Range + T Size Selection Method Increases PacBio HiFi Read Lengths

sage science

Application Note: PippinHT

Victoria Bunting, Jayson Talag, Seunghee Lee, Dario Copetti
Arizona Genomics Institute, University of Arizona

THE UNIVERSITY OF ARIZONA
COLLEGE OF AGRICULTURE & LIFE SCIENCES
Arizona Genomics Institute

Background

A key feature of PacBio library construction for HiFi sequencing is the combination of shearing and size-selection techniques to limit library size distributions to a 15-18 kb target range. During HiFi sequencing, each SMRTbell molecule is read multiples times, giving a consensus accuracy of Q30 or more (≥99.9% accuracy) for each molecule. Since the average number of subreads per SMRTbell is determined by the polymerase read length, for maximum HiFi read yield the library size distribution must be controlled to balance insert read length and quality scores to acceptable limits.

The current HiFi library construction workflow utilizes shearing with the Diagenode Megaruptor 3 to reduce the input DNA size distribution down to 10-20 kb with a mode in the 15-18 kb size window. After SMRTbell construction, a final size selection step is performed mainly to eliminate shorter library elements (<10 kb). Although PacBio promotes magnetic bead size selection protocols for this final size selection, such protocols do not provide as stringent elimination of short fragments as gel-based technologies, such as those provided by the Sage Science BluePippin, PippinHT, and SageELF instruments. Moreover, it is difficult to control the size range of selection with bead methods.

To address the need for improved size-selection for PacBio HiFi libraries, Sage Science recently introduced a new programming mode, called "Range + T", and a new cassette definition with improved resolution over the size range of 9-30 kb. With the Range + T programming mode, users set a starting size threshold for initiation of product collection, and a collection time (usually 5-30 minutes). Sage has found this programming mode allows better control over product size distribution in the size range relevant for PacBio HiFi libraries. The intention of the Range + T mode is to give users a programming method which can be easily customized for different sheared input DNAs. The advantages are a) quantitative removal of short SMRTbells below the low MW cutoff value, b) increased control over the upper size limit of the library, and c) increased potential to adjust the breadth of the final library by varying low or high MW cutoffs (or both).

Testing PippinHT Range + T software for plant whole genome sequencing at the Arizona Genomics Institute

The Arizona Genomics Institute (AGI) was given early access to the new Sage Science PippinHT Range + T mode (software version v1.14-Cassette Definition set 17, with cassette definition "0.75% 9-30kb R+T 75E") to evaluate its capability for increasing the average size of PacBio HiFi libraries. Longer HiFi read length is potentially important for plant genomics because of the higher complexity (size, repeats, polyploidy) of plant genomes.

The first experiment tested the effect of different elution times on output recovery and size distribution. Wheat genomic DNA was sheared with Megaruptor 3 to a range between 13 to 25 kb (mean size 19 kb), and size selection was performed on the PippinHT. DNA elution started at 15 kb and carried out for 13 to 30 minutes. Figure 1 shows the traces of the DNA fractions of each sample. The longer elution time had a clear effect on the size distribution of the recovered DNA, and a considerable fraction of the output was greater than 40 kb when the 30-minute elution was used. The overlayed Femto Pulse traces highlight the accuracy and precision of the Range + T mode (**Figure 1**). The collection starting point was very close across all gel lanes, and the size distribution of the product was correlated with elution time. The 30 minute elution time allowed recovery of most of the HMW DNA that was present in the sheared samples.

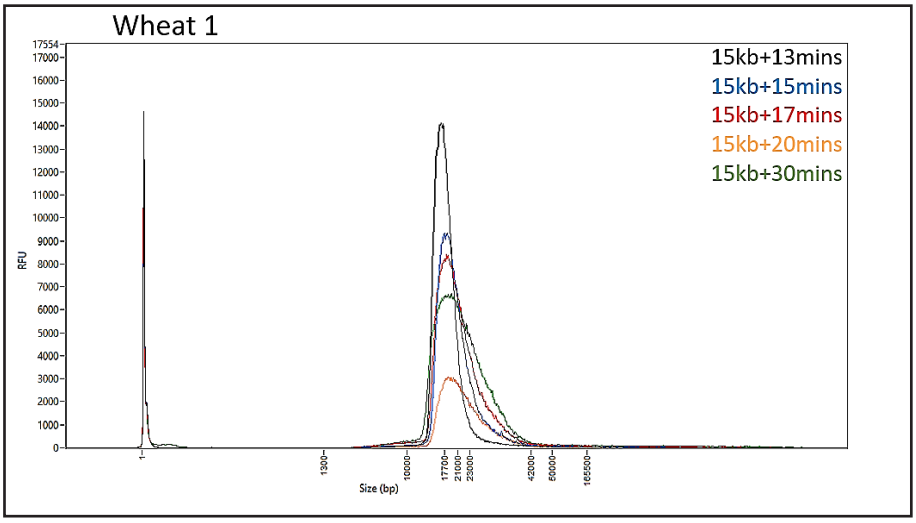

Figure 1. Adjustment of size range of collection using PippinHT Range + T mode

Furthermore, short library elements were dramatically reduced by the PippinHT size selection. This reduction in short library fragments can be seen in **Figure 2**, where size profiles of sheared (in black and blue, the two sheared wheat DNA samples that were pooled for size selection) and size-selected product (red trace) are overlayed. It is apparent how DNA smaller than the size-selection cutoff was dramatically reduced in the product, whereas most of the HMW DNA was recovered. The clean product size distribution was also reflected in the read length profile of the resulting PacBio sequencing run (see **Figure 2** inset).

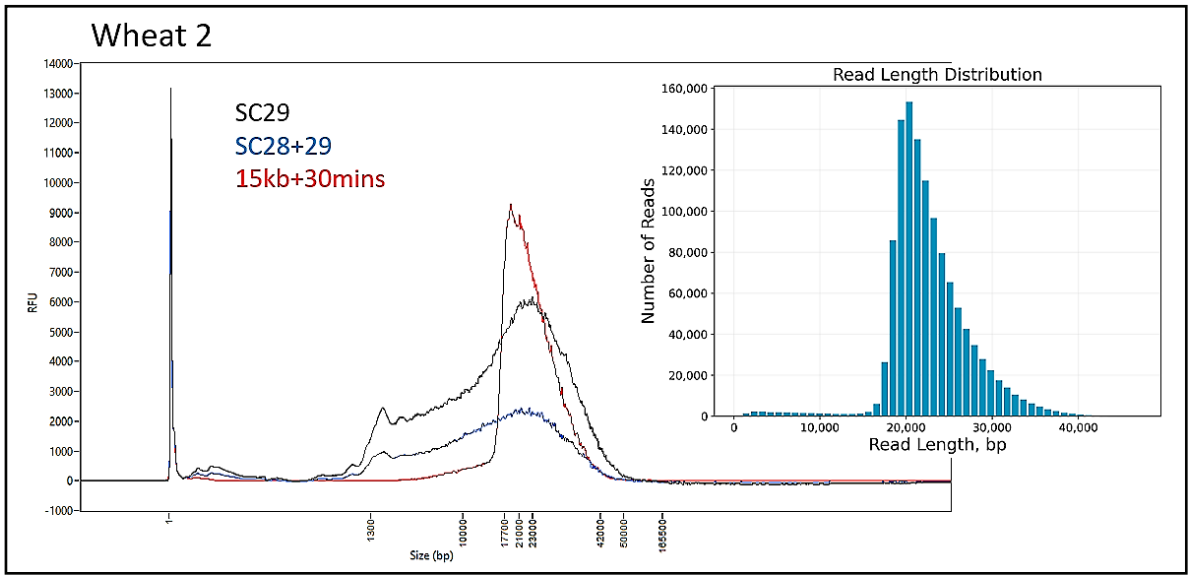

Figure 2. Stringent reduction of small SMRTbells using PippinHT Range + T mode at 15kb threshold.

DNA size-selection using the PippinHT was reproducible with respect to yield and size distribution. On 16 size selections from wheat (2), cassava (8), and maize (6) samples, recovery ranged from 17 to 45% of input sheared DNA (**Table 1**). The DNA from a set of 8 cassava samples was sheared to a target range of 13-25 kb (mean size ~18 kb) and DNA was selected at a low cutoff of 13 kb and eluted for 30 minutes. Size selection recovery ranged between 33 to 45%, and the mean library size was 19-20 kb with the longest DNA fragments reaching 40 kb. Femto Pulse traces of the eight samples showed accurate and consistent lower cutoffs (**Figure 3**). Mean HiFi read lengths were at 18-19 kb (**Table 1**), in agreement with the mean size of the selected DNAs.

Sample	Loaded DNA		R+T size selection		Recovered DNA			Mean read length (kb)
	mean size (kb)	amount (ug)	start size (kb)	elution time (min)	mean size (kb)	amount (ug)	% recovered	
Wheat 1	19.1 and 19.8	6.00	15	30	22.5	1.95	32.5	21.0
Wheat 2	19.3 and 19.0	6.00	15	30	24.7	1.26	20.3	22.6
Cassava 1	18.4	2.40	13	30	19.2	0.92	38.5	18.7
Cassava 2	18.1	2.80	13	30	19.2	1.22	43.4	18.1
Cassava 3	18.5	2.80	13	30	19.2	1.03	36.7	18.5
Cassava 4	18.9	2.80	13	30	20.0	1.27	45.3	19.0
Cassava 5	18.2	2.80	13	30	19.4	1.05	37.5	18.2
Cassava 6	18.3	2.80	13	30	19.1	0.94	33.5	18.5
Cassava 7	18.3	2.40	13	30	20.6	0.79	32.8	19.1
Cassava 8	18.6	3.00	13	30	20.9	1.09	36.3	19.4
Maize K	24.8	1.40	17	30	27.1	0.29	20.7	22.4
Maize A	25.2	1.40	17	30	27.1	0.58	41.3	22.9
Maize K	24.8	2.80	18	30	26.3	0.49	17.5	22.6
Maize M	23.6	2.80	18	30	25.2	0.85	30.3	22.5
Maize M	23.6	2.80	20	30	26.7	0.89	31.8	23.2
Maize A	25.2	2.80	20	30	29.0	0.85	30.5	24.5

Table 1. PippinHT Range + T recovery, product size, mean HiFi subread length

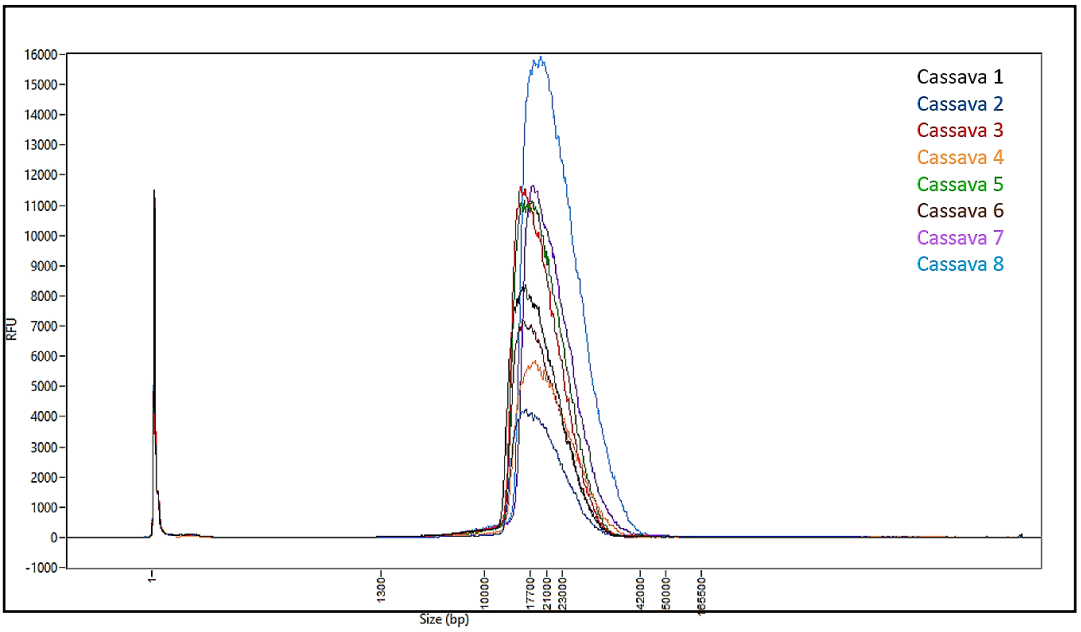

Figure 3. Reproducibility of PippinHT size selections at 13 kb R+T threshold

To assess the capability of the PippinHT Range+T software in size-selection at higher molecular weights, genomic DNA of three maize samples were sheared to a 15-35 kb range (mean size 23-25 kb, **Figure 4)**. The samples were size selected at three different cutoff values ranging from 17 to 20 kb and eluted for 30 minutes (**Table 1**). DNA recovery rates varied from 17 to 41% with a detectable maximum fragment length of ~ 50 kb (**Figure 4**). The wider range and lower amount of recovered DNA of these samples was a consequence of the higher cutoff value adopted and the size distribution of the sheared DNA. Increasing the low cutoff size produced a distinctly larger output size distribution and consequently, a longer mean HiFi read length (**Table 1** and **Figure 5**).

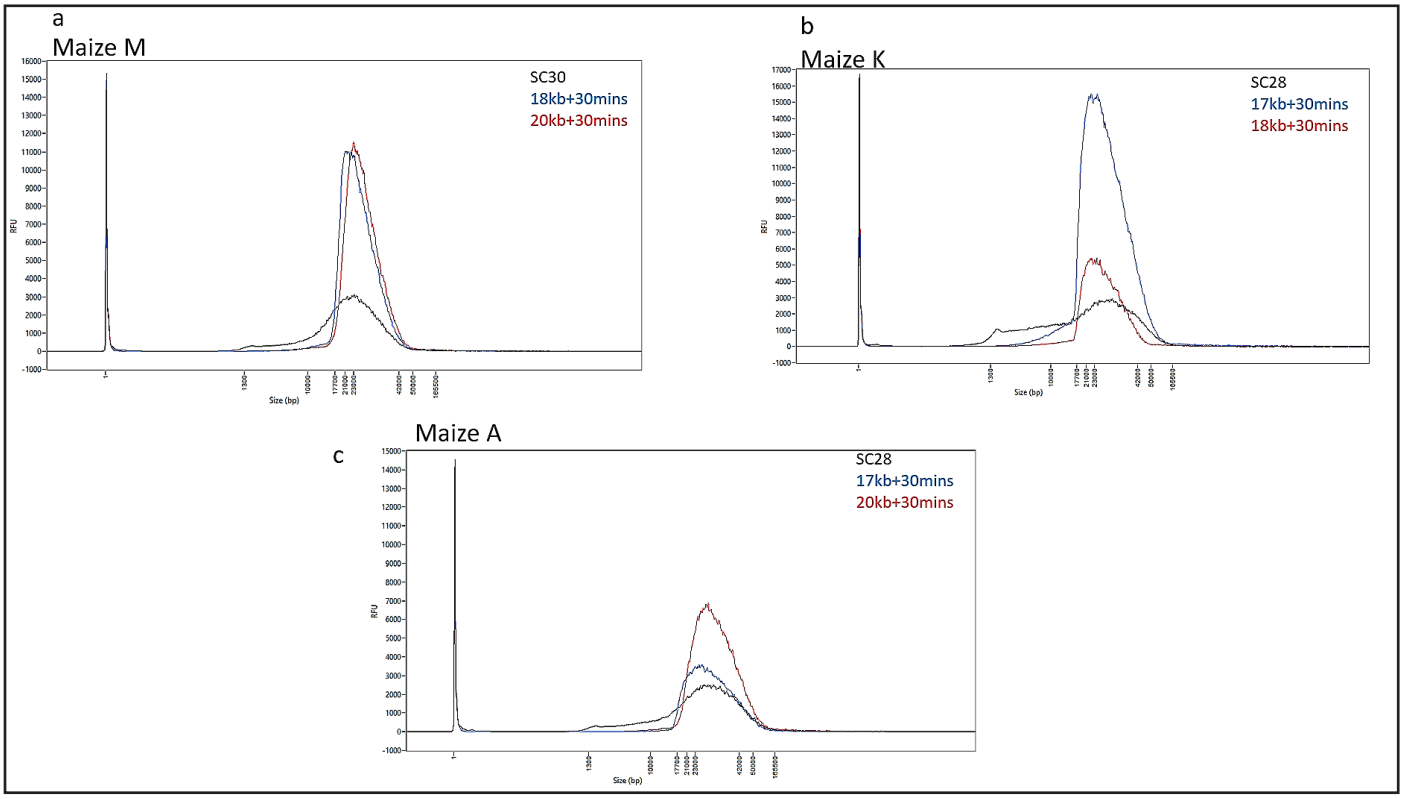

Figure 4. PippinHT Range +T selections at sizes >15kb

Effect of longer SMRTbell libraries on HiFi sequencing yield and quality scores

In order to evaluate the balance between sequencing yield, read quality, and mean HiFi read length, we compared data from 25 libraries size-selected at 9-13 kb with the current PippinHT software (PippinHT Software v1.13, Cassette Definition Set 16; blue symbols in **Figure 5**), and 59 libraries size-selected at 15 kb or above using the early access PippinHT Range+T software (orange symbols in **Figure 5**). As discussed above, libraries generated using Range + T size selection above a 15 kb cutoff consistently resulted in mean read lengths above 21-22 kb (wheat and maize samples), often with a noticeable amount of reads in the 30-40 kb range (**Figure 2** inset). By comparison, library generated using size selections at 9-13 kb sizes resulted in mean HiFi read lengths between 13 to 17 kb. Importantly, these data (**Figure 5a**) indicate that there is minimal to no decrease in HiFi sequence yield with increasing mean HiFi subread length over the range of 13 kb to 22 kb. Since the PacBio sequencing chemistry has a finite polymerase read length, shorter SMRTbell libraries had higher QV scores than longer libraries, but the lowest mean QV score for the longer libraries was still quite good at 28, and 88% of the longer libraries (orange symbols) had mean QV scores of 30 or greater (range of mean QV scores 32-35 for short libraries; range 28-33 for the longer libraries, **Figure 5.b-c**). To better illustrate the quality of the longer libraries, **Figure 6** shows the read quality distribution for one maize M sample that had an average subread length of 22.6 kb, where ~40% of the reads scored Q30 or higher.

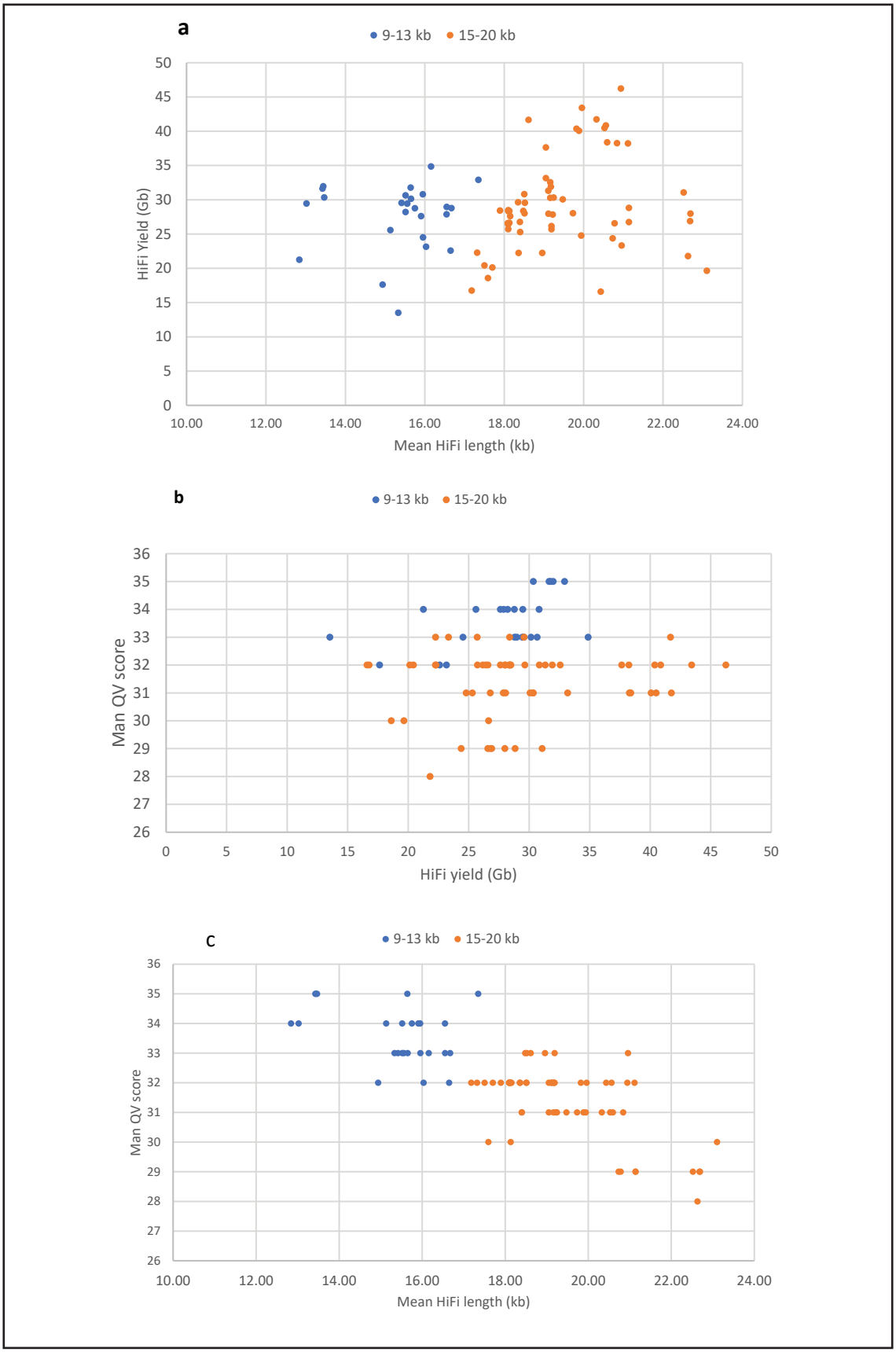

Figure 5. HiFi sequencing yield, mean QV data from libraries using different PippinHT Range + T size-selection settings.

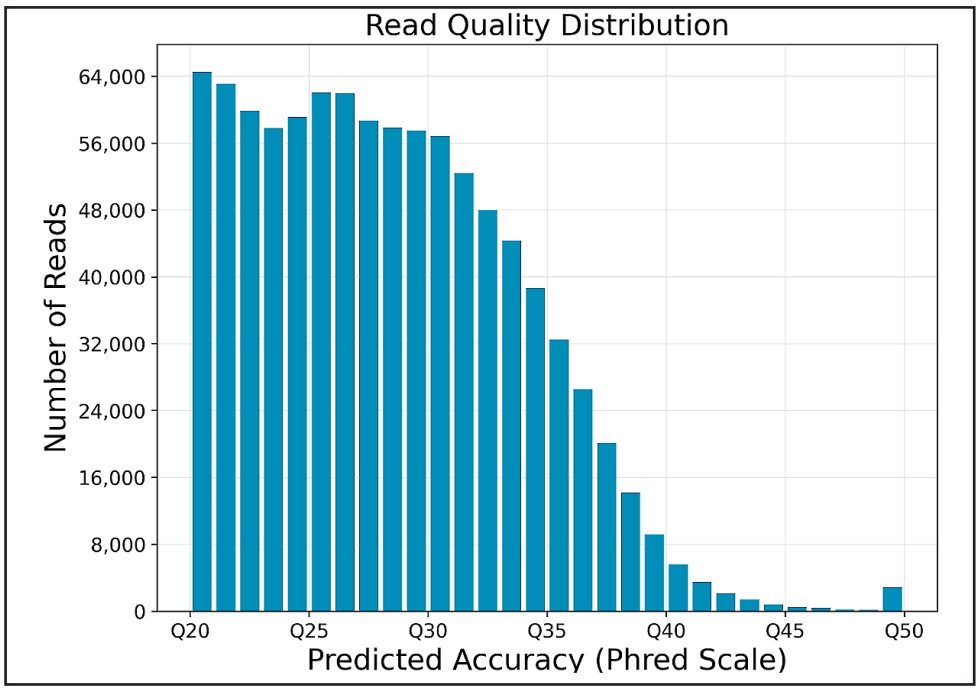

Figure 6. Read quality distribution, PippinHT Range+T SS, mean HiFi subread length 22.6 kb, Zea mays library.

Impact of longer PippinHT size-selection on HiFi assembly quality

To model the impact of longer reads towards more complete and contiguous genome assemblies, we assembled three datasets created from a Range + T size-selected PacBio HiFi library (Maize M sample of **Table 1**). The parent library had 18× coverage with an average HiFi read length of 20.6 kb. Three datasets of different features and equal coverage were generated as described below (**Table 2**):

- "Small" dataset: 4000 bp were removed from the 3' of all reads. This resulted in a subset of the reads that preserves the shape of the original read distribution, but the entire distribution is shifted to shorter size, with a mean length of 16 kb.
- "Long" dataset: the original 20 kb library was downsampled to ~14.5× coverage to match the number of bases of the "Small" dataset.
- "Narrow" dataset: 3' terminal bases of all the longest reads were truncated to a value of approximately 17 kb to simulate a SMRTbell library with a narrow distribution. The overall coverage was also reduced to match the other two datasets.

While the "Small" dataset was representing the current standard for PacBio library prep, the "Long" dataset was used to prove the benefits of a higher and wider library size selection allowed by the Range + T mode. Additionally, the "Narrow" dataset was developed to assess the contribution of the longest reads in improving assembly metrics. Each dataset was assembled independently with HiFiasm. Total assembly size did not change and matched the expected 2.4 Gb value, meaning that the amount of bases was not limiting in obtaining a complete assembly. As expected, the "Long" dataset had the lowest number of contigs, highest mean contig length and highest Nx values (**Table 2**) when compared to the assemblies obtained with the other read formats.

	Small	Long	Narrow
Reads (#)	2,080,374	1,672,324	2,091,595
Mean read length (kb)	16.72	20.61	16.58
Total bases (Gb)	34.83	34.52	34.78
Haploid genome coverage (x)	14.51	14.39	14.49
Total size (Gb)	2.40	2.40	2.46
Contigs (#)	1,732	1,438	5,536
Mean contig length (Mb)	1.38	1.67	0.44
Longest contig (Mb)	57.42	91.18	81.88
N50 (Mb)	11.09	13.56	10.98
L50 (#)	62	52	62
N70 (Mb)	6.61	8.59	6.53
L70 (#)	119	95	121
N90 (Mb)	2.64	2.62	1.49
L90 (#)	231	192	260

Table 2. Assembly statistics, Maize M down-sampled datasets

Conclusion

The use of the new Sage Science PippinHT Range + T software with the 9-30kb cassette definition enables flexible, reproducible size selection of PacBio HiFi libraries with average subread lengths pushing beyond the 20 kb current standard. Additionally, the new mode allows the use to control the upper size range of the library.

Our efforts prove that, when coupled with Megaruptor 3-sheared input DNA, samples selected with Sage Science Pippin HT using the new Range + T mode have a larger average size, negligible residual short inserts, and are enriched in >30 kb fragments.

The larger libraries did not show a tradeoff with SMRT cell yield and had only (expected) slightly lower QV scores than smaller HiFi libraries. In some cases, the larger insert size yielded around 45 Gb per SMRT cell – a value impossible to reach with libraries of shorter inserts. Such "large and wide" libraries result in as much or better yield per SMRT cell than standard, 16 kb, narrow libraries, and contribute to more contiguous assemblies of large and complex genomes.