Metagenomics assembly and binning

GastroPak Bioinformatics Workshop

DR CHRISTOPHER QUINCE

Earlham/Quadram Group Leader

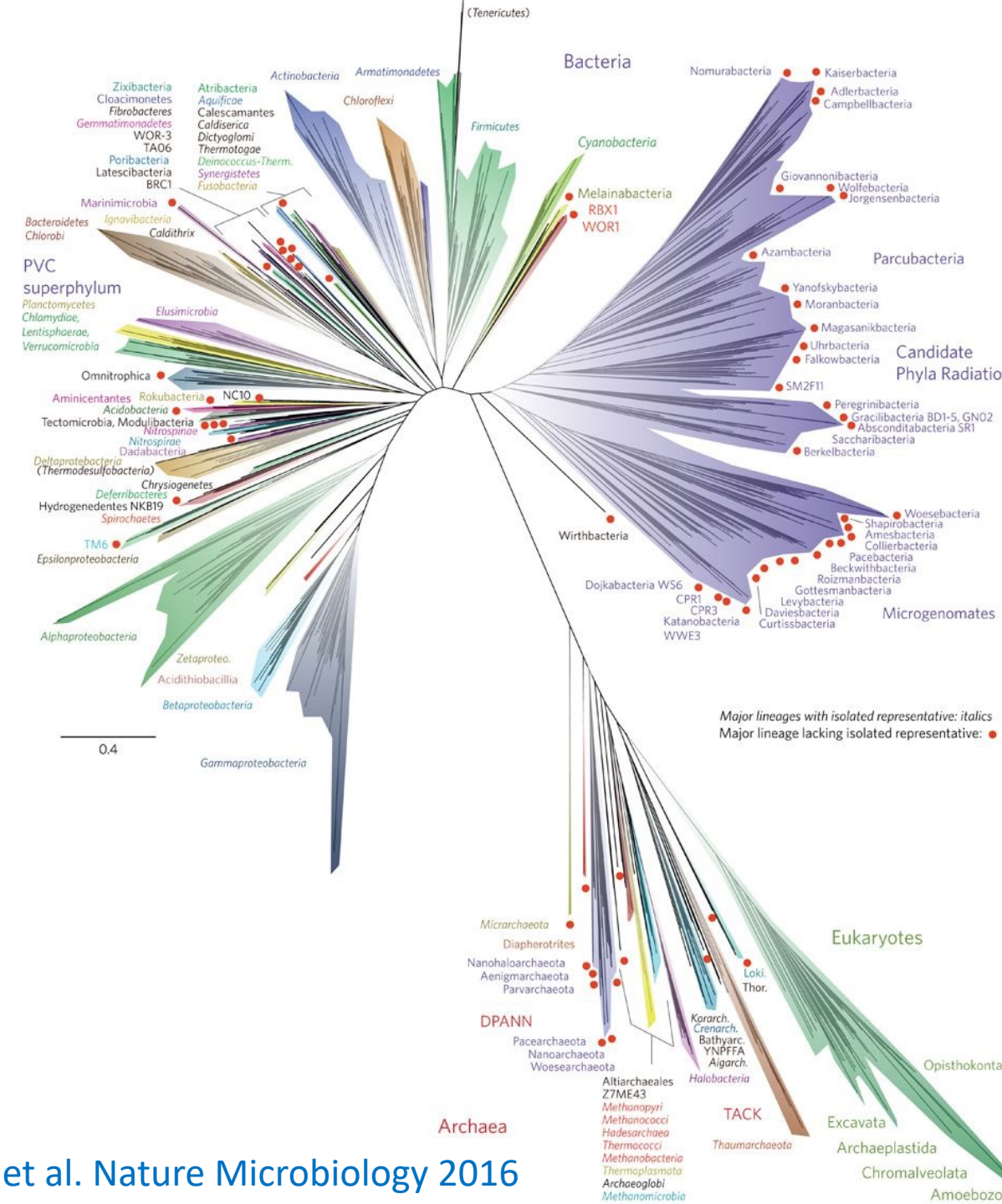
University of East Anglia

Quadram Institute
Science Health Food Innovation

Earlham Institute
Decoding Living Systems

Introduction

- Focus on automated pipelines for genome resolved metagenomics and applications to large scale data sets
- Metagenome assembled genomes have rewritten our understanding of environmental microbial diversity
- Large scale binning of human microbiome (Nature Milestones in human microbiota research 2019)

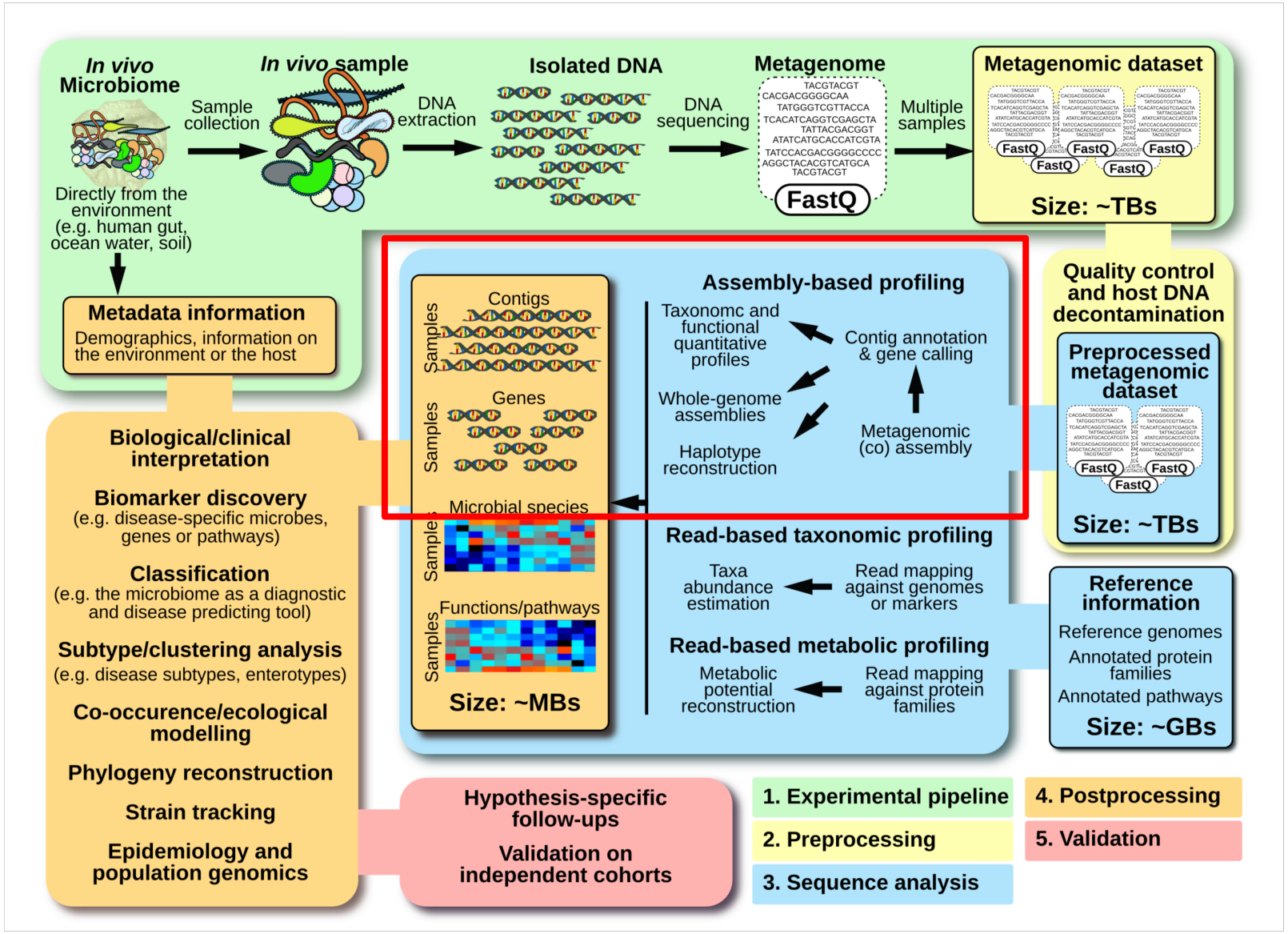


Hug et al. Nature Microbiology 2016

Overview

- Brief recap of short-read metagenomics assembly
- Present our pipelines for assembly-based metagenomics
- Application to CICRA dataset

Bioinformatics for shotgun metagenomics

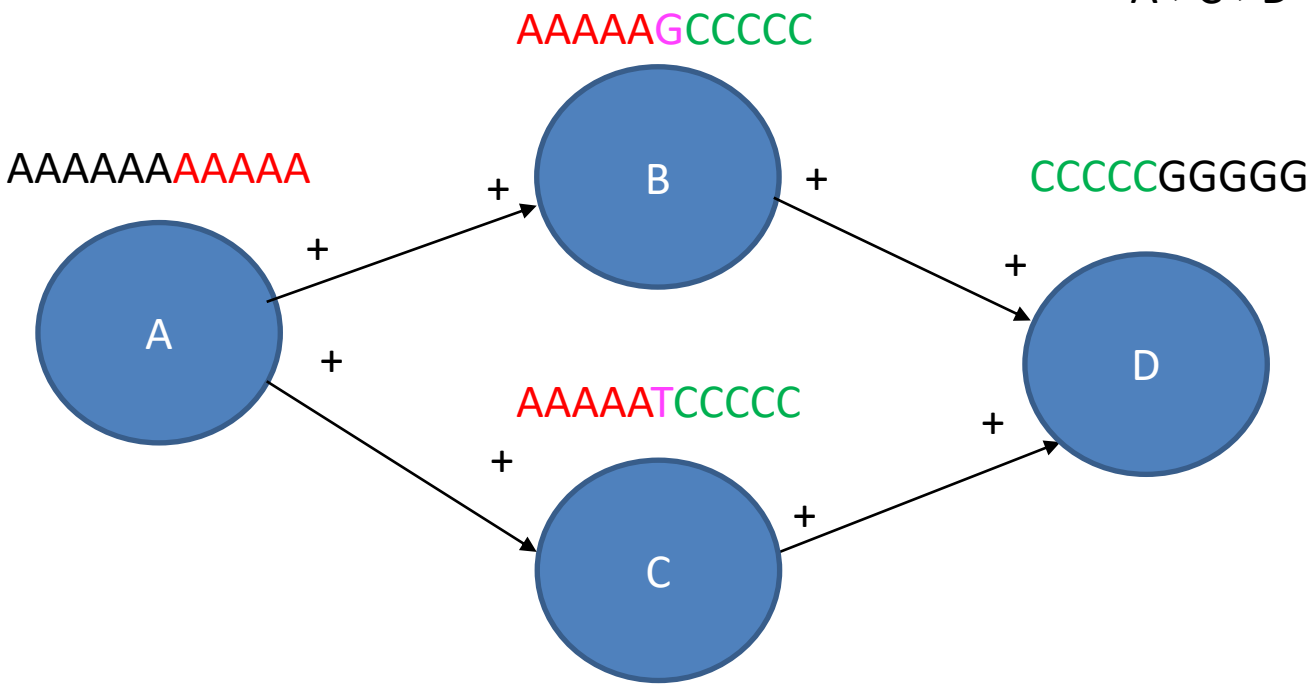


Quince et al Nat. Biotech 2017 http://rdcu.be/vMYk

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

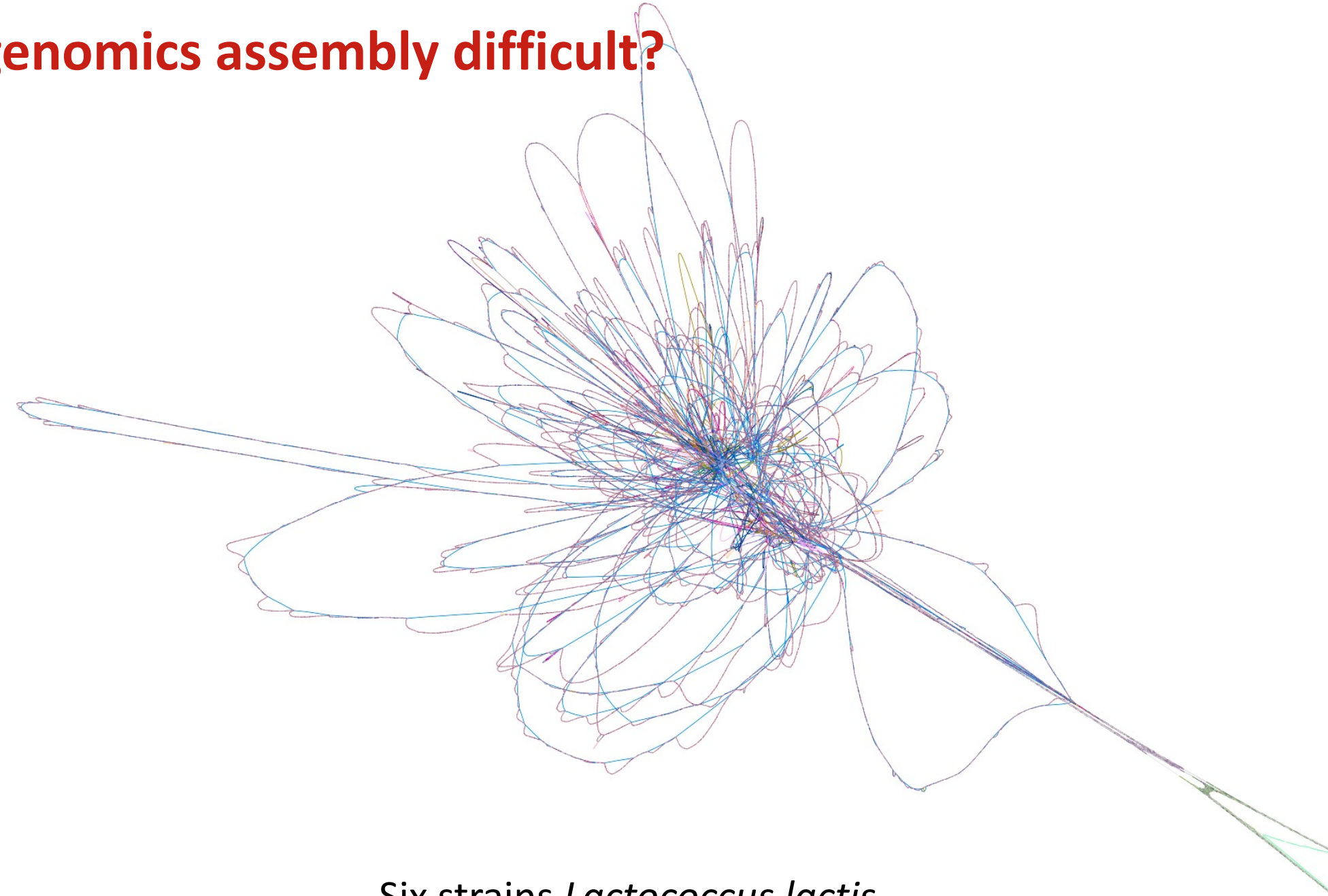
A->B->C = AAAAAAAAAAGCCCCCGGGGGG
A->C->D = AAAAAAAAAATCCCCCGGGGGG



Earlham Institute

Why is metagenomics assembly difficult?

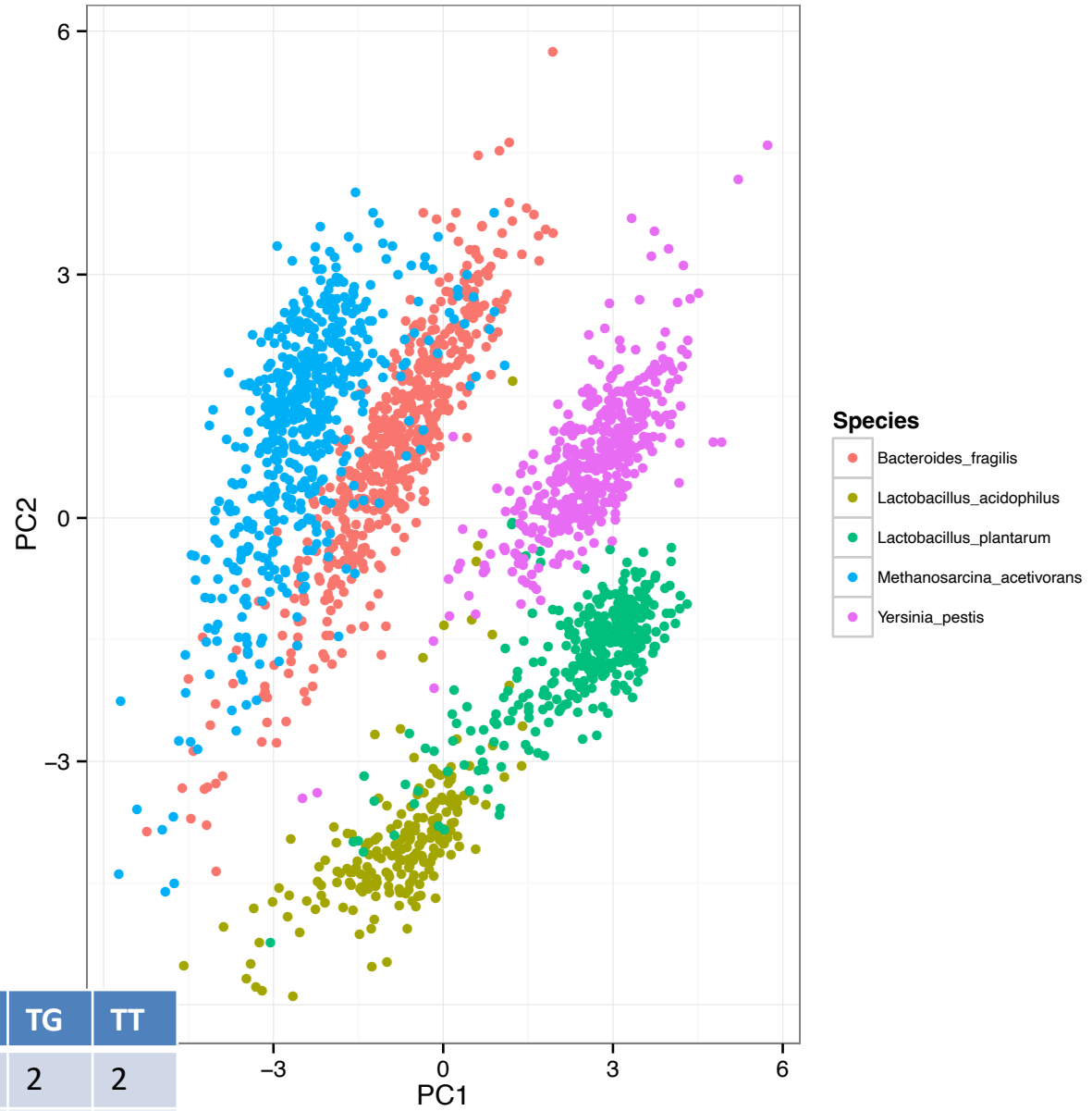
- Strains result in complex graphs
- Contigs are simply linear portions of graph post-simplification
- Use shared features across contigs to cluster into metagenome assembled genomes or MAGs



Six strains *Lactococcus lactis*

Contig clustering by sequence composition

- Represent 10kbp microbial genome fragments as tetramers then they cluster by species



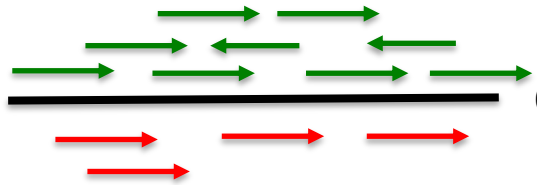
ACTTGCCACCTGCCTT

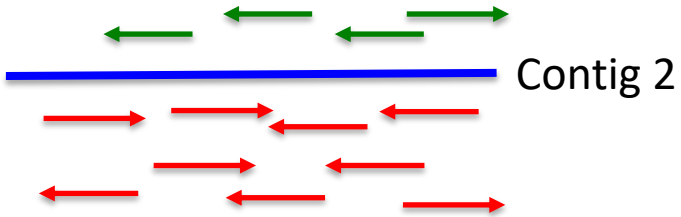
AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GT	GG	TA	TC	TG	TT
0	2	0	0	1	3	0	3	0	2	0	0	0	0	2	2
0	2	0	0	0	2	0	2	1	1	0	1	1	0	1	2
0	0	0	0	0	3	0	4	0	2	0	0	0	0	3	2

ACTGGCCTTTGACTA
CTTGCCTCCTGCCTTG

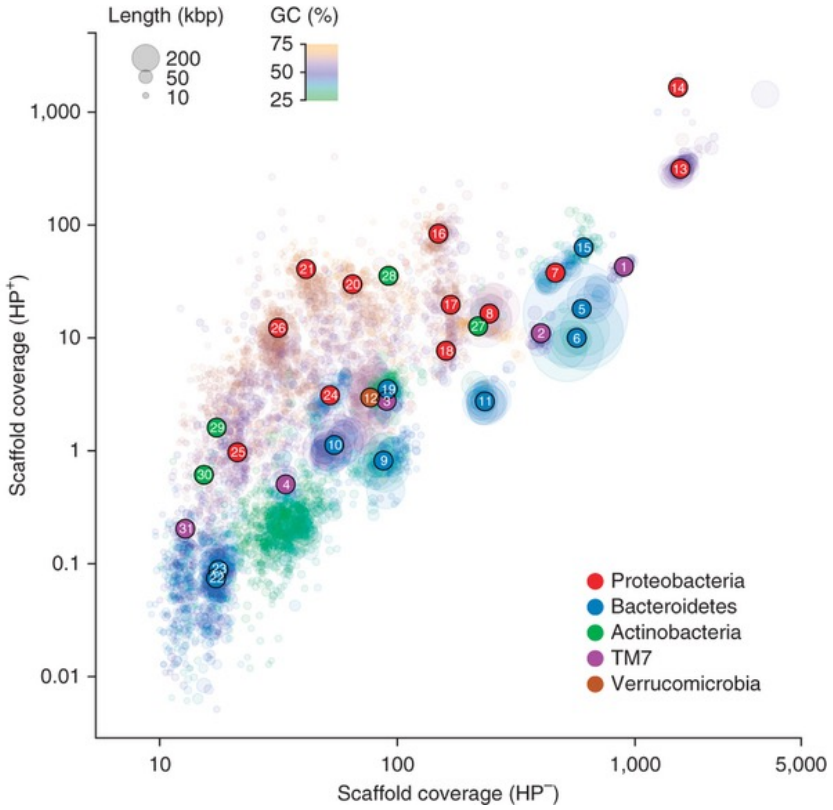
Contig clustering by differential coverage

$$\langle c \rangle = \frac{RN}{L}$$





- R is read length
- N is number of reads
- L is genome length

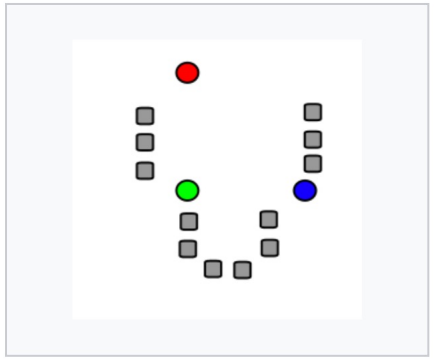


Albertsen et al. Nature Biotech. 2014

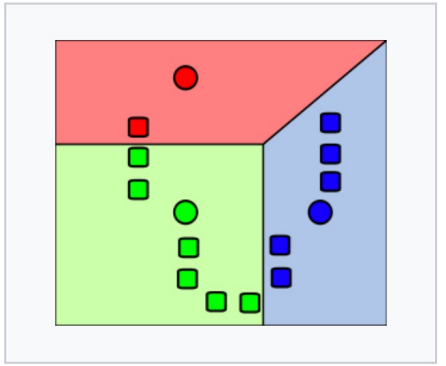
Earlham Institute

Simple algorithms for automatic binning

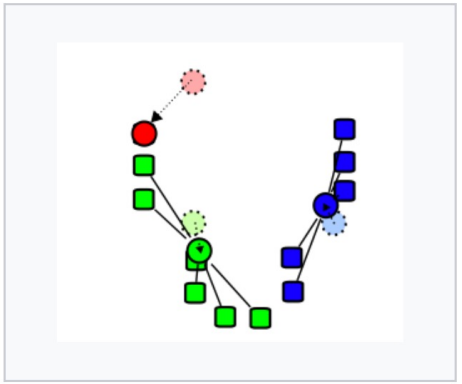
- Use shared features across contigs to cluster into bins:
  - Composition – tetramer frequencies
  - Coverage across multiple samples
- Simplest clustering algorithms depend only on a distance metric
- Example hierarchical clustering in ANVIO
- Partitional clustering – e.g. kmeans (MetaBAT Kang et al. PeerJ 2015)
- How to define number of clusters
- Model based clustering using statistical models



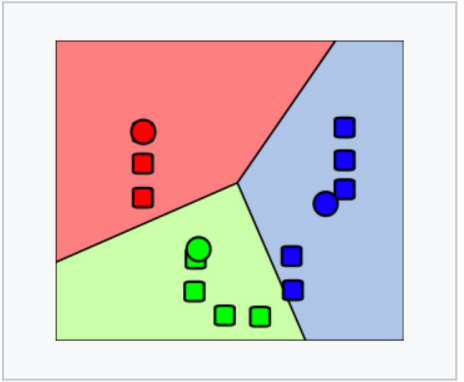
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



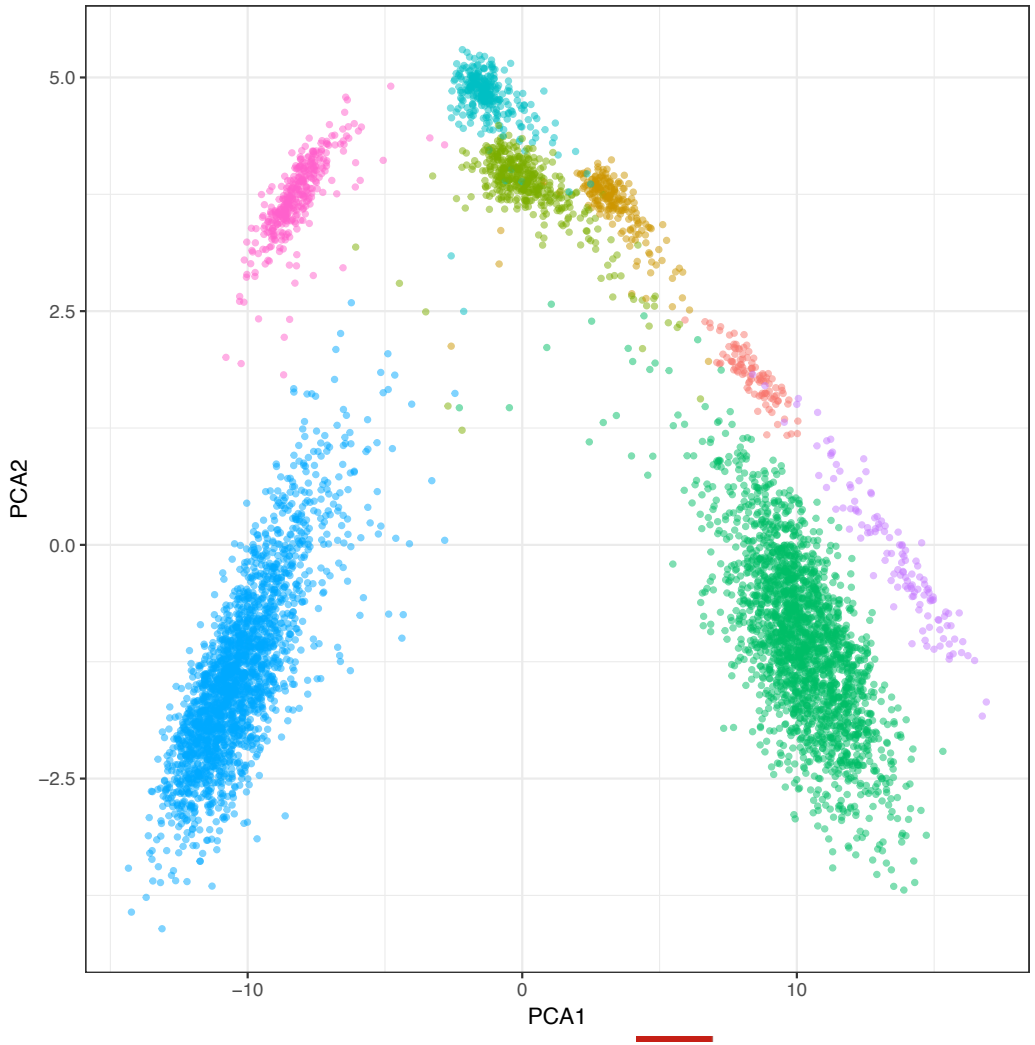
3. The centroid of each of the *k* clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

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

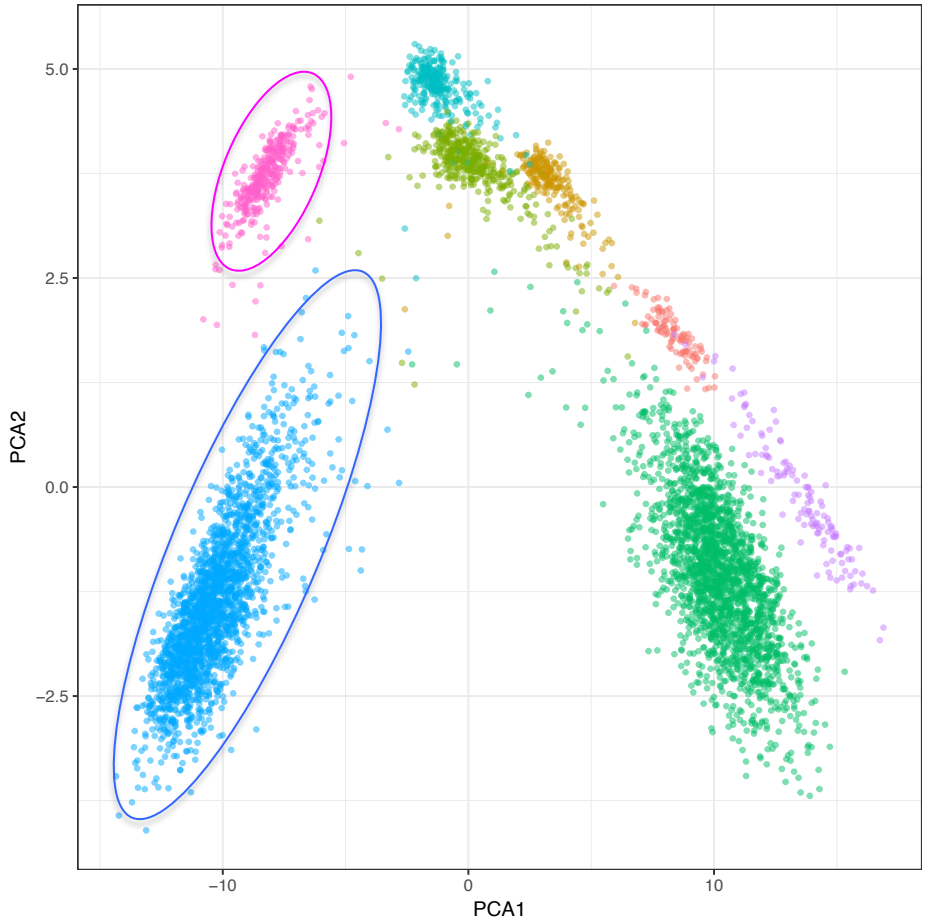
- Data pre-processing:
  - Perform coassembly across all samples
  - Fragment contigs greater than 10kb
  - Map reads back to *N* contigs to get mean coverage of contig in each of *S* samples
  - Generate k-mer frequency vector dimension *V* for each contig
  - Add pseudo-counts, normalise coverage and k-mer frequencies, join and log-transform
  - Perform PCA keep *D* dimensions that explains 90% of variance
  - One vector for each contig *n = 1* to *N*



stitute

CONCOCT algorithm

- Cluster with Gaussian Mixture Model
- Variational Bayes to select number of components through automatic relevance determination (ARD)

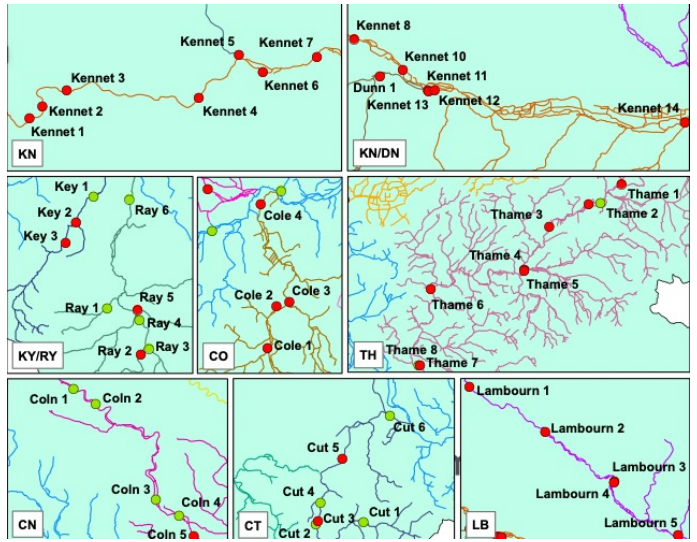


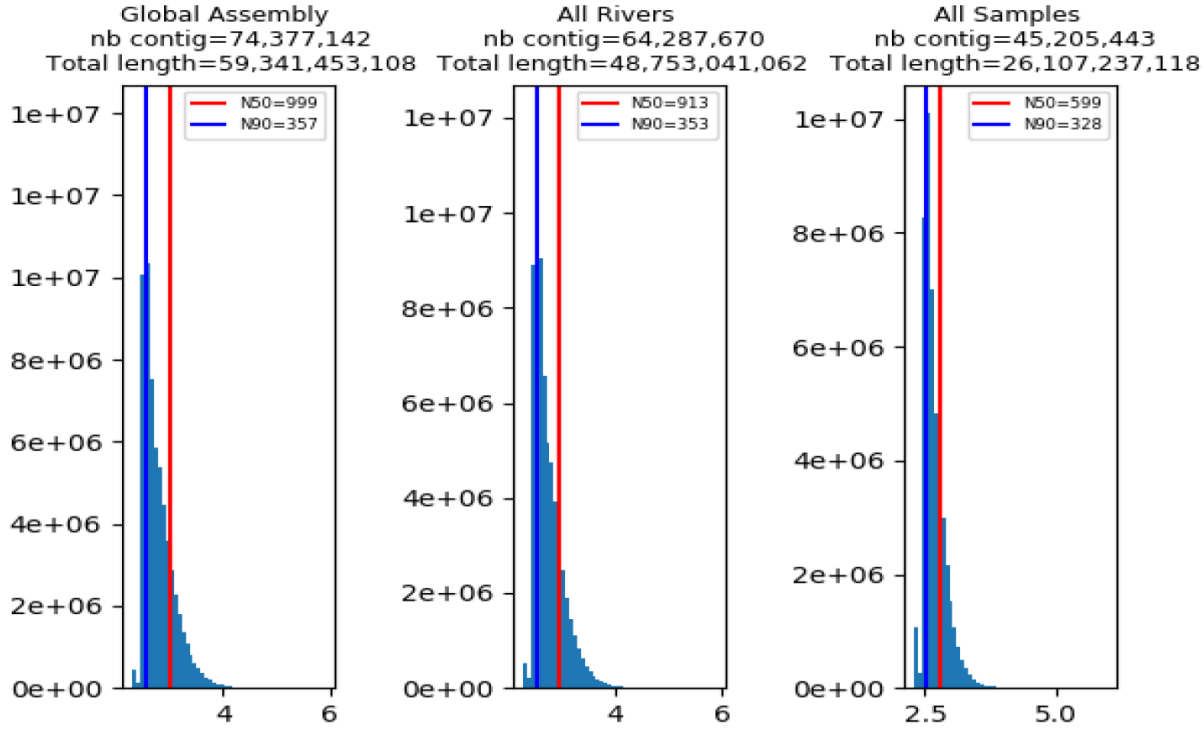
Current binning algorithms

- Crowded field now e.g. GroopM (Imelfort et al. PeerJ 2014), MetaBAT2 (Kang et al. 2019), MaxBin2 (Wu et al. Bioinf. 2015), COCACOLA (Lu et al. 2017), VAMB (Nissen et al. 2018)
- CONCOCT enabled some important discoveries: Asgard archea (Zaremba-Niedzwiedzka et al. Nature 2017) and Commamox (Pinto et al. mSphere 2016)
- Still competitive as a binner and less biased to prokaryote bins than more recent algorithms, effective for viruses and eukaryotes
- New algorithms generally not using new information exception GraphBin (Mallawaarachchi et al. 2020) and MetaCoAG (Mallawaarachchi and Vin 2021)
- Evaluate bins using single-copy core genes and assign to metagenome assembled genomes or MAGs (CheckM, EukCC): remember these just estimate MAG qualities

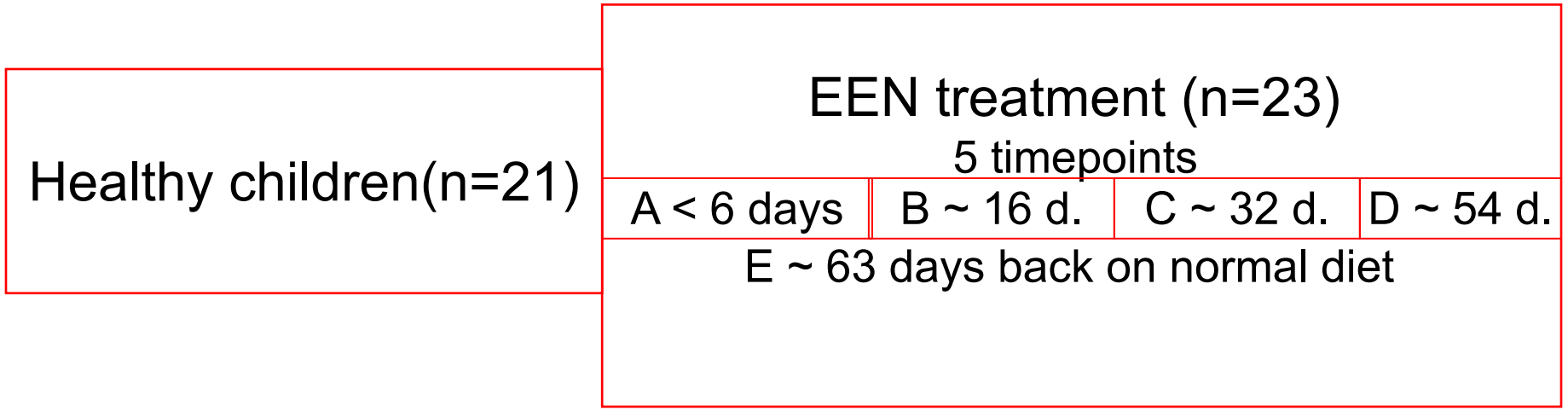
When to perform coassemblies vs single sample?

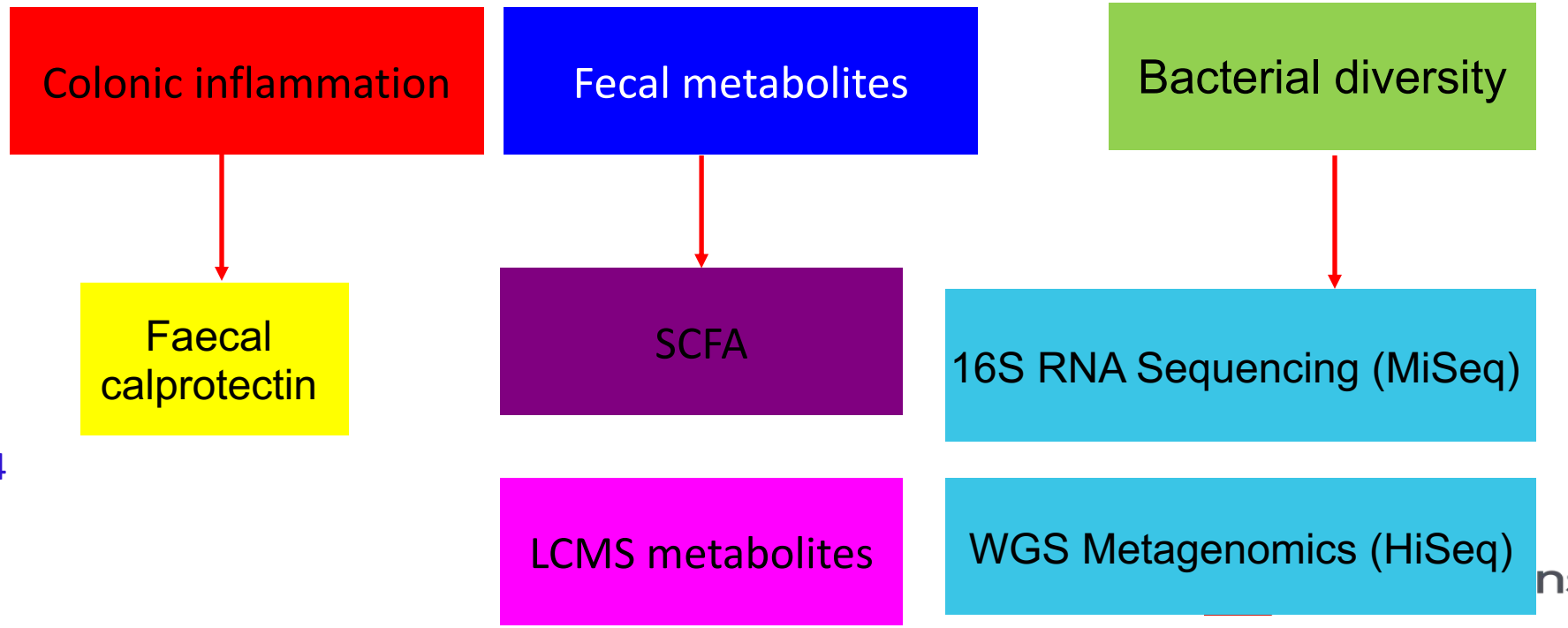
- Coassemble samples from the same community
- Access low coverage organisms
- Single sample assemblies if different strains in each sample
- To obtain maximal no. of MAGs combine assemblies and binning algorithms (Metahood, MetaWrap)





CICRA Study design



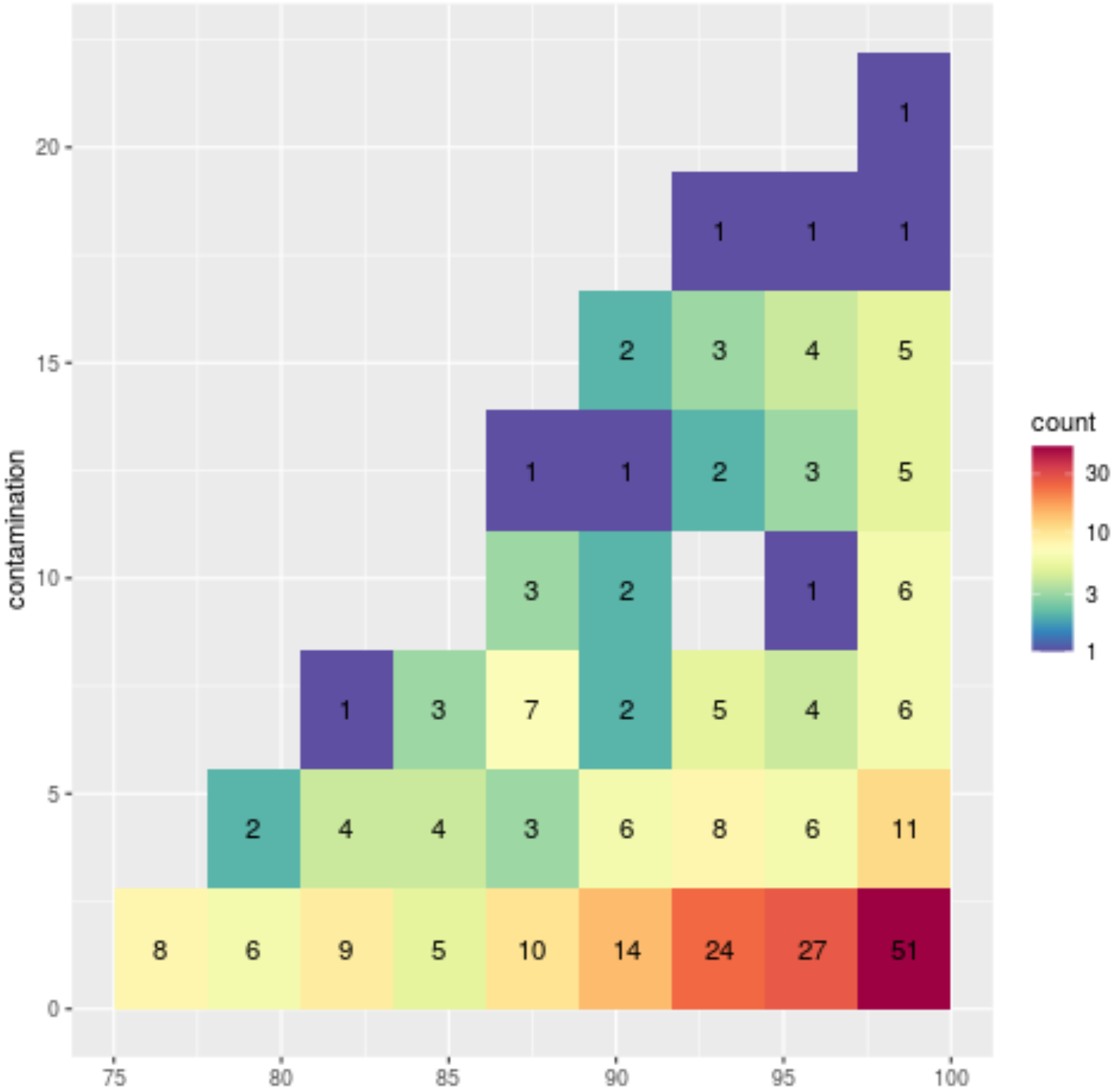


Decreased Butyrate and *Faecalibacterium prausnitzii* (qPCR) with treatment (Gerasimidis et al. Inflamm Bowel Dis. 2014 May;20(5):861-71.)

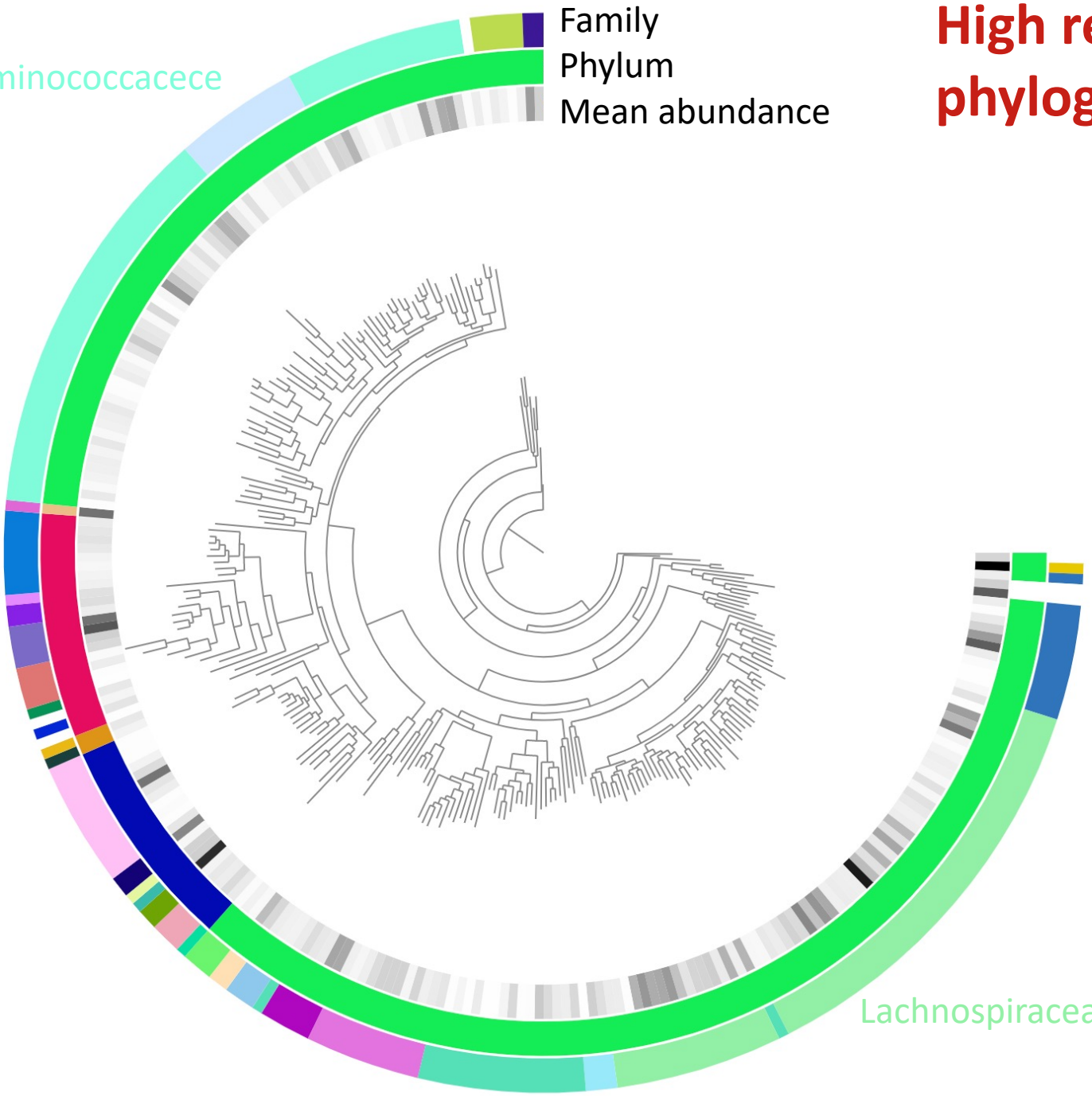
nstitute

Assembly and binning of pediatric Crohn's samples

- Coassembly total length: 3.06 Gbp N50: 1909 bp
- 636061 contig fragments > 1kbp
- Use panel 36 single-copy core genes
- 268 75% complete metagenome assembled genomes (MAGs)
- Estimate assembly contains 591 prok. genomes
- Median 61% of reads map onto assembly and 25% to MAGs



High resolution single copy core gene phylogeny of MAGs

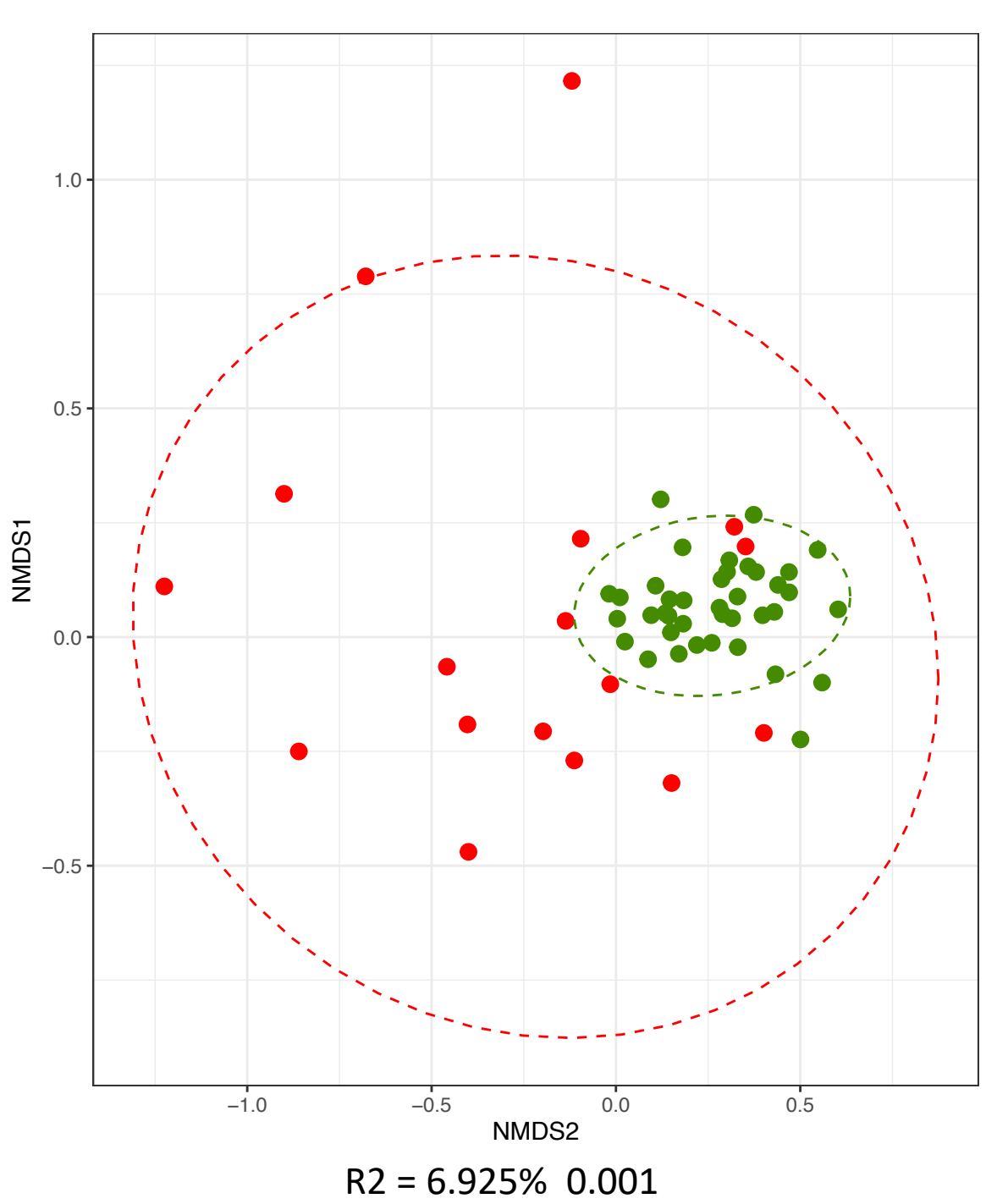


Phyla

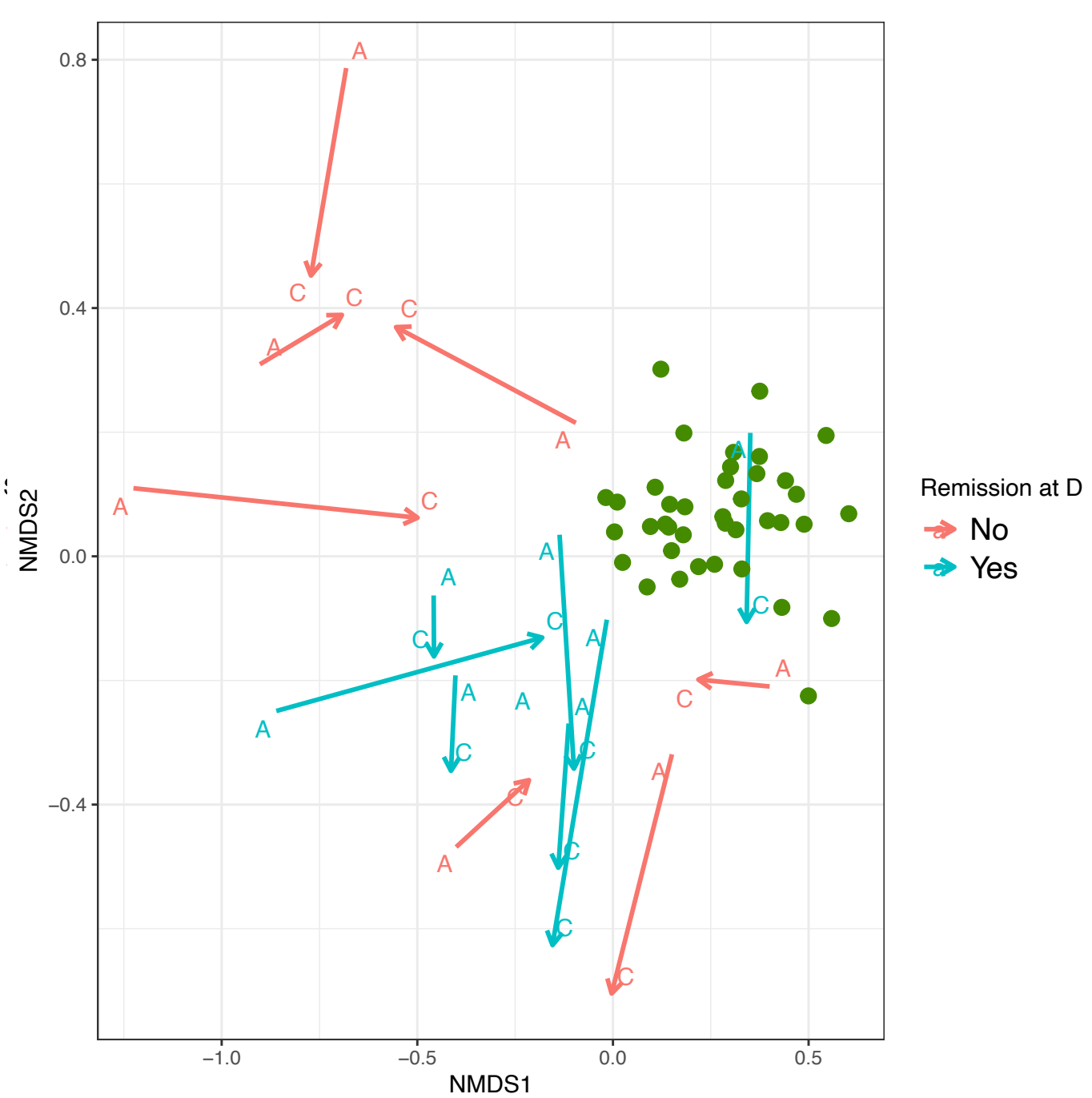
- Firmicutes (190)
- Bacteroidetes (23)
- Actinobacteria (22)
- Proteobacteria (2)
- Verrucomicrobia (1)
- None (2)

Families

- Lachnospiraceae (56)
- Ruminococcaceae (55)
- Clostridiaceae (18)
- Eggerthellaceae (12)
- Erysipelotrichaceae (12)
- Oscillospiraceae (12)
- Clostrid. XIII (11)
- Rikenellaceae (8)
- Peptostreptococcaceae (5)
- Lactobacillaceae (5)
- Prevotellaceae (4)
- Bacteroidaceae (4)
- Acidaminococcaceae (3)
- Peptoniphilaceae (3)
- Actinomycetaceae (3)
- Hungateiclostridiaceae (3)
- Coriobacteriaceae (2)
- Bifidobacteriaceae (2)
- Tannerellaceae (2)
- Veillonellaceae (2)
- Streptococcaceae (2)
- Desulfovibrionaceae (1)
- Enterococcaceae (1)
- Odoribacteraceae (1)
- Porphyromonadaceae (1)
- Barnesiellaceae (1)
- Propionibacteriaceae (1)
- Akkermansiaceae (1)
- Atopobiaceae (1)
- Sutterellaceae (1)
- Dietziaceae (1)
- None (6)

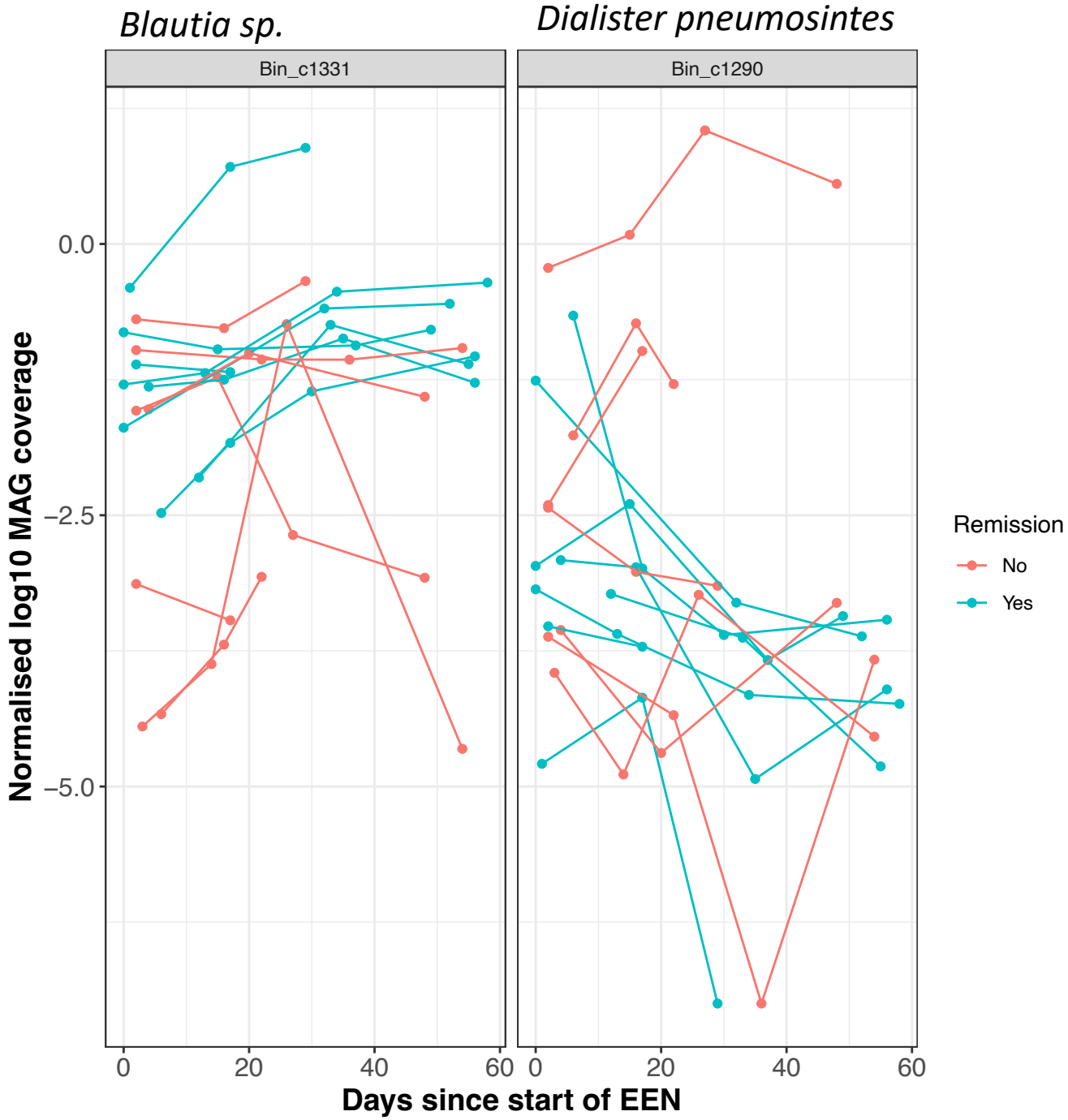

NMDS1
NMDS2
1.0
0.5
0.0
−0.5
−1.0
−0.5
0.0
0.5

R2 = 6.925% 0.001


NMDS2
NMDS1
0.8
0.4
0.0
−0.4
−1.0
−0.5
0.0
0.5
Remission at D
No
Yes

Changes in abundance during treatment

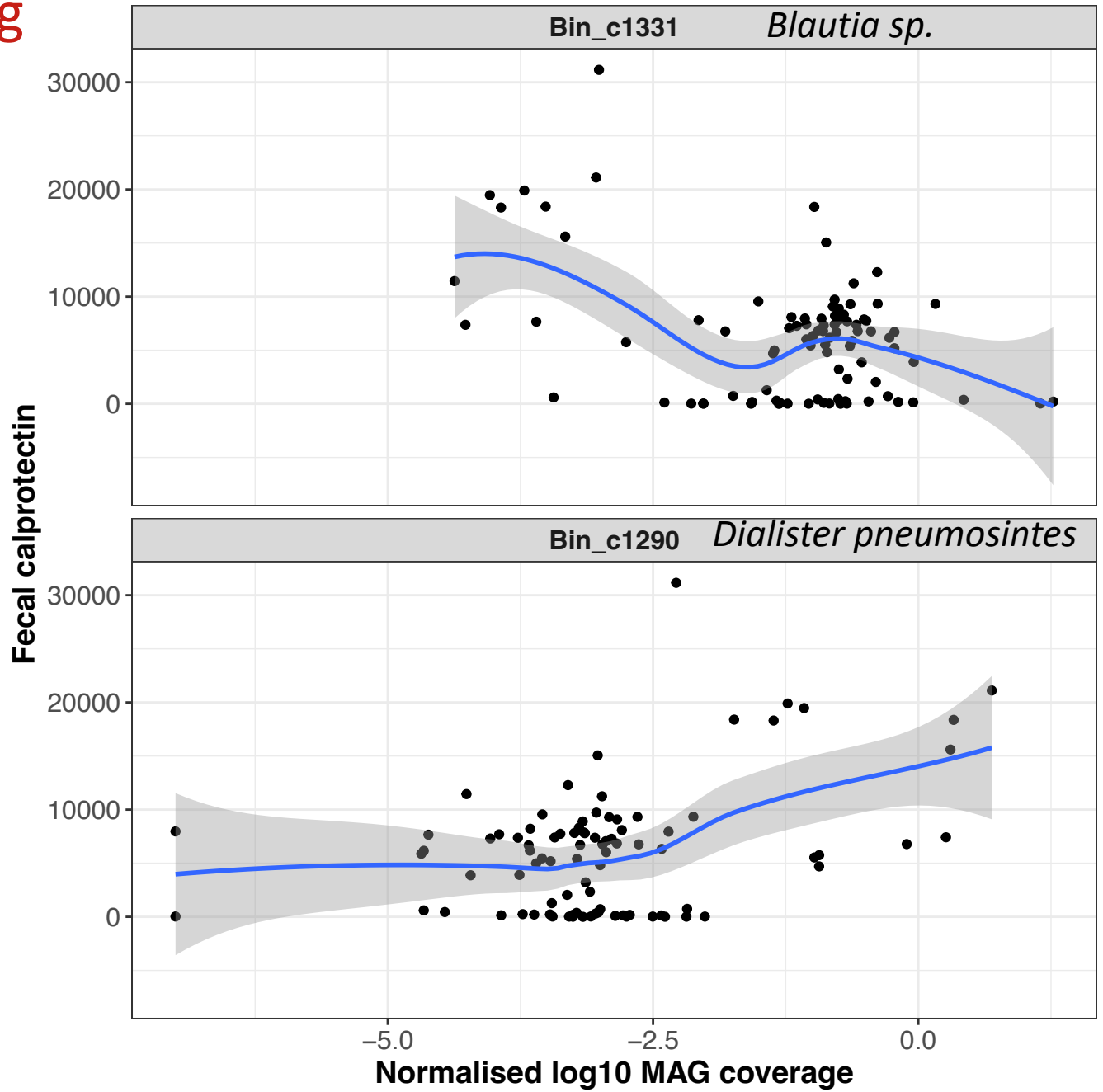
- Looked for consistent trends in abundance across subjects that entered remission
- 35 out of 268 MAGs changed significantly
- Four increased!

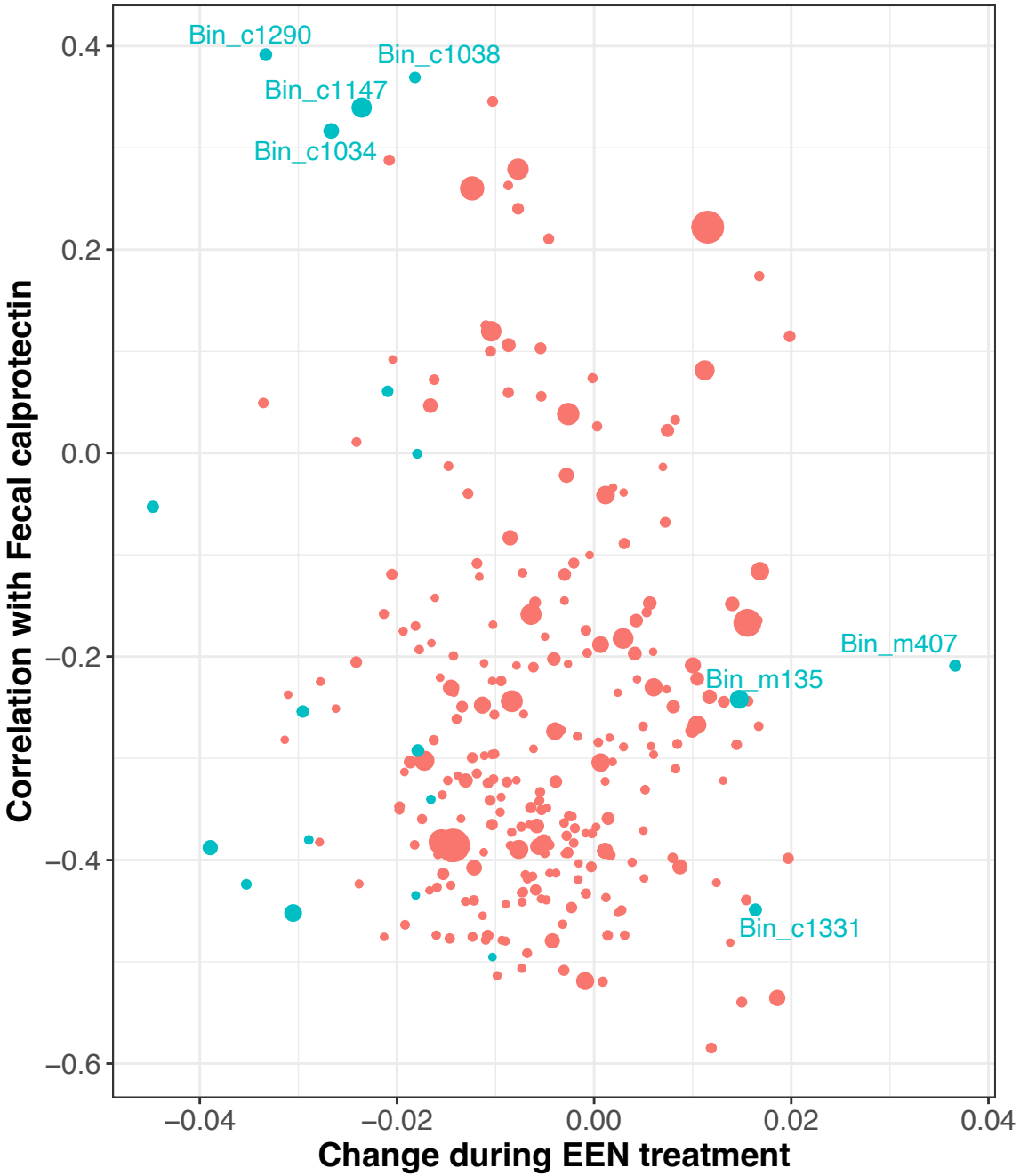


Correlation with calprotectin during treatment

- MAGs can explain 48.99% of variation in Calprotectin
- 236 MAGs q < 0.05 but only 9 positively

	p	pa	r	Genus
1	1.16e−09	3.11e−07	−0.585	g__UBA7160
2	3.39e−08	4.03e−06	−0.540	g__Clostridium_A
3	4.51e−08	4.03e−06	−0.536	g__Ruminiclostridium_C
4	1.31e−07	7.36e−06	−0.520	g__UBA7096
5	1.37e−07	7.36e−06	−0.519	g__Fusicatenibacter
6	1.93e−07	8.61e−06	−0.514	g__UBA11524
7	2.68e−07	1.02e−05	−0.508	g__CAG−56
8	3.04e−07	1.02e−05	−0.506	g__Eubacterium_F
9	5.99e−07	1.78e−05	−0.495	g__QAMH01
10	7.51e−07	2.01e−05	−0.492	g__Roseburia
11	1.24e−04	4.62e−04	0.392	g__Dialister_A
12	3.16e−04	9.31e−04	0.369	g__Abiotrophia
13	7.93e−04	1.93e−03	0.346	g__Gemella
14	9.94e−04	2.32e−03	0.340	g__Peptostreptococcus
15	2.24e−03	4.61e−03	0.317	g__Lancefieldella

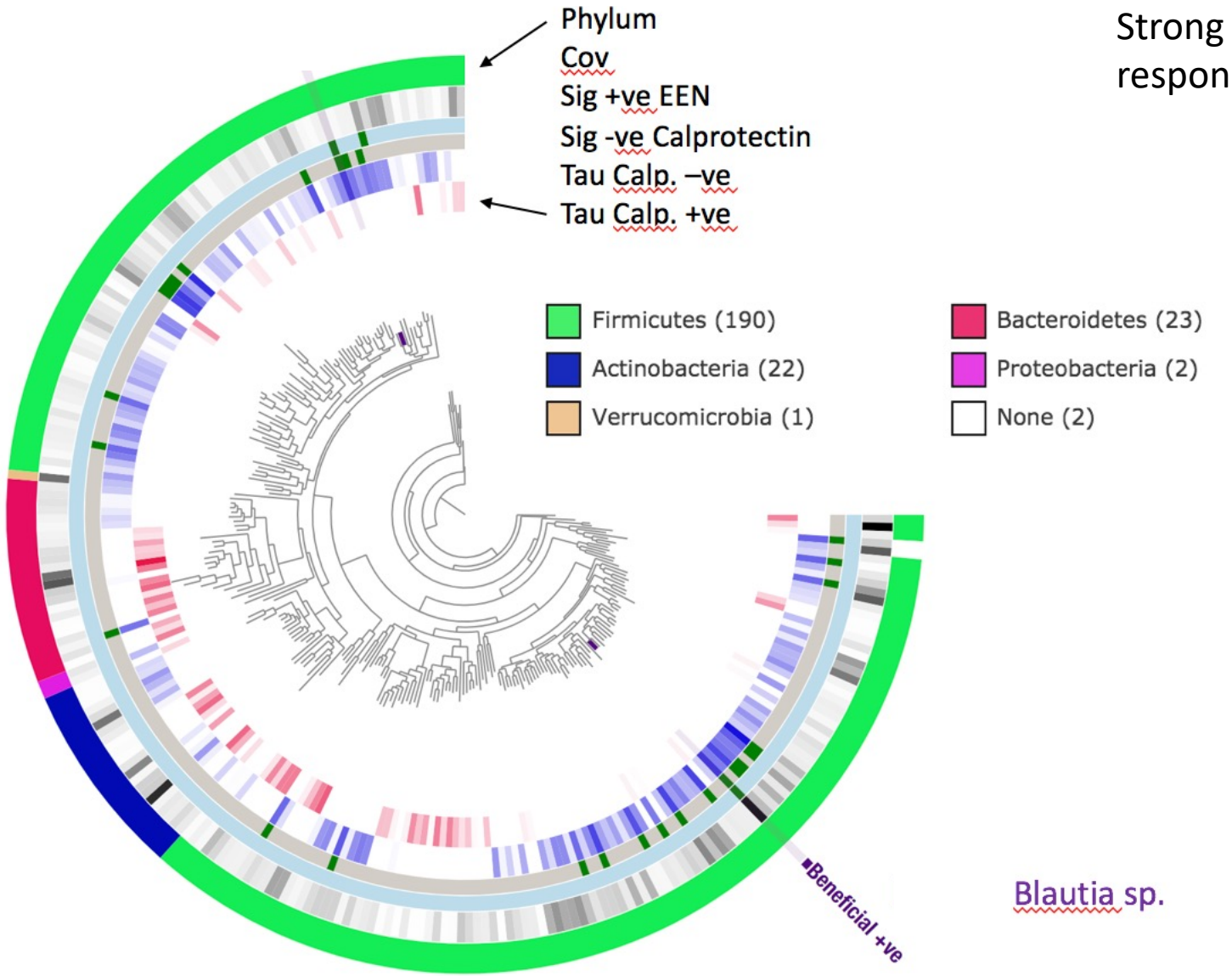




Bin_c1290 Dialister pneumosintes, Bin_c1147 Peptostreptococcus stomatis, Bin_c1038 Abiotrophia sp., Bin_c1034 Lancefieldella sp000564995

Bin_c1331 Blautia sp.

Earlham Institute



Strong phylogenetic signal of response to inflammation …

Blautia sp.

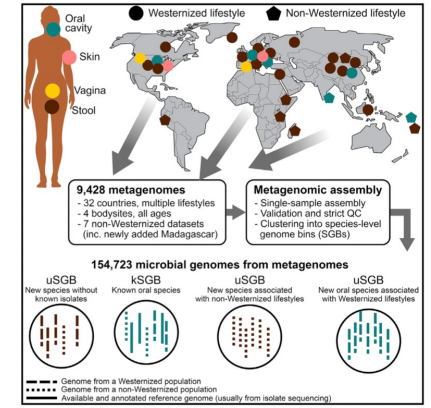
Overview – Large-scale AMR analysis

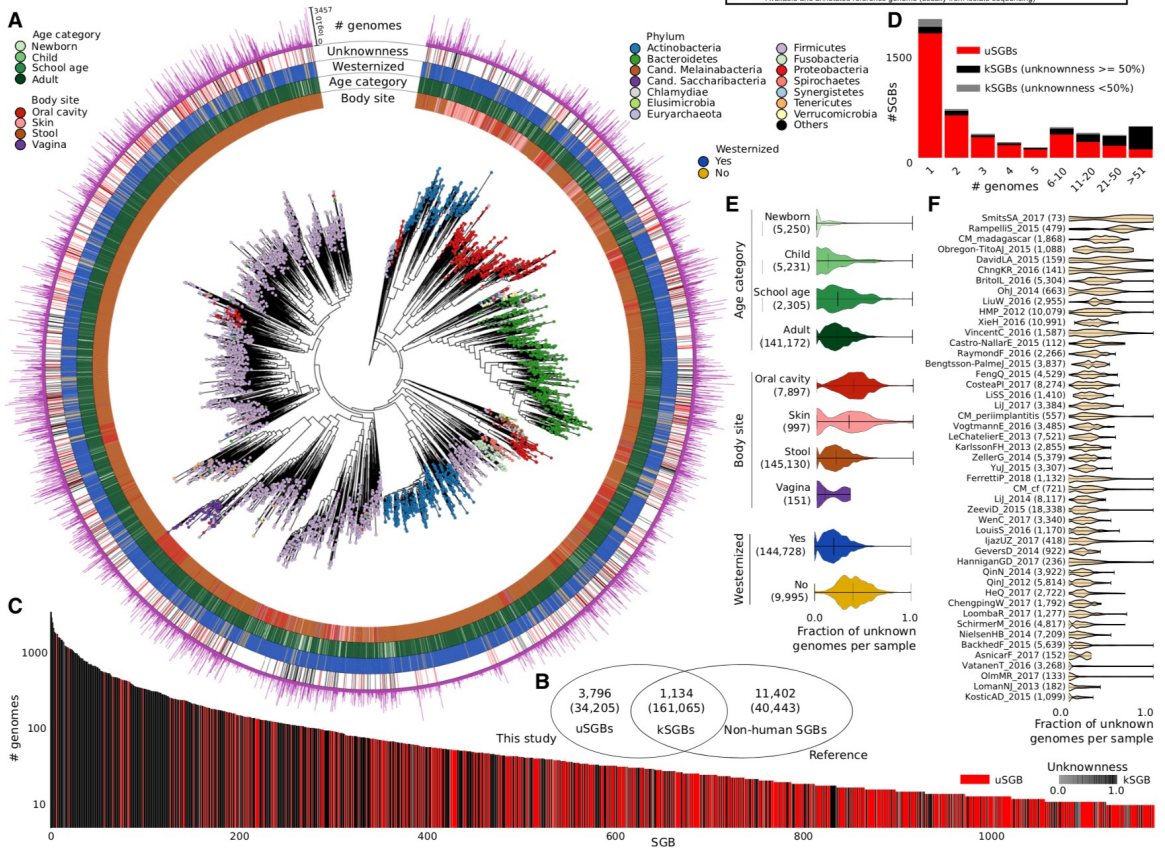
- Antibiotic resistance is a major threat to global health
- Microbiome is a potential reservoir of ARGs
- Numerous studies on short-term impact of antibiotics on the gut microbiome
- Hypothesis that on the population-level antibiotic consumption may impact mobilome in healthy individuals not taking antibiotics
- Similar to earlier studies e.g. Forslund et al. 2014 but at far greater scale using 8,972 microbiome samples
- Link to large-scale Pasolli et al. MAG collection to associate ARG to taxa and quantify mobilisation

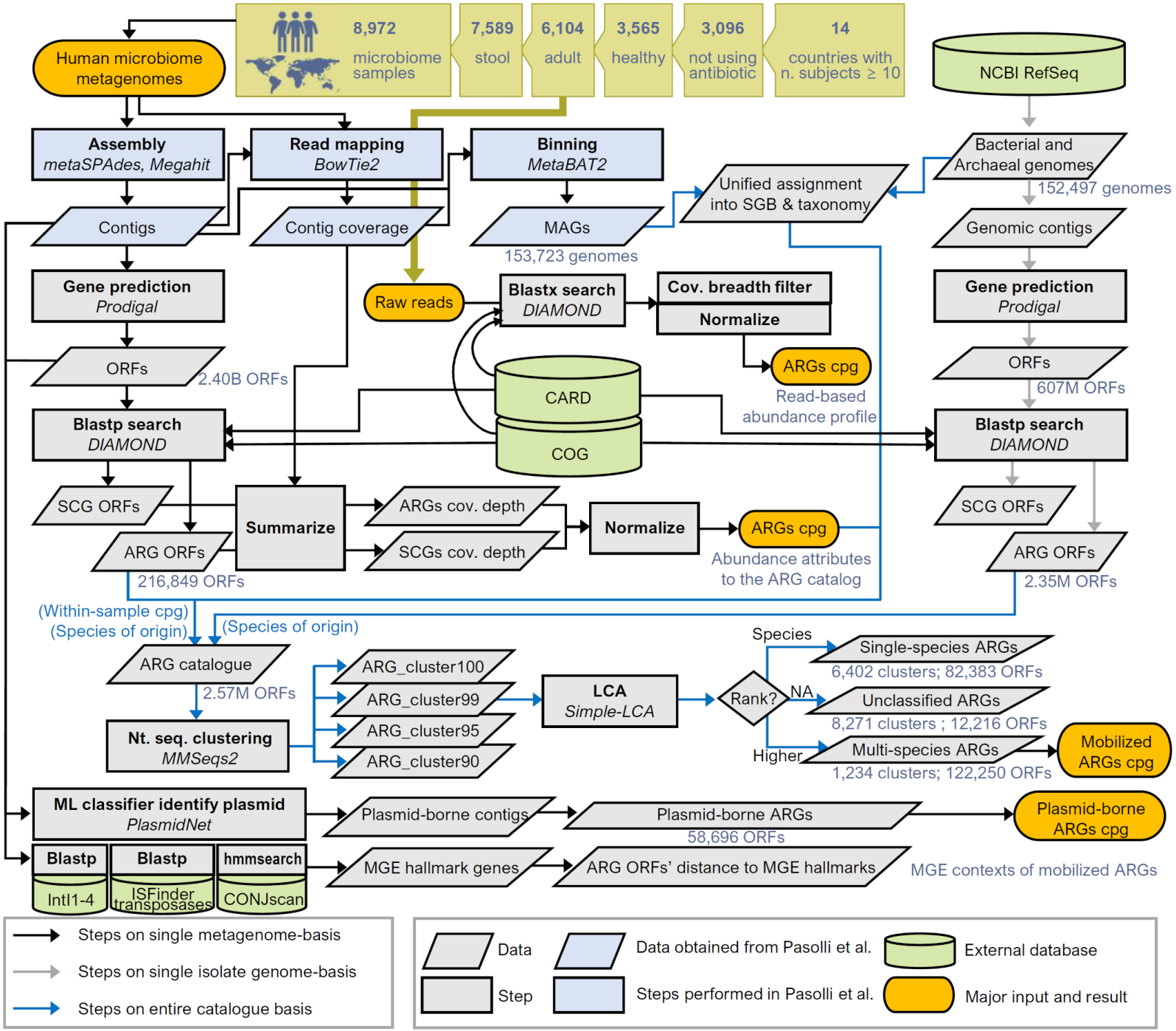
Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle

Authors

Edoardo Pasolli, Francesco Asnicar, Serena Manara, ..., Christopher Quince, Curtis Huttenhower, Nicola Segata







Dr Kihyun Lee – Chung-An University

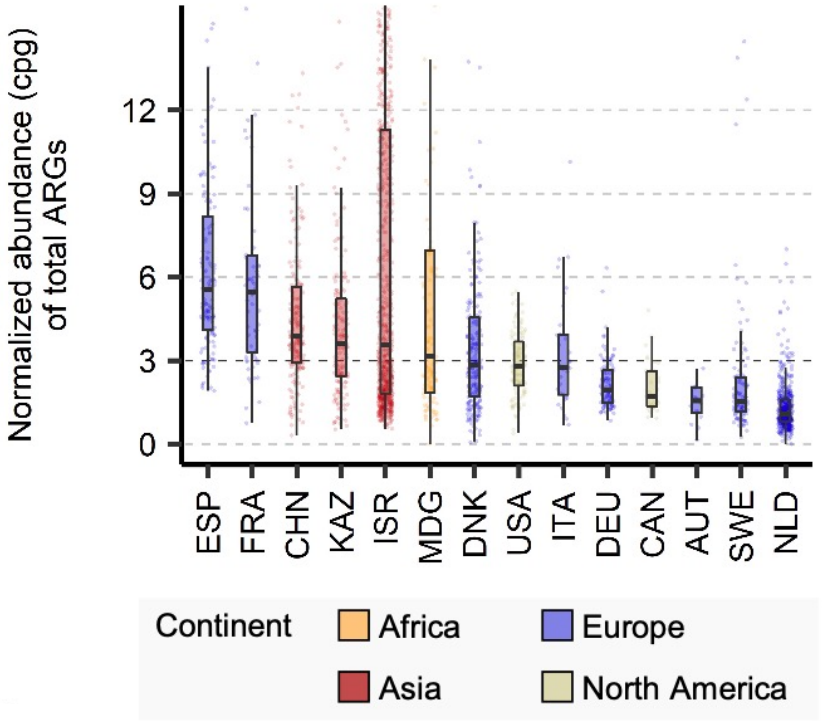
152,497 isolate genomes
153,723 MAGs
2.35M reference ARG ORFs

2.4 billion metagenome ORFs
216,849 ARG metagenome ORFs

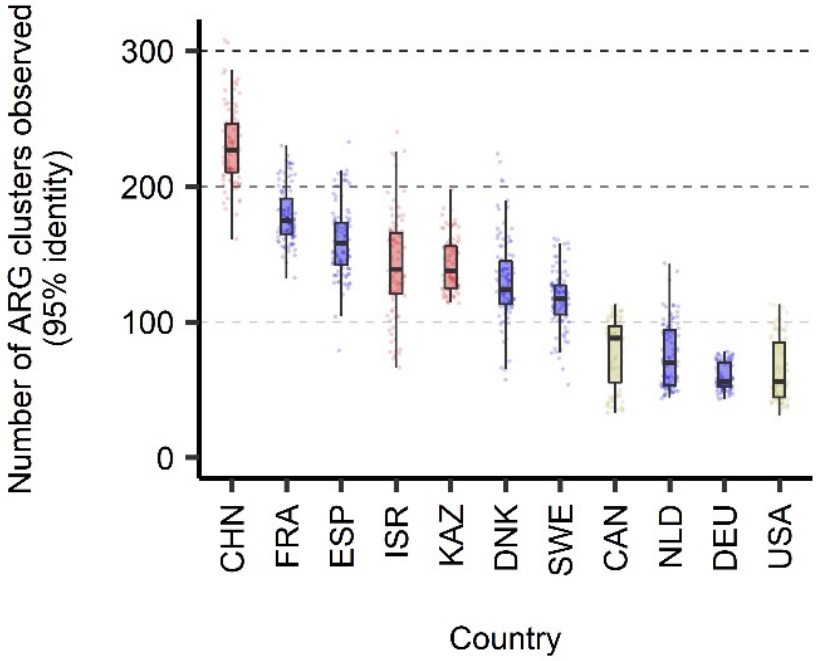
Samples span 20 countries

Earlham Institute

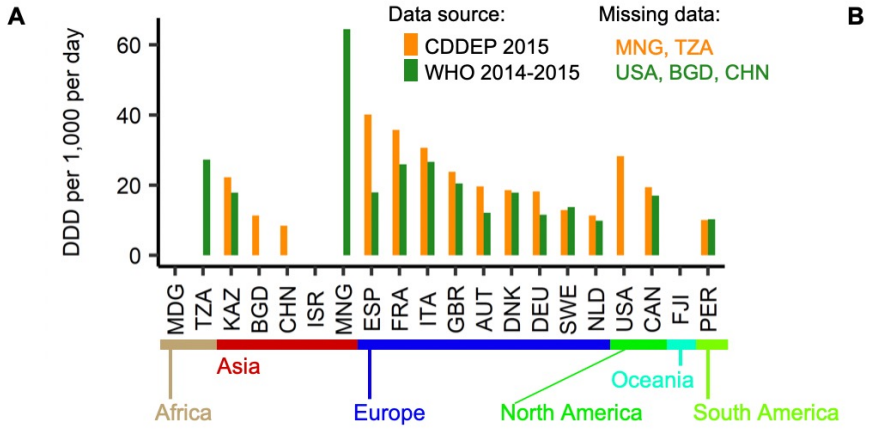
ARG richness and abundance in healthy antibiotic free human gut samples varies across countries (n = 3096)

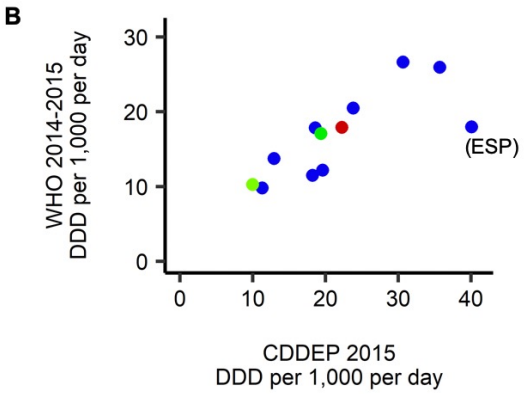


- Highest Spain median 5.56
- Lowest Netherlands median 1.08
- 5.5-fold difference
- Kruskal Wallis p value = 1.6e − 65



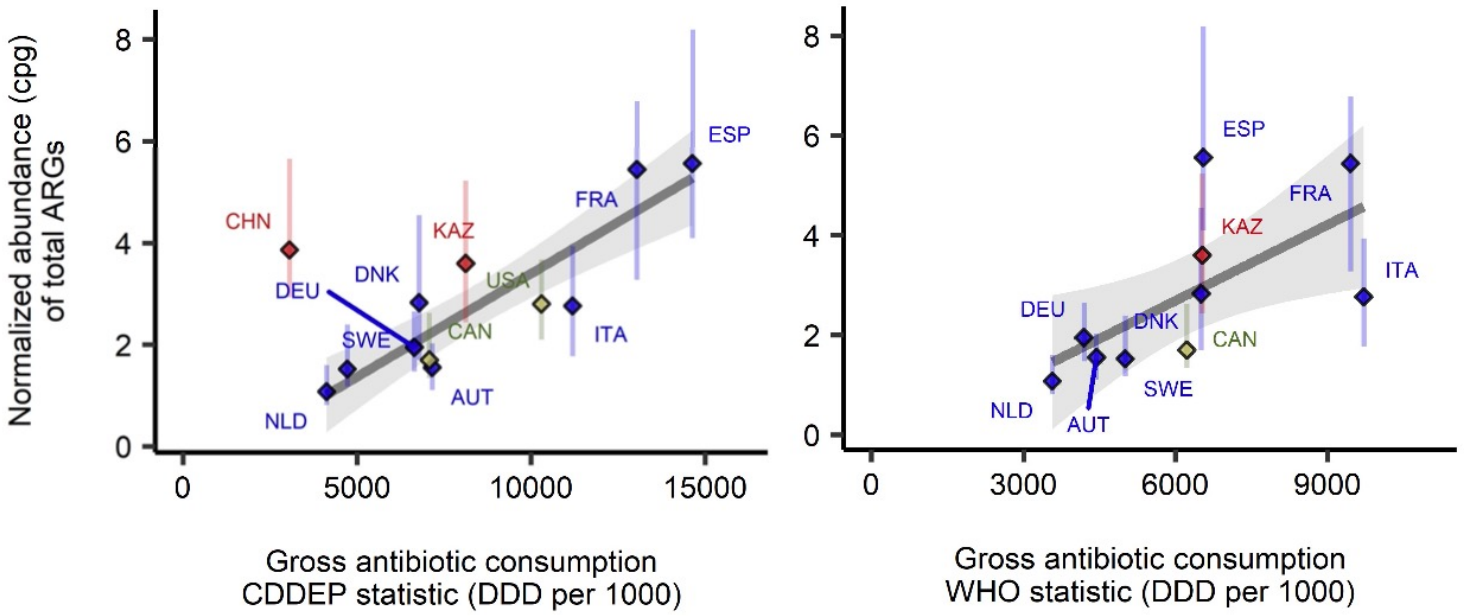
- Highest China median 227
- Lowest USA median 56
- 4.1-fold difference
- Kruskal Wallis p value = 7.7E-183





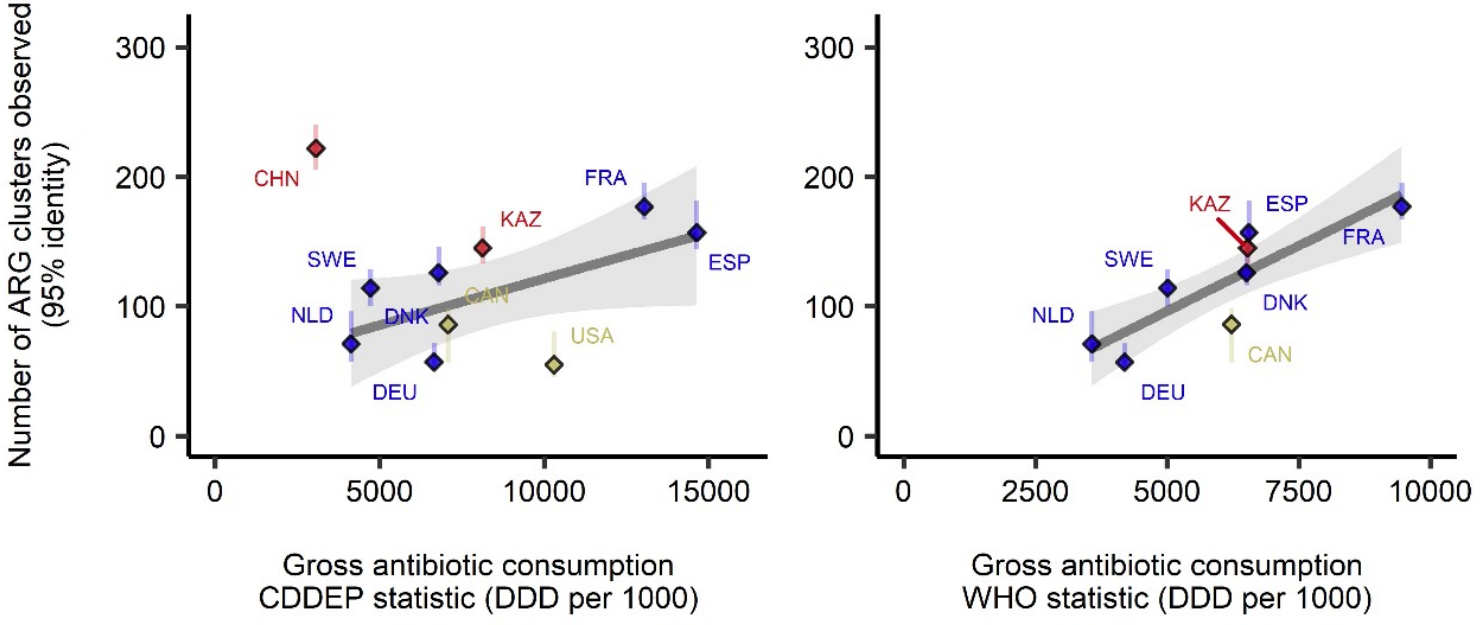
r = 0.89, p = 0.23e-5 (exc. China)

r = 0.65, p = 0.04



r = 0.57, p = 0.11

r = 0.85, p = 0.0073

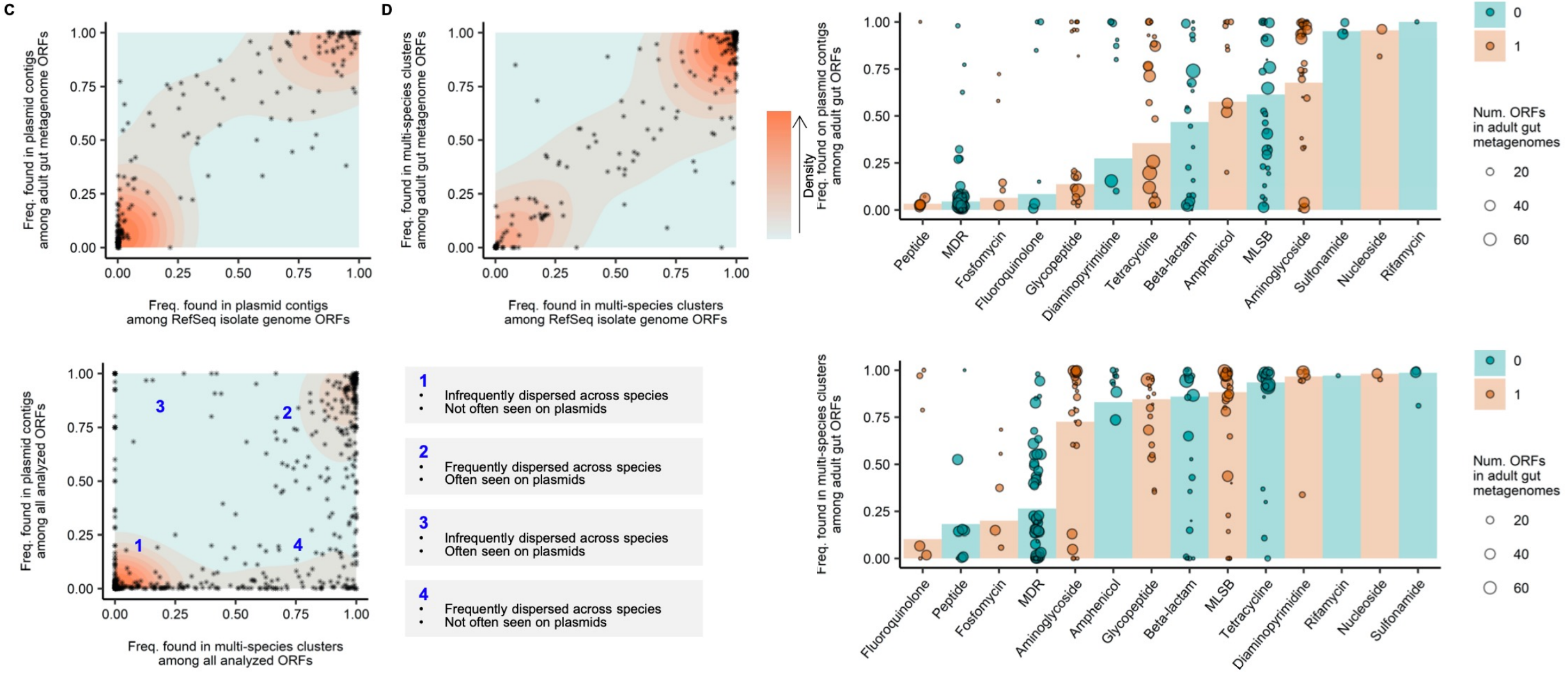


When separated by antibiotic class only observed significant correlations for beta-lactams

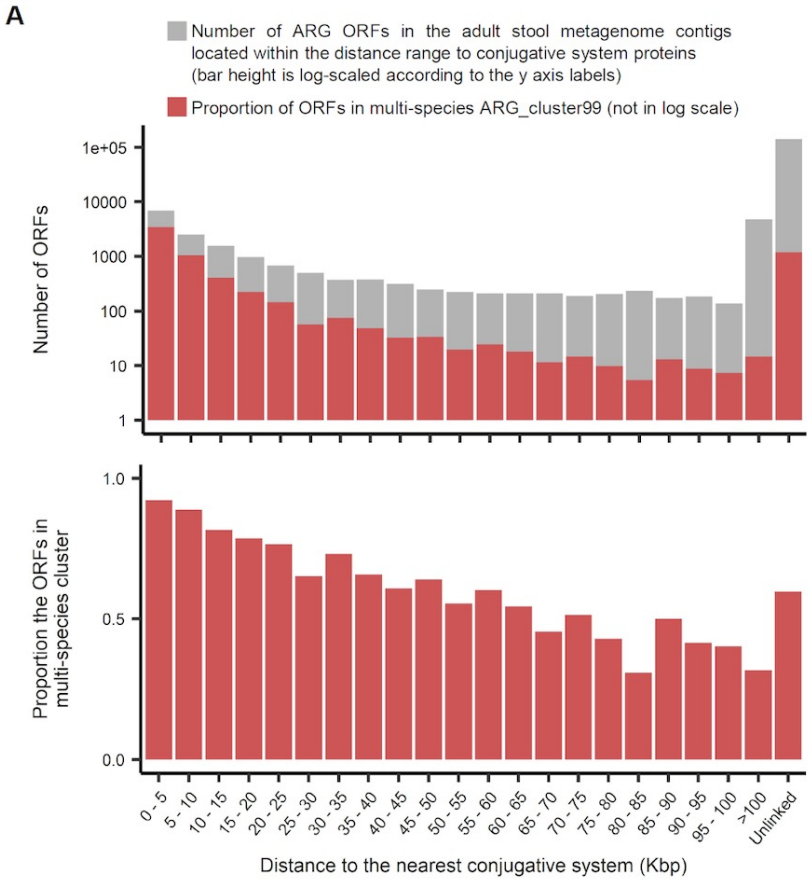
Identifying mobilome

- Can assign 54.5% of 99% ANI ARG clusters to MAGs or RefSeq genomes
- 10% of these were assigned to multiple species – hence mobile
- Mobile multi-species ARGs constituted the majority of each individual's gut resistome: 87% of within-sample ARG richness and 96% of the total ARG abundance (cpg) per individual.
- Combine with machine learning method that uses contig sequence characteristics to classify as plasmid or genomic
- To identify ARGs possibly mobilised by other MGEs, we searched contigs for hallmark genes of:
  - Insertion sequence (IS) elements
  - Conjugative elements (e.g., ICEs)
  - Class 1 integrons

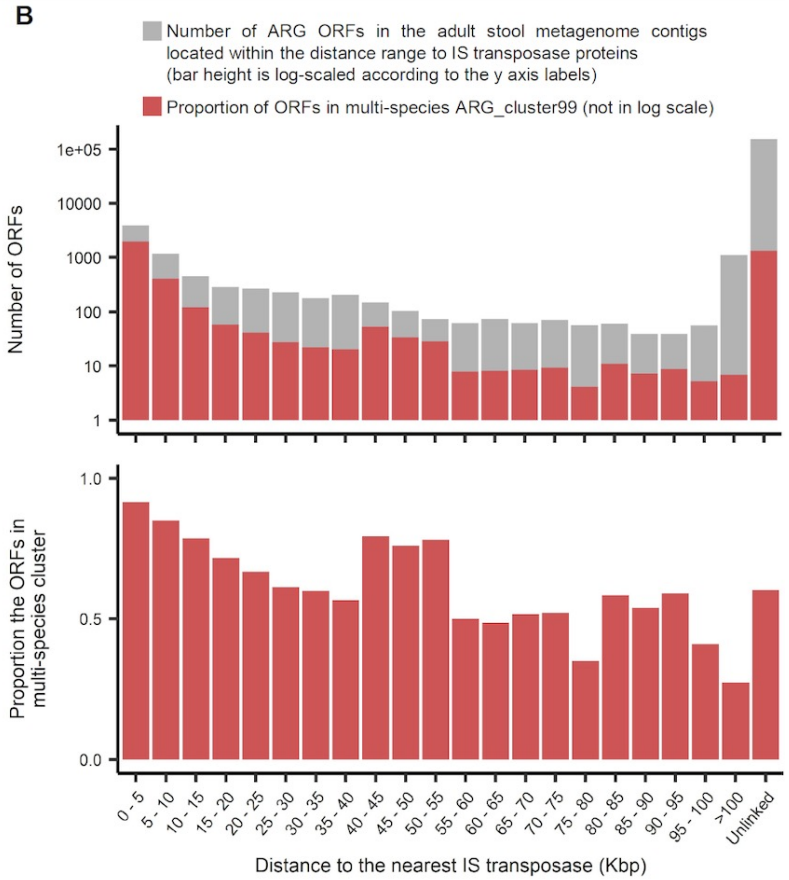
Comparison of multi-species vs. plasmid ARGs across families



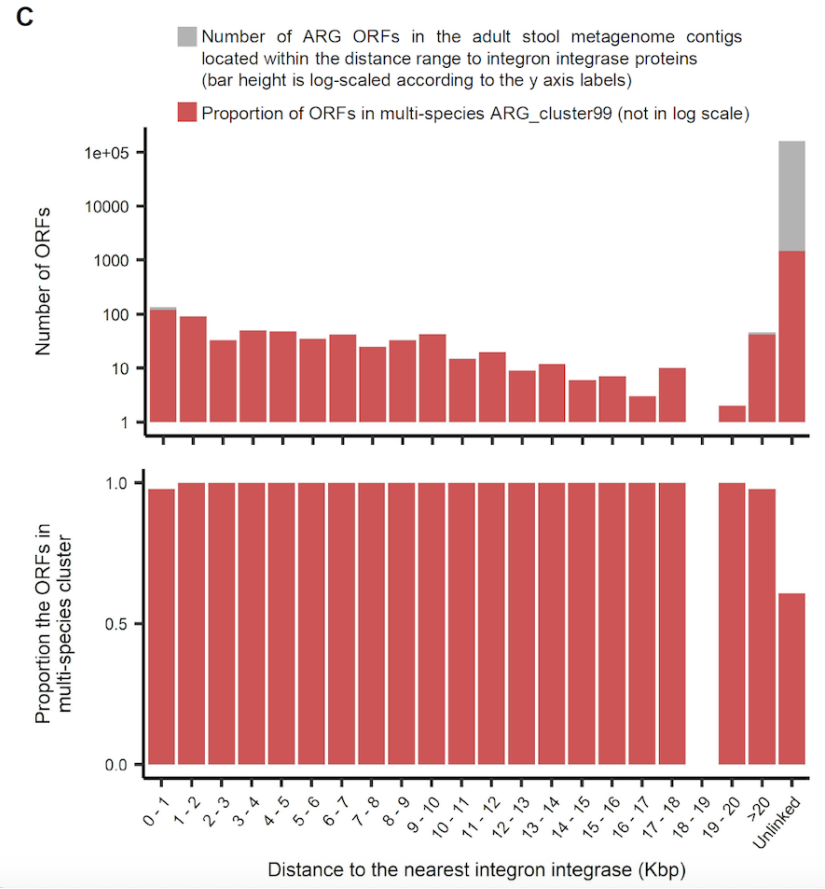
Conjugative systems



Transposase

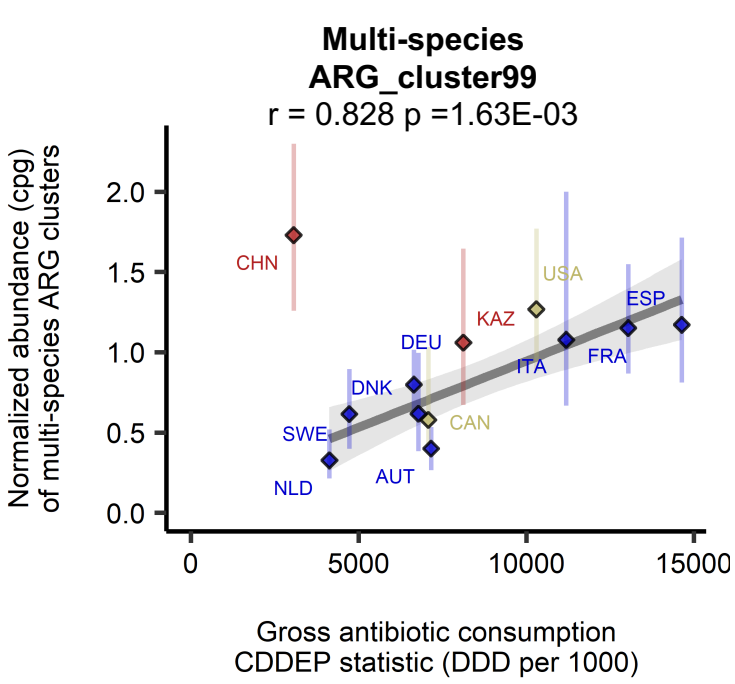


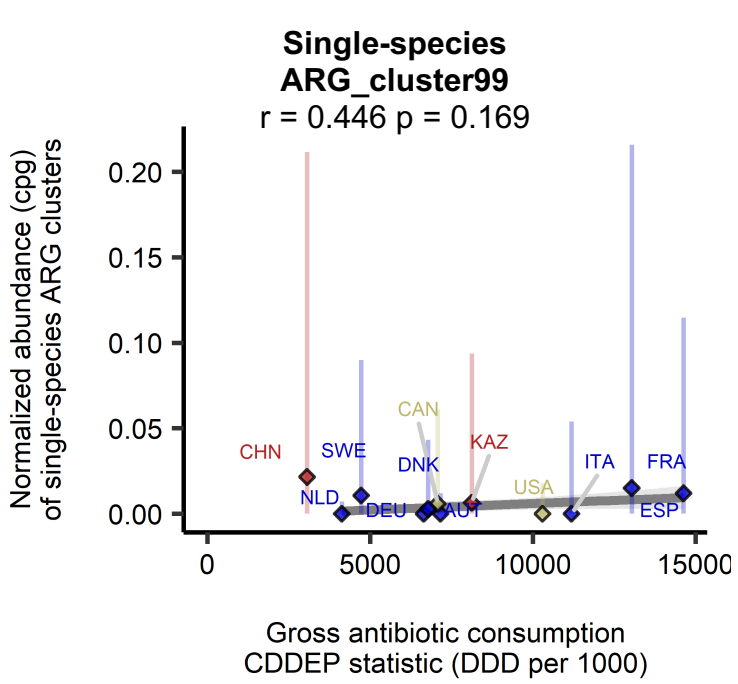
Integrons

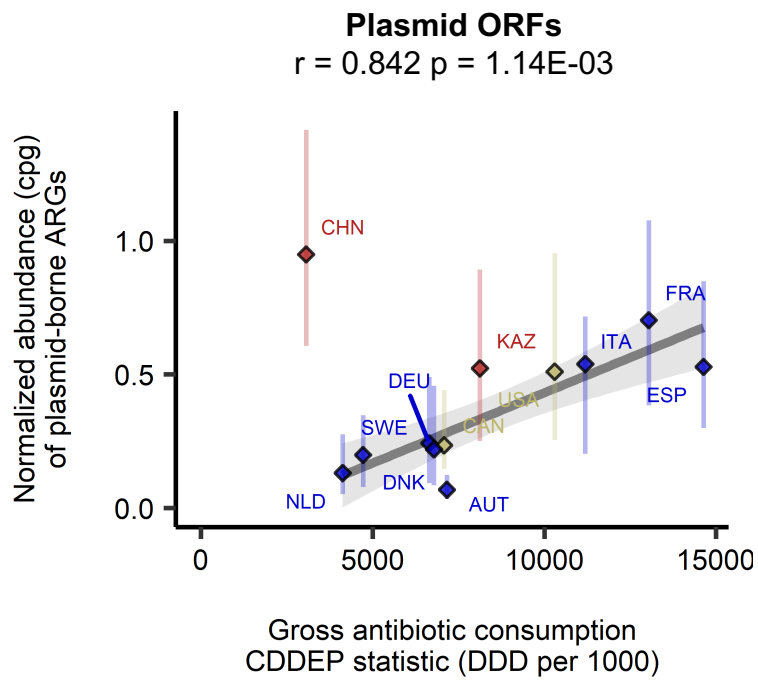


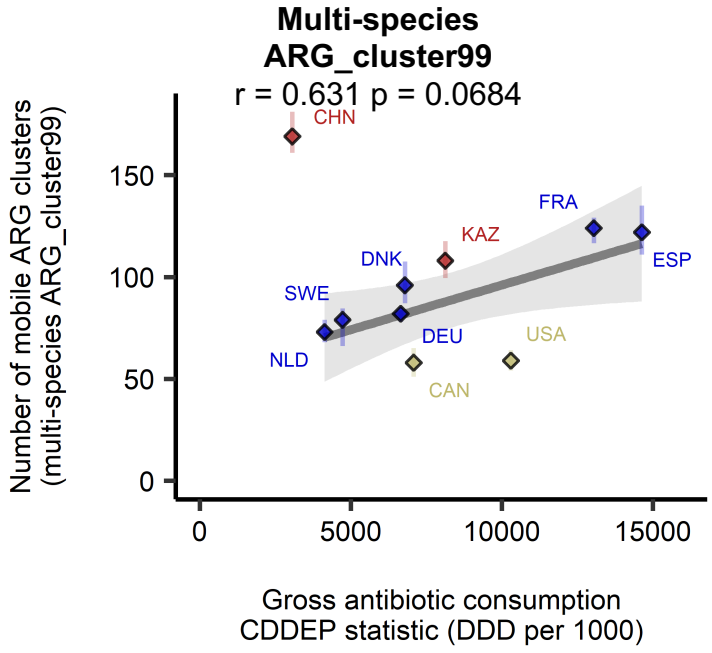
- 36.2% of the adult gut resistome ORFs were associated with at least one type of MGE
  - 27.1% on plasmids; 4.4% IS transposases; 10.2% conjugative systems; 0.3% integrons
- 89.1% of the plasmid-borne ARGs were found within multi-species clusters vs 49.0% of non-plasmid-borne ARGs, 80.9% of ICE-associated, 82.1% IS transposases, 99.4% of the ARGs located ≤ 10 Kbp from integron integreases

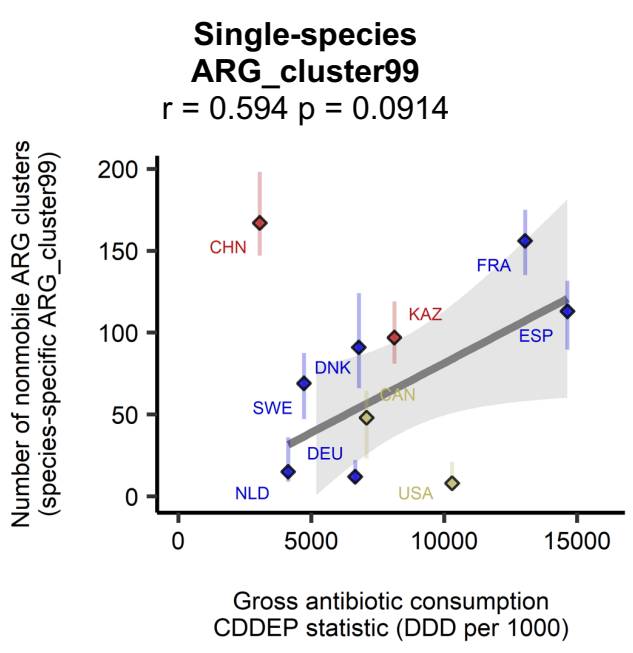
Earlham Institute

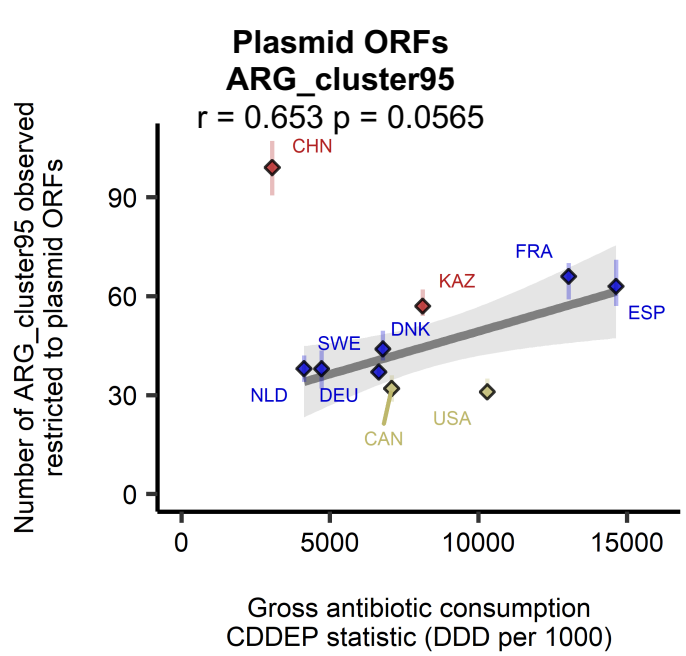

Multi-species
ARG_cluster99
r = 0.828 p =1.63E-03
Normalized abundance (cpg)
of multi-species ARG clusters
Gross antibiotic consumption
CDDEP statistic (DDD per 1000)


Single-species
ARG_cluster99
r = 0.446 p = 0.169
Normalized abundance (cpg)
of single-species ARG clusters
Gross antibiotic consumption
CDDEP statistic (DDD per 1000)


Plasmid ORFs
r = 0.842 p = 1.14E-03
Normalized abundance (cpg)
of plasmid-borne ARGs
Gross antibiotic consumption
CDDEP statistic (DDD per 1000)

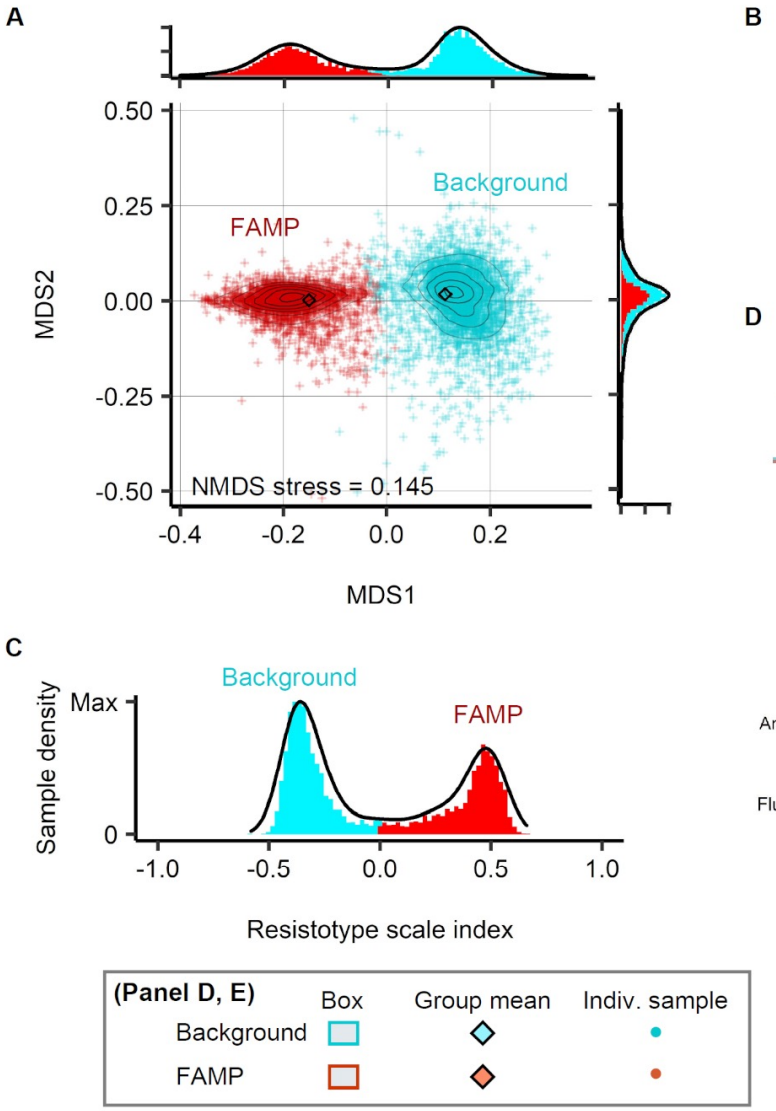

Multi-species
ARG_cluster99
r = 0.631 p = 0.0684
Number of mobile ARG clusters
(multi-species ARG_cluster99)
Gross antibiotic consumption
CDDEP statistic (DDD per 1000)

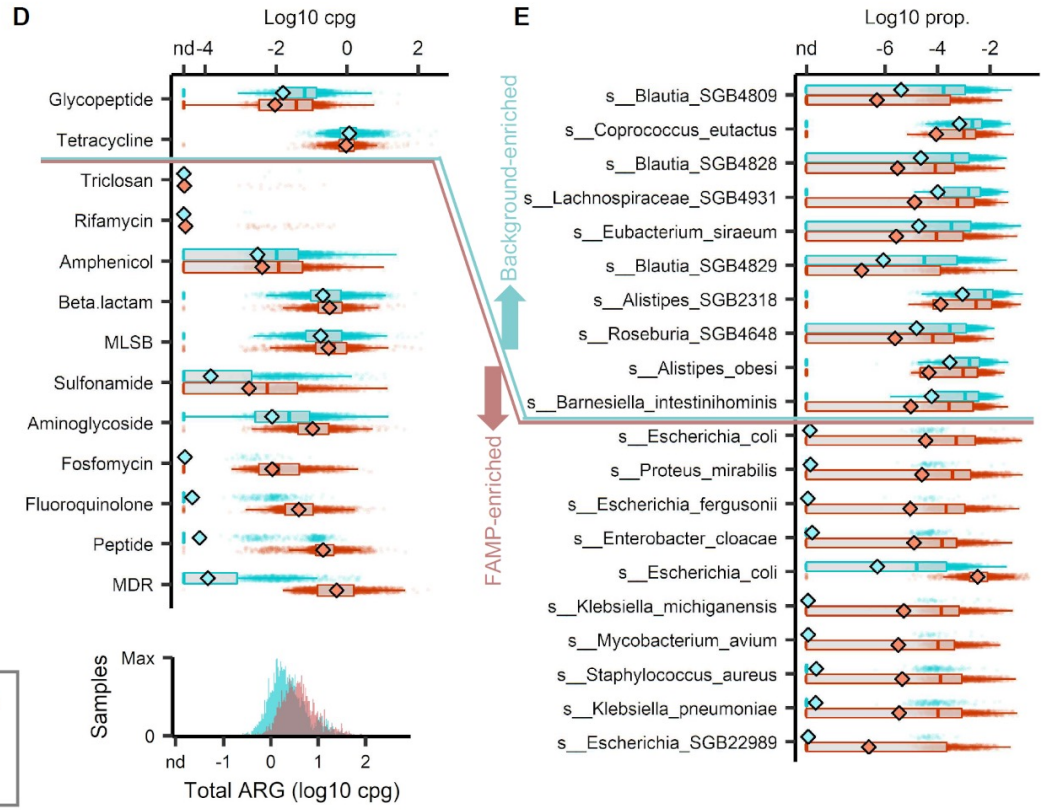

Single-species
ARG_cluster99
r = 0.594 p = 0.0914
Number of nonmobile ARG clusters
(species-specific ARG_cluster99)
Gross antibiotic consumption
CDDEP statistic (DDD per 1000)

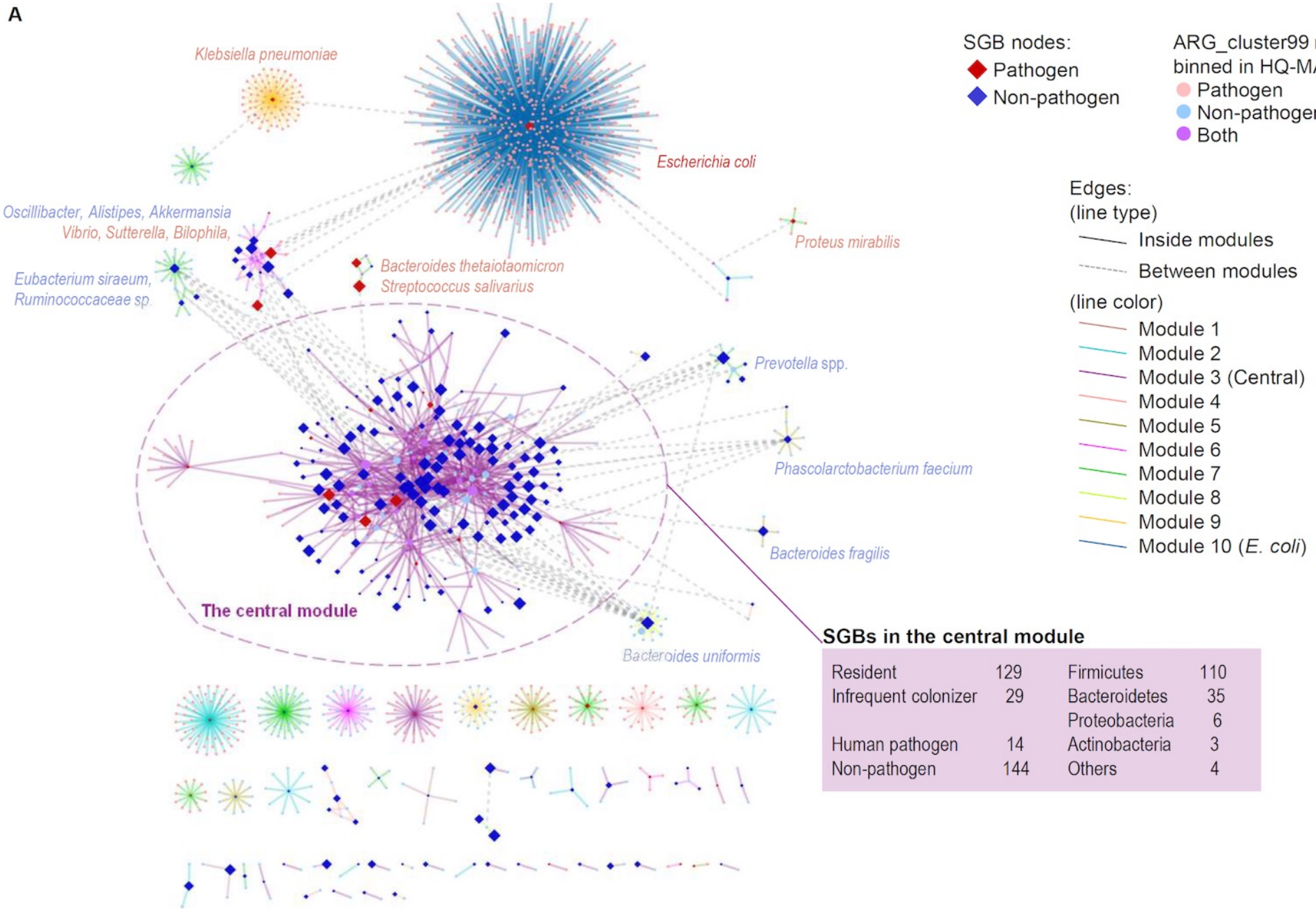

Plasmid ORFs
ARG_cluster95
r = 0.653 p = 0.0565
Number of ARG_cluster95 observed
restricted to plasmid ORFs
Gross antibiotic consumption
CDDEP statistic (DDD per 1000)

Resistotypes in the human gut microbiome

- Represent each sample as profile over 422 ARG families
- Observe two clusters of profiles
- Background in 55.4% of samples (66.1% of healthy)
- FAMP – enriched for Floroquinolone, Fosfomycin, Aminogycoside, Peptide and MDR ARGs





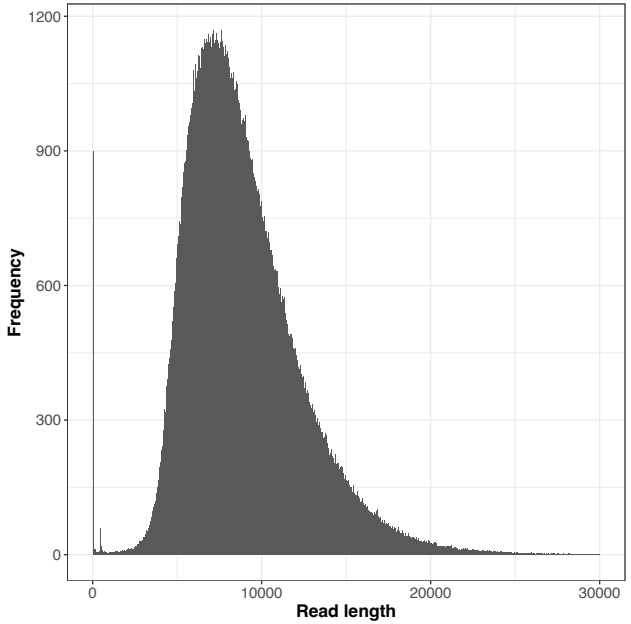


SGBs in the central module

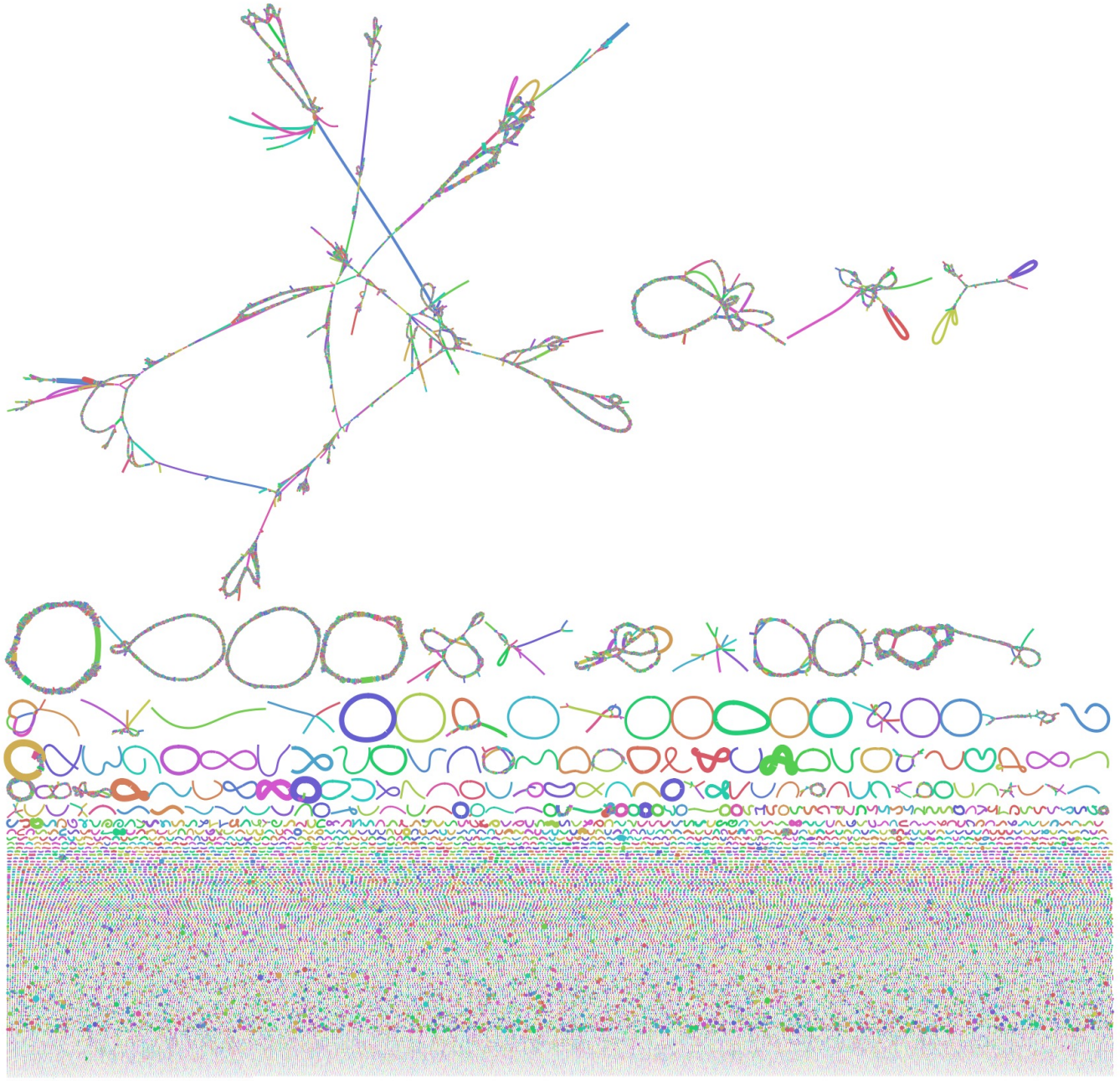
Resident	129	Firmicutes	110
Infrequent colonizer	29	Bacteroidetes	35
		Proteobacteria	6
Human pathogen	14	Actinobacteria	3
Non-pathogen	144	Others	4

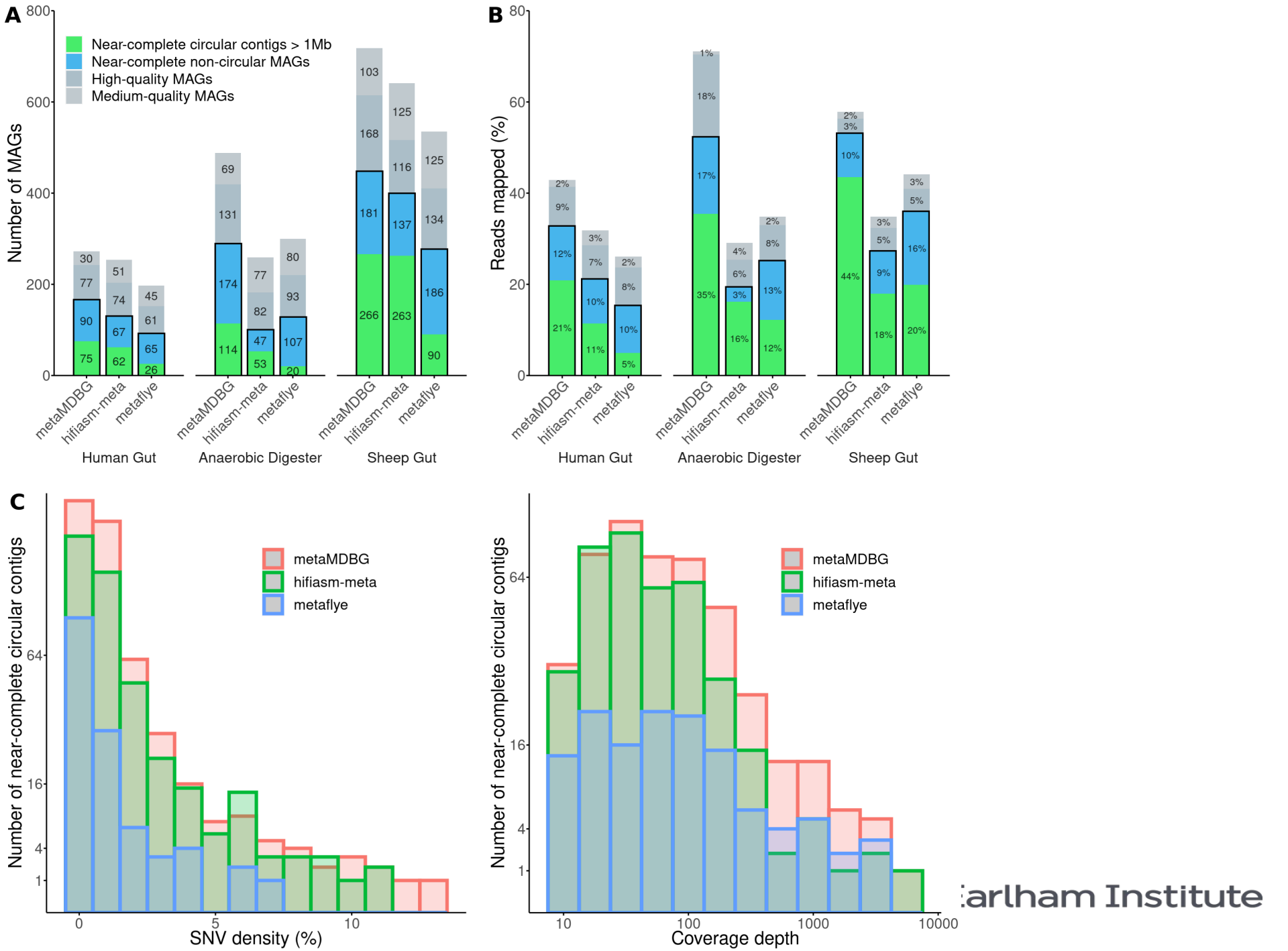
Assembling HiFi PacBio data AD2.1 – week 20 PacBio assembly

- 64.69 Gbp of sequence, mean length 8.8kbp



- Hifiasm-meta assembly:
  - sequence #: 91013 total length: 2.87 Gbp max length: 3,964,305 N50: 38,174 N90: 14691
- MetaFlye assembly
  - Sequence #: 27805 total length: 1.12 Gbp max length: 6,170,979 N50: 183,625 N90: 19,399




A
Near-complete circular contigs > 1Mb
Near-complete non-circular MAGs
High-quality MAGs
Medium-quality MAGs
Number of MAGs
metaMDBG
hifiasm-meta
metaflye
Human Gut
Anaerobic Digester
Sheep Gut
B
Reads mapped (%)
C
Number of near-complete circular contigs
SNV density (%)
Coverage depth

Summary

- Automated binning enables high-throughput genome resolved metagenomics on individual studies
- Underlying principles are data transformation followed by clustering and bin evaluation to generate MAGs
- Can link MAG properties to community dynamics