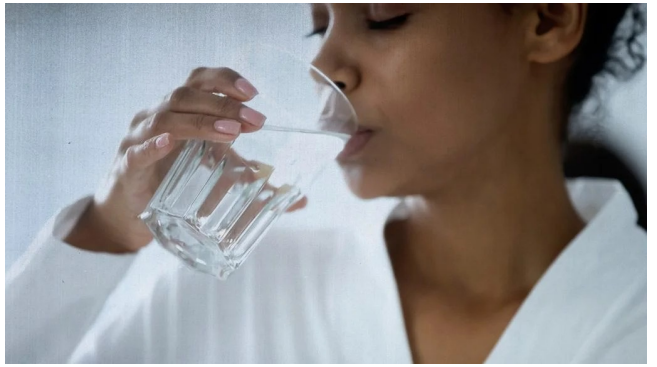
# Environmental Microbiome Analysis

Jeseth Delgado Vela, PhD
Assistant Professor
Department of Civil and Environmental Engineering
Howard University

*Duke University as of August 1, 2023*

CINEMA PARIS

RICE UNIVERSITY Global
Unconventional Wisdom

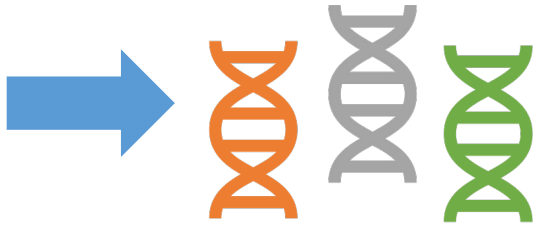
# Icebreaker

Menti.com

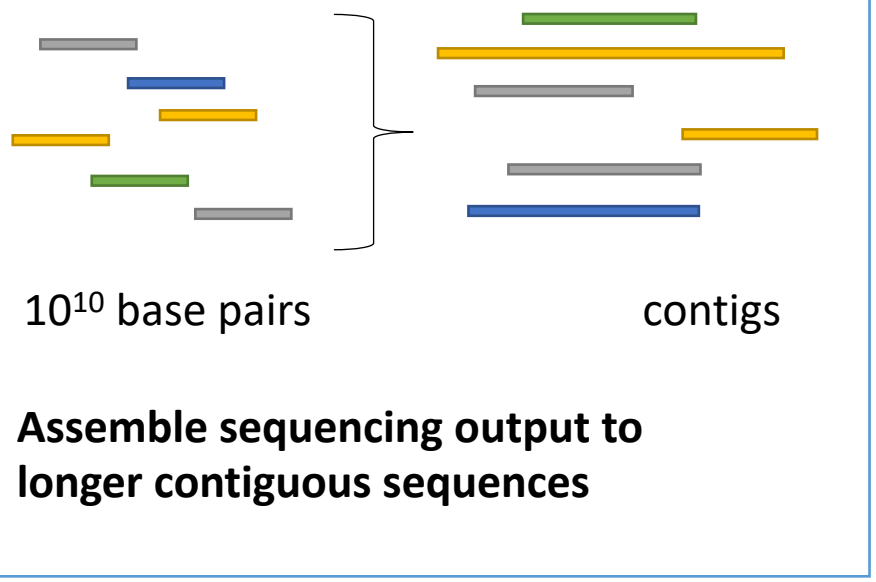
Code: 3506 9500

# Metagenome analysis can be typically includes assembly.

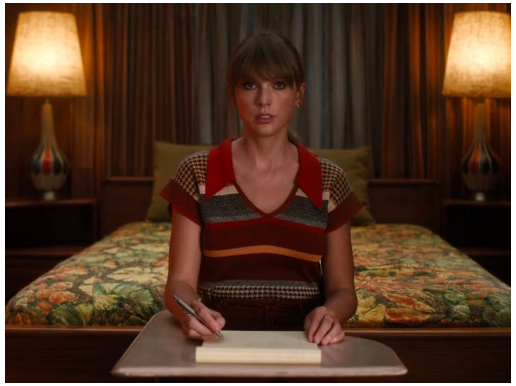

**Collect samples**



**Extract and sequence all DNA from sample**



# de Bruijn graph assembly is a typical approach.



*'Cause the players gonna play, play, play, play, play*
*And the haters gonna hate, hate, hate, hate, hate*
*Baby, I'm just gonna shake, shake, shake, shake, shake*
*I shake it off, I shake it off*

*Heartbreakers gonna break, break, break, break, break*
*And the fakers gonna fake, fake, fake, fake, fake*
*Baby, I'm just gonna shake, shake, shake, shake, shake*
*I **shake it off**, I shake it off*

Use much longer kmers
Step increase in kmer size and consolidate de Bruijn Graphs

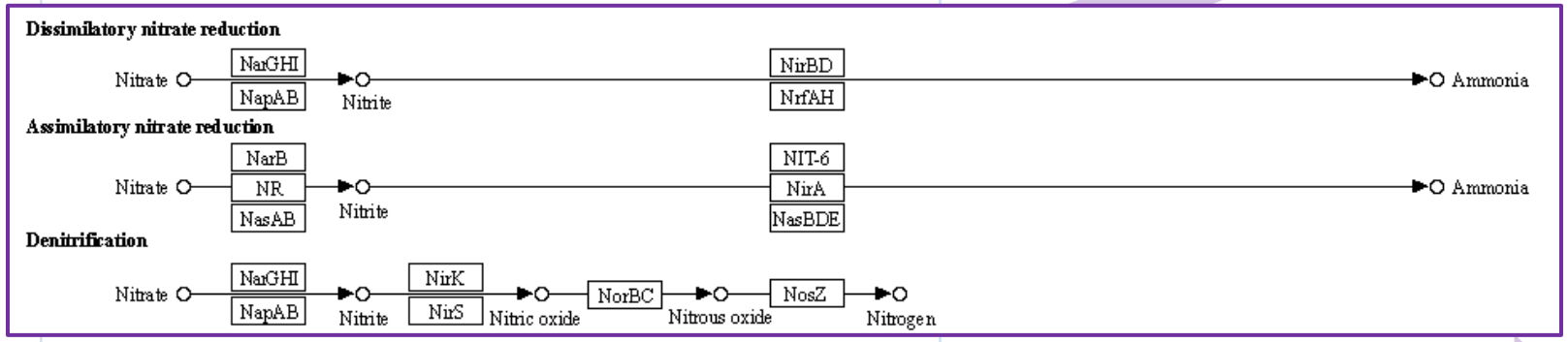
- Resources for deBruin graph assembly approach: https://youtu.be/OY9Q_rUCGDw
- https://www.cs.jhu.edu/~langmea/resources/lecture_notes/19_assembly_dbg2_v2.pdf

# Metagenome analysis can be **gene-** or **genome-centric**.

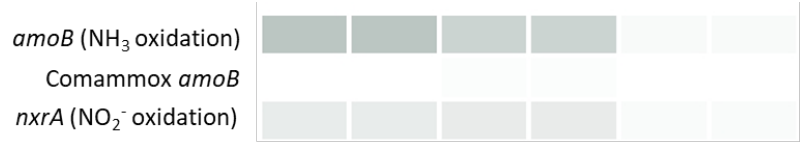
Collect samples

Extract and sequence

What is the functional capacity of individual community members?

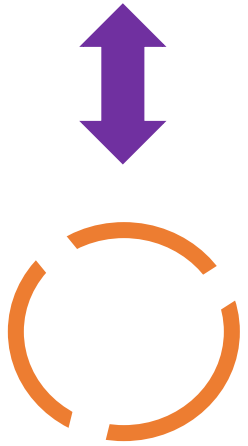



Assemble sequencing output to longer contiguous sequences




**Annotate sequences to develop functional analysis**

**Gene centric:** what functions are catalyzed?

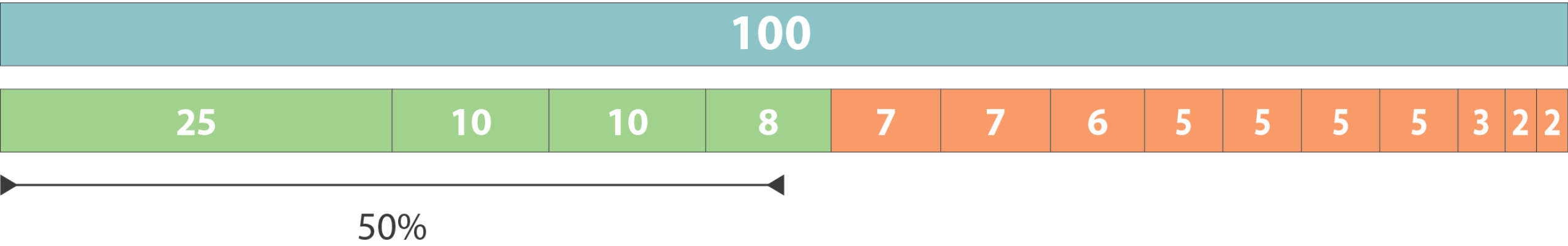


**Bin sequences to develop metagenomic assembled genomes**

**Genome centric:** what organisms are present?

# *Activity: evaluate quality and characteristics of a wastewater metagenomic assembly*

On Github, look up QUAST folder; report.html




**N50=8; L50=4**

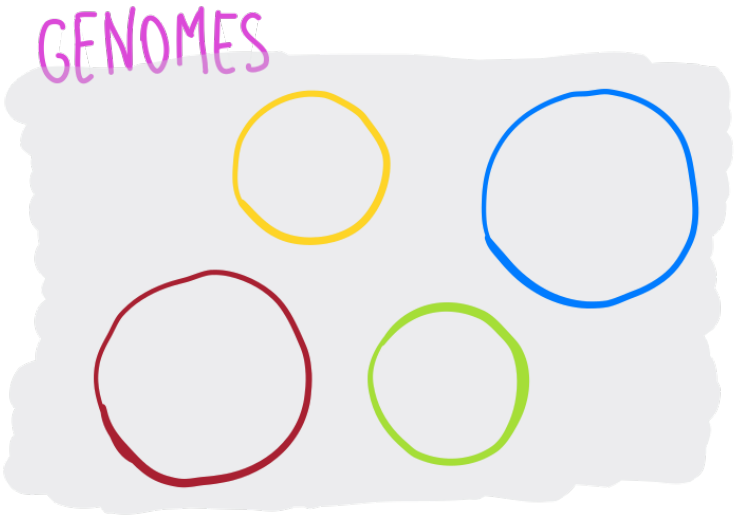
https://training.galaxyproject.org/training-material/topics/assembly/tutorials/assembly-quality-control/slides-plain.html

# **Genome-centric** approaches are challenging, but can help piece together novel genomic capabilities.

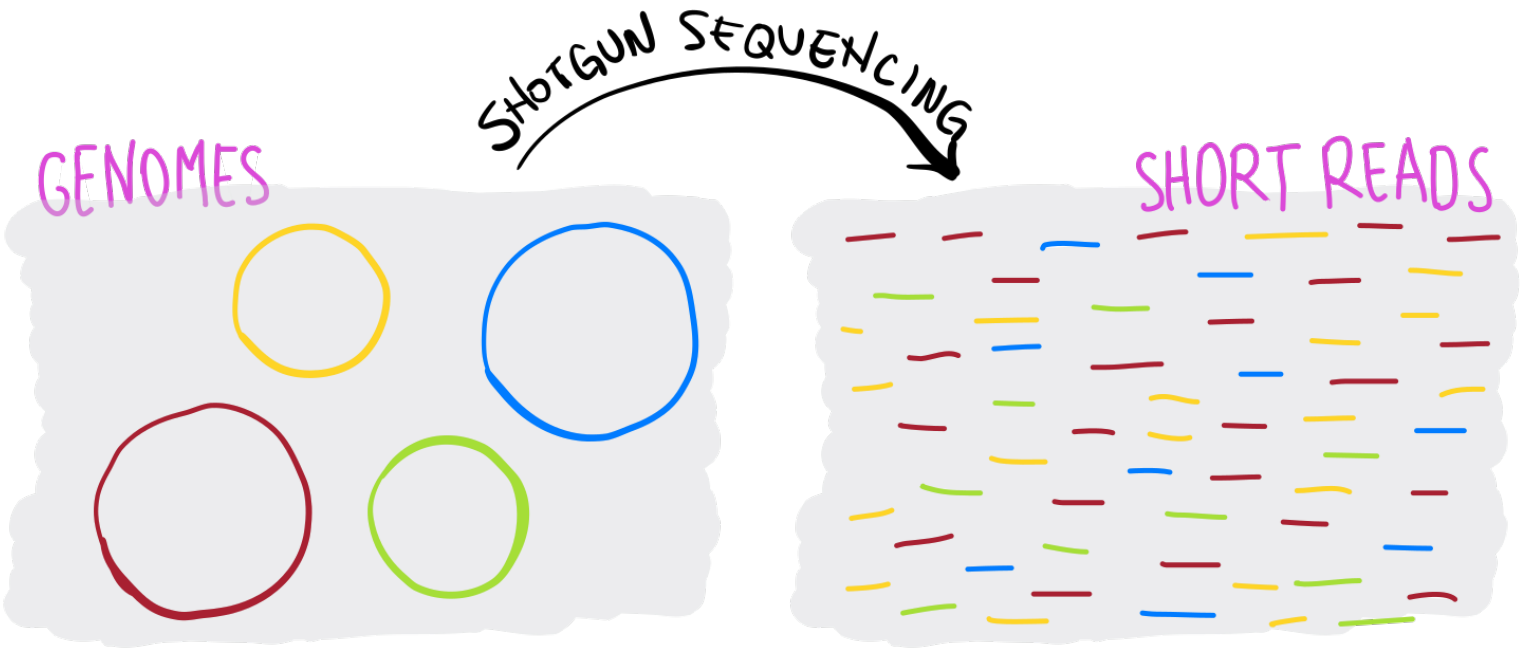
Main rationale for focus on genome reconstruction:

Communities are grouping of individuals, aka cells, where functional traits are linked within the constraints of a cellular envelope.

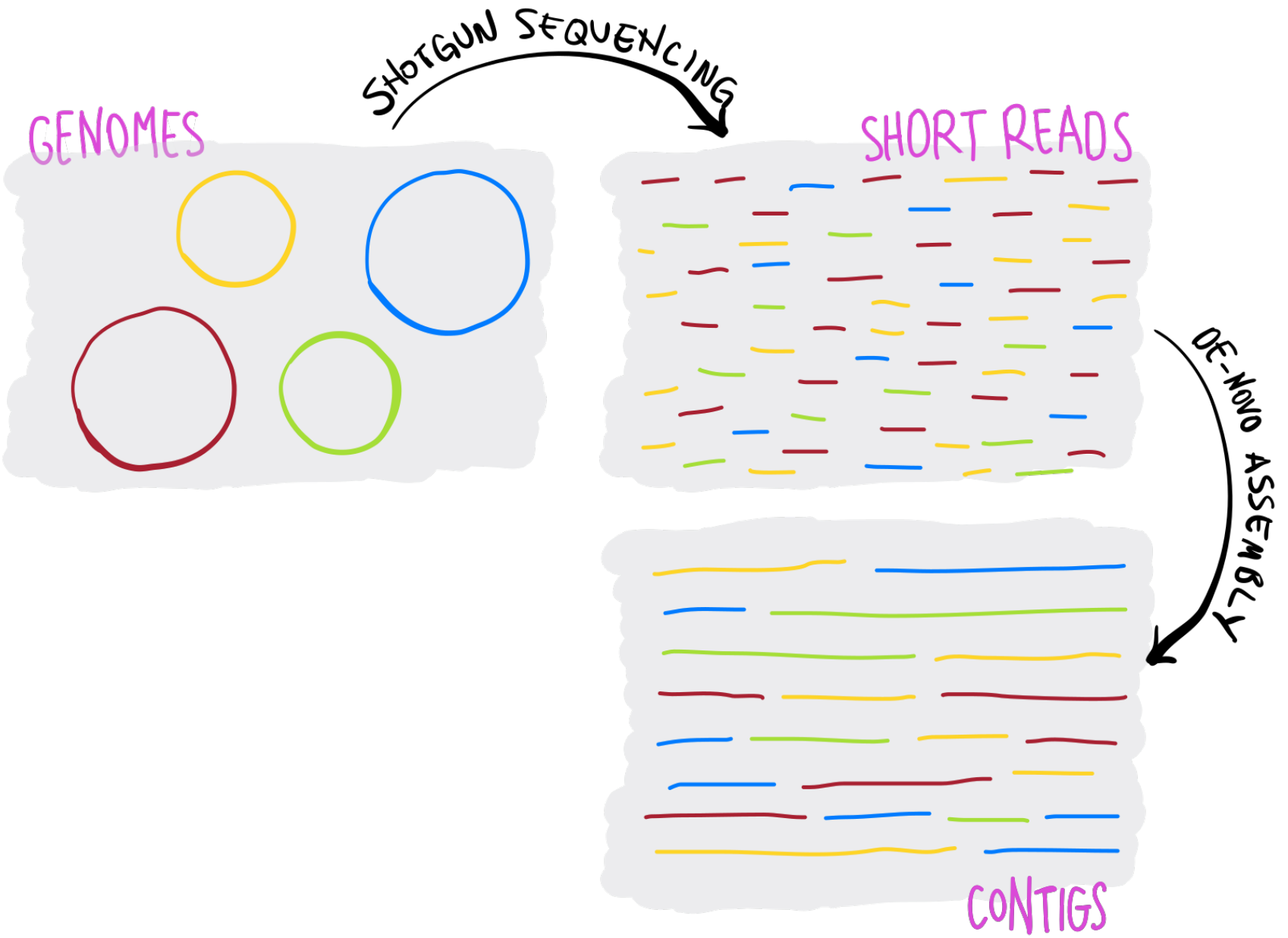
Communities are not piles of genes.



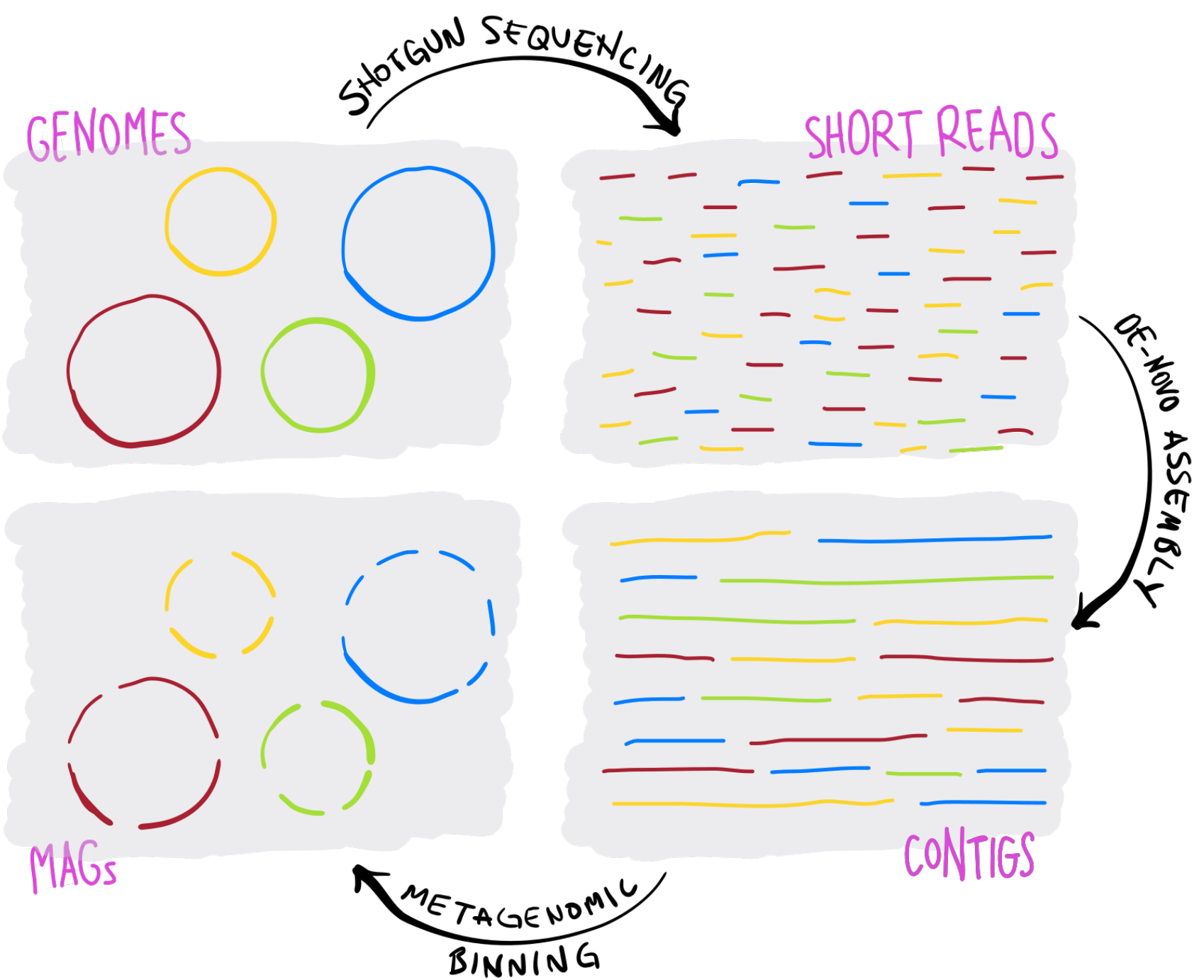

Slides courtesy of https://merenlab.org/momics



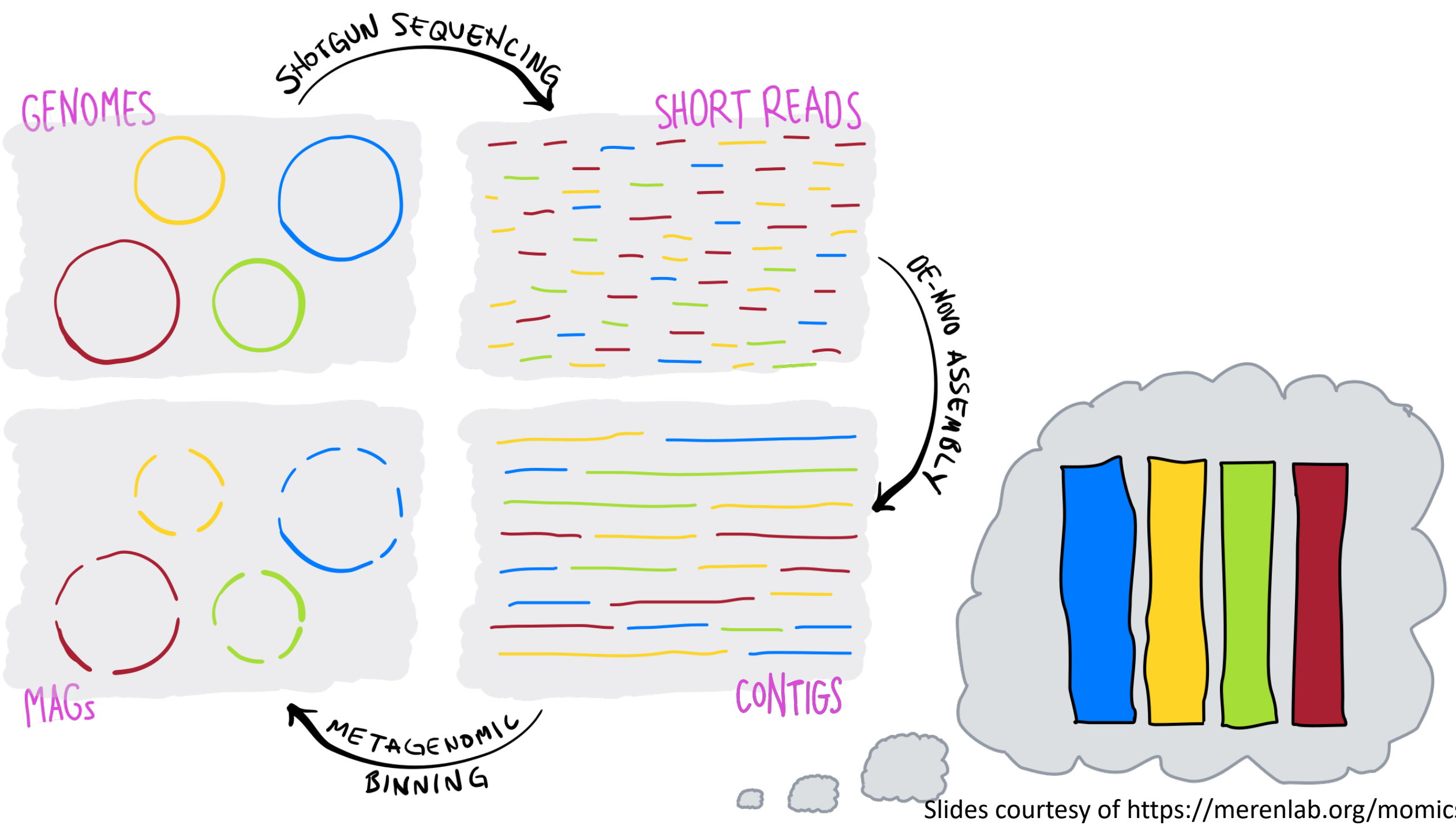

Slides courtesy of https://merenlab.org/momics



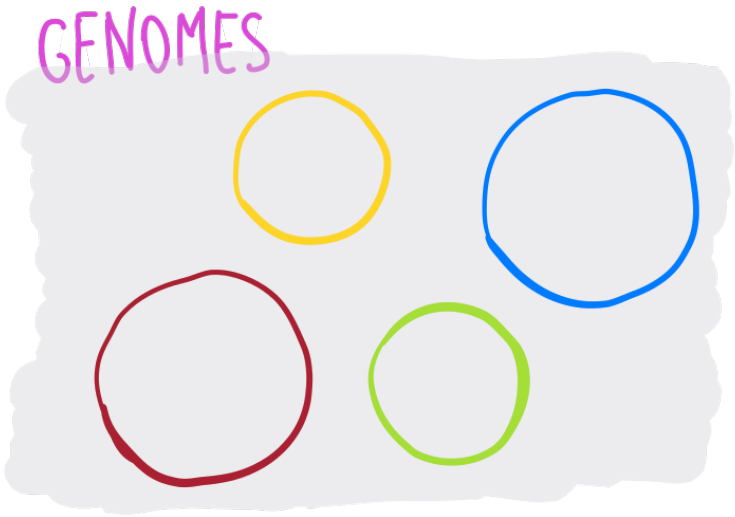

Slides courtesy of https://merenlab.org/momics

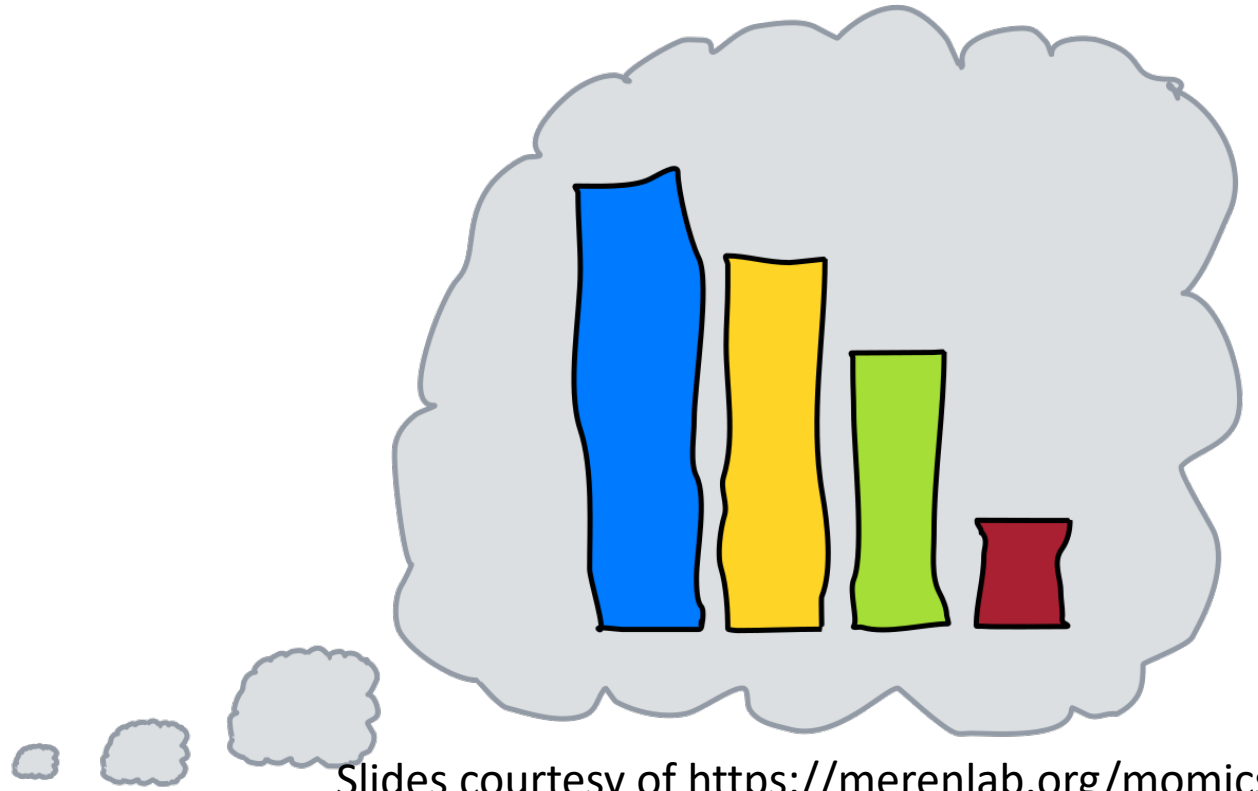



Slides courtesy of https://merenlab.org/momics

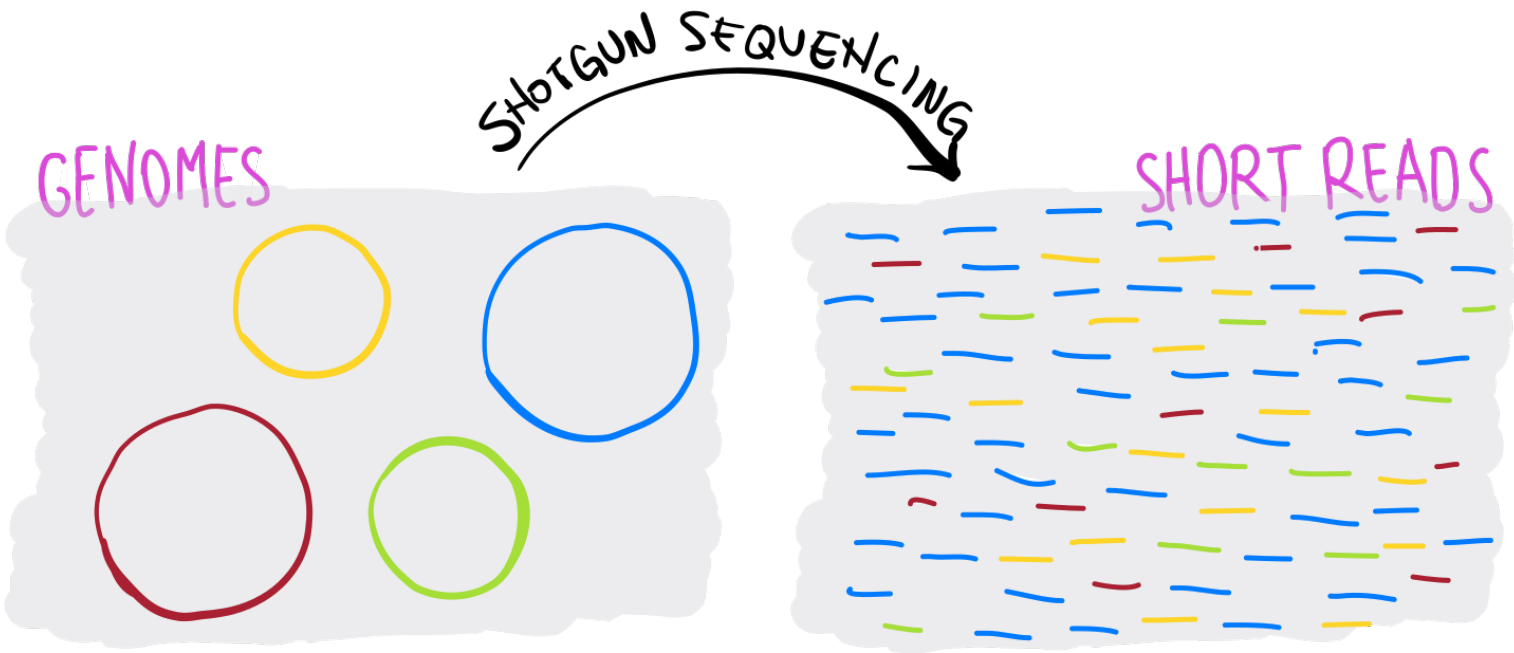


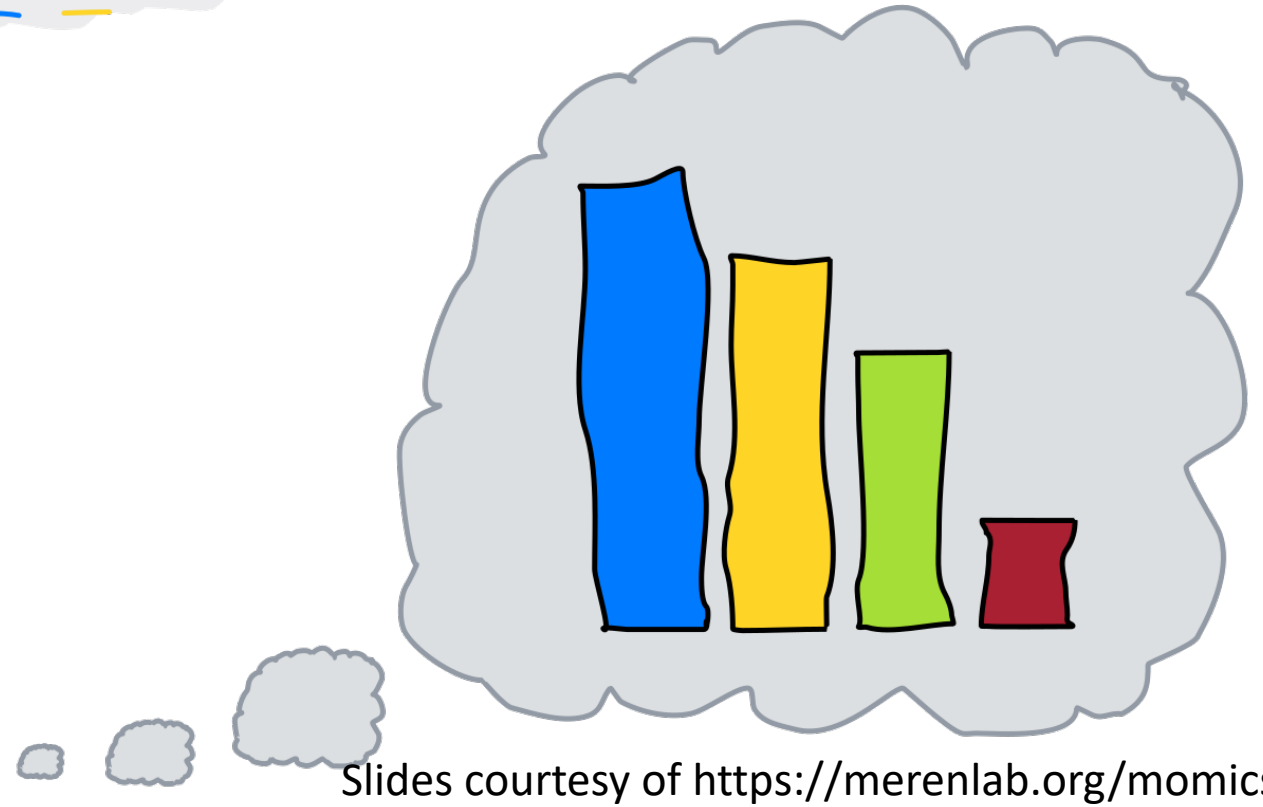

Slides courtesy of https://merenlab.org/momics



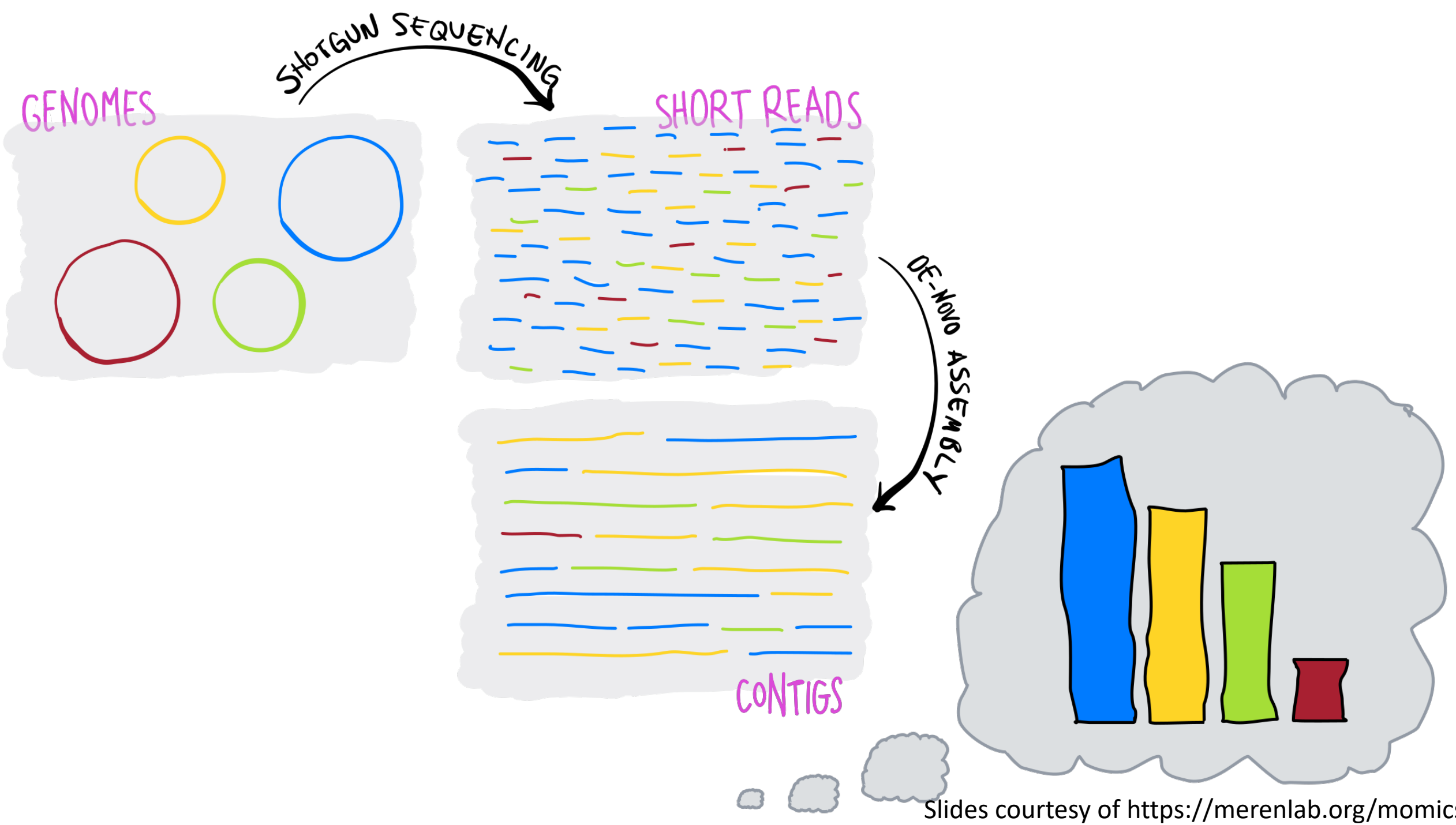



Slides courtesy of https://merenlab.org/momics

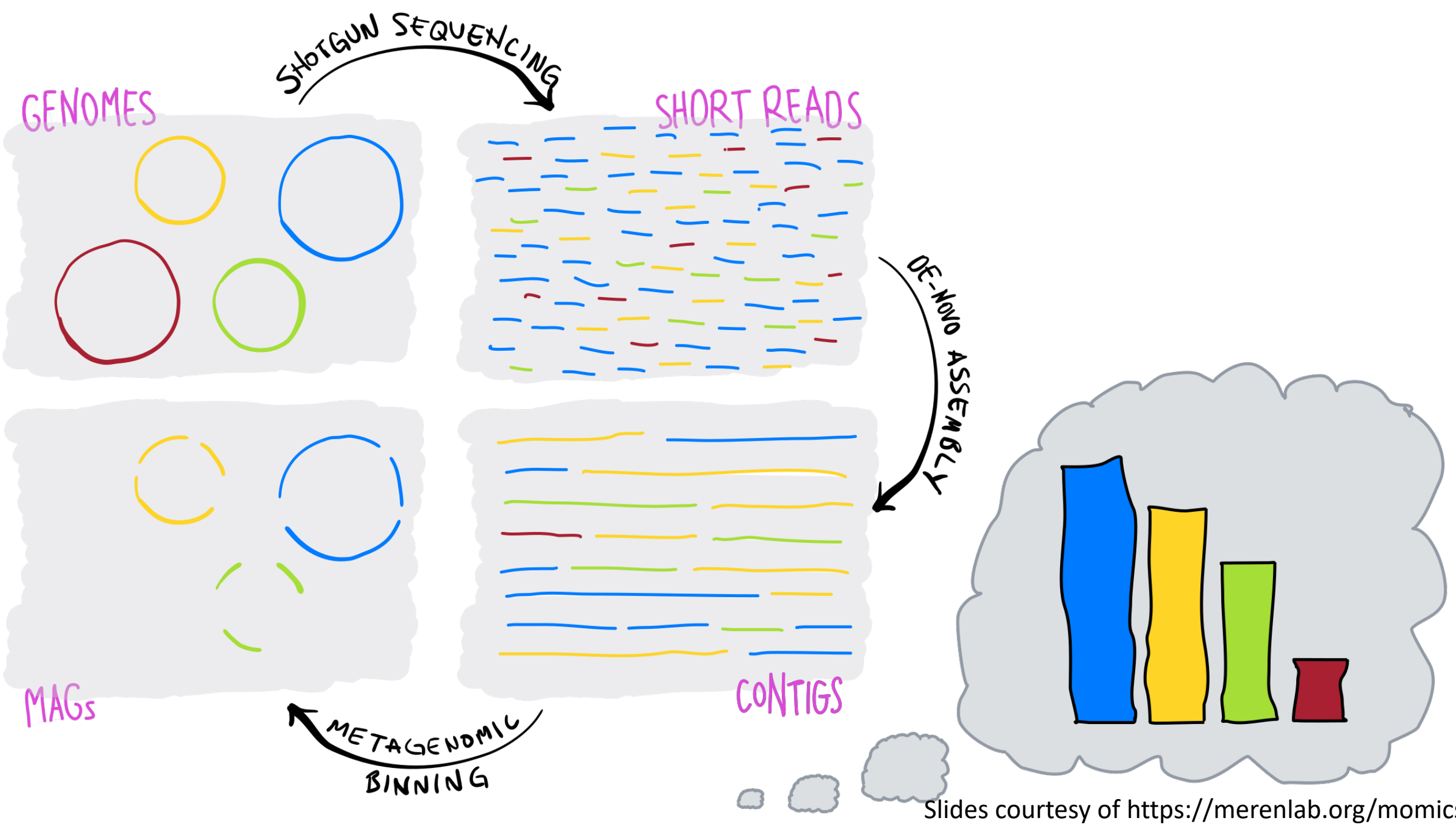


Slides courtesy of https://merenlab.org/momics




Slides courtesy of https://merenlab.org/momics

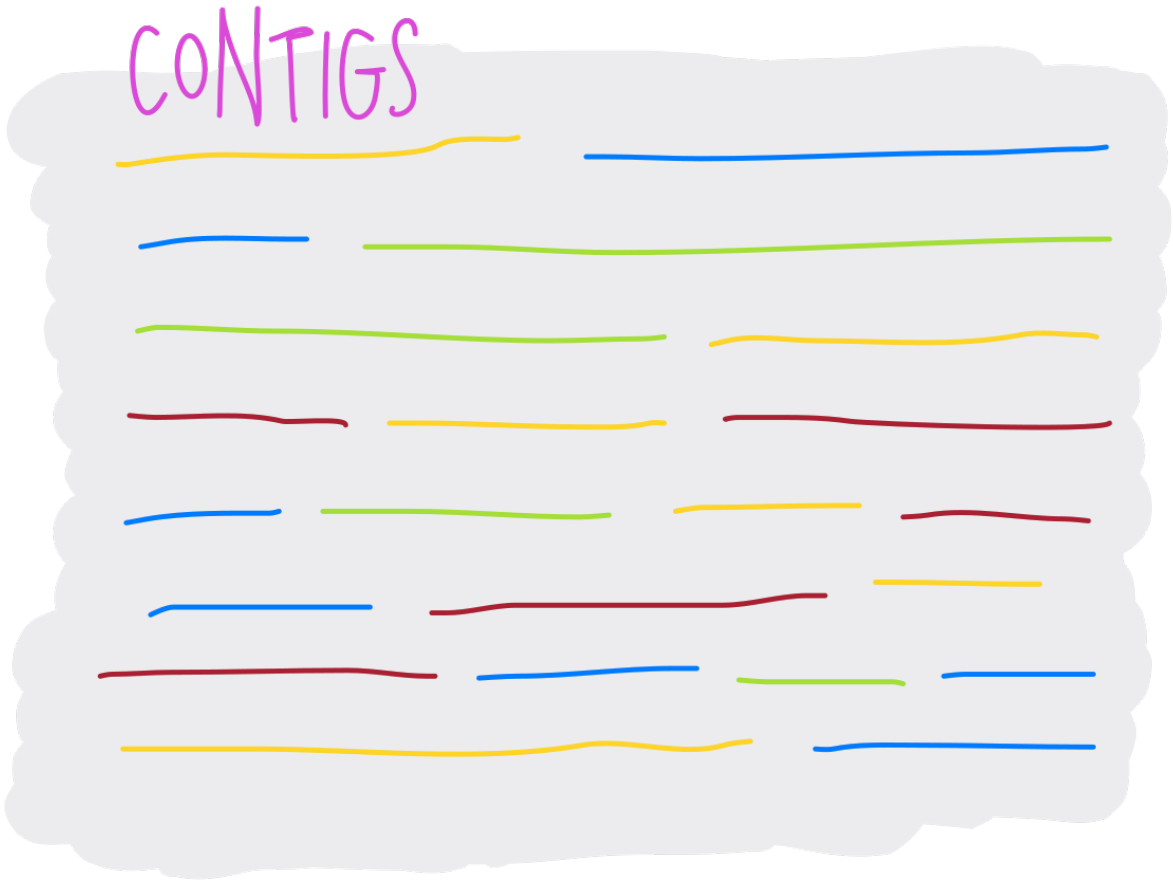



Slides courtesy of https://merenlab.org/momics

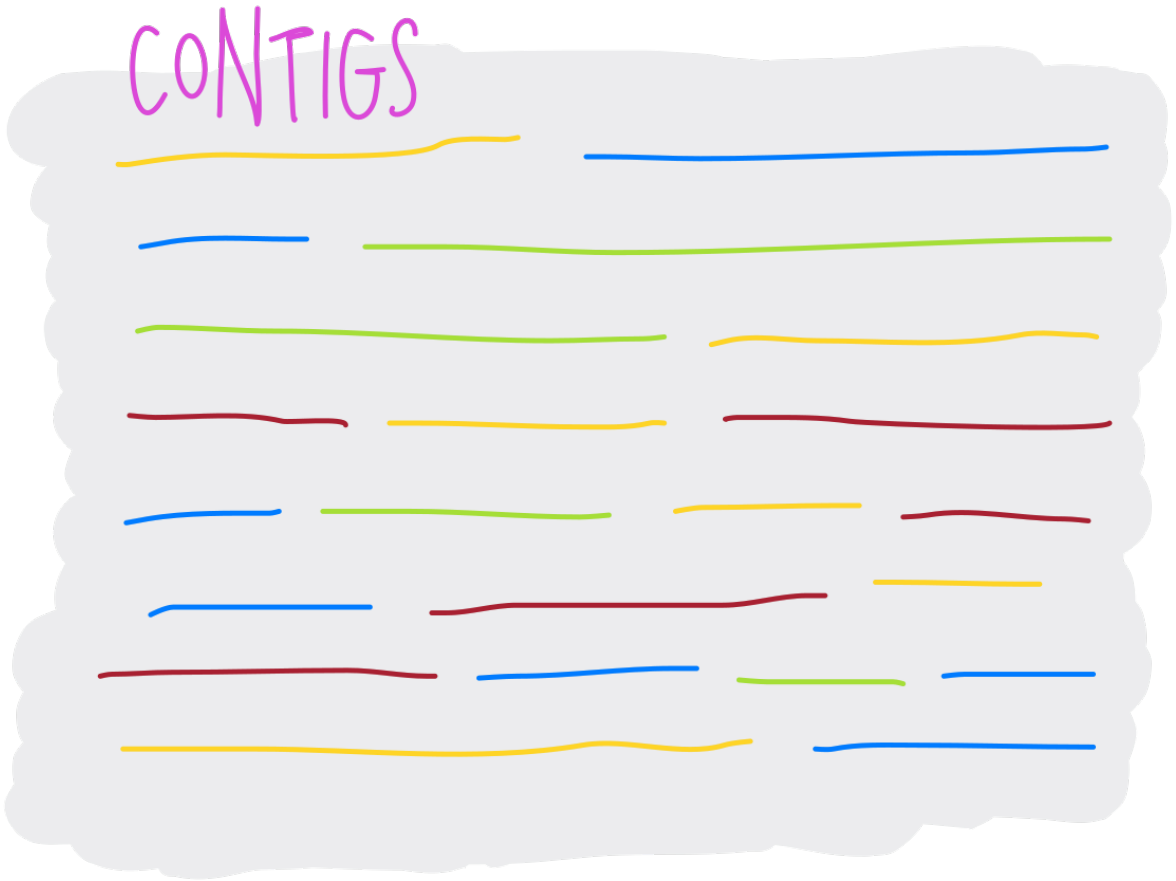
# Low abundance organisms require deeper sequencing for genome reconstruction.

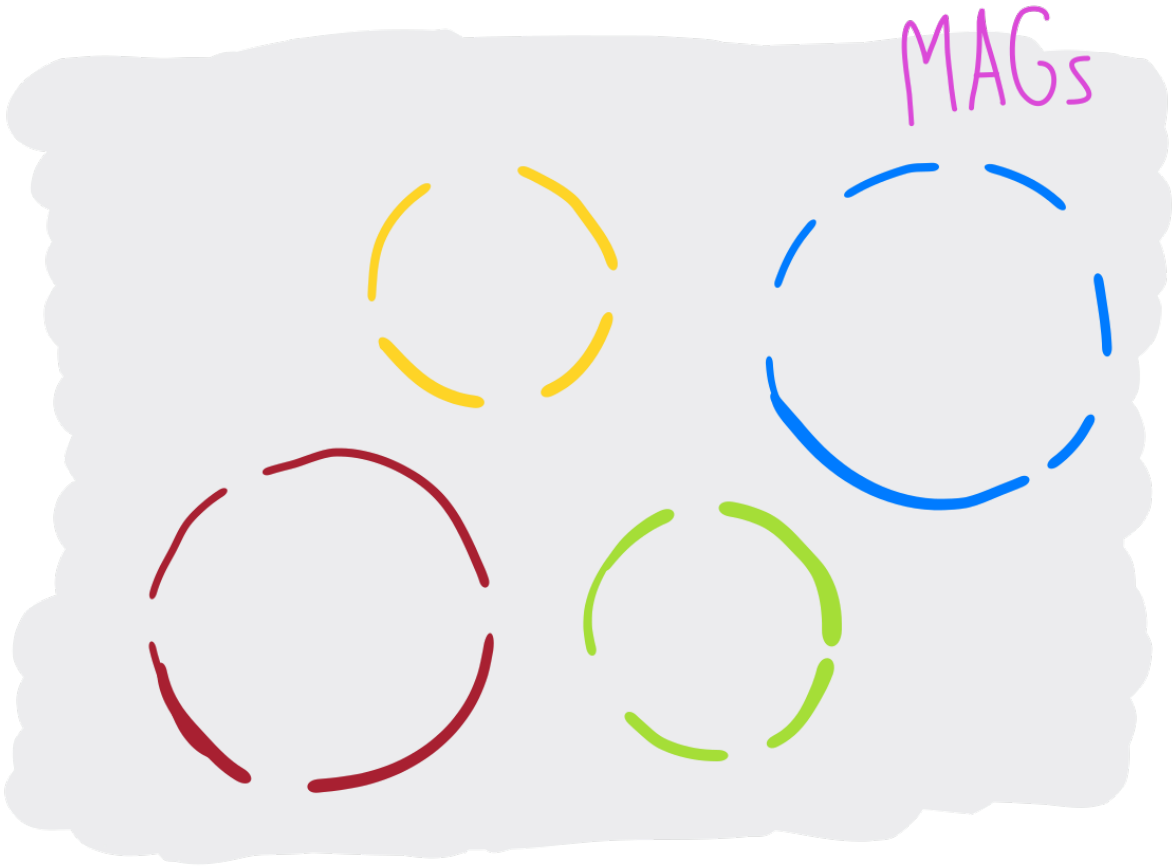
If you want to obtain 20X sequence coverage of an organism present at 0.1% of the community, assuming an average genome size of 3 Mbp, how many Gbp of sequence do you need to acquire?

[20x is minimum coverage needed to allow genome reconstruction in complex samples – Luo, ISMEJ 2012]

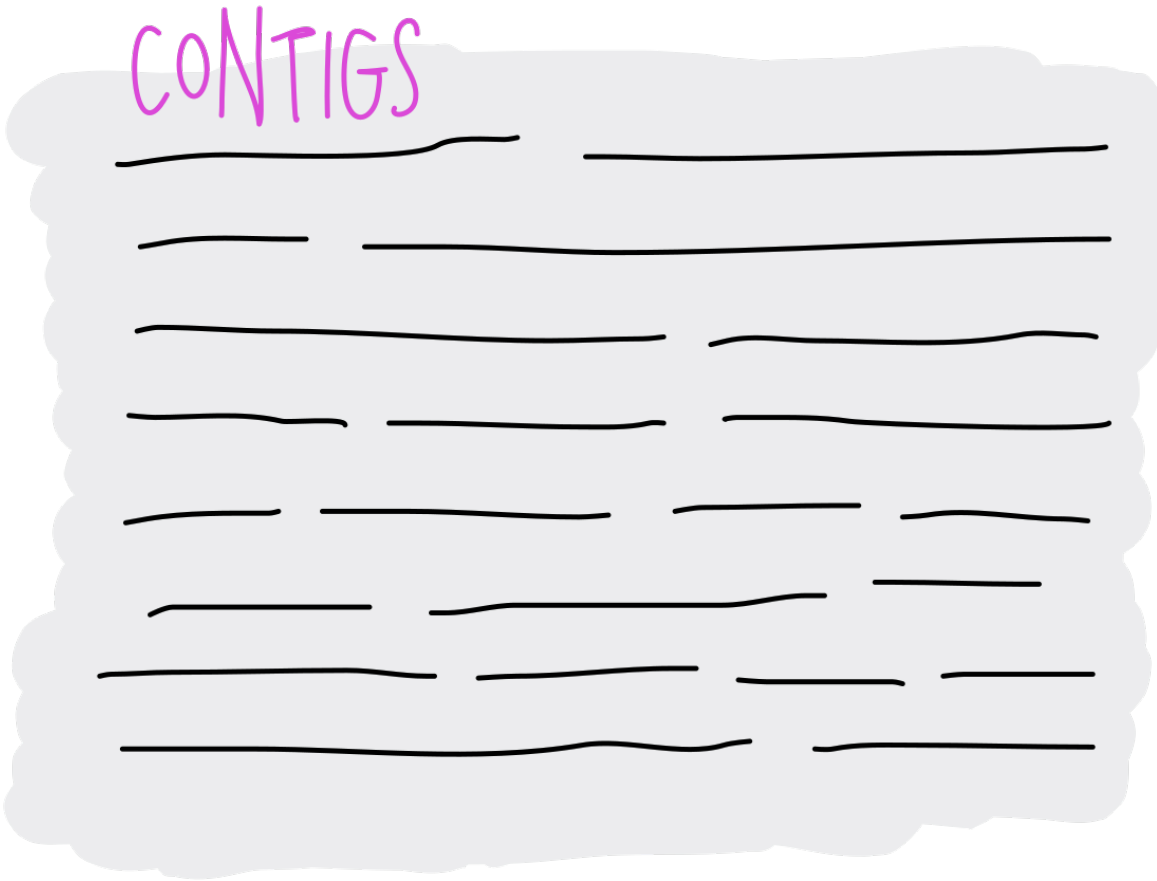


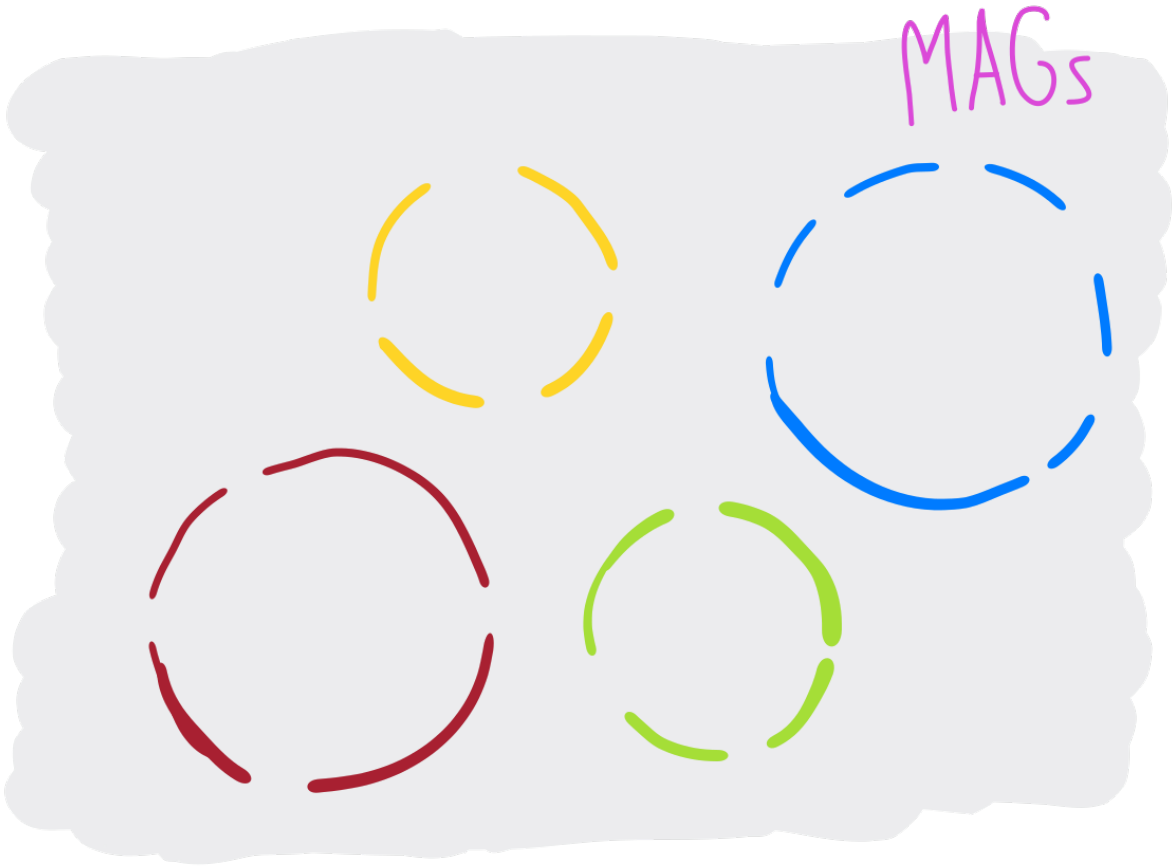

Slides courtesy of https://merenlab.org/momics



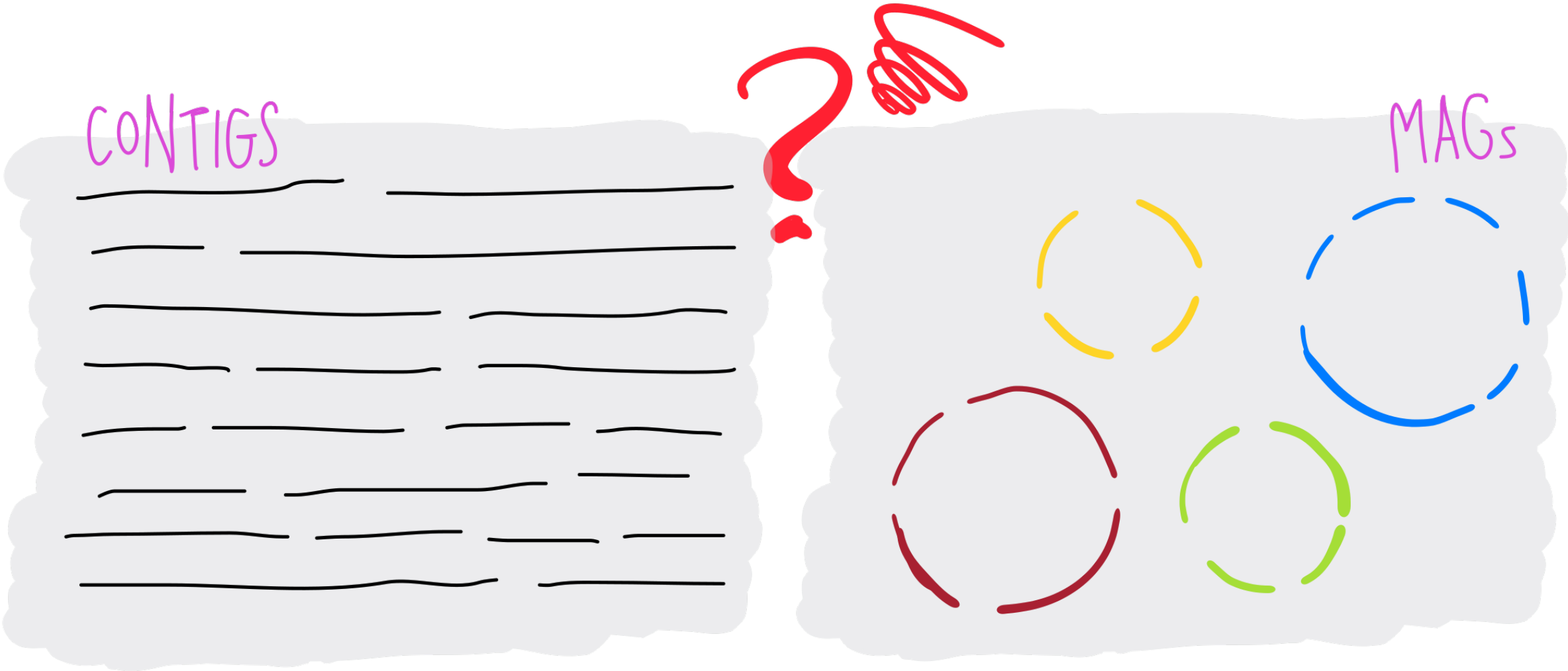




Slides courtesy of https://merenlab.org/momics




Slides courtesy of https://merenlab.org/momics

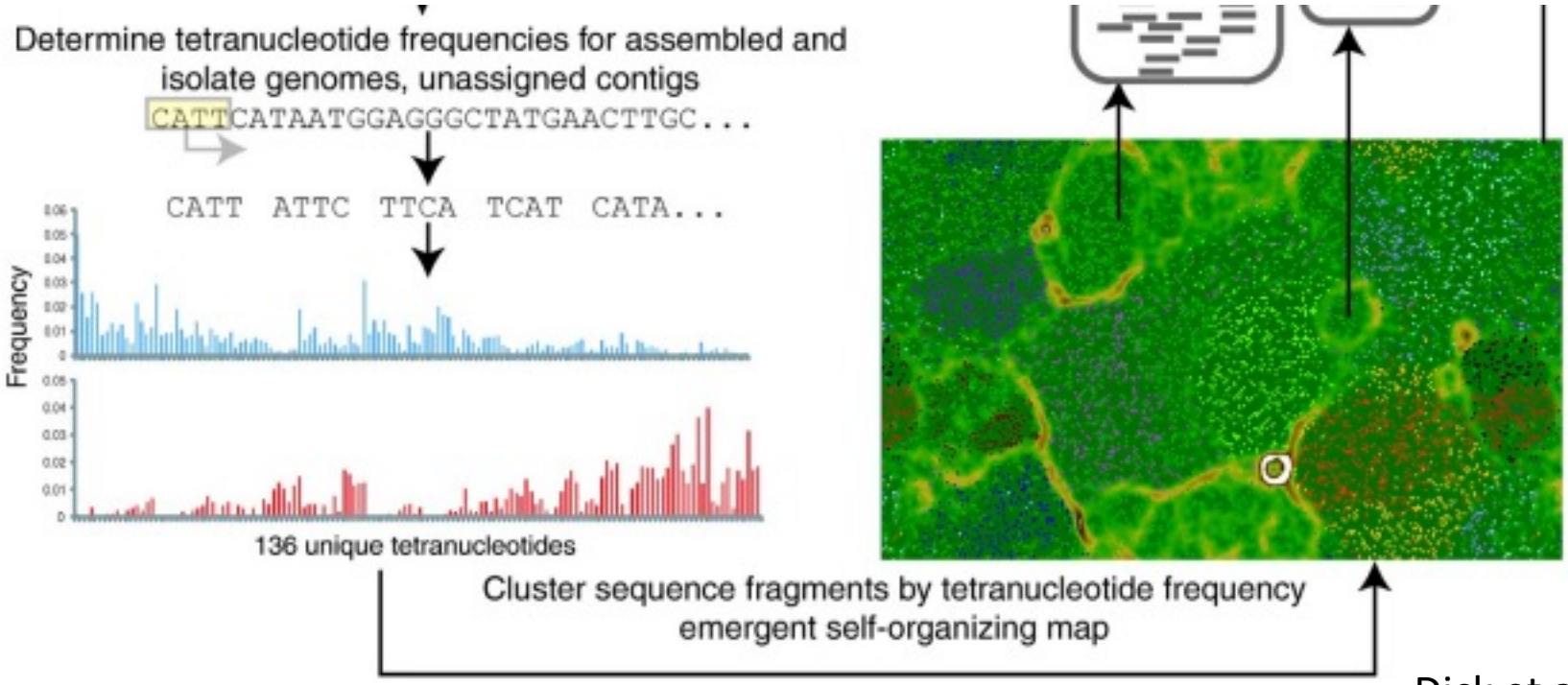



Slides courtesy of https://merenlab.org/momics

# BREAK

**Codon bias**: Multiple codons for the same amino acid are not used equally.

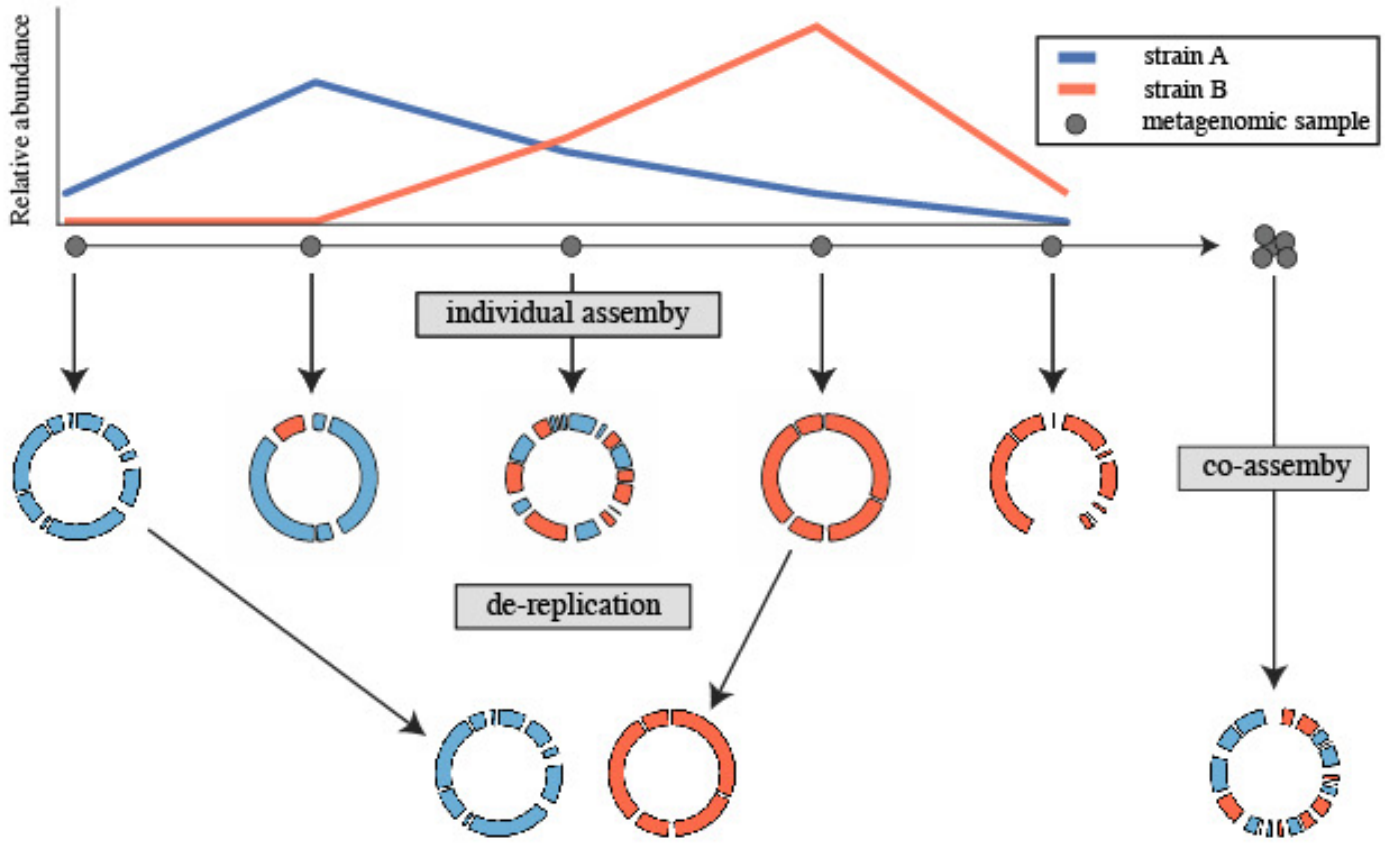
- varies between organisms
- correlated with tRNA availability/concentration
- leveraged for genome binning




Dick et al. (2009)

# OK, I want to do genome centric analysis and made a bunch of MAGs… now what?

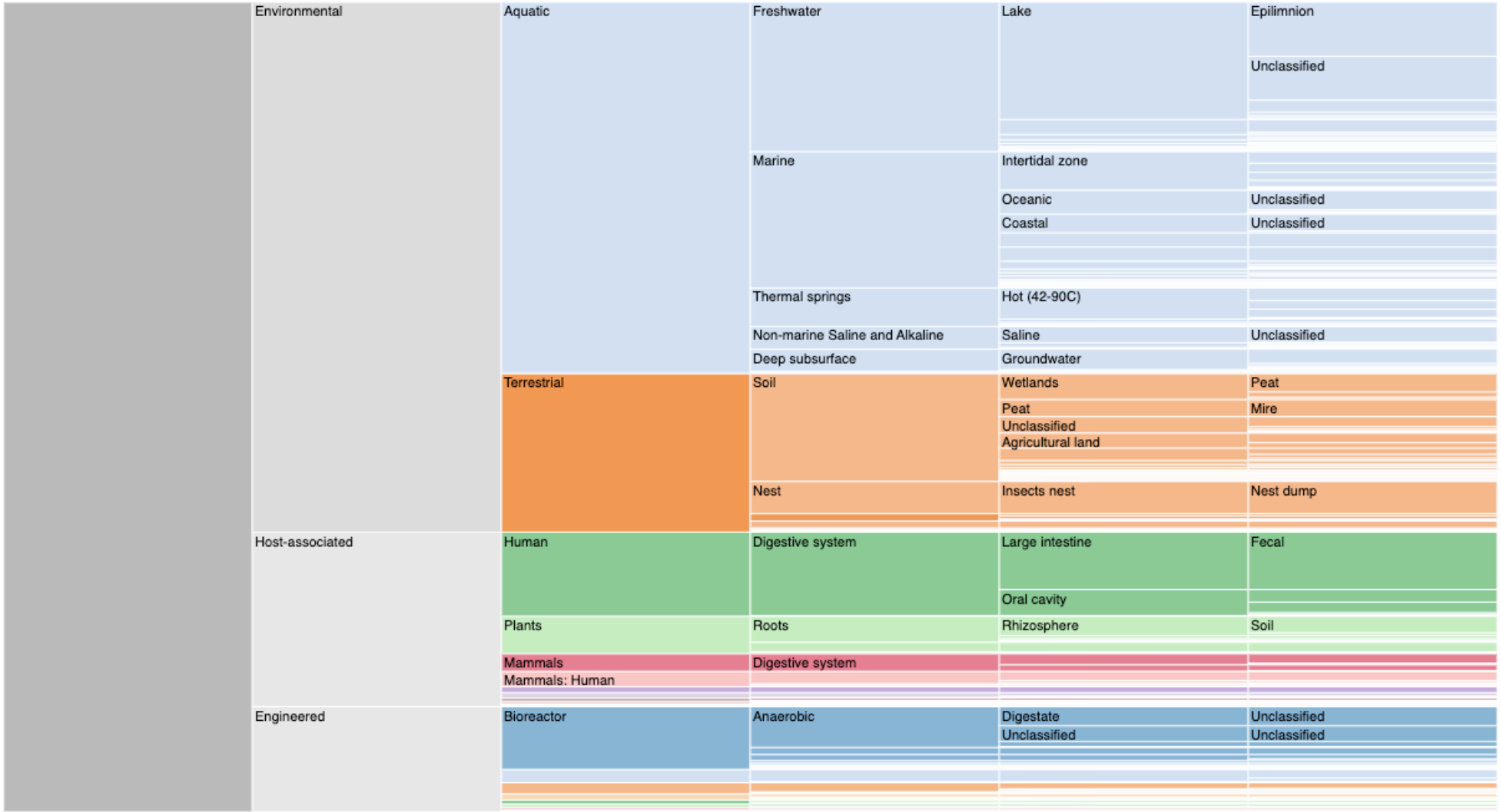
- Evaluate quality: completeness & contamination
  - Based on the number of known single copy genes in the bin
  - Too many- MAG is contaminated; too few- MAG is contaminated
- Dereplicate based on quality




https://drep.readthedocs.io/en/latest/overview.html

# Compare your MAGs to other pure culture genomes… or other MAGs!

Img.jgi.doe.gov




d metagenome binning tool, MetaBAT
coverage information, and parameter '-
nsensus taxonomic lineage is then
MG taxonomy of individual contigs, and
consensus are automatically removed.
are estimated based on the recovery of
CheckM, and are reported according to
ompliant Minimum Information on a
nally, the taxonomic lineage per bin is
member scaffolds taxonomic lineage,
ins classified as Medium (MQ) or High
2) quality bins are further explored for
genome quality (completeness and
on single copy marker gene sets using
n and less than 10% contamination are

Bins by Ecosystem

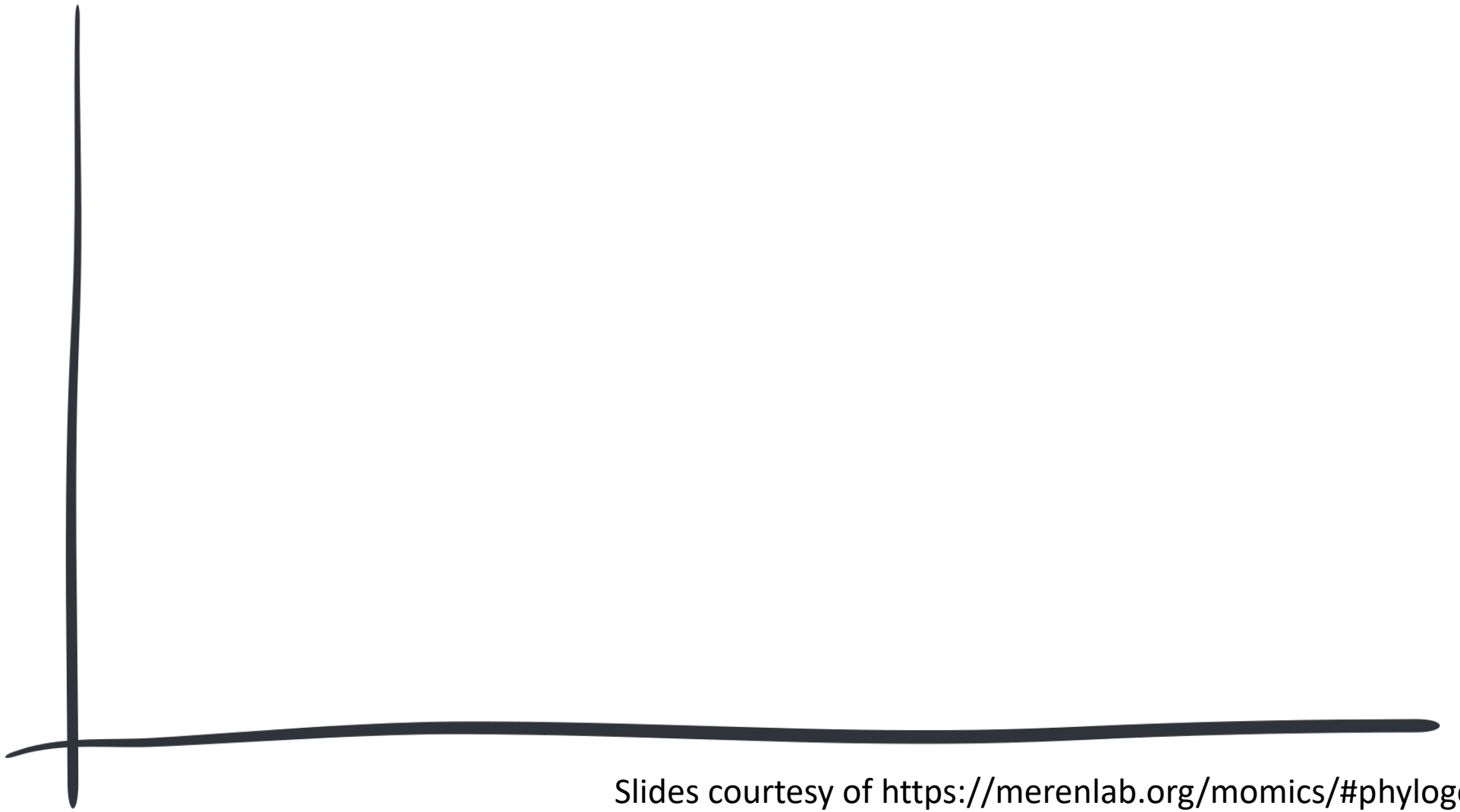
# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS

Slides courtesy of https://merenlab.org/momics/#phylogenomics

# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)

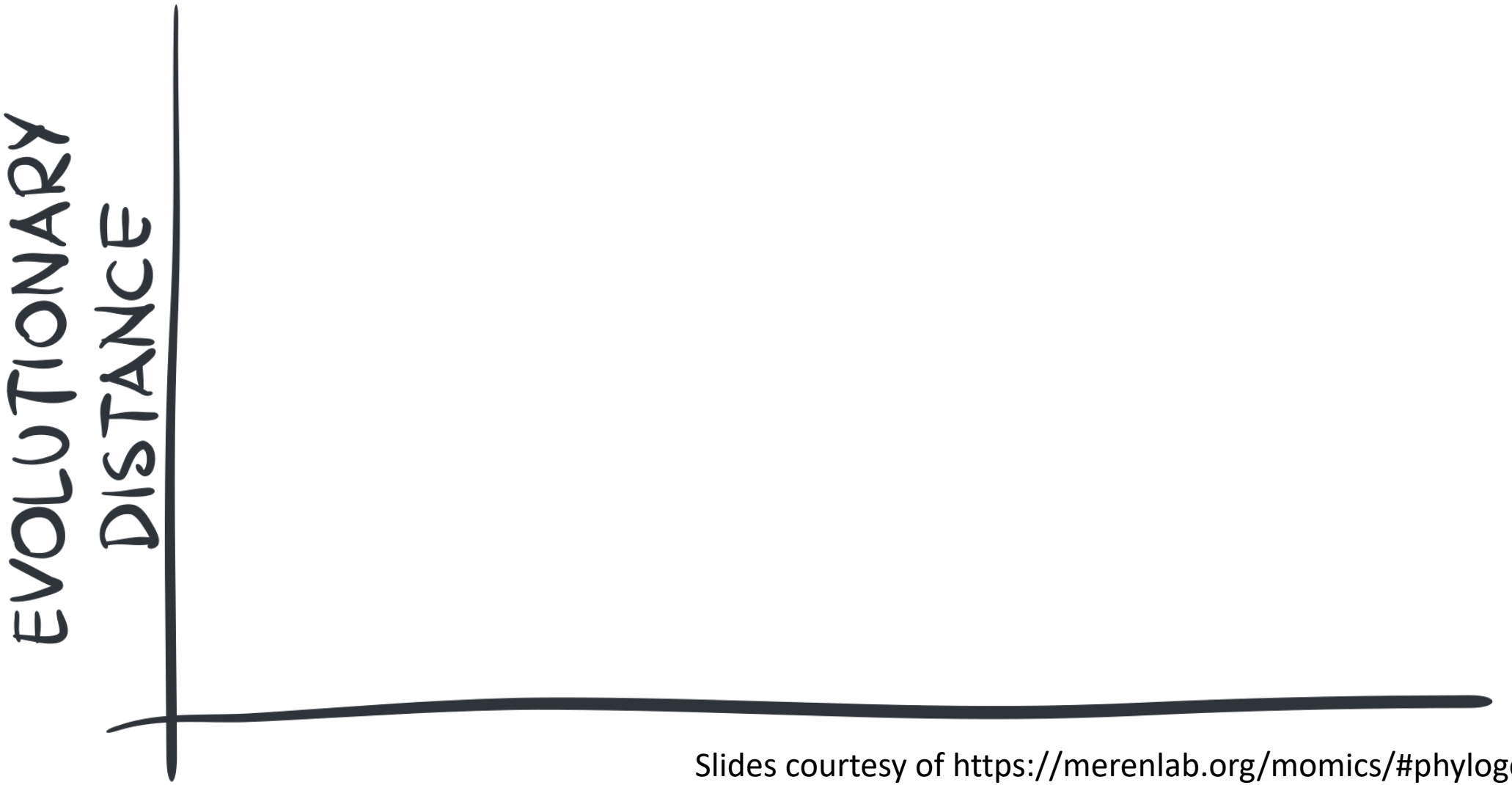
Slides courtesy of https://merenlab.org/momics/#phylogenomics

# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)



Slides courtesy of https://merenlab.org/momics/#phylogenomics

# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)

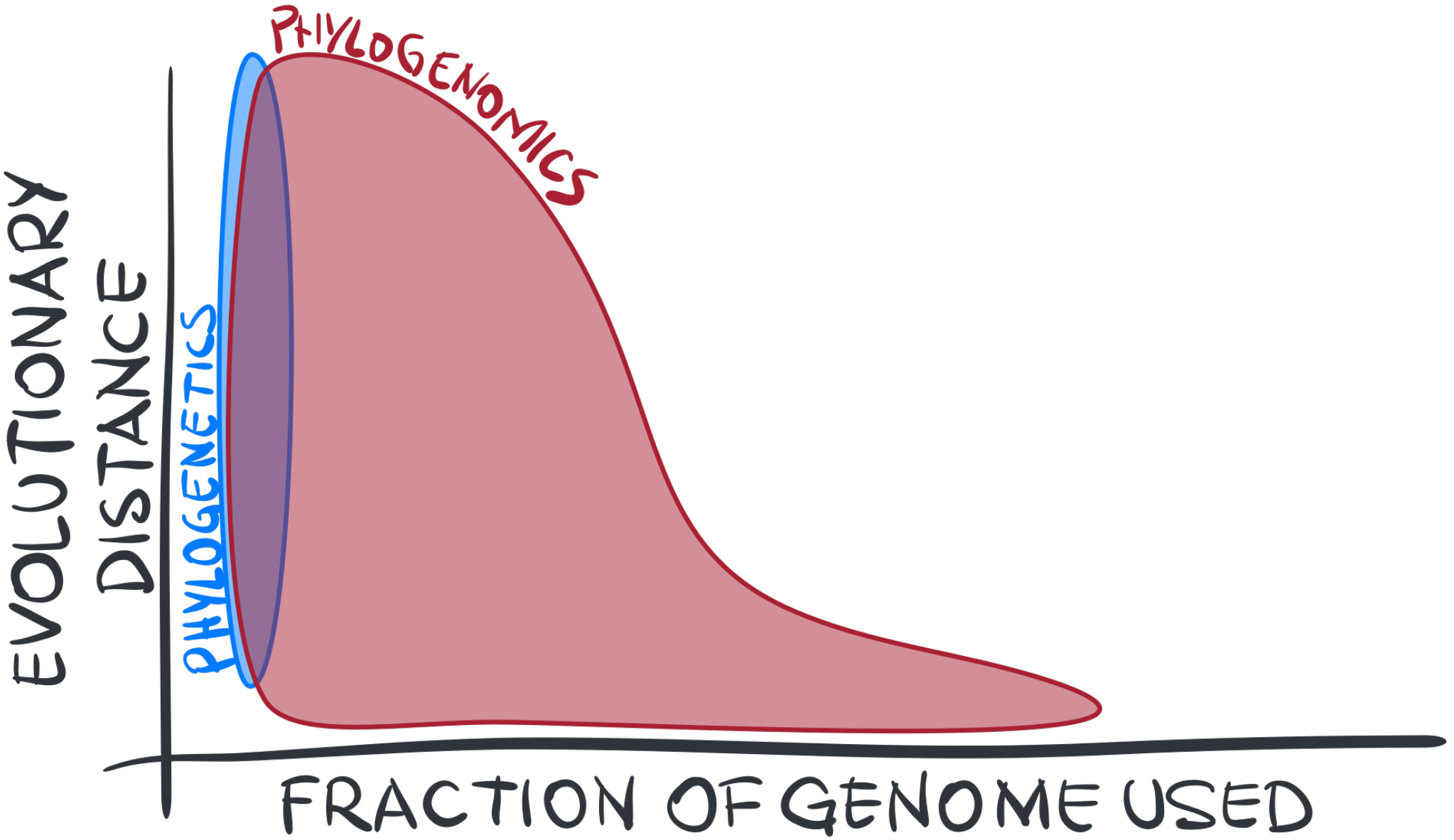



Slides courtesy of https://merenlab.org/momics/#phylogenomics

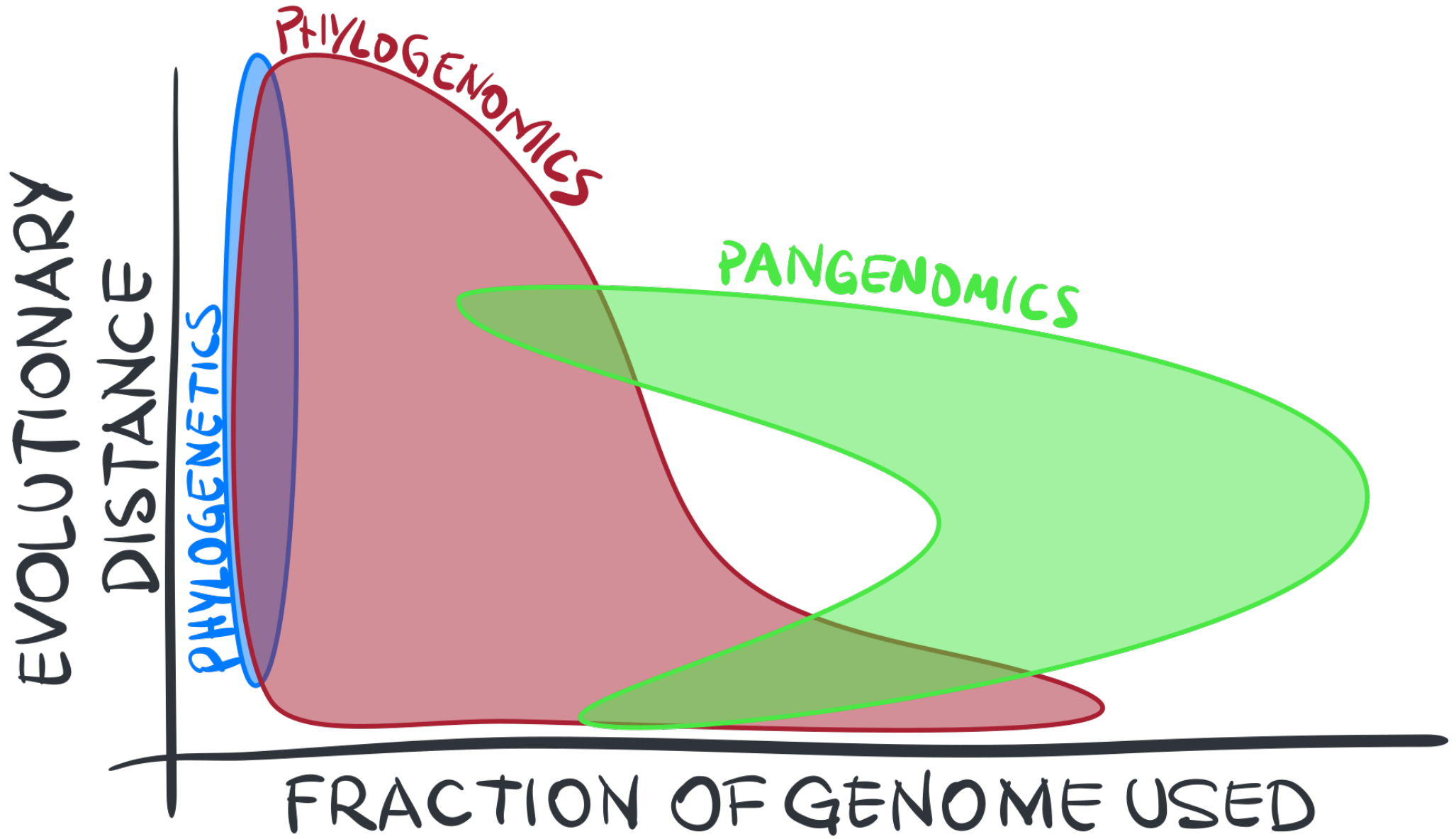
# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)

# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)

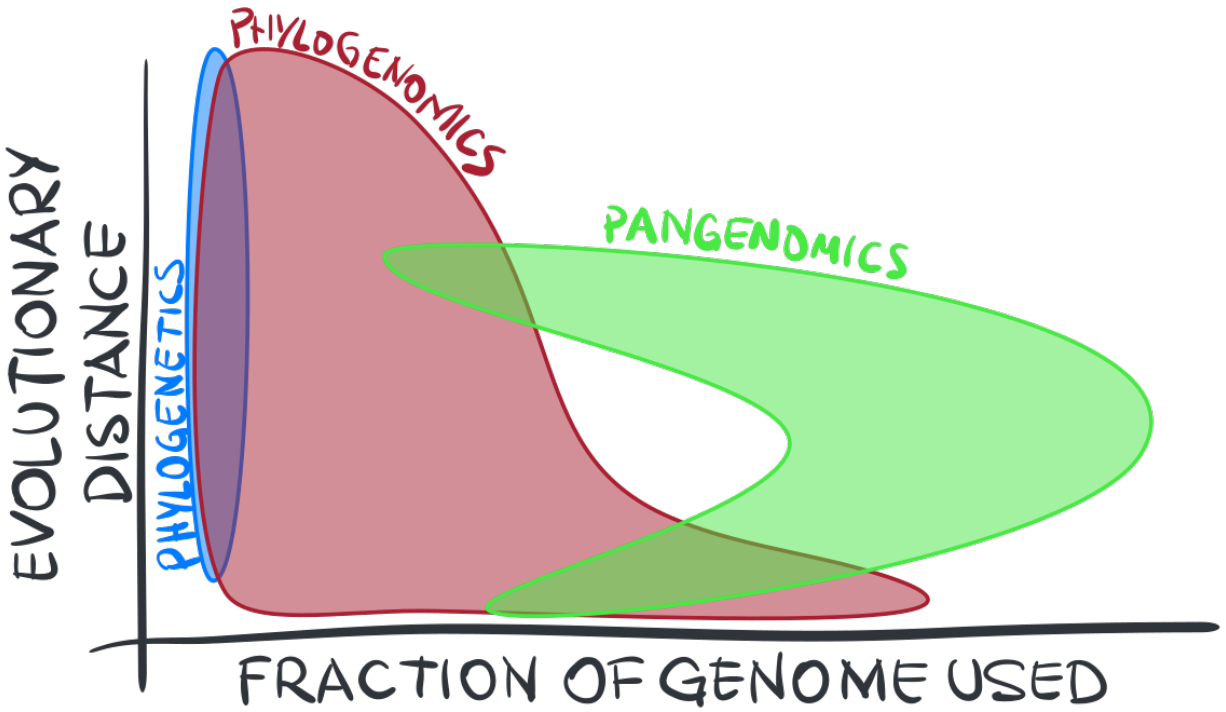
# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)



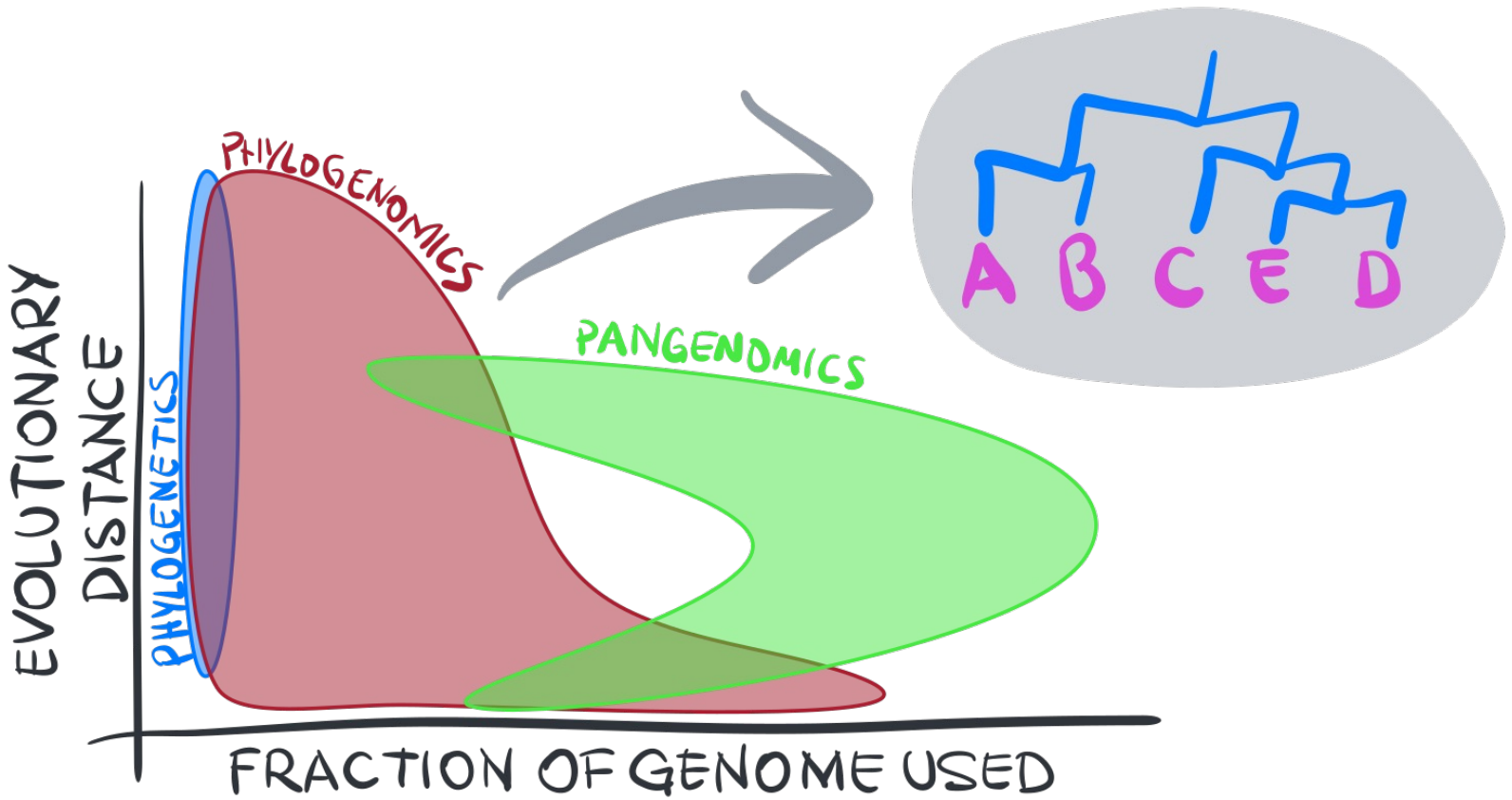
# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)



# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)

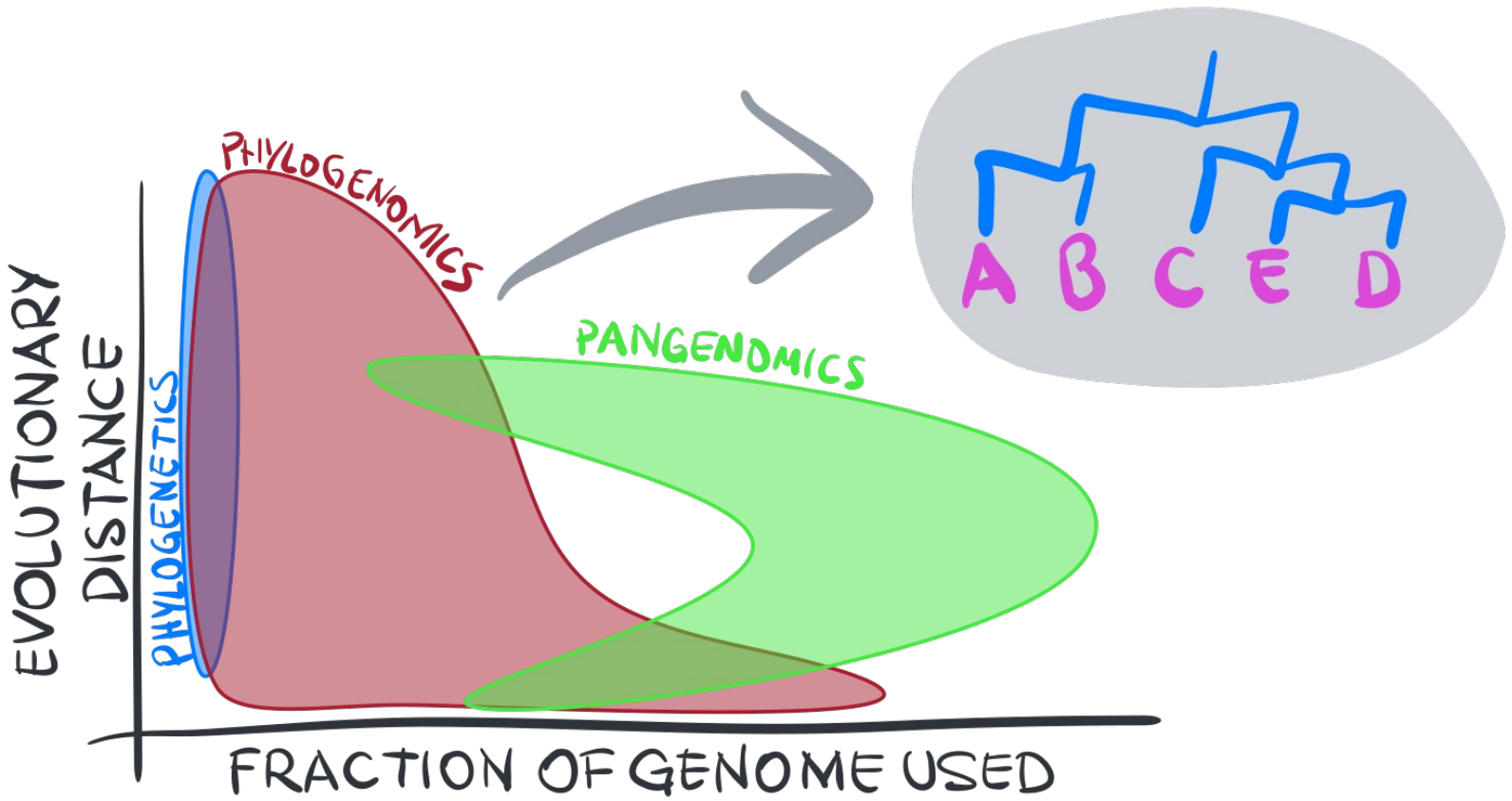


# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)

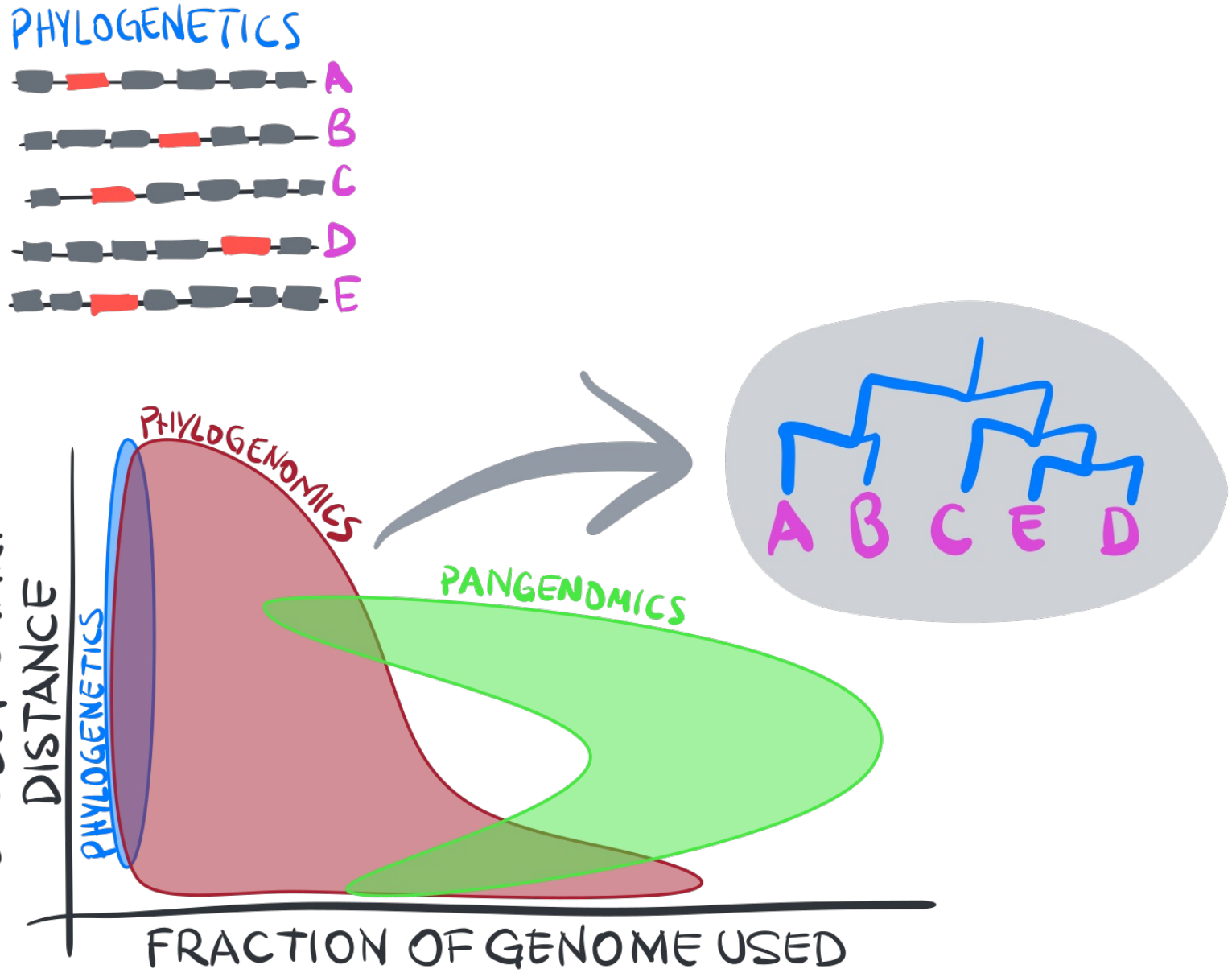


# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)

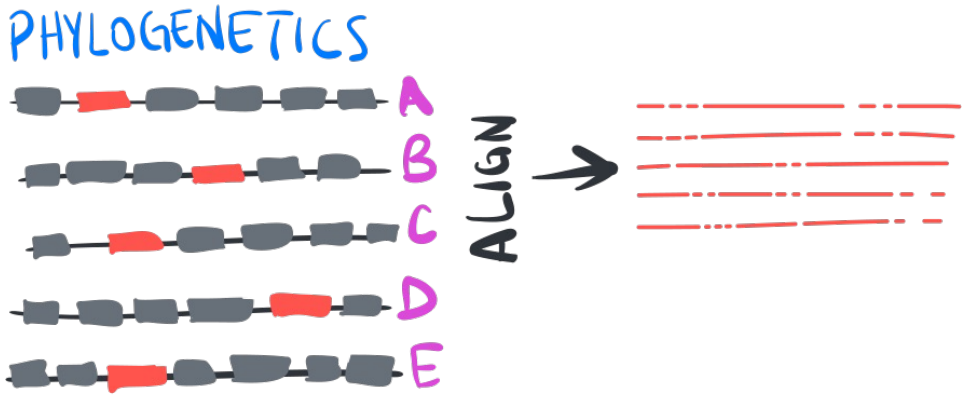
PHYLOGENETICS

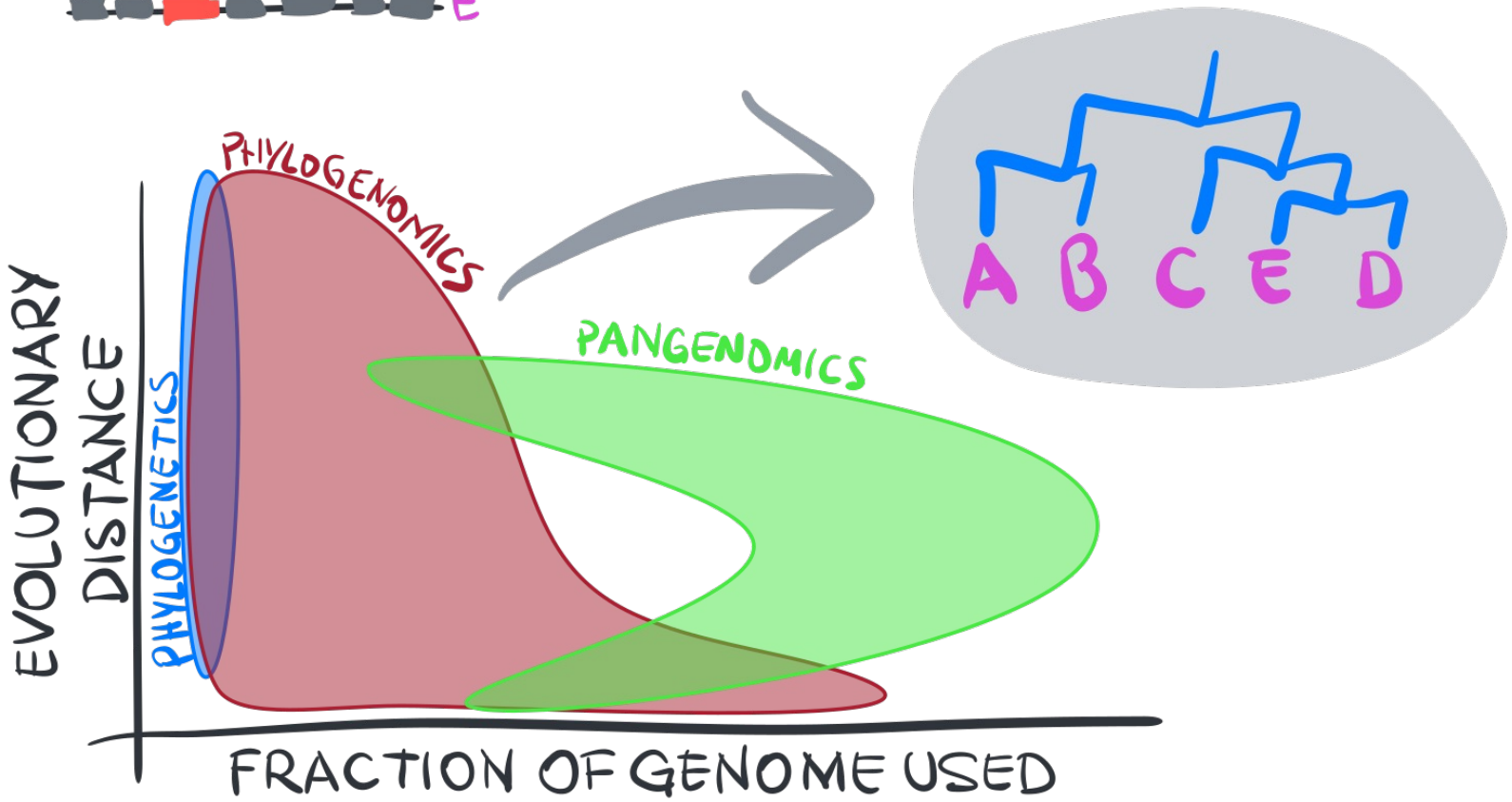


# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)

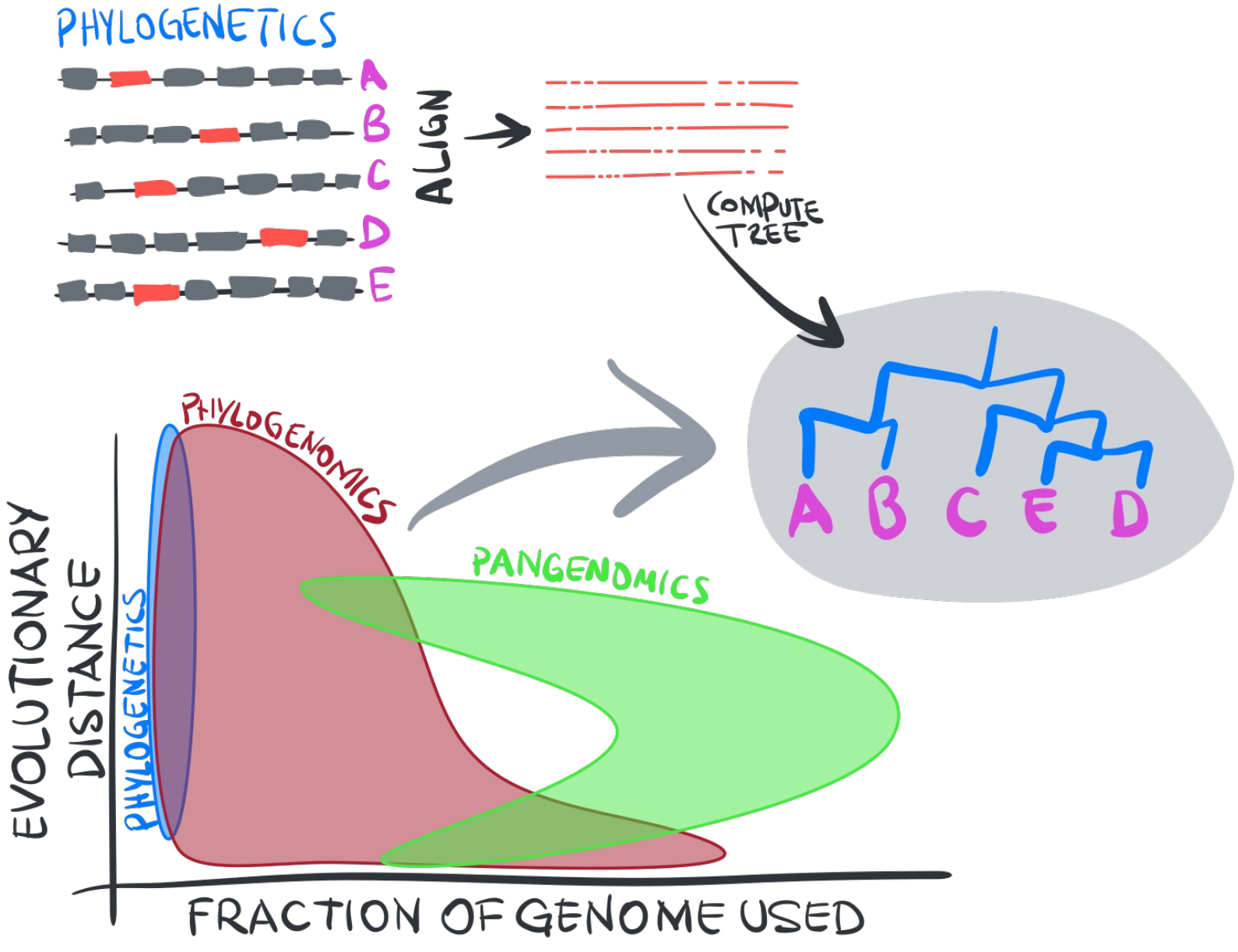


# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)



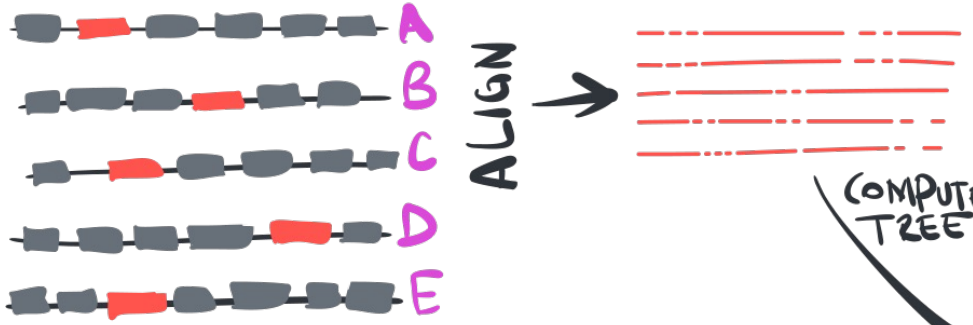



# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)



# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)

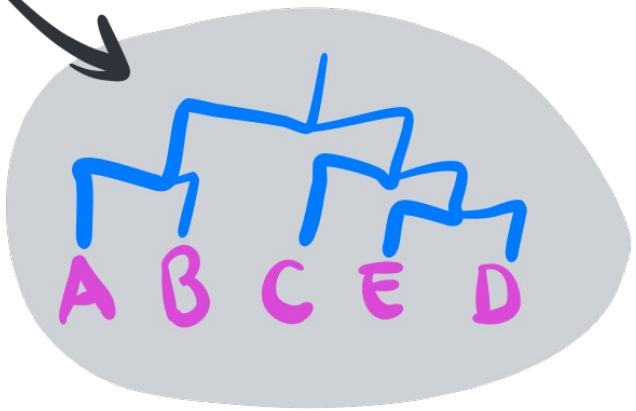
## PHYLOGENETICS



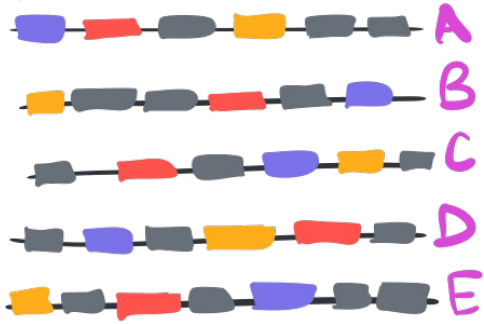

ALIGN

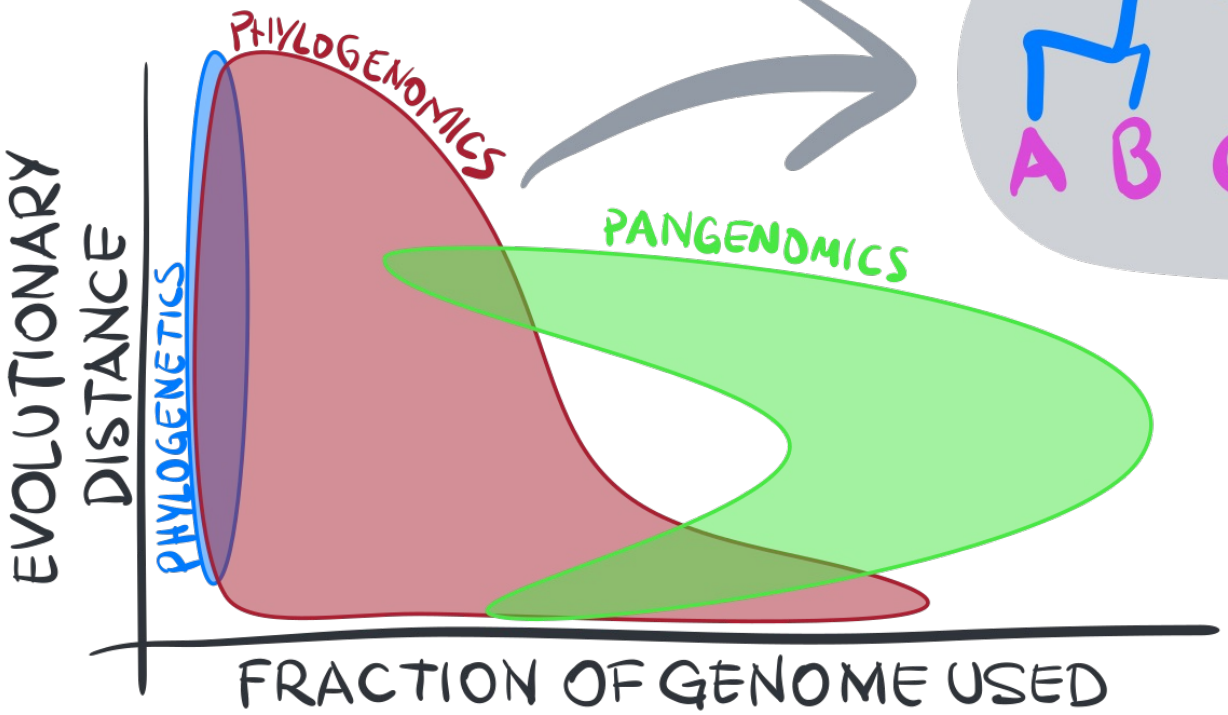


COMPUTE TREE

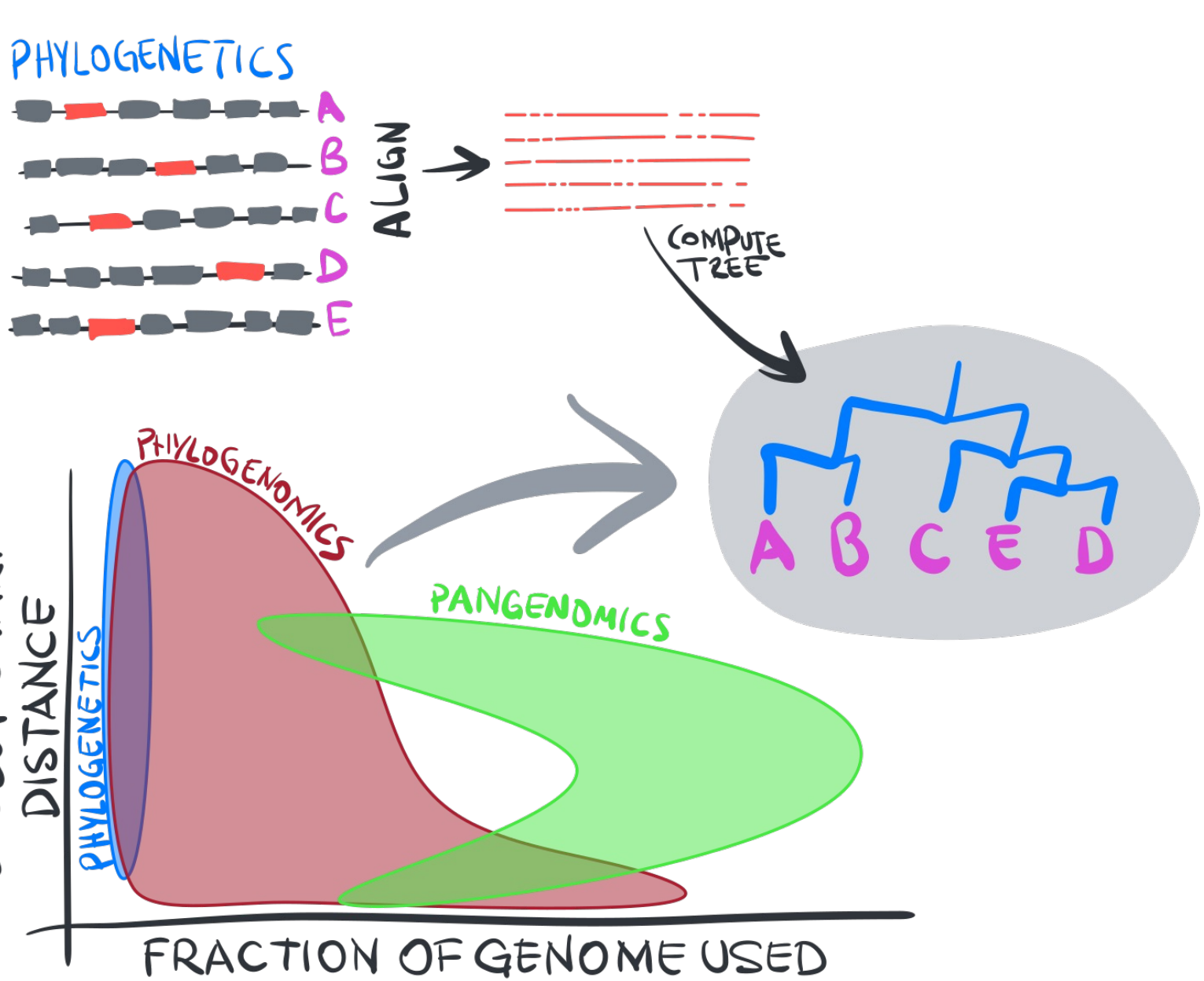


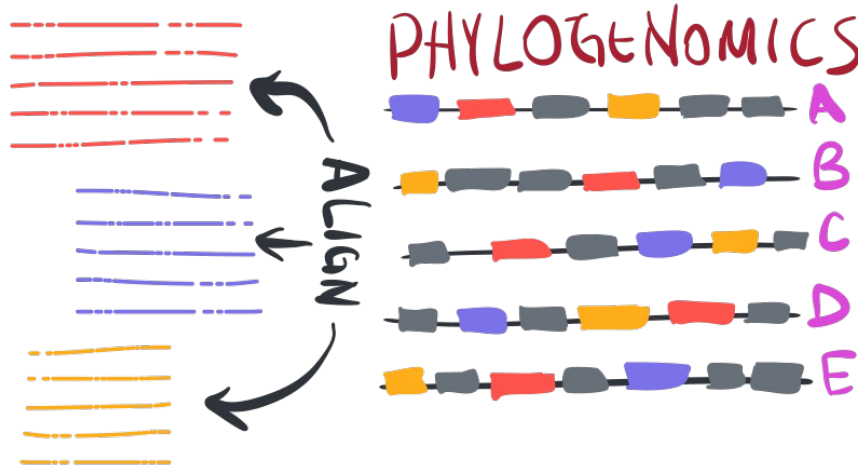

## PHYLOGENOMICS



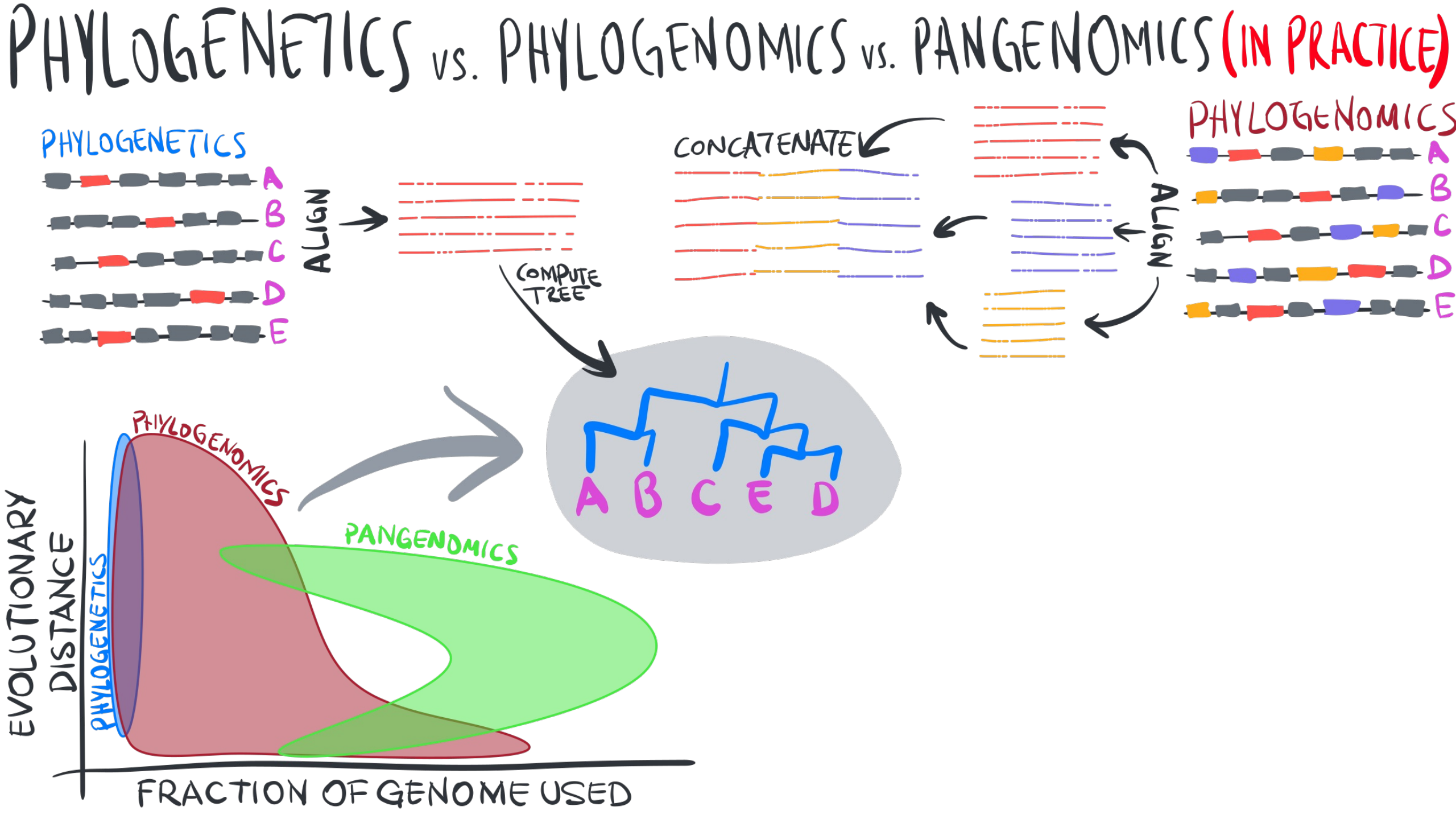



# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)

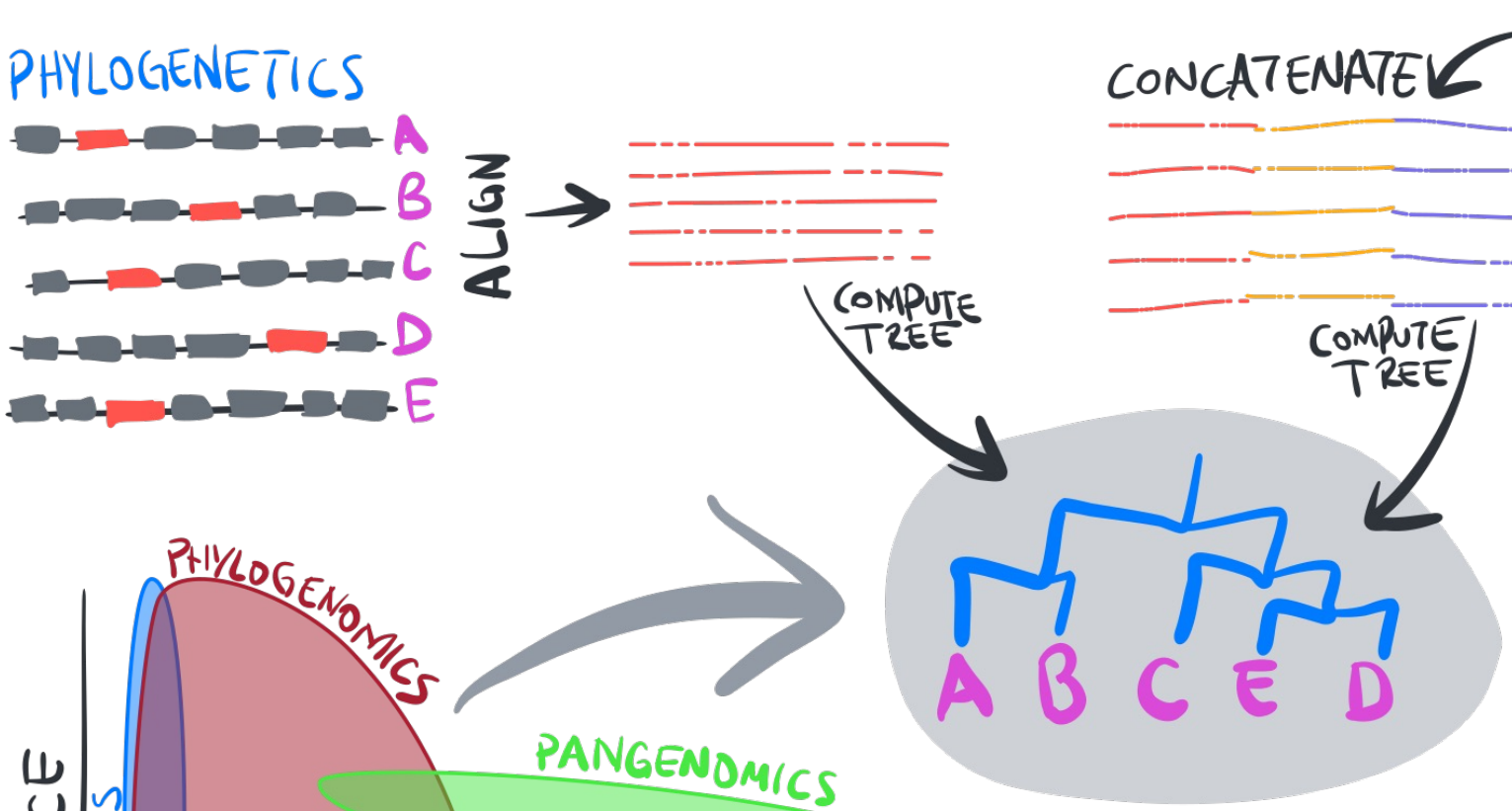


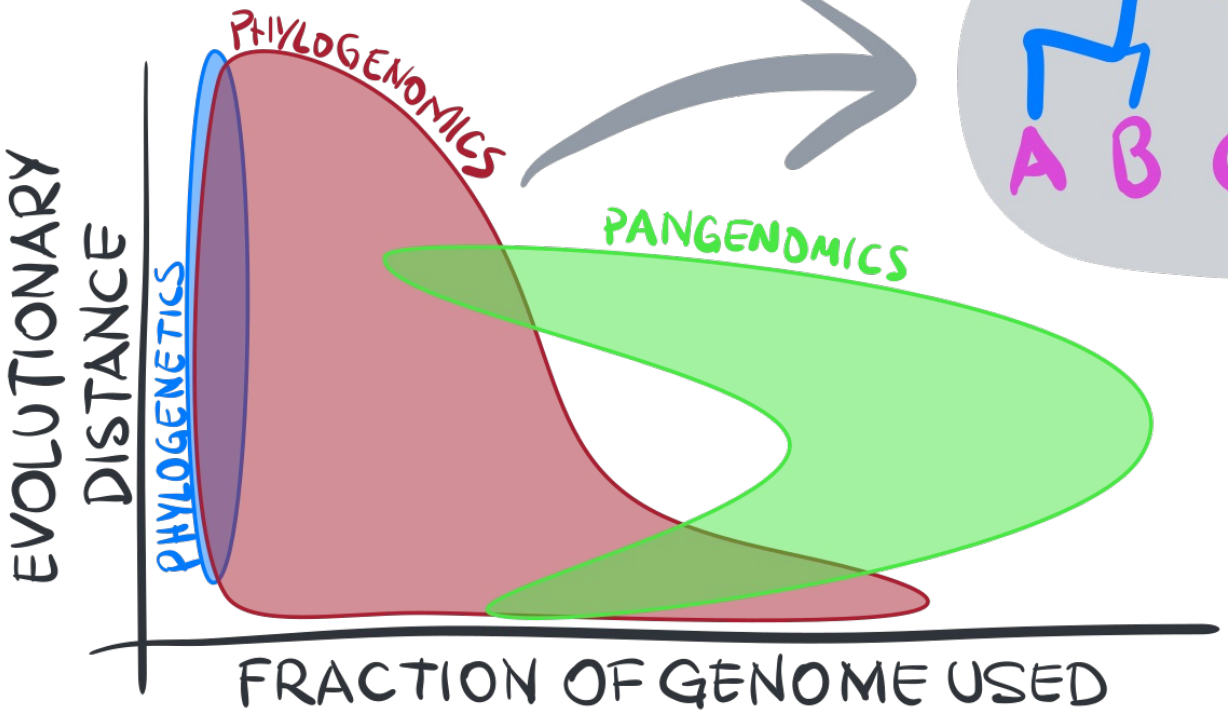



# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)

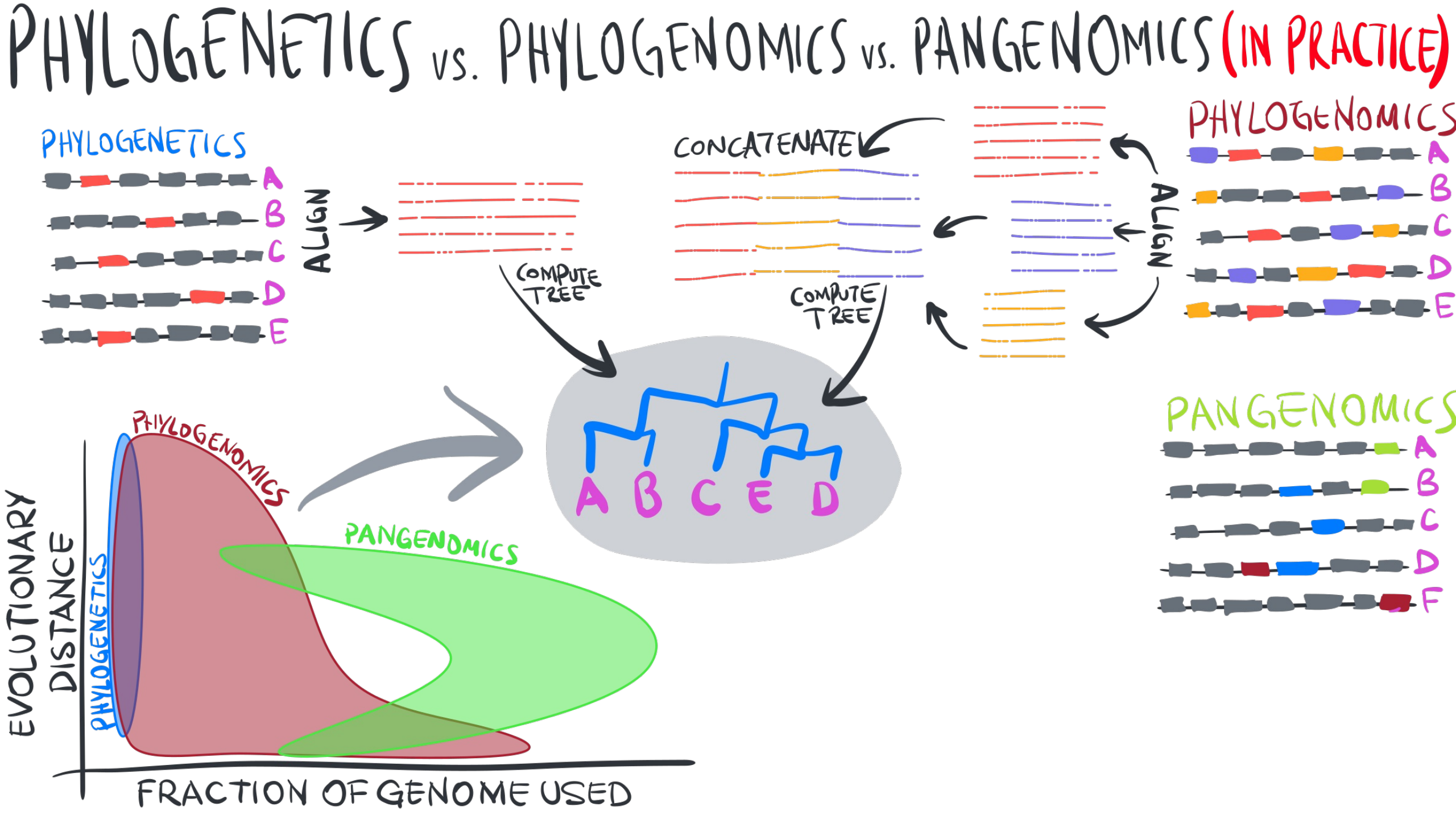


# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)

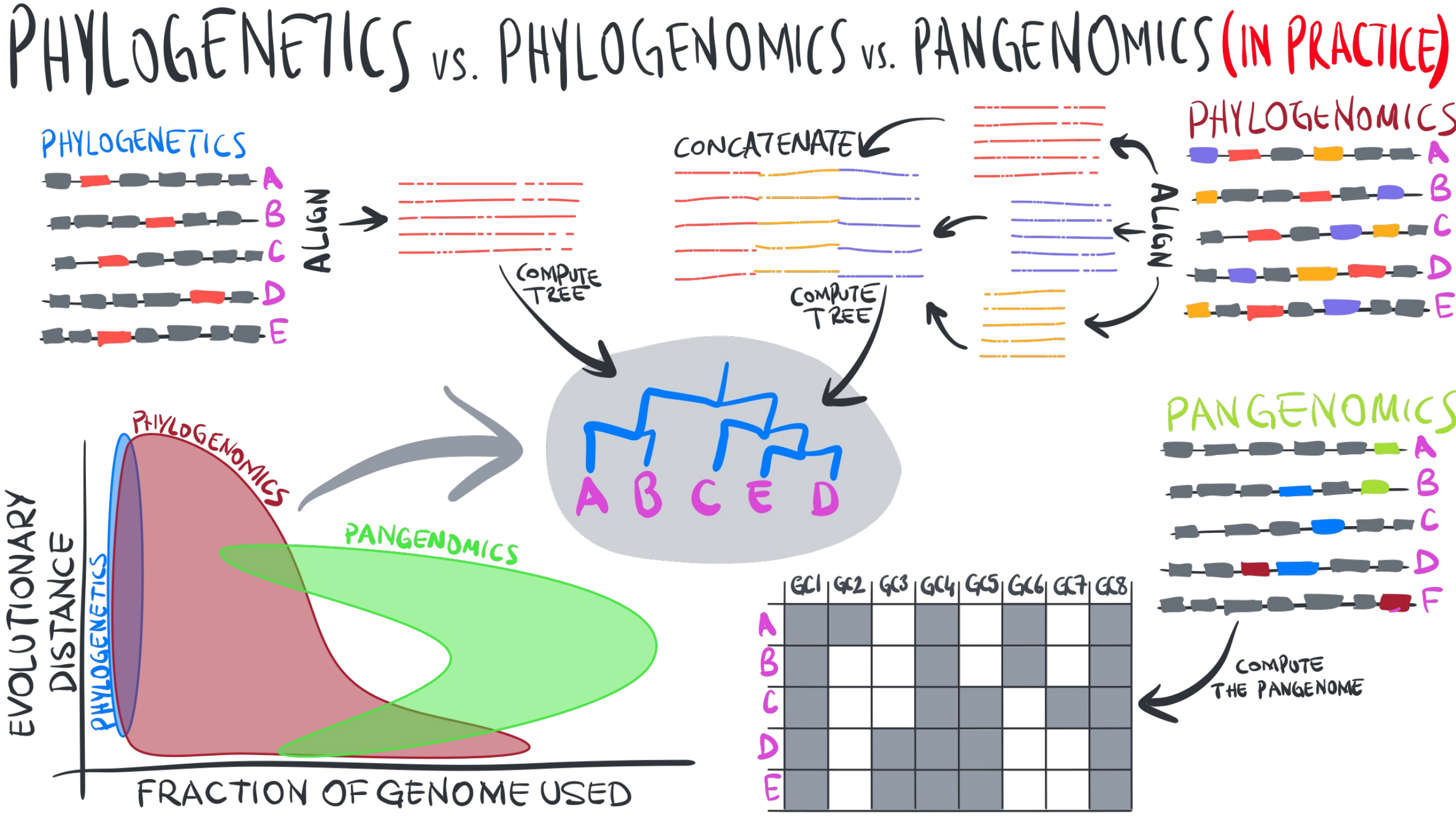


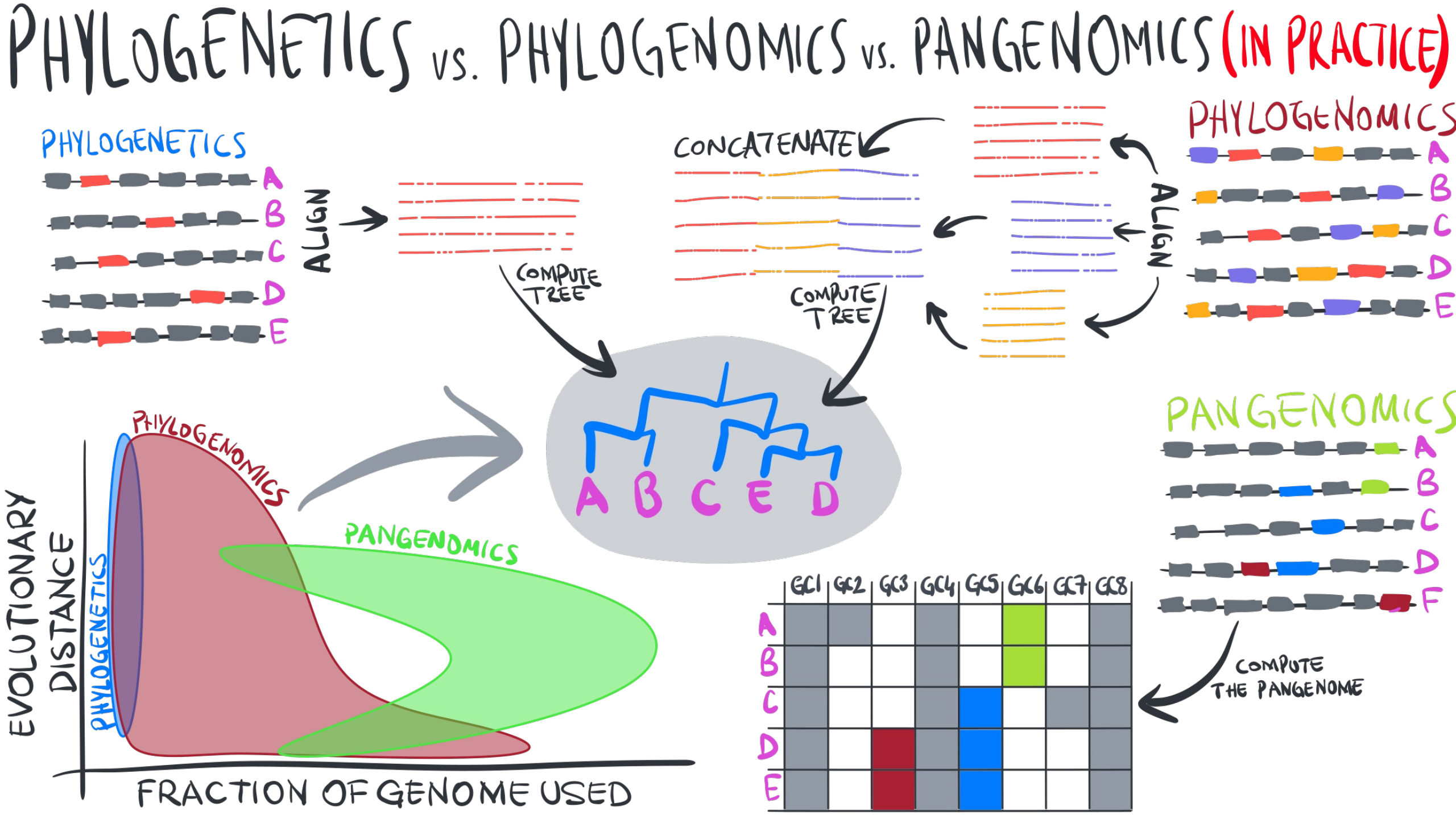
# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)



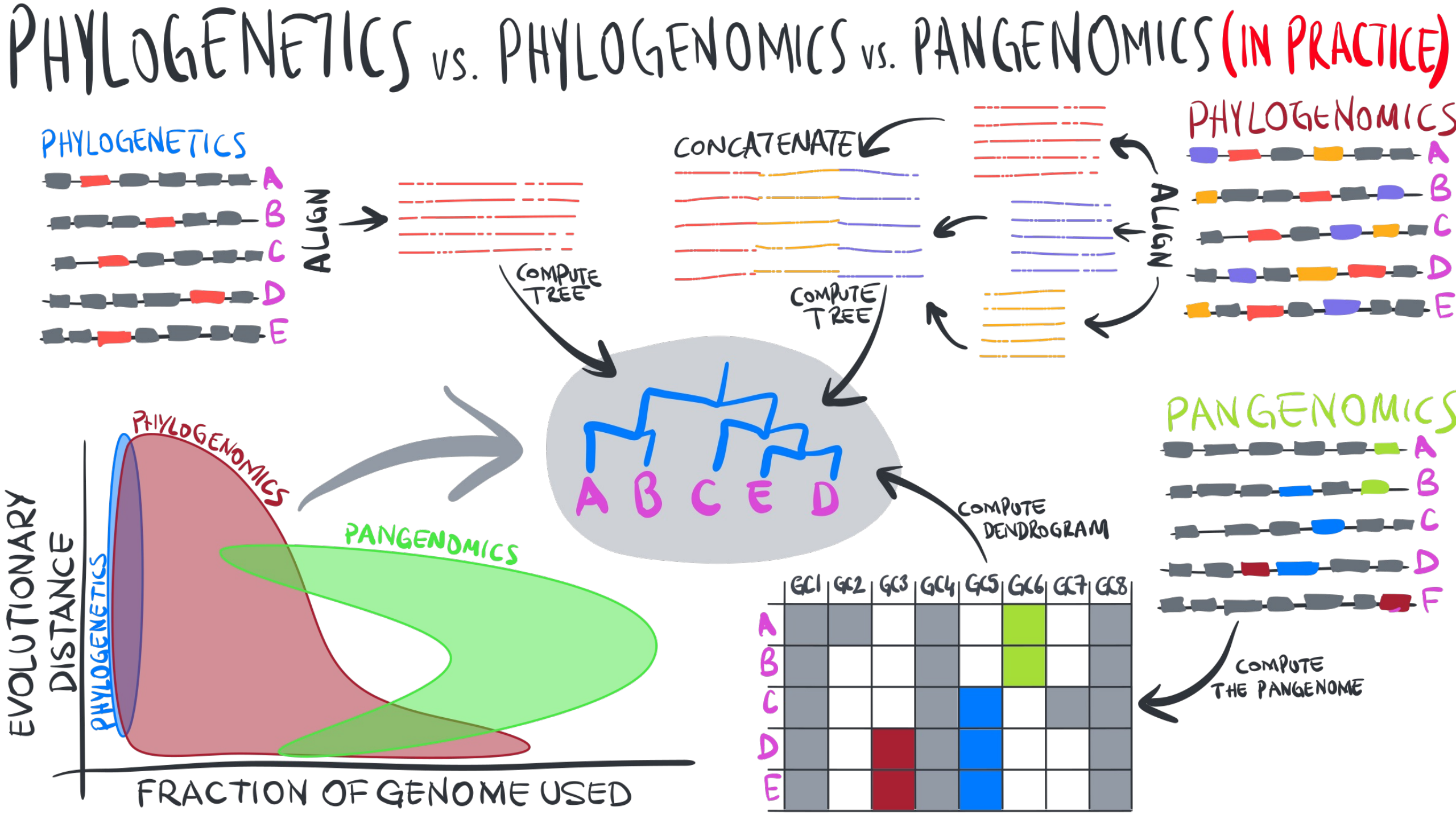
# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)



# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)



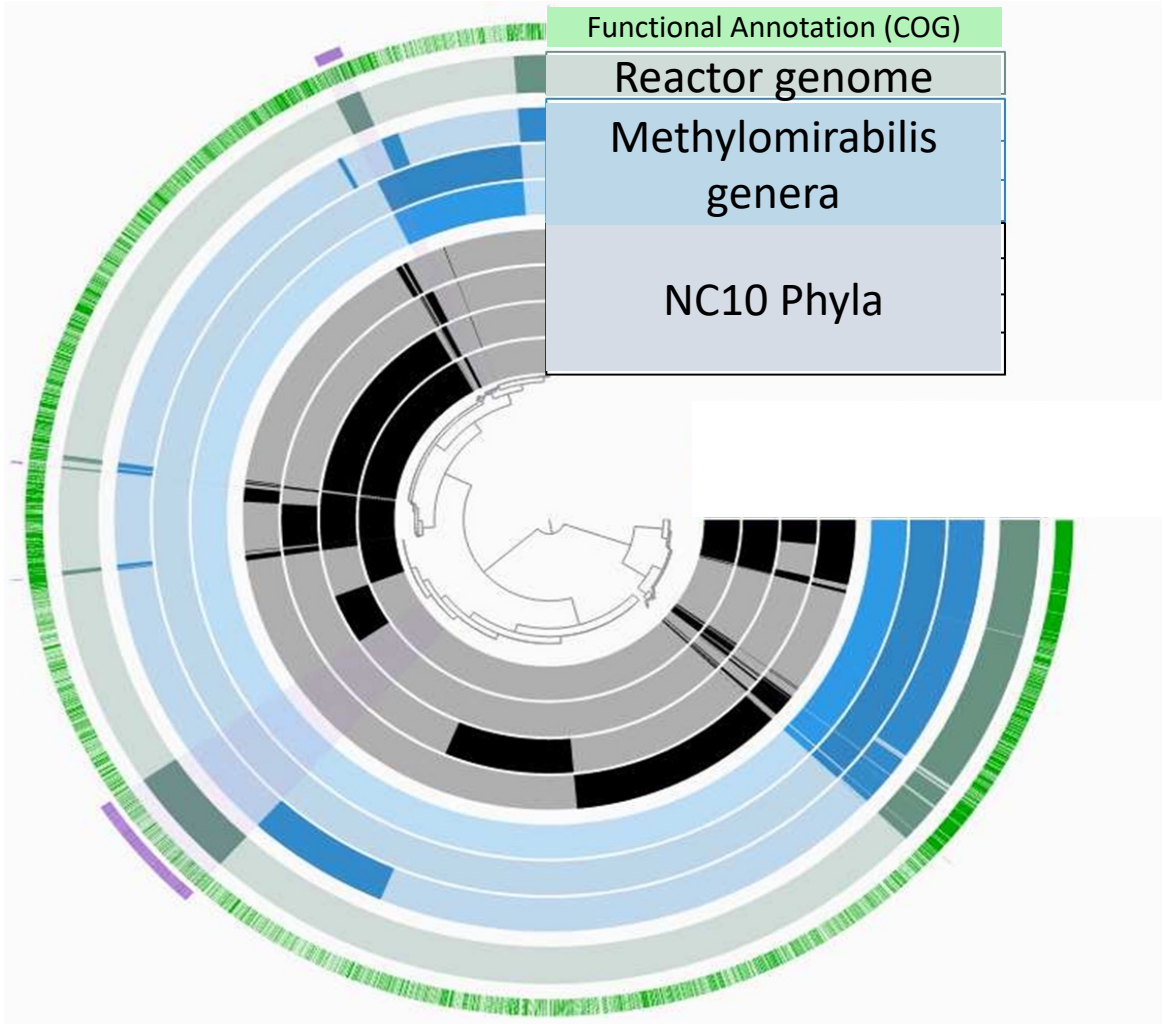
# PHYLOGENETICS vs. PHYLOGENOMICS vs. PANGENOMICS (IN PRACTICE)



# Example pangenome: *Ca.* Methylomirabilis oxyfera recovered from a bioreactor.

**Reactor Genome features**

- 95% complete, 3% contamination
- 16S rRNA gene 97% identical to *Ca.* Methylomirabilis oxyfera
- Average nucleotide identity 80% identical to *Ca.* Methylomirabilis oxyfera



# Hands on activity- Gene centric analysis of antibiotic resistance genes in a wastewater assembly.

- See read me and files on GitHub
- Part 1: Gene centric analysis using the MEGARES resistance database.
- Part 2: Evaluate MAG quality from this wastewater assembly

# Hands on activity- linking gene- and genome-centric information

- Which of the MAGs had AMR capacity?
- What further analysis and quality control do you think is needed from this analysis?
- If you have extra time: find a MAG from an environment of interest to you using JGI. What are the functional capabilities that are of interest to you? How can you analyze it using this tool?