

**Metagenomics  
Summer School 2020**

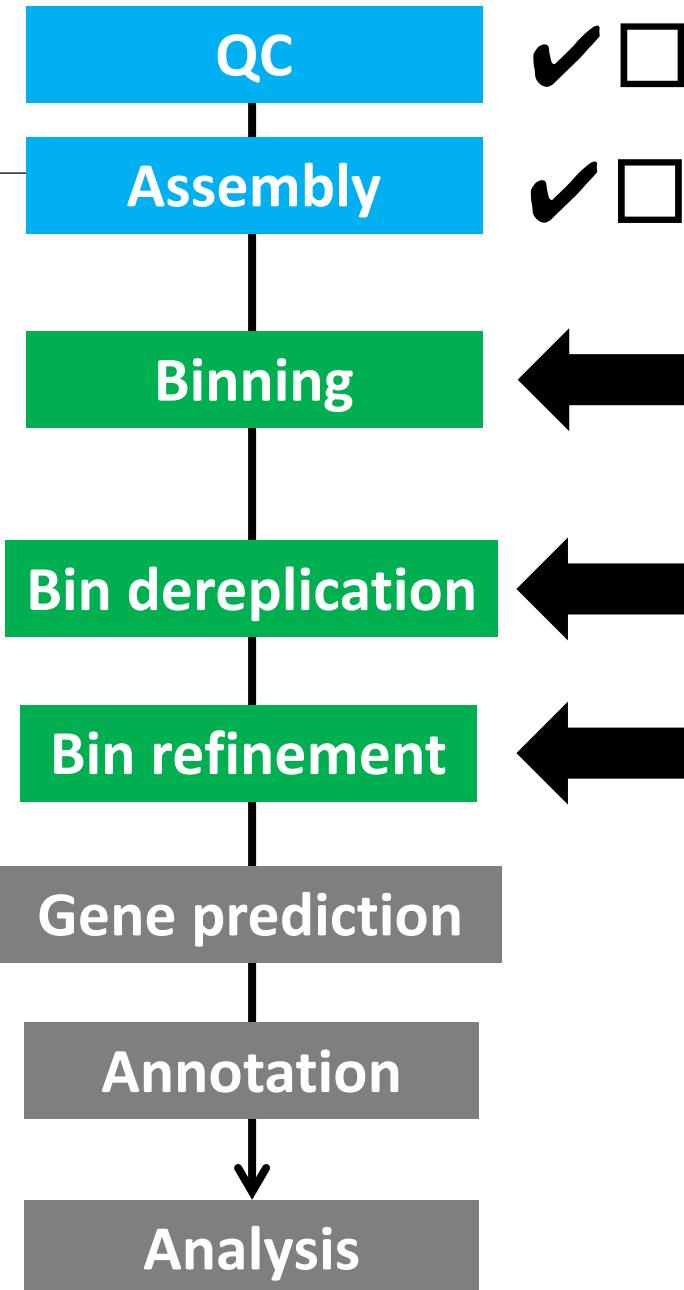


# **Day 2**

**Binning, binning, and more binning!**

# Day overview

- Same time format as day 1
- Goals:
  - Evaluate assembly results
  - Bin contigs
  - Bin refinement



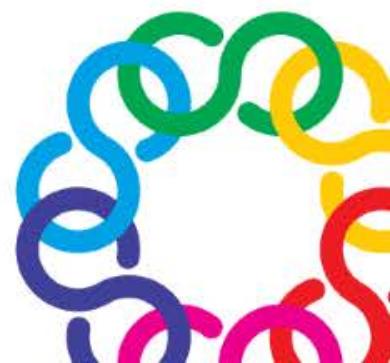
# Binning



# Overview of binning

---

- Key parameters
- History
- Strategies for binning



# Overview of binning

---

## Key parameters:

- G+C content (compositional signature)
- Tetranucleotide frequencies (composition signature)

**ATCG TCGG CGGC**

- Genome coverage
- Differential coverage
- Phylogenetic affiliation of:
  - predicted proteins in a contig
  - core genes within a bin



# Tetranucleotide frequencies

---

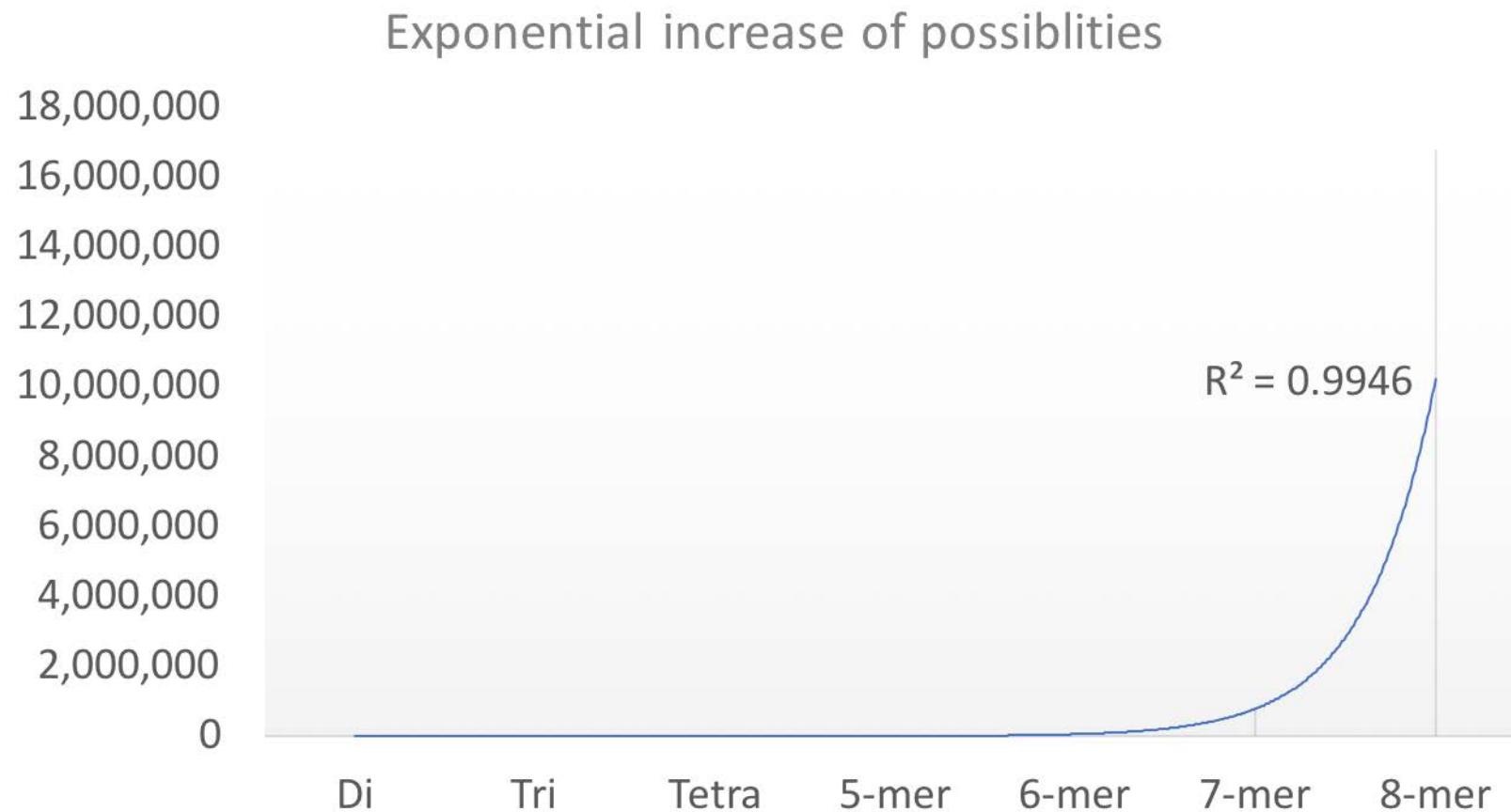
- Tend to be conserved within genomes
- Can use 2 to 8 mers to detect genomic signatures
- Longer oligomers have increased sensitivity
- Longer oligomers require longer lengths of sequence for sampling saturation
- The number of possibilities rises exponentially with length

Tetra-SOM of prokaryotic and eukaryotic genomes artificially combined, yield inter and intra species separation (Abe et al., 2003, Genome Research).

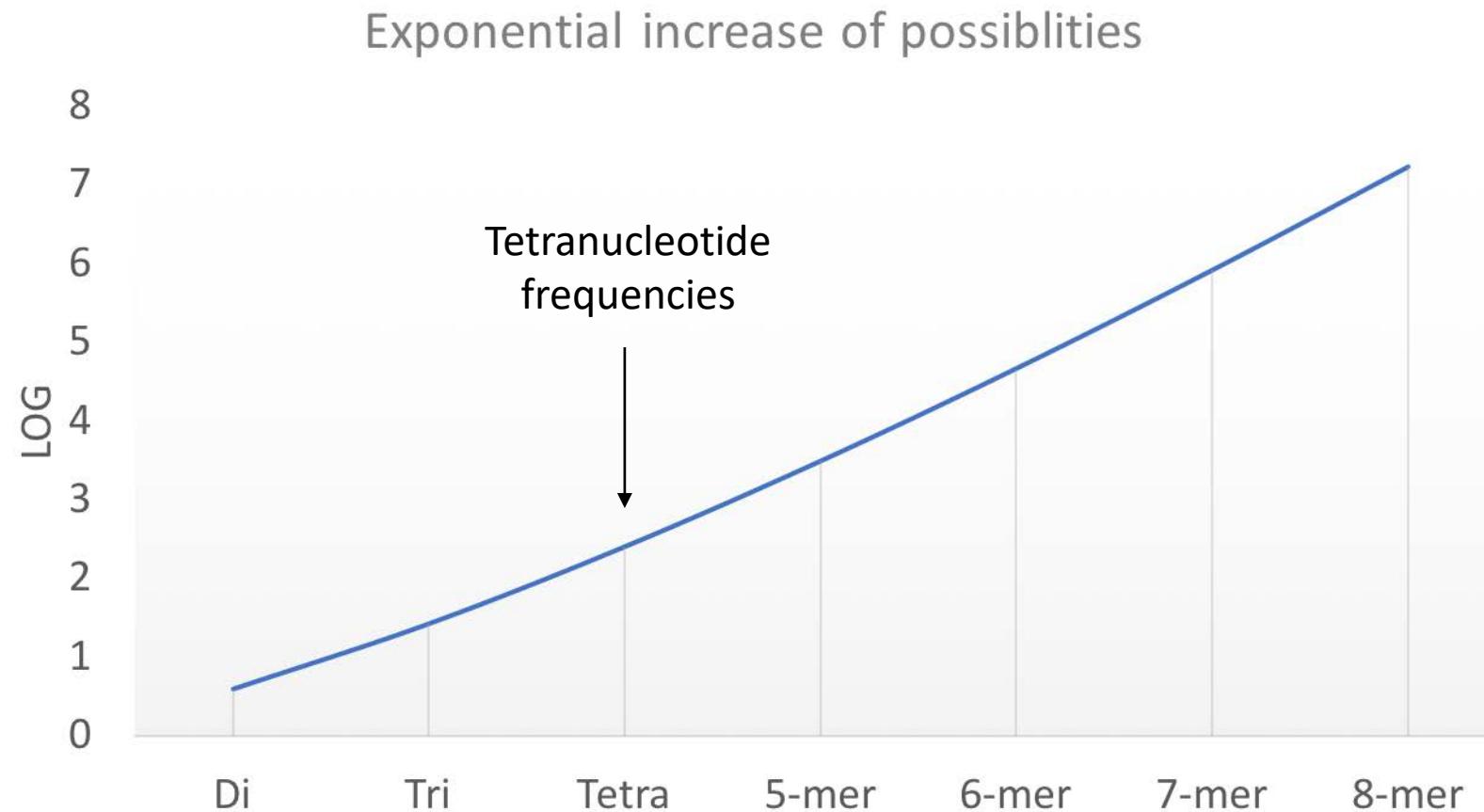


# Discrimination with different $k$ -mers

---



# Discrimination with different $k$ -mers



# Possible tetranucleotide freqs

---

$4 \times 4 \times 4 \times 4 = 256$  possible tetranucleotides in DNA

16 are palindromic (e.g. reverse complement of AATT is AATT)

240 include distinct reverse complement sequences (e.g. AAAA = TTTT)

(240 / 2) + (16) = 136 unique tetranucleotides (when removing reverse complements with distinct sequences)



# Tetranucleotide frequencies

---

## Procedure:

- Fragment contigs to uniform lengths (e.g. 5 kb), use a sliding window for calculations, or normalized frequencies by contig length
- Count fragment tetranucleotide frequencies for forward or reverse complement



# Tetranucleotide frequencies

---

ATCGGCTAA



# Tetranucleotide frequencies

---

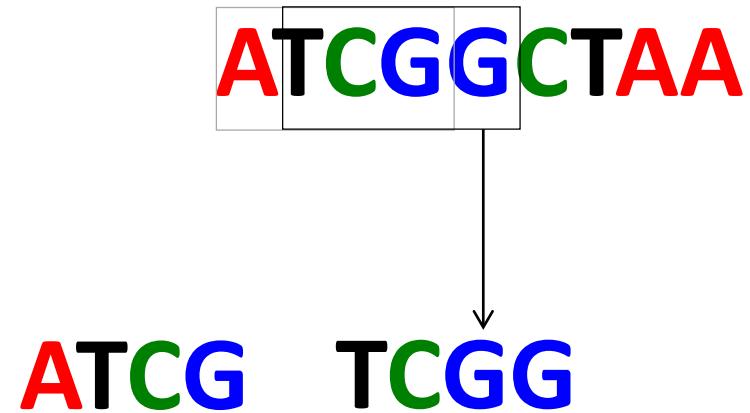
ATCGGGCTAA

ATCG



# Tetranucleotide frequencies

---



# Tetranucleotide frequencies

---

The diagram illustrates the extraction of tetranucleotides from a sequence. At the top, the sequence **ATCGGCTAA** is shown in black text. Above the second nucleotide, 'C', there is a small grid consisting of four empty white boxes stacked vertically. An arrow points from this grid to the sequence below. The sequence is broken down into three tetranucleotides: **ATCG**, **TCGG**, and **CGGC**. The first two nucleotides of each tetranucleotide are colored red, while the last two are blue. The third nucleotide of each tetranucleotide is green.

**ATCGGCTAA**

**ATCG   TCGG   CGGC**



# Tetranucleotide frequencies

---

The diagram illustrates the extraction of tetranucleotides from a sequence. At the top, the sequence **ATCGGCTAA** is shown in black text. Above the second nucleotide ('C'), there is a small grid consisting of four empty white boxes arranged horizontally. An arrow points from this grid down to the sequence below. The sequence is then broken down into three tetranucleotides: **ATCG**, **TCGG**, and **CGGC**. Below each of these three tetranucleotides is the number **1**, indicating their frequency or count.

**ATCGGCTAA**

**ATCG   TCGG   CGGC**

1        1        1



# Possible tetranucleotide freqs

Example for ESOM using esomana software

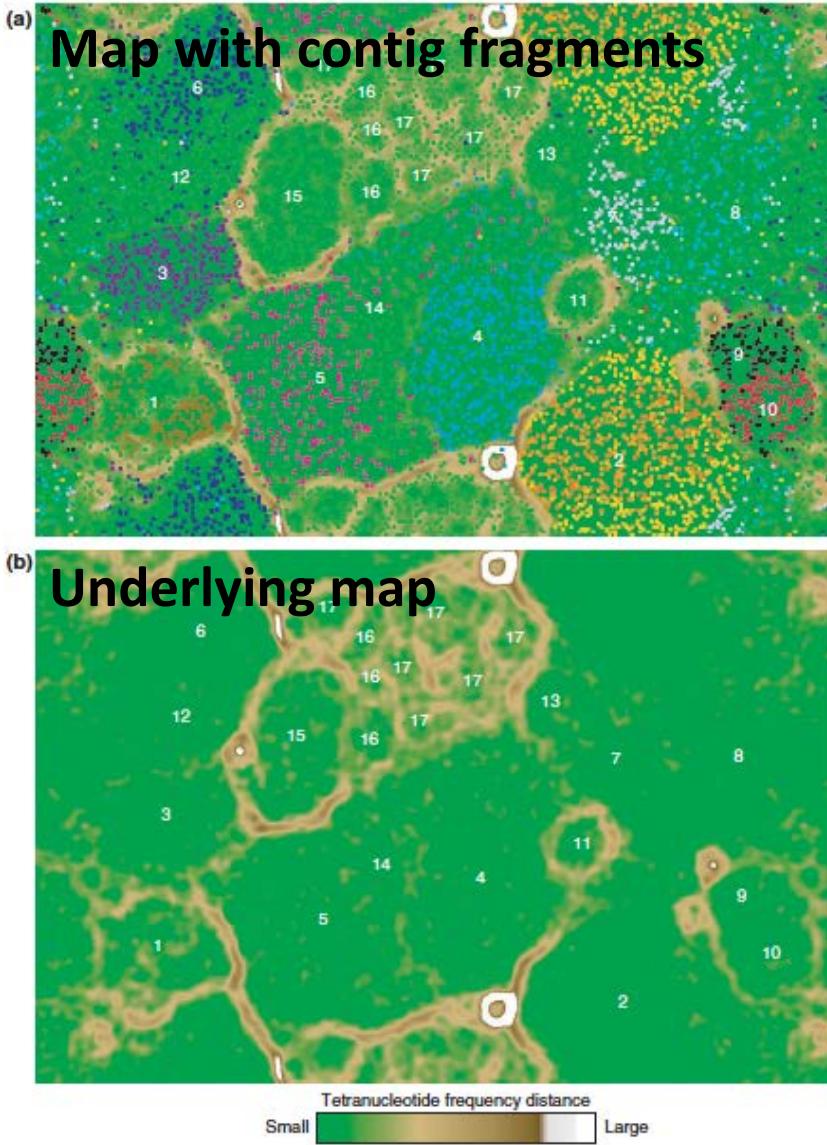
	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	% Key	AAAA	AAAT	AAAC	AAAG	AATA	AATT	AATC	AATG	AACA	AACT	AACC	AACG	AAG
2	1	0.0225338	0.01151728	0.01402103	0.01552328	0.00751127	0.00350526	0.01452178	0.00901352	0.01101652	0.00550826	0.00751127	0.00751127	0.00
3	2	0.01452178	0.01502253	0.01402103	0.01301953	0.00901352	0.00751127	0.00751127	0.01352028	0.01301953	0.00851277	0.00751127	0.00650976	0.00
4	3	0.01702554	0.01301953	0.01802704	0.01001502	0.01352028	0.00650976	0.00701052	0.00650976	0.01101652	0.00851277	0.01201803	0.00951427	0.00
5	4	0.01652479	0.01101652	0.01051577	0.01452178	0.01352028	0.00400601	0.00751127	0.00901352	0.00951427	0.00951427	0.00951427	0.00650976	0.00
6	5	0.00801202	0.01101652	0.01001502	0.00951427	0.00901352	0.00300451	0.00801202	0.01151728	0.00901352	0.00901352	0.00901352	0.00500751	0.00
7	6	0.01251878	0.01352028	0.01151728	0.01051577	0.01201803	0.00300451	0.00851277	0.01001502	0.01402103	0.00250376	0.01101652	0.00801202	0.00
8	7	0.01402103	0.01101652	0.01251878	0.01802704	0.01151728	0.00150225	0.01251878	0.01402103	0.01151728	0.00400601	0.00901352	0.00600901	0.01
9	8	0.01402103	0.01101652	0.01001502	0.01352028	0.00851277	0.00500751	0.00801202	0.01251878	0.01201803	0.00500751	0.01001502	0.00901352	0.00
10	9	0.01552328	0.01301953	0.01402103	0.01301953	0.00801202	0.00450676	0.01051577	0.01201803	0.00550826	0.00400601	0.01352028	0.01001502	0.01
11	10	0.01201803	0.00701052	0.01402103	0.01201803	0.00350526	0.00500751	0.00951427	0.01151728	0.00450676	0.00650976	0.01301953	0.01051577	0.01
12	11	0.02704056	0.01652479	0.01201803	0.01552328	0.00751127	0.00500751	0.00751127	0.01001502	0.00901352	0.00650976	0.01051577	0.00801202	0.01
13	12	0.01352028	0.01452178	0.01151728	0.01352028	0.00901352	0.00851277	0.01151728	0.01151728	0.01051577	0.00751127	0.00901352	0.00751127	0.01
14	13	0.01702554	0.01201803	0.01051577	0.01301953	0.00901352	0.00600901	0.00801202	0.01051577	0.00901352	0.00851277	0.01201803	0.00901352	0.00
15	14	0.03204807	0.02303455	0.01402103	0.01352028	0.01051577	0.00550826	0.01301953	0.01201803	0.00951427	0.00650976	0.00951427	0.00751127	0.00
16	15	0.01602404	0.01101652	0.01502253	0.01001502	0.00701052	0.00400601	0.00751127	0.01151728	0.01151728	0.00751127	0.01051577	0.00500751	0.00
17	16	0.01051577	0.01251878	0.01352028	0.00751127	0.01301953	0.00650976	0.00801202	0.00951427	0.01201803	0.00650976	0.00901352	0.00851277	0.00
18	17	0.01552328	0.01502253	0.01301953	0.01251878	0.01201803	0.00500751	0.01101652	0.00801202	0.01101652	0.00600901	0.01151728	0.00450676	0.00
19	18	0.0225338	0.01952929	0.01151728	0.01151728	0.00650976	0.00600901	0.01101652	0.01251878	0.00801202	0.00650976	0.01101652	0.01151728	0.01
20	19	0.0235353	0.01452178	0.01502253	0.01602404	0.01001502	0.00150225	0.01301953	0.01301953	0.01001502	0.00801202	0.01201803	0.00951427	0.01
21	20	0.02203305	0.01352028	0.01101652	0.01251878	0.00801202	0.00400601	0.01001502	0.01402103	0.01151728	0.00550826	0.00901352	0.00901352	0.00
22	21	0.01852779	0.01402103	0.01352028	0.01301953	0.00650976	0.00500751	0.00851277	0.01101652	0.00851277	0.00600901	0.00951427	0.00751127	0.00
23	22	0.01502253	0.01101652	0.01352028	0.00901352	0.01151728	0.00550826	0.01101652	0.01352028	0.01101652	0.00801202	0.01201803	0.00450676	0.00
24	23	0.02403605	0.01652479	0.01051577	0.01702554	0.00901352	0.00650976	0.01101652	0.01352028	0.00650976	0.00801202	0.00550826	0.00851277	0.00
25	24	0.01452178	0.01402103	0.00951427	0.01151728	0.01301953	0.00801202	0.00851277	0.01051577	0.01151728	0.00951427	0.01352028	0.00500751	0.00
26	25	0.01802704	0.00951427	0.01201803	0.01251878	0.01301953	0.00550826	0.00550826	0.01001502	0.01201803	0.00901352	0.01051577	0.00851277	0.00
27	26	0.01301953	0.01201803	0.00851277	0.01552328	0.00751127	0.00500751	0.01101652	0.01352028	0.00951427	0.00801202	0.00650976	0.00751127	0.01
28	27	0.02954432	0.01902854	0.01301953	0.01552328	0.01402103	0.00701052	0.01101652	0.01301953	0.00951427	0.00801202	0.00751127	0.00400601	0.01
29	28	0.02103155	0.01502253	0.01301953	0.01502253	0.01352028	0.00701052	0.00951427	0.01151728	0.00600901	0.00751127	0.01101652	0.00851277	0.00



# ESOM and Tetranucleotide frequencies

(Dick et al., 2009, Genome Biology)

Acid mine drainage genomes

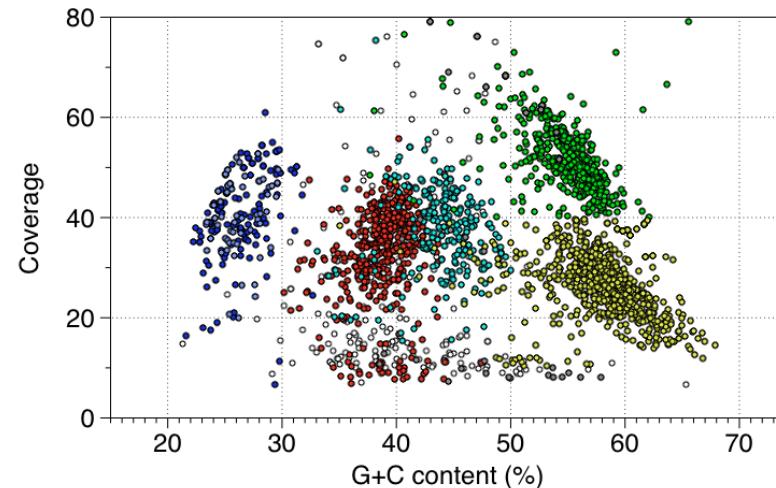


- Emergent Self Organizing Map (ESOM)
- Created using the Databionics tool esomana (<http://databionic-esom.sourceforge.net/>)
- Unsupervised neural network algorithm
- Allows outliers (e.g. partially sampled genomes)
- Determines clusters as supported by the data

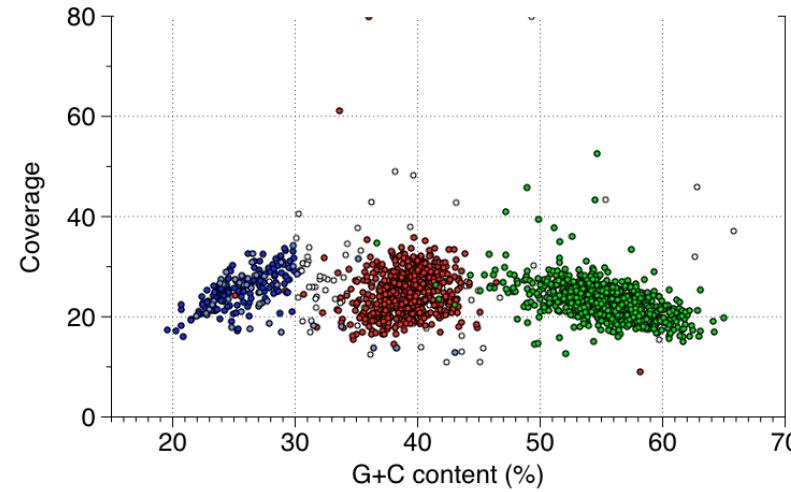




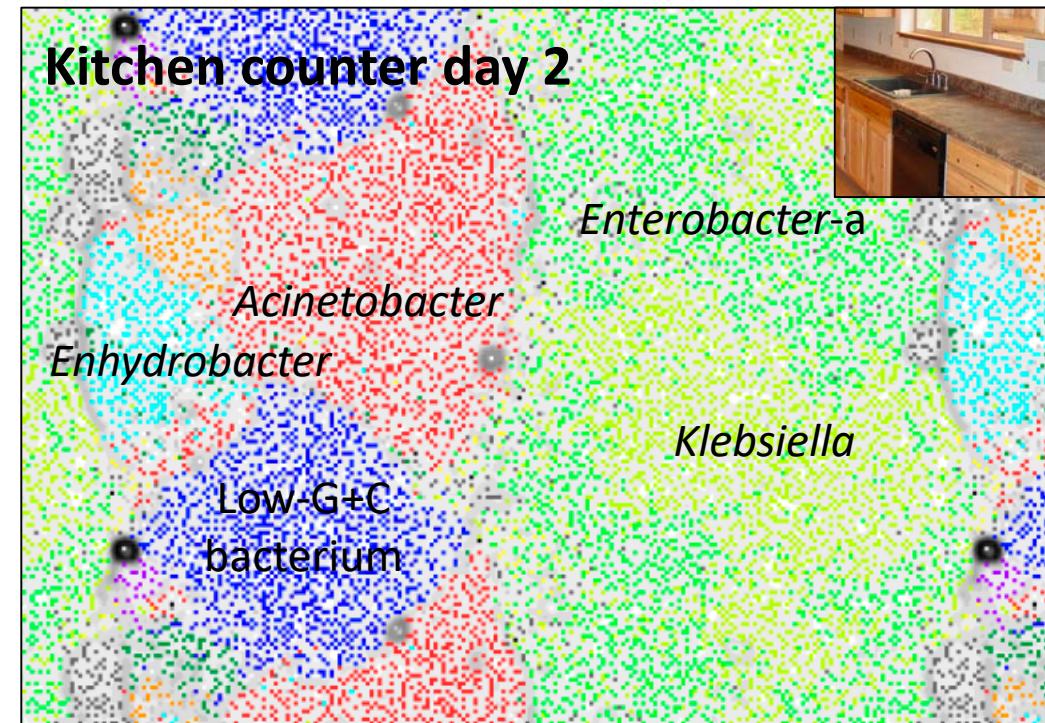
## Kitchen counter day 2



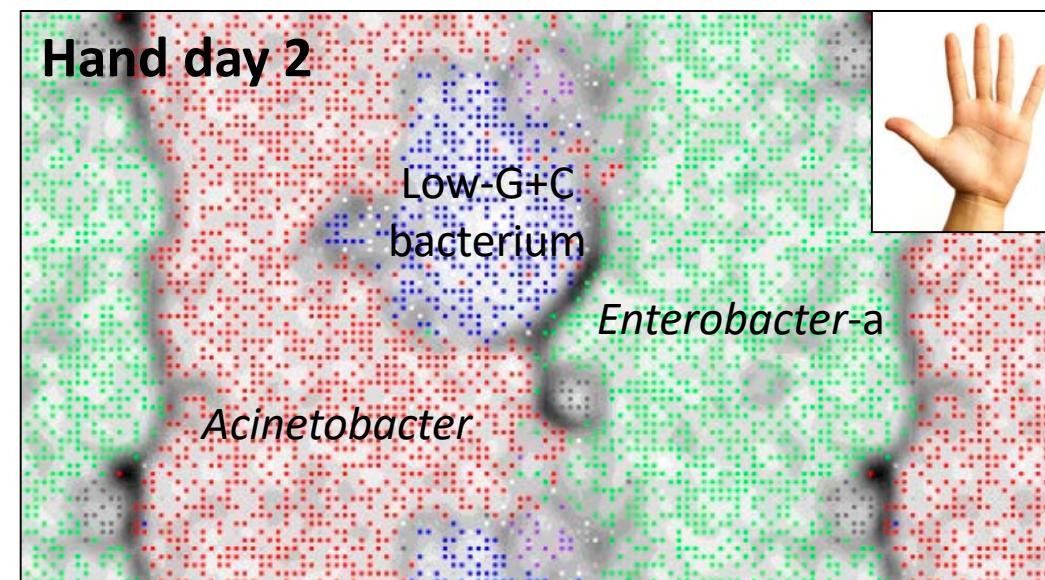
## Hand day 2



## Kitchen counter day 2



## Hand day 2

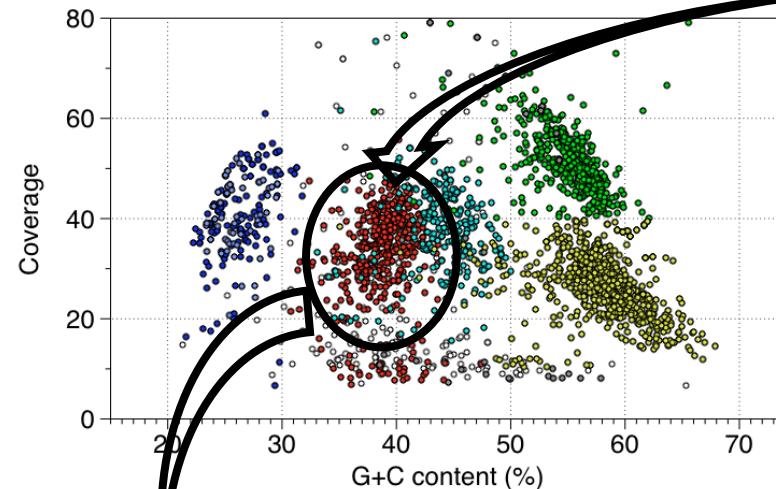


Cluster  
separation of  
distinct  
genera or  
species

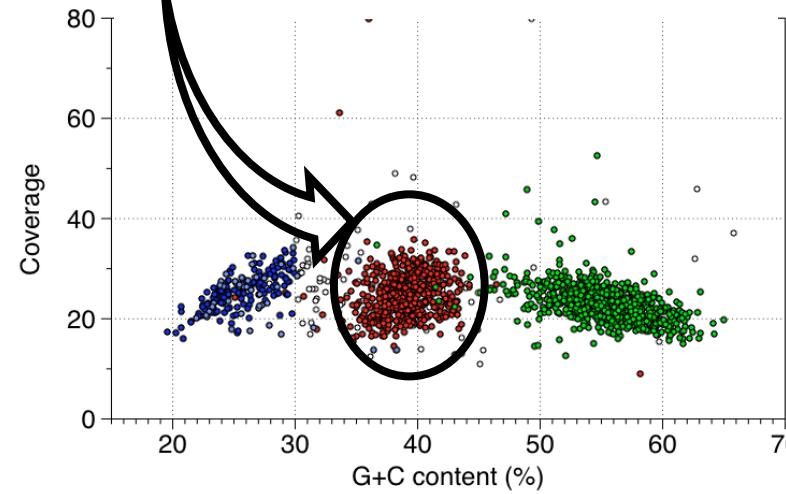




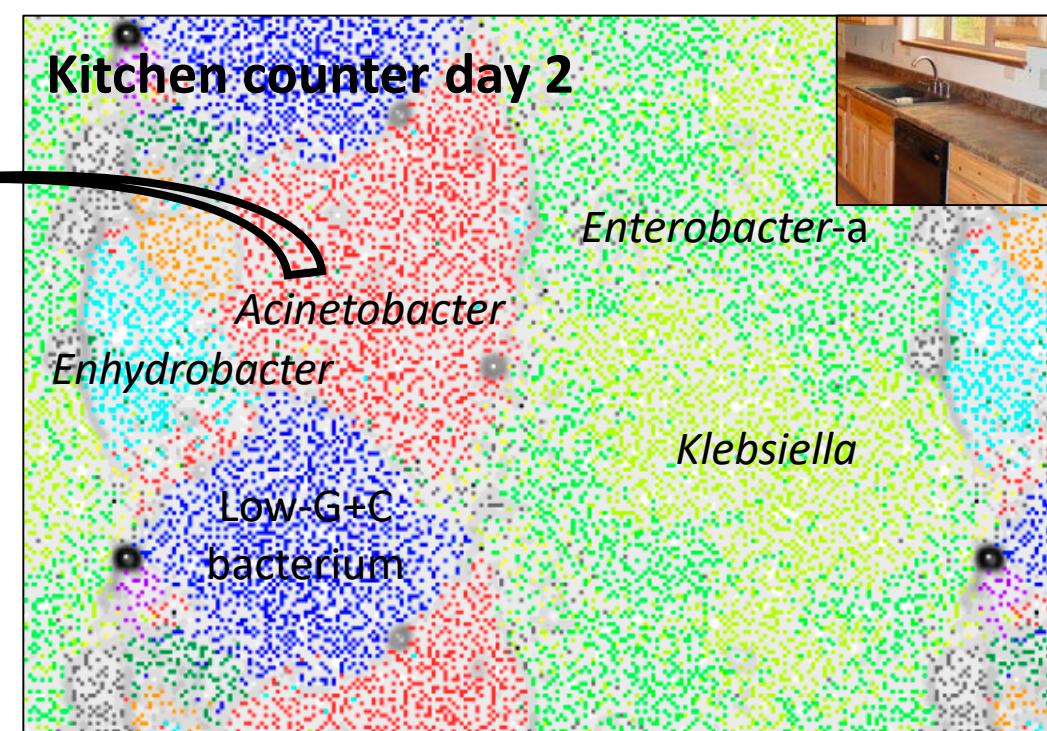
Kitchen counter day 2



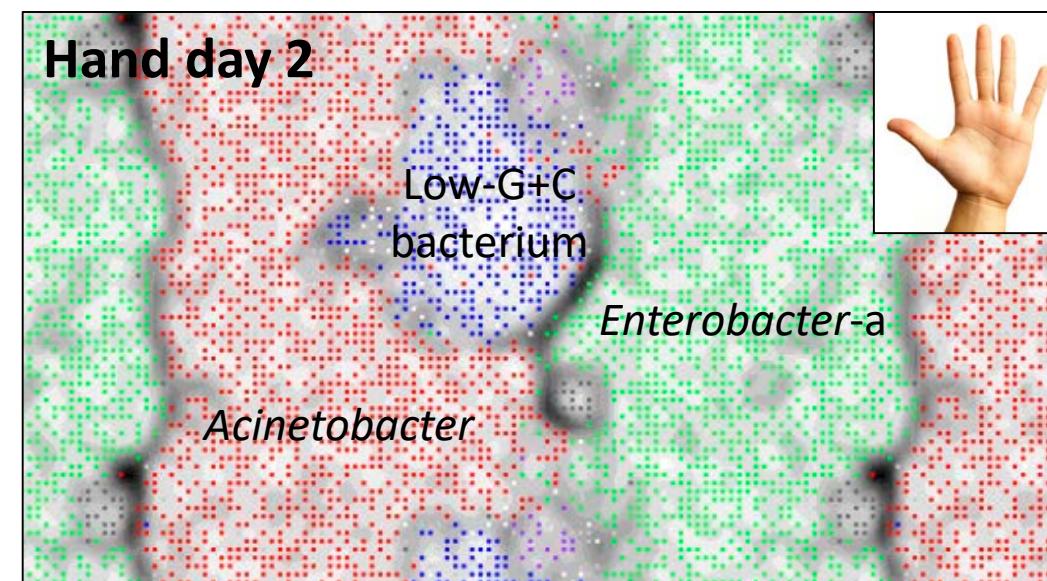
Hand day 2



Kitchen counter day 2



Hand day 2

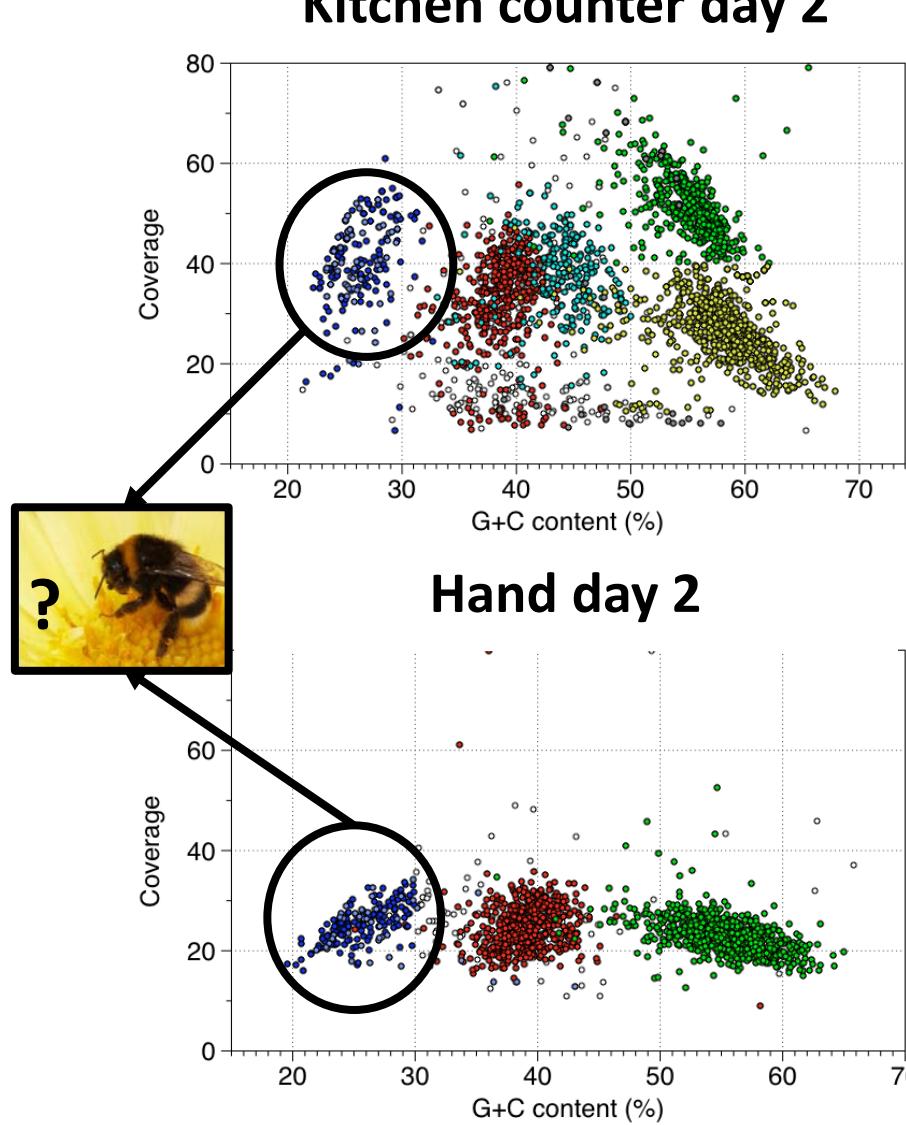
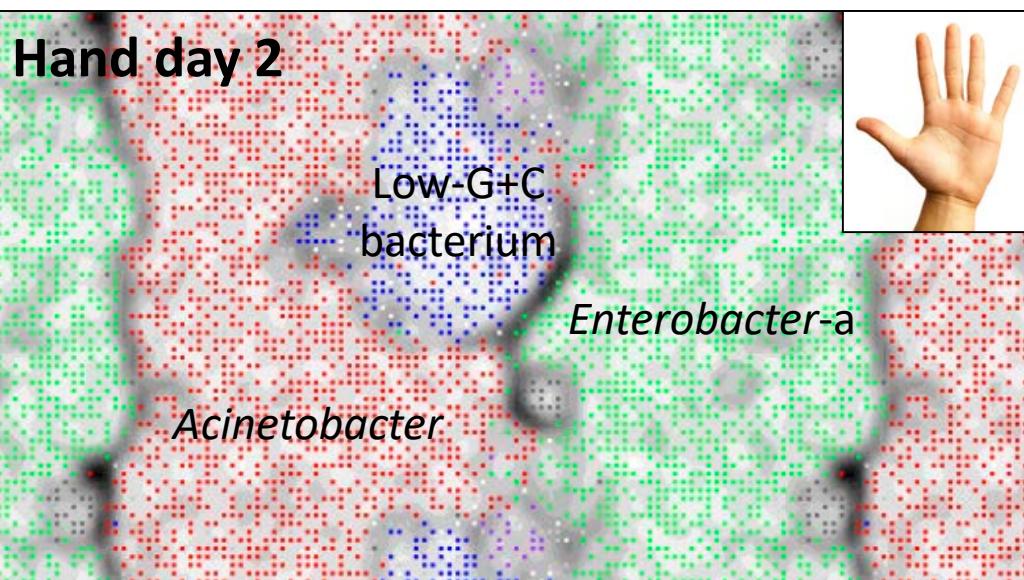
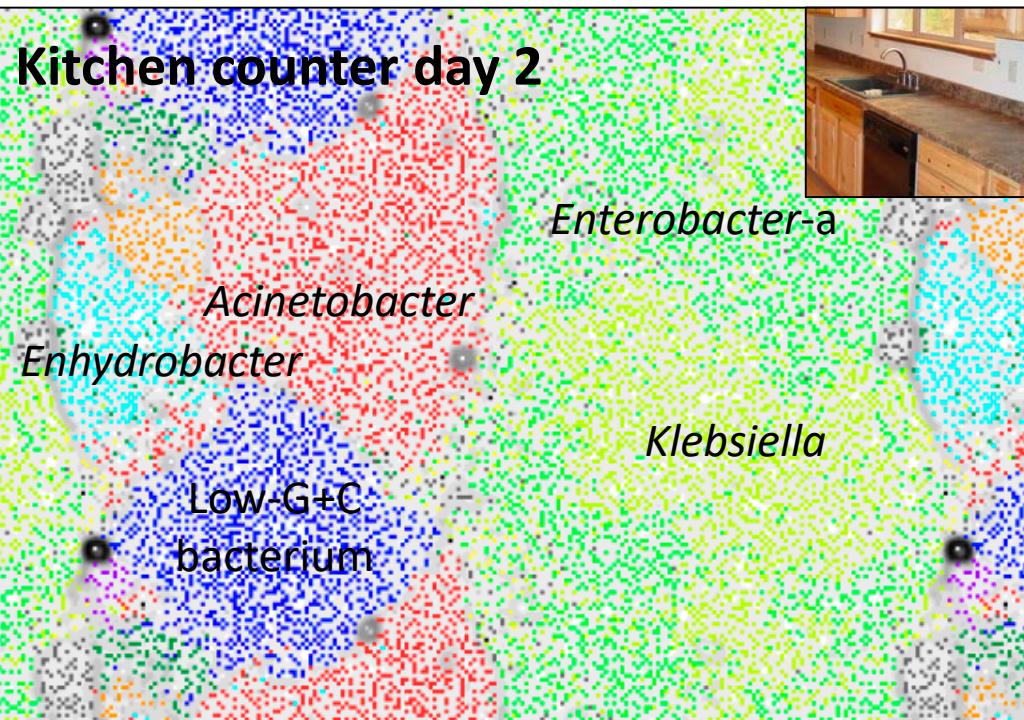


Cluster separation of distinct genera or species





**Cluster separation of distinct genera or species**



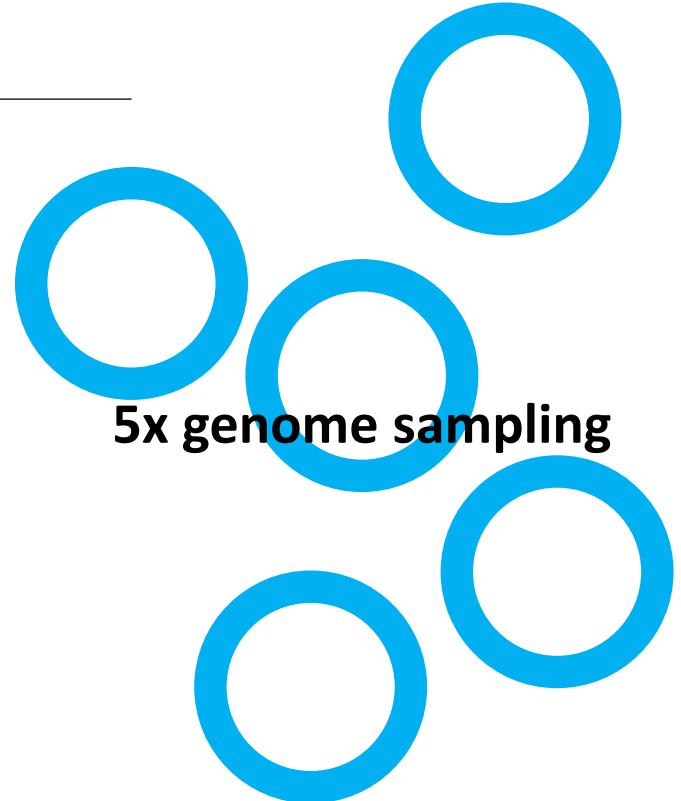
Recombinase A (RecA) ~70% ID to *Gilliamella apicola*



# Coverage

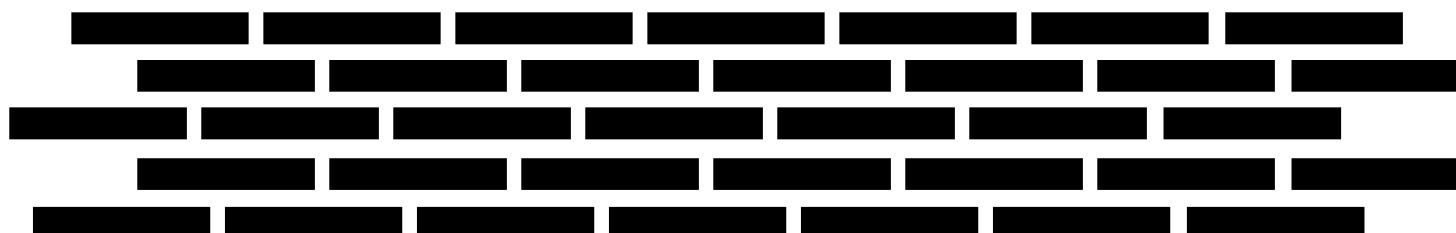
---

- Coverage = Number of times genome sequenced
- Map reads to genome

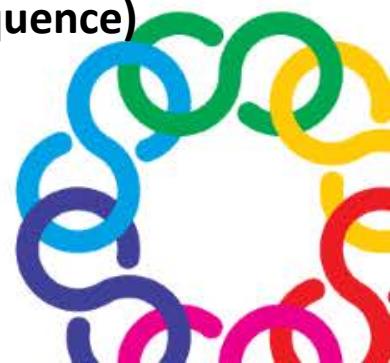


ATGGGCTAAAGATAGGCCTAATAGGTACTGGGATCCAAGCT

Assembled contig (consensus sequence)

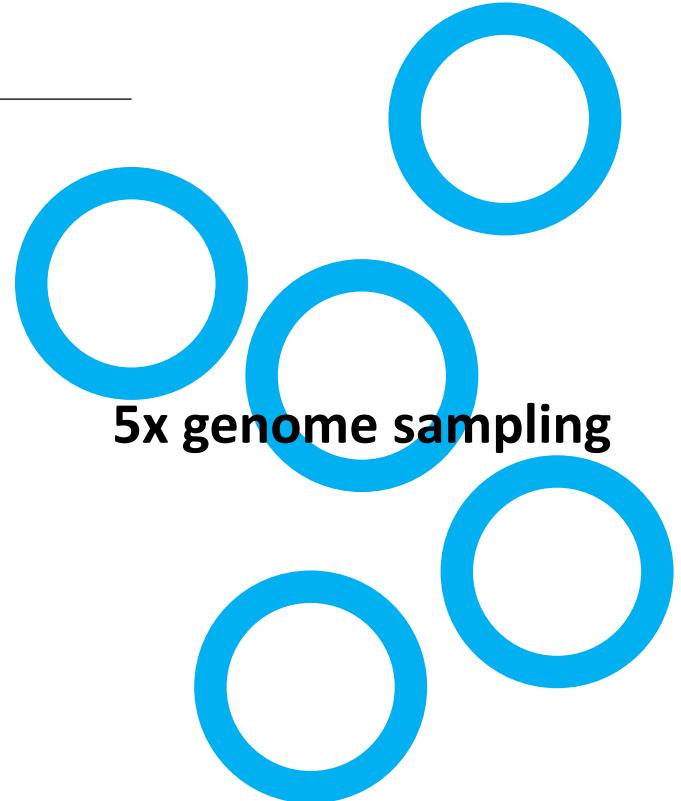


Mapped reads at 5x coverage



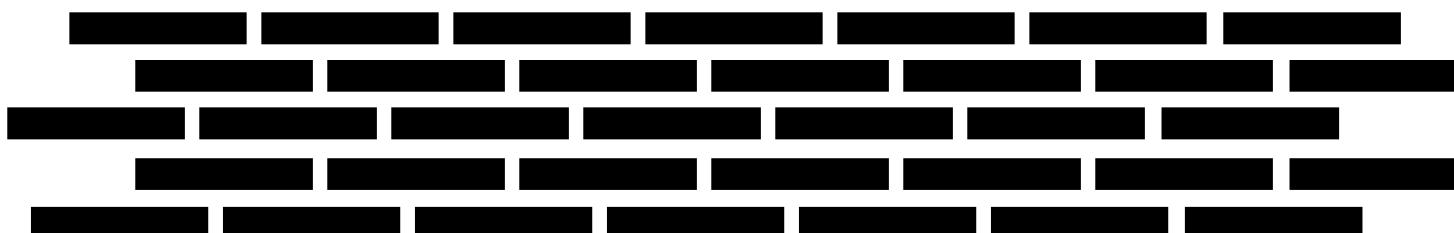
# Coverage

- Use mapping file (e.g. bowtie sam file)
- Count reads per mapped per contig
- Calculate contig (or bin) coverage:
  - $(\text{read length} * \text{number reads}) / \text{contig (or bin) length}$
  - Normalize to library size (if multiple samples)

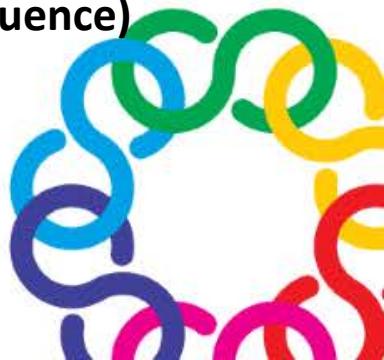


ATGGGCTAAAGATAGGCCTAATAGGTACTGGGATCCAAGCT

Assembled contig (consensus sequence)



Mapped reads at 5x coverage



# Coverage viewing: Tablet



(<http://tablet.hutton.ac.uk/en/latest/index.html#>)



# Read mapping: sam file contents

---

```

HWI-ST1035:107:COLBYACXX:3:2308:18248:182773      163   NODE_2076_length_3089_cov_9.764973  1497  255  101M =  

      1511 115  GATAATGCCACGTCCCACCCATGCCACAGGGCGAGCACTTGCCTCTGGAGCACGGGCATTCCAACCGCCTTCGGACGCAACGAATGAGGAAGGTG  
  CCCFFFFHHHHHIIJJJJJJJJJJJJJJJJJJFJJJJJJGJJJJJJFHFFFDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDC  
XA:::1  MD:Z:2G42T55  NM:::2  
  

HWI-ST1035:107:COLBYACXX:3:2308:18248:182773      83    NODE_2076_length_3089_cov_9.764973  1511  255  101M =  

      1497 -115  CACCATGCCACAAGGCCGAGCAGTGCCTCTGGAGACAGGGCATTCCAACCGCCTTCGGACGCAATGAGGAAGGTGCGAGAGCAGCAA  
BDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDDD  
XA:::0  MD:Z:31T69  NM:::1  
  

HWI-ST1035:107:COLBYACXX:3:2308:13412:183356      99    NODE_389_length_2884_cov_10.890083  2682  255  101M =  

      2735 154  ATCTCGGGCTTCGAGATCCAGCAGCTGGCGGCTTGGCGAGCTATTGACAGCTGGTGGTTGAGAACTATCGTGAAACGCAGCAGCCTGCG  
@@@DFFFFHHHHHIIJJJJJJJJJJJJJJJJJJGIEHFFEDEDCEDDBCCCD@BB:@ABDDCDDDDDDDDDDDDDDBDDEDCCDDDDDDDDDDDD  
XA:::0  MD:Z:46G54  NM:::1  
  

HWI-ST1035:107:COLBYACXX:3:2308:13412:183356      147   NODE_389_length_2884_cov_10.890083  2735  255  101M =  

      2682 -154  AGCTTGGTGGTTGAGAACTATCGGTGGAACGCAGCAGCTGTGGCGACTGCAGCGGGTGTCCGCTCTGCAAGTCAGCAGCAGCACACA  
DDBDDDDDBBA>DDCDEEDCDDDDDDDDA<25DDDDDDDDDDDDDDDDDDDDDDDDDDDDDDFFHHIIJJJJJJJJJJJJJJJJJJJJJJJJJJJJJHHHGF  
XA:::0  MD:Z:101  NM:::0  
  

HWI-ST1035:107:COLBYACXX:3:2308:16904:185061      163   NODE_1144_length_1992_cov_11.747992  579   579  101M =  

      592 114  TCTACCTTATCGGGCAGGCCGTAAGTGATGCCGCGGGATGACGATGTAATCCCGCAGATAACGTTGCTTCGAAAAAGTTCAACTGTACC  
@??DBD? DF<CCCBDEHBH@AF?D?P?D?D?DBFFGFHGHHFCB@C@BD?P?@C>C@BD?P?D?D?D?D?D?D@?9ABC@C@D@?D?D?D?D?  
MD:Z:85T0C14 NM:::2  XA:::0  
  

HWI-ST1035:107:COLBYACXX:3:2308:16904:185061      83    NODE_1144_length_1992_cov_11.747992  592   255  101M =  

      579 -114  GCGGGAGCCGGTAAGTGATGCCGCCGGATGACGATGTAATCCCGCAGATAACGTTGCTTCGAAAAAGTTCAACTGTACCGCCCTTGAAG  
#####? CC@@ACCB@BB<5><?>CBBBBBCECCC@?B;95BBB9<<C?CC@?C?<@AACED@GGEIHFCH@GIG@;7IGFDHBDDDA;D@?@ XA:::1  
MD:Z:2A0C68T0C27  NM:::4

```

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,255}	Query template NAME
2	FLAG	Int	[0,2 <sup>16</sup> -1]	bitwise FLAG
3	RNAME	String	\*  [!-()+-<>-~] [!-~]*	Reference sequence NAME
4	POS	Int	[0,2 <sup>31</sup> -1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2 <sup>8</sup> -1]	MAPping Quality
6	CIGAR	String	\*  ([0-9]+[MIDNSHPX=])+	CIGAR string
7	RNEXT	String	\* =  [!-()+-<>-~] [!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0,2 <sup>31</sup> -1]	Position of the mate/next read
9	TLEN	Int	[-2 <sup>31</sup> +1,2 <sup>31</sup> -1]	observed Template LENgth
10	SEQ	String	\*  [A-Za-z.=.]+	segment SEQuence
11	QUAL	String	[!-~]+	ASCII of Phred-scaled base QUALity+33



# Task: Binning

---

[Go to Github MSS webpage](#)

Tasks:

- Short contig removal ✓
- Read mapping
- Automated binning (using Metabat and Maxbin)
- Pick-the-winner bin dereplication (using DAS Tool)
- Bin evaluation (using CheckM)

**NOTE:** we usually also use CONCOCT which tends to produce a smaller number of high quality bins. Running it is a little more involved, so it's excluded from the MSS, but we encourage everyone to consider using this tool.



# Binning (continued)



# Differential coverage

---

## Infant gut ESOM

(Time series differential coverage ESOM,  
Sharon et al., 2013, Genome Research)

(Extraction bias differential coverage,  
Albertson et al., 2013, Nature Biotechnology)



# Differential coverage: Extraction bias

---

- Activated sludge DNA was extracted 2 different ways
- Extraction bias resulted in different relative abundances of microbial DNA
- Differential coverage helped separate 12 genomes, including 4 TM7 candidate phyla
- How to guide:  
<https://github.com/MadsAlbertsen/multi-metagenome>

(Albertson et al., 2013, Nature Biotechnology)



# Phylogeny

---

- Classify bins based on core gene matches (**16S rRNA genes are typically not recovered!!!**)
- Manual binning or inspection:
  - use protein sequence matches to reference genome database
  - e.g. uniref90 predicted protein database from UniProt  
(<http://www.uniprot.org/uniref/>)
  - many are not phylogenetically conserved
  - works best for longer contigs with many genes



node	gene	g+C	coverage	length	%ID	e-value	bin	taxa	lineage
contig-75_1_-	2	0.49	32.34137	118,738	92	1.10E-144	alpha-Rhodobacterales	Rhodobacteraceae RepID=Q169T6	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	82	0.5	32.34137	118,738	65	5.50E-79	alpha-Rhodobacterales	Roseobacter RepID=Q16CQ2_ROSD	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	38	0.5	32.34137	118,738	57	4.50E-95	alpha-Rhodobacterales	Silicibacter lacuscaerulensis ITI-115	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	21	0.49	32.34137	118,738	78	2.70E-213	alpha-Rhodobacterales	Roseovarius nubinhibens ISM RepID	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	83	0.47	32.34137	118,738	68	1.80E-76	alpha-Rhodobacterales	Pseudoxanthomonas spadix (strain	Bacteria; Proteobacteria; Gammaproteobacteri
contig-75_1_-	100	0.46	32.34137	118,738	60	3.50E-70	alpha-Rhodobacterales	Oceanicola batsensis HTCC2597 Rep	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	104	0.5	32.34137	118,738	55	1.90E-57	alpha-Rhodobacterales	Roseovarius sp. 217 RepID=A3VXP4	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	113	0.52	32.34137	118,738	79	0	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC22	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	39	0.51	32.34137	118,738	79	5.20E-107	alpha-Rhodobacterales	Ruegeria pomeroyi (strain ATCC 700	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	59	0.5	32.34137	118,738	66	5.70E-177	alpha-Rhodobacterales	Ruegeria pomeroyi (strain ATCC 700	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	64	0.51	32.34137	118,738	77	2.40E-153	alpha-Rhodobacterales	Ruegeria pomeroyi (strain ATCC 700	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	65	0.49	32.34137	118,738	77	2.00E-155	alpha-Rhodobacterales	Roseobacter sp. SK209-2-6 RepID=A	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	49	0.48	32.34137	118,738	83	5.70E-120	alpha-Rhodobacterales	Rhodobacteraceae bacterium HTCC	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	14	0.5	32.34137	118,738	37	5.80E-43	alpha-Rhodobacterales	Limnobacter sp. MED105 RepID=A6	Bacteria; Proteobacteria; Betaproteobacteria; B
contig-75_1_-	32	0.47	32.34137	118,738	55	5.20E-134	alpha-Rhodobacterales	Silicibacter lacuscaerulensis ITI-115	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	80	0.52	32.34137	118,738	49	4.20E-94	alpha-Rhodobacterales	Thalassiothrix sp. R2A62 RepID=C7U	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	110	0.52	32.34137	118,738	63	2.70E-134	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC22	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	55	0.5	32.34137	118,738	73	1.50E-184	alpha-Rhodobacterales	Celeribacter baekdemonensis B30 Rep	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	81	0.49	32.34137	118,738	63	4.50E-73	alpha-Rhodobacterales	Acaryochloris marina (strain MBIC 1	Bacteria; Cyanobacteria; Oscillatoriophycideae;
contig-75_1_-	16	0.5	32.34137	118,738	80	0	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC22	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	47	0.51	32.34137	118,738	76	3.40E-123	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC22	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	48	0.52	32.34137	118,738	79	9.60E-127	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC22	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	30	0.49	32.34137	118,738	67	0	alpha-Rhodobacterales	Dinoroseobacter shibae (strain DFL	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	77	0.5	32.34137	118,738	63	2.30E-136	alpha-Rhodobacterales	Rhodobacter sp. SW2 RepID=C8RZY	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	28	0.5	32.34137	118,738	75	1.40E-195	alpha-Rhodobacterales	Rhodobacteraceae RepID=I7ES19_PI	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	106	0.5	32.34137	118,738	87	0	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC22	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	67	0.49	32.34137	118,738	91	5.50E-139	alpha-Rhodobacterales	Alphaproteobacteria RepID=Q1GDV	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_-	70	0.44	32.34137	118,738	43	2.30E-82	alpha-Rhodobacterales	Roseobacter RepID=Q160N7_ROSD	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_6_-	32	0.41	45.95991	106,871	47	1.60E-91	delta-a1	Clostridium thermocellum RepID=A	Bacteria; Firmicutes; Clostridia; Clostridiales; Clo
contig-75_6_-	47	0.34	45.95991	106,871	48	8.60E-56	delta-a1	Nostoc sp. PCC 7107 RepID=K9QA8	Bacteria; Cyanobacteria; Nostocales; Nostocace
contig-75_6_-	62	0.42	45.95991	106,871	41	1.40E-41	delta-a1	Sulfurihydrogenibium RepID=B2V76	Bacteria; Aquificae; Aquificales; Hydrogenother
contig-75_6_-	48	0.37	45.95991	106,871	49	2.70E-63	delta-a1	Candidatus Kuenenia stuttgartiensis	Bacteria; Planctomycetes; Planctomycetia; Can
contig-75_6_-	44	0.35	45.95991	106,871	35	5.00E-26	delta-a1	uncultured bacterium RepID=K2CVX	Bacteria; Firmicutes; Clostridia; Clostridiales
contig-75_6_-	5	0.42	45.95991	106,871	58	9.40E-234	delta-a1	Desulfobacula toluolica (strain DSM	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_-	18	0.45	45.95991	106,871	66	1.90E-168	delta-a1	Desulfobacterium autotrophicum (s	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_-	60	0.4	45.95991	106,871	61	2.30E-147	delta-a1	Desulfatibacterium alkenivorans (stra	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_-	6	0.37	45.95991	106,871	41	2.00E-147	delta-a1	uncultured bacterium RepID=K2D67	Bacteria; Firmicutes; Clostridia; Clostridiales
contig-75_6_-	10	0.38	45.95991	106,871	31	2.90E-128	delta-a1	Desulfobacula toluolica (strain DSM	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_-	58	0.37	45.95991	106,871	50	2.90E-102	delta-a1	Cyanotheces RepID=B1WVA7_CYAAE	Bacteria; Cyanobacteria; Oscillatoriophycideae;
contig-75_6_-	63	0.4	45.95991	106,871	55	1.60E-77	delta-a1	Methylomicrobium album BG8 Rep	Bacteria; Proteobacteria; Gammaproteobacteri
contig-75_6_-	7	0.43	45.95991	106,871	29	3.80E-66	delta-a1	Cyanotheces sp. (strain PCC 7822) Re	Bacteria; Cyanobacteria; Oscillatoriophycideae;
contig-75_6_-	20	0.39	45.95991	106,871	32	9.30E-26	delta-a1	Desulfobacula toluolica (strain DSM	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_-	34	0.43	45.95991	106,871	69	5.70E-23	delta-a1	Desulfobacter postgatei 2ac9 RepID	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_-	35	0.42	45.95991	106,871	64	2.70E-276	delta-a1	Desulfobacter postgatei 2ac9 RepID	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_-	12	0.39	45.95991	106,871	53	4.20E-31	delta-a1	Desulfobacterium autotrophicum (s	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_-	14	0.38	45.95991	106,871	50	3.70E-16	delta-a1	Desulfobacterium autotrophicum (s	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_-	15	0.42	45.95991	106,871	48	3.70E-126	delta-a1	Desulfobacterium autotrophicum (s	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_-	13	0.41	45.95991	106,871	58	9.70E-45	delta-a1	Desulfobacterium autotrophicum (s	Bacteria; Proteobacteria; Deltaproteobacteria; I



contig 1

contig 2

node	gene	g+C	coverage	length	%ID	e-value	bin	taxa	lineage
contig-75_1_	2	0.49	32.34137	118,738	92	1.10E-144	alpha-Rhodobacterales	Rhodobacteraceae RepID=Q169T6	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	82	0.5	32.34137	118,738	65	5.50E-79	alpha-Rhodobacterales	Roseobacter RepID=Q16CQ2_ROSD	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	38	0.5	32.34137	118,738	57	4.50E-95	alpha-Rhodobacterales	Silicibacter lacuscaerulensis ITI-115	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	21	0.49	32.34137	118,738	78	2.70E-213	alpha-Rhodobacterales	Roseovarius nubinhibens ISM RepID	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	83	0.47	32.34137	118,738	68	1.80E-76	alpha-Rhodobacterales	Pseudoxanthomonas spadix (strain	Bacteria; Proteobacteria; Gammaproteobacteri
contig-75_1_	100	0.46	32.34137	118,738	60	3.50E-70	alpha-Rhodobacterales	Oceanicola batsensis HTCC2597 Rep	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	104	0.5	32.34137	118,738	55	1.90E-57	alpha-Rhodobacterales	Roseovarius sp. 217 RepID=A3VXP4	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	113	0.52	32.34137	118,738	79	0	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC22	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	39	0.51	32.34137	118,738	79	5.20E-107	alpha-Rhodobacterales	Ruegeria pomeroyi (strain ATCC 700	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	59	0.5	32.34137	118,738	66	5.70E-177	alpha-Rhodobacterales	Ruegeria pomeroyi (strain ATCC 700	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	64	0.51	32.34137	118,738	77	2.40E-153	alpha-Rhodobacterales	Ruegeria pomeroyi (strain ATCC 700	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	65	0.49	32.34137	118,738	77	2.00E-155	alpha-Rhodobacterales	Roseobacter sp. SK209-2-6 RepID=A	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	49	0.48	32.34137	118,738	83	5.70E-120	alpha-Rhodobacterales	Rhodobacteraceae bacterium HTCC	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	14	0.5	32.34137	118,738	37	5.80E-43	alpha-Rhodobacterales	Limnobacter sp. MED105 RepID=A6	Bacteria; Proteobacteria; Betaproteobacteria; B
contig-75_1_	32	0.47	32.34137	118,738	55	5.20E-134	alpha-Rhodobacterales	Silicibacter lacuscaerulensis ITI-115	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	80	0.52	32.34137	118,738	49	4.20E-94	alpha-Rhodobacterales	Thalassibium sp. R2A62 RepID=C7	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	110	0.52	32.34137	118,738	63	2.70E-134	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC22	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	55	0.5	32.34137	118,738	73	1.50E-184	alpha-Rhodobacterales	Celeribacter baekdomensis B30 Rep	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	81	0.49	32.34137	118,738	63	4.50E-73	alpha-Rhodobacterales	Acaryochloris marina (strain MBIC 1	Bacteria; Cyanobacteria; Oscillatoriophycideae;
contig-75_1_	16	0.5	32.34137	118,738	80	0	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC22	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	47	0.51	32.34137	118,738	76	3.40E-123	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC22	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	48	0.52	32.34137	118,738	79	9.60E-127	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC22	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	30	0.49	32.34137	118,738	67	0	alpha-Rhodobacterales	Dinoroseobacter shibae (strain DFL	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	77	0.5	32.34137	118,738	63	2.30E-136	alpha-Rhodobacterales	Rhodobacter sp. SW2 RepID=C8RZY	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	28	0.5	32.34137	118,738	75	1.40E-195	alpha-Rhodobacterales	Rhodobacteraceae RepID=I7ES19_PI	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	106	0.5	32.34137	118,738	87	0	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC22	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	67	0.49	32.34137	118,738	91	5.50E-139	alpha-Rhodobacterales	Alphaproteobacteria RepID=Q1GDV	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	70	0.44	32.34137	118,738	43	2.30E-82	alpha-Rhodobacterales	Roseobacter RepID=Q160N7_ROSD	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_6_	32	0.41	45.95991	106,871	47	1.60E-91	delta-a1	Clostridium thermocellum RepID=A	Bacteria; Firmicutes; Clostridia; Clostridiales; Clo
contig-75_6_	47	0.34	45.95991	106,871	48	8.60E-56	delta-a1	Nostoc sp. PCC 7107 RepID=K9QA8	Bacteria; Cyanobacteria; Nostocales; Nostocace
contig-75_6_	62	0.42	45.95991	106,871	41	1.40E-41	delta-a1	Sulfurihydrogenibium RepID=B2V76	Bacteria; Aquificae; Aquificales; Hydrogenother
contig-75_6_	48	0.37	45.95991	106,871	49	2.70E-63	delta-a1	Candidatus Kuenenia stuttgartiensis	Bacteria; Planctomycetes; Planctomycetia; Can
contig-75_6_	44	0.35	45.95991	106,871	35	5.00E-26	delta-a1	uncultured bacterium RepID=K2CVX	Bacteria; Firmicutes; Clostridia; Clostridiales
contig-75_6_	5	0.42	45.95991	106,871	58	9.40E-234	delta-a1	Desulfovacula toluolica (strain DSM	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_	18	0.45	45.95991	106,871	66	1.90E-168	delta-a1	Desulfovacterium autotrophicum (s	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_	60	0.4	45.95991	106,871	61	2.30E-147	delta-a1	Desulfatibacillus alkenivorans (stra	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_	6	0.37	45.95991	106,871	41	2.00E-147	delta-a1	uncultured bacterium RepID=K2D67	Bacteria; Firmicutes; Clostridia; Clostridiales
contig-75_6_	10	0.38	45.95991	106,871	31	2.90E-128	delta-a1	Desulfovacula toluolica (strain DSM	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_	58	0.37	45.95991	106,871	50	2.90E-102	delta-a1	Cyanothecae RepID=B1WVA7_CYAAE	Bacteria; Cyanobacteria; Oscillatoriophycideae;
contig-75_6_	63	0.4	45.95991	106,871	55	1.60E-77	delta-a1	Methylomicrobium album BG8 Rep	Bacteria; Proteobacteria; Gammaproteobacteri
contig-75_6_	7	0.43	45.95991	106,871	29	3.80E-66	delta-a1	Cyanothecae sp. (strain PCC 7822) Re	Bacteria; Cyanobacteria; Oscillatoriophycideae;
contig-75_6_	20	0.39	45.95991	106,871	32	9.30E-26	delta-a1	Desulfovacula toluolica (strain DSM	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_	34	0.43	45.95991	106,871	69	5.70E-23	delta-a1	Desulfovacter postgatei 2ac9 RepID	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_	35	0.42	45.95991	106,871	64	2.70E-276	delta-a1	Desulfovacter postgatei 2ac9 RepID	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_	12	0.39	45.95991	106,871	53	4.20E-31	delta-a1	Desulfovacterium autotrophicum (s	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_	14	0.38	45.95991	106,871	50	3.70E-16	delta-a1	Desulfovacterium autotrophicum (s	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_	15	0.42	45.95991	106,871	48	3.70E-126	delta-a1	Desulfovacterium autotrophicum (s	Bacteria; Proteobacteria; Deltaproteobacteria; I
contig-75_6_	13	0.41	45.95991	106,871	58	9.70E-45	delta-a1	Desulfovacterium autotrophicum (s	Bacteria; Proteobacteria; Deltaproteobacteria; I



contig 1

node	gene	g+C	coverage	length	%ID	e-value	bin	taxa	lineage
contig-75_1_	2	0.49	32.34137	118,738	92	1.10E-144	alpha-Rhodobacterales	Rhodobacteraceae RepID=Q169T6	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	82	0.5	32.34137	118,738	65	5.50E-79	alpha-Rhodobacterales	Roseobacter RepID=Q16CQ2_ROSD	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	38	0.5	32.34137	118,738	57	4.50E-95	alpha-Rhodobacterales	Silicibacter lacuscaerulensis ITI-115	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	21	0.49	32.34137	118,738	78	2.70E-213	alpha-Rhodobacterales	Roseovarius nubinhibens ISM RepID	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	83	0.47	32.34137	118,738	68	1.80E-76	alpha-Rhodobacterales	Pseudoxanthomonas spadix (strain	Bacteria; Proteobacteria; Gammaproteobacter
contig-75_1_	100	0.46	32.34137	118,738	60	3.50E-70	alpha-Rhodobacterales	Oceanicola batsensis HTCC2597 Rep	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	104	0.5	32.34137	118,738	55	1.90E-57	alpha-Rhodobacterales	Roseovarius sp. 217 RepID=A3VXP4	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	113	0.52	32.34137	118,738	79	0	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC2	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	39	0.51	32.34137	118,738	79	5.20E-107	alpha-Rhodobacterales	Ruegeria pomeroyi (strain ATCC 70	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	59	0.5	32.34137	118,738	66	5.70E-177	alpha-Rhodobacterales	Ruegeria pomeroyi (strain ATCC 70	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	64	0.51	32.34137	118,738	77	2.40E-153	alpha-Rhodobacterales	Ruegeria pomeroyi (strain ATCC 70	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	65	0.49	32.34137	118,738	77	2.00E-155	alpha-Rhodobacterales	Roseobacter sp. SK209-2-6 RepID=A	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	49	0.48	32.34137	118,738	83	5.70E-120	alpha-Rhodobacterales	Rhodobacteraceae bacterium HTCC	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	14	0.5	32.34137	118,738	37	5.80E-43	alpha-Rhodobacterales	Limnobacter sp. MED105 RepID=A6	Bacteria; Proteobacteria; Betaproteobacteria;
contig-75_1_	32	0.47	32.34137	118,738	55	5.20E-134	alpha-Rhodobacterales	Silicibacter lacuscaerulensis ITI-115	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	80	0.52	32.34137	118,738	49	4.20E-94	alpha-Rhodobacterales	Thalassiothrix sp. R2A62 RepID=C7	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	110	0.52	32.34137	118,738	63	2.70E-134	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC2	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	55	0.5	32.34137	118,738	73	1.50E-184	alpha-Rhodobacterales	Celeribacter baekdomensis B30 Rep	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	81	0.49	32.34137	118,738	63	4.50E-73	alpha-Rhodobacterales	Acaryochloris marina (strain MBIC 1	Bacteria; Cyanobacteria; Oscillatoriophycideae
contig-75_1_	16	0.5	32.34137	118,738	80	0	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC2	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	47	0.51	32.34137	118,738	76	3.40E-123	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC2	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	48	0.52	32.34137	118,738	79	9.60E-127	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC2	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	30	0.49	32.34137	118,738	67	0	alpha-Rhodobacterales	Dinoroseobacter shibae (strain DFL	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	77	0.5	32.34137	118,738	63	2.30E-136	alpha-Rhodobacterales	Rhodobacter sp. SW2 RepID=C8RZY	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	28	0.5	32.34137	118,738	75	1.40E-195	alpha-Rhodobacterales	Rhodobacteraceae RepID=I7ESI9_P	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	106	0.5	32.34137	118,738	87	0	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC2	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	67	0.49	32.34137	118,738	91	5.50E-139	alpha-Rhodobacterales	Alphaproteobacteria RepID=Q1GDV	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	70	0.44	32.34137	118,738	43	2.30E-82	alpha-Rhodobacterales	Roseobacter RepID=Q160N7_ROSD	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_6_	32	0.41	45.95991	106,871	47	1.60E-91	delta-a1	Clostridium thermocellum RepID=A	Bacteria; Firmicutes; Clostridia; Clostridiales; Clo
contig-75_6_	47	0.34	45.95991	106,871	48	8.60E-56	delta-a1	Nostoc sp. PCC 7107 RepID=K9QA8	Bacteria; Cyanobacteria; Nostocales; Nostocac
contig-75_6_	62	0.42	45.95991	106,871	41	1.40E-41	delta-a1	Sulfurihydrogenibium RepID=B2V76	Bacteria; Aquificae; Aquificales; Hydrogenothe
contig-75_6_	48	0.37	45.95991	106,871	49	2.70E-63	delta-a1	Candidatus Kuenenia stuttgartiens	Bacteria; Planctomycetes; Planctomycetia; Can
contig-75_6_	44	0.35	45.95991	106,871	35	5.00E-26	delta-a1	uncultured bacterium RepID=K2CV2	Bacteria; Firmicutes; Clostridia; Clostridiales
contig-75_6_	5	0.42	45.95991	106,871	58	9.40E-234	delta-a1	Desulfovacula toluolica (strain DSM	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	18	0.45	45.95991	106,871	66	1.90E-168	delta-a1	Desulfovacterium autotrophicum (s	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	60	0.4	45.95991	106,871	61	2.30E-147	delta-a1	Desulfatibacterium alkenivorans (stra	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	6	0.37	45.95991	106,871	41	2.00E-147	delta-a1	uncultured bacterium RepID=K2D67	Bacteria; Firmicutes; Clostridia; Clostridiales
contig-75_6_	10	0.38	45.95991	106,871	31	2.90E-128	delta-a1	Desulfovacula toluolica (strain DSM	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	58	0.37	45.95991	106,871	50	2.90E-102	delta-a1	Cyanothecaceae RepID=B1WVA7_CYAA	Bacteria; Cyanobacteria; Oscillatoriophycideae
contig-75_6_	63	0.4	45.95991	106,871	55	1.60E-77	delta-a1	Methylomicrobium album BG8 Rep	Bacteria; Proteobacteria; Gammaproteobacter
contig-75_6_	7	0.43	45.95991	106,871	29	3.80E-66	delta-a1	Cyanothecaceae sp. (strain PCC 7822) R	Bacteria; Cyanobacteria; Oscillatoriophycideae
contig-75_6_	20	0.39	45.95991	106,871	32	9.30E-26	delta-a1	Desulfovacula toluolica (strain DSM	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	34	0.43	45.95991	106,871	69	5.70E-23	delta-a1	Desulfovacter postgatei 2ac9 RepID	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	35	0.42	45.95991	106,871	64	2.70E-276	delta-a1	Desulfovacter postgatei 2ac9 RepID	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	12	0.39	45.95991	106,871	53	4.20E-31	delta-a1	Desulfovacterium autotrophicum (s	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	14	0.38	45.95991	106,871	50	3.70E-16	delta-a1	Desulfovacterium autotrophicum (s	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	15	0.42	45.95991	106,871	48	3.70E-126	delta-a1	Desulfovacterium autotrophicum (s	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	13	0.41	45.95991	106,871	58	9.70E-45	delta-a1	Desulfovacterium autotrophicum (s	Bacteria; Proteobacteria; Deltaproteobacteria;

G+C = 50%

Coverage = 32 x

Delta proteobacteria

G+C = 40%

Coverage = 46 x



node	gene	g+C	coverage	length	%ID	e-value	bin	taxa	lineage
contig-75_1_	2	0.49	32.34137	118,738	92	1.10E-144	alpha-Rhodobacterales	Rhodobacteraceae RepID=Q169T6	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	82	0.5	32.34137	118,738	65	5.50E-79	alpha-Rhodobacterales	Roseobacter RepID=Q16CQ2_ROSD	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	38	0.5	32.34137	118,738	57	4.50E-95	alpha-Rhodobacterales	Silicibacter lacuscaerulensis ITI-115	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	21	0.49	32.34137	118,738	78	2.70E-213	alpha-Rhodobacterales	Roseovarius nubinhibens ISM RepID	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	83	0.47	32.34137	118,738	68	1.80E-76	alpha-Rhodobacterales	Pseudoxanthomonas spadix (strain	Bacteria; Proteobacteria; Gammaproteobacter
contig-75_1_	100	0.46	32.34137	118,738	60	3.50E-70	alpha-Rhodobacterales	Oceanicola batsensis HTCC2597 Rep	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	104	0.5	32.34137	118,738	55	1.90E-57	alpha-Rhodobacterales	Roseovarius sp. 217 RepID=A3VXP4	Bacteria; Proteobacteria; Alphaproteobacteria;

contig 1

contig-75_1_	47	0.51	32.34137	118,738	76	3.40E-123	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC22	Bacteria; Proteobacteria; Alphaproteobacteria;	
contig-75_1_	48	0.52	32.34137	118,738	79	9.60E-127	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC22	Bacteria; Proteobacteria; Alphaproteobacteria;	
contig-75_1_	30	0.49	32.34137	118,738	67		0	alpha-Rhodobacterales	Dinoroseobacter shibae (strain DFL	
contig-75_1_	77	0.5	32.34137	118,738	63	2.30E-136	alpha-Rhodobacterales	Rhodobacter sp. SW2 RepID=C8RZY	Bacteria; Proteobacteria; Alphaproteobacteria;	
contig-75_1_	28	0.5	32.34137	118,738	75	1.40E-195	alpha-Rhodobacterales	Rhodobacteraceae RepID=I7ESI9_PH	Bacteria; Proteobacteria; Alphaproteobacteria;	
contig-75_1_	106	0.5	32.34137	118,738	87		0	alpha-Rhodobacterales	Rhodobacterales bacterium HTCC22	Bacteria; Proteobacteria; Alphaproteobacteria;
contig-75_1_	67	0.49	32.34137	118,738	91	5.50E-139	alpha-Rhodobacterales	Alphaproteobacteria RepID=Q1GDV	Bacteria; Proteobacteria; Alphaproteobacteria;	
contig-75_1_	70	0.44	32.34137	118,738	43	2.30E-82	alpha-Rhodobacterales	Roseobacter RepID=Q160N7_ROSD	Bacteria; Proteobacteria; Alphaproteobacteria;	
contig-75_6_	32	0.41	45.95991	106,871	47	1.60E-91	delta-a1	Clostridium thermocellum RepID=ATCC25759	Bacteria; Firmicutes; Clostridia; Clostridiales; Clostridiales	
contig-75_6_	47	0.34	45.95991	106,871	48	8.60E-56	delta-a1	Nostoc sp. PCC 7107 RepID=K9QA81	Bacteria; Cyanobacteria; Nostocales; Nostocales	
contig-75_6_	62	0.42	45.95991	106,871	41	1.40E-41	delta-a1	Sulfurihydrogenibium RepID=B2V76	Bacteria; Aquificae; Aquificales; Hydrogenotheres	
contig-75_6_	48	0.37	45.95991	106,871	49	2.70E-63	delta-a1	Candidatus Kuenenia stuttgartiensis	Bacteria; Planctomycetes; Planctomycetia; Canthobiontia	
contig-75_6_	44	0.35	45.95991	106,871	35	5.00E-26	delta-a1	uncultured bacterium RepID=K2CVX	Bacteria; Firmicutes; Clostridia; Clostridiales	
contig-75_6_	5	0.42	45.95991	106,871	58	9.40E-234	delta-a1	Desulfobacula toluolica (strain DSM	Bacteria; Proteobacteria; Deltaproteobacteria;	
contig-75_6_	18	0.45	45.95991	106,871	66	1.90E-168	delta-a1	Desulfobacterium autotrophicum (str	Bacteria; Proteobacteria; Deltaproteobacteria;	
contig-75_6_	60	0.4	45.95991	106,871	61	2.30E-147	delta-a1	Desulfatibacillus alkenivorans (stra	Bacteria; Proteobacteria; Deltaproteobacteria;	
contig-75_6_	6	0.37	45.95991	106,871	41	2.00E-147	delta-a1	uncultured bacterium RepID=K2D67	Bacteria; Firmicutes; Clostridia; Clostridiales	
contig-75_6_	10	0.38	45.95991	106,871	31	2.90E-128	delta-a1	Desulfobacula toluolica (strain DSM	Bacteria; Proteobacteria; Deltaproteobacteria;	

contig 2

contig-75_6_	60	0.4	45.95991	106,871	61	2.30E-147	delta-a1	Desulfatibacillus alkenivorans (stra	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	6	0.37	45.95991	106,871	41	2.00E-147	delta-a1	uncultured bacterium RepID=K2D67	Bacteria; Firmicutes; Clostridia; Clostridiales
contig-75_6_	10	0.38	45.95991	106,871	31	2.90E-128	delta-a1	Desulfobacula toluolica (strain DSM	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	58	0.37	45.95991	106,871	50	2.90E-102	delta-a1	Cyanothecaceae RepID=B1WVA7_CYAA	Bacteria; Cyanobacteria; Oscillatoriophycideae
contig-75_6_	63	0.4	45.95991	106,871	55	1.60E-77	delta-a1	Methylomicrobium album BG8 Rep	Bacteria; Proteobacteria; Gammaproteobacteria
contig-75_6_	7	0.43	45.95991	106,871	29	3.80E-66	delta-a1	Cyanothecaceae RepID=PCC 7822_R	Bacteria; Cyanobacteria; Oscillatoriophycideae
contig-75_6_	20	0.39	45.95991	106,871	32	9.30E-26	delta-a1	Desulfobacula toluolica (strain DSM	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	34	0.43	45.95991	106,871	69	5.70E-23	delta-a1	Desulfobacter postgatei 2ac9 Rep	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	35	0.42	45.95991	106,871	64	2.70E-276	delta-a1	Desulfobacter postgatei 2ac9 Rep	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	12	0.39	45.95991	106,871	53	4.20E-31	delta-a1	Desulfobacterium autotrophicum (str	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	14	0.38	45.95991	106,871	50	3.70E-16	delta-a1	Desulfobacterium autotrophicum (str	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	15	0.42	45.95991	106,871	48	3.70E-126	delta-a1	Desulfobacterium autotrophicum (str	Bacteria; Proteobacteria; Deltaproteobacteria;
contig-75_6_	13	0.41	45.95991	106,871	58	9.70E-45	delta-a1	Desulfobacterium autotrophicum (str	Bacteria; Proteobacteria; Deltaproteobacteria;



# Short contig removal

---

**Shorter contigs are harder to bin based on:**

- **Compositional signatures**
- **Phylogeny affiliation to reference databases**

**There is no rule, but we typically remove contigs <2 kbp long from assemblies. Others remove contigs <1 kbp long. This typically removes a large number of contigs, which tend to be:**

- **problematic to bin, or**
- **contain truncated genes**



# Manual vs automated binning

---

## Manual:

- traditional
- laborious
- highly effective

## Automated:

- fast
- highly variable results between tools
- can co-bin highly abundant contigs from distinct genomes
- tend to discard small genomes (e.g. viruses)

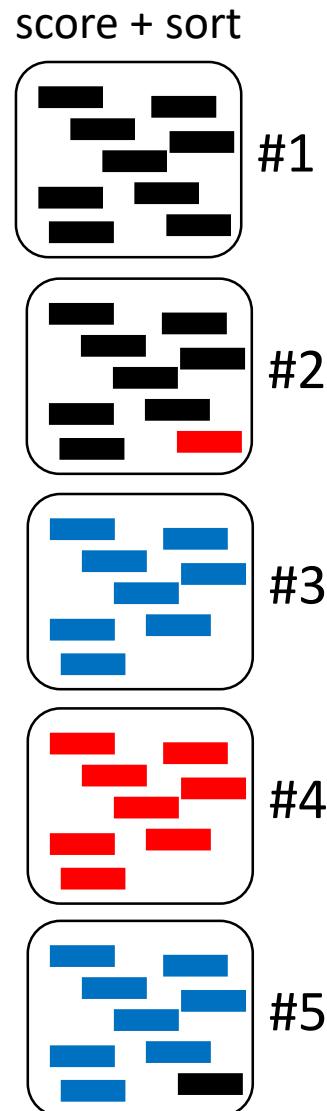


# Multi-binning strategy: pick winner

repeat

DAS Tool:

1. Score bins
2. Sort from best to worse
3. Pick top bin
4. Delete identical contigs in other bins

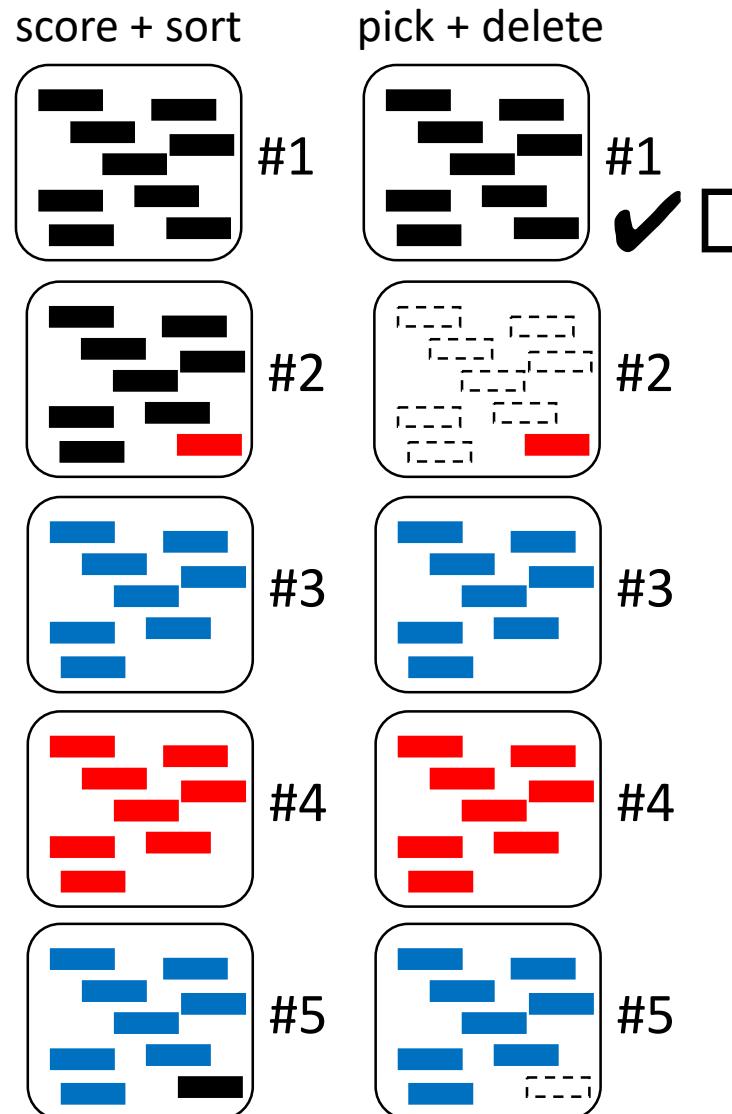


# Multi-binning strategy: pick winner

repeat

DAS Tool:

1. Score bins
2. Sort from best to worse
3. Pick top bin
4. Delete identical contigs in other bins

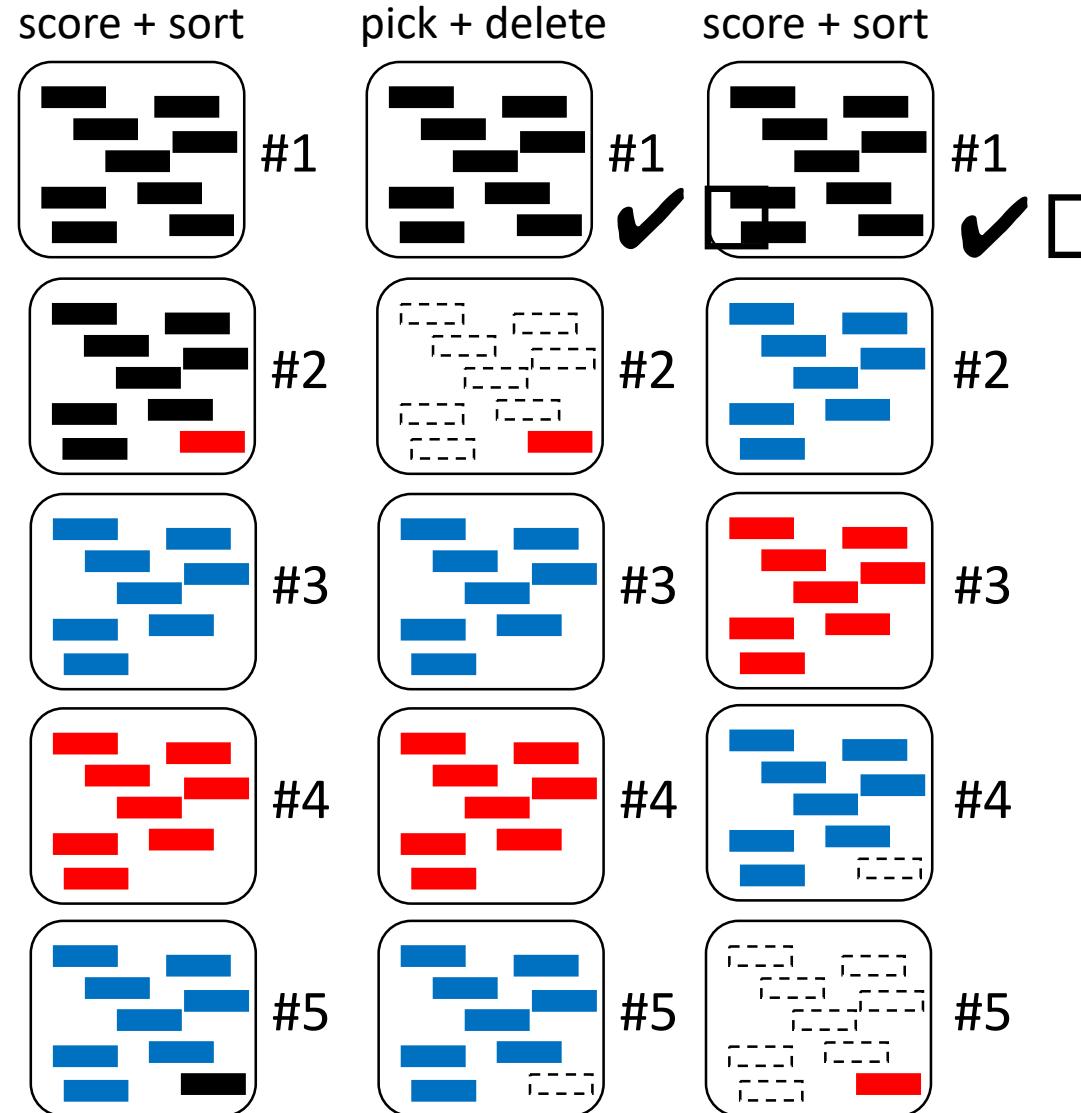


# Multi-binning strategy: pick winner

repeat

DAS Tool:

1. Score bins
2. Sort from best to worse
3. Pick top bin
4. Delete identical contigs in other bins

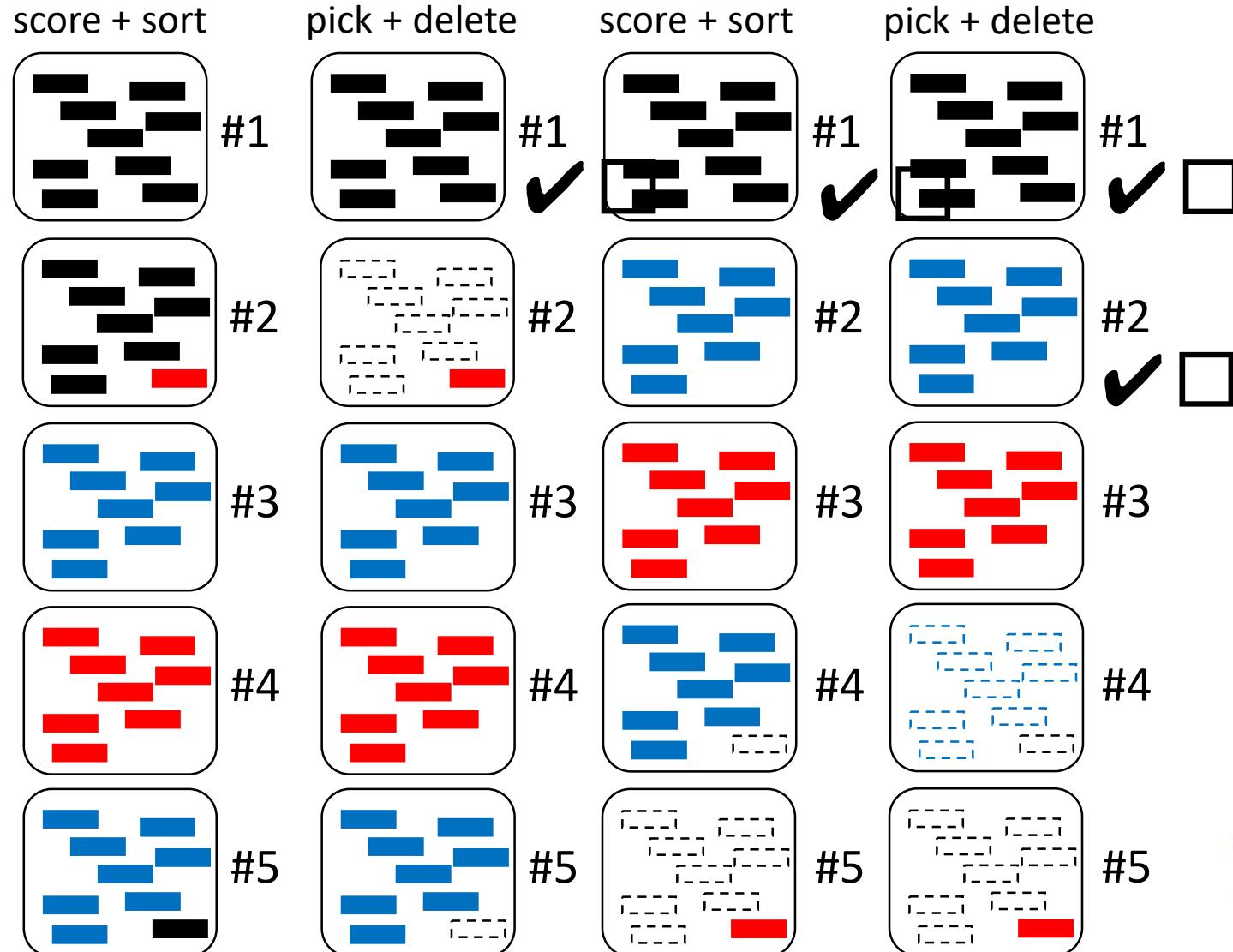


# Multi-binning strategy: pick winner

repeat

DAS Tool:

1. Score bins
2. Sort from best to worse
3. Pick top bin
4. Delete identical contigs in other bins



# Bin evaluation

---

Determine presence of a set of single copy core genes:

- Present/Missing = completeness/incompleteness
- Duplicates = strain heterogeneity (co-binned closely-related genomes)
- Phyla mismatches (duplicates or unique) = contamination

Example: CheckM (<https://ecogenomics.github.io/CheckM/>)



# Estimating genome completeness

---

- **Completeness** = percentage of single copy core genes present out of a given set
- Is “often” equivalent to expected genome size based on close relatives
- **CAVEAT:** genome sizes can vary substantially between close relatives



# Automated binning!!!

---

MetaBAT (<https://bitbucket.org/berkeleylab/metabat>):

- Partial multi-parameter binning tool
- Combines:
  - tetranucleotide frequencies
  - (Differential) coverage
- Iterative binning using a modified k-medoid (like k-means) clustering algorithm: iteratively forms clusters, keeps large clusters, dissolves small clusters, and then attempts to recruit ex-small cluster contigs to large clusters
- Does not use taxonomy

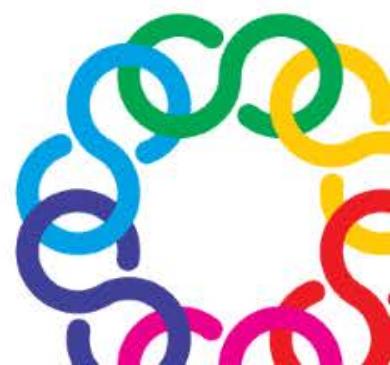


# Automated binning!!!

---

**CONCOCT** (<https://github.com/BinPro/CONCOCT>):

- Partial multi-parameter binning tool
- Combines:
  - tetranucleotide frequencies
  - (Differential) coverage
- Gaussian mixture models used to cluster contigs
- Cluster evaluation: checks bins for a set of single copy core genes to determine bin completeness and purity
- Does not use taxonomy



# Task: Binning

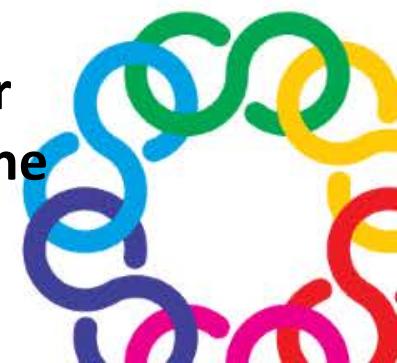
---

[Go to Github MSS webpage](#)

Tasks:

- ✓  Short contig removal
- ✓  Read mapping
  - Automated binning (using Metabat and Maxbin)
  - Pick-the-winner bin dereplication (using DAS Tool)
  - Bin evaluation (using CheckM)

**NOTE:** we usually also use CONCOCT which tends to produce a smaller number of high quality bins. Running it is a little more involved, so it's excluded from the MSS, but we encourage everyone to consider using this tool.



# Binning: other considerations



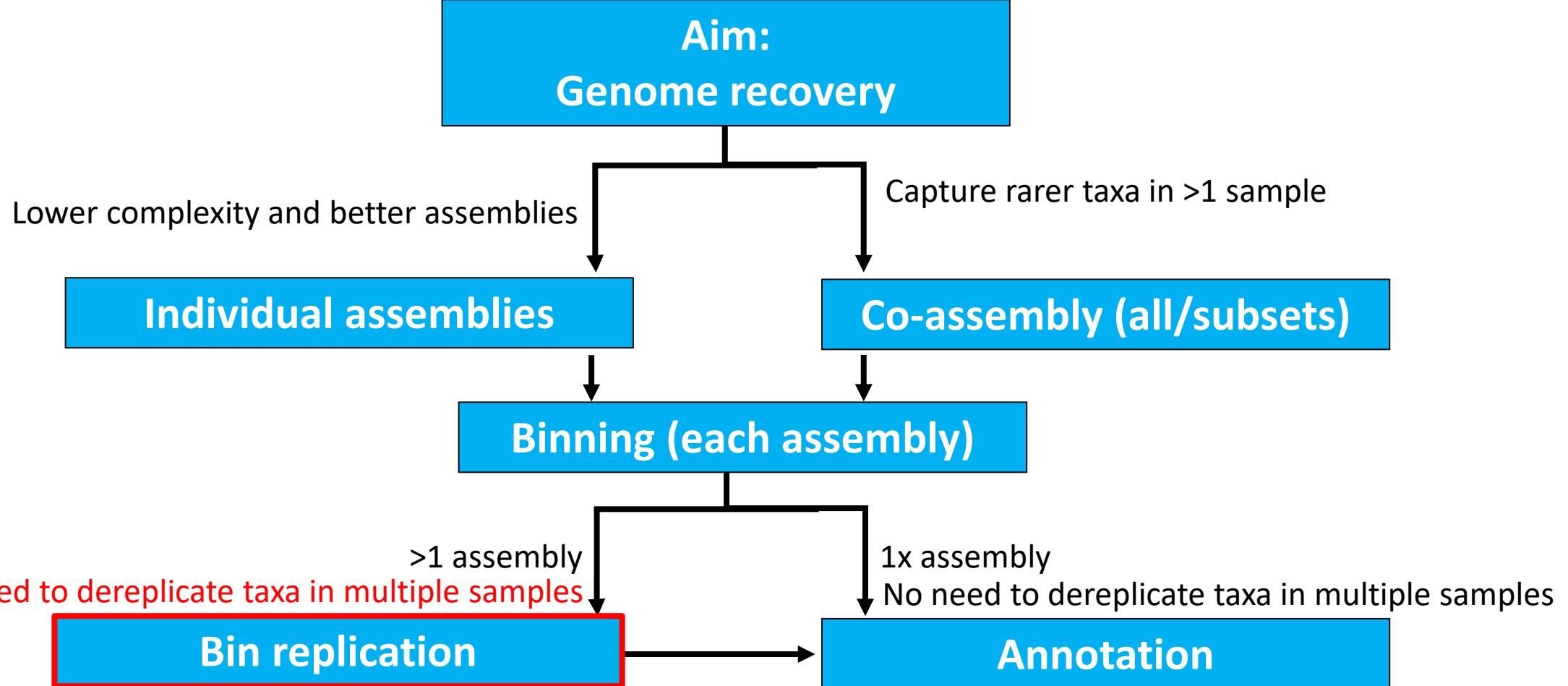
# Binning: other considerations

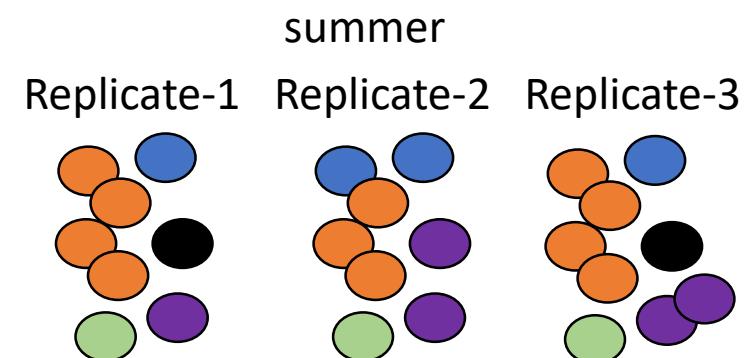
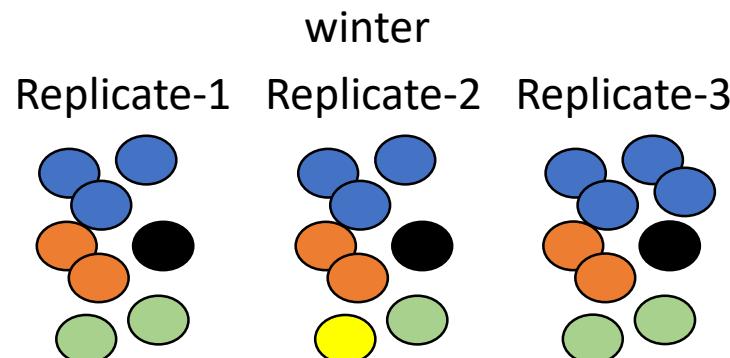
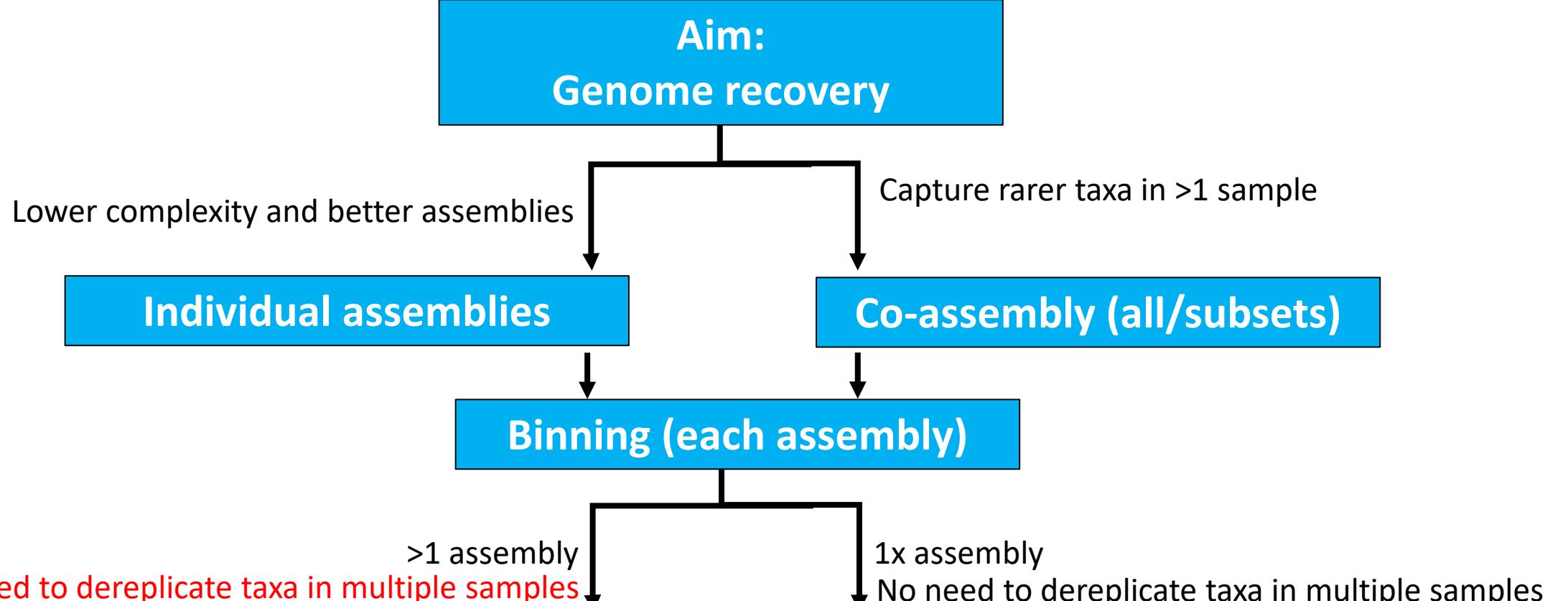
---

Dealing with:

- **Duplicate bins across multiple assemblies (samples)**
- **Viral and eukaryotic bins**
- **Unassembled data**
- **Organisms that possess minimal genomes**



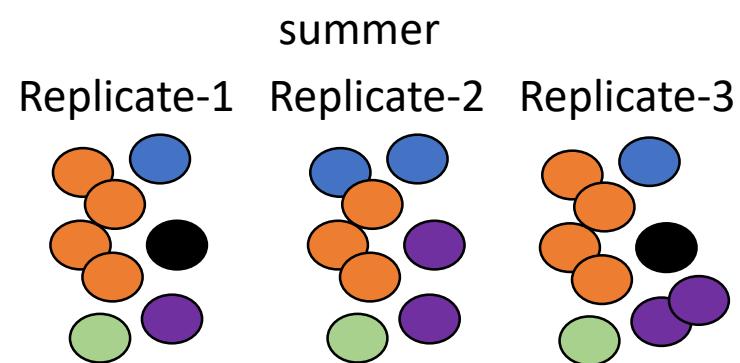
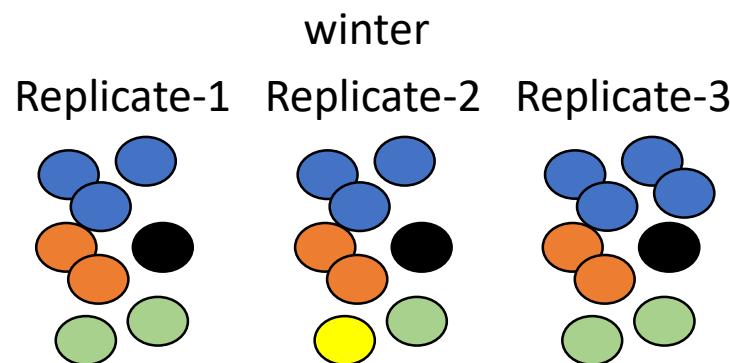




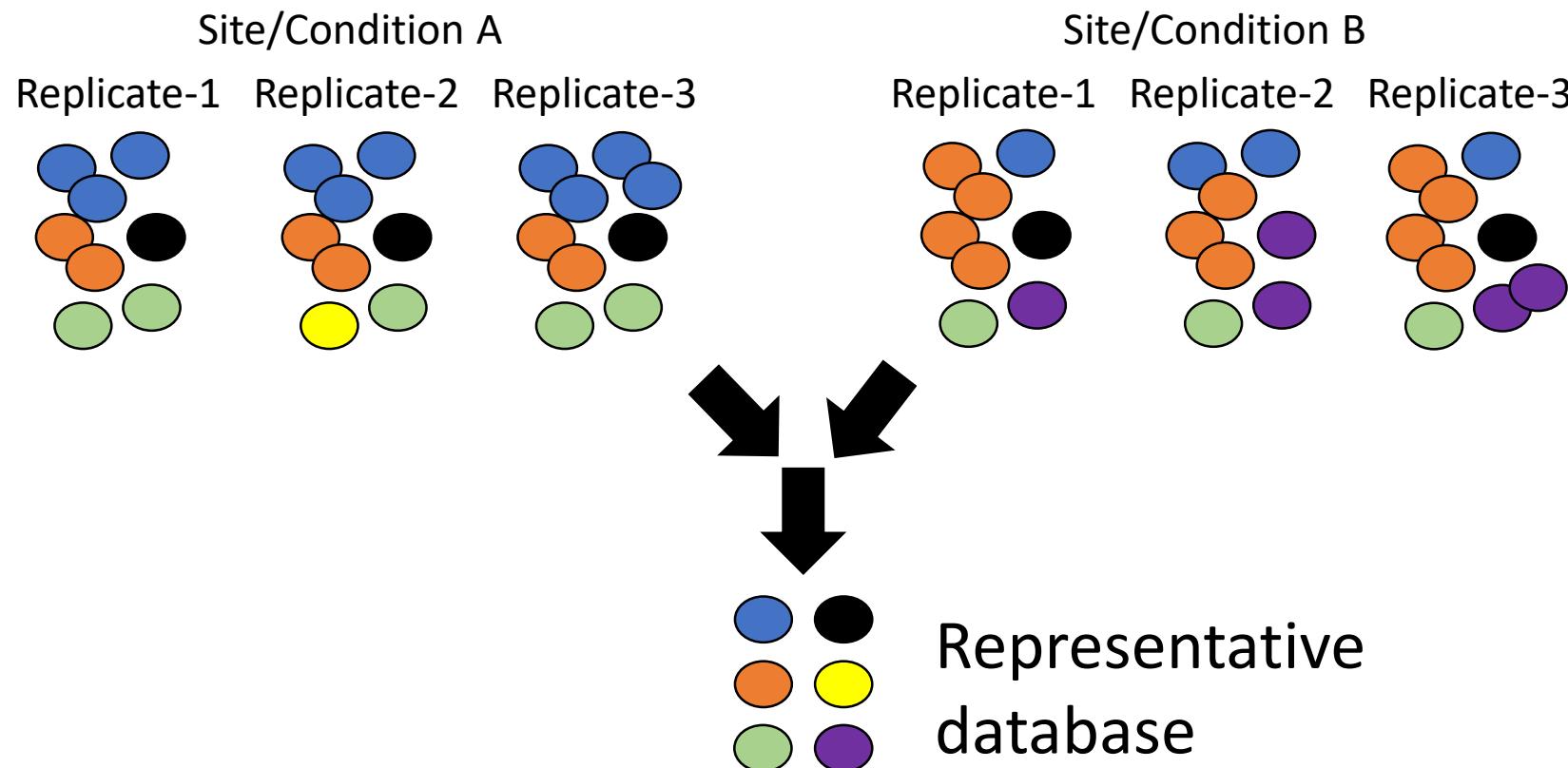
# Multiple samples: duplicate genomes

---

- Remove repeat genomes
- For example, separate assemblies of the communities below would yield duplicate genomes for: 



# Multiple samples: duplicate genomes



**Goal:**

- work with 1 genomic database
- generate sample-genome coverage table
- if applicable, map transcript data and generate counts table

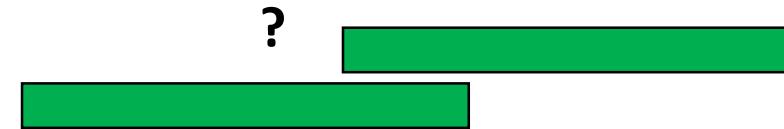


# Multiple samples: duplicate genomes

---

**Option 1:** search contigs between datasets

- Problem: dealing with partially overlapping contigs

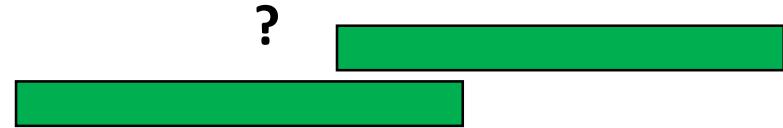


# Multiple samples: duplicate genomes

---

**Option 1:** search contigs between datasets

- Problem: dealing with partially overlapping contigs



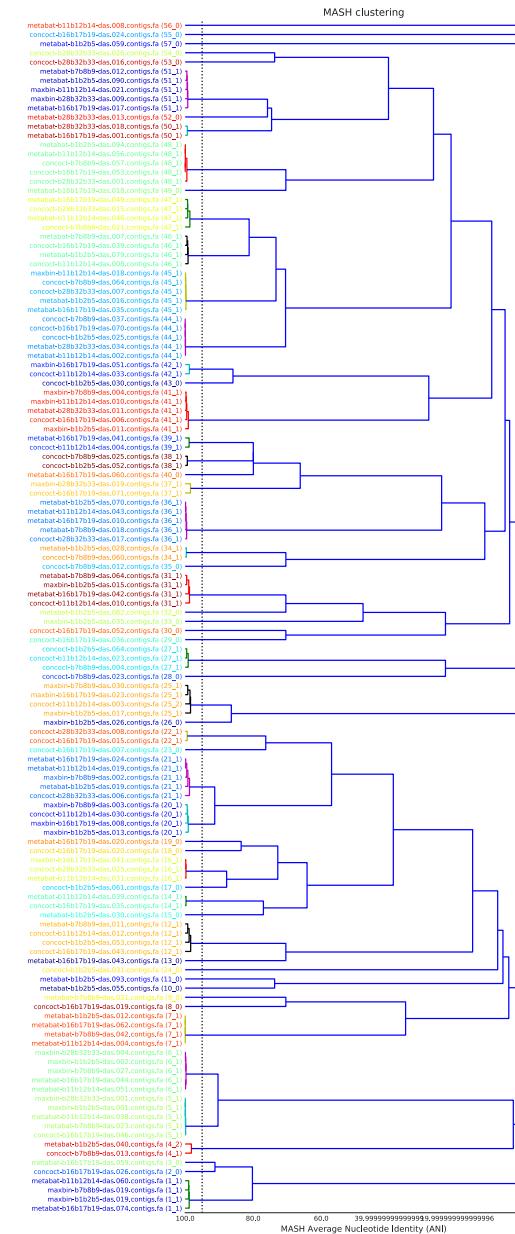
**Option 2:** search bins between datasets (e.g. dRep, Olm et al., 2017, ISME J)

- Step 1: Fast pairwise MASH search of bins → identify clusters
- Step 2: Slow gANI (average nucleotide identity) pairwise search of bins in primary clusters



## **STEP 1**

## MASH primary cluster creation

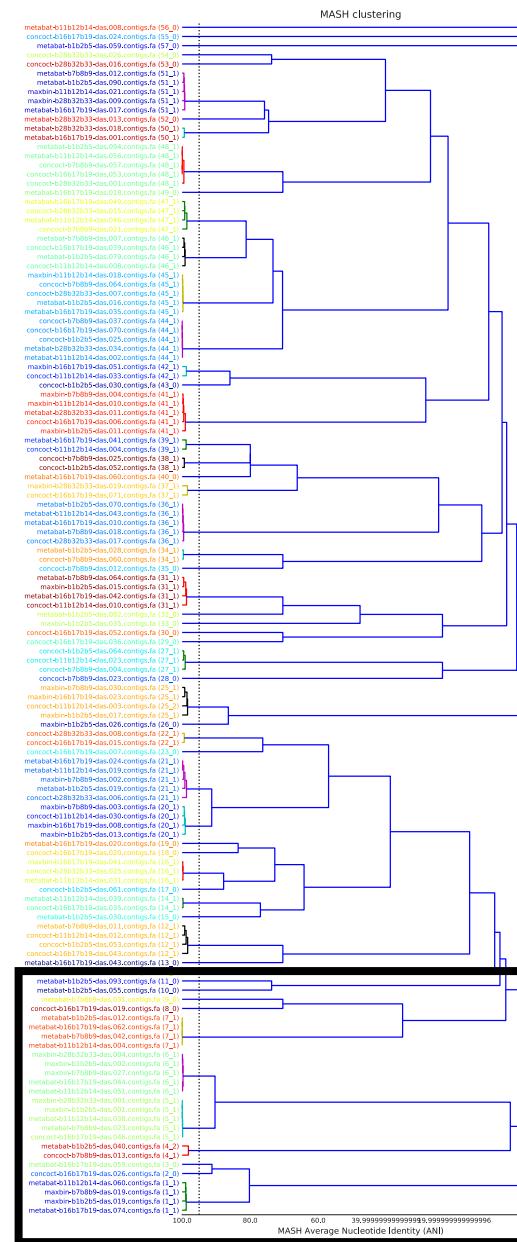


# Genomes from a cyanobacterial bloom (Wai-iti River, Nelson)



# STEP 1

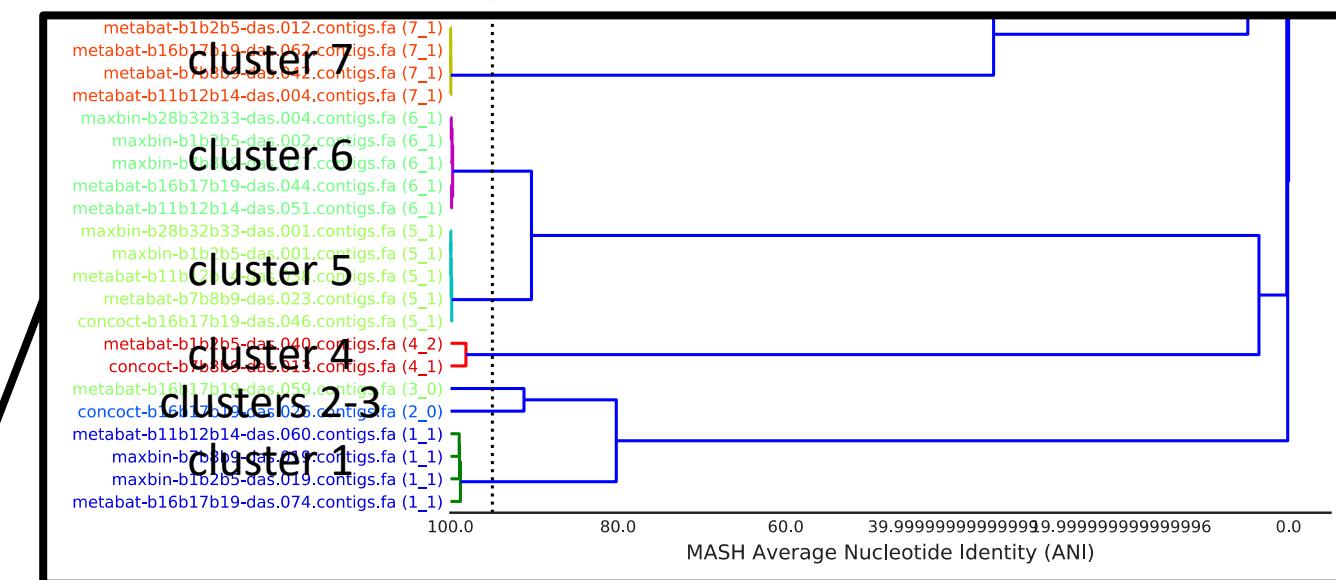
## MASH primary cluster creation



# Genomes from a cyanobacterial bloom (Wai-it River, Nelson)

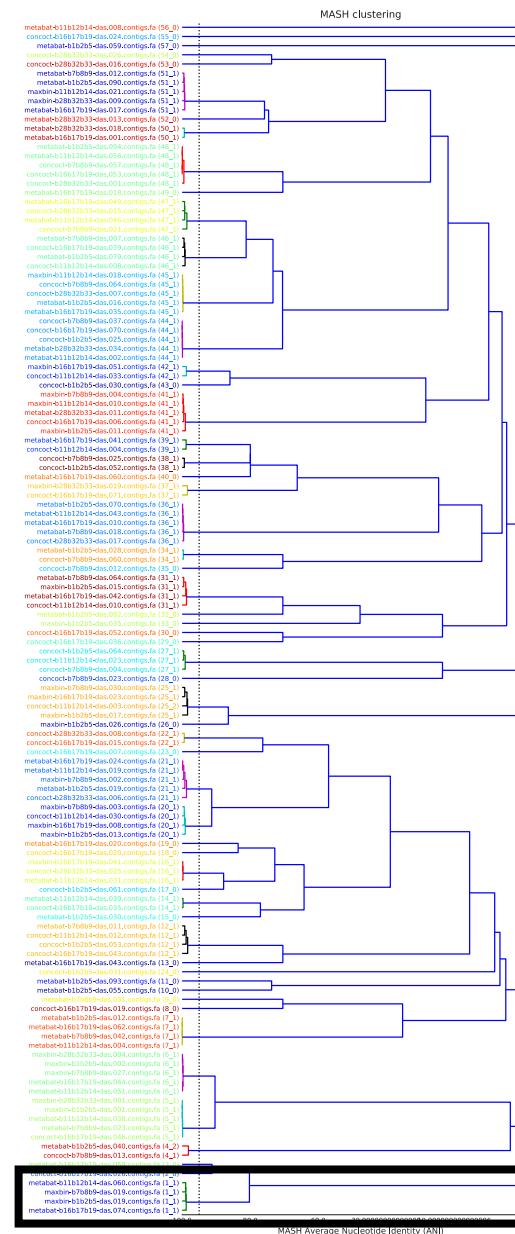


90% ANI cluster cutoff



## STEP 1

MASH primary cluster creation



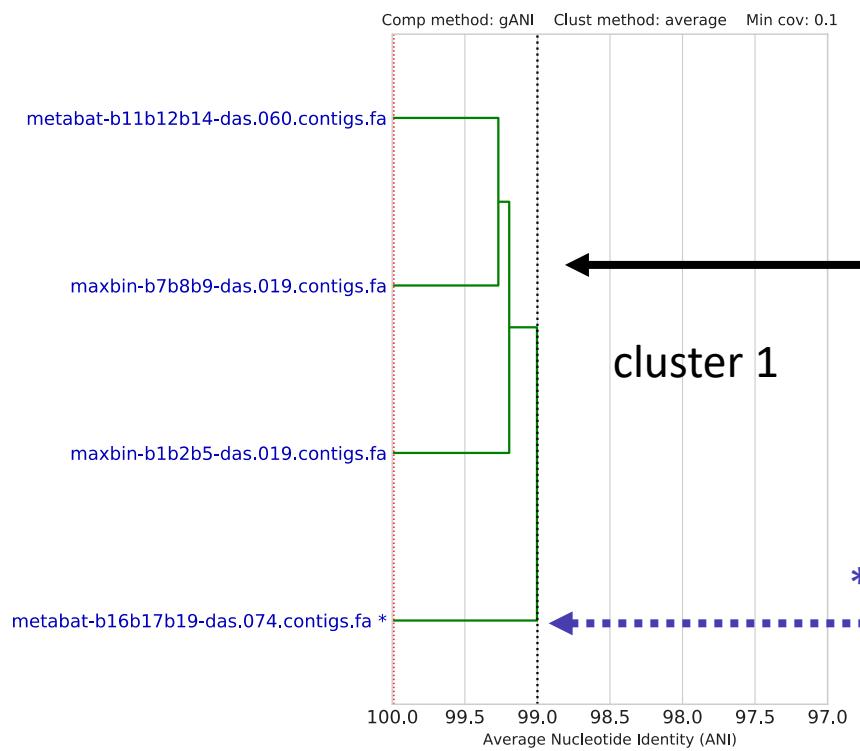
Genomes from a cyanobacterial bloom  
(Wai-iti River, Nelson)



## STEP 2

Secondary clustering

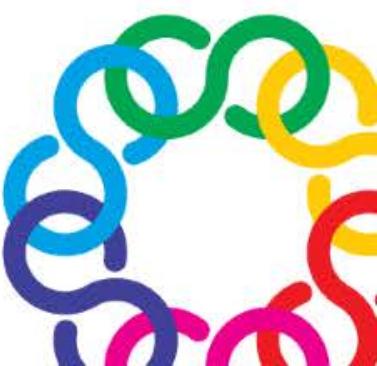
99% gANI cluster cutoff



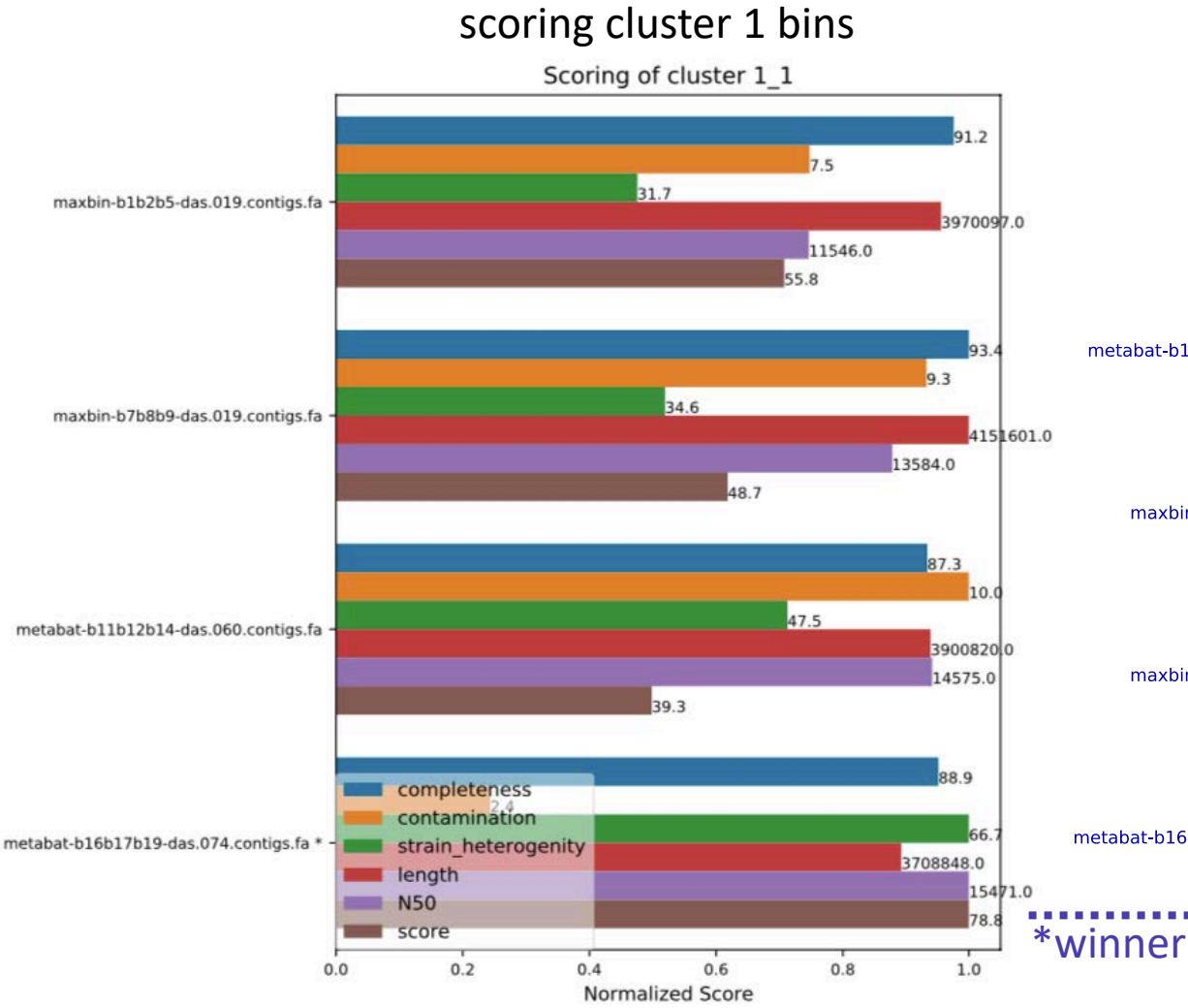
Re-cluster  
cluster (gANI  
method)

- All >99% similar →
- DerePLICATE
  - Pick best scoring bin

\*winner

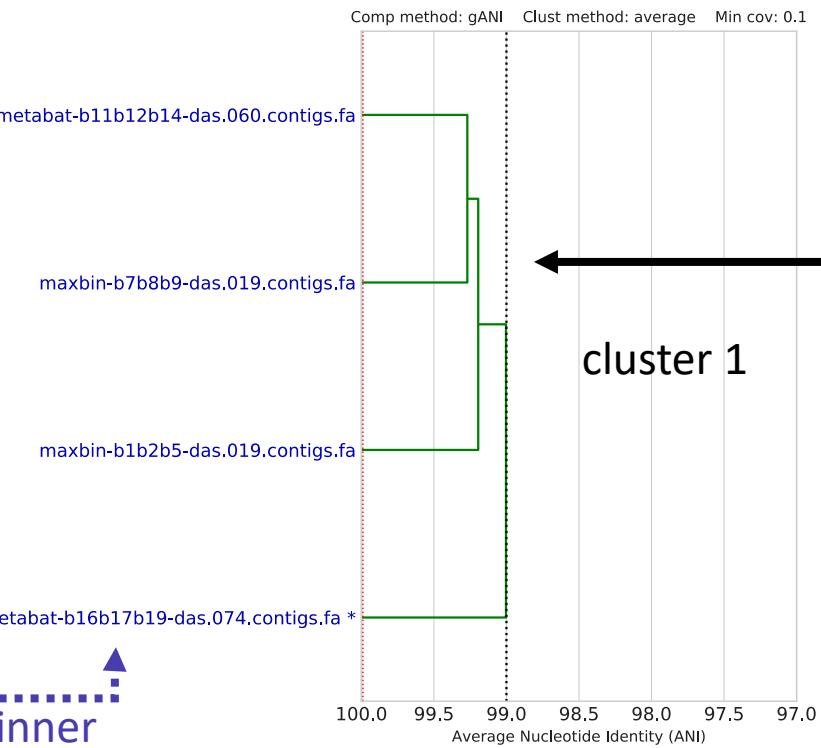


# Genomes from a cyanobacterial bloom (Wai-iti River, Nelson)



## STEP 2 Secondary clustering

99% ANI cluster cutoff



- All >99% similar →
- DerePLICATE
  - Pick best scoring bin

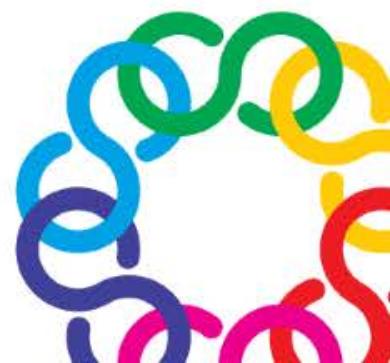


# Eukaryotic and viral bins

---

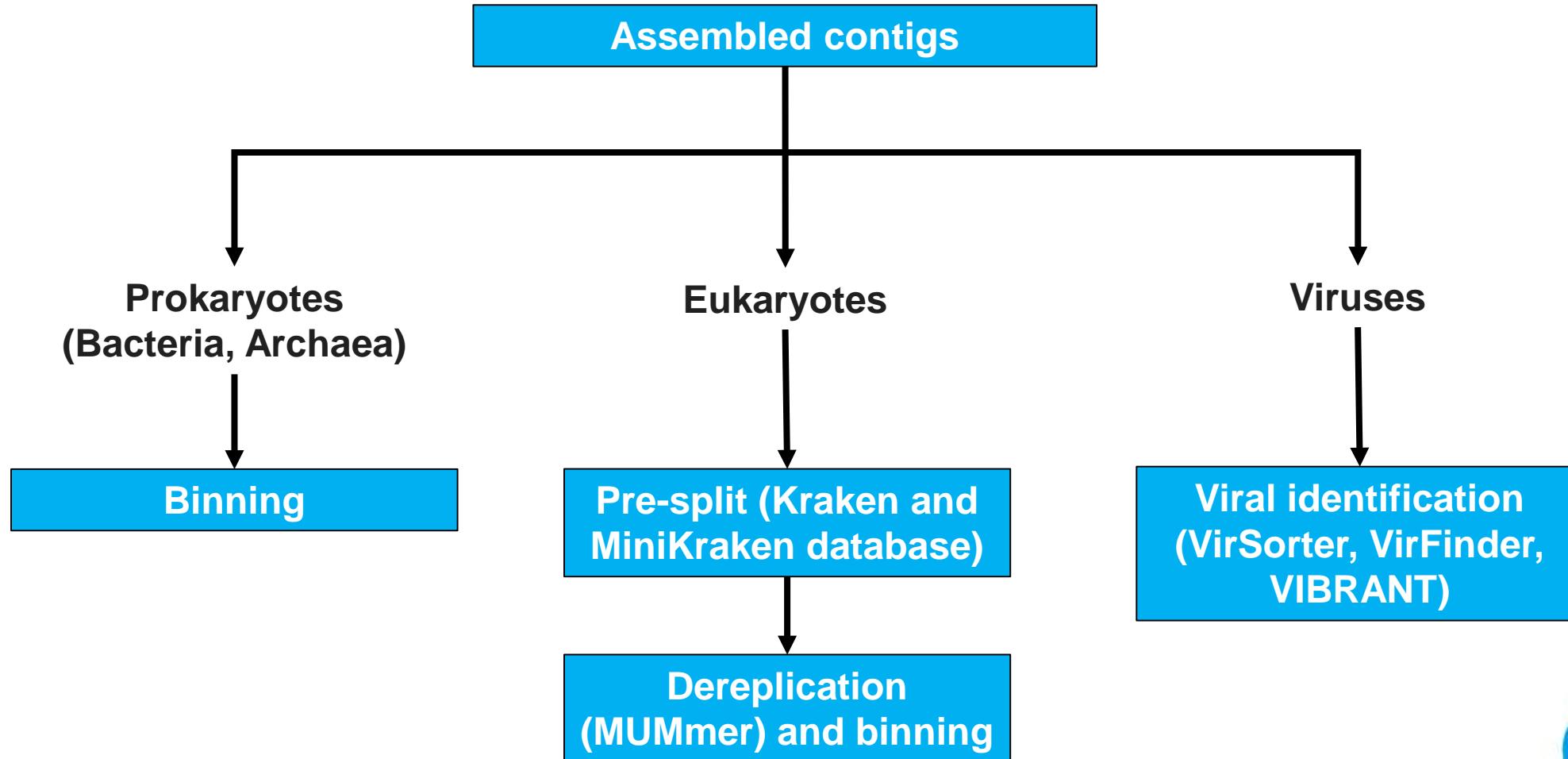
**Not included** in output of tools using prokaryotic single copy core genes for bin evaluation:

- CheckM
- DAS Tool
- dRep

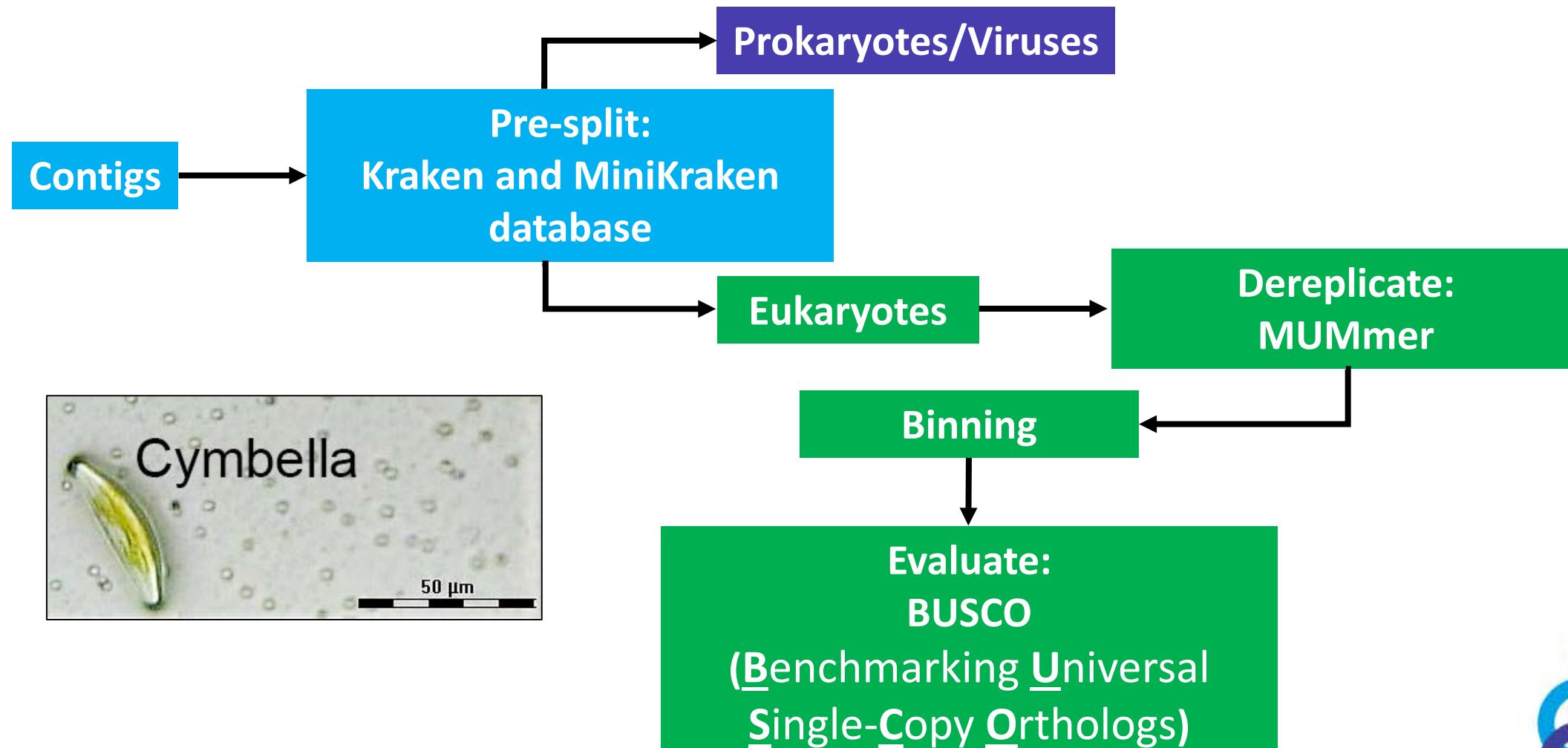


# Recovering genomes

---



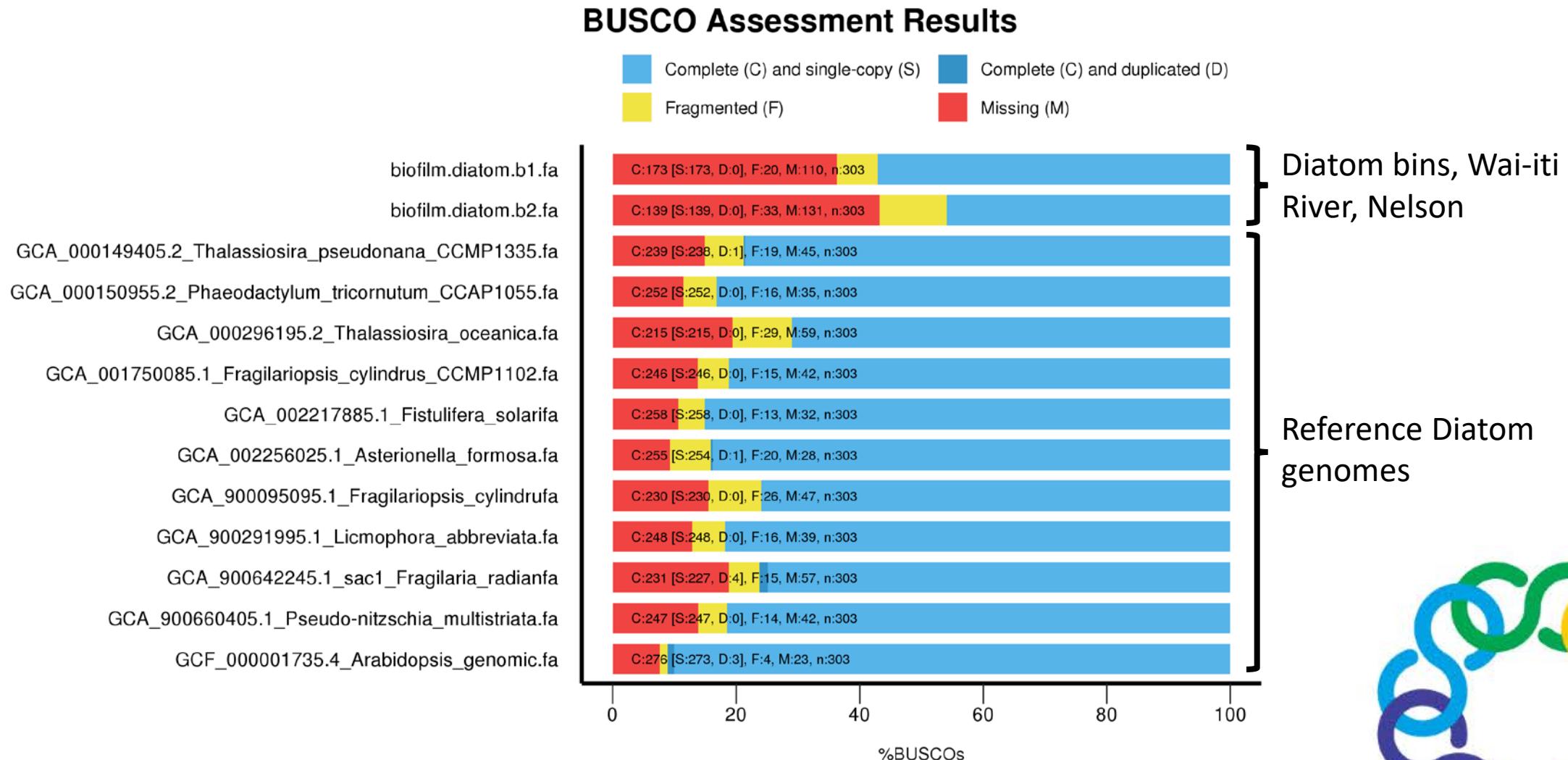
# Eukaryotic and viral bins



<https://busco.ezlab.org/>

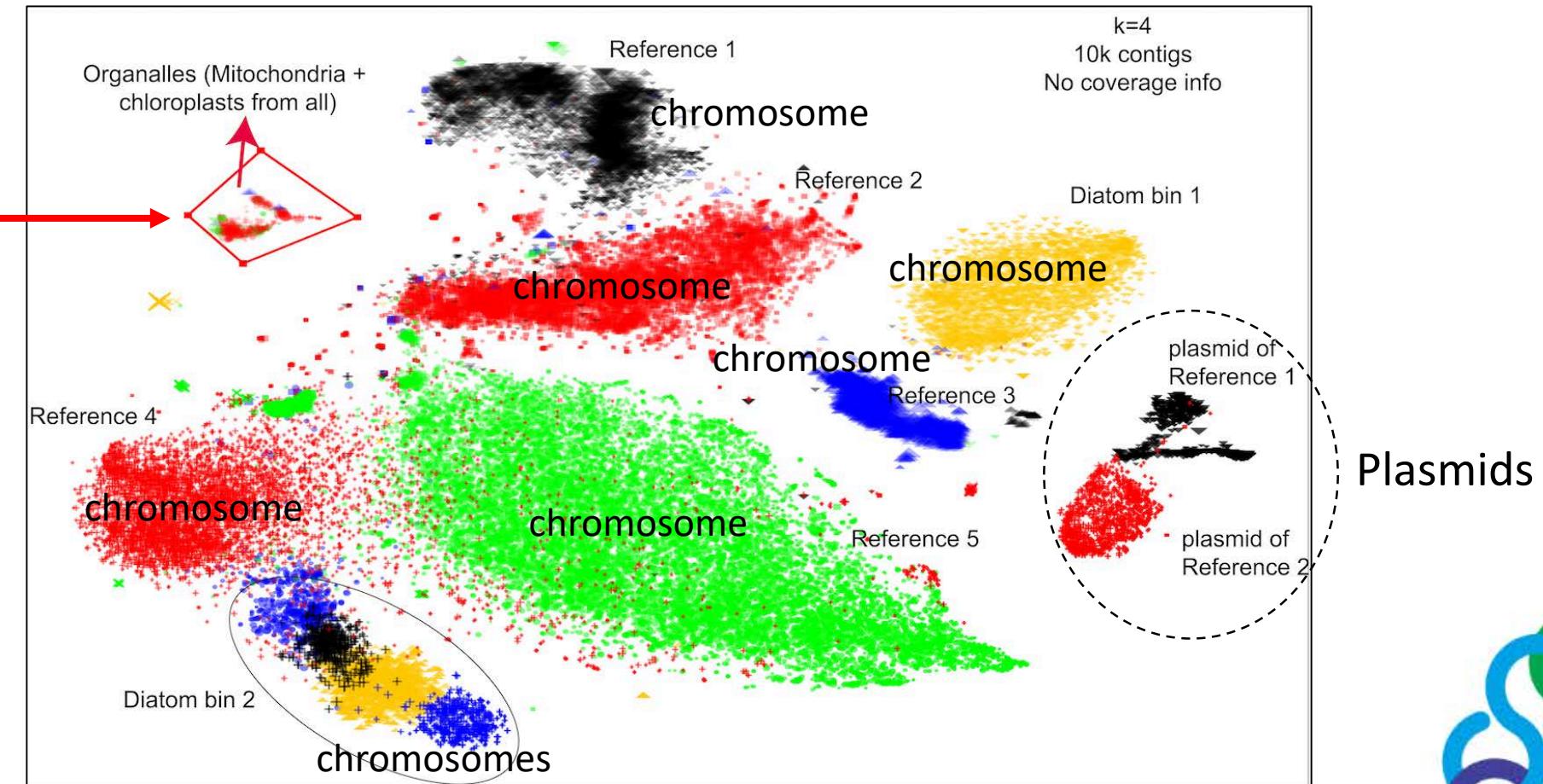


# SCG assessment of diatom genomes



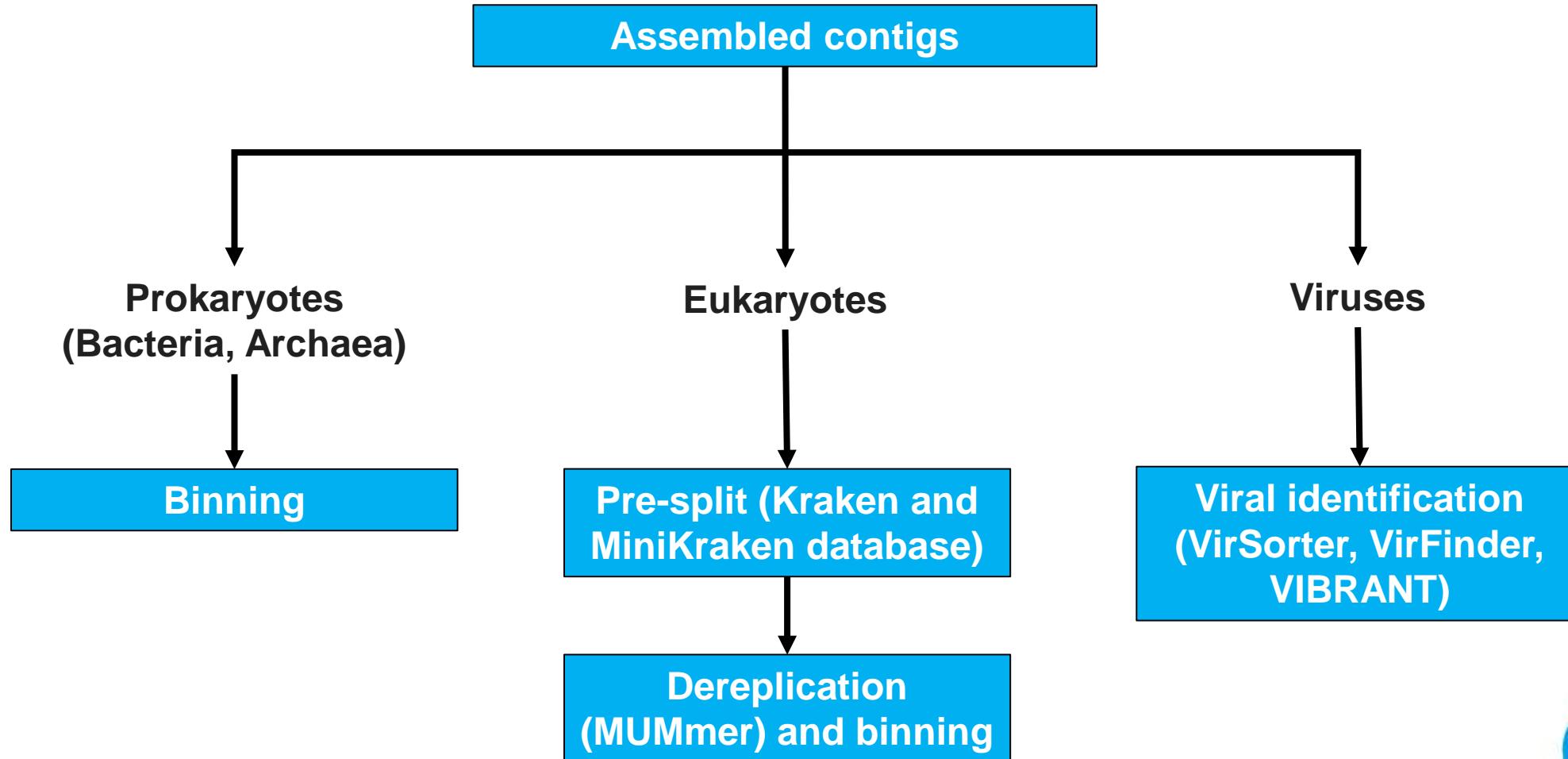
# Caveat: Organelles and plasmids

Mitochondria  
and chloroplast  
contigs have  
distinct  
compositional  
signatures from  
chromosomes



# Recovering genomes

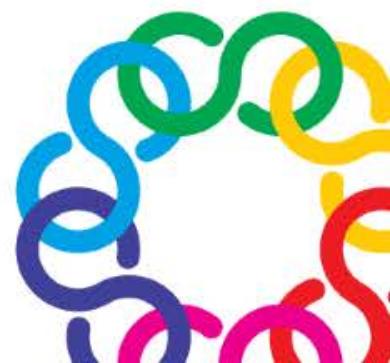
---



# Minimal genomes

---

- What if you're working with tiny bacteria or archaea, with small genomes?
- Some lack certain single copy core genes (SCCGs), e.g. certain ribosomal proteins



# Minimal genomes

---

- Potential outcome:
  - **unexpectedly poor completeness scores!**
  - **can be eliminated during dereplication steps due to poor scores**
- Solution:
  - recalculate scores
  - exclude troublesome SCCGs
- Example: Patescibacteria superphylum
  - identify reduced set of 43 single copy genes (Brown et al., 2015, Nature)
  - 100 NZ groundwater Patescibacteria genomes: average completeness increased from 60% to 85%



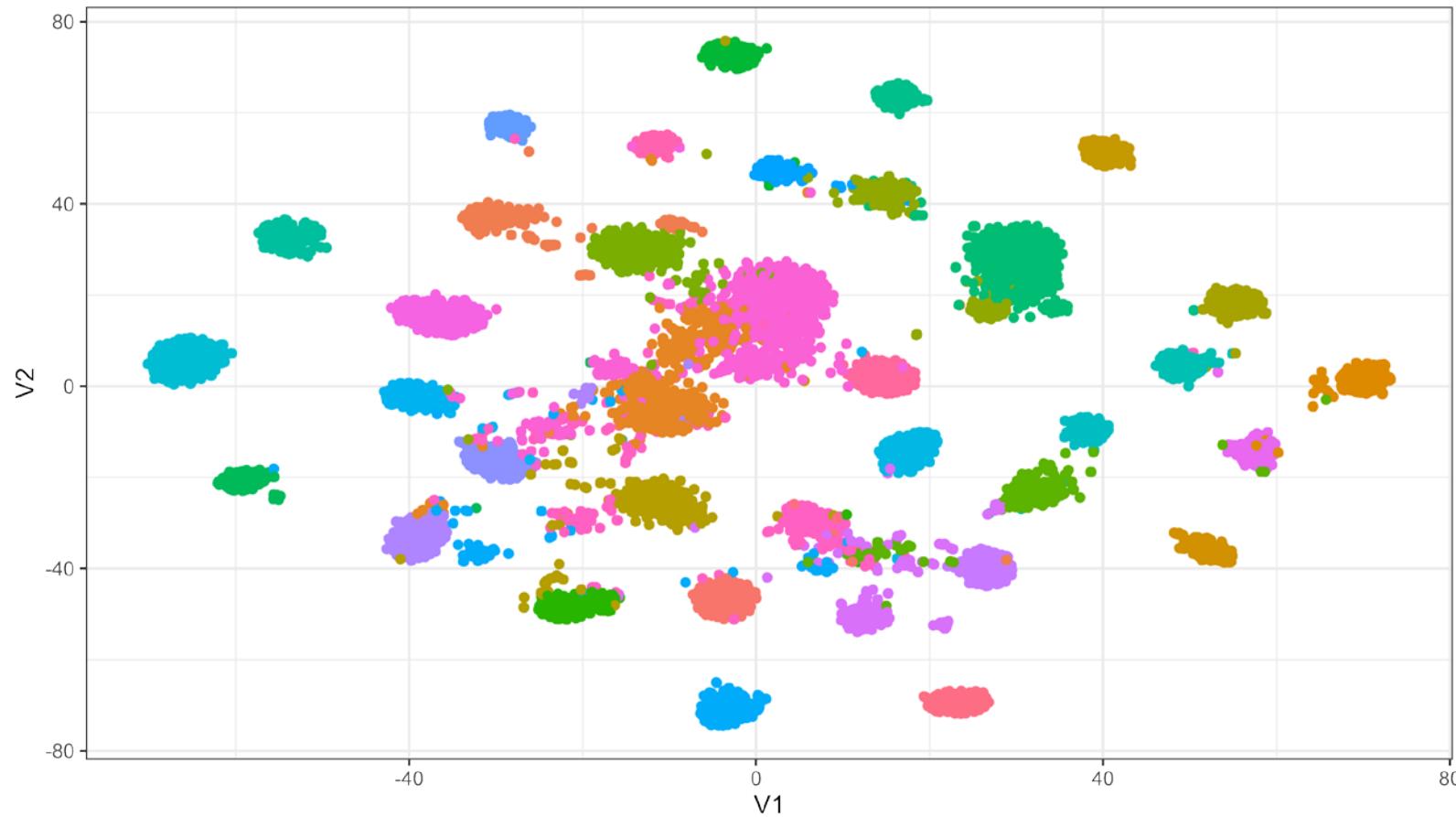
# Bin refinement



# VizBin

---

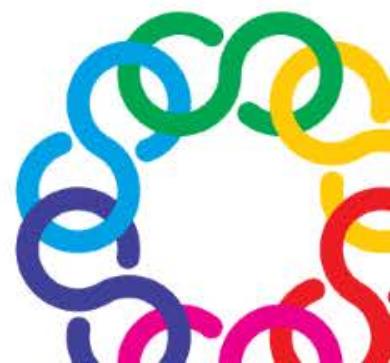
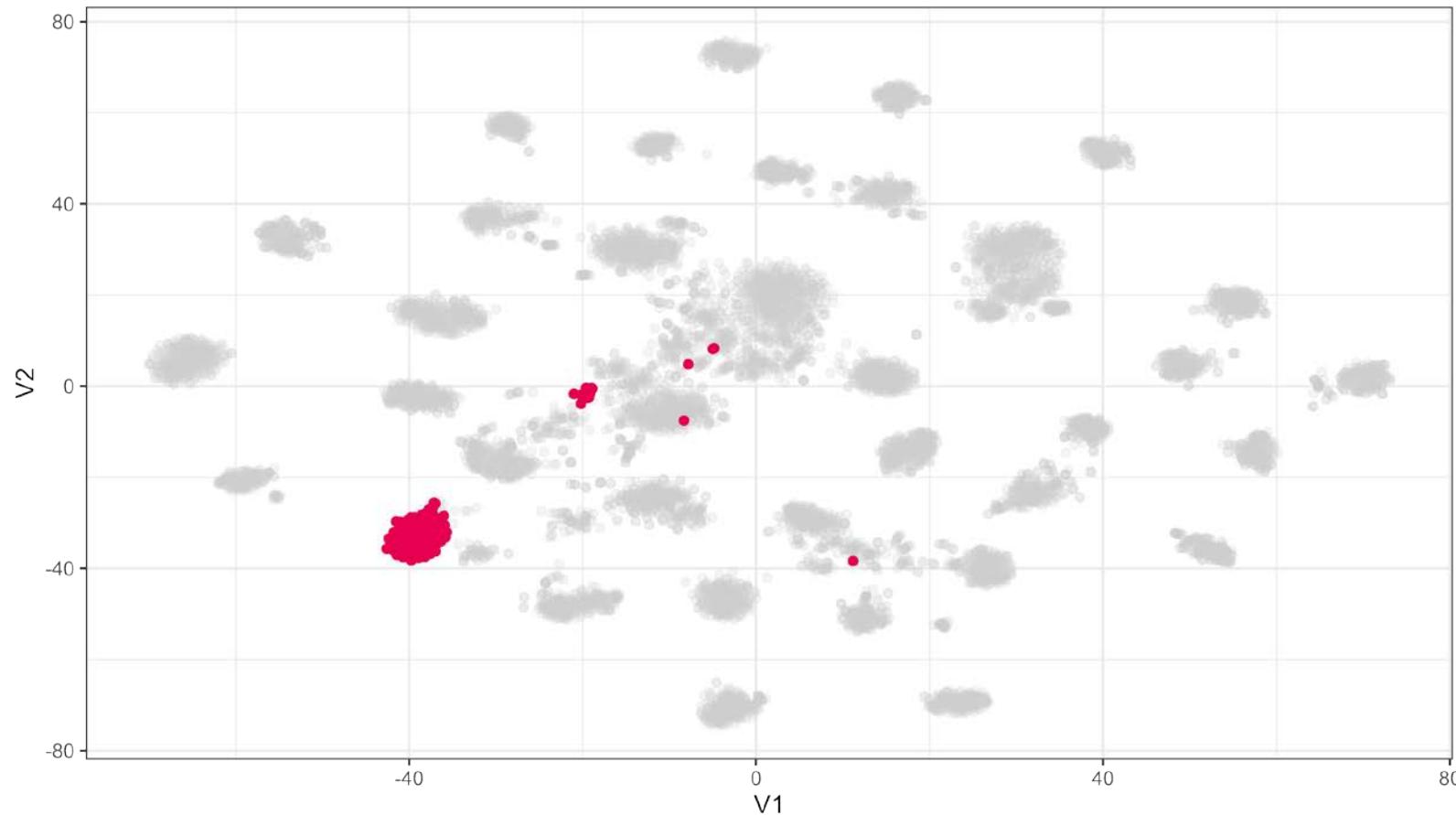
- Inspect bin assignments



# VizBin

---

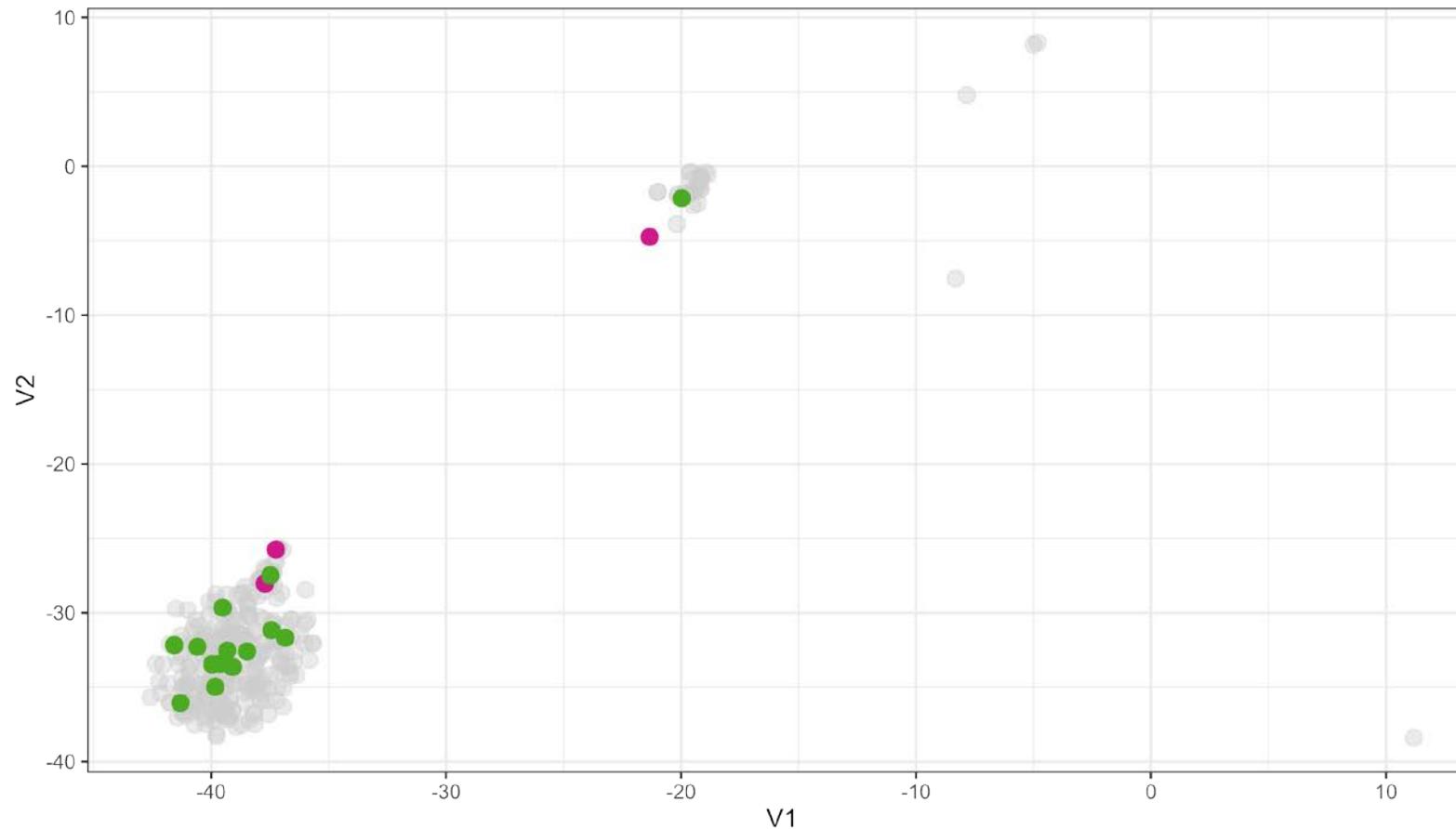
- Use graphical user interface to assign bins or reassign contigs



# VizBin

---

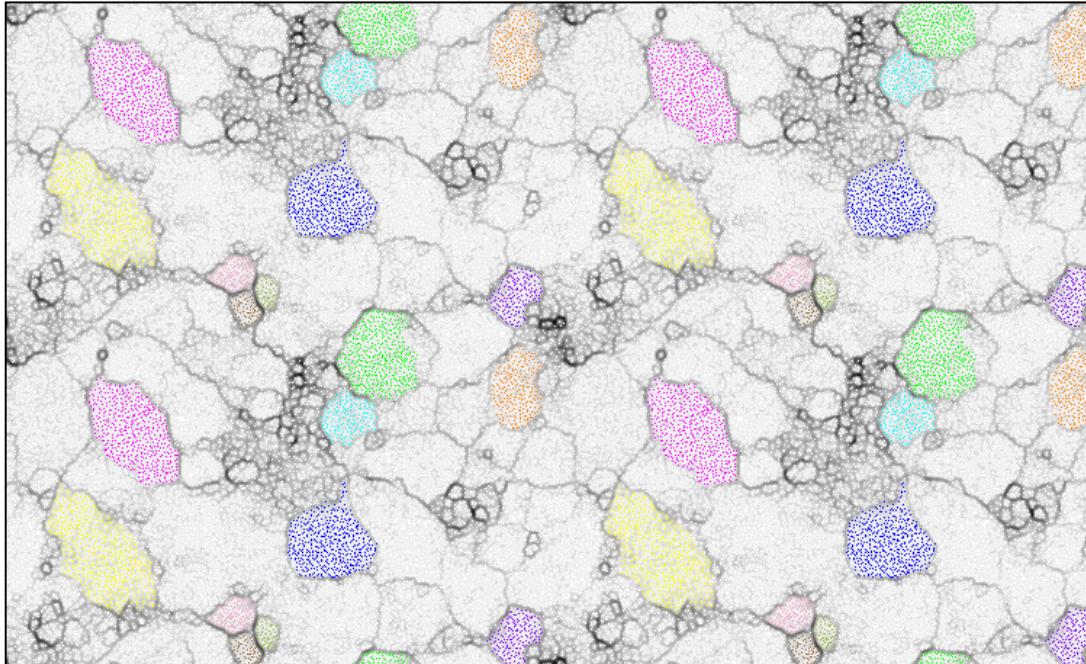
- Identify contigs with unstable placement



# Alternative: ESOM

---

- Emergent Self Organizing Map (ESOM)
- esomana: <http://databionic-esom.sourceforge.net/user.html>

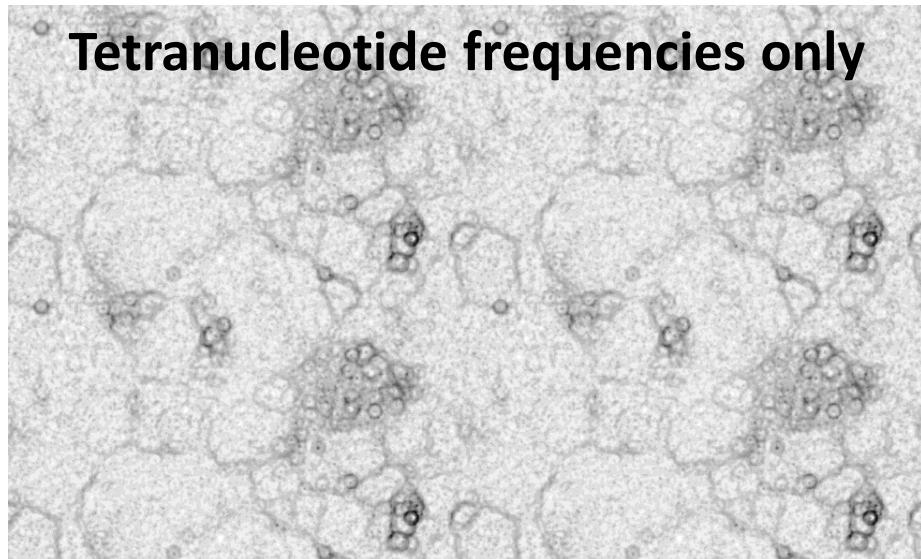


Gulf of Mexico seafloor communities (Handley et al., 2017, ISME J)

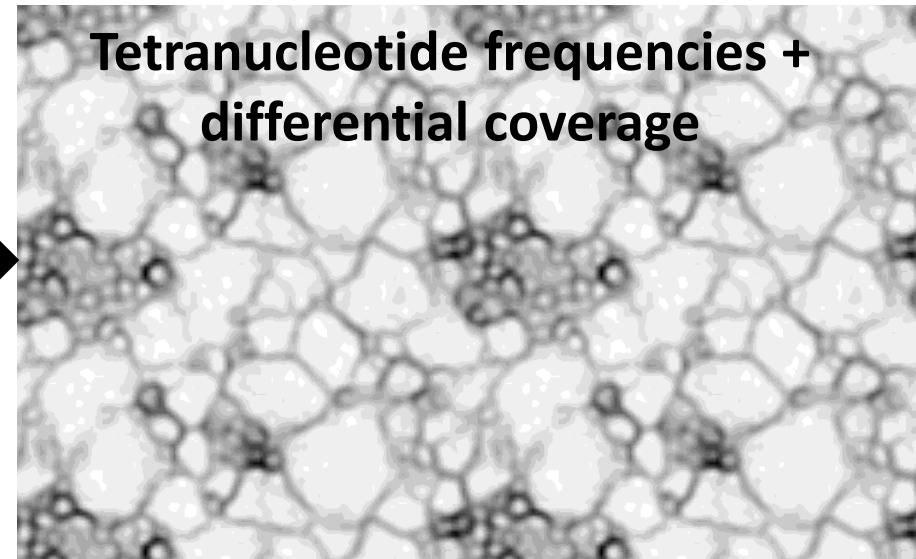


# Alternative: ESOM

- **Flexible: use whatever data you want, e.g.: tetranucleotide frequencies, coverage, both**



Indistinct bin boundaries between highly similar genomes

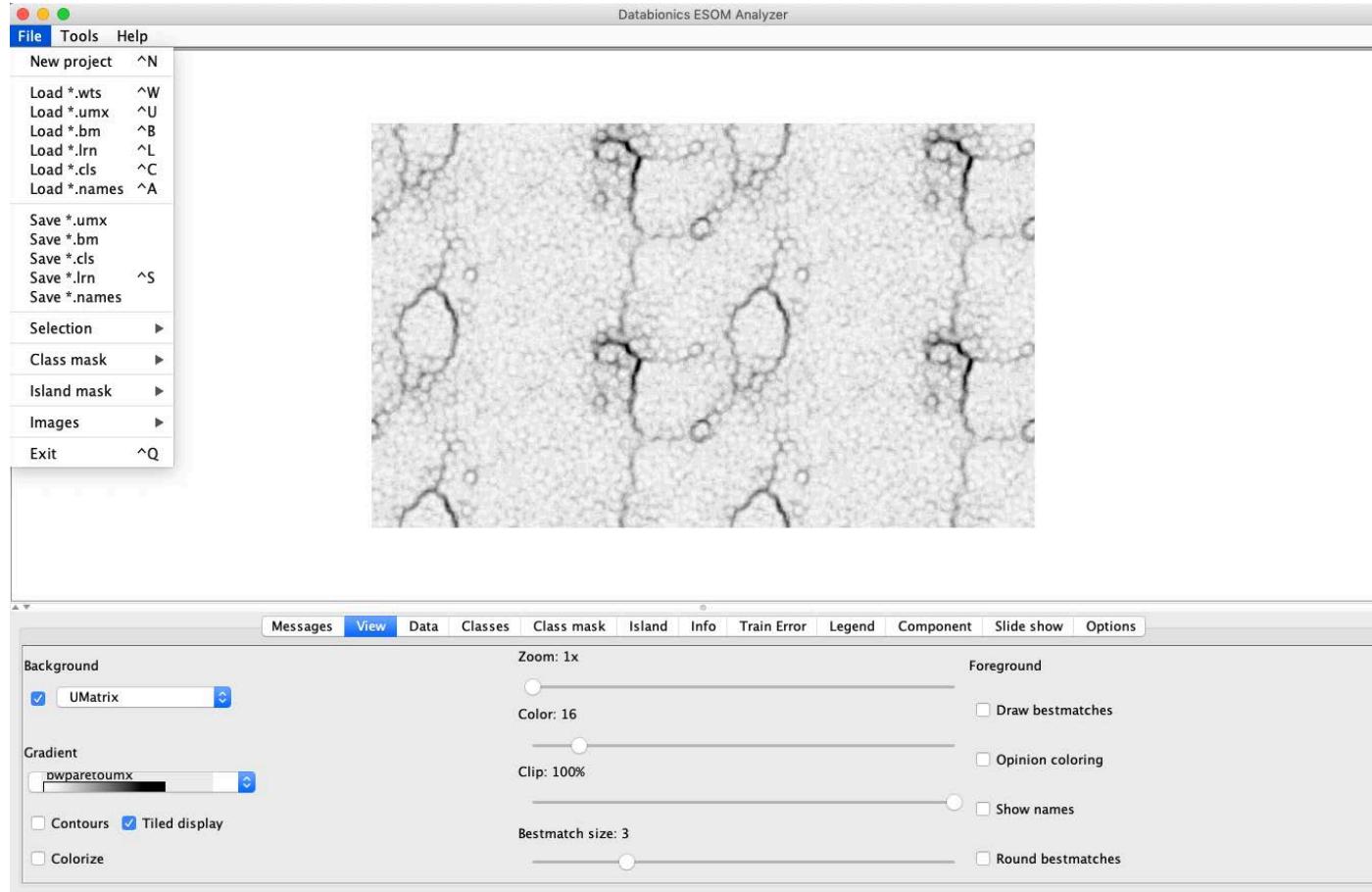


Clear bin boundaries between highly similar genomes using spatial gradient data



# Alternative: ESOM

## Topographic map of clusters

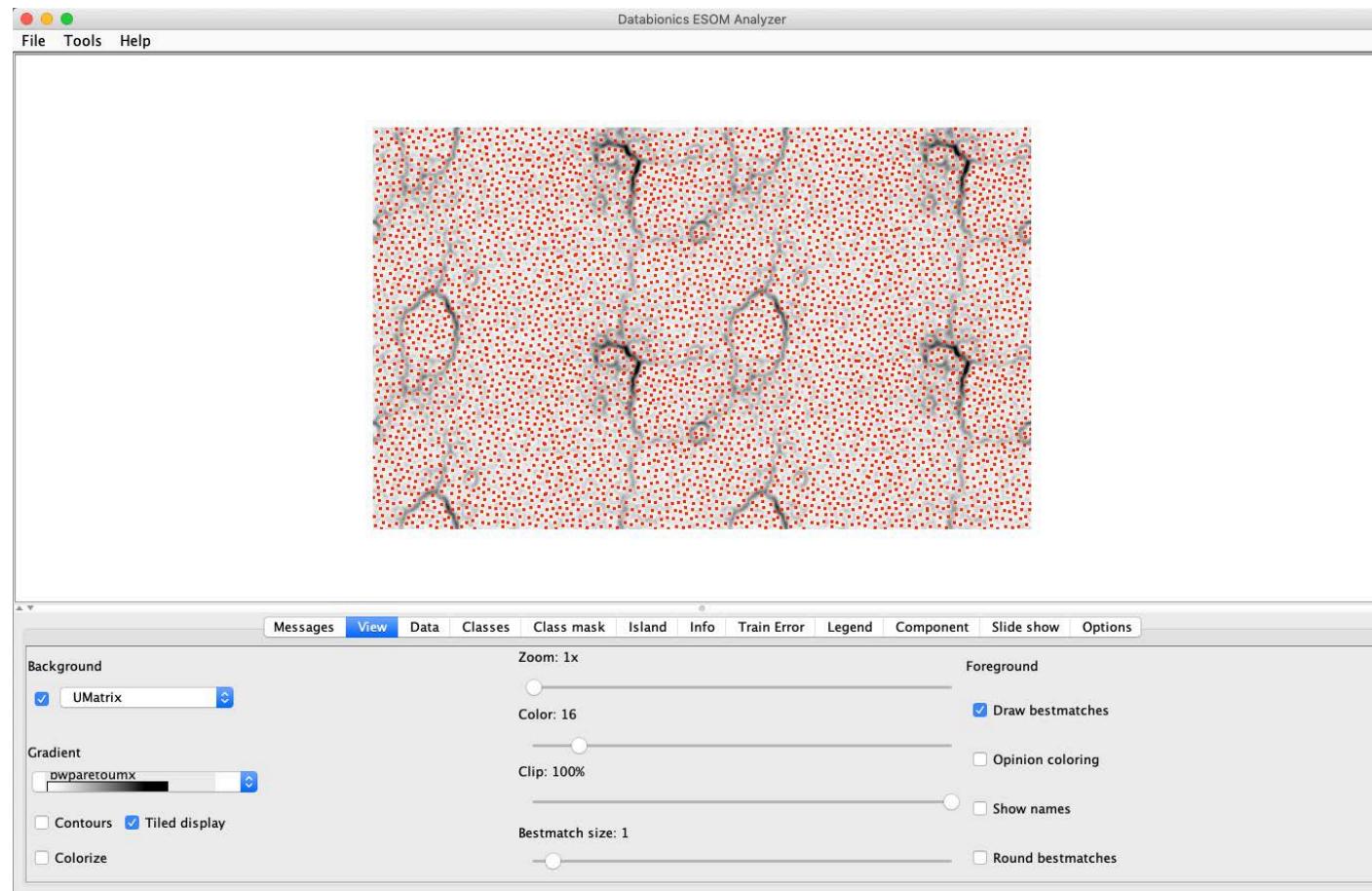


- Uses graphical user interface
- Must supply own prepared data (e.g. pre-calculate tetranucleotide frequencies)
- Dark lines = bin boundaries
- Strong lines = strong bin divisions



# Alternative: ESOM

Map with contig fragments shown in red

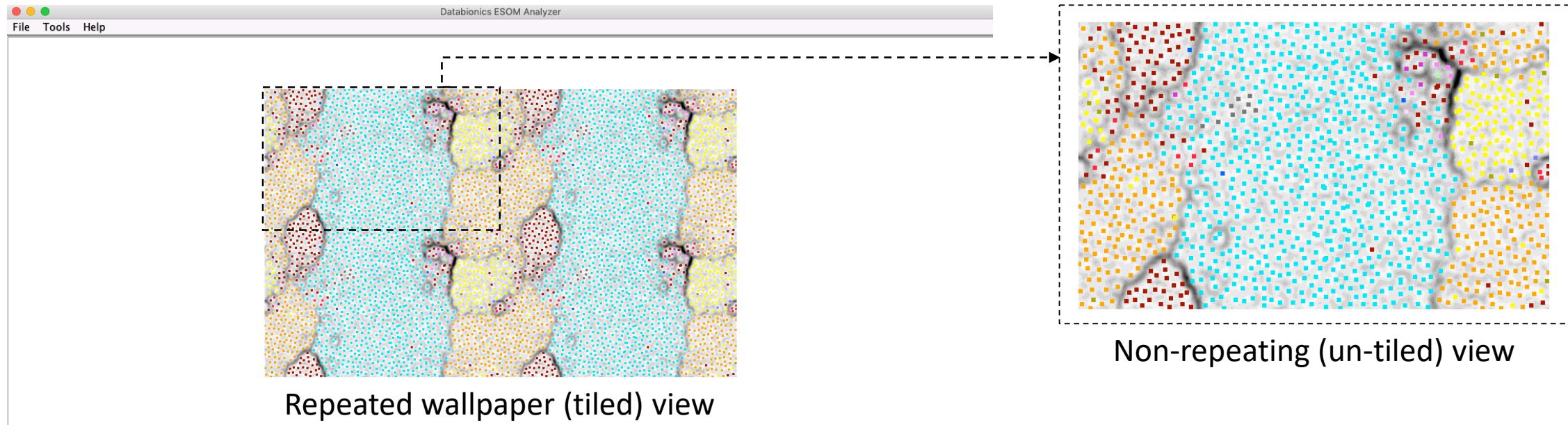


- Uses graphical user interface
- Must supply own prepared data (e.g. pre-calculate tetranucleotide frequencies)
- Dark lines = bin boundaries
- Strong lines = strong bin divisions

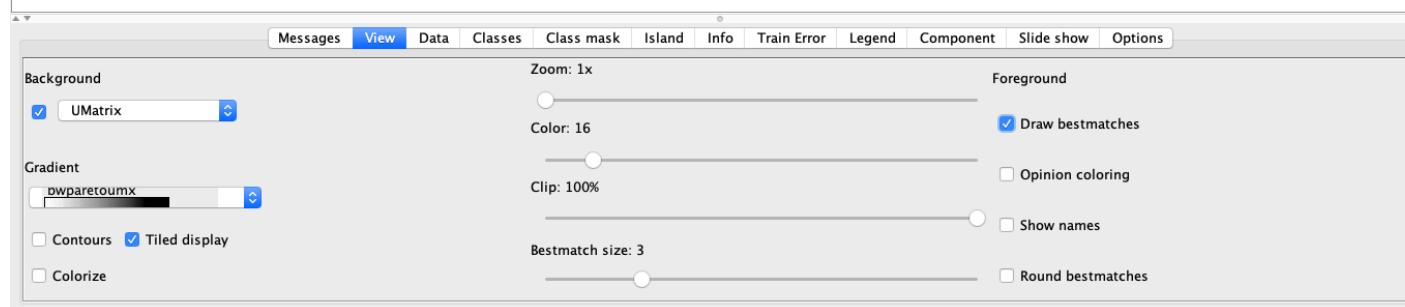


# Alternative: ESOM

Map with contig fragments shown and coloured by bin assignment



Non-repeating (un-tiled) view



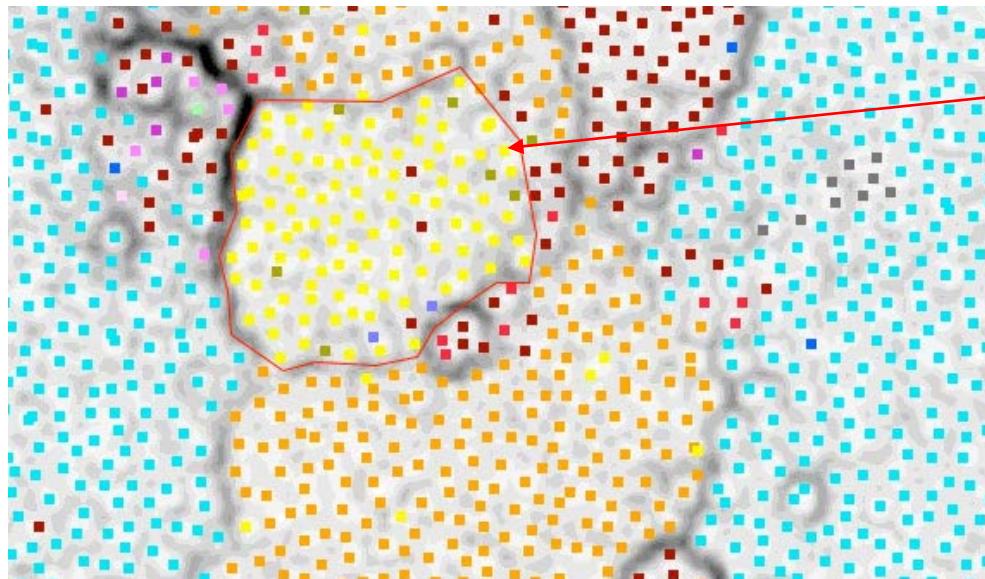
Colorado groundwater communities (Handley et al., 2013, ISME J)



# Alternative: ESOM

---

Map with contig fragments shown and coloured by bin assignment

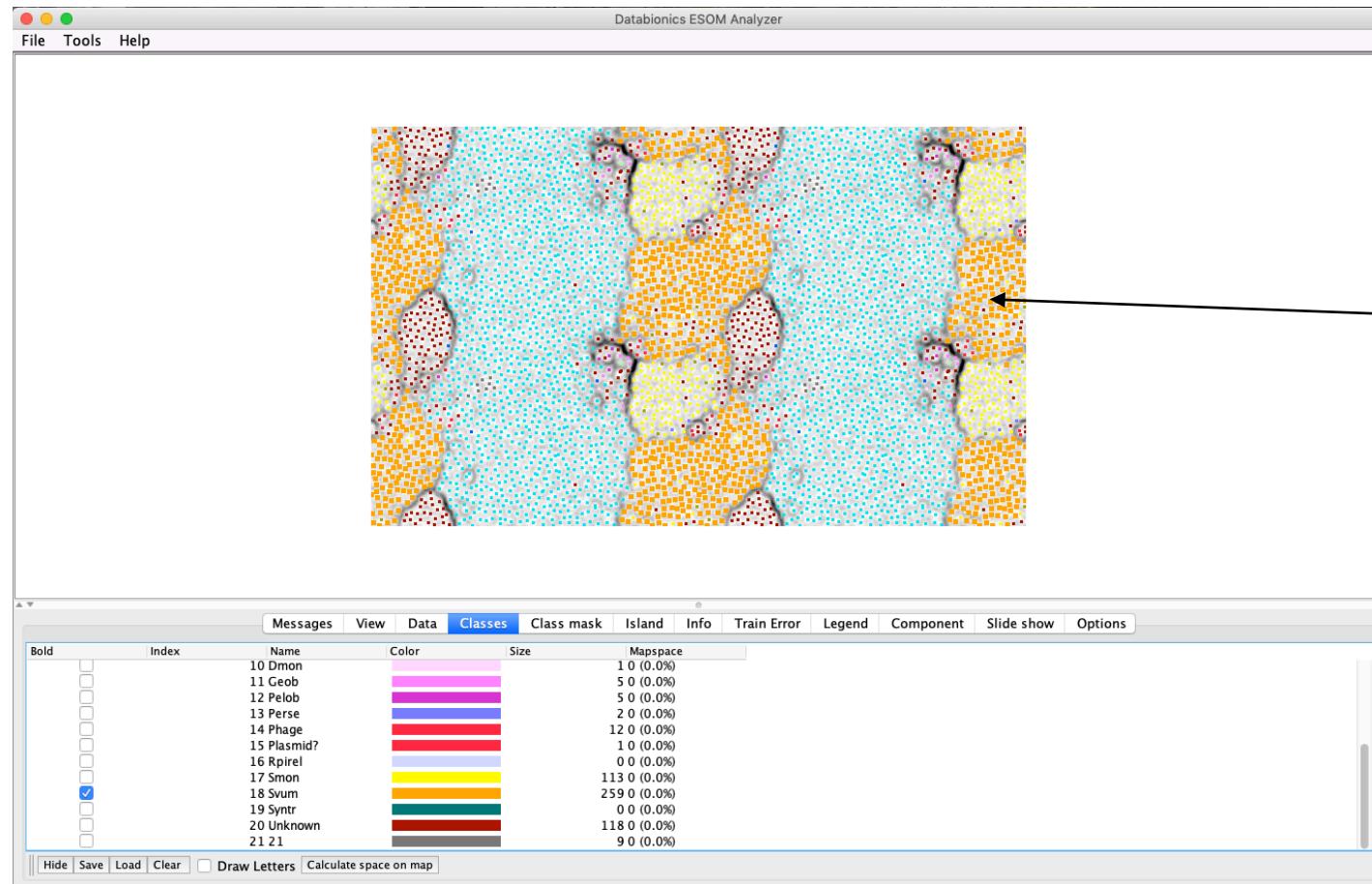


Like VizBin, bins are selected by manually drawing around boundary



# Alternative: ESOM

Map with contig fragments shown and coloured by bin assignment



- Select pre-assigned bin to highlight contig fragments
- Choose/change bin colours
- Example: *Sulfurovum* bin highlighted



# Task: Work with VizBin

---

**Use VizBin to:**

- Prepare input files for VizBin
- Project high-dimensional data down into a 2D plot
- Pick refined bins



# Viruses: Identification

---

[Go to Github MSS webpage](#)

- Submit the VIBRANT slurm job in ex10\_viruses to run overnight

