PACIFIC BIOSCIENCES®

Getting the most out of your PacBio® Libraries with Size Selection

Susana Wang, John Harting, Elizabeth Tseng, Richard Hall, Primo Baybayan
Pacific Biosciences, Menlo Park, CA

Introduction

PacBio® RS II sequencing chemistries provide read lengths beyond 20 kb with high consensus accuracy. The long read lengths of P4-C2 chemistry and demonstrated consensus accuracy of 99.999% are ideal for applications such as *de novo* assembly, targeted sequencing and isoform sequencing. The recently launched P5-C3 chemistry generates even longer reads with N50* often >10,000 bp, making it the best choice for scaffolding and spanning structural rearrangements. With these chemistry advances, PacBio's read length performance is now primarily determined by the SMRTbell™ library itself.

Size selection of a high-quality, sheared 20 kb library using the BluePippin™ System has been demonstrated to increase the N50 read length by as much as 5 kb with C3 chemistry. BluePippin size selection or a more stringent AMPure® PB selection cutoff can be used to recover long fragments from degraded genomic material. The selection of chemistries, P4-C2 versus P5-C3, is highly dependent on the final size distribution of the SMRTbell library and experimental goals.

PacBio's long read lengths also allow for the sequencing of full-length cDNA libraries at single-molecule resolution. However, longer transcripts are difficult to detect due to lower abundance, amplification bias, and preferential loading of smaller SMRTbell constructs. Without size selection, most sequenced transcripts are 1-1.5 kb. Size selection dramatically increases the number of transcripts >1.5kb, and is essential for >3kb transcripts.

** N50>X defined as half of the data in reads with length greater than X bp.*

Large Genome Scaffolding with P5-C3

- The new P5 polymerase and C3 Chemistry combined with 3-hr data collection are ideal for generating long reads for gap closing or scaffolding large genomes such as the highly repetitive Maize genome
- Maximum long read benefits of P5-C3 can be achieved by constructing and sequencing 20 kb SMRTbell library size-selected using the BluePippin™ system

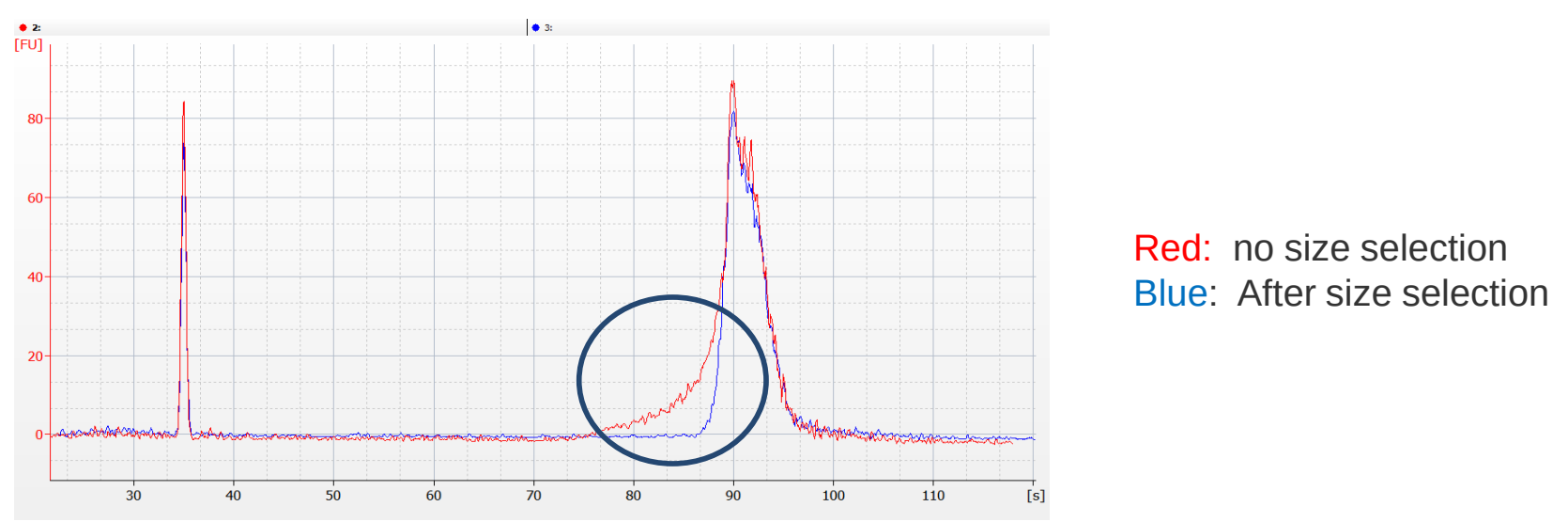

Figure 1. Size selection with the BluePippin system using a cutoff threshold between 10 kb to 50 kb removes short insert SMRTbells (<10kb). Removal of short insert SMRTbells is key to generating long subread lengths with P5-C3.

Bases	N50 subread length	95th Percentile	Longest subread length	Longest polymerase read length
20.2 Gb	10,838	20,240	35,964	43,521

Table 1. P5-C3 sequencing metrics of size-selected 20 kb Maize library. 50% of the data come from reads greater than 10,838 bp

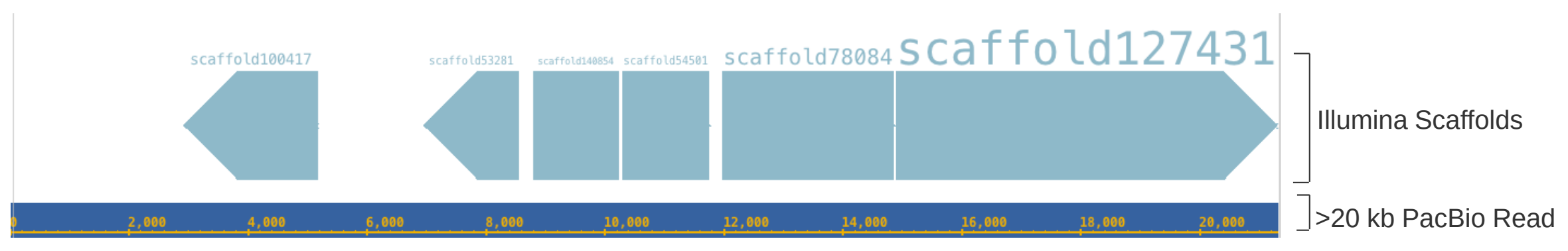

Figure 2. An example of one >20 kb PacBio P5-C3 read that spans 6 contigs from an Illumina® assembly with >1.6 million scaffolds.

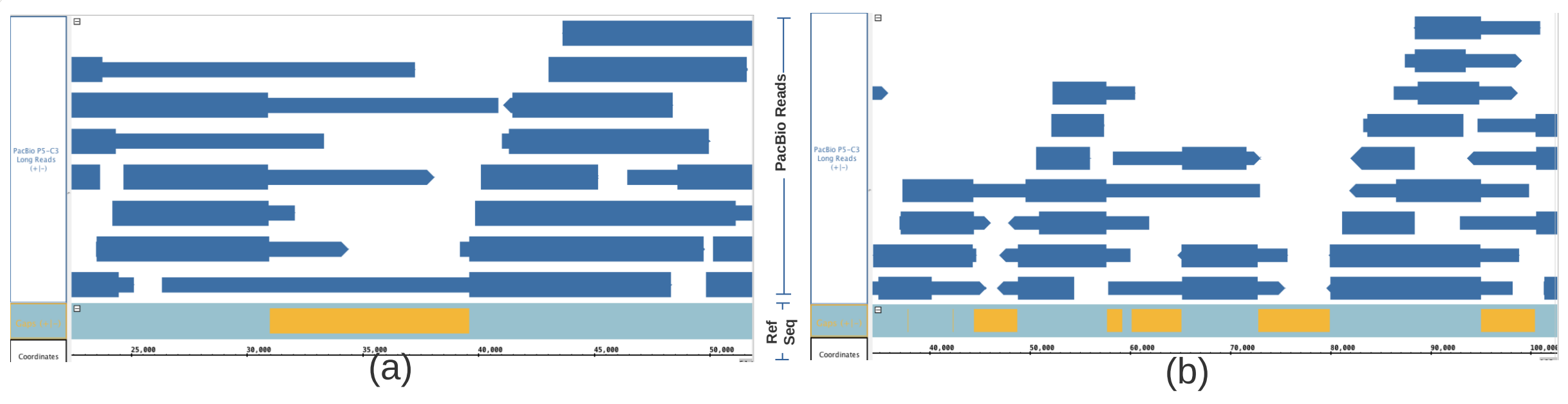

Figure 3. Examples of P5-C3 reads spanning across gaps in Illumina scaffolds. Gaps were filled using PBJelly[1] for high accuracy. The gap in (a) is ~10 kb, while (b) contains multiple gaps in a 60 kb region of a larger scaffold. Thick/thin bars indicate mapped/unmapped regions of a single read.

- With 5X coverage of PacBio reads, the initial total gap size of 254 Mb was reduced down to 145 Mb, a 43% reduction in the upper 50th percentile of scaffolds.
- Additional coverage of ~25X of P5-C3 is necessary to further improve assembly of this highly repetitive genome.

See also Poster P044 – Latest Sequencing Chemistry Performance on *Arabidopsis* Genome.

Size Selecting Degraded Samples

- When input DNA is already fragmented to the desired size or smaller, shearing is not necessary and may further reduce library insert size.
- For these samples, it is very important to remove shorter fragments that will be much less beneficial in assembly. The BluePippin system is the preferred method of size selection, if sample quantity is sufficient.
- Shown below, SMRTbell libraries from partially degraded samples can be successfully prepared and sequenced in the PacBio RS II to generate long read lengths

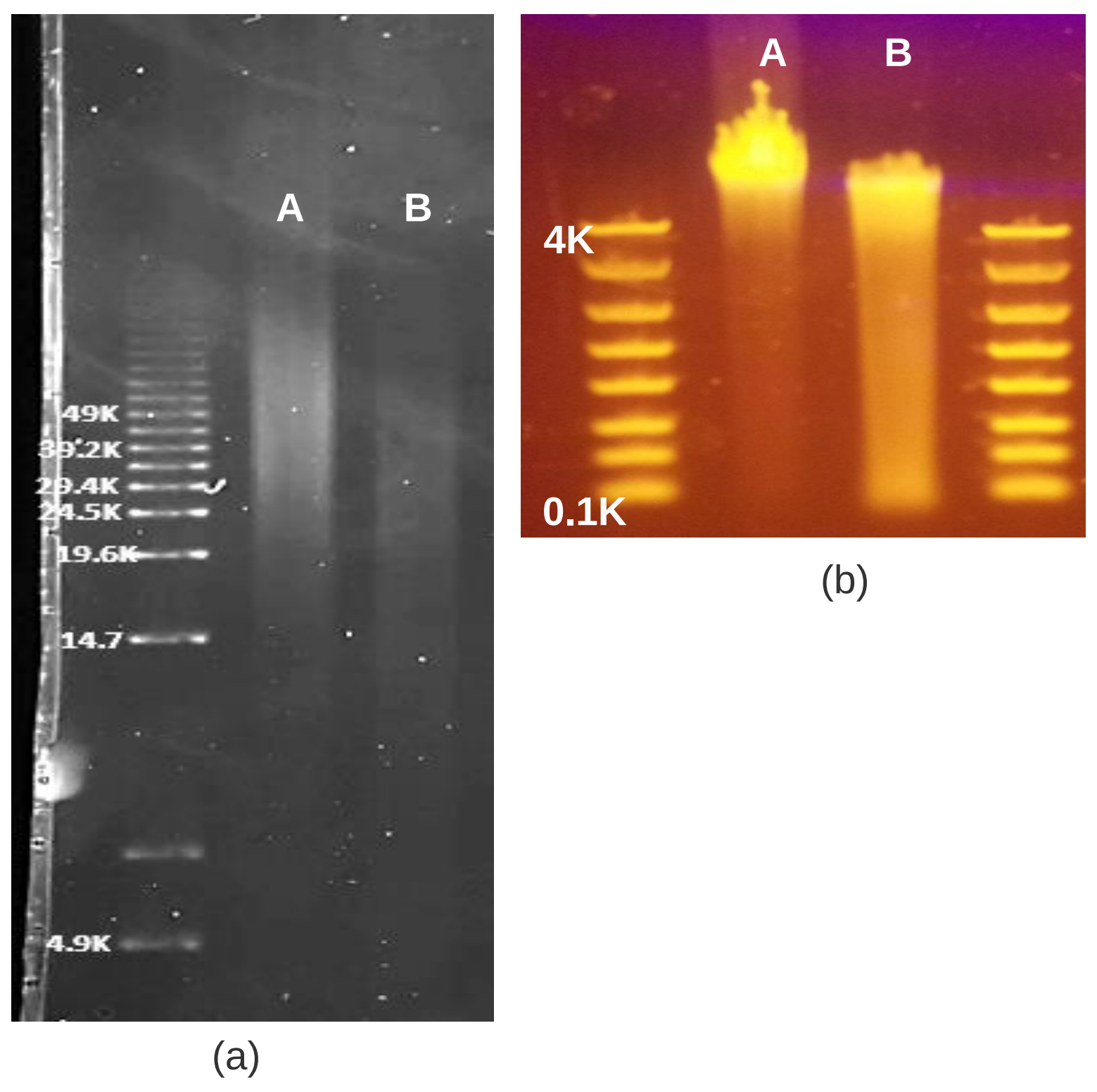

Figure 4. The quality of the genomic DNA from two *M.tuberculosis* strains (A and B) were analyzed on Chef Mapper® from Bio-Rad (a) and a Lonza™ FlashGel® (b). Both samples show degradation as evidenced by the smears on both gels.

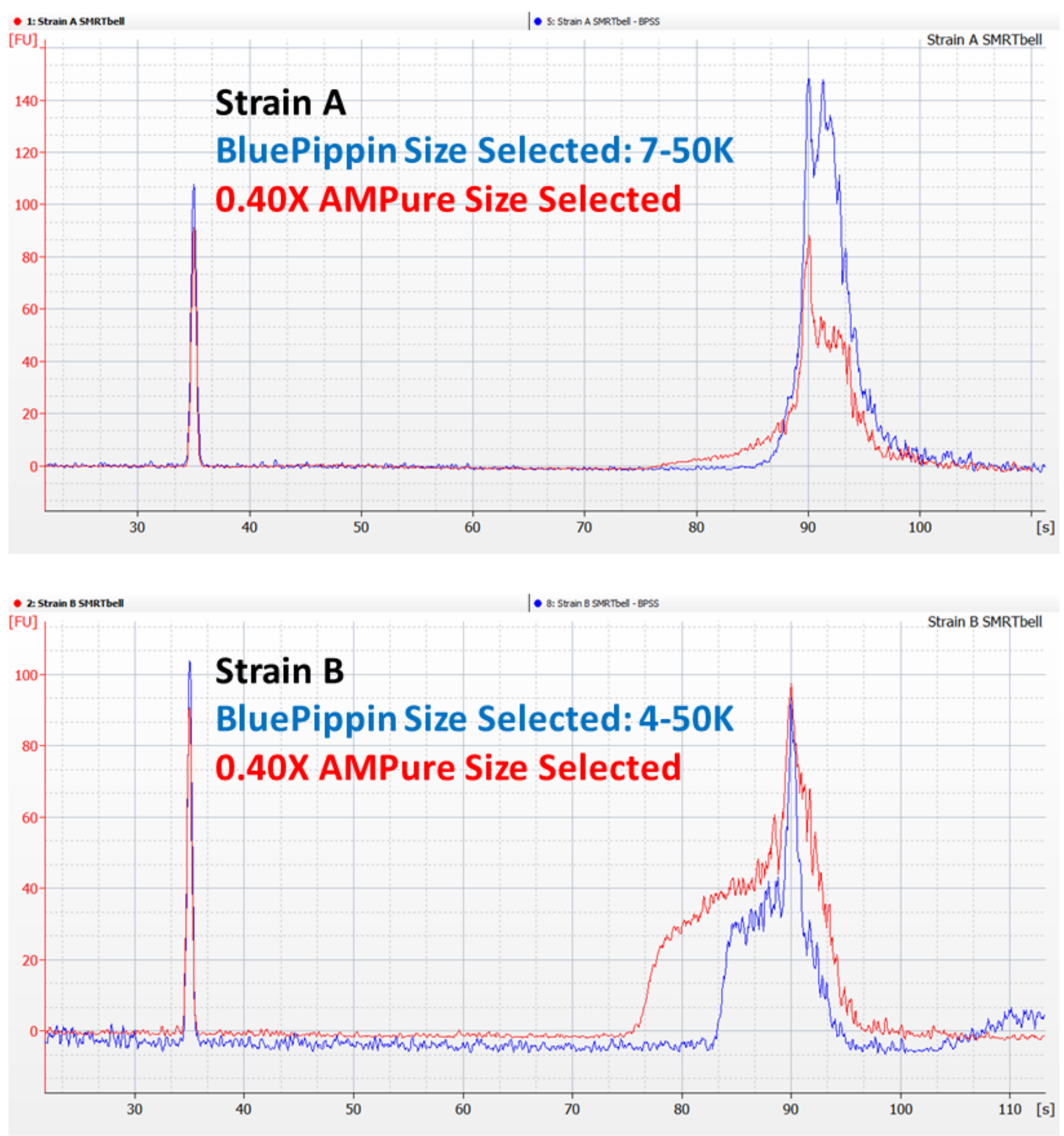

Figure 5. Distribution of fragment lengths in two 20 kb *M.tuberculosis* libraries following size selection with either AMPure PB beads (red) or the BluePippin system (blue). Since the genomic DNA were partially degraded, SMRTbell Libraries were constructed without additional fragmentation.

		Expected # Contigs	# Contigs	# SMRT® Cells	Max PreAssembled Read	PreAssembled Average RL	PreAssembled N50
Strain A	AMPure PB	1	9	2	13,587	3,134	3,390
	BluePippin 7-50K	1	2	2	17,243	4,871	7,004
Strain B	AMPure PB	2*	40	2	8,646	2,473	2,623
	BluePippin 4-50K	2*	8	2	19,208	7,828	8,841

Table 2. Assembly statistics for two *M.tuberculosis* strains prepared with AMPure PB and BluePippin size selection. The N50 preassembled read length of BluePippin improved by a factor of >2 resulting in fewer contigs compared to the AMPure PB purified libraries. Data was assembled using the HGAP[2] assembly method.

*Non-clonal population—extra contig reflects 1.6kb deletion in one sub-population.

Maximizing Long Reads in Iso-Seq Sequencing

- Long read lengths allow sequencing of transcript isoforms from high-quality poly(A) RNA using PacBio's Iso-Seq method.
- Full-length, intact transcripts are defined by the detection of both 5' and 3' PCR primers.
- To capture full diversity of transcripts, we recommend three size fractions for each cDNA sample: 1-2 kb, 2-3 kb, and 3-6 kb.
- Size selection can be performed by excision from agarose gels (traditional gel cuts), or with the BluePippin system

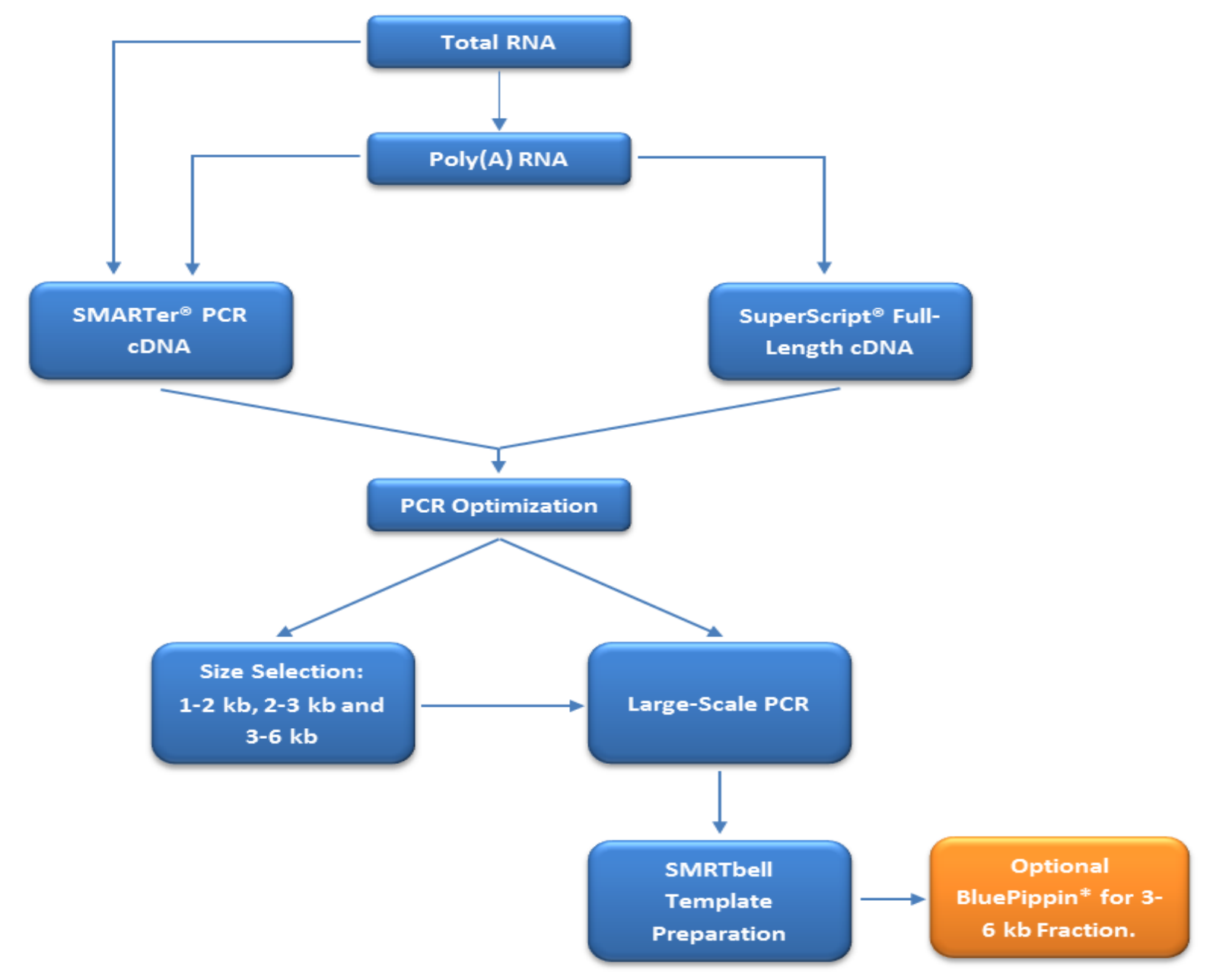
Total RNA
Poly(A) RNA
SMARTer® PCR cDNA
SuperScript® Full-Length cDNA
PCR Optimization
Size Selection: 1-2 kb, 2-3 kb and 3-6 kb
Large-Scale PCR
SMRTbell Template Preparation
Optional BluePippin* for 3-6 kb Fraction.

Figure 6. Workflow overview of the Iso-Seq library preparation procedure (available in SampleNet) http://www.smrtcommunity.com/SampleNet

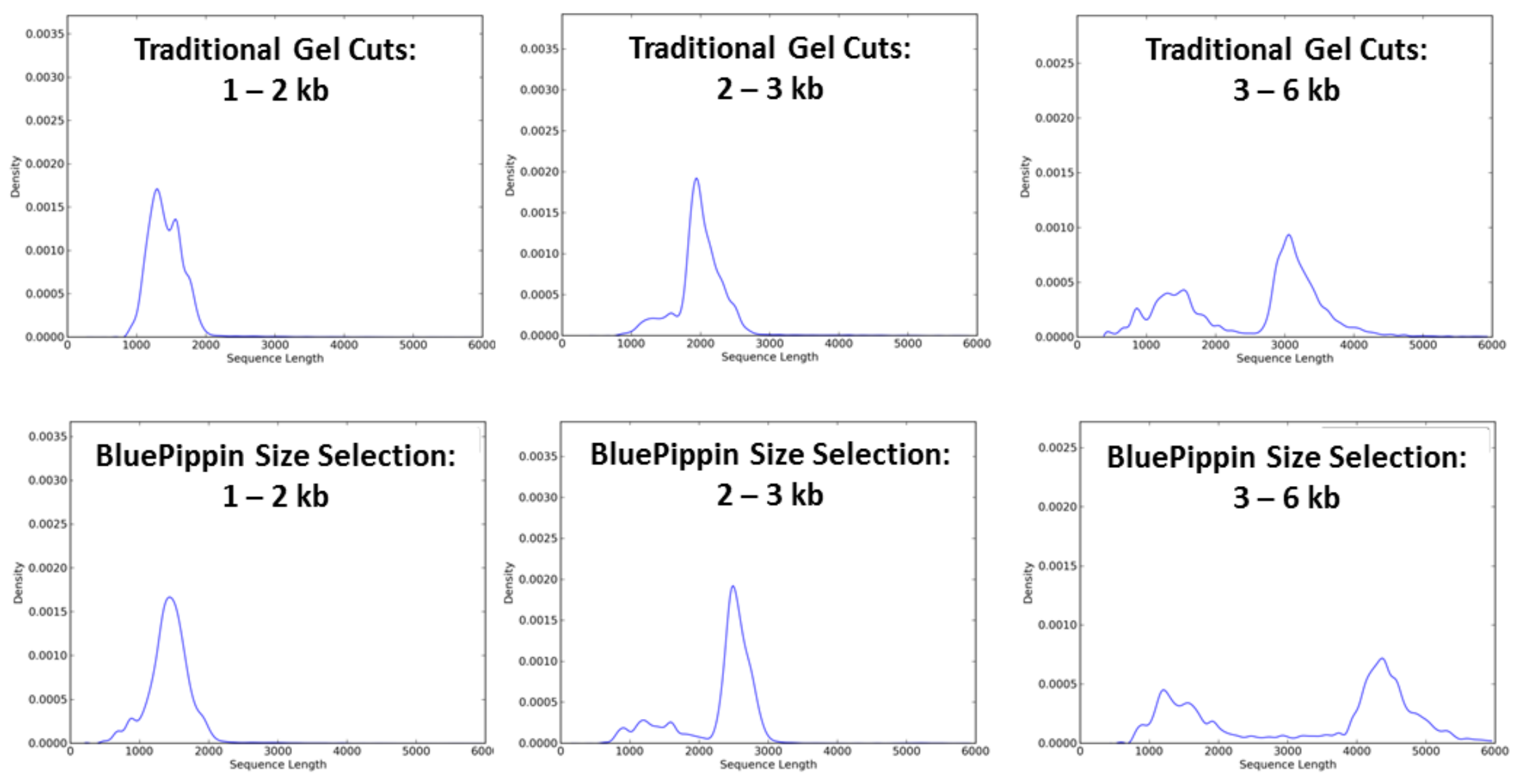

Figure 7. Distribution of sequence lengths of three size fractions, using two sizing strategies. BluePippin size-selection contains slightly longer transcripts compared to traditional gel cuts. Libraries were sequenced with P4 polymerase-C2 sequencing chemistry.

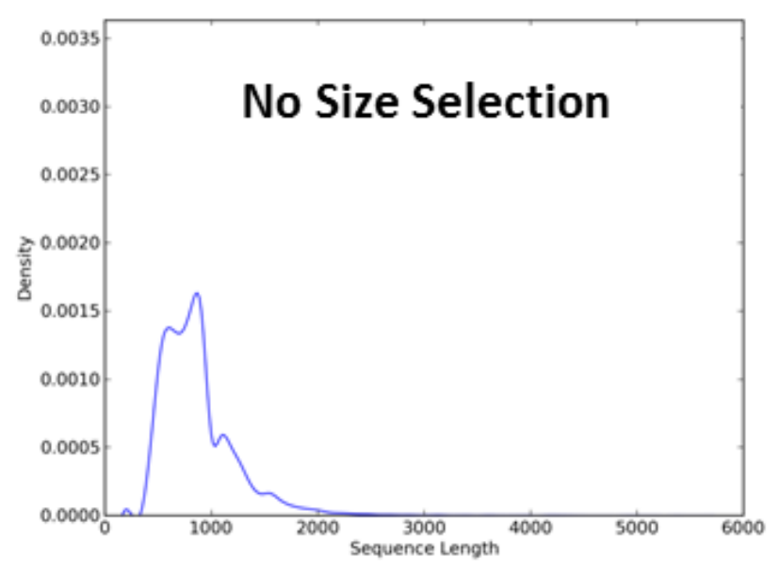

Figure 8. Distribution of sequence lengths with no size selection. Depending on your application needs, size selection may not be required. Without size selection, the majority of sequences will be from transcripts 1-1.5 kb in length.

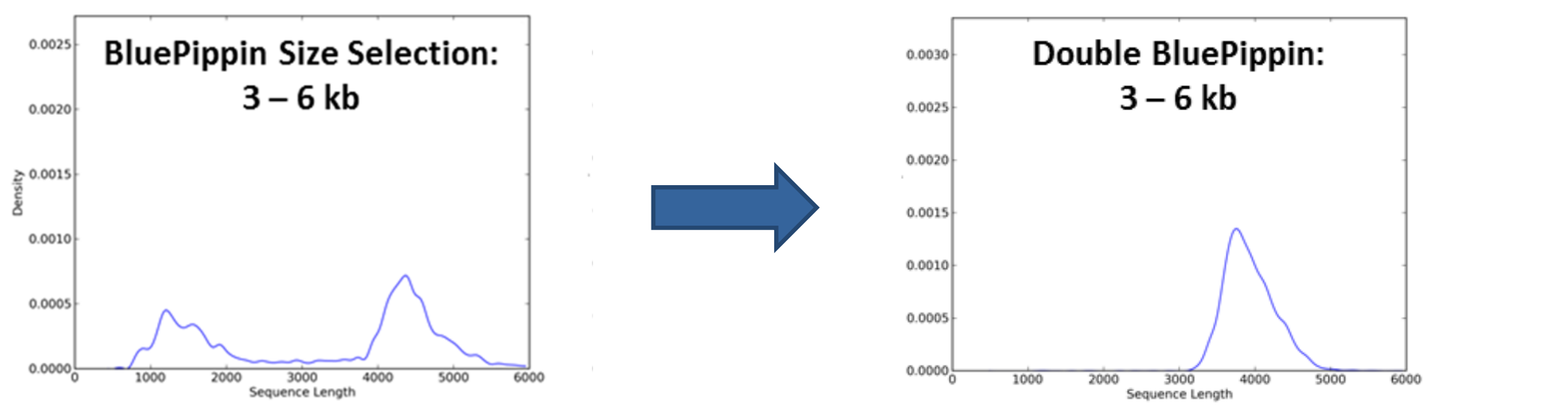

Figure 9. If your interest is to sequence only long transcripts, a second round of size selection of the 3-6 kb SMRTbell library using BluePippin eliminates the majority of short transcripts.

Conclusions

Long continuous read lengths are essential for applications such as *de novo* assembly and Isoform Sequencing (Iso-Seq).

- The combination of P5 polymerase-C3 chemistry with a high quality BluePippin size-selected library has resulted in N50 subread length > 10.8 kb, enabling gap closure or contig scaffolding for complex genomes such as maize.
- Size selection using the BluePippin system greatly increases insert sequence lengths and assembly results from partially degraded gDNA, even with a low cutoff such as 4 kb.
- With Iso-Seq, size selection of transcripts allows the detection of isoforms up to 6 kb. With no size selection, the average transcript size is generally 1-1.5 kb. (See Poster P043)

Acknowledgements

The authors would like to thank our collaborator, Mark A. Mikel et al. from University of Illinois at Urbana-Champaign and Ellen Paxinos and Jenny Gu, Pacific Biosciences, for their contributions to this poster

REFERENCE:
[1] English AC, *et al*. "Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology." *PLoS ONE*, 7(11):e47768 (2012).
[2] Chin et al. (2013) Nonhybrid, finished microbial genome assemblies from long read SMRT sequencing data. *Nature Methods* 10, 563-669 doi:10.1038/nmeth.2474