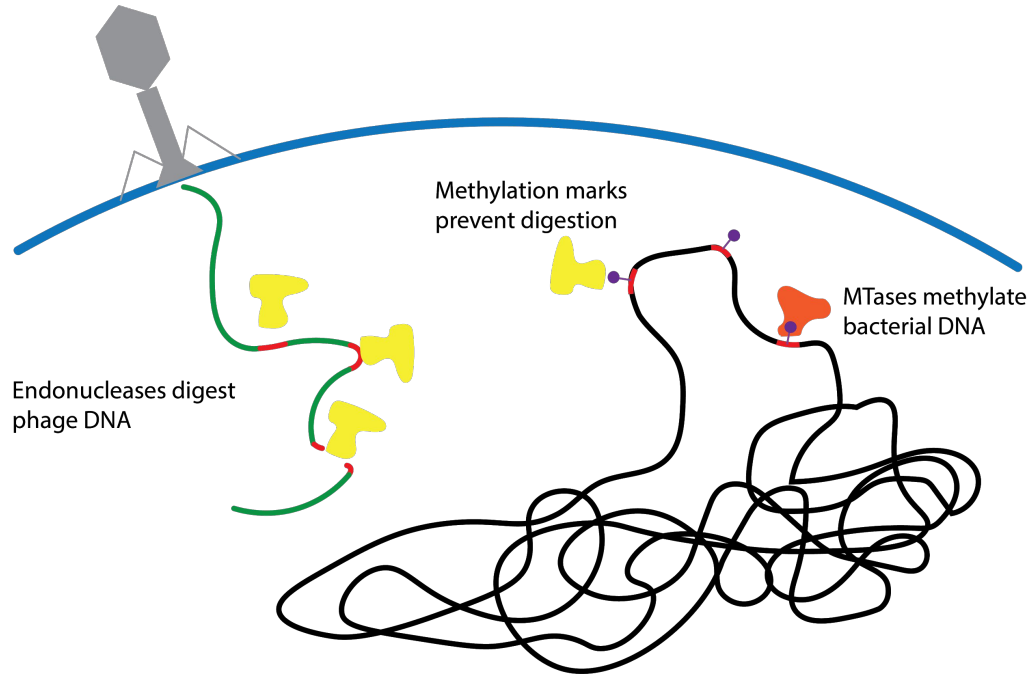


# Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation

John Beaulaurier, Shijia Zhu, Gintaras Deikus, Ilaria Mogno, Xue-Song Zhang, Austin Davis-Richardson, Ronald Canepa, Eric W Triplett, Jeremiah J Faith, Robert Sebra, Eric E Schadt & Gang Fang

# Bacterial Methylation

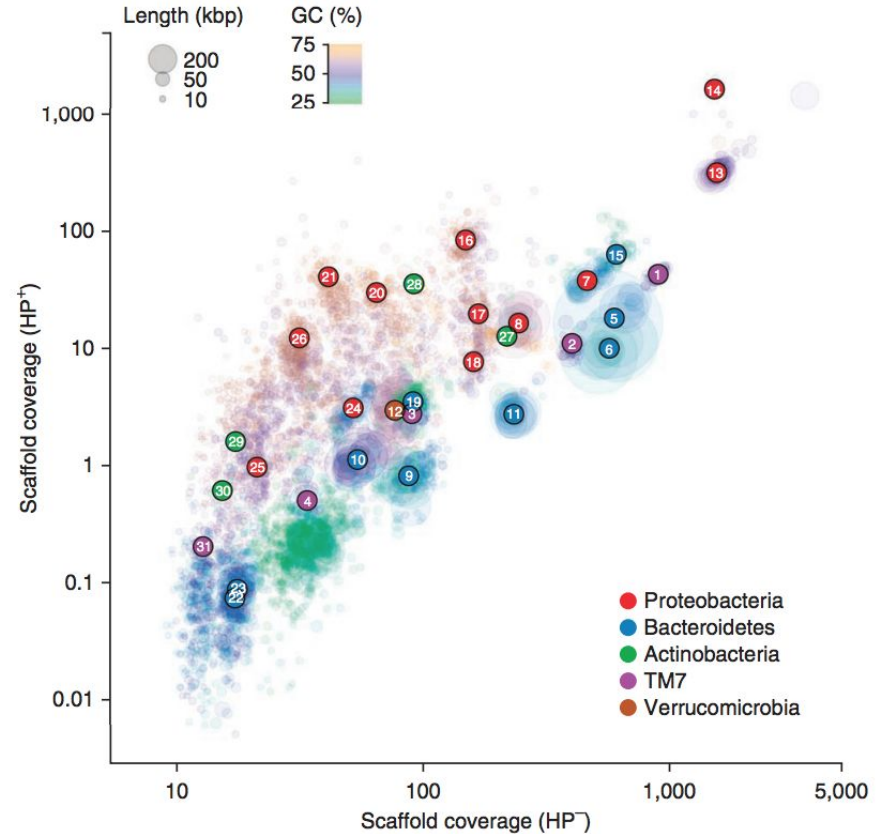


## → Restriction-Methylation System

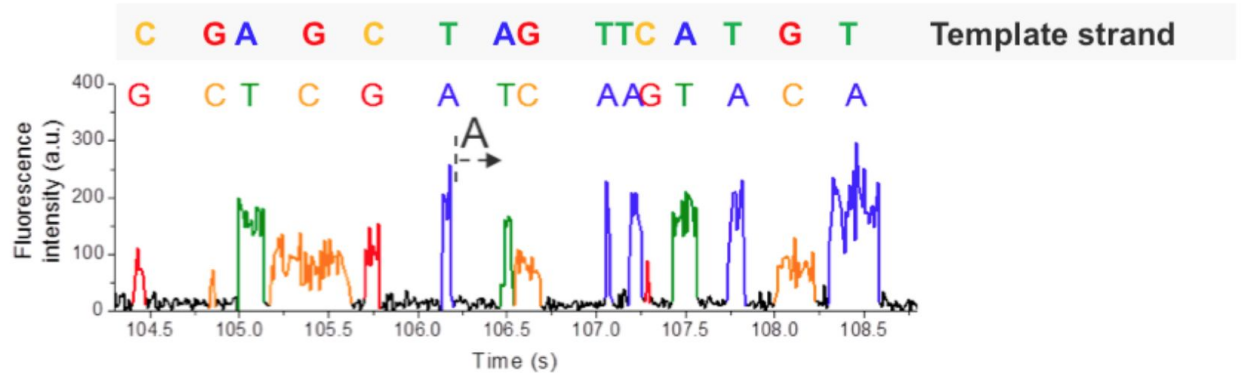
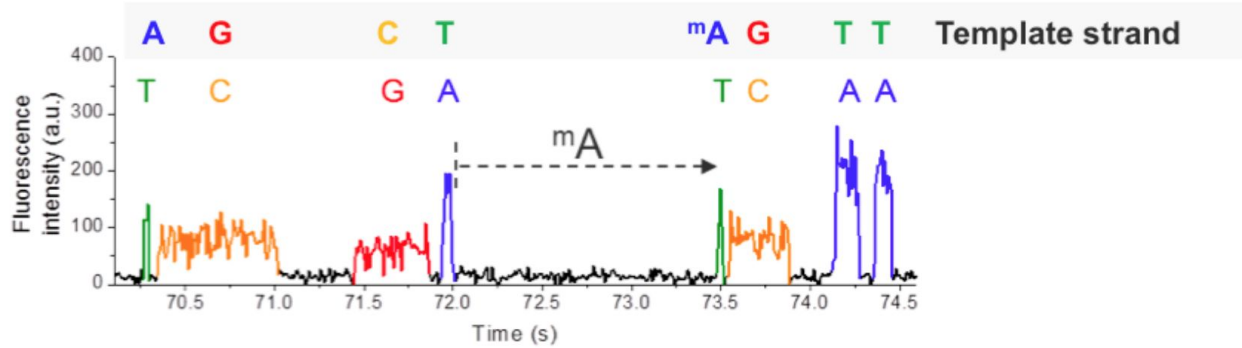
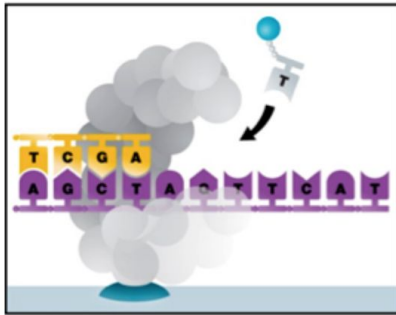
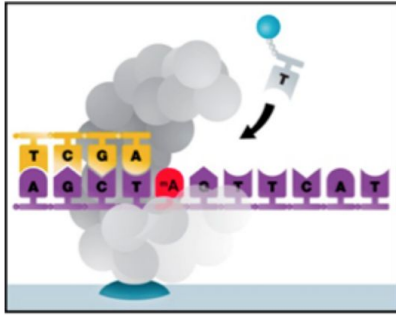
- ◆ Thought to have developed as a defense against viruses
- ◆ Pairs of methyltransferases and endonucleases bind at the same motifs
- ◆ Methyl marks prevent auto-digestion
- ◆ Variety of motifs can be methylated in bacteria, not just CG

# Metagenomic Binning

- Group contigs that look like they came from the same species
- ◆ Sequence composition
  - ◆ Differential coverage (right)
    - Likely to fail for plasmids :(
  - ◆ Methylation



# Inter-pulse duration (IPD)



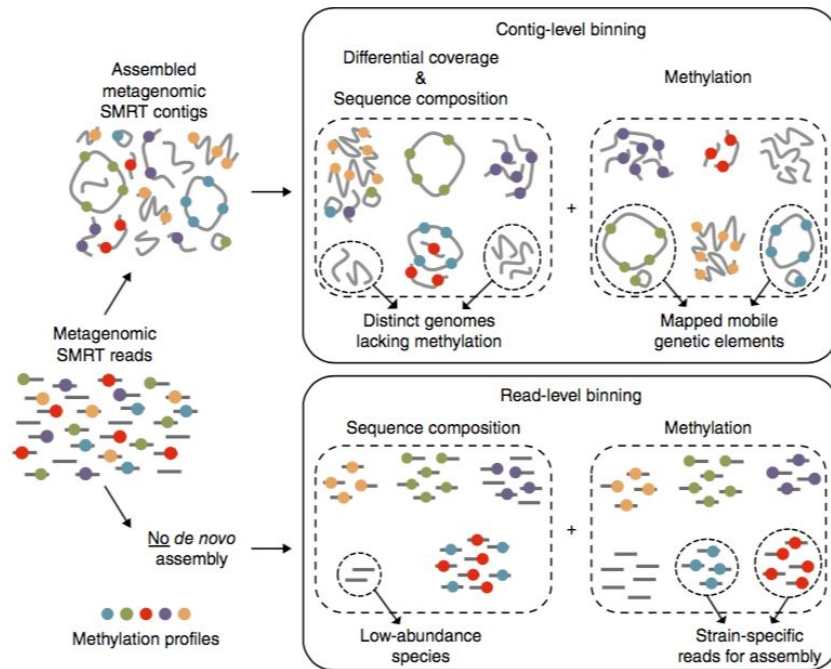
# Using Methylation

→ Tag each contig/read with a methylation profile

- ◆ Profiles are methylation scores for a set of motifs

$$R_{ij}^o = \frac{1}{\sum_{s=1}^S M_s} \sum_{s=1}^S \sum_{m=1}^{M_s} nIPD_{ms}$$

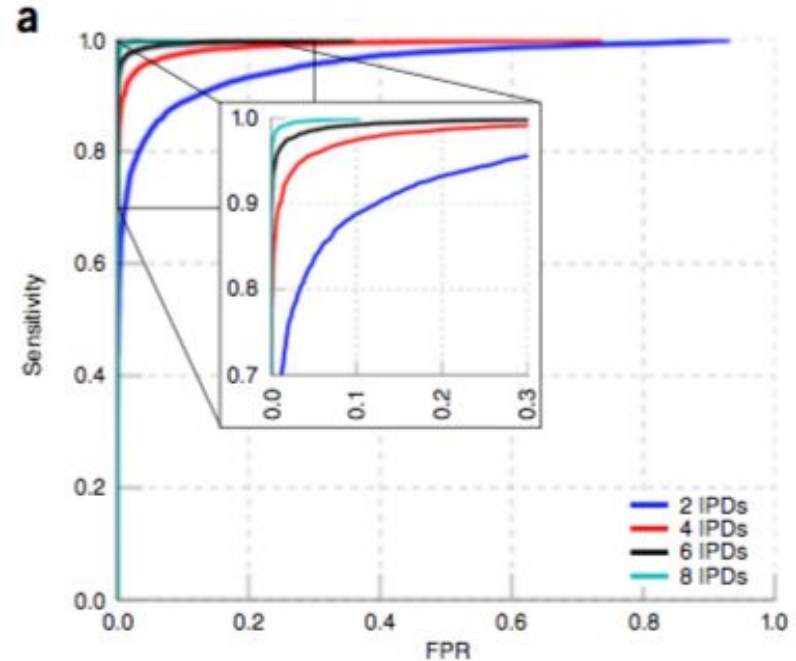
Scores are essentially a mean of IPD values at across a read/contig at the motif in question



# Contig/read classification

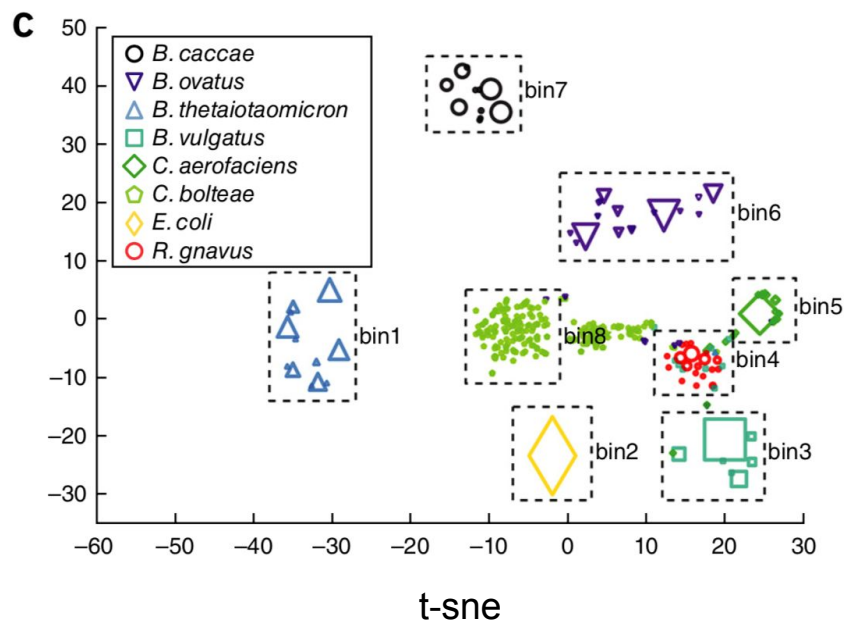
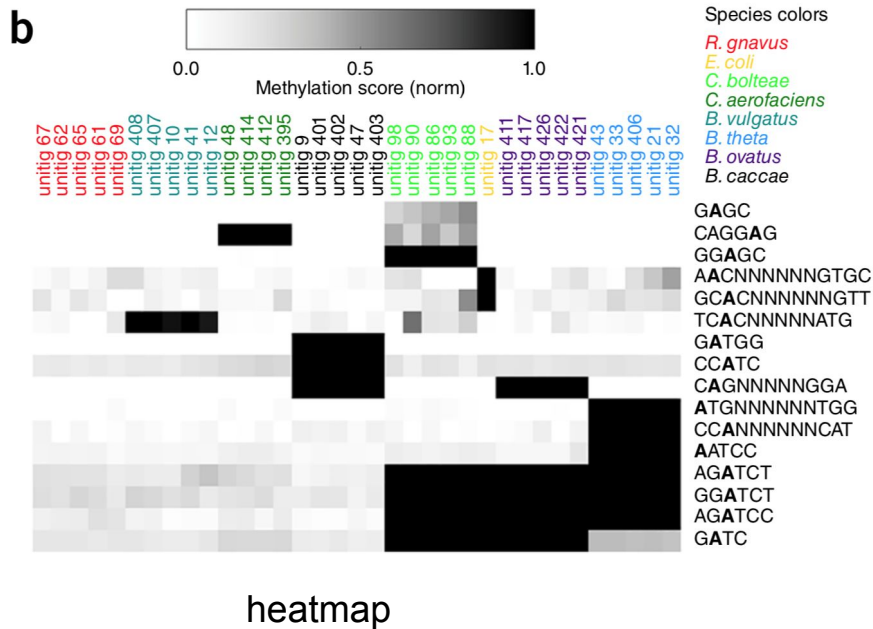
→ Classification power increases as more motifs/IPD values appear on the sequence

(a) Receiver operating characteristic (ROC) curve illustrating the power to classify a contig as methylated (N6-methyladenine, 6mA) or non-methylated regarding a specific sequence motif, as a function of the number of IPD values available for the motif sites on the contig. FPR, false-positive rate.



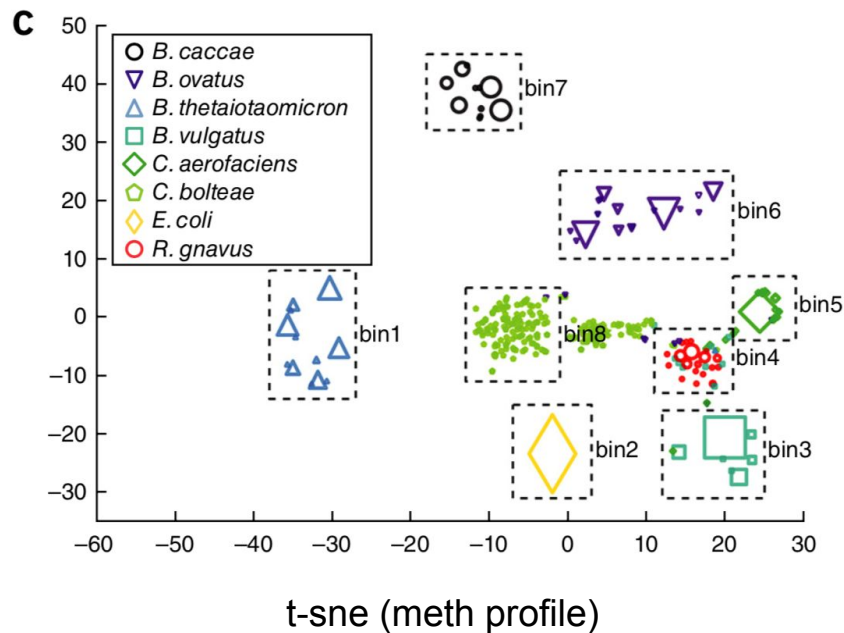
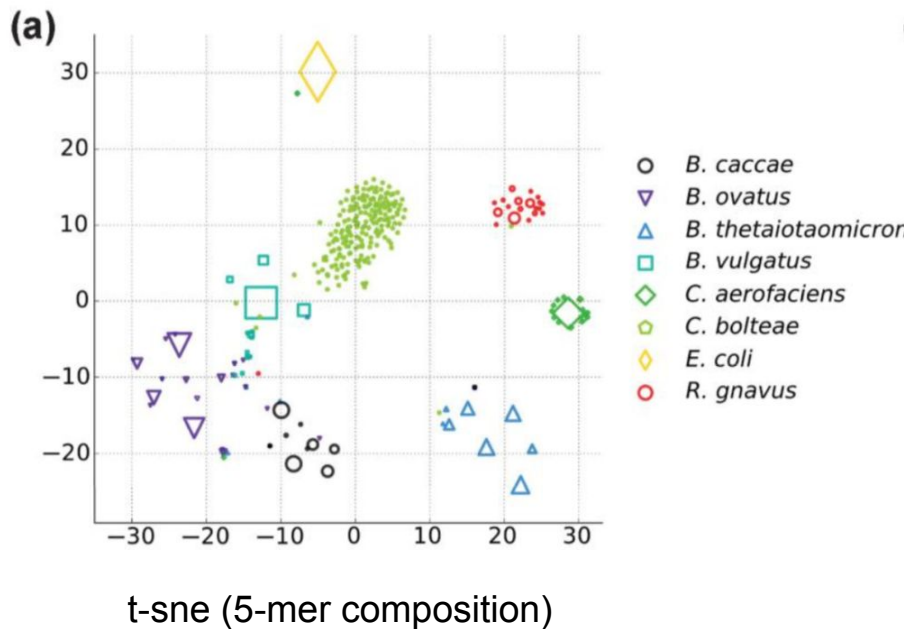
# Binning

Synthetic mixture of reads from 8 species

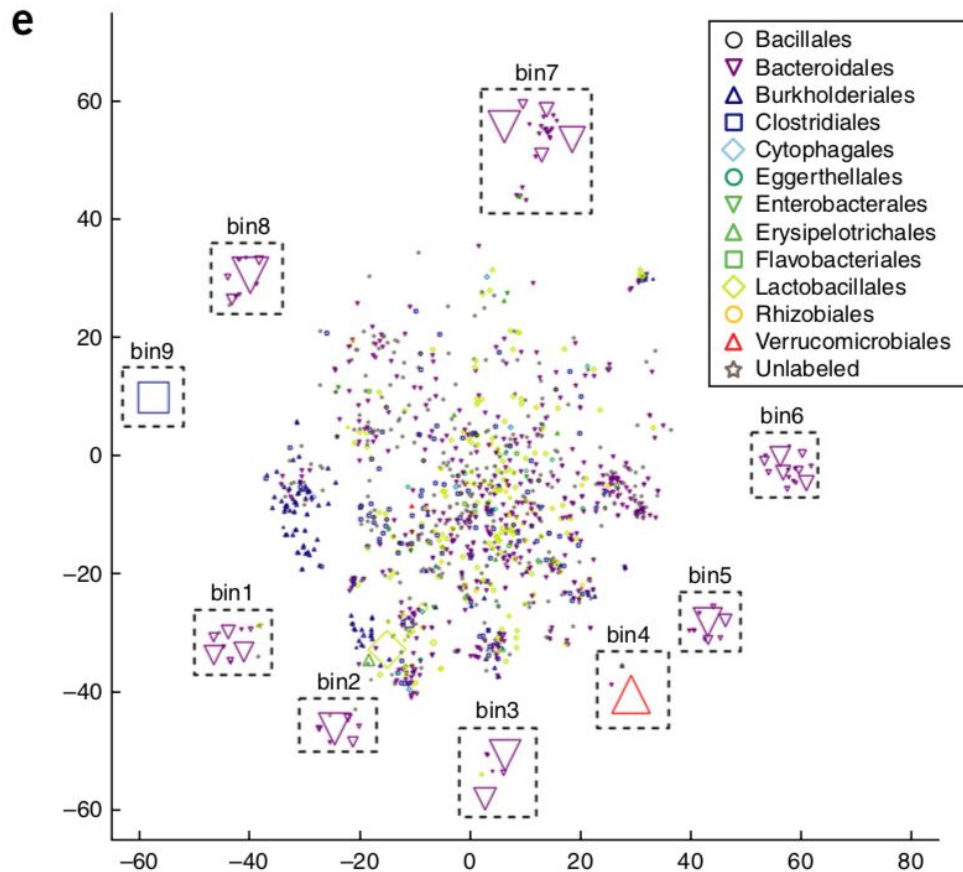


# Binning

Improved separation of *Bacteroides* genus compared to k-mer profiles







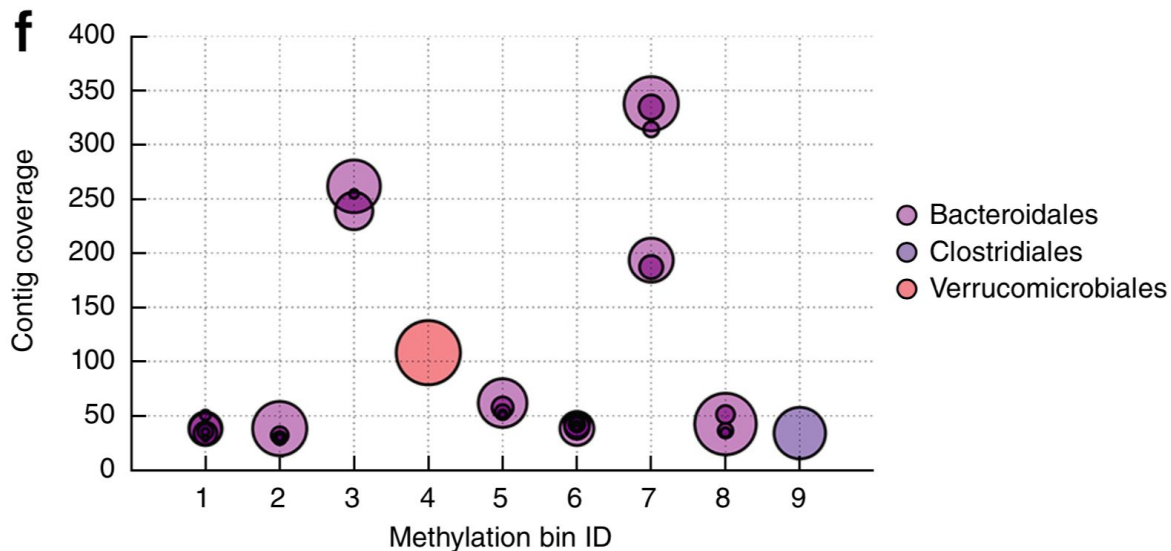
# Mouse Microbiome

(e) t-SNE projection of metagenomic contigs assembled from SMRT reads of an adult mouse gut microbiome, organized according to differing methylation profiles across 38 sequence motifs in the sample. Labeled bins denote genome-scale assemblies with distinct methylation profiles

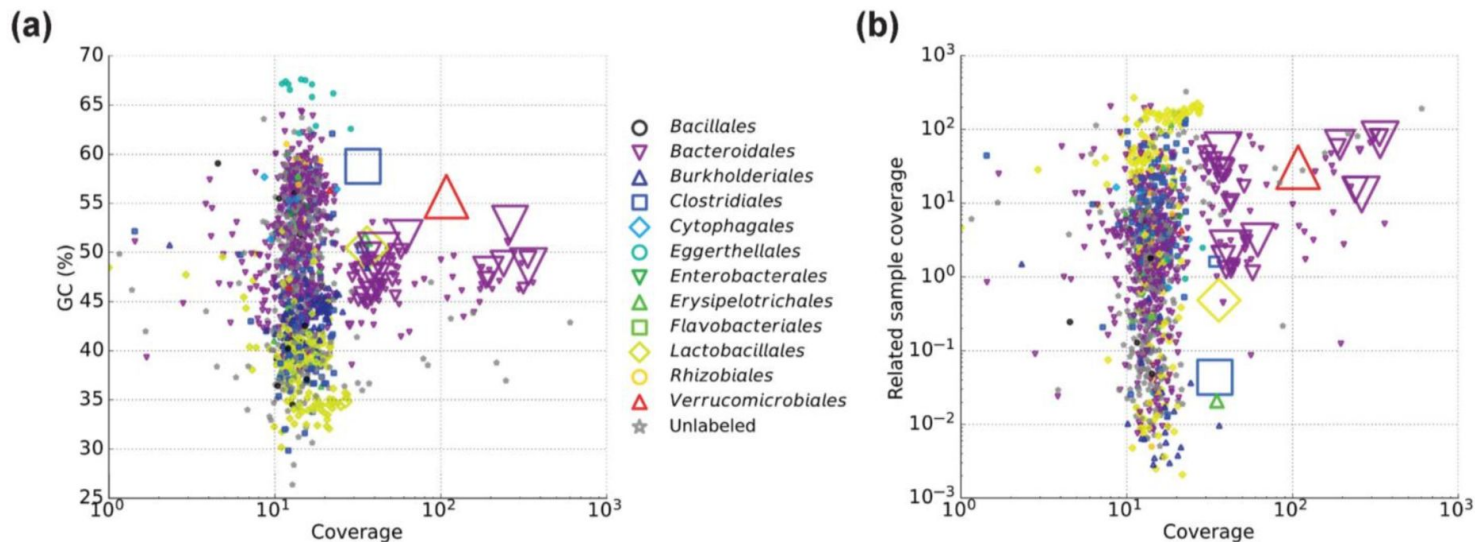
# Adding coverage

→ Meth bin 7 thought to be two different genomes

- ◆ Substantial contamination found using CheckM



# Coverage and sequence characteristic

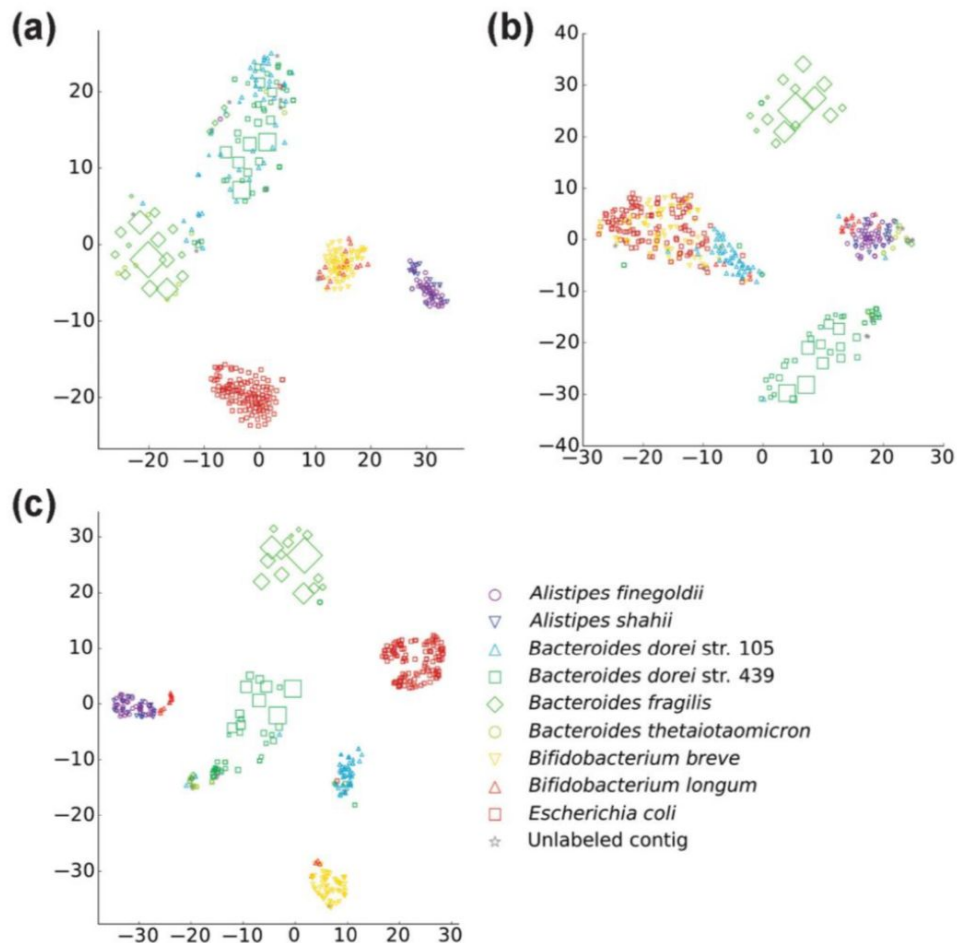


Bins don't appear to resolve as well using just coverage and/or GC content

# A hybrid approach

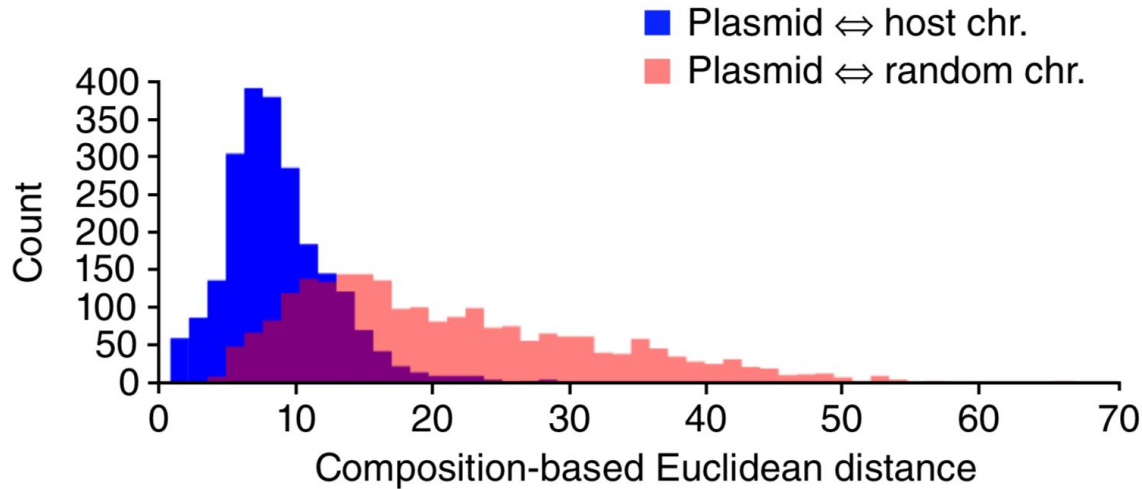
- a. 5-mer freqs only
- b. Meth only
- c. Harmony in the universe  
(still some genus level  
confusion tho)

(infant gut microbiome data)



# Who owns this plasmid?

**a**



→ Use coverage?

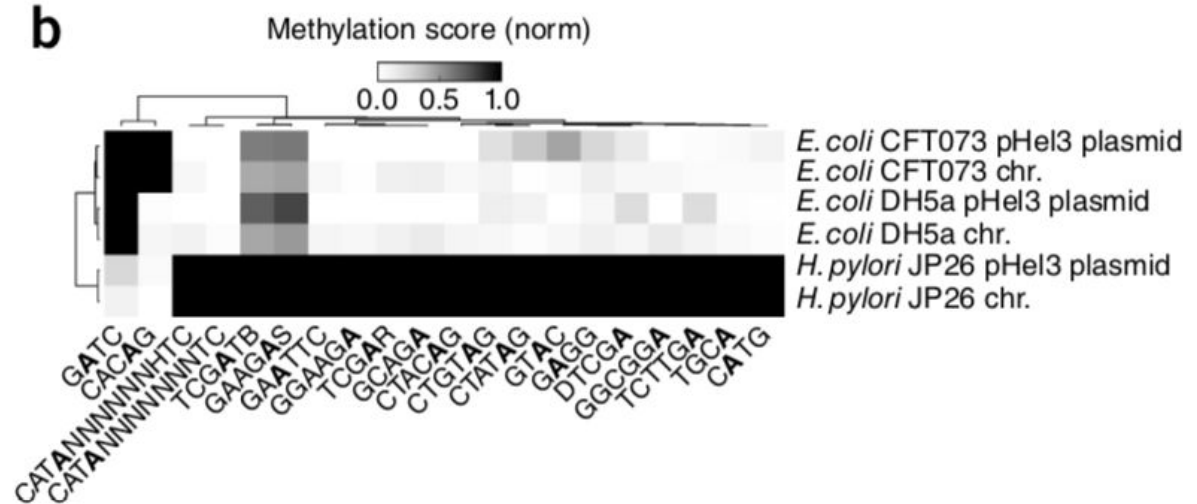
- ◆ Plasmids replicate independently from chromosomes

→ 5-mer freqs?

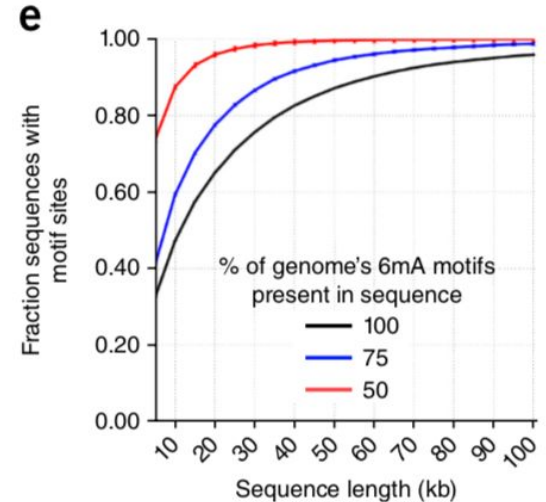
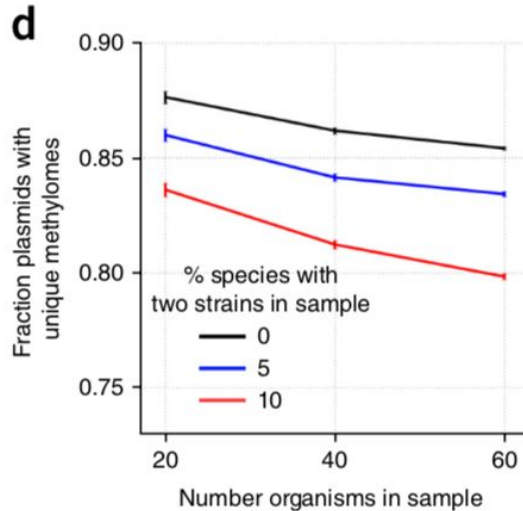
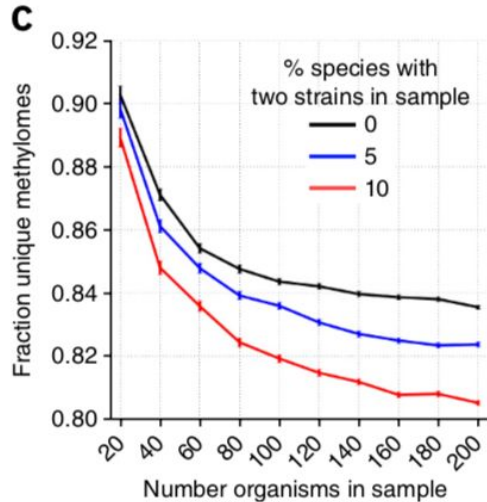
- ◆ Sort of?
- ◆ Plasmids shared between species through conjugation could be a problem

# Using methylation

Plasmids transformed into new hosts take on the host's methylation profile.

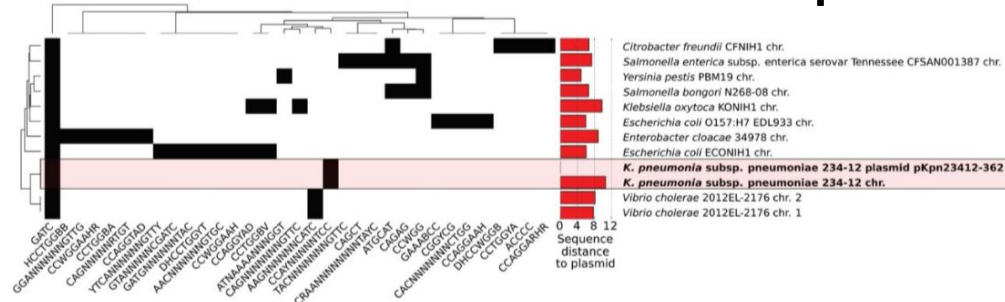


# Using methylation: a consideration

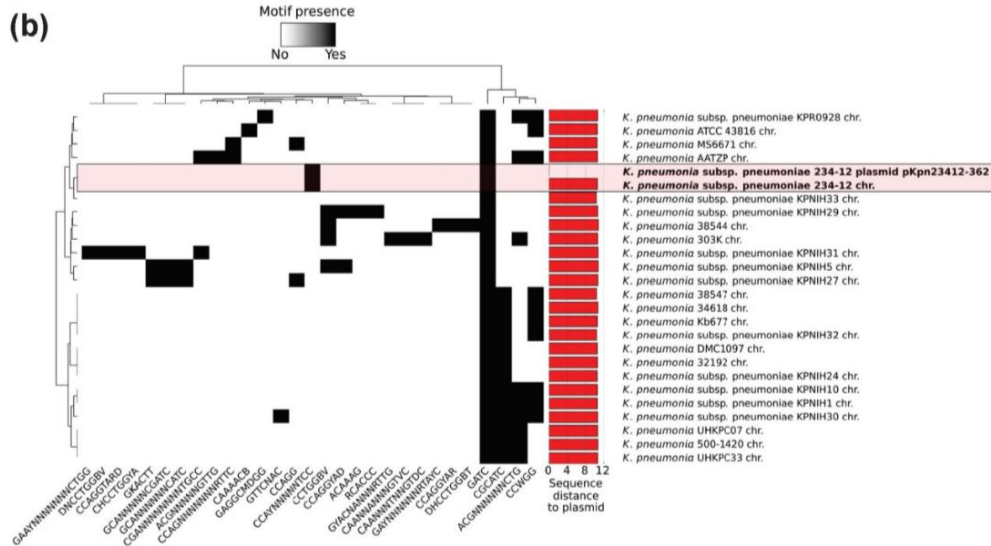


- Methylome sharing makes plasmid ID difficult - becomes a problem at high richness.
- Small plasmids might not have enough motifs to discriminate well.

(a)



(b)



# A particularly pernicious plasmid

→ pKpn23412-362 in *Klebsiella pneumoniae*

◆ encodes 13 antibiotic resistance genes!

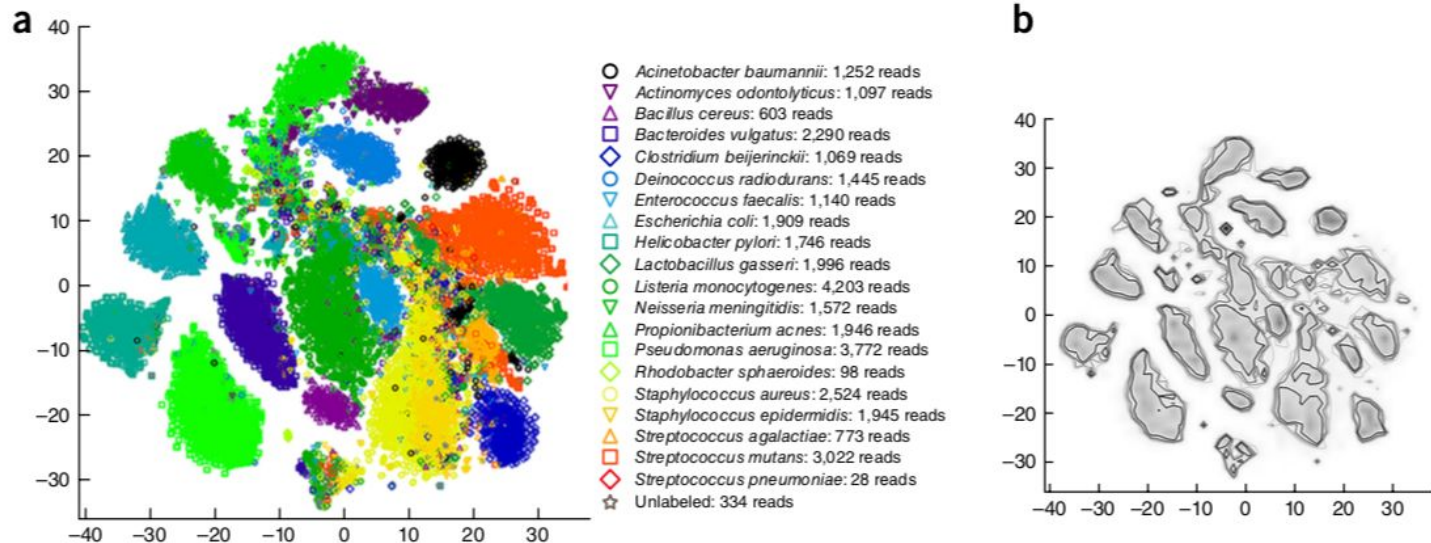
→ An examination of REBASE (so why is motif presence a gradient?)

→ Mouse ubiome data: 8/19 MGEs were conclusively attributed to bins based on meth profile.



# Read binning

→ For low coverage, difficulty assembling

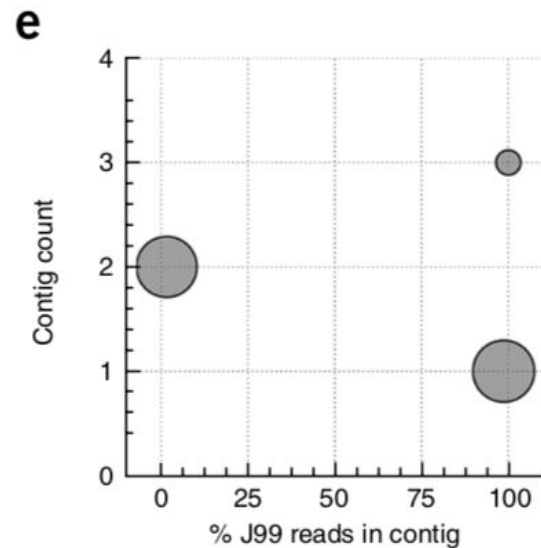
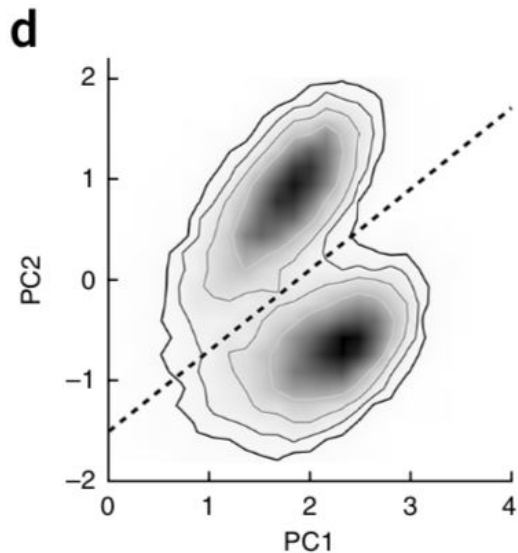
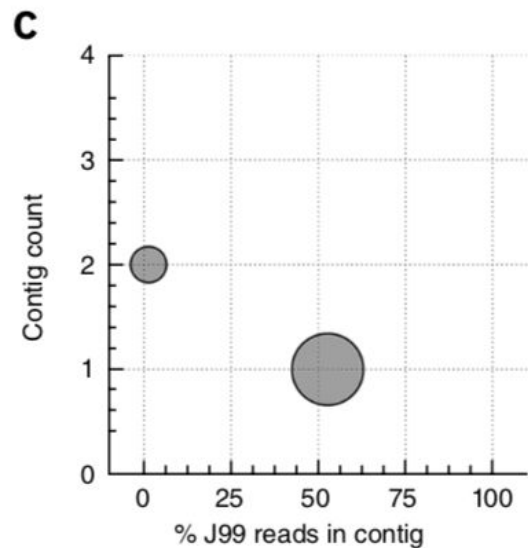


t-SNE of 5-mer freqs and 2D hist of the same

# Read segregation prior to assembly

→ 2 strains (J99 and 26695) of *H. pylori* with different meth profiles

- ◆ Bulk assembly results in a highly chimeric contig
- ◆ Meth based read segregation results in strain specific assemblies
- ◆ What is 'contig count'?

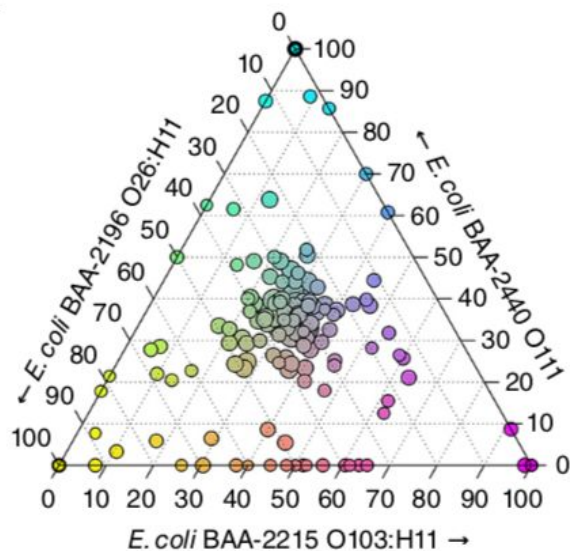


# Read segregation prior to assembly

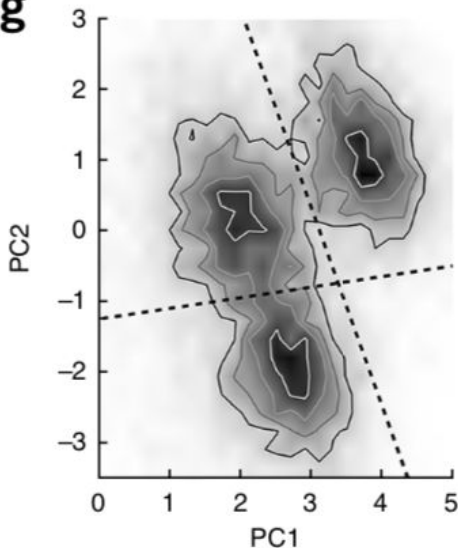
→ 3 strains of ecoli with different meth profiles

- ◆ Needed error correction step
- ◆ Is each circle a contig?

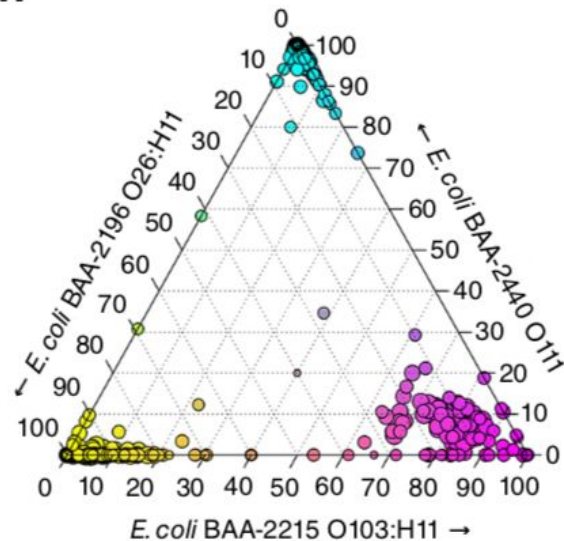
**f**



**g**



**h**



# References

1. John Beaulaurier, 'Metagenomic binning and association of plasmids with bacterial host genomes using DNA methylation' *Nature Biotechnology* (January 2018)
2. Mads Albertsen, 'Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes' *Nature Biotechnology* (May 2013)