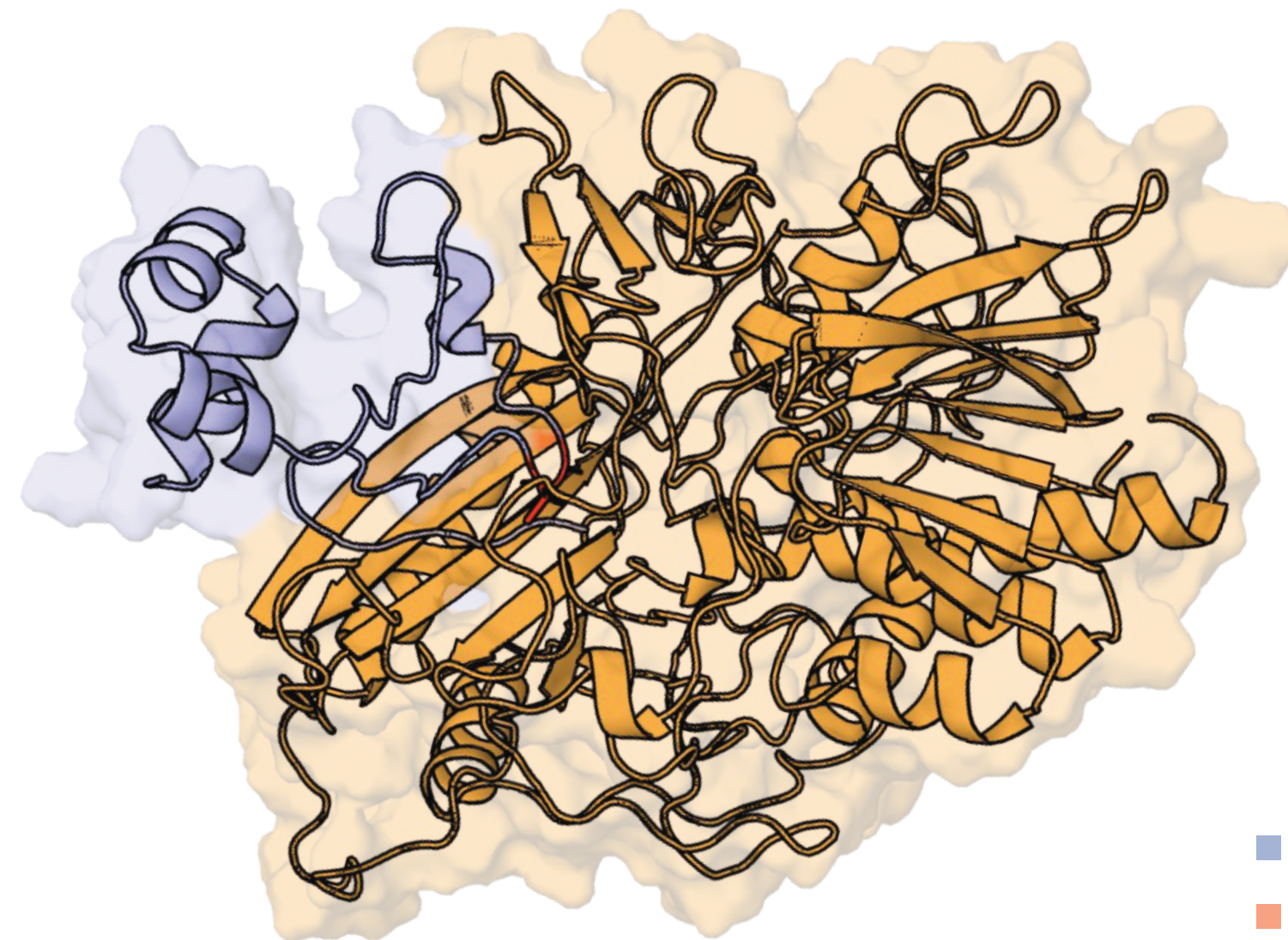
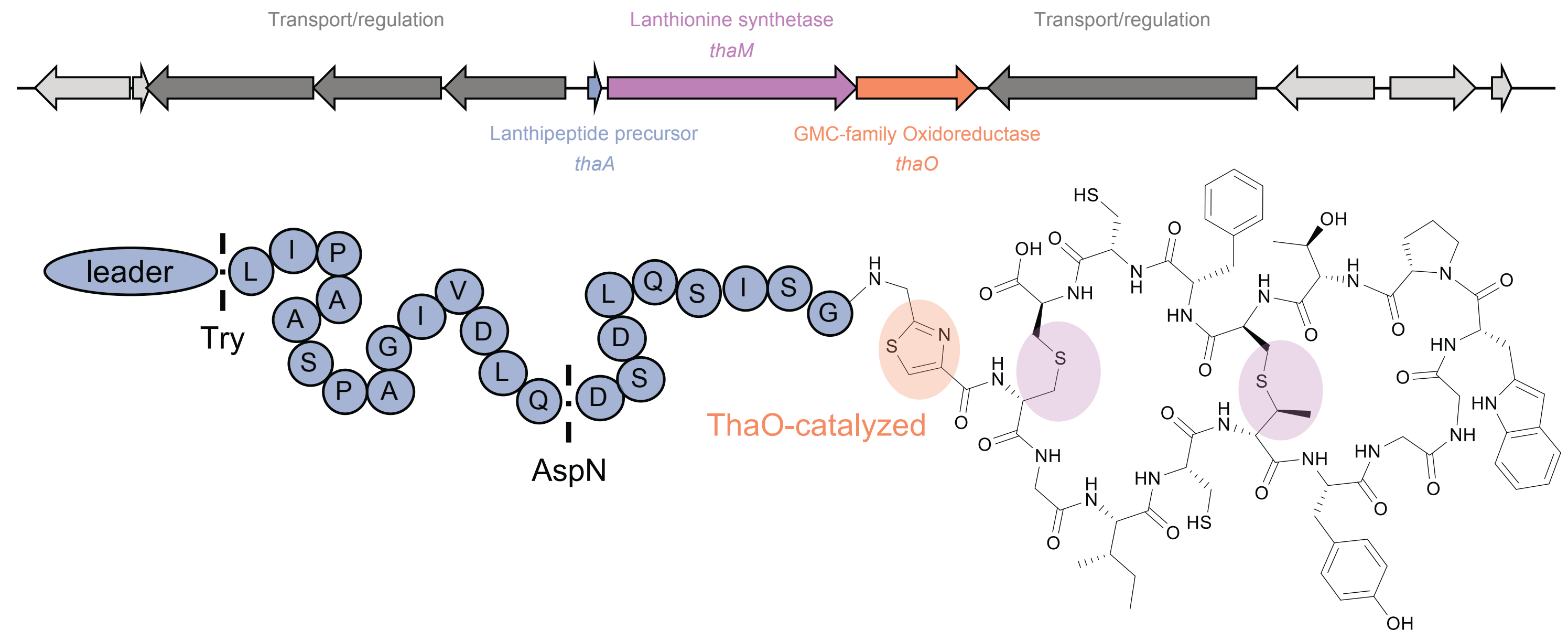
└─**MAG48 (species - 98.9% id.)**

└─Acanthopleuribacteraceae (family)

└─MAG20 (species - 100% id.)

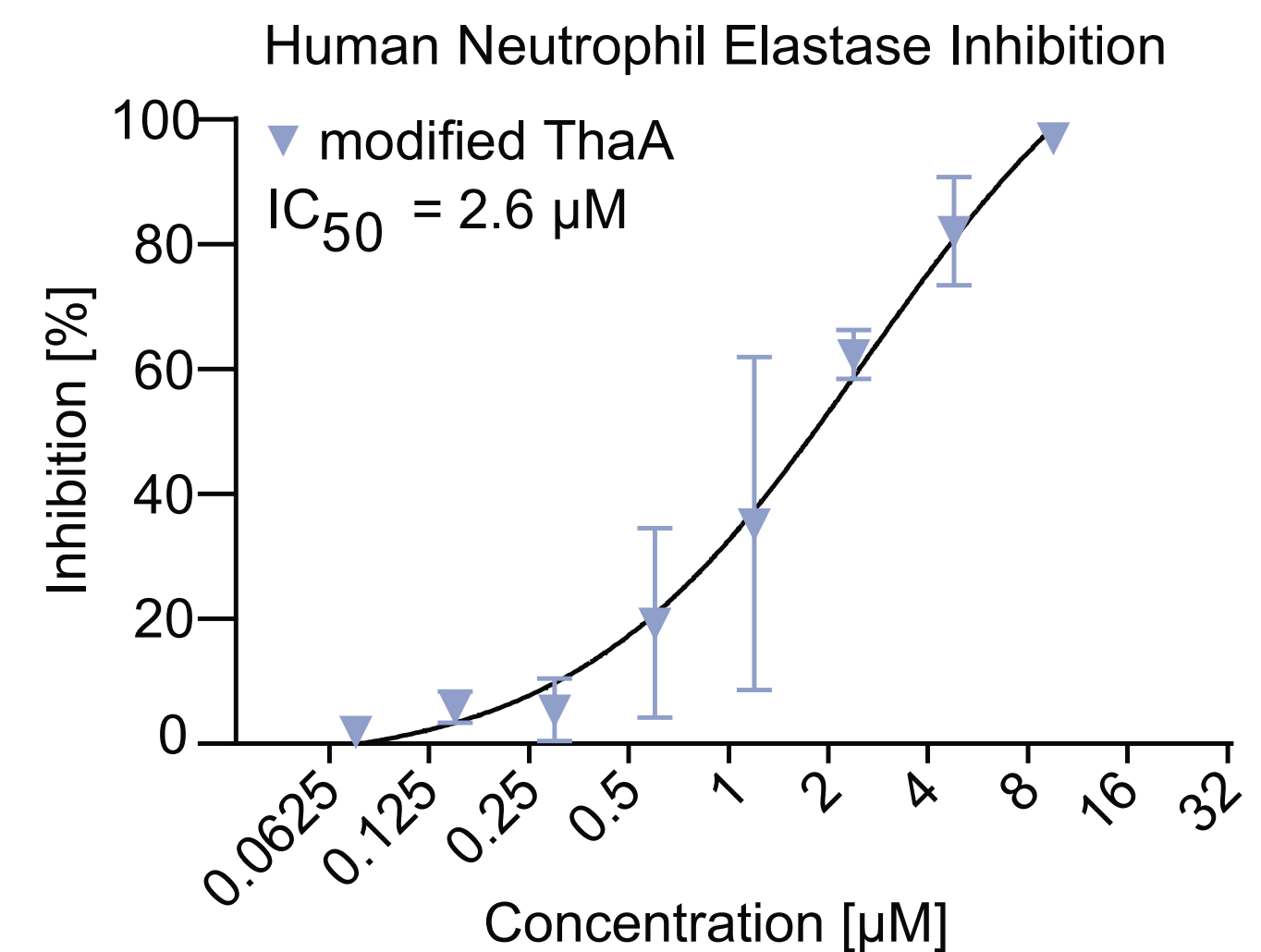
└─Sulfidibacter corallicola (species - 100% id.)

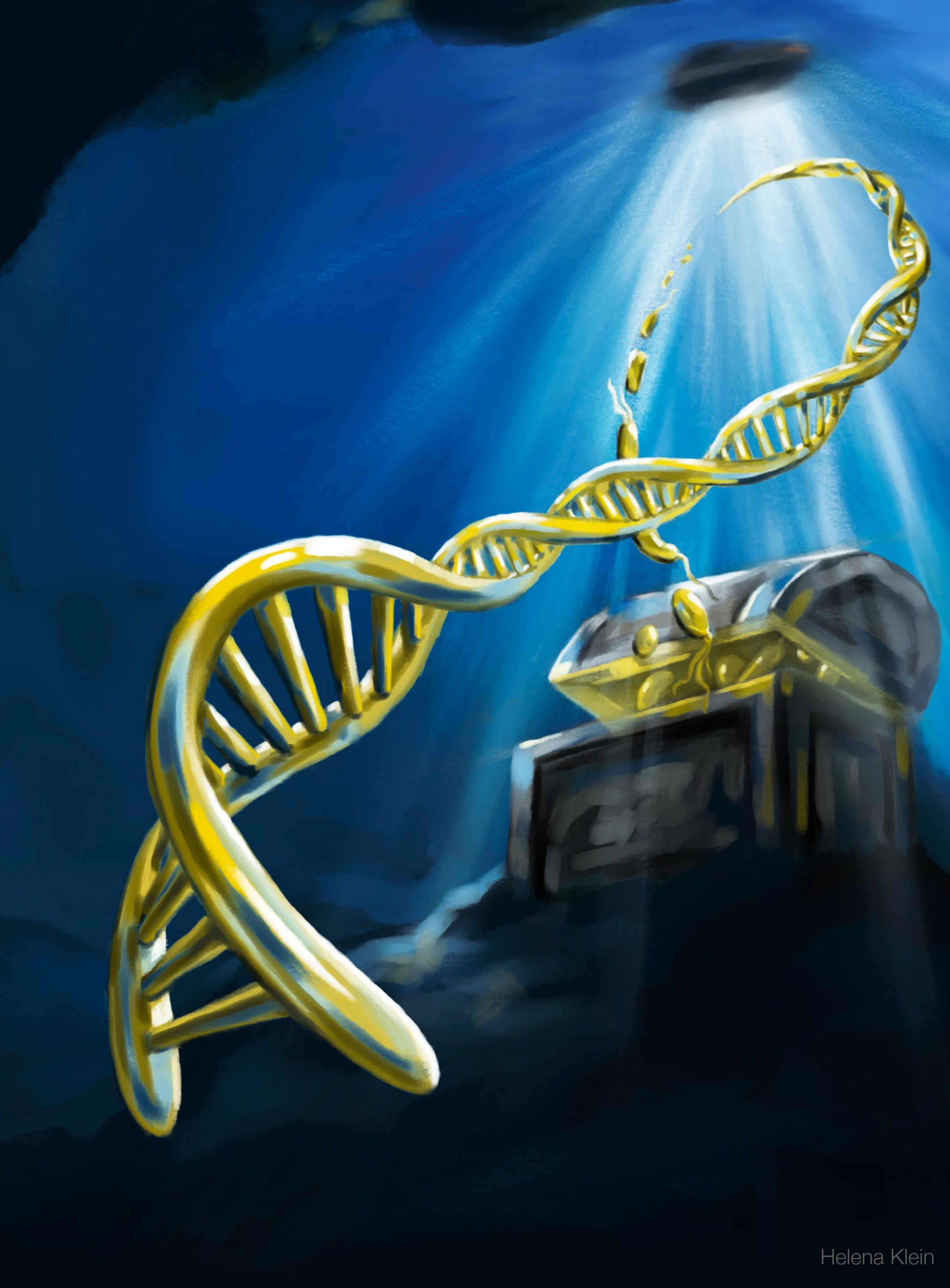
└─Acanthopleuribacter pedis (species - 99.5% id.)



ATP-independent
thiazole formation

■ ThaA
■ ThaO





Unlocked the biosynthetic potential of the ocean and coral microbiome

Unveiled uncharacterised BGC-rich lineages for the discovery of new pathways and compounds

Manifold consequences of current and projected reef biodiversity loss

