A method for sub-species taxonomic resolution of bacterial microbiomes with ONT-sequenced ribosomal DNA.

Chris Woodruff

22 October 2025

Table of Contents

SECTION: Introduction and Context for this Work

- Aim: Achieving strain-level amplicon-based bacterial microbiome profiling with nanopore-quality reads.
- What species(and, ideally, strains of species) are present and in what relative abundance?
- Nanopore sequencing - noisy, long reads.
- Denoising to give Amplicon Sequence Variants (ASVs).
- Look for matches against an appropriate database.
- Can then determine what organisms are present and, relatively, how much of each.
- Have developed a method for analysing matches to profile bacterial microbiota.
- Paucity of high quality nanopore-sequenced amplicon data publicly available.
- Good simulated data has allowed markedly more complete evaluation of the profiling method.

Problems and a Solution

- Publicly available 16S and 23S rDNA genes that were ONT nanopore-sequenced with state-of-the-art R10.4.1 pores were not available.
- Extracted such sequences from published
  - ⋆ WGS data [**?**] for an 8 bacterial species whole genome sequencing study.
  - ⋆ Partial 16S-ITS-23S ribosomal DNA sequences from preferred primer pairs for extraction [**?**]
- Demonstrated strain-level resolution over 2 orders of relative abundance based on this data.
- Limitations in generality as the mock microbiomes of these studies had all bacteria differing at the genus level.
- Simulation seen as only current approach to more demanding assessment of method.
- Badread simulator of Ryan Wick [**?**] seemed to be the most suitable simulator
- Also considered squiggle-based simulators (Beslic et al. 2025 [**?**], Gamaarachi et al. 2024 [**?**]) and Nanosim, [**?**].

Key Data Requirements

- Reference library of rDNA sequences of bacteria relevant to the microbiomes of interest.
- Libraries of 16S, 23S and rDNA operons representative of nanopore-sequenced microbiomes
- A procedure to process such libraries to identify at strain, or near-strain, resolution what organisms are present and their relative abundances.
- For each strain or species of the database have all operons of that strain identified in the database.

The Human Gut Microbiome

- WEHI is a medical research institution.
- Human gut microbiome is being perceived as of increasing relevance to physical and mental health.
- Rapid development in knowledge of composition of healthy and disease state microbiomes.
- King et al (2019) [**?**] proposed a set of bacteria and their relative abundances as an attempt to move to some standardisation of a healthy human gut.
- I took this as a basis for my current work.
- This "standard" gut microbiome has 160 different strains of bacteria and their relative abundances - not all identified at the strain level.
- Used a subset of strains and rank ordering of abundances to specify a mock microbiome all at the strain level.

SECTION: Results

Real Reads Source 1 - Sereika et al.

- Have 16S and 23S data, no rrn data.
- Used an pper bound on read mean error rate of 0.01 errors/base
- For 16S have 23136 fastq record, reduced to 6430 after quality and length filtering.
- For 23S have 23613 fastq record, reduced to 5334 after quality and length filtering.

Read Quality for Denoising

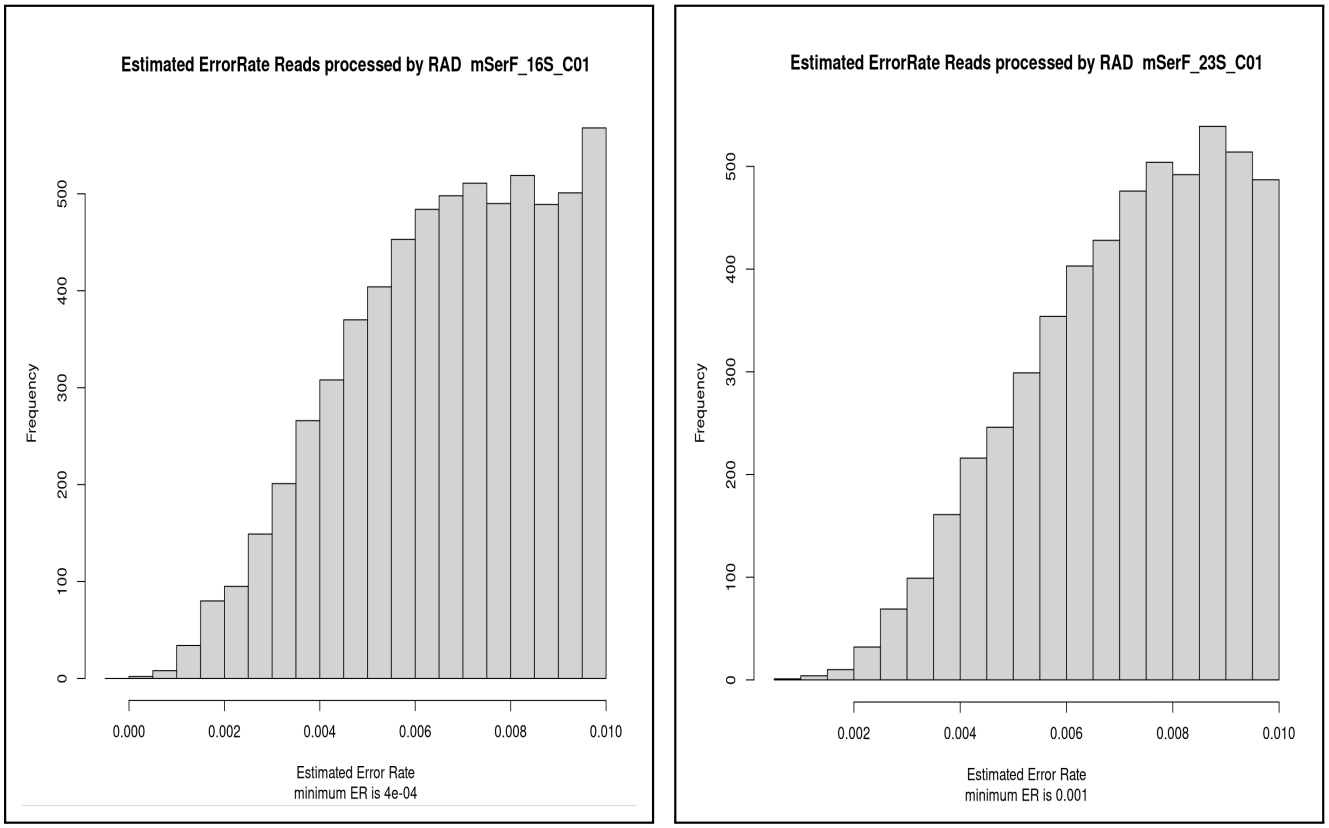

Figure 1: Error rate of reads processed by the RAD denoiser from the 16S and 23S rRNA gene sequences extracted from Sereika et al.'s WGS of the Zymo D6322 mock microbiome.

Identification 16S

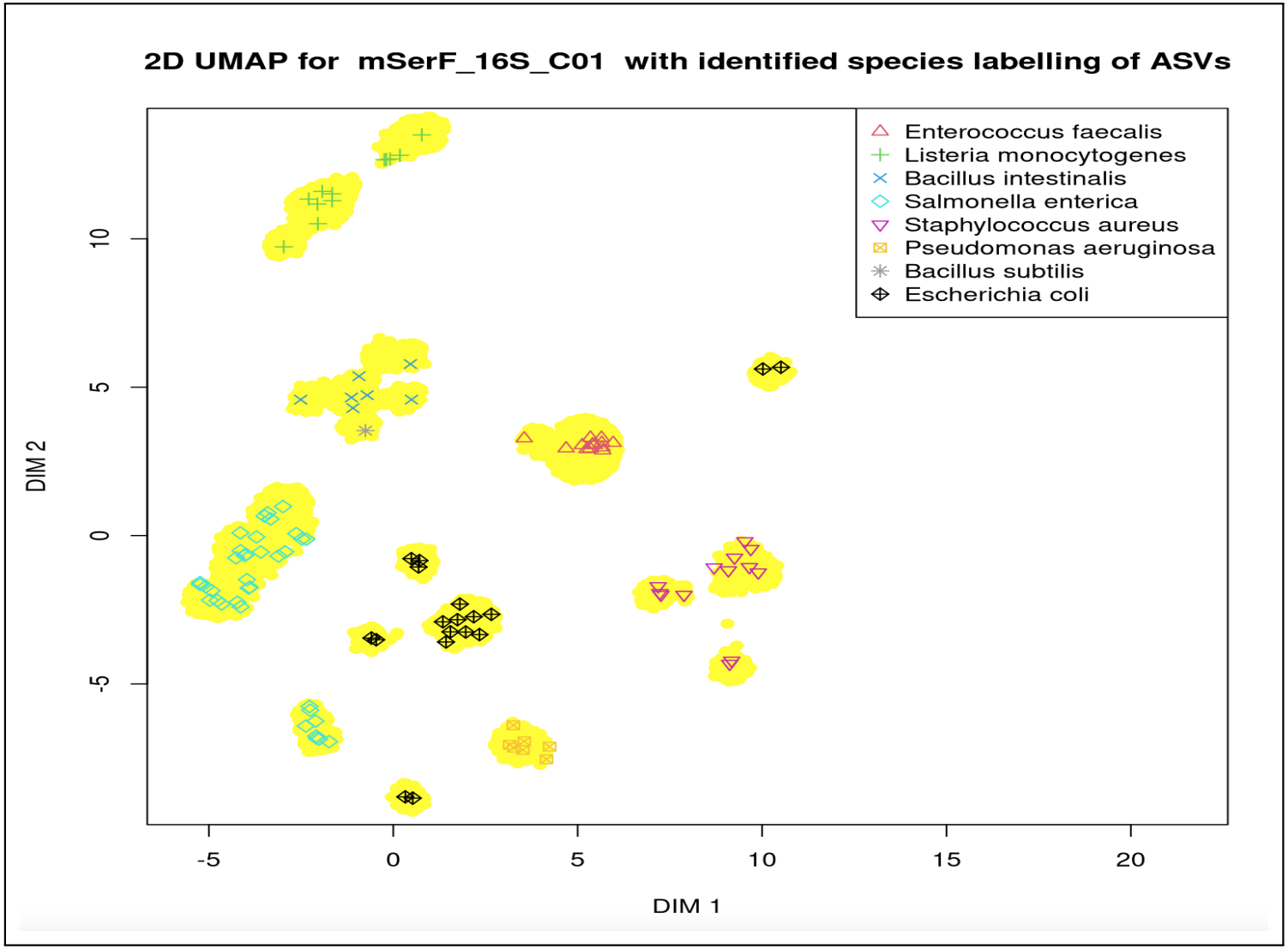

Figure 2: 2D UMAP representation of the relationship between ASVs and their associated read clusters for the 16S rRNA gene sequences extracted from the Sereika et al. D6322 dataset.

Identification 23S

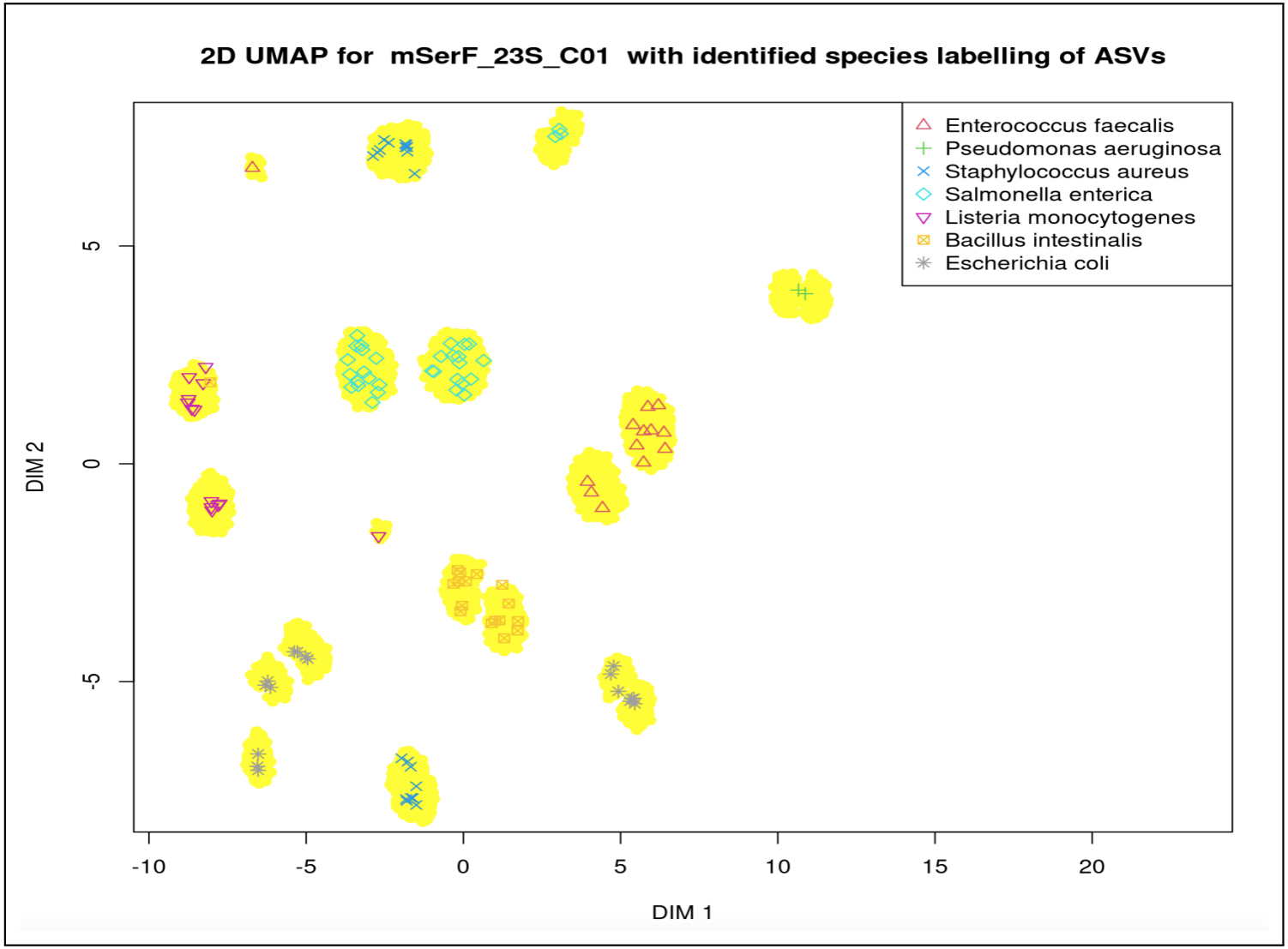

Figure 3: 2D UMAP representation of the relationship between ASVs and their associated read clusters for the 23S rRNA gene sequences extracted from the Sereika et al. D6322 dataset.

Quantification 16S

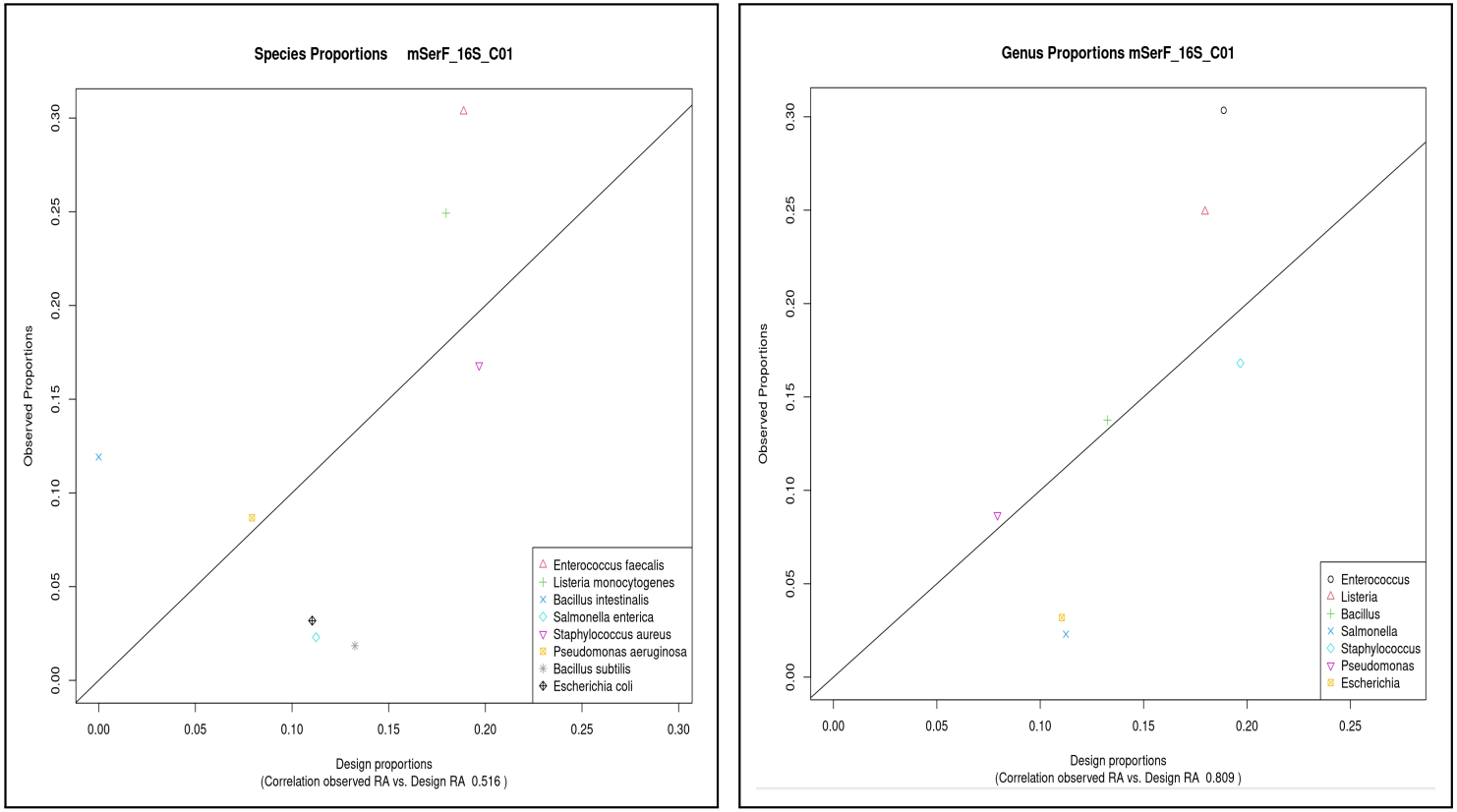

Figure 4: Relation between the observed species and genus relative abundances and those of the designed microbiome as determined from 16S rRNA genes. ASVs receiving labelling of non-design species (invalid) have a design abundance of zero. Thus both missed and invalid species (or genera or strains) are plotted and easily identified.

Quantification 23S

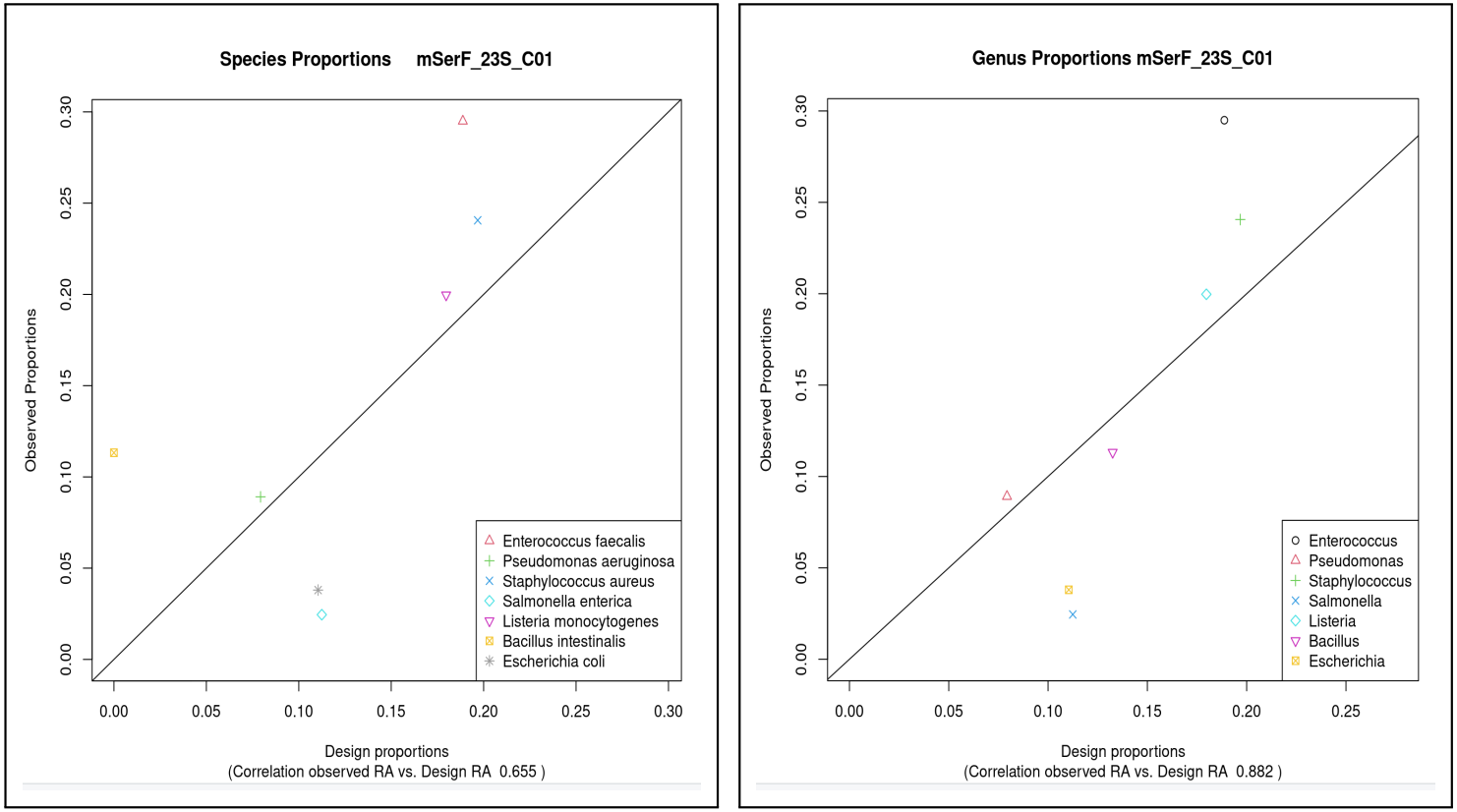

Figure 5: Relation between the observed species and genus relative abundances and those of the designed microbiome as determined from 23S rRNA genes. ASVs receiving labelling of non-design species (invalid) have a design abundance of zero. Thus both missed and invalid species (or genera or strains) are plotted and easily identified.

Summary Sereika results

- Data quality: Very good - some reads wih error rates ¡0.005
- Identification: Very good - 16S species-level has 1 invalid species (Bacillus intestinalis). Due to ties in g score (100%)
- Quantification: Moderate - noisy, but +ve correlation of observed with design.

Real Reads Source 2 - Srinivas et al.

- Two mock microbiomes from this study were considered
  - ATCC "even"; 10 bacterial strains, all of different genus; provided as equal masses of genomic DNA for each strain.
  - Zymobiomics D6311; 7 bacterial strains, all of different genus; provided as masses forming geometric series, with ratio 1:10 - called a "logarithmic" microbiome.
- For each microbiome 1 of 4 pairs of start and end primers to define the 16S-ITS-23S amplicons was used.
- Identify datasets as SA1, SA2, SA3 and SA4 for the ATCC source, and as SZ1, SZ2,.. for the Zymobiomics source.

Read Quality for Denoising, SA1, SA2

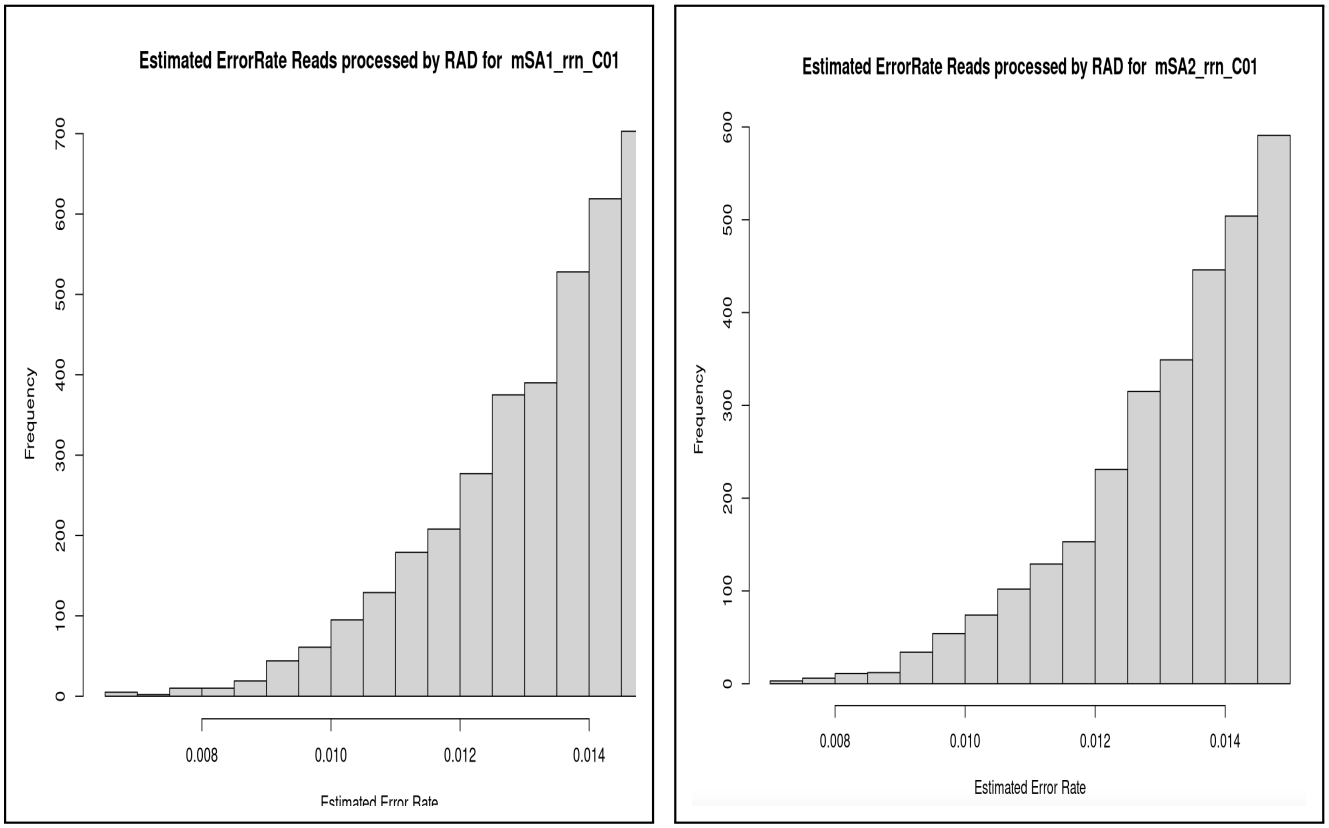

Figure 6: Error rate of reads processed by the RAD denoiser from the 16S-ITS-23S rRNA operons of the ATCC mock microbiomes using primer pairs 1 (left) and 2 (right).

Read Quality for Denoising, SZ1, SZ2

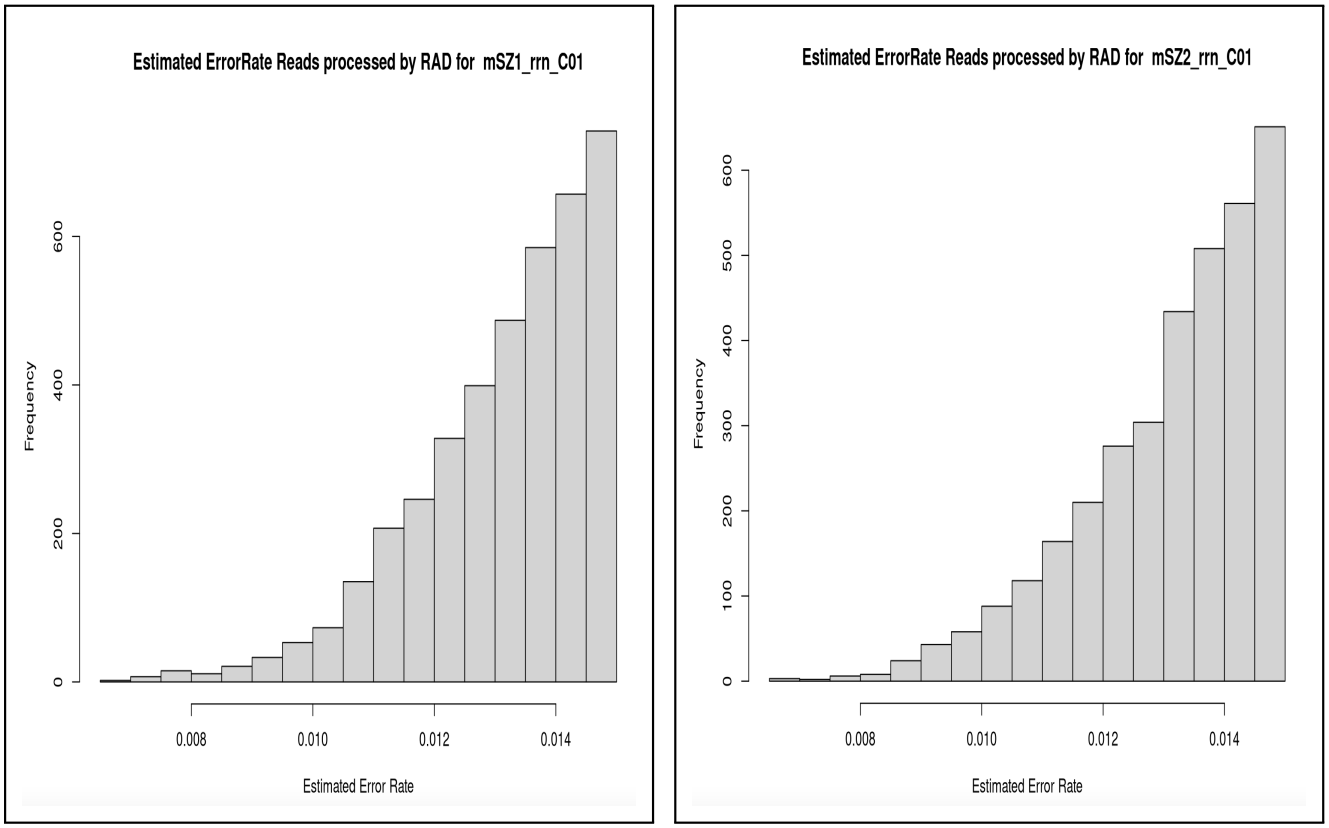

Figure 7: Error rate of reads processed by the RAD denoiser from the 16S-ITS-23S rRNA operons of the Zymo D6311 mock microbiomes using primer pairs 1 (left) and 2 (right).

Read Quality for Denoising - Notes

- Figures **??**,**??** show very few reads with error rate below 0.008. The reads of these datasets are markedly inferior to those from Sereika et al..
- Table **??** presents detail on the results of denoising and blast alignment.
- Some entries have bracketed numbers indicating the total number of unique genera or species that were labelled, with the preceding unbracketed number being the number of valid labellings.
- ASVs were cut(removed) because their alignment was too short or the alignment score fell below a specified threshold.
- Comparison with Sereika datasets shows the importance of data quality extending below a mean read error rate of about 0.008 errors per base.

Identification - Data

Table 1: Srinivas et al. D6322 ATCC and Zymo D6311 mock microbiome. Table column headers Cut= number of ASVs removed from analysis, UniqG=unique genera, UniqSp=unique Species, UniqOps=unique rrn operons

Data	Reads	ASV	Cut	UniqG	UniqSp	UniqOps
SA1	3654	13	1	4	4	7
SA2	3014	14	1	6(7)	6(7)	10
SA3	1139	8	0	1(2)	1(2)	4
SA4	3264	20	2	4(5)	4(7)	16
SZ1	4001	18	0	2	2	5
SZ2	3458	19	0	2	2	6
SZ3	1406	11	3	1	1	1
SZ4	4872	35	0	2	2	5

Identification - Visualisation

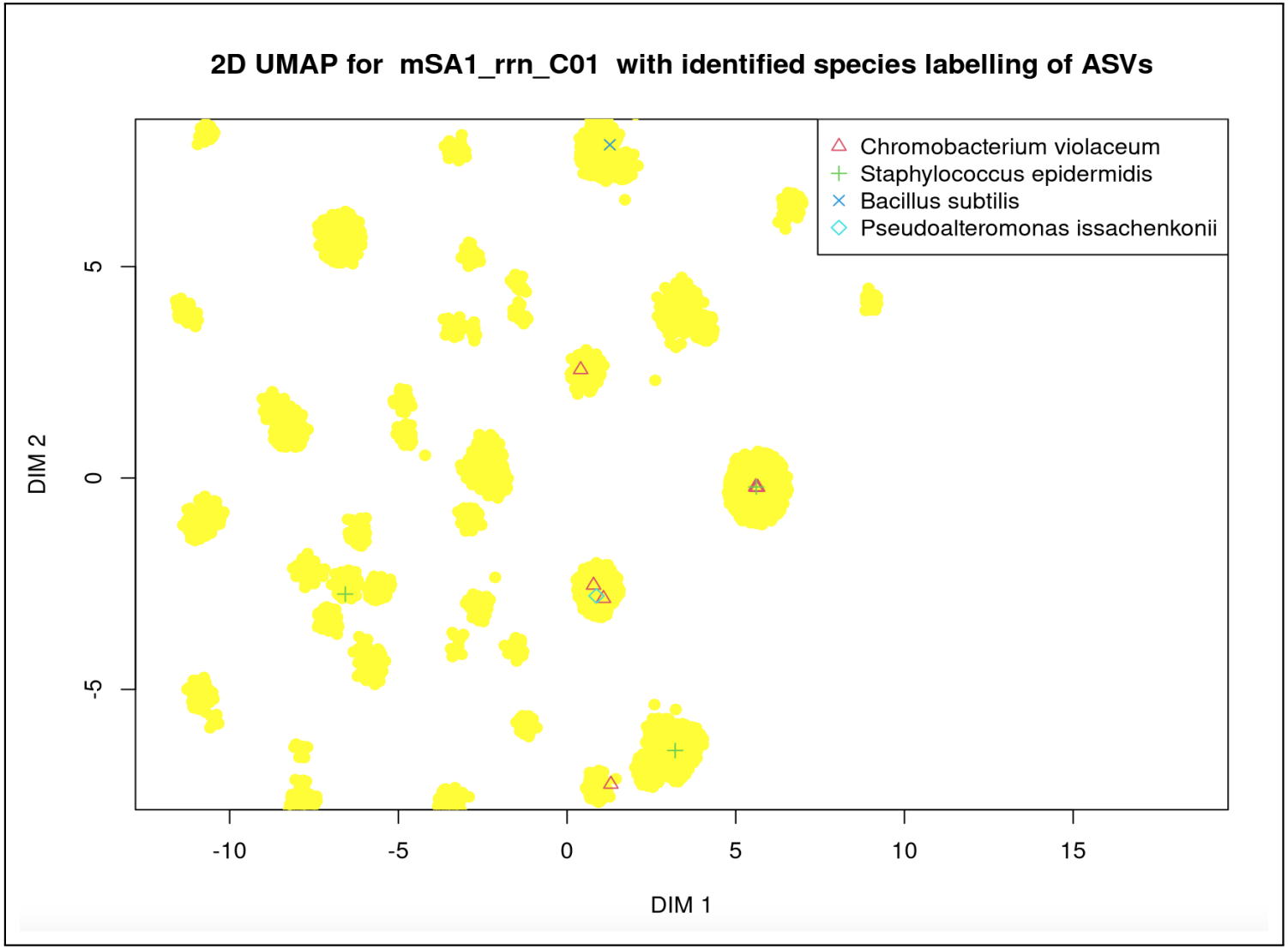

Figure 8: 2D UMAP representation of ASVs and their associated read clusters for the 16S-ITS-23S sequences, primer pair 1, from the Srinivas et al. study for the ATCC even microbiome.

Identification Notes

- Table **??** shows that there is more success with the ATCC even microbiome than the Zymo logarithmic microbiome
- The data is unable to support identification of any component in the Zymo microbiome other than the 2 most abundant - despite the taxonomic level of differentiation being genus.
- The following figures provided visualisation for SA1, SZ1 and SA2. SA2 was chosen as it is perhaps the best-performing of the Srinivas datasets on identification.
- In all figures most read clusters have no ASV associated with them. Hence our quantification method will be wildly inaccurate as most reads are not counted.

Identification - Visualisation

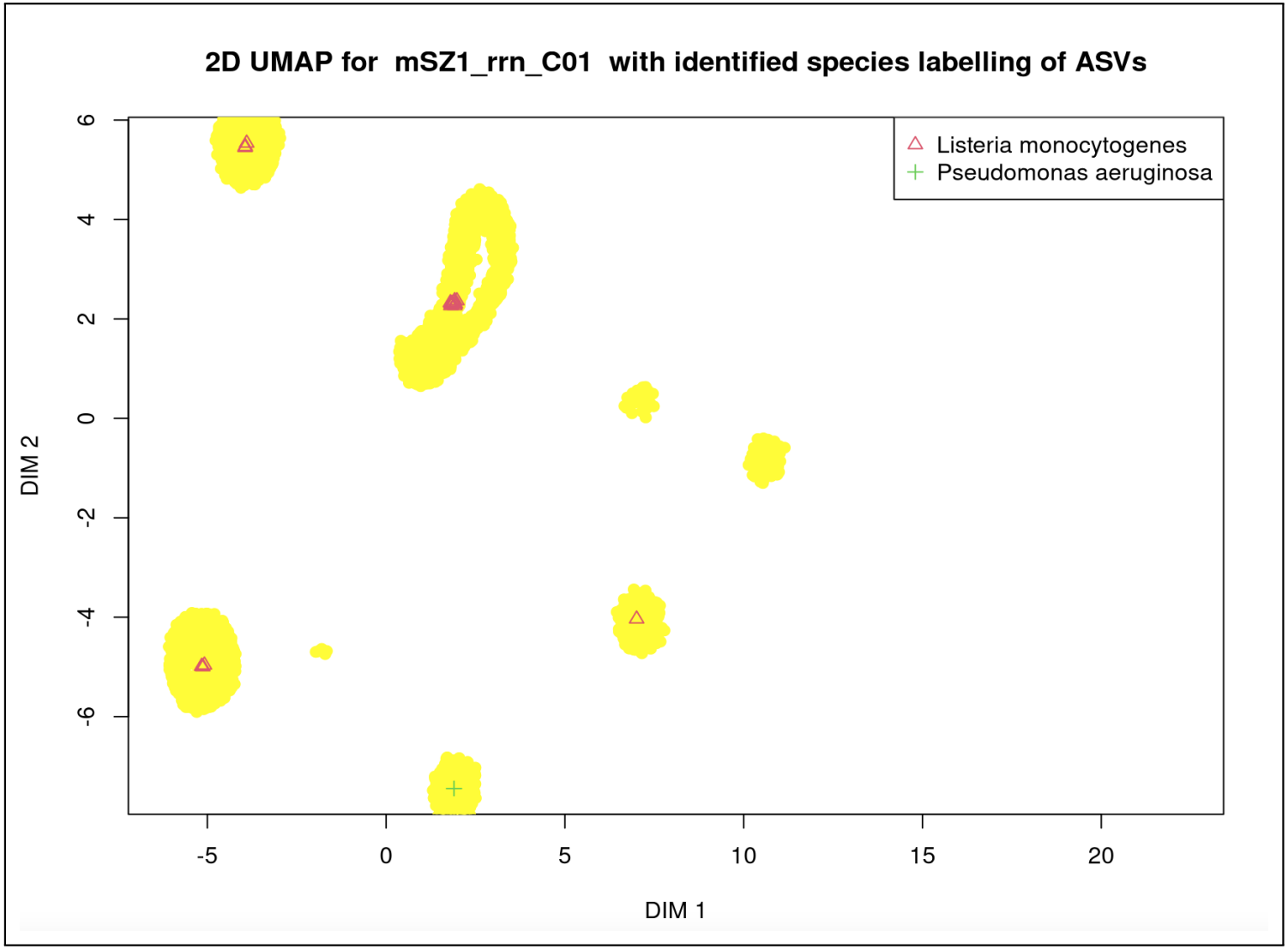

Figure 9: 2D UMAP representation of ASVs and their associated read clusters for the 16S-ITS-23S sequences, primer pair 1, from the Srinivas et al. study for the Zymo D6311 log microbiome.

Identification - Visualisation

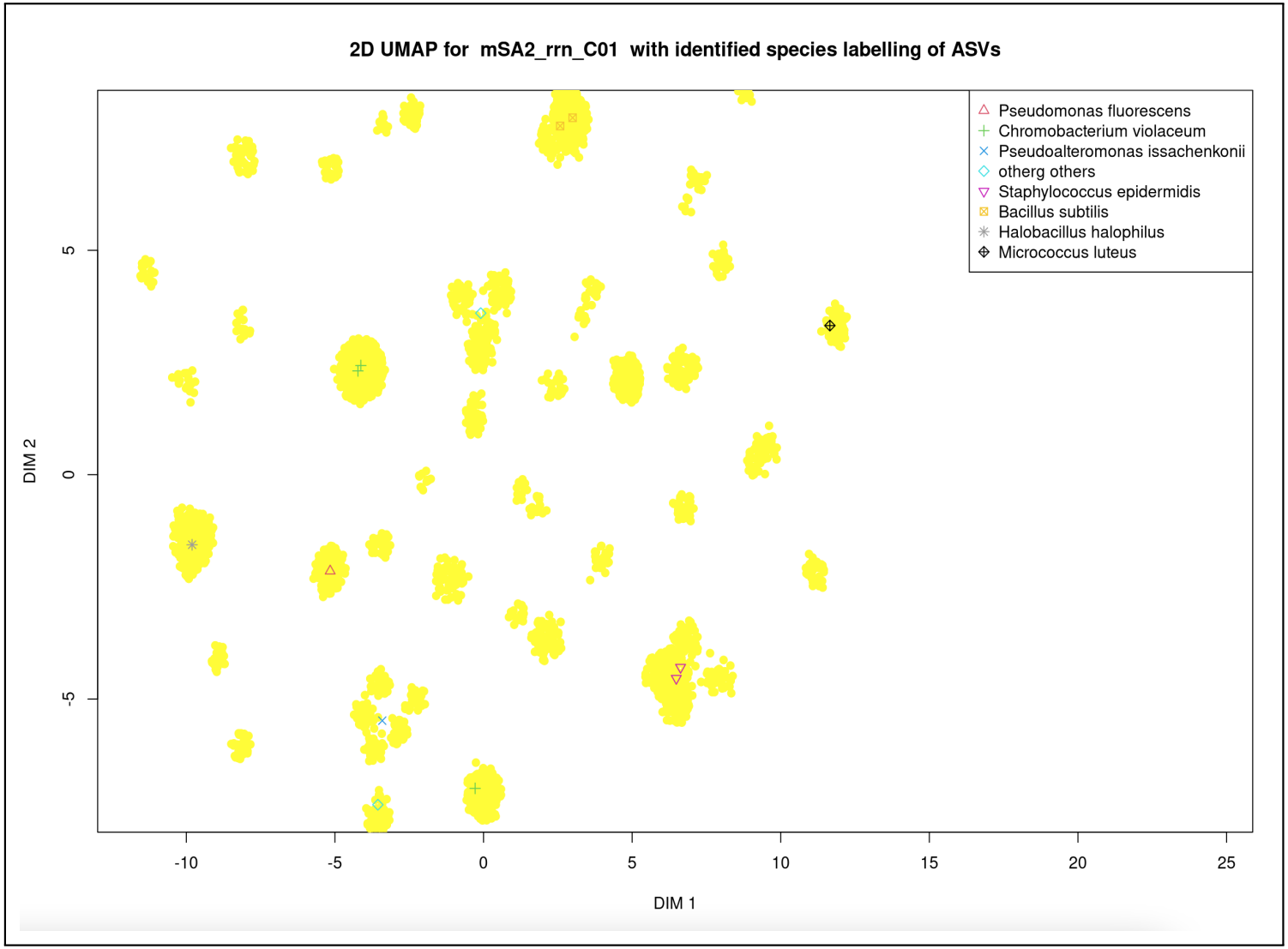

Figure 10: 2D UMAP representation of ASVs and their associated read clusters for the 16S-ITS-23S sequences, primer pair 2, from the Srinivas et al. study for the ATCC even microbiome.

Quantification - Visualisation

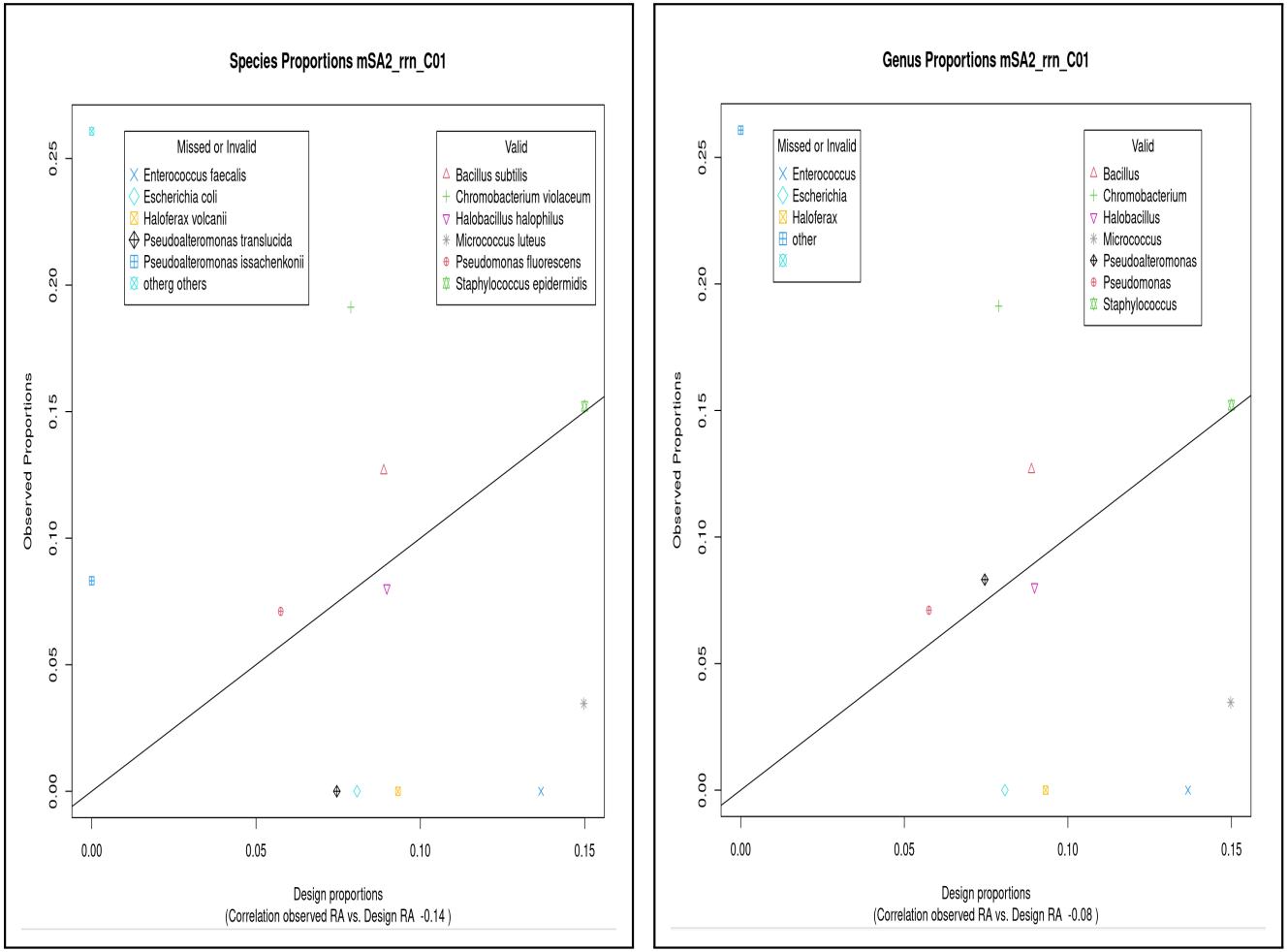

Figure 11: Observed species and genus relative abundance vs. designed microbiome as determined from the 16S-ITS-23S sequences extracted using primer pair 2 for the ATCC mock microbiome.

Quantification Notes

- The plots include both strains that are in the design but not observed (classified as "missed", and lying along the zero line for observed value), and those observed strains that are not actually in the design (classified as "invalid", and lying along the zero line for design value).
- The correlation between observed and design relative abundances is computed based only on those points representing observed strains that do exist in the designed microbiome.
- As expected , quantification was close to meaningless.
- Note that there can be more than one ASV for a single operon - for instance two perfect alignments but of different length. Figure **??** has 2 quite close ASVs in 3 read clusters. There are 13 operons shown but the analysis shows 10 unique operons. Almost certainly these 3 pairs of close ASVs correspond to 3 operons with 2 ASVs associated with each.

Simulated Reads Source - King + Wick

- From King et al.'s 160 strains, identified 59 that could be matched at strain level in the Walsh et al.'s Refseq-derived rrn operon database.
- Using the average relative abundance(RA) values of King the strains were rank-ordered and given relative abundances reportedf by King then scaled to sum to 1. Final values ranged from approximately 0.5 to 0.0001.
- The set of 59 strains have a total of 288 operons. For each operon a sufficiently large number of reads were simulated to allow construction of a mock microbiome with approximately 50000 or 25000 reads per operon for the most abundant strain (*Phocaeicola vulgatus ATCC 8482)*.
- Two qualities of this mock microbiome, labelled C01 and C11, were generated using Wick's badread code. Also, 3 types of sequences were used - 16S, 23S rRNA genes of 16S-ITS-23S (rrn) operons.

Read Quality for Denoising

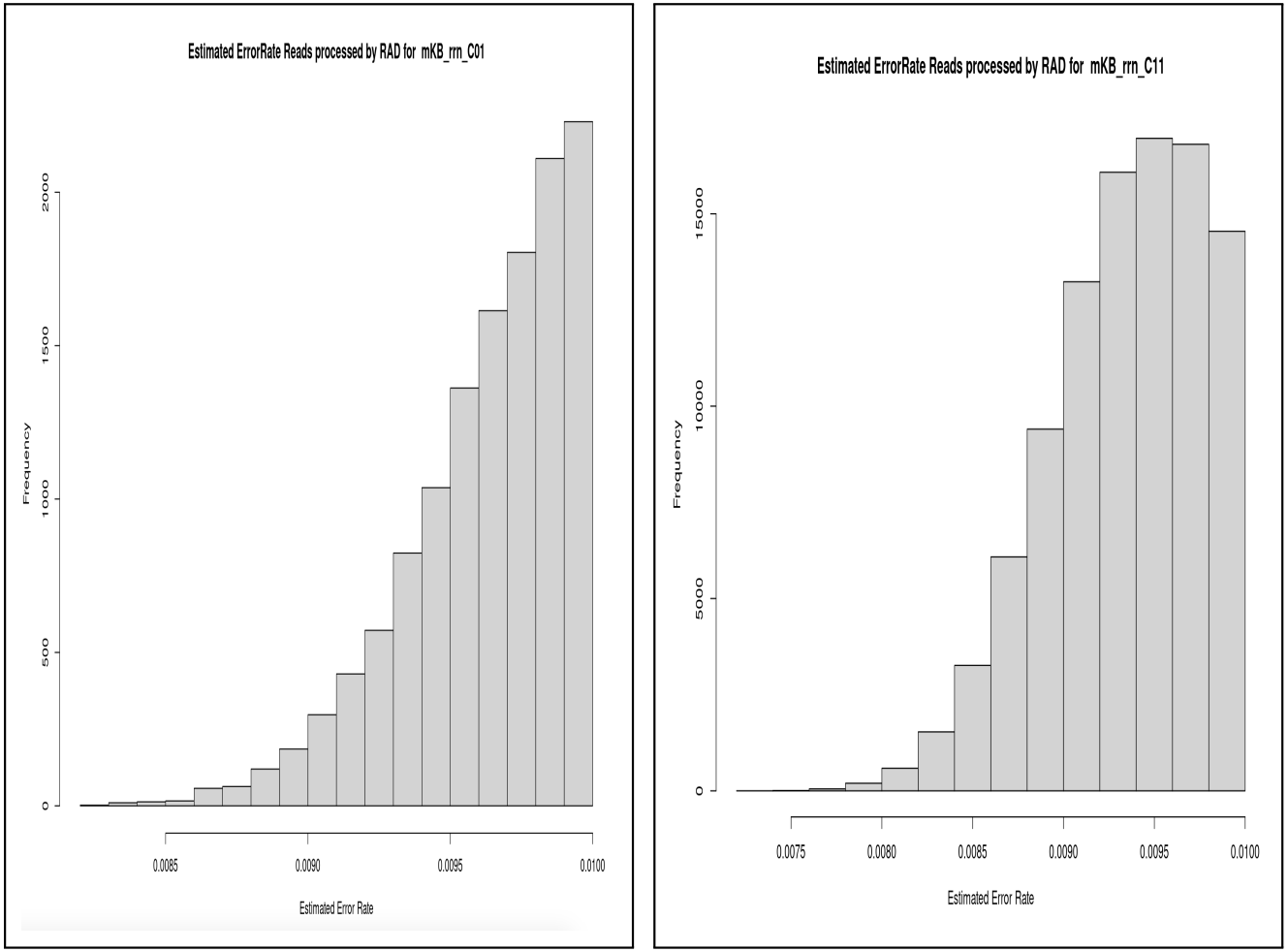

Figure 12: Error rate of reads processed by the RAD denoiser from the badread-simulated reads based on the King et al. healthy human gut Knowledge Base. The left-hand plot is from dataset C01 the right-hand from C11 .

Quality - Notes

- There is a clear improvement from C01 to C11 in the distribution of read error rate.
- The C11 distribution of reads is closer than the Sereika data, but lacks the reads below an error rate of about 0.0075.
- C11 is clearly better than the Srinivas datasets, while C01, while also better, is more similar.
- Overall, the simulated data quality falls between the data quality of the two real reads sources.
- `mKB_rrn_C01` has 12746 reads input to RAD denoising, and 197 ASVs formed.
- `mKB_rrn_C11` has 98762 reads input to RAD denoising, and 523 ASVs formed.

Identification - Visualisation

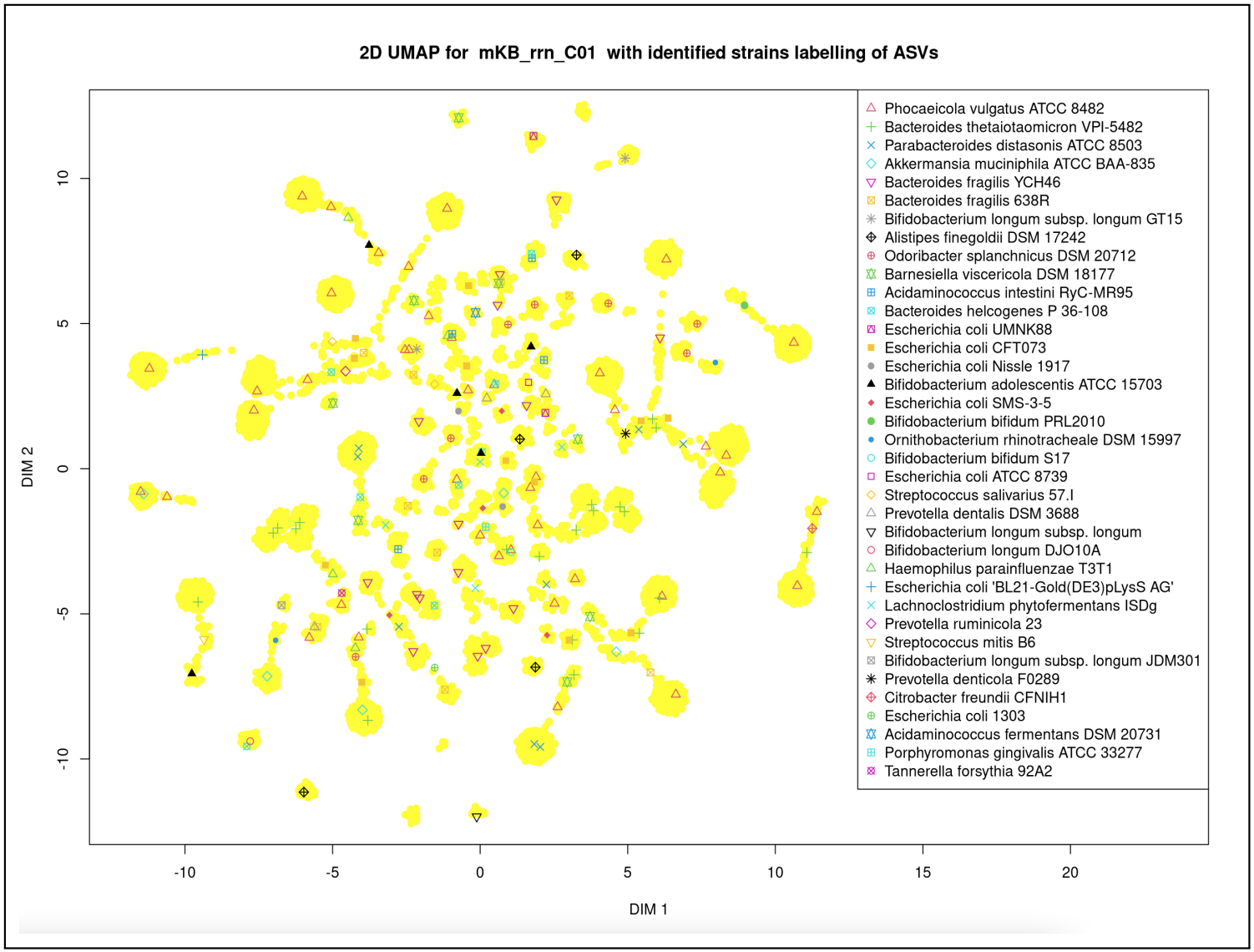

Figure 13: 2D UMAP of dataset `mKB_rrn_C01`

Identification - Visualisation

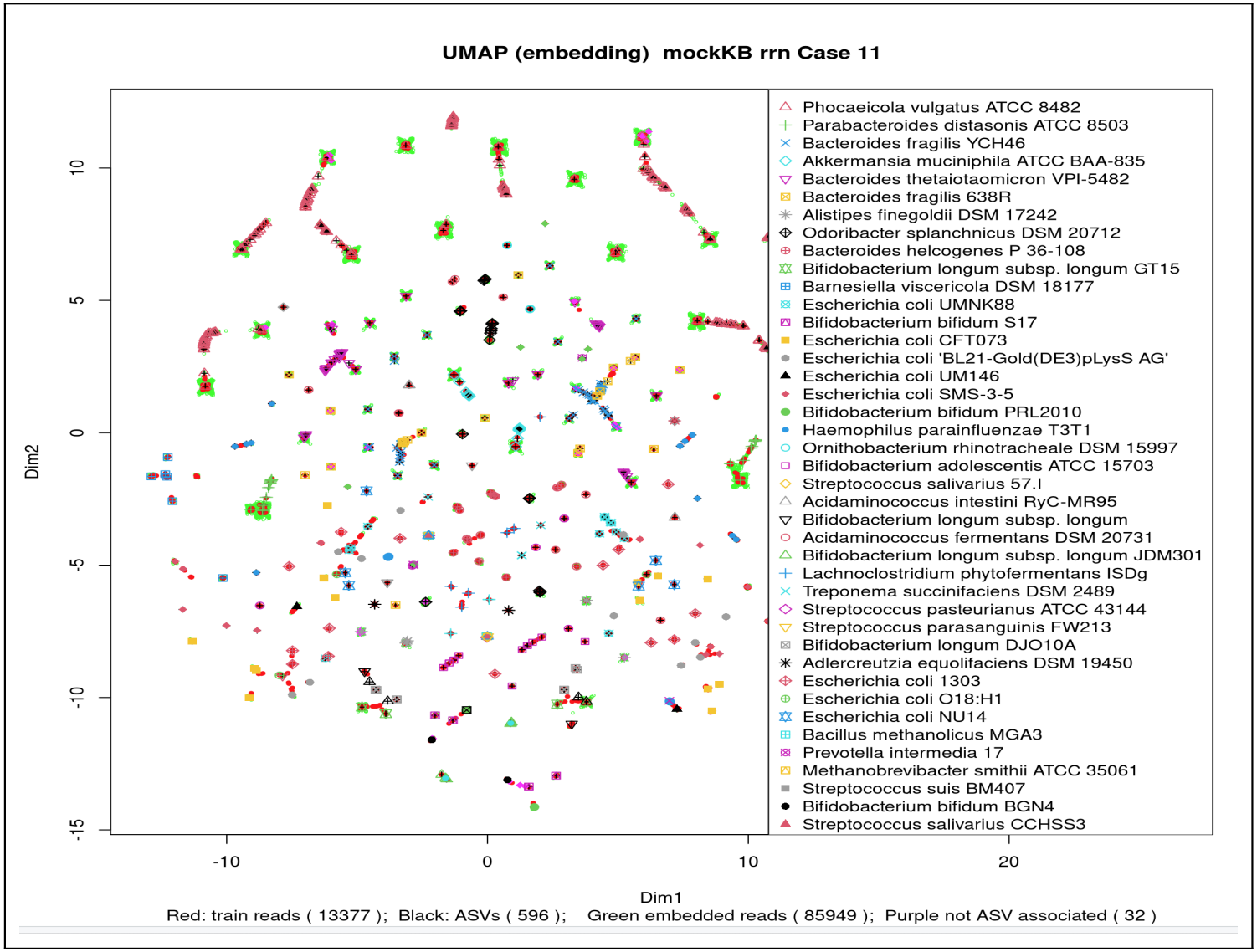

Figure 14: 2D UMAP for dataset `mKB_rrn_C10`, with embedding of most reads of high-read-count ASVs

Multi-strain Species Visualisations

- This microbiome has multiple species that have more than one strain present.
- ASVs and their associated reads were selected for each of 4 species - *Bifidobacterium longum*, *Bifidobacterium breve*, *Escherichia coli* and *Bacteroides fragilis*.
- Each of these, separately, had a 2D UMAP generated and they are presented next.
- These visualisations are key evidence for the strain-level resolution that can be obtained.

Identification - *Bifidobacterium longum* and *Bifidobacterium bifidum*

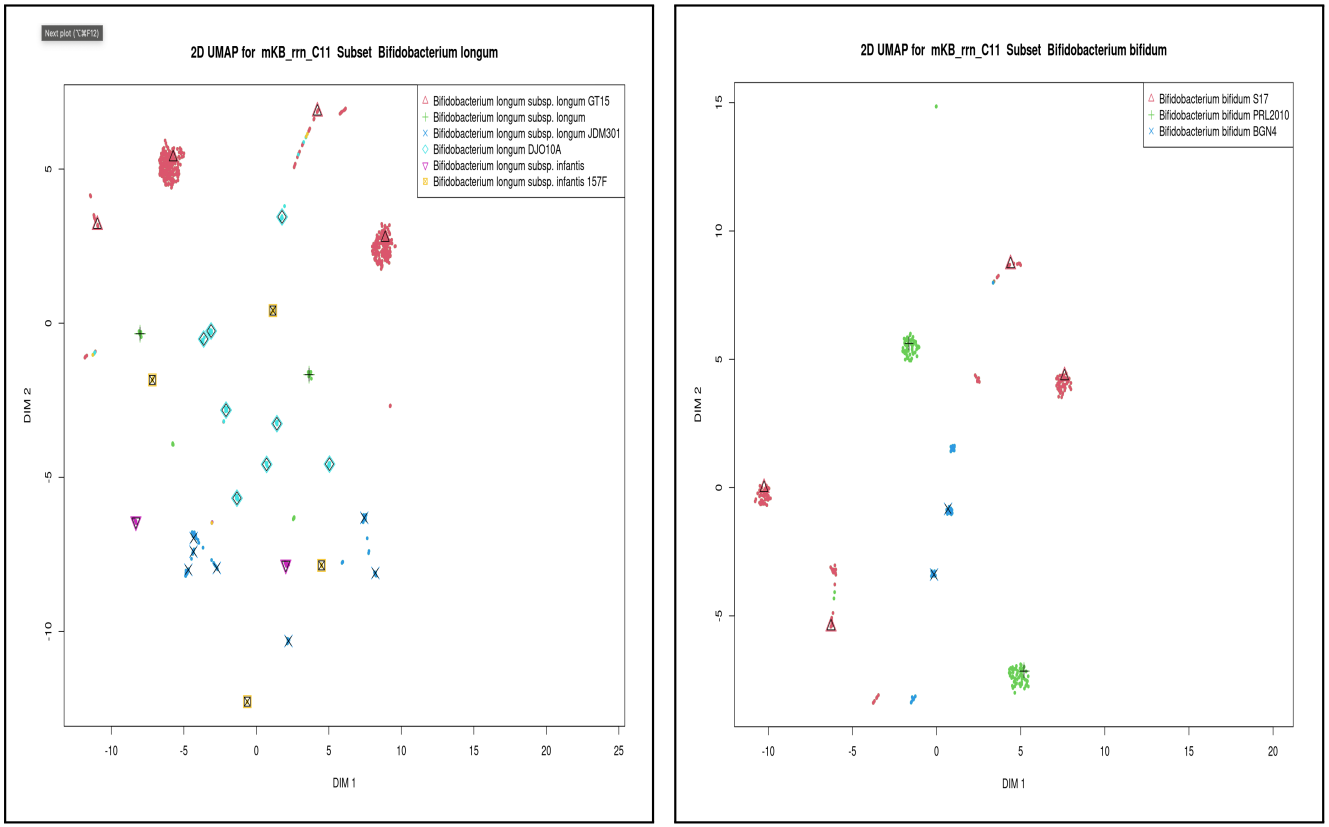

Figure 15: 2D UMAP representation of (left) *Bifidobacterium longum* strains and (right) *Bifidobacterium bifidum* strains.

Identification - *Bacteroides fragilis* and *Escherichia coli*

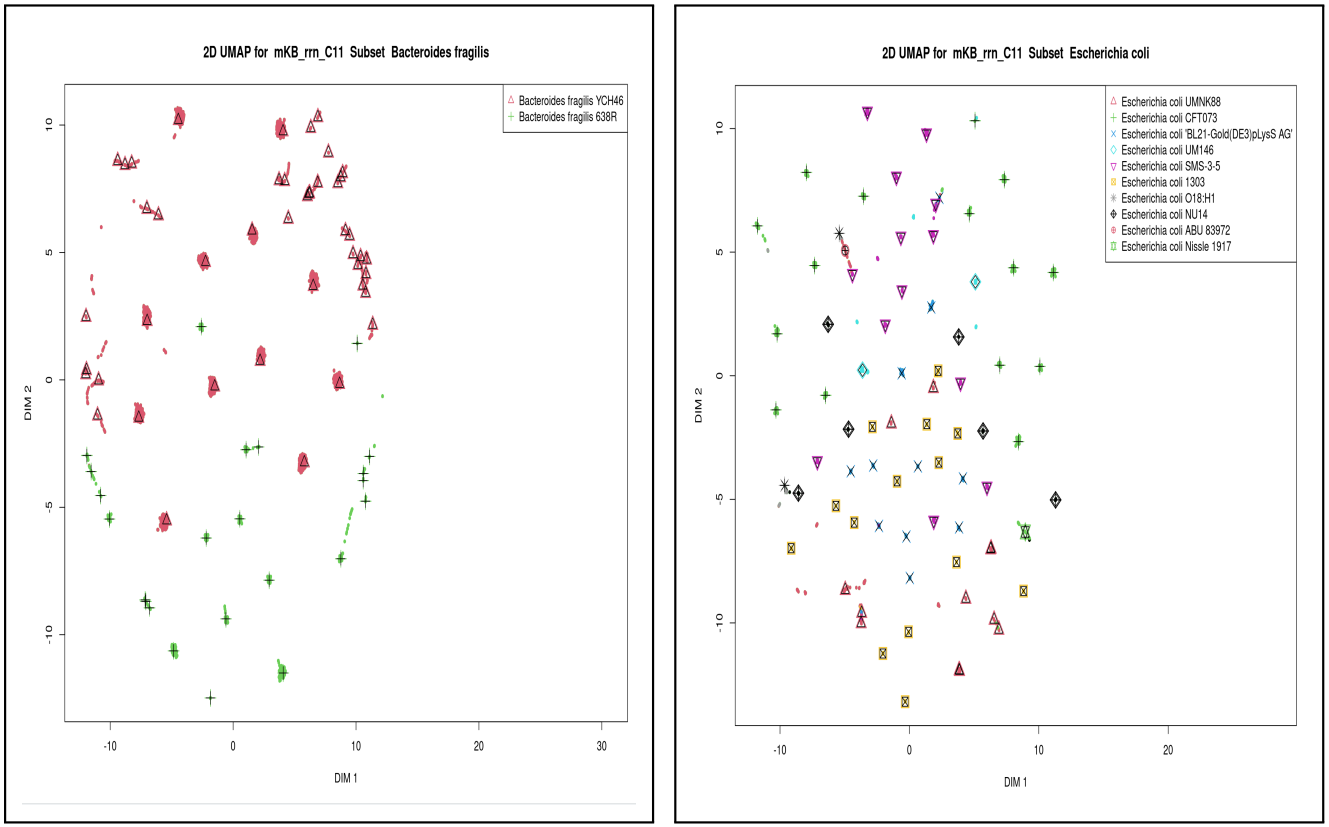

Figure 16: 2D UMAP representation of (left) *Bacteroides fragilis* strains and (right) *Escherichia coli* strains.

Identification - Summary

- There a 4 *Bifidobacterium longum* strains observed 3 of which are in the design. Two less specific database entryies(subspecies *longum* and *infantis)* are also listed, and and an invalid strain identification *infantis 157F given.*
- *There a 2 Bifidobacterium bifidum* strains observed, both of which are in the design. No invalid strains are listed.
- The *Bacillus fragilis* plot has 2 strains observed, both of which are in the design. No invalid strains are listed.
- For *Escherichia coli* the plot shows 9 strains, but 2 of those are not in the designed mock microbiome. One strain that was in the design was not identified.
- Relative abundances of the strains correctly identified in these multi-strain species range from 0.070 to 0.0017 or rank order 5 through to 33, while missed strains have RA ranks of 30, 37, 40 and 51.

Quantification - strain. mKB_rrn_C11

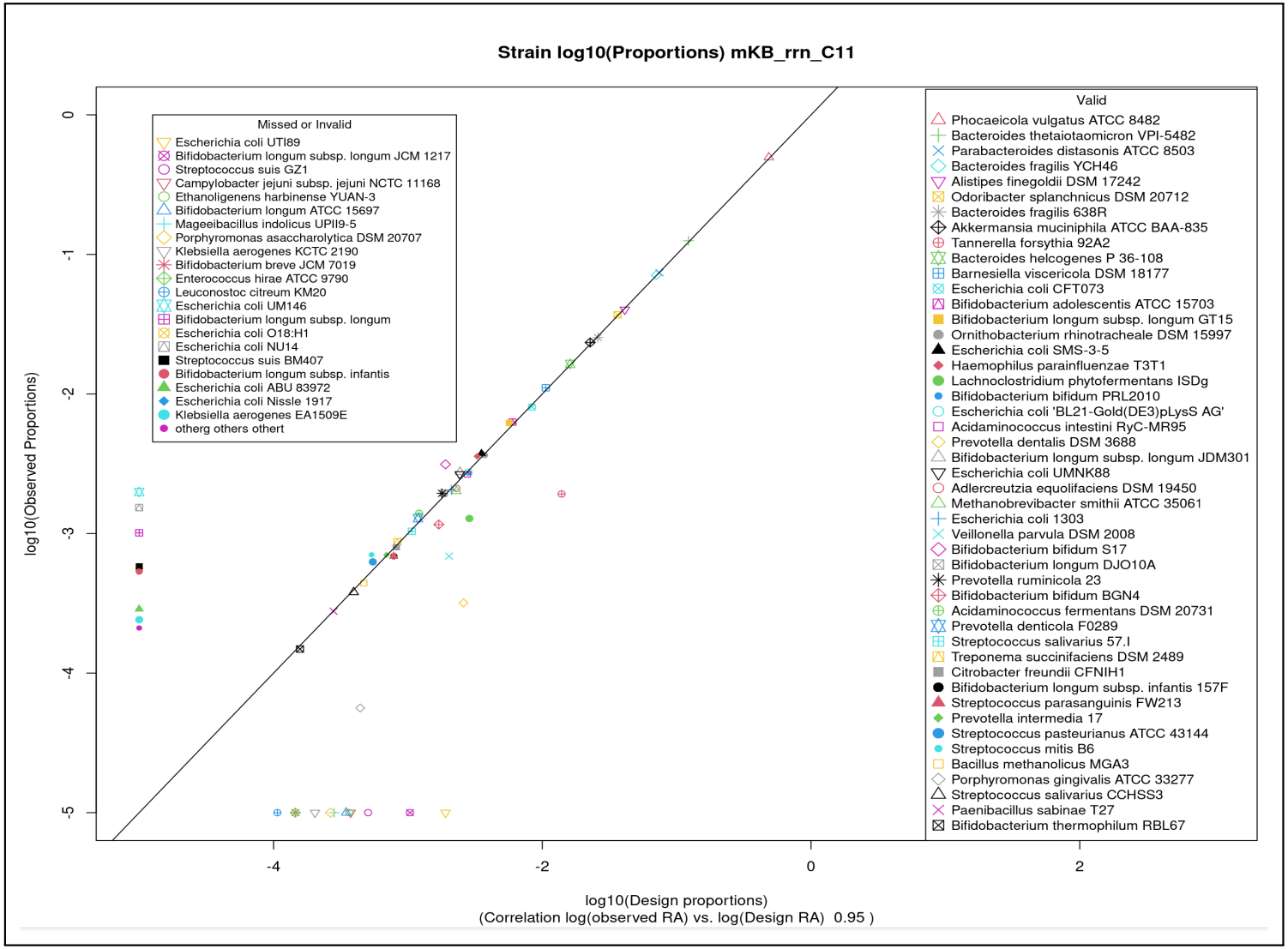

Figure 17: Relation between the observed strains' relative abundance and that of the designed microbiome as determined from badread-simulated full rrn sequences.

Quantification - species. mKB_rrn_C11

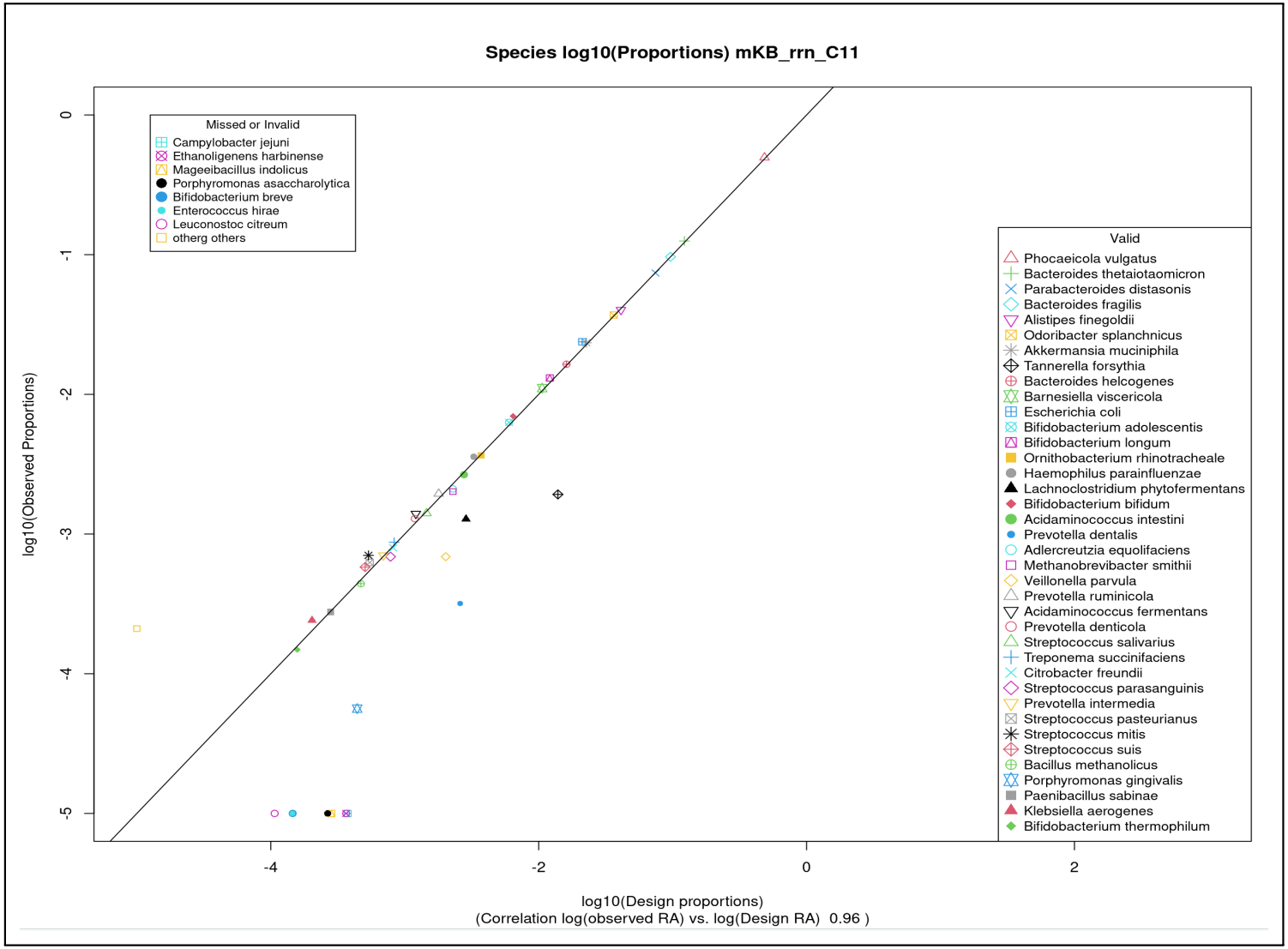

Figure 18: Relation between the observed species' relative abundance and that of the designed microbiome as determined from badread-simulated full rrn sequences.

Quantification - genus. mKB_rrn_C11

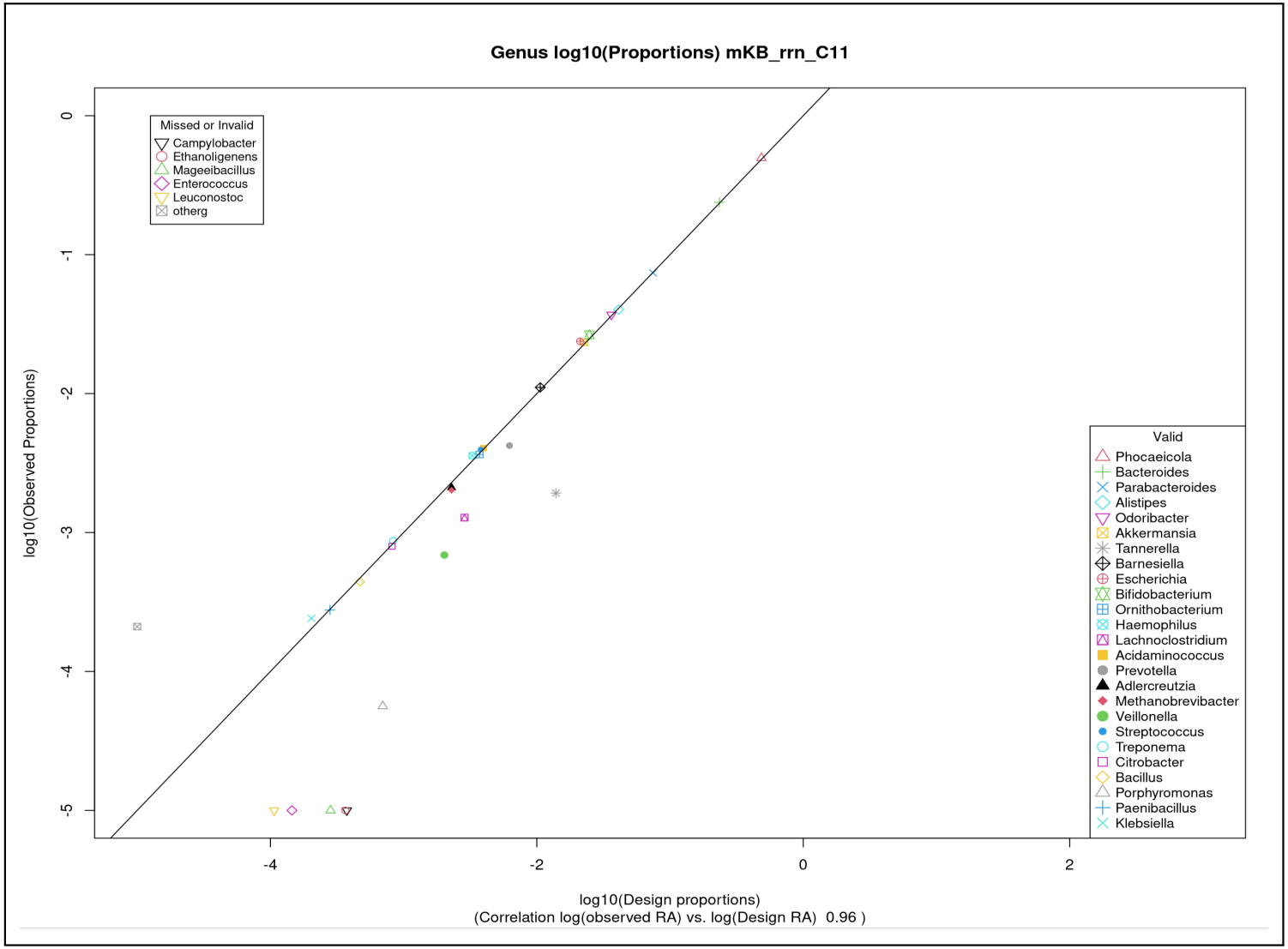

Figure 19: Relation between the observed genera's relative abundance and that of the designed microbiome as determined from badread-simulated full rrn sequences.

Quantification - summary

- Observed abundances range over approximately 3 orders of magnitude.
- Observed abundances are strongly correlated with the design abundances.

Outcomes

- These results show that, with appropriate processing of good quality read libraries from current ONT nanopore sequencing technology, strain-level taxonomic resolution of bacterial microbiomes is feasible.
- Furthermore, such resolution is achieved for strains of species that constitute less than 1% of the microbial cellular content - implying that the individual strains themselves are a smaller component of the microbiome.
- The analysis has already demonstrated the value of good simulation and its appropriate use. This is without the multitude of additional analyses that could usefully be implemented via simulation.

Limitations

- The real data lacked taxonomic depth and consisted of a small number of bacteria in the mock microbiomes.
- The databases used, despite being recently published, are quite small by comparison with databases for analyses limited to higher taxonomic level such as species level.
- Ribosomal RNA features by themselves are not adequate for accurate identification of very close strain. This is particulalry pertinent when it comes to medical uses of engineered strains, or the use of probiotics - especially if they are to be used for medical treatment purposes (e.g. gut dysbiosis in infants).

Final Comment

- Thanks to
  - ⋆ Terry for providing ongoing critiquing and support of what I attempt.
  - ⋆ An unknown reviewer whose review was very valuable in encouraging me to take on this simulation task.
  - ⋆ Calum Walsh, Doherty Institute, for information about the GROND database plus more.
  - ⋆ Ben Murrell, Karolinska Institute, for assistance with using RAD.
  - ⋆ Denis Beslic, Robert Koch Institute, Berlin, for help with seq2squiggle use.
  - ⋆ My former student, Zhengming Zhang, for 23S in-silico primer identification and verification.
  - ⋆ All of you for listening and questioning.

Denis Beslic, Martin Kucklick, Susanne Engelmann, Stephan Fuchs, Bernhard Y. Renard, and Nils Ko€rber. End-to-end simulation of nanopore sequencing signals with feed-forward transformers. *Bioinformatics*, 41(1), January 2025.

Hasindu Gamaarachchi, James M. Ferguson, Hiruna Samarakoon, Kisaru Liyanage, and Ira W. Deveson. Simulation of nanopore sequencing signal data with tunable parameters. *Genome Research*, 34:778–783, May 202.

Charles H. King, Hiral Desai, Allison C. Sylvetsky andJonathan LoTempio, Shant Ayanyan, Jill Carrie, Keith A. Crandall, Brian C. Fochtman, Lusine Gasparyan, Naila Gulzar, Paul Howell, Najy Issa, Konstantinos Krampis, Lopa Mishra, Hiroki Morizono, Joseph R. Pisegna, Shuyun Rao, Yao Ren, Vahan Simonyan, Krista Smith, Sharanjit VedBrat, Michael D. Yao, and Raja Mazumder.

Baseline human gut microbiota profile in healthy people and standard reporting template.
PLoS One, 14(9):e0206484, September 2019.

Mantas Sereika, Rasmus Hansen Kirkegaard, Søren Michael Karst, Thomas Yssing Michaelsen, Emil Aarre Sørensen, Rasmus Dam Wollenberg, and Mads Albertsen.
Oxford nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing.
Nature Methods, 19:823–826, July 2022.

Meghana Srinivas, Calum J. Walsh, Fiona Crispie, Orla O'Sullivan, Paul D. Cotter, Douwe van Sinderen, and John G. Kenny.
Evaluating the efficiency of 16S-ITS-23S operon sequencing for species level resolution in microbial communities.
Scientific Reports, 15:2822, January 2025.

Ryan Wick.
Badread: simulation of error-prone long reads.
Journal of Open Source Software, 4(36):1316, 2019.

Chen Yang, Justin Chu, RenÃ© L Warren, and InanÃ§ Birol.
Nanosim: nanopore sequence read simulator based on statistical characterization.
Gigascience, 6(4):1–9, 2017.

Sereika Sub-sampled Datasets

Table 2: Sub-sampling factors on reads from species in the Sereika-derived 16S and 23S rRNA gene datasets used to generate four mock microbiomes for each such gene that cover a range of either 50:1 or 100:1 in relative cellular abundance of the species. Species key: Bs=*Bacillus subtilis* , Ef=*Enterococcus faecalis* , Ec=*Escherichia coli* , Lm=*Listeria monocytogenes* , Pa=*Pseudomonas aeruginosa* , Se=*Salmonella enterica* , Sa=*Staphylococcus aureus* .

Dataset	Bs	Ef	Ec	Lm	Pa	Se	Sa
Sub1	0.1	1	1	1	1	0.5	0.2
Sub2	0.5	0.01	0.1	1	1	1	1
Sub3	1	0.02	1	1	1	1	0.02
Sub4	1	1	0.1	1	1	1	0.02

Quantification 16S Sub1 Sub2

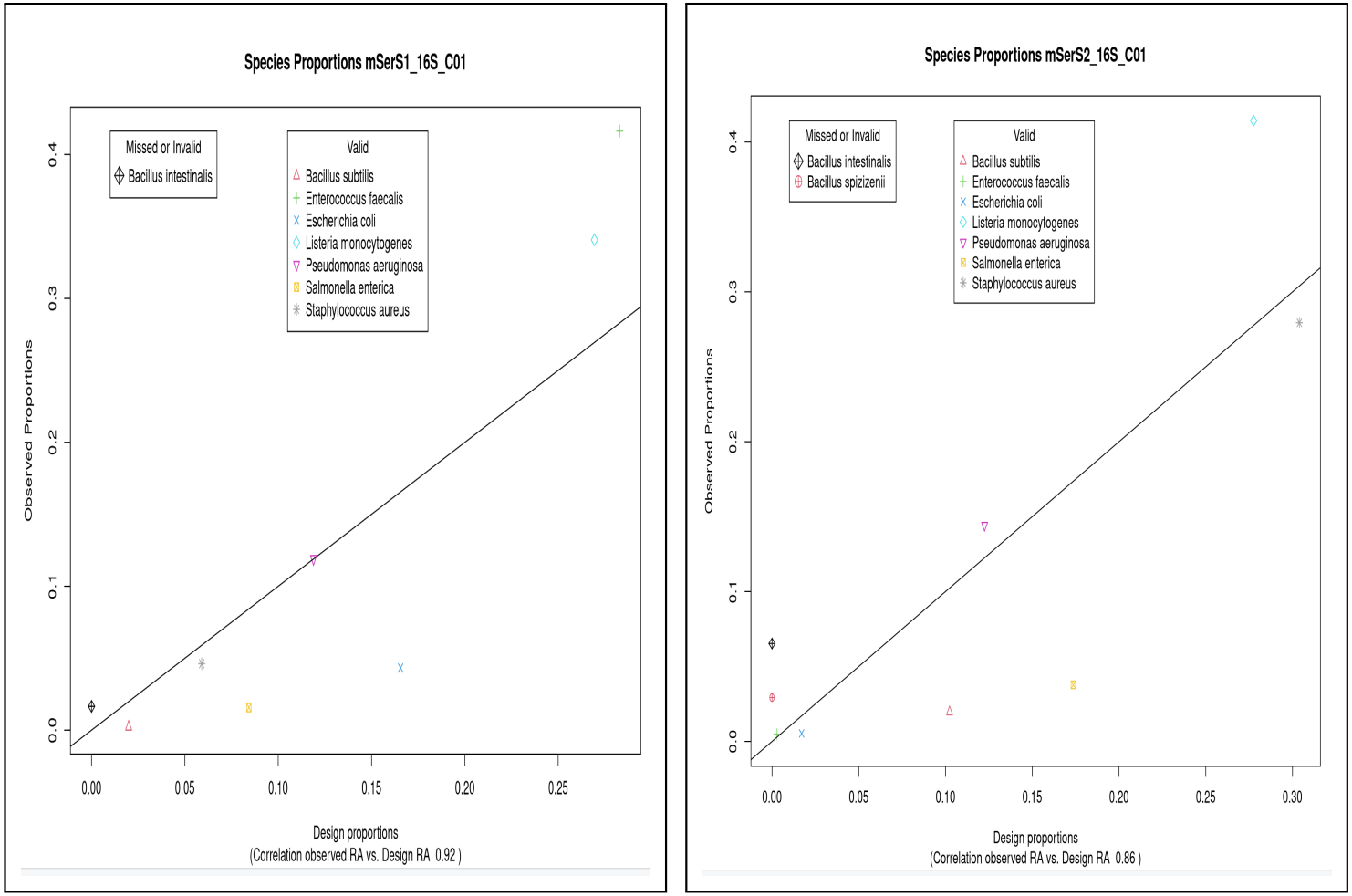

Figure 20: Relation between the observed species relative abundances and those of the designed microbiomes for sub-samples Sub1 and 2 as determined from 16S rRNA genes.

Quantification 16S Sub3 Sub4

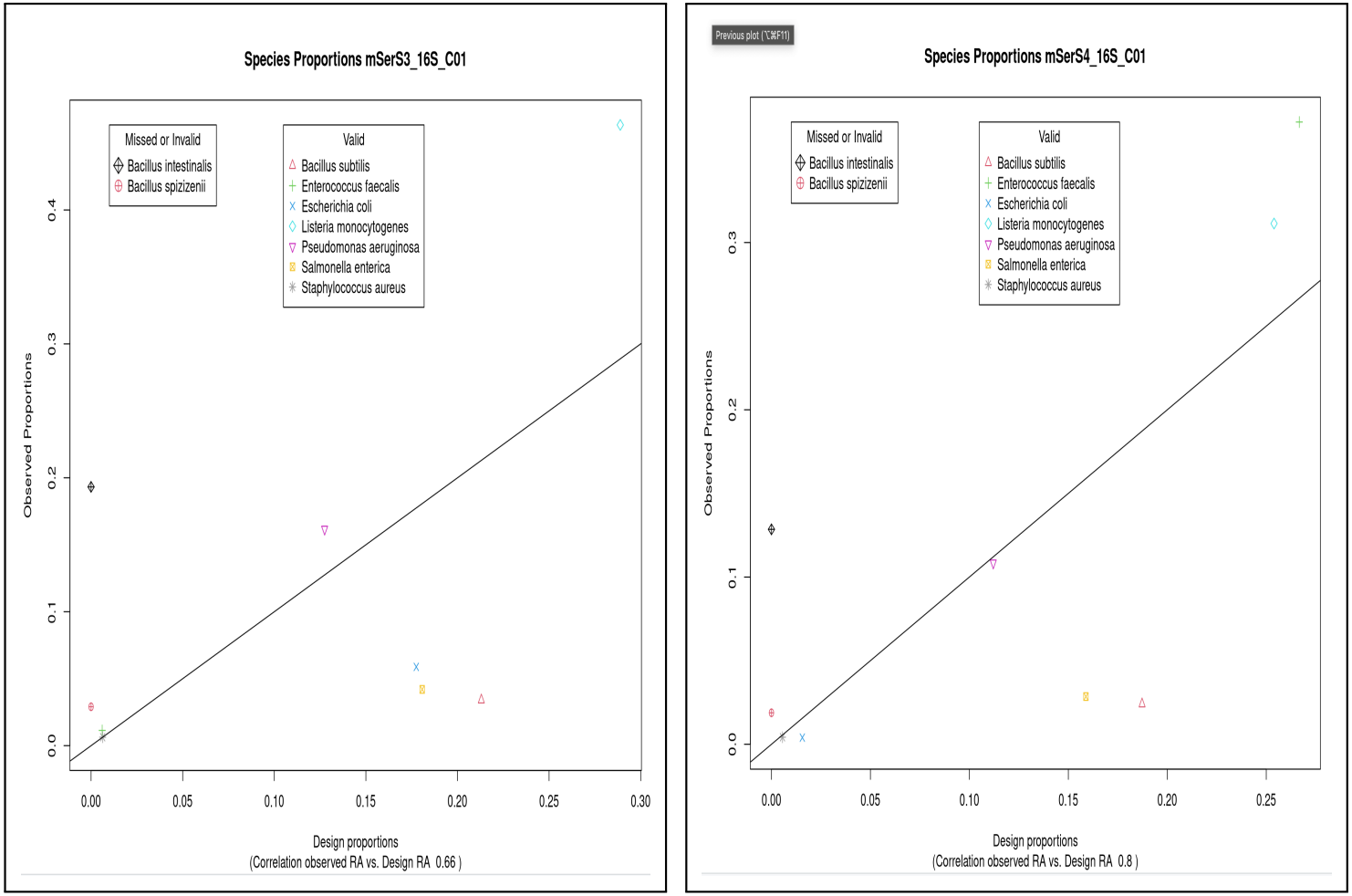

Figure 21: Relation between the observed species relative abundances and those of the designed microbiomes for sub-samples Sub 3 and 4 as determined from 16S rRNA genes.

Quantification 23S Sub1 Sub2

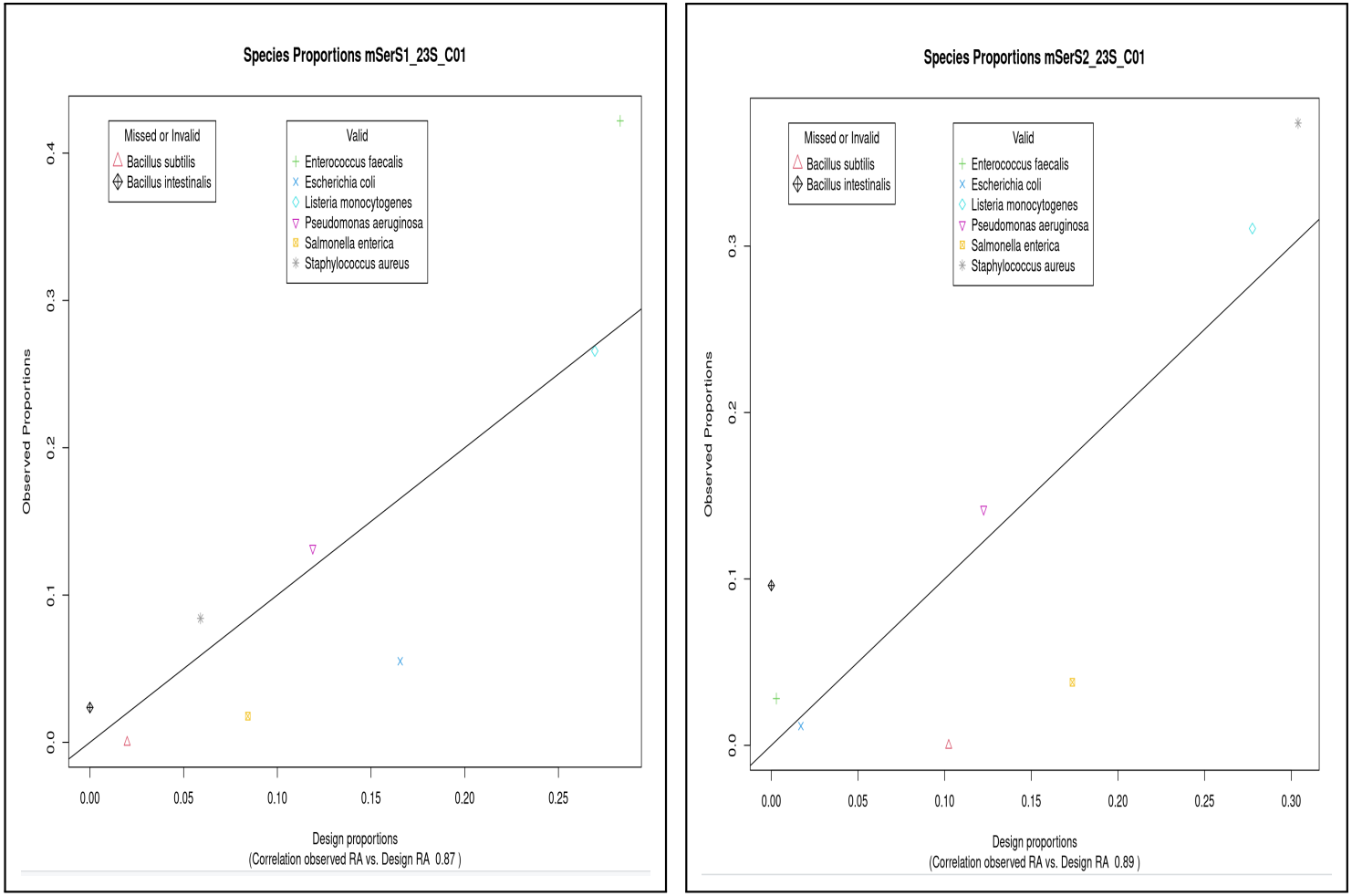

Figure 22: Relation between the observed species relative abundances and those of the designed microbiomes for sub-samples Sub1 and 2 as determined from 23S rRNA genes.

Quantification 23S Sub3 Sub4

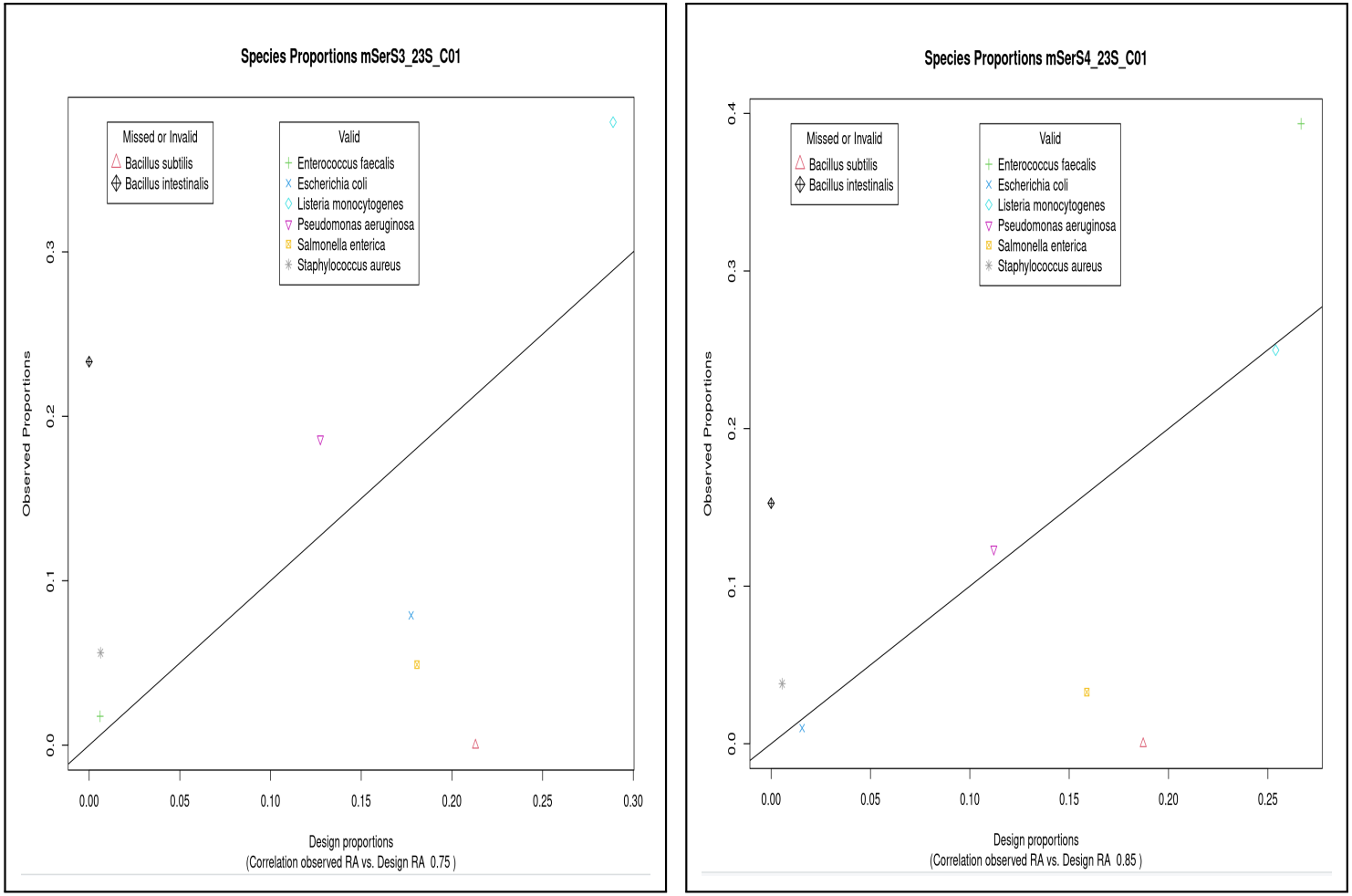

Figure 23: Relation between the observed species relative abundances and those of the designed microbiomes for sub-samples Sub 3 and 4 as determined from 23S rRNA genes.