

Genome resolved metagenomics II: De novo Extraction of Strains from MetAgeNomes

Christopher Quince
Warwick Medical School

Strain resolution from short read metagenome data

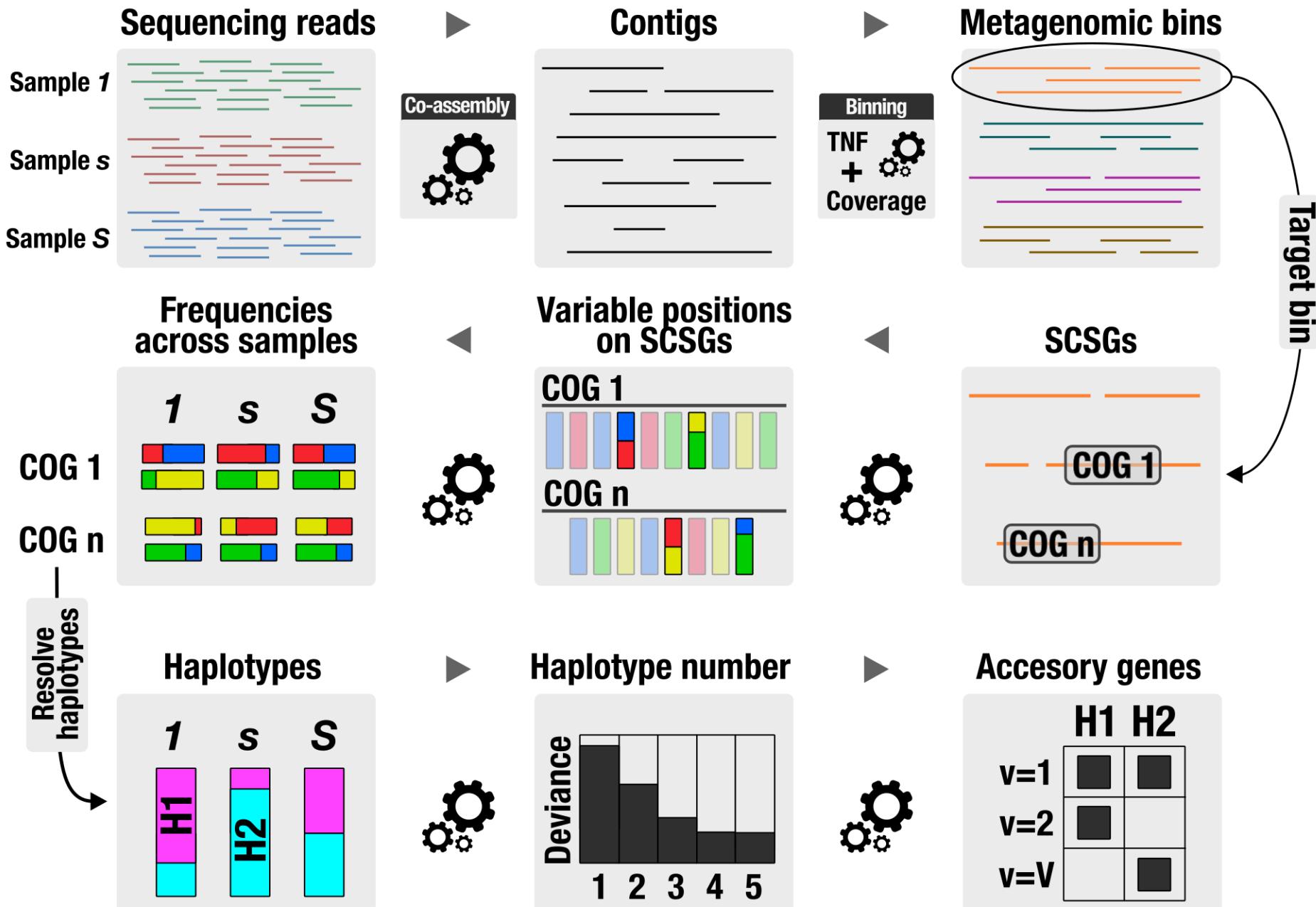
- Co-occurrence binners can cluster contigs into species genomes
- Additional variation exists within species both nucleotide variants on shared genes and variation in the accessory genome
- Can call per sample consensus e.g. MIDAS and StrainPhlAn ([Nayfach et al. Genome Res 2016](#), [Truong et al. Genome Biology 2017](#))
- Can use co-occurrence across multiple samples to resolve strain mixtures *de novo* after mapping to references:
 - Constrains identifies microbial strains in metagenomic datasets ([Luo et al. Nature Biotech. 2015](#))
 - A Bayesian Approach to Inferring the Phylogenetic Structure of Communities from Metagenomic Data ([O'Brien et al. Genetics 2014](#))
- Can we do this entirely *de novo* using contigs and MAGs?

DESMAN: De novo Extraction of Strains from MetAgeNomes

- **DESMAN** <https://github.com/chrisquince/DESMAN>:
 - 1) Map reads onto core genes on contigs
 - 2) Gibbs sampler to infer strain haplotypes and abundances
 - 3) Determine copy count of accessory genome

bioRxiv: <http://biorxiv.org/content/early/2016/09/06/073825>





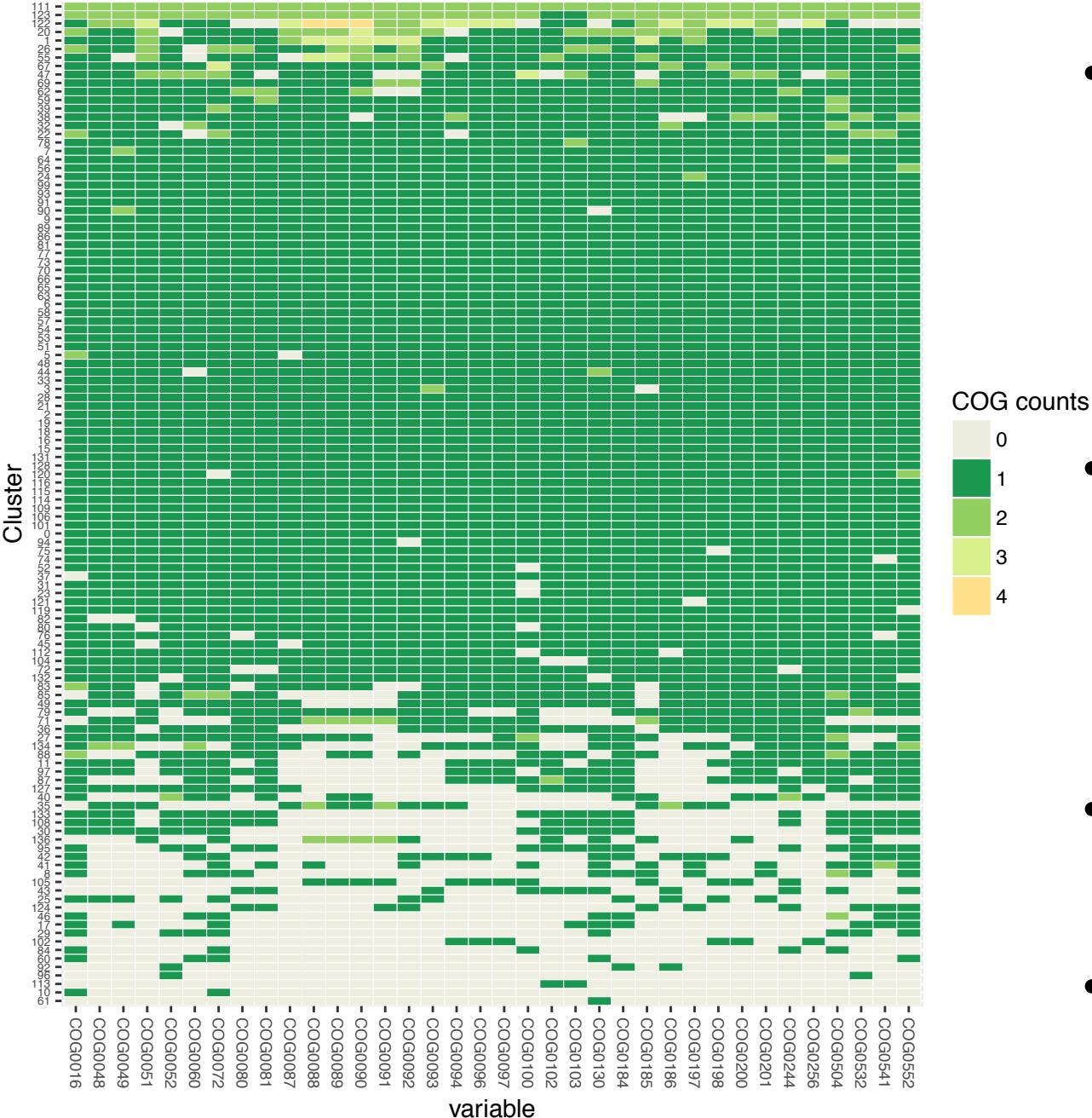
Complex synthetic community

- 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>

Concoct binning

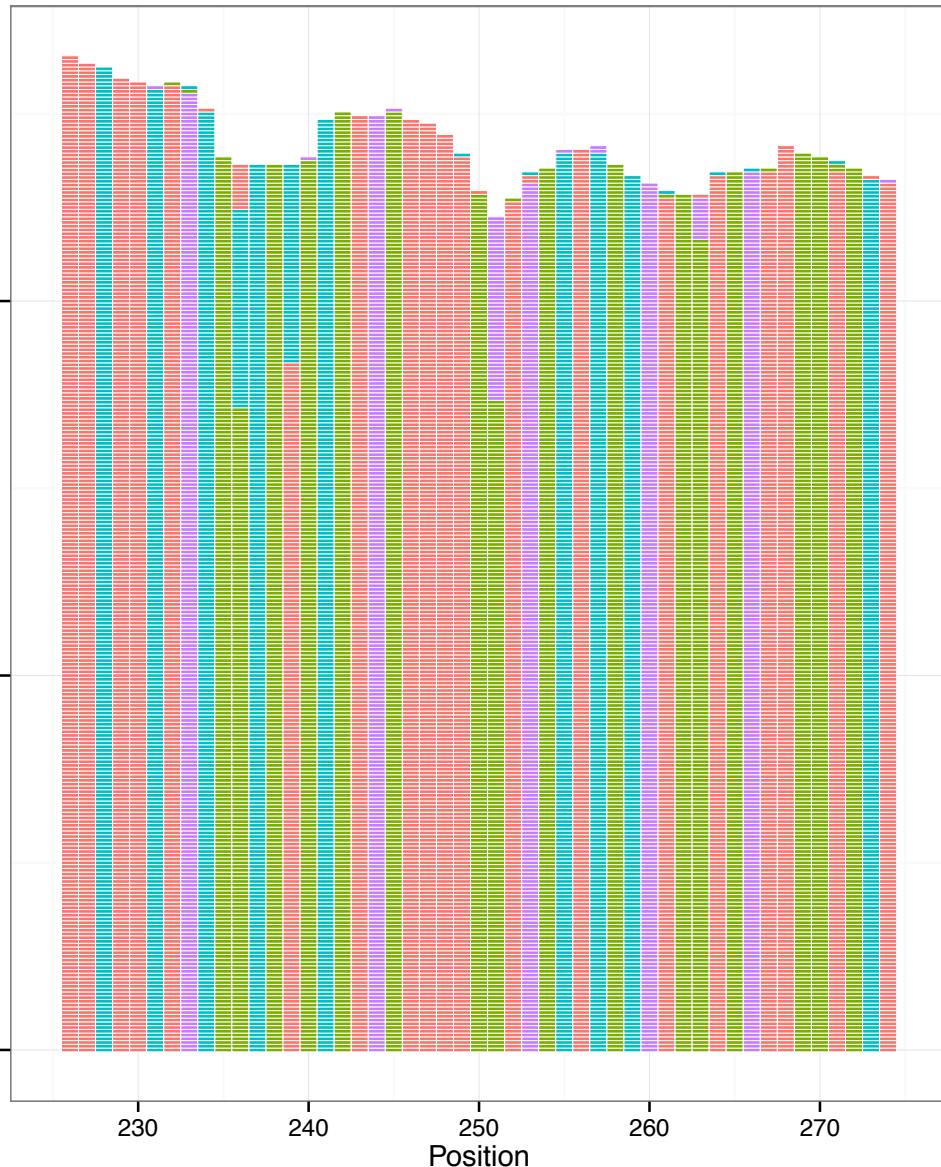
- Coassembly gave 74,580 contig fragments with a total length of 409 Mbp as compared to 687 Mbp for all 210 genomes
- CONCOCT2 generated 137 clusters (species recall of 86.1% and a precision of 98.2%)
- 75 clusters with 75% of SCGs in single copy
- Map clusters onto reference species



DESMAN analysis using single-copy core genes

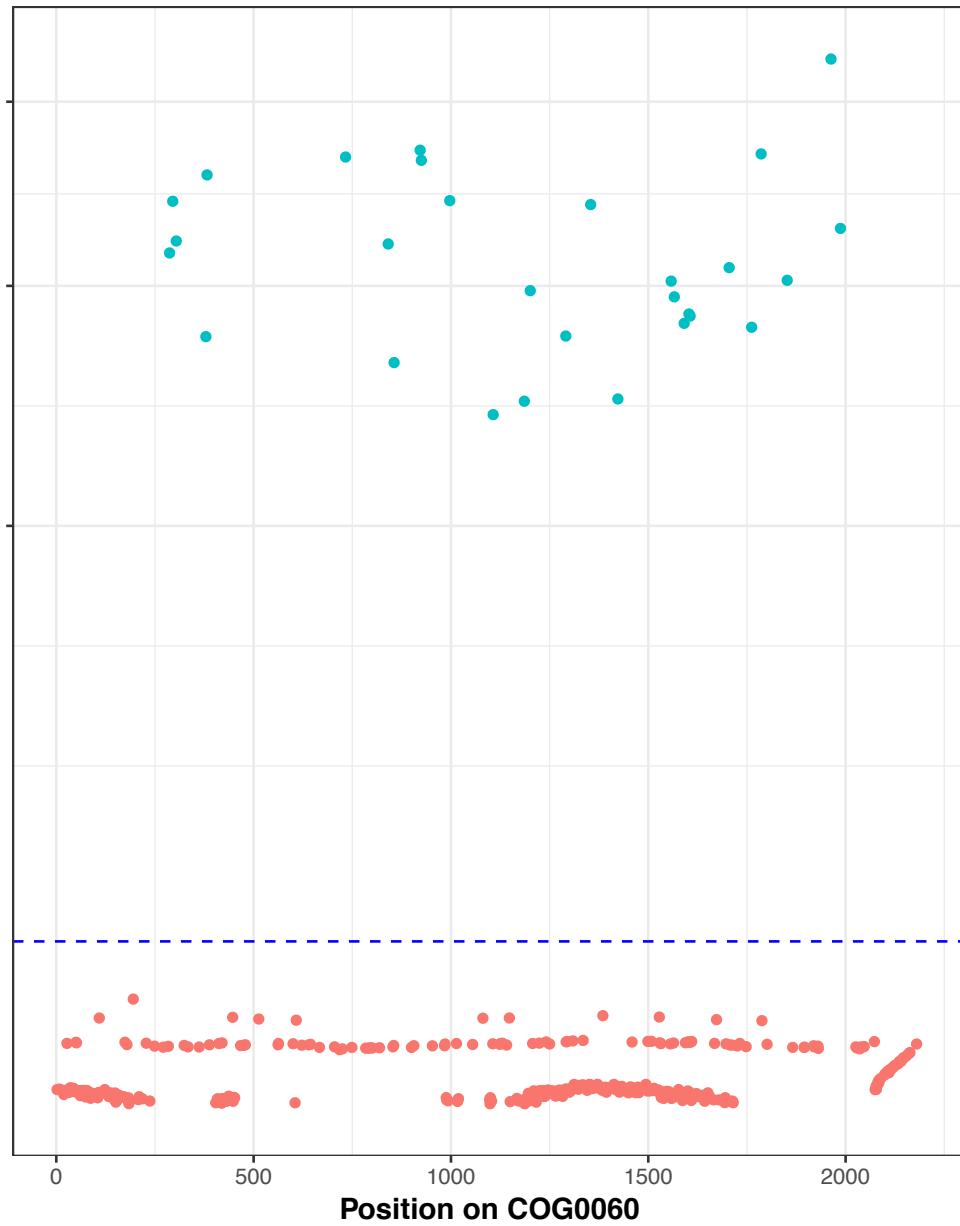
- Process each of the 75 – 75% complete clusters separately
- Map reads onto 36 SCGs
- If target taxa is known and cultured custom set of species specific single-copy core genes (SSCGs) could be used instead

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:
$$\mathcal{E}_{a,b}$$
- 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

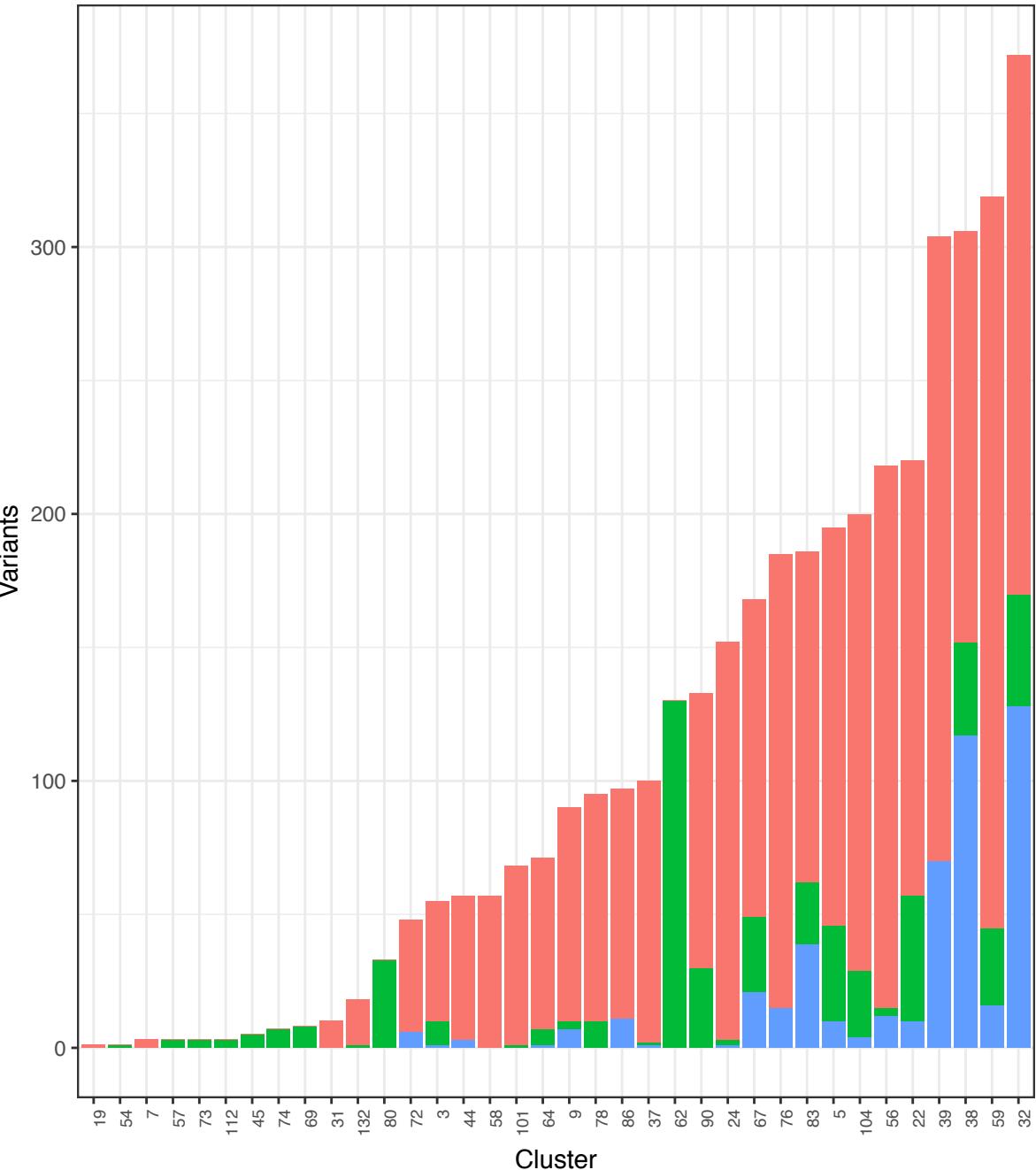
Over all 27 SCGs and 26015 positions with $q < 0.001$

Predicted

	False	True
False	25015	1
True	1	98

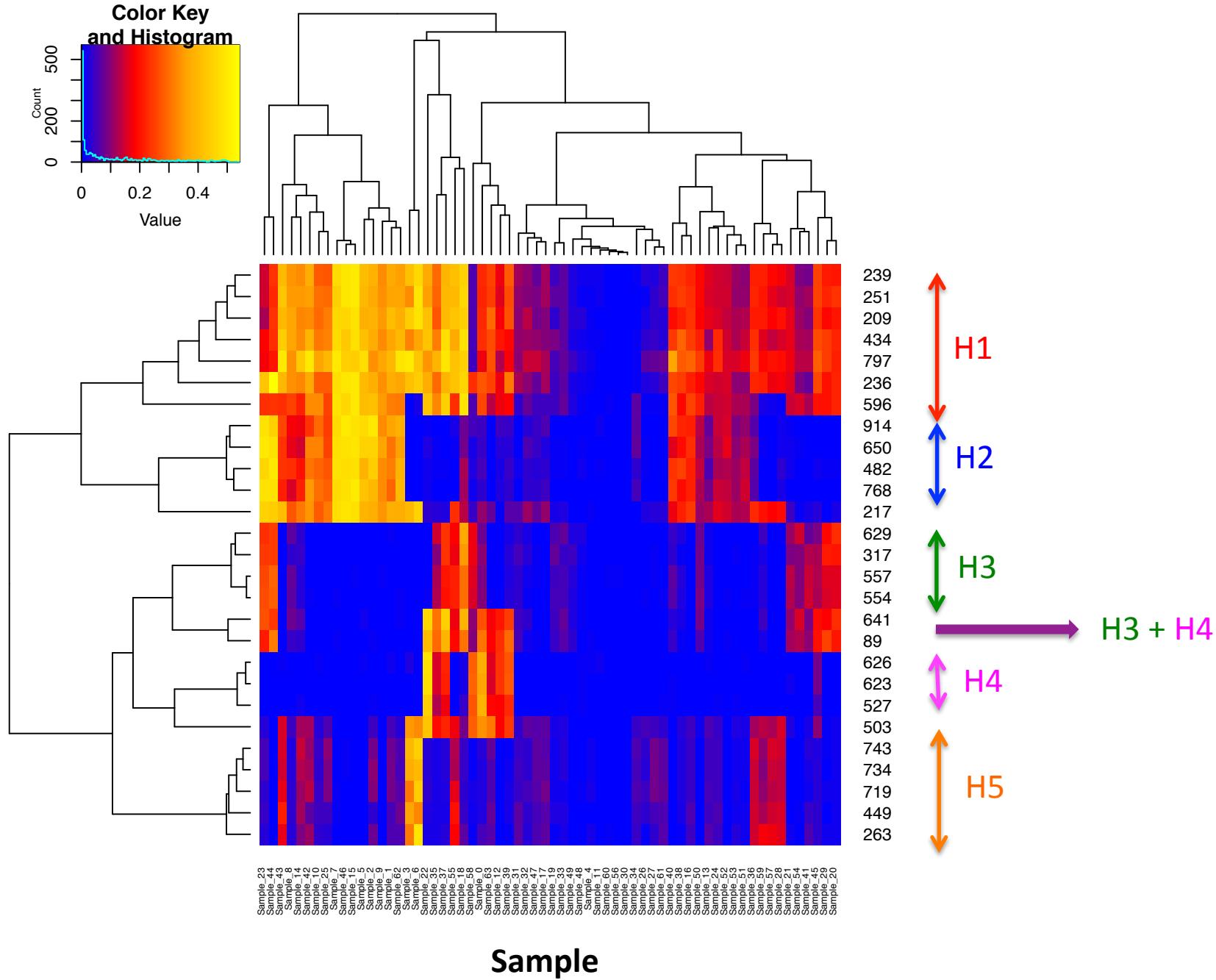
Recall = 98.9%, Precision = 98.9%

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)

Linking variants by co-occurrence



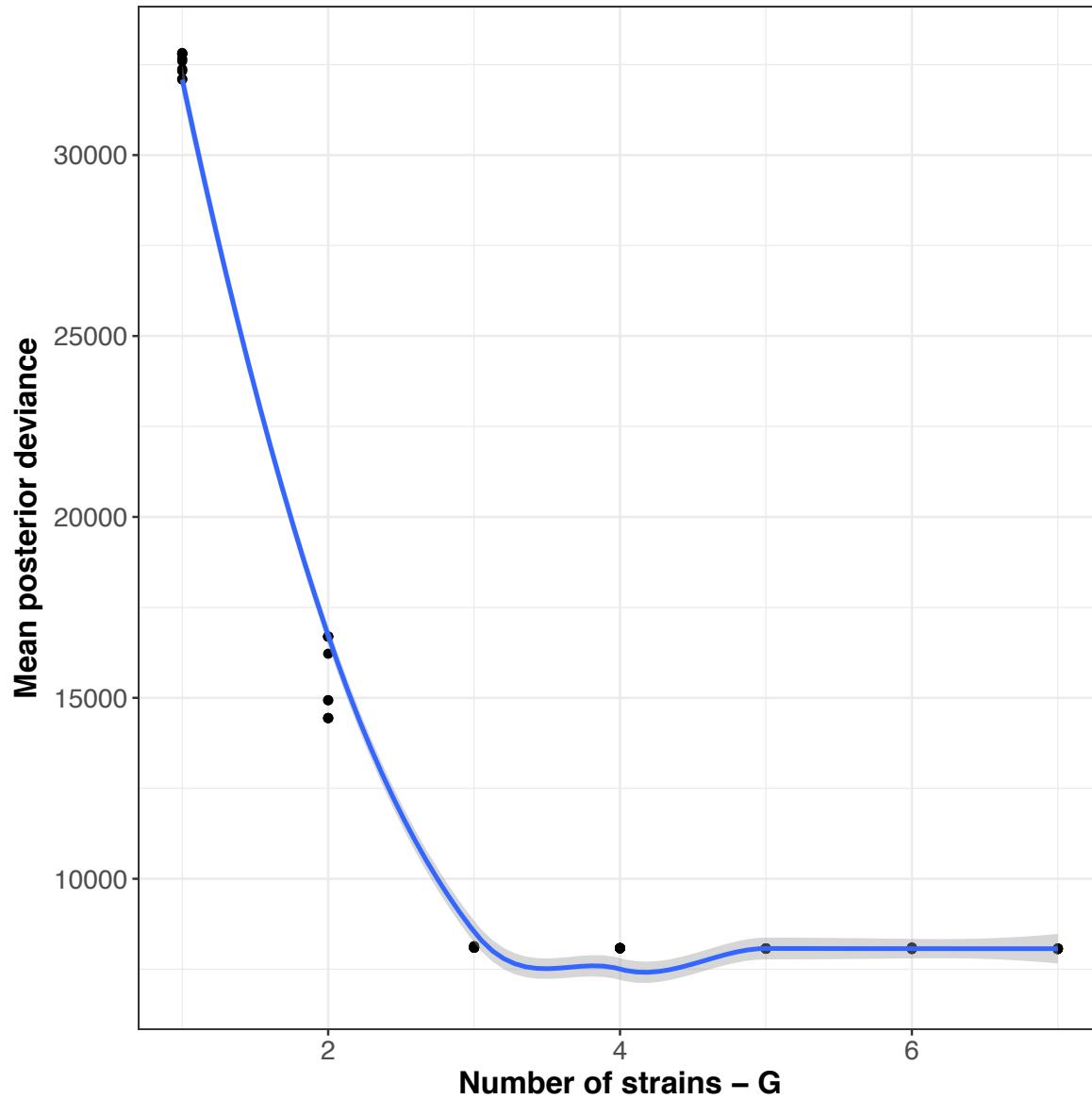
Resolving haplotypes and their abundance (intuitive)

- Assume that each haplotype has a characteristic frequency profile across samples
- Cannot cluster variants into haplotypes since clustering assumes that each sample derives from a single cluster
- Instead the variant frequencies at a given position is a summation of underlying haplotype profiles

Gibbs sampling algorithm

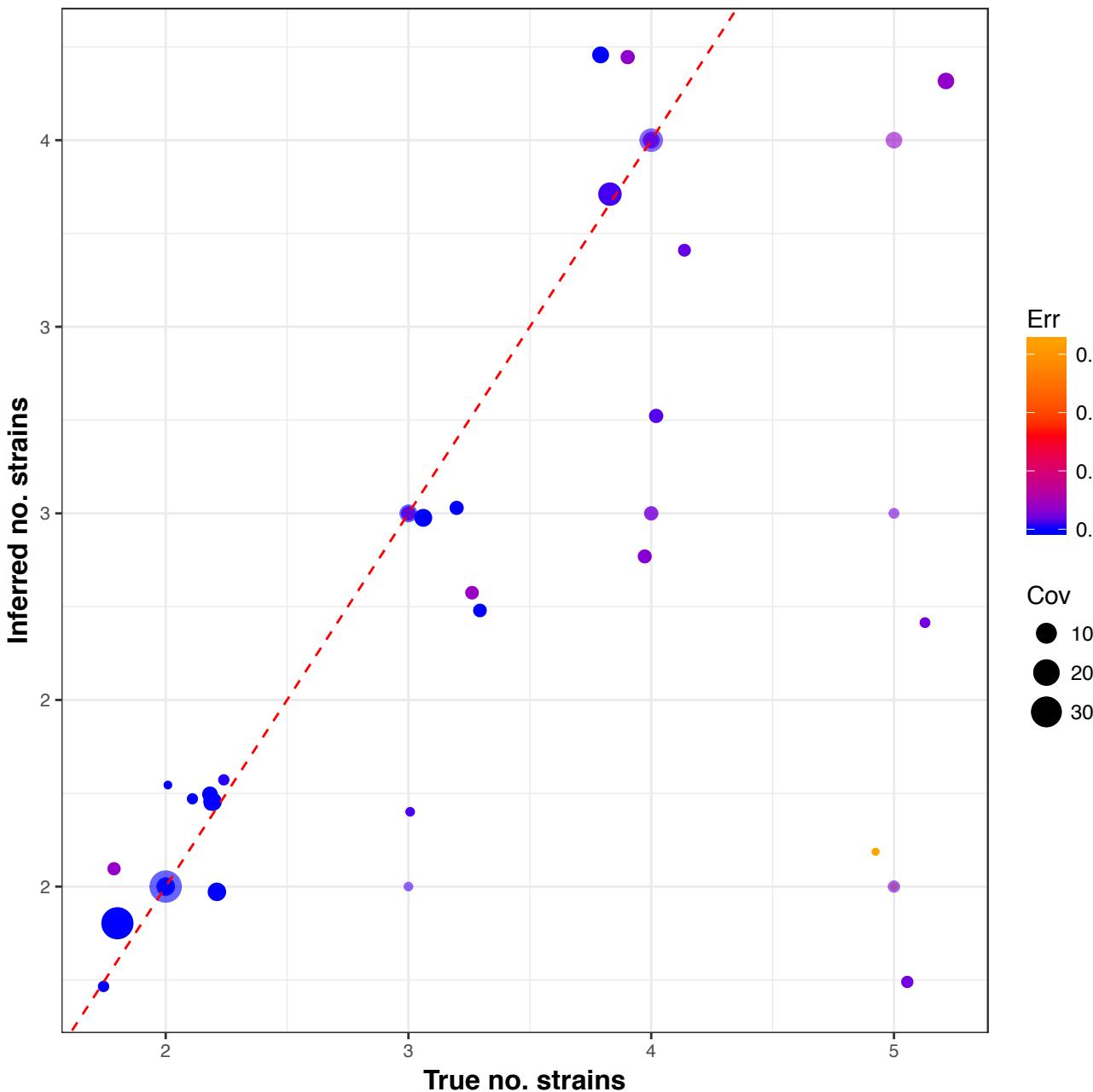
- Assume bases are independent leads to binomial probability at each position
- Devise an iterative algorithm that successively samples:
 - Haplotypes
 - Relative frequencies
 - Error rates
- Bayesian algorithm generates distribution of fitted values
- Negative log-likelihood describes overall fit
-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

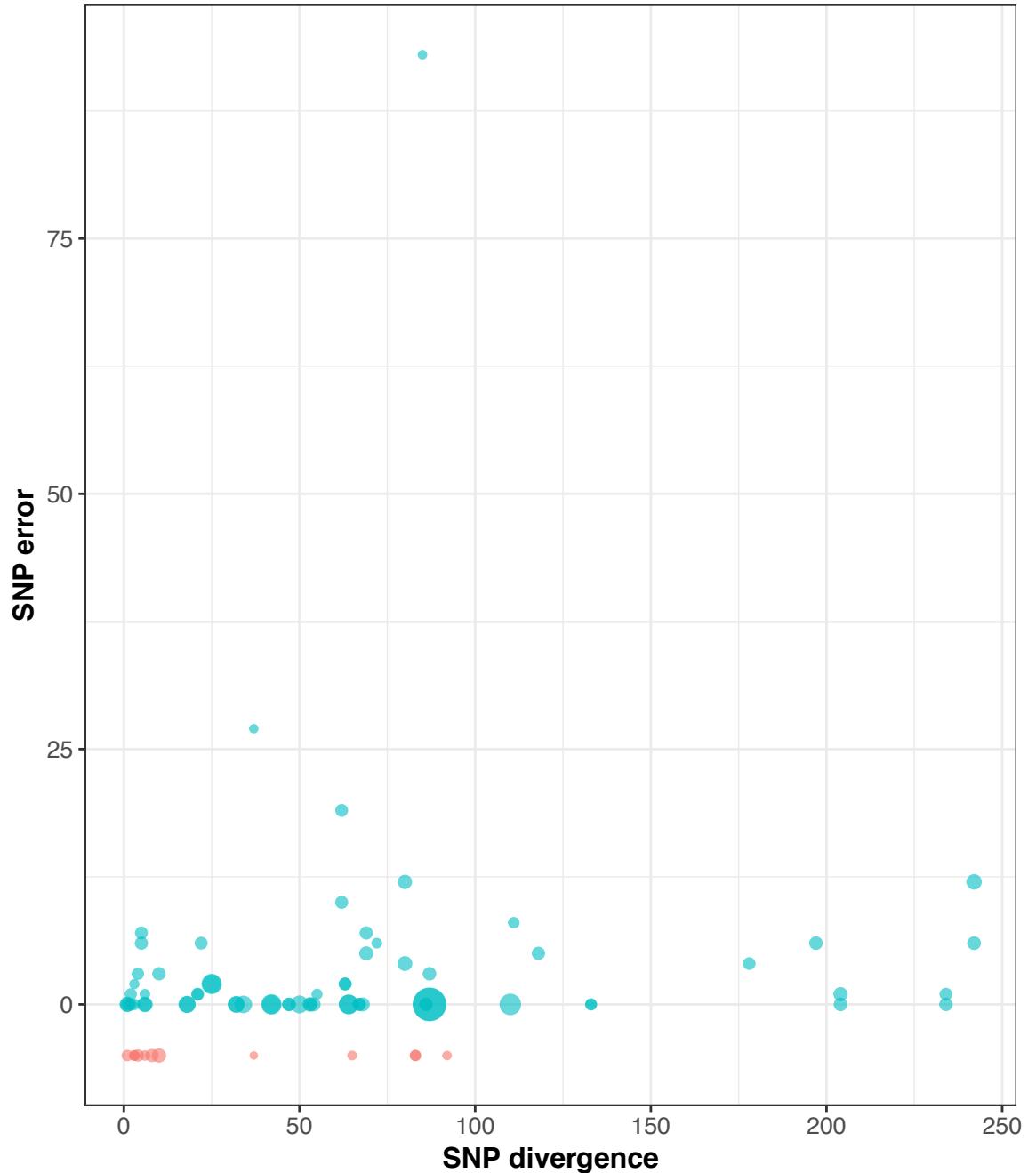


Model fit fails to improve after 3 haplotypes

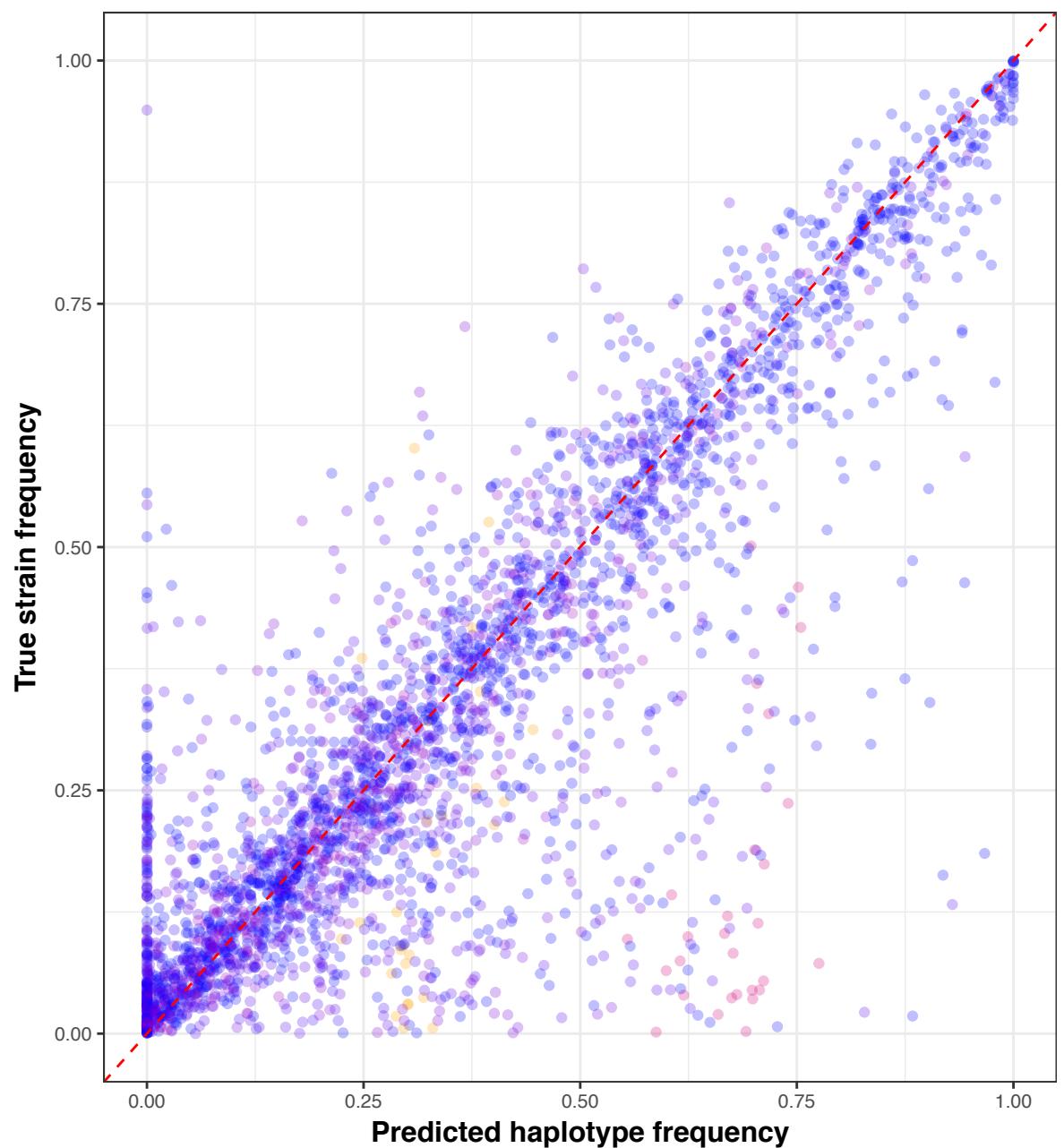
The best fit 3 haplotype run had each haplotype map onto a different reference strain with no errors



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

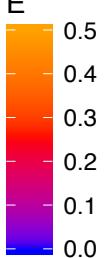


- Mean SNV error rate below 1% for 15/25 (60%) of clusters
 - Median of 0.25% and a mean of 2.38%
 - No correlation between error rate and either the number of variants in the cluster or coverage
 - We find a positive relationship between detection (67 out of 79 detected) and individual strain coverage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



Linear regression: slope
0.820, adjusted R-
squared 0.741, p-value:
 $< 2.2\text{e-}16$

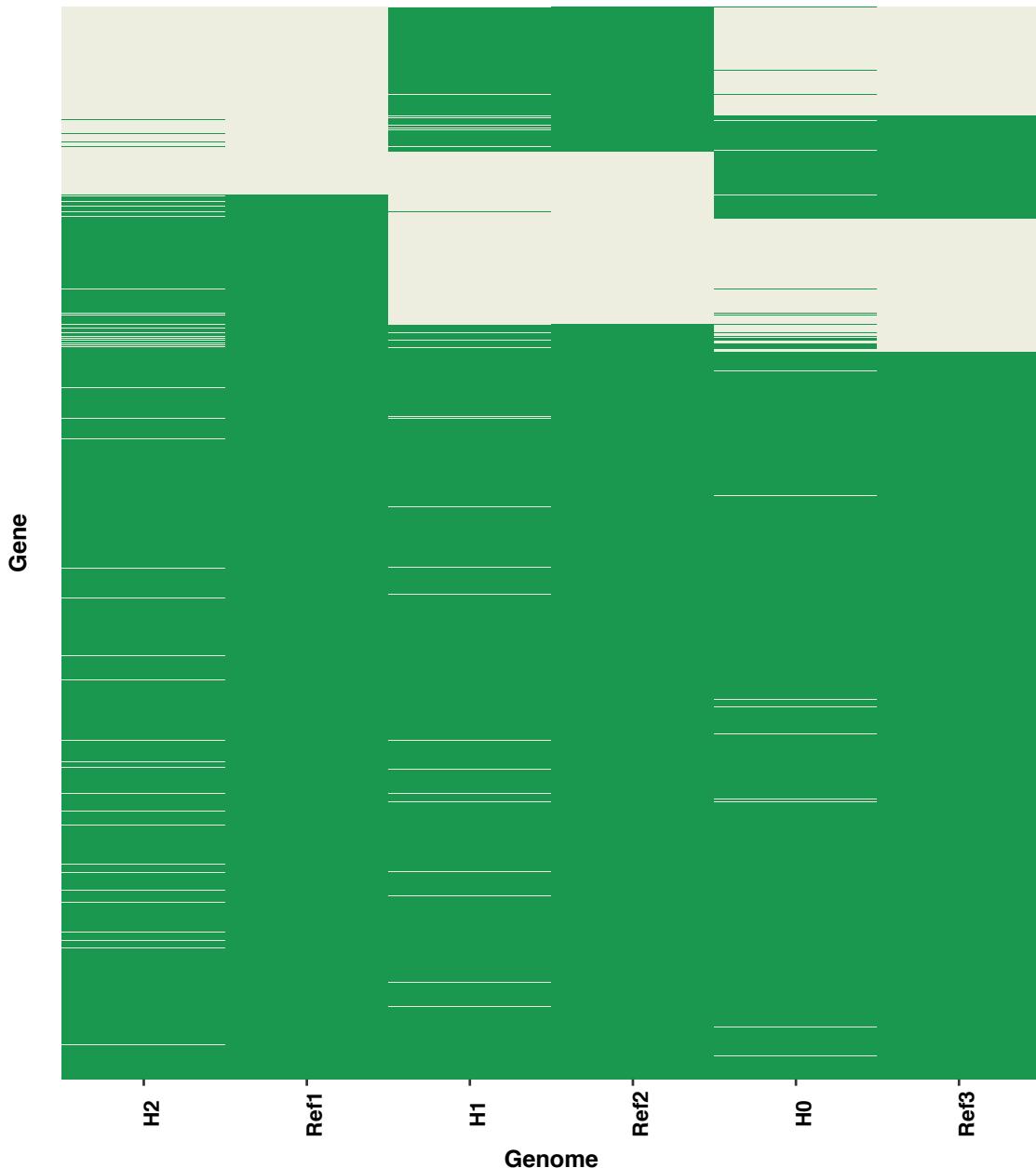
Haplotypes with $E < 0.01$, slope 0.853,
adjusted R-squared
0.810, p-value: $< 2.2\text{e-}16$



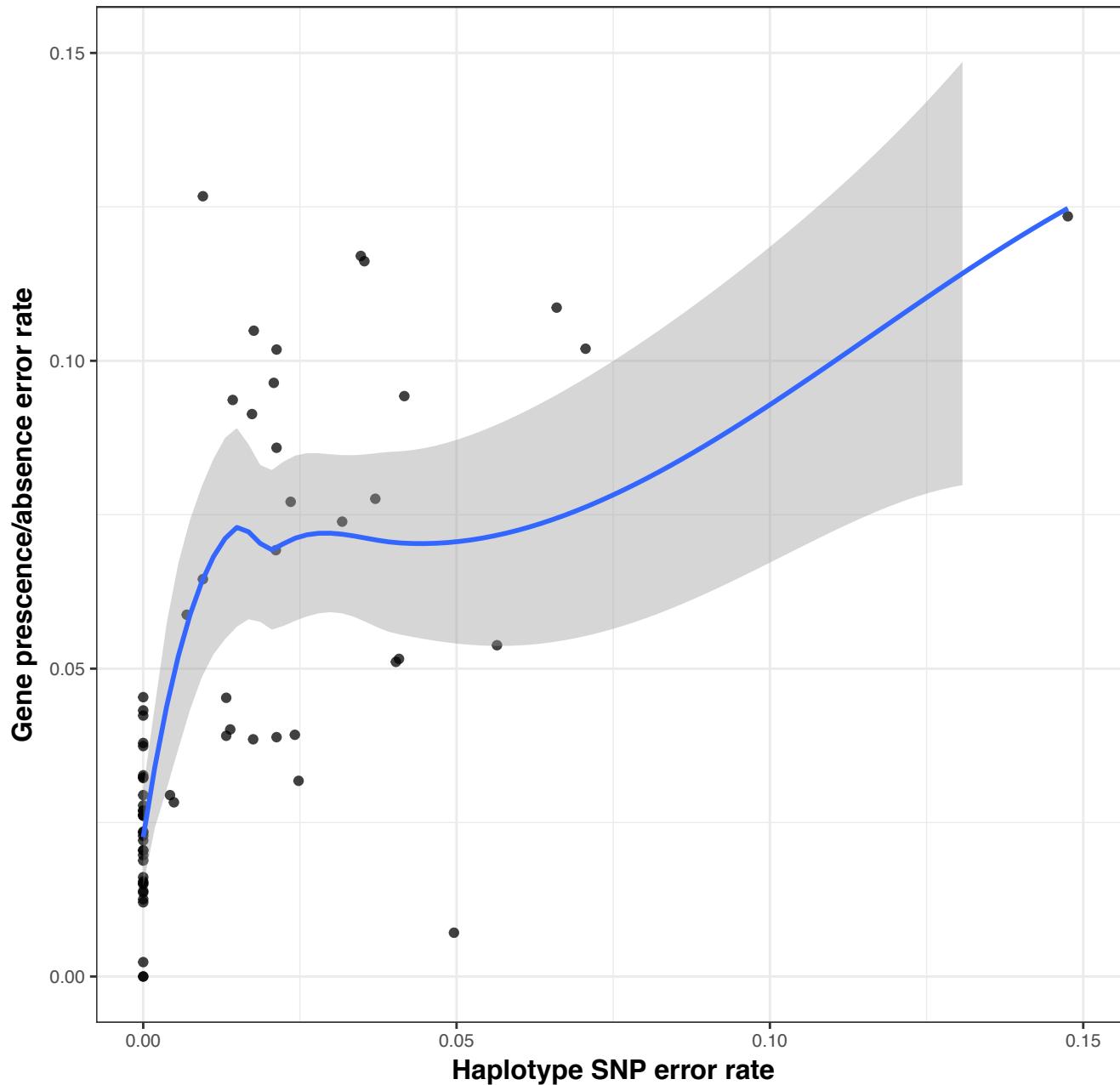
Performance of Lineage (O'Brien et al. Genetics 2014)

- On the complex mock community applied to SCGs DESMAN consistently 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
- Constraints 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%

Expanded Granular Sludge-Bed Laboratory Bioreactors (EGSB)

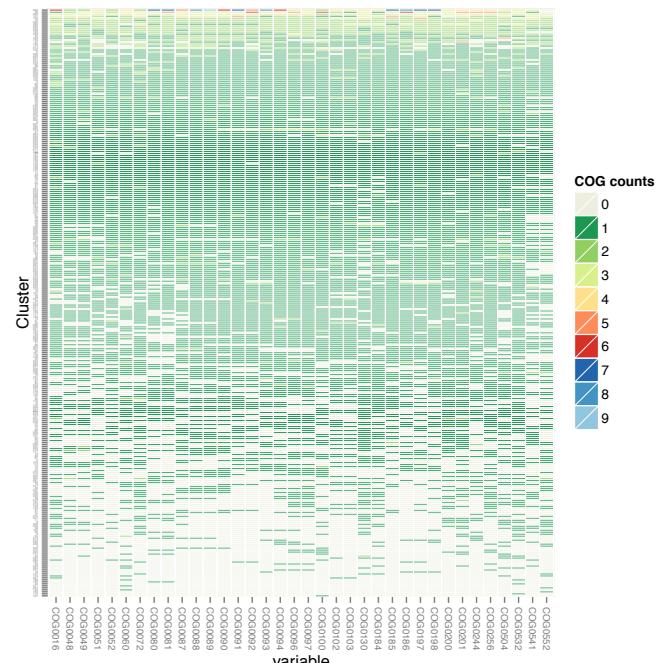
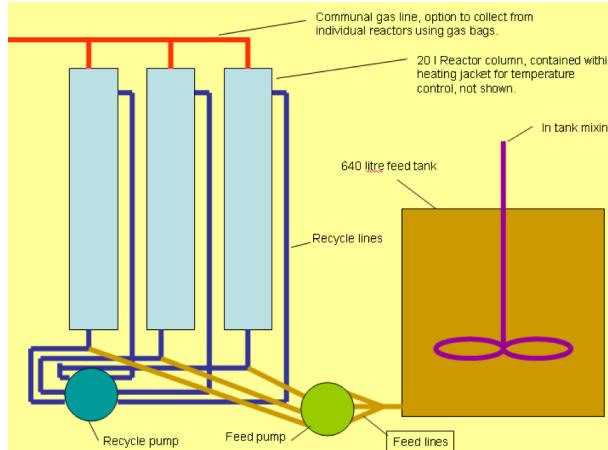


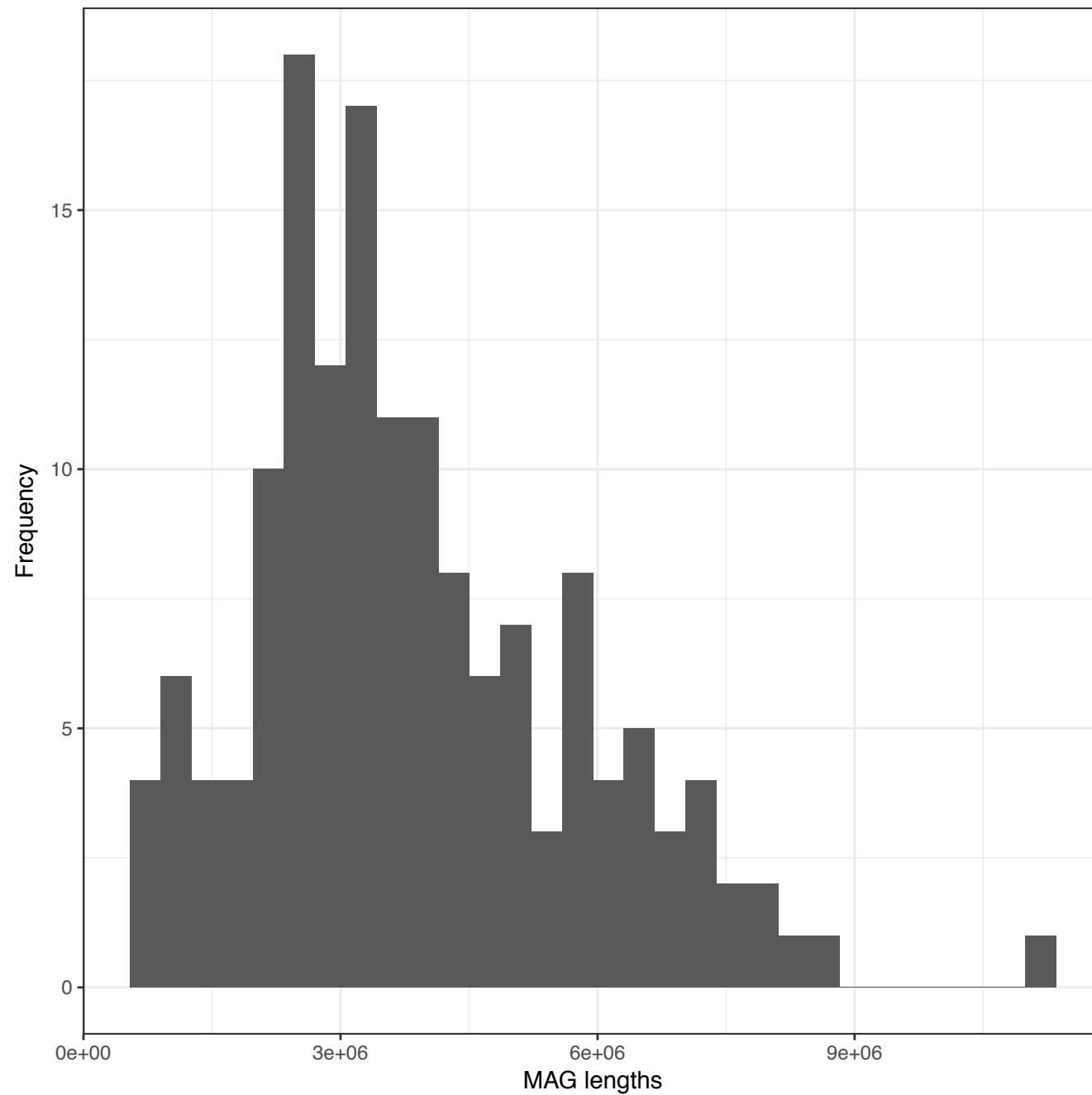
NUI Galway
OÉ Gaillimh

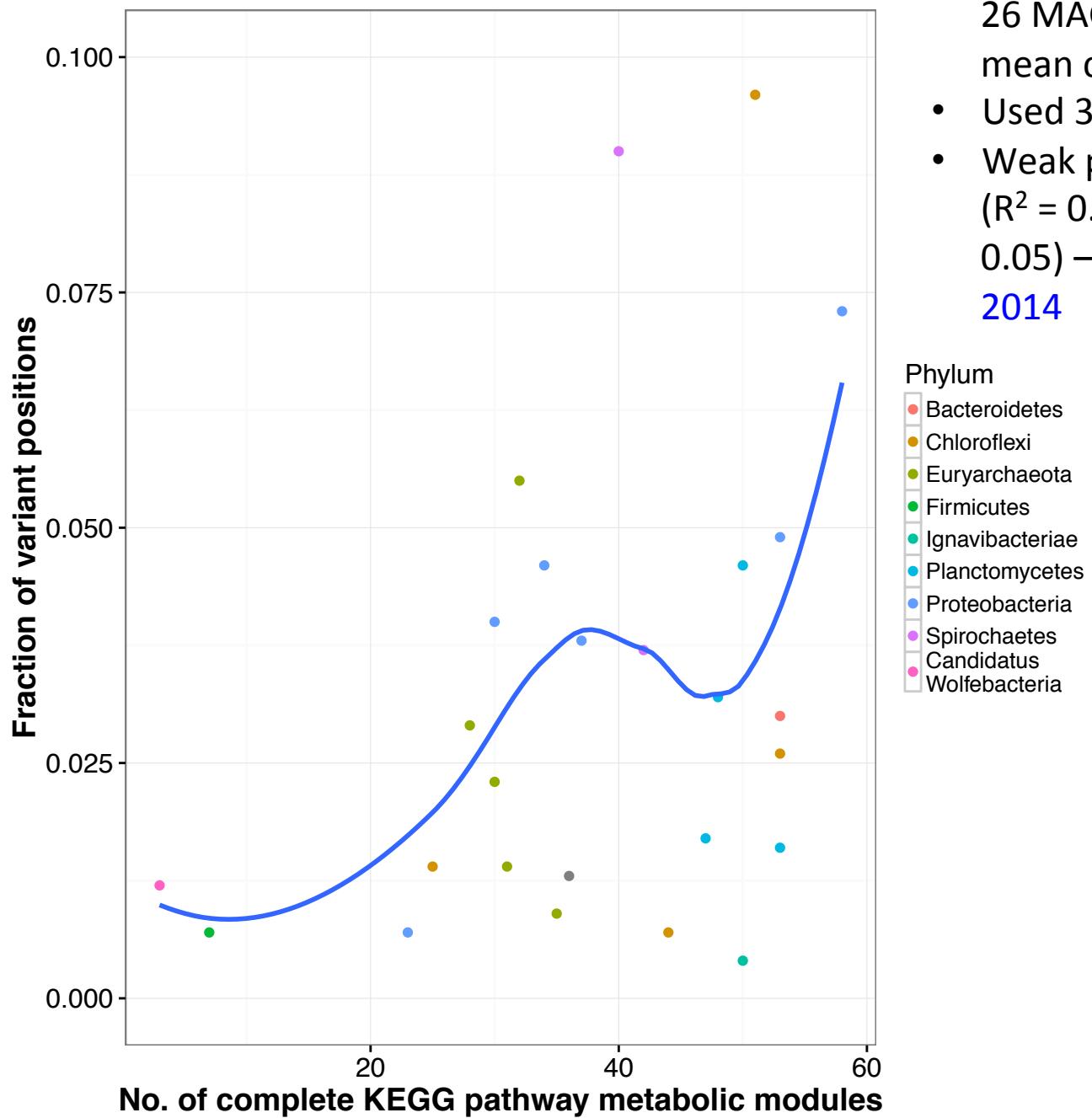


University
of Glasgow

- 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
- 355 CONCOCT clusters, 152/153 – 70% pure and complete





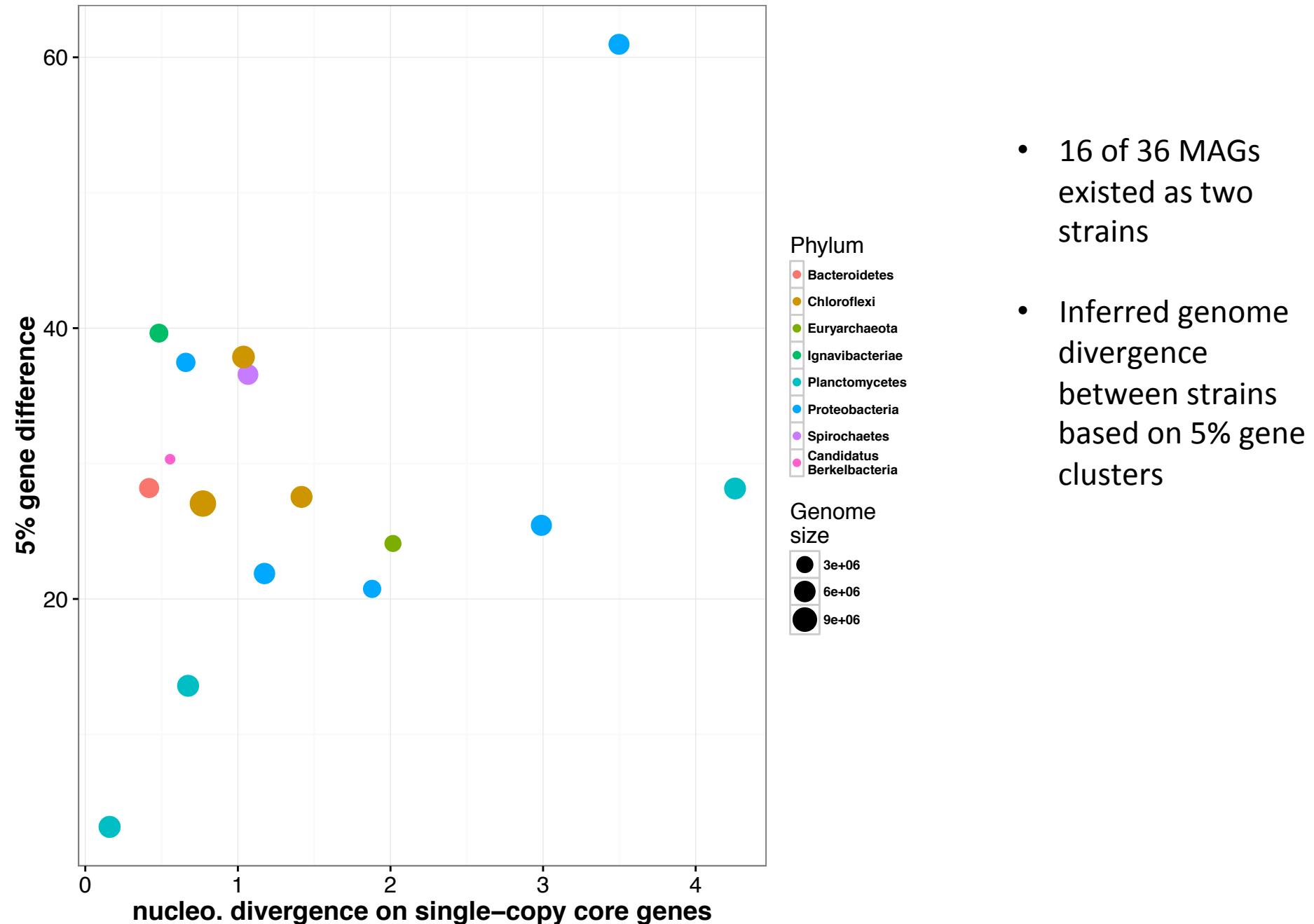


- Applied strain resolution to 26 MAGs with biomass mean cov. > 75.
- Used 36 SCGs
- Weak positive relationship ($R^2 = 0.1132$, p-value = 0.05) – c.f. [Muller et al. 2014](#)

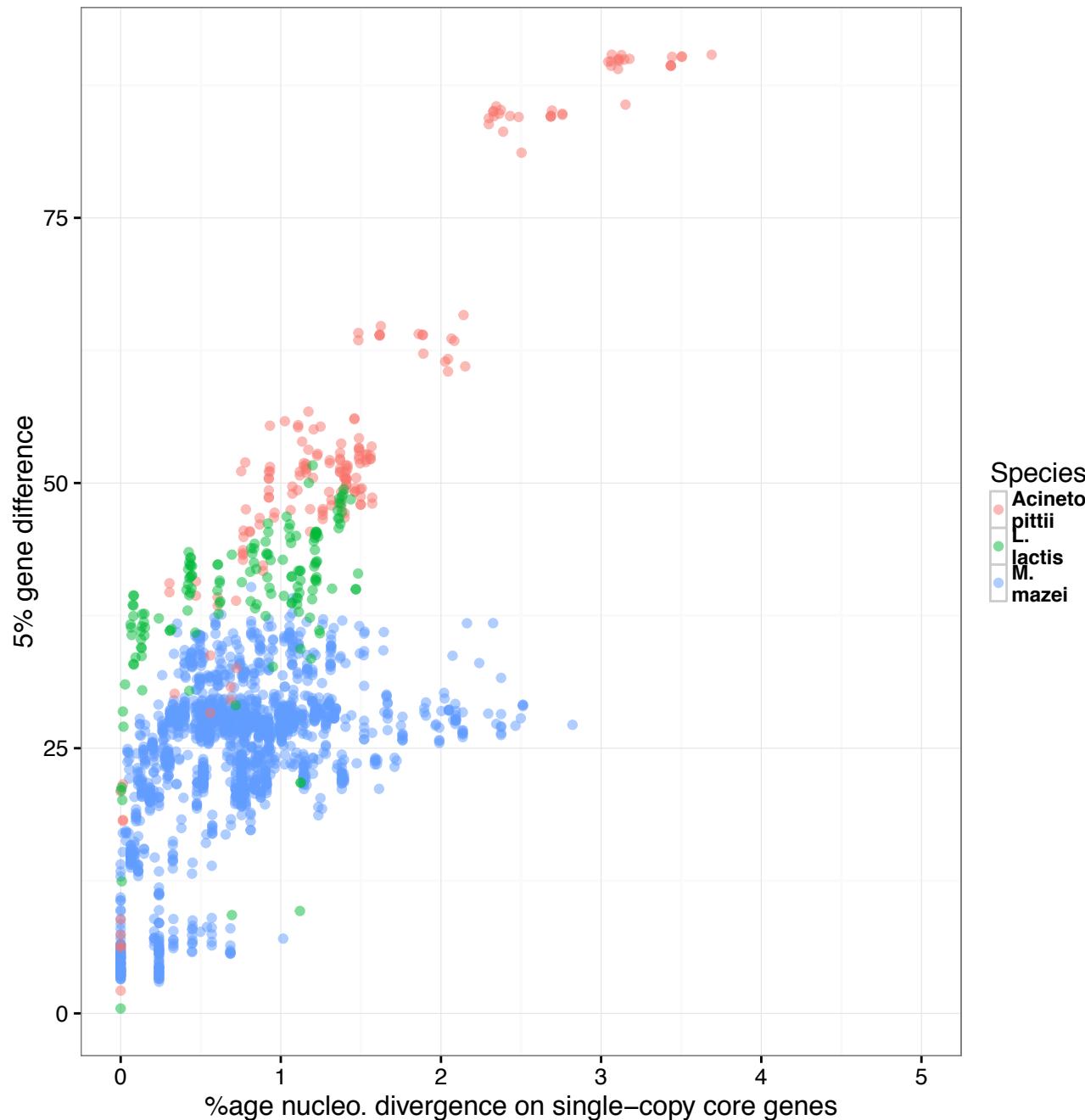
Phylum

- Bacteroidetes
- Chloroflexi
- Euryarchaeota
- Firmicutes
- Ignavibacteriae
- Planctomycetes
- Proteobacteria
- Spirochaetes
- Candidatus
- Wolfebacteria

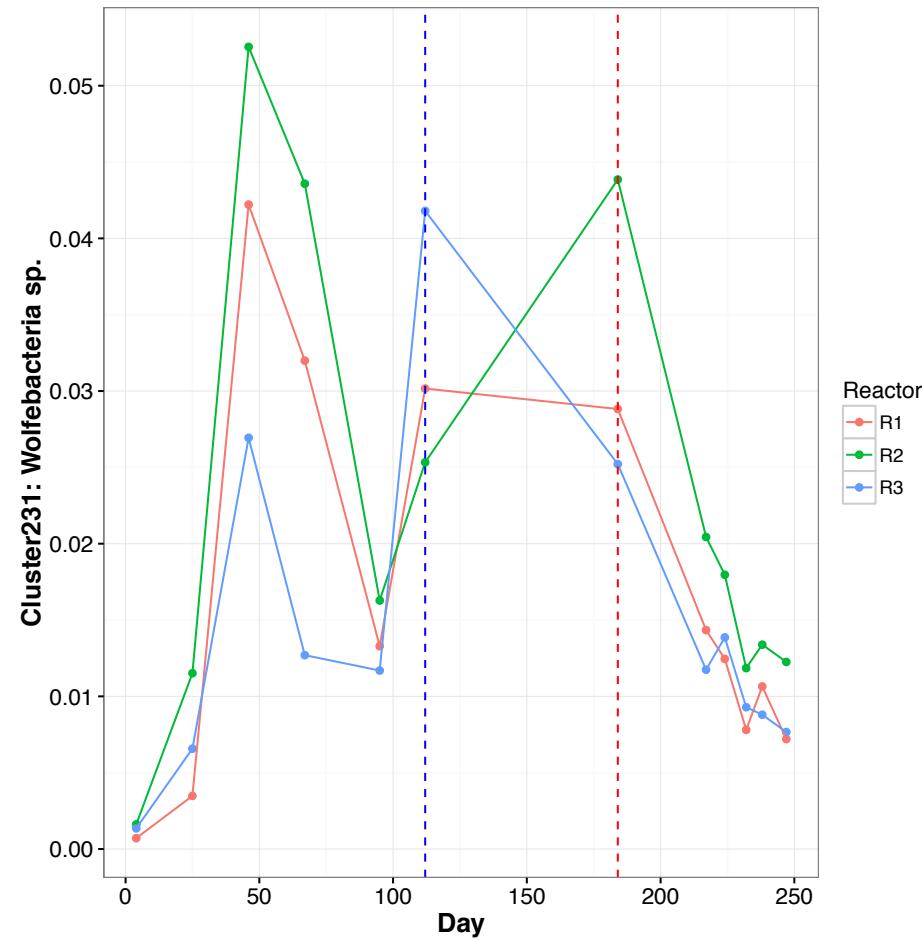
AD strain analysis



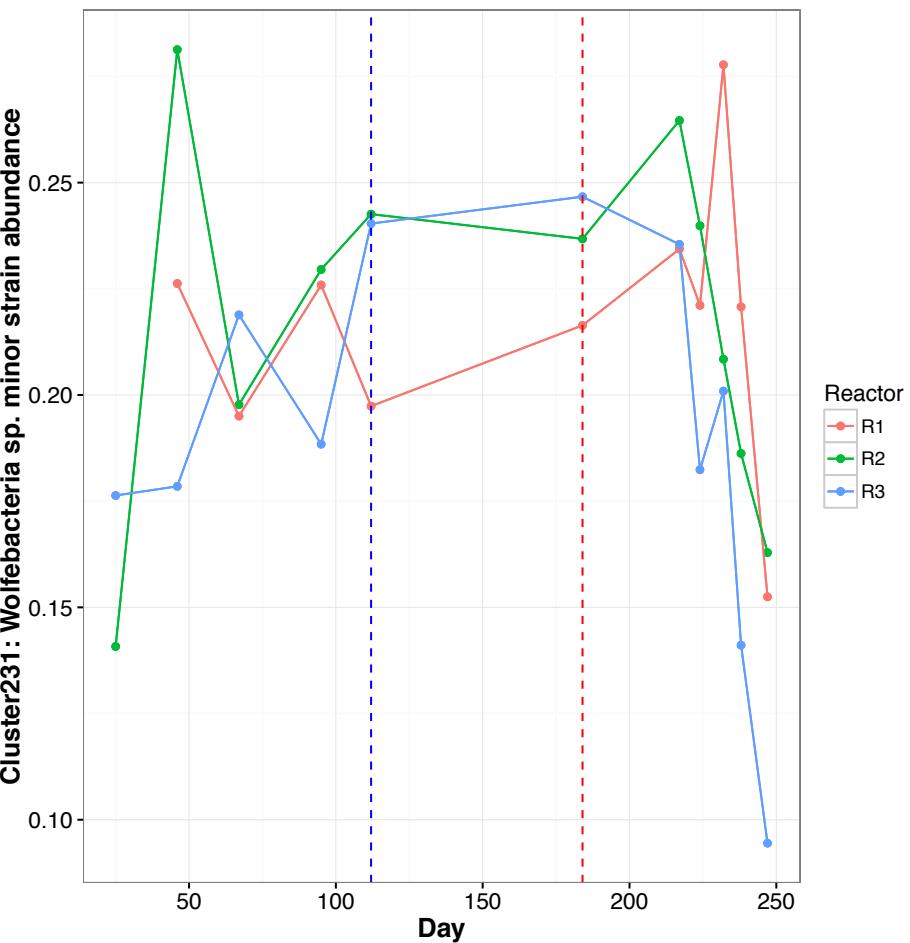
Comparison to real environmental organisms



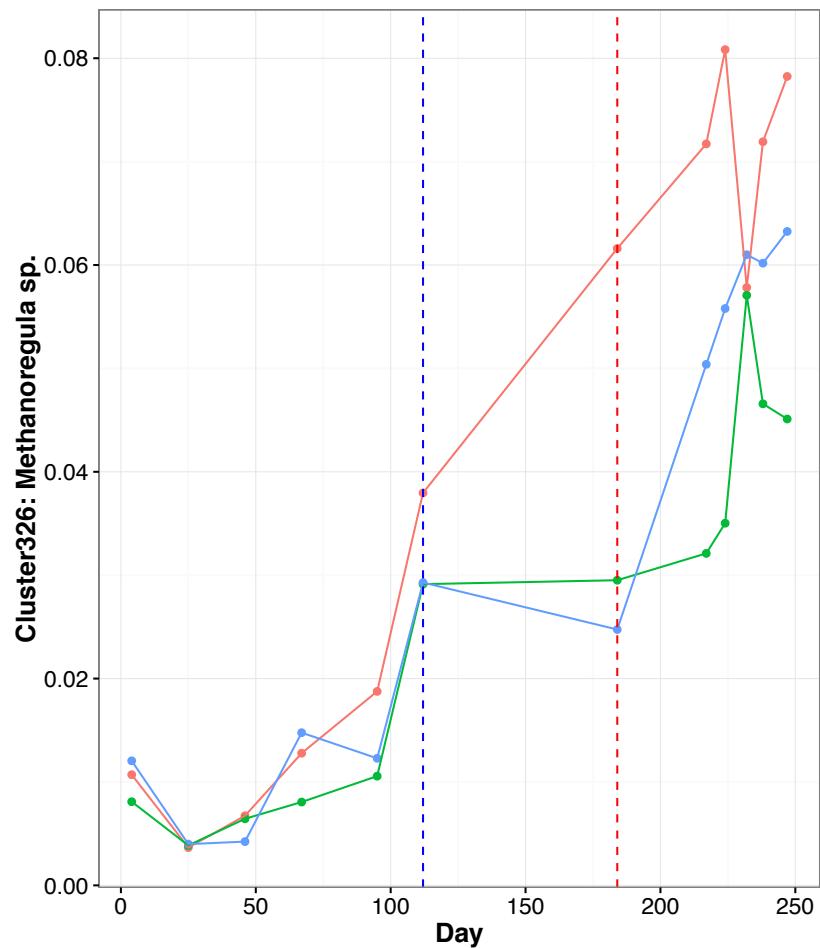
MAG abundance



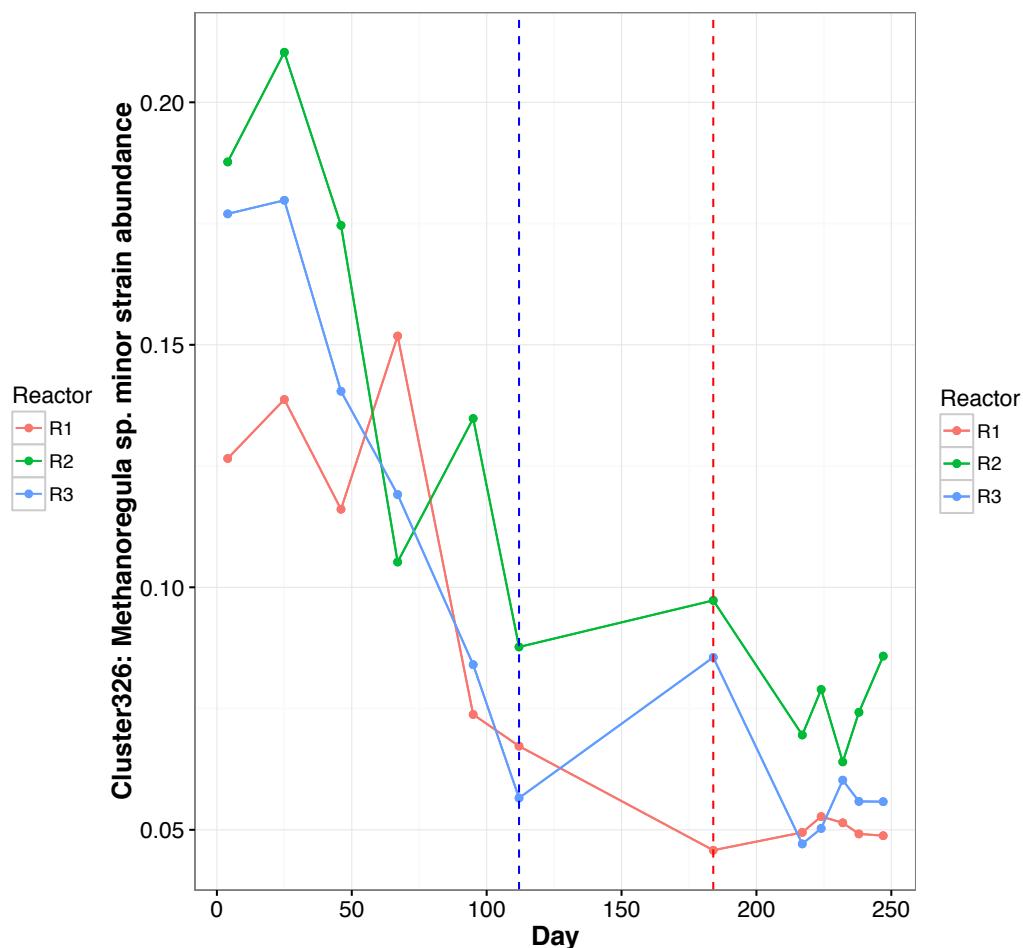
Strain abundance



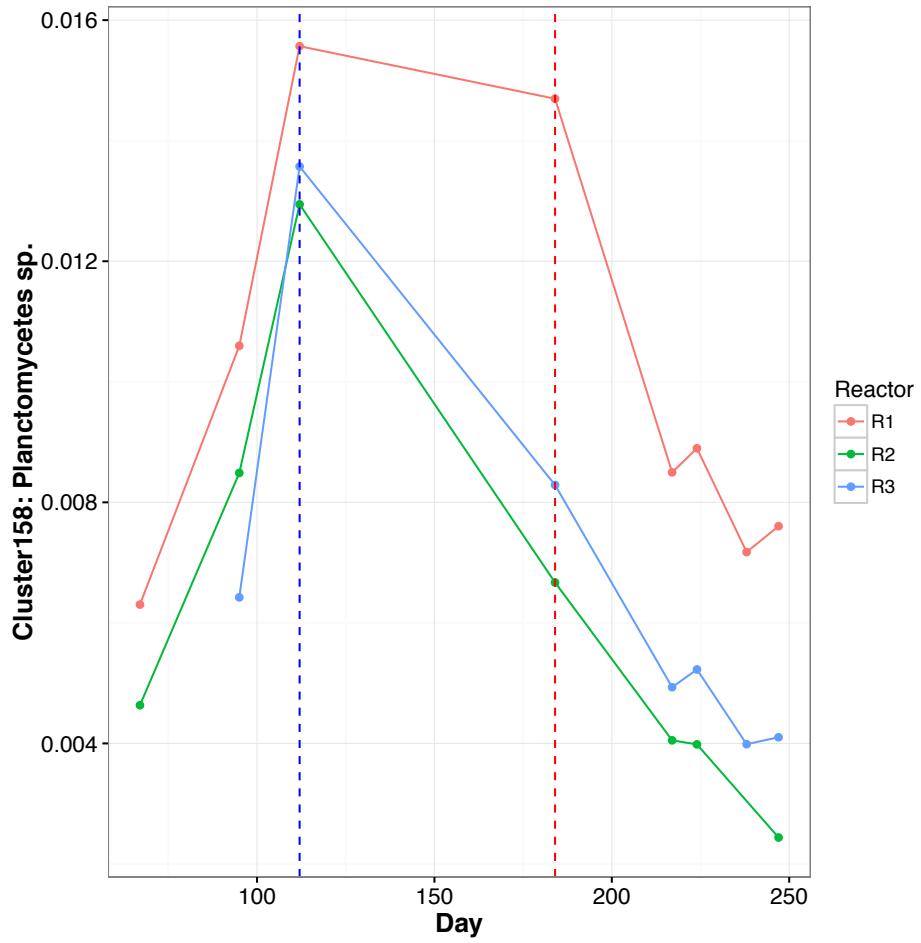
MAG abundance



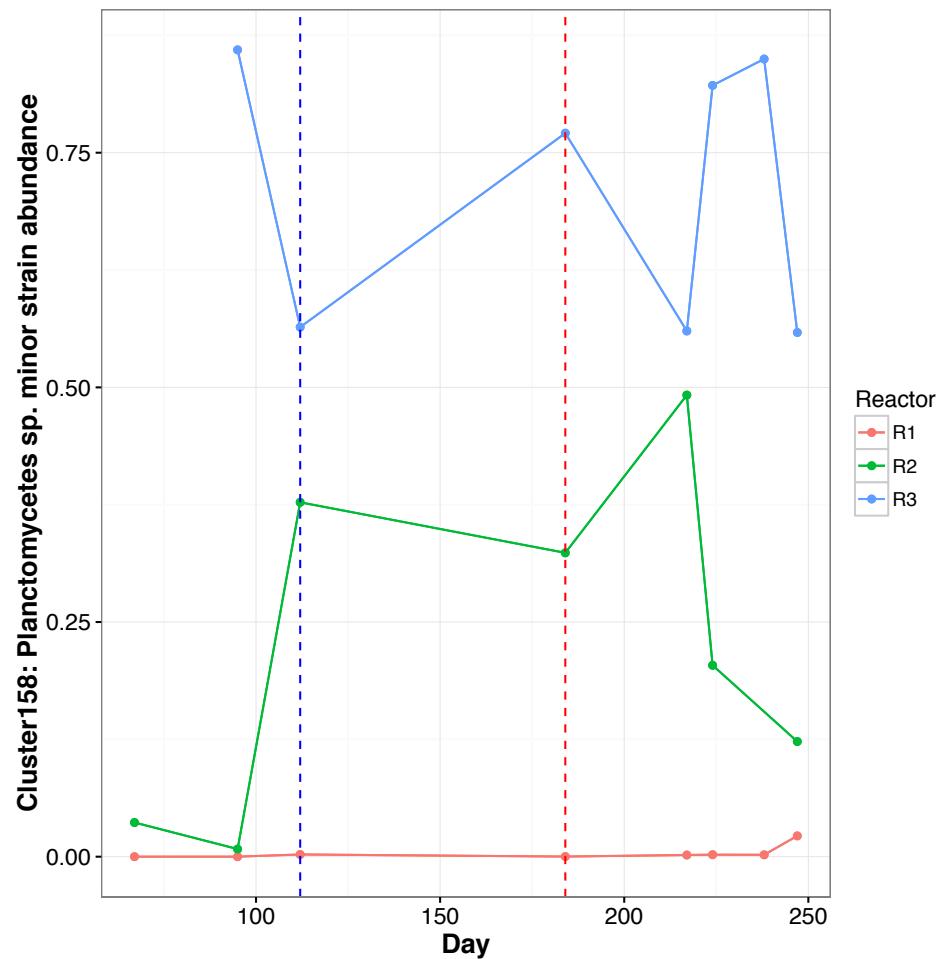
Strain abundance

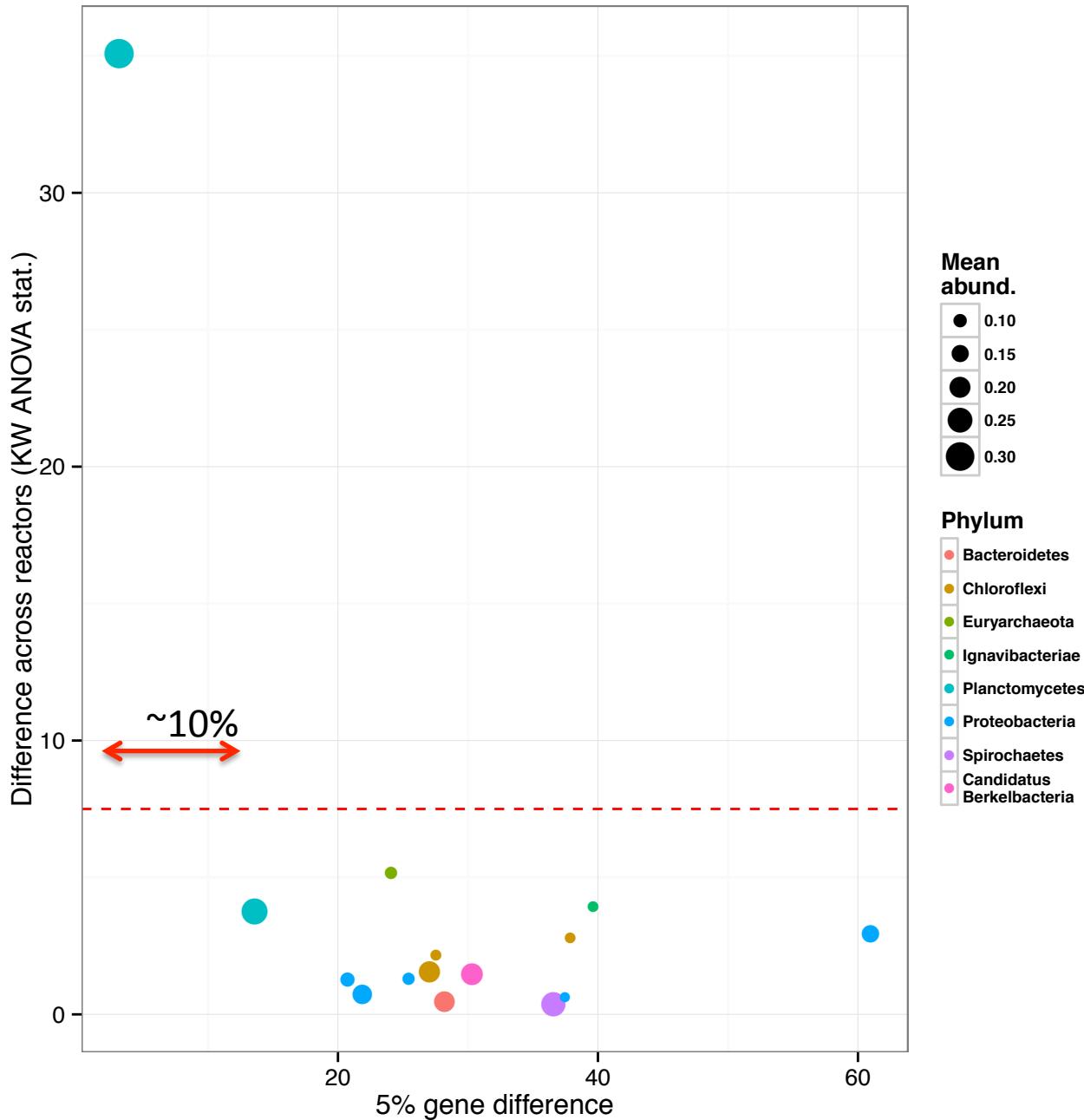


MAG abundance



Strain abundance

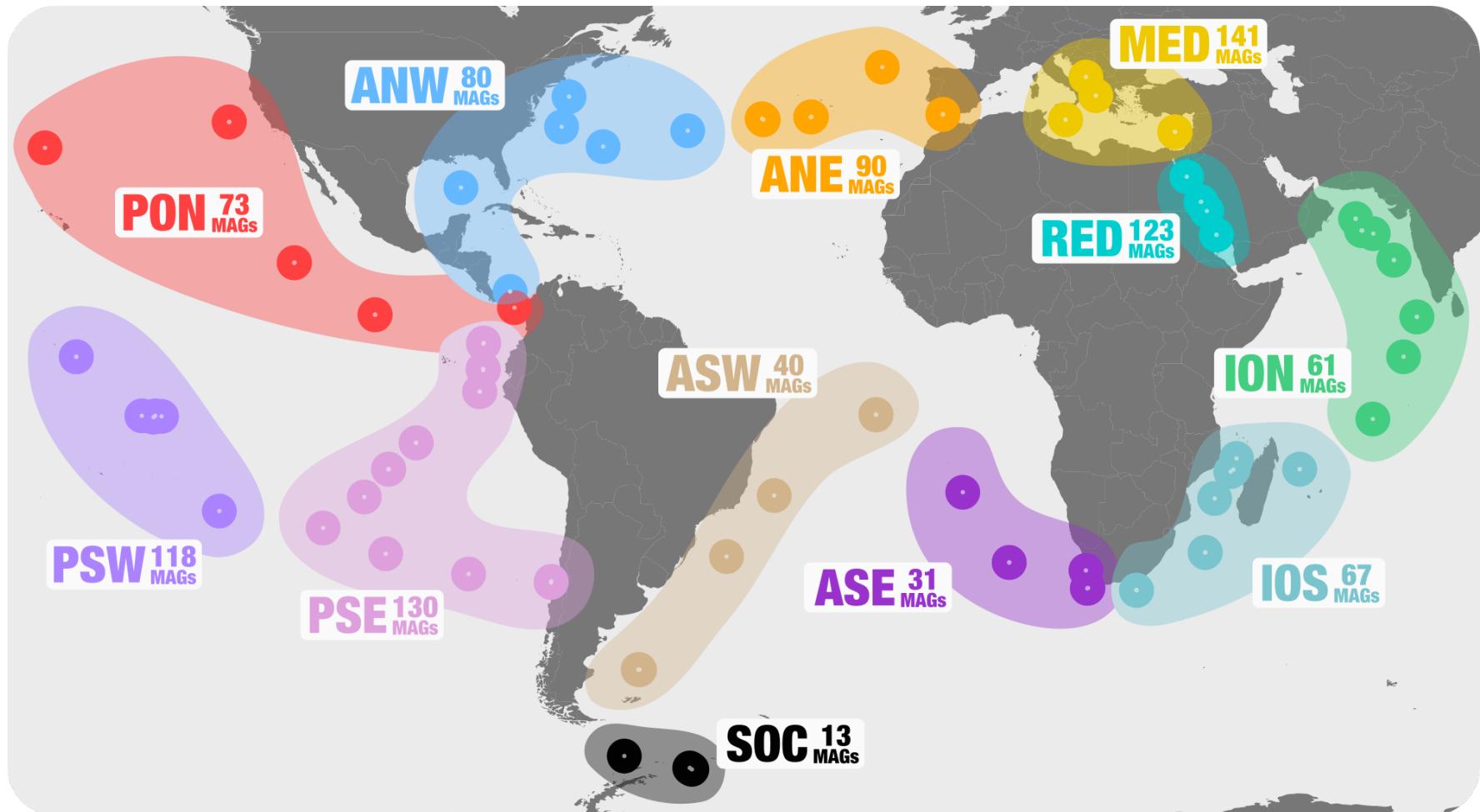




Hypotheses:

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

TARA Oceans: 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

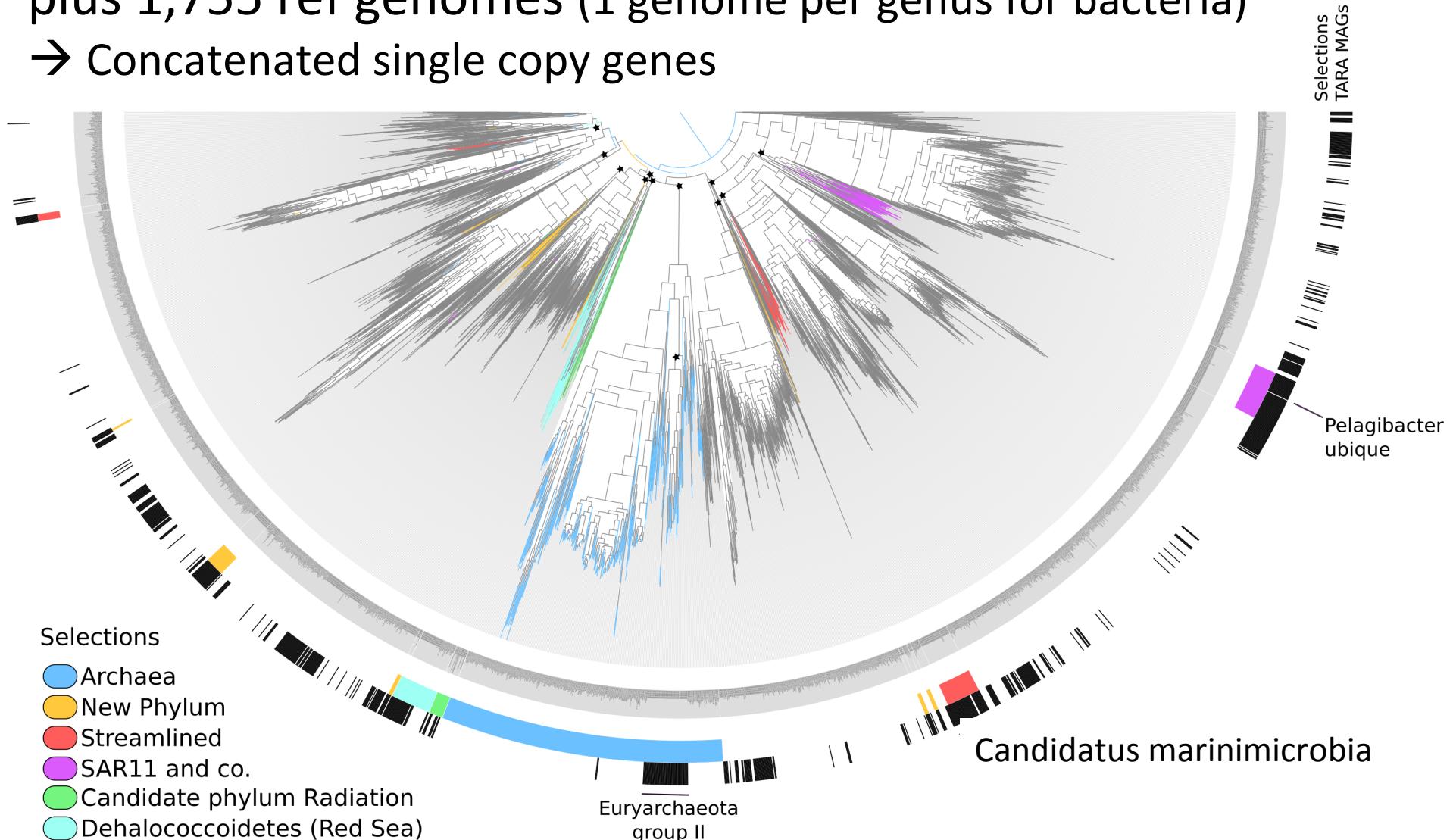
Delmont et al Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in the surface ocean <http://biorxiv.org/content/early/2017/04/23/129791>



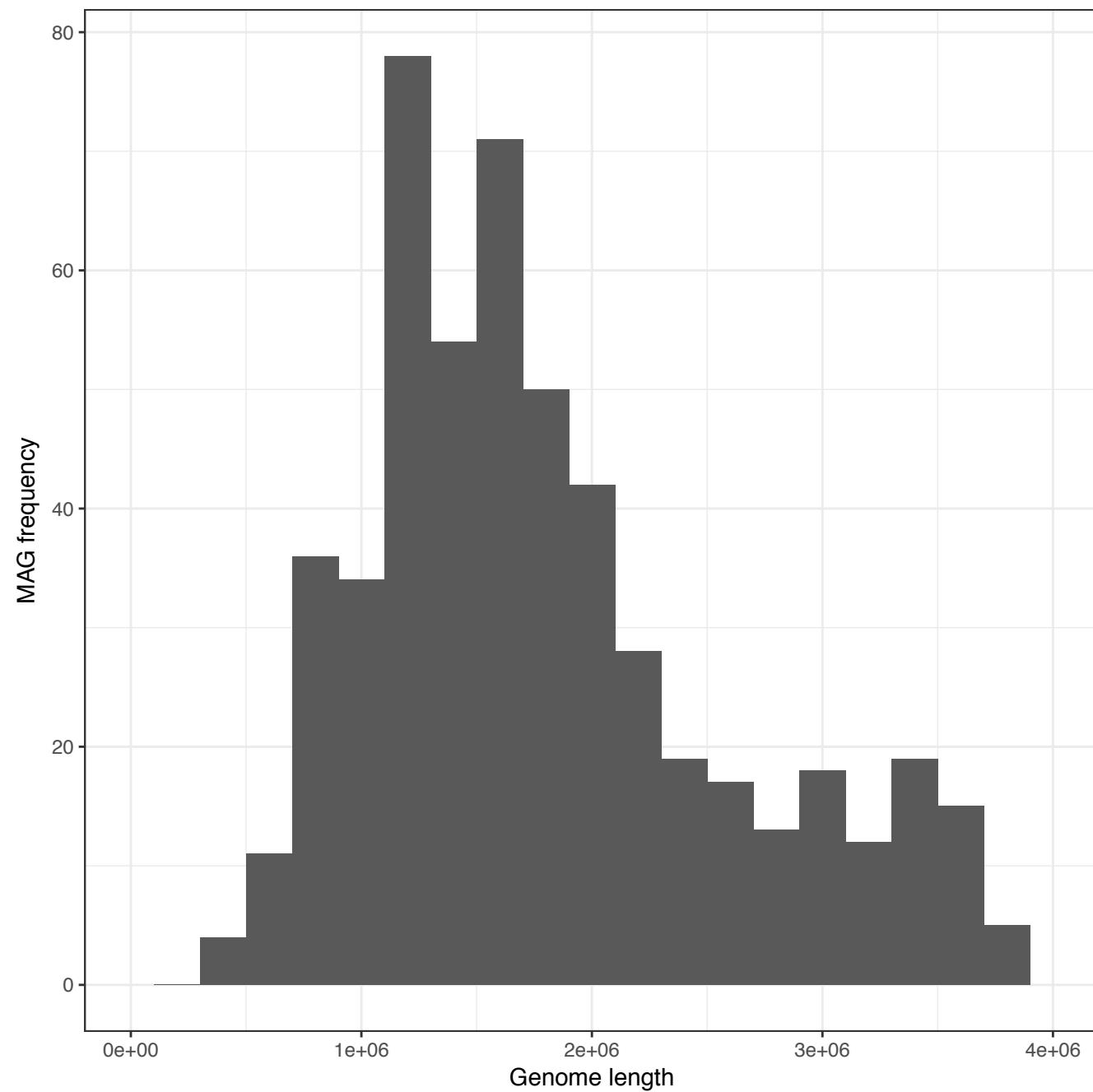
Phylogenetic analysis of 660 MAGs

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

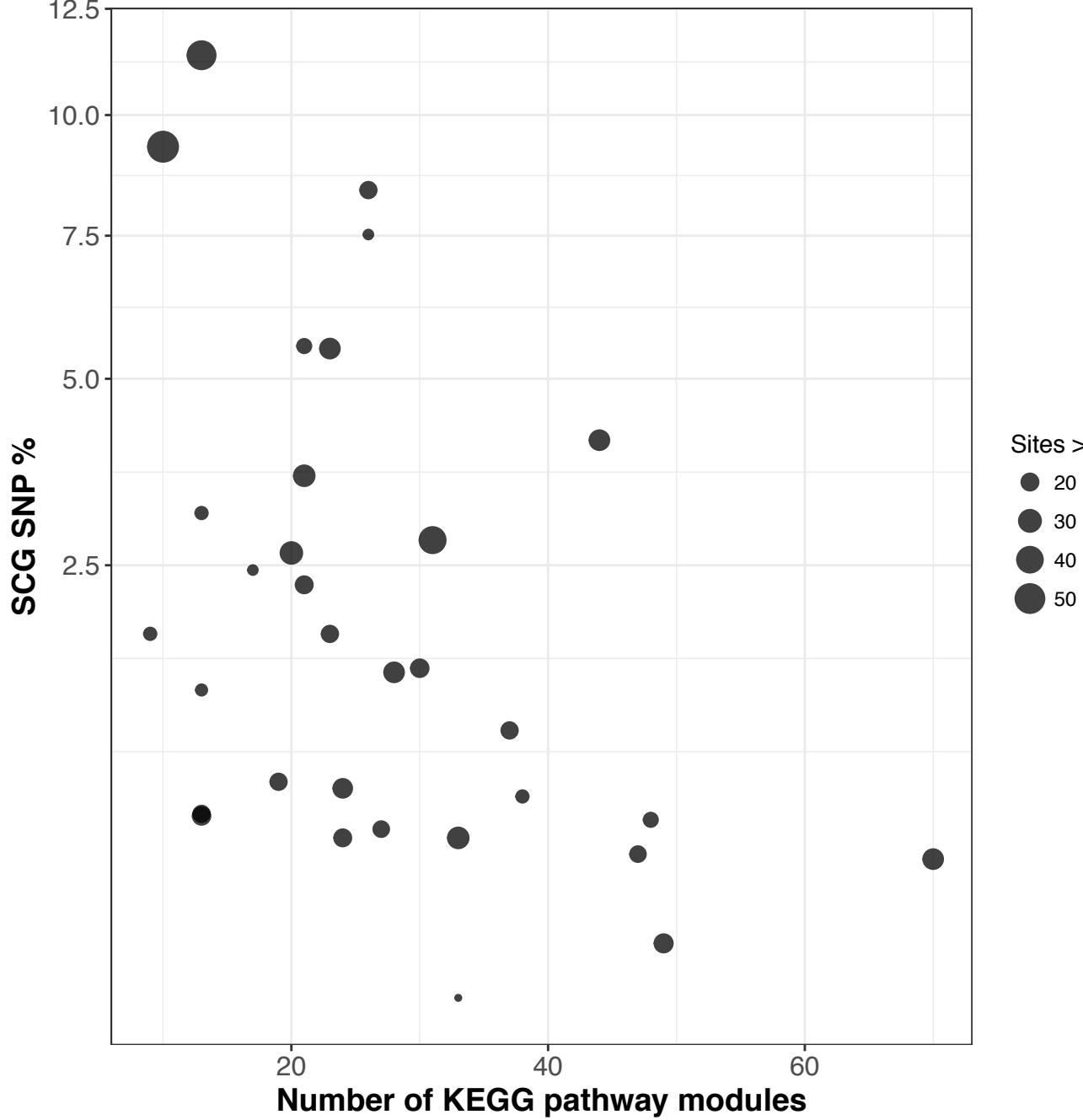
→ Concatenated single copy genes



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>



Tara MAG variants



Consider 32 MAGs with cov. > 100.

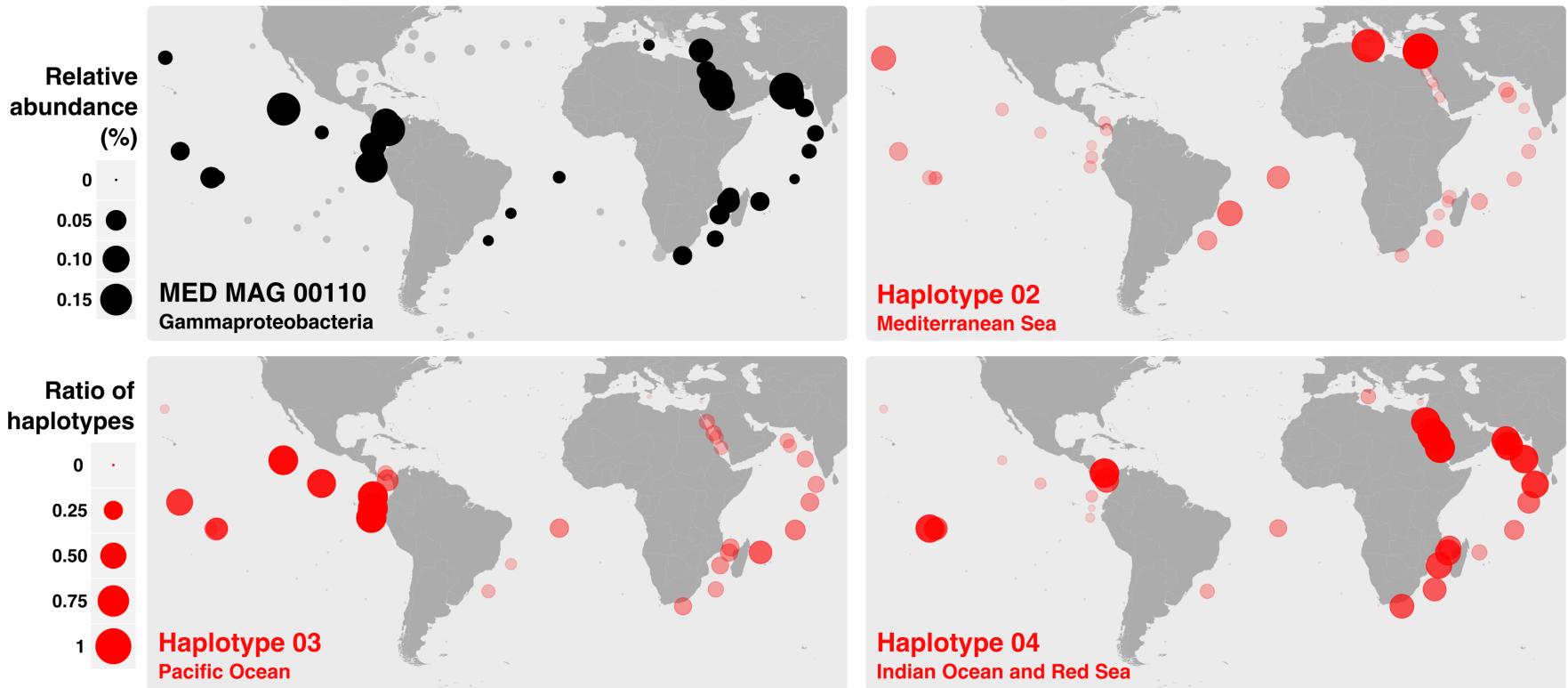
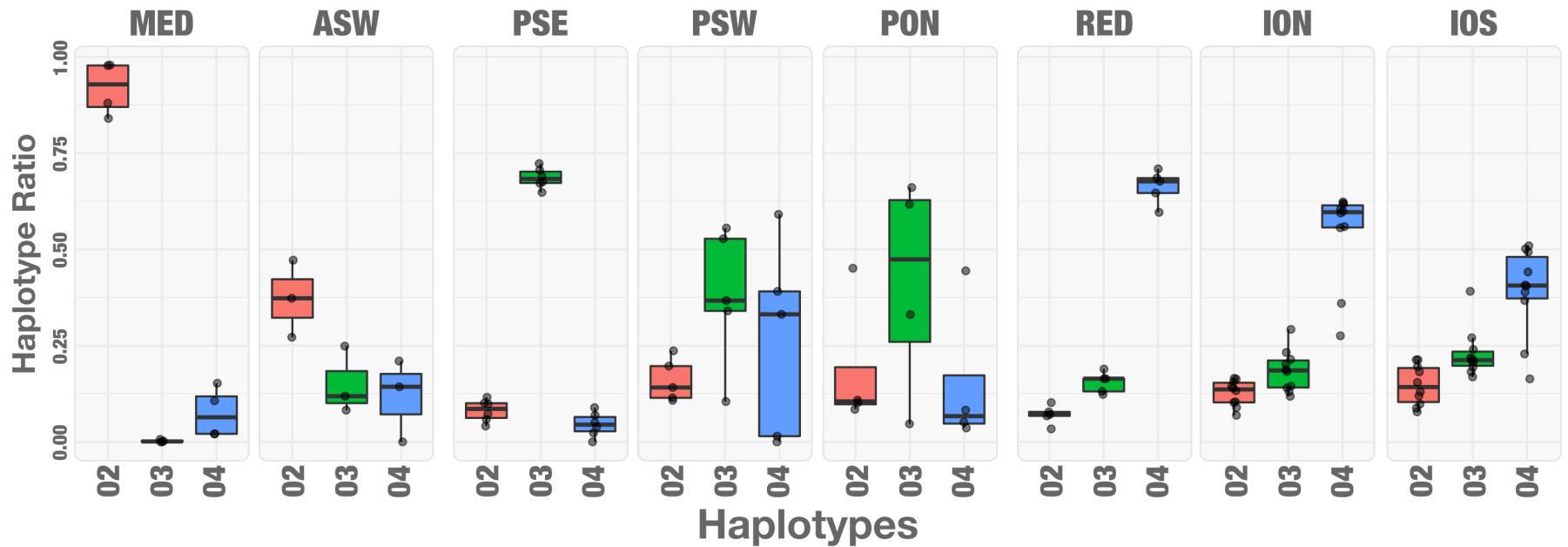
The SNP frequency was independent of MAG coverage

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

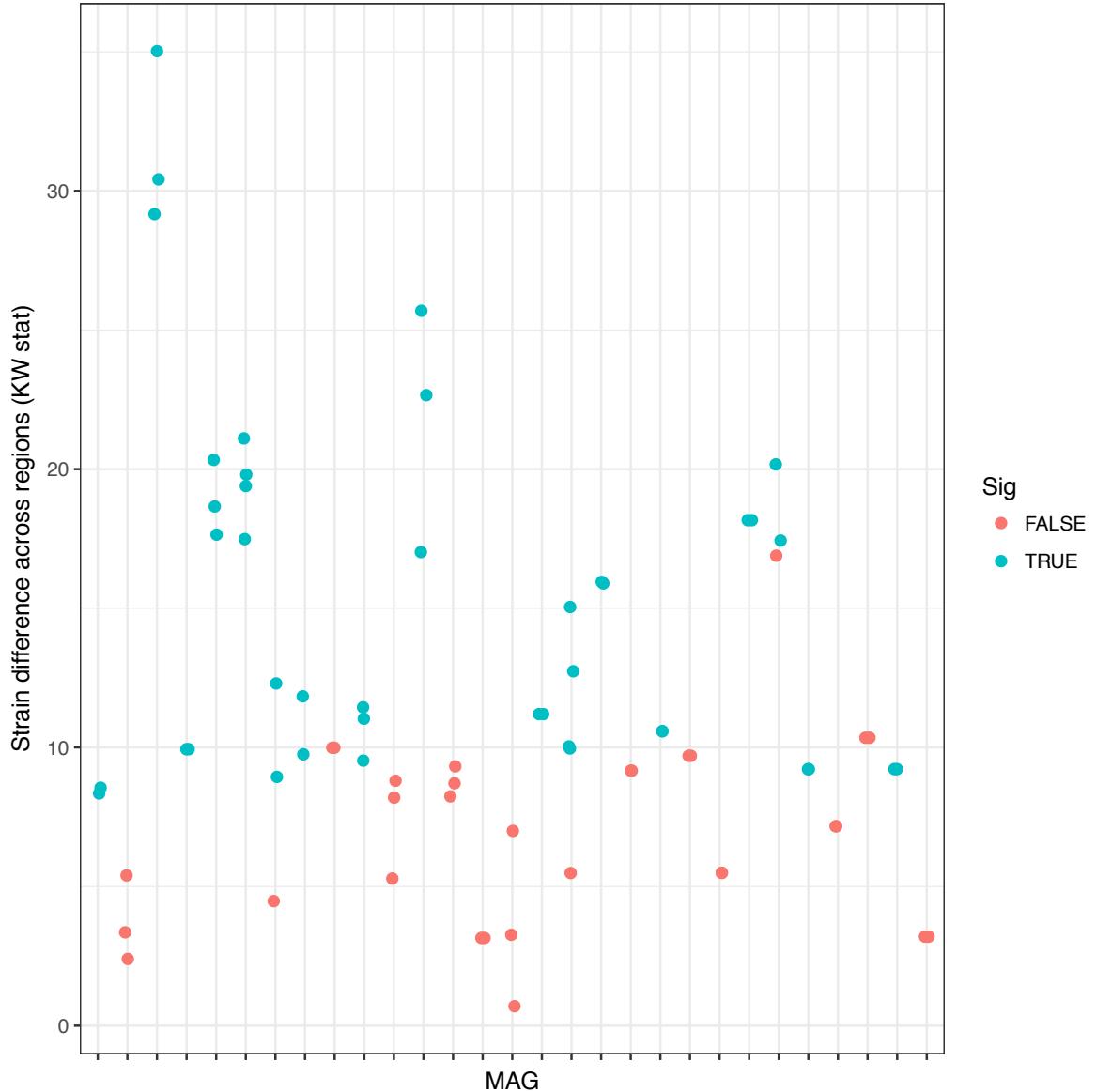
Stronger negative relationship with number of KEGG Pathway modules encoded in the genome (Spearman's p-value = 0.0045)

TARA DESMAN analysis

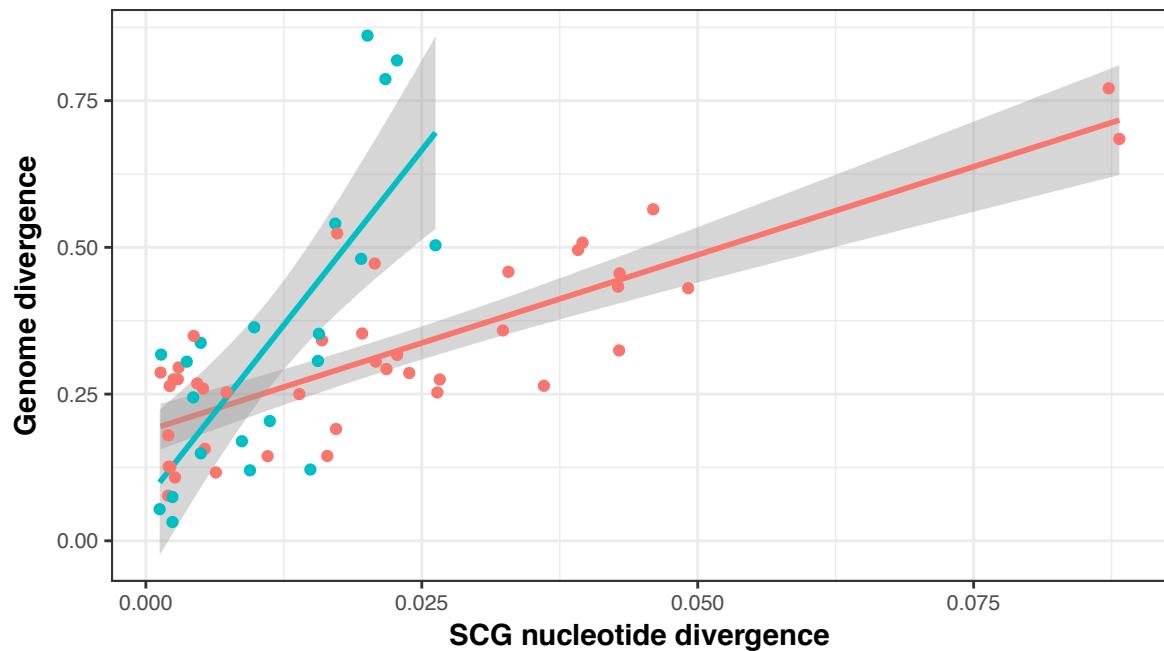
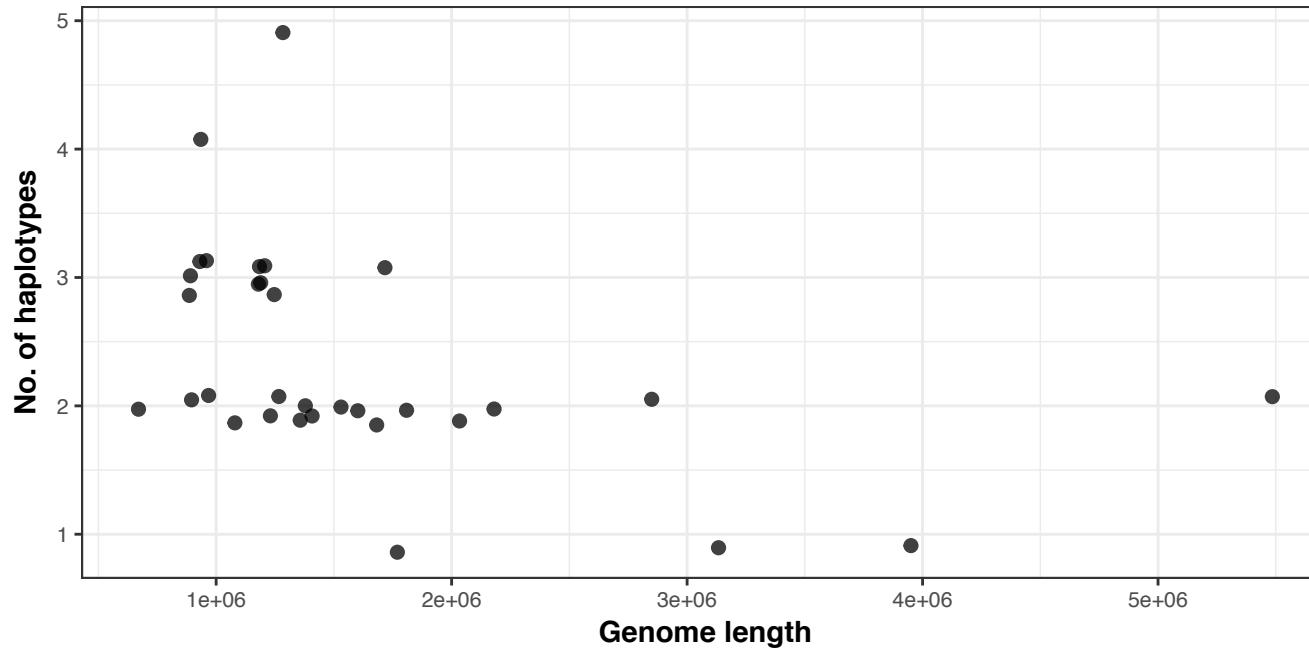
- Out of the 32 MAGs tested for haplotypes 29/32 had strain variation (1->3 2->17 3->10 4->1 5->1)
- The haplotypes were geographically localised e.g. TARA MAG 00110 a Proteobacteria with a highly streamlined 890,789 bp genome
 - Large group of uncultured organisms (relatives *Candidatus Evansia muelleri* 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



Majority of haplotypes across all MAGs showed correlation with geographic region



- $42/73 = 57\%$ of inferred strains had a significant correlation with geographic region
- $29/73 = 40\%$ correlated with mean sea temperature
- Ordering MAGs by length there is a trend to more geographic signal although not significant ($p\text{-value} = 0.1753$)



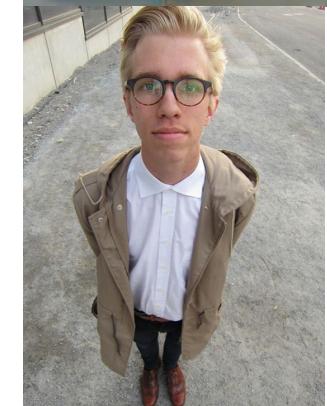
Summary

- DESMAN increases the amount of biologically relevant information that can be extracted from an assembly:
 - Resolve abundant strains *de novo* with high accuracy from core genome
 - Obtain accessory gene complement
 - <https://github.com/chrisquince/DESMAN>
- Implemented on Mosaic (info@mosaicbiome.com or ksoh@its.jnj.com)
- Tara Oceans:
 - Adaptive strain diversity endemic in the Oceans
 - Stream lined genomes associated with increased genetic diversity and strain diversity
 - Evidence of rapid relative genome divergence for very streamlined < 1Mbp organisms
- Future directions:
 - Strain resolution directly on the assembly graph
 - Long read technologies (Nanopore and Synthetic Long Reads)

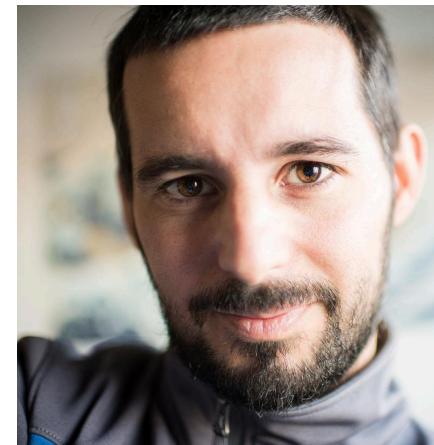
Stephanie Connelly + Gavin Collins + Nick Loman + Joshua Quick
Seung Gu (U. of Glasgow) (U. of Birmingham)



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



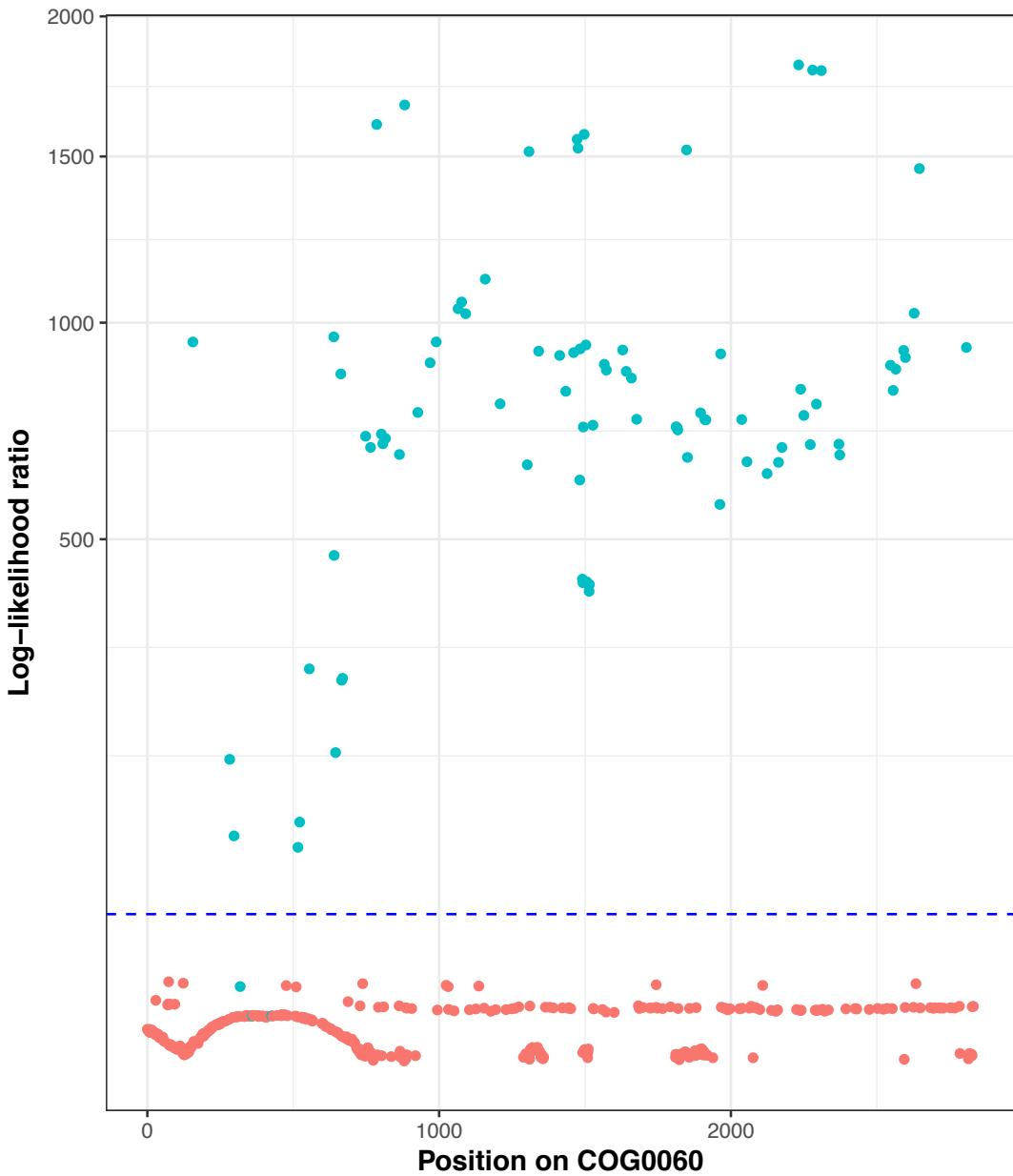
A. Murat
Eren and
Tom
Delmont
(Chicago)



STAMPS CONCOCT/DESMAN Tutorial

<https://github.com/chrisquince/Ebame3/DESMAN.md>

Variant prediction for Cluster 59 -> *Pseudomonas chlororaphis*



COG0060 predict 98 variant positions which includes 98 out of 99 true variants

Over all 34 SCGs and 23914 positions with $q < 0.001$

Predicted

Variant	False	True
False	23595	29
True	16	274

Recall = 94%, Precision = 90.4%