

# Bringing Order to Chaos: Metagenomics Starting from Raw Reads

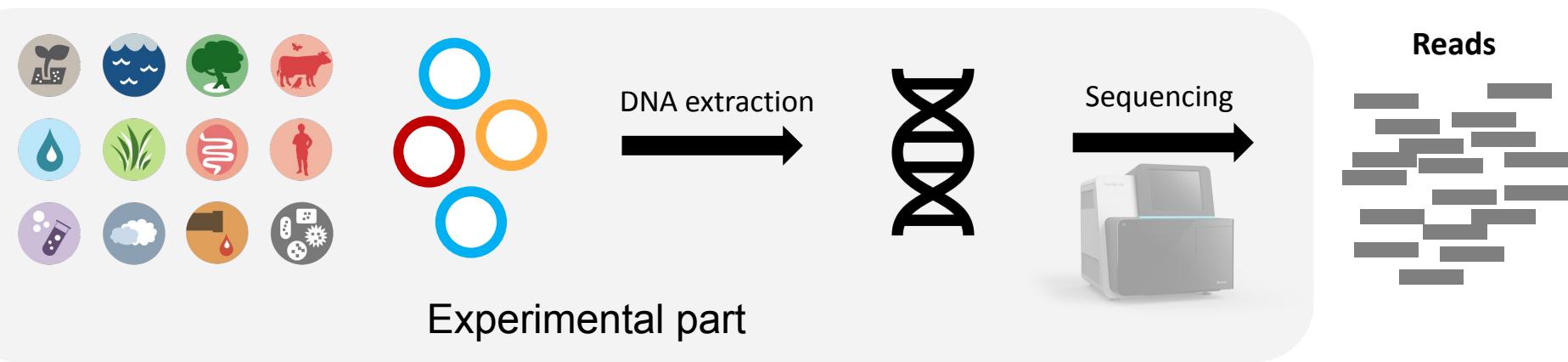


DALL-E

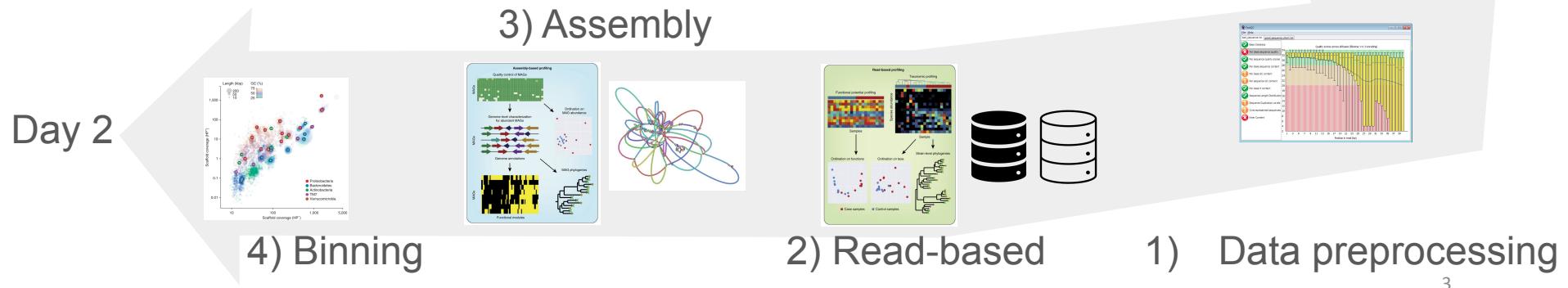
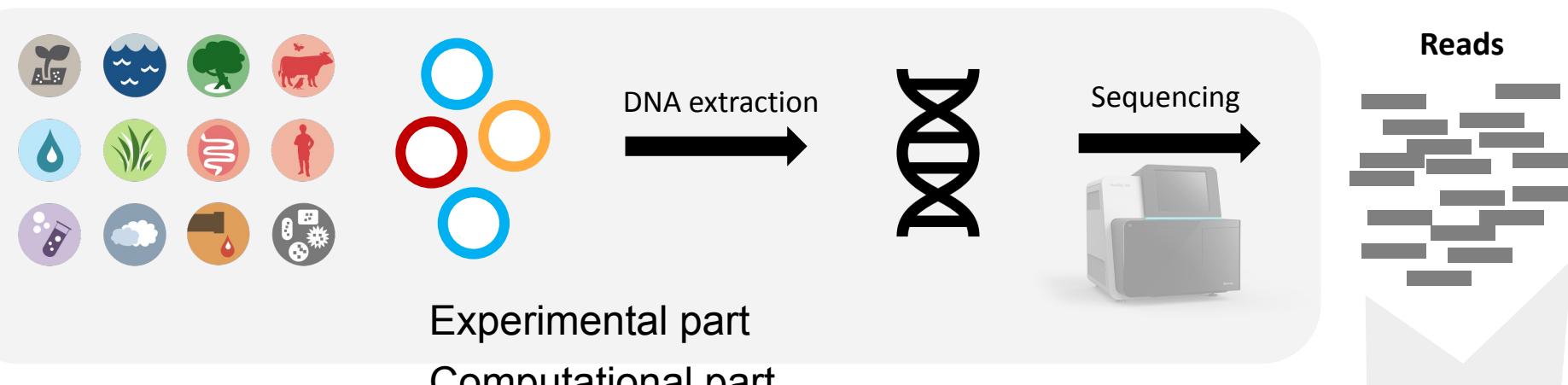
Johannes Björk  
(UMCG/RUG)

Chrats Melkonian  
(WUR & UU; Bioinformatics)

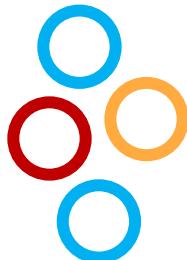
# Metagenomics: Overview



# Metagenomics: Overview (today's focus)



# Metagenomics workflow



DNA extraction



Sequencing

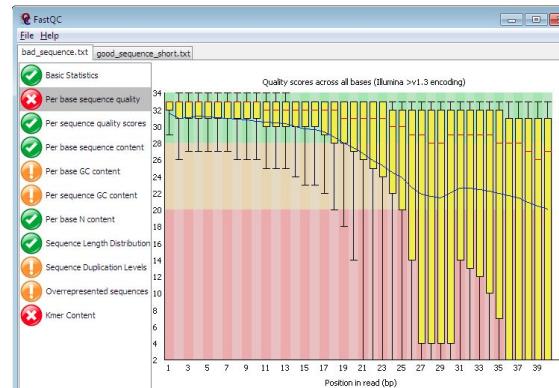


Raw reads

- 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

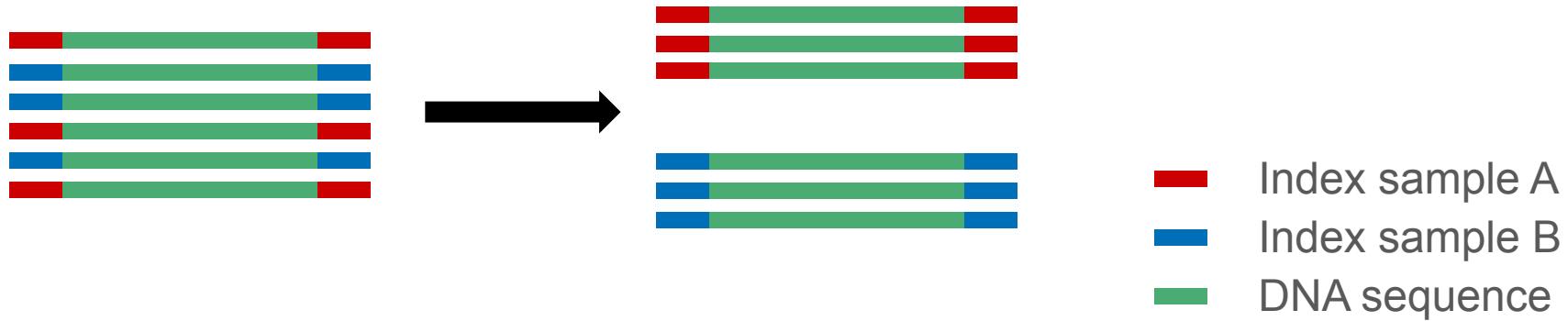


- Index sample A
- Index sample B
- DNA sequence

# 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.

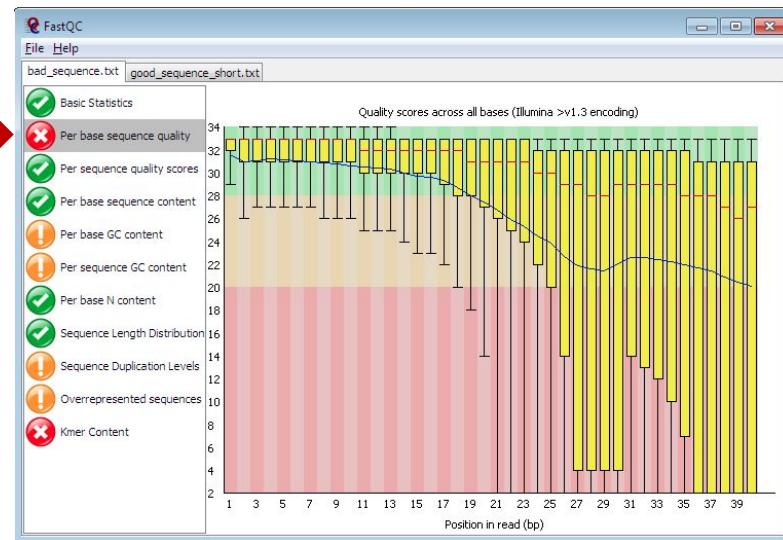
Recommended tools:  
Trimmomatic, Cutadapt, bbduk, fastp

After adapter removal, rerun QC on the fastq files

Keep an eye out for polyA and polyG sequences

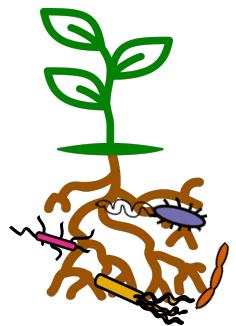
# Data preprocessing - Quality filter

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%

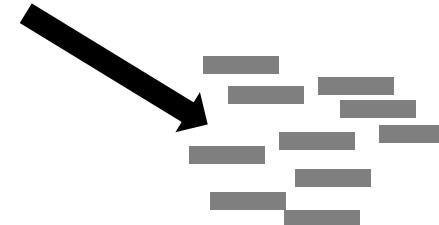
# Data preprocessing - Host (& other) removal



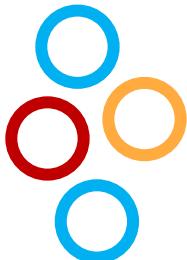
Rhizosphere  
Microbiome



e.g. Bowtie2, bbduk,  
KneadData, HoCoRT



# Metagenomics: Read-based



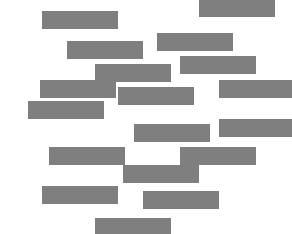
DNA extraction



Sequencing

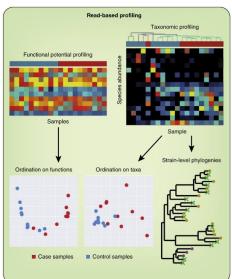


Reads



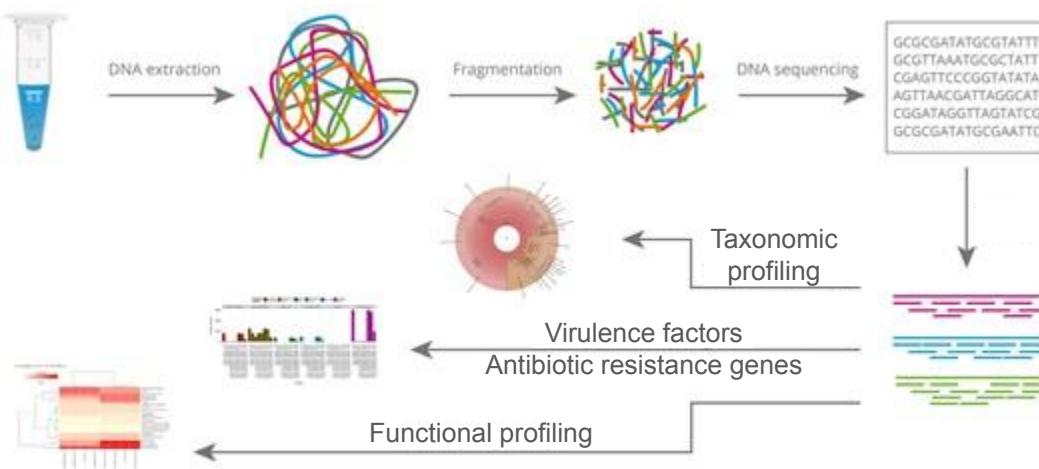
100-150 bp

The read based:  
Mapping into  
databases



Read-based

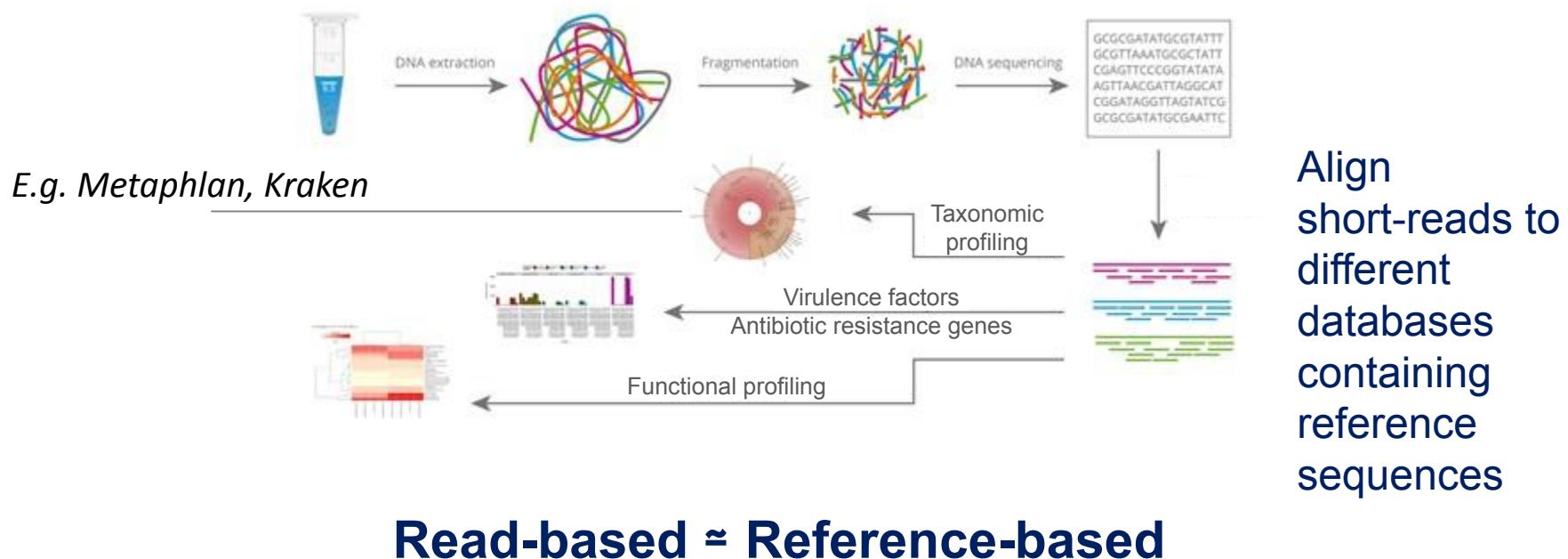
# Read-based approaches



Align short-reads to different databases containing reference sequences

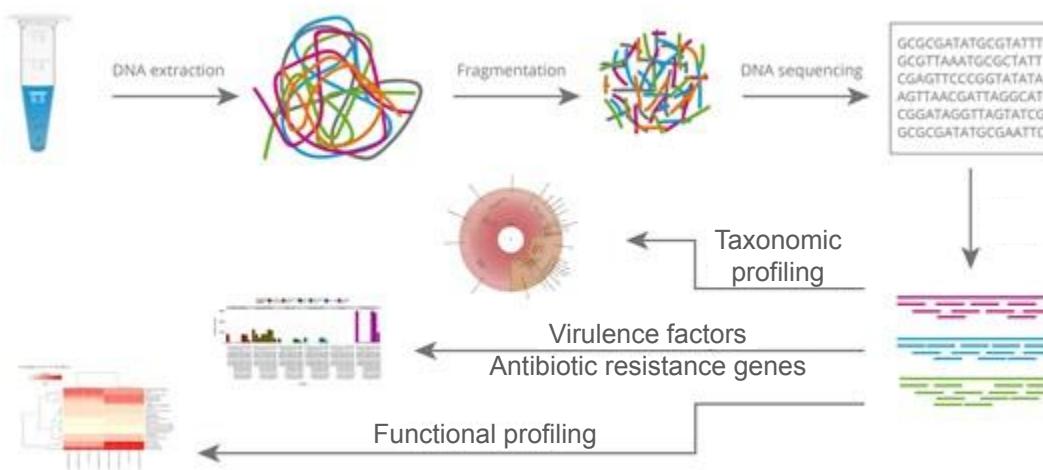
**Read-based ≈ Reference-based**

# Read-based approaches



# Read-based approaches

E.g. HUMAnN,  
MetaPathways

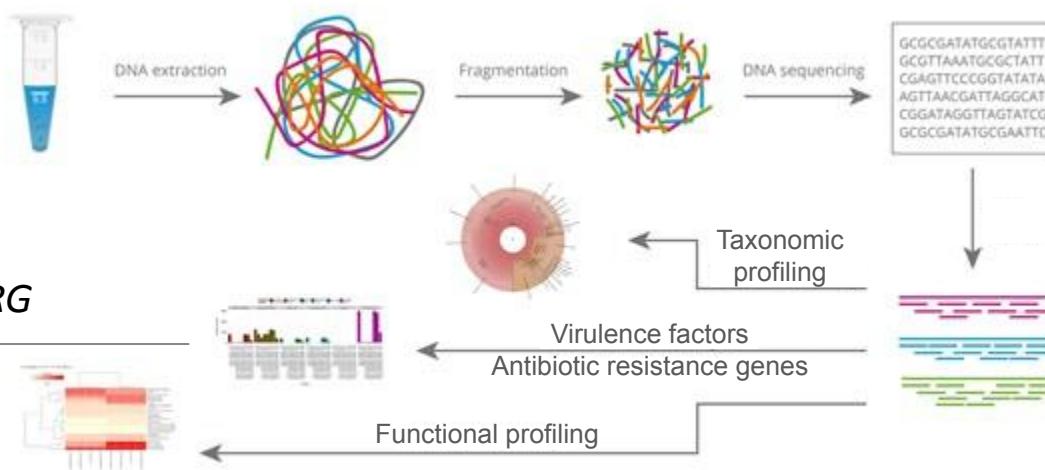


Align short-reads to different databases containing reference sequences

**Read-based ≈ Reference-based**

# Read-based approaches

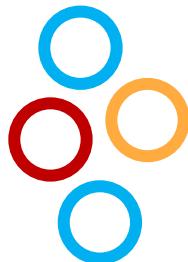
E.g. *ShortBRED*,  
*ResFinder*, *DeepARG*



Align short-reads to different databases containing reference sequences

**Read-based ≈ Reference-based**

# Metagenomics workflow: Assembly-based



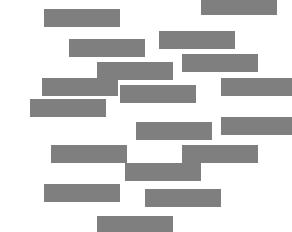
DNA extraction



Sequencing



Reads



100-150 bp

Assembly

Contigs



1000+ bp

Assembly-based

# The Assembly Problem:

Library of books



Shred all books

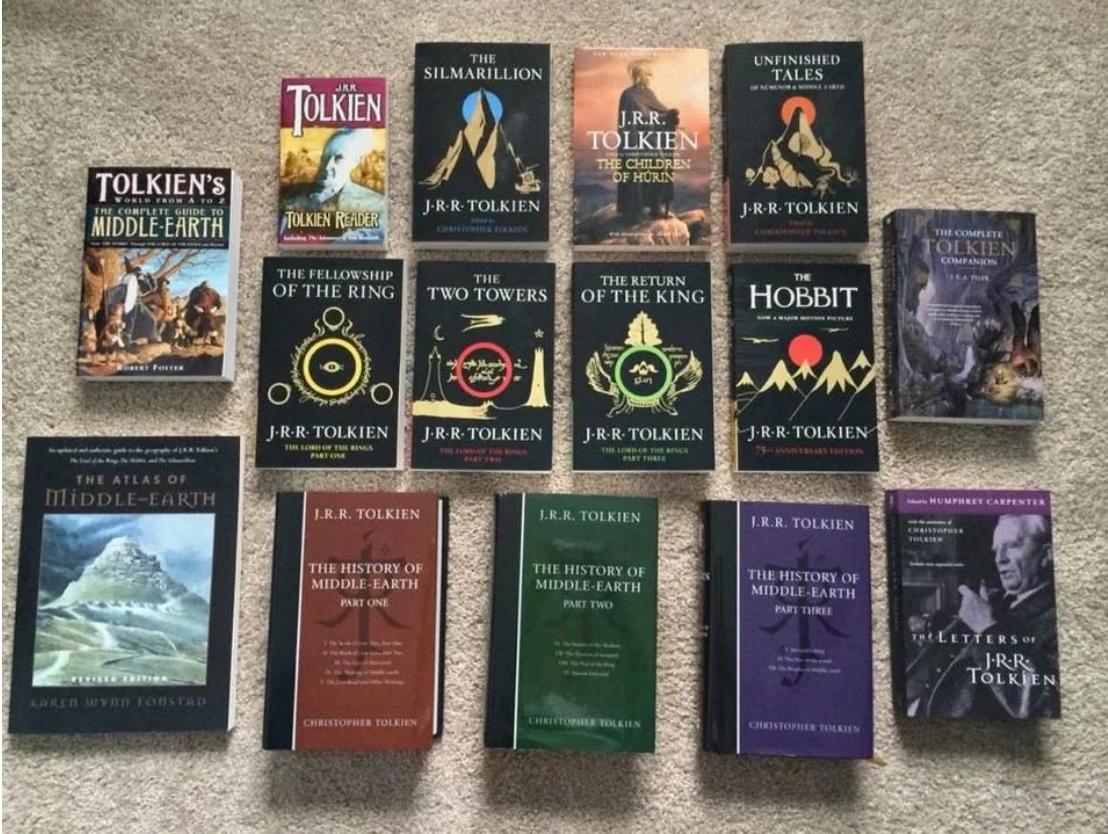


Sequencing

Reconstruct each book



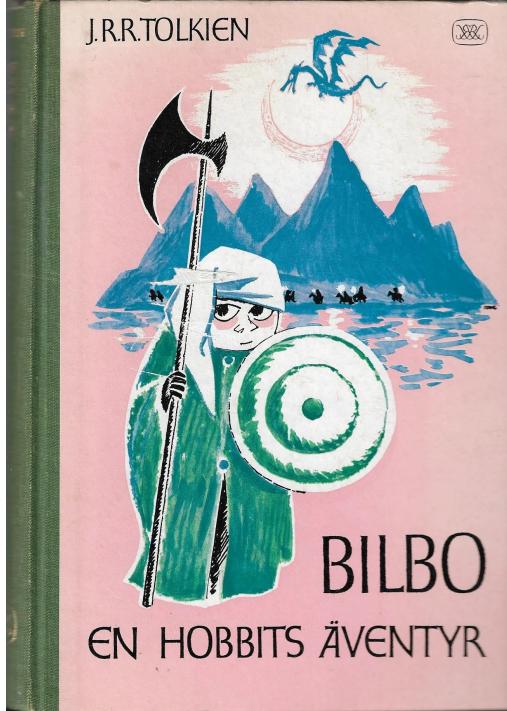
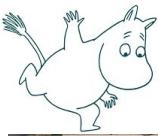
Assembly



Repetitive content makes the reconstruction more difficult

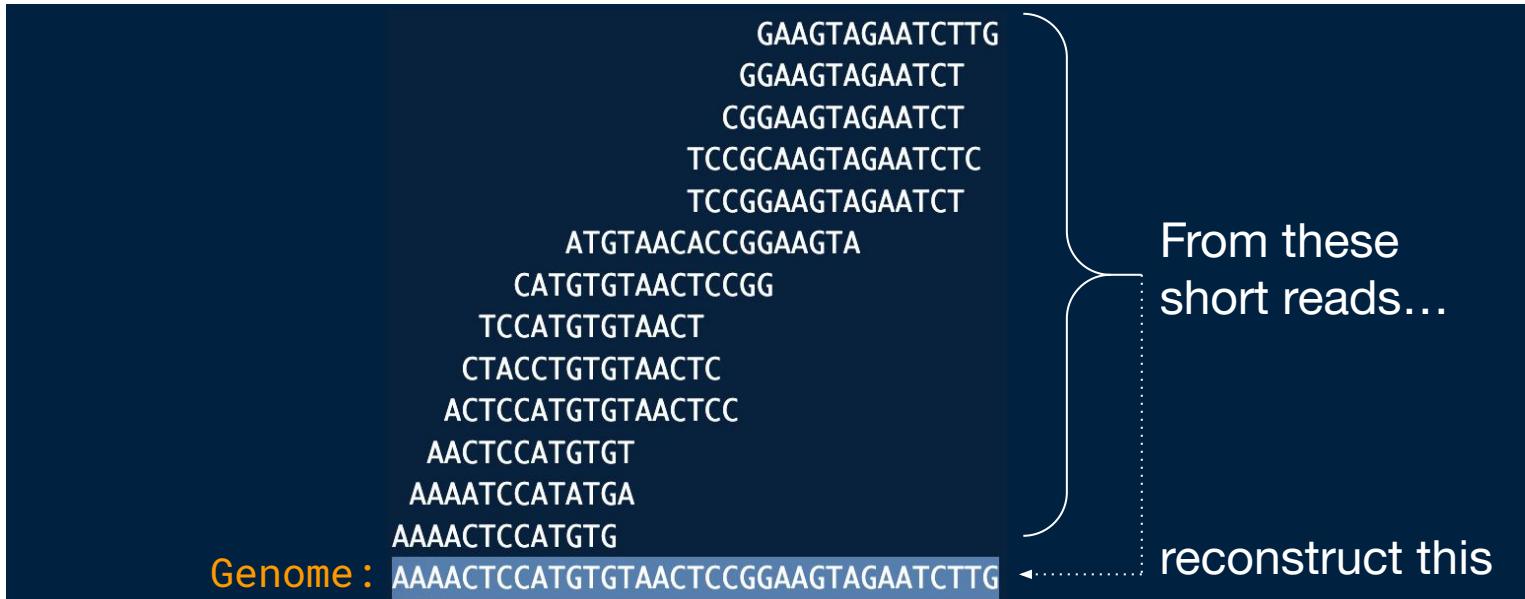


**Misprints or damaged fragments make reconstruction more difficult**



**Rare books are difficult or impossible to reconstruct**

# Back to DNA sequences...



GAAGTAGAATCTTG  
GGAAGTAGAATCT  
CGGAAGTAGAATCT  
TCCGCAAGTAGAATCTC  
TCCGGAAGTAGAATCT  
ATGTAACACCGGAAGTA  
CATGTGTAACTCCGG  
TCCATGTGTAACT  
CTACCTGTGTAACTC  
ACTCCATGTGTAACTCC  
AACTCCATGTGT  
AAAATCCATATGA  
AAAACTCCATGTG  
**Genome :** AAAACTCCATGTGTAACTCCGGAAAGTAGAATCTTG

From these short reads...

reconstruct this

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

Genome : AAAACTCCATGTGTAACCTCCGGAAAGTAGAATCTTG

AAAACTCCATGTG  
ATGTAACACCGGAAGTA  
CATGTGTAACCTCCGG  
ACTCCATGTGTAACCTCC  
TCCGGAAGTAGAATCT  
CTACCTGTGTAACTC  
AAAATCCATATGA  
AACTCCATGTGT  
GGAAGTAGAATCT  
TCCGCAAGTAGAATCTC  
TCCATGTGTAACT  
CGGAAGTAGAATCT  
GAAGTAGAATCTG

From these  
short reads...

reconstruct this

**Reality: Reads are scrambled ...**

AAAACCTCCATGTG  
ATGTAACACCGGAAGTA  
CATGTGTAACTCCGG  
ACTCCATGTGTAACTCC  
TCCGGAAGTAGAACATCT  
CTACCTGTGTAACTC  
AAAATCCATATGA  
AACTCCATGTGT  
GGAAGTAGAACATCT  
TCCGCAAGTAGAACATCTC  
TCCATGTGTAACT  
CGGAAGTAGAACATCT  
GAAGTAGAACATTG

Genome : ?????????????????????????????????? ←.....

From these  
short reads...

reconstruct this

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

**ONE DOES NOT SIMPLY**

**ASSEMBLE A GENOME**



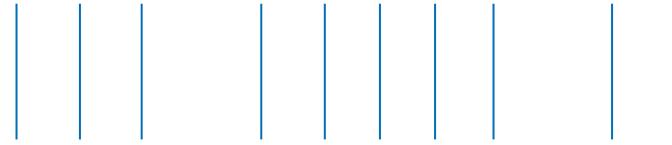
# How do we stitch together reads into contigs?

Read A: **AAA**ACTCCATGTG****

Read B: **AAA**ATCCATATGA****

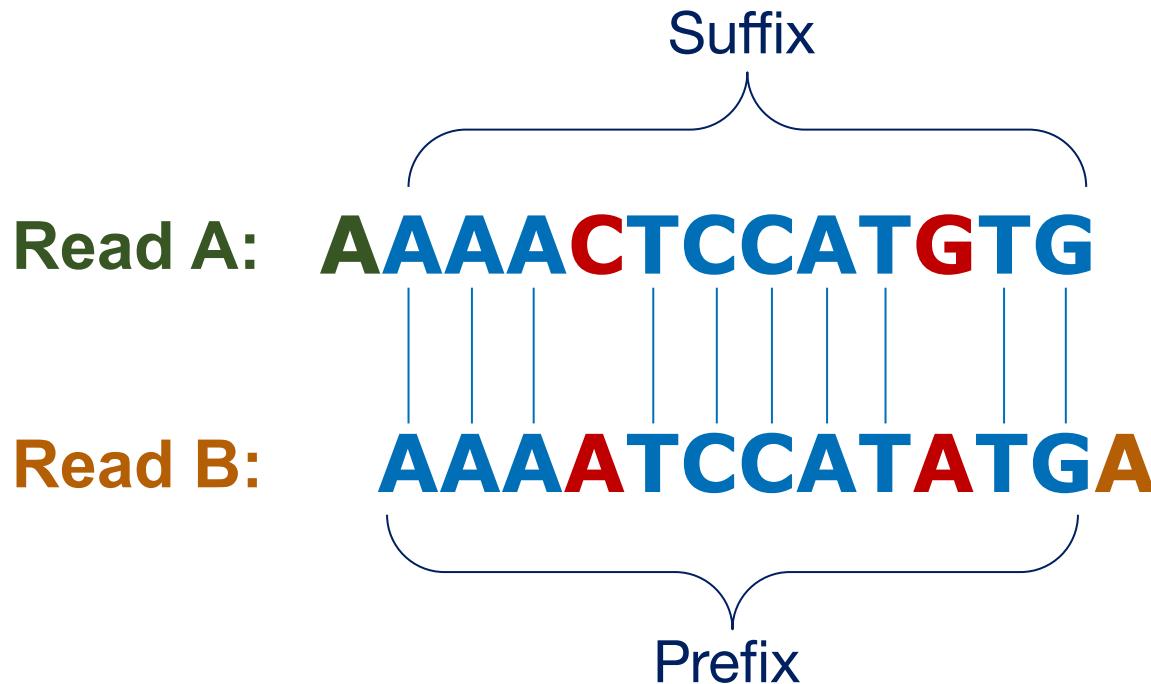
# How do we stitch together reads into contigs?

Read A: AAAA**C**TCC**A**T**G**TG



Read B: AAAA**A**TCC**C**AT**A**TGA

# 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

## Genome:

AAGTAGAATCTTG  
GGAAGTAGAATCTTG  
GGAAGTATAATCTTG  
CGGAAGTAGAAT  
CGGAAGTAGAATC  
TAACTACGGCAGTAGAG  
TGTAACTCCGGAAAGTAG  
TGTGTATCTCCC  
TGTGTAACTCCG  
CATGTGTAACTCCGG  
CTCCATGTGTAAAC  
ACTCCATGTGTAAAC  
AACTCCATGTGTAA  
AAAACACCATCTGA  
AAAACTCCATGT  
AAAACTCCATGTGTAACTCCGGAAAGTAGAATCTTG  
GGAAGTAGAATCTTG  
TAACTCAGGAAGTAG  
GTGTAACTCCGGA  
TCCATCTGTAACTCC  
AAAACACCATGTGT

## More coverage

## Less coverage

# More coverage leads to more and longer overlaps

Genome : AAAACTCCATGTGTAACCTCGGAAGTAGAATCTTG  
                  GGAAGTACAATCTTG  
                  TAACTCAGGAAGTAG  
                  GTGTAACCTCGGA  
                  TCCATCTGTAACTCC  
                  AAAACTCCATGTGT  
  
                  Coverage=3

AAGTAGAATCTTG  
GGAAGTACAATCTTG  
GGAAGTATAATCTTG  
CGGAAGTACAAT  
CGGAAGTACAATC  
TAACTCAGGAAGTAG  
TGTAACCTCGGAAGTAG  
TGTGTAACTCCCG  
TGTGTAACTCCG  
CATGTGTAACTCCGG  
CTCCATGTGTAAAC  
ACTCCATGTGTAAAC  
AACTCCATGTGTA  
AAAACACCACATCTGA  
AAAACTCCATGT  
  
                  Coverage=8

Average coverage:  
 $207/35 \approx 6\text{-fold}$

Average coverage:  
 $70/35 = 2\text{-fold}$

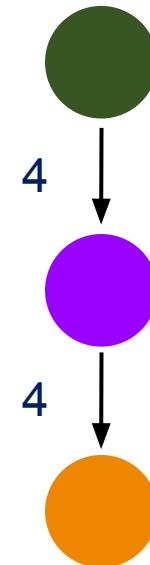
# Overlap graph

Read A: **ACGTA**

Read B: **CGTAC**

Read C: **GTACA**

Contig: **ACGTACA**



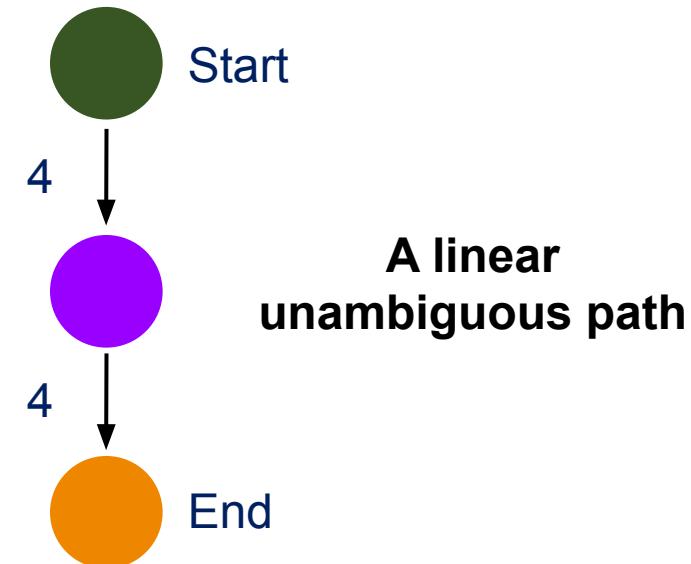
# Overlap graph

Read A: **ACGTA**

Read B: **CGTAC**

Read C: **GTACA**

Contig: **ACGTACA**



# Overlap graph

# Read A: ACGTA

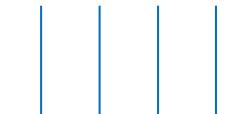
**Read B:** CGTAC

# Read C: CGTACA

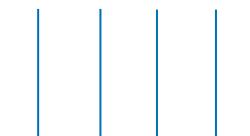
**Contig: ACGTACGTA (With repeats)**

# Overlap graph

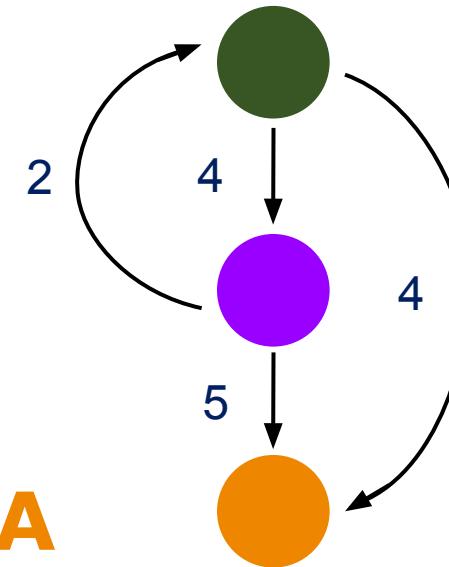
Read A: **A**CGTA****



Read B: ****CGTAC**C**



Read C: ****CGTACA****



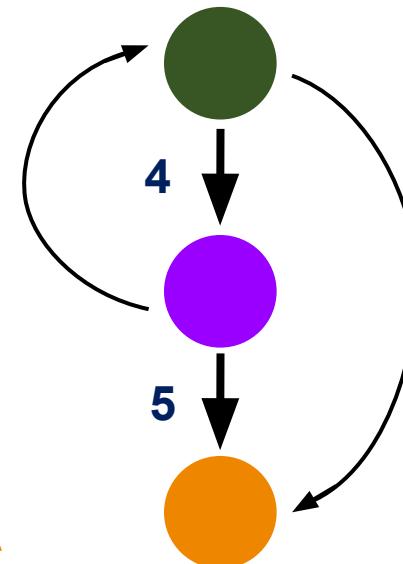
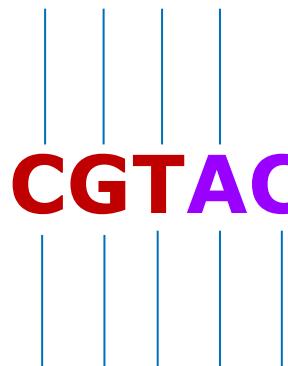
The introduced repeat "CGT" appears in multiple reads, causing branching paths in the overlap graph

# Overlap graph

Read A: **ACGTA**

Read B: **CGTAC**

Read C: **CGTACA**



Path along the  
edges with  
highest overlap

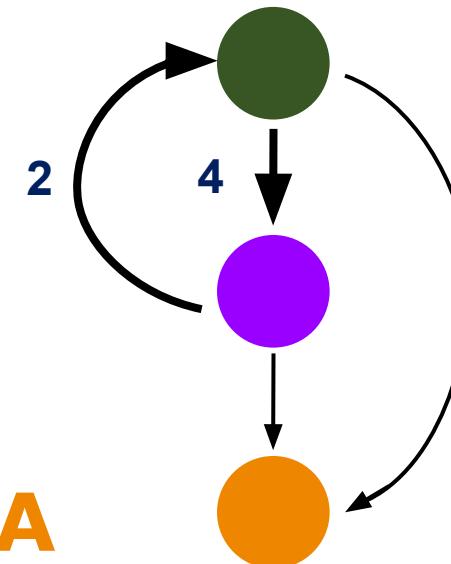
Not the original contig **ACGTACA**

# Overlap graph

Read A: **A**CGTA****

Read B: ****CGTAC**C**

Read C: ****CGTACA****



Alternative path

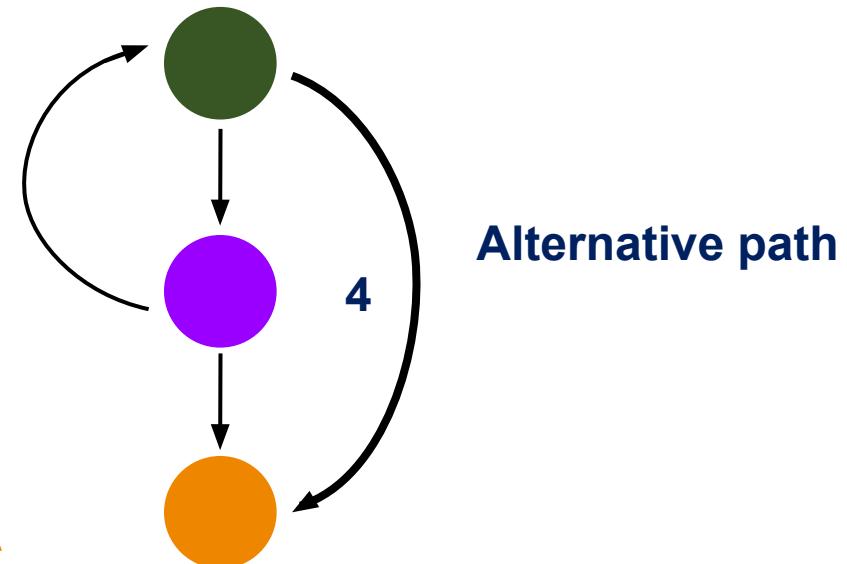
Original contig **ACGTACGTA**

# Overlap graph

Read A: **A**CGTA****

Read B: ****CGTAC**C**

Read C: ****CGTACA****



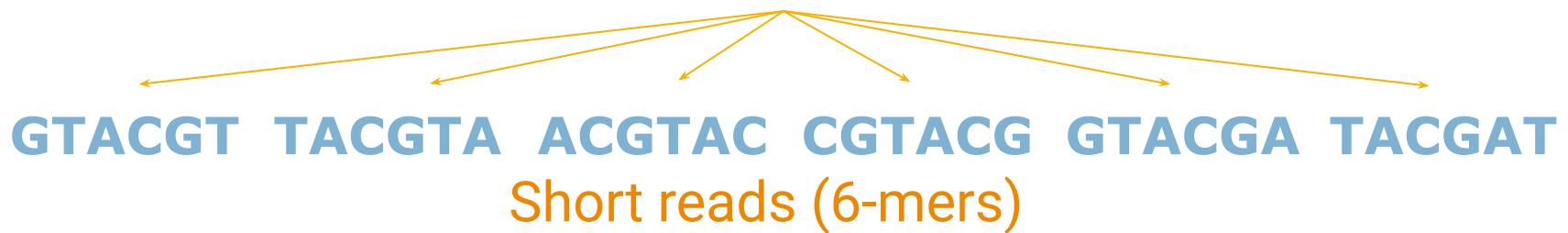
Original contig **ACGTACGTA**

## Third Law of Assembly

*Repeats make assembly difficult*

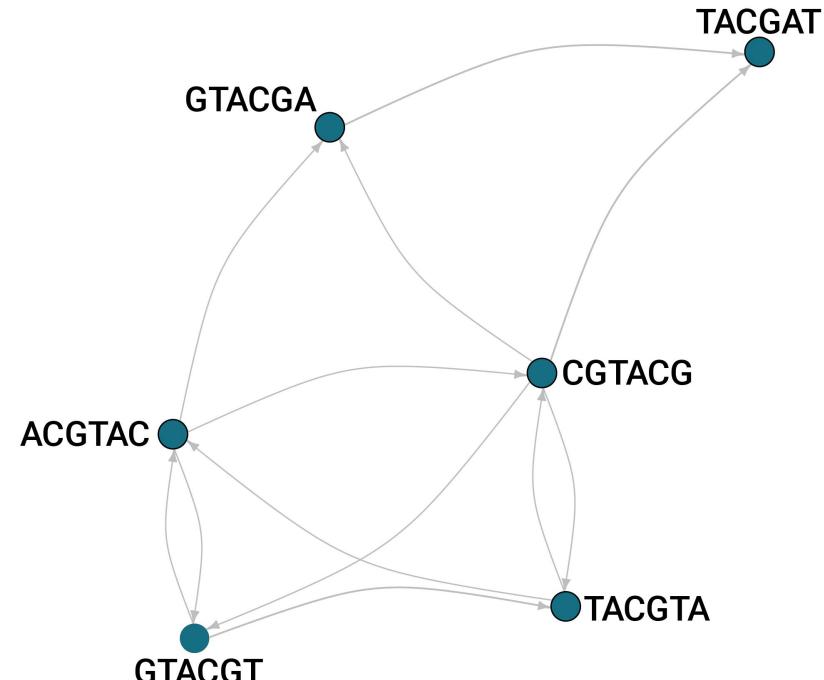
# Overlap graph

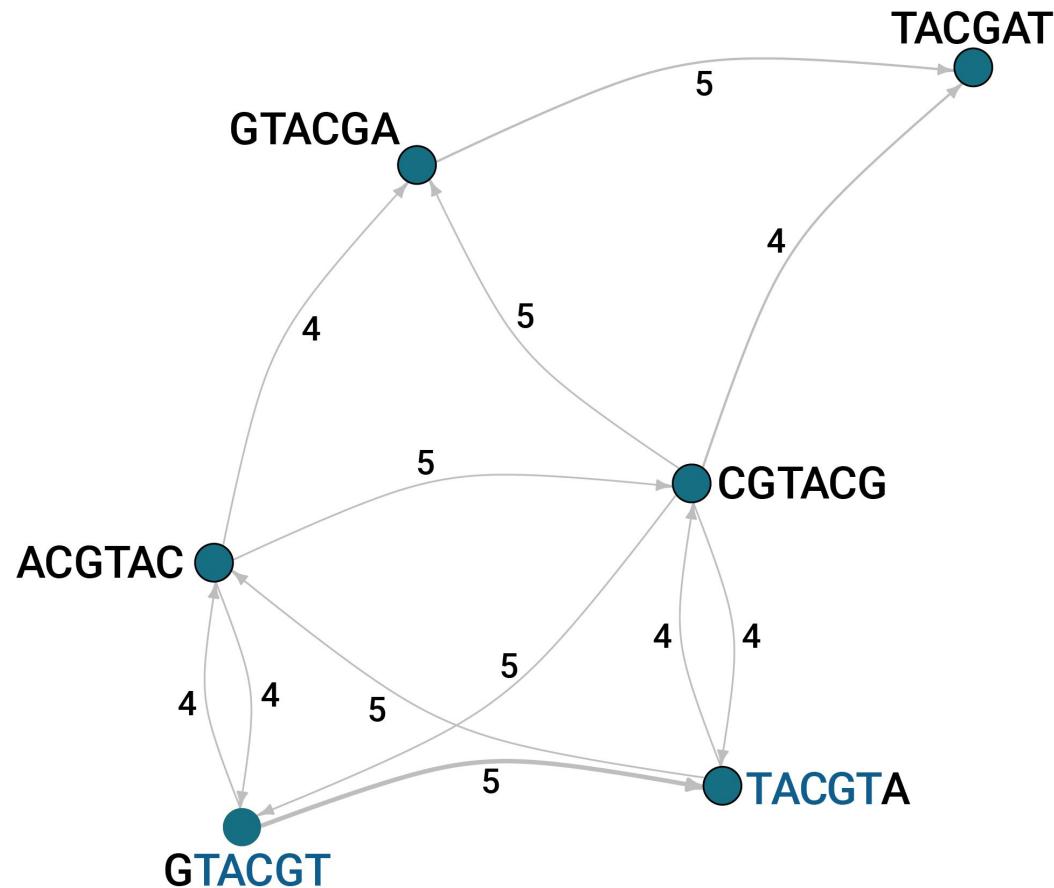
Original contig **GTACGTACGAT**

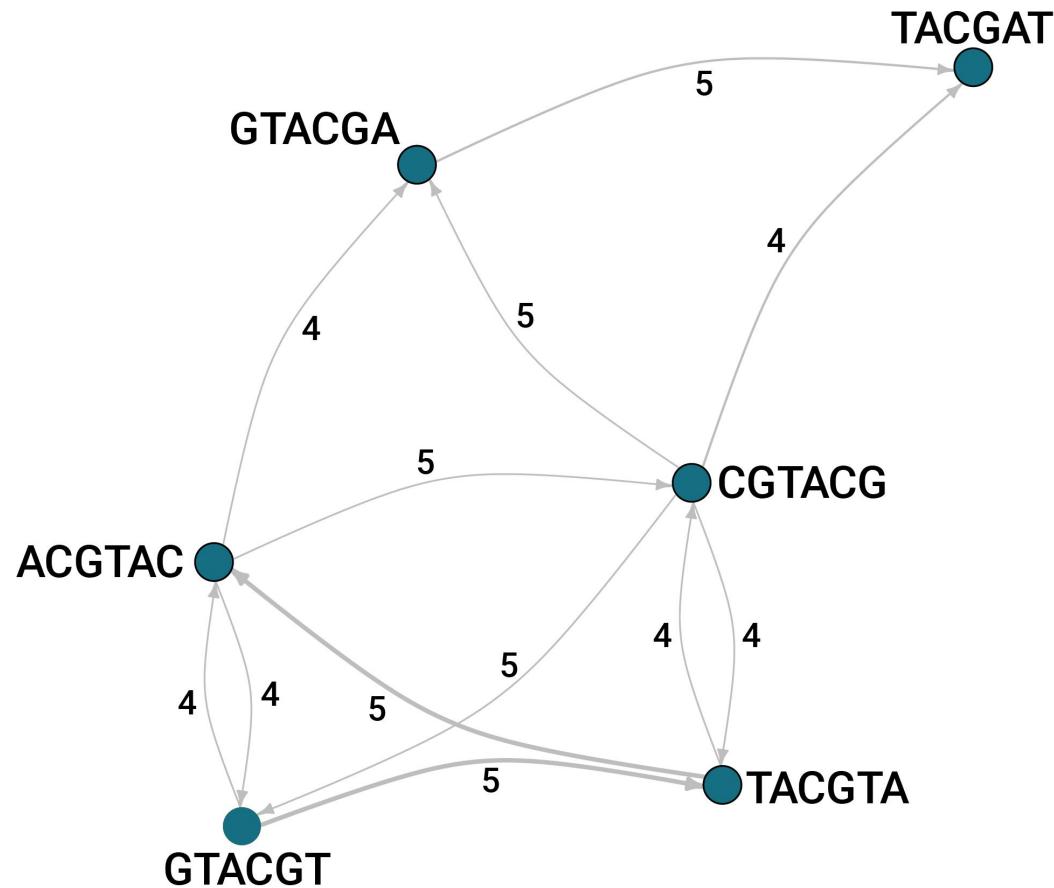


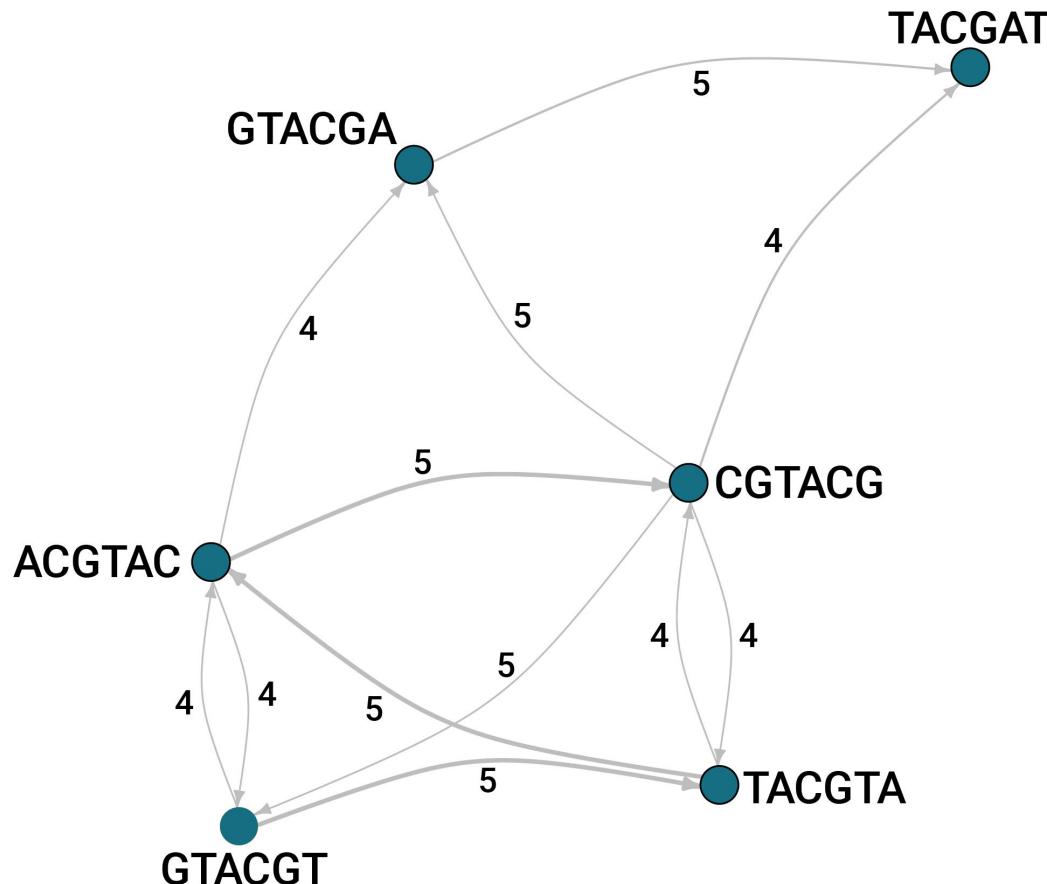
# Overlap graph

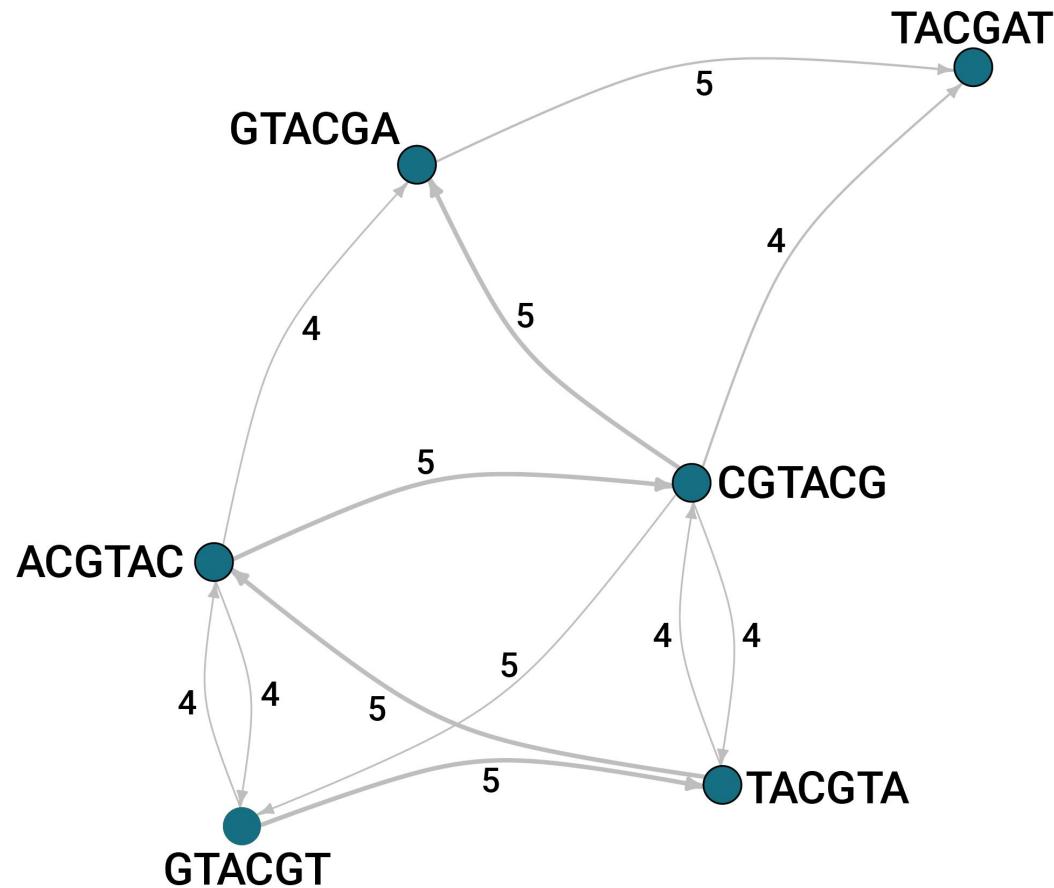
- Each read is a node
- Draw an edge between A and B if suffix of A overlaps with prefix of B
- Contigs are reconstructed by walking along unambiguous paths
- Remove cycles, and at branching paths continue on the edge with the highest overlap

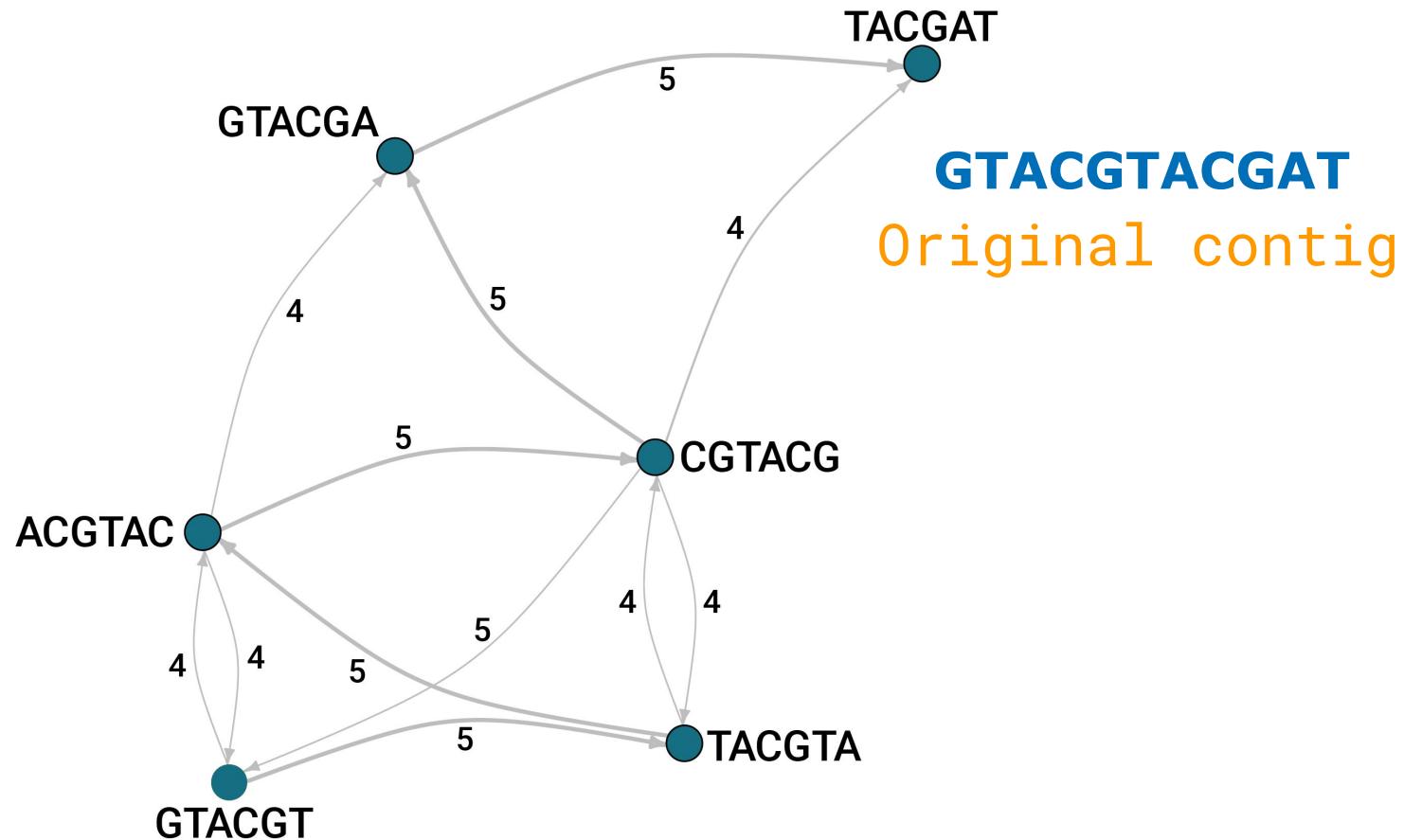


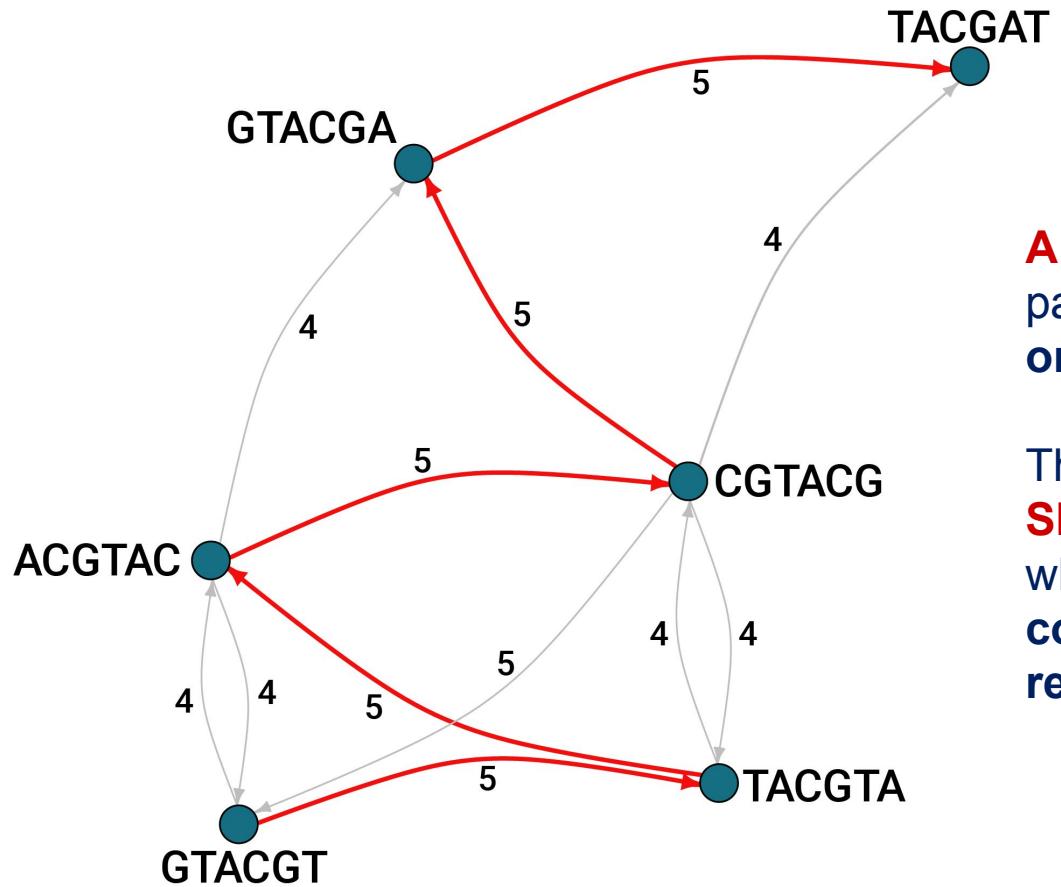










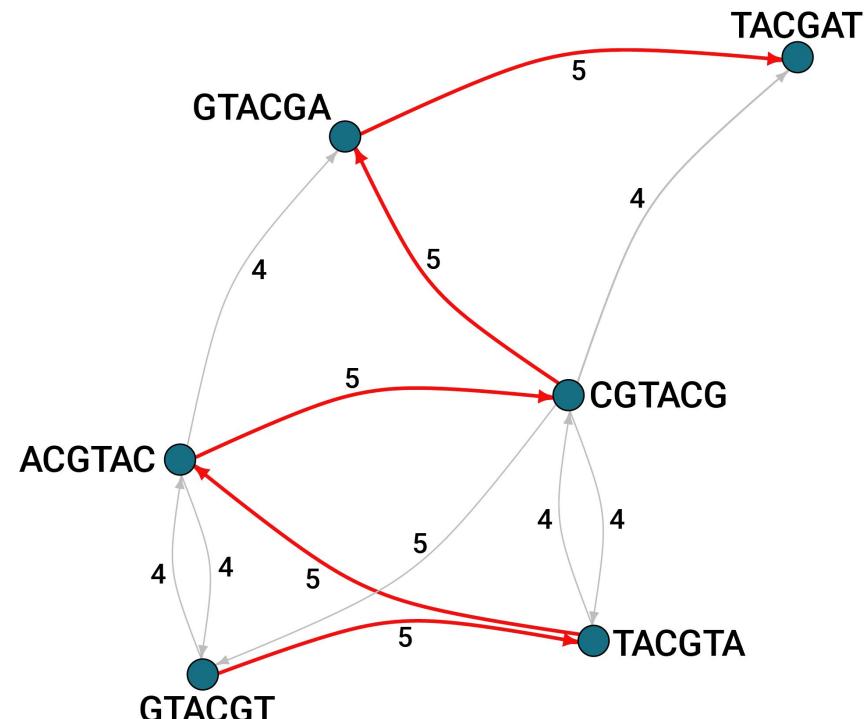


**A Hamiltonian path** in a graph is a path that visits each **node exactly once**

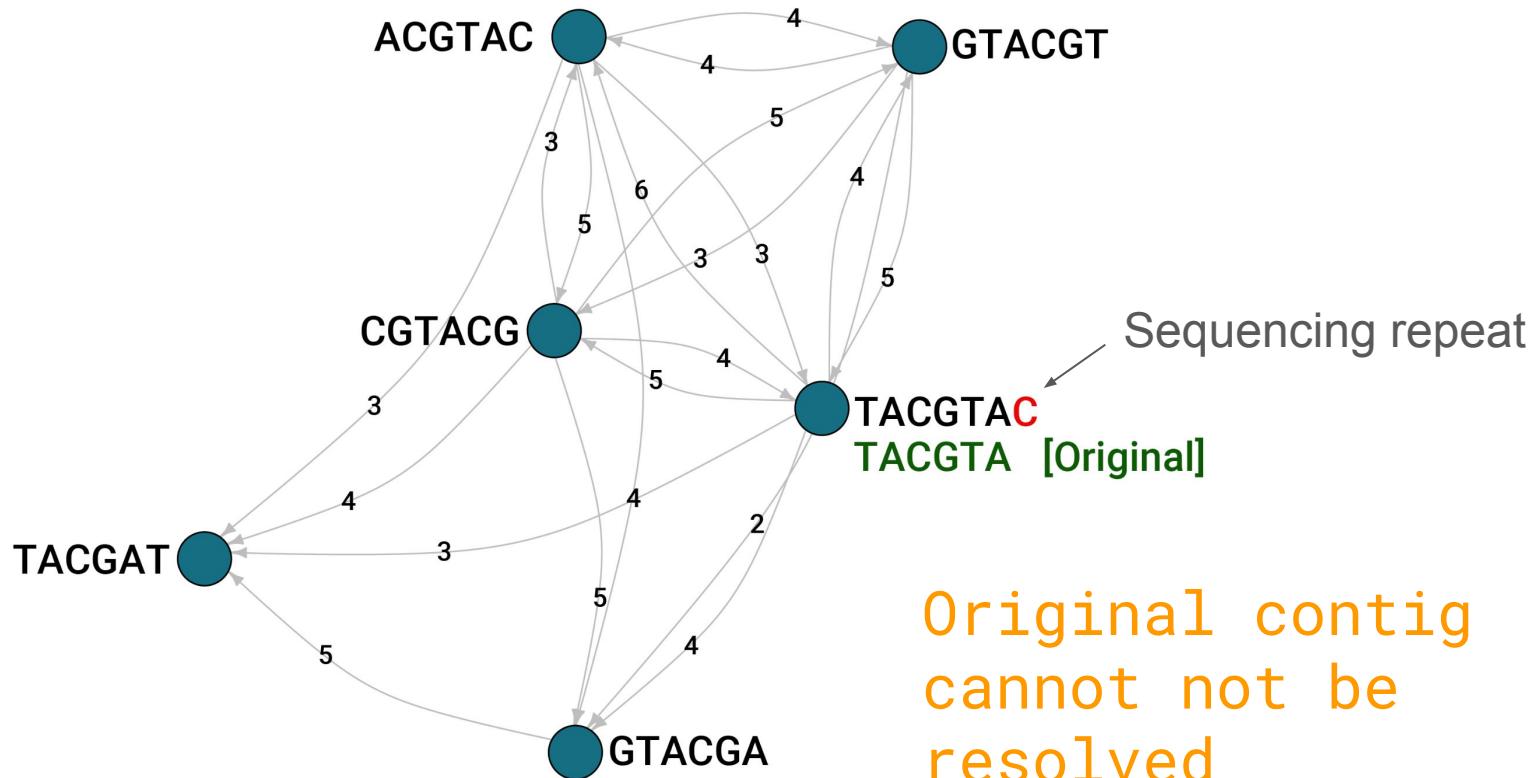
This path (in this case) is also the **Shortest Common Superstring** which **represents the most compact way to cover all the reads, minimizing redundancy**

# 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

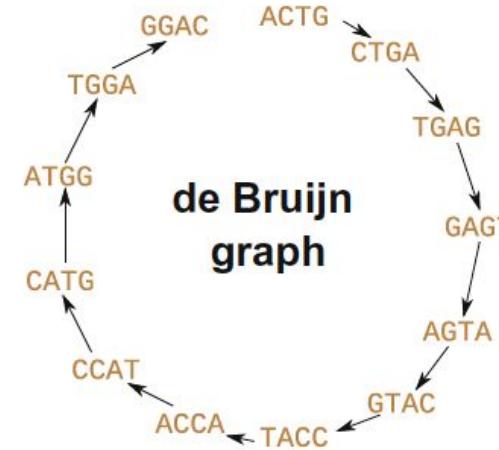


# Modern short-read assemblers use the de Bruijn graph

Reference genome ACTGAGTACCATGGAC

Sequenced reads ACTGAGTAC  
CTGAGTACCAT  
GAGTACCATGGAC

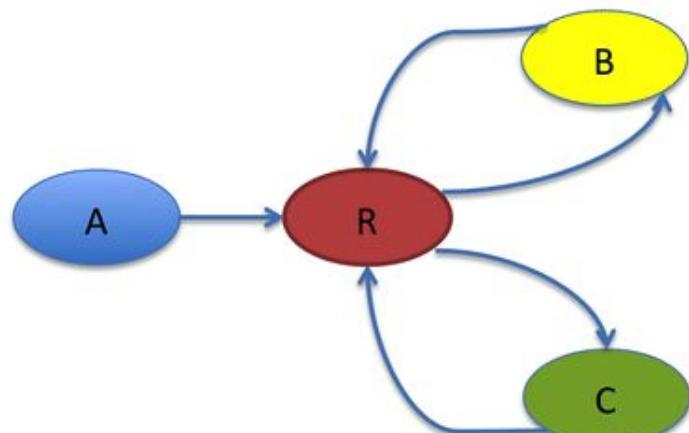
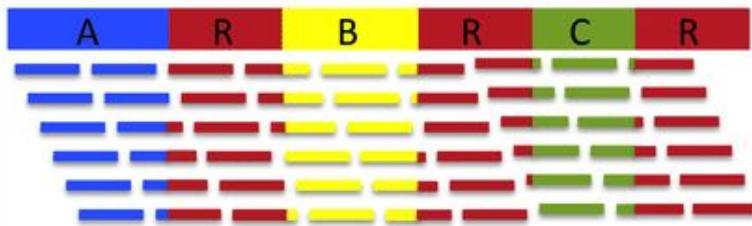
k-mers ACTG TACC GGAC  
CTGA ACCA  
TGAG CCAT  
GAGT CATG  
AGTA ATGG  
GTAC TGGA



Scales linearly instead of quadratically

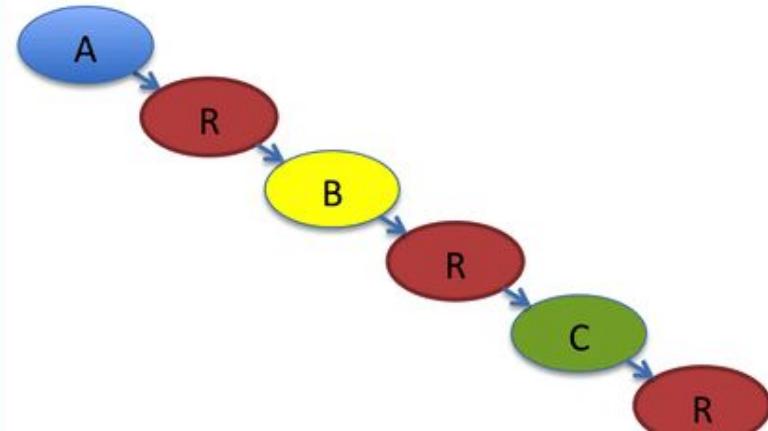
## Short Read Assembly

(read length < repeat length)



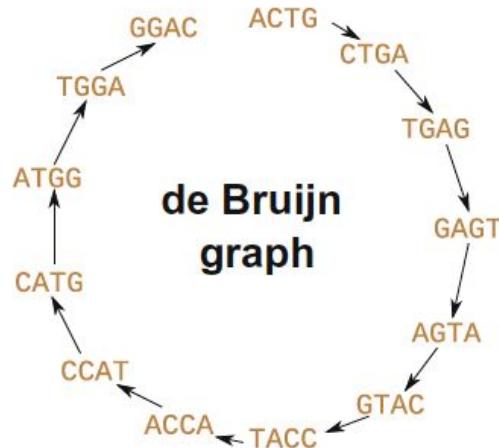
## Long Read Assembly

(read length > repeat length)



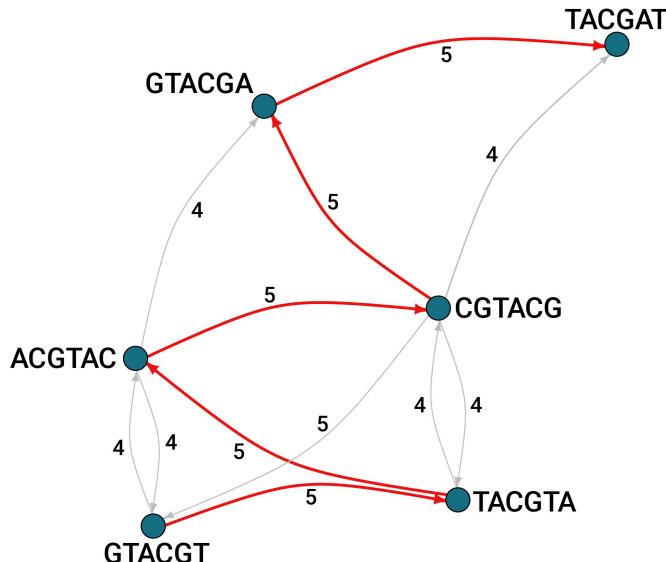
## Short Read Assembly

(read length < repeat length)



## Long Read Assembly

(read length > repeat length)

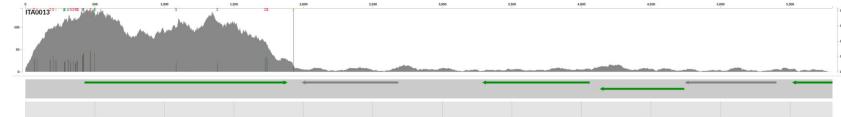


Overlap-Layout-Consensus

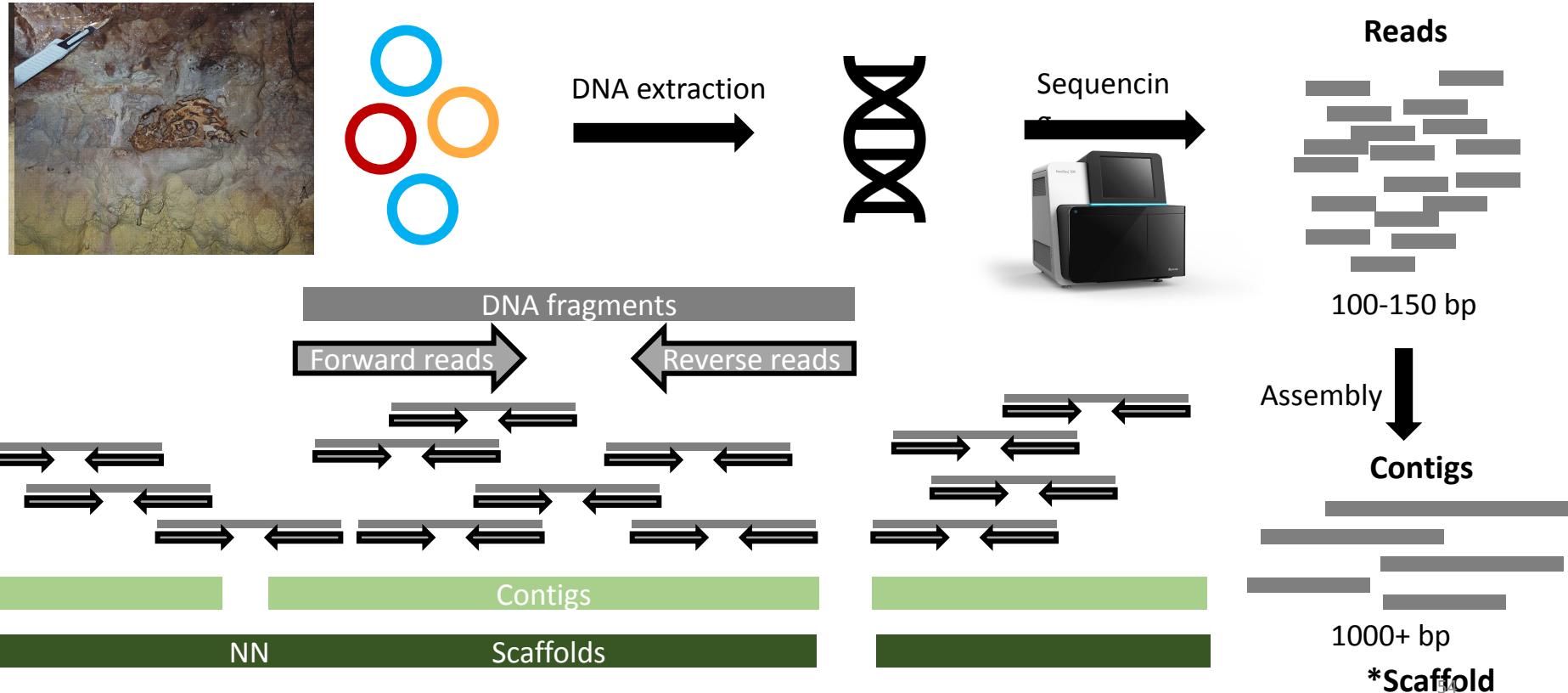
# Assessing assemblies

Quast

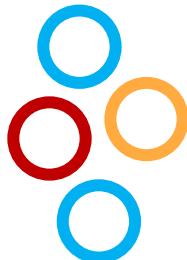
- **Number** of contigs
- **Average/median** contig length
- **Min/Max** contig length
- **N50:** The length of the contigs which covers 50% of genome
- **Coverage:** The number of reads that align to a contig and can be used as an indicator of confidence in the contig
- **Read recruitment:** Percentage of all reads mapped back to the assembly
- **Evenness in depth along contig.**



# Metagenomics workflow: Coverage



# Metagenomics workflow: Coverage



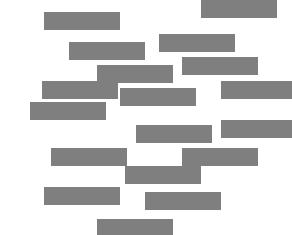
DNA extraction



Sequencing



Reads



100-150  
bp

Mapping

Taxonomy

Function

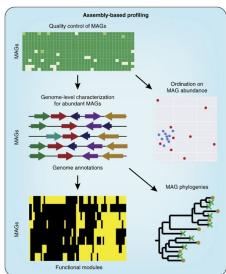
Genome

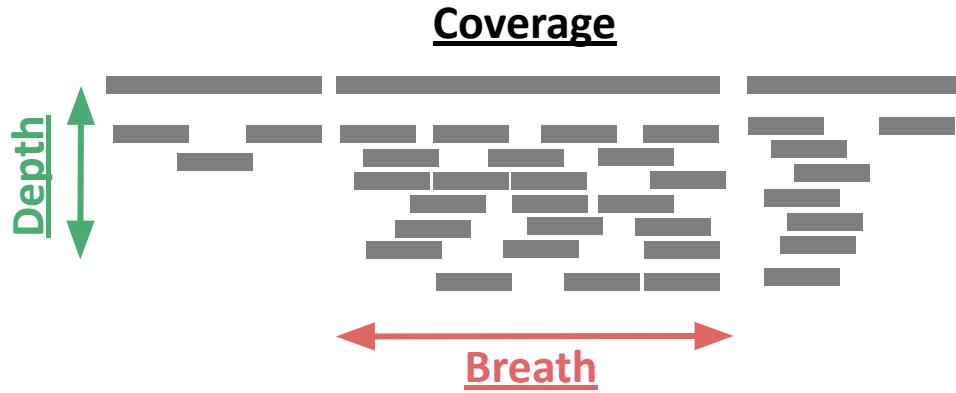
Assembly

Scaffolds

1000+ bp

e.g.  
Bowtie2  
BWA  
SAMtools  
...





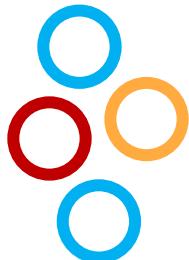
### Depth of coverage (mapping depth)

- 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



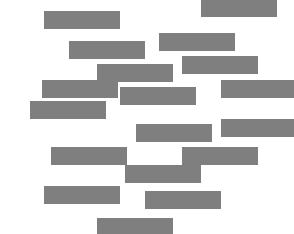
DNA extraction



Sequencing



Reads



100-150 bp

Mapping

Taxonomy

Function

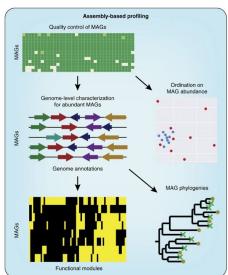
Genome

Assembly

Scaffolds

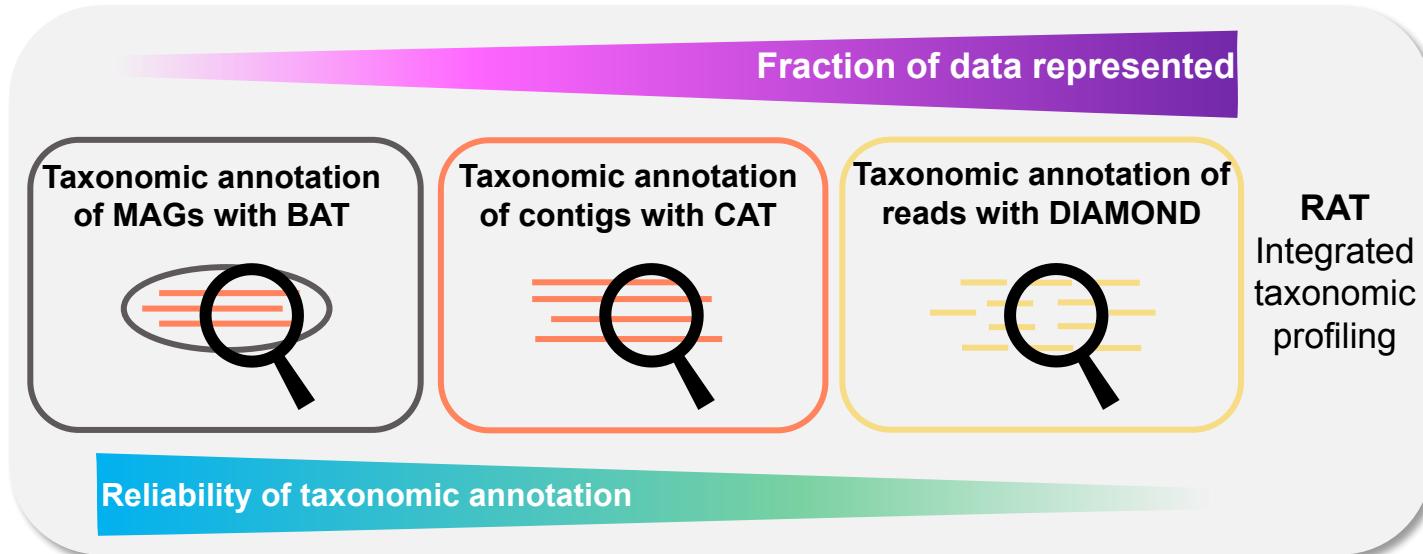
1000+ bp

Stay tuned: Tuesday October 15th





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

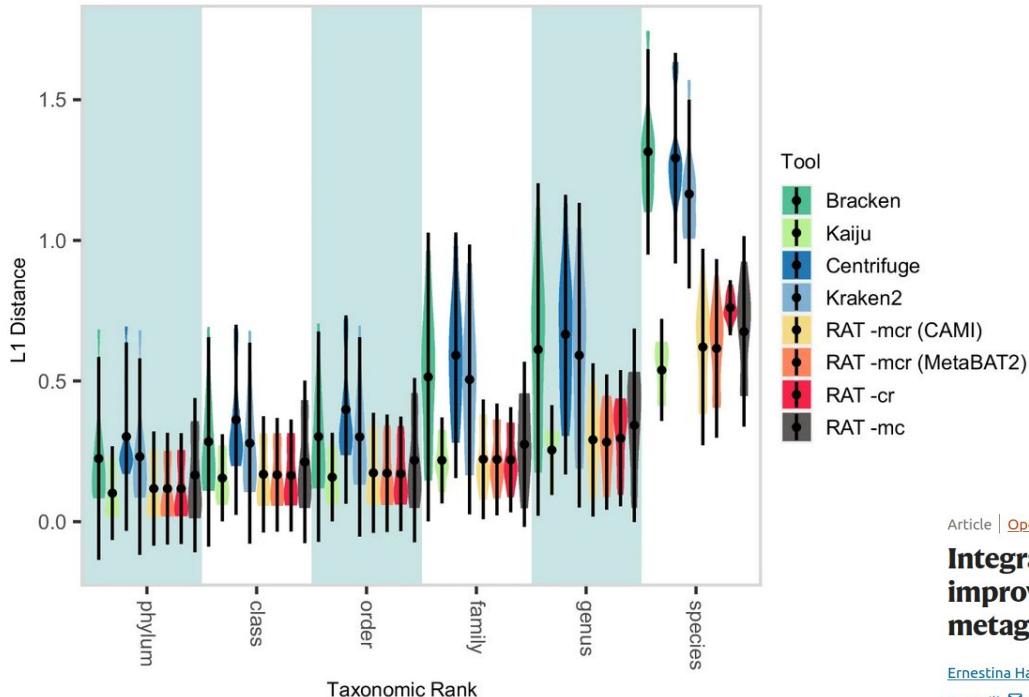


+ Function

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



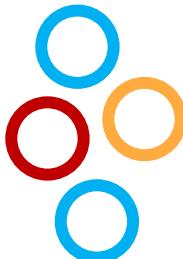
Article | [Open access](#) | Published: 20 April 2024

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

[Ernestina Hauptfeld](#), [Nikolaos Pappas](#), [Sandra van Iwaarden](#), [Bastien L. Snoek](#), [Andrea Aldas-Vargas](#), [Bas E. Dutilh](#) & [F. A. Bastiaan von Meijenfeldt](#)

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

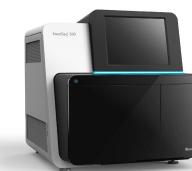
# Metagenomics workflow: Function



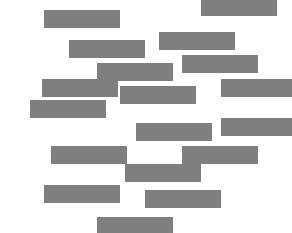
DNA extraction



Sequencing



Reads



100-150 bp

Assembly

Mapping

Taxonom

Function

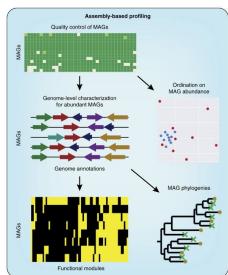
Genome

↓

Scaffolds

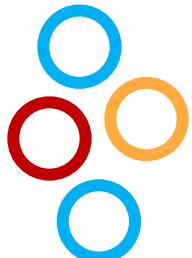
↓

1000+ bp



Stay tuned: Tuesday October 15th

# Metagenomics workflow: Assembled-based analysis



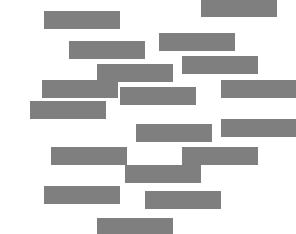
DNA extraction



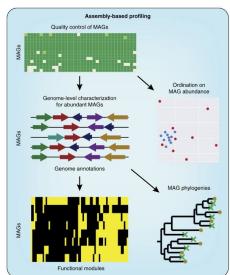
Sequencing



Reads



100-150 bp



Assembled-based

Mapping

Taxonom

Function

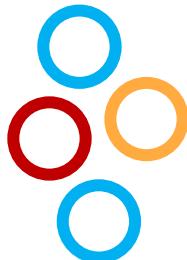
Genome

Assembly

Scaffolds

1000+ bp

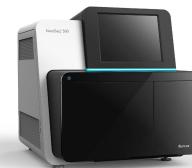
# Metagenomics workflow: Binning



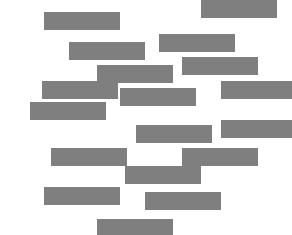
DNA extraction



Sequencing



Reads



100-150 bp

Assembly

Mapping

Taxonom

Function

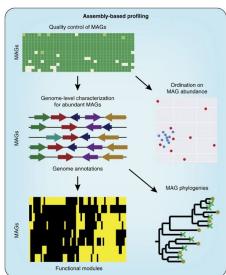
Genome

↓

Scaffolds



1000+ bp



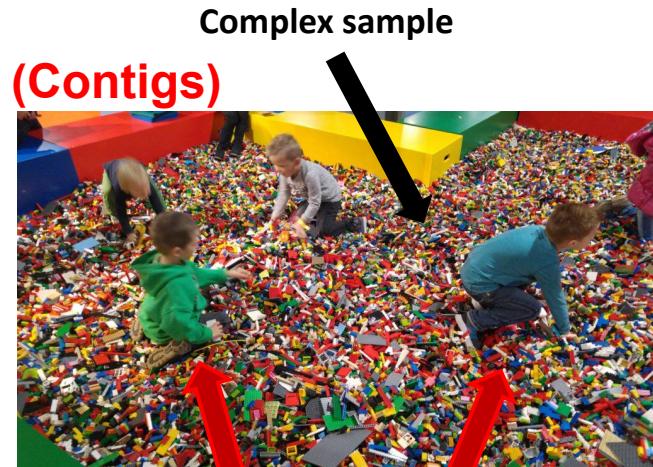
# Binning = Separation of genomes from metagenomes

- Who is there and what can every individual do?

**PI:** How difficult would it be?



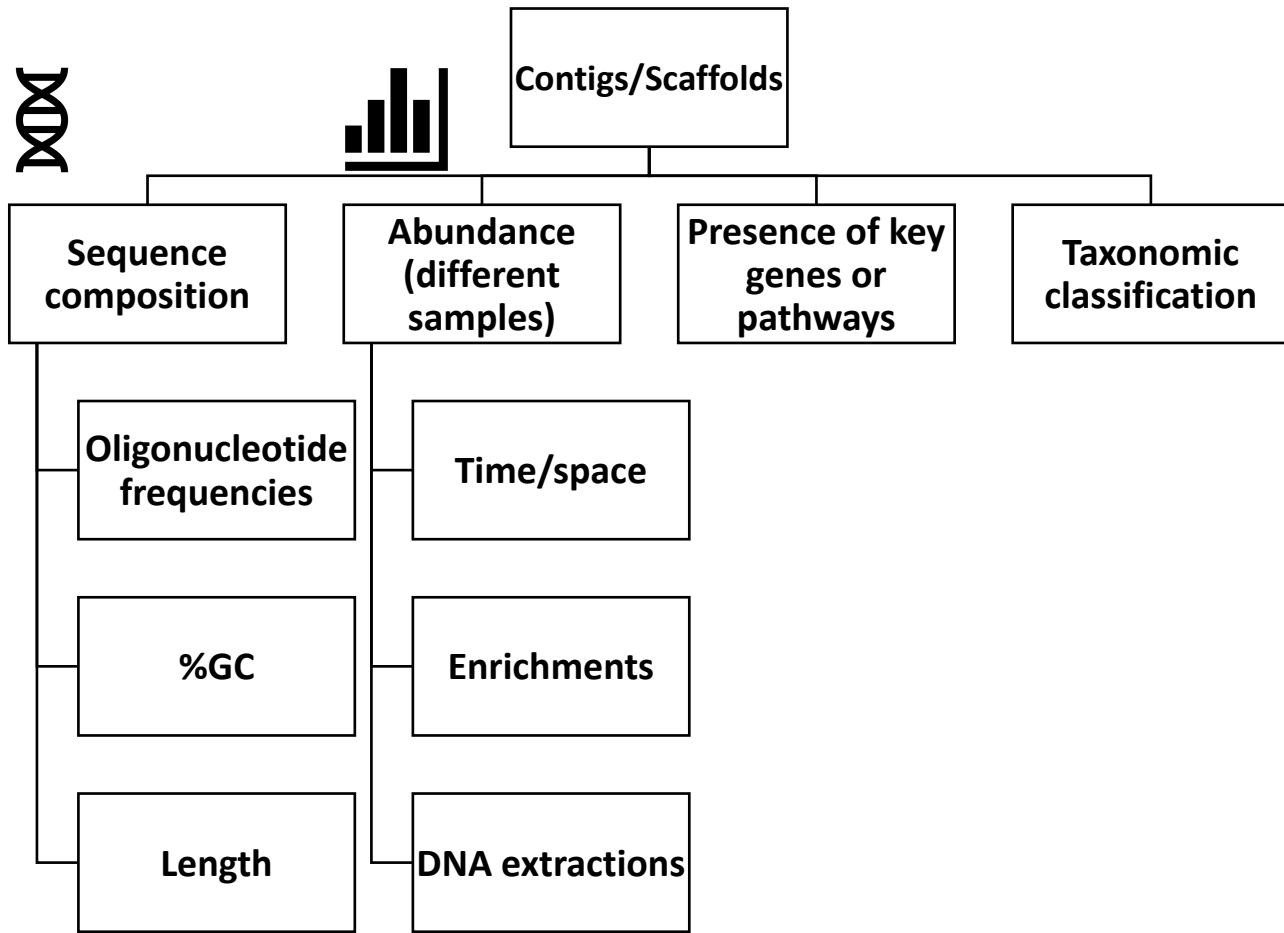
**PostDoc:** It's challenging yet fun, and there are plenty of standardized methods available to help!

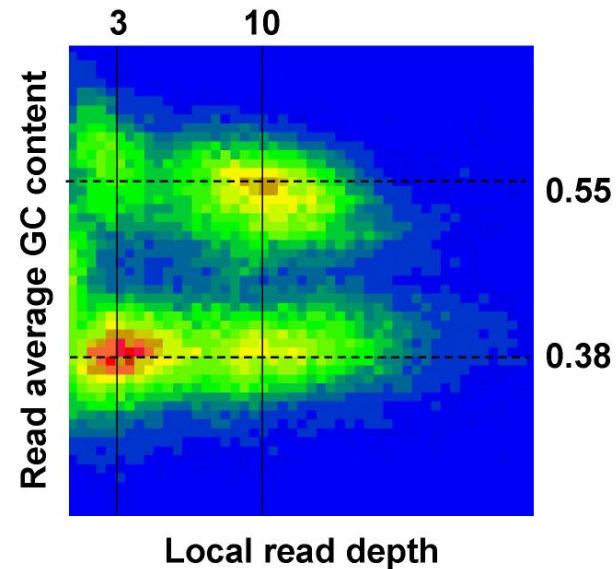
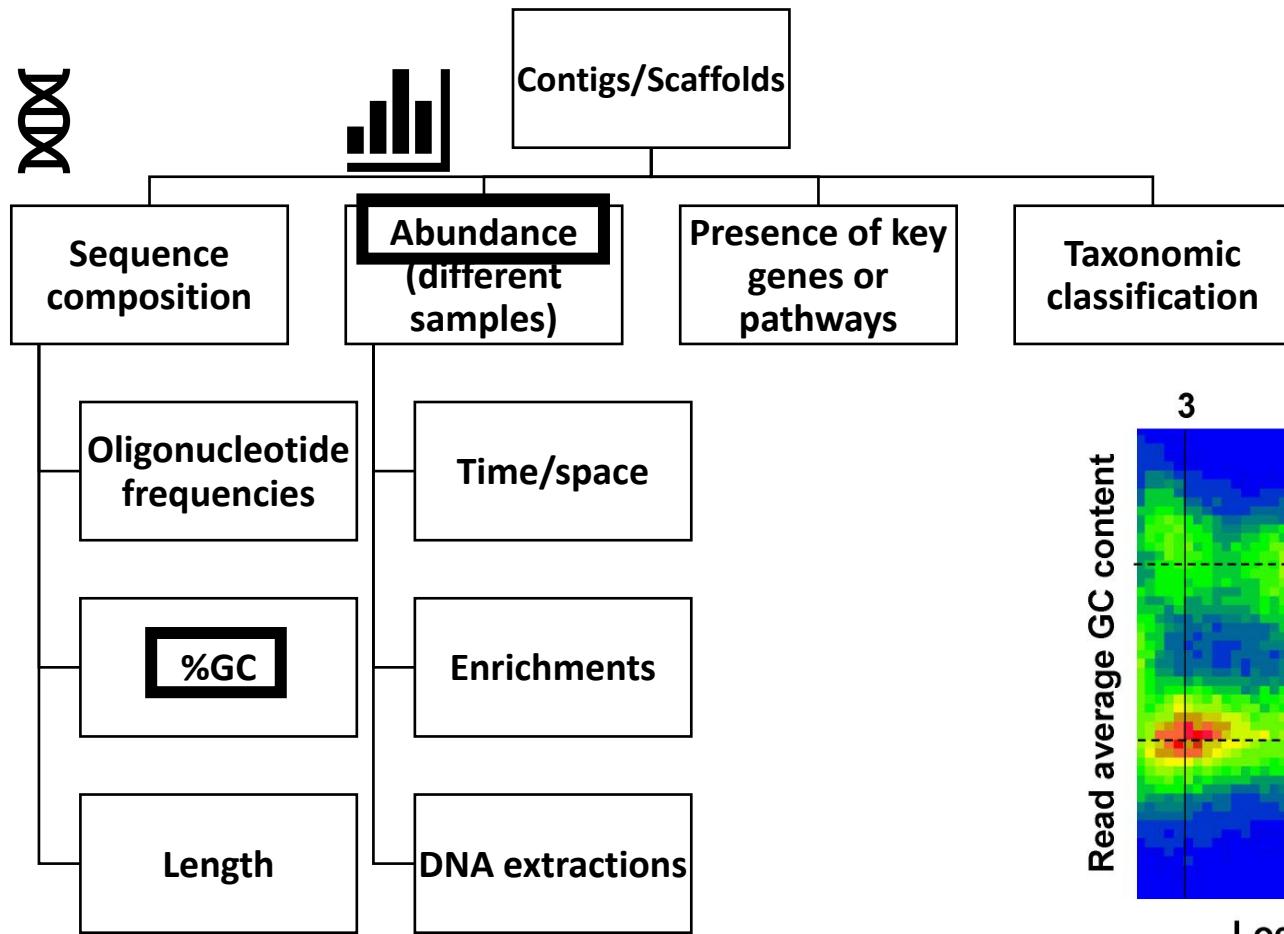


**PhD candidates &  
MSc interns**

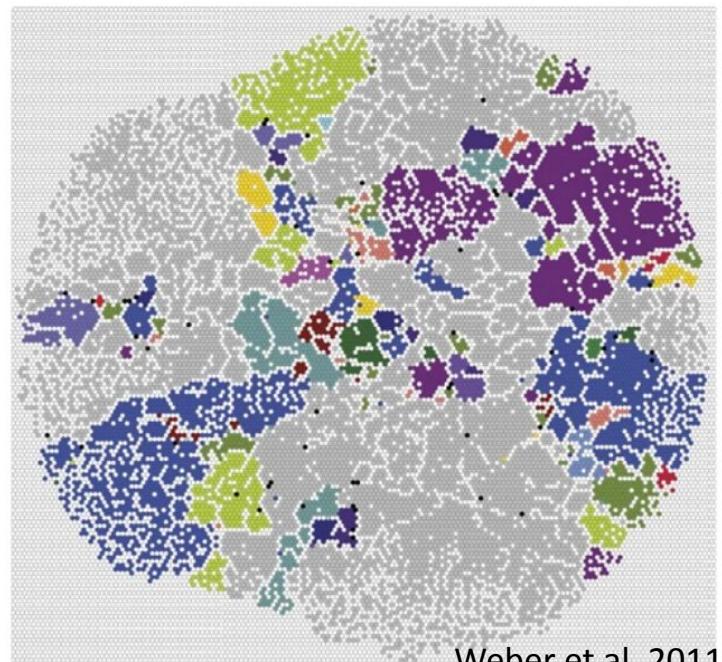
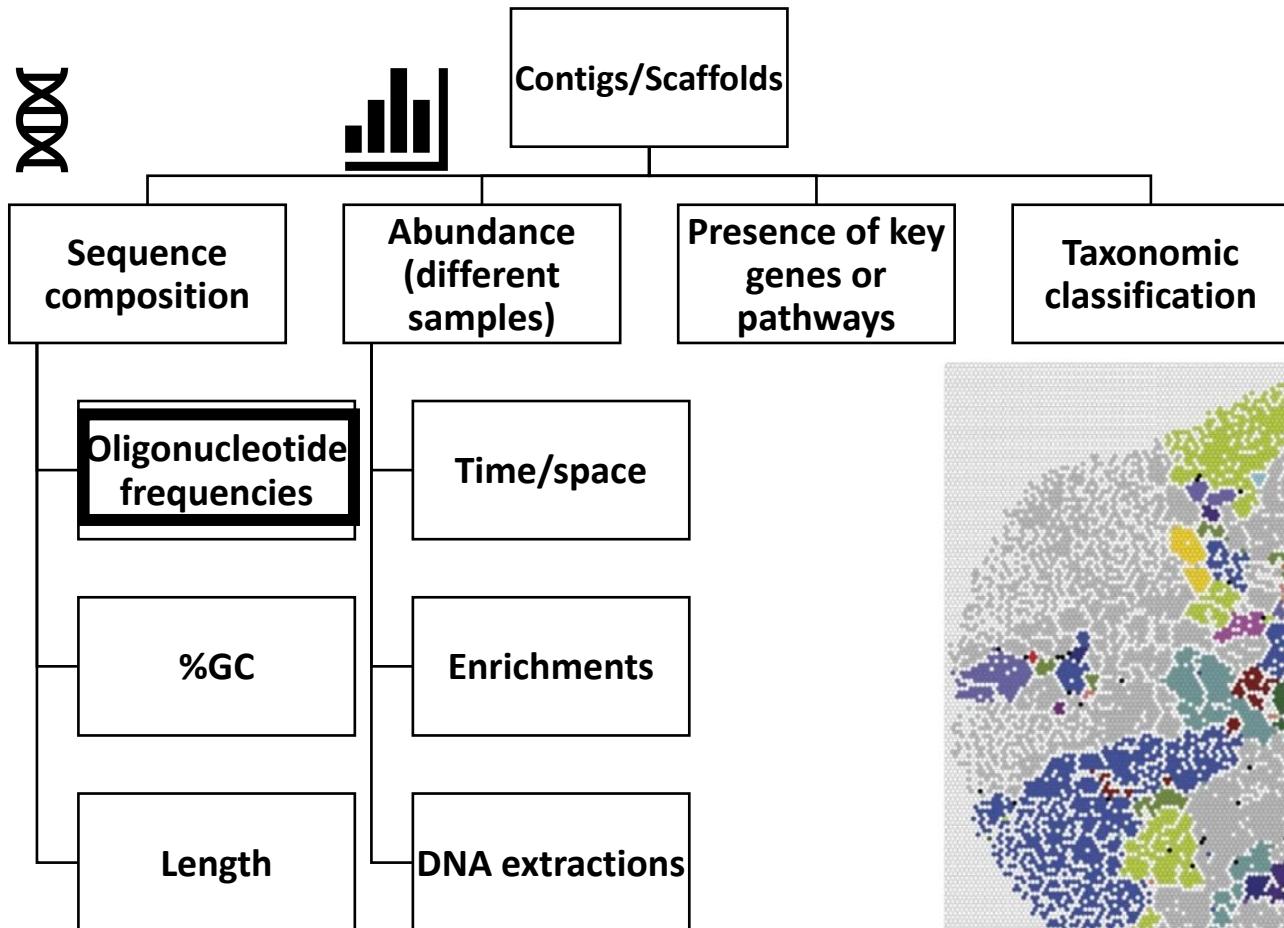
Binning  
→

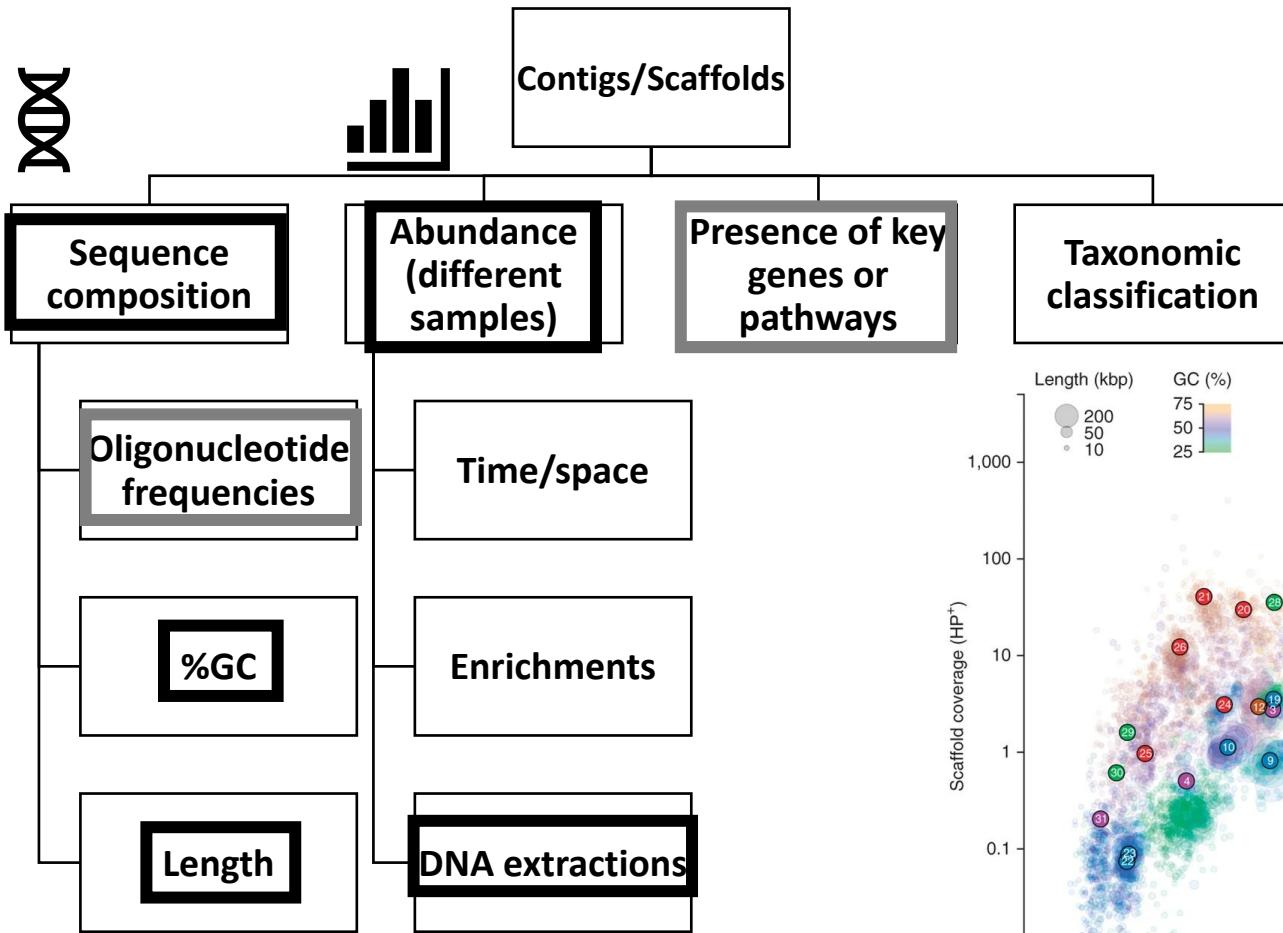




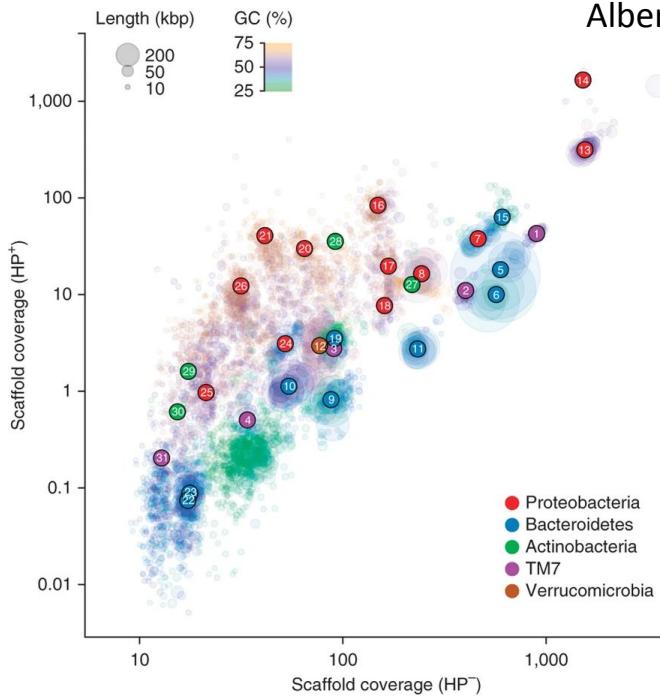


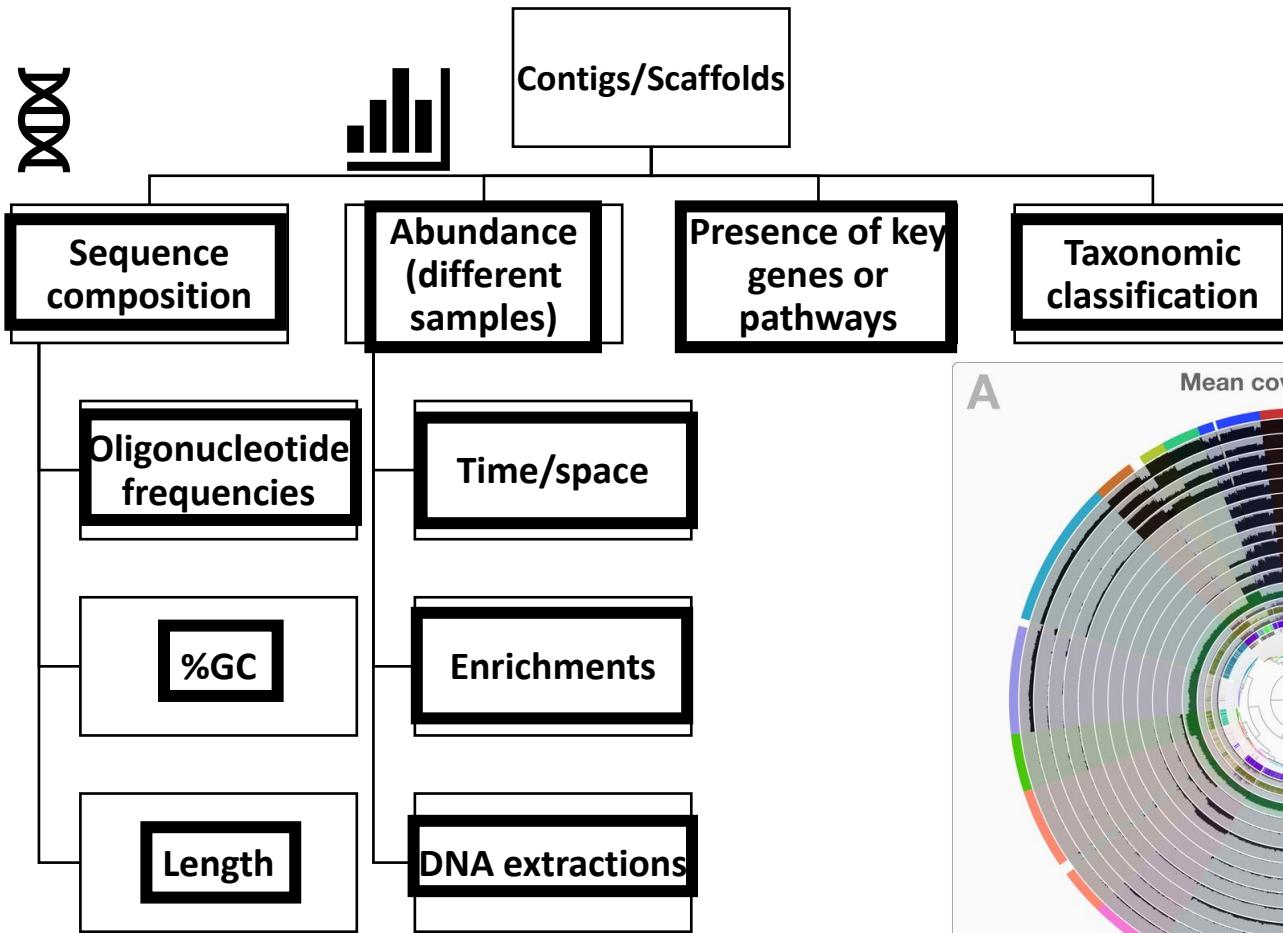
Tyson et al. 2004



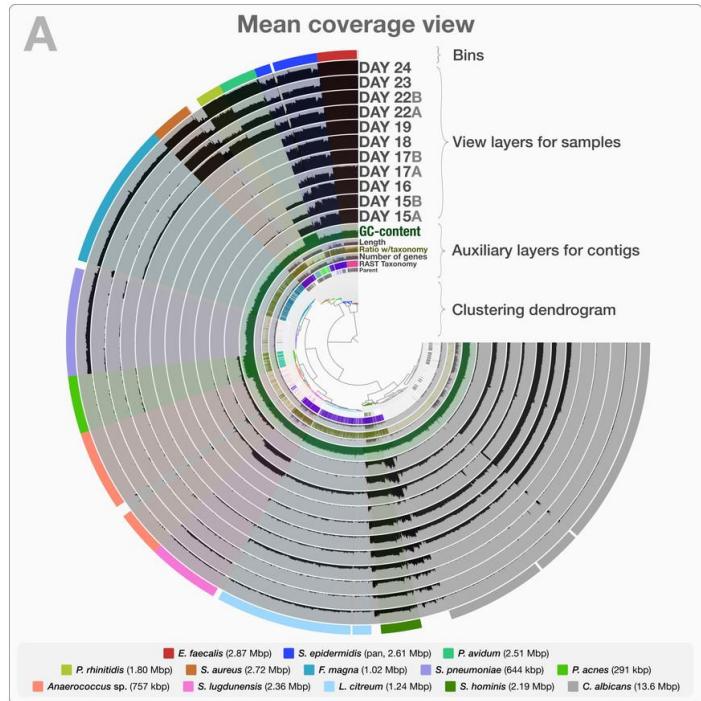


Albertsen et.al, 2013





Eren et.al, 2015



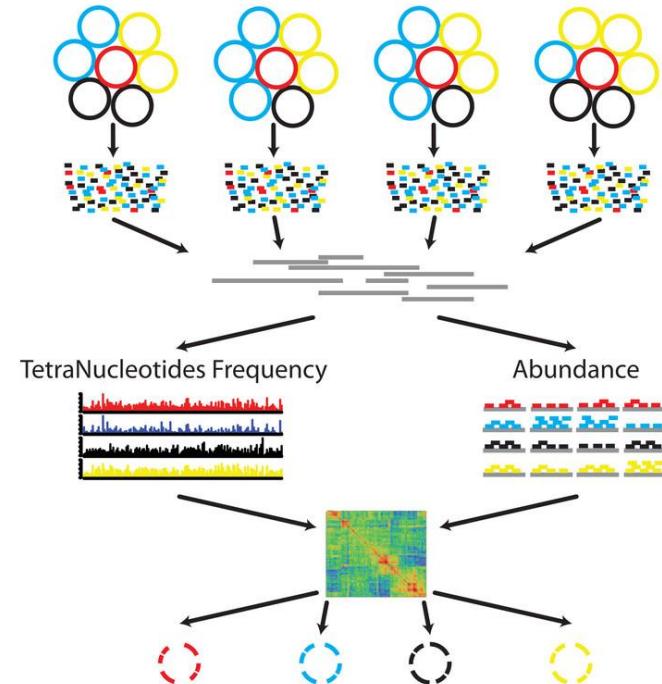
# Automatic tools for binning

Tools:

MetaBAT2  
Maxbin2  
CONCOCT

Aggregate multiple binning results:  
e.i. DASTool

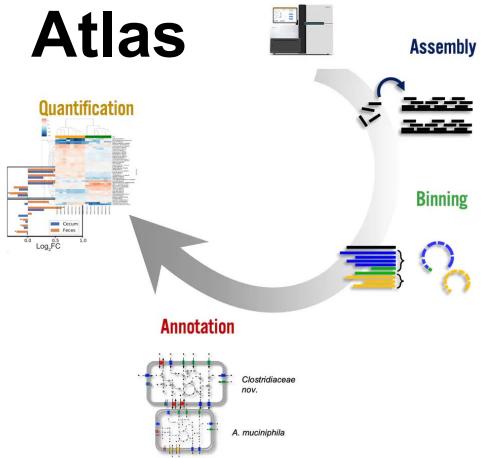
New kids in the playground:  
Vamb  
SemiBin  
MetaDecoder



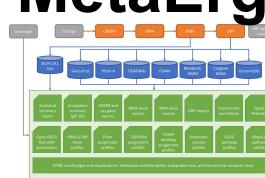
Overview of the MetaBAT pipeline.

# (Semi)Automatic pipelines

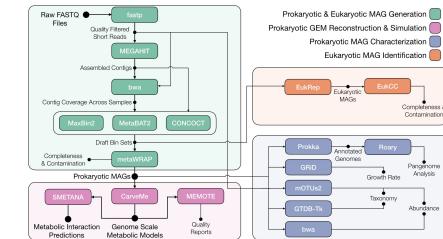
## Atlas



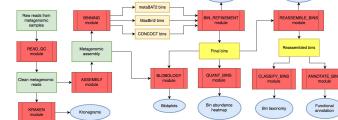
## MetaErg



## metaGEM

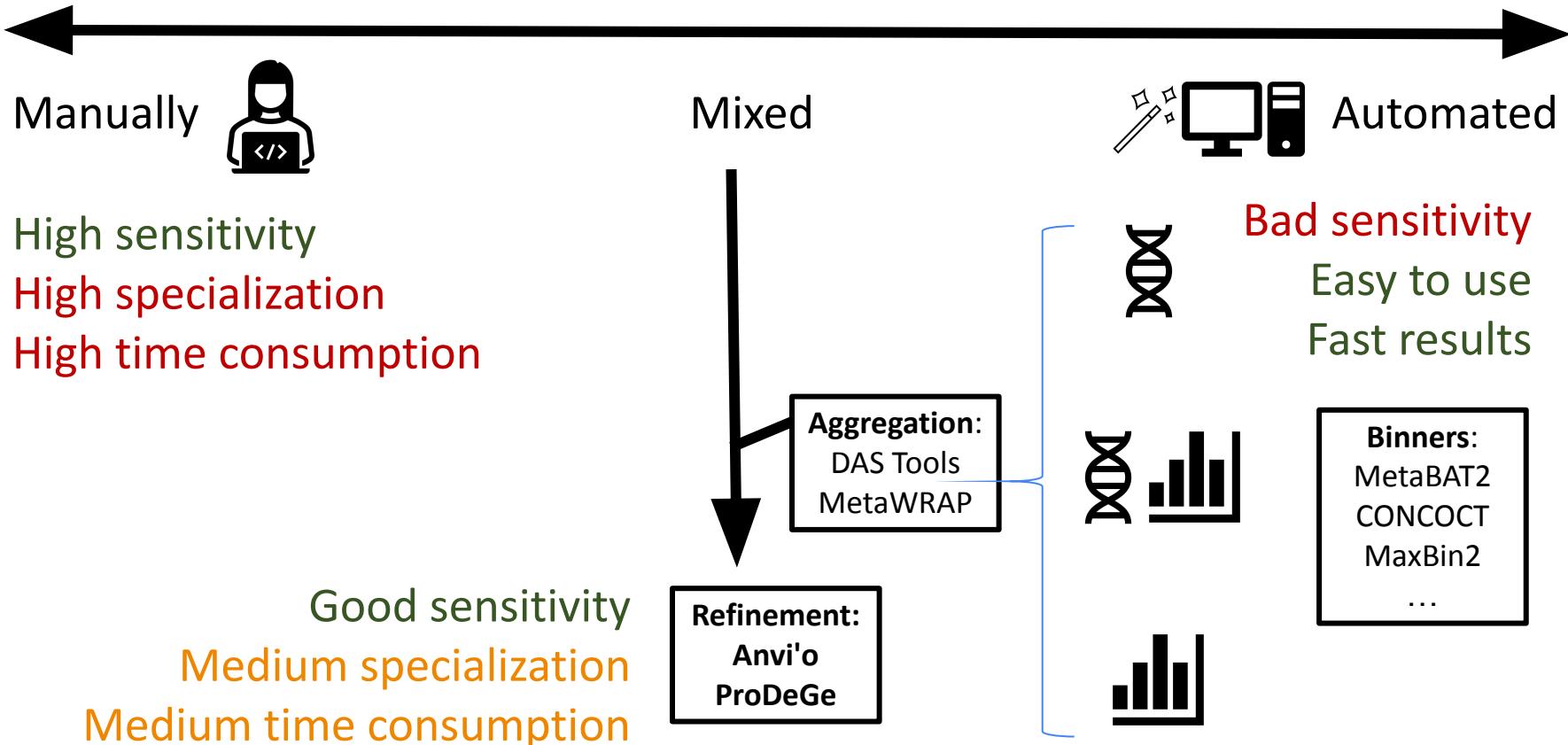


## MetaWRAP



>Standardization of metagenomics is (not) required!

# In practice



# MAG quality assessment

- Single-copy marker genes

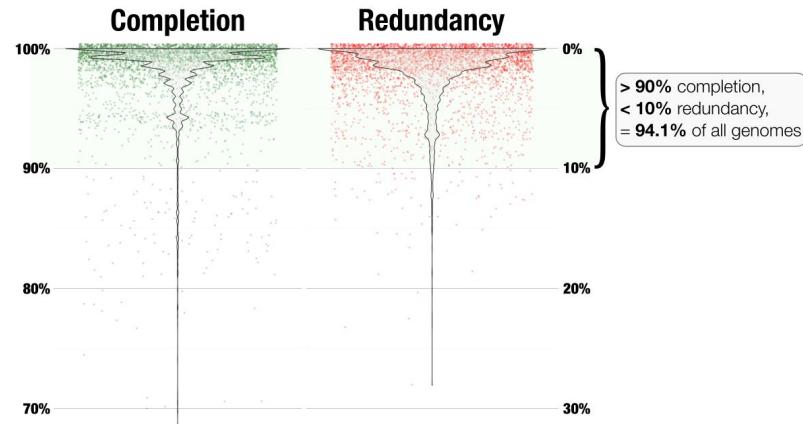
## 1. Completeness/completion

Marker genes are expected to be present in all bacteria

## 2. Contamination/redundancy

Single-copy genes are expected to be only once

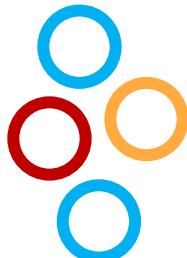
4,022 closed genomes from NCBI



Golden standard: CheckM

<https://merenlab.org>

# Metagenomics workflow: MAG-based analysis



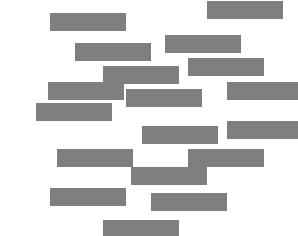
DNA extraction  
→



Sequencing  
→



Reads



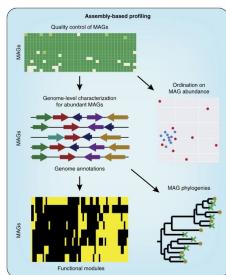
100-150  
bp

Assembly  
→  
Scaffolds



1000+ bp 73

MAG-based



Mapping  
Taxonom  
Function  
Genome

# 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	Medium	Low
Genome-resolved metabolism	High	Low
Expert manual supervision	High	Low/Medium
Integration with microbiomics	High	None

Choose according to your question!

Quince et.al., 2017

# Which tools to pick?

## Microbiome COSI



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/>). 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 taxon across several multi-sample data sets from different environments. This includes a marine data set (ended), a high-strain diversity data set 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

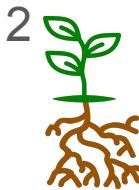
# Metagenomics is a great tool but...

- 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
- Toooooooooooooo much data....
  - That's why you are here!

# Challenges of getting genomes from metagenomes across environments



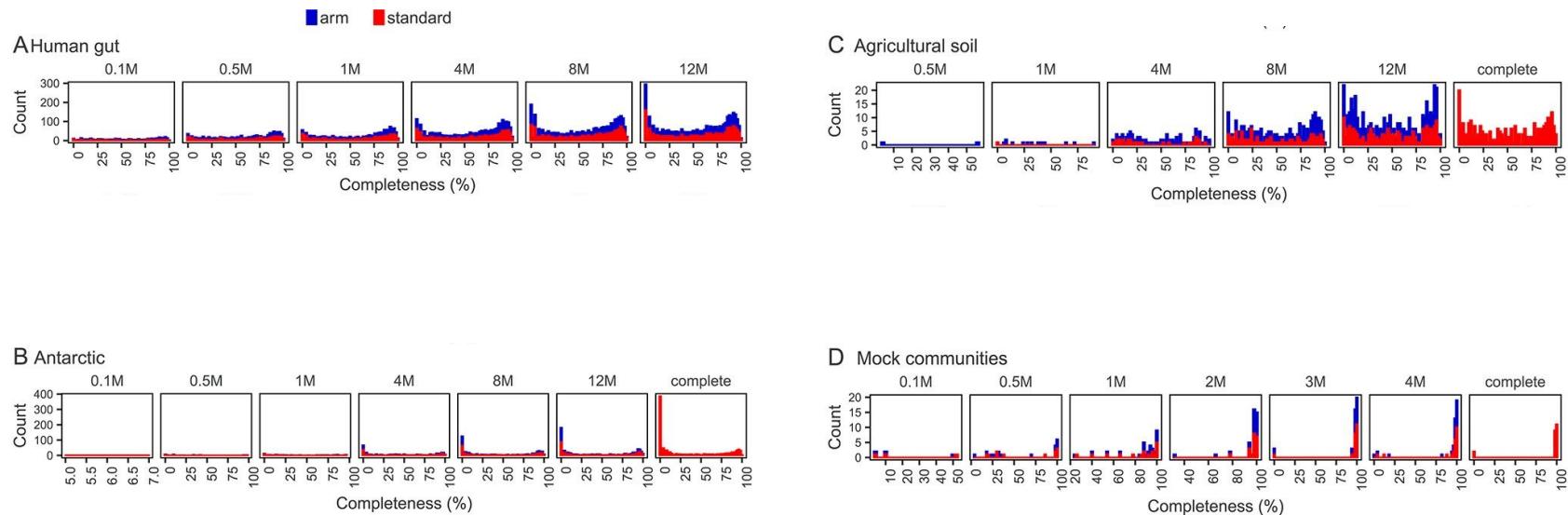
- 1) Which of those environments have the highest diversity?
- 2) From which we can get most MAGs?



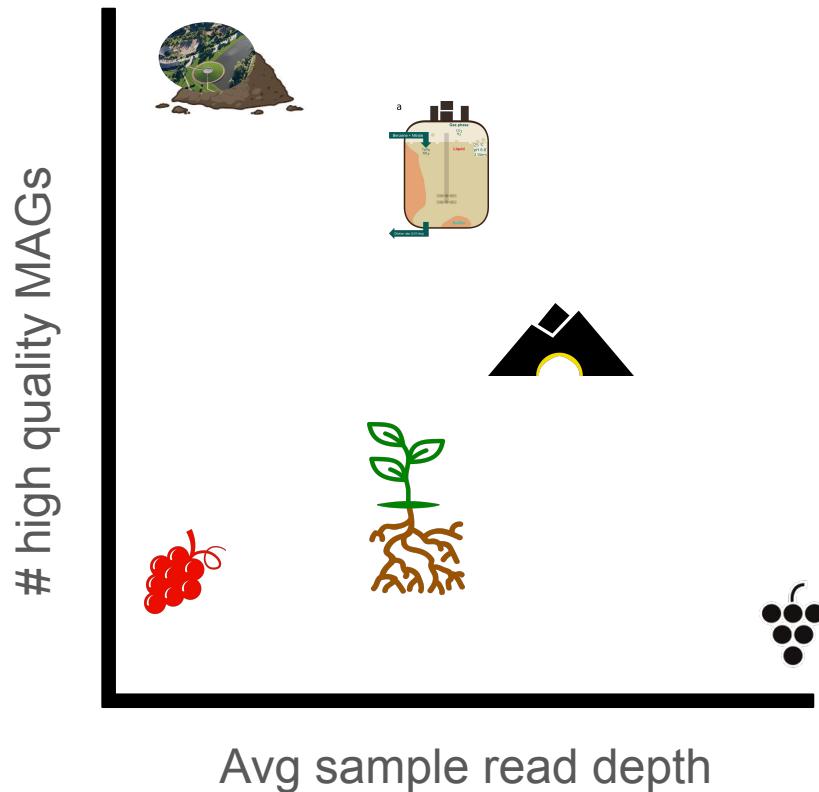
\*High quality:  
> 90% completeness &  
< 10% contamination

MAG:  
metagenome-assembled  
genome

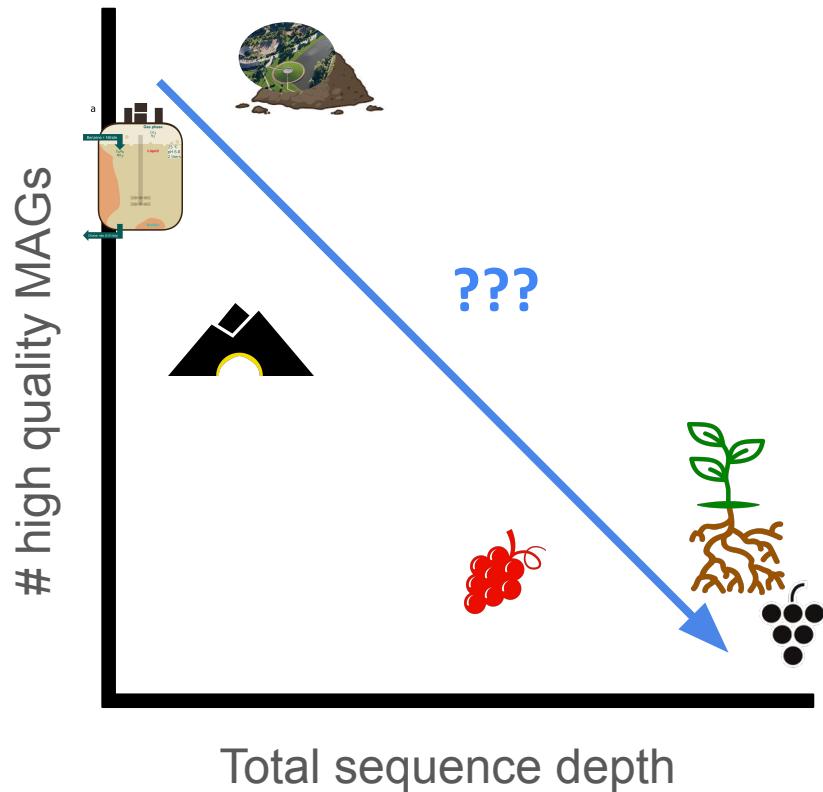
# Challenges of getting genomes from metagenomes across environments: The more data, the better?



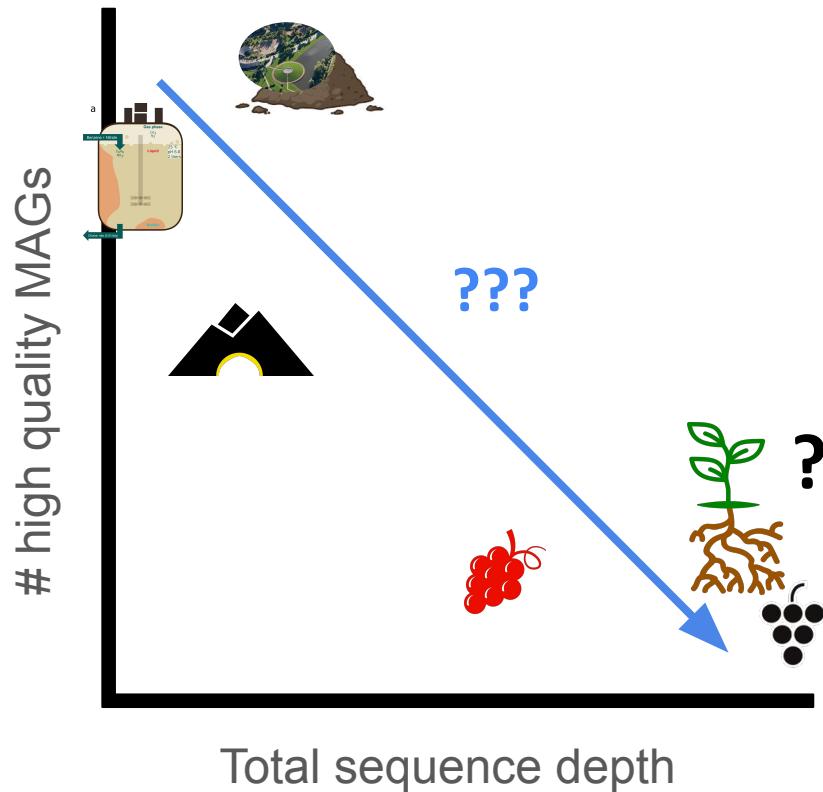
# Challenges of getting genomes from metagenomes across environments



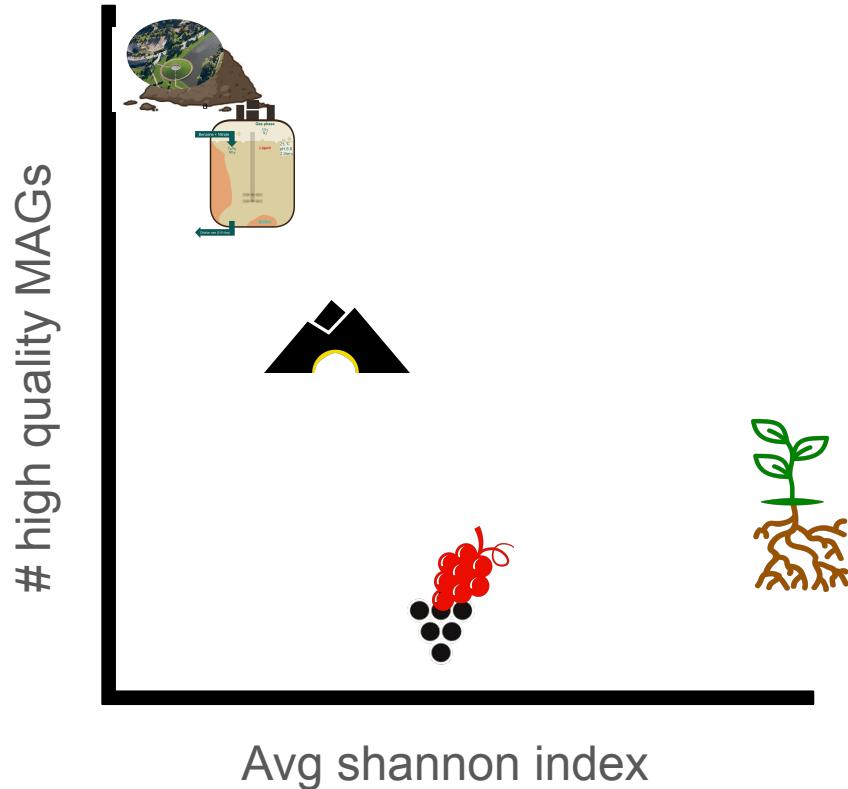
# Challenges of getting genomes from metagenomes across environments



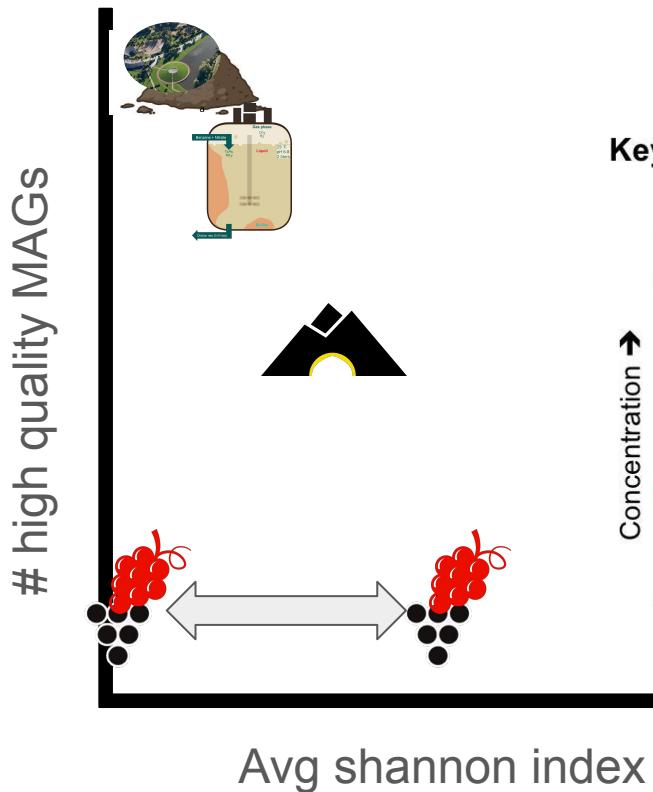
# Challenges of getting genomes from metagenomes across environments



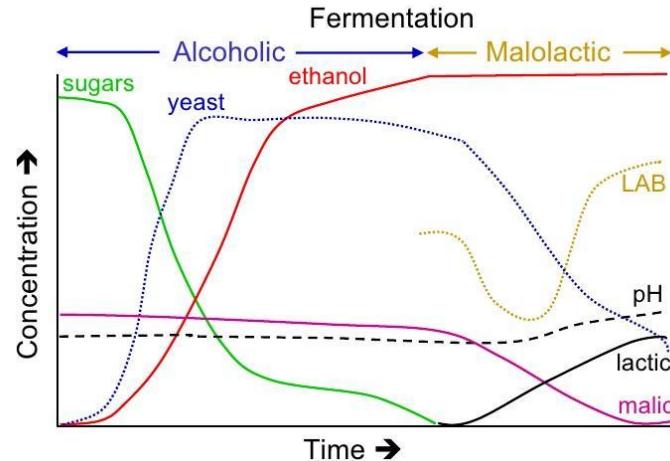
# Challenges of getting genomes from metagenomes across environments



# Challenges of getting genomes from metagenomes across environments



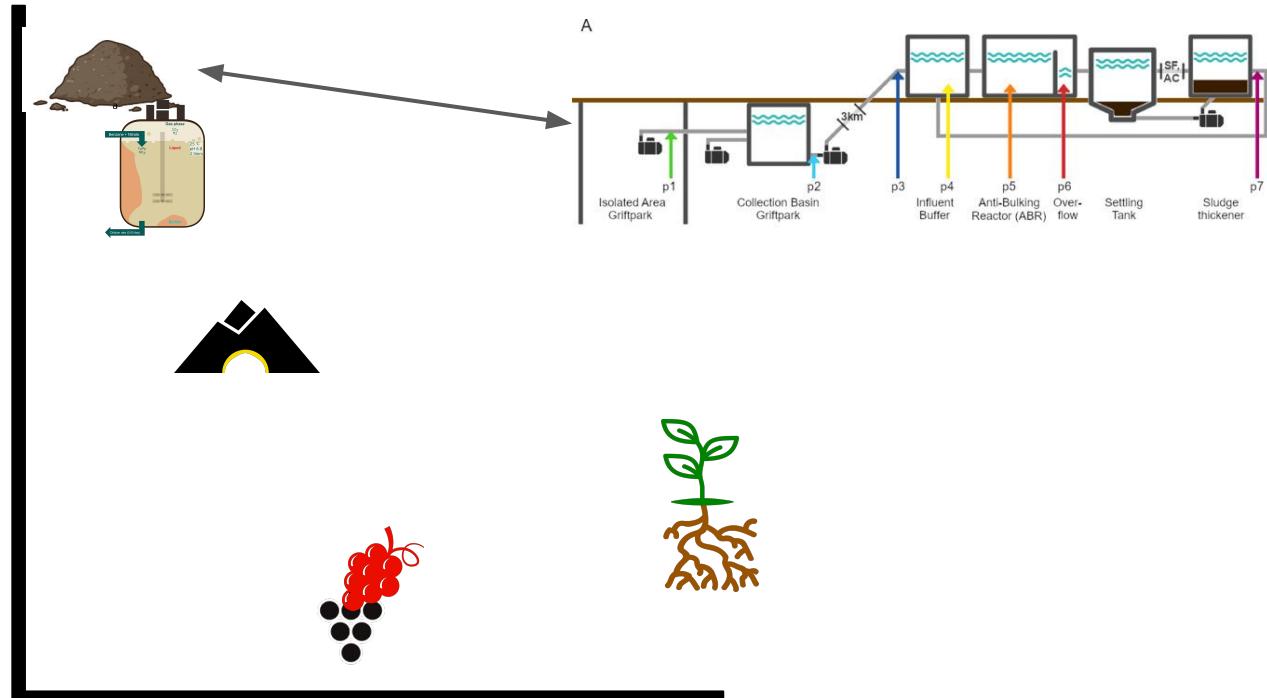
## Key events in winemaking



# Challenges of getting genomes from metagenomes across environments



# high quality MAGs



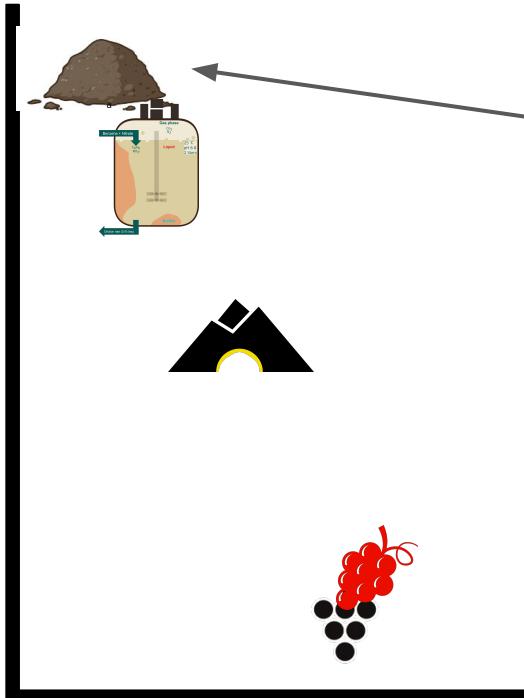
Avg shannon index

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>

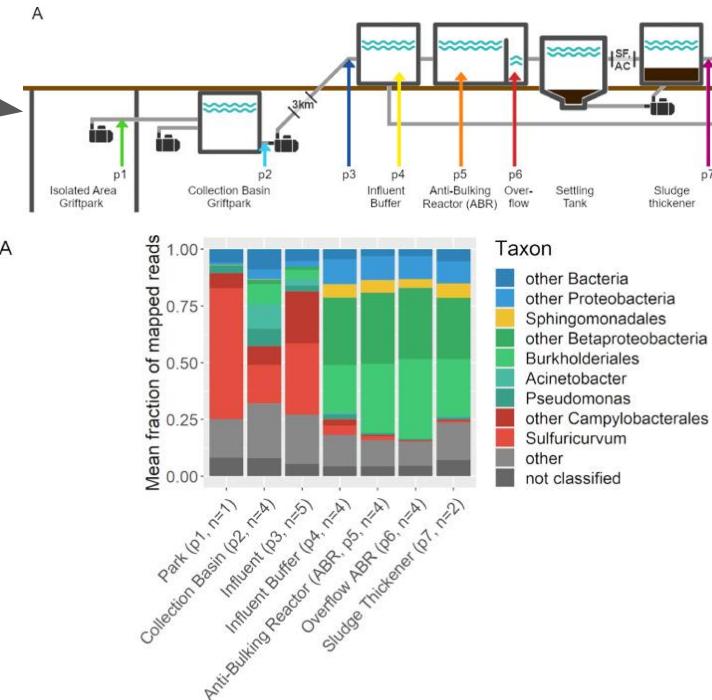
# Challenges of getting genomes from metagenomes across environments



# high quality MAGs

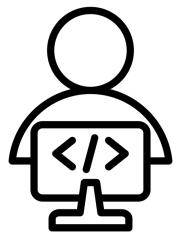


Avg shannon index



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



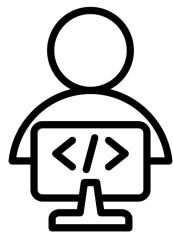
Connect to JupyterHub:

[https://bioinformatics.nl/biosb\\_metagenomics](https://bioinformatics.nl/biosb_metagenomics)

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

By Mattias de Hollander

# Practicals



Connect to JupyterHub:

[https://bioinformatics.nl/biosb\\_metagenomics](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**

