From raw reads to contigs

Biol-217
22 January - 2 February 2024

Dr. Cynthia M. Chibani

Outline: Intro to Metagenomics and Assembly

- Why Sequencing and Metagenomics?
- Different approaches
- Quality control
- Assembly




Studying microbes in different environments

- Why microbiology?
    - Understand the microbial world and processes in the environment
    - Discover new tools for biotechnology:
        - antibiotics, mimicry, biogas, green chemicals

- Why metagenomics?

  Metagenomics can access the complete microbiota (cultivable and uncultivable)
    - How many different taxa are there?
    - What are they doing?
    - How are they doing it?
    - What influence do they have on the ecosystem/biosphere?

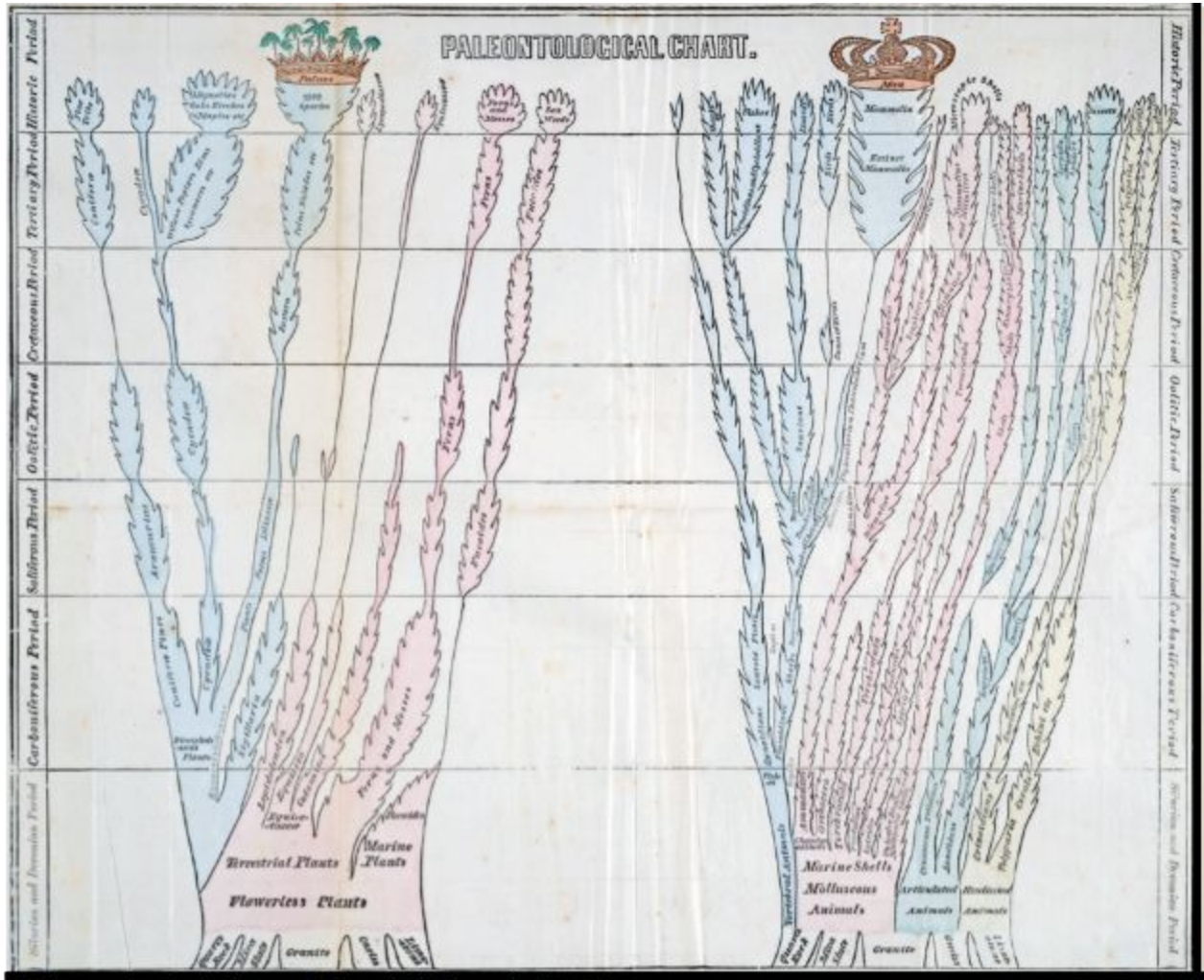
How many kinds of living beings are there?

Tree of Life, Charley Harper

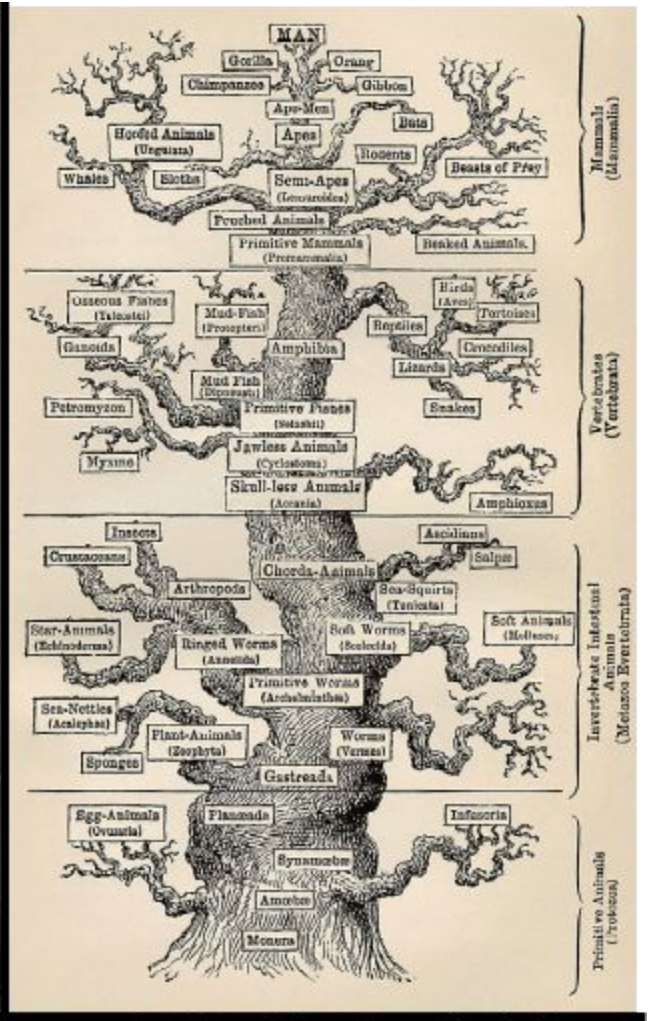
Group	Examples (given by Aristotle)	Blood	Legs	Soul (Rational, Sensitive, Vegetative)	Qualities (Hot–Cold, Wet–Dry)
Man	Man	with blood	2 legs	R, S, V	Hot, Wet
Live-bearing tetrapods	Cat, hare	with blood	4 legs	S, V	Hot, Wet
Cetaceans	Dolphin, whale	with blood	none	S, V	Hot, Wet
Birds	Bee-eater, nightjar	with blood	2 legs	S, V	Hot, Wet, except Dry eggs
Egg-laying tetrapods	Chameleon, crocodile	with blood	4 legs	S, V	Cold, Wet except scales, eggs
Snakes	Water snake, Ottoman viper	with blood	none	S, V	Cold, Wet except scales, eggs
Egg-laying fishes	Sea bass, parrotfish	with blood	none	S, V	Cold, Wet, including eggs
(Among egg-laying fishes): placental selachians	Shark, skate	with blood	none	S, V	Cold, Wet, but placenta like tetrapods
Crustaceans	Shrimp, crab	without	many legs	S, V	Cold, Wet except shell
Cephalopods	Squid, octopus	without	tentacles	S, V	Cold, Wet
Hard-shelled animals	Cockle, trumpet snail	without	none	S, V	Cold, Dry (mineral shell)
Larva-bearing Insects	Ant, cicada	without	6 legs	S, V	Cold, Dry
Spontaneously-generating	Sponges, worms	without	none	S, V	Cold, Wet or Dry, from earth
Plants	Fig	without	none	V	Cold, Dry
Minerals	Iron	without	none	none	Cold, Dry

A detail from 'The School of Athens' by Raphael (1509)

Aristotle's 'Scala naturae', ~350 BC



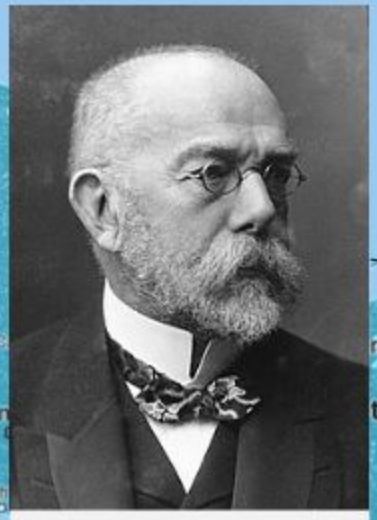
"Paleontological Chart", Edward Hitchcock (1840)



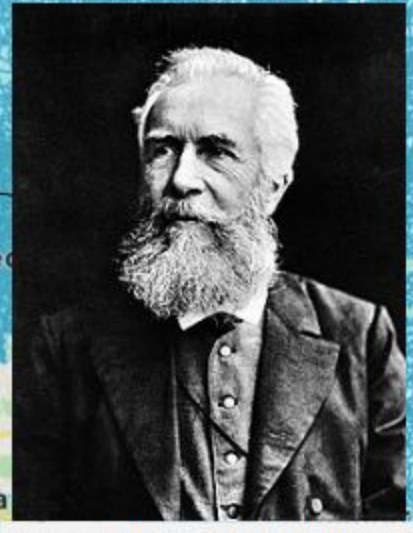
"Evolution of Man", Ernst Haeckel (1879)

2 hr 45 min (260,5 km) via A9

13 hr 39 min (263,4 km) via B2



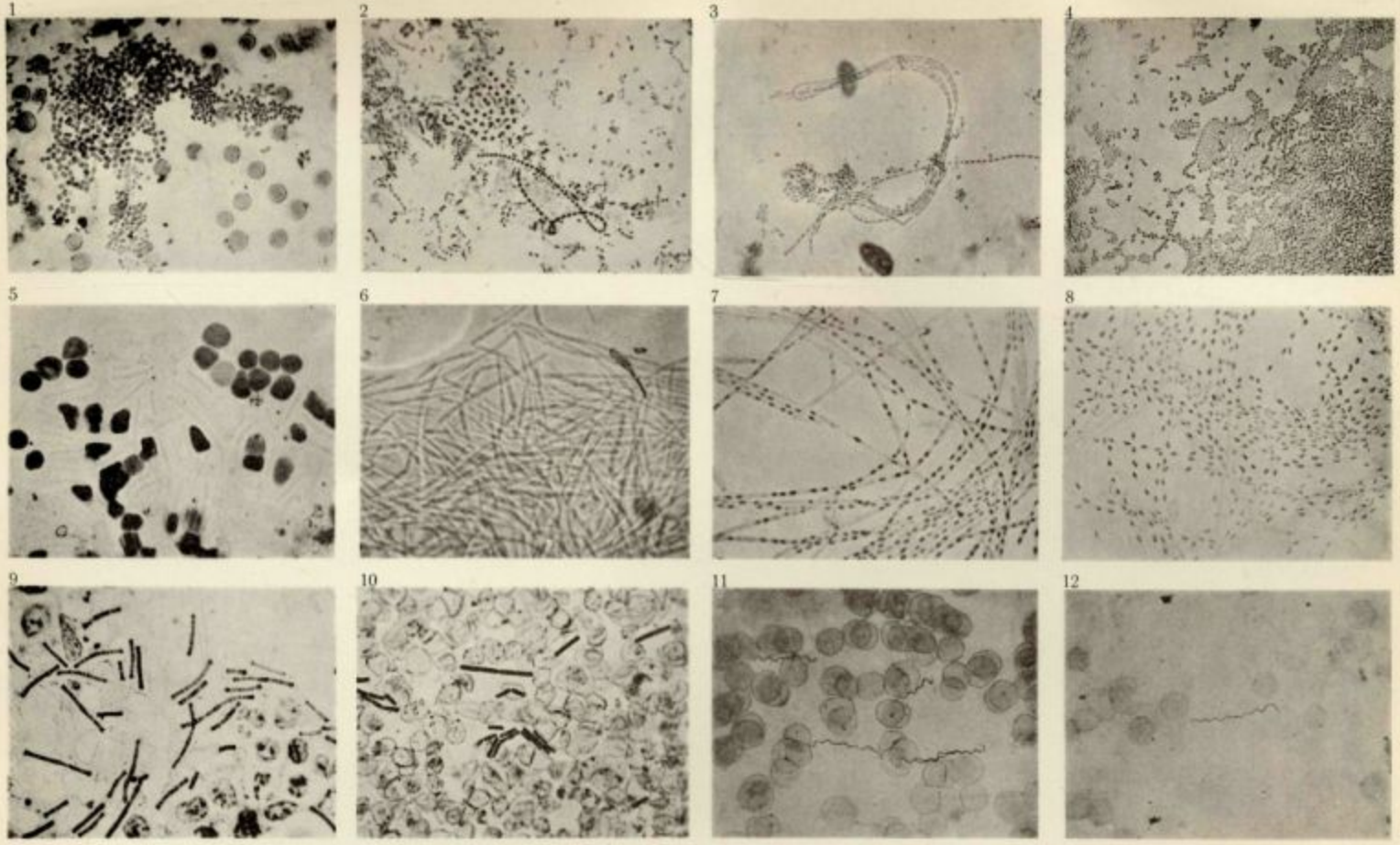
Heinrich Hermann
Robert Koch



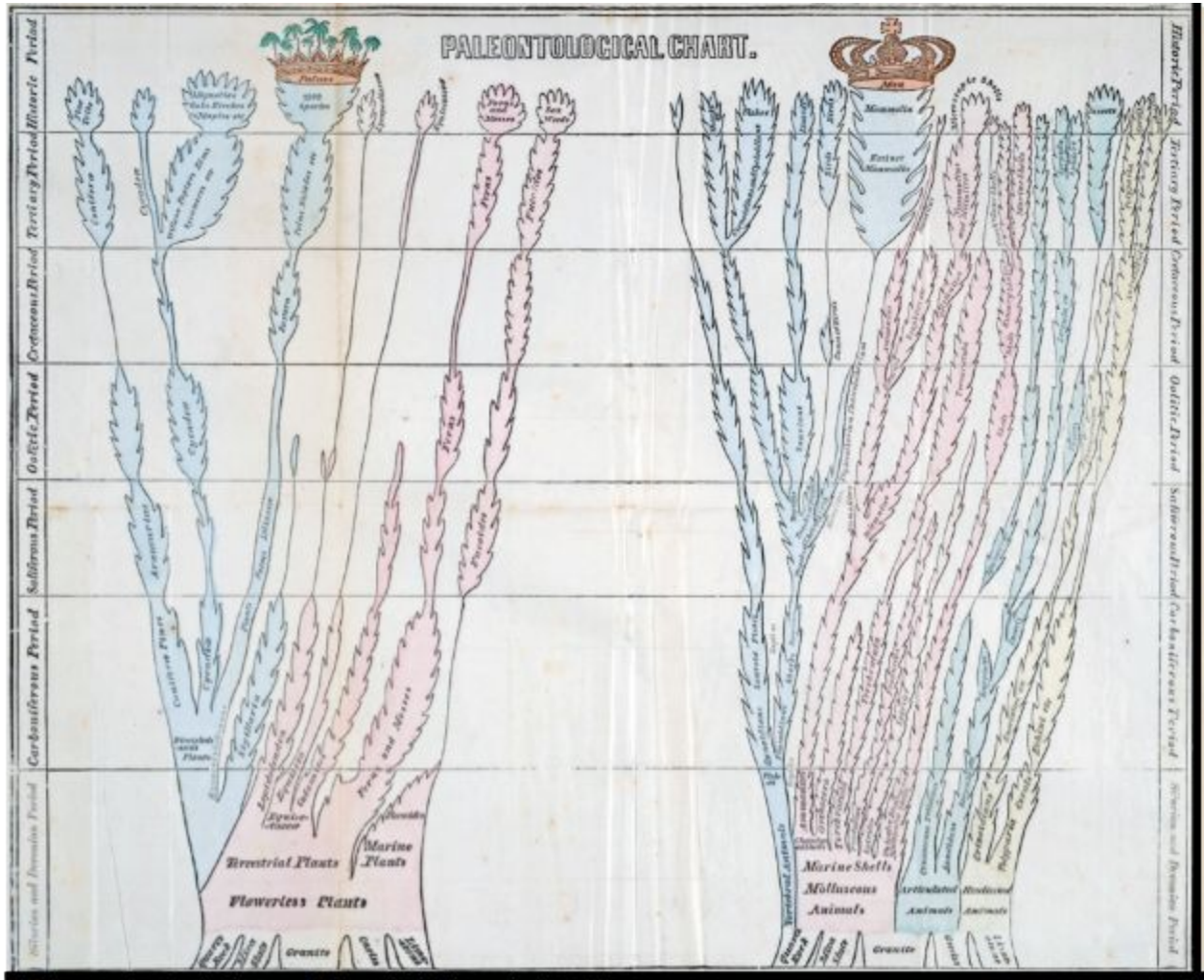
Ernst Heinrich Philipp
August Haeckel

In 1880's Germany

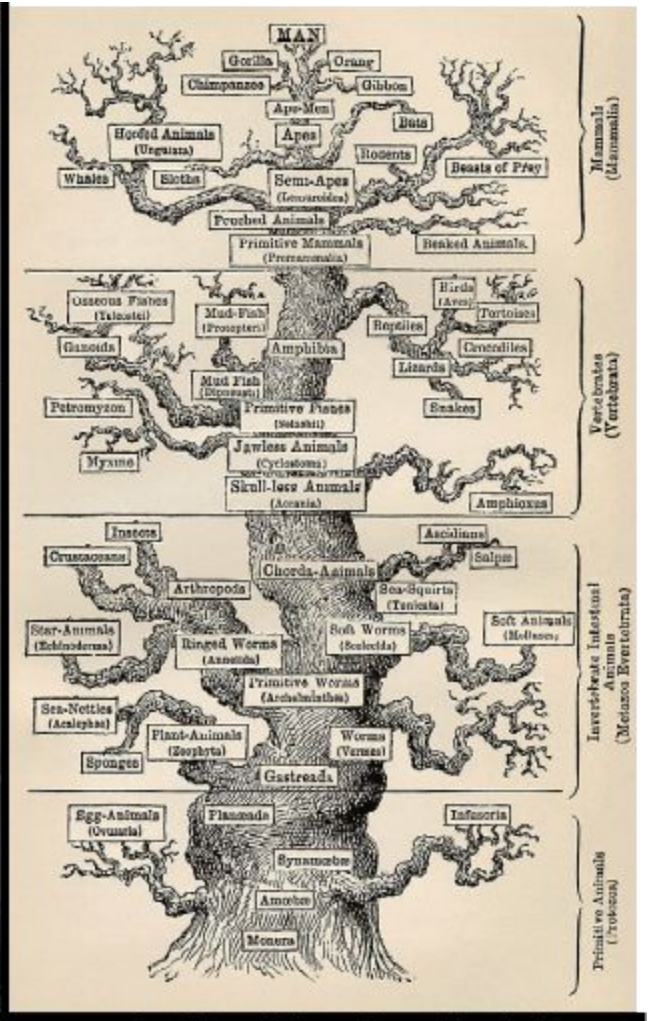
United Kingdom
Isle of Man
Leeds
Liverpool
ENGLAND
WALES
Oxford
Bristol
Cardiff
Plymouth
Southampton
English Channel
Guernsey
Jersey
Paris
France
La Rochelle
Hamburg
Bremen
Szczecin
Poland
Warsaw
Poznan
Łódź
Berlin
Hanover
Wolfsburg
Brunswick
Magdeburg
Dortmund
Leipzig
Dresden
Cologne
Germany
Frankfurt
Prague
Luxembourg
Mannheim
Nuremberg
Stuttgart
Strasbourg
Augsburg
Munich
Salzburg
Austria
Zürich
Liechtenstein
Switzerland



Verfahren zur Untersuchung, zum Konservieren etc. der Bakterien,



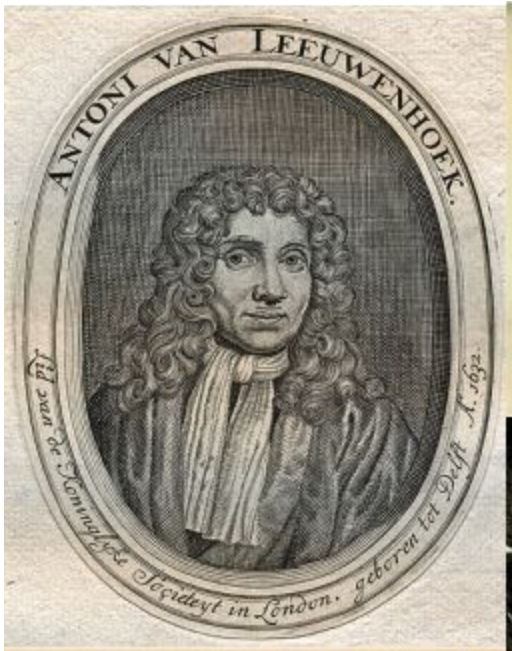
"Paleontological Chart", Edward Hitchcock (1840)



"Evolution of Man", Ernst Haeckel (1879)

Image by Marcin Zemla and Manfred Auer, JBEI

CDC Public Health Image Library


ANTONI VAN LEEUWENHOEK.
Lid van de Konincklijcke Societeyt in Londen, geboren tot Delft A. 1632.

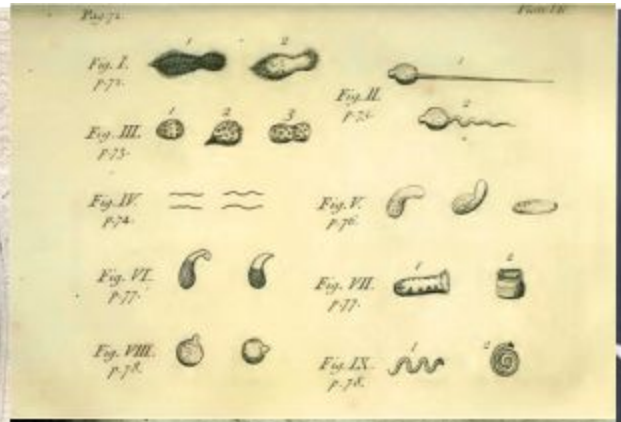
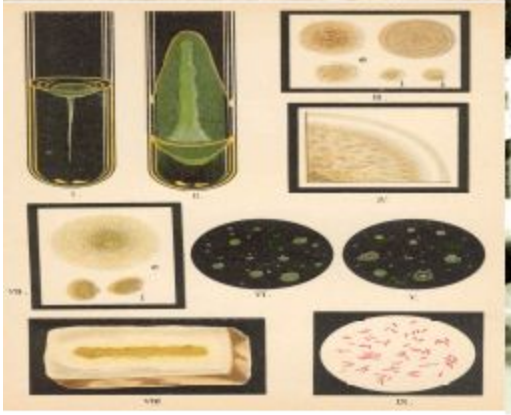
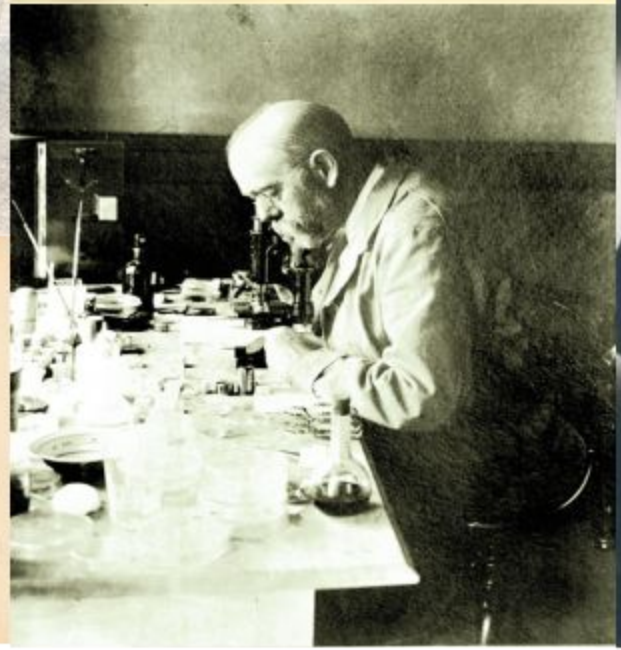
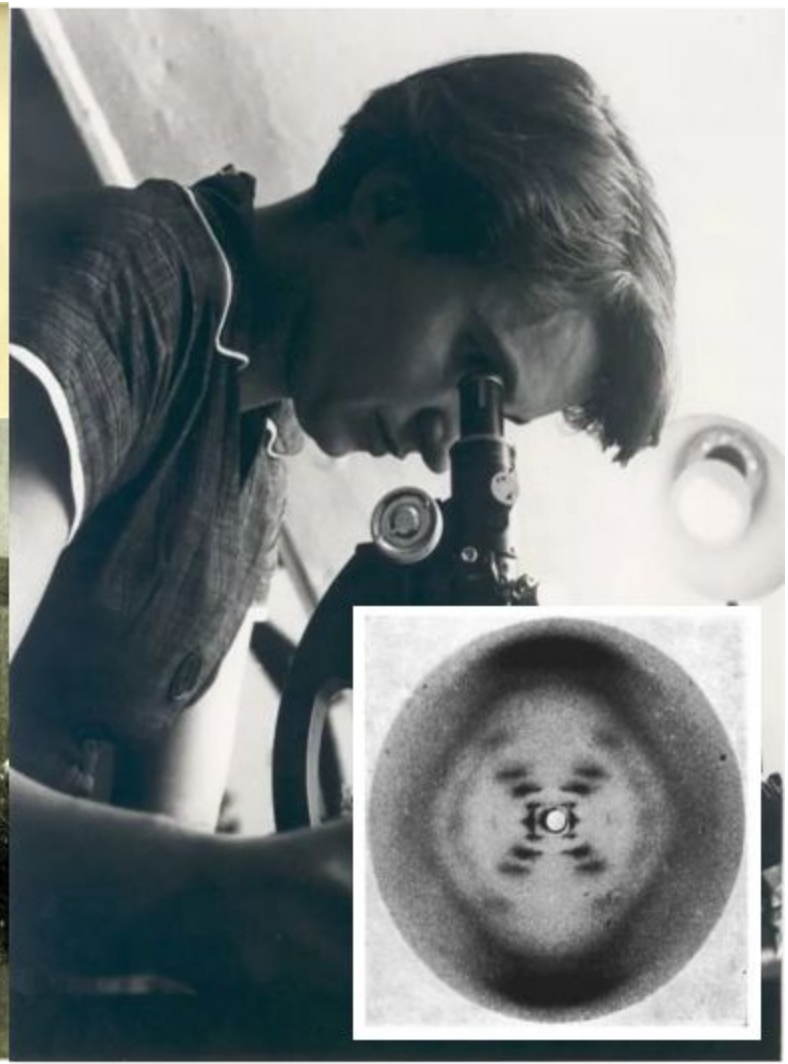
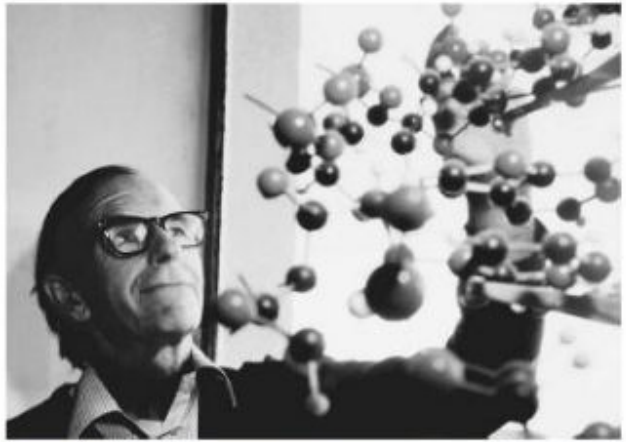


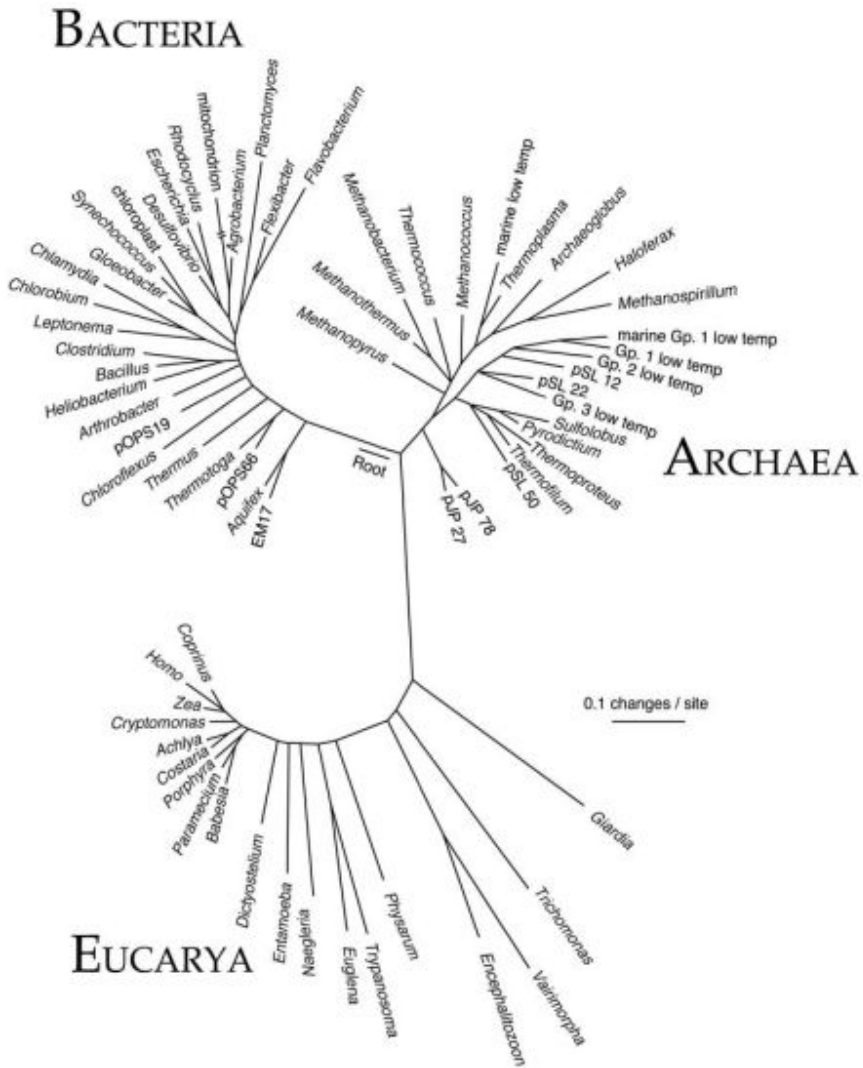
Fig. I.
Fig. II.
Fig. III.
Fig. IV.
Fig. V.
Fig. VI.
Fig. VII.
Fig. VIII.
Fig. IX.










BACTERIA
Planctomyces
mitochondrion
Agrobacterium
Flexibacter
Flavobacterium
Rhodocyclus
Escherichia
Desulfovibrio
chloroplast
Synechococcus
Gloeobacter
Chlamydia
Chlorobium
Leptonema
Clostridium
Bacillus
Heliobacterium
Arthrobacter
pOPS19
Chloroflexus
Thermus
Thermotoga
pOPS66
Aquifex
EM17
Root
Methanobacterium
Thermococcus
Methanothermus
Methanopyrus
Methanococcus
marine low temp
Thermoplasma
Archaeoglobus
Haloferax
Methanospirillum
marine Gp. 1 low temp
Gp. 1 low temp
Gp. 2 low temp
pSL 12
pSL 22
Gp. 3 low temp
Sulfolobus
Pyrodictium
Thermoproteus
Thermofilum
pSL 50
pJP 78
pJP 27
ARCHAEA
0.1 changes / site
Coprinus
Homo
Zea
Cryptomonas
Achlya
Costaria
Porphyra
Paramecium
Babesia
Dictyostelium
Entamoeba
Naegleria
Euglena
Trypanosoma
Physarum
Giardia
Trichomonas
Vairimorpha
Encephalitozoon
EUCARYA

Cultivable microbes from the environment

Marine water (0.001 – 0.1%)

Soil (0.3 – 1 %)

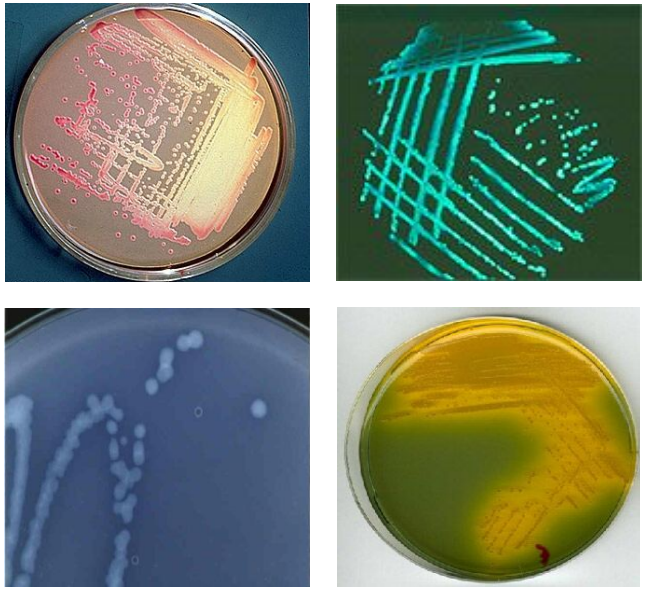
Hot spring (0.1 – 5%)

Fresh water (0.01 – 0.25 %)

- 72,463 type strains are in the strain collections (DSMZ)
- 10^12 are estimated to still be out there

Problem: Only a small percentage of the microorganisms found in the environment can be cultivated

Classical approach: Isolation and characterization



≠

Microbial community

Difficult to find ideal culturing conditions for, e.g: syntrophs, symbionts, slow growers, particle associated taxa

⇒ Cultures do not reflect the diversity of microbes in nature, or the behaviour of microbial communities

What is “Metagenomics”?

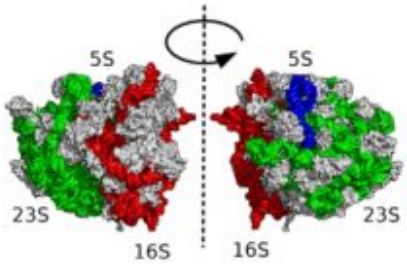
- Cultivation-independent method based on short or long-read sequencing
- Metagenomics means the application of modern genomic techniques (sequencing) to study the entirety of a microbial community straight from their natural environment.
- Metagenomics bypasses the need for isolation and lab cultivation of individual species, consortia
- Provides information regarding community metabolism, interactions, and processes

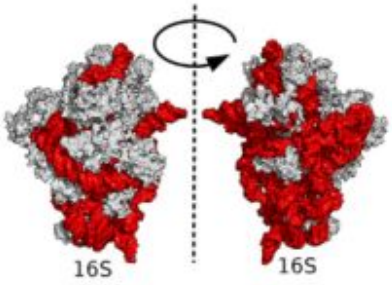
What can Metagenomics *not* do?

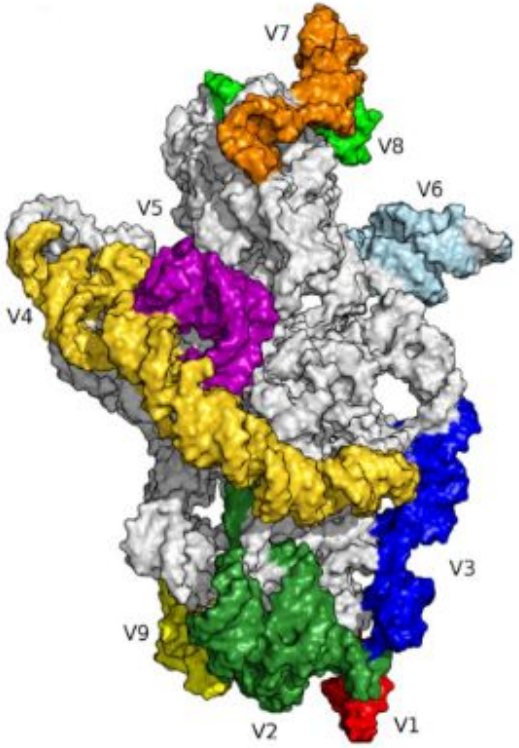
- Provides proportions of microbes in a given environment, not real numbers

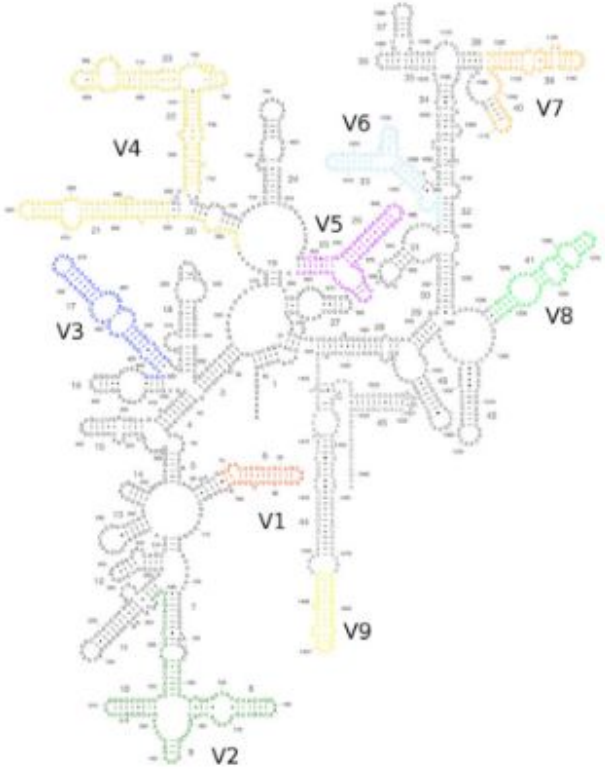
- Based mostly on known information on gene functions and metabolic network
  - Currently 40-60% of genes are still hypothetical
  - A metagenome is only as good as the databases it uses

- Current developments:
  - Computational methods, such as machine learning are being used to hypothesize functions for unknown genes and proteins



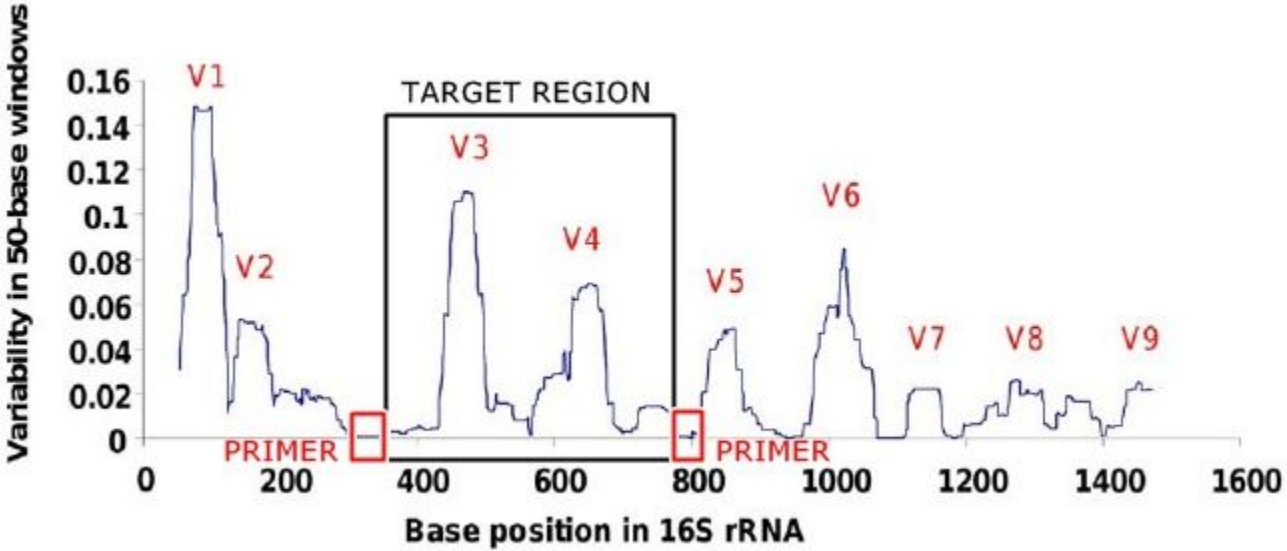






Adapted from Ramazotti & Bacci (2018), doi:10.1016/B978-0-08-102268-9.00005-7

16S as a phylogenetic marker

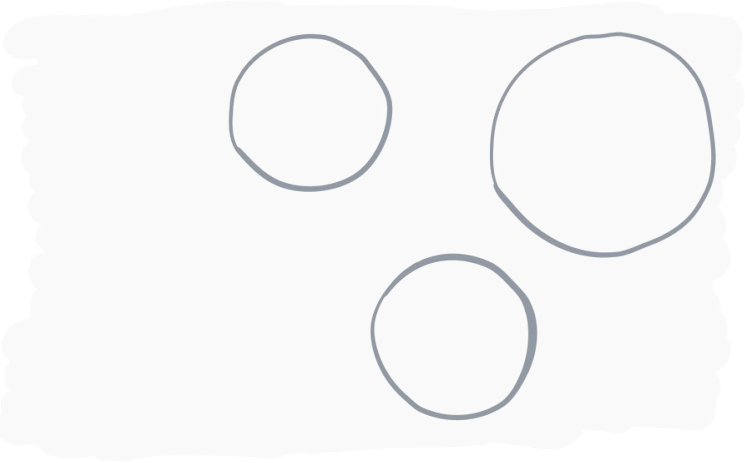


- Widely conserved (bacteria, archaea): easy to target across all bacteria
- 9 hyper-variable regions, flanked by conserved sequences: distinguish between genus

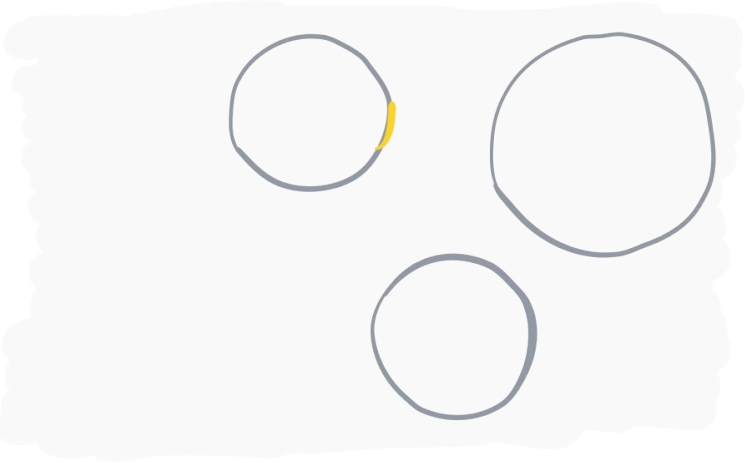
https://www.slideshare.net/bcbbslides/introduction-to-16s-microbiome-analysis

AMPLICON SEQUENCING

http://merenlab.org/momics



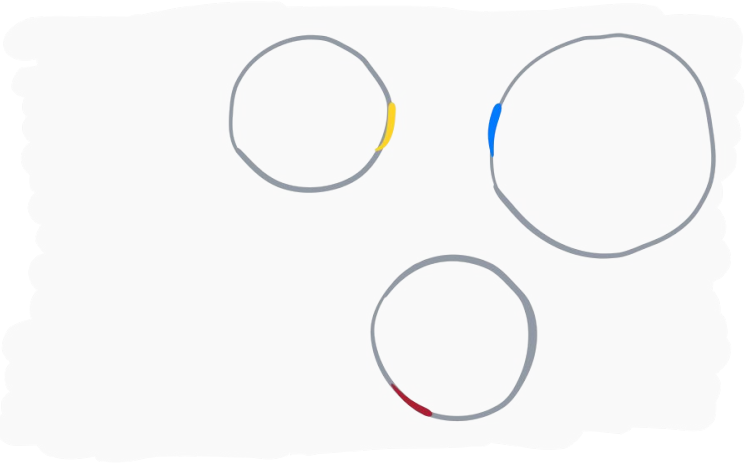
AMPLICON SEQUENCING



AMPLICON SEQUENCING

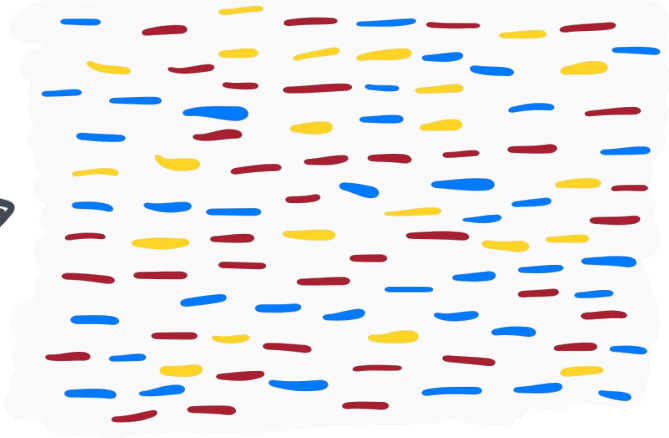

AMPLICON
SEQUENCING

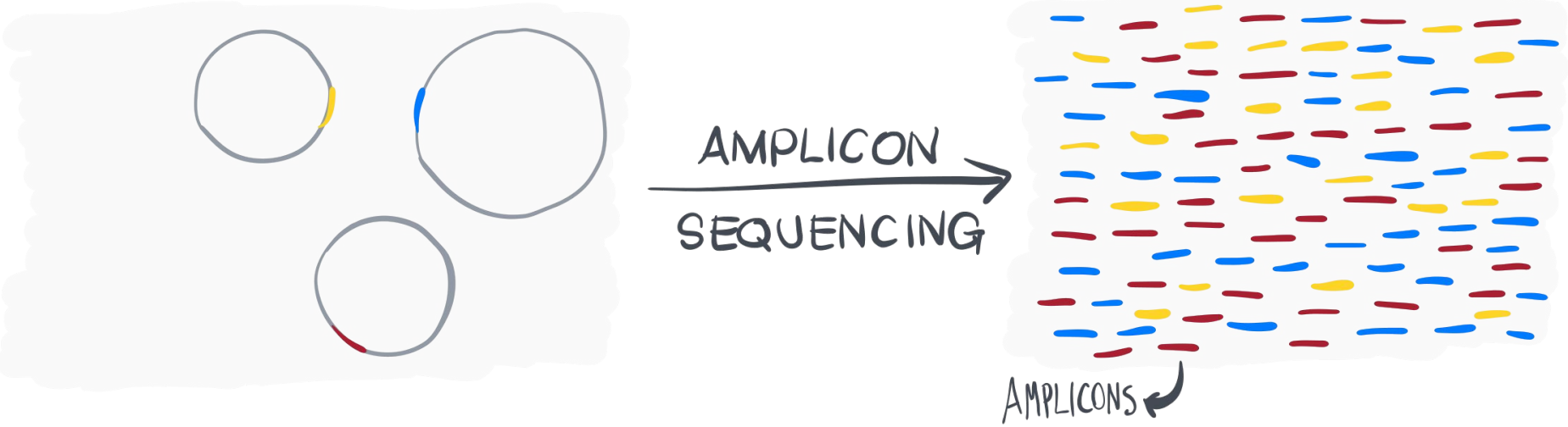

AMPLICON
SEQUENCING



AMPLICON

SEQUENCING



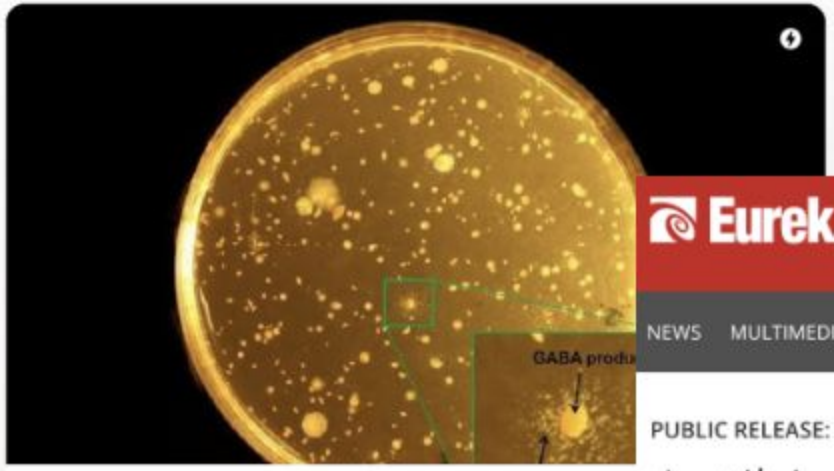

AMPLICON
SEQUENCING
AMPLICONS

MEDICAL NEWS TODAY

Breast cancer: Bacte deficiency linked wit onset

By Ana Sandoiu | Published Mon 9 Oct 2017

Researchers examined the bact makeup of breast tissue in wom with breast cancer and found th has insufficie bacterial gen *Methylobacte*



Researchers Link a Gut Microbe to De

Scientists are learning more about how specific bacteria in our guts are connected to various asp

LabRoots, Inc. • 1 day ago

Can Microbe Altruism?

If gut bacteria can sway their hosts to be selfless, it could answer a riddle that go...

Quanta Magazine • 1 year ago

EurekAlert! | AAAS

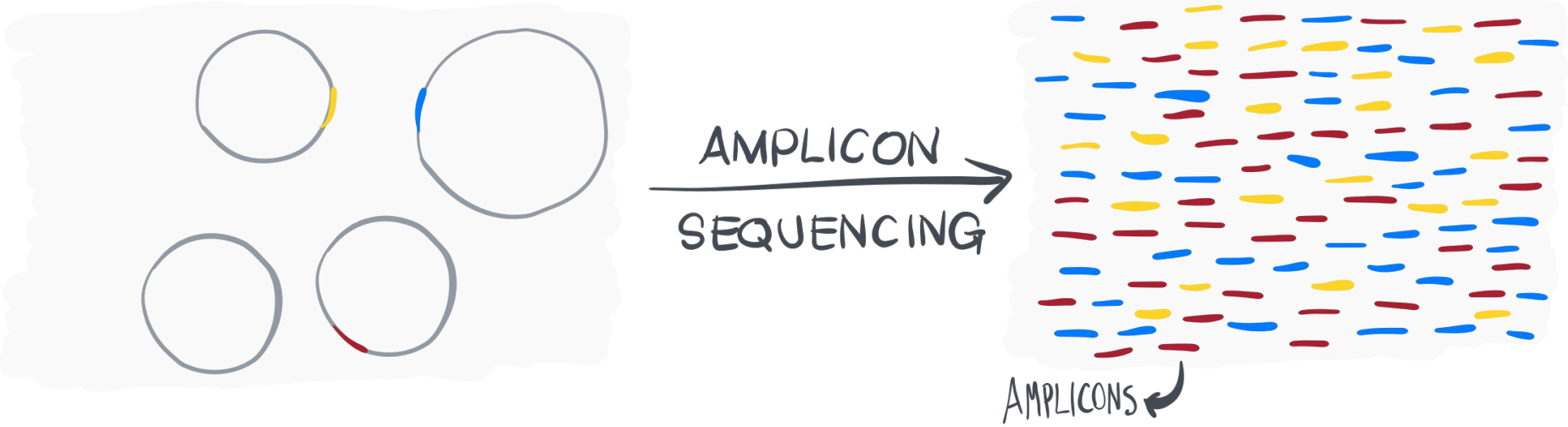
NEWS MULTIMEDIA MEETINGS ABOUT LOGIN

PUBLIC RELEASE: 21-JUL-2016

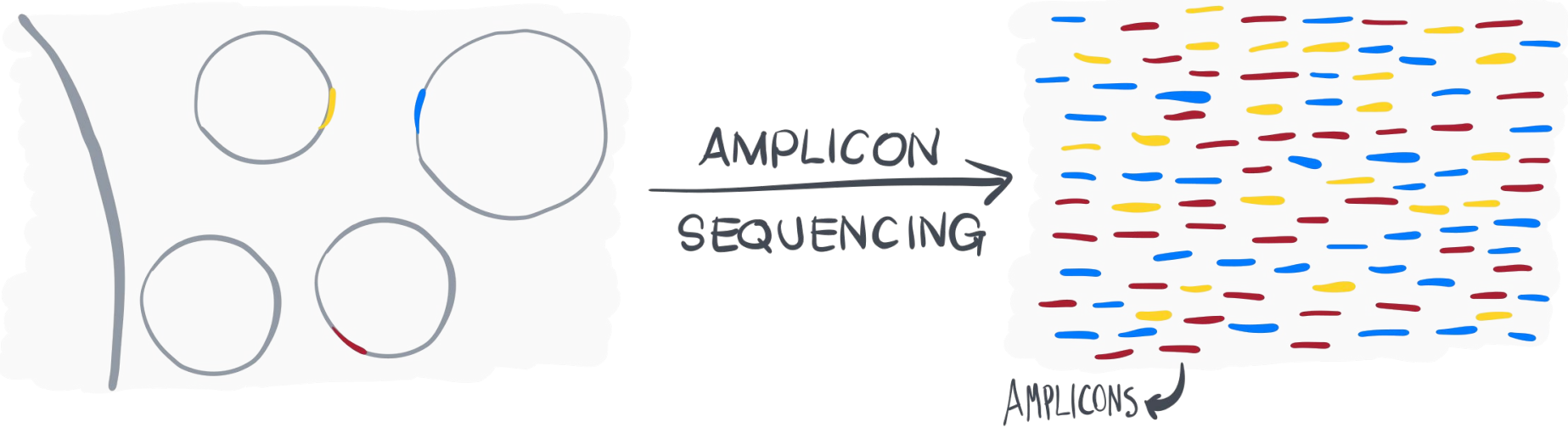
Antibiotics weaken Alzheimer's disease progression through changes in the gut microbiome

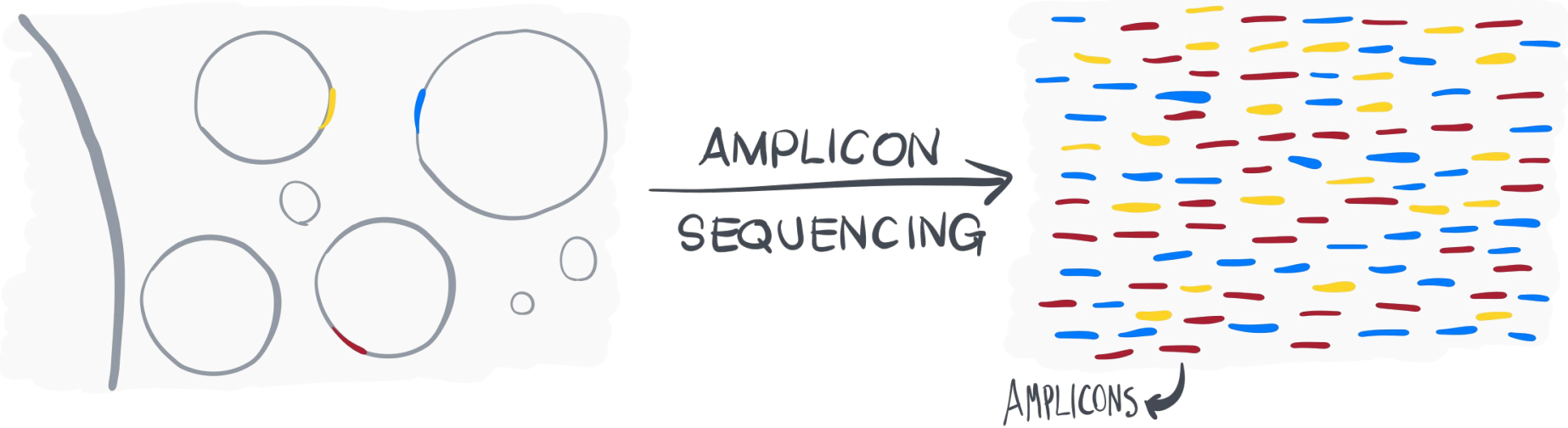
Long-term antibiotic treatment in mice decreases levels of disease-causing plaques and enhances neuroinflammatory activity of microglial cells

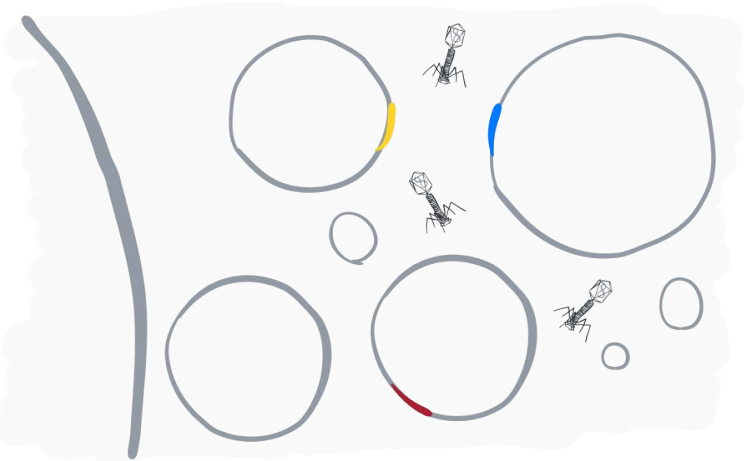
UNIVERSITY OF CHICAGO MEDICAL CENTER



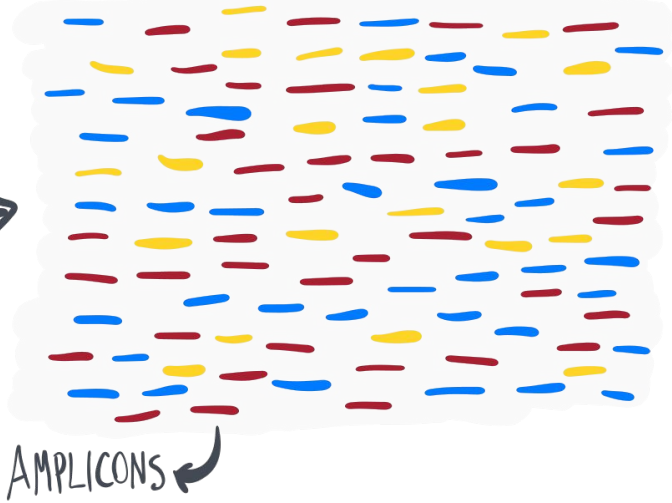
What is missing in amplicon sequencing? What are we not capturing?

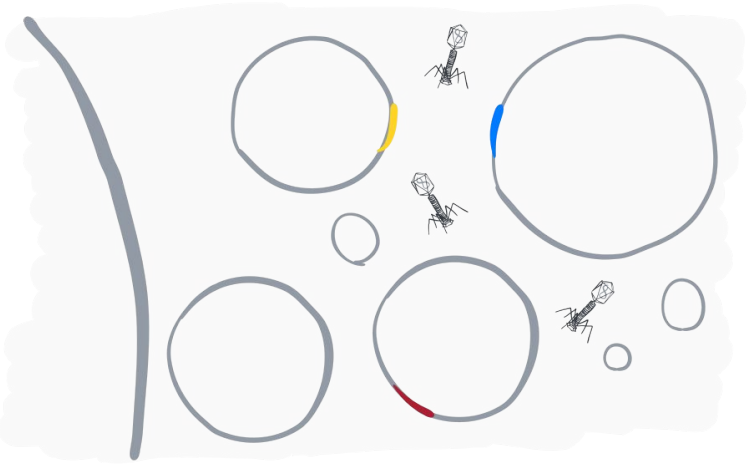

AMPLICON
SEQUENCING
AMPLICONS


AMPLICON
SEQUENCING
AMPLICONS

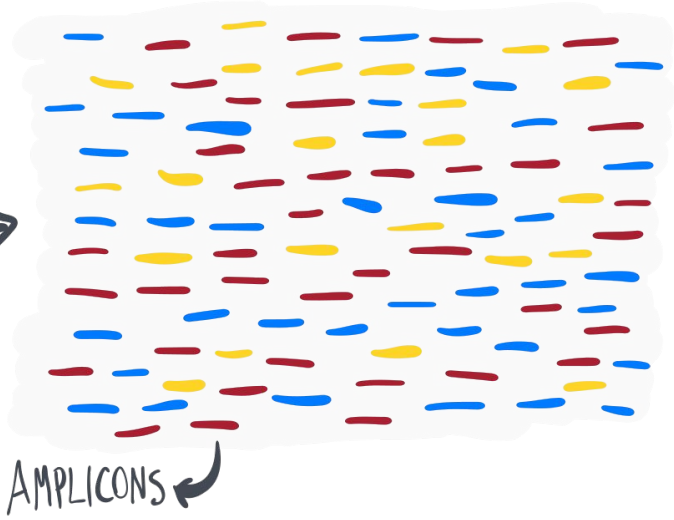


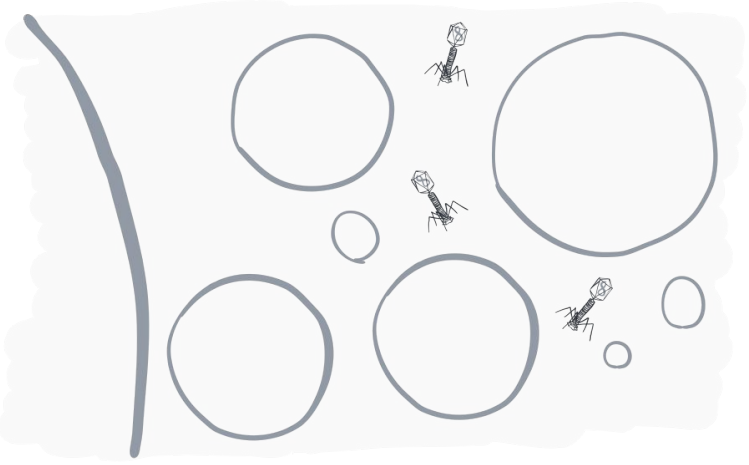
AMPLICON SEQUENCING

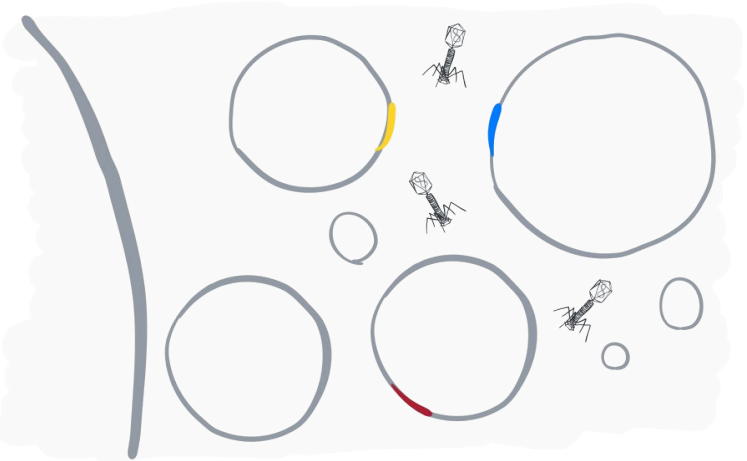




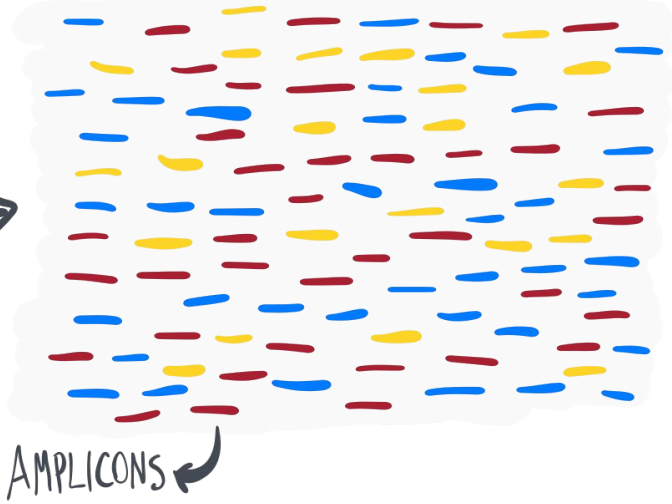
AMPLICON SEQUENCING

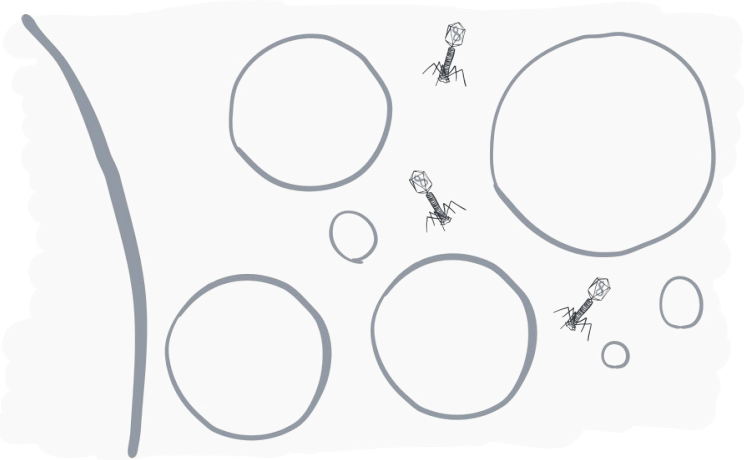




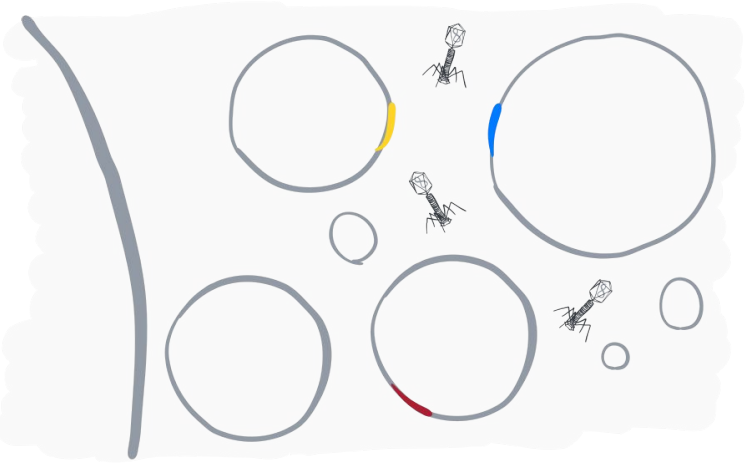


AMPLICON SEQUENCING

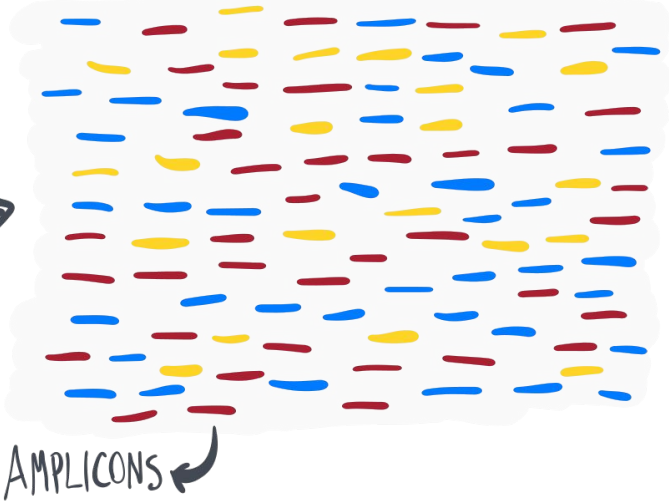




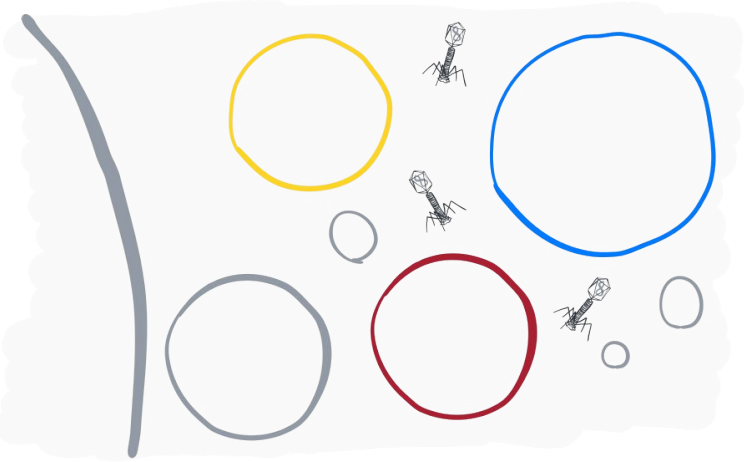
SHOTGUN SEQUENCING



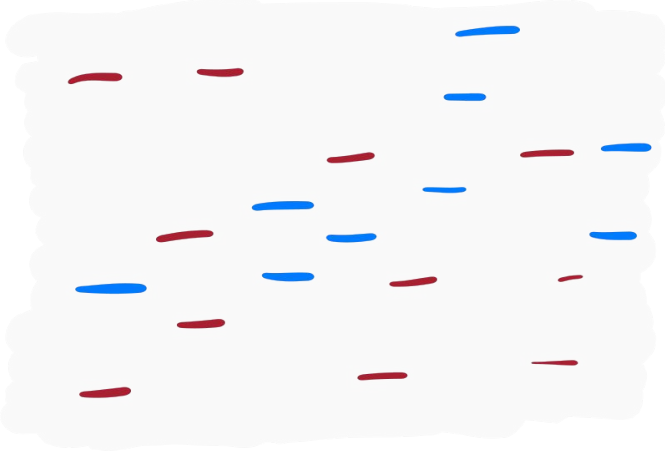
AMPLICON SEQUENCING

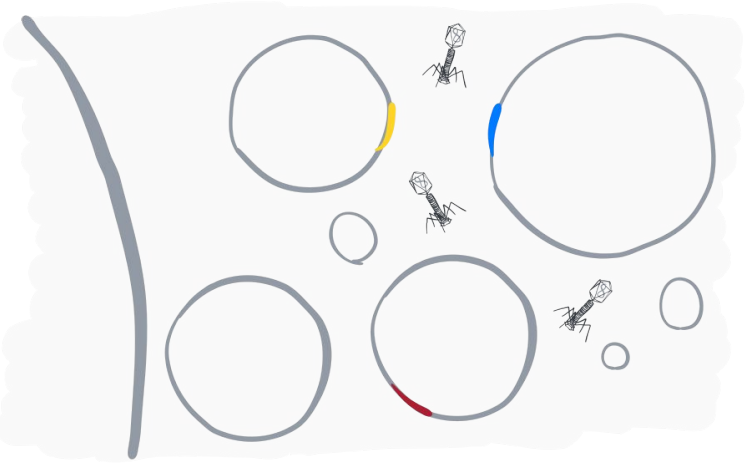


AMPLICONS

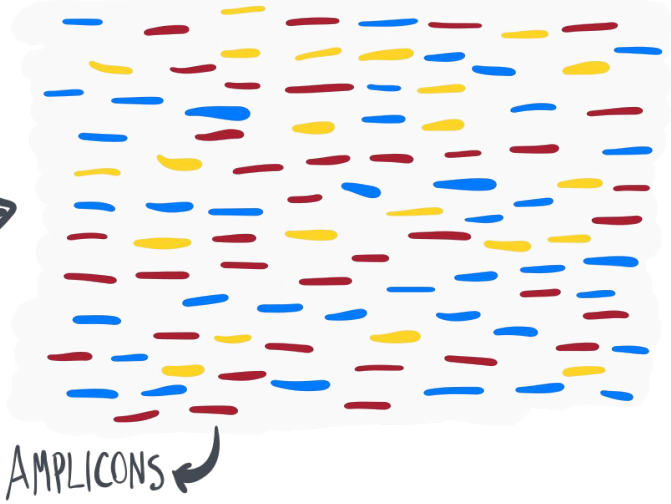


SHOTGUN SEQUENCING

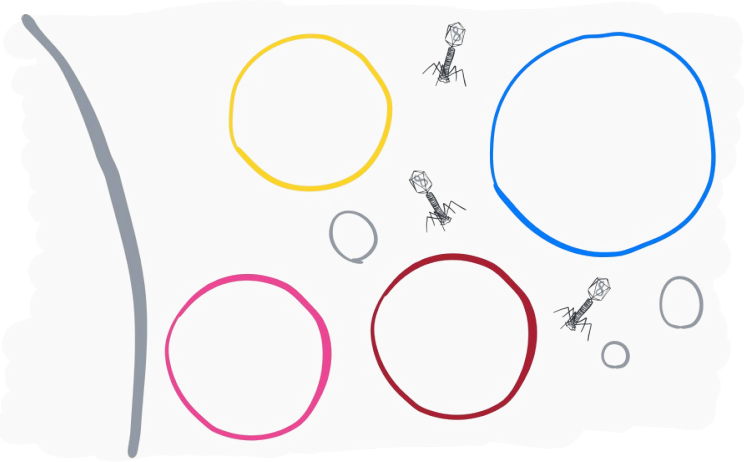




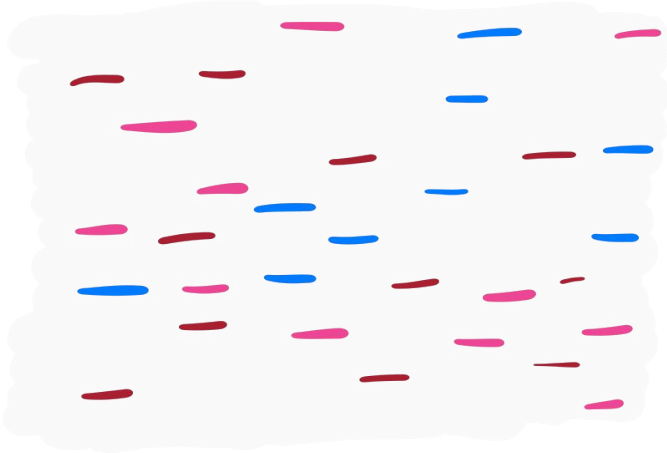
AMPLICON SEQUENCING

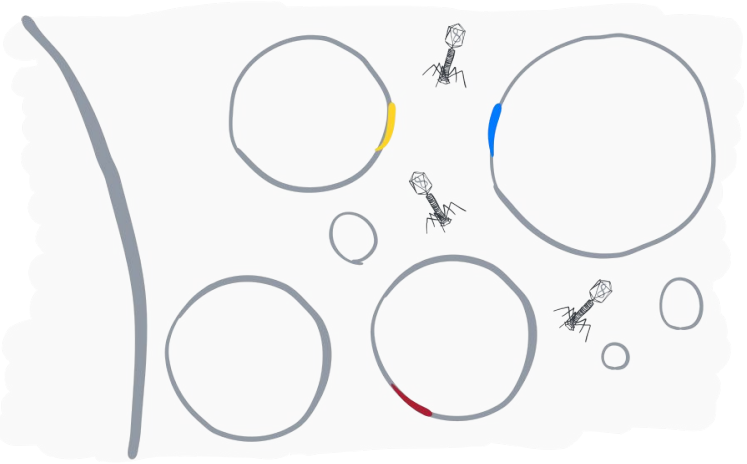


AMPLICONS

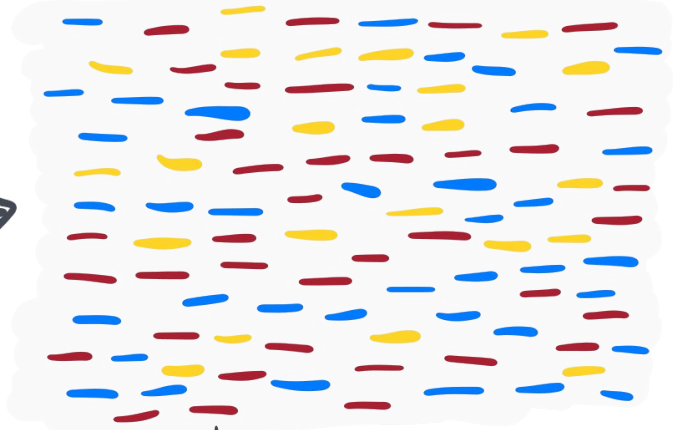


SHOTGUN SEQUENCING

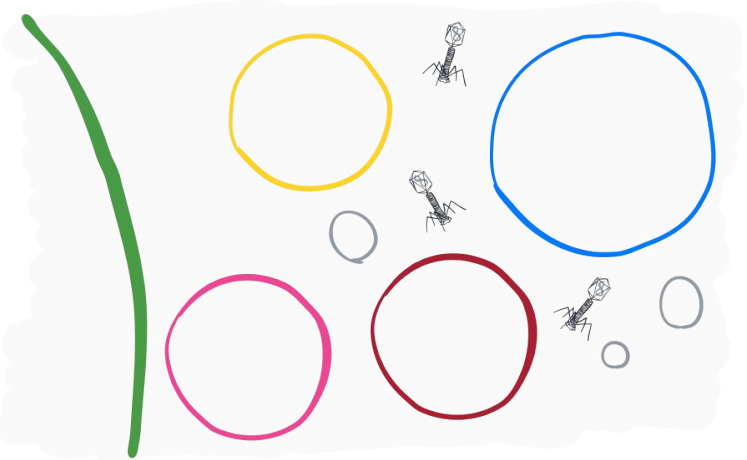




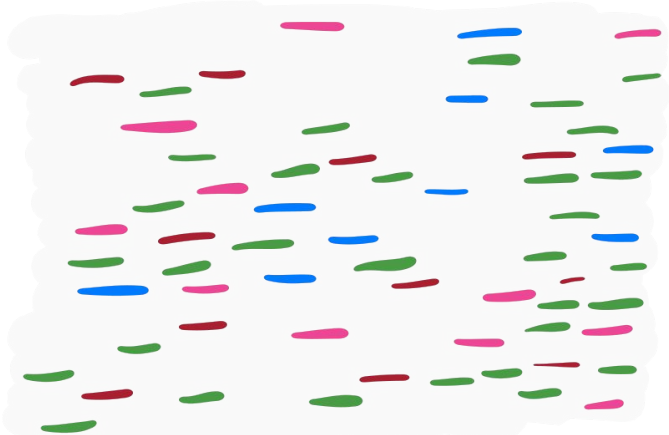
AMPLICON SEQUENCING



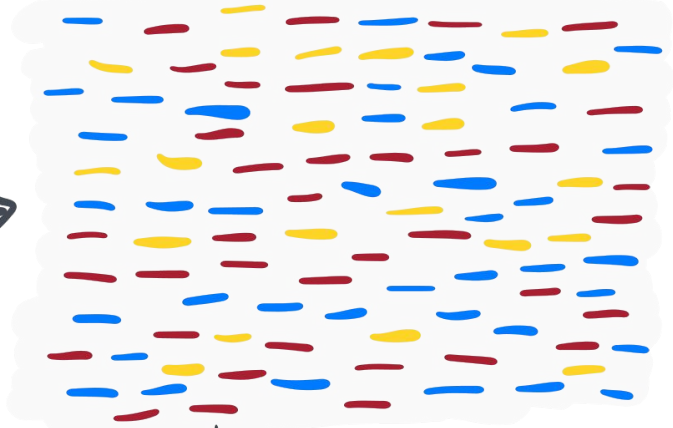
AMPLICONS



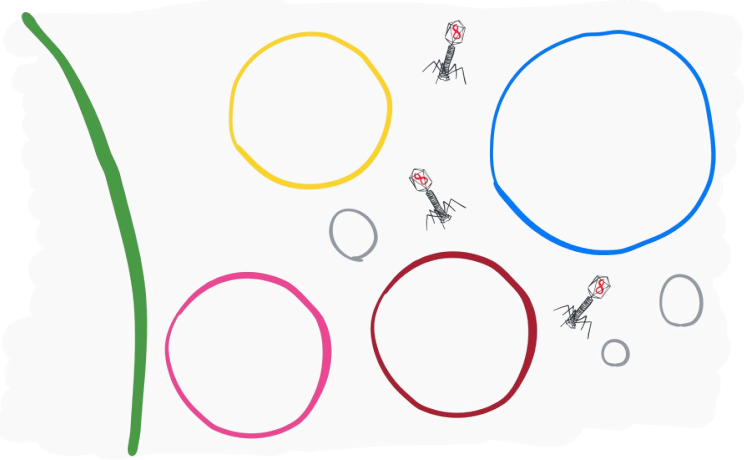
SHOTGUN SEQUENCING



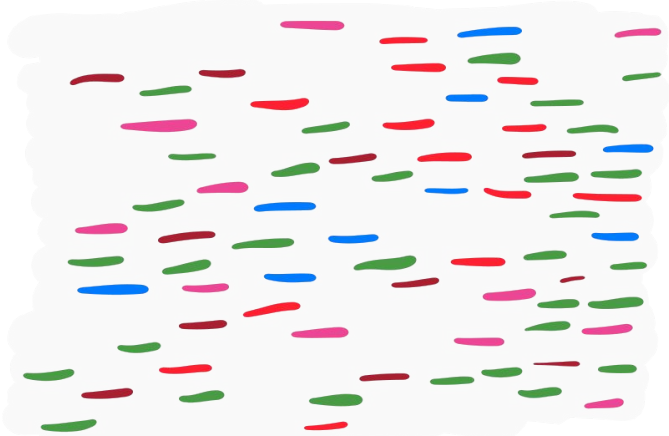
AMPLICON SEQUENCING

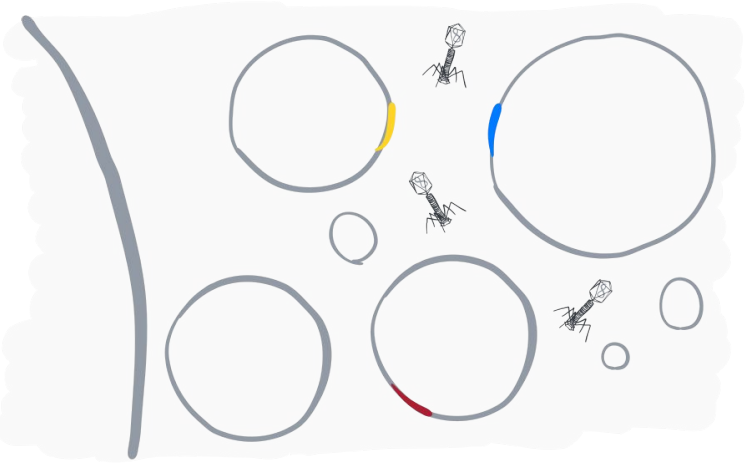


AMPLICONS

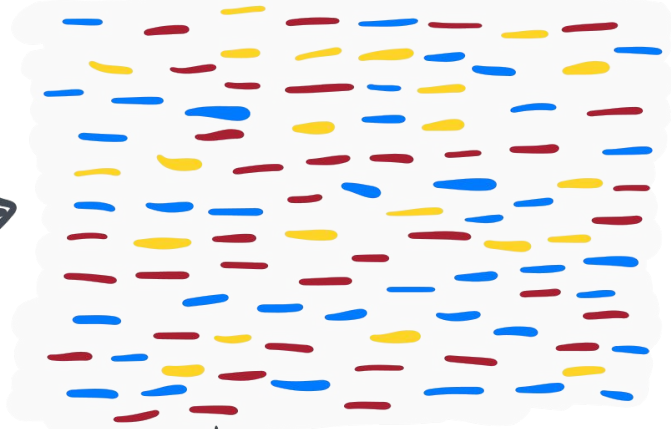


SHOTGUN SEQUENCING

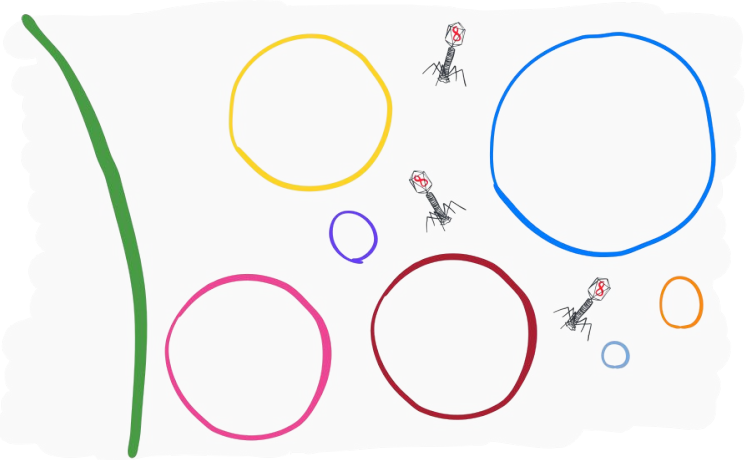




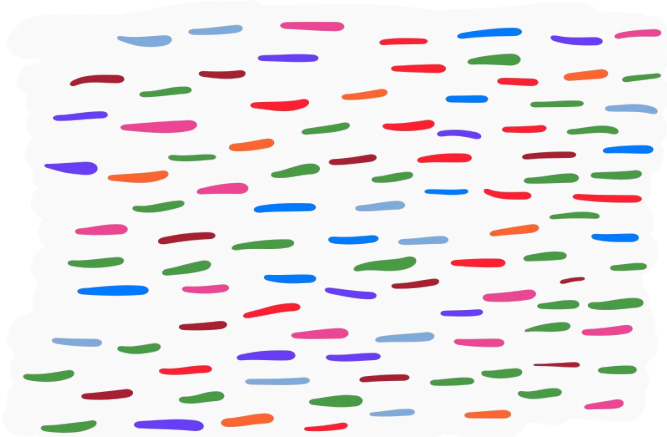
AMPLICON SEQUENCING

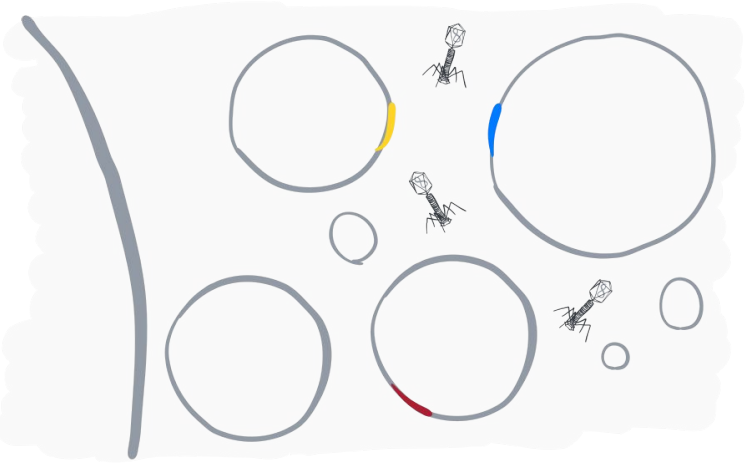


AMPLICONS

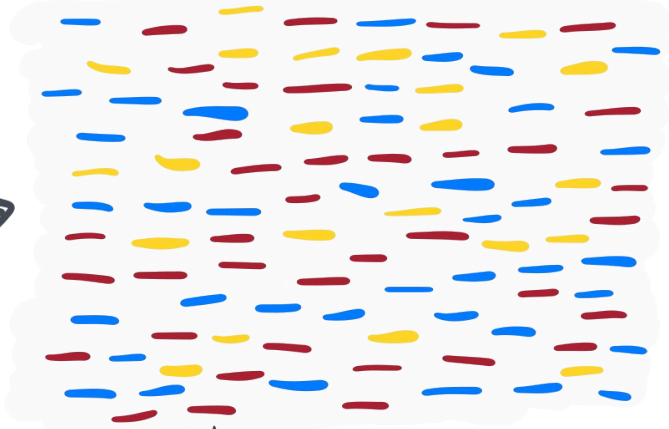


SHOTGUN SEQUENCING



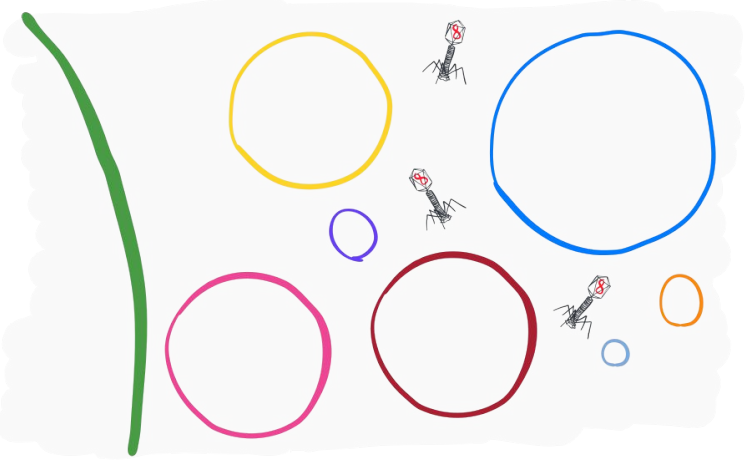


AMPLICON SEQUENCING

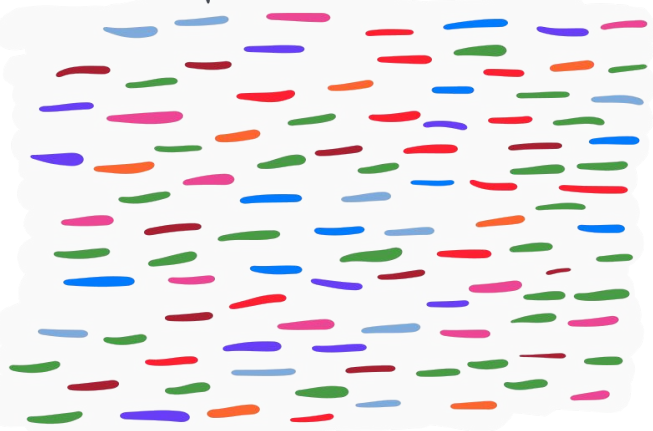


AMPLICONS

METAGENOMIC SHORT READS



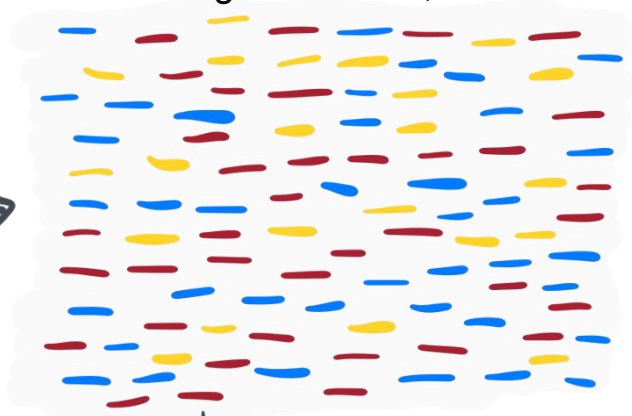
SHOTGUN SEQUENCING



Sequence based (bacterial) diversity

Marker genes
e.g. 16S rRNA, ITS

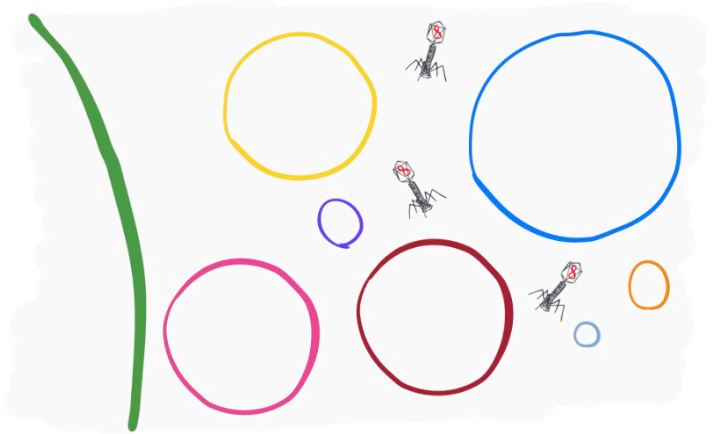
AMPLICON SEQUENCING



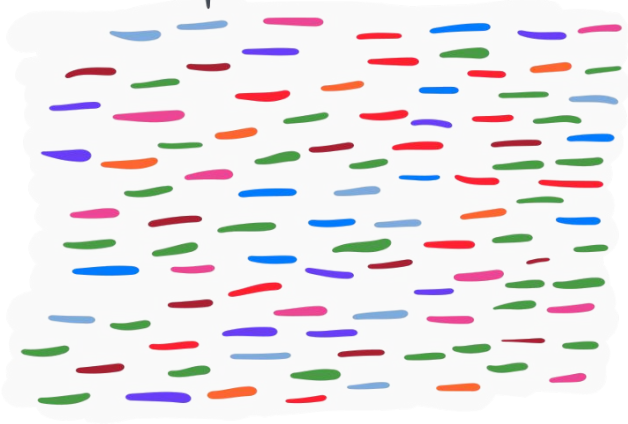
AMPLICONS

METAGENOMIC SHORT READS

Amplicons ≠ Metagenomics



SHOTGUN SEQUENCING

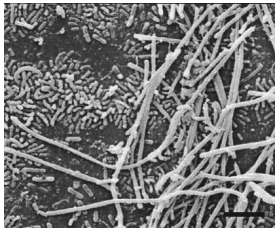


http://merenlab.org/momics

Typical metagenomic workflow

Wetlab

Environmental Sample

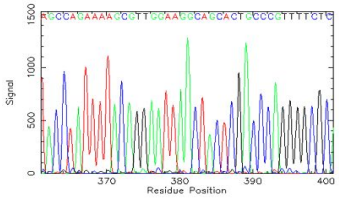


Extract DNA/ RNA/Protein

Sheared Size selection

Library Preparation

High throughput sequencing

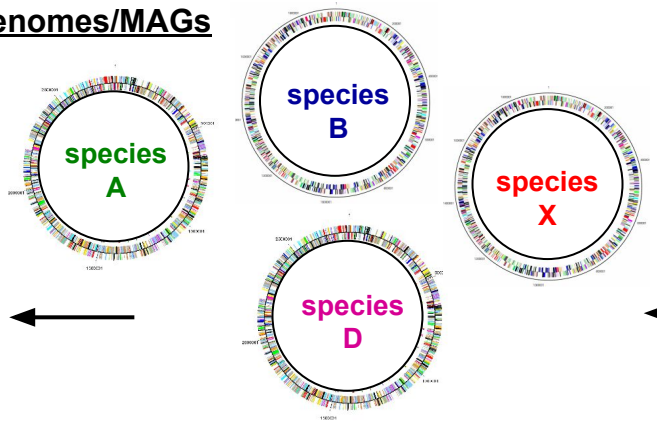


short or long reads

Bioinformatic processing:

- Quality control
- Assembly
- Read mapping

Genomes/MAGs

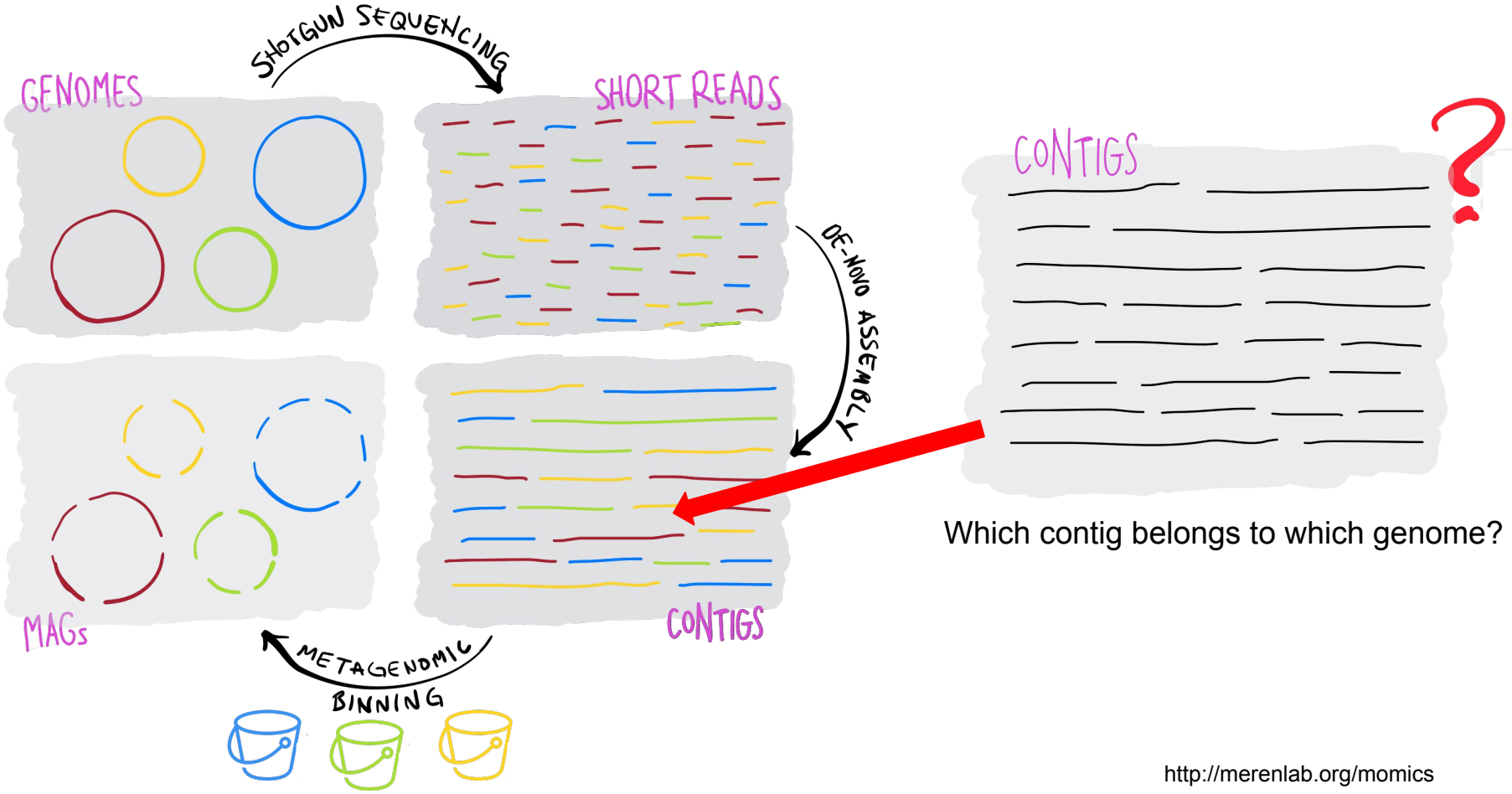


Downstream analysis

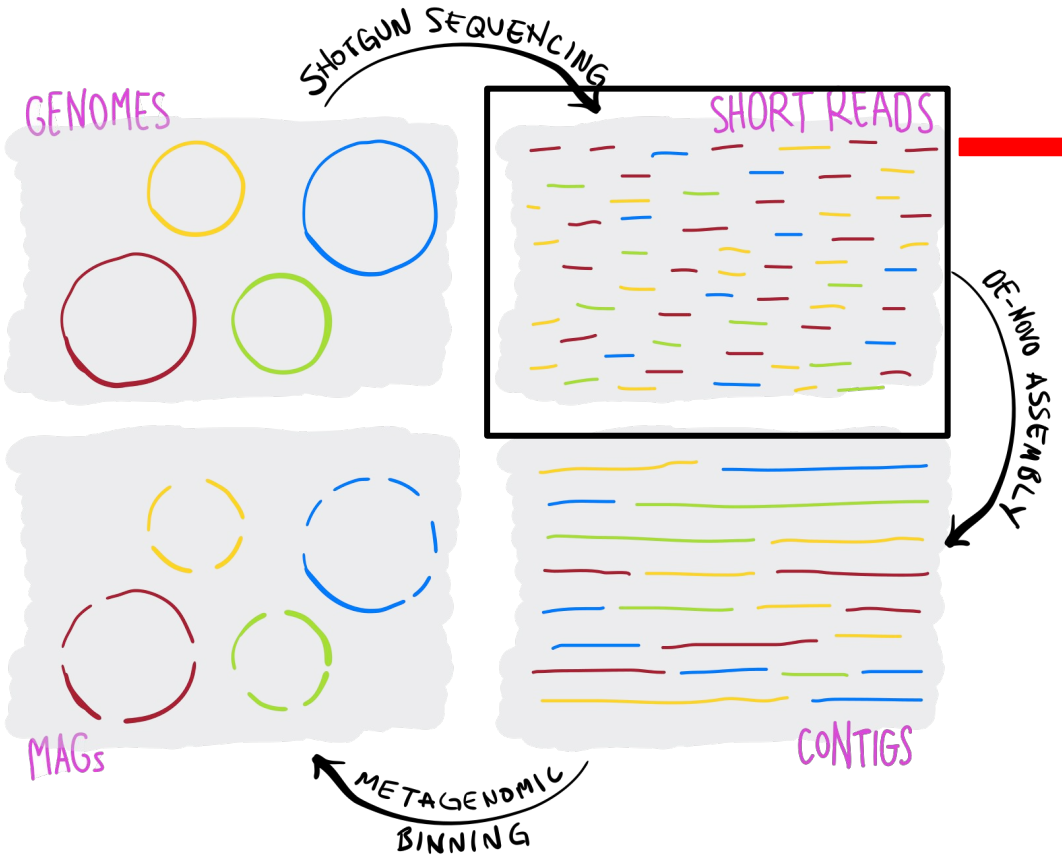
- Abundance estimation
- Gene/Protein annotation

Drylab

What are **MAG**s (**M**etagenome **A**ssembled **G**enomes)?



Step 1: Quality Control (QC) of Short Reads



Step at which to perform QC

What do we need QC for?

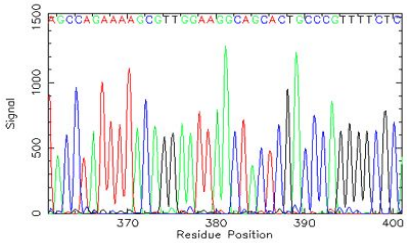
Removal of wet-lab derived errors:

- Sequencing artifacts Erroneous nucleotide readout from sample storage, DNA extraction, library prep and sequencing signals
- PCR duplicates Amplification of multiple sequences from the same DNA fragment
- Chimeras False priming resulting in mixed sequences

- Quality trimming Phred score > 20, often start and end of sequence

- Removal of common contaminants:
  - *phiX174_virus:* Quality control for sequencing run (balanced GC, small genome)
  - *Adapters:* Binding sites between sequence and sequening array
  - *Barcodes:* Identification sequence for each sample (multiplexing)
- Host removal
  - e.g.: chm13v2.0 Human genome sequence

The Phred score



Each fluorescent signal in a sequencing trace file is assigned a quality score based on ist intensity, the Phred score

The phred quality score is the negative ratio of the **error probability** to the reference level of P=1 expressed in Decibel (dB).

Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%

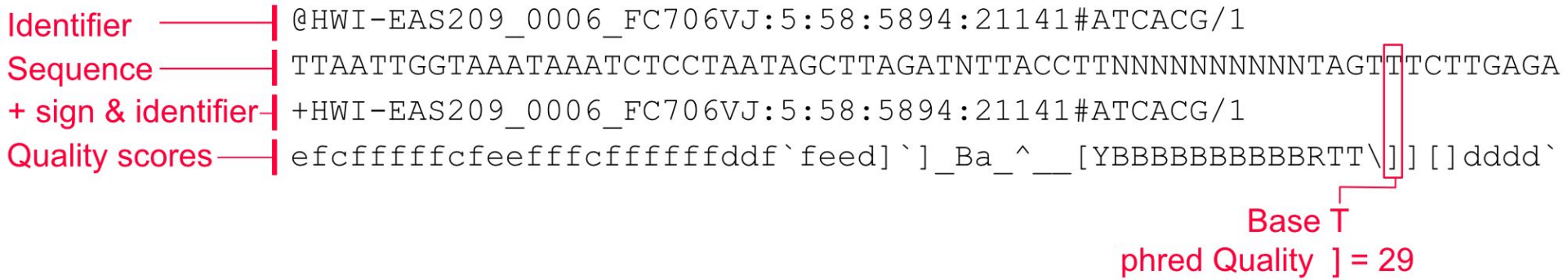
More info and training on QC at:

https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html

e.g. Phred assigns a quality score of 30 to each base:

- The chances that this base is called incorrectly are 1 in 1000.

Fastq format: ID, Sequence and quality

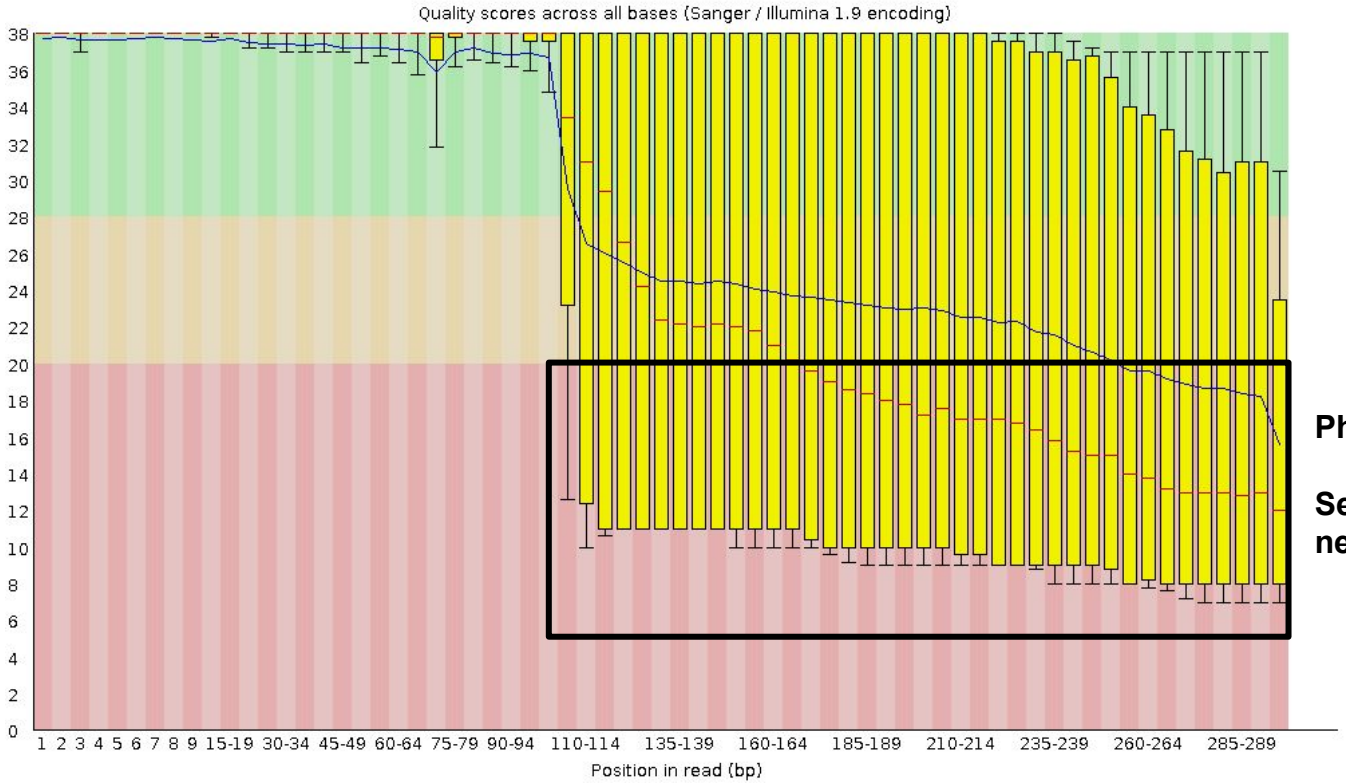


- **Line 1:** @Sequence_ID, technical specifications, Read1 or 2
- **Line 2:** Raw Sequence & Barcode
- **Line 3:** + (additional sequence ID optional) (Location of cluster on flow-cell, R1 or R2)
- **Line 4:** Phred Quality scores (in ASCII*) for the sequence in line 2

- Read 1 and 2 will have the same starting header,1 or 2 at the end identifies the partner read

* **ASCII:** American Standard Code for Information Interchange

Phred score visualization (fastqc)

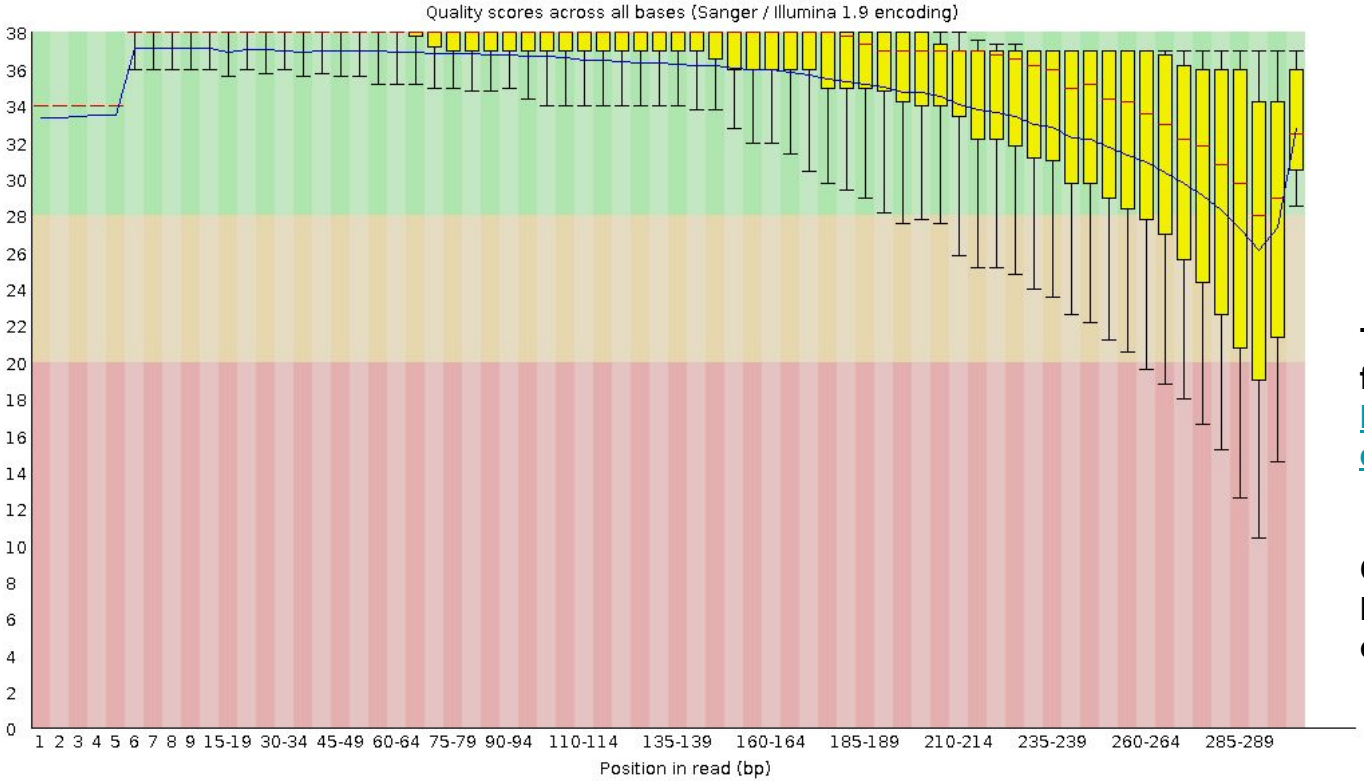


Phred quality score < 20

Sequence trimming necessary

Sequence quality decreases at the end of a read, due to signal decay and phasing (chemistry)

Post-trimming

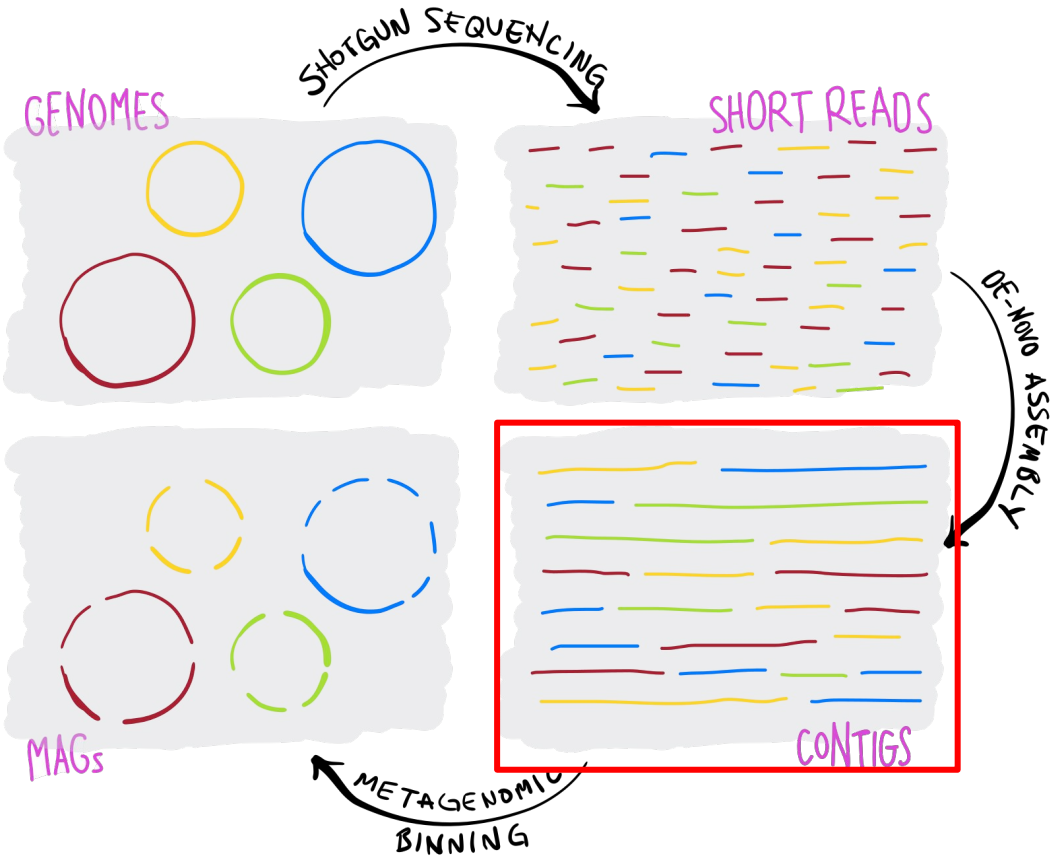


- **Trimming based on phred score**
- **Removal of adapter & barcodes**
- **Removal of duplicates**

Texample tool:
fastp
https://github.com/OpenGene/fastp

Customisation of barcodes, host contamination, phiX often possible

Step 2: Assembly

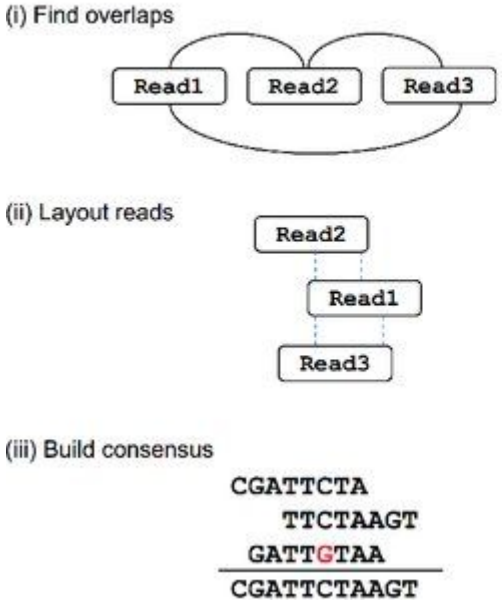


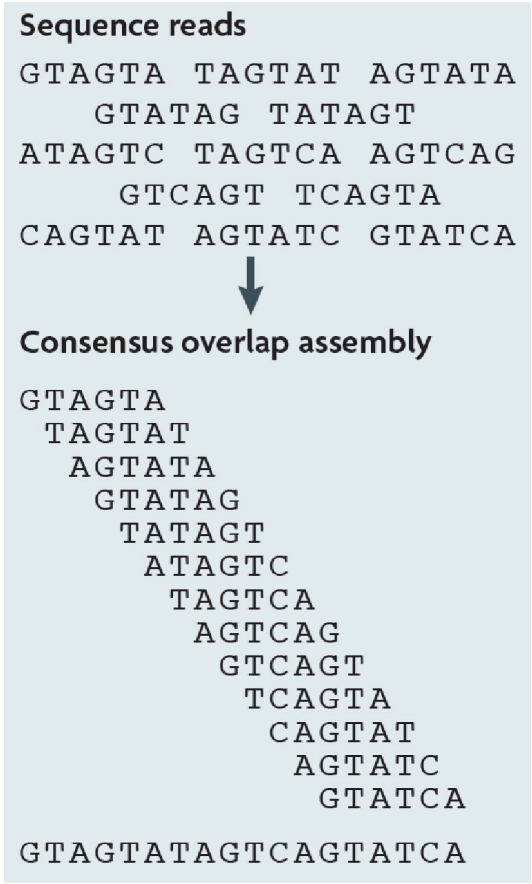
Assembly & Co-assembly:

- **Reference-guided assembly:**
  Uses known sequence as basis.
  Maps/alignes reads onto existing assembly

- **De-Novo assembly:**
  Combines reads into contigs based on sequence overlap
  a) Overlap-Layout-Consensus assembly
  b) De-Brujn-graph assembly

Overlap-Layout-Consensus Assembly

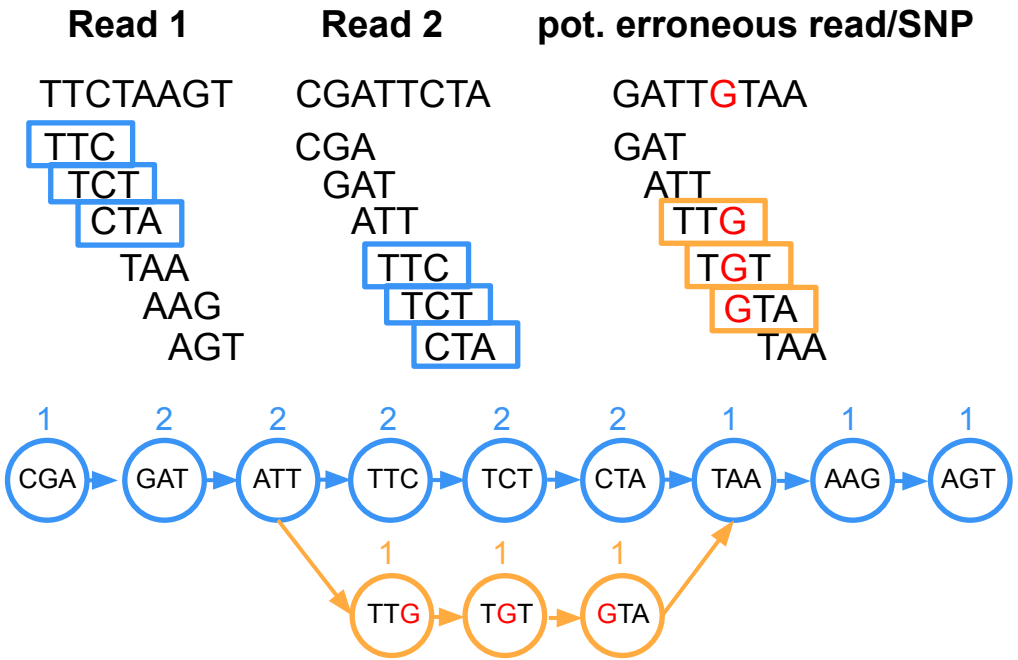




- **Assembly is greatly complicated by sequence heterogeneity and gaps in coverage due to undersampling**
- **Long repetitive regions are troublesome for this approach**
- **Unassembled reads that do not find an overlap are called "singletons"**

Ayling, M., Clark, M.D. and Leggett, R.M., 2020. New approaches for metagenome assembly with short reads. *Briefings in bioinformatics*, *21*(2), pp.584-594.

de Bruijn Graph Assembly



Final contig: CGATTCTAAGT

Step 1: Split sequence into k-mers using a sliding window, e.g: 3

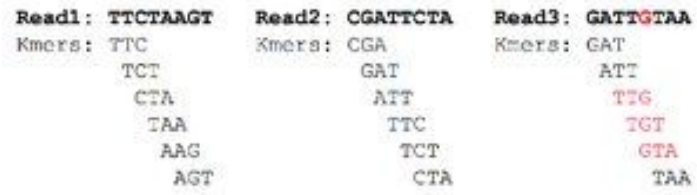
Step 2: Build graph based on k-mer overlap: Multiple possibilities lead to branching of the graph

Step 3: Each k-mer is counted.
Starting from the first k-mer of a sequence, it is read based on frequency of k-mers.

Ayling, M., Clark, M.D. and Leggett, R.M., 2020. New approaches for metagenome assembly with short reads. *Briefings in bioinformatics*, *21*(2), pp.584-594.

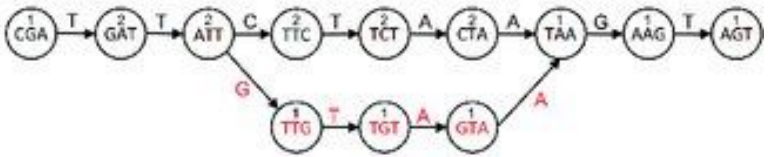
De Bruijn Graph Exercise

(i) Make kmers



k=3, sliding window along sequence

(ii) Build graph



Graph including weights (or frequency of kmer)

(iii) Walk graph and output contigs

CGATTCTAAGT

Final sequence using most abundant kmers

Ayling, M., Clark, M.D. and Leggett, R.M., 2020. New approaches for metagenome assembly with short reads. *Briefings in bioinformatics*, *21*(2), pp.584-594.

de Bruijn Graph Assembly Exercise

Reads: ATGCTA
GCTAGC
TAGCAC
GCACAT
ACATGC

Task: Fill out the k-mer matrix and draw the de Bruijn graph

Tip: Our sequence comes full circle.

k-mers:

	ATG	TGC	GCT	CTA	TAG	AGC	GCA	CAC	ACA	CAT
read1										
read2										
read3										
read4										
read5										
total										

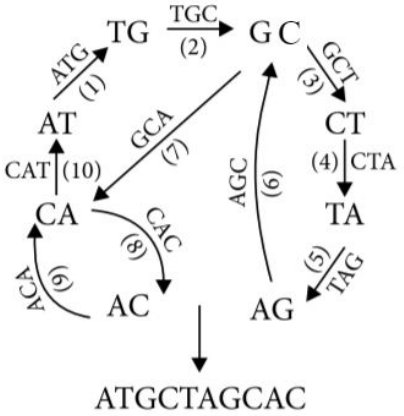
https://towardsdatascience.com/genome-assembly-using-de-bruijn-graphs-69570efcc270

de Bruijn Graph Assembly Solution

Reads: ATGCTA
GCTAGC
TAGCAC
GCACAT
ACATGC

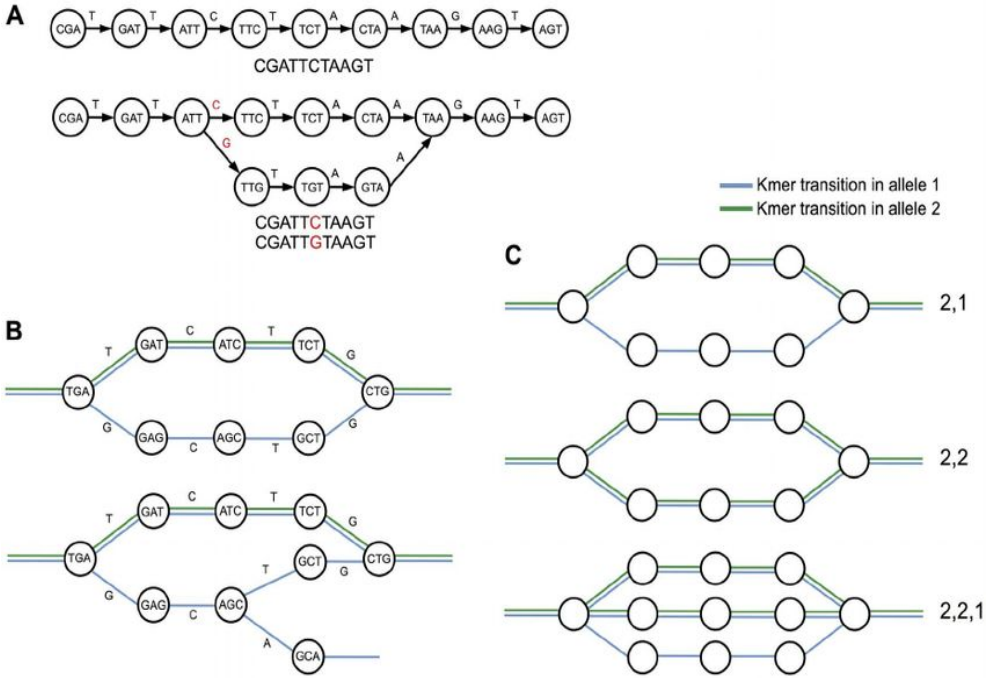
k-mers:

	ATG	TGC	GCT	CTA	TAG	AGC	GCA	CAC	ACA	CAT
read1	1	1	1	1						
read2			1	1	1	1				
read3					1	1	1	1		
read4							1	1	1	1
read5	1	1							1	1
total	2	2	2	2	2	2	2	2	2	2



https://towardsdatascience.com/genome-assembly-using-de-bruijn-graphs-69570efcc270

Bubbles in the de Bruijn graph



Causes for bubbles:

- Strain-level variation – Single nucleotide polymorphisms (SNPs)
- Unresolved repeats:
  - Transposable Elements
  - CRISPRs
  - rRNAs
  - …
- Erroneous reads

Leggett, R.M. and MacLean, D., 2014. Reference-free SNP detection: dealing with the data deluge. *Bmc Genomics*, *15*(4), pp.1-7.

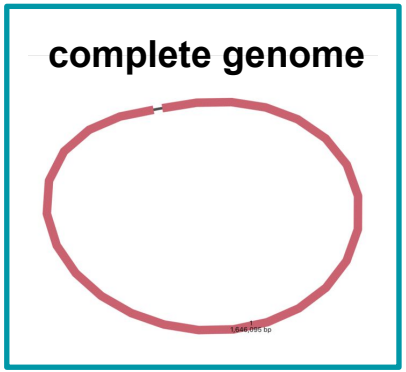
Bubbles in the de Bruijn graph

Methods to resolve bubbles:

- Increase of k-mer size
- Use of a combination of k-mer sizes:
- Typical k-mers sizes used in (meta-)Spades (assembly software):
    - k21 k33 k55 k77 k99 k127
    - Iterative evaluation of contigs

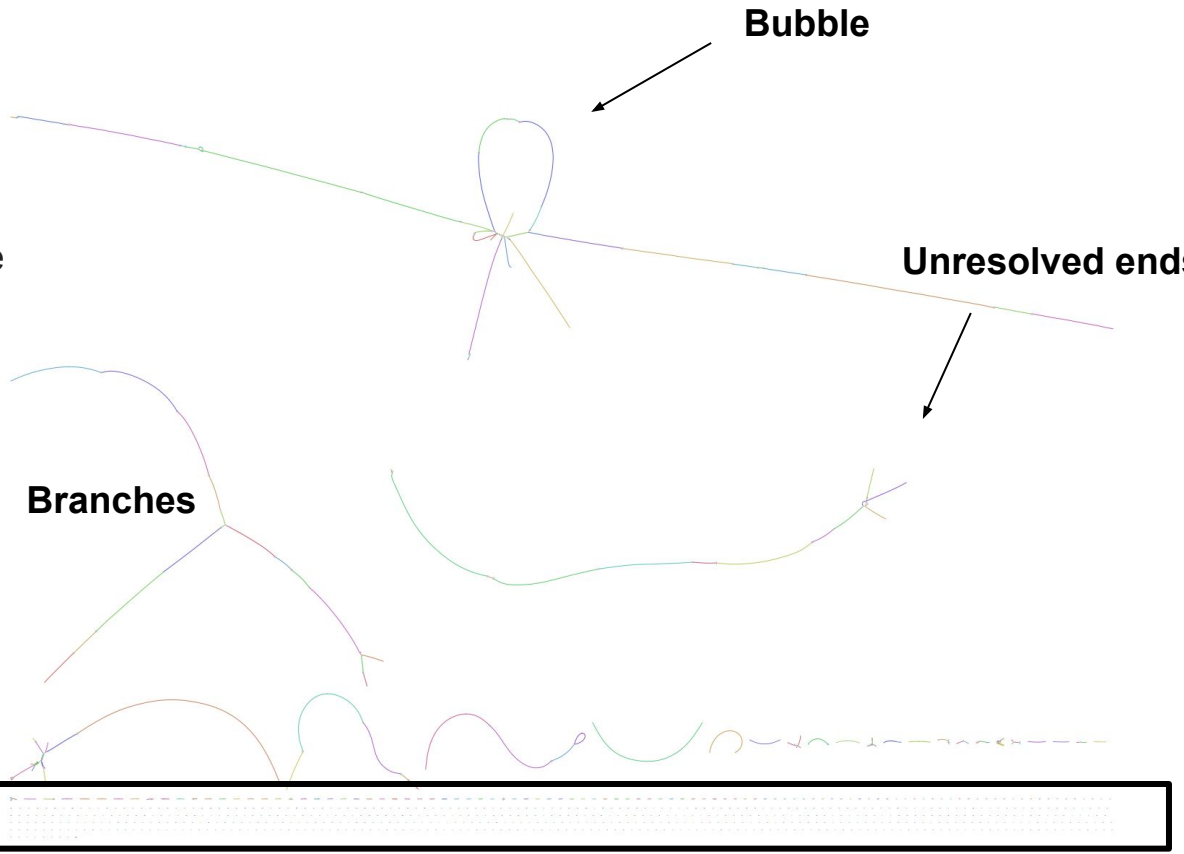
    □ Program chooses best assembly
- Additional long reads can resolve bridges
    - e.g. PacBio or Oxford Nanopore

    □ Hybrig assembly



What does an assembly look like?

- For a **metagenomic** assembly, we expect **many short contigs**
- Gapless assembly of multiple contigs forms genome **scaffolds**



Tool: Bandage by rrwick

Important Omics Vocabulary

- Sequencing and quality control:
  - **Intensity**: Light recorded from a base during sequencing
  - **Basecalling**: Process of turning light intensities (chromatogram) into a letter-sequence (GATACA)
  - **Read 1 & 2:** Forward and reverse read
  - **Fastq:** Common sequence file format including sample ID, sequence and quality scores
- Quality control:
  - **Phred Score:** Quality score assigned to a sequence
  - **Duplications:** Multiple sequences of the same original template (due to PCR cycling)
  - **Barcode:** Sequence identifying a sample (used for multiplexing, aka. pooling of samples in a run)
  - **Adapter:** Sequence linking DNA template and Sequencing-array
- Assembly/Co-Assembly:
  - **De Bruijn Graph:** Graph representing overlaps between sequences through k-mers
  - **K-mer:** Multimer of part of a sequence used for de Bruijn graph assembly
  - **Bubbles and/or bridges:** Alternative options in de Bruijn graph assembly due to SNPs
  - **SNPs:** Single nucleotide polymorphisms
  - **Contigs:** Contiguous sequence of DNA assembled from short or long reads
  - **Scaffolds:** Longer sequence assembled (gap-free) from multiple contigs

https://anvio.org/vocabulary/

Lecture resources

- Ayling, M., Clark, M.D. and Leggett, R.M., 2020. New approaches for metagenome assembly with short reads. Briefings in bioinformatics, 21(2), pp.584-594.

- Leggett, R.M. and MacLean, D., 2014. Reference-free SNP detection: dealing with the data deluge. *Bmc Genomics*, *15*(4), pp.1-7.

- Introduction to metagenomics: https://training.galaxyproject.org/trainingmaterial/topics/metagenomics/slides/introduction.html#1

- Video: de Bruijn graph assembly for DNA sequences (RobEdwards), https://www.youtube.com/watch?v=zmZvlNglAU0

- Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G., 2013. **QUAST**: quality assessment tool for genome assemblies. Bioinformatics, 29(8), pp.1072-1075.

- http://merenlab.org/momics
- https://anvio.org/vocabulary/

Metric to asses assembly quality

- No. of contigs
- Largest contig
- Total Length: makes sense in the context of genome assembly
- N50: defines assembly quality in terms of contiguity
- L50: count of smallest number of contigs whose length sum makes up half of genome size
- GC content (%): n(G+C)/length of the assembly

For a full overview of quality metrics check out this reference:

Gurevich, A., Saveliev, V., Vyahhi, N. and Tesler, G., 2013. **QUAST**: quality assessment tool for genome assemblies. Bioinformatics, 29(8), pp.1072-1075.

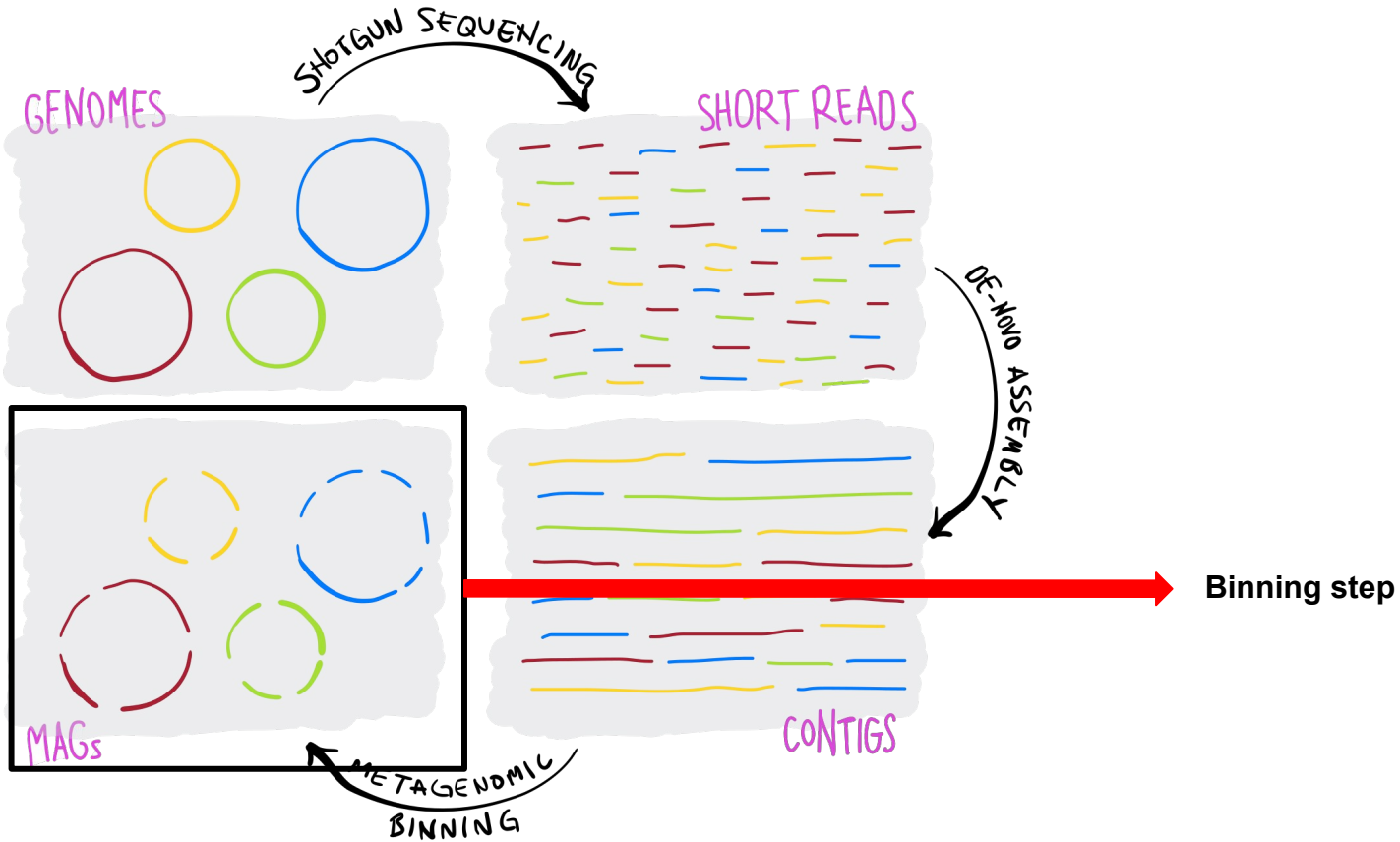
Any

From contigs to bins and quality assessment

Biol-217
22 January - 2 February 2024

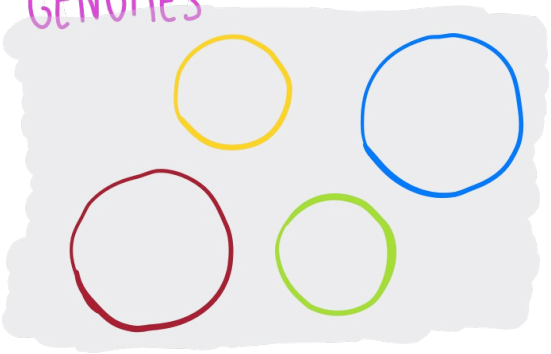
Dr. Cynthia M. Chibani

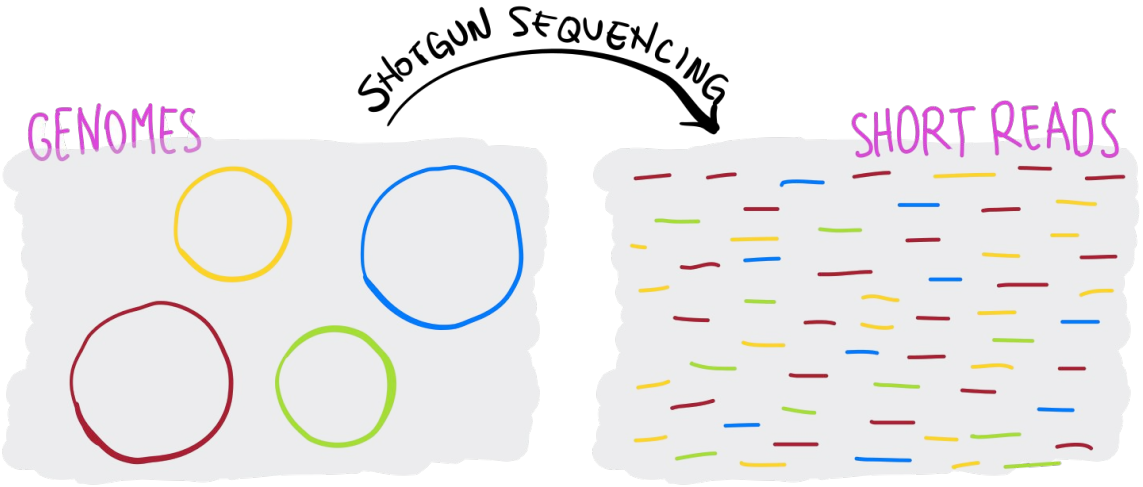
Binning

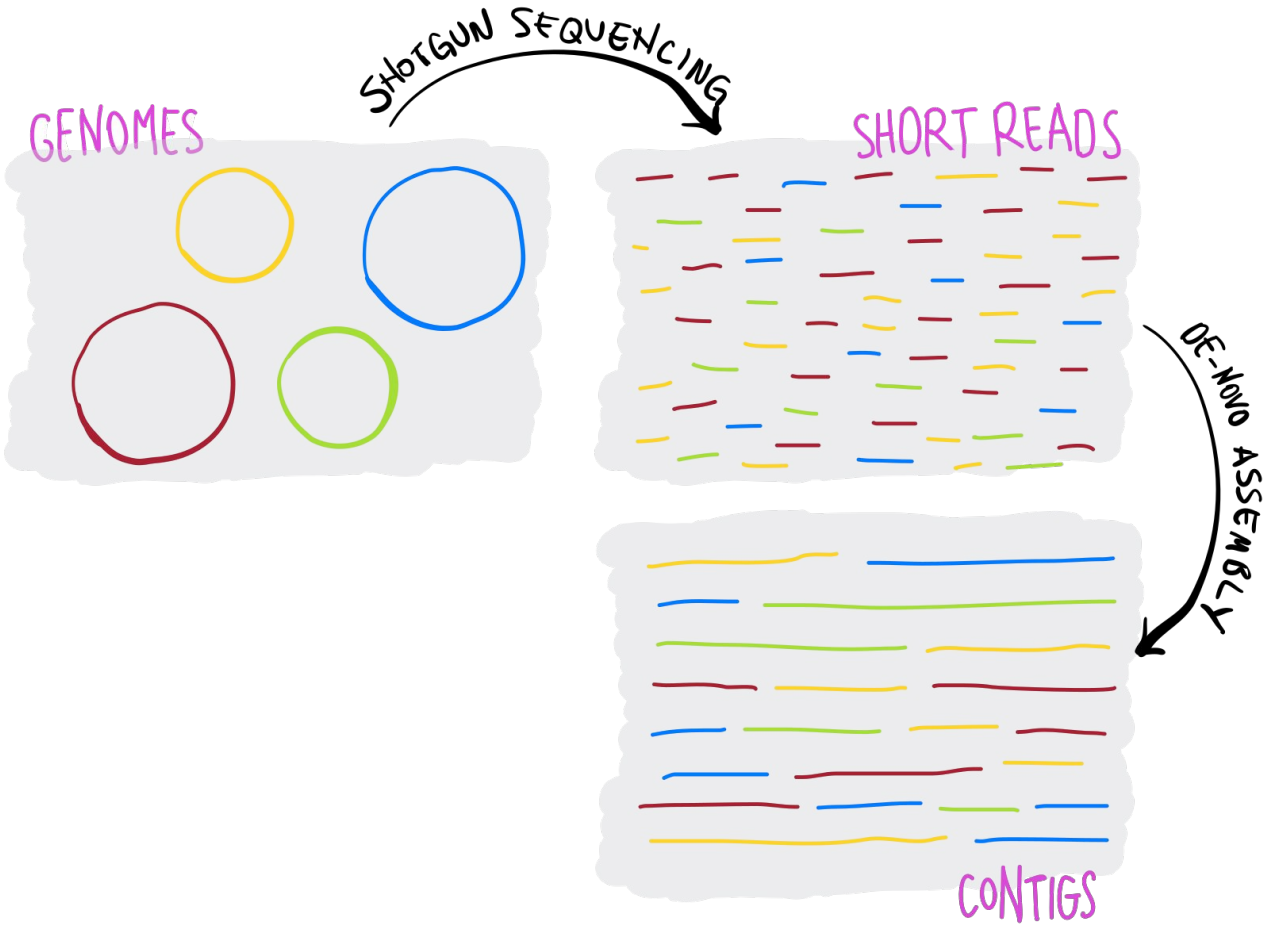


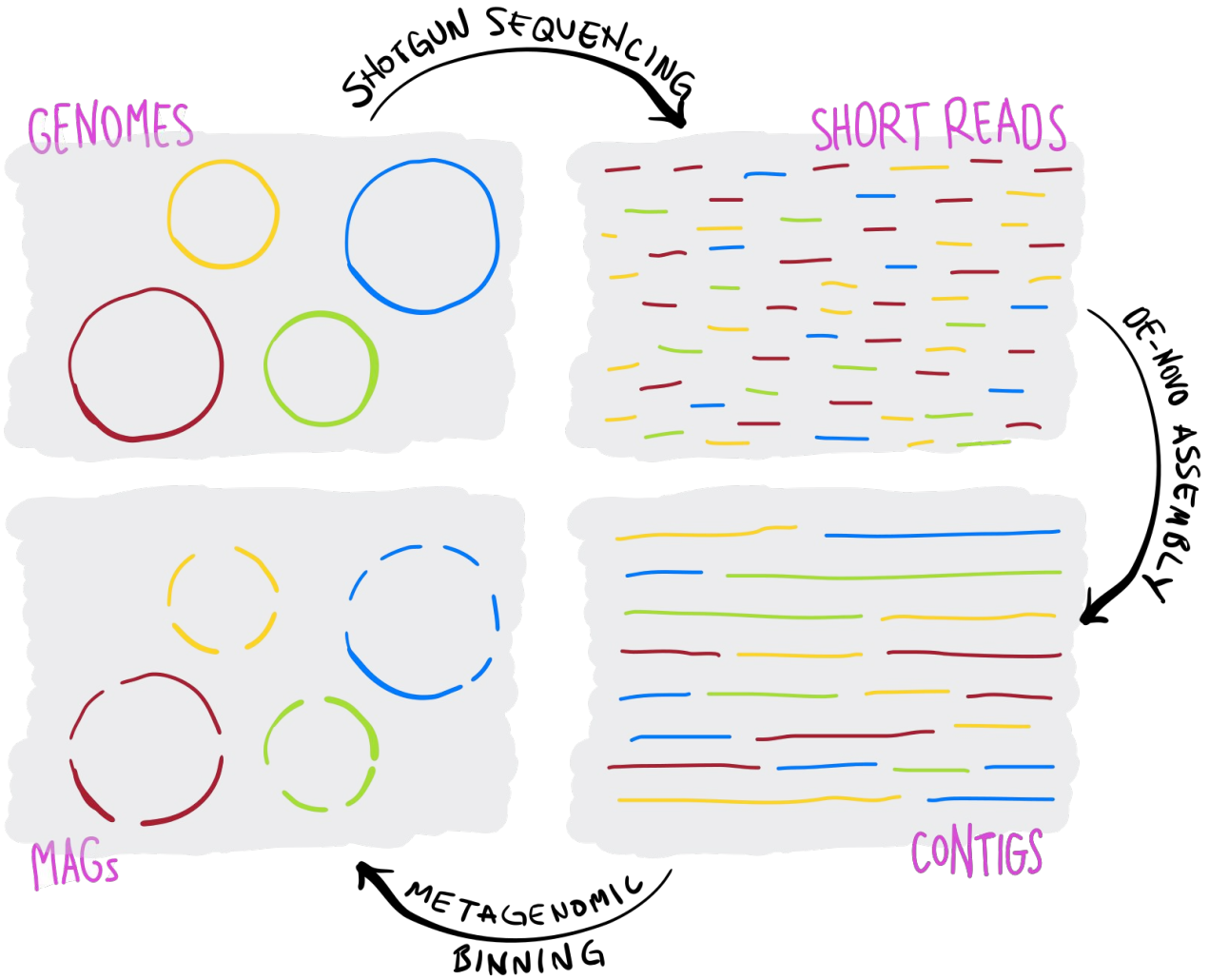
http://merenlab.org/momics

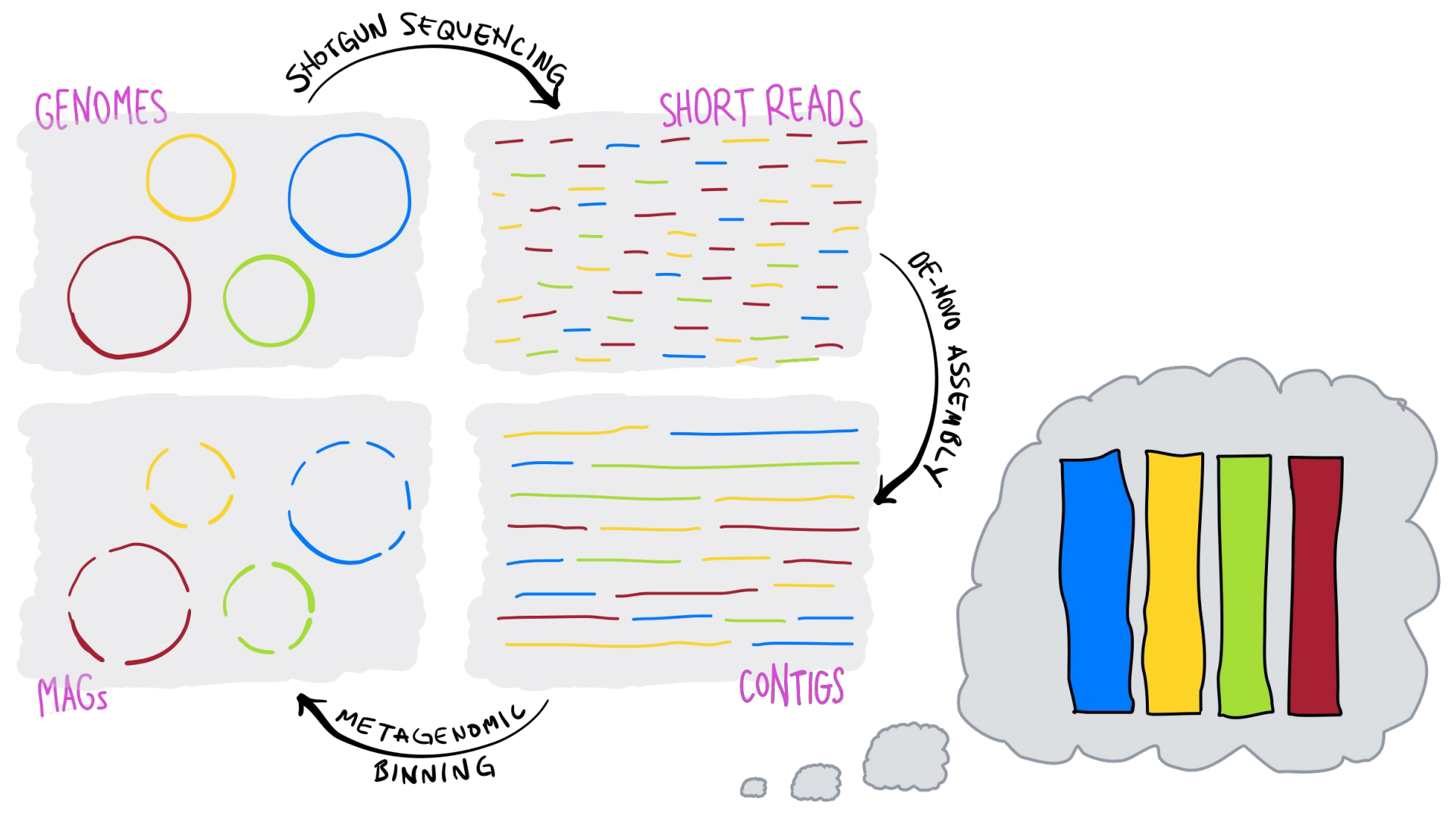
GENOMES



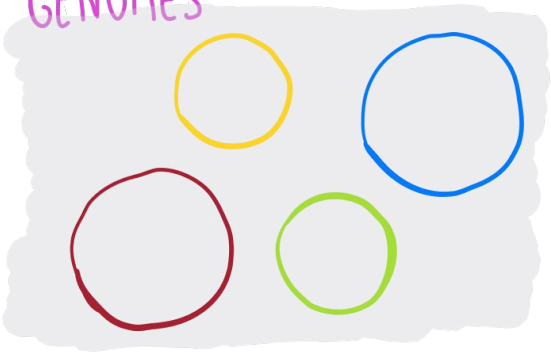

SHOTGUN SEQUENCING
GENOMES
SHORT READS

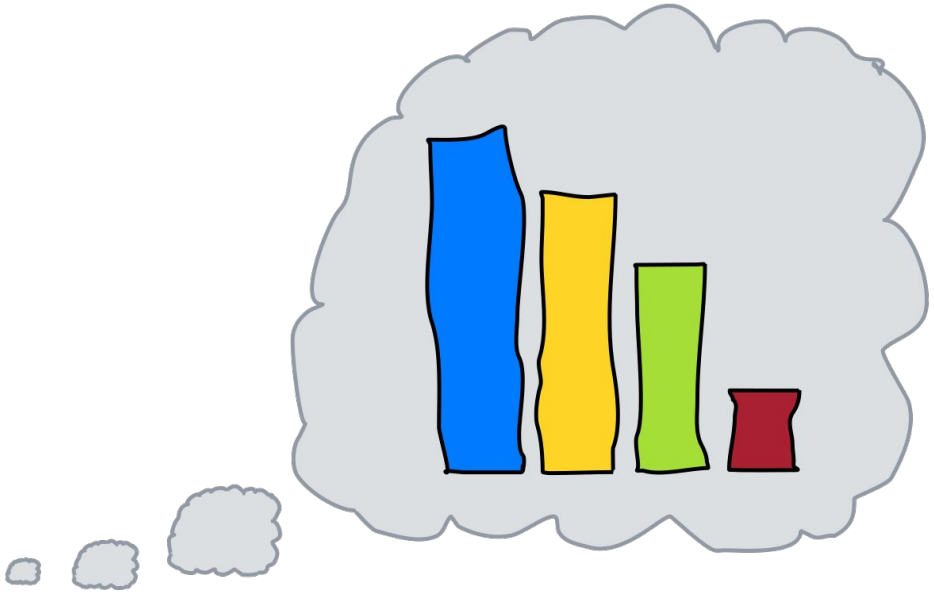

SHOTGUN SEQUENCING
GENOMES
SHORT READS
DE-NOVO ASSEMBLY
CONTIGS

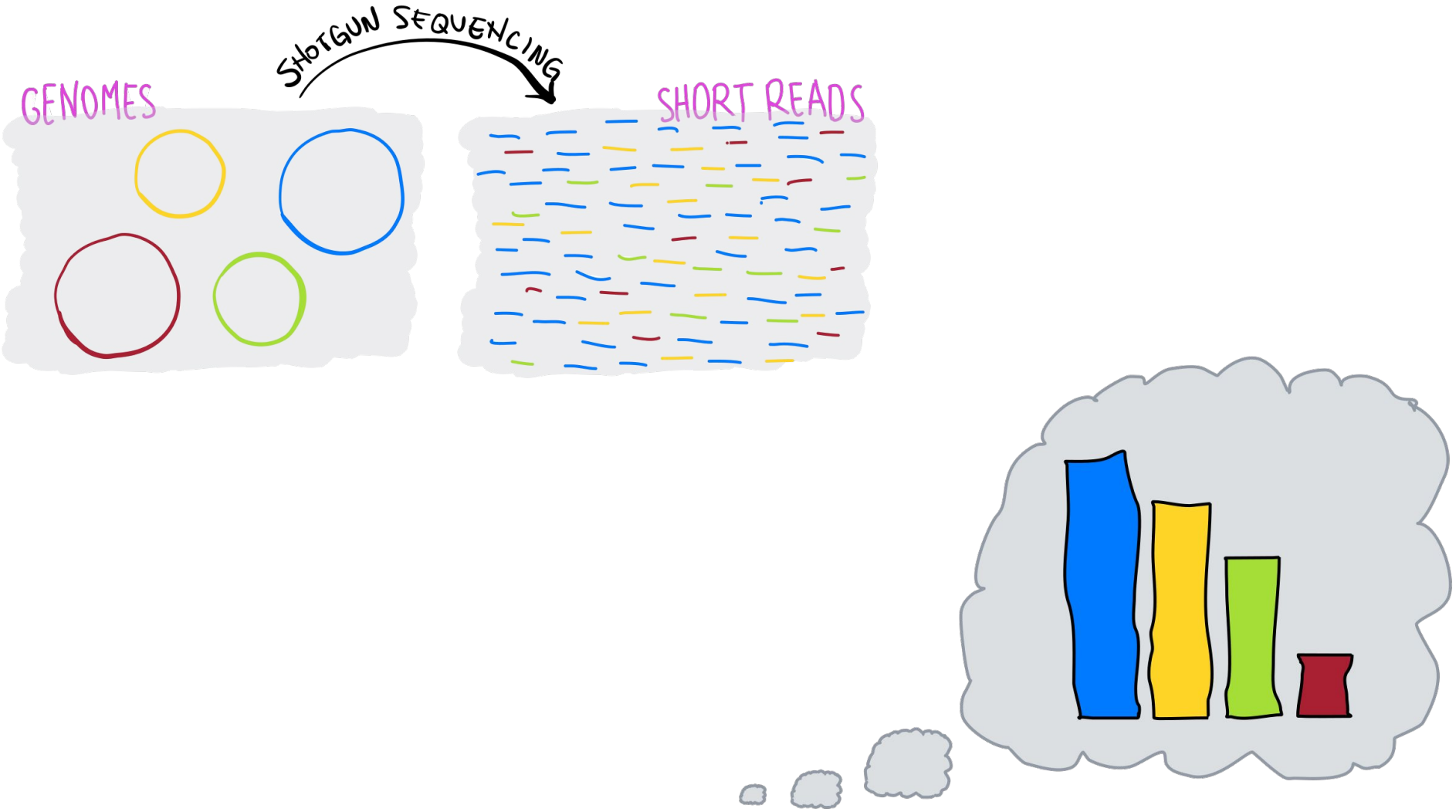

SHOTGUN SEQUENCING
GENOMES
SHORT READS
DE-NOVO ASSEMBLY
MAGs
CONTIGS
METAGENOMIC
BINNING

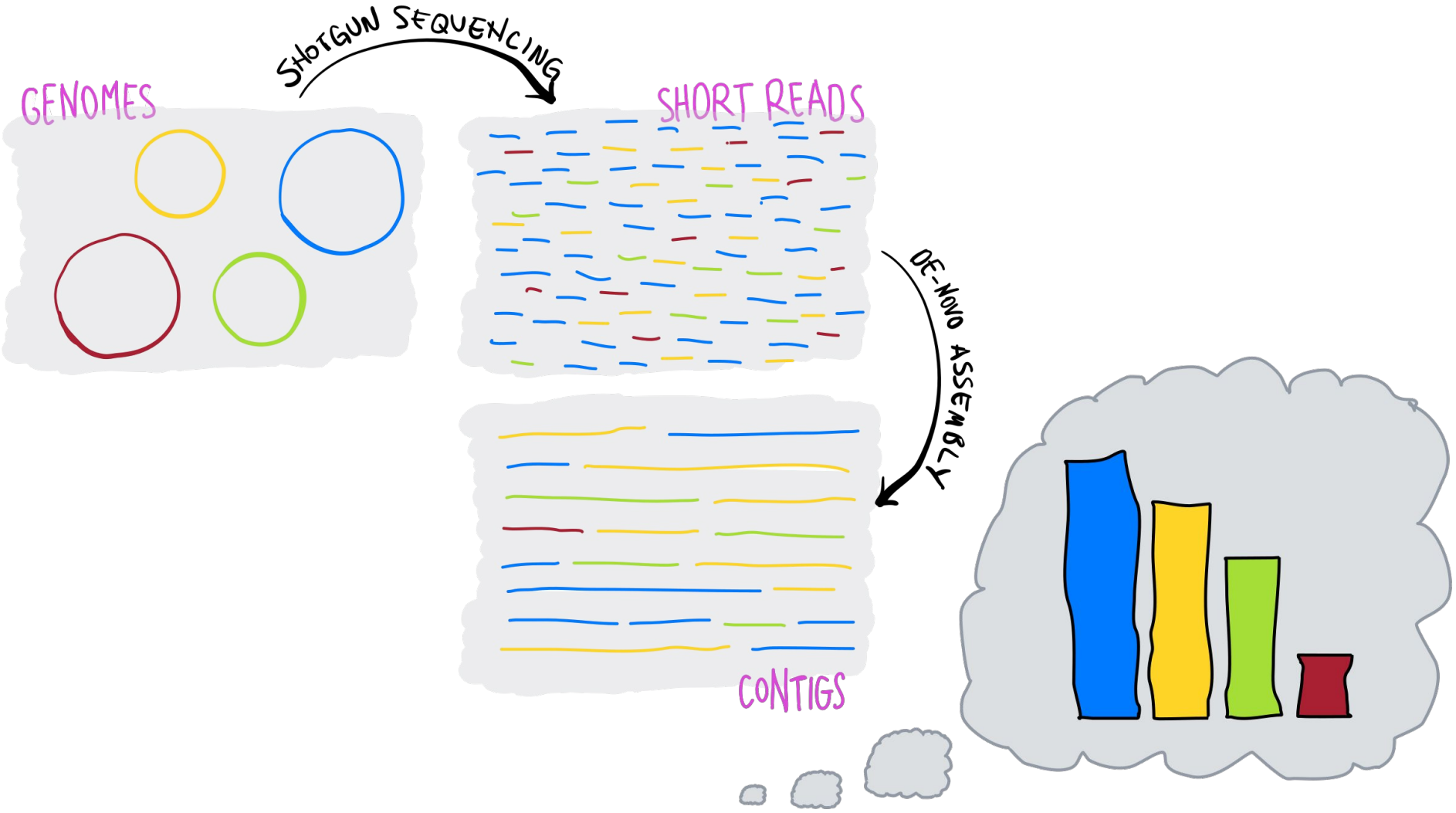

SHOTGUN SEQUENCING
GENOMES
SHORT READS
DE-NOVO ASSEMBLY
MAGs
CONTIGS
METAGENOMIC
BINNING

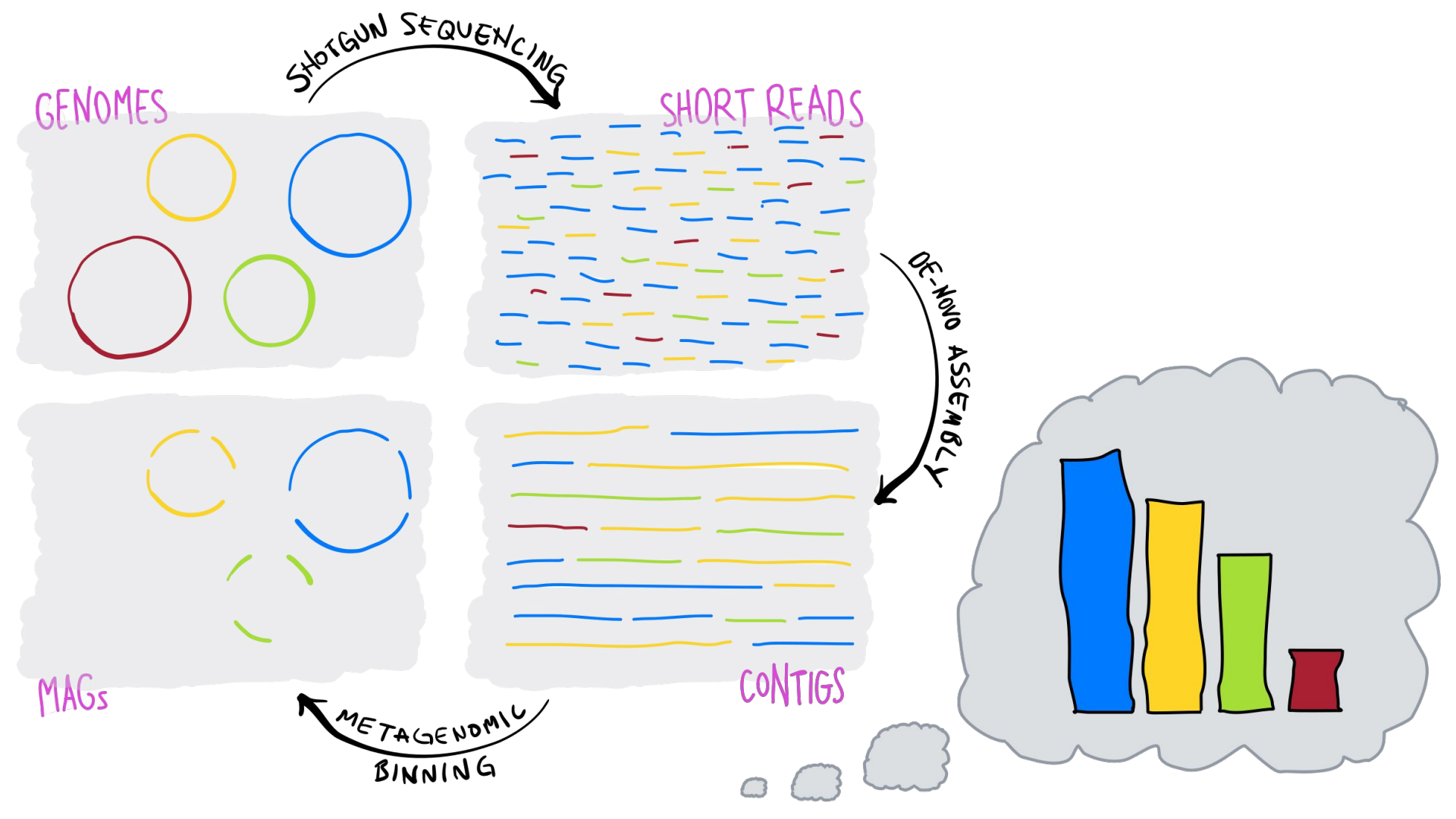
GENOMES

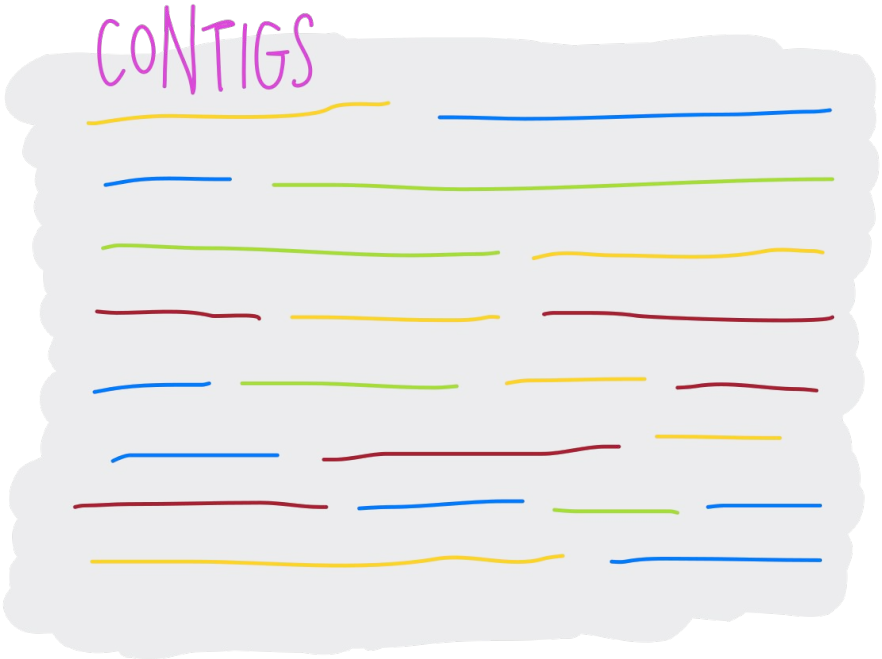


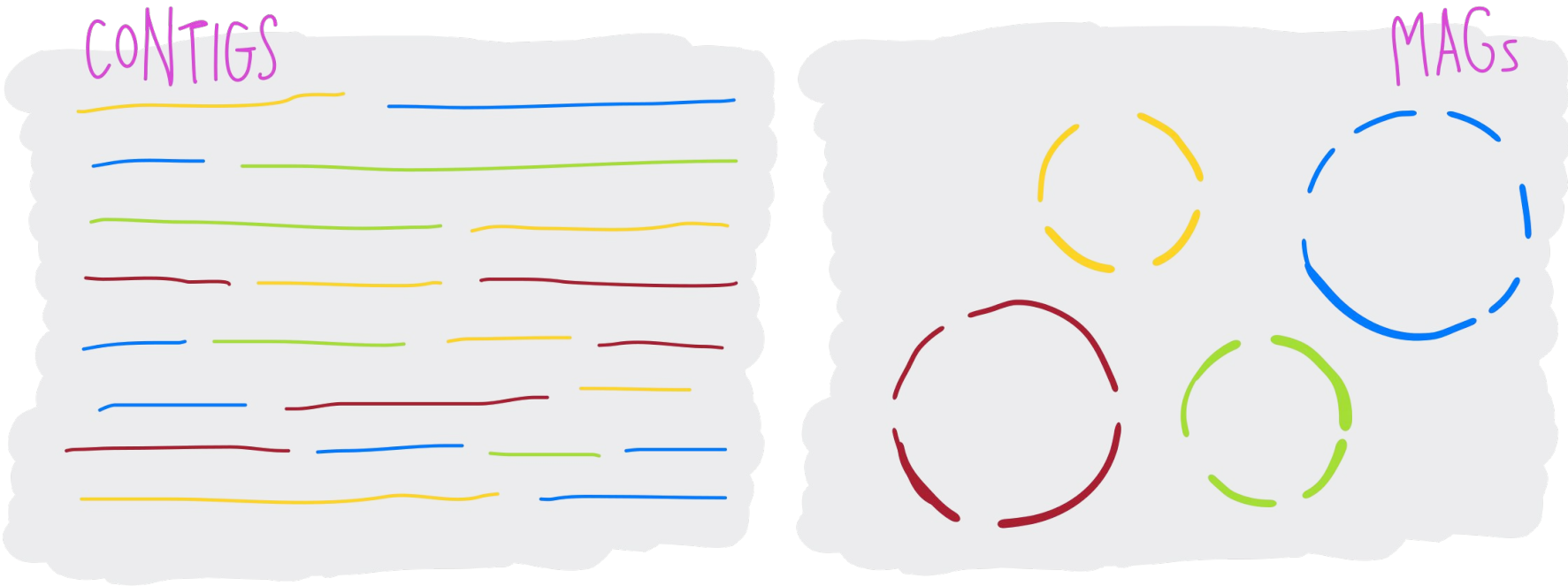



SHOTGUN SEQUENCING
GENOMES
SHORT READS

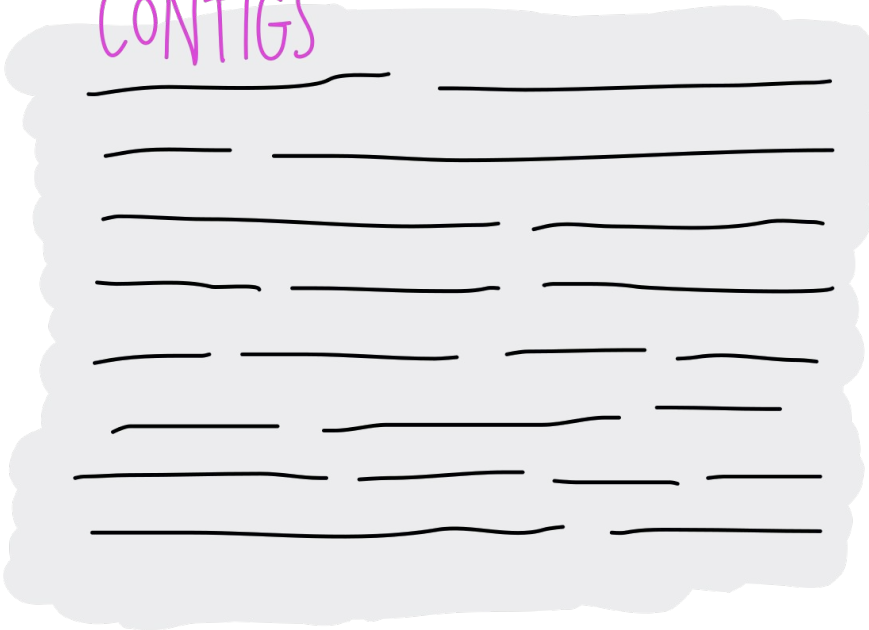

SHOTGUN SEQUENCING
GENOMES
SHORT READS
DE-NOVO ASSEMBLY
CONTIGS


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

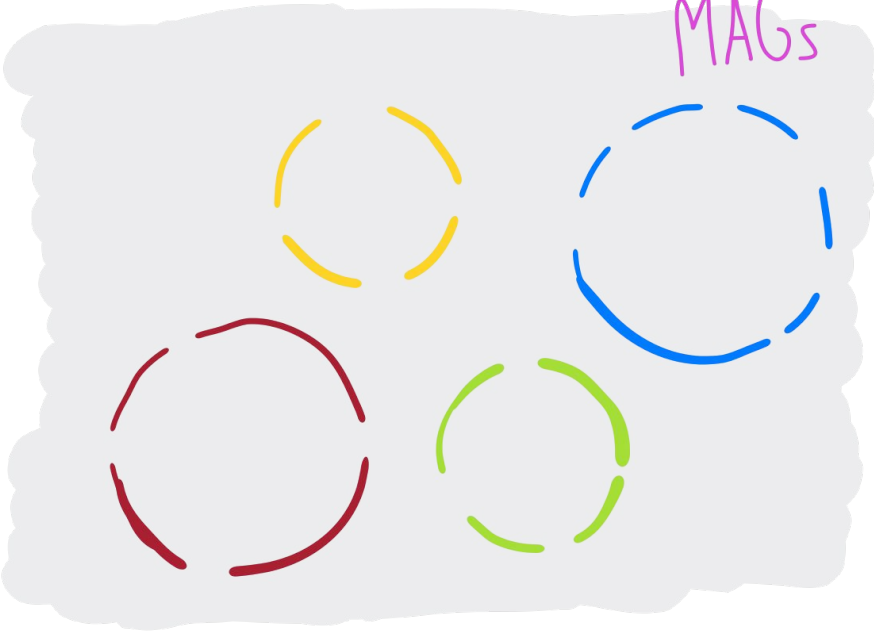

CONTIGS

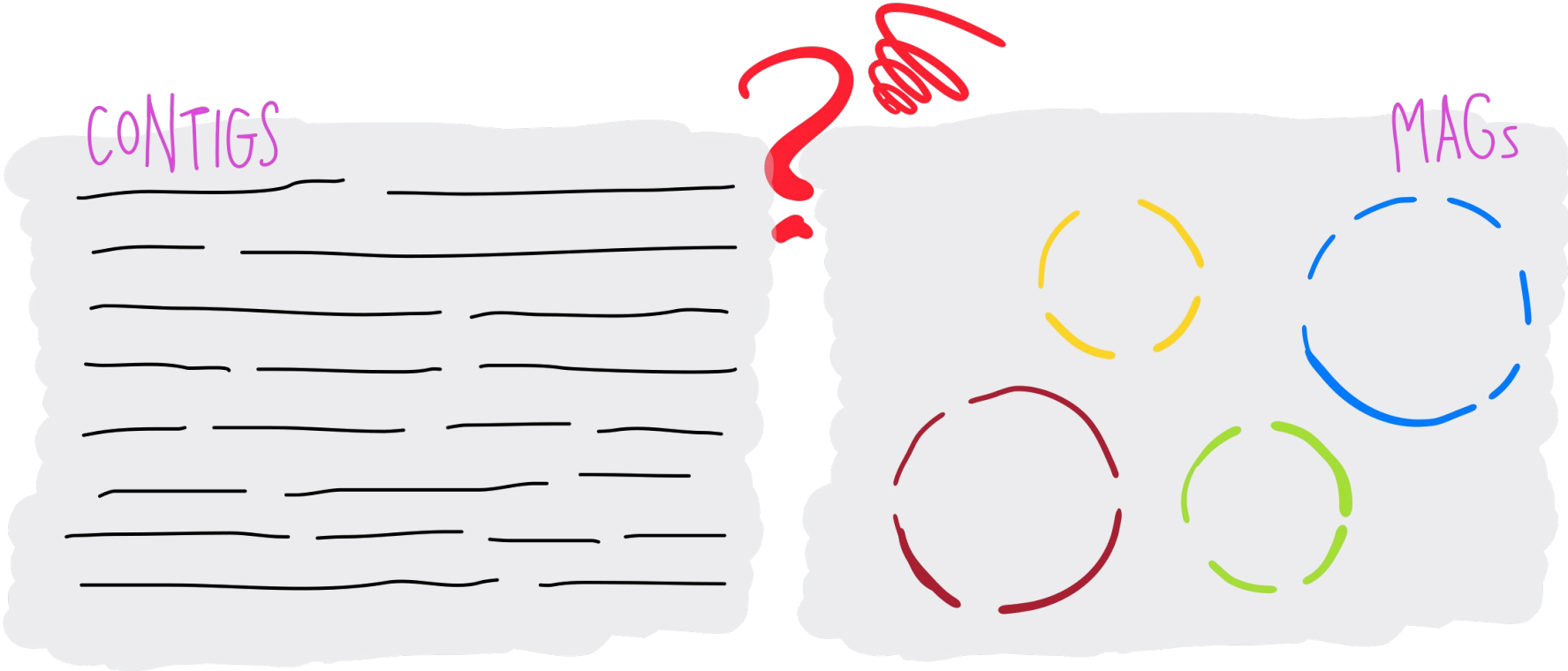

CONTIGS
MAGs

CONTIGS



MAGs




CONTIGS
MAGs

Community structure and metabolism through reconstruction of microbial genomes from the environment

Gene W. Tyson, Jarrod Chapman, Philip Hugenholtz, Eric E. Allen, Rachna J. Ram, Paul M. Richardson, Victor V. Solovyev, Edward M. Rubin, Daniel S. Rokhsar & Jillian F. Banfield

A new view of the tree of life

Laura A. Hug, Brett J. Baker, Karthik Anantharaman, Christopher T. Brown, Alexander J. Probst, Cindy J. Castelle, Cristina N. Butterfield, Alex W. Hernsdorf, Yuki Amano, Kotaro Ise, Yohey Suzuki, Natasha Dudek, David A. Relman, Kari M. Finstad, Ronald Amundson, Brian C. Thomas & Jillian F. Banfield

Unusual biology across a group comprising more than 15% of domain Bacteria

Christopher T. Brown, Laura A. Hug, Brian C. Thomas, Itai Sharon, Cindy J. Castelle, Andrea Singh, Michael J. Wilkins, Kelly C. Wrighton, Kenneth H. Williams & Jillian F. Banfield

Population Genomic Analysis of Strain Variation in *Leptospirillum* Group II Bacteria Involved in Acid Mine Drainage Formation

Sheri L Simmons, Genevieve DiBartolo, Vincent J Denef, Daniela S. Aliaga Goltsman, Michael P Thelen, Jillian F Banfield

Jill Banfield, University of California Berkeley

THE JOURNAL OF BIOLOGICAL CHEMISTRY
Vol. 236, No. 3, March 1961
Printed in U.S.A.

Enzymatic Synthesis of Deoxyribonucleic Acid

VIII. FREQUENCIES OF NEAREST NEIGHBOR BASE SEQUENCES IN DEOXYRIBONUCLEIC ACID

JOHN JOSSE,* A. D. KAISER, AND ARTHUR KORNBERG

From the Department of Biochemistry, Stanford University School of Medicine, Palo Alto, California

(Received for publication, October 4, 1960)

Determination of deoxyribonucleotide sequence in a deoxyribonucleic acid molecule is important from both the chemical and genetic points of view. It is also essential for answering the question of whether DNA synthesized *in vitro* by polymerase (1, 2) is a faithful copy of the nucleotide[1] sequence of the primer DNA. Although enzymatically synthesized DNA has the same over-all nucleotide composition as the particular primer DNA (3), it could not be inferred that this synthesis is a replication of the nucleotide sequences of the primer.

Because of the limitations of present methods, complete se-

(8). All of the labeled substrates contained P^{32} in the phosphate esterified to the sugar; they were prepared as described previously (1). The DNA-synthesizing enzyme was prepared from the polymerase, Fraction VII, described elsewhere (1); this enzyme was refractionated with diethylaminoethyl cellulose, yielding a preparation with a specific activity of 500 units per mg of protein. Micrococcal DNase was prepared according to Cunningham *et al.* (9); the final fraction had a specific activity of 7500 units per mg of protein.[2] Calf spleen phosphodiesterase was isolated by Hilmoe's procedure (10); the purified preparation

In the studies to be reported here, we have derived the frequencies of the 16 possible nearest neighbor pairs in a variety of DNA's by the technique of enzymatic incorporation of 5′-P^{32}-labeled nucleotides into DNA and then degradation of the DNA into 3′-nucleotides. Briefly, we have found that: (*a*) each DNA directs the synthesis of a product which has a unique and non-random pattern of the 16 nearest neighbor frequencies; (*b*) the DNA synthesized has the same nearest neighbor frequencies whether the primer is native DNA or enzymatically prepared DNA containing only traces of the original native DNA; and (*c*) the pattern of nearest neighbor frequencies in every case involves both base-pairing of adenine to thymine and of guanine to cytosine between sister strands of DNA, and opposite "polarity" of the two strands as proposed in the Watson and Crick model (7).

A bit of history

DNA RESEARCH 5, 251–259 (1998)

Genes from Nine Genomes Are Separated into Their Organisms in the Dinucleotide Composition Space

Hiroshi NAKASHIMA,[1,*] Motonori OTA,[2] Ken NISHIKAWA,[2] and Tatsuo OOI[3]

School of Health Sciences, Faculty of Medicine, Kanazawa University, 5-11-80 Kodatsuno, Kanazawa 920-0942, Japan,[1] Center for Information Biology, National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, Japan,[2] and Kyoto Women's University, Kitahiyoshi-cho 35, Higashiyama-ku, Kyoto 605, Japan[3]

(Received 2 September 1998)

Abstract

A set of 16 kinds of dinucleotide compositions was used to analyze the protein-encoding nucleotide sequences in nine complete genomes: *Escherichia coli*, *Haemophilus influenzae*, *Helicobacter pylori*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Synechocystis* sp., *Methanococcus jannaschii*, *Archaeoglobus fulgidus*, and *Saccharomyces cerevisiae*. The dinucleotide composition was significantly different between the organisms. The distribution of genes from an organism was clustered around its center in the dinucleotide composition space. The genes from closely related organisms such as Gram-negative bacteria, mycoplasma species and eukaryotes showed some overlap in the space. The genes from nine complete genomes together with those from human were discriminated into respective clusters with 80% accuracy using the dinucleotide composition alone. The composition data estimated from a whole genome was close to that obtained from genes, indicating that the characteristic feature of dinucleotides holds not only for protein coding regions but also noncoding regions. When a dendrogram was constructed from the disposition of the clusters in the dinucleotide space, it resembled the real phylogenetic tree. Thus, the distinct feature observed in the dinucleotide composition may reflect the phylogenetic relationship of organisms.

Key words: separation of genes; dinucleotide frequency; phylogenetic tree

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

GTTTTGGCATGATAAGGAGTTTCTTTTGTGCTTC

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	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	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

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

ACTTGCGCAGTCGCGCATTACGCGTAGTGGAATAA

	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										
K										
M										

k=2

GGAGCGTTTTATTTAGTACCGGTTTTGAAGTTAAC

	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										
M										

k=2

GCCGCGAGCGGCCCGCGCCGGCTTCGGCGCCGCAC

	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										

k=2

GGGCCTGCGCCGGTCCAGTCACCCGGCTGCGACCT

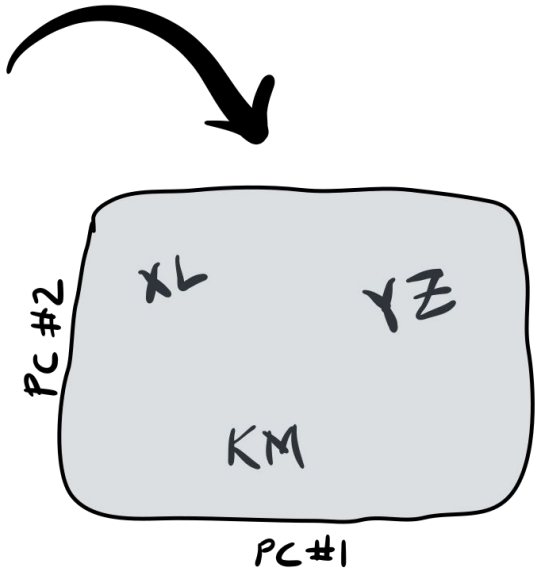
	AA	AC	AG	GA	CA	CC	CG	GC	AT	TA
X	11	3	4	4	5	2	0	2	2	1
Y	4	5	2	4	5	4	4	3	2	1
Z	4	5	3	2	4	1	5	5	2	3
L	11	6	3	2	2	3	2	1	1	4
K	1	1	2	2	1	8	9	10	0	0
M	0	4	4	3	4	10	4	5	0	0

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

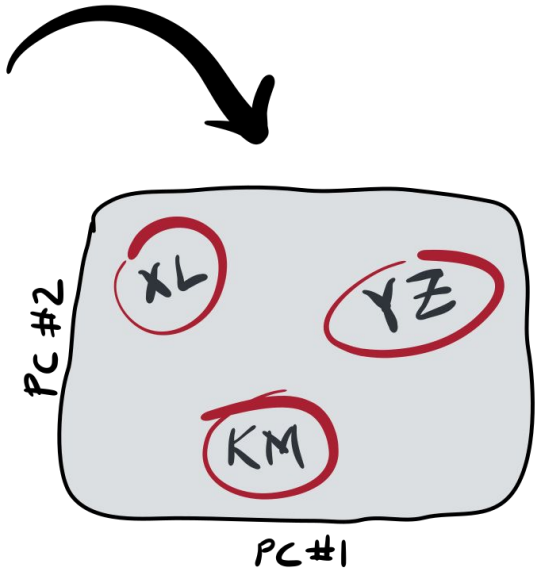
k=2

	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

A bit of history

DNA RESEARCH 5, 251–259 (1998)

Genes from Nine Genomes Are Separated into Their Organisms in the Dinucleotide Composition Space

Hiroshi NAKASHIMA,[1,*] Motonori OTA,[2] Ken NISHIKAWA,[2] and Tatsuo OOI[3]

School of Health Sciences, Faculty of Medicine, Kanazawa University, 5-11-80 Kodatsuno, Kanazawa 920-0942, Japan,[1] Center for Information Biology, National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, Japan,[2] and Kyoto Women's University, Kitahiyoshi-cho 35, Higashiyama-ku, Kyoto 605, Japan[3]

(Received 2 September 1998)

Abstract

A set of 16 kinds of dinucleotide compositions was used to analyze the protein-encoding nucleotide sequences in nine complete genomes: *Escherichia coli*, *Haemophilus influenzae*, *Helicobacter pylori*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Synechocystis* sp., *Methanococcus jannaschii*, *Archaeoglobus fulgidus*, and *Saccharomyces cerevisiae*. The dinucleotide composition was significantly different between the organisms. The distribution of genes from an organism was clustered around its center in the dinucleotide composition space. The genes from closely related organisms such as Gram-negative bacteria, mycoplasma species and eukaryotes showed some overlap in the space. The genes from nine complete genomes together with those from human were discriminated into respective clusters with 80% accuracy using the dinucleotide composition alone. The composition data estimated from a whole genome was close to that obtained from genes, indicating that the characteristic feature of dinucleotides holds not only for protein coding regions but also noncoding regions. When a dendrogram was constructed from the disposition of the clusters in the dinucleotide space, it resembled the real phylogenetic tree. Thus, the distinct feature observed in the dinucleotide composition may reflect the phylogenetic relationship of organisms.

Key words: separation of genes; dinucleotide frequency; phylogenetic tree

A bit of history

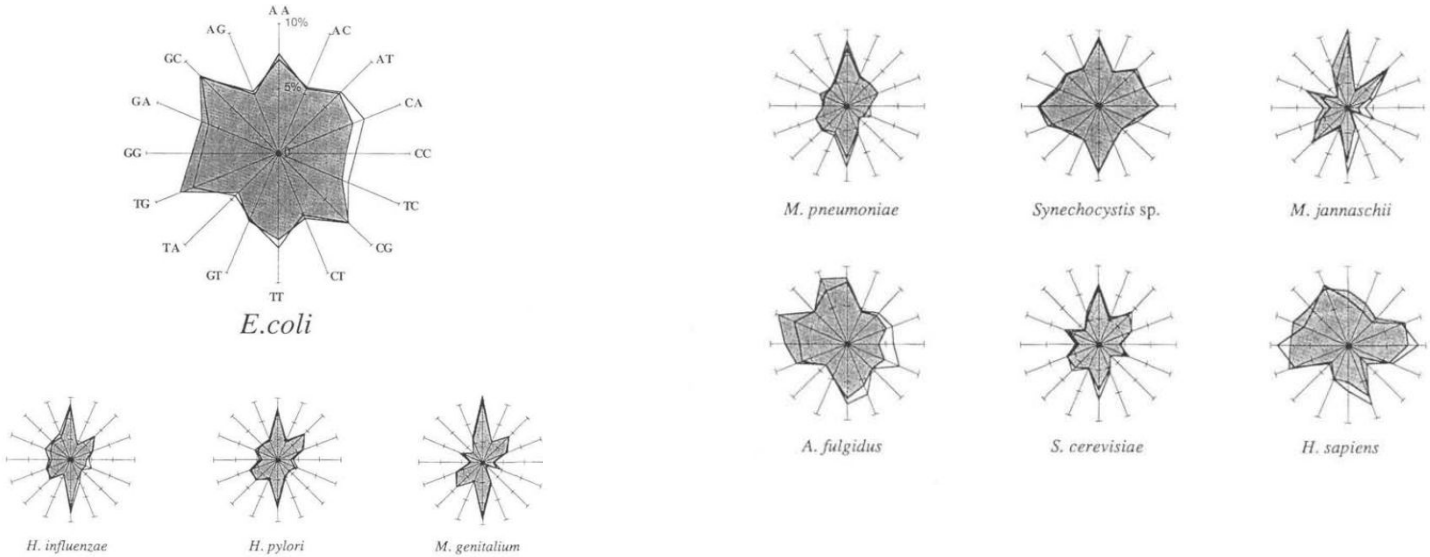


Figure 1. Star-diagrams presenting the dinucleotide composition. The mean compositions (%) over all genes (shaded) and the entire genome (non-shaded) are plotted. The data for human are exceptions (see Sequence data section), depicted here as references. The radial axes of 16 dinucleotides are allotted so that the complementary dinucleotides, AA/TT, AC/GT, etc., occupy counter positions along the circle. Complementary pairs should have equivalent amounts in the total composition over a genome. Note that the scale is different for each diagram. The innermost broken circle indicates the 5% level.

A bit of history

Peter A. Noble[1]
Robert W. Citek[2]
Oladele A. Ogunseitan[3]

[1]**Belle W. Baruch Institute for Marine Biology and Coastal Research, University of South Carolina, Columbia, SC, USA**
[2]**Department of Soil and Environmental Science, University of California at Riverside, Riverside, CA, USA**
[3]**Department of Environmental Analysis and Design, University of California at Irvine, Irvine, CA, USA**

Tetranucleotide frequencies in microbial genomes

A computational strategy for determining the variability of long DNA sequences in microbial genomes is described. Composite portraits of bacterial genomes were obtained by computing tetranucleotide frequencies of sections of genomic DNA, converting the frequencies to color images and arranging the images according to their genetic position. The resulting images revealed that the tetranucleotide frequencies of genomic DNA sequences are highly conserved. Sections that were visibly different from those of the rest of the genome contained ribosomal RNA, bacteriophage, or undefined coding regions and had corresponding differences in the variances of tetranucleotide frequencies and GC content. Comparison of nine completely sequenced bacterial genomes showed that there was a nonlinear relationship between variances of the tetranucleotide frequencies and GC content, with the highest variances occurring in DNA sequences with low GC contents (less than 0.30 mol). High variances were also observed in DNA sequences having high GC contents (greater than 0.60 mol), but to a much lesser extent than DNA sequences having low GC contents. Differences in the tetranucleotide frequencies may be due to the mechanisms of intercellular genetic exchange and/or processes involved in maintaining intracellular genetic stability. Identification of sections that were different from those of the rest of the genome may provide information on the evolution and plasticity of bacterial genomes.

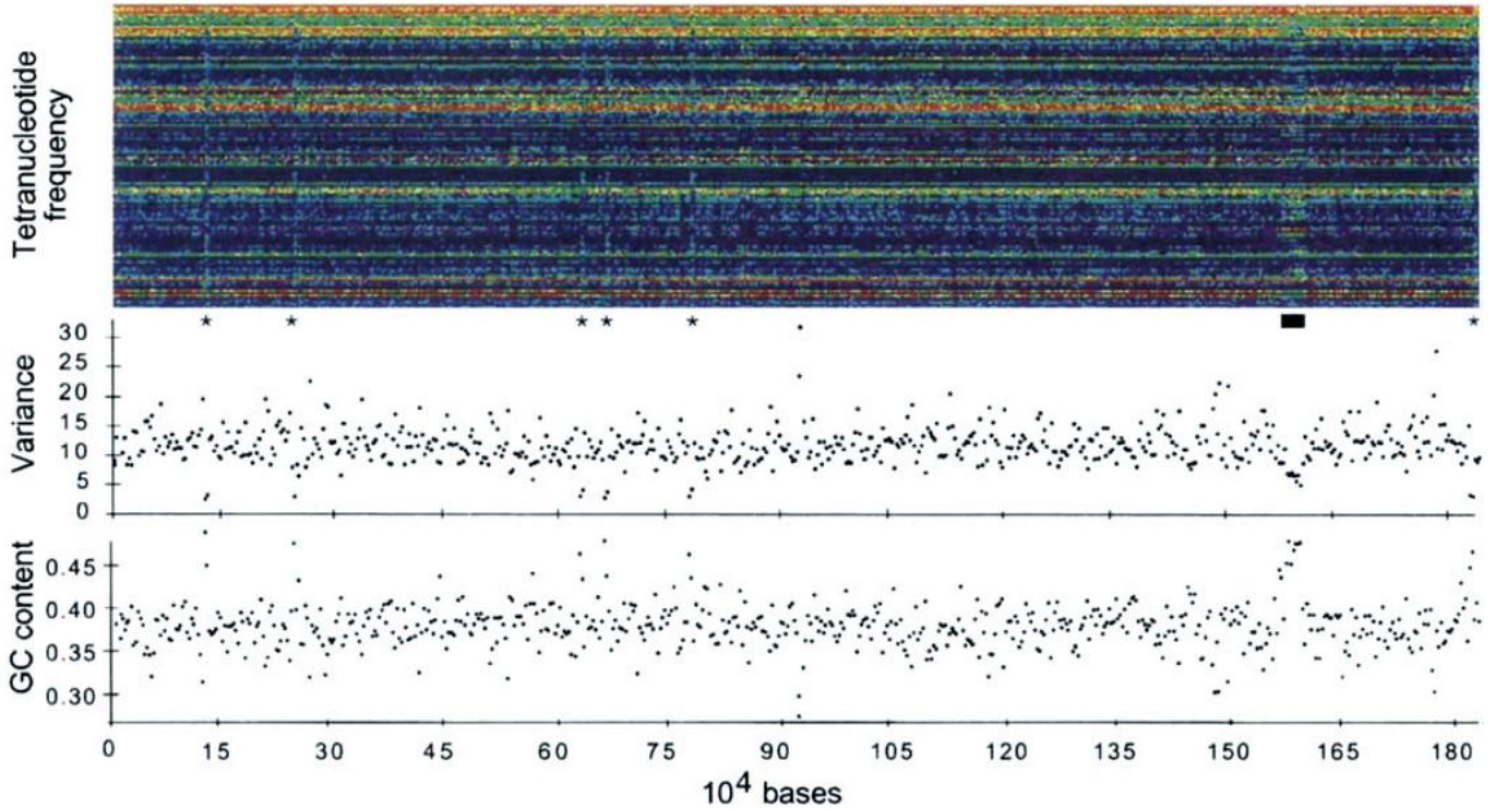
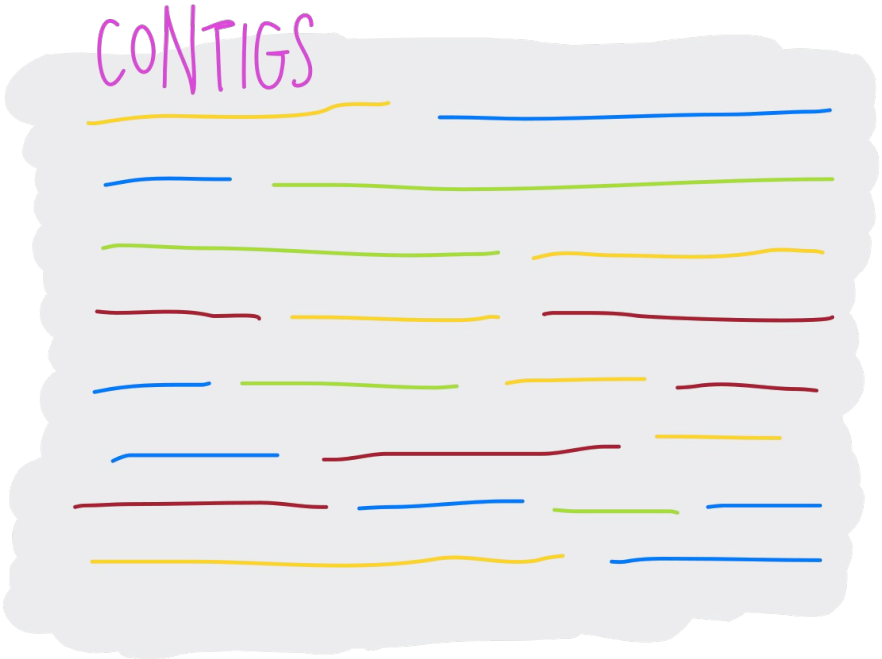
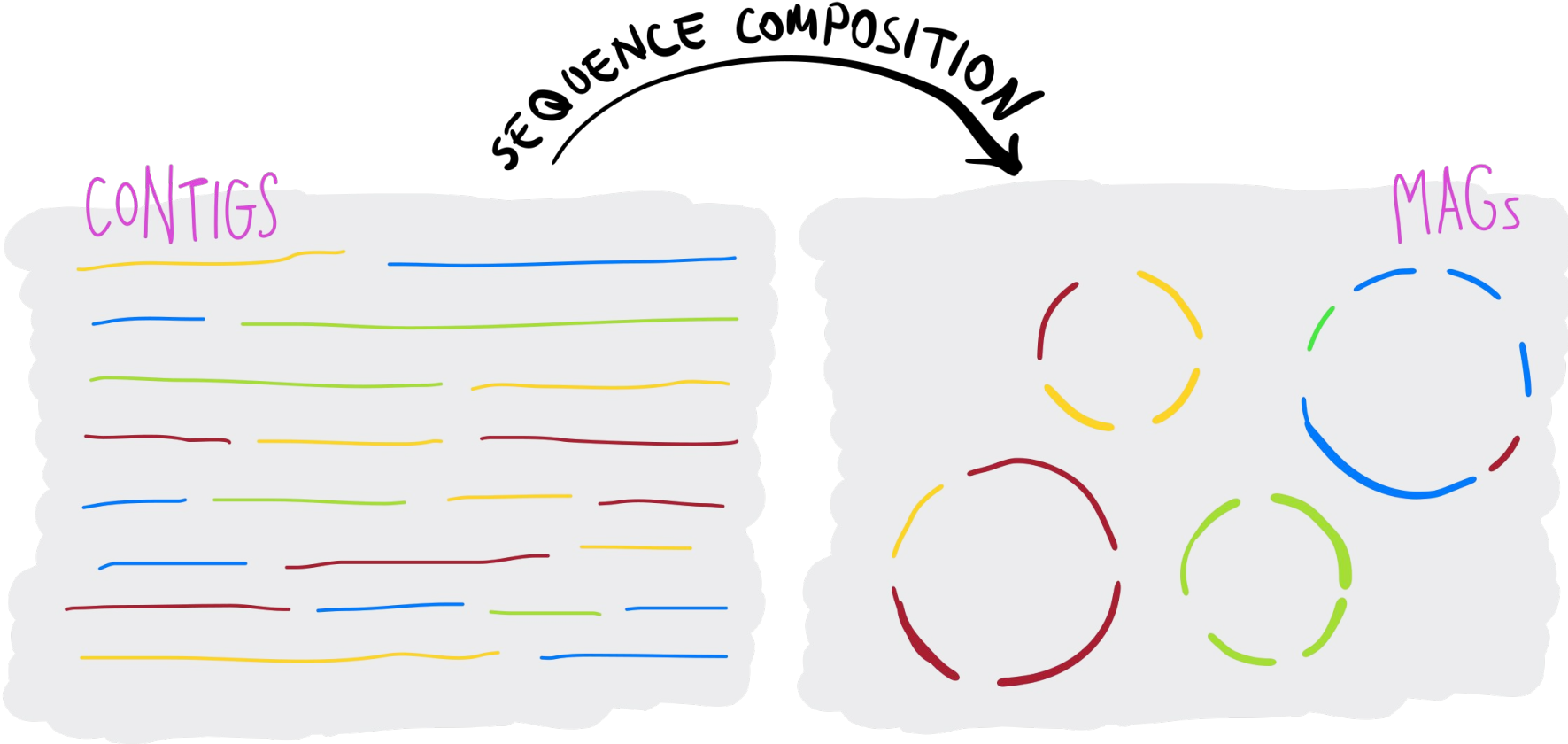


Figure 1. Fingerprints, variances of tetranucleotide frequencies, and GC values of sections of the *Haemophilus influenzae* Rd genome are consecutively ordered from the *Not*I restriction site [9]. Each column of the color image represents the fingerprint obtained from the analysis of one DNA sequence (*i.e.*, a 3000 bp section). Each row represents the frequency of a specific tetranucleotide and its complement. Tetranucleotides are arranged alphabetically on the *y*-axis. Each tetranucleotide is represented by a box, whose color is determined by its frequency, ranging from purple (low) to red (high). A star (*) identifies sections containing ribosomal RNA. The black bar identifies the location of the cryptic Mu-like bacteriophage. The variance and GC values were computed from the analysis of one section.


CONTIGS

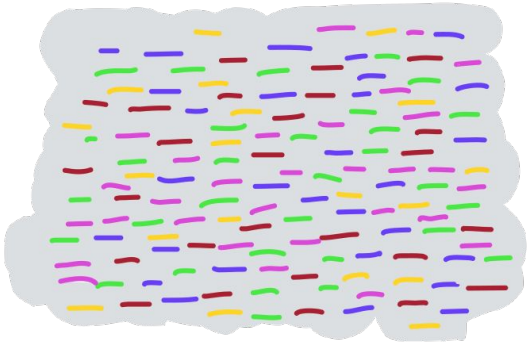

SEQUENCE COMPOSITION
CONTIGS
MAGs

CONTIG #1

CONTIG #2

CONTIG #1

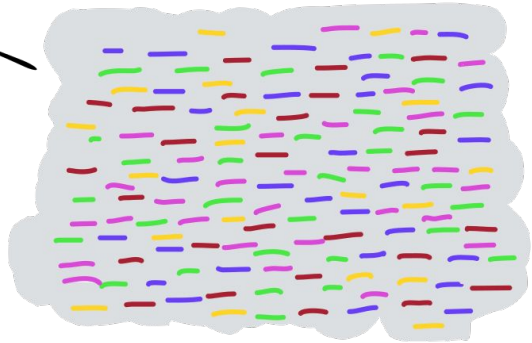
CONTIG #2



METAGENOMIC READS

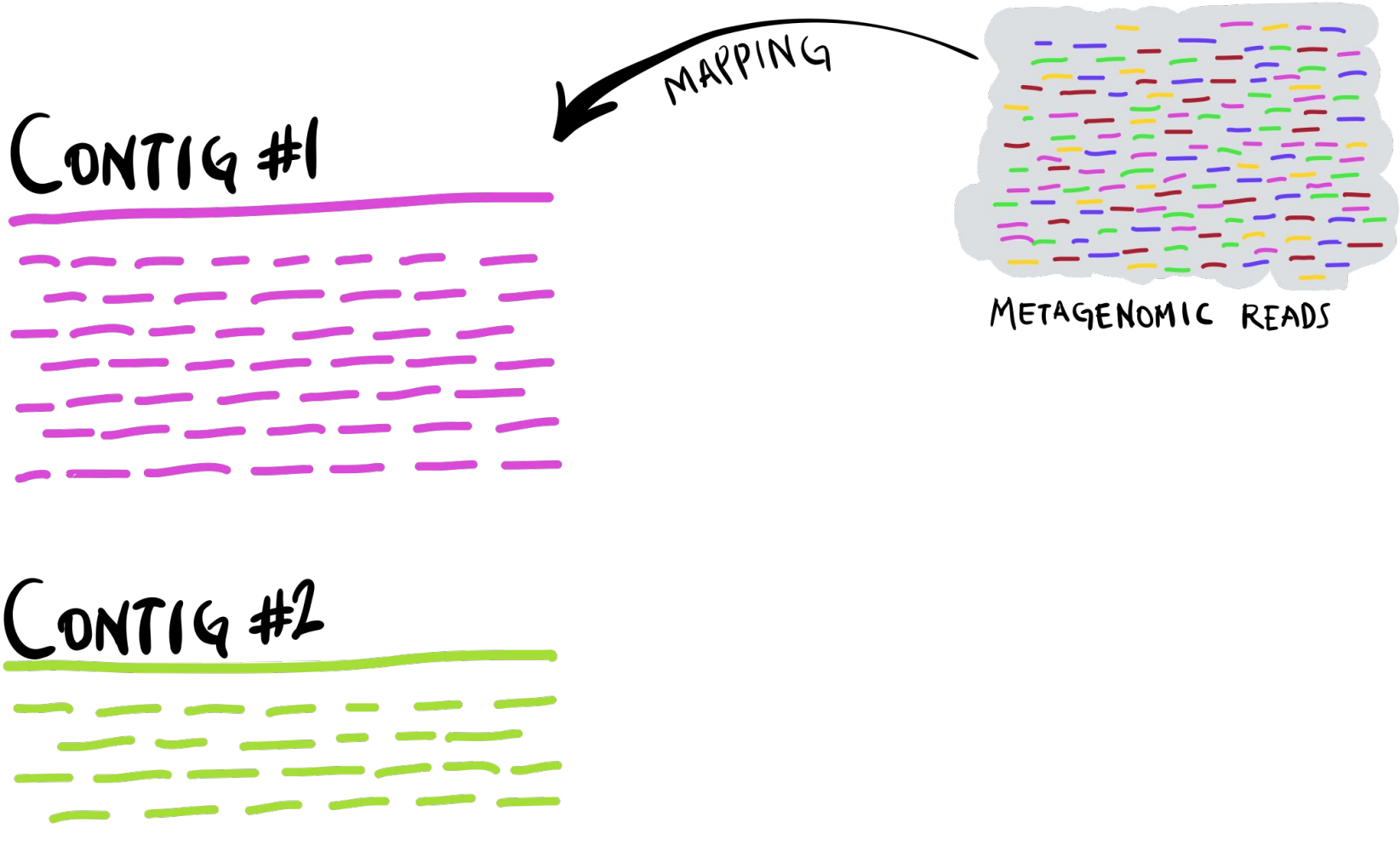
CONTIG #1

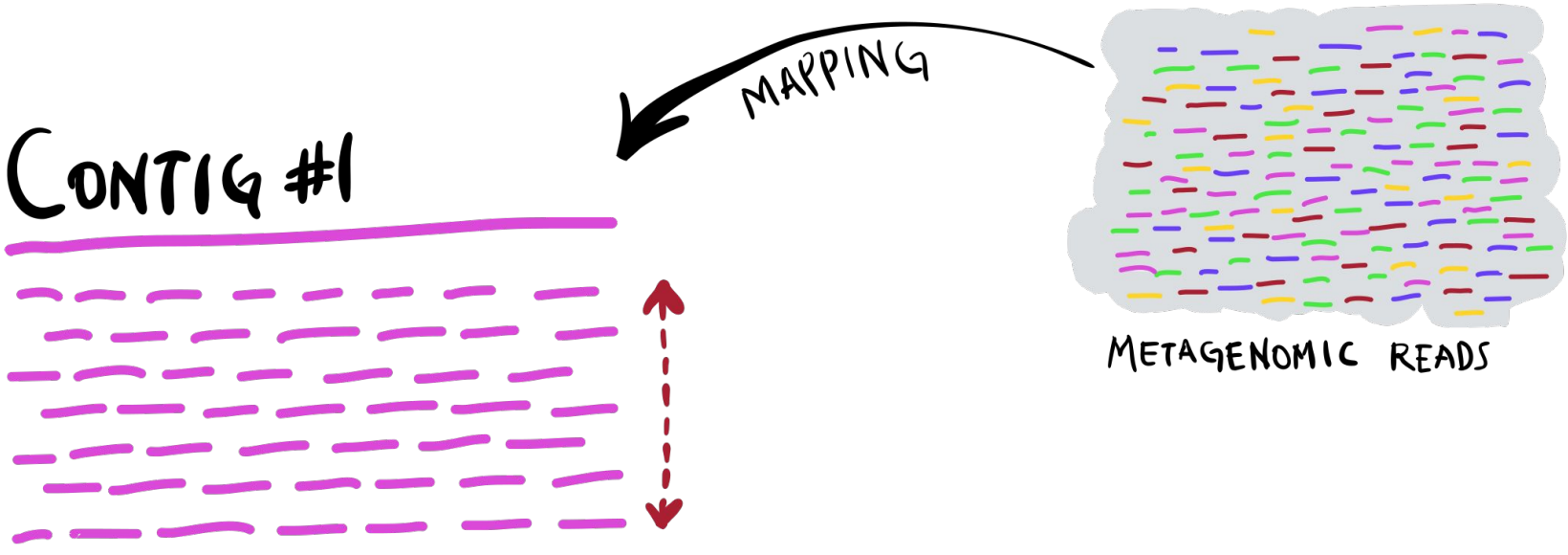
MAPPING

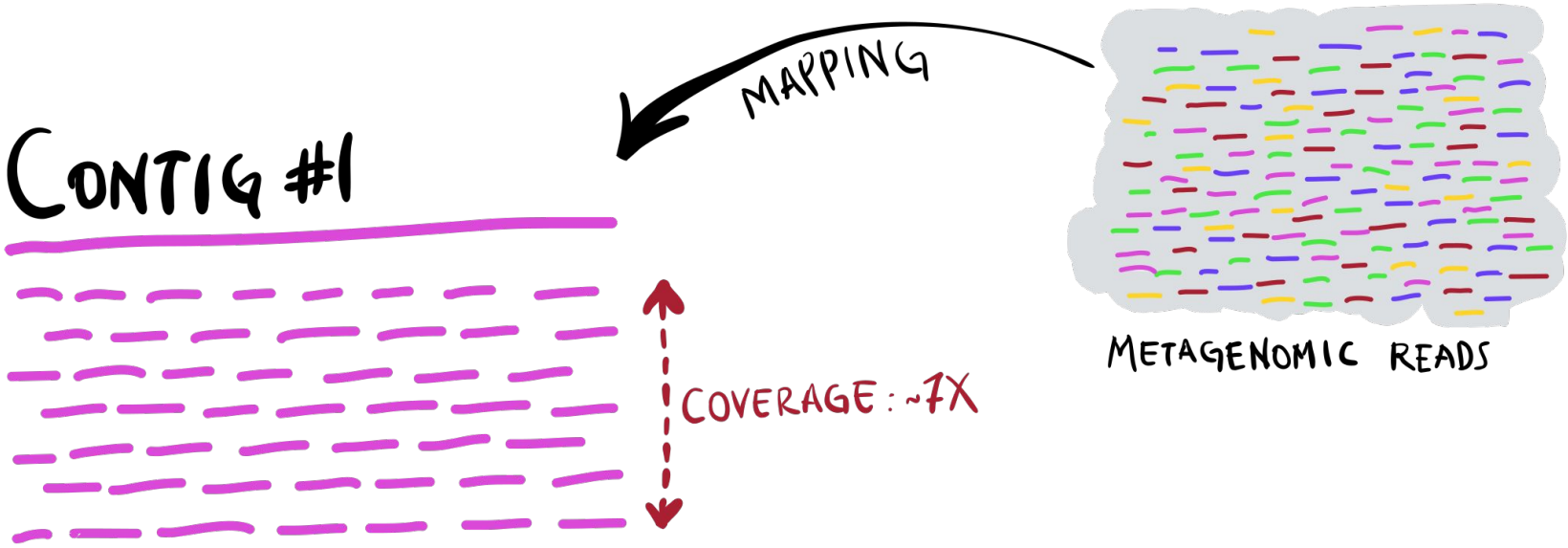


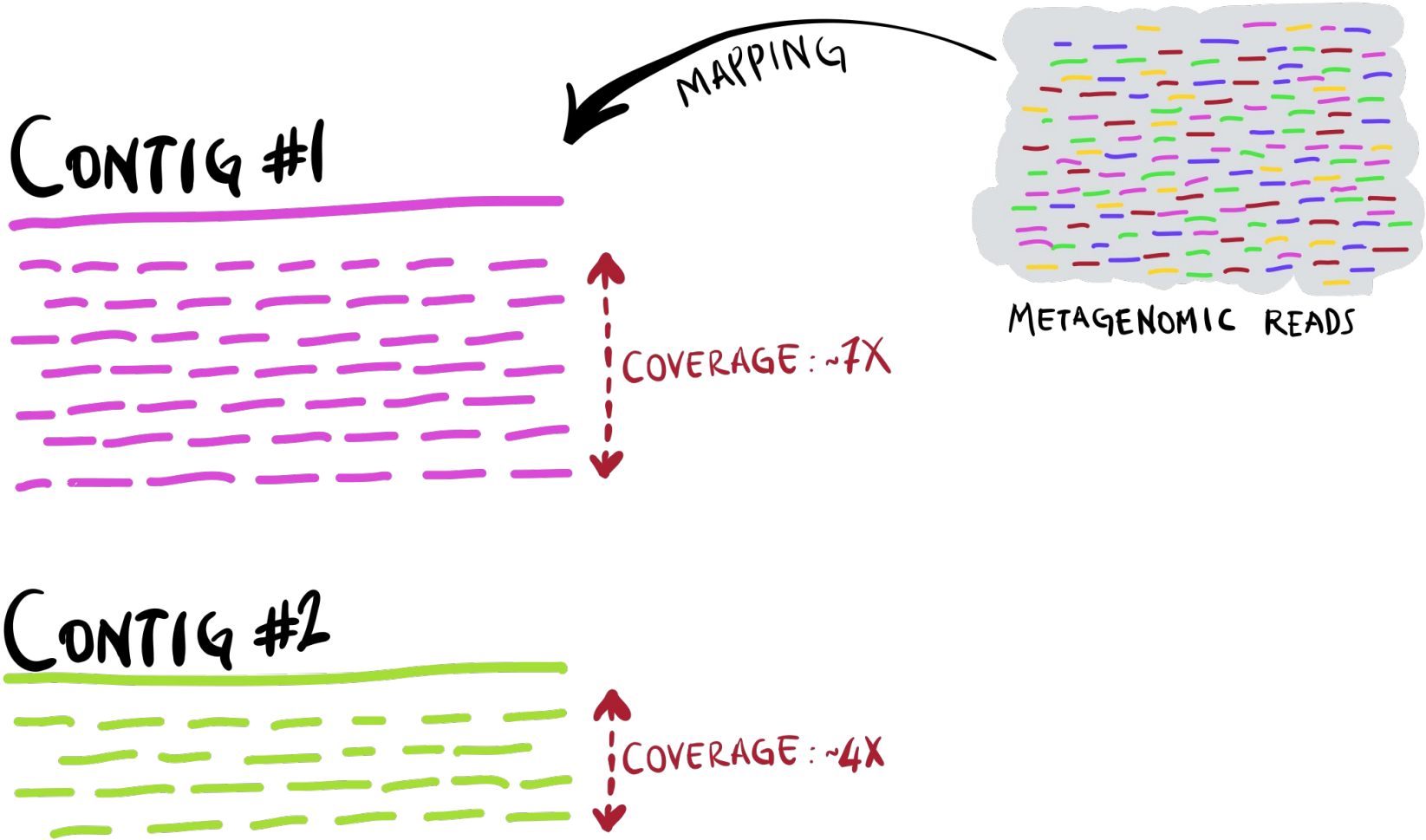
METAGENOMIC READS

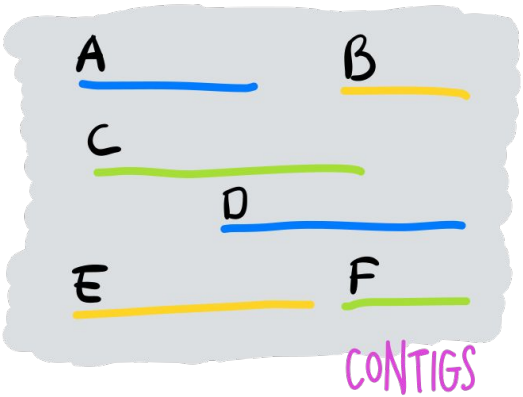
CONTIG #2

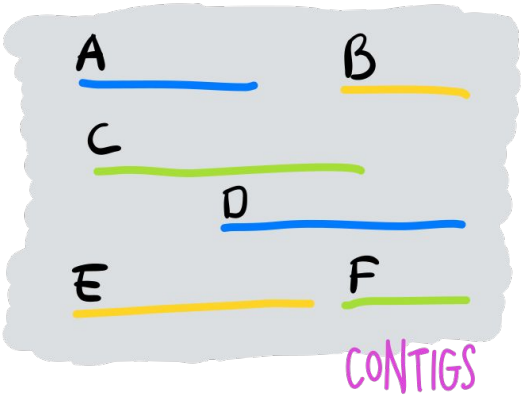

MAPPING
METAGENOMIC READS
CONTIG #1
CONTIG #2


MAPPING
METAGENOMIC READS
CONTIG #1
CONTIG #2


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


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

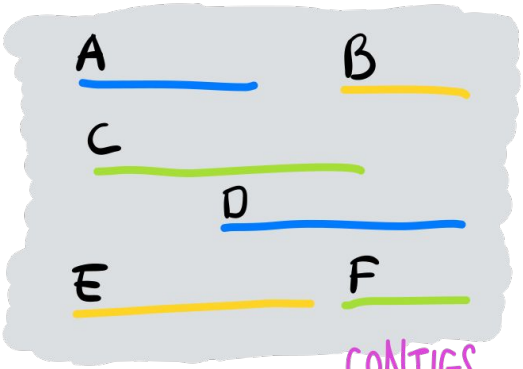

A
B
C
D
E
F
CONTIGS



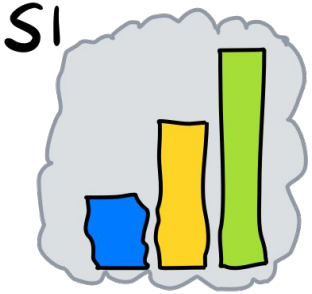
S1

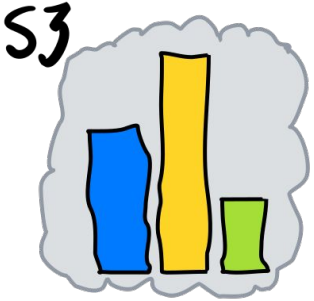
S2

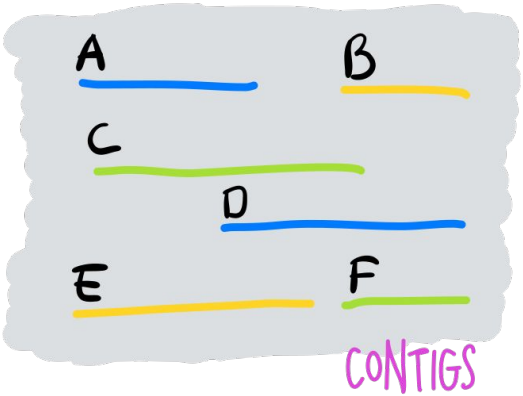
S3

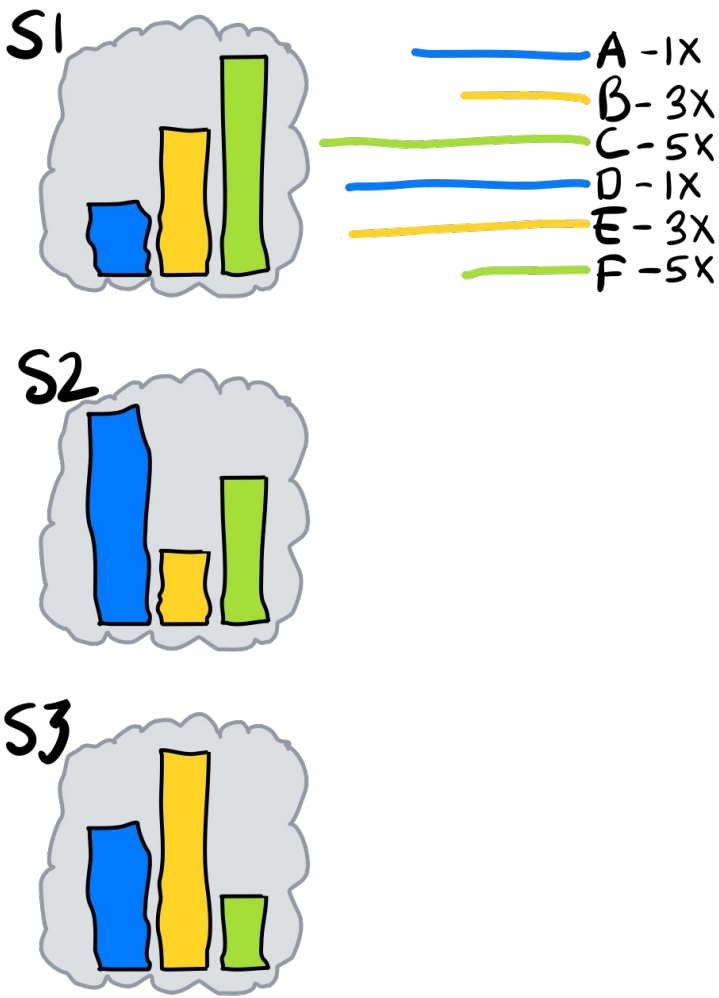


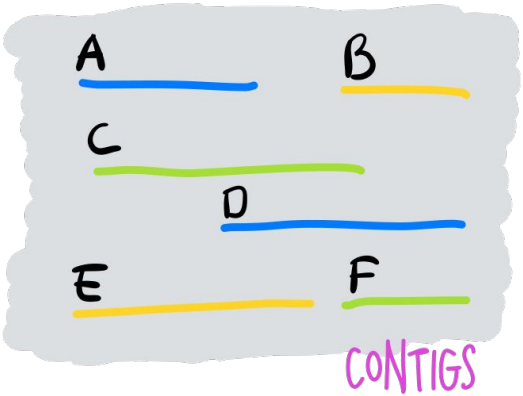
CONTIGS

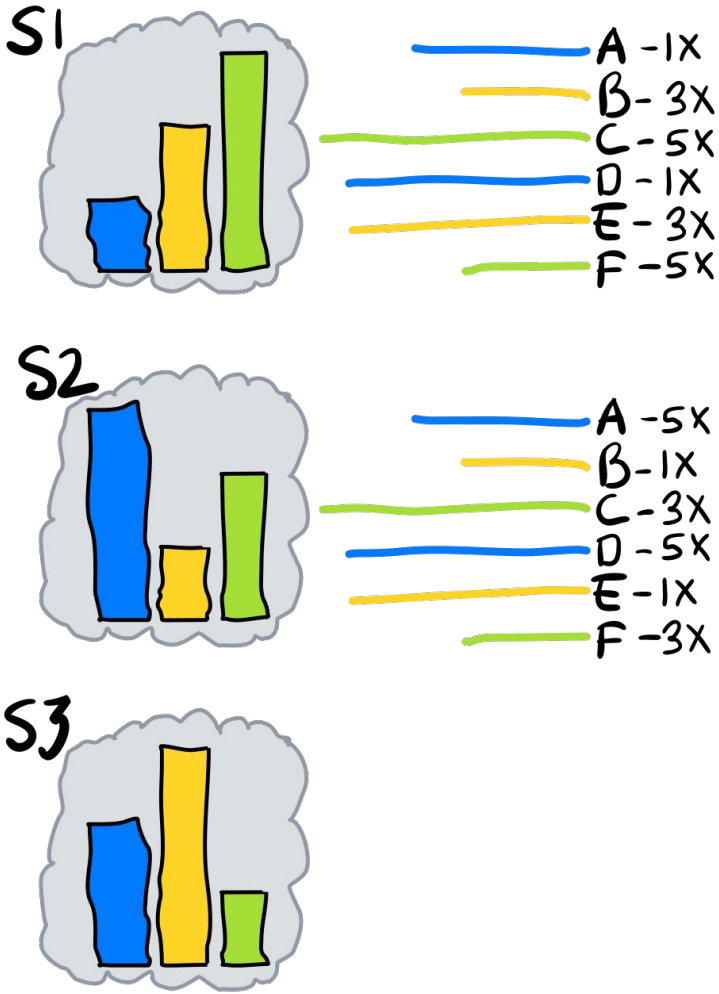


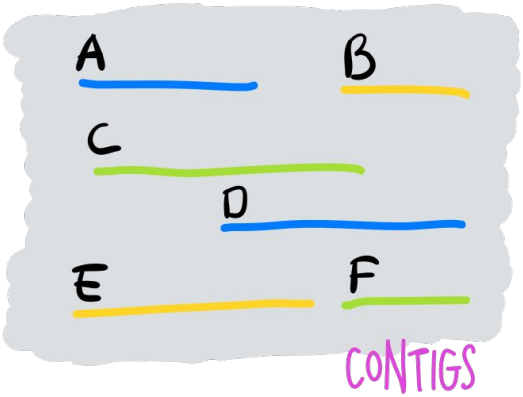


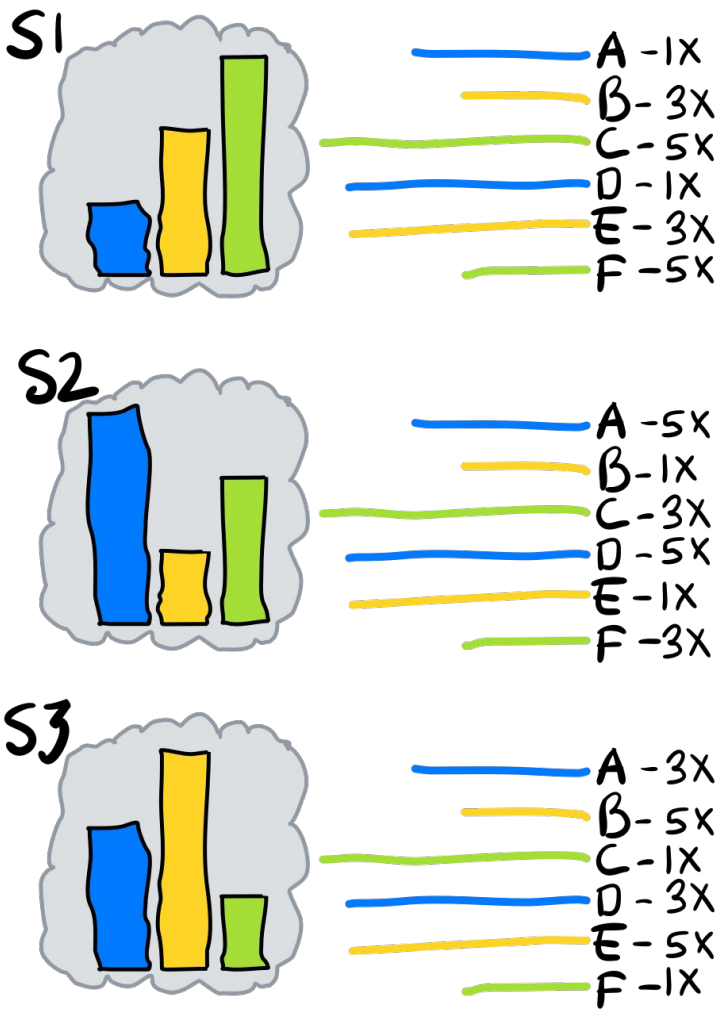

A
B
C
D
E
F
CONTIGS

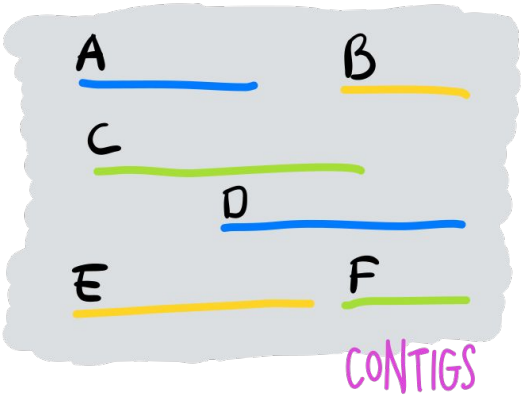

S1
A -1X
B- 3X
C- 5X
D -1X
E- 3X
F -5X
S2
S3

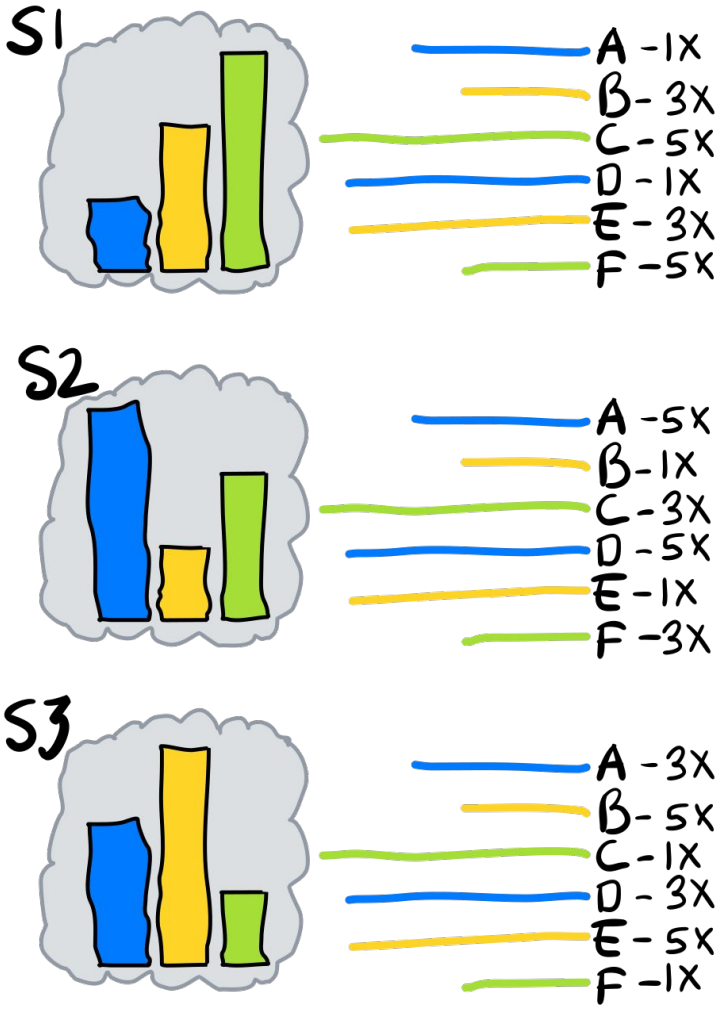

A
B
C
D
E
F
CONTIGS


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

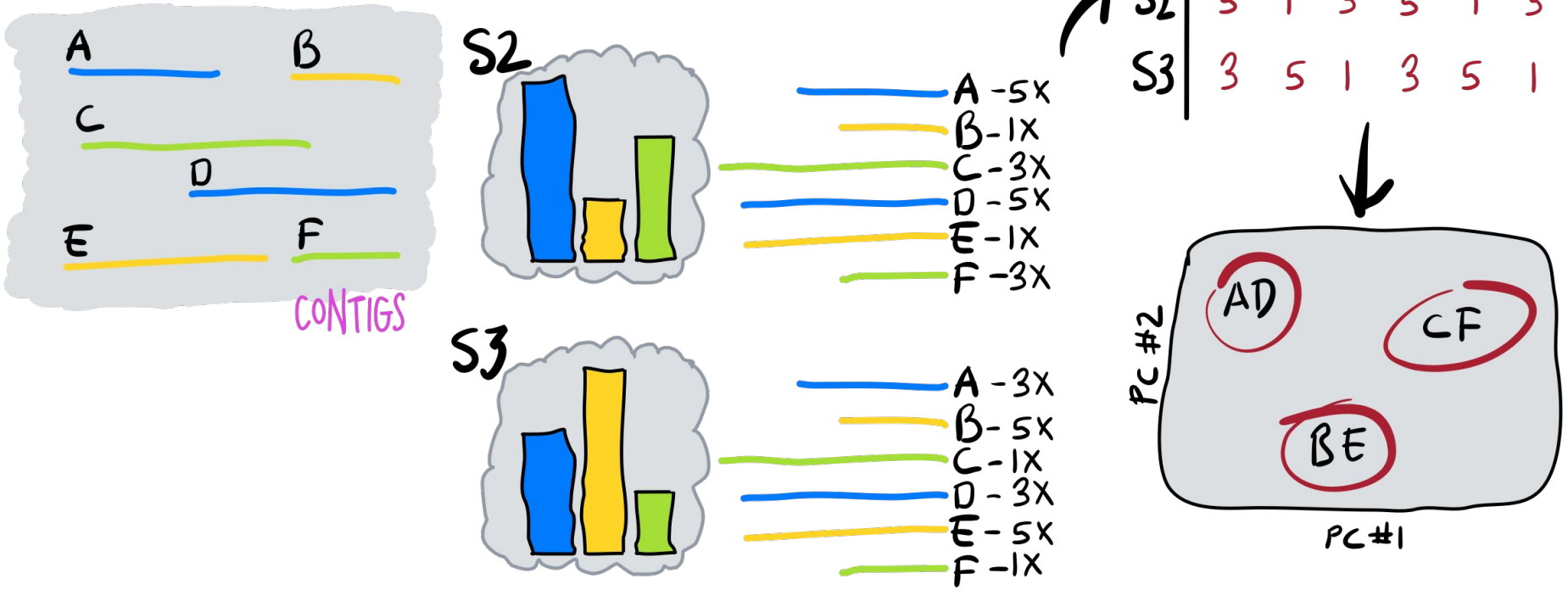

A
B
C
D
E
F
CONTIGS

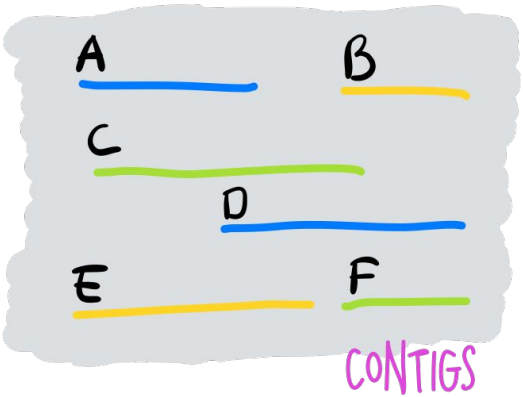

S1
A -1X
B- 3X
C- 5X
D -1X
E- 3X
F -5X
S2
A -5X
B-1X
C-3X
D-5X
E-1X
F -3X
S3
A -3X
B- 5X
C-1X
D- 3X
E- 5X
F -1X

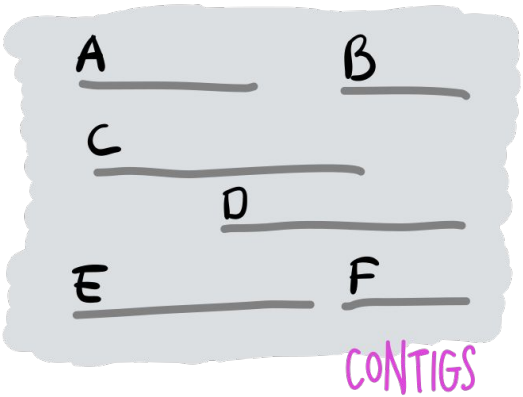


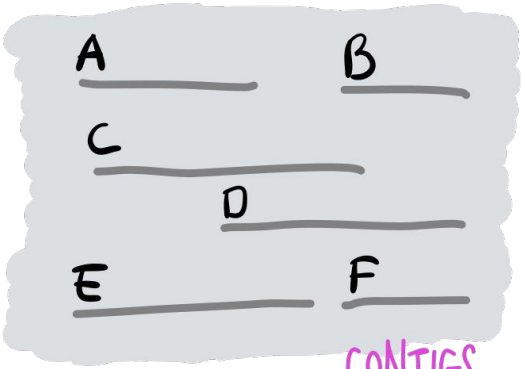


	A	B	C	D	E	F
S1	1	3	5	1	3	5
S2	5	1	3	5	1	3
S3	3	5	1	3	5	1


A
B
C
D
E
F
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


A
B
C
D
E
F
CONTIGS


A
B
C
D
E
F
CONTIGS

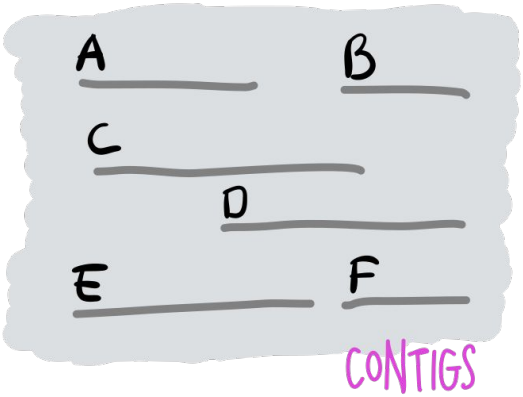


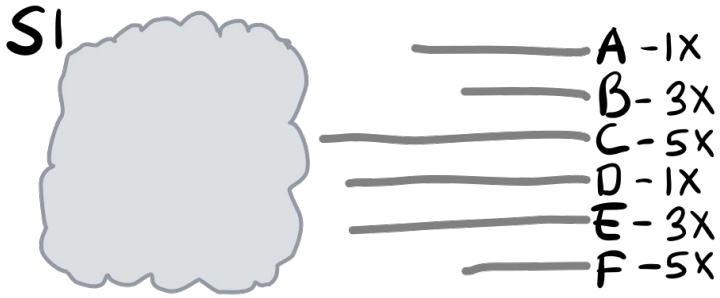
CONTIGS

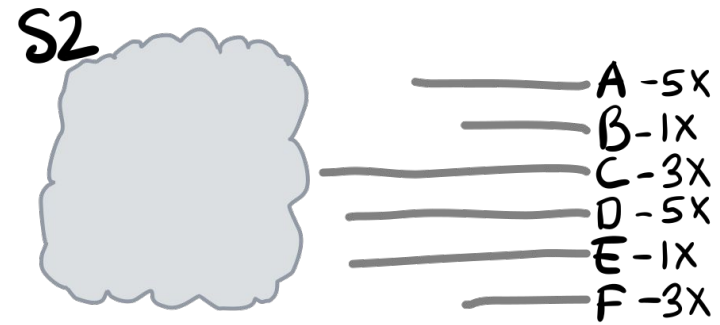
S1

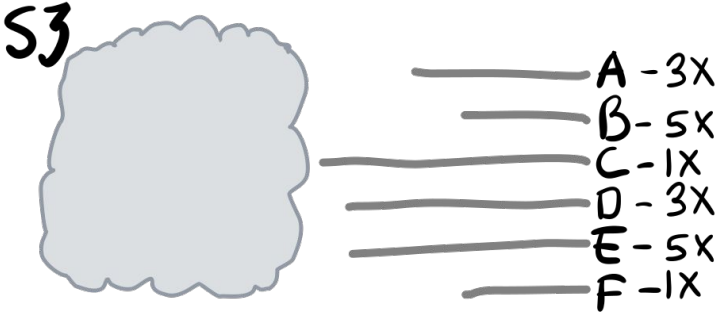
S2

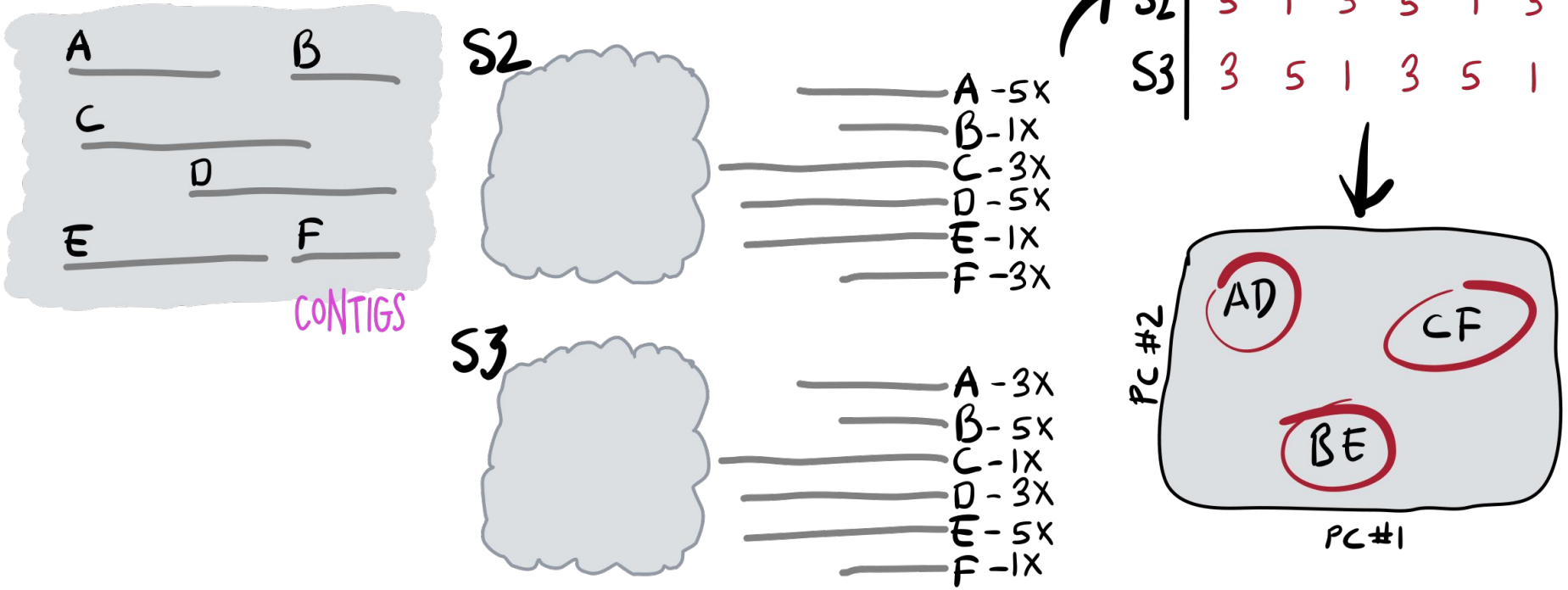
S3

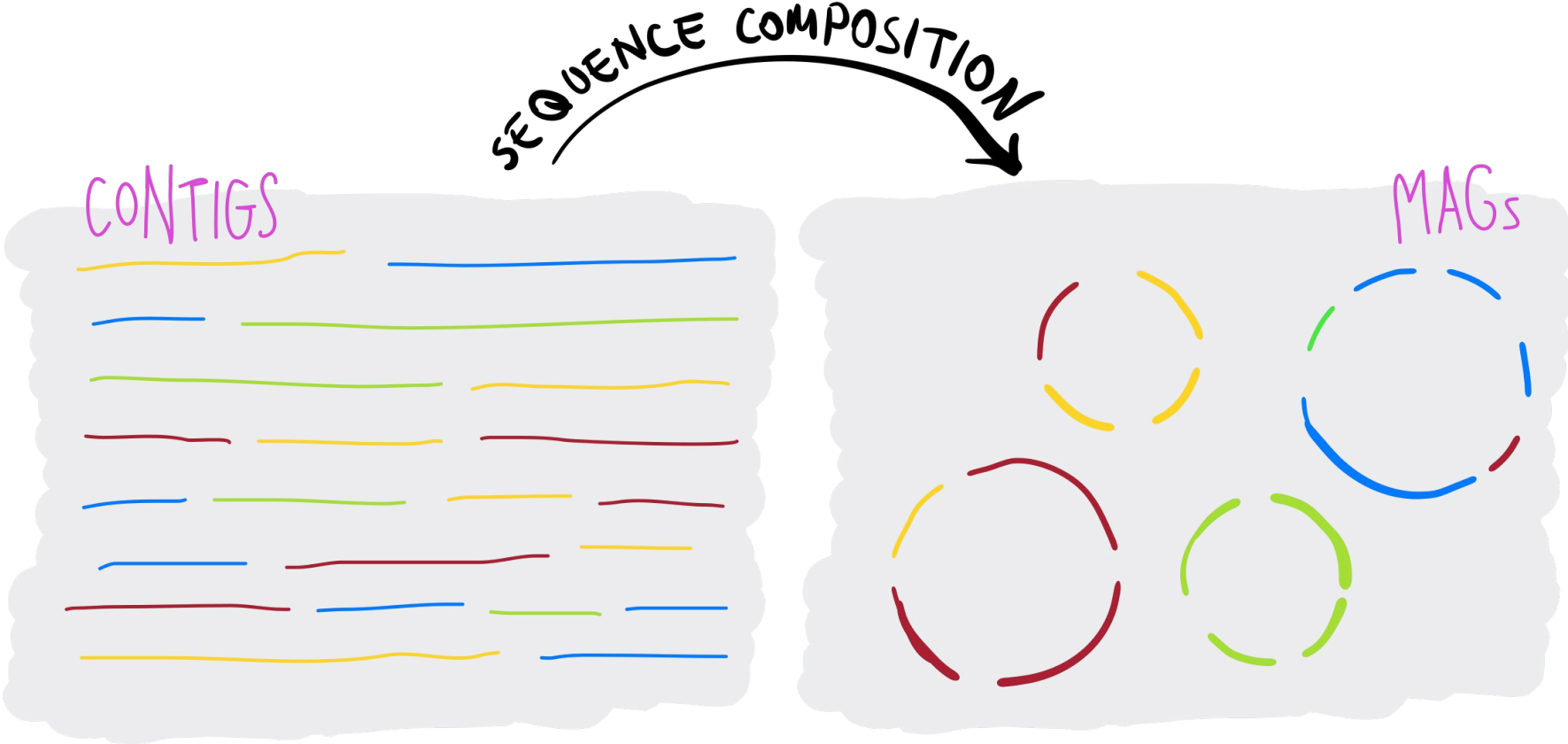

A
B
C
D
E
F
CONTIGS

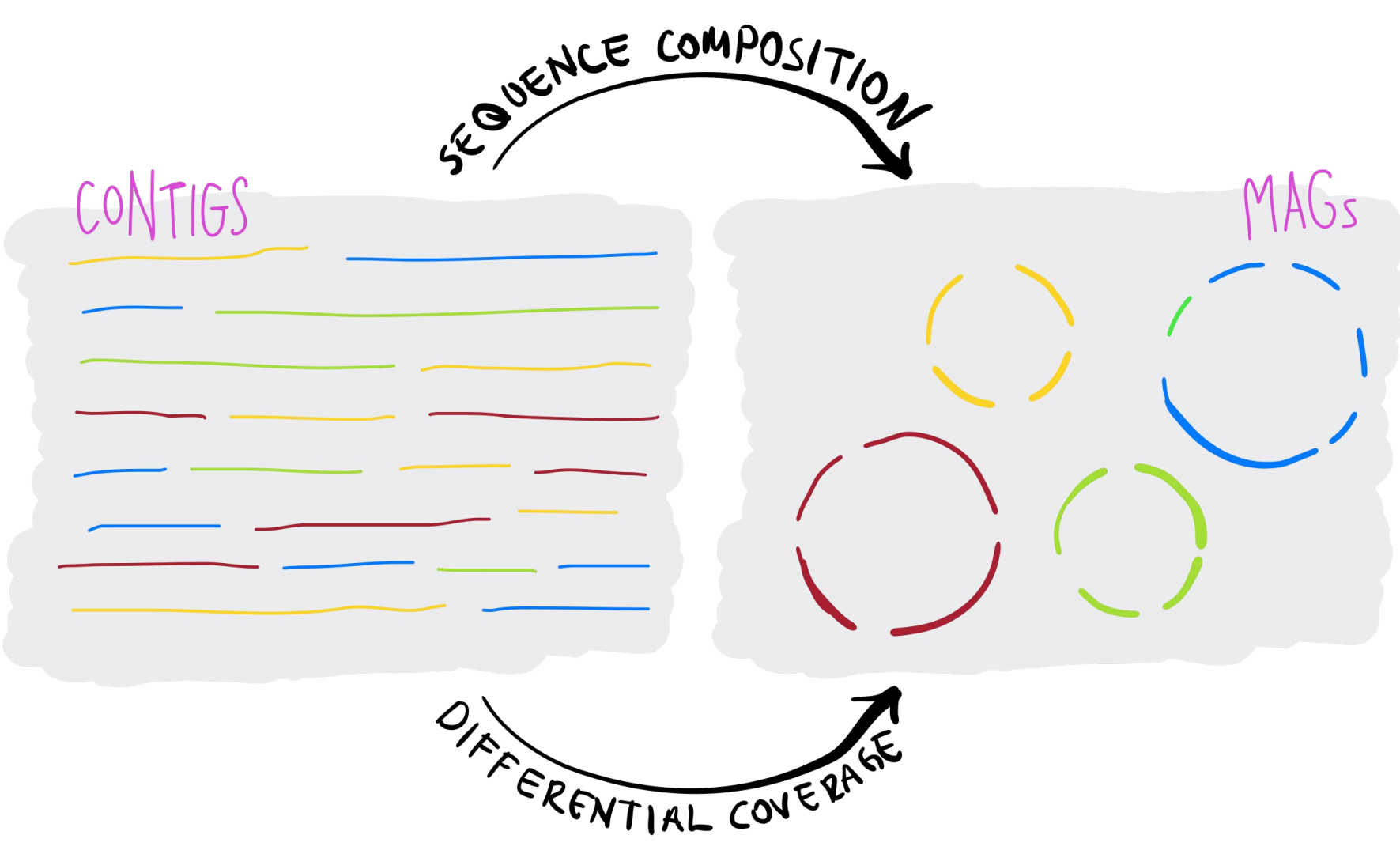

S1
A -1X
B- 3X
C- 5X
D -1X
E- 3X
F -5X


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
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


SEQUENCE COMPOSITION
CONTIGS
MAGs


SEQUENCE COMPOSITION
CONTIGS
MAGs
DIFFERENTIAL COVERAGE

Binning

- Metagenomic binning is the process of clustering sequences into clusters corresponding to operational taxonomic units of some level.
- They are reference-based binning and reference-free binning.
    - Reference-based binning methods align sequences to databases of reference genomes and determines the taxonomic group to which the sequence belongs to.
    - Reference-free binning methods make use of sequence information, without any prior knowledge and group sequences into unlabelled bins.
        - If closely related reference genomes are lacking, binning has to be conducted in an unsupervised fashion.

Basis for binning

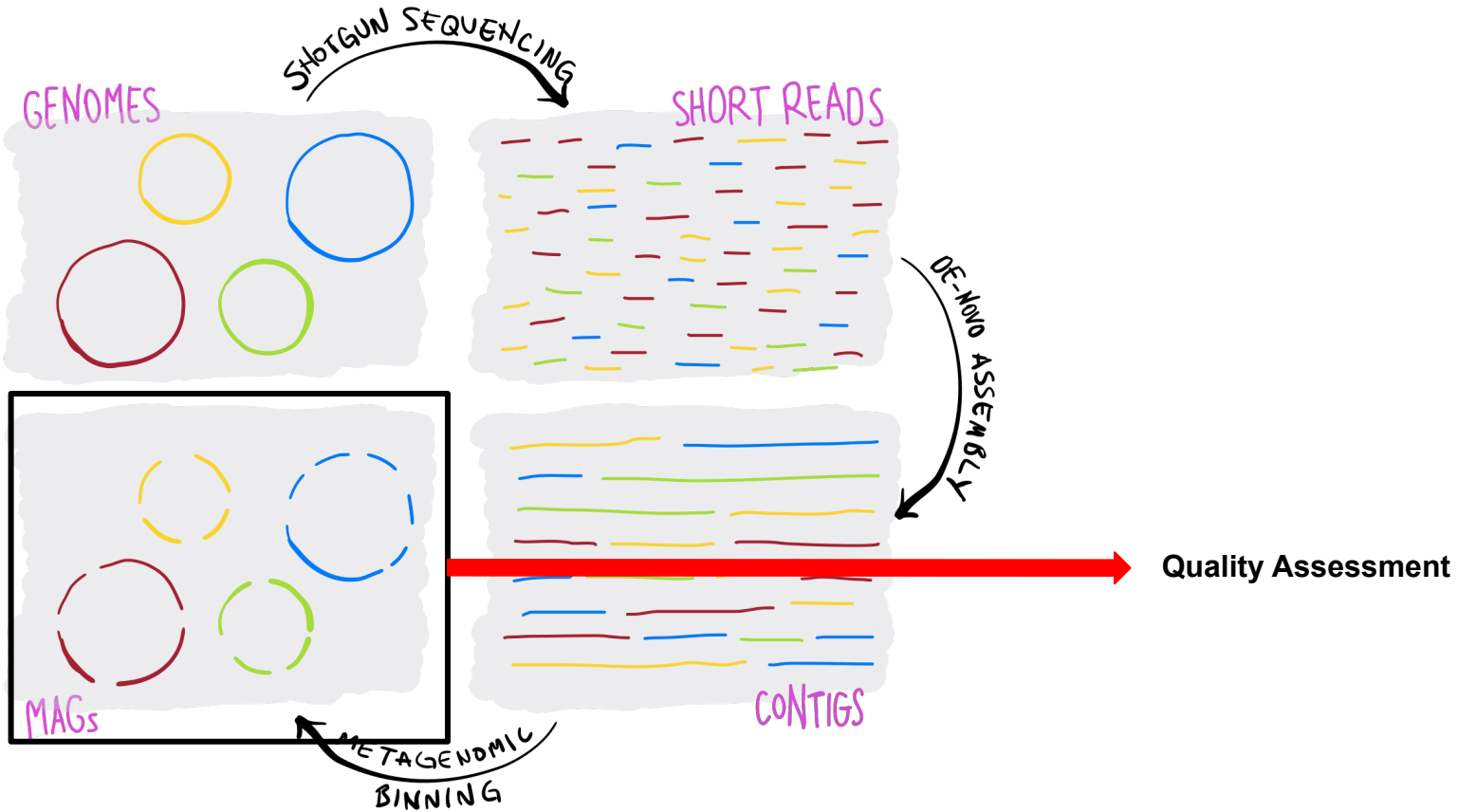
Genomic signatures to group together contigs that we don’t know the identity of into draft genomes:

- GC content
- Tetranucleotide composition (or kmer composition)
- Coverage data in multiple samples

Variations dependent on these features:

- normalized tetra-nucleotide frequency (TNF) scores,
- a graph structure and an iterative graph partitioning procedure for clustering
- linkage information from paired-end reads

MAGs QC


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

Metrics for MAGs quality assessment

- Completeness & Contamination
  - Identify and count universal single copy genes (SCGs) found in all known life, and in only one copy
  - Mainly consist of genes encoding for ribosomal proteins and other housekeeping genes
    - Completeness = number of unique SCGs present within the bin / number of total unique SCGs
    - Contamination = how many SCGs are present in multiple copies
- The presence of tRNAs for the standard 20 amino acids
- Plots depicting key genomic characteristics (e.g., GC, coding density)
  - highlight sequences outside the expected distributions of a typical genome

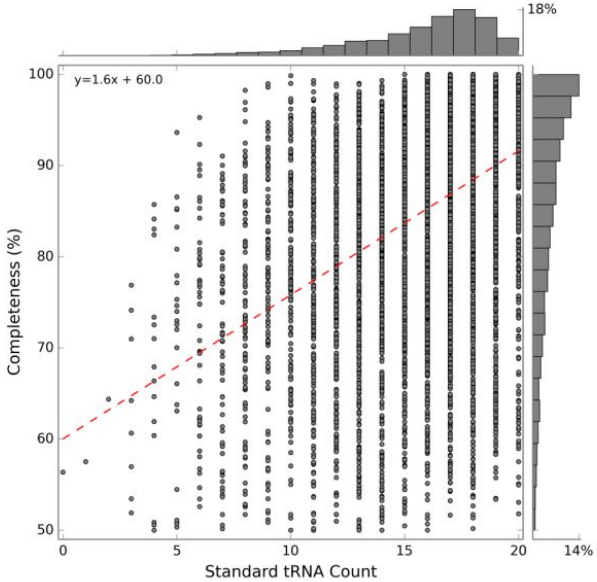
16S rRNA

A great challenge related to MAGs is their lack of 16S rRNA sequences.

- Skewed species abundance,
- high 16S sequence similarity, and
- high volumes of short-reads data

cause major difficulties for assembling the sequences of this gene, frequently rendering these genomes incomplete.

Genome completeness and identified tRNAs



Supplementary Figure 2. Correlation between number of tRNAs for each of the 20 standard amino acids and estimated genome completeness for the 7,903 UBA genomes. The correlation is positive (R^2=0.17) though the presence/absence of specific tRNAs is a poor overall metric for estimating completeness.

Parks, D.H., Rinke, C., Chuvochina, M., Chaumeil, P.A., Woodcroft, B.J., Evans, P.N., Hugenholtz, P. and Tyson, G.W., 2017. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nature microbiology, 2(11), pp.1533-1542.

Minimum information about a MAG (MIMAG) standards

Criterion	Description
	Finished
Assembly quality	Single contiguous sequence without gaps or ambiguities with a consensus error rate equivalent to Q50 or better
	High-quality draft
Assembly quality	Multiple fragments where gaps span repetitive regions. Presence of the 23S, 16S, and 5S rRNA genes and at least 18 tRNAs.
Completion	>90%
Contamination	<5%
	Medium-quality draft
Assembly quality	Many fragments with little to no review of assembly other than reporting of standard assembly statistics.
Completion	≥50%
Contamination	<10%
	Low-quality draft
Assembly quality	Many fragments with little to no review of assembly other than reporting of standard assembly statistics.
Completion	<50%
Contamination	<10%

Bowers, R.M., Kyrpides, N.C., Stepanauskas, R., Harmon-Smith, M., Doud, D., Reddy, T.B.K., Schulz, F., Jarett, J., Rivers, A.R., Eloe-Fadrosh, E.A. and Tringe, S.G., 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. Nature biotechnology, 35(8), pp.725-731.

Strain heterogeneity for HQ MAGs

Identify assemblies resulting from strain mixtures even when the strains were very closely related

- Map reads against MAGs from the same sample
- Determine dominant and non-dominant alleles over all protein coding nucleotides
- a position is considered as non-polymorphic if the dominant allele frequency was >80%

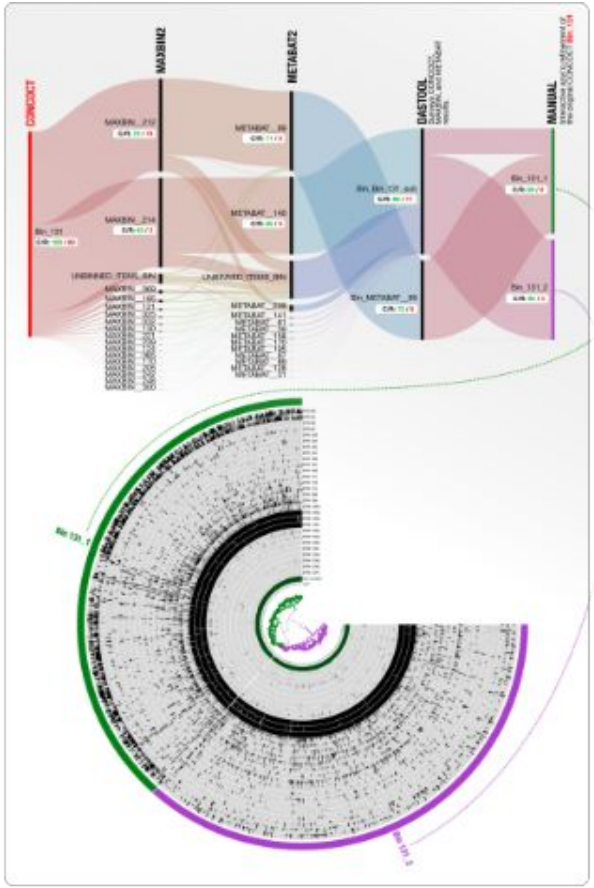
<0.5% polymorphic positions add on as a QC for HQ MAGs

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P. and Collado, M.C., 2019. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell, 176(3), pp.649-662.

Summary MAGs Quality assessment

Genome quality is mainly assessed based on

- fragmentation (i.e., the size distribution of assembled contigs, with “closed” genomes as the optimum)
- completeness (the fraction of the source genome captured)
- contamination (“surplus” genomic fragments originating from other sources), frequently estimated based on ubiquitous and single-copy marker genes (SCGs)



Visualizing the fate of contigs across metagenomic binning algorithms

a post by **A. Murat Eren (Meren)**

Web Email Twitter LinkedIn Github ORCiD

and **Jarrod J. Scott**

Web Email

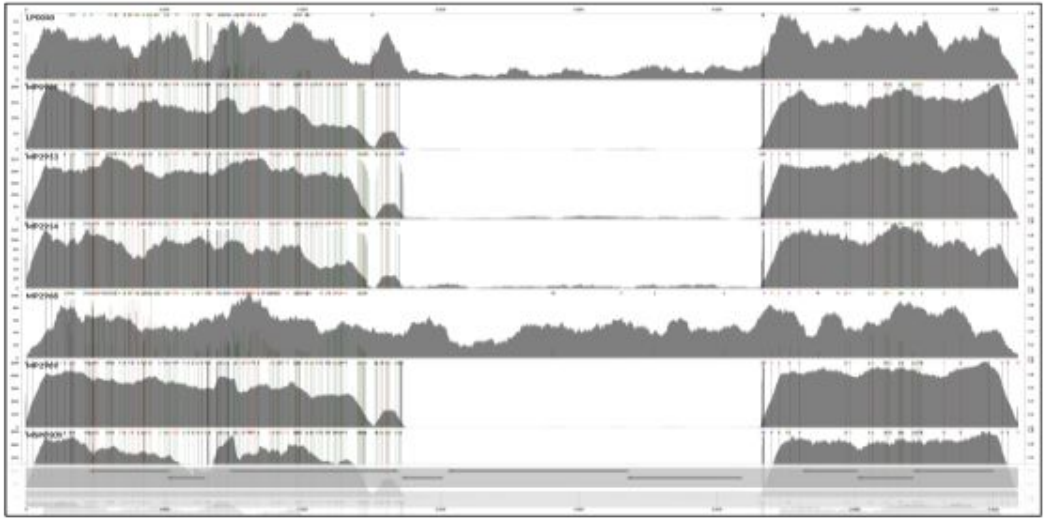
Visualizing contig coverages to better understand microbial population structures

a post by **Emily Fogarty**

Email Github ORCiD

and **Ryan Moore**

Twitter LinkedIn Github



merenlab.org/

Important Omics Vocabulary

Binning

Tetranucleotide frequency

GC content

Read recruitment

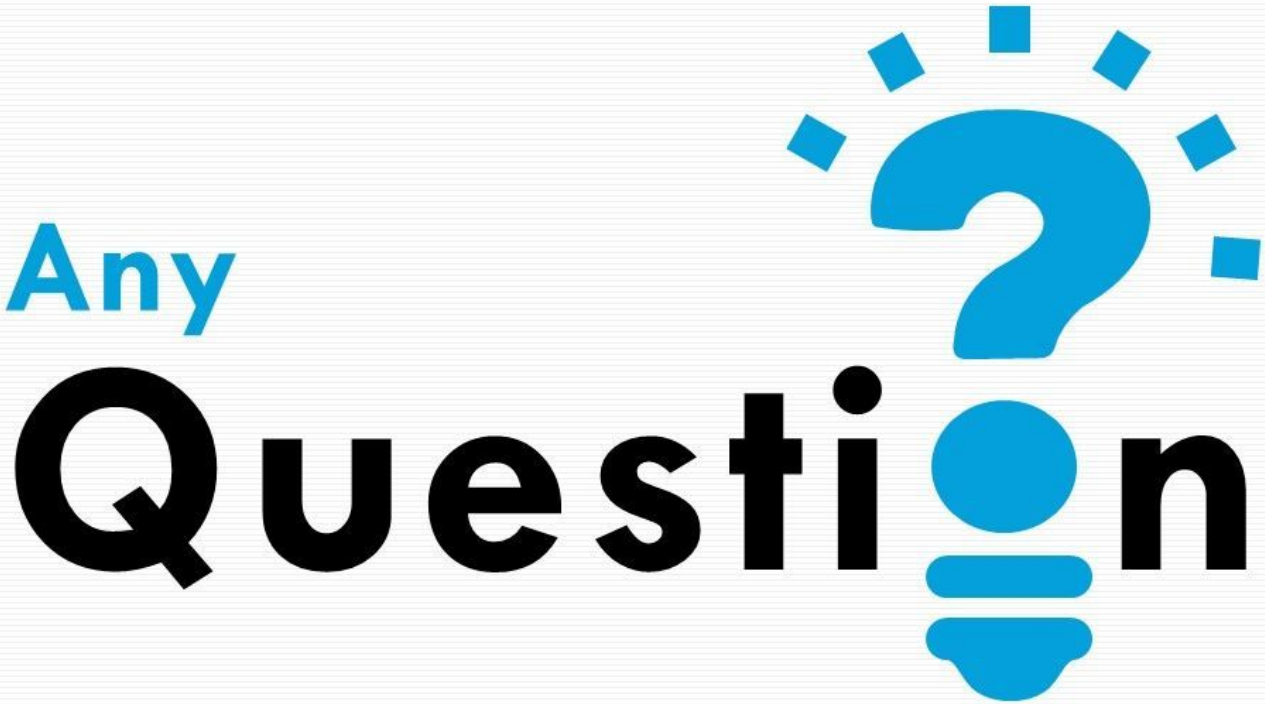
Genome coverage

MAGs

MAGs QC ≠ Reads QC

MIMAG

https://anvio.org/vocabulary/


Any
Question

From bins to species and abundance estimation

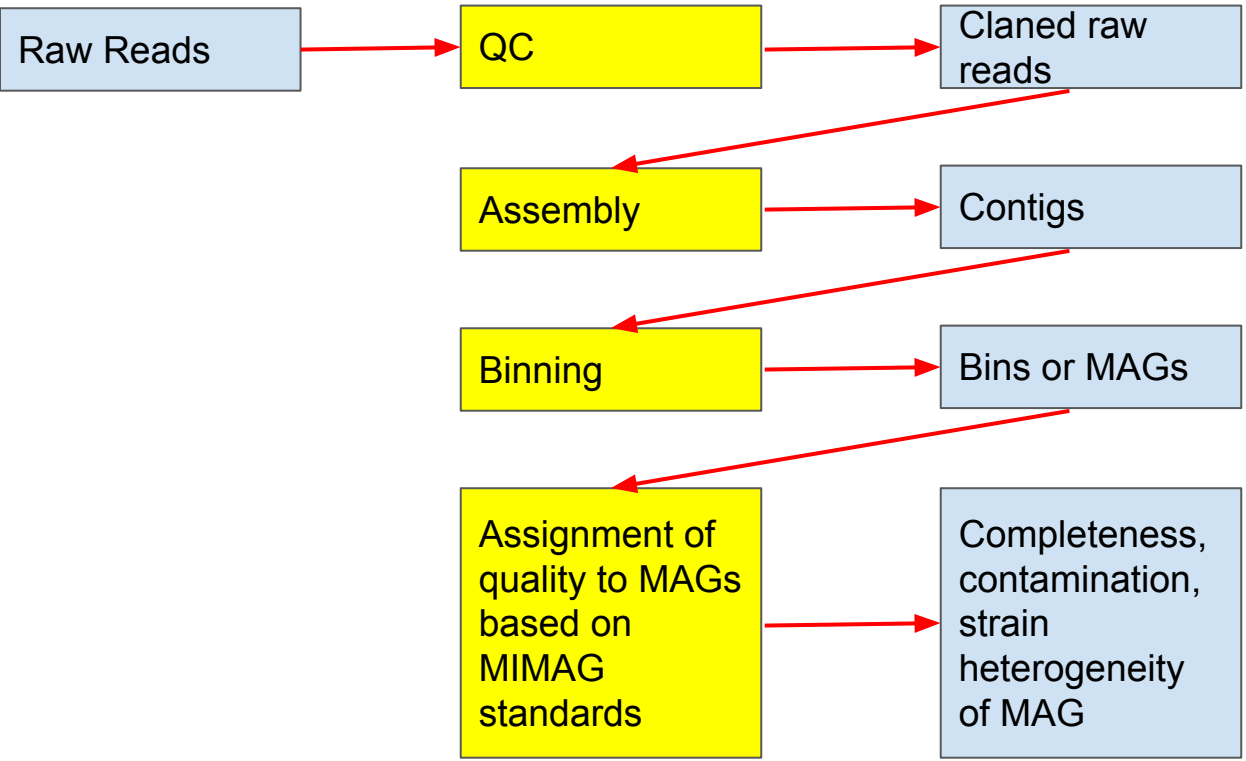
Biol-217
22 January - 2 February 2024

Dr. Cynthia M. Chibani

Outline

- Bin refinement
- Bin reassembly
- Detection of chimeras
- Bins dereplication
- Taxonomic assignment
- Depth of coverage
- Genome abundances

Short Summary



Contaminations in MAGs

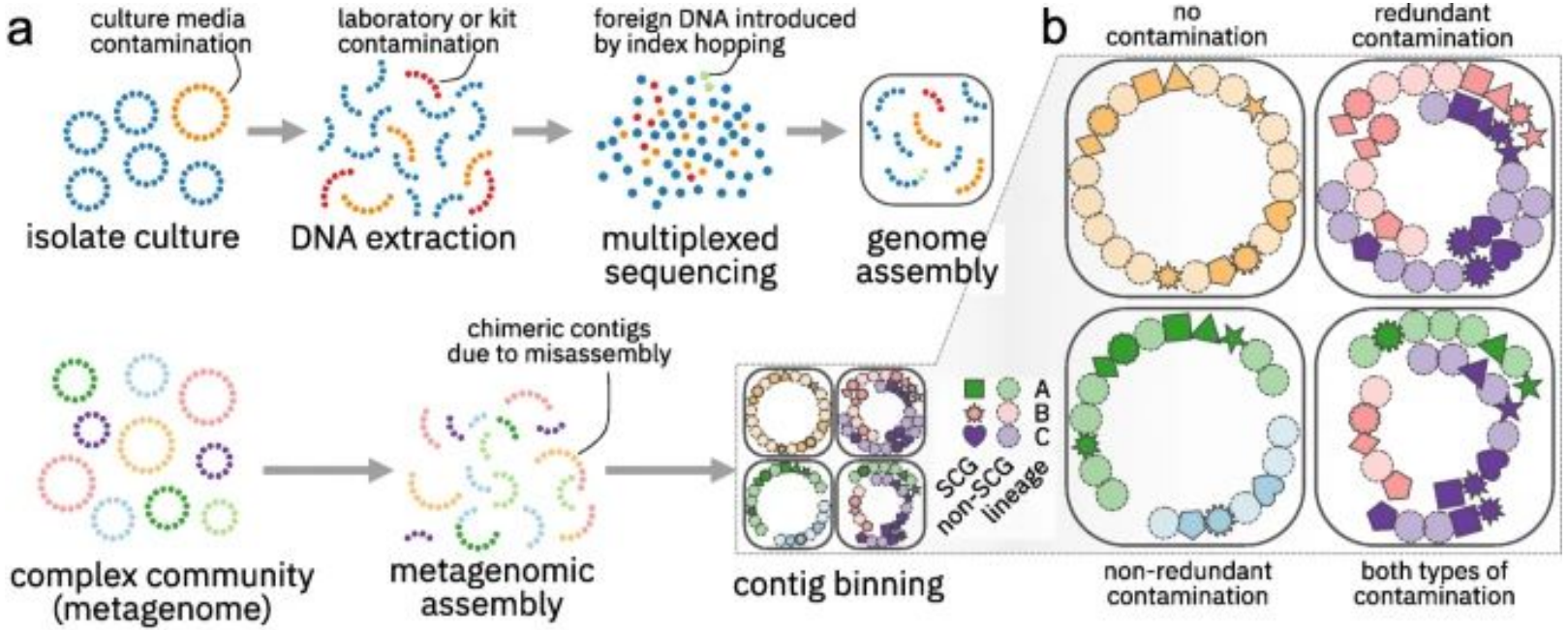
Errors in isolate-derived genomes

- introduced during physical sample processing
    - contamination of reagents or culture media.

Errors in MAGs are expected to be computational

- misassembly (relatively rare)
- mis-binning (major source of errors)

Contaminations in MAGs



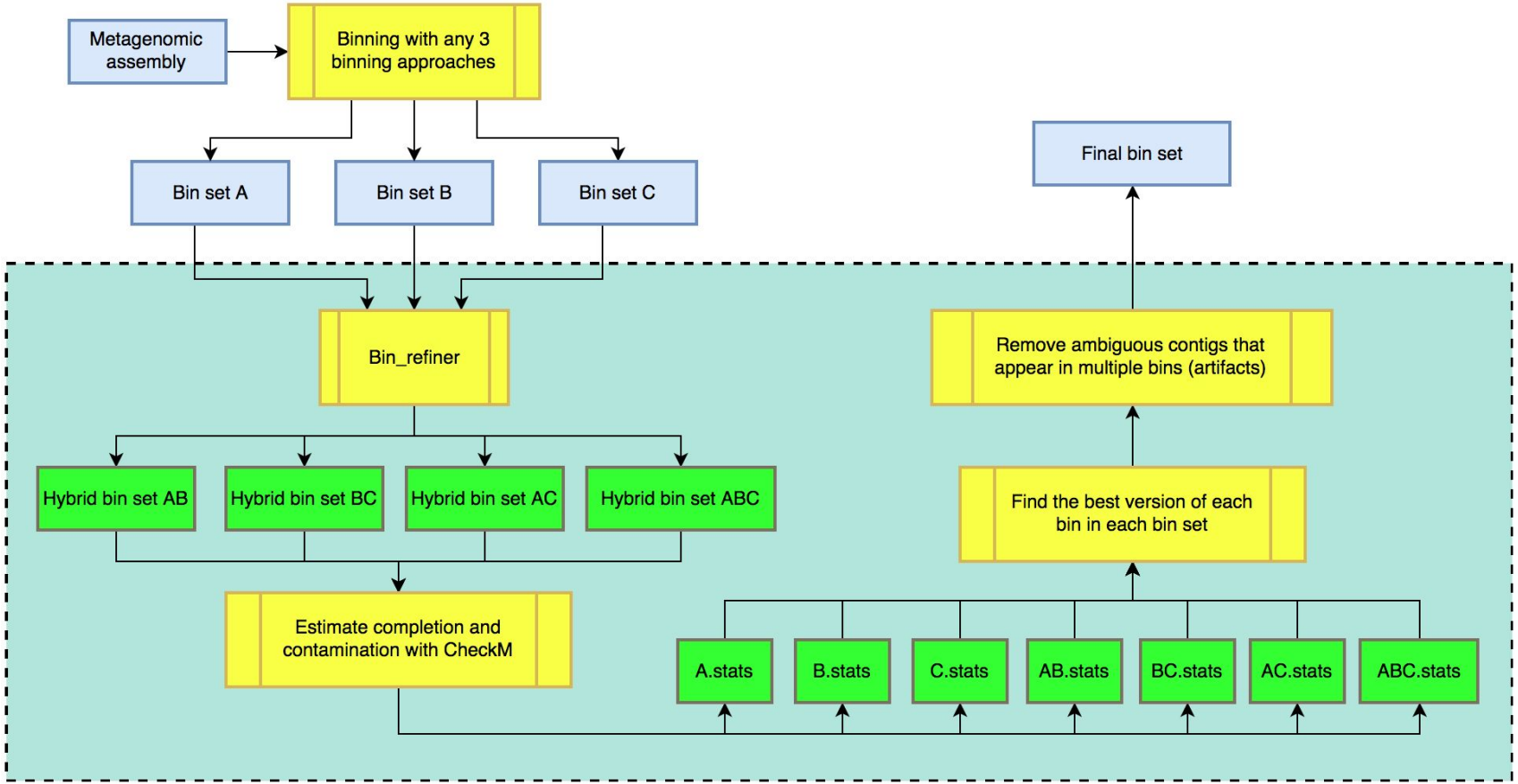
Orakov, A., Fullam, A., Coelho, L.P., Khedkar, S., Szklarczyk, D., Mende, D.R., Schmidt, T.S. and Bork, P., 2021. GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome biology, 22(1), pp.1-19.

Ways to improve Bin quality

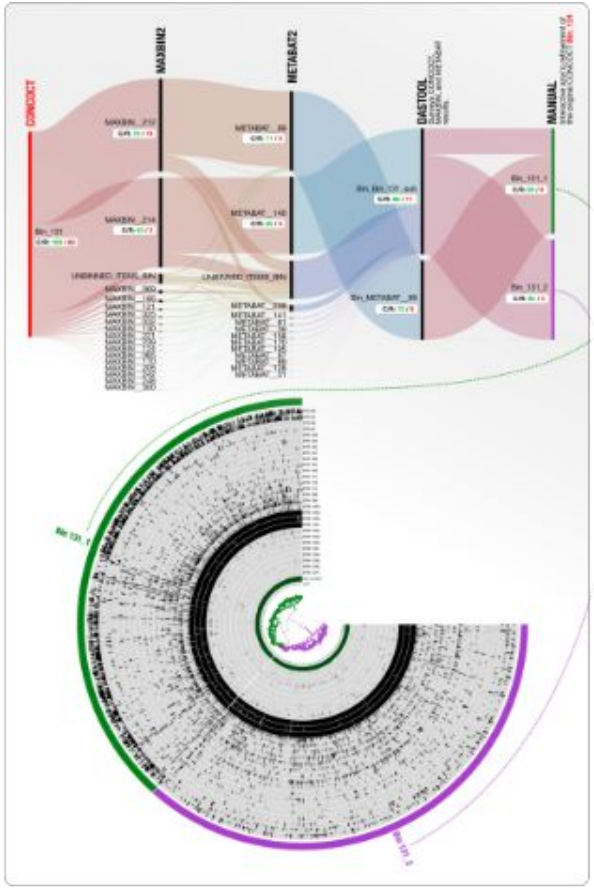
- Bin refinement
  - Use multiple binners
  - Consolidate results

Uritskiy, G.V., DiRuggiero, J. and Taylor, J., 2018. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome, 6(1), pp.1-13.

Bin Refinement



Uritskiy, G.V., DiRuggiero, J. and Taylor, J., 2018. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome, 6(1), pp.1-13.



Visualizing the fate of contigs across metagenomic binning algorithms

a post by **A. Murat Eren (Meren)**

Web Email Twitter LinkedIn Github ORCiD

and **Jarrod J. Scott**

Web Email

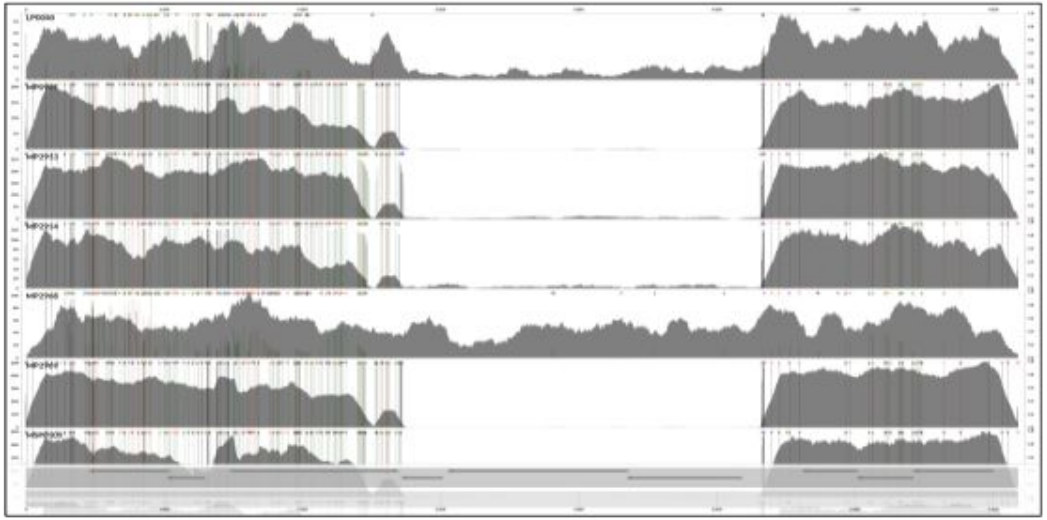
Visualizing contig coverages to better understand microbial population structures

a post by **Emily Fogarty**

Email Github ORCiD

and **Ryan Moore**

Twitter LinkedIn Github



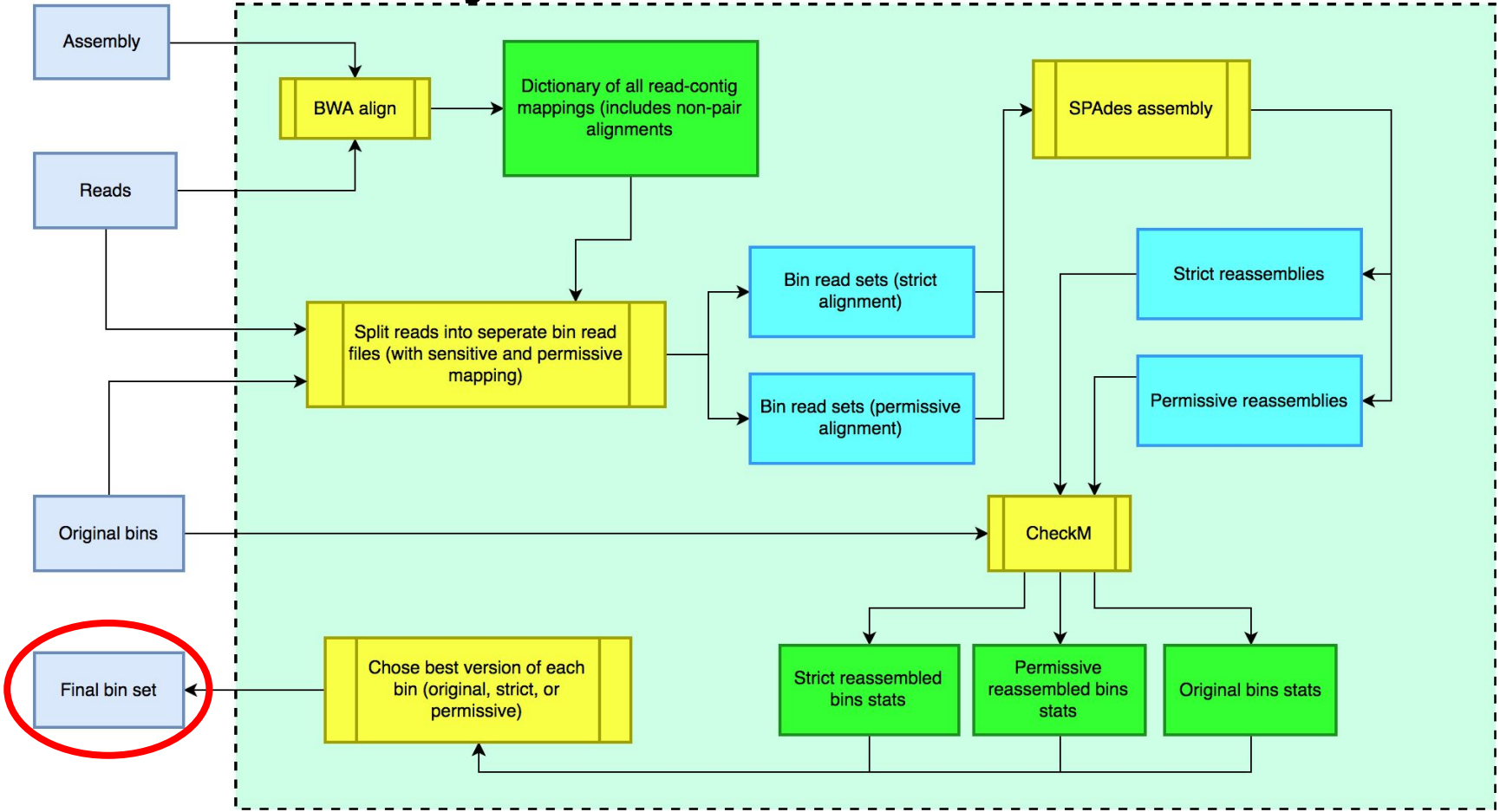
merenlab.org/

Ways to improve Bin quality

- Bin refinement
  - Use multiple binners
  - Consolidate results
- Bin reassembly
  - Align raw reads to bins
  - Extracting reads that belong to a given bin and assembling them separately from the rest of the metagenome
- Re-evaluate completeness and contamination of each bin

Uritskiy, G.V., DiRuggiero, J. and Taylor, J., 2018. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome, 6(1), pp.1-13.

Bin Reassembly



Uritskiy, G.V., DiRuggiero, J. and Taylor, J., 2018. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome, 6(1), pp.1-13.

Contaminations in MAGs

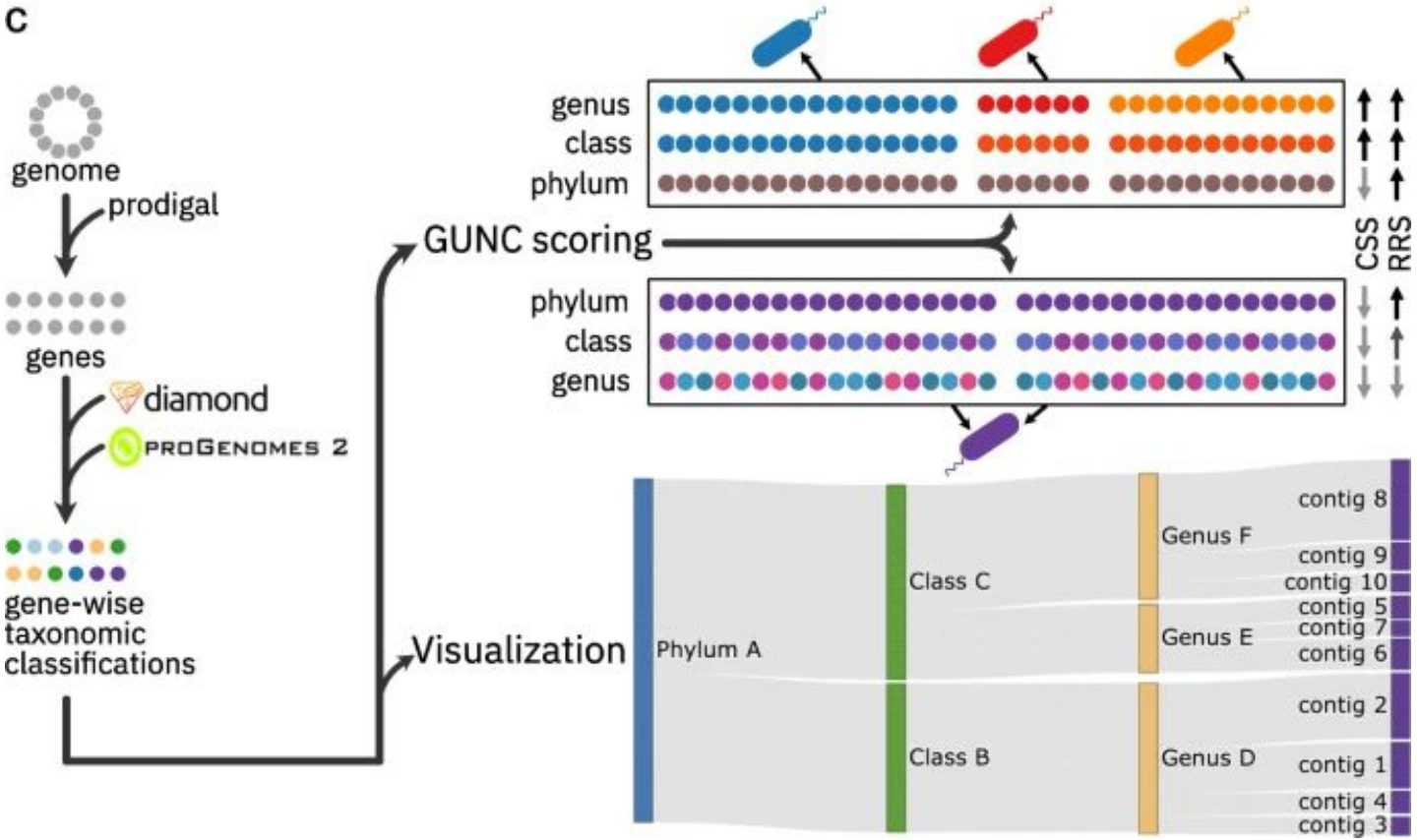
MAGs can be taxonomically resolved to strain level.

Beneficial in undersampled environments where reference genomic coverage is scarce

Contaminating fragments are detrimental to biological interpretation

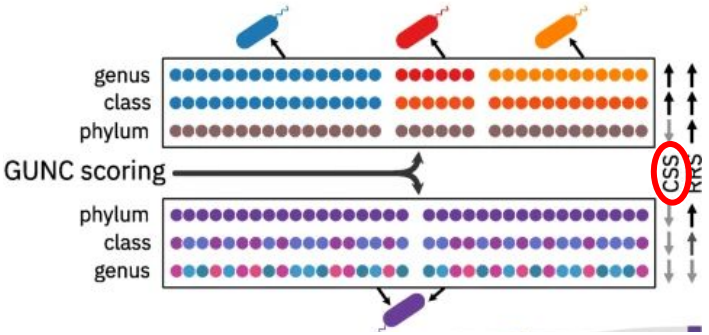
- cause false inferences about a genome’s functional repertoire
- Or structure.

Chimera detection in MAGs



Orakov, A., Fullam, A., Coelho, L.P., Khedkar, S., Szklarczyk, D., Mende, D.R., Schmidt, T.S. and Bork, P., 2021. GUNC: detection of chimerism and contamination in prokaryotic genomes. Genome biology, 22(1), pp.1-19.

Clade separation score (CSS)



Measures

- how diverse the taxonomic assignments are within each contig
- normalized to the diversity across the whole genome
- and normalized to the expected entropy when there is no relationship between taxonomic labels across contigs

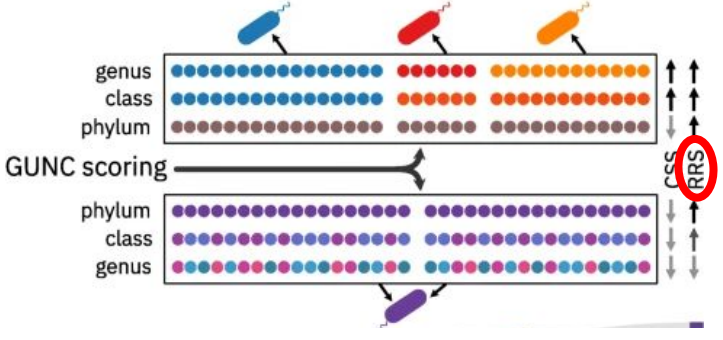
0< CSS <1

- A genome composed of contigs that are internally homogeneous, but disagree with each other, then the metric will return a value closer to 1

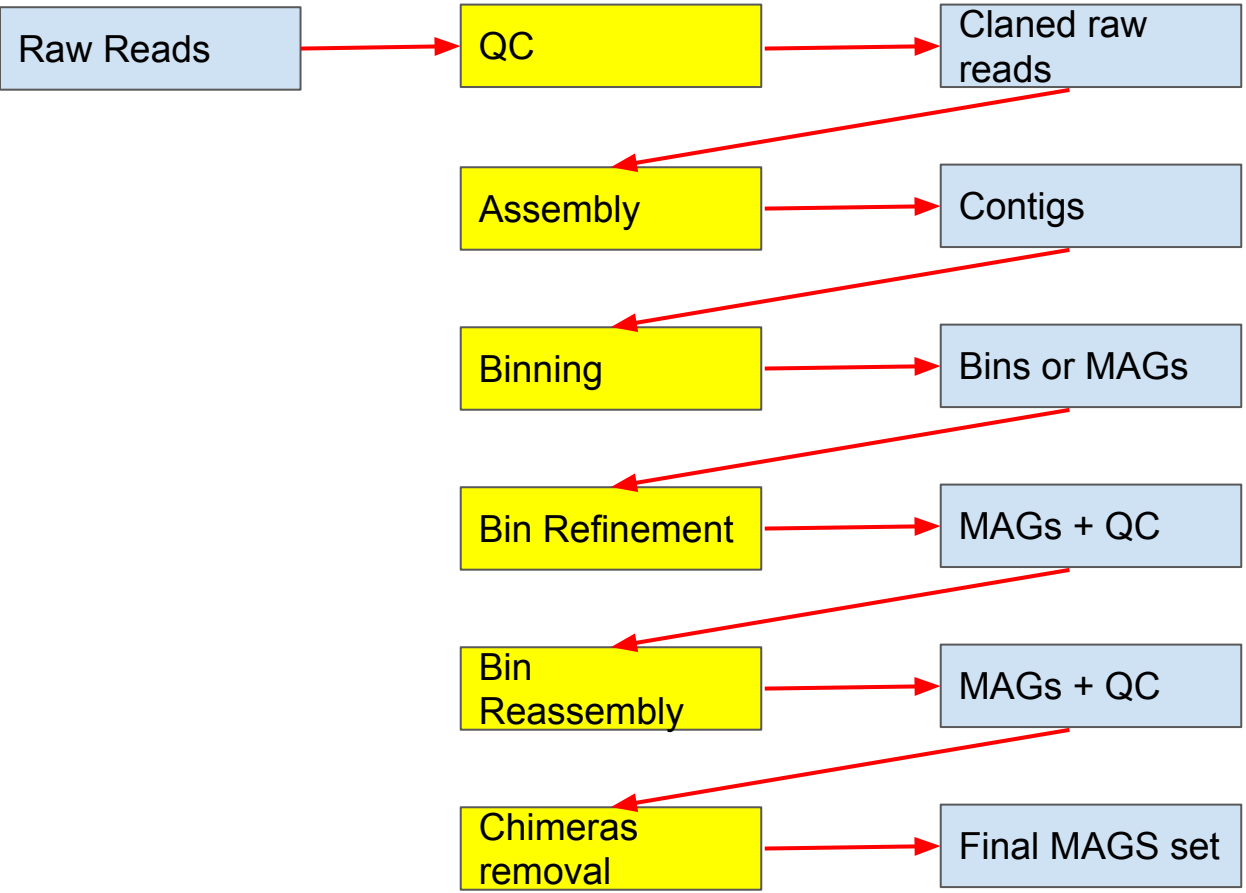
- A genome with all genes assigning to the same taxonomy, i.e., free of contamination, will be assigned a CSS score of 0

Reference representation score (RSS)

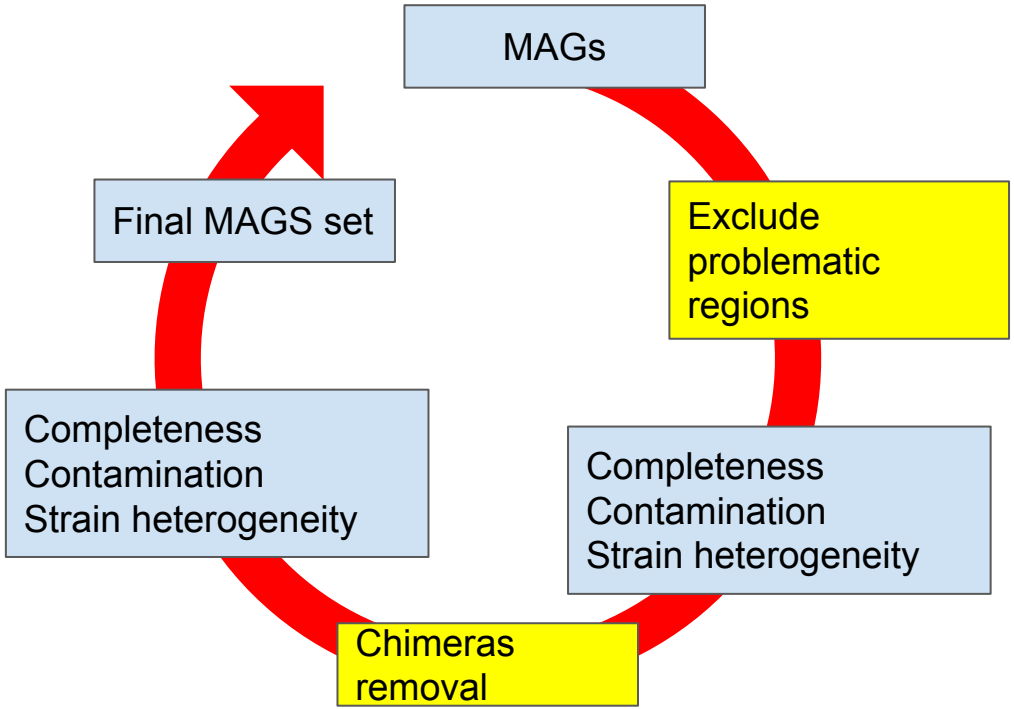
- RSS is based on the average identity of query genes to the reference and the number of spurious mappings
  - High RRS values ⇒ genomes map confidently into the reference space at a given taxonomic level
  - Low RRS ⇒ a lineage is “novel” relative to the reference



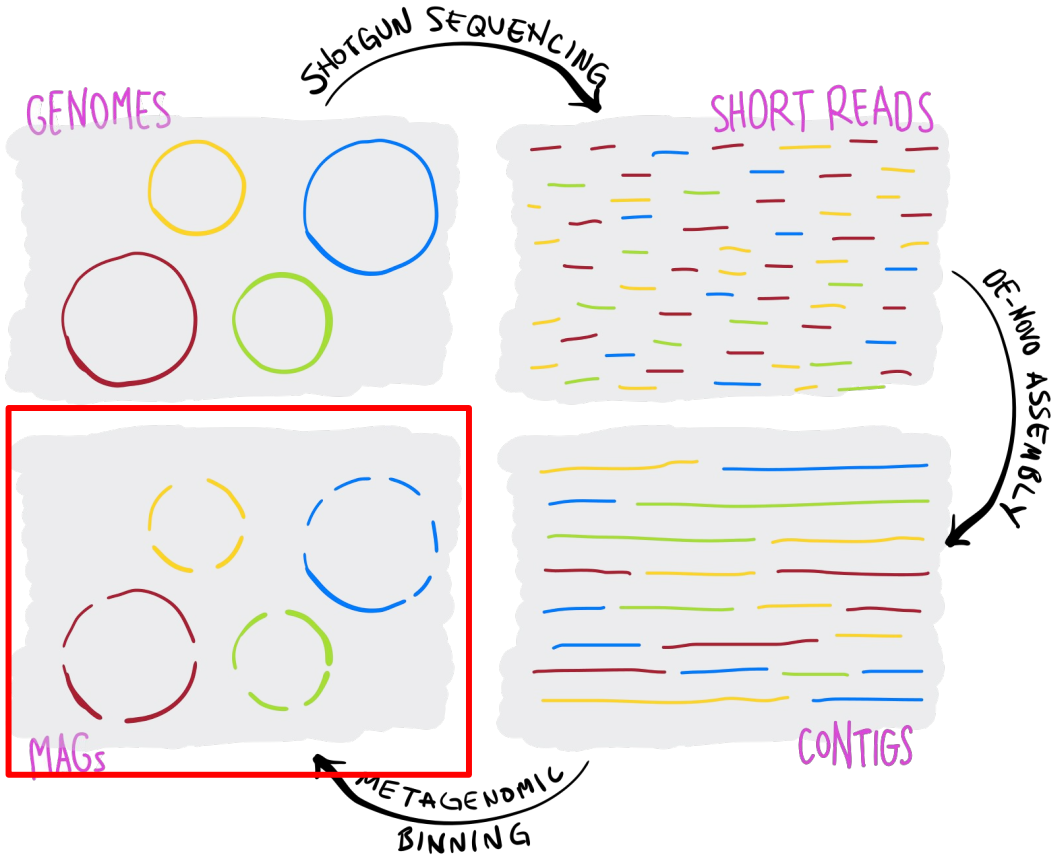
Short Summary



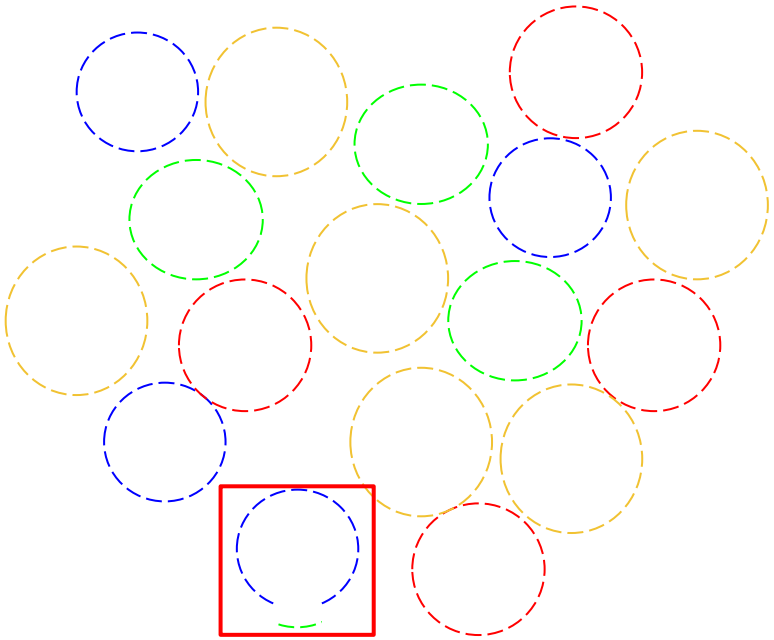
Bin refinement in the tutorial



Final MAGS set



Final MAGS in practice



- If you assemble per sequenced sample
  ⇒ Multiple MAGs with a mix of different qualities of probably the same genomes

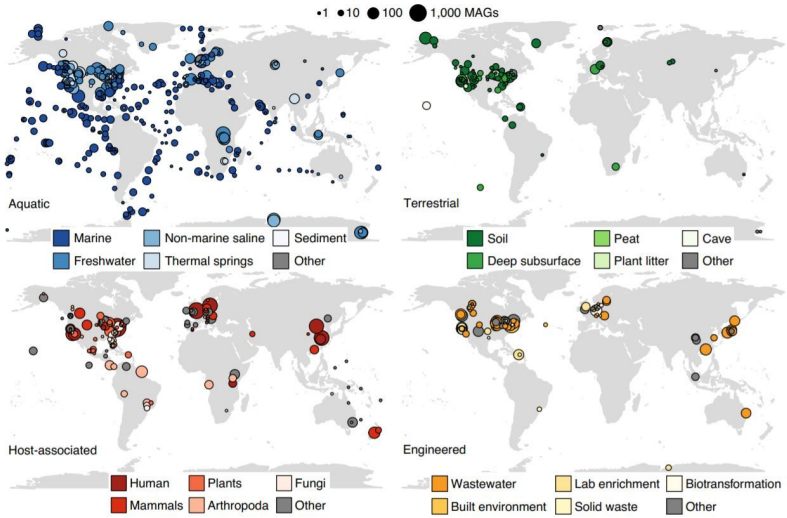
- If you co-assemble (meaning concatenate the raw reads) or co-bin (concatenate assemblies and then bin) then probably one MAG per “group” assembled
  - Only possible when samples are extremely similar

  ⇒ Depending on the purpose of your study

How MAGs advance taxonomy

MAGs have led to the discovery of novel deep-branching lineages previously eluding cultivation-based approaches

- Asgardarchaeota
- Bacterial Candidate Phyla Radiation
- 12,000 New Species of Microbes



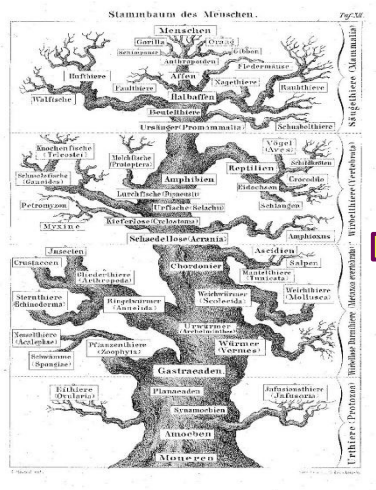
Nayfach, S., Roux, S., Seshadri, R., Udwary, D., Varghese, N., Schulz, F., Wu, D., Paez-Espino, D., Chen, I.M., Huntemann, M. and Palaniappan, K., 2021. A genomic catalog of Earth's microbiomes. *Nature biotechnology*, *39*(4), pp.499-509.

Taxonomic identification

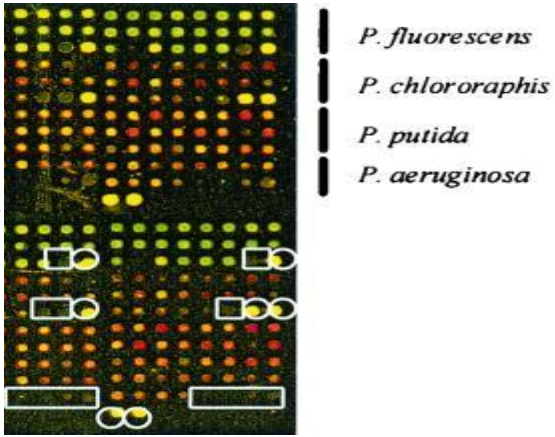
Traditional approach

Initially: Phenotypic observations

Later: DNA:DNA hybridization



Haeckel (1884) *Anthropogenie oder Entwickelungsgeschichte des Menschen. Gemeinverständliche wissenschaftliche Vorträge über die Grundzüge der menschlichen Keimes-*



J.-C. Cho, J. M. Tiedje (2001)
10.1128/AEM.67.8.3677–3682.2001

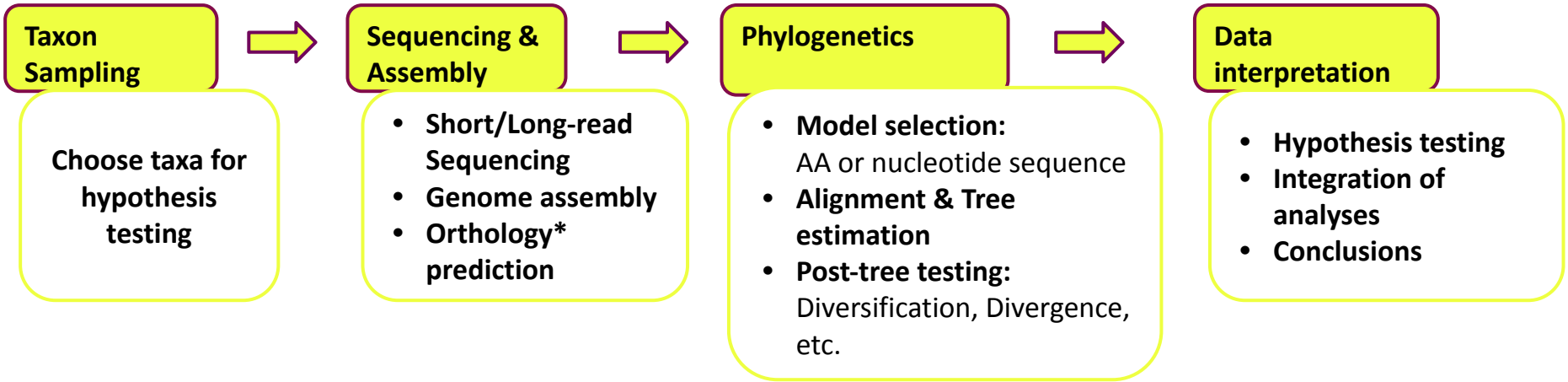
Present-Day approach

- Computational classification of taxonomy: **Phylogenomics**
- Based on **sequencing data** of partial or complete genomes
- **Nucleotide-based**:
  Average **N**ucleotide **I**dentity
- **Protein-based**:
  Relative **E**volutionary **D**ivergence

Application of Phylogenomics

Phylogenomics combines the fields of genomics and evolution to understand how taxa/genes are related, where they come from and where they might be going.

Phylogenetics: use of individual/small amount of genes, e.g. 16S based tree
Phylogenomics: use of a large number of genes/proteins



*Ortholog: homologous sequence descended from the same ancestral sequence. Divergence due to evolutionary speciation.

Adapted from: Young & Gillung, 2020: https://doi.org/10.1111/syen.12406

Taxonomic assignment

A robust taxonomy is needed to

- accurately describe microbial diversity,
- interpret metagenomic data
- provide a common language for communicating scientific results

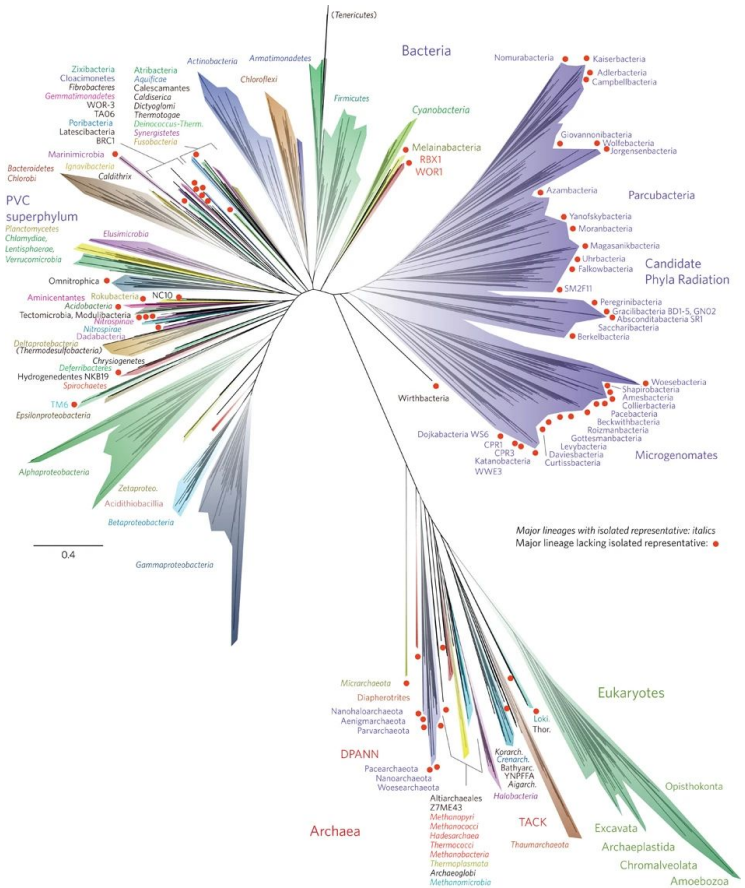
Sequence-based phylogenetic trees provide a framework for the development of a taxonomy that takes into account both evolutionary relationships and differing rates of evolution

Microbial taxonomies based on 16S rRNA gene relationships have several limitations

Protein-based phylogenomics

- Calculation of **R**elative **E**volutionary **D**ivergence/association
- Alignment of concatenated AA sequences
- Calculation of tree using e.g. Maximum Likelihood algorithm
- **Targets:** Genomes and MAGs
- Suitable for any assembly as long as required proteins are present

Step 1: Get AA sequences from marker genes
Step 2: Concatenate all AA sequences



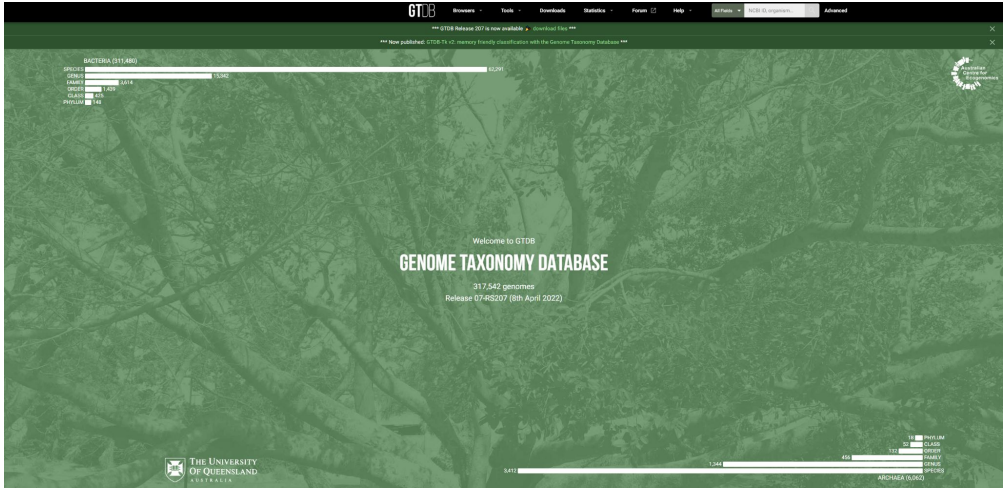
Hug et al., 2016:
https://www.nature.com/articles/nmicrobiol201648

The Genome Taxonomy DataBase “GTDB” taxonomy

- Taxonomic groups in this classification describe monophyletic lineages of similar phylogenetic depth after normalization for lineage-specific rates of evolution

- The GTDB uses relative evolutionary divergence (RED) to delineate higher-rank taxa and average nucleotide identity (ANI) to delineate species clusters

The GTDB developers curate the taxonomy **biannually** since 2017 to incorporate new genomes and proposed taxonomic groups, while retaining a phylogenetically consistent classification



The Genome Taxonomy DataBase “GTDB” taxonomy

Published 2017

- Only 18% of taxon names in the GTDB taxonomy above the rank of species have been validly published
- Further 19% have been proposed but not validated
- The remaining 63% are currently nonstandard placeholder names

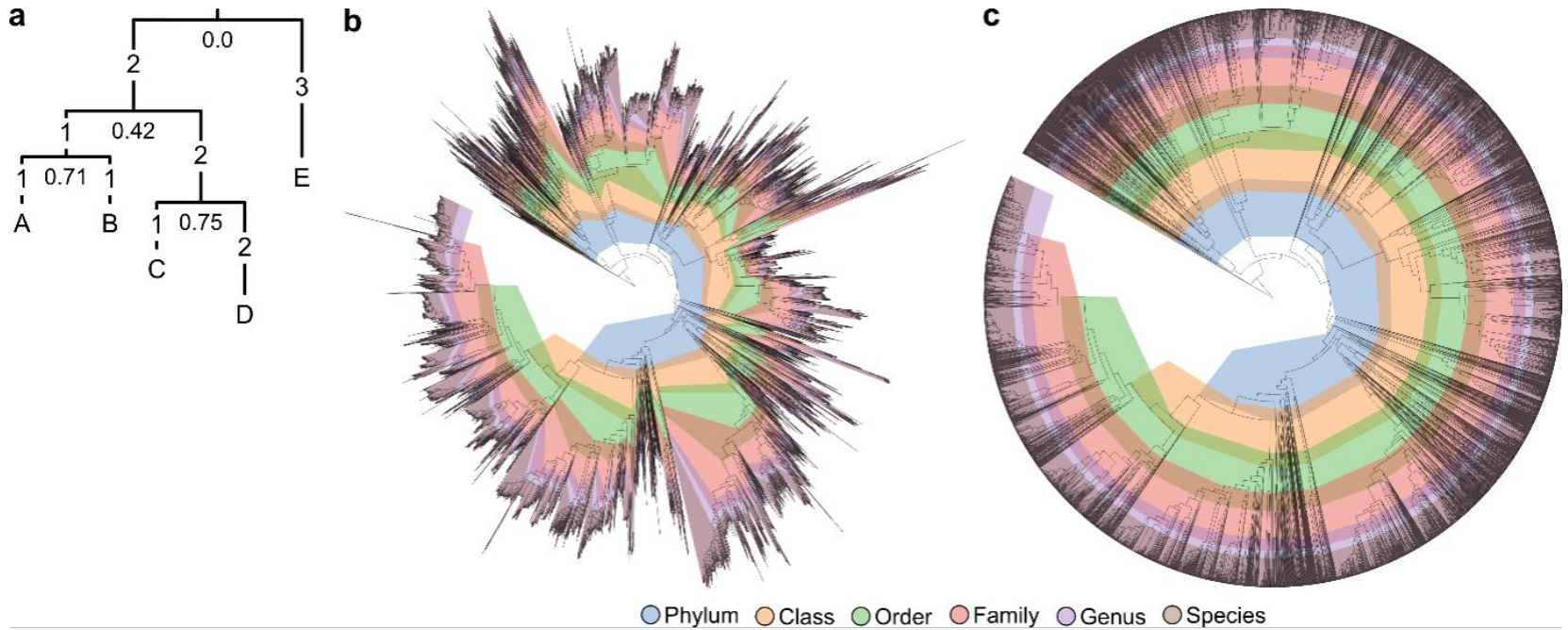
Published 2023 (growth by over 270%, MAGs account for the majority of taxonomic diversity)

- ~ 20.2 of bacterial taxa are MAGs or SAGs of the 402,709 bacterial genomes organized into 85,205 species clusters
- ~70% of archaeal taxa are lacking an available cultured representative

⇒ indicating the scope of the task remaining to produce a fully standardized taxonomy consisting of validated names

Relative Evolutionary Divergence

- Initially no method for higher taxonomic rank assignment:



Parks et al. (2018): doi:10.1038/nbt.4229

Relative Evolutionary Divergence

- Initially no method for higher taxonomic rank assignment:

Phylum:
Class:
Order:
Family:
Genus:
Speci

?

es: ANI

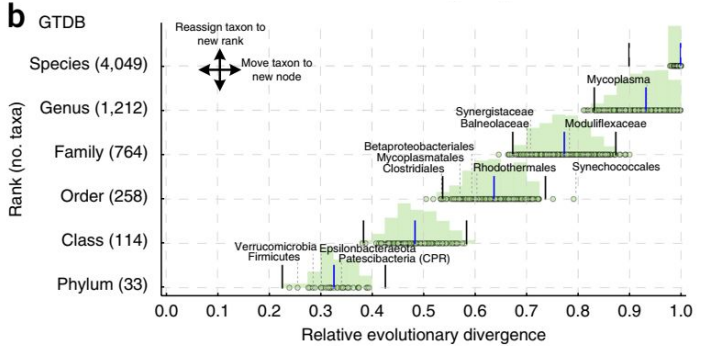
Relative Evolutionary Divergence

Past → **Present**

relative time

- Median RED-values of each taxonomic rank used for taxonomic sorting

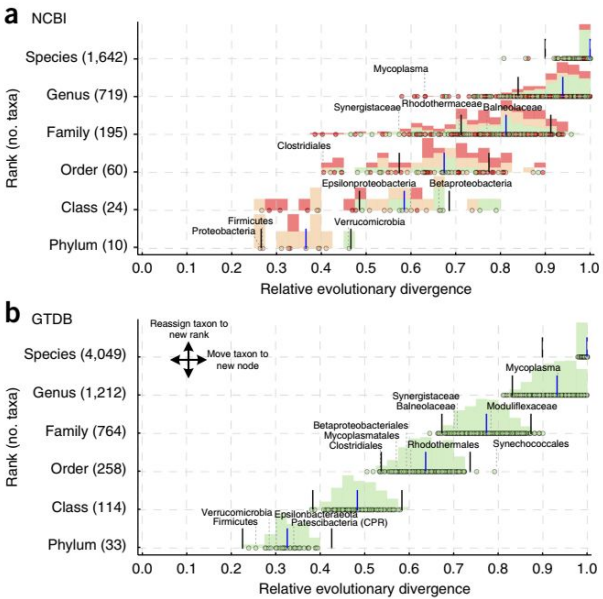
Common ancestor **Extant species**



Parks et al. (2018): doi:10.1038/nbt.4229

The Genome Taxonomy Database: GTDB

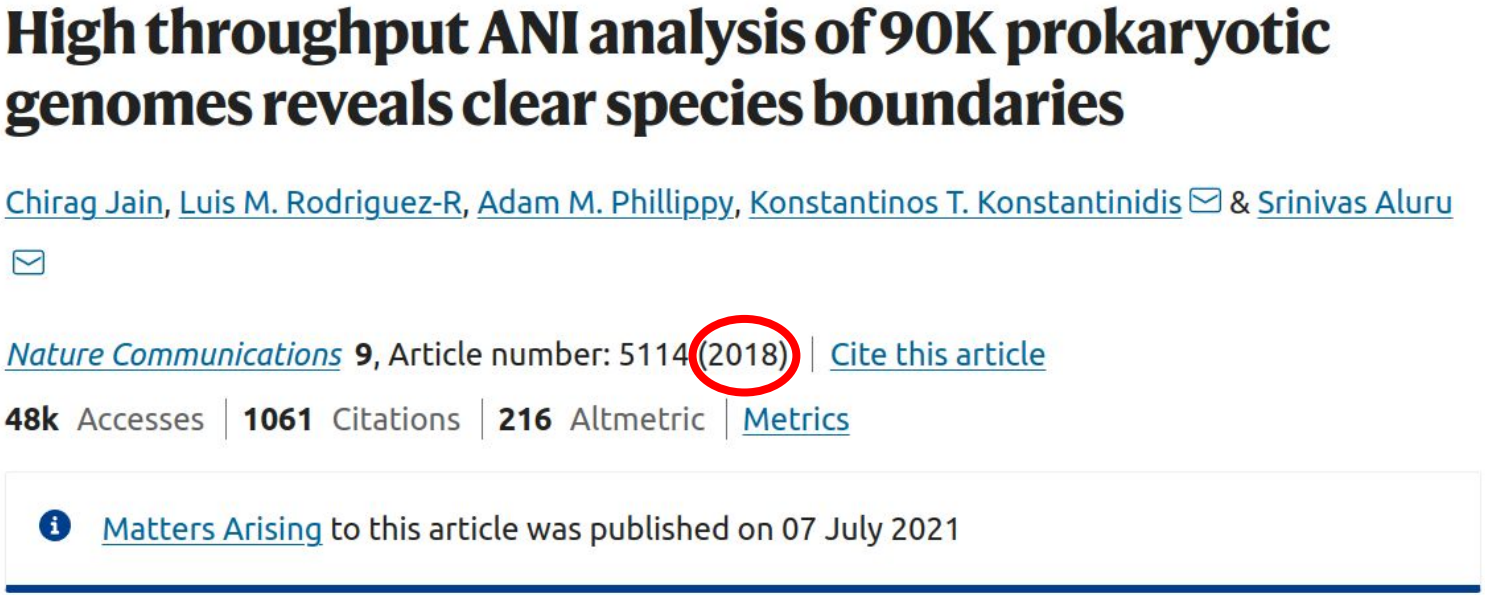
- Effort to create a **unified nomenclature concept**
- GTDB combines complete **genomes** and **MAGs**
- 120 different bacterial and 53 archaeal ubiquitous SCG markers



Taxonomic rank classification higher than species

How to assign species to MAGs?

Is there a continuum of genetic diversity among genomes, or is there a clear species boundaries?



High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries

Chirag Jain, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis & Srinivas Aluru

Nature Communications **9**, Article number: 5114 (2018) | Cite this article

48k Accesses | **1061** Citations | **216** Altmetric | Metrics

Matters Arising to this article was published on 07 July 2021

Whole genome similarity – ANI

- A similarity index between a given **pair of genomes** that can be applicable to prokaryotic organisms
  - **independently of their G+C content**
- Average amount of shared nucleotides between all shared genes (coding regions) of two genomes
- Determination of species-level taxonomy based on **A**verage **N**ucleotide **I**dentity
  - **W**idely used to delineate archaeal and bacterial species
- Computational version of **DNA:DNA** hybridization
- **Complete** and/or **high quality** draft **genomes** necessary
- **Used mostly for cultured representatives**

Species cut-off at ≥95% ANI

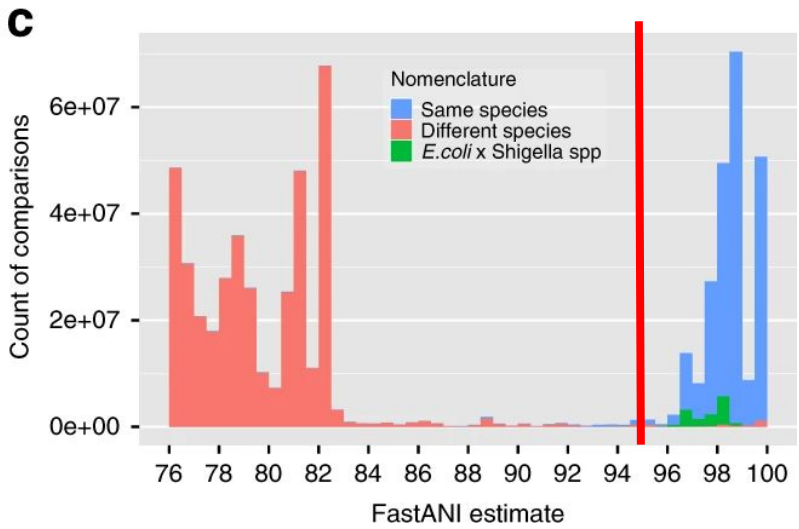


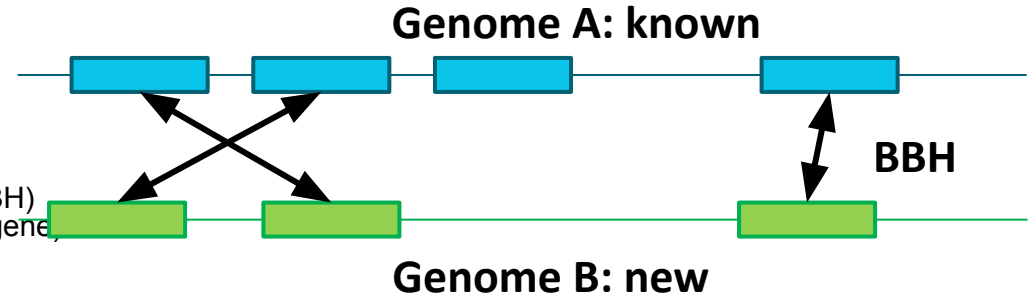
Image: Jain et al. 2018: doi: 10.1038/s41467-018-07641-9

Calculation of Average Nucleotide Identity

Step 1: Sequence similarity search

BLAST: Basic Local Alignment Search Tool

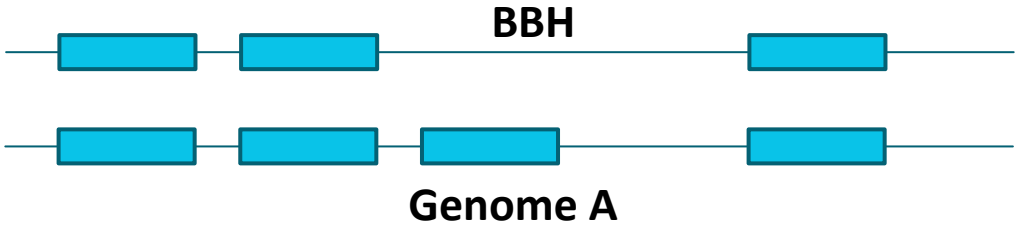
- Creates an alignment based on k-mers
- Alignments are scored and **B**i-directional **B**est **H**its (BBH) kept (> 70% identity & > 70% coverage of the shorter gene)



Step 2: Calculate **A**lignment **F**raction:

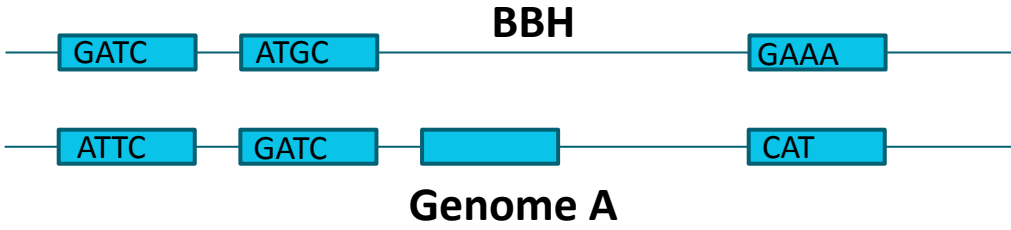
$$AF_{G1 \rightarrow G2} = \frac{\text{lengths of BBH genes}}{\sum \text{length of genes in genome 1}}$$

- Keep only BBHs
- Calculate Ratio:
  BBH-length:length of all genes in A



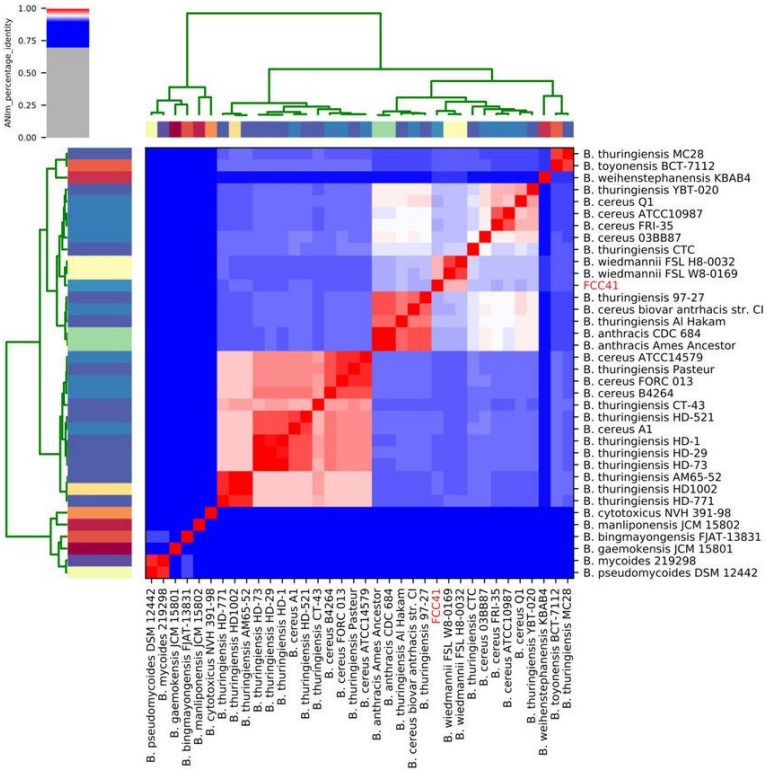
Step 3: Calculate **A**verage **N**ucleotide **I**dentity

$$gANI_{G1 \rightarrow G2} = \frac{\sum_{bbh} (\text{Percent Identity} * \text{Alignment length})}{\text{lengths of BBH genes}}$$



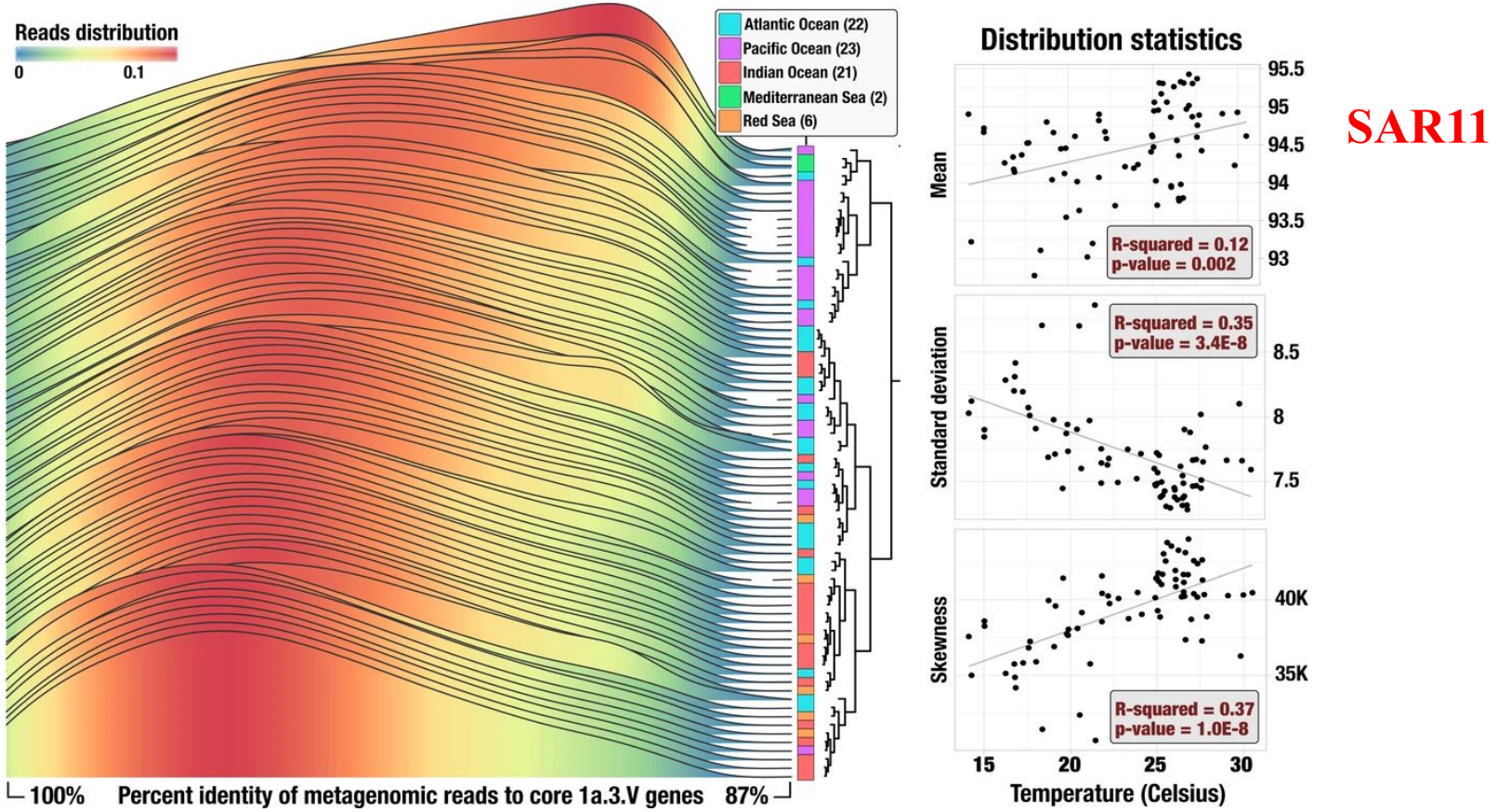
Sequence-based species boundaries

Alignment Fraction: > 70%
Species boundary ANI: ≥ 95%
Identification of novel species



Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T. and Aluru, S., 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature communications*, *9*(1), pp.1-8.

ANI 95% not for most but not all species



SAR11

Delmont, T.O…. and Eren, A.M., 2019. Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. *Elife*, *8*, p.e46497.

GTDB taxonomy assignment

- The bac120 data set represents ~4% of an average bacterial genome and is comparable to other bacterial domain marker sets
- Having inferred the concatenated protein phylogeny, the tree with group names is annotated by using the NCBI taxonomy standardized to seven ranks

Examples:

- **d__**Bacteria; **p__**Cyanobacteria; **c__**; **o__**Nostocales; **f__**Nostocaceae; **g__**Trichormus; **s__**Trichormus azollae
- d__Bacteria;p__Firmicutes;c__Bacilli;**o__ML615J-28;f__CAG-698;g__DTU067;s__DTU067 sp001512995**
- d__Bacteria;p__Firmicutes_A;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__UBA3947;**s__**

Alphanumeric names nonstandard placeholders are to be replaced with standard validated names in due course

GTDB taxonomy assignment - Not Perfect!

- According to the GTDB release 207v2 and release 214, *Methanosarcina* belongs to *Halobacteriota*

  d__Archaea;p__Halobacteriota;c__Methanosarcinia;o__Methanosarcinales;f__Methanosarcinaceae;g__Methanosarcina;s__Methanosarcina flavescens

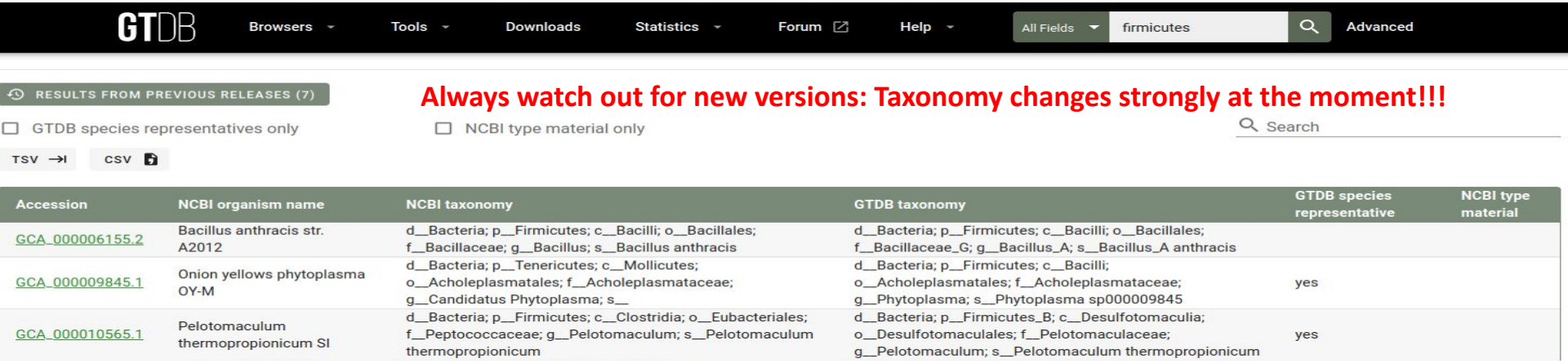
- no Euryarchaeota phylum on GTDB

- according to NCBI and encyclopedia of life, Seqcode and LPSN Methanosarcina is classified as Archaea;Euryarchaeota;Methanomicrobia;Methanosarcinales;Methanosarcinacaea

- according to the latest tree of life, *Halobacteria* is a genus closely related to *Methanomicrobia* and *Methanosarcina*

Taxonomy is a bit of a mess….

- Different databases use varying nomenclature
- **Metagenomics** gold-standard is the **GTDB** (genome taxonomy database) and the **SeqCode***
- **Isolate** species are recognized by The Bacterial Code (**ICNP**)



Accession	NCBI organism name	NCBI taxonomy	GTDB taxonomy	GTDB species representative	NCBI type material
GCA_000006155.2	Bacillus anthracis str. A2012	d__Bacteria; p__Firmicutes; c__Bacilli; o__Bacillales; f__Bacillaceae; g__Bacillus; s__Bacillus anthracis	d__Bacteria; p__Firmicutes; c__Bacilli; o__Bacillales; f__Bacillaceae_G; g__Bacillus_A; s__Bacillus_A anthracis		
GCA_000009845.1	Onion yellows phytoplasma OY-M	d__Bacteria; p__Tenericutes; c__Mollicutes; o__Acholeplasmatales; f__Acholeplasmataceae; g__Candidatus Phytoplasma; s__	d__Bacteria; p__Firmicutes; c__Bacilli; o__Acholeplasmatales; f__Acholeplasmataceae; g__Phytoplasma; s__Phytoplasma sp000009845	yes	
GCA_000010565.1	Pelotomaculum thermopropionicum SI	d__Bacteria; p__Firmicutes; c__Clostridia; o__Eubacteriales; f__Peptococcaceae; g__Pelotomaculum; s__Pelotomaculum thermopropionicum	d__Bacteria; p__Firmicutes_B; c__Desulfotomaculia; o__Desulfotomaculales; f__Pelotomaculaceae; g__Pelotomaculum; s__Pelotomaculum thermopropionicum	yes	

*SeqCode: https://doi.org/10.1038/s41564-022-01214-9

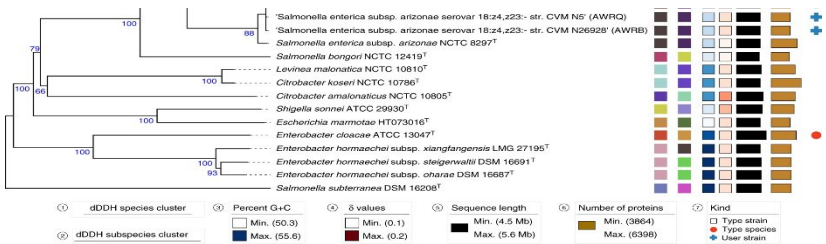
External comparisons – Online resources for comparison

TYGS

- Type strain database (DSZM)*
- Phylogenetic trees based on *Genome BLAST Distance Phylogeny*
- Additional: 16S rRNA comparison
- **Based on cultured microorganisms!**

JSpeciesWs

- Curated non-redundant genome database including: ENA, GenBank (NCBI) and DNA Database of Japan*



*https://tygs.dsmz.de/; Image: https://www.nature.com/articles/s41467-019-10210-3; https://jspecies.ribohost.com/jspeciesws/#home

Phylogenomic comparison

Species boundary 95%

Identification of novel taxa (high resolution)

Most reliable on complete and HQ Genomes

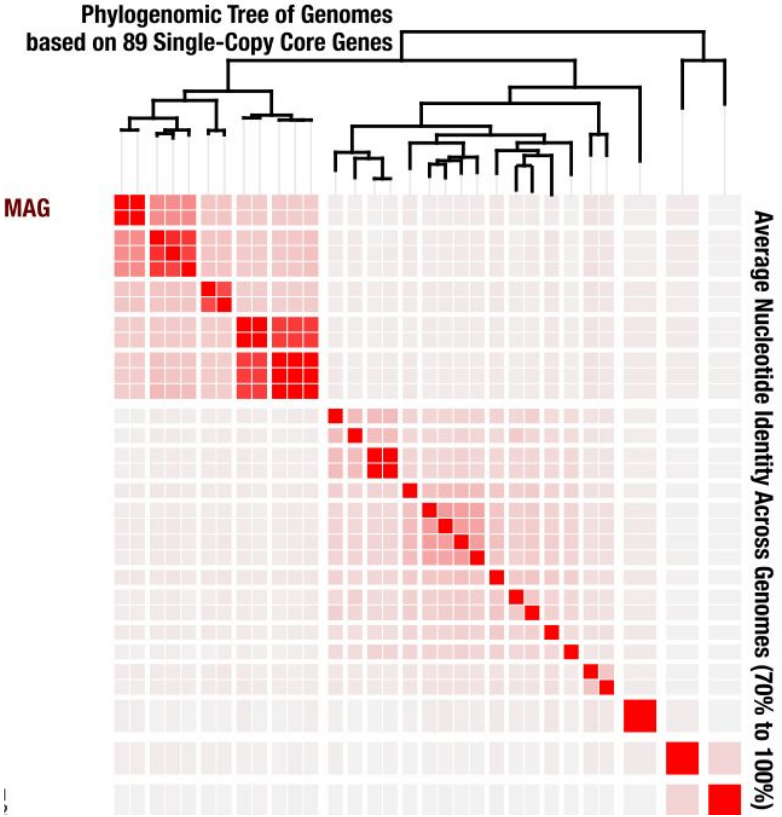


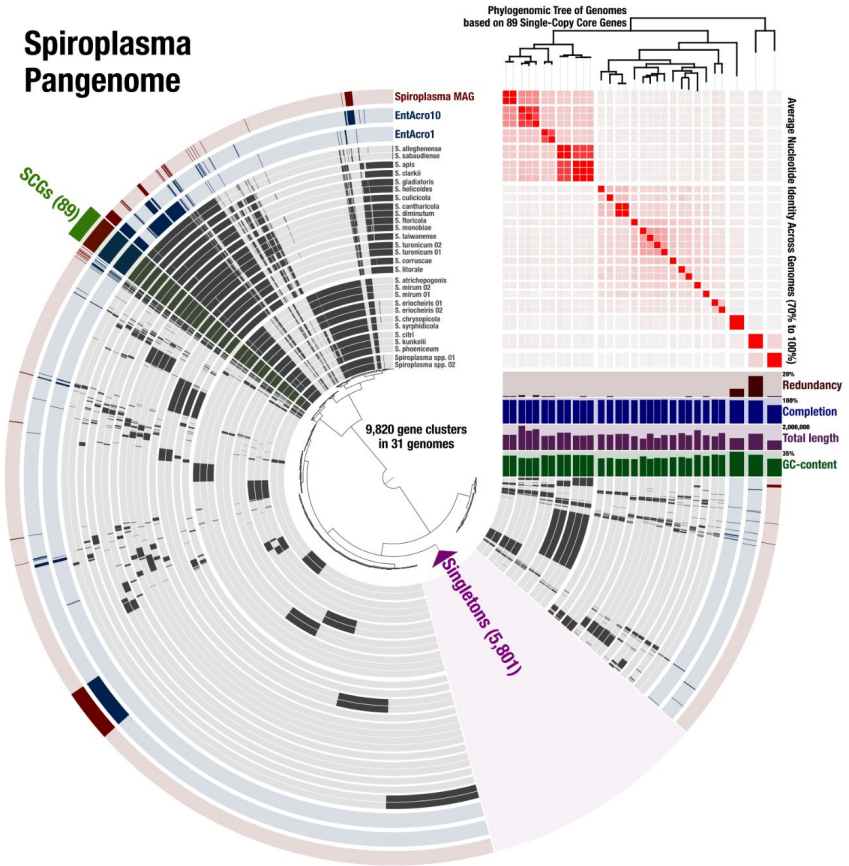
Image:
https://merenlab.org/data/spiroplasma-pangenome/

Evolutionary divergence over time

Based on conserved proteins

Phylogeny of MAGs and Genomes

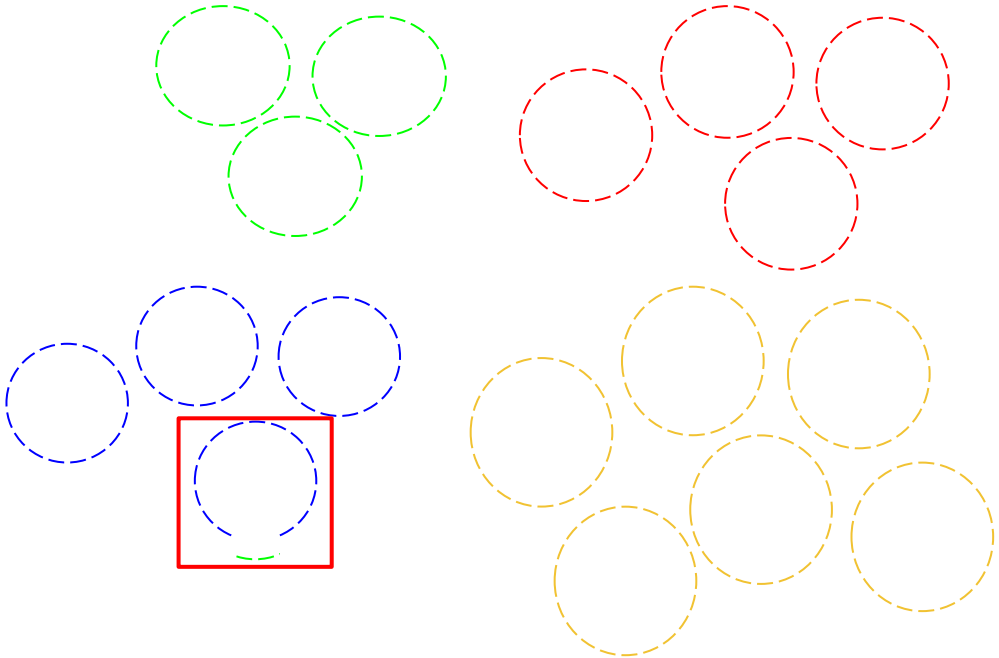
Pan- and Phylogenomics



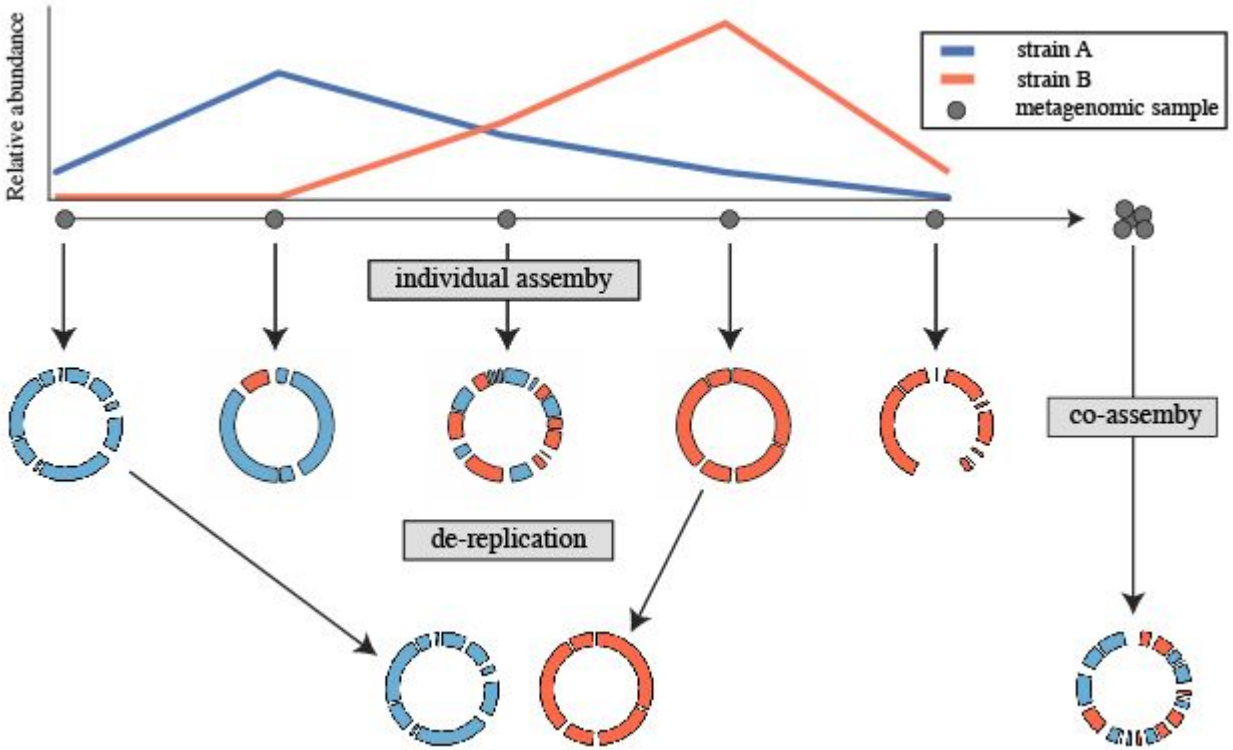
Pangenomics: Next Week

Image: https://merenlab.org/data/spiroplasma-pangenome/

Sorting MAGs based on ANI



Dereplication



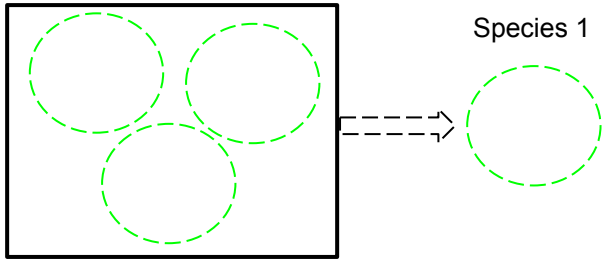
Olm, M.R., Brown, C.T., Brooks, B. and Banfield, J.F., 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *The ISME journal*, *11*(12), pp.2864-2868.

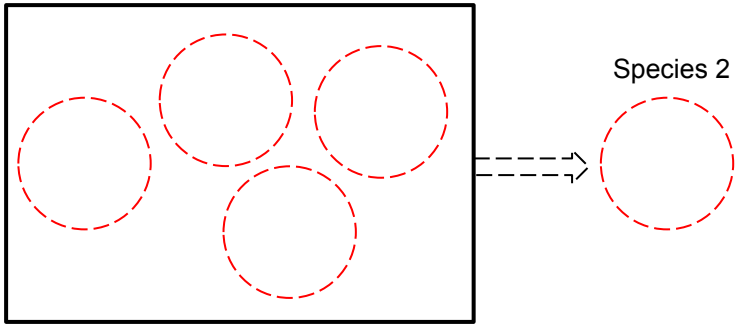
Cluster representatives

Dereplicating MAGs at 95% identity and 80% AF

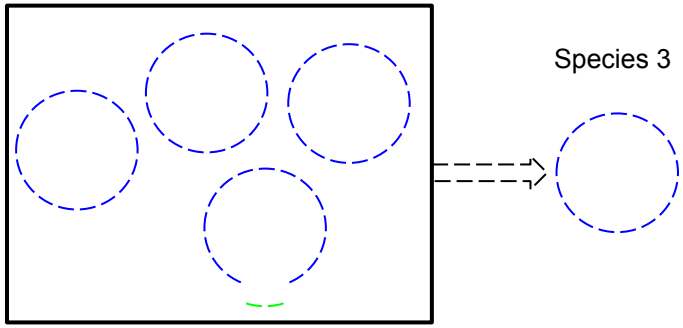
- Species level dereplication
- MAG with the best quality is chosen as the species cluster representative
- If complete genome available in the dataset, it would be preferred to a MAG

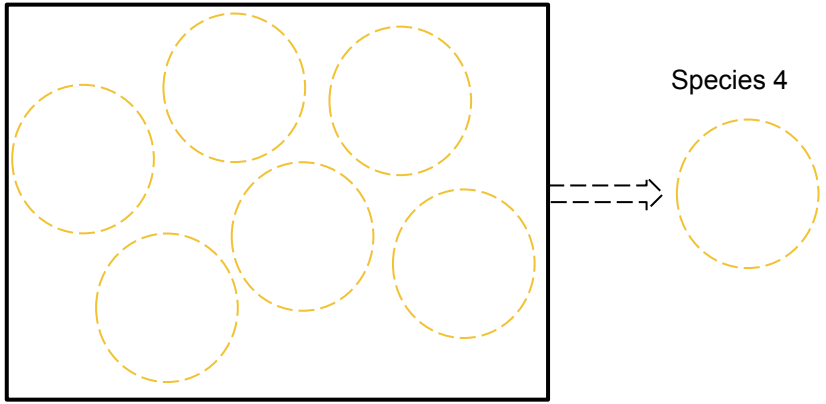
MAGs dereplication



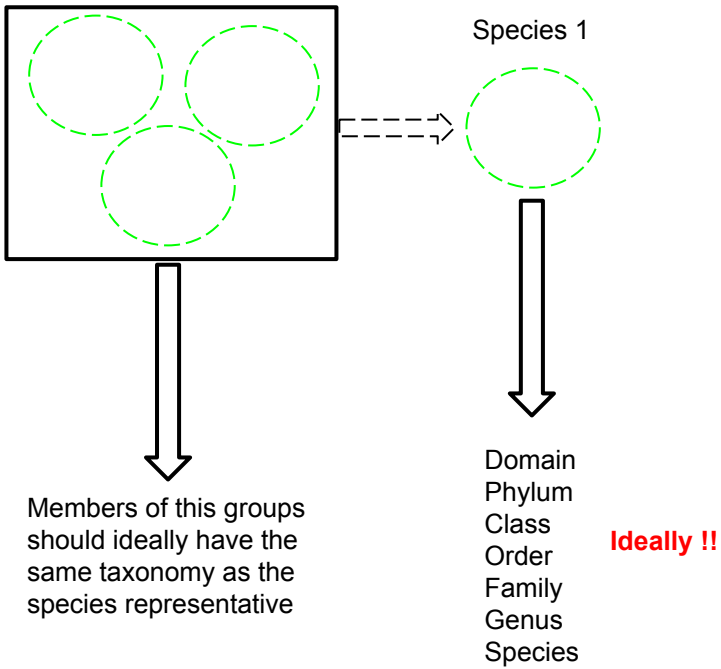


Best Quality MAG ⇒ species representative !!

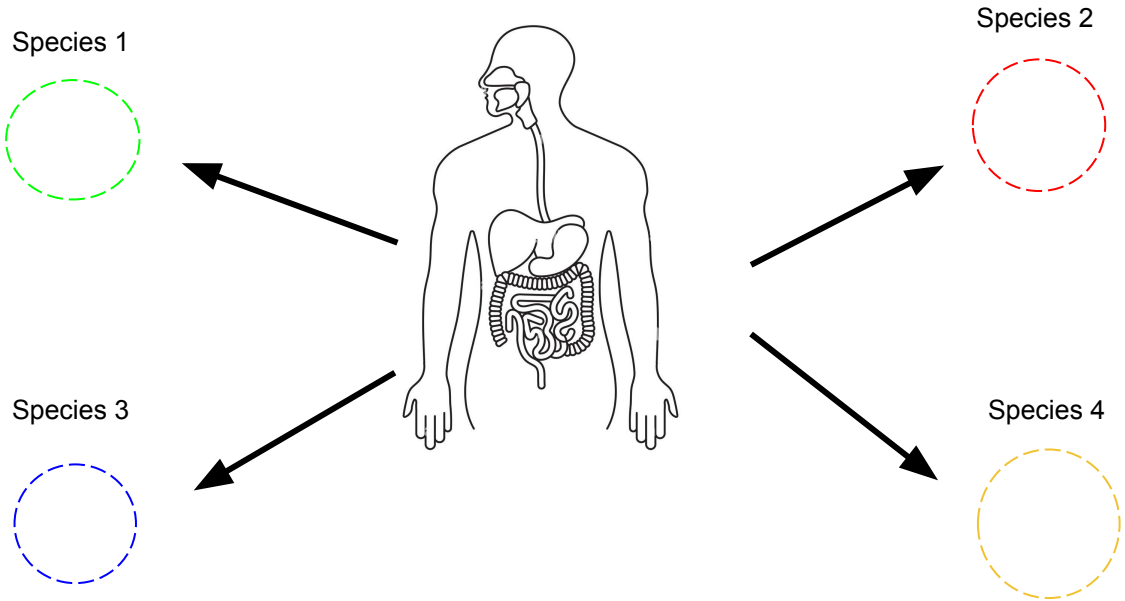




Taxonomic assignment



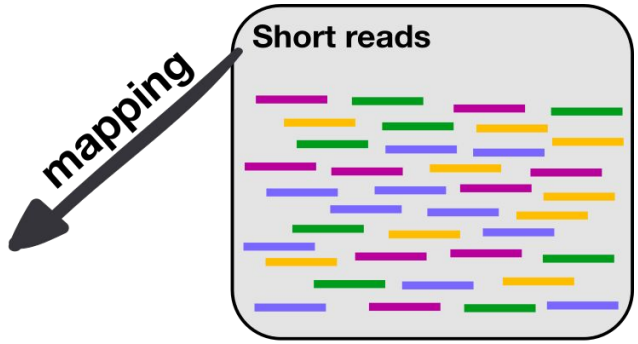
Abundance of a species in a sample per sampling point



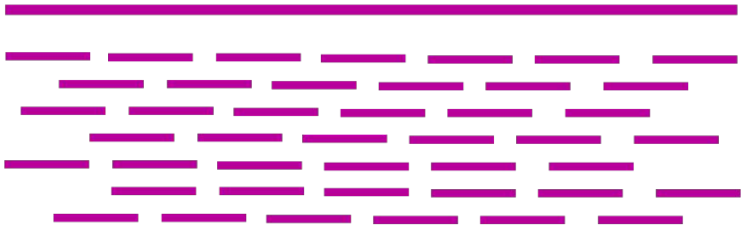
These classified species are not occurring once in a sampled environment.
How can we estimate their abundance???

Read recruitment contig basis

Remember!! Basis for binning



Contig[1]

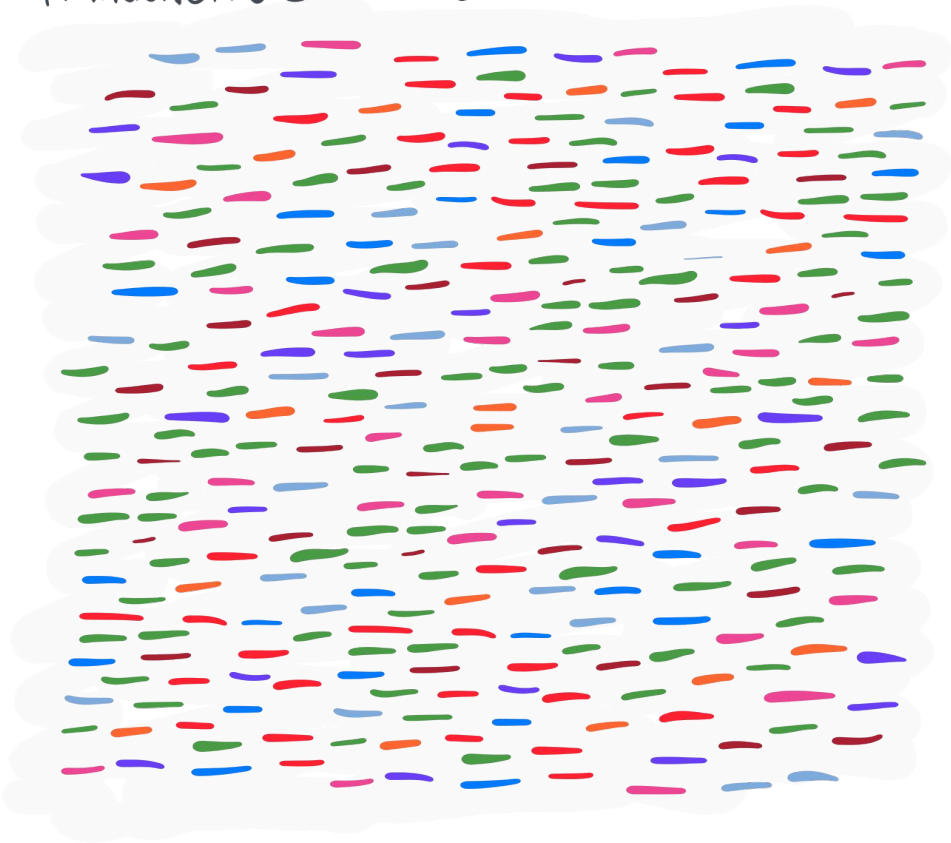


Coverage: ~7X

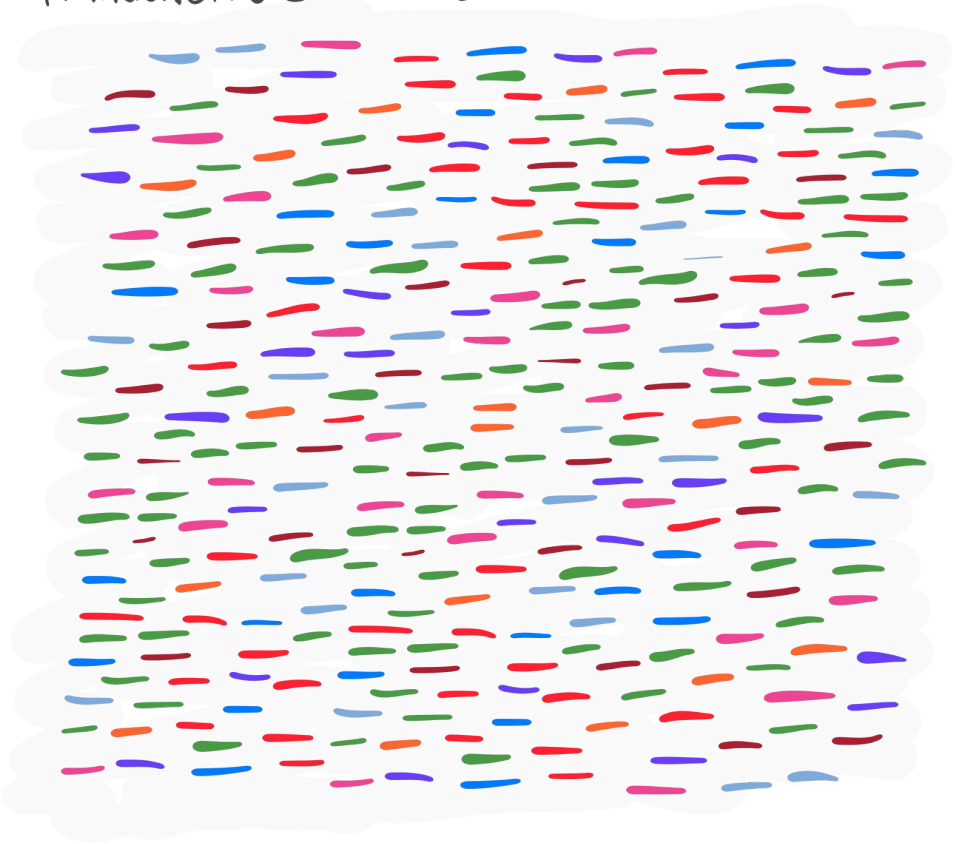
Contig[2]

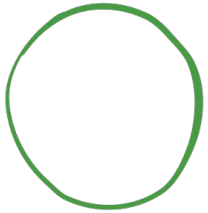
Coverage: ~4X

METAGENOMIC SHORT READS

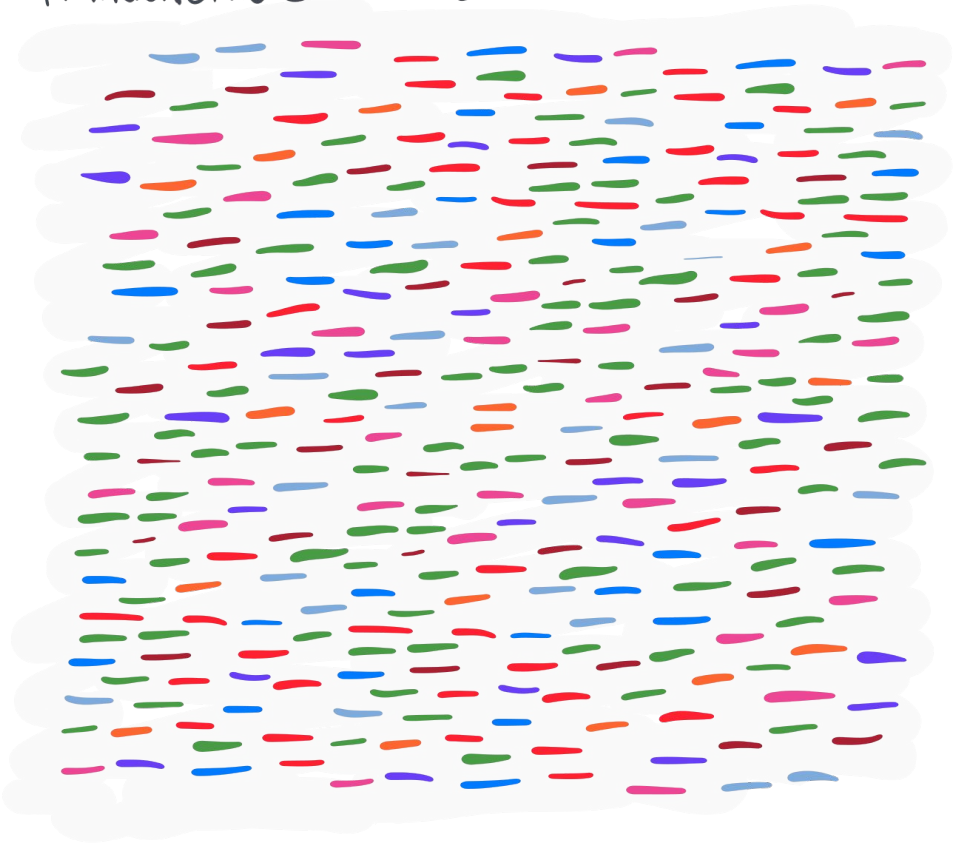


METAGENOMIC SHORT READS

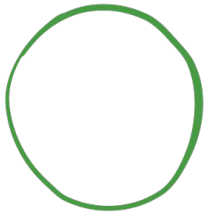




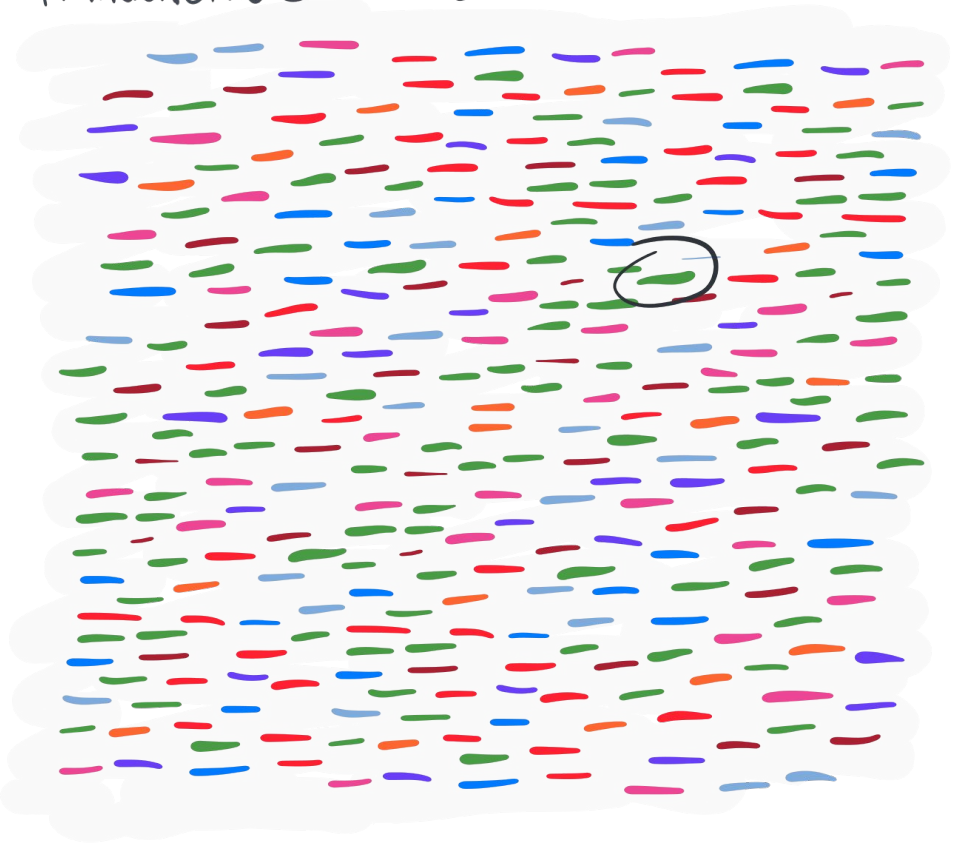
METAGENOMIC SHORT READS



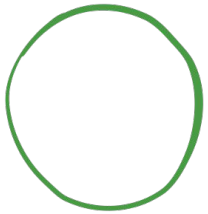
READ RECRUITMENT

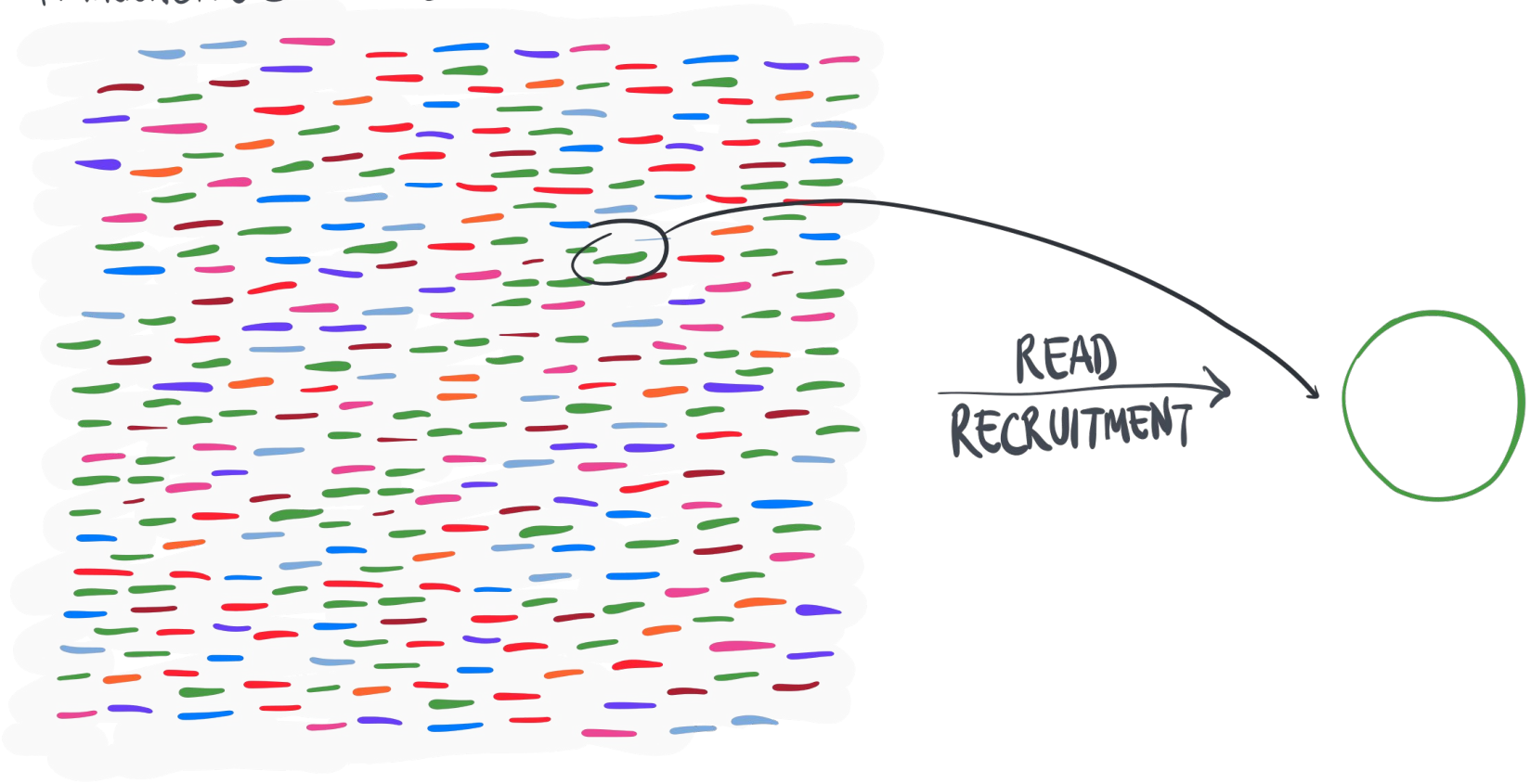


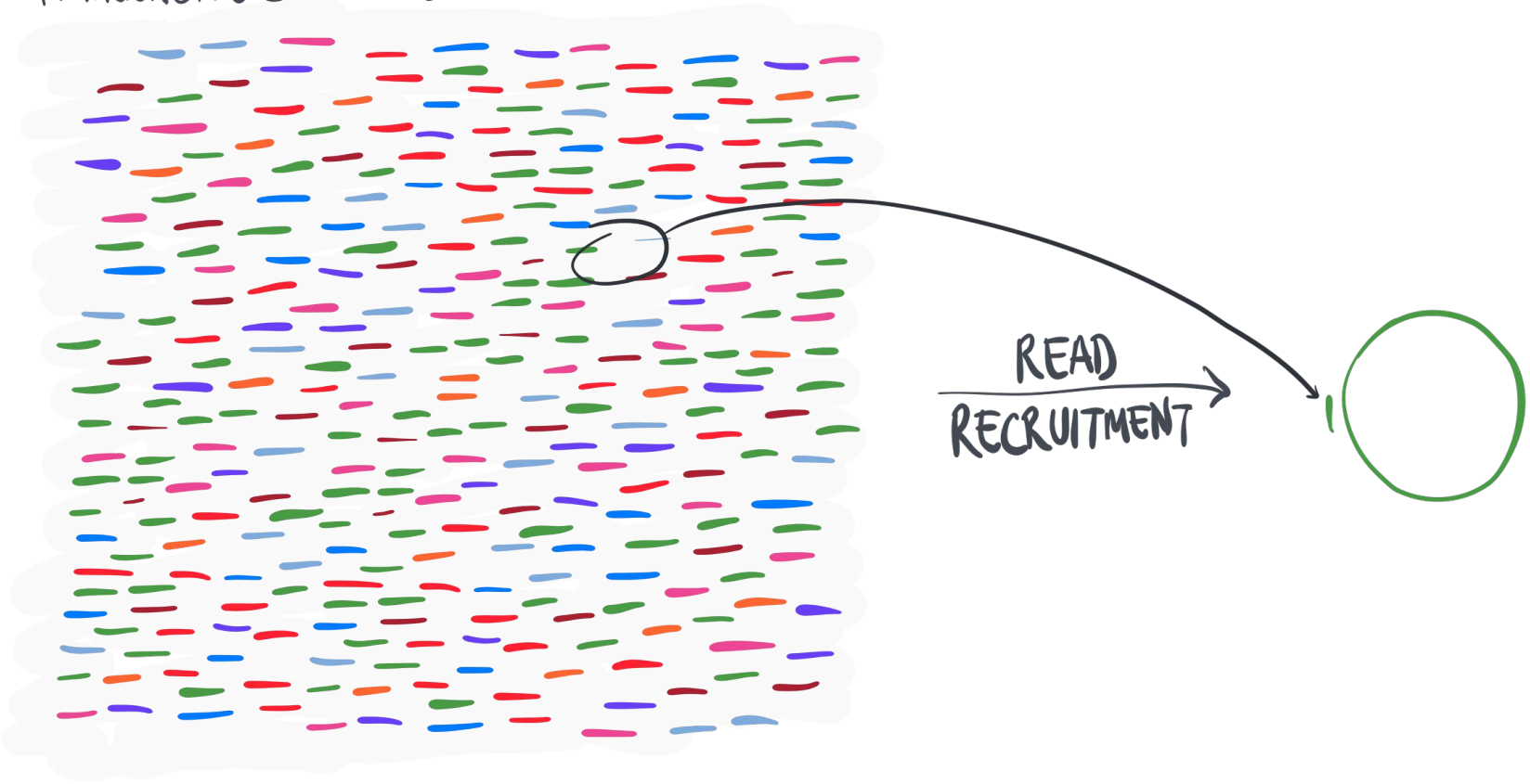
METAGENOMIC SHORT READS



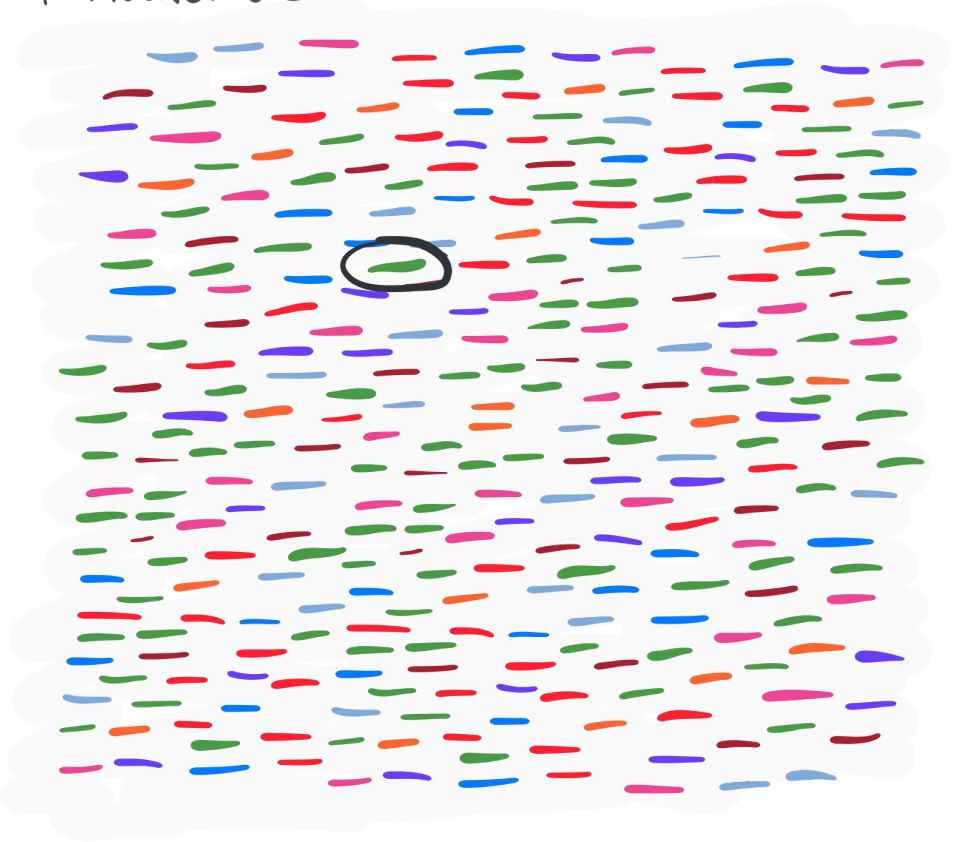
READ RECRUITMENT



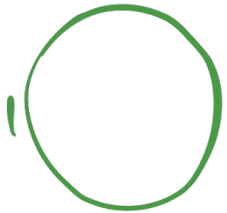

METAGENOMIC SHORT READS
READ
RECRUITMENT

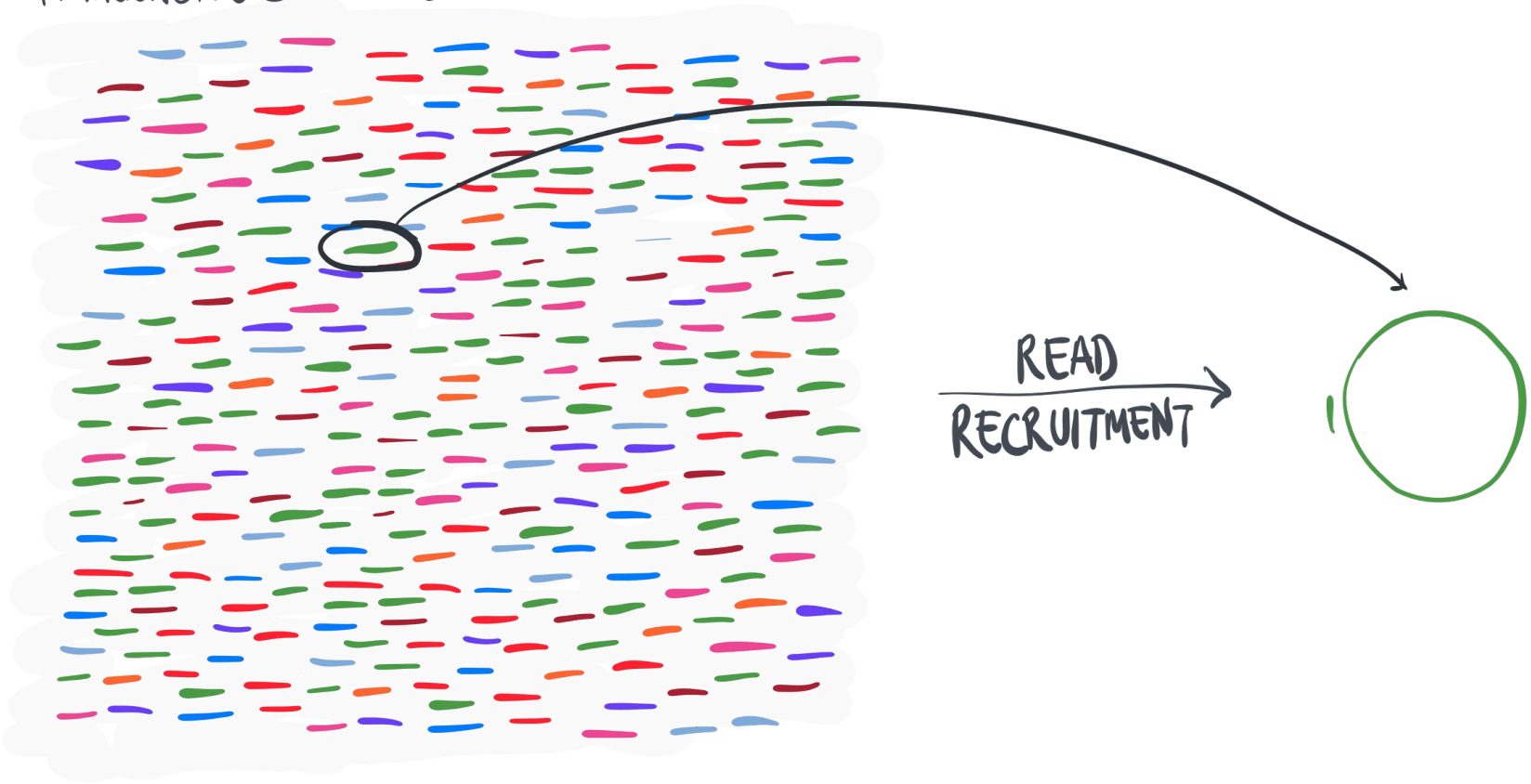

METAGENOMIC SHORT READS
READ
RECRUITMENT

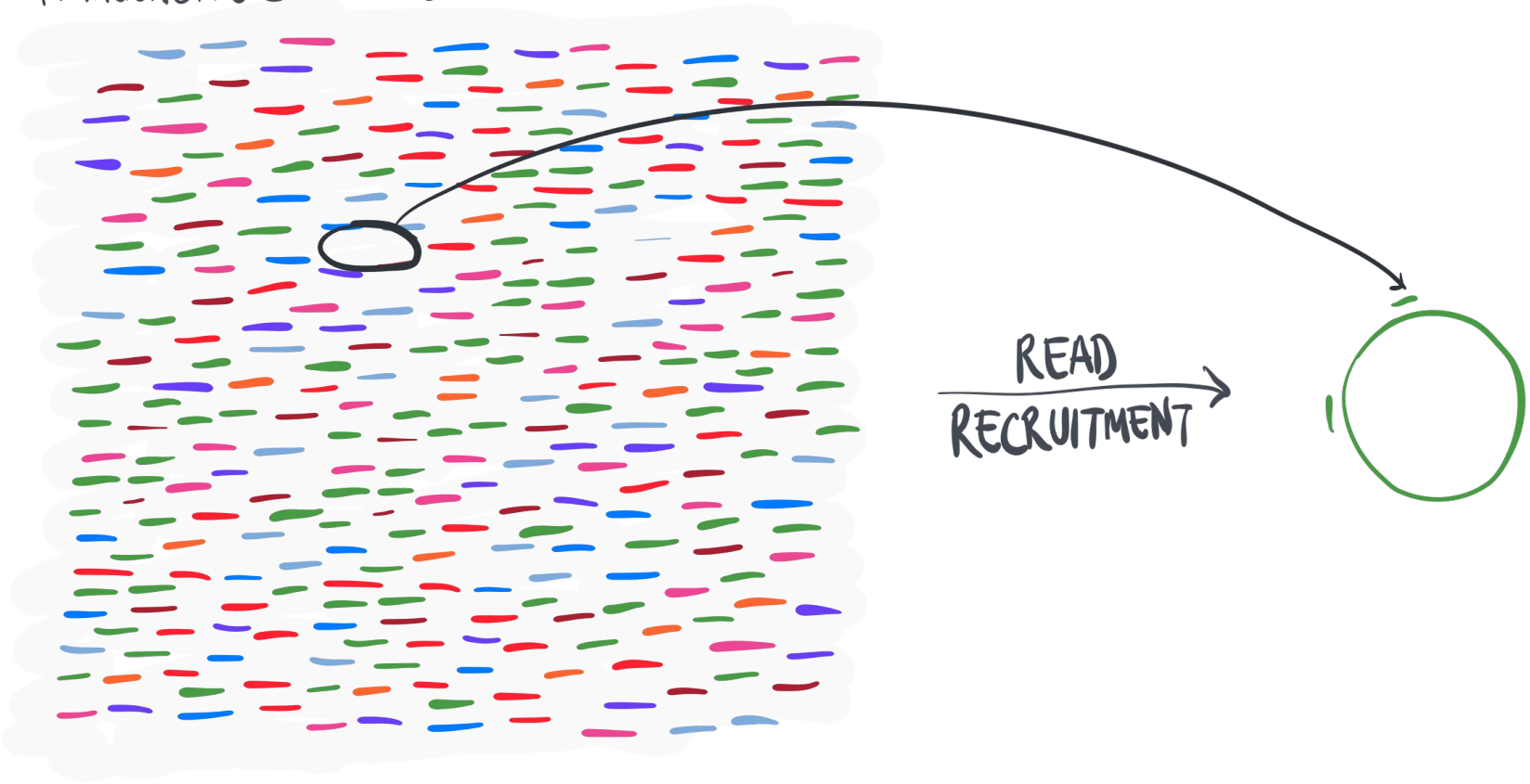
METAGENOMIC SHORT READS



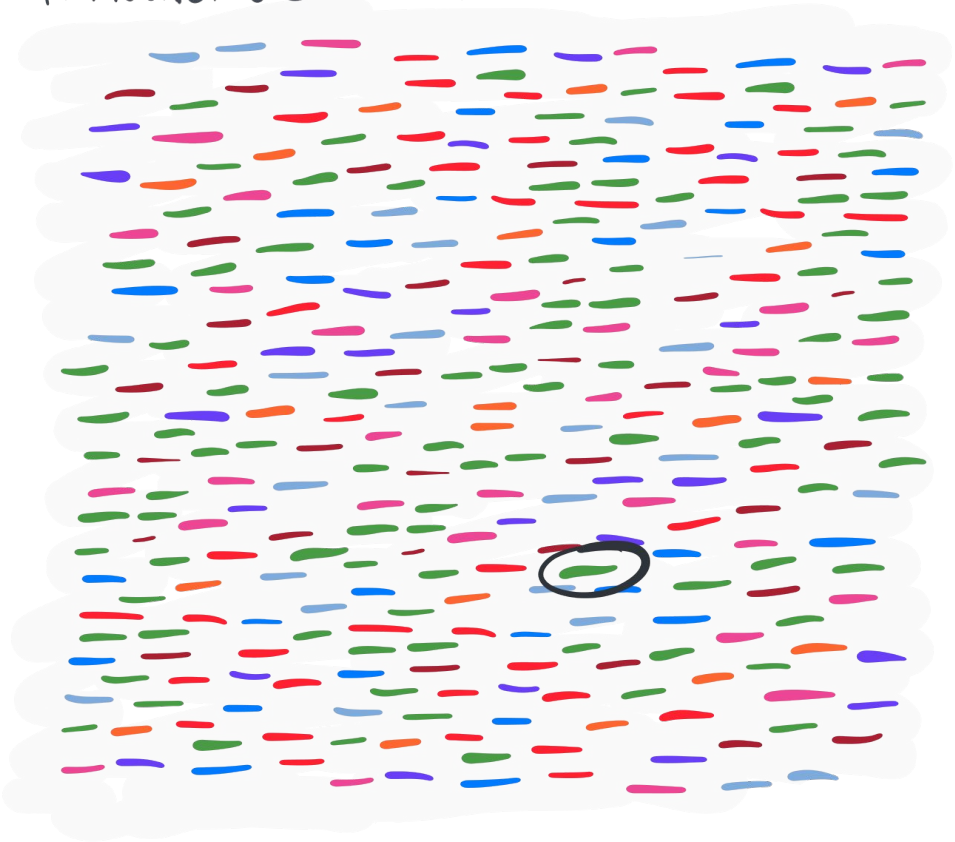
READ RECRUITMENT



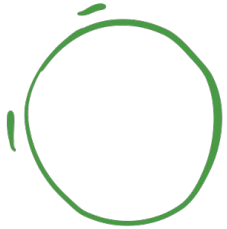

METAGENOMIC SHORT READS
READ
RECRUITMENT

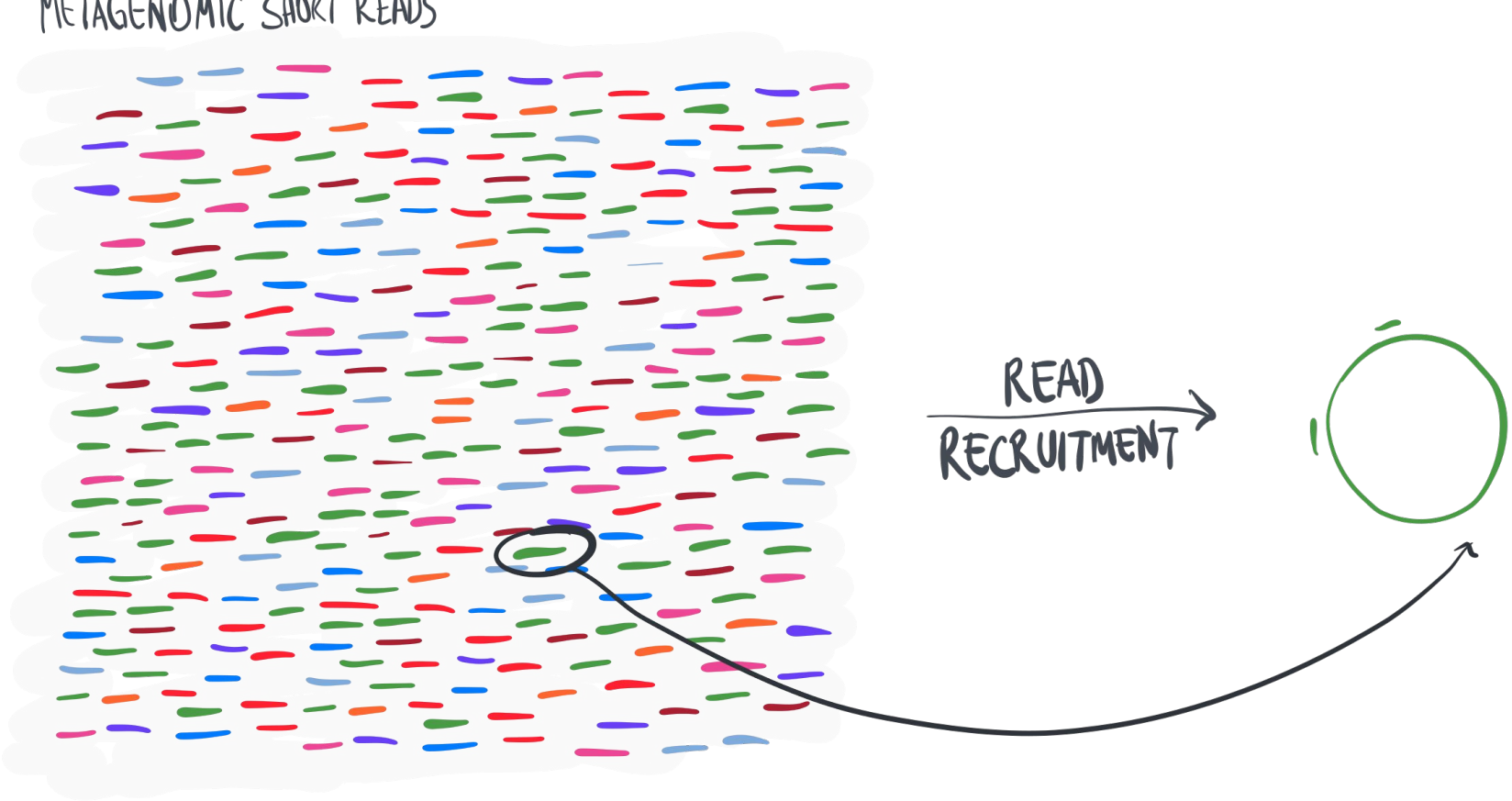

METAGENOMIC SHORT READS
READ
RECRUITMENT

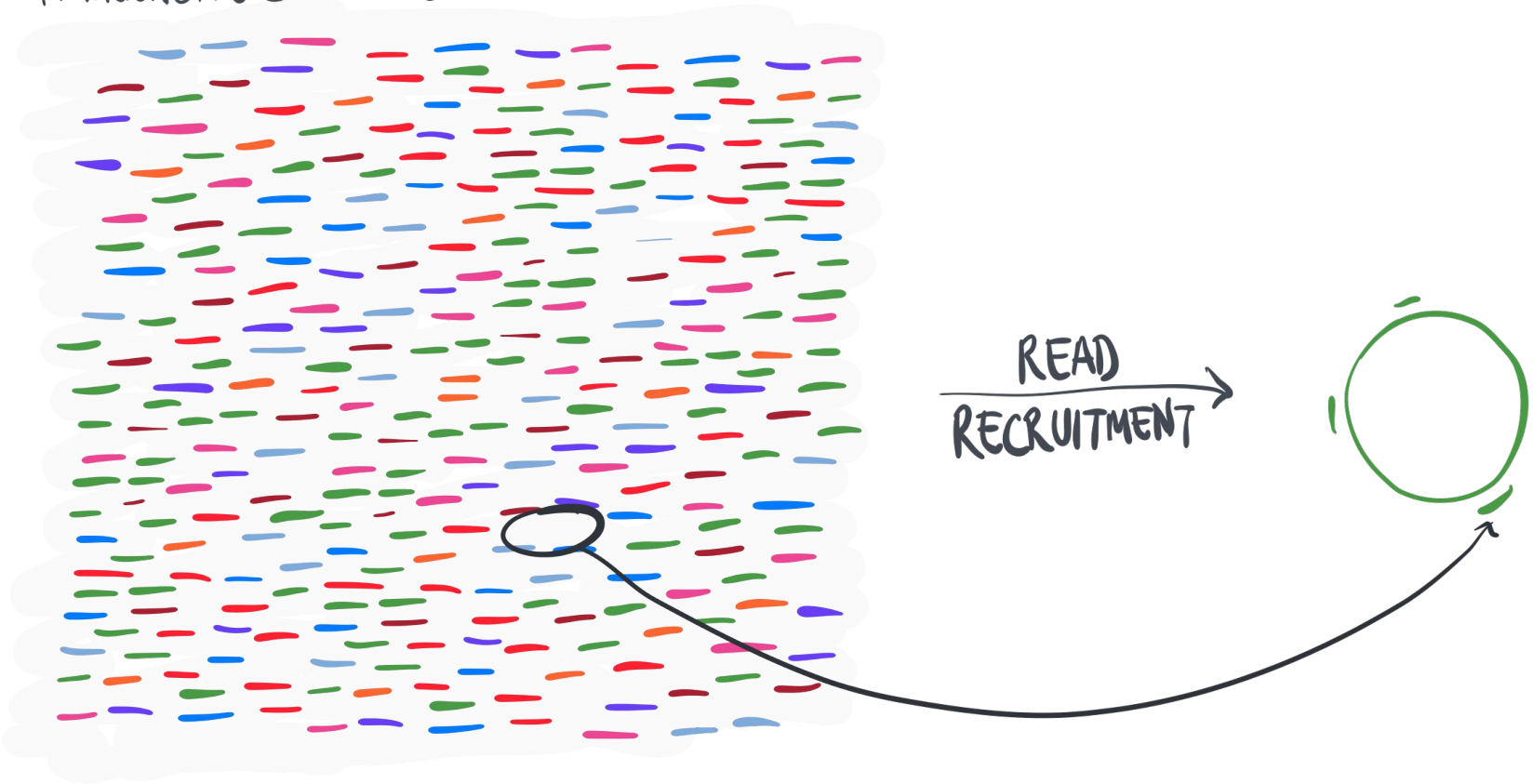
METAGENOMIC SHORT READS



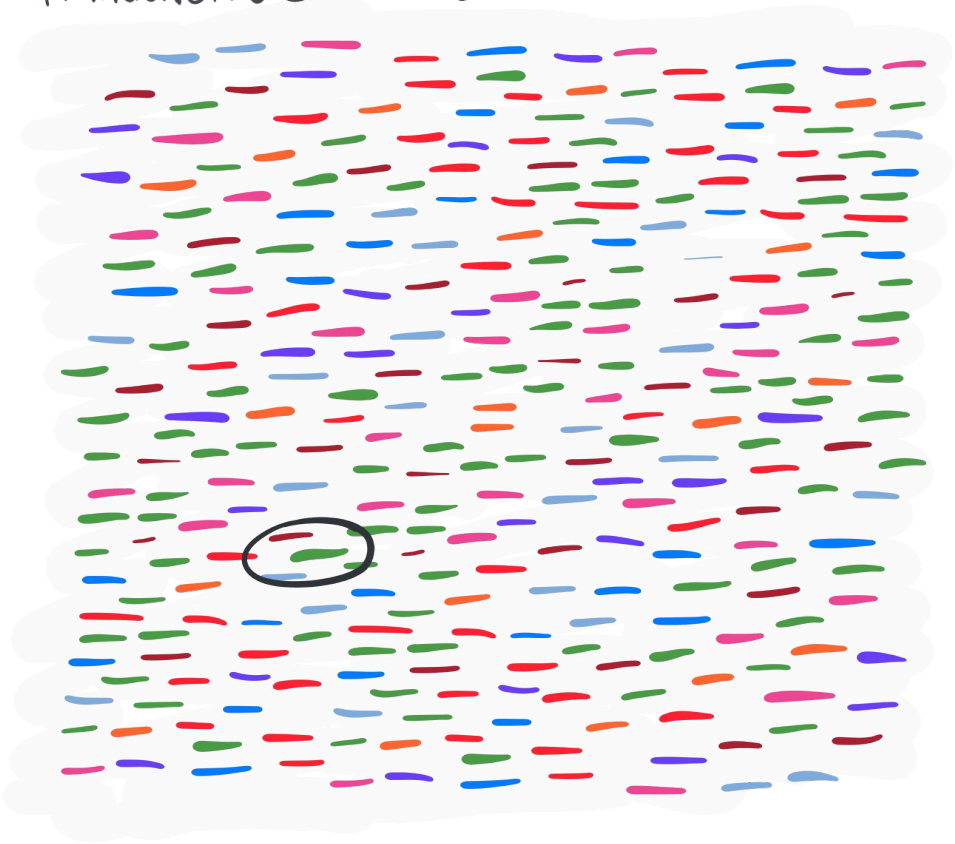
READ RECRUITMENT



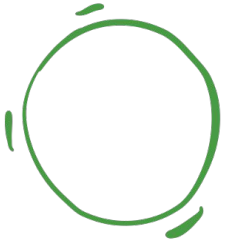

METAGENOMIC SHORT READS
READ
RECRUITMENT

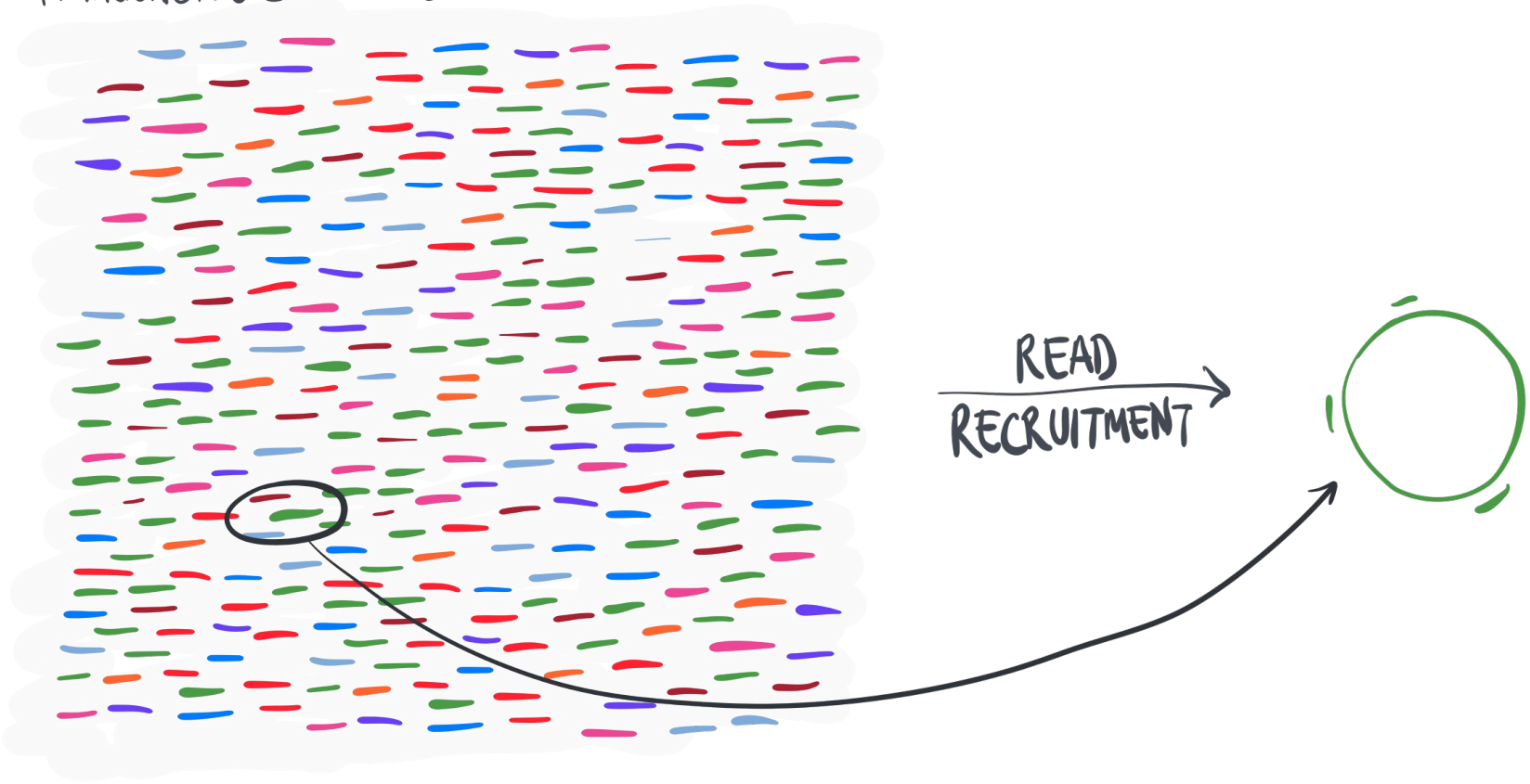

METAGENOMIC SHORT READS
READ
RECRUITMENT

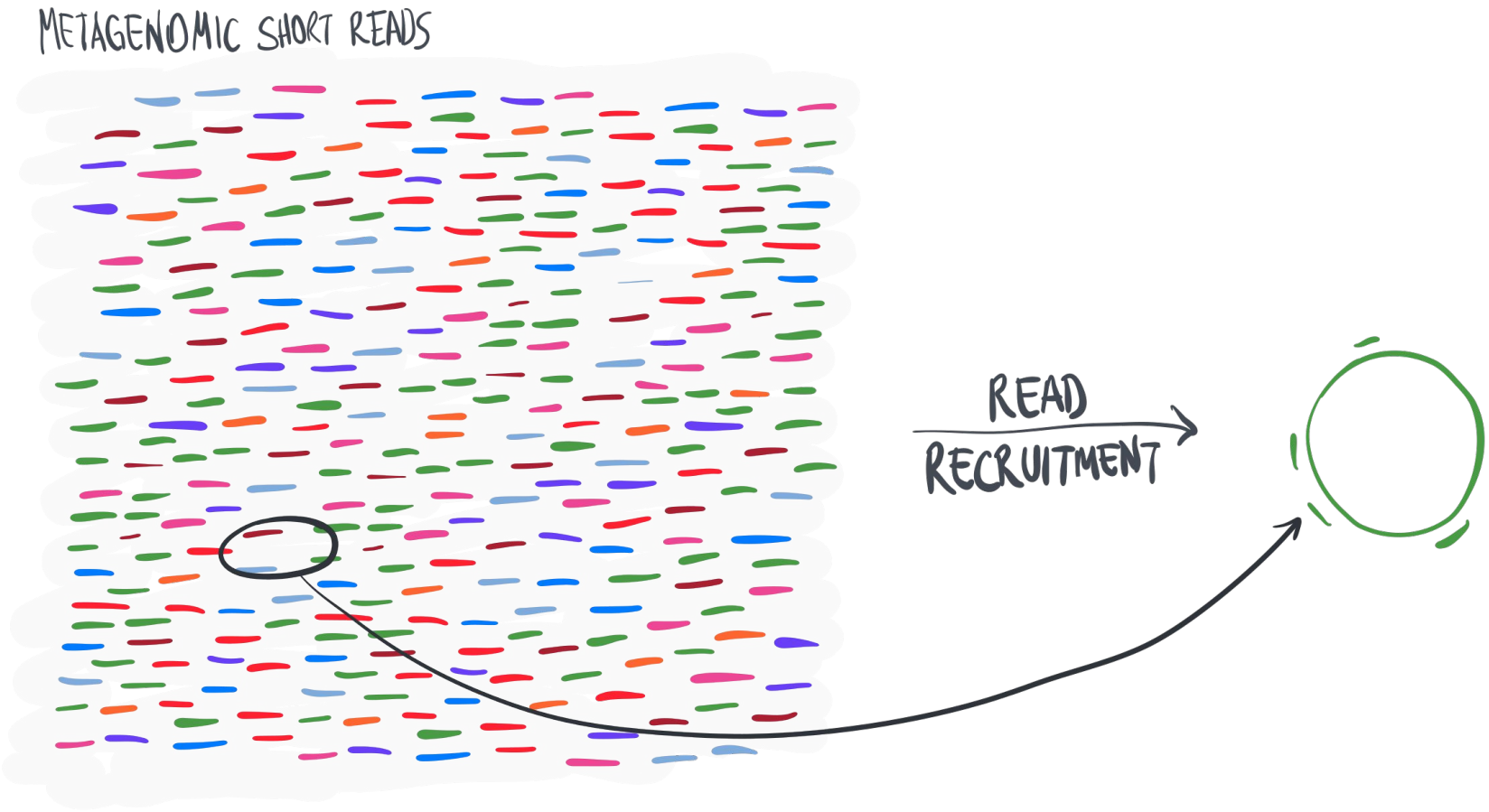
METAGENOMIC SHORT READS



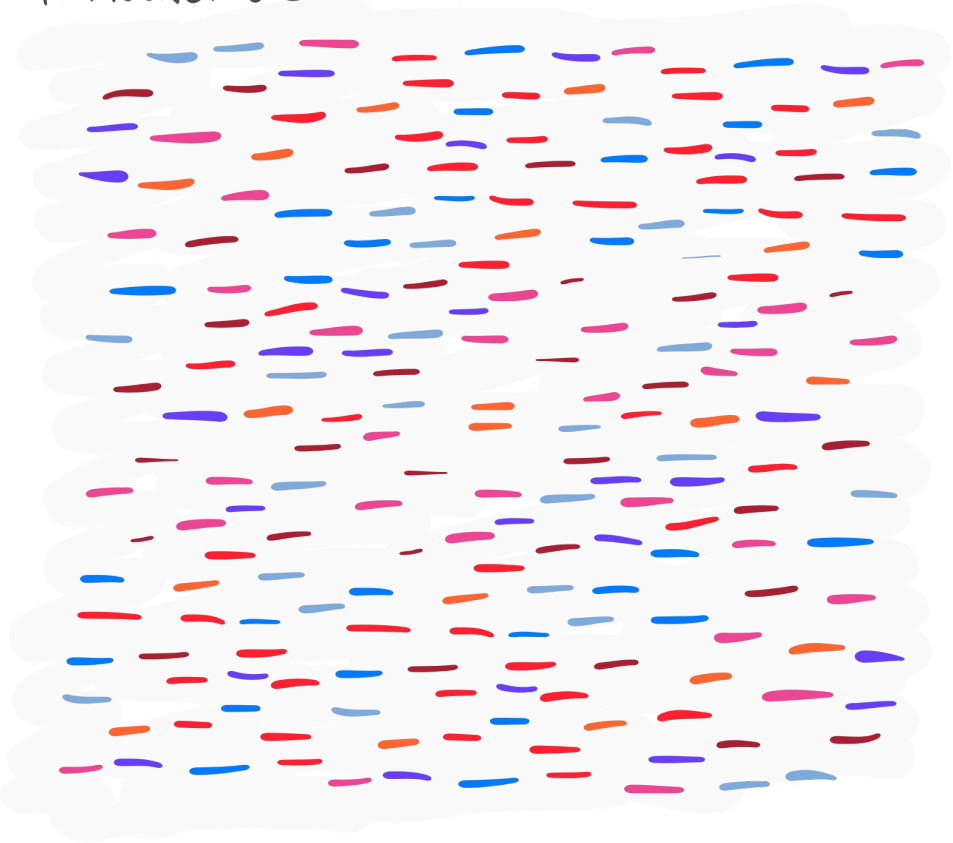
READ RECRUITMENT



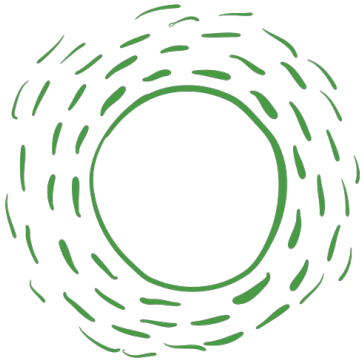

METAGENOMIC SHORT READS
READ
RECRUITMENT


METAGENOMIC SHORT READS
READ
RECRUITMENT

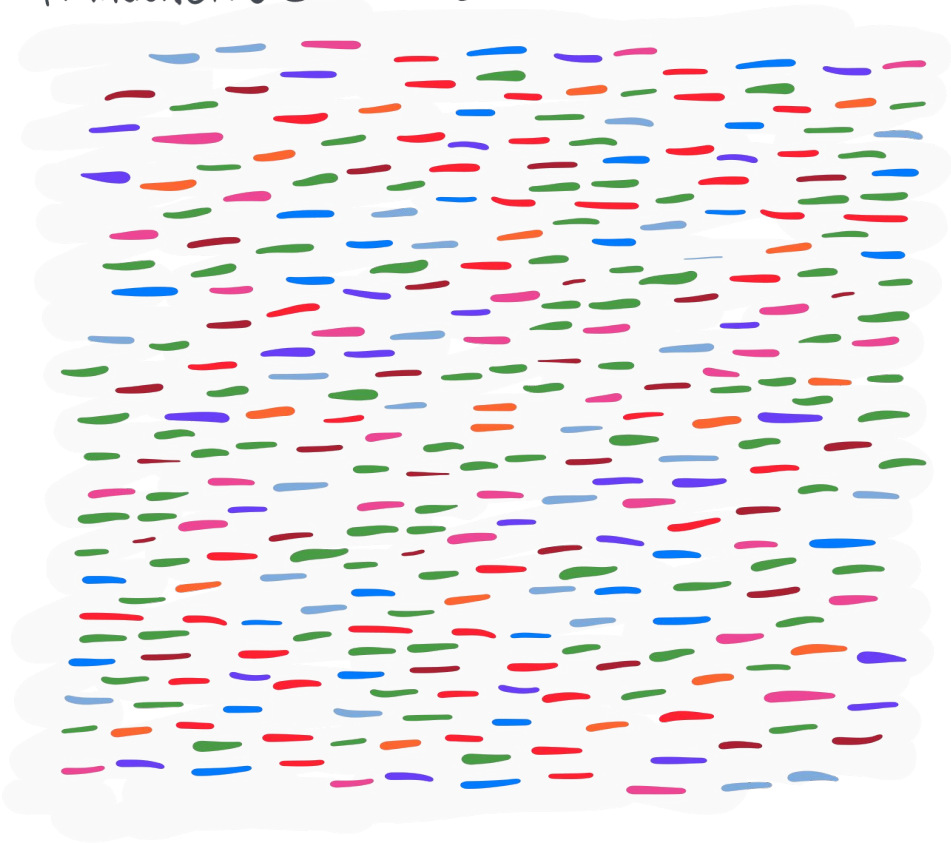
METAGENOMIC SHORT READS



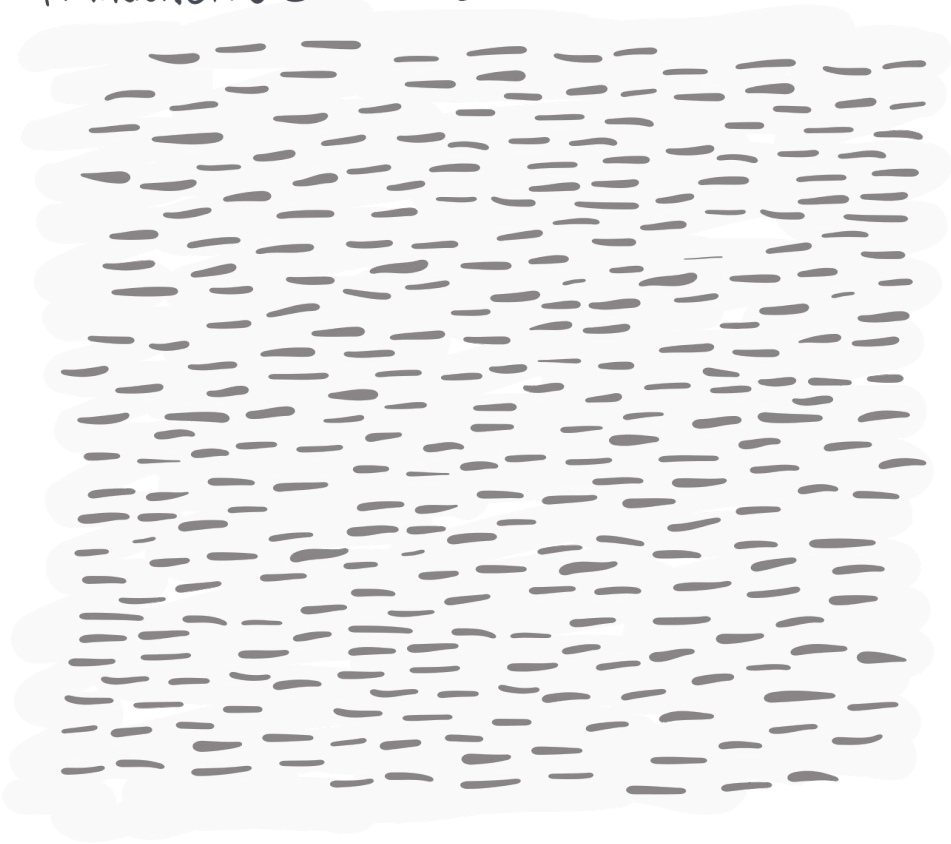
READ RECRUITMENT



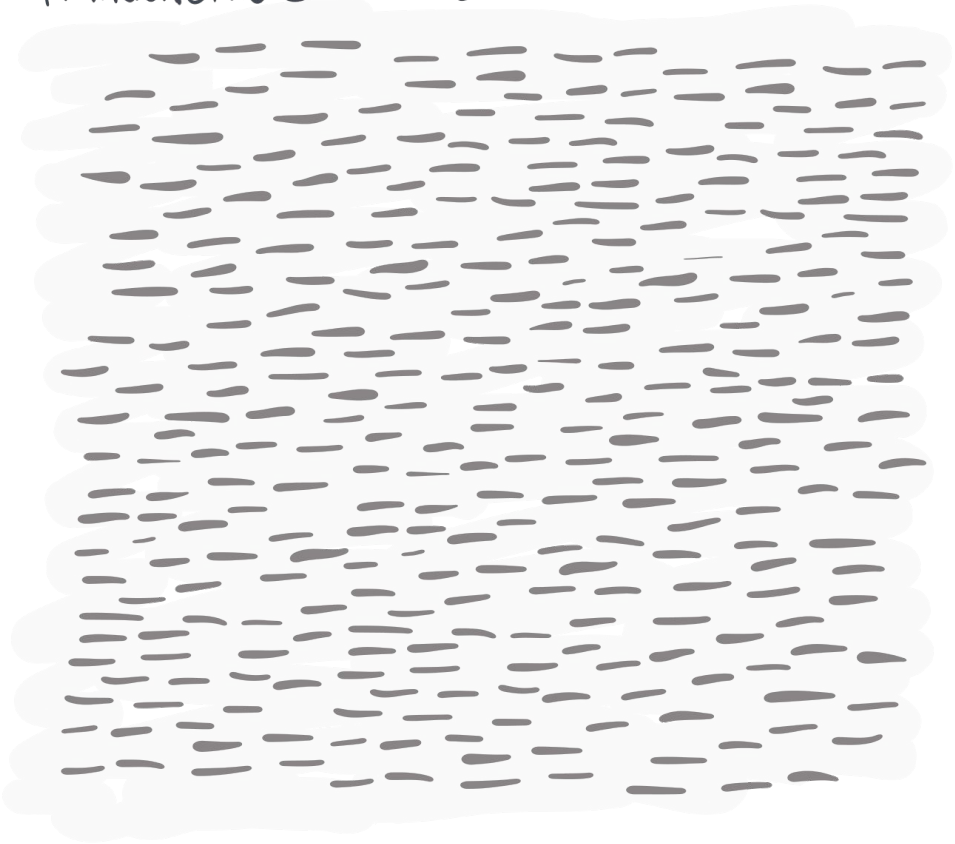
METAGENOMIC SHORT READS

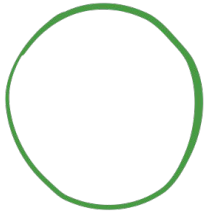


METAGENOMIC SHORT READS

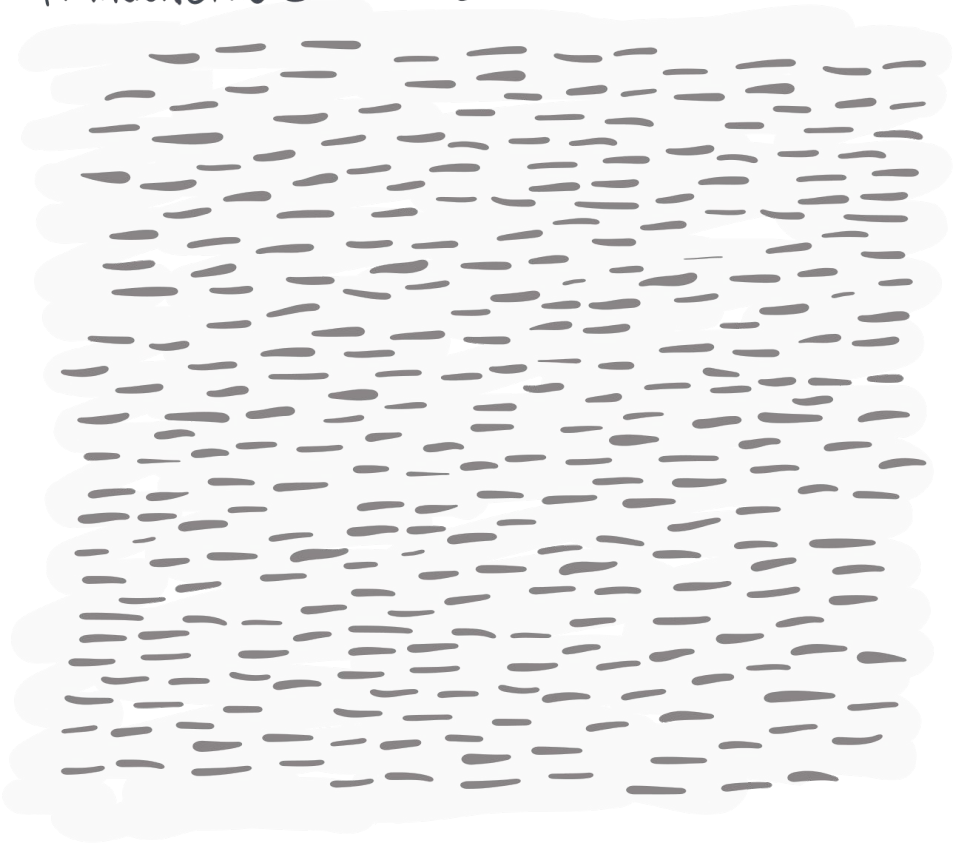


METAGENOMIC SHORT READS

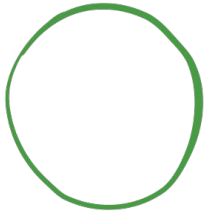




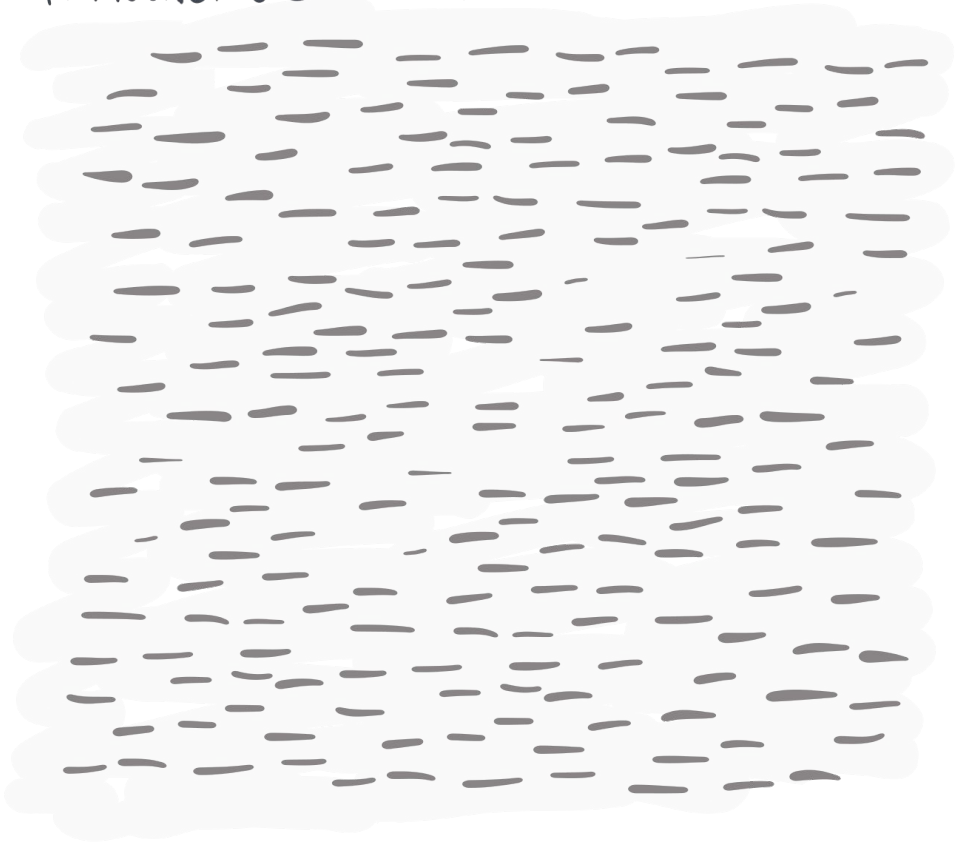
METAGENOMIC SHORT READS



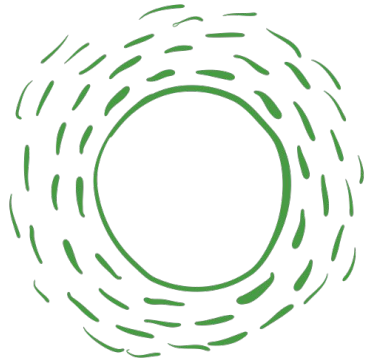
READ RECRUITMENT



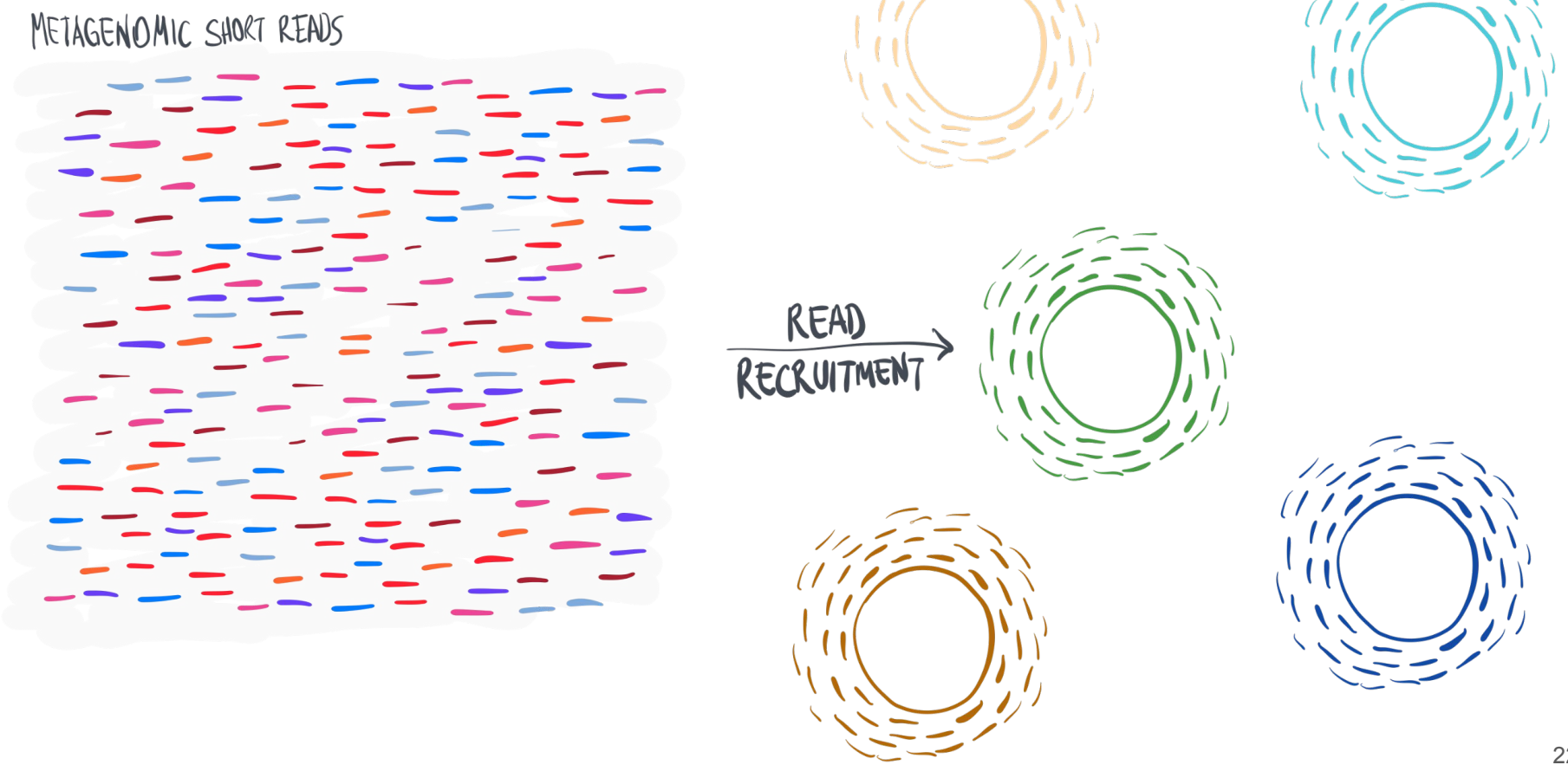
METAGENOMIC SHORT READS



READ RECRUITMENT



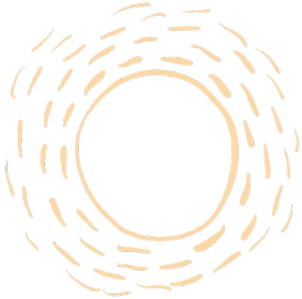
Read recruitment



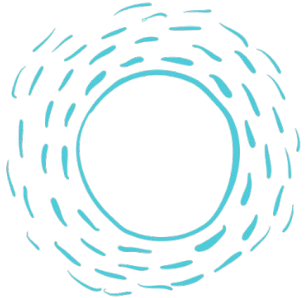
Read recruitment

READ RECRUITMENT

Species 1 ⇒ 10x coverage



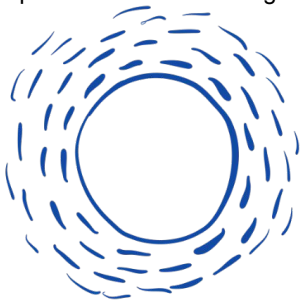
Species 2 ⇒ 15x coverage



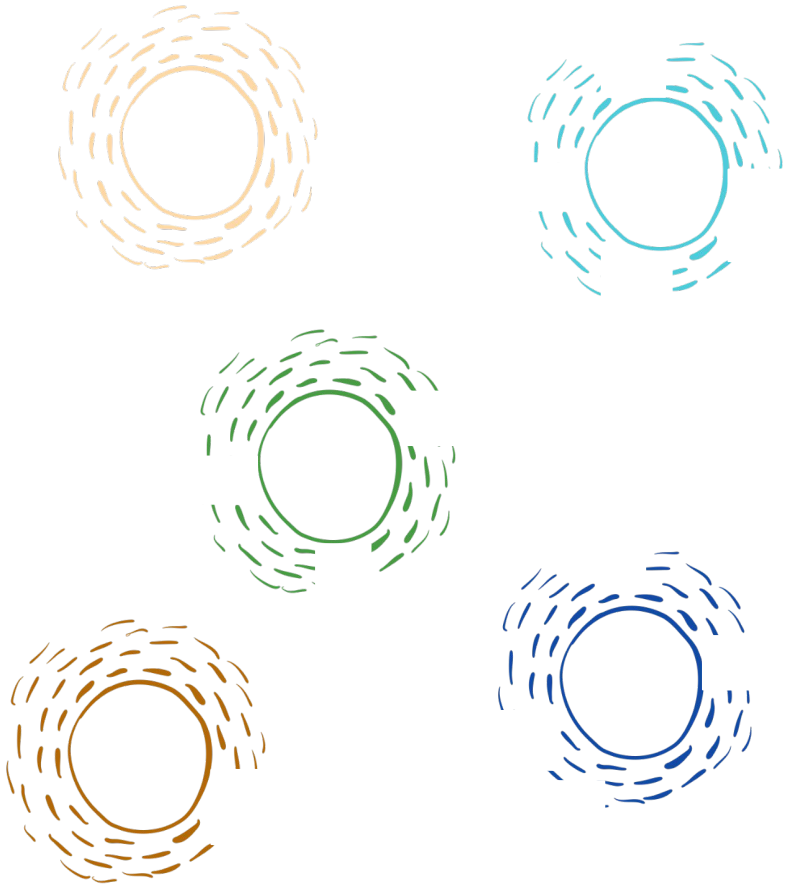
Species 3 ⇒ 5x coverage

Species 4 ⇒ 20x coverage

Species 4 ⇒ 15x coverage



Read recruitment



What is the % of MAG bps covered by reads in one sample or in different samples?

So the fraction of the MAG covered is an important consideration.

Normally cut off 80% covered in reads

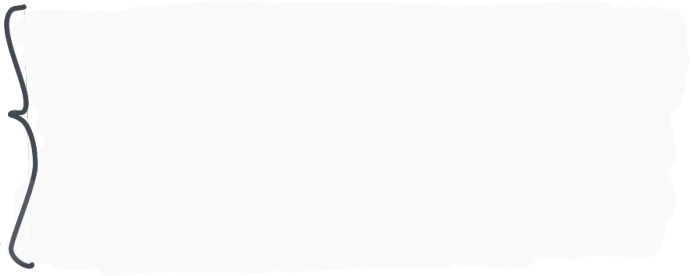

'omics
The power of metagenomic read recruitment
to see the 'obvious'
http://merenlab.org/momics

Metagenomic Read Recruitment

A REFERENCE CONTEXT

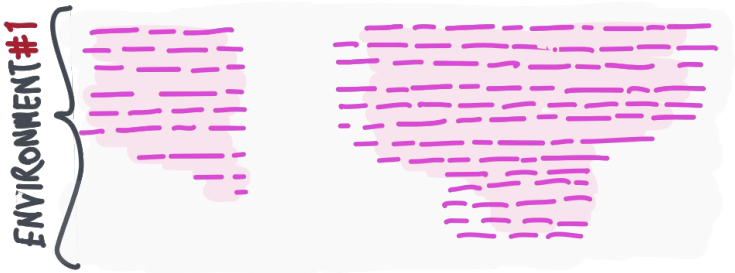
A REFERENCE CONTEXT

ENVIRONMENT #1



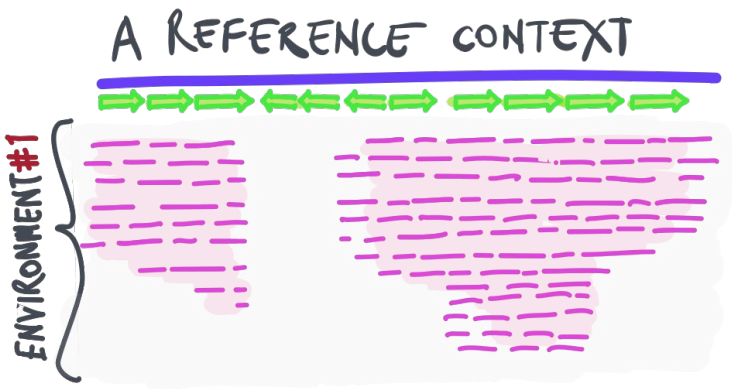
METAGENOMIC READ RECRUITMENT

A Reference context



Metagenomic Read Recruitment

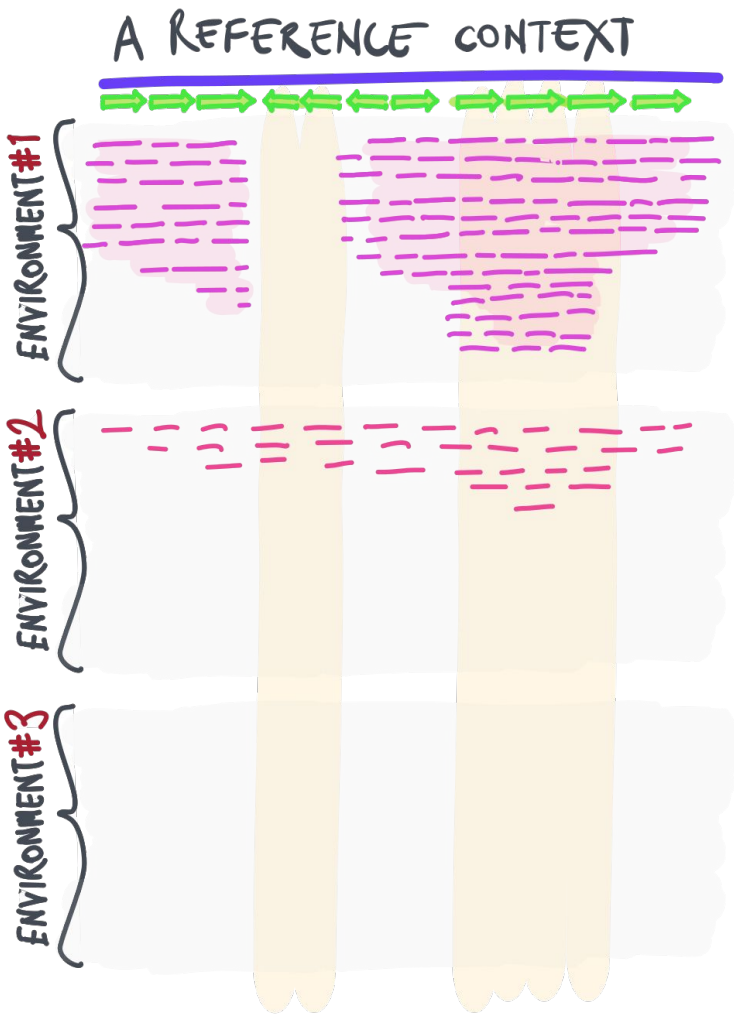
METAGENOMIC READ RECRUITMENT



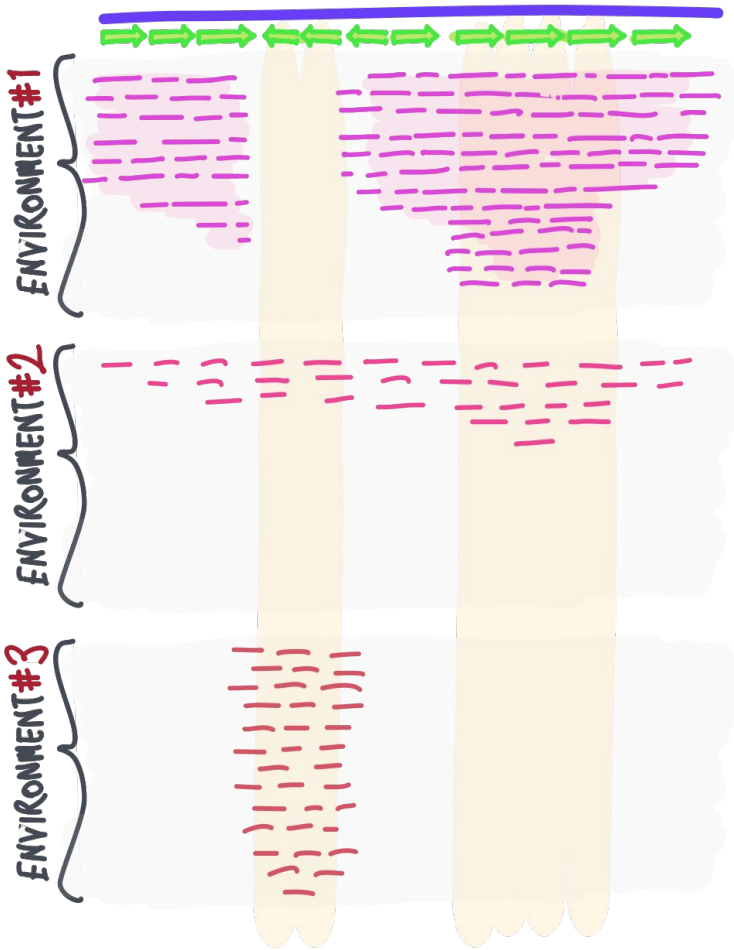
METAGENOMIC READ RECRUITMENT

METAGENOMIC READ RECRUITMENT

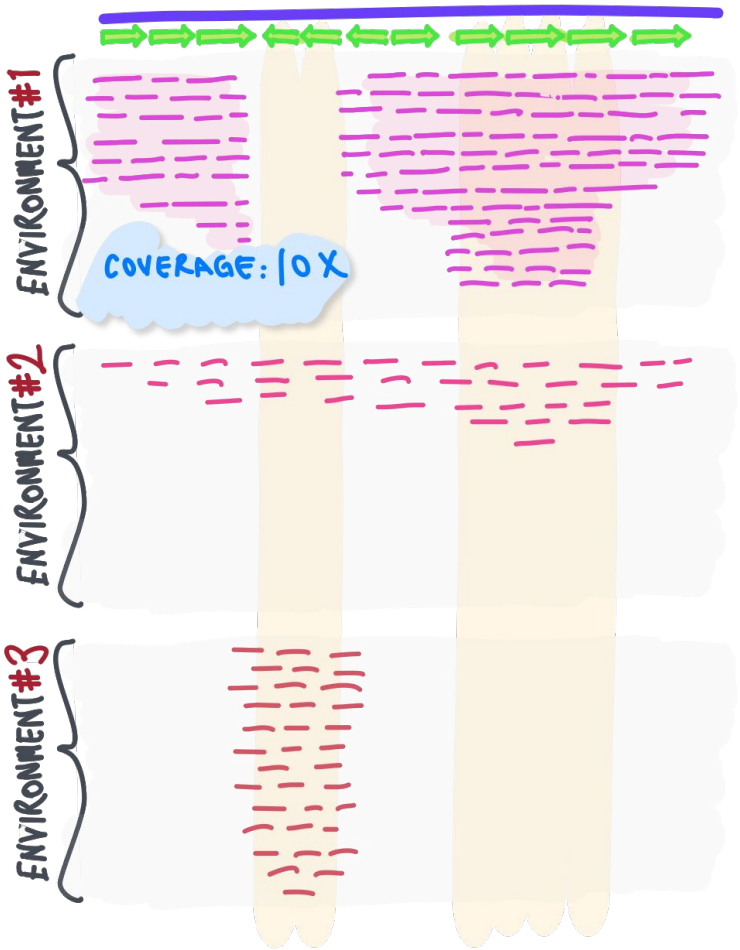
Metagenomic Read Recruitment



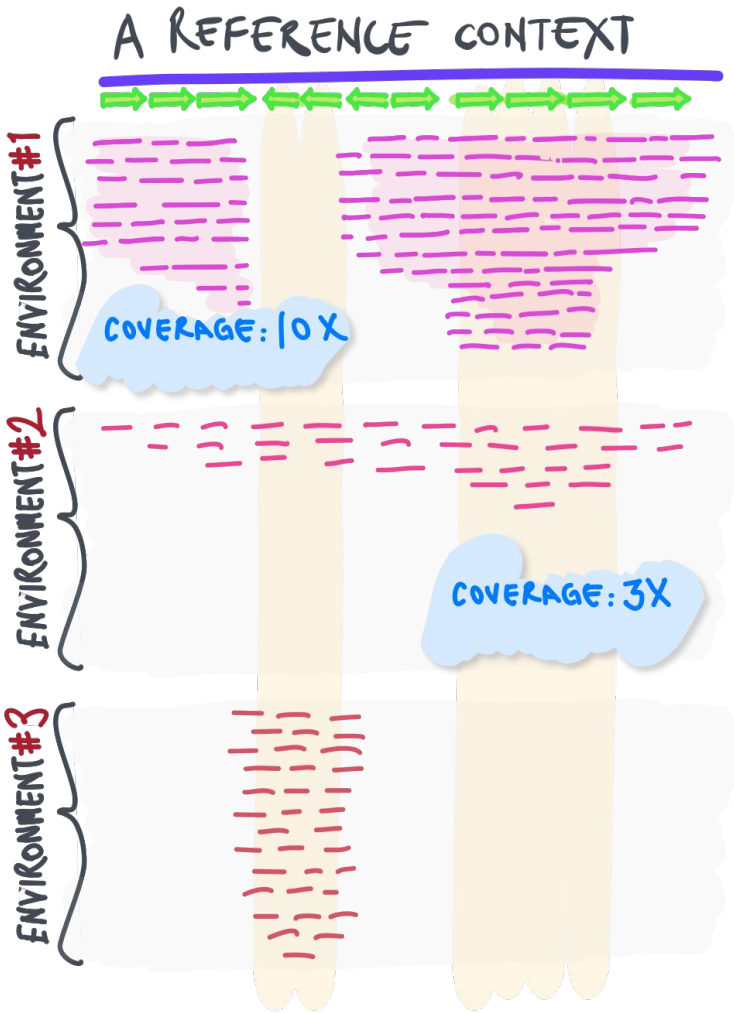
METAGENOMIC READ RECRUITMENT



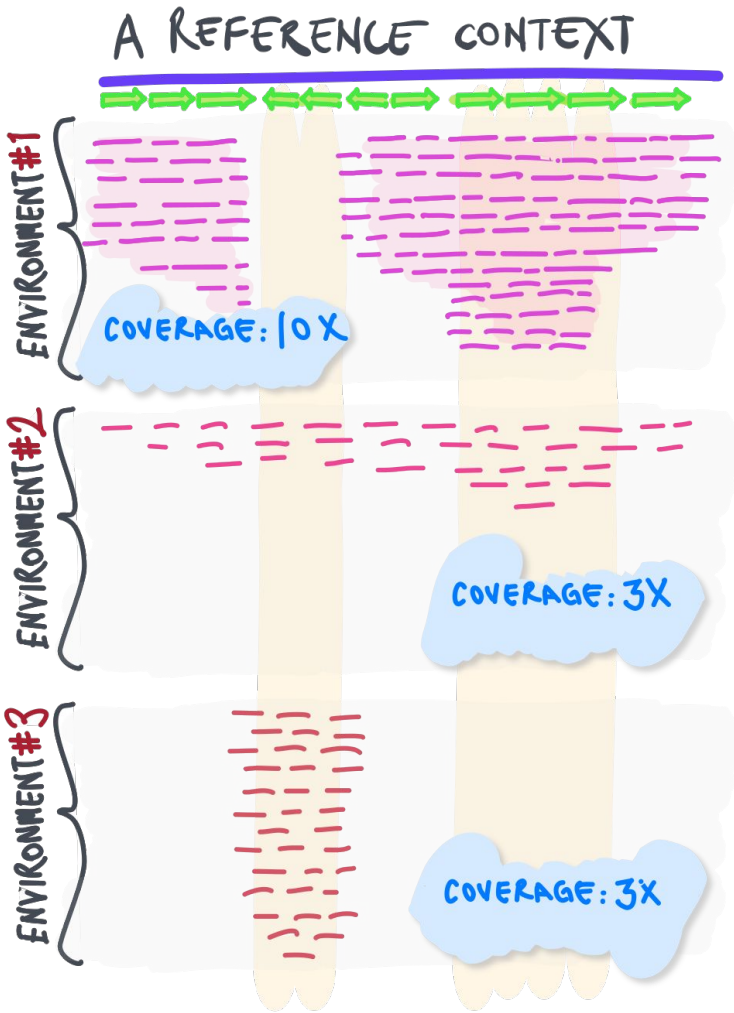
Metagenomic Read Recruitment



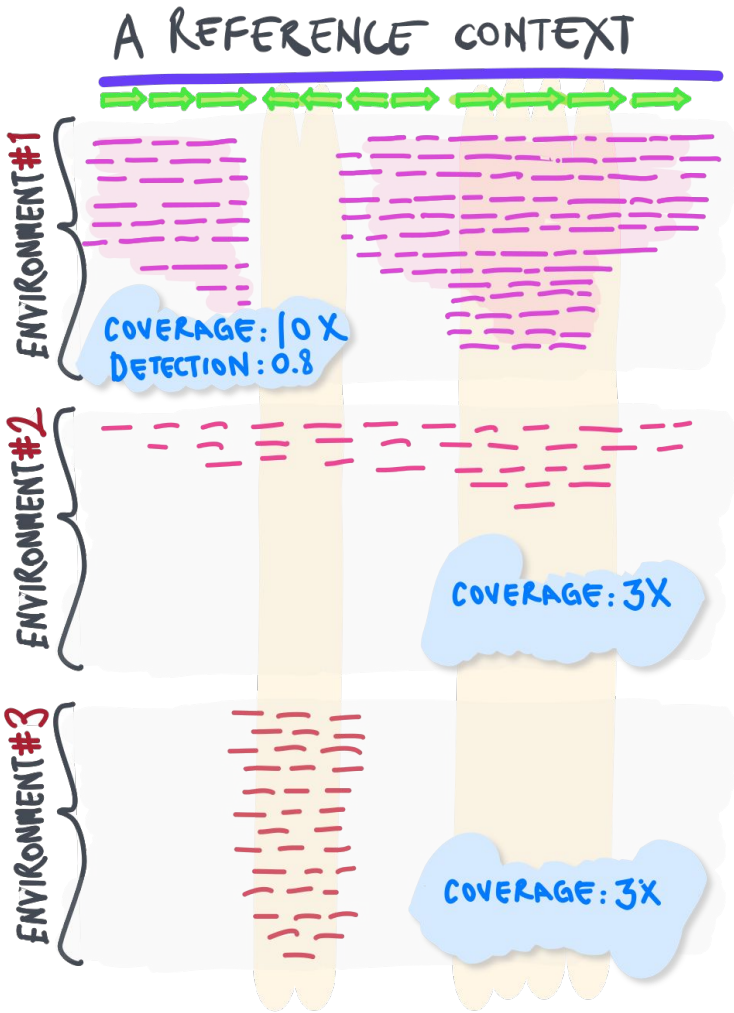
METAGENOMIC READ RECRUITMENT



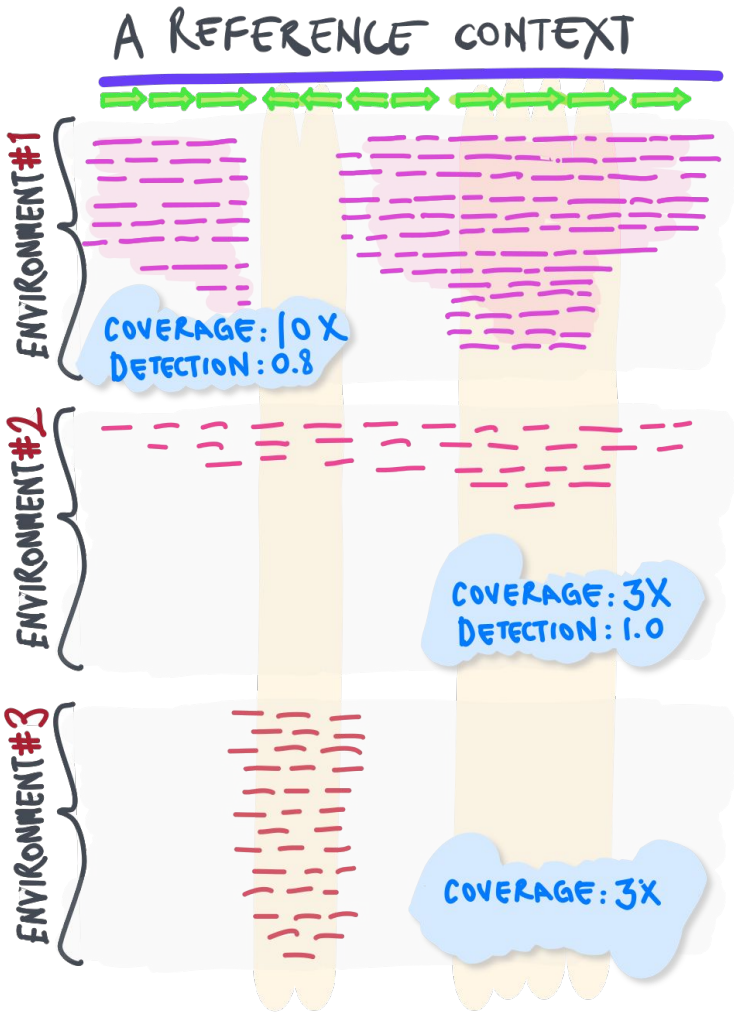
METAGENOMIC READ RECRUITMENT



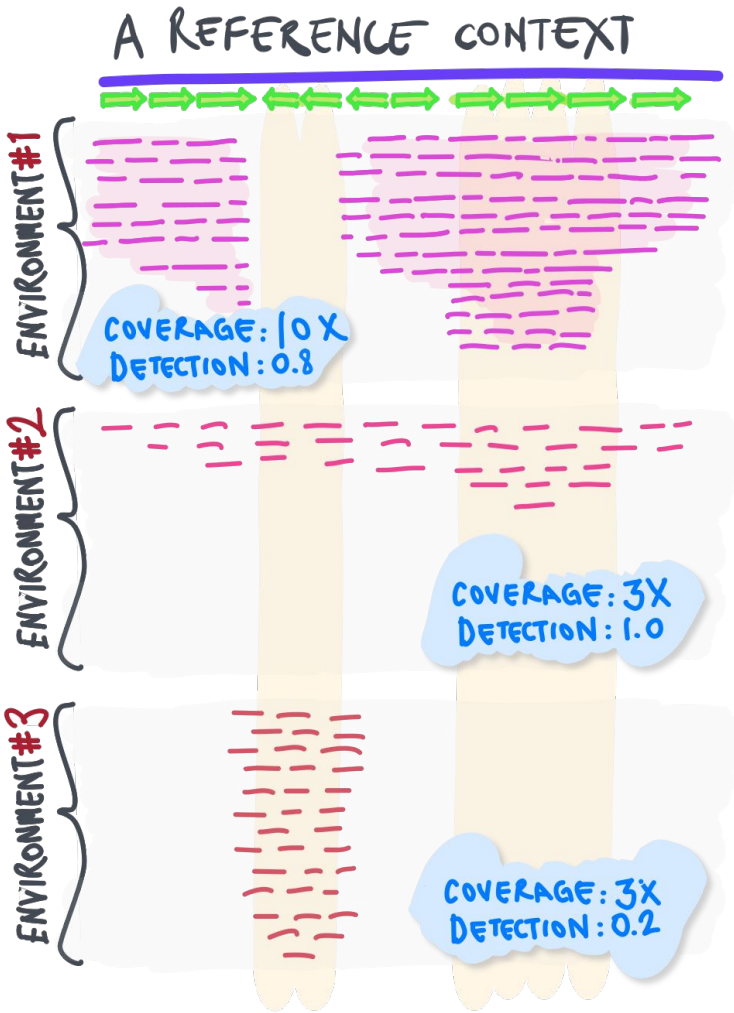
METAGENOMIC READ RECRUITMENT



Metagenomic Read Recruitment

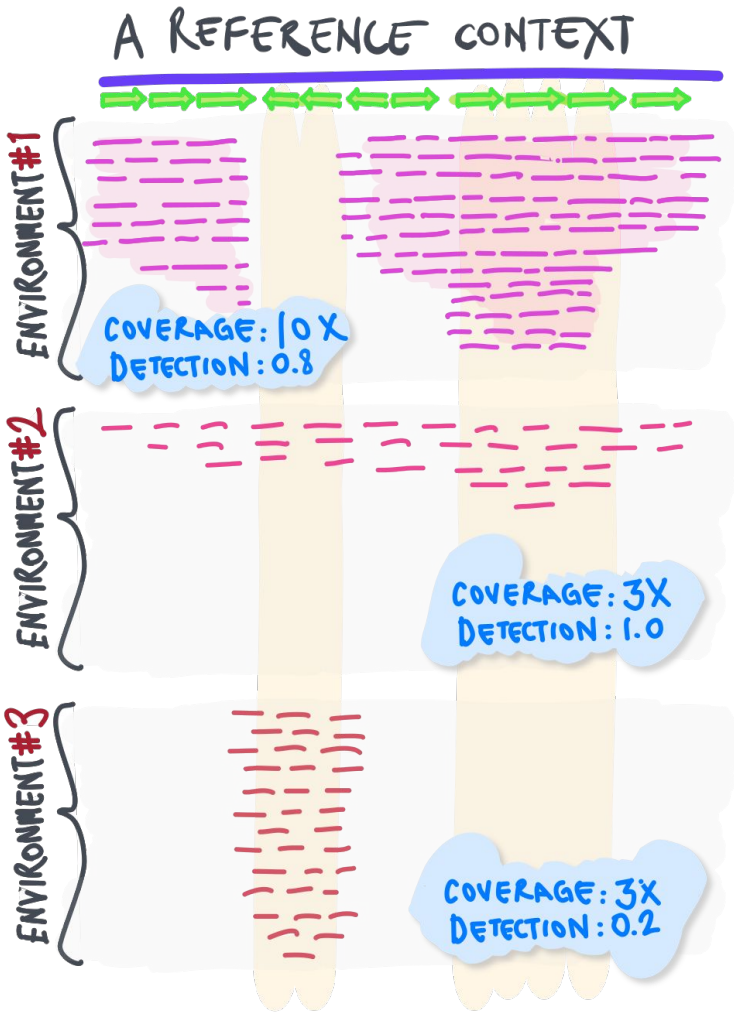


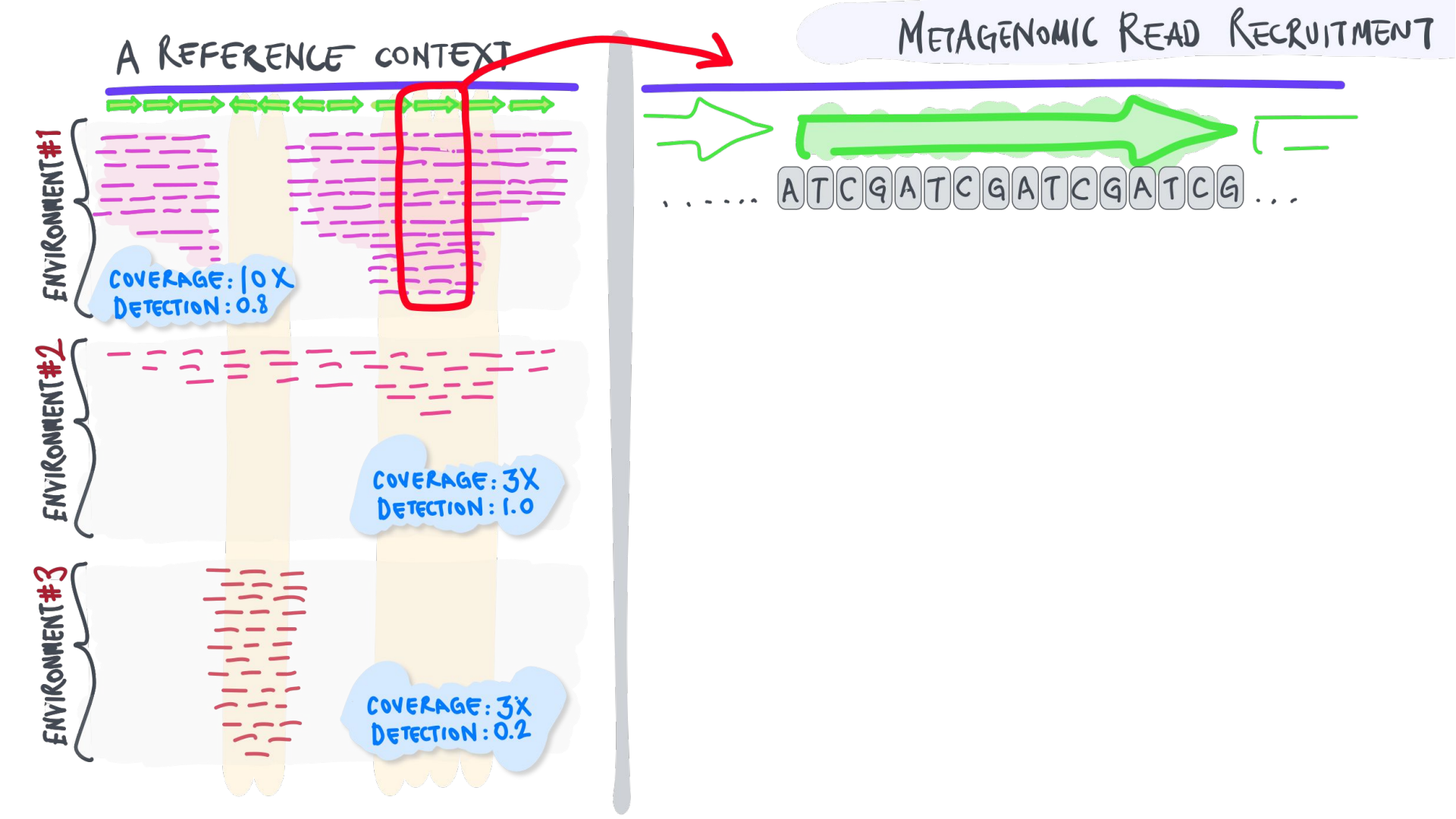
Metagenomic Read Recruitment

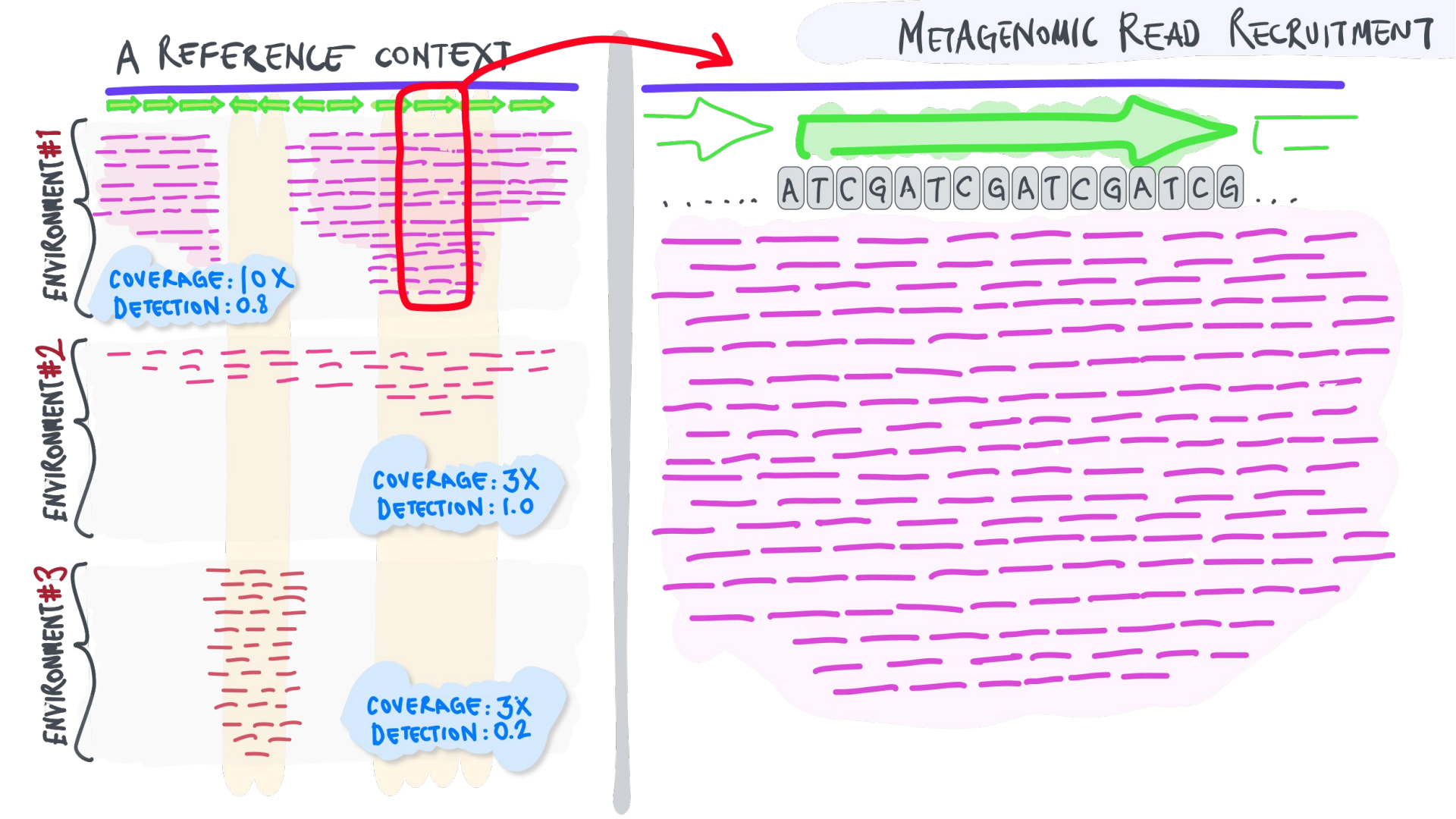


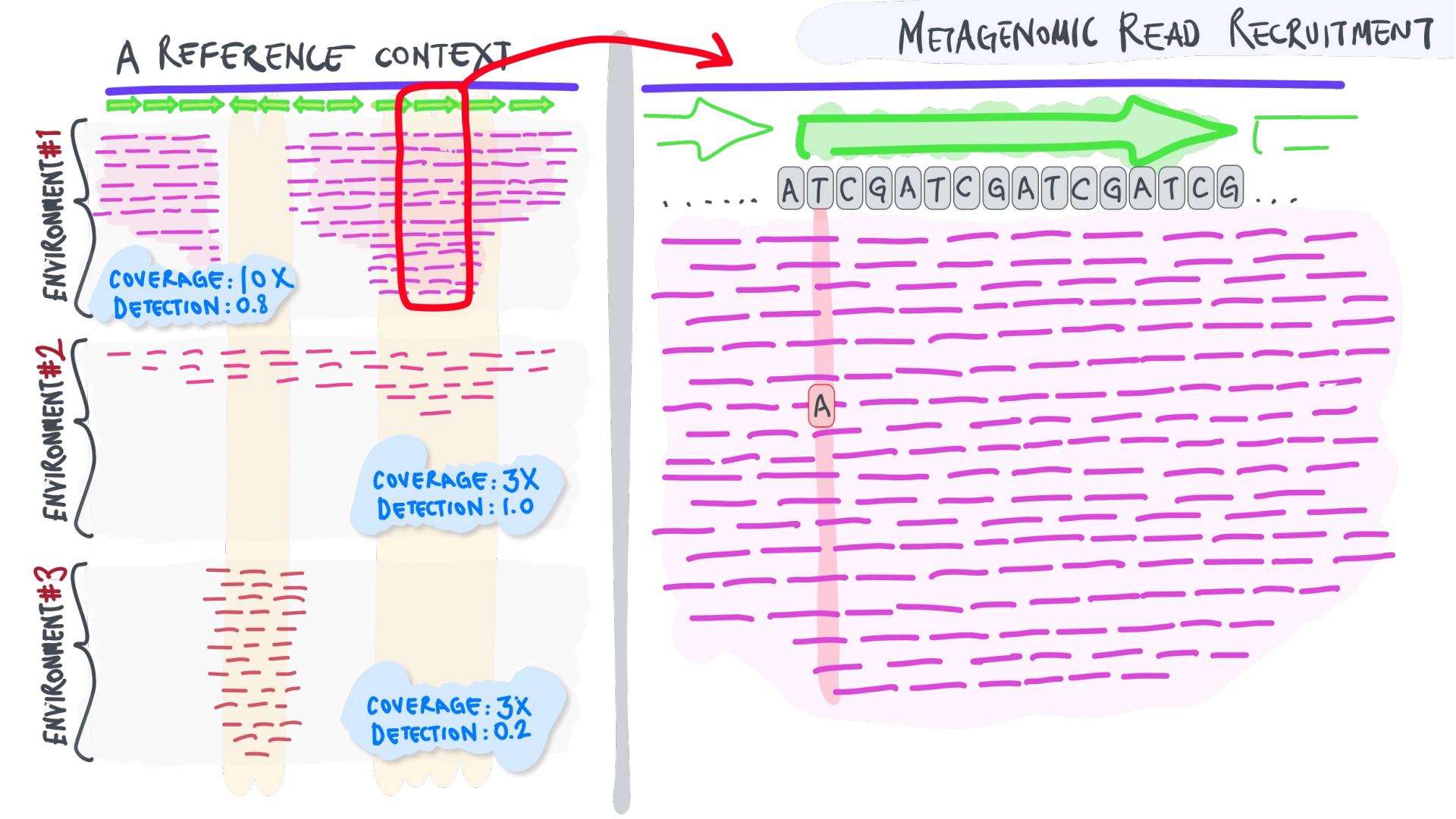

'omics
The power of metagenomic read recruitment
to see the 'subtle'
http://merenlab.org/momics

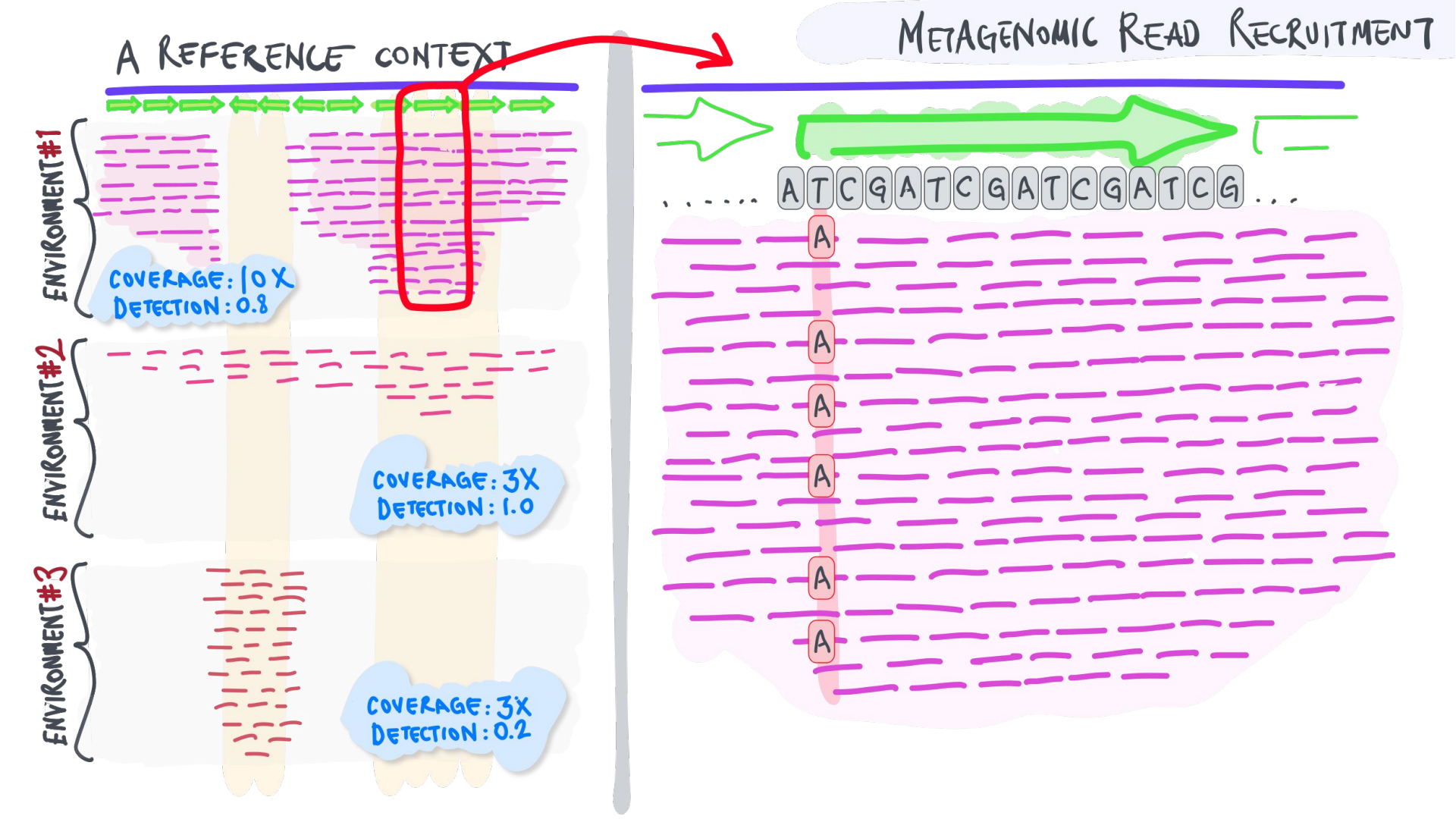
Metagenomic Read Recruitment

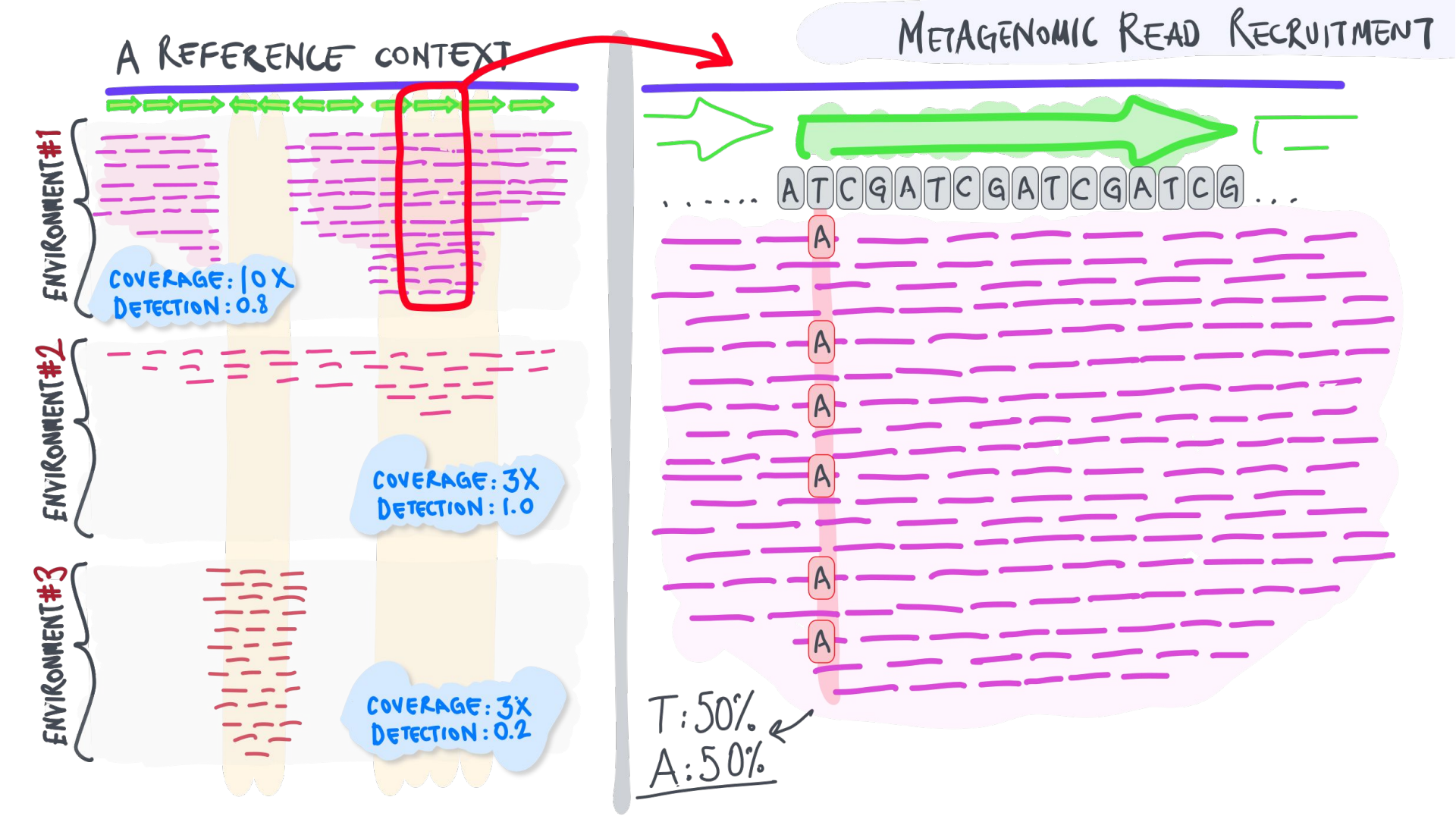


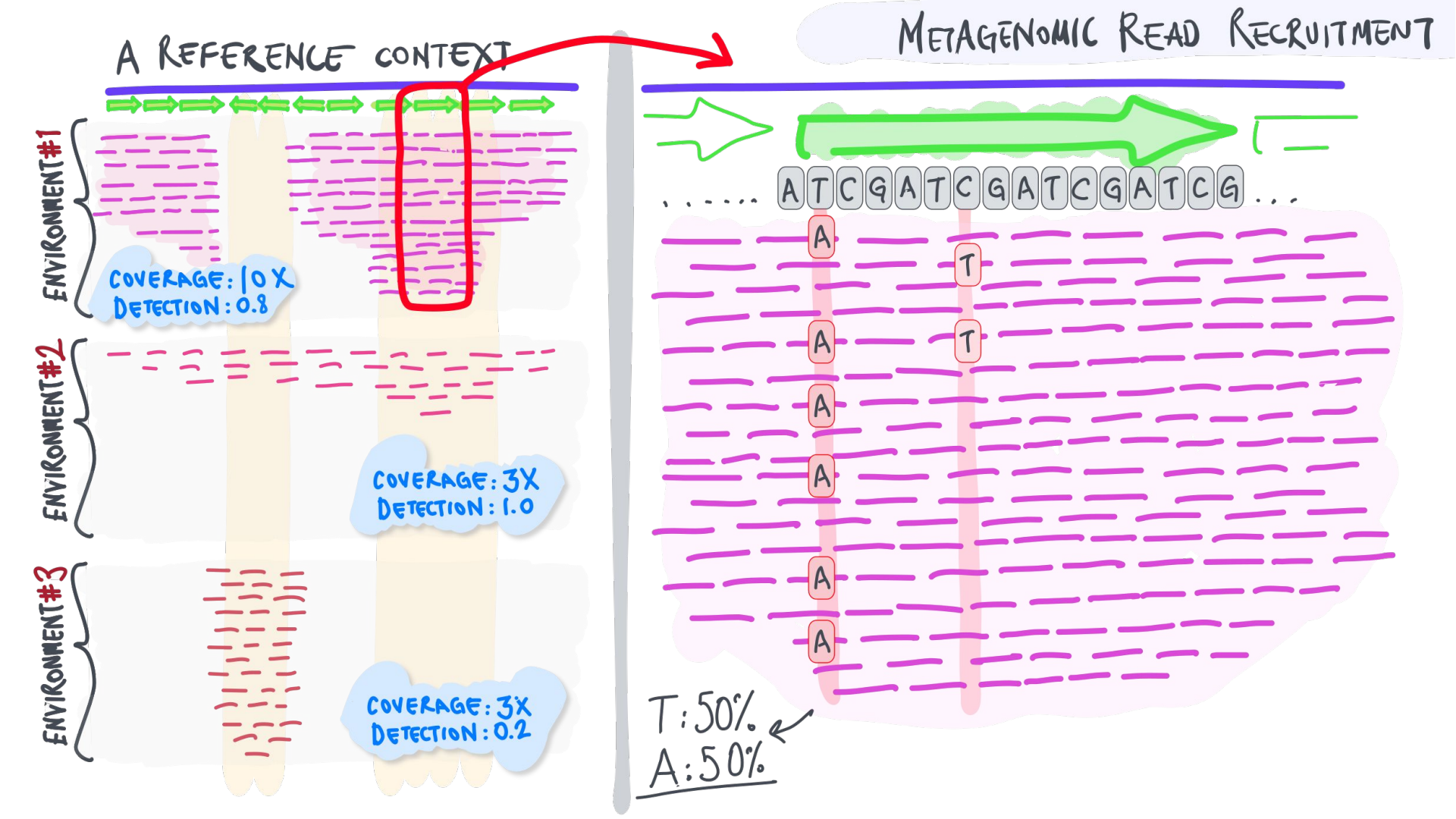

A REFERENCE CONTEXT
METAGENOMIC READ RECRUITMENT
ENVIRONMENT#1
COVERAGE: 10X
DETECTION: 0.8
ENVIRONMENT#2
COVERAGE: 3X
DETECTION: 1.0
ENVIRONMENT#3
COVERAGE: 3X
DETECTION: 0.2
. . . . . A T C G A T C G A T C G A T C G . . .

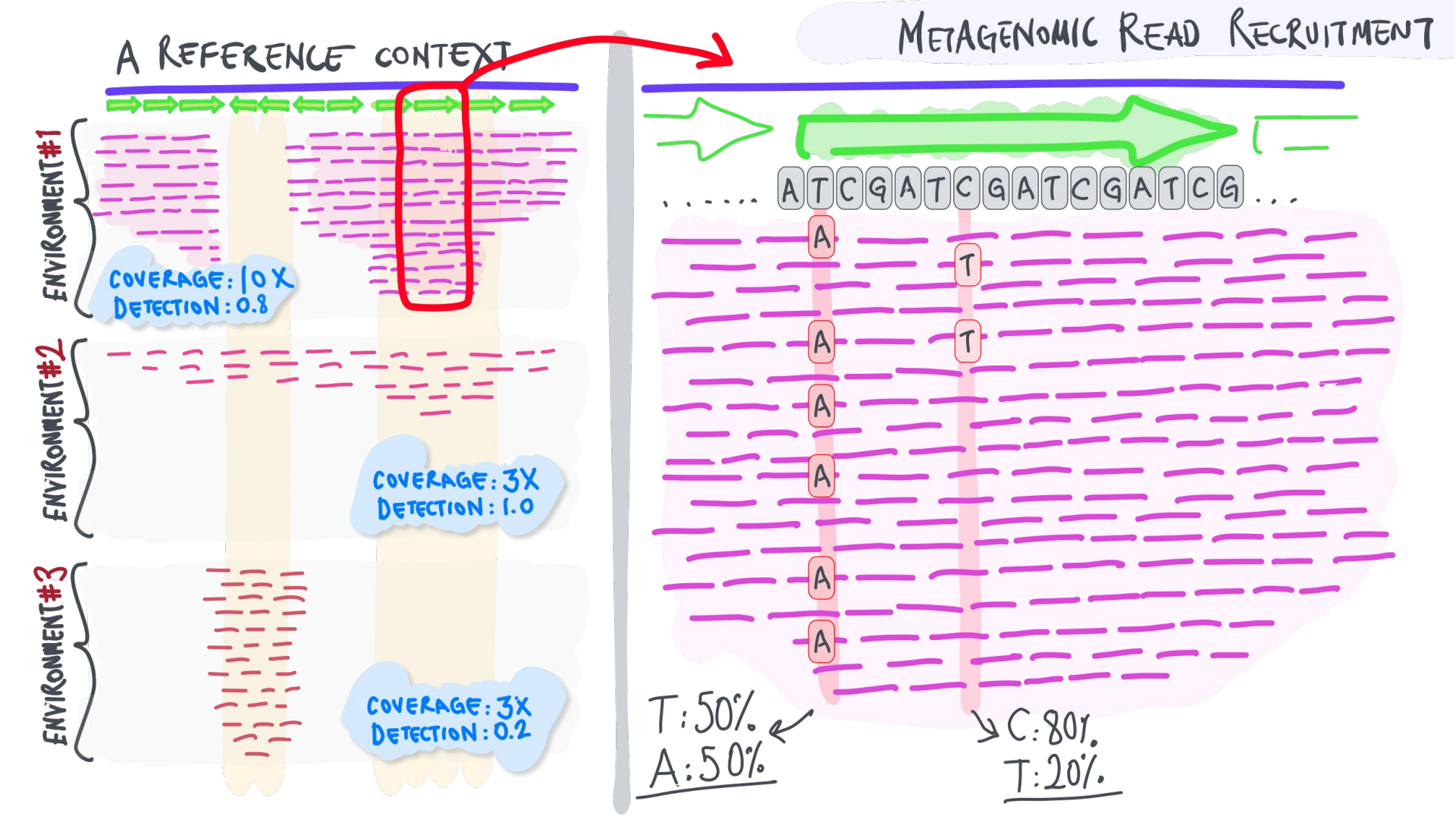

A REFERENCE CONTEXT
METAGENOMIC READ RECRUITMENT
ENVIRONMENT#1
COVERAGE: 10X
DETECTION: 0.8
ENVIRONMENT#2
COVERAGE: 3X
DETECTION: 1.0
ENVIRONMENT#3
COVERAGE: 3X
DETECTION: 0.2
ATCGATCGATCGATCG

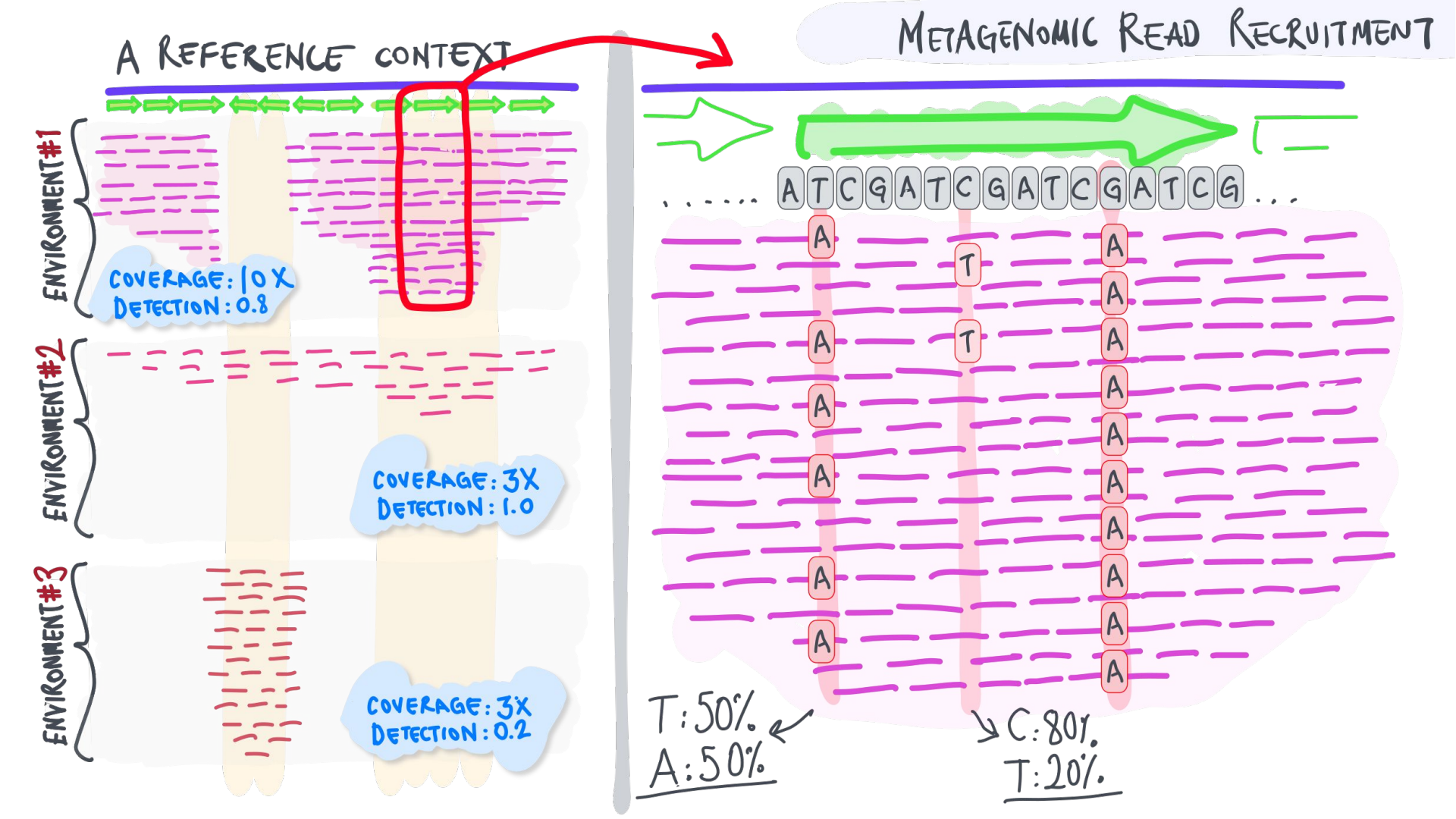

A REFERENCE CONTEXT
METAGENOMIC READ RECRUITMENT
ENVIRONMENT#1
COVERAGE: 10X
DETECTION: 0.8
ENVIRONMENT#2
COVERAGE: 3X
DETECTION: 1.0
ENVIRONMENT#3
COVERAGE: 3X
DETECTION: 0.2
ATCGATCGATCGATCG
A

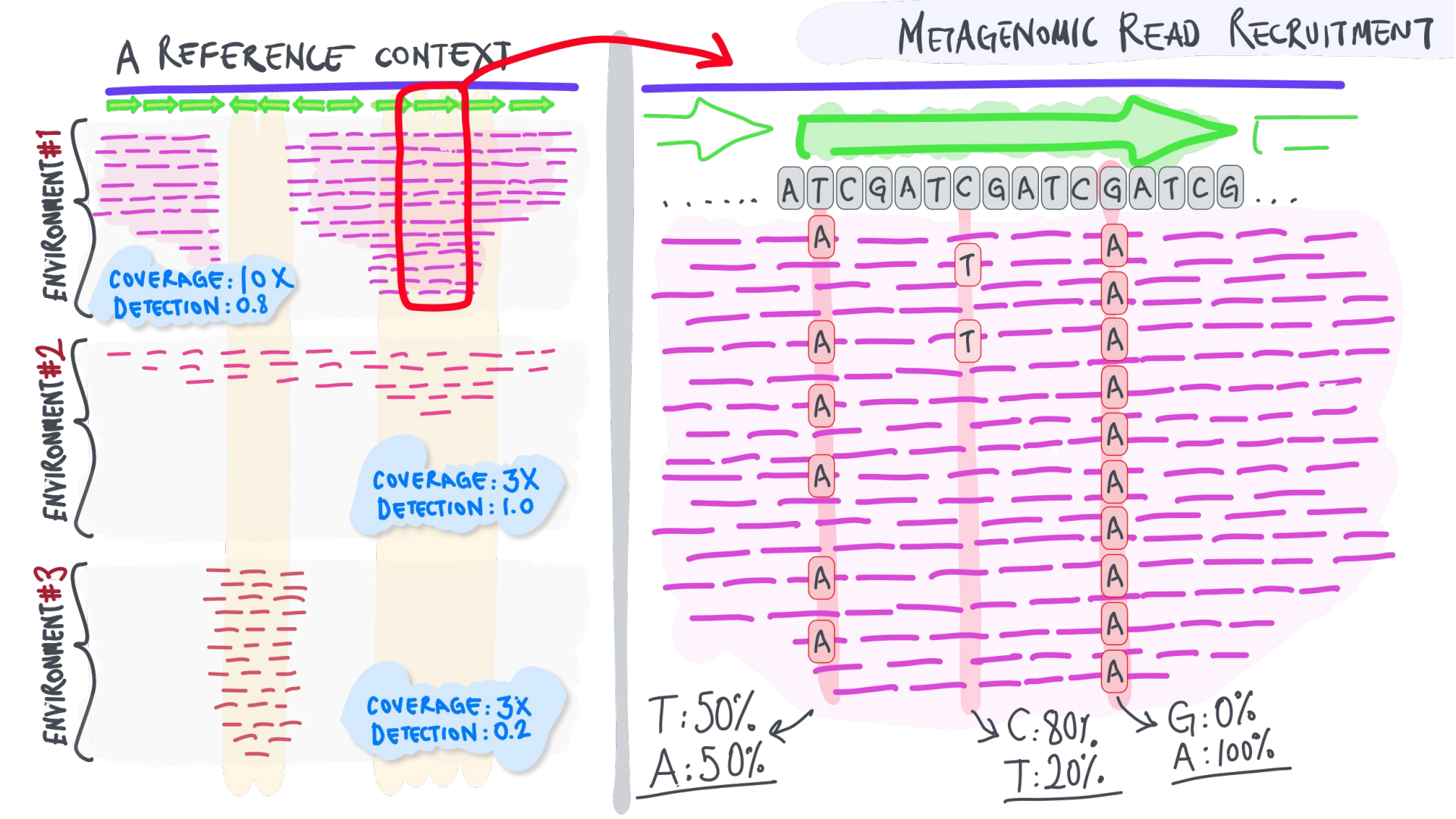

A REFERENCE CONTEXT
METAGENOMIC READ RECRUITMENT
ENVIRONMENT#1
COVERAGE: 10X
DETECTION: 0.8
ENVIRONMENT#2
COVERAGE: 3X
DETECTION: 1.0
ENVIRONMENT#3
COVERAGE: 3X
DETECTION: 0.2
ATCGATCGATCGATCG
A
A
A
A
A
A


A REFERENCE CONTEXT
METAGENOMIC READ RECRUITMENT
ENVIRONMENT#1
COVERAGE: 10X
DETECTION: 0.8
ENVIRONMENT#2
COVERAGE: 3X
DETECTION: 1.0
ENVIRONMENT#3
COVERAGE: 3X
DETECTION: 0.2
ATCGATCGATCGATCG
T: 50%
A: 50%


A REFERENCE CONTEXT
METAGENOMIC READ RECRUITMENT
ENVIRONMENT#1
COVERAGE: 10X
DETECTION: 0.8
ENVIRONMENT#2
COVERAGE: 3X
DETECTION: 1.0
ENVIRONMENT#3
COVERAGE: 3X
DETECTION: 0.2
. . . . . A T C G A T C G A T C G A T C G . . .
A
T
A
T
A
A
A
A
T: 50%
A: 50%


A REFERENCE CONTEXT
METAGENOMIC READ RECRUITMENT
ENVIRONMENT#1
COVERAGE: 10X
DETECTION: 0.8
ENVIRONMENT#2
COVERAGE: 3X
DETECTION: 1.0
ENVIRONMENT#3
COVERAGE: 3X
DETECTION: 0.2
. . . . . A T C G A T C G A T C G A T C G . . .
A
T
A
T
A
A
A
A
A
T: 50%
A: 50%
C: 80%
T: 20%


A REFERENCE CONTEXT
METAGENOMIC READ RECRUITMENT
ENVIRONMENT#1
COVERAGE: 10X
DETECTION: 0.8
ENVIRONMENT#2
COVERAGE: 3X
DETECTION: 1.0
ENVIRONMENT#3
COVERAGE: 3X
DETECTION: 0.2
. . . . . A T C G A T C G A T C G A T C G . . .
T: 50%
A: 50%
C: 80%
T: 20%


A REFERENCE CONTEXT
METAGENOMIC READ RECRUITMENT
ENVIRONMENT#1
COVERAGE: 10X
DETECTION: 0.8
ENVIRONMENT#2
COVERAGE: 3X
DETECTION: 1.0
ENVIRONMENT#3
COVERAGE: 3X
DETECTION: 0.2
. . . . . A T C G A T C G A T C G A T C G . . .
T: 50%
A: 50%
C: 80%
T: 20%
G: 0%
A: 100%

Strain heterogeneity for HQ MAGs

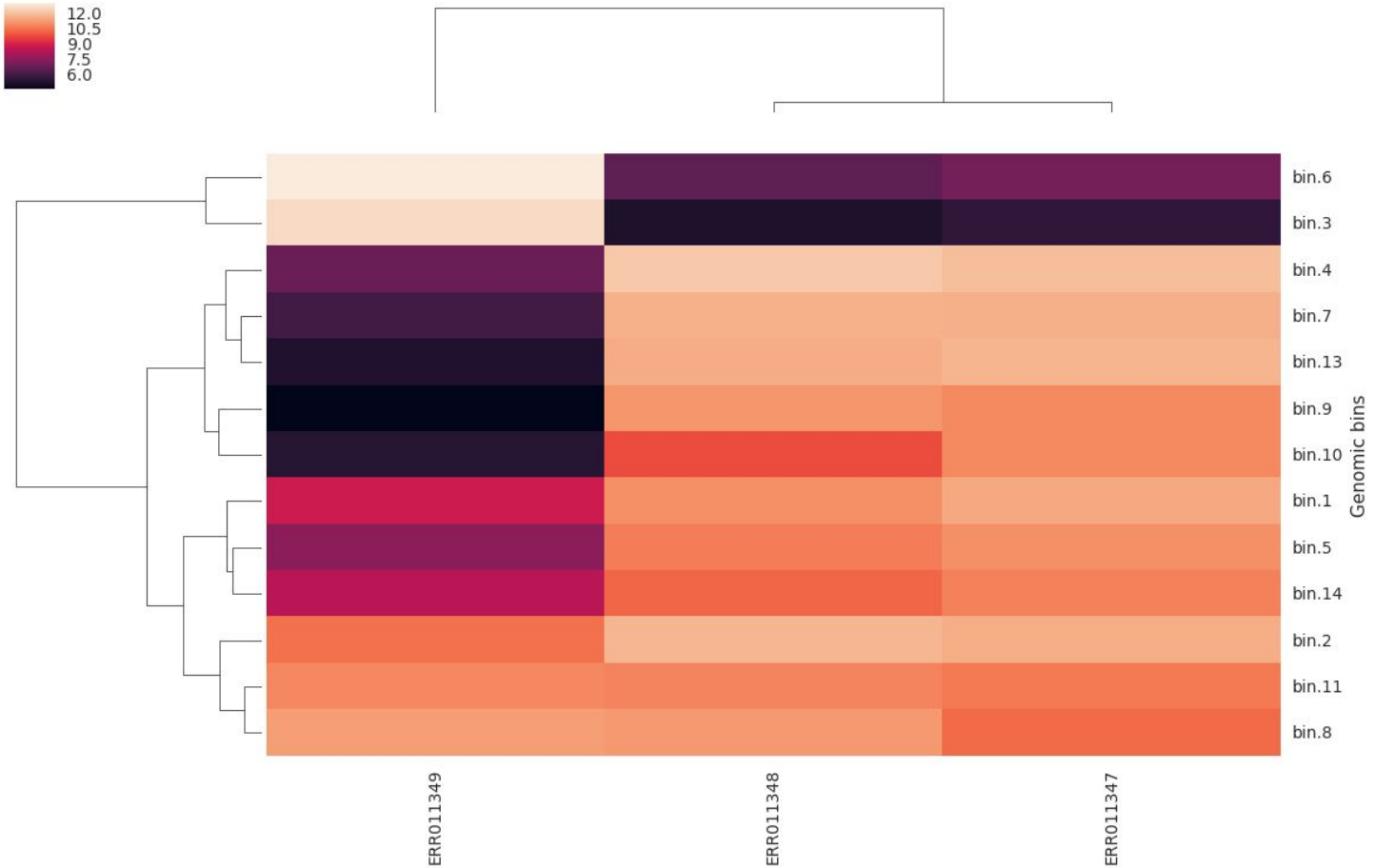
Identify assemblies resulting from strain mixtures even when the strains were very closely related

- Map reads against MAGs from the same sample
- Determine dominant and non-dominant alleles over all protein coding nucleotides
- a position is considered as non-polymorphic if the dominant allele frequency was >80%

<0.5% polymorphic positions add on as a QC for HQ MAGs

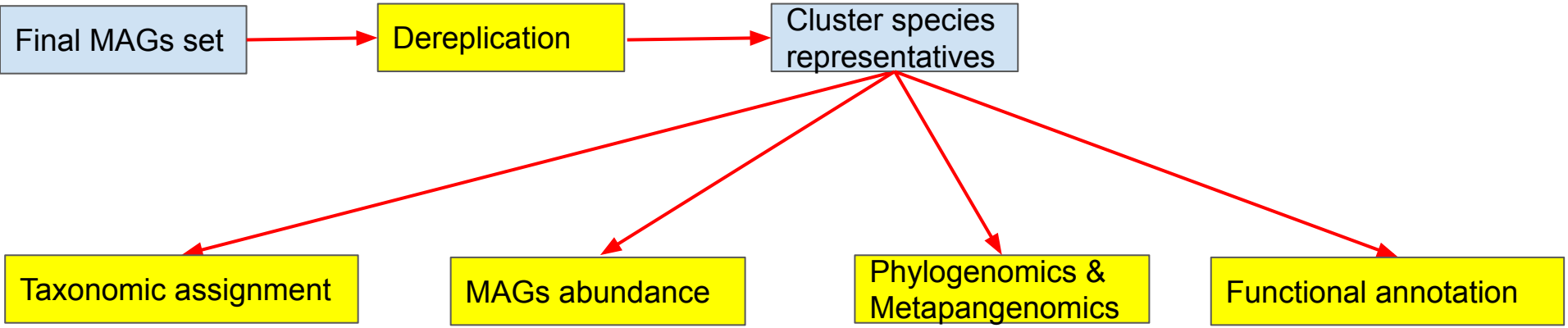
Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P. and Collado, M.C., 2019. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. Cell, 176(3), pp.649-662.

MAGs abundance



Uritskiy, G.V., DiRuggiero, J. and Taylor, J., 2018. MetaWRAP—a flexible pipeline for genome-resolved metagenomic data analysis. Microbiome, 6(1), pp.1-13.

Short Summary

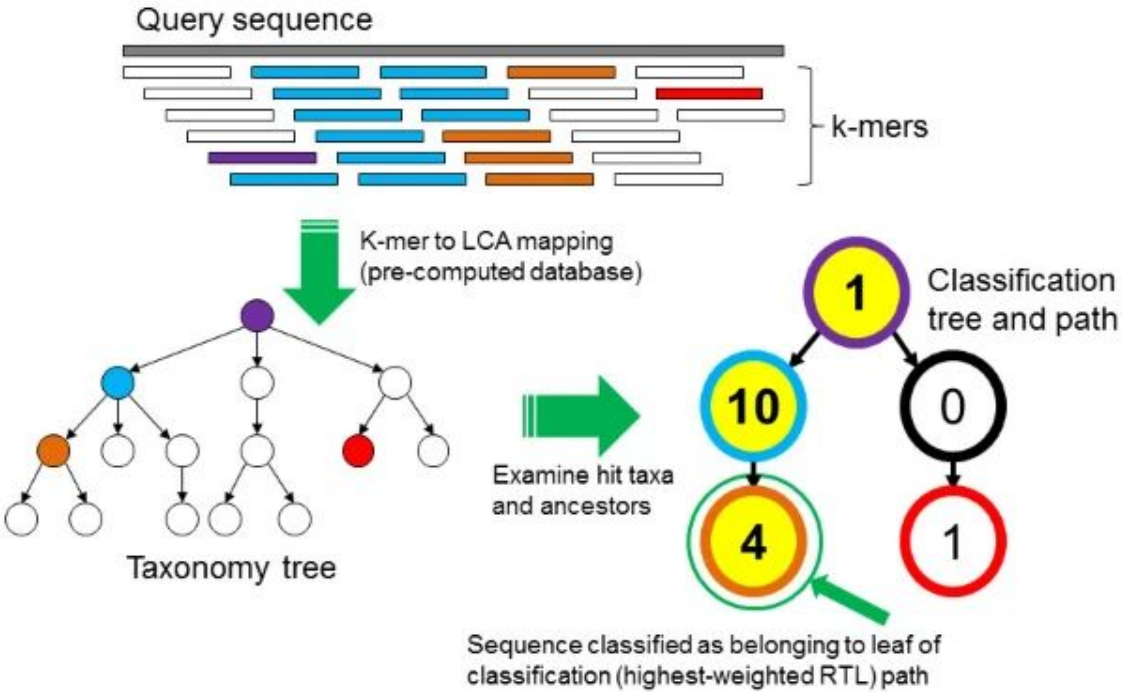


NEXT DAYS!!!

Read profiling (Community taxonomic composition)

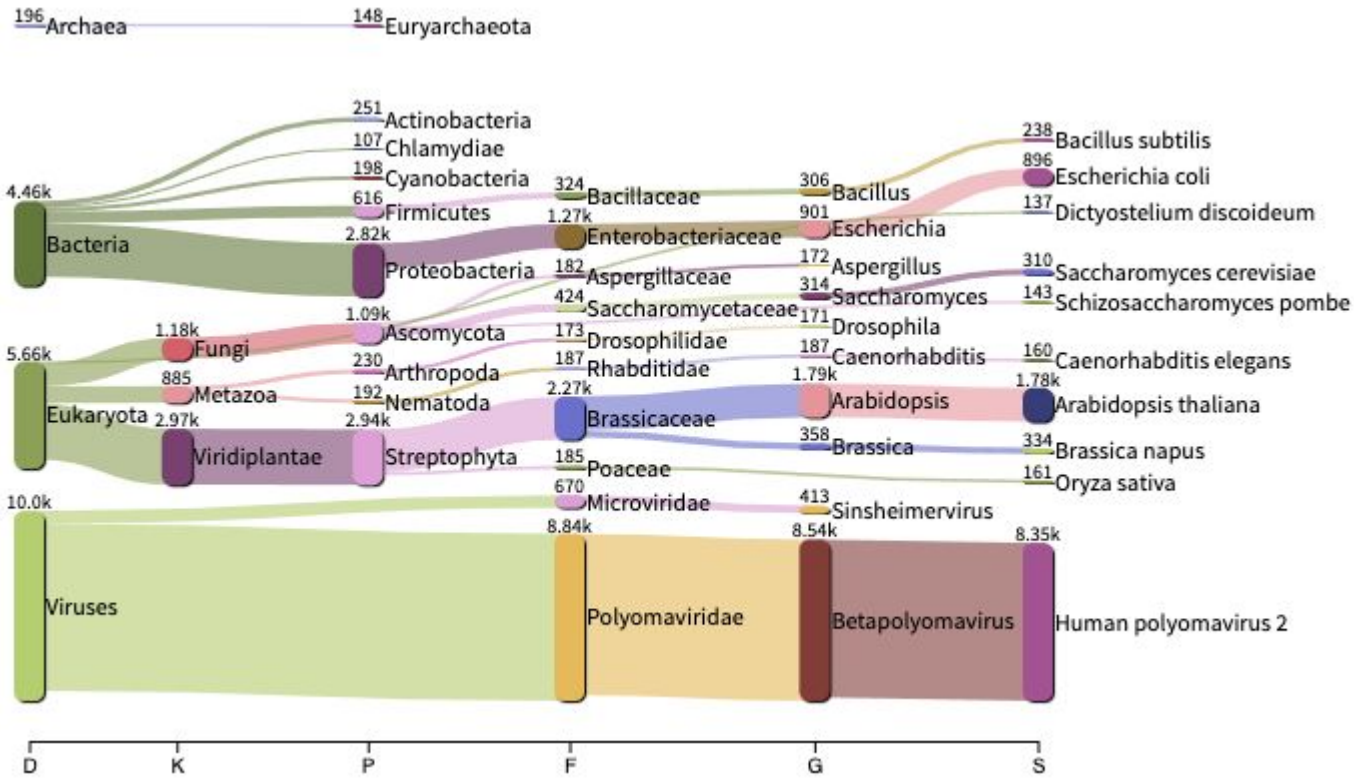
- Profiling tools classify the metagenome as a whole community which reports how much of each taxon is present in the metagenome
  - ⇒ Classify each read in your sample against common databases
- LCA: lowest common ancestor is a robust taxonomic label for unknown sequences
- Sequences are classified by
  - querying the database for each k-mer in a sequence,
  - then using the resulting set of LCA taxa to determine an appropriate label for the sequence

The Kraken sequence classification algorithm



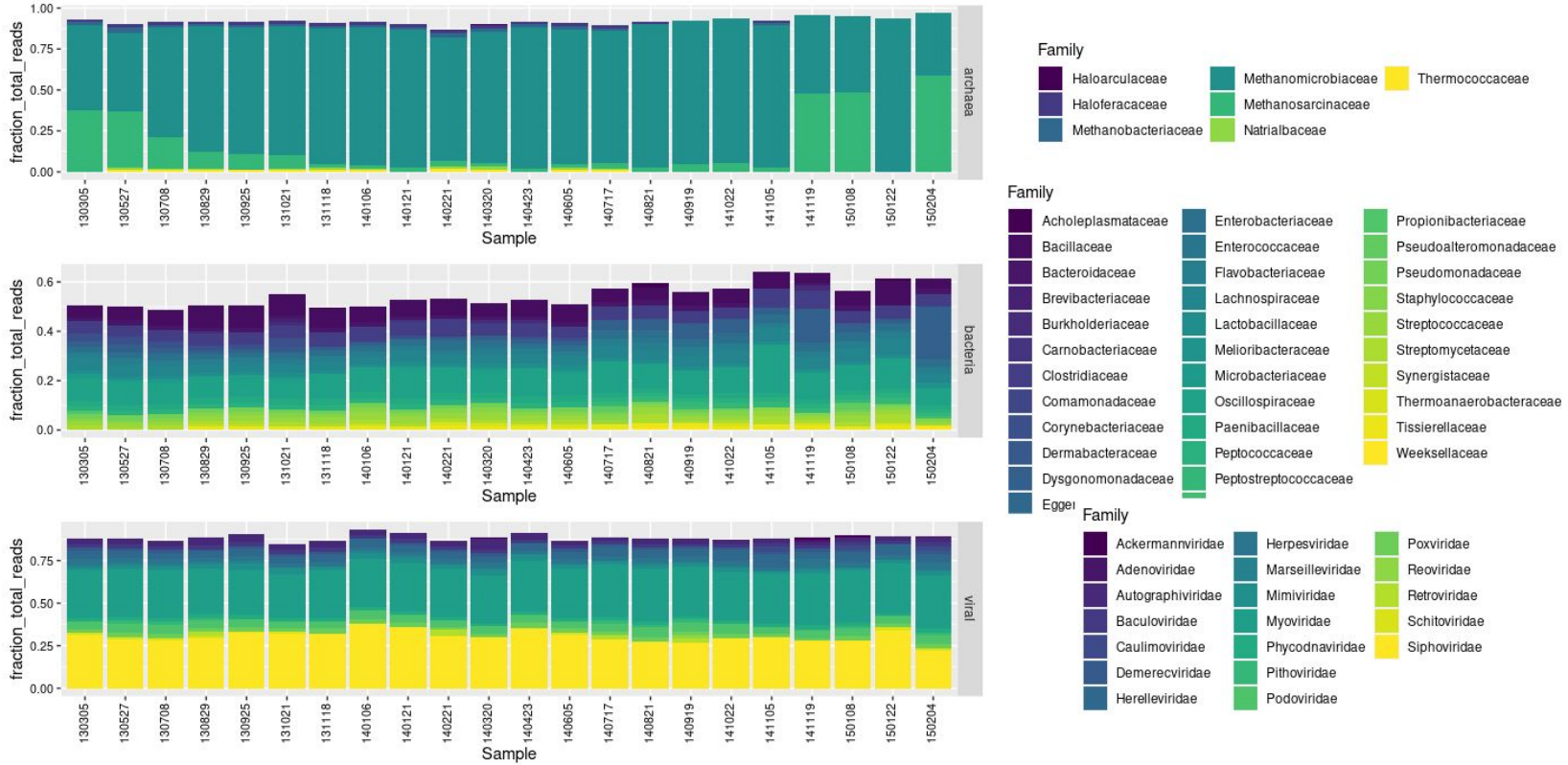
Wood, D.E. and Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, *15*(3), pp.1-12.

Read Profiling Report visualization



https://github.com/soedinglab/MMseqs2/wiki#taxonomy-assignment-using-mmseqs-taxonomy

Read profiling from 22 biogas reactor samples



Important Omics Vocabulary

Bin refinement

Bin reassembly

Chimeras in MAGs

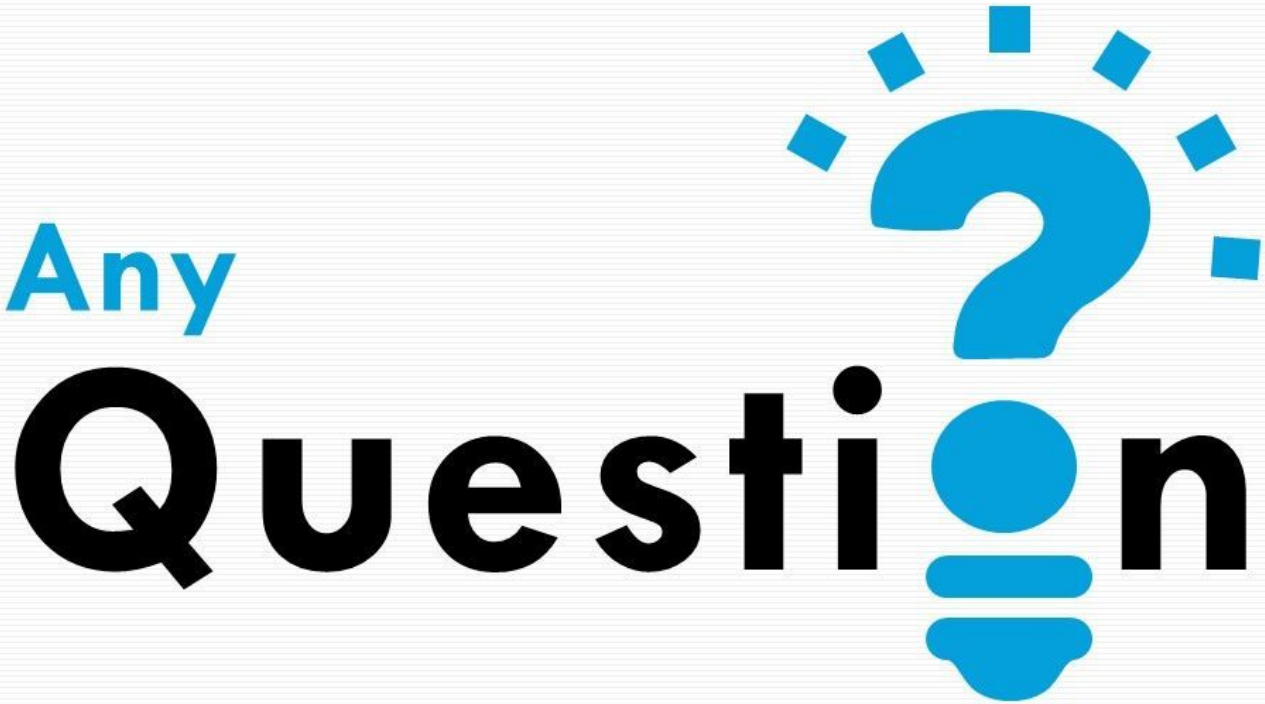
MAGs dereplication

MAGs Taxonomic assignment (GTDB)

Species abundance

LCA and community profiling

https://anvio.org/vocabulary/


Any
Question