Bringing Order to Chaos: Metagenomics Starting from Raw Reads

DALL-E

Johannes Björk (UMCG/RUG)

Chrats Melkonian (WUR & UU; Bioinformatics)

Computational Metagenomics - BioSB research school (14/10/2024)

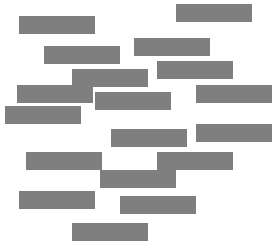
Metagenomics: Overview

DNA extraction

Sequencing

Experimental part

Reads

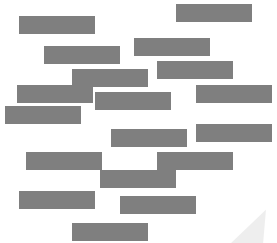


Metagenomics: Overview (today's focus)

DNA extraction

Sequencing

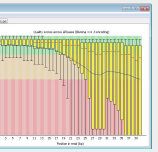
Reads



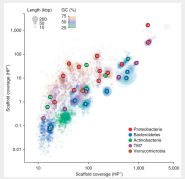
Experimental part

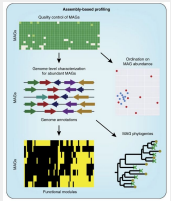
Computational part

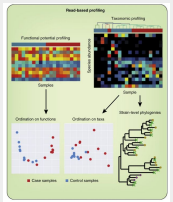
3) Assembly



Day 2







4) Binning

2) Read-based

1) Data preprocessing

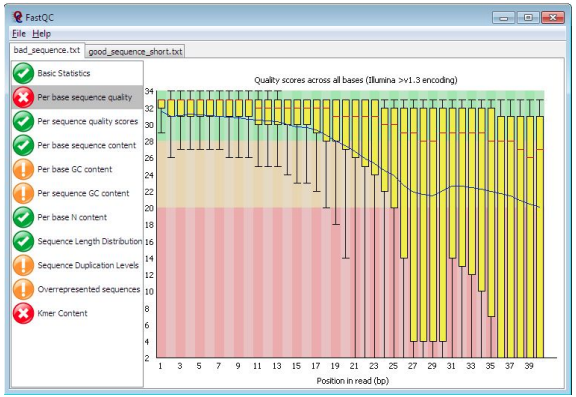
Metagenomics workflow

DNA extraction

Sequencing

Raw reads

- **Exploration**
  e.g., FASTQC
- **Processing**



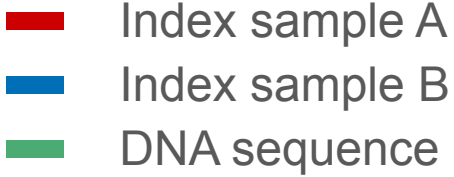
Data preprocessing - Adapters

Adapters are nucleotide sequences placed at either one or both ends of the DNA fragments that are being sequenced

They are composed of 3 sections:

- Sequencer binding site (illumina)
- Multiplexing index (P5-P7)
- Sequencing primer binding site (illumina)

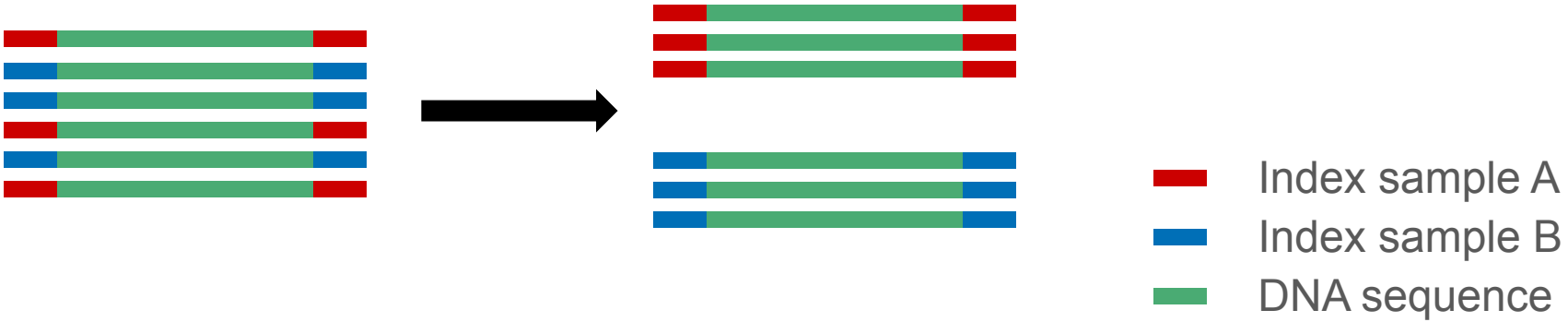
They are necessary for sequencing but should be removed early on in data pre-processing steps



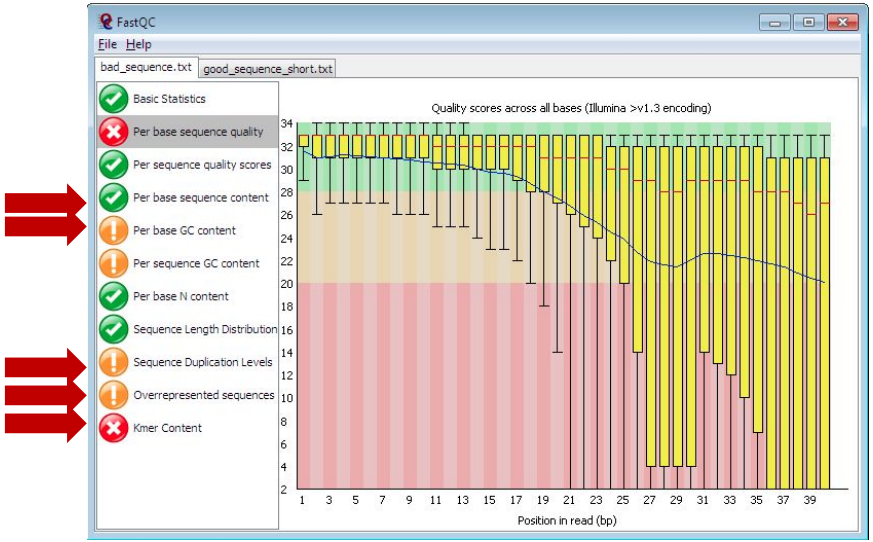
Data preprocessing - Demultiplexing

Demultiplexing tools: Sabre, iDemux etc..

Generally performed by sequencing companies before sending the data. Good to know what it is to be able to spot it in QC.



Data preprocessing - Adapter trimming



From QC data you may notice that adapters are still present in your sequence. You should remove them either by providing the adapter sequence or using a de-novo search.

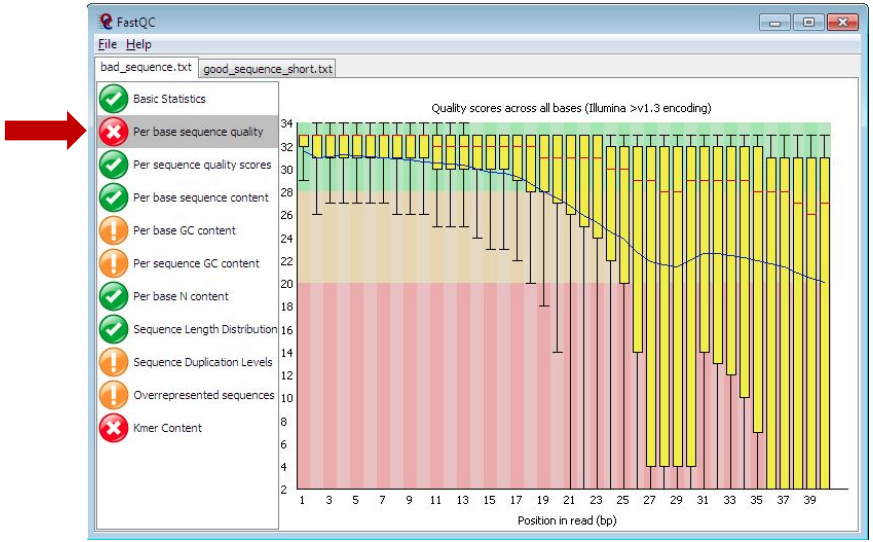
Recommended tools:
Trimmomatic, Cutadapt, bbduk, fastp

After adapter removal, rerun QC on the fastq files

Keep an eye out for polyA and polyG sequences

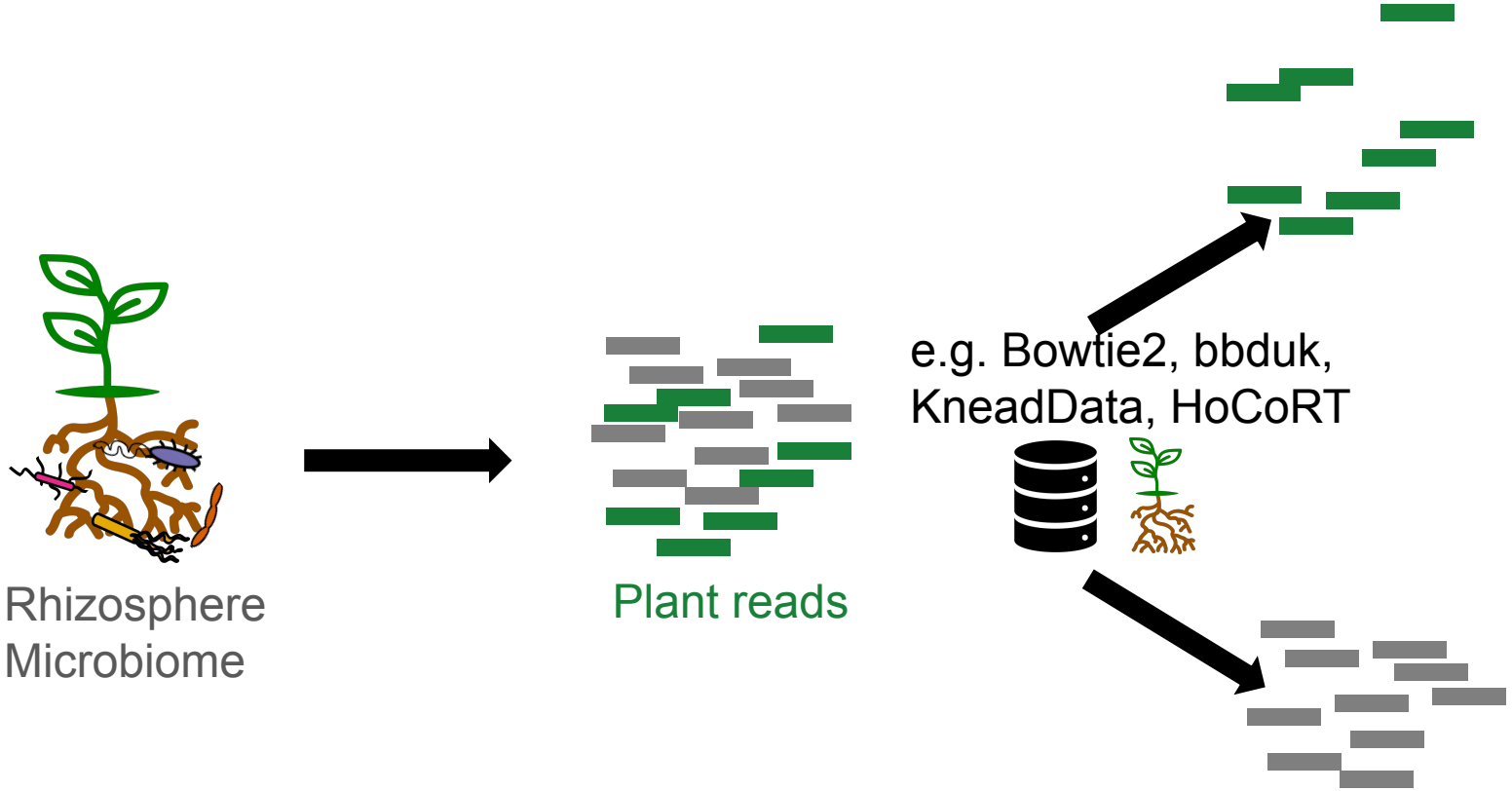
Data preprocessing - Quality filter

Phred quality score – Logarithmic score representing the quality of a nucleotide



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%

Data preprocessing - Host (& other) removal



Metagenomics: Read-based

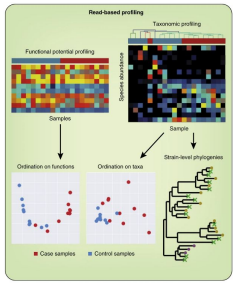
DNA extraction

Sequencing

Reads

100-150 bp

The read based: Mapping into databases

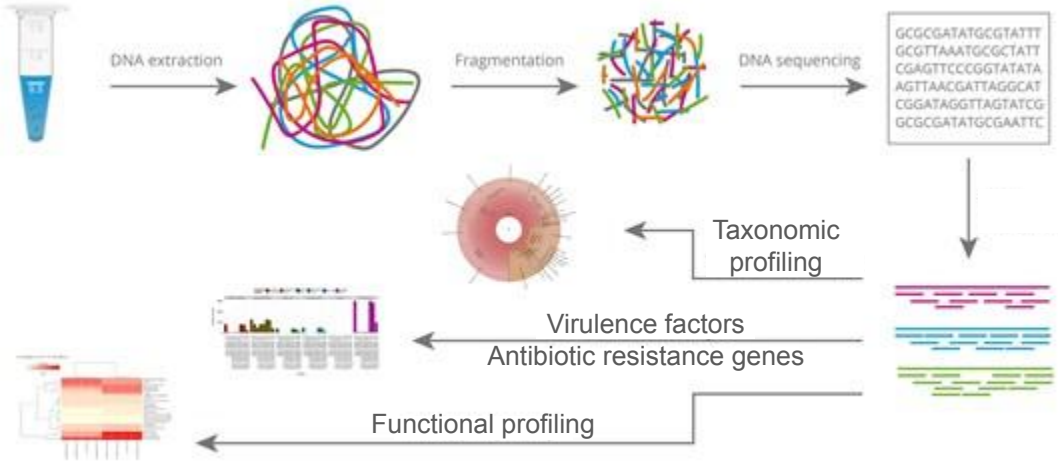


Read-based

NCBI

UniProt

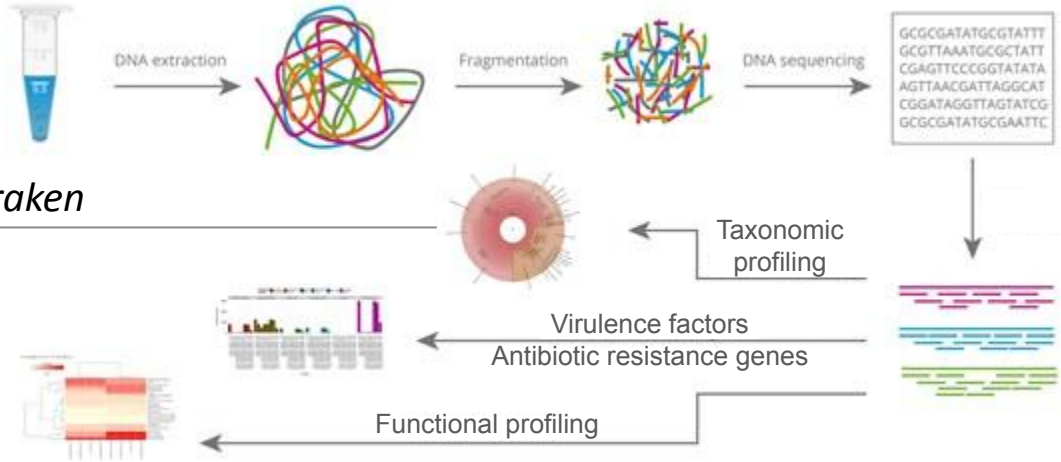
Read-based approaches



Align short-reads to different databases containing reference sequences

Read-based ≃ Reference-based

Read-based approaches

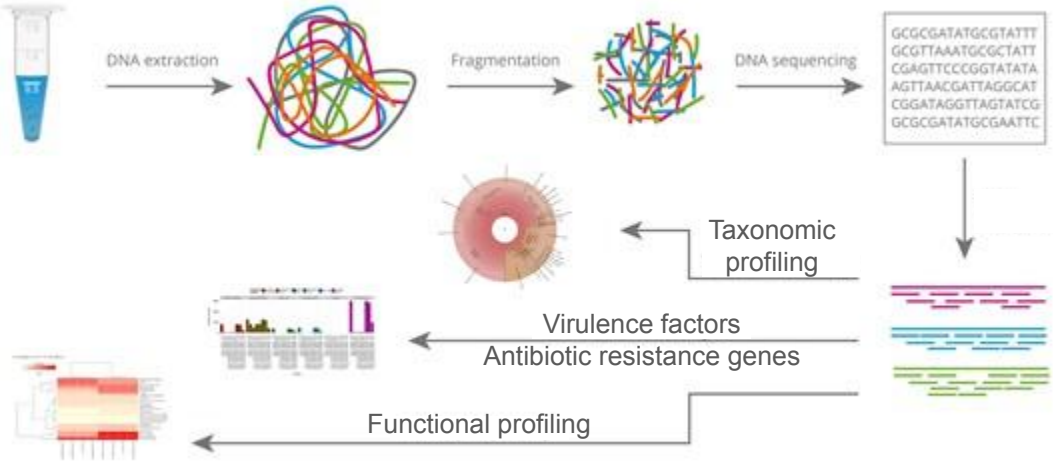


E.g. Metaphlan, Kraken

Align short-reads to different databases containing reference sequences

Read-based ≃ Reference-based

Read-based approaches

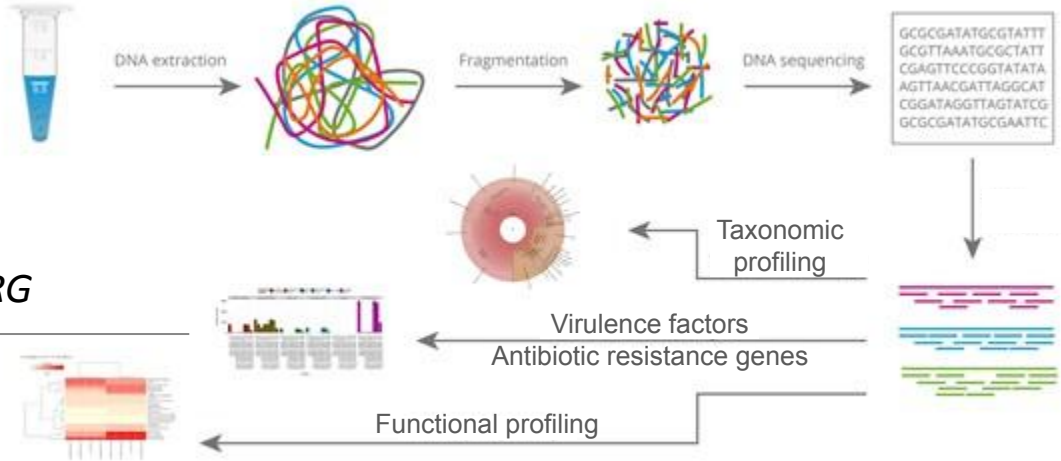


E.g. HUMAnN, MetaPathways

Align short-reads to different databases containing reference sequences

Read-based ≃ Reference-based

Read-based approaches



E.g. ShortBRED, ResFinder, DeepARG

Align short-reads to different databases containing reference sequences

Read-based ≃ Reference-based

Metagenomics workflow: Assembly-based

DNA extraction

Sequencing

Reads

100-150 bp

Assembly

Contigs

1000+ bp

Assembly-based

The Assembly Problem:

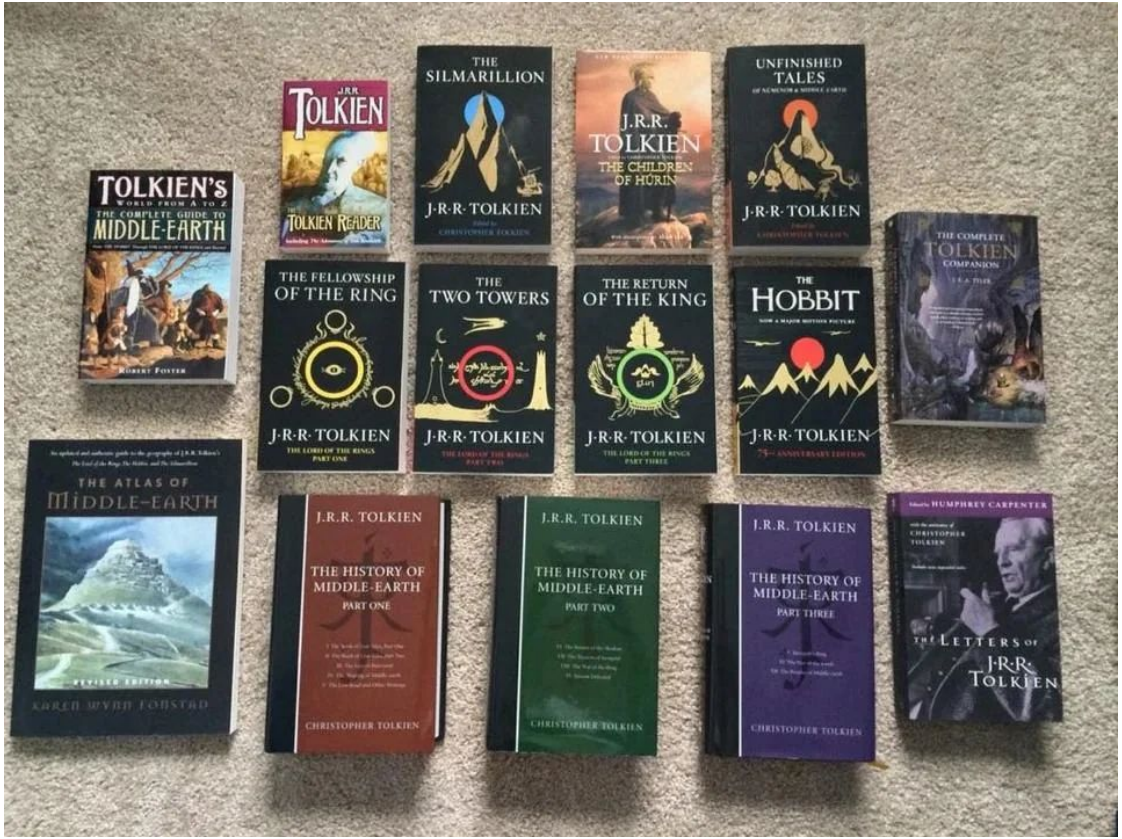
Library of books

Sequencing

Shred all books

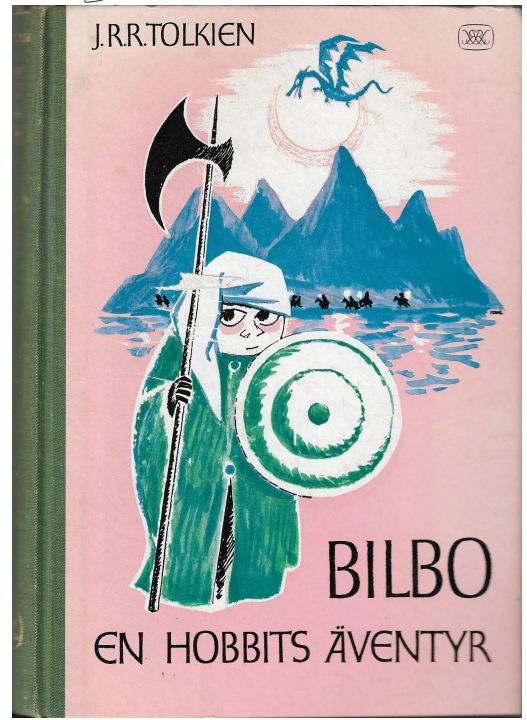
Assembly

Reconstruct each book



Repetitive content makes the reconstruction more difficult

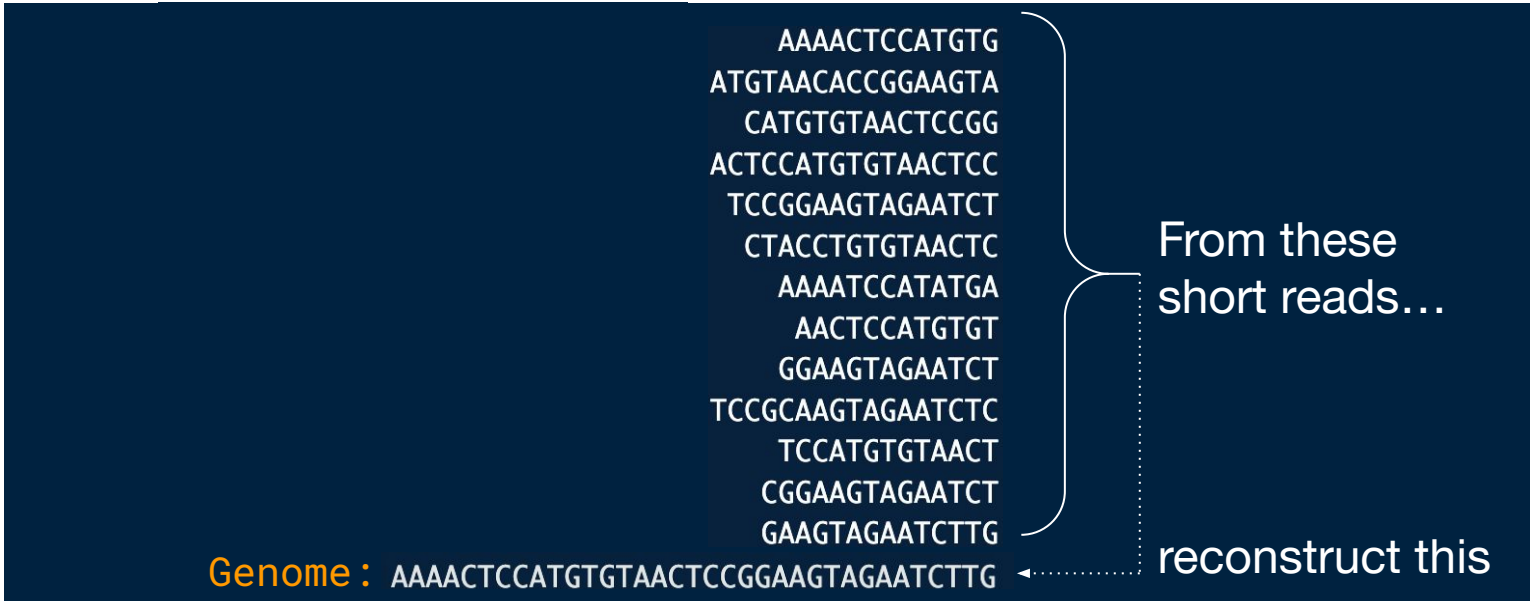
Misprints or damaged fragments make reconstruction more difficult



Rare books are difficult or impossible to reconstruct

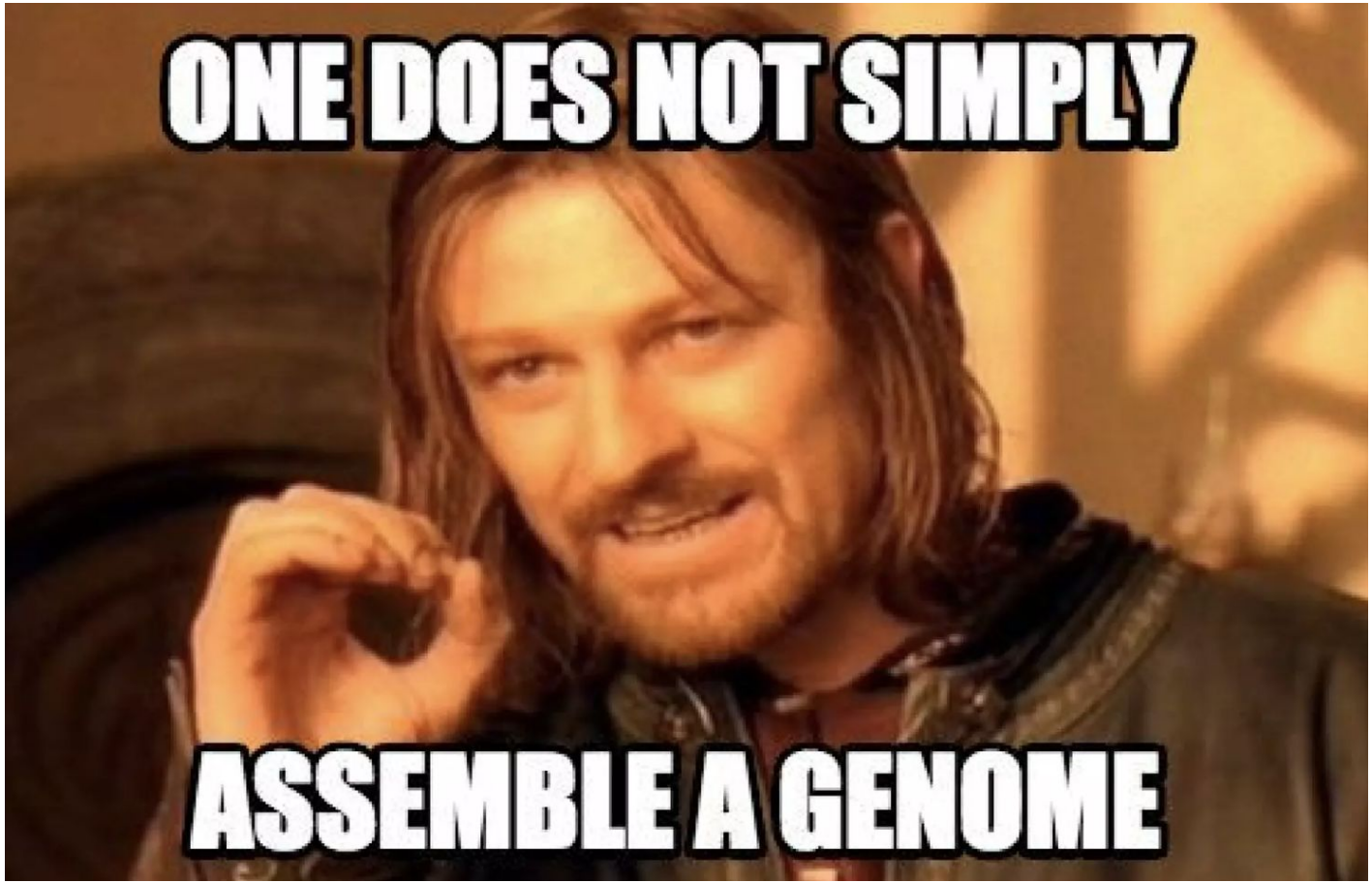
Back to DNA sequences…

Problem: We don’t know where the reads came from in respect to the genome sequence



Reality: Reads are scrambled …

Reality: Reads are scrambled AND we don't know the genome sequence

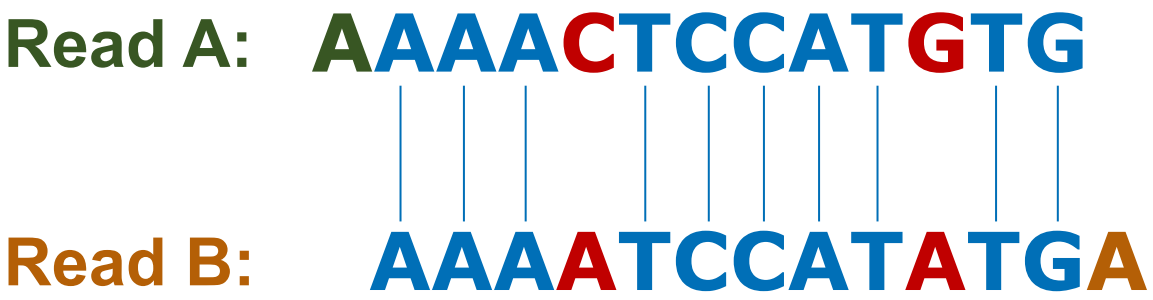

ONE DOES NOT SIMPLY
ASSEMBLE A GENOME

How do we stitch together reads into contigs?

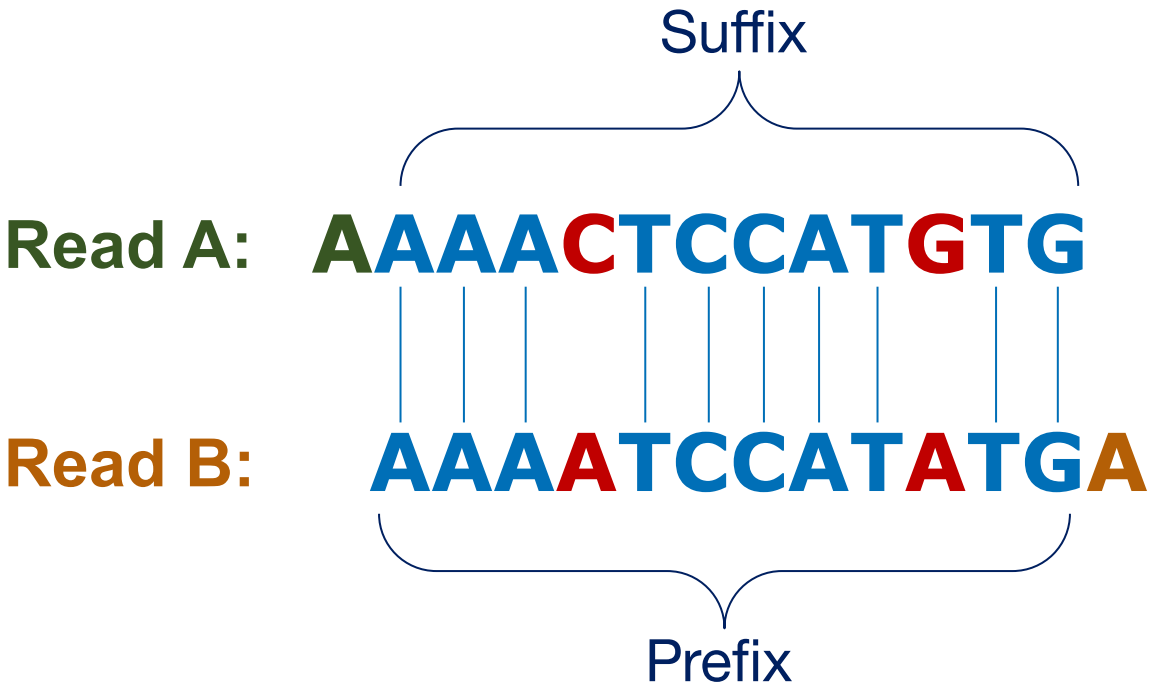
Read A: **AAAACTCCATGTG**

Read B: **AAAATCCATATGA**

How do we stitch together reads into contigs?



Suffix - Prefix Overlap



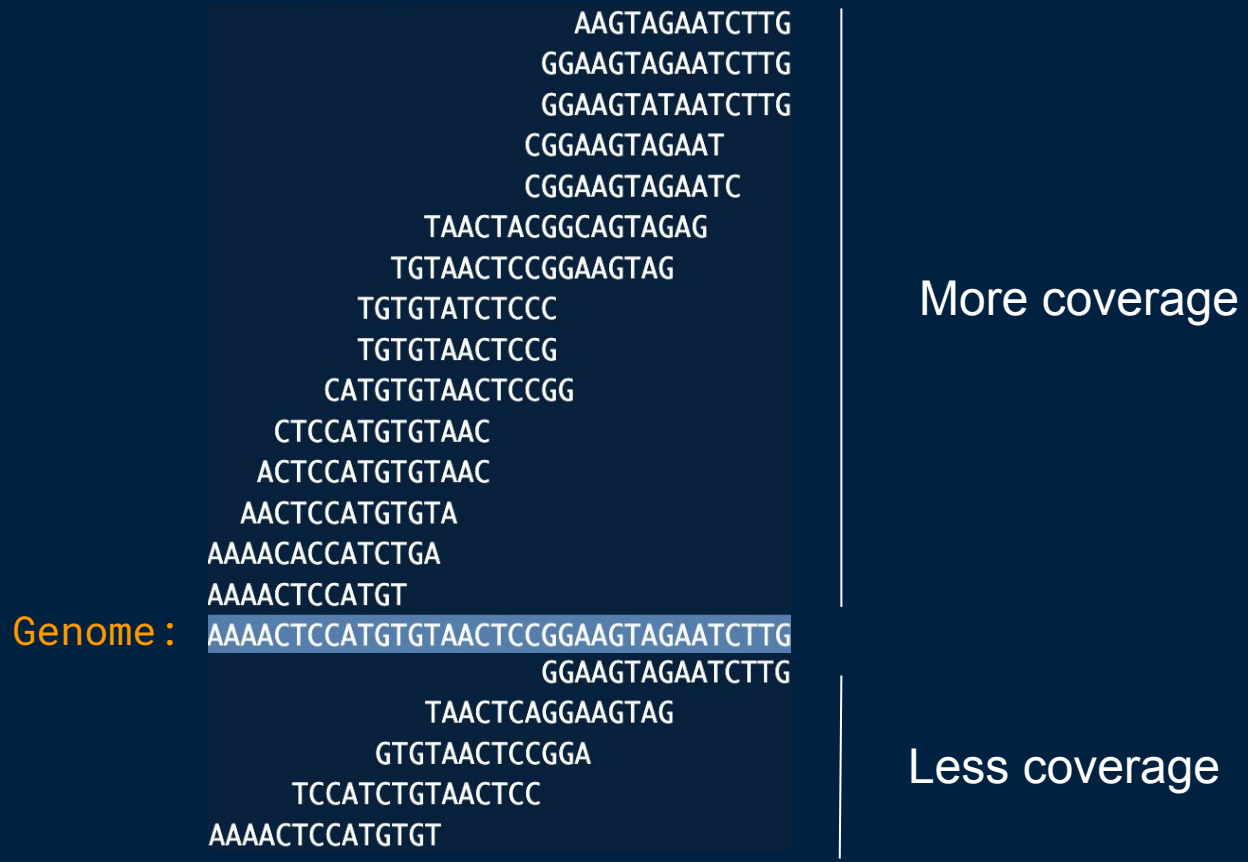
First Law of Assembly

If a suffix of read A is similar to a prefix of read B… then A and B might overlap in the genome

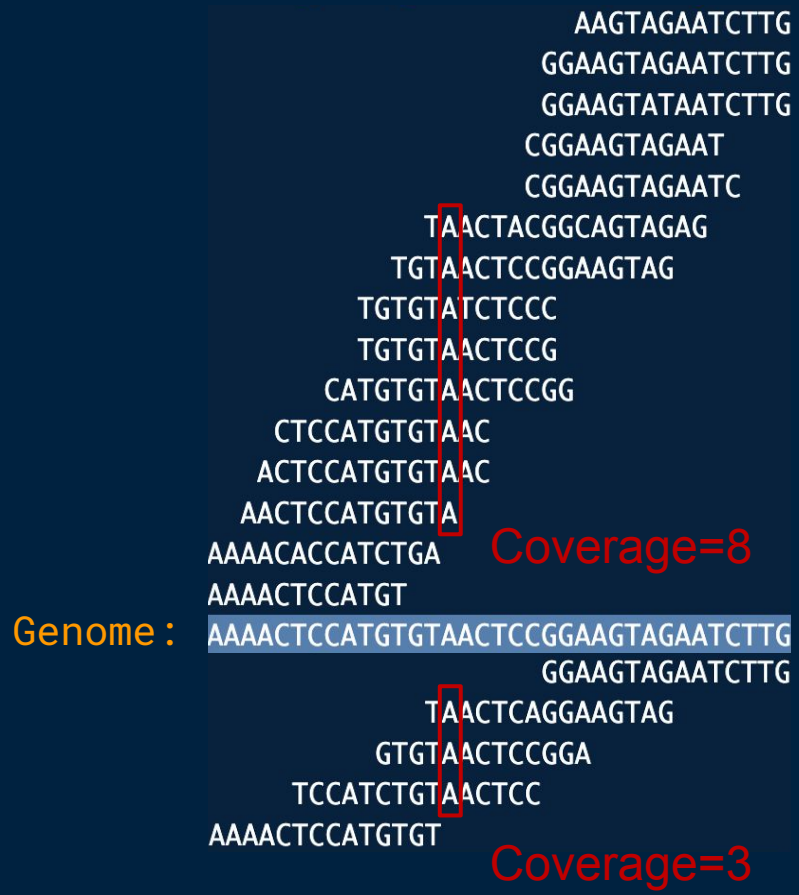
Second Law of Assembly

More coverage leads to more and longer overlaps

More coverage leads to more and longer overlaps



More coverage leads to more and longer overlaps



Average coverage:
207/35 ≈ **6-fold**

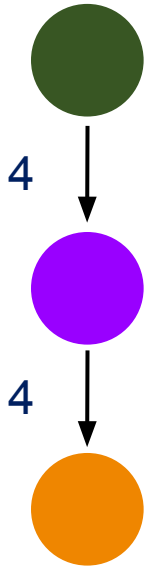
Average coverage:
70/35 = **2-fold**

Overlap graph

Read A: ACGTA

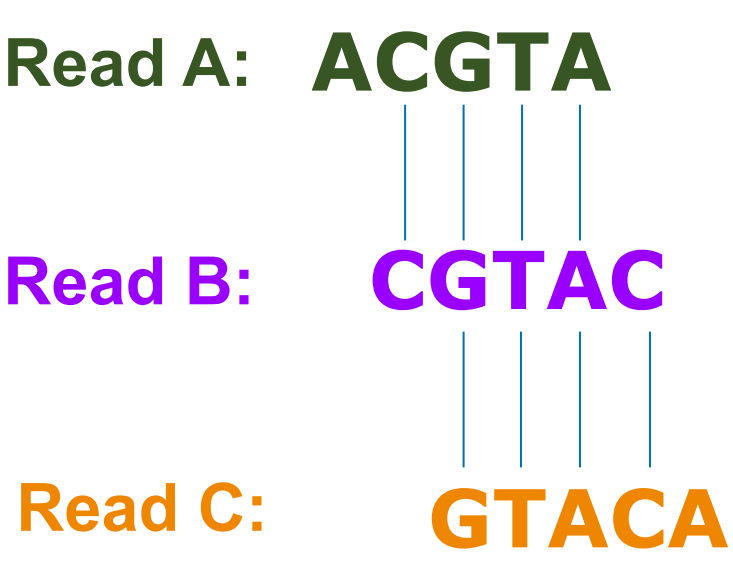
Read B: CGTAC

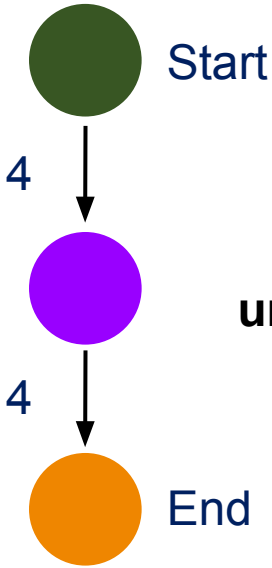
Read C: GTACA



Contig: ACGTACA

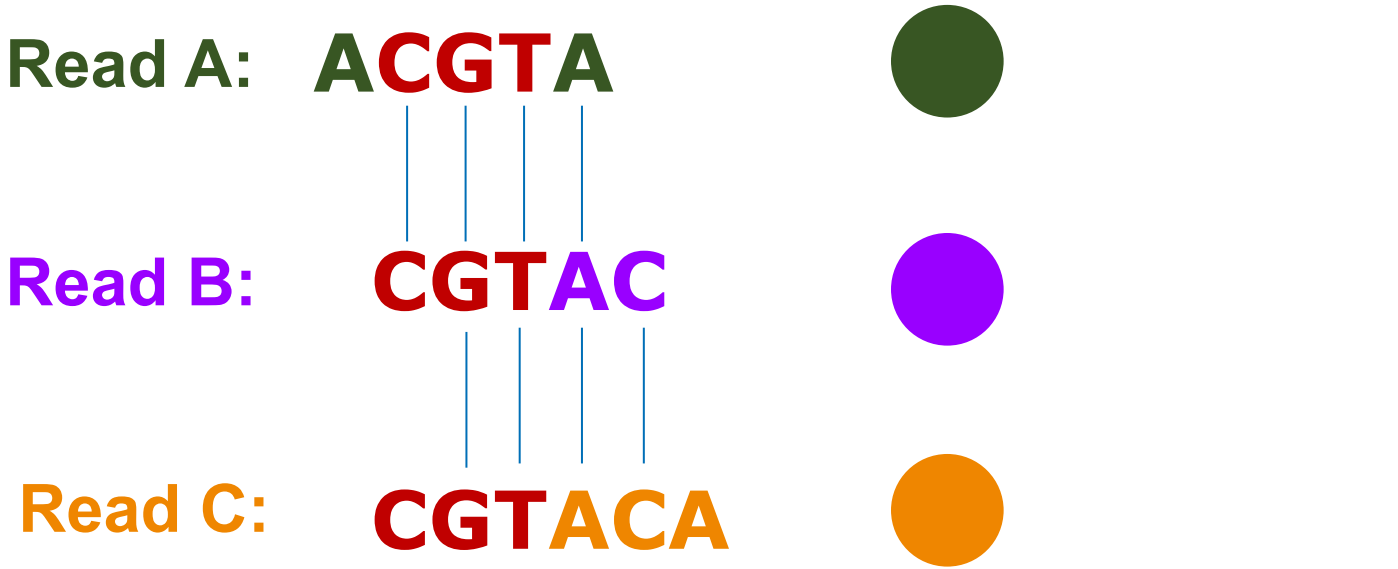
Overlap graph





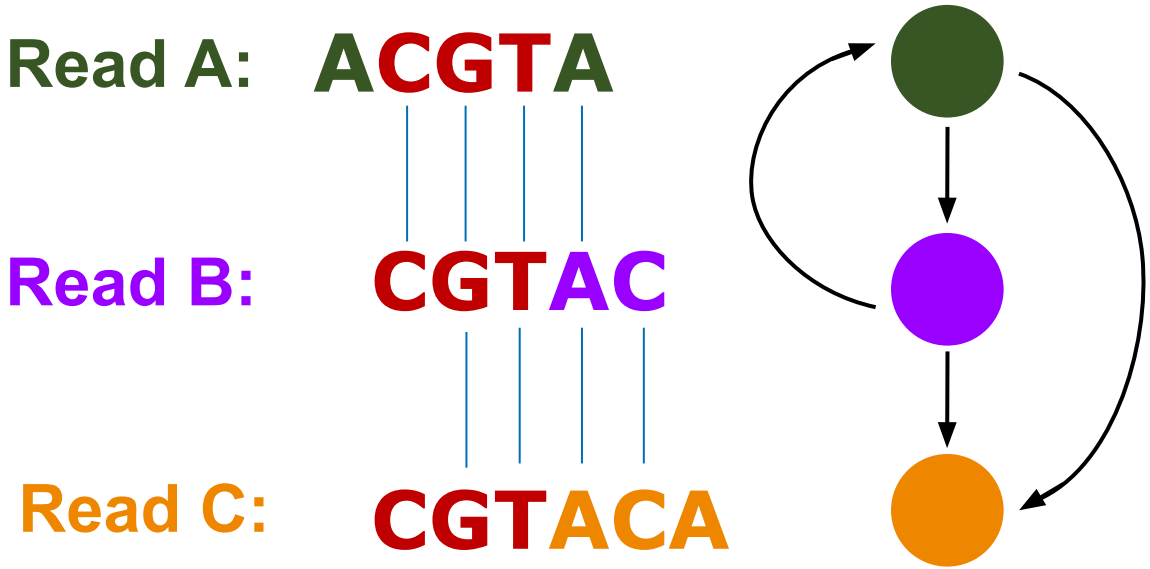
A linear unambiguous path

Overlap graph



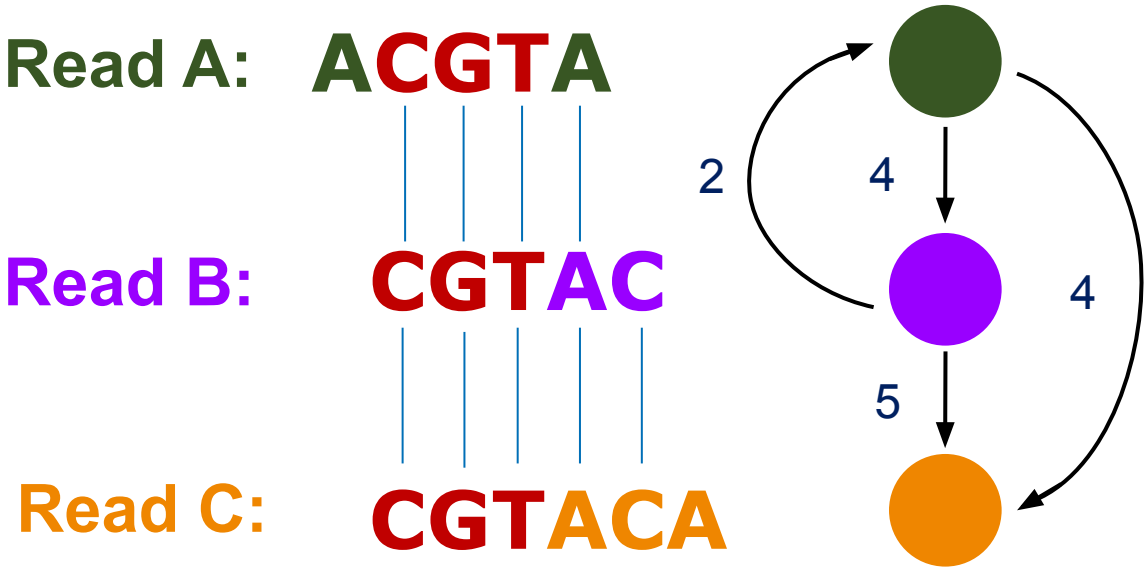
Contig: ACGTACGTA (With repeats)

Overlap graph



The introduced repeat "CGT" appears in multiple reads, causing branching paths in the overlap graph

Overlap graph



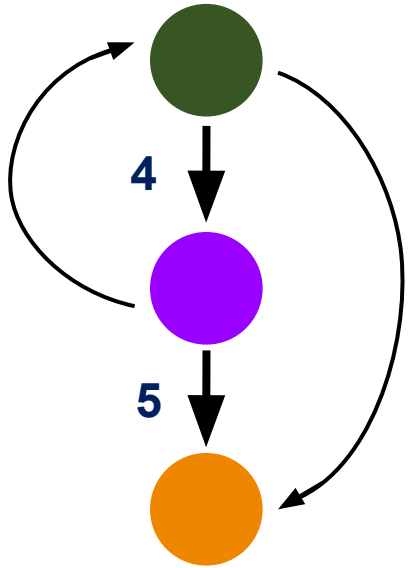
The introduced repeat "CGT" appears in multiple reads, causing branching paths in the overlap graph

Overlap graph

Read A: ACGTA

Read B: CGTAC

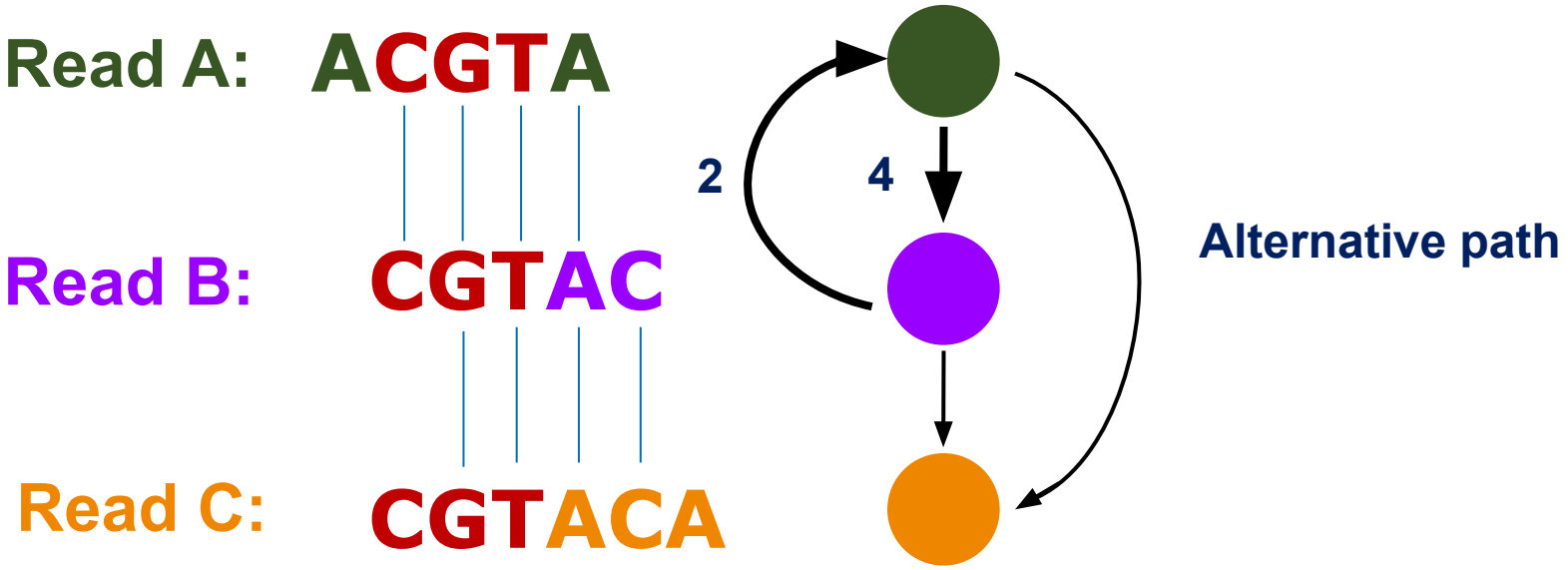
Read C: CGTACA



Path along the edges with highest overlap

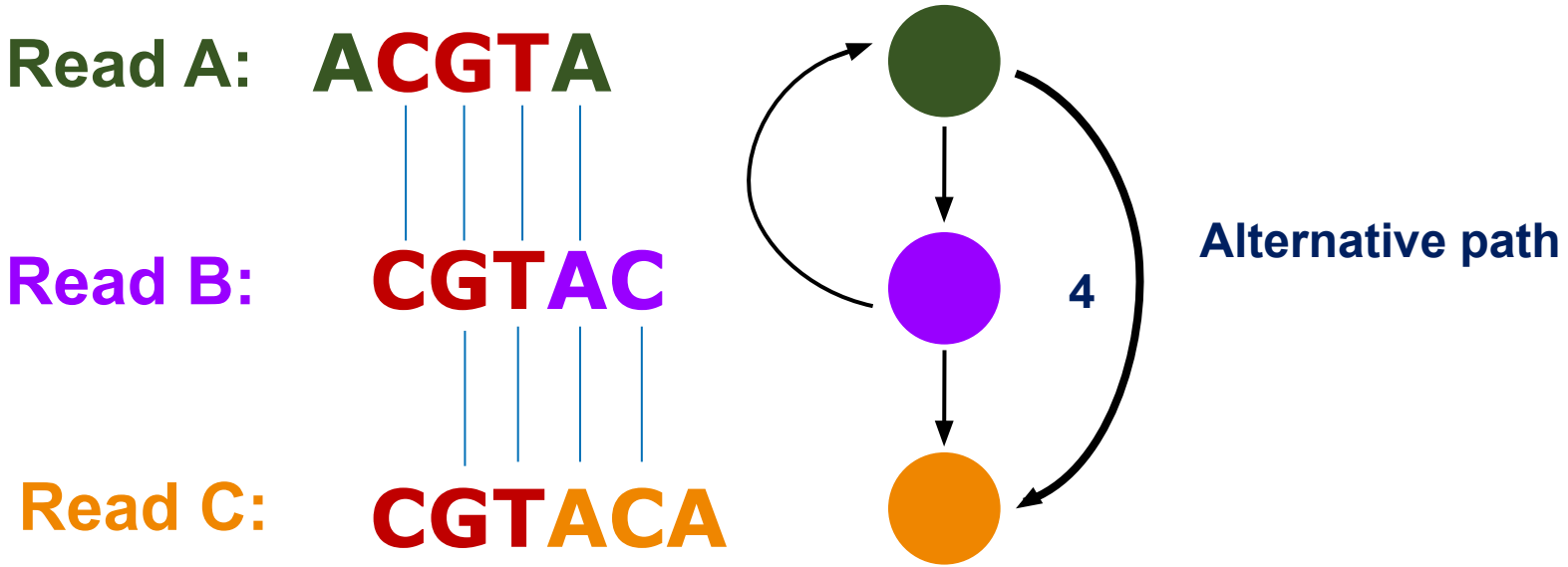
Not the original contig ACGTACA

Overlap graph



Original contig ACGTACGTA

Overlap graph



Not the original contig ACGTACA

Third Law of Assembly

Repeats make assembly difficult

Overlap graph

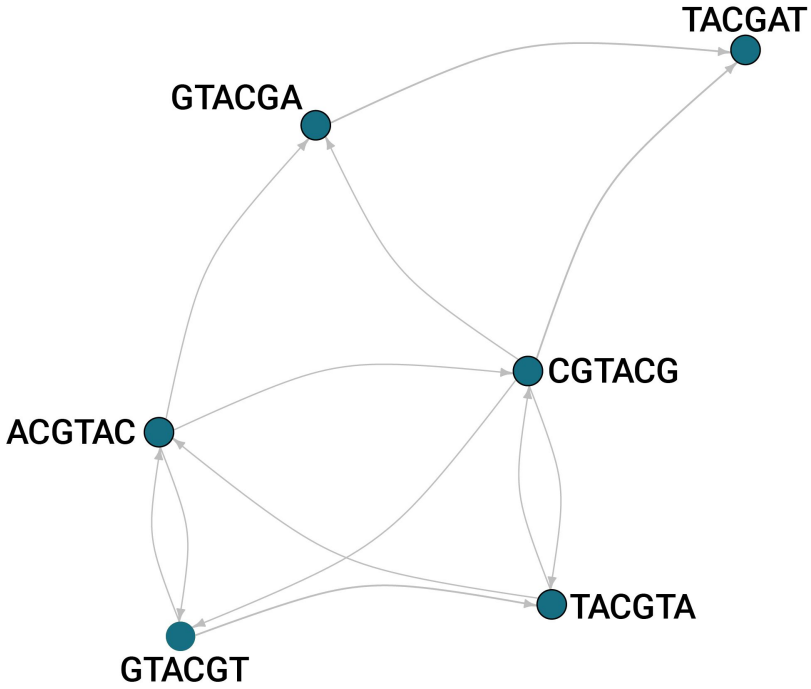
Original contig **GTACGTACGAT**

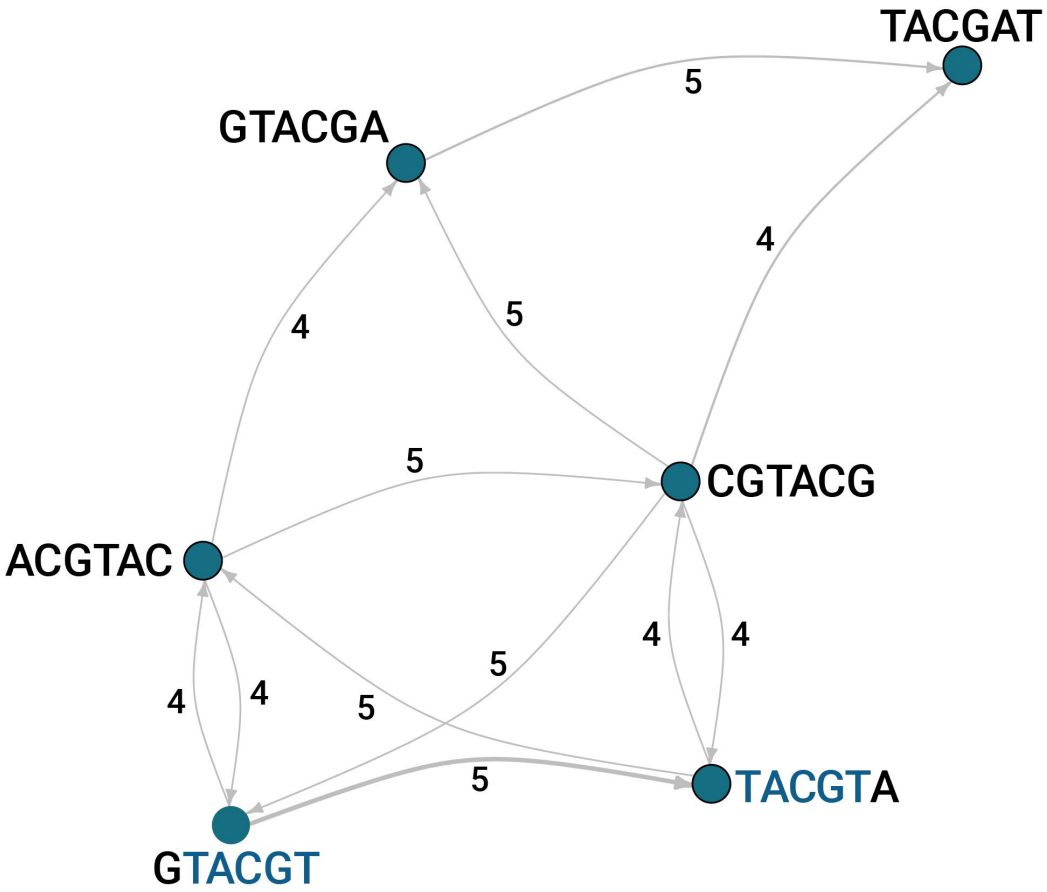
GTACGT TACGTA ACGTAC CGTACG GTACGA TACGAT

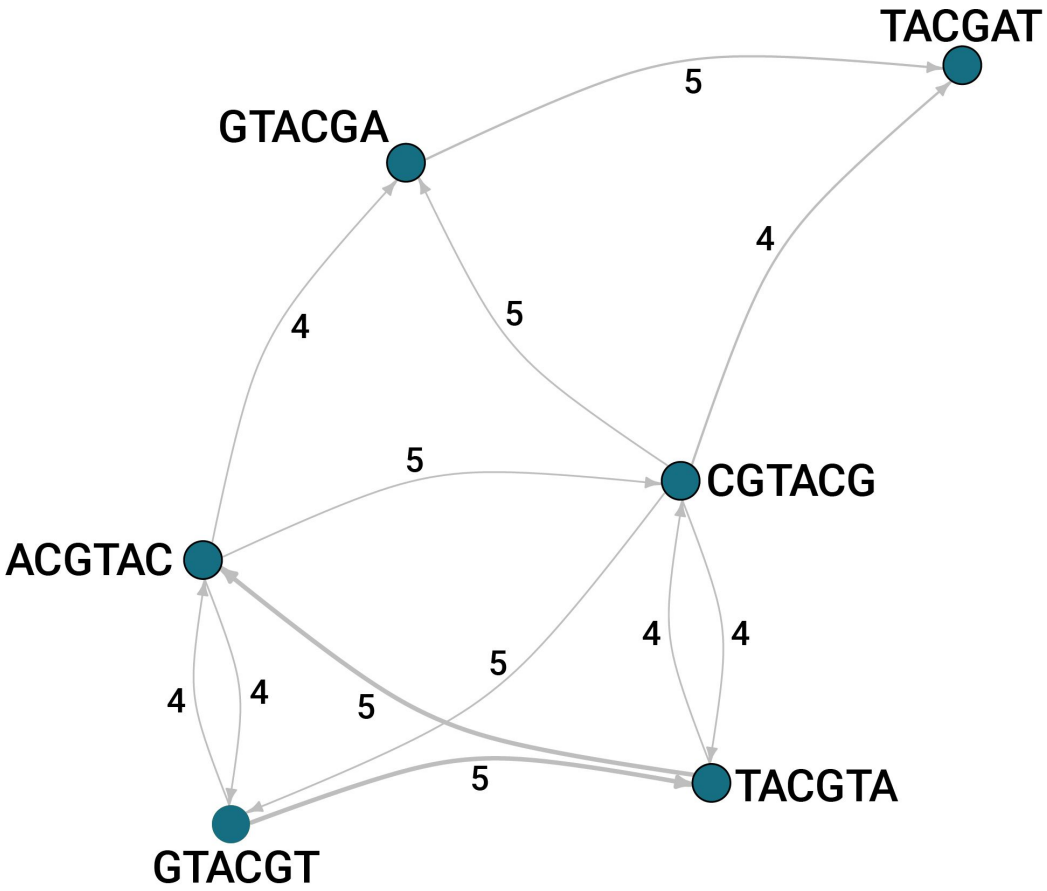
Short reads (6-mers)

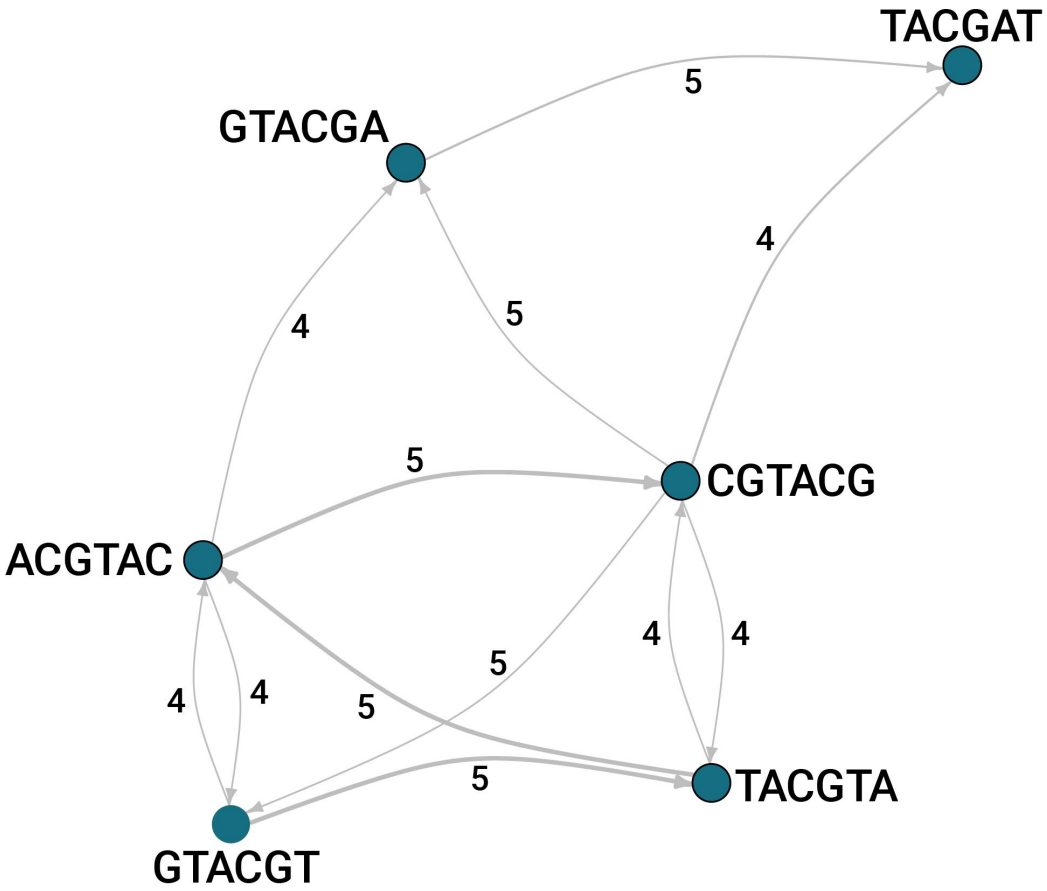
Overlap graph

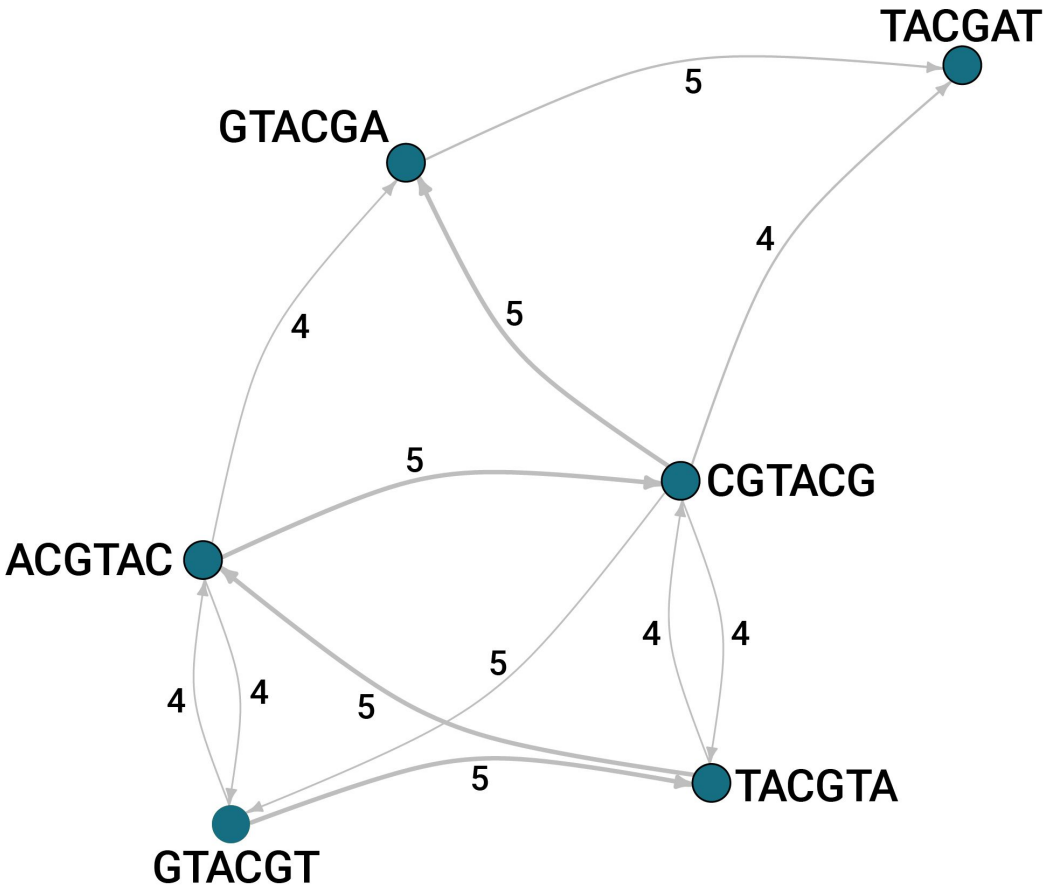
- Each read is a node
- Draw an edge between A and B if suffix of A overlaps with prefix of B
- Contigs are reconstructed by walking along unambiguous paths
- Remove cycles, and at branching paths continue on the edge with the highest overlap

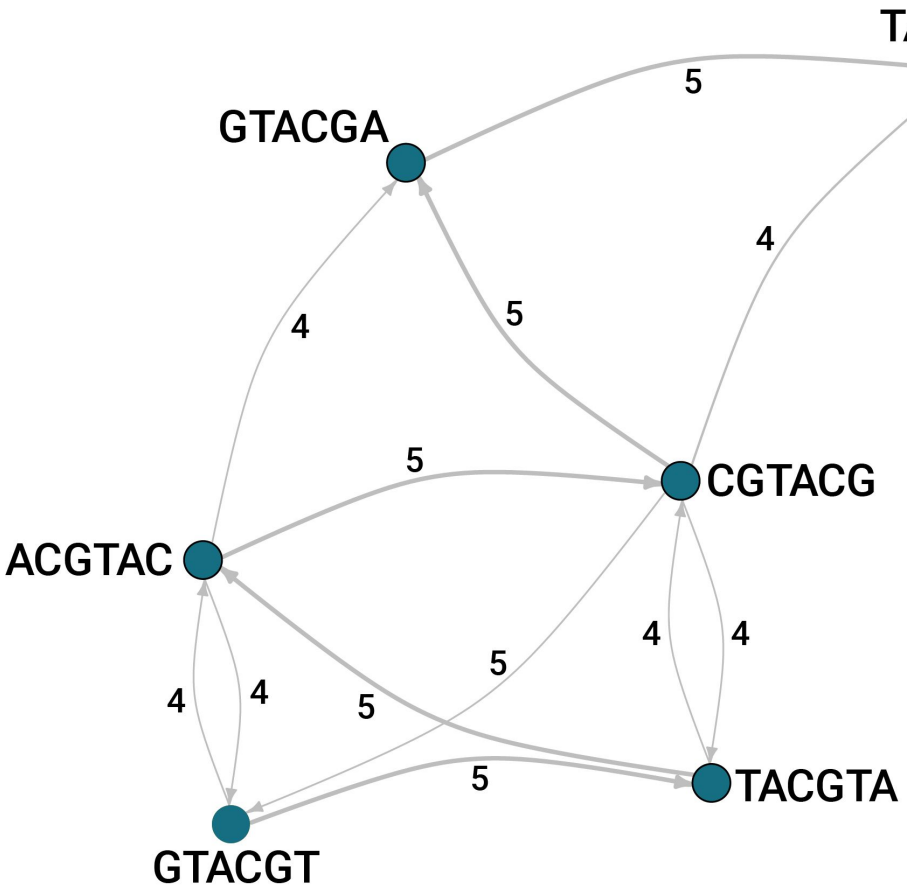



TACGAT
5
GTACGA
4
5
4
5
CGTACG
ACGTAC
4
4
5
4
4
5
5
TACGTA
GTACGT


TACGAT
GTACGA
5
4
5
4
5
CGTACG
ACGTAC
4
4
5
4
4
5
5
TACGTA
GTACGT

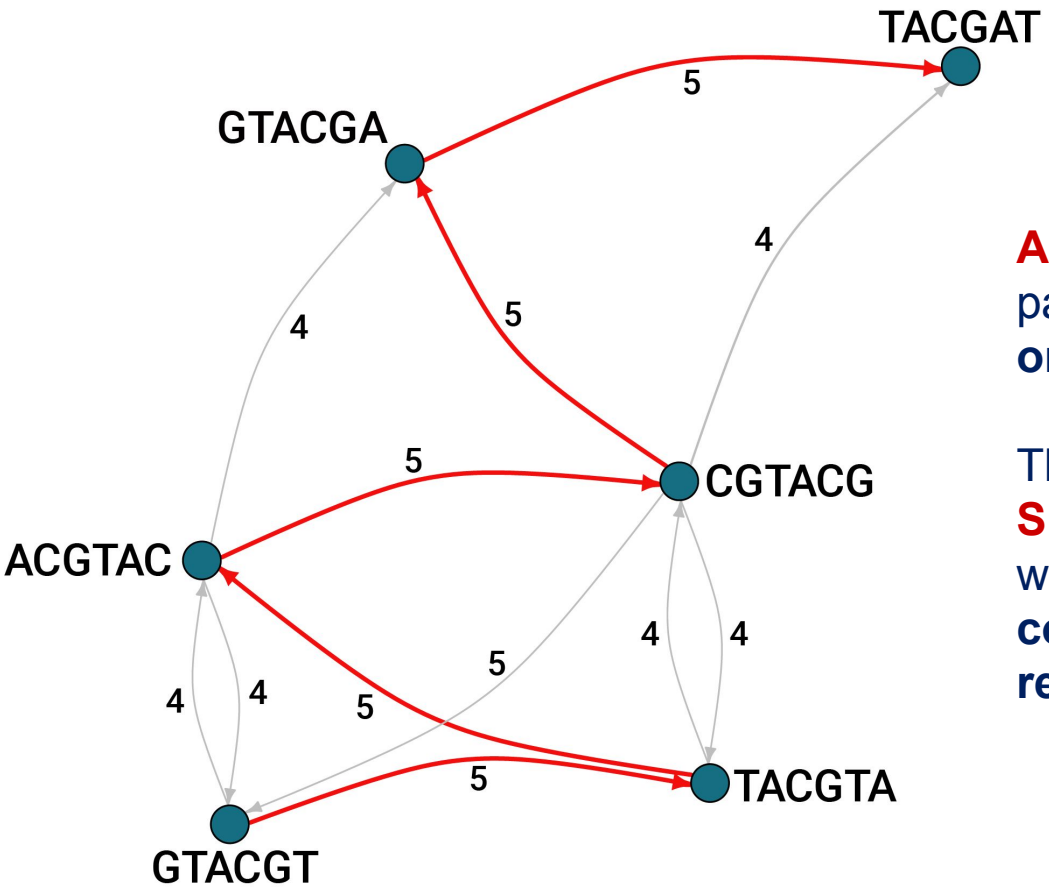

TACGAT
GTACGA
5
4
5
4
5
CGTACG
ACGTAC
4
4
5
4
4
5
5
TACGTA
GTACGT


TACGAT
GTACGA
5
4
5
4
5
CGTACG
ACGTAC
4
4
5
4
4
5
5
TACGTA
GTACGT



GTACGTACGAT

Original contig

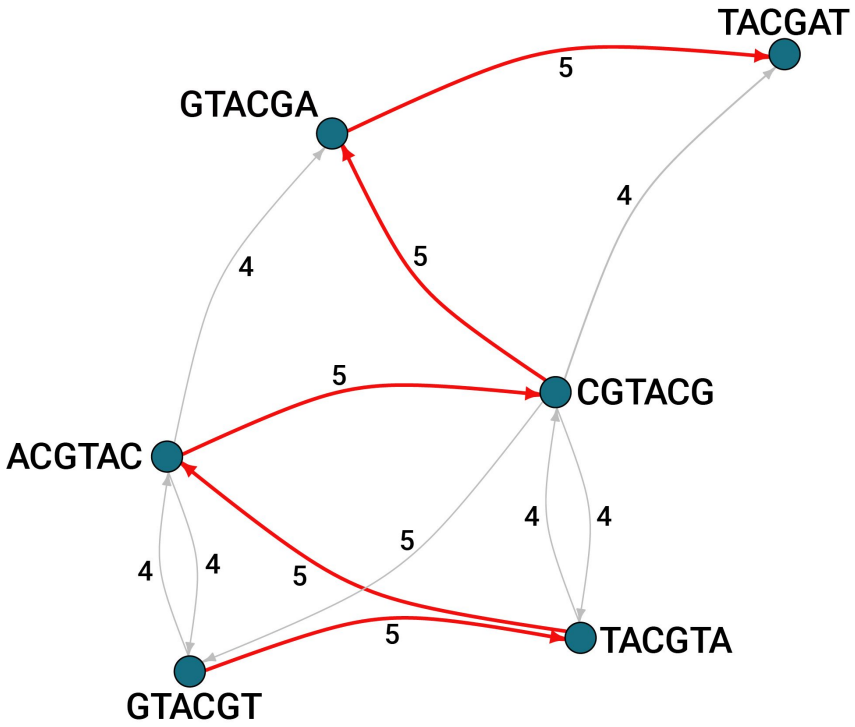


A Hamiltonian path in a graph is a path that visits each **node exactly once**

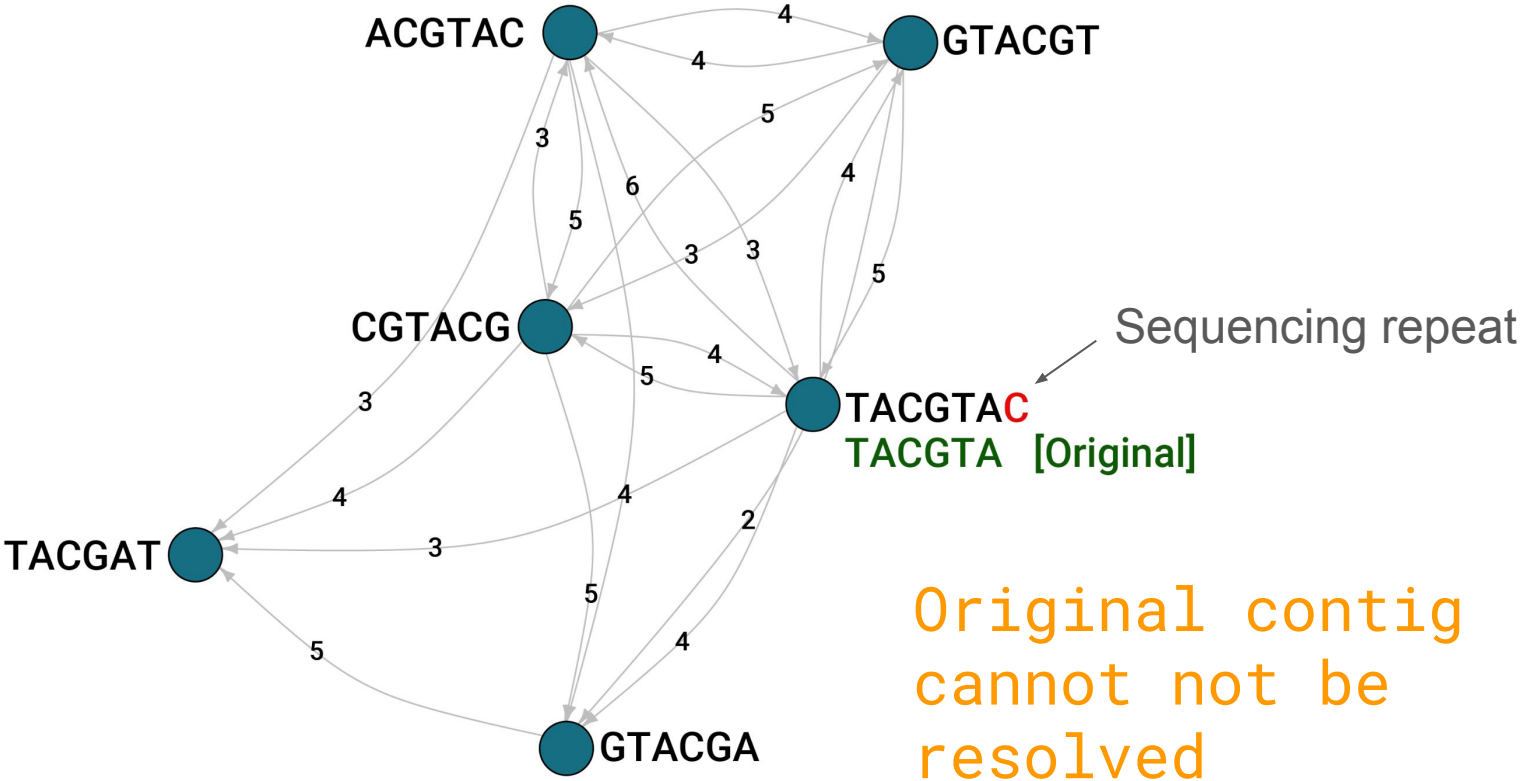
This path (in this case) is also the **Shortest Common Superstring** which **represents the most compact way to cover all the reads, minimizing redundancy**

Overlap graphs & SCS are not feasible on short-reads

- **Quadratic Complexity in Pairwise Comparisons:** given N reads, this results in N * (N - 1) comparisons, which scales quadratically with the number of reads
- **Finding the Hamiltonian path that gives the exact SCS is NP-hard**
- **Sequencing repeats and errors create ambiguous overlaps**



Repeats introduce ambiguities



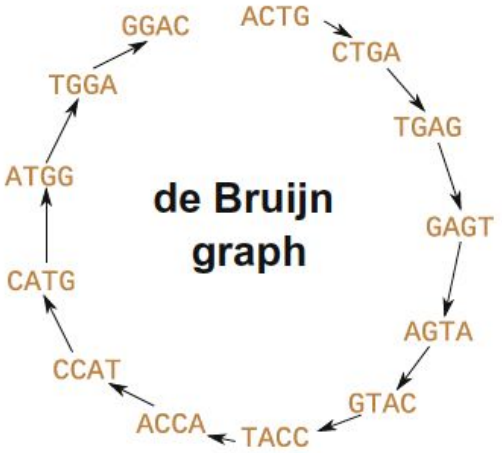
Original contig cannot not be resolved

Modern short-read assemblers use the de Bruijn graph

Reference genome ACTGAGTACCATGGAC

Sequenced reads
ACTGAGTAC
CTGAGTACCAT
GAGTACCATGGAC

k-mers
ACTG TACC GGAC
CTGA ACCA
TGAG CCAT
GAGT CATG
AGTA ATGG
GTAC TGGA



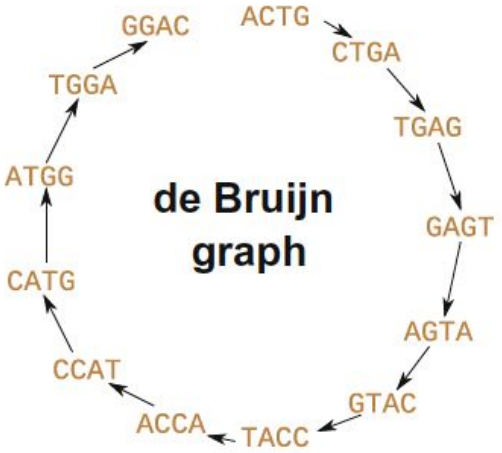
Scales linearly instead of quadratically

Modern short-read assemblers use the de Bruijn graph

Reference genome ACTGAGTACCATGGAC

Sequenced reads
ACTGAGTAC
CTGAGTACCAT
GAGTACCATGGAC

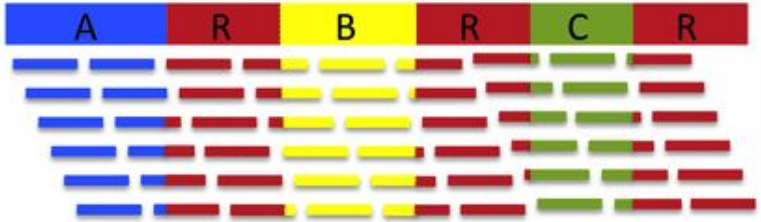
k-mers
ACTG TACC GGAC
CTGA ACCA
TGAG CCAT
GAGT CATG
AGTA ATGG
GTAC TGGA

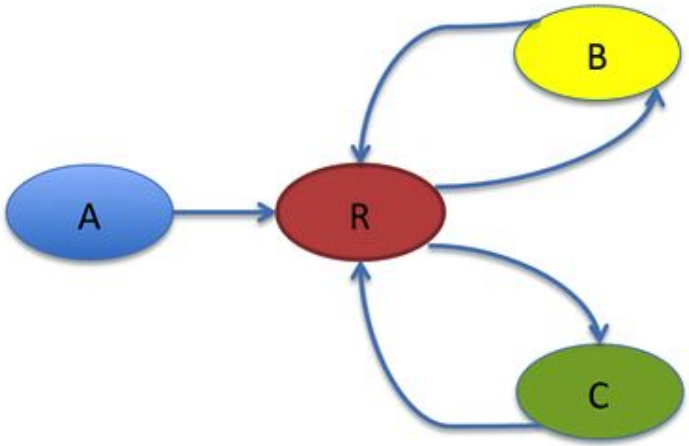


Eulerian path: Each edge is visited exactly once

Short Read Assembly

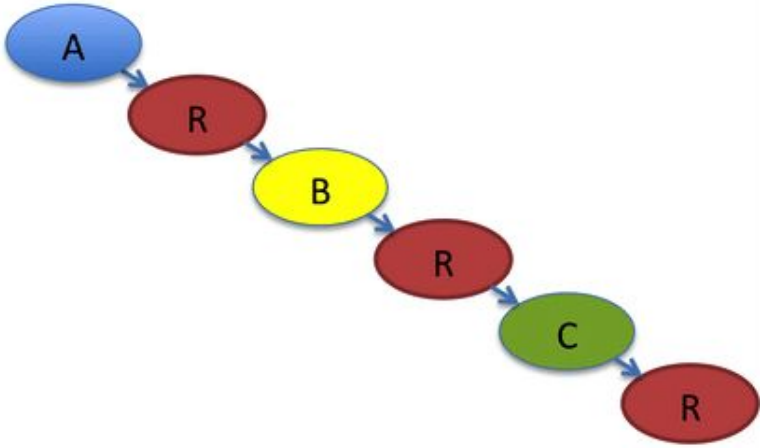
(read length < repeat length)





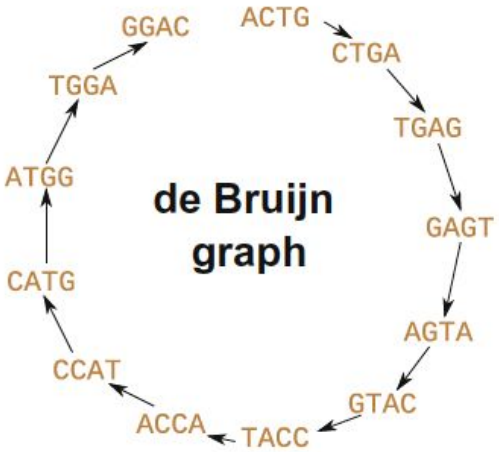
Long Read Assembly

(read length > repeat length)



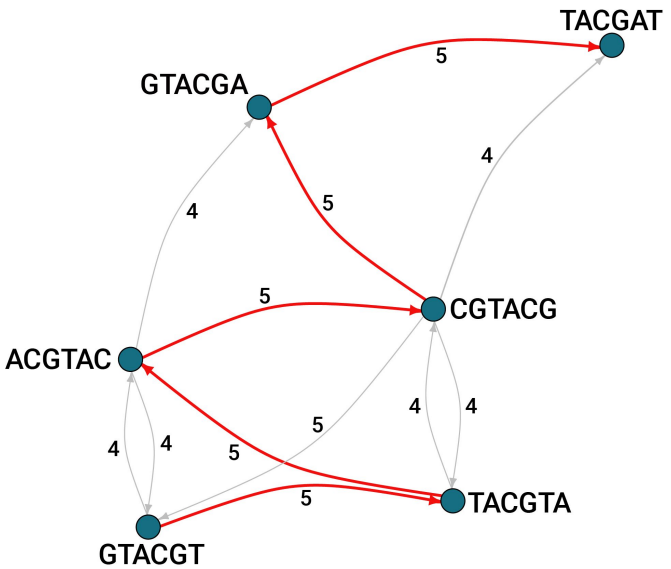
Short Read Assembly

(read length < repeat length)



Long Read Assembly

(read length > repeat length)

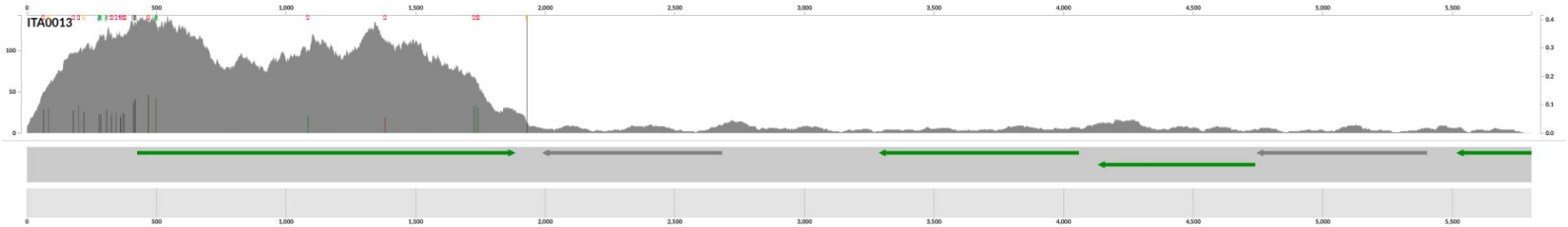


Overlap-Layout-Consensus

Assessing assemblies

Quast

- **Number** of contigs
- **Average/median** contig length
- **Min/Max** contig length
- **N50:** The length of the contigs which covers 50% of genome
- **Read recruitment:** Percentage of all reads mapped back to the assembly
- **Evenness in depth along contig**



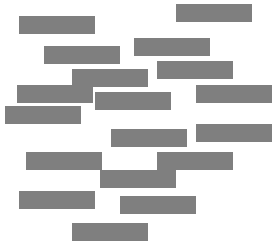
https://merenlab.org

Metagenomics workflow: Coverage

DNA extraction

Sequencin

Reads



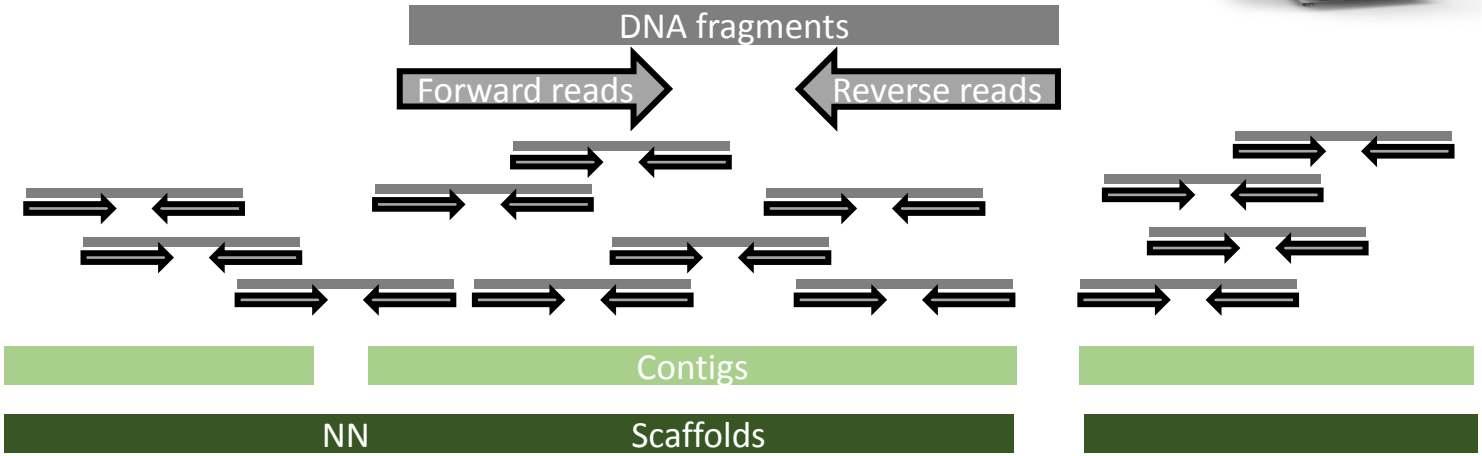
100-150 bp

Assembly

Contigs

1000+ bp

***Scaffold**

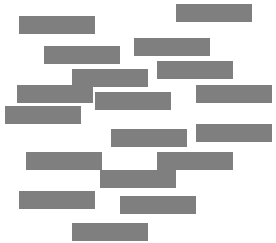


Metagenomics workflow: Coverage

DNA extraction

Sequencin

Reads



100-150 bp

Assembly

Scaffolds

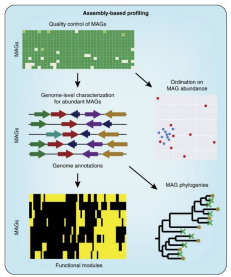
1000+ bp

Mapping

Taxonomy

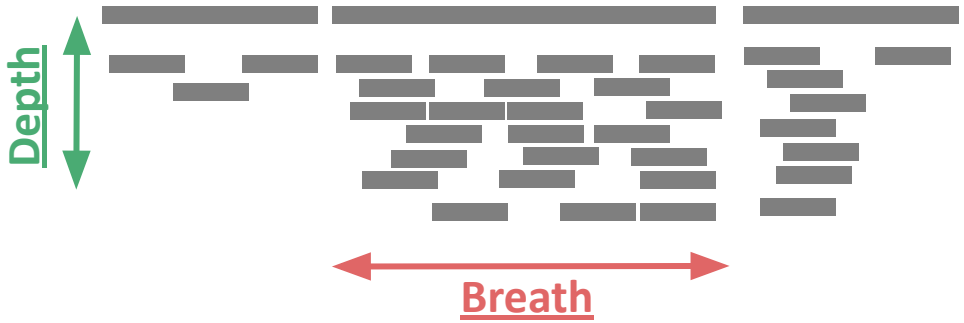
Function

Genome



e.g.
Bowtie2
BWA
SAMtools
…

Coverage



Depth of coverage (mapping depth)

- Average number of times each nucleotide is covered in the assembly
  - Estimate to the abundance of a sequence in the sample

Breath of coverage (covered length)

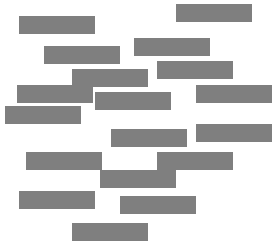
- Percentage of bases of a targeted genome that are covered with a certain depth
  - Metagenomic assembly quality – percentage of data included in the assembly
  - Identify chimeric regions

Metagenomics workflow: Taxonomy

DNA extraction

Sequencing

Reads

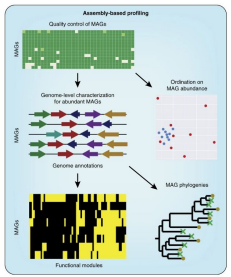


100-150 bp

Assembly

Scaffolds

1000+ bp



Stay tuned: Tuesday October 15th

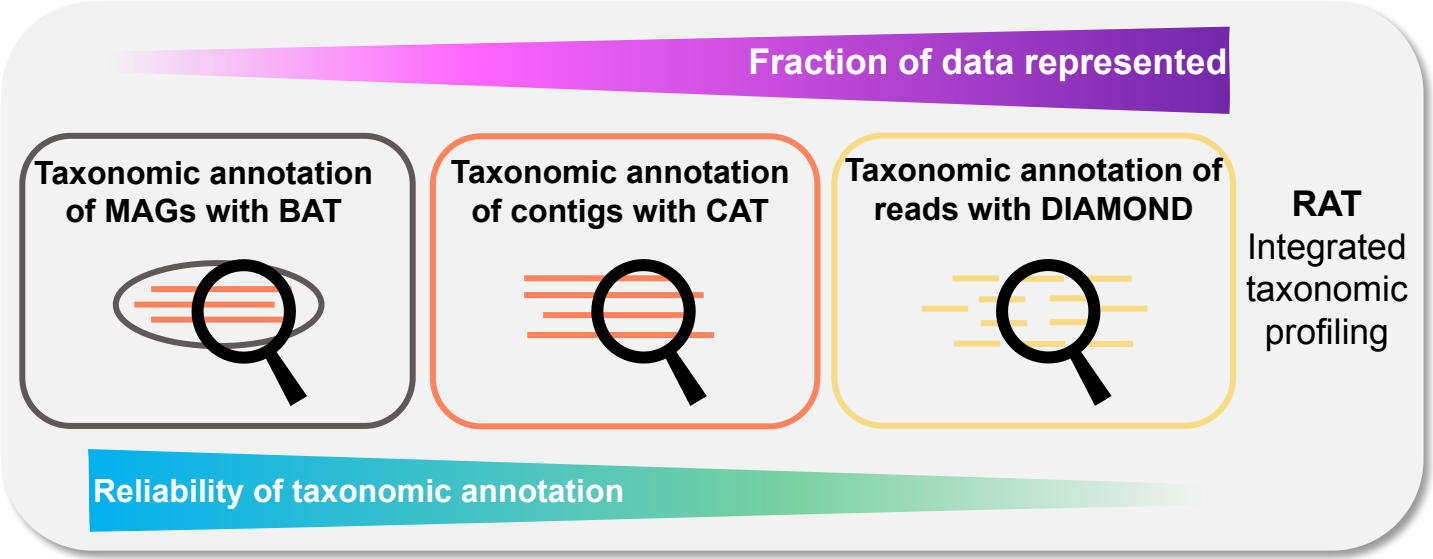
Mapping

Taxonomy

Function

Genome

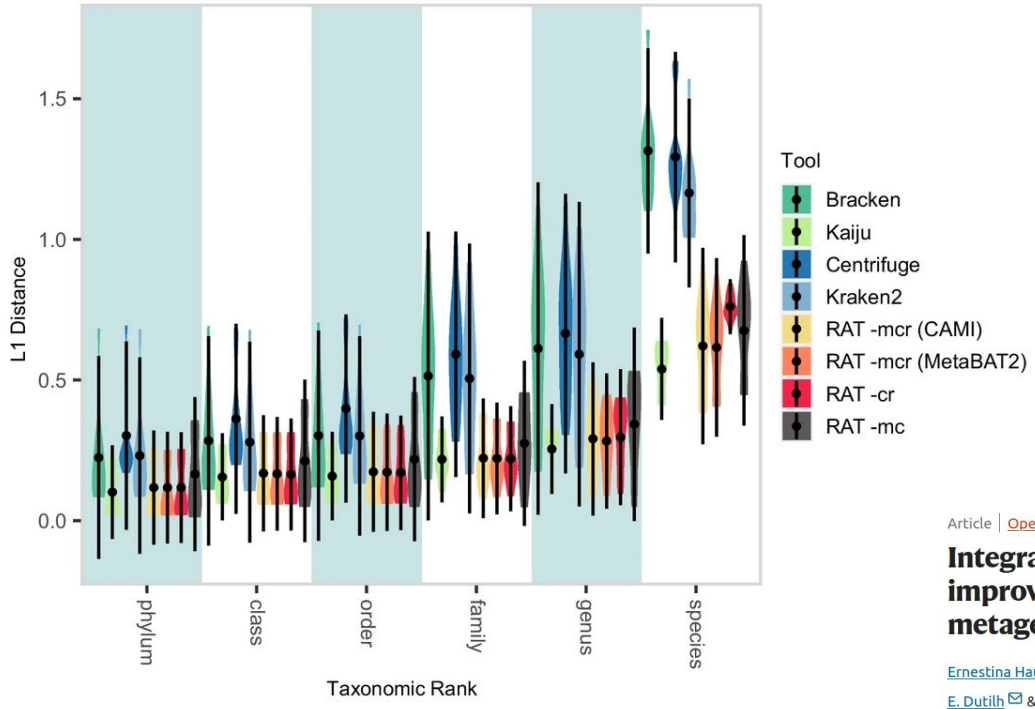
A new tool, RAT, expanding taxonomy assignment on all three levels



+ Function

20 metagenomic classifiers compared: Simon et.al, 2019 (https://doi.org/10.1016/j.cell.2019.07.010)

A new tool, RAT, expanding taxonomy assignment on all three levels



Article | Open access | Published: 20 April 2024

Integrating taxonomic signals from MAGs and contigs improves read annotation and taxonomic profiling of metagenomes

Ernestina Hauptfeld, Nikolaos Pappas, Sandra van Iwaarden, Basten L. Snoek, Andrea Aldas-Vargas, Bas E. Dutilh & F. A. Bastiaan von Meijenfeldt

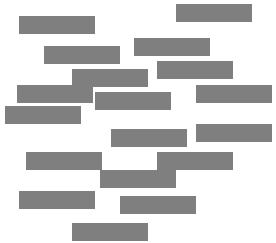
Nature Communications **15**, Article number: 3373 (2024) | Cite this article

Metagenomics workflow: Function

DNA extraction

Sequencing

Reads



100-150 bp

Assembly

Scaffolds

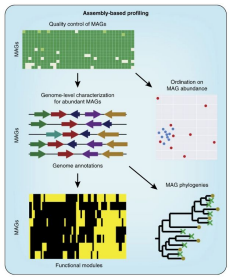
1000+ bp

Mapping

Taxonom

Function

Genome



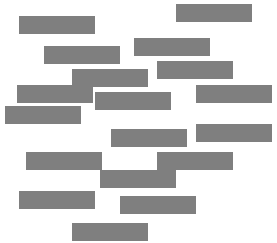
Stay tuned: Tuesday October 15th

Metagenomics workflow: Assembled-based analysis

DNA extraction

Sequencing

Reads



100-150 bp

Assembly

Scaffolds

1000+ bp

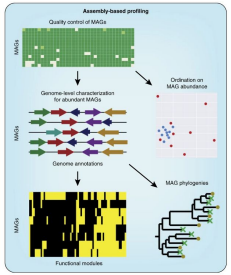
Mapping

Taxonom

Function

Genome

Assembled-based

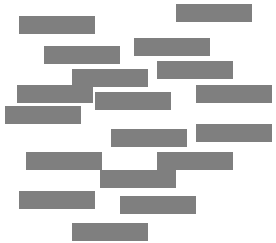


Metagenomics workflow: Binning

DNA extraction

Sequencing

Reads



100-150 bp

Assembly

Scaffolds

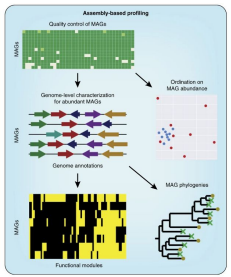
1000+ bp

Mapping

Taxonom

Function

Genome



Binning = Separation of genomes from metagenomes

- Who is there and what can every **individual** do?

PI: How difficult would it be?

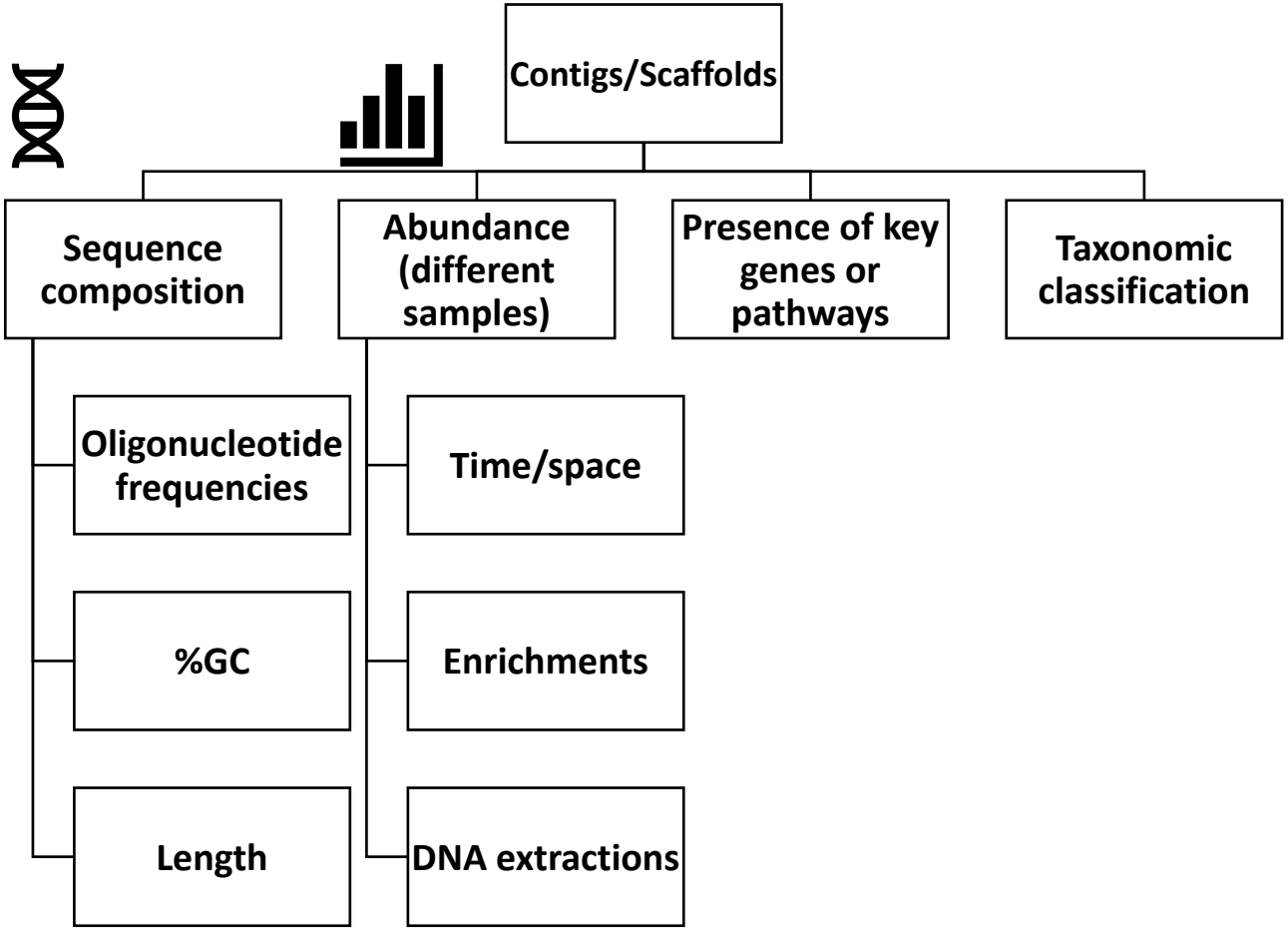
PostDoc: It's challenging yet fun, and there are plenty of standardized methods available to help!

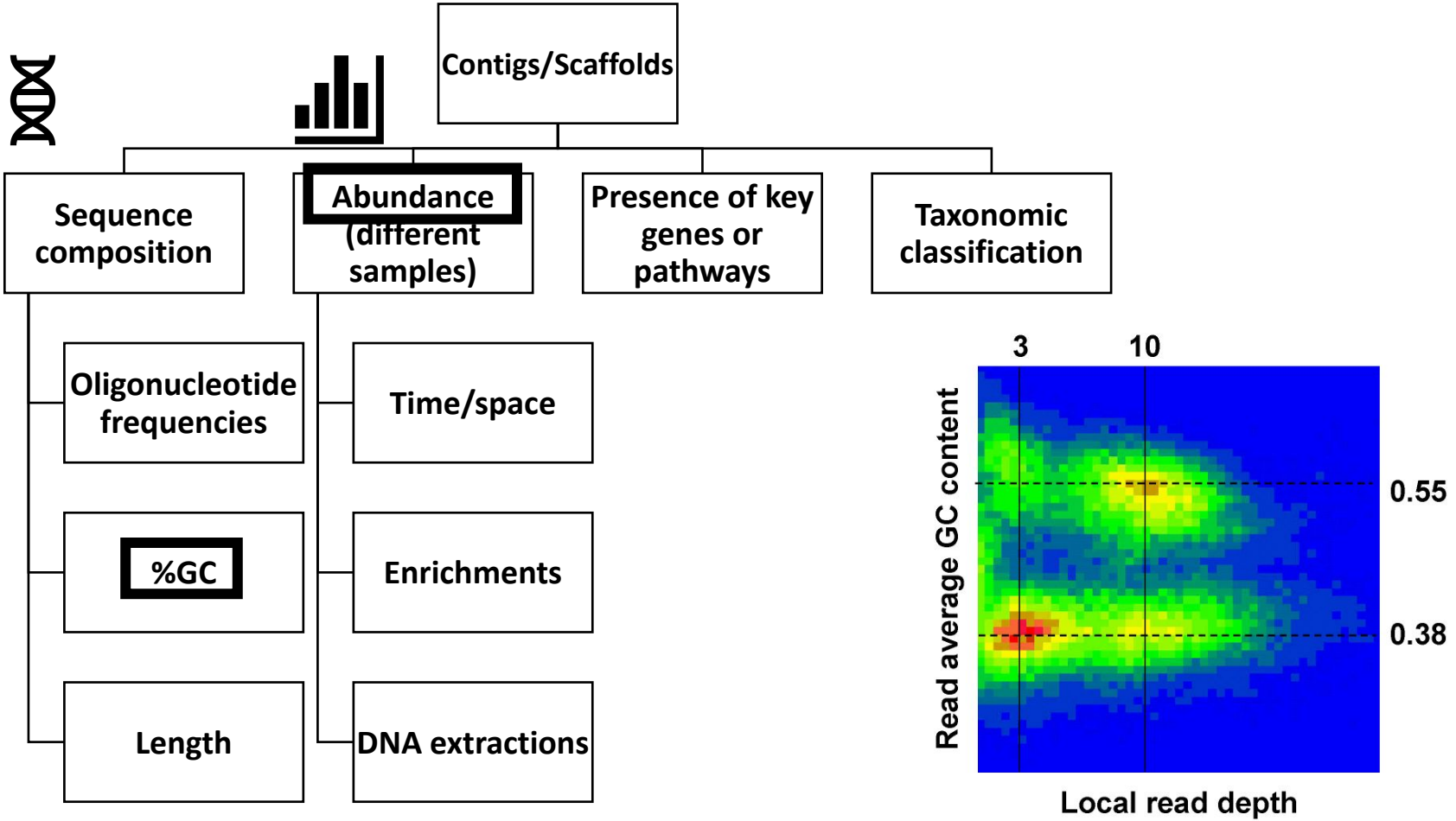
Complex sample

(Contigs)

PhD candidates & MSc interns

Binning


Contigs/Scaffolds
Sequence composition
Abundance (different samples)
Presence of key genes or pathways
Taxonomic classification
Oligonucleotide frequencies
%GC
Length
Time/space
Enrichments
DNA extractions



Tyson et al. 2004

Contigs/Scaffolds

Sequence composition

Abundance (different samples)

Presence of key genes or pathways

Taxonomic classification

Oligonucleotide frequencies

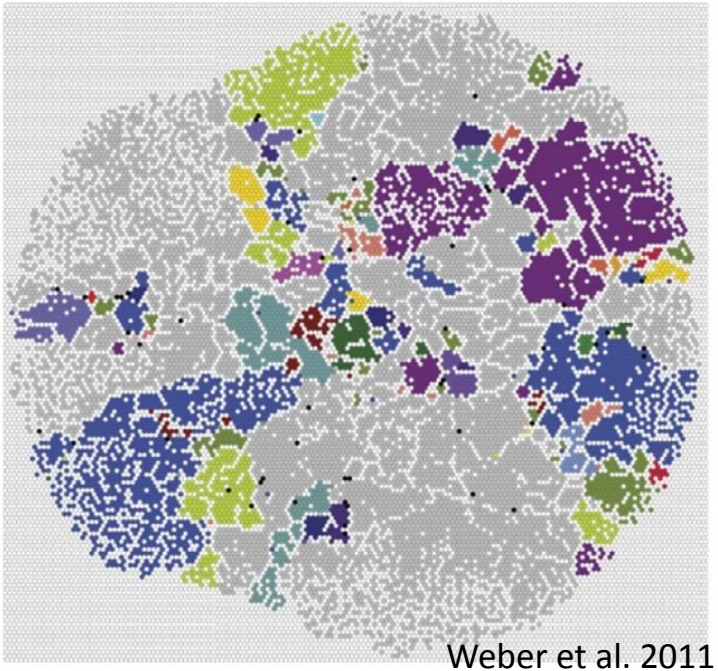
%GC

Length

Time/space

Enrichments

DNA extractions



Weber et al. 2011

Contigs/Scaffolds

Sequence composition

Abundance (different samples)

Presence of key genes or pathways

Taxonomic classification

Oligonucleotide frequencies

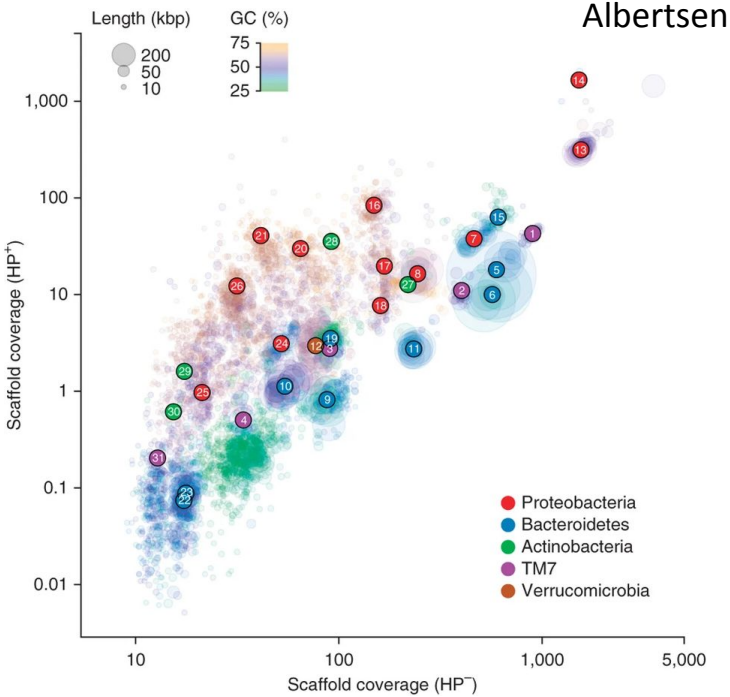
%GC

Length

Time/space

Enrichments

DNA extractions



Albertsen et.al, 2013

Contigs/Scaffolds

Sequence composition

Abundance (different samples)

Presence of key genes or pathways

Taxonomic classification

Oligonucleotide frequencies

%GC

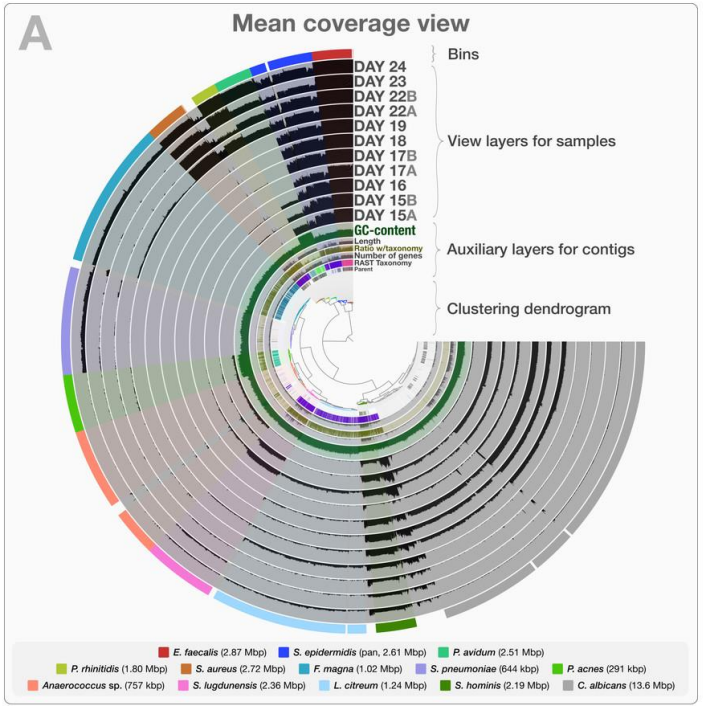
Length

Time/space

Enrichments

DNA extractions

Eren et.al, 2015

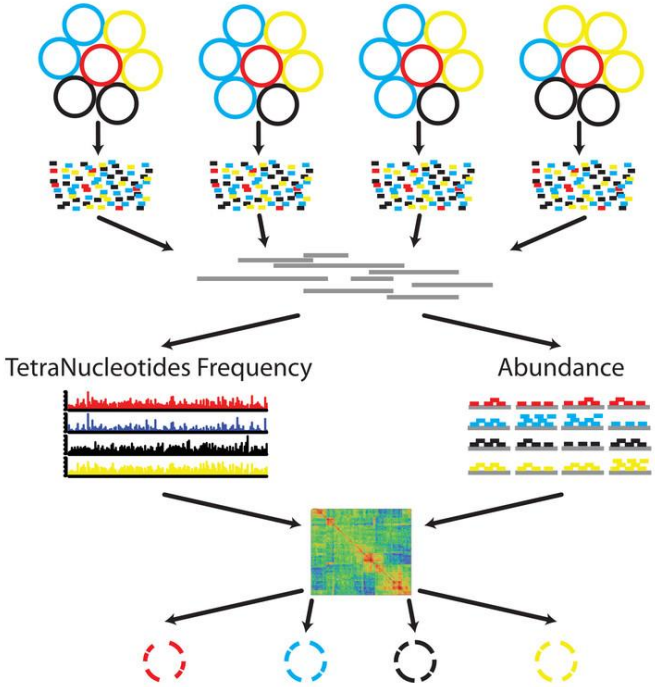


Automatic tools for binning

Tools:
MetaBAT2
Maxbin2
CONCOCT

Aggregate multiple binning results:
e.i. DASTool

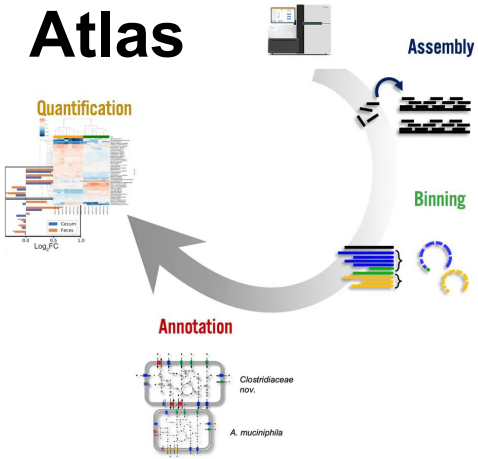
New kids in the playground:
Vamb
SemiBin
MetaDecoder



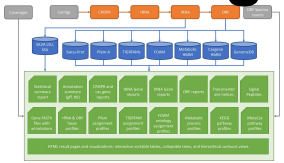
Overview of the MetaBAT pipeline.

(Semi)Automatic pipelines

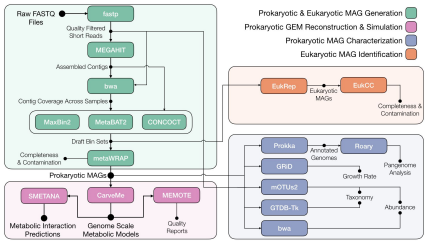
Atlas



MetaErg



metaGEM



MetaWRAP

>Standardization of metagenomics is (not) required!

In practice

Manually

Mixed

Automated

High sensitivity
High specialization
High time consumption

Aggregation:
DAS Tools
MetaWRAP

Bad sensitivity
Easy to use
Fast results

Binners:
MetaBAT2
CONCOCT
MaxBin2
…

Good sensitivity
Medium specialization
Medium time consumption

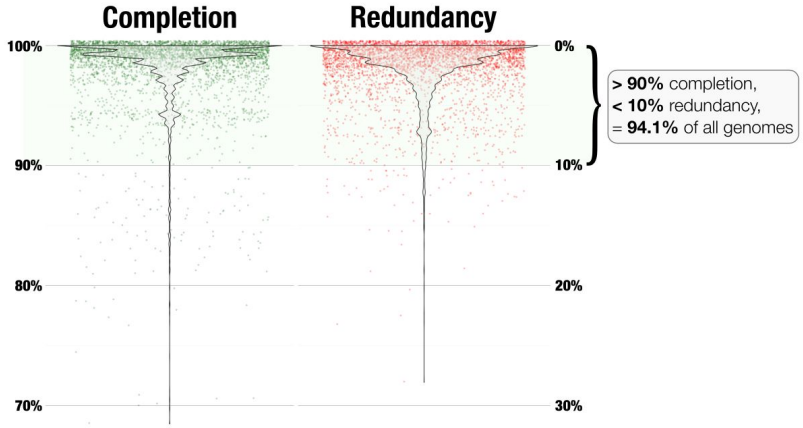
Refinement:
Anvi'o
ProDeGe

MAG quality assessment

- Single-copy marker genes

1. Completeness/completion
   Marker genes are expected to be present in all bacteria
2. Contamination/redundancy
   Single-copy genes are expected to be only once

4,022 closed genomes from NCBI



https://merenlab.org

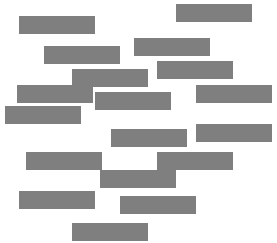
Golden standard: CheckM

Metagenomics workflow: MAG-based analysis

DNA extraction

Sequencin

Reads



100-150 bp

Assembly

Scaffolds

1000+ bp

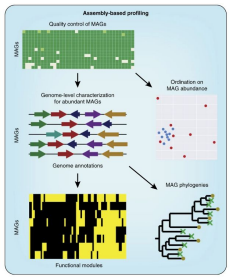
Mapping

Taxonom

Function

Genome

MAG-based



MAG/assembly-based vs. read-based v2

Criteria:	MAG/Assembly-based analysis	Read-based analysis ('mapping')
Comprehensiveness	Low/Medium	Low/Medium/High
Community complexity	Low/Medium	High
Novelty	High	None
Computational burden	High	Low
Genome-resolved metabolism	High	Low
Expert manual supervision	High	Low/Medium
Integration with microbial genomics	High	None

Quince et.al., 2017

MAG/assembly-based vs. read-based v2

Criteria:	MAG/Assembly-based analysis	Read-based analysis ('mapping')
Comprehensiveness	Low/Medium	Low/Medium/High
Community complexity	Low/Medium	High
Novelty	High	None
Computational burden	High	Low
Genome-resolved metabolism	High	Low
Expert manual supervision	High	Low/Medium
Integration with microbial genomics	High	None

Choose according to your question!

Quince et.al., 2017

Which tools to pick?

Microbiome COSI

MICROBIOME
COMMUNITY OF SPECIAL INTEREST

HOME ISMB 2024 CAMI COMMUNITY

CAMI2

CAMI II Challenge Information

We proudly announce the beginning of the second round of challenges of the Initiative for the Critical Assessment of Metagenome Interpretation (CAMI) and release of the official challenge data sets!

Over the last two years, we received valuable feedback from the community on important challenges in the field and how to design interesting new data sets and challenges. We incorporated many of your suggestions, thanks again! For you to familiarize with data set types and formats, additional exemplary data sets together with accompanying standards of truth have already been made available over the last months. Two multisample "toy" data sets representing microbial communities from different human body sites and from mouse gut are already provided to allow participants to prepare for the challenges (https://data.cami-challenge.org/participate). These practice data sets are generated from known genomes, and therefore reference-based methods (e.g., using genome databases for their analysis) might perform better here than for real shotgun metagenomic data, where a substantial portion of microbial community members have not been sequenced.

The second CAMI challenge datasets will therefore again include new genomes from taxa (at different evolutionary distances) not found in public databases. Furthermore, a new focus will be on establishing the value of long sequencing reads for microbiome research, with data sets providing both long- and short-read data. Lastly, a clinical pathogen discovery challenge will be offered, mimicking an emergency diagnostic situation in the clinic.

Specifically, the second round of CAMI challenges comprise a metagenome assembly, a genome binning, a taxonomic binning and a ta across several multi-sample data sets from different environments. This includes a marine data set (ended), a high-strain diversity data s pathogen detection challenge (ended). A new round of challenges on a rhizosphere data set has just started in early 2020!

We are looking forward to receiving your submissions!

The CAMI Team

https://www.nature.com/articles/s41592-022-01431-4

Metagenomics is a great tool but...

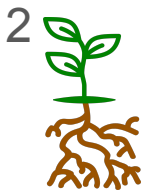
- Abundance is qualitative
  - Not easy to be quantitative with microbial communities
    - ?Integrate metagenomics/barcoding with qPCR, DNA spiking, flow-cytometry and microscopy?
- We are measuring the DNA content, therefore viable & non viable cells
  - RNA, CFUs (if culturable)
- We investigate potential functionality, not activity
  - Multi-omics: Adding layers of information (RNA, protein, metabolites)
- No clue on spatial organization
  - Microscopy

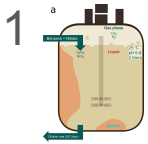
Metagenomics is a great tool but...

- Abundance is qualitative
  - Not easy to be quantitative with microbial communities
    - ?Integrate metagenomics/barcoding with qPCR, DNA spiking, flow-cytometry and microscopy?
- We are measuring the DNA content, therefore viable & non viable cells
  - RNA, CFUs (if culturable)
- We investigate potential functionality, not activity
  - Multi-omics: Adding layers of information (RNA, protein, metabolites)
- No clue on spatial organization
  - Microscopy
- Toooooooooooo much data....
  - That's why you are here!

Challenges of getting genomes from metagenomes across environments

1) Which of those environments have the highest diversity?
2) From which we can get most MAGs?

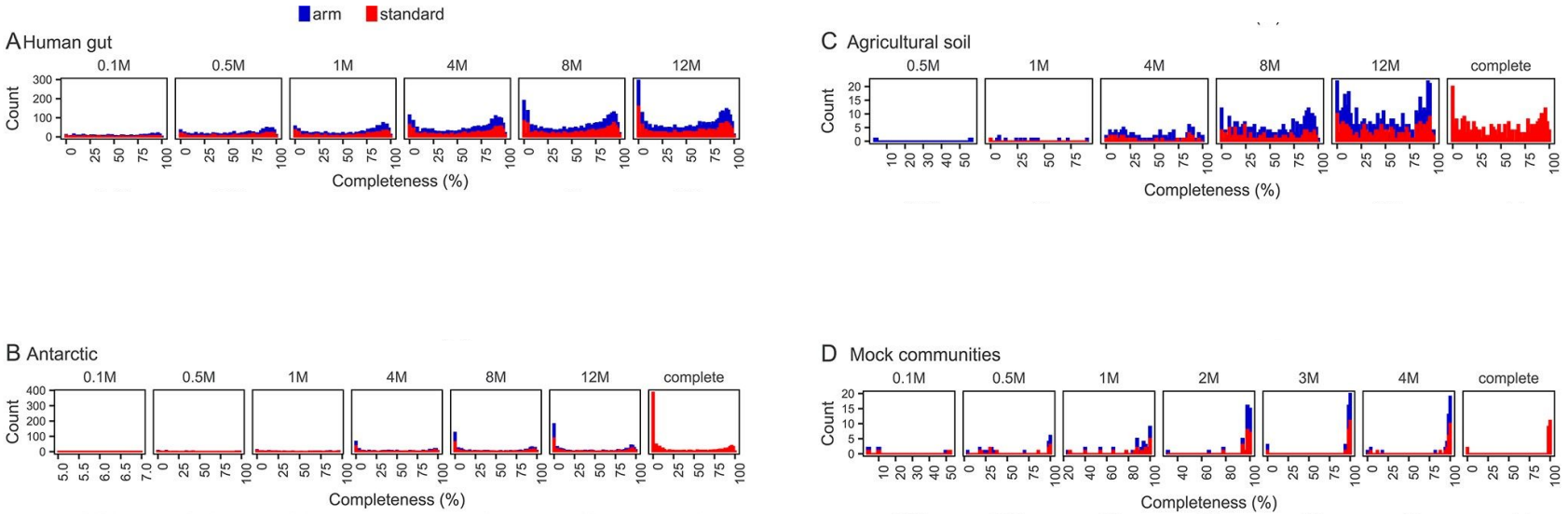




*High quality:
> 90% completeness &
< 10% contamination

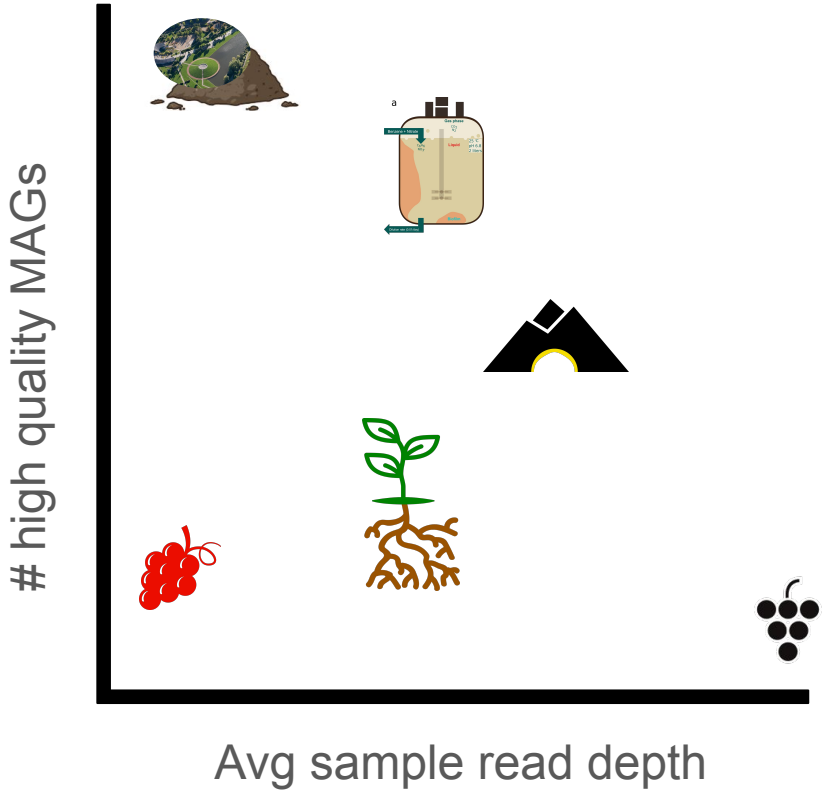
MAG:
metagenome-assembled genome

Challenges of getting genomes from metagenomes across environments: **The more data, the better?**

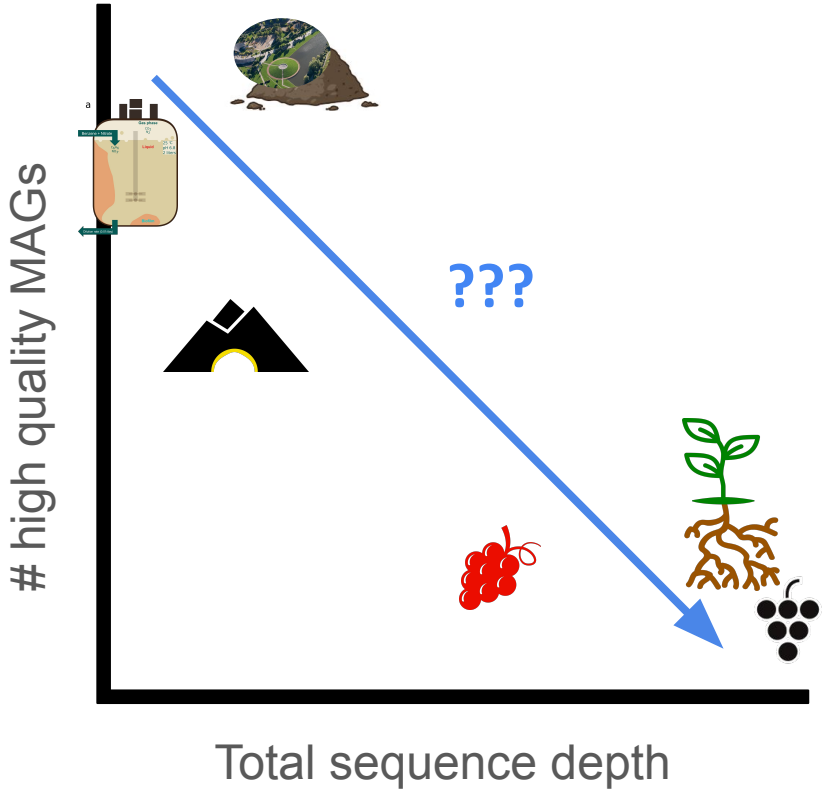


Julien Tremblay, Lars Schreiber, Charles W Greer, High-resolution shotgun metagenomics: the more data, the better?, *Briefings in Bioinformatics*, Volume 23, Issue 6, November 2022, bbac443, https://doi.org/10.1093/bib/bbac443

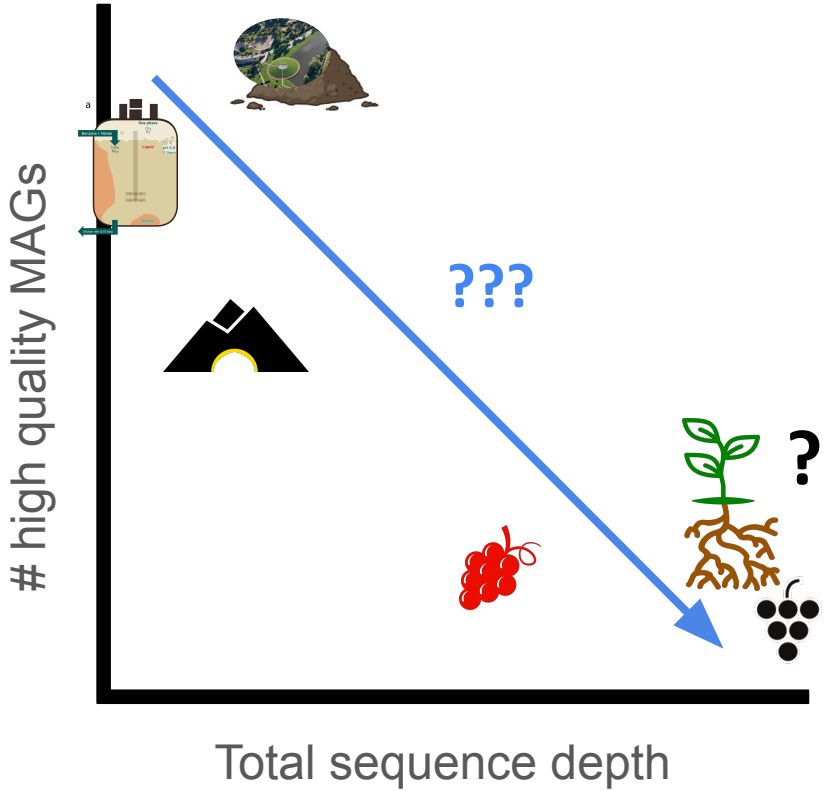
Challenges of getting genomes from metagenomes across environments



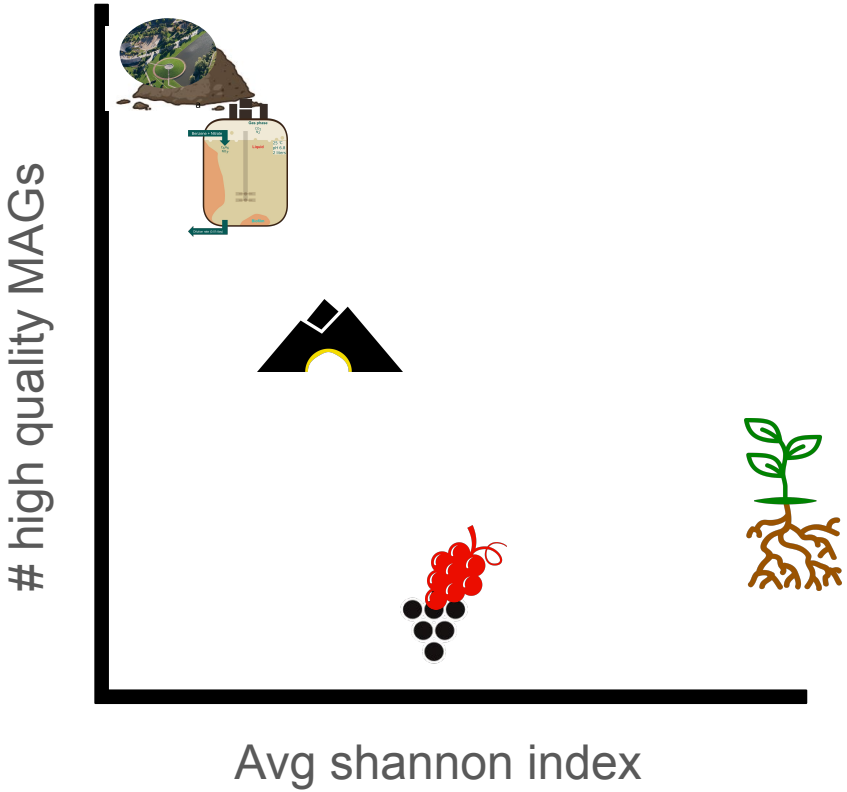
Challenges of getting genomes from metagenomes across environments



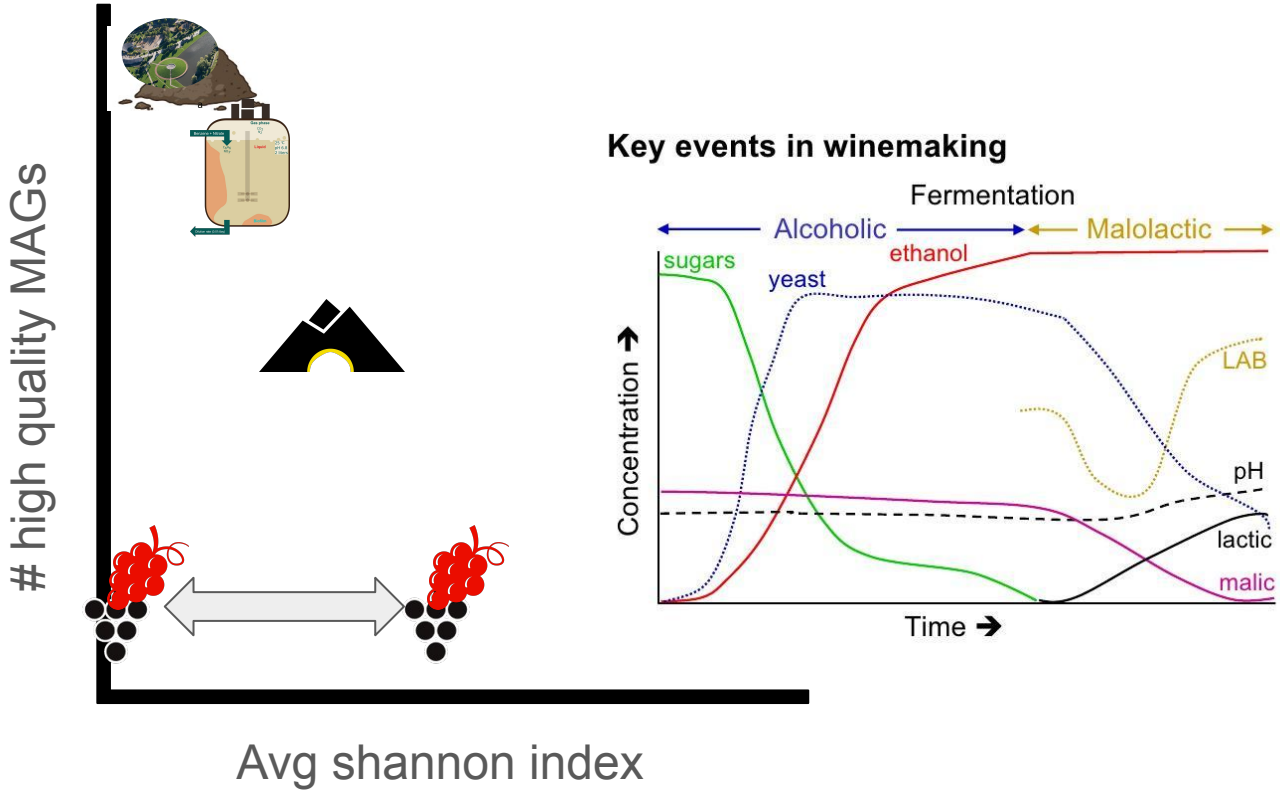
Challenges of getting genomes from metagenomes across environments



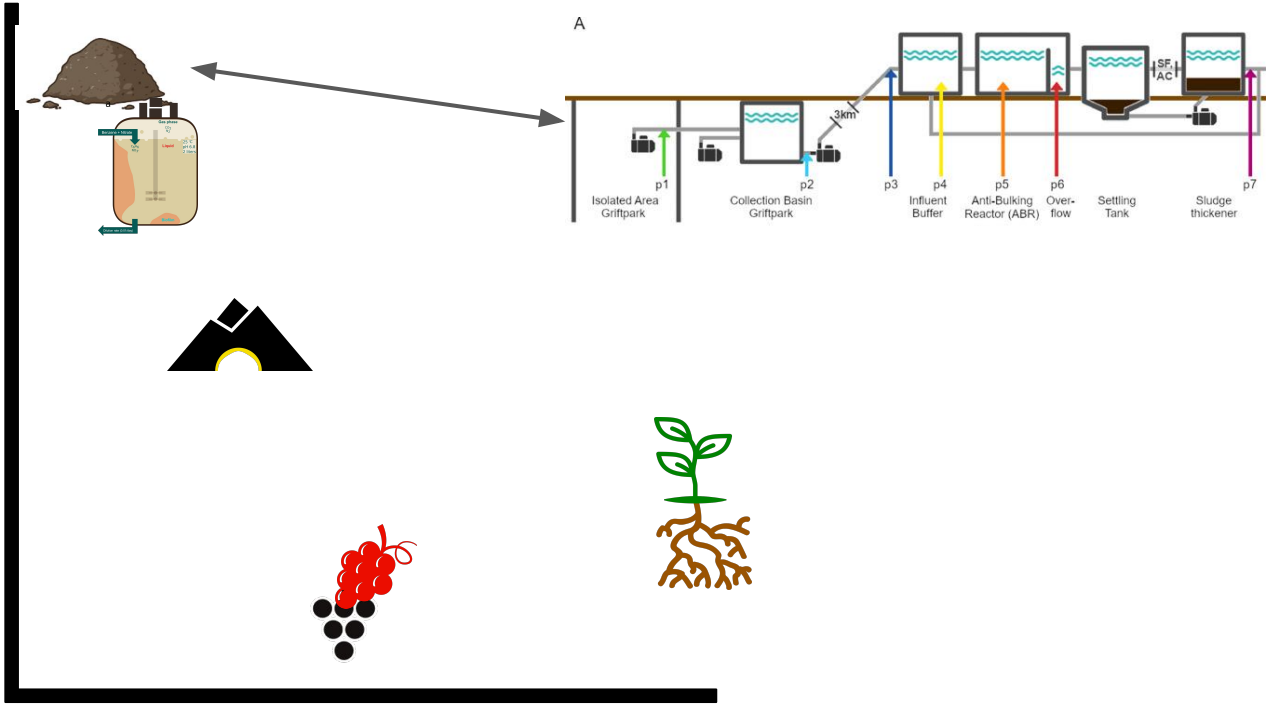
Challenges of getting genomes from metagenomes across environments



Challenges of getting genomes from metagenomes across environments

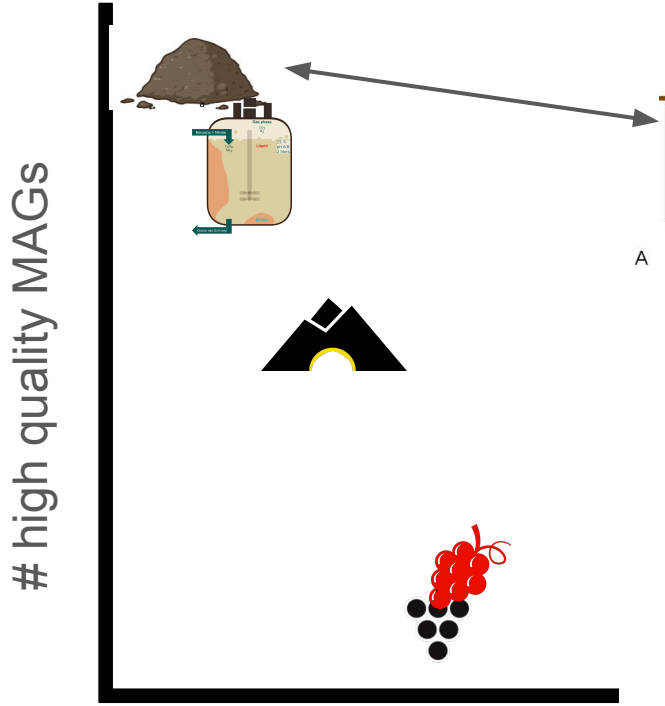


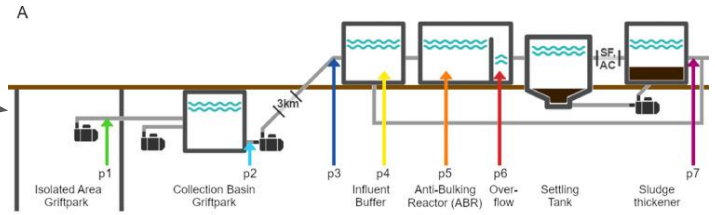
Challenges of getting genomes from metagenomes across environments

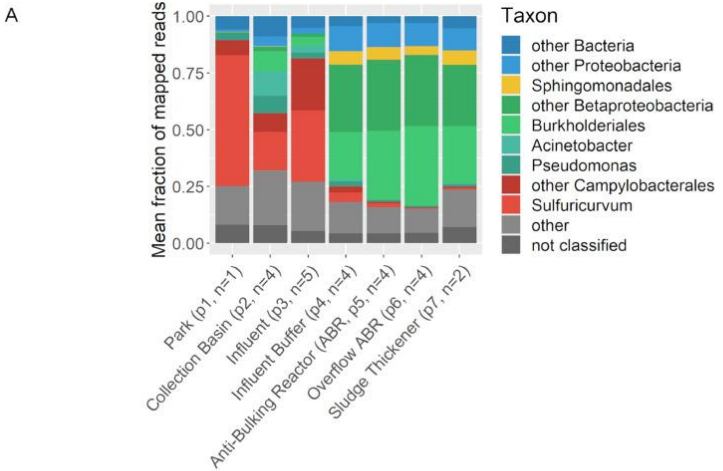


Hauptfeld, E., et. al. (2022). A metagenomic portrait of the microbial community responsible for two decades of bioremediation of poly-contaminated groundwater. In Water Research (Vol. 221, p. 118767). Elsevier BV. https://doi.org/10.1016/j.watres.2022.118767

Challenges of getting genomes from metagenomes across environments

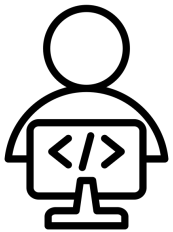






Hauptfeld, E., et. al. (2022). A metagenomic portrait of the microbial community responsible for two decades of bioremediation of poly-contaminated groundwater. In Water Research (Vol. 221, p. 118767). Elsevier BV. https://doi.org/10.1016/j.watres.2022.118767

Practicals

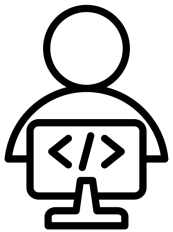


Connect to JupyterHub:
https://bioinformatics.nl/biosb_metagenomics

>https://mdehollander.github.io/biosb-metagenomics/

By Mattias de Hollander

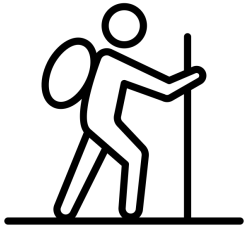
Practicals



Connect to JupyterHub:
https://bioinformatics.nl/biosb_metagenomics

>https://mdehollander.github.io/biosb-metagenomics/

By Mattias de Hollander



Feeling adventurous?
Explore the microbiome of a deadly toxic cave
-> **Cave expedition tab**