Metagenomics Core Control Set: Insights

Evaluating Compatibility of the Metagenomics Core Control Set with PacBio Long-Read Sequencing

Summary

Standardization remains a significant challenge in metagenomic sequencing, particularly as new platforms emerge with distinct chemistries, introducing additional variability in sequencing metrics. Synthetic spike-in controls provide a framework by anchoring sequencing experiments to known internal benchmarks. The Sequins Metagenomics Core Control Set was developed as a platform-agnostic ladder of synthetic DNA molecules designed to support quantitative benchmarking in shotgun metagenomics. This study evaluated the compatibility of the Control Set with a Pacific Biosciences (PacBio) long-read sequencing workflow, specifically using Ampli-Fi on the Revio platform, and assessed the impact of size selection on control performance. As expected, PacBio size selection enriched for longer molecules, providing a useful test of whether the retained longer Sequins controls could preserve quantitative performance. Importantly, with length-aware analysis, ladder linearity and run-specific limit of detection assessment were maintained. Together, these findings reinforce the reliability of the Control Set as a built-in quantitative reference for PacBio workflows, while supporting its broader utility in long-read metagenomic sequencing.

Background

Next-generation sequencing platforms differ substantially in sequencing chemistry, read length, throughput, and error profiles. While short-read technologies remain widely used for high-throughput metagenomics, long-read sequencing platforms, such as PacBio, offer distinct advantages for microbial applications. These include improved genome assembly, resolution of repetitive and structurally complex regions, more accurate reconstruction of microbial genomes, and enhanced characterization of mixed and diverse communities. As long-read technologies mature and are increasingly adopted in metagenomics, laboratories are integrating them alongside or in place of, existing short-read workflows.

The adoption of long-read sequencing introduces new sources of variability, including changes in library preparation, size selection, molecule length distributions, and platform-specific biases. In addition, laboratories may transition between long-read platforms, sequencing chemistries, or software versions over time. Maintaining quantitative consistency and analytical confidence across these transitions requires a robust internal control framework that can track platform- and workflow-specific effects while preserving interpretable quantitative benchmarks.

The Sequins Control Set was designed to provide such a framework. The Control Set comprises a defined ladder of synthetic DNA molecules spanning multiple abundance levels, enabling quantitative benchmarking of sequencing performance within each run. Because Sequins are synthetic and mirrored, they can be spiked into any metagenomic sample and distinguished from native microbial DNA during analysis, independent of sample

composition or microbial species present in the sample. This makes the Control Set inherently species-agnostic and suitable for use across diverse microbial communities, hosts, and environments.

Importantly, the Control Set supports routine quantitative applications beyond platform evaluation. The abundance ladder enables direct assessment of linearity, supports run-specific limit of detection (LoD) determination, and provides a stable reference for normalization and quality control across samples, runs, and laboratories. These capabilities are particularly valuable in long-read metagenomics, where changes in molecule length distributions or size-selection strategies can impact quantitative performance if not appropriately controlled.

Introduction to Sequins

Sequins (sequencing spike-ins) are synthetic nucleic acid controls that directly mirror naturally occurring sequences. Because Sequins retain the same nucleotide composition as the natural sequence, they enable accurate representation of genomic complexity without compromising the integrity of the sample or the results. Sequins perform equivalently throughout sequencing workflows, providing a true measure of control.

Sequins' innovative design enables production of synthetic mirrored sequences that directly represent almost any genomic feature, in any organism with a reference genome. Common and ecologically significant features, such as microbial markers or virulence-associated genes and features from analytically challenging regions of metagenomes, can be represented. By combining Sequins in precise stoichiometric ratios, quantification of low-abundance species can also be assessed. Sequins are simply spiked-in to a sample prior to library preparation and progressed together through a workflow (Figure 1). Sequins controls can then be distinguished from native microbial DNA in subsequent data by virtue of their synthetic sequence. This enables normalization and comparison between samples, runs, laboratories, chemistries, and sequencers.

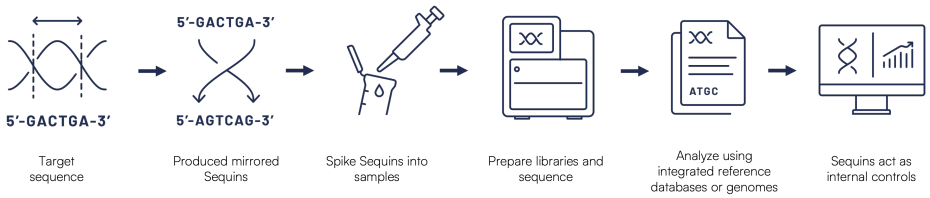

Figure 1. Schematic showing the design and use of Sequins in an NGS workflow.

Methods

Metagenomics Core Control Set Content

The Metagenomics Core Control Set is a synthetic spike-in standard comprising 65 Sequins molecules representing more than 50 microbial species across bacteria, archaea and eukaryotes. Sequins are distributed across 13 defined abundance levels, spanning 1-fold (lowest abundance) to 32,768-fold (highest abundance), with five Sequins per level to form a quantitative ladder, with GC content ranging from 27-71% and fragment sizes from 982 bp to 6,228 bp.

Study Design

Naturally occurring microbial community profiles were generated from three dog fecal samples. DNA was extracted using the ZymoBIOMICS DNA Miniprep kit (Zymo Research) according to the manufacturer's protocol. Extracted DNA from each sample was divided into three experimental groups: 1% (mass/mass spike-in relative to sample DNA input) Control Set spike-in[1] without size selection, 1% Control Set spike-in[1] with size selection, and 10% Control Set spike-in[1] with size selection. As such, a total of nine canine metagenomics samples were analyzed.

Sequencing and Analysis

Libraries were prepared for each of the nine experimental samples and distributed across two SMRT Cells to obtain sufficient reads. Libraries were prepared using the standard PacBio Ampli-Fi protocol (103-648-000) for size-selection conditions, with a protocol modification for the non-size-selected condition. Sequencing was performed on the PacBio Revio system. PacBio reads were processed with fastplong and aligned to the Control Set decoy chromosome reference (provided with the product) using minimap2, followed by abundance quantification with CoverM.

Results

Sequencing Metrics

The inclusion of the Control Set did not adversely affect total read output at any of the tested conditions: 1% spike-in without size-selection, 1% spike-in with size-selection or 10% with size-selection (Figure 2A). As expected, size-selected libraries showed modest enrichment of longer reads (Figure 2B). Mapping rates to Sequins closely reflected the proportion of spike-in added (Figure 2C), as expected.

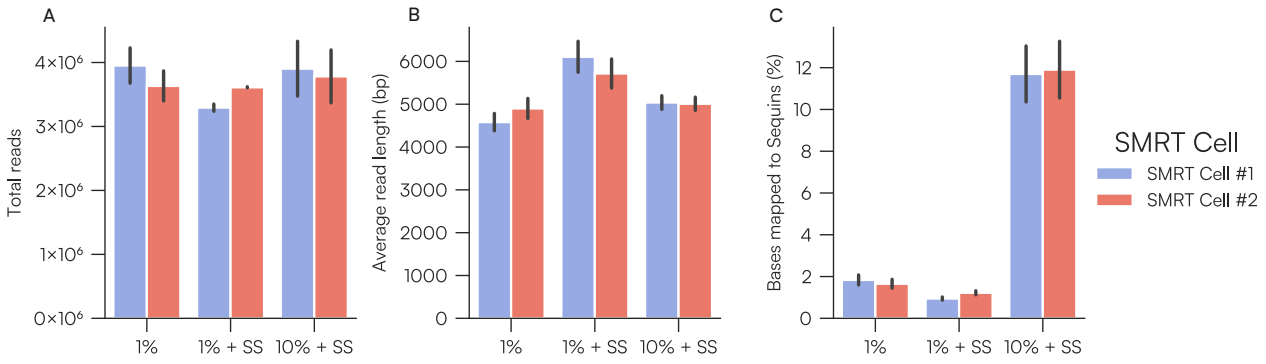

Figure 2: Summary of sequencing metrics. (A) Total read output was comparable across libraries prepared with 1% spike-in without size-selection, 1% spike-in with size-selection, and 10% spike-in with size-selection, indicating consistent sequencing yield across conditions. **(B)** Average read length increased modestly in size-selected libraries relative to non-size-selected libraries, consistent with enrichment for longer reads. **(C)** The proportion of total bases mapped to Sequins increased with spike-in input, supporting input-dependent recovery of the Control Set. *Size-selection (SS).*

[1] Spike-in as a percentage of total DNA input

Ladder Linearity

Analysis of Sequins Control Set data without size-selection demonstrated an R^2 of 0.837 at a 1% spike-in, with most Sequins control targets detected (at least 58 out of 65) (Figure 3). When size-selection was applied at the same spike-in level, ladder linearity decreased (R^2 = 0.610), and up to 18 Sequins controls were undetected. Increasing the Sequins spike-in to 10% with size-selection resulted in only partial improvement in recovery, although up to nine Sequins controls remained undetected and overall linearity remained reduced (R^2 = 0.600). These results indicate that size selection introduced a systematic length-dependent bias that was not fully corrected by increasing spike-in input.

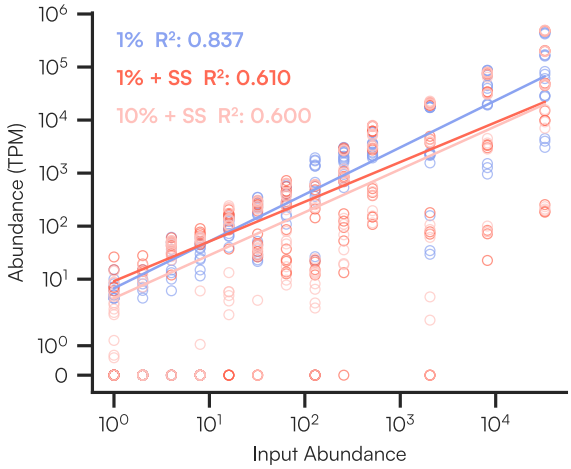

Figure 3: Sequins ladder performance with and without size selection. Observed abundance is plotted against expected input abundance for each condition. Regression analyses indicate higher linearity in the absence of size selection (R^2 = 0.837) compared with size-selected libraries (R^2 = 0.610 and 0.600). *Transcripts per million (TPM), Size-selection (SS). Sequins with an observed abundance of zero are shown for completeness but excluded from regression analyses.*

Optimization by Length Filtering

Excluding Sequins controls shorter than 1,500 bp in size-selected samples resulted in modest improvement in ladder linearity (R^2 values of 0.766 and 0.765 for 1% spike-in with size-selection and 10% with size-selection, respectively; Figure 4A). In contrast, applying a threshold of 2,800 bp markedly improved ladder performance, achieving R^2 values up to 0.961 while retaining 32 Sequins across the 12 abundance levels out of the total number of Sequins (n=65) across the 13 abundance levels included in the Control Set (Figure 4B). Increasing the threshold further to 3,000 bp did not provide additional gains in linearity but substantially reduced ladder resolution with only 11 Sequins left representing 10 abundance levels on the ladder (Figure 4C).

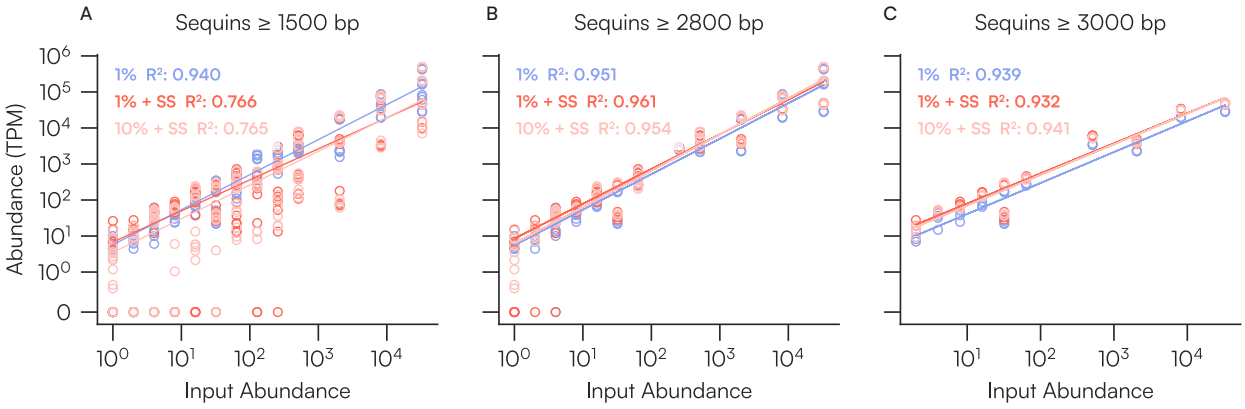

Figure 4: Effect of length threshold optimization on the Control Set ladder performance. (A) Applying ≥1,500 bp length threshold, low R^2 values were observed in size-selected conditions, reflecting reduced quantitative performance as compared to the non-size-selected condition. **(B)** Applying a ≥2,800 bp length threshold improved ladder linearity across all conditions with R^2 of 0.961 and 0.954 for size-selection conditions. **(C)** Increasing the threshold to ≥3,000 did not further improve ladder linearity. *Transcripts per million (TPM), Size-selection (SS). Sequins with an observed abundance of 0 are shown for completeness but excluded from regression analyses.*

Limit of Detection Analyses

Ladder linearity was optimal when a ≥2,800 bp threshold was applied. Limit of detection (LoD) analyses were therefore performed using this length-filtered subset across all three experimental conditions (Table 1). Inclusion of the Control Set in each sample enabled run-specific operational LoD estimation. For the 1% spike-in, the LoD corresponded to 714 gc/µL at the 8-fold abundance in both size-selected and non size-selected libraries. In contrast, for the 10% spike-in with size-selection, the LoD corresponded to 818 gc/µL at the 1-fold abundance level. This reflects the higher absolute input of Sequins in the 10% spike-in condition and should therefore be interpreted as workflow- and input-specific LoD rather than a platform-intrinsic value.

Table 1. Ladder performance and limit of detection (LoD) across spike-in conditions.

Condition	R^2 (before → after threshold optimization*)	LoD abundance level	LoD (gc/µL)
1% spike-in, no size selection	0.837 → 0.951	8-fold	714
1% spike-in, with size selection	0.610 → 0.961	8-fold	714
10% spike-in, with size selection	0.600 → 0.954	1-fold	818

*Threshold optimization was performed by excluding Sequins molecules shorter than 2,800 bp from analyses. This length-aware analysis improved ladder linearity across conditions while preserving run-specific LoD estimation for the retained ladder subset.

Note: LoD was defined as the lowest concentration level on the ladder for which all Sequins controls were detected.

Discussion

This study demonstrates that the Control Set maintains robust quantitative performance when applied to PacBio long-read metagenomic workflows, despite the fundamental differences between long-read and short-read sequencing in molecule length distributions, error profiles, and library preparation strategies.

Across the evaluated conditions, the Control Set retained its core utility as an internal benchmark, including reproducible detection across abundance levels, quantitative ladder behavior, and accurate, run-specific limit of detection estimation. The primary variable influencing ladder performance in PacBio workflows was molecule length, with size selection preferentially reducing representation of shorter Sequins molecules. Importantly, this effect was systematic and readily addressed through length-based filtering during analysis. Applying a ≥2,800 bb threshold restored high ladder linearity in the retained subset while maintaining sufficient ladder resolution to support quantitative benchmarking and LoD estimation.

A key strength of the Control Set is its independence from sample composition and microbial species. As a synthetic, sequence-distinct abundance ladder, the Control Set can be spiked into any metagenomic sample and unambiguously distinguished from native DNA, enabling consistent quantification regardless of host background or community complexity. This species-agnostic design supports its use not only during platform evaluation, but also in routine sequencing operations, where it provides a stable internal reference for inter-run normalization, longitudinal performance monitoring, and workflow quality control.

In the context of long-read metagenomics - where advantages such as improved genome assembly, resolution of complex microbial communities, and enhanced structural insight are balanced by evolving chemistries and analytical pipelines - the availability of a commutable, platform-agnostic control is particularly valuable. The results presented here establish the Control Set as a reliable internal reference for PacBio workflows, enabling laboratories to adopt long-read sequencing while maintaining quantitative continuity across platforms and technologies.

Conclusion

These findings support the use of the Sequins Metagenomics Core Control Set as a foundational control for metagenomic sequencing using PacBio technology. Its compatibility with long-read technology, combined with established performance in short-read workflows (Hardwick et al. 2017; Davis et al. 2025; Langenfeld et al. 2025), positions the Control Set as a species- and platform-agnostic control standard for quantitative benchmarking, LoD assessment, and sequencing workflow quality control in both research and clinical metagenomics.

Implications For Researchers

The availability of a single, validated control set that functions across platforms enables researchers to benchmark performance across sequencing technologies, normalize metagenomic datasets for reproducibility and confidently compare studies regardless of platform choice. The Control Set therefore provides a foundation for rigorous and reproducible metagenomics in both research and clinical settings.

FAQs

How does size selection in PacBio workflows impact Control Set performance?

Size selection preferentially removes shorter Sequins molecules, affecting ladder representation and quantitative linearity. As such, the Control Set performance is impacted but this effect is predictable, systematic and can be mitigated, as outlined in the next question.

Can Control Set still be used with PacBio size-selected libraries?

Yes. Although shorter Sequins molecules may be underrepresented following size selection, quantitative performance can be recovered by applying length-based filtering during analysis. Restricting analysis to Sequins molecules ≥2,800 bp restores high ladder linearity while retaining sufficient ladder resolution for quantitative benchmarking and LoD estimation. For convenience, the Control Set bioinformatics bundle supplied by Sequins, includes a file containing the subset of Sequins IDs corresponding to molecules ≥2,800 bp.

What spike-in percentage should be used?

1% spike-in (mass/mass) is sufficient for routine use and run-specific LoD determination. 10% may improve detection of lower-abundance ladder levels in size-selected libraries but does not fully compensate for length bias.

Does Control Set affect PacBio sequencing output or read quality?

No measurable impact on total read output or sequencing performance has been observed in evaluations undertaken by Sequins.

How is LoD determined?

Sequins defines LoD as the lowest ladder level at which all Sequins molecules are detected within a run. When included in individual samples, the Control Set enables LoD determination at both the sample- and run-level.

Product Information and Ordering

Please visit our website to download the latest version of the Metagenomics Core Control Set Product Sheet. The Core Control Set is available in 24 sample and 48 sample format kits (Table 2).

Table 2. Metagenomics Core Control Set kit information.

Product	Catalog Number	Description
Metagenomics Core Control Set	PN-10008	Metagenomics Core Control Set (24 samples*)
Metagenomics Core Control Set	PN-10009	Metagenomics Core Control Set (48 samples*)

*based on a 1% spike-in; 250 ng DNA input

Bioinformatics Resources

Sequins provides a growing library of practical bioinformatics tutorials on our website to support users in integrating Sequins into their analysis workflows. These tutorials include step-by-step example approaches, curated datasets, and guidance on interpretation. Visit our website to access our tutorials.

Ordering and Support

enquiries@sequins.bio | support@sequins.bio

Key Publications

Davis, B.C., P.J. Vikesland, and A. Pruden, Evaluating Quantitative Metagenomics for Environmental Monitoring of Antibiotic Resistance and Establishing Detection Limits. *Environ Sci Technol*. 59, 6192-6202 (2025)

Hardwick SA, Deveson IW. Mercer TR. Reference standards for next-generation sequencing. *Nat Rev Genet*. 2017 Aug:18(8):473-484 (2017)

Hertramph, T.L., Dorda, M., Pallenberg, S.T. et al. Effects of elexacaftor/tezacaftor/ivacaftor on the nasal microbial metagenome in cystic fibrosis. *Microbiol Spectr* 14, e00601-26 (2026)

Langenfeld, K., Hegarty, B., Vidaurri, S., et al. Development of a quantitative metagenomic approach to establish quantitative limits and its application to viruses. *Nucleic Acids Res.* 53:gkaf118 (2025)