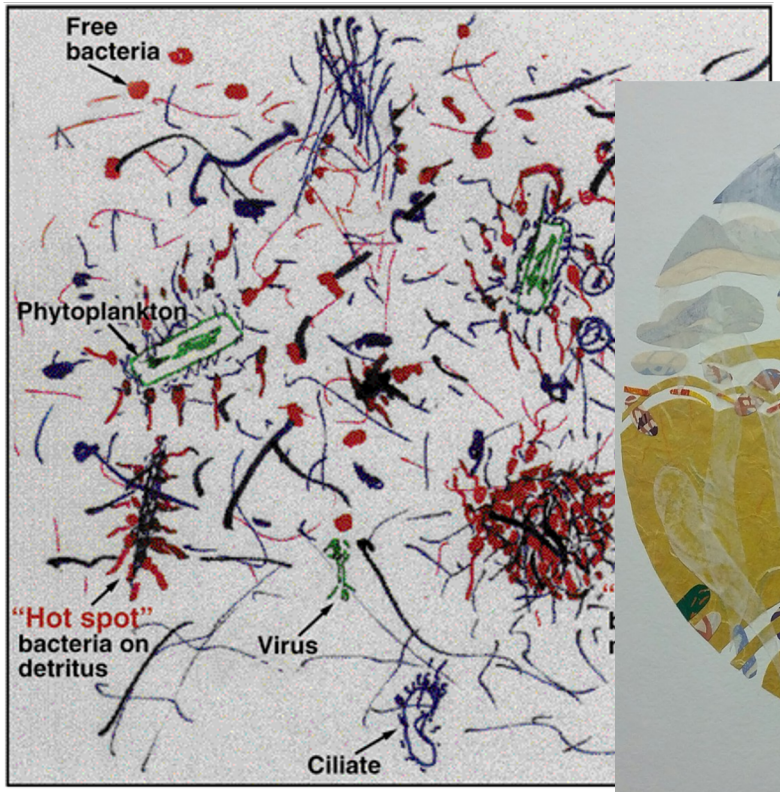
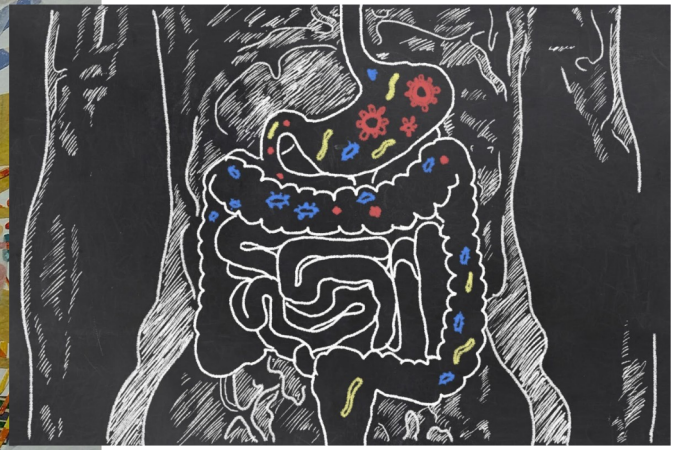
Metagenomics Lesson 8



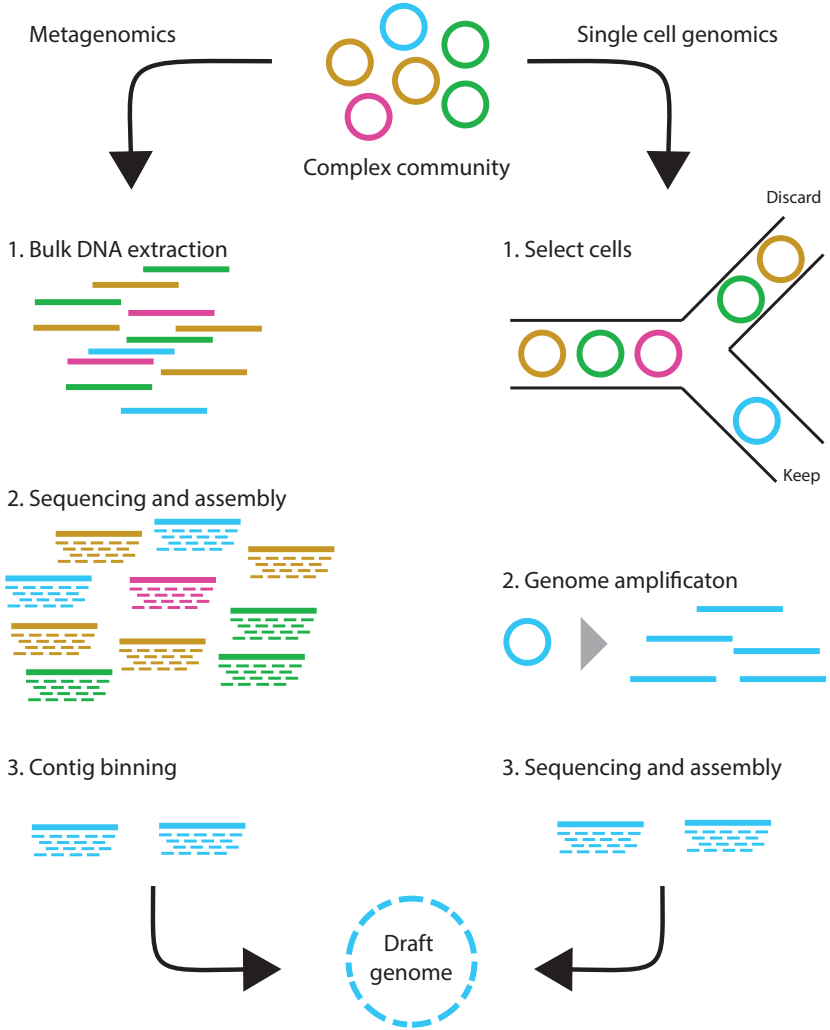


BVCN

biovcnet.github.io

Lesson 8 – Binning

1. What & why
2. How does binning work
3. Methods & tools
4. Complementary approaches
5. Demo

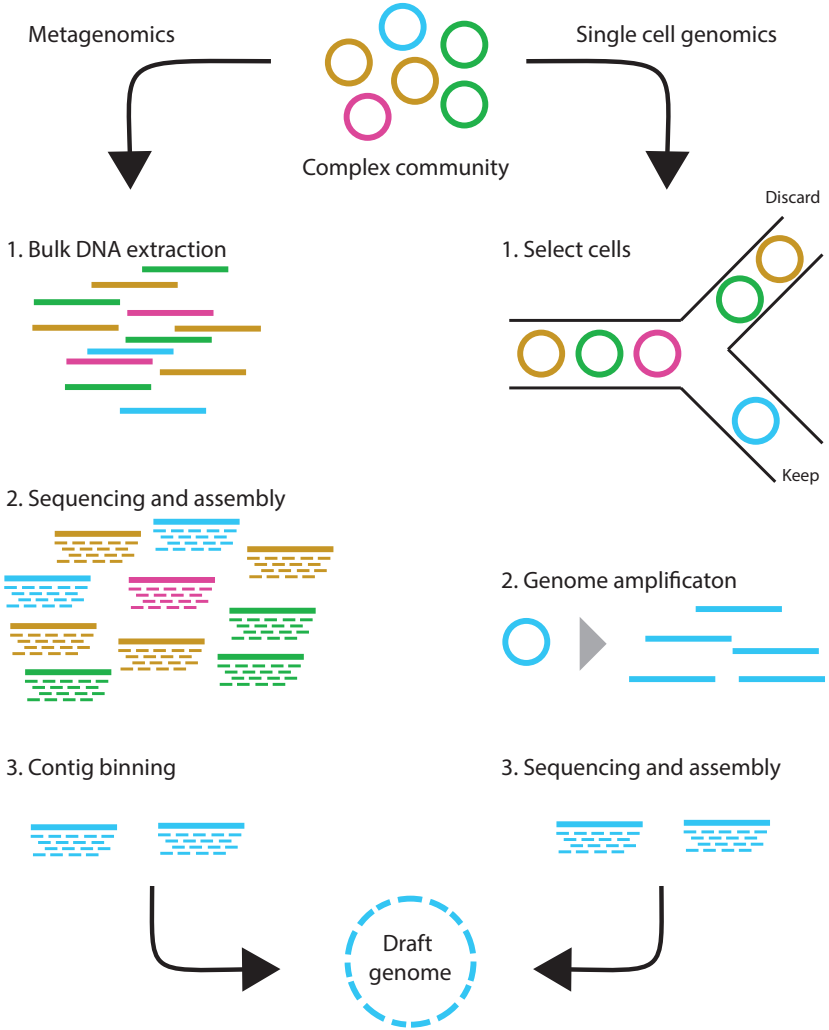


What is binning?

Grouping sequenced & assembled DNA sequences from a metagenome into separate draft genomes.

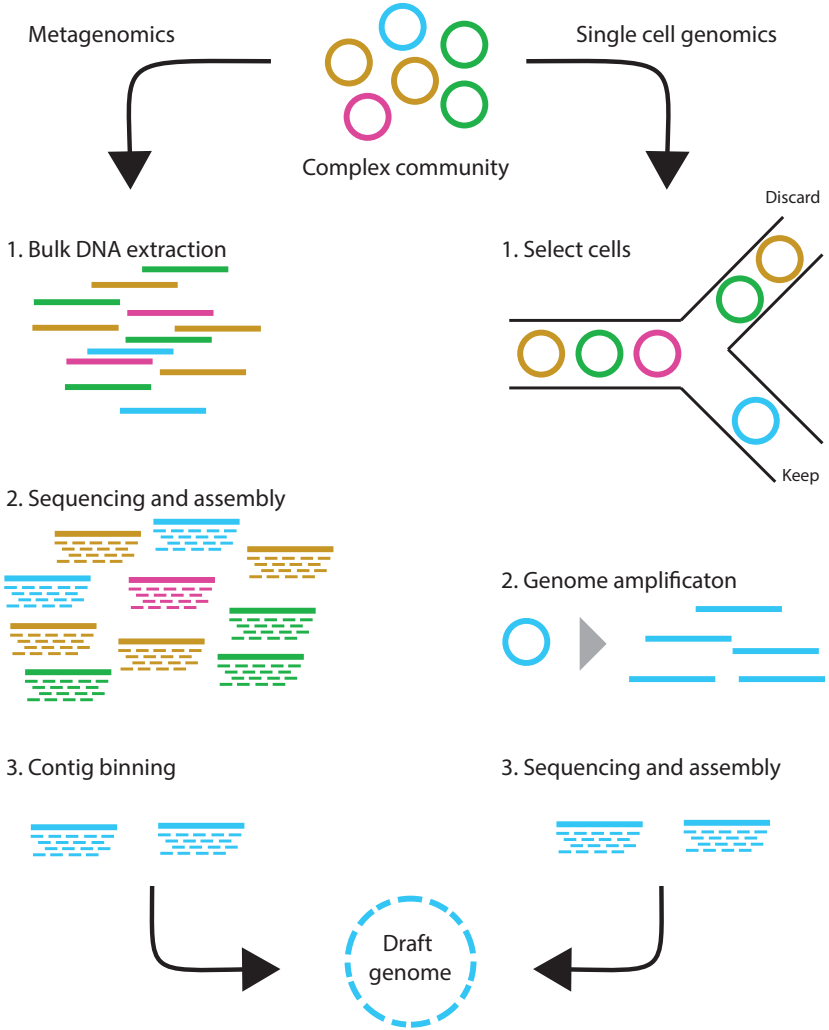
Currently largely restricted to Bacteria & Archaea

Possibly a “transient problem”



Why do binning?

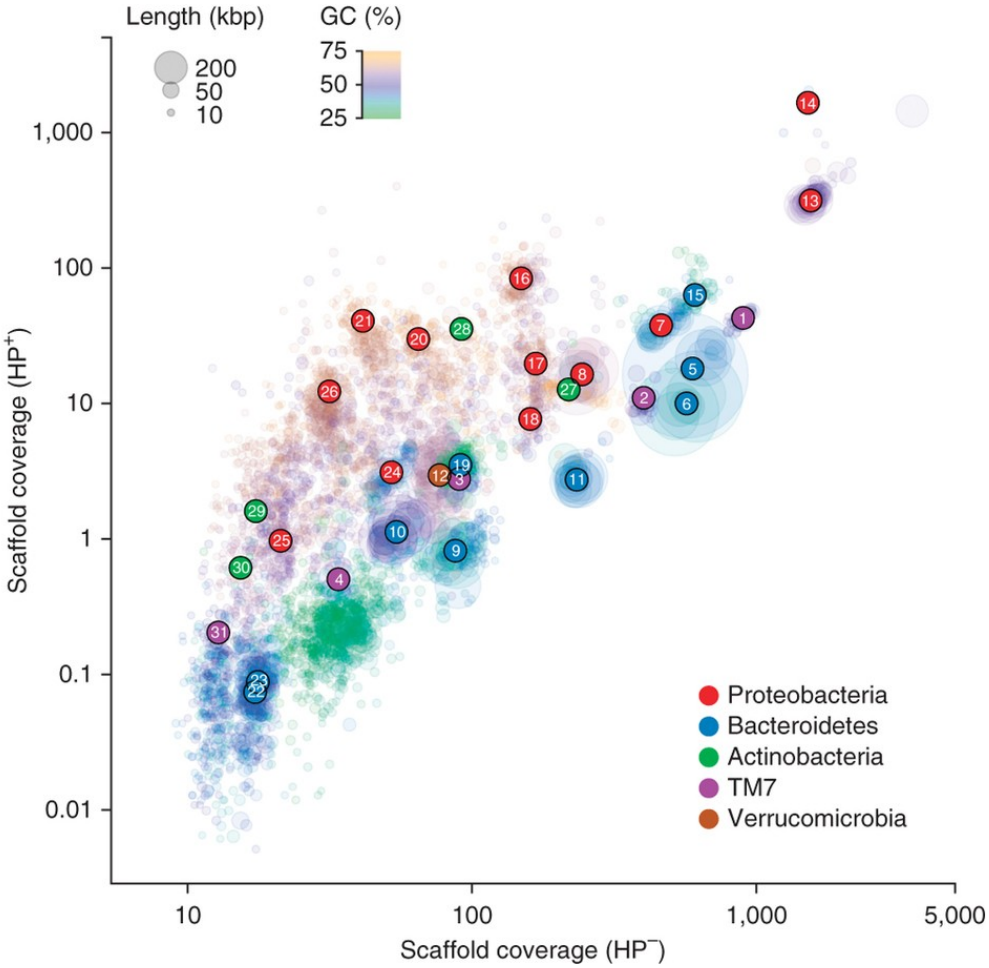
- Discovery of novel organisms
- Insight in trait and organism evolution
- Predicting community interactions
- Physiology hypothesis generation



How does binning work?

Takes advantage of conserved features across the genome.

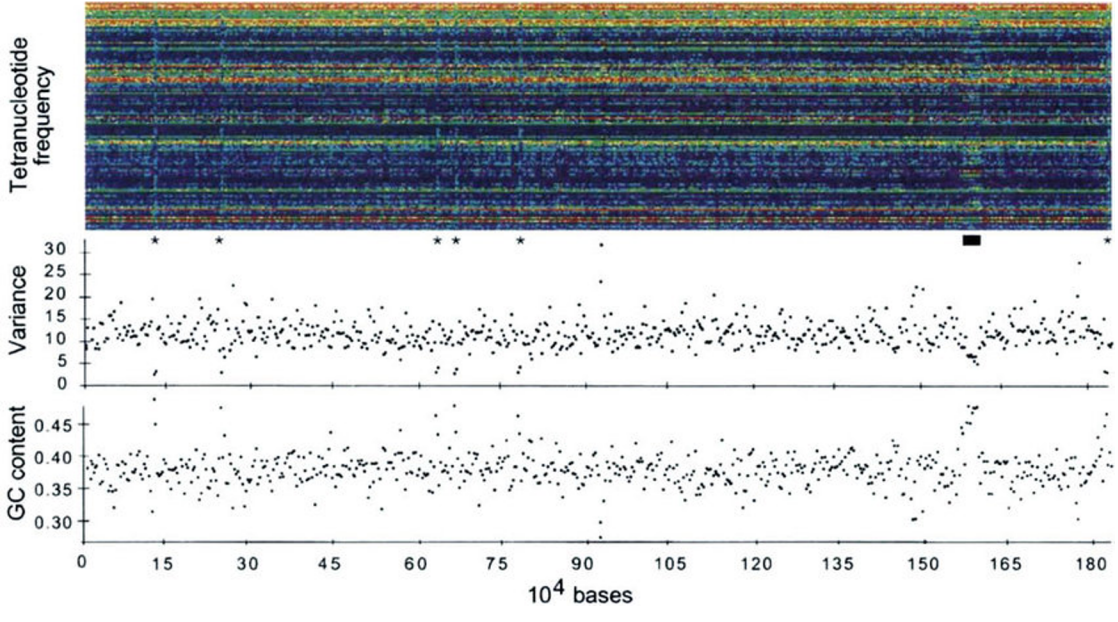
- Sequence composition
- Abundance of the DNA
- Taxonomic affiliation (of encoded features)



Lesson 2: Sequence composition is conserved in a genome

At the level of 4-mers (aka “tetranucleotide frequency”), bacterial genomes have conserved “fingerprints” which are used in metagenomic binning tools and to detect horizontal gene transfer events.

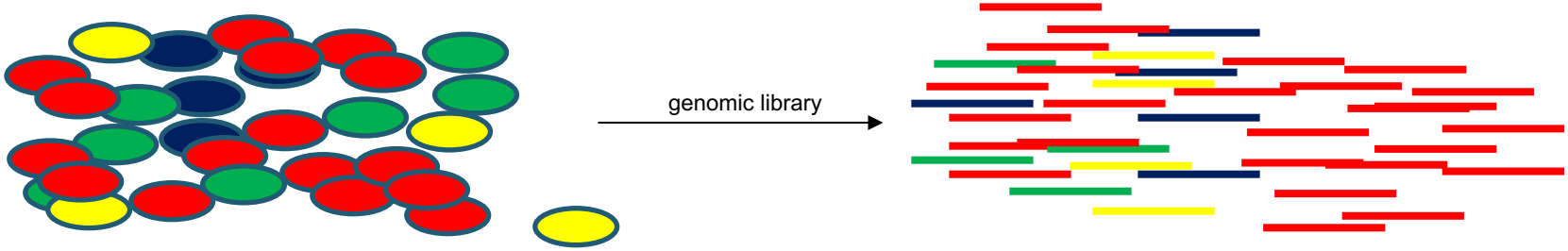
This figure shows an *H. influenzae* genome, where each column is the 4-mer fingerprint of a 3000 bp region and each row is a 4-mer colored from low abundance (purple) to high (red).



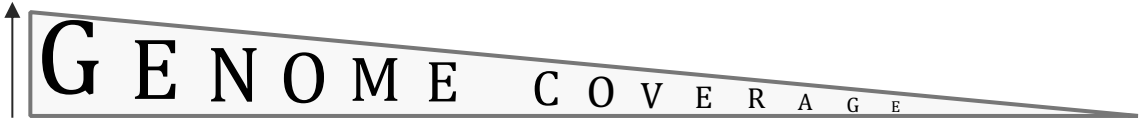
Starred (*) sections contain ribosomal RNA. The black bar identifies a bacteriophage.

Noble et al. (1998).

Lesson 5: Coverage depends on genome relative abundance



4 dominant species

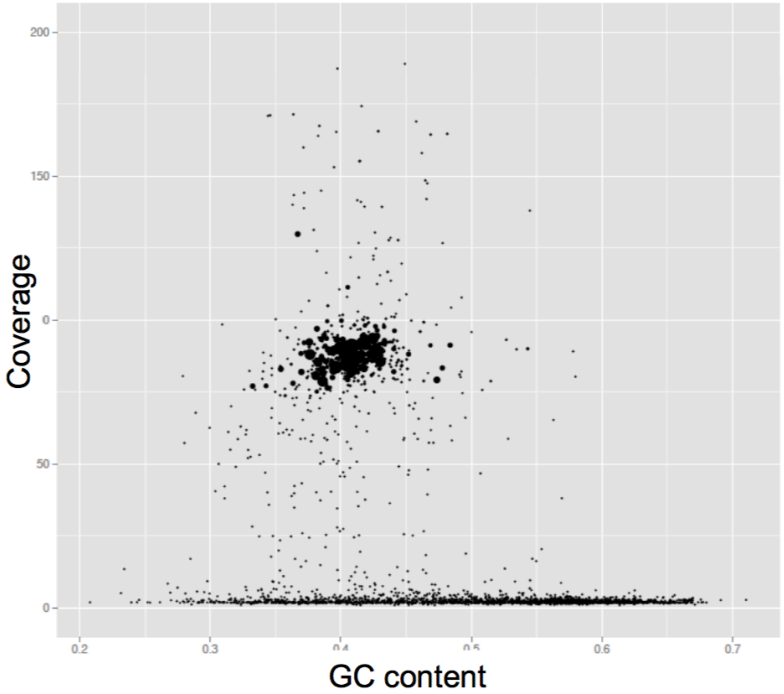


Abundant species >> moderately abundant species > rare species

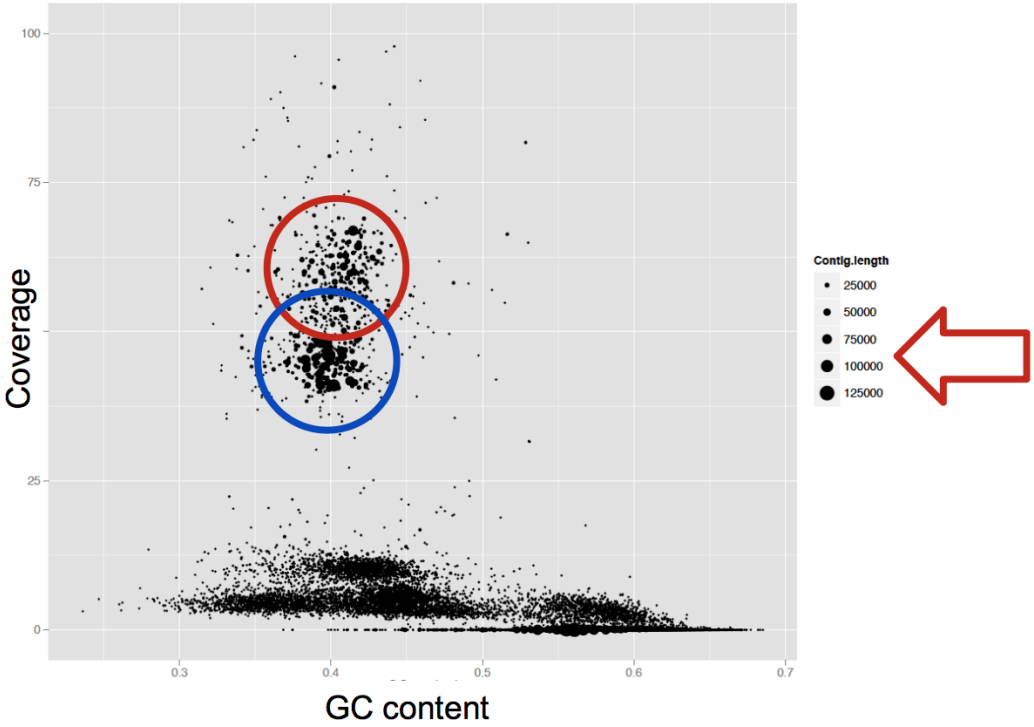
10X coverage of species 1
0.5X coverage of species 4

Abundant groups: Draft Genome Assembly Possible
Rare groups: Only Read-Based Analysis Possible

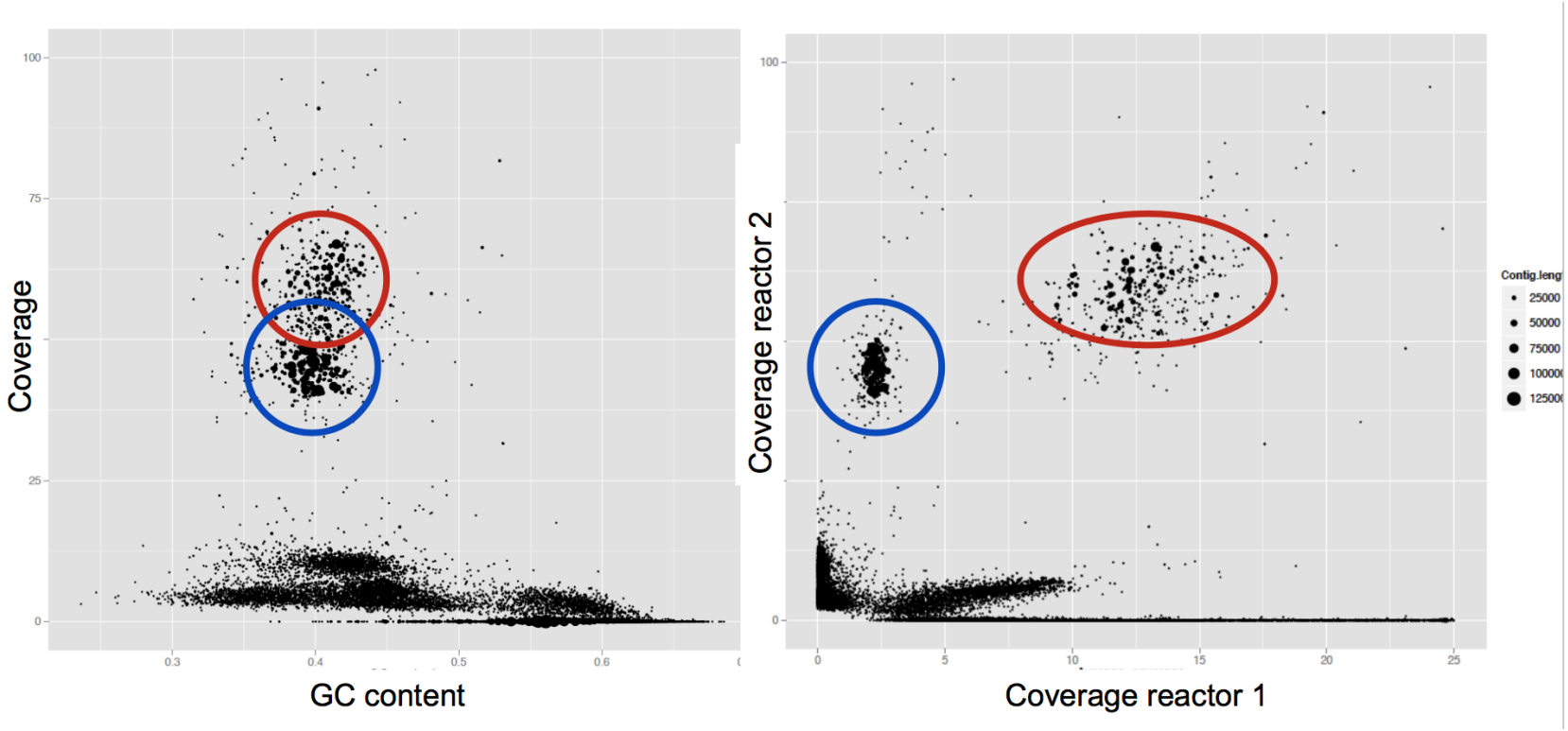
In practice: binning by sequence composition and abundance



In practice: binning by differential abundance



In practice: binning by differential abundance

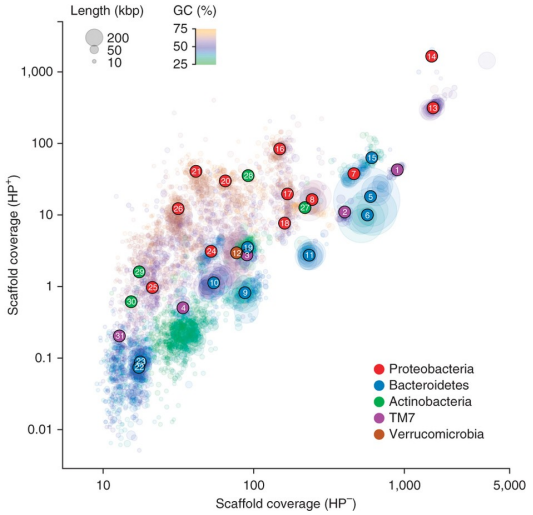


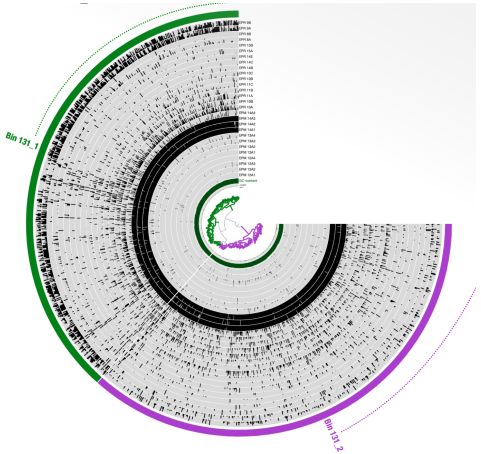
Binning approaches: “Manual”

- mmgenome (Albertsen et al 2013)
- anvi’o (Eren et al 2015)
- ESOM (Dick et al 2009)
- Composition based matrix & dimensionality reduction (tSNE/PCA/UMAP)

- Low throughput but tailored decisions per bin

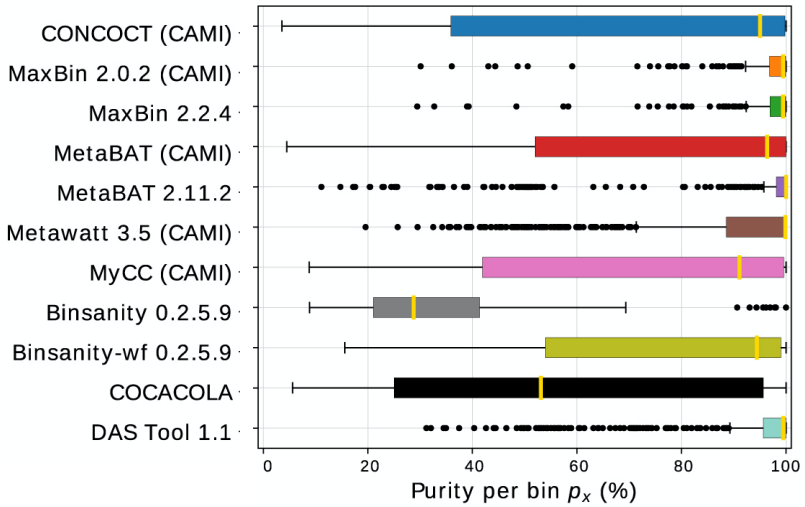
- Reproducibility can be challenging

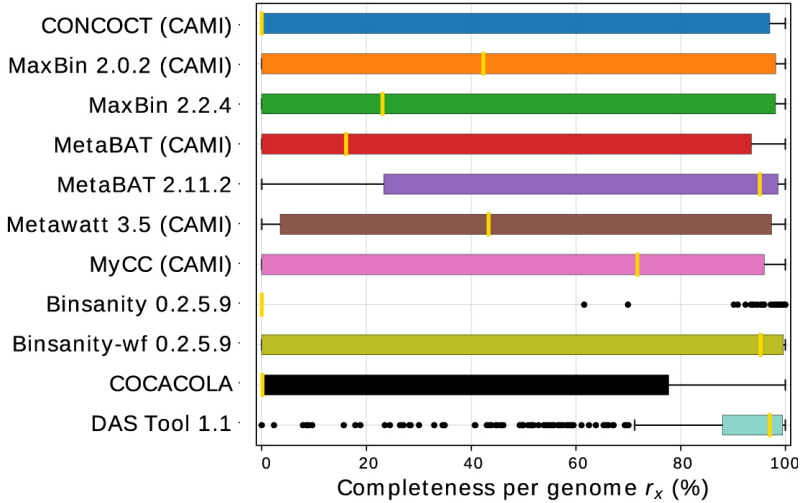




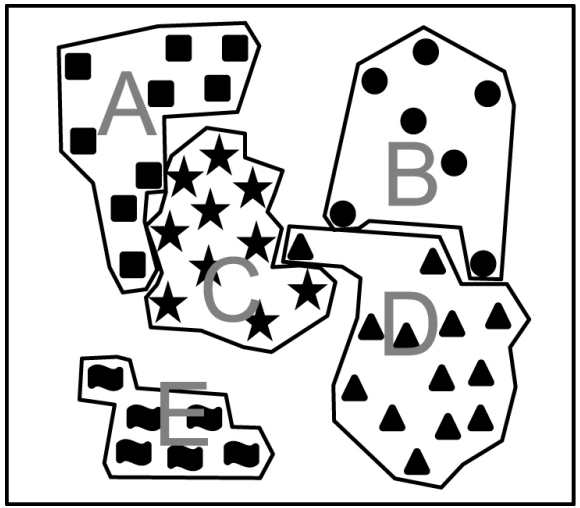
Binning approaches: Automated

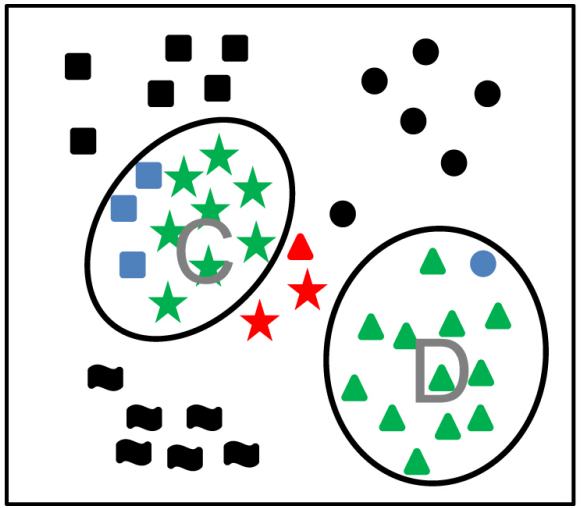
- Many tools, often relatively easy to run.
- Use the same underlying information, but weigh different
- Work best with higher number of samples
- Subject of recent large scale evaluations (e.g. Meyer et al 2018)





Possible errors in binning





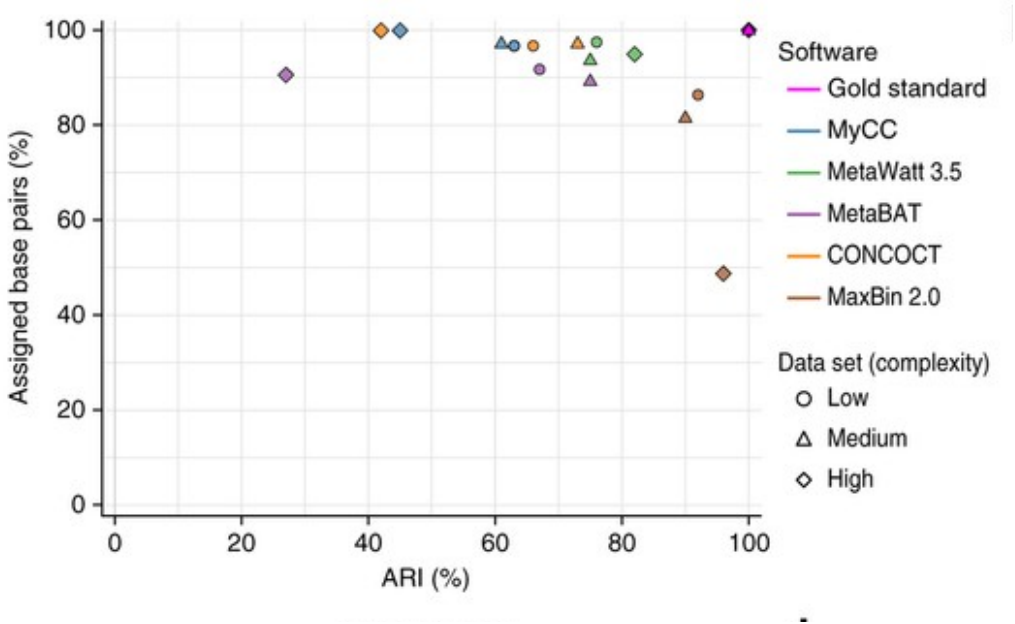
Read or contig of genome A
Read or contig of genome B
Read or contig of genome C
Read or contig of genome D
Read or contig of genome E
True positives (*TP*)
False positives (*FP*)
False negatives (*FN*)

$$ARI = \frac{\sum_{x,y}\binom{m_{x,y}}{2} - \frac{\sum_x \binom{m_{x,.}}{2} \sum_y \binom{m_{.,y}}{2}}{\binom{m}{2}}}{\frac{1}{2}\left[\sum_x \binom{m_{x,.}}{2} + \sum_y \binom{m_{.,y}}{2}\right] - \frac{\sum_x \binom{m_{x,.}}{2} \sum_y \binom{m_{.,y}}{2}}{\binom{m}{2}}},$$

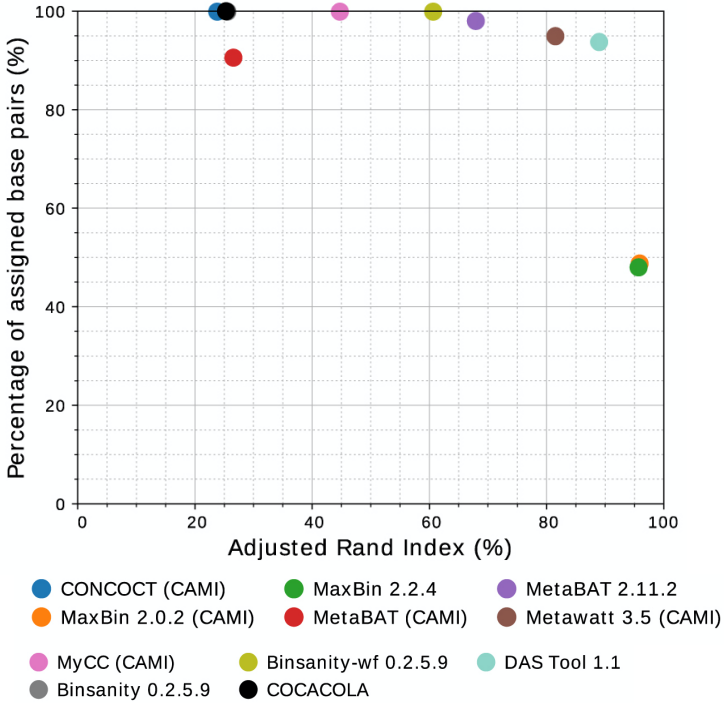
Meyer et al 2018

Evaluating binning approaches

CAMI (Szcyrba et al 2017)



AMBER (Meyer et al 2018)

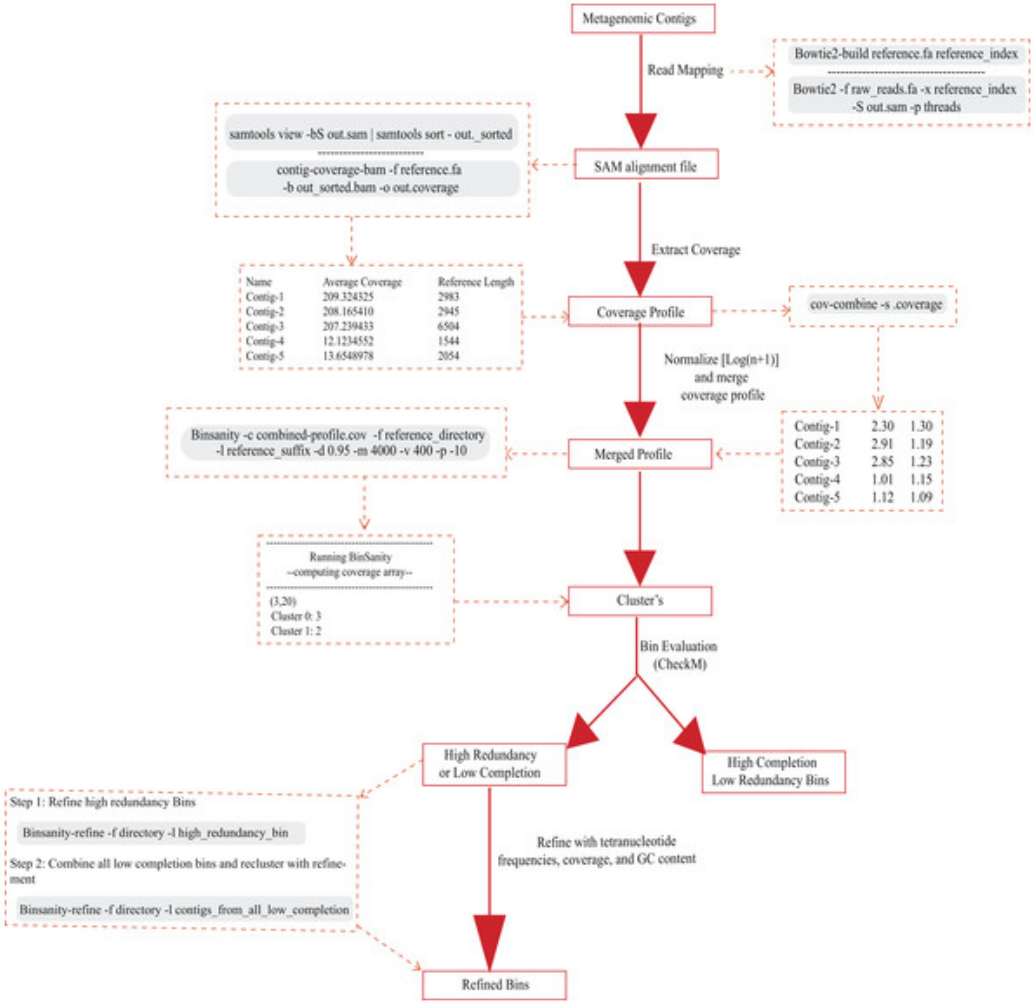


Binning approaches: Metabat 2 (Kang et al 2019)

- Based on tetranucleotide frequency (TNF) and coverage (ABD)
- Contigs are assigned a score S
  $S = \sqrt{TNF^{(1-w)} * ABD^{w} * COR}$
  where $w = nABD/(nABD + 1)$
- Builds a graph of contigs initially based on TNF
  Contigs = nodes, similarity = edges
- Binning by graph partitioning: modified label propagation algorithm
- Post partitioning recruiting of: short contigs (< 2500bp) & contigs in small bins (< 200k)
- Criterion: avg S similarity to contigs in bin is greater than avg S similarity between contigs in bin

Binning approaches: Binsanity (Graham et al 2019)

- Coverage based
- Affinity propagation: Each contig is tested as “cluster center”
- Evaluation of bin quality
- Bin refinement with composition if needed

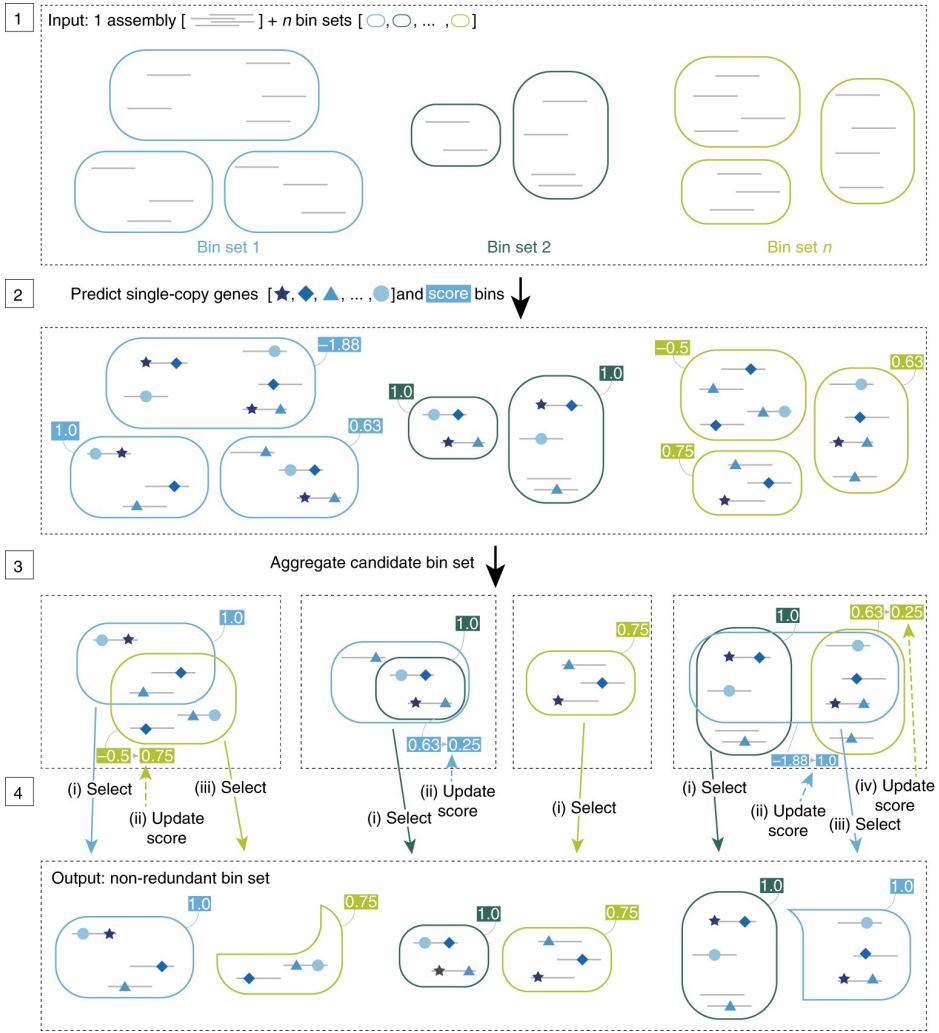


Binning approaches: DAS Tool (Sieber et al 2018)

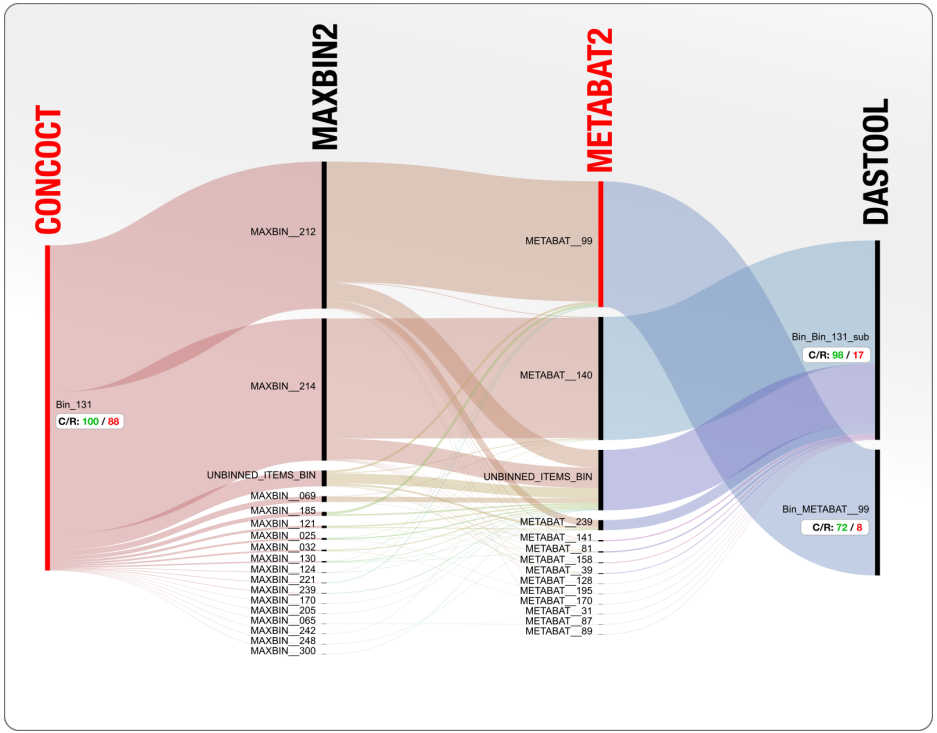
- Binning aggregator
- Scores each bin based on single copy markers:

$$S_b = \frac{\text{uSCG}}{\text{rSCG}} - b\frac{\text{dSCG}}{\text{uSCG}} - c\frac{\Sigma\text{SCG} - \text{uSCG}}{\text{rSCG}}$$

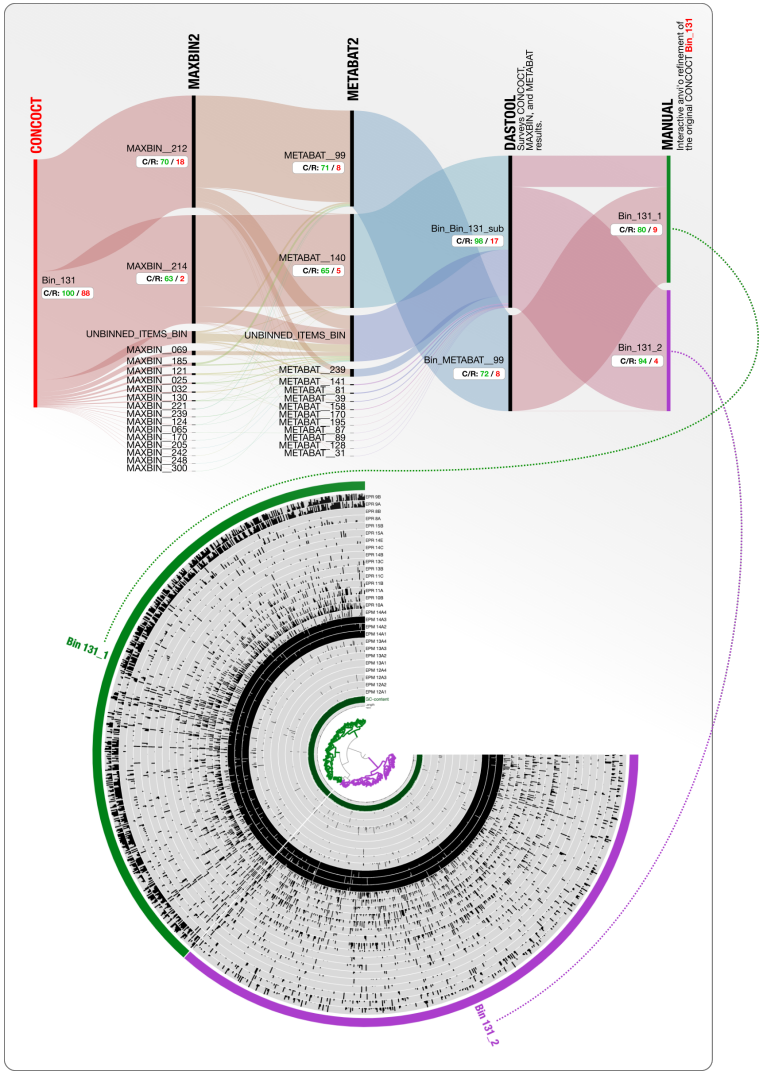
- Picks bin based on:
  Score > N50 > total length
- Rescores new bin set -> repeats



More binning evaluations:

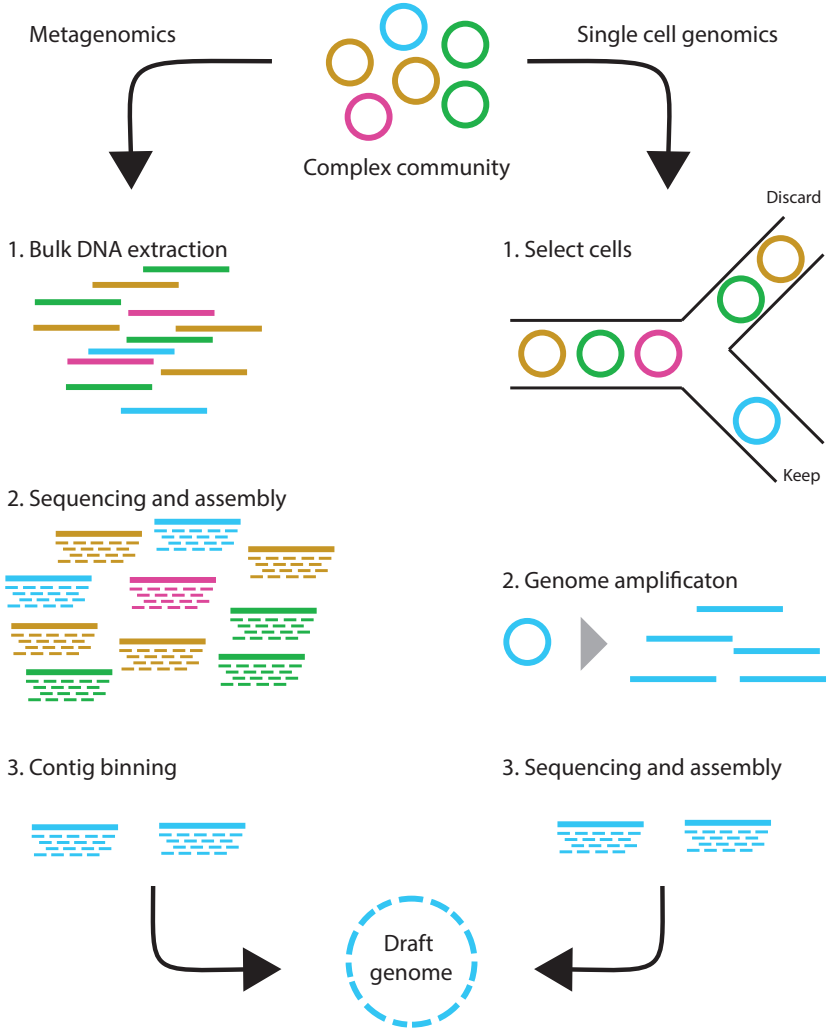


http://merenlab.org/2020/01/02/visualizing-metagenomic-bins/



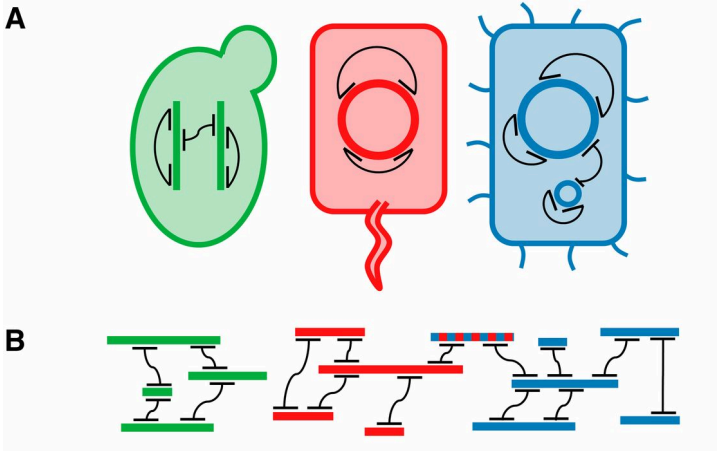
Complementary methods: Single cell genomics

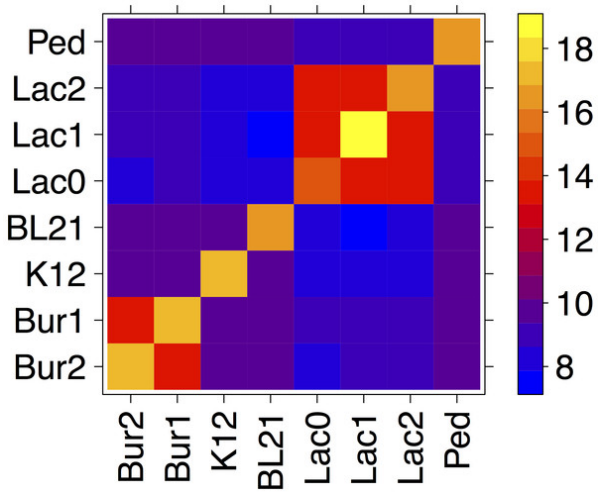
- Sorting of individual cells & MDA amplification
- Pros: individual genotypes, population variation, rare cells
- Cons: poor completeness (but improving), more lab intensive, lower throughput
- Much info available, could be it's own topic



Complementary methods: Hi-C metagenomics

- Crosslinking replicons before lysis.
- Proximity linkage: digest and re-ligate crosslinked DNA
- Sequence proximity linked DNA and untreated sample. Use links to bin assembly from untreated sample
- Can resolve individual replicons with a cell.
  (Beitel et al 2014, Burton et al 2014)

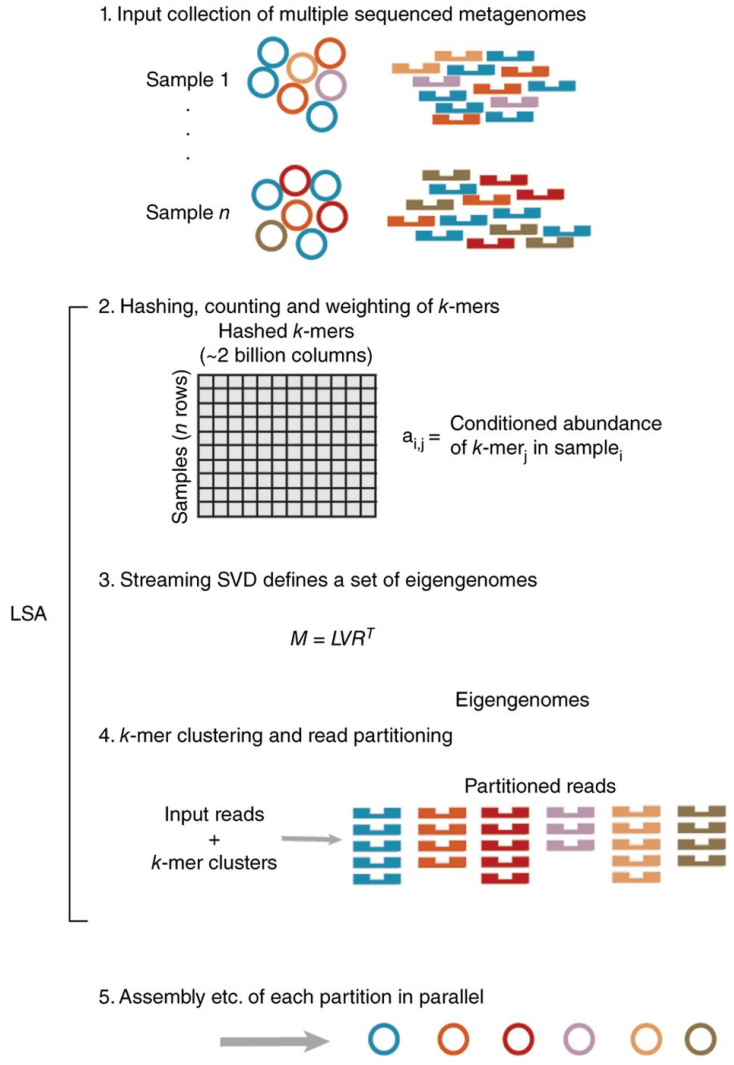




Analogous methods: Read partitioning

Using long k-mer presence/abundance to group reads into partitions.

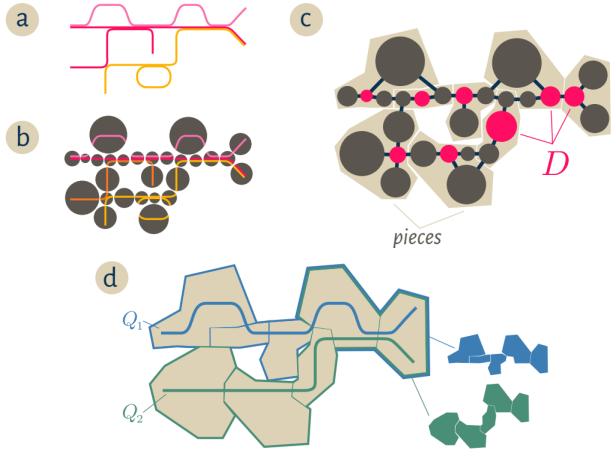
- Early binning tools
  (e.g. MetaCluster 5.0, Wang et al 2012)
- LSA
  (Cleary et al. 2015)
- LSH - Sparse coding
  (Kyrgyzov et al. 2020)

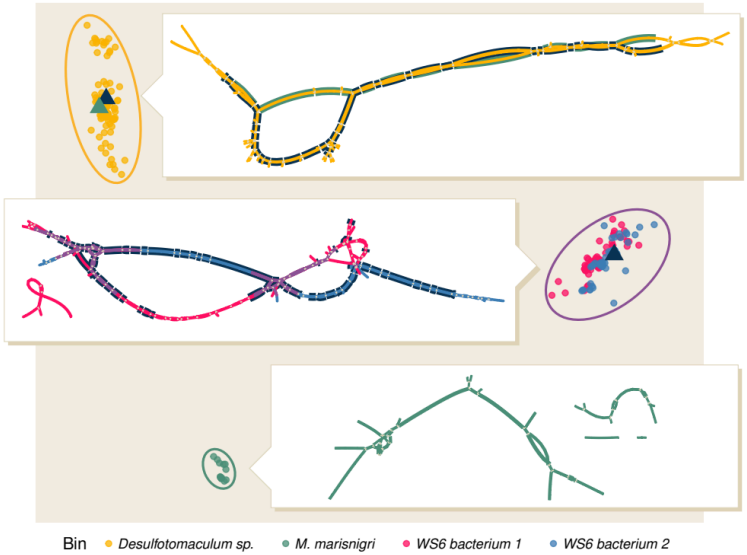


Analogous methods: Assembly graph query

Rather than “flattening” the assembly graph into contigs, directly extract the relevant areas from the graph

- (possibly) emerging approach
- Implementation: Spacegraphcats (Brown et al 2020)
- Allows deconvoluting strain variation





Binning approaches:
Demo

- Short review of:
  Metabat/Maxbin/CONCOCT/Binsanity

- Demo of DAS tool using example data