Microbial whole-genome sequencing using the Singular Genomics G4™ Sequencing Platform and the Roche KAPA EvoPlus Kit

Authors

Sandra Theall
Associate Application Scientist

Sarah Trusiak*
Senior Application Scientist

Mariana Fitarelli-Kiehl
Senior Application Scientist

Lindsey Cambria
Support Team Manager

Spencer Debenport*
Applications Team Manager

Rachel Kasinskas
Director of Support & Applications

Jieqiong Dai
Senior Bioinformatics Scientist

Nikita Raymond Dsouza
Bioinformatics Scientist

Alejandro Quiroz Zarate
Bioinformatics Manager

Sequencing and Life Science
Roche Diagnostics Corporation
Wilmington, MA, USA

**Author no longer an employee of Roche*

Sabrina Shore
Associate Director, Sequencing Applications

Ryan Shultzaberger
Director, Bioinformatics

Chrystal Day
Scientist, Sequencing Applications

Timothy Looney
Senior Director, Scientific Affairs

Martin M Fabani
Vice President, Sequencing Applications

Yu Xiang
Senior Scientist, Bioinformatics

Singular Genomics Systems, Inc.
San Diego, CA

Date of first publication:
July 28, 2023

Publication date of this version:
July 28, 2023

High-throughput next-generation sequencing (NGS) has revolutionized the field of molecular biology. However, as the impact of NGS on human health increases, so does the need for economical and accelerated sequencing results. The G4™ Sequencing Platform from Singular Genomics is an innovative benchtop sequencer that delivers highly accurate results with single-day turnaround and lower cost-per-base. This application note describes the integration of Roche's KAPA EvoPlus Kit with Singular Genomics' unique dual indices and Universal Library Prep Adapter for microbial whole-genome sequencing on the G4 Platform. Libraries prepared with KAPA EvoPlus Kits and sequenced on the G4 show high coverage uniformity across a broad spectrum of genomic GC content, sufficient number of contigs and contig lengths to facilitate de novo assembly, and expected levels of start-site complexity; together, these highlight the compatibility of these methods with a wide range of genomic material.

Introduction

Whole-genome sequencing (WGS) has become a standard tool in biomedical research, enabling researchers to address a wide variety of challenges in public health and diagnostics.[1] To facilitate more widespread use of WGS and other next-generation sequencing (NGS) applications, sequencing platforms and technologies have continuously evolved to offer higher throughputs, greater accuracy, and lower cost-per-base.[2] In addition, pre-sequencing workflows can be streamlined by using NGS library preparation methods that include enzymatic fragmentation of input DNA, thus enabling efficient use of the expanded sequencing power offered by new NGS platforms. The KAPA EvoPlus Kit—the most recent addition to the Roche library preparation portfolio—offers a streamlined enzymatic-fragmentation library preparation solution that requires no master-mix setup or EDTA treatment steps.

The Singular Genomics G4™ Sequencing Platform is an innovative benchtop sequencer combining novel 4-color Rapid sequencing by synthesis (SBS) chemistry with advanced engineering to provide single-day turnaround times for a broad range of applications. The G4's ability to deliver fast results and run 1-4 flow cells in parallel, each with 4 independently addressable lanes, enables laboratories to conduct highly efficient operations. More information about G4 specifications, such as run time, accuracy, and quality metrics, can be found on the Singular Genomics website.

This application note describes the preparation of high-quality libraries for the G4 using KAPA EvoPlus Kits with the Singular Genomics Universal Library Prep Adapter, Cleave Enzyme, and unique dual-indexed (UDI) primers. Sequencing metrics are presented for bacterial WGS libraries created from a mixed-DNA input sample containing DNA from three bacterial species that are relevant to human health, and which represent a broad range of genomic GC contents.

Materials and Methods

Input Sample:

Three microbial species that are relevant to human health were used to model a broad range of genomic GC contents: *Clostridium difficile* (29% GC), *Escherichia coli* (51% GC), and *Bordetella pertussis* (68% GC). Bacterial genomic DNA was obtained from the American Type Culture Collection (ATCC). Strains and accession numbers were as follows: *C. difficile* (Hall and O'Toole) Prevot, strain 630 (BAA-1382D-5); *E. coli* (Migula) Castellani and Chalmers, strain MG1655 (700926D-5) and *B. pertussis* (Bergey, et al.) Moreno-Lopez, strain Tohama 1 (BAA-589D-5). The genomic DNA of each bacterial species was equally mixed by mass, and triplicates of 10 ng, 50 ng and 100 ng of the resulting mixed-DNA sample were used as input into library preparation.

Library Preparation and Sequencing

Triplicate libraries were prepared from the microbial genomic DNA mixture using three input amounts: 10 ng, 50 ng and 100 ng. Libraries were prepared with the KAPA EvoPlus Kit and the Singular Genomics (SG) Universal Library Prep Adapter. Following enzymatic fragmentation of the input DNA for 25 min at 37°C (targeting a mode fragment length of 300 bp), DNA fragments were ligated to the universal adapters (550 nM) for 15 min at 20°C.

To open the cleavable sites on the universal adapter, which is a stem-loop adapter, the cleavage reaction was performed by adding 3 µL of Singular Genomics' Cleave Enzyme, 25 µL of buffer and 3.5 µL of nuclease-free water to each library and incubating at 37°C for 10 minutes, followed immediately by 67°C incubation for 30 minutes. A 0.8X bead cleanup using KAPA HyperPure beads was performed to remove residual reagents from previous steps. Indices were incorporated via PCR using the SG UDI Primers Set A [1-24] at 2 µM as follows: 6 cycles of PCR for 10 ng libraries, 4 cycles of PCR for 50 ng libraries, and 3 cycles of PCR for 100 ng libraries. A 1.0X bead cleanup using KAPA HyperPure beads was performed to remove residual PCR amplification reagents. Libraries were quantified with Qubit 1X HS dsDNA assay and library phenotypes validated with Agilent Bioanalyzer HS DNA assay.

Sequencing was carried out in a single F2 flowcell using all 4 lanes, with 150-cycle paired reads and dual 12 bp index reads.

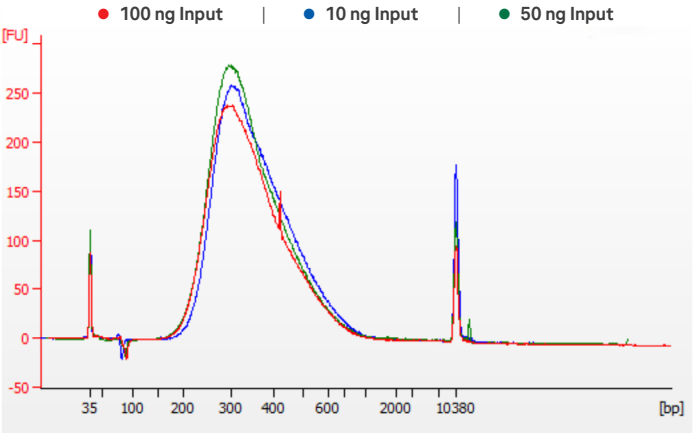
Data Analysis: Adapter and quality trimming was performed using Cutadapt (v. 4.1). Reads were aligned with BWA MEM (v. 0.7.12) and downsampled to the lowest common number of reads (3,000,000) using Seqtk (v. 1.3). GC bias was calculated using Picard CollectGCBiasMetrics (v. 2.27.0), and coverage was calculated with Mosdepth (v. 0.3.3). De novo assembly was performed using Spades (v. 3.15.2), and metrics were collected using Quast (v. 5.0.1). The genomic regions of alignment start sites were extracted using Bedtools (v. 2.30.0), and start site complexity was analyzed using Fastqc (v. 0.11.9).

Results and Discussion

Library QC Metrics (Pre-sequencing)

To determine whether the final amplified libraries displayed appropriate phenotypes and yielded sufficient material for sequencing, libraries were analyzed for size and concentration **(Figure 1)**. Yield was calculated using average library size (Bioanalyzer) and concentration (Qubit). All libraries yielded sufficient material for sequencing on the Singular Genomics G4 Sequencing Platform.

Panel A.



Panel B.

Library	Average (bp)	std dev (bp)
10 ng input	360	8
50 ng input	352	1
100 ng input	350	4

Panel C.

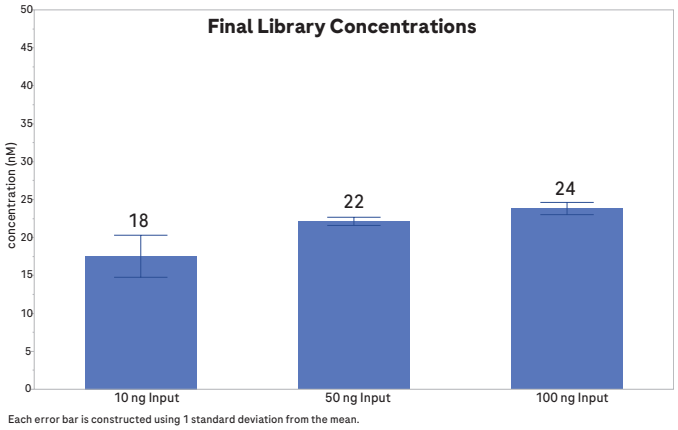


Figure 1. Pre-sequencing metrics for amplified libraries created using 10, 50, or 100 ng of mixed-DNA input. Panel A. Electropherograms generated with the Bioanalyzer High Sensitivity DNA Kit (Agilent Technologies). Representative data are shown for three libraries (one library per set of triplicate libraries created with each input mass). Libraries were diluted 1:10 prior to analysis. **Panel B.** Average library size (with standard deviation), calculated using all libraries (undiluted) across each input mass condition. **Panel C.** Average concentrations of amplified, undiluted libraries determined with Qubit 1x HS dsDNA Assay (Invitrogen). All libraries yielded sufficient material for sequencing. For each individual library, concentrations from all triplicate libraries across each input mass condition (Qubit) and average library size (Bioanalyzer) were used to determine sufficiency for sequencing.

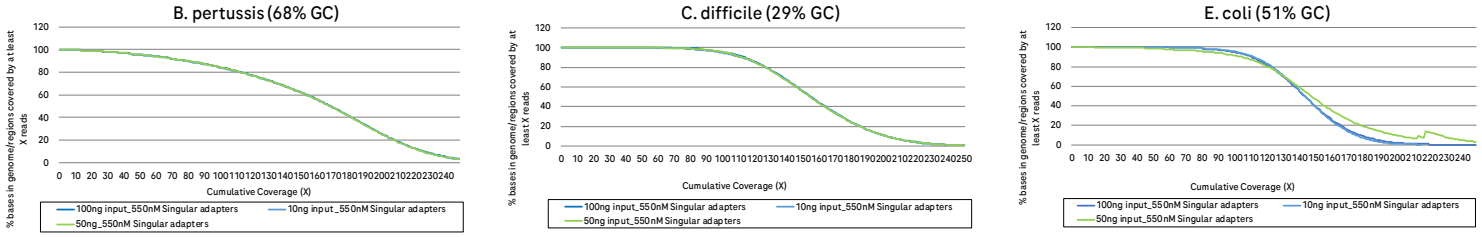


Figure 2. Coverage uniformity plots. Coverage uniformity describes how evenly the target regions are represented. Data for all libraries were downsampled to 3 million reads and coverage calculated using Mosdepth.

Run performance metrics

Flowcell throughput was 208M paired reads and the percentage basecalls ≥ Q30 was 87.2% and 92.0% for R1 and R2, respectively.

Sequencing analysis metrics

Coverage uniformity

The KAPA EvoPlus workflow with the SG Universal Library Prep Adapter yielded highly similar coverage profiles for all three bacterial species across all input mass amounts, demonstrating flexibility across a range of inputs for a wide range of GC contents **(Figure 2)**.

GC bias

GC bias is a potential artifact in NGS that presents as uneven coverage of regions with high or low GC content in a genome; GC bias can also lead to coverage gaps, greatly reducing the amount of information available for analysis.[3] GC bias is a greater challenge for genomes with extreme levels of GC content (such as above 75% or below 15%) and can be introduced at several steps of the NGS workflow, including: library preparation (fragmentation, adapter ligation, or library amplification), sequencing, and data analysis.

The bacterial genomes used in this study represent a wide range of genomic GC content. GC coverage for these libraries was assessed by calculating the normalized coverage across GC content using 100 bp-window bins, shown in normalized GC coverage plots in **Figure 3**.

Libraries generated with KAPA EvoPlus and sequenced on the Singular Genomics G4 Sequencing Platform showed similar coverage uniformity **(Figure 3)** for bacterial species with low- and mid-range genomic GC content *(C. difficile and E. coli). B. pertussis* yielded libraries with reduced representation of high-GC areas; this is typical for high-GC content genomes.

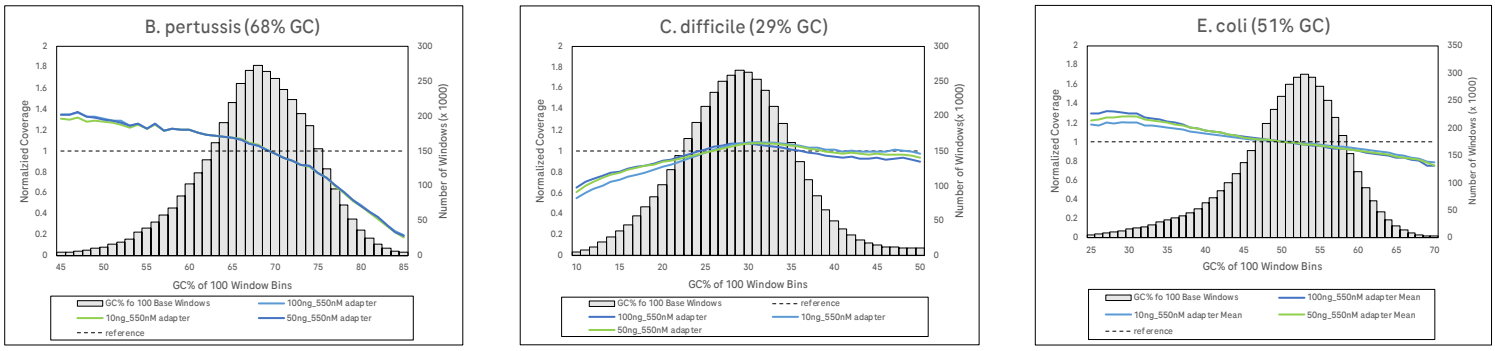


Figure 3. GC coverage plot. GC-bias plots were generated with Picard CollectGCBiasMetrics. Gray histograms represent the distribution of GC content for each bacterial genome, calculated for the reference sequence in 100 bp bins; note that the histogram curve shifts to the right or left for each species. GC-bias for each workflow was assessed by plotting the normalized coverage for each bin (colored lines), as average of three replicate libraries. If all sample-to-data processes were completely unbiased, all bins would be equally represented and the plot for each workflow would be a horizontal distribution centered on a normalized coverage of 1 (on the Y axis).

De novo Assembly

Microbial WGS may sometimes require de novo assembly—for instance, when a reference genome is not available or novel genes are being examined. The assembly process begins with shearing genomic DNA into consistent and appropriate fragment sizes for sequencing, requiring a library construction method that provides consistent and tunable fragmentation such as KAPA EvoPlus.[4] Subsequently, the sequencing reads generated from the fragmented DNA are assembled by identifying large overlapping read segments to produce a contig. A scaffold is then constructed by linking two or more joined-oriented contigs, then combined to form a chromosome. The quality of the inferred chromosome sequence is impacted by the "holes" between the scaffolds and the length of the contigs. Quantitatively the quality of the assembly can be inferred by the number of contigs, length of the longest contig and N50.

The NGS results for all prepared libraries were compared across the above described metrics **(Figure 4)**. Since coverage uniformity facilitates assembly, GC bias from samples containing extremes in GC content poses a challenge.[3] Thus, it was expected that the more GC-rich *B. pertussis* genome yielded more contigs that were also shorter in length, likely due to the assembler treating these regions as repetitive elements. The less GC-rich genomes *C.difficile* and *E.coli* translated to fewer and longer contigs.

Start Site Complexity

Start site complexity plots show the nucleotide content of all aligned reads in a 40-bp window (-10 to +30 bp) relative to the alignment start. Start site bias, which can be introduced during fragmentation, potentially impacts library diversity (i.e. number of unique reads representing each genome position). As seen in **Figure 5**, all libraries performed as expected based on the overall GC content of each bacterial genome, across all input amounts tested.

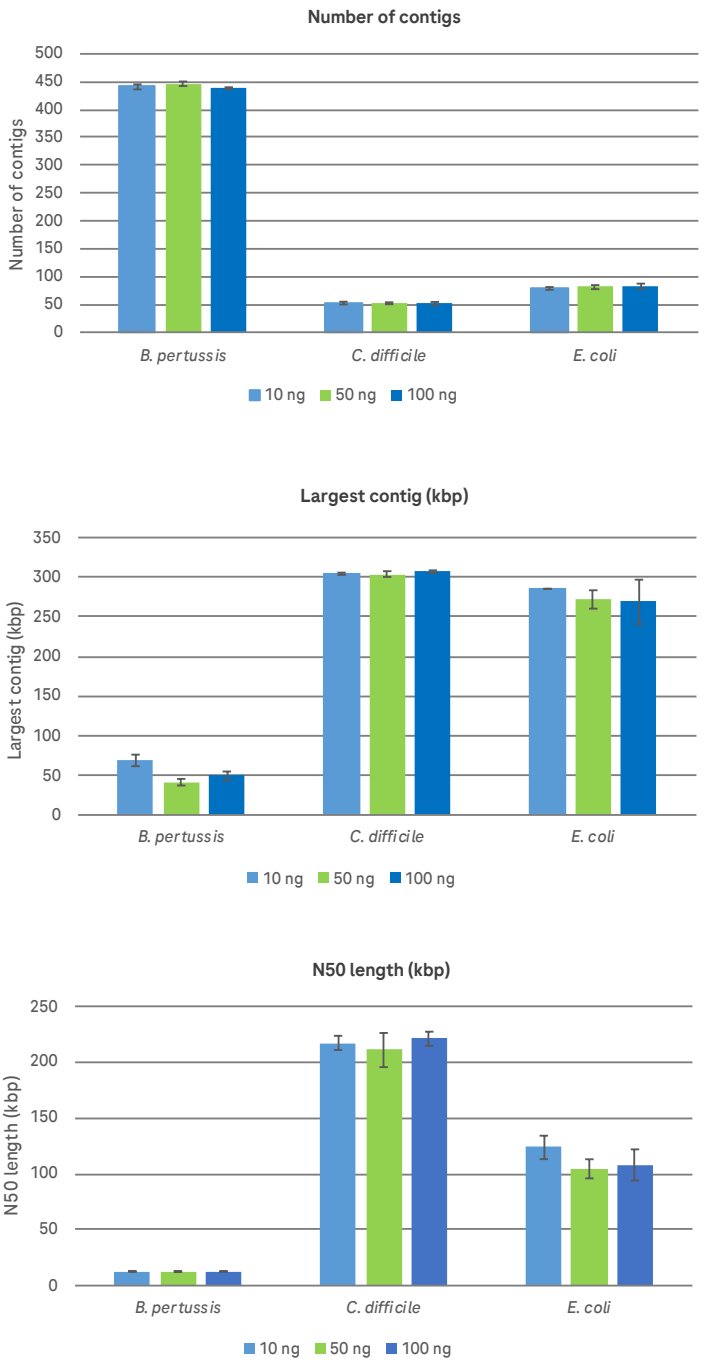


Figure 4. De novo assembly metrics. The KAPA EvoPlus workflow was assessed with respect to three key de novo assembly metrics. **Number of contigs**. In general, a large number of contigs implies the inferred assembly will have a large number of "holes". High-quality assemblies present a low number of contigs. **Largest contig**. In combination with the number of contigs and N50, this metric provides an assessment of the quality of the recovered genome. **N50**. N50 is defined as the contig sequence length at 50% of the length distribution of all the contigs.

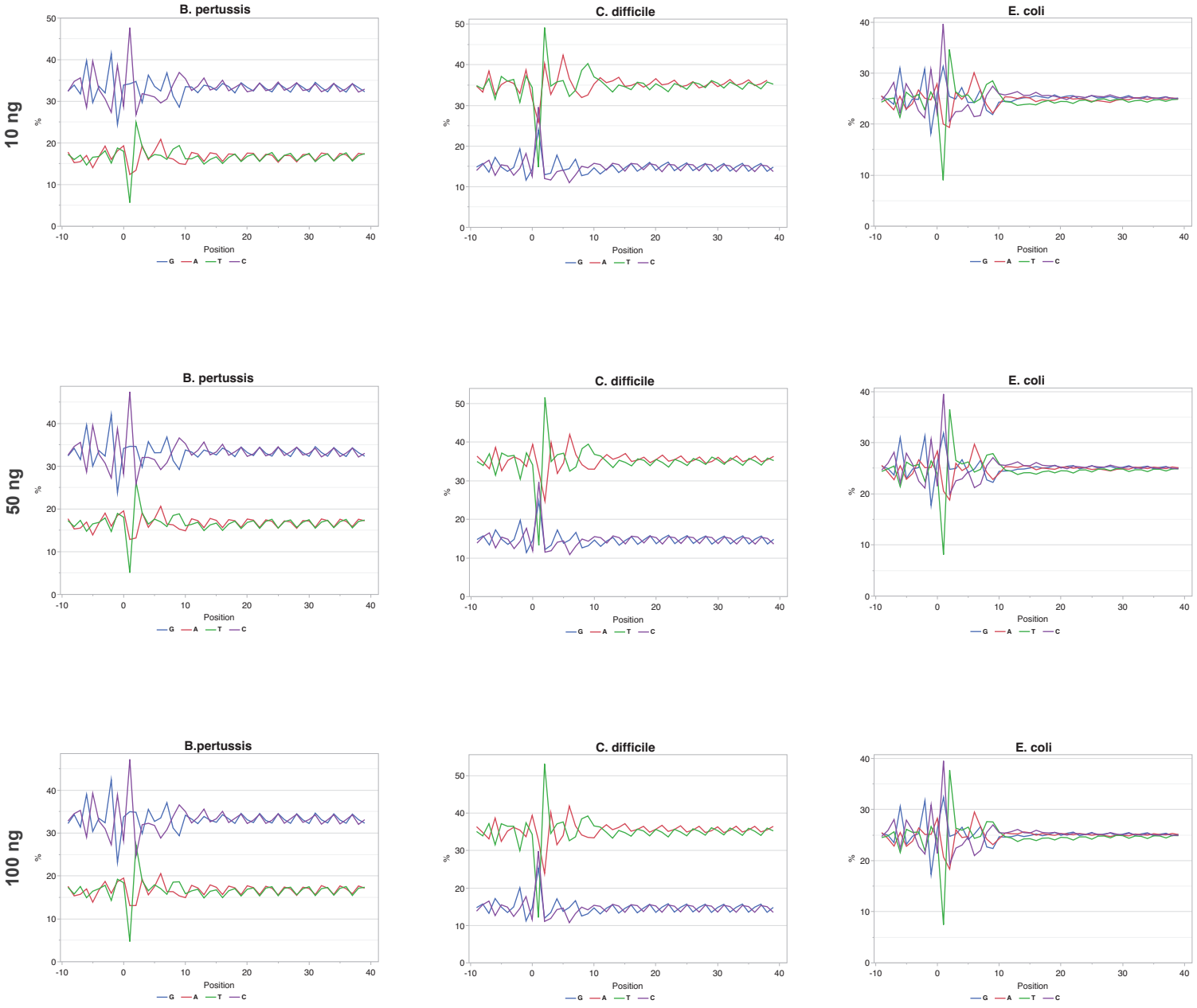


Figure 5. Start site complexity plots. Nucleotide content over a 40 bp window (-10 to +30 bp relative to read alignment start) for *B. pertussis* (68% GC), *C. difficile* (29% GC), and *E. coli* (51% GC), for libraries prepared with the KAPA EvoPlus workflow and Singular Genomics' unique dual indices and Universal Library Prep Adapter. If all three library construction processes (fragmentation, adapter ligation, and library amplification) as well as sequencing and data analysis were completely unbiased, each base (A, C, G, and T) would be represented by a perfectly flat, horizontal line with a y-axis value corresponding to the average prevalence of that nucleotide in the genome. For example, the A and T plots for *C. difficile* (29% genomic GC content and 71% AT content) would be overlayed, and have a value of ~35% for each position, whereas the C and G plots would both have a value of ~15% at each position.

Conclusion

KAPA EvoPlus Kits enable the preparation of high-quality DNA libraries for the Singular Genomics G4™ Platform. Libraries prepared using KAPA EvoPlus are able to incorporate the SG Universal Library Prep Adapter, enabling compatibility with the G4 Sequencing Platform and maintaining quality across key metrics for microbial WGS. Microbial WGS samples prepared with KAPA EvoPlus Kits and sequenced on the G4 platform show high uniformity of coverage across a broad spectrum of genomic GC content, acceptable number of contigs and contig lengths to facilitate de novo assembly, and expected levels of start-site complexity; these results highlight the compatibility of these methods with a wide range of genomic material.

In summary, this new workflow—the streamlined convenience and robust performance of KAPA EvoPlus Kits and Singular Genomics' unique dual indices and Universal Library Prep Adapter—yields high-quality data when the resulting libraries are sequenced on the Singular Genomics G4 Platform, demonstrating that this library preparation workflow yields fast, flexible, and accurate microbial whole-genome sequencing.

References

1. Park, S. T., & Kim, J. (2016). Trends in Next-Generation Sequencing and a New Era for Whole Genome Sequencing. International Neurourology Journal, 20 (Suppl 2), S76–S83. https://doi.org/10.5213/inj.1632742.371
2. Roche Sequencing & Life Science. Application Note. Miller B, et al. A novel, single-tube enzymatic fragmentation and library construction method enables fast turnaround times and improved data quality for microbial whole-genome sequencing. December 2017.
3. Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., & Hwang, C. C. (2013). Effects of GC bias in next-generation-sequencing data on de novo genome assembly. PLOS ONE, 8(4), e62856. https://doi.org/10.1371/journal.pone.0062856
4. Roche Sequencing Solutions, Inc. Application Note. Adams N, et al. KAPA EvoPlus Kits: Continued evolution sets a new standard in high-performance, streamlined library preparation for a wide range of applications. May 2022.

This page intentionally left blank

For more information about Roche KAPA EvoPlus Kits,
please visit: **go.roche.com/GetEvoPlus**

Published by:
Roche Sequencing and Life Science
9115 Hague Road
Indianapolis, IN 46256

sequencing.roche.com

Project name: Microbial whole genome sequencing using the G4 Sequencing Platform and the Roche KAPA EvoPlus Kit
For Research Use Only. Not for use in diagnostic procedures.
KAPA EVOPLUS and KAPA HYPERPREP are trademarks of Roche.
G4 and Singular Genomics are trademarks of Singular Genomics Systems, Inc.
All other product names and trademarks are the property of their respective owners.

MC-US-13173 APP118002 A751 08/23