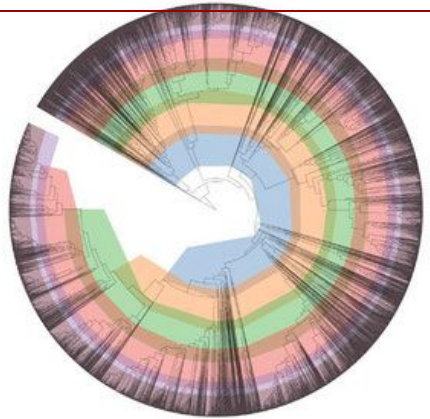
ENVIRONMENTAL METAGENOMICS

Physalia course, online, 11-15 November 2024

MAG QC & Taxonomic annotation

Nikolay Oskolkov, Lund University, NBIS SciLifeLab
Luis Pedro Coelho, Queensland University of Technology



Physalia Courses

NB: original course material courtesy:
Dr. Antti Karkman, University of Helsinki
Dr. Igor Pessi, Finnish Environment Institute (SYKE)

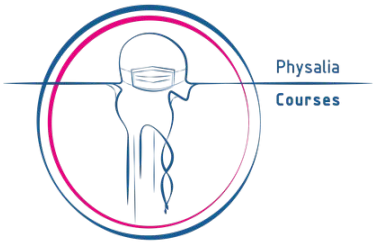
You got MAGs! Now what?

First step: QC

- How good are the bins you got?

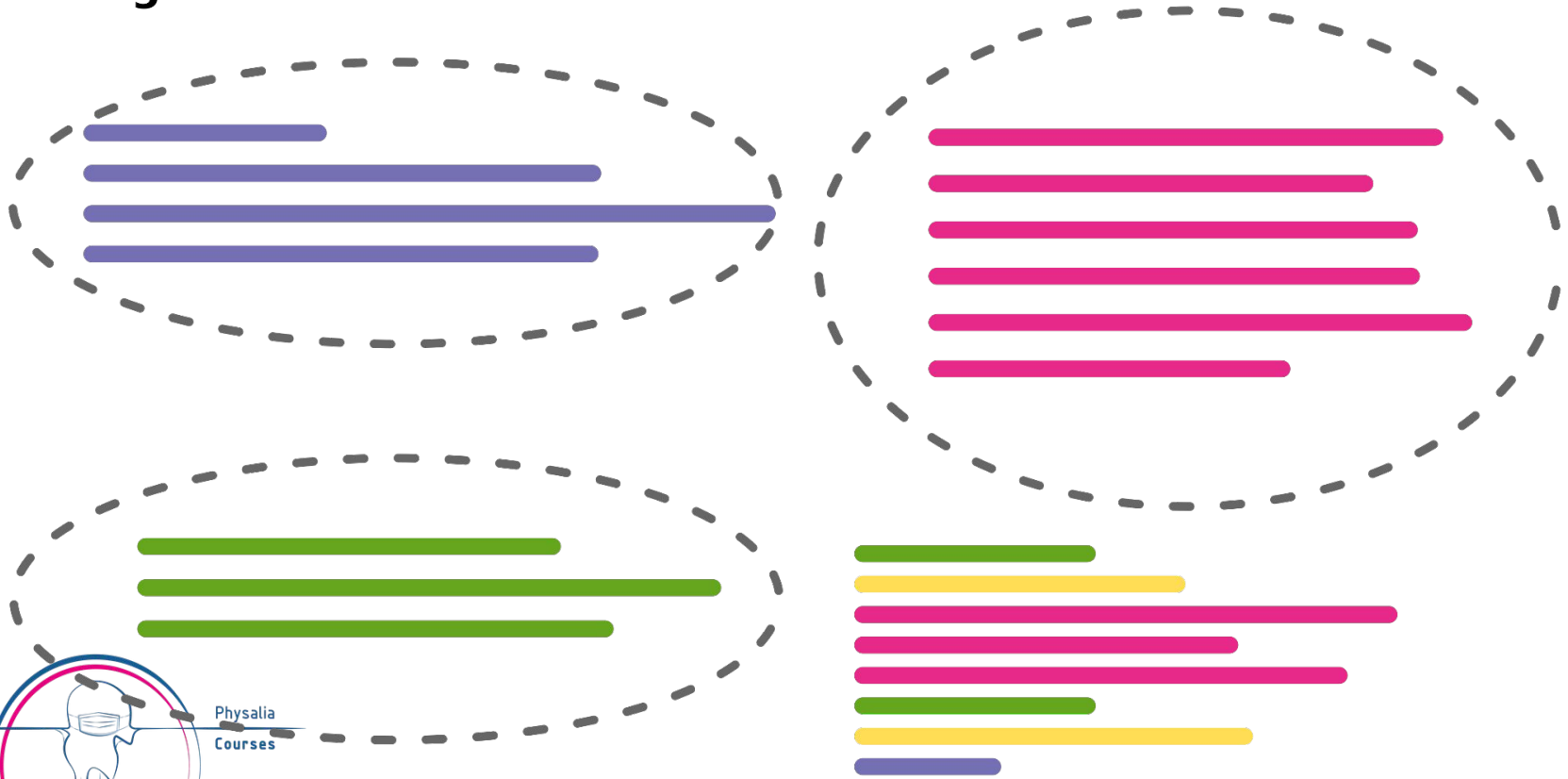
Other steps

- Annotation
- *Dereplication*
- *Abundance estimations*
- *Comparison with existing data*

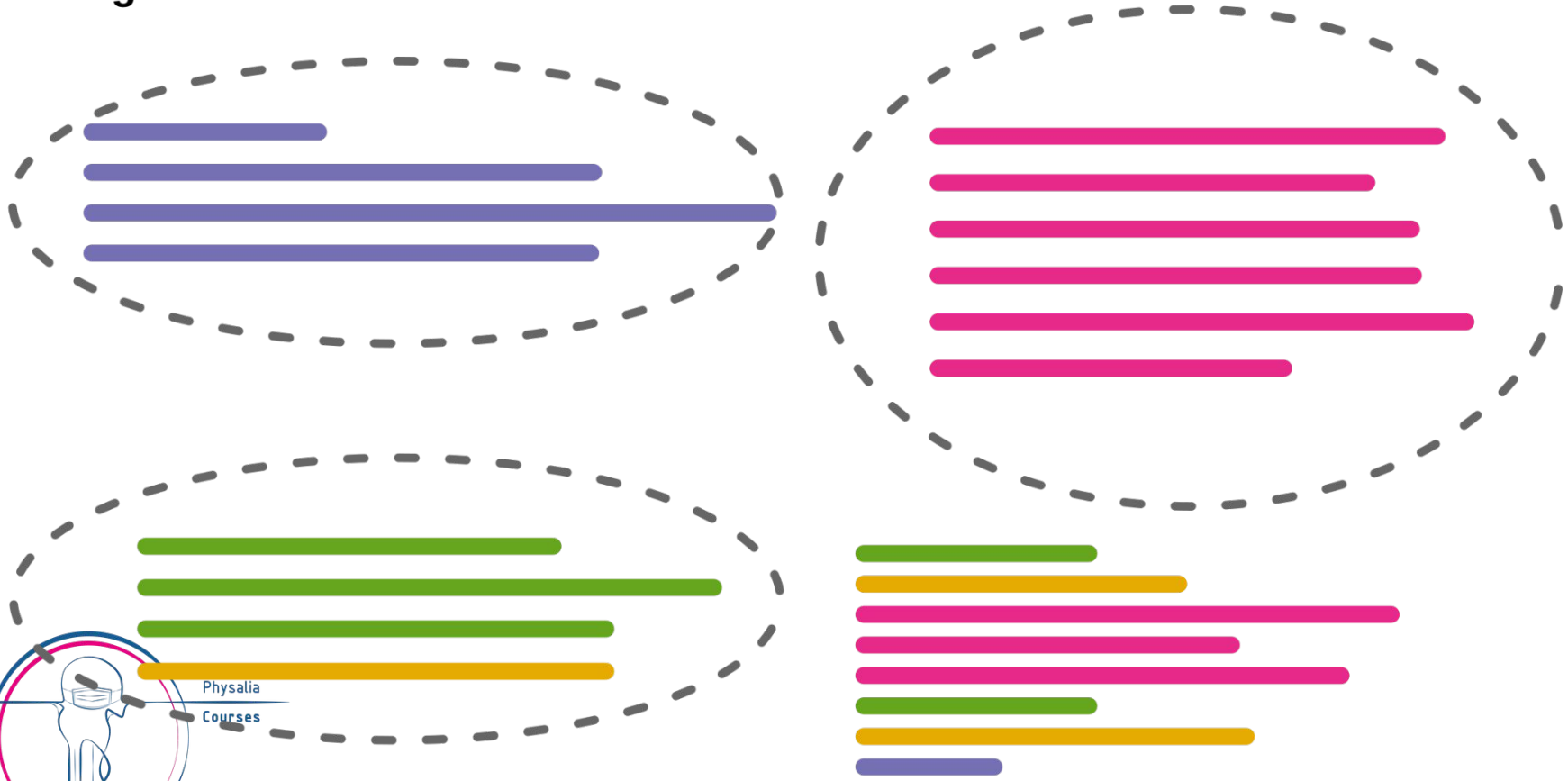


Congrats!
You got big FASTA files!

Binning



Binning



Errors

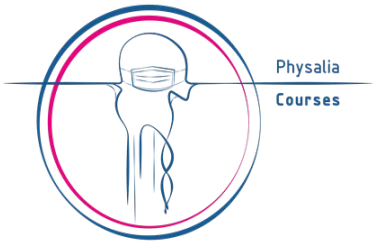
Contamination (false positive): a bin has contigs that do not belong there
In**completeness** (false negatives): a bin is missing contigs

What is a good enough genome?

- **High**: 90% complete, <5% contaminated
5S, 16S, & 23S rRNA genes present
18 different tRNA genes present

- **Medium**: 50% complete, <10% contaminated

- **Low:**50% complete, <10% contaminated



Single copy marker genes methods

Orthologous Group	Av. Length	Annotation	Genes in Prok.	Genes in Euk.	Total Genes
COG0012	380	Predicted GTPase, probable translation factor	171	30	201
COG0016	423	Phenylalanine-tRNA synthethase alpha subunit	168	42	210
COG0018†	548	Arginyl-tRNA synthetase	175	45	220
COG0048	137	Ribosomal protein S12	168	48	216
COG0049	182	Ribosomal protein S7	169	41	210
COG0052	240	Ribosomal protein S2	168	79	247
COG0060*	956	Isoleucyl-tRNA synthetase	172	42	214
COG0080	154	Ribosomal protein L11	170	61	231
COG0081	230	Ribosomal protein L1	168	61	229
COG0085†	1138	DNA-directed RNA polymerase, beta subunit	178	60	238
COG0087	288	Ribosomal protein L3	168	54	222
COG0091	157	Ribosomal protein L22	168	75	243
COG0092	240	Ribosomal protein S3	168	30	198
COG0093	130	Ribosomal protein L14	168	41	209
COG0094	182	Ribosomal protein L5	169	36	205
COG0096	131	Ribosomal protein S8	168	55	223
COG0097	177	Ribosomal protein L6P/L9E	168	65	233
COG0098	220	Ribosomal protein S5	168	110	278
COG0099‡	133	Ribosomal protein S13	168	49	217
COG0100	145	Ribosomal protein S11	169	51	220
COG0102	167	Ribosomal protein L13	168	54	222
COG0103	172	Ribosomal protein S9	168	52	220
COG0124*	472	Histidyl-tRNA synthetase	178	31	209
COG0143*†	646	Methionyl-tRNA synthetase	180	35	215
COG0172	442	Seryl-tRNA synthetase	177	37	214
COG0184	154	Ribosomal protein S15P/S13E	168	41	209
COG0186	122	Ribosomal protein S17	170	46	216
COG0197	175	Ribosomal protein L16/L10E	168	54	222
COG0200	166	Ribosomal protein L15	168	70	238
COG0201	445	Preprotein translocase subunit SecY	178	37	215
COG0202	323	DNA-directed RNA polymerase, alpha subunit	171	45	216
COG0256	178	Ribosomal protein L18	168	50	218
COG0495	854	Leucyl-tRNA synthetase	172	43	215
COG0522	199	Ribosomal protein S4 and related proteins	174	46	220
COG0525*‡	880	Valyl-tRNA synthetase	169	37	206
COG0533	375	Metal-dependent proteases with chaperone activity	168	35	203

Basic machinery of life genes (ribosomal)

Are universal and appear only once *(mostly)*

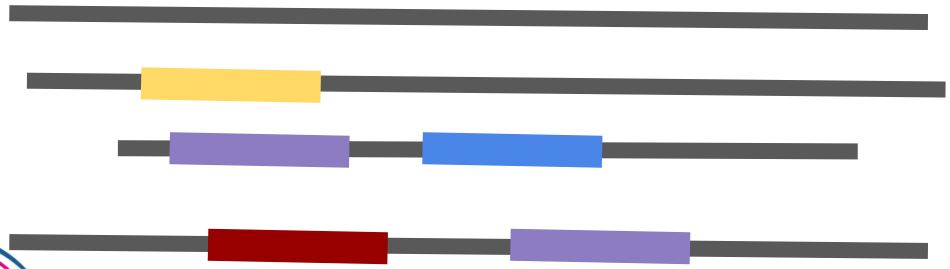
Many different sets have been proposed

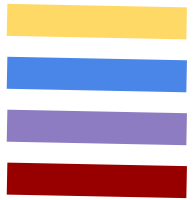
On the left, from (Ciccarelli et al., Science, 2006)

checkM1

Marker gene based:, a good genome has

1. **All** single copy marker genes
2. **No** single copy marker gene appears twice





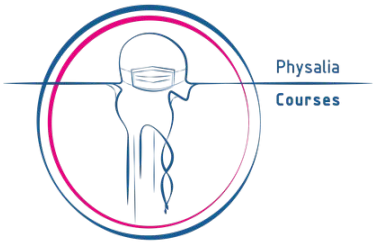
4 marker genes

Physalia
Courses

Other methods for QC I: checkM2

CheckM2, which uses machine learning

Different intuition: genes form groups and so seeing gene A1 means you should expect A2



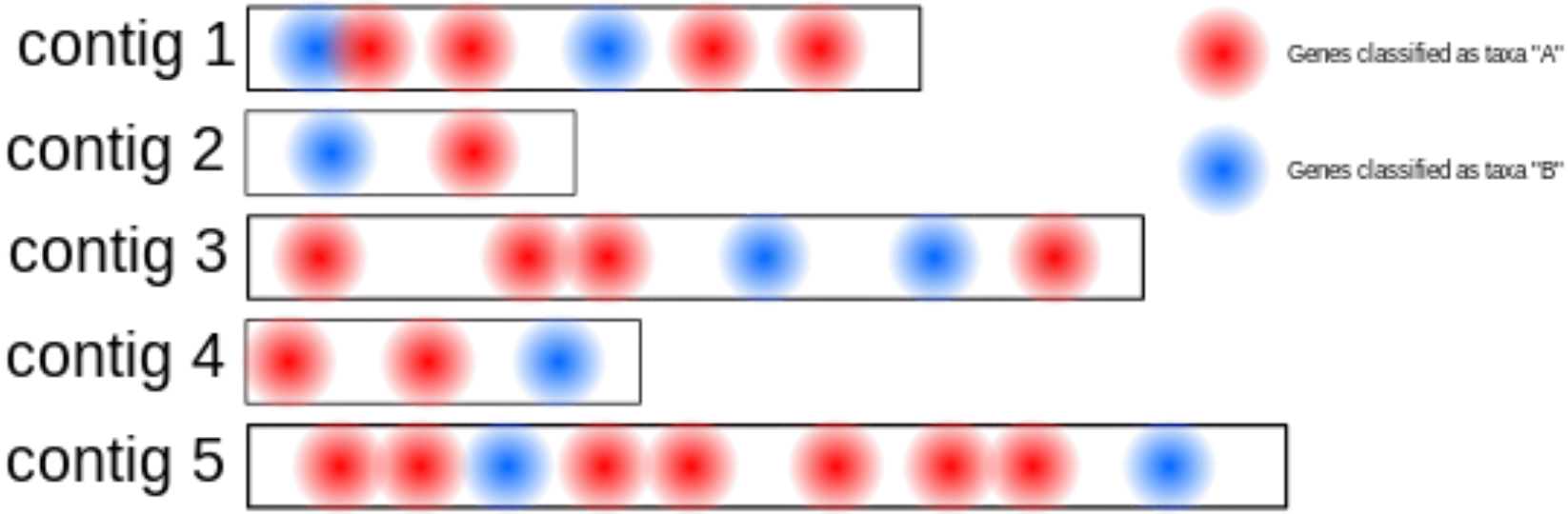
Article | Published: 27 July 2023

CheckM2: a rapid, scalable and accurate tool for assessing microbial genome quality using machine learning

Alex Chklovski, Donovan H. Parks, Ben J. Woodcroft & Gene W. Tyson

Nature Methods **20**, 1203–1212 (2023) | Cite this article

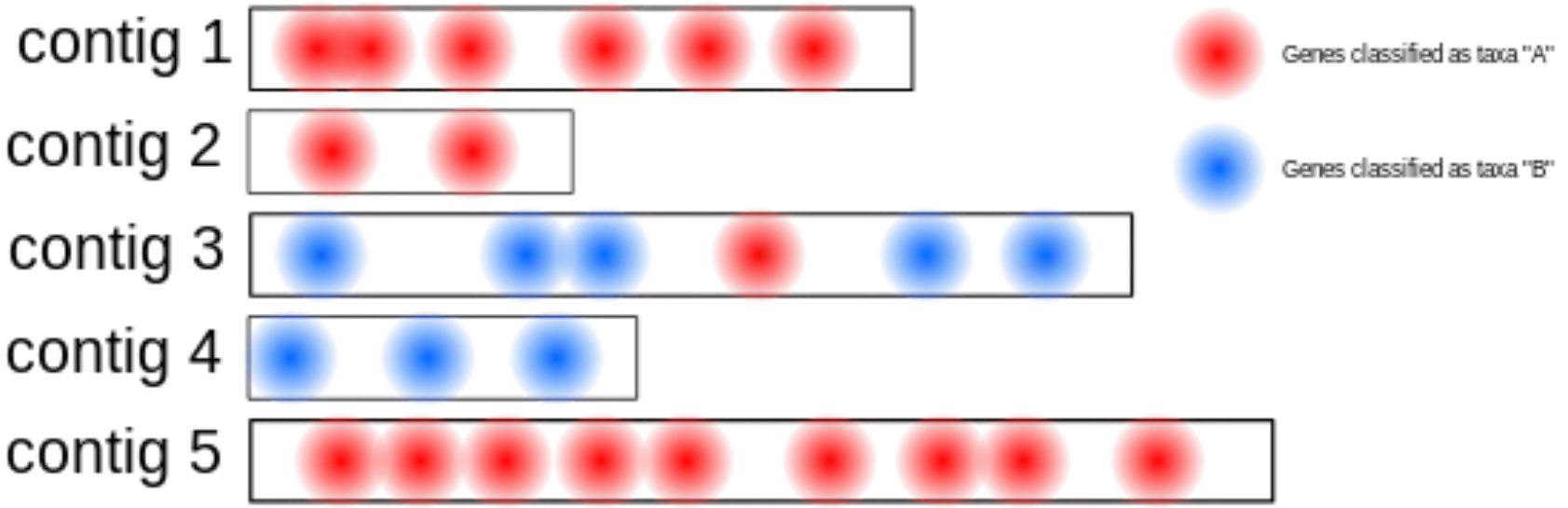
Other methods for QC II: GUNC



Physalia Courses

From https://www.big-data-biology.org/blog/2021/02/12/gunc/

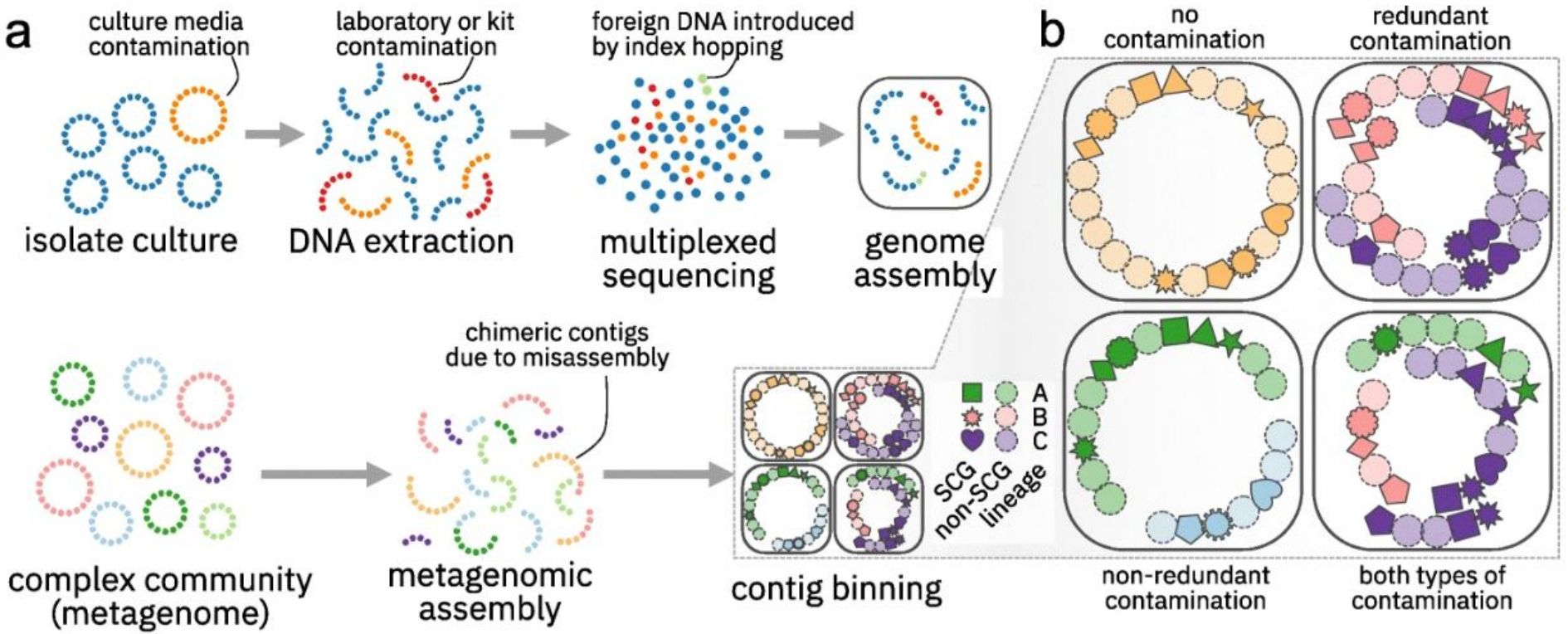
Other methods for QC II: GUNC



Physalia Courses

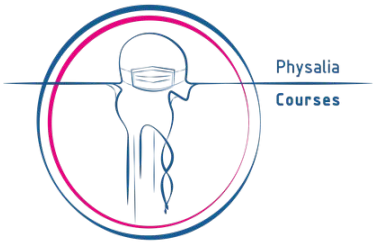
From https://www.big-data-biology.org/blog/2021/02/12/gunc/

From: GUNC: detection of chimerism and contamination in prokaryotic genomes



Limitations of current binning/QC methods

- Non-chromosomal elements
  - Plasmids can be very important for function/strain specificity, not captured by most methods
  - Very active area of research right now
- Species that are distant from reference genomes/*"weird"* species
  - checkM2 will always report low completeness for Allobaculum (likely others)
- What to do about *Microeukaryotes*?
  - Binning won't work well
  - Some methods work only for prokaryotes (e.g., because they use prokaryotic marker genes)



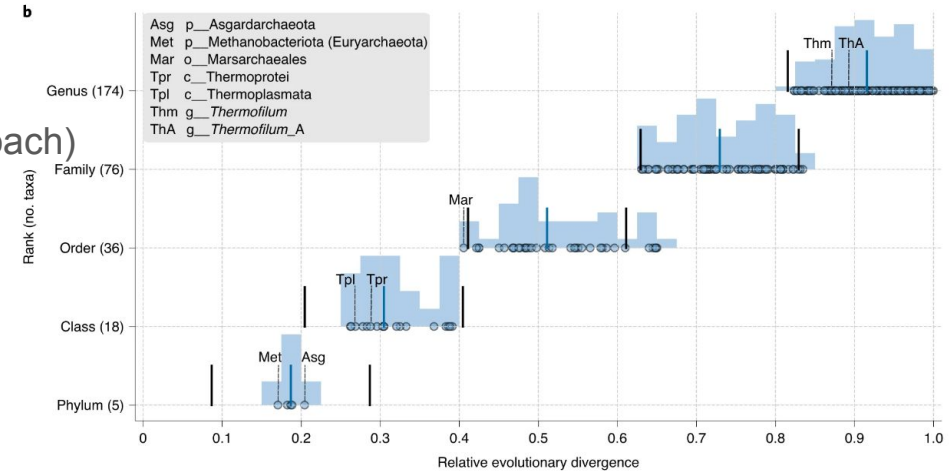
Taxonomic annotation: the GTDB

What species/genus/… is this genome from?

- GTDB: Genome Taxonomy Database
- https://gtdb.ecogenomic.org/
- Very important
  - There are different versions!
  - (NCBI takes the living document approach)
- Purely genomic based
  - *Shigella* is just a funny *E. coli*

  see (Parks et al., bioRxiv, 2021)

Below: Fig 1b in
(Rinke et al., Nat Micro, 2021)

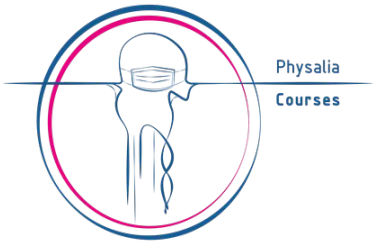


Physalia
Courses

An important topic we do not cover in depth

Multiple sample topics

1. Multi-sample binning
2. Co-assembly
3. Dereplication

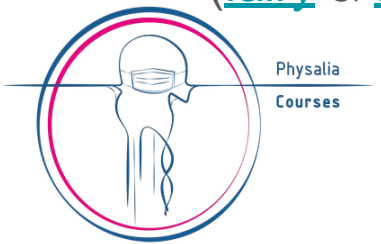


Multi sample binning

- Best results
  But very slow

Alternatives

1. Concatenate
   (VAMB & SemiBin)
2. Choose samples cleverly
   (Bin chicken — preprint only)
3. Better tools
   (fairy & strobealign-aemb – 2024 tools, AEMB still unpublished)



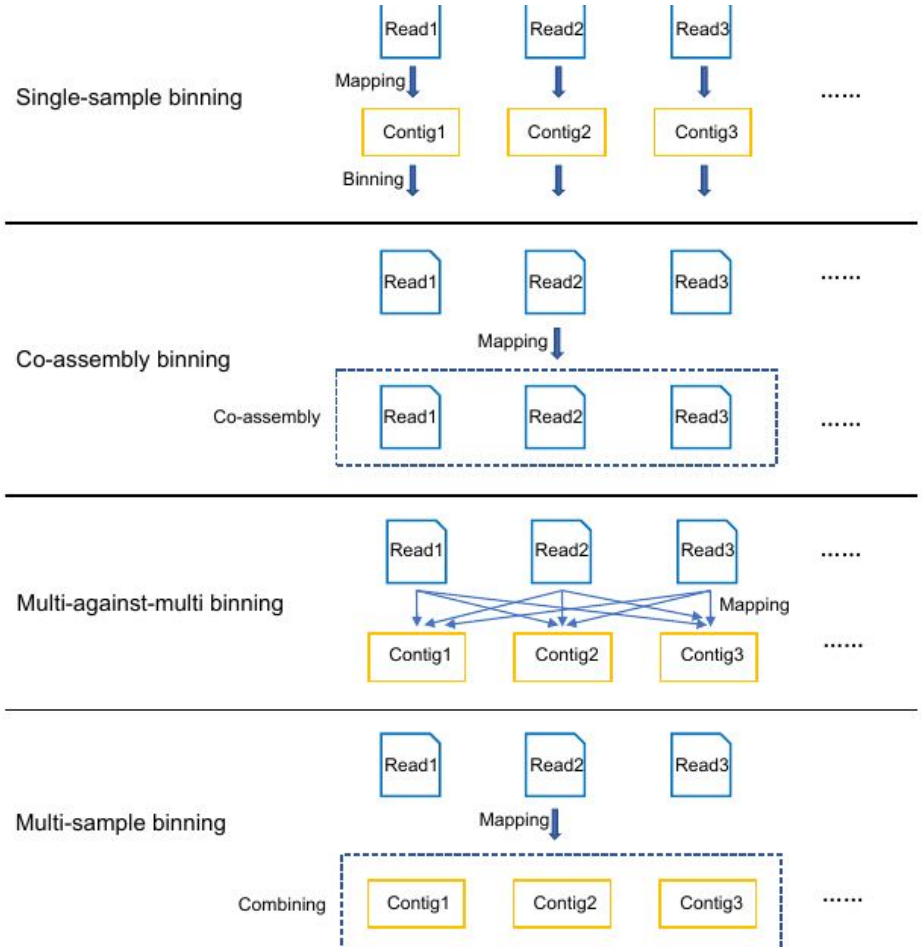
Brief Communication | Published: 29 June 2023

A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination

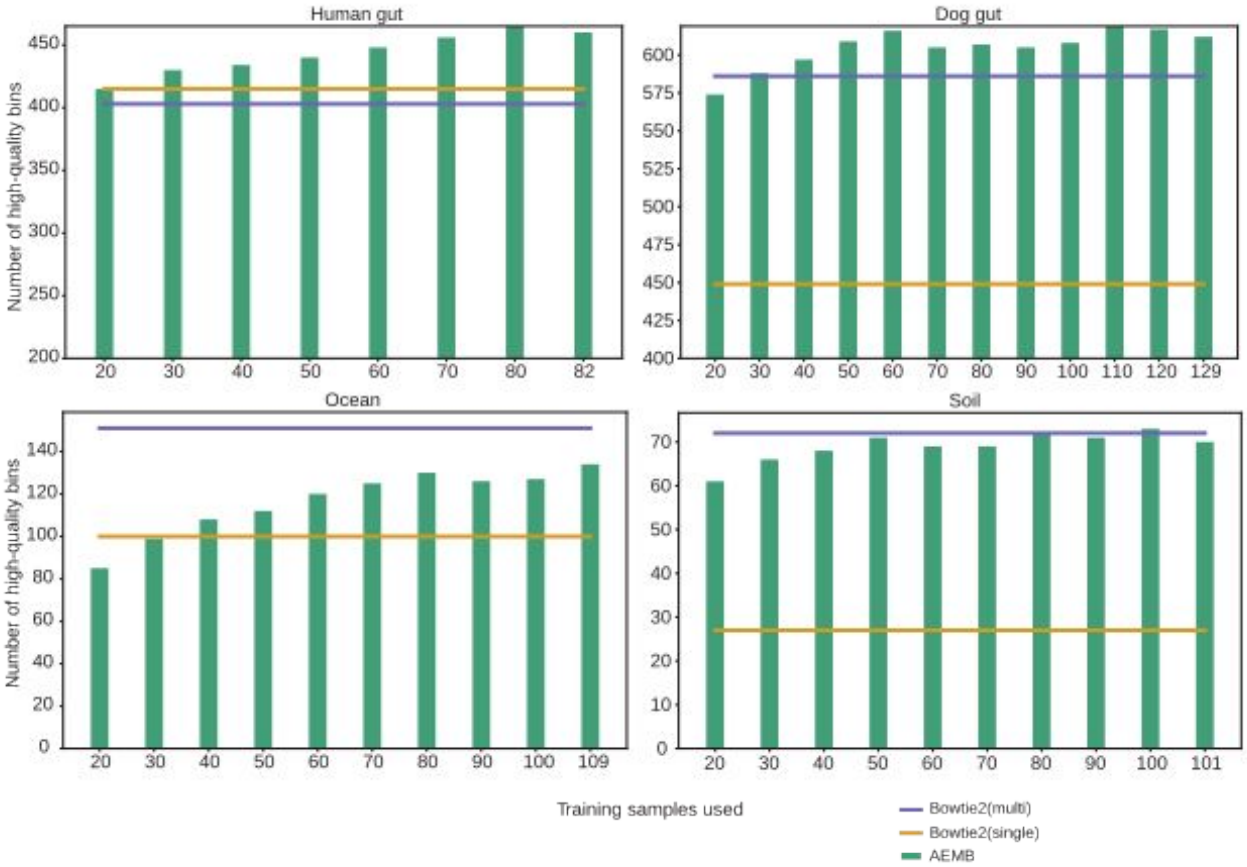
Jennifer Mattock & Mick Watson

Nature Methods **20**, 1170–1173 (2023) | Cite this article

Different modes



Physalia
Courses



Physalia
Courses

Dereplication

1. If you have got multiple samples
    a. 95% ANI with dRep
2. If you have got multiple MAGs from the same sample
    a. SemiBin2
    b. MetaBAT2
    c. VAMB
    d. Metadecoder
    e. …