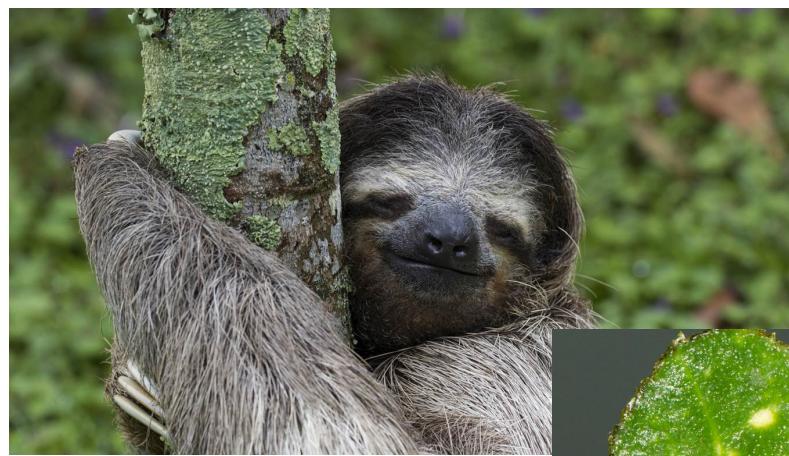


Introduction to Metagenomic Binning

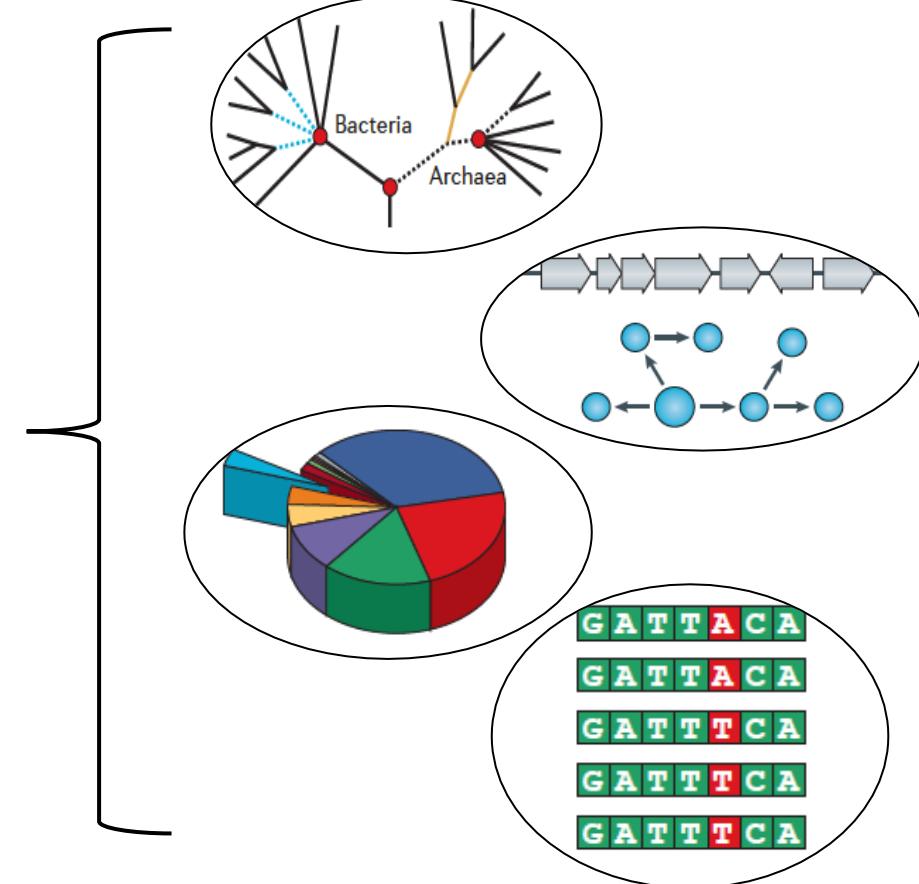
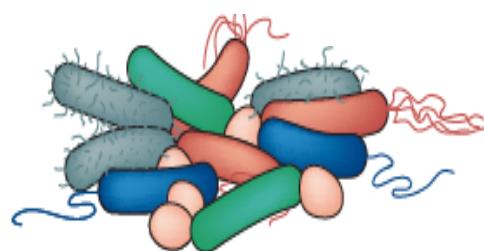
Elizabeth McDaniel
[@lizilla93](https://twitter.com/lizilla93)



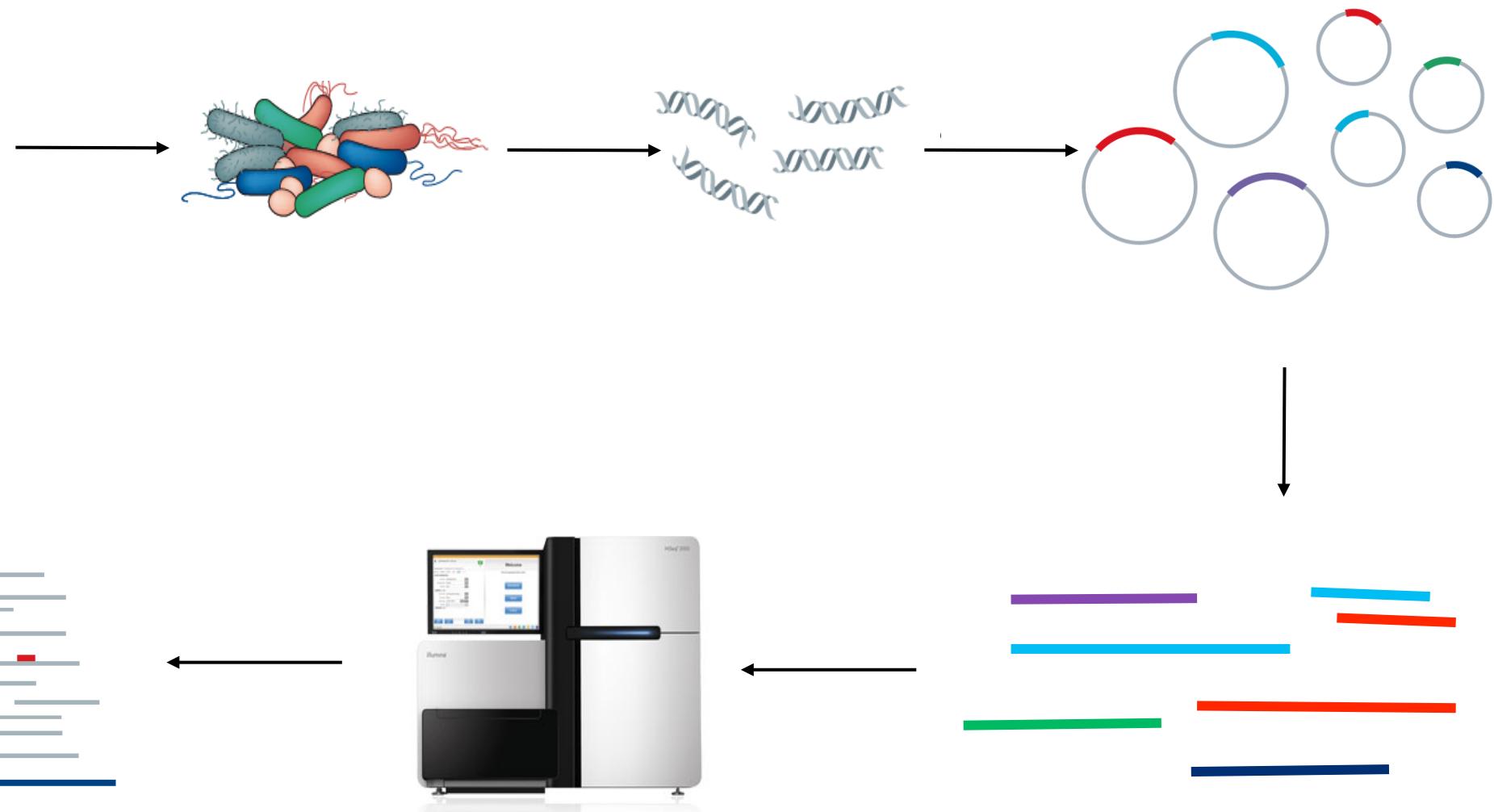
Microbial Ecology



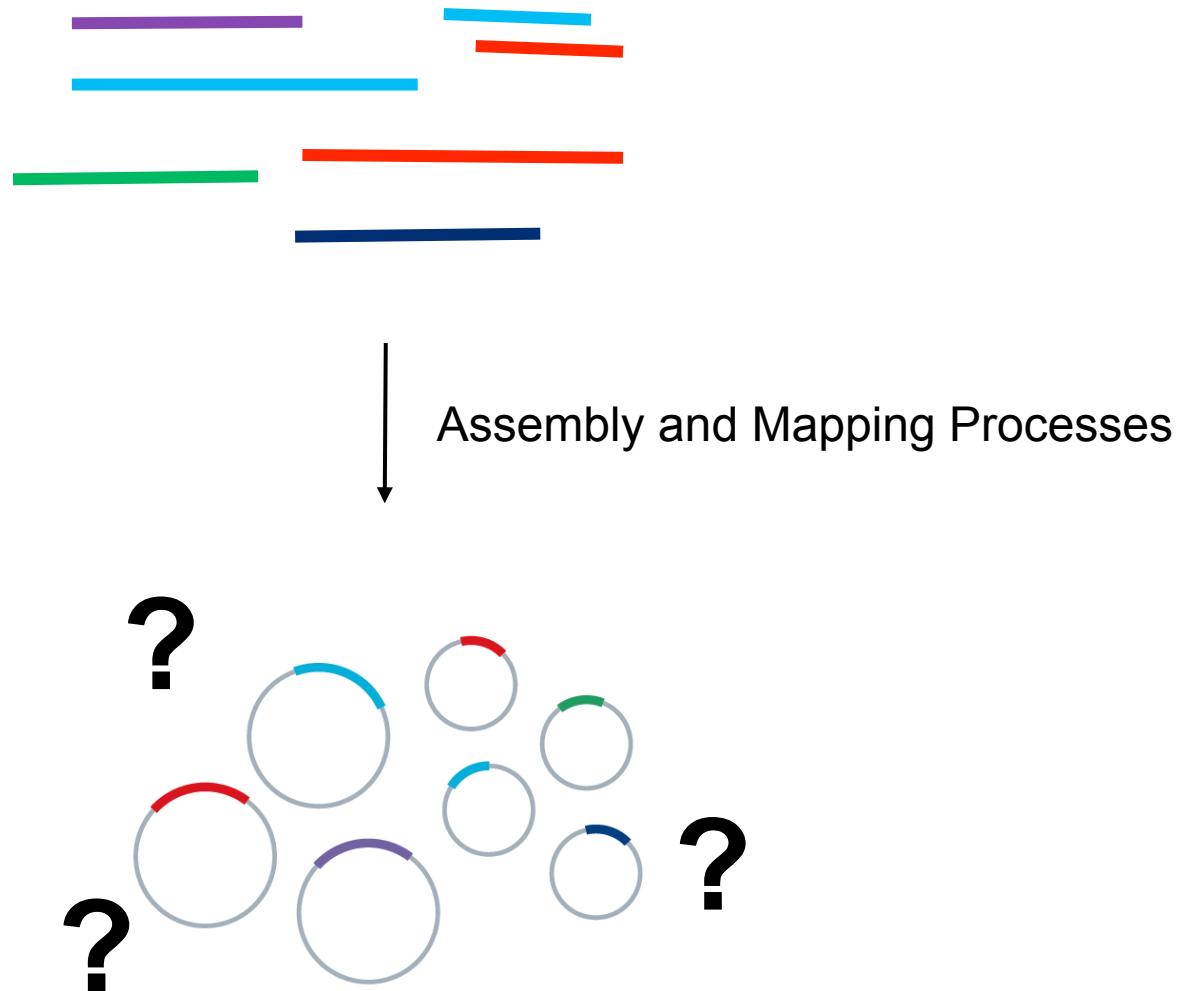
Microbial Ecology



Metagenome Assembled Genomes



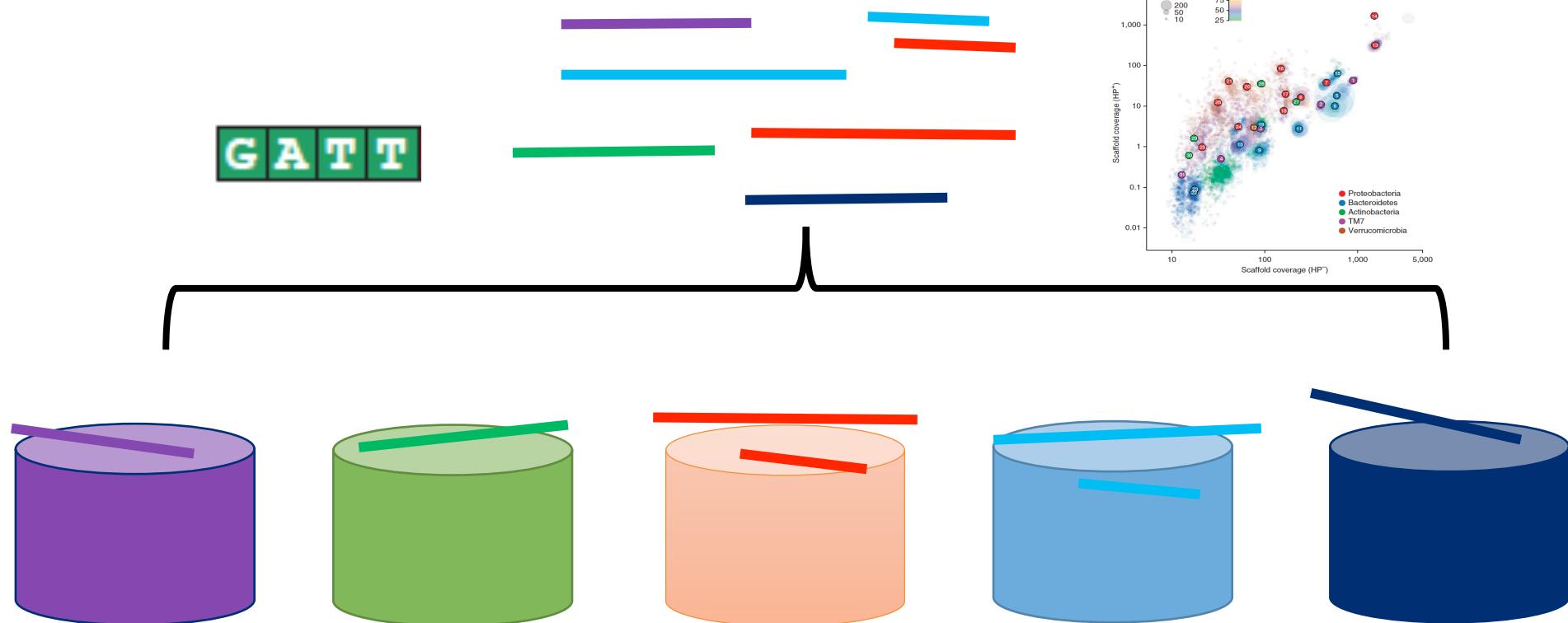
Metagenome Assembled Genomes



Binning

Place the sequence in the correct “bin” or OTU

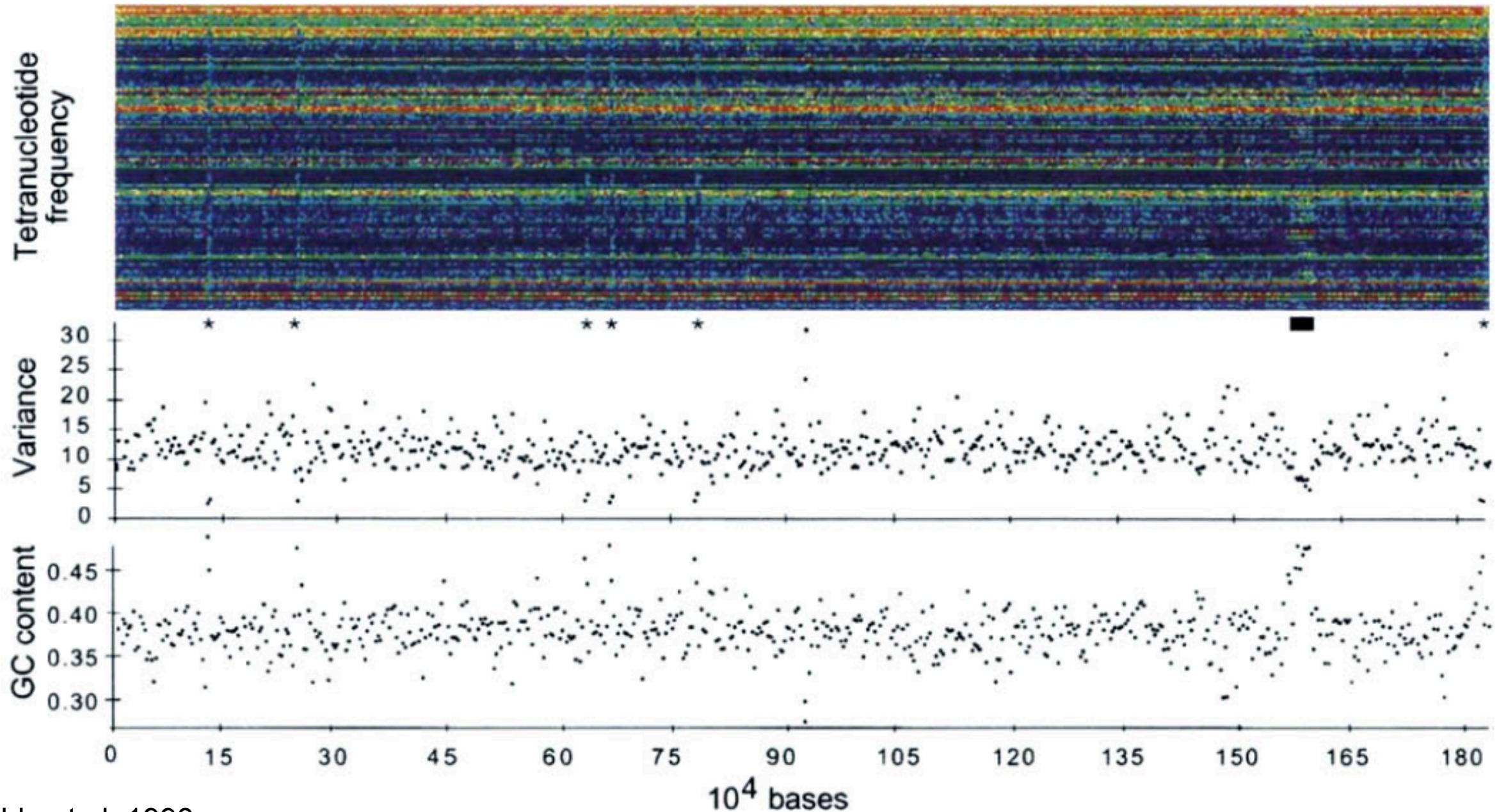
- Composition based
- Differential Coverage based



Composition Based Binning

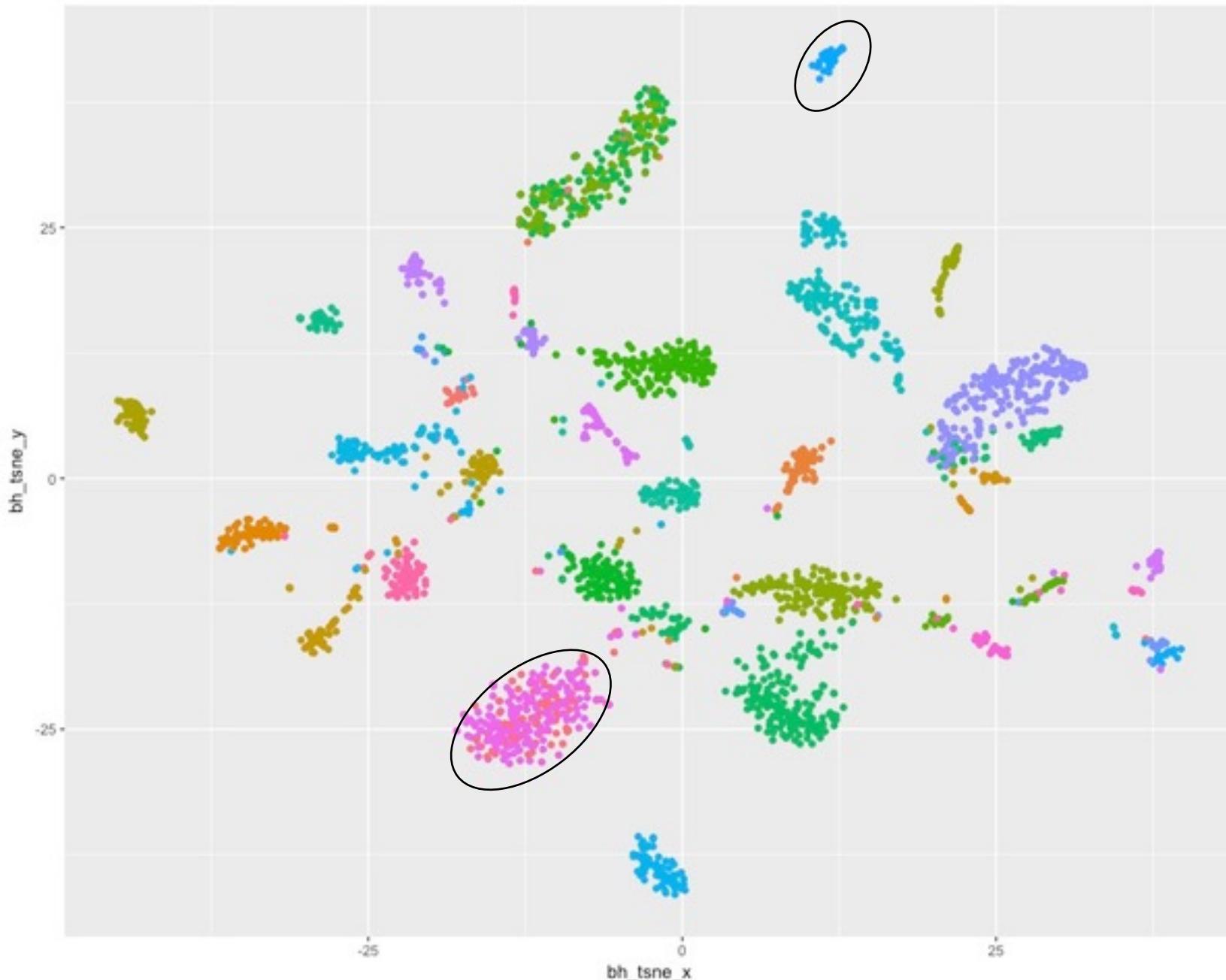
- Sequence signature k-mer (di-nucleotide, tetra-nucleotide) throughout the entire genome
- Non-random
- Non-linearly related to GC content
- Tetranucleotide frequency (TNF) used to bin short reads to OTUs



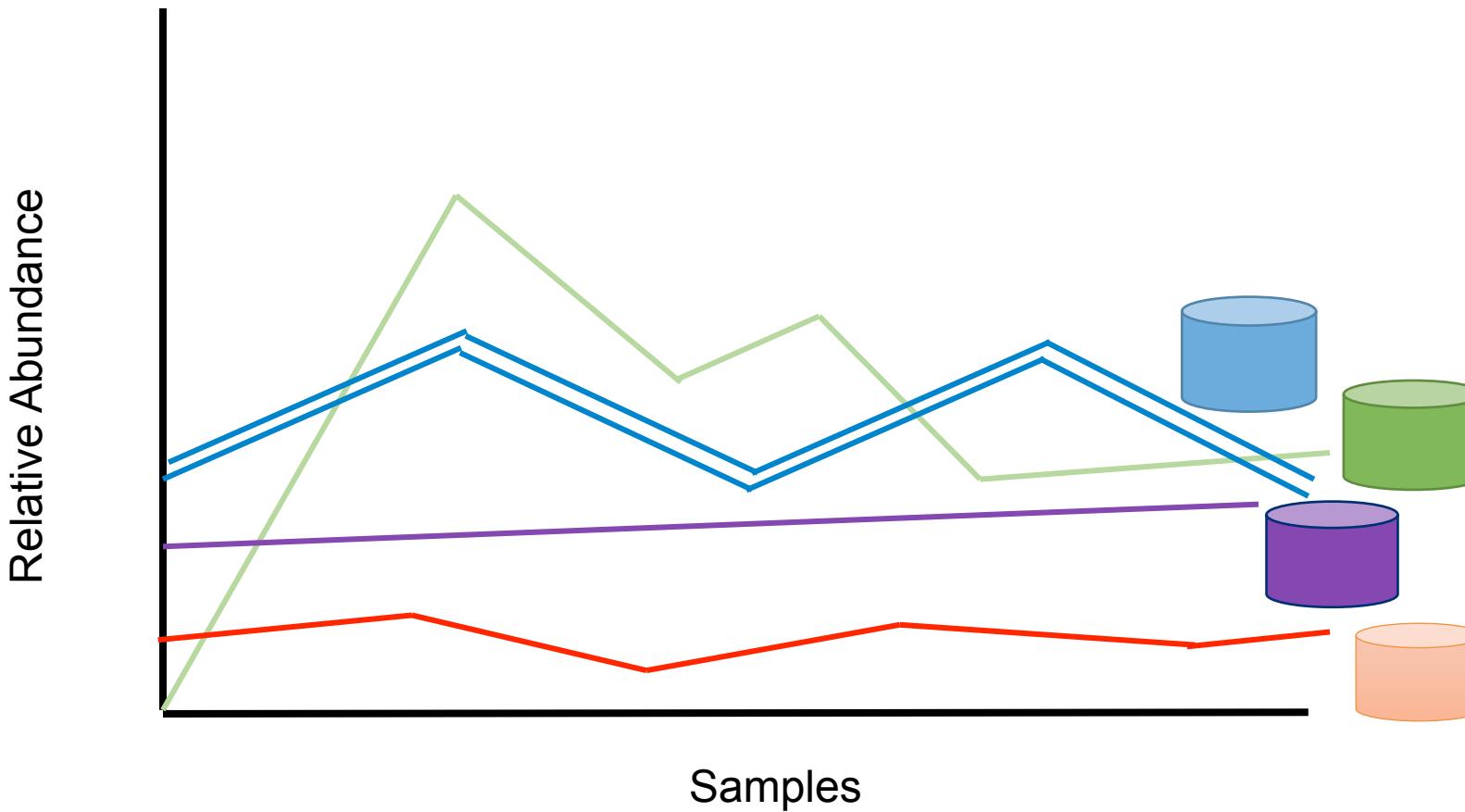


Noble et al. 1998





Differential Coverage Based Binning



A Really, Really Crude Way to Think About This:

[https://www.youtube.com/watch?
v=DxbNRiH3ghc](https://www.youtube.com/watch?v=DxbNRiH3ghc)



Binning Algorithms

Different Algorithms take different parameters into account:

- **CONCOCT**: Sequence composition and coverage (Alneberg et al 2013)
- Ray Meta: k-mer frequencies (Boisvert et al. 2012)
- MetaVelvet: extended from single genome assembler Velvet – k-mer connectivity and then coverage difference
- Albertson et al (unnamed binning workflow): differential abundance and then k-mer frequencies
- SPAdes: considers *pairs* of k-mers (*k-bimers*), developed for single cell sequencing (Bankevich 2012)



Completeness vs. Redundancy

Completeness =
$$\frac{\text{\# of single copy genes found}}{\text{\# of single copy genes searched for}}$$

Redundancy =
$$\frac{\text{\# of single copy genes found more than once}}{\text{\# of single copy genes searched for}}$$



Completeness vs. Redundancy

Completeness = $\frac{100}{\text{complete}}$ = ~ 72 %

139

Redundancy = $\frac{5}{139}$ = ~ 3.6 %

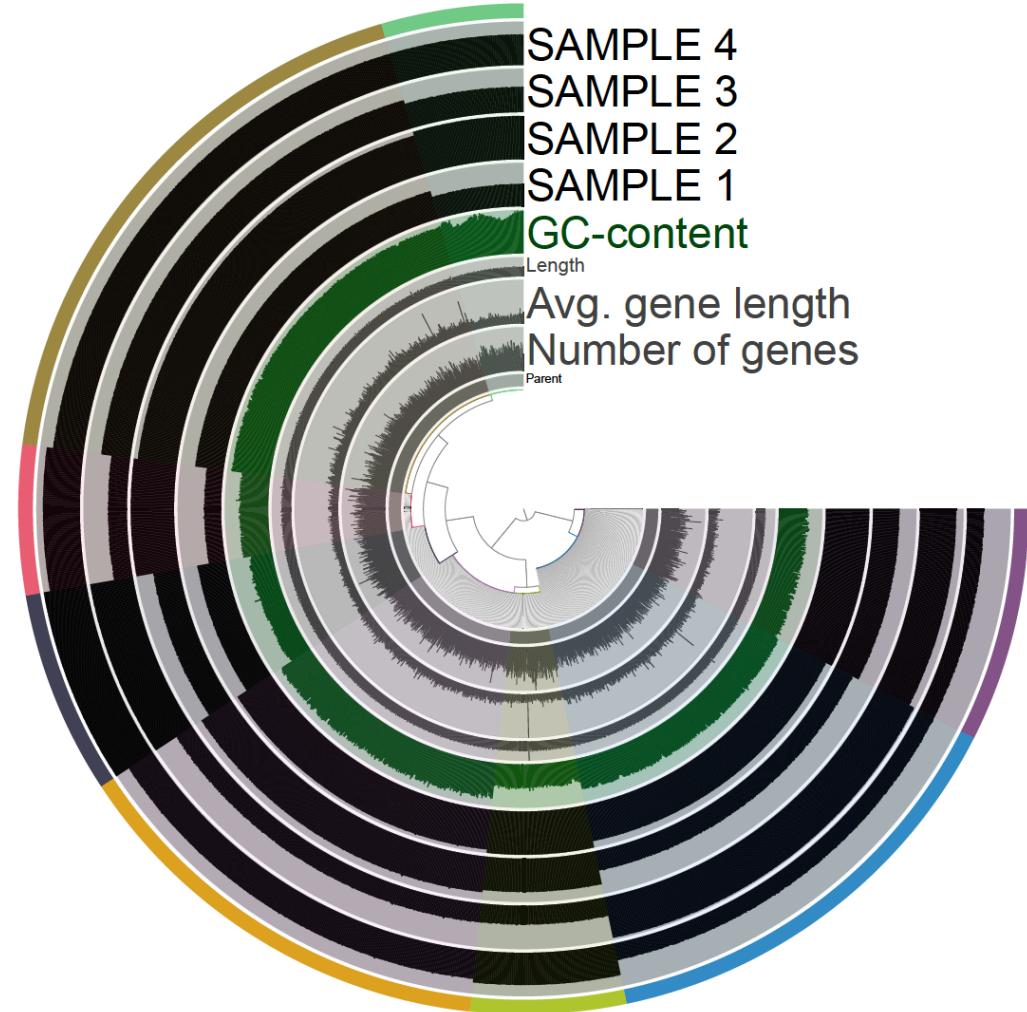
139





Completeness/contamination estimates by manually curating your bins

	% comp. / % cont.	Default Bacterial HMM Profiles	Rinke et al. Archaeal HMM Profile
Archaea	<i>Acidilobus saccharovorans</i>	63% / 6%	96% / 4%
	<i>Archaeoglobus profundus</i>	67% / 7%	99% / 5%
	<i>Acidianus hospitalis</i>	63% / 5%	95% / 4%
	<i>Nitrosopumilus maritimus</i>	67% / 5%	93% / 8%
	<i>Methanobacterium formicicum</i>	66% / 7%	98% / 7%
Bacteria	<i>Planctomyces brasiliensis</i>	97% / 4%	50% / 6%
	<i>Escherichia coli</i>	99% / 4%	49% / 5%
	<i>Alteromonas macleodii</i>	99% / 6%	51% / 7%



Single Copy Core Genes

- To estimate completeness, the contigs are scanned using previously published bacterial/archaeal single-copy collections

```
$ anvi-run-hmms -c contigs.db
```

- Running this script utilizes bacterial/archaeal HMM profiles in the codebase (Campbell et al 2013, Rinke et al. 2013)
- From this point, you can refine your bins within anvi-summarize and anvi-refine to add a “human guided” aspect to your binning workflow



Questions?

