Application Note

Human whole-genome sequencing

Authors

Jacqueline Meyer
Support & Training Senior Specialist
Global Customer Support

Heather Whitehorn
Support & Training Team Leader
Global Customer Support

Maryke Appel
Sr. International Product Manager

Roche Sequencing Solutions
Cape Town, South Africa and
Pleasanton, CA, USA

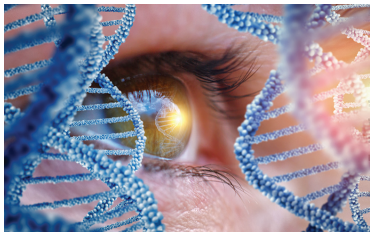
Jennifer Pavlica
Applications Team Manager

Roche Sequencing & Life Science
Wilmington, MA, USA

KAPA HyperPrep Kits offer a flexible, high-efficiency library preparation solution for PCR-free human whole-genome sequencing

Routine human whole-genome sequencing (WGS) requires robust and streamlined PCR-free library preparation protocols that can be tailored to ensure optimal sequencing results on production-scale Illumina® sequencers. KAPA HyperPrep Kits, combined with KAPA Dual-Indexed Adapters, KAPA Pure Beads and KAPA Library Quantification Kits, provide a complete sample prep solution for efficient human WGS on Illumina HiSeq X® and NovaSeq™ instruments.

Introduction



HiSeq X and NovaSeq 6000 sequencers from Illumina utilize sequencing technology improvements that enable significant reductions in per-base cost. This has stimulated global investment in population-based human whole-genome sequencing (WGS); for the discovery of new biomarkers and drug targets, and to advance our understanding of human diseases, fitness and longevity.

Broad-based access to, and routine application of next-generation sequencing (NGS) continue to drive the evolution of the other two components of the sequencing value chain, namely sample preparation and data analysis/reporting. Providers of sample preparation solutions have been focusing on streamlining library construction methods to facilitate automation, reduce turnaround time, and improve reproducibility. In addition, chemistries have been optimized to achieve higher conversion of input DNA to adapter-ligated library fragments. This enables higher success rates with PCR-free protocols, from lower inputs and samples of variable quality.

KAPA HyperPrep Kits offer a very efficient, automation-ready, single-tube library construction protocol. Extensive chemistry and protocol optimization has enabled high conversion rates, thereby expanding the pool of samples that can be successfully processed for a variety of sequencing applications.[1-5] The flexible protocol can be fine-tuned to optimize performance with specific sample types or cohorts, or to meet operational requirements. A suite of accessory and complementary products, such as KAPA Dual-Indexed Adapters, KAPA Pure Beads, KAPA HiFi Library Amplification Kits and KAPA Library Quantification Kits, rounds out the sample preparation workflow.

In this Application Note, we demonstrate the benefits of our complete sample prep solution for human WGS, which include higher library construction efficiency, greater flexibility to support high-throughput pipelines, and the convenience of service and support from a single supplier. Important experimental considerations are discussed, and two strategies for the preparation of high-quality, PCR-free human WGS libraries are outlined. Comprehensive data for libraries prepared from a HapMap sample (NA12878, 500 ng input), generated on both the HiSeq X and NovaSeq 6000 instruments, are presented.

Important experimental considerations

DNA fragmentation

Despite recent advances in non-mechanical fragmentation methods (e.g., tagmentation and enzymatic fragmentation), Covaris® shearing is routinely used for human WGS library construction. Nevertheless, shearing protocols specifically optimized for human WGS are rarely included in library preparation reagent instruction manuals. Even when Covaris shearing parameters are provided, little to no information is given about the potential impact of external factors such as shearing volume (Figure 1), the temperature of the water bath, and the concentration and/or viscosity of the input DNA. Suboptimal shearing can be corrected with size selection (see below), but the preferred approach is to empirically test (and optimize) shearing protocols before precious samples are processed.

In our experience, fragmentation in a 130 µL volume yields more reproducible results (Figure 1), and is recommended when input DNA is not limited. This provides the opportunity to perform post-fragmentation size selection prior to library construction. Size-selected DNA should be re-quantified, and diluted to the appropriate concentration and volume required for the first step in the library construction process. If an excess of input DNA is not available, it is best to fragment in a 50 µL volume, and to carefully recover and directly transfer the fragmented material to the end repair/A-tailing reaction. In this case, size selection may be performed after the post-ligation cleanup. The flexible KAPA HyperPrep protocol allows for both approaches, which were compared in this study.

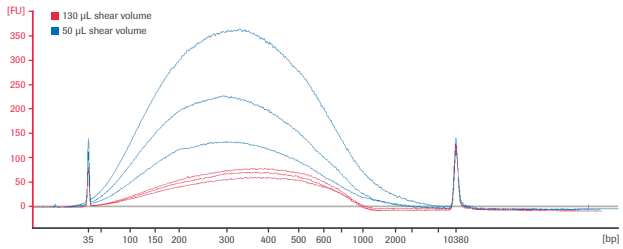


Figure 1. Shearing volume impacts the outcome of fragmentation. Triplicate 500 ng aliquots of NA12878 human genomic DNA (Coriell Institute) were diluted in 10 mM Tris-HCl (pH 8.0), to a final volume of either 130 µL (red curves) or 50 µL (blue curves). The total volume of each DNA sample was transferred to a Covaris MicroTUBE (AFA Fiber 6 x 16 mm with Pre-Slit Snap-Cap). DNA was fragmented with a Covaris E220 instrument, using parameters previously optimized for a mean peak size of 350 bp. Shearing in the smaller volume resulted in a more variable mean peak size, which was ~50 bp shorter than expected. Recovery from the larger shearing volume was more reproducible.

Size selection

Library insert size requirements vary widely for different NGS applications. For sequencing on HiSeq X® and NovaSeq™ instruments, narrow insert size distributions (in the range of 300 – 650 bp), and sequencing-ready libraries free of short fragments, unligated adapter and adapter-dimer are required. This is essential to ensure optimal cluster generation, mitigate the potential impact of index misassignment,[6] and facilitate data analysis.

Bead-based reagents are commonly used for size selection in NGS library preparation. “Dual size selection” or “double-sided cleanups” consist of a first and second “cut”, performed with different bead-to-sample volume ratios. The first ratio determines the upper size limit of the size-selected DNA, whereas the second determines the lower size limit. While size selection results in a much narrower final library size distribution, it comes at the cost of a significant amount of DNA. This can have a profound impact on library yield and complexity, particularly if the size distribution of sheared DNA does not correspond well to the desired insert size distribution of the final library.

Because Covaris shearing yields a relatively broad size distribution around mode fragment lengths of 300 – 400 bp, size selection is inevitable when preparing libraries for human WGS. Tunable size selection (Figure 2) may be performed with KAPA Pure Beads (or a similar product, e.g., AMPure® XP reagent from Beckman Coulter).

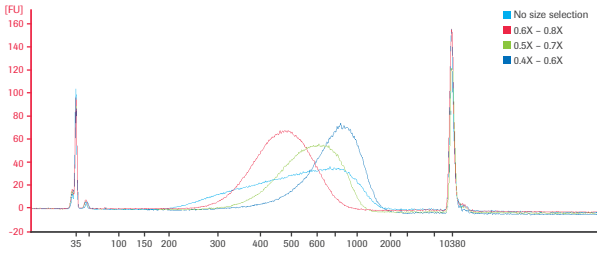


Figure 2. Size selection can be tuned to achieve the desired insert size distribution. Libraries were prepared with the KAPA HyperPrep Kit, using the PCR-free HyperPrep PL protocol described in Materials and methods. The turquoise curve corresponds to a library prepared without size selection. The final library fragment size distribution can be modified by employing different parameters for the post-ligation size selection, e.g. 0.6X – 0.8X (red curve); 0.5X – 0.7X (green curve), or 0.4X – 0.6X (blue curve). Areas under the curves are not indicative of final library yields and concentrations, as an aliquot of each library was amplified to enable accurate insert size determination.

The KAPA HyperPrep Kit supports both post-fragmentation and post-ligation size selection for PCR-free workflows. Each strategy has potential advantages and disadvantages. When input DNA is size selected after fragmentation, library construction is performed with fragments that are already free of very short fragments, and have a narrower size distribution. However, experience has shown the recovery of size-selected DNA after fragmentation to often be less efficient as compared to later in the protocol. Additionally, two post-ligation cleanups may be required to effectively exclude adapter-dimers (which cluster efficiently), and unused adapter (to mitigate index misassignment). Post-ligation size selection obviates the need for a second post-ligation cleanup step. The library insert size distribution can be optimized, and unwanted adapter species reduced to desired levels by using a combination of one “single-sided” and one “double-sided” post-ligation cleanup. This can shorten the overall library preparation time by 20 – 30 min, and results in higher final library yields (see **Results**). Post-ligation size selection is therefore recommended if an excess of input DNA is not available, or if there is a requirement to reduce the DNA input for a PCR-free workflow.

Adapters

KAPA Dual-Indexed Adapters have the same design as TruSeq® DNA PCR-Free HT dual-indexed adapters (different 8-nt sequencing barcodes on the P5 and P7 adapter oligos; eight unique P5 indices x 12 unique P7 indices = 96 combinations). They are compatible with HiSeq X and NovaSeq instruments, and are functionally tested in an Illumina® library prep workflow to confirm high library construction efficiency. In addition, each lot of KAPA Dual-Indexed Adapters is assayed for barcode

cross-contamination. Typically, no barcoded adapter oligo is contaminated with more than 0.01% of any other barcoded oligo.

If human WGS libraries are pooled for sequencing on either of the abovementioned instruments (which both employ exclusion amplification or "ExAmp" cluster generation chemistry on patterned flow cells), best practices should be employed to mitigate the potential impact of index misassignment (index "hopping"). These include efficient removal of free, unligated adapters from library preps (see previous section); not storing library pools for extended periods prior to sequencing; employing unique index combinations (i.e., not using any i5 or i7 index more than once in a pool); and employing both the P5 and P7 indices for sequencing and demultiplexing.[6-8]

To PCR or not to PCR?

The human genome contains elements that are notoriously difficult to amplify and sequence. These include repetitive sequences, regions of extreme GC content (<25% and >75%), and low-complexity regions. PCR-free library prep has become the gold standard for large-scale human WGS projects, as it eliminates an important source of amplification-associated bias, and results in improved coverage uniformity and higher overall coverage depth.[9]

Given the requirement for size selection, PCR-free library prep requires higher inputs, and very efficient conversion of input DNA to adapter-ligated molecules. We selected 500 ng as the input for this study (as this is in the range typically used in real-life pipelines). We have, however, previously shown that PCR-free libraries with a final concentration in the range of 2 – 5 nM can be prepared from as little as 100 ng of high-quality human gDNA using the KAPA HyperPrep Kit.[10]

Library quantification

Accurate quantification of NGS libraries is essential to ensure that (i) libraries are accurately normalized and pooled for multiplexed sequencing, and (ii) that individual libraries or library pools are diluted to the optimal concentration for cluster generation. Standard library quantification methods have a number of disadvantages, particularly when used to quantify libraries produced in PCR-free workflows that do not include an enrichment step for sequencing-competent molecules. Most notably, fluorometry (employed in Qubit™/PicoGreen® assays), spectrophotometry (on which the Nanodrop™ instrument is based) and electrophoretic assays (e.g., those performed using an Agilent Bioanalyzer or TapeStation) measure total nucleic acid concentrations. In contrast, qPCR is inherently well-suited for NGS library quantification, as it measures only those library fragments that can serve as templates during cluster generation. A comprehensive discussion of the over- or under-quantification of libraries with non-qPCR based methods falls outside the scope of this Application Note, but may be found elsewhere.[11-12] Moreover, because qPCR is extremely sensitive, it allows for the quantification of dilute libraries and consumes very small amounts of library.

KAPA Library Quantification Kits provide a complete solution (a pre-diluted set of DNA standards, KAPA SYBR® FAST qPCR Master Mix and quantification primers) for the absolute, qPCR-based quantification of PCR-free human WGS libraries.

One objection to the use of SYBR Green-I based qPCR assays for NGS library quantification is that the average fragment length is needed for library concentration calculations. In PCR-free workflows it is difficult to obtain accurate average fragment sizes from electrophoretic systems, as molecules flanked by adapters with long single-stranded terminals migrate anomalously in gel matrices, thereby appearing to be longer than they truly are (Figure 3). Easy workarounds for this problem include the following:

- Use the average length of the fragmented DNA plus the total length of the two adapters (usually ~120 bp) as an estimate for the average library fragment size in concentration calculations. This approach is only feasible if the size selection parameters were optimized to preserve the size distribution of the fragmented DNA.
- Amplify a small aliquot of the PCR-free library for 2 – 5 cycles prior to electrophoretic analysis. Amplification will render all molecules fully double-stranded, and yield a reliable size determination from the electrophoretic assay.
- Subject the product of the library **quantification** reaction to electrophoretic analysis. The library quantification reaction is performed for 35 cycles, and contains artifacts resulting from reagent depletion toward the end of the assay. Nevertheless, the mode fragment size of the qPCR product provides a better approximation of the average library fragment size than an unamplified library (Figure 3). Since systematic under- or over-estimation of library concentration is likely when using this approach, it is important to remember that the relationship between calculated library concentration and cluster density has to be determined empirically for each specific library prep workflow and sequencing system.[11-12]

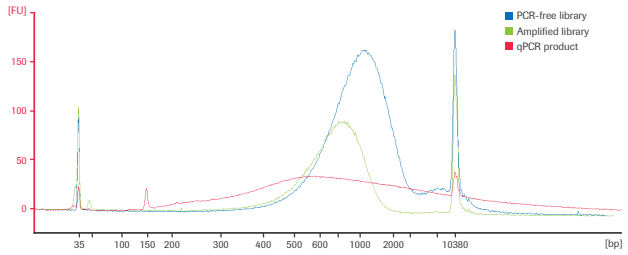


Figure 3. Methods for determining the true average fragment size distribution of human WGS libraries produced in PCR-free workflows, for qPCR-based library concentration calculations. Libraries were prepared from DNA sheared to a mode fragment length of ~650 bp with the KAPA HyperPrep Kit (post-ligation size selection workflow, as described in Materials and methods). Unamplified libraries (blue curve) have a significantly longer apparent average fragment length, due to the anomalous migration of inserts flanked by adapters with long single-stranded terminals. Amplification (for 5 cycles) of a small aliquot (5 µL) of a PCR-free library results in reliable fragment length determination (green curve). The product of the library quantification assay (red curve) may provide a reasonable estimate if it is not feasible to amplify a portion of the library for the purpose of analysis.

A final benefit of the KAPA Library Quantification assay is that it offers a sensitive means for the detection of adapter-dimer or library fragments with very short insert sizes (which cluster preferentially). Residual levels of adapter-dimer can be assessed by including a standard melt curve analysis at the end of the assay. Please refer to the **KAPA Library Quantification Kit (Illumina Platforms) Technical Data Sheet**[11] for details and an example. Should the levels of adapter-dimer in individual libraries be of concern, an additional 0.8X bead cleanup should be performed before proceeding to sequencing.

Materials and methods

Experimental design

Whole human shotgun libraries were prepared from a characterized HapMap sample (NA12878; Coriell Institute of Biomedical Research), using two library construction kits (Table 1 and Figure 4 on p. 5), namely the KAPA HyperPrep Kit (Roche) and the TruSeq® DNA PCR-Free Library Prep Kit (Illumina®). With the KAPA HyperPrep Kit, two protocols—one employing post-fragmentation size selection (PF), and one including size selection after the post-ligation cleanup (PL)—were used. Detailed, step-by-step protocols may be found in a separate Tech Note.[13] TruSeq libraries were prepared according to the manufacturer's recommended protocol. Replicate libraries from each workflow were pooled for sequencing on an Illumina HiSeq® X instrument, as well as a NovaSeq™ 6000 instrument with an S2 flow cell. The three library construction methods were compared with respect to key library construction, sequencing, assembly and variant calling metrics.

Library construction

DNA shearing. To facilitate downstream analysis and data comparison, the shearing parameters provided in the TruSeq Protocol Guide (optimized for a median fragment length of 350 bp) were used for all three workflows. NA12878 DNA was diluted to the appropriate concentrations (Figure 4), and transferred to a Covaris® MicroTUBE (AFA Fiber 6 x 16 mm with Pre-Slit Snap-Cap). Shearing was performed with a Covaris E220 Focused Ultrasonicator using the following settings: duty factor: 5%; peak incident power: 175 W; time: 50 s; cycles per burst: 200; power mode: frequency sweeping; temperature of water bath: 6°C. Fifty microliters of each sheared DNA sample were recovered for library construction.

HyperPrep PF workflow. Post-fragmentation size selection of DNA sheared in 130 µL volumes (500 ng per library) was performed with KAPA Pure Beads (Roche), using a bead-to-sample volume ratio of 0.6X for the first cut and 0.8X for the second cut. These ratios were previously optimized to yield fragmented DNA with a mode size of ~450 bp. To optimize recovery, a heated incubation step (at 37°C for 10 min) was utilized for the final elution of size-selected DNA (in 10 mM Tris-HCl, pH 8.0). Library construction was performed using the standard KAPA HyperPrep protocol, with 5 µL of 15 µM KAPA Dual-Indexed Adapter per library (and a unique i5/i7 index combination for each replicate). Although an optional, second post-ligation cleanup may be included, the standard protocol (one 0.8X post-ligation cleanup) was employed in this study.

HyperPrep PL workflow. DNA sheared in 50 µL volumes (500 ng per library) was transferred directly to the end repair/A-tailing reaction. The standard KAPA HyperPrep protocol was executed, with the same adapter strategy as for the PF workflow. Following the 0.8X post-ligation cleanup, a 0.5X – 0.7X size selection was performed. These parameters were also previously optimized for a mode insert size of ~450 bp. Since adapter-ligated libraries are longer than fragmented DNA, the bead-to-sample volume ratios for post-ligation size selection are lower than for post-fragmentation size selection when targeting the same final insert size.

TruSeq workflow. DNA sheared in 50 µL volumes (1 µg per library) was used directly for library construction. The standard protocol was followed without modifications. Illumina-supplied dual-indexed adapters and cleanup beads were used.

Library quantification. All final, adapter-ligated libraries were quantified with the KAPA Library Quantification Kit for Illumina platforms, using a LightCycler® 480 instrument (Roche). Library dilutions (1/1,000) were made with an epMotion® 5075 automated liquid handling workstation (Eppendorf), and assayed in triplicate.

Table 1. Library construction methodologies used in this study

No.	Abbreviation (replicates)	Fragmentation method	Library preparation kit	Adapters	Size selection	Protocol time	Hands-on tme	Total time
1A	**HyperPrep PF** (*n*=3)	Covaris shearing, using settings provided in TruSeq DNA PCR-Free Protocol Guide (for 350 bp inserts)	KAPA HyperPrep Kit (Roche)	KAPA Dual-Indexed Adapters	Post-fragmentation (0.6X – 0.8X)	2 h	1 h	3 h
1B	**HyperPrep PL** (*n*=3)				Post-ligation (0.5X – 0.7X)	2 h	1 h	3 h
2	**TruSeq** (*n*=4)		TruSeq DNA PCR-free Library Prep Kit (Illumina)	TruSeq DNA PCR-Free HT dual-indexed adapters	After end repair, using vendor's recommendations for a 350 bp insert size	2 h 30 min	1 h 40 min	4 h 10 min

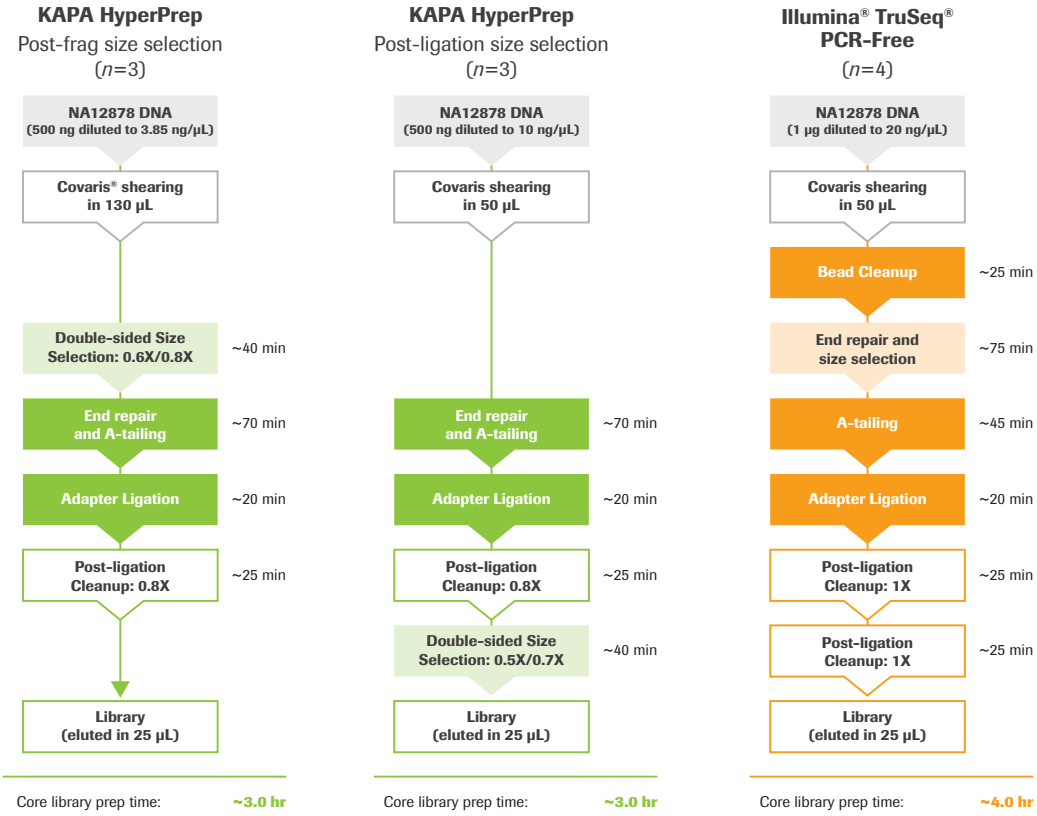


Figure 4. PCR-free library construction workflows employed in this study. Replicate aliquots of NA12878 human genomic DNA (500 ng for the two HyperPrep workflows; 1 µg for the TruSeq workflow) were sheared in using the same Covaris instrument parameters for all workflows (as described in Materials and Methods). The number of replicate libraries prepared for each workflow is indicated at the top. Size selection may be performed at different stages of the KAPA HyperPrep workflow, whereas the TruSeq PCR-free protocol includes size selection after end repair. Core library prep times do not include DNA quantification, Covaris shearing or library QC (quantification and size distribution assessment). For sequencing, a library pool was created from all the available replicate libraries for each workflow.

Library fragment size assessment. To determine true fragment size distributions, a 5 µL-aliquot of each library was amplified for 5 cycles using KAPA HiFi HotStart ReadyMix and KAPA Library Amplification Primer Mix for Illumina (Roche). Amplified libraries were subjected to a 1X post-amplification cleanup with KAPA Pure Beads. Library size distributions were confirmed with a 2100 Bioanalyzer instrument and High Sensitivity DNA Kit (Agilent Technologies).

Sequencing (HiSeq X). Libraries were normalized to a concentration of 3.5 nM where possible (libraries with a final concentration <3.5 nM were used as is). Equal volumes of the replicate libraries available for each workflow were combined to generate a pool for that workflow. Each pool was loaded in a separate lane of an Illumina HiSeq X® instrument, for 2 x 150 bp paired-end sequencing. The loading concentration for each pool is included Table 4 on p. 7.

Sequencing (NovaSeq). Once the HiSeq X run was set up, the remainder of each of the three library pools was normalized to a concentration of 2.5 nM. Equal volumes of these pools (as well as one more human WGS library pool, also normalized to 2.5 nM) were combined to generate 75 µL of material. This was mixed with 75 µL of another library pool, comprised of unrelated human PCR-free WGS and exome libraries. The final pool was loaded at a concentration of 500 pM on an Illumina NovaSeq™ 6000 instrument (S2 flow cell), for 2 x 150 bp paired-end sequencing.

Sequencing reports. Standard HiSeq X and NovaSeq sequencing reports were generated to obtain general sequencing metrics.

Alignment and downsampling. This workflow included sequence alignment to build GRCh38 with BWA-MEM, marking of duplicates with Picard, base quality recalibration with GATK, and lossless conversion to CRAM format with Samtools. A second iteration of the alignment was run after downsampling with Picard DownsampleSam to ~75 billion base pairs (Gb) per sample, for a comparison of variant calls (see below).

Quality control analysis. The steps for this workflow included generation of library insert size, alignment, GC bias and genome coverage metrics with Picard. Samtools v1.3.1 flagstat was run to summarize alignment metrics.

Variant calling. Germline variation was evaluated using GATK HaplotypeCaller version 3.5, which was run on each chromosome. gVCF format variant files were converted to VCF using GATK SelectVariants and all chromosomes combined with CatVariants. Low-quality variant calls, Genoytpe Quality (GQ) <30, were excluded using GATK VariantFiltration and SelectVariants.

Variant evaluation. The resulting VCFs were decomposed into a single alternate allele per line using vt (https://github.com/atks/vt.git; commit bcee48d6ec1dcaf3d0ea975efae209f8ec49eaa6). Decompose (-s), indels were normalized using vt normalize and duplicate calls were collapsed using vt uniq. Finally, indels and SNPs were separated out into separate files for evaluation. Variants were compared to the Genome-in-a-Bottle NA12878 gold standard (v2.19) using GATK 3.5 GenotypeConcordance.

Data analysis tools and specifications are summarized in the **Appendix**.

Results and discussion

Library construction metrics

The Illumina® TruSeq® DNA PCR-Free Library Prep Kit Protocol Guide provides Covaris® shearing parameters for an average insert size of 350 bp. Combined with the size selection performed after end repair, libraries with an average fragment size distribution of ~470 bp is expected. The parameters were, however, found to reproducibly yield adapter-ligated libraries with an average size distribution in the range of 590 – 630 bp (~140 bp longer than expected). Size-selection parameters for the two HyperPrep workflows were adjusted accordingly.

A portion of each library was amplified to produce the electropherograms shown in Figure 5. The two HyperPrep workflows yielded libraries with a very similar and consistent fragment size distribution, slightly shorter than that of the TruSeq libraries.

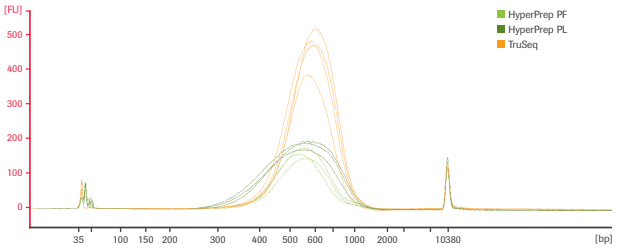


Figure 5. True size distributions of replicate NA12878 libraries generated using the HyperPrep PF (light green), HyperPrep PL (dark green) and TruSeq (orange) workflows. As described in Materials and methods, a 5 µL-aliquot of each library was amplified for 5 cycles to enable accurate fragment size distribution with the 2100 Bioanalyzer instrument and High Sensitivity DNA Kit (Agilent Technologies). All three protocols yielded reproducible size distributions, and libraries that appear to be free of adapter-dimer and unligated adapter.

Average fragment lengths from the electrophoretic analysis *versus* mean insert sizes calculated from trimmed, aligned reads are compared in Table 2. The HyperPrep workflow with post-ligation size selection (HyperPrep PL) returned a slightly larger deviation in mean insert sizes than the two workflows in which size selection was performed earlier in the library construction process.

Table 2. Average library fragment lengths, determined by electrophoretic analysis, compared to mean insert sizes from sequencing data*

Workflow	Average fragment length (bp)	Mean insert size (bp)	
		HiSeq® X	NovaSeq™
HyperPrep PF	580 ±14 bp	392.0 ±102.6 bp	395.3 ±101.4 bp
HyperPrep PL	571 ±2 bp	366.1 ±134.7 bp	369.7 ±134.6 bp
TruSeq	610 ±17 bp	438.4 ±112.2 bp	434.2 ±108.9 bp

*Average fragment lengths determined by electrophoretic analysis of amplified libraries are inclusive of adapters, whereas mean insert sizes calculated from sequencing metrics are not.

Average library concentrations and overall conversion rates (inclusive of losses incurred during fragmentation and size selection) for replicate libraries, obtained from each of the three PCR-free workflows, are given in Figure 6.

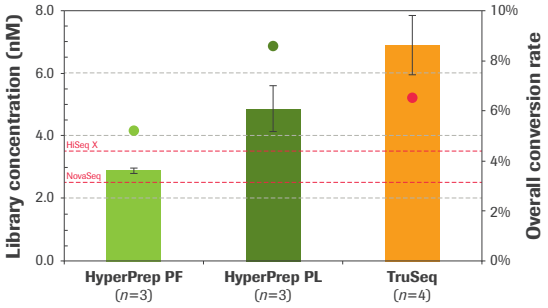


Figure 6. Final library concentrations (bars) and overall conversion rates (dots). The HyperPrep PL workflow (dark green; 500 ng input into fragmentation, post-ligation size selection) returned the highest conversion rate (9%), followed by the TruSeq workflow (orange; 1 µg input into fragmentation, 7% conversion). The HyperPrep PF workflow (light green; 500 ng input into fragmentation, post-fragmentation size selection, 5% conversion) produced the least concentrated libraries, but also the least variation in final library concentration. The red lines indicate the preferred library working concentrations for the HiSeq X and NovaSeq 6000 instruments, respectively.

The HyperPrep PF workflow (with post-fragmentation size selection) produced the most consistent results (only 3% difference in the concentrations of replicate libraries), presumably because library construction is performed with fragments that have already been size selected. This workflow also produced the lowest final library concentrations. This was expected, as experience has shown the recovery of fragmented DNA to be less efficient than the recovery of DNA fragments when size selection is performed later in the library construction workflow. A workflow with post-fragmentation size selection should therefore only be considered when an excess of input DNA is available, or if it is feasible to recover final libraries in a smaller volume.

The HyperPrep PL workflow (with post-ligation size selection) yielded libraries with a final concentration well above the preferred working concentration for both the HiSeq X and NovaSeq 6000 instruments (3.5 nM and 2.5 nM, respectively). This workflow returned the highest overall conversion rate, with 9% of the input into fragmentation converted to sequencing-ready library. Since the KAPA HyperPrep Kit typically converts 40 – 60% of input DNA (into the end repair/A-tailing reaction) to adapter-ligated library when this input exceeds 100 ng, and no size selection is performed, the combined sample loss due to fragmentation and size selection was estimated at 30 – 50%.

The TruSeq workflow yielded the highest final library concentrations (from double the input than the other workflows), but a lower overall conversion rate (7%) than the HyperPrep PL workflow. Since the same fragmentation protocol was used throughout, this suggests larger losses during size selection and/or a less efficient core library construction process. In terms of reproducibility, the TruSeq and HyperPrep PF workflows performed similarly (15% and 14% deviation between the concentrations of replicate libraries, respectively).

Data on file.

Sequencing metrics

A summary of the sequence data generated in this study is given in Table 3, whereas alignment and coverage statistics for the three library pools are summarized in Table 4. A very even amount of data was obtained from the HiSeq X® run across the three pools.

The amount of data obtained for each of the three library pools from the NovaSeq™ run was, however, highly imbalanced. Since the NovaSeq pool was derived from the normalized pools prepared for the HiSeq X run, error during subsequent dilution and/or pooling was assumed to be the likely cause. Because of this uneven distribution of reads and resulting coverage, data were downsampled to the lowest coverage level (75 Gb of data per pool, equivalent to 20 – 22X coverage) for variant calling.

The data confirmed that all three workflows produced high-quality libraries. There was a good correlation between data generated on the two sequencing instruments.

GC-bias plots

GC-bias plots are shown in Figure 7 on p. 8. For context, GC-bias plots for PCR-free WGS libraries generated from real-life blood and saliva samples are also included. Coverage uniformity was highly similar for the NA12878 libraries generated with the different library preparation workflows in this study. Inter-workflow and inter-sequencer variation in the normalized coverage for genomic regions with very low (<25%) and very high (>70%) GC content does not appear to be significant when considered in the context of much larger sample sets.

Variant calling and concordance analysis

Variant calling results, and concordance to the Genome-in-a-Bottle NA12878 gold standard (v2.19) set are given in Figure 8 on p. 8. All three library preparation methods yielded similar sensitivity, false discovery and discordance rates. Results obtained from the two sequencing instruments again correlated very well. Variant concordance results are also represented in the form of Venn diagrams in Figure 9.

Table 3. Sequencing statistics

Workflow	Lane	Loading conc. (pM)	Total Gb PF	% PF Clusters	Average Q-score (Read 1)	Average Q-score (Read 2)	% >Q30 (Read 1)	% >Q30 (Read 2)
HiSeq X								
HyperPrep PF pool	5	288	137.6	73	38.62	37.07	93	88
HyperPrep PL pool	6	350	134.8	72	38.70	37.29	93	88
TruSeq pool	7	350	133.7	71	38.84	37.36	94	89
NovaSeq								
Pool*	N/A	500	1369	75	ND	ND	93	90

*37.5% of the S2 flow cell was occupied by libraries from this study. The rest of the pool consisted of unrelated PCR-free human WGS libraries and exome libraries.
PF=passed filter. Additional statistics for the NovaSeq run were: %PhiX aligned=1.94; % PhiX error rate (R1) 0.40; % PhiX error rate (R2)=0.45.

Table 4. Alignment and coverage statistics

Workflow	Total Gb PF*	PF reads	% PF reads aligned	% PF reads mapping in improper pairs	% Chimeras	Mean coverage
HiSeq X						
HyperPrep PF	137.6	863,400,098	99.85%	2.4%	1.1%	34.2
HyperPrep PL	134.8	831,643,698	99.80%	1.7%	0.8%	32.1
TruSeq	133.7	840,221,726	99.88%	1.8%	0.6%	32.2
NovaSeq**						
HyperPrep PF	86	552,203,222	99.80%	2.3%	1.0%	23.3
HyperPrep PL	164	1,052,427,404	99.87%	1.5%	0.7%	42.0
TruSeq	89	575,462,464	99.90%	1.7%	0.6%	24.3

*PF=Passed filter.
**NovaSeq data yield was calculated after demultiplexing. Uneven data yields across the four library pools were attributed to dilution error. Data were downsampled to 75 Gb per pool for variant calling.

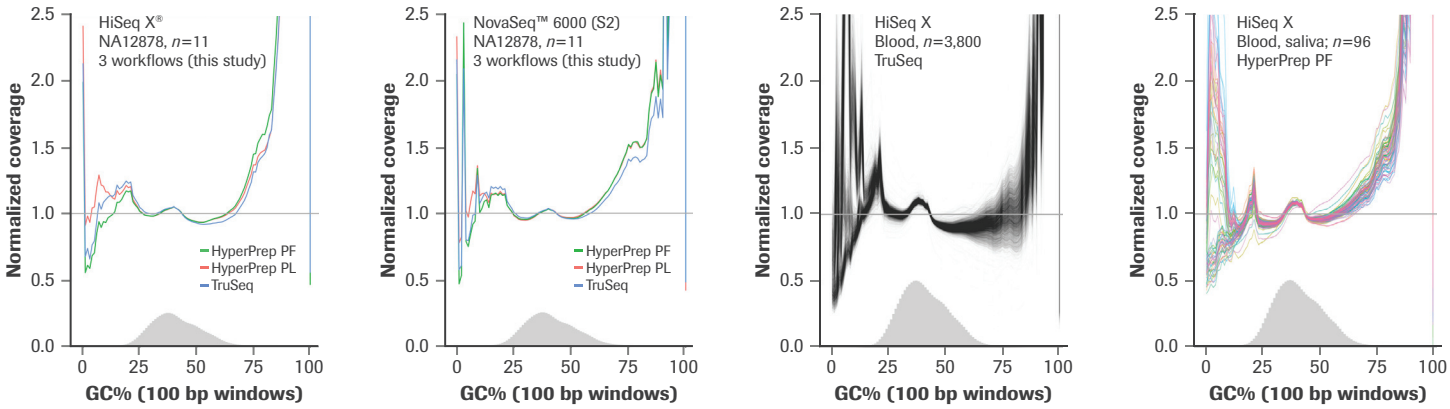


Figure 7. GC-bias plots for PCR-free human WGS libaries generated with different workflows, from different sample types and on different sequencing instruments. The two plots on the left are for libraries generated in this study. The two plots on the right are courtesy of the McDonnell Genome Institute at Washington University (St. Louis, MO) and represent libraries prepared and sequenced in their high-throughput human WGS pipeline, which now utilizes the KAPA HyperPrep Kit for library preparation. The sequencer, number of samples represented and library prep workflow used for each data set are indicated above each plot. Plots were generated with R and ggplot2.

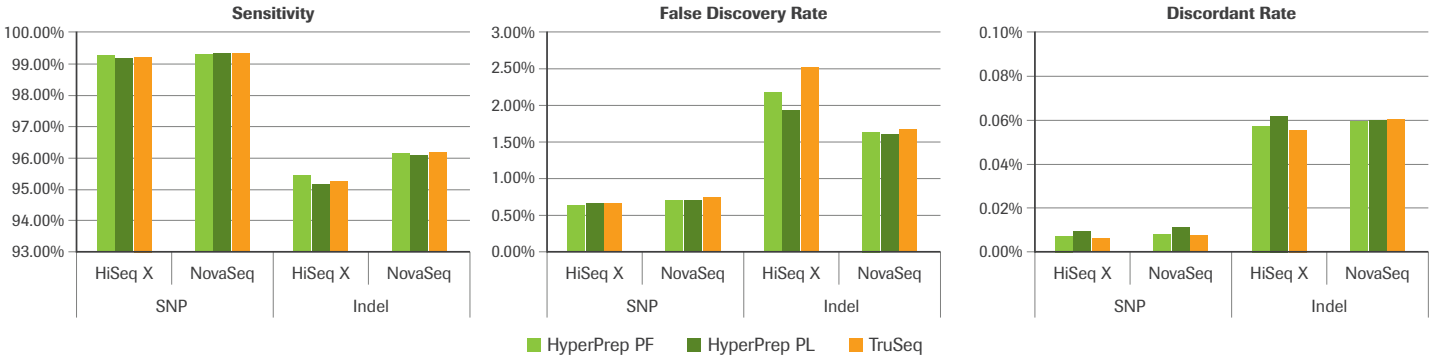


Figure 8. Variant analysis and concordance results for PCR-free human WGS libraries generated with different library preparation workflows. Comparable results were obtained for all three library pools, on both sequencing instruments.

Figure 9. Variant and concordance analysis for different workflows and instruments. The total number of variants (SNPs and indels) included in the analysis is given below each diagram, and the number of variants contained in each set or union is expressed as a percentage of the total. The concordance between workflows was slightly lower than the concordance between identical library pools sequenced on the two different instruments. Individual workflows yielded 2.6 – 3.0% unique variants, which were not subjected to further analysis. Diagrams were generated with R Library VennDiagram (https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-35).

Data on file.

Conclusions

Specific recommendations and validated, step-by-step protocols[13] for the construction of PCR-free human WGS libraries with the KAPA HyperPrep Kit and accessory reagents from Roche were generated in this study. Libraries prepared from 500 ng inputs of a commercial preparation of NA12878 human gDNA yielded high-quality sequencing data. Alignment, coverage, and variant-calling statistics confirmed the sample prep solution from Roche to be highly suitable for routine human WGS on HiSeq® X and NovaSeq™ instruments.

In this study, the KAPA HyperPrep Kit yielded PCR-free human WGS libraries of comparable quality to those generated with the TruSeq® DNA PCR-Free Library Prep Kit (Illumina®); from half the amount of input DNA, and with choice of two protocols that are both 25% shorter than the TruSeq protocol. In a different study, the KAPA HyperPrep Kit, was found to offer significant improvements in sequencing data quality over an established TruSeq® PCR-free workflow; with respect to sensitivity, specificity, and reproducibility when calling indels and CNVs.[14]

As demonstrated in this Application Note, Roche's sample prep solution for PCR-free human WGS offers the following specific benefits:

- The KAPA HyperPrep workflow with post-ligation size selection achieved the **highest conversion rate** and returned very consistent library construction metrics (yields and fragment size distribution). Because of its robust performance, KAPA HyperPrep has become the preferred library prep solution for high-throughput, automated, PCR-free human WGS pipelines (Figure 7 and ref. 14). High core library construction efficiency offers the potential to further reduce DNA input for PCR-free workflows, or to process more challenging samples.
- The KAPA HyperPrep workflow is **more streamlined and flexible** than the TruSeq protocol from Illumina. The results from this study confirm that high-quality data can be generated with different strategies, which may be tailored to different scenarios. Roche provides support for solutions that best meet end-user needs, based on extensive experience and a deep understanding of the parameters that impact the efficiency of sample preparation.
- Roche offers a **complete sample prep solution for PCR-free human WGS,** including a qPCR-based library quantification kit. Our single-supplier solution not only facilitates and streamlines ordering and inventory management, but also guarantees support for the entire workflow from input DNA to sequencing-ready library.

Acknowledgments

The authors wish to thank Robert Fulton, Catrina Fronick, and other staff from the McDonnell Genome Institute at Washington University for project management and sequencing services for this project, and for the historical data included in Figure 7. In particular, we wish to thank Matt Cordes and Lisa Cook, for generating the HiSeq X and NovaSeq data, respectively; as well as Jason Walker and Tiandao Li, for providing sequencing quality metrics for and performing all of the data analysis.

References

1. Pan W, Gu W, Nagpal S, et al. Brain Tumor Mutations Detected in Cerebral Spinal Fluid. Clin Chem. 2015;61(3):514. doi: 10.1373/clinchem.2014.235457
2. Kis O, Kaedbey R, Chow S, et al. Circulating tumour DNA sequence analysis as an alternative to multiple myeloma bone marrow aspirates. Nat Commun. 2017;8(May):15086. doi: 10.1038/ncomms15086
3. Song C-X, Yin S, Ma L, et al. 5-Hydroxymethylcytosine signatures in cell-free DNA provide information about tumor types and stages. Cell Res. 2017;27:1231. doi:10.1038/cr.2017.106
4. Kader T, Goode DL, Wong SQ, et al. Copy number analysis by low coverage whole genome sequencing using ultra low-input DNA from formalin-fixed paraffin embedded tumor tissue. Genome Med 2016;68:121. doi: 10.1186/s13073-016-0375-z
5. Hiranuma N, Liu J, Song C, et al. Cis-Compound Mutations are Prevalent in Triple Negative Breast Cancer and Can Drive Tumor Progression. bioRxiv. 2016. doi: http://dx.doi.org/10.1101/085316
6. Illumina. Effects of Index Misassignment on Multiplexing and Downstream Analysis. 2017. Accessed January 2018.
7. Kircher M, Sawyer S, Meyer M (2012). Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res. 2012;40(1):e3. doi:10.1093/nar/gkr771
8. 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. bioRxiv. 2017. doi: http://dx.doi.org/10.1101/200790
9. Dionne D, Hegarty R, Abreu J, et al. Enabling population-scale PCR-free whole genome sequencing. Poster presented at AGBT 2016.
10. Appel M, Van Rooyen B, Meyer J, et al. The impact of enzymatic fragmentation and limited library amplification on data quality for human whole-genome libraries sequenced on the HiSeq X. Poster presented at SFAF 2016.
11. Roche. KAPA Library Quantification Kit (Illumina platforms) Technical Data Sheet. 2017. Accessed January 2018.
12. Roche. KAPA Library Quantification Technical Guide. 2014. Accessed January 2018.
13. Roche. Sequencing Solutions Technical Note. How To ... Construct human whole-genome shotgun libraries using KAPA HyperPrep. 2018.
14. Tao J, Idrisoglu S, Dinger ME, et al. Evaluation of PCR-free whole genome sequencing for clinical diagnostics. Poster presented at HGSA 2017.

Appendix

Data analysis tools and specifications

Table A1. Data analysis tools and specifications

Process	Program	Version	Description/parameters
Alignments			
Align	BWA-MEM	0.7.15	Reference build38 (GRCh38DH) Opt: -K 10,000,000 -t -p -Y
MarkDuplicates	Picard	2.4.1	–
Sort	Sambamba	0.6.4	–
BQSR	GATK Base-Recalibrator	3.6	dbSNP 138, Known Indels, Mills and 1000G Indels. GATK hg38 Resource Bundle
Apply BQSR	GATK PrintReads	3.6	–
Convert to CRAM	Samtools	1.3.1	–
Quality Control			
CollectInsert-SizeMetrics	Picard	2.14.0	–
Collect-Alignment-Summary-Metrics	Picard	2.14.0	–
Collect-GcBias-Metrics	Picard	2.14.0	–
CollectWgs-Metrics	Picard	2.14.0	Autosomal Chromosome Intervals
Flagstat	Samtools	1.3.1	–
FREEMIX	verifyBamID	1.1.3	Omni 2.5M SNP from 1000G, verifyBamID and Omni VCF Download
Variant Detection			
Haplotype Caller	GATK	3.5	-ERC GVCF -GQB 5 -GQB 20 -GQB 60 Autosomal, Sex, and MT Chromosomes as intervals
SelectVariants	GATK	3.6	–
CatVariants	GATK	3.6	–
VariantFiltration	GATK	3.6	–

Published by:

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

sequencing.roche.com

SEQ100220 04/2018