Application Note

High-performance, streamlined library prep

Authors

Nicolette Adams
Principal Scientist I – Research and Early Development

Clare Garrard
Senior Bioinformatics Scientist

Jacqueline Meyer
Applications Technical Product Manager – Global Customer Support Sample Prep Reagents

Samantha Dockrall
International Product Manager – Sample Prep Reagents

Roche Dia Cape Town, South Africa

Heather Maasdorp
International Business Leader – Sample Prep Reagents

Roche Molecular Labs, Pleasanton, CA, USA

Date of publication:
May 2022

KAPA EvoPlus Kits: Continued evolution sets a new standard in high-performance, streamlined library preparation for a wide range of applications

Ligation-based library preparation is routinely used across a wide range of NGS applications. Protocols that employ enzymatic DNA fragmentation offer significant workflow advantages, but have shortcomings. KAPA EvoPlus Kits offer a more streamlined workflow, more robust fragmentation, improved stability, more convenient reagent formats, and reduced sequencing artefacts to support high-performance, fast, and scalable library preparation for short-read sequencing applications.

Introduction

Perpetual advances in next-generation sequencing (NGS) technology and computational biology continue to create a demand for library preparation solutions that are flexible, fast, and scalable—without sacrificing performance. We set an industry standard in 2014 with the launch of the KAPA HyperPrep Kit, which offered a streamlined, single-tube protocol *and* significant improvements in library construction efficiency over the best homebrew and commercial protocols available at that time.[1] KAPA HyperPlus Kits with integrated enzymatic fragmentation followed in 2015, obviating the need for mechanical DNA shearing and enabling even higher conversion rates.[2,3] This supported fully automated workflows and robust, high-performance library preparation for a wide range of applications, including somatic variant analysis from FFPE samples and—more recently—epidemiological, biochemical, and immunological studies of the SARS-CoV-2 virus. To date, KAPA HyperPrep and KAPA HyperPlus Kits have been referenced in more than 950 peer-reviewed publications.[4]

Notwithstanding the broad adoption of KAPA HyperPrep and KAPA HyperPlus Kits, there is always room for improvement. In the case of KAPA HyperPlus Kits, lower stability of the enzymatic fragmentation module and sensitivity to EDTA in input DNA presented challenges to implementation in certain settings. In addition, studies published since 2019 have identified sequencing artefacts attributable to the enzymatic fragmentation cocktails included in several commercially available library prep kits.[5,6]

KAPA EvoPlus Kits represent the latest milestone in the continued evolution of KAPA library preparation chemistry. Novel enzyme and buffer formulations enable an even more streamlined workflow, more robust fragmentation in the presence of EDTA and other buffer components, and improved stability that enables pre-mixed and plated reagents. In addition, sequencing artefacts have been reduced to levels comparable to those achieved with ligation-based library preparation from mechanically sheared input DNA, and significantly lower than artefact levels produced with kits from other suppliers. This boosts confidence in sequencing results across a wide range of sample types and applications, whilst facilitating adoption in higher throughput and automated sample preparation pipelines.

Workflow improvements

Combined enzymatic fragmentation and A-tailing enables further streamlining and reduces hands-on time

With KAPA HyperPlus Kits, adapter-ligated libraries are constructed in a single tube, but enzymatic fragmentation, end repair/A-tailing, and adapter ligation are three discrete steps requiring individual reagent additions. In the KAPA EvoPlus protocol (Figure 1), all pre-ligation enzymatic activities are combined into a "FragTail" reaction that requires the pipetting of a single, ready-to-use reagent mix. Once the two-step (37°C/55°C) FragTail incubation has been completed, a ligation mix and adapters are added. The ligation incubation (15 min at 20°C) is followed by a single-sided, bead-based post-ligation cleanup, and adapter-ligated libraries are recovered from the same tube in which input DNA was dispensed. As for the KAPA HyperPrep and KAPA HyperPlus workflows, library amplification (with the gold standard KAPA HiFi HotStart ReadyMix) is optional when using a KAPA EvoPlus Kit, and additional single- or double-sided cleanups may be incorporated as needed for a specific sample type and/or sequencing application.

The total time needed to prepare a batch of libraries with the KAPA EvoPlus Kit is similar to that for the KAPA HyperPlus Kit (~2 hours for PCR-free workflows and 2.5 hours for workflows that require library amplification), but KAPA EvoPlus hands-on time is shorter (and walkaway time longer) due to fewer reagent manipulations and faster fragmentation.

Reduced sensitivity to EDTA and other buffer components supports robust, tunable DNA fragmentation

Preservatives such as EDTA and sodium azide (NaN_3) are commonly included in DNA storage/dilution buffers. These preservatives inhibit the activity of many enzymes used in molecular biology and are known to inhibit first-generation enzymatic fragmentation cocktails, leading to inconsistent fragmentation within or between sample batches. Depending on the application, over- or under-fragmentation may negatively impact sequencing metrics and cost, or confound assembly and data analysis. The potential impacts of these buffer components may be mitigated by the inclusion of counter-inhibitors (e.g., the Conditioning Solution included in KAPA HyperPlus Kits), but this complicates sample processing, particularly in high-throughput settings.

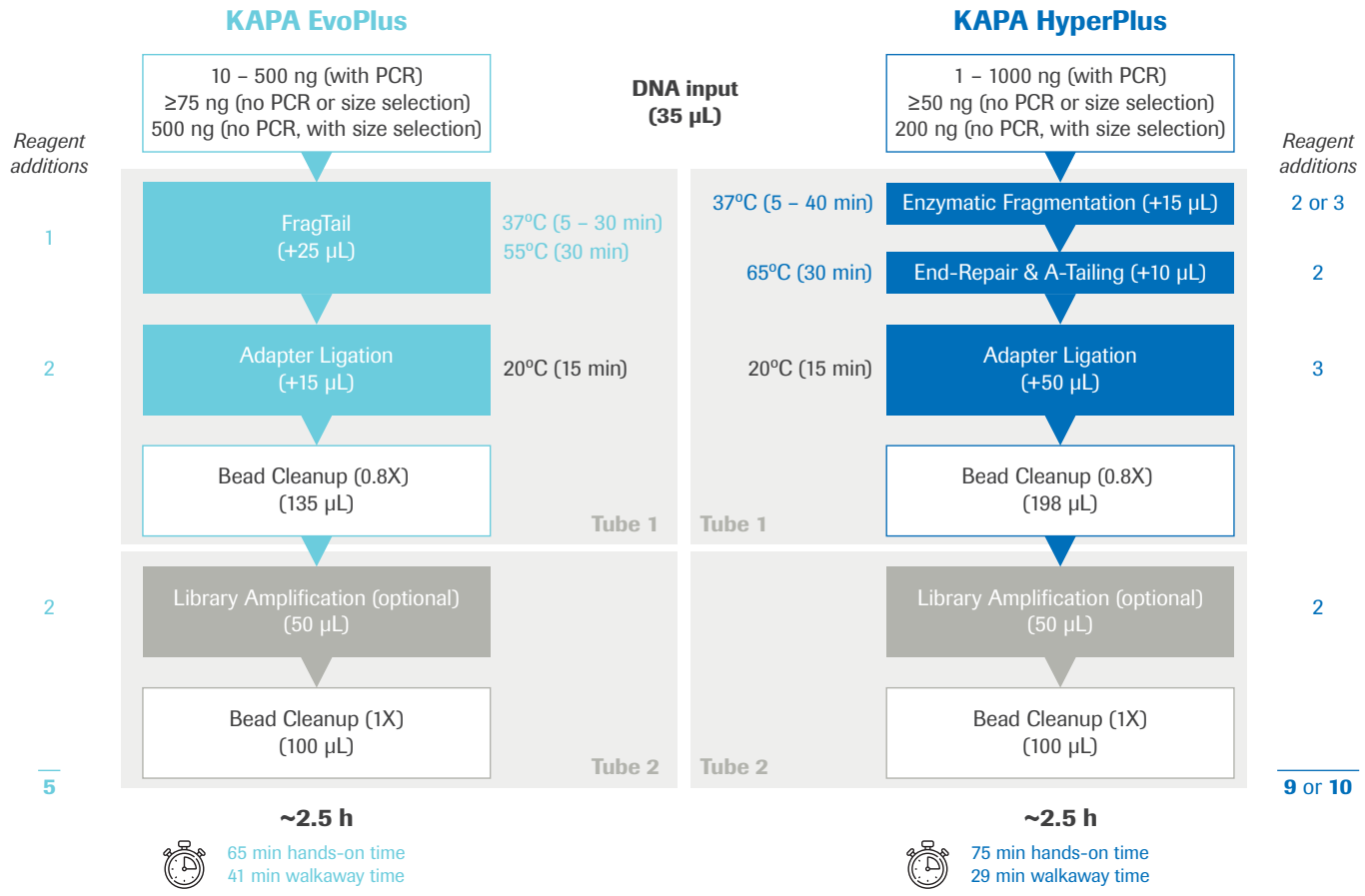

Figure 1. Comparison between the KAPA EvoPlus and KAPA HyperPlus workflows. KAPA EvoPlus Kits are fully validated with 10 – 500 ng of input DNA. Lower inputs may be used if the available number of genome equivalents is sufficient for the application, but will require optimization of the fragmentation (37°C incubation) time. Minimum recommended inputs for PCR-free workflows (with or without an optional post-ligation size selection) are higher with the KAPA EvoPlus Kit. This is a consequence of lower overall yields, resulting from trade-offs to achieve improved reagent stability and reduce sequencing artefacts (see subsequent sections for more details). In both protocols, adapter-ligated libraries are recovered from the same tube in which input DNA was dispensed, but the KAPA EvoPlus Kit requires fewer reagent additions and less fragmentation and hands-on time. The KAPA EvoPlus Ligation Mix has a unique composition, allowing for a significantly reduced ligation volume. Both KAPA EvoPlus and KAPA HyperPlus Kits are fully compatible with full-length KAPA Unique Dual-Indexed Adapters and the "stubby" KAPA Universal Adapter plus KAPA UDI Primer Mixes. KAPA HyperPure Beads are recommended for the KAPA EvoPlus workflow, but KAPA Pure Beads may also be used. KAPA HiFi HotStart ReadyMix has not undergone any changes as a result of the development of the KAPA EvoPlus Kit, and is included in standard (as opposed to PCR-Free) KAPA EvoPlus Kits. KAPA Library Amp Primer Mix is recommended for standard Illumina® sequencing applications, and is available separately.

The second-generation KAPA EvoPlus fragmentation chemistry was specifically formulated to be less sensitive to EDTA and other buffer compounds. As shown in Figure 2 (A and B), consistent fragmentation results are obtained with input DNA in Tris-based buffers (10 mM Tris-HCl, pH 7.0 – 9.0) containing ≤2 mM EDTA or <0.1% NaN_3. Higher EDTA concentrations (up to 5 mM in the input DNA) can be tolerated, but will require longer fragmentation times to achieve the desired fragment size distribution.

The KAPA EvoPlus fragmentation chemistry produces robust and reproducible results across a 10- to 50-fold input range (depending on genome complexity; Figure 2C) and DNA of variable quality (Figure 2D). As shown in Figure 2E, fragment size is controlled by incubation time at the standard incubation temperature of 37°C. Depending on the workflow and sequencing application, fragmentation times may be reduced to obtain shorter library insert sizes (and a tighter size distribution), or increased for higher inputs and/or to overcome fragmentation inhibition. Since the KAPA EvoPlus FragTail chemistry is more effective, shorter incubation times at 37°C are needed to achieve a specific mean/median fragment size, as compared to the time required when using the KAPA HyperPlus Kit (not shown). If optimal incubation times become inconveniently short (e.g., <5 min), especially when working with lower inputs or less complex templates, fragmentation times may be extended by reducing the temperature for the first stage of the FragTail incubation (from 37°C to ≤25°C; Figure 2F).

Evolved reagent formulations allow for stable, ready-to-use master mixes and expanded reagent formats

Enzymatic fragmentation cocktails contain a number of enzymes, some of which require different chemical and physical parameters for optimal activity and stability. This impacts protocol design, as well as the stability of reagent master mixes and the overall shelf-life of kits.

Improved reagent stability was a key driver in the development of the KAPA EvoPlus chemistry. In addition to supporting a more streamlined protocol, novel reagent formulations for all of the core library construction steps enabled ready-to-use master mixes, longer shelf-life, and expanded reagent formats (Table 1). This provides more flexibility in both manual and automated library preparation pipelines, improves ease-of-use, reduces risk and waste, and facilitates reagent inventory management.

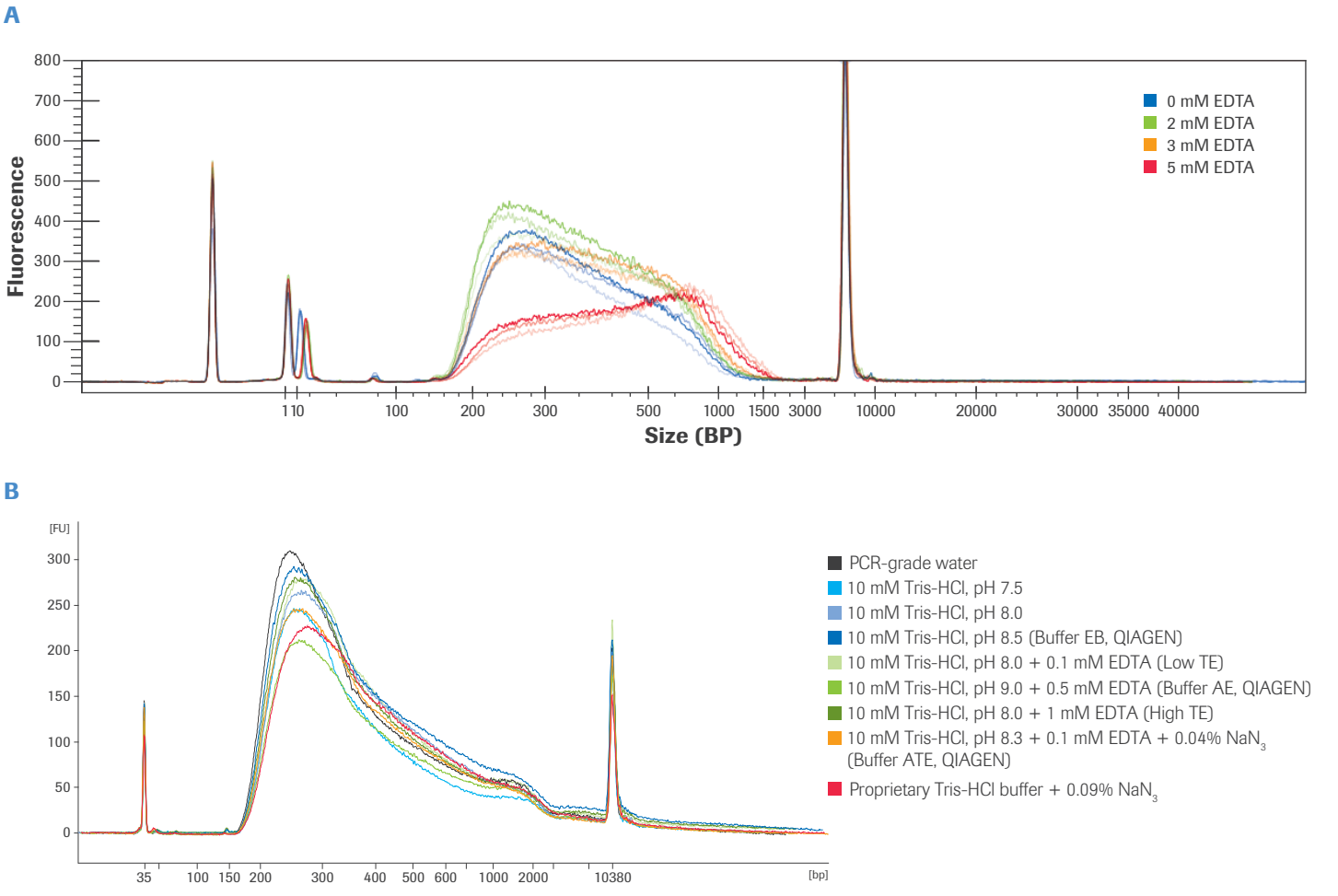

Figure 2. Fragment size distributions of KAPA EvoPlus libraries prepared from different DNA inputs in different buffer chemistries, with different enzymatic fragmentation parameters. (A) PCR-free libraries prepared from high-quality *E. coli* K-12 genomic DNA (100 ng inputs in 10 mM Tris-HCl, pH 8.0 + 0 – 5 mM EDTA; 15 min fragmentation at 37°C). An aliquot of each library was amplified for 3 cycles prior to electrophoretic analysis to obtain accurate fragment size distributions. **(B)** PCR-free libraries prepared from high-quality human buffy coat genomic DNA (100 ng inputs, in PCR-grade water or 10 mM Tris-HCl buffers with different pH and/or buffer components, as indicated). DNA was fragmented for 25 min at 37°C. An aliquot of each library was amplified for 8 cycles prior to electrophoretic analysis. Additional experimental details may be found in *Materials and methods*.

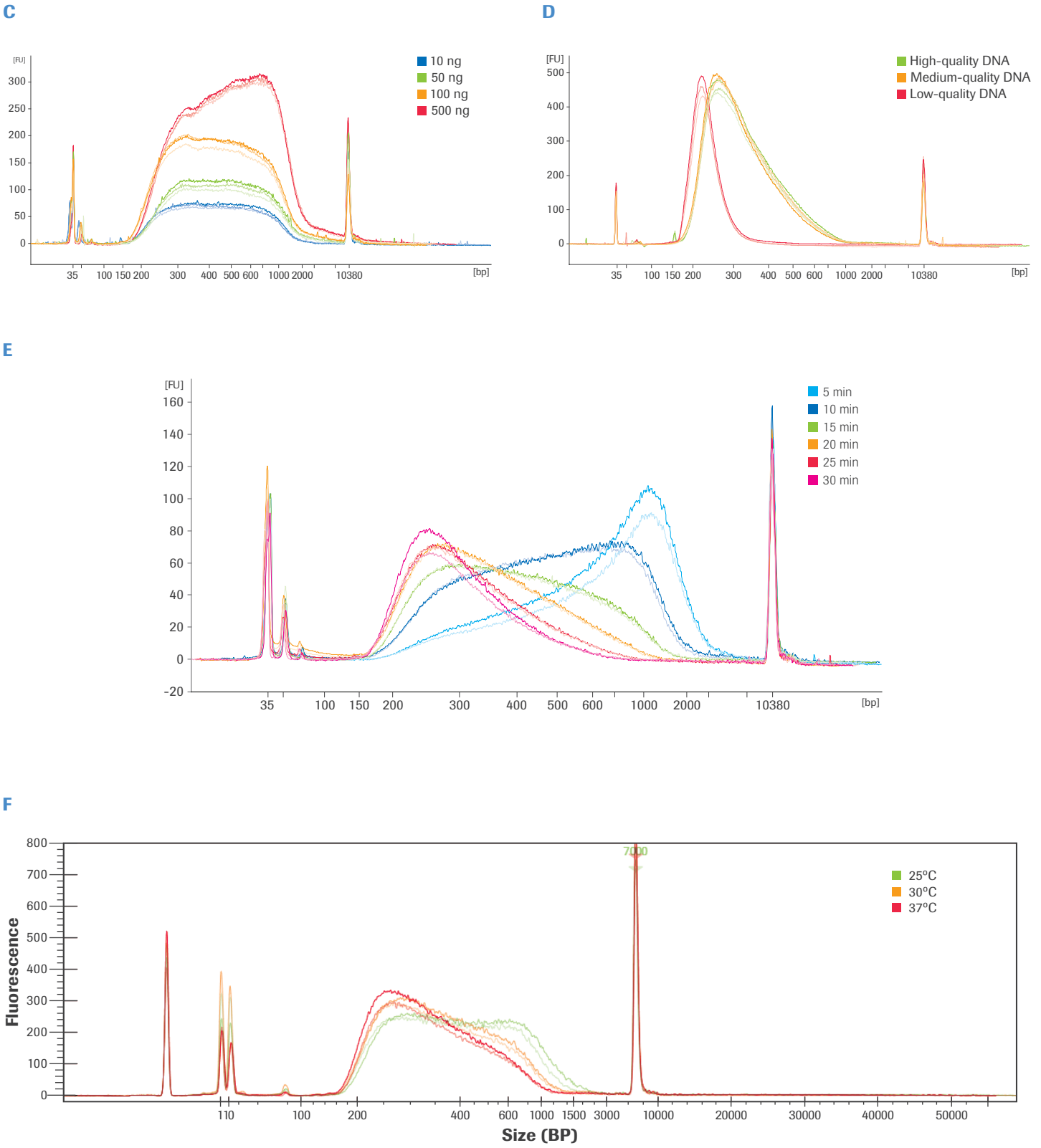

Figure 2. Fragment size distributions of KAPA EvoPlus libraries prepared from different DNA inputs in different buffer chemistries, with different enzymatic fragmentation parameters (continued). (C) Libraries prepared from human NA12878 genomic DNA (10 – 500 ng; fragmented for 10 min at 37°C to target a mean fragment size of 350 bp), and amplified for 2 cycles (10 ng input) or 5 cycles (50 – 500 ng inputs). **(D)** Libraries prepared from high-quality human NA12878 genomic DNA, or medium- to low-quality DNA extracted from fresh-frozen (FF) or FFPE tissue (100 ng; 20 min fragmentation at 37°C for downstream target enrichment). FFPE libraries were amplified for 8 cycles and all other libraries for 6 cycles. **(E)** PCR-free libraries prepared from *E. coli* K-12 genomic DNA fragmented for 5 – 30 min at 37°C (100 ng input, aliquots amplified for 5 cycles prior to electrophoresis). **(F)** Libraries prepared from *E. coli* K-12 genomic DNA (100 ng), fragmented for 15 min at 25, 30, or 37°C. An aliquot of each library was amplified for 3 cycles prior to size assessment. Additional experimental details may be found in *Materials and methods*. Unless otherwise stated, input DNA was of a high quality, and prepared in 10 mM Tris-HCl, pH 8.0.

Data on file.

Table 1. Features and benefits of improved KAPA EvoPlus reagent formulations

Feature	Benefits
Combined FragTail reaction	▪ More streamlined protocol with less hands-on and more walkaway time
Ready-to-use master mixes	▪ No preparation time (other than preparation of input DNA) ▪ Fewer reagent additions reduces hands-on time and the probability of error and inconsistency ▪ Less plasticware required to process a batch of samples, translating to lower per-sample cost and less plastic waste
Improved stability	▪ Kits can withstand more freeze-thaw cycles (20) ▪ Improved stability at room temperature (24 h) reduces distribution-related risks and simplifies use ▪ Longer shelf-life (18 months at -15 to -25°C) facilitates reagent inventory management
Expanded reagent formats	▪ Improved stability enables larger pack sizes (384 rxn kits) ▪ Standard (tube-based) or plated reagents (96 rxn kits) provide support for manual and automated library preparation pipelines ▪ Plated reagents with pierceable and peelable seals and generous overages support automated liquid handling ▪ All reagents are barcoded for traceability and compatibility with LIMS systems

Performance improvements

KAPA EvoPlus chemistry reduces sequencing artefacts

Every stage of an analytical workflow—sample preparation, the analysis itself, and data processing—has the potential to distort the information captured in the biological sample as a result of experimental biases and artefacts. This impacts variant analysis and conclusions about molecular and cellular structure, function, interactions, regulation, and dynamics. In NGS workflows, artefacts have been attributed to formalin fixation,[7] Covaris® shearing,[8] transposases (used in tagmentation-based library preparation),[9] ligases,[10] library amplification,[11,12] and sequencing barcode misassignment[13,14]—to name but a few. Identification and characterization of these biases and artefacts have lead to improvements in sample preparation and sequencing technologies, and continue to drive the development of computational tools to mitigate their impacts.

Artefacts attributed to the enzyme cocktails used for enzymatic DNA fragmentation have recently been described (Table 2). Findings from these studies were taken into consideration during the development of the KAPA EvoPlus chemistry. With the new kit, sequencing artefacts have been substantially reduced—in most cases to levels comparable to those achieved with ligation-based library preparation from Covaris-sheared DNA, and significantly

Table 2. Sequencing artefacts associated with enzymatic DNA fragmentation cocktails used in NGS library prep

Artefact	Explanation/definition	Implications	References
False SNV and indel variant calls	Single nucleotide variants (SNVs) and small insertions or deletions (indels), identified as the result of misalignment between sequencing data to a reference that originated from sample processing, sequencing error, or bioinformatic analysis.	Incorrect interpretation of experimental data, leading to inaccurate conclusions about biological phenomena, molecular mechanisms, and biomarkers. In clinical settings, this may have significant consequences for the diagnosis, treatment, and management of diseases.	5, 15, 16
Strand Split Artefact Reads (SSARs)	Artefact reads made up of at least two distinct biological sequences that align to different areas of the reference genome, on opposite strands of DNA that are in close proximity to one another.	Increases the frequency of sequence bias, base error rates, and false positive detection of SNVs and copy number variants (CNVs).	6, 7
Soft-clipped bases	Portions of a read found at either/both of the 5'- and 3'-ends that do not align to the reference and are ignored during alignment. In contrast, hard-clipped bases are removed prior to alignment. Soft-clipped bases may be associated with short hairpins, SSARs, or inversions.	Incorrectly trimmed reads may lead to misassignment of reads, particularly to repetitive regions. This may result in incorrect characterization of variants, especially structural variants (SVs).	17, 18
Chimeras*	Artefact corresponding to a sequence that does not exist in nature. Chimeric reads consist of at least two distinct biological sequences which align to different areas of the reference genome in the same orientation.	Inaccurate representation of biological diversity and inaccurate variant profiling.	7, 19

*May result from enzymatic fragmentation, ligation, and/or library amplification.

lower than those obtained with kits from other suppliers (Figure 3). These improvements in sequencing data quality translate to fewer false variant calls (for single nucleotide variants, indels, and structural variants), higher confidence in results, improved sensitivity, and better sequencing economy.

Higher data quality and enhanced reagent stability achieved with the new chemistry comes at some cost. In direct comparisons with the KAPA HyperPlus Kit, lower library yields are typically observed for the KAPA EvoPlus Kit. It is important to note that KAPA HyperPlus post-ligation yield (or the normalized metric: conversion rate) may be inflated due to higher levels of molecular artefacts (data not shown). Nevertheless, higher inputs are recommended when using the KAPA EvoPlus Kit—particularly in PCR-free workflows, and for applications where double-sided size selection is performed (before or after library amplification) to obtain narrower fragment size distributions.

Applications

KAPA EvoPlus Kits are compatible with any NGS sample preparation workflow that requires DNA fragmentation and the addition of platform-specific adapters by ligation. This includes:

- human and microbial whole genome sequencing (WGS);
- the preparation of shotgun libraries for hybrid capture-based target enrichment (exomes and panels), or for metagenomic sequencing;
- the preparation of libraries from full-length cDNA, long amplicons, plasmids and other vectors used in synthetic and molecular biology applications; and
- low-pass sequencing for NGS-based genotyping.

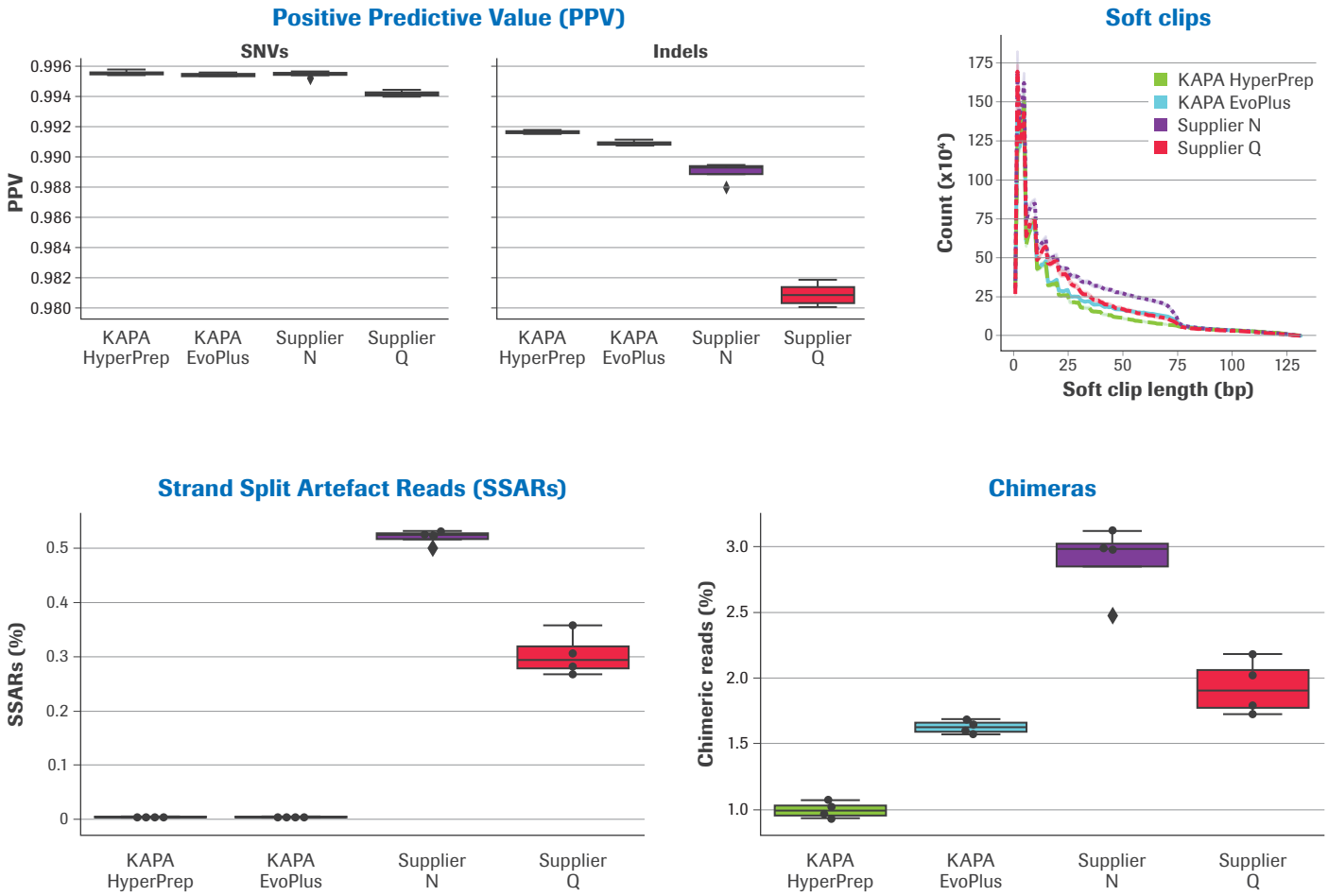

Figure 3. Sequencing artefacts observed in human WGS libraries, prepared with the KAPA EvoPlus Kit and other ligation-based library preparation kits. PCR-free libraries were prepared from 500 ng inputs of NA12878 DNA (*n*=8) with a double-sided size selection after ligation. Libraries prepared with the KAPA HyperPrep Kit (from Covaris®-sheared DNA) represent the gold standard. For all other libraries (prepared with the KAPA EvoPlus Kit and kits from Suppliers N and Q), input DNA was fragmented enzymatically. Library preparation, sequencing, and data analysis were performed as described in *Materials and methods*. With the second-generation KAPA EvoPlus enzymatic fragmentation chemistry, false positive SNV and indel calls, SSARs, and soft-clipped bases (base count and soft clip lengths) equal or approach levels observed in libraries prepared from mechanically sheared DNA, and are significantly lower than those obtained with kits from other suppliers. The levels of chimeric reads, which may originate from enzymatic fragmentation or ligation, are also greatly reduced relative to kits from Suppliers N and Q. Higher levels of chimeric reads were observed in libraries with shorter insert sizes (longer fragmentation time) and those prepared from higher inputs (longer ligation times; data not shown).

Data on file.

Human whole genome sequencing

Consistent library metrics (including fragment size distributions, yields, and concentrations) are needed to achieve the sequencing data quality requirements for large-scale human WGS projects, particularly those utilizing PCR-free workflows. The KAPA EvoPlus Kit supports robust and reproducible human WGS library construction from a wide range of inputs in a variety of commonly used storage/dilution buffers (Figures 2B – 2D). This obviates the need to fine-tune parameters for individual samples, and enables fully automated workflows. More importantly, reduced levels of sequencing artefacts (i.e., less data trimming, Figure 3) and more uniform coverage compared to kits from other suppliers (Figure 4A) improve sequencing economy, and support variant calling performance comparable to that achievable with mechanically sheared DNA (Figure 4B).

Microbial whole genome sequencing

Microbial genomes are as diverse as the environments in which they are found, ranging from a few kilobases (small viruses) to hundreds of megabases (complex fungi), and GC contents ranging from <15% to 75%. This poses a significant challenge to microbial WGS, especially when reference genomes are unavailable or of poor quality.

The KAPA EvoPlus chemistry is designed to address known sources of bias during library preparation. Improvements to enzymatic fragmentation and ligation chemistry, combined with the tried-and-tested performance of the KAPA HiFi enzyme (for workflows that require library amplification), offer uniform coverage across extreme genomes (Figure 5). This does not only improve confidence in microbial identification, but also facilitates functional and comparative genomics (Figure 6).

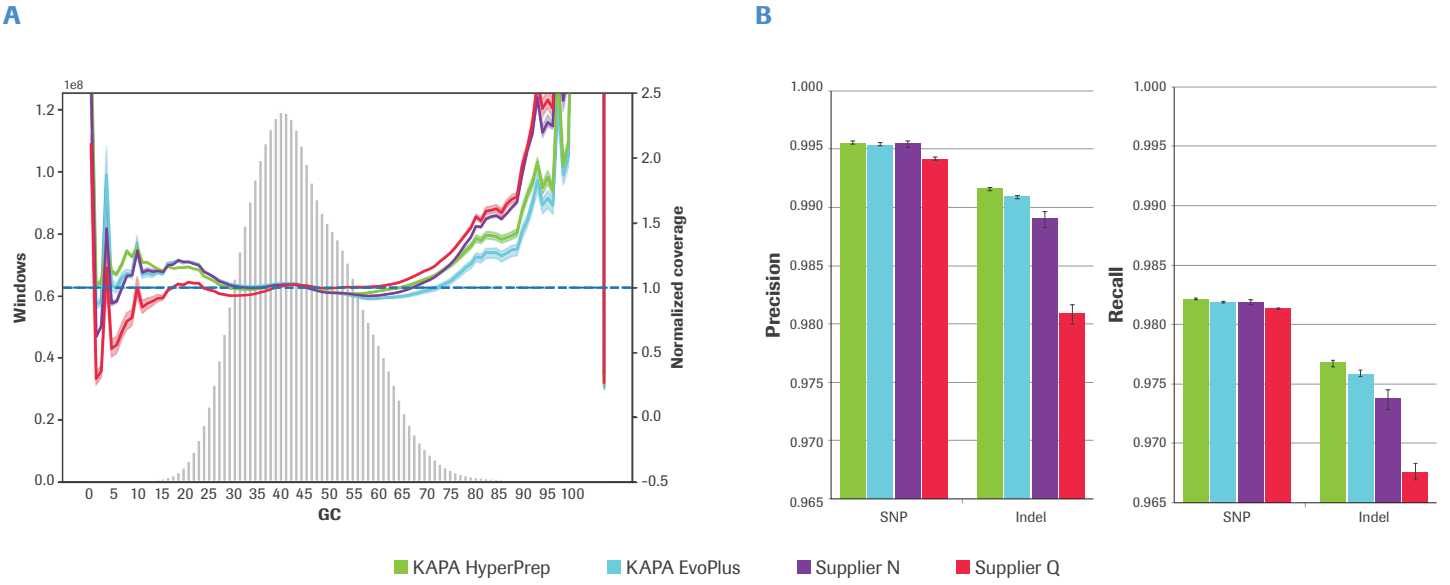

Figure 4. KAPA EvoPlus Kits match gold-standard performance in human WGS. (A) GC bias plot (generated from 42X average coverage data) for PCR-free human WGS libraries prepared from 500 ng inputs of NA12878 DNA (n=8) with the KAPA HyperPrep Kit (mechanically fragmented DNA; gold standard), the KAPA EvoPlus Kit, and two kits from other suppliers. Library preparation, sequencing, and data analysis were performed as described in *Materials and methods.* Gray histograms represent the GC content distribution of the human genome (41% GC) calculated in 100-bp bins. **(B)** Precision and recall for the same libraries, compared to the NA12878 truth data set. The evolved KAPA EvoPlus chemistry offers the best of both worlds: gold-standard performance with a fully automatable workflow, whereas data quality lag behind with enzymatic fragmentation kits from other suppliers.

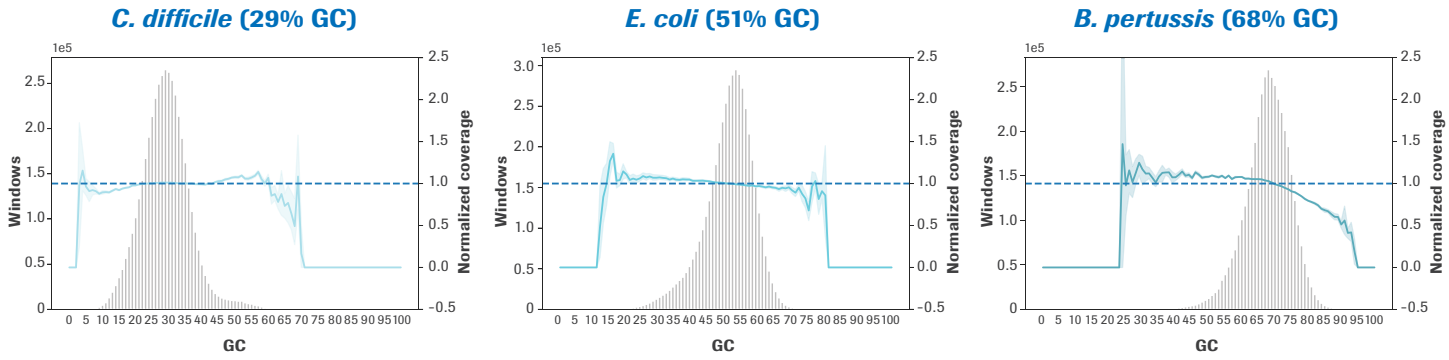

Figure 5. KAPA EvoPlus Kits supports high-performance microbial WGS. GC bias plots for whole genome libraries prepared from commercial DNA preparations of three bacterial species with diverse genome GC content. Library preparation, sequencing, and data analysis were performed as described in *Materials and methods.* Near-perfect coverage was obtained for *C. difficile* and *E. coli*, suggesting minimal bias attributed to enzymatic fragmentation, adapter ligation, or library amplification. Reduced coverage of the extremely GC-rich (>75% GC) bins of *B. pertussis* is partially attributed to inherent limitations of the sequencing technology.

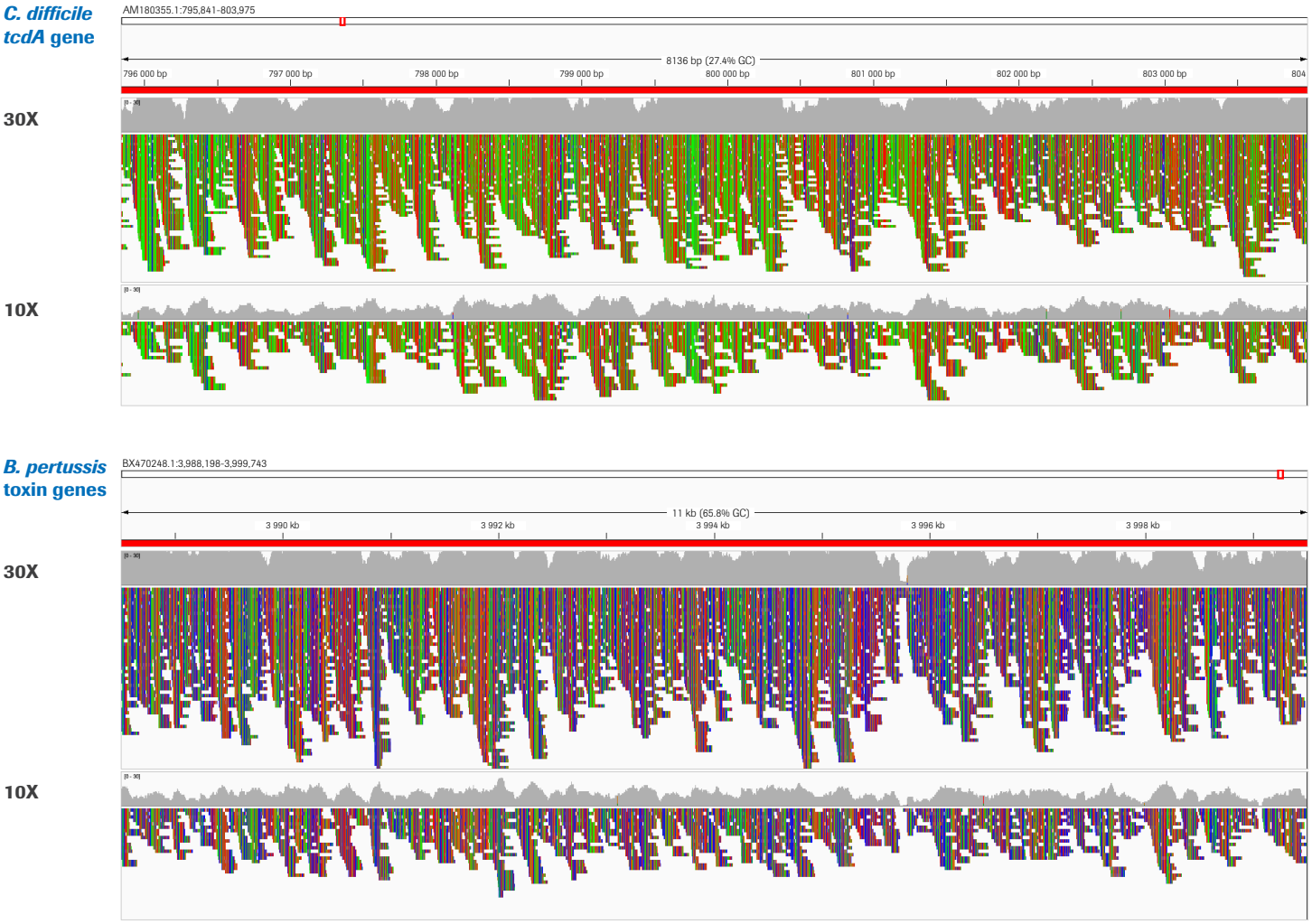

Figure 6. Uniform coverage of extreme bacterial genomes facilitates functional genetics. IGV snapshots of a 8,136-bp (27.4% GC) portion of the *C. difficile* Toxin A-encoding gene (top, GenBank protein accession number CAJ67494) and an 11-kb region (65.8% GC) of the *B. pertussis* genome encoding pertussis toxin genes (bottom, GenBank protein accession numbers CAE44038 – CAE44051), at 30X and 10X average coverage depths. Each horizontal bar represents a trimmed, aligned read; colored to reflect the nucleotide sequence (A=green, T=red, C=blue, G=yellow). Gray plots indicate coverage depth (constrained to the 0 – 30X range). At 30X average coverage, these challenging regions are covered with high uniformity. Even with downsampling to an average coverage depth of 10X, near-complete (albeit less uniform) coverage of both regions is still achieved. This facilitates annotation and functional analysis, which may be complicated or impossible when library construction inefficiencies and bias result in substantial coverage gaps.

Conclusions

NGS library preparation has come a long way since the days when less than 1% of the micrograms of DNA required for library construction was represented in the final sequencing data. KAPA EvoPlus Kits build on our proud tradition of improving library and sequence data quality, while reducing turnaround times and supporting higher throughput—to achieve predictable and desired outcomes from a larger pool of samples across ever-evolving applications.

Key features of KAPA EvoPlus Kits highlighted in this study include:

- a fast, highly streamlined, and fully automatable workflow, requiring only three steps to convert input DNA into sequencing-ready libraries;
- robust enzymatic fragmentation across variable DNA inputs and quality, even in the presence of EDTA and other inhibitors found in commonly used DNA storage/dilution buffers;
- low-bias, low-artefact library construction, which enables high-confidence variant calling and improved sequencing economy;
- automation-friendly reagent formulations and formats that require less reagent additions and less plasticware;
- improved reagent stability to streamline reagent inventory management;
- flexibility across sample types, library construction parameters (e.g., size selection and amplification), and applications; and
- compatibility with tried-and-tested KAPA accessory reagents (adapters, purification beads, and the KAPA HiFi enzyme for library amplification).

As such, KAPA EvoPlus Kits set a new standard in high-performance, simple, and efficient library preparation for short-read sequencing applications.

Materials and methods

1. DNA samples

Libraries were constructed from human or bacterial DNA, prepared and processed as described below:

- **Figure 2A, 2E, and 2F:** *Escherichia coli* genomic DNA was extracted from a liquid culture of strain K-12 using an in-house lysozyme/SDS method, and resuspended in 10 mM Tris-HCl, pH 8.0. For libraries shown in Figure 2A, EDTA was added to 100 ng DNA inputs to a final concentration of 1 – 5 mM (in 1 mM increments).
- **Figure 2B:** Human genomic DNA extracted from buffy coats (Roche PN: 11691112001) was obtained from MilliporeSigma. DNA was received in 10 mM Tris-HCl, pH 8.0 + 1 mM EDTA. Individual aliquots were subjected to a 3X bead cleanup using KAPA HyperPure Beads. Each aliquot was recovered and diluted in a different buffer (as indicated in the figure legend).
- **Figures 2C, 2D, 3, and 4:** High-quality human genomic DNA (PN: NA12878) was obtained from the Coriell Institute for Medical Research. The preparation was specified to have an 1.65 <A260/A280 <2.1 and a DNA Integrity Number (DIN) ≥7 (as determined with an Agilent 4200 TapeStation System and Genomic DNA ScreenTape assay). Concentrated DNA was received in 10 mM Tris-HCl, pH 8.5 + 1 mM EDTA, and diluted in 10 mM Tris-HCl, pH 8.0 to the appropriate concentration for library construction. The EDTA concentration in diluted DNA was negligible.
- **Figure 2D:** Human genomic DNA was extracted from fresh-frozen or FFPE tissue using the KAPA NGS DNA Extraction Kit (Roche PN: 09189823001 or 09190023001). DNA was recovered and diluted in 10 mM Tris-HCl, pH 8.0.
- **Figures 5 and 6:** High-quality, dried genomic DNA from the following bacteria were obtained from the American Type Culture Collection (ATCC): *Escherichia coli* (Migula) Castellani and Chalmers, strain MG1655 (PN: 700926D-5), *Bordetella pertussis* (Bergey, et al.) Moreno-Lopez, strain Tohama 1 (PN: BAA-589D-5), and *Clostridioides difficile* (Prevot) Lawson, et al., strain 630 (PN: BAA-1382D-5). DNA was resuspended and diluted in 10 mM Tris-HCl, pH 8.0.

DNA was quantified using a Qubit® Fluorometer and dsDNA HS Assay Kit (ThermoFisher Scientific). DNA quality was assessed in terms of the proportion of high-molecular weight DNA observed by agarose gel electrophoresis—except in the case of DNA extracted from fresh-frozen or FFPE tissue, which was assayed using the KAPA NGS FFPE QC Kit (Roche PN: 09217193001).

Where applicable, genomic DNA was mechanically sheared with a Covaris® M220 Focused Ultrasonicator, using Covaris MicroTUBES (AFA Fiber 6 x 16 mm with Pre-Slit Snap-Cap) according to the manufacturer's recommended protocols.

2. Library construction

Libraries were constructed from different DNA preparations with the KAPA EvoPlus Kit (Roche PN: 09420037001, 09420053001, 420339001, or 09420428001) or ligation-based library preparation kits with integrated enzymatic fragmentation from two other suppliers (Kits N and Q), following recommended protocols unless otherwise stated. The KAPA HyperPrep Kit (Roche PN: 07962312001, 07962347001, or 07962363001) was used to prepare libraries from Covaris-sheared DNA. Three to eight technical replicates were prepared for every library construction condition. Data shown in some figures were limited to a subset of representative replicates to facilitate interpretation.

The enzymatic fragmentation parameters used for each experiment are given in figure captions. All libraries were prepared with KAPA Unique Dual-Indexed Adapters (Roche PN: 08861919702) and KAPA HyperPure Beads (Roche PN: 08963835001, 08963843001, 08963851001, 08963878001, or 08963860001). The human WGS libraries shown in Figures 3 and 4 were subjected to a double-sided post-ligation size selection (0.5X – 0.7X for all kits, except for Kit N, which could only accommodate a 0.25X – 0.35X size selection due to protocol constraints).

Library amplification was performed with the amplification reagents included in each kit, using recommended cycling protocols and the number of amplification cycles indicated in figure captions. Aliquots of PCR-free libraries were amplified for a limited number of cycles (as indicated in figure captions) prior to electrophoretic analysis to avoid anomalous migration (and overestimated mean fragment sizes) as a result of single-stranded adapter terminals.[20]

3. Electrophoretic analysis

Library fragment size distributions were assessed with a LabChip® GX Touch® HT Nucleic Acid Analyzer (PerkinElmer) or a 2100 Bioanalyzer System (Agilent Technologies) using appropriate reagent kits according to recommended protocols. Mean fragment sizes were calculated using each system's analysis software.

4. Sequencing

Paired-end sequencing (2 x 150 bp) was performed on the Illumina® platform. Human whole genome sequencing (Figures 3 and 4) was performed on an Illumina® NovaSeq® 6000 instrument, using an S4 flow cell and NovaSeq 6000 S4 Reagent Kit v1.5 (300 cycles). Bacterial whole genome sequencing (Figures 5 and 6) was performed on an Illumina NextSeq® 500 instrument using a NextSeq 500/550 High Output Kit v2.5 (300 Cycles).

5. Bioinformatic analysis

Conversion of BCL files to FASTQ format and demultiplexing were performed with bcl2fastq (v2.19.1.403). Trimming (of adapter sequences, terminal polyG tracts (resulting from dark cycles), and primer dimers) was performed with Cutadapt (v1.18). Sections of

reads with Q<28 (as determined with FastQC v0.11.8) were also trimmed, and trimmed reads <30 bp were discarded. BWA-MEM (v0.7.17) was used for alignment and Samtools view (v1.9) for downsampling.

- **Figure 3:** reads for all libraries were downsampled to the lowest number of aligned reads (an average of approximately 928,927,000 mapped reads across all samples) in order to obtain comparable metrics.

  Positive Predictive Values (precision; True Positives / (True Positives + False Positives)) were calculated using GATK HaplotypeCaller (version 4.0.10.0). Hap.py was used to benchmark variant calls against the gold standard NA12878 truth dataset (obtained from ftp://ftp-trace.ncbi.nih.gov/giab/ftp/release/NA12878_HG001/).

  Soft clips were identified and characterized with an in-house Pysam script that uses CIGAR strings to count the number of soft clips of different lengths.

  Strand Split Artefact Reads (SSARs) were identified and characterized with an in-house Pysam script, which counts the percentage of reads with a strand-split artefact on read1, read2, or read1 and read2. The following strand split artefact read criteria were applied: (i) must be a primary mapped read, (ii) must be mapped to a single chromosome, (iii) mapping quality must be >20, (iv) must have an SA tag, (v) must map to both positive and negative strands, and (vi) mapping distance between positive and negative strand must be ≤500 bp.

  Chimeras were identified and characterized with Picard CollectMultipleMetrics (version 2.18.14).

- **Figure 4:** downsampling was performed as described for Figure 3. GC coverage plots were generated with Picard CollectGcBiasMetrics (version 2.18.14).

  Precision (True Positives / (True Positives + False Positives)) and recall (True Positives / (True Positives + False Negatives)) were calculated using GATK HaplotypeCaller (version 4.0.10.0). Hap.py was used to benchmark variant calls against the gold standard NA12878 truth dataset (obtained from ftp://ftp-trace.ncbi.nih.gov/giab/ftp/release/NA12878_HG001/).

- **Figure 5:** downsampling was performed across all samples to achieve average coverage of 30X or 10X. GC coverage plots were generated with Picard CollectGcBiasMetrics (version 2.18.14).

- **Figure 6:** alignments were visualized using IGV (version 2.12.3). Gene regions of interest were selected from feature tables.

Graphs were generated with Python 3.7.4: pandas 0.25.1, seaborn 0.9.0 and matplotlib 3.1.1.

Acknowledgments

The authors wish to thank Albert Abrie, Peter Ristow, William Haylett, Cumisa Mlandu, Etienne Slabbert, and Stephanie Cornelissen (Research and Early Development – Sample Prep Reagents, Roche Dia Cape Town, South Africa); Ruhan Slabbert, Claire Steward, Rene Goosen, Dennis Lin, Marlize Jordaan, Yolandi Snyman, Janine Blanckenberg, Olona Hlahane, Nabeela Khan, Akhona Ngcatshe, Anneke Postma, and Chris Visser (Research and Development – Sample Prep Reagents Roche Dia Cape Town, South Africa) for their contributions to the development of the KAPA EvoPlus Kit. Nishtha Jhilmeet (Complaints Investigation and Resolution, Roche Dia Cape Town, South Africa) and Yuri Yakushko, Patricia Do Dinh, and Rafael Szczepanowski (EMEA Regional Customer Support Center, Roche Diagnostics Deutschland GmbH, Mannheim, Germany) are thanked for assisting with experimental work. We also wish to thank Du Toit Schabort and Michael Berry (Bioinformatics, Roche Dia Cape Town, South Africa) for computational analysis.

References

1. Roche Sequencing Solutions. KAPA HyperPrep Brochure. https://sequencing.roche.com/content/dam/rochesequence/worldwide/resources/brochure-kapa-hyperprep-kits-SEQ100003.pdf. Accessed 27 April 2022.
2. Roche Sequencing Solutions. KAPA HyperPlus Brochure. https://sequencing.roche.com/content/dam/rochesequence/worldwide/resources/KAPA_HyperPlus_Brochure.pdf. Accessed 27 April 2022.
3. Aigran L, Gu Y, Quail, MA. Quantitation of next generation sequencing library preparation protocol efficiencies using droplet digital PCR assays – a systematic comparison of DNA library preparation kits for Illumina sequencing. *BMC Genomics* 2016;17(458). doi: 10.1186/s12864-016-2757-4.
4. Internal publications database.
5. Tanaka N, Takahara A, Hagio T, et al. Sequencing artifacts derived from a library preparation method using enzymatic fragmentation. *PLOS ONE* 2019;15(e0227427). doi: 10.1371/journal.pone.0227427.
6. Gregory T, Ngankeu A, Orwick S, et al. Characterization and mitigation of fragmentation enzyme-induced dual stranded artifacts. *NAR Genomics and Bioinformatics* 2020;2(lqaa070). doi: 10.1093/nargab/lqaa070.
7. Haile S, Corbett RD, Bilobram S, et al. Sources of erroneous sequences and artifact chimeric reads in next generation sequencing of genomic DNA from formalin-fixed paraffin-embedded samples. *Nucleic Acids Res.* 2019;47(e12). doi: 10.1093/nar/gky1142.
8. Costello M, Pugh TJ, Fennell TJ, et al. Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic Acids Res.* 2013;41(e67). doi: 10.1093/nar/gks1443.
9. Kia A, Gloeckner C, Osothprarop T, et al. Improved genome sequencing using an engineered transposase. *BMC Biotechnol.* 2017;17(6). doi: 10.1186/s12896-016-0326-1.
10. Seguin-Orlando A, Schubert M, Clary J, et al. Ligation bias in Illumina next-generation DNA libraries: implications for sequencing ancient genomes. *PLOS ONE* 2013;8(e78575). doi: 10.1371/journal.pone.0078575.
11. Quail M, Otto T, Gu Y, et al. Optimal enzymes for amplifying sequencing libraries. *Nat. Methods* 2012;9(10). doi: 10.1038/nmeth.1814.
12. Aird D, Ross MG, Chen W-S, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* 2011; 12(R18). doi: 10.1186/gb-2011-12-2-r18.
13. Costello M, Fleharty M, Abreu J, et al. Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms. *BMC Genomics* 2018;19(332). doi: 10.1186/s12864-018-4703-0.
14. Illumina, Inc. https://www.illumina.com/content/dam/illumina-marketing/documents/products/whitepapers/index-hopping-white-paper-770-2017-004.pdf?linkId=36607862. Effects of Index Misassignment on Multiplexing and Downstream Analysis. Accessed 27 April 2022.
15. Ribarska T, Bjørnstad PM, Sundaram AYM, et al. Optimization of enzymatic fragmentation is crucial to maximize genome coverage: a comparison of library preparation methods for Illumina sequencing. *BMC Genomics* 2022;23(92). doi: 10.1186/s12864-022-08316-y.
16. MacArthur D, Manolio T, Dimmock D, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature* 2014; 508(469). doi: 10.1038/nature13127.
17. Suzuki S, Yasuda T, Shiraishi Y, et al. ClipCrop: a tool for detecting structural variations with single-base resolution using soft-clipping information. *BMC Bioinformatics* 2011;12(S7). doi: 10.1186/1471-2105-12-S14-S7.
18. Landman SR, Hwang TH, Silverstein KAT, et al. SHEAR: sample heterogeneity estimation and assembly by reference. *BMC Genomics* 2014; 15(84). doi: 10.1186/1471-2164-15-84.
19. Schloss PD, Gevers D, and Westcott SL. Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *PLOS ONE* 2011;6(e27310). doi: 10.1371/journal.pone.0027310.
20. Meyer J, Whitehorn H, Appel M, et al. https://sequencing.roche.com/content/dam/rochesequence/worldwide/resources/applicationnote-human-whole-genome-sequencing-SEQ100220.pdf. KAPA HyperPrep Kits offer a flexible, high-efficiency library preparation solution for PCR-free human whole-genome sequencing. Roche Application Note, SEQ100183. Accessed 27 April 2022.

Published by:

Roche Sequencing Solutions, Inc.
4300 Hacienda Drive
Pleasanton, CA 94588

sequencing.roche.com

MC--09191 05/2022