revvity

WHITE PAPER

Improving long-read next generation sequencing (NGS): What you should know about sample and library preparation.

For research use only. Not for use in diagnostic procedures.

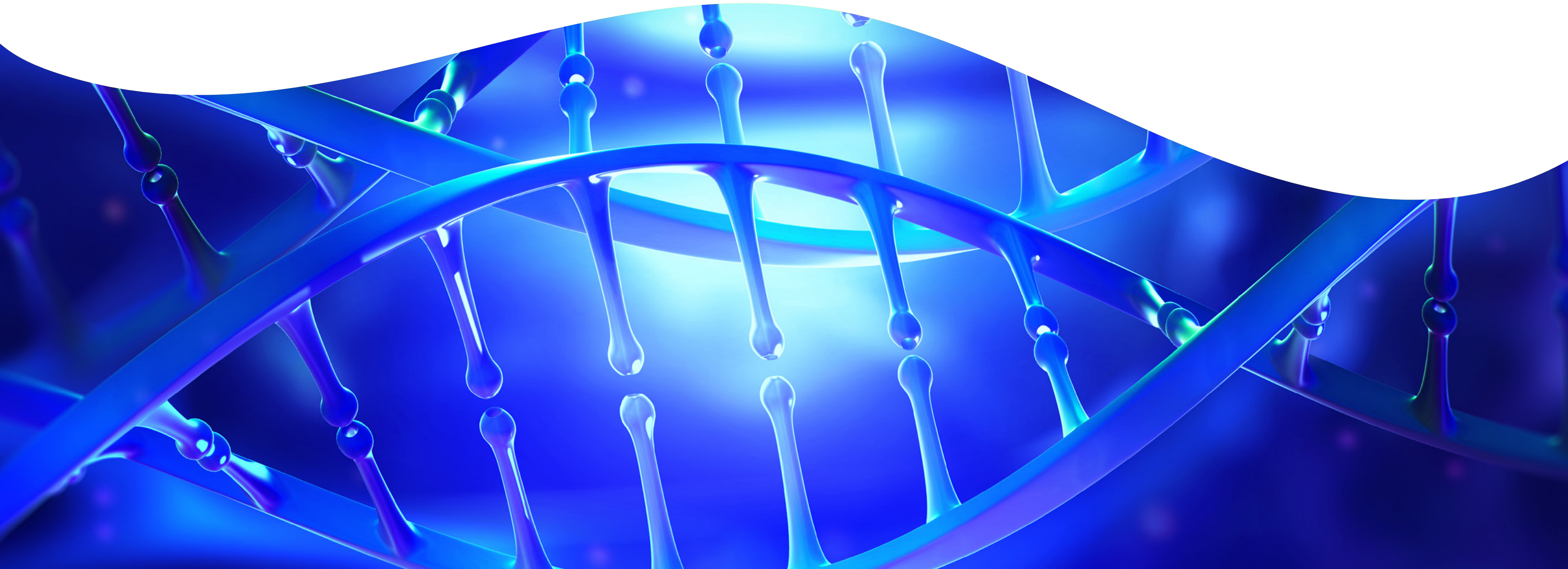
Introduction

During recent years, long-read, single-molecule DNA sequencing has ascended to become a foundational technology in genomic research. With the ability to provide a more comprehensive view of the genome, the technology has been applied to resolve some of the most challenging areas of human genetics. It also has been effective identifying complex structural variants and analyzing among the first telomere-to-telomere assemblies of whole chromosomes.[1]

While long-read sequencing is becoming common in genetic research, it introduces new requirements for sample and DNA library preparation. This scientific brief describes the workflows that ensure long-read sequencing projects will generate accurate results. It explores the major problems that can arise when samples are improperly processed. In addition, the results of two long-read sequencing applications are reviewed that demonstrate best-practice in library preparation.

What is long-read sequencing and why is it important?

Long-read sequencing is a technique that enables the sequencing of much longer DNA fragments than traditional short-read sequencing methods. While short reads can capture the majority of genetic variation, long-read sequencing allows the detection of complex structural variants that may be difficult to detect with short reads.



Long-read sequencing immediately addresses one of the main challenges faced by short-read sequencing. It can sequence an entire single molecule — eliminating amplification bias and generating a reasonable length to overlap a sequence for better sequence assembly.[2]

By overcoming the hurdles inherent to the length and complexity of many genomes, such as missing parts or errors, long-read sequencing produces complete, accurate sequences.

The fragment size for long reads is typically 10,000 base pairs, much longer than the typical 400 base pairs found in short reads. In addition, long-read sequencing can determine the nucleotide sequence of much longer DNA sequences, typically between 10,000 to 100,000 base pairs. This eliminates the need for DNA cleavage and amplification methodologies in short-read sequencing techniques.

Common long-read sequencing technologies

The two most widely used commercial long-read technologies are provided by Pacific Biosciences and Oxford Nanopore Technologies. PacBio's® Single Molecule Real-Time (SMRT™) sequencing has an average read length of about 20 kb with >99.9% accuracy for HiFi™ reads. PacBio's HiFi sequencing method is a new type of long-read sequencing technology with accuracy on par with short reads and Sanger sequencing.[3]

Oxford Nanopore Technologies' nanopore sequencing provides an average read length of about 100 kb for ultra-long reads with about 99% accuracy for R10.4.[4]

The advantages of long-read DNA sequencing

Long-read technology can help resolve challenging regions of the genome by sequencing thousands of bases to:

- Resolve traditionally difficult to map genes or regions of the genome, such as those containing highly variable or repetitive elements
- Perform phased sequencing to identify co-inherited alleles, haplotype information, and phase de novo mutations
- Generate long reads for de novo assembly and genome finishing applications

The technology allows the laboratory to achieve improved accuracy for repeated sequences and copy number variations. It permits more accurate detection of a large number of mutations, plus optimizes DNA extraction protocols. The technology is rapid, affordable, and in some cases, portable.[5]

Most popular applications

The longer read lengths are responsible for enabling the vast number of discoveries in reading genomes, transcriptomes and epigenomes in humans and other species. Long-read data has revealed many previously dark regions of the genome, such as telomeres, which we now know are important in our ability to understand cancer and aging.[5] It has enabled ongoing, active method development for long-read data analysis tasks, ranging from identifying different bases and chemical modifications in DNA and RNA to genome variation detection.

The technology offers solutions that solve some of the present challenges in metagenomic assembly by providing greater sequence overlap, simplification of the assembly process, and assembly quality improvement (e.g., N50, the weighted mid-point of the read length distribution of a sequencing run). Long-read sequencing also permits direct assembly of complete or circular bacterial chromosomes from metagenomes.[6]

Long-read sample and library preparation

Long-read DNA sequencing's sample and library preparation workflows encompass significant differences than their short-read counterparts. Correctly performing these work steps are essential for the generation of meaningful data. The following are some important differences:

- Many times, long-read applications do not require amplification — eliminating the inconsistencies that can result from this work step.
- Fragmentation allows the user to establish the DNA size at the beginning of the process. For long reads, different enzymes may be used up front along with reagents, and fragmentation temperature timing may vary.
- The type of sample being processed also is a key difference. Depending on quality control (QC) sizing, long-read monitoring spikes occur at 25,000 base pairs or more on a library versus a 500 base-pair curve for short reads.

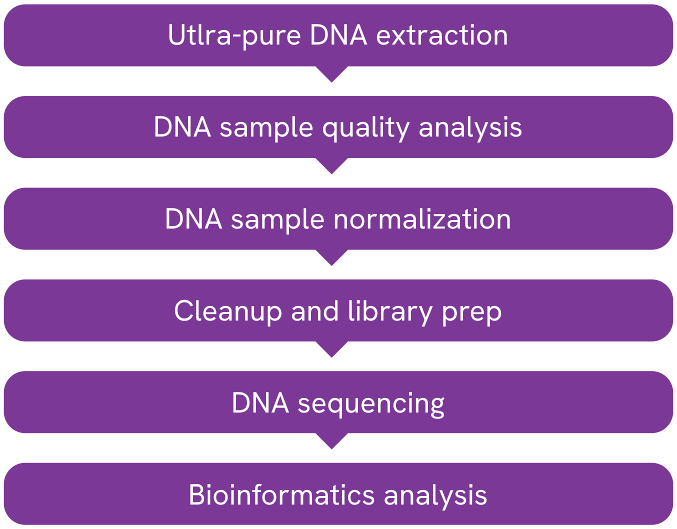


Figure 1: The steps in the long-read sequencing sample and library preparation workflow.

Four things you should know to ensure high-quality data

There are four key factors that will ensure high-quality data is generated from long-read sequencing.

1. Extract ultra-pure, high-molecular weight (HMW) DNA

A well-planned DNA extraction approach for HMW DNA isolation is critical to achieving the highest-quality data. DNA purity is the single greatest factor affecting the success of a sequencing experiment. Degraded samples are more likely to have significant DNA damage that can reduce read length and yield during sequencing, plus have overall lower insert sizes. It is essential to use an extraction process that yields ultra-pure DNA.

For example, Revvity's chemagen™ technology provides an effective solution for challenges associated with the isolation of HMW gDNA from dried blood spot (DBS) samples. It eliminates the traditional alkaline elution, lysis, high-temperature, shaking, and centrifugation steps that can result in nucleic acid denaturation and excessive fragmentation. The chemagen automated magnetic separation procedure uses magnetic particles based on polyvinyl alcohol (M-PVA Magnetic Beads) to isolate and purify nucleic acid from DBS. The beads' high affinity for nucleic acids and low protein binding results in very pure and long DNA/RNA elution.[7]

2. Avoid sample contamination

When selecting the sample type to process, a cell-dense tissue with minimal potential contaminants should be preferred, e.g., using tissues such blood, brain, kidney, or muscle for vertebrate applications.[8] A contaminated sample can not only affect enzymatic reactions in library preparation, but also the efficiency of pore occupation in sequencing — thus decreasing data yield. Contamination also can occur when washes are not stringent enough or carryover from salts is present in the sample.

To maximize read length and quality, it is essential that the DNA sample:[9]

- Is double-stranded
- Has not undergone multiple freeze-thaw cycles
- Has not been exposed to high temperatures (e.g., > 65° C for 15 min) or pH extremes (<6 or >9)
- Does not contain insoluble material or RNA contamination
- Has not been exposed to intercalating fluorescent dyes or ultraviolet radiation.
- Does not contain denaturants (e.g., guanidinium salts or phenol), detergents (e.g., SDS or Triton X100) or chelating agents
- Does not have any low molecular weight fragments

3. Conduct careful, gentle library preparation

For long-read applications, particular attention must be given to how the sample is mixed during library prep to avoid shearing the DNA. It is recommended to keep the mixing slow and not as aggressive on certain steps, such as mixing beads with samples.

Another major challenge in long-read sequencing is the avoidance of DNA shearing of easy-to-lyse organisms while still achieving lysis of difficult-to-lyse organisms.[10] Shearing forces must not be present that will fragment the DNA.

4. Use sequence-qualified library preparation methods: the fast track from bench to data

Some automation vendors and kit suppliers collaborate to produce, test, and QC DNA/RNA libraries, and perform analysis to ensure the library quality and metrics comply with their highest standards. Implementing automation with a sequence-qualified method eliminates much of the on-site application development, field service, and support work that is required by a custom or unqualified automated method. Validated for accuracy and performance, sequence-qualified automated library methods considerably reduce the time from system install to actual sequencing — from months to as few as five days from installation.

Revvity presently has three sequence-qualified library methods available for long-read applications with more coming in the future:

- PacBio HiFiViral™ SARS-CoV-2 Automated on Sciclone® G3 NGSx Workstation
- PacBio SMRTbell Express Template Prep Kit 2.0 on Sciclone G3 NGSx Workstation
- PacBio SMRTbell® Prep Kit 3.0 on Sciclone G3 NGSx Workstation
- Oxford Nanopore Technologies Midnight Protocol [in process]
- Oxford Nanopore Technologies Ligation Sequencing Kit V14 (in process)

[Sidebar]

PacBio SMRTbell Prep Kit 3.0 on Sciclone G3 NGSx Workstation

PacBio's SMRTbell Prep Kit 3.0 automated on Revvity's Sciclone G3 NGSx workstation offers a high throughput workflow for long-read library preparation. This solution reduces hands-on-time, human error, and variability to generate SMRTbell HiFi libraries for sequencing on the Revio™ or Sequel IIe systems.[11]

In testing, the HiFi Library prep application on the Sciclone workstation consisted of five steps: end repair/A tailing, adapter ligation, 1X bead cleanup, nuclease treatment, and 1X bead cleanup. The total processing time for the 48-sample library prep, including incubations, was approximately 4.5 hours.

The full workflow using the 1X SMRTbell cleanup option produced libraries with an average yield of 54.91 ng/µL (figure 7) with an average of 38.93% recovery, which is within the expected range of the kit. The 3-plex pool on Revio yielded a total of 98 Gb of HiFi bases with mean read lengths >19 kb for each of the 3 demultiplexed samples (Table 1 and Figure 2). Whereas the 44-plex pool on Sequel II yielded a total 35 Gb, with good coverage balance across all 44 samples (Figure 1).

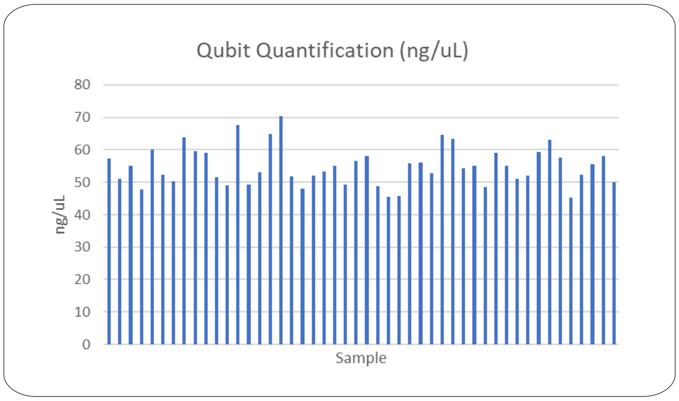


Figure 1. SMRTbell library yields (ng/µL)

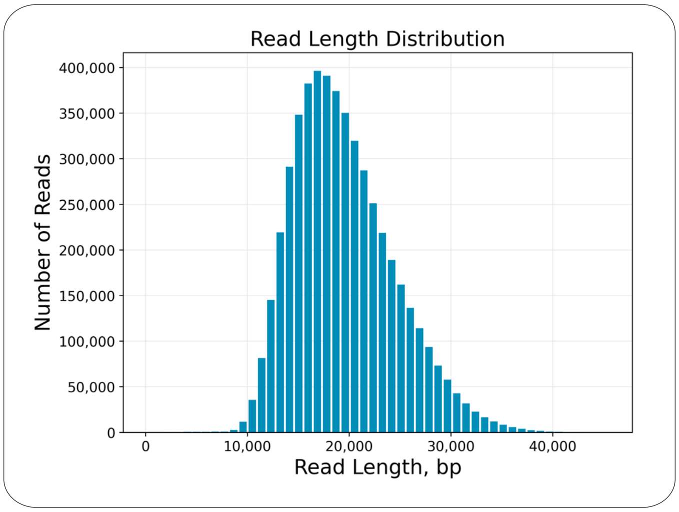


Figure 2. HiFi read length distribution of three sample pool sizes selected on the Sage Science Pippin HT and sequenced on a single Revio SMRT Cell. Average HiFi read length was 19,468 bp.

Table 1. Revio sequencing metrics for the three sequenced barcoded samples prepared on the Sciclone G3 NGSx system. The run achieved a total of 98 Gb of barcoded HiFi data, and > 10 fold coverage across the human genome for each sample.

Sample	Barcode	HiFi Reads	HiFi Read Length mean	HiFi Read Quality mean	HiFi Yield (Gb)	Fold coverage per sample
HG001 C5	bc2035	1,684,696	19,488 bp	Q27	32.83	10.5
HG001 E5	bc2037	1,665,826	19,517 bp	Q27	32.51	10.4
HG001 A7	bc2049	1,690,932	19,407 bp	Q27	32.82	10.5
NA	No barcode	43,647	19,166 bp	Q24	00.84	na
Total barcoded yield and coverage					98.16 Gb	31.4

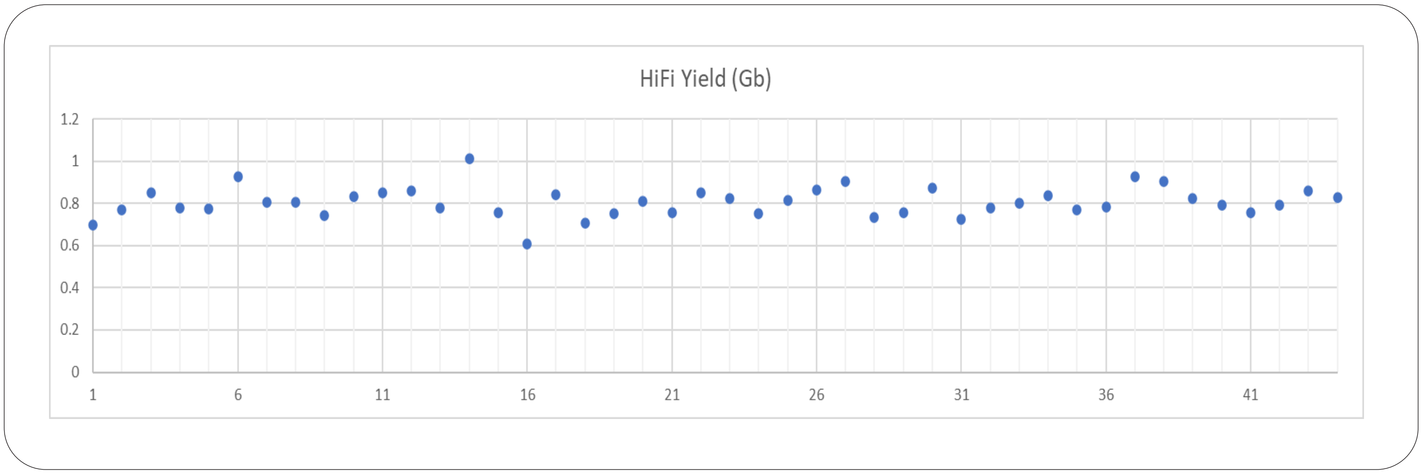


Figure 3. HiFi yield per barcoded library prepared on the Sciclone G3 NGSx system.

[Sidebar]

PacBio HiFiViral™ SARS-CoV-2 automated on Sciclone G3 NGSx Workstation

The PacBio HiFiViral™ SARS-CoV-2 kit was developed as a scalable solution with increased resilience against virus mutations, designed for use on the PacBio Sequel lle system. The kit was automated on Revvity's Sciclone G3 NGSx workstation to enable a high-throughput workflow from cDNA synthesis through library construction. PacBio successfully processed different control inputs on the Sciclone G3 NGSx workstation. The resulting libraries aligned with QC and sequencing metrics expected for the kit.[12]

The full workflow consisted of two applications on the Sciclone workstation with user touch points for off-deck incubation and reagent plating. The first application took the samples through cDNA amplification and the second application processed the pooled amplified cDNA through library construction.

Results: the full workflow produced PacBio SMRTbell® libraries with yields from 8.76 ng/µL to 79.1 ng/µL (figure 1), averaging 55.6 ng/ µL. All samples were within the target peak of > 700 bp.

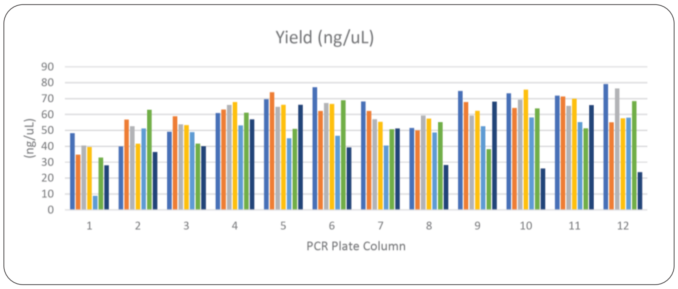
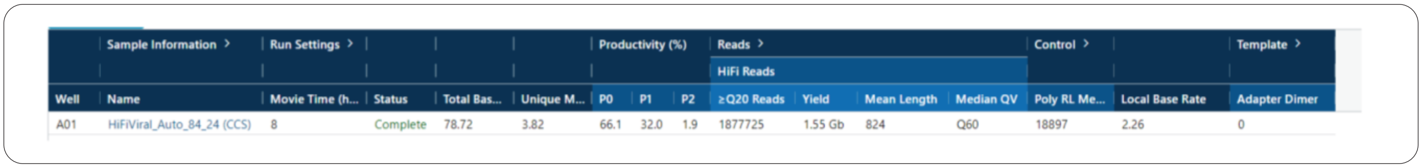


Figure 1: PacBio SMRTbell library yields

Figure 2 shows the primary analysis results for two PacBio HiFiViral libraries on the PacBio Sequel lle sequencer. With %P1 32%, the libraries generated yield of 1.55 Gb with a mean subread length of 18.9 kb.



	Sample Information >	Run Settings >				Productivity (%)			Reads >				Control >		Template >
									HiFi Reads						
Well	Name	Movie Time (h...	Status	Total Bas...	Unique M...	P0	P1	P2	≥Q20 Reads	Yield	Mean Length	Median QV	Poly RL Me...	Local Base Rate	Adapter Dimer
A01	HiFiViral_Auto_84_24 (CCS)	8	Complete	78.72	3.82	66.1	32.0	1.9	1877725	1.55 Gb	824	Q60	18897	2.26	0

Figure 2: The results for control samples on a Sequel lle system prepared by Sciclone G3 NGSx workstation.

The PacBio HiFiViral SARS-CoV-2 kit provided a robust, simple-to-use, scalable, cost-effective solution for sequencing the entire SARS-CoV-2 genome. Automating the PacBio HiFiViral kit on the Sciclone G3 NGSx workstation not only reduced hands-on time and sample variability but also reduced the overall project cost. The Sciclone G3 system is intuitive and simple to use with its interface-guided workflow set-up and step tracking.

Conclusion

Long-read sequencing introduces new requirements for sample and DNA library preparation, encompassing significant differences than their short-read counterparts. These include the elimination of the amplification step, plus a fragmentation step that allows the researcher to establish the DNA size at process inception. Depending on QC sizing, long-read monitoring spikes occur at 25,000 base pairs or more on a library versus a 500 base-pair curve for short reads.

When conducting long-read sample and library preparation, there are four key issues that must be considered:

1. Extract ultra-pure, high-molecular weight (HMW) DNA using platforms based on gentle rotational resuspension technology
2. Avoid practices that can introduce sample contamination
3. Conduct careful, gentle library preparation by minimizing shearing forces during library preparation with gentle pipetting
4. Use sequence-qualified library preparation methods that will dramatically shorten bench-to-data times and complexity

By embracing the correct methodology for processing long-read samples, researchers can expect more accurate sequencing data and higher throughput.

References

[1] A complete reference genome improves analysis of human genetic variation, Science, April 2022.

[2, 5] CD Genomics Website, cd-genomics.com.

[3] Method of the Year 2022: long-read sequencing, Nature Methods, January 12, 2023.

[4] Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing, Elsevier Computational and Structural Biotechnology Journal, Volume 21, 2023.

[6, 10] Optimizing Sample Preparation for Metagenomic Assembly using Long-Read Sequencing, Zymo Research.

[7] Comparison of Automated Nucleic Acid Purification Systems on High Molecular Weight (HMW) DNA Extraction Efficiency, Revvity Application Note, 2023.

[8] PacBio Website, pacb.com.

[9] Genome Sample Requirements, Long-Read Sample Services, University of New South Wales, Ramaciotti Centre for Genomics.

[11] PacBio SMRTbell® Prep Kit 3.0 on Sciclone G3 NGSx Workstation Application Note, 2023.

[12] PacBio HiFiViral™ SARS-CoV-2 Automated on Sciclone G3 NGSx Workstation Application Note, 2022.

Revvity, Inc.
940 Winter Street
Waltham, MA 02451 USA

(800) 762-4000
www.revvity.com

For a complete listing of our global offices, visit www.revvity.com