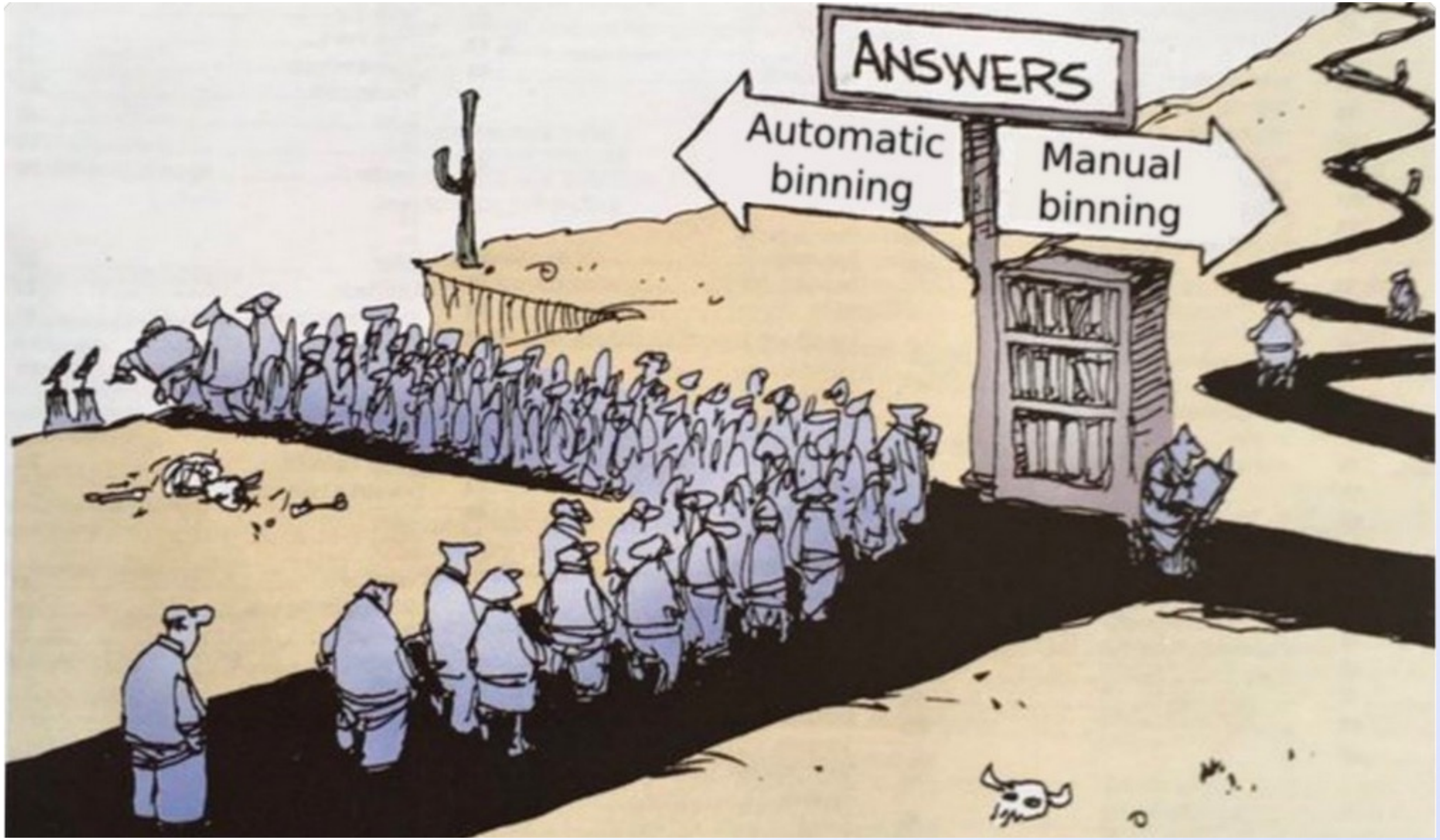
# Module V – Metagenomic binning and MAGs



Tom Delmont

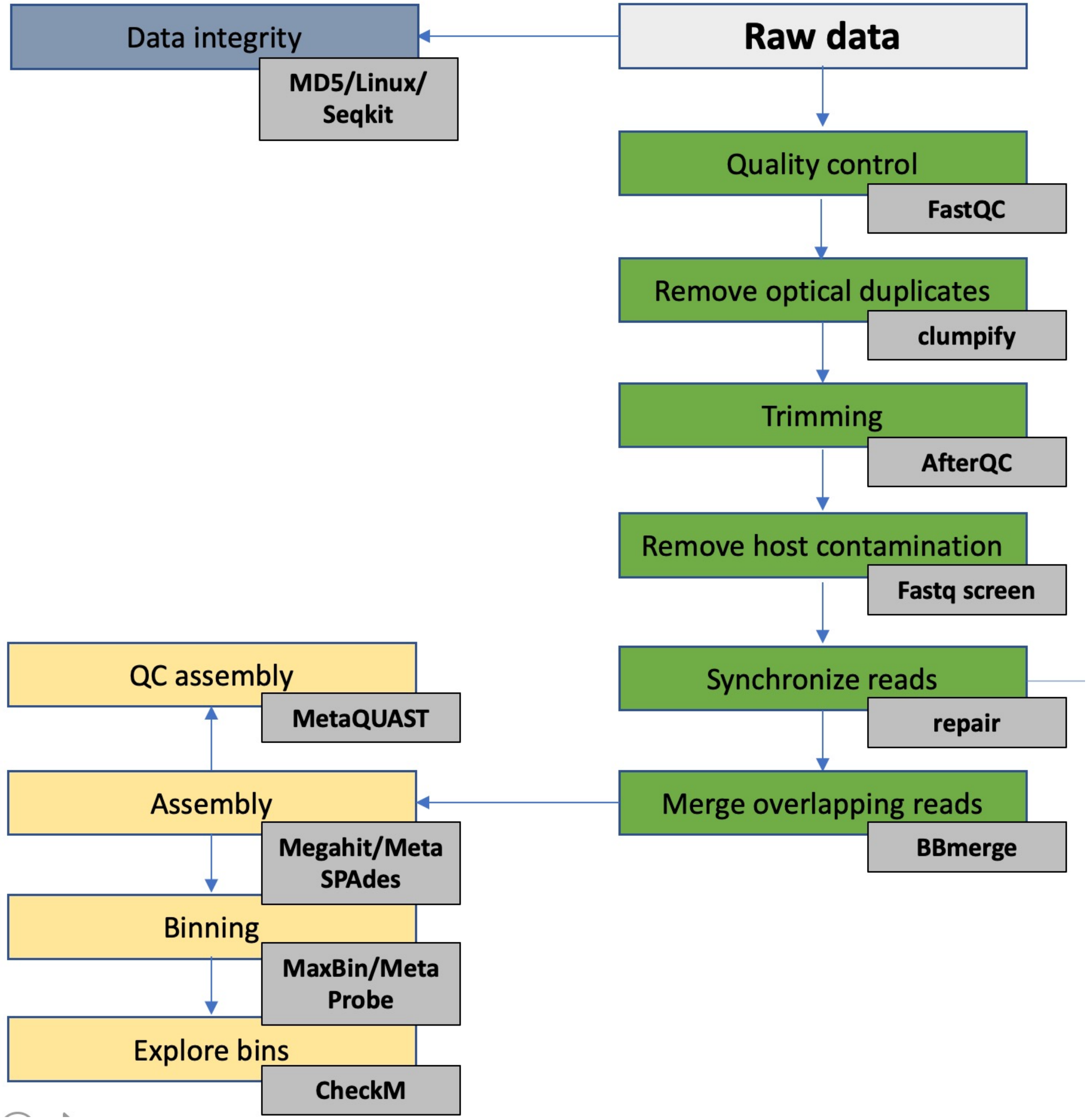
# Overview

Binning of contigs

Binning of reads

Evaluate metagenomic bins

Taxonomic classification of bins

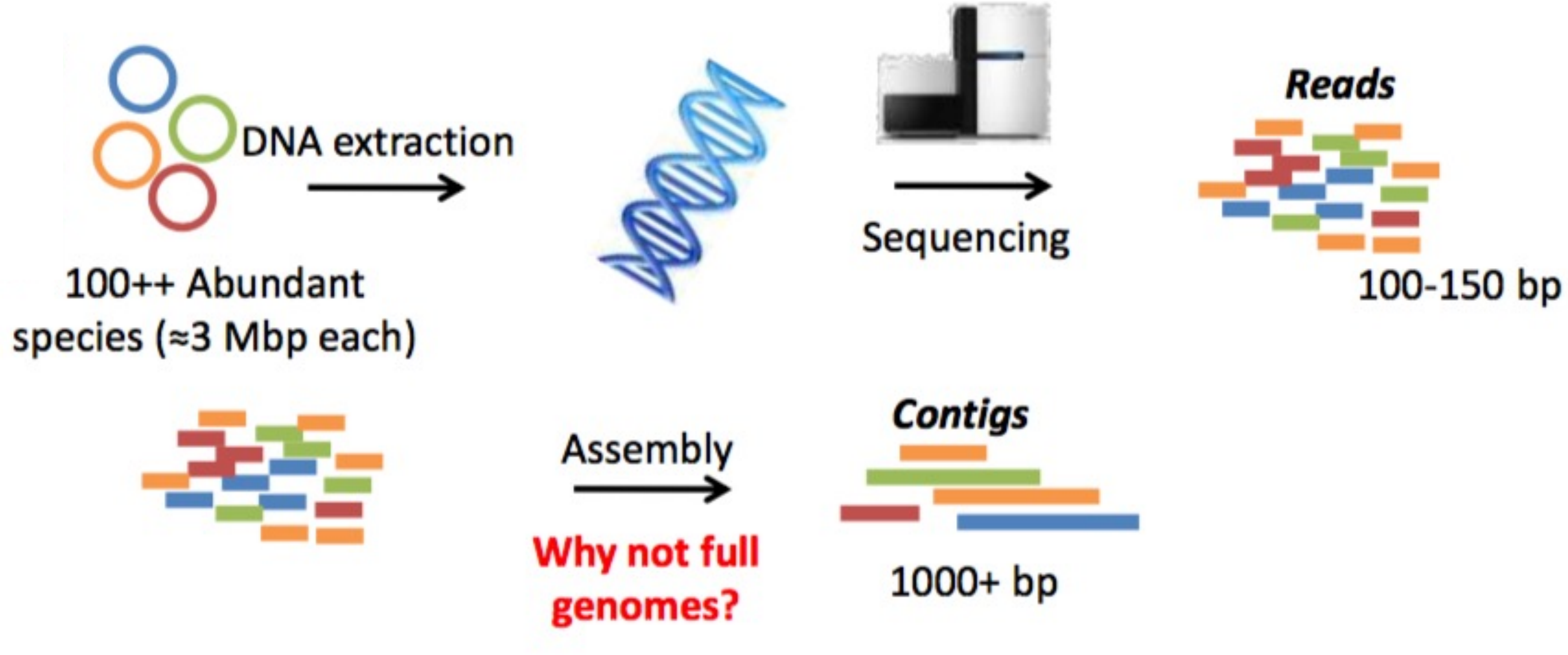


# Recap: obtaining a genome sequence from a metagenomes

## Metagenomic Assembled Genomes (MAGs)

Trying to reconstruct the individual genomes of a mixture of DNA from an entire population

Metagenomic assemblies will still be highly fragmented - Binning



www.slideshare.net Mads Albertsen, University of Vienna

# Binning

Method to sort data values into a smaller groups or “bins”

For example to group animals into more taxon-specific bins

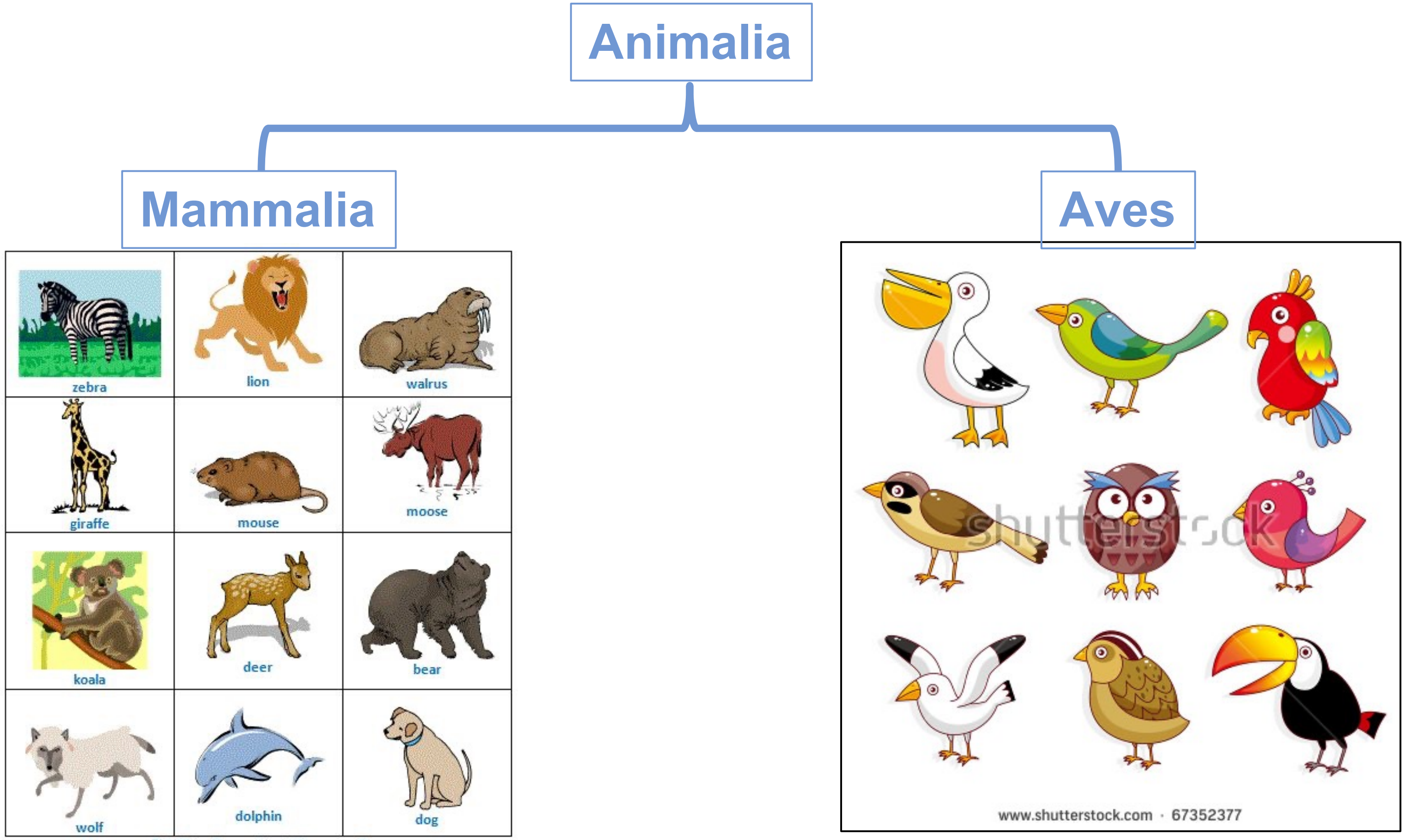
Jungle Animals Real Life Wallpapers Hd

# Binning

Method to sort data values into a smaller groups or “bins”

For example to group animals into more taxon-specific bins

Various taxonomic levels: All belong to Kingdom = *Animalia*, but Class = *Aves* AND *Mammalia*



Animalia

Mammalia

Aves

©All Rights Reserved Lovin2Learn.com™

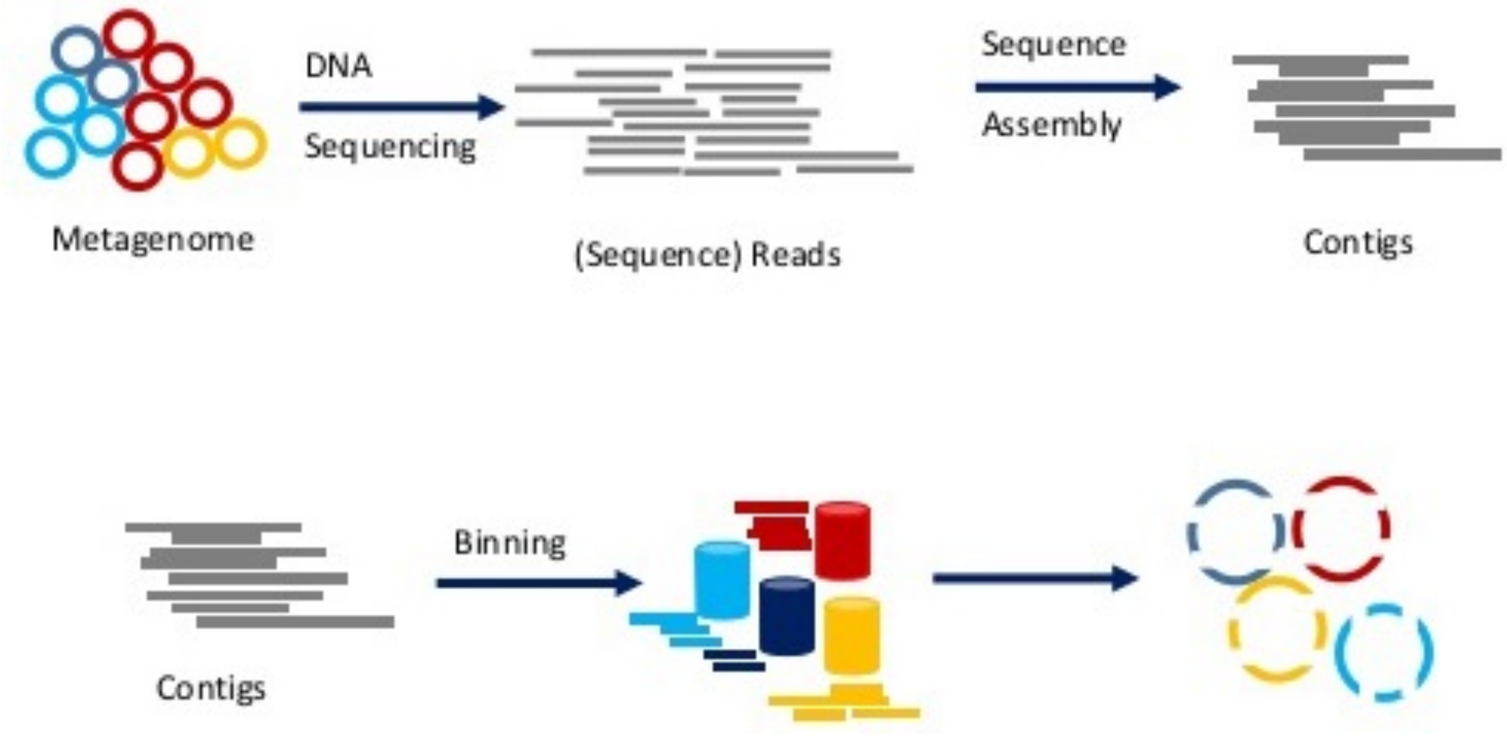
www.shutterstock.com · 67352377

# Binning - metagenomics

Group contigs or reads belonging to the same specie

Group nucleotide sequences based on **composition**

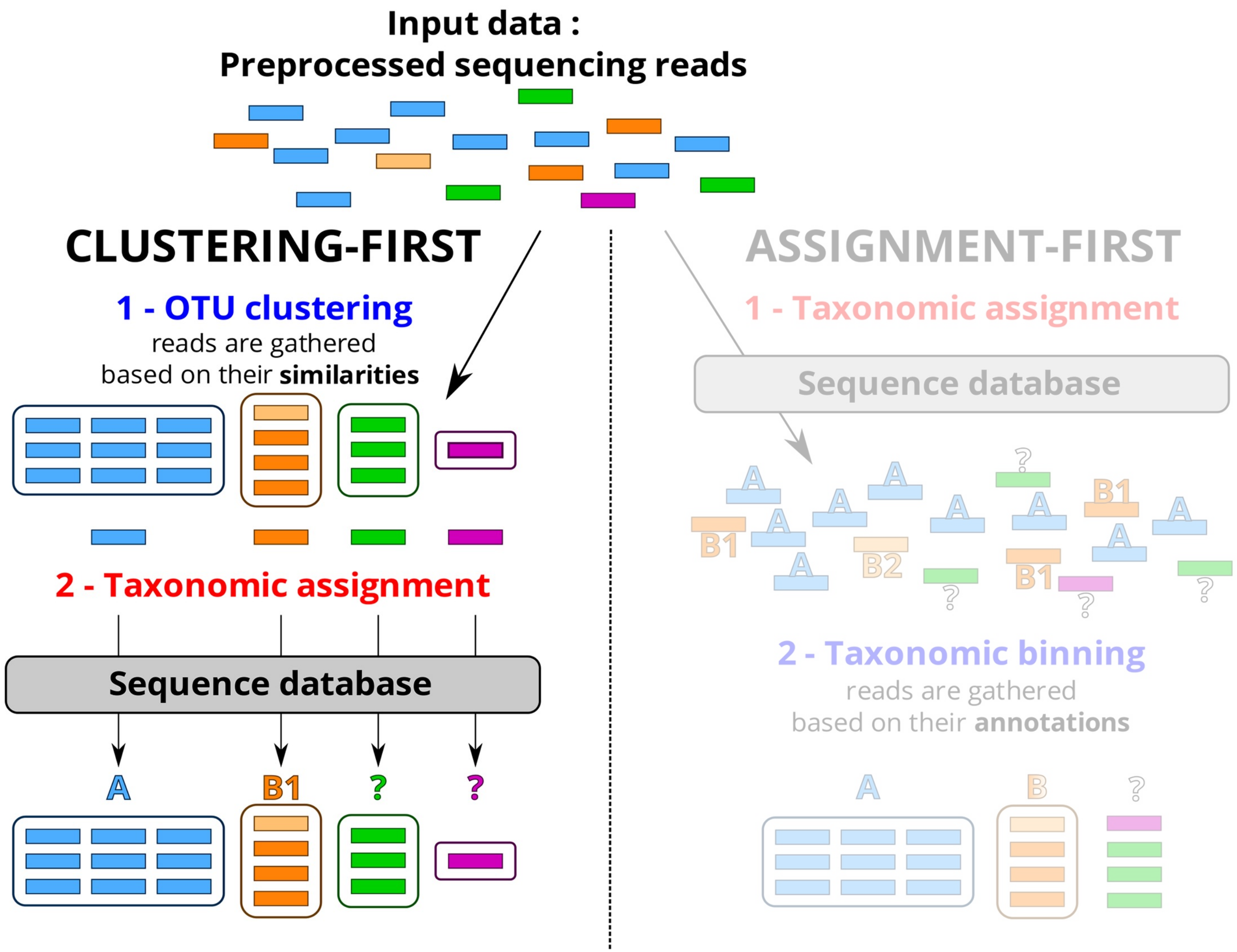
Group nucleotide sequences based on **abundance**



SlideShare / CoMet: Coverage and Composition based binning of Metagenomes

# Two types of binning strategies

Taxonomy dependent and taxonomy independent strategies



Modified from: Siegwald et al. / PLoS ONE 12(1): e0169563. https://doi.org/10.1371/journal.pone.0169563

# Binning – Taxonomy independent methods

Referred to as un-supervised

Enables the discovery of new microbial of new organisms

Two types of features used for classification

**Sequence composition based**

- Assumption that the genome composition is unique for each taxon
- DNA fragments from the same genome are more similar than those from different genomes
- Cluster formation being defined by k-mer composition

**Abundance based**

- Coverage reflecting abundance of given tax
- Cluster formation being defined by k-mer abundance

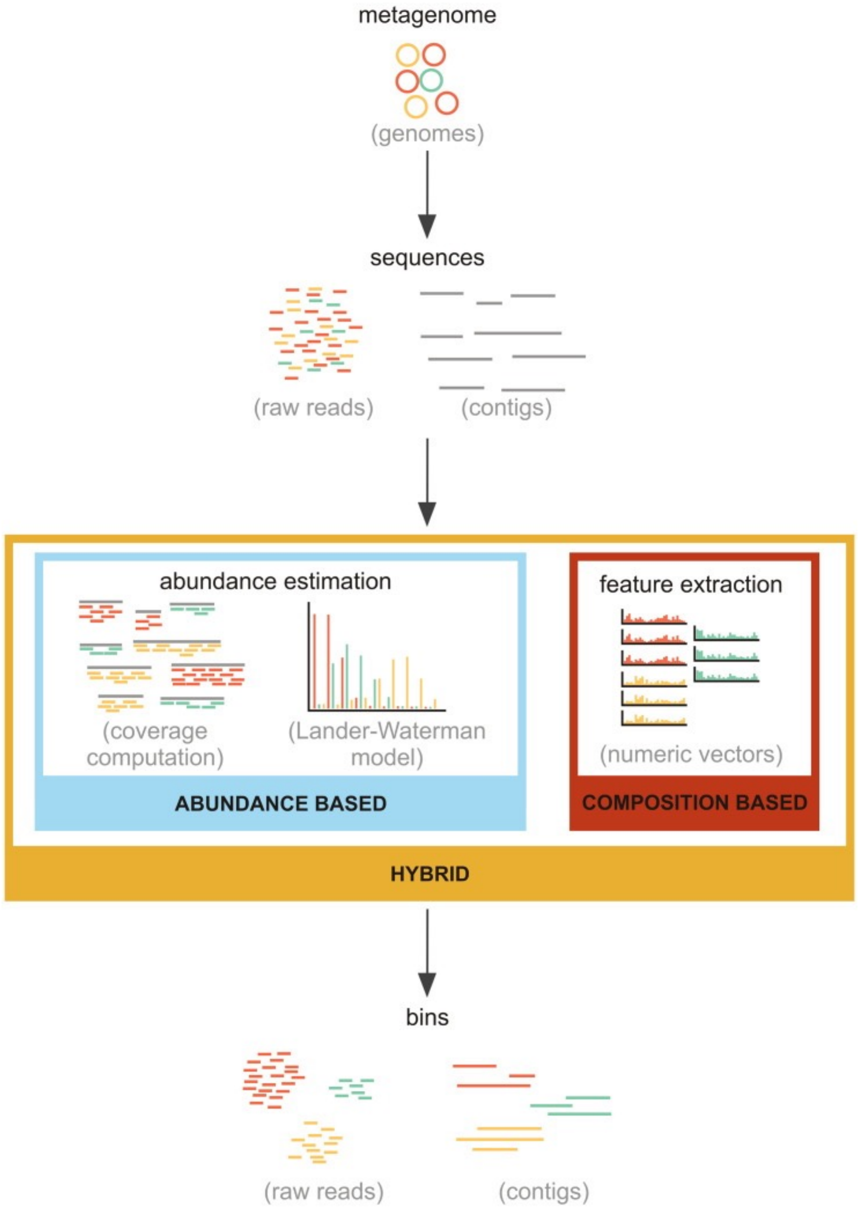
# Binning – Taxonomy independent methods

Hybrid binning

- Combine abundance and composition
- Give more accurate binning results

Can be performed on either sequence reads or assembled contigs

Potentially separate subspecies into individual bins



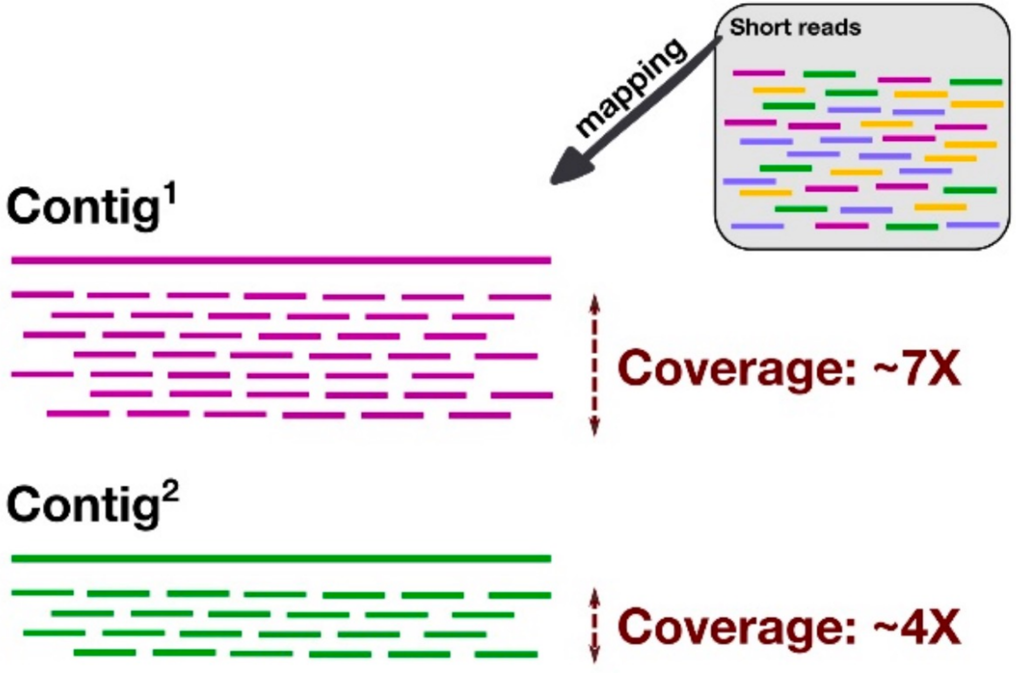
K. Sedlar et al. / Computational and Structural Biotechnology Journal 15 (2017)

# How are nucleotide sequences binned?

Abundance based binning

Also called coverage based binning

Sequences originating from the same specie will have similar abundance in the sample



SlideShare / A. Murat Eren, Intro to metagenomic binning

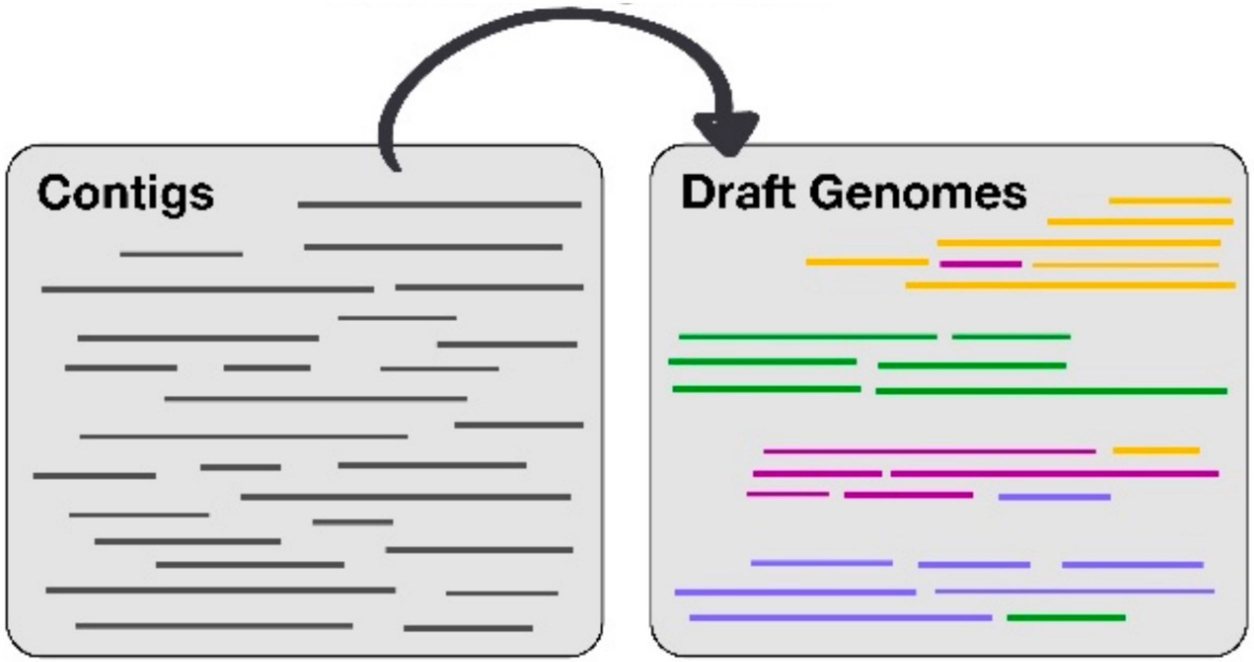
# How are nucleotide sequences binned?

## Composition based binning

Genomic signatures have been shown to display a species-specific pattern

GC content is simple and commonly used genomic signature

More widely used genomic signature is tetranucleotide frequencies



SlideShare / A. Murat Eren, Intro to metagenomic binning

# Composition based binning

Computation of tetranucleotide frequencies (k-mer =4)

Seq1: AATTCCGG

# Composition based binning

## Computation of tetranucleotide frequencies (k-mer =4)

```
Seq1: AATTCCGG
      AATT
       ATTC
        TTCC
         TCCG
          CCGG
```

# Composition based binning

## Computation of tetranucleotide frequencies (k-mer =4)

Seq1: AATTCCGG

AATT<br>
ATTC<br>
TTCC<br>
TCCG<br>
CCGG

|  | AATT | ATTC | TTCC | TCCG | CCGG |
|---|---|---|---|---|---|
| Seq1 |  |  |  |  |  |

# Composition based binning

## Computation of tetranucleotide frequencies (k-mer =4)

Seq1: AATTCCGG

AATT

ATTC

TTCC

TCCG

CCGG

| | AATT | ATTC | TTCC | TCCG | CCGG |
|---|---|---|---|---|---|
| Seq1 | 1 | 1 | 1 | 1 | 1 |

# Composition based binning

## Computation of tetranucleotide frequencies (k-mer =4)

Seq1: AATTCCGG
AATT
ATTC
TTCC
TCCG
CCGG

Seq2: AATTAAGG
AATT
ATTA
TTAA
TAAG
AAGG

| | AATT | ATTC | TTCC | TCCG | CCGG | ATTA | TTAA | TAAG | AAGG |
|---|---|---|---|---|---|---|---|---|---|
| Seq1 | 1 | 1 | 1 | 1 | 1 | | | | |
| Seq2 | 1 | | | | | 1 | 1 | 1 | 1 |

# Composition based binning

## Computation of tetranucleotide frequencies (k-mer =4)

Seq1: AATTCCGG
AATT
ATTC
TTCC
TCCG
CCGG

Seq2: AATTAAGG
AATT
ATTA
TTAA
TAAG
AAGG

Seq3: AAGGAAGG
AAGG
AGGA
GGAA
GAAG
AAGG

Seq4: AATTAATT
AATT
ATTA
TTAA
TAAT
AATT

Seq5: GGAAGGAA
GGAA
GAAG
AAGG
AGGA
GGAA

| | AATT | ATTC | TTCC | TCCG | CCGG | ATTA | TTAA | TAAG | AAGG | TCCC | AGGA | GGAA | GAAG | TAAT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Seq1 | 1 | 1 | 1 | 1 | 1 | | | | | | | | | |
| Seq2 | 1 | | | | | 1 | 1 | 1 | 1 | | | | | |
| Seq3 | | | | | | | | | 2 | | 1 | 1 | 1 | |
| Seq4 | 2 | | | | | 1 | 1 | | | | | | | 1 |
| Seq5 | | | | | | | | | 1 | | 1 | 2 | 1 | |

# Composition based binning

## Computation of tetranucleotide frequencies (k-mer =4)

Seq1: AATTCCGG
AATT
ATTC
TTCC
TCCG
CCGG

Seq2: AATTAAGG
AATT
ATTA
TTAA
TAAG
AAGG

Seq3: AAGGAAGG
AAGG
AGGA
GGAA
GAAG
AAGG

Seq4: AATTAATT
AATT
ATTA
TTAA
TAAT
AATT

Seq5: GGAAGGAA
GGAA
GAAG
AAGG
AGGA
GGAA

| | AATT | ATTC | TTCC | TCCG | CCGG | ATTA | TTAA | TAAG | AAGG | TCCC | AGGA | GGAA | GAAG | TAAT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Seq1 | 1 | 1 | 1 | 1 | 1 | | | | | | | | | |
| Seq2 | 1 | | | | | 1 | 1 | 1 | 1 | | | | | |
| Seq3 | | | | | | | | | 2 | | 1 | 1 | 1 | |
| Seq4 | 2 | | | | | 1 | 1 | | | | | | | 1 |
| Seq5 | | | | | | | | | 1 | | 1 | 2 | 1 | |

# Composition based binning

## Computation of tetranucleotide frequencies (k-mer =4)

Seq1: AATTCCGG
AATT
ATTC
TTCC
TCCG
CCGG

Seq2: AATTAAGG
AATT
ATTA
TTAA
TAAG
AAGG

Seq3: AAGGAAGG
AAGG
AGGA
GGAA
GAAG
AAGG

Seq4: AATTAATT
AATT
ATTA
TTAA
TAAT
AATT

Seq5: GGAAGGAA
GGAA
GAAG
AAGG
AGGA
GGAA

| | AATT | ATTC | TTCC | TCCG | CCGG | ATTA | TTAA | TAAG | AAGG | TCCC | AGGA | GGAA | GAAG | TAAT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Seq1 | 1 | 1 | 1 | 1 | 1 | | | | | | | | | |
| Seq2 | 1 | | | | | 1 | 1 | 1 | 1 | | | | | |
| Seq4 | 2 | | | | | 1 | 1 | | | | | | | 1 |
| Seq3 | | | | | | | | | 2 | | 1 | 1 | 1 | |
| Seq5 | | | | | | | | | 1 | | 1 | 2 | 1 | |

# Composition based binning

## Computation of tetranucleotide frequencies (k-mer =4)

Seq1: AATTCCGG
AATT
ATTC
TTCC
TCCG
CCGG

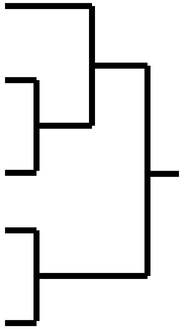
Seq2: AATTAAGG
AATT
ATTA
TTAA
TAAG
AAGG

Seq3: AAGGAAGG
AAGG
AGGA
GGAA
GAAG
AAGG

Seq4: AATTAATT
AATT
ATTA
TTAA
TAAT
AATT

Seq5: GGAAGGAA
GGAA
GAAG
AAGG
AGGA
GGAA

| | AATT | ATTC | TTCC | TCCG | CCGG | ATTA | TTAA | TAAG | AAGG | TCCC | AGGA | GGAA | GAAG | TAAT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Seq1 | 1 | 1 | 1 | 1 | 1 | | | | | | | | | |
| Seq2 | 1 | | | | | 1 | 1 | 1 | 1 | | | | | |
| Seq4 | 2 | | | | | 1 | 1 | | | | | | | 1 |
| Seq3 | | | | | | | | | 2 | | 1 | 1 | 1 | |
| Seq5 | | | | | | | | | 1 | | 1 | 2 | 1 | |



# Composition based binning

## Computation of tetranucleotide frequencies (k-mer =4)

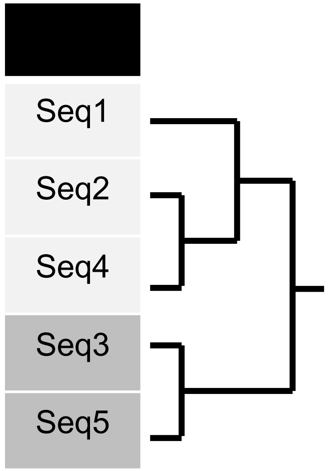
Seq1: AATTCCGG
AATT
ATTC
TTCC
TCCG
CCGG

Seq2: AATTAAGG
AATT
ATTA
TTAA
TAAG
AAGG

Seq3: AAGGAAGG
AAGG
AGGA
GGAA
GAAG
AAGG

Seq4: AATTAATT
AATT
ATTA
TTAA
TAAT
AATT

Seq5: GGAAGGAA
GGAA
GAAG
AAGG
AGGA
GGAA



# Composition based binning

## Computation of tetranucleotide frequencies (k-mer =4)

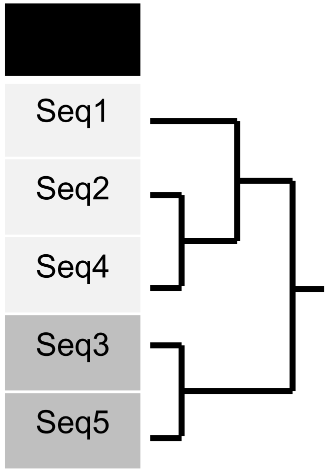
Seq1: AATTCCGG
AATT
ATTC
TTCC
TCCG
CCGG

Seq2: AATTAAGG
AATT
ATTA
TTAA
TAAG
AAGG

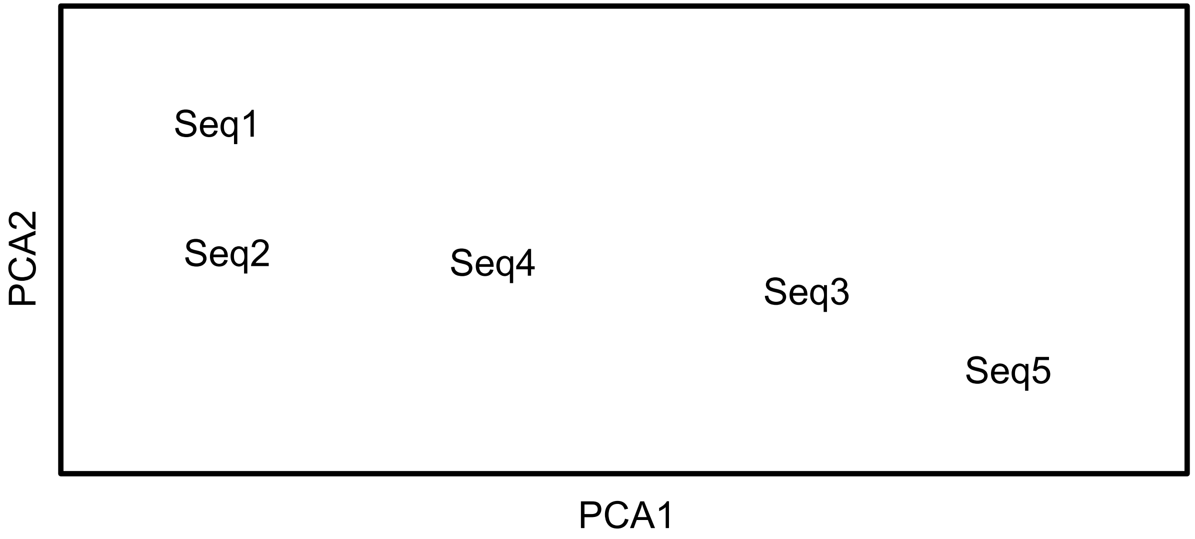
Seq3: AAGGAAGG
AAGG
AGGA
GGAA
GAAG
AAGG

Seq4: AATTAATT
AATT
ATTA
TTAA
TAAT
AATT

Seq5: GGAAGGAA
GGAA
GAAG
AAGG
AGGA
GGAA



PCA Analysis



# Composition based binning

## Computation of tetranucleotide frequencies (k-mer =4)

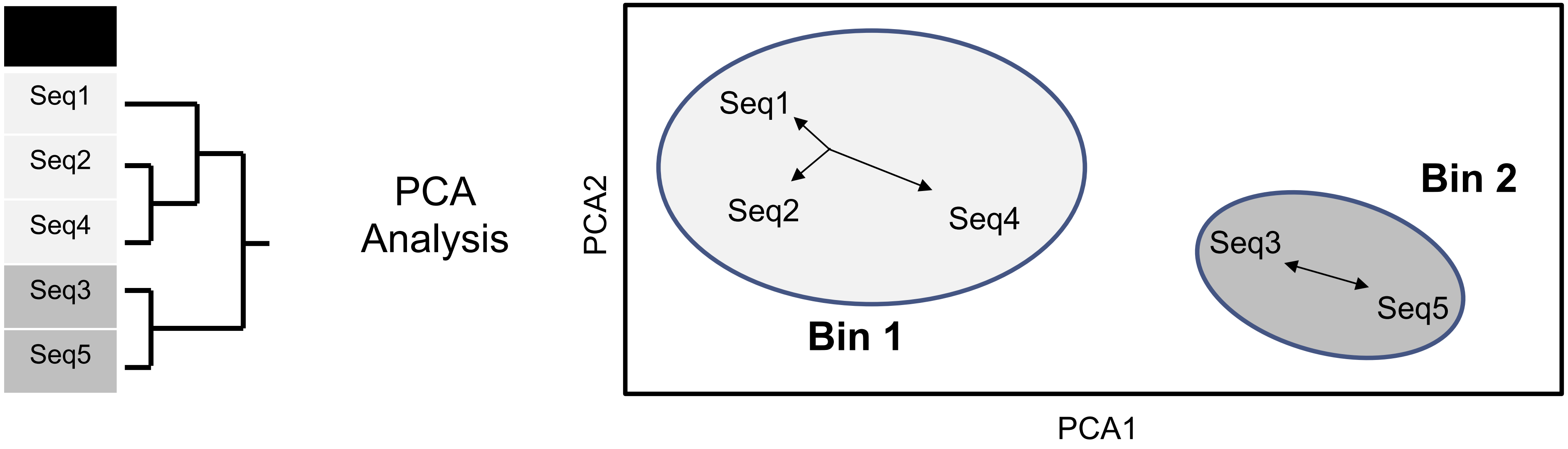
Seq1: AATTCCGG
AATT
ATTC
TTCC
TCCG
CCGG

Seq2: AATTAAGG
AATT
ATTA
TTAA
TAAG
AAGG

Seq3: AAGGAAGG
AAGG
AGGA
GGAA
GAAG
AAGG

Seq4: AATTAATT
AATT
ATTA
TTAA
TAAT
AATT

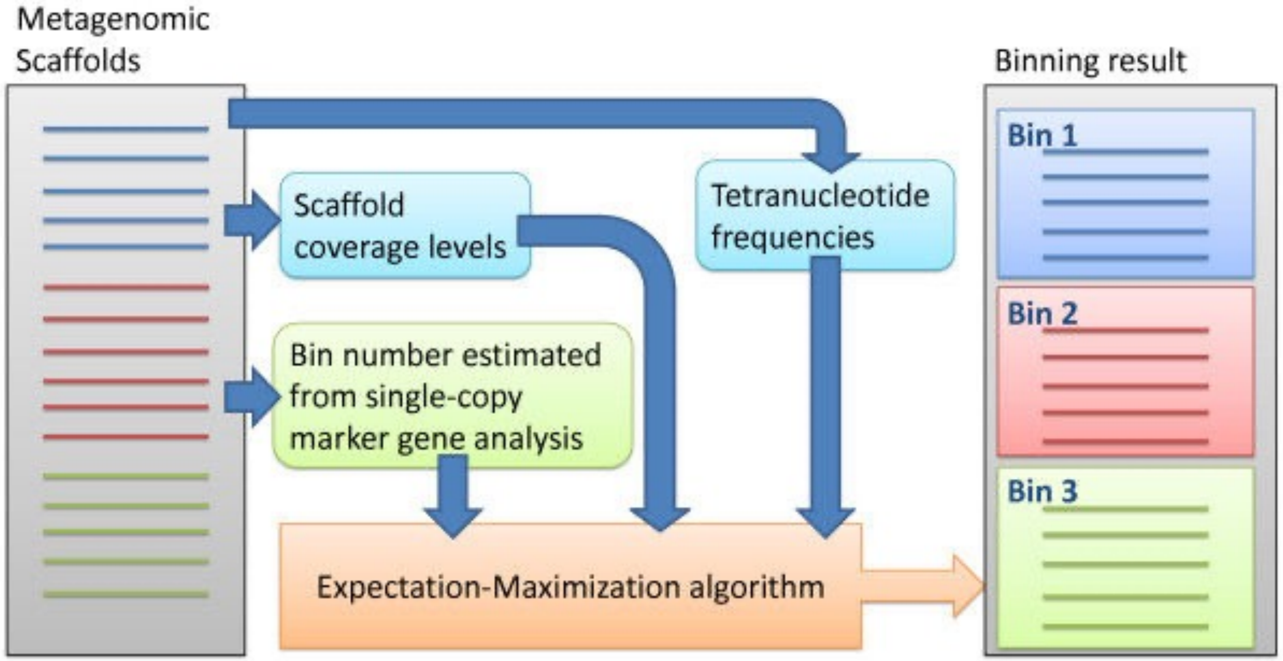
Seq5: GGAAGGAA
GGAA
GAAG
AAGG
AGGA
GGAA



# Taxonomy independent binning of contigs - MaxBin

Binning of assembled contigs using an expectation-maximization algorithm

- Bins are predicted from initial identification of marker genes in assembled sequences
- Tetranucleotide frequencies and scaffold coverages are combined to organize metagenomic sequences into individual bins
- Estimation of genome completeness – 107 marker genes

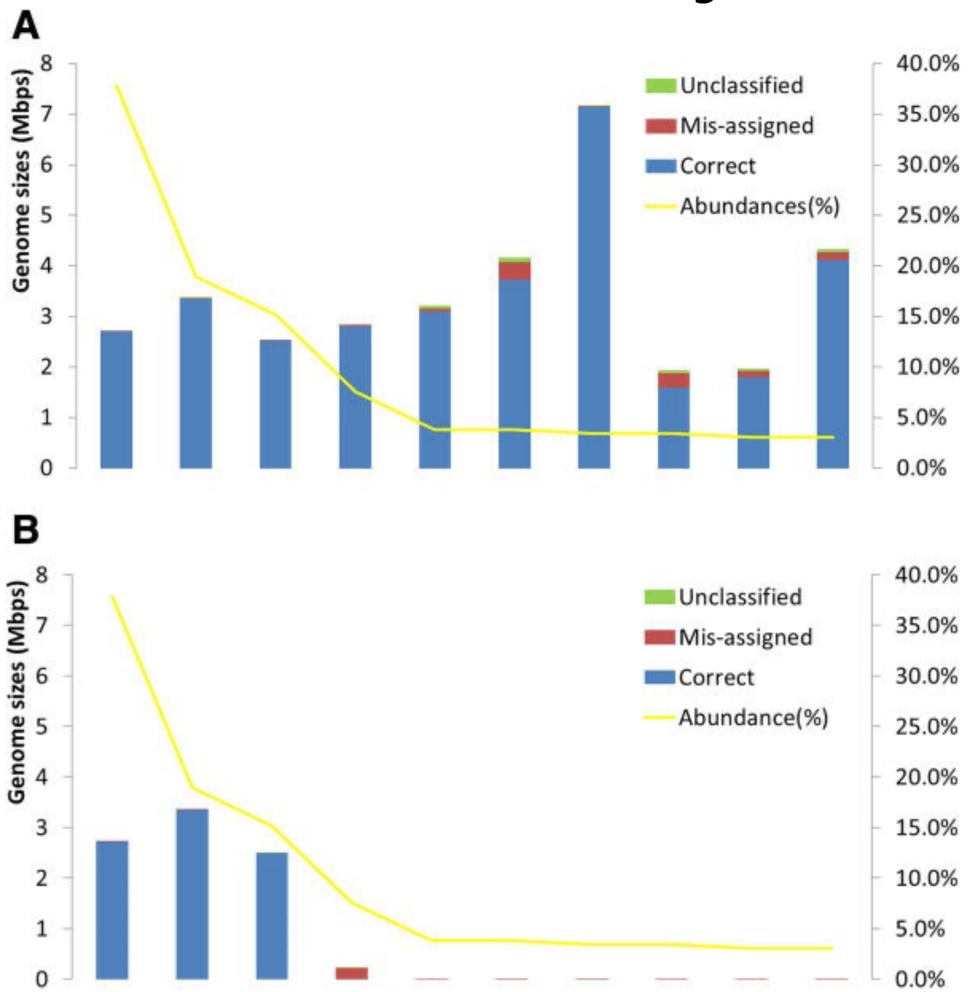

Metagenomic Scaffolds
Scaffold coverage levels
Bin number estimated from single-copy marker gene analysis
Tetranucleotide frequencies
Expectation-Maximization algorithm
Binning result
Bin 1
Bin 2
Bin 3

Wu et al., Microbiome 2014 2:26

# MaxBin - preformance

Sequencing depths highly affect the results

10-genome simulated datasets - 20X versus 80X coverage



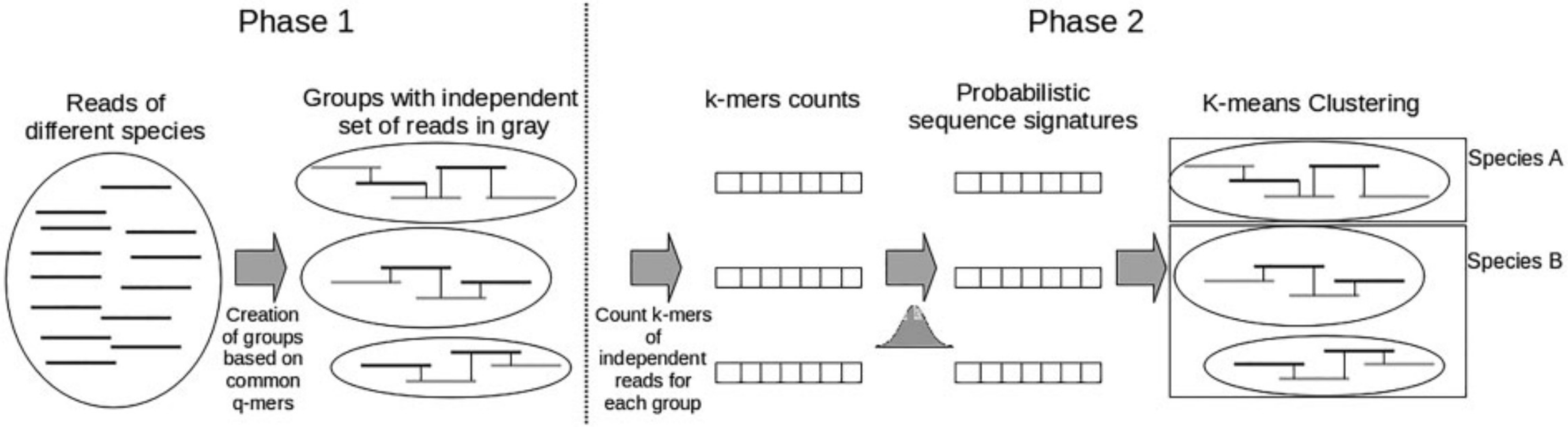
Wu et al., Microbiome 2014 2:26

# Taxonomy independent binning of sequence reads- MetaProb

Assembly-assisted tool for binning of reads

Phase 1 groups overlapping reads into groups

Phase 2 builds the probabilistic sequence signatures of independent reads and merges the groups into clusters



Phase 1
Phase 2
Reads of different species
Groups with independent set of reads in gray
Creation of groups based on common q-mers
k-mers counts
Count k-mers of independent reads for each group
Probabilistic sequence signatures
K-means Clustering
Species A
Species B

S. Girotto et al. / Bioinformatics, Volume 32, Issue 17, 1 September 2016, Pages i567–i575

# Binning results for the CAMI data sets

Investigated performance when recovering individual genome bins

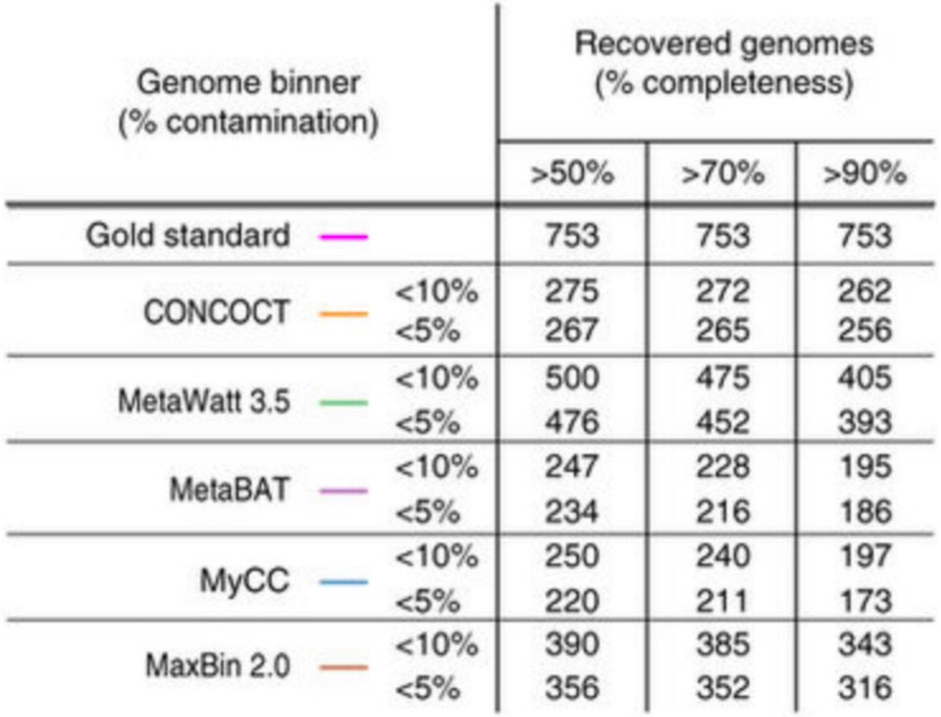
Large variation:

- Average genome completeness (34% to 80%)
- Average purity (70% to 97%)

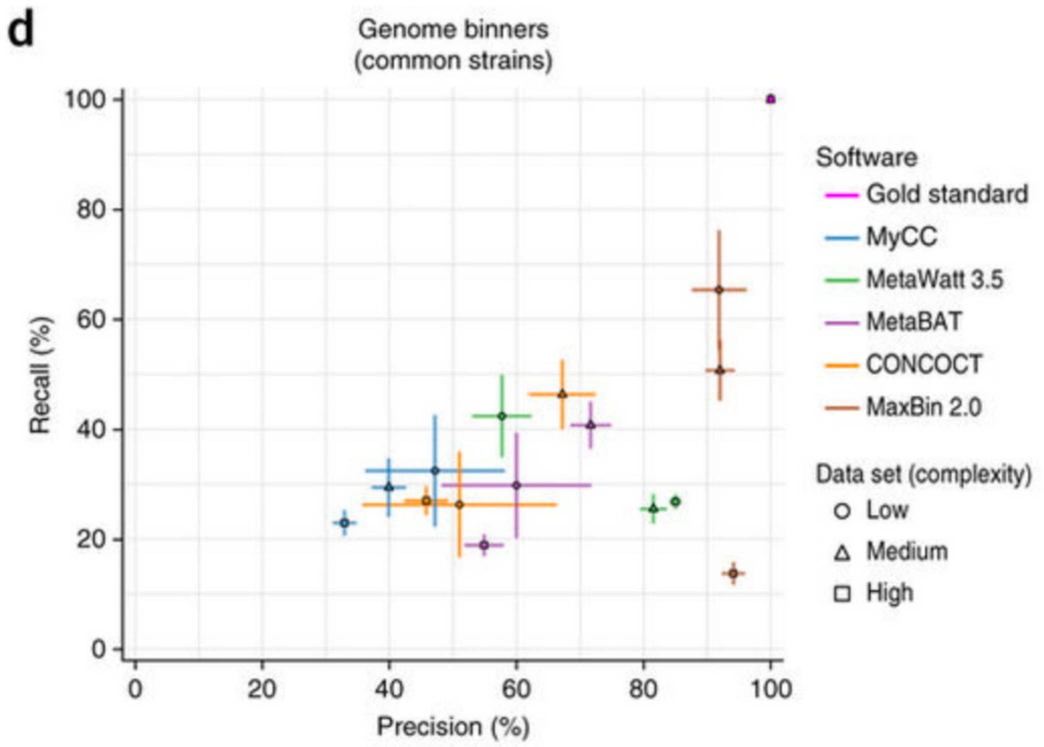
nature methods

Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software

Alexander Sczyrba ✉, Peter Hofmann […] Alice C McHardy ✉



| Genome binner (% contamination) | | Recovered genomes (% completeness) | | |
|---|---|---|---|---|
| | | >50% | >70% | >90% |
| Gold standard — | | 753 | 753 | 753 |
| CONCOCT — | <10% | 275 | 272 | 262 |
| | <5% | 267 | 265 | 256 |
| MetaWatt 3.5 — | <10% | 500 | 475 | 405 |
| | <5% | 476 | 452 | 393 |
| MetaBAT — | <10% | 247 | 228 | 195 |
| | <5% | 234 | 216 | 186 |
| MyCC — | <10% | 250 | 240 | 197 |
| | <5% | 220 | 211 | 173 |
| MaxBin 2.0 — | <10% | 390 | 385 | 343 |
| | <5% | 356 | 352 | 316 |



# Selecting a binning method

Highly dependent on the sample and the aim of the project and available resources

The length of the metagenomic sequences – key factor

- Ultra-short sequences (75 bp) - assembly step becomes a necessity
- Short length sequences (200-400 bp)- alignment-based or hybrid binning methods
- Long length sequences - alignment-based as well as composition-based binning methods

Are you aiming to identify novel un-culturable species?

- Human microbiota
  - Most species are known
  - Presence or absence of one or several species?
- Environmental samples
  - Most species are unknown

# Refinement of bins - RefineM

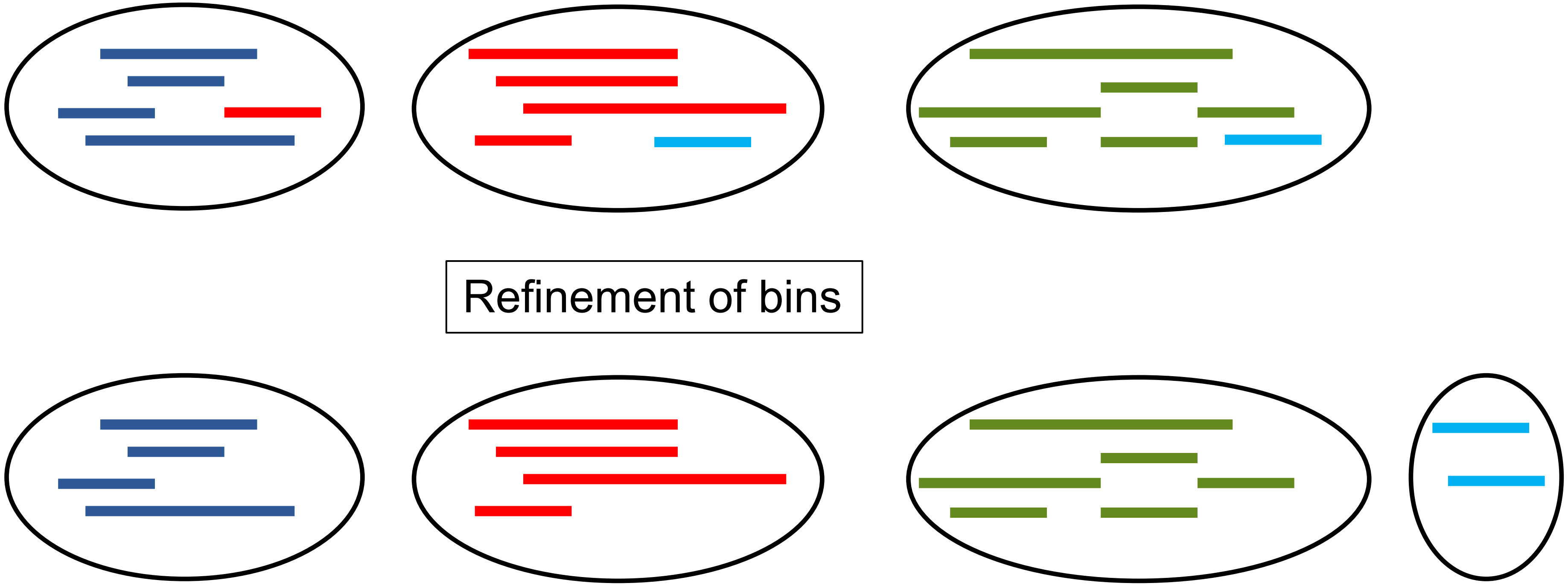
Methods for outlier filtering reduces the total number of contigs being binned

Deviating GC

Deviating tetranucleotide composition

Deviating coverage depth

Identify contigs with a coding density suggestive of a Eukaryotic origin



Refinement of bins

# Estimate completeness and contamination of MAGs

Assembly statistics

- Total size of MAG (sum of contigs in bin)
- Contig size (N50 value)

Presence and absence of lineage-specific genes

Presence of 20 standard tRNAs

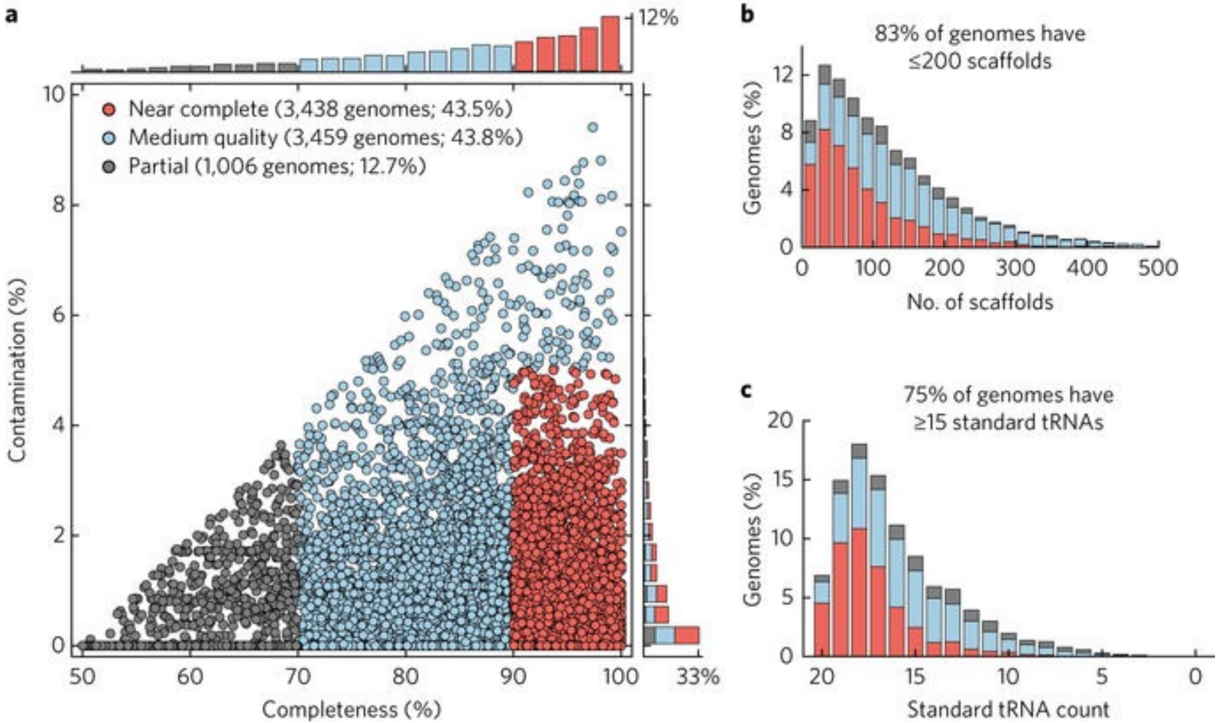
# Estimate completeness and contamination of MAGs

CheckM assess the quality of genomes recovered from metagenomes

Estimate genome completeness and contamination

Using collocated single-copy marker genes within a phylogenetic lineage

- Bacteria: 104 markers organized into 58 sets
- Archaea: 150 markers organized into 108 sets

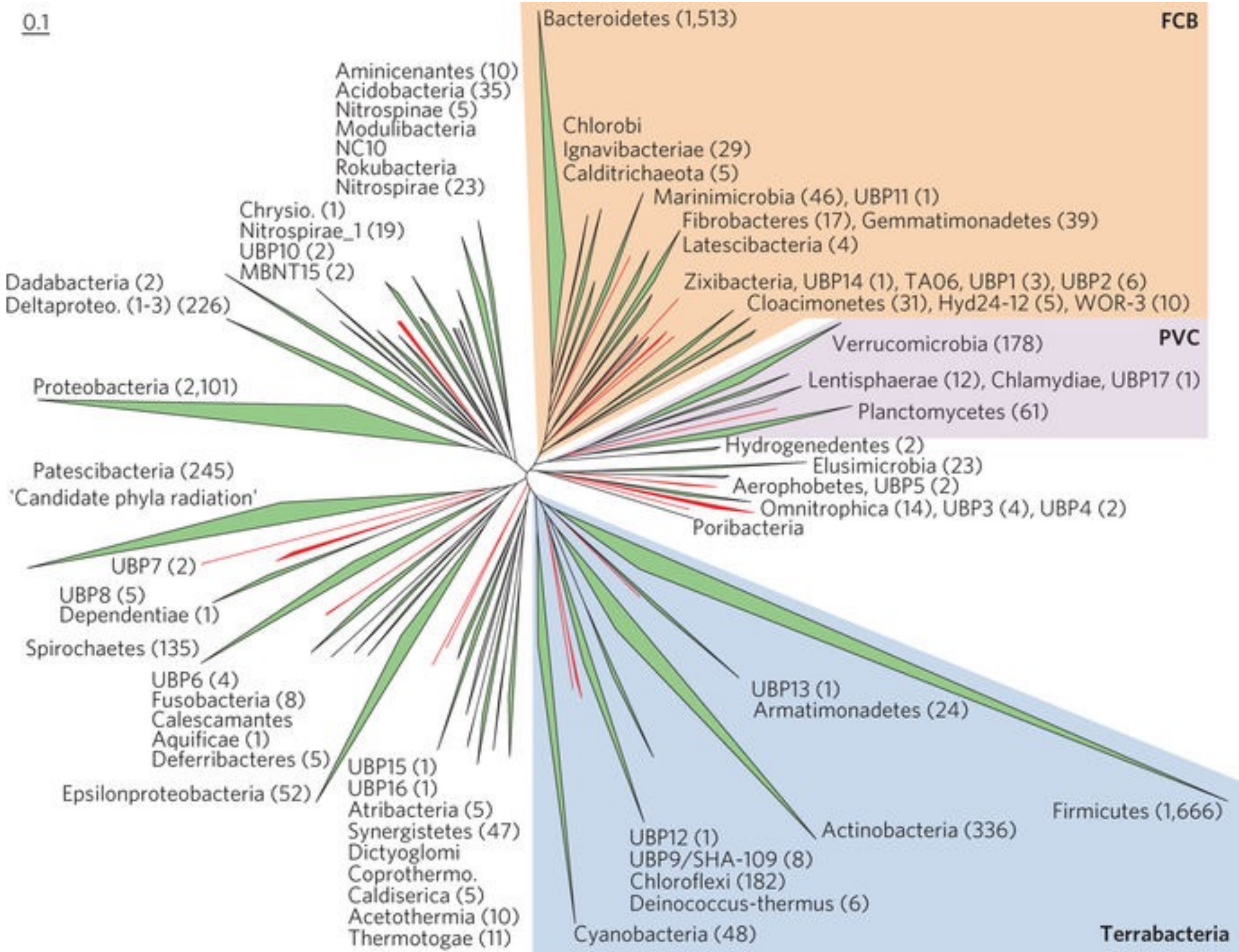


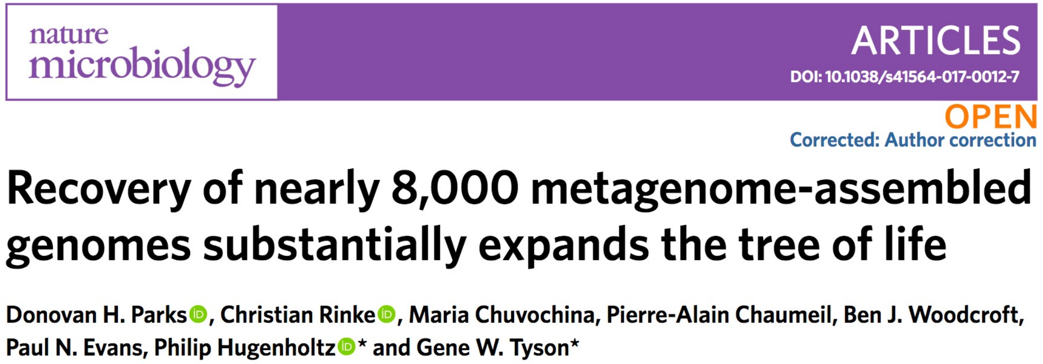
Parks et al. / *Nat Microbiol. 2017 Nov;2(11):1533-1542.*

# Example of important outcome from MAGs

Analysed 1500 metagenomics datasets

First genomes from 17 bacterial phyla and 3 archaeal phyla




nature microbiology

ARTICLES

DOI: 10.1038/s41564-017-0012-7

OPEN

Corrected: Author correction

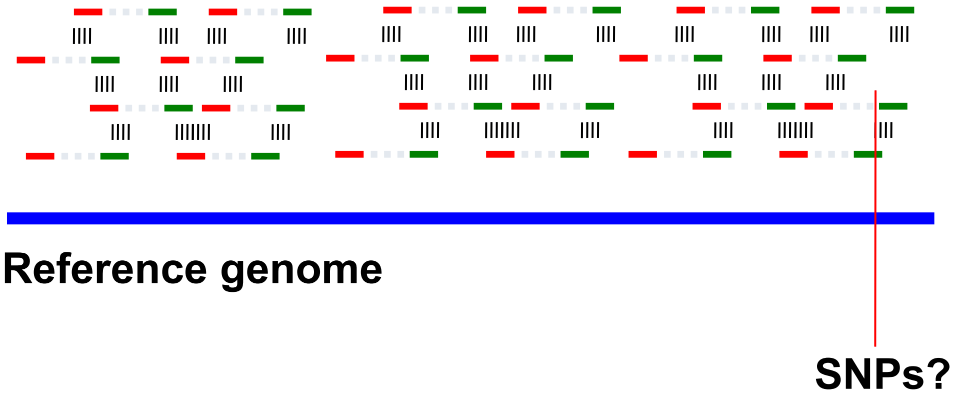
**Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life**

Donovan H. Parks, Christian Rinke, Maria Chuvochina, Pierre-Alain Chaumeil, Ben J. Woodcroft, Paul N. Evans, Philip Hugenholtz* and Gene W. Tyson*

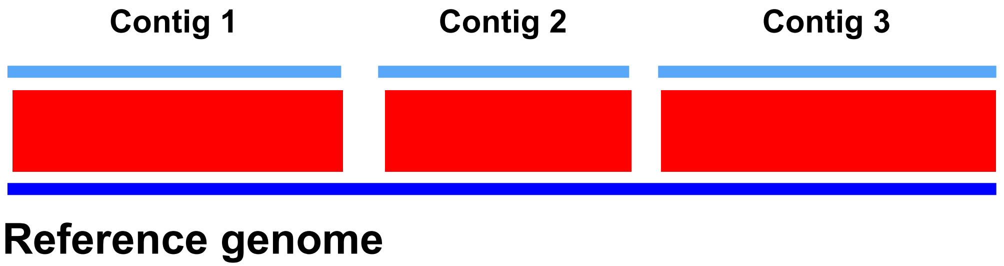
# Two slides on mapping against a reference sequence

Two methods for mapping against a reference sequence

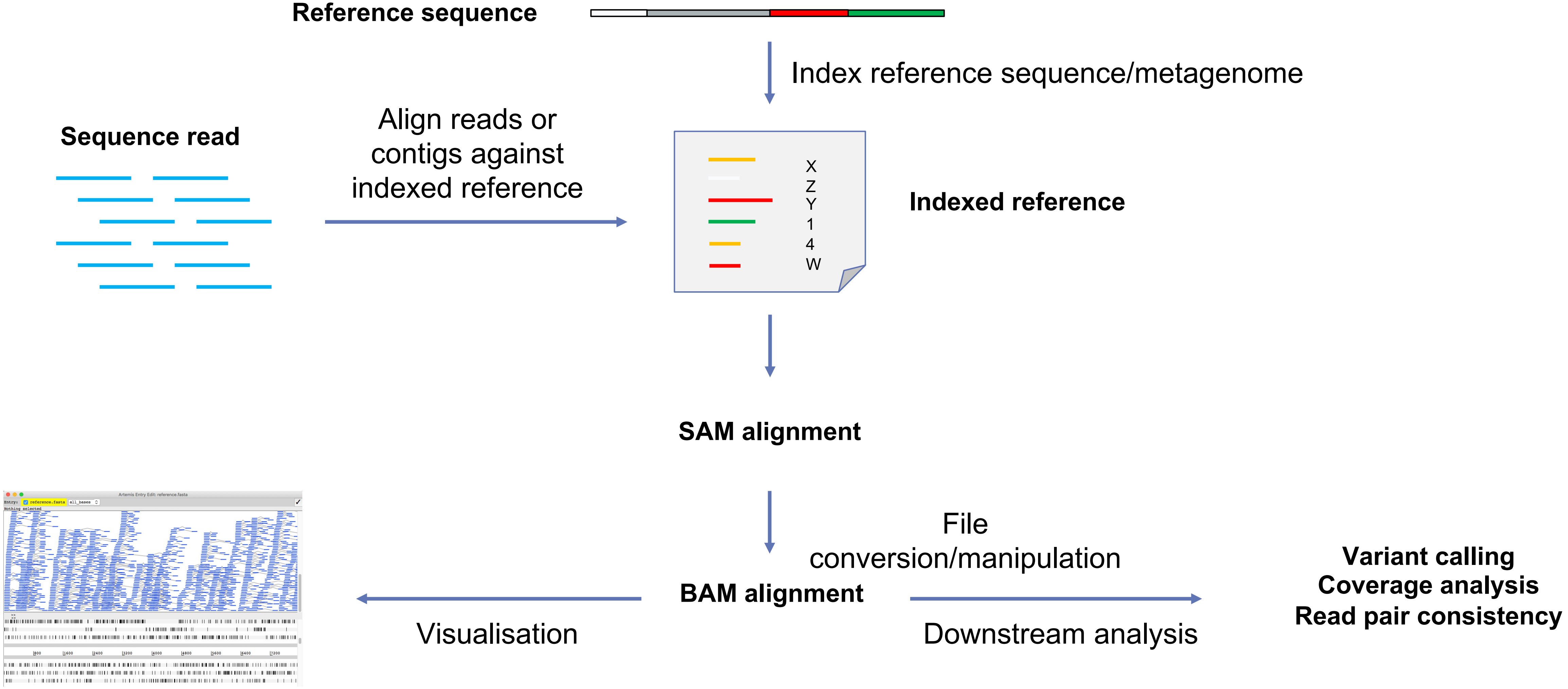
## Aligning reads against reference



## Aligning contigs against reference



# Overall read mapping process:



Reference sequence

Index reference sequence/metagenome

Sequence read

Align reads or contigs against indexed reference

Indexed reference

SAM alignment

File conversion/manipulation

BAM alignment

Visualisation

Downstream analysis

Variant calling
Coverage analysis
Read pair consistency