Enzymatic Determinants of Read Length in Ultra-Low Input HiFi Assembly: A Systematic Polymerase Evaluation under the Ampli-Fi Workflow

Tamara Akentjew[1], Javier Gimpel[1], Javier Cáceres[1], Catalina Valdivia[1], Julián Saez[1], Manuel Rozas[1].
[1]Kura Biotech, Puerto Varas, Los Lagos, Chile

Learn more

Introduction

Amplification-based HiFi workflows enable genomic sequencing from nanogram-scale inputs, but the enzymatic amplification step introduces a processivity bias that compresses library fragment distributions toward shorter reads, imposing a read length ceiling that limits assembly contiguity. Furthermore, sequence-dependent GC bias can lead to uneven coverage, making it difficult to achieve representative assemblies from ultra-low (1 ng) inputs. Understanding how different polymerases influence these factors is essential for optimizing ultra-low input protocols.

Objective

To characterize the impact of polymerase on read-length distribution and GC representation, identifying enzymatic parameters that overcome library compression, enabling high-contiguity HiFi assembly from ultra-low (1 ng) DNA inputs.

Methods

Genomic Benchmarking

Escherichia coli and *Staphylococcus epidermidis* were selected as genomic benchmarks to characterize polymerase performance and assembly outcomes under ultra-low input constraints.

Library Preparation & Amplification

Five high-processivity polymerases: DeCodi-Fi™ Long&Complex, KOD Xtreme™ Hot Start DNA Polymerase, Q5® Hot Start High-Fidelity DNA Polymerase, PrimeSTAR® GXL DNA Polymerase, and LongAmp® Taq DNA Polymerase were evaluated using 1 ng of genomic DNA under the Ampli-Fi workflow across three replicates.

Analytical Quality Control

Library yield and fragment size distribution were characterized via agarose gel electrophoresis and Femto Pulse sizing.

Sequencing & Bioinformatics Analysis

Libraries were sequenced on PacBio platforms, and de novo assemblies were generated using “flye”. Key metrics including mean read length, contig N50, genome coverage, and gene completeness (via compleasm) were quantified to evaluate the relationship between enzymatic processivity and assembly contiguity.

References: PacBio. (2025). Ampli-Fi: Amplifying genomic DNA for SMRTbell library preparation and HiFi sequencing (103-648-000 REV04).

Pre-Sequencing Library Quality Control

Pre-sequencing QC was conducted via agarose gel electrophoresis to characterize library yield and size distribution across the evaluated polymerases.

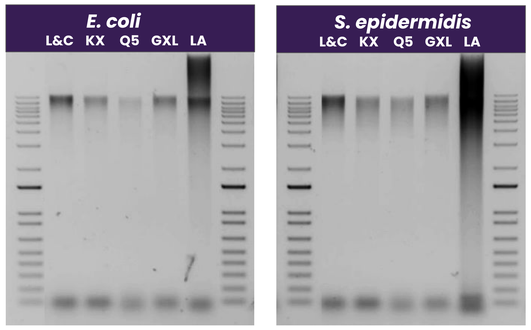

Figure 1: Library amplification QC. Comparison of *E. coli* and *S. epidermidis* library yields using five polymerases: DeCodi-Fi™ Long&Complex (L&C), KOD Xtreme™ Hot Start DNA Polymerase (KX), Q5® Hot Start High-Fidelity DNA Polymerase (Q5), PrimeSTAR® GXL DNA Polymerase (GXL), and LongAmp® Taq DNA Polymerase (LA). Libraries were generated using the PacBio’s Ampli-Fi protocol with 1 ng of DNA input per 50 µL reaction and a standardized 14-cycle amplification program. Samples were analyzed via agarose gel electrophoresis to characterize size distribution and yield prior to PacBio sequencing.

DeCodi-Fi™ Long&Complex produced the highest yield with a distinct distribution shift toward higher molecular weight fragments. KOD Xtreme and PrimeSTAR GXL showed standard library smears, while Q5 resulted in lower yield and smaller size distribution. LongAmp exhibited significant non-specific amplification and was subsequently excluded from sequencing.

Comparative Sequencing and Assembly Metrics

Table 1: Sequencing and assembly metrics for *E. coli* library.

Escherichia coli	>Q30 (%)	Mean length	N50	Gene completeness
KOD Xtreme	97	5.9 kb	3,539 kb	92.47%
DeCodi-F™ Long&Complex	97	6.7 kb	4,585 kb	92.40%
Q5	98	4.3 kb	1,610 kb	92.37%
PRIMESTAR GXL	97	5.4 kb	4,585 kb	92.43%

Table 2: Sequencing and assembly metrics for *S. epidermidis* library.

Staphylococcus epidermidis	>Q30 (%)	Mean length	N50	Gene completeness
KOD Xtreme	97	7.9 kb	1,673 kb	98.40%
DeCodi-Fi™ Long&Complex	97	8.5 kb	2,145 kb	98.37%
Q5	98	5.6 kb	1,716 kb	98.43%
PRIMESTAR GXL	98	6,4 kb	2,142 kb	98.40%

Gene completeness was assessed using BUSCO with the enterobacterales_odb12 dataset for E. coli and the lactobacillales_odb12 dataset for S. epidermidis. To ensure compatibility, E. coli samples were normalized to a uniform depth of 63X, and S. epidermidis samples to a depth of 66X, corresponding to the lowest coverage sample for each genome.

Analysis of *E. coli* and *S. epidermidis* benchmarks revealed a direct correlation between enzymatic processivity and assembly success. DeCodi-Fi™ Long&Complex consistently achieved the highest mean read lengths (6.7 kb and 8.5 kb, respectively), facilitating chromosome-level reconstruction with an N50 of 4,585 kb and 2,145 kb. Notably, while PrimeSTAR GXL matched the N50 (4,585 kb in *E. coli* and 2,142 kb in *S. epdermis),* it did so with shorter mean read lengths than DeCodi-Fi™ Long&Complex , highlighting different efficiency profiles in fragment extension.

In contrast, Q5 was hindered by severe fragment length compression, leading to a significant collapse in assembly contiguity. This resulted in highly fragmented outputs, with the *E. coli* N50 dropping to 1,610 kb. These results establish that enzymatic processivity is the primary driver required to generate the long HiFi reads necessary for resolving complex genomic architectures in ultra-low input workflows.

GC Content vs. Coverage Plot

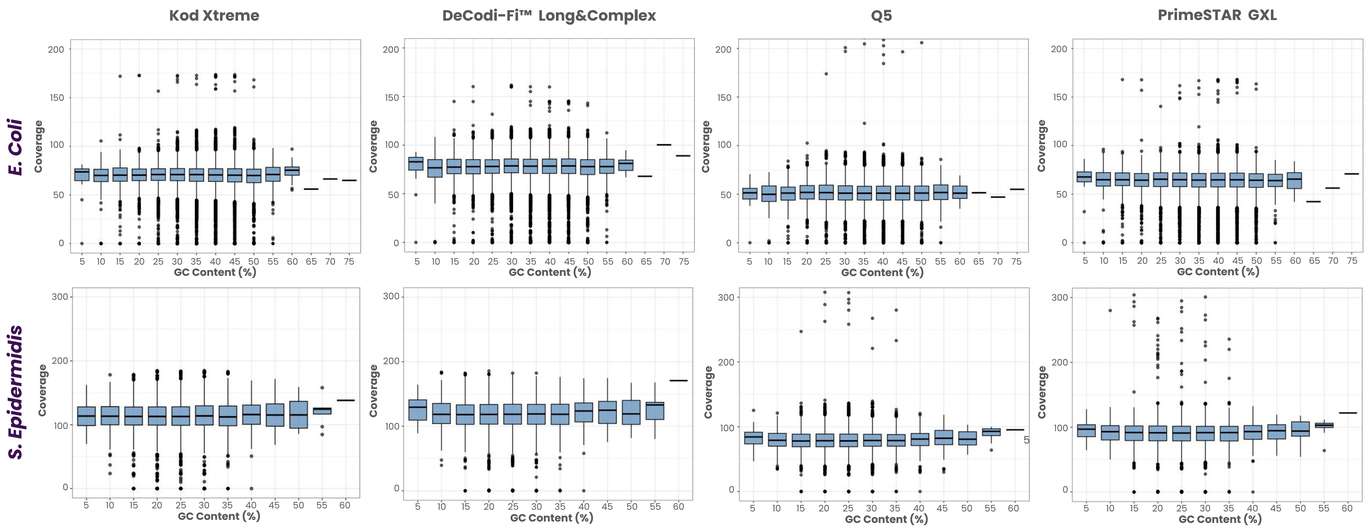

Figure 3: Genomic coverage distribution across GC-content gradients. Box plots illustrate read coverage for *E. coli* (top panels) and *S. epidermidis* (bottom panels) across varying GC-content percentages. To account for differences in sequencing depth, coverage was normalized by downsampling to the lowest common read count per genome following reference-based alignment.

By achieving high, uniform coverage across both genomes, DeCodi-Fi™ Long&Complex and KOD Xtreme effectively minimized GC-coverage bias, satisfying a critical Ampli-Fi requirement for the unbiased representation of difficult-to-sequence genomic regions.

Read Length Distribution

DeCodi-Fi™ Long&Complex verification:

Read length distributions for DeCodi-Fi™ Long&Complex were evaluated across independent triplicates. The overlapping profiles indicate high technical consistency between replicates, confirming that the fragment size distribution is a stable result of the amplification protocol.

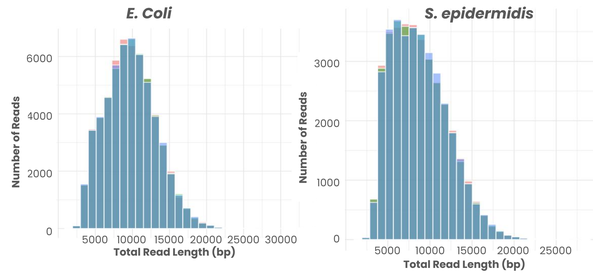

Figure 2: Read length distributions for DeCodi-Fi™ Long&Complex. Length (bp) is plotted against read frequency for triplicate samples of *E. coli* (A) and *S. epidermis* (B). Sample 1: red, sample 2: green, sample 3: blue.

Conclusions

Ultra-Low Input Validation

Four polymerases (DeCodi-Fi™ Long&Complex, KOD Xtreme, Q5, and PrimeSTAR GXL) successfully generated sufficient yield from from just 1 ng of input DNA, meeting the protocol's minimum requirements for challenging or limited samples.

Polymerase Performance

Across both benchmarks, DeCodi-Fi™ Long&Complex and KOD Xtreme delivered the longest mean reads, directly resulting in superior assembly contiguity. DeCodi-Fi™ Long&Complex achieved the highest mean length and N50 values for both *E. coli* (4,585 kb) and *S. epidermidis* (2,145 kb).

In contrast, Q5 despite high raw quality (98% >Q30), its limited processivity produced shorter reads and fragmented assemblies. While all enzymes maintained high gene completeness (~92–98%), these results establish enzymatic processivity as the decisive factor for resolving complex genomic architectures in ultra-low input workflows.

Enzymatic Control of Read Length

Within the Ampli-Fi workflow, read length is a variable that depends largely on the choice of enzyme. Our results demonstrate that enzymatic processivity is the fundamental determinant of assembly contiguity. While high-fidelity enzymes like Q5 offer accuracy, their low processivity causes excessive library fragmentation. By utilizing high-processivity polymerases instead, we can prevent library compression and achieve superior assembly continuity without increasing DNA input requirements.