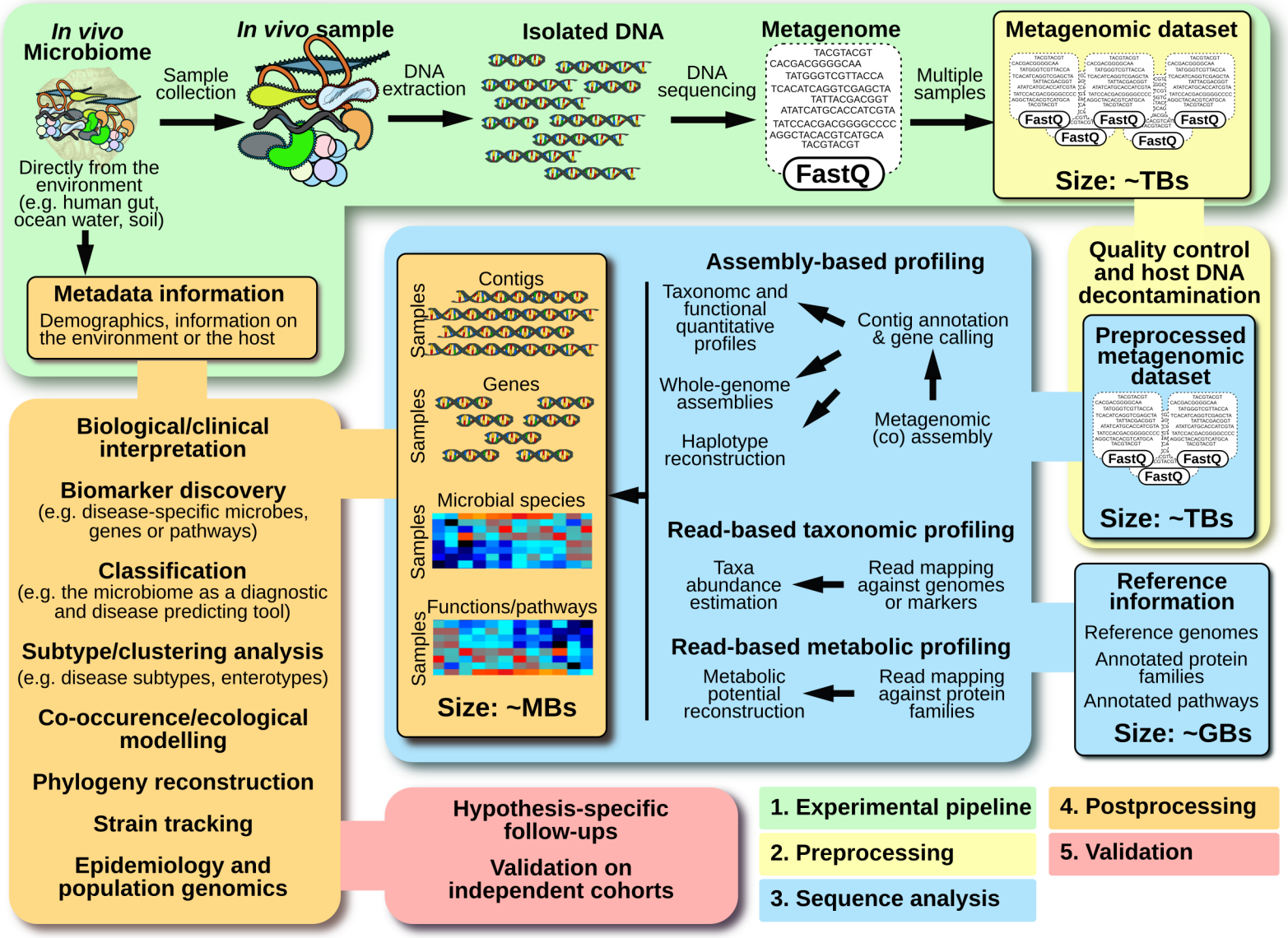
Contig Binning

Christopher Quince
Warwick Medical School

Introduction

- Read based metagenome analysis throws away a lot of information
- Lose link between function and populations
- In theory that information is present even in short read next generation sequence data
- Genome resolved metagenomics aims to extract genome complements and their abundances

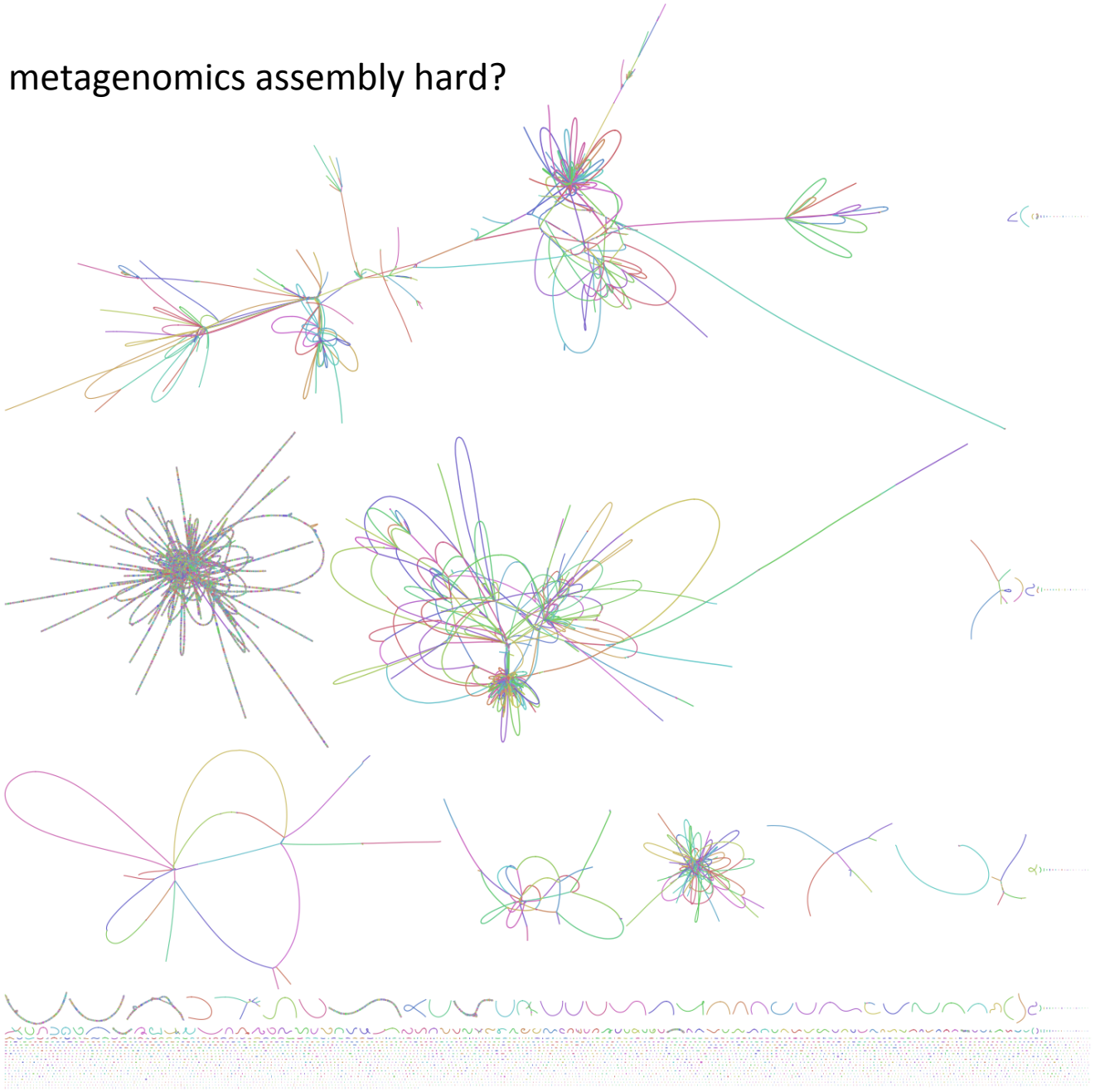


Quince et al Nat. Biotech http://rdcu.be/vMYk

Overview

- Contig binning using composition and co-occurrence (Today)
- De novo strain resolution (Tomorrow)
- Two examples:
  - Anaerobic digestion (AD) reactor time series
  - Tara oceans

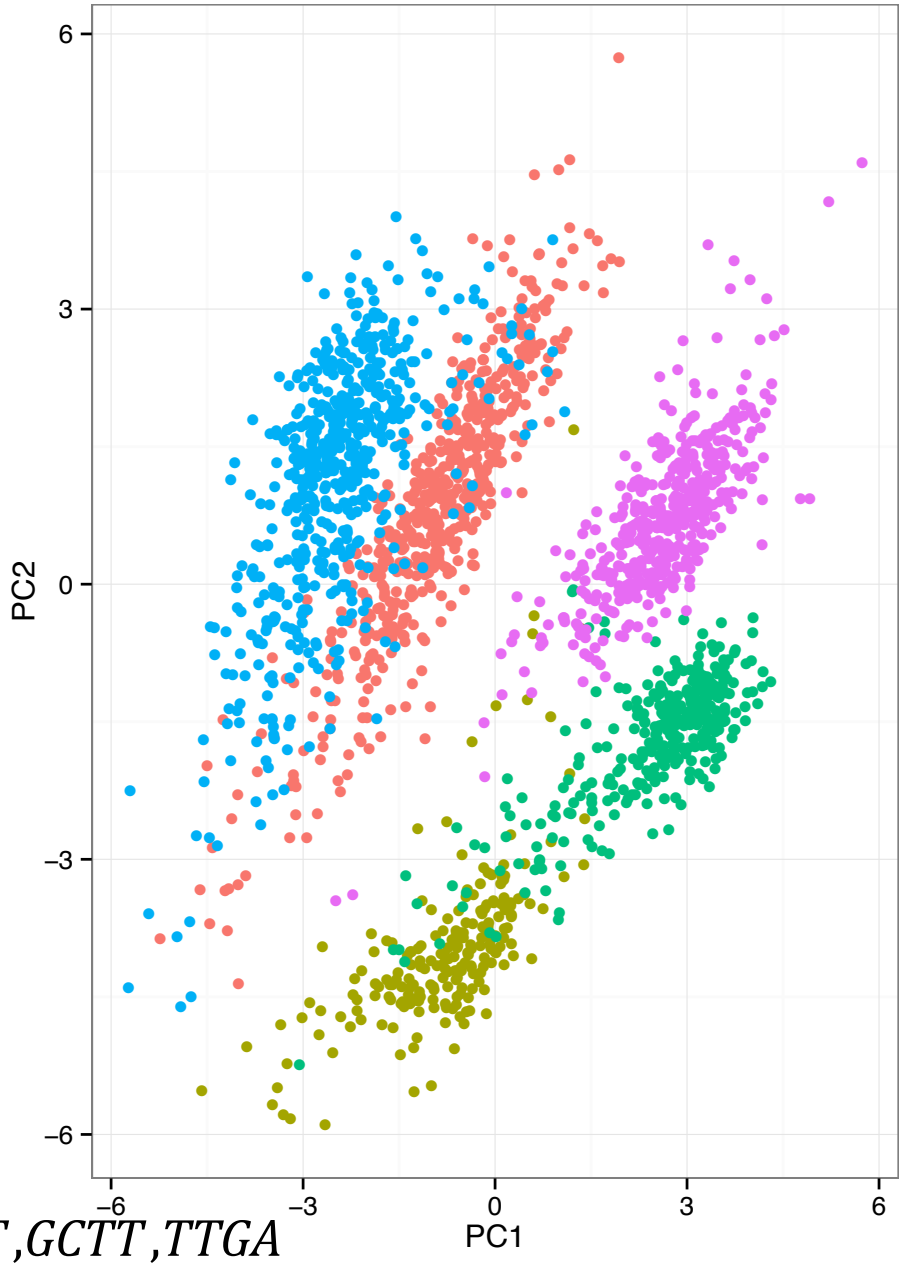
Why is metagenomics assembly hard?

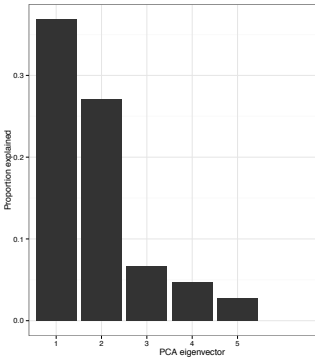


Use contig binning to cluster contigs back into strain/species genomes

Contig clustering by sequence composition

1) Different species have characteristic signatures of overlapping kmer frequencies (Sandberg et al. Genome Res. 2001)
2) Fragmented five genomes into 10kb
3) Counted tetramers and added pseudocount
4) Calculated log-proportions
5) Generated PCA

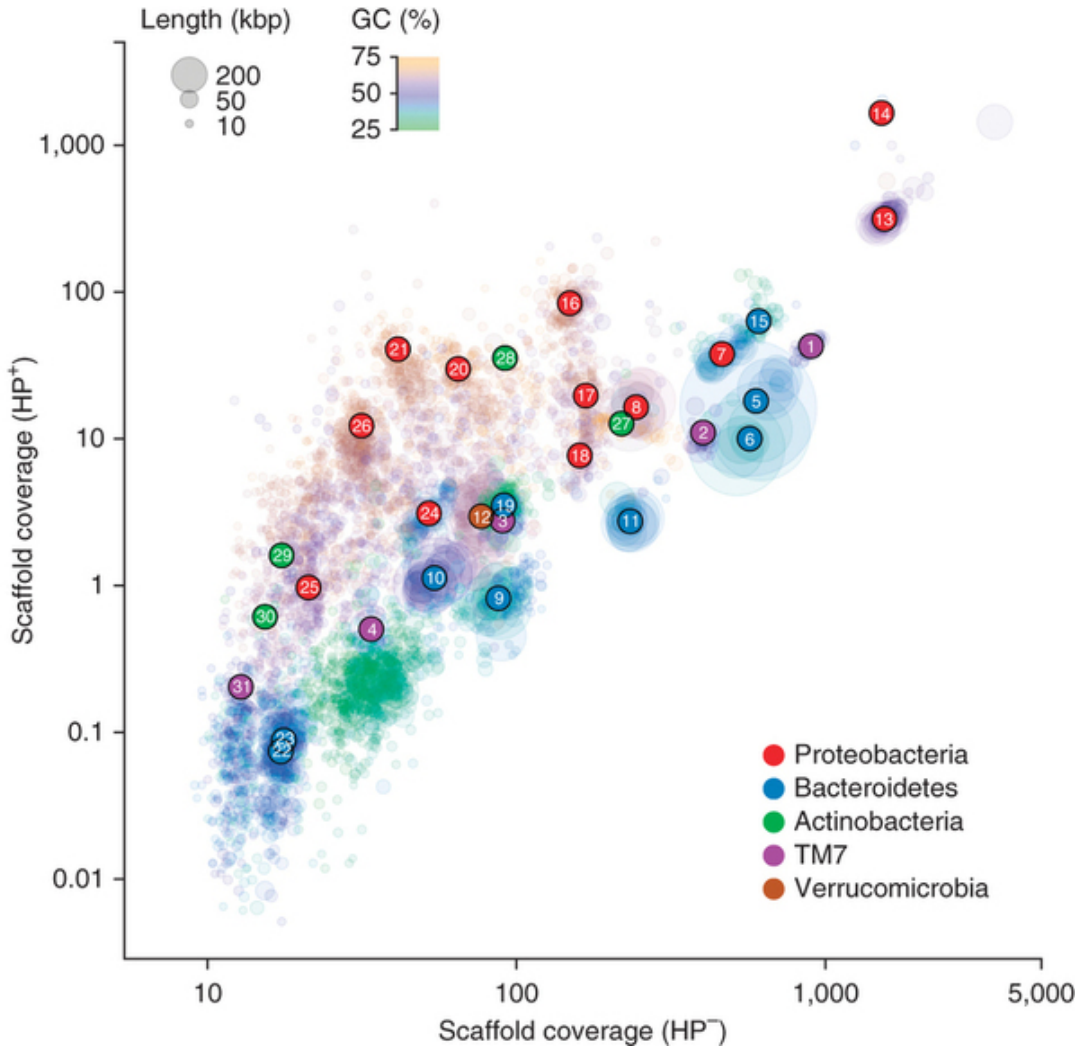




$$sequence = \left(\overline{CTGGCTTGA}\right)$$

$$4-mers: 2\times CTGG, TGGC, GGCT, GCTT, TTGA$$

Albertsen et al. “Genome sequences of rare, uncultured bacteria obtained by differential coverage binning of multiple metagenomes” Nature Biotech. 2014



Overview of methods for clustering contigs

Automatic composition based

- CompostBin (Chatterji et al. RECOMB 2008): PCA + normalized cut clustering algorithm
- LikelyBin (Kisyluk et al. BMC Bioinf. 2009): Nested likelihood model
- Scimm and MetaWatt (Kelly et al. BMC Bioinformatics 2010, Straus et al. Front Microbiol 2012): Interpolated Markov Models

Human-input co-occurrence based

- ESOM (Sharon et al. Genome Res 2013)
- ANVIO (Meren et al. 2015…)
- Albertsen et al.2014

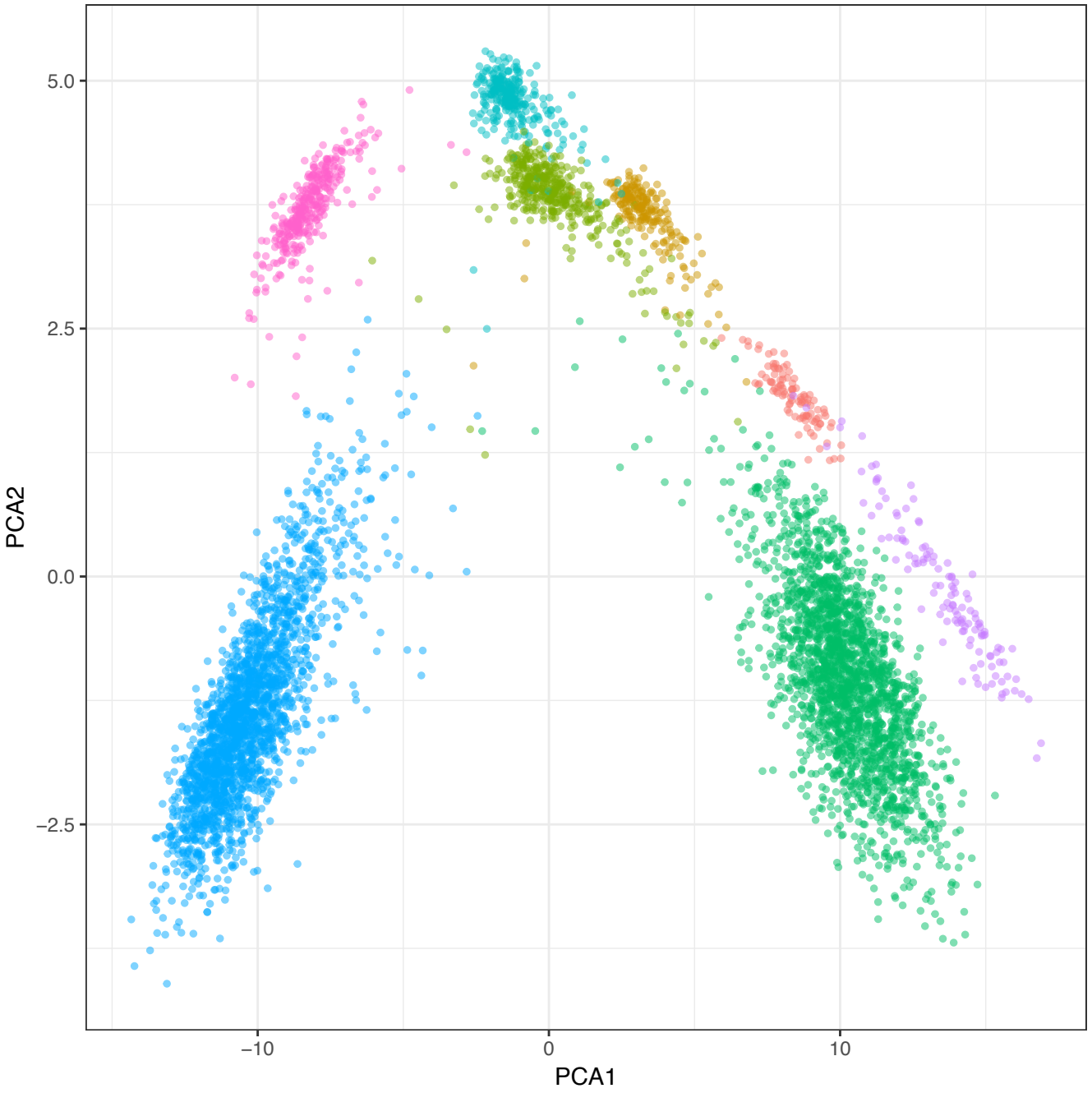
Automatic co-occurrence (and composition) based

- MetaHit gene cataologue (Almeida et al. Nat. Biotech. 2014)
- GroopM (M. Imelfort et al. PeerJ preprints 2014)
- CONCOCT (Alneberg et al. Nat. Methods 2014)
- MetaBAT (Kang et al. PeerJ 2015)
- MaxBin 2 (Wu et al. Bioinf. 2015)

CONCOCT: **C**lustering c**ON**tigs on **CO**verage and **C**ompos**T**ion

Data pre-processing

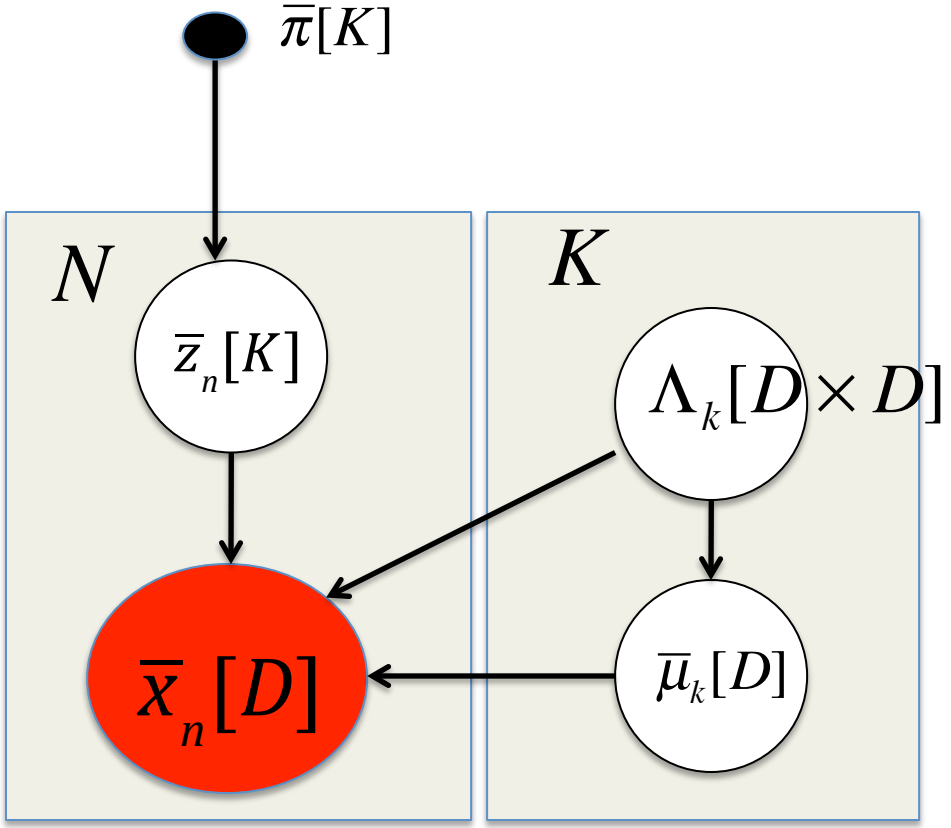
- Perform coassembly across all samples
- Fragment contigs greater than 10kb and map reads back to *N* contigs to get mean coverage of contig in each of *M* samples
- Generate k-mer frequency vector for each contig
- Add pseudo-counts, normalise coverage and k-mer frequencies, join and log-transform
- Perform PCA keep *D* dimensions that explain 90% of variance



D = 32

CONCOCT: Contig Clustering

- Describe each contig cluster in data set as a Gaussian with full covariance
- Complete data set derives from a mixture of *K* of these components
- Variational Bayes to select number of components
- In practice an Expectation-Maximisation (EM) algorithm



Other binning algorithms

- MetaBAT (Kang et al. PeerJ 2015): Modified k-medoid clustering algorithm using coverage and composition distances parameterised from existing bacterial genomes
- MaxBin2 (Wu et al. Bioinf. 2015): EM algorithm operating on untransformed kmer frequencies and coverages (maybe uses core genes)
- MetaWatt and MaxBin2 best performance in CAMI challenge but CAMI examples had limited sample number (max 5):

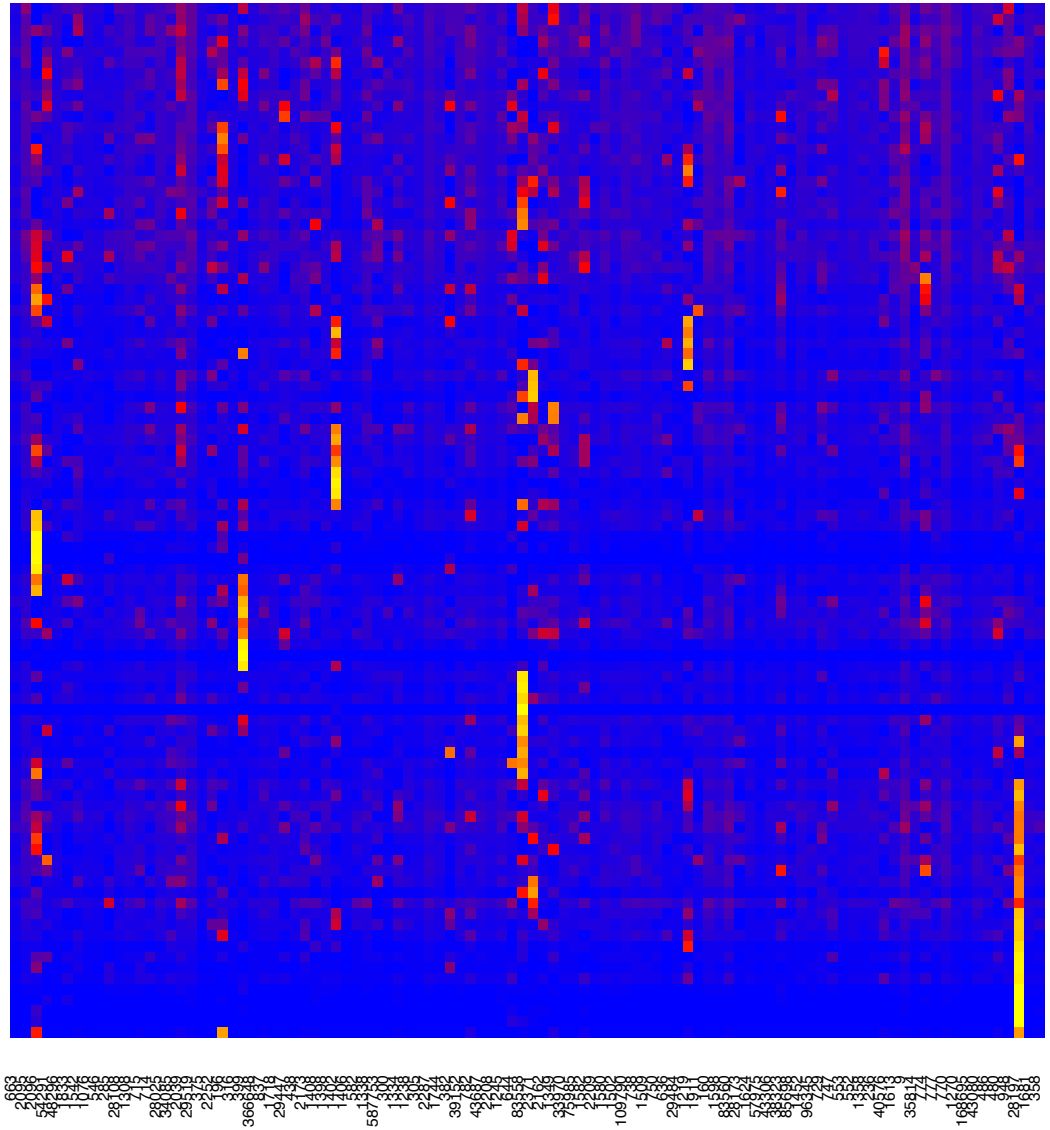
http://www.biorxiv.org/content/early/2017/01/09/099127

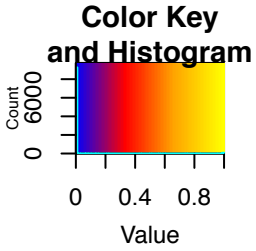
Complex synthetic community

- 100 different species and 210 NCBI genomes
- 10 ten separate phyla, 49 families, and 74 genera
- Species strain frequency distribution of (1:50, 2:20, 3:10, 4:10, 5:10)
- Simulated 96 samples of 6.25 million 2X150 bp paired end reads using ART: 1 HiSeq 2500 high output: https://github.com/chrisquince/StrainMetaSim

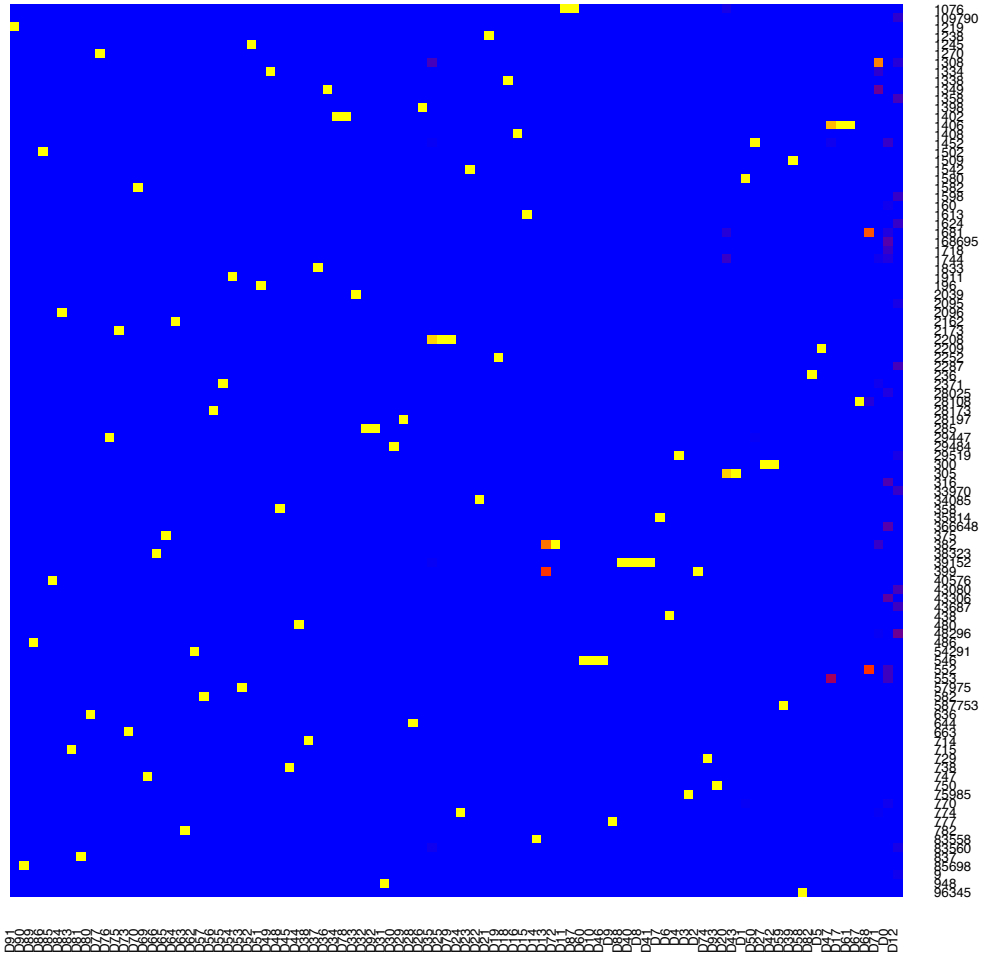
Complex synthetic community

- Log-normal distribution for species across samples
- Dirichlet for strains within a species
- Total species coverage ranges from 44.16 to 12490 with median 242.80
- Coassembly with megahit gave N50 11,940 bp
- 74,580 contig fragments with a total length of 409 Mbp
- 687 Mbp for all 210 reference genomes



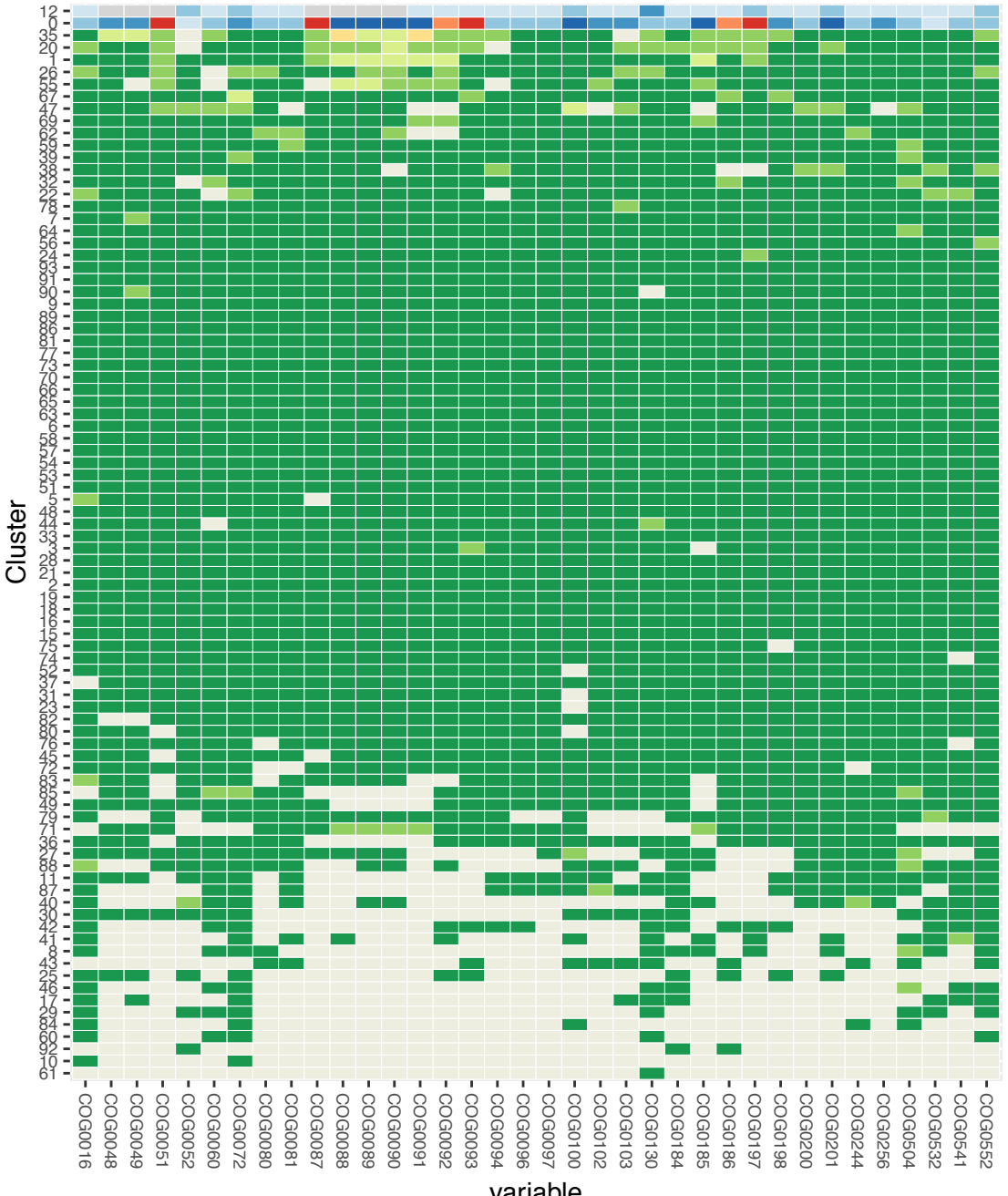


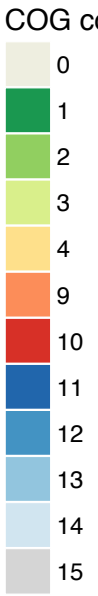
Concoct binning



- CONCOCT generated 94 clusters
- Compare contig clusters to true species assignments
- Species recall of 87%, precision of 84%, Adjusted Rand index ARI = 0.545

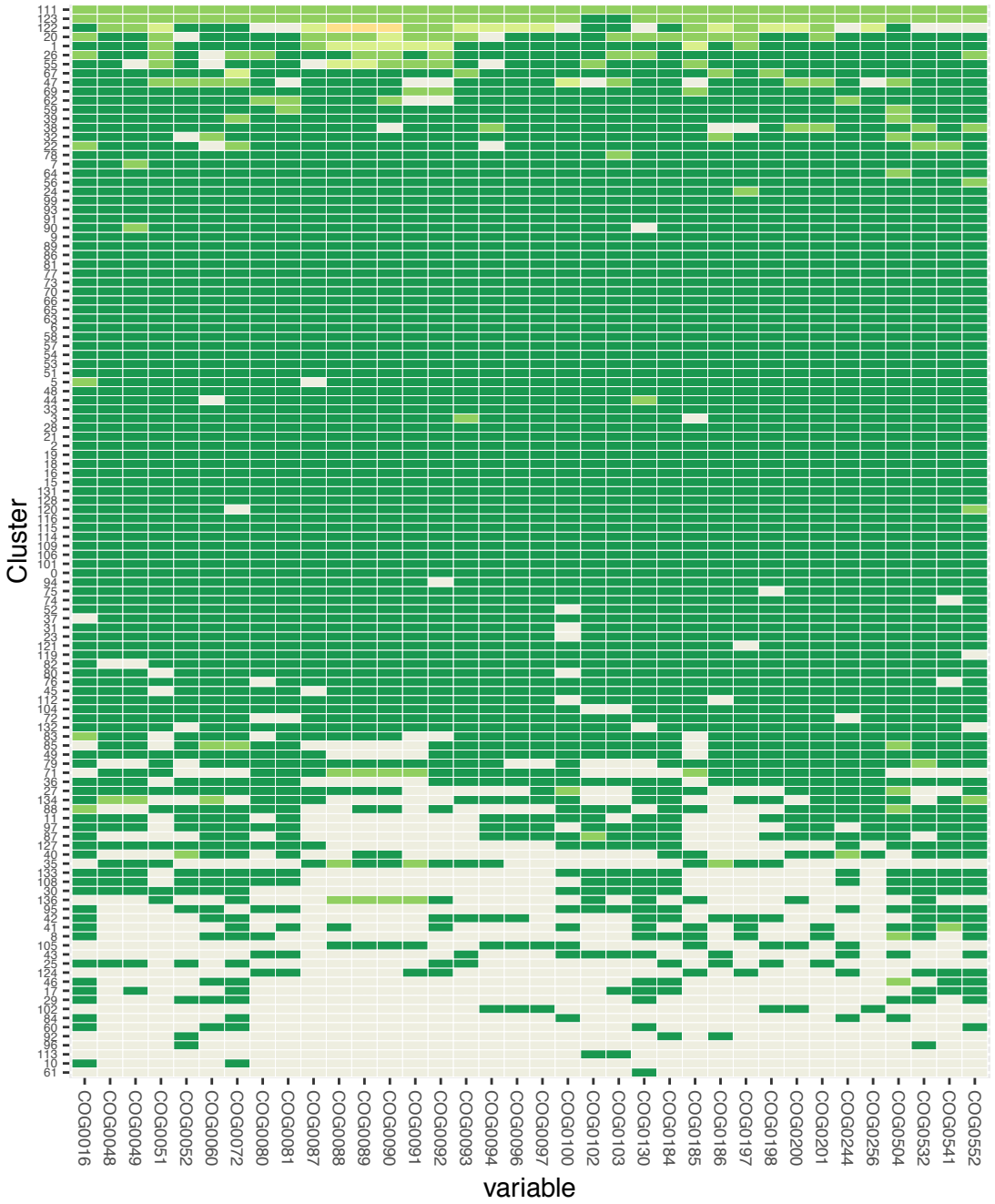
Concoct single-copy core genes





- 58 clusters with 75% of SCGs in single copy

Concoct2 refinement



- Reran CONCOCT on each cluster with median SCG no. greater than 2
- CONCOCT2 generated 137 clusters (species recall of 86.1% and a precision of 98.2%, ARI = 0.83)
- 75 clusters with 75% of SCGs in single copy

Complex synthetic community comparison

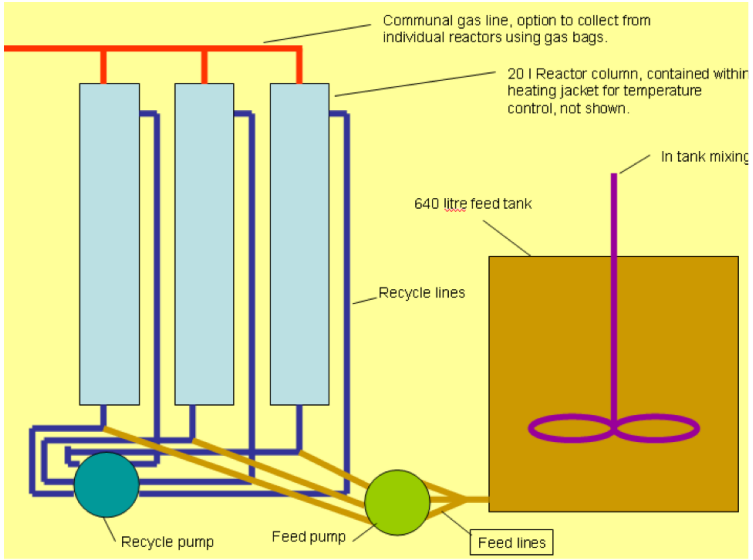
- MetaBAT: minimum contig length 1500bp obtains (82% recall, 98% precision, ARI 0.78) but only 57 75% complete genomes
- MaxBin2: 86% recall, 92% precision, ARI 0.78 and 71 complete genomes

Expanded Granular Sludge-Bed Laboratory Bioreactors (EGSB)

- Seed from industrial EGSB bioreactor treating distillery waste
- Applied a series of step-wise engineered changes to triplicate reactors:

1) Treating low-strength distillery waste (600mg/l COD) at 37degC for 12 weeks
2) Treating low-strength distillery waste (600mg/l COD) at 15degC for 13 weeks
3) Treating low-strength SYNTHES (500mg/l COD) at 15 degC for 6 weeks

- Sequenced 95 reactor samples – 521,492,655 2X125 bp reads



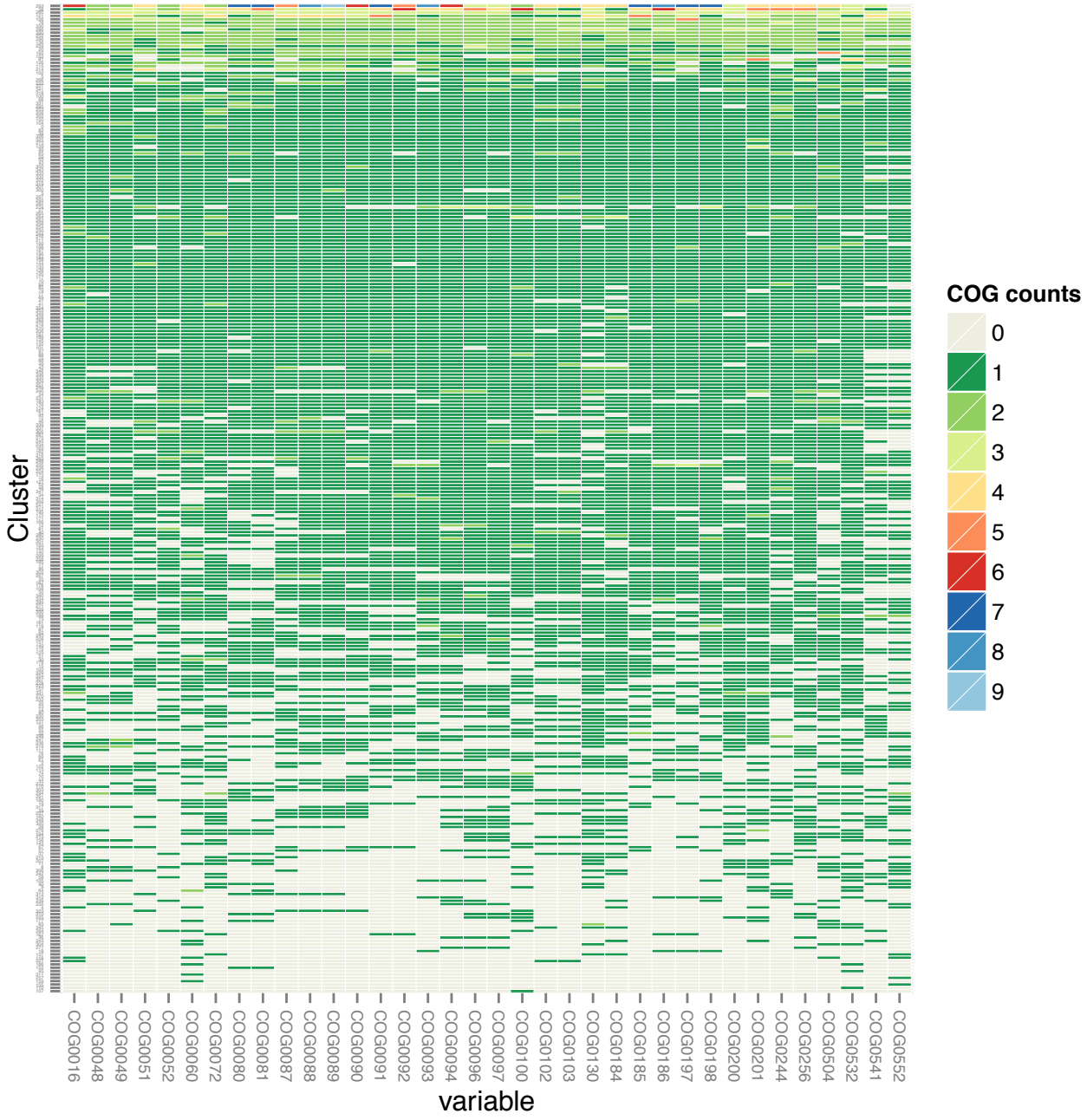
Metagenomics AD sequencing

Assembled with Ray generated 393,230 contigs > 1kb, total length 1,394 Mb
186,081 > 2kb

355 CONCOCT clusters

152 – 70% pure and complete from analysis of single copy core COGs – refer to these as metagenome assembled genomes (MAGs)

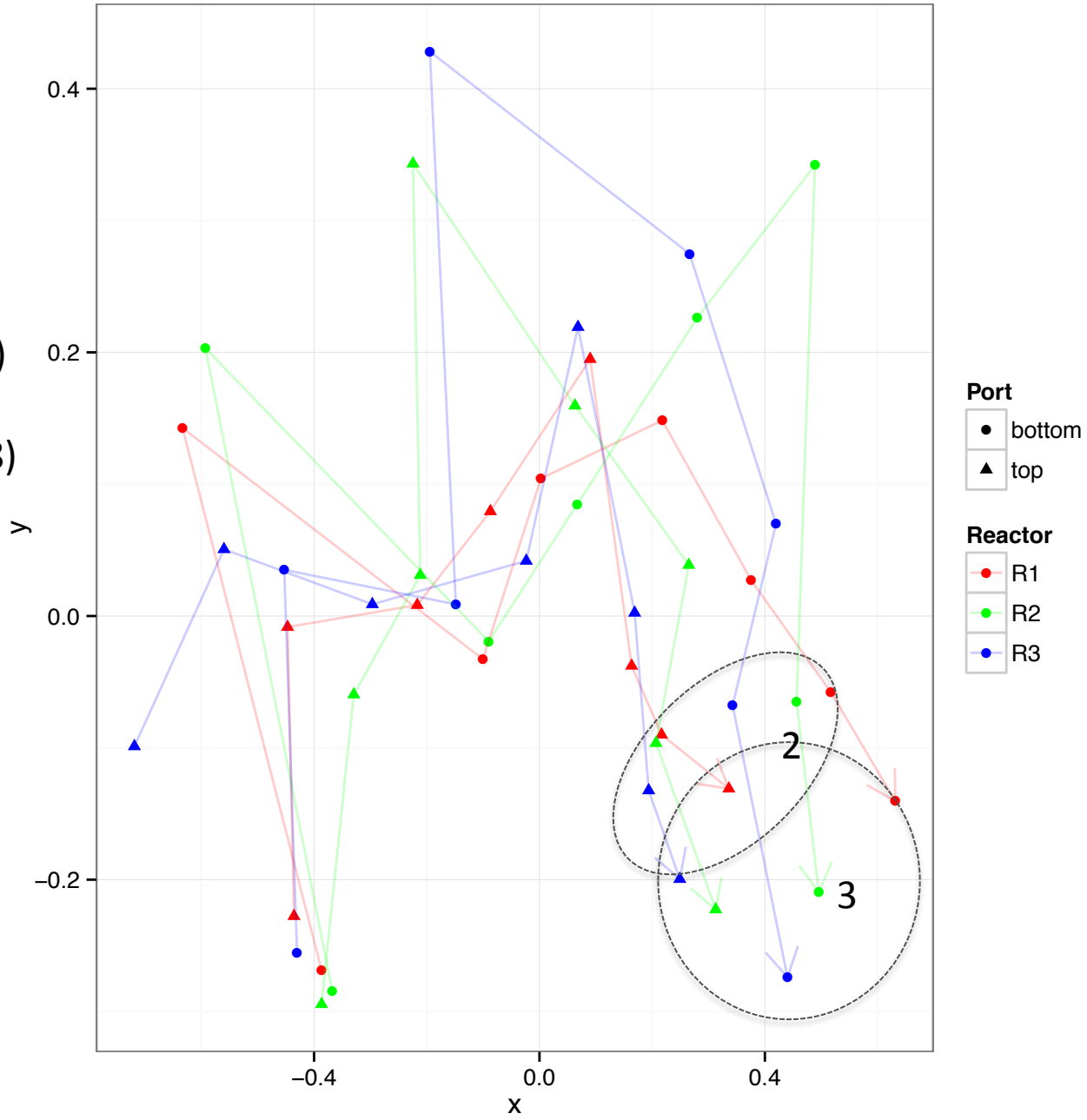
Metabat - 123
MaxBin2 - 123

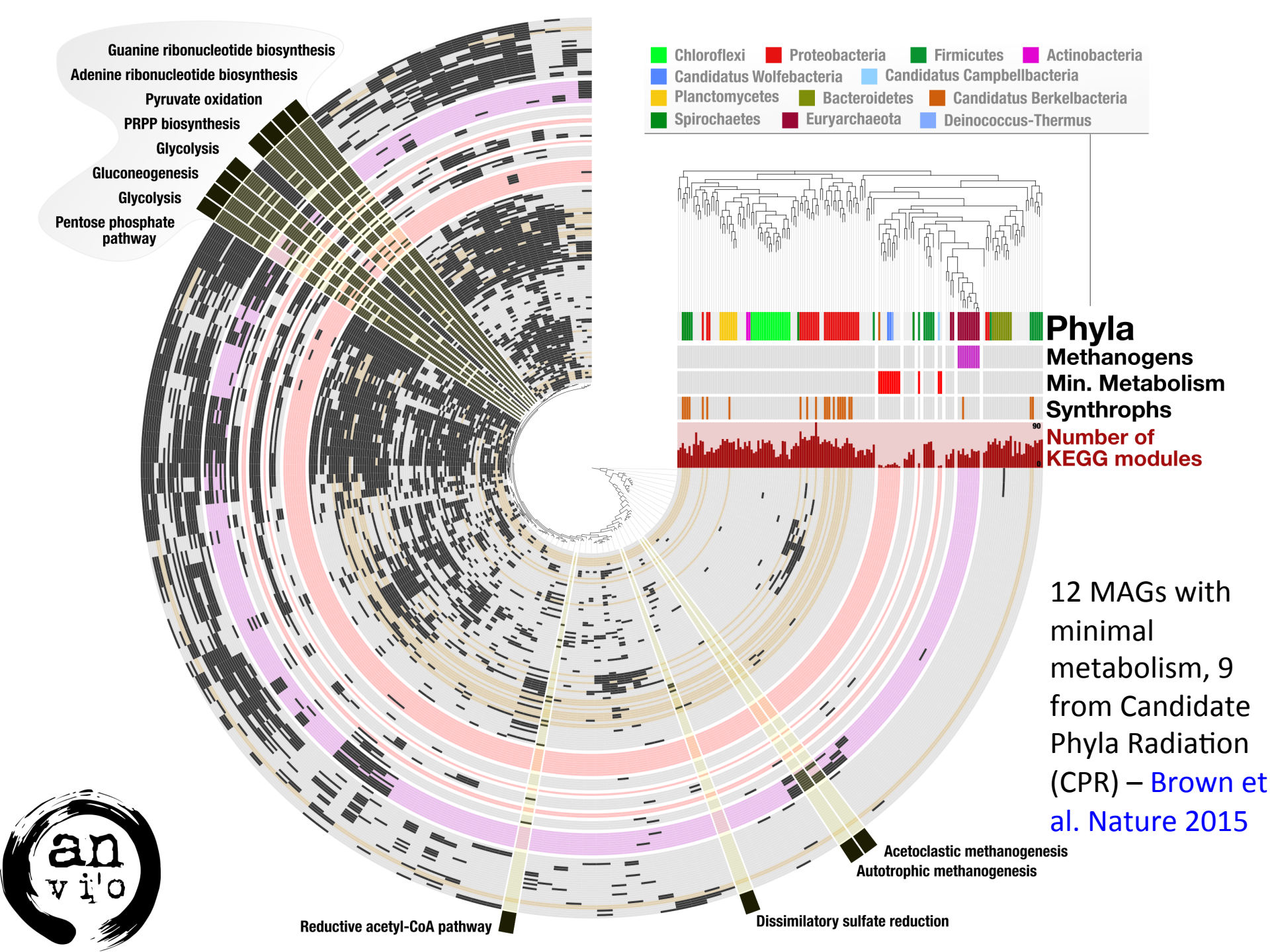


Based on NMDS of mean cluster abundances observe community succession over time

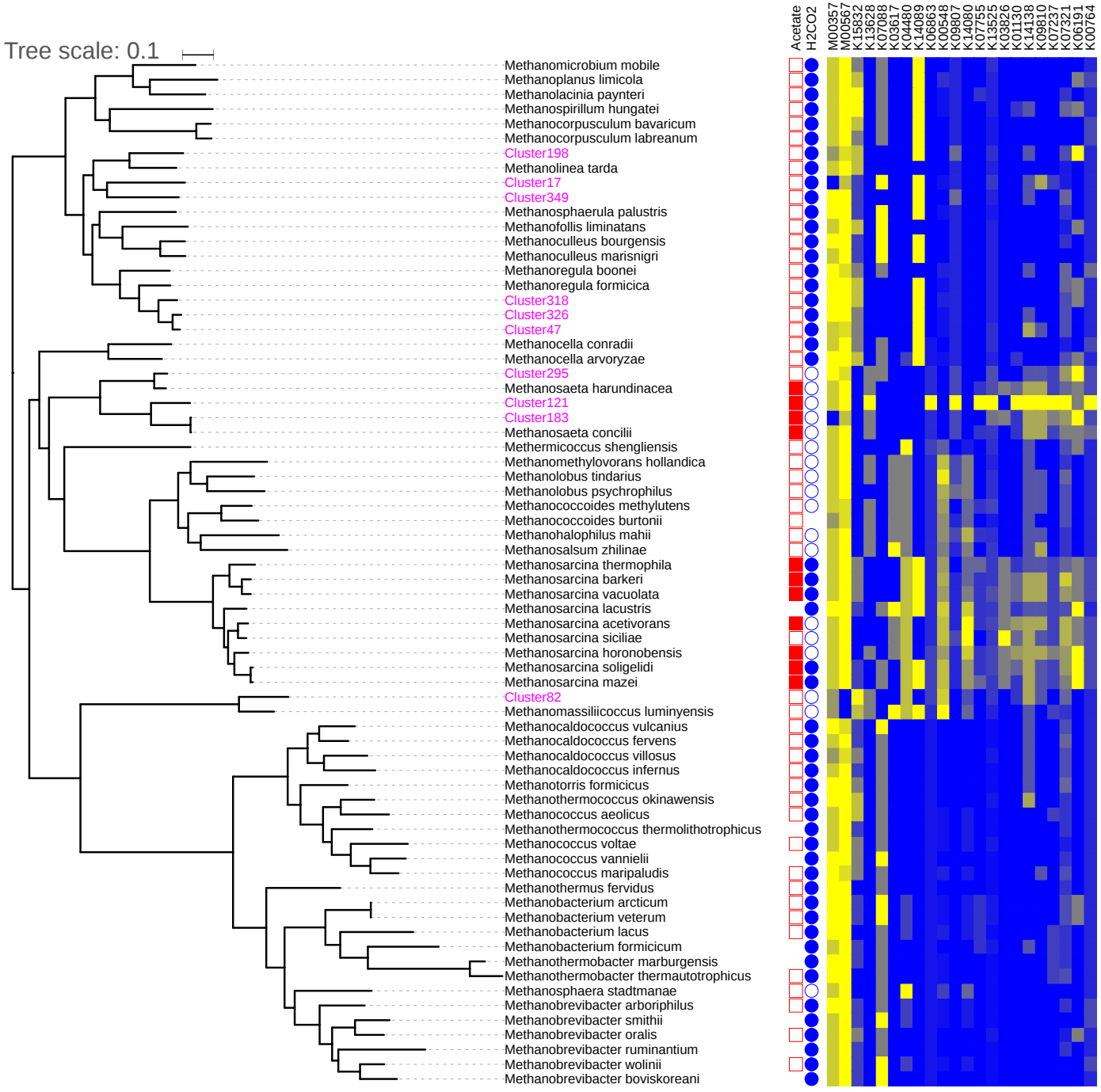
Time explained 36% (p = 0.01)

Reactor marginal 8% (p = 0.08)





12 MAGs with minimal metabolism, 9 from Candidate Phyla Radiation (CPR) – Brown et al. Nature 2015



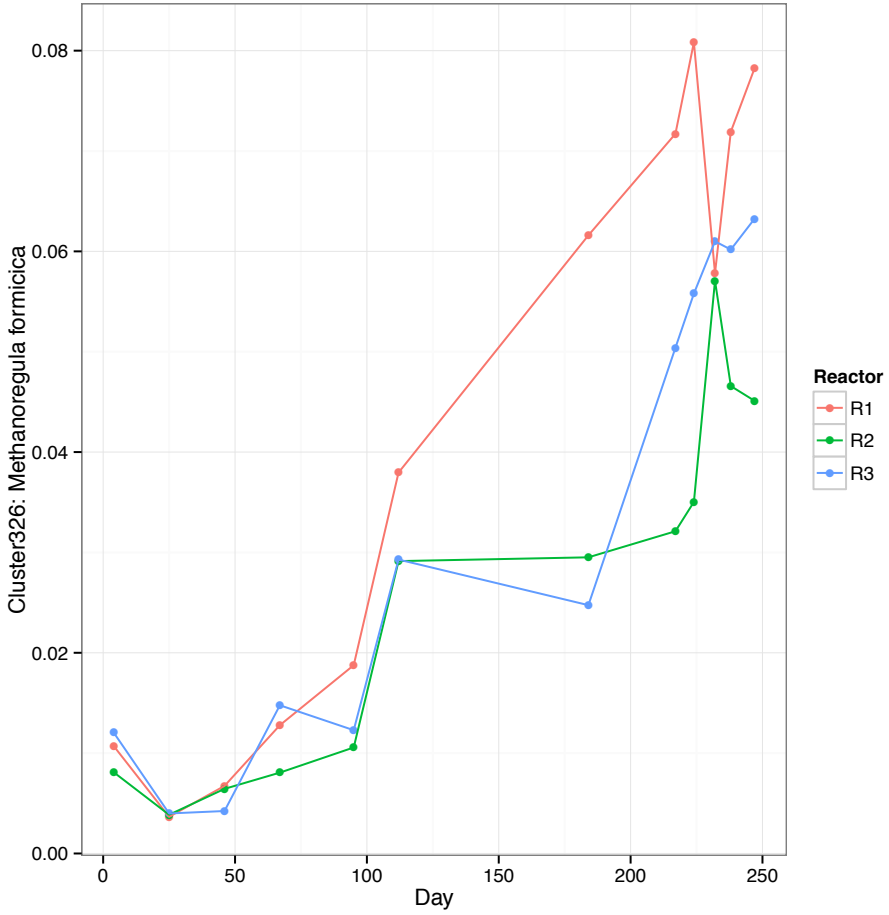
Trained Random Forest classifier on known methanogen genomes with known substrates

Predict known autotrophs or acetoclasts with 5% error rate

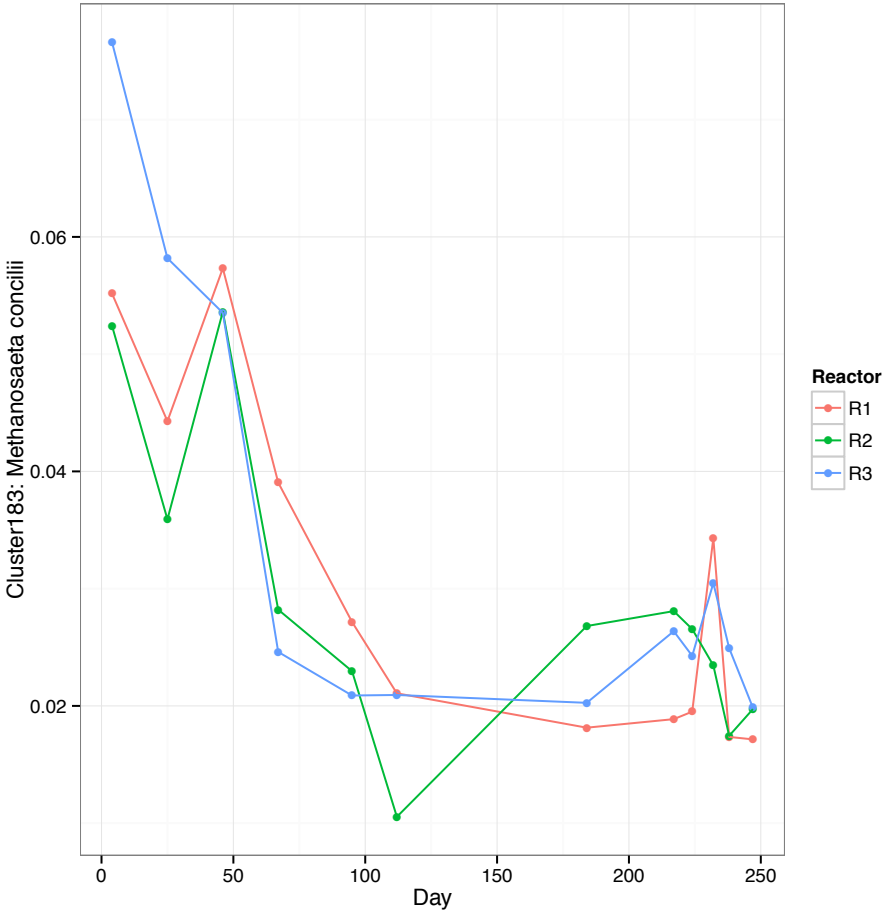
Apply to MAGs

Also provides information on what orthologs may be involved in methanogenesis

Autotroph

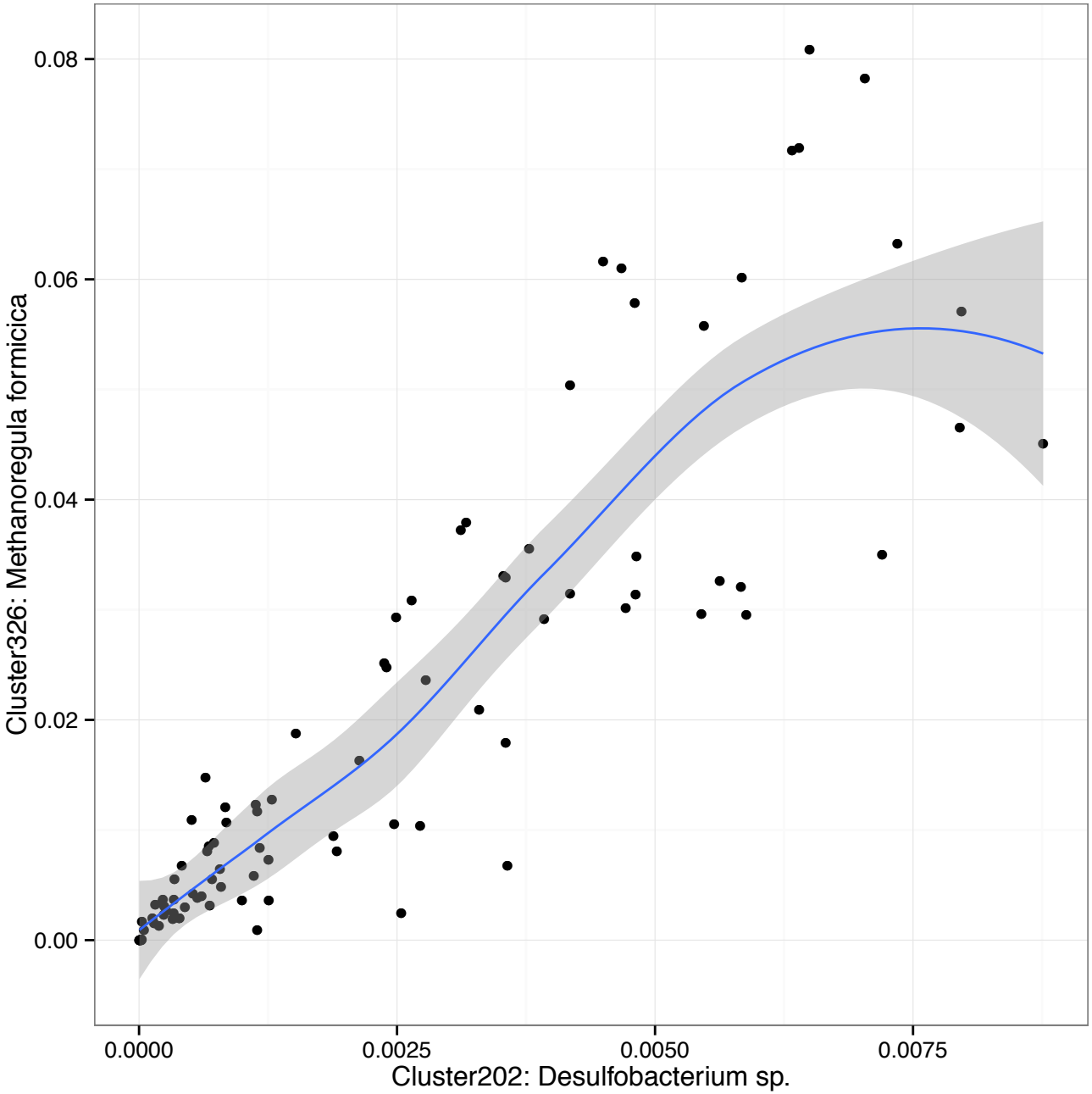


Acetoclast



Shift from acetoclastic to autotrophic methanogenesis over time

Driven by syntrophies…

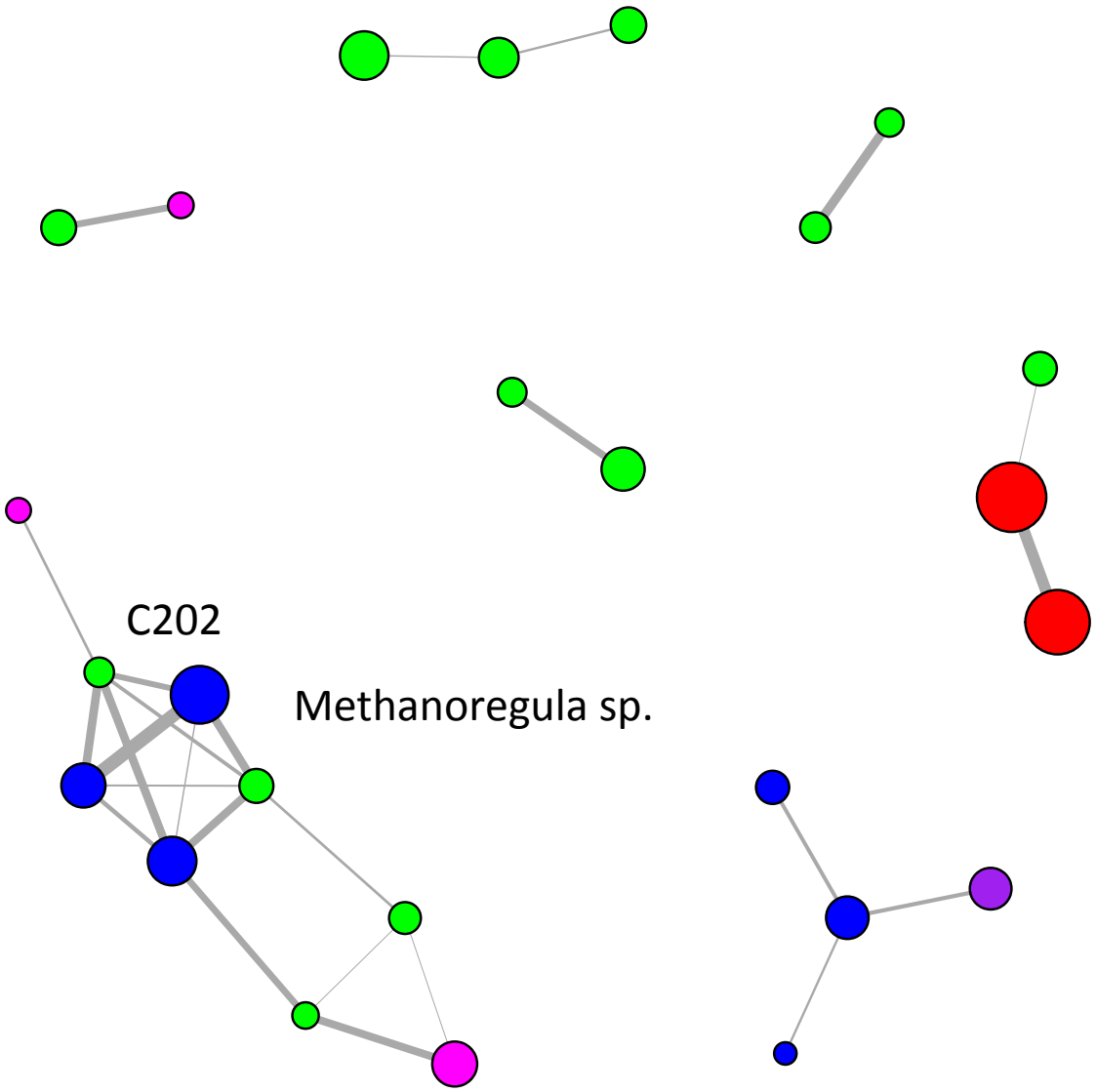


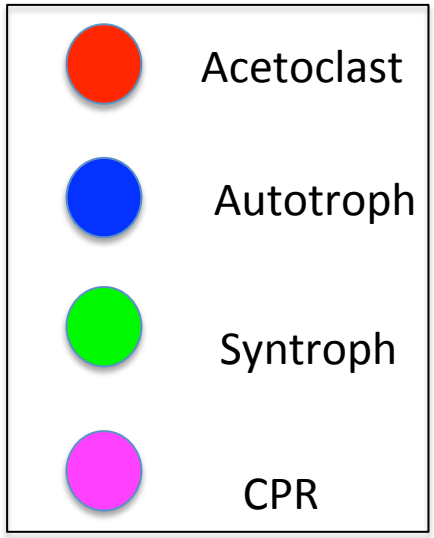
Possible syntroph

Cluster202: Desulfobacterium sp.

Sulphate reducer but possesses 3 copies of ko:K02380 Formate dehydrogenase maturation protein FdhE characteristic of syntrophs (Worm et al. 2014)

Evidence of other non-SRB syntrophs e.g. propionate degraders: *Syntrophomonas wolfei*

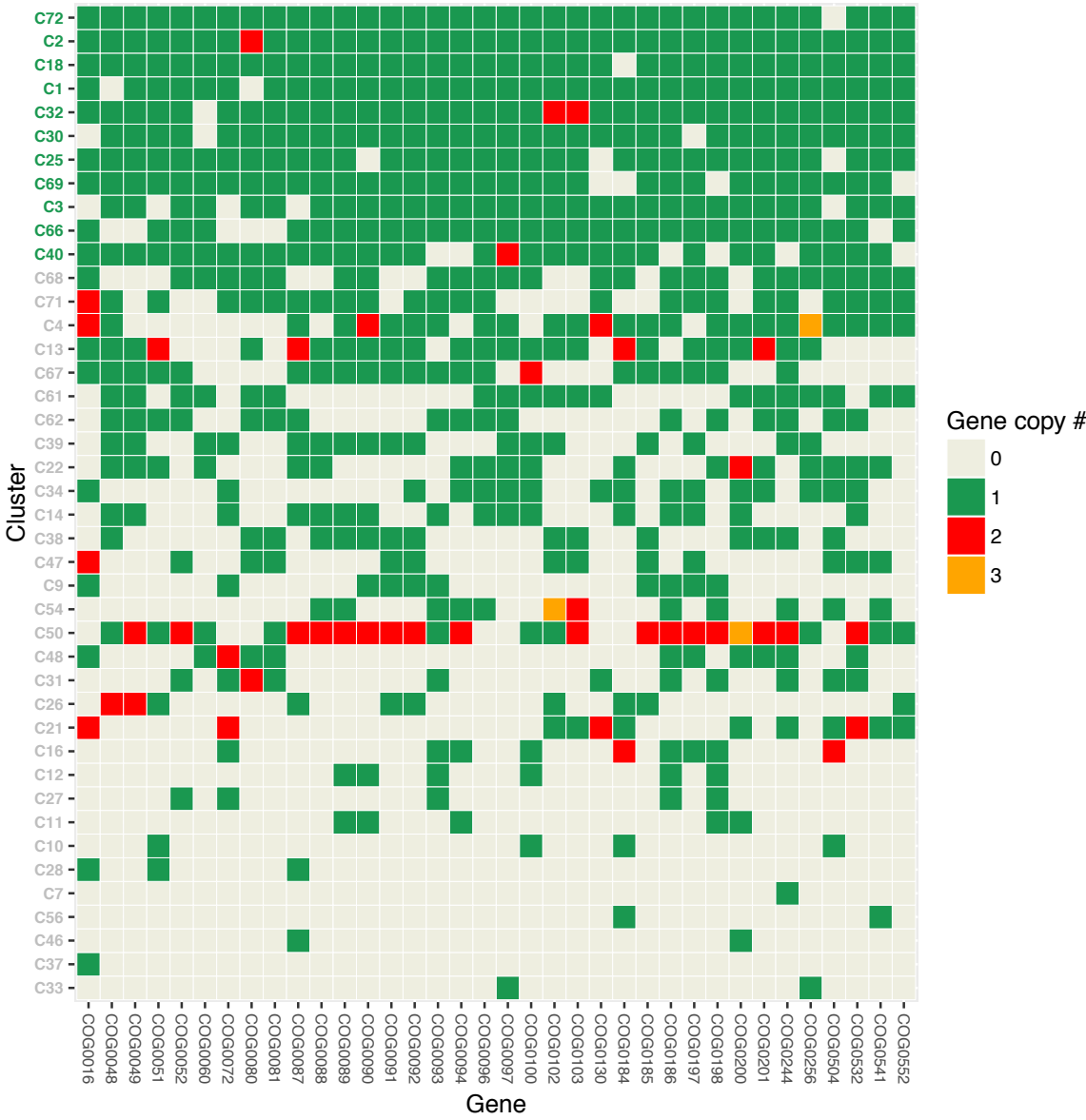




Network formed from very strong positive correlations only $r > 0.8$

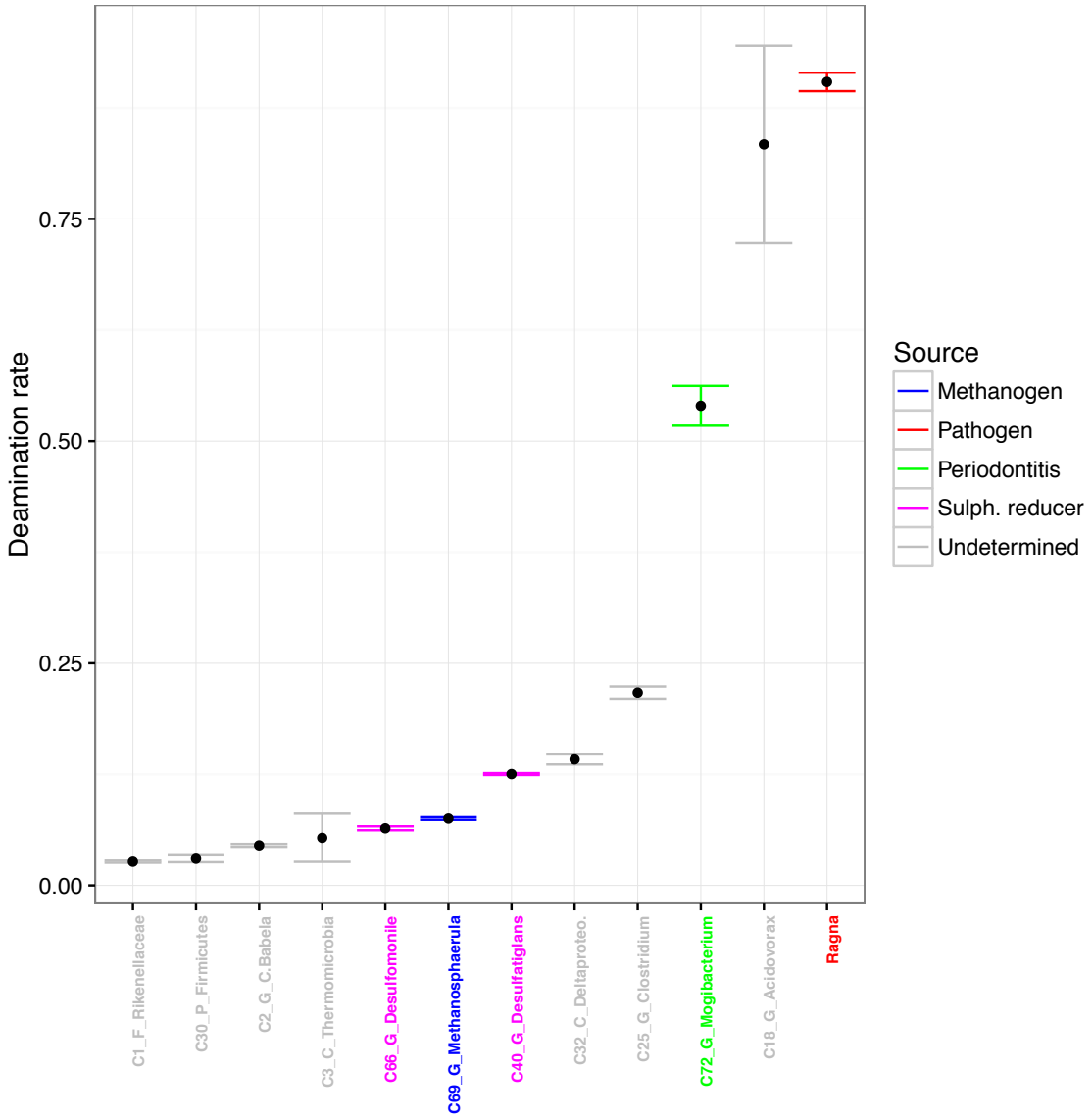
Reconstructing microbial genomes from ancient DNA samples

- 1.49 billion non-human reads from 9 libraries were assembled into 38,703 contigs
  - total length: 86 Mbp
  - N50: 2.4 kbp
- Clustered into 76 bins using CONCOCT
- 11 bins were 75% complete bacterial genomes



DNA damage in microbial genomes

- Calculated deamination rates (increase with age) using mapDamage
- Sulphate reducing bacteria (SRB) and methanogen derive from acidic water logged soil
- Mogibacterium associated with periodontitis
- Mogibacterium found almost exclusively in calculus
- Environmental organisms across all samples



Summary

- CONCOCT can accurately and automatically resolve species genomes from metagenomics data
- Effectively we can now do high throughput genomics direct from metagenomes
  - Compared to 16S rRNA gene: unbiased and comprehensive, resolve entire genome
  - Compared to read based metagenomics: allows better resolution of function as we demonstrated with methanogen pathways
  - Link metabolism with ecology inferring syntrophies etc.
  - Don't have complete assembled genomes but the gene composition is what we are really interested in anyhow