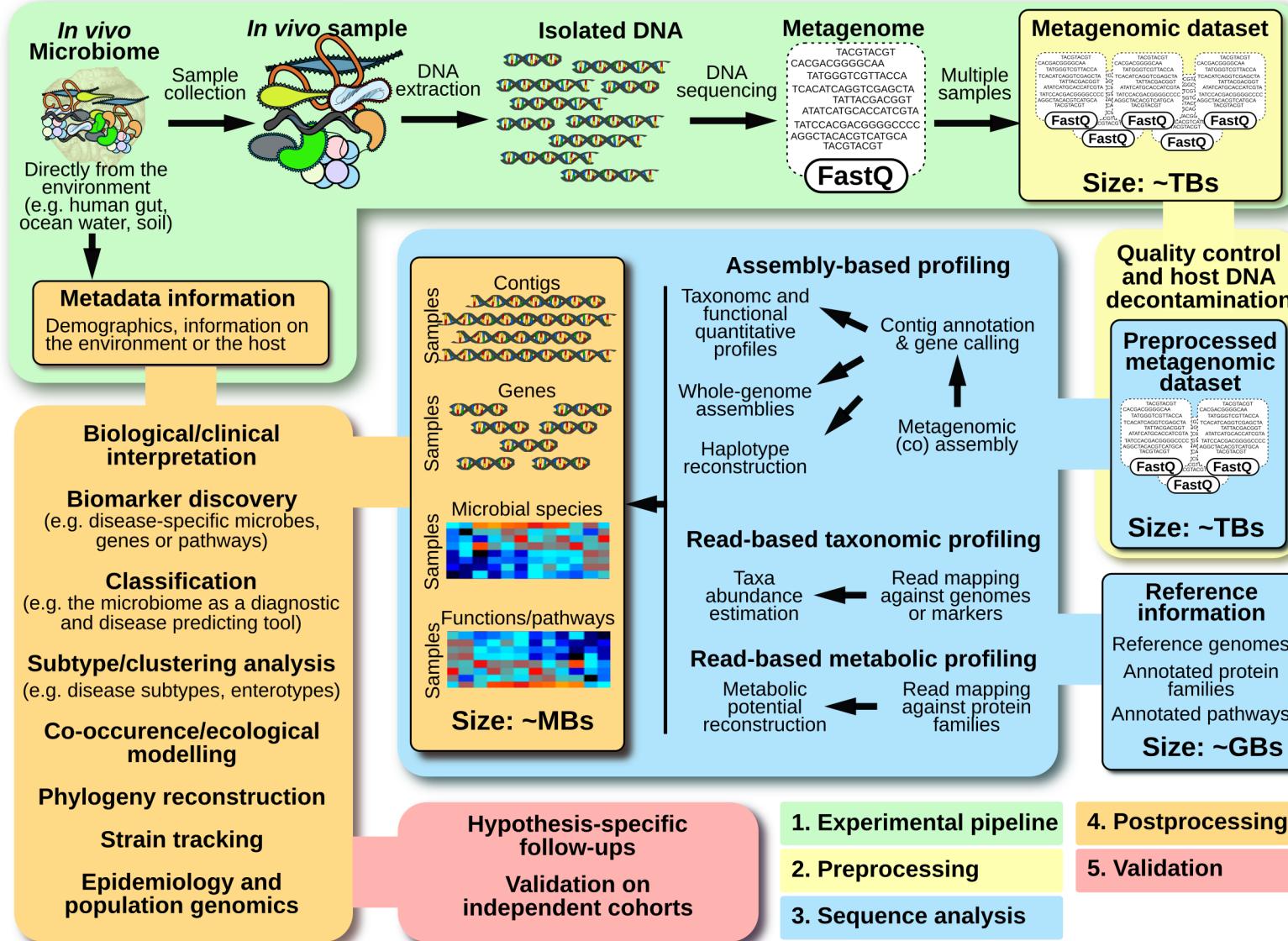


Contig Binning

Christopher Quince
Warwick Medical School

Introduction

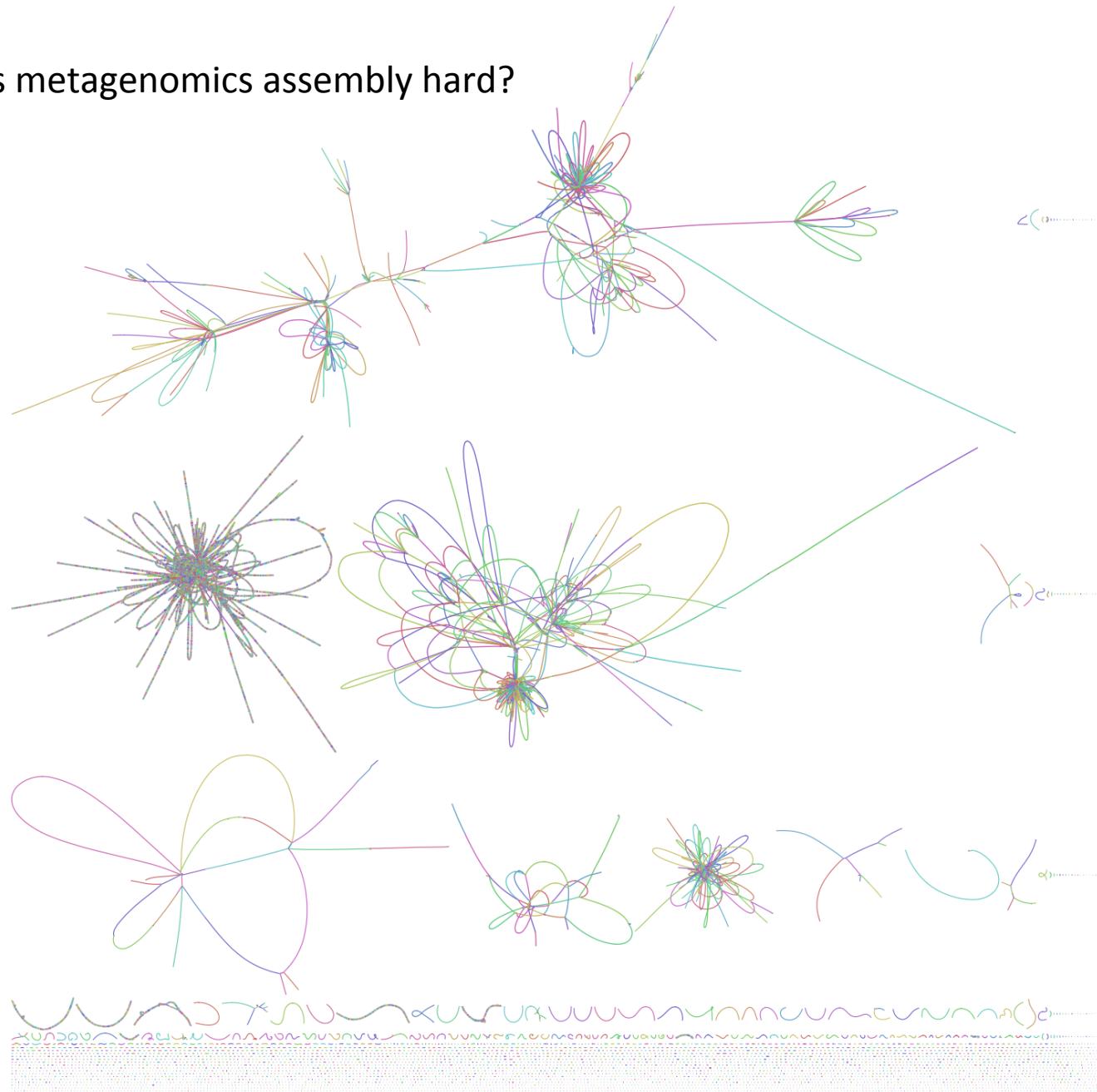
- Read based metagenome analysis throws away a lot of information
- Lose link between function and populations
- In theory that information is present even in short read next generation sequence data
- Genome resolved metagenomics aims to extract genome complements and their abundances



Overview

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

Why is metagenomics assembly hard?



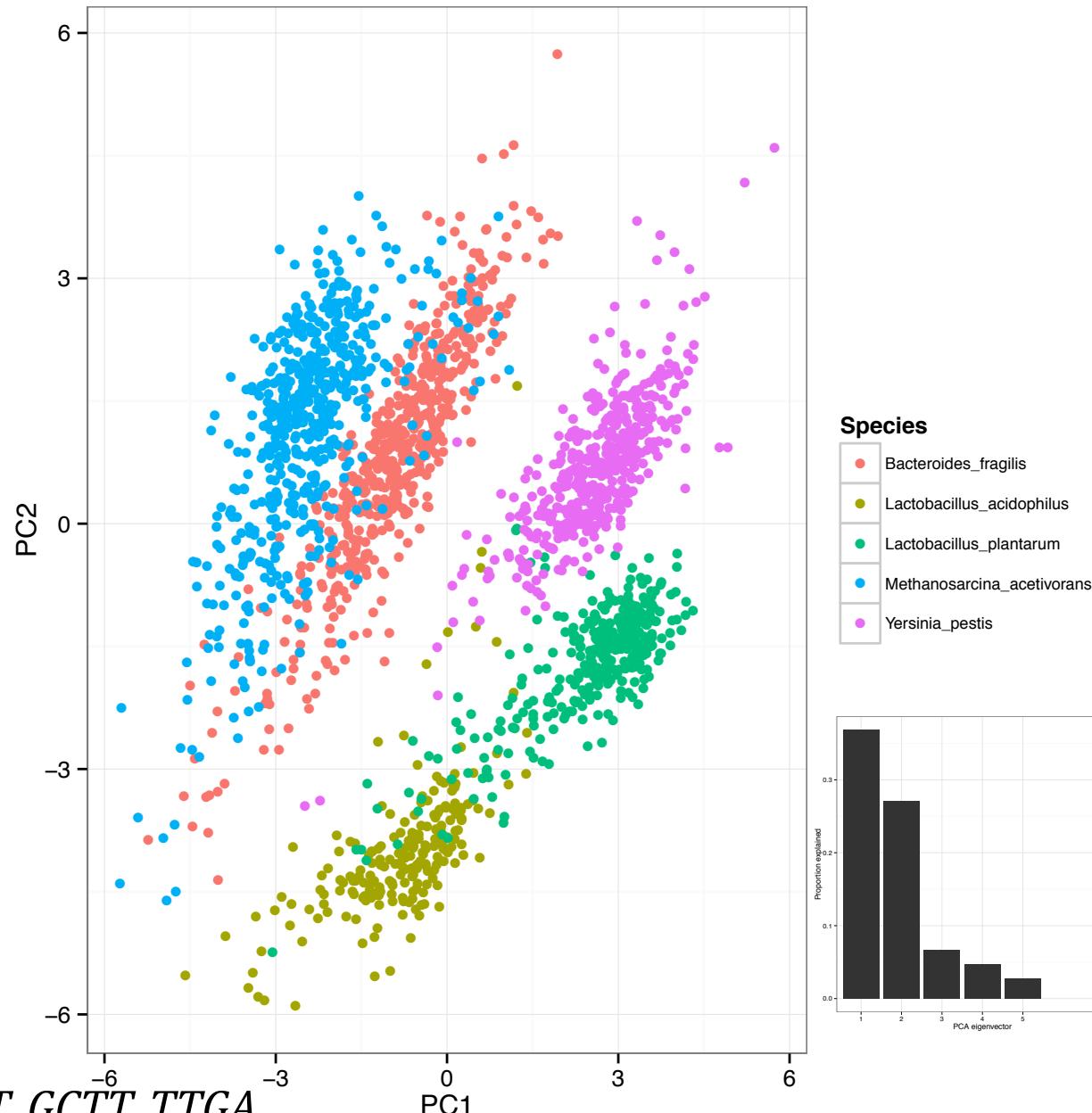
Use contig binning to **cluster** contigs back into strain/species genomes

Contig clustering by sequence composition

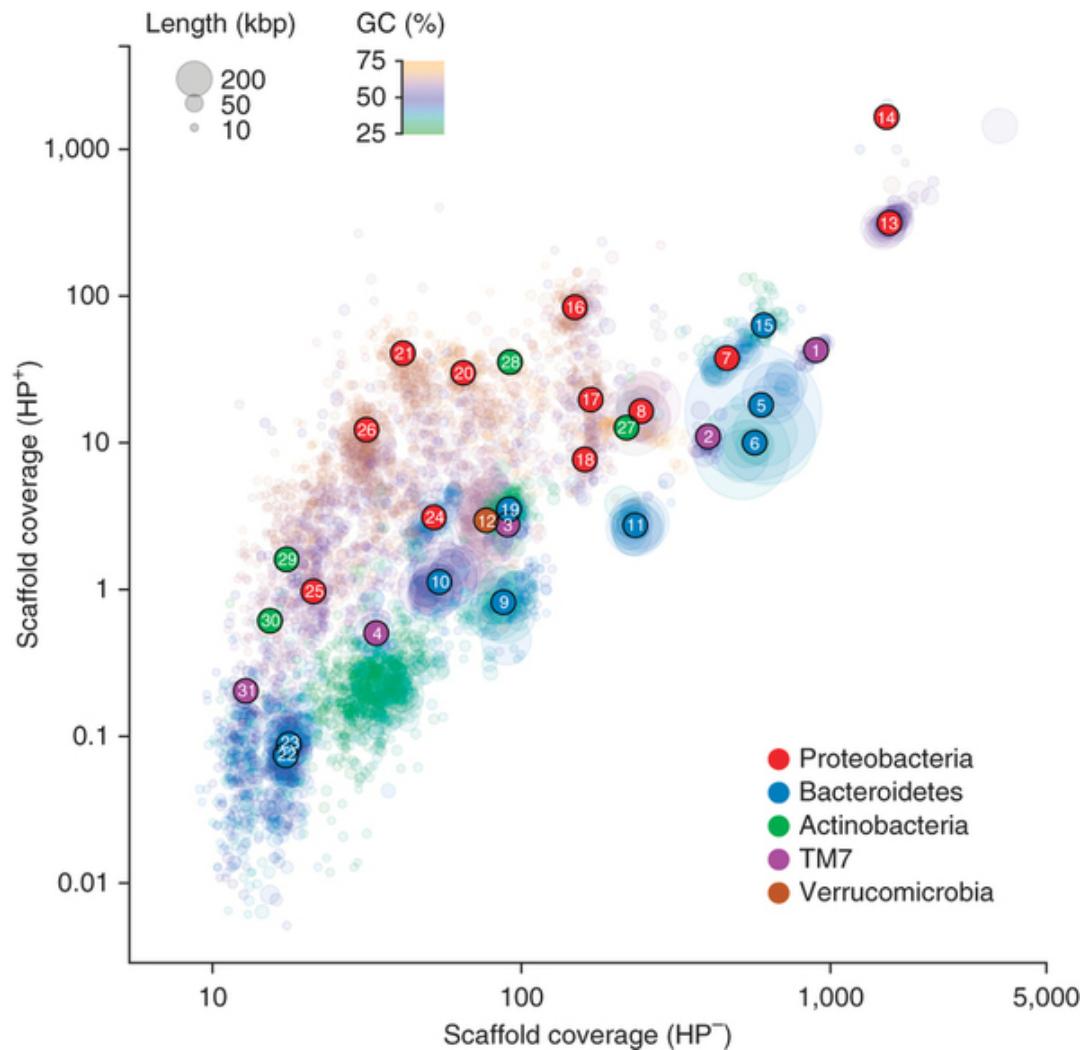
- 1) Different species have characteristic signatures of overlapping kmer frequencies ([Sandberg et al. Genome Res. 2001](#))
- 2) Fragmented five genomes into 10kb
- 3) Counted tetramers and added pseudocount
- 4) Calculated log-proportions
- 5) Generated PCA

sequence = $(\overline{CTGG} \overline{CTTG})$

4-mers: 2×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 catalogue ([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](#))



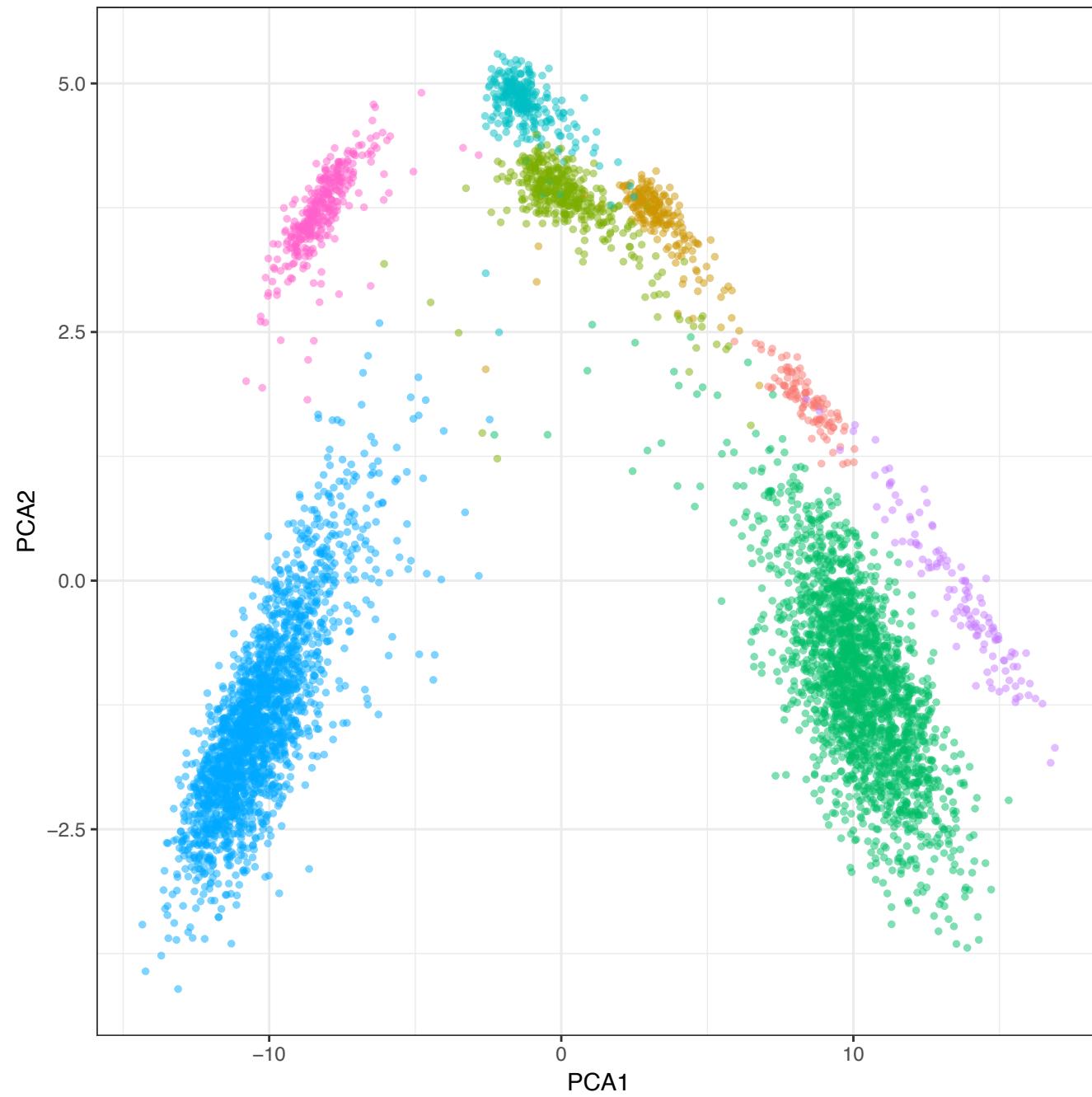
CONCOCT: Clustering cONtigs on COverage and ComposiTion



Data pre-processing

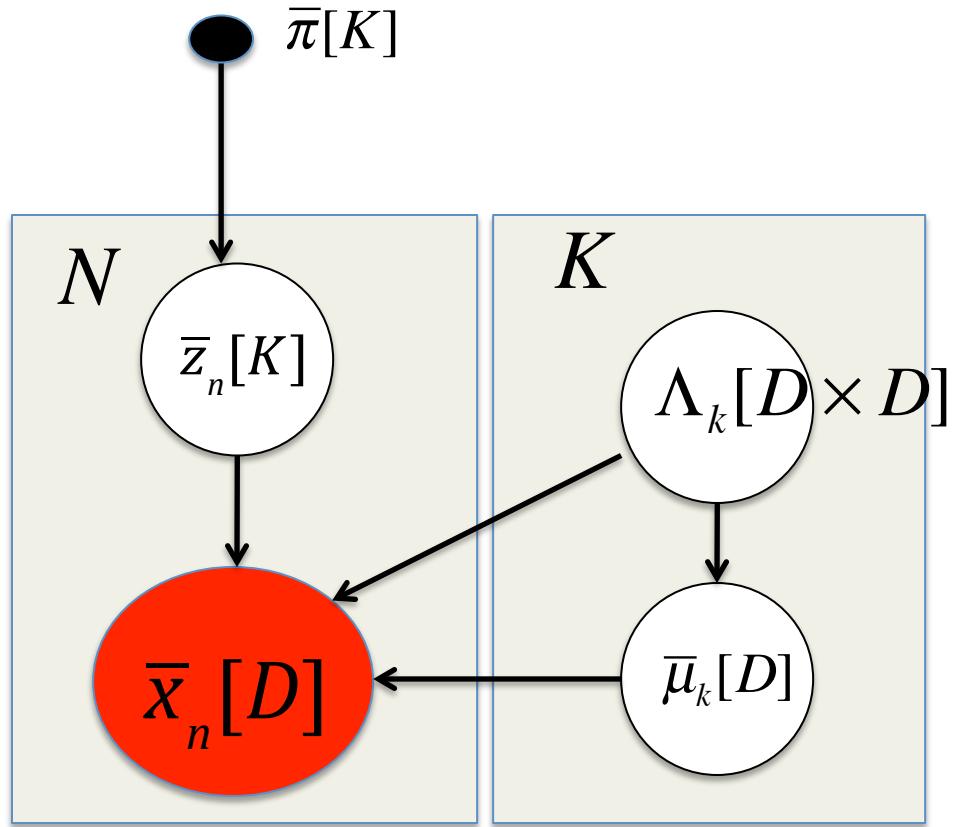
- Perform coassembly across all samples
- Fragment contigs greater than 10kb and map reads back to N contigs to get mean coverage of contig in each of M samples
- Generate k-mer frequency vector for each contig
- Add pseudo-counts, normalise coverage and k-mer frequencies, join and log-transform
- Perform 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](http://www.biorxiv.org/content/early/2017/01/09/099127)

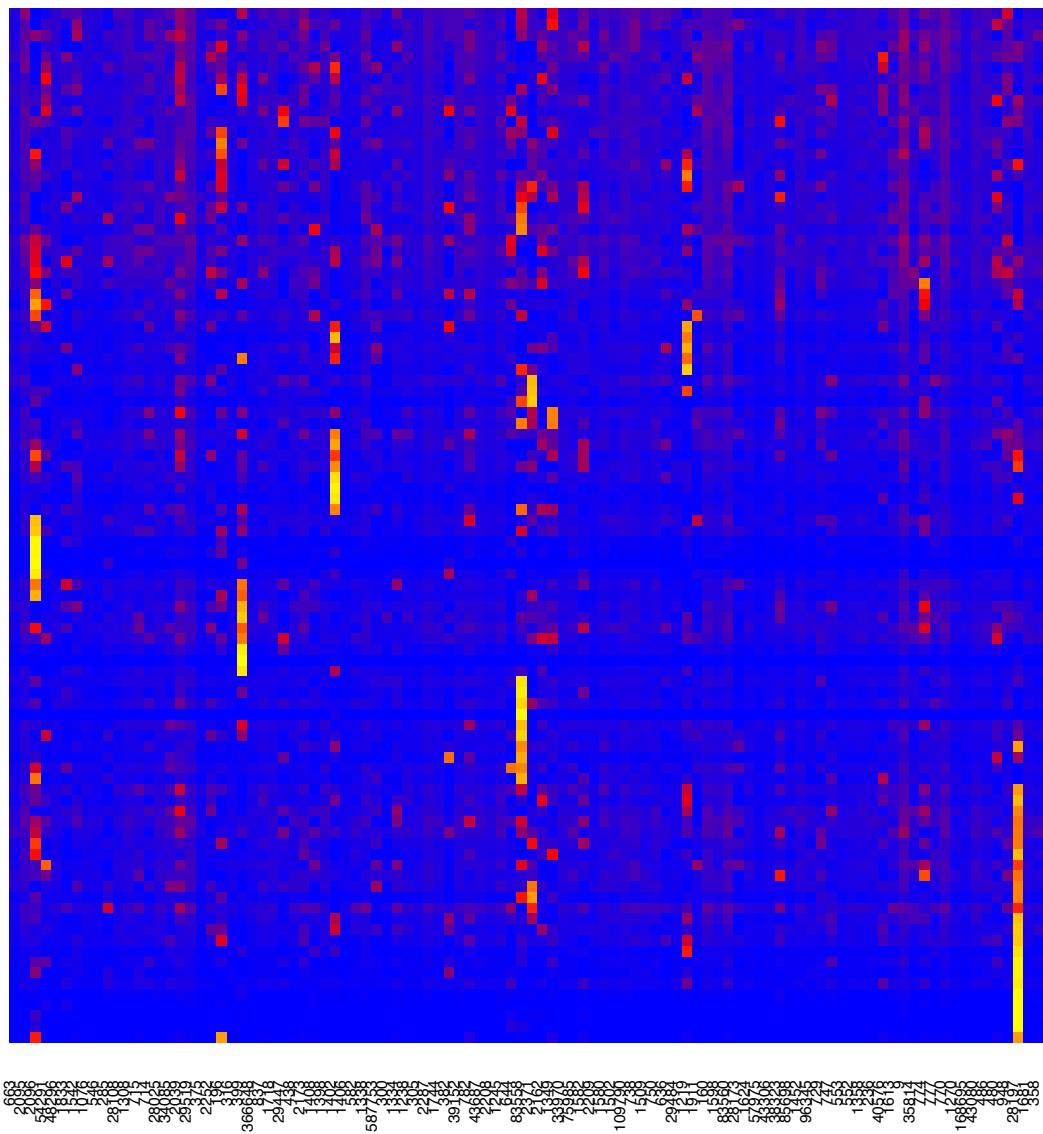
Complex synthetic community

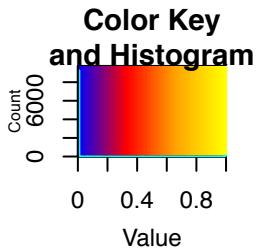
- 100 different species and 210 NCBI genomes
- 10 ten separate phyla, 49 families, and 74 genera
- Species strain frequency distribution of (1:50, 2:20, 3:10, 4:10, 5:10)
- Simulated 96 samples of 6.25 million 2X150 bp paired end reads using ART: 1 HiSeq 2500 high output:

<https://github.com/chrisquince/StrainMetaSim>

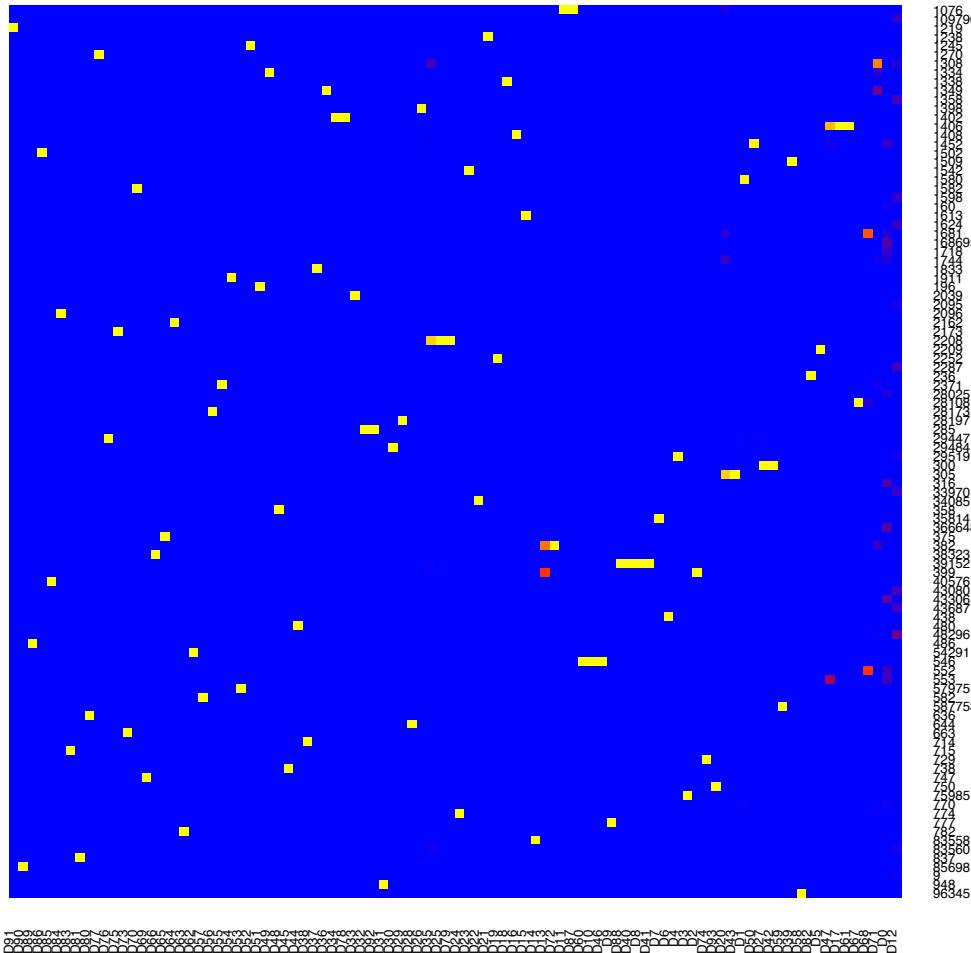
Complex synthetic community

- 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



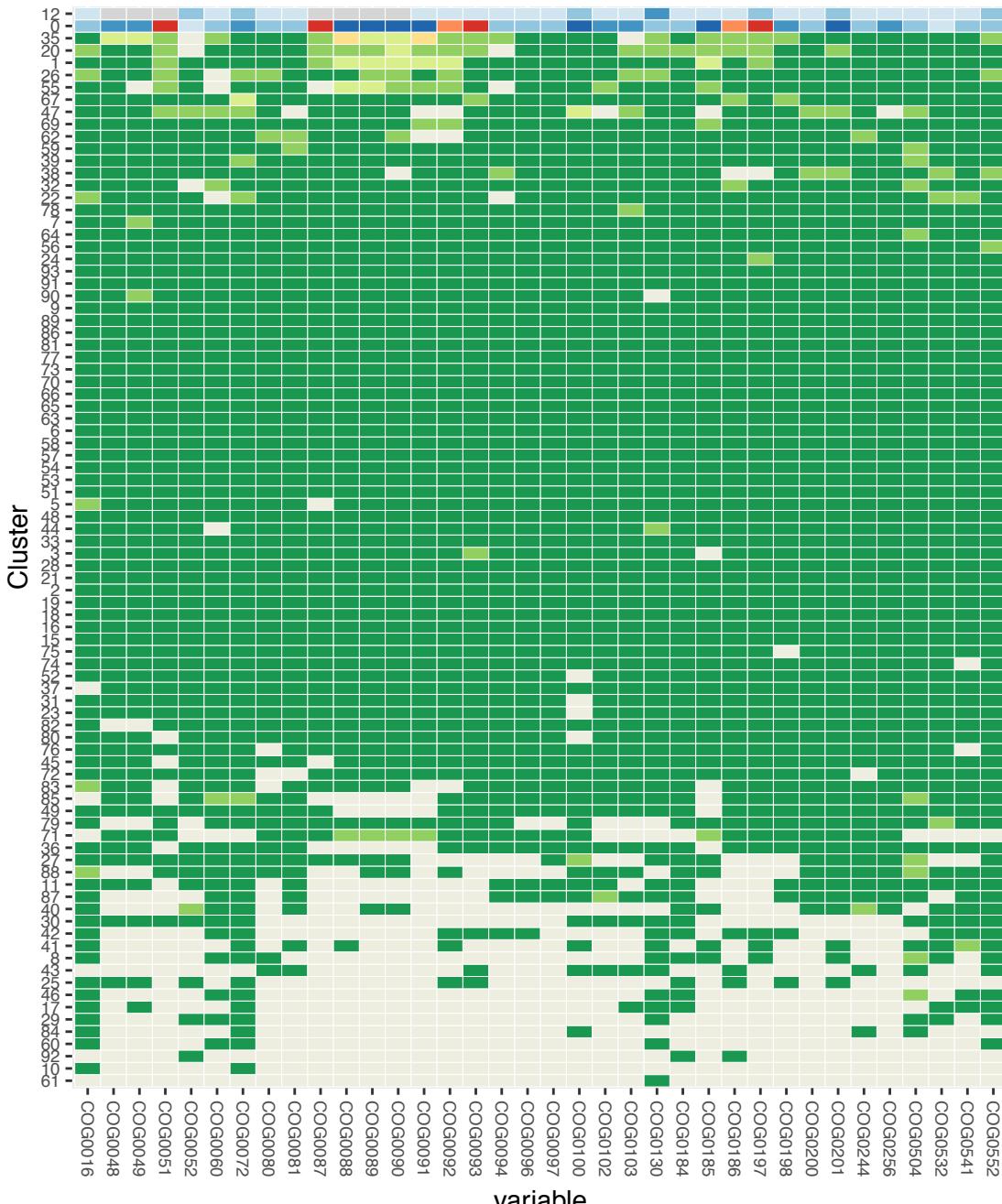


Concoct binning



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

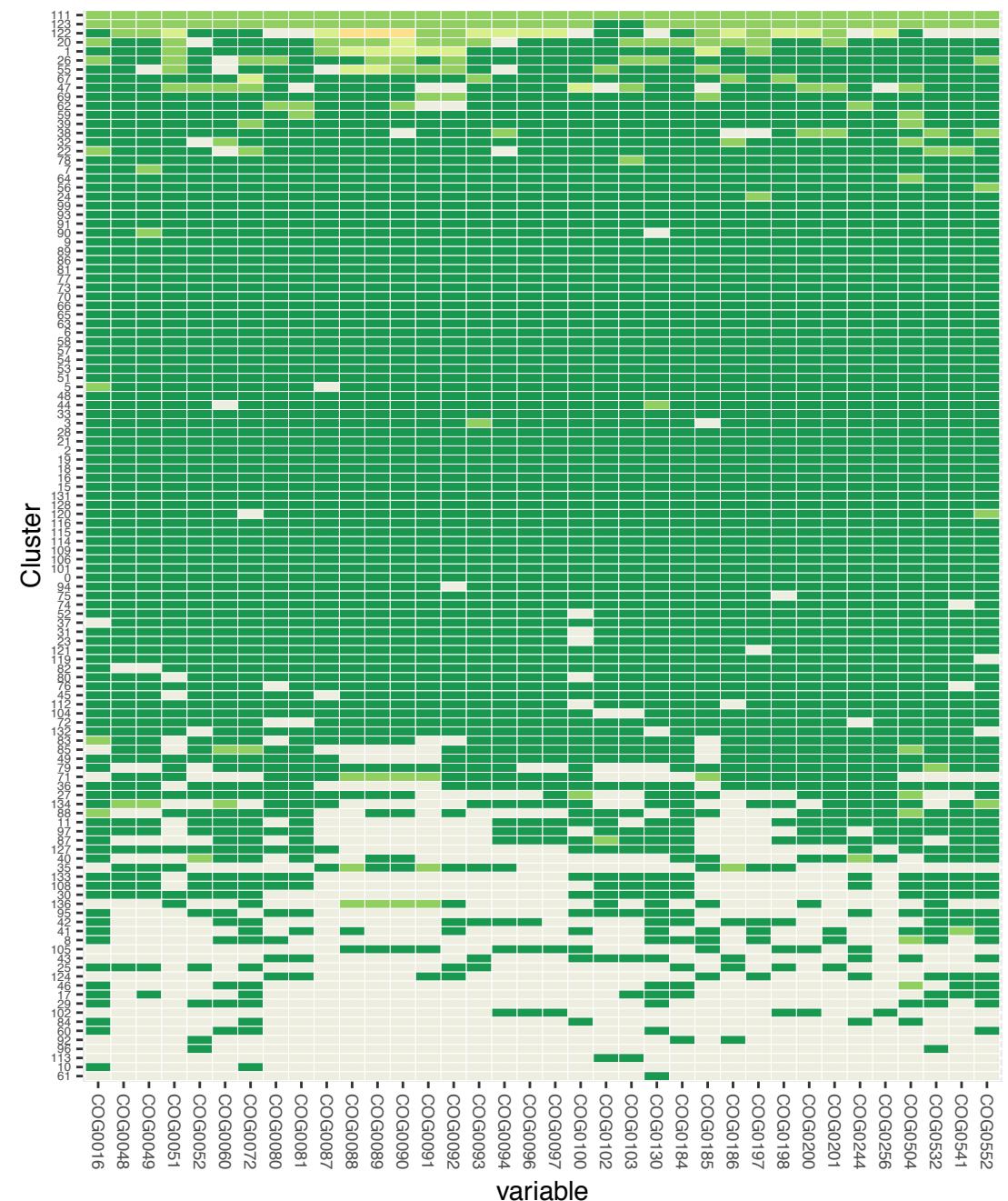
Concoct single-copy core genes



- 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

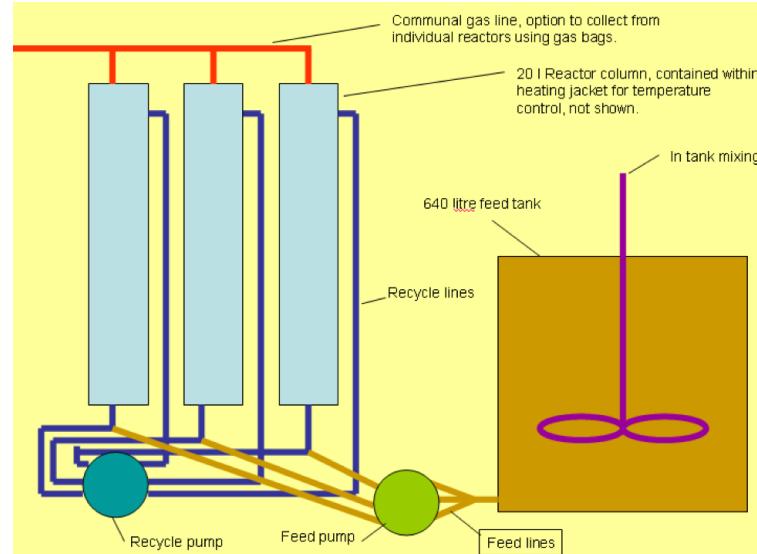


Complex synthetic community comparison

- 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



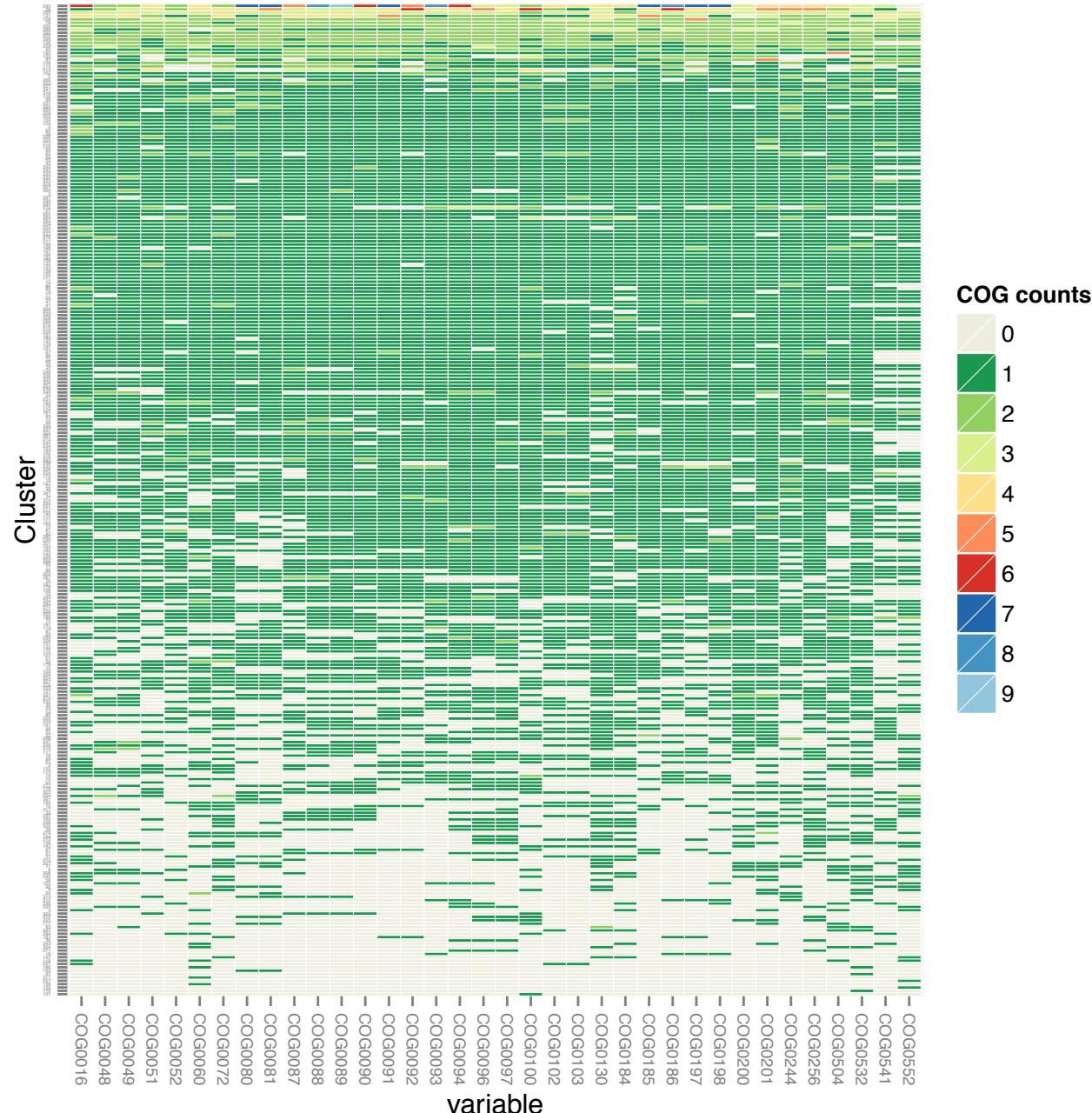
Metagenomics AD sequencing

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)

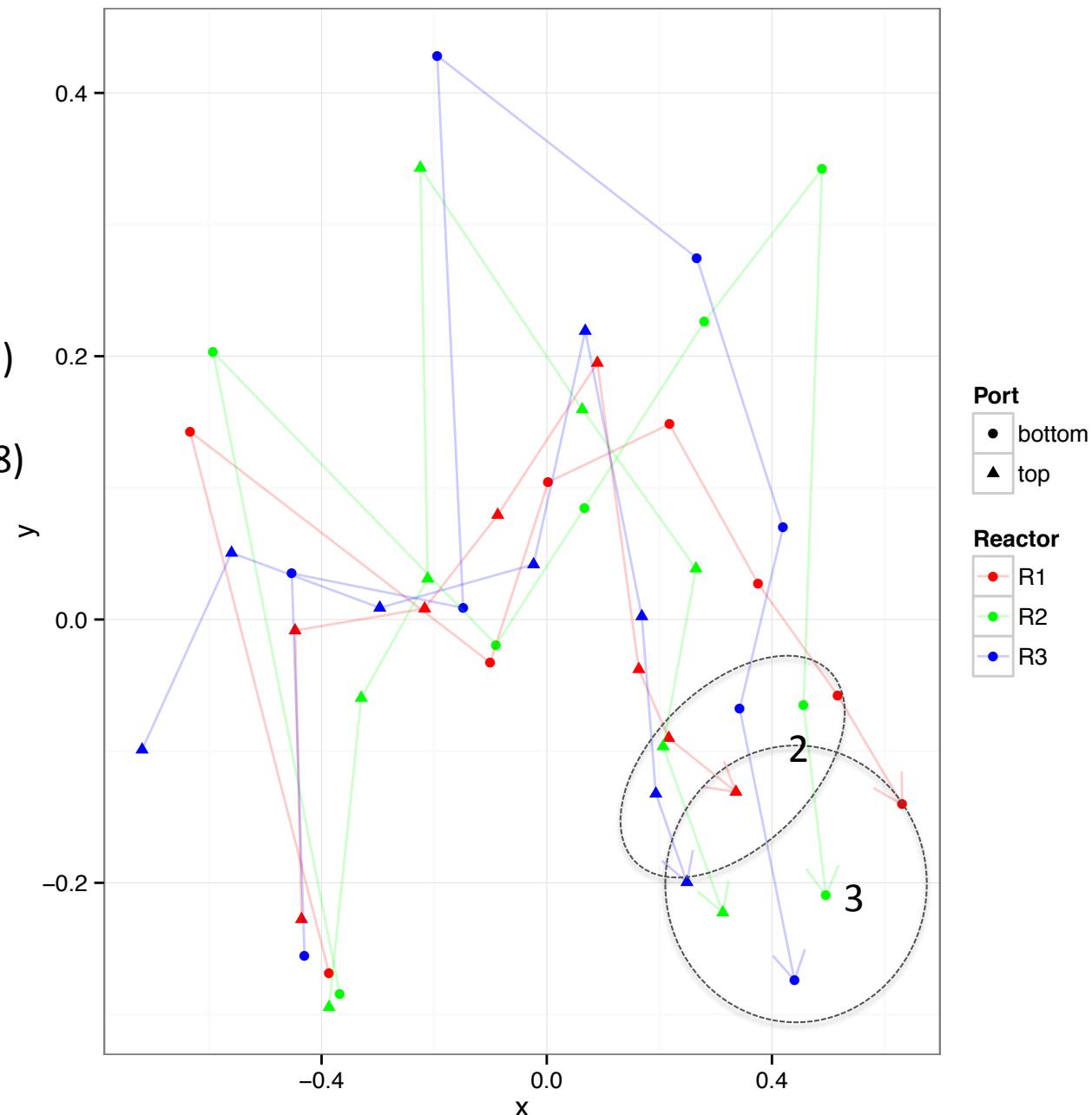
Metabat - 123
MaxBin2 - 123

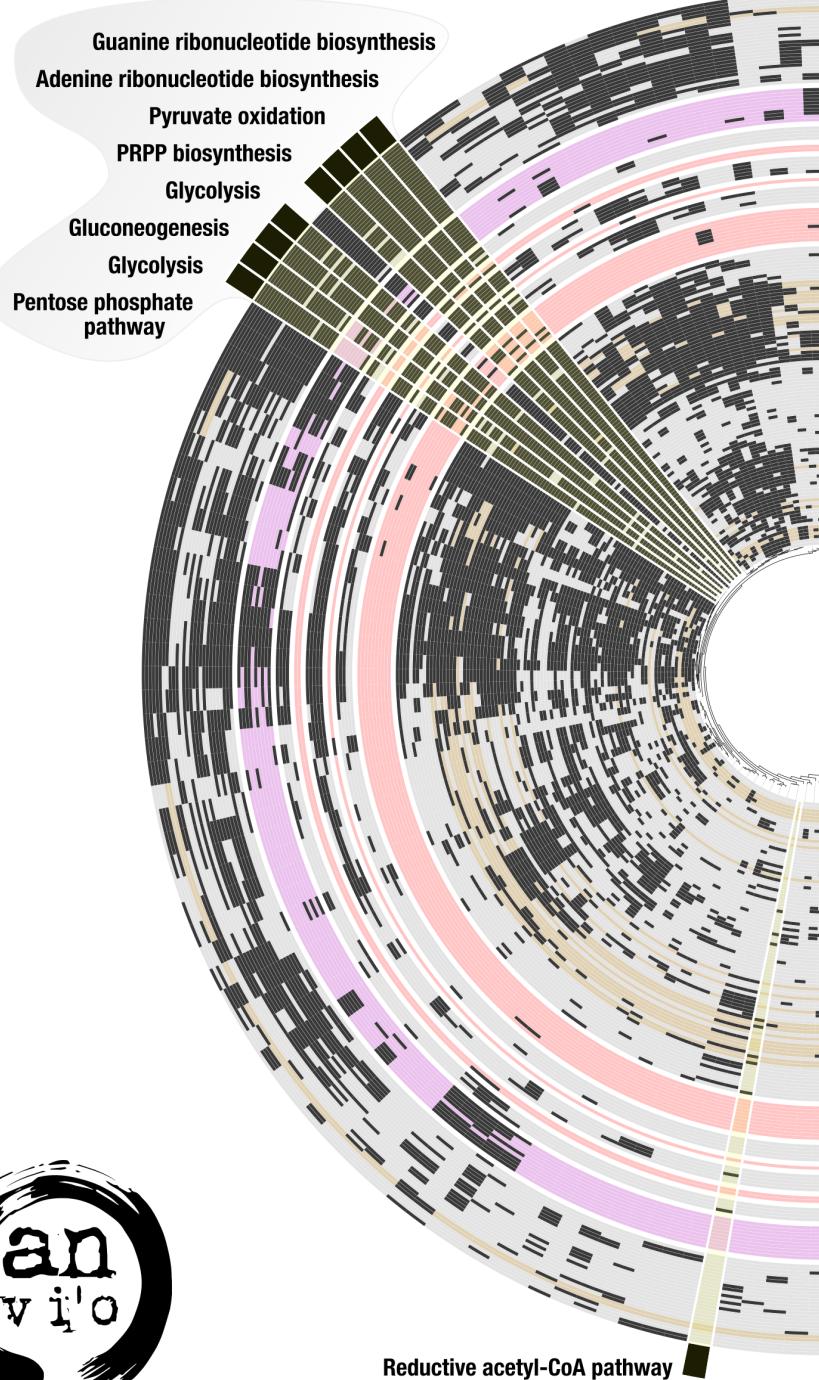


Based on NMDS of mean cluster abundances observe community succession over time

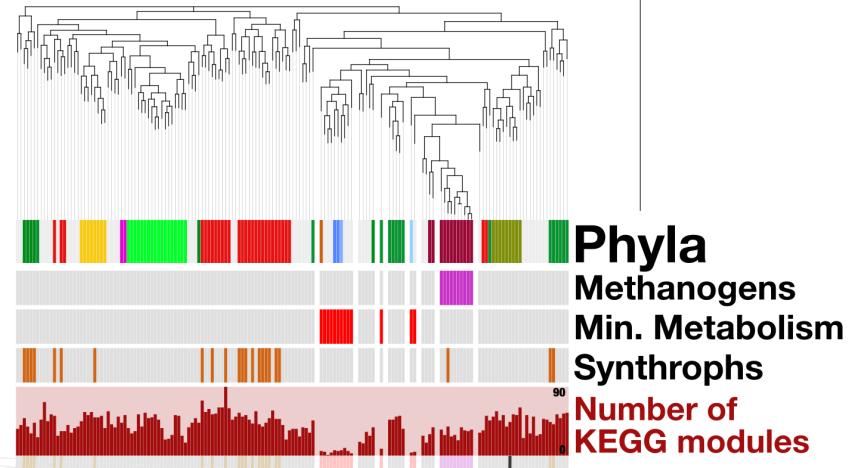
Time explained 36% ($p = 0.01$)

Reactor marginal 8% ($p = 0.08$)





Chloroflexi Proteobacteria Firmicutes Actinobacteria
Candidatus Wolfebacteria Candidatus Campbellbacteria
Planctomycetes Bacteroidetes Candidatus Berkelbacteria
Spirochaetes Euryarchaeota Deinococcus-Thermus

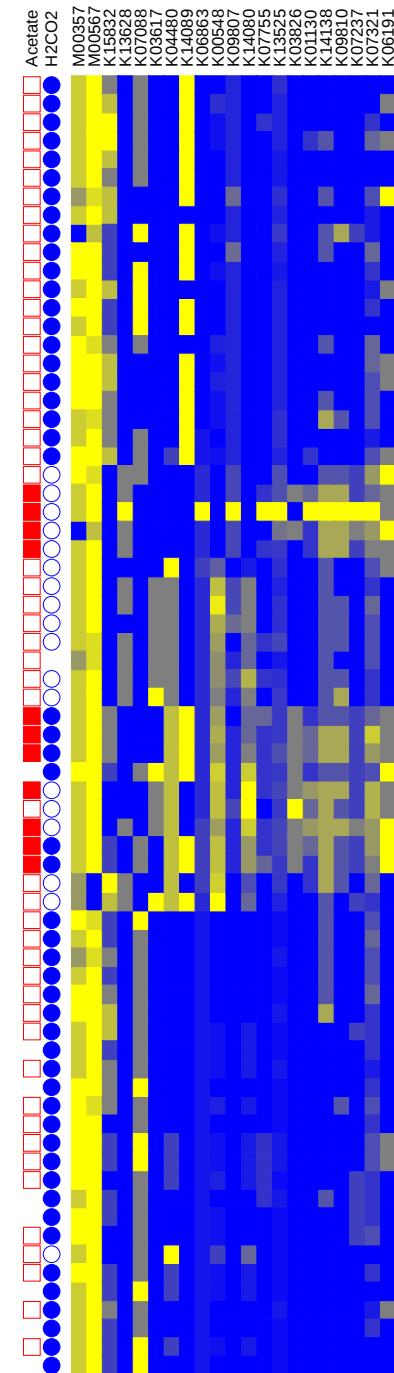
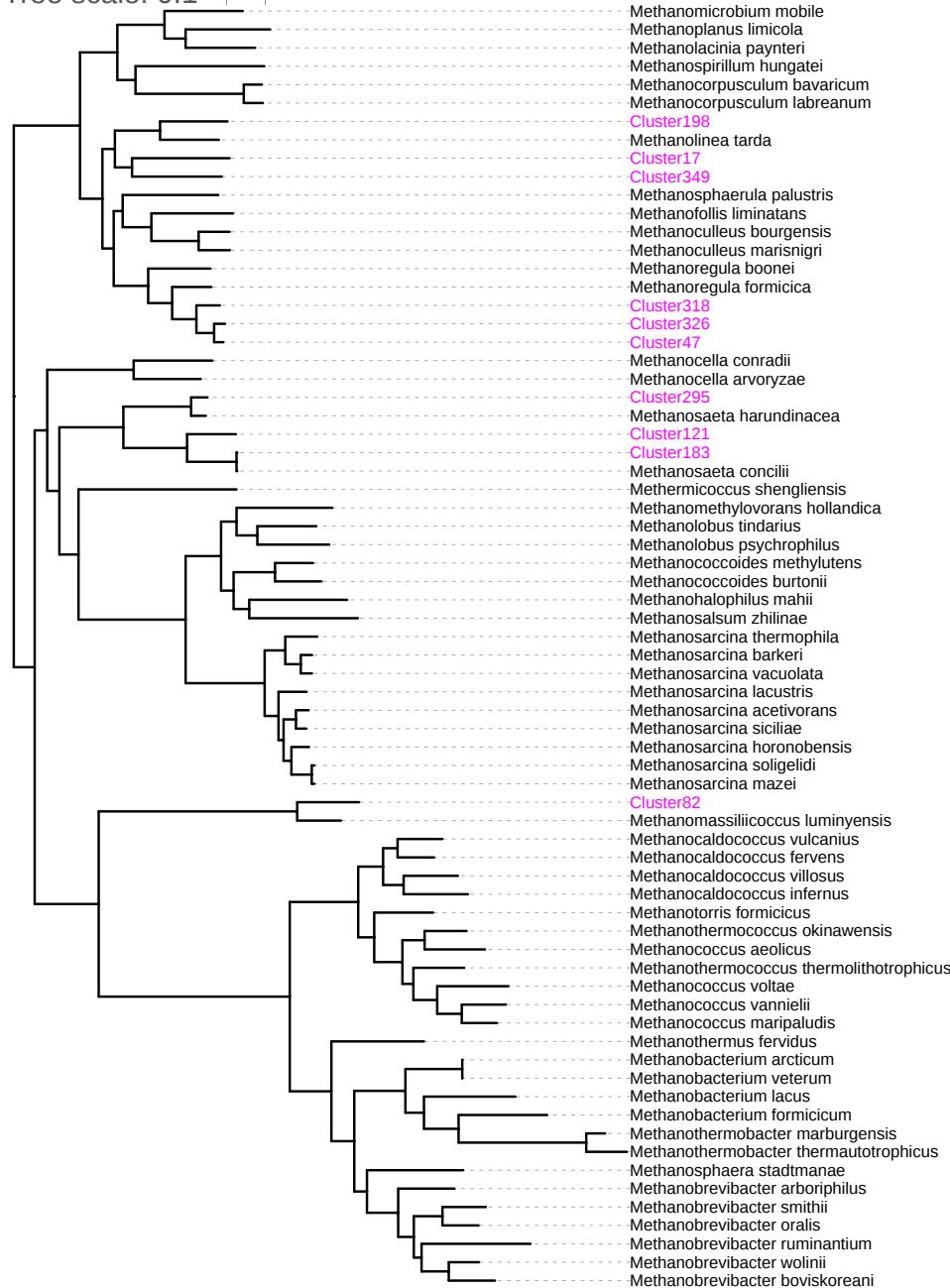


12 MAGs with minimal metabolism, 9 from Candidate Phyla Radiation (CPR) – Brown et al. Nature 2015

Acetoclastic methanogenesis
Autotrophic methanogenesis

Dissimilatory sulfate reduction

Tree scale: 0.1



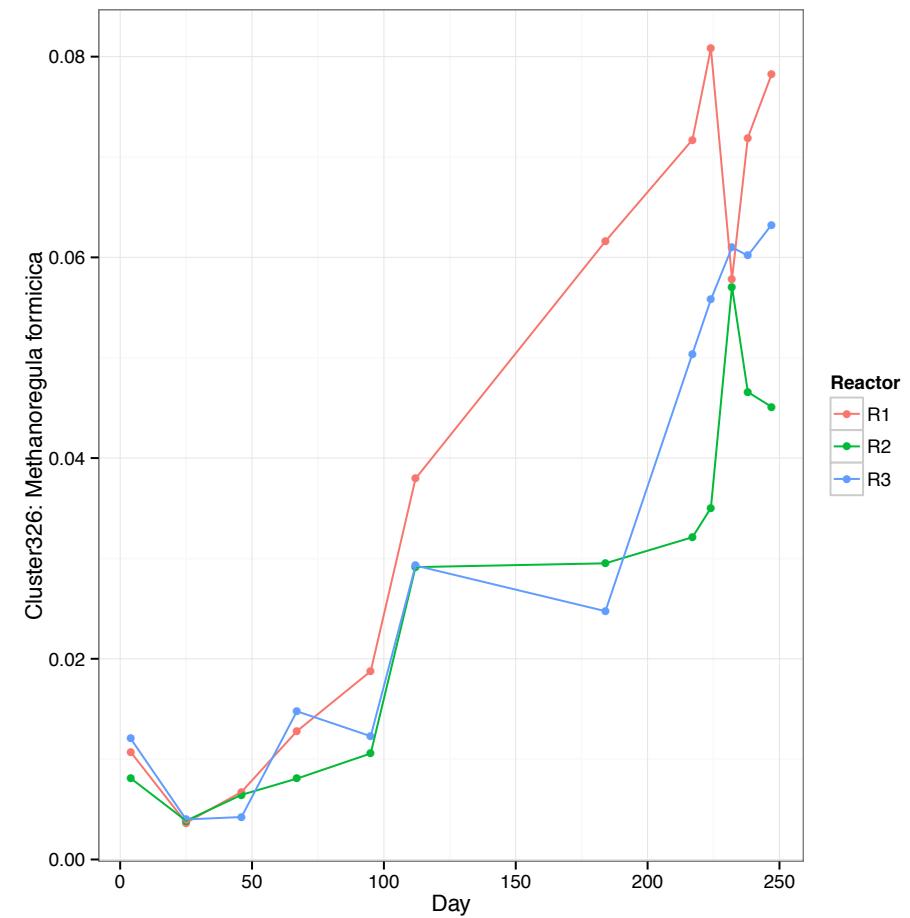
Trained Random Forest classifier on known methanogen genomes with known substrates

Predict known autotrophs or acetoclasts with 5% error rate

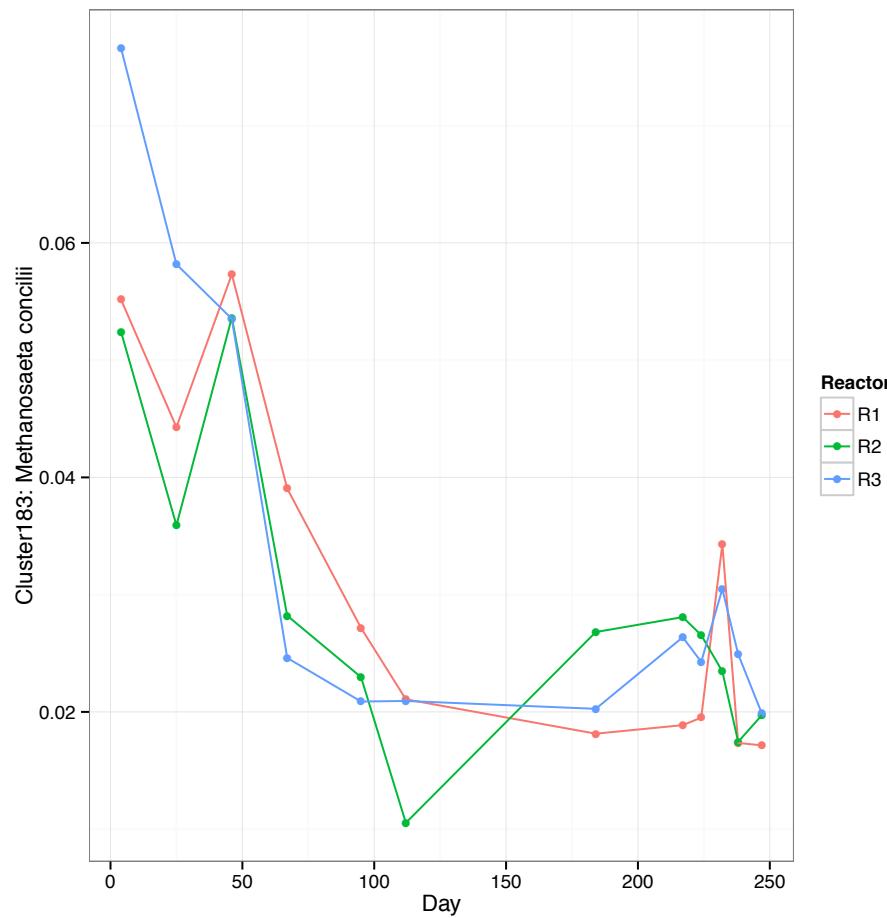
Apply to MAGs

Also provides information on what orthologs may be involved in methanogenesis

Autotroph



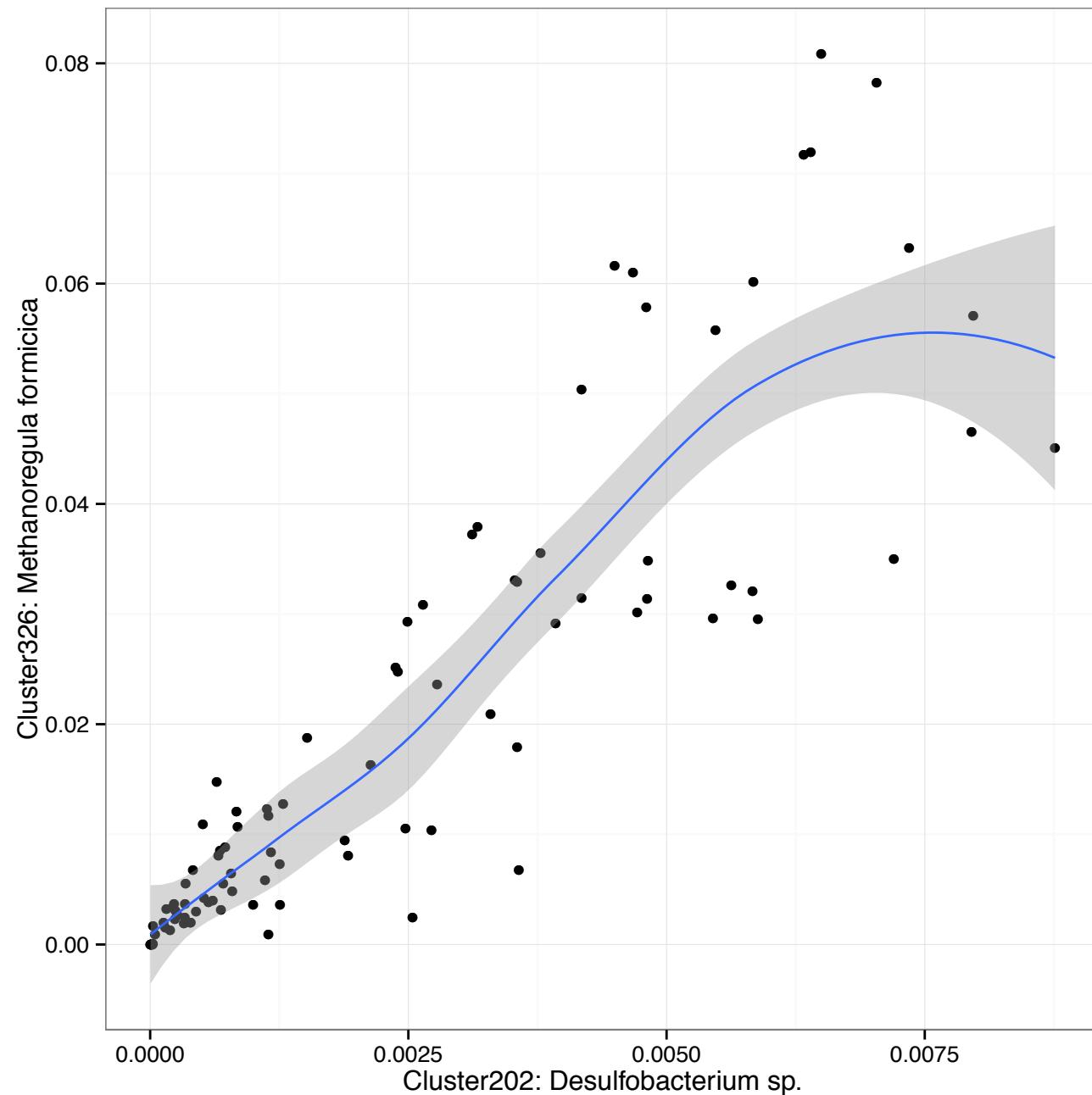
Acetoclast



Shift from acetoclastic to autotrophic methanogenesis over time

Driven by syntrophies...

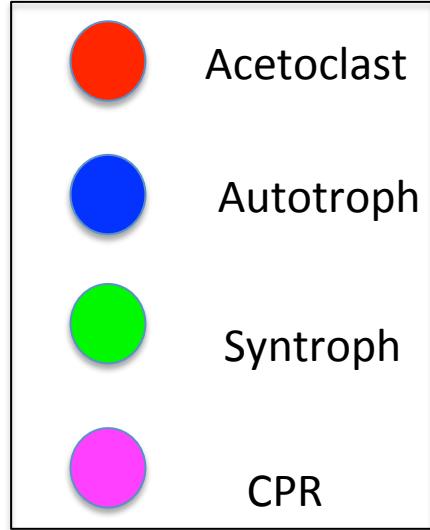
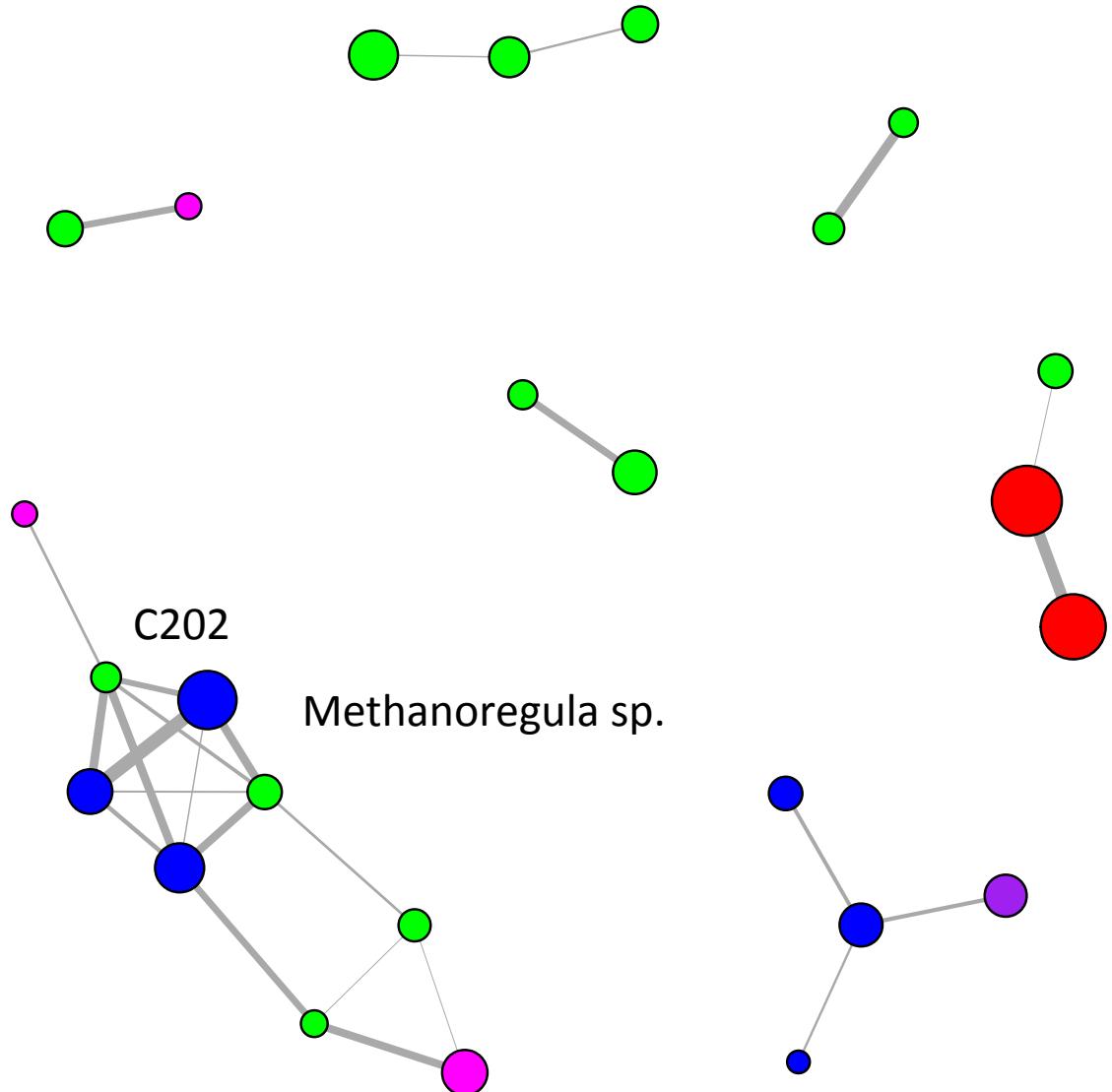
Possible syntroph



Cluster202:
Desulfobacterium
sp.

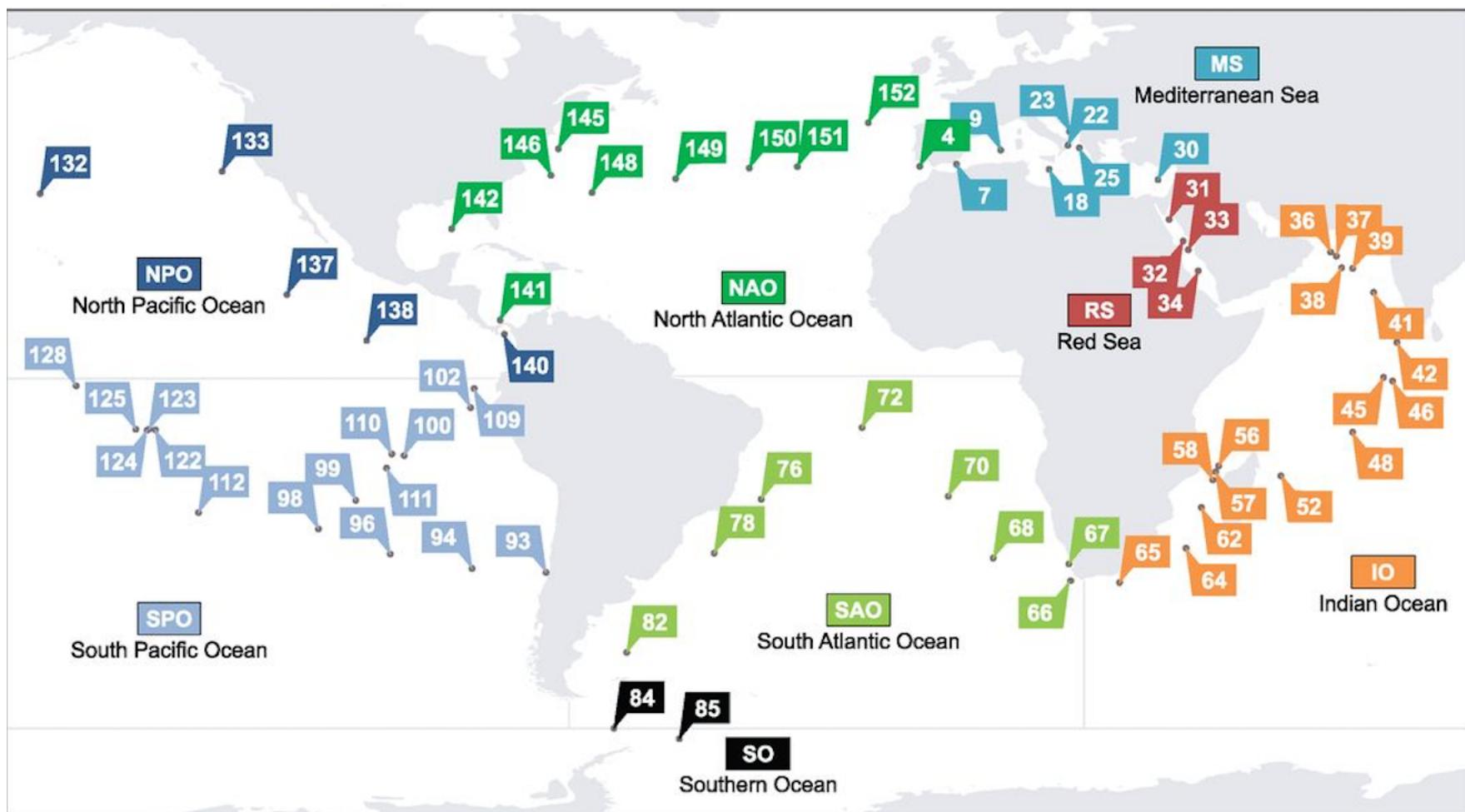
Sulphate reducer
but possesses 3
copies of
ko:K02380 Formate
dehydrogenase
maturation protein
FdhE characteristic
of syntrophs
(Worm et al. 2014)

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



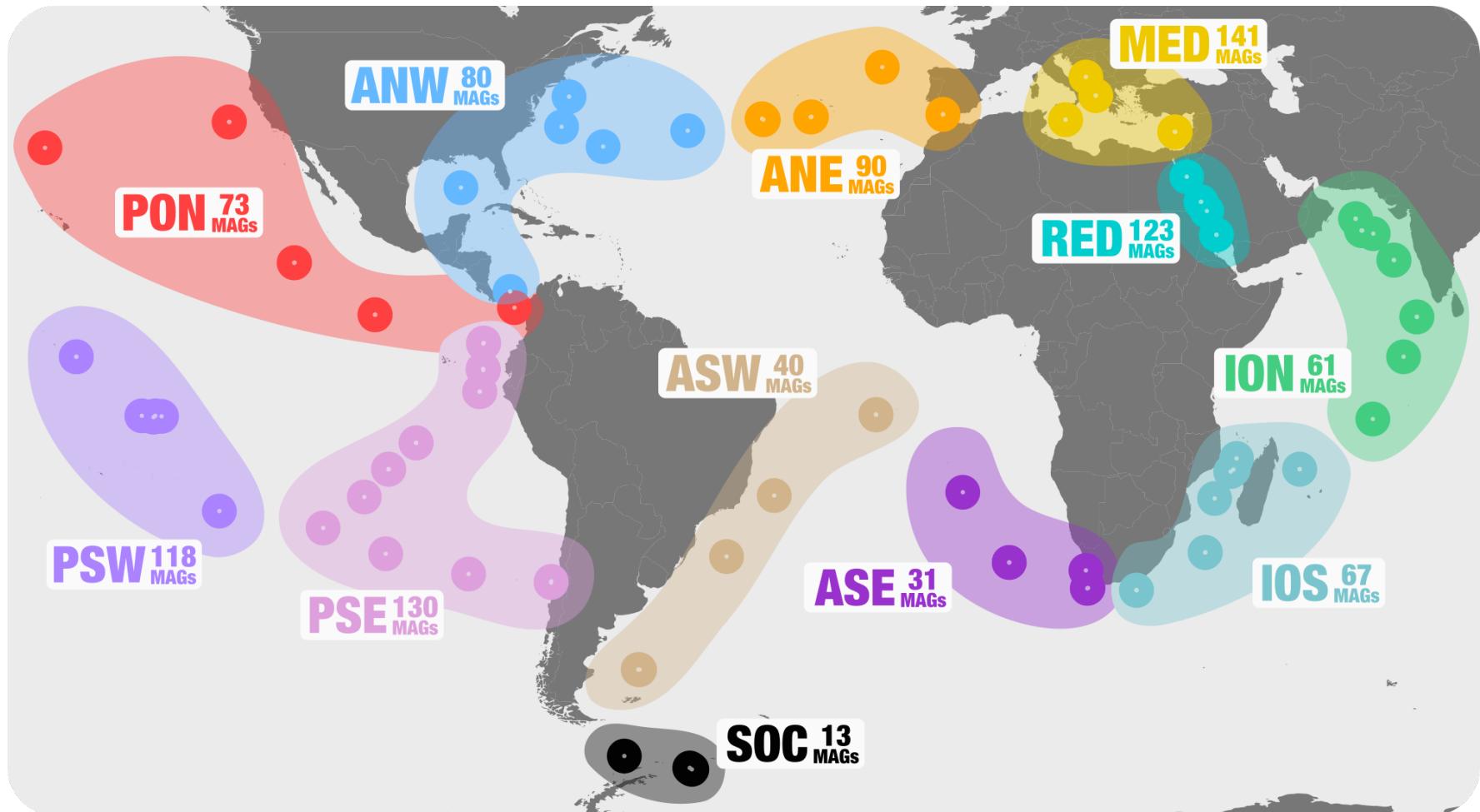
Network
formed from
very strong
positive
correlations
only $r > 0.8$

TARA Oceans sampling sites: Sunagawa et al. Science 2015



The 93 TARA Oceans metagenomes we analyzed represent the planktonic size fraction (0.2-3 μ m) of 61 surface samples and 32 samples from the deep chlorophyll maximum layer of the water column

TARA Oceans: Sunagawa et al. Science 2015



- 93 metagenomes from the planktonic size fraction for which we performed 12 metagenomic co-assemblies
- Generated 957 non-redundant MAGs encompassing the three domains of life

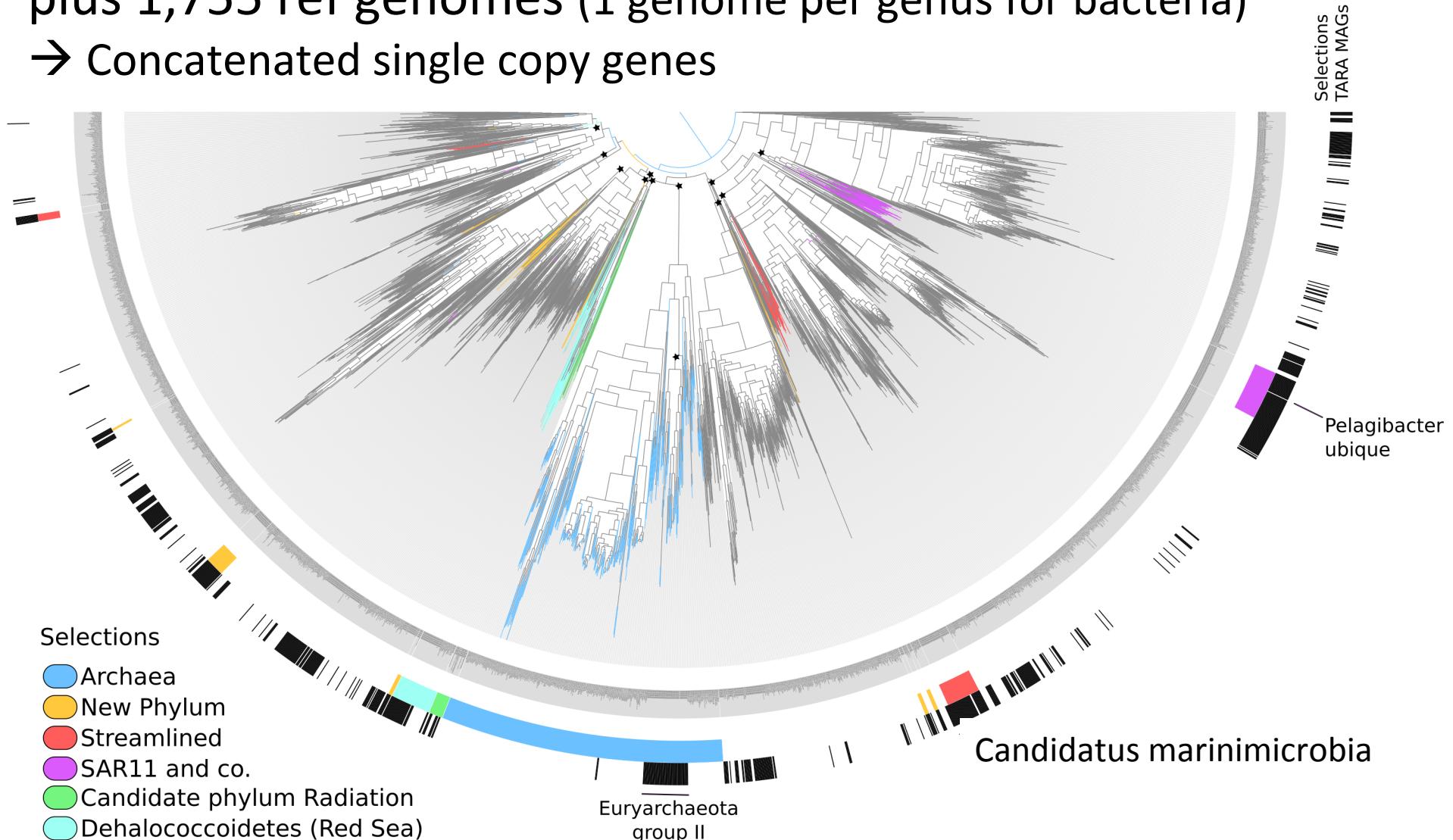
Delmont et al Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in the surface ocean <http://biorxiv.org/content/early/2017/04/23/129791>



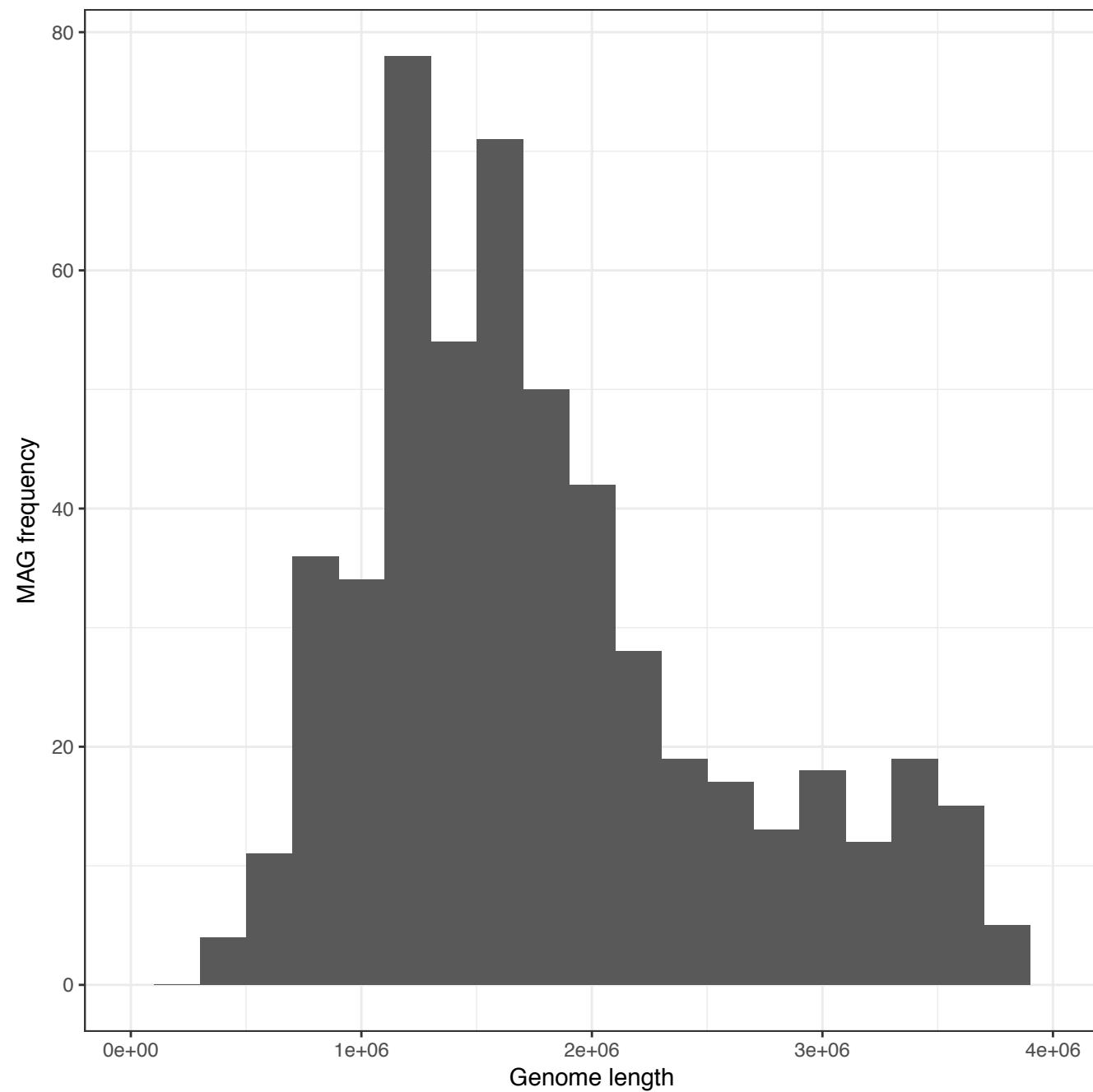
Phylogenetic analysis of 660 MAGs

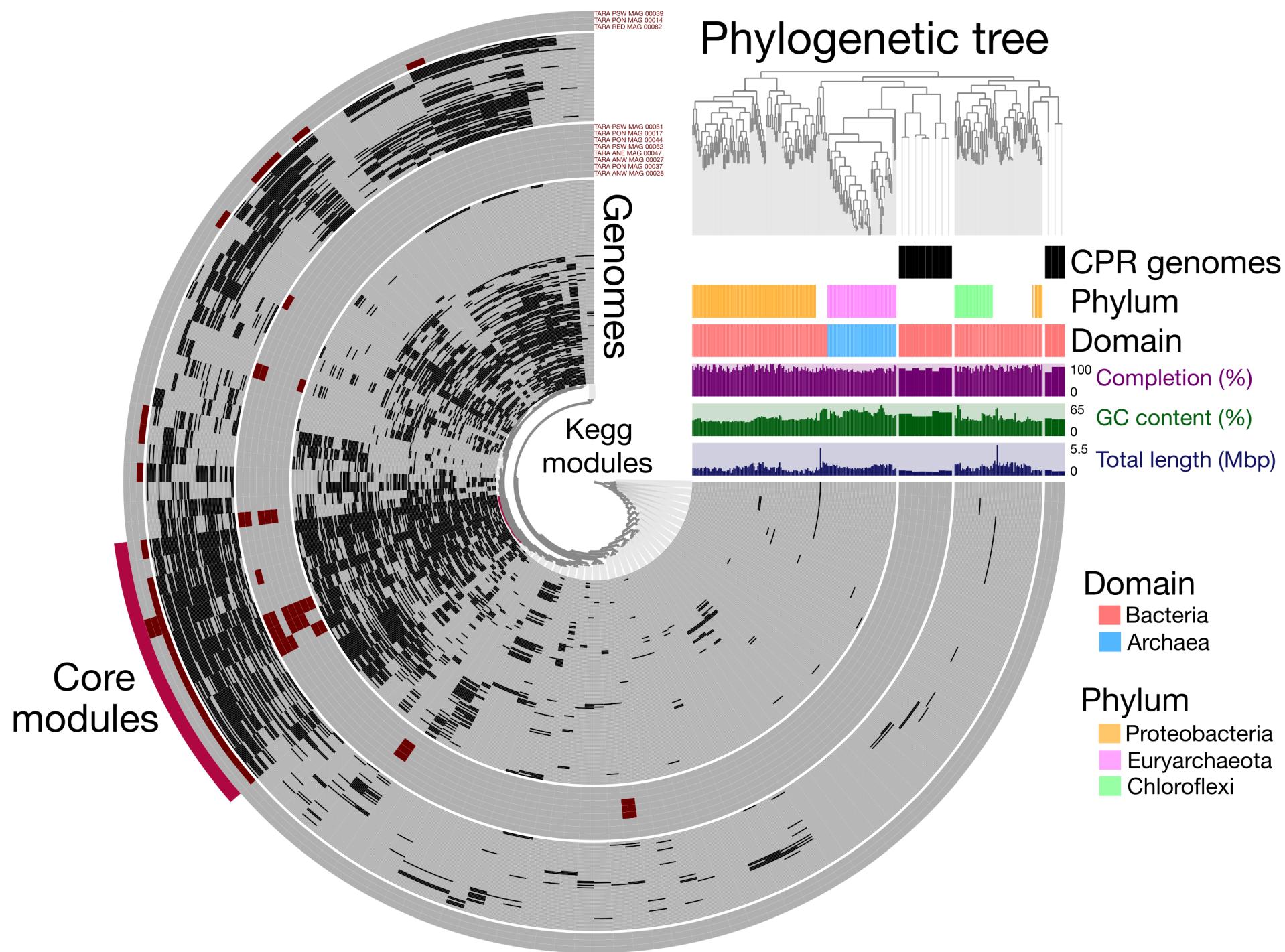
plus 1,755 ref genomes (1 genome per genus for bacteria)

→ Concatenated single copy genes



Delmont et al Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in the surface ocean <http://biorxiv.org/content/early/2017/04/23/129791>





Oceanic heterotrophic bacterial diazotrophs (HBDs)

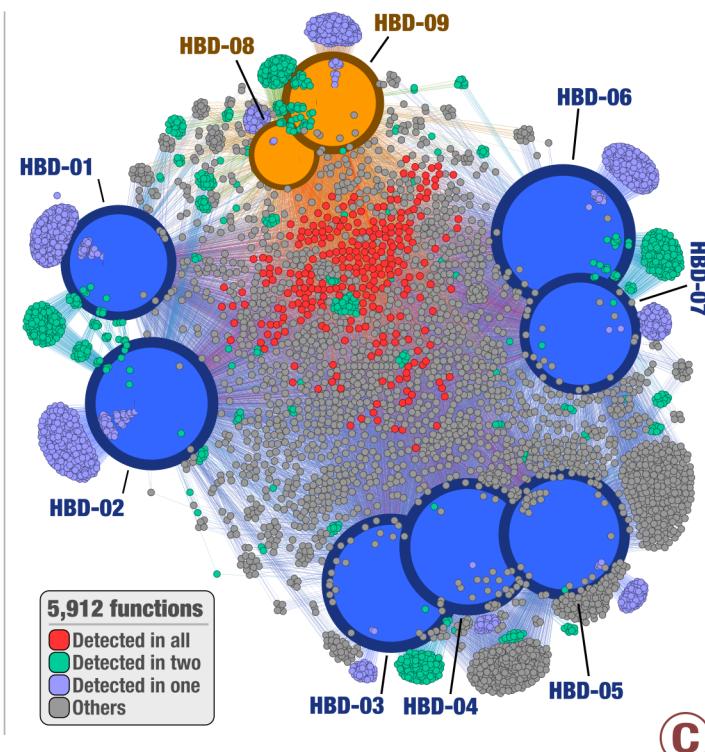
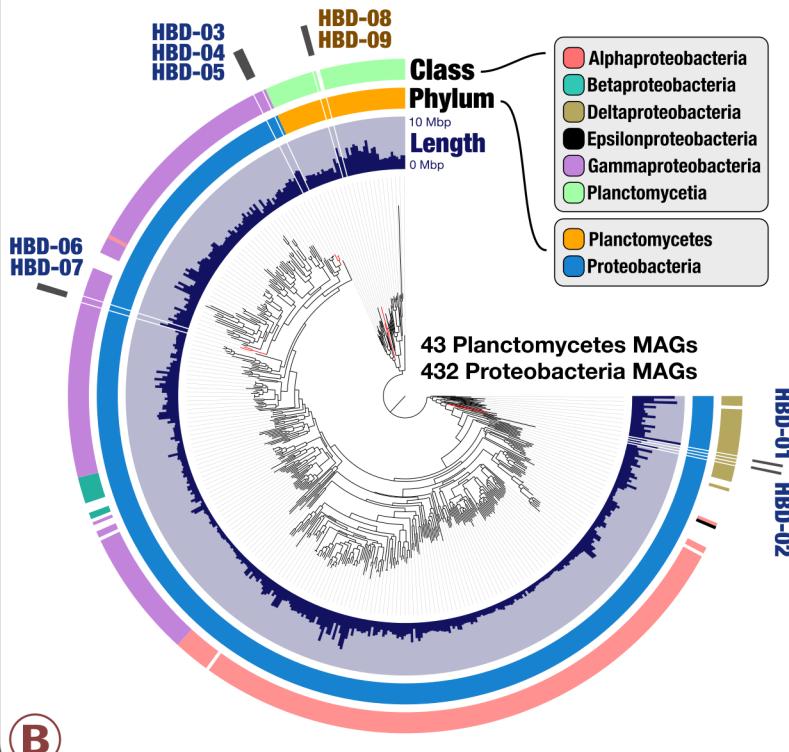
- Nitrogen fixation in the surface regulates microbial primary productivity and the sequestration of carbon through the biological pump
- Cyanobacteria populations have long been thought to represent the main suppliers of the bio-available nitrogen
- Nitrogenase reductase gene surveys revealed the existence of non-cyanobacterial populations that can also fix nitrogen - heterotrophic bacterial diazotrophs (HBDs)

Delmont et al Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in the surface ocean <http://biorxiv.org/content/early/2017/04/23/129791>

- Search for MAGs containing the catalytic (*nifH*, *nifD*, *nifK*) and the biosynthetic proteins (*nifE*, *nifN* and *nifB*) required for nitrogen fixation
- One of six Cyanobacteria
- Nine other MAGs possessed all six genes

A

	Region of Recovery	Length (Mbp)	N50	Num Contigs	Num Genes	GC Content	Percent Compl.	Percent Redund.	Phylum	Class	Order	Family	Genus
HBD-01	PSW	3.67	48,153	118	3,592	52.6%	97.7%	4.4%	Proteobacteria	Deltaproteo.	Desulfovibronales	Desulfovibrionaceae	Desulfovibrio
HBD-02	PSW	6.00	20,964	405	5,385	53.1%	97.1%	5.9%	Proteobacteria	Deltaproteo.	Desulfobacterales	Desulfobacteraceae	-
HBD-03	ION	4.47	57,949	110	4,155	52.4%	97.5%	8.1%	Proteobacteria	Gammaproteo.	Oceanospirillales	Oceanospirillaceae	-
HBD-04	PON	4.29	48,897	138	3,957	52.4%	89.7%	6.1%	Proteobacteria	Gammaproteo.	Oceanospirillales	Oceanospirillaceae	-
HBD-05	PSE	4.15	65,098	94	3,867	53.3%	47.7%	5.8%	Proteobacteria	Gammaproteo.	Oceanospirillales	Oceanospirillaceae	-
HBD-06	ANW	5.49	76,792	112	4,842	54.2%	98.1%	5.6%	Proteobacteria	Gammaproteo.	Pseudomonadales	-	-
HBD-07	ANW	3.99	10,488	487	3,585	58.7%	91.2%	4.3%	Proteobacteria	Gammaproteo.	Pseudomonadales	-	-
HBD-08	PSW	4.03	10,413	480	3,327	52.6%	33.5%	6.0%	Planctomycetes	-	-	-	-
HBD-09	PSW	5.86	79,495	113	4,872	50.0%	97.3%	4.6%	Planctomycetes	Planctomycetia	Planctomycetales	Planctomycetaceae	-

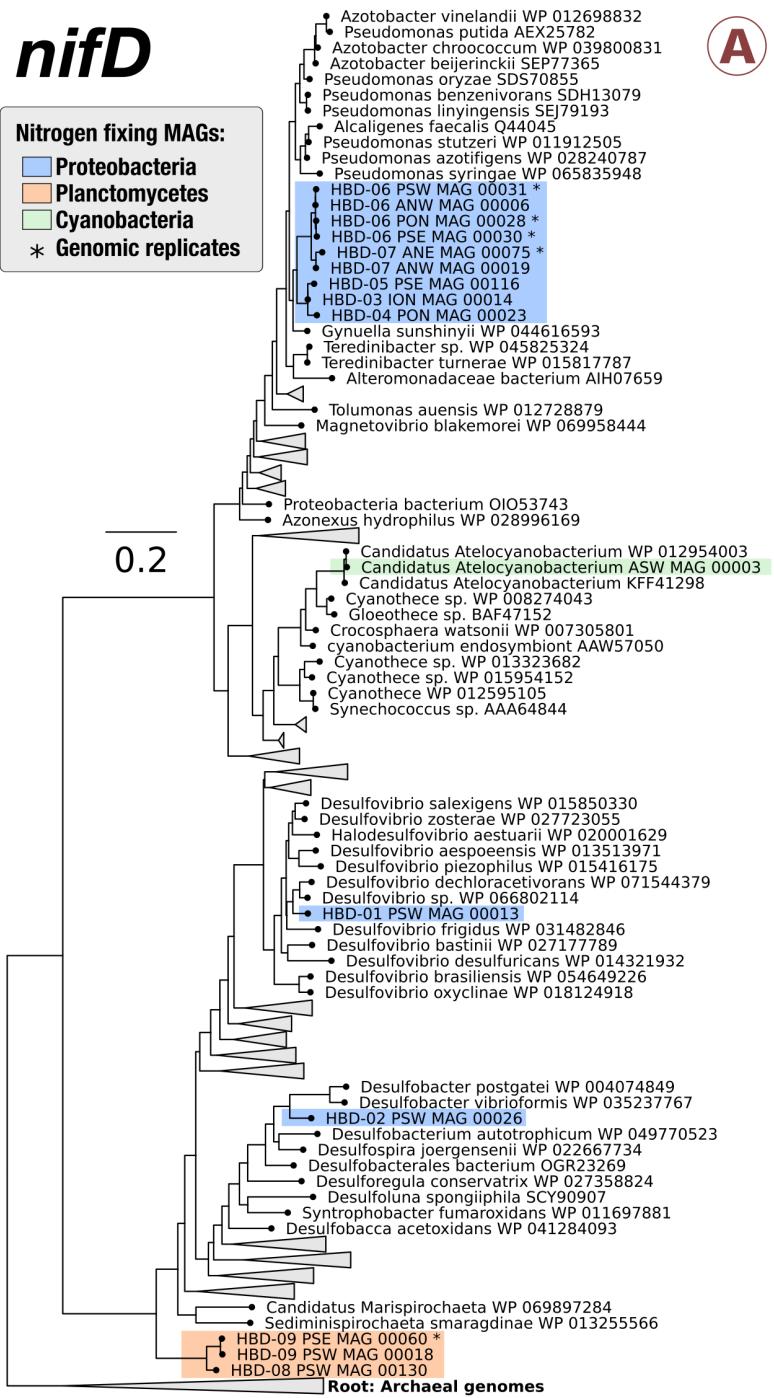


nifD

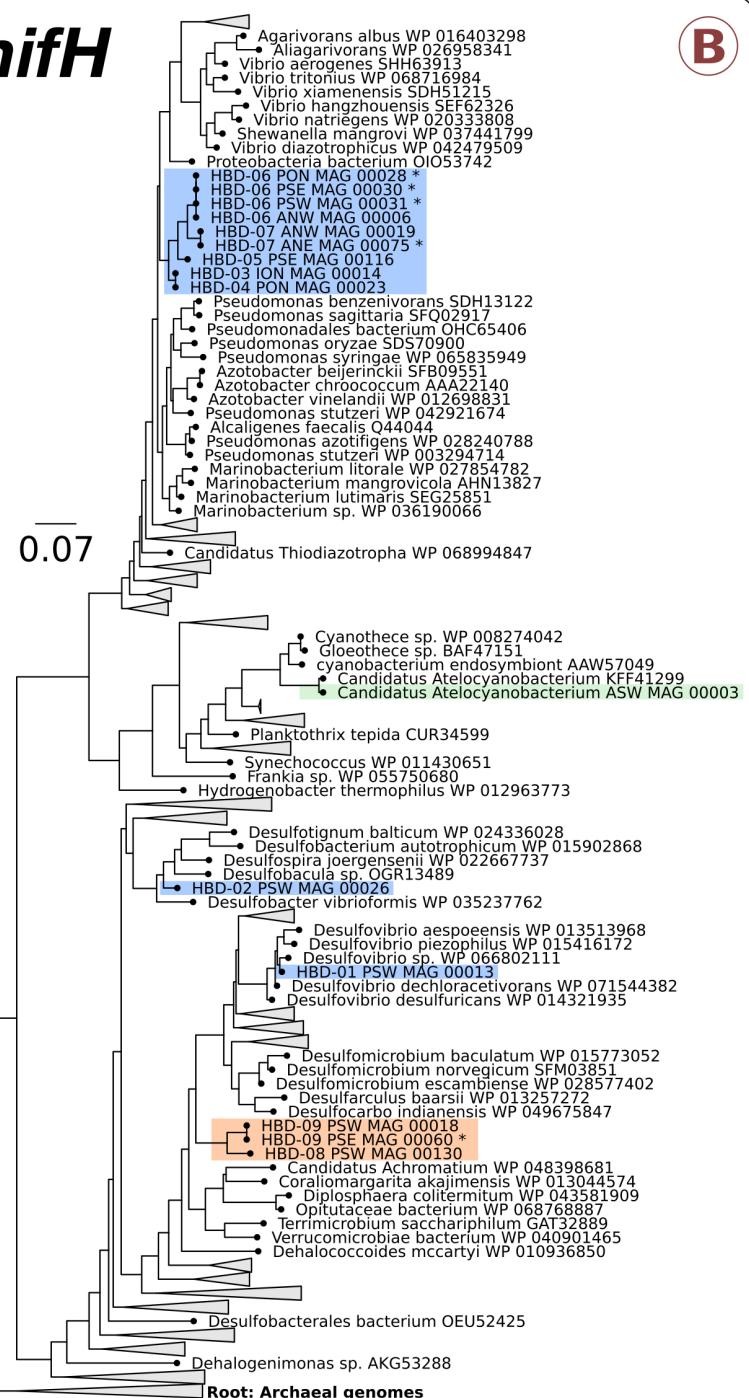
A

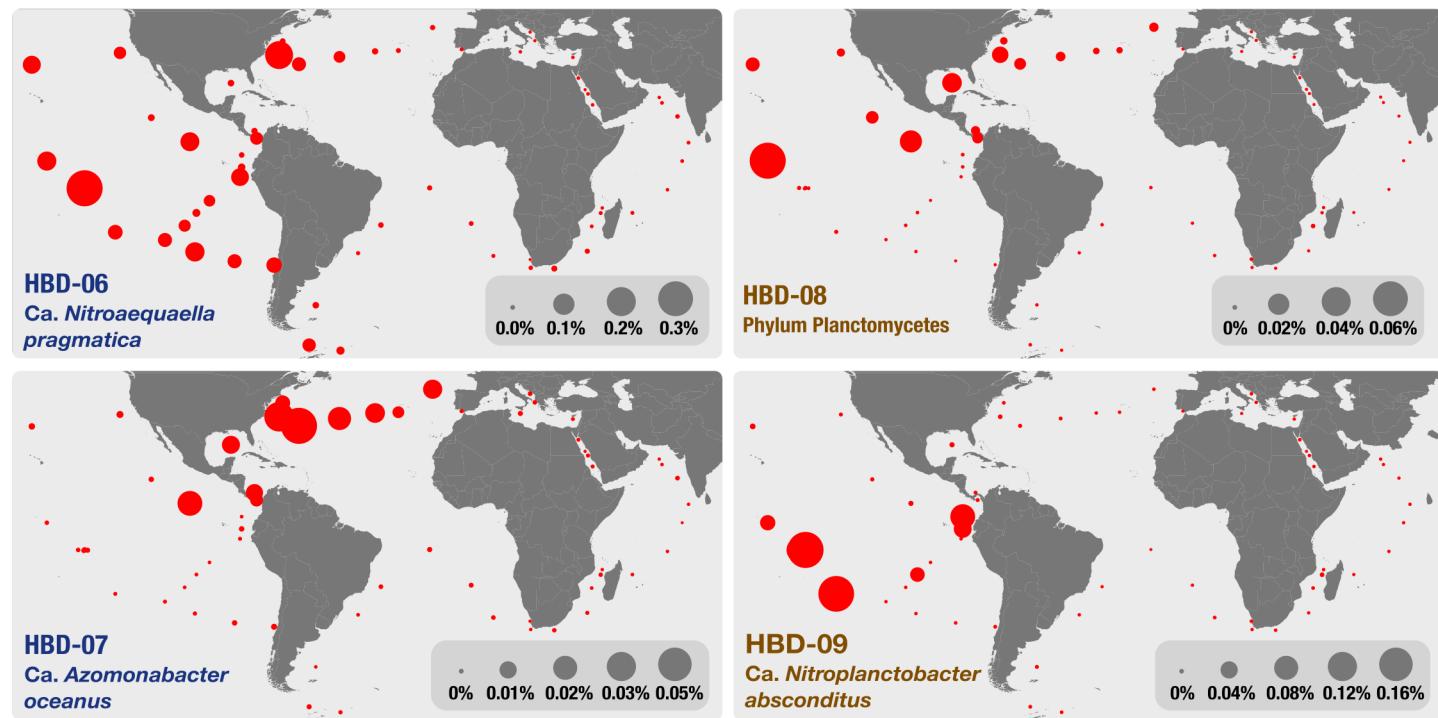
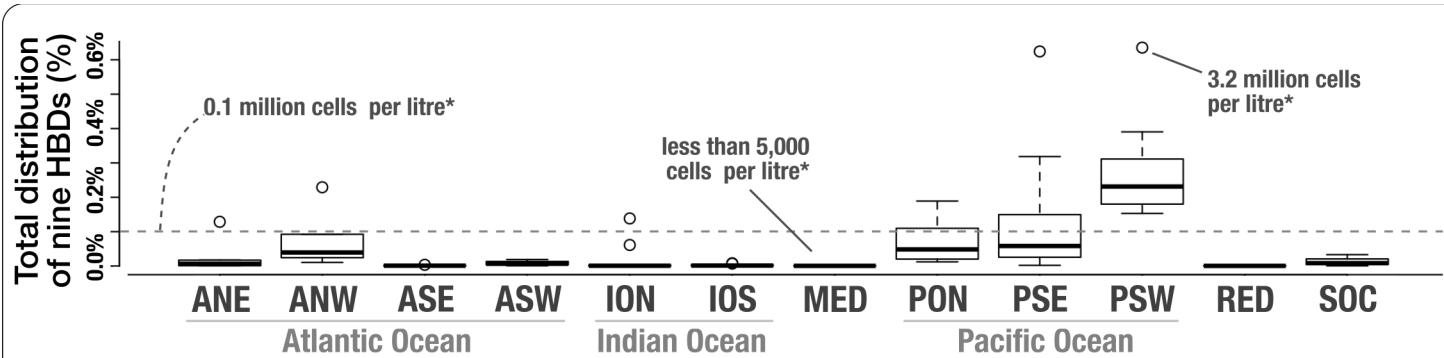
Nitrogen fixing MAGs:

- Proteobacteria
- Planctomycetes
- Cyanobacteria
- * Genomic replicates



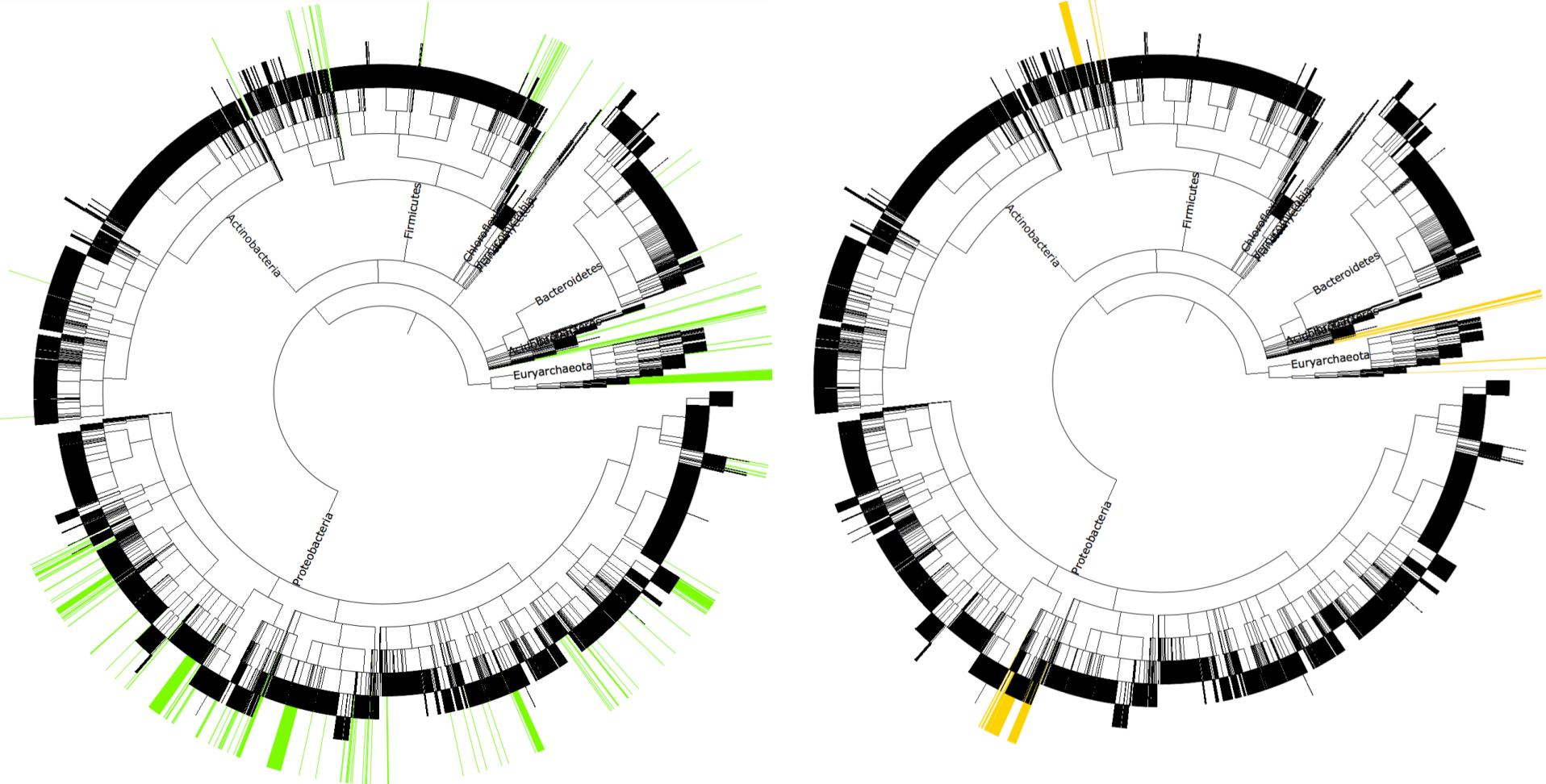
nifH

B



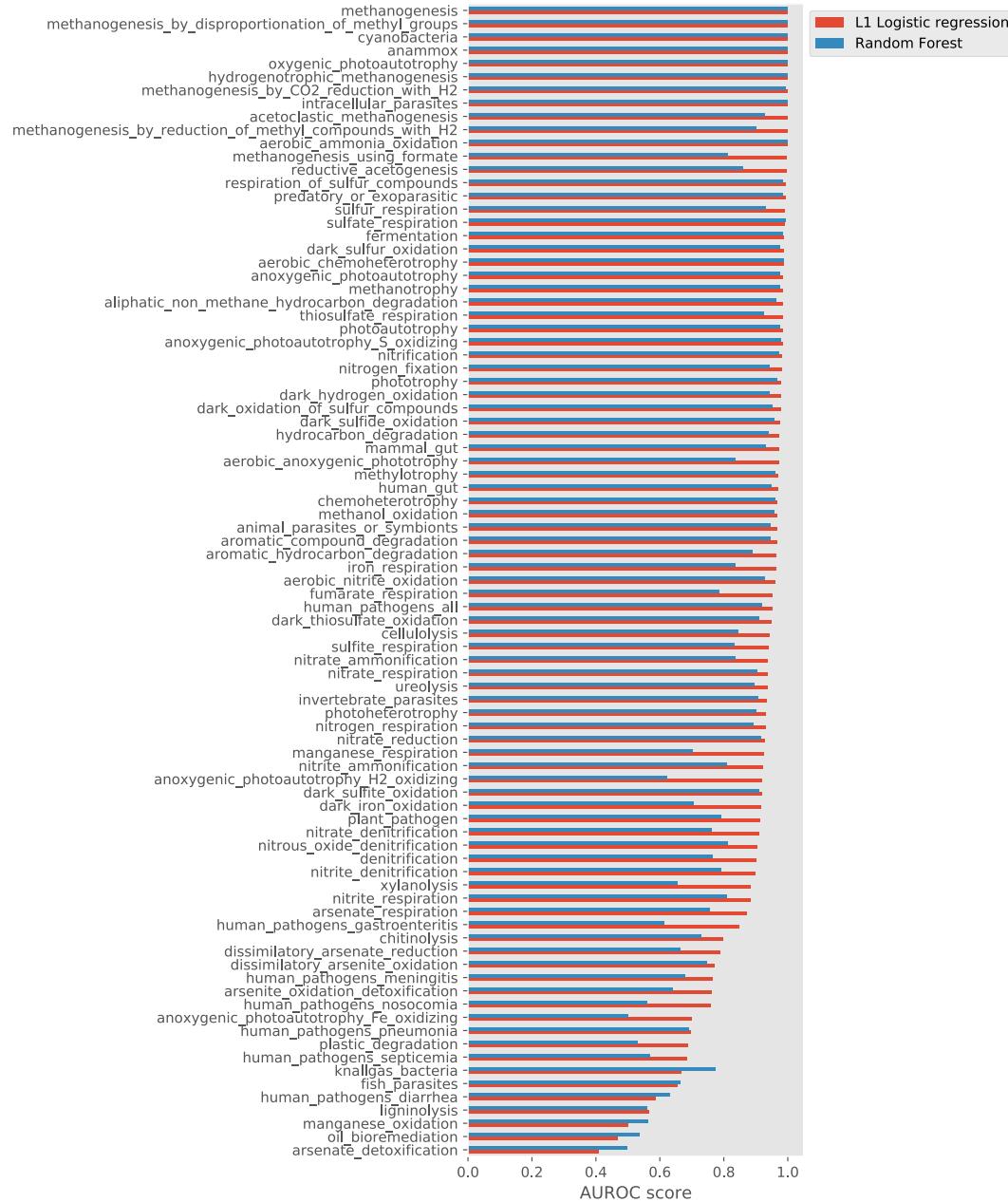
MAG trait inference

- Expanded machine learning strategy to infer 84 traits from Faprotax database ([Louca et al. Science 2016](#))
- Mapped these onto 5,418 genomes with 9,102 kegg orthologs



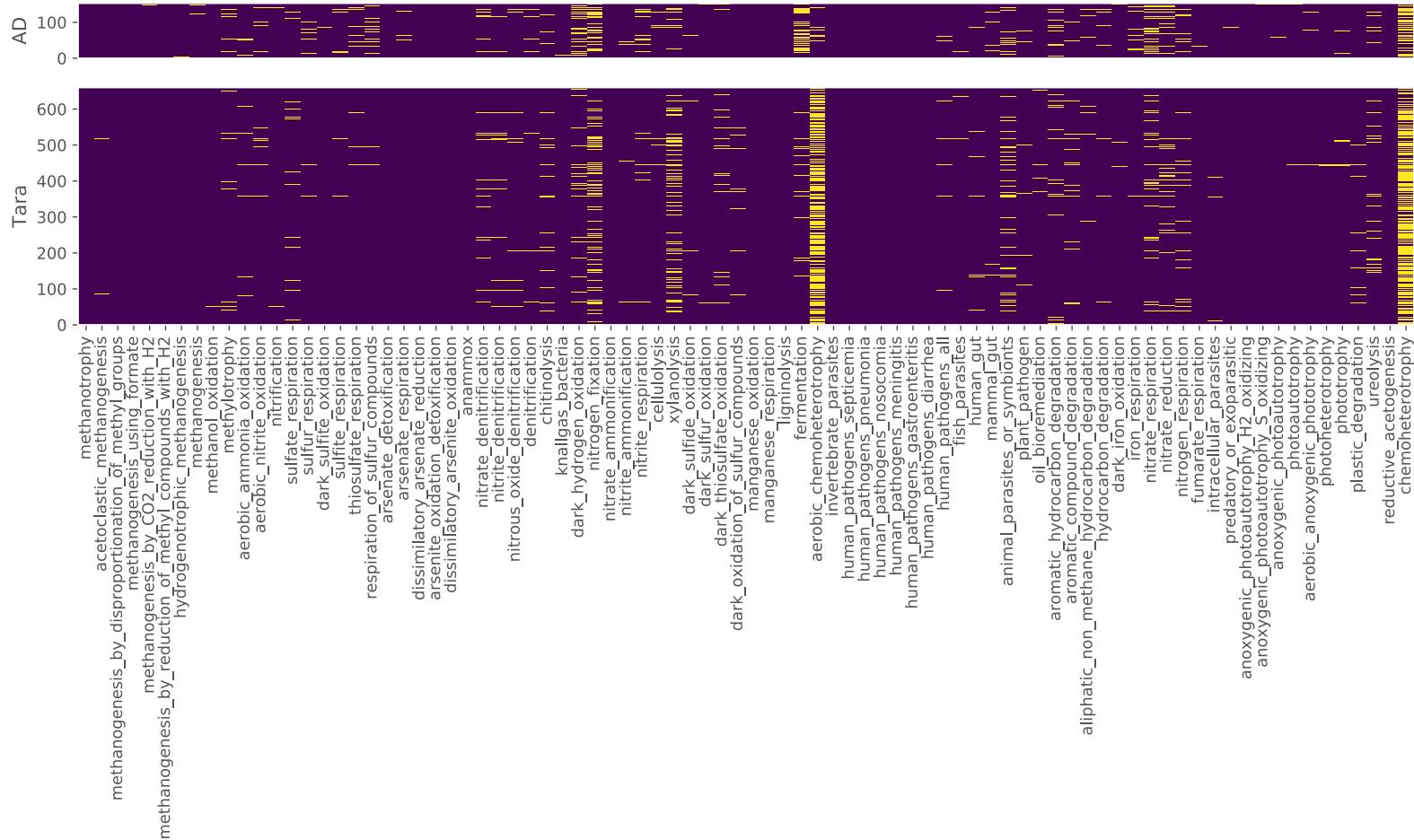
Nitrate respiration

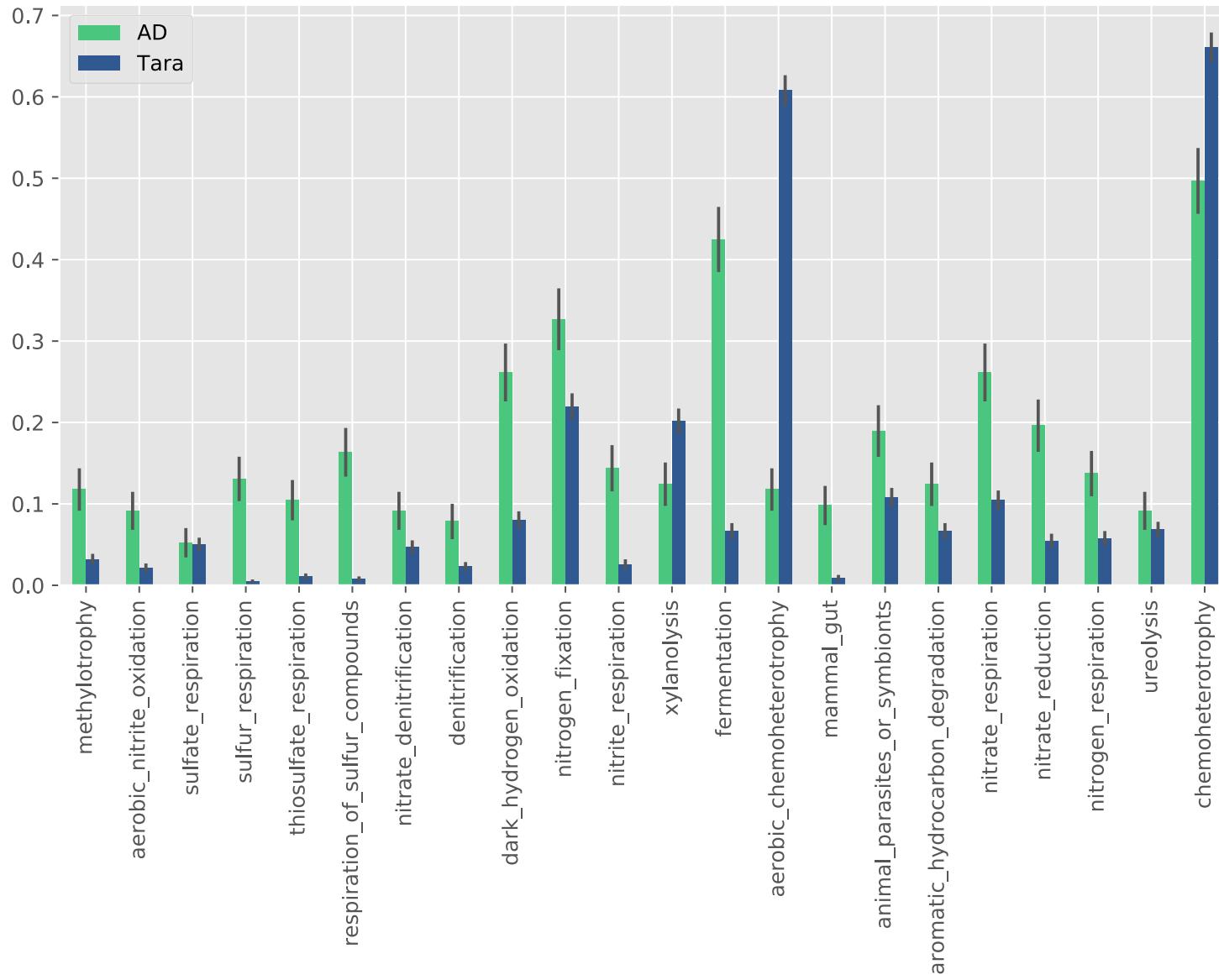
Sulfate respiration



Lasso logistic regression performed best overall

Sulfate respiration		
KO	Weight	Description
K03421	0.263	methyl-coenzyme M reductase subunit C
K14109	0.226	energy-converting hydrogenase A subunit R
K14094	0.156	energy-converting hydrogenase A subunit C
K00401	0.131	methyl-coenzyme M reductase beta subunit [EC:2.8.4.1]
K14097	0.093	energy-converting hydrogenase A subunit F
K00440	0.091	coenzyme F420 hydrogenase subunit alpha [EC:1.12.98.1]
K06862	0.045	energy-converting hydrogenase B subunit Q
K16204	0.038	seco-amyrin synthase [EC:5.4.99.52 5.4.99.54]
K11099	0.033	small nuclear ribonucleoprotein G
K14098	0.027	energy-converting hydrogenase A subunit G
K09613	0.026	COP9 signalosome complex subunit 5 [EC:3.4.--]
K14093	0.022	energy-converting hydrogenase A subunit B
K08074	0.013	ADP-dependent glucokinase [EC:2.7.1.147]
K05181	0.013	gamma-aminobutyric acid receptor subunit beta
K09493	0.013	T-complex protein 1 subunit alpha
K06612	0.013	alpha-N-acetyl-neuraminate alpha-2;8-sialyltransferase (sialyltransferase 8B) [EC:2.4.99.-]
K17278	0.011	membrane-associated progesterone receptor component
K02938	0.003	large subunit ribosomal protein L8e
K00442	0.003	coenzyme F420 hydrogenase subunit delta
K14096	0.003	energy-converting hydrogenase A subunit E





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

Stephanie Connelly + Gavin Collins + Nick Loman + Joshua Quick
Seung Gu (U. of Glasgow) (U. of Birmingham)



Brynjar Smári Bjarnason
+ Johannes Alneberg +
Ino de Bruijn +
Anders Andersson (KTH,
Stockholm)



Fred Farell



A. Murat
Eren and
Tom
Delmont
(Chicago)

