Judit Szarvas

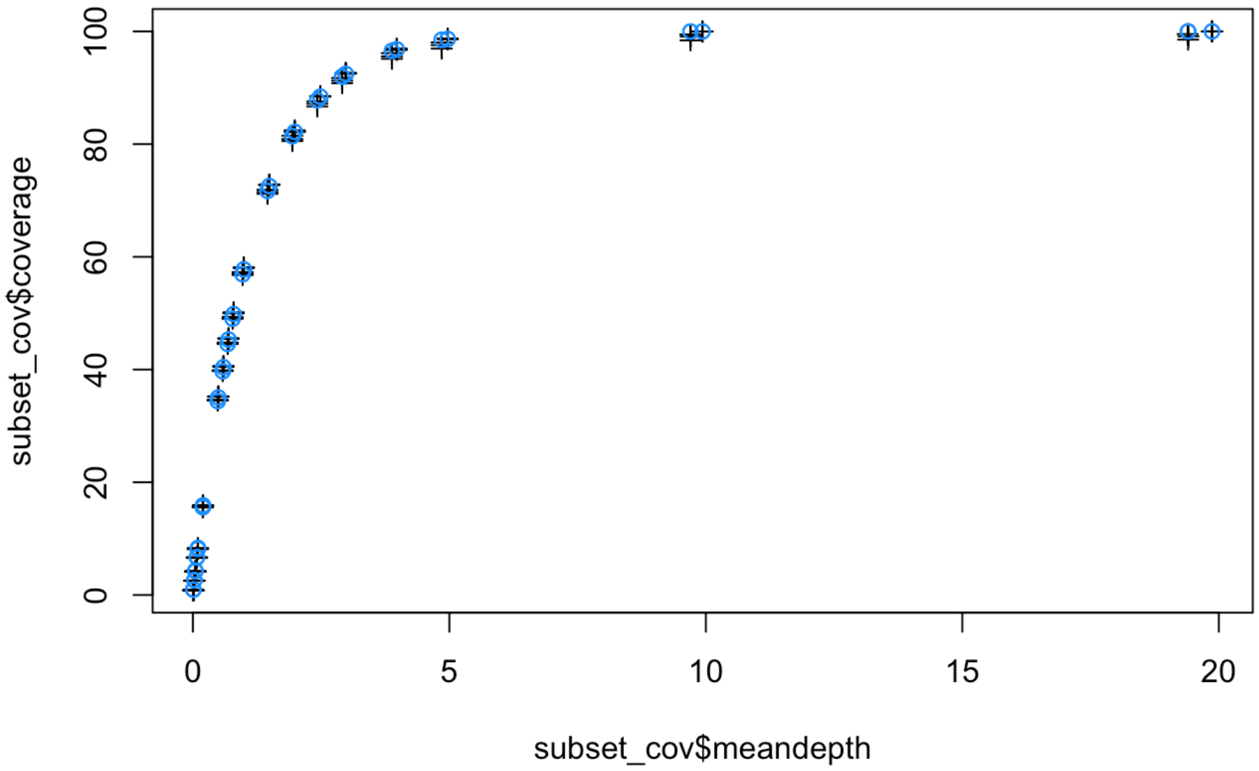
Is this species in this metagenomic sample? Part II: Benchmarking

Recap

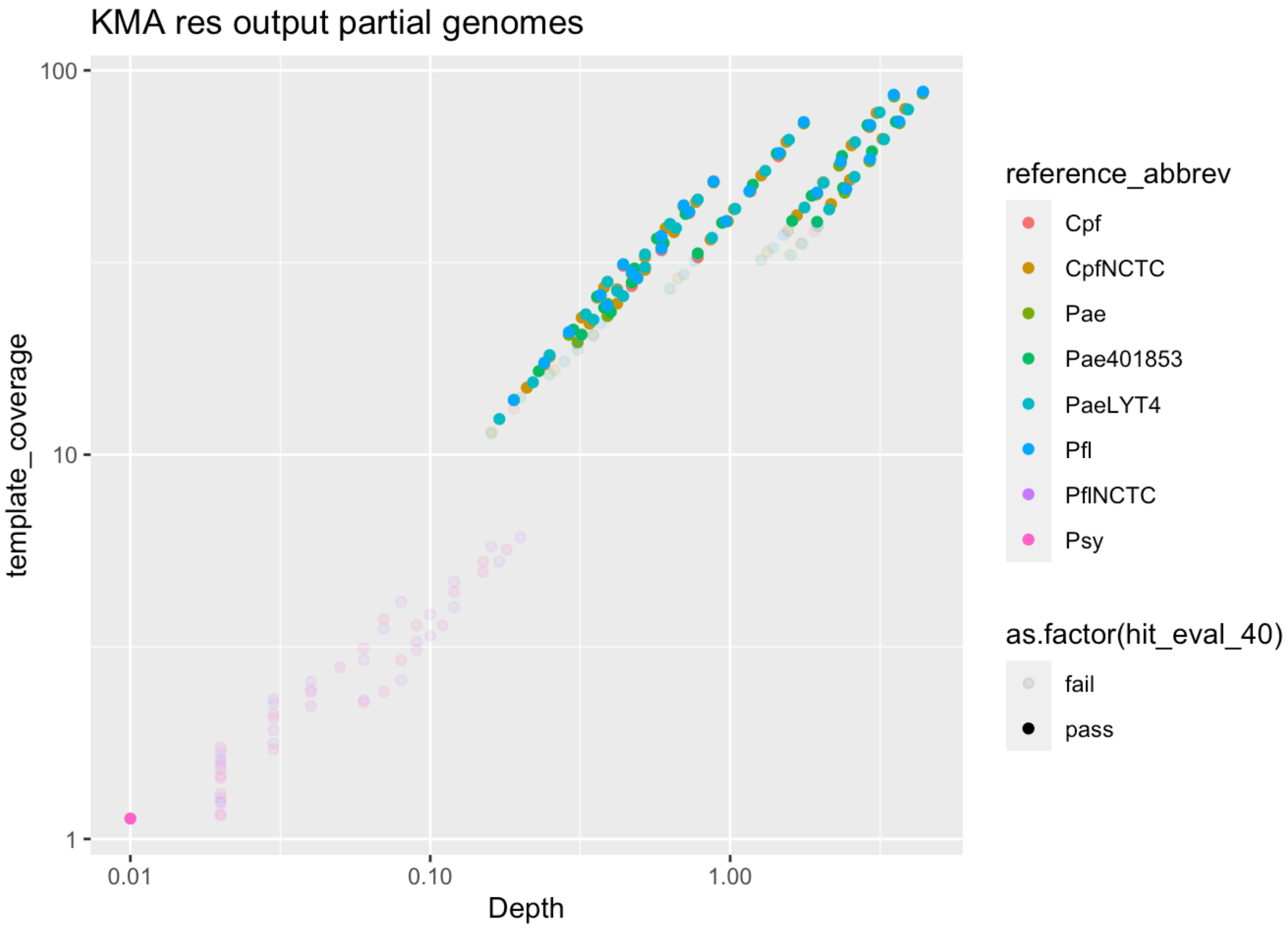
Species level classification using whole reference genomes

Read alignment coverage is a function of depth

$$coverage = (-0.9987808 * exp(-0.8670195 * depth) + 1) * 100$$



Filters out low ANI% alignments in metagenomes



Minimum depth required to avoid KMA rounding effects

96.7 ANI% lowest accepted hit

Benchmarking

How does it perform on true metagenomics samples?

Zymo D6331 gut metagenome mock community

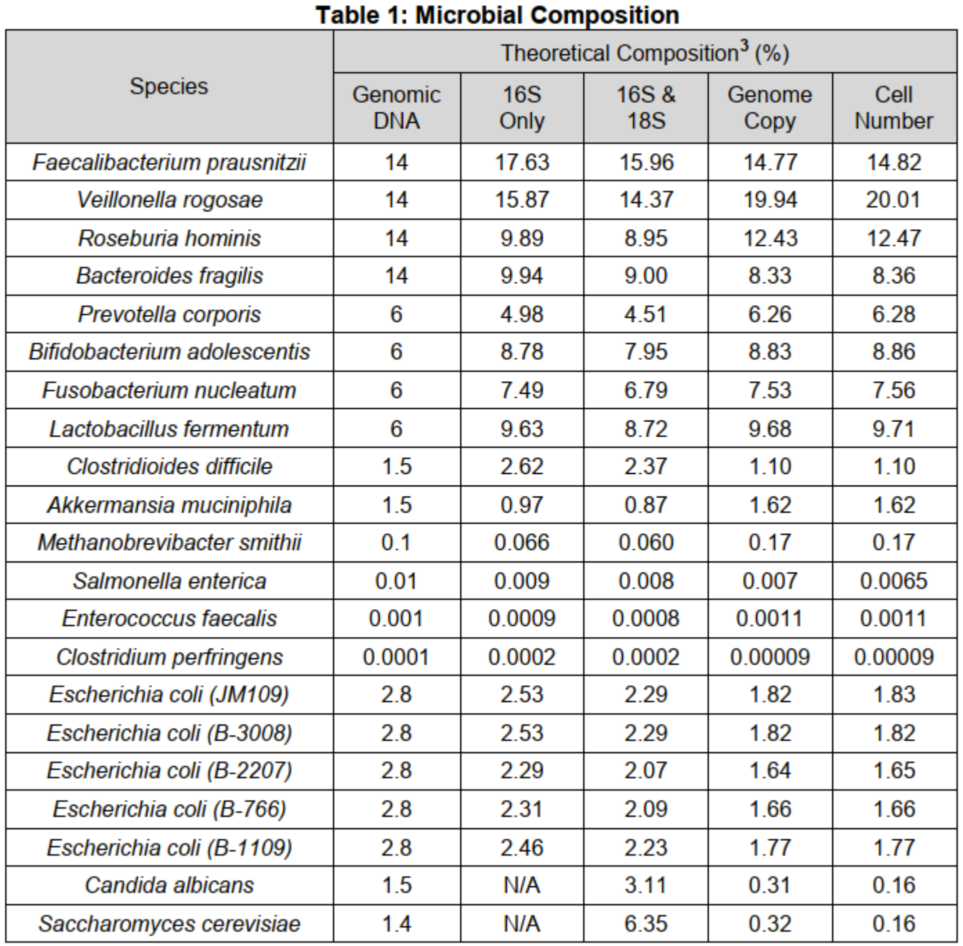


Table 1: Microbial Composition

Species	Theoretical Composition[3] (%)				
	Genomic DNA	16S Only	16S & 18S	Genome Copy	Cell Number
Faecalibacterium prausnitzii	14	17.63	15.96	14.77	14.82
Veillonella rogosae	14	15.87	14.37	19.94	20.01
Roseburia hominis	14	9.89	8.95	12.43	12.47
Bacteroides fragilis	14	9.94	9.00	8.33	8.36
Prevotella corporis	6	4.98	4.51	6.26	6.28
Bifidobacterium adolescentis	6	8.78	7.95	8.83	8.86
Fusobacterium nucleatum	6	7.49	6.79	7.53	7.56
Lactobacillus fermentum	6	9.63	8.72	9.68	9.71
Clostridioides difficile	1.5	2.62	2.37	1.10	1.10
Akkermansia muciniphila	1.5	0.97	0.87	1.62	1.62
Methanobrevibacter smithii	0.1	0.066	0.060	0.17	0.17
Salmonella enterica	0.01	0.009	0.008	0.007	0.0065
Enterococcus faecalis	0.001	0.0009	0.0008	0.0011	0.0011
Clostridium perfringens	0.0001	0.0002	0.0002	0.00009	0.00009
Escherichia coli (JM109)	2.8	2.53	2.29	1.82	1.83
Escherichia coli (B-3008)	2.8	2.53	2.29	1.82	1.82
Escherichia coli (B-2207)	2.8	2.29	2.07	1.64	1.65
Escherichia coli (B-766)	2.8	2.31	2.09	1.66	1.66
Escherichia coli (B-1109)	2.8	2.46	2.23	1.77	1.77
Candida albicans	1.5	N/A	3.11	0.31	0.16
Saccharomyces cerevisiae	1.4	N/A	6.35	0.32	0.16

21 strains as cells:

- 15 bacterial sp.
- 1 archeal sp.
- 2 fungal sp.

Zymo D6331 gut metagenome mock community

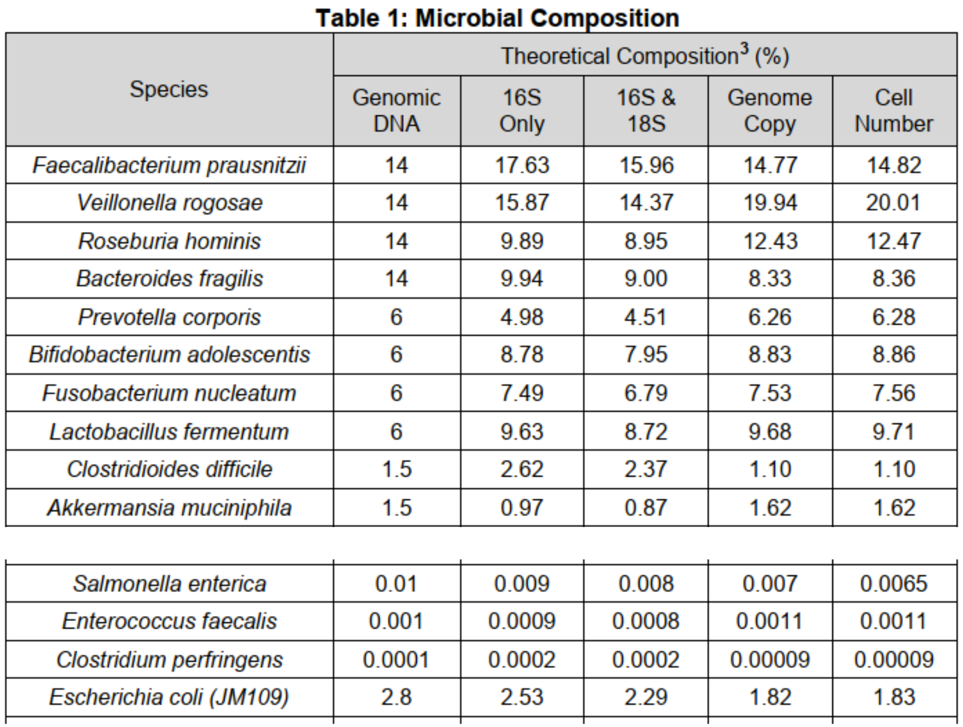


Table 1: Microbial Composition

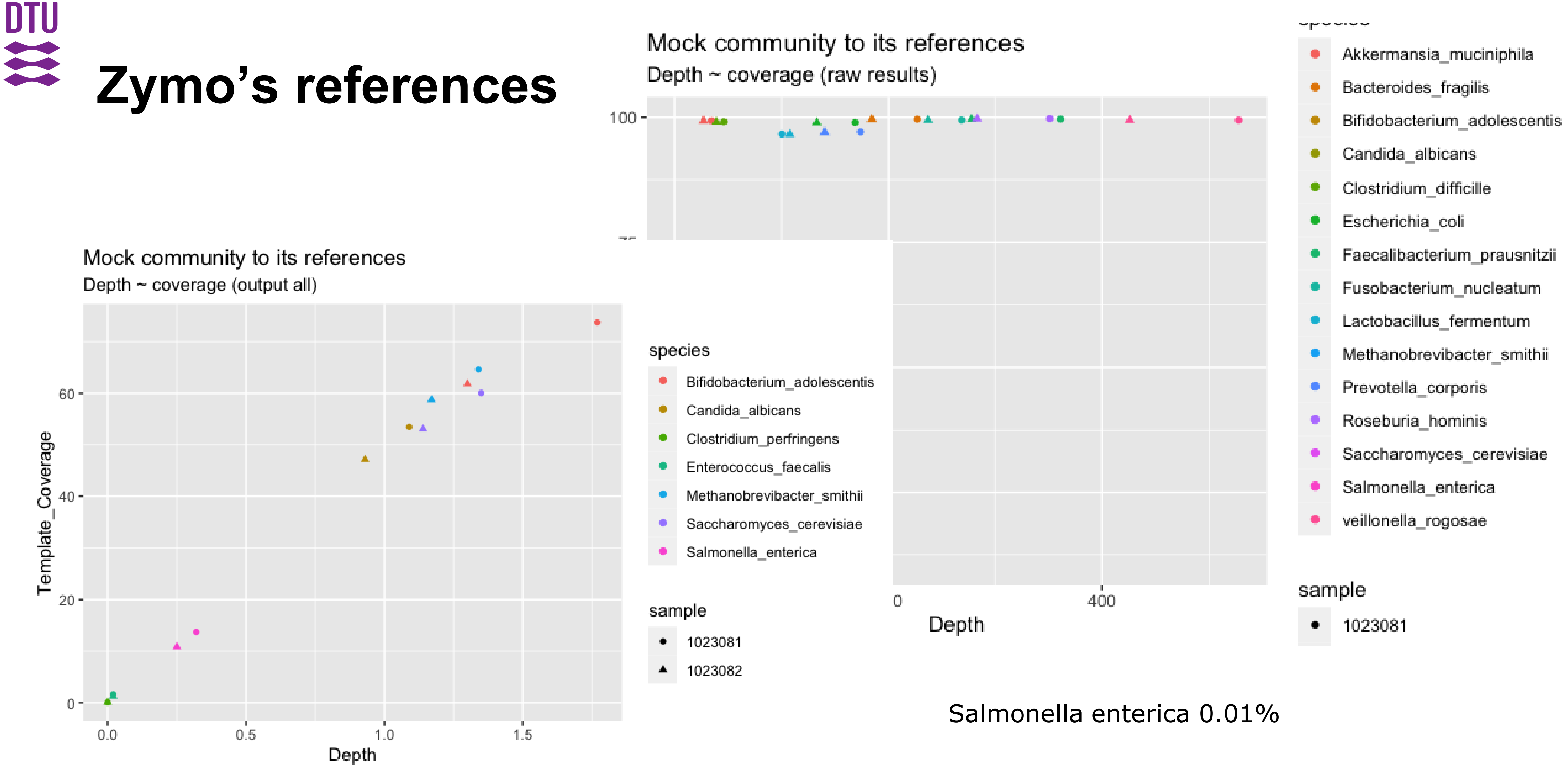
Species	Theoretical Composition[3] (%)				
	Genomic DNA	16S Only	16S & 18S	Genome Copy	Cell Number
Faecalibacterium prausnitzii	14	17.63	15.96	14.77	14.82
Veillonella rogosae	14	15.87	14.37	19.94	20.01
Roseburia hominis	14	9.89	8.95	12.43	12.47
Bacteroides fragilis	14	9.94	9.00	8.33	8.36
Prevotella corporis	6	4.98	4.51	6.26	6.28
Bifidobacterium adolescentis	6	8.78	7.95	8.83	8.86
Fusobacterium nucleatum	6	7.49	6.79	7.53	7.56
Lactobacillus fermentum	6	9.63	8.72	9.68	9.71
Clostridioides difficile	1.5	2.62	2.37	1.10	1.10
Akkermansia muciniphila	1.5	0.97	0.87	1.62	1.62
Salmonella enterica	0.01	0.009	0.008	0.007	0.0065
Enterococcus faecalis	0.001	0.0009	0.0008	0.0011	0.0011
Clostridium perfringens	0.0001	0.0002	0.0002	0.00009	0.00009
Escherichia coli (JM109)	2.8	2.53	2.29	1.82	1.83

21 strains as cells:

- 15 bacterial sp.
  - 9 phyla
  - 13 families

	Illumina		**ONT**
	Mock 1	*Mock 2*	*Mock*
Qual. bases (MB)	9613	9007	252

Zymo's references



Mock community to its references

Depth ~ coverage (raw results)

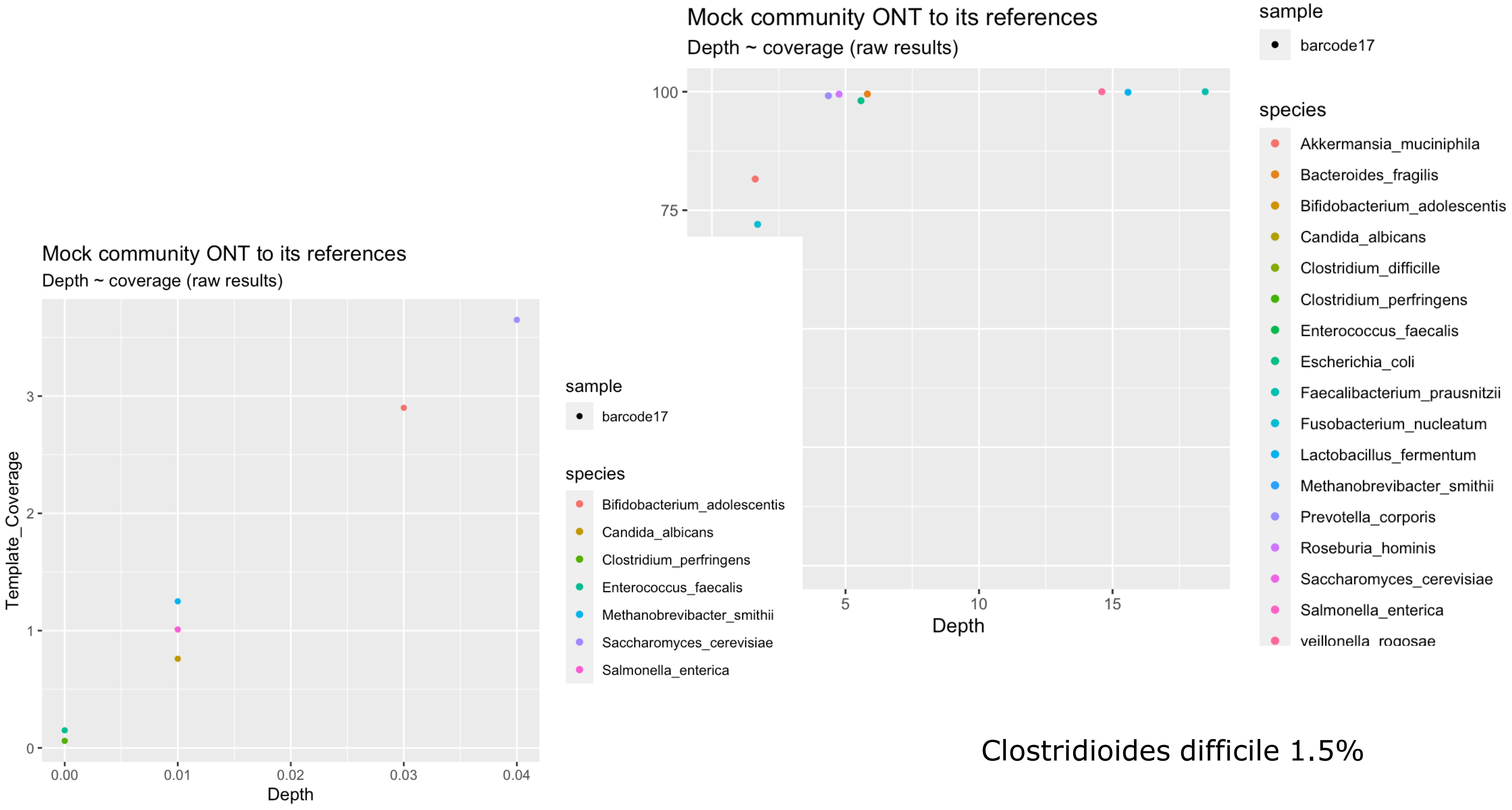
Mock community to its references

Depth ~ coverage (output all)

Salmonella enterica 0.01%

Enterococcus faecalis	0.001%
Clostridium perfringens	0.0001%

Zymo's references - Nanopore



Mock community ONT to its references

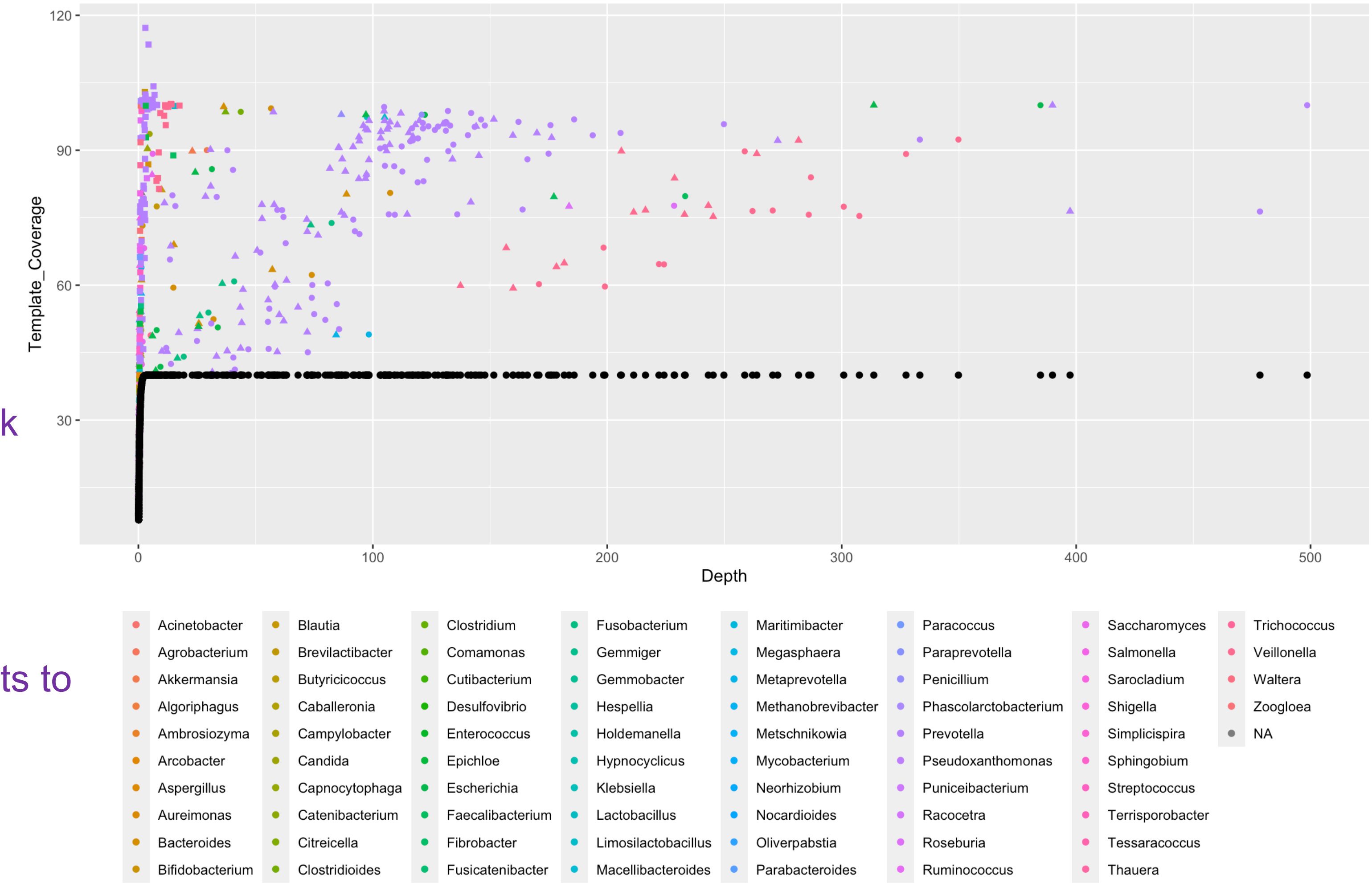
Depth ~ coverage (raw results)

Mock community ONT to its references

Depth ~ coverage (raw results)

Clostridioides difficile 1.5%

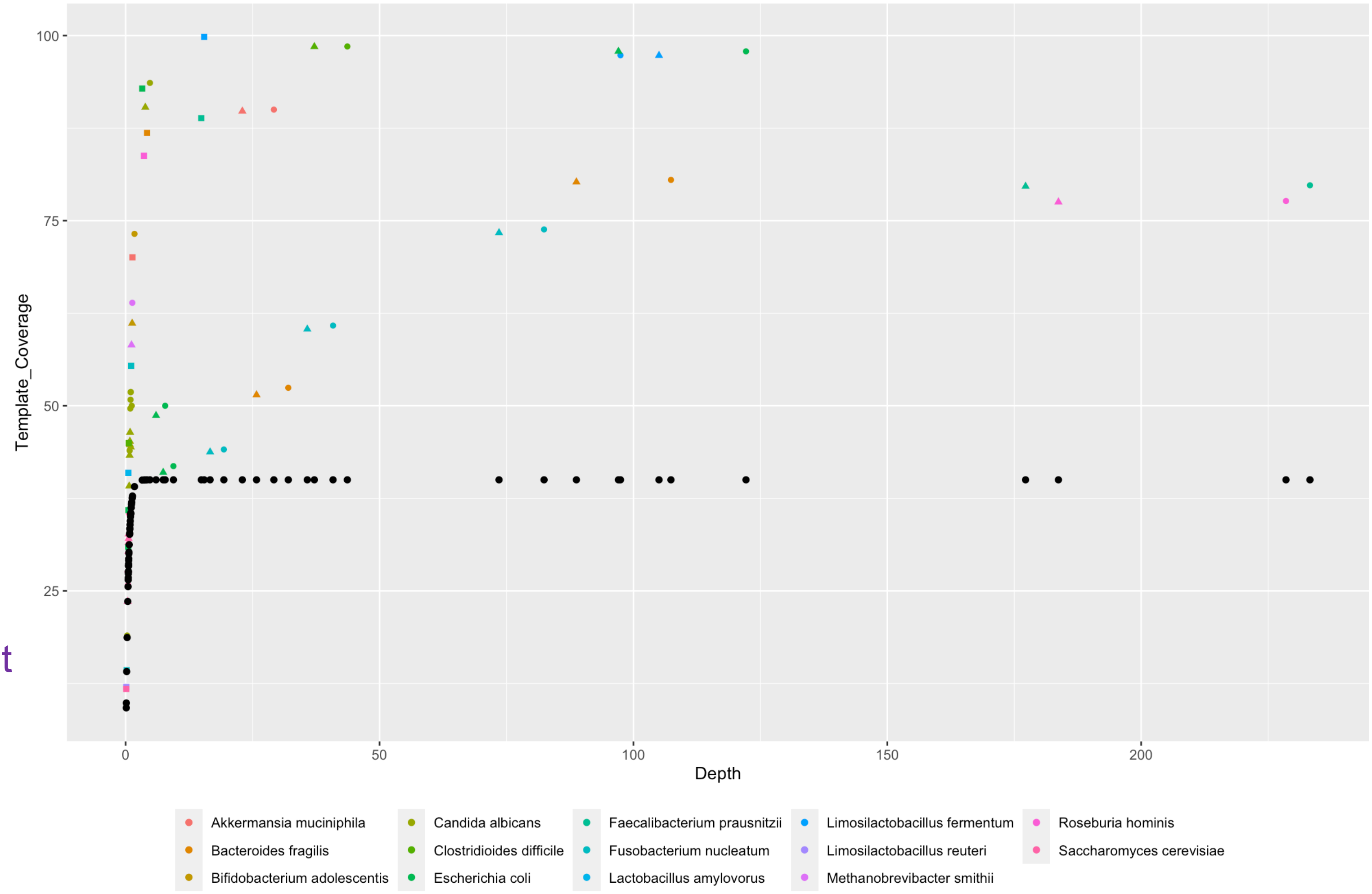
CGE “Genomic” database (RefSeq genomes)



Filtering does not work on separate contigs and plasmids

Many false positive hits to many genera

CGE “Genomic” database - complete chromosomes



Filtered results

C. difficile is the lowest abundance hit 1.5%

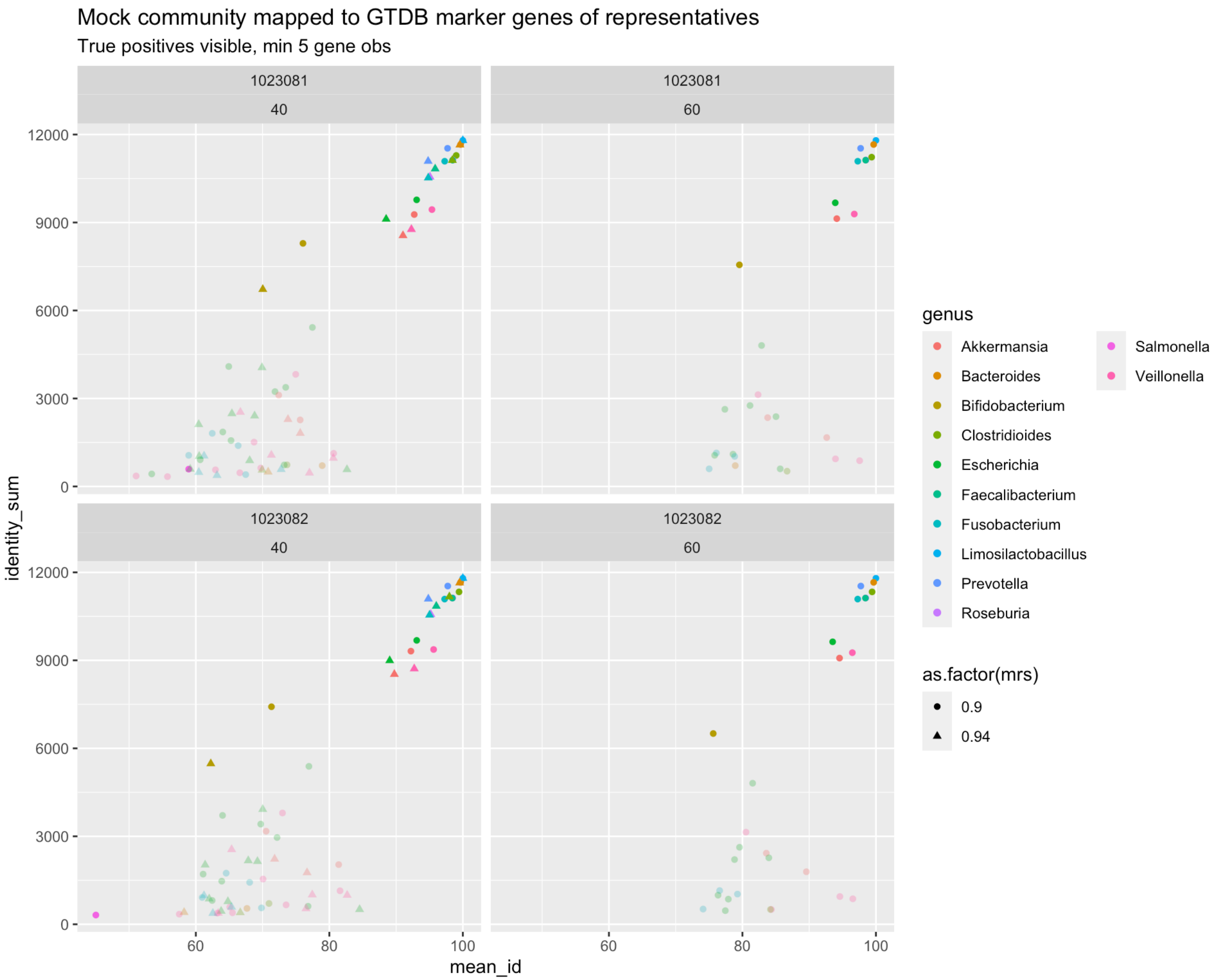
GTDB v207 bacterial single copy marker genes

- 120 loci
- pre-filtering:
  - individual loci min. covered 40 or 60%
  - min. 5 loci per species detected

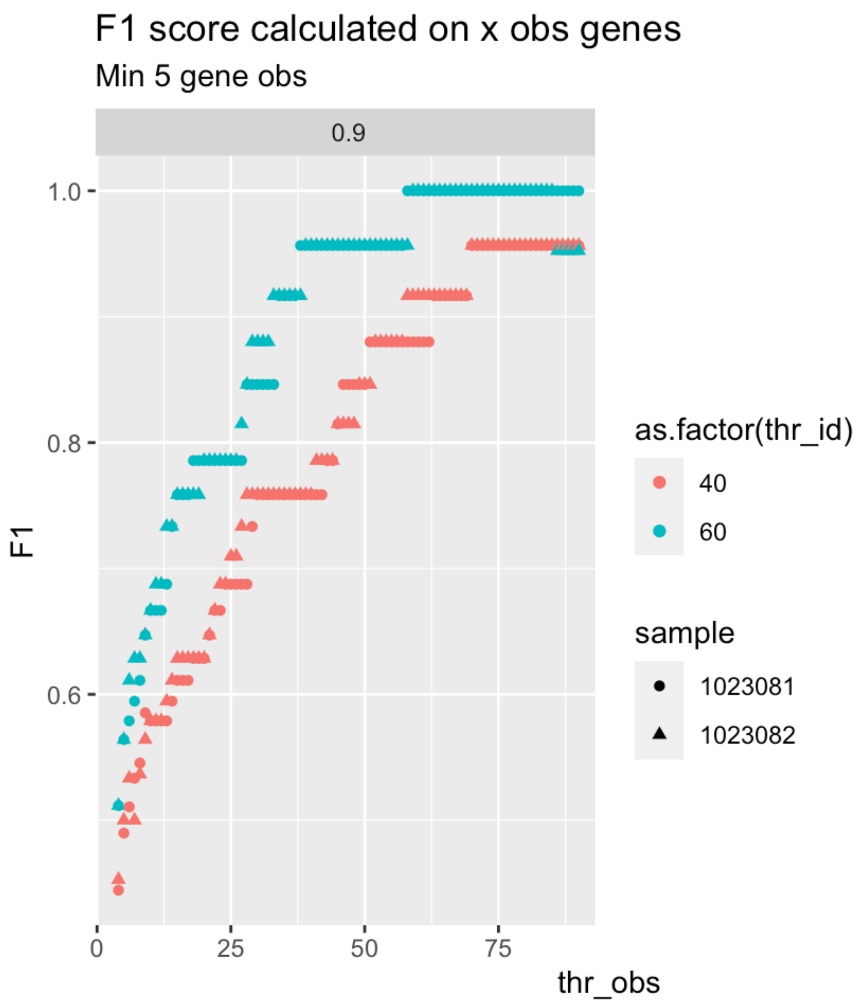
GTDB v207 bacterial single copy marker genes

Raw results:

High abundance true positives separated from false positives

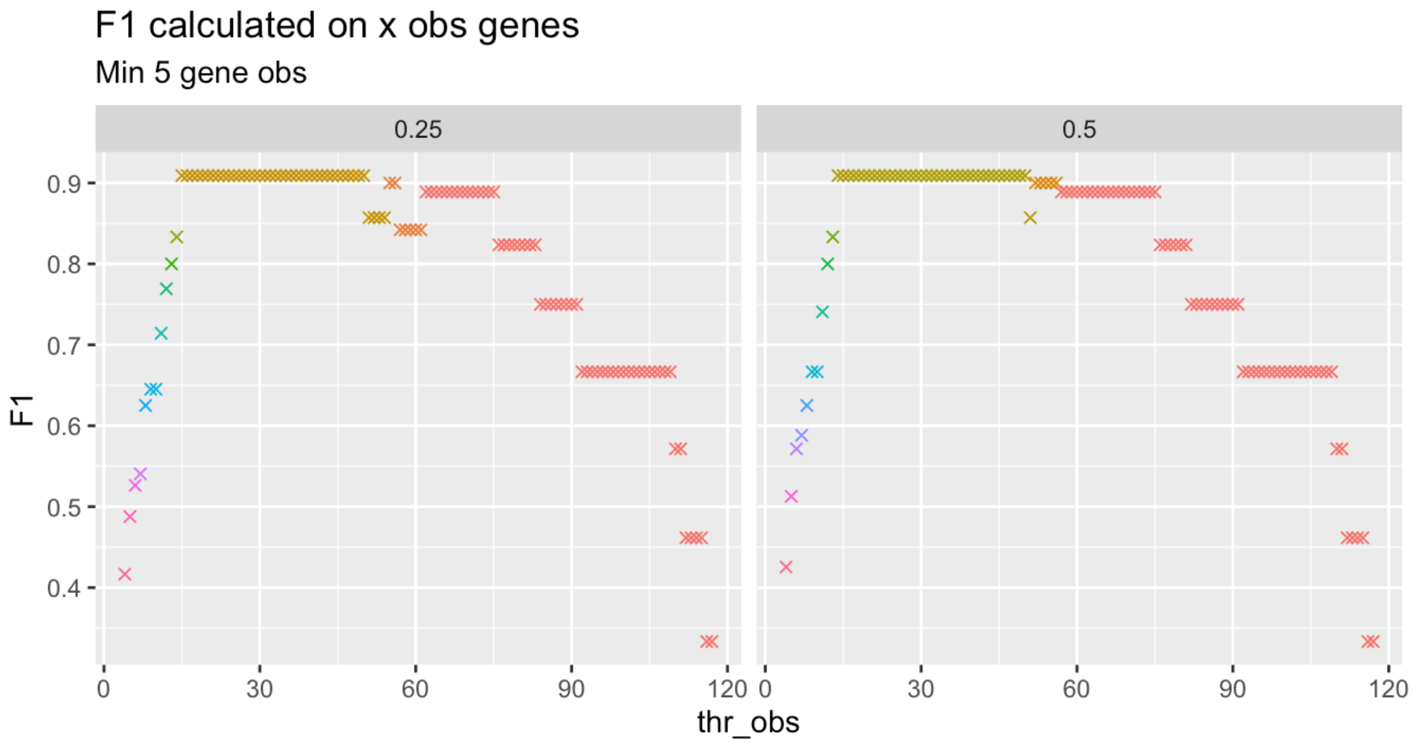


GTDB v207 bacterial single copy marker genes



Requiring min. 55 loci to be detected gives few FPs and max F1

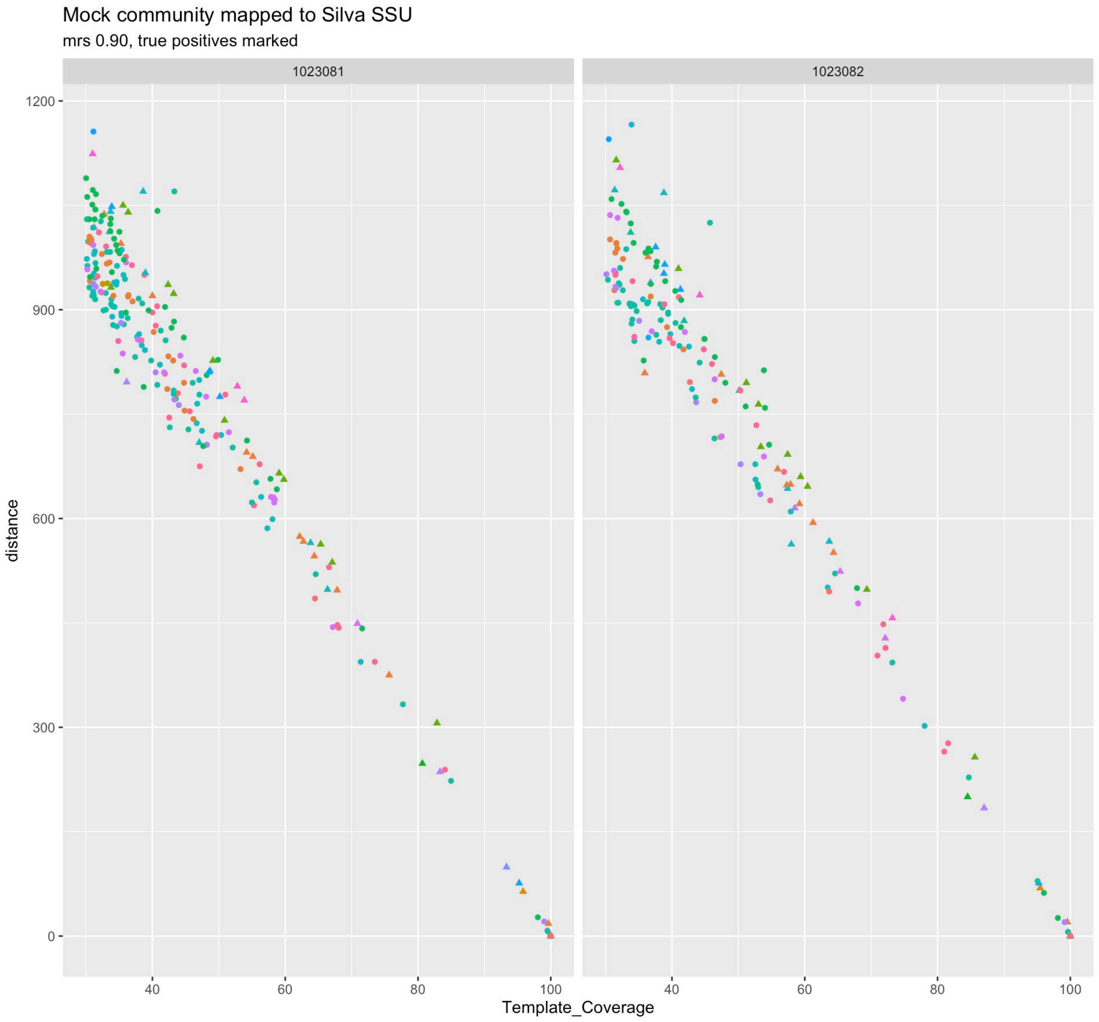
GTDB v207 - Nanopore

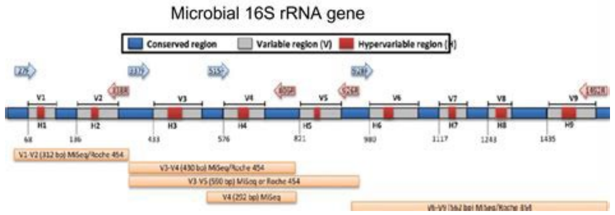


Min. 55 loci works for the stricter mapping

sample	no_TP	no_TN	no_FN	no_FP	recall	precision	F1
1023081	12	0	3	2	0.8	0.857	0.828
1023082	12	0	3	2	0.8	0.857	0.828
barcode17	9	0	6	1	0.6	0.900	0.720

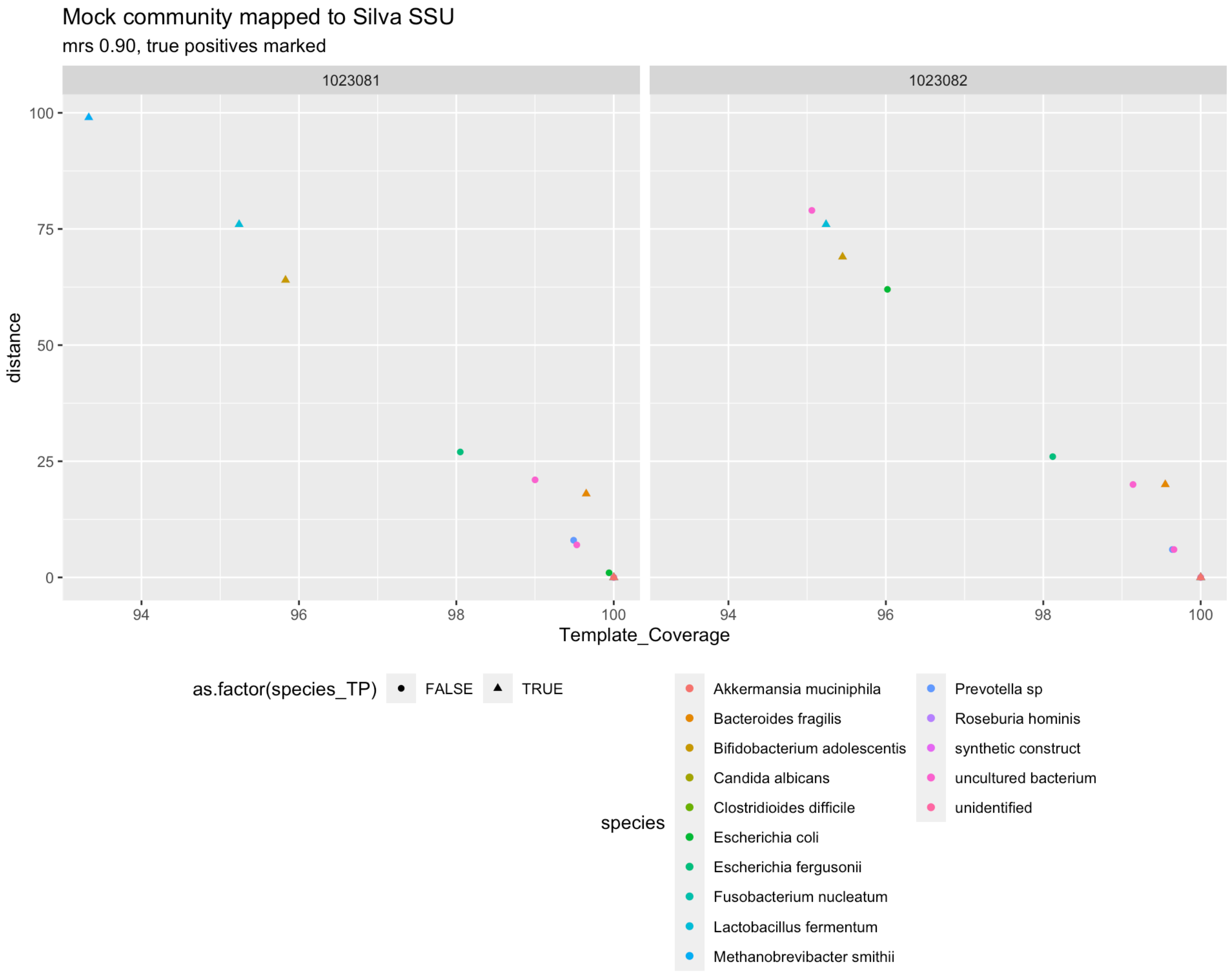
Silva SSU redundant database





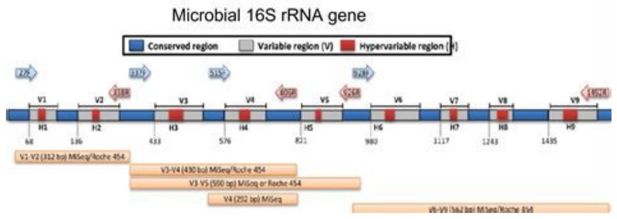
False positive hits to other genera

Silva SSU redundant database



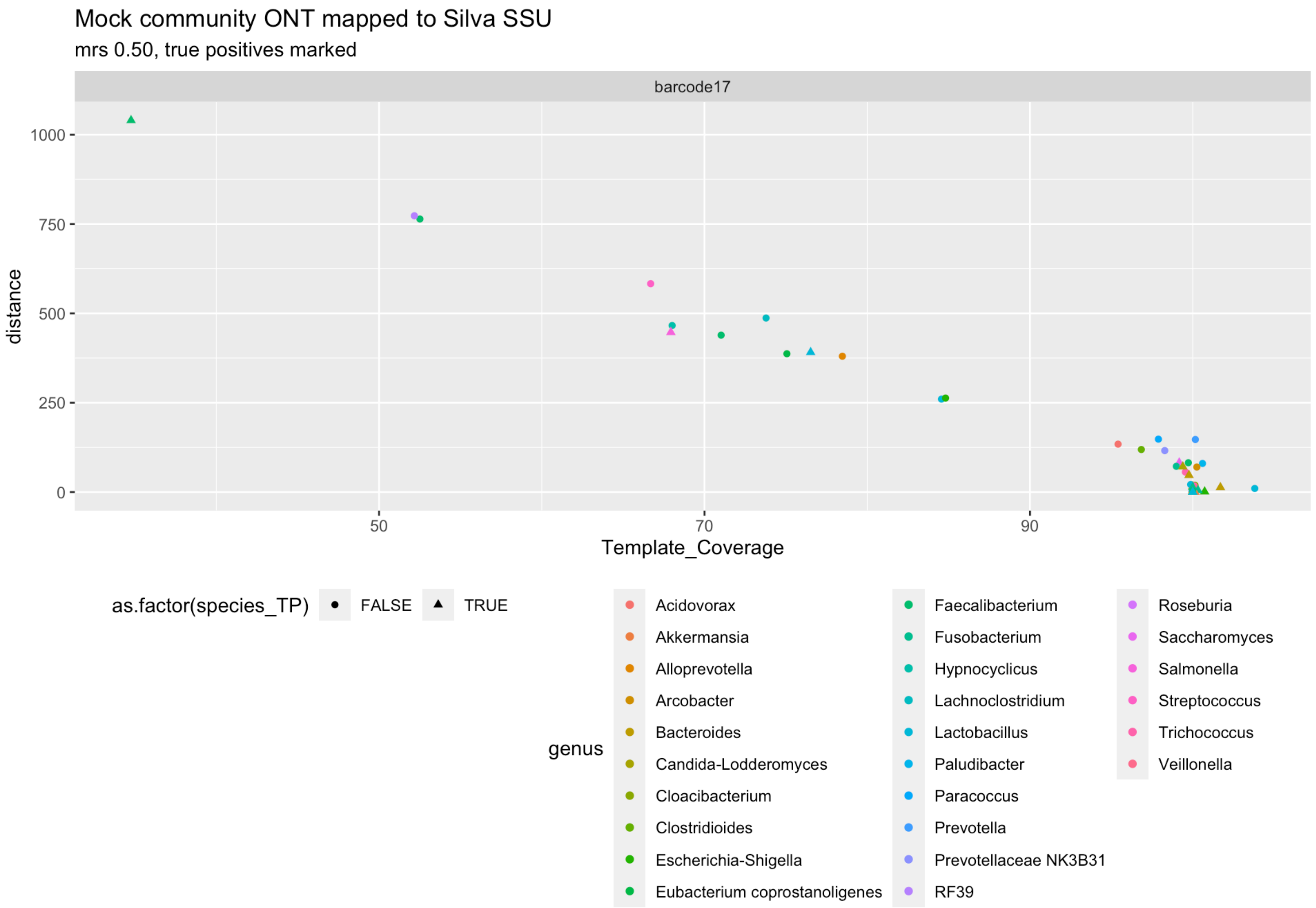
Mock community mapped to Silva SSU

mrs 0.90, true positives marked

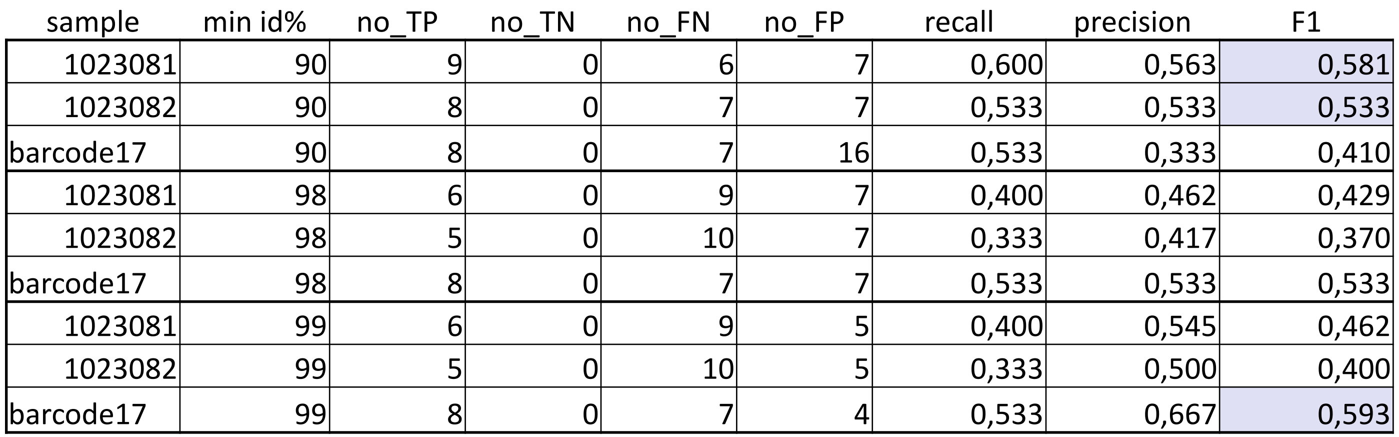


Microbial 16S rRNA gene

Silva SSU - Nanopore



Silva SSU redundant database



sample	min id%	no_TP	no_TN	no_FN	no_FP	recall	precision	F1
1023081	90	9	0	6	7	0,600	0,563	0,581
1023082	90	8	0	7	7	0,533	0,533	0,533
barcode17	90	8	0	7	16	0,533	0,333	0,410
1023081	98	6	0	9	7	0,400	0,462	0,429
1023082	98	5	0	10	7	0,333	0,417	0,370
barcode17	98	8	0	7	7	0,533	0,533	0,533
1023081	99	6	0	9	5	0,400	0,545	0,462
1023082	99	5	0	10	5	0,333	0,500	0,400
barcode17	99	8	0	7	4	0,533	0,667	0,593

min. 90 id% gives best F1 score for Illumina, but for ONT it’s 99 id%

Summary

	Illumina		ONT
	Mock 1	**Mock 2**	**Mock**
Qual bases (MB)	9613	9007	252
	Lowest genomic DNA (%) classified		
Mock reference chr	0,01	0,01	1,5
GTDB bac120	1,5	1,5	6
GTDB ar53	0,1	0,1	-
Silva SSU	1,5	1,5	1,5
Genomic chr (excl. draft genomes)	0,1	0,1	0,1
	F1 score of species level classification based on mock composition, excl. eukaryotes		
GTDB bac120+ar53 (min 46% markers hit)	0,83	0,83	0,72
Silva SSU (min 90id%)	0,58	0,53	0,41
Silva SSU (min 98id%)	0,43	0,37	0,53
Silva SSU (min 99id%)	0,46	0,40	0,59
Genomic (excl. draft genomes, coverage)	0,83	0,83	0,67

DTU