

Kinnex 16S Full-Length Profiling on PacBio Revio

Accuracy, Reproducibility, and Species-Level Resolution

VIB · Nucleomics Core

March 2026

Contents

1. Introduction	1
2. Dataset and Bioinformatics	3
3. Results	4
4. Discussion	10
5. Summary	12
References	12
Code Availability	12
Appendix — Glossary of Technical Terms	13
Acknowledgements	15

1. Introduction

1.1 General Description of the Kinnex 16S Method

The PacBio Kinnex 16S workflow is a full-length amplicon approach targeting the entire bacterial 16S rRNA gene (V1-V9, ~1.5 kb), followed by molecular concatenation into longer SMRTbell libraries for CCS long-read sequencing. The method is based on MAS-Seq concatenation, increasing effective per-cell throughput by 8-12-fold over regular full-length 16S libraries. Combinatorial indexing supports up to 384 samples via barcoded primers, or up to 1,536 samples when combined with 4-plex Kinnex adapters, at DNA inputs as low as 1-2 ng per sample [1,2].

The concatenated library peaks around 18-19 kb, and the Revio platform delivers ~50-60 million deconcatenated 16S reads per SMRT Cell. At high multiplexing, per-sample sequencing cost becomes directly comparable to a V3-V4 short-read run, while delivering substantially richer taxonomic information.

1.2 Library Preparation

Based on the PacBio checklist (103-238-800, Rev06, Jun 2025), the workflow proceeds as follows:

1. Full-length 16S PCR with dual-barcoded primers.
2. Amplicon pooling and SMRTbell bead cleanup.
3. Kinnex PCR (12-way concatenation) and cleanup.
4. Array formation via adapter ligation, enzymatic digestion, DNA repair.
5. Nuclease treatment, final cleanup, QC (Qubit + fragment analysis).
6. Polymerase binding, cleanup, and SMRT Cell loading.

Expected size at each checkpoint: ~1,500 bp monomer after initial PCR; ~18-19 kb final concatenated library. The protocol includes multiple safe stopping points. The HiFi-16S workflow applies read-level quality filtering and subsamples to 20,000 S-reads per sample by default; higher depths can be set via the subsample parameter for complex community surveys.

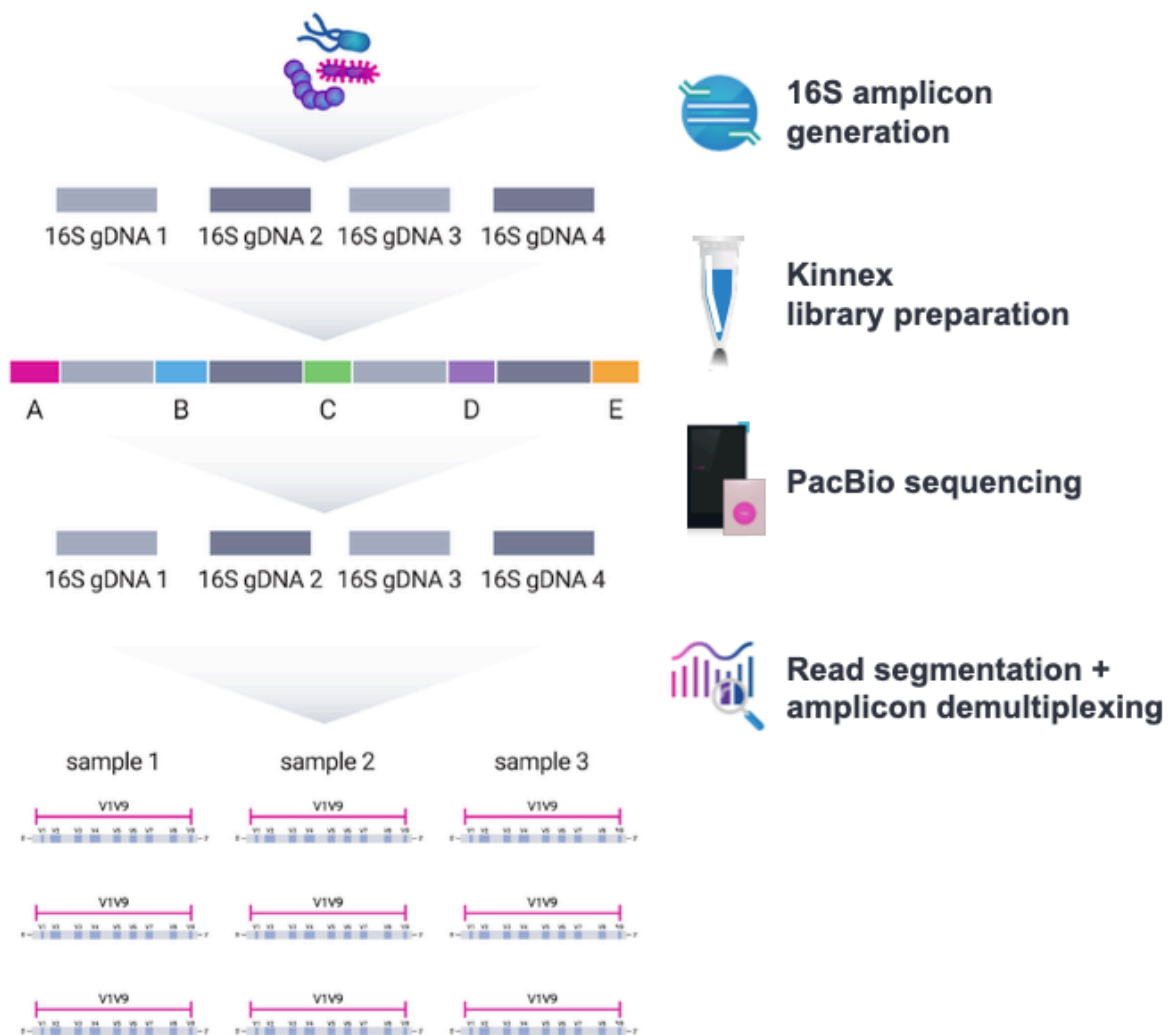


Figure 1: Kinnex library preparation — overview

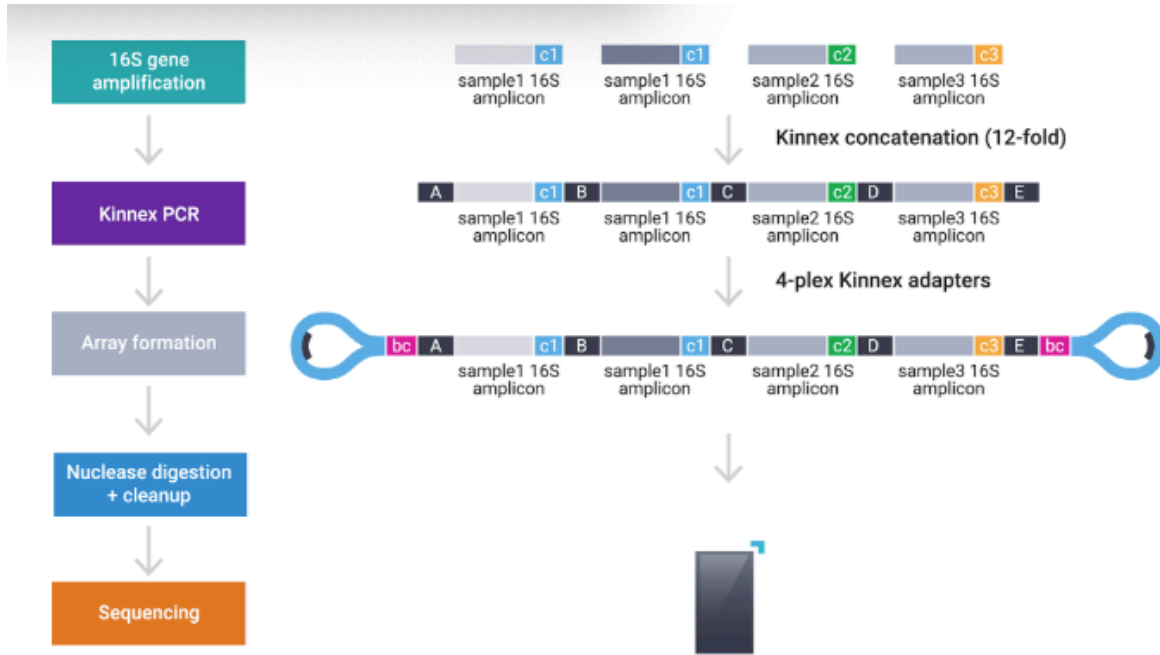


Figure 2: Kinnex library preparation — array assembly detail

1.3 Full-Length 16S vs Short-Read V3-V4

Early long-read sequencing platforms were hampered by high per-base error rates, making them poorly suited for applications requiring accurate sequence discrimination. The introduction of HiFi sequencing on PacBio instruments changed this fundamentally: by generating consensus reads from multiple passes over the same molecule, HiFi achieves per-base accuracy exceeding 99.9% — matching or surpassing that of short-read sequencing — while retaining the read-length advantage of long-read technology.

Short-read 16S surveys commonly target the V3-V4 hypervariable sub-region (~464 bp with the widely used 341F/785R primer pair), a choice historically driven by short-read read-length constraints [3]. V3-V4 performs well for genus-level community profiling but lacks the discriminatory power to separate closely related species whose V3-V4 sequences are identical or near-identical.

Full-length V1-V9 amplicons encode all nine hypervariable regions in a single molecule, providing cumulative sequence diversity that is sufficient for robust species-level — and in many cases strain-level — assignment [4]. The additional variable regions (particularly V1, V2 and V6-V9) are the primary source of this discriminatory advantage. Representative examples of species pairs that are unresolvable by V3-V4 but separable with full-length 16S include *Staphylococcus aureus* / *S. epidermidis*, *Streptococcus agalactiae* / *S. mutans*, and most *Enterococcus*, *Lactobacillus*, and *Clostridium* species pairs sharing a common habitat.

2. Dataset and Bioinformatics

2.1 Public Mock-Community Dataset

To evaluate the method’s accuracy and reproducibility, a public benchmark dataset comprising four well-characterised mock communities was analysed [7]:

Table 1. Mock community groups analysed in this study.

Group	Product	Composition	n replicates
ATCC-even	ATCC MSA-1002	20 species, equimolar genomic DNA	8
ATCC-staggered	ATCC MSA-1003	Same 20 species, 4-fold dilution series	8
Zymo-Gut	Zymo D6331	21 defined gut-representative strains	8
Zymo-Fecal	Zymo D6323	Pooled human fecal DNA (complex, undefined)	8

The 32 samples were processed through the PacBio HiFi-16S Nextflow pipeline [8] with the GreenGenes 2 (2024-09) taxonomy database [5,6]. To assess the effect of sequencing depth on taxonomic recovery, four subsampling levels were tested: 2,000, 20,000, 40,000, and 100,000 S-reads per sample (hereafter 2k, 20k, 40k, and 100k).

2.2 Bioinformatics Pipeline

DADA2 was used for denoising and ASV generation within the HiFi-16S workflow [9]. Taxonomy was assigned by VSEARCH against a reference database; the pipeline supports three options:

- **GG2** (default): `gg2_2024_09_trainset` — GreenGenes2 2024-09 build
- **SILVA**: `silva_nr99_v138.2` — SILVA non-redundant 99% database v138.2
- **GTDB**: `gtdb_bac120_r220` — GTDB bacterial release r220

All results presented here use the GG2 2024-09 database. All depth comparisons are based on the per-depth merged frequency/taxonomy tables (`best_tax_merged_freq_tax.tsv`). Relative abundances were computed by normalising raw read counts to 100% per sample after averaging replicates within each group.

3. Results

3.1 Composition Profiles and Effect of Sequencing Depth

Figure 3 shows the mean relative abundance composition for all four mock community types across four sequencing depths (2k, 20k, 40k, 100k), averaged over eight replicates per group. Species representing less than 1% of every bar are pooled into a single “merged taxa” category.

Generated by `scripts/plot_freq_vs_depth.R` (abundance cutoff: 1%; see Code Availability).

The two ATCC mock communities display highly stable composition profiles across all four depths: the dominant species are correctly called even at 2k reads, and the relative ordering of taxa does not change with depth. This reflects the simple, defined nature of these mocks (20 species covering virtually all reads) and confirms that the DADA2 denoising step resolves the community reliably regardless of input depth for defined communities.

The Zymo-Gut mock (21 defined strains) also shows consistent composition from 2k to 100k, though some low-abundance strains are inconsistently detected at 2k. The 40k and 100k profiles are essentially identical for both Zymo-Gut and the ATCC mocks, confirming that 20k is already sufficient for defined low-complexity communities.

The Zymo-Fecal mock (real pooled human feces) shows a qualitatively different pattern: while the dominant taxa are stable across depths, the number of detectable low-abundance species increases substantially with depth (see §3.4 and Figure 4), making sequencing depth a critical variable for complex community surveys.



Figure 3: 100%-normalised stacked barplots showing mean taxon abundance (averaged across 8 replicates) for each mock community group at 2k, 20k, 40k, and 100k S-reads. Species below 1% in all bars are pooled as “merged taxa” (white/outlined).

3.2 Species-Level Accuracy in the ATCC Even Mock

The ATCC MSA-1002 mock provides a defined ground truth: 20 bacterial species at equimolar genomic DNA concentrations. At 100k reads, all 20 expected species were recovered above the 1% relative abundance threshold. At 2k reads, 19/20 species were recovered; the missing taxon is one of the lowest 16S-represented species in the mock (borderline in read count after denoising at this depth).

Observed proportions deviate from the theoretical 5% per species because genomic DNA equimolarity does not translate directly to 16S read equimolarity: each species carries a different number of 16S gene copies per genome (ranging from 1 to >10 across bacteria), so observed relative abundances reflect both input DNA and 16S copy number simultaneously. This is a fundamental property of all 16S amplicon surveys, not a pipeline artefact [10].

Table 2. ATCC MSA-1002 species — observed relative abundance and inter-replicate CV at 100k reads.

Species	Observed mean (%)	CV (%)
<i>Clostridium</i> sp. (beijerinckii clade)	9.0	4.7
<i>Streptococcus agalactiae</i>	8.1	5.0
<i>Lactobacillus gasseri</i>	7.3	4.6
<i>Streptococcus mutans</i>	6.7	6.8
<i>Phocaeicola vulgatus</i>	6.5	7.0
<i>Neisseria meningitidis</i>	6.4	7.7
<i>Staphylococcus epidermidis</i>	6.2	4.7
<i>Deinococcus radiodurans</i>	6.1	5.0
<i>Acinetobacter baumannii</i>	6.0	17.9
<i>Cereibacter sphaeroides</i>	5.2	5.3
<i>Staphylococcus aureus</i>	4.8	4.5
<i>Porphyromonas gingivalis</i>	4.6	4.7
<i>Pseudomonas aeruginosa</i>	4.0	18.1
<i>Helicobacter pylori</i>	4.0	6.1
<i>Enterococcus faecalis</i>	3.3	5.9
<i>Cutibacterium acnes</i>	2.7	5.1
<i>Escherichia coli</i> (complex)	2.0	23.4
<i>Bacillus</i> sp.	2.0	6.2
<i>Bifidobacterium adolescentis</i>	1.3	14.6
Unclassified ASV	1.0	11.6

Observed mean: arithmetic mean of relative abundance (%) computed from per-replicate VSEARCH frequency tables across 8 replicates. CV (coefficient of variation): $SD / mean \times 100$, expressing inter-replicate variability as a percentage of the mean — lower values indicate higher reproducibility.

Several observations are noteworthy:

- **16S gene copy number bias.** Species with a low copy number (*Acinetobacter*, *Pseudomonas*, *Deinococcus*) appear below the expected 5% and show higher inter-replicate CV, consistent with stochastic effects at low read counts. *Bifidobacterium adolescentis*, a known single-copy species, shows the most pronounced under-representation. This is a fundamental property of all 16S amplicon surveys, not a pipeline artefact [10].
- **Sub-species resolution.** The *E. coli* entry is split by GreenGenes 2 across three distinct sequence variants (*E. coli*, *E. albertii*, *E. boydii*), reflecting strain-level granularity encoded in the V1 and V6–V9 hypervariable regions. Their combined abundance sums to the expected range.
- **Congeneric species separation.** Both *Streptococcus* species (*S. agalactiae* and *S. mutans*) and both *Staphylococcus* species (*S. aureus* and *S. epidermidis*) are unam-

biguously separated in every replicate — a separation routinely lost in V3-V4 data for these genus pairs [4].

3.3 Reproducibility Across Eight Replicates

Inter-replicate reproducibility is high across all four mock types:

- **ATCC-even at 100k:** median CV of **6%** across the 20 dominant species; 17/20 species show CV < 10% (Table 2).
- **High-CV outliers** (*Acinetobacter*, *Pseudomonas*, *E. coli* complex) are precisely the taxa with lowest 16S read counts due to low gene copy numbers. Stochastic sampling variance is inherently higher at low counts — this reflects a biological property, not a platform limitation.
- **Across all four mock types:** dominant taxa (>2% mean) display tight within-group clustering at all depths. Variability increases for rare community members, a natural consequence of Poisson sampling at low counts.

3.4 Effect of Sequencing Depth on Rare Taxon Sensitivity and Diversity

The Zymo-Fecal mock (D6323, pooled human feces) is a complex, continuously distributed community with no fixed theoretical composition, directly mirroring real clinical or environmental microbiome samples. Two complementary analyses quantify the effect of sequencing depth: species-level detection at two abundance thresholds, and per-sample alpha diversity metrics.

Rare taxon sensitivity Figure 4 lowers the display cutoff to 0.1% to reveal the rare taxon tier that is hidden at the 1% threshold used in Figure 3.

Generated by scripts/plot_freq_vs_depth.R (abundance cutoff: 0.1%; see Code Availability).

Table 3. Species-level detection in Zymo-Fecal (D6323) at each sequencing depth.

Depth	Species >1%	Species 0.1–1%	Total detectable
2k	24	29	53
20k	26	82	108
40k	26	83	109
100k	24	89	113

Species counts are the number of unique taxa detected above each relative-abundance threshold (>1% or 0.1–1% of per-sample total reads) in the merged frequency table; “Total detectable” is their sum. Values reflect the union across all 8 Zymo-Fecal replicates at each depth.

Key observations from the species-level data:

- **Dominant community (>1%)** is stable at 24–26 species across all depths. The backbone community structure is captured even at 2k reads.
- **Rare tier (0.1–1%)** triples from 29 species at 2k to 82 at 20k — the single biggest gain. From 40k to 100k the increment is marginal (83→89 species), indicating near-saturation.
- **Biologically relevant rare taxa** include slow-growing anaerobes, health-associated butyrate producers (*Faecalibacterium*, *Roseburia*), and low-prevalence pathogens that are entirely invisible at 2k depth.

Alpha diversity by sequencing depth Beyond taxon detection thresholds, per-sample alpha diversity metrics capture the full signal, including taxa below 0.1%. **Figures 5 and 6** show observed ASV richness and Shannon diversity (H') across all four groups and depths.

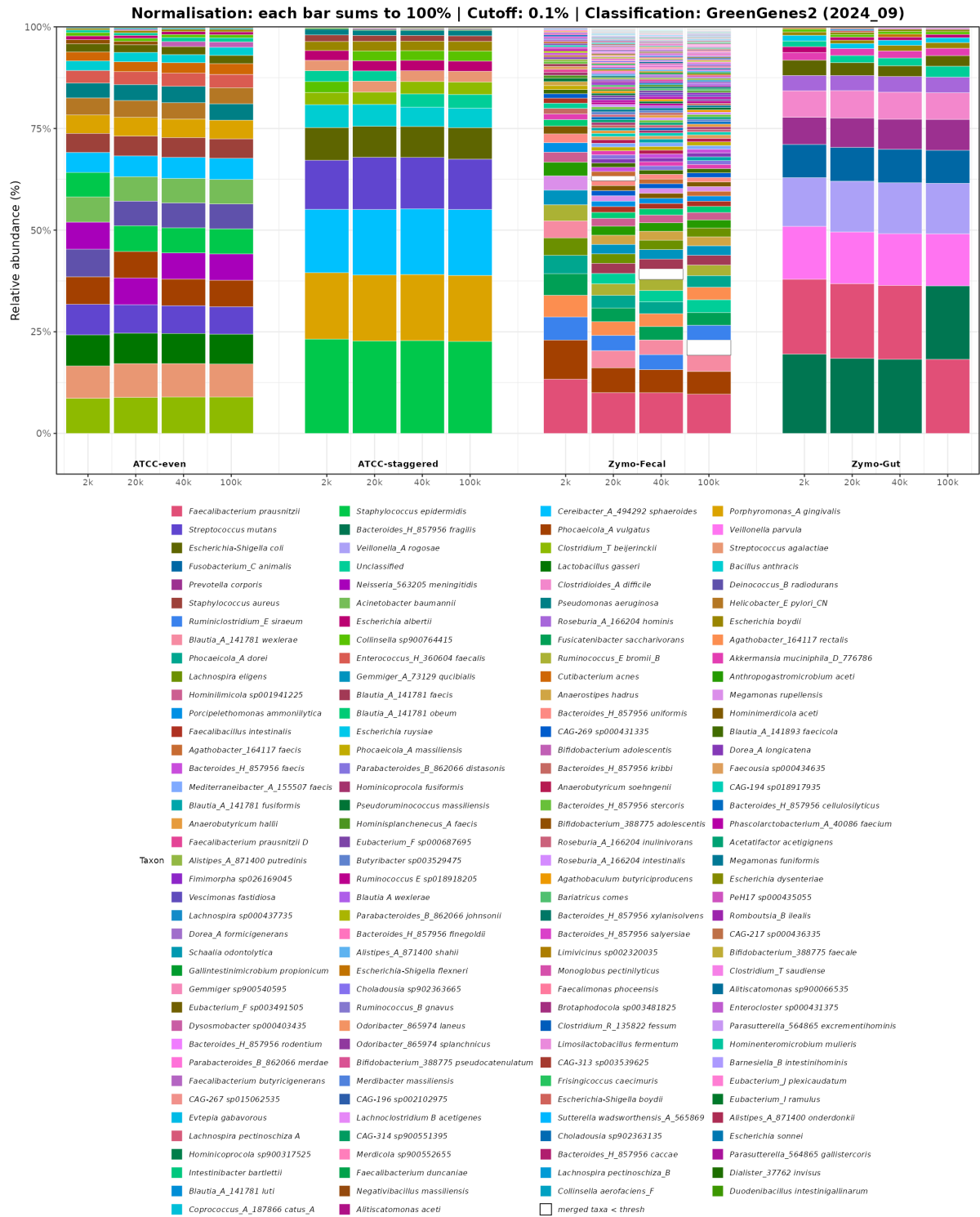


Figure 4: Same data as Figure 3 with the taxon cutoff lowered to 0.1%, revealing the rare taxon tier. The large “merged taxa” fraction (white/outlined) at 2k reflects species not yet detected — not species truly absent. The Zymo-Fecal panel shows the most pronounced depth-sensitivity, with a dramatic reduction in “merged taxa” from 2k to 20k.

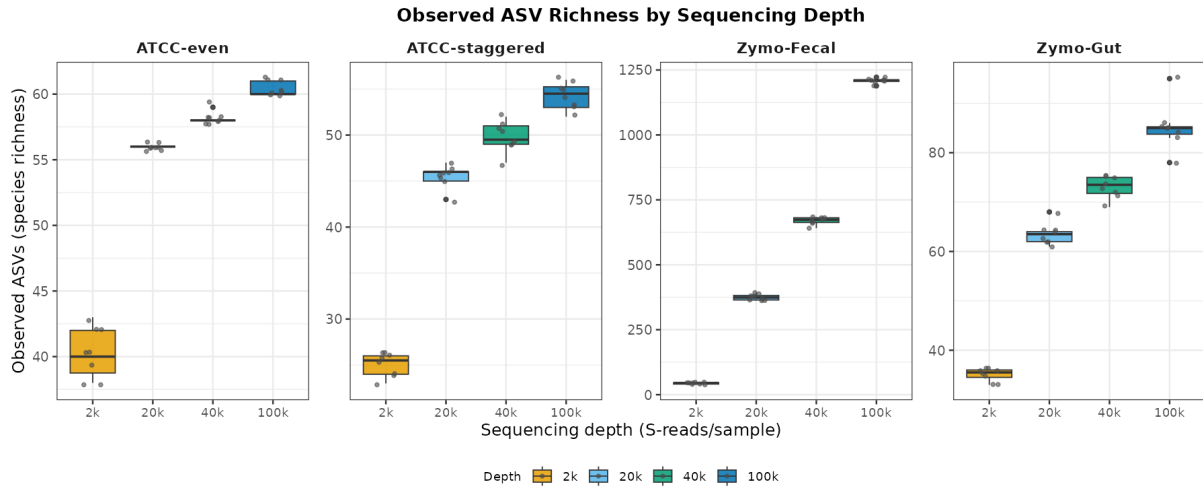


Figure 5: Observed ASV richness (number of unique ASVs per sample) as a function of sequencing depth for each mock group. Boxplots show the distribution across 8 replicates; individual points are overlaid. Note the divergent behaviour between defined communities (ATCC, Zymo-Gut) and the complex Zymo-Fecal community.

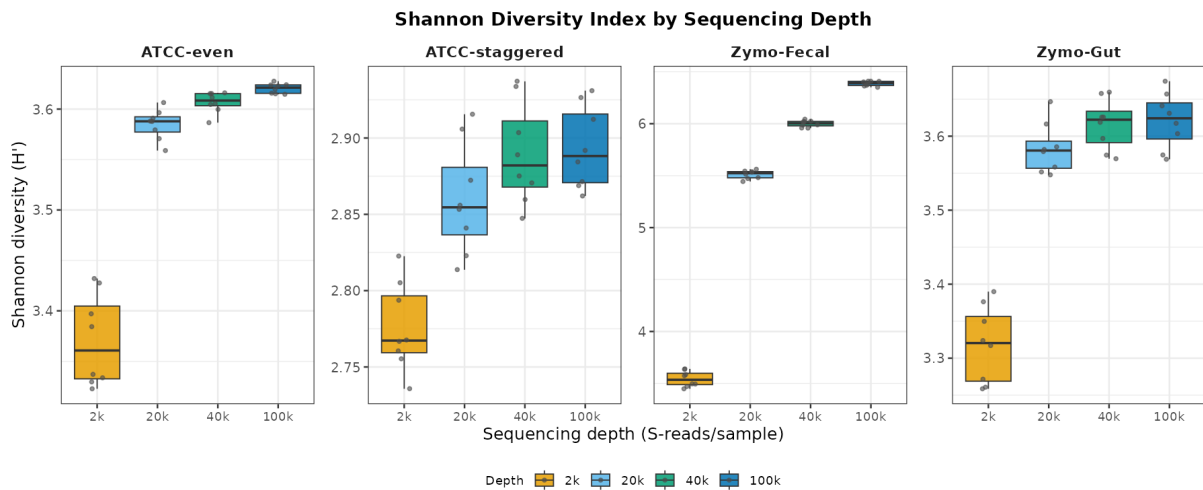


Figure 6: Shannon diversity index (H') per sample across depths. Shannon is weighted by abundance and therefore less sensitive to rare ASV inflation than observed richness. The convergence of the curves for defined communities above 20k contrasts sharply with the continued rise in Zymo-Fecal.

Figures 5 and 6 generated by `scripts/plot_alpha_diversity.R` (see Code Availability).

Table 4. Alpha diversity by group and depth (mean $\pm$ SD across 8 replicates).

Group	Depth	Observed ASVs	Shannon H'
ATCC-even	2k	40.2 $\pm$ 1.9	3.37 $\pm$ 0.05
ATCC-even	20k	56.0 $\pm$ 0.0	3.58 $\pm$ 0.02
ATCC-even	40k	58.1 $\pm$ 0.4	3.61 $\pm$ 0.01
ATCC-even	100k	60.4 $\pm$ 0.5	3.62 $\pm$ 0.01
ATCC-staggered	2k	25.0 $\pm$ 1.2	2.78 $\pm$ 0.03
ATCC-staggered	20k	45.5 $\pm$ 1.2	2.86 $\pm$ 0.04
ATCC-staggered	40k	49.8 $\pm$ 1.6	2.89 $\pm$ 0.03
ATCC-staggered	100k	54.2 $\pm$ 1.5	2.89 $\pm$ 0.03
Zymo-Gut	2k	35.0 $\pm$ 1.3	3.32 $\pm$ 0.05
Zymo-Gut	20k	63.5 $\pm$ 2.1	3.58 $\pm$ 0.03
Zymo-Gut	40k	73.0 $\pm$ 2.2	3.62 $\pm$ 0.03
Zymo-Gut	100k	85.1 $\pm$ 4.7	3.62 $\pm$ 0.04
Zymo-Fecal	2k	43.8 $\pm$ 3.6	3.54 $\pm$ 0.08
Zymo-Fecal	20k	375 $\pm$ 12	5.51 $\pm$ 0.04
Zymo-Fecal	40k	670 $\pm$ 14	6.00 $\pm$ 0.03
Zymo-Fecal	100k	1208 $\pm$ 10	6.39 $\pm$ 0.02

Observed ASVs: count of unique amplicon sequence variants with read count > 0 in a given sample; values are mean $\pm$ SD across 8 replicates. Shannon H': Shannon entropy index $H' = -\sum p_i \ln(p_i)$, where p_i is the relative abundance of ASV i ; measures both richness and evenness, with higher values indicating greater diversity. Both metrics were computed from pipeline output by the analysis scripts accompanying this report.

Key observations from the alpha diversity data:

- **Defined communities (ATCC-even, ATCC-staggered, Zymo-Gut):** Shannon diversity reaches a plateau by 20k (H' changes by < 0.05 between 20k and 100k for all three groups). Observed ASV richness continues to rise modestly with depth, driven by technical sub-species ASV splitting rather than discovery of genuinely new community members.
- **Complex community (Zymo-Fecal):** Observed richness rises 27-fold from 2k (44 ASVs) to 100k (1,208 ASVs), and Shannon H' increases from 3.54 to 6.39 — nearly doubling — with no sign of saturation. Each additional depth increment reveals genuine biological signal.
- **Between-group divergence:** At 2k, all four groups show similar observed richness (25–44 ASVs). By 100k, Zymo-Fecal is 14–20 $\times$ richer than any defined mock, underscoring the qualitatively different nature of complex community surveys.
- **Practical implication:** Shannon diversity is the more robust metric for community comparisons across depths because it is weighted by abundance and is therefore less inflated by rare ASV splitting. For gut microbiome surveys targeting diversity differences between groups, 20–40k reads provides stable Shannon estimates; higher depth is necessary when surveying richness or rare community members.

4. Discussion

4.1 Species-Level Resolution: the Core Advantage of Full-Length 16S

The analysis of four mock communities confirms that the Kinnex 16S workflow achieves reliable species-level resolution in a single amplicon workflow:

- **V1-V9 cumulative diversity.** The full-length molecule encodes all nine hypervariable regions simultaneously. No single sub-region amplicon can provide equivalent discriminatory power.

- **Congeneric separation.** Both *Staphylococcus* species (*S. aureus* / *S. epidermidis*) and both *Streptococcus* species (*S. agalactiae* / *S. mutans*) are separated in every replicate — a routine failure point of V3-V4 surveys [4].
- **Strain-level granularity.** GreenGenes 2 resolves the *E. coli* entry into three distinct sequence variants, illustrating V1 and V6-V9 strain discrimination.
- **Clinical and ecological relevance.** For applications where species identity is essential — pathogen discrimination, probiotic tracking, microbiome-disease association — this is not a marginal improvement but a qualitative change in what the data supports.

4.2 Recommended Sequencing Depth

The data from this benchmark support the following depth guidelines:

- **2k reads:** Recovers the dominant community (>1%) for defined mocks. Misses borderline low-abundance species and most of the rare tier. Suitable only for rapid screening or high-level profiling of simple communities.
- **20k reads:** Recommended minimum for defined or low-complexity communities. Shannon diversity is stable from this depth onward (H' change < 0.05 from 20k to 100k for ATCC and Zymo-Gut). Rare-tier detection for complex communities jumps from 29 to 82 species at this depth — the largest single gain.
- **40k reads:** Practical saturation point for species-level detection in the 0.1–1% abundance range of complex gut communities (83 species vs 89 at 100k). Shannon diversity continues to rise in complex communities, making this an intermediate choice when budget constrains depth.
- **100k reads:** Maximum characterisation of complex gut or environmental communities. Observed ASV richness in Zymo-Fecal reaches 1,208 ASVs (vs 44 at 2k), and Shannon $H' = 6.39$ (vs 3.54). No saturation is observed, meaning further depth would continue to reveal community members. This depth is recommended for discovery-oriented surveys, metagenome-comparable profiling, and rare-event detection.

4.3 Revo Throughput and Per-Sample Economics

At the highest Kinnex multiplexing configuration (up to 1,536 samples per SMRT Cell using 4-plex Kinnex adapters combined with barcoded primers), the per-sample sequencing cost on Revo is directly comparable to a standard V3-V4 short-read run. The Kinnex approach therefore removes the traditional cost barrier between short-read and full-length 16S, making species-level microbiome profiling accessible at population scale without financial trade-off.

4.4 Limitations and Considerations

16S amplicon surveys in general — full-length or partial — carry several inherent limitations that apply equally to Kinnex 16S:

- **Primer amplification bias.** Universal 16S primers amplify some taxa more efficiently than others, introducing systematic biases that are independent of sequencing technology.
- **Database-dependent taxonomy.** Classification accuracy depends on the completeness of the reference database (GreenGenes 2 here). Novel or poorly-represented taxa may receive imprecise or genus-level assignments.
- **16S gene copy number.** Variable copy numbers per genome (ranging from 1 to >10 across bacteria) mean that observed relative abundances reflect both input DNA and copy number simultaneously. The ATCC-even mock illustrates this clearly: equimolar genomic input does not produce equimolar 16S read output. This is a fundamental property of all 16S amplicon surveys, not a pipeline artefact [10].
- **Absolute abundance.** Users planning absolute-abundance interpretations should consider spike-in controls or paired metagenomic validation. Within-sample relative comparisons and community-level statistics (alpha/beta diversity, differential abundance) are robust to copy-number bias and represent the primary use case for Kinnex

5. Summary

Table 5. Summary of key benchmark metrics.

Metric	Value
Mock communities analysed	4 types × 8 replicates
Sequencing depths tested	2k, 20k, 40k, 100k reads/sample
ATCC-even species recovered at 100k	20/20
ATCC-even species recovered at 2k	19/20
Inter-replicate CV (dominant species, 100k)	median 6%, max 23%
Zymo-Fecal rare species at 2k vs 100k	29 vs 89 (0.1–1% tier)
Recommended depth (complex gut)	40–100k S-reads/sample

Full-length 16S on PacBio Revio with Kinnex concatenation delivers species-level taxonomic resolution, high inter-replicate reproducibility, and broad rare-taxon sensitivity in complex communities — at per-sample costs now on par with short-read V3-V4 approaches.

References

1. PacBio. *Preparing Kinnex libraries from 16S rRNA amplicons: Procedure & checklist* (103-238-800 Rev06, Jun 2025).
2. PacBio. *Kinnex 16S rRNA kit for full-length 16S sequencing: Application note* (102-326-601 Rev01, Dec 2023).
3. Klindworth A, et al. *Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies*. *Nucleic Acids Res.* 2013;41(1):e1. doi:10.1093/nar/gks808.
4. Johnson JS, et al. *Evaluation of 16S rRNA gene sequencing for species and strain-level microbiome analysis*. *Nat Commun.* 2019;10:5029. doi:10.1038/s41467-019-13036-1.
5. ATCC. *MSA-1002 and MSA-1003 Microbiome Standards Technical Datasheet*. Manassas, VA: ATCC; 2022.
6. McDonald D, Jiang Y, Balaban M, et al. *Greengenes2 unifies microbial data in a single reference tree*. *Nat Biotechnol.* 2024;42:715–718. doi:10.1038/s41587-023-01845-1. Database build 2024-09 available at: https://zenodo.org/records/14169078/files/gg2_2024_09_toSpecies_trainset.fa.gz
7. PacBio. *Kinnex 16S public dataset*. Available at: <https://downloads.paccloud.com/public/dataset/Kinnex-16S/> (accessed March 2026).
8. PacBio. *HiFi-16S-workflow: A SMRT bioinformatics workflow for full-length 16S rRNA amplicon analysis*. GitHub. Available at: <https://github.com/PacificBiosciences/HiFi-16S-workflow> (accessed March 2026).
9. Callahan BJ, et al. *DADA2: High-resolution sample inference from Illumina amplicon data*. *Nat Methods.* 2016;13:581–583. doi:10.1038/nmeth.3869.
10. Kembel SW, et al. *Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance*. *PLoS Comput Biol.* 2012;8(10):e1002743. doi:10.1371/journal.pcbi.1002743.

Code Availability

All R scripts used to generate the figures and derived metrics in this report are provided as supplementary material alongside the document. They can be run from the project root

using a standard R installation with the `ggplot2`, `dplyr`, and `tidyr` packages.

- **`scripts/plot_freq_vs_depth.R`** — 100%-normalised stacked barplots of taxon composition across sequencing depths, with configurable abundance cutoff. Input: per-depth `best_tax_merged_freq_tax.tsv`. Output: Figures 3 & 4.
- **`scripts/plot_alpha_diversity.R`** — Per-sample observed ASV richness and Shannon H' across depths; group summary statistics. Input: per-depth `best_tax_merged_freq_tax.tsv`. Output: Figures 5 & 6; Table 4.

Appendix — Glossary of Technical Terms

16S rRNA gene A gene present in all bacteria and archaea that encodes the RNA component of the small ribosomal subunit. Because it contains both conserved regions (useful for broad detection) and variable regions (useful for identification), it is the standard marker gene for bacterial community profiling.

Hypervariable regions (V1-V9) Nine short stretches within the 16S rRNA gene where the DNA sequence varies greatly between species. The more variable regions are included, the better one can distinguish closely related organisms. Short-read methods typically target only V3-V4; the Kinnex approach captures all nine simultaneously.

Amplicon A DNA fragment produced by PCR amplification of a specific target region. In 16S profiling, the amplicon is the ~1,500 bp copy of the full-length 16S gene generated from each bacterium in the sample.

PCR (Polymerase Chain Reaction) A laboratory technique that copies a specific DNA region millions of times, making it detectable by sequencing. Barcoded primers (short DNA sequences flanking the target) can be included so that reads from different samples can be identified and separated after sequencing.

SMRTbell library The sequencing-ready DNA construct used on PacBio instruments. It consists of the DNA insert flanked by hairpin adapters that allow the sequencing polymerase to read the same molecule multiple times, improving accuracy.

MAS-Seq / Kinnex concatenation A strategy that ligates multiple short amplicons end-to-end into a single long molecule before sequencing. This increases the number of amplicons read per sequencing pass (8–12-fold), dramatically improving cost efficiency. The Kinnex kit is the commercial implementation of MAS-Seq for 16S.

HiFi reads / CCS (Circular Consensus Sequencing) PacBio sequencing reads generated by passing the sequencing polymerase around a circular SMRTbell molecule multiple times. Each pass corrects errors from the previous one; the consensus of all passes yields a HiFi read with >99.9% accuracy despite the long read length.

SMRT Cell The consumable flow cell loaded into a PacBio Revio instrument. Each SMRT Cell contains millions of microscopic wells (zero-mode waveguides) where individual sequencing reactions take place simultaneously.

Revio PacBio's current high-throughput long-read sequencing instrument. It delivers approximately 50–60 million deconcatenated 16S reads per SMRT Cell when used with the Kinnex workflow.

S-reads (segmented reads) After sequencing, Kinnex concatenated molecules are computationally split back into their individual amplicon components. These individual segments are called S-reads; they represent single 16S amplicons derived from a single bacterium in the sample.

Combinatorial barcoding A multiplexing strategy where each sample receives a unique combination of short DNA tags (barcodes) attached to the PCR primers. After sequencing, reads are sorted back to their sample of origin by reading the barcode. The Kinnex system supports up to 1,536 samples per SMRT Cell using this approach.

DADA2 A bioinformatics algorithm that models sequencing errors statistically and corrects them to produce exact amplicon sequence variants (ASVs). It distinguishes true biological sequences from sequencing artefacts without collapsing similar sequences into arbitrary clusters.

ASV (Amplicon Sequence Variant) A unique, exact DNA sequence recovered from a sequencing run after error correction. Unlike the older OTU approach (which groups sequences at 97% similarity), ASVs represent single-nucleotide resolution variants and provide finer taxonomic discrimination.

OTU (Operational Taxonomic Unit) An older way of grouping 16S sequences by clustering those that are $\geq 97\%$ similar together. Less precise than ASVs; included here for context when comparing Kinnex to legacy short-read workflows.

VSEARCH An open-source tool used to compare ASV sequences against a reference database and assign taxonomy. It finds the closest matching reference sequence and reports the corresponding species name.

GreenGenes 2 (GG2) A curated reference database of 16S rRNA sequences linked to a consistent bacterial taxonomy. The 2024-09 build is used here for assigning species names to ASVs.

SILVA / GTDB Alternative reference databases for 16S taxonomy. SILVA is a long-established general-purpose database; GTDB (Genome Taxonomy Database) uses whole-genome phylogeny to define species boundaries. All three databases are supported by the HiFi-16S pipeline.

Mock community A laboratory standard consisting of a defined set of microorganisms mixed at known proportions. Running a mock community through the same workflow as real samples allows objective measurement of accuracy, reproducibility, and bias.

Relative abundance The proportion of reads assigned to a given taxon expressed as a percentage of the total reads in a sample. Because sequencing measures proportions rather than absolute cell counts, all 16S results are inherently relative.

16S gene copy number The number of copies of the 16S gene per bacterial genome, which varies from 1 to more than 10 across species. Because PCR amplifies every copy, species with more gene copies generate disproportionately more reads, causing a systematic bias in observed relative abundances. This is a fundamental property of all 16S methods, not a technical error.

CV (Coefficient of Variation) A measure of variability expressed as the standard deviation divided by the mean, multiplied by 100 to give a percentage. A low CV means that repeated measurements (replicates) give very similar results; a high CV indicates greater variability. Used here to quantify inter-replicate reproducibility.

Alpha diversity A measure of species diversity within a single sample. Two common metrics are used in this report: observed ASV richness (the count of distinct ASVs) and Shannon H' (which accounts for both the number of species and their relative abundances).

Shannon diversity index (H') A widely used alpha diversity metric calculated as $H' = -\sum p_i \ln(p_i)$, where p_i is the relative abundance of each ASV. Higher values indicate a community that is both species-rich and evenly distributed. Because it is weighted by abundance, it is less sensitive to rare species than simple richness counts.

Subsampling (rarefaction) Randomly drawing a fixed number of reads from each sample to make sequencing depth comparable across samples. Used here to test how community profiles change as depth increases from 2,000 to 100,000 S-reads per sample.

Nextflow A workflow management framework that chains bioinformatics tools into reproducible, scalable pipelines. The PacBio HiFi-16S workflow is implemented in Nextflow.

Acknowledgements

This document was written by Stéphane Plaisance (Nucleomics Core) with the assistance of Claude (Anthropic), an AI language model, for drafting, editing, and literature review.

Document version: 1.0 Last updated: march 16, 2026