White Paper

Agilent
Trusted Answers

The Importance of Quality Control in NGS Library Preparation Workflows: A Review of Applications using the Agilent Automated Electrophoresis Portfolio

Author

Whitney Pike, Chava Pocernich, and Steve Siembieda
Agilent Technologies, Inc.

Abstract

Next-generation sequencing (NGS) is an essential tool in molecular biology laboratories for the analysis of nucleic acid samples in numerous basic, translational, and clinical research settings. Preparation of sample libraries is a critical step of the NGS workflow, and can be a time-consuming, labor-intensive, and costly process. Quality control (QC) of input samples, at intermediate steps of the preparation process, and of the final libraries before sequencing, can help save time and resources by identifying samples that are of poor quality or of insufficient concentration to yield successful sequence data. Sample quality information can also aid in troubleshooting or optimizing library preparation protocols. The Agilent automated electrophoresis portfolio offers several instruments for QC analysis of nucleic acids, including the Bioanalyzer system, Fragment Analyzer systems, TapeStation systems, and Femto Pulse system. Ideal for determining the size, concentration, and molarity of samples throughout the NGS library preparation process, the automated electrophoresis instruments have been well cited throughout scientific literature.

Introduction

Today, NGS platforms leverage many sequencing technologies that use a variety of chemistries, enabling the sequencing of even the most complex samples. These technologies can be classified by the sequencing read length into short-read and long-read sequencing. Regardless of the type of sequencing, all NGS platforms depend upon the creation of a high-quality library to ensure successful sequencing results. Typically, an NGS library starts with a DNA or RNA sample of a specific size that then undergoes enzymatic treatments specific to the library preparation protocol and the sequencing platform. In general, the steps required for preparation of an NGS library include fragmentation of the input sample, ligation of a platform-specific sequencing adapter to the sample, and amplification of the library to produce enough material for sequencing. As the quality of the final sequencing results is highly dependent upon the quality of the library, it is advantageous to perform QC checks at various points throughout the library preparation workflow. QC checks are recommended prior to initiating the library preparation protocol, after fragmentation, adapter ligation, and post-enrichment or the final library (Figure 1). These QC checks give valuable insight into the sample size, concentration, and integrity. High-quality samples are necessary for successful sequencing. Performing QC can confirm that a sample is of sufficient quality to proceed with downstream analysis, or is of poor quality and would likely not yield robust results. Thus, QC checkpoints are crucial steps in the library preparation workflow, saving researchers valuable time and resources.

The automated electrophoresis portfolio from Agilent provides several instruments dedicated to quality control of samples, with reagent kits specialized for a variety of sample types throughout the NGS library preparation workflow. These instruments include the Bioanalyzer system, the Fragment Analyzer systems, the TapeStation systems, and the Femto Pulse system. Each instrument offers specific benefits suited to meet a variety of individual laboratory needs, such as throughput, sensitivity, speed, and resolution. Together, the automated electrophoresis instruments provide superior quantification, qualification, and sizing of nucleic acids to allow for confident assessment of samples throughout the NGS workflow. Visualization of integrity, contamination, and degradation of each sample is provided through electropherogram and digital gel images. This review discusses where QC steps can be easily implemented throughout the NGS library preparation workflow and provides examples of publications that have utilized the automated electrophoresis instruments at each specific QC checkpoint.

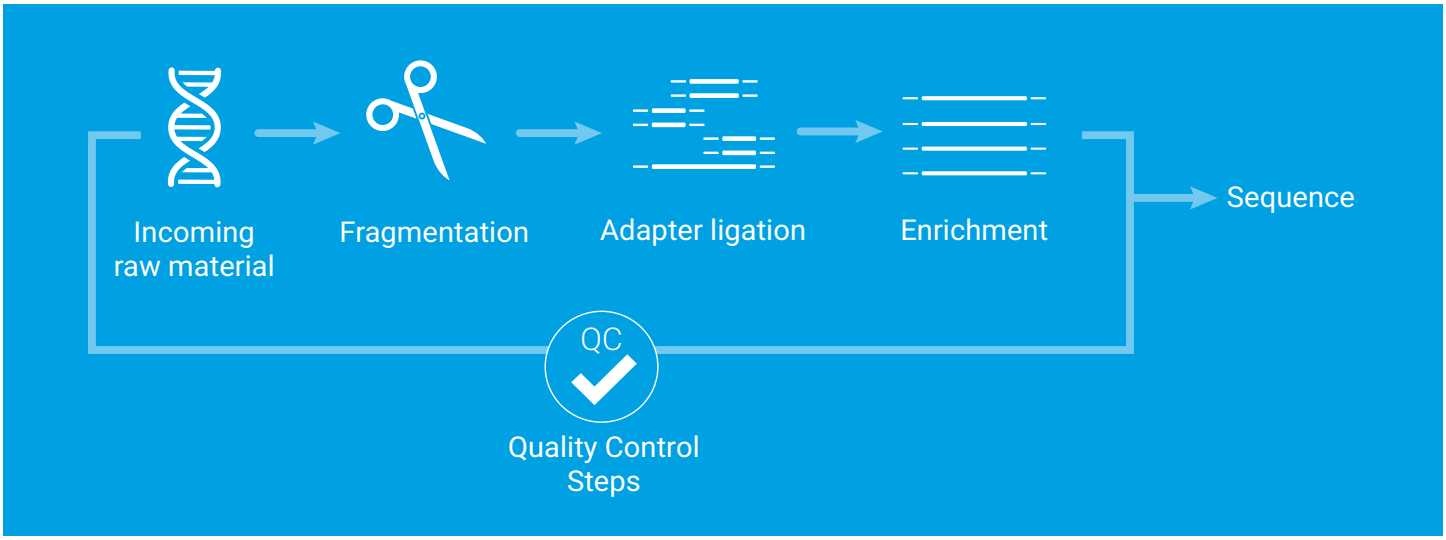


Figure 1. Recommended steps for quality control during NGS library preparation.

Quality Control of Input Material for NGS Library Preparation

One of the most important QC steps in the library preparation process occurs before the NGS workflow begins. This critical QC step is the analysis of the initial material to be used, regardless of whether it is DNA or RNA. Ensuring good quality of the starting material is the first step to preparing a quality library and obtaining successful sequencing results. Most NGS library preparation kits recommend a specific input concentration and size for optimal sequencing results. If the starting material is degraded or of low quality, it may be difficult, if not impossible, to create a library from the sample. Several studies have investigated the correlation between sample quality and downstream analysis.

QC of input DNA

Storage conditions of the starting genomic DNA (gDNA) material can affect its integrity. To understand the storage temperature effect on gDNA, Permenter et al.[1] examined the quality of gDNA samples stored for 1 to 130 days at room temperature compared to 4 °C in different tube types. To analyze degradation, they visualized the sample integrity on an electropherogram from a TapeStation system and calculated degradation as the area under the curve between 150 to 10,000 bp. Intact samples showed a large portion of higher molecular weight material above 10,000 bp indicating intact gDNA, while fragmented or degraded samples displayed a larger amount of sample smeared throughout the assigned degradation area (see publication for example). In this example, the samples with the highest quality were the ones that had been stored in refrigerated EDTA tubes. Downstream analyses, including both DNA fingerprinting assays and SNP analysis, showed that the quality of the gDNA correlated with the quality of the data. This gives valuable insight into proper handling of gDNA and the importance of high-quality starting material to be used for sensitive applications such as NGS.

When developing new methods, it is important to understand how any change, big or small, can affect the sample. For example, in Zhong et al.[2], immunoprecipitated DNA was purified with different methods and reagents, then the quality of the resulting DNA was examined for use in ChIP-seq libraries and sequencing. The size of the different purified DNA samples was analyzed using the Fragment Analyzer system (Figure 2). The results indicated that purification with SPRI bead-based reagents eliminated fragments below 100 bp compared to the column-based extractions and phenol chloroform (PC) method, which recovered fragments less than 35 bp. Since ChIP-seq workflows traditionally fragment and size-select DNA between 100 to 300 bp for sequencing, the column-based and SPRI-bead-based methods of purification did not affect sequencing coverage. However, the authors note that for applications requiring smaller fragments below 100 bp, the sample purification methods and the resulting sample size should be considered when planning experiments.

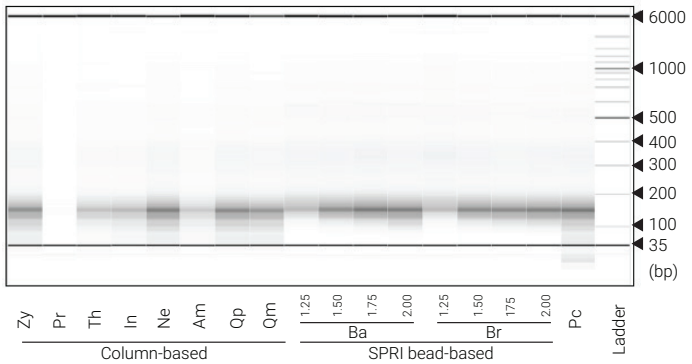


Figure 2. DNA from de-crosslinked chromatin was purified with different reagents and kits, then analyzed using an Agilent Fragment Analyzer system. This figure has been modified and reproduced from Zhong et al.[2]. *Abbreviation: Zy, ChIP DNA Clean & Concentrator (Zymo Research); Pr, Wizard SV Gel and PCR Clean-Up System (Promega); Th, GeneJET PCR Purification Kit (Thermo Fisher Scientific); In, PureLink PCR Purification Kit (Invitrogen); Ne, Monarch PCR & DNA Cleanup Kit (New England Biolabs); Am, Chromatin IP DNA Purification Kit (Active Motif); Qp, QIAquick PCR Purification Kit (Qiagen); Qm, MinElute PCR Purification Kit (Qiagen); Ba, Agencourt AMPure XP kit (Beckman, chromatin to beads ratio from 1:1.25 to 1:2); Br, RNAClean XP kit (Beckman, chromatin to beads ratio from 1:1.25 to 1:2); PC, phenol/ chloroform extraction.*

High-quality nucleic acids are necessary for successful library preparations and sequencing results. To aid in identifying samples of adequate quality, Agilent has developed several quality scores, which are specific to the different automated electrophoresis platforms. These quality metrics provide the user with a reliable assessment of the integrity of a sample, saving time and money by reducing human error and variation between user assessments, while easily identifying unfit samples that would lead to poor results. For example, to evaluate sheared and gDNA, the Fragment Analyzer and Femto Pulse systems utilize the Genomic Quality Number (GQN). For this score, a user-defined threshold specific to the application is applied to a sample. The GQN value is then calculated based on the fraction of the total measured concentration of the sample that lies above the specified size threshold. The GQN scores samples on a scale of 0 to 10, with a score of 0 indicating that none of the sample exceeds the threshold. A score of 10 indicates that 100% of the sample lies above the threshold. Alternately, the DNA integrity number (DIN) is a user-independent objective score of gDNA quality provided by the TapeStation systems. With the DIN, each sample is assigned a score from 1 to 10. A high DIN indicates highly intact gDNA, while a low DIN suggests a strongly degraded gDNA sample.

Formalin-fixed, paraffin-embedded (FFPE) tissues are particularly difficult samples because they have undergone chemical fixation methods to allow for long-term preservation. Nucleic acids can be extracted from these tissues, but the quality is often lower than that of fresh tissue due to the preservation technique, causing cross-linking and degradation. The lower quality can make it difficult to use FFPE samples for sensitive downstream analyses, such as NGS. However, with the ability to determine the quality of the samples and make informed decisions about how to optimize the library preparation protocol, sequencing of FFPE samples can be successful. For example, in Muscarella et al.[3], the TapeStation DIN was applied to FFPE gDNA from different extraction methods to help determine a reliable NGS workflow for oncology studies. The authors state that *"The DIN makes the interpretation of electropherograms easier, facilitates the comparison of samples and provides the repeatability of experiments and quantitation of high-quality genomic DNA for NGS technology."*[3] All samples extracted with manual or automated kits displayed a DIN of >4.4, with no significant difference in quality between the kits. Phenol-chloroform extracted samples were of significantly lower quality, with one sample at a DIN of 1.9 and two samples so degraded that a DIN could not be calculated. Representative examples from each kit and the average DIN for each method are shown in Figure 3. Samples from each extraction method were used for NGS library preparation, and the resulting libraries were also evaluated for quality using the TapeStation system. Each library exhibited similar amplicon patterns, with uniform distribution of concentrations. While all samples gave good sequencing results, the authors showed that the FFPE gDNA extracted with automated methods saved time and expensive reagents, displayed a higher quality over manual preparations, and resulted in higher quality sequencing data.

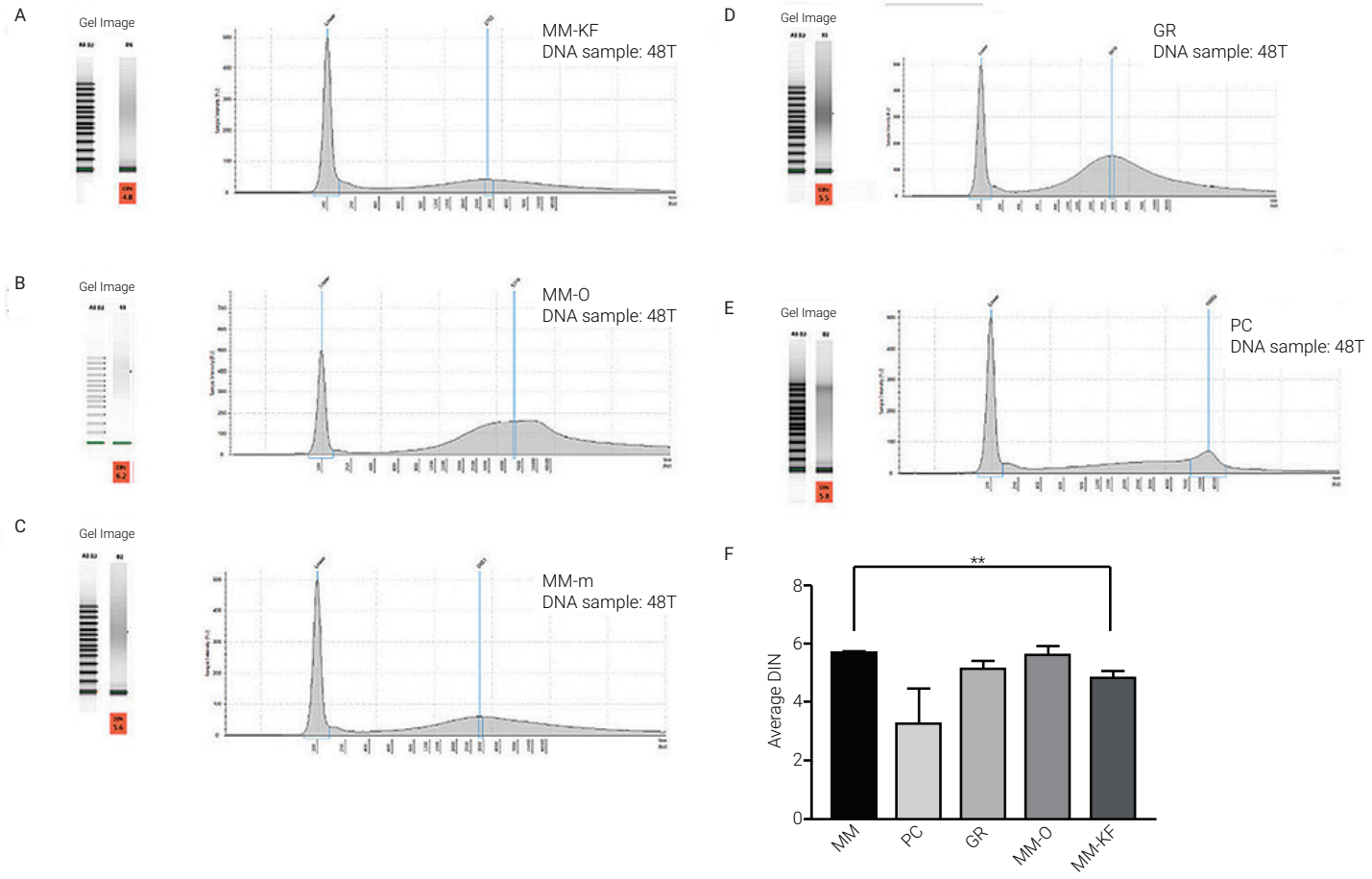


Figure 3. FFPE DNA was extracted from the same sample using five extraction methods and analyzed on an Agilent TapeStation system. (A-E) Representative profiles of the DNA extracted from a single sample, 48T, with: (A) MM-KF, MagMAX FFPE Total Nucleic Acid Isolation Kit (Thermo Fisher), automatic procedure using King Fisher Duo (Thermo Fisher); (B) MM-O, MagMAX FFPE Total Nucleic Acid Isolation Kit (Thermo Fisher), automatic procedure using OMNIA Prima (MASMEC S.p.A); (C) MM-m, MagMAX FFPE Total Nucleic Acid Isolation Kit (Thermo Fisher), manual procedure; (D) GR, GeneRead DNA FFPE Kit (Qiagen); and (E) PC, phenol-chloroform method. (F) The average DIN values for FFPE DNA extracted from six different samples. ** $p < 0.001$, t-test. This figure has been modified and reproduced from Muscarella et al.[3].

In recent years, long-read sequencing technologies have gained prevalence for applications such as de novo genome assembly. Thus, the extraction and treatment of high molecular weight (HMW) gDNA is of utmost importance to avoid unintentional fragmentation or degradation of the sample. The Femto Pulse system is uniquely situated to analyze the quality of HMW DNA, with the ability to separate samples through 165 kb in just a few hours. Klingström et al.[4] discussed proper handling of HMW gDNA and use of the Femto Pulse system to analyze samples of large size to ensure their integrity. In their example, high-quality gDNA is characterized by a single sharp peak at >120 kb, with small smears emanating from either side, indicating portions of the sample that were either slightly degraded (left side) or protected from fragmentation (right side) compared to the majority of the sample (Figure 4).

Depending on the initial sample size, some long read sequencing methods are able to sequence long fragments without subjecting them to fragmentation and amplification steps that may introduce bias. Kingan et al.[5] used the Femto Pulse system to QC their input gDNA material and the resulting NGS library. For the long-read library preparation protocol, it was important that their starting material was greater than 20 kb (generally around 40 kb), and not contain fragments below 5 kb, to avoid shearing and size-selection steps. Because of the extended sizing range up to 165 kb on the Femto Pulse system (Figure 5), they were able to identify samples that meet these requirements, eliminate these steps, and proceed with library preparation with a low input concentration.

QC of input RNA

RNA sequencing requires high-quality RNA as input material. However, RNA is a fragile biomolecule that can be easily degraded. Storage temperature, freeze/thaw events, extraction methods, and presence of ubiquitous RNases can all impact the integrity of RNA, making QC of input RNA a crucial step before use in sensitive downstream applications. RNA integrity can be assessed using the RNA integrity number (RIN) from the Bioanalyzer system, and the equivalent quality metrics developed for the other automated electrophoresis instruments: the RNA integrity number equivalent (RIN^e) from the TapeStation systems[6], or the RNA quality number (RQN) from the Fragment Analyzer[7] and Femto Pulse systems. Each of the RNA quality metrics considers the entire electrophoretic separation of the RNA sample, including the ratio of the ribosomal fragments and the presence or absence of degradation products.

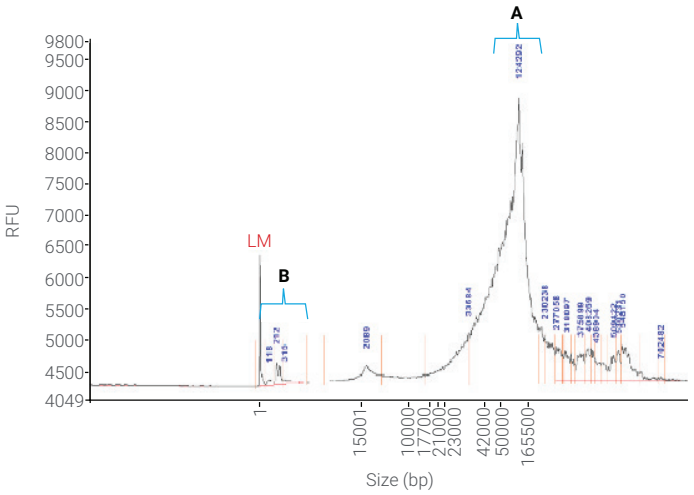


Figure 4. High molecular weight DNA was analyzed on the Agilent Femto Pulse system to ensure the quality of the sample was sufficient for long read sequencing. (A) The main fragment displayed a size of >120 kb, indicating high quality DNA. (B) Small fragments to the left of the main smear indicate only slight degradation of the sample. This figure has been modified and reproduced from Klingström et al.[4].

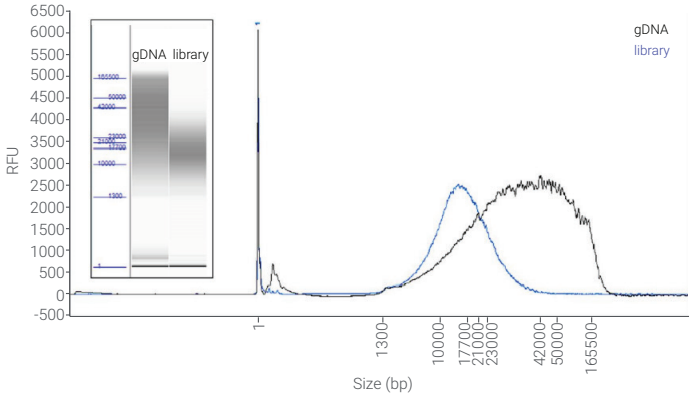


Figure 5. gDNA was extracted from a single *Anopheles coluzzii* mosquito and used for PacBio Single Molecule, Real-Time (SMRT) Sequencing, with a modified SMRTbell library preparation protocol. The size and concentration of both (A) the gDNA and (B) the final library were assessed on the Agilent Femto Pulse system before sequencing. This figure has been modified and reproduced from Kingan et al.[5].

Many sequencing laboratories consider good quality RNA to have a RIN score above or equal to 7.0, or perform studies of their own to establish the sample quality criteria and standards that will be employed in their workflows. For example, Haller et al.[8] evaluated several RNA samples using a variety of methods. The samples were classified as *"pass"* or *"fail"* to indicate whether the sample qualified for use in downstream microarray and qRT-PCR experiments. Among the analysis methods, the RIN score was used as *"a tool for standardization of RNA quality control, in a user-independent, automated and reliable procedure. The RIN takes into account decreases of signal intensities for the two ribosomal bands and increases in the presence of shorter fragments between the two peaks and below the 18S band."*[8] Each of the samples with a *"passing"* grade displayed a RIN of 6.6 or higher, with an average of 7.5 (see publication for examples). As NGS technologies have evolved, the high passing RIN score has also been used as a determining factor for achieving successful library preparation and sequencing results.

While the RIN is the most established standard for RNA QC, several studies have been performed demonstrating the equivalences of the RIN to the RINe [6], and the RIN to the RQN[7]. By providing information about the integrity of a sample, the RNA quality metrics allow users to make informed decisions about how to proceed with samples that are not ideal. These decisions can include how to optimize protocols for slightly degraded samples. Alpern et al.[9] aimed to improve upon an existing protocol for a unique RNA-Seq approach: bulk RNA barcoding and sequencing (BRB-seq). The authors used a Fragment Analyzer system to assess both the initial RNA integrity and the resulting NGS libraries after changes to the method were introduced. RNA samples were fragmented to different levels and evaluated using BRB-seq to estimate differential gene expression. As shown in the overlay (Figure 6), the RQN decreased as the RNA was increasingly fragmented. Another protocol change that was aided by quality assessment involved the authors comparing commercially available reagents to in-house prepared reagents. In one example, they observed an increase in library yield on the Fragment Analyzer system with their in-house tagmentation enzymes (Tn5-A/B, Tn5-B/B), allowing them to reduce the costs of their library preparations (Figure 7). These QC checkpoints gave them the knowledge they needed to optimize their BRB-seq protocol for degraded samples and allowed them to successfully generate an NGS library that was able to detect greater than 75% of the differentially expressed genes seen in the intact sample. Overall, the authors were able to optimize their NGS library preparation workflow based on quality assessment so that good sequencing results could be achieved even from samples with low RQN values.

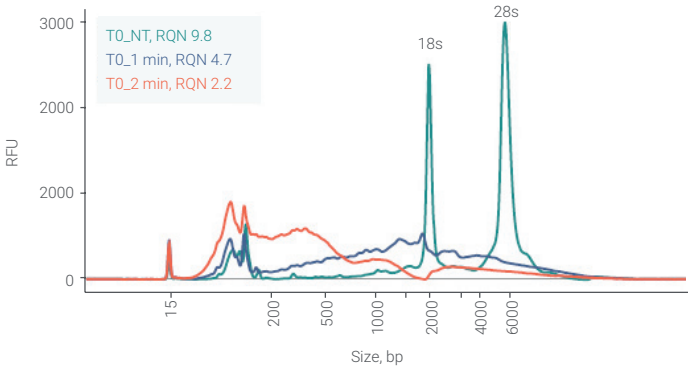


Figure 6. Samples were intentionally degraded for 1 or 2 minutes in order to evaluate the performance of BRB-seq with fragmented RNA. The RNA was evaluated using an Agilent Fragment Analyzer system, and the RQN was compared between the intact (NT: green trace) and degraded (1 min: blue trace; 2 min: red trace) samples. T0: pre-adipocytes. This figure has been modified and reproduced from Alpern et al.[9].

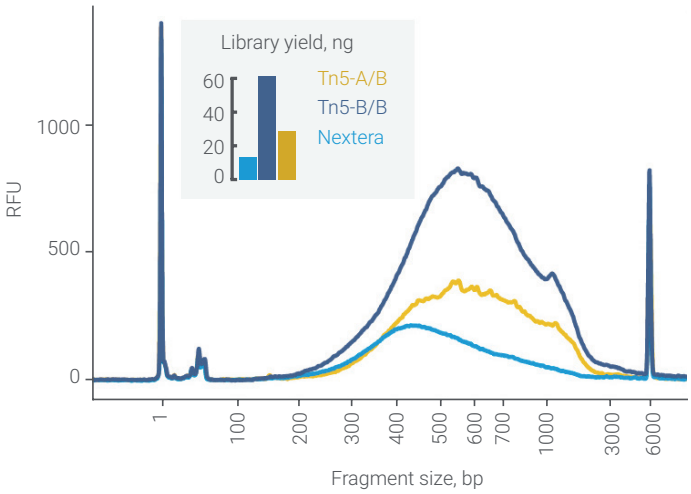
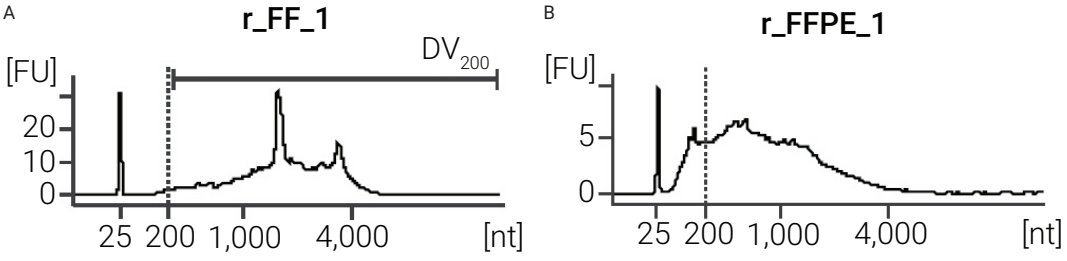


Figure 7. NGS libraries were prepared with different Tn5 tagmentation enzymes (Tn5-A/B and Tn5-B/B made in-house, and Illumina Nextera), and the profile and yield of the libraries were analyzed with an Agilent Fragment Analyzer system. This figure has been modified and reproduced from Alpern et al.[9].

QC of FFPE RNA

Recently, more labs have been sequencing with samples that are of lower quality or slightly degraded due to extraction methods or storage conditions. FFPE is a common way to preserve samples for later use, due to its ability to protect cellular structure and tissue morphology. However, it can also introduce complications, such as chemical modification of nucleic acids, rendering the sample challenging for NGS analysis. The automated electrophoresis instruments enable quality assessment of FFPE RNA through the DV_{200} metric, used specifically to look at the percentage of RNA larger than 200 nt in the sample as an indication of the level of RNA degradation. With the DV_{200} score, researchers can make decisions about how to modify their NGS workflows if necessary. Common adjustments include increasing or decreasing the shearing time to obtain the desired size or excluding a sample that may not be of good enough quality to use. For example, Wimmer et al.[10] used the Bioanalyzer system and the DV_{200} to evaluate the quality of several fresh-frozen (FF) and FFPE RNA samples from rat (Figure 8), human, and mice, and compared the samples using microarray and RNA-Seq. As shown in the electropherogram, the rat FF samples showed clear 18S and 28S peaks, with an average RIN of 6.1. In contrast, the rRNA peaks from FFPE RNA are not distinguishable, as the sample appears as a large plateau with a RIN of 2.3, indicative of degraded RNA.

The long plateau ranges from 100 to 4,000 bp, indicating the presence of some long RNA fragments. The DV_{200} score was used as an alternative approach to measure sample quality. Although the average DV_{200} score of the FFPE samples was lower at 76% compared to the 97% from the FF samples, the FFPE score was still indicative of acceptable material for Illumina sequencing (DV_{200} > 70%). The authors state, *"RNA integrity analysis using RIN (or equivalents), as routinely done for fresh and FF samples, cannot be reliably applied to FFPE material since all RNA species are innately fragmented due to formalin fixation. Instead, the DV_{200}, the percentage of RNA fragments with a length of more than 200 nucleotides, might be a viable alternative. Reportedly, a DV_{200} greater than 70% indicates high-quality RNA, while RNA samples with values between 50 and 70% are of medium quality and demand the use of higher input volumes for transcriptome analysis. RNA samples with DV_{200} values below 30% have been suggested to be too degraded for further experiments. In our study, we were able to achieve DV_{200} values clearly above the proposed threshold for high-quality RNA despite very low tissue input."*[10] The DV_{200} score provided by the Bioanalyzer system allowed the authors to identify samples for sequencing that would have otherwise been discarded.



C

	RIN	DV_{200}
r_FF_1	6.1	97
r_FF_2	6.7	97
r_FF_3	5.4	96
Mean ± SD	6.1 ± 0.7	97 ± 1
r_FFPE_1	2.2	81
r_FFPE_2	2.4	72
r_FFPE_3	2.2	76
Mean ± SD	2.3 ± 0.1	76 ± 5

Figure 8. (A) Fresh frozen (FF) and (B) FFPE RNA prepared from rat (r) microdissected samples were analyzed on an Agilent Bioanalyzer system. (G) The RIN and DV_{200} from multiple samples were reported to assess sample quality. This figure has been modified and reproduced from Wimmer et al.[10].

QC of Cell-free DNA

Analysis of cell-free DNA (cfDNA) has recently gained more prevalence as an input material for NGS, as in clinical research it enables the collection of nucleic acid materials through minimally invasive methods. Each of the automated electrophoresis instruments can be used to examine the cfDNA profile, including sizing and quantification. Electrophoretic separation of cfDNA typically displays a profile with three peaks, indicative of nucleosomal fragments. Mendioroz et al.[11] investigated potential epigenetic biomarkers for neurodegeneration by extracting cfDNA from amyotrophic lateral sclerosis (ALS) patients. The concentration, quality, and size distribution of the cfDNA was assessed with a Fragment Analyzer system (Figure 9) prior to sequencing. All samples showed a typical pattern of fragmented cfDNA and were successfully used for downstream pyrosequencing and bisulfite cloning sequencing to identify a differentially methylated region of a gene in ALS patients compared to controls.

While each of the automated electrophoresis instruments can be used for analysis of cfDNA, the TapeStation system is the only one with a dedicated cfDNA assay and validated quality metric. The %cfDNA score can be applied to cfDNA samples analyzed with the cfDNA ScreenTape assay in order to qualify the amount of cfDNA present in a sample relative to any gDNA contamination, which could hinder downstream sequencing analyses.

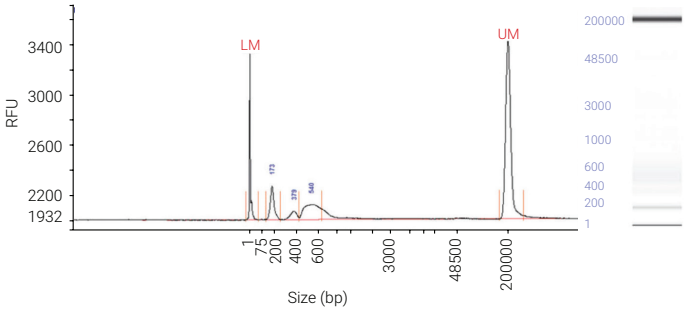


Figure 9. cfDNA analyzed on an Agilent Fragment Analyzer system displays three distinct peaks at approximately 165, 350, and 565 bp, indicative of nucleosomal fragmentation. This figure has been modified and reproduced from Mendioroz et al.[11].

Quality Control Throughout NGS Library Preparation

Once appropriate samples for NGS library preparation have been identified, several steps are taken to turn them into a suitable library for sequencing. To achieve good, reliable sequencing data, QC checkpoints at various points through the workflow are highly recommended to ensure the creation of a high-quality library. Some of these checkpoints include QC of the sample following fragmentation and adapter ligation.

Fragmentation quality assessment

Fragmentation plays a key role in library preparation by shearing large samples down to smaller sized fragments. These fragments are then converted into a library that can be read by the specific sequencing technology. The size distribution required for library preparation differs for each library preparation kit and sequencing type. Thus, QC analysis is required to ensure that the sample is of the correct length before proceeding with library preparation. Any of the automated electrophoresis instruments are well suited for this purpose. For example, with ChIP-seq libraries, the initial samples are first fragmented to a size of 200 to 500 bp. However, fragmentation methods are not always optimized, leading to the presence of smaller and larger fragments remaining in the sample. Blecher-Gonen et al.[12], discussed how these larger fragments interfere with sequencing and suggest performing a size selection step after library pooling to eliminate them. Analysis of the samples before and after size selection using the TapeStation systems confirms exclusion of the larger fragments after size selection[12]. Sequencing results demonstrated that size selection improved cluster generation and led to an increase in the number of peaks obtained from sequencing compared to the non-size selected samples.

Fragmentation of samples can be performed through various methods, including shearing and enzymatic procedures. Shearing through sonication involves an acoustic probe that focuses high-frequency, short-wavelength bursts of energy on the sample to introduce random cuts into the DNA. However, this process of sonication can be variable and requires careful calibration of multiple settings, including the intensity and frequency of the acoustic waves, to produce a consistent size distribution. In Tramontano et al.[13], the authors sheared their samples with a variety of sonication settings to optimize their shearing protocol. Each fragmented sample was analyzed on the Fragment Analyzer for comparison, with the goal to improve upon their overall NGS library preparation workflow for amplicon sequencing. *"The flexibility and throughput of amplicon sequencing is increased through fragmentation of PCR products. More bases can be interrogated from a PCR product, and fragmentation into small pieces allows use of a range of different short-read sequencing platforms."*[13] Thus, they tested a variety of different sonication parameters on PCR amplicons to identify the method that gave the best fragmentation distribution (Figure 10). The refined parameters were then applied to pooled amplicons to be used for sequencing (Figure 11). A comparison of sequencing data from fragmented and nonfragmented PCR amplicons showed an increase in coverage and more consistent data.

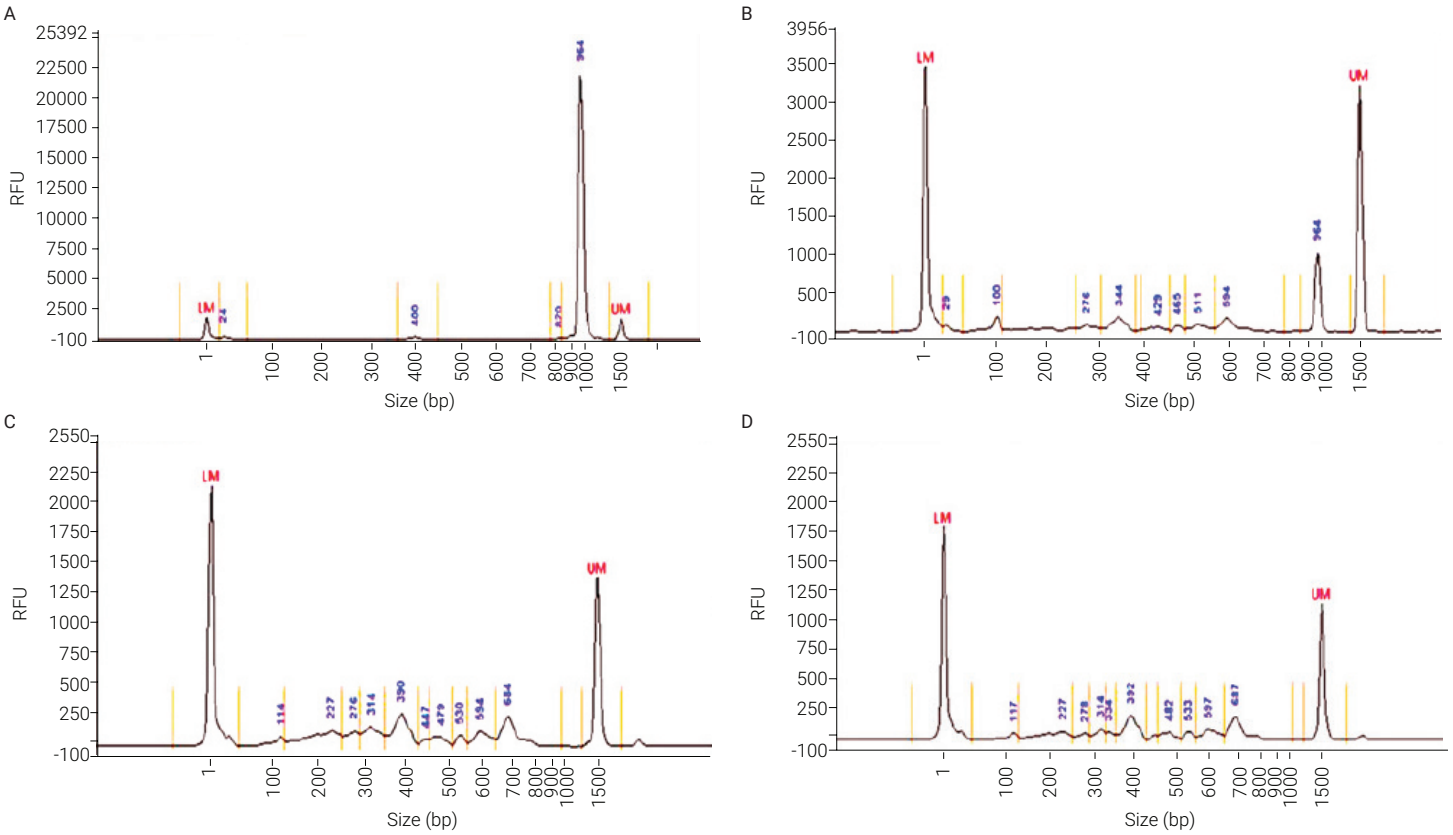


Figure 10. PCR amplicons were fragmented to a size of 350 to 500 bp with a variety of sonication parameters to obtain the appropriate samples for sequencing. The extent of fragmentation was evaluated using an Agilent Fragment Analyzer system. (A) The non-fragmented PCR amplicon at 964 bp. (B) The fragmentation pattern obtained from sonication test one. The large peak leftover at 964 bp indicates incomplete fragmentation. (C and D) Technical replicates of sonication test 13, which displayed a broad fragmentation pattern with a peak at approximately 390 bp, ideal for downstream sequencing. This figure has been modified and reproduced from Tramontano et al.[13].

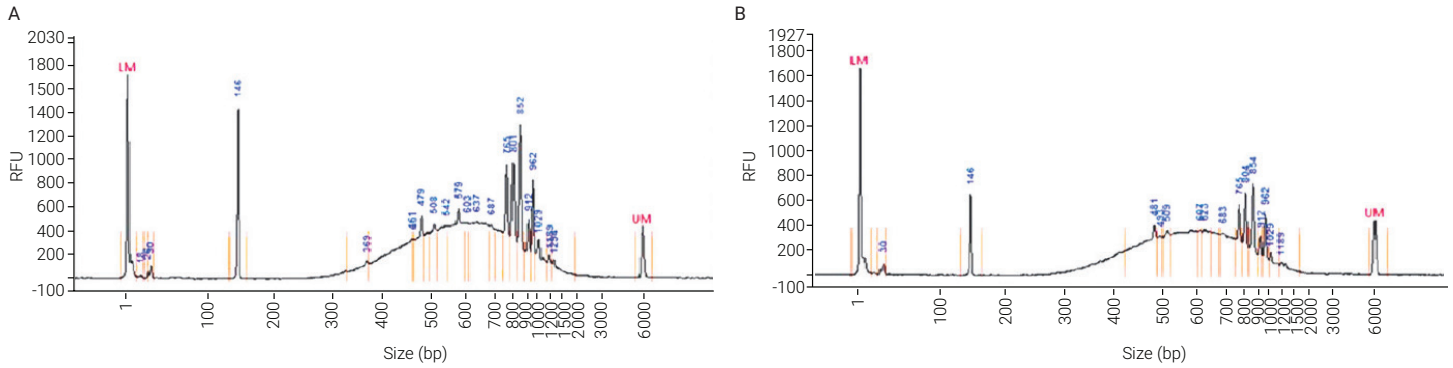


Figure 11. Representative examples of two NGS libraries (A,B) prepared from pooled PCR amplicons were analyzed on an Agilent Fragment Analyzer system. This figure has been modified and reproduced from Tramontano et al.[13].

Alternately, enzymatic fragmentation involves the use of enzymes to create double stranded breaks in the DNA. While this method can be easily scaled to high-throughput experiments and is relatively easy to use, it comes with concerns regarding sequence bias due to enzyme recognition sites on the DNA. Enzymatic protocols also have to be optimized for sample type, buffer, and concentration. For instance, Aguirre et al.[14] applied ddRADseq (double digest restriction site-associated DNA sequencing) to a non-model organism for genotyping. Because standard protocols are written for model organisms, optimization of the protocol at several steps was required to obtain the most reliable data. Part of their protocol optimization included selecting a pair of restriction enzymes that would give them an optimum size range for their sequencing platform. The Fragment Analyzer systems were utilized for QC of the samples before and after enzymatic fragmentation, as well as for QC of the final library before sequencing. Figure 12 shows the large variation in size distribution displayed after fragmentation with two different enzyme sets. The homogenous pattern displayed by the Fragment Analyzer and analysis of the smear range from 350 to 600 bp aided the authors in determining which enzyme set to use for their protocol. In this case, the SphI-MboI pair (Figure 12A) produced the largest percentage of sample within the necessary size needed for sequencing compared to the PstI-MspI pair (Figure 12B).

With the advent of specialized sequencing types, a common question is which method of fragmentation to use. Lan et al.[15] used NGS for high-throughput HLA typing. To optimize their protocol, they compared traditional TruSeq Nano (which uses Covaris sonication for shearing) to transposase-based (a step that combines enzymatic fragmentation and adapter ligation into a single reaction) Nextera library preparation methods and examined the sequencing coverage bias and accuracy of genotyping calls. Coverage, or read depth, is an incredibly important metric in genotyping, as it refers to the number of times a nucleotide is represented in the aligned reads. Coverage is therefore directly related to the specificity of variant detection, thereby impacting genotyping calls. Libraries prepared with both technologies were validated and quantified with a TapeStation system[15].

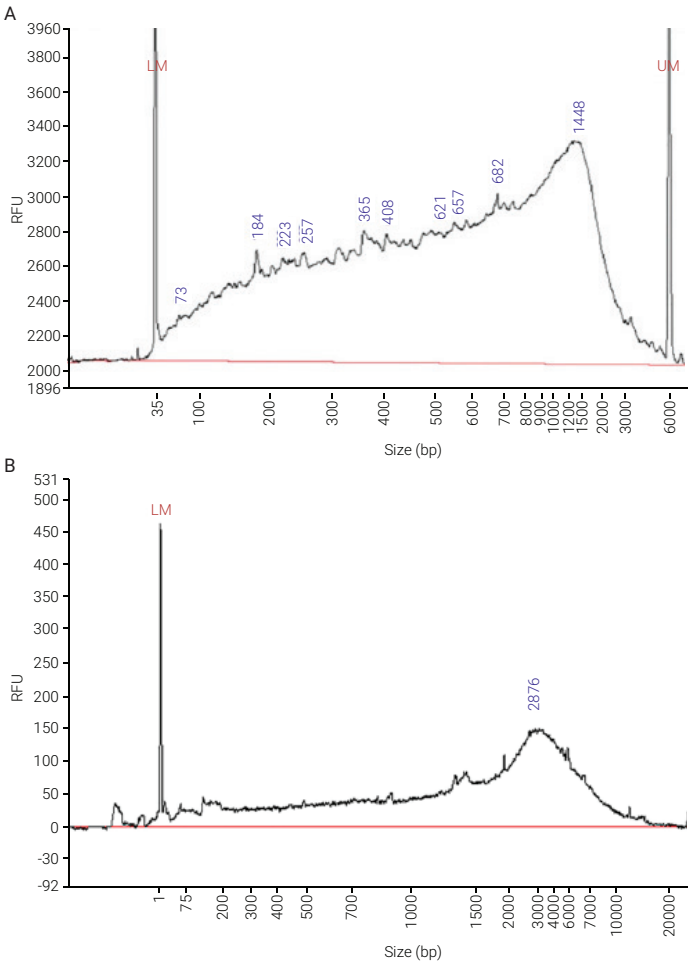


Figure 12. An Agilent Fragment Analyzer system was used to analyze *in vitro* digestions of *E. dunnii* gDNA with different enzyme pairs: (A) SphI-MboI and (B) PstI-MspI. The SphI-Mbo digestion resulted in a larger number of small-sized fragments in the size range necessary for size selection and sequencing. This figure has been modified and reproduced from Aguirre et al.[14].

The distribution of the Nextera libraries was more dispersed than the TruSeq libraries, resulting in slightly larger sizes on the electropherograms. However, after sequencing and trimming, the median insert size of the aligned reads was similar for all library preparations. Overall, the transposase methods saw more coverage bias, and examination of the coverage ratio for different alleles showed a large variation between all library preparation protocols tested, indicating some coverage bias with all methods. This paper demonstrates why different fragmentation methods are used for different sequencing types and highlights why method optimization is crucial for sensitive NGS applications.

cDNA preparation for RNA-Seq libraries

An essential step in the preparation of RNA-Seq libraries is to separate the mRNA from the total RNA and convert it into cDNA (complementary DNA) via reverse transcription for subsequent library preparation steps and sequencing. Since the cDNA is transcribed from mRNA, the library will contain only the coding regions for the expressed genes of the organism or tissue the RNA was collected from. Thus, common QC steps in the process are to examine the quality of both the total RNA prior to reverse transcription and the resulting cDNA before proceeding with the rest of library preparation. Misra et al.[16] detail a protocol for single cell RNA-Seq (scRNA-seq) using FACS (fluorescent activated cell sorting) to isolate Arabidopsis sperm cells. In short, RNA from single-sorted sperm cells undergoes reverse transcription, pre-amplification, and cDNA purification. Among other tests to confirm sample quality, size distribution of the cDNA was confirmed on a Fragment Analyzer system with the High Sensitivity NGS kit before proceeding with subsequent library preparation steps and sequencing. The electropherogram of the cDNA shows two peaks with a broad base, and an average size of 1,500 to 2,000 bp (Figure 13A,B). The same kit was used on the Fragment Analyzer system to qualify the final single-cell library before sequencing (Figure 13D). Since obtaining single cells from plants and isolating cDNA with very little cellular content can be quite difficult, these QC steps allowed them to choose only the best cDNA samples and libraries for sequencing.

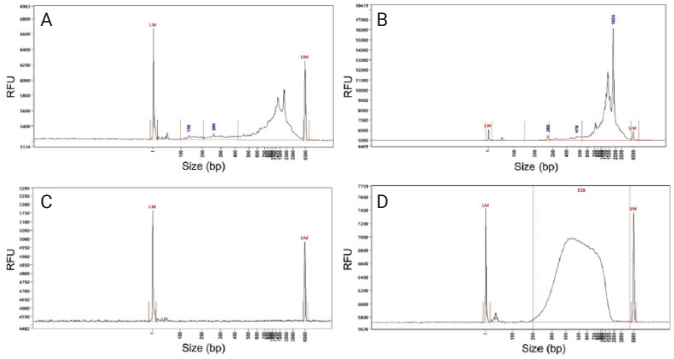


Figure 13. Various single-cell RNA-seq library preparation steps undergo QC with an Agilent Fragment Analyzer system. Shown are electropherograms from representative steps: (A) cDNA amplification from a single cell. (B) cDNA amplification from a bulk sample as a positive control. (C) A negative control, with no cells. (D) The final NGS library. This figure has been modified and reproduced from Misra et al.[16].

Adapter ligation

Following sample preparation, the next steps of NGS library preparation generally include end-repair, adenylation, and adapter ligation. Adapter are platform-specific sequences that enable recognition of individual fragments during sequencing. QC after adapter ligation ensures successful ligation of the adapters to the sample. When analyzed with an automated electrophoresis instrument, successful ligation is indicated by a slight shift in distribution compared to the fragmented sample, with the adapter ligated library showing a larger average size. Poor ligation efficiency is known to result in sample loss, which can be detrimental for sensitive sequencing experiments where the input material is limited, such as cfDNA. As such, there have been many attempts to optimize the adapter ligation step in the NGS library preparation process. In a study by Nix et al.[17], the ligation efficiency of cfDNA with single-stranded and duplex adapters was examined to optimize the library preparation protocol and reduce sample loss. Figure 14 demonstrates how the TapeStation system was used to QC all the checkpoints in the workflow: the adapters (B), input material (C), post ligation (D), and post-PCR samples (E). As shown in the electropherogram image, the sensitivity of the TapeStation allowed for visualization of multiple peaks within the post ligation sample, identified as free adapter, unligated product, single-end ligated product, and dual-end ligated product. The ligation efficiency post ligation was determined by quantifying the molarity within the regions of each peak. The ligation efficiency was then calculated as the percentage of dual-end ligated product within all ligation products. Comparing the post-PCR sample to the input DNA shows a slight size shift, with the post-PCR peak being slightly larger as a result of successful adapter ligation. Overall, the authors utilized the TapeStation to determine that for their protocol, the ligation efficiency of the duplex adapter was higher than the single adapter.

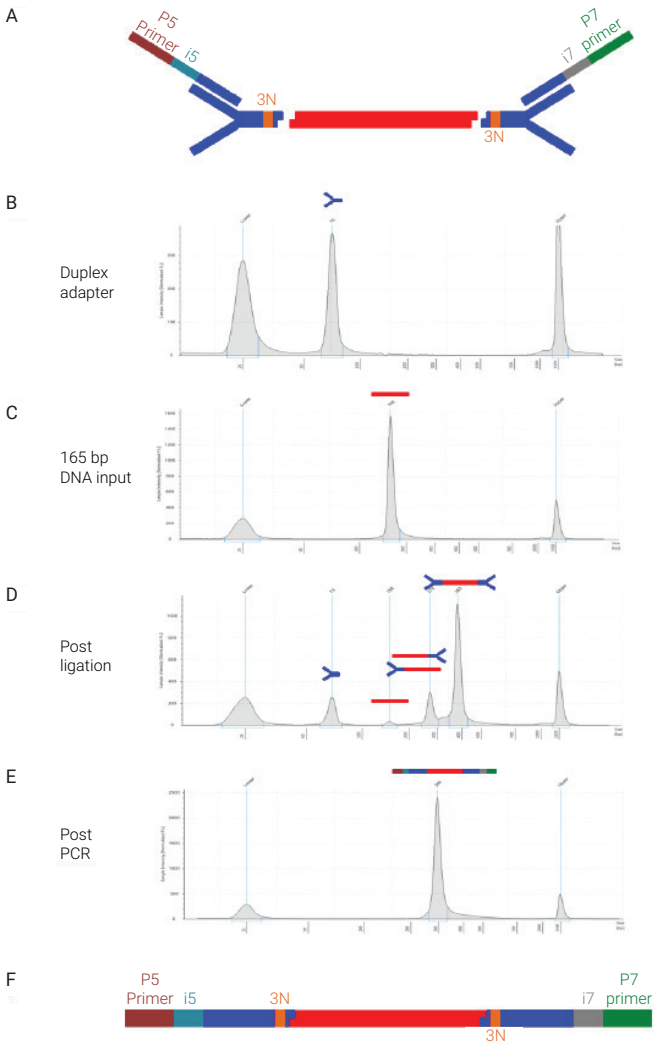


Figure 14. QC of adapter ligation using an Agilent TapeStation system. (A) Schematic of the insert (red) and duplex adapter (blue). (B) An electropherogram of the duplex adapter, composed of single-stranded DNA, displays a peak at 75 bp. (C) The input DNA at 165 bp. (D) Using B and C as references, the four peaks displayed in the post-ligation QC can be identified as unligated adapter, unligated insert, single-end ligation products, and dual-end ligation products. (E) The sample and (F) schematic of the 165 bp insert with dual-end adapters post-PCR. This figure has been modified and reproduced from Nix et al.[17].

Quality Control of Final NGS Libraries

The final step of library preparation is PCR amplification, which allows for enrichment of adapter-ligated fragments and generates enough copies of the library for sequencing. This step can also be used to add an index to multiplex multiple samples together for a single sequencing run. While the size and shape of the final library is dependent upon the library preparation kit used, visualization and analysis of all types of libraries is possible with any of the automated electrophoresis platforms. Any additional peaks other than the library smear indicate the presence of excess primer or adapter dimers. Analysis at this stage can also help determine if a library has been over- or underamplified and aids in pooling of libraries for sequencing. QC of the final library ensures a high-quality preparation and allows for determination of the size, concentration, and molarity of the library.

Library amplification

In the final steps of NGS library preparation, the adapter-ligated fragments prepared in the first steps are amplified through PCR. During targeted amplicon sequencing, two PCR reactions are performed. The first reaction is to amplify only the targets of interest, while the second enrichment is to add indices to the amplicons. Only those fragments containing the indices will anneal to the flow cell during sequencing. The overall success of amplicon sequencing thus depends on how well the PCR reactions work.

Each individual amplicon must be sufficiently amplified with a separate index ligated to each amplicon to differentiate them during the bioinformatics process. When handling challenging samples, such as those with low concentrations or contaminants, the PCR process must be optimized to generate the best reactions and obtain good sequencing results. Sidstedt et al.[18] investigated the impact that two known PCR inhibitors have on forensic STR (short tandem repeat) analysis using targeted MPS (massively parallel sequencing) techniques. Forensic samples may contain molecules, such as these inhibitors, that hinder DNA polymerization, leading to low template availability for downstream analysis. Thus, having an optimized PCR protocol is of utmost necessity for successful sequencing of these samples. By analyzing the size and concentration of the amplicons and the resulting NGS libraries with the Fragment Analyzer system, the authors showed that the inhibitors lowered the efficiency of the initial PCR. This resulted in a lower read count during sequencing. However, they were able to counteract the decrease in efficiency through the addition of BSA to improve the PCR chemistry. Analysis of challenging samples with the automated electrophoresis instruments can help improve NGS library preparation workflows, helping researchers save time, costs, and precious samples.

Dimers

Dimers are artifacts of a PCR reaction, in which the primers or adapters anneal to copies of themselves instead of to the template. This can occur for many reasons, for example low template concentration, excess primer, suboptimal reaction temperatures, or overamplification. When samples containing dimers are analyzed with the automated electrophoresis instruments, the resulting electropherogram pattern will display a peak to the left of the main library smear. The size of this peak will depend upon the length of the primer but will generally be approximately 100 bp or less. The high sensitivity provided by the instruments allows even small amounts of dimer to be seen with ease, and the percent of dimer present in a sample can be easily calculated. Dimers are a common issue, and if seen on an electropherogram, indicate that an additional clean-up step should be performed to eliminate the artifacts before sequencing. Recommendations from core facilities state that libraries will not be accepted for sequencing unless they are made up of less than 0.5% adapter dimer. This is because shorter fragments such as primer dimers bind more efficiently to the flow cell than the larger fragments of the library. If even 5% of a library is composed of dimers, they can consume more than 50 to 60% of the sequencing reads, resulting in a failed sequencing run[19,20].

Marchal et al.[21] present a protocol for sequencing newly synthesized DNA, in which they emphasize the effects of PCR artifacts on the sequencing results, stating that *"overamplification with primer depletion causes PCR artifacts and should be prevented."* The authors suggest that re-amplification is necessary if the concentration is too low, but should be avoided, as it can increase the risk of PCR bias. Several examples of *"good and bad quality libraries"* were analyzed on a Bioanalyzer system. Generally, a library is considered *"good"* when it displays a bell-shaped smear with a fairly tight distribution within the required size range (Figure 15A). A *"bad"* library displays additional peaks outside of the library smear (Figure 15B), indicative of an adapter dimer. If overamplification has occurred, the peak height of the main library is doubled, and an extraneous smear appears after the library[22] (Figure 15C). Since dimers can use up a large portion of sequencing reads, it is incredibly important to remove as much as possible from the library. This can be done through purification steps, and refining protocols to either decrease the concentration of adapters and primers used or optimizing the number of amplification cycles used for future library preparations.

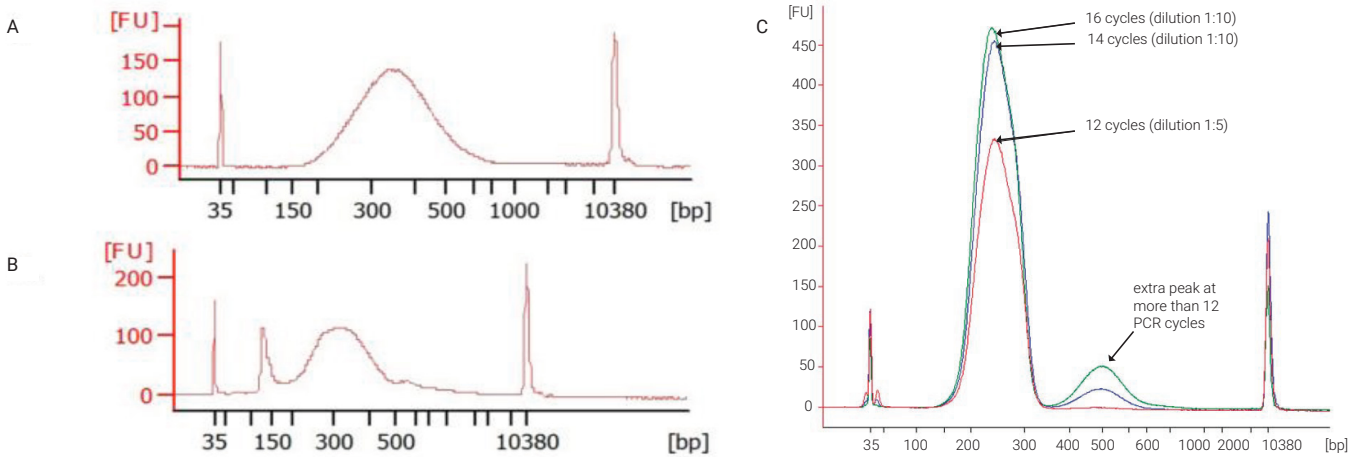


Figure 15. NGS library QC using the Agilent Bioanalyzer system. (A) An example of a good library. (B) An example of a library with remaining adapter dimers around 150 bp. (C) NGS libraries were intentionally over-amplified, resulting in a small extraneous peak at twice the size of the main library peak. This figure has been modified and reproduced from Marchal et al.[21].

Sizing, quantification and molarity

Determining molarity is an important part of the final library QC, as the sample will need to be diluted to the appropriate molarity and possibly pooled with other libraries before loading onto the sequencer. Both size and concentration are used to calculate molarity, so ensuring accurate quantification and sizing, as well as performing a quality check of the library, is key to successful sequencing. Sample size can be easily and accurately determined with the automated electrophoresis instruments, and the distribution of the sample is indicative of the library quality. The concentration of the sample is also calculated with the analysis software corresponding to each instrument. In a study aiming to make improvements to their library preparation protocol, the authors (Quail et al.[23]) showed that they got a better cluster density when they quantify their libraries with the Bioanalyzer than with traditional spectrophotometry methods. Accurate flow cell loading is of utmost importance to sequencing results as stated by the authors: *"the concentration of DNA going into the flow cell governs the number of clusters produced…there is an optimal concentration range of DNA that will yield clusters in the optimal density range, enabling the maximum amount of data to be obtained."*[23] Overestimating the concentration of a library can lead to too few clusters, making it uneconomical to sequence. Underestimation of the concentration generates too high cluster density, resulting in overlapping clusters, and thereby reducing the amount of data obtained from sequencing. For example, Quail et al.[23] showed that using spectrophotometry for quantification led to inconsistent cluster density, likely due to the presence of adapter dimers and unextended dimers that cannot be differentiated from the library. Because the automated electrophoresis instruments separate a sample by size, analysis with the Bioanalyzer DNA 1000 kit allowed for distinction between the intended amplicon and the dimers, permitting a more accurate quantification of the library. With the additional benefit of sizing analysis, they were able to calculate the molarity of the library in a single QC checkpoint and generate a more consistent cluster density upon flow cell loading. Each of the automated electrophoresis instruments and their corresponding analysis kits provides accurate and reliable sizing, quantification, and molarity, making the portfolio ideal for QC analysis of NGS libraries.

Automated Electrophoresis Throughout NGS Library Preparation

The automated electrophoresis instruments from Agilent, including the Bioanalyzer, Fragment Analyzer, TapeStation, and Femto Pulse systems, are ideal for quality control throughout the entire NGS library preparation workflow. They help ensure good sample quality for successful sequencing by analyzing initial samples, at many checkpoints through the process, and the final library.

The use of the automated electrophoresis instruments through the NGS workflow has been well documented in literature. For example, Marosy et al.[24] utilized the Bioanalyzer system at three checkpoints throughout their protocol for generating exome-enriched libraries from FFPE samples. First, following fragmentation to ensure the size distribution and sample concentration before proceeding with the subsequent library steps. Second, the precapture amplification reaction to confirm the concentration and size shift due to the ligated adapters. Finally, they use the Bioanalyzer to determine the molarity of the final library following the postcapture amplification step, in order to properly dilute the sample for flow-cell loading. QC at these crucial steps allows them to make decisions about how to proceed with their library preparation workflow and avoid losing precious samples, reagents, and time.

Each of the automated electrophoresis instruments offers different reagents kits for a variety of sample types, provides accurate and reliable sizing, quantification, and molarity, and utilizes quality metrics for an unbiased assessment of sample integrity. A benefit of the instruments is the simple switch between applications with the different reagent kits based on the type and size of the sample. For example, as referenced previously, Muscarella et al.[3] uses the TapeStation with the Genomic DNA ScreenTape to determine the DIN of their samples before proceeding with library preparation. They then use the High Sensitivity D5000 ScreenTape assay to do QC of the final library before sequencing. By performing quality checks at different steps, they ensure that only the best samples and libraries are sequenced.

The automated electrophoresis instruments can also be used as a tool to help researchers optimize their workflows or make custom protocols for samples that cannot be processed with universal kits. Aguirre et al.[14] combined many aspects of several protocols and optimized the protocol to successfully sequence a unique species. To aid in the optimization of the protocol, they utilized the Fragment Analyzer systems to QC the initial samples and throughout the entire library preparation workflow for quality, size, and concentration (Figure 12).

Summary

Quality control steps throughout the entire library preparation workflow are essential to NGS library preparation and enable successful sequencing results. The automated electrophoresis instruments from Agilent offer a portfolio of reagents kits that are compatible with a variety of sample types and sizes, providing accurate and reliable sizing, quantification, and molarity. Additionally, each instrument utilizes specific quality metrics optimized for each platform, for an unbiased assessment of sample integrity. Although the quality metric calculations are specific for each instrument type, the resulting quality scores are highly comparable across platforms. As evidenced throughout many publications, the automated electrophoresis instruments are ideal for performing these important QC steps, not only for traditional NGS library preparation, but also for troubleshooting and optimizing protocols, and developing custom workflows for challenging or unique samples.

References

1. Permenter, J.; Ishwar, A.; Rounsavall, A.; Smith, M.; Faske, J.; Sailey, C. J.; and Alfaro, M. P. Quantitative Analysis of Genomic DNA Degradation in Whole Blood Under Various Storage Conditions for Molecular Diagnostic Testing. *Mol. Cell Probes* **2015**, 29, 449–453.
2. Zhong, J.; Ye, Z.; Lenz, S. W.; Clark, C. R.; Bharucha, A.; Farrugia, G.; Robertson, K. D.; Zhang, Z.; Ordog, T.; and Lee, J. H. Purification of Nanogram-Range Immunoprecipitated DNA in ChIP-seq Application. *BMC Genomics* **2017**, 18, 985. https://creativecommons.org/licenses/by/4.0/
3. Muscarella, L. A.; Fabrizio, F. P.; De Bonis, M.; Mancini, M. T.; Balsamo, T.; Graziano, P.; Centra, F.; Sparaneo, A.; Trombetta, D.; Bonfitto, A.; Scagliusi, V.; Larizza, P.; Capoluongo, E. D.; and Fazio, V. M. Automated Workflow for Somatic and Germline Next Generation Sequencing Analysis in Routine Clinical Cancer Diagnostics. *Cancers* **2019**, 11, 1691. https://creativecommons.org/licenses/by/4.0/
4. Klingström, T.; Bongcam-Rudloff, E.; and Pettersson, O. V. A Comprehensive Model of DNA Fragmentation for the Preservation of High Molecular Weight DNA. *bioRxiv online* **2018**. https://creativecommons.org/licenses/by/4.0/
5. Kingan, S. B.; Heaton, H.; Cudini, J.; Lambert, C. C.; Baybayan, P.; Galvin, B. D.; Durbin, R.; Korlach, J.; and Lawniczak, M. A High-Quality De novo Genome Assembly from a Single Mosquito Using PacBio Sequencing. *Genes* **2019**, 10, 62. https://creativecommons.org/licenses/by/4.0/
6. Performance Characteristics of the RNA and the High Sensitivity RNA ScreenTape Assays for the 4150 TapeStation system. Agilent Technologies *technical overview*, publication number 5994-1038EN, **2019**.
7. Comparison of RIN and RQN for the Agilent Bioanalyzer and the Fragment Analyzer Systems. Agilent Technologies *technical overview*, publication number 5994-1860EN, **2020**.
8. Haller, A. C.; Kanakapalli, D.; Walter, R.; Alhasan, S.; Eliason, J. F.; and Everson, R. B. Transcriptional Profiling of Degraded RNA in Cryopreserved and Fixed Tissue Samples Obtained at Autopsy. *BMC Clin. Pathol*. **2006**, 6.
9. Alpern, D.; Gardeux, V.; Russeil, J.; Mangeat, B.; Meireles-Filho, A. C. A.; Breysse, R.; Hacker, D.; and Deplancke, B. BRB-seq: Ultra-Affordable High-Throughput Transcriptomics Enabled by Bulk RNA Barcoding and Sequencing. *Genome Biol.* **2019**, 20, 71. https://creativecommons.org/licenses/by/4.0/
10. Wimmer, I.; Tröscher, A. R.; Brunner, F.; Rubino, S. J.; Bien, C. G.; Weiner, H. L.; Lassmann, H.; and Bauer, J. Systematic Evaluation of RNA Quality, Microarray Data Reliability and Pathway Analysis in Fresh, Fresh Frozen and Formalin-Fixed Paraffin-Embedded Tissue Samples. *Sci. Rep.* **2018**, 8, 6351. https://creativecommons.org/licenses/by/4.0/
11. Mendioroz, M.; Martínez-Merino, L.; Blanco-Luquin, I.; Urdánoz, A.; Roldán, M.; and Jericó, I. Liquid Biopsy: A New Source of Candidate Biomarkers in Amyotrophic Lateral Sclerosis. *Ann. Clin. Transl. Neurol.* **2018**, 5, 763–768.
12. Blecher-Gonen, R.; Barnett-Itzhaki, Z.; Jaitin, D.; Amann-Zalcenstein, D.; Lara-Astiaso, D.; and Amit, I. High-Throughput Chromatin Immunoprecipitation for Genome-Wide Mapping of in vivo Protein-DNA Interactions and Epigenomic States. *Nat. Protoc*. **2013**, 8, 539–554.
13. Tramontano, A.; Jarc, L.; Jankowicz-Cieslak, J.; Hofinger, B. J.; Gajek, K.; Szurman-Zubrzycka, M.; Szarejko, I.; Ingelbrecht, I.; and Till, B. J. Fragmentation of Pooled PCR Products for Highly Multiplexed TILLING. *G3: Genes, Genomes, Genetics* **2019**, 9, 2657. https://creativecommons.org/licenses/by/4.0/
14. Aguirre, N. C.; Filippi, C. V.; Zaina, G.; Rivas, J. G.; Acuña, C. V.; Villalba, P. V.; García, M. N.; González, S.; Rivarola, M.; Martínez, M. C.; Puebla, A. F.; Morgante, M.; Hopp, H. E.; Paniego, N. B.; Marcucci Poltri, S. N. Optimizing ddRADseq in Non-Model Species: A Case Study in Eucalyptus dunnii Maiden. *Agronomy* **2019**, 9, 484. https://creativecommons.org/licenses/by/4.0/
15. Lan, J. H.; Yin, Y.; Reed, E. F.; Moua, K.; Thomas, K.; and Zhang, Q. Impact of Three Illumina Library Construction Methods on GC Bias and HLA Genotype Calling. *Human Immunol*. **2015**, 76, 166–175.

16. Misra, C. S.; Santos, M. R.; Rafael-Fernandes, M.; Martins, N. P.; Monteiro, M.; and Becker, J. D. Transcriptomics of Arabidopsis Sperm Cells at Single-Cell Resolution. *Plant Reprod*. **2019**, 32, 29–38.

17. Nix, D. A.; Hellwig, S.; Conley, C.; Thomas, A.; Fuertes, C. L.; Hamil, C. L.; Bhetariya, P. J.; Garrido-Laguna, I.; Marth, G. T.; Bronner, M. P.; and Underhill, H. R. The Stochastic Nature of Errors in Next-Generation Sequencing of Circulating Cell-Free DNA. *PLoS One* **2020**, 15. https://creativecommons.org/licenses/by/4.0/

18. Sidstedt, M.; Steffen, C. R.; Kiesler, K. M.; Vallone, P. M.; Rådström, P.; and Hedman, J. The Impact of Common PCR Inhibitors on Forensic MPS Analysis. *Forensic Sci. Int. Gen*. **2019**, 40, 182–191.

19. Illumina Library Sequencing Services. https://dnatech.genomecenter.ucdavis.edu/illumina-library-sequencing (accessed Jun 17, 2020).

20. Genome Sequencing Services Center. http://med.stanford.edu/gssc/faq.html (accessed Jun 17, 2020).

21. Marchal, C.; Sasaki, T.; Vera, D.; Wilson; K.; Sima, J.; Rivera-Mulia J. C.; Trevilla-García, C.; Nogues, C.; Nafie, E.; and Gilbert, D. M. Genome-Wide Analysis of Replication Timing by Next-Generation Sequencing with E/L Repli-seq. *Nat. Protoc*. **2018**, 13, 819–839.

22. Quality Control of NGS Libraries with Daisy Chain Molecules. Agilent Technologies *application note*, publication number 5994-2233EN, **2020**

23. Quail, M. A.; Kozarewa, I.; Smith, F.; Scally, A.; Stephens, P. J.; Durbin, R.; Swerdlow, H.; and Turner, D. J. A Large Genome Centre's Improvements to the Illumina Sequencing System. *Nat. Methods* **2008**, 5, 1005–1010.

24. Marosy, B. A.; Craig, B. D.; Hetrick, K. N.; Witmer, P. D.; Ling, H.; Griffith, S. M.; Myers, B.; Ostrander, E. A.; Stanford, J. L.; Brody, L. C.; and Doheny, K. F. Generating Exome Enriched Sequencing Libraries from Formalin-Fixed, Paraffin-Embedded Tissue DNA for Next-Generation Sequencing. Curr. Protoc. Hum. Genet. **2017**, 92(1), 18.10.1–18.10.25.

www.agilent.com/automated-electrophoresis

PR7000-7429

Printed in the USA, September 30, 2020
5994-2281EN

Agilent
Trusted Answers