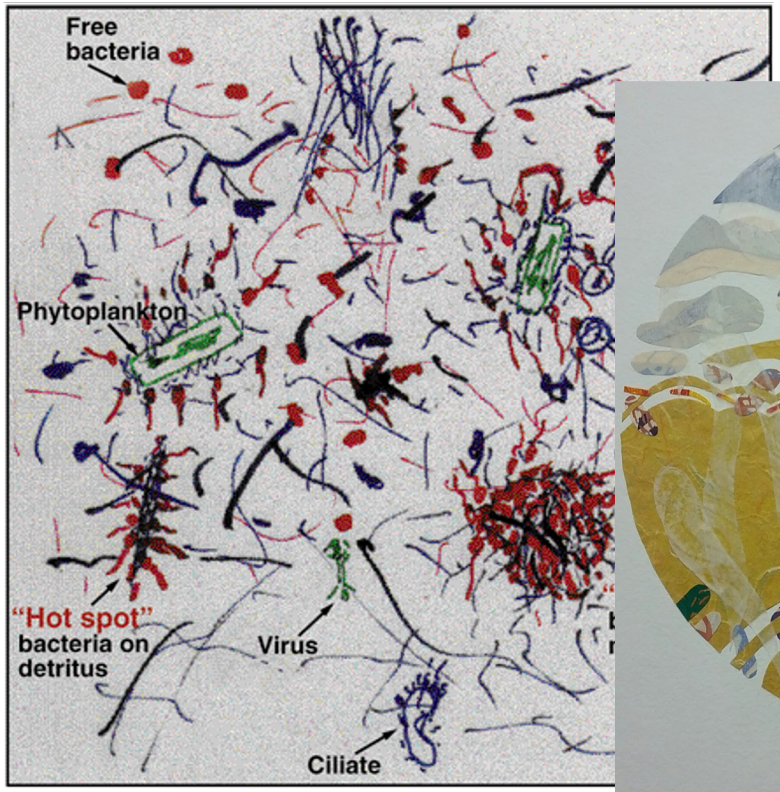
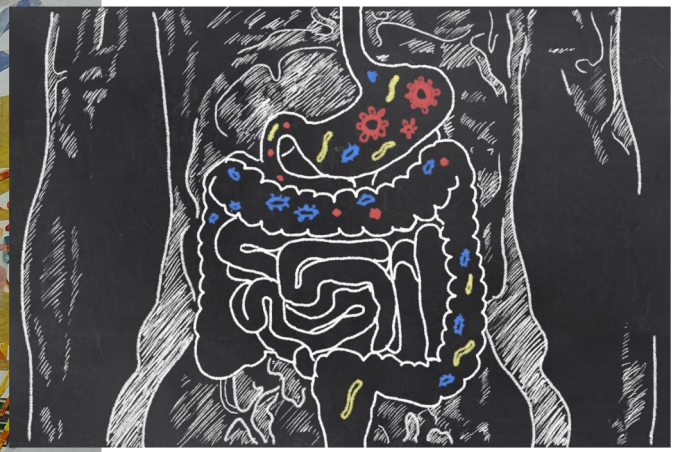
Metagenomics Lesson 9



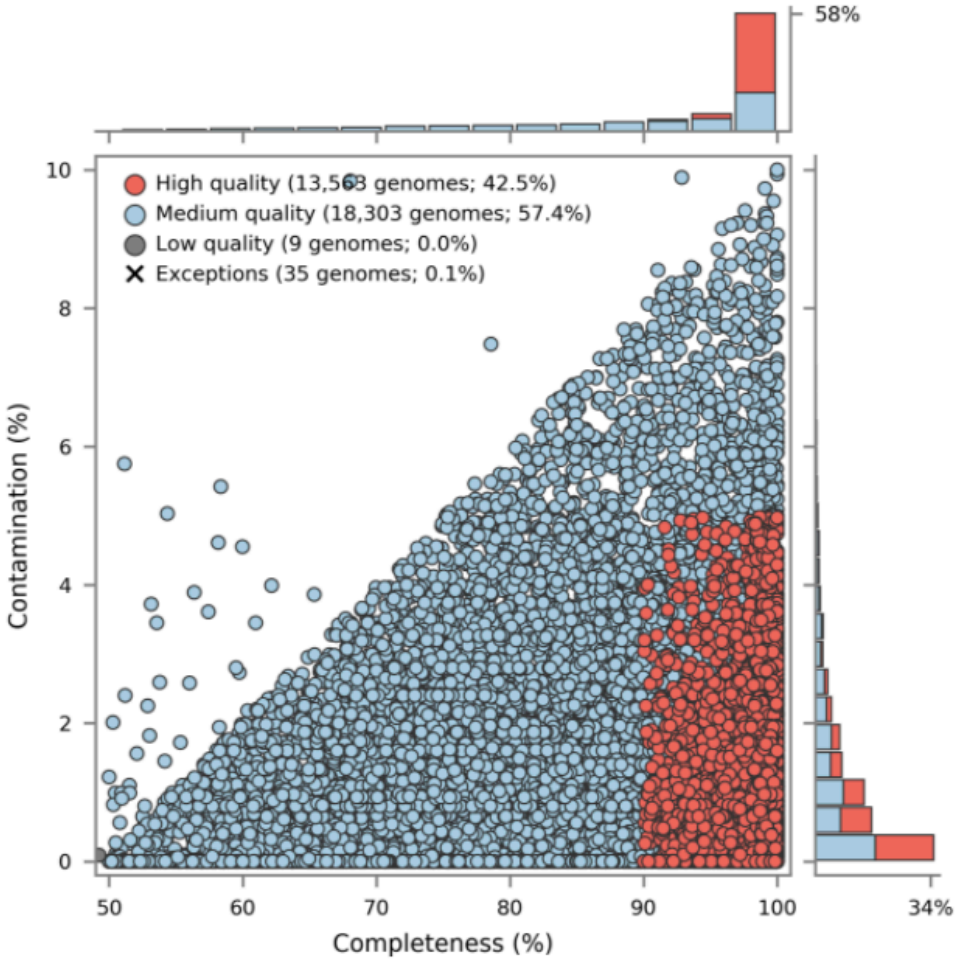


BVCN

biovcnet.github.io

Lesson 9 – Bin quality

1. What & why
2. Marker genes
3. CheckM: marker genes in practice
4. Beyond marker genes
5. CheckM Demo



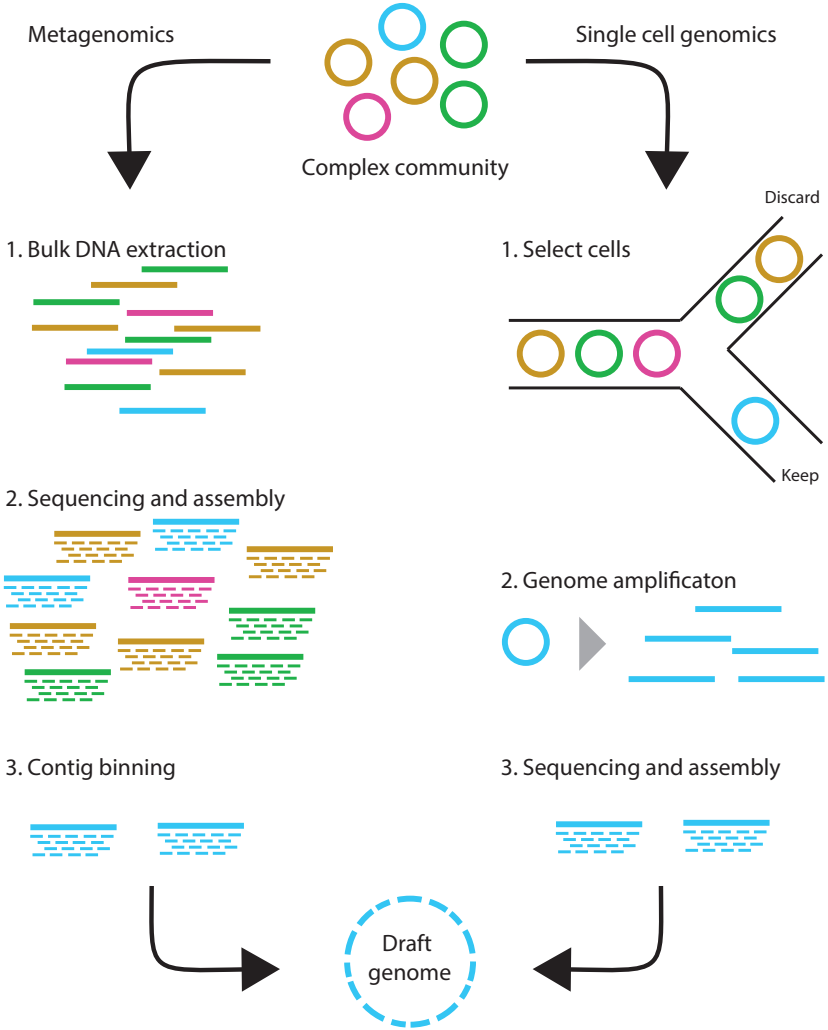
https://gtdb.ecogenomic.org/stats

Recap Lesson 8: What is binning?

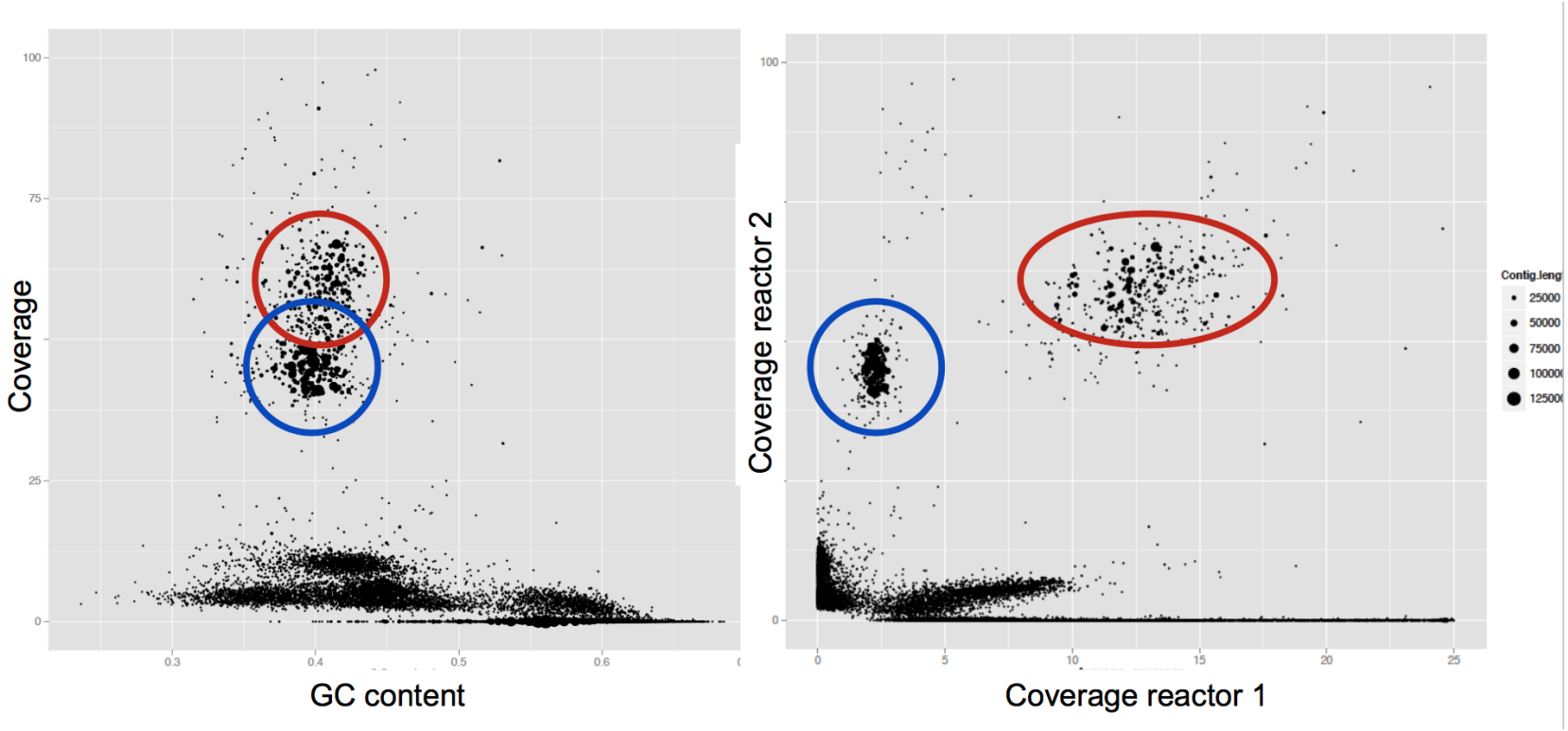
Grouping sequenced & assembled DNA sequences from a metagenome into separate draft genomes.

Currently largely restricted to Bacteria & Archaea

Possibly a “transient problem”

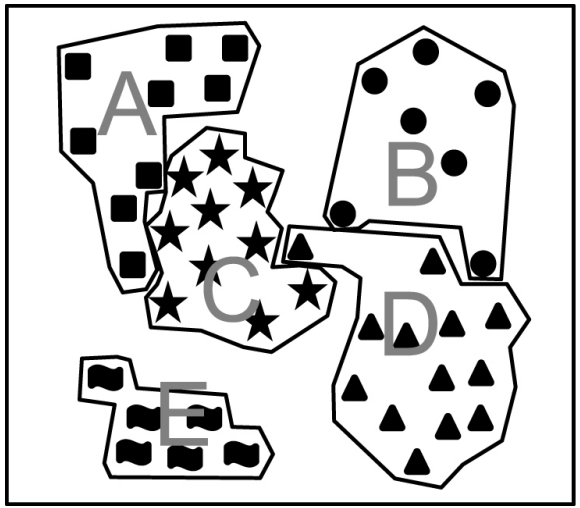


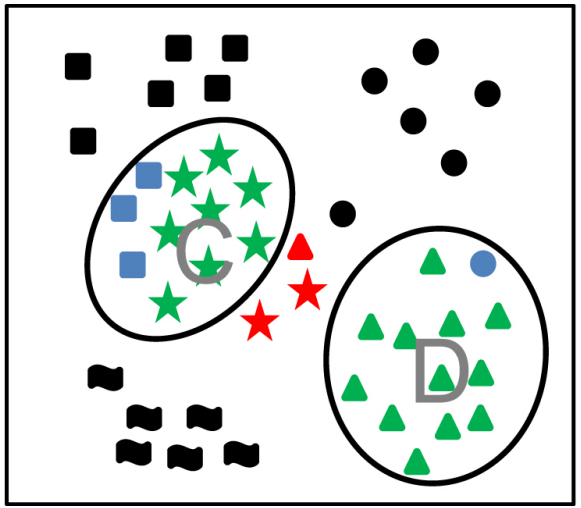
Recap lesson 8: binning by differential abundance



data from Speth et al 2015

Recap lesson 8: Possible errors in binning





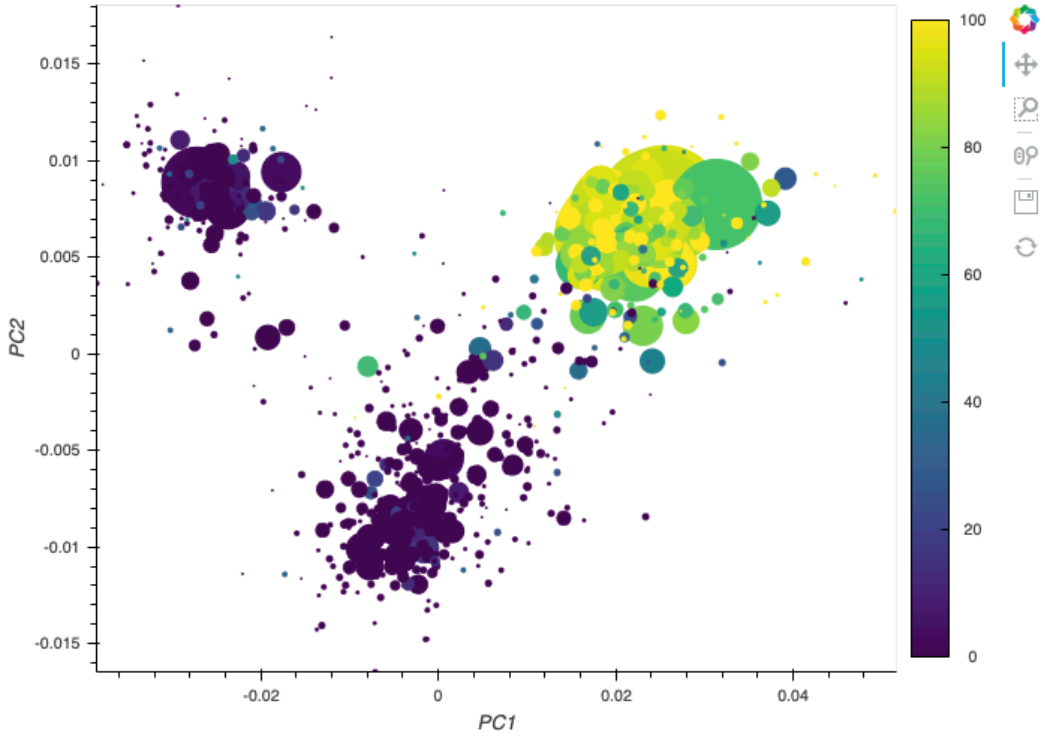
Read or contig of genome A
Read or contig of genome B
Read or contig of genome C
Read or contig of genome D
Read or contig of genome E
True positives (*TP*)
False positives (*FP*)
False negatives (*FN*)

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

Meyer et al 2018

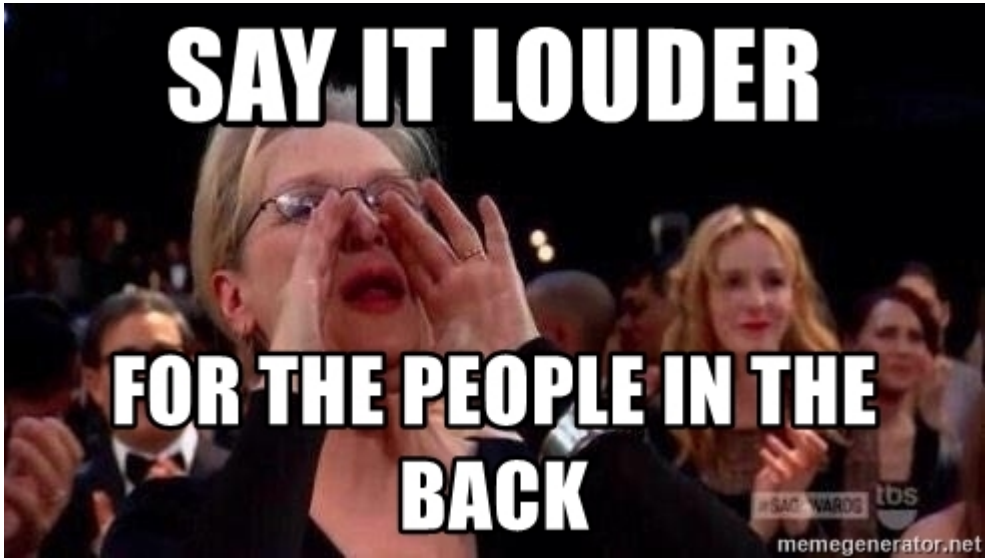
Consequences of poor quality data

- Database rot
- Hinders comparative & pan-genomics analysis
- Erroneous taxonomic affiliation of proteins



Speth et al ongoing work

Assessing bin quality is SUPER important



Early days of bin quality assessment

Dupont et al. 2012

Using the Comprehensive Microbial Resource as a database, 107 hidden Markov models (HMMs) that hit only one gene in greater than 95% of bacterial genomes were identified. […] These HMMs were then used to search the SAR86 genomes, and the **percentage of the total found was extrapolated for the entire genome.**

Albertsen 2013:

Currently, **there is no best practice for validation of population genome assemblies** from metagenomic data in terms of completeness and potential contamination with other species.

Hess et al. 2011

To estimate the completeness of the largest potential microbial draft genomes identified through this approach, **we first determined the most likely phylogenetic order from which each of these bins was derived.** For each of these orders, we used all available sequenced reference genomes to identify a **minimal set of core genes that are present in all members of this order**.

To address the possibility that the completeness of individual draft genomes was overestimated as a result of binning of scaffolds derived from multiple organisms, we further **validated their authenticity by copy number analysis of genes that were present only in single copy in all reference genomes of the respective phylogenetic orde**r.

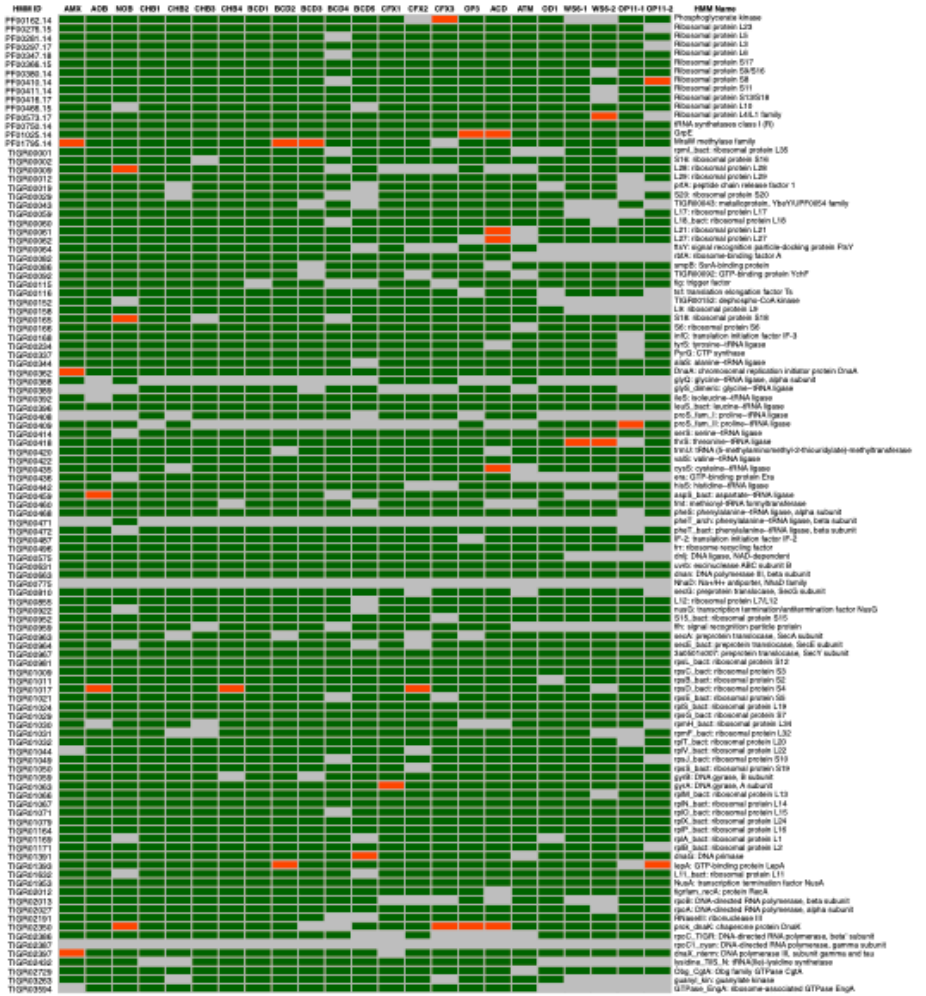
Example of marker gene presence/absence

Presence (green)
Absence (grey)
Redundancy (red)

Caveats:
- Not all genes present in all phyla

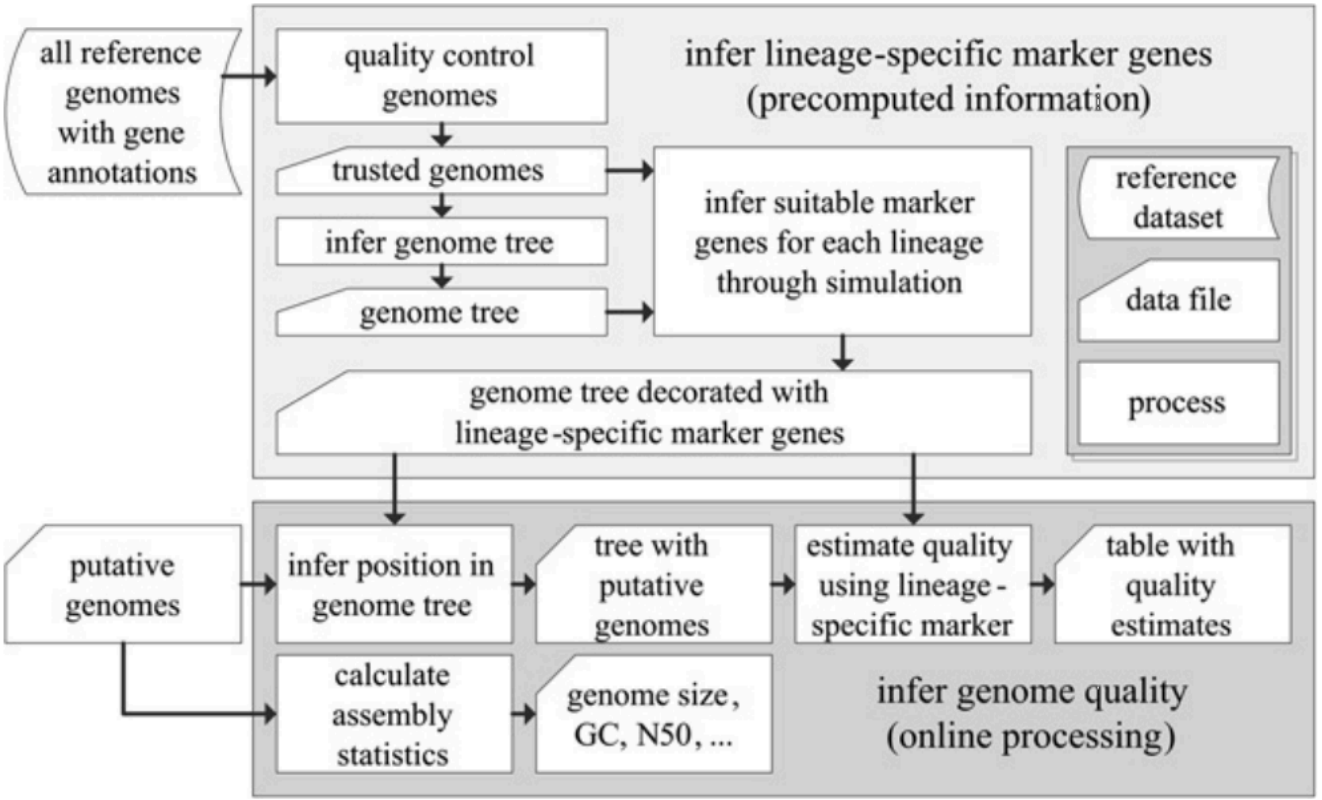
- Redundancy =/= contamination

Speth et al. 2016



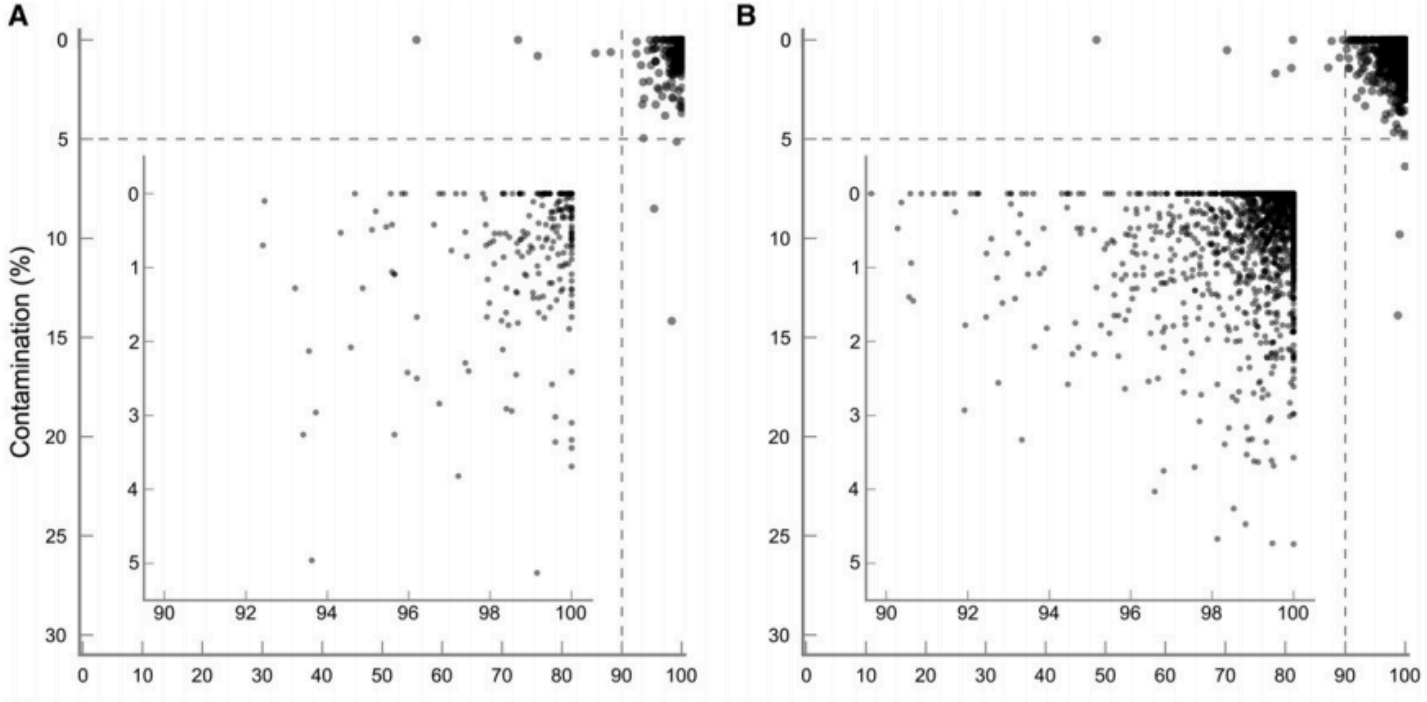
Quality assessment: CheckM (Parks et al 2015)

Overview:



Quality assessment: CheckM (Parks et al 2015)

CheckM on isolate genomes



Community standards for bin quality

- GSC: Bowers *et al*. 2017

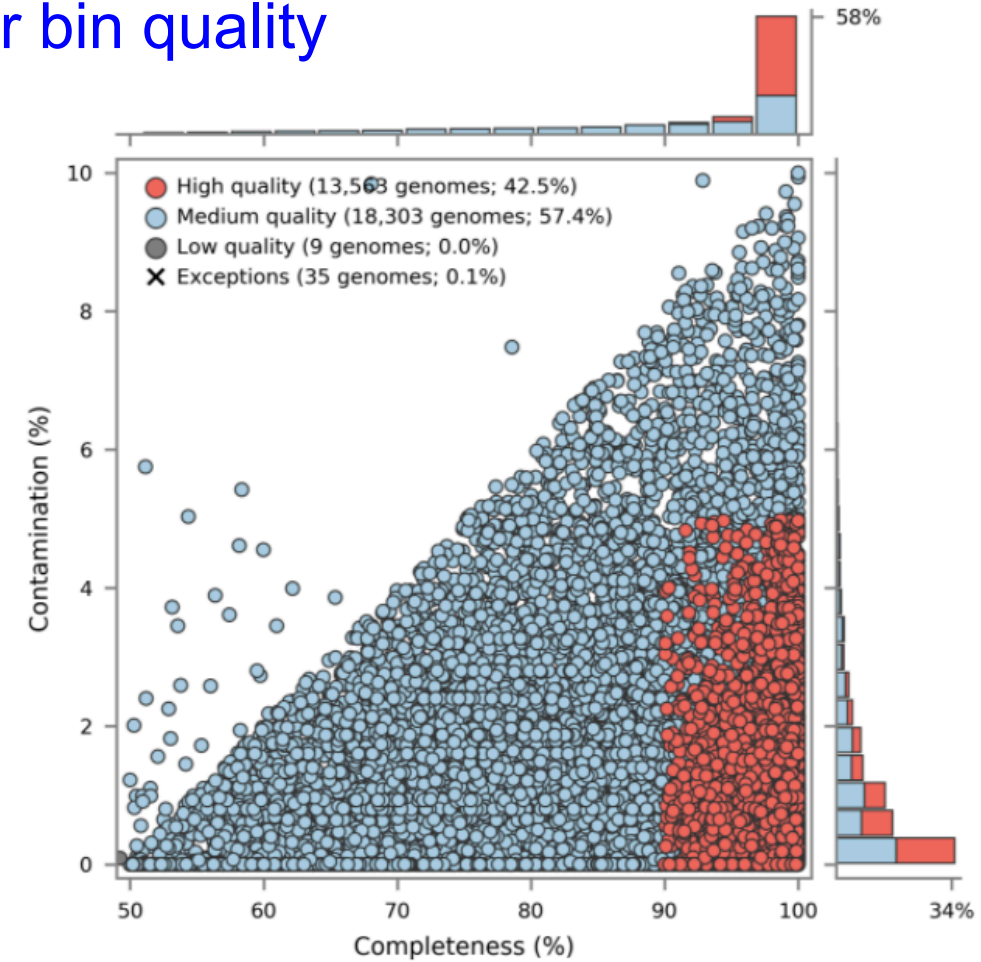
  - High:
  ≥90% complete, <5% contaminated
  5S, 16S, & 23S rRNA genes present
  ≥18 different tRNA genes present

  - Medium
  ≥50% complete, <10% contaminated

  - Low
  ≥50% complete, <10% contaminated

- GTDB

https://gtdb.ecogenomic.org/stats



Limitations of marker genes

Marker genes are not homogenously distributed

There's always exceptions to "universal" or "single copy"

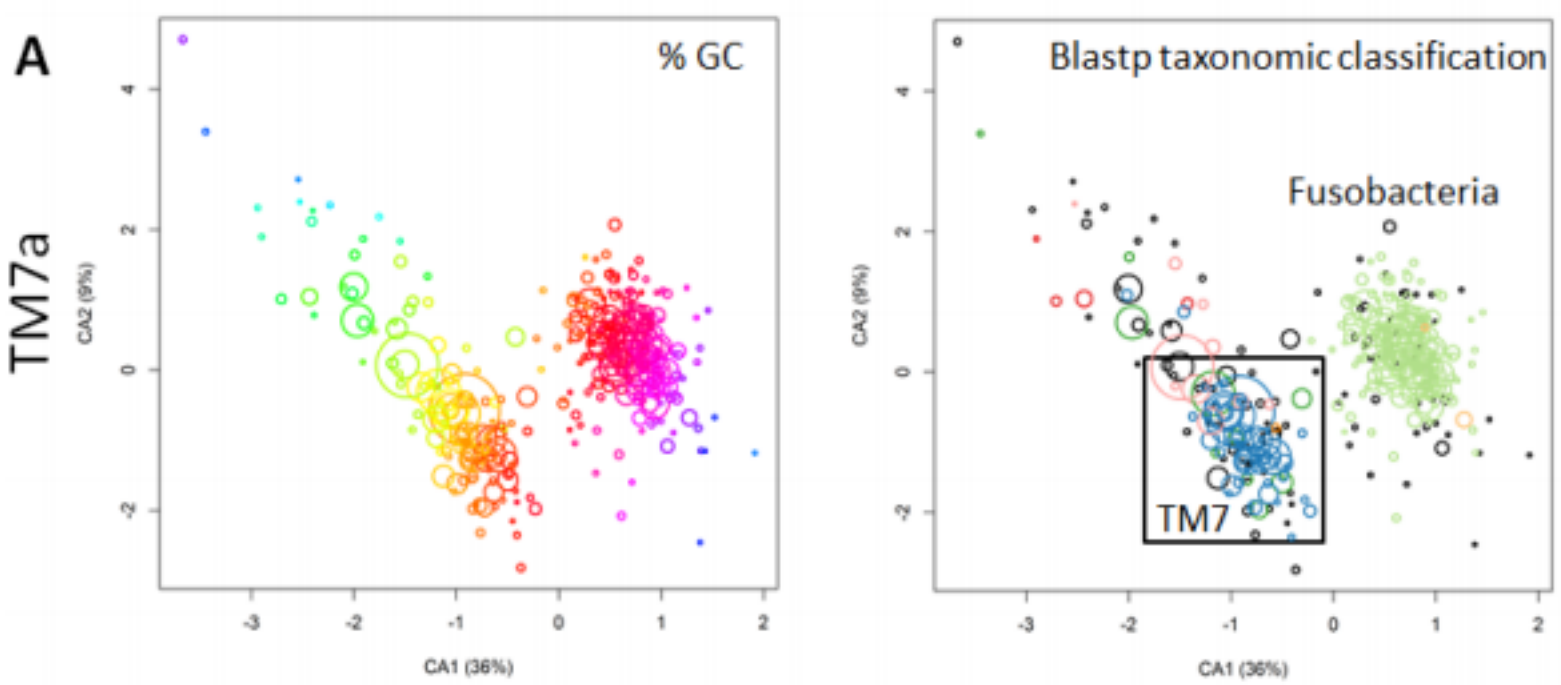
Marker genes are by definition part of the core genome, so completeness estimates miss any pan-genome content

CheckM contains some tools to evaluate bins beyond marker genes

Beyond marker genes

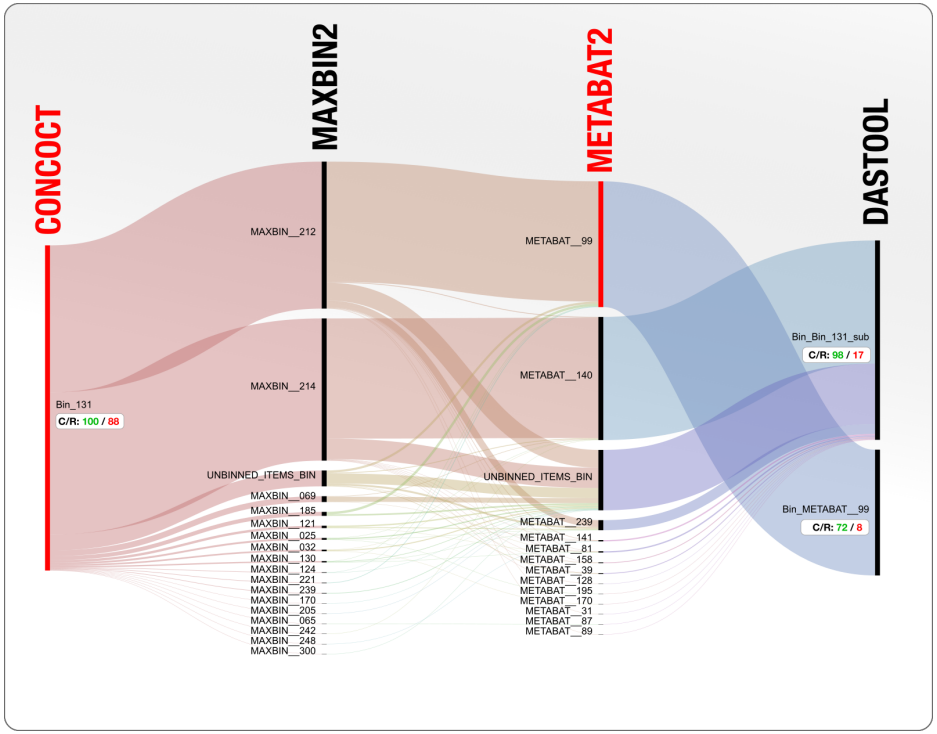
```
Reference distribution plots:
  gc_plot       -> Create GC histogram and delta-GC plot
  coding_plot   -> Create coding density (CD) histogram and delta-CD plot
  tetra_plot    -> Create tetranucleotide distance (TD) histogram and delta-TD plot
  dist_plot     -> Create image with GC, CD, and TD distribution plots together

General plots:
  nx_plot       -> Create Nx-plots
  len_hist      -> Sequence length histogram
  marker_plot   -> Plot position of marker genes on sequences
  gc_bias_plot -> Plot bin coverage as a function of GC

Bin exploration and modification:
  unique        -> Ensure no sequences are assigned to multiple bins
  merge         -> Identify bins with complementary sets of marker genes
  outliers      -> [Experimental] Identify outlier in bins relative to reference distributions
  modify        -> [Experimental] Modify sequences in a bin

Utility functions:
  unbinned      -> Identify unbinned sequences
  coverage      -> Calculate coverage of sequences
  tetra         -> Calculate tetranucleotide signature of sequences
  profile       -> Calculate percentage of reads mapped to each bin
  ssu_finder    -> Identify SSU (16S/18S) rRNAs in sequences
```

Beyond marker genes

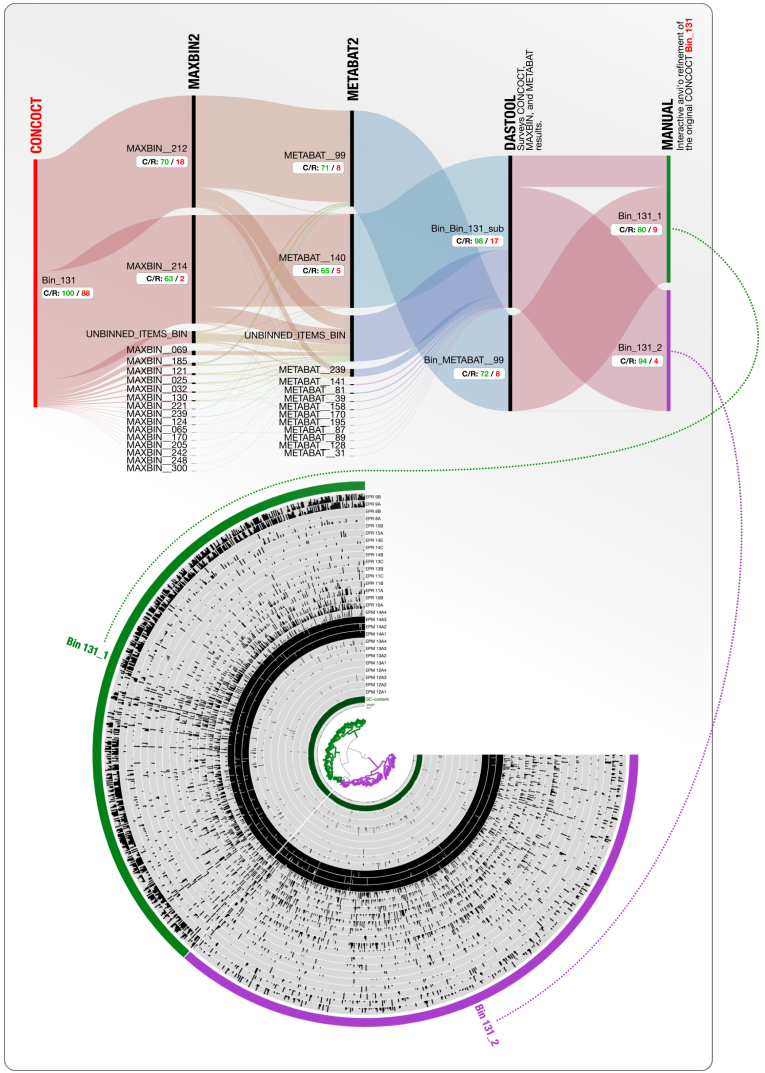


Albertsen *et al.* 2013

Beyond marker genes



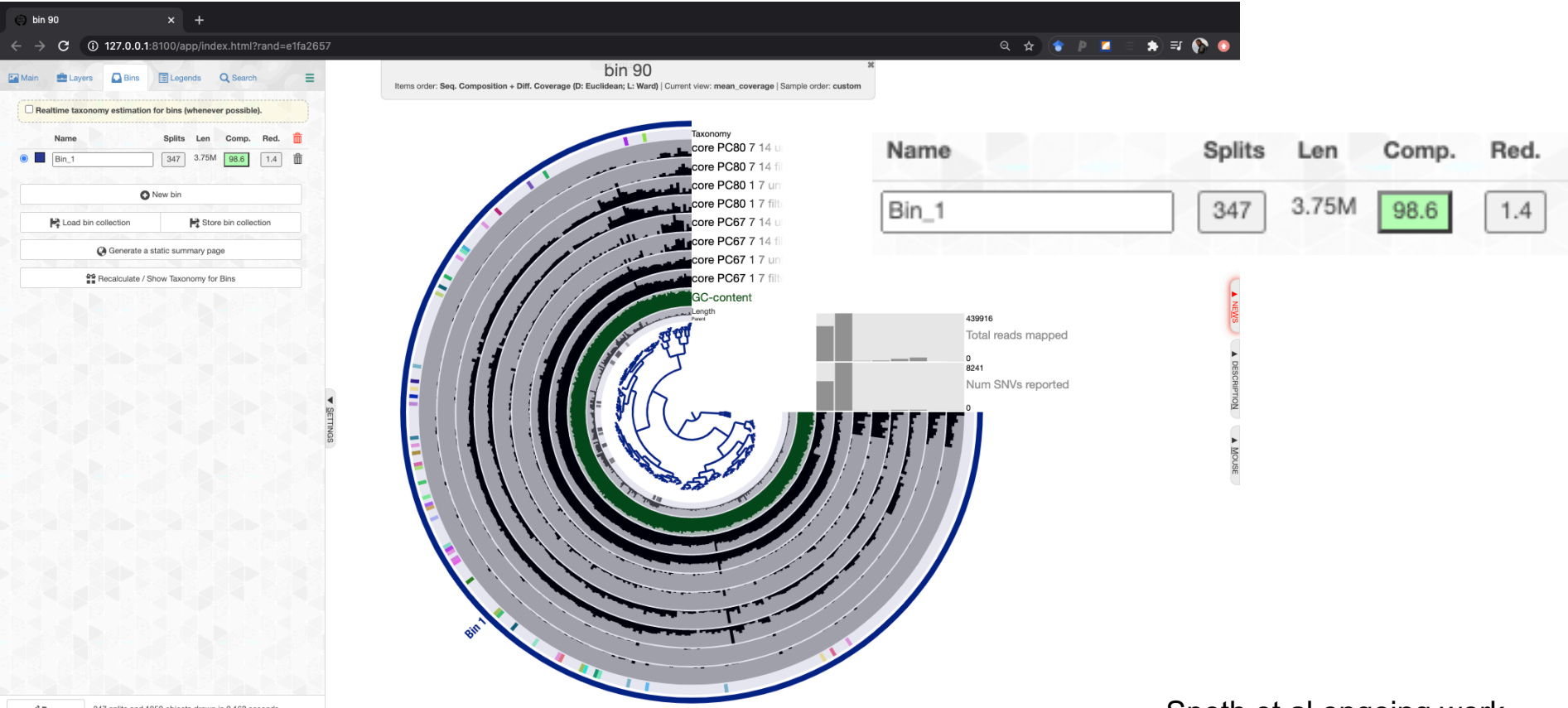
http://merenlab.org/2020/01/02/visualizing-metagenomic-bins/



Beyond marker genes

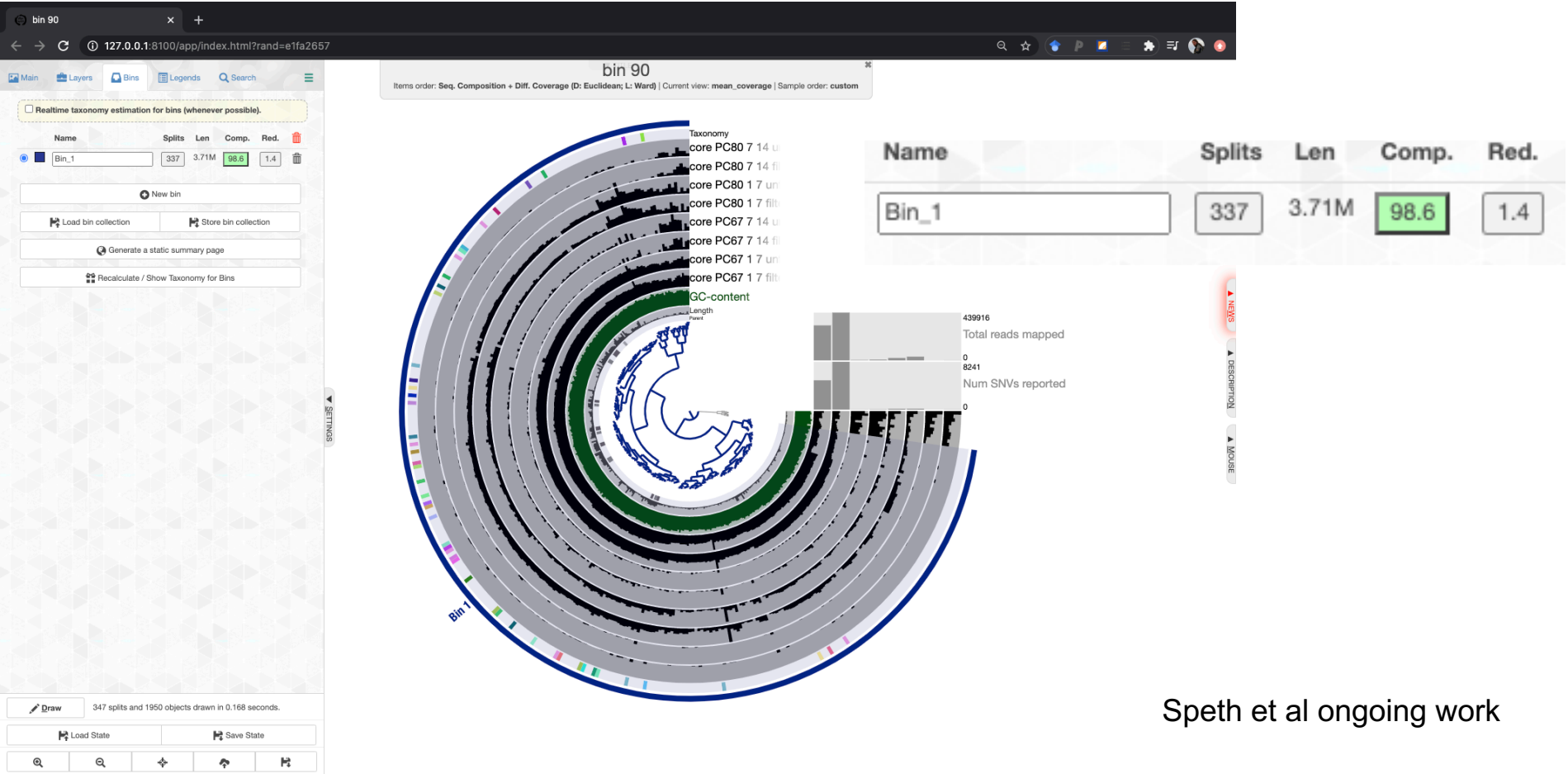
A little exercise
providing no answers
but hopefully makes you think

Beyond marker genes



Speth et al ongoing work

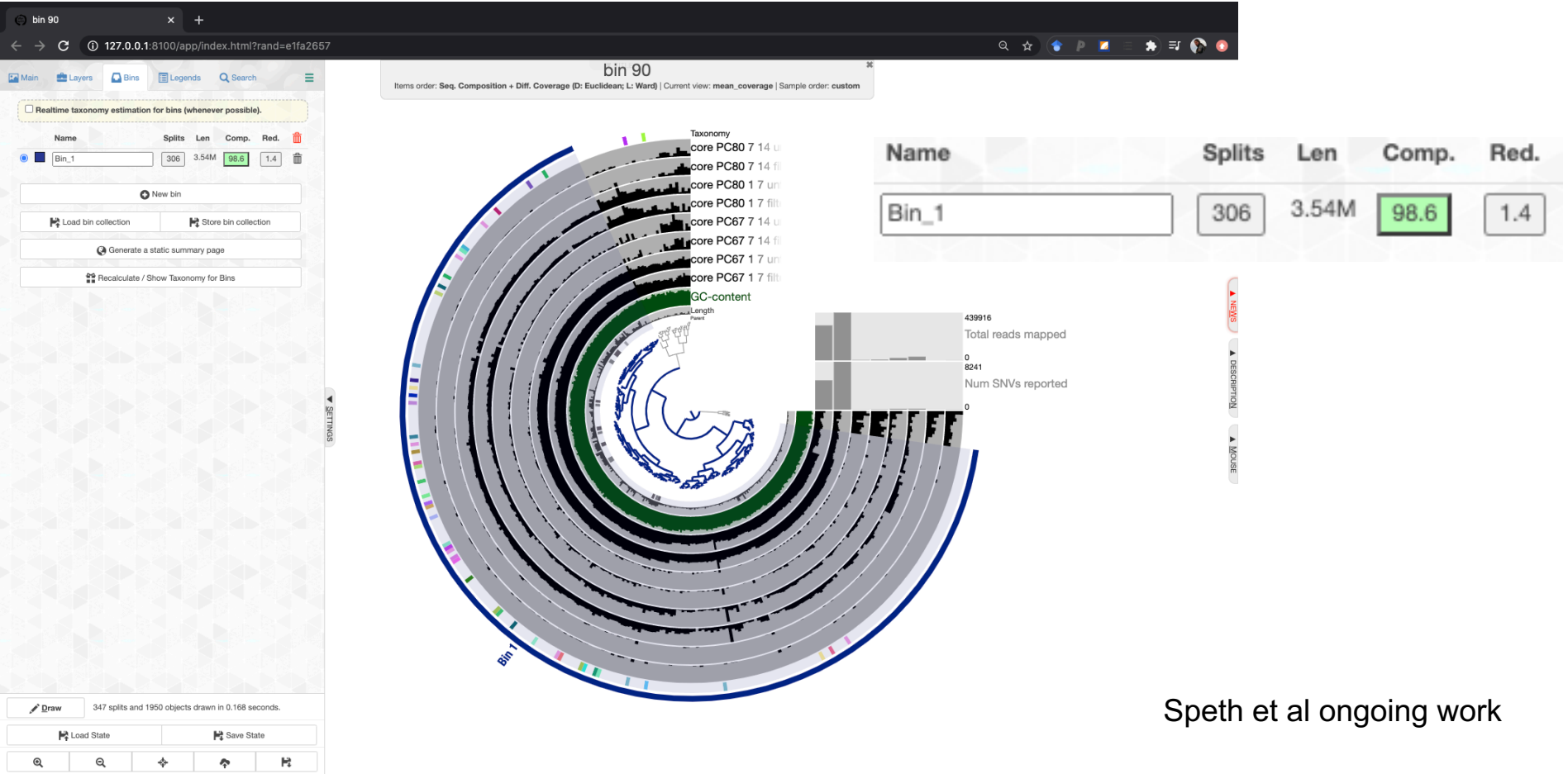
Beyond marker genes



Name	Splits	Len	Comp.	Red.
Bin_1	337	3.71M	98.6	1.4

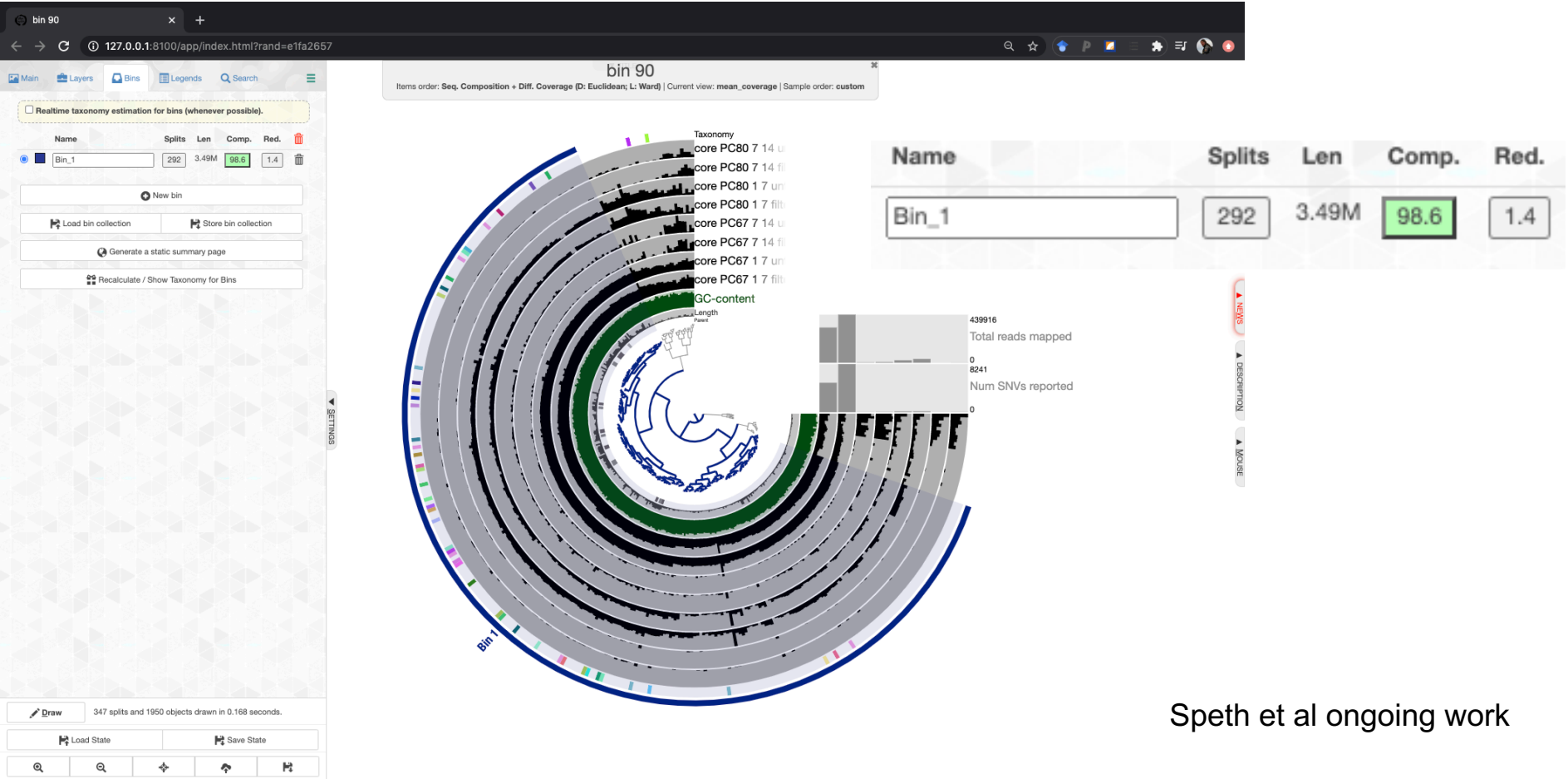
Speth et al ongoing work

Beyond marker genes



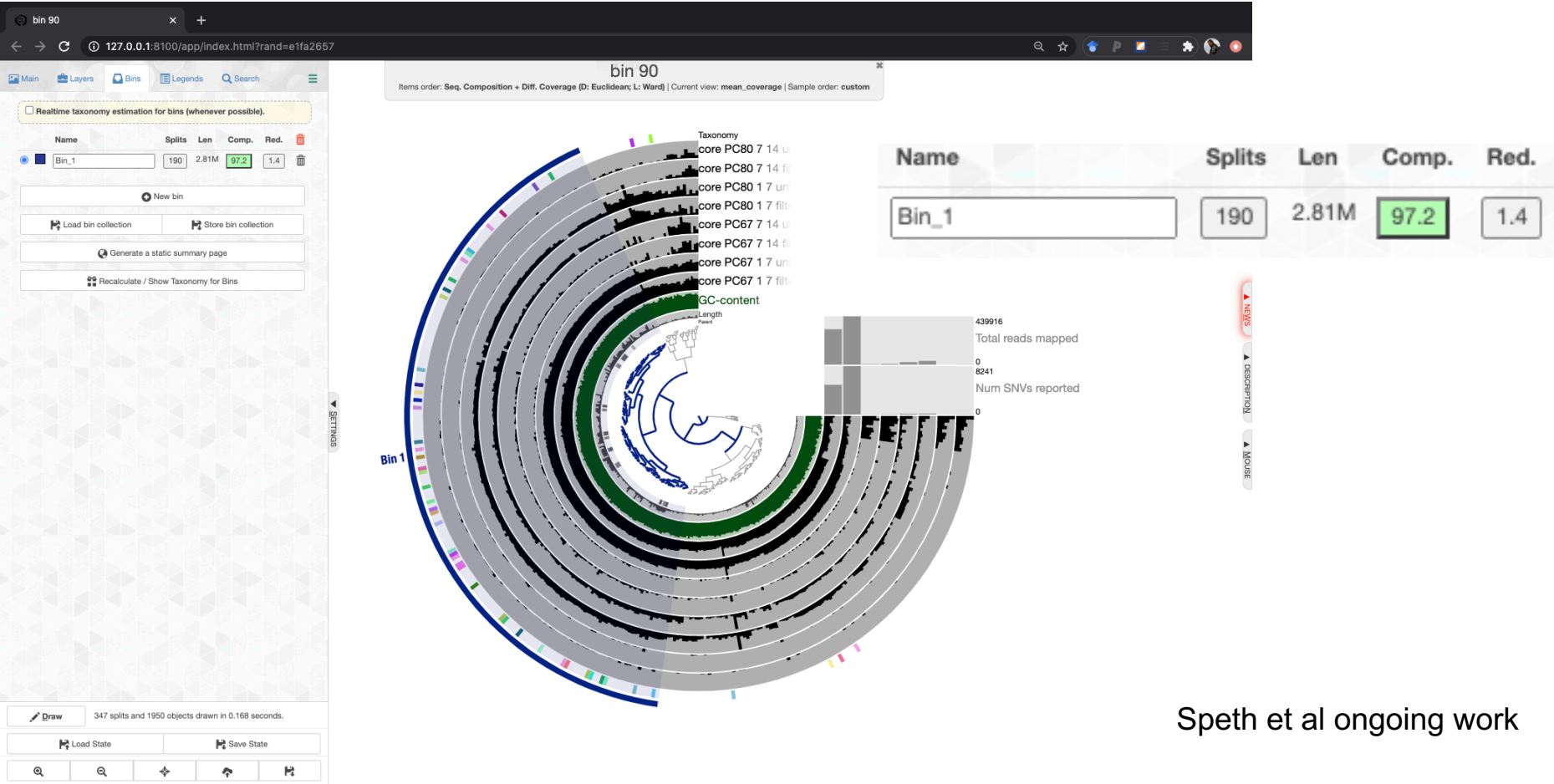
Speth et al ongoing work

Beyond marker genes



Speth et al ongoing work

Beyond marker genes



Name	Splits	Len	Comp.	Red.
Bin_1	190	2.81M	97.2	1.4

Speth et al ongoing work

Beyond marker genes

- We progressively dropped the number of splits in the bin from 347 to 292 (16% reduction), and the number of bases from 3.75Mb to 3.49Mb (7% reduction, ~250 genes) without changing completeness or contamination
- We then, in a giant step, dropped the number of splits from 292 to 190 (35% reduction), and the number of bases from 3.49Mb to 2.81Mb (19% reduction, ~700 genes) and only changed lowered completeness 1.4%.
- The difference between the start and finish is 45% of the splits, and 25% of the bases, corresponding to 1000 genes…
- What is the right answer here?

Take home messages

- Genome bins/MAGs/SAGs live in databases long after publication and **will** be used in comparative studies.
- It is our **responsibility** to the community to curate the data we submit to databases as well as we reasonably can.
- Available tools are great, but there is **no magic bullet** for bin quality assessment.

Questions?

Next up:
hands-on CheckM tutorial
https://github.com/biovcnet/bvcn-binder-checkm