Application Note

Genomics

Agilent

Trusted Answers

Preventive Control of Sequencing Through the Insert with the Agilent 5200 Fragment Analyzer System

Authors

Vera Rykalina and Kyle Luttgeharm, Agilent Technologies

Abstract

The Agilent Fragment Analyzer systems, together with the Agilent NGS Fragment kits and the companion Agilent ProSize data analysis software, offer a well-established solution for sample quality control (QC) in next-generation sequencing (NGS) workflows. This application note expands the capabilities of the ProSize smear analysis and demonstrates how its functionality can be further applied to avert a specific sequencing issue called sequencing through the insert. In this instance, the insert length is shorter than the optimal insert size resulting in a portion or the entire adapter being sequenced along with the insert. This data is then introduced into the sequencing read and in some cases can increase the background noise decreasing the quality of the overall sequencing run. The %Total value of the ProSize software allows a user to estimate the percentage of the library that will be sequenced through the insert based on the planned sequencing run method and electrophoretic profile of the library. Easy and comprehensive smear analysis enables the user to determine an optimal read length and thereby minimize the number of unwanted bases present in the sequencing reads, saving expensive reagents and time for auxiliary data processing.

Introduction

Library preparation in major NGS workflows is a procedure required to convert nucleic acid samples of interest into a platform-specific format. This generally includes fragmentation of input material, ligation of known sequences (adapters) to the resulting fragments (inserts), and amplification of the ligated constructs in the case of PCR-based protocols. Thus, a final library is multiple inserts of various length flanked by adapters meant to act as priming points for the sequencing chemistry reactions. Analysis of a library using an electrophoretic profile is an essential step in a library preparation process[1]. Quality control based on electrophoretic traces not only ensures that an obtained size distribution of the library meets the application requirements, but also enables the user to make an informed decision regarding which sequencing run method to be used.

When designing a sequencing experiment, it is important to consider both the library insert size and the desired sequencing read length. Ideally, the fragments composing the library should be optimized to an insert size greater than the anticipated sequencing read length (Figure 1A). Sequencing past the optimal insert length can result in a so-called sequencing through the insert event where a portion or the entire adapter is sequenced and data is introduced into the read (Figure 1B). These adapter sequences should be removed bioinformatically as they dramatically impact downstream data processing such as sequencing alignment or *de novo* assembly. Sequencing through the insert can have even more severe outcomes when the insert size is extremely short. In this case, the sequencing reaction will continue beyond the adapter and into the surface of the flow cell. The imaging system will be detecting noise from the unincorporated bases that will cause low-quality basecalls at the end of the read (Figure 1C).

To prevent adapter read through and thereby maximize quality of sequencing data, careful analysis of the library sizing profile should be considered. This application note describes how the electrophoretic QC of the final library and assessment of critical size ranges based on a desired run method can be performed using the Agilent Fragment Analyzer system and the Agilent ProSize data analysis software. In pursuit of a comprehensive demonstration of the QC capabilities of the ProSize smear analysis in relation to sequencing through the insert, several hypothetical sequencing scenarios are discussed in this study.

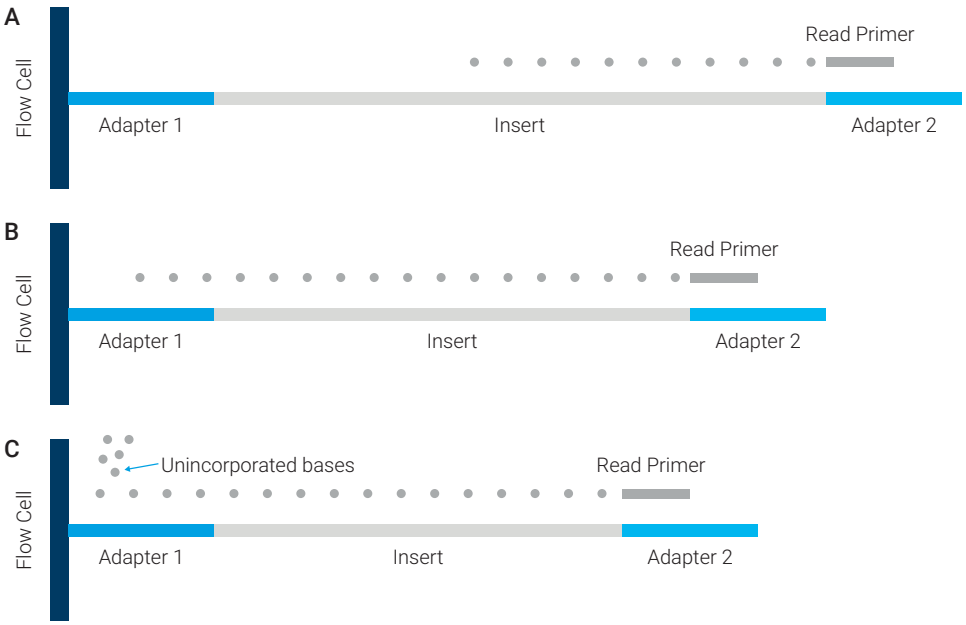

Figure 1. Schematic representation of sequencing reaction scenarios. (A) Optimal read length. (B) Sequencing through the insert. (C) Sequencing through the adapter.

Experimental

Sample preparation

The NGS library used in this study was constructed by a standard library preparation protocol. This included shearing of input genomic DNA (100 ng), ligation of adapters to fragmented DNA molecules, and amplification of the ligated material to achieve a desired concentration. The post-amplified sample was subjected to 1X bead-based cleanup and eluted in 10 mM Tris-HCl (pH 8.0). The final library was quantified spectrofluorometrically and used to prepare a set of two-fold serial dilutions.

Smear analysis

The Agilent 5200 Fragment Analyzer system (p/n M5310AA) and the Agilent HS NGS Fragment kit (1-6000 bp) (p/n DNF-474) were used for electrophoretic analysis of the dilution series. Sample preparation was carried out according to the Agilent quick guide instructions[2]. Smear analysis of all samples was performed using the Agilent ProSize data analysis software (version 4.0.0.3).

Estimation of optimal and extremely short sizes for a library

Optimal and extremely short sizes for an individual library can be easily defined using an intended read length and the lengths of the adapters used. Table 1 summarizes these sizes for three most used sequencing run methods.

Table 1. Computation of the critical sizes for a final library.

Number of Cycles	Optimal Insert Size	Optimal Final Library Size (Optimal Insert + Ad 1* + Ad 2*)	Extremely Short Insert (Optimal Insert - Ad 1*)	Extremely Short Final Library Size (Extremely Short Insert + Ad 1* + Ad 2*)
150	>150 bp	>286 bp	≤84 bp	≤220 bp
250	>250 bp	>386 bp	≤184 bp	≤320 bp
300	>300 bp	>436 bp	≤234 bp	≤370 bp

**Assumes adapter 1 (Ad 1) = 66 bp and adapter 2 (Ad 2) = 70 bp.*

Results and discussion

Smear analysis

The Fragment Analyzer together with the ProSize data analysis software allow for a convenient and reliable smear analysis. The results can be obtained as an electropherogram, digital gel image, and smear analysis table. Figure 2 illustrates an example of the region setting for a library analyzed with the HS NGS Fragment kit. The region was established based on the library electropherogram trace using the ProSize functionality. In the example used, a range of 110 to 1,350 bp was set to cover the entire distribution of the library. The smear analysis determined that the library has an average size of 333 bp. In addition to average size, concentration, molarity, and %CV, the results table provides a %Total value that corresponds to the percentage of the sample distributed within the specified region. As shown in Figure 2, 99% of the sample is covered by the defined region relative to the entire trace area.

ID	Range	ng/µL	% Total	nmole/L	Avg. Size	%CV
A1: S1 - 4 ng/µL	110 bp to 1350 bp	4.6470	99.0	22.9669	333	23.16

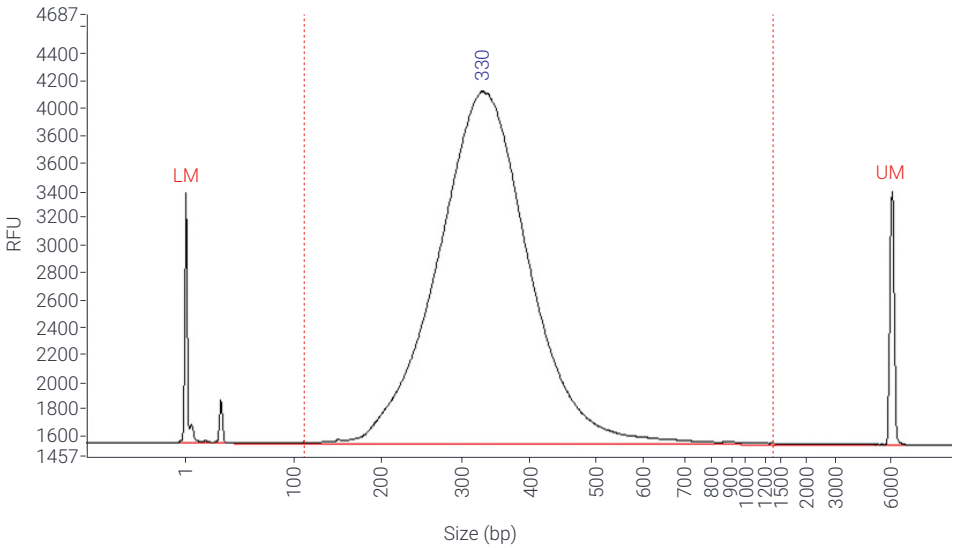

Figure 2. Smear analysis of the library performed with the Agilent ProSize data analysis software. The library distribution starts at 110 bp and extends to 1,350 bp. The average size determined by smear analysis is 333 bp. Considering that adapters contribute about 136 bp to the final library size, the average insert length is 197 bp.

Assessment of final library sizing profile

Electrophoretic quality control steps are highly recommended during library preparation. The steps not only confirm that sample quality and quantity meet the requirements specified by the protocol but they also reveal and address potential issues prior to downstream analysis. One such issue is sequencing through the insert, which can happen when the majority of the library products have an extremely short insert size for a desired read length. As was shown in Table 1, the optimal size of a final library as well as determining an insert that is too short can be easily derived based on actual length of both adapters and an intended length of the sequencing read.

To demonstrate how ProSize smear analysis can be applied to determine a proper read length for sequencing based on the library size, a sequencing ready NGS library was analyzed on the 5200 Fragment Analyzer system with the HS NGS Fragment kit. Assuming a standard 2 × 150 bp sequencing run, several critical smear ranges were assessed for this library.

As was mentioned earlier, a preferable sequencing outcome is achieved when an insert size is greater than the length of the desired read (insert size > 150 bp). If applied to a final library, this size will be increased by 136 bp, pertaining to the total length of both adapters (66 bp + 70 bp). Setting the smear range in ProSize from 286 to 1,350 bp provides a %Total value representing the portion of the library where only the insert is sequenced (Figure 3A, green area). For this particular library, the %Total is 73.7%. By shifting the left boundary of this region to 220 bp, a user can estimate the library fraction (20.9%; Δ %Total = $\%Total_{220\text{-}1350}$ - $\%Total_{286\text{-}1350}$) sequenced through the insert and partially into the adapter region (Figure 3B, orange area). Moving the left boundary further to 110 bp, represents a negligible portion of the library (4.1%; Δ %Total = $\%Total_{110\text{-}1350}$ - $\%Total_{220\text{-}1350}$) that has an extremely short insert size for a 150 bp read length (Figure 3C, red area). DNA fragments in this region would result in sequencing through the entire adapter and incorporating unspecific base pairs, as represented in Figure 1C. Overall, this example library would be a good fit for a 150 read length based on the fact that 73.7% of the inserts are longer than the read length and would result in quality sequencing data.

ID	Range	ng/µL	% Total	nmole/L	Avg. Size	%CV
A2: S1 - 2 ng/µL	110 bp to 1350 bp	2.3296	98.7	11.5454	332	23.30
	220 bp to 1350 bp	2.2339	94.6	10.8748	338	21.67
	286 bp to 1350 bp	1.7412	73.7	7.9547	360	18.82

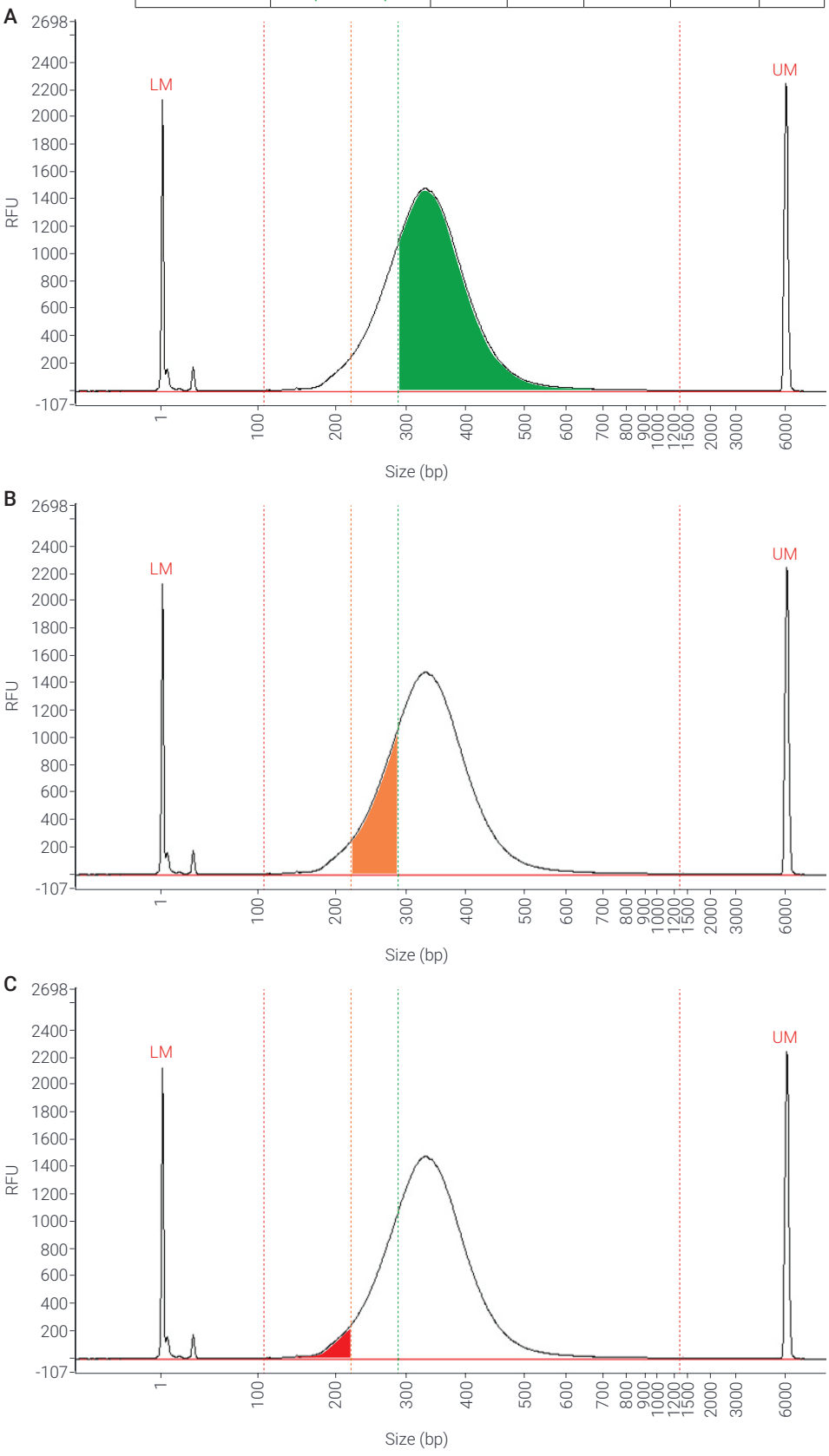

Figure 3. Electrophoretic analysis of the final library. Multiple ranges were set to assess a portion of the library (A) with an optimal insert size, (B) sequencing through the insert, and (C) through the adapter region. A 2 × 150 bp run setup is the optimal read length for this library.

Smear analysis of a library dilution series

Consistency and reliability of the ProSize smear analysis feature throughout the concentration range of the HS NGS Fragment kit was demonstrated by varying the library concentration and assessing the percentage of the library that could be sequenced through the adapter. Serial dilutions of the library were prepared and analyzed using the Fragment Analyzer system and the HS NGS Fragment kit. Smear analysis was performed by setting identical ranges for each library concentration and assuming a 2 × 150 bp sequencing run as in the previous section. For this sequencing run, DNA fragments plus adapters smaller than 220 bp in total length, will have an extremely short insert of < 84 bp after taking into account the adapter lengths, resulting in sequencing beyond the adapter region (Table 1). As reported in the result table generated for each concentration, the %Total remained consistent throughout the dilution series, predicting about 4% of the library that will be sequenced through the adapter introducing unincorporated bases (Figure 4).

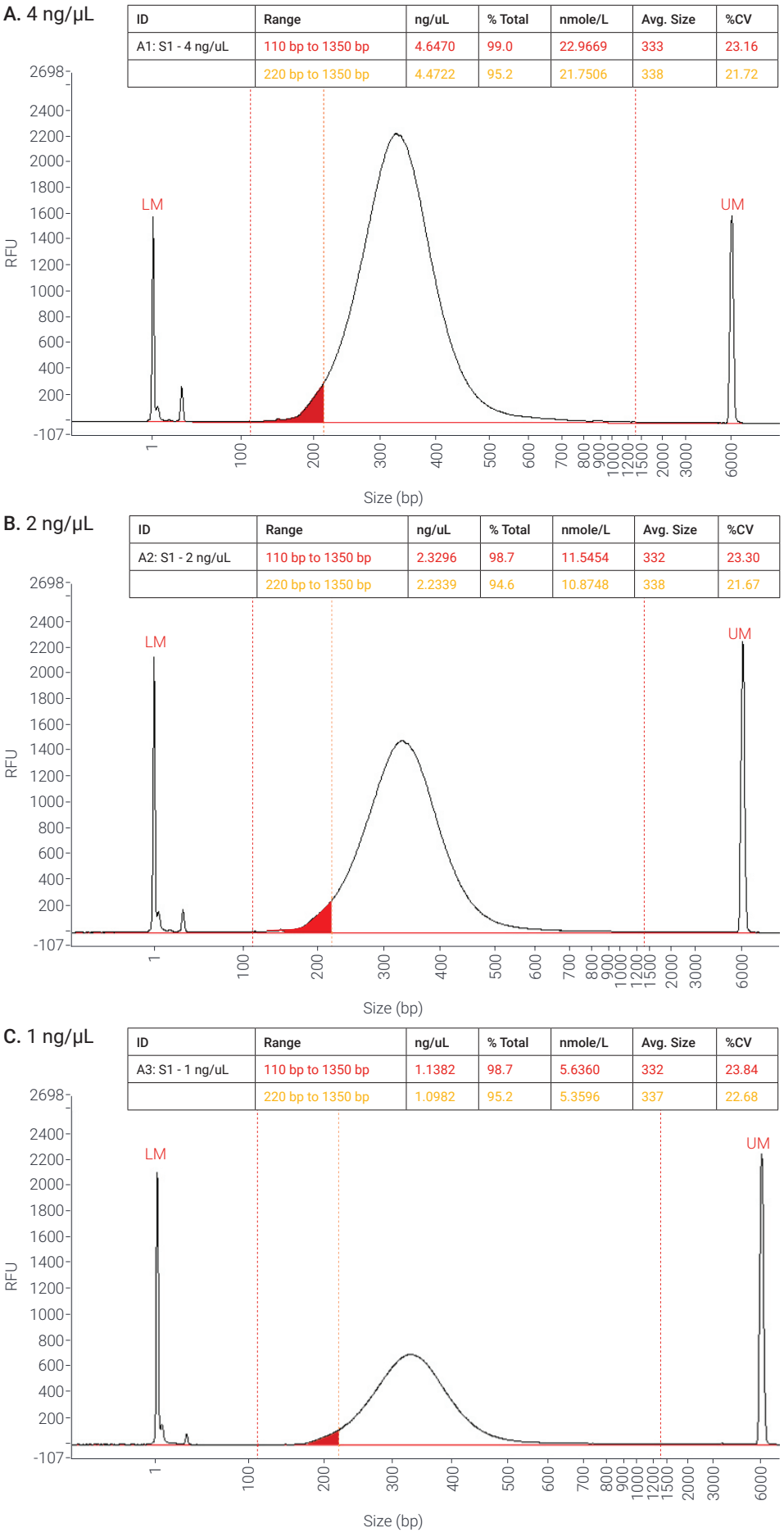

Figure 4. Smear analysis of the dilution series. Library concentrations are equal to (A) 4 ng/µL, (B) 2 ng/µL, and (C) 1 ng/µL, respectively.

Sequencing through the insert in relation to the read length

Sequencing through the insert should be a serious concern especially with longer read lengths. The higher the number of sequencing cycles to be performed, the more important it is to check a library for the proper insert length during routine QC processes. Figure 5 shows how the portion of the library sequenced through the adapter increases along with the number of the intended sequencing cycles or read length. For example, for a hypothetical run with the shortest read length of 150 bp, only 4.1% of the library would have an extremely short insert size where sequencing can run into the flow cell (Figure 5A, orange area). As this is just a negligible portion of a sample, it is unlikely to affect the overall sequencing data. However, if sequencing was performed using a 250 bp run, about 45% of library products would result in sequencing through the insert and past the adapter (Figure 5B, green area). Finally, for a long 300 bp run, 73.4% of the library would be classified as having extremely short insert size of 234 bp or less (Figure 5C, blue area). Based on the smear analysis, it is reasonable to conclude that for this particular library, a run with a read length of 150 bp is the most suitable sequencing method.

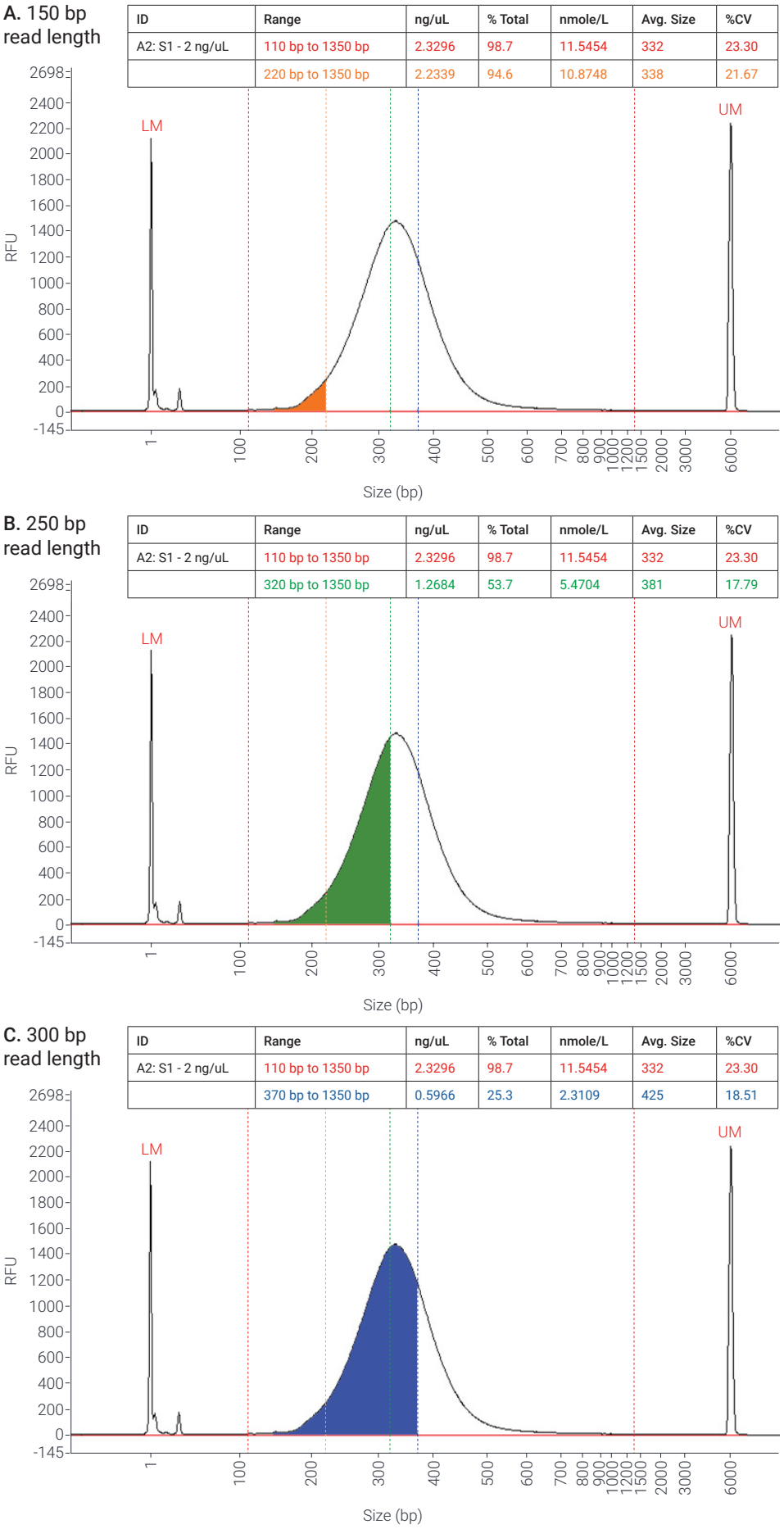

Figure 5. Library fractions representing extremely short insert sizes depending on the read length selected: (A) 150 bp, (B) 250 bp, and (C) 300 bp respectively.

Conclusion

When designing a sequencing experiment, it is important to select an appropriate sequencing run method. In fact, a suitable read length for a given library can ensure that only the insert is sequenced, thus minimizing the efforts to remove unwanted adapter bases from a resulting read. This application note highlighted the sample QC capabilities of the Agilent 5200 Fragment Analyzer system and ProSize smear analysis as a useful tool to avert sequencing through the insert. The %Total value of the smear analysis allows a user to estimate the percentage of the library with extremely short inserts for an intended run and adjust a read length accordingly, avoiding sequencing through the insert. Accessory analysis to confirm whether an electrophoretic profile of the library corresponds to a desired run method as a part of the routine sample QC operations enables optimal sequencing decision making, thereby saving time and resources on the way to successful sequencing results.

References

1. Monitoring Library Preparation for Next-Generation Sequencing in Systems Biology Omics *Analysis. Agilent Technologies application note*, publication number 5994-0946EN, **2019**.
2. Agilent DNF-474 HS NGS Fragment Kit Quick Guide for Fragment Analyzer Systems. *Agilent Technologies*, publication number SD-AT000134, **2021**.

www.agilent.com/genomics/fragment-analyzer

Printed in the USA, August 15, 2021
5994-3985EN

Agilent
Trusted Answers