Strain resolution from metagenomes

Dr Christopher Quince

University of Warwick

Introduction

- What is strain resolution
- Why is it important?
- De novo strain resolution vs reference based

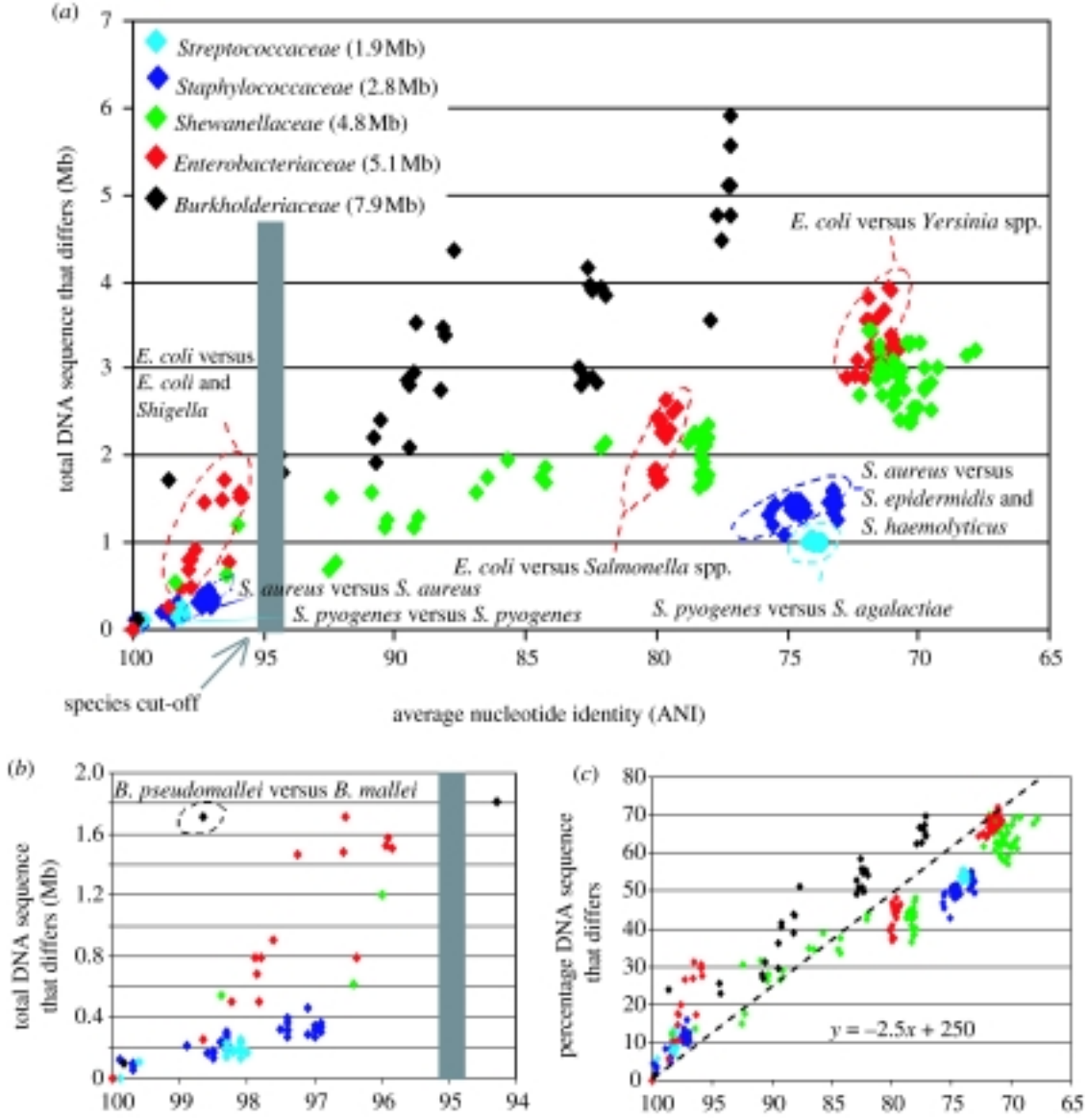
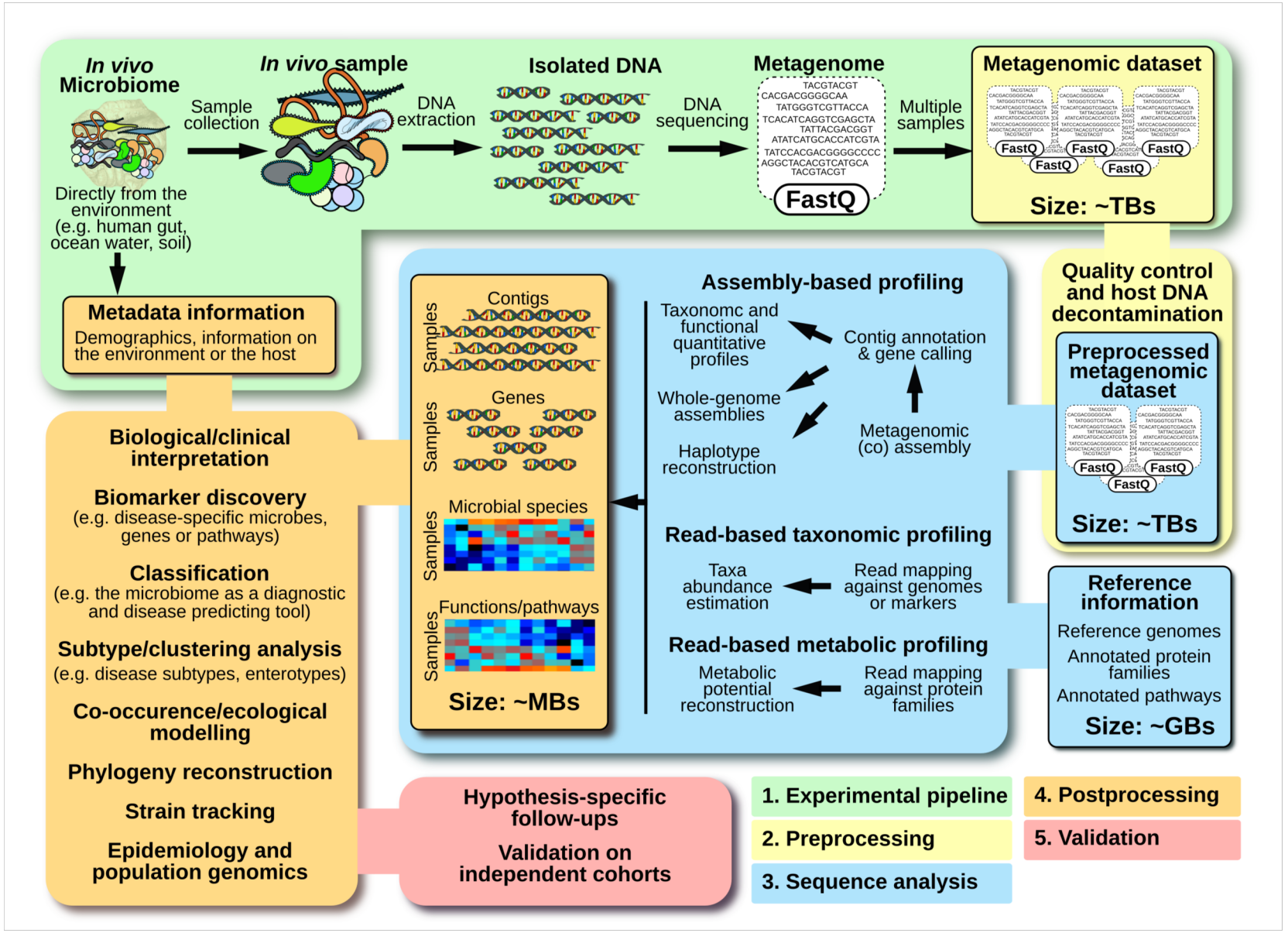


Figure 1: 'The bacterial species definition in the genomic era' Konstantinidis et al. 2005

Overview

- De novo vs reference based strain resolution
- DESMAN: Resolving strains using variant frequencies across multiple samples
  - Application to FMT
  - AD reactors
  - TARA Oceans
- STRONG: Strain resolution on graphs



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

De novo vs. reference based strain resolution

- Reference based strain resolution can only be used for well studied organisms:
  - https://enterobase.warwick.ac.uk/ (200,000+ Salmonella genomes, 100000+ E coli/Shigella)
- Given set of reference genomes some methods exist that will profile from reads to below the species level
  - Pathoscope, mSWEEP, SPARSE - https://github.com/zheminzhou/SPARSE
- Why can’t we simply assemble strain genomes de novo?

Why is metagenomics assembly difficult?

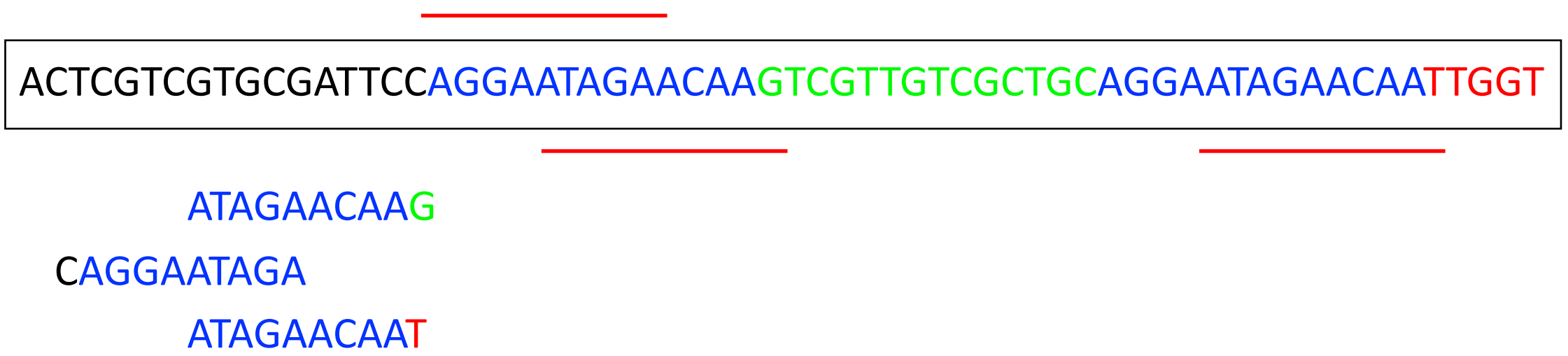
- Assembly consists of looking for overlaps between short reads

Read 1: ACTCGTCGTG
Read 2: GTGCGATTCC

ACTCGTCGTG
GTGCGATTCC

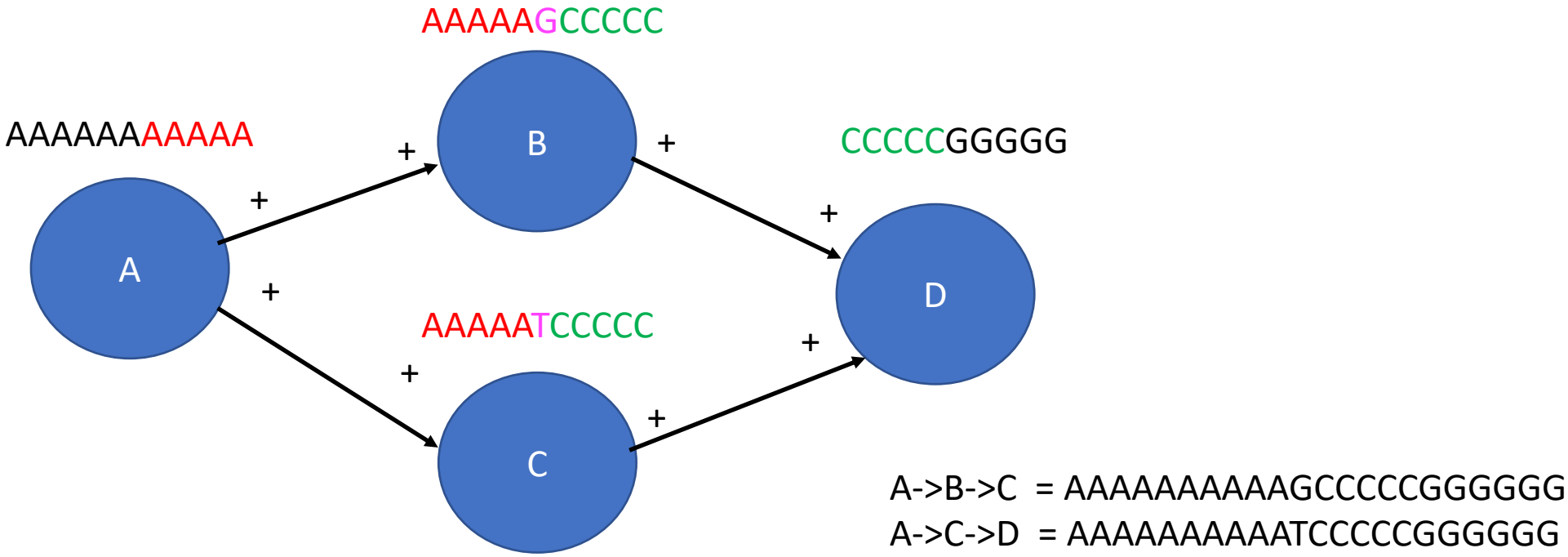
Contig: ACTCGTCGTGCGATTCC

- Challenge arises because of ‘repeat’ regions on genomes that exceed read length

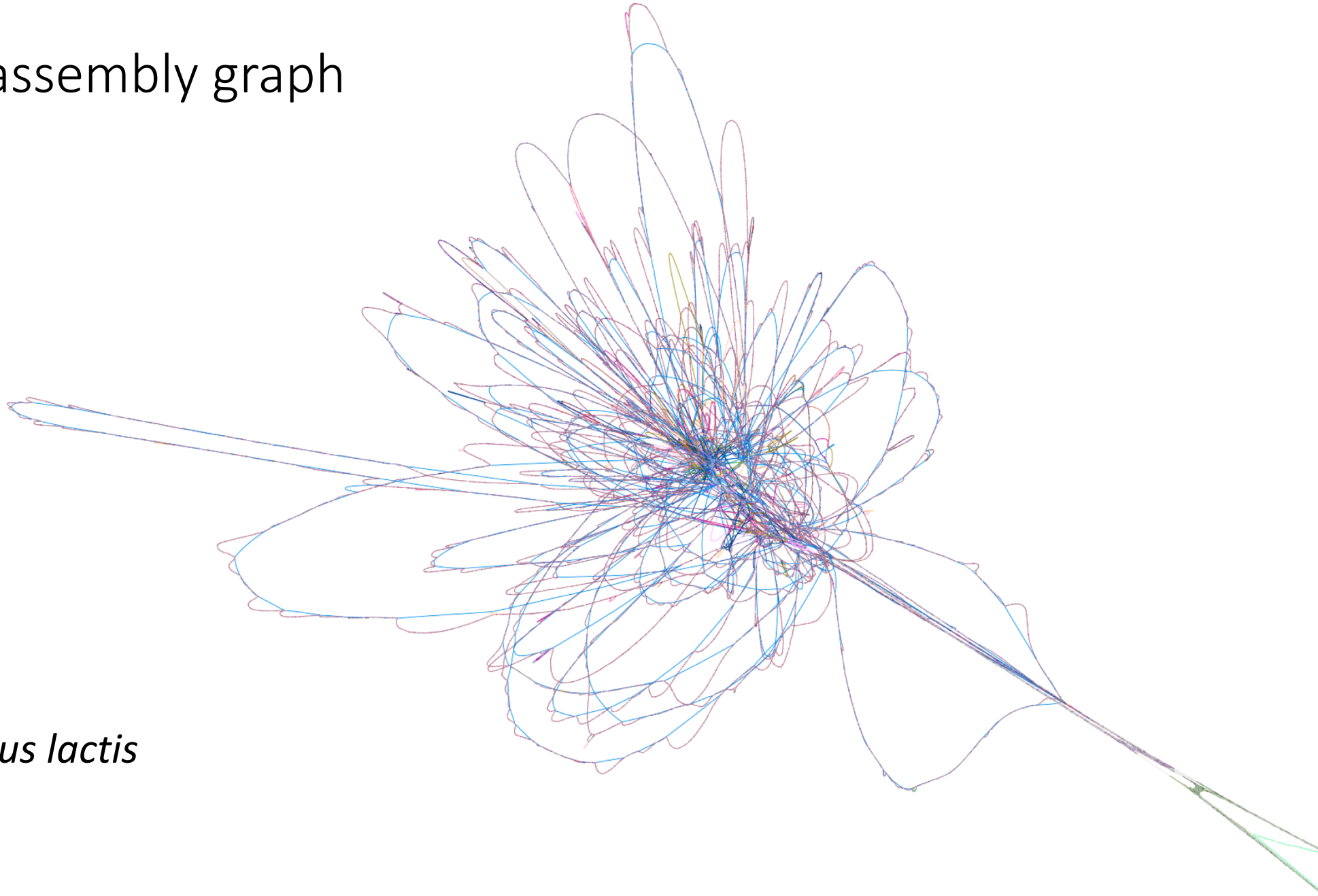


Assembly graph represents uncertainty

- Produced by all assemblers following compaction de Bruijn graphs or removal of transitive reads string graphs
- Nodes represent sequence (unitigs) and edges overlaps
- Represents fundamental uncertainty in possible paths and hence sequences



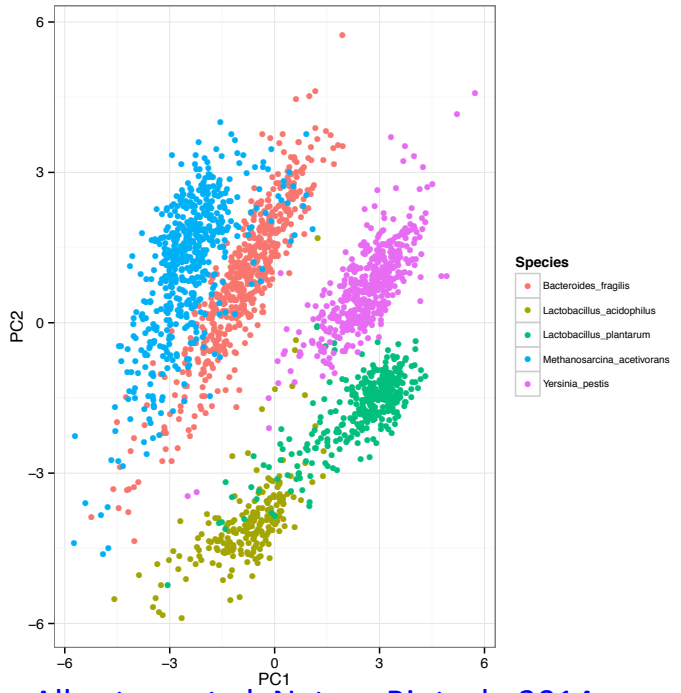
Metagenomics assembly graph



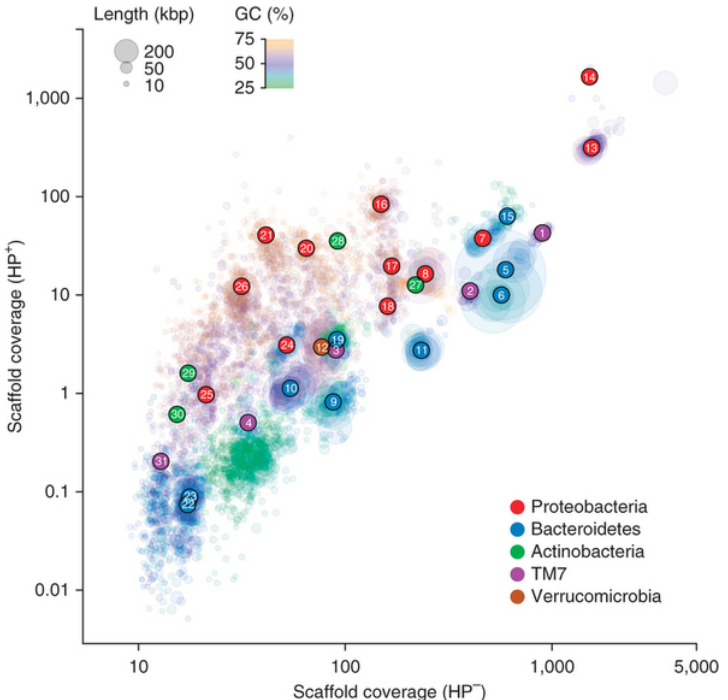
Six strains *Lactococcus lactis*

Contig clustering

- Use shared features across contigs to cluster into metagenome assembled genomes or MAGs:
  1) Composition – tetramer frequencies
  2) Coverage across multiple samples
- Algorithms:
  1) Human input e.g. ESOM and ANVIO (Meren et al. PeerJ 2015)
  2) Automatic e.g. CONCOCT (Alneberg et al. Nat. Methods 2014), MetaBAT (Kang et al. PeerJ 2015), MaxBin2 (Wu et al. Bioinf. 20150



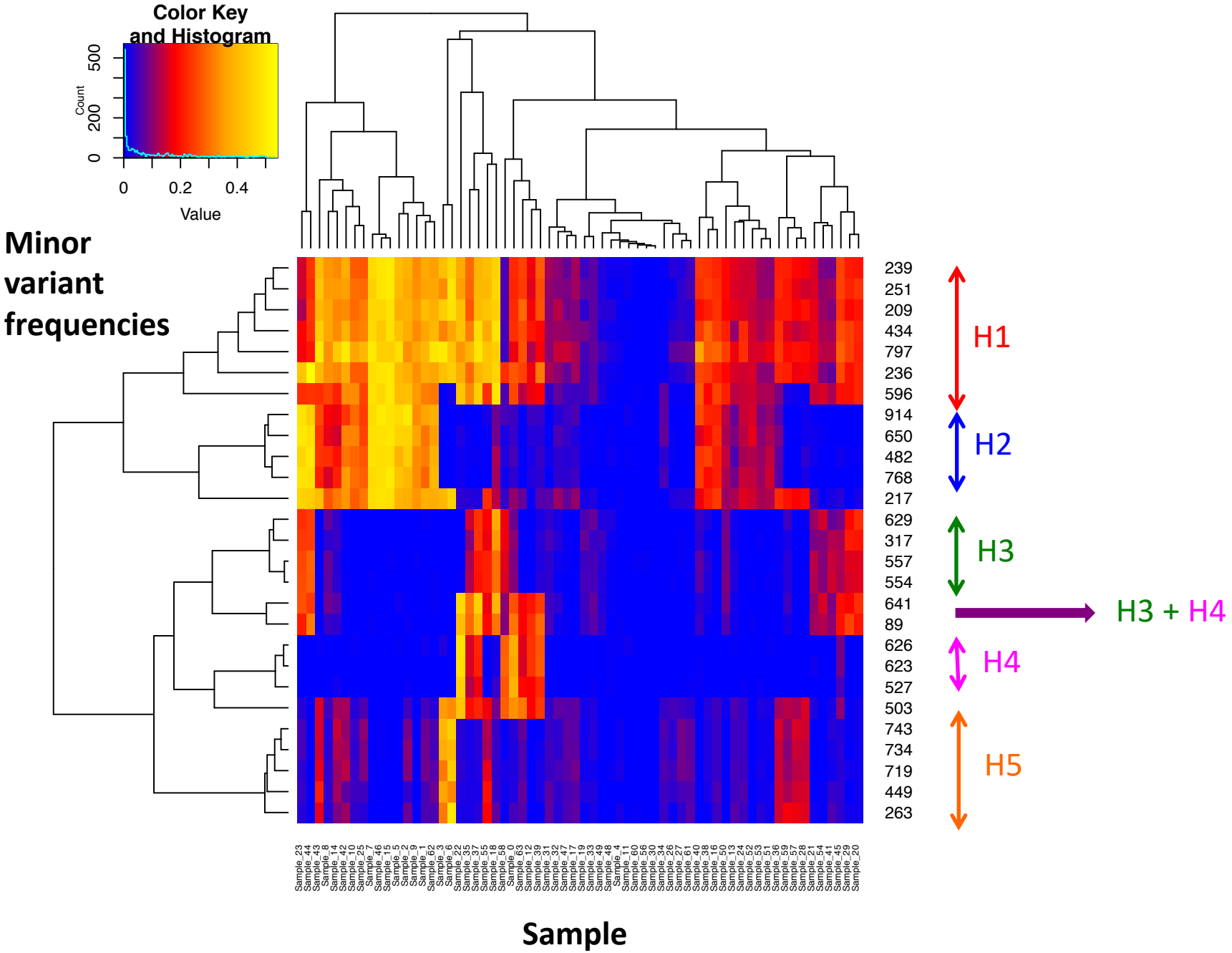
Albertsen et al. Nature Biotech. 2014



Resolving intra-MAG population diversity

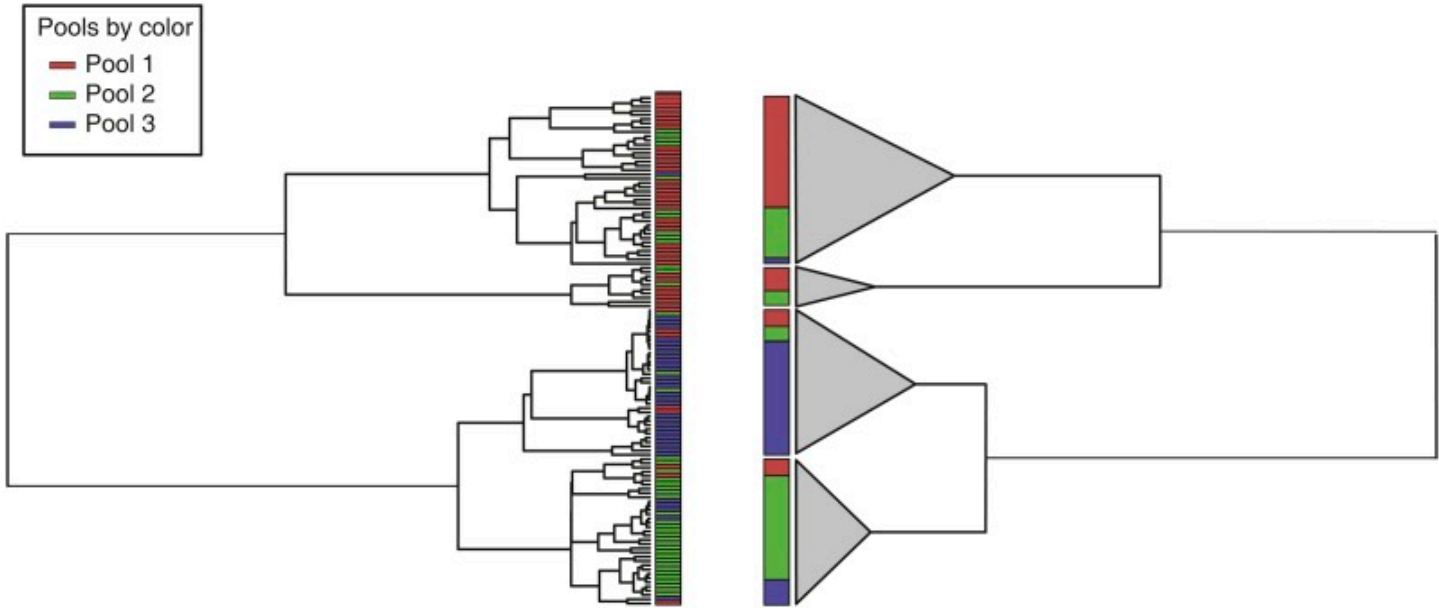
- What level of phylogenetic resolution does a MAG correspond to?
  - Does it differ if bins generated from coassembly vs single sample assembly
- Additional variation exists within MAGs both nucleotide variants on shared genes and variation in the accessory genome
- Variation of two sources, segregating variation within a population and sub-populations
- Methods exist using co-occurrence across multiple samples to resolve sub-populations *de novo* after mapping to references:
  - Constrains (Luo et al. Nature Biotech. 2015)
  - Lineage (O'Brien et al. Genetics 2014)
- In DESMAN we extend this to contigs

Linking variants by co-occurrence



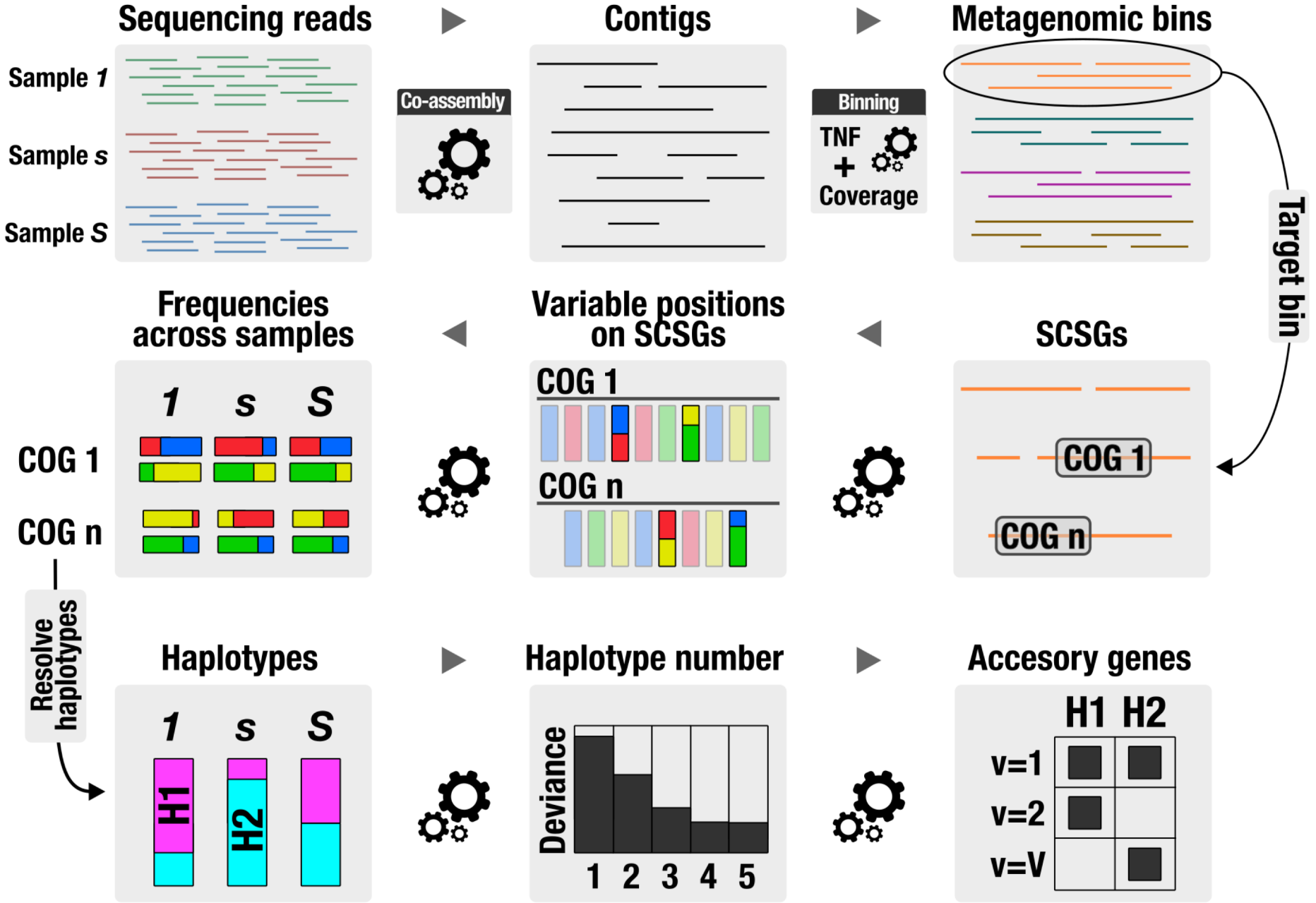
What are DESMAN sub-populations?

- Signal resolvable through SNV co-occurrence:
  - Distinct populations within a MAG
  - Finer scale variation if it is structured
- What cannot be resolved:
  - Unstructured fine scale variation
- Clonal expansions, niche selection and mutation will generate structured variation but recombination and migration will dilute it
- Ecologically relevant patterns should be resolvable but not always to actual haplotypes or genomes…



O' Brien et al. Genetics 2014

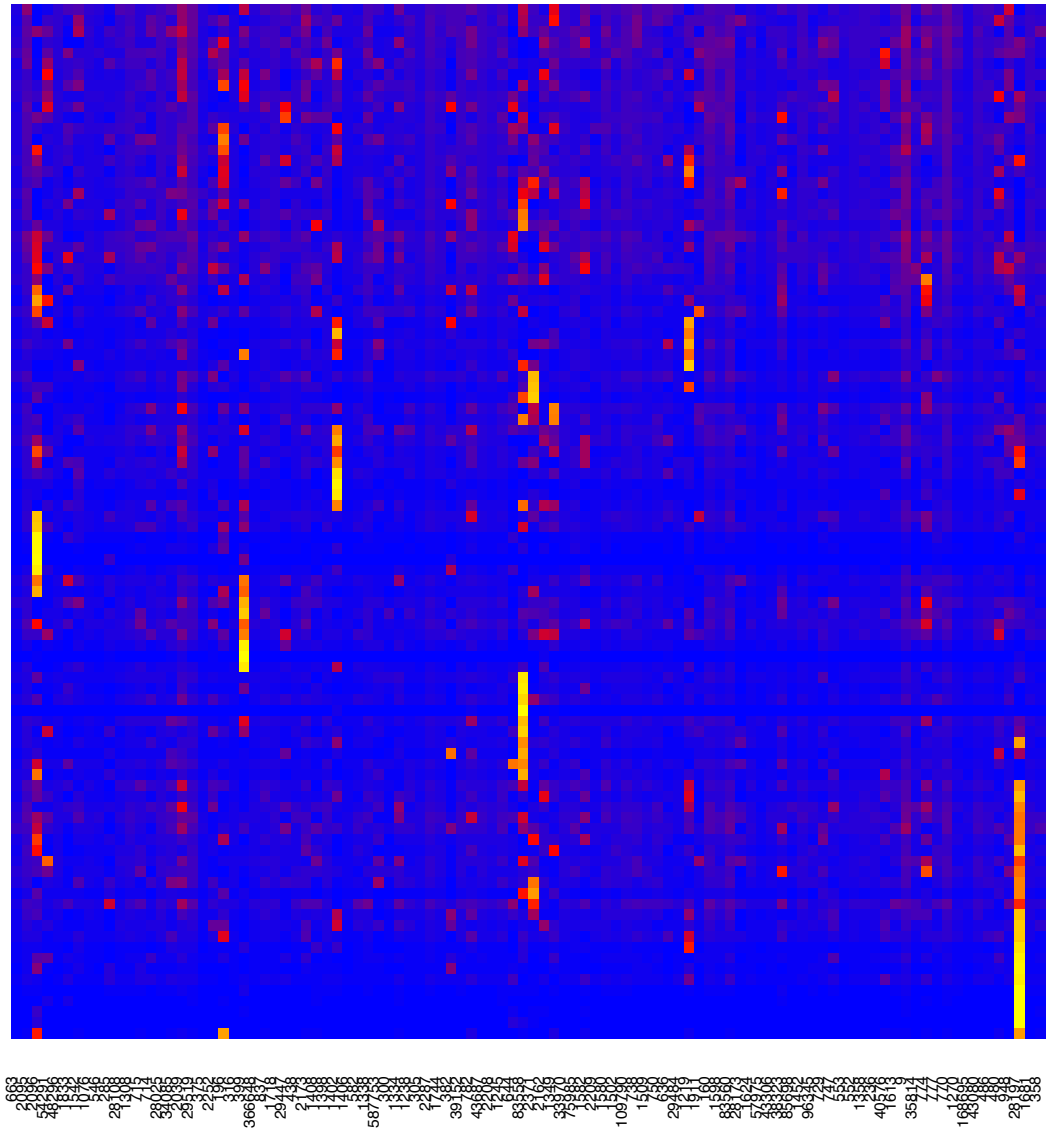
DESMAN https://github.com/chrisquince/DESMAN



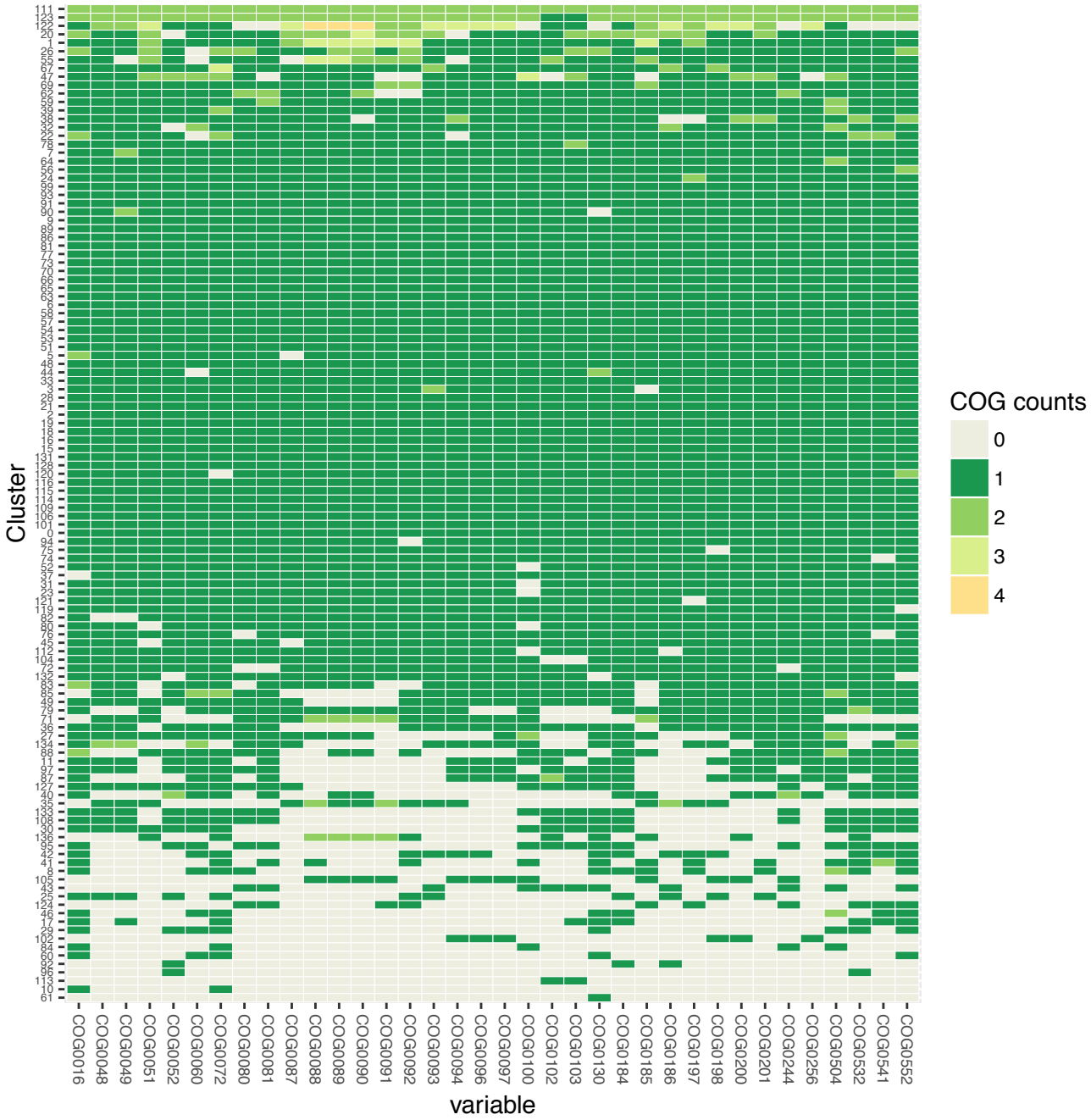
Quince, …, Eren Genome Biology 2017 18:181 https://doi.org/10.1186/s13059-017-1309-9

Complex synthetic community

- 100 different species and 210 NCBI genomesSpecies 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
- Log-normal distribution for species across samples and 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 vs 687 Mbp for all 210 reference genomes

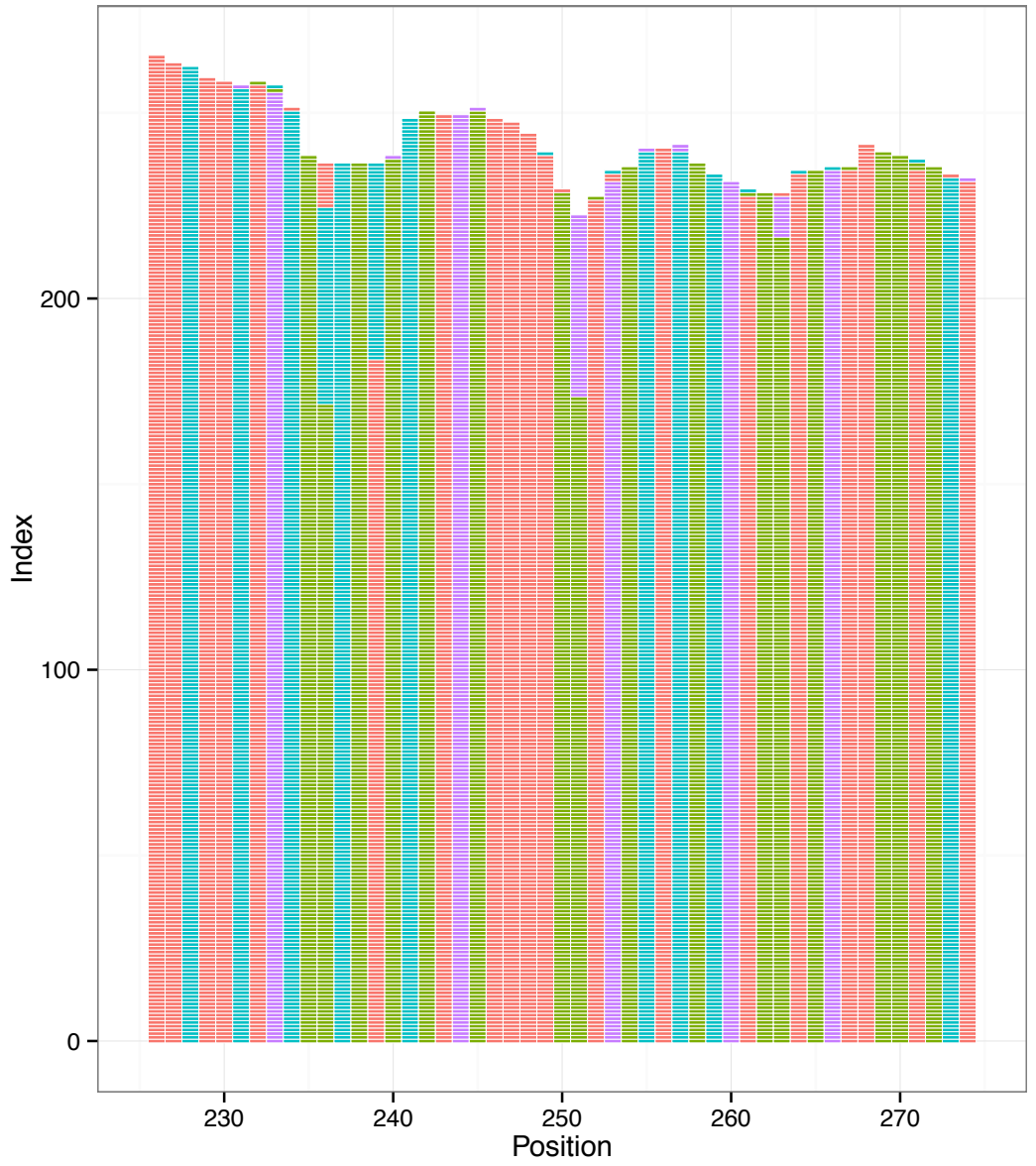


Concoct binning



- Coassembly gave 74,580 contig fragments with a total length of 409 Mbp (687 Mbp for all 210 genomes)
- CONCOCT generated 137 clusters (species recall of 86.1% and a precision of 98.2%)
- 75 clusters with 75% of SCGs in single copy

Determining variant positions on SCGs

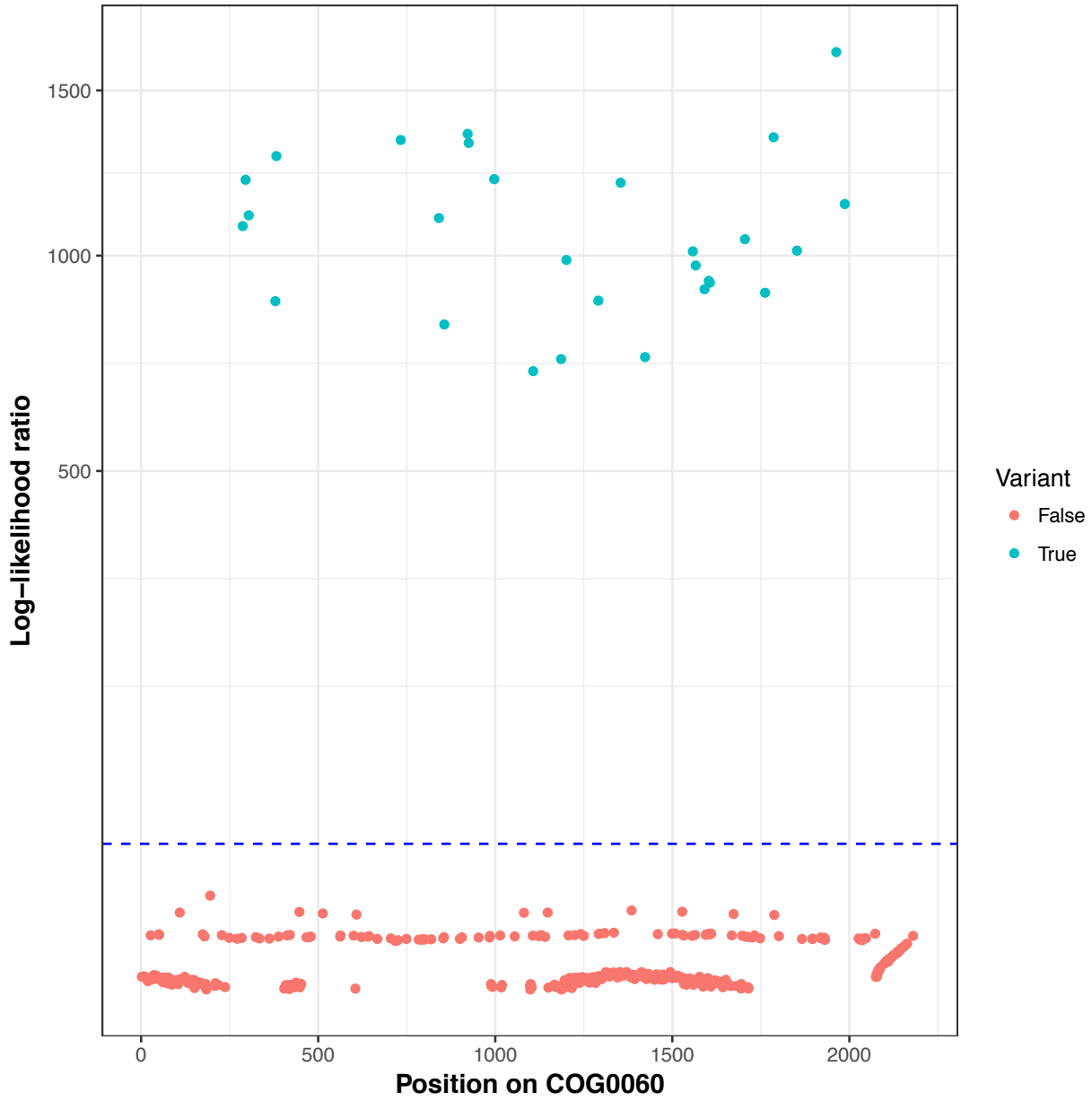


- Ignore distribution across samples just ask if minority bases could be created by errors alone
- Assume errors are position independent with true base *a* generating observed base *b* with probability:

$$\varepsilon_{a,b}$$

- Likelihood ratio test comparing null hypothesis that there is one variant present against two variants

Variant prediction for Cluster 37 -> Rhodococcus erythropolis



COG0060 predict 28 variant positions all correctly

Over all 27 SCGs and 26015 positions with q < 0.001

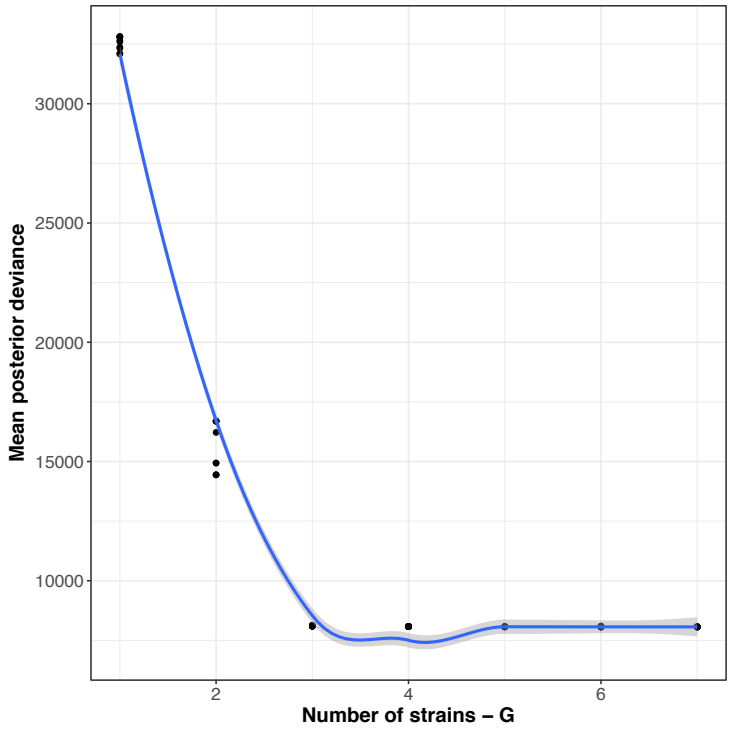
Predicted

	False	True
False	25015	1
True	1	98

Recall = 98.9%, Precision = 98.9%

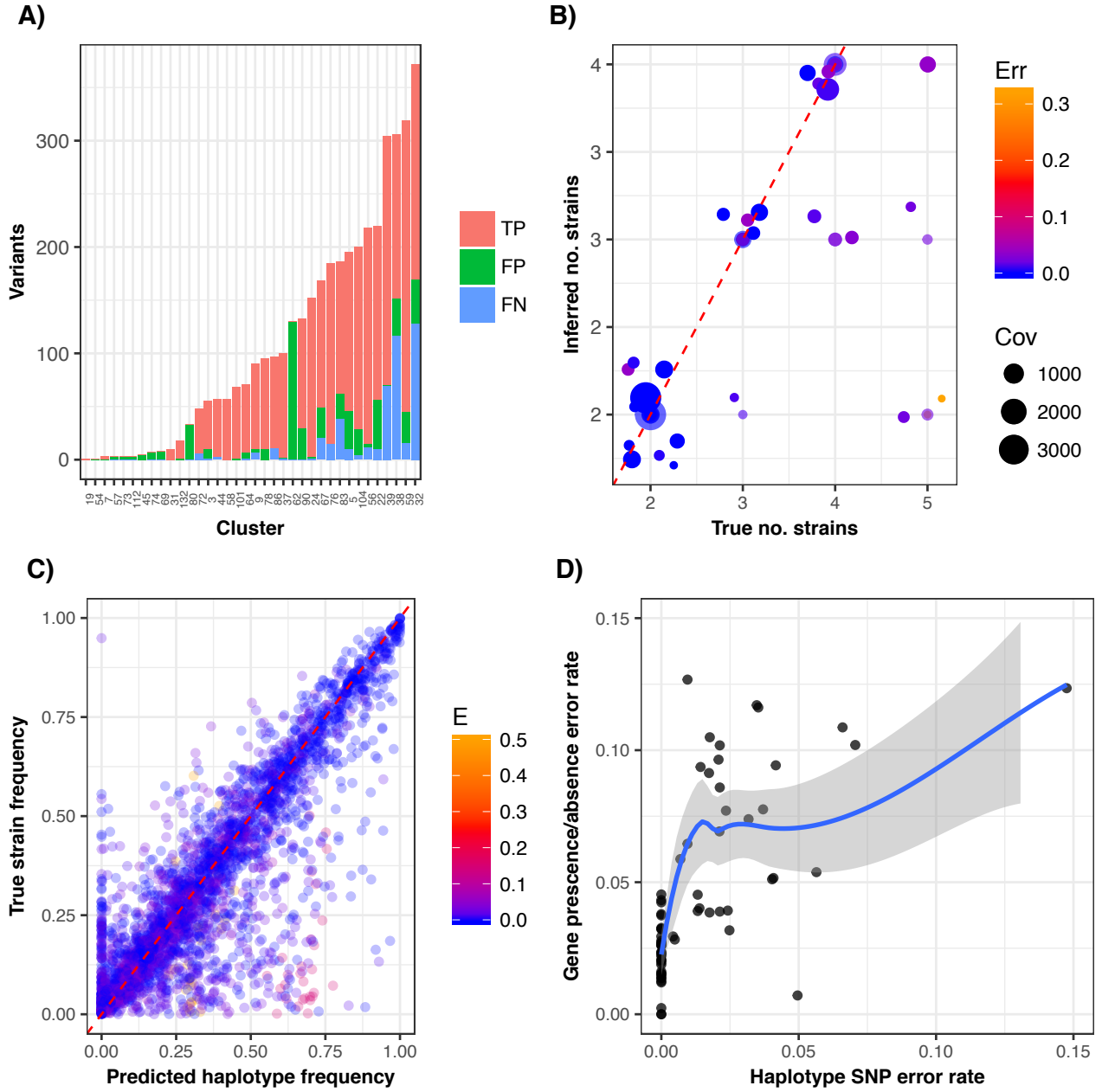
Gibbs sampling algorithm

- Assume bases are independent leads to binomial probability at each position
- Devise an iterative algorithm that successively samples:
  - Haplotypes
  - Relative frequencies
  - Error rates
- Bayesian algorithm generates distribution of fitted values
- Negative log-likelihood describes overall fit
- Heuristic for determining optimum haplotype number:
  - Determine haplotype number when fractional reduction in deviance below some value
  - Find value below or equal to this which gives the most strains with relative abundance > 5% and mean SNP uncertainty < 10%



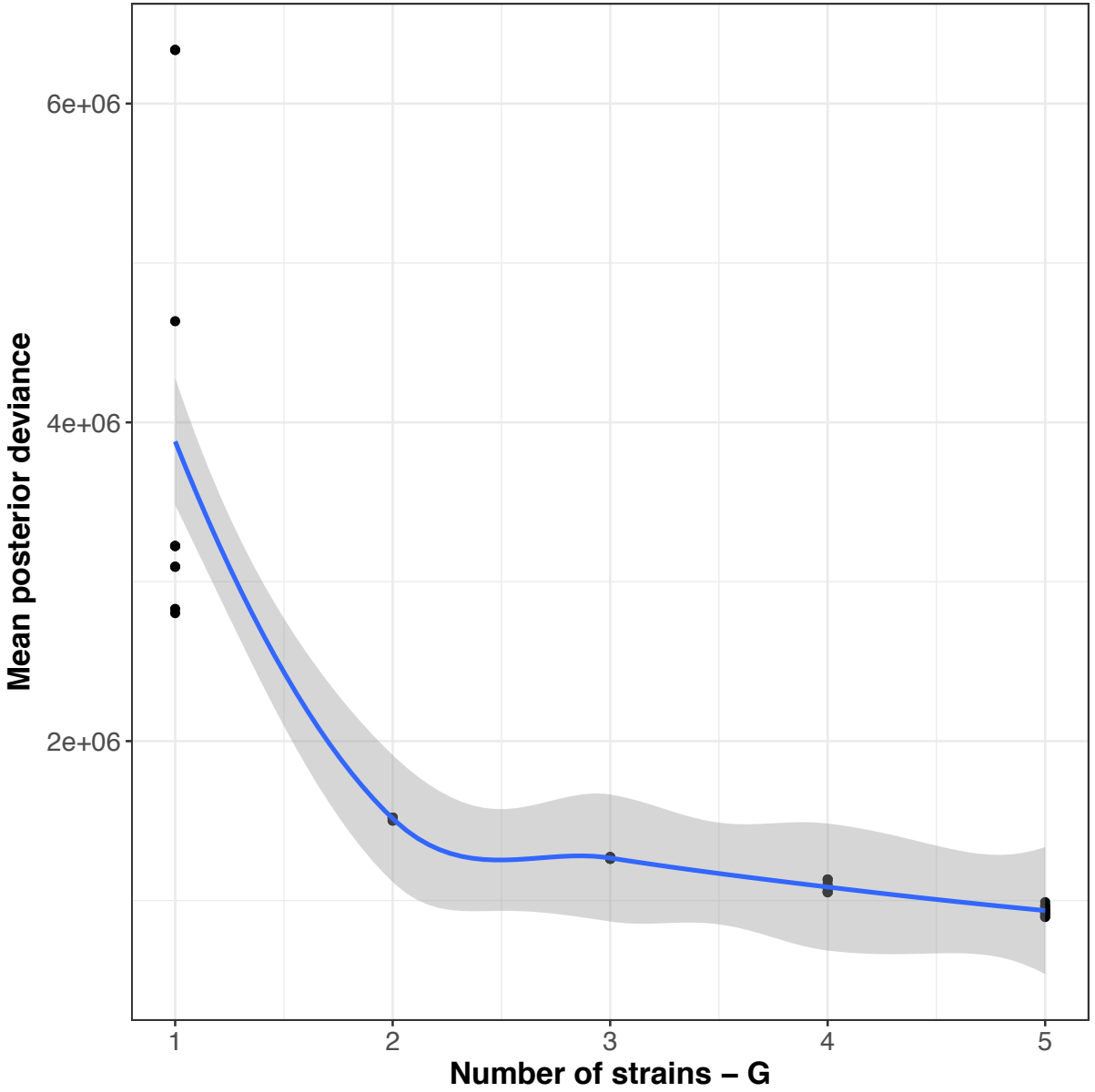
Complex synthetic community

- 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%.



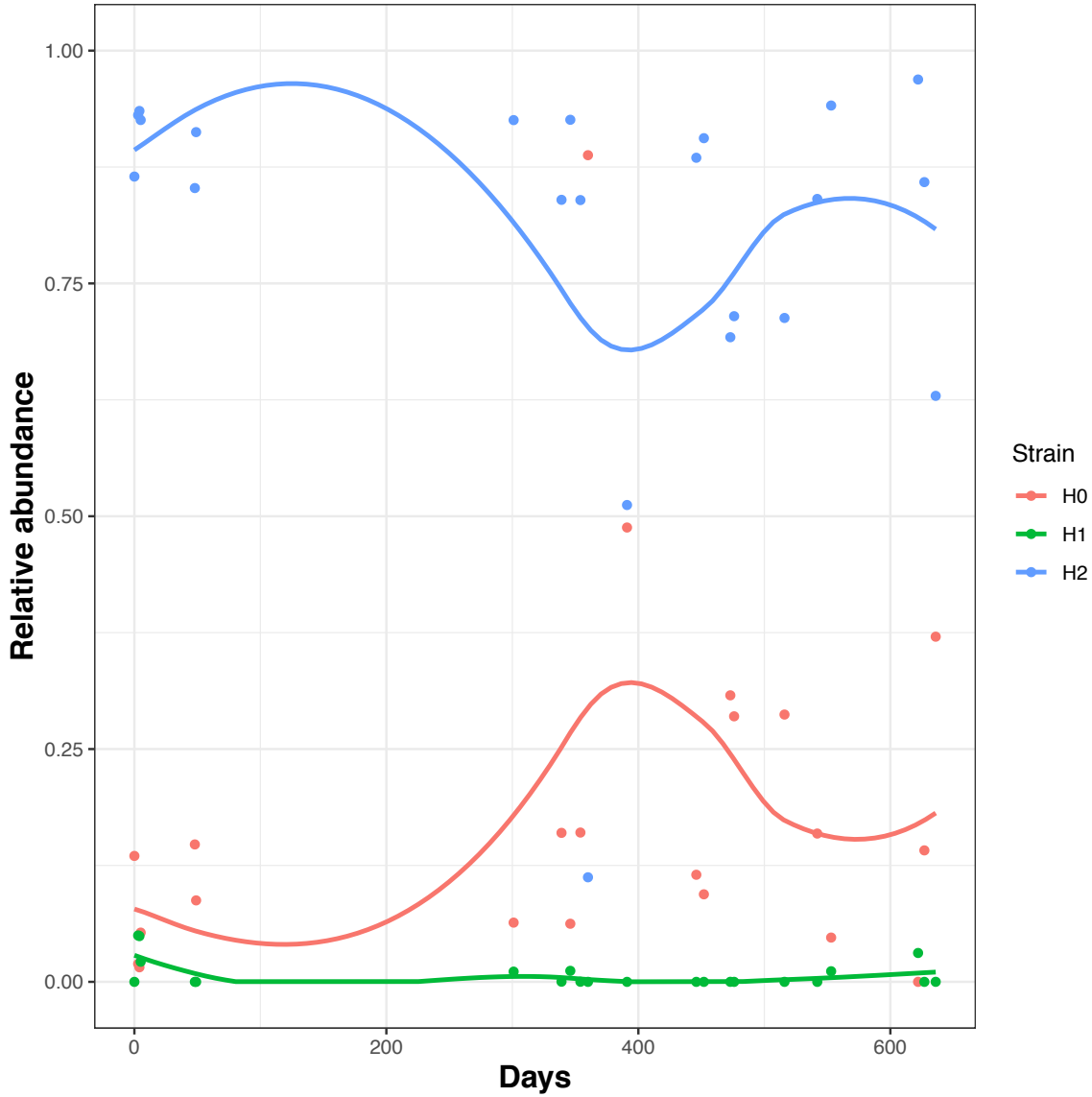
FMT DESMAN CP/R03 strain analysis

(Andrea Watson and A. Murat Eren)

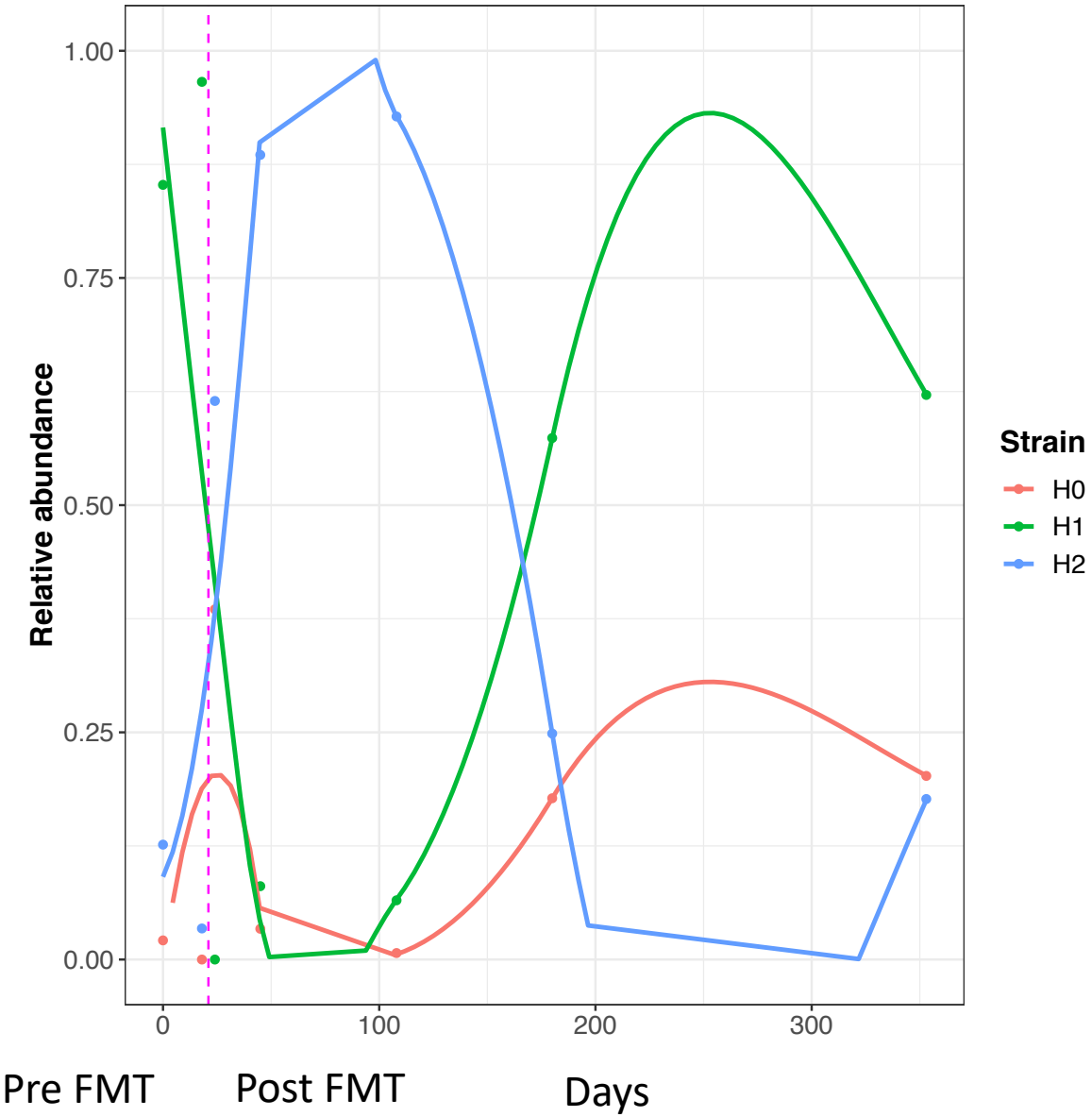


- Healthy donor, CP, 24 samples 2 years
- Recipients with *C. difficile* infection received FMT through colonoscopy
- One sample from pre-FMT, and 5 samples taken post-FMT over the course of a year
- Detected variants on one Anvi'o MAG assigned to Allistipes Finegoldii
- Detected 35,910 variants on 2.8 Mbp ~ 1.2%
- Applied DESMAN to 5,000 variant positions increasing strain numbers from 1 to 5
- 3 strains selected as best fit
- Varying between 23 – 53% of SNV positions

CP Donor

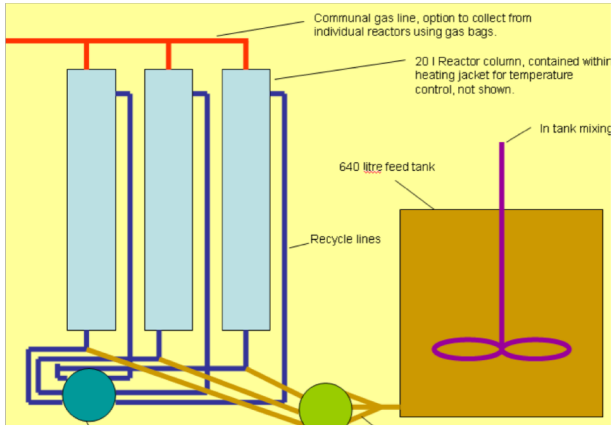


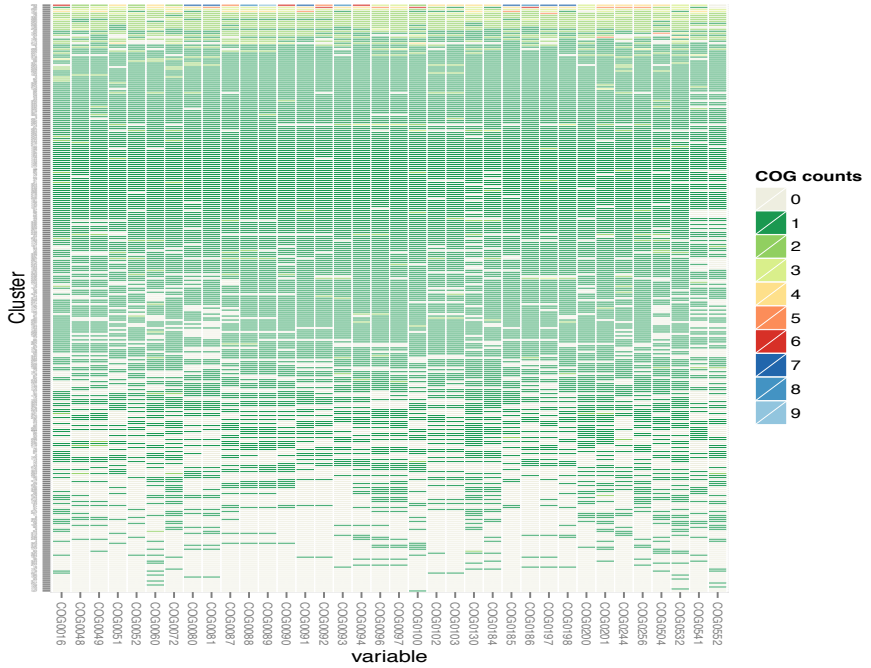
Recipient R03

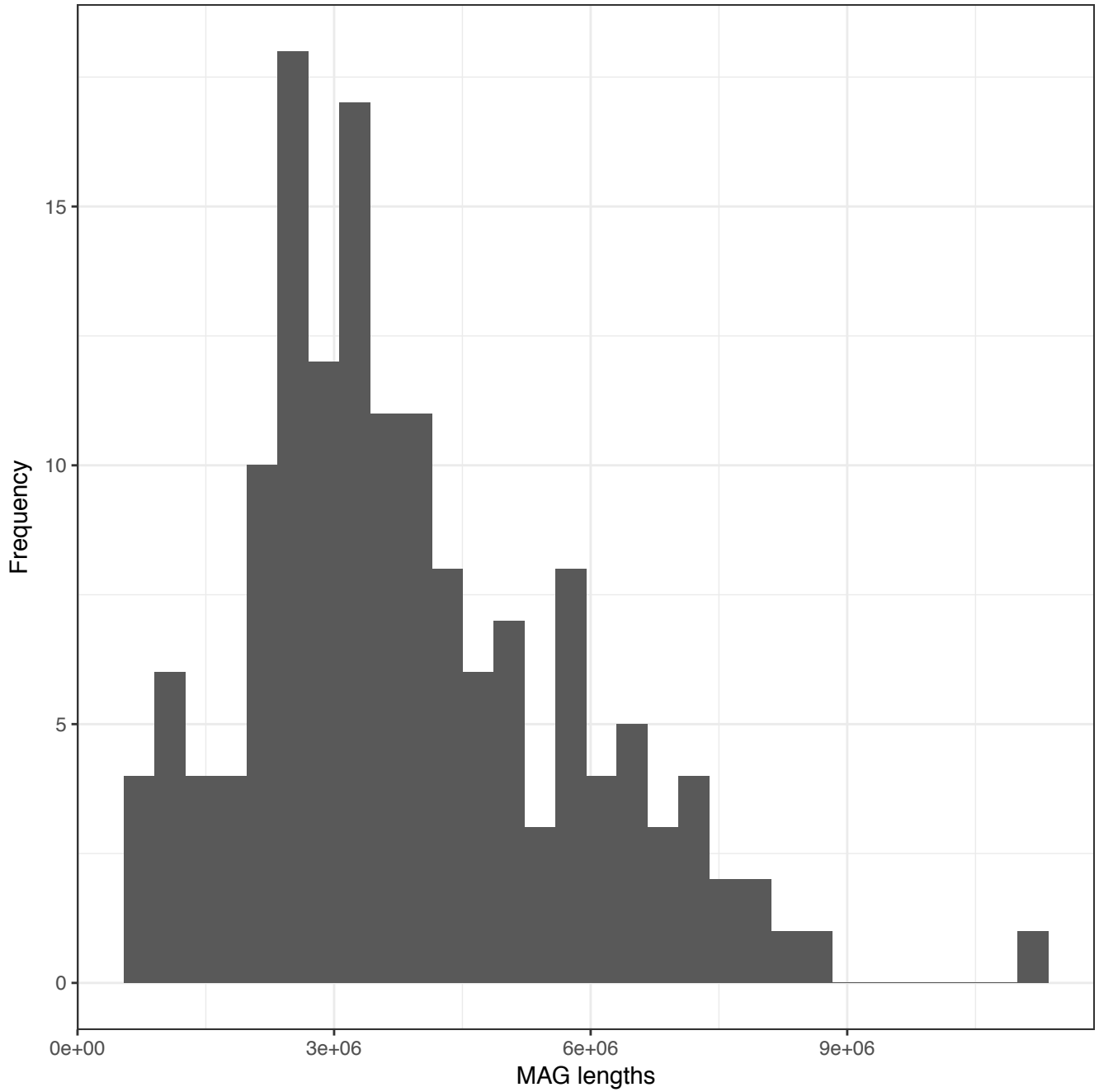


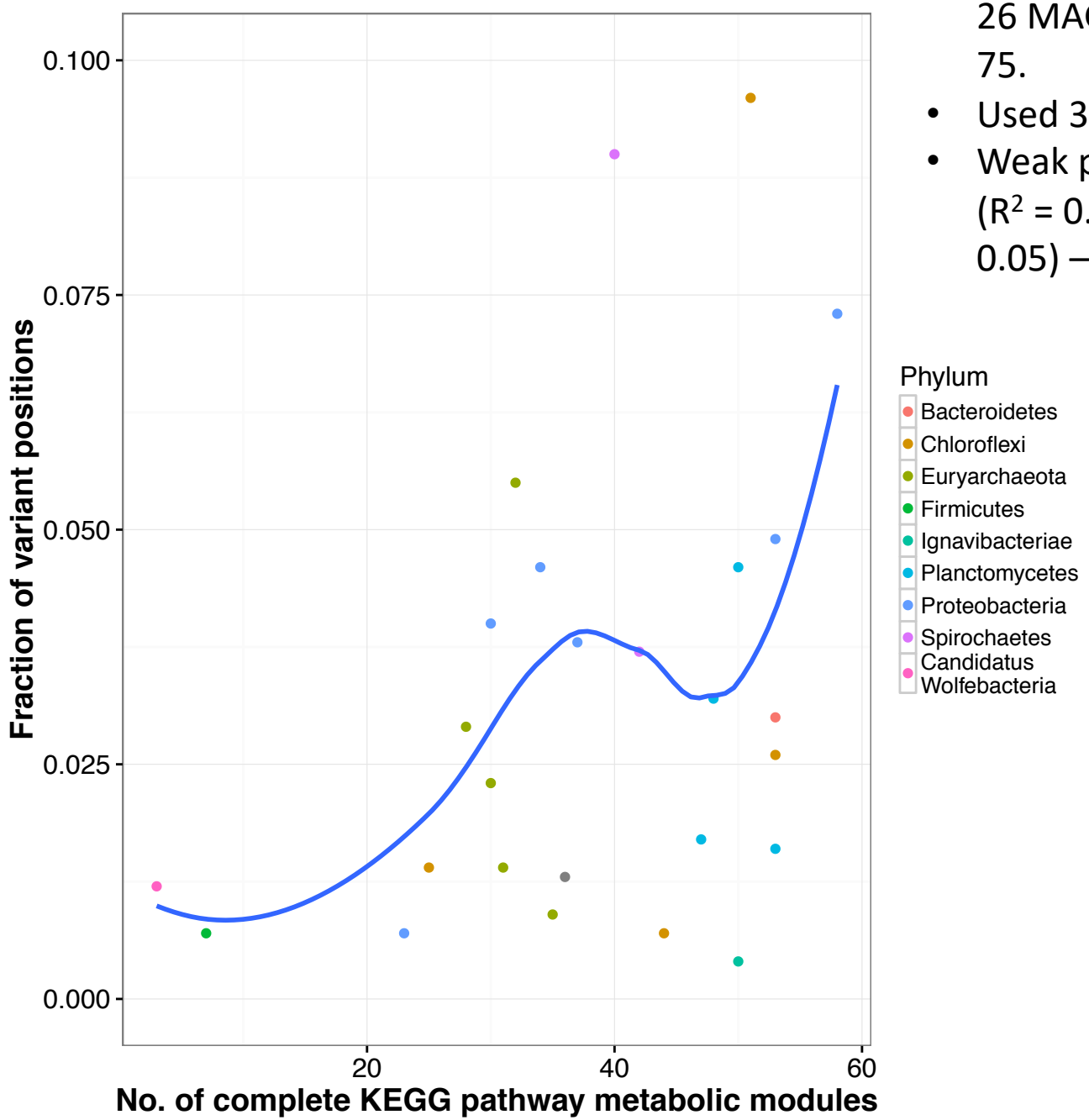
Expanded Granular Sludge-Bed Laboratory Bioreactors (EGSB)

- Seed from industrial EGSB bioreactor treating distillery waste
- 3 reactors run for approximately 3 months
- Sequenced 95 reactor samples approximately biweekly – 521,492,655 2X125 bp reads
- 355 CONCOCT clusters, 152/153 – 70% pure and complete



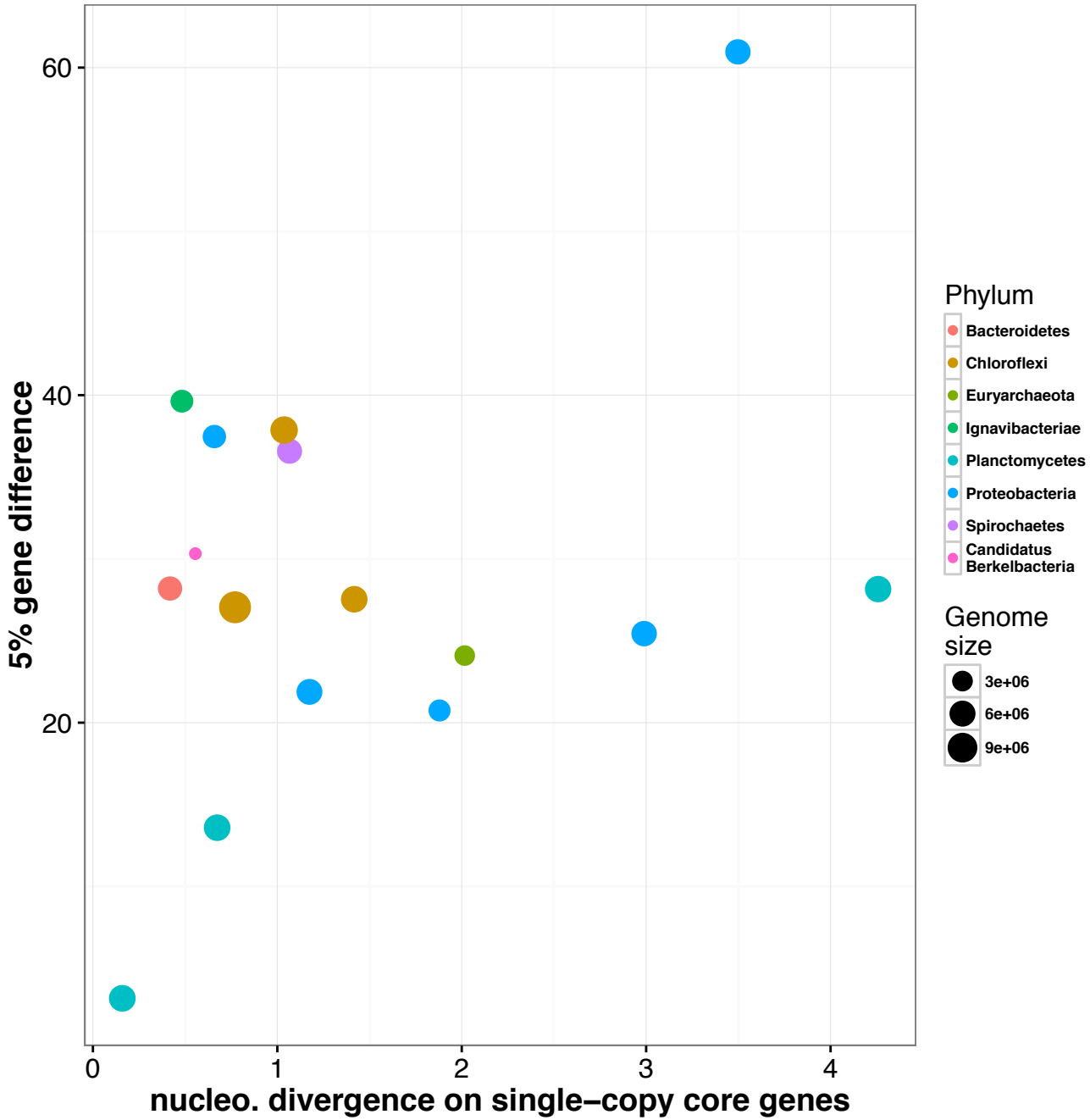



Frequency
15
10
5
0
0e+00
3e+06
6e+06
9e+06
MAG lengths



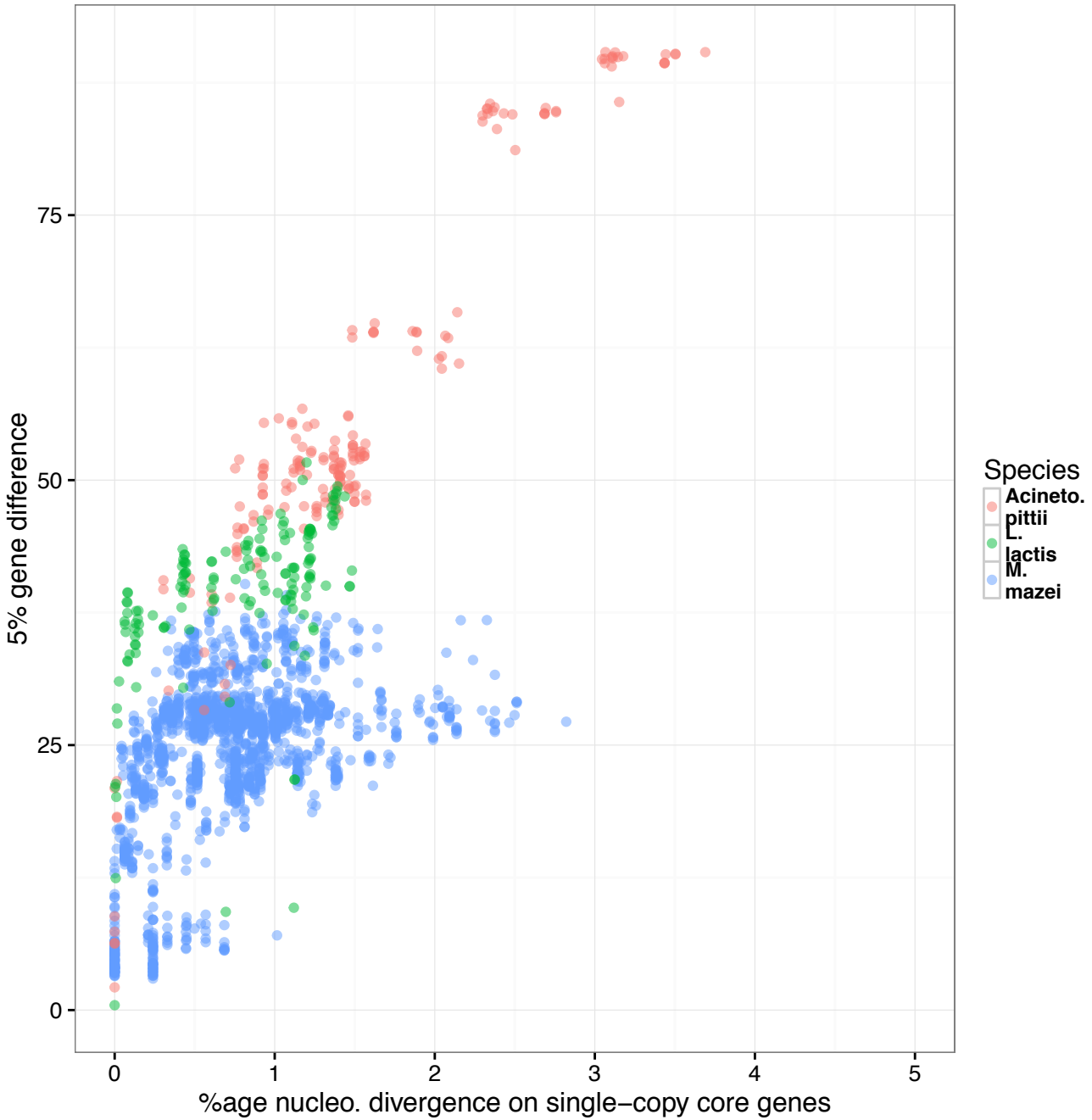
- Applied strain resolution to 26 MAGs with total cov. > 75.
- Used 36 SCGs
- Weak positive relationship (R^2 = 0.1132, p-value = 0.05) – c.f. Muller et al. 2014

AD strain analysis

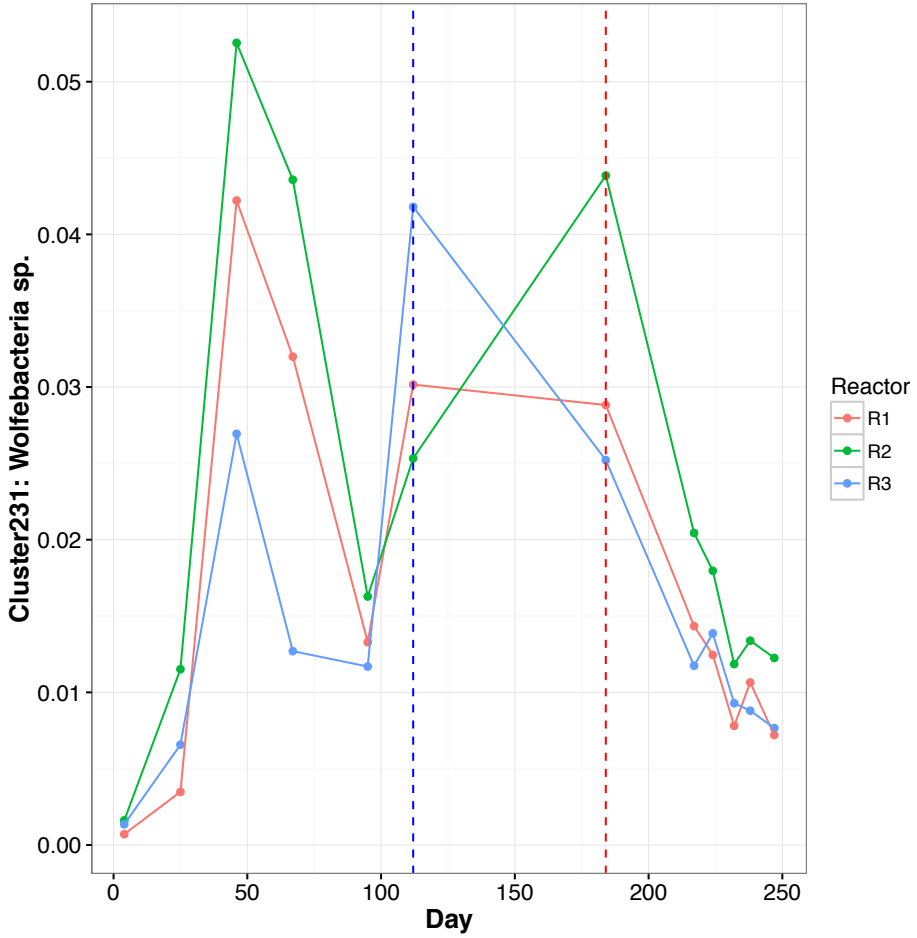


- 16 of 36 MAGs existed as two strains
- Inferred genome divergence between strains based on 5% gene clusters

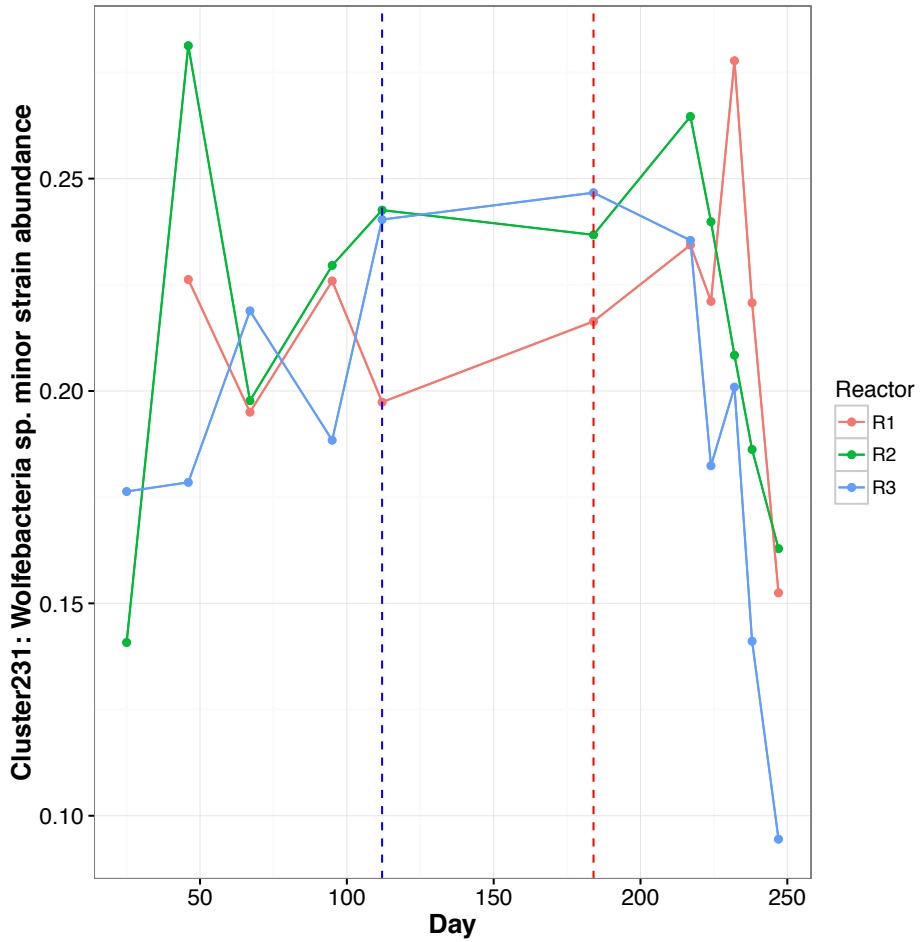
Comparison to real environmental organisms



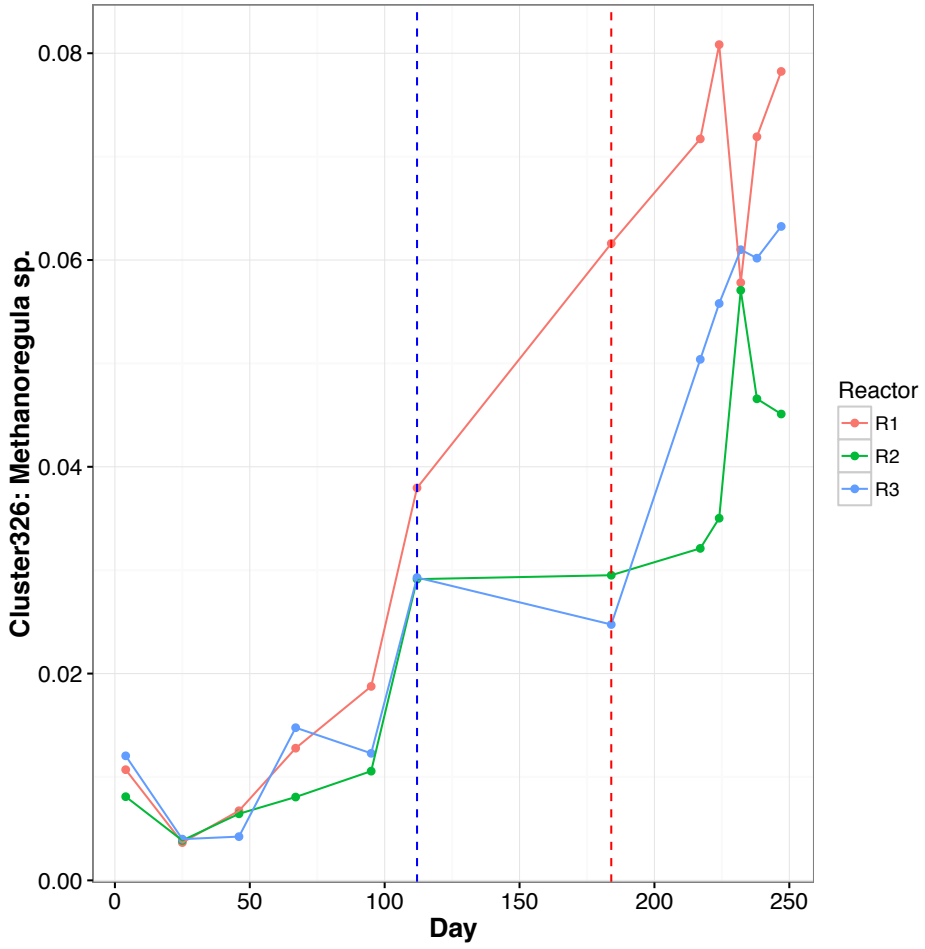
MAG abundance



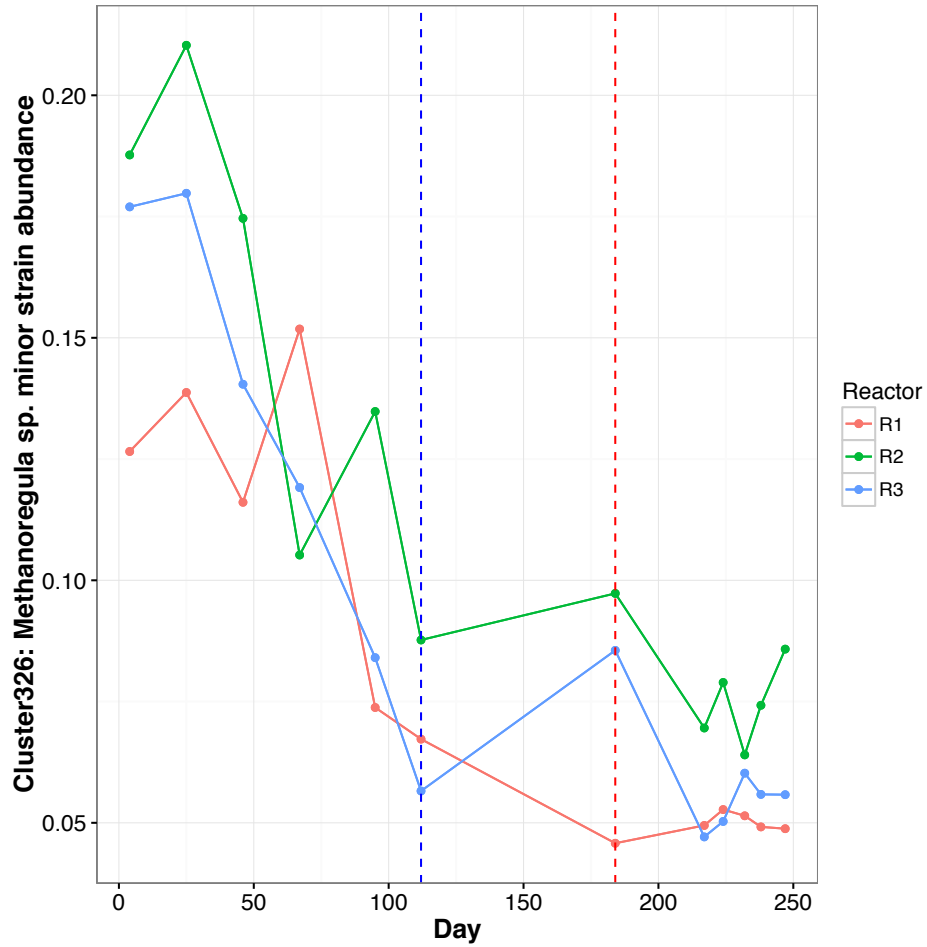
Strain abundance



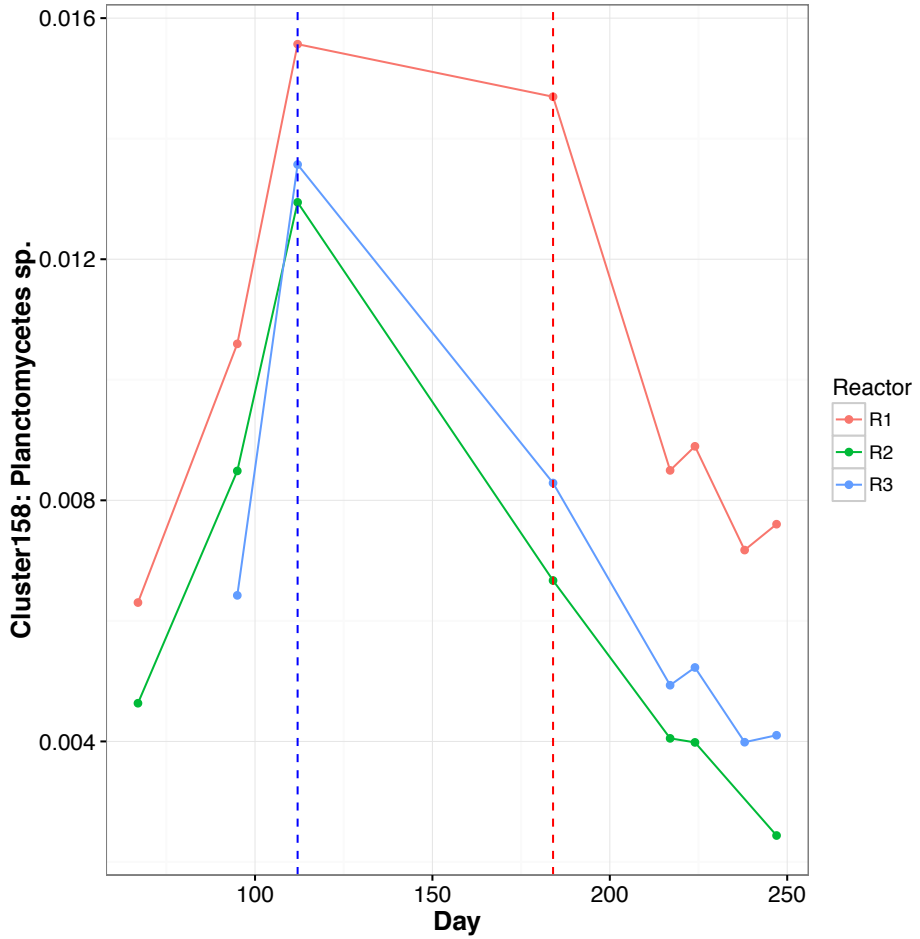
MAG abundance



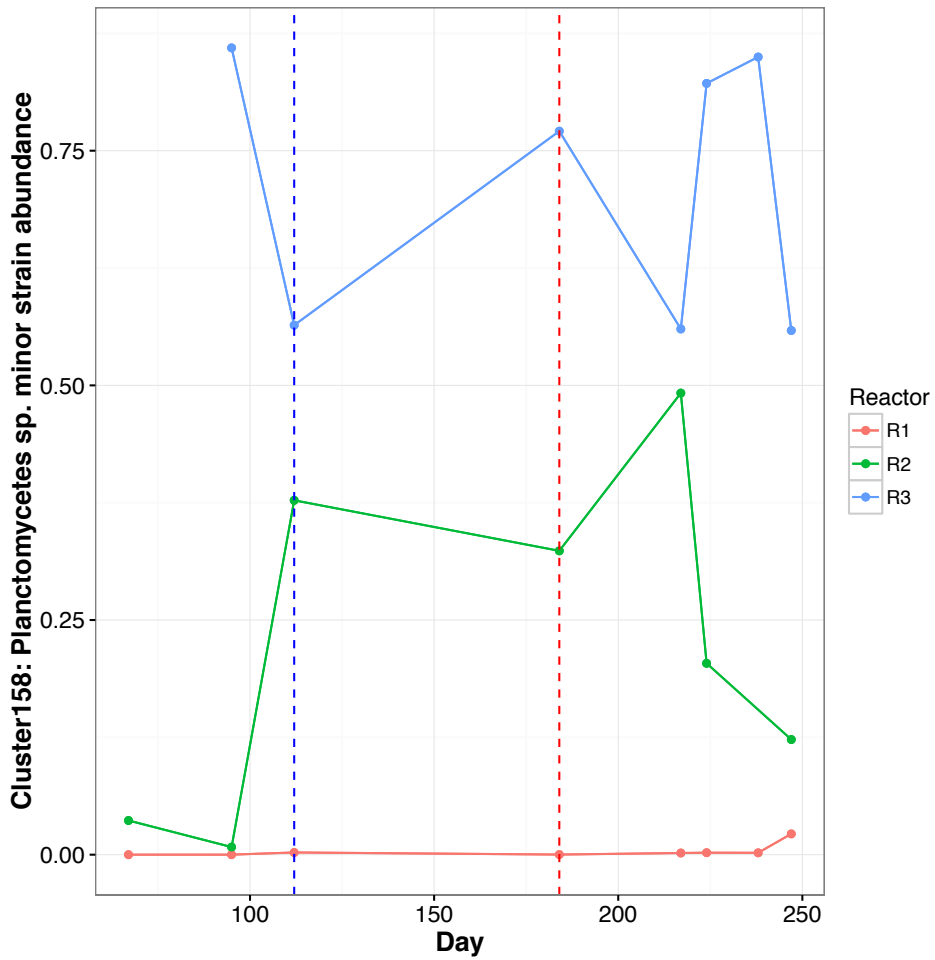
Strain abundance

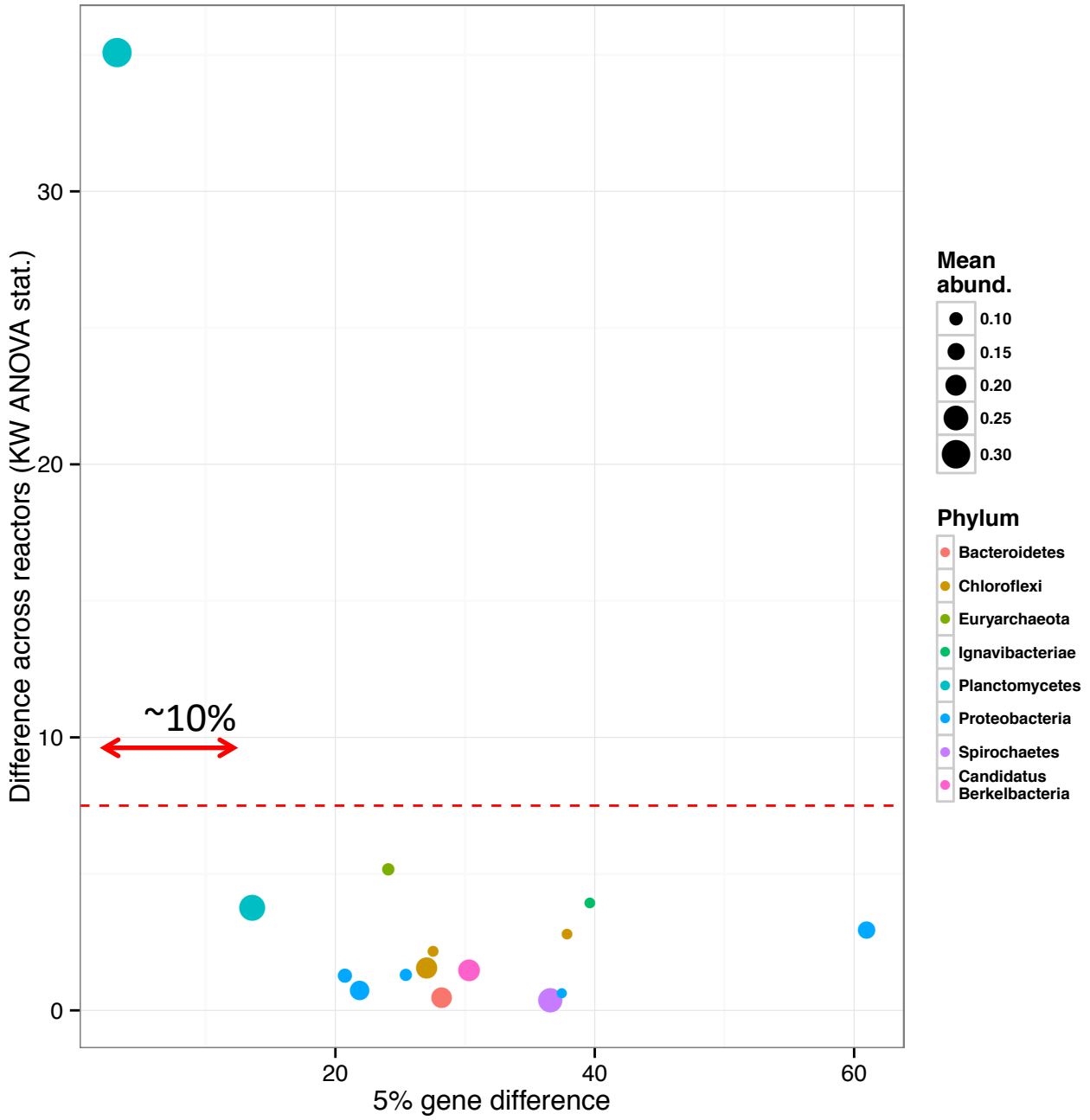


MAG abundance



Strain abundance

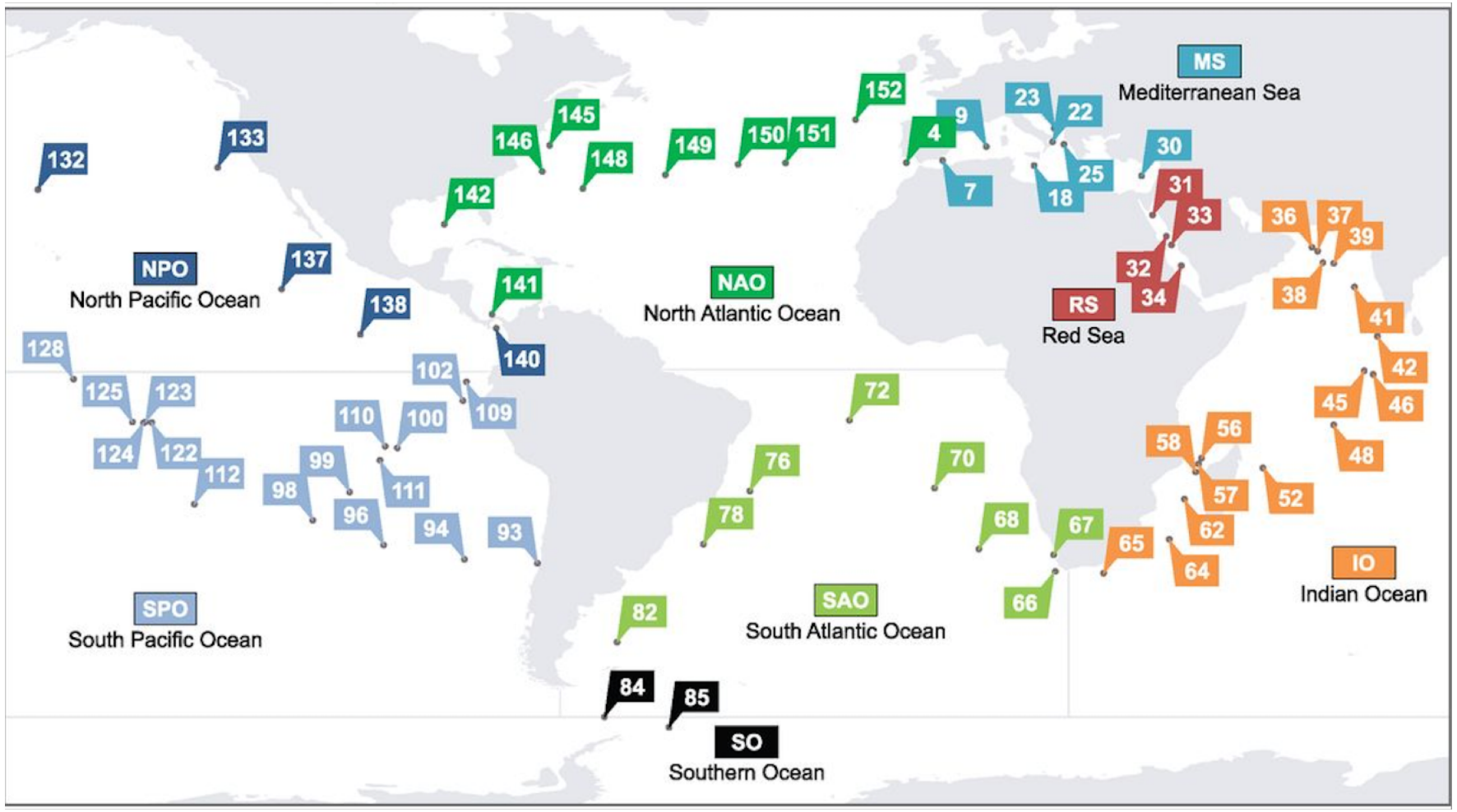




Hypotheses:

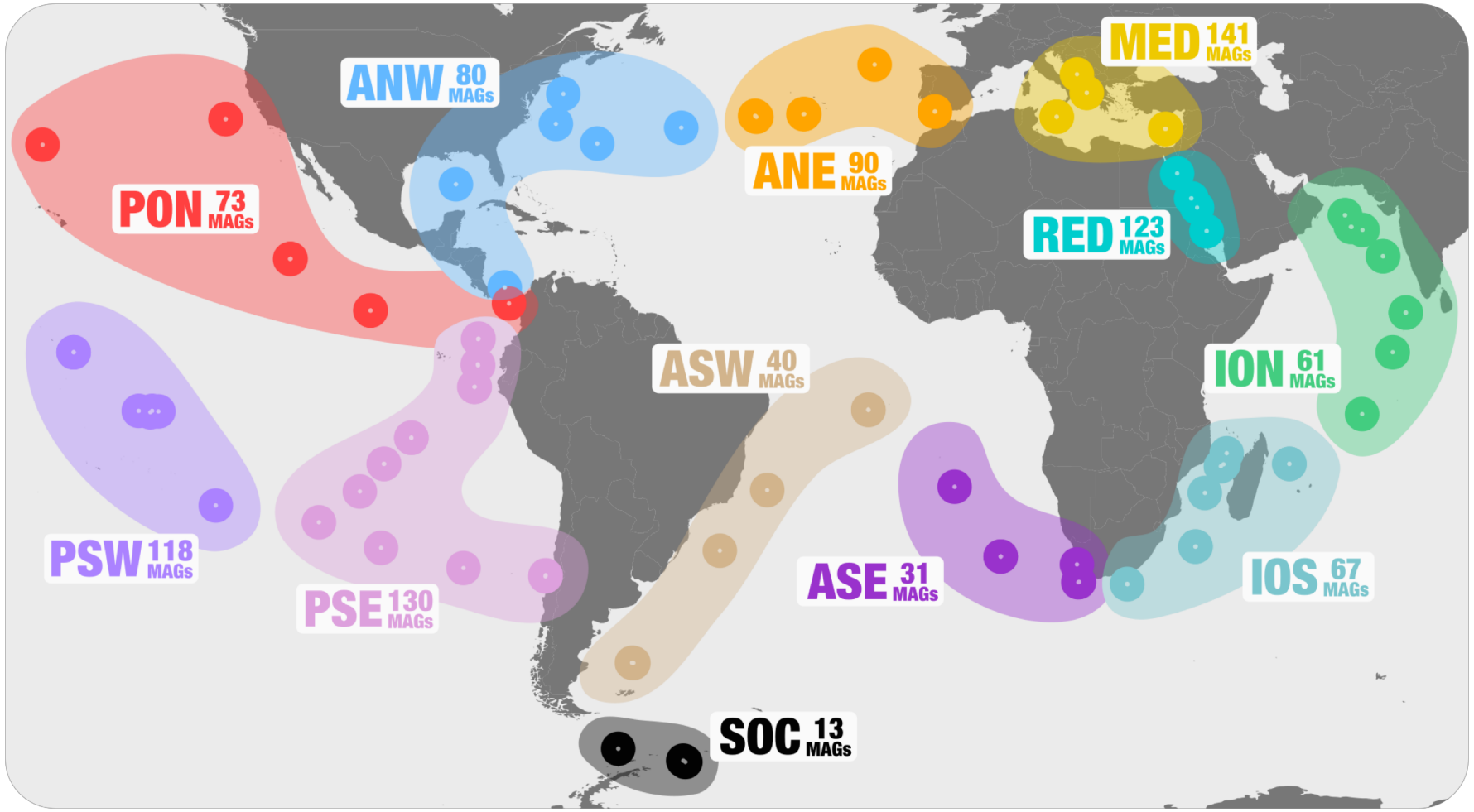
- Wide-spread niche partitioning within microbial species
- Ecological redundancy requires strains have similar genomes

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



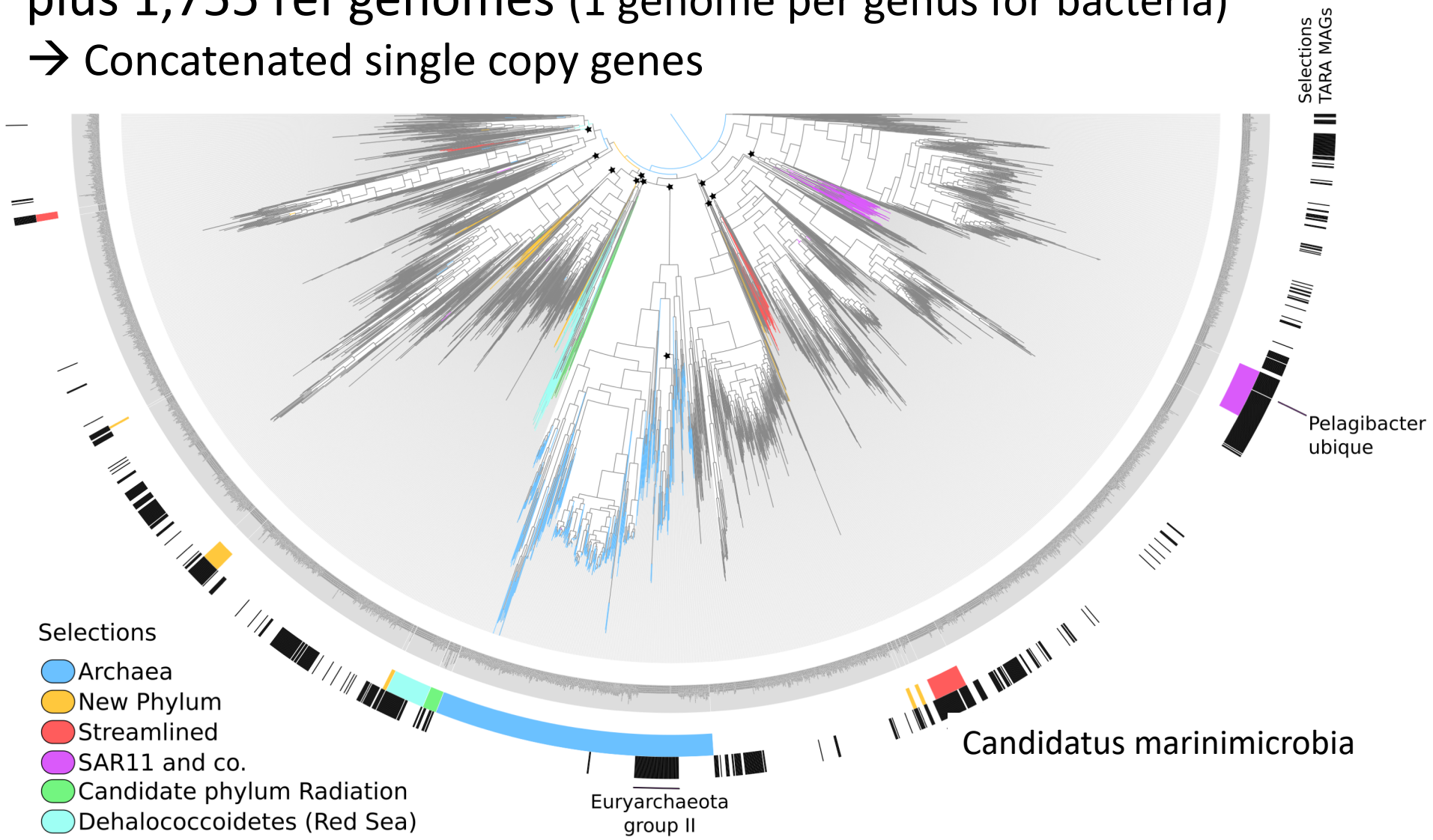
- 93 metagenomes from the planktonic size fraction for which we performed 12 metagenomic co-assemblies
- Generated 957 non-redundant MAGs encompassing the three domains of life

  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

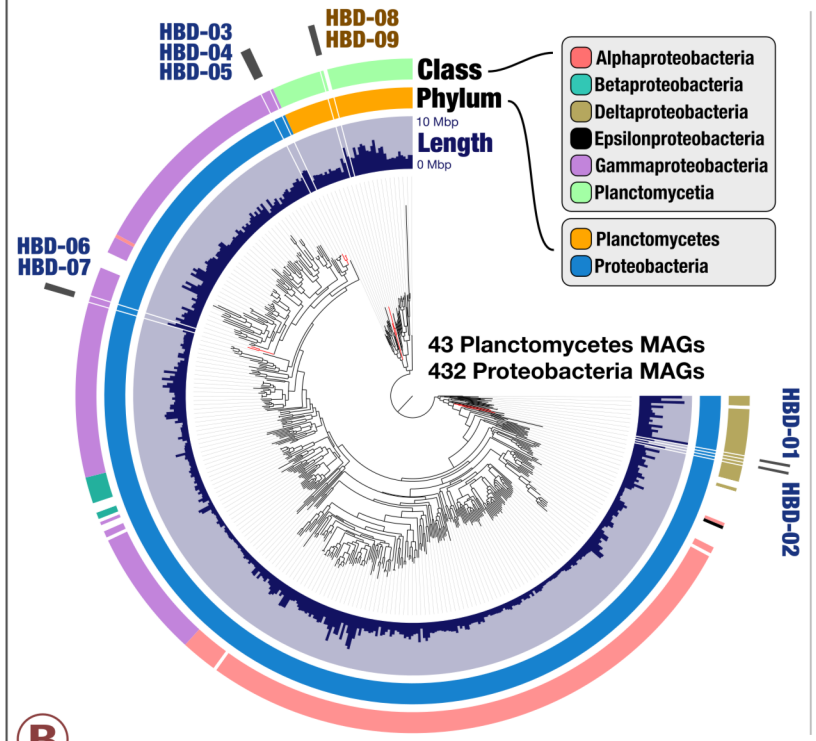


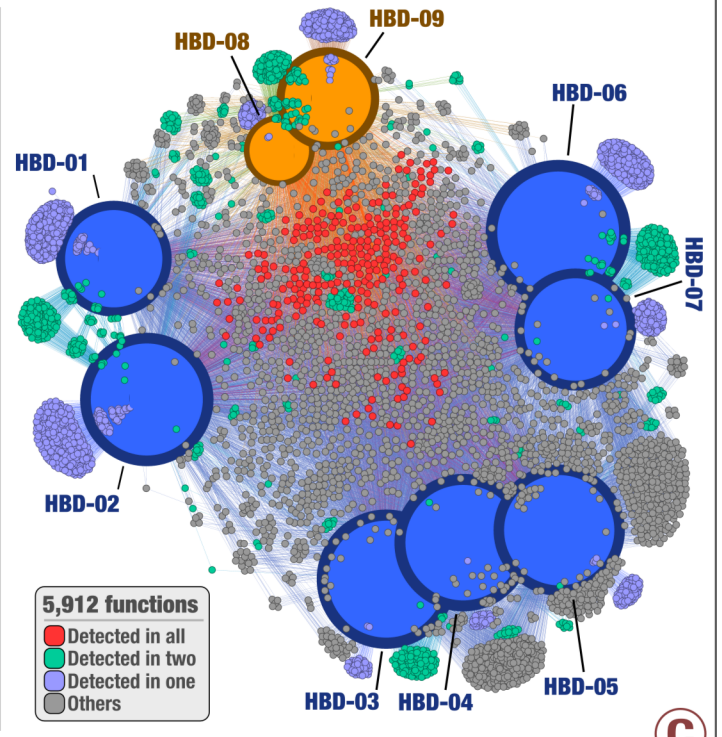
Oceanic heterotrophic bacterial diazotrophs (HBDs)

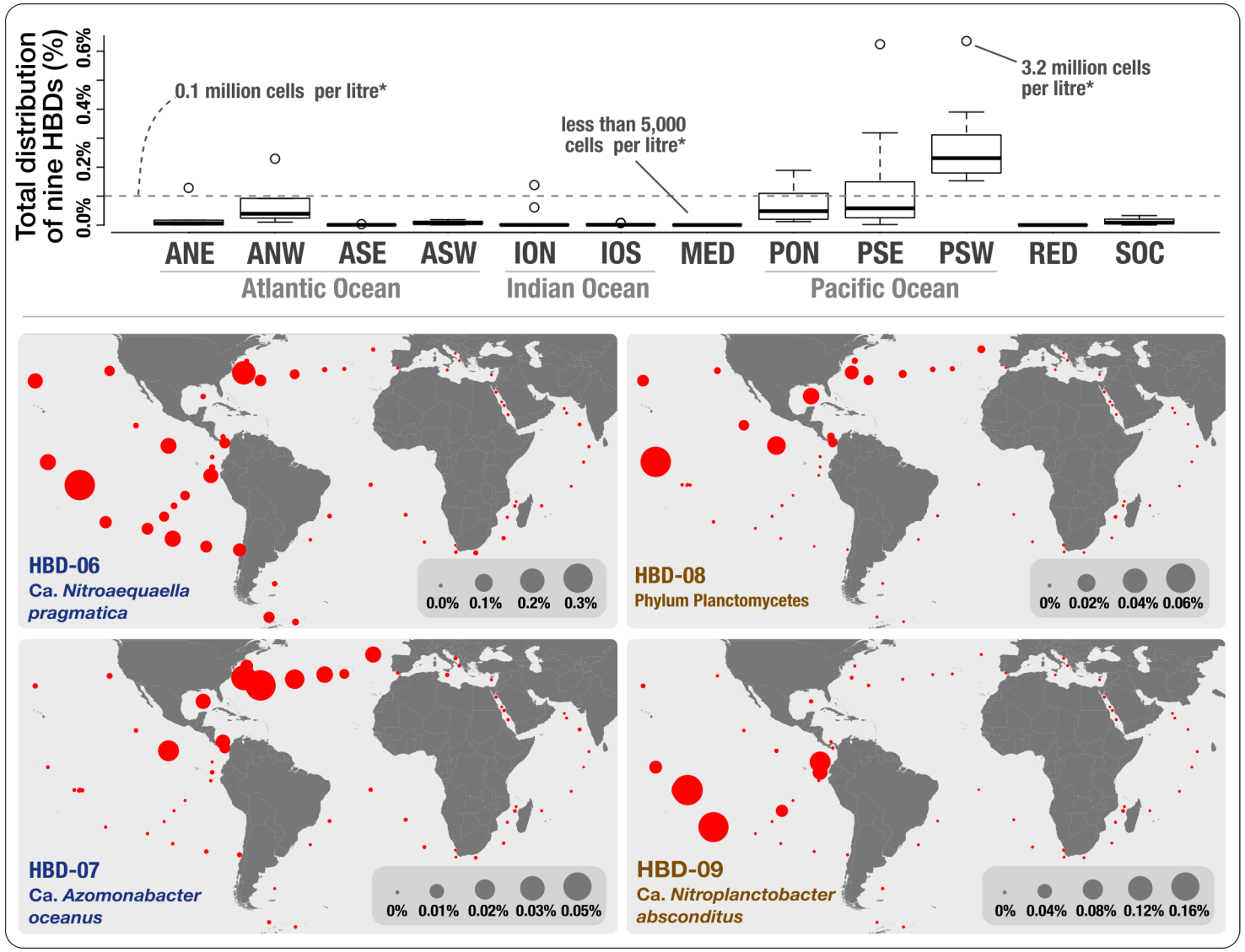
- Nitrogen fixation often limiting factor to growth in Ocean
- Cyanobacteria populations have long been thought to represent the main oceanic nitrogen fixers
- Nitrogenase reductase gene surveys revealed the existence of heterotrophic bacterial diazotrophs (HBDs)
- Search for MAGs containing the catalytic (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

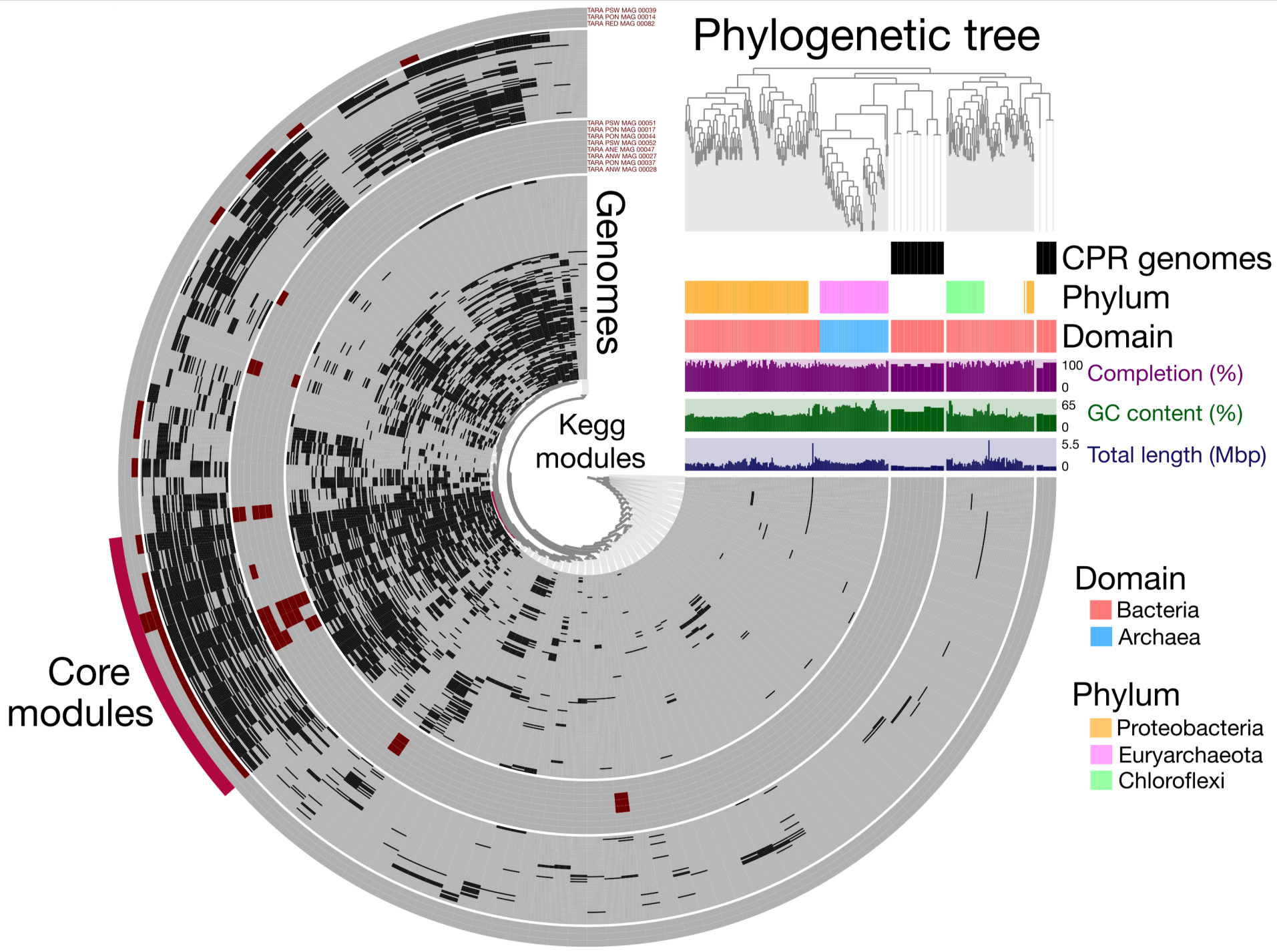
A

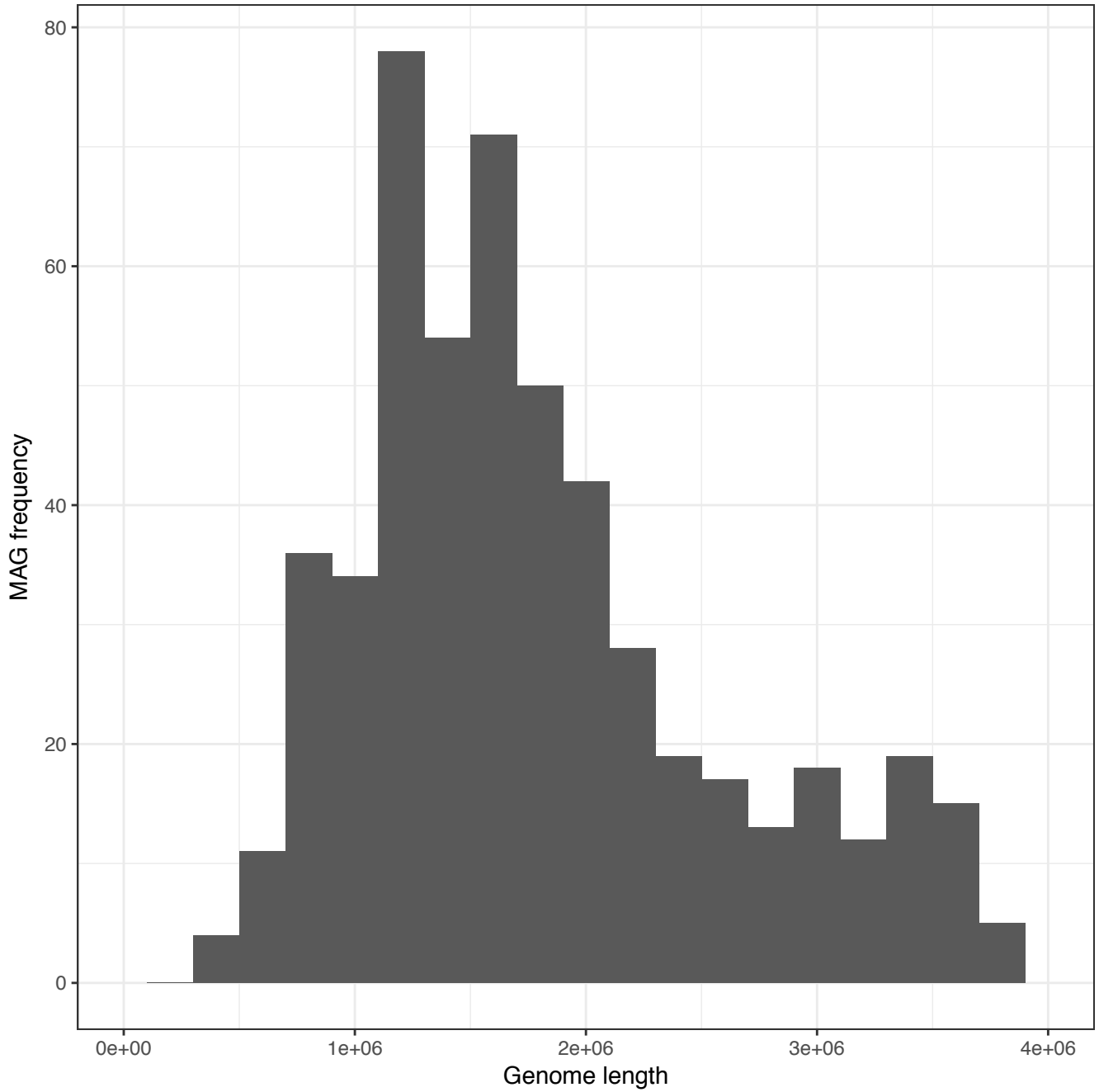
	Region of Recovery	Length (Mbp)	N50	Num Contigs	Num Genes	GC Content	Percent Compl.	Percent Redund.	Phylum	Class	Order	Family	Genus
HBD-01	PSW	3.67	48,153	118	3,592	52.6%	97.7%	4.4%	Proteobacteria	Deltaproteo.	Desulfovibrionales	Desulfovibrionaceae	Desulfovibrio
HBD-02	PSW	6.00	20,964	405	5,385	53.1%	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	-



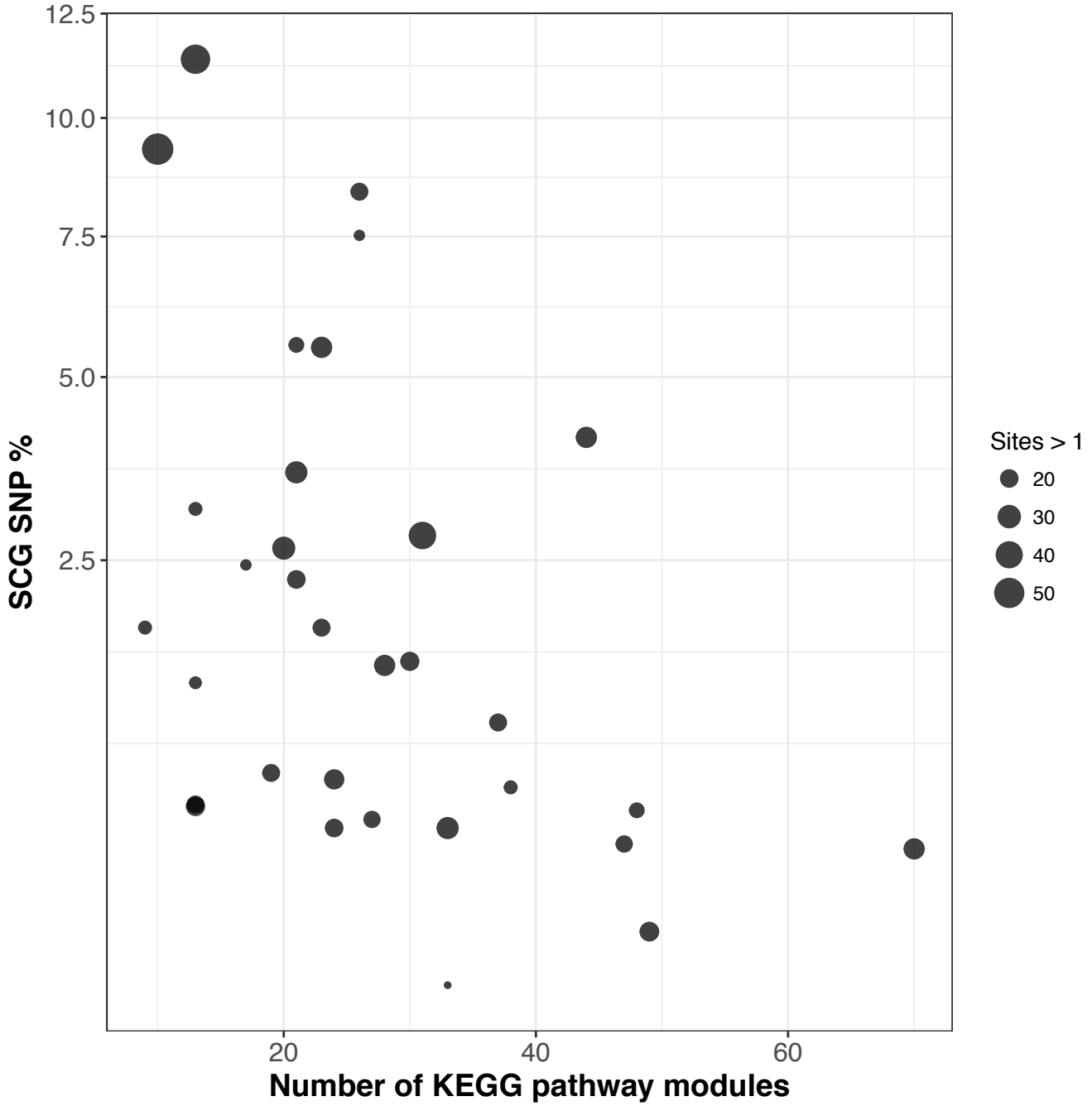



Total distribution of nine HBDs (%)
0.0%
0.2%
0.4%
0.6%
0.1 million cells per litre*
less than 5,000 cells per litre*
3.2 million cells per litre*
ANE
ANW
ASE
ASW
ION
IOS
MED
PON
PSE
PSW
RED
SOC
Atlantic Ocean
Indian Ocean
Pacific Ocean
HBD-06
Ca. Nitroaequaella pragmatica
0.0% 0.1% 0.2% 0.3%
HBD-08
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%


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

Tara MAG variants



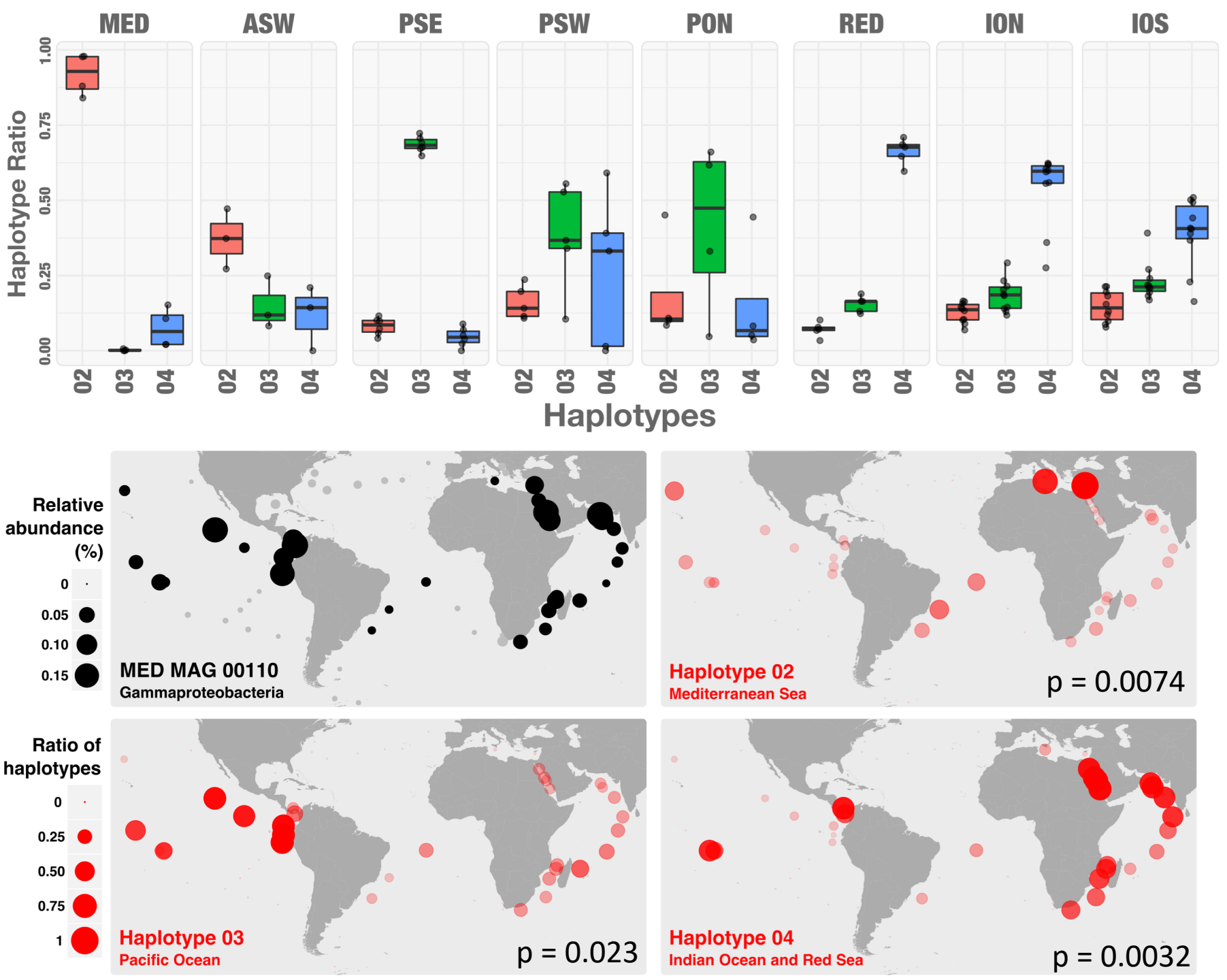
Consider 32 MAGs with cov. > 100.

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

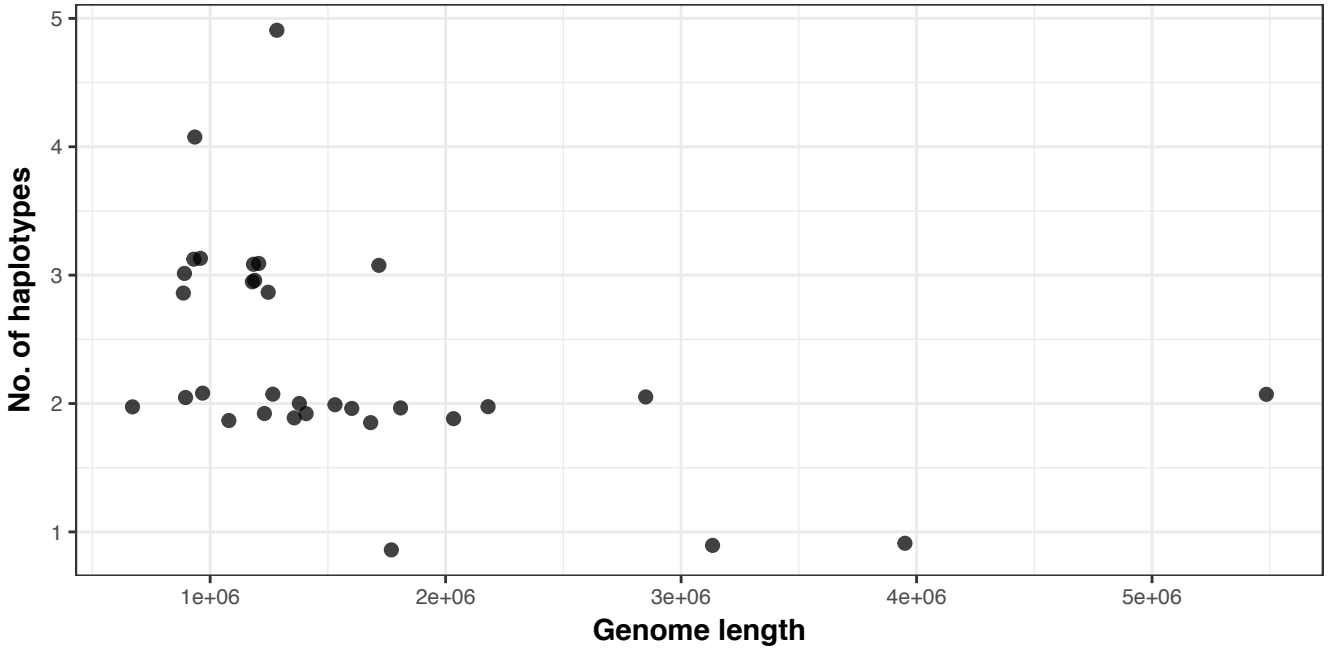
Stronger negative relationship with number of KEGG Pathway modules (Spearman’s p = 0.0045)

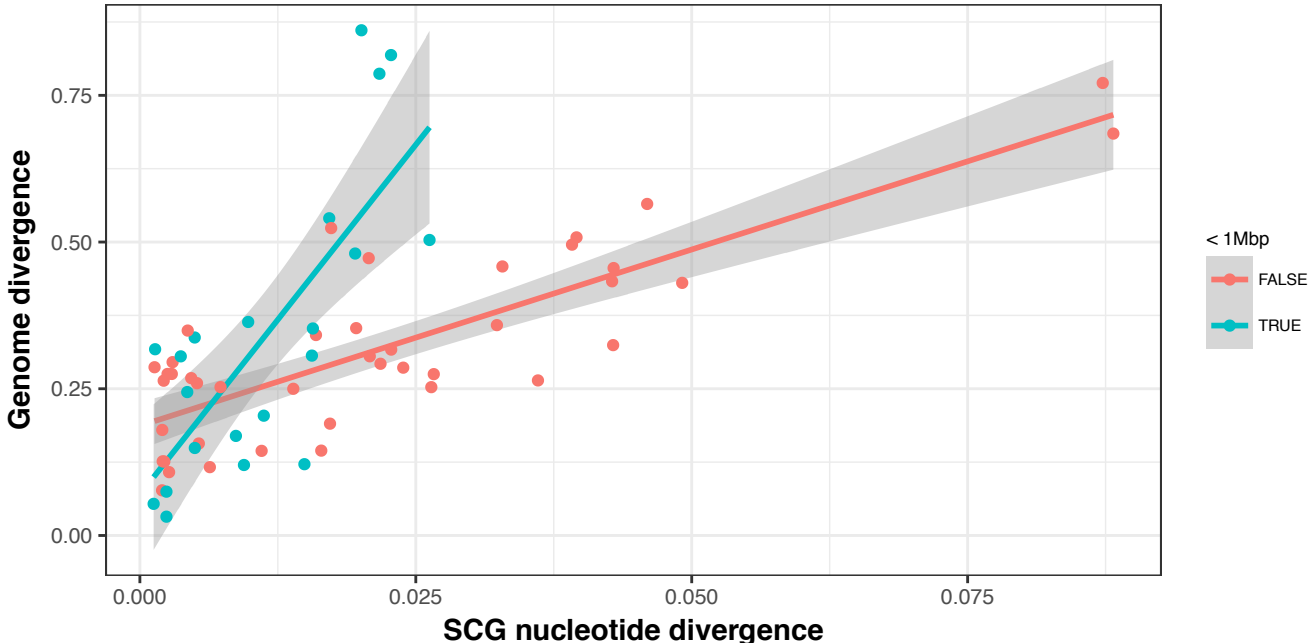
TARA DESMAN analysis

- 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 MED 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)

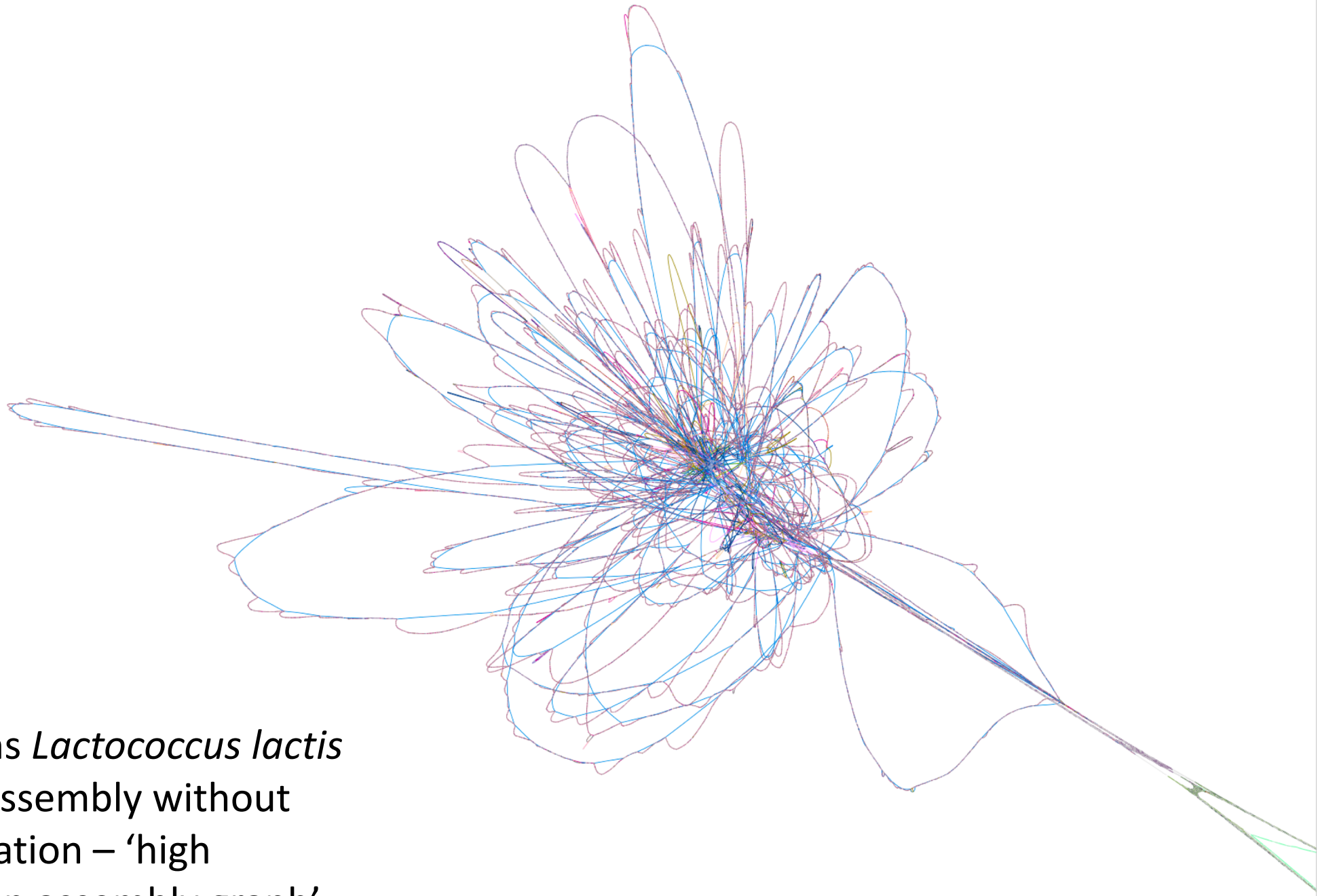
Summary

- DESMAN resolves subpopulations reliably from time series data e.g. FMT and AD reactors
- In cross-sectional studies e.g. Tara resolves ecologically relevant patterns
- There will always be limitations on the accuracy of long range variant linkage that can be resolved from metagenomes with short reads
- Long read sequencing may transform situation
- However, we can probably do better using graphs…

Part II: Metagenomics strain resolution on assembly graphs

Introduction

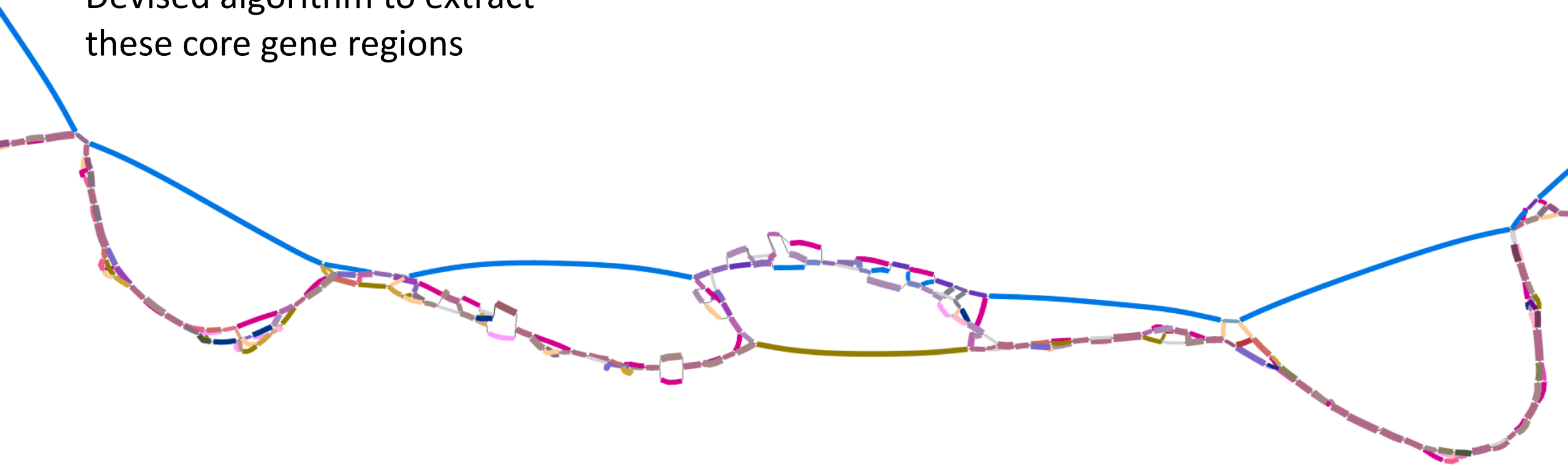
- Local *de novo* resolution of strains is possible using linkage of variants in short reads (e.g. PredictHaplo Prabhakaran et al. 2014, Hansel and Gretel Nicholls et al. 2018)
- Methods for bacterial genome scale de novo strain resolution utilize co-occurrence of variants across multiple samples (Lineages O'Brien et al. Genetics 2014, Constrains Luo et al. 2015, DESMAN Quince et al. Genome Biology 2017)
- All methods map reads onto a linear sequence either a reference or consensus contig
- Resolving strains directly on assembly graphs but utilizing co-occurrence would avoid mapping, allow more complex variant structure and incorporate read information (Brown et al. 2018 spacegraphcats)

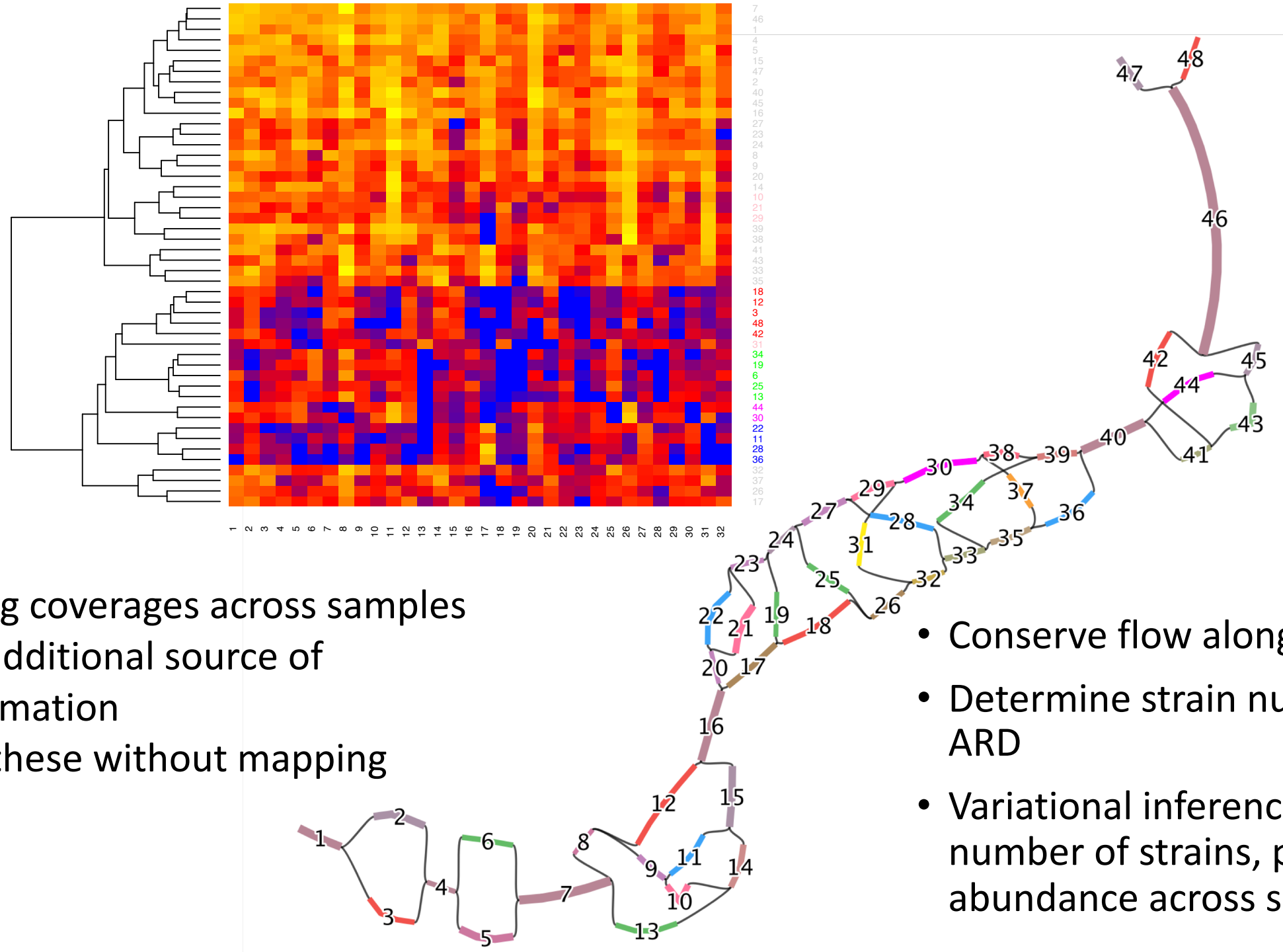


Six strains *Lactococcus lactis* spades assembly without simplification – ‘high resolution assembly graph’

Amplify region around
COG0060
Has simple ‘linear’
structure…

Devised algorithm to extract
these core gene regions

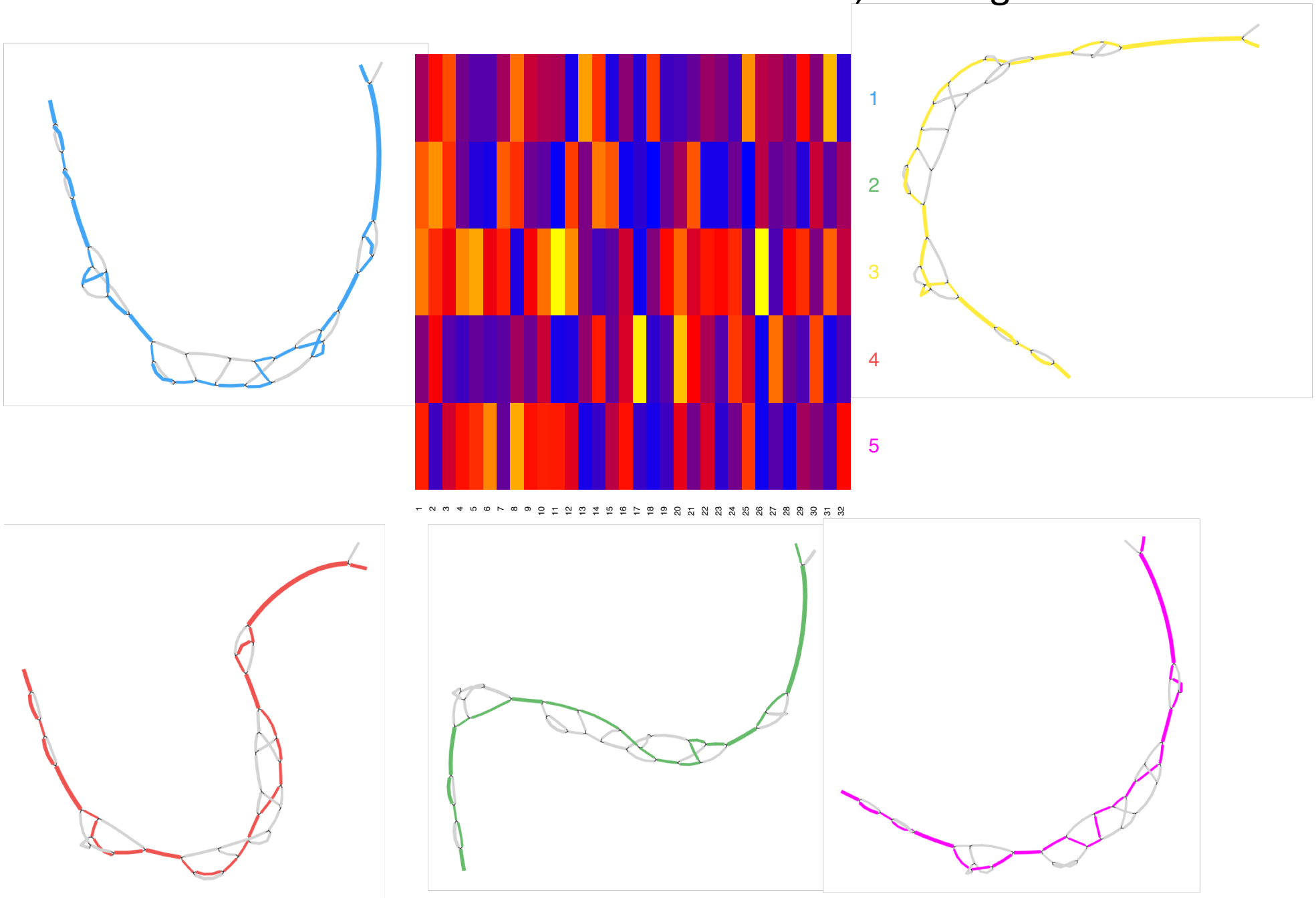




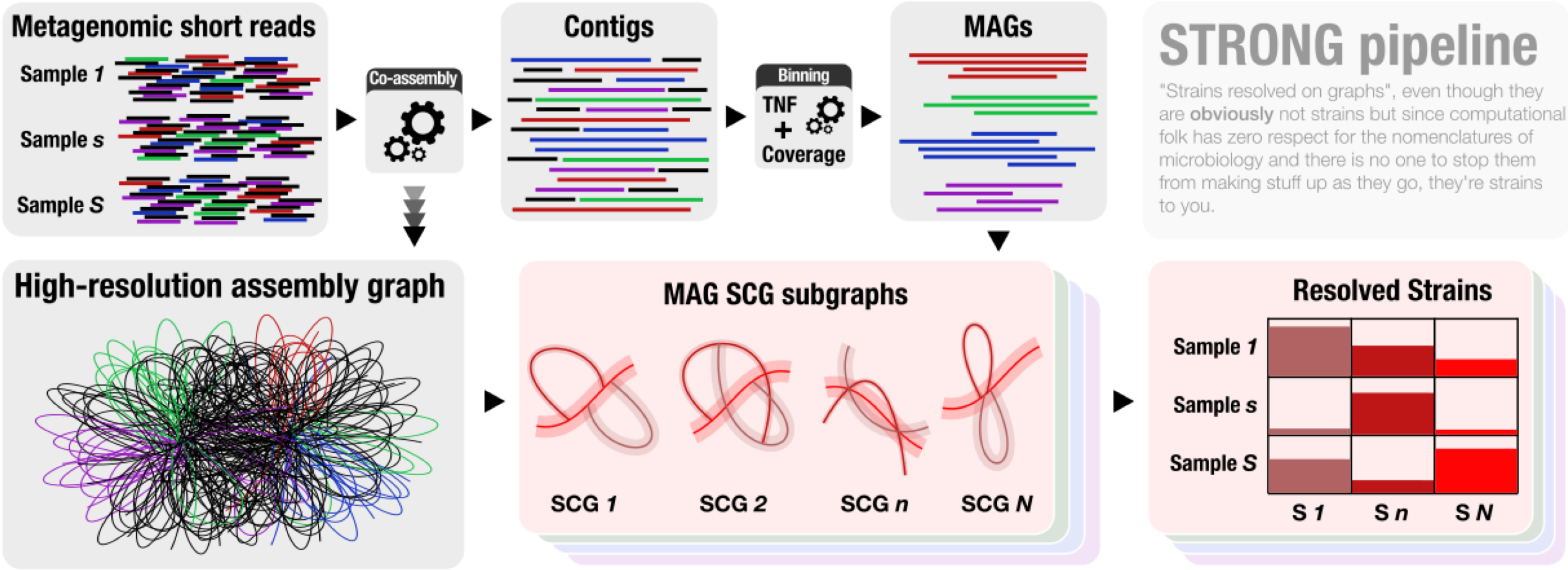
- Unitig coverages across samples are additional source of information
- Get these without mapping

- Conserve flow along graph
- Determine strain number with ARD
- Variational inference to infer number of strains, paths, and abundance across samples

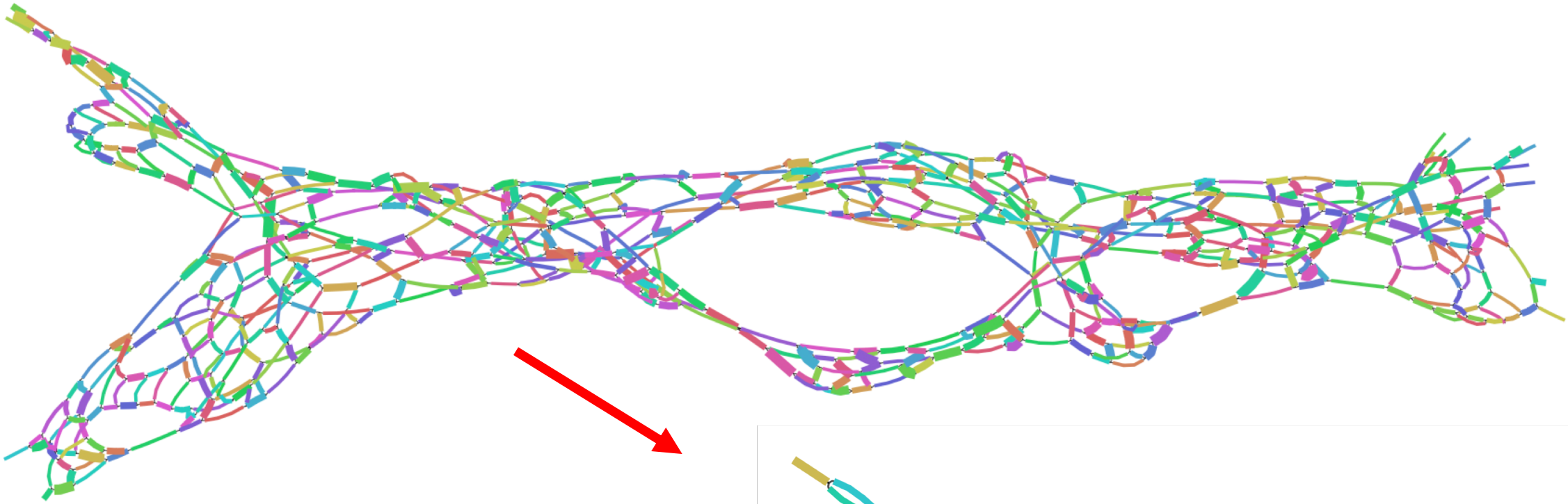
Five inferred strains match five references with zero errors, coverage match R^2 = 95%



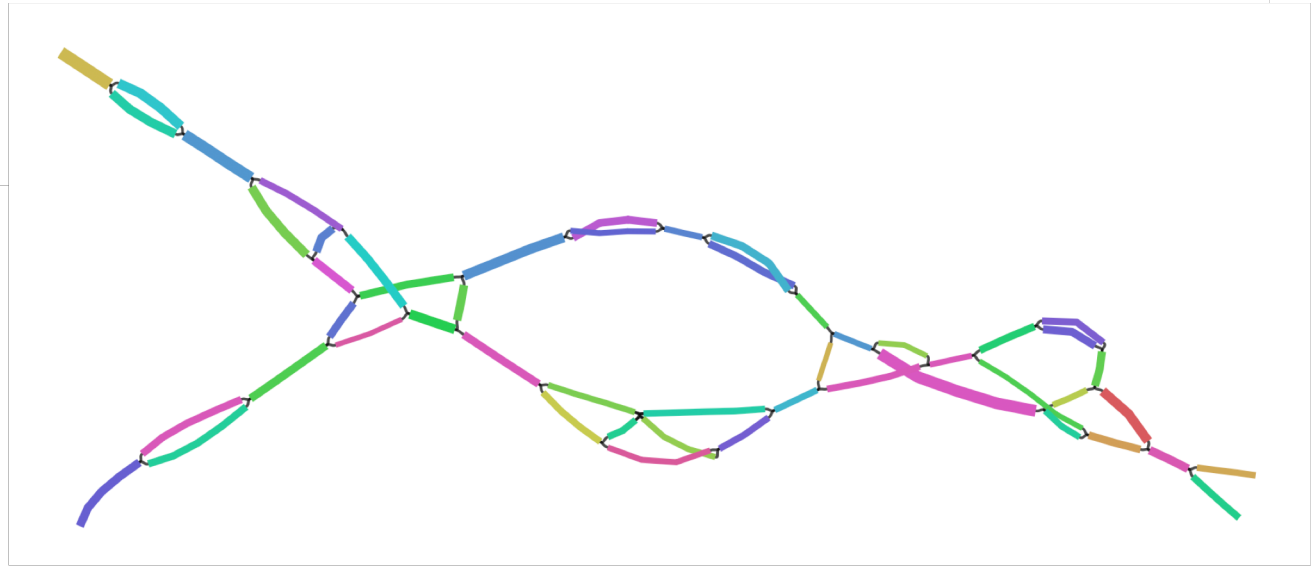
STRONG https://github.com/chrisquince/STRONG
(appearing soon)



Second stage of graph simplification

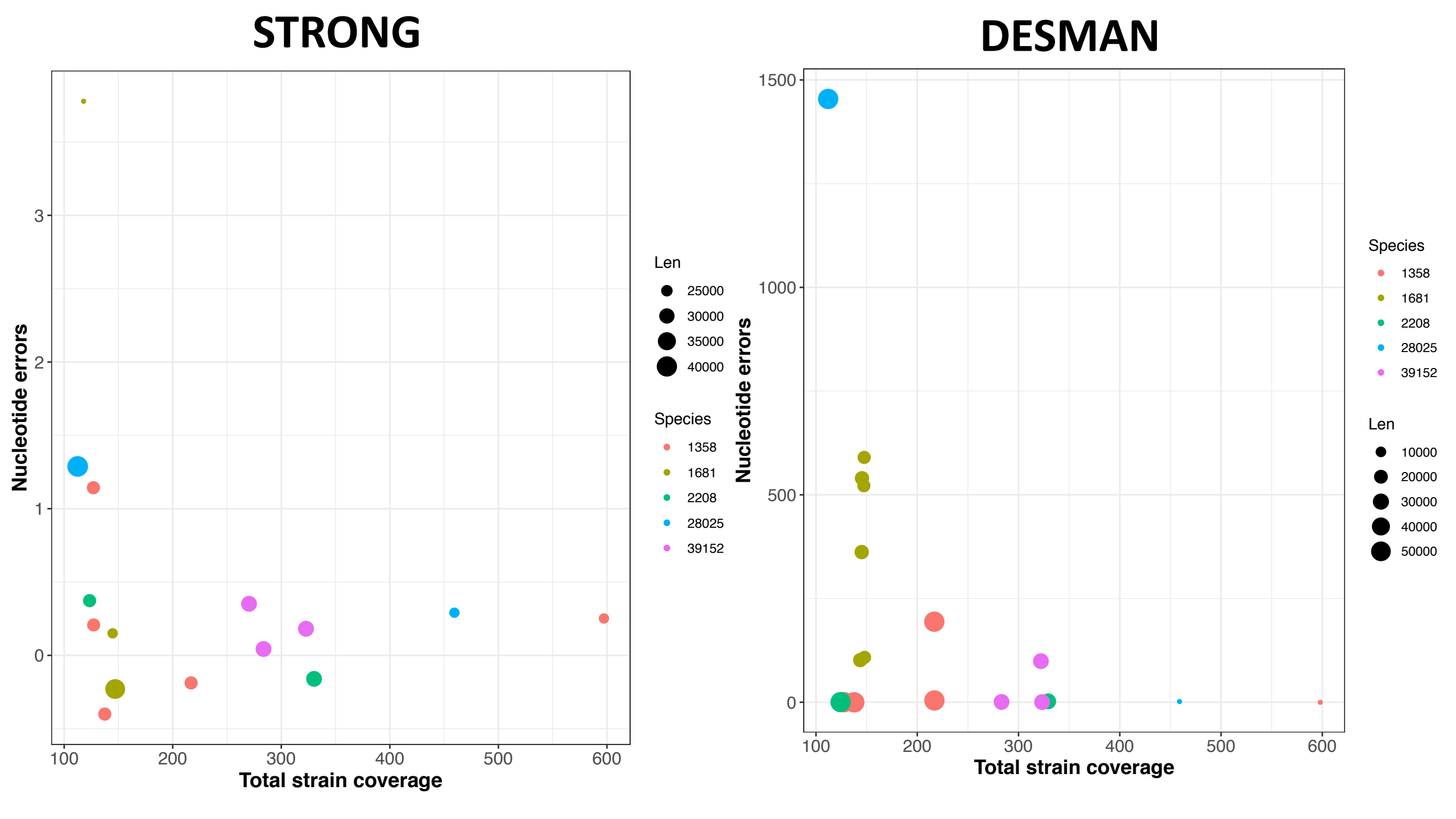


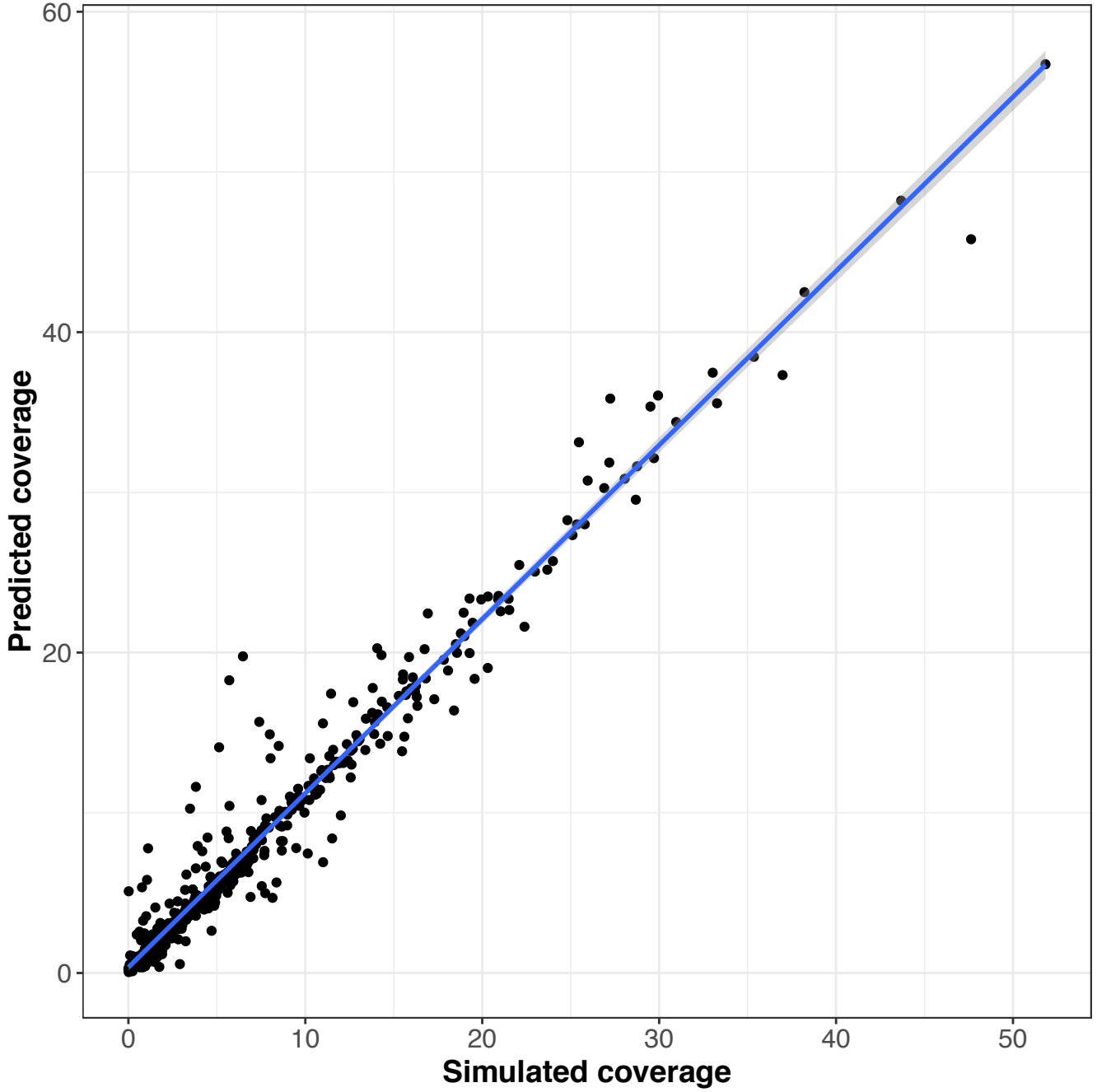
Use coverage estimates from bins to further simplify graphs



Simple synthetic community

- 5 species, 20 strains (6,4,4,4,2) generated 2.5 million 2X150bp reads per sample
- Obtained 4 ‘good’ clusters from CONCOCT
- Applying STRONG resolved 16 (6,4,4,2) strains in those clusters with almost no errors across a concatenation of 21,786 – 40,928 bp core COGS regions
- How does it compare to DESMAN?


STRONG
DESMAN
Nucleotide errors
Total strain coverage
0
1
2
3
100
200
300
400
500
600
Len
25000
30000
35000
40000
Species
1358
1681
2208
28025
39152
0
500
1000
1500
Len
10000
20000
30000
40000
50000

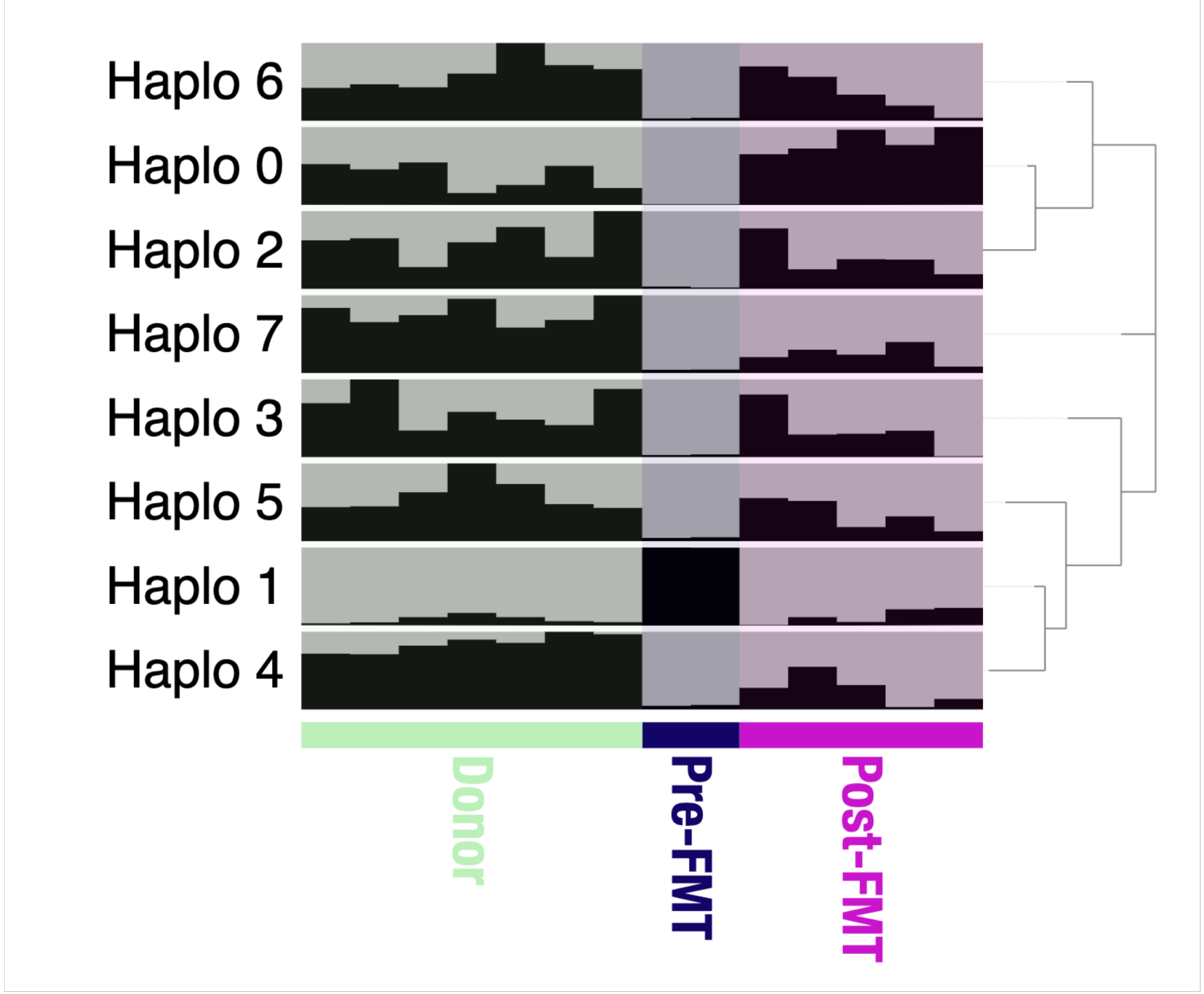


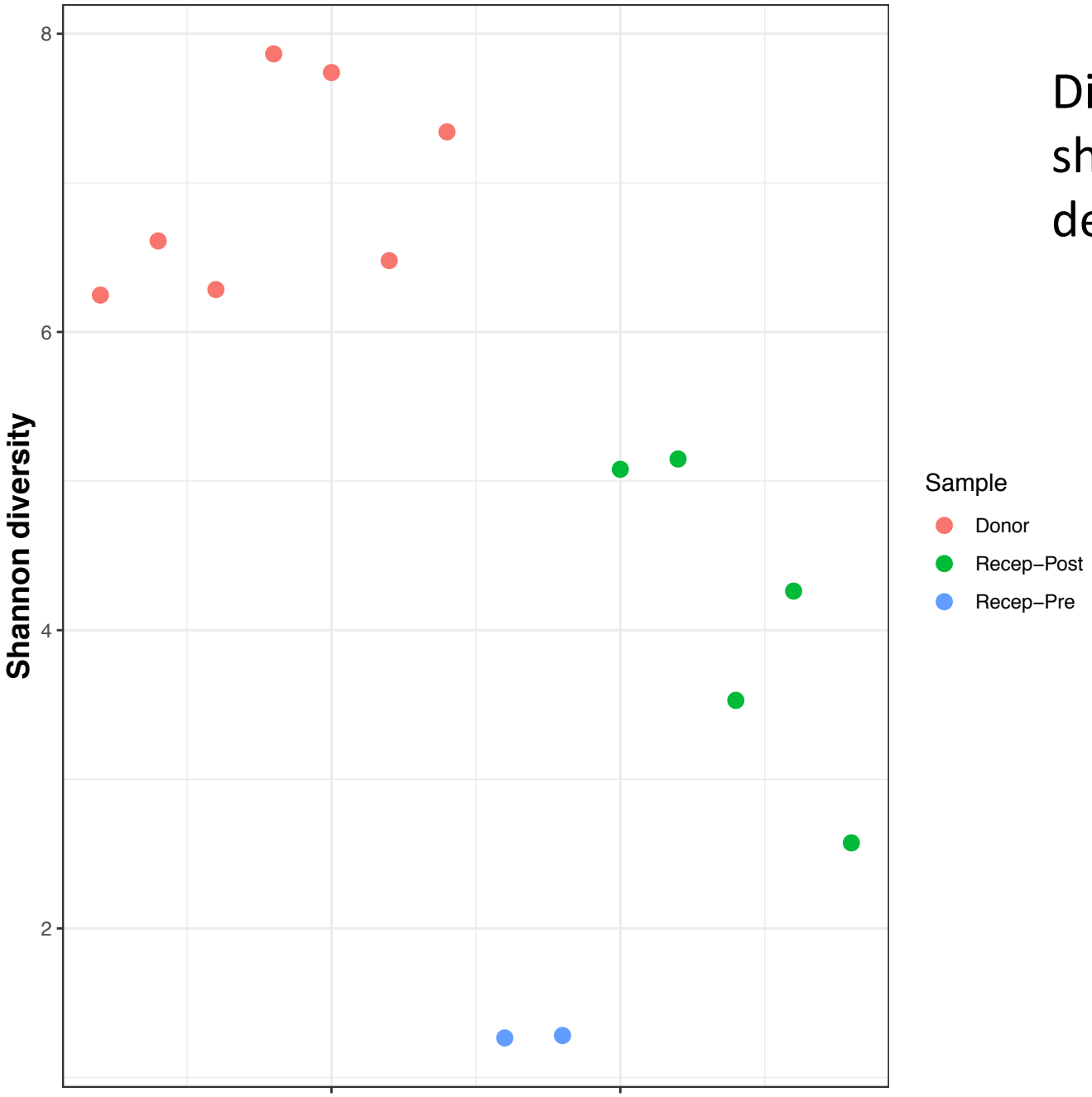
- Good correlation between true and inferred coverages ~ 96%
- Consistent bias though ~ 10%
- With 16 samples only one species was binned but we could construct 3 strains in that with error rate < 0.001%

FMT CP/R03 strain analysis

(Andrea Watson and A. Murat Eren)

- Healthy donor, 24 samples 2 years, only analysed 7 prior to FMT
- Recipients with *C. difficile* infection received FMT through colonoscopy
- Two samples from pre-FMT, and 5 samples taken post-FMT over the course of a year
- STRONG spades co-assembly gave 990 Mbp, max contig 674,730bp with N50: 6,759 bp
- 81 bins resolved accurately
- Applied strain resolution to Bin_275 assigned to Allistipes finegoldii

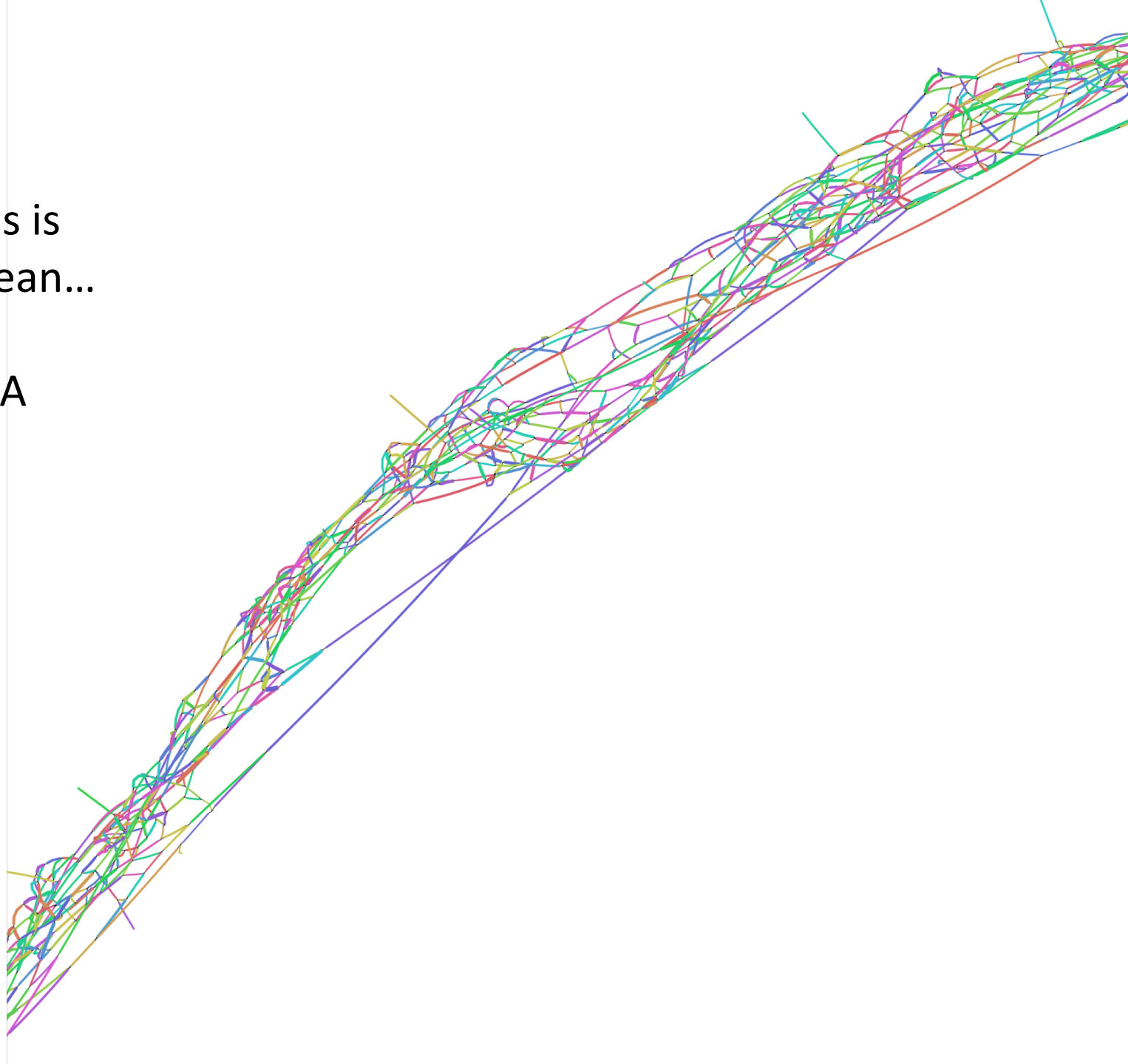

Haplo 6
Haplo 0
Haplo 2
Haplo 7
Haplo 3
Haplo 5
Haplo 1
Haplo 4
Donor
Pre-FMT
Post-FMT



Diversity in the single MAG shows increase post-FMT but declines over time

Graph structure of gut organisms is quite simple what about the Ocean…

Streamlined Proteobacteria TARA MED MAG 00110



Summary

- Can resolve strain haplotypes on short read assembly graphs
- Link these across multiple core genes
- Seems to reveal novel biology now applying it to larger data sets and validating with long reads
- Next step the accessory genome
- Some organisms will still likely be intractable

Acknowledgements

- Medical Research Council funding and CLIMB, BBSRC/Unilever
- Tom Delmont and A. Murat Eren (U. of Chicago)
- Sergey Nurk (St Petersburg), Aaron Darling (UTS) and Sebastian Raguideau (Warwick)