Application Note

Genomics

Agilent

Trusted Answers

Quality Control in Illumina Sequencing Workflows Using the TapeStation System

Authors

Nicolle Diessl, Ute Ernst, Angela Schulz, and Stephan Wolf
DKFZ Genomics and Proteomics Core Facility, High Throughput Sequencing Unit,
Heidelberg, Germany

Eva Graf
Agilent Technologies,
Waldbronn, Germany

Abstract

This Application Note describes quality control (QC) steps in various Illumina next-generation sequencing (NGS) workflows provided as a service by the German Cancer Research Center (DKFZ) Genomics and Proteomics Core Facility. The QC steps can be performed using the Agilent 4200 TapeStation system with the ScreenTape portfolio. Appropriate reference samples are included as positive controls to monitor the individual steps during library preparation. Representative QC data of these reference samples serve as positive examples for successful completion of critical steps of the most frequently used NGS library protocols.

Introduction

The DKFZ is the largest biomedical research institution in Germany. The High Throughput Sequencing Unit of the DKFZ Genomics and Proteomics Core Facility provides sequencing services to national and international cancer genome projects using several Illumina paired-end and single-read sequencing platforms. The service includes QC of starting material, and library preparation including QC, clustering, sequencing, and data analysis. Various sequencing applications are provided such as whole genome sequencing (WGS), exome sequencing (WES), targeted resequencing, RNA sequencing (RNA-Seq), and sequencing of protein-binding regions (ChIP-Seq).

The selected protocols are suitable to process high-quality DNA and RNA starting material, as well as more challenging samples with low integrity or concentration. The sequencing core facility subjects all samples to an incoming QC upon receipt. During library preparation, additional QC steps are performed to monitor critical passages of the workflow. Lastly, the quality of final libraries is assessed before pooling the samples for sequencing. To verify the success of library preparation, at least one positive control is processed parallel per batch of samples at the DKFZ sequencing core facility. This Application Note describes representative QC data of these reference samples analyzed with the 4200 TapeStation system.

Experimental

Materials

The HiSeq 2000, HiSeq 2500, HiSeq 4000, HiSeq X, NovaSeq, MiSeq, and NextSeq systems and kits for sequencing were obtained from Illumina (San Diego, CA, USA). The Agilent 4200 TapeStation system (G2991AA) in combination with Agilent genomic DNA ScreenTape (p/n 5067-5365) with reagents (p/n 5067-5366), D1000 ScreenTape (p/n 5067-5582) with reagents (p/n 5067-5583), High Sensitivity D1000 (HS D1000) ScreenTape (p/n 5067-5584) with reagents (p/n 5067-5585), and RNA ScreenTape (p/n 5067-5576) with reagents (p/n 5067-5577) from Agilent Technologies (Santa Clara, CA, USA) was used for sample quality control. TruSeq Nano and TruSeq Stranded mRNA kits from Illumina (San Diego, CA, USA), Agilent SureSelect^{XT} Human All Exon v5 kit (p/n 5190-6210) from Agilent Technologies (Santa Clara, CA, USA), and NEBNext ChIP-Seq from New England Biolabs (Ipswich, MA, USA) were used for library preparation. The Qubit 2.0 instrument or Filtermax F3 microplate reader obtained from Molecular Devices (San José, CA, USA) was used for DNA quantification. The Covaris E220 and LE220 instruments from Covaris (Woburn, MA, USA) were used for shearing. The Mastercycler Pro from Eppendorf (Hamburg, Germany), Thermocycler TProfessional from Analytik Jena (Jena, Germany), and GeneAmp PCR Systems from Applied Biosystems were used for PCR amplification.

Reference samples

As positive control samples, Human DNA (NA12891 and NA12878) was obtained from Coriell Institute (Camden, NJ, USA), Human DNA for Whole-Genome Variant Assessment (RM 8398) from NIST (Gaithersburg, MD, USA), and Human Genomic DNA from Roche Diagnostics GmbH (Mannheim, Germany). Universal Human Reference RNA was obtained from Agilent Technologies (Santa Clara, CA, USA), and First Choice Human Brain Reference total RNA was obtained from ThermoFisher Scientific (Waltham, MA, USA).

Methods

Quality control of starting material: The 4200 TapeStation system with the genomic DNA ScreenTape assay and RNA ScreenTape assay was used for sample integrity assessment of DNA and RNA starting material. The quantification of DNA or RNA samples at the sequencing core facility was performed using the Qubit assay based on fluorescence detection.

Library preparation: In general, the Agilent SureSelect^{XT} protocol[1] was used for WES from genomic DNA starting material. The D1000 ScreenTape assay was used for intermediate QC steps of the SureSelect^{XT} protocol. The TruSeq Nano protocol[2] was used to generate WGS libraries. The High Sensitivity D1000 ScreenTape assay was used to analyze sheared DNA of the TruSeq Nano protocol. The NEBNext ChIP-Seq protocol[3] was used for ChIP-Seq of protein-linked DNA. The TruSeq Stranded mRNA protocol[4] was used for RNA-Seq.

Custom WGS from WES: WGS libraries were processed according to the SureSelect^{XT} protocol[1], as described in Chapter 3: Sample Preparation. Instead of hybridization and capturing, 100 ng per sample were amplified with six PCR cycles using appropriate index primers and the postcapture PCR cycling program according to Chapter 5, step 1 of the SureSelect^{XT} protocol. The indexed library was purified as described in step 2, using 1.8x magnetic beads.

Custom TruSeq Nano for FFPE samples: The TruSeq Nano DNA protocol[2] was slightly modified to process formalin-fixed paraffin-embedded (FFPE)-derived samples. Briefly, 100 ng of gDNA were sheared to 150–200 bp fragments by prolonged incubation at Covaris (when using LE220, FFPE samples were ultrasonicated for 480 seconds instead of 130 seconds; when using E220, shearing for FFPE samples was increased to 480 seconds instead of 45 seconds). End repair, adenylation of 3' ends, and ligation of adapters were performed as described in the standard protocol. Adapter-ligated DNA samples were enriched with 10 PCR cycles instead of eight cycles.

Custom ChIP-Seq: The generation of adaptor-ligated libraries was performed in accordance with the NEBNext ChIP-Seq protocol[3]. The PCR enrichment was modified, using 10 µL of adapter-ligated DNA sample and 10 µL of High-Fidelity 2X PCR Master Mix for the PCR reaction.

QC and sequencing of generated libraries: Final libraries were analyzed with the D1000 ScreenTape assay according to the assay guide to evaluate the library size. The molarity was calculated by the maximum peak size of the Agilent TapeStation system and the concentration measured by Qubit or Filtermax microplate fluorometer. In general, libraries of customer samples were normalized to 10 nM, equimolar pooled, transferred to a sequencing flow cell, and loaded onto the appropriate Illumina sequencer. Reference standards were used to control successful library preparation, and were only sequenced for testing, validation, or troubleshooting purposes.

Results and discussion

DNA sequencing

QC of genomic DNA (gDNA): gDNA is used as the starting material for SureSelect^{XT} and TruSeq Nano protocols. The integrity of the gDNA critically affects the success of library preparation and sequencing. The Agilent gDNA ScreenTape assay offers an objective DNA Integrity Number (DIN) for the assessment of gDNA integrity (Figure 1), resulting in a numerical value from 1 (degraded) to 10 (intact)[5]. gDNA originating from FFPE tissue is typically partially degraded, resulting in DIN values below 6 (Figure 2). Optimized protocols allow for successful preparation of FFPE material, and usually require adaption of input amount, shearing incubation time, and increased PCR cycles dependent on sample integrity. Empirical studies accounted for a DIN threshold of 7.0 as acceptance criteria to ensure successful library preparation with DNA samples[6]. The sequencing core facility cannot guarantee a high coverage rate for DNA material with DIN below 7, and processes these samples only after approval by the client. When processing FFPE DNA, other facilities observed good sequencing results using starting material with a DIN >3 and optimized protocols[1].

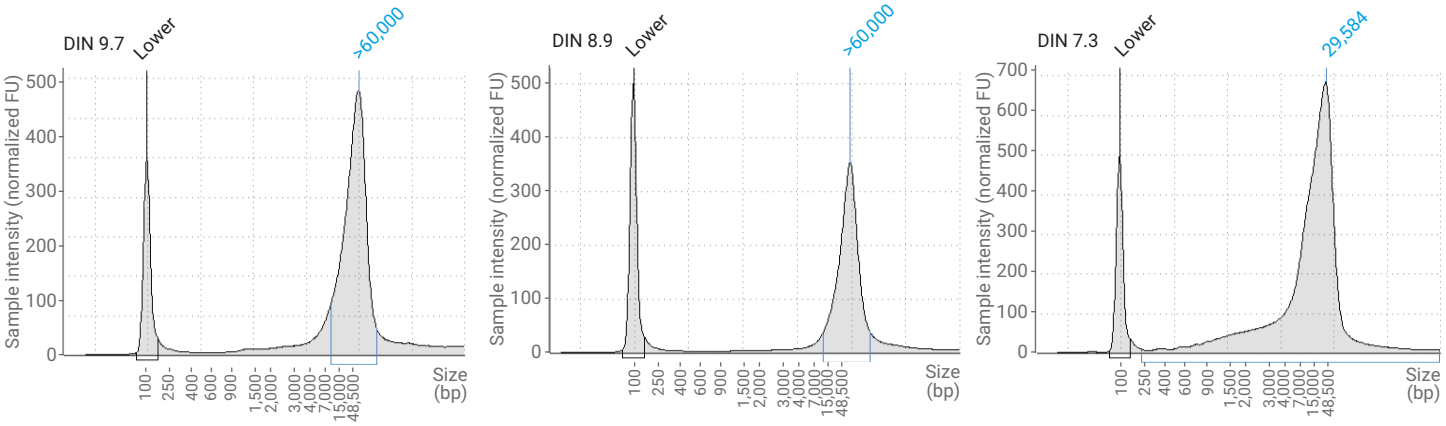

Figure 1. Electropherograms of reference gDNA analyzed with the gDNA ScreenTape assay. With a DIN above 7.0, the samples all qualify as starting material for DNA library preparation workflows.

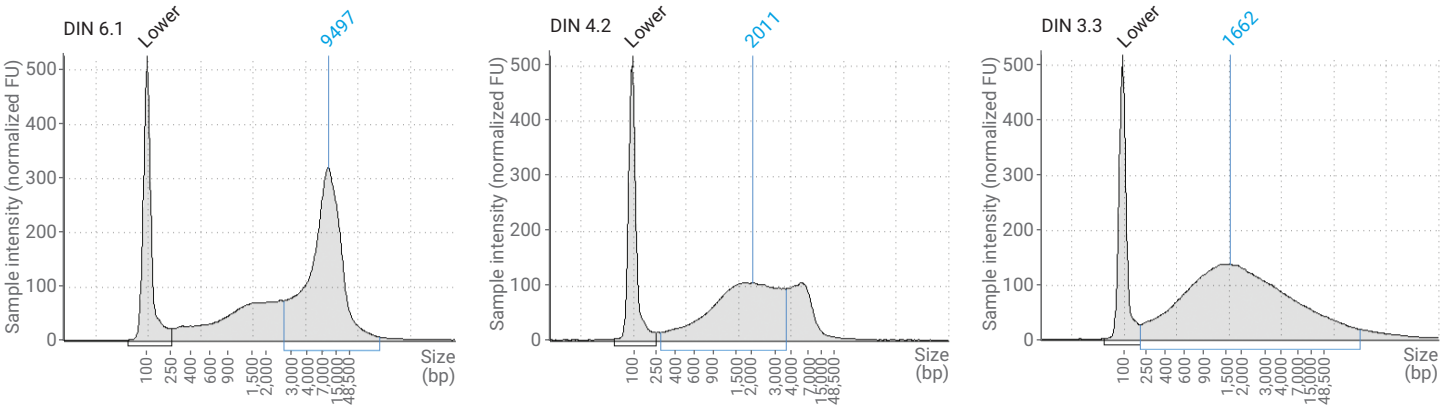

Figure 2. Electropherogram patterns of gDNA with various degradation levels. The DIN of all samples is below the general quality threshold of 7.0, which is typical for FFPE-derived DNA.

SureSelect^{XT}: NGS target enrichment enables a deep analysis of specific regions to identify causal genetic variants of complex conditions. The SureSelect^{XT} protocol[1] is designed to create libraries with enriched targeted regions of the genome for sequencing with Illumina paired-end platforms. For each sample to be sequenced, an individual indexed library is prepared. The sequencing core facility implemented four QC steps[7], starting with gDNA input material (Figures 1 and 2). The two intermediate QC steps include evaluation of smear size after shearing as well as before capturing, and are carried out for all samples in the sequencing core facility using the D1000 ScreenTape assay. The expected size range of the maximum peak of sheared DNA is 150–200 bp (Figure 3). For precapture DNA, a larger maximum peak size of 225–275 bp is expected due to adapter ligation (Figure 4). In addition to sizing, the concentration is evaluated during the same analysis. This is essential, as the subsequent hybridization step requires a DNA concentration above 221 ng/µL.

The last QC step of the SureSelect^{XT} protocol implies the qualification of the final library with the D1000 ScreenTape assay prior to pooling (Figure 5). Due to the addition of index sequences, another size shift is expected between precapture (Figure 4) and final library (Figure 5). The peak maximum of the final library is expected to be positioned between 250 and 350 bp. A minimum concentration of 2 ng/µL is expected for successfully generated final libraries[7]. High-quality samples display a symmetric peak within the expected size range of the precapture product or final library without additional peaks in the electropherogram.

Examples for poor-quality libraries generated during SureSelect^{XT} workflows and possible corrective actions were recently published[8].

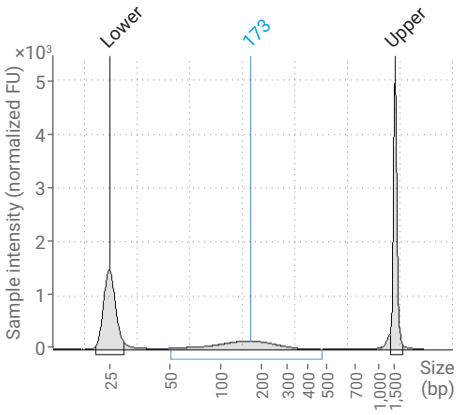

Figure 3. Size evaluation of a sheared DNA sample during the SureSelect^{XT} workflow with the D1000 ScreenTape assay. The sheared DNA shows a maximum peak size of 173 bp, which matches the expected size range of 150 to 200 bp.

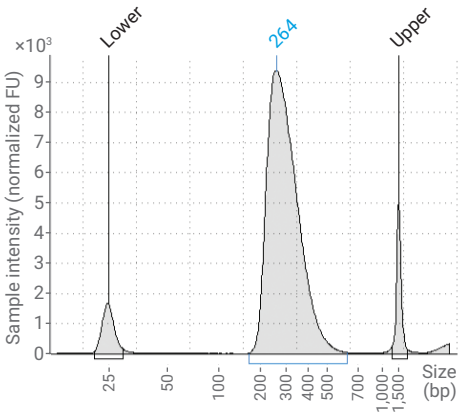

Figure 4. Precapture DNA of the SureSelect^{XT} workflow analyzed with the D1000 ScreenTape assay with a maximum peak size of 264 bp. The expected maximum peak size ranges from 225 to 275 bp.

Custom whole genome sequencing

WGS allows identification of unknown conserved alterations of the genome, and serves as a reference sequence for selected samples that require deep sequencing of the exome by WES. The standard SureSelect^{XT} protocol[1] is designed to generate WES libraries by target enrichment. The sequencing core facility created a modified protocol that allows splitting samples during the SureSelect^{XT} workflow into WES and WGS libraries. With this procedure, whole genome and whole exome sequencing data can be achieved from a single aliquot of starting material. The final libraries are qualified using the 4200 TapeStation system (Figure 6) with the same requirements for size and quantity as in the standard protocol.

Illumina TruSeq Nano: The TruSeq

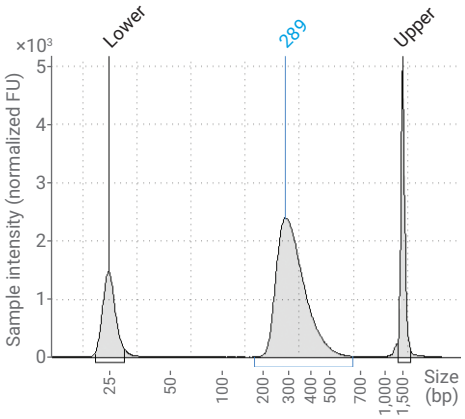

Figure 5. Final library of a SureSelect^{XT} workflow analyzed with the D1000 ScreenTape assay. The electropherogram shows a peak with a maximum at 289 bp, within the acceptable size range of 250 –350 bp.

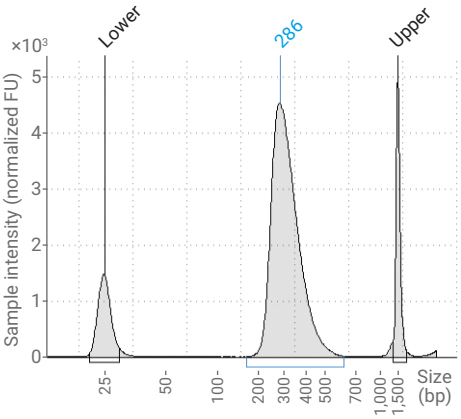

Figure 6. Final library of the SureSelect^{XT} workflow modified for WGS sequencing analyzed with the D1000 ScreenTape assay. The electropherogram shows a peak with a maximum at 286 bp, which matches the acceptable size range of 250–350 bp.

Nano DNA workflow[2] generates WGS libraries, and was designed for samples with limited available DNA. The workflow comprises three QC steps, the first being quality assessment of gDNA starting material, as shown in Figures 1 and 2. After initial QC, 100 ng of gDNA are used for ultrasonic fragmentation to create a 350 bp insert size. The size distribution of the fragmented DNA is evaluated using the High Sensitivity D1000 ScreenTape assay (Figure 7), which is suited for QC of low sample amounts. A maximum peak size between 280 and 480 bp indicates successful fragmentation of intact gDNA (Figure 7A). For the processing of FFPE-derived DNA samples, the sequencing core facility uses an optimized protocol including prolonged shearing and more PCR cycles. The custom workflow generates sheared FFPE DNA with a maximum peak size of 150–200 bp (Figure 7B).

Effectively fragmented samples of either gDNA or FFPE starting material are end-repaired, and, in the case of gDNA, size selected. After adenylation of 3' ends and adapter ligation, the libraries are enriched to generate the final product. The final libraries are evaluated with the D1000 ScreenTape assay to verify the expected size shift. A size shift of 120 bp for single-index adaptors, and 130 bp for dual-index adaptors in comparison to the sheared DNA is expected. Consequently, for 350 bp libraries from gDNA with high integrity as starting material, the expected maximum peak size of the final library is approximately 470 bp for single-indexed libraries, or 480 bp for dual-indexed libraries, respectively (Figure 8A). With respect to FFPE starting material, the expected maximum peak size of the final library is approximately 320 bp (Figure 8B). Electropherogram profiles of high-quality end products show a single broad peak at the expected size range. High-quality final libraries are normalized and pooled before the hybridization to the sequencing flow cell.

NEBNext ChIP-Seq: Chromatin

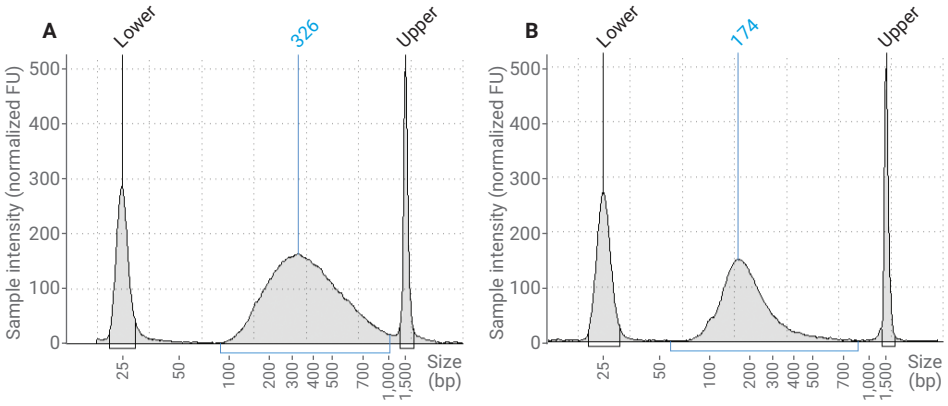

Figure 7. Fragmented DNA generated with the TruSeq Nano protocol using a Covaris Ultrasonicator and analyzed with the High Sensitivity D1000 ScreenTape assay. A) Sheared gDNA sample with a maximum peak size of 326 bp. B) Sheared DNA derived from FFPE tissue with a maximum peak size of 174 bp.

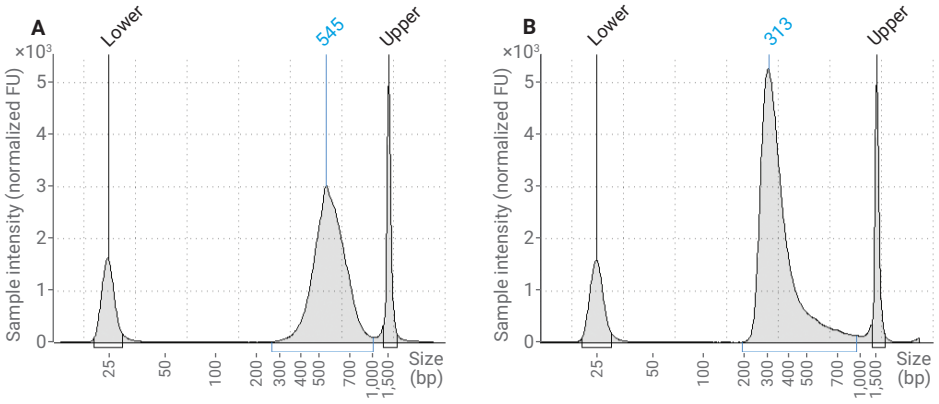

Figure 8. End-products of the TruSeq Nano protocol analyzed with the D1000 ScreenTape assay. A) Final library of a gDNA reference sample with a maximum peak size of 545 bp. For 350 bp libraries, the expected maximum peak size of the final libraries ranges from 460 to 600 bp. B) Final library of a DNA sample originating from FFPE material with a maximum peak size of 313 bp. A size shift of 120 bp resp. 130 bp for the final library compared to the sheared DNA sample is expected.

immunoprecipitation (ChIP) sequencing is an NGS method combining ChIP with massive parallel sequencing to reveal binding sites of DNA-associated proteins. The starting material for the NEBNext ChIP-Seq workflow[3] is chromatin-immunoprecipitated DNA unlinked from protein. The workflow requires two QC steps, the first being quality assessment of the unlinked DNA, and the second the quality assessment of the final library. After incoming QC of the unlinked DNA starting material (Figure 9), the samples are end-repaired, followed by dA-tailing and adaptor ligation. The adapter-ligated libraries undergo a bead-based size selection. The conditions for the size selection are optimized for specific fragment lengths, which are 150, 200, 250, 300, or 400 bp. The insert size of the starting material is determined (Figure 9), and samples are collated to the closest available fragment length for the size selection step.

End products of the ChIP workflow are analyzed with the D1000 ScreenTape assay to verify the expected total library size according to the size selection step (Figure 10).

The final library includes inserts and adapters; therefore, a size shift of 120 bp for single-index adaptors, and 130 bp for dual-index adaptors is expected compared to the starting material (Figure 10A). Figure 10B shows an example of a final library with 494 bp, generated from starting material with approximately 300 bp. Larger libraries may show an increased size shift, since the bead-based size selection is not as accurate. The sizing result is used to calculate the molarity of the libraries, which are then normalized, pooled, and sequenced.

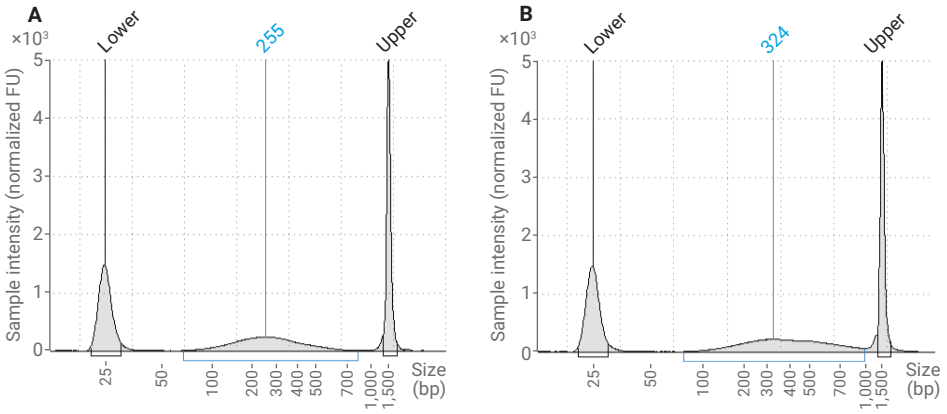

Figure 9. Size determination of starting material for ChIP sequencing with the D1000 ScreenTape assay. A) Example for a sample with an insert size of 255 bp, which was used for the protocol optimized for 250 bp. B) Example for a dsChIP-Seq DNA with an insert size of 324 bp, assigned to the protocol optimized for 300 bp.

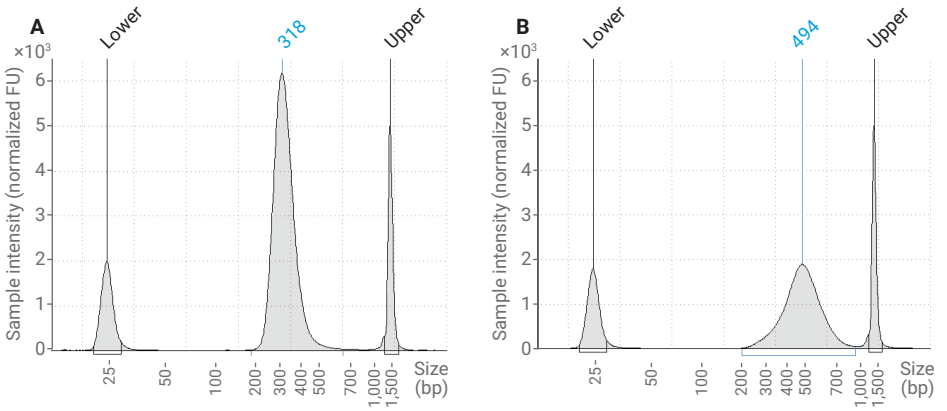

Figure 10. Final library of the ChIP-Seq workflow analyzed with the D1000 ScreenTape assay. Successful libraries show a narrow library distribution with a peak size of 120–130 bp larger than the starting material. A) Example for a final library with a maximum peak size of 318 bp, correlating to an insert size of 200 bp. B) Example of a final library with a maximum peak size of 494 bp, generated with starting material with an approximate size of 300 bp.

RNA sequencing

QC of RNA: RNA is even more subject to degradation than DNA due to the ubiquitous presence of RNase and its more fragile single-stranded structure. Therefore, monitoring the integrity of starting material is indispensable, and it is highly advisable to process a reference sample as positive control throughout the library preparation and sequencing. With the Agilent RNA ScreenTape assay, the RNA integrity number equivalent (RIN^e) delivers an objective assessment of the integrity of RNA starting material. RIN^e has been proven to be equivalent to the widely accepted quality metric RIN[9]. The fragmentation conditions of RNA-seq protocols used by the sequencing core facility are optimized for high-quality RNA; more precisely, a RIN^e of 8.0 or higher is recommended for successful library preparation[4]. Figure 11 shows two samples close to this threshold, one passing and one failing the quality requirement. The use of degraded RNA can result in low yield, over-representation of 3' ends of the RNA molecules, or failure of the protocol.

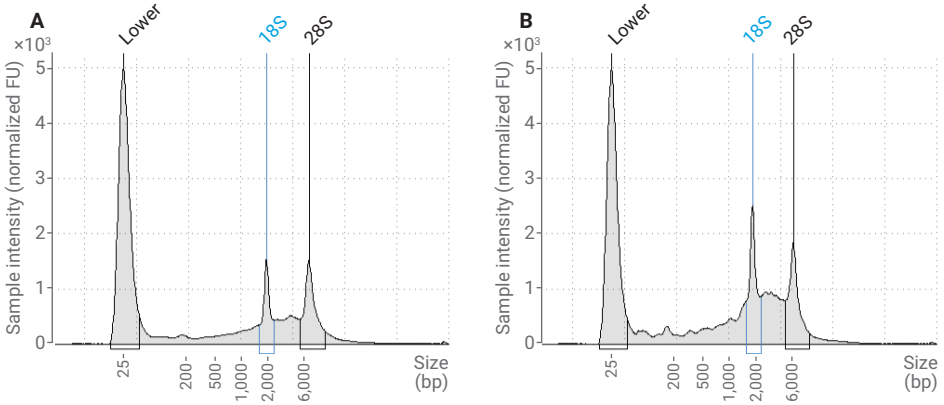

Figure 11. RNA integrity analysis of RNA reference samples with the RNA ScreenTape assay. A) This RNA sample with a RIN^e of 8.1 qualifies for RNA-Seq library preparation. B) RNA sample with a RINe of 7.8, missing the RIN^e threshold of 8.0.

TruSeq Stranded mRNA: RNA-Seq is applied for transcriptome sequencing and gene expression analysis. The TruSeq Stranded mRNA workflow[4] requires total RNA for starting material. It is essential to monitor the integrity of total RNA as starting material in RNA-seq for reliable sequencing data (Figure 11). During the first step in the workflow, poly-A RNA molecules are captured by poly-T oligo magnetic beads. Fragmentation of mRNA is achieved by cleavage with divalent cations. Fragmented mRNA is transcribed into first-strand cDNA using reverse transcriptase and random primers, followed by second-strand cDNA synthesis and subsequent ligation of the adapter. The products are purified, and PCR-amplified to create the final cDNA library (Figure 12), which is analyzed with the D1000 ScreenTape assay[4]. A high-quality library showing a single symmetric peak with a maximum between 250 and 350 bp is suitable for subsequent cluster generation and sequencing. The exact library size depends on the sample. The molarity of the end products is calculated using the concentration result of the Qubit fluorometer or the microplate reader.

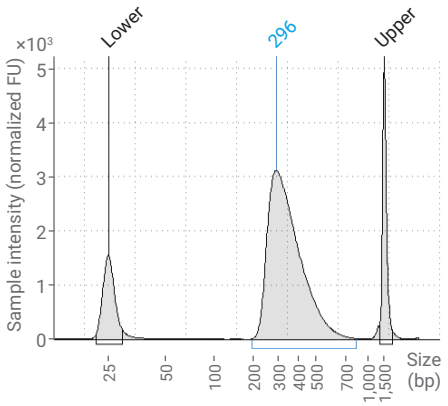

Figure 12. Final cDNA library product of a TruSeq Stranded mRNA workflow analyzed with the D1000 ScreenTape assay. The maximum peak size is typically at approximately 300 bp, as shown in this electropherogram of a reference sample.

Conclusions

Sample QC enables monitoring of the success of library preparation for various Illumina NGS applications, minimizing the risk of producing unreliable sequencing data due to poor sample quality. The DKFZ sequencing core facility successfully implemented the 4200 TapeStation system together with the ScreenTape portfolio for quality control at key steps of multiple Illumina sequencing workflows. For the core facility as a service provider, it is essential to determine whether the input material provided by clients is fit for purpose. At the sequencing core facility, DNA or RNA starting material sample integrity is assessed with the 4200 TapeStation system, and quantification is currently performed using the Qubit assay. However, the gDNA and RNA ScreenTape assays also provide quantitative data, revealing the sample concentration within the same QC step. The intermediate QC steps are specifically useful during the establishment of a new library preparation protocol and for troubleshooting established protocols. Reference samples as positive controls are used to identify deviations, which allows for timely implementation of corrective actions in case of failure. To create optimum cluster densities during sequencing, it is crucial to accurately quantify the final libraries of all workflows. Currently, the molarity determination of final libraries is validated in the sequencing core facility with an external calculation step using the sizing results of the D1000 ScreenTape assay together with the concentration results of the Qubit fluorometer or microplate reader. However, the TapeStation Analysis software also provides molarity data using the region function, which allows to directly evaluate library size and molarity within a single QC step.

References

1. Agilent SureSelectXT Target Enrichment System for Illumina Paired-End Multiplexed Sequencing, *Agilent Technologies User Manual*, publication number G7530-90000, **2017**.
2. TruSeq Nano DNA Library Prep Reference Guide, publication number #15041110 Rev. D, **2015**.
3. NEBNext ChIP-Seq Library Prep Master Mix Set for Illumina Instruction Manual, publication number #E6240S/L, **2016**.
4. TruSeq® Stranded mRNA Sample Preparation Guide, publication number #15031047 Rev. E, **2013**.
5. DNA Integrity Number (DIN) with the Agilent 2200 TapeStation System and the Agilent Genomic DNA ScreenTape Assay, *Agilent Technologies Application Note*, publication number 5991-5258EN, **2015**.
6. Use of the Agilent 4200 TapeStation System for Sample Quality Control in the Whole Exome Sequencing Workflow at the German Cancer Research Center (DKFZ), *Agilent Technologies Application Note*, publication number 5991-7615EN, **2016**.
7. The DNA Integrity Number (DIN) Provided by the Genomic DNA ScreenTape Assay Allows for Streamlining of NGS on FFPE Tissue Samples, *Agilent Technologies Application Note*, publication number 5991-5360EN, **2017**.
8. Sample Quality Control in Agilent NGS Solutions, *Agilent Technologies Application Note*, publication number 5994-0127EN, **2018**.
9. Comparison of RIN and RIN^e algorithms for the Agilent Bioanalyzer and the Agilent 2200 TapeStation systems, *Agilent Technologies Technical Overview*, publication number 5990-9613EN, **2016**.

www.agilent.com/chem

Printed in the USA, October 23, 2018
5994-0327EN

Agilent
Trusted Answers