Resolving population diversity from metagenomes

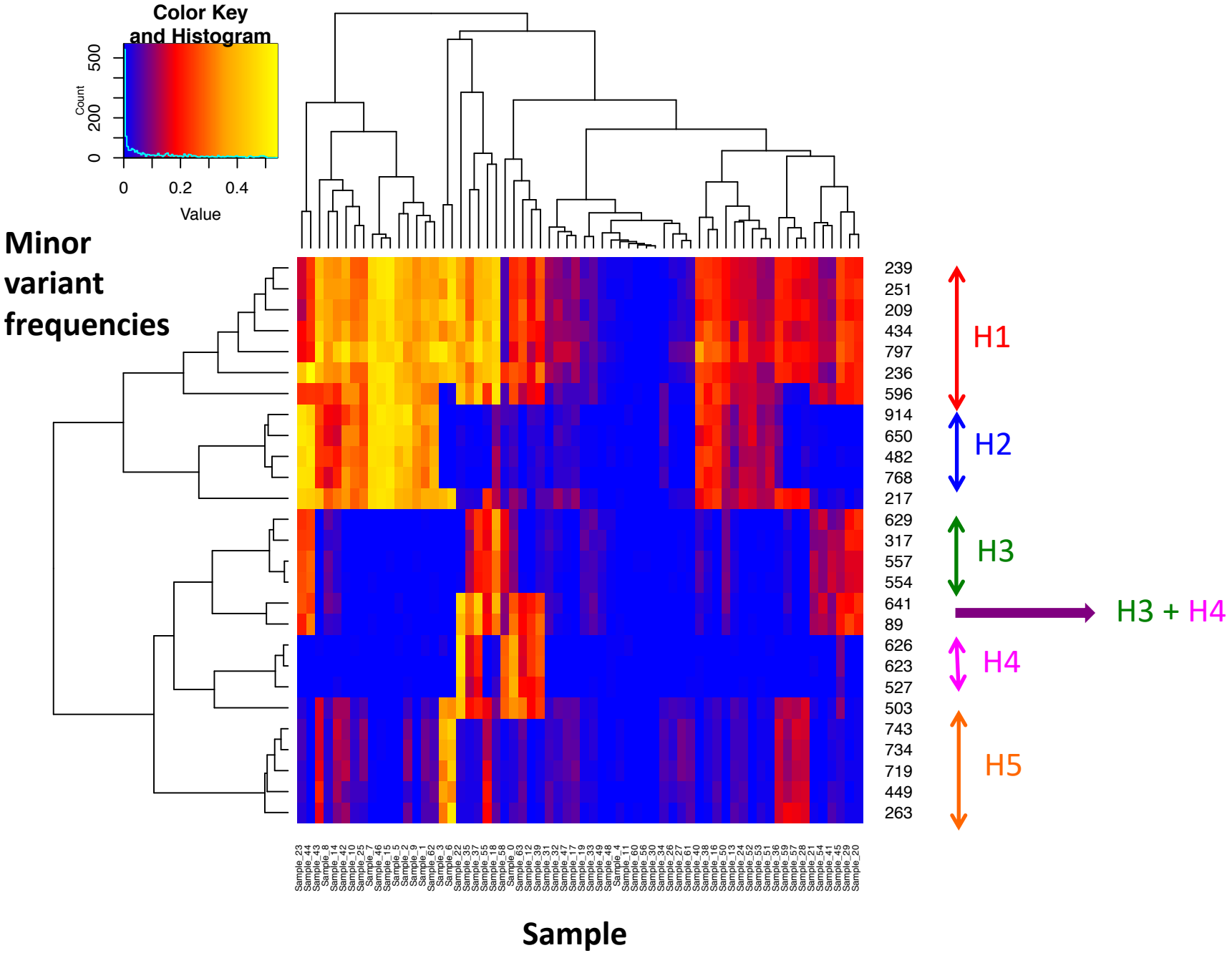
Dr Christopher Quince

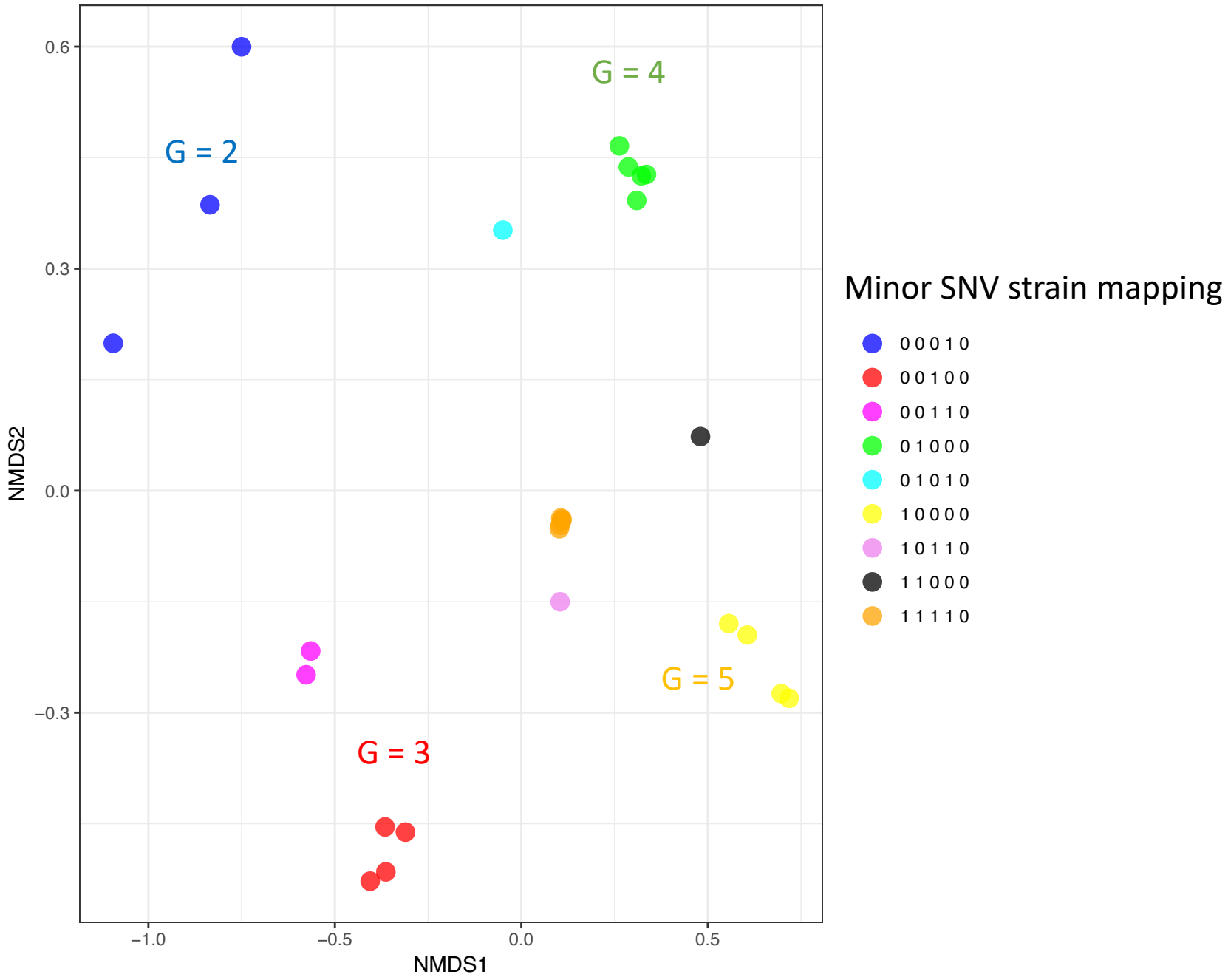
University of Warwick

Resolving intra-MAG population diversity

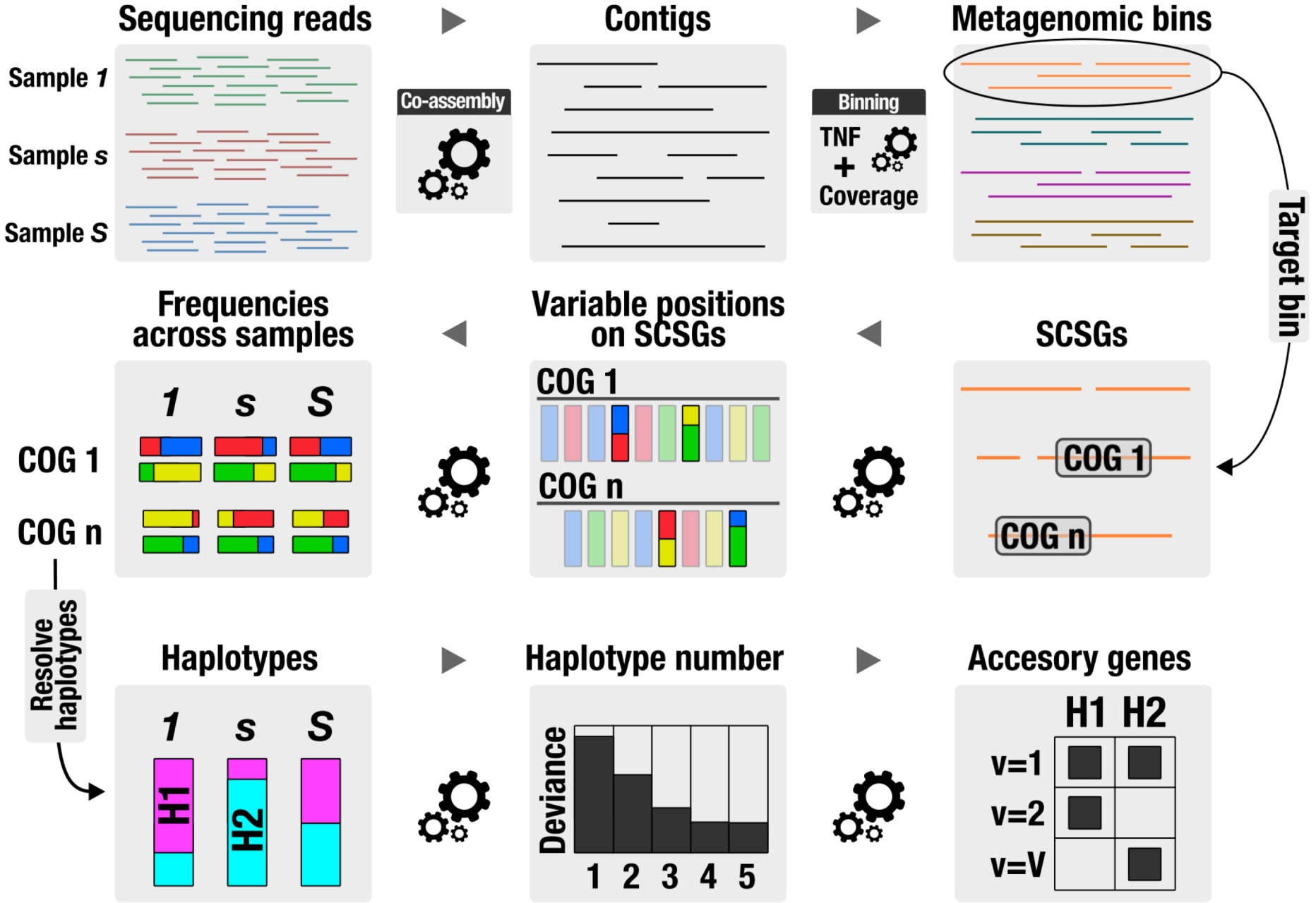
- Additional variation exists within MAGs both nucleotide variants on shared genes and variation in the accessory genome
- Variation of two sources, segregating variation within a population and sub-populations
- Read based haplotype resolution cannot span contigs or regions of low variation
- Methods 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)

Linking variants by co-occurrence




G = 2
G = 4
G = 3
G = 5
NMDS1
NMDS2
0.6
0.3
0.0
−0.3
−1.0
−0.5
0.0
0.5
Minor SNV strain mapping
0 0 0 1 0
0 0 1 0 0
0 0 1 1 0
0 1 0 0 0
0 1 0 1 0
1 0 0 0 0
1 0 1 1 0
1 1 0 0 0
1 1 1 1 0

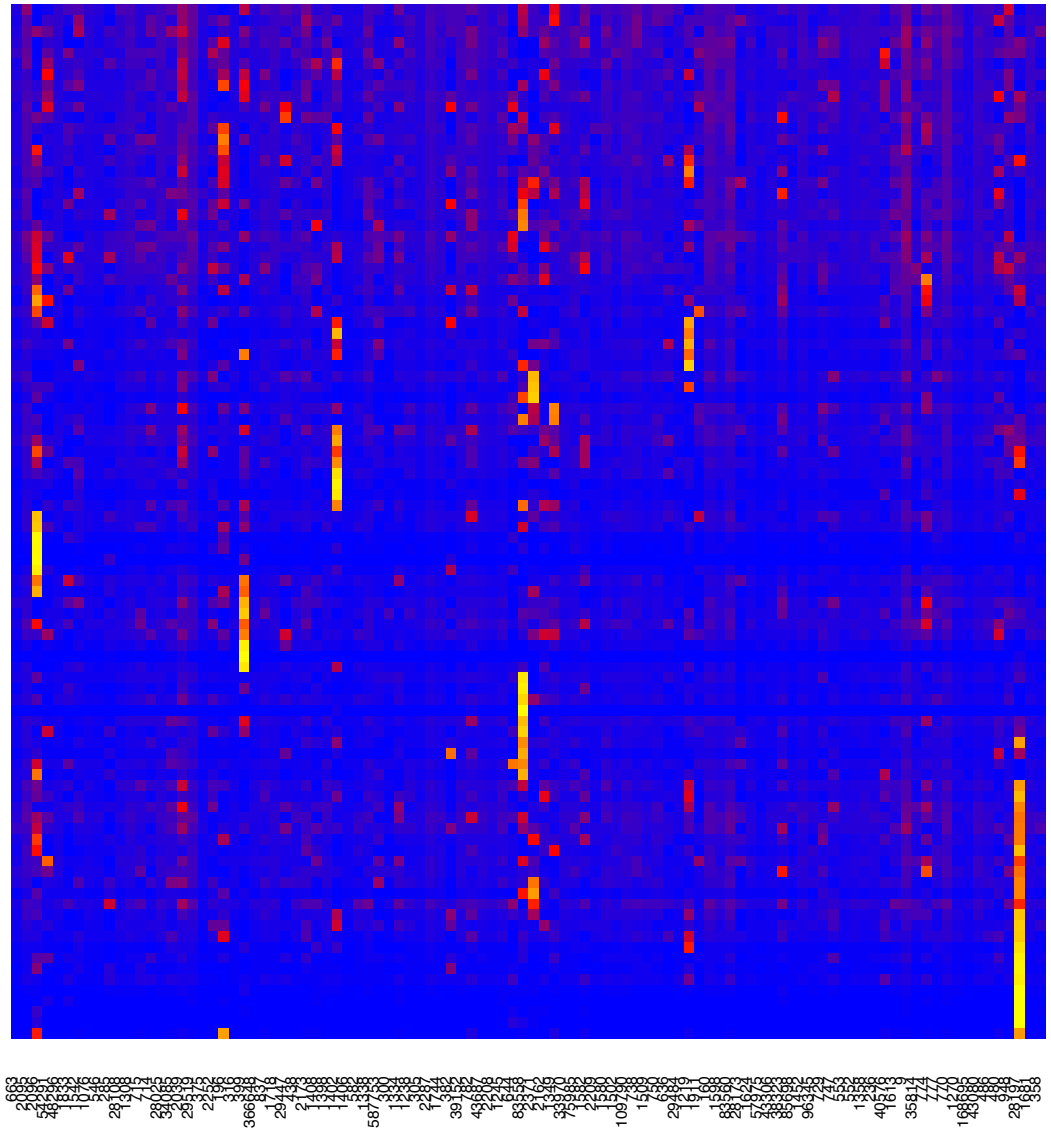
DESMAN https://github.com/chrisquince/DESMAN



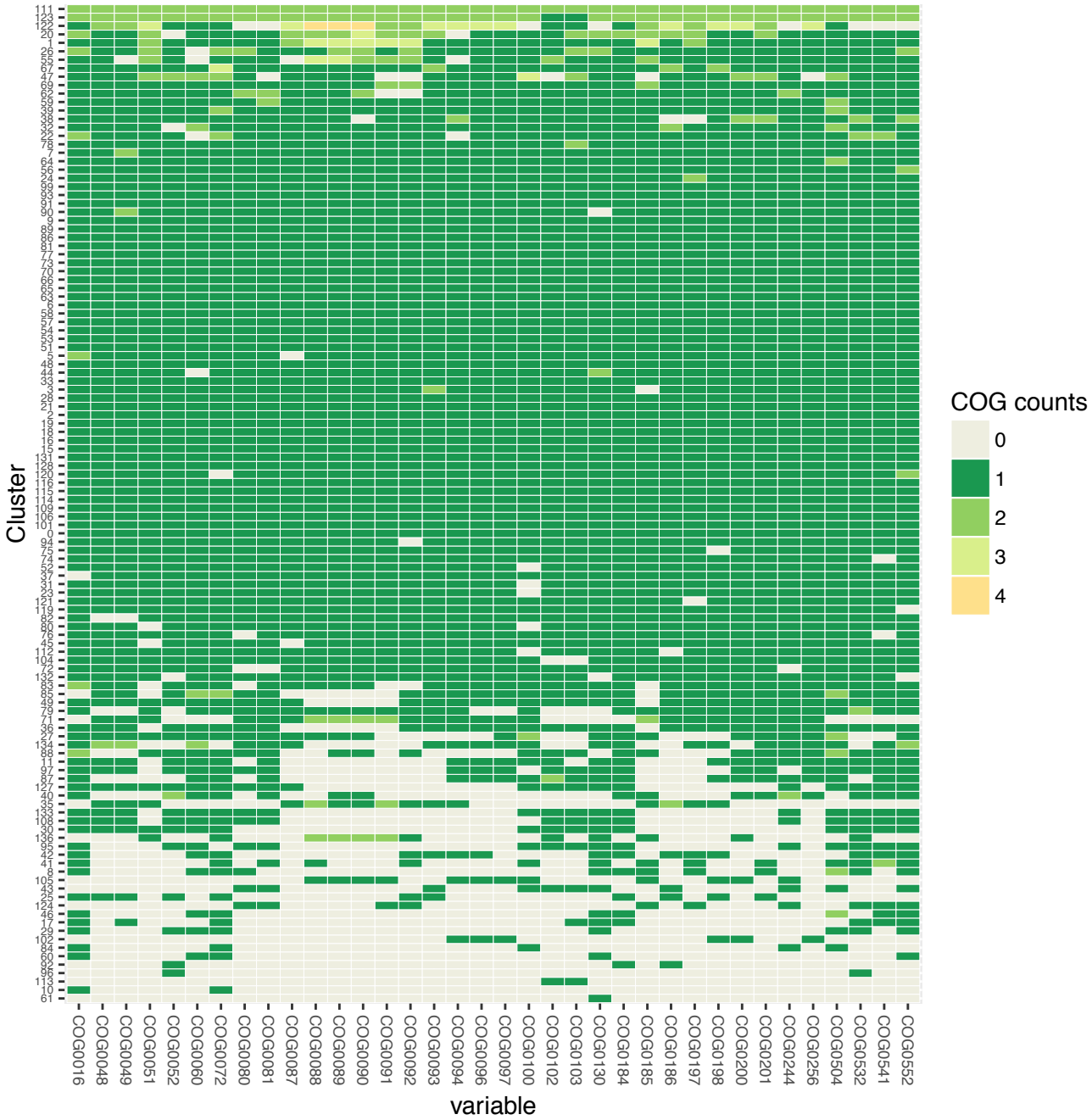
Quince, …, Eren Genome Biology 2017 18:181 https://doi.org/10.1186/s13059-017-1309-9

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

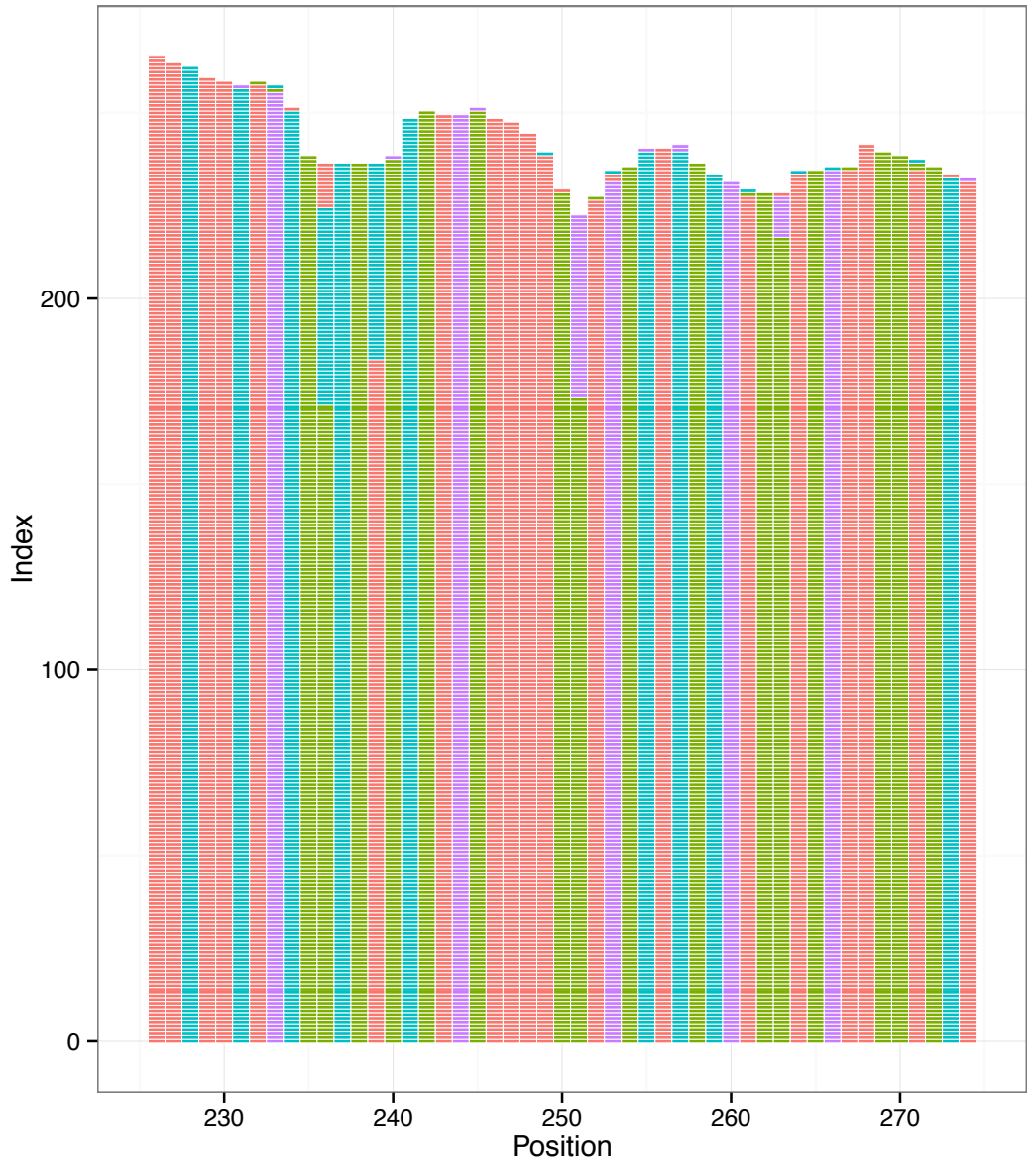


- Coassembly gave 74,580 contig fragments with a total length of 409 Mbp (687 Mbp for all 210 genomes)
- CONCOCT generated 137 clusters (species recall of 86.1% and a precision of 98.2%)
- 75 clusters with 75% of SCGs in single copy

DESMAN analysis using single-copy core genes

- Process each of the 75 – 75% complete clusters separately
- Map reads onto 36 SCGs
- If target taxa is known and cultured custom set of species specific single-copy core genes (SSCGs) could be used instead

Determining variant positions on SCGs

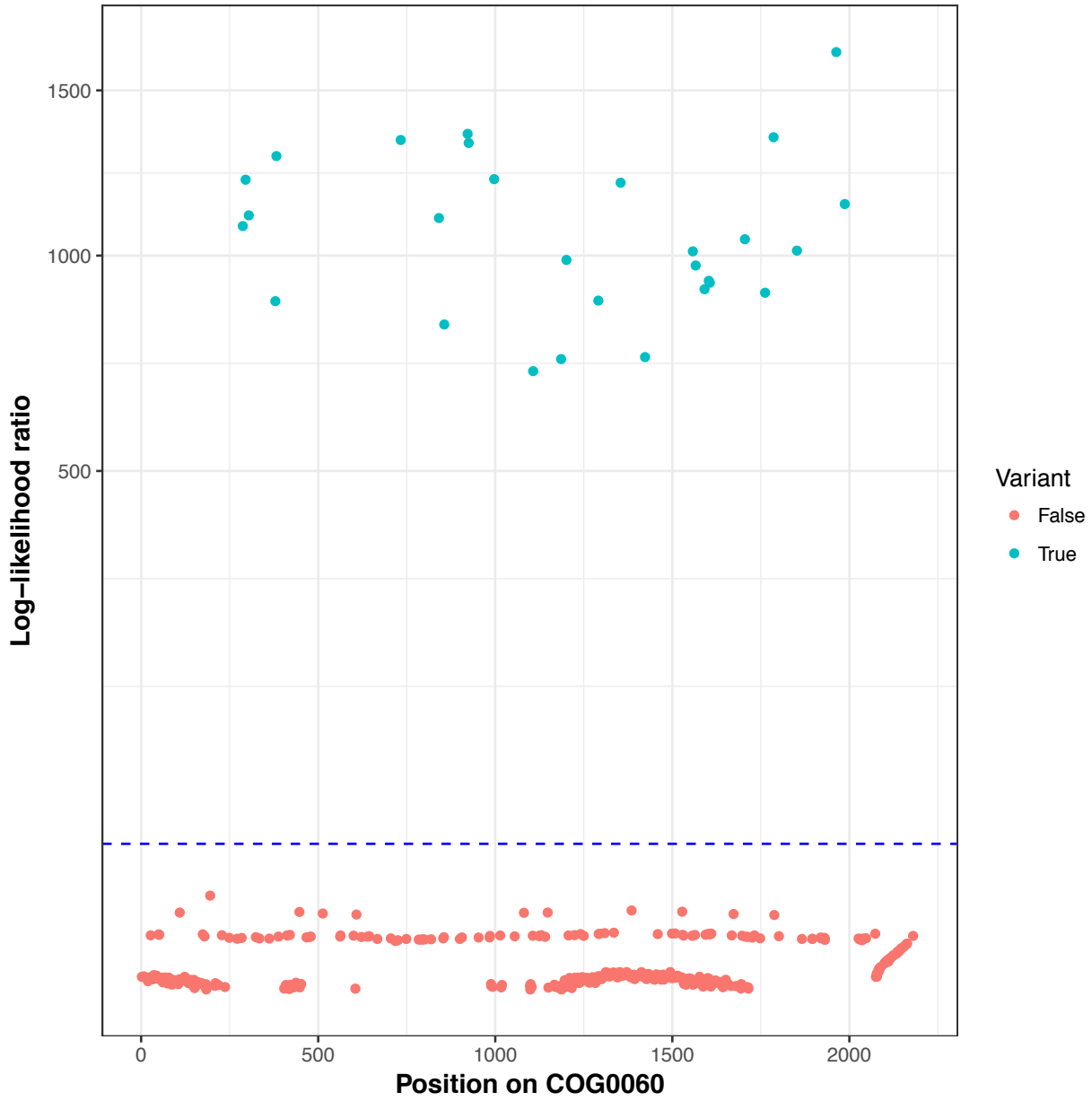


- Ignore distribution across samples just ask if minority bases could be created by errors alone
- Assume errors are position independent with true base *a* generating observed base *b* with probability:

$$\varepsilon_{a,b}$$

- 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

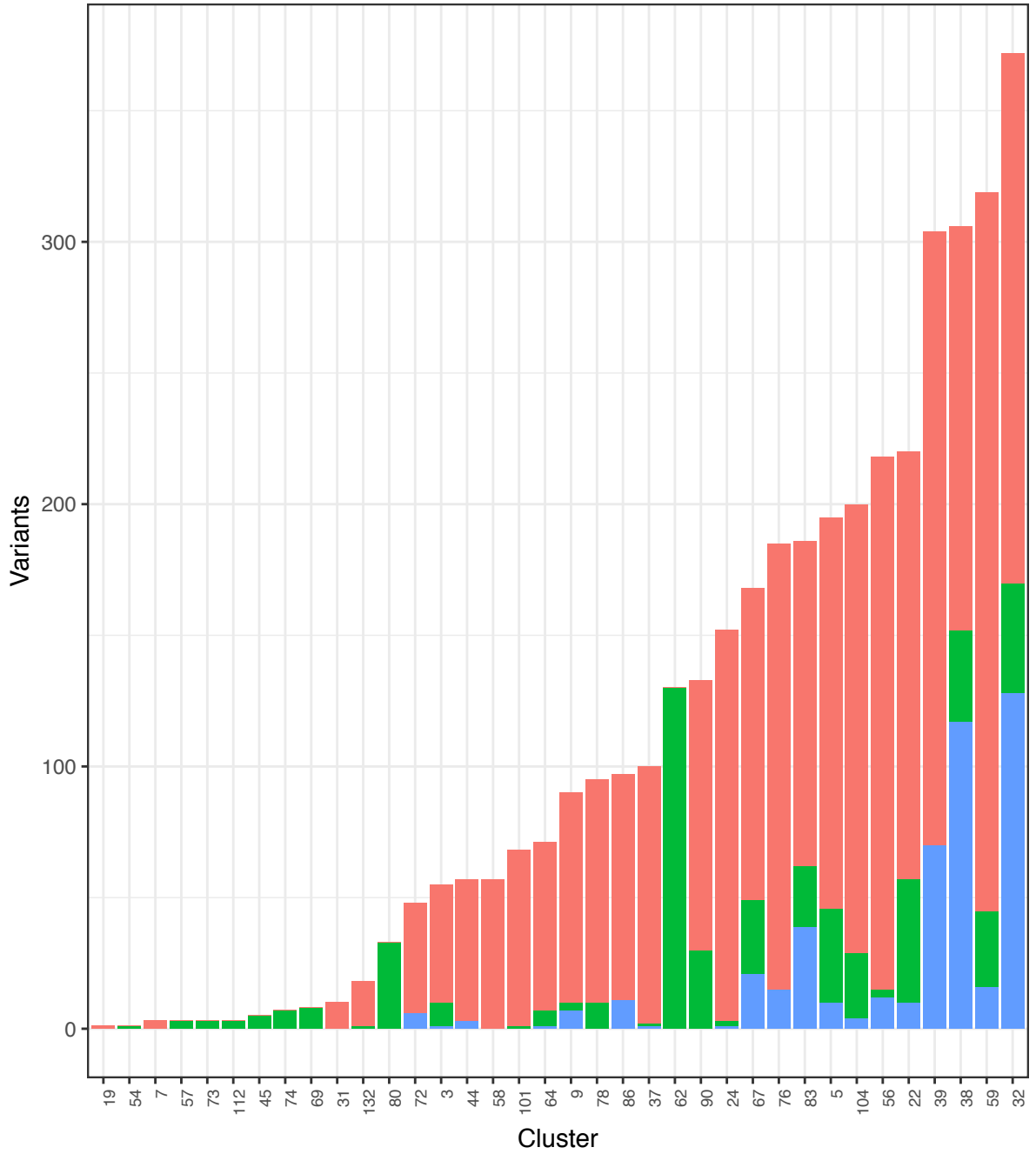
Over all 27 SCGs and 26015 positions with q < 0.001

Predicted

	False	True
False	25015	1
True	1	98

Recall = 98.9%, Precision = 98.9%

Complex mock variant results

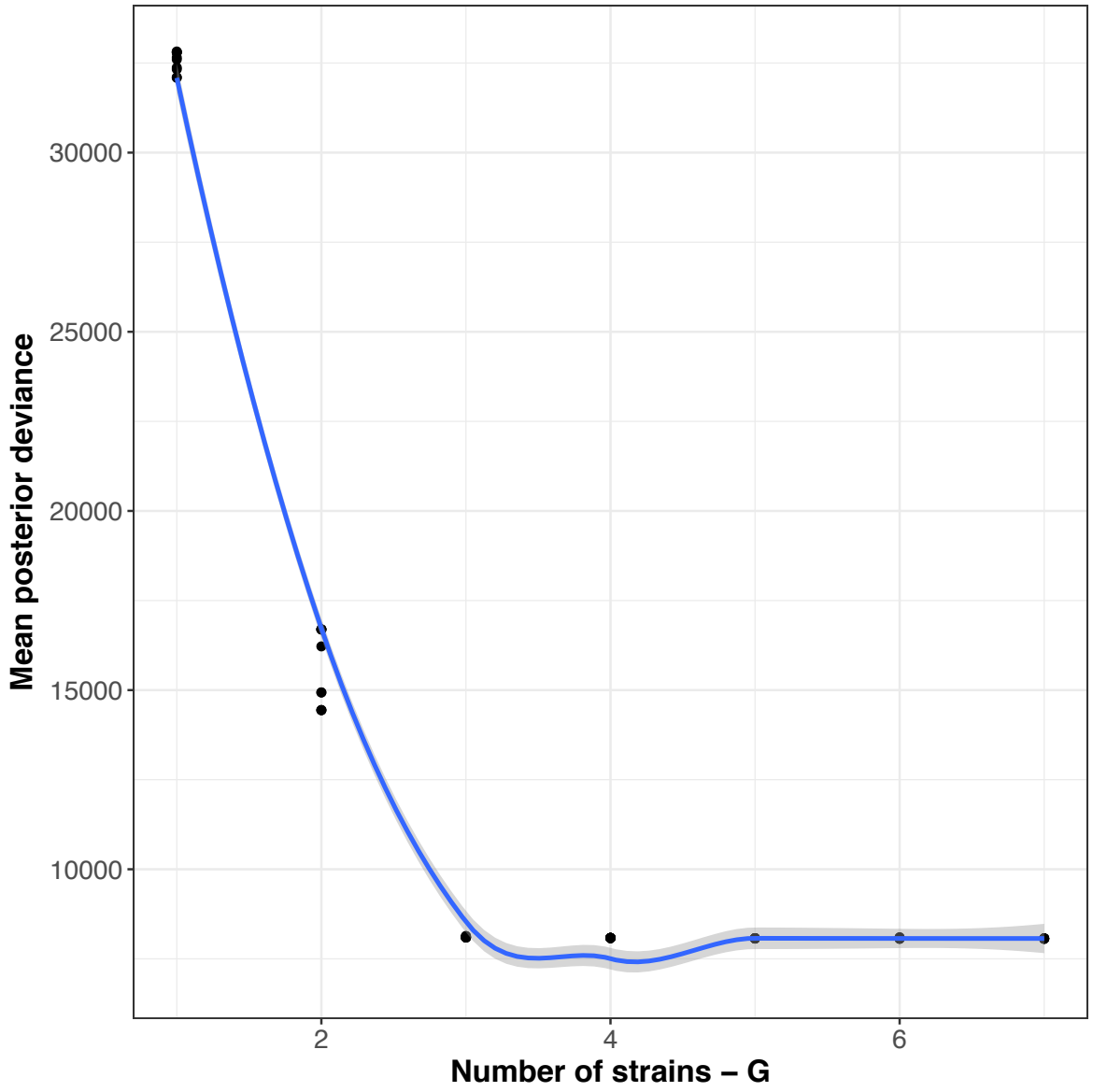


- Filtered SCGs for coverage outliers (medians SCGs 35 to 30)
- Of the 75 clusters we predicted variants in 36, including 27 of the 29 that should have had SNPs
- Over those 27 clusters we predicted a median of 99 variants per cluster, with a mean precision of 92.32% and a mean recall of 91.85%
- 25 of the 27 clusters had at least five variants (DESMAN)

Gibbs sampling algorithm

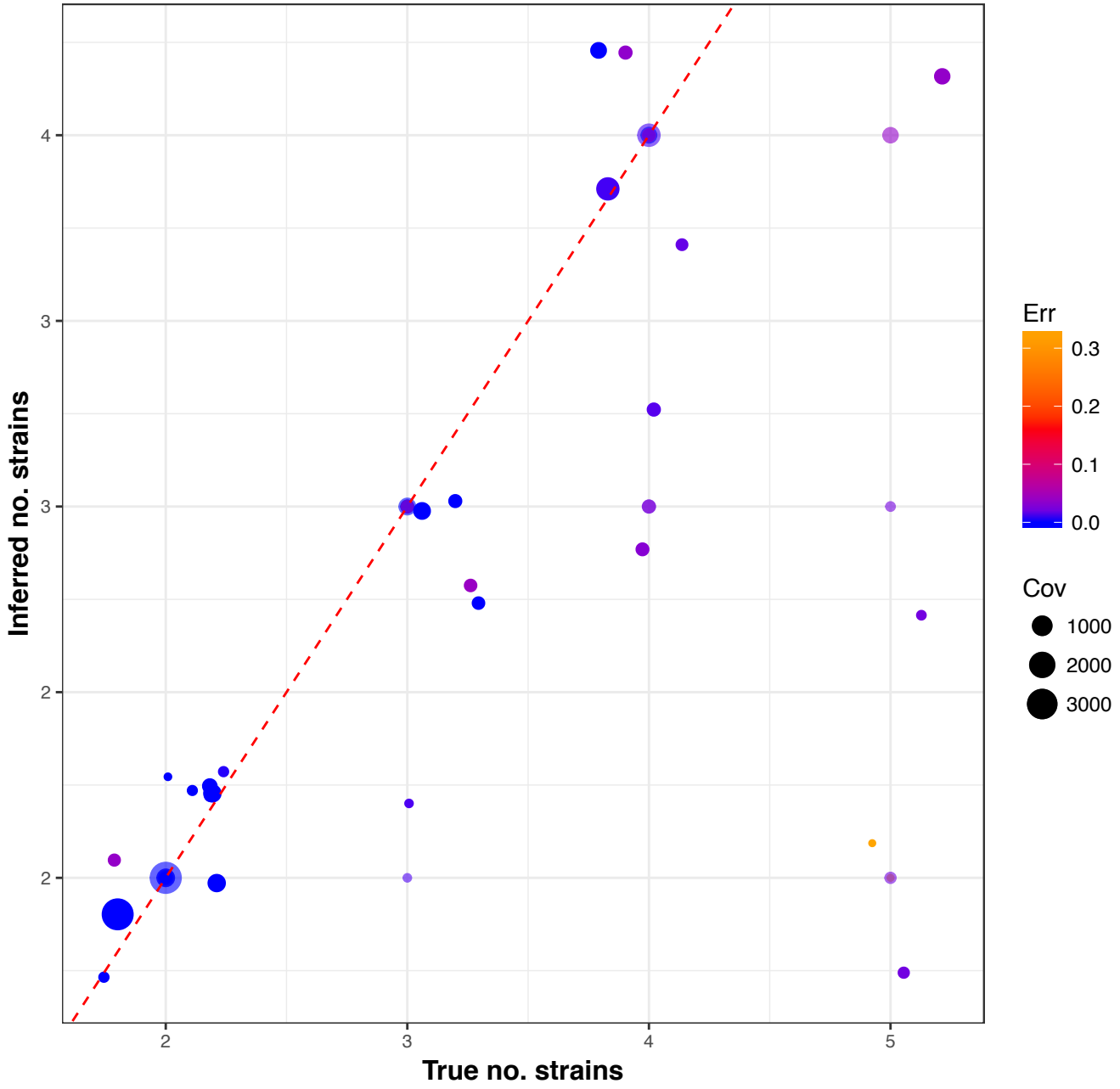
- Assume bases are independent leads to binomial probability at each position
- Devise an iterative algorithm that successively samples:
  - Haplotypes
  - Relative frequencies
  - Error rates
- Bayesian algorithm generates distribution of fitted values
- Negative log-likelihood describes overall fit
- 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%

Results for Cluster 37

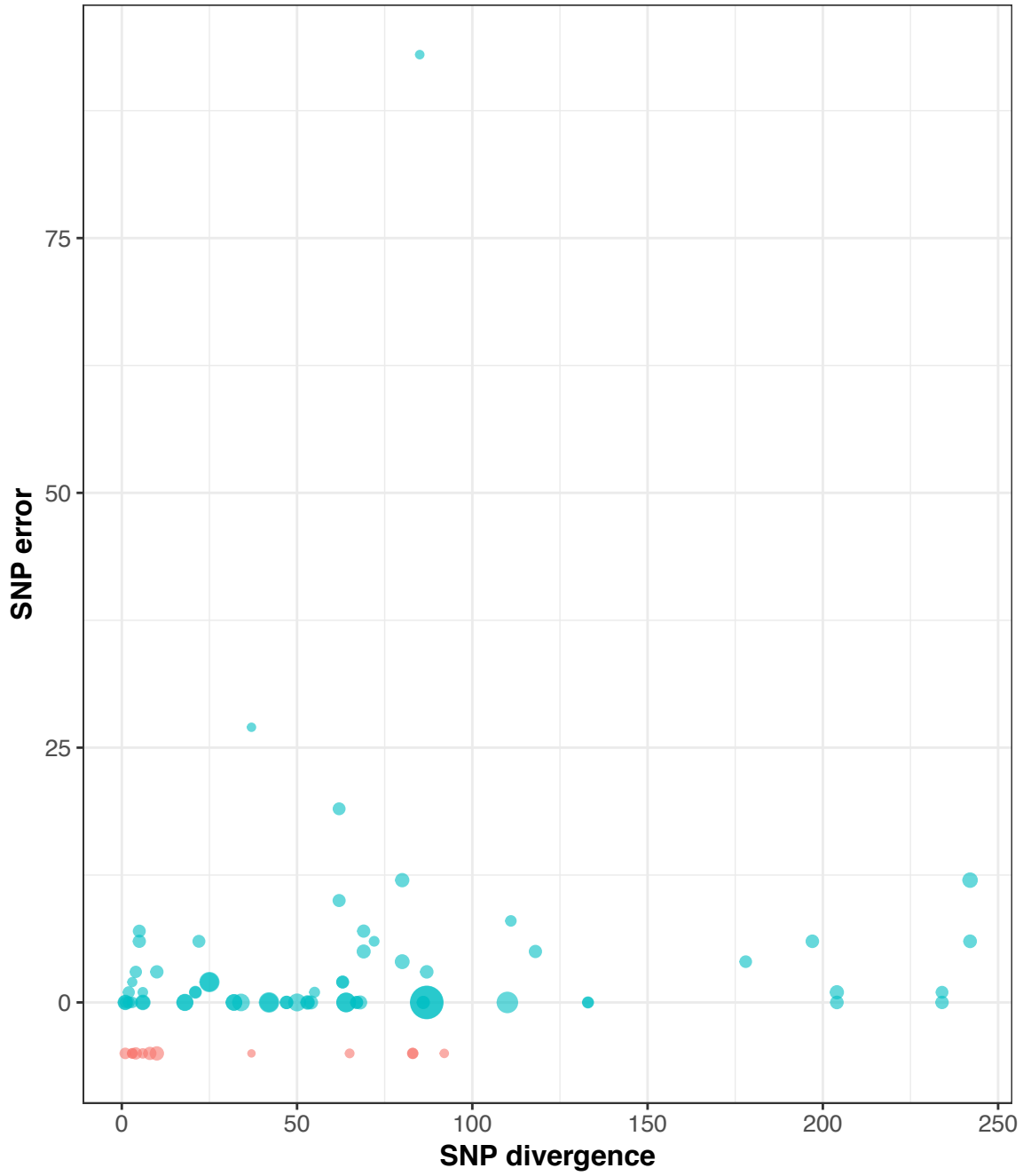


Model fit fails to improve after 3 haplotypes

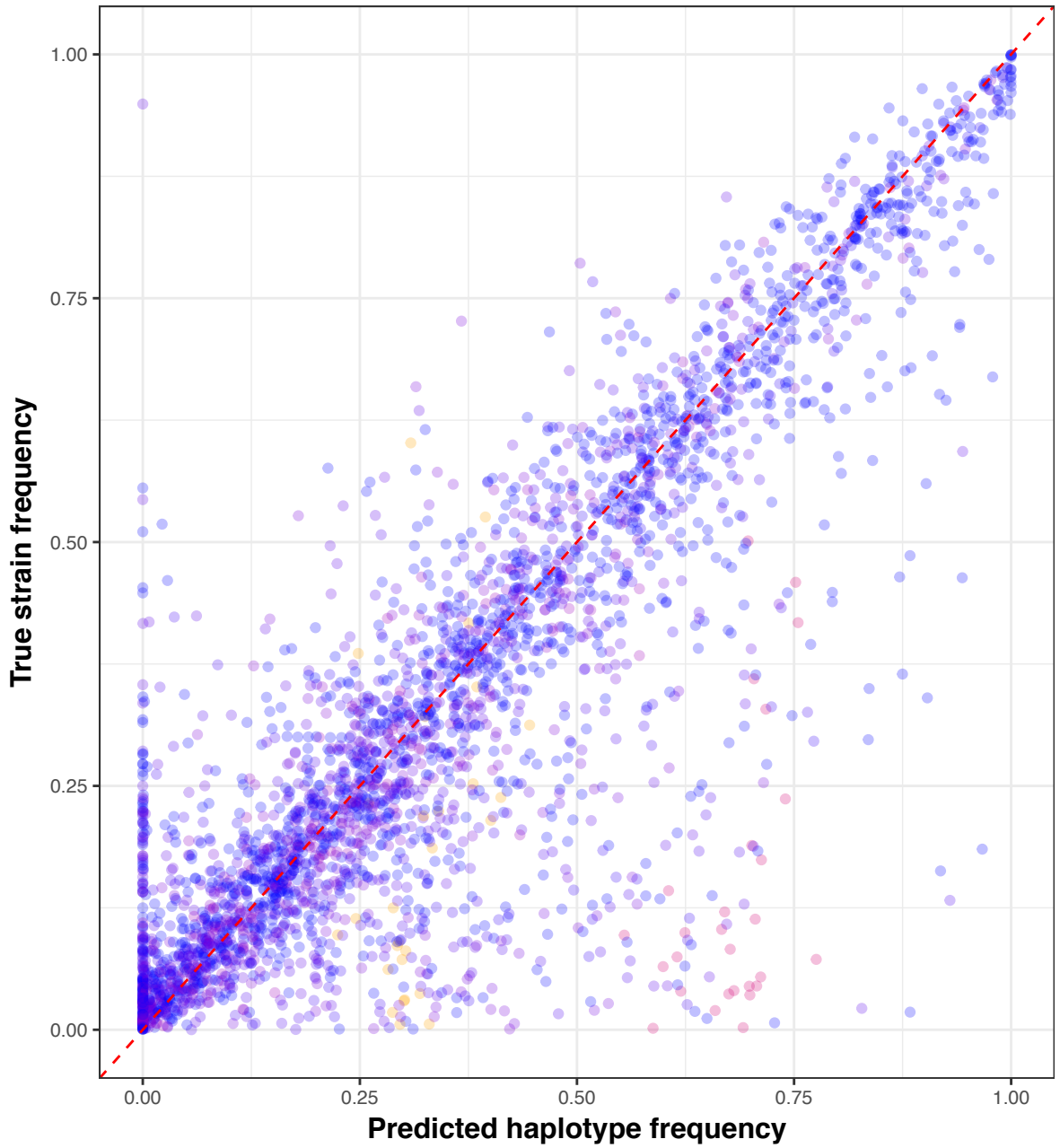
The best fit 3 haplotype run had each haplotype map onto a different reference strain with no errors



- Predicted the correct haplotype number for 18/25 (72%) of the clusters
- For 22/25 (88%) prediction was within one of true value

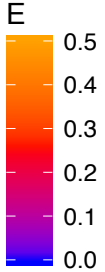


- Mean SNV error rate below 1% for 15/25 (60%) of clusters
- Median of 0.25% and and a mean of 2.38%
- No correlation between error rate and either the number of variants in the cluster or coverage
- We find a positive relationship between detection (67 out of 79 detected) and individual strain coverage (p-value = 0.0035)



Linear regression: slope 0.820, adjusted R-squared 0.741, p-value: < 2.2e-16

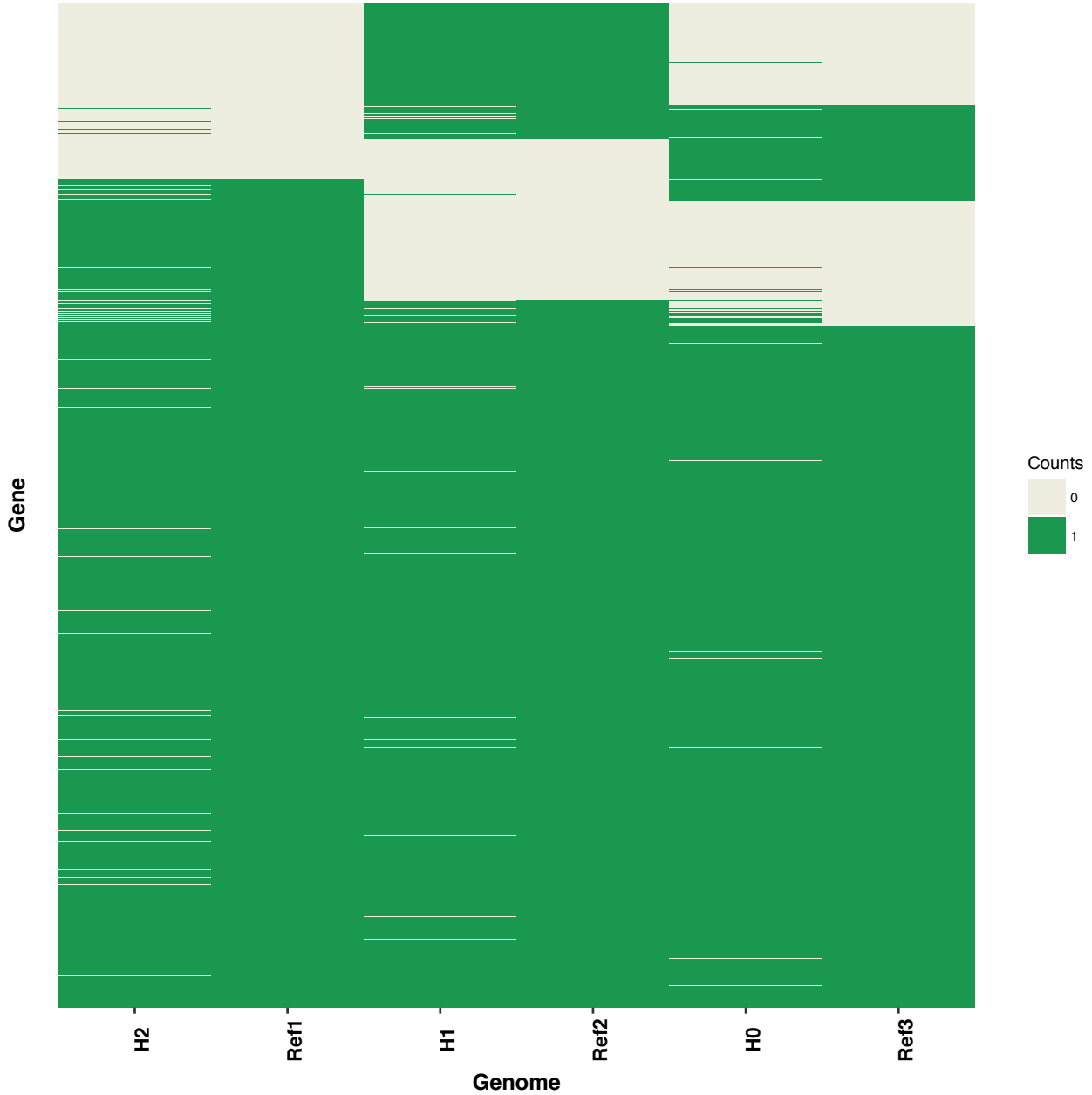
Haplotypes with E < 0.01, slope 0.853, adjusted R-squared 0.810, p-value: < 2.2e-16



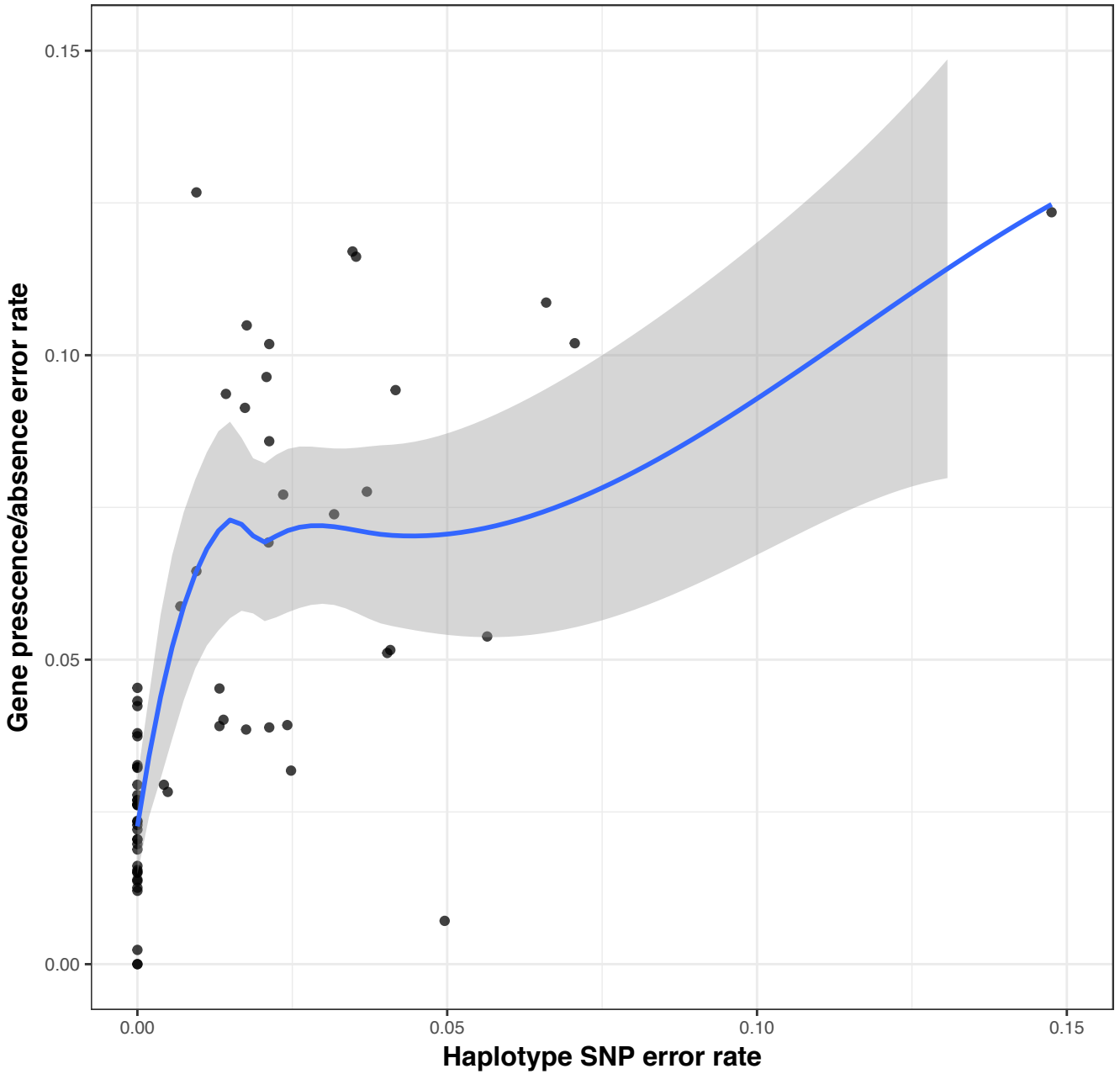
Performance of Lineage (O'Brien et al. Genetics 2014)

- On the complex mock community applied to SCGs DESMAN consistently (but marginally) outperformed Lineage:
  - Haplotype number correct in 18 rather than 15 MAGs
  - SCG SNV median and mean error rates were also lower at 0.25% and 2.38% for DESMAN vs. 0.641% and 3.583% respectively for Lineage (p-value = 0.06)
- On a simpler 5 strain E. coli 20 genome mock DESMAN did dramatically better 99.58% vs 76.32% accuracy
- Difference more complex problem 6,444 variants on 372 single-copy species genes
- Constrains would not run with more than ~30 samples

Cluster37: Resolving the accessory genome



- Don’t know whether an accessory gene is present or absent in a strain
- This changes the effective strain proportions
- We infer the gene presence/absences assuming the strain proportions and error matrix are fixed at their posterior mean values from core genes
- Overall accuracy: 96-97% of gene presence/absence correct for three haplotypes

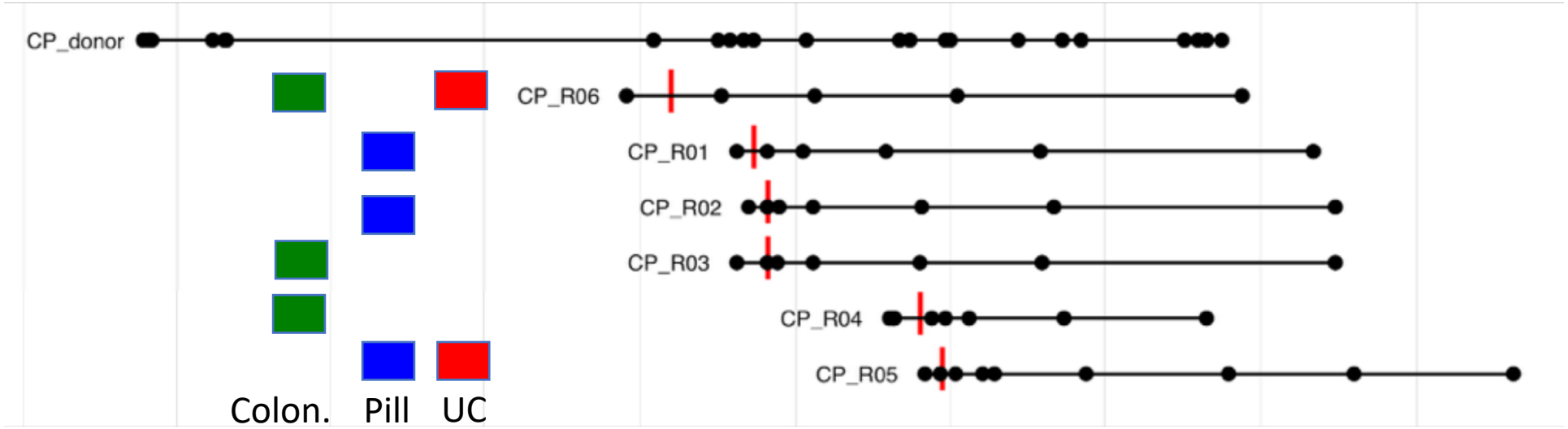


- Averaged over 67 detected haplotypes gene prediction accuracy was 94.9% (median 96.26%)
- Increased to 97.39% for the 39 haplotypes that we predicted with error rate < 1%

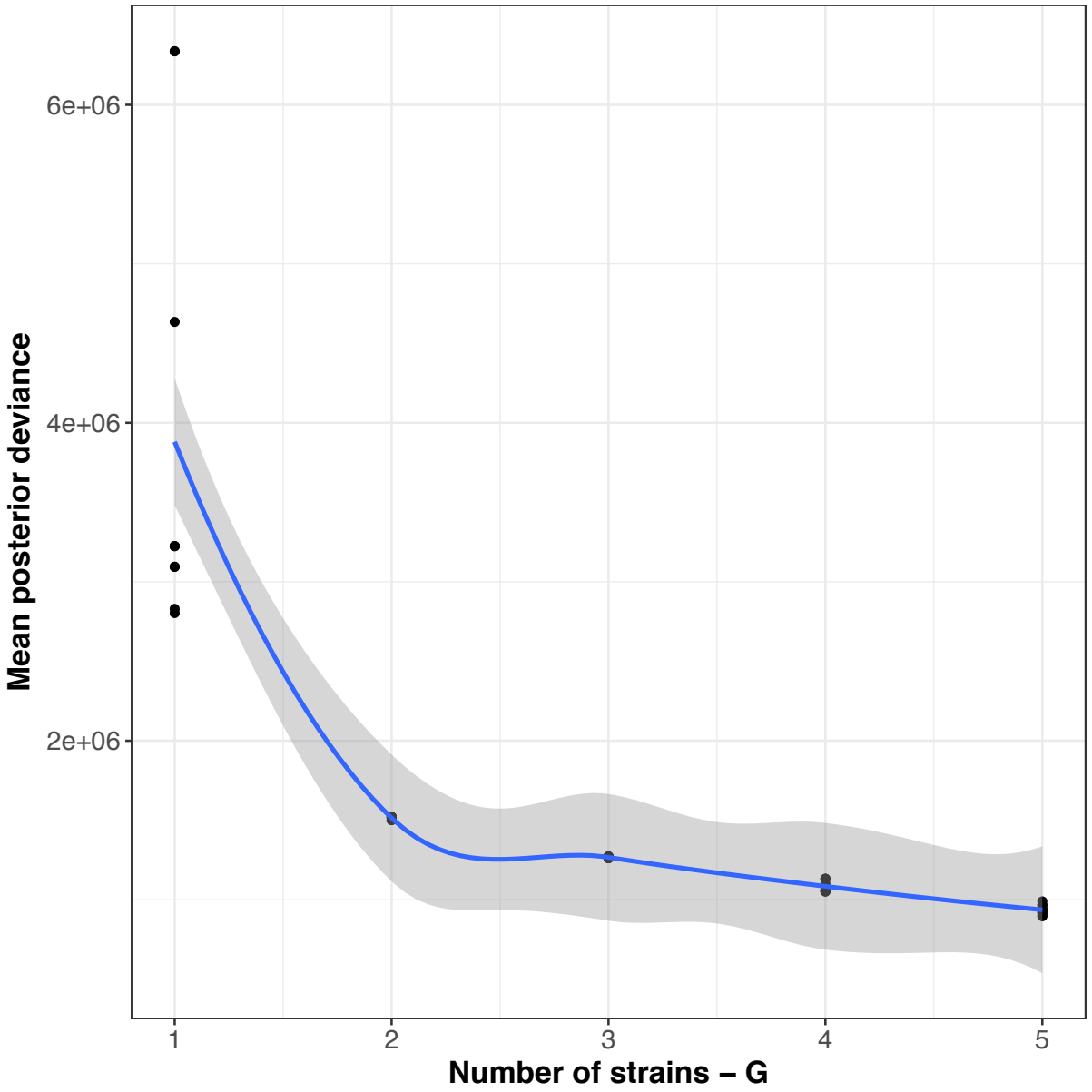
Fecal Microbiome Transplants

(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.

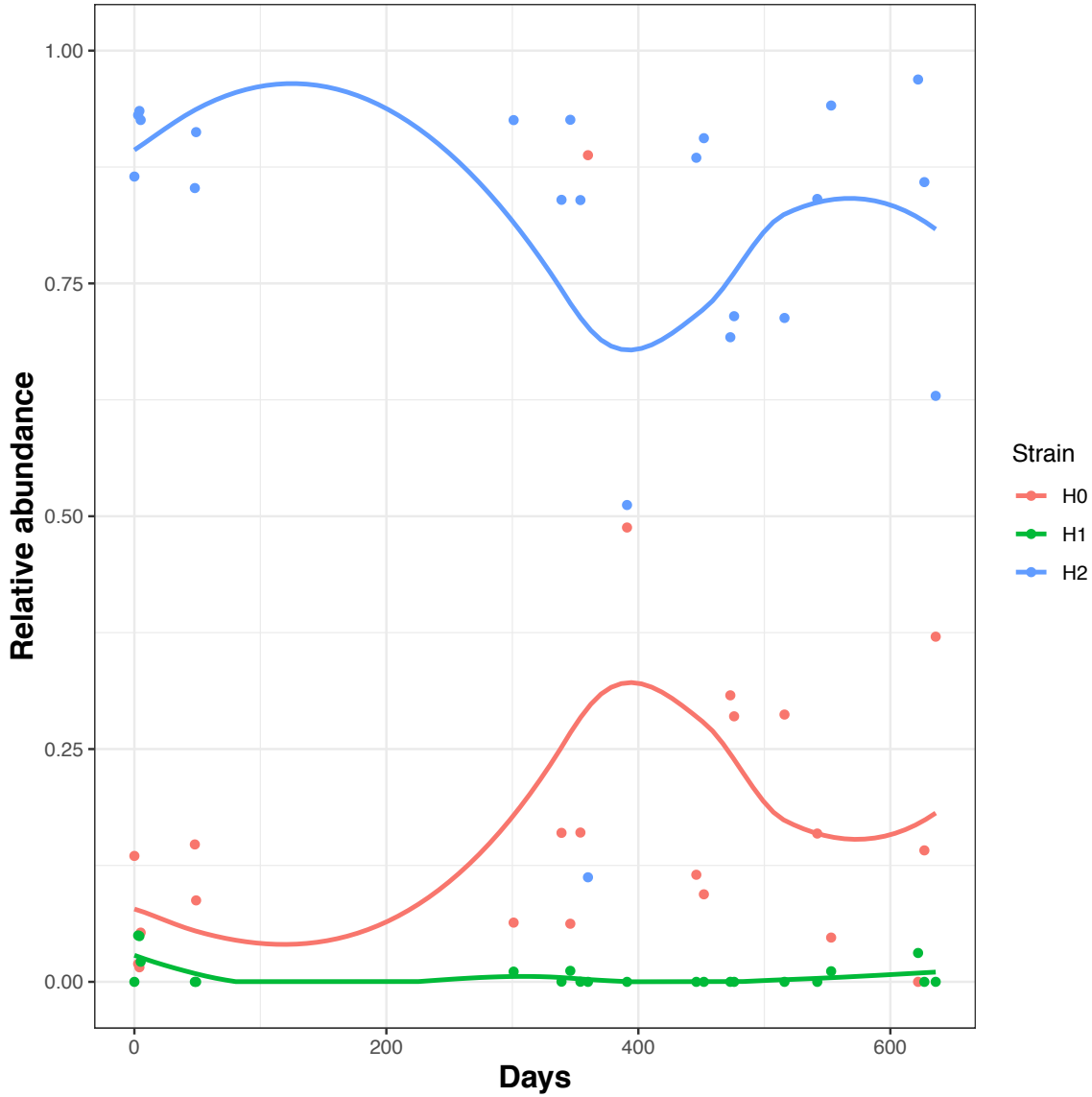


FMT DESMAN CP/R03 strain analysis

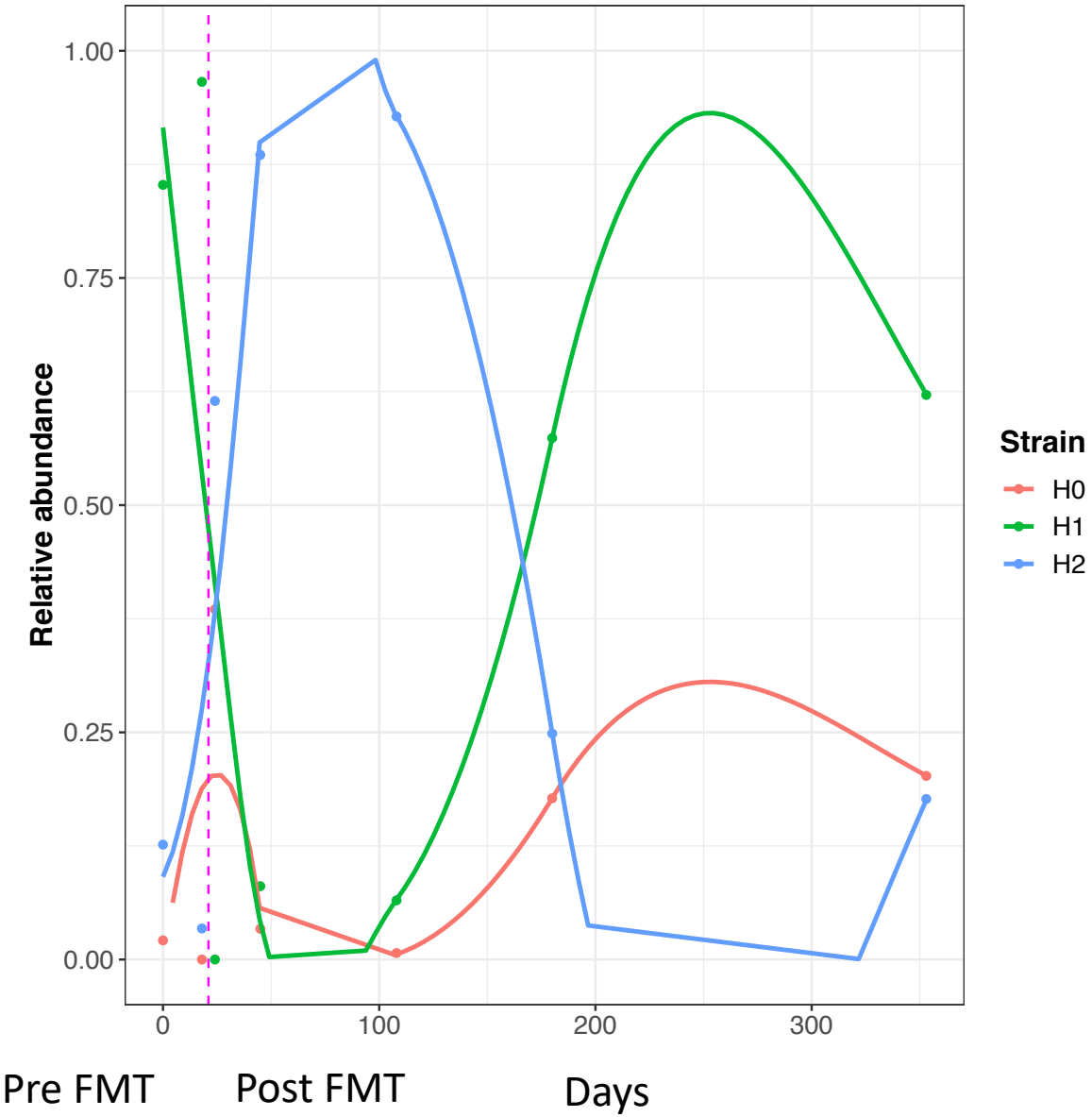


- Detected variants on one Anvi'o MAG assigned to Allistipes Finegoldii
- Detected 35,910 variants on 2.8 Mbp ~ 1.2%
- Applied DESMAN to 5,000 variant positions increasing strain numbers from 1 to 5
- 3 strains selected as best fit
- Varying between 23 – 53% of SNV positions

CP Donor

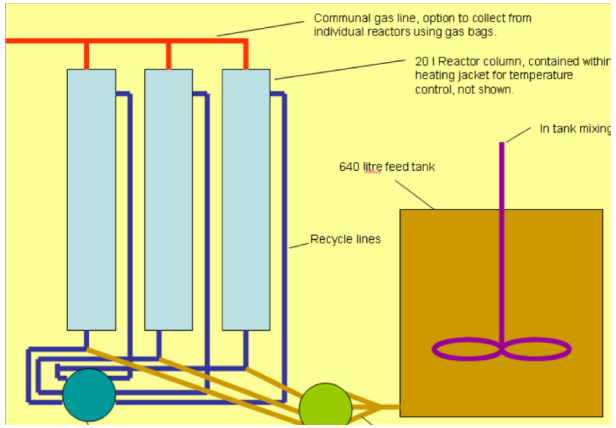


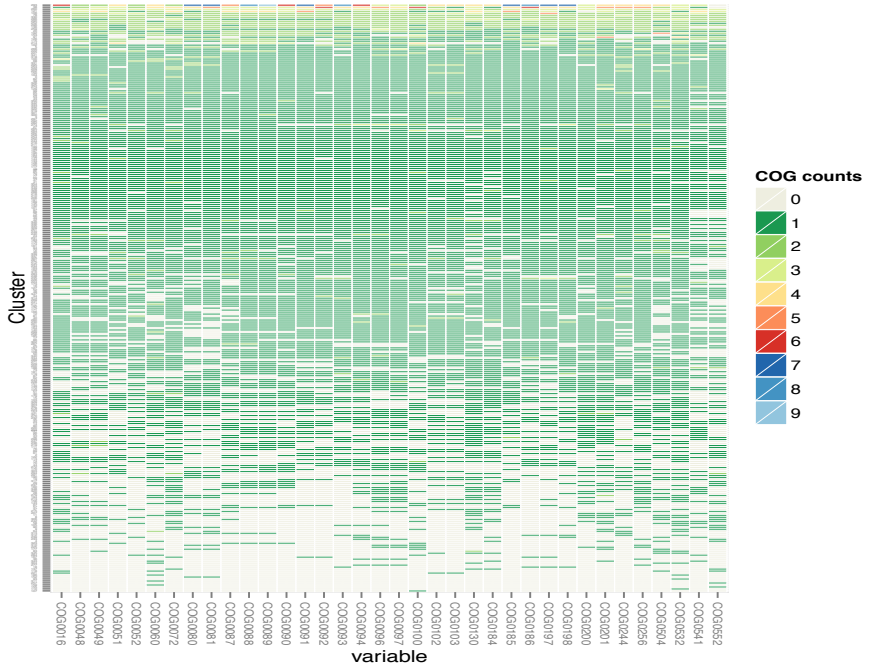
Recipient R03

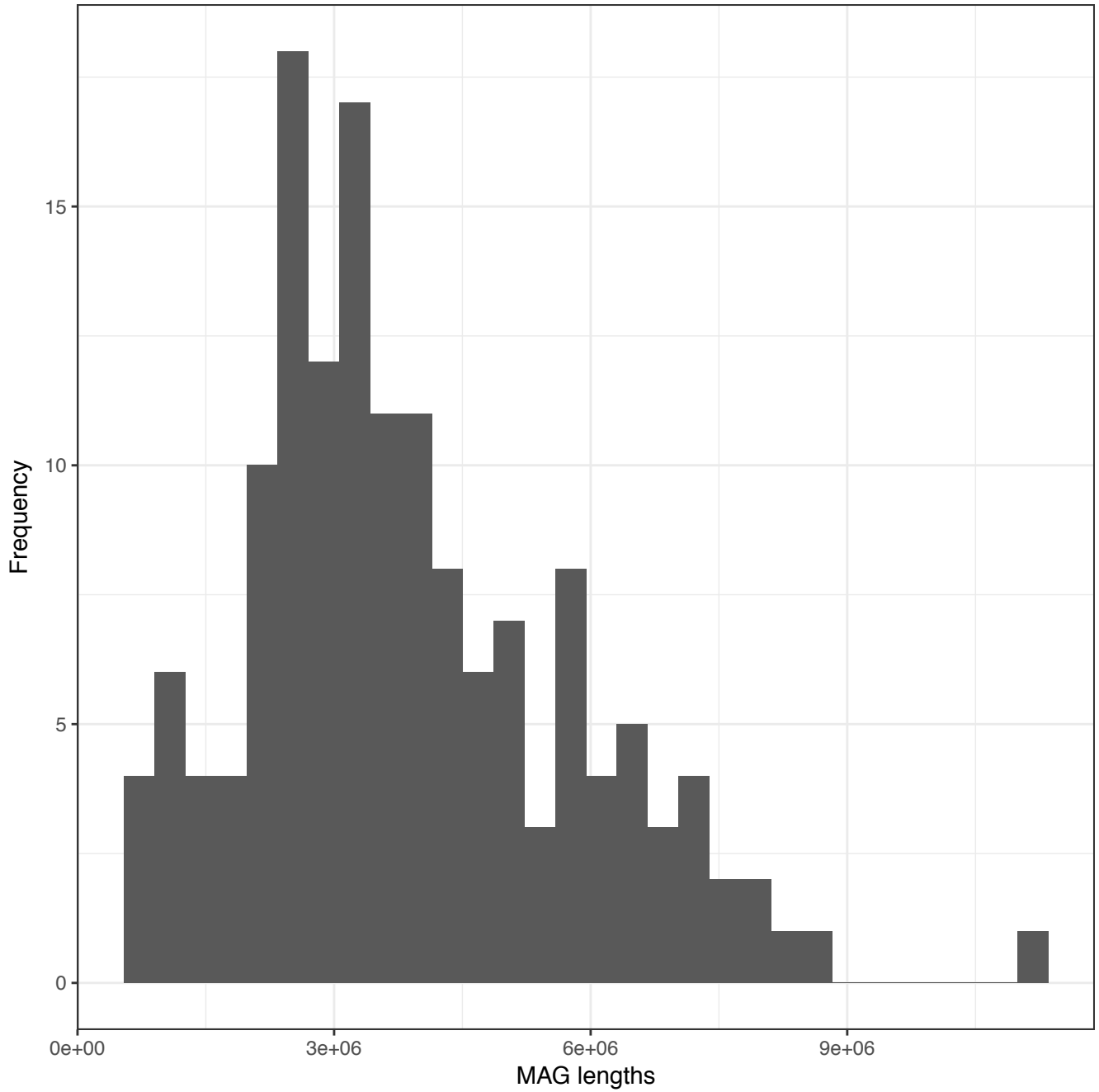


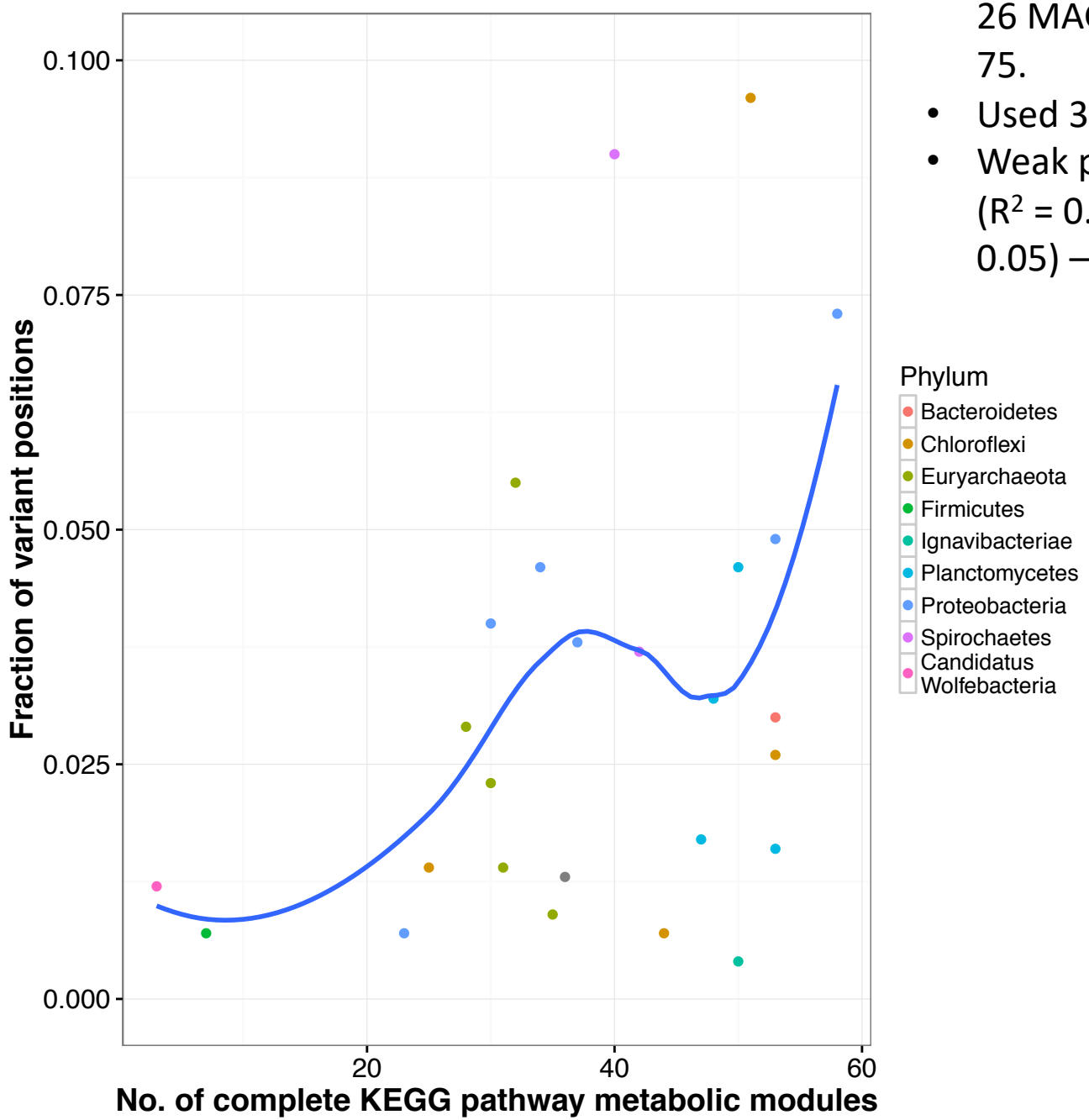
Expanded Granular Sludge-Bed Laboratory Bioreactors (EGSB)

- 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



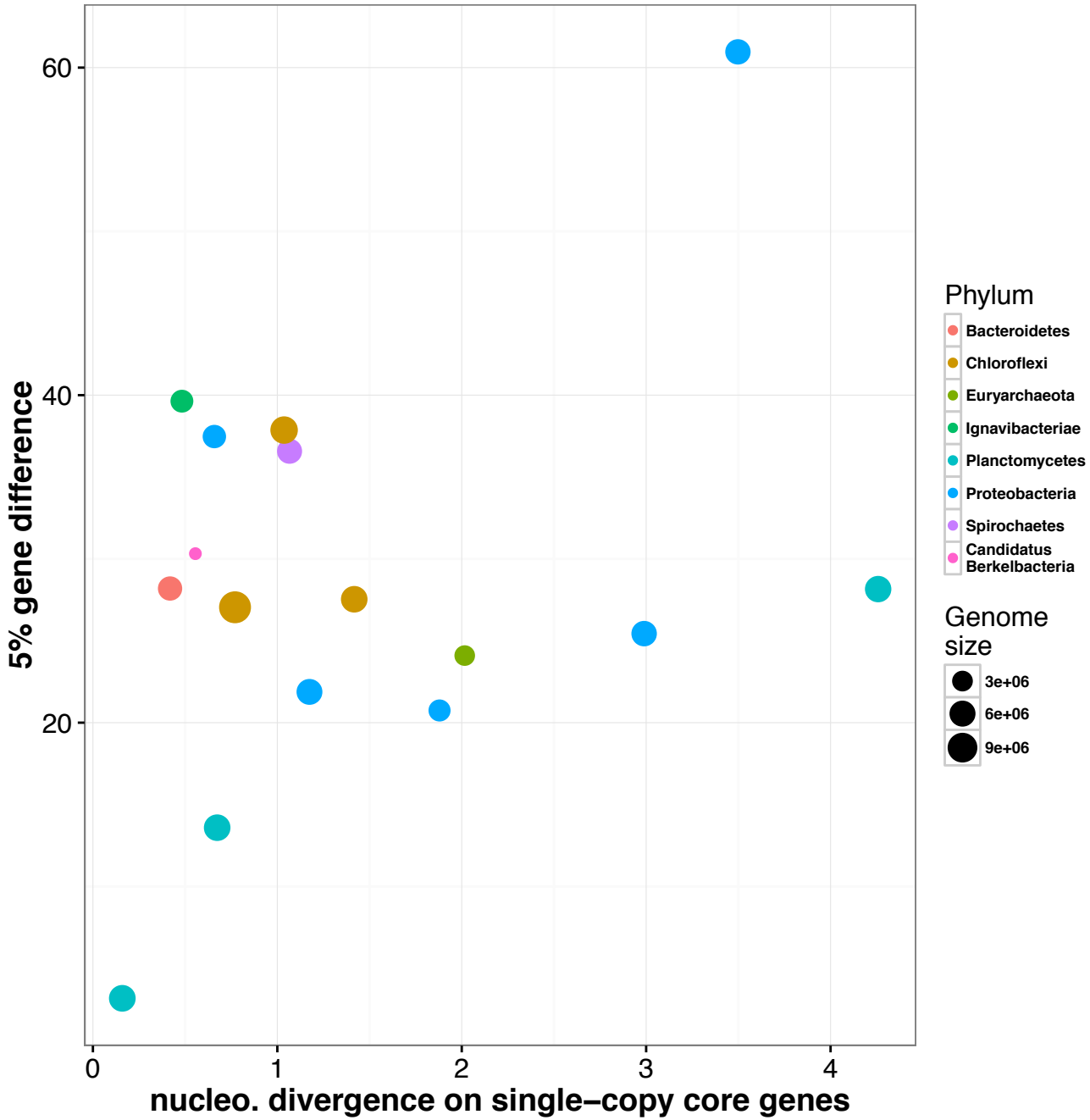



Frequency
15
10
5
0
0e+00
3e+06
6e+06
9e+06
MAG lengths



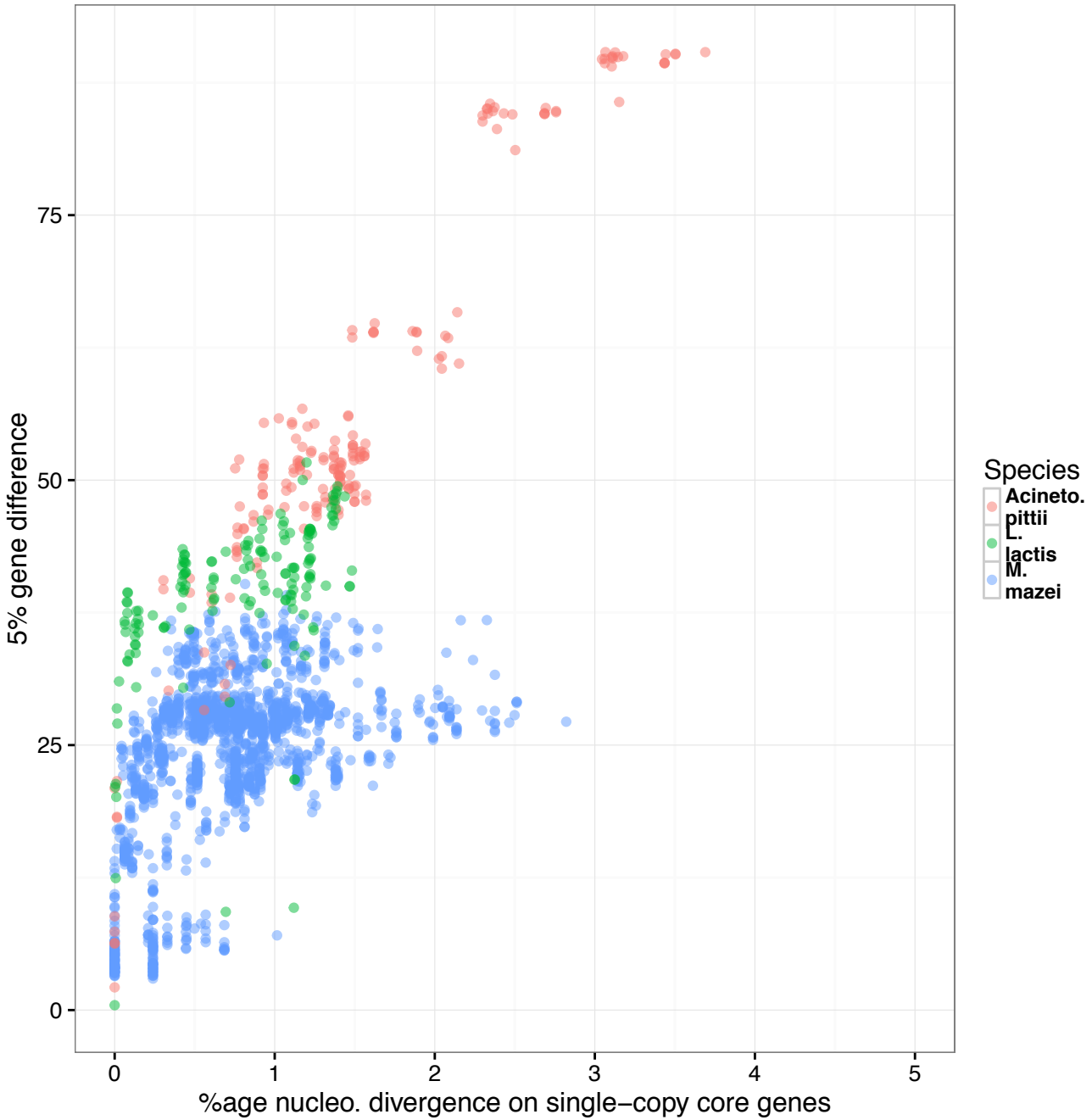
- Applied strain resolution to 26 MAGs with total cov. > 75.
- Used 36 SCGs
- Weak positive relationship (R^2 = 0.1132, p-value = 0.05) – c.f. Muller et al. 2014

AD strain analysis

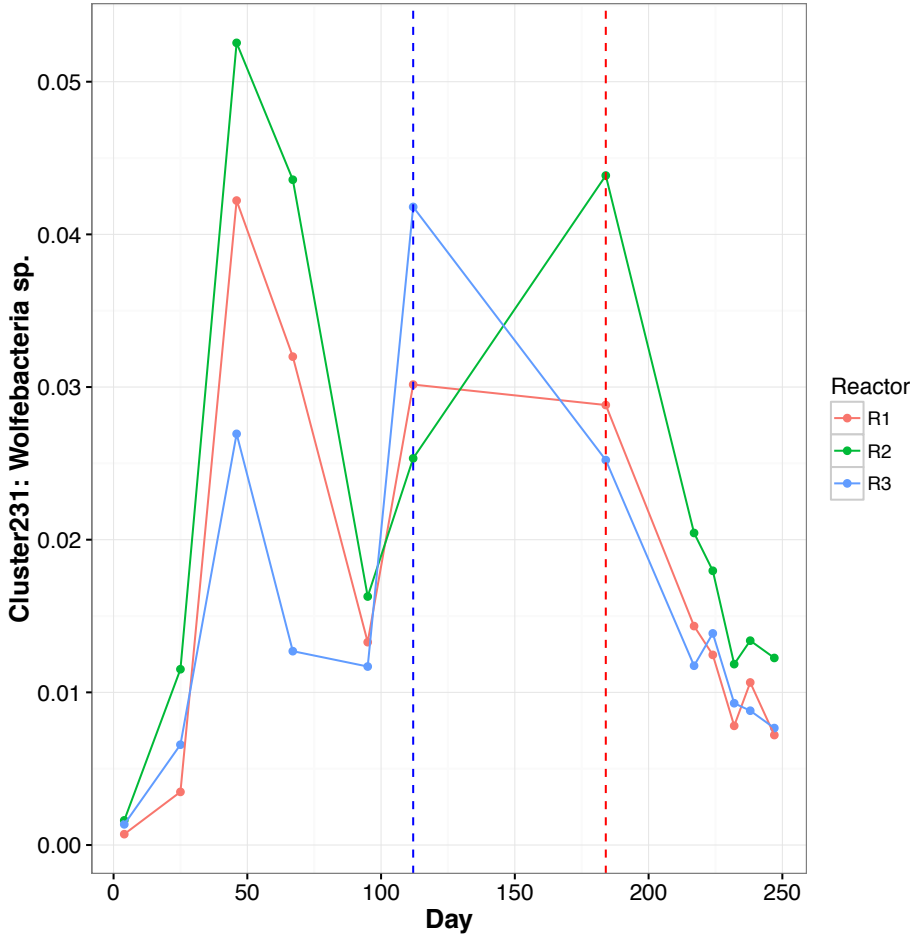


- 16 of 36 MAGs existed as two strains
- Inferred genome divergence between strains based on 5% gene clusters

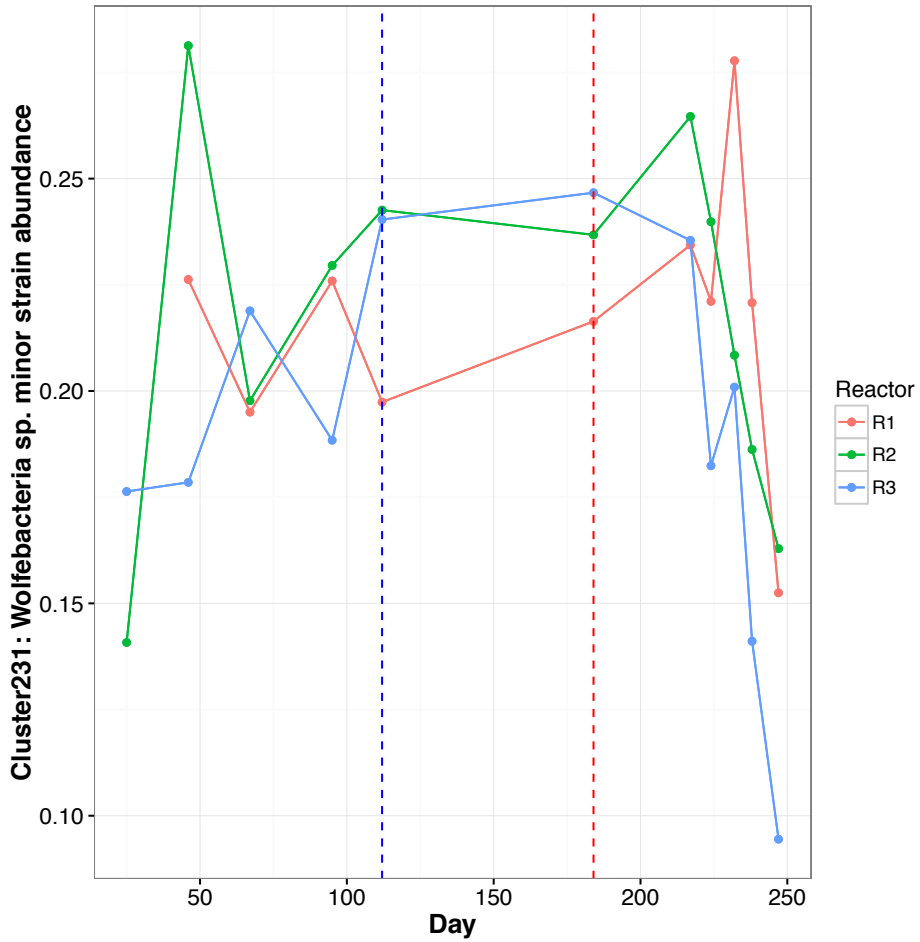
Comparison to real environmental organisms



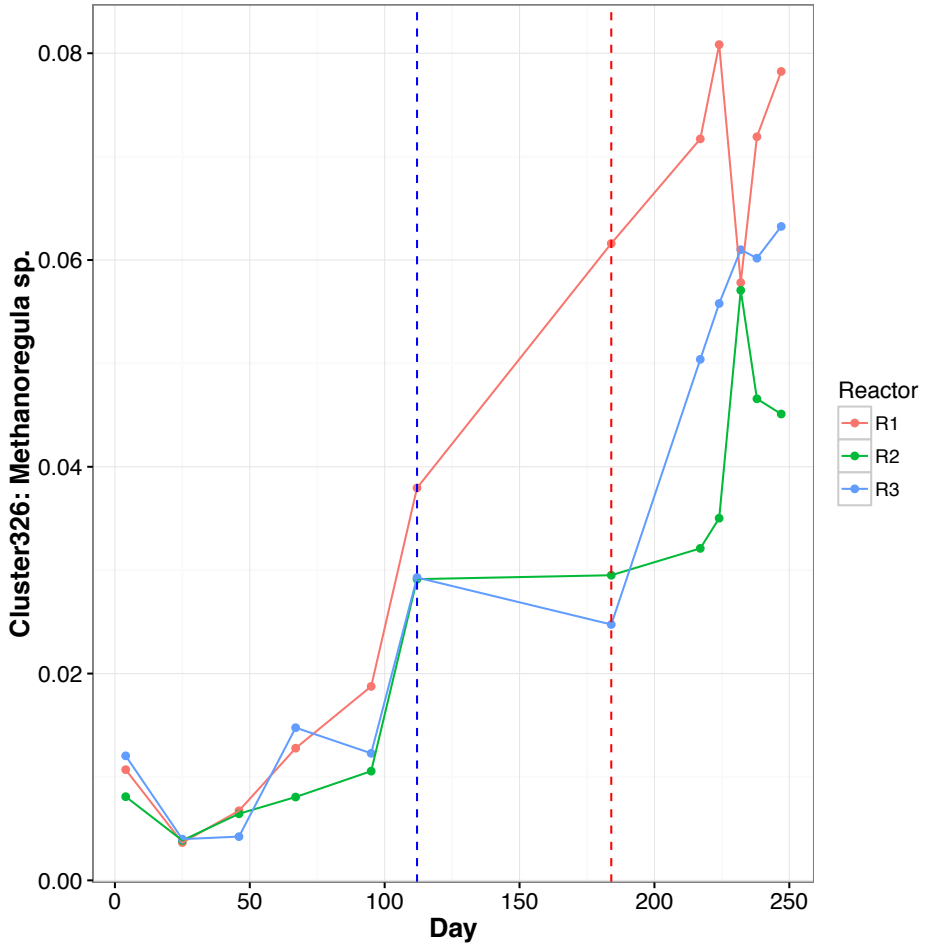
MAG abundance



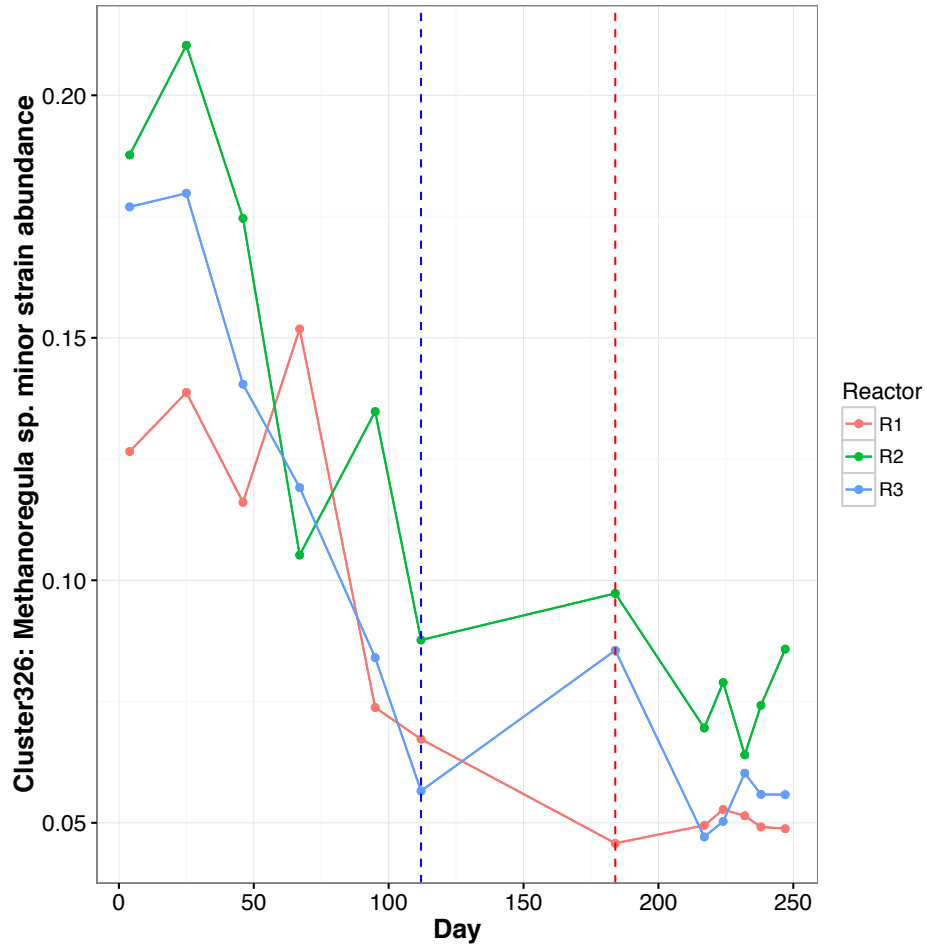
Strain abundance



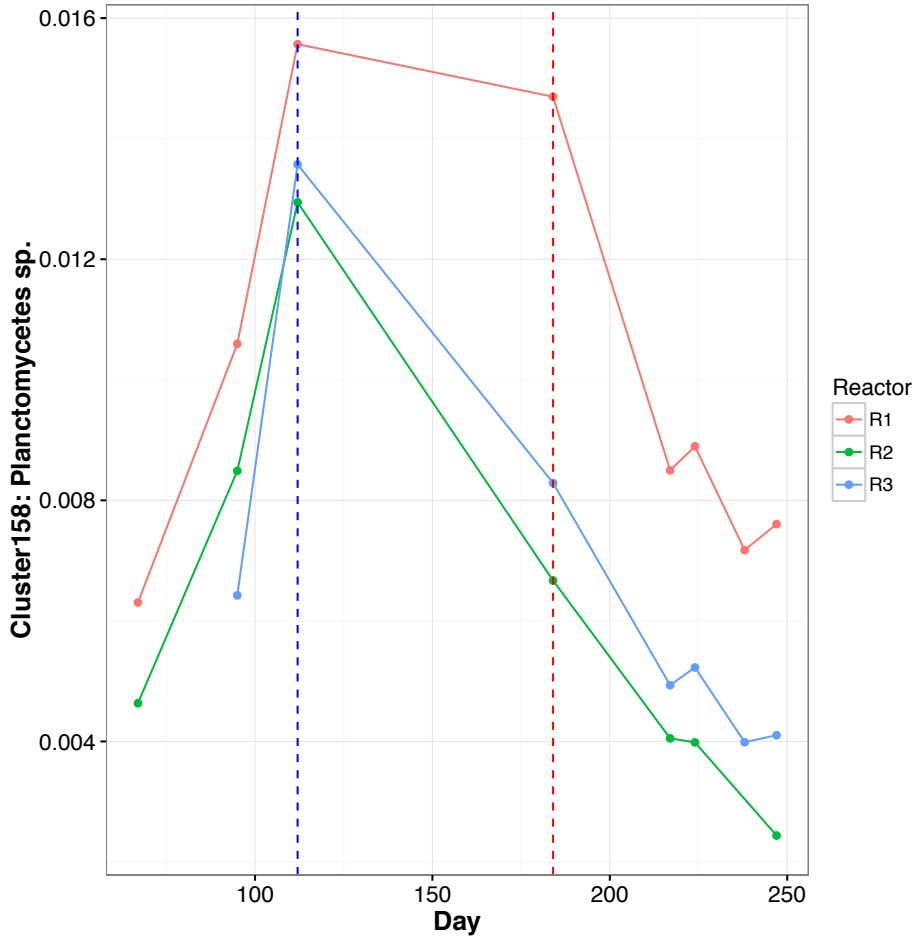
MAG abundance



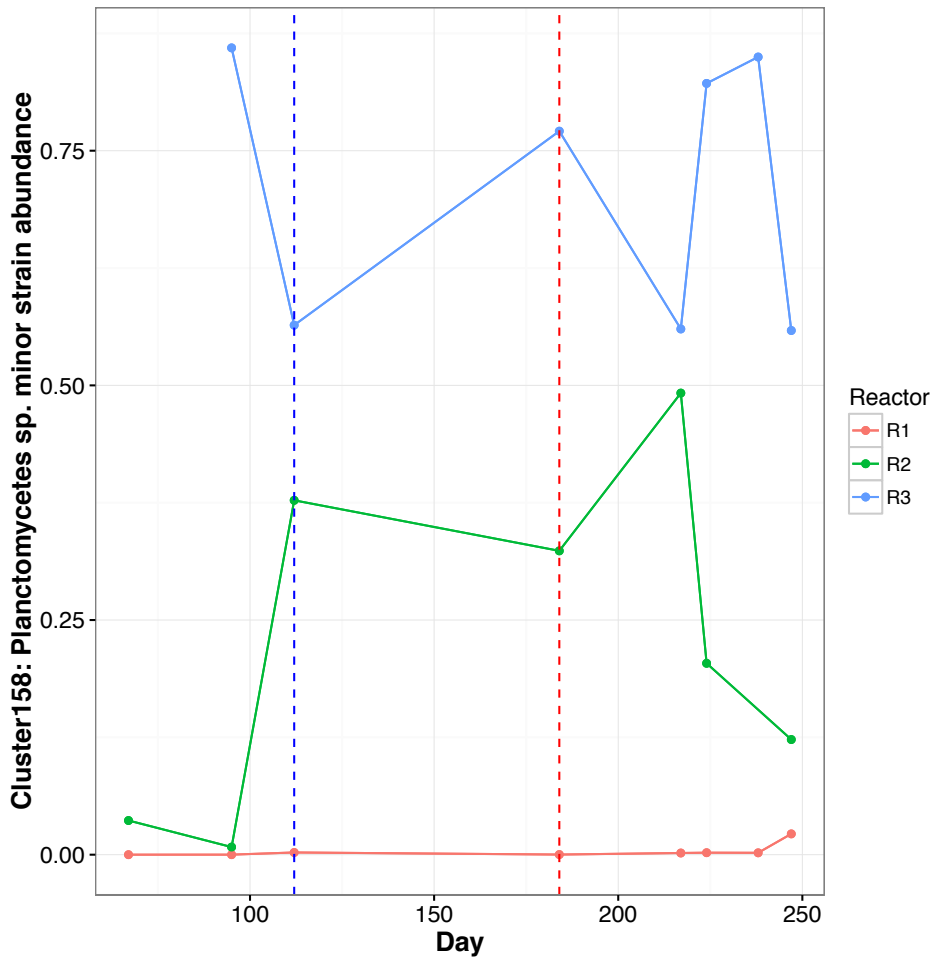
Strain abundance

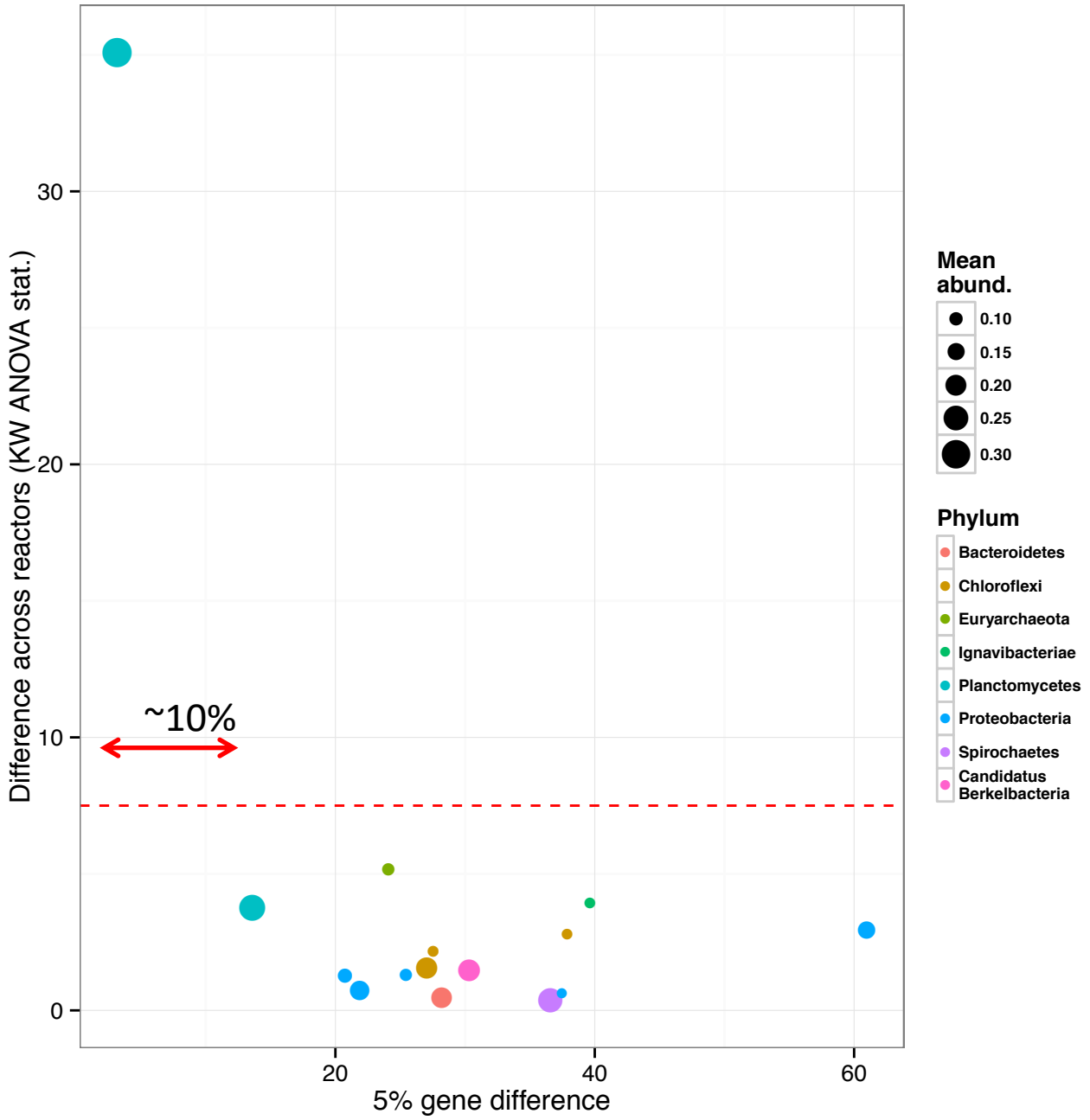


MAG abundance



Strain abundance

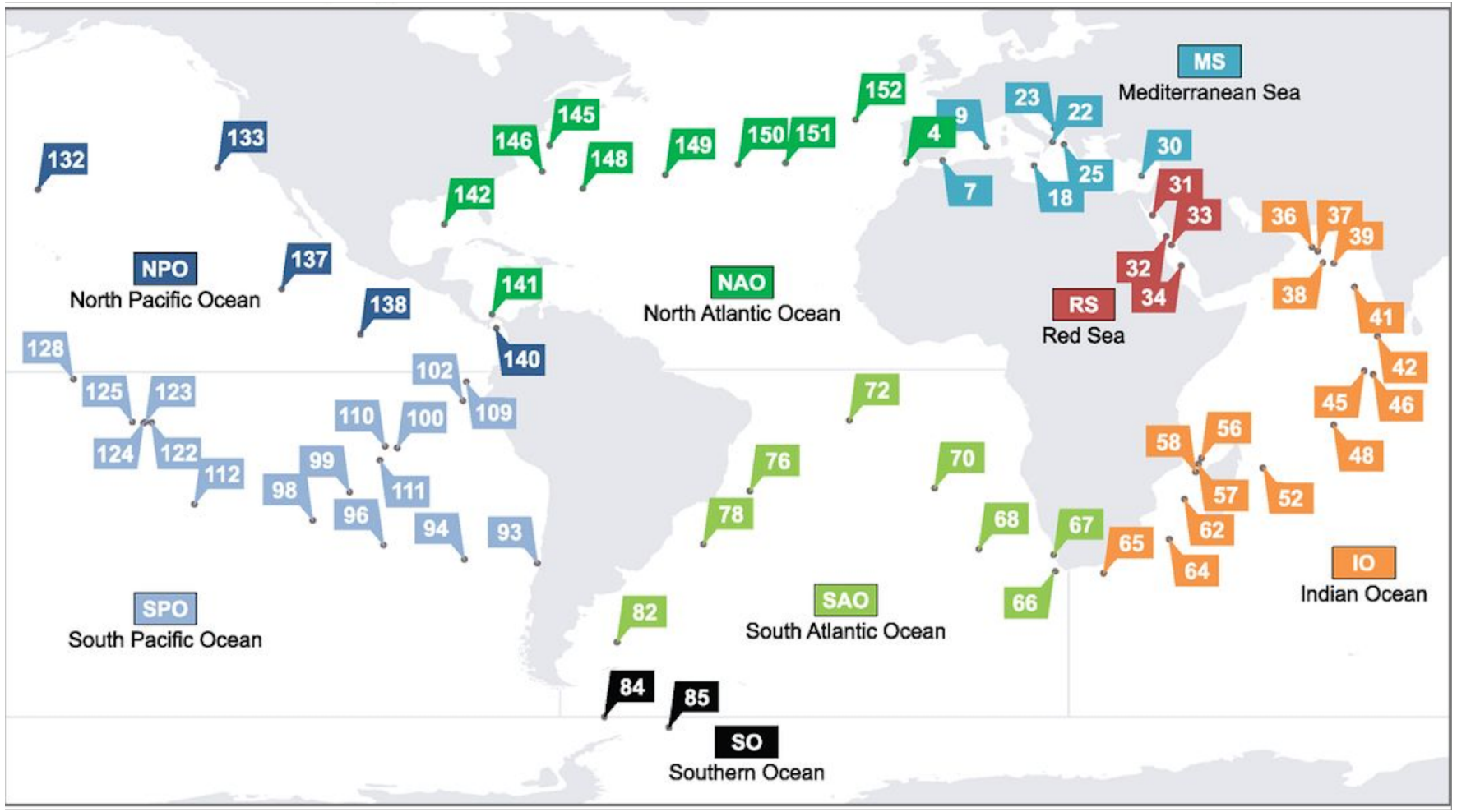




Hypotheses:

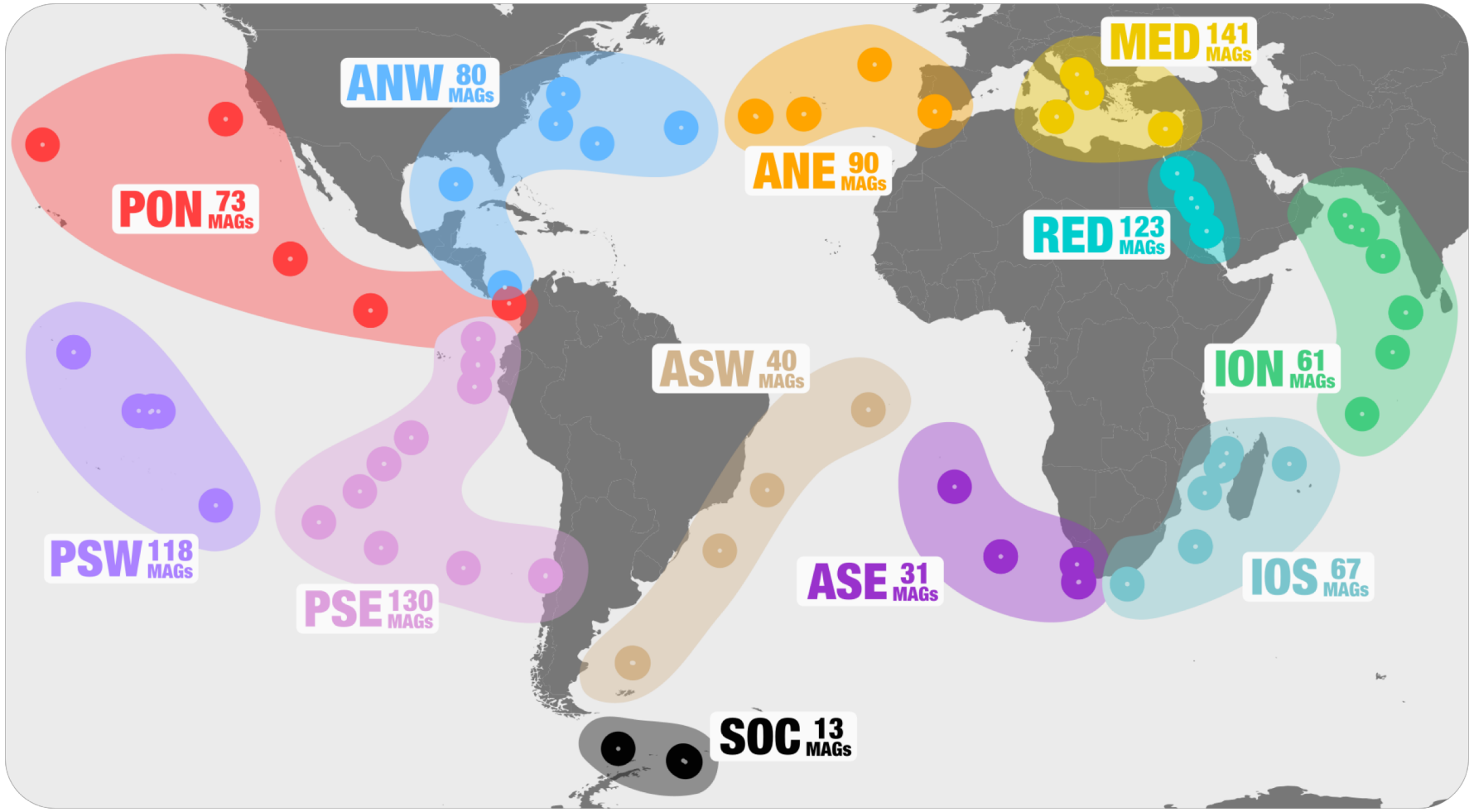
- Wide-spread niche partitioning within microbial species
- Ecological redundancy requires strains have similar genomes

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



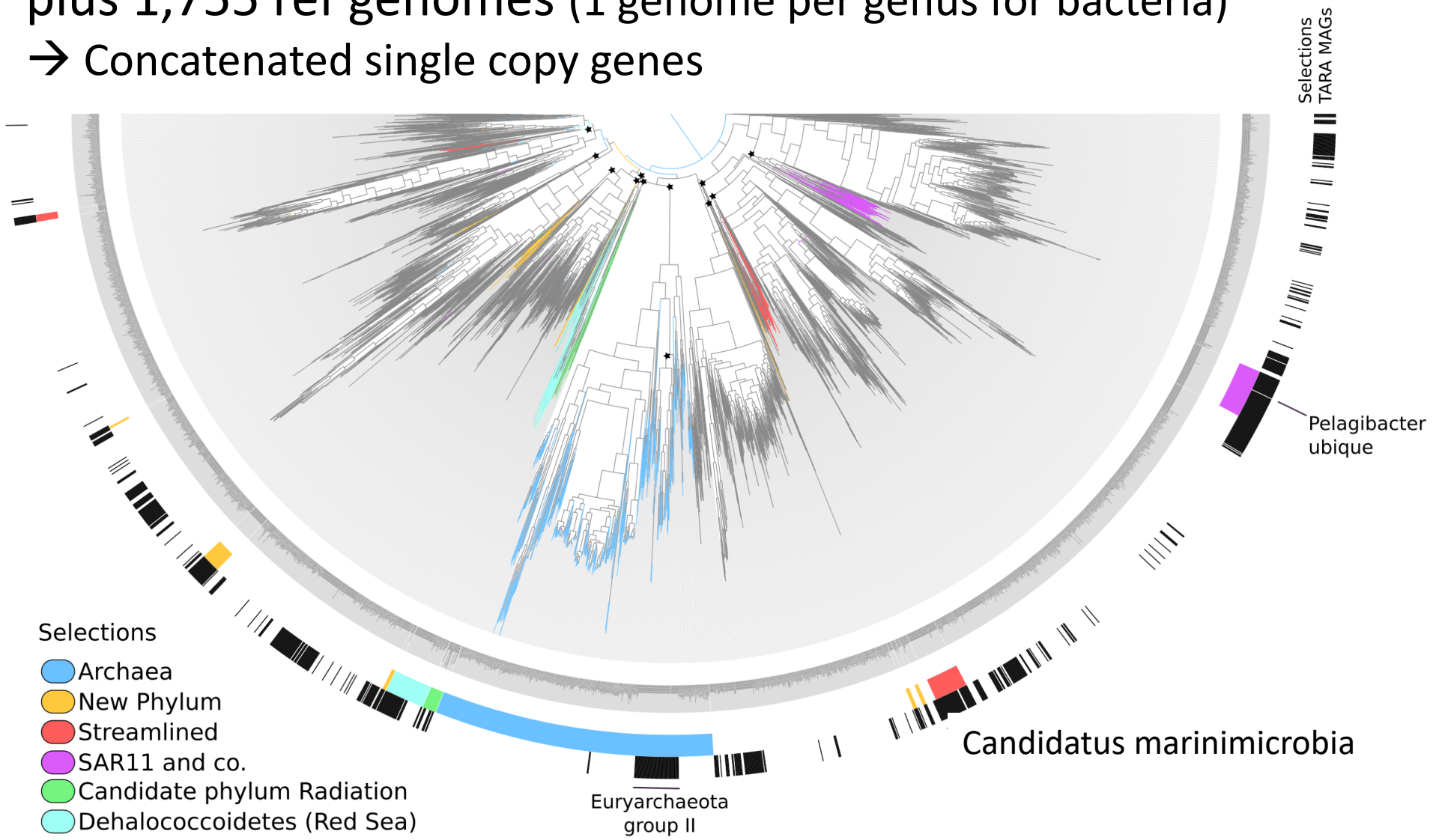
- 93 metagenomes from the planktonic size fraction for which we performed 12 metagenomic co-assemblies
- Generated 957 non-redundant MAGs encompassing the three domains of life

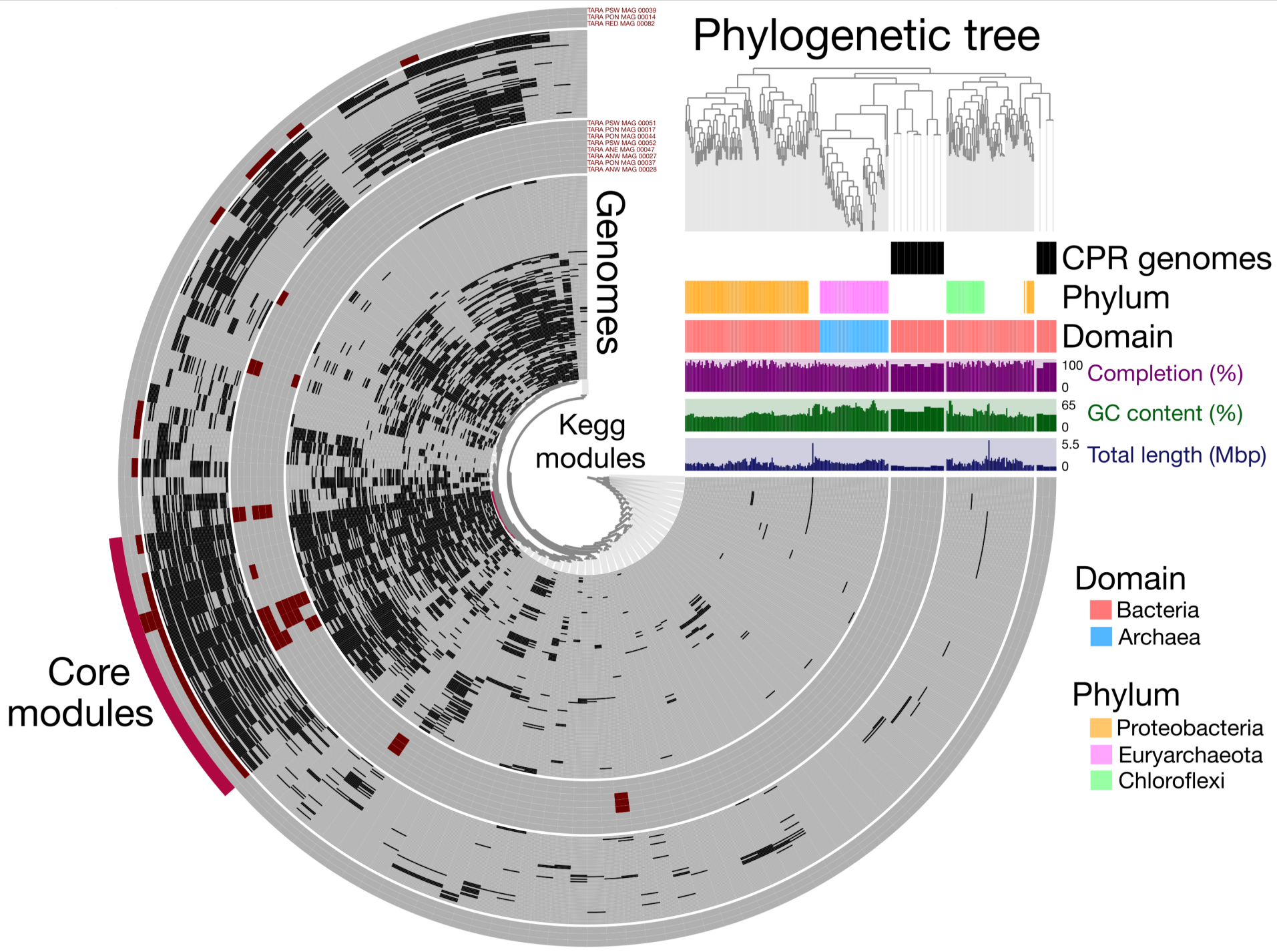
  Delmont, Quince, …, Eren “Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in the surface ocean” Nature Microbiology 2018

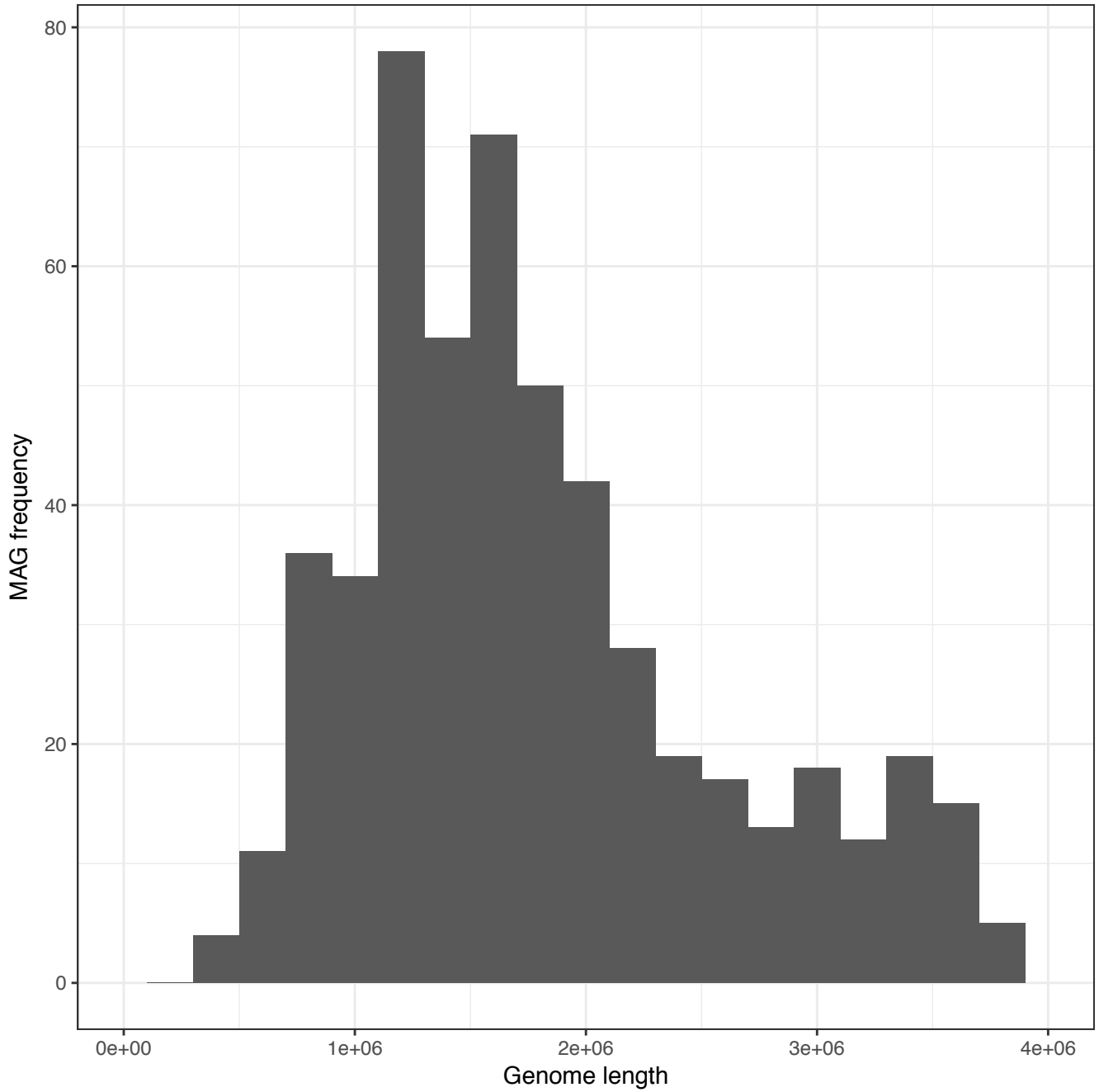
Phylogenetic analysis of 660 MAGs

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

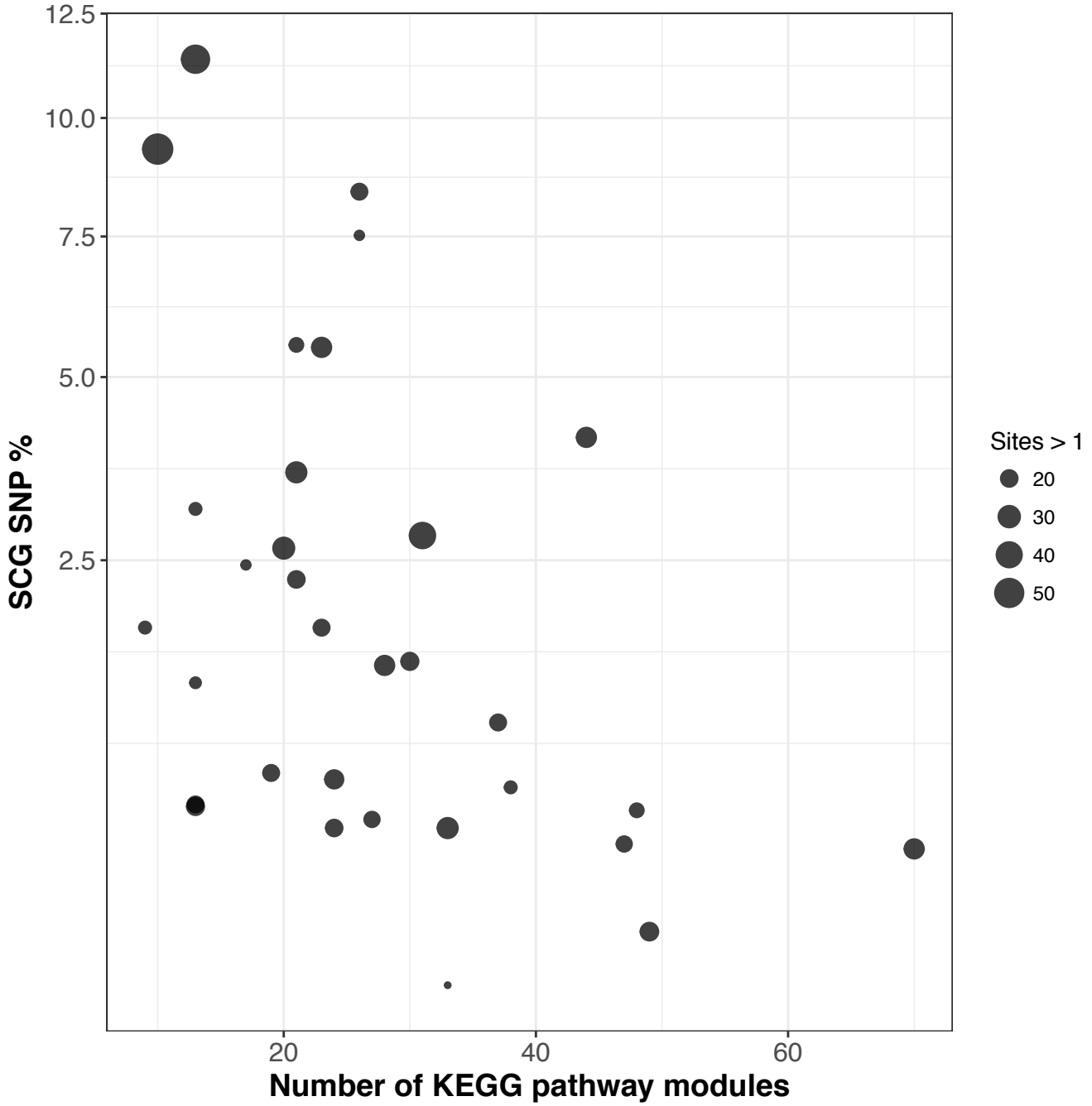
→ Concatenated single copy genes




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



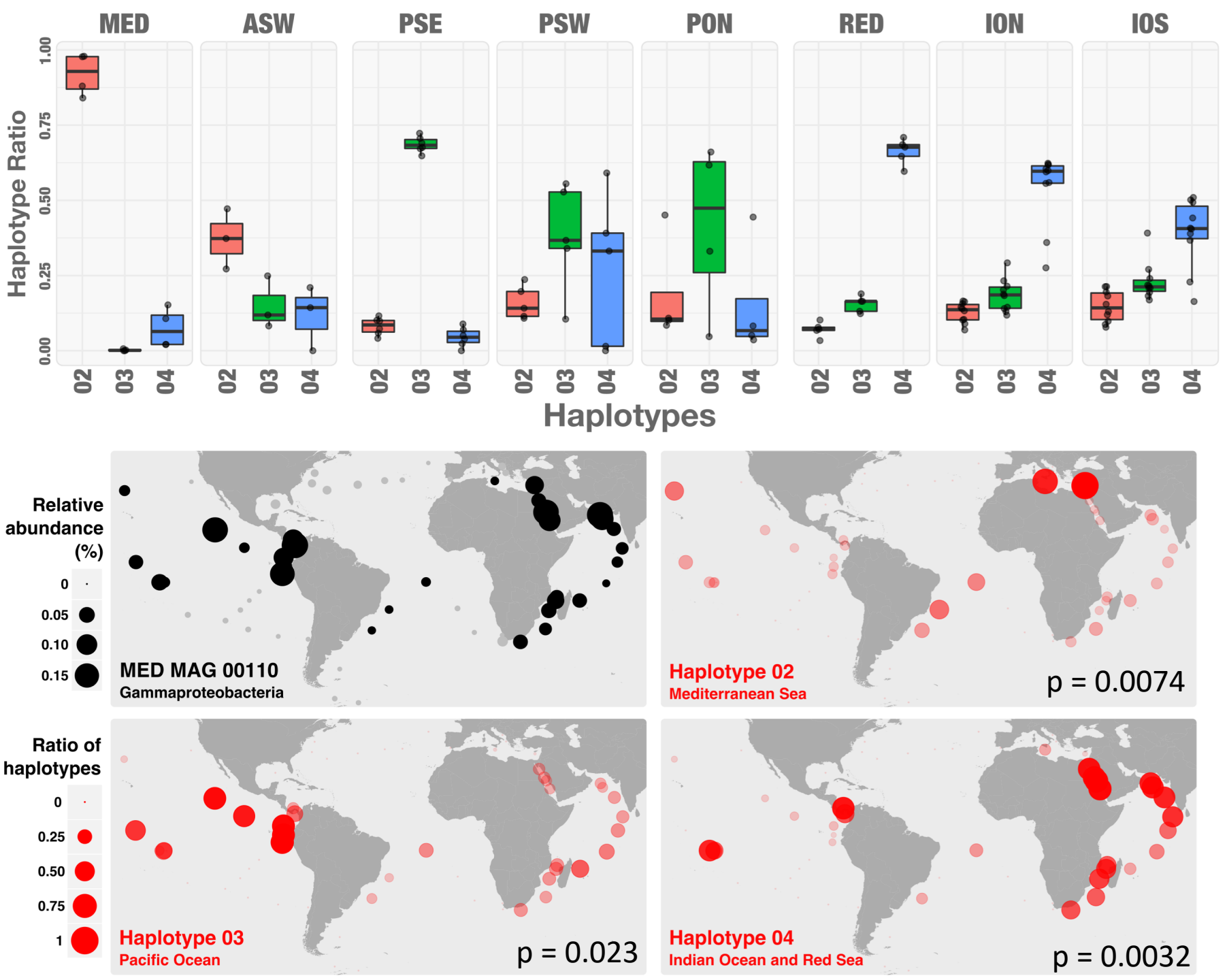
Consider 32 MAGs with cov. > 100.

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

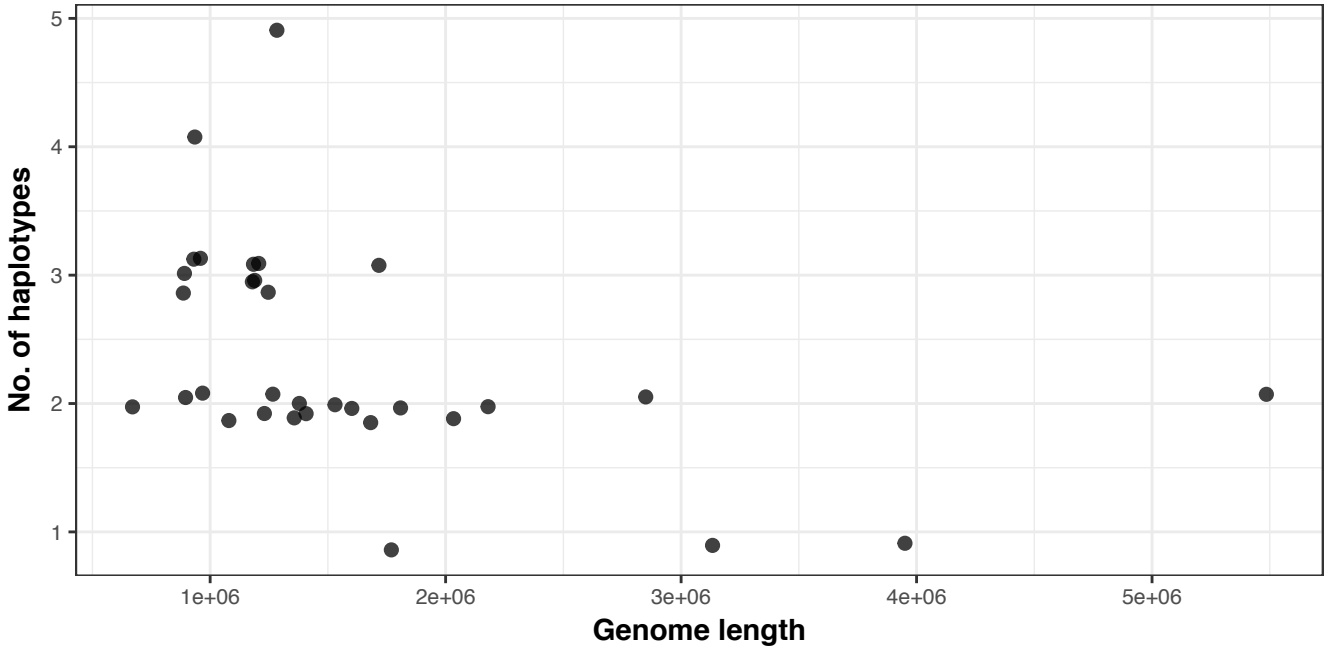
Stronger negative relationship with number of KEGG Pathway modules (Spearman’s p = 0.0045)

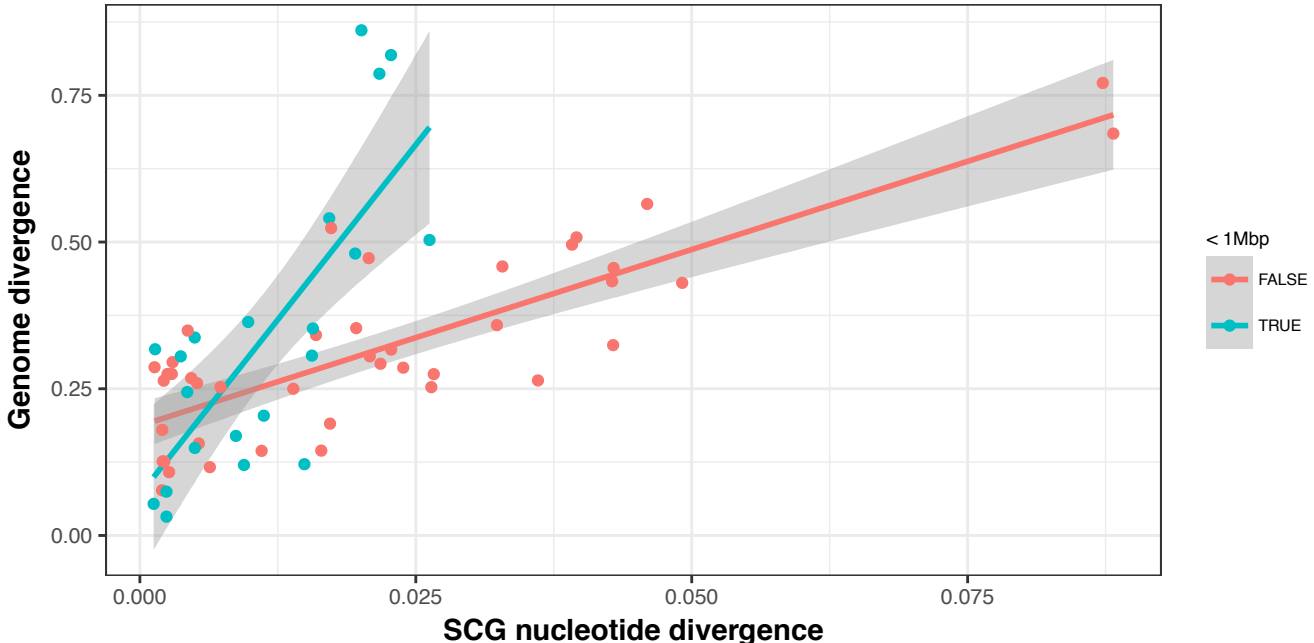
TARA DESMAN analysis

- 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

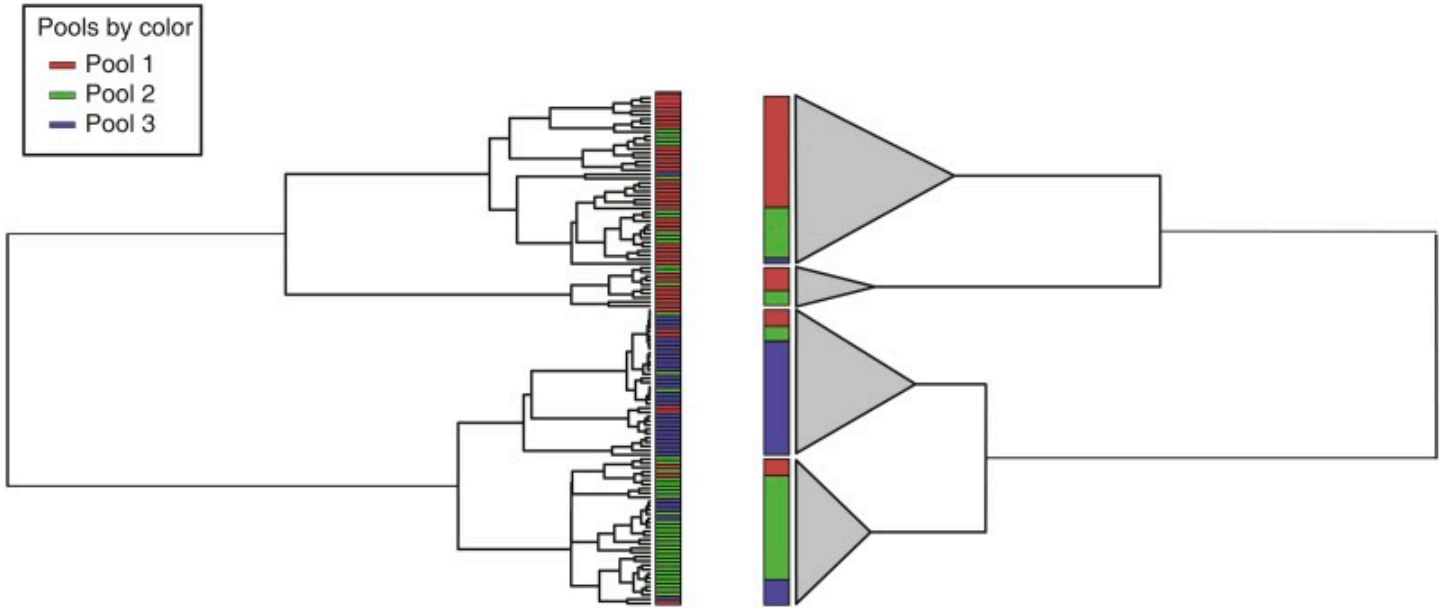




- Significant negative correlation between strain number in MAG and genome length (p = 0.000068)
- 42/73 = 57% of inferred strains had a significant correlation with geographic region
- Significant interaction (p = 3.51e-06) between rate of whole genome divergence relative to core gene divergence and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

What are DESMAN sub-populations?

- Signal resolvable through SNV co-occurrence:
  - Distinct populations within a MAG
  - Finer scale variation if it is structured
  - 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 necessarily to actual haplotypes or genomes…



O' Brien et al. Genetics 2014

Summary

- There will always be limitations on the accuracy of long range variant linkage that can be resolved from metagenomes
- True even when long reads become widely available
- However, it seems that ecologically relevant information is tractable and we can probably do better using graphs

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)