

## Introduction

Advancements in sequencing technologies have made metagenomic analyses of complex microbial samples routine and accessible. Mock communities of known composition are often run in parallel to allow for accurate data evaluation and to facilitate cross-study and inter-lab comparisons, yet they lack the microbial diversity of real-world samples. The **ZymoBIOMICS Fecal Reference with TruMatrix Technology** (D6323) is a highly diverse pooled human fecal reference that provides a truly complex alternative to mock communities. However, the microbial content of this standard is only partially characterized, and species level composition remains underexplored. Here, we explore the content of this sample using highly accurate long-read sequencing.

## Methods

### PacBio HiFi sequencing

- Shotgun metagenomics: 4 SMRT Cells (8M) on the Sequel IIe system
  - 11.9 million HiFi reads with a mean length of ~8 kb, for a total of 88.3 Gb of data. HiFi read median QV was 41, representing >99.99% accuracy
  - Down-sampled the full dataset to investigate effects on analyses and ran additional SMRT Cell 8M to test 48-plex barcoded samples to confirm
- Full-length 16S sequencing (V1-V9)

### Metagenome taxonomic and functional profiling

- Profiling was performed using **DIAMOND**<sup>1</sup> and **MEGAN-LR**<sup>2</sup> with the NCBI-nr protein database
- Analysis was automated with the PacBio **Taxonomic-Functional-Profiling-Protein** workflow (available on github: [PacificBiosciences/pb-metagenomics-tools](https://github.com/PacificBiosciences/pb-metagenomics-tools)), with long-read settings and filtering optimized for high precision species detection<sup>3</sup>

### Full-length 16S taxonomic profiling

- Profiling was performed using **QIIME 2**<sup>4</sup> and **DADA2**<sup>5</sup> with the GTDB database
- Analysis was automated with the PacBio **Full-length 16S analysis** workflow that is available on github: [PacificBiosciences/pb-16S-nf](https://github.com/PacificBiosciences/pb-16S-nf)

### Metagenome assembly

- Assembly was performed using **hifiasm-meta**<sup>6</sup>
- The PacBio **HiFi-MAG-Pipeline** was used to identify high-quality metagenome assembled genomes (HQ MAGs) using a **circular-aware binning** strategy

### Phase Genomics ProxiMeta metagenome deconvolution

- PacBio HiFi assembly + ProxiMeta Hi-C binning

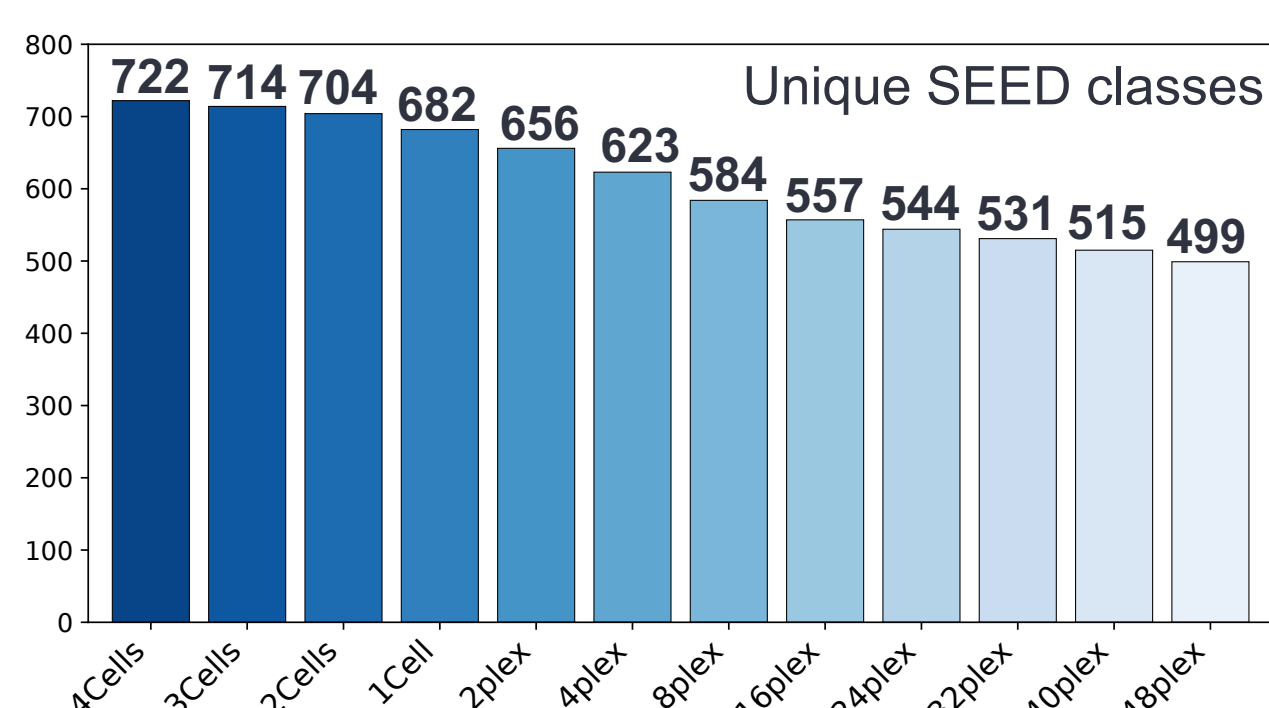
## Metagenome functional profiling

Functional profiling using **DIAMOND** and **MEGAN-LR** resulted in:

- ~92% of reads received at least one functional annotation
- Over 66.9 million functional annotations across all databases (Table 1)
- Small decrease in number of unique classes with decreasing data levels (Fig. 1)

Database	Total annotations	Unique classes	Annotations per read (mean)
EC	13.1 million	2,714	2.3
eggNOG	11.3 million	2,714	3.2
InterPro2GO	25.1 million	17,428	4.1
SEED	17.4 million	722	2.2

**Table 1. Functional annotations.** A summary of functional annotations derived from the four databases used by MEGAN-LR, based on protein alignments inferred with DIAMOND.

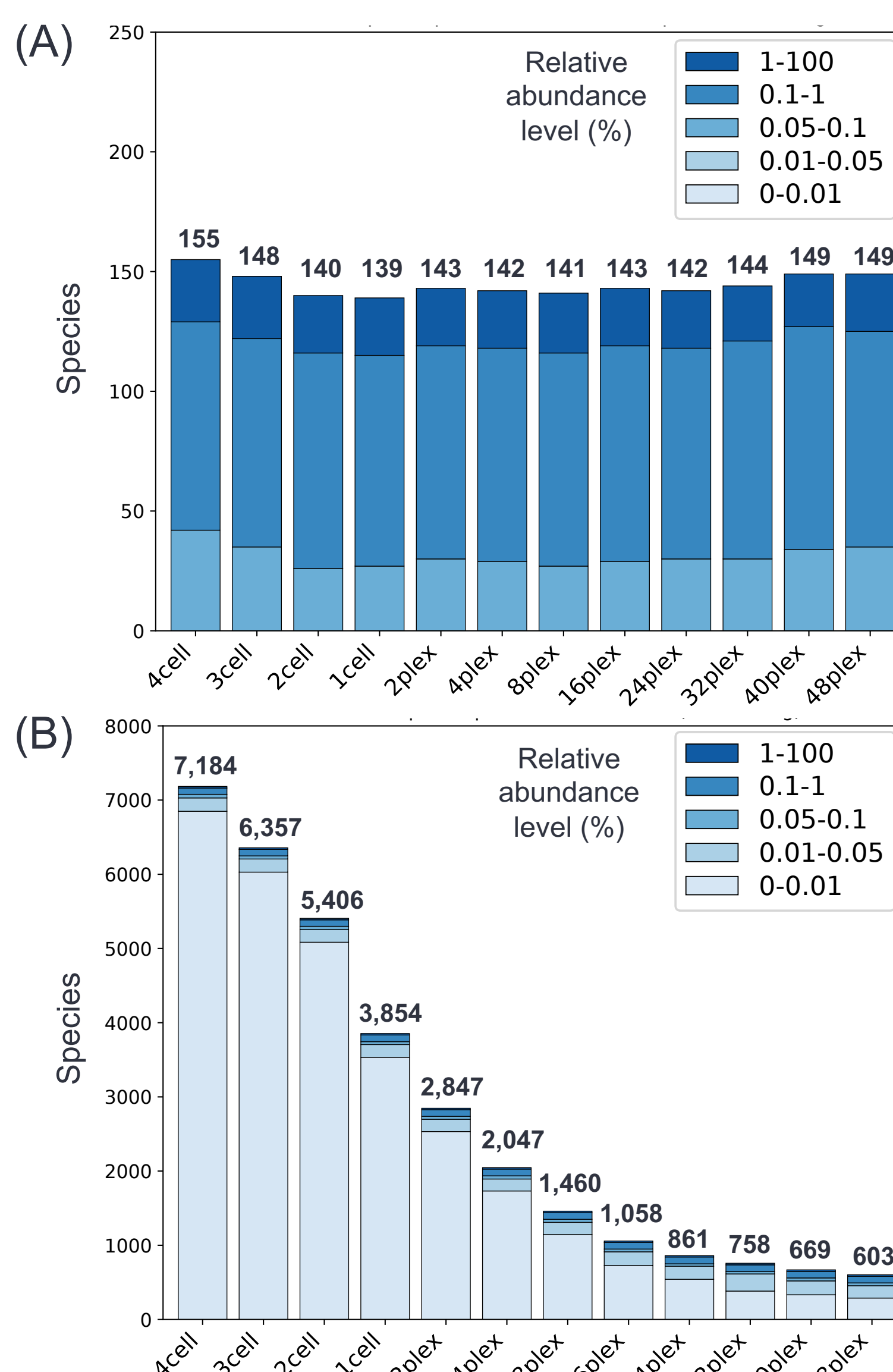


**Figure 1. Total data vs. unique functional classes.** A summary of the number of unique SEED classes assigned across reads, based on total data. At the lowest data level, 70% of the unique classes from the full dataset were detected. The trend was similar for the other three functional databases.

## Metagenome taxonomic profiling

Taxonomic profiling using **DIAMOND** and **MEGAN-LR** resulted in:

- Detection of 155 species (80 genera) in high precision mode (Fig. 2A)
- Detection of 7,184 species (~2,000 genera) in low precision mode (Fig. 2B)
- Consistent profiles across data levels using high precision mode (Figs. 2, 3)

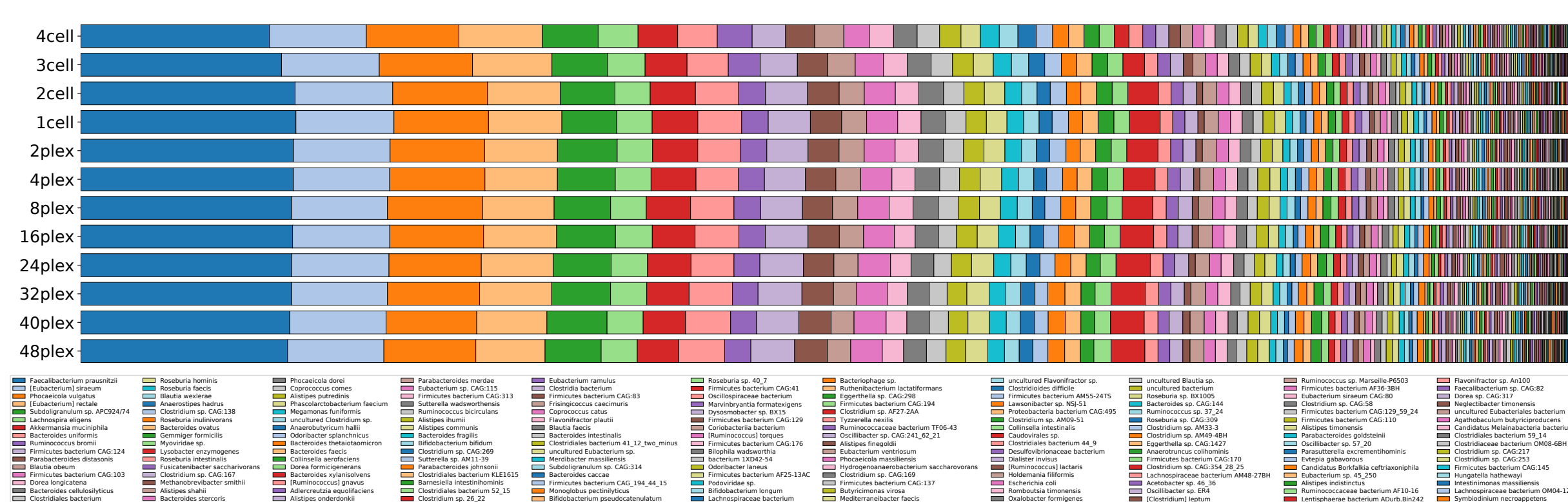


**Figure 2. Species detection.**

A summary of the number of species detected across down-sampled data levels. Colors indicate the number of species detected in each relative abundance category, and the total number of species is shown on top. The full dataset is shown on the left, with decreasing data levels to the right.

(A) Results for the high precision filtering mode, in which a minimum threshold of reads must be assigned to a particular species to report it. The lower limit for detection in this mode is ~0.03% relative abundance.

(B) Results for the low precision mode, in which no threshold filtering is applied. Here, species can be represented by a single matched read. Across data levels, nearly 95% of assignments are at the ultra-low (<0.01%) or very low (0.01–0.05%) abundance levels (85% and 10%, respectively).

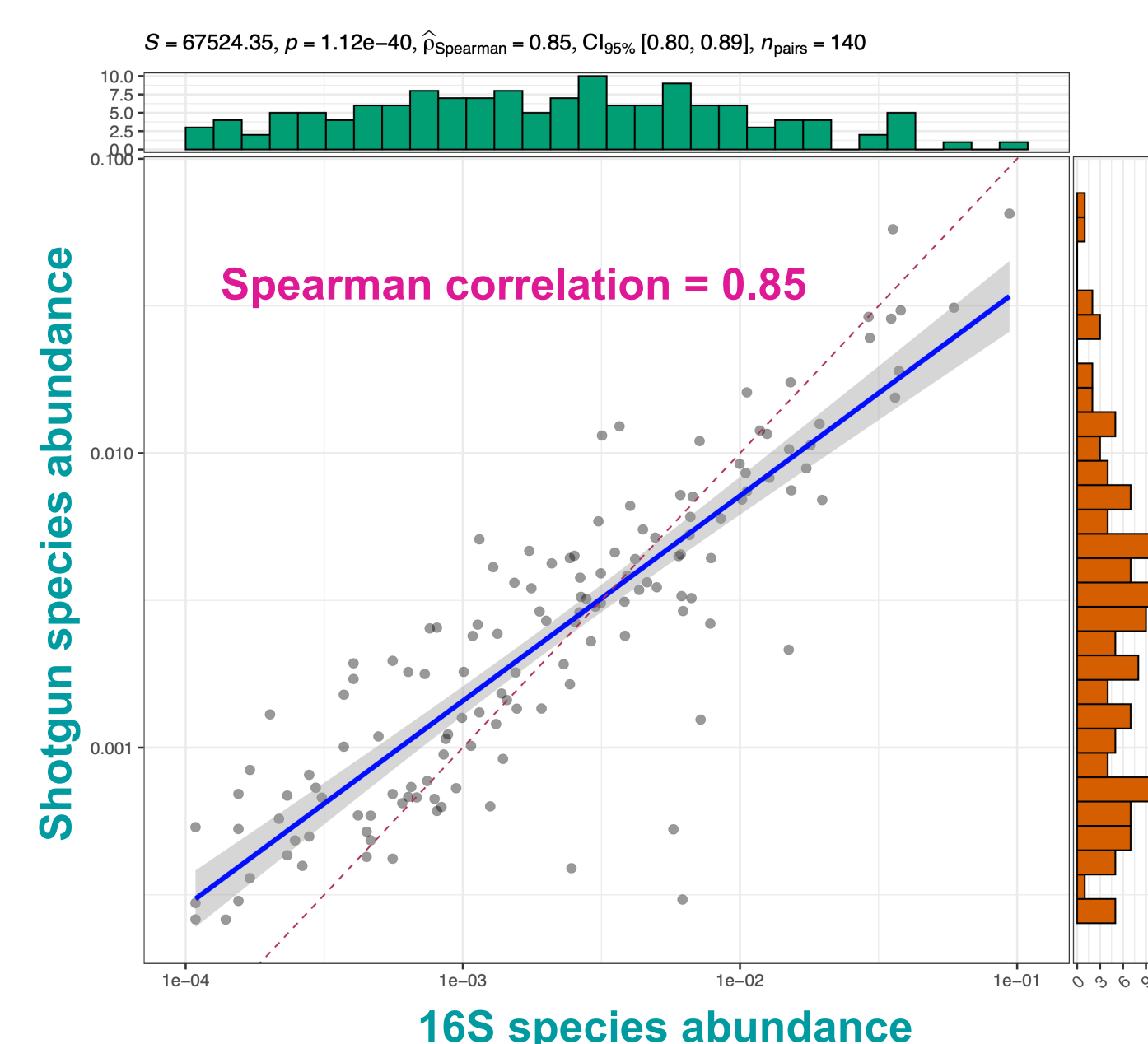


**Figure 3. Relative abundance profiles.** A comparison of the relative abundances of species across different data levels. Within a row, each color represents a distinct species and the width of the bar indicates its relative proportion in the community. The abundance profiles result from the high precision filtering mode and display high similarity across 0.5–88 Gb of data. The 48-plex barcoded run yielded similar results (not shown).

## Full-length 16S taxonomic profiling

Taxonomic profiling using **QIIME 2** and **DADA2** resulted in:

- Detection of 205 species
- 140 species** classified in both dataset accounts for **88% of assigned 16S reads and 74% of assigned shotgun reads**
- Unique species are all <2%
- Spearman correlation is 0.85 at the species level (Fig. 4)
- Down-sampled 16S (denoised) reads to ~64k reads
- Shotgun metagenomics data down-sampled to simulate 48-plex (~60k reads)

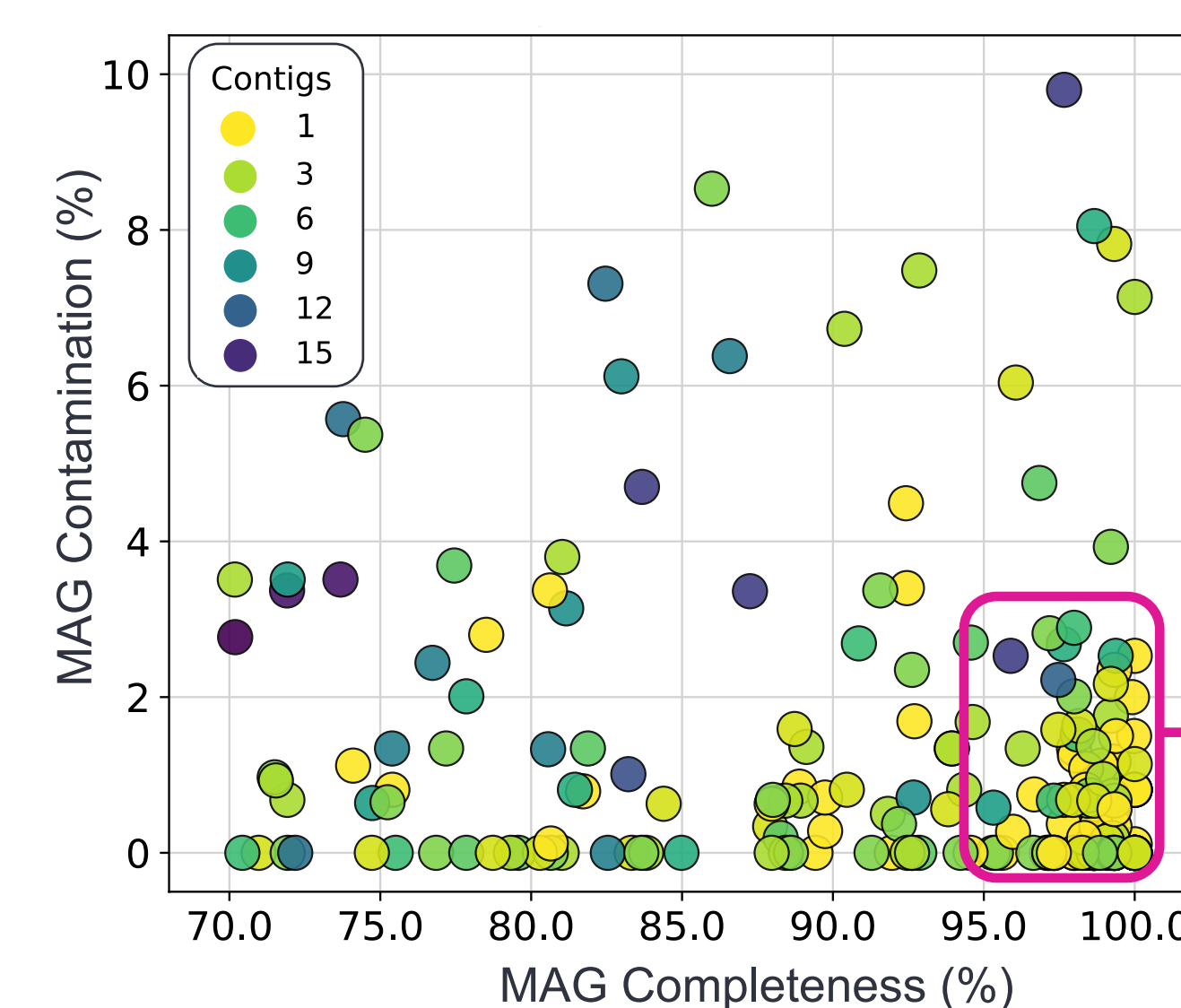


**Figure 4. 16S species abundance vs shotgun species abundance.** A correlation of the relative abundances of species between 16S and shotgun metagenomics classifications. At the species level, correlation is 0.85. At the genus level, correlation increases to 0.90, and account for 87% of assigned shotgun reads.

## Metagenome assembly

Assembly with **hifiasm-meta** and evaluation with **HiFi-MAG-Pipeline** produced:

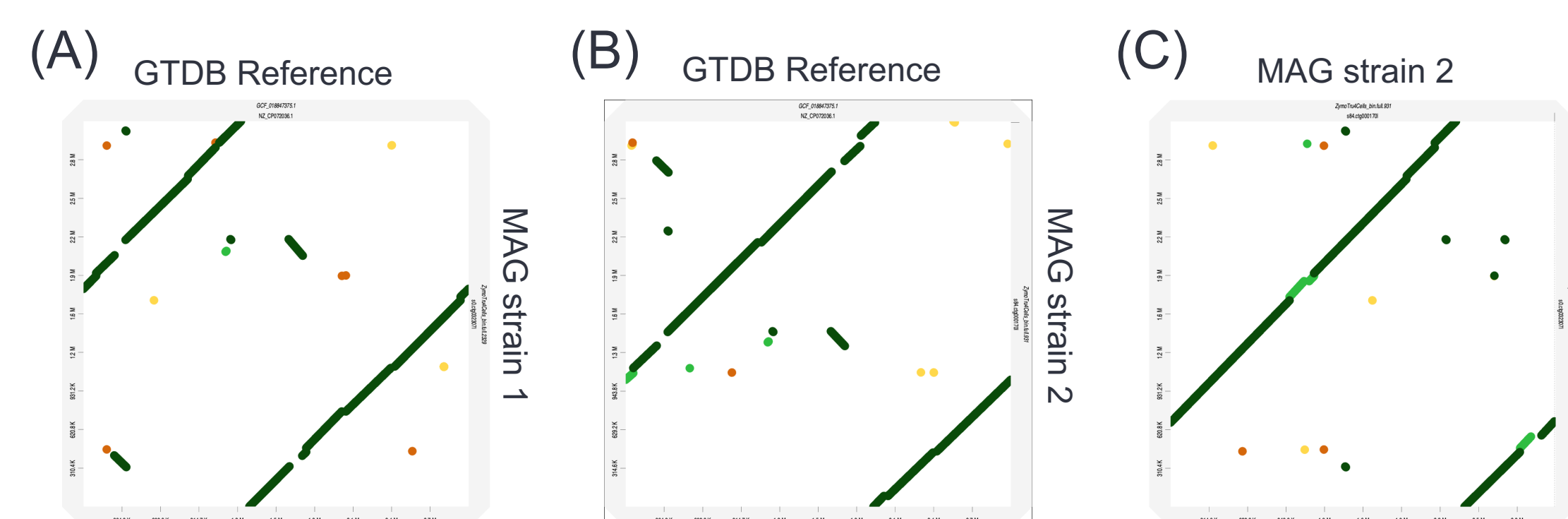
- ~2600 genome bins
- 199 total HQ-MAGs; 72 are single contig, 102 are >95% complete (Fig. 5)
- HQ-MAGs from 164 species, 114 genera, and 43 families
- 28 species represented by 2–3 MAGs (e.g., strain-level variation; Fig. 6)



**Figure 5. MAG characteristics.** Completeness versus contamination scores for the 199 HQ-MAGs, as evaluated by CheckM. Each dot represents a MAG, and colors indicate the number of contigs the MAG contains. We found 102 HQ-MAGs (51%) displayed >95% completeness. Furthermore, 54 HQ-MAGs (27%) were >95% complete and composed of a single (and often circular) contig.

**Exceptionally high-quality MAGs**

- 102 MAGs in this zone
- 54 are single contig



**Figure 6. Strain variation.** We assembled two highly complete MAGs for *Akkermansia muciniphila*\_B. The D-GENIES dotplots show strain variation across comparisons, including between (A) the Genome Taxonomy Database (GTDB) reference and MAG 1, (B) GTDB and MAG 2, and (C) between the MAGs.

## ProxiMeta metagenome deconvolution

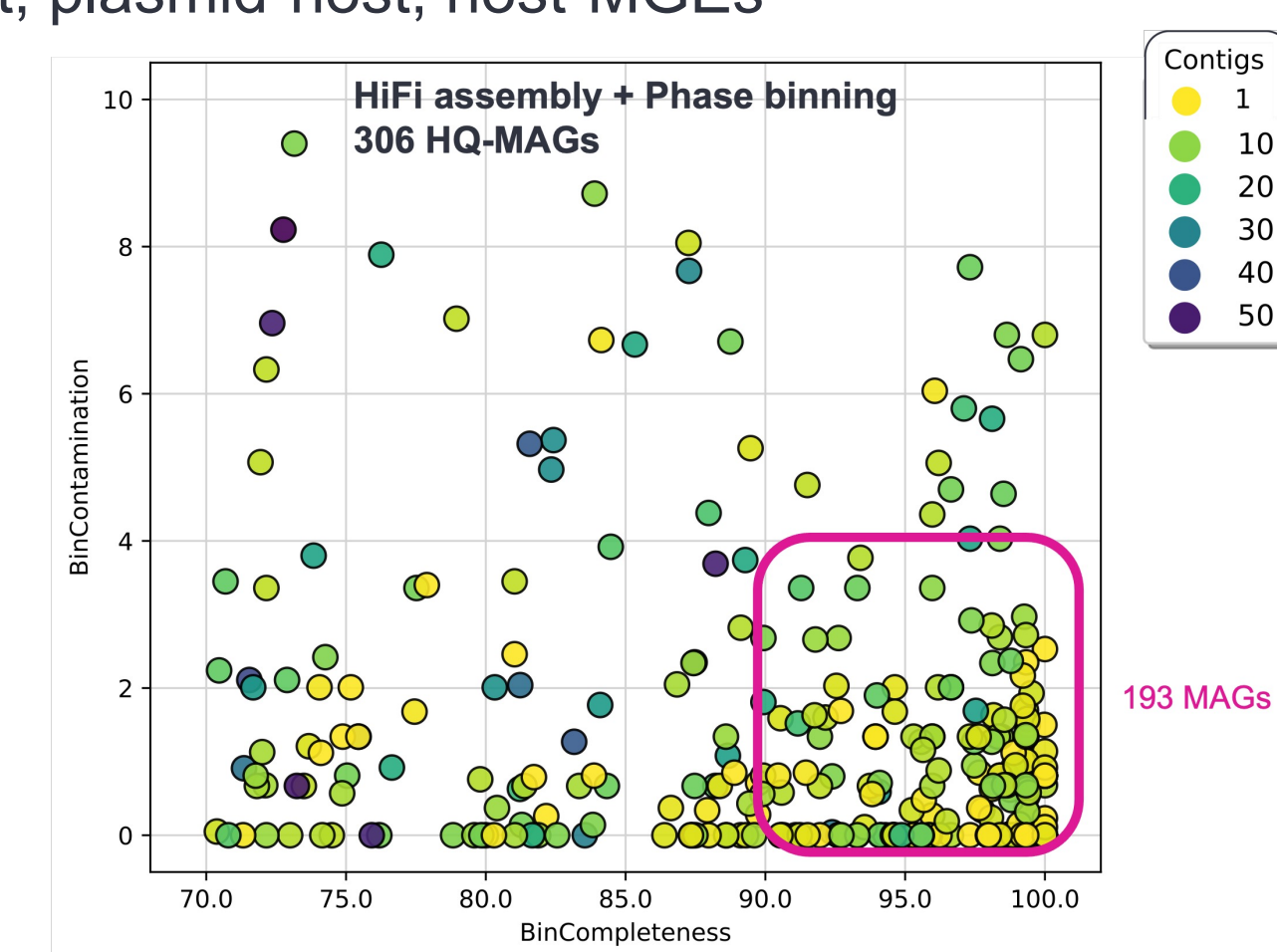
Assembly with **hifiasm-meta** and binning with Phase Genomics

ProxiMeta produced:

- 306 total HQ-MAGs (>70% complete; Fig. 7)
- 193 total HQ-MAGs (>90% complete)
- BGCs, AMR gene-host, virus-host, plasmid-host, host-MGEs

### HiFi assembly + Phase binning MAG characteristics.

Completeness versus contamination scores for the 306 HQ-MAGs, as evaluated by CheckM. Each dot represents a MAG, and colors indicate the number of contigs the MAG contains. We found 193 HQ-MAGs displayed >90% completeness.



## Summary

### Metagenome taxonomic and functional profiling

- Detection of 155 species in high-precision mode
- Up to 7,184 species detected without filtering
- Over 67 million functional annotations (from 12 million reads)
- Consistent taxonomic profiles across data levels

### Metagenome assembly

- Assembled 199 high-quality MAGs
- 54 MAGs are single contig and >95% complete (reference quality)
- With 1 SMRT Cell, recovered 110 high-quality MAGs

### PacBio HiFi + ProxiMeta Hi-C

- 306 high-quality MAGs
- BGCs, AMR gene-host, virus-host, plasmid-host, host-MGEs

### Full-length 16S taxonomic profiling

- High concordance with shotgun metagenome profiles

[PacificBiosciences / pb-metagenomics-tools](https://github.com/PacificBiosciences/pb-metagenomics-tools)



## References

- Buchfink B, et al. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, 12, 59–60.
- Huson DH, et al. (2018). MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biology Direct*, 13, 6.
- Portik DM, et al. (2021). Evaluation of taxonomic profiling methods for long-read shotgun metagenomic sequencing datasets. *bioRxiv*, doi: 10.1101/2022.01.31.478527
- Bolyen E, et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, 37, 852–857.
- Callahan BJ, et al. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, 13, 581–583.
- Feng X, et al. (2022). Metagenome assembly of high-fidelity long reads with hifiasm-meta. *Nature Methods*, Online Early, doi: 10.1038/s41592-022-01478-3