

Sample Quality Control in Agilent NGS Solutions

Authors

Madhurima Biswas
and Shweta Sharma
Agilent Technologies, Inc.
La Jolla, CA USA

Weiwei Liu, Tracy Liu,
and David Weiss
Agilent Technologies, Inc.
Santa Clara, CA USA

Abstract

Next Generation Sequencing (NGS) is a valuable, high-throughput tool for rapid generation of massive scale sequencing data. This technology takes advantage of the ability to sequence multiple short fragments in a parallel fashion, thereby generating large amounts of reads that can be compiled into a consensus sequence, representing the genome being sequenced. High quality NGS data rely on the quality of the short fragment library. Agilent has developed a full suite of tools for the generation of NGS libraries from DNA and RNA with or without target enrichment of specific regions in the genome or transcriptome. Accurate quality assessment of the samples throughout the workflow is key to overall success. Agilent offers two automated electrophoresis platforms, the TapeStation and Bioanalyzer systems, that can easily be integrated into the NGS library generation workflows.

This Application Note gives further insight into the significance of sample quality control (QC), and provides examples of expected and anomalous electropherogram profiles throughout the library preparation and enrichment process on both automated electrophoresis platforms.

Introduction

The Agilent NGS product portfolio enables effective generation of high-quality DNA and cDNA libraries. These libraries can then be subjected to target enrichment using long oligo libraries. This target enrichment technology permits increased representation and focused sequencing of the regions of interest. Agilent offers three established target enrichment solutions including SureSelect^{XT}, SureSelect^{XT2} based on a multistep ligation method, SureSelect^{QXT} based on a transposase-mediated tagmentation method, and HaloPlex and HaloPlex^{HS} based on a restriction enzyme based

amplicon method. The SureSelect^{XT2} workflow enables precapture pooling of 8 or 16 samples depending on the capture size. HaloPlex and HaloPlex^{HS} leverage restriction digestion-mediated fragmentation and fast workflow as well as molecular barcodes in the case of HaloPlex^{HS} for rare variant detection.

It is imperative to accurately quantify and monitor the quality of the sample as it progresses through various workflows. Target enrichment workflows of SureSelect and HaloPlex integrate the TapeStation and Bioanalyzer systems for quantitative and qualitative assessment of samples at various key steps (Figure 1).

Materials and methods

The 4200 TapeStation (G2991AA) and 2200 TapeStation (G2965AA) systems with D1000 ScreenTape (5067-5582) and reagents (5067-5583), High Sensitivity D1000 (HS D1000) ScreenTape (5067-5584) and reagents (5067-5585), Genomic DNA ScreenTape (5067-5365) and reagents (5067-5366), 2100 Bioanalyzer system (G2939BA) with DNA 1000 kit (5607-1504) and High Sensitivity DNA (HS DNA) kit (5067-4626), SureSelect^{XT} (G9611A), SureSelect^{XT2} (G9621A), SureSelect^{QXT} (G9681A), HaloPlex (G9901C), and HaloPlex^{HS} (G9931C) reagent kits were obtained from Agilent Technologies. Quantity determination of DNA libraries was measured using the region functionality. Unless stated, the manufacturer's protocols and guidelines were followed.

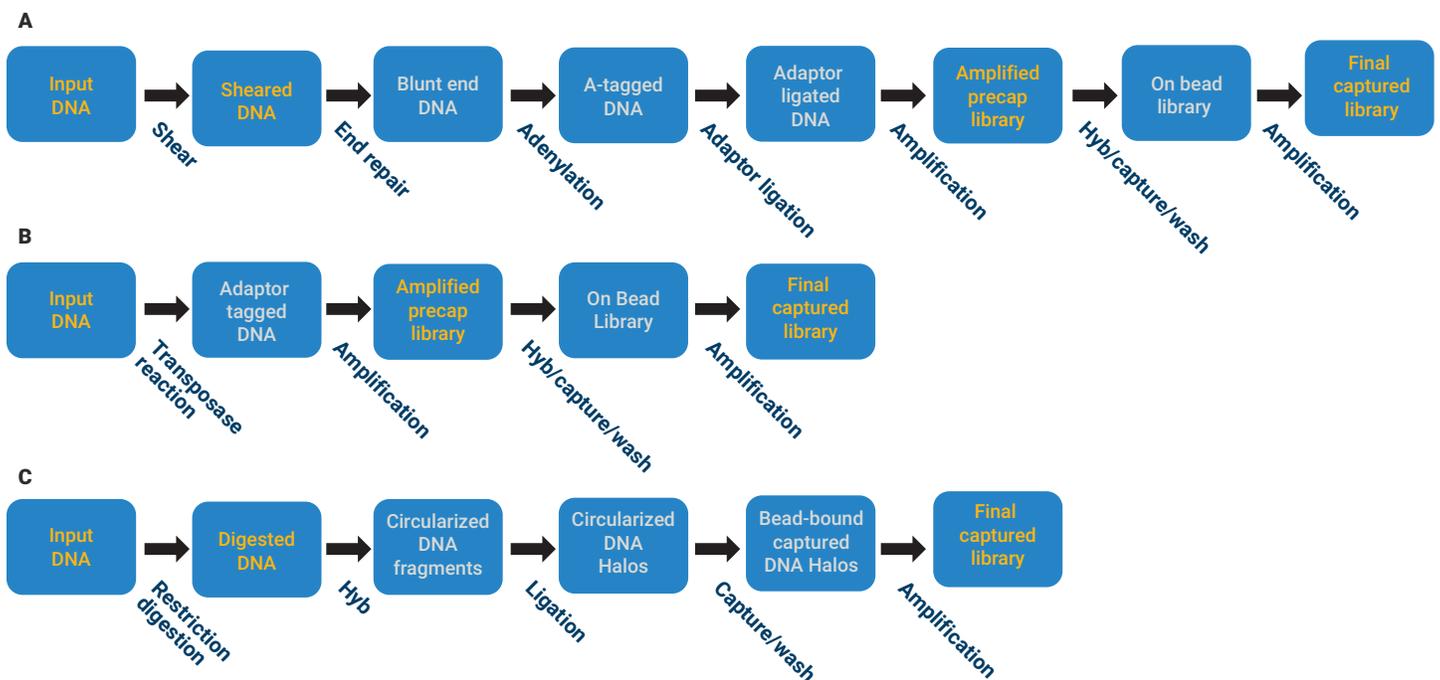


Figure 1. Schematic workflow overview of Agilent target enrichment systems with recommended QC checkpoints highlighted in yellow using either the TapeStation or the Bioanalyzer system. A) The SureSelect^{XT} and SureSelect^{XT2} library preparation workflow; QC in the post shear step is optional. B) The SureSelect^{QXT} library preparation workflow. C) The HaloPlex and HaloPlex^{HS} library preparation workflow. Enzymatic digestion is tested by analyzing enrichment control DNA (ECD) digest, included in the kit.

Results and Discussion

SureSelect^{XT} and SureSelect^{XT2}

Multistep adaptor ligation library preparation methods such as SureSelect^{XT} and SureSelect^{XT2} have four main QC points in the overall library preparation and capture workflow: input DNA, post shearing, amplified precapture library, and amplified post capture library (Figure 1A)¹. Each of these QC points provides information critical for successful sequencing library generation.

Quality assessment of the input material has a significant impact on downstream processing. It is important to determine the average fragment integrity of the starting material so that any necessary modifications to the protocol can be made to account for overly degraded DNA. Depending upon the sample type, length can be assessed on either the TapeStation or the Bioanalyzer system. For example, if genomic DNA (gDNA) is the starting material, its assessment is best accomplished on the TapeStation system using the Agilent Genomic DNA ScreenTape assay and software generated DNA Integrity Number (DIN)². DIN allows numerical assessment of gDNA integrity, with 1 representing the most degraded and 10 being the most intact (Figure 2). In most cases, a DIN greater than 7 is acceptable for the next step of library preparation^{3,4}. Where sample degradation cannot be avoided (for example if derived from formalin-fixed paraffin-embedded (FFPE) tissue), lower DIN scores necessitate greater initial DNA input, a higher number of PCR cycles, and deeper sequencing.

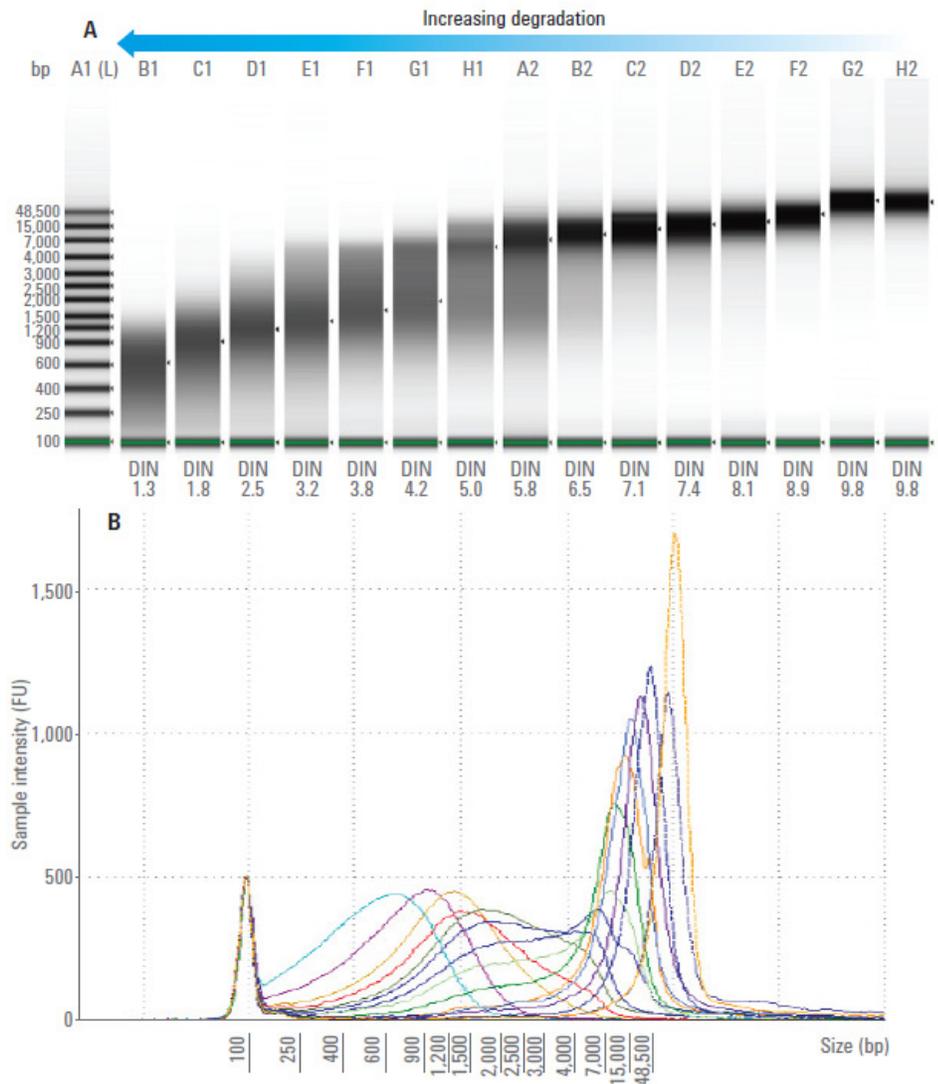


Figure 2. Assessment of gDNA integrity of a sample series with increasing degradation using the Genomic DNA ScreenTape assay on the TapeStation system. A) Gel view is shown with DIN, indicating the gDNA integrity of each sample. High-quality gDNA samples show a single, high molecular weight band. Degraded samples show up as a smear. B) Electropherogram overlay of gDNA samples with varying levels of degradation. High-quality gDNA samples exhibit a sharp peak, and degraded samples a broad smear.

Input gDNA meeting quality standards is sheared to generate fragments appropriate for the type of experiment. The SureSelect workflows use a Covaris ultrasonic shearing system to fragment gDNA. The sheared gDNA library is then analyzed on TapeStation or Bioanalyzer systems to check that the fragments give a single smear distribution with the median size ranging between 150 to 200 bp (Figure 3). Assessment of the size and uniformity of shearing through this step is critical in establishing the baseline for size comparison in subsequent steps of the SureSelect target enrichment protocol. Uneven shearing (Figures 4A and 4B) may result from poor quality input DNA or DNA suspended in incorrect buffer, other than TE. It may also happen from commonly encountered instrumental issues such as:

- Low sample volume (Figure 4B)
- Allowing an air gap to interfere with fragmentation
- Water level too high or low
- Insufficient degassing
- Temperature of water bath outside 6–8 °C

Instrument issues leading to incomplete fragmentation may also result in extended size range (Figure 4C). Too much input DNA used for shearing leads to inadequate shearing and peak tailing (Figure 4D). Adherence to the recommended amounts of starting material, optimized for Covaris shearing, helps to avert such consequences.

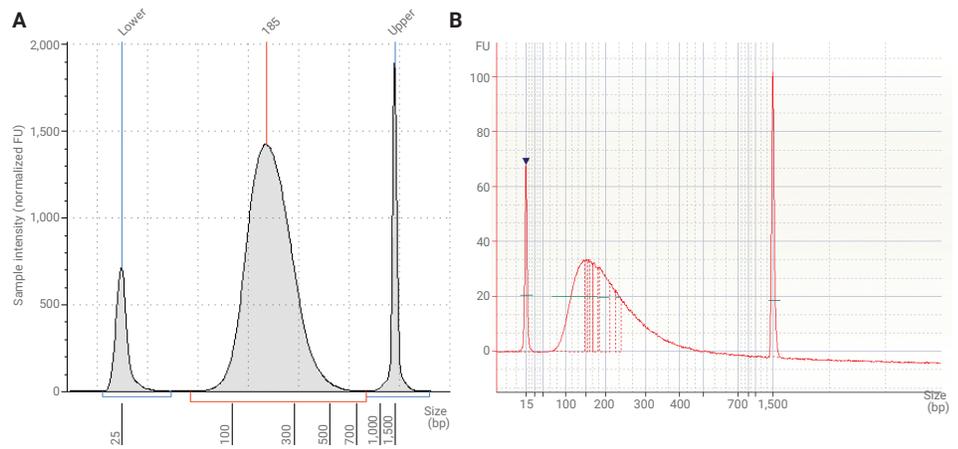


Figure 3. Electropherogram pattern of sheared DNA, showing maximum peak size between 150 and 200 bp. A) Electropherogram of sheared gDNA separated with the D1000 ScreenTape assay and the TapeStation system. B) Electropherogram of sheared gDNA analyzed with the DNA 1000 assay and the Bioanalyzer system.

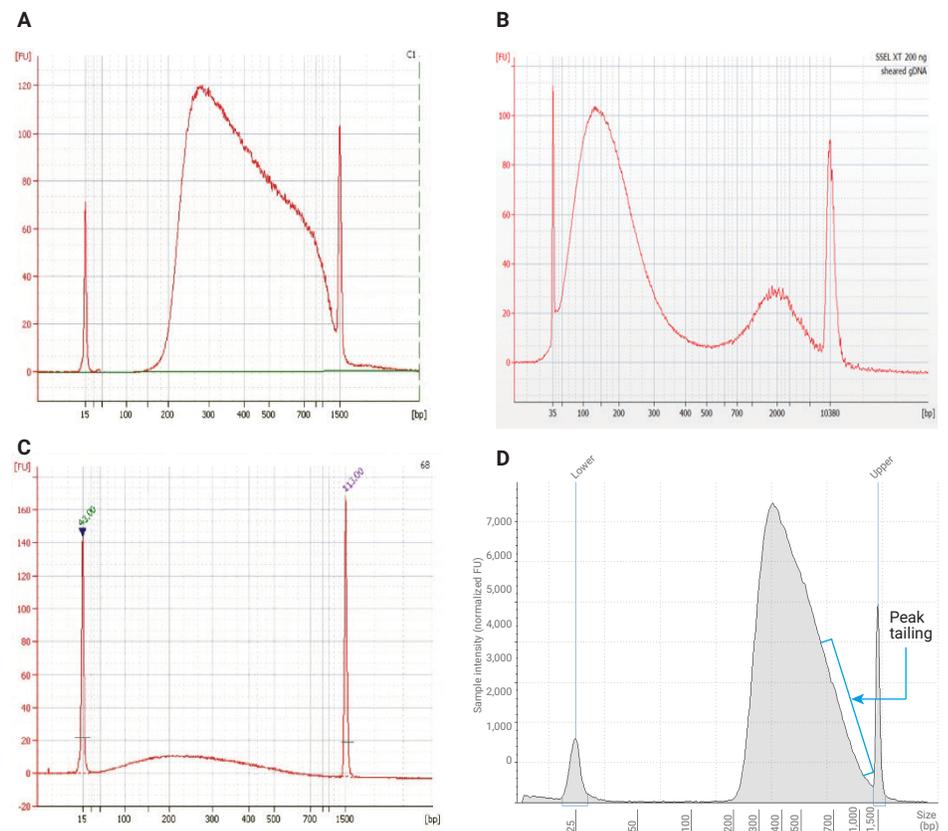


Figure 4. Electropherogram pattern of abnormal sheared gDNA samples. A to C show electropherograms obtained with the Bioanalyzer system. A) Sample showing uneven shearing analyzed with the DNA 1000 assay B) Sample with additional large peak analyzed with the HS DNA assay. C) Electropherogram showing extended size range obtained with the DNA 1000 assay. D) Peak tailing of a sample analyzed with the D1000 ScreenTape assay and TapeStation system.

Once samples have been sheared, they are then end-repaired, adenylated, and adaptor-ligated. After the adaptor ligation is completed, the libraries are amplified. Next, these amplified adaptor-ligated libraries, or amplified precaptured libraries, are assessed on the TapeStation or Bioanalyzer system for sample QC. At this step, a size shift of approximately 60–80 bp with SureSelect^{XT} (Figure 5) and 80–100 bp with SureSelect^{XT2} (Figure 6) with respect to post sheared library is expected.

The difference in size results from the indexes attached to adaptors in SureSelect^{XT2}, which adds extra bases compared to SureSelect^{XT}. Adaptor ligation is usually not 100 % efficient, so the D1000 ScreenTape assay with the TapeStation system, or the DNA 1000 assay, with the Bioanalyzer system can be used to visualize the size shift after successful ligation of adaptors, as inefficient adapter ligation will result in reduced library complexity after PCR¹. The expected size change indicates a successful addition of the adaptors as well as an optimal enrichment of those adaptor-ligated fragments through amplification. During this QC step, it is important to quantify the library based on the expected size range to ensure that there is an adequate amount as per the chosen protocol for the subsequent step of library capture; also look for the presence of excess adaptors, which should be less than 10 % of the main library (Figure 7).

Free adaptors are typically from inefficient ligation due to too much input DNA, reagent issues, or suboptimal incubation temperatures for ligation. During the amplification step, PCR artifacts from overamplification (Figures 8A and 8B)⁵ and primer-dimers (Figure 8E) may be observed. When working with small amounts of DNA, primer-dimers can be removed by performing additional SPRI bead cleanup steps or gel-based size selection.

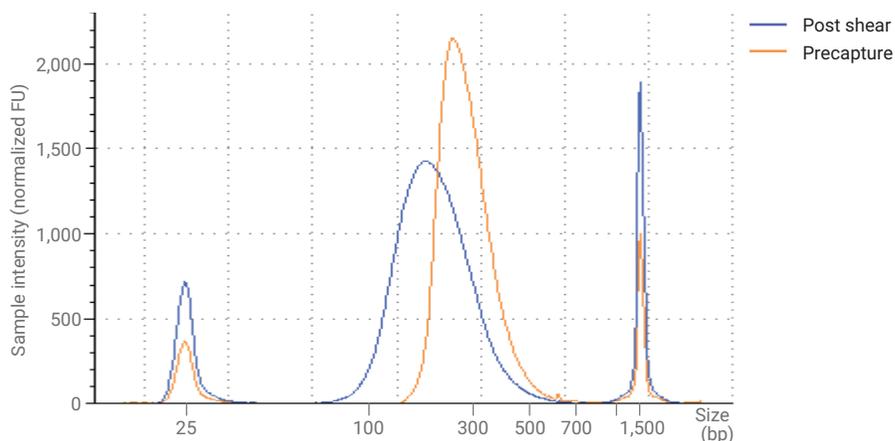


Figure 5. Electropherogram overlay of purified sheared (blue) and purified adaptor-ligated (orange) libraries showing a size shift of 60–80 bp in SureSelect^{XT} amplified precapture libraries with respect to post sheared libraries analyzed with the D1000 ScreenTape assay and the TapeStation system.

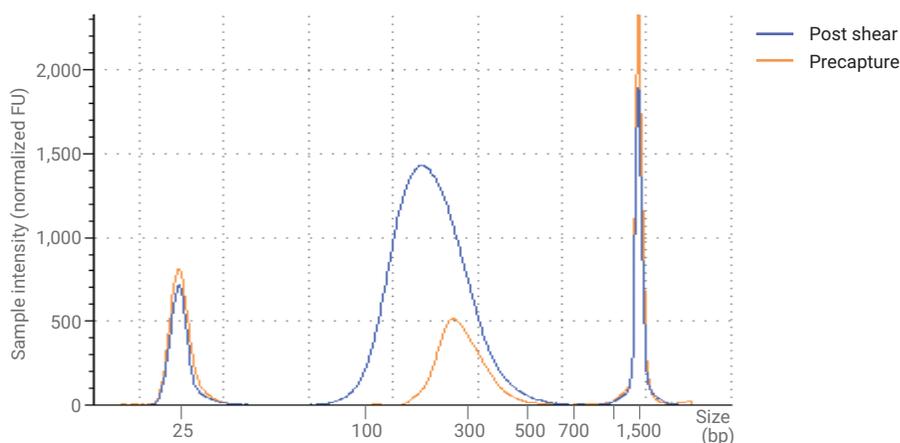


Figure 6. Electropherogram overlay of purified sheared (blue) and purified adaptor-ligated (orange) libraries showing a size shift of 80–100 bp in SureSelect^{XT2} amplified precapture libraries with respect to post sheared libraries obtained with the D1000 ScreenTape assay and the TapeStation system.

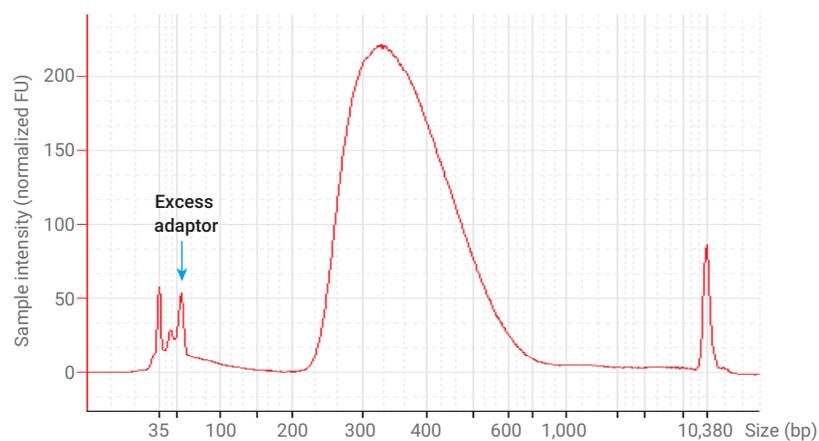


Figure 7. Electropherogram pattern of adaptor ligated precapture amplified library from SureSelect target enrichment workflow showing excess adaptor, next to the lower marker of the HS DNA assay analyzed with the Bioanalyzer system.

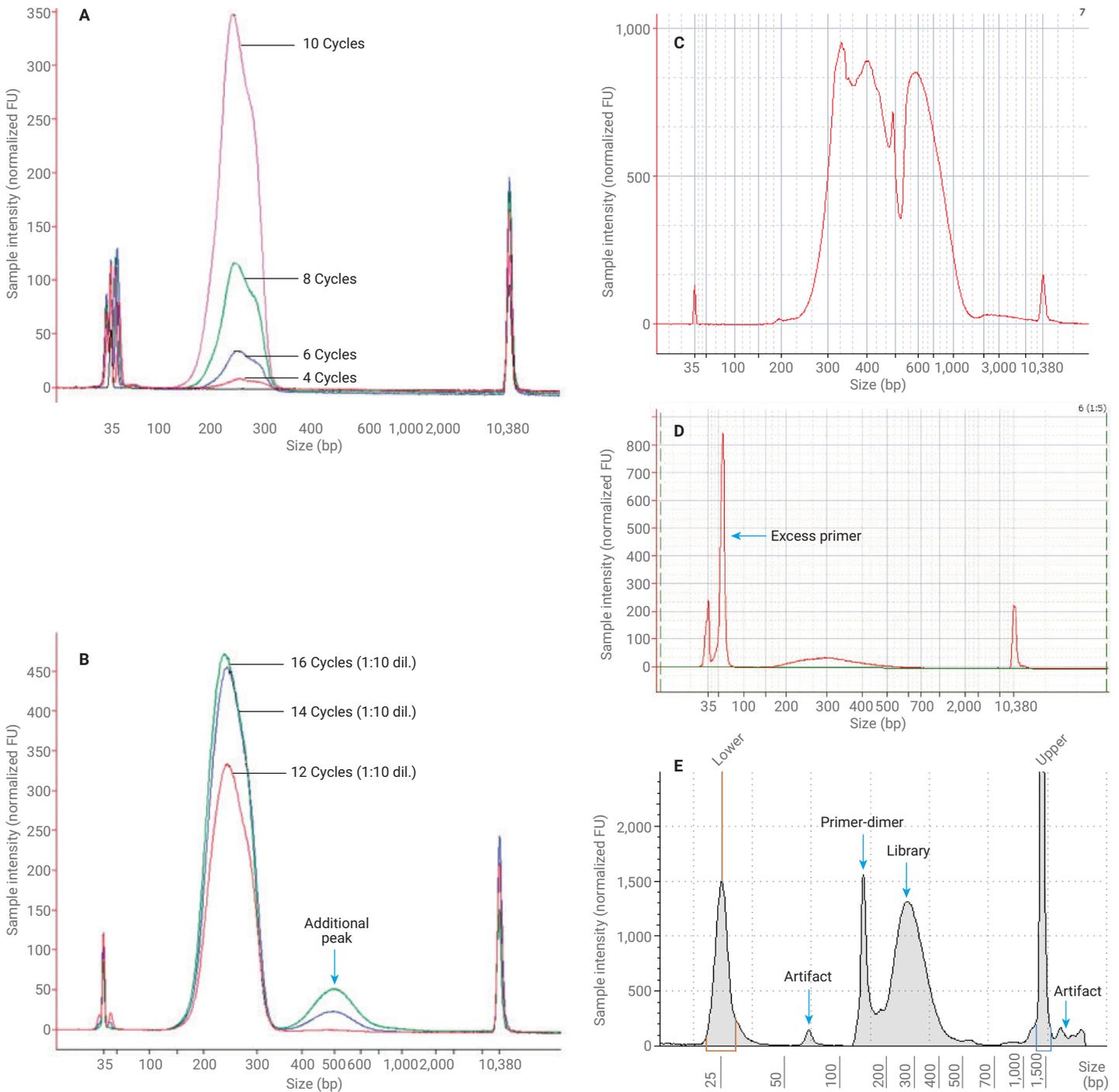


Figure 8. Electropherogram pattern of artifacts in PCR-amplified DNA libraries derived from the SureSelect Target Enrichment workflow. A to D show electropherograms obtained with the HS DNA assay and the Bioanalyzer system. A) Overlay of electropherograms after 4 to 10 PCR cycles as well as TE buffer blank (black) showing progressive increase in DNA concentration with higher cycle number. B) Overlay of electropherograms obtained after 12 to 16 PCR cycles. When using 10 or more cycles, the DNA concentration is outside the quantitative range of the HS DNA assay. Therefore, samples should be diluted with TE buffer in the indicated dilution ratios prior to the analysis on the Bioanalyzer system. Loss of linear increase of DNA concentration, saturation of DNA, and additional peak seen at 14 and higher cycles. C) Split peaks from overloading the Bioanalyzer system. D) Sample showing excess primer and low yield. E) Library analyzed with the HS D1000 ScreenTape assay and the TapeStation system showing primer-dimers and PCR artifacts. Primer-dimers show up as a sharp peak below the expected library size, and artifacts can be found throughout the trace.

Reducing the PCR cycle number can help prevent spurious peaks and artifacts from overamplification. Inefficient PCR cycling may result in low yield and an excess of primers (Figure 8D). This can be a result of poor adapter ligation, low DNA quality, inefficient bead cleanup, use of too few cycles, or a thermocycler that is not well calibrated. Figure 8C shows that overloading DNA on Bioanalyzer chips can cause a split peak.

The last QC point is to determine the size distribution of the final library, that is the amplified post captured library, because sequencing those that do not fall within the recommended size distribution may lead to poor enrichment of targeted regions. These libraries are a result of the hybridization of the amplified precaptured library with a pool of probes, followed by a subsequent capture by streptavidin-coated magnetic beads. These bound hybridized pairs are then washed through a series of various stringency washes to remove nonspecific sequences, and are then amplified. This amplification serves to ensure that a sufficient amount of library has been generated for sequencing, and in the case of SureSelect^{XT}, to add sample index for multiplexing. Adding the indexes at this point results in an additional 20–30 bp shift in the sample peak (Figure 9).

Issues similar to the primer dimerization and over-amplification described above can be observed at this step too. Additionally, bead carryover from post PCR DNA cleanup can impact successful assessment of the libraries. Residual beads can comigrate with the upper marker, and may affect sizing and quantitation of samples (Figure 10). Using a strong magnet for bead separation and pipetting carefully to avoid disturbance of the beads is helpful in such situations. Low yields after post capture PCR may indicate suboptimal hybridization conditions,

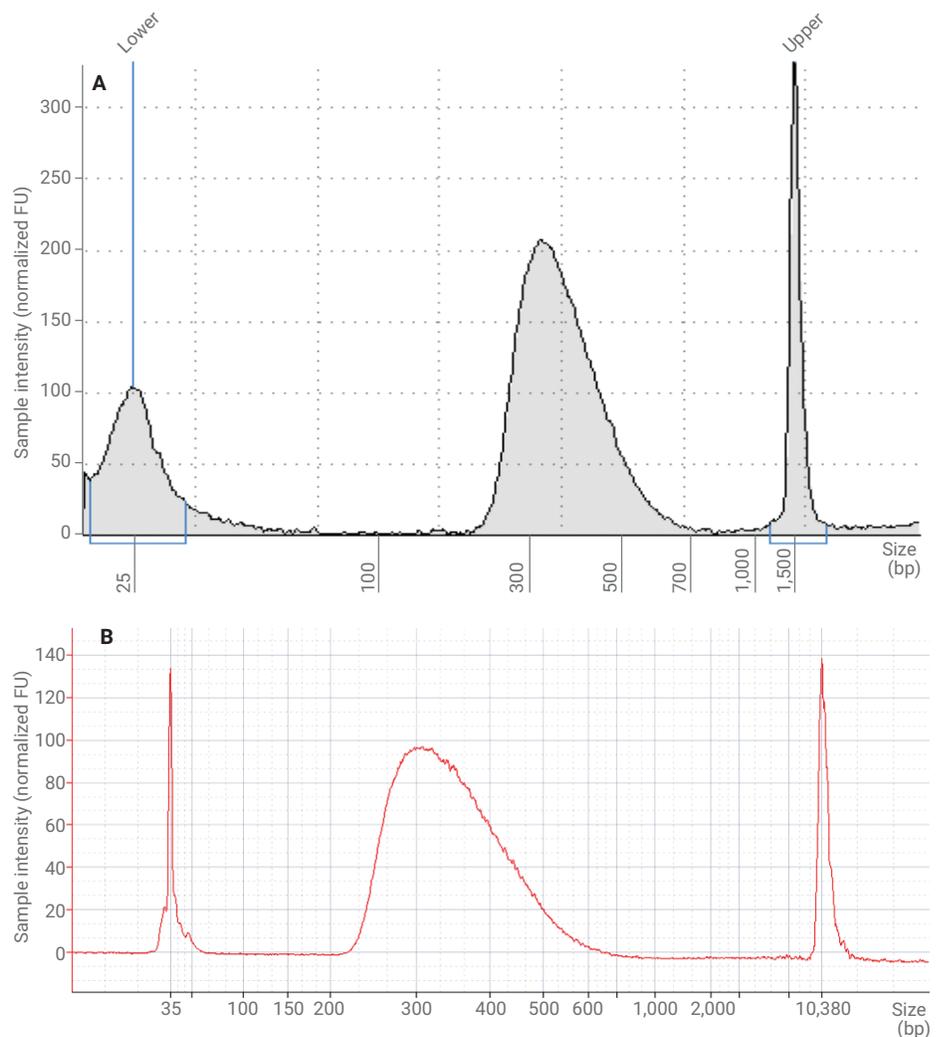


Figure 9. Electropherogram pattern of SureSelect^{XT} and SureSelect^{XT2} amplified post capture libraries showing maximum peak size between 250 and 350 bp. A) Post captured amplified DNA library separated with the TapeStation HS D1000 ScreenTape assay and the TapeStation system. B) Post captured amplified DNA library analyzed with the HS DNA assay and the Bioanalyzer system.

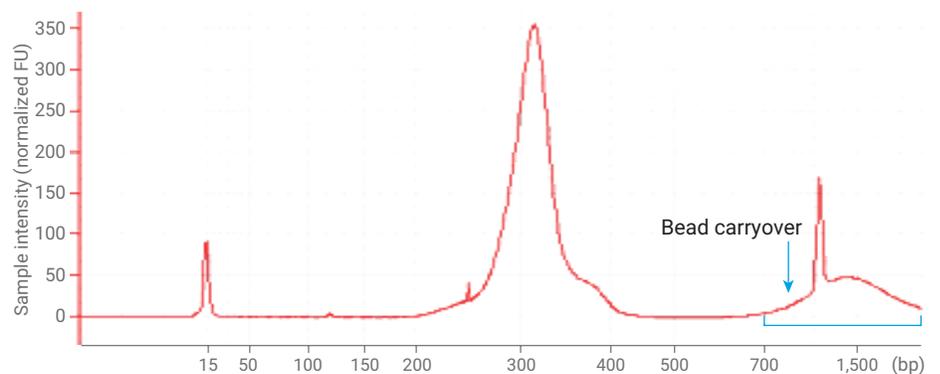


Figure 10. Electropherogram pattern of post capture library from SureSelect Target Enrichment workflow showing magnetic bead carryover, obtained with the DNA 1000 assay and the Bioanalyzer system.

along with other reasons discussed earlier (Figure 8D). It is useful to ensure that excessive evaporation has not occurred during the hybridization step, and proceed with sequencing only when sufficient DNA is present. Users often pool library samples at this stage based on molar concentrations for sequencing applications. Data obtained with the Bioanalyzer and TapeStation systems provide information on molarity along with size distribution, and can be used to estimate percent primer-dimer, which are critical parameters for assessment of final library.

SureSelect^{QXT}

SureSelect^{QXT} is a transposase-mediated library preparation method (Figure 1B), which also has three main QC checkpoints such as SureSelect^{XT} and SureSelect^{XT2}. DNA integrity is essential for transposase-based library preparation methods due to the need of the transposase to cut and ligate at two sites on a single strand of DNA. SureSelect^{QXT} is not recommended for DNA of poor integrity such as DNA from FFPE samples. For SureSelect^{QXT}, a DIN >8 is recommended for gDNA inputs, as represented in lanes E2 to H2 of Figure 2A. The QC point after precapture PCR will assess the success of the transposase reaction, as demonstrated in Figures 11A and 11B. Variation of fragmentation profiles may also be observed, as shown in Figure 11C. Peaks appearing at sizes lower than 245 bp may indicate a lower input amount than estimated, and may be associated with increased duplicates in the sequencing data. In contrast, a peak DNA fragment size significantly greater than 325 bp may indicate too much gDNA in the fragmentation reaction, and may be associated with decreased percent-on-target performance in sequencing results (Figure 11C)⁶.

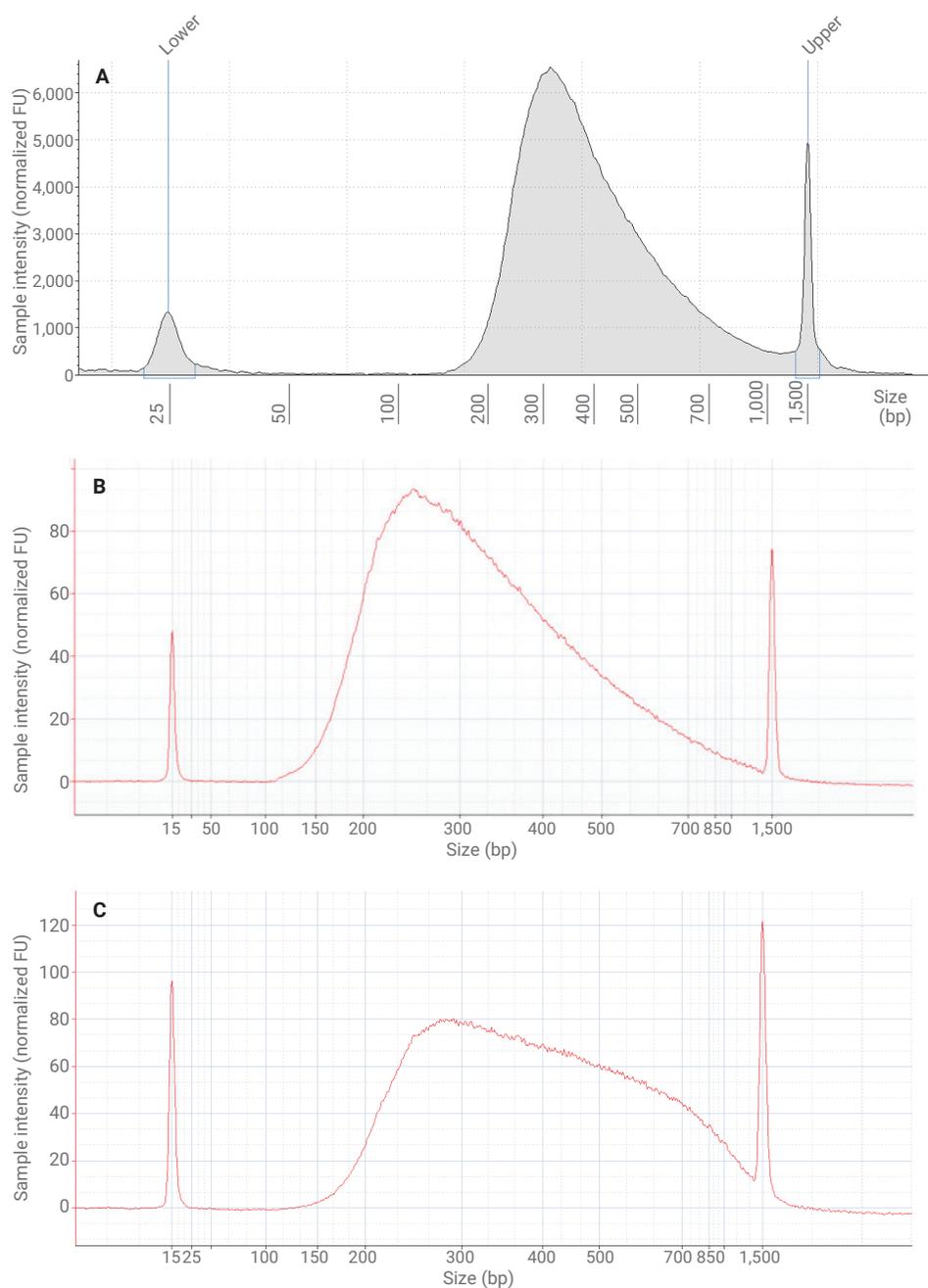


Figure 11. Electropherogram pattern of SureSelect^{QXT} amplified precapture libraries showing maximum peak size between 245 and 325 bp. A) Precapture amplified DNA library separated with the D1000 ScreenTape assay and the TapeStation system. B) Precapture amplified DNA library analyzed with the DNA 1000 assay and the Bioanalyzer system. C) Variation in precapture amplified DNA library with DNA 1000 assay.

In a successful experiment, the final library shows a peak between 325 bp and 450 bp (Figure 12). Similar to SureSelect^{XT} and SureSelect^{XT2} workflows, besides ensuring sufficient yield for sequencing, examine the sample for signs of primer-dimers and over-amplification.

HaloPlex and HaloPlex^{HS}

The HaloPlex and HaloPlex^{HS} target enrichment technology uses an amplicon-based approach, unlike the random fragmentation used in the SureSelect workflows. The initial human gDNA sample is individually digested by eight unique pairs of restriction enzymes generating fragments (amplicons) with expected sequences. These fragments are subsequently pooled and combined with probes designed to hybridize to the ends of their target fragments, forming a circular structure known as a Halo (Figure 13). Then, the indexing cassette, carrying the index sequence and partial adaptor sequence, is guided by the probe to ligate between the ends of the target fragment. This generates a ring fragment as the template for PCR, producing the amplified target amplicon with full adaptor sequences.

Integrity of the input DNA is key to the success of HaloPlex and HaloPlex^{HS} library preparation, which relies on the cutting of intact gDNA to generate target fragments. The use of high-quality gDNA samples verified by OD 260/280 ratio and size distribution on gel electrophoresis as stated in protocol is required. The use of FFPE DNA is only recommended after incorporating modifications based on the ddCq quality score obtained using the Agilent NGS FFPE QC kit (G9700-90000).

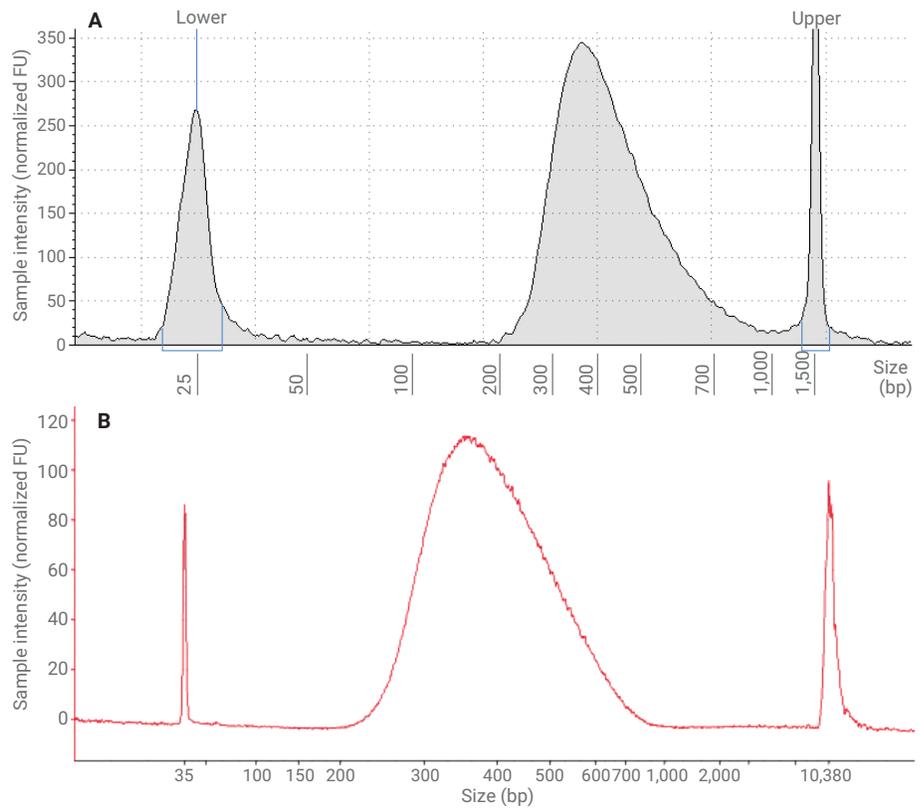


Figure 12. Electropherogram pattern of SureSelect^{QXT} amplified post capture libraries showing maximum peak size between 325 and 450 bp. A) Post capture amplified DNA library separated with the HS D1000 ScreenTape assay and the TapeStation system. B) Post capture amplified DNA library with the HS DNA assay and the Bioanalyzer system.

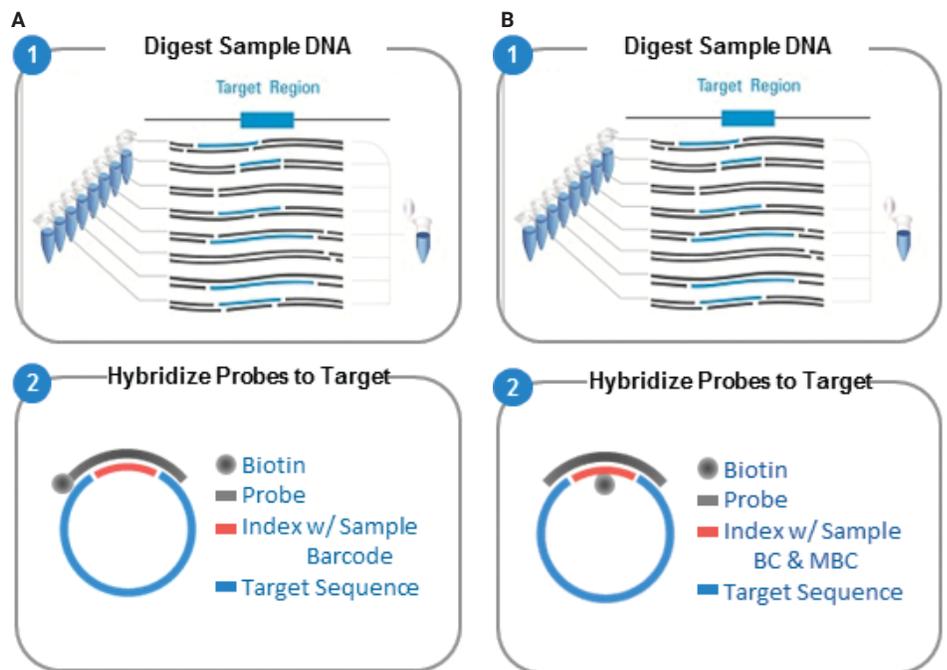


Figure 13. Schematic of a Halo structure from HaloPlex (A) and HaloPlex^{HS} (B) workflows formed by target DNA fragment, probe, and index sequence.

The restriction digestion of gDNA produces a smear, making it difficult to judge the success of digestion. Therefore, an Enrichment Control DNA (ECD) sample is provided to be digested along with experimental samples to verify the activity of the restriction enzymes. The ECD control is a mixture of high-quality gDNA and

an 800 bp PCR product that contains restriction sites for all the enzymes used. To validate the restriction digestion (Figure 14), an aliquot of each of the eight double-digests of the ECD control, along with an aliquot of undigested ECD control, is run on the TapeStation or Bioanalyzer system.

Partial digestion results in a minor band, absence of any one of the expected bands, or the presence of a major undigested band, which may have an impact on the enrichment outcome (Figure 15).

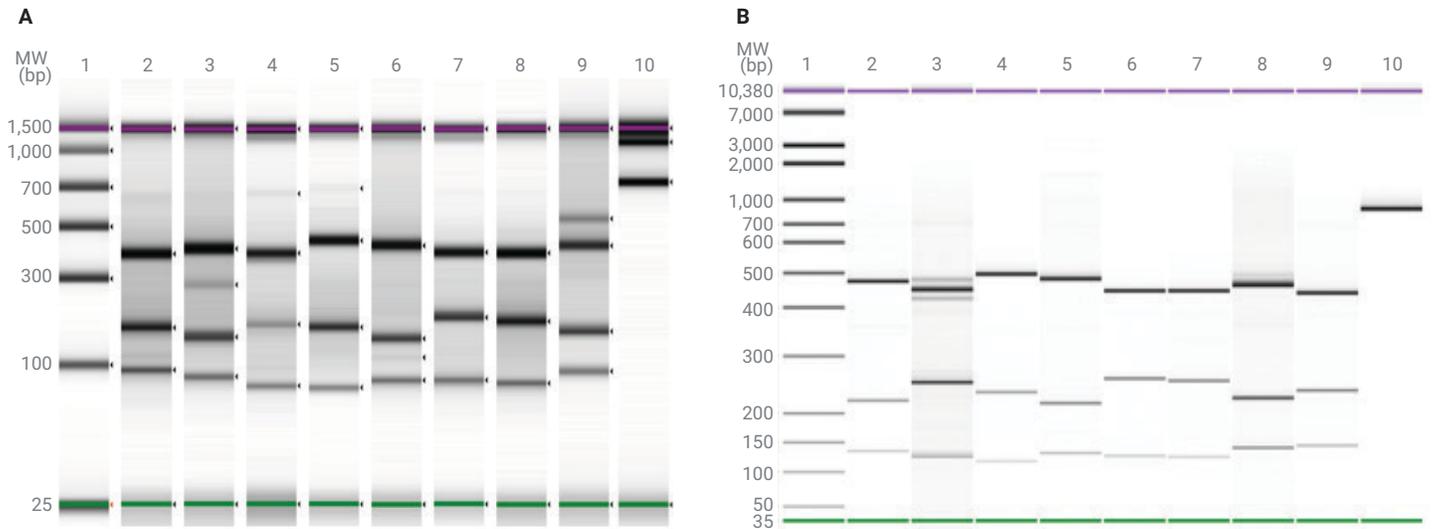


Figure 14. Gel images of ECD digest from HaloPlex Target Enrichment workflow. Lane 1: DNA ladder, Lanes 2–9: ECD digestion reactions A–H, Lane 10: Undigested Enrichment Control DNA. The successfully digested sample shows three prominent bands at approximately 125, 225, and 450 bp on top of a smear of gDNA between 100 and 2,500 bp. The precise sizes of the three bands will differ after digestion in each of the eight restriction enzyme master mixes. It is normal to observe additional minor bands with relative abundance similar to lane 3. A) ECD restriction digestion by the HS D1000 ScreenTape assay and the TapeStation system. B) ECD restriction digestion by the HS DNA assay and the Bioanalyzer system

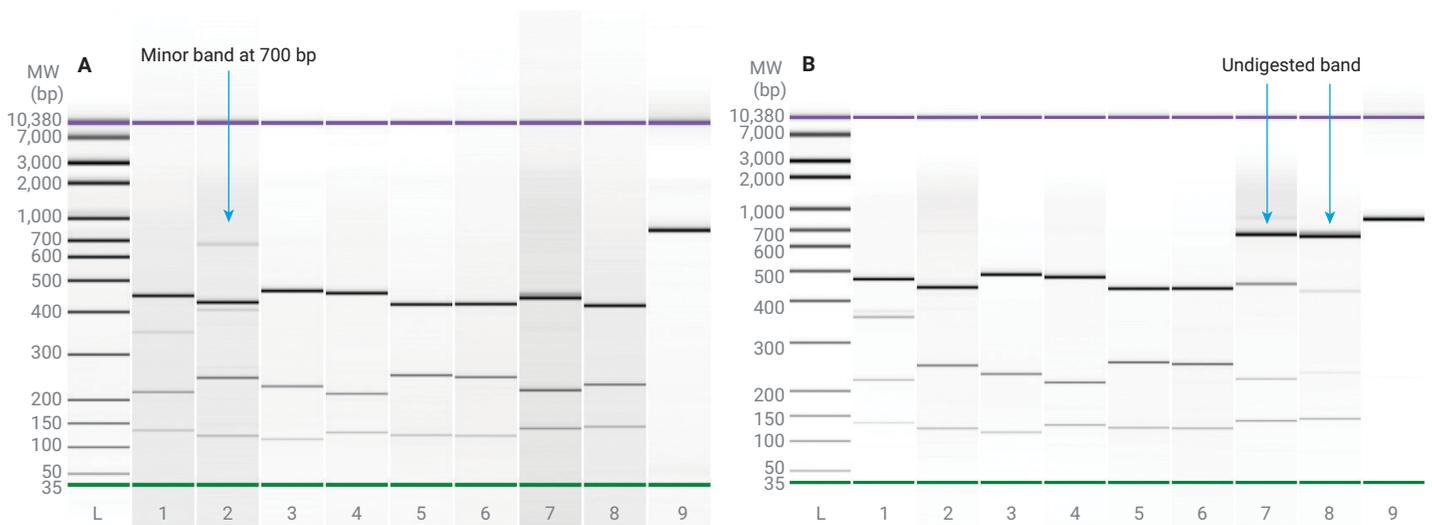


Figure 15. Gel images of partial ECD digest from HaloPlex Target Enrichment workflow analyzed with the HS DNA assay and the Bioanalyzer system. A) Minor band at 700 bp in lane 2. B) Major undigested band in lanes 7 and 8.

The final libraries of HaloPlex and HaloPlex^{HS} workflows show a profile with a characteristic smear in the range of 175 to 625 bp (Figures 16A and 16B) and 190 to 545 bp, respectively (Figures 16C

and 16D). The appearance of the profile may vary due to specific library designs as well as to the overall quality of the input material. Each new lot of a custom kit is provided with a Bioanalyzer library

trace image generated in-house as part of the QC process for specific HaloPlex and HaloPlex^{HS} custom kits, using a high-quality DNA sample.

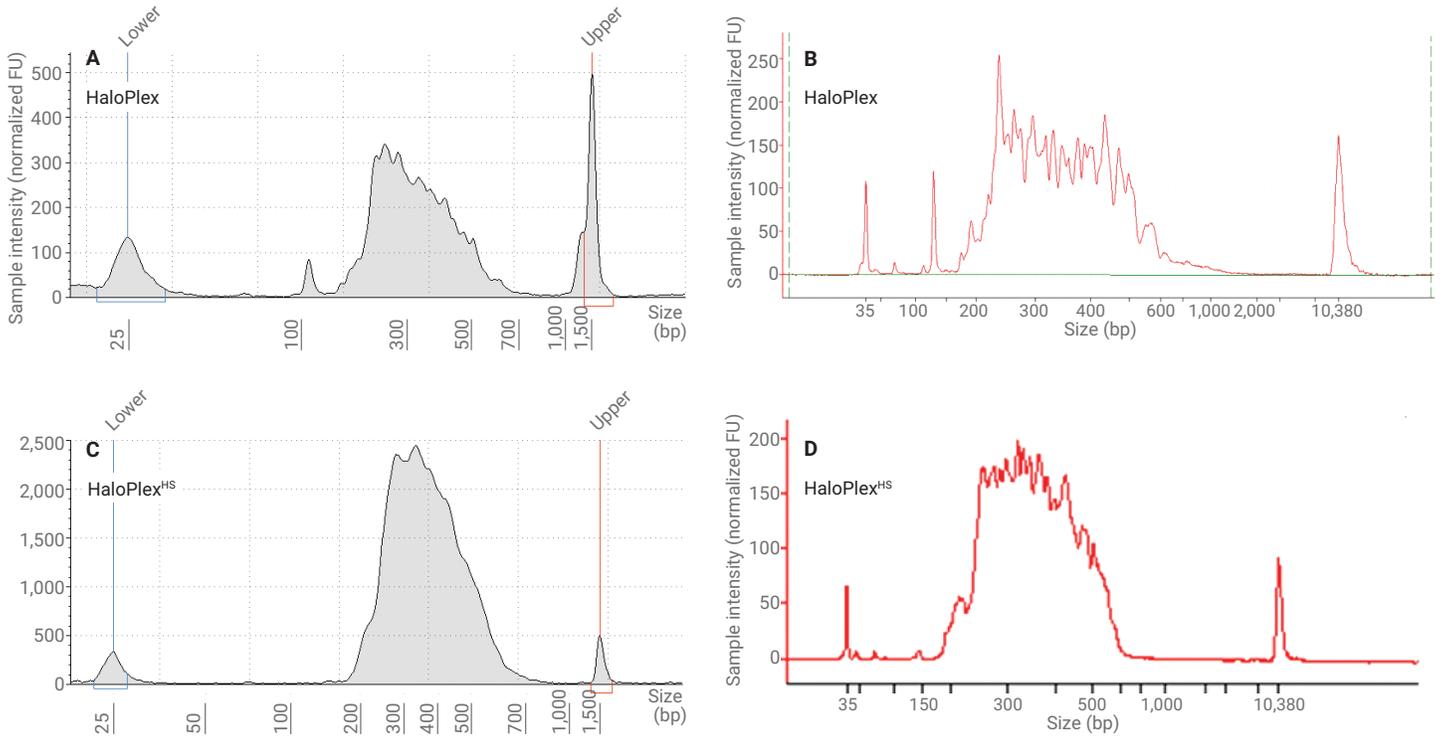


Figure 16. Electropherogram pattern of HaloPlex and HaloPlex^{HS} final libraries. (A) and (B) HaloPlex library pattern with smear in the range of 175 to 625 bp. (C) and (D) HaloPlex^{HS} library with smear in the range of 190 to 545 bp. The libraries were analyzed with the HS D1000 ScreenTape assay and the TapeStation system (A and C) or the HS DNA assay and the Bioanalyzer system (B and D).

A common issue seen in HaloPlex and HaloPlex^{HS} libraries is artifacts with sizes less than 150 bp (Figure 17). Artifact peaks of approximately 120 bp for HaloPlex and approximately 140 bp for HaloPlex^{HS} are related to primer-dimers that can cluster and consume

sequencing real estate, and should not be included in the quantitation of the library. If the peak is greater than 10 % of the total product generated, removal of the primer through an additional round of AMPure purification is highly recommended.

A sharp decline at the high molecular weight end of HaloPlex post capture libraries in comparison to the expected library pattern may be seen from poor-quality FFPE DNA samples, or user introduced inadequacies to the library preparation protocol (Figure 18).

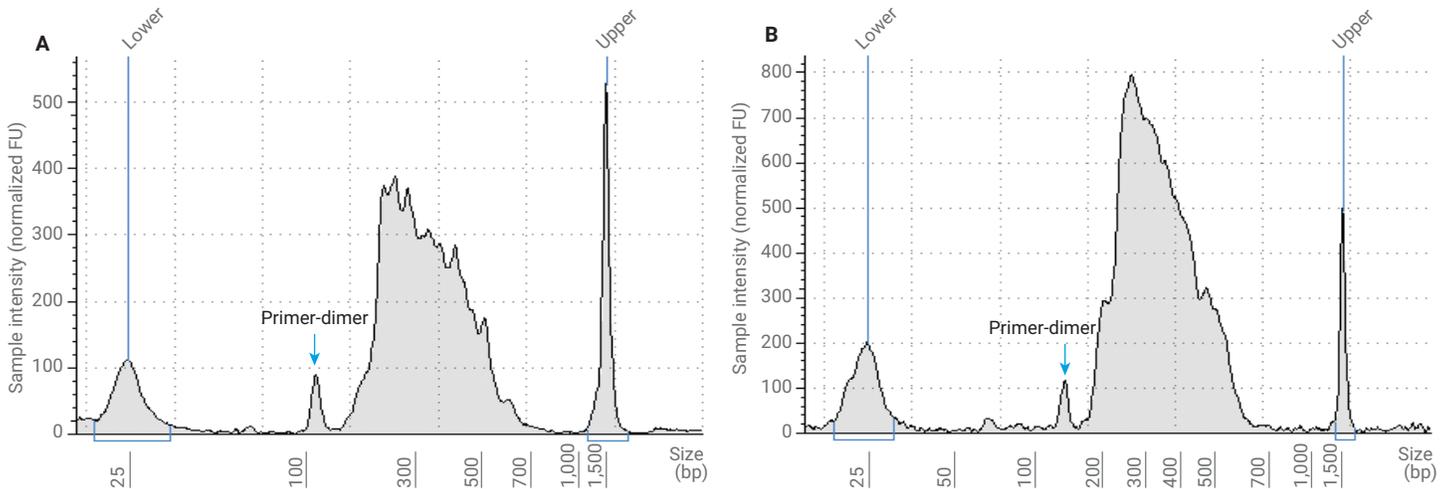


Figure 17. Electropherogram pattern of primer-dimers in HaloPlex and HaloPlex^{HS} final library obtained with the HS D1000 ScreenTape assay and the TapeStation system. A) HaloPlex library showing primer-dimer at 120 bp. B) HaloPlex^{HS} library with primer-dimer at 145 bp.

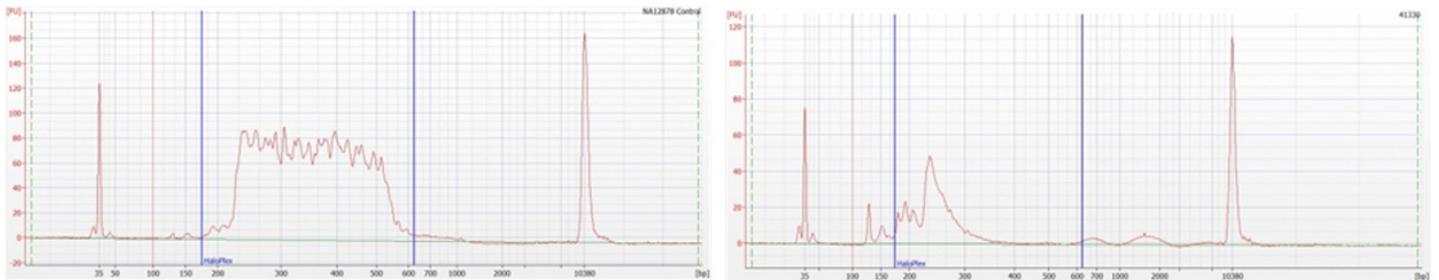


Figure 18. Electropherogram pattern of FFPE DNA library from HaloPlex target enrichment protocol analyzed with HS DNA assay and the Bioanalyzer system. A) Normal FFPE DNA library profile. B) Sharp decline at high molecular weight seen from poor-quality FFPE samples, improper mixing of individual digested reactions into the hybridization solution leading to inadequate inactivation of restriction enzymes, or over-drying of AMPure beads at the time of hybridization buffer removal in the HaloPlex^{HS} workflow.

Conclusion

To ensure successful NGS library preparation, it is important to monitor samples at various steps from acquisition to sequencing library. Knowing the expected outcome at different QC steps provides users a reference point and confidence in generating high-quality libraries throughout the process. This is key to securing reliable sequencing data. Additionally, it is important to understand the potential deviations and their causes, which will facilitate rapid assessment and rectification. This Application Note addresses the above, and demonstrates the utility of two automated electrophoretic platforms, TapeStation and Bioanalyzer systems, at various QC check-points throughout the Agilent NGS product portfolio. Note that these representations are not all-encompassing, and it is possible to observe results different from those discussed here.

References

1. A. Padmanaban, End to End Sample Quality Control for Next Generation Sequencing Library Preparation and SureSelect Target Enrichment on the Agilent 2200 TapeStation System, *Agilent Technologies Application Note*, publication number 5991-3654EN, **2014**.
2. M. Gassman, B. McHoull, DNA Integrity Number (DIN) with the Agilent 2200 TapeStation System and the Agilent Genomic DNA ScreenTape Assay, *Agilent Technologies Technical Overview*, publication number 5991-5258EN, **2014**.
3. J. Petersen, *et al.* 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**.
4. A. Padmanaban, DNA Integrity Number (DIN) For the Assessment of Genomic DNA Samples in Real-Time Quantitative PCR (qPCR) Experiments, *Agilent Technologies Application Note*, publication number 5991-6368EN, **2015**.
5. K. Gromadski, *et al.* Improving sample quality for SureSelect target enrichment and next-generation sequencing with the High Sensitivity DNA kit, *Agilent Technologies Application Note*, publication number 5990-5008EN, **2009**.
6. E. Graf, Quality Control for Agilent SureSelect^{QXT} WGS Library Preparation, *Agilent Technologies Application Note*, publication number 5991-8191EN, **2017**.

www.agilent.com/chem

For Research Use Only. Not for use in diagnostic procedures.

This information is subject to change without notice.

© Agilent Technologies, Inc. 2018
Printed in the USA, December 1, 2018
5994-0127EN

