Application Note

Genomics

Agilent
Trusted Answers

Comparison of Agilent Femto Pulse System Sizing with Long-Read Sequencing Read Length

Authors

Whitney Pike and Vera Rykalina,
Agilent Technologies

Abstract

Long-read sequencing results can be maximized by loading only long fragments onto the sequencer, thereby eliminating any issues with preferential sequencing of smaller fragments. This can be achieved through size selection to exclude the portion of the sample below a specified threshold. Sheared samples, with and without size selection, were analyzed with the Agilent Femto Pulse system, and then sequenced on the Oxford Nanopore Technologies MinION. The size distribution of the samples reported by the Femto Pulse was similar to the distribution of the sequencing read lengths. The Femto Pulse confirmed effective size selection, which was further confirmed by the sequencing results.

Introduction

Routine, robust methods for long-read sequencing are of great interest to the sequencing community. These methods have the potential to allow researchers to analyze long stretches of DNA (and sometimes entire genes) without the requirement of preamplification, and can also enable analysis of repetitive or complex regions that are otherwise difficult to map. Recent advances in long-read sequencing technology have addressed several inherent challenges, resulting in increased read lengths, throughput, and accuracy. However, certain limitations remain, such as the tendency to preferentially sequence smaller fragments[1].

One strategy to improve long-read sequencing results is size selection of the sample, which can help eliminate smaller-sized fragments prior to sequencing, and thus increase the average read lengths generated by the sequencer[2]. The Blue Pippin system from Sage Science enables targeted size selection. High-pass filtering kits for the Blue Pippin are beneficial for long-read sequencing, as they allow the user to collect all the sample above a specified threshold, enabling analysis of only the fragment length of interest. Size selection therefore makes it possible to maximize the fragment lengths that are sequenced by selectively loading long fragments onto the sequencer.

Quality control (QC) of input DNA for sequencing is important for assessing the integrity and average size of a sample and can help determine the cutoff size for size selection. These important QC checks can easily be performed using the Femto Pulse, an automated pulsed-field capillary electrophoresis system for sizing high molecular weight (HMW) genomic DNA (gDNA).

This application note demonstrates how the Femto Pulse was used in a long-read sequencing workflow. First, QC of the initial gDNA with the Femto Pulse was performed to ensure that the sample was of good integrity and high molecular weight. Next, the sheared gDNA was analyzed using the Femto Pulse to determine the size-selection cutoff to be used. Finally, the Femto Pulse was utilized to assess the quality of the size-selected DNA and confirm that the fragments below the threshold size were successfully eliminated. The sheared and size-selected samples were used to prepare libraries for sequencing with the Oxford Nanopore Technologies MinION. The mean sequencing read length was compared with the average size of the sheared and size-selected samples reported by the Femto Pulse prior to library preparation (Figure 1).

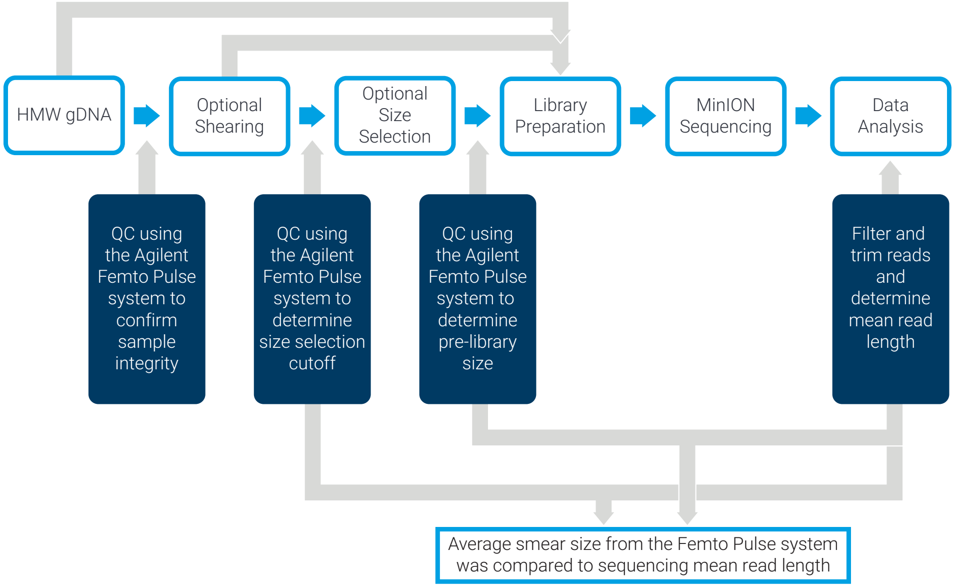

HMW gDNA
Optional Shearing
Optional Size Selection
Library Preparation
MinION Sequencing
Data Analysis
QC using the Agilent Femto Pulse system to confirm sample integrity
QC using the Agilent Femto Pulse system to determine size selection cutoff
QC using the Agilent Femto Pulse system to determine pre-library size
Filter and trim reads and determine mean read length
Average smear size from the Femto Pulse system was compared to sequencing mean read length

Figure 1. Experimental design with the typical MinION workflow (white boxes) and quality control steps (blue boxes). The Agilent Femto Pulse system was utilized throughout the workflow to determine the integrity and average smear size of the input gDNA, sheared gDNA, and size-selected gDNA. Libraries were prepared from the sheared and size-selected samples and sequenced with the MinION. The average smear sizes reported by the Femto Pulse were compared to the mean read lengths of the sequencing results. HMW: high molecular weight. QC: quality control.

Experimental

Sample shearing

Human gDNA was obtained from Promega (p/n G3041) and sheared to smaller sizes for further analysis. Shearing was performed using two methods: needle and syringe for larger sizes, or g-TUBE (Covaris, p/n 010145) for smaller sizes. A 25-gauge needle was utilized to shear a sample of 50 µL at 150 ng/µL to an average size of approximately 40 kb by aspirating 40 times. A g-TUBE was used to shear samples following the manufacturer's protocol for 20 kb. Briefly, 150 µL of gDNA at 100 ng/µL was placed into the g-TUBE and centrifuged with an Eppendorf MiniSpin at 7,600 rpm in 60 second intervals until the entire sample passed through the filter. The tube was then inverted and centrifuged again until the entire sample was in the lid, and the sample was transferred to a new tube. The sheared and unsheared samples were analyzed using the Agilent Femto Pulse system (p/n M5330AA) and the Agilent Genomic DNA 165 kb kit (p/n FP-1002)[3] to confirm sizing after shearing. A smear analysis tool in the ProSize data analysis software was utilized to determine the average size of each sample.

Size selection

Size selection was performed using the Sage Science Blue Pippin with high-pass methods[4] in order to collect the sample above a specified threshold.
40 kb sheared samples were selected using the 40 kb cutoff method (Sage Science, p/n PAC-30KB; 0.75% DF Marker U1 high-pass 30–40 kb v3 cassette definition). 20 kb sheared samples were size selected with the 15 kb cutoff method (Sage Science, p/n PAC-20KB; 0.75% DF marker S1 high-pass 15–20 kb cassette definition). The size-selected samples were analyzed on the Femto Pulse with the gDNA 165 kb kit to confirm sizing.

Library preparation and sequencing

Libraries were prepared for four samples using the Oxford Nanopore Technologies MinION with the Ligation Sequencing kit (Oxford Nanopore Technologies, p/n SQK-LSK109) according to the manufacturer's specifications[5], using 500 ng DNA input. The samples included two sheared samples, with average sizes of 20 and 40 kb, and two samples that were sheared and then size selected with cutoffs of 15 and 20 kb. Libraries were sequenced using the Oxford Nanopore Technologies MinION equipped with MinION Spot On Flow Cells (version R9.4.1, p/n FLO-MIN106D). The 15 kb size-selected and 20 kb sheared samples were sequenced on one flow cell, following the manufacturer's protocol for washing the flow cell between runs. Similarly, the 40 kb size-selected and 40 kb sheared samples (data not shown) were sequenced on a second flow cell.

Sequencing data analysis

Long-read sequencing data was generated by the MinION sequencing device and base called in real time using the MinKNOW (v20.06.5) software provided by Oxford Nanopore Technologies. Additionally, the reads were filtered based on the MinKNOW quality metric and split into 'pass' and 'fail' categories. The 'passed' reads were further processed for data analysis using the NanoPack package[6]. Briefly, the NanoLyse tool was utilized to filter fastq files to remove reads mapping to the lambda phage genome (Nanopore DNA control standard, DCS). Nanopore library adapter sequences were identified and removed from the reads using Porechop (v0.2.1, https://github.com/rrwick/Porechop). Finally, to generate sequencing read statistical summary reports and weighted histograms, the NanoPlot tool from the NanoPack suite was applied to each library. Quality data from the NanoPlot statistical analysis following filtering and trimming of the sequencing reads is shown in Table 1.

Table 1. Sequencing metrics and statistical analysis of the sequencing runs. N50: the length at which half of the nucleotides in an assembly belongs to reads equal or larger than that length. Number of reads above Q#: The number of reads that remain after removing those that fall below a failure threshold specified by the Q number (7, 10, 12, or 15), and is indicative of the accuracy of a sequencing run.

	MinION Sequencing Read Data Summary (After Trimming and Filtering)										
Sample	Mean read length (bp)	Mean read quality	Median read length (bp)	Median read quality	No. of reads	Read length N50	Total bases (in millions)	No. of reads above Q7	No. of reads above Q10	No. of reads above Q12	No. of reads above Q15
20 kb sheared	9,578.9	11.0	8,587	11.2	115,228	12,152	1,103	115,227	91,416	24,420	14
15 kb size selected	16,658.8	11.0	15,578	11.2	58,413	17,547	973	58,409	47,903	11,682	5
40 kb size selected	36,259.8	10.8	36,928	11.0	8,235	42,996	298	8,235	6,147	1,318	0

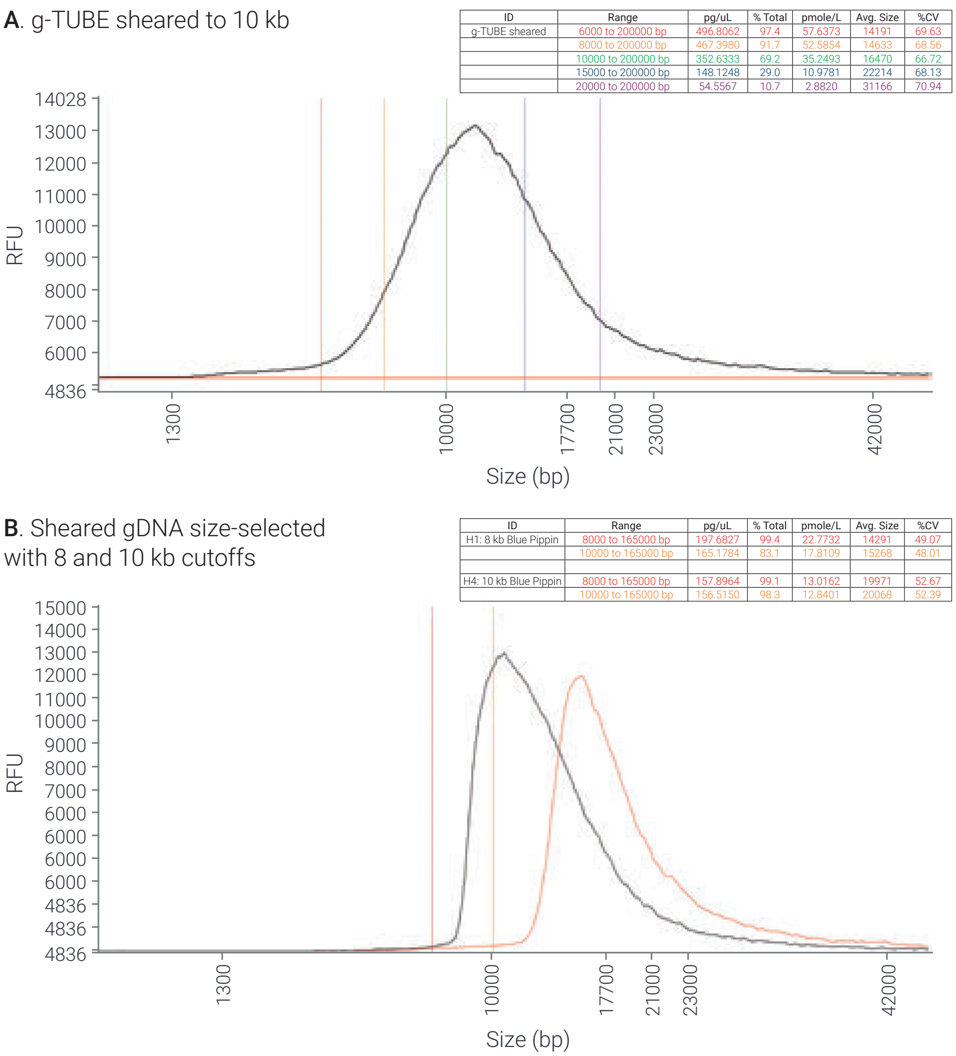
Results and discussion

Quality control of sheared gDNA to determine size-selection cutoff

Initial QC of gDNA to determine sample integrity and size provides details about a sample that can be informative to downstream applications. The optimized pulsed-field method of the gDNA 165 kb kit for the Femto Pulse is ideal for fast and automated analysis of HMW DNA through 165 kb, providing high sizing accuracy. A smear analysis tool allows for different ranges to be set to identify the portion of a sample that falls within a specific base pair range. Visualizing the sizing range of the sample on an electropherogram can help ensure that a sample is of sufficient quality for downstream applications, and help determine size selection and shearing needs. For example, as shown in Figure 2A, a gDNA sample was sheared to an average size of about 14 kb, with the sample distributed between approximately 2 and 30 kb. Implementing multiple smear ranges can identify the total percentage of the sample that is within the size range of interest and aid in determining how much of a sample will be lost during the size-selection process. Size selection with the Blue Pippin high-pass collection methods eliminates the fragments below a specific cutoff. In this example, setting the size selection too small, such as 6 kb, is unlikely to eliminate much of the sample, while setting it too big, such as 20 kb, would eliminate almost all of the sample. Size selection for this sample was thus performed at 8 and 10 kb for comparison. Thus, smear analysis can help identify where a cutoff should be set for optimal size selection of a sample and the amount of sample that would remain after size selection.

Quality control of size-selected gDNA to confirm selection

Analysis with the Femto Pulse can confirm that size selection was successful by comparing the size distribution of a sample before and after size selection. For example, Figure 2B shows a sample that has undergone size selection with the Blue Pippin. Analysis of this sample with the Femto Pulse system confirmed that the Blue Pippin effectively eliminated the portion of the sample that was below the specified cutoff (Figure 2B, black trace: 8 kb cutoff, red trace: 10 kb cutoff), when compared to the sample prior to size selection (Figure 2A). Additionally, the Femto Pulse reported the average smear size of the remaining sample following size selection. The average size of the size-selected smear will appear larger than the non-size-selected sample. Since the smaller fragments are eliminated from the smear, the distribution of the size-selected sample is shifted to the right and the average size is increased. In addition to smear size, simple visual inspection of the sample distribution and comparison of the sample before and after size selection is the best indicator of a successful size selection.



A. g-TUBE sheared to 10 kb

ID	Range	pg/uL	% Total	pmole/L	Avg. Size	%CV
g-TUBE sheared	6000 to 200000 bp	496.8062	97.4	57.6373	14191	69.63
	8000 to 200000 bp	467.3980	91.7	52.5854	14633	68.56
	10000 to 200000 bp	352.6333	69.2	35.2493	16470	66.72
	15000 to 200000 bp	148.1248	29.0	10.9781	22214	68.13
	20000 to 200000 bp	54.5567	10.7	2.8820	31166	70.94

B. Sheared gDNA size-selected with 8 and 10 kb cutoffs

ID	Range	pg/uL	% Total	pmole/L	Avg. Size	%CV
H1: 8 kb Blue Pippin	8000 to 165000 bp	197.6827	99.4	22.7732	14291	49.07
	10000 to 165000 bp	165.1784	83.1	17.8109	15268	48.01
H4: 10 kb Blue Pippin	8000 to 165000 bp	157.8964	99.1	13.0162	19971	52.67
	10000 to 165000 bp	156.5150	98.3	12.8401	20068	52.39

Figure 2. The Agilent Femto Pulse system can help identify size-selection cutoffs and confirms successful size selection of gDNA. (A) gDNA was sheared to an average size of approximately 10 kb and analyzed on the Femto Pulse. Multiple smear analysis ranges were set to determine an appropriate size-selection threshold for the sample, based on the sample size distribution and the % total (inset table). (B) Size selection using the Sage Science Blue Pippin was performed with cutoff thresholds set at both 8 kb (black) and 10 kb (red). Analysis with the Femto Pulse confirmed that size selection successfully eliminated the portion of the sample below the cutoff size (indicated by the red (8 kb) and orange (10 kb) lines), as shown by the electropherogram and the increased average size of the sample.

Comparison of average smear size and mean sequencing read length

Short-read next-generation sequencing (NGS) technologies primarily rely on generating numerous short reads that can be utilized for reference-based or *de novo* assemblies using complex bioinformatics approaches. Long-read sequencing technologies aim to overcome some of the limitations of short-read sequencing. For instance, long reads can be beneficial for resolving regions of the genome that include large stretches of repetitive regions and where short reads cannot be mapped uniquely. Long reads can identify specific transcript isoforms by reading through entire RNA fragments. In addition, long-read sequencing can eliminate amplification and consequently the amplification bias that is often associated with short-read platforms. Long-read sequencing with the MinION (Oxford Nanopore Technologies) provides real-time data of the entire fragment length being sequenced, simplifying assembly of larger genomes or detection of structural variants. With any sequencing platform, the process of library preparation converts a sample into a technology-specific format. For the MinION, this involves ligating an adapter that includes a tether and a motor protein to the sample. The library is brought to the surface of a nanopore protein and the DNA is translocated through the nanopore as it is sequenced. As individual nucleotides pass through the nanopore, electrical current changes are monitored, and the signal is decoded in real time into the sequence of the nucleic acid. Since only the DNA, and not the motor or tether proteins, is translocated through the nanopore, the size of the sequenced DNA fragment should be the same as the size of the input DNA.

While DNA samples can be sequenced on the MinION in their native state, Oxford Nanopore Technologies does provide recommendations for optional shearing and size selection steps, which are thought to help increase the read lengths and the quality of the sequencing. The addition of the motor and tether proteins to the DNA during library preparation makes sizing of the final library unreliable, so the recommended QC checkpoints prior to sequencing include the input gDNA and the optional shearing and size-selection steps (Figure 1). QC of the initial gDNA was performed to ensure that the sample was intact and of high quality (Figure 3). An aliquot of the gDNA was then sheared using a g-TUBE following the protocol for a 20 kb shear. Half of the sheared sample was retained (20 kb g-TUBE), while the other half underwent size selection with a 15 kb cutoff using the Blue Pippin (15 kb BP). QC of the sheared (Figure 4A) and size-selected (Figure 4B) samples was performed on the Femto Pulse with the 165 kb method prior to library preparation and sequencing on the Oxford Nanopore Technologies MinION. The mean size and distribution of the read lengths was determined (Figure 4C, D), and compared pairwise to the average smear size and size distribution of the samples reported by the Femto Pulse following shearing and size selection (Figure 4A, B).

As seen in the Femto Pulse electropherograms, the sheared sample has a somewhat bell-shaped curve, with tailing towards the right side. A small amount of sample smearing to the left side is indicative of the smaller molecular weight portion of the sample (Figure 4A). The read length histograms generated from the sequencing data show a similar distribution pattern to the Femto Pulse results, with a sharp peak tailing to the right, as well as the presence of smaller fragments to the left of the main peak (Figure 4C). As smaller samples are preferentially sequenced on the MinION, the amount of reads of smaller sizes appears slightly disproportionate in comparison to the Femto Pulse, and the mean sequence read length of the sheared sample is smaller than the average smear size of the DNA reported by the Femto Pulse (Figure 4E).

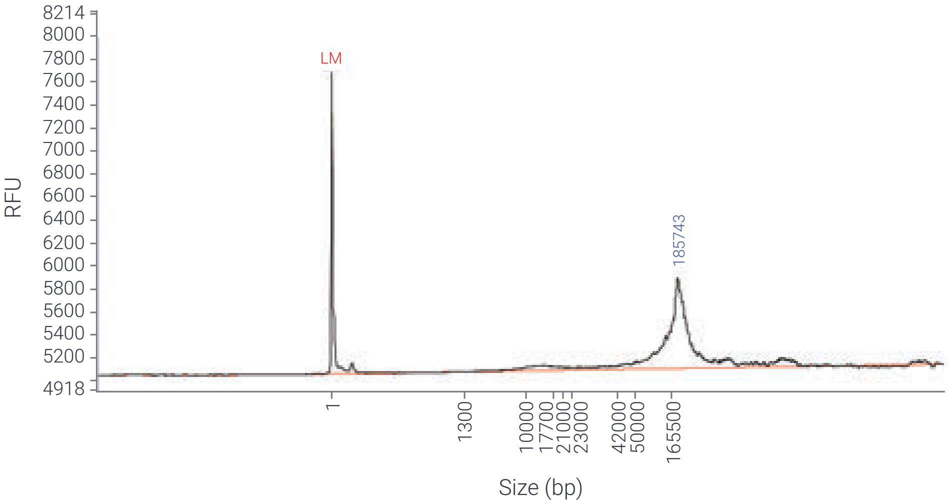


Figure 3. Quality control of gDNA using the Agilent Femto Pulse system. The size and integrity of the sample was confirmed prior to shearing and size selection for a long-read sequencing workflow. The electropherogram of the sample displays a sharp peak at approximately 185 kb with only small amounts of smearing to either side, indicating highly intact gDNA.

The Femto Pulse electropherogram of the size-selected sample also displays a smear tailing to the right side. However, on the left-hand side, the size-selected sample shows a sharp cutoff with the smaller-sized fragments of the sample no longer present, indicating a successful size selection (Figure 4B). The weighted read length histogram of the size-selected sample also demonstrates depletion of the sample below 15 kb (Figure 4D). The average size of the DNA reported by the Femto Pulse and the mean sequencing read length are similar (Figure 4E). Agreement between the sequencing results and the Femto Pulse size distributions confirms successful size selection.

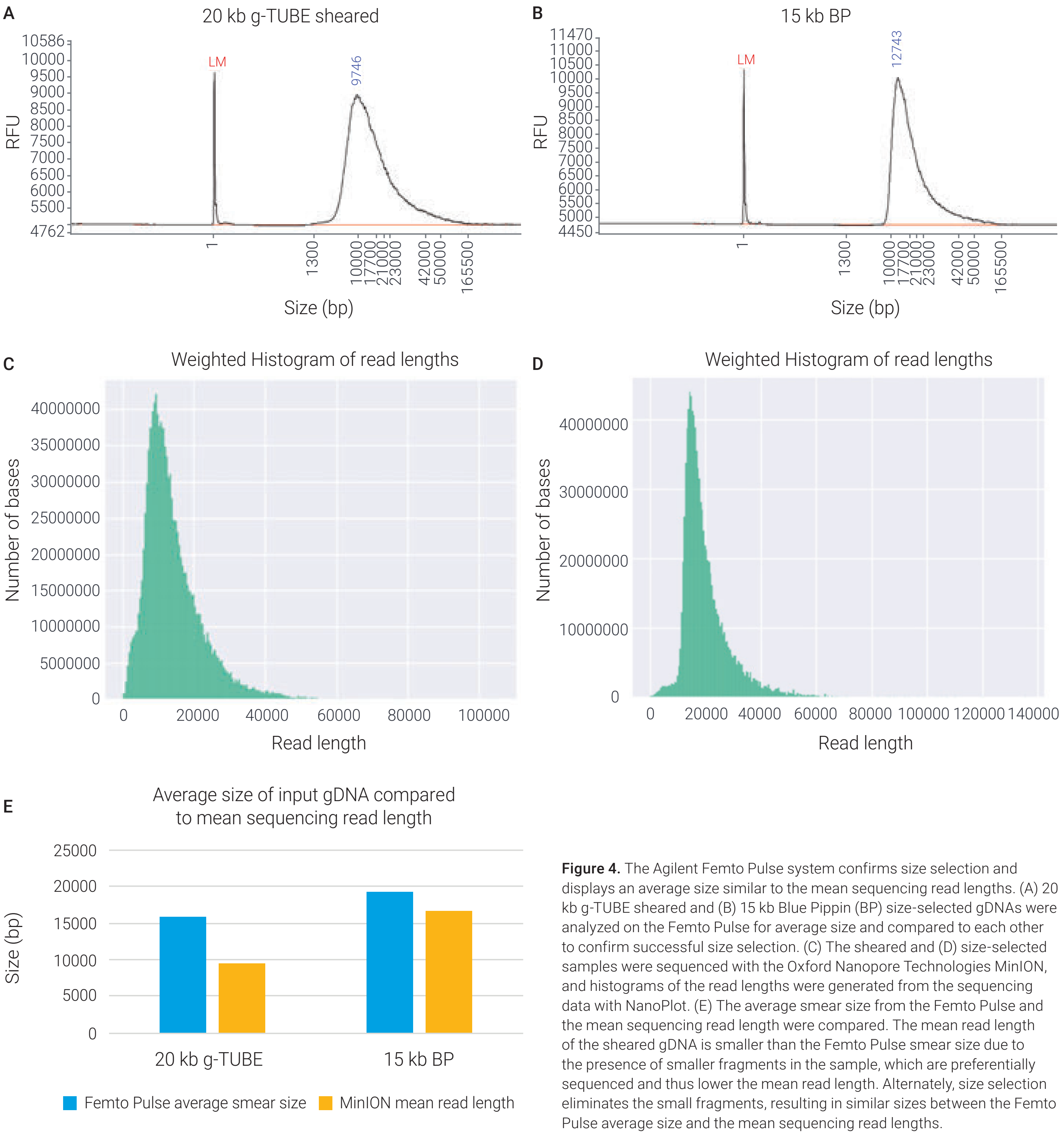


Figure 4. The Agilent Femto Pulse system confirms size selection and displays an average size similar to the mean sequencing read lengths. (A) 20 kb g-TUBE sheared and (B) 15 kb Blue Pippin (BP) size-selected gDNAs were analyzed on the Femto Pulse for average size and compared to each other to confirm successful size selection. (C) The sheared and (D) size-selected samples were sequenced with the Oxford Nanopore Technologies MinION, and histograms of the read lengths were generated from the sequencing data with NanoPlot. (E) The average smear size from the Femto Pulse and the mean sequencing read length were compared. The mean read length of the sheared gDNA is smaller than the Femto Pulse smear size due to the presence of smaller fragments in the sample, which are preferentially sequenced and thus lower the mean read length. Alternately, size selection eliminates the small fragments, resulting in similar sizes between the Femto Pulse average size and the mean sequencing read lengths.

To further examine the relationship between the Femto Pulse smear analysis and the MinION sequencing read lengths, a second sample was sheared and size selected with a 40 kb cutoff. The size-selected sample was analyzed on the Femto Pulse prior to sequencing (Figure 5A). While the majority of the sample below 40 kb was successfully omitted, a small portion of sample remained to the left of the peak, between 30 and 40 kb. This pattern was reproducible amongst several replicates (n = 4), indicating that the size-selection cutoff may not be as sharp at this larger size, compared to the 15 kb size-selected sample. The sequencing results showed the presence of some fragments under 20 kb (Figure 5B), indicative of even smaller fragments that are preferentially sequenced because of their small size. The average size of the sample analyzed on the Femto Pulse was therefore larger than the mean sequencing read length (Figure 5C). However, the size distribution of the sample on the Femto Pulse confirmed size selection, and this was further confirmed by the sequencing read length histogram, which displayed a similar size distribution pattern to the Femto Pulse (Figure 5B).

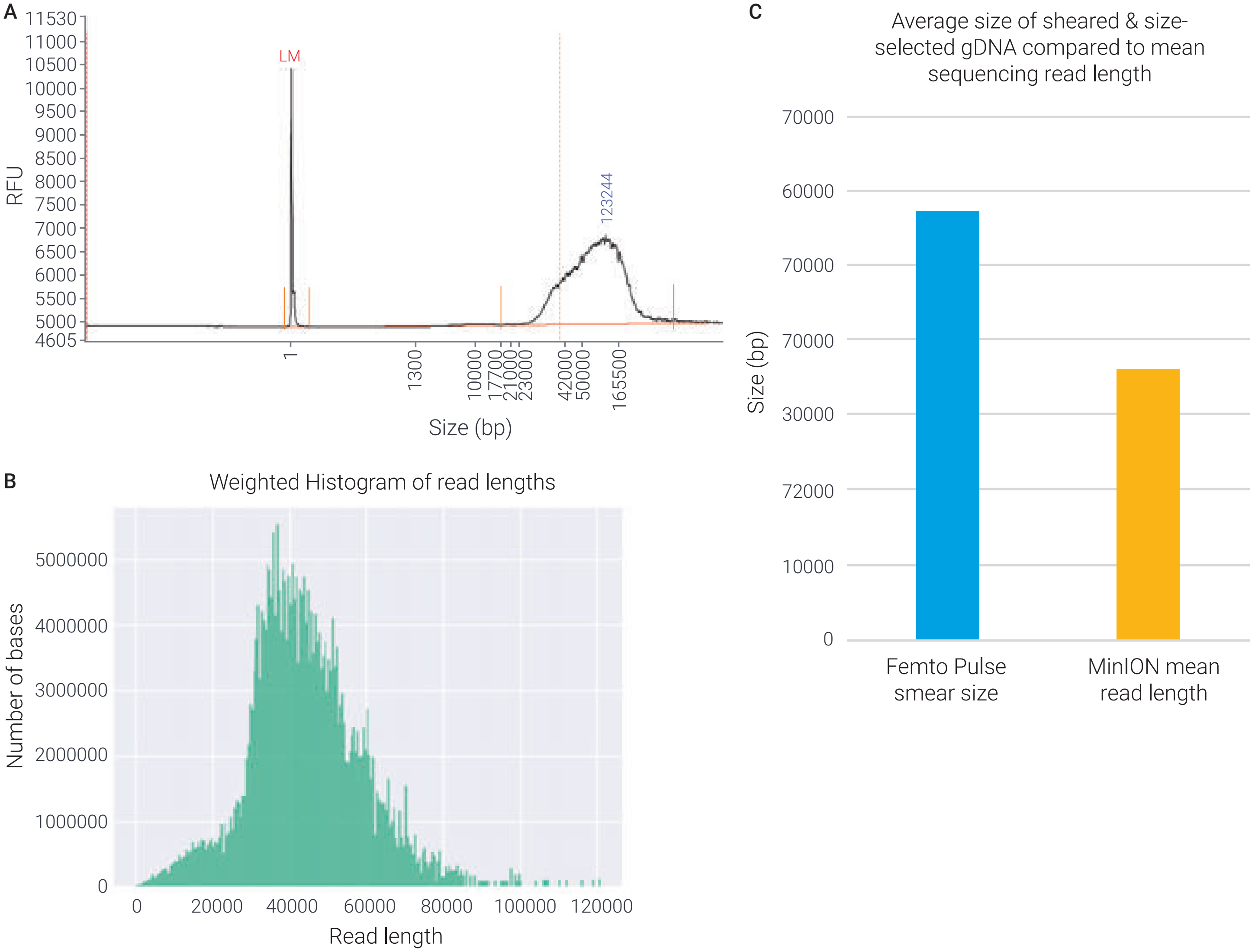


Figure 5. Size selection and sequencing of a larger gDNA sample. A gDNA sample was sheared and size selected using a 40 kb cutoff. (A) The Agilent Femto Pulse system was used to confirm size selection, with most of the sample sizing above 40 kb (red line). (B) A weighted histogram of the Oxford Nanopore Technologies MinION sequencing results confirms the Femto Pulse distribution, with most of the reads around 40 kb and above. The smaller sizes shown are indicative of the presence of smaller fragment lengths in the sample that may be preferentially sequenced. (C) Comparison of the average smear size reported by the Femto Pulse and the mean read length reported by NanoPlot analysis of the sequencing results.

Conclusion

While long-read sequencing methods provide a number of advantages over short-read NGS approaches, they can be challenging due to the preferential sequencing of short fragments. In this study, the Agilent Femto Pulse system was used to address these challenges by successfully analyzing gDNA quality and size. This crucial information can help determine appropriate size-selection thresholds and aid in quality assessment for samples prior to downstream analysis. The distribution of the size-selected sample on the Femto Pulse electropherogram confirmed successful size selection, with the elimination of smaller fragments. Successful size selection was further confirmed by sequencing results using the Oxford Nanopore Technologies MinION. The mean sequencing read lengths aligned well with the average smear sizes reported by the Femto Pulse prior to library preparation, indicating the sizing accuracy of the Femto Pulse.

References

1. De Roeck, A.; De Coster, W.; Bossaerts, L.; Cacace, R.; De Pooter, T.; Van Dongen, J.; D'Hert, S.; De Rijk, P.; Strazisar, M.; Van Broeckhoven, C.; Sleegers, K. NanoSatellite: Accurate Characterization of Expanded Tandem Repeat Length and Sequence through Whole Genome Long-Read Sequencing on PromethION. *Genome Biol*. **2019**, *20* (1), 239.
2. Schalamun, M.; Nagar, R.; Kainer, D.; Beavan, E.; Eccles, D.; Rathjen, J. P.; Lanfear, R.; Schwessinger, B. Harnessing the MinION: An Example of How to Establish Long-Read Sequencing in a Laboratory Using Challenging Plant Tissue from *Eucalyptus pauciflora*. *Mol. Ecol. Resour*. **2019**, *19* (1), 77–89.
3. Agilent Genomic DNA 165 kb Kit Quick Guide for Femto Pulse System. *Agilent Technologies kit guide*, publication number SD-AT000141, **2020**.
4. Blue Pippin User Guide for High-Pass DNA Size Selection. *Sage Science*, document number 460047, **2018**.
5. Nanopore Protocol for Genomic DNA by Ligation (SQK-LSK109). *Oxford Nanopore Technologies*, version GDX_9095_v109_revB_24Jan2020, **2020**.
6. De Coster, W.; D'Hert, S.; Schultz, D. T.; Cruts, M.; Van Broeckhoven, C. NanoPack: Visualizing and Processing Long-Read Sequencing Data. *Bioinformatics* **2018**, *34* (15), 2666–2669.

www.agilent.com

For Research Use Only. Not for use in diagnostic procedures.
PR7000-7796

This information is subject to change without notice.

Printed in the USA, April 1, 2021
5994-3078EN

Agilent
Trusted Answers