Application Note - Considerations for Using the Low and Ultra-Low DNA Input Workflows for Whole Genome Sequencing

Introduction

As the foundation for scientific discoveries in genetic diversity, sequencing data must be accurate and complete. With highly accurate long-read sequencing, or HiFi sequencing, there is no longer a compromise between read length and accuracy. HiFi sequencing enables some of the highest quality *de novo* genome assemblies available today as well as comprehensive variant detection in human samples.

PacBio® HiFi libraries constructed using our standard library workflows require at least 3 µg of DNA input per 1 Gb of genome length, or ~10 µg for a human sample. For some samples it is not possible to extract this amount of DNA for sequencing. For samples where between 300 ng and 3 ug of DNA is available, the Low DNA Input Workflow enables users to generate high-quality genome assemblies of small-bodied organisms. For samples where even less DNA is available (as low as 5 ng), the amplification-based Ultra-Low DNA Input Workflow is available.

With three different workflows for HiFi sequencing (Table 1), there is a solution for sequencing genomes of all types of organisms.

Choosing a Workflow

We recommend considering the genome assembly project as a whole, from DNA extraction to bioinformatics, to establish your experimental design.

Where possible, the standard HiFi workflow run on the Sequel® II System gives you the highest quality results for both genome assembly and human variant detection projects. However, if you are sample-limited, the Low and Ultra-Low DNA Input Workflows will still provide excellent results.

	Standard HiFi Sequencing	Low DNA Input Sequencing 2-Plex	Low DNA Input Sequencing Single Sample	Ultra-Low DNA Input Sequencing
Minimum DNA Input	>3 µg / 1 Gb genome	300 ng for each genome	400 ng	5 ng
Amplification Based?	No	No	No	Yes
Genome Size Limit	N/A	600 Mb for each genome	1 Gb	500 Mb
Supported Applications	*de novo* Assembly Human Variant Detection	*de novo* Assembly	*de novo* Assembly	*de novo* Assembly Human Variant Detection

Table 1. Details of standard, low DNA input, and ultra-low DNA input HiFi sequencing workflows on the Sequel II System.

Whole Genome Sequencing for *de novo* Assembly

For *de novo* genome assembly projects, consider the size of the genome to be sequenced as well as the amount of DNA available when choosing a workflow. The minimum DNA amount for the Low DNA Input Workflow is 300 ng for a 2-plex project where each genome can be up to ~600 Mb in size. If you have multiple genomes of interest that fit within these DNA and genome size requirements, this is an efficient and cost-effective option. If the genome is slightly larger, up to 1 Gb in size, and you are able to extract ≥400 ng of DNA from the organism, the single-sample Low DNA Input Workflow is the appropriate workflow. Both single-sample and 2-plex workflows can be found in the Low DNA Input Protocol.

Small arthropod samples that are unable to produce the required 300 ng - 400 ng of input DNA and have a genome size up to ~500 Mb can be sequenced with the amplification-based Ultra-Low DNA Input Workflow. This workflow utilizes a simple, all-in-one SMRTbell® gDNA Sample Amplification Kit for enrichment prior to SMRTbell library preparation.

Variant Detection Using Human Samples

For human variant detection projects, the main consideration is the input amount of DNA available. If your sample does not meet the requirements for the standard HiFi sequencing workflow, the Ultra-Low DNA Input Workflow is the next best option. With 5 ng - 20 ng of DNA, a whole human genome can be sequenced to ~15-fold HIFi coverage with two SMRT® Cells 8M and used for variant detection.

Importance of DNA Quality

In all cases, you must start with high molecular weight genomic DNA (gDNA): ≥30 kb for Low DNA Input and ≥20 kb for the Ultra-Low DNA Input Workflow. With high molecular weight gDNA, the gDNA sample can be sheared reliably to meet the required workflow-specific fragment distribution. Furthermore, lower quality gDNA may contain damages that impact sequencing results. Samples with DNA damage should be re-extracted to help ensure optimal sequence data yield and quality. The best gDNA extraction method will depend on your sample type and additional guidance can be found in this Technical Note.

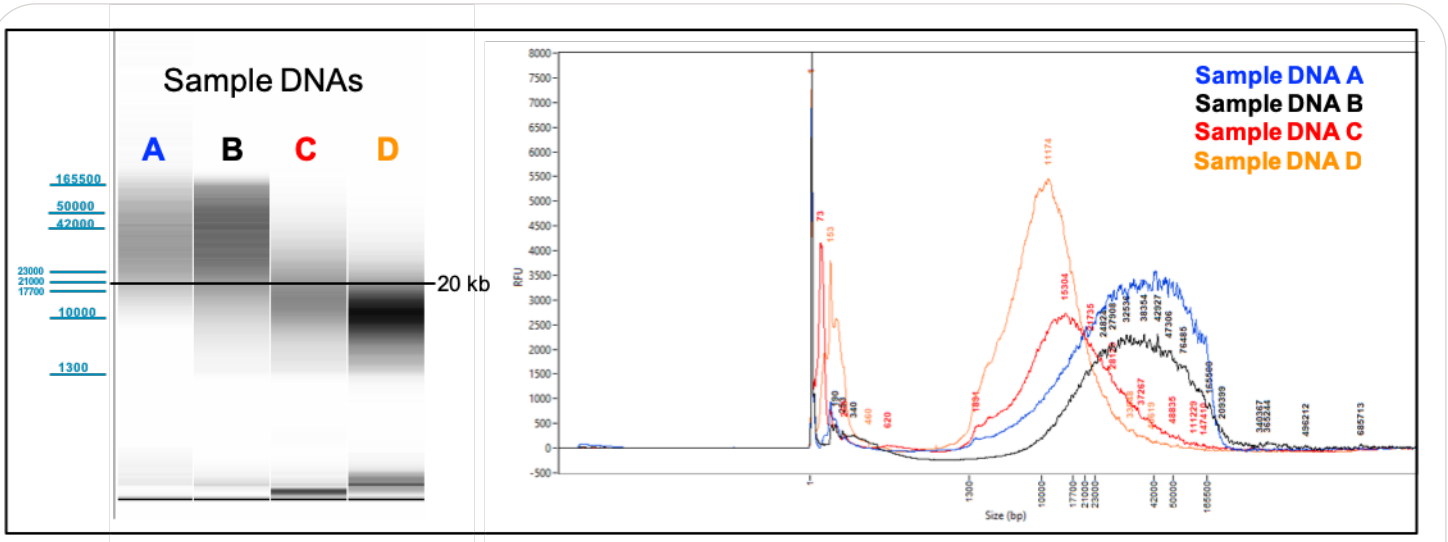

Figure 1. Femto Pulse gel images and traces of four gDNA samples. Sample A contains fragments with a majority of gDNA >20 kb, with minimal fragments <20 kb, and is considered suitable for use with Low and Ultra Low DNA Input Protocols. Sample B has a majority of fragments >20 kb but also has a smear of DNA <10 kb, putting some risk on continuing into library preparation. Samples C and D are too fragmented with a majority of gDNA <20 kb and would not be recommended for SMRTbell library construction or sequencing.

Detailed Considerations for the Low DNA Input Workflow

Sample and Library Preparation

DNA quality will be the primary driver in determining the success of your low DNA input genome assembly project because there is no stringent size-selection step to remove library fragments <10 kb. For best results, the starting gDNA should be >30 kb. To determine the gDNA size distribution, we recommend using the Femto Pulse System from Agilent to enable DNA size analysis from only 500 pg of input material. If you have sufficient gDNA (100 ng - 150 ng for sizing QC plus sufficient DNA as listed in Table 1 for library preparation) you may be able to evaluate the size distribution via other methods (CHEF Mapper System from BioRad and Pippin Pulse System from Sage Science).

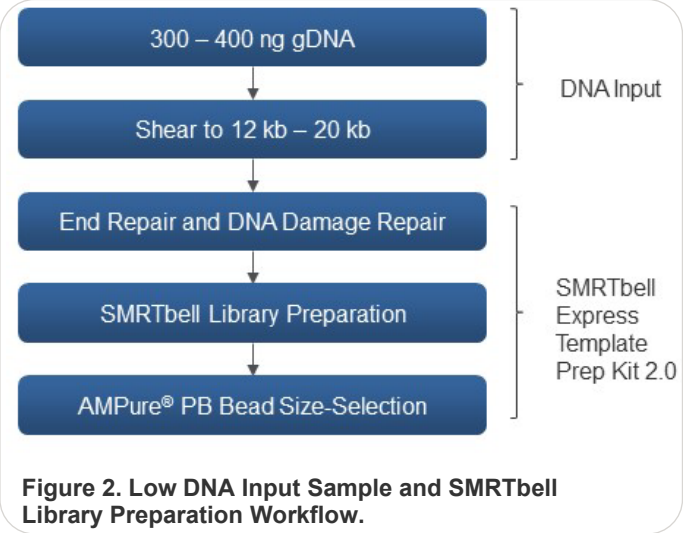

Figure 2. Low DNA Input Sample and SMRTbell Library Preparation Workflow.

To sequence a single low DNA input sample, ≥400 ng of input gDNA is required for the Sequel II System and the target DNA shear size distribution is 12 kb - 20 kb using the Megaruptor system. The library preparation for single low DNA input samples is similar to the standard HiFi library preparation workflow with the addition of an AMPure® PB Bead size-selection step at the end (instead of the BluePippin or SageELF Systems for size selection). The SMRTbell library is purified and size selected using AMPure PB Beads to remove <3 kb SMRTbell templates. Nuclease treatment of the SMRTbell library is not required.

To sequence two low DNA input samples in a single SMRT Cell 8M, ≥300 ng of input gDNA is required for each sample. The target gDNA shear size distribution is 12 kb - 20 kb using the Megaruptor system. Each sample will independently go through library preparation up to ligation with barcoded overhang adapters (Barcoded Overhang Adapter Kit 8A or 8B). Following ligation, the two barcoded SMRTbell libraries are treated with nucleases to remove damaged or partial SMRTbell templates prior to pooling. Finally, the two-plex SMRTbell library is purified and size selected using AMPure PB Beads to remove <3 kb SMRTbell templates.

Data Analysis

After sequencing the final low DNA input library to the desired HiFi coverage depth (10-15-fold per haplotype is recommended), the HiFi reads can be treated like any other HiFi dataset for downstream assembly. For *de novo* genome assembly there are many assembly tools available for HiFi reads, such as IPA, the newest assembler from PacBio, as well as HiCanu, hifiasm, and others. Each tool has their own unique value, but ultimately the choice is yours, and in our experience they all work well.

Low DNA Input Workflow Case Study

Example: Multiplexed Mosquito Genome Assembly

Starting from 230 ng each from two different *Anopheles gambiae* DNA preps (kindly donated by Mara Lawniczak), individual gDNA samples were sheared to 13 kb using the Megaruptor 3 System. For each sample, single-strand overhangs are removed before going through DNA Damage Repair and End-Repair/A-tailing. Barcoded overhang adapters are ligated to each sample separately. Following ligation, the two SMRTbell libraries were treated with nucleases prior to pooling. The library was sequenced on a single SMRT Cell 8M, generating 24-fold HiFi read coverage per sample, and the resulting HiFi data were assembled with hifiasm. HiFi reads were deposited in NCBI under BioProject PRJNA643270.

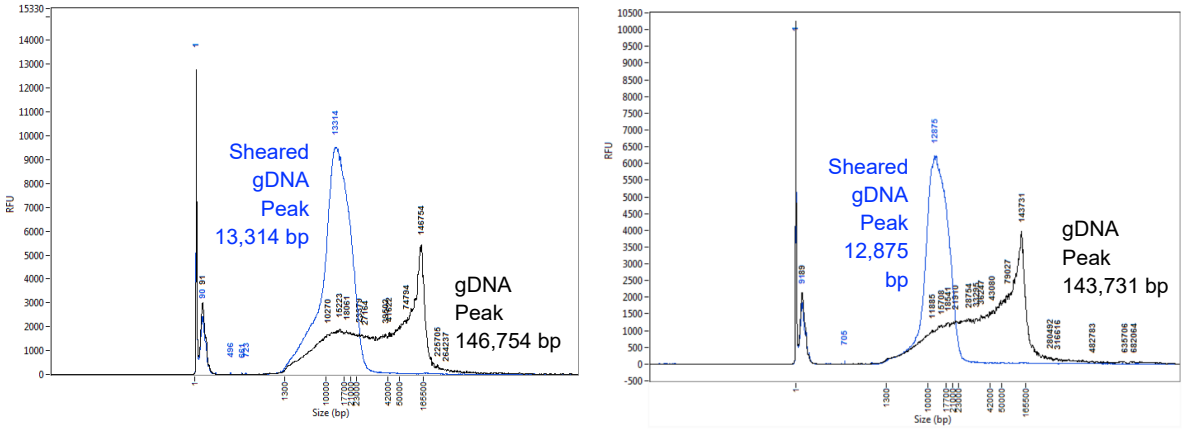

Figure 3. Femto Pulse trace of Sample 1 (left) and Sample 2 (right) Mosquito genomic DNA before (black) and after (blue) shearing.

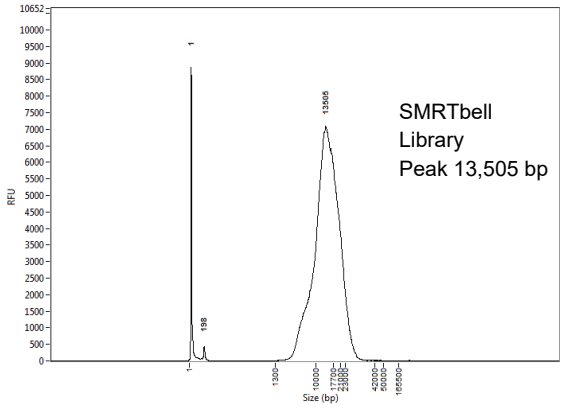

Figure 4. Femto Pulse trace of two-plex Mosquito library after AMPure PB bead size selection.

Sequencing Results for Mosquito two-plex		
	Sample 1	**Sample 2**
HiFi Data Yield	6.85 Gb	6.55 Gb
Mean HiFi Read Length	9,184 bp	7,720 bp
Median HiFi Read Quality	Q38	Q40

Table 2. Sequencing results for two-plex of mosquito low DNA input samples.

Assembly Results for Mosquito two-plex		
	Sample 1	**Sample 2**
Assembly Size	278 Mb	278 Mb
Contig N50	15.7 Mb	8.6 Mb
BUSCO Complete	98.5 %	98.8 %

Table 3. Assembly results for two-plex of mosquito low DNA input samples.

Detailed Considerations for the Ultra-Low DNA Input Workflow

PCR Amplification

Unlike standard PacBio libraries, SMRTbell libraries produced with the Ultra-Low DNA Input Protocol utilize amplification to increase the amount of available DNA template material prior to sequencing.

Because it uses PCR, the Ultra-Low DNA Input Protocol is subject to the limitations of PCR, (ie.low processivity in high-GC regions). Therefore, high-GC regions may be under-represented in a dataset generated with the Ultra-Low DNA Input Workflow.

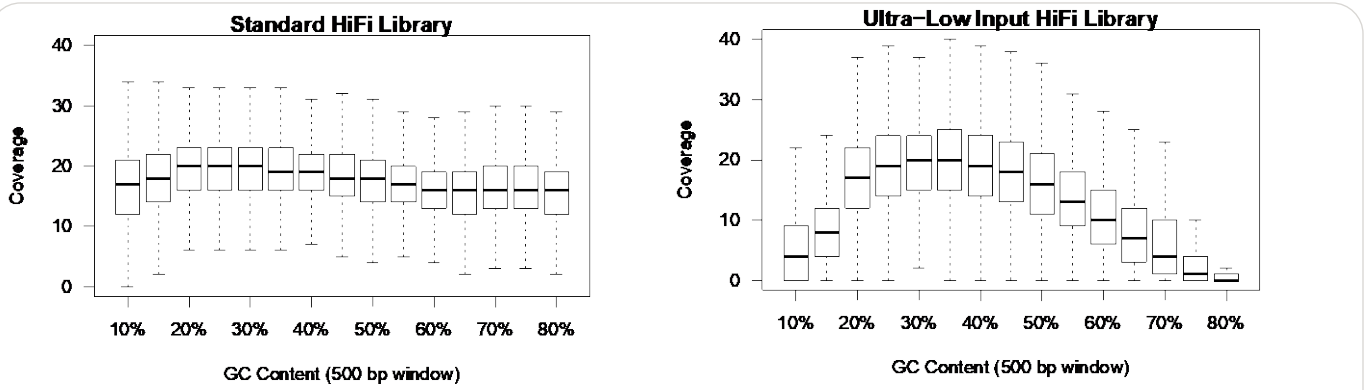

Figure 5 – Coverage vs GC content plots show even coverage (no GC bias) from a standard PacBio library prep, compared to the less even coverage across GC range with the PCR-based ultra-low DNA input library prep.

The impacts of the use of PCR amplification include:

- If the genome you are trying to sequence contains high-GC regions, a genome assembly may not be as complete or contiguous.
- Variant recall performance may be lower in a human genome compared to an unamplified dataset.
- There will be PCR duplicate reads in an ultra-low DNA input dataset. These are removed with the "Mark PCR Duplicates" application during analysis and there may be a reduction in the overall coverage of unique data.
- PCR introduces errors, mostly homopolymer length changes and dinucleotide repeat compression, which may affect the resulting accuracy of genome assemblies or precision and recall of small variants.

Therefore, when designing your genome assembly or variant detection project with the Ultra-Low DNA Input Workflow, align your expectations with the limitations of PCR in mind.

Sample and Library Preparation

DNA quality will be the primary driver in determining the success of your ultra-low DNA input project, whether embarking on a genome assembly or aiming to detect variants across a human genome. For best results, the starting gDNA should be >20 kb. To determine the DNA size distribution, we recommend using the Femto Pulse System from Agilent to enable DNA size analysis from only 500 pg of input material. If you have sufficient DNA (100-150 ng for instrument usage plus sufficient DNA for library preparation) you may be able to evaluate the size distribution via other methods (CHEF Mapper® System from BioRad and Pippin Pulse™ System from Sage Science).

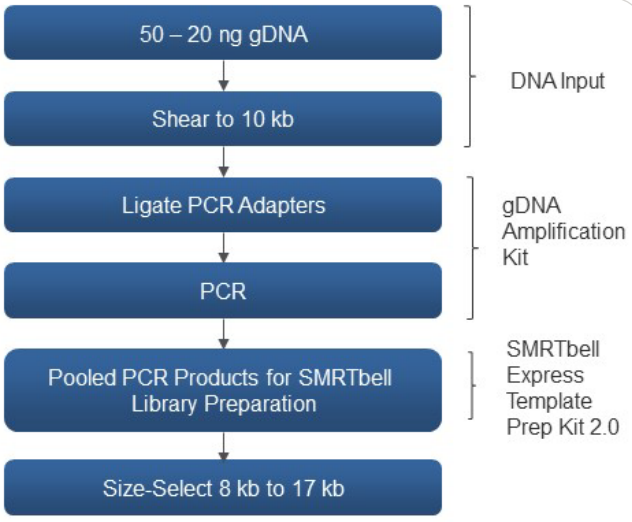

Figure 6. Ultra-low DNA input SMRTbell library preparation workflow

Starting with at least 5 ng of input gDNA, we recommend shearing to approximately 10 kb using either the Megaruptor by Diagenode or the g-TUBE by Covaris.

Under- or over-shearing of DNA will result in poor amplification and may impact the yield of the final BluePippin size-selected library. Therefore, it may take multiple test shears to obtain the optimal DNA size.

The library preparation for ultra-low DNA input samples is similar to the standard HiFi SMRTbell library construction with the addition of an up-front amplification step using the SMRTbell gDNA Sample Amplification Kit.

After the gDNA is sheared to approximately 10 kb, it is amplified using two parallel PCR reactions that have been optimized for varying GC content ranges. The amplification protocol specifies a range of 13-18 cycles of amplification. To reduce the rate of PCR duplication, we recommend that you initially start with 13 cycles and if the yields for each PCR reaction do not reach the required minimum amounts to proceed with SMRTbell library construction, run an additional 2-5 cycles of PCR amplification. For constructing ultra-low DNA input SMRTbell libraries, a minimum of 500 ng of pooled amplified sample (PCR reaction 1 + PCR reaction 2) is required to generate sufficient SMRTbell library material to run 1 Sequel II SMRT Cell 8M. To generate sufficient library material to run 2 Sequel SMRT Cells 8M, we recommend starting with approximately 800 ng of pooled amplified sample for library construction. If you consistently find that you need to add additional PCR amplification cycles to meet the above requirements, you may start with a higher number of cycles for future samples.

To assess whether the amplification worked as expected, we recommend assessing the size distribution of the amplified DNA products using the Femto Pulse System before proceeding to SMRTbell library preparation. After successful amplification, the DNA products from the two parallel PCR reactions are pooled together in equal mass quantities. The pooled amplified DNA can then be constructed into a SMRTbell library as a single sample using SMRTbell Express Template Prep Kit 2.0.

To reduce the amount of short insert SMRTbell templates from the library, size-selection is required. We recommend the use of the BluePippin system from Sage Science to remove SMRTbell templates less than 8 kb in size. Lastly, to verify the quality of the size-selected library, we recommend assessing the final library size distribution by using the Femto Pulse System.

Data Analysis

After sequencing the final ultra-low DNA input library to desired coverage depth (>30-fold is recommended for *de novo* assembly and >15-fold is recommended for human variant detection), the HiFi reads must go through two

preliminary processing steps before they can be used for *de novo* assembly or variant calling: trimming of PCR adapter sequences and removal of PCR-duplicate reads.

In SMRT® Link, two analysis workflows are available to accomplish the aforementioned steps: "Trim gDNA Amplification Adapters" and "Mark PCR Duplicates". Users who prefer command-line tools from pbbioconda can use `lima` for PCR adapter trimming and `pbmarkdups` for PCR duplicate removal. The primer adapter sequence required for command-line trimming is: "`AAGCAGTGGTATCAACGCAGAGTACT`".

Successful datasets typically have >98% of HiFi reads with the adapter sequence present and less than 10% PCR duplicate reads. Samples with higher coverage may have higher observed PCR duplication rates and will consequently have a higher percentage of data removed during processing.

For *de novo* genome assembly projects, start with the trimmed, deduplicated HiFi reads. There are many assembly tools available for HiFi reads, such as IPA, the newest assembler from PacBio, as well as HiCanu, hifiasm, and others. Each tool has their own unique value, but ultimately the choice is yours, and in our experience they all work well.

For structural variant (≥50 bp) detection projects, start with the trimmed, deduplicated HiFi data set and run the "Structural Variant Calling" application in SMRT Analysis. Under the advanced settings make the following parameter changes to optimize performance from this amplified data type:

- Minimum % of Reads that Support Variant (any one sample): 30
- Minimum Reads that Support Variant (any one sample): 2
- Minimum Reads that Support Variant (total over all samples): 2

For small variant (<50 bp) detection, start with the trimmed, deduplicated HiFi dataset and run the "Mapping" application in SMRT® Analysis. The aligned BAM produced by this application is compatible with the DeepVariant v0.10.0 PacBio model for germline small variant calling.

Ultra-Low DNA Input Workflow Case Studies

The Ultra Low DNA Input Workflow supports two whole genome sequencing applications: small arthropod *de novo* genome assembly and human variant calling.

Example 1: *Drosophila* Genome Assembly

A total of 10 ng of gDNA from a single *Drosophila melanogaster* insect was sheared to 10 kb using a g-TUBE device and amplification was performed with the SMRTbell gDNA Sample Amplification Kit using a total 17 cycles for PCR Reaction Mix 1 and 13 cycles for PCR Reaction Mix 2. A SMRTbell library was generated using SMRTbell Express Template Prep Kit 2.0 and sizeselected with the BluePippin system to give an average insert length of 11 kb. The library was sequenced on a single SMRT Cell 8M, producing 26.8 Gb of HiFi data. HiFi reads were adapter-trimmed and PCR deduplicated, resulting in 183-fold processed coverage of the *Drosophila* genome. HiFi reads were down sampled to give 30-fold coverage and then assembled with hifiasm. Adapter-trimmed and PCR deduplicated HiFi reads were deposited in NCBI under BioProject PRJNA657245.

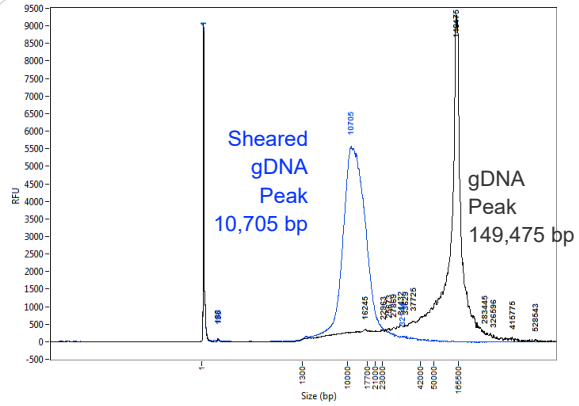

Figure 7 – Femto Pulse trace of *Drosophila melanogaster* DNA before and after shearing.

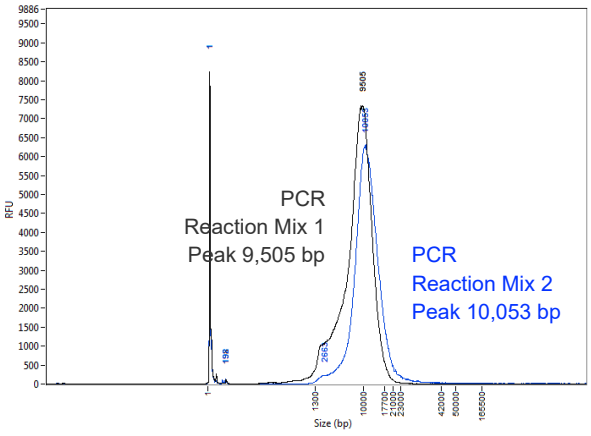

Figure 8 – Femto Pulse trace of *Drosophila melanogaster* amplified DNA.

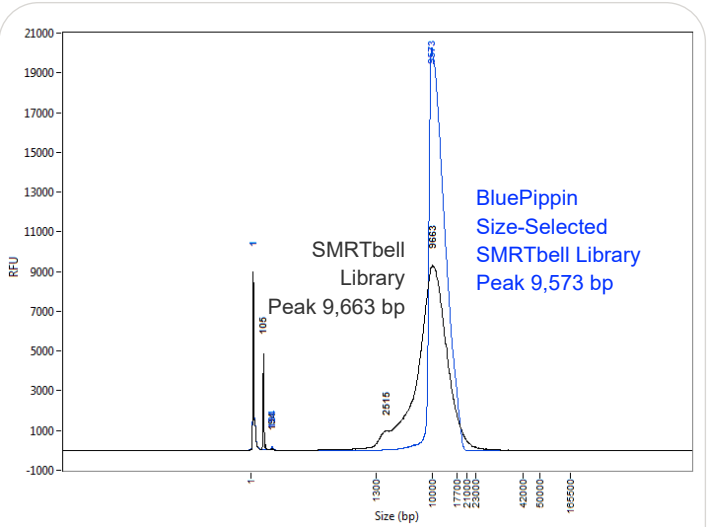

Figure 9 – Femto Pulse trace of *Drosophila melanogaster* library before and after size selection.

Sequencing Results for *Drosophila melanogaster*	
HiFi Data Yield	26.77 Gb
Mean HiFi Read Length	11,153 bp
Median HiFi Read Quality	Q34
PCR Adapter Percentage	99.85 %
PCR Duplication Rate	3.78 %
Processed HiFi Coverage for 140 Mb Genome	183-fold

Table 4. Sequencing and data processing results for *Drosophila melanogaster* ultra-low DNA input library.

Assembly Results for *Drosophila melanogaster*	
Downsampled HiFi Coverage	30-fold
Assembly Size	147 Mb
Contig N50	8.3 Mb
BUSCO Complete	98.6 %

Table 5. Assembly results for *Drosophila melanogaster* ultra-low DNA input library.

Example 2: HG002 Human Variant Detection

Genomic DNA was extracted from cell pellets of the HG002 cell line from Coriell using the Lucigen Masterpure kit and sheared to 10 kb using a g-TUBE device. Amplification with the gDNA amplification kit was performed with 5 ng of sheared gDNA and 18 cycles for PCR Reaction Mix 1 and 17 cycles for PCR Reaction Mix 2. A SMRTbell library was generated using SMRTbell Express Template Prep Kit 2.0 and size selected with the BluePippin system to give an average insert length of 11.4 kb. The library was sequenced on two SMRT Cells 8M and the resulting HiFi data were trimmed and deduplicated, resulting in 10-fold HiFi coverage per SMRT Cell 8M. The reads were aligned to the GRCh37 reference genome with pbmm2. Small variants were called with DeepVariant v0.10.0 and benchmarked with Hap.py against the Genome in a Bottle v3.3.2 small variant benchmark for HG002. Structural variants were called with PBSV 2.2.2 and evaluated against the Genome in a Bottle v0.6 SV benchmark using Truvari. Adapter-trimmed and PCR deduplicated HiFi reads were deposited in NCBI under BioProject PRJNA657245.

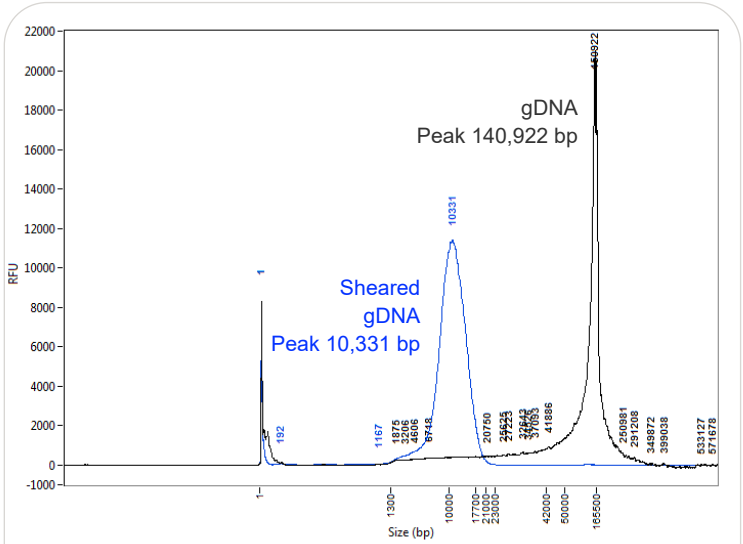

Figure 10 – Femto Pulse trace of HG002 DNA before (black) and after (blue) shearing.

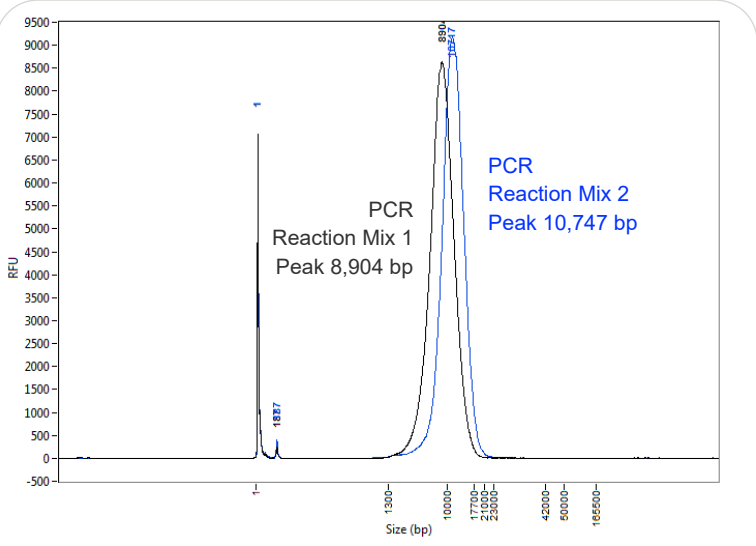

Figure 11 – Femto Pulse trace of HG002 amplified DNA from PCR Reaction Mix 1 (black) and PCR Reaction Mix 2 (blue).

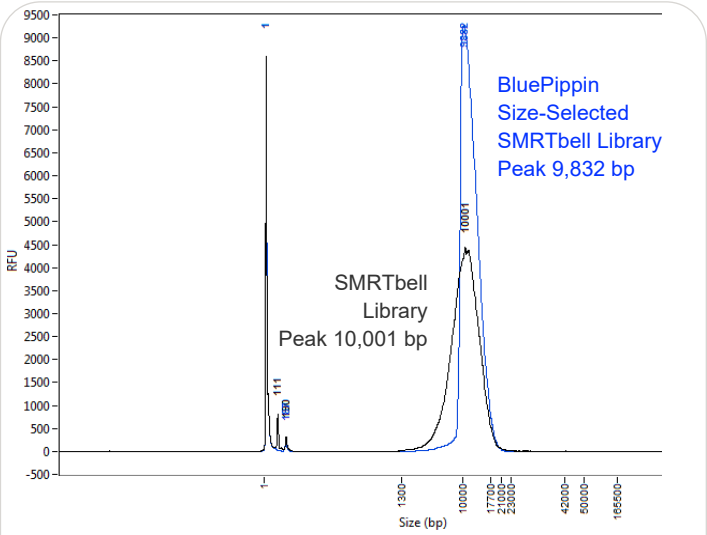

Figure 12 – Femto Pulse trace of HG002 library before (black) and after (blue) size selection.

Sequencing results for HG002 (human)	SMRT Cell A	SMRT Cell B
HiFi Data Yield	32.22 Gb	30.77 Gb
Mean HiFi Read Length	10,909	10,999
Median HiFi Read Quality	Q33	Q37
PCR Adapter Percentage	99.90 %	99.89 %
PCR Duplicate Rate	6.89 %	4.32 %
Processed HiFi Coverage for 3 Gb Genome	10-fold	10-fold

Table 6. Sequencing and data processing results for a human (HG002) ultra-low DNA input library sequenced on two SMRT Cells 8M.

Variant Detection Benchmarking Results for HG002 (human)	Variant Caller	Precision	Recall
Single Nucleotide Variants	DeepVariant	99.6 %	99.0 %
Indels (<50 bp)	DeepVariant	84.7 %	90.3 %
Structural Variants (≥50 bp)	PBSV	95.4 %	84.3 %

Table 7. Variant calling results for a human (HG002) ultra-low DNA input library sequenced on two SMRT Cells 8M.

Conclusions

The value of highly accurate long reads as a starting point for genomic studies is now understood, but hurdles in obtaining this data exist for samples that have low DNA quantities. The Low and Ultra-Low DNA Input Workflows described here can be applied to samples with input DNA amounts as low as 5 ng and put high-quality genome assemblies and human variant detection studies within reach for many sample types that would otherwise be challenging to analyze using other long-read sequencing technologies.

Resources:

Procedure & Checklist – Preparing HiFi Libraries from Low DNA Input Using SMRTbell Express Template Prep Kit 2.0

Procedure & Checklist – Preparing HiFi SMRTbell Libraries from Ultra Low DNA Input

Technical Note: DNA Prep for PacBio HiFi Sequencing – Extraction and Quality Control

PacBio Consumable Part Numbers:

Part Number	Item
100-938-900	SMRTbell Express Template Prep Kit 2.0
101-980-000	SMRTbell gDNA Sample Amplification Kit

PN 101-995-900 Version 01 (September 2020)