

# DAWG Workshop 3

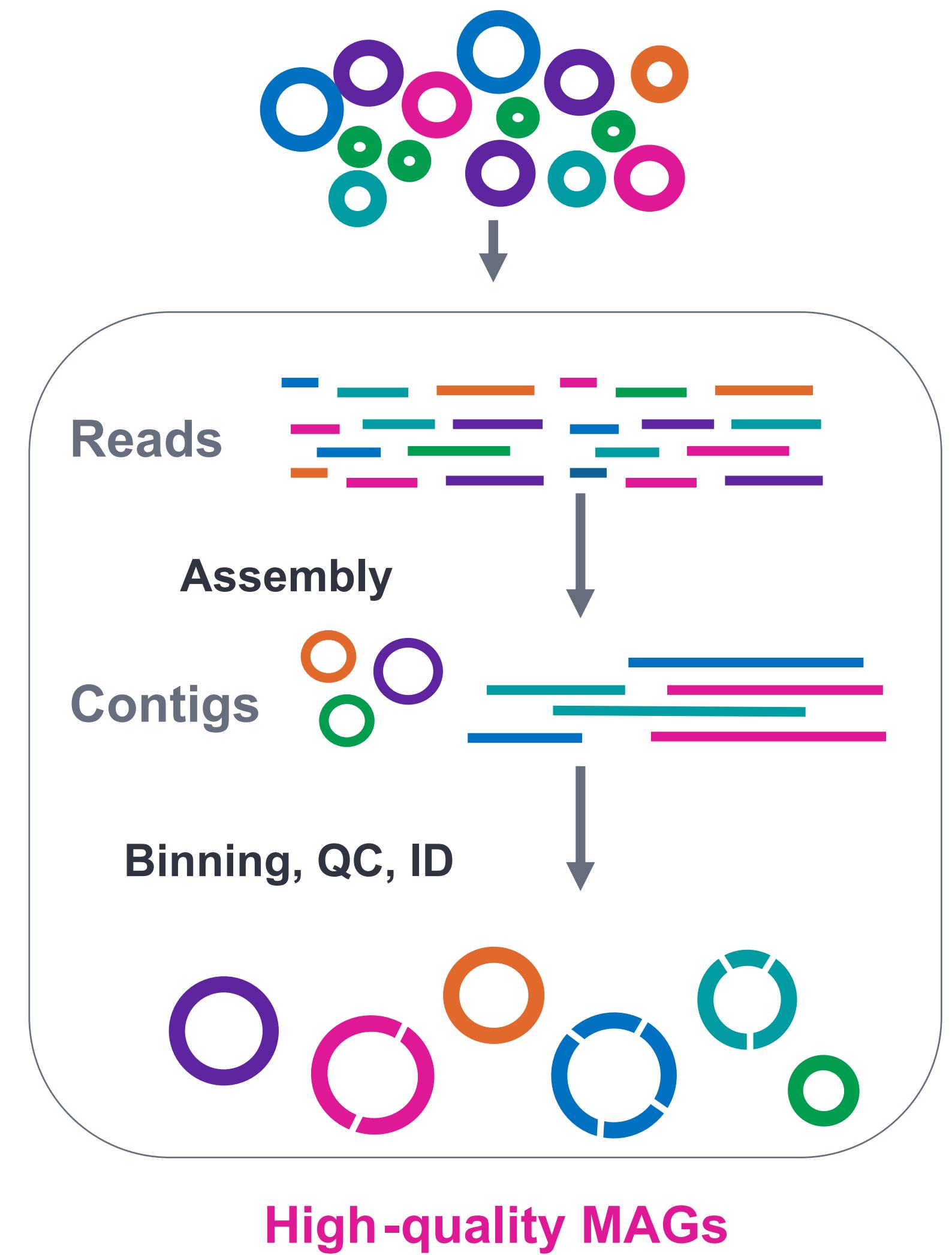
Metagenomic Sequencing  
Tree of Life Workshop Series

Dec 5<sup>th</sup>, 2025



# Workshop Pipeline

- QUALITY CONTROL (QC)
- ASSEMBLY
- MAPPING & COVERAGE
- BINNING
- QUALITY ASSESSMENT
- ANNOTATION & VISUALIZATION



QC

FastP  

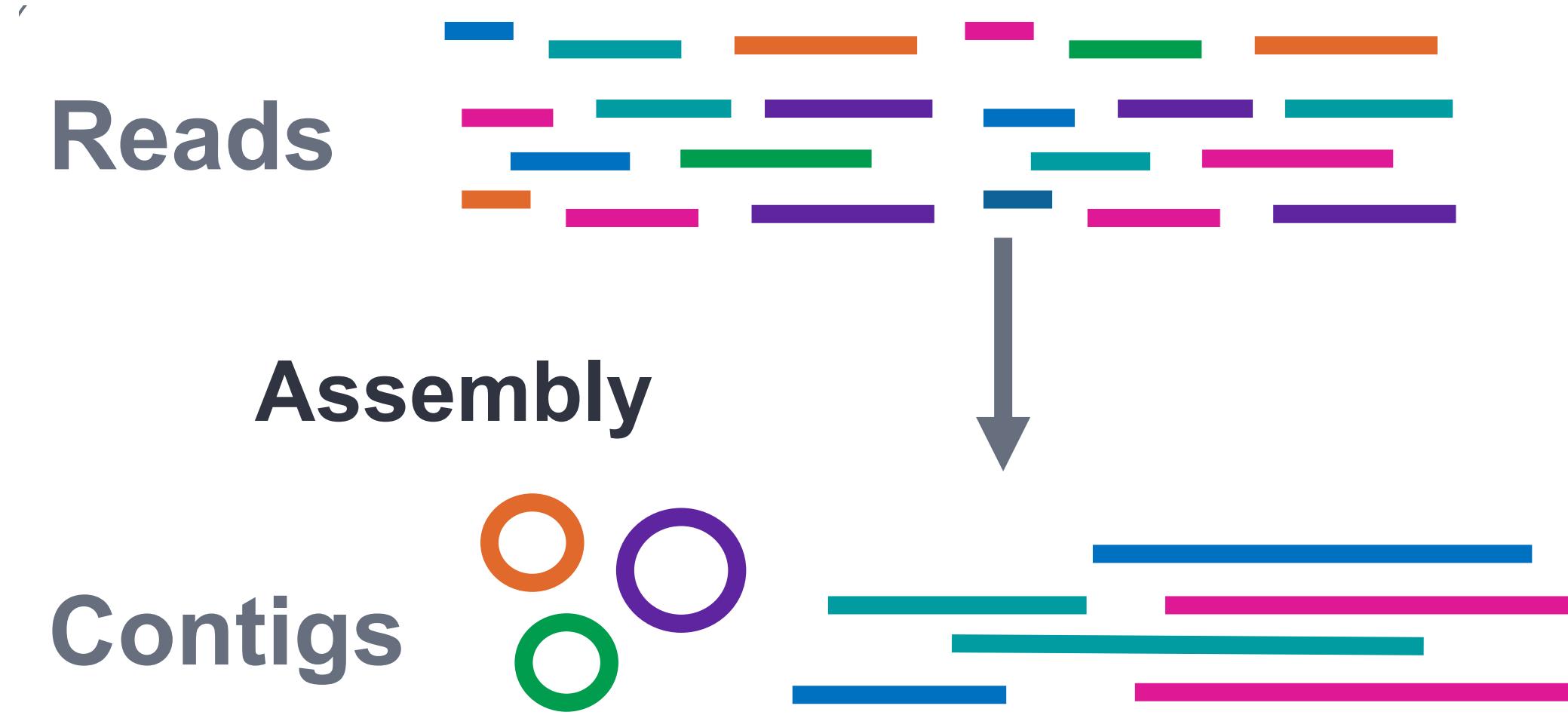

- Used for read trimming and quality control
- Removes adapter sequences
- Trims low-quality bases
- Removes reads that are too short

# ASSEMBLY

MEGAHIT

INPUT: Raw reads (FastQ files)

OUTPUT: Assembled contigs (Fasta file)



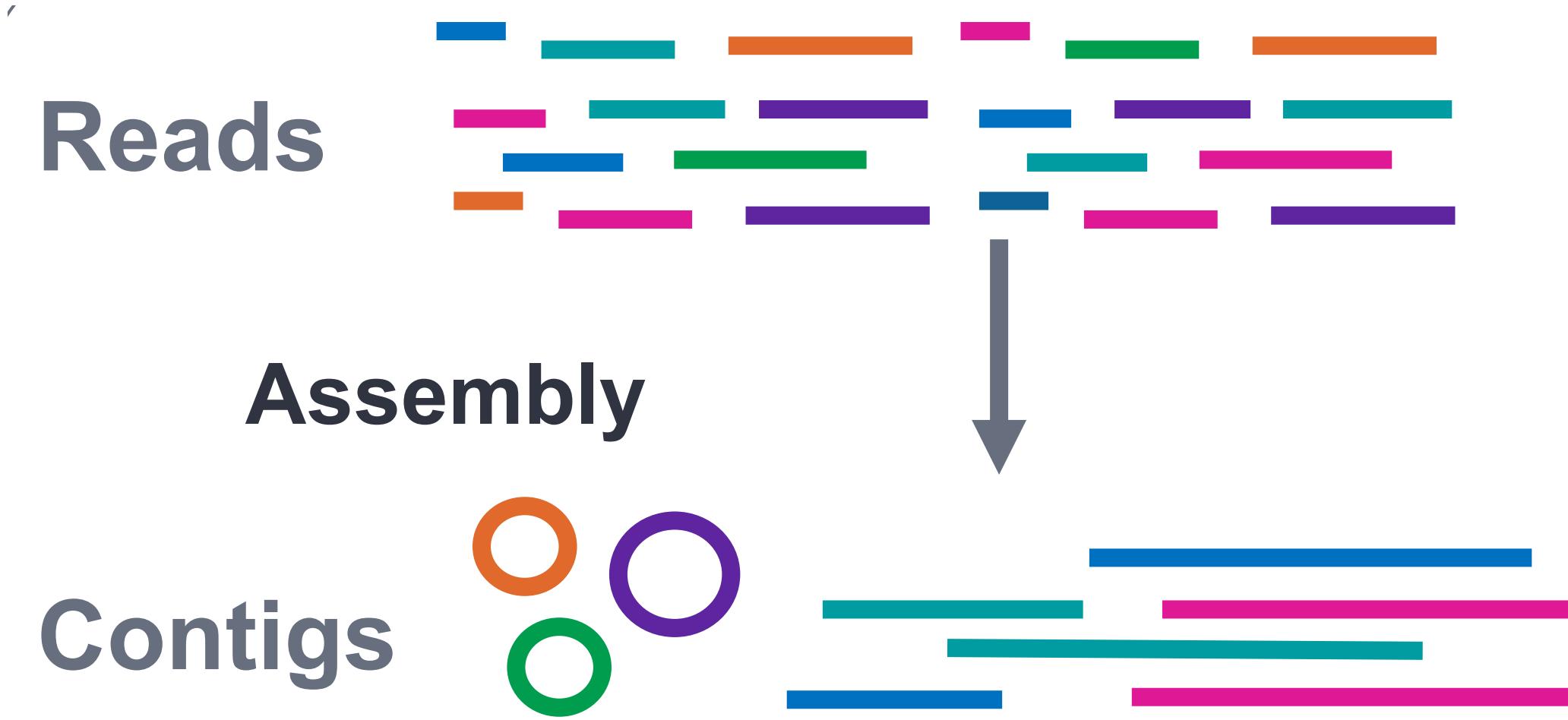
# ASSEMBLY

MEGAHIT

INPUT: Raw reads (FastQ files)

OUTPUT: Assembled contigs (Fasta file)

Longer contiguous  
sequences



## MAPPING

Bowtie2

INPUT: Raw reads (FastQ files) &  
Assembled contigs (Fasta files)

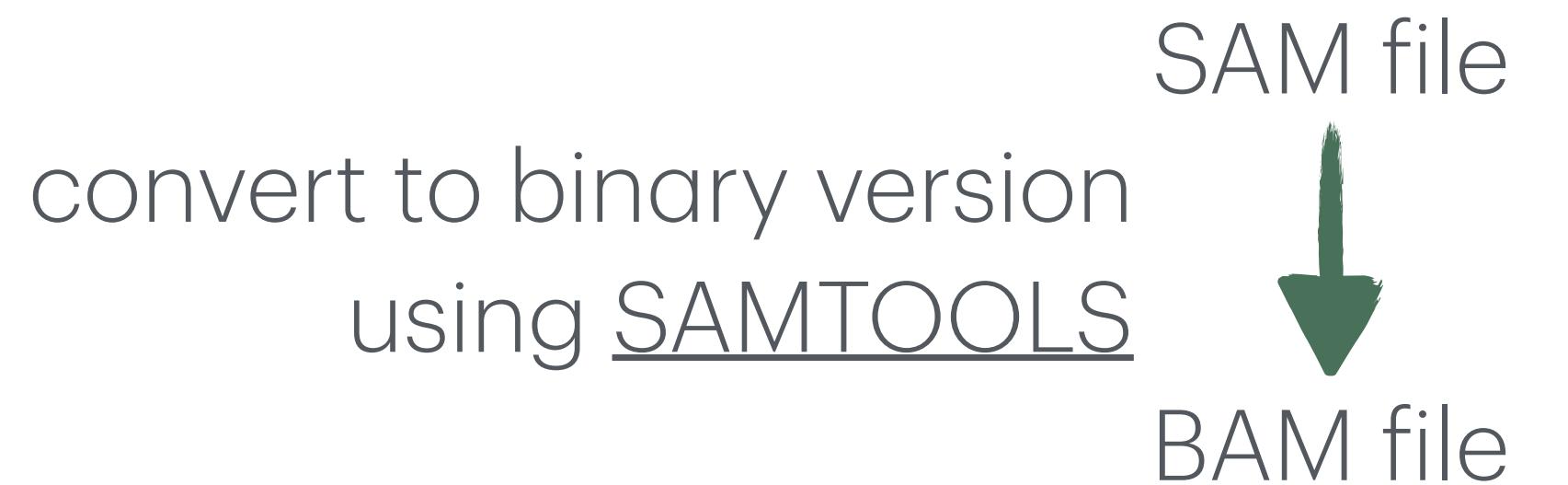
OUTPUT: Aligned reads (BAM  
file) and a subsequent depth/  
coverage file

## MAPPING

Bowtie2

INPUT: Raw reads (FastQ files) &  
Assembled contigs (Fasta files)

OUTPUT: Aligned reads (BAM  
file) and a subsequent depth/  
coverage file



# MAPPING

## Bowtie2

INPUT: Raw reads (FastQ files) & Assembled contigs (Fasta files)

OUTPUT: Aligned reads (BAM file) and a subsequent depth/coverage file

SAM file  
convert to binary version  
using SAMTOOLS  
BAM file

**A**

Coor	12345678901234	20	30	40
ref	AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGGCCAT			
+r001/1	TTAGATAAAGGATA*CTG			
+r002	aaaAGATAA*GGATA			
+r003	gcctaAGCTAA			
+r004	ATAGCT.....TCAGC			
-r003	ttagctTAGGC			
-r001/2	CAGCGGCAT			

**B**

@HD VN:1.5 SO:coordinate		Header section		QUAL (read quality; * meaning such information is not available)	
@SQ SN:ref LN:45					
r001	99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *				
r002	0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAAGGATA *				
r003	0 ref 9 30 5S6M * 0 0 GCCTAACGCTAA * SA:Z:ref,29,-,6H5M,17,0;				
r004	0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *				
r003	2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;				
r001	147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1				

Annotations for Panel B:

- Header section:** @HD VN:1.5 SO:coordinate, @SQ SN:ref LN:45
- Alignment section:** The main body of the SAM file.
- Optional fields in the format of TAG:TYPE:VALUE:** QUAL, NM:i:1, SA:Z:ref,29,-,6H5M,17,0; SA:Z:ref,9,+,5S6M,30,1;
- Fields explained:**

  - QNAME:** (query template name, aka. read ID)
  - FLAG:** (indicates alignment information about the read, e.g. paired, aligned, etc.)
  - RNAME:** (reference sequence name, e.g. chromosome /transcript id)
  - POS:** (1-based position)
  - MAPQ:** (mapping quality)
  - CIGAR:** (summary of alignment, e.g. insertion, deletion)
  - RNEXT:** (reference sequence name of the primary alignment of the NEXT read; for paired-end sequencing, NEXT read is the paired read; corresponding to the RNAME column)
  - PNEXT:** (Position of the primary alignment of the NEXT read in the template; corresponding to the POS column)
  - TLEN:** (the number of bases covered by the reads from the same fragment. In this particular case, it's 45 - 7 + 1 = 39 as highlighted in Panel A). Sign: plus for leftmost read, and minus for rightmost read
  - SEQ:** (read sequence)

## COVERAGE

### jgi\_summarize\_bam\_contig\_depths

---

#### **jgi\_summarize\_bam\_contig\_depths**

Script used to calculate the depth of each contig across multiple BAM files.

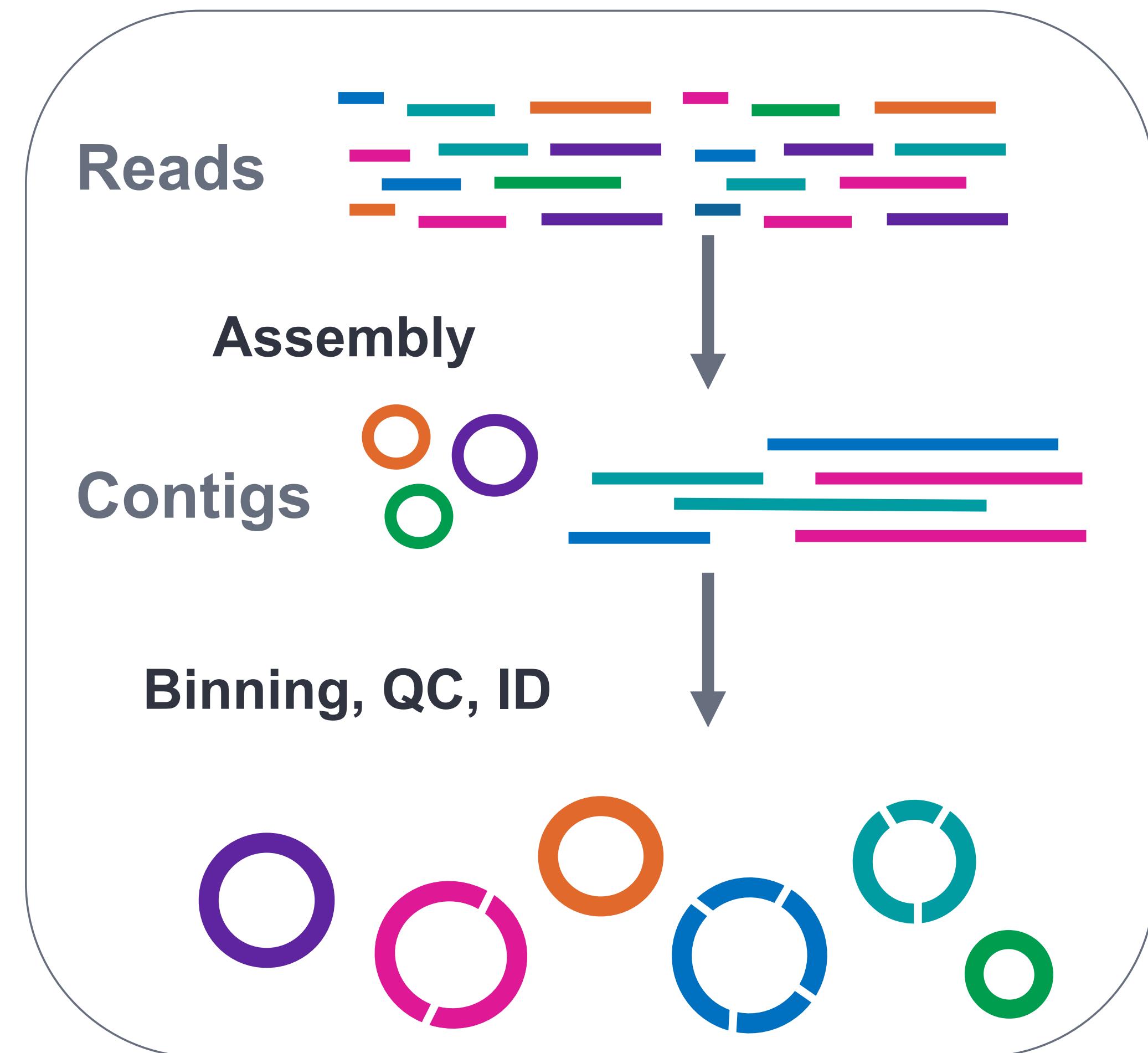
It determines the average number of reads that align to each contig.

This depth information is crucial for binning -> grouping contigs into potential genome bins based on coverage similarity. Input for **METABAT**.

## BINNING

METABAT2

INPUT: Contigs (FASTA files) and coverage/depth information

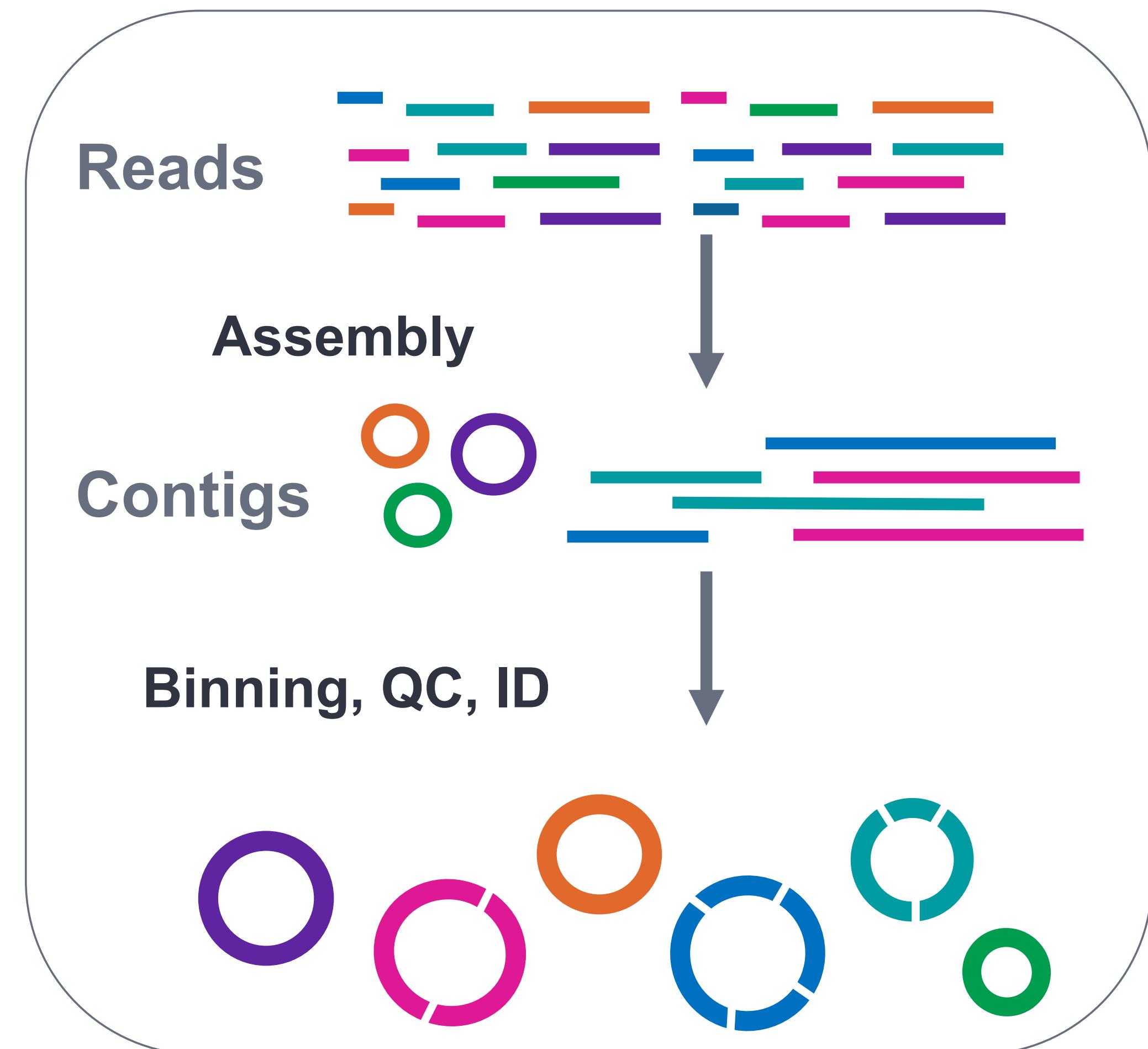


## BINNING

## METABAT2

INPUT: Contigs (FASTA files) and coverage/depth information

OUTPUT: Draft MAGs (multiple FASTA files, one for each "bin")



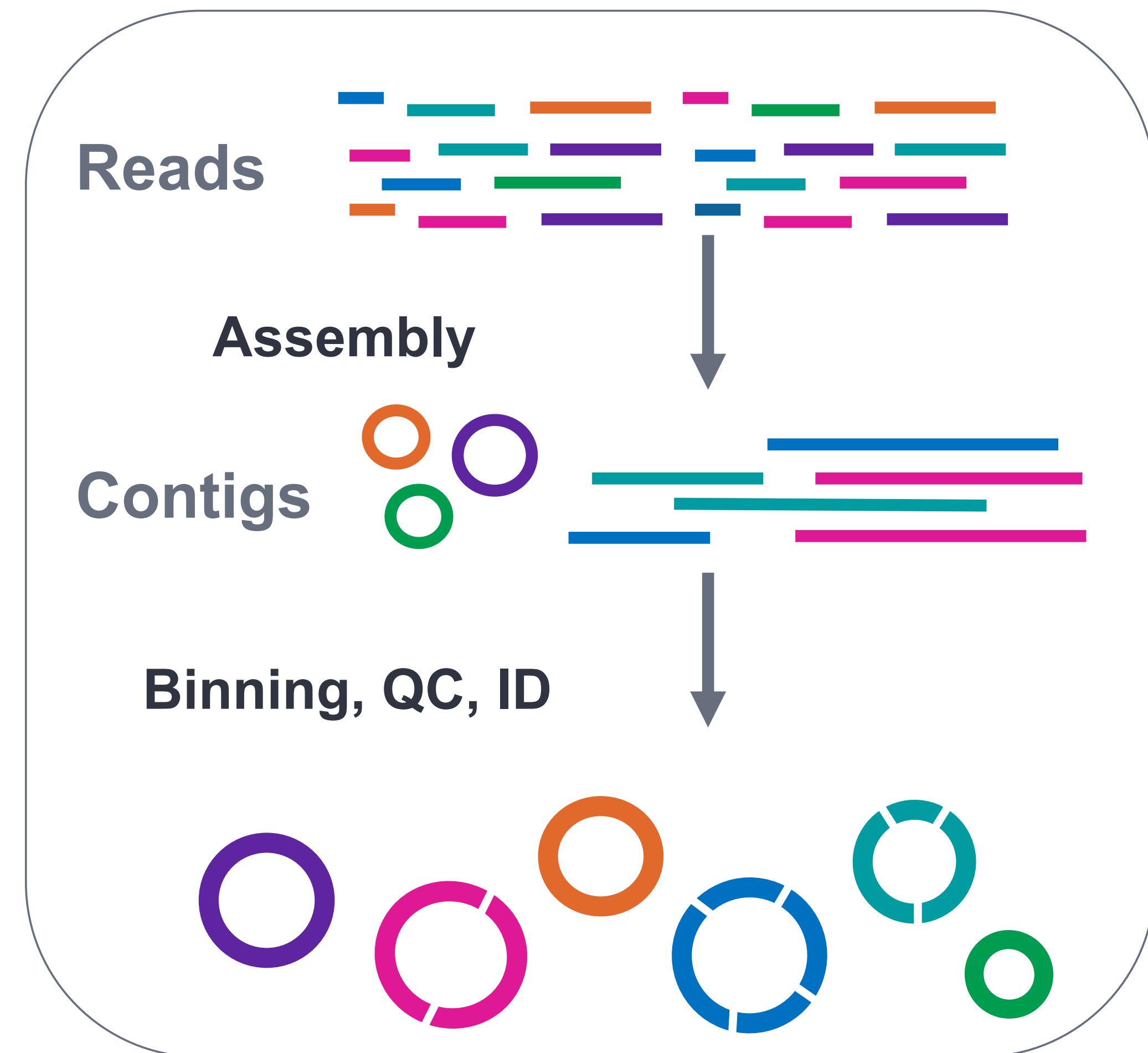
# BINNING

INPUT: Contigs (FASTA files) and coverage/depth information

OUTPUT: Draft MAGs (multiple FASTA files, one for each "bin")

AIM: group contigs into species

# METABAT2



QUALITY ASSESSMENT

CheckM



INPUT: Draft MAGs/bins (FASTA files)

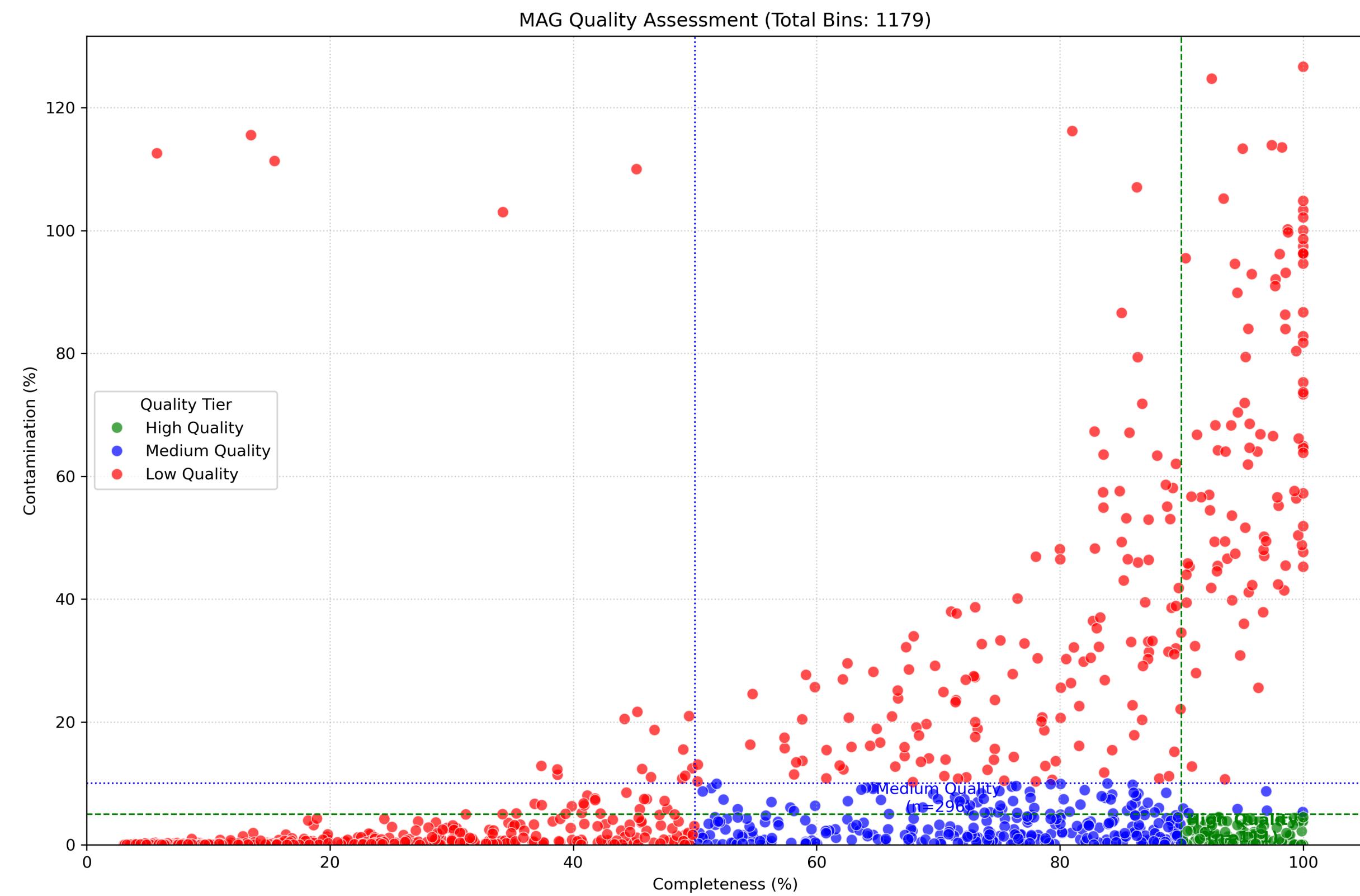
OUTPUT: Summary table with quality scores.

# QUALITY ASSESSMENT

CheckM

INPUT: Draft MAGs/bins (FASTA files)

OUTPUT: Summary table with quality scores.



- ☒ ANNOTATION and VISUALIZATION

- ✓ ANNOTATION and VISUALIZATION

BAKTA & Proksee



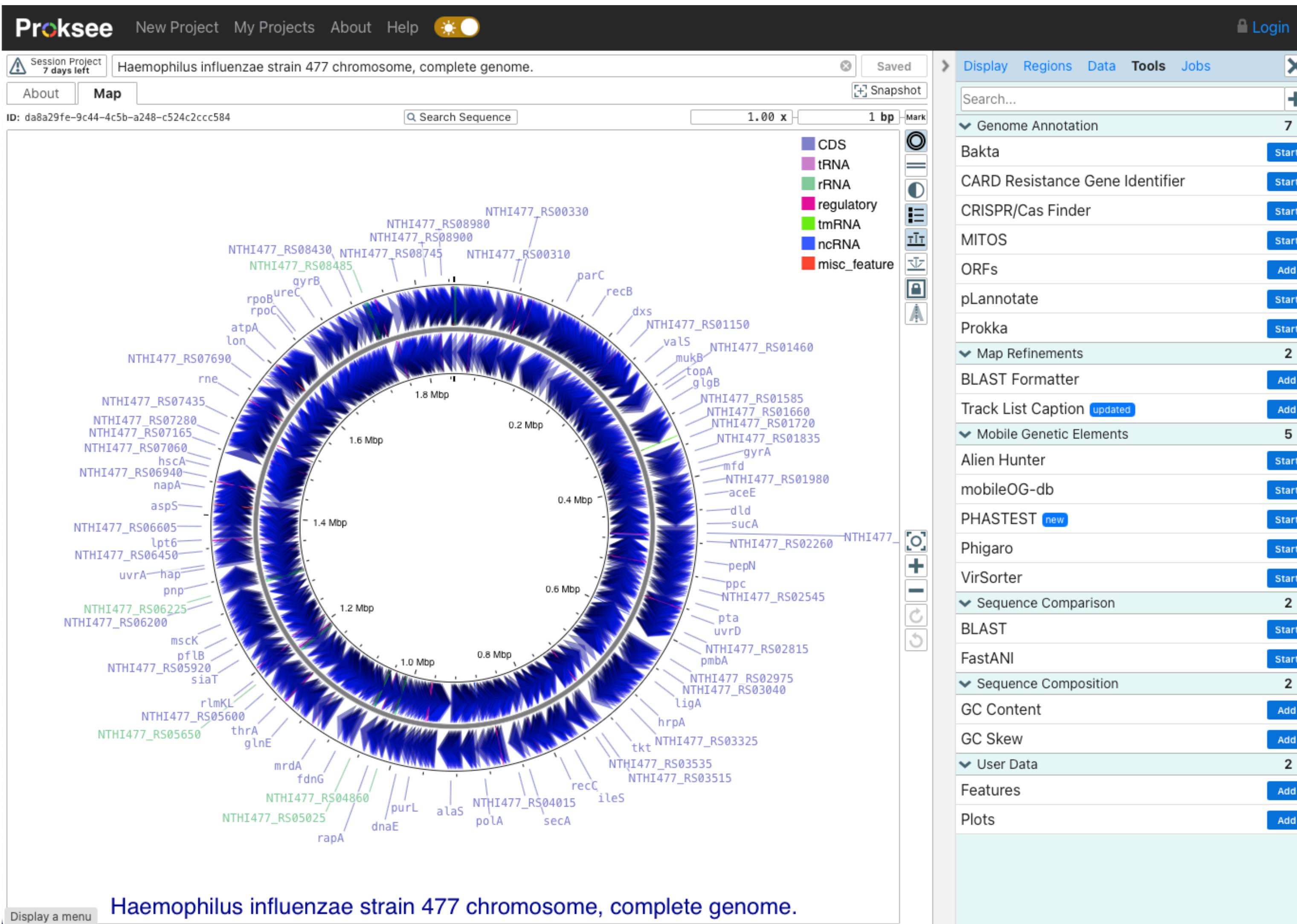
- ✓ ANNOTATION and VISUALIZATION

BAKTA & Proksee



INPUT: High-quality MAGs (FASTA files)

# ✓ ANNOTATION and VISUALIZATION



# BAKTA & Proksee

INPUT: High-quality MAGs (FASTA files)

- ☒ ANNOTATION & VISUALIZATION

ANNOTATION & VISUALIZATION

Anvi'o



## ANNOTATION & VISUALIZATION

Anvi'o

- Check bin quality
- Assembles reads into contigs
- Annotates MAGs
- Explore data
- Allows users to manually curate and refine genome bins and performs various other 'omics analyses, including comparative genomics and phylogenomics

# ✓ ANNOTATION & VISUALIZATION

Anvi'o

- Check bin quality
- Assembles reads into contigs
- Annotates MAGs
- Explore data
- Allows users to manually curate and refine genome bins and performs various other 'omics analyses, including comparative genomics and phylogenomics

