Metagenomics:

Contig Binning, Taxonomic Assignment & Validation

STAMPS – Day 7

July 25, 2024

Michael Nute & Todd Treangen

mike.nute@gmail.com

mn56@rice.edu

Approximate Agenda

9:00am to 9:10am:	Intro/kickoff (Todd)
9:10pm to 9:35am:	Binning lecture (Mike)
9:35am to 9:55am:	Binning tutorial (Todd & Mike)
9:55am to 10:10pm:	We have genomes/genome bins, now what? (Todd)
10:10am to 10:35am:	Break (Group Photo at Lillie)
10:35am to 10:50am:	Phylogenetics + MSA lecture (Mike)
10:50am to 11:00am:	MSA game (Todd)
11:00am to 11:35am:	Parsnp/strain analysis lecture (Mike)
11:35am to 11:55am:	Parsnp tutorial (Mike & Todd)
11:55am to Noon:	Emu advertisement (Mike)

Metagenomic Contig Binning

You have completed a de novo metagenomic assembly…

…we’re sorry, but the insights are in another castle.

Review: Metagenome Assembly

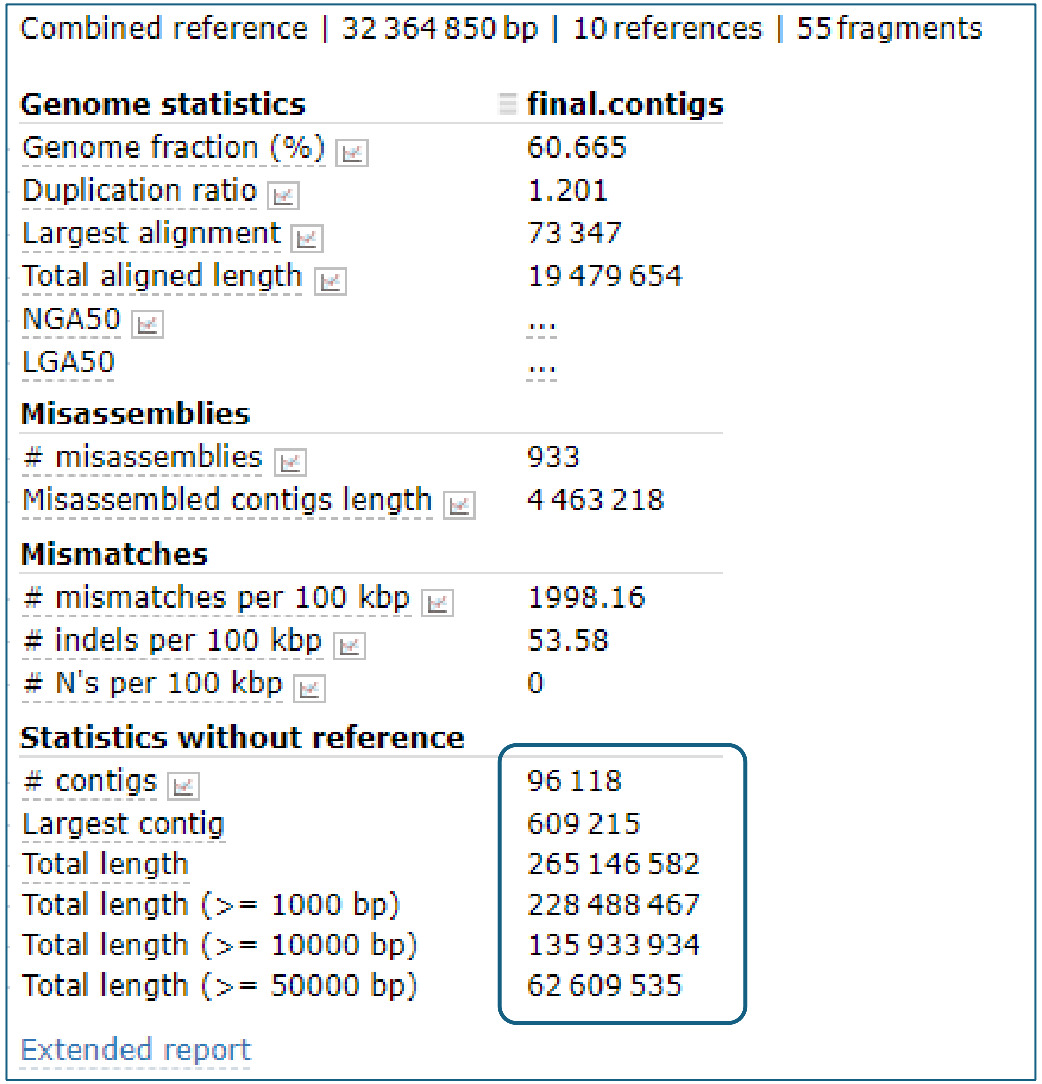
- Metagenome assembly produces contigs.
- Example (right):
  - SRA Run ID: SRR27117388
  - Human Stool Sample
  - 49,897,298 paired-end reads
  - 300bp per read (NovaSeq 6000)

Decent depth & length

- Assembly Results (MEGAHIT):
  - 179,415 contigs
  - (96k at right → min-length=500)
  - For contigs>500bp: Mean-length=2,758

A LOT of contigs!

Now What?

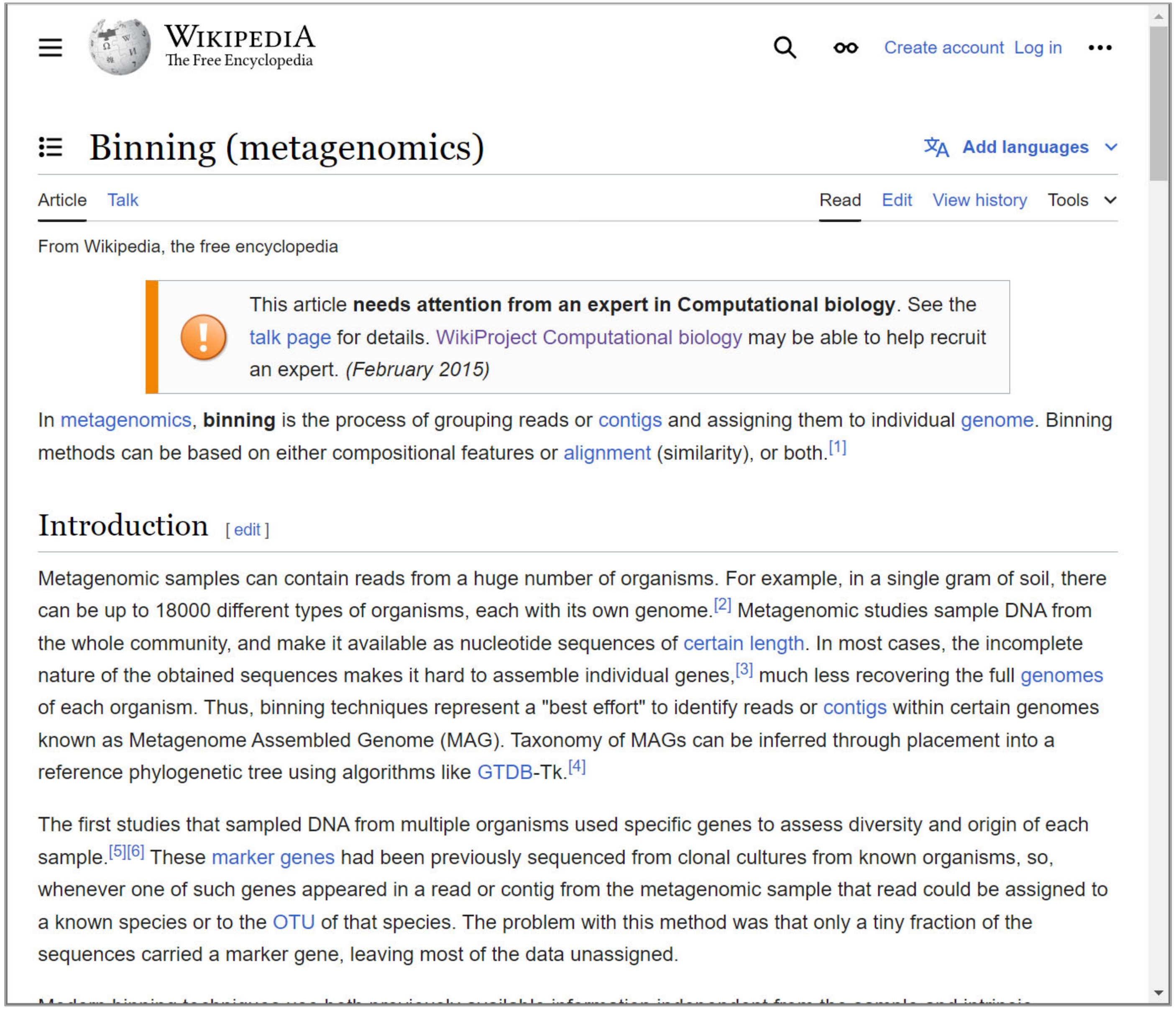


Combined reference | 32 364 850 bp | 10 references | 55 fragments

Genome statistics	**final.contigs**
Genome fraction (%)	60.665
Duplication ratio	1.201
Largest alignment	73 347
Total aligned length	19 479 654
NGA50	...
LGA50	...
Misassemblies	
# misassemblies	933
Misassembled contigs length	4 463 218
Mismatches	
# mismatches per 100 kbp	1998.16
# indels per 100 kbp	53.58
# N's per 100 kbp	0
Statistics without reference	
# contigs	96 118
Largest contig	609 215
Total length	265 146 582
Total length (>= 1000 bp)	228 488 467
Total length (>= 10000 bp)	135 933 934
Total length (>= 50000 bp)	62 609 535

Extended report

Next Step: Contig Binning



Binning (metagenomics)

From Wikipedia, the free encyclopedia

This article **needs attention from an expert in Computational biology**. See the talk page for details. WikiProject Computational biology may be able to help recruit an expert. *(February 2015)*

In metagenomics, **binning** is the process of grouping reads or contigs and assigning them to individual genome. Binning methods can be based on either compositional features or alignment (similarity), or both.[1]

Introduction [edit]

Metagenomic samples can contain reads from a huge number of organisms. For example, in a single gram of soil, there can be up to 18000 different types of organisms, each with its own genome.[2] Metagenomic studies sample DNA from the whole community, and make it available as nucleotide sequences of certain length. In most cases, the incomplete nature of the obtained sequences makes it hard to assemble individual genes,[3] much less recovering the full genomes of each organism. Thus, binning techniques represent a "best effort" to identify reads or contigs within certain genomes known as Metagenome Assembled Genome (MAG). Taxonomy of MAGs can be inferred through placement into a reference phylogenetic tree using algorithms like GTDB-Tk.[4]

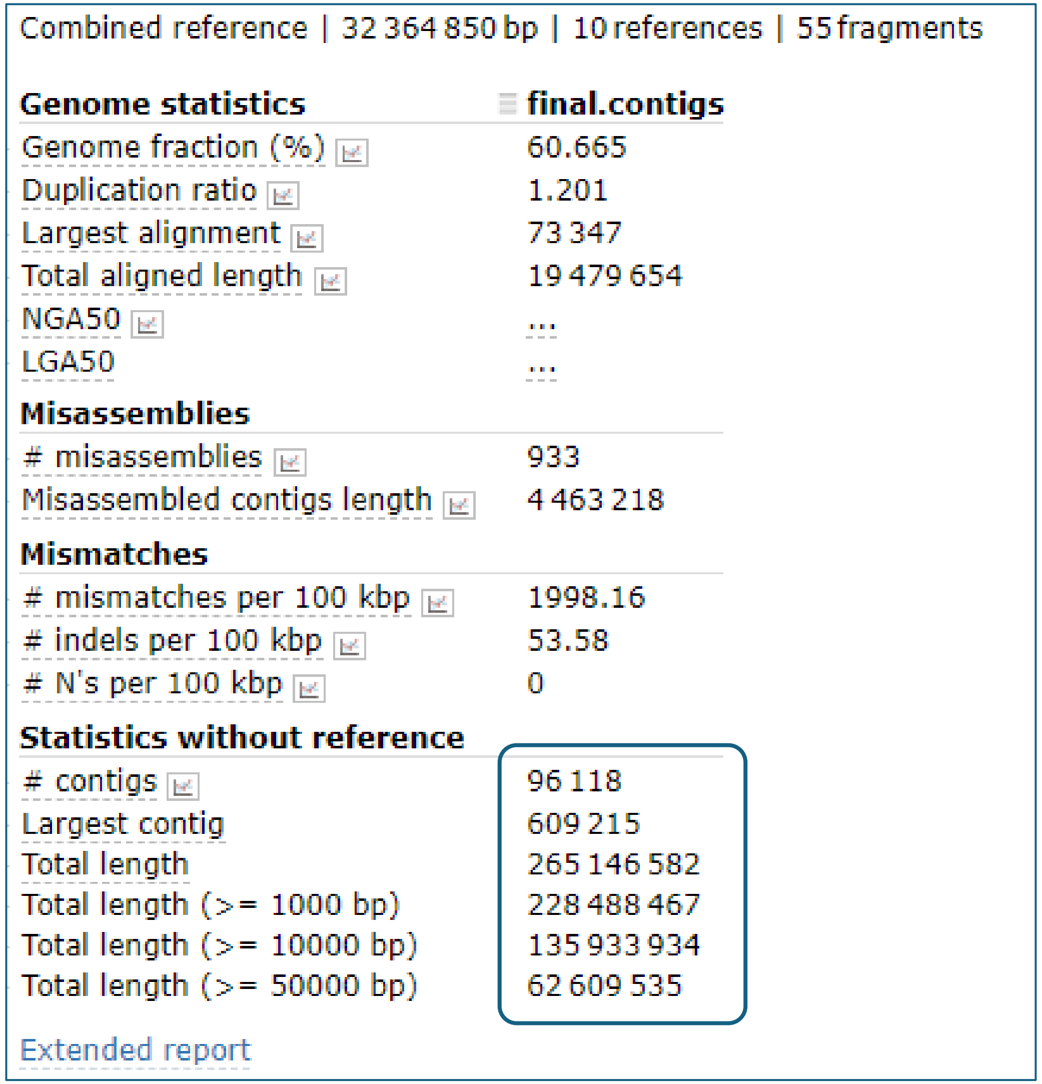
The first studies that sampled DNA from multiple organisms used specific genes to assess diversity and origin of each sample.[5][6] These marker genes had been previously sequenced from clonal cultures from known organisms, so, whenever one of such genes appeared in a read or contig from the metagenomic sample that read could be assigned to a known species or to the OTU of that species. The problem with this method was that only a tiny fraction of the sequences carried a marker gene, leaving most of the data unassigned.

Modern binning techniques use both previously available information independent from the sample and intrinsic

...frankly a decent description of it.

Things we want to know…

- Which of these contigs are part of the same genome?
- What organism does each of those genomes belong to?
- (other things)



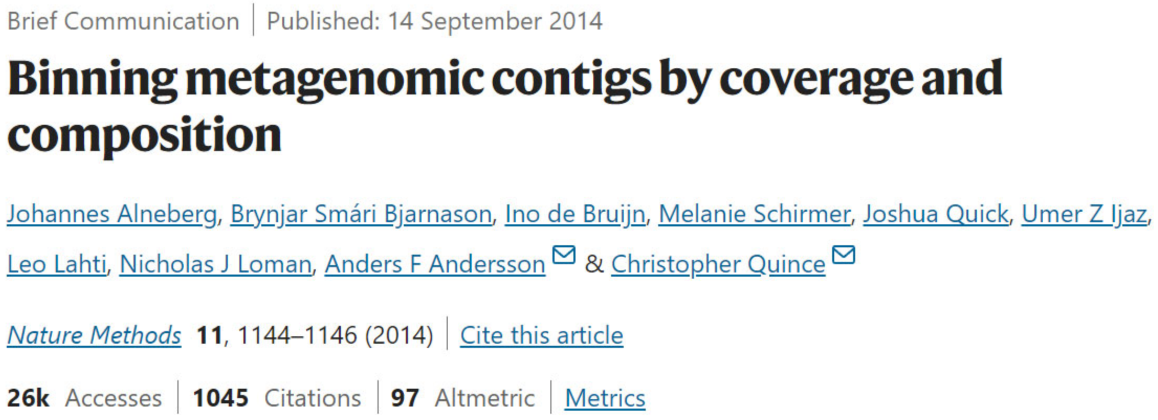
Combined reference | 32 364 850 bp | 10 references | 55 fragments

Genome statistics	final.contigs
Genome fraction (%)	60.665
Duplication ratio	1.201
Largest alignment	73 347
Total aligned length	19 479 654
NGA50	...
LGA50	...
Misassemblies	
# misassemblies	933
Misassembled contigs length	4 463 218
Mismatches	
# mismatches per 100 kbp	1998.16
# indels per 100 kbp	53.58
# N's per 100 kbp	0
Statistics without reference	
# contigs	96 118
Largest contig	609 215
Total length	265 146 582
Total length (>= 1000 bp)	228 488 467
Total length (>= 10000 bp)	135 933 934
Total length (>= 50000 bp)	62 609 535

Extended report

Which contigs go together

- How do we know?
- Clues:
  - Read coverage
  - K-mer composition
  - Alignment to reference genomes
  - Paired-end read linkage
  - Assembly graph properties
  - *…among others*



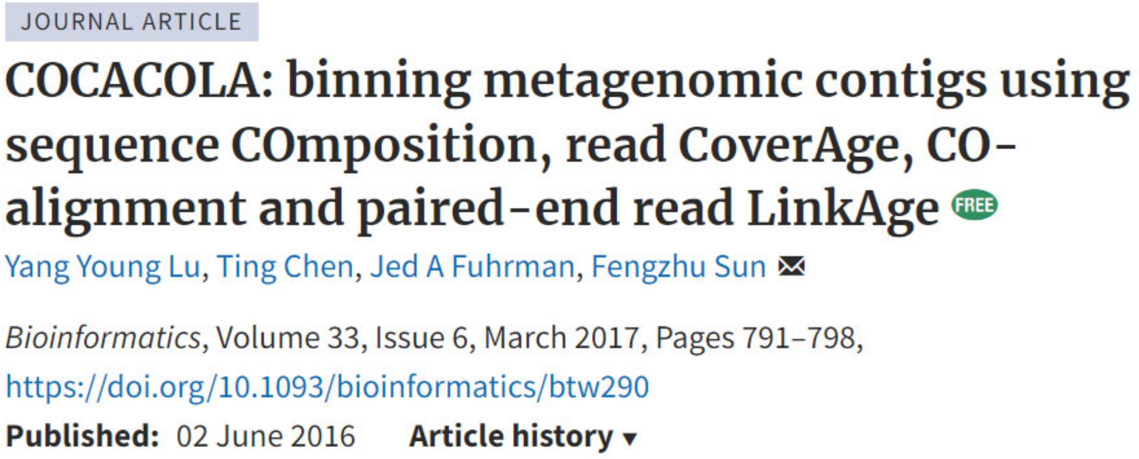
Brief Communication | Published: 14 September 2014

Binning metagenomic contigs by coverage and composition

Johannes Alneberg, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson & Christopher Quince

Nature Methods **11**, 1144–1146 (2014) | Cite this article

26k Accesses | **1045** Citations | **97** Altmetric | Metrics



JOURNAL ARTICLE

COCACOLA: binning metagenomic contigs using sequence COmposition, read CoverAge, CO-alignment and paired-end read LinkAge FREE

Yang Young Lu, Ting Chen, Jed A Fuhrman, Fengzhu Sun

Bioinformatics, Volume 33, Issue 6, March 2017, Pages 791–798, https://doi.org/10.1093/bioinformatics/btw290

Published: 02 June 2016 **Article history** ▾

Binning Clues: Coverage & Composition

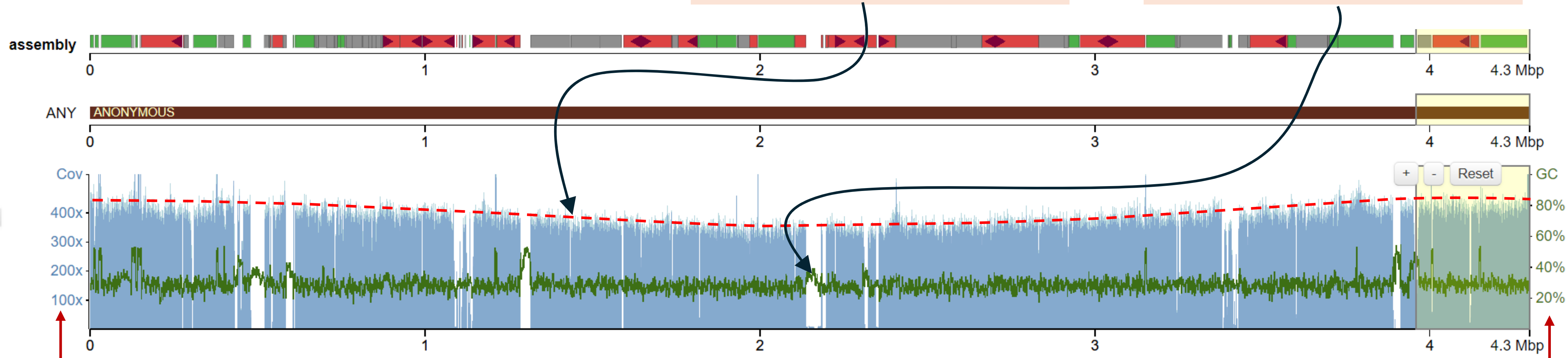
- *Example (below):*
  - *C. Difficile* clinical isolate
  - WGS de-novo assembly.
  - Comparison to CD630 reference

Read coverage is approximately constant over the length of the genome…

- Except for ±5% variation in sinusoidal pattern (WHY?)

GC % in reference is consistent EXCEPT for regions where clinical coverage is zero.

- Possible contamination or misassembly in reference?



Blue = Read Coverage

Green = % GC

Binning Clues: Reference Co-Alignment & Pair-End Linkage

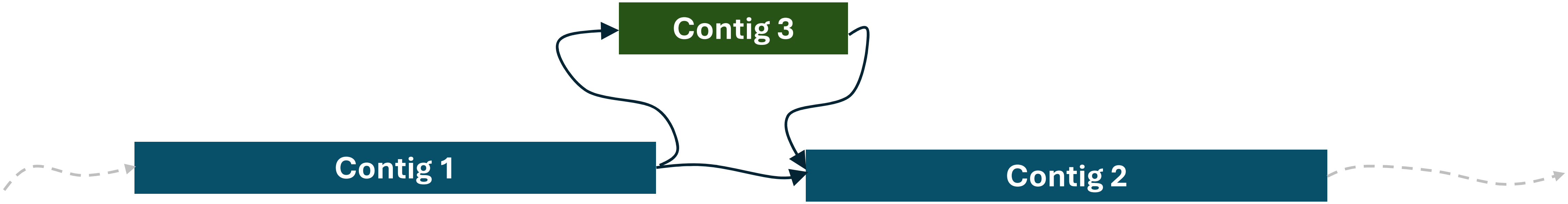
- **Reference co-alignment:**
  - Contigs align to the same reference genome
  - Pretty good sign, especially if they don’t overlap
  - Note:
    - *This means that really, any taxonomic classifier can be a “contig binning” algorithm if you use it that way, but that doesn’t mean it’s a very good one…*

- **Paired-End Linkage:**
  - Read pair-mates map to two different contigs
  - More = more likely the same genome

Should be self-explanatory why these would indicate contigs come from same organism…

Binning Clues: Assembly Graph

Example: Bubble



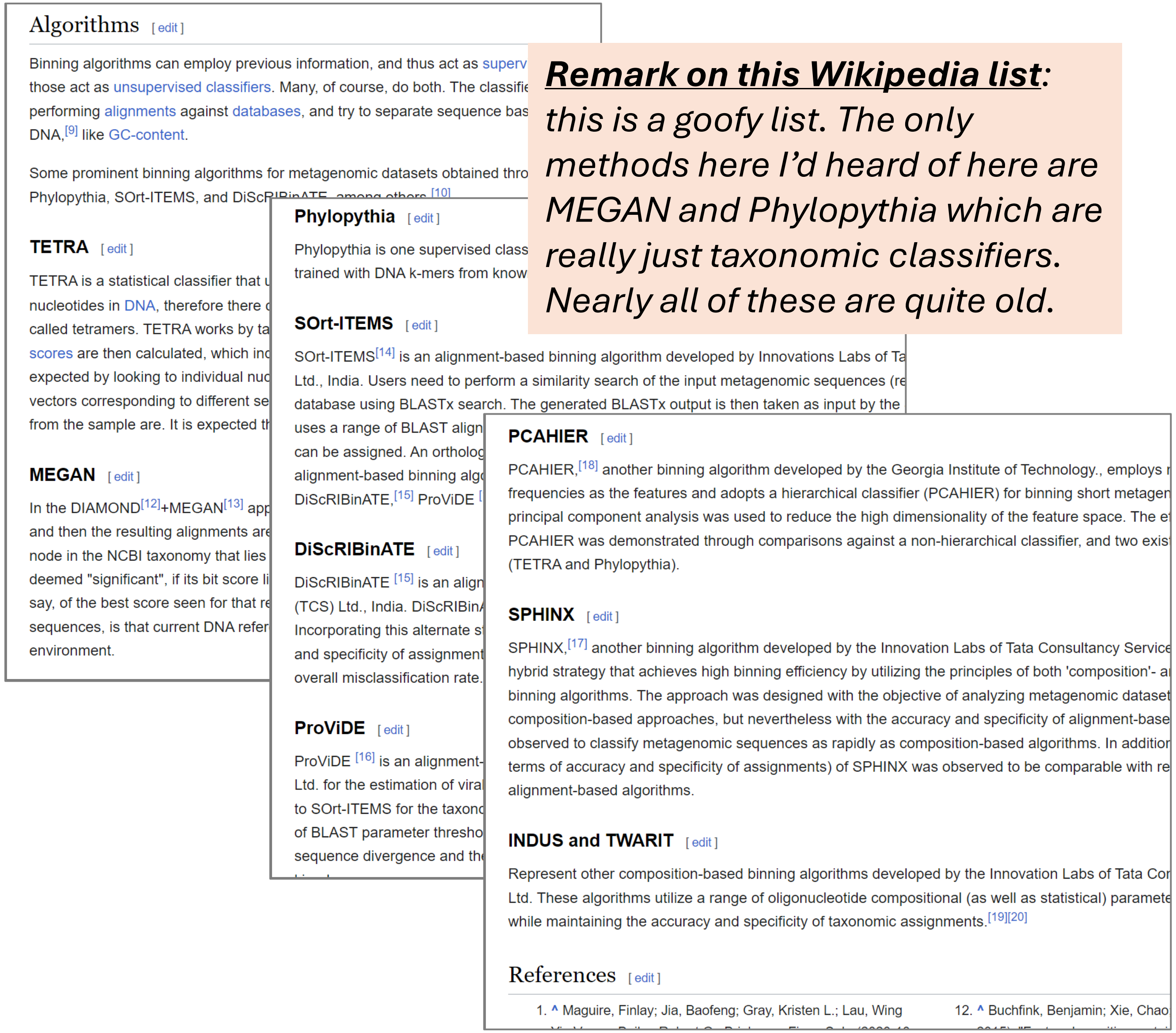
What does this tell us?

Should Contigs 1, 2 & 3 be in the same bin?
Two different bins?

Contig Binning & Validation

Brass Tacks

Contig Binning Software List (Partial)



Algorithms [edit]

Binning algorithms can employ previous information, and thus act as superv
those act as unsupervised classifiers. Many, of course, do both. The classifie
performing alignments against databases, and try to separate sequence bas
DNA,[9] like GC-content.

Some prominent binning algorithms for metagenomic datasets obtained thro
Phylopythia, SOrt-ITEMS, and DiScRIBinATE, among others.[10]

TETRA [edit]

TETRA is a statistical classifier that u
nucleotides in DNA, therefore there c
called tetramers. TETRA works by ta
scores are then calculated, which inc
expected by looking to individual nuc
vectors corresponding to different se
from the sample are. It is expected th

MEGAN [edit]

In the DIAMOND[12]+MEGAN[13] app
and then the resulting alignments are
node in the NCBI taxonomy that lies
deemed "significant", if its bit score li
say, of the best score seen for that re
sequences, is that current DNA refer
environment.

Phylopythia [edit]

Phylopythia is one supervised class
trained with DNA k-mers from know

SOrt-ITEMS [edit]

SOrt-ITEMS[14] is an alignment-based binning algorithm developed by Innovations Labs of Ta
Ltd., India. Users need to perform a similarity search of the input metagenomic sequences (re
database using BLASTx search. The generated BLASTx output is then taken as input by the
uses a range of BLAST align
can be assigned. An ortholog
alignment-based binning alg
DiScRIBinATE,[15] ProViDE [

DiScRIBinATE [edit]

DiScRIBinATE [15] is an align
(TCS) Ltd., India. DiScRIBinA
Incorporating this alternate s
and specificity of assignment
overall misclassification rate.

ProViDE [edit]

ProViDE [16] is an alignment-
Ltd. for the estimation of viral
to SOrt-ITEMS for the taxono
of BLAST parameter thresho
sequence divergence and th

PCAHIER [edit]

PCAHIER,[18] another binning algorithm developed by the Georgia Institute of Technology., employs r
frequencies as the features and adopts a hierarchical classifier (PCAHIER) for binning short metagen
principal component analysis was used to reduce the high dimensionality of the feature space. The e
PCAHIER was demonstrated through comparisons against a non-hierarchical classifier, and two exis
(TETRA and Phylopythia).

SPHINX [edit]

SPHINX,[17] another binning algorithm developed by the Innovation Labs of Tata Consultancy Service
hybrid strategy that achieves high binning efficiency by utilizing the principles of both 'composition'- a
binning algorithms. The approach was designed with the objective of analyzing metagenomic dataset
composition-based approaches, but nevertheless with the accuracy and specificity of alignment-base
observed to classify metagenomic sequences as rapidly as composition-based algorithms. In additio
terms of accuracy and specificity of assignments) of SPHINX was observed to be comparable with re
alignment-based algorithms.

INDUS and TWARIT [edit]

Represent other composition-based binning algorithms developed by the Innovation Labs of Tata Co
Ltd. These algorithms utilize a range of oligonucleotide compositional (as well as statistical) paramete
while maintaining the accuracy and specificity of taxonomic assignments.[19][20]

References [edit]

1. ^ Maguire, Finlay; Jia, Baofeng; Gray, Kristen L.; Lau, Wing

12. ^ Buchfink, Benjamin; Xie, Chao

Remark on this Wikipedia list*: this is a goofy list. The only methods here I'd heard of here are MEGAN and Phylopythia which are really just taxonomic classifiers. Nearly all of these are quite old.*

More modern list:

- MetaBAT2 (2019)
- CONCOCT (2014)
- COCACOLA (2017)
- VAMB (2022)
- MyCC (2016)
- MaxBin 2.0 (2015)
- MetaWatt-3.5 (2012)
- MetaWRAP (2018)
- Autometa (2019)
- MetaBinner (2023)
- UltraBinner (2023)

...and probably a dozen more since 2020

Remarks on More Modern Binners

More modern list:

Very different core algorithm from original MetaBAT (2016), added assembly graph & other features

- MetaBAT2 (2019)

All largely based primarily on:
1) Coverage
2) Kmer composition

- CONCOCT (2014)
- MyCC (2016)
- MaxBin 2.0 (2015)
- MetaWatt-3.5 (2012)

Added co-alignment & paired-end read linkage

- COCACOLA (2017)
- MetaWRAP (2018)
- Autometa (2019)

Deep-Learning method that got a lot of attention

- VAMB (2022)

Heavy use of single-copy marker genes to refine bins

- MetaBinner (2023)

MetaBinner + CONCOCT + MetaBAT2

- UltraBinner (2023)

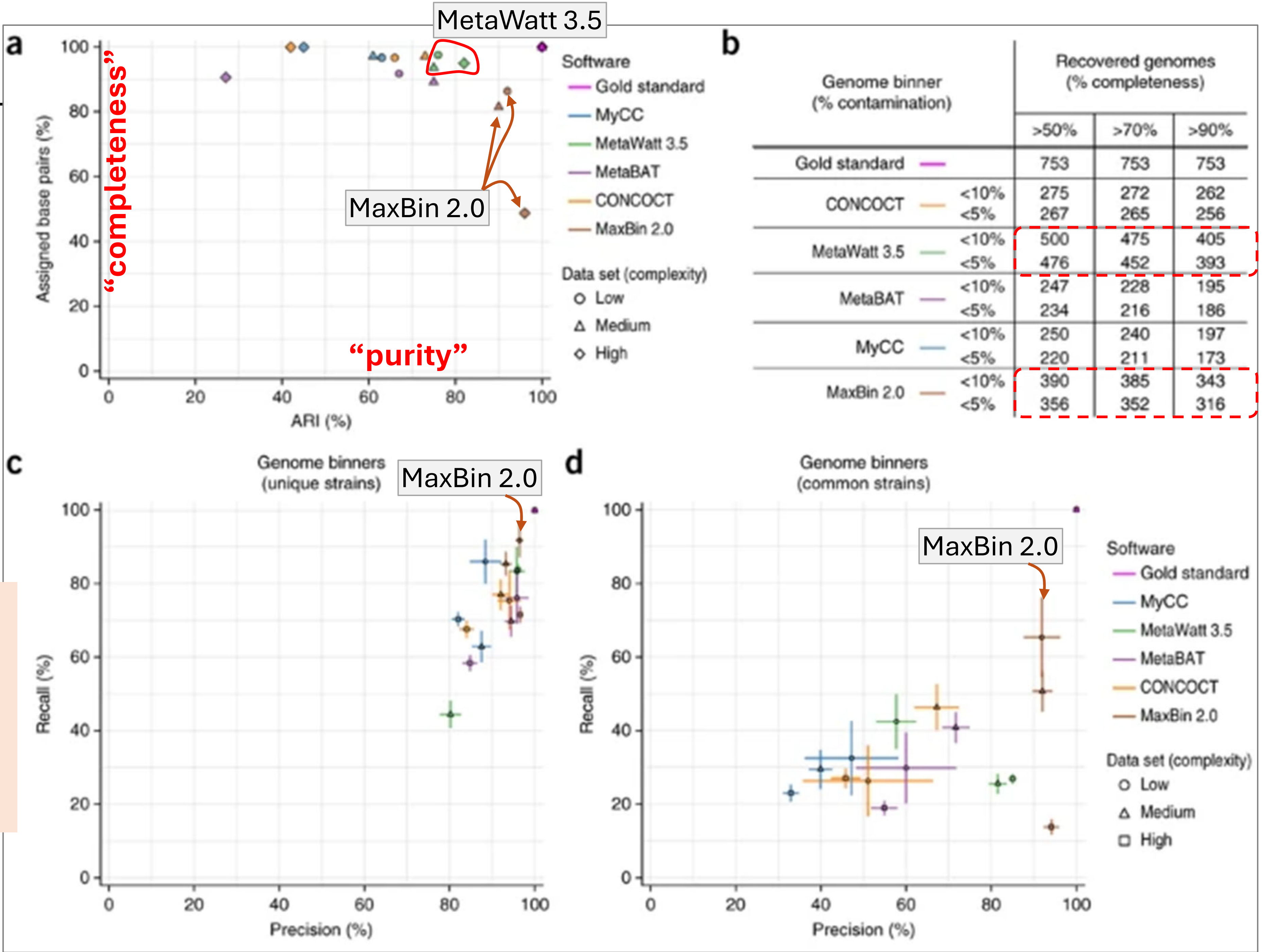
CAMI 1 (2017)

CAMI: **C**ritical **A**ssessment of **M**etagenome **I**ntepretation *(Sczyrba, et al., 2017)*

ARI = adjusted Rand Index. (% of contig pairs correctly co-assigned).

...hard not to conclude that MaxBin was the winner here, with honorable mention to MetaWatt for best sensitivity...

Fastest: MetaBAT



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

Image credit: Sczyrba, et al. Nat. Meth. (2017), Figure 2

CAMI 2 (2022)

CAMI-2

- 3 different datasets reflecting “modern” challenges
  - novel microbes (Plant)
  - strain diversity
  - high synteny (i.e. short contigs) (Marine)
- 2 Types of contigs
  - “gold-standard”
    - short+long read assembly
  - “megahit”
    - short reads only

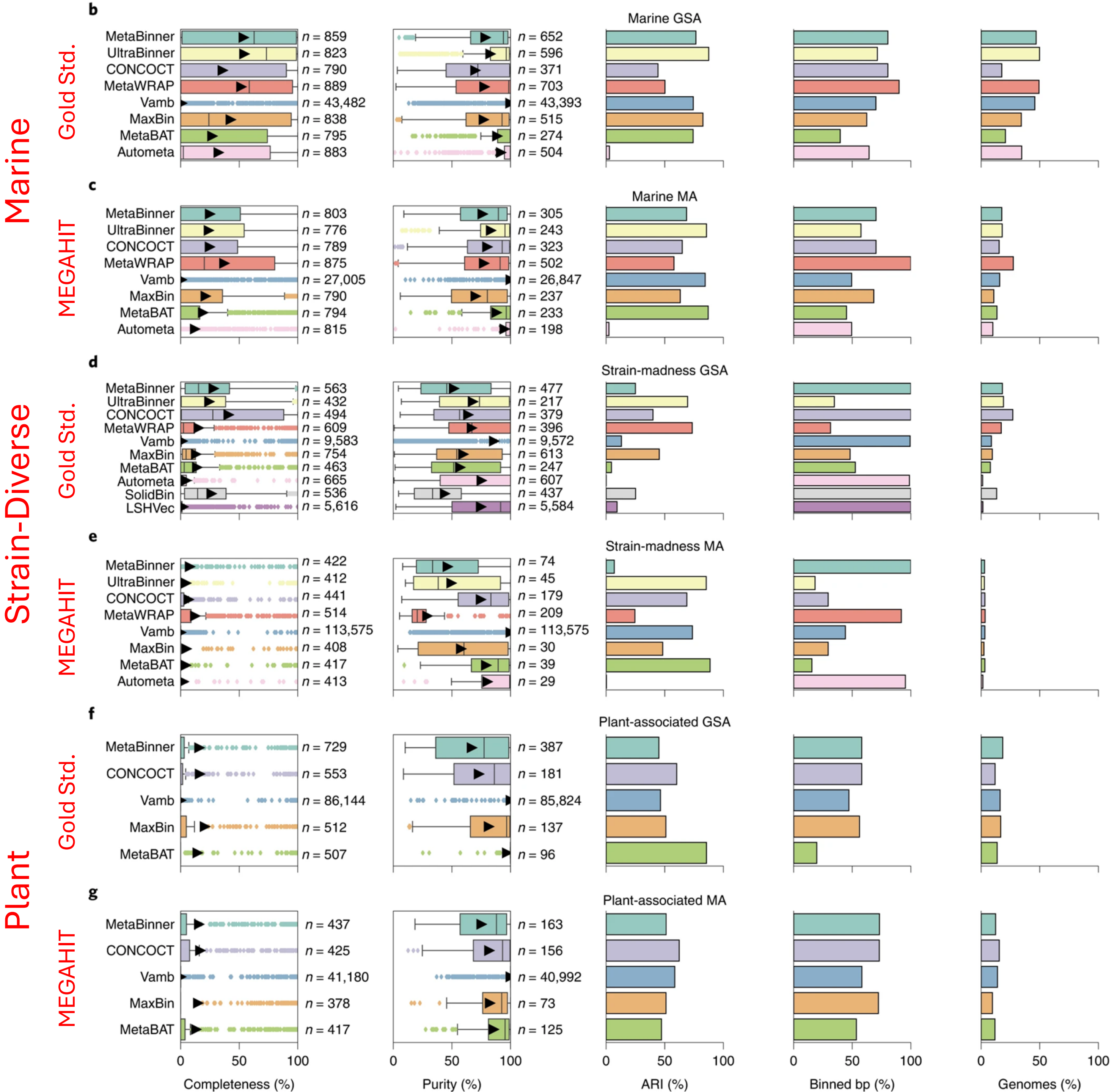


Image credit: Meyer, et al. Nat. Meth. (2022), Figure 2

CAMI 2 (2022)

- Far less clear which method is optimal overall
- MetaBinner/UltraBinner seem to do reliably well ("ensemble" methods)
  - CONCOCT also quite well.
- VAMB is an outlier:
  - Very high purity, Very low completeness
  - "Partial" binning
  - Recall the bubble example
- MetaBAT:
  - Does reasonably well, not best
  - By far lowest compute cost (time/RAM)
- Several methods unable to run on larger datasets

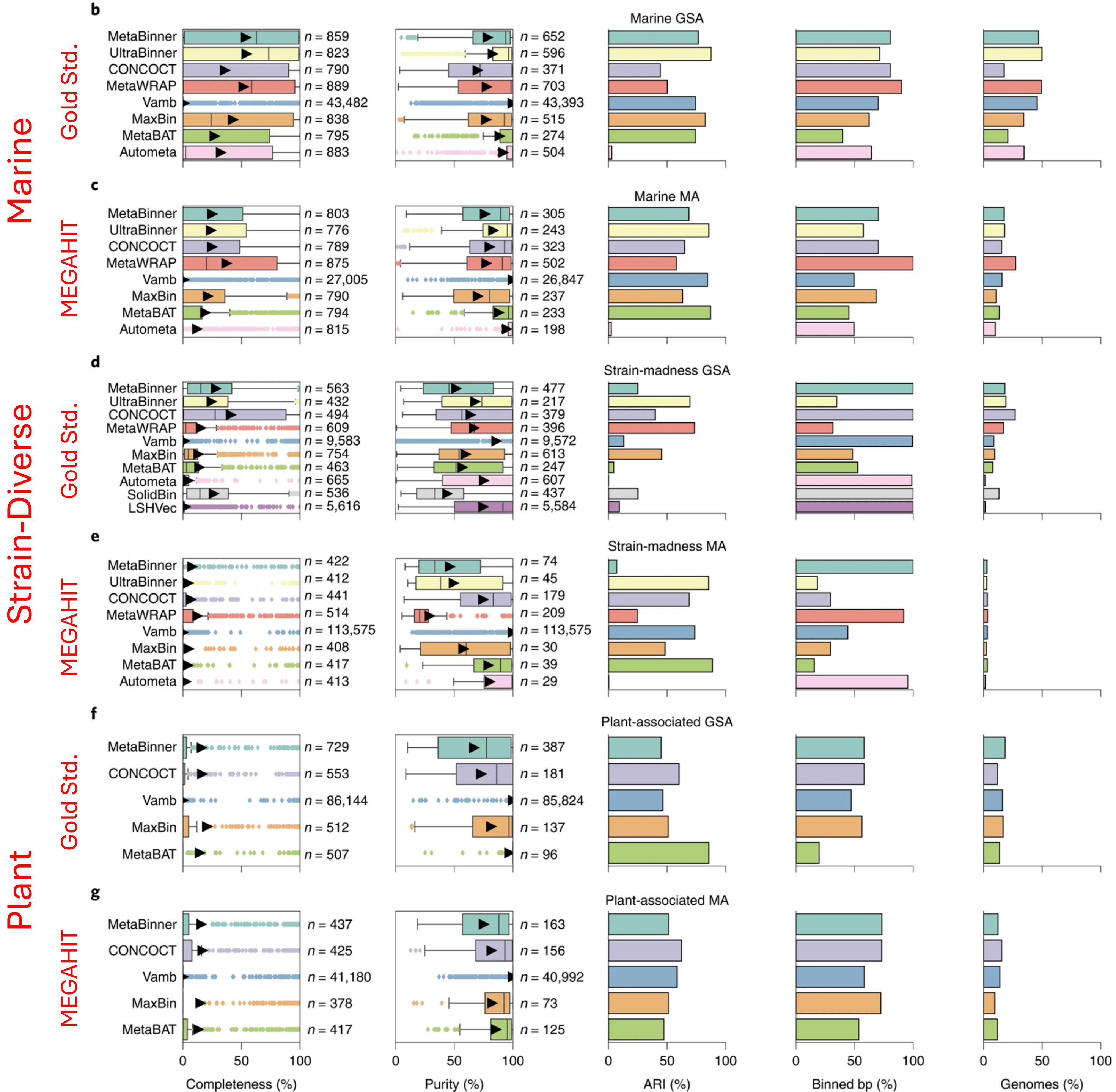


Image credit: Sczyrba, et al. Nat. Meth. (2017), Figure 2

CAMI 2 (2022): Runtime & Memory Usage

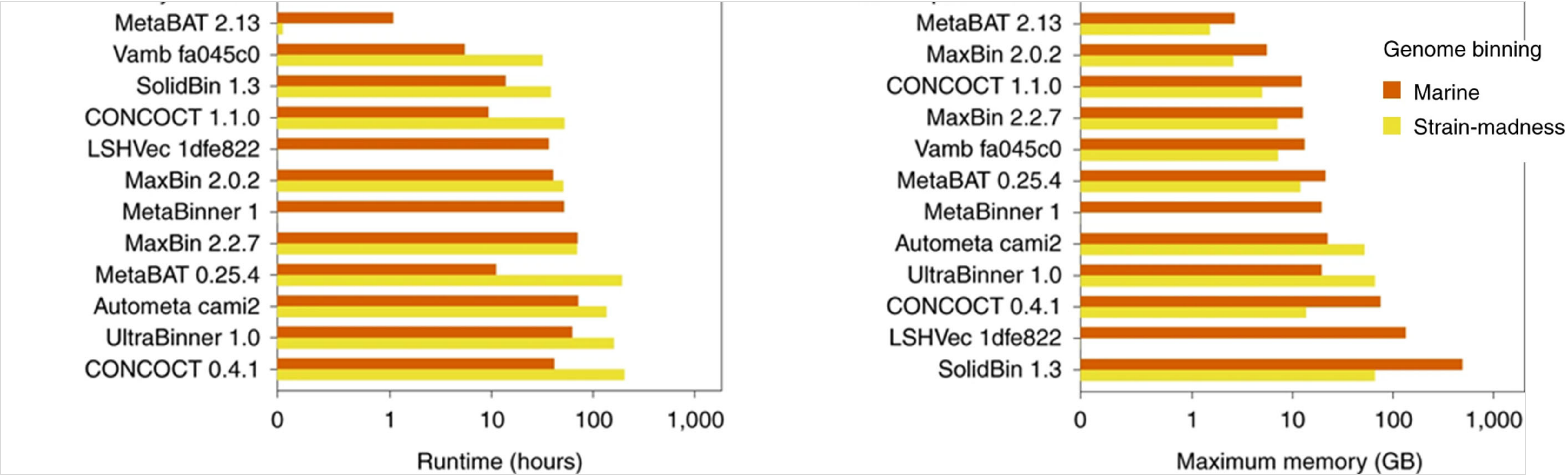
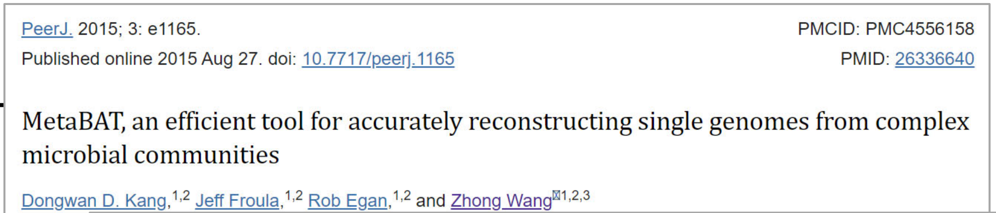


Image credit: Sczyrba, et al. Nat. Meth. (2017), Figure 5

Contig Binning Software: MetaBAT2



PeerJ. 2015; 3: e1165. PMCID: PMC4556158

Published online 2015 Aug 27. doi: 10.7717/peerj.1165 PMID: 26336640

MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities

Dongwan D. Kang,[1,2] Jeff Froula,[1,2] Rob Egan,[1,2] and Zhong Wang✉[1,2,3]

PeerJ. 2019; 7: e7359. PMCID: PMC6662567

Published online 2019 Jul 26. doi: 10.7717/peerj.7359 PMID: 31388474

MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies

Dongwan D. Kang,[1] Feng Li,[2] Edward Kirton,[1] Ashleigh Thomas,[1] Rob Egan,[1] Hong An,[2] and Zhong Wang✉[1,3,4]

- Inputs:
  - Contig file
  - Coverage file
- Parameters (default value):

Argument	Definition	Default Value	Notes
`-m`	Minimum size of a contig for binning (must be ≥1500)	2500	CAMI-2 ran both 1500, 2500 and found little difference
`--maxP`	% of 'good' contigs considered for binning	95%	Greater ⟹ more sensitivity
`--minS`	Minimum score of a edge for binning (from 1 to 99)	60	Greater ⟹ more specificity
`--maxEdges`	Maximum number of edges per node	200	Greater ⟹ more sensitivity
`--pTNF`	4-mer probability cutoff for building 4-mer graph	0	Used to skip a preparation step, for speed.
`-x --minCV`	Minimum mean coverage of a contig in each library for binning.	1	Could be adjusted based on expectations/needs vis-à-vis depth
`-s --minClsSize`	Minimum size of a bin as the output	200kbp	

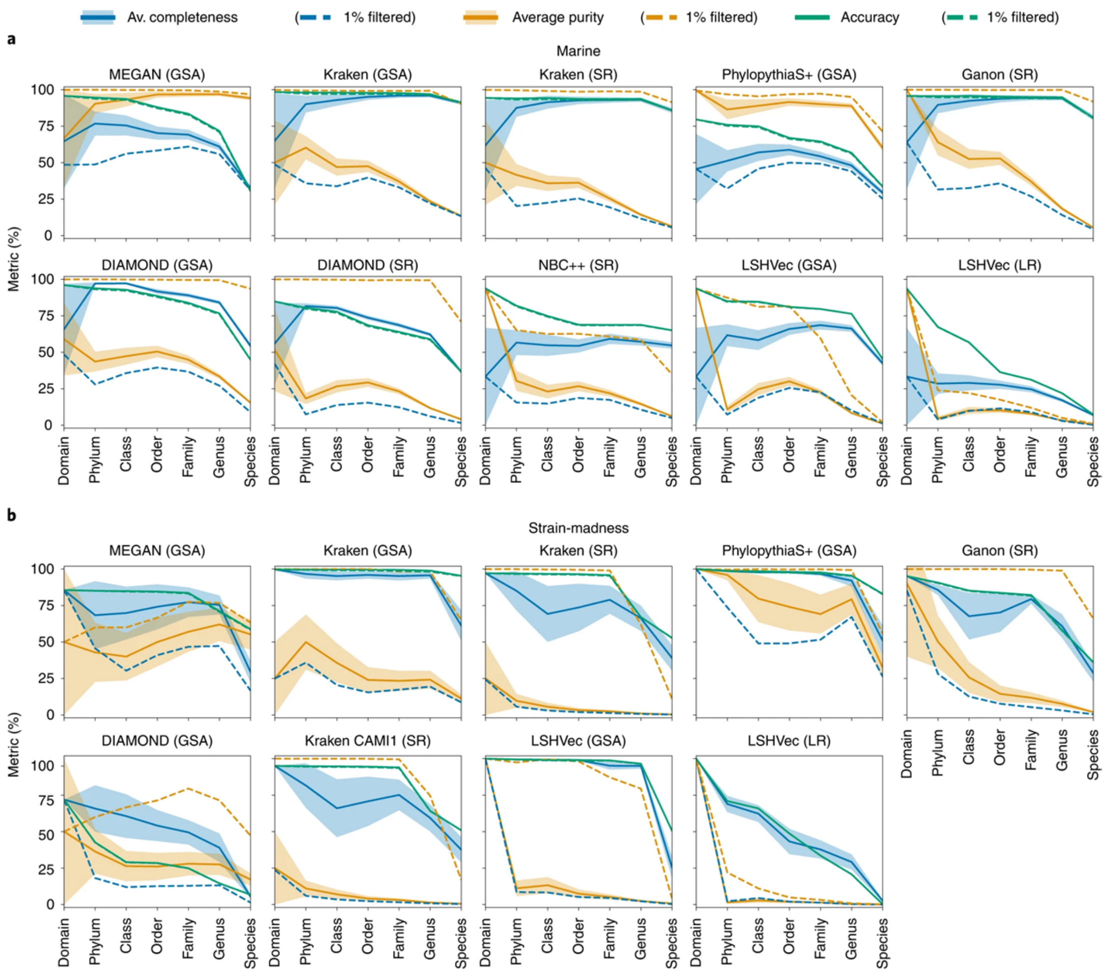
Contig Bin Validation: CheckM2

- Original Idea (CheckM, i.e. version 1):
  - Use *lineage-specific, single-copy marker genes* (à la MetaPhlAn) to estimate how complete the bin is.
  - Issue: only well-studied lineages have good database of markers
- New Idea (CheckM2):
  - Use a fancier machine learning algorithm to predict bin quality
- Output:

Bin #	Completeness (%)	Contamination (%)	Coding Density	Contig N50	Average Gene Length	Genome Size	GC Content	Total Coding Sequences	Total Contigs	Max Contig Length
1	45.5	0.0	0.885	609,215	336.2	896,313	53%	788	3	609,215
2	10.7	0.0	0.859	34,272	324.0	273,576	60%	242	13	56,825
3	75.6	21.3	0.878	13,874	296.7	2,692,053	58%	2,660	247	100,008
4	73.0	0.0	0.907	123,875	351.4	2,083,794	38%	1,796	31	196,080
5	92.2	3.6	0.886	182,793	363.5	2,626,142	60%	2,136	32	355,138
6	24.2	0.0	0.865	4,241	272.0	475,717	64%	506	113	14,380
7	10.5	0.0	0.903	8,236	302.3	247,455	61%	247	34	17,322
8	58.6	0.8	0.907	4,656	268.2	1,372,490	64%	1,550	302	13,961
9	88.0	5.5	0.883	28,581	311.1	3,020,618	58%	2,861	146	107,535
10	38.7	0.2	0.889	81,426	327.6	885,202	63%	802	16	111,004
11	61.0	0.0	0.899	162,638	331.1	1,125,146	54%	1,020	8	265,330

Taxonomic Assignment

- Note that neither MetaBAT nor CheckM2 assign taxonomy to each of our bins.
- Taxon assignment is pretty simple, many ways to do it:
  - BLAST
  - Sourmash
  - Kraken2
  - *…any of the other tools evaluated in CAMI-2:*
    - *MEGAN*
    - *DIAMOND*
    - *Ganon*
    - *PhyloPythiaS+*
    - *NBC++*



Nota Bene: *This step has not been included in the tutorial because most of these programs require a large database to be downloaded or built in advance. It should be simple to do this using Sourmash, however, based on that material from earlier in the course…*

Image credit: Sczyrba, et al. Nat. Meth. (2017), Figure 3

Any Questions?

Things to think about:

- What is contig binning and what is the goal?
- What information can we use to bin contigs together?
- How many contig binning methods are there?
- What is the best contig binner?
- Is VAMB any good?
- What does CheckM2 produce?
- Why didn’t I add taxon assignment to the tutorial?

Binning & Validation Tutorial

- Tasks:
  1. Find the outputs from the QC/Assembly tutorial:
     a) MEGAHIT Contigs
     b) Coverage information (.bam file)
     c) MetaBAT bins
  2. Run CheckM2 on the bins from (c)
  3. Run MetaBAT2 on the contigs/coverage, see if you get the same # of bins
  4. Run CheckM2 on the new bins & compare to (4)
  5. Run CheckM2 on some bins from a Human stool sample (provided)
- Left as an exercise:
  6. Assign taxonomy to one of these bins using Sourmash
- Other bash-scripting concepts:
  - Setting & using bash variables
  - Absolute vs. Relative paths

https://github.com/MGNute/stamps_2024_assembly_tutorial/blob/main/binning_validation.md

Thank You

Please don’t hesitate to follow up with me after STAMPS if I can help or go through any of this material again.