NOV2013, Paul Coupland, Liz Sheridan

Wellcome Trust Sanger Institute

pc10@sanger.ac.uk

Needle shearing DNA for PacBio >20 kb libraries.

This protocol outlines a shearing technique we use for preparing long, >20 kb, DNA fragments for PacBio library prep using a small gauge needle. This decades-old technique has shown to provide a SMRTbell library with a template size and, importantly, distribution, that has allowed bacterial de novo assembly up to 5 Mb, from a single SMRTcell (~55K reads) of data, to a contig : {chromosome, plasmid} ratio of 1:1. Only SPRI-bead size-selection was performed in the final step of library prep, using a ratio of 0.375 X (i.e., 100 ul SMRT bell : 37.5 ul PB-SPRI beads). Genome complexity is of course a major driving factor for de novo assembly; in this protocol we use a *Salmonella* strain as the example.

As with all hydrodynamic shearing methods there are some assumptions about the quality (purity and fragment degradation) of genomic DNA (gDNA). If the gDNA is already 'contaminated' with smaller fragments then no shearing protocol will fix this; the final assembly will be adversely affected by the small fragments in the SMRTbell library. In these instances more stringent size-selection will be necessary, the current recommendation being Blue-pippin. However, if your gDNA passes initial fragment size QC, then the technique described here is hopefully a useful alternative to the G-tube + Blue-pippin protocol and will provide comparable data and assemblies with a cost and time saving.

Initial QC.

gDNA is assessed with the Agilent Tape Station (Figures. 1,2) if the peak is assessed to be >60 kb with no observable small fragment contamination then we will attempt to sequence and assemble from one SMRTcell of data, e.g, ~50K reads, sub-read N50 > 9.5 kb, most data in HQ region (i.e., the SMRTcell was not overloaded).

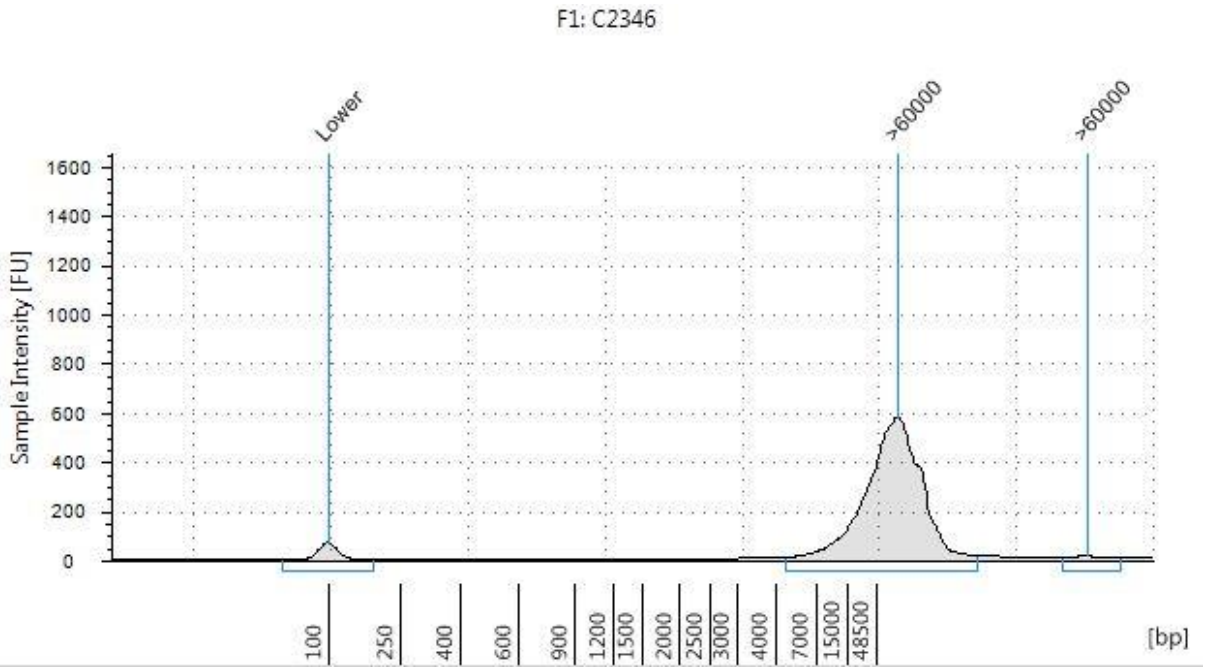


Figure 1: gDNA that PASSES fragment degradation QC and will be needle sheared.

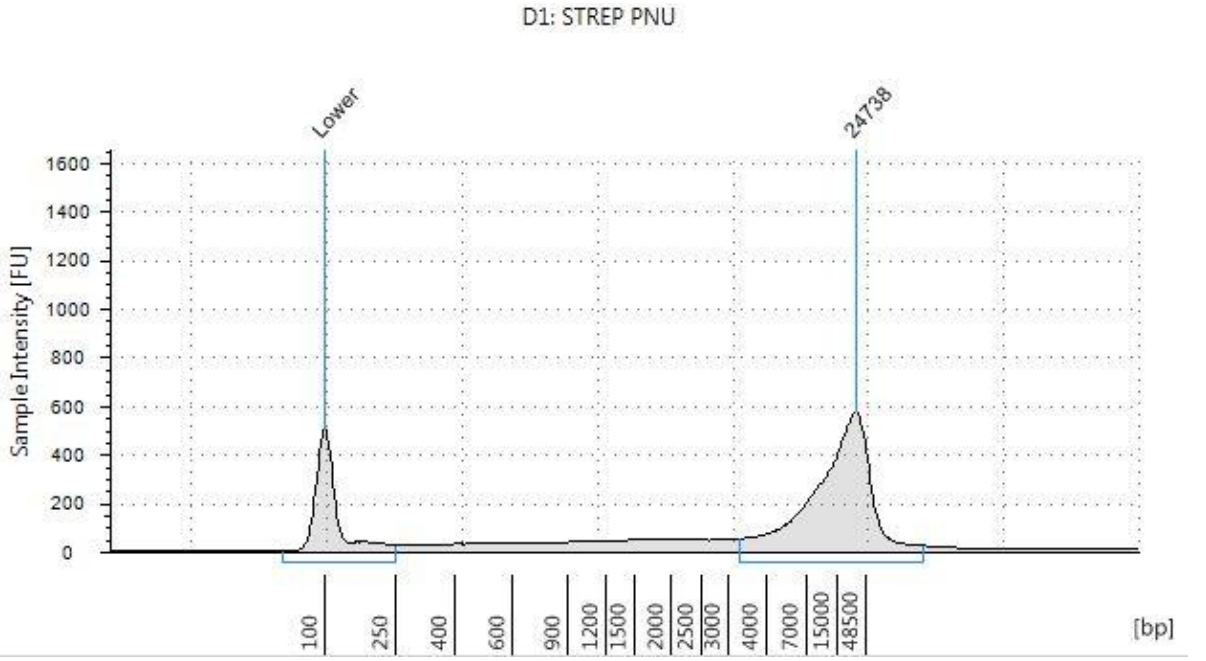


Figure 2: gDNA that FAILS fragment degradation QC. This sample doesn't need shearing further at least, but those small fragments are going to cause assembly problems. Gel size selection is very likely to be necessary.

Needle Shearing.

1. Place 2 – 10 ug gDNA diluted to 100ul with EB buffer into a clean 1.5ul LoBind tube.
2. Attach a 26 gauge blunt end needle (ThermoFisher UK Ltd HCA-413-030Y GC Syringe Replacement Parts 26g , 51mm) onto a 1ml luer-loc syringe.

3. Place tip of the needle into the bottom of the 1.5 ul LoBind sample tube and draw the plunger up to the 0.3 ml line on syringe. Once you see all the liquid from the bottom of the tube drawn into the needle press the plunger down to expel the sample back into the tube (no need to use excessive force*) repeat 4 - 5 times. The number of passes / repeats can be reduced for less concentrated DNA, you may need to experiment to find the best parameters in your own lab**.
4. Run samples on Agilent Tapestation to ensure accurate shearing

*/** some notes to add. It is difficult to express 'excessive force' in a protocol but we simply mean to say there is nothing abnormal about the aspiration speed. Encouragingly, we also found that you can monitor the shearing process after each pass, i.e., 1st pass = ~50 kb, 2nd pass = ~35 kb etc. and that with this needle we achieved ~25 kb after four or five passes. The final size didn't appear to decrease significantly with extra passes, we tested up to 10 passes, which suggests it's a robust technique too, but one that 'can' be monitored if the sample is limited / precious. We've done no testing yet to find if there is a correlation with GC content.

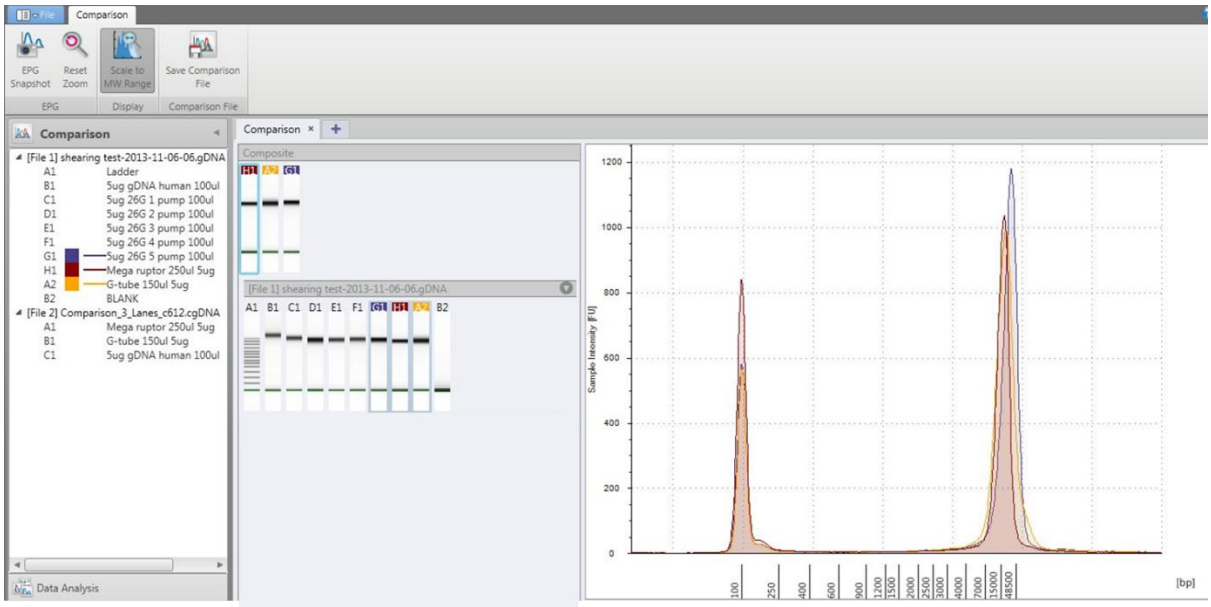
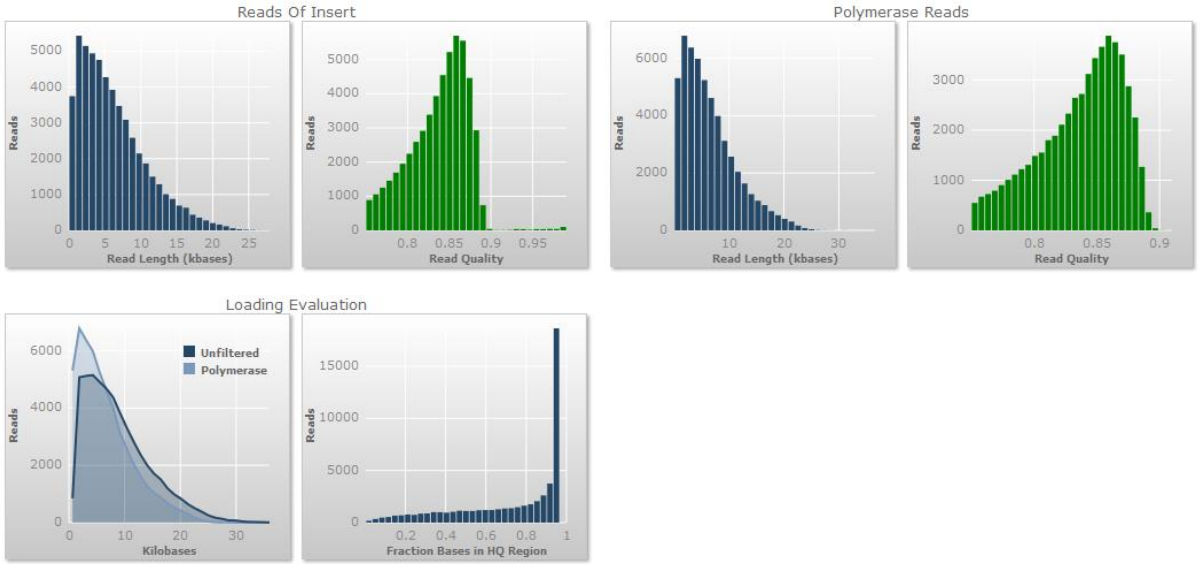


Figure 3: 5 ug >60 kb gDNA (gel image lane B1) sheared with a 26G needle, one pass (lane C1) two passes (lane D1) three passes (lane E1) etc. For comparison the graph shows the same sample sheared with G-tube (yellow) MegaRuptor, manufactured by Diagenode (red) and a 26G needle (purple).

The comparison in Figure 3 shows that the three techniques we've used so far for >20 kb PacBio libraries are relatively comparable, the G-tube gives a slightly wider distribution, the Mega Ruptor appears tightest, while the 26G needle lies in-between. Most importantly however, the needle shearing hasn't created smaller fragments; a problem for assembly, and the problem that size-selection is be required to fix.

Run Metrics.

Sequencing three cells of the same SMRTbell library generated sequence data with metrics shown in figure 4.



Collection Protocol	Insert Size (bases)	Stage Start	SMRT Cells	Total Reads	Polymerase Reads			Reads Of Insert		
					Length	Quality	Mbases	Length	Quality	Mbases
MagBead Standard Seq v2	10000	Yes	1	53038	6729	0.84	356.9	6404	0.84	339.7
MagBead Standard Seq v2	10000	Yes	1	52231	7470	0.84	390.2	6971	0.84	364.1
MagBead Standard Seq v2	10000	Yes	1	45438	7773	0.84	353.2	7305	0.84	331.9

Figure 4: Sequencing metrics for three SMRT cells (top panel showing distributions of one SMRTcell in detail).

Filtered sub-reads metrics

Using the SMRTcell with 53038 polymerase reads, and filtering with the parameters shown in the box below, generates the sub-read population shown in Figure 5.

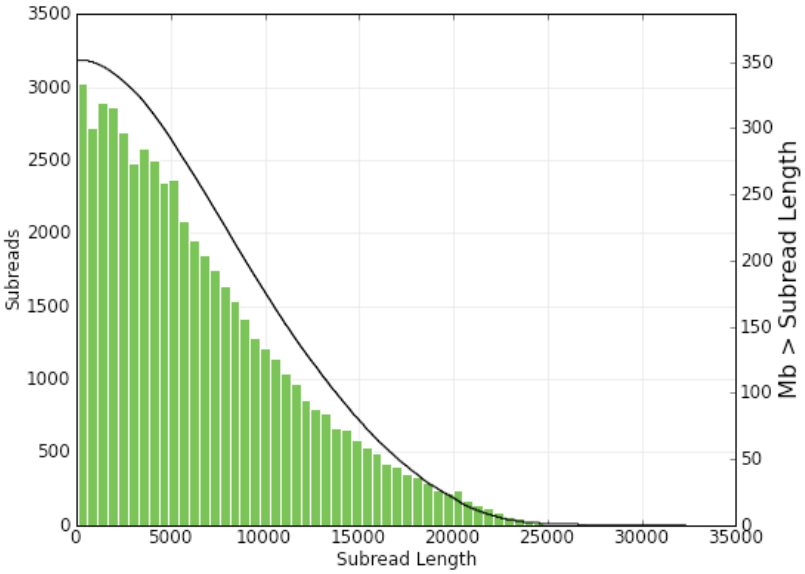
SFilter v1

Minimum Subread Length 250

Minimum Polymerase Read Quality 0.75

Minimum Polymerase Read Length 700

Description: This module filters reads based on a minimum subread length, polymerase read quality and polymerase read length.



Mean Subread length	6,307	N50	8,914
Total Number of Bases	353,978,332	Number of Reads	56,124

Figure 5: Sub-read population and distribution

Final Assembly metrics

Single SMRTcell Draft assembly of a salmonella strain using HGAP.2

Total length	No. contigs	Mean length	Max length	Min length	N50	N50n	MeanGC
5198243	3	1732747.667	5075345	12725	5075345	1	51.65

Contigs in draft assembly	length
0 scf7180000000006	5075345
1 scf7180000000007	12725
2 scf7180000000008	110173

After polishing the 12.7 Kb contig is removed, as almost no filtered_subreads map to scf7180000000007. This is therefore, assumed to be an assembly error that is easily identifiable with quiver, leaving the plasmid and genome in single contigs.