NOVEL IMPROVEMENTS TO THE ILLUMINA TRUSEQ INDEXED LIBRARY CONSTRUCTION, AMPLIFICATION AND QUANTIFICATION PROTOCOLS FOR OPTIMIZED MULTIPLEXED SEQUENCING

GAVIN J. RUSH[1], ERIC VAN DER WALT[1], JACOB KITZMAN[2], LIESL NOACH[1], ZAYED ALBERTYN[3], COLIN HERCUS[3], CHARLIE LEE[2], JAY SHENDURE[2], JOHN F. FOSKETT III[1], PAUL J. MCEWAN[1]

1 KAPA BIOSYSTEMS, 600 WEST CUMMINGS PARK, SUITE 2250, WOBURN, MA 01801 | 2 UNIVERSITY OF WASHINGTON, DEPARTMENT OF GENOME SCIENCES, FOEGE BUILDING S-250, BOX 355065, 3720 15TH AVE NE , SEATTLE WA 98195 | 3 NOVOCRAFT, C-23A-5, TWO SQUARE, SECTION 19, 46300 PETALING JAYA, SELANGOR, MALAYSIA

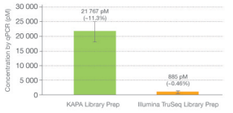
SAMPLE PREPARATION

INTRODUCTION

Dramatic improvements in commercial Next Generation Sequencing (NGS) platforms have resulted in spectacular reductions in the cost-per-base of DNA sequencing. Until recently, the primary focus for innovation has been on improvements to the core sequencing technologies, with optimization of sample preparation playing a secondary role. The exponential gains in sequencing capacity have simultaneously led to higher sample throughput, placing increasing emphasis on the importance of improved library construction protocols for multiplexed sample sequencing. While major commercial NGS systems all require the construction of similar libraries via analogous workflows, some protocols and/or reagents offer significant advantages over others, and end-users must choose among numerous alternative methods and reagents for sample preparation. We re-sequenced the *Staphylococcus aureus*, *Escherichia coli*, and *Mycobacterium tuberculosis* genomes to compare the standard Illumina TruSeq sample preparation reagents and workflow with a number of innovative improvements including: alternative library preparation reagents and protocols; automated fragment size selection; real-time library amplification; amplification-free sequencing; and accurate qPCR library quantification for sample pooling and multiplexed sequencing.

SPECIES	GENBANK ID	GENOME SIZE (bp)	% GC
S. aureus subsp. *aureus* TW20	FN433596	3,043,210	33%
E. coli DH10B	NC_010473	4,686,137	51%
M. tuberculosis H37Rv	AL123456	4,411,532	65%

Sheared gDNA by nebulizing to generate fragments 200 – 500 bp in length.

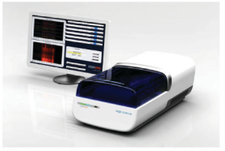


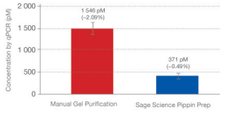
Yield of adaptor-ligated library molecules before size selection.

Sheared genomic DNA was prepared in bulk by nebulization and identical starting material was used for each library (1 µg of gDNA from *S. aureus*, *E. coli*, or *M. tuberculosis*). Libraries were constructed using reagents and procedures (see detailed methods on left-hand side) from KAPA Biosystems (green; 9 libraries), or using the Illumina TruSeq DNA Sample Prep Kit and Low Throughput Protocol (orange; 15 libraries). Libraries were quantified by qPCR before size selection using the KAPA Biosystems Library Quantification Kit according to the recommended protocol. The estimated percentage of starting material that was converted to useful, adaptor-ligated (PCR-amplifiable) library molecules is provided.

The KAPA Biosystems library construction reagents and protocol produced ~25-fold more adaptor-ligated library fragments from the same amount of starting material than did the standard Illumina TruSeq DNA Sample Prep Kit. Important differences between the two library construction procedures assessed here include: Combined enzyme and reaction buffer "master mix" formulations (TruSeq) vs. separate enzyme and reaction buffer (KAPA); and the absence of an Ampure XP Bead clean-up between A-Tailing and Adaptor Ligation in the TruSeq Protocol.

MANUAL GEL EXTRACTION	AUTOMATED SIZE SELECTION
Requires manual gel excision and downstream extraction of the sample from the agarose gel.	Automated elution of sample material into buffer compatible with downstream workflows (PCR amplification, flow cell amplification).
Actual DNA fragment length distribution is broader and less reproducible.	Selects tight, accurate, and reproducible DNA fragment size ranges.
Manual gel electrophoresis requires longer run time and more manual steps.	Significant time and labor savings from quicker run times and automated elution.
Contamination of adaptor-dimers and other low molecular weight artifacts leads to lower data quality.	Reduction of low molecular weight contamination reduces library artifacts.
Potential for cross-contamination amongst samples on a single gel.	Reduced potential for cross-contamination amongst samples.

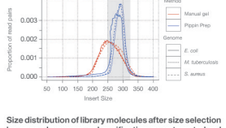




Yield of adaptor-ligated library molecules after size selection.

Starting with 1 µg of either *S. aureus*, *E. coli*, or *M. tuberculosis* gDNA, we constructed duplicate libraries using reagents and procedures from the KAPA DNA Library Prep Kit (see detailed methods on left-hand side). Duplicate libraries were pooled, and the three resulting libraries (*S. aureus*, *E. coli*, or *M. tuberculosis*) were each split equally before size selection either by manual gel purification (Qiagen QIAquick Gel Extraction Kit) or by Pippin Prep automated gel purification (Sage Science). Libraries were quantified by qPCR using the KAPA Library Quantification Kit according to the recommended protocol. The estimated percentage of starting material converted to useful, adaptor-ligated (PCR-amplifiable) library molecules is provided.

Size selection by manual gel purification yielded ~4-fold more DNA than size selection by automated gel purification using the Sage Science Pippin Prep apparatus. To a large extent this may be accounted for by the fact that the Pippin Prep method yielded a much narrower size-range of library fragments (see below), despite careful manual gel purification targeting an equivalent size range of library molecules.



Size distribution of library molecules after size selection by manual agarose gel purification or automated gel purification with Sage Science Pippin Prep.

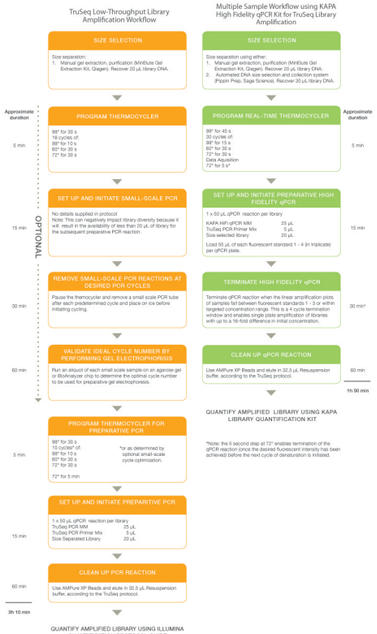
Starting with 1 µg of either *S. aureus*, *E. coli*, or *M. tuberculosis* sheared gDNA, we constructed duplicate libraries using reagents and procedures from the KAPA DNA Library Prep Kit (see detailed methods on left-hand side). Duplicate libraries were pooled, and the three resulting libraries (*S. aureus*, *E. coli*, or *M. tuberculosis*) were each split equally before size selection either by manual gel excision and purification (Qiagen QIAquick Gel Extraction Kit) or by Pippin Prep automated gel purification (Sage Science). We set the Pippin Prep instrument to collect library molecules in the 370 bp – 450 bp range including adaptors, and we attempted to select the same range via manual agarose gel purification. TruSeq adaptors are 121 bp in total, suggesting targeted insert sizes were ~250 bp – 330 bp (broken lines). Paired reads (2 x 75 bp) were used to determine actual insert sizes.

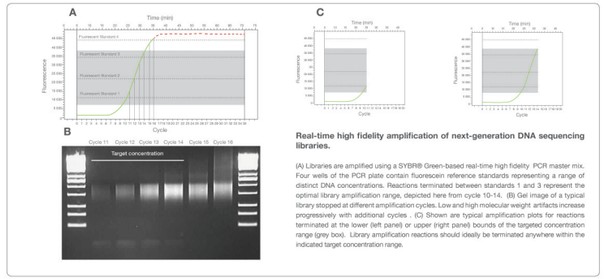
AMPLIFICATION

RATIONALE FOR REAL-TIME HIGH FIDELITY AMPLIFICATION OF NEXT-GENERATION DNA SEQUENCING LIBRARIES

High fidelity PCR is used to selectively enrich library fragments carrying appropriate adaptor sequences and to amplify the amount of DNA prior to sequencing. During PCR enrichment of libraries, a subset of library molecules are amplified with reduced efficiencies, introducing bias and resulting in uneven sequence coverage. GC content is known to be an important factor in NGS library amplification bias, and different PCR enzymes and buffer formulations are likely to show individual strengths and weaknesses in this regard. Furthermore, such biases – along with other artifacts such as PCR-induced errors, adaptor dimers, PCR duplicates, and chimeras – are exacerbated by over-amplification, while under-amplification results in insufficient yields. Inherent uncertainty in the outcome of end-point PCR often demands downstream validation of library quality by electrophoresis.

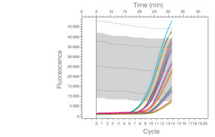
Here we present a high fidelity, real-time PCR method for rapid and convenient enrichment and amplification of libraries. The benefits of this approach include: 1) automatable workflows, 2) built-in quality metrics for each enriched library, eliminating expensive and time-consuming post-enrichment gel electrophoresis, 3) precise control over the number of PCR cycles required for optimal amplification, 4) a quality control metric for identifying inconsistencies in library preparation, and 5) seamless integration with KAPA qPCR Library Quantification kits.





Real-time high fidelity amplification of next-generation DNA sequencing libraries.

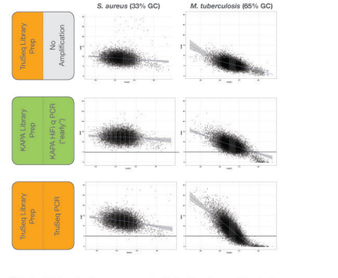
(A) Libraries are amplified using a SYBR® Green-based real-time high fidelity PCR master mix. Four wells of the PCR plate contain fluorescein reference standards representing a range of distinct DNA concentrations. Reactions terminated between standards 1 and 3 represent the optimal library amplification range, depicted here from cycle 10-14. (B) Gel image of a typical library stopped at different amplification cycles. Low and high molecular weight artifacts increase progressively with additional cycles . (C) Shown are typical amplification plots for reactions terminated at the lower (left panel) or upper (right panel) bounds of the targeted concentration range (grey box). Library amplification reactions should ideally be terminated anywhere within the indicated target concentration range.



Example of multiplexed real-time high fidelity amplification.

20 libraries, spanning a ~64-fold concentration range (6 cycles), were simultaneously amplified and terminated after 14 cycles. 14 of the 20 libraries fall within the targeted amplification range. The remaining 6 libraries could either be used as is, noting that they may be outside the optimal concentration range, or they could be re-amplified individually or in high- or low-concentration groups.

DISADVANTAGES ASSOCIATED WITH THE EXISTING TRUSEQ WORKFLOW		ADVANTAGES OF HIGH FIDELITY REAL-TIME PCR
Cycle number optimization	Fixed cycle number amplification	High fidelity real-time PCR
Wastage of library material required for cycle number optimization leads to loss of library diversity. Inconsistencies in reaction volume scale-up can cause variable results. Longer protocol involves more time and labor. Gel electrophoresis steps are not amenable to automation.	Variability in the nature and/or concentration of the input DNA can result in under- or over-amplification exacerbating amplification bias and artifacts. Post-amplification gel electrophoresis QC is not amenable to automation. Ideally, libraries should be minimally amplified to a low final concentration (~3.5 µM), not detectable via standard gel electrophoresis.	Built-in real-time quality metrics (concentration range) for each amplified DNA library. Real-time PCR is amenable to automation. Precise control over PCR cycle number required for optimal amplification. Seamless integration with qPCR-based library quantification. KAPA HiFi DNA Polymerase is less prone to amplification bias due to high- or low-GC content.



Effect of GC content on coverage depth for libraries amplified using KAPA HiFi qPCR Master Mix or Illumina TruSeq PCR Master Mix.

Indexed libraries were prepared from identical sheared *S. aureus* (33% GC; left panels) and *M. tuberculosis* (65% GC; right panels) gDNA using either the KAPA DNA Library Prep Kit, or the Illumina TruSeq DNA Sample Prep Kit, and then amplified using the indicated PCR reagents before paired-end sequencing (2 x 75 bp). After filtering and aligning read pairs to reference sequences, 250 000 read pairs were randomly sampled for each genome, and scatter plots of mean sequence coverage depth vs. GC content were generated by analyzing 250 bp windows.

For the AT-rich *S. aureus* genome (left panel), none of the samples showed gross amplification bias compared to the unamplified control sample. GC-rich *M. tuberculosis* sequences (right panel) in the library constructed and amplified using Illumina TruSeq reagents are under-represented in the sequencing data. In contrast, the library prepared using KAPA Biosystems reagents yielded coverage across the range of GC-content that is almost indistinguishable from that of the unamplified control, indicating that amplification with KAPA HiFi qPCR Master Mix introduced minimal additional GC-dependent coverage bias.

Percentage of the *M. tuberculosis* genome not represented at all (top), or at ≥5x coverage (bottom) when using various library construction and PCR reagents.

Indexed libraries were prepared from identical sheared *M. tuberculosis* (65% GC) gDNA using either the KAPA DNA Library Prep Kit, or the Illumina TruSeq DNA Sample Prep Kit, and then amplified using the indicated PCR reagents before paired-end sequencing (2 x 75 bp). Libraries were quantified before and after amplification using the KAPA Library Quantification Kit to determine the number of doublings in each case. After filtering and aligning read pairs to reference sequences, 250 000 read pairs (~8.5x coverage) were randomly sampled for each genome.

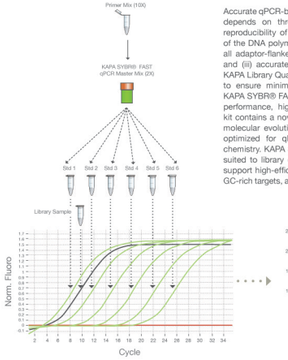
LIBRARY QUANTIFICATION

KAPA LIBRARY QUANTIFICATION KIT FOR OPTIMAL SAMPLE MULTIPLEXING

Accurate quantification of the number of amplifiable molecules in a library is critical to the outcome of sequencing results on the Illumina Genome Analyzer next-generation sequencing platform. Overestimation of library concentration results in lower cluster density after bridge PCR. Underestimation of library concentration results in too many clusters on the flow cell, which can lead to poor cluster resolution. Both scenarios result in suboptimal sequencing capacity. qPCR is widely regarded as the gold standard for accurate quantification of DNA libraries as it is the only technique capable of measuring the number of amplifiable molecules. The broad dynamic range of qPCR also enables accurate quantification of extremely dilute libraries.

Here we demonstrate the use of the system to:

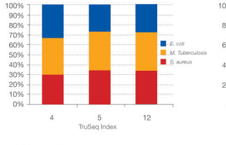
1. Generate Illumina TruSeq indexed libraries comprising equimolar concentrations of 3 bacterial genomes spanning a range of average GC contents (*S. aureus*, 33% GC; *E. coli*, 51% GC; and *M. tuberculosis*, 65% GC).
2. Pool multiple Illumina TruSeq indexed libraries in equimolar concentrations for equal representation in sequence data.



Accurate qPCR-based library quantification ultimately depends on three factors: (i) the accuracy and reproducibility of the standards used, (ii) the ability of the DNA polymerase used in the qPCR to amplify all adaptor-flanked molecules with equal efficiency, and (iii) accurate and reproducible liquid handling. KAPA Library Quantification Kits are rigorously tested to ensure minimal lot-to-lot variation. In addition, KAPA SYBR® FAST qPCR Kits are designed for high performance, high-throughput, real-time PCR. The kit contains a novel DNA polymerase engineered via molecular evolution, resulting in an unique enzyme optimized for qPCR using SYBR® Green I dye chemistry. KAPA SYBR® FAST qPCR Kits are ideally suited to library quantification applications, as they support high-efficiency amplification of both AT- and GC-rich targets, and of fragments up to 1 kb in length.

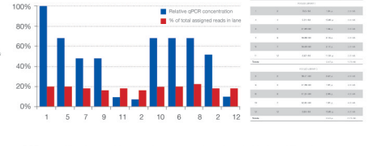
KAPA Library Quantification Kit Workflow.

KAPA Library Quantification Kits comprise DNA Standards (six 10-fold dilutions) and 10X Primer Premix, paired with KAPA SYBR® FAST qPCR Kits to accurately quantify the number of amplifiable molecules in an Illumina GA library. The 452 bp KAPA Illumina GA DNA Standard consists of a linear DNA fragment flanked by qPCR primer binding sites. Quantification is achieved by inference from a standard curve generated using the six DNA Standards.



qPCR quantification enables equal representation of different sample types within indexed libraries.

For each index (TruSeq 4, 5, and 12) we constructed three separate libraries (*S. aureus*, 33% GC; *E. coli*, 51% GC; and *M. tuberculosis*, 65% GC). Individual libraries were quantified using the KAPA Library Quantification Kit and for each index the libraries were pooled to achieve equimolar representation for each genome. The results indicate that quantification is reliable for samples with a wide range of GC content.



qPCR quantification enables equal representation of pooled indexed libraries.

Eleven indexed Illumina TruSeq libraries were quantified by qPCR using the KAPA Library Quantification Kit according to the recommended protocol, and then combined to achieve equal final concentrations in two separate pools for multiplexed sequencing on different flow-cell lanes. The eleven libraries ranged ~11-fold in concentration from 0.67 pM to 7.65 pM, while representation of each index varied between 90% and 127% of expected assigned reads per lane.