Next-Generation Size Selection for Optimized Long-Read Sequencing Workflows

Yourgene Health's Ranger Technology can enable precise size selection to prepare PacBio SMRTbell DNA libraries for high-fidelity sequencing

By Joanne Mason, PhD

Whole genome analysis plays a critical role in the development of life-saving diagnostics, therapeutics, and vaccines, with growing interest in noncoding regions and structural variants, which may impact gene activity and provide novel insights into disease pathogenesis.

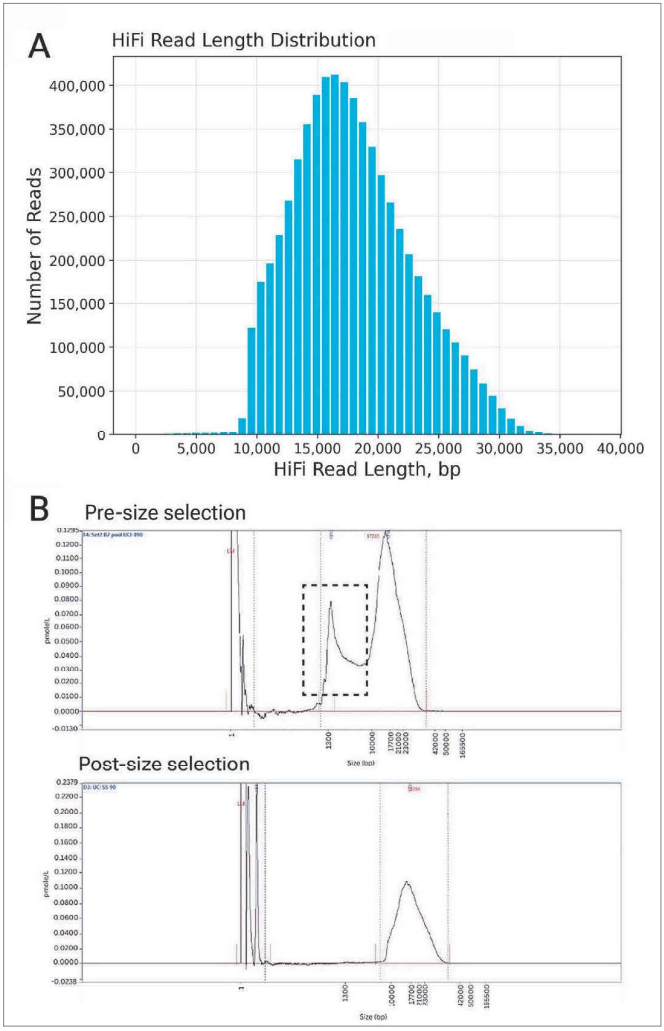

Figure 1. Representative HiFi read length distribution (A) and comparison of pre- and post-size selection on the Agilent Femto Pulse system (B) for 10 kb cutoff demonstrates effective removal of 10 kb fragments.

Some genomic regions are more difficult to sequence. Centromeres and telomeres contain highly repetitive sequences; regions that are AT- or GC-rich respond poorly to the amplification protocols required by some platforms; and palindromic sequences or hairpin structures are difficult to denature, making such regions challenging for sequencing tools that include a denaturation step. In addition, long reads that sequence through repeat regions and structural rearrangements make it easier to define changes in the genome that we have typically found challenging to reconstruct with the bioinformatics tools associated with short-read sequencing technologies.

Single-molecule real-time long-read sequencing

Long-read sequencing, in particular, could help to advance genomics by resolving some of the most challenging regions of the human genome, discerning previously inaccessible locations and making sense of the deserts of noncoding material. Through the utilization of longer fragments, the full spectrum of genetic variation could be revealed, offering greater context and giving opportunities for the discovery of novel mechanisms of disease.

Long-read sequencing provides several advantages compared to other next-generation sequencing methods—the most obvious being the actual sequence length. It can also manage regions containing repeats and structural variants, and it can define methylation from native DNA samples.

Long-read sequencing produces genomic data by generating individual reads ranging from 1,000 to 20,000 nucleotides or more in length, while most short-read sequencing technologies use fragments that are 50 to 300 bases long and often lose vital genetic information through the amplification process.

Single Molecule Real-Time (SMRT) sequencing from PacBio uses single molecules of DNA immobilized on a nanofluidic chip in miniaturized wells, known as zero-mode waveguides (ZMWs). A polymerase incorporates labeled nucleotides, and light emission is measured in real time.

The SMRTbell library format involves capping both sides of the DNA fragments with ligated hairpin adapters, where the sequencing primers attach. This creates a circular template for the polymerase to navigate. These can be constructed for libraries of varying insert lengths—from 250 bp to greater than 25,000 bp.

Circular consensus sequencing—the ability to sequence the same

DNA molecule multiple times and thereby generate long high-fidelity (HiFi) reads—is unique to the SMRT system and helps to maximize accuracy by cross-referencing each copy of the molecule against numerous other copies.

Overcoming DNA fragmentation challenges

All DNA is prone to fragmentation, whether it is derived from a biological matrix or created during gene synthesis; thus, any DNA sample will contain a range of fragment sizes. To really exploit the true benefits of long-read sequencing, it is necessary to remove these shorter fragments, which might otherwise be sequenced preferentially.

DNA size selection can exclude short fragments, maximizing data yields by ensuring that those fragments with the most informational content are not blocked from accessing detection centers (for example, ZMWs) by shorter DNA fragments.

Next-generation size-selection solutions

Starting with clean, appropriate-length fragments for HiFi reads can accelerate research by reducing the computation and data processing time needed post-sequencing. Ranger Technology from Yourgene Health is a patent-protected process for automating electrophoresis-based DNA analysis and size selection. Its fluorescence machine vision system and image analysis algorithms provide real-time interpretation of the DNA separation process.

For size-selection protocols, the voltage is modulated per channel to control migration of DNA and allow for the synchronized arrival of the desired fragment sizes at extraction wells. Recovery yields are consistently over 70%, while extraction volumes can be kept to under 50 µL volumes to ensure a high concentration of DNA for downstream processing.

Ranger Technology powers a variety of DNA sample preparation platforms. This includes LightBench, a three-in-one instrument offering automated DNA size selection, fragment length analysis, and fluorometric quantification.

The LightBench adheres to the Standardization in Lab Automation (SiLA 2) specification. As such, it can be operated manually or integrated into a fully automated workflow with third-party liquid handlers, providing a scalable size-selection solution for long-read sequencing.

Enriching long-read sequencing libraries

Next-generation size selection was assessed in an experimental benchmarking of 84 human genomic DNA samples from fresh blood for HiFi long-read sequencing library preparation. The Ranger Technology software was set up according to the manufacturer's recommendations for LightBench. The samples were combined with Dual Dye Loading Buffer containing a 7 kb marker (CG-14000-31-31), mixed thoroughly and centrifuged to remove any air bubbles. The cassette was prepared, and the excess buffer was removed from the reservoir before loading the samples into the wells. Following electrophoresis, samples were extracted using the In-Channel Filter array. The size-selected fractions were then bead-cleaned, and the concentration and fragment-length distributions were assessed prior to sequencing.

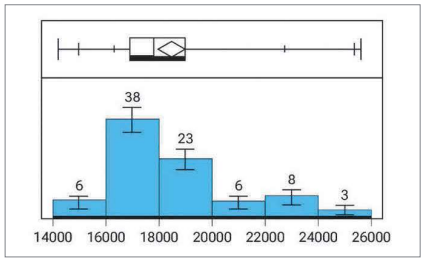

Figure 2. Mean read length distribution across 84 samples on the Revio system.

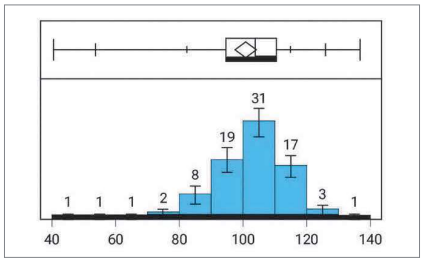

Figure 3. HiFi yield (Gb) distribution across 84 libraries on the Revio system.

Using the LightBench enabled researchers to achieve consistent yields for 8 kb and 10 kb cutoffs across all samples with zero failures. The 8 kb cutoff yielded an average recovery of about 83% of the input SMRTbell library, while the 10 kb cutoff recovered an average of about 67% (*Figure 1*).

After library preparation and size selection with the LightBench, 84 samples were sequenced on the Revio platform (PacBio). The mean HiFi read length was 18,517.5 bp (*Figure 2*—max: 25,613 bp; min: 14,202 bp), and the mean HiFi data yield was 101.27 Gb/SMRT cell (*Figure 3*).

These results show that by excluding smaller fragments, long-read inputs are optimized to deliver increased mean HiFi read lengths and data yields of more than 100 Gb, demonstrating that the LightBench is an effective size-selection method on the Revio system.

Revolutionizing long-read sequencing workflows

Whole genome sequencing use cases are characterized by a need for high data yields to achieve appropriate depth of coverage.

Next-generation size selection enabled increased recovery of long fragments (>10 kb) and simultaneously eliminated smaller fragments that would be preferentially sequenced. Automated, flexible, and high-throughput options are also essential to accommodate demand and ensure enough material is recovered for sequencing. Precise and accurate sizing also helps to reduce sample-to-sample and run-to-run variability.

Utilization of next-generation size-selection instruments that deliver dynamic target enrichment of DNA offer precision, scale, and automation to enable faster, lower-cost workflows that are crucial for maximizing data yields from long-read sequencing. ■

Joanne Mason, PhD, is the chief scientific officer at Yourgene Health (part of the Novacyt group of companies). Website: www.yourgenehealth.com.