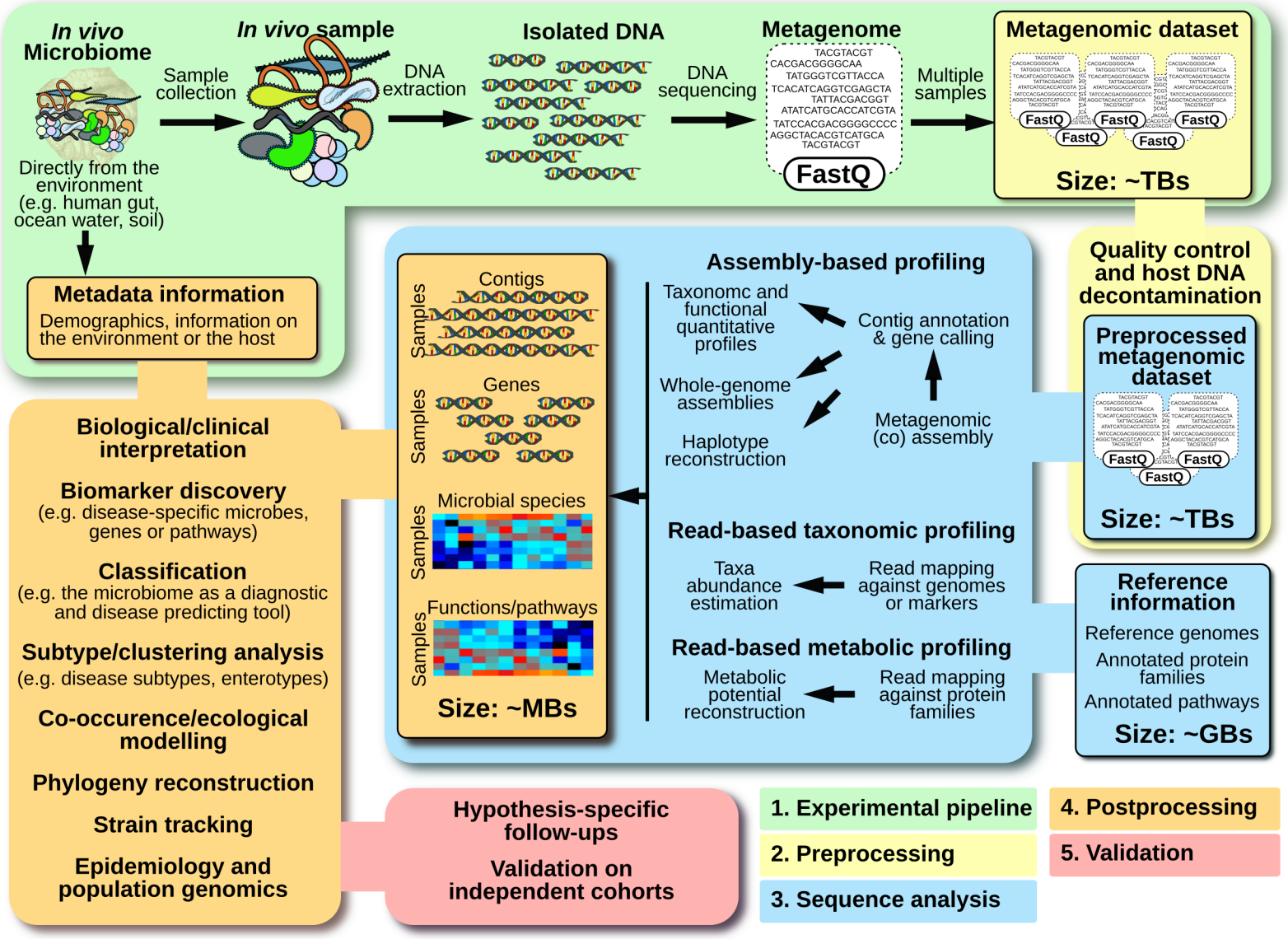
Species and strain resolution from metagenomic coassemblies

Dr Christopher Quince

University of Warwick

Introduction

- Metagenomics provides a culture free method to survey an entire microbial community
- Contig binning from coassemblies has allowed *de novo* species extraction:
  - Candidate Phyla radiation (Brown et al. Nature 2015)
  - Comammox (Maartje et al. 2015)
- Reference based methods exist for high resolution strain profiling (Pathoscope – Hong et al. 2014; SPARSE - Zhou et al. 2018)
- Still cannot obtain 'population genomics' data *de novo* from metagenomes



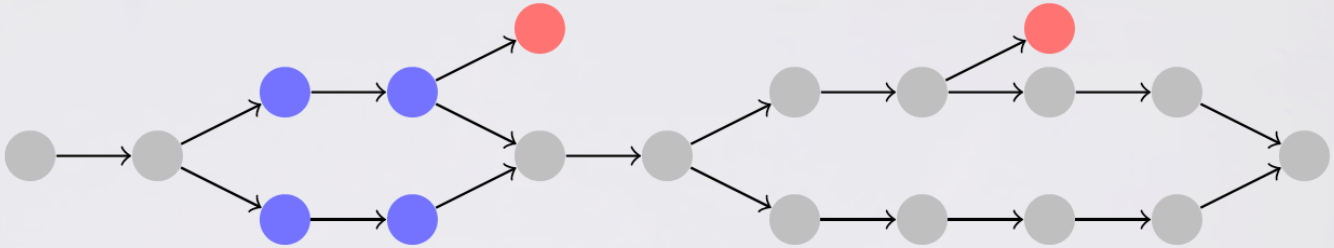
Quince et al Nat. Biotech 2017 http://rdcu.be/vMYk

Overview

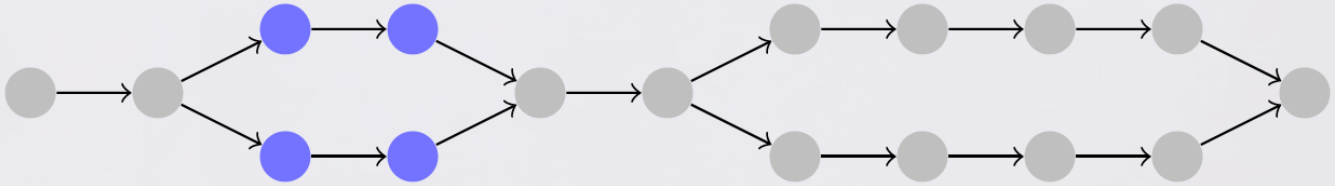
- Introduction: Why is metagenomics coassembly hard?
- Part I: CONCOCT, contig binning
- Part II: DESMAN, resolving intra-MAG diversity

How a metagenome assembler generally works

1) de Bruijn **graph** construction



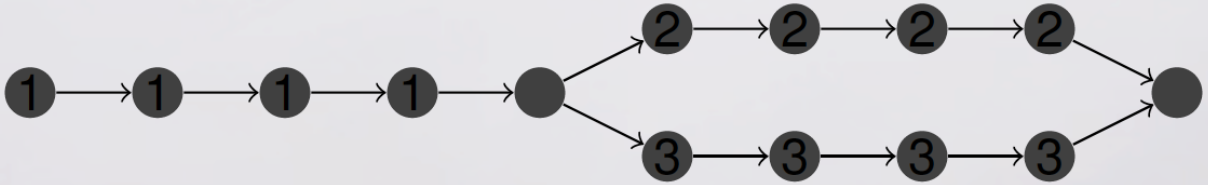
2) Likely sequencing errors are removed.



3) Variations (e.g. SNPs, similar repetitions) are removed.

→ **Skipped in Strains #1**

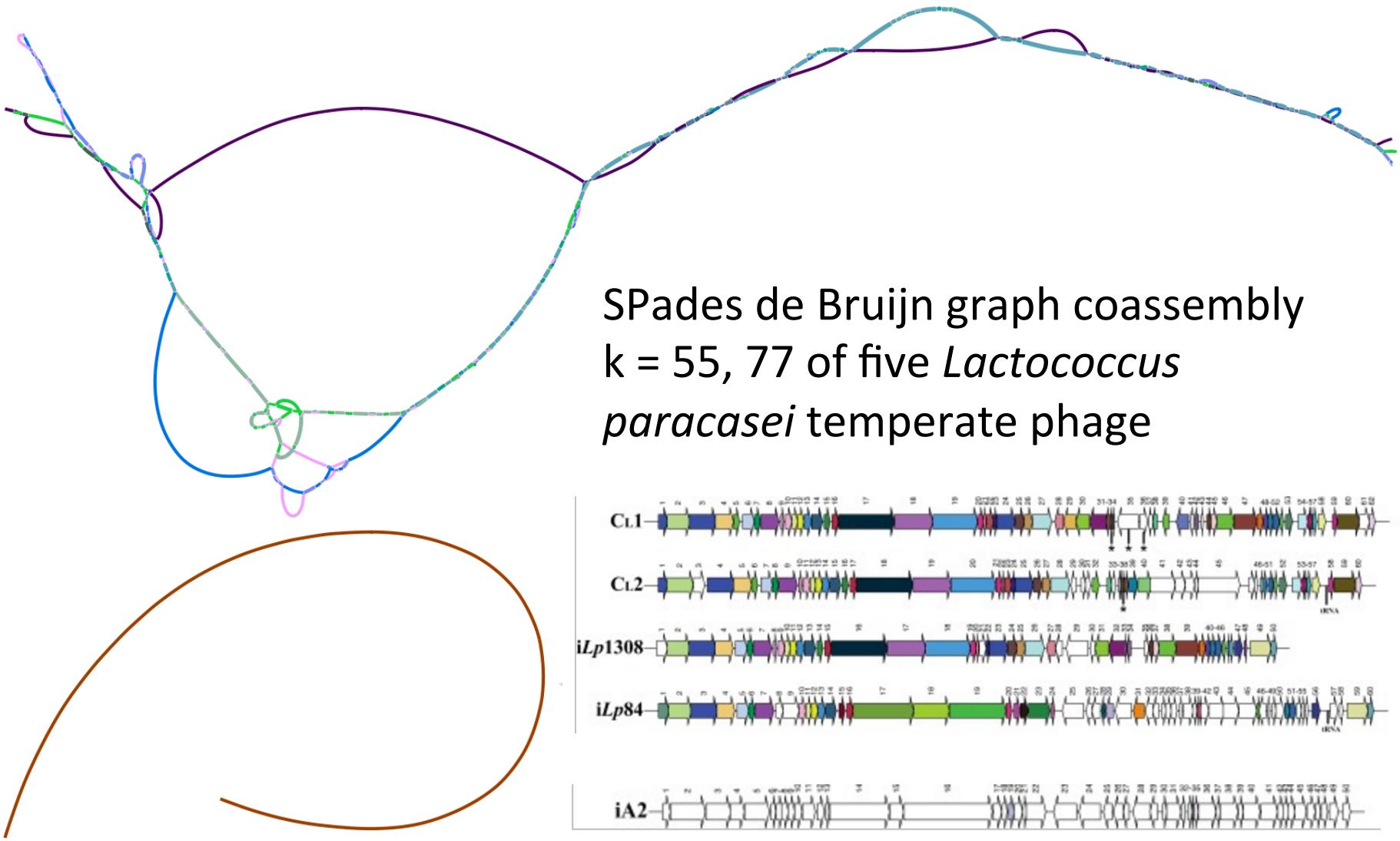
4) **Simple paths** (i.e. contigs) are returned.



5) Extra steps: repeat-resolving, scaffolding (**not done in Minia**)

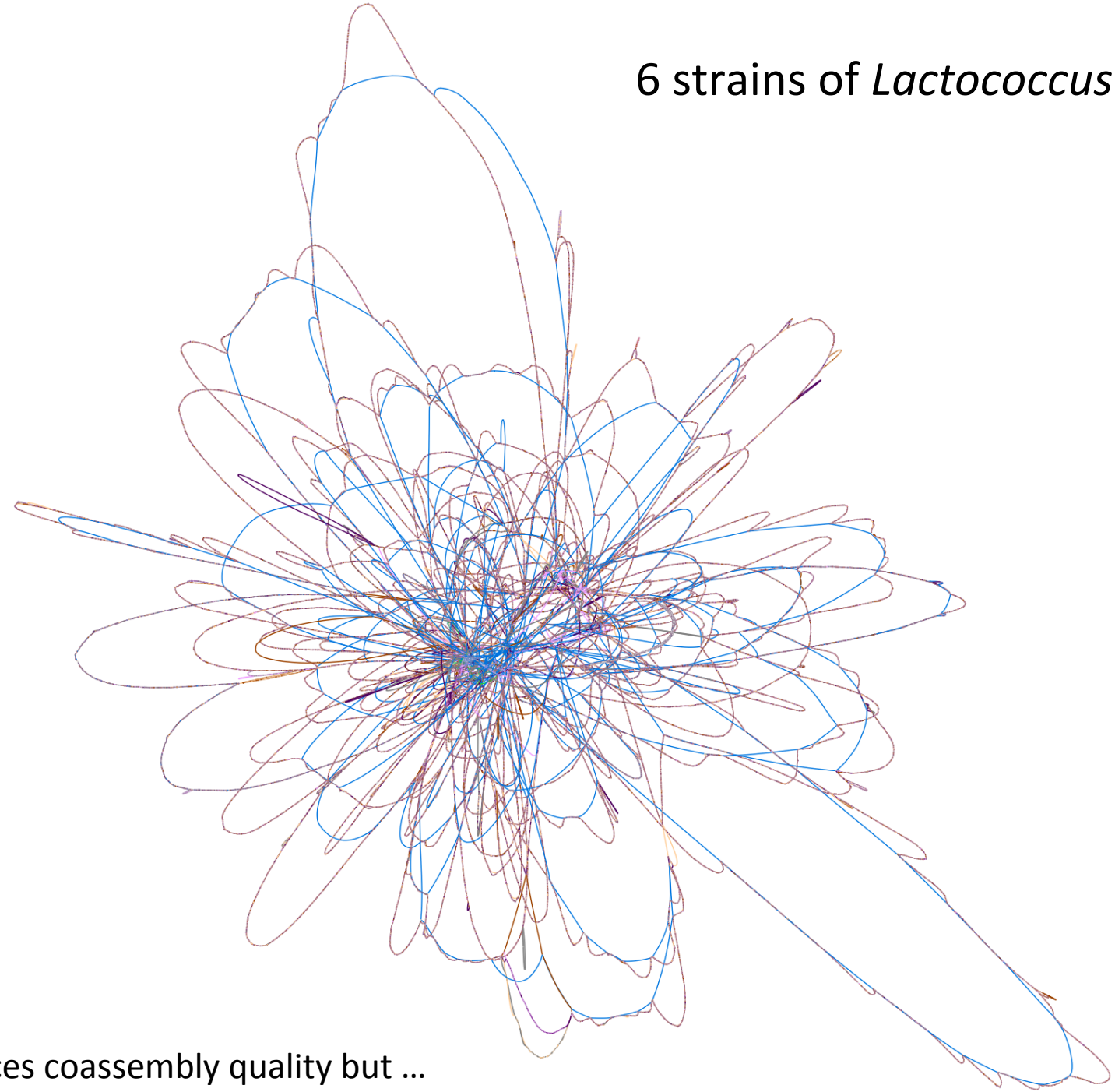
Slide provided by Rayan Chikhi

Strain variation leads to complex assembly graphs…



SPades de Bruijn graph coassembly k = 55, 77 of five *Lactococcus paracasei* temperate phage

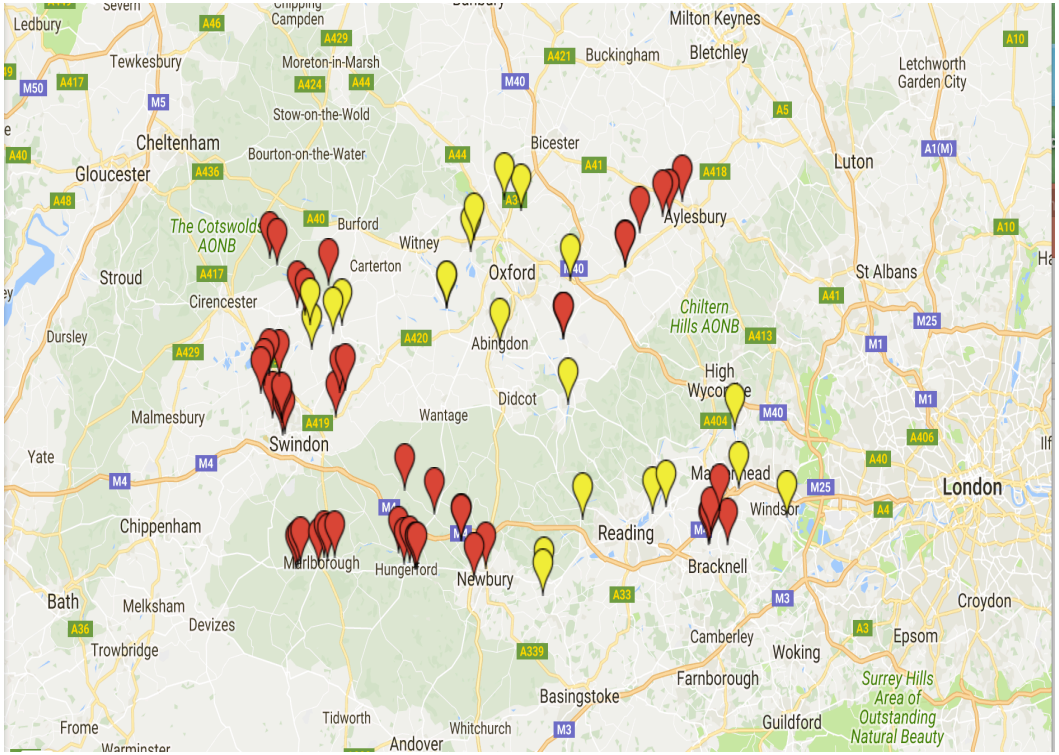
6 strains of *Lactococcus lactis*



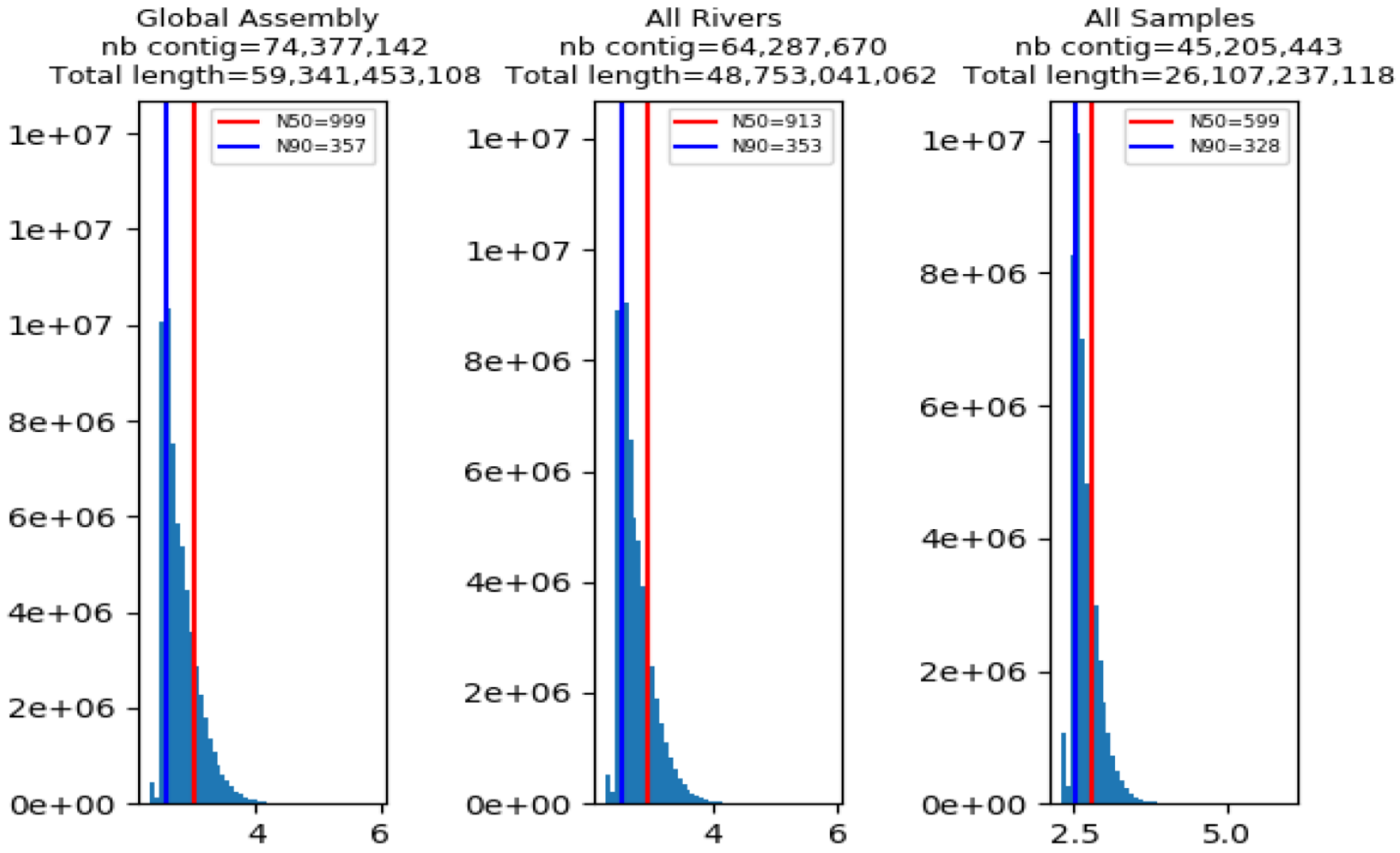
This reduces coassembly quality but …

Thames AMRG survey

- 20 sites located along 7 tributaries of Thames catchment were chosen for sampling across three seasons and with up to three biological replicates per site



- Total 171 samples
- 30 – 114.7 million 2X150 bp reads per sample
- 2.3 Tbp



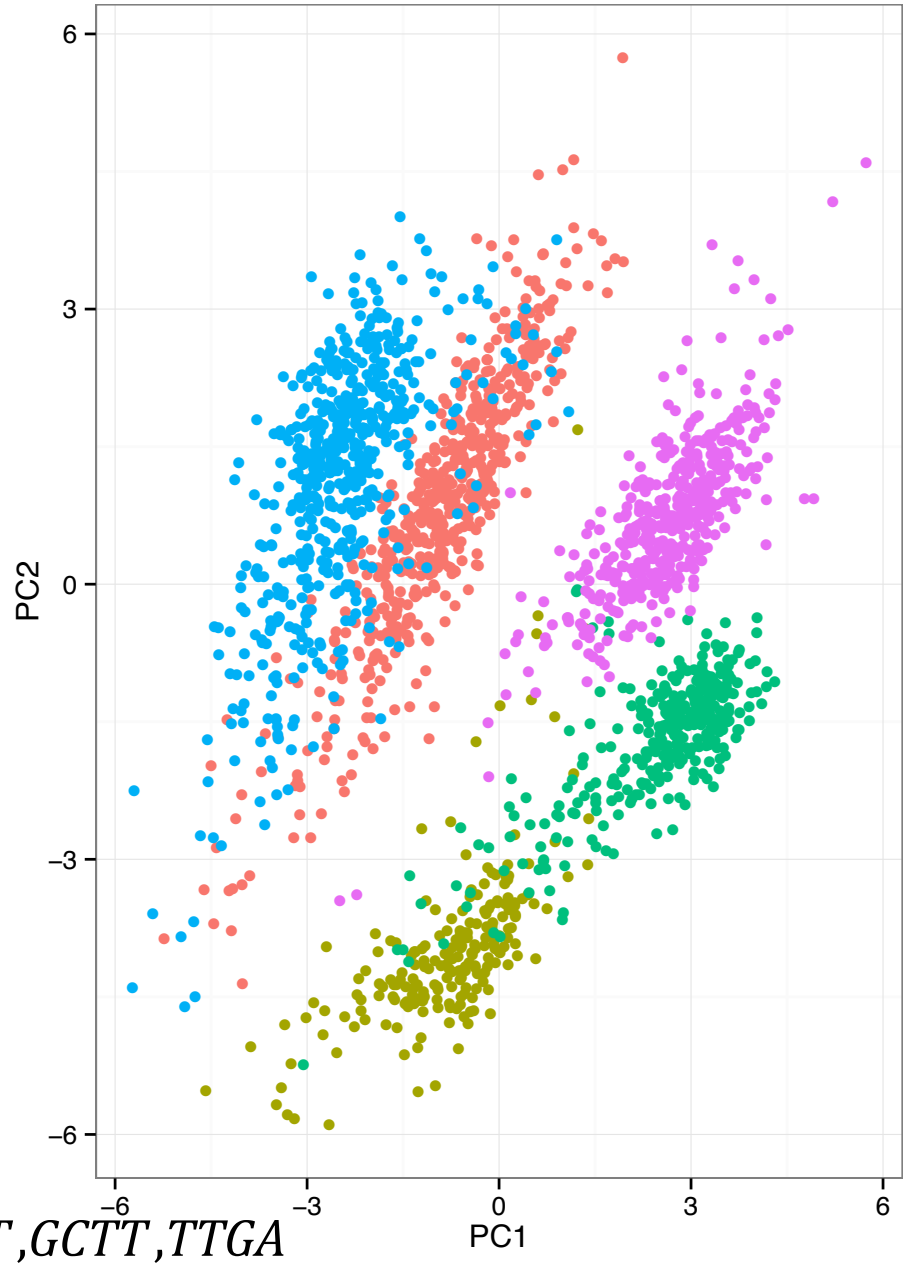
Performed assemblies with megahit

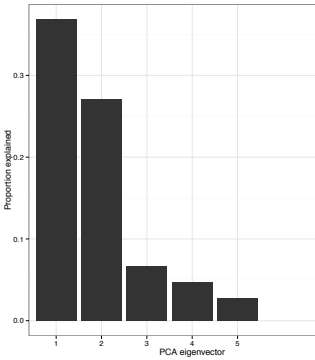
PART I:
CONCOCT

Contig binning

Contig clustering by sequence composition

Key to contig binning is to find features that are shared by sequence from the same genome…

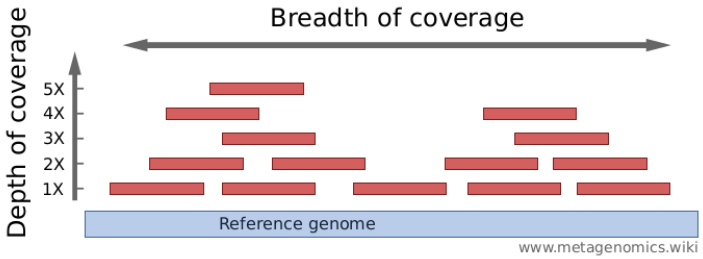


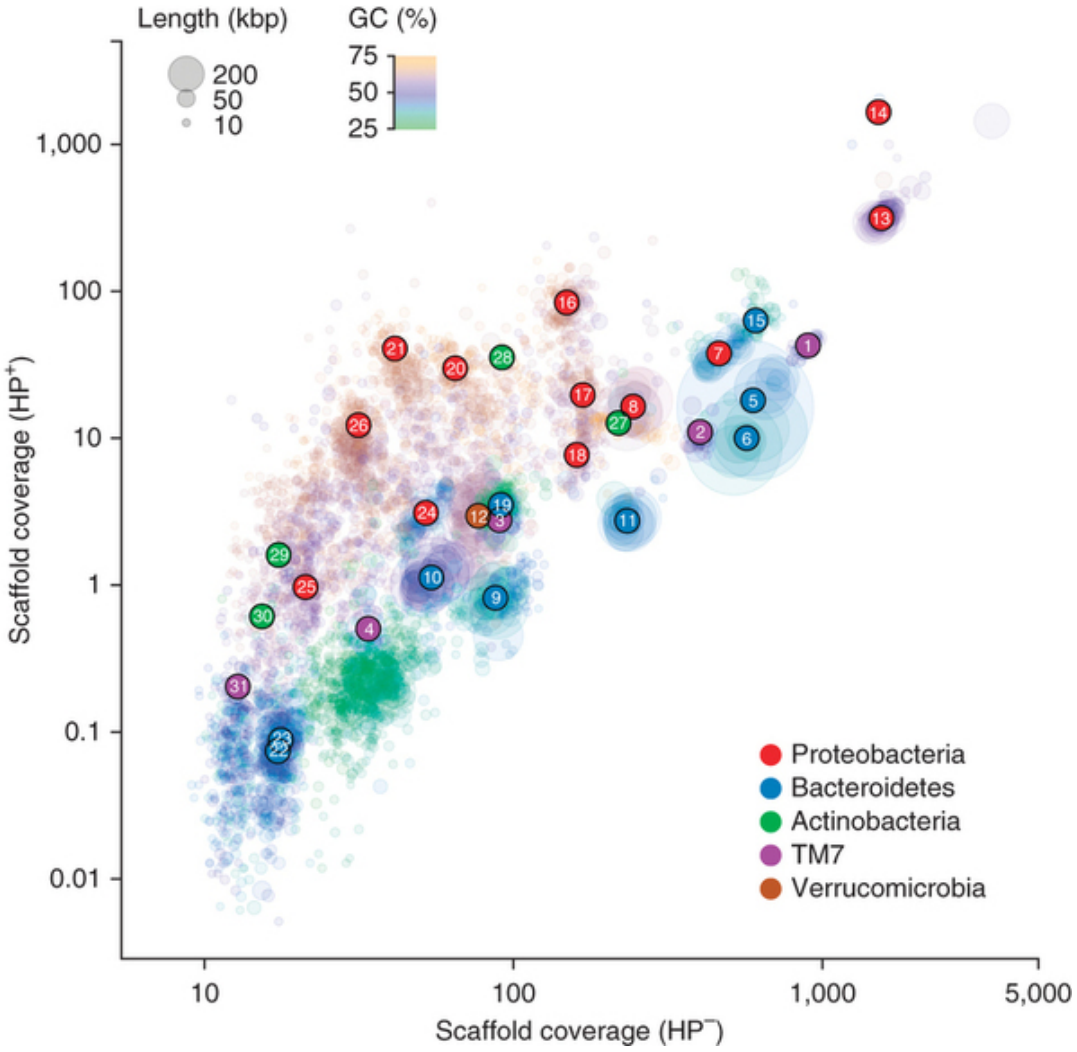


$$sequence = \left(\overline{CTGGCTTGA}\right)$$

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

Contig clustering by differential coverage (Albertsen et al. Nature Biotech. 2014)





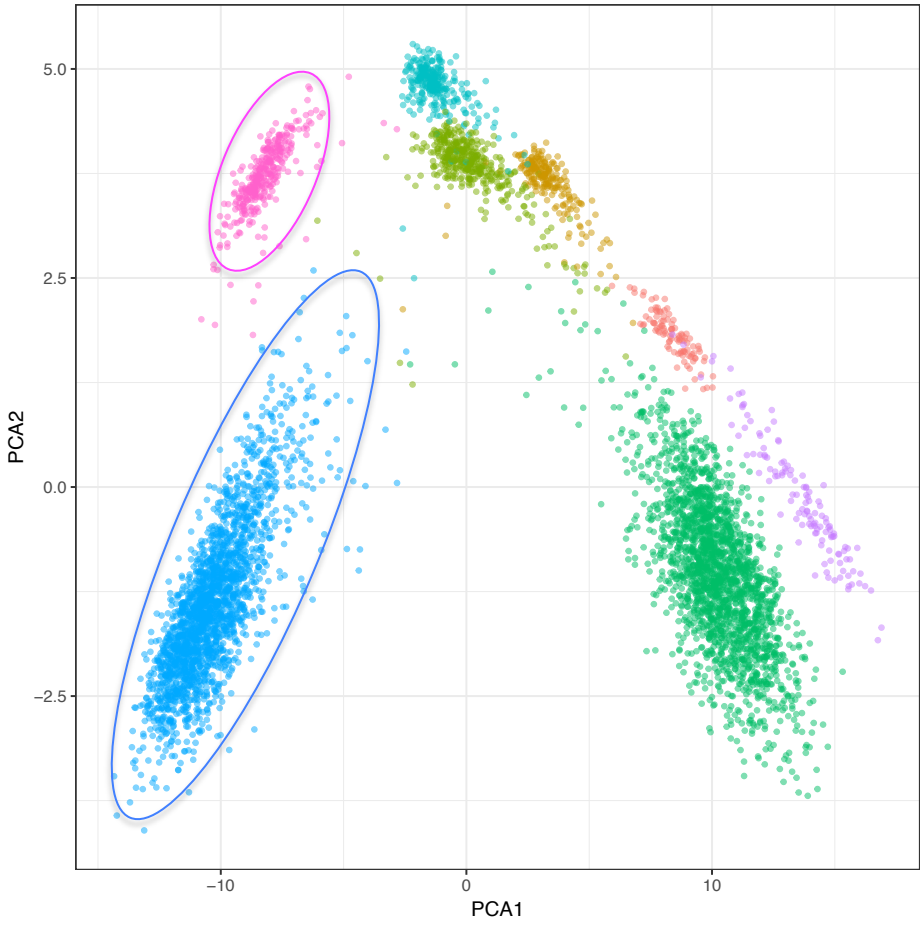
CONCOCT: **C**lustering c**ON**tigs on **CO**verage and **C**ompos**T**ion

(Alneberg et al. Nat. Methods 2014)

- Albertsen's strategy and ESOMs (Sharon et al. 2013) required human supervision
- Created CONCOCT as the first? automatic binner to combine coverage and and composition
- Crowded field now e.g. GroopM (Imelfort et al. PeerJ 2014), MetaBAT (Kang et al. PeerJ 2015), MaxBin2 (Wu et al. Bioinf. 2015), COCACOLA (Lu et al. 2017)
- CONCOCT still used and is highly cited, first step in the second generation human supervised clustering platform Anvi'o (Eren et al. PeerJ 2015)

CONCOCT algorithm

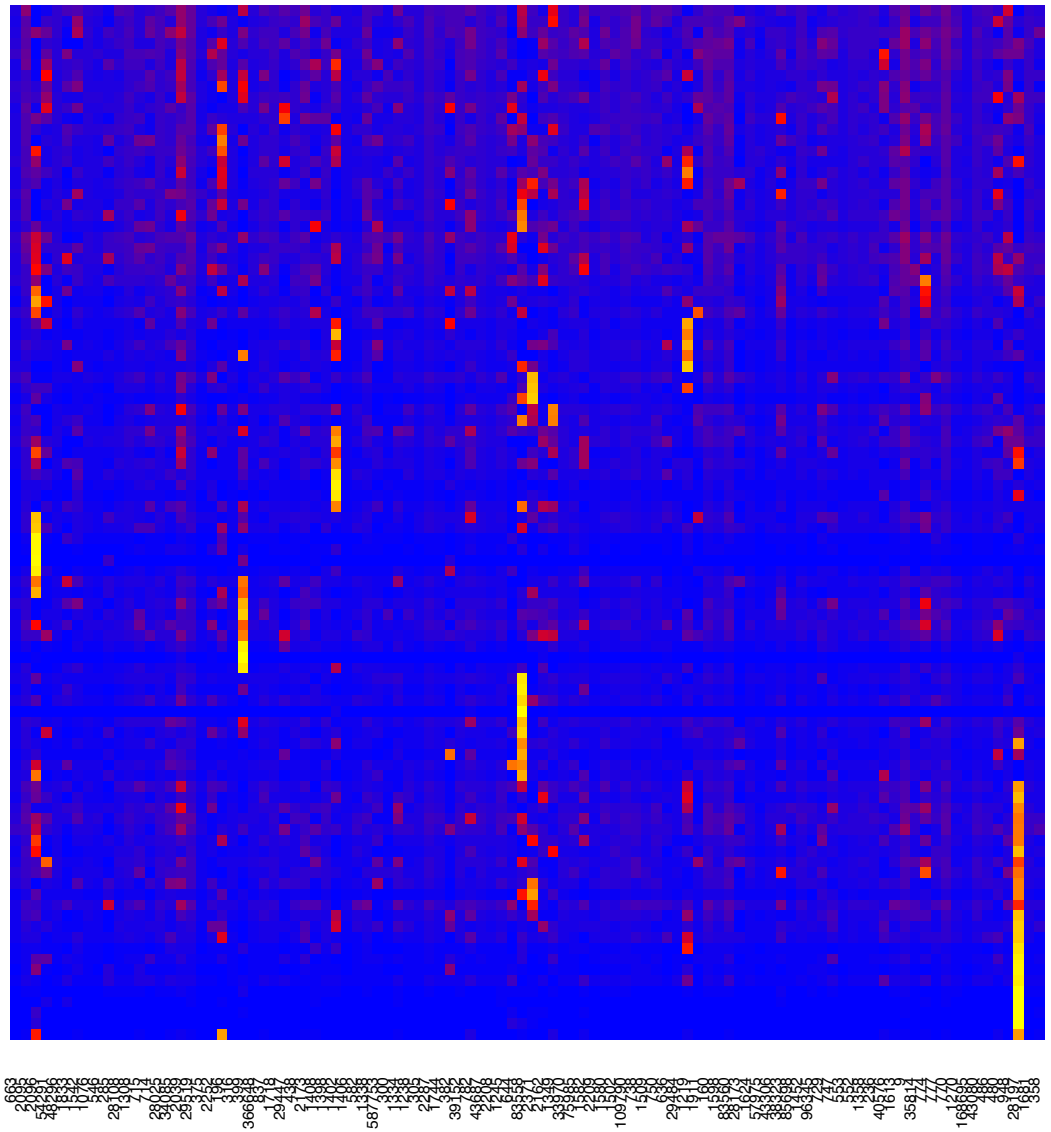
- Perform coassembly across all samples
- Map reads back to get coverage depth in each sample
- Use coverage and composition but perform PCA to reduce dimensionality
- Cluster with Gaussian Mixture Model
- Variational Bayes to select number of components through automatic relevance determination (ARD)
- Added mulithreading and refinement step in development branch



Metagenomic binner	Description	Original publication
CONCOCT	Binner using differential coverage, tetranucleotide frequencies, paired-end linkage	Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., … Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, *11*(11), 1144–1146.
CONCOCT with refinement step	Binner using differential coverage, tetranucleotide frequencies, paired-end linkage	-
MaxBin 2.0	Binner using multi-sample coverage, tentranucleotide frequencies	Wu, Y. W., Simmons, B. A., & Singer, S. W. (2015). MaxBin 2.0: An automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics*, *32*(4), 605–607.
MetaBAT	Binner using multi-sample coverage, tetranucleotide frequencies, paired-end	Kang, D. D., Froula, J., Egan, R., & Wang, Z. (2015). MetaBAT, an efficient tool for accurately

Strain diverse synthetic community

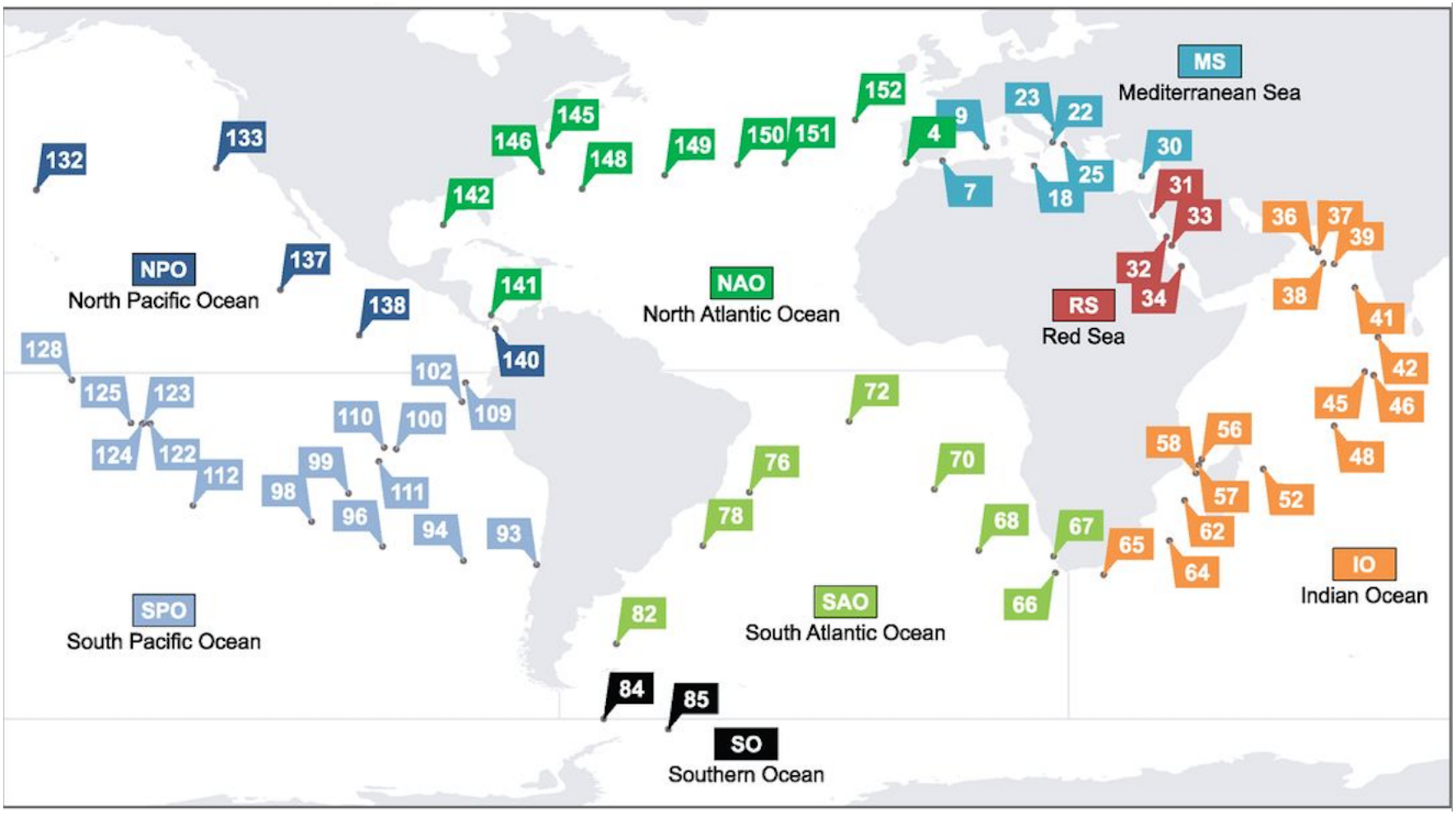
- 100 different species and 210 genomes (1:50, 2:20, 3:10, 4:10, 5:10)
- Simulated 96 samples of 6.25 million 2X150 bp HiSeq paired end reads (https://github.com/chrisquince/StrainMetaSim)
- Realistic species and strain coverages across samples
- Coassembly with megahit
- Randomly select 10, 20, 30 and 60* samples
- Evaluate with adjusted Rand index and number of complete genomes



Results

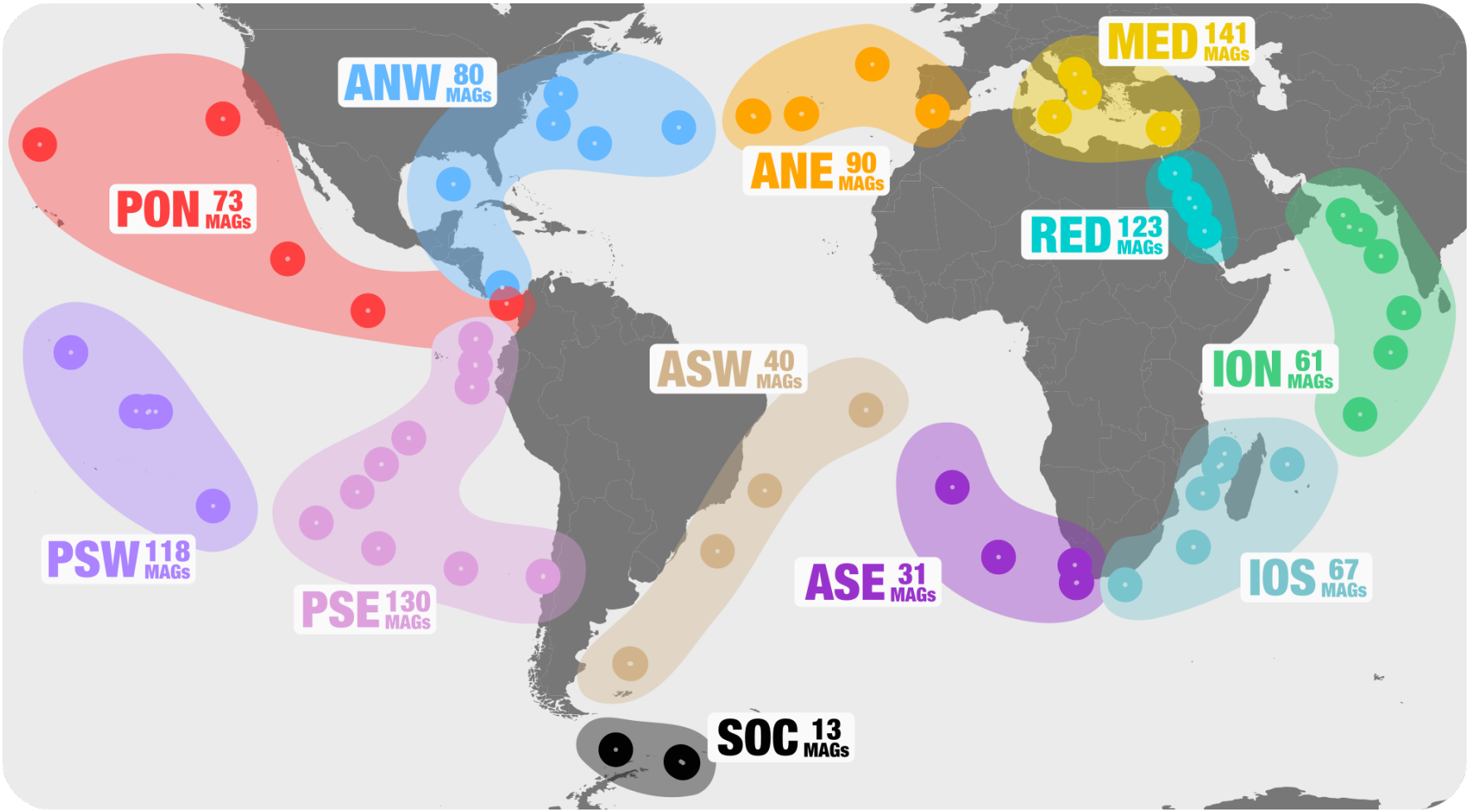
Sample size	Metagenomic binner	Contigs binned	Number of bins (75% COGs)	ARI
N=10	CONCOCT	73458	170 (59)	0.8086
	CONCOCT refinement	**73458**	**130 (66)**	**0.8197**
	MAXBIN 2.0	72512	110 (48)	0.7105
	MetaBAT	71603	123 (48)	0.2895
	GATTAGA	73379	95 (39)	0.4881
N=20	CONCOCT	79863	130 (52)	0.7800
	CONCOCT refinement	**79863**	**165 (70)**	**0.8061**
	MAXBIN 2.0	79377	116 (64)	0.7398
	MetaBAT	76875	137 (45)	0.2768
	GATTAGA	79778	87 (42)	0.3638
N=30	CONCOCT	78382	115 (53)	0.6456
	CONCOCT refinement	**78382**	**165 (75)**	**0.7848**
	MAXBIN 2.0	77983	117 (54)	0.7579
	MetaBAT	74936	136 (46)	0.2836
	GATTAGA	78431	89 (44)	0.50988
N=60	CONCOCT	75486	104 (58)	0.6806
	CONCOCT refinement	**75486**	**152 (79)**	**0.7944**
	MAXBIN 2.0	74936	117(70)	0.7688
	MetaBAT	72625	136 (43)	0.2967

TARA Oceans sampling sites: Sunagawa et al. Science 2015



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

TARA Oceans: Sunagawa et al. Science 2015



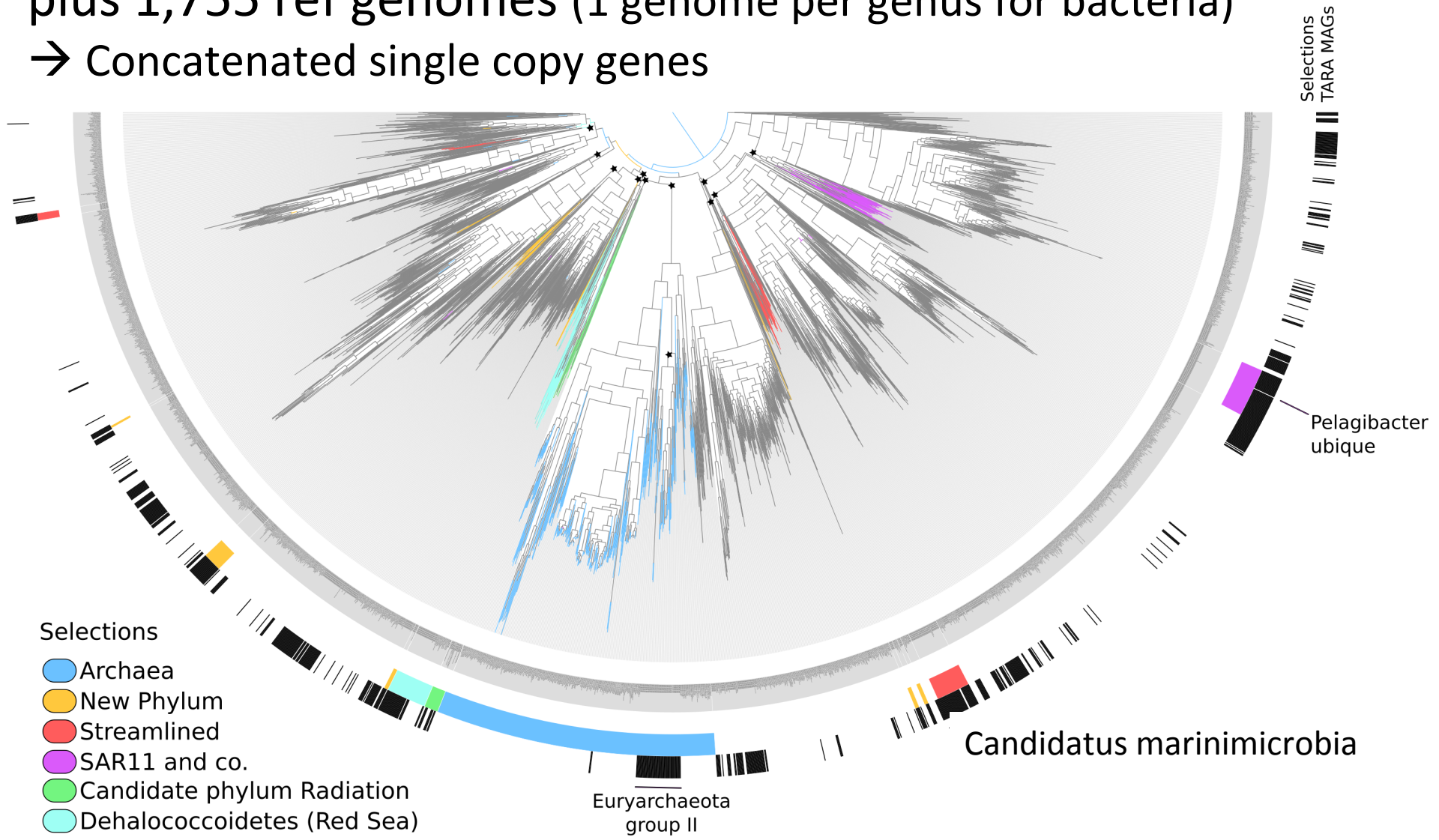
- 93 metagenomes from the planktonic size fraction for which we performed 12 metagenomic co-assemblies
- Generated 957 non-redundant MAGs encompassing the three domains of life

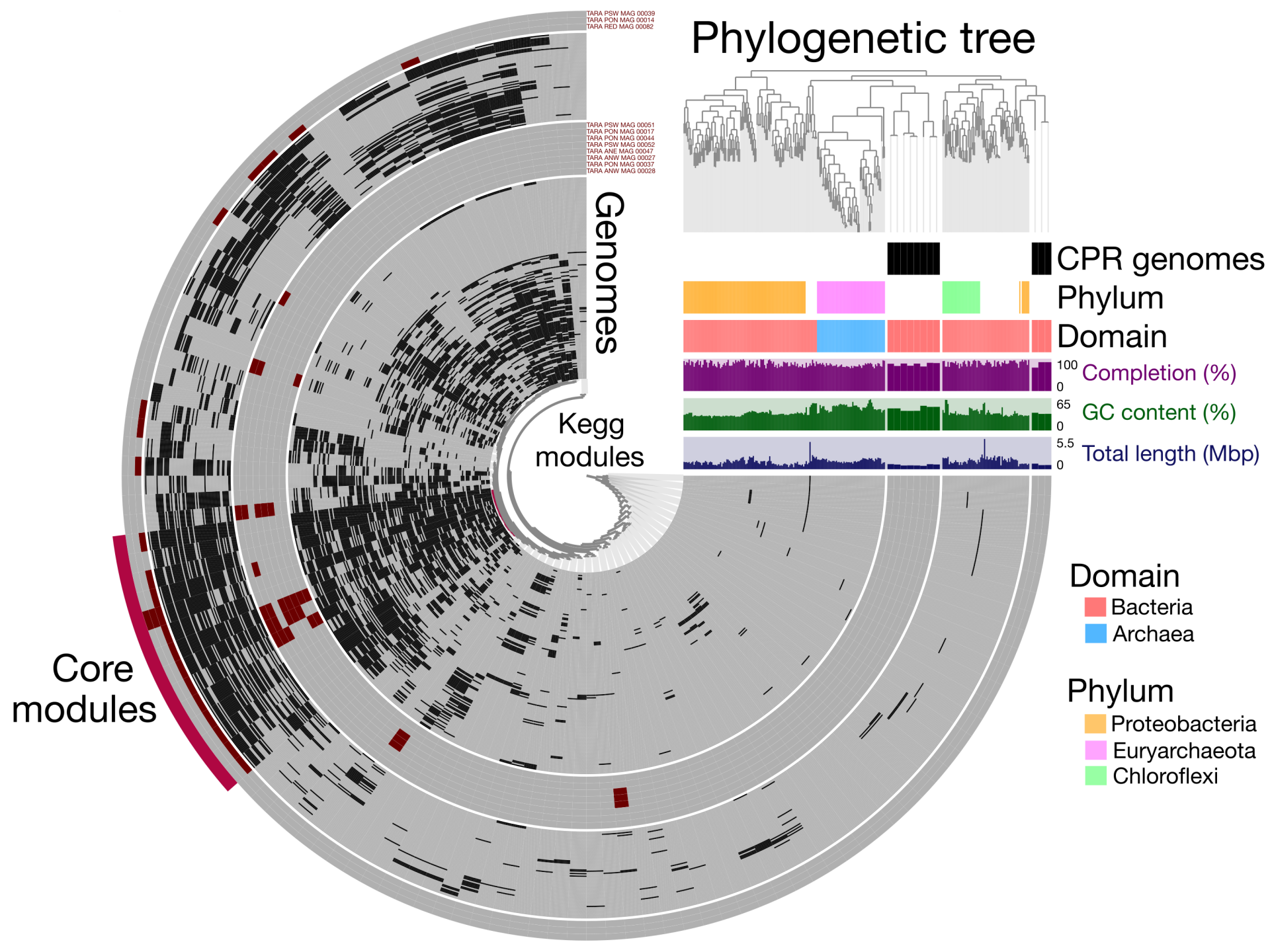
  Delmont, Quince, …, Eren "Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in the surface ocean" Nature Microbiology 2018

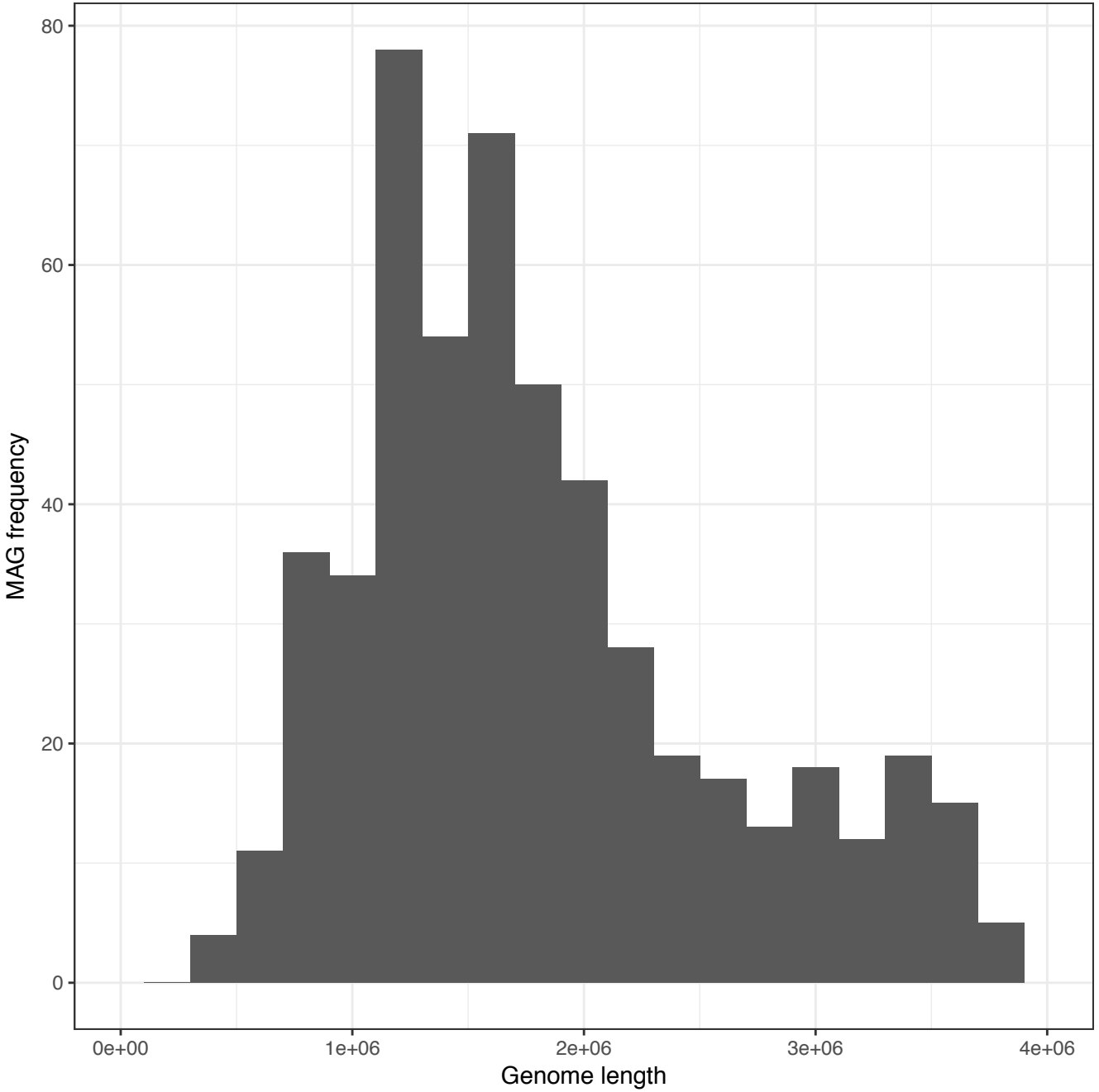
Phylogenetic analysis of 660 MAGs

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

→ Concatenated single copy genes




Phylogenetic tree
TARA PSW MAG 00039
TARA PON MAG 00014
TARA RED MAG 00082
TARA PSW MAG 00051
TARA PON MAG 00017
TARA PON MAG 00044
TARA PSW MAG 00052
TARA ANE MAG 00047
TARA ANW MAG 00027
TARA PON MAG 00037
TARA ANW MAG 00028
Genomes
Kegg modules
Core modules
CPR genomes
Phylum
Domain
100
0
Completion (%)
65
0
GC content (%)
5.5
0
Total length (Mbp)
Domain
Bacteria
Archaea
Phylum
Proteobacteria
Euryarchaeota
Chloroflexi


80
60
40
20
0
MAG frequency
0e+00
1e+06
2e+06
3e+06
4e+06
Genome length

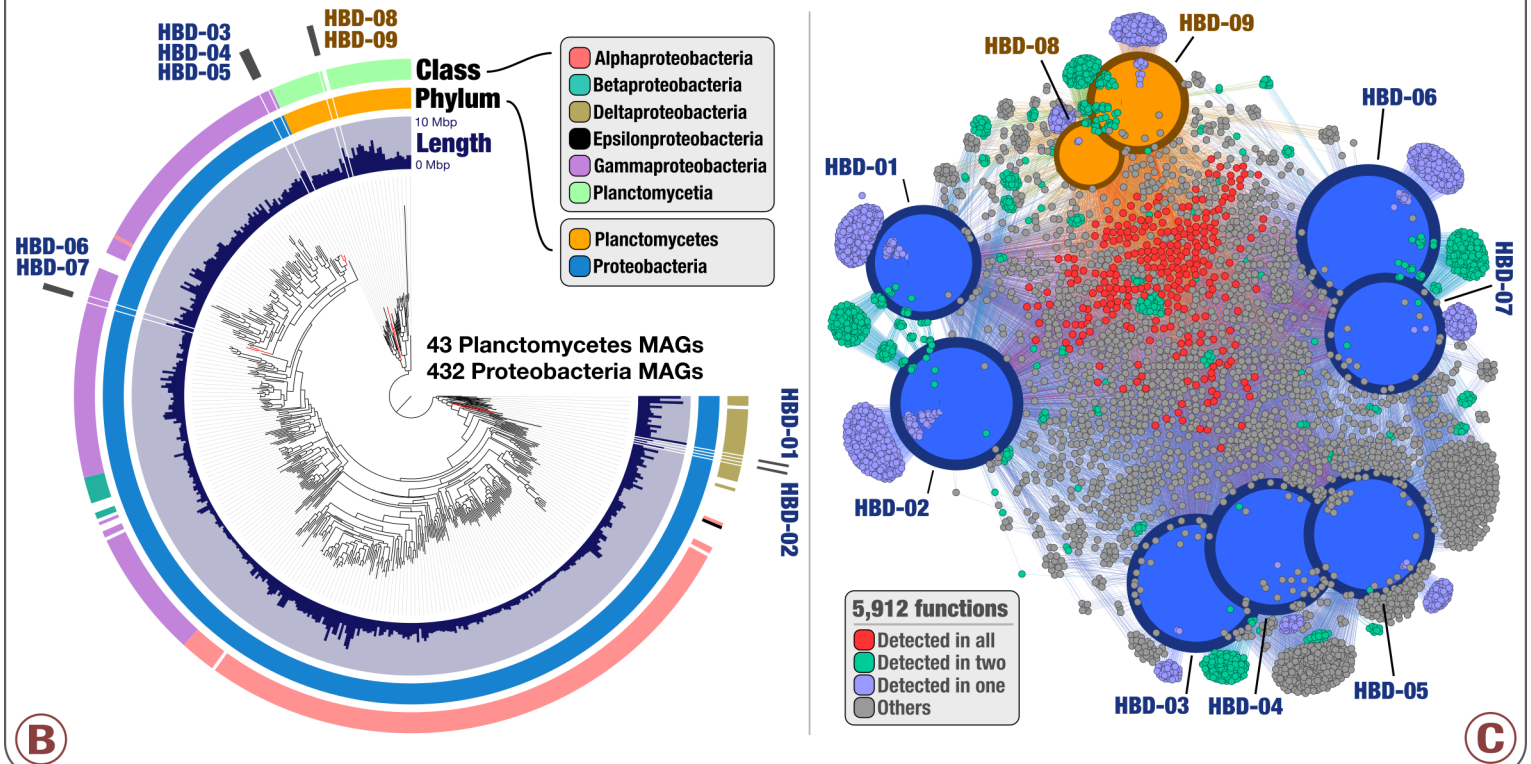
Oceanic heterotrophic bacterial diazotrophs (HBDs)

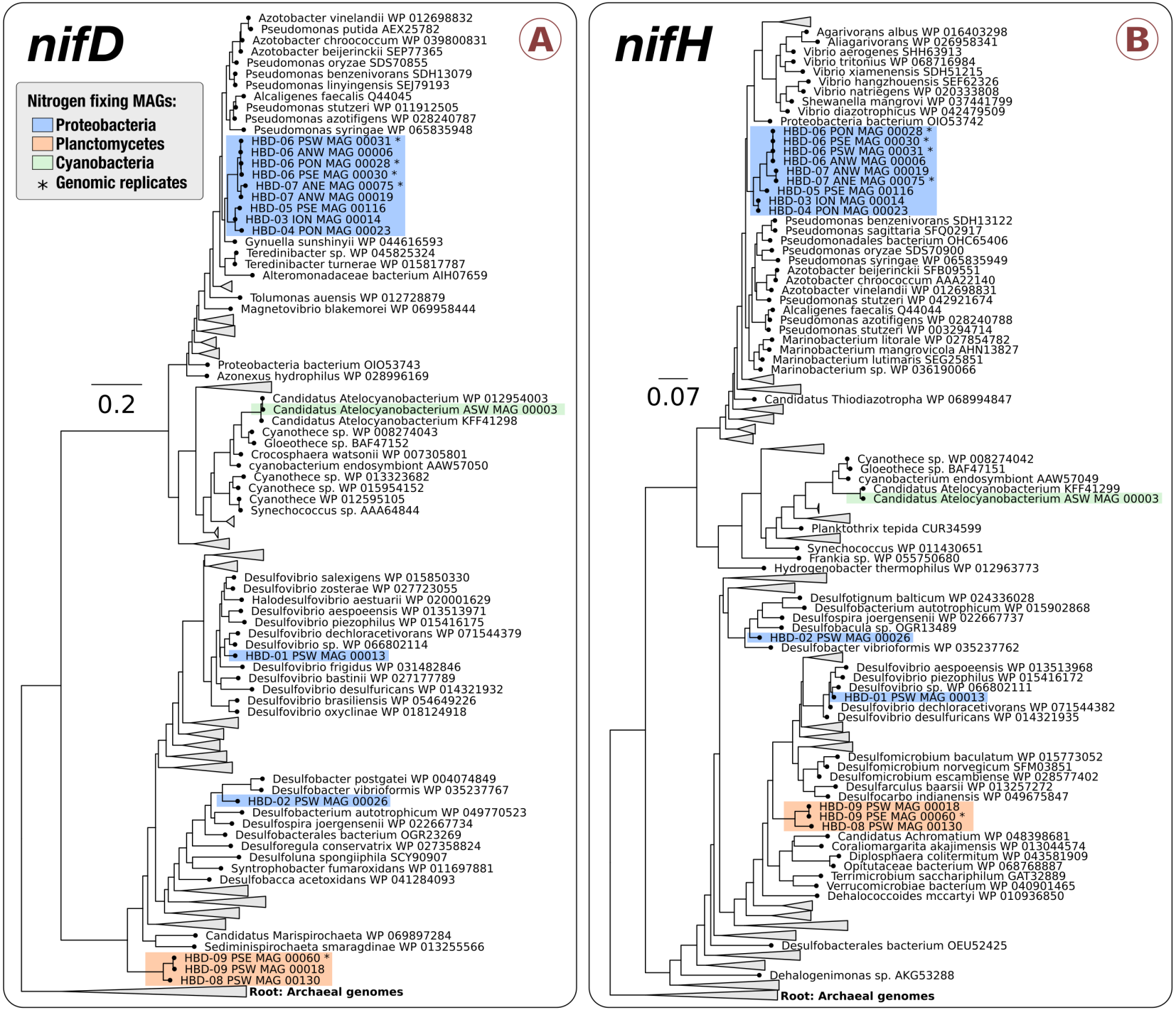
- Nitrogen fixation in the surface regulates microbial primary productivity and the sequestration of carbon through the biological pump
- Cyanobacteria populations have long been thought to represent the main suppliers of the bio-available nitrogen
- Nitrogenase reductase gene surveys revealed the existence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

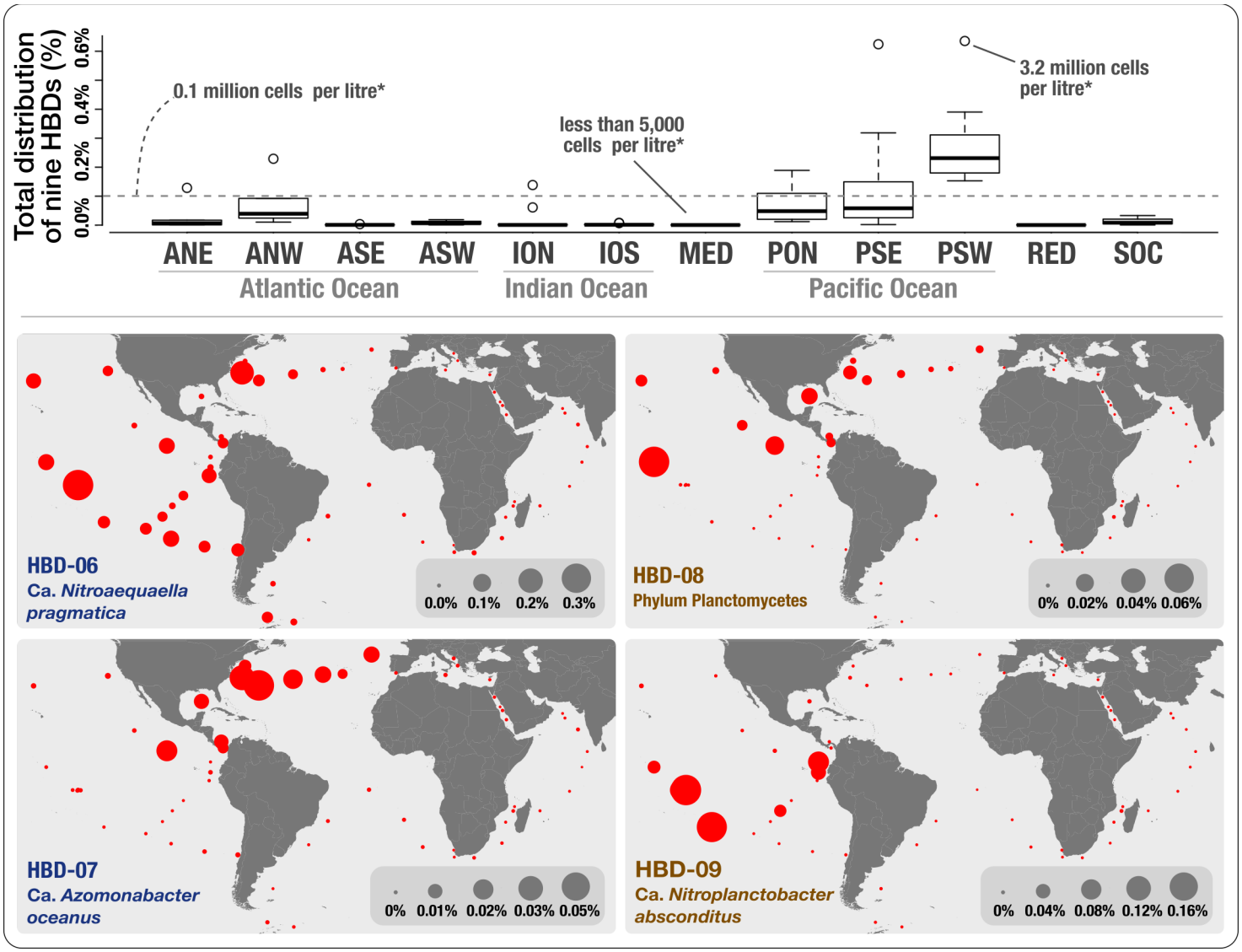
- Search for MAGs containing the catalytic (nifH, nifD, nifK) and the biosynthetic proteins (nifE, nifN and nifB) required for nitrogen fixation
- One of six Cyanobacteria
- Nine other MAGs possessed all six genes

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.	Desulfovibrionales	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
0.2
Azotobacter vinelandii WP 012698832
Pseudomonas putida AEX25782
Azotobacter chroococcum WP 039800831
Azotobacter beijerinckii SEP77365
Pseudomonas oryzae SDS70855
Pseudomonas benzenivorans SDH13079
Pseudomonas linyingensis SEJ79193
Alcaligenes faecalis Q44045
Pseudomonas stutzeri WP 011912505
Pseudomonas azotifigens WP 028240787
Pseudomonas syringae WP 065835948
HBD-06 PSW MAG 00031 *
HBD-06 ANW MAG 00006
HBD-06 PON MAG 00028 *
HBD-06 PSE MAG 00030 *
HBD-07 ANE MAG 00075 *
HBD-07 ANW MAG 00019
HBD-05 PSE MAG 00116
HBD-03 ION MAG 00014
HBD-04 PON MAG 00023
Gynuella sunshinyii WP 044616593
Teredinibacter sp. WP 045825324
Teredinibacter turnerae WP 015817787
Alteromonadaceae bacterium AIH07659
Tolumonas auensis WP 012728879
Magnetovibrio blakemorei WP 069958444
Proteobacteria bacterium OIO53743
Azonexus hydrophilus WP 028996169
Candidatus Atelocyanobacterium WP 012954003
Candidatus Atelocyanobacterium ASW MAG 00003
Candidatus Atelocyanobacterium KFF41298
Cyanothece sp. WP 008274043
Gloeothece sp. BAF47152
Crocosphaera watsonii WP 007305801
cyanobacterium endosymbiont AAW57050
Cyanothece sp. WP 013323682
Cyanothece sp. WP 015954152
Cyanothece WP 012595105
Synechococcus sp. AAA64844
Desulfovibrio salexigens WP 015850330
Desulfovibrio zosterae WP 027723055
Halodesulfovibrio aestuarii WP 020001629
Desulfovibrio aespoeensis WP 013513971
Desulfovibrio piezophilus WP 015416175
Desulfovibrio dechloracetivorans WP 071544379
Desulfovibrio sp. WP 066802114
HBD-01 PSW MAG 00013
Desulfovibrio frigidus WP 031482846
Desulfovibrio bastinii WP 027177789
Desulfovibrio desulfuricans WP 014321932
Desulfovibrio brasiliensis WP 054649226
Desulfovibrio oxyclinae WP 018124918
Desulfobacter postgatei WP 004074849
Desulfobacter vibrioformis WP 035237767
HBD-02 PSW MAG 00026
Desulfobacterium autotrophicum WP 049770523
Desulfospira joergensenii WP 022667734
Desulfobacterales bacterium OGR23269
Desulforegula conservatrix WP 027358824
Desulfoluna spongiiphila SCY90907
Syntrophobacter fumaroxidans WP 011697881
Desulfobacca acetoxidans WP 041284093
Candidatus Marispirochaeta WP 069897284
Sediminispirochaeta smaragdinae WP 013255566
HBD-09 PSE MAG 00060 *
HBD-09 PSW MAG 00018
HBD-08 PSW MAG 00130
Root: Archaeal genomes
nifH
B
0.07
Agarivorans albus WP 016403298
Aliagarivorans WP 026958341
Vibrio aerogenes SHH63913
Vibrio tritonius WP 068716984
Vibrio xiamenensis SDH51215
Vibrio hangzhouensis SEF62326
Vibrio natriegens WP 020333808
Shewanella mangrovi WP 037441799
Vibrio diazotrophicus WP 042479509
Proteobacteria bacterium OIO53742
HBD-06 PON MAG 00028 *
HBD-06 PSE MAG 00030 *
HBD-06 PSW MAG 00031 *
HBD-06 ANW MAG 00006
HBD-07 ANW MAG 00019
HBD-07 ANE MAG 00075 *
HBD-05 PSE MAG 00116
HBD-03 ION MAG 00014
HBD-04 PON MAG 00023
Pseudomonas benzenivorans SDH13122
Pseudomonas sagittaria SFQ02917
Pseudomonadales bacterium OHC65406
Pseudomonas oryzae SDS70900
Pseudomonas syringae WP 065835949
Azotobacter beijerinckii SFB09551
Azotobacter chroococcum AAA22140
Azotobacter vinelandii WP 012698831
Pseudomonas stutzeri WP 042921674
Alcaligenes faecalis Q44044
Pseudomonas azotifigens WP 028240788
Pseudomonas stutzeri WP 003294714
Marinobacterium litorale WP 027854782
Marinobacterium mangrovicola AHN13827
Marinobacterium lutimaris SEG25851
Marinobacterium sp. WP 036190066
Candidatus Thiodiazotropha WP 068994847
Cyanothece sp. WP 008274042
Gloeothece sp. BAF47151
cyanobacterium endosymbiont AAW57049
Candidatus Atelocyanobacterium KFF41299
Candidatus Atelocyanobacterium ASW MAG 00003
Planktothrix tepida CUR34599
Synechococcus WP 011430651
Frankia sp. WP 055750680
Hydrogenobacter thermophilus WP 012963773
Desulfotignum balticum WP 024336028
Desulfobacterium autotrophicum WP 015902868
Desulfospira joergensenii WP 022667737
Desulfobacula sp. OGR13489
HBD-02 PSW MAG 00026
Desulfobacter vibrioformis WP 035237762
Desulfovibrio aespoeensis WP 013513968
Desulfovibrio piezophilus WP 015416172
Desulfovibrio sp. WP 066802111
HBD-01 PSW MAG 00013
Desulfovibrio dechloracetivorans WP 071544382
Desulfovibrio desulfuricans WP 014321935
Desulfomicrobium baculatum WP 015773052
Desulfomicrobium norvegicum SFM03851
Desulfomicrobium escambiense WP 028577402
Desulfarculus baarsii WP 013257272
Desulfocarbo indianensis WP 049675847
HBD-09 PSW MAG 00018
HBD-09 PSE MAG 00060 *
HBD-08 PSW MAG 00130
Candidatus Achromatium WP 048398681
Coraliomargarita akajimensis WP 013044574
Diplosphaera colitermitum WP 043581909
Opitutaceae bacterium WP 068768887
Terrimicrobium sacchariphilum GAT32889
Verrucomicrobiae bacterium WP 040901465
Dehalococcoides mccartyi WP 010936850
Desulfobacterales bacterium OEU52425
Dehalogenimonas sp. AKG53288
Root: Archaeal genomes


Total distribution of nine HBDs (%)
0.0%
0.2%
0.4%
0.6%
0.1 million cells per litre*
less than 5,000 cells per litre*
3.2 million cells per litre*
ANE
ANW
ASE
ASW
ION
IOS
MED
PON
PSE
PSW
RED
SOC
Atlantic Ocean
Indian Ocean
Pacific Ocean
HBD-06
Ca. Nitroaequaella pragmatica
0.0% 0.1% 0.2% 0.3%
HBD-08
Phylum Planctomycetes
0% 0.02% 0.04% 0.06%
HBD-07
Ca. Azomonabacter oceanus
0% 0.01% 0.02% 0.03% 0.05%
HBD-09
Ca. Nitroplanctobacter absconditus
0% 0.04% 0.08% 0.12% 0.16%

PART II:

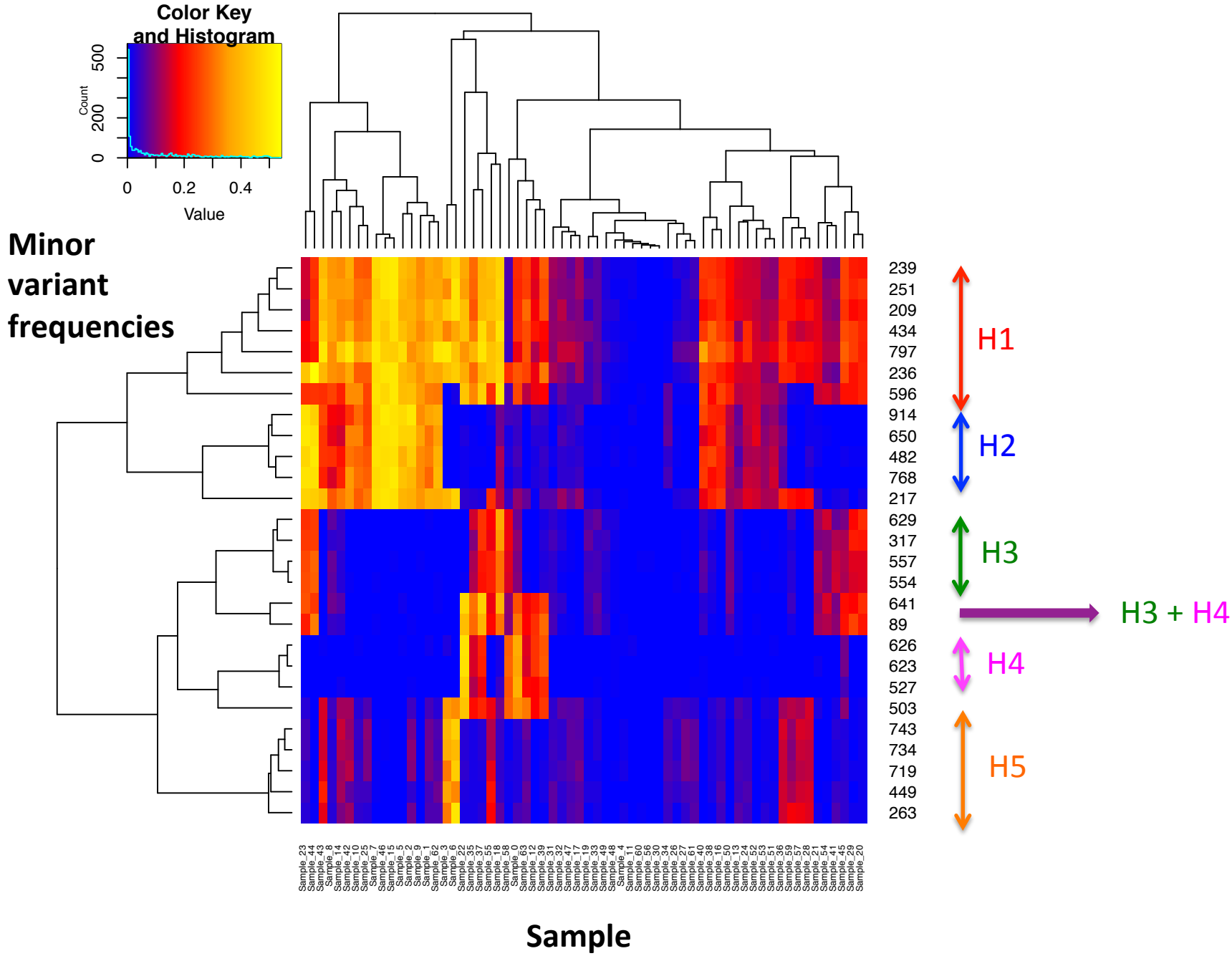
DESMAN

Resolving Intra-MAG diversity

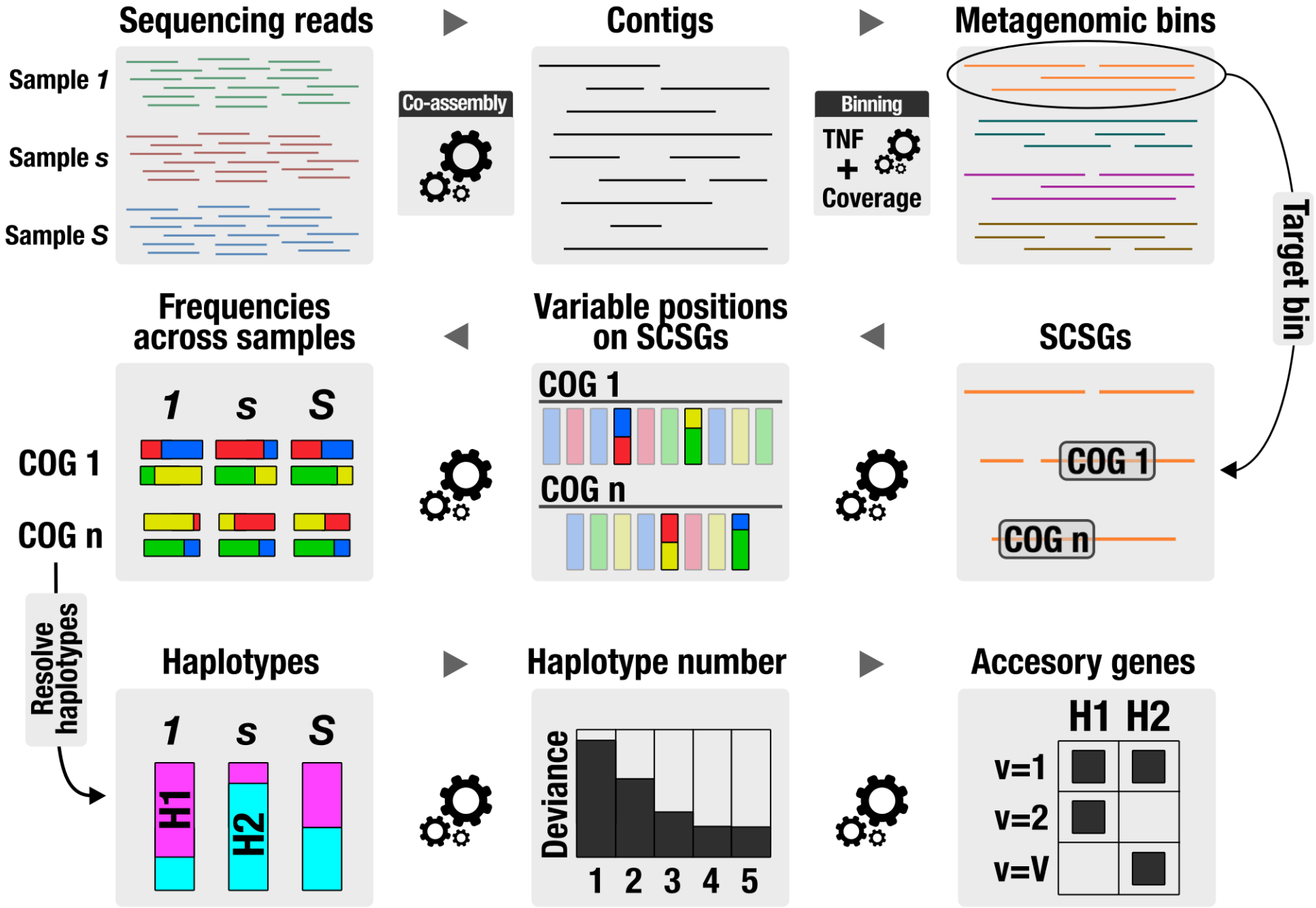
Resolving intra-MAG diversity

- Additional variation exists within MAGs both nucleotide variants on shared genes and variation in the accessory genome
- Read based haplotype resolution cannot span contigs or regions of low variation
- Methods exist using co-occurrence across multiple samples to resolve strain mixtures *de novo* after mapping to references:
  - Constrains (Luo et al. Nature Biotech. 2015)
  - Lineage (O'Brien et al. Genetics 2014)

Linking variants by co-occurrence



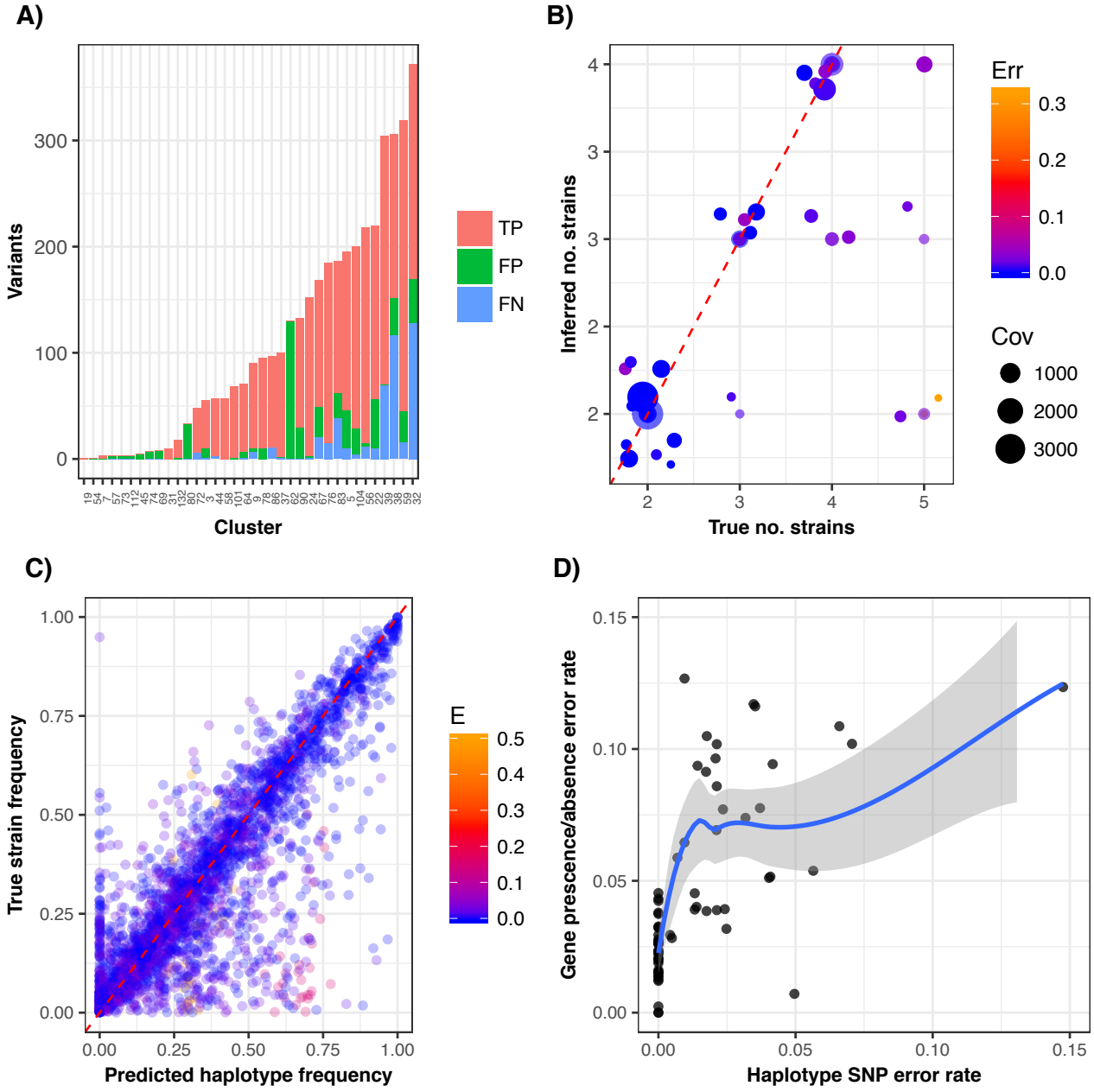
DESMAN https://github.com/chrisquince/DESMAN



Quince, ..., Eren Genome Biology 2017 18:181 https://doi.org/10.1186/s13059-017-1309-9

Complex synthetic community

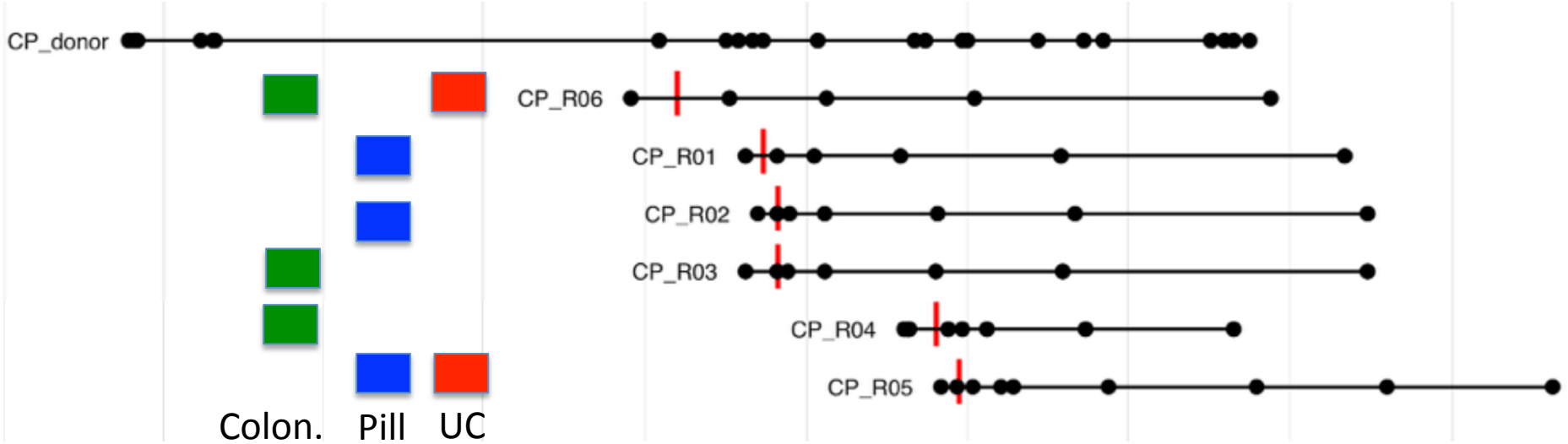
- Detected variants with a mean precision of 92.32% and a mean recall of 91.85%
- Predicted the correct haplotype number for 18/25 (72%) of the clusters
- SNV error rate: median of 0.25% and a mean of 2.38%
- True frequency against predicted gave a slope of 0.820 (R-squared 0.741, p < 2.2e-16)
- Overall accessory gene prediction accuracy was 95.7%.



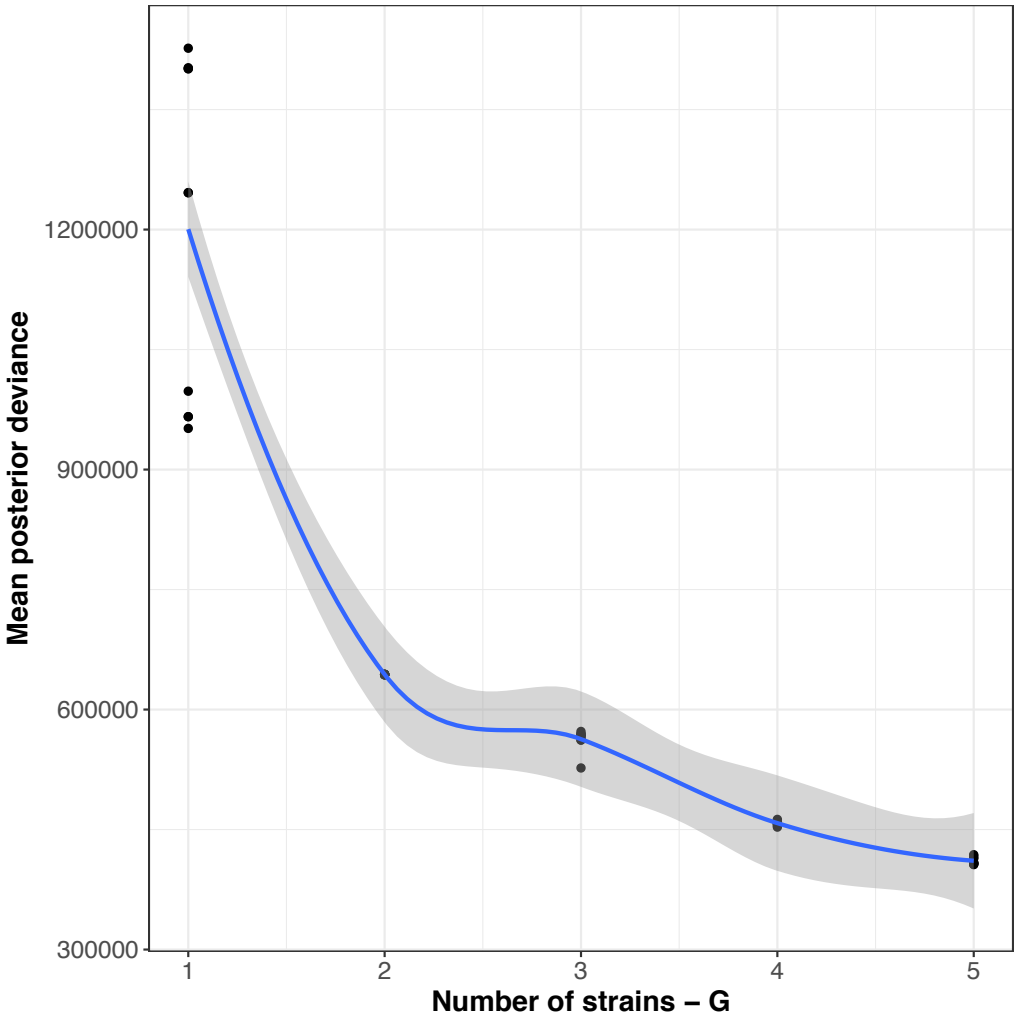
Fecal Microbiome Transplants

(Andrea Watson, A. Murat Eren)

- Healthy donors, CP, 24 samples 2 years
- 6 recipients, with *C. difficile* infection, and 2 also diagnosed with ulcerative colitis.
- Half received FMT through pill, and the other half received FMT through colonoscopy.
- For each of the recipients we have at least one sample from pre-FMT, and a median of 5 samples taken post-FMT over the course of approximately a year.



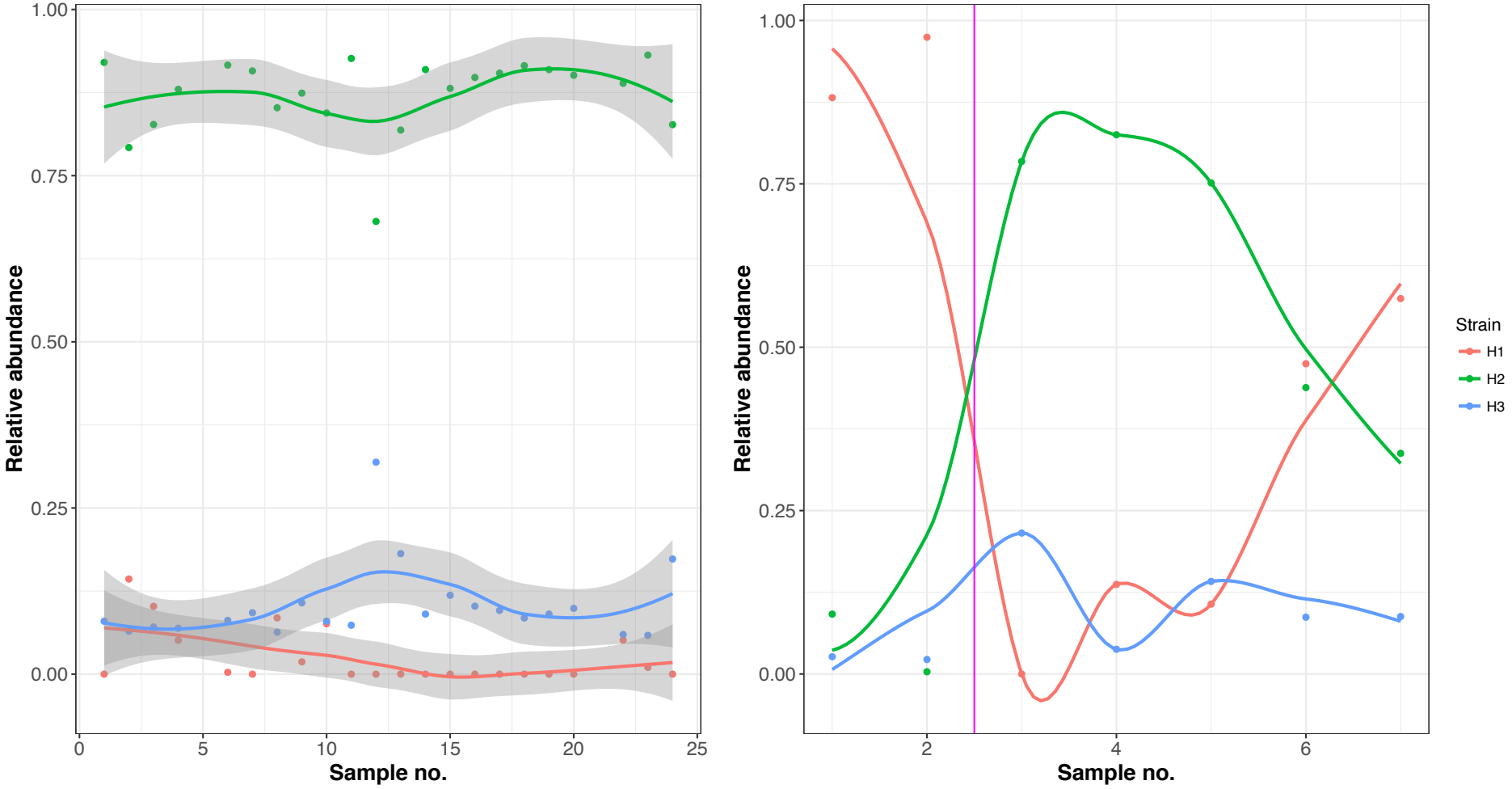
FMT DESMAN strain analysis



- Detected variants on one Anvi'o MAG assigned to Allistipes Finegoldii
- Detected 41,544 variants on 2.8 Mbp
- Applied DESMAN to 1,000 variant positions increasing strain numbers from 1 to 5
- 4 strains selected as best fit, three of which were 'reproducible'
- Varying between 35 – 50% of SNV positions

CP Donor

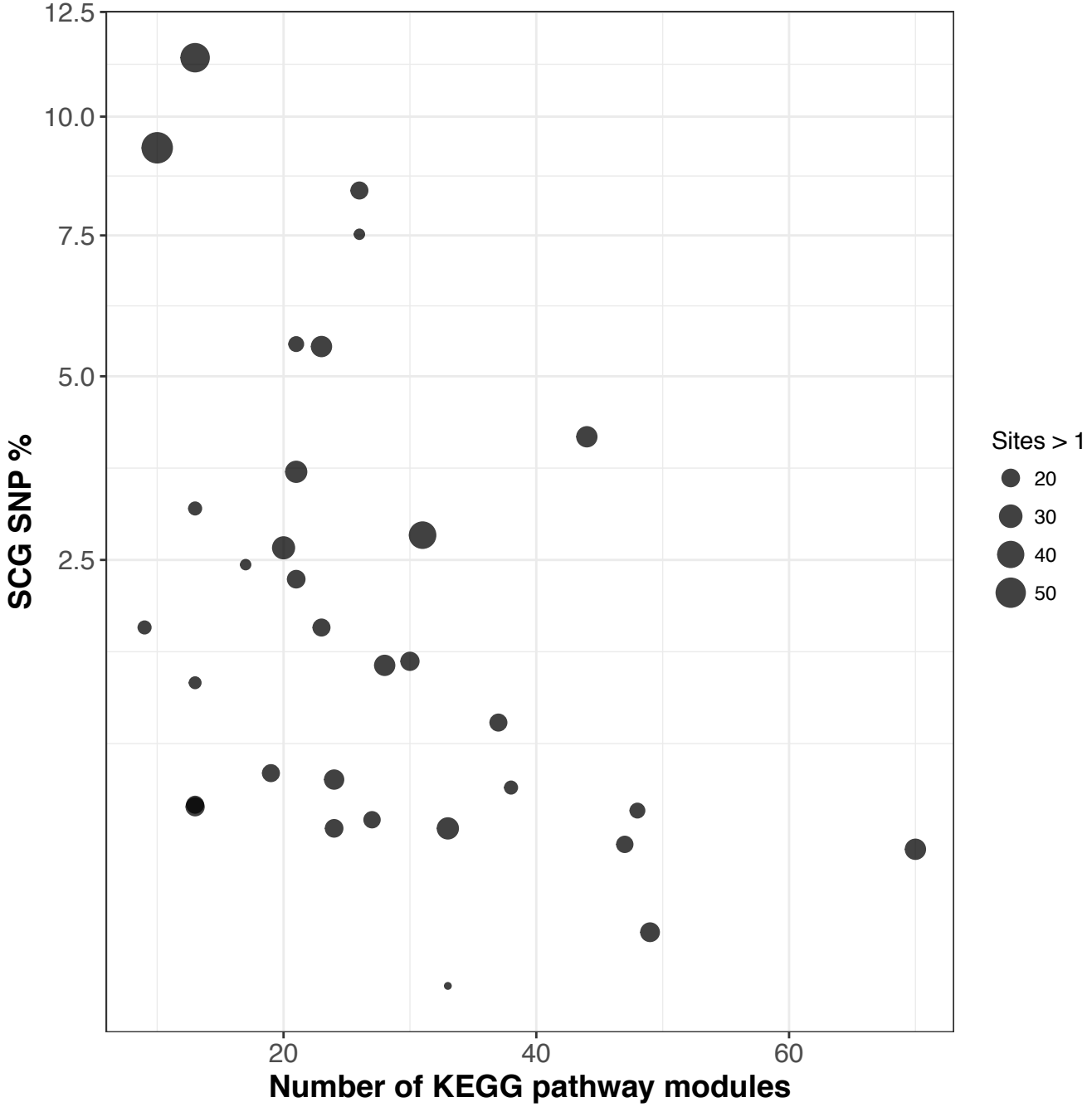
Recipient R03



Pre FMT

Post FMT

Tara MAG variants



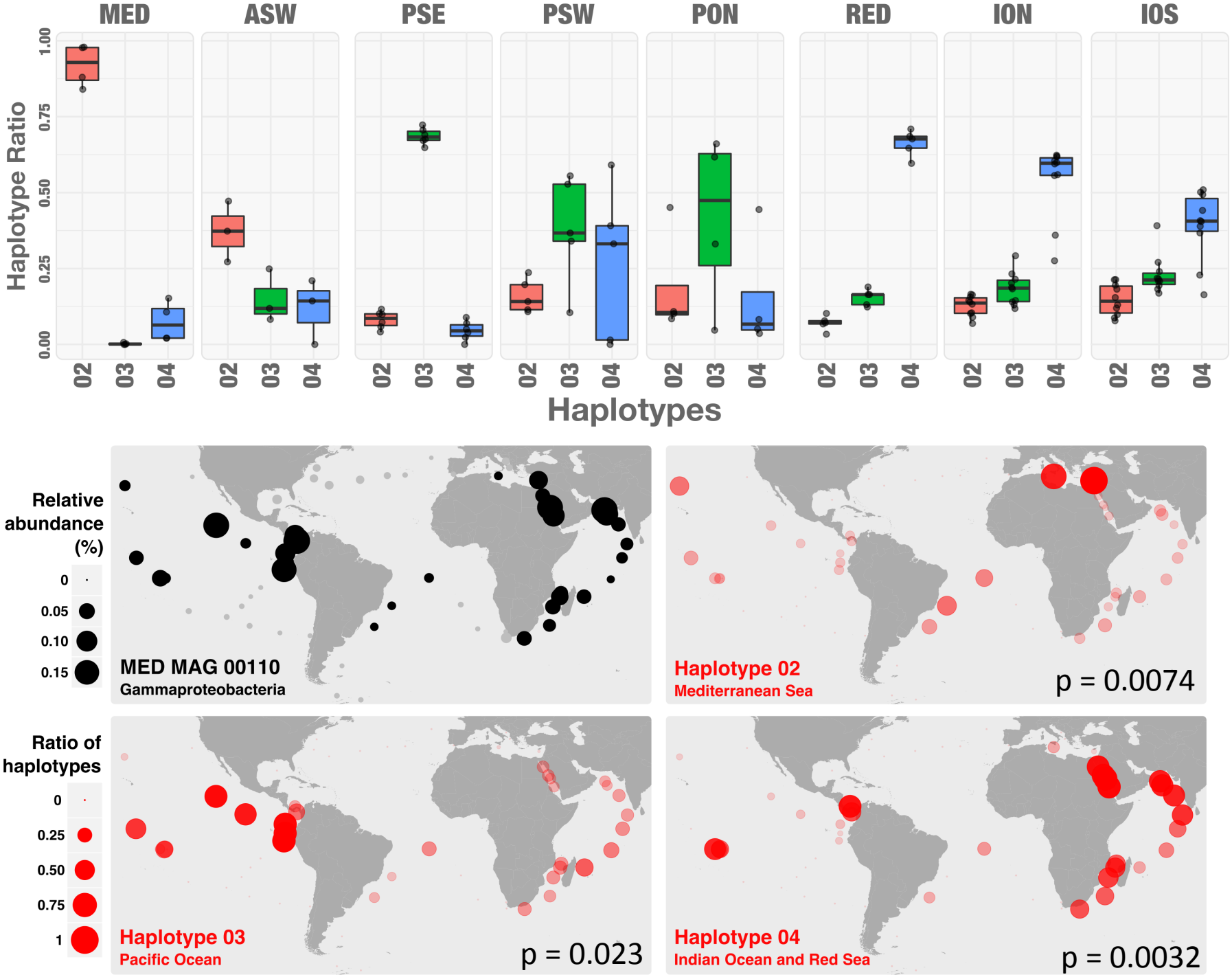
Consider 32 MAGs with cov. > 100.

Observe a negative correlation with genome length (Spearman's p = 0.016)

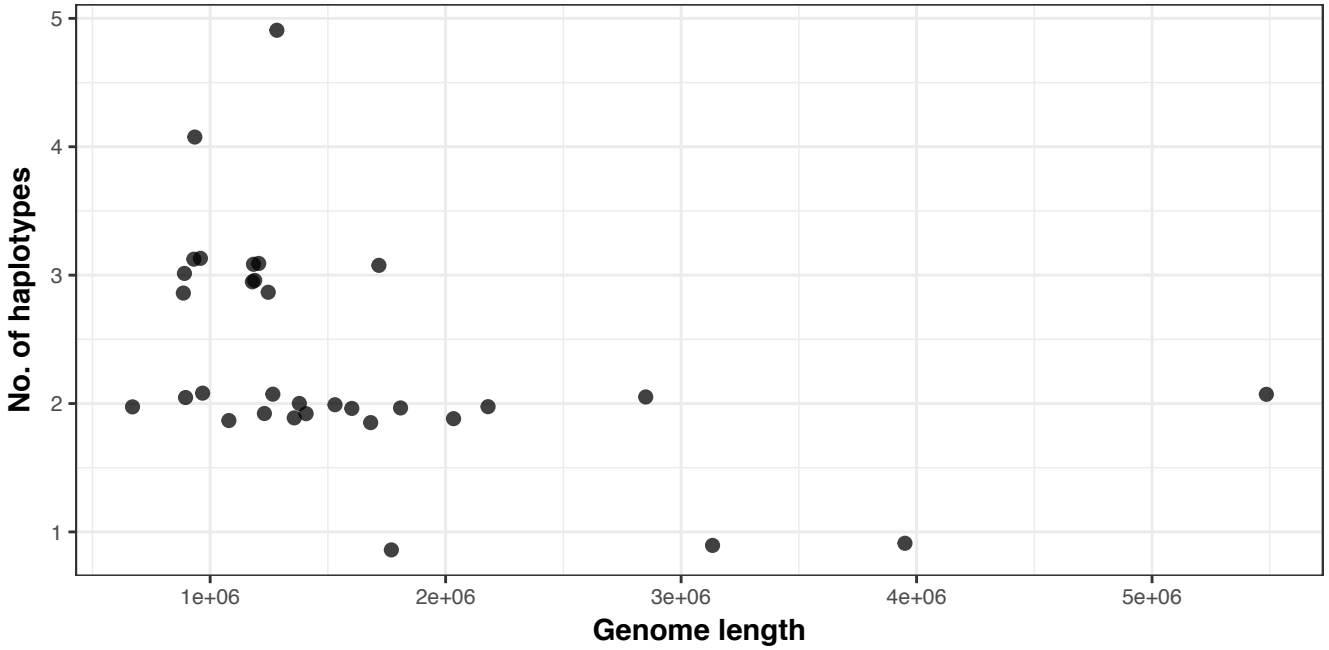
Stronger negative relationship with number of KEGG Pathway modules (Spearman’s p = 0.0045)

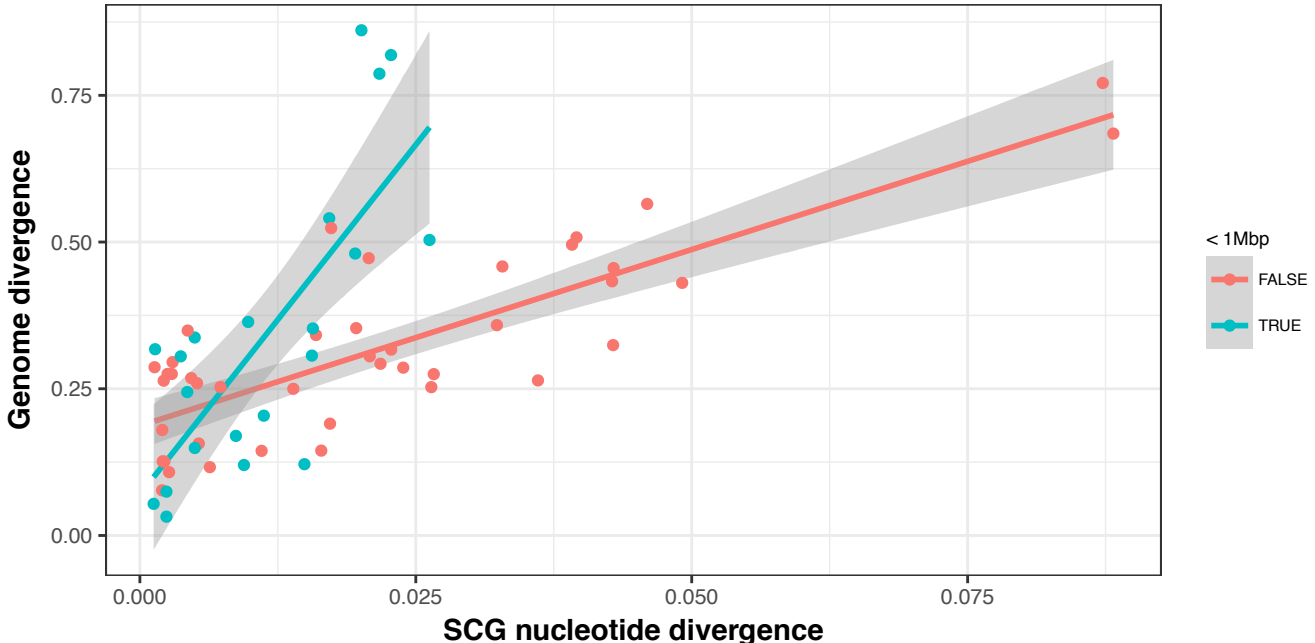
TARA DESMAN analysis

- Out of the 32 MAGs tested for haplotypes 29/32 had strain variation (1->3 2->17 3->10 4->1 5->1)
- The haplotypes were geographically localised e.g. TARA MAG 00110 a Proteobacteria with a highly streamlined 890,789 bp genome
  - Large group of uncultured organisms (relatives *Candidatus Evansia muelleri* and *Riesia pediculicola*)
  - Three haplotypes that differed by around 2% ANI on core genes and between 79-86% of accessory genes at 5% ANI clusters


MED
ASW
PSE
PSW
PON
RED
ION
IOS
Haplotype Ratio
0.00
0.25
0.50
0.75
1.00
02
03
04
Haplotypes
Relative abundance (%)
0
0.05
0.10
0.15
MED MAG 00110
Gammaproteobacteria
Haplotype 02
Mediterranean Sea
p = 0.0074
Ratio of haplotypes
0
0.25
0.50
0.75
1
Haplotype 03
Pacific Ocean
p = 0.023
Haplotype 04
Indian Ocean and Red Sea
p = 0.0032

TARA DESMAN results





- Significant negative correlation between strain number in MAG and genome length (p = 0.000068)
- 42/73 = 57% of inferred strains had a significant correlation with geographic region
- Significant interaction (p = 3.51e-06) between rate of whole genome divergence relative to core gene divergence and highly streamlined genome (<1Mb)

Summary

- Strain diversity is endemic in environmental and host associated communities
- Now have the tools to generate hundreds or even thousands of genomes from metagenomes
- CONCOCT followed by DESMAN can resolve both species and strain diversity *de novo* but with limitations
- The question is then what will we do with all this data!
  - Machine learning based trait inference from genomes https://www.biorxiv.org/content/early/2018/04/25/307157

Acknowledgements

- Medical Research Council funding and CLIMB, BBSRC/Unilever
- Tom Delmont and A. Murat Eren (U. of Chicago)
- Leonidas Souliotis and Sebastian Raguideau (Warwick)