A High-Quality *de novo* Genome Assembly from a Single Mosquito using PacBio Sequencing

Primo Baybayan[1], Haynes Heaton[2], Juliana Cudini[2], Nancy Holroyd[2], Alan Tracey[2], Christine C. Lambert[1], Sarah Kingan[1], Brendan Galvin[1], Jonas Korlach[1], Matthew Berriman[2], and Mara K. N. Lawniczak[2]

1. Pacific Biosciences, Menlo Park, CA, USA
2. Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK

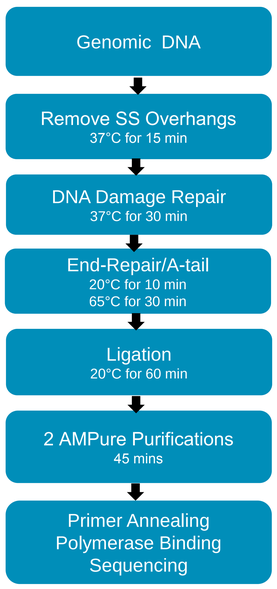
Abstract

A high-quality reference genome is an essential tool for studies of plant and animal genomics. PacBio Single Molecule, Real-Time (SMRT) Sequencing generates long reads with uniform coverage and high consensus accuracy, making it a powerful technology for *de novo* genome assembly. While PacBio is the core technology for many large genome initiatives, relatively high DNA input requirements (3 µg for standard library protocol) have placed PacBio out of reach for many projects on small, non-inbred organisms that may have lower DNA content.

Here we present high-quality *de novo* genome assemblies from single invertebrate individuals for two different species: the *Anopheles coluzzii* mosquito and the *Schistosoma mansoni* parasitic flatworm. A modified SMRTbell library construction protocol without DNA shearing and size selection was used to generate a SMRTbell library from just 150 ng of starting genomic DNA. The libraries were run on the Sequel System with chemistry v3.0 and software v6.0, generating a range of 21-32 Gb of sequence per SMRT Cell with 20-hour movies (10-12 Gb for 10-hour movies), and followed by diploid *de novo* genome assembly with FALCON-Unzip. The resulting assemblies had high contiguity (contig N50s over 3 Mb for both species) and completeness (as determined by conserved BUSCO gene analysis). We were also able to resolve maternal and paternal haplotypes for 1/3 of the genome in both cases.

By sequencing and assembling material from a single diploid individual, only two haplotypes are present, simplifying the assembly process compared to samples from multiple pooled individuals. This new low-input approach puts PacBio-based assemblies in reach for small, highly heterozygous organisms that comprise much of the diversity of life. The method presented here can be applied to samples with starting DNA amounts around 150 ng per 250 Mb – 600 Mb genome size.

SMRTbell Library Preparation Workflow



- MagAttract HMW DNA Kit yields DNA suitable for low-input library construction
- Single-tube, addition-only workflow minimizing DNA loss
- Yield >50%, >4 SMRT Cells
- ~3.5 hours preparation time
- No size-selection is required

DNA Requirements

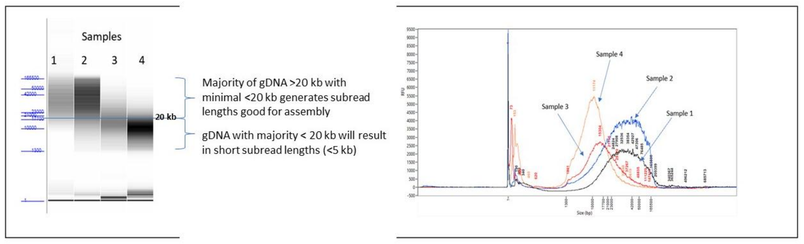
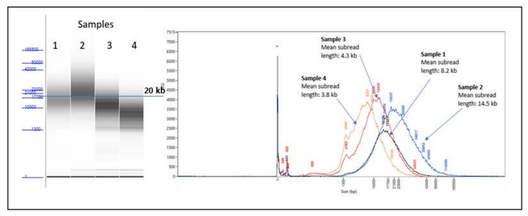


Figure 1. Quality of DNA for library construction. FEMTO *Pulse* gel images and traces of 4 gDNA samples. DNA 1 and 2 contain fragments with majority of DNA >20 kb (minimal fragments <20 kb) suitable for generating long reads for *de novo* assembly. DNA 3 and 4 are too fragmented with majority of DNA <20 kb and not recommended for library construction and sequencing.



Sample	Library Size	Subread Length	N50 Subread Length
1	17 kb	8.2 kb	12 kb
2	22 kb	14.5 kb	13 kb
3	12 kb	4.3 kb	7 kb
4	9 kb	3.8 kb	6 kb

Figure 2. Impact of <10 kb short insert SMRTbell templates on subread length. FEMTO *Pulse* gel images and traces of 4 SMRTbell libraries. The majority of inserts for libraries 3 and 4 are <10 kb, resulting in ~4 kb subread lengths.

Example: *Anopheles coluzzii*

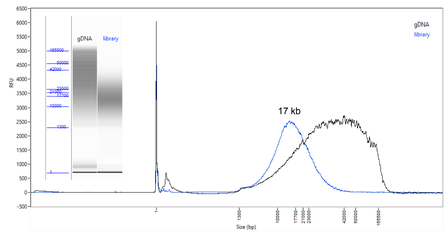


Figure 3. *Anopheles coluzzii* input DNA and final SMRTbell library. Majority of gDNA is >20 kb and the final SMRTbell library is 17 kb. 100 ng input gDNA.

Loading Conc.	Total Yield (Gb)	Unique Mol. Yield (Gb)	N50 Polymerase Read Length	N50 Subread Length	P0	P1	P2
5 pM	24.1	4.5	116,615	12,978	26.0%	60.1%	13.9%
5 pM	23.6	4.5	114,807	13,132	27.1%	59.0%	14.0%
6 pM	25.0	3.9	122,898	12,751	35.3%	53.1%	11.7%

Table 1. Run statistics for the three Sequel SMRT Cells run from the library (Chemistry 3.0, SW 6.0, 20 hour movie/SMRT Cell). A total of 45-fold unique molecular coverage (UMC) was generated for assembly.

		PacBio	Sanger assembly
Primary Contigs	Total Length	251 Mb	224 Mb
	No. contigs	206	27,063
	Contig N50	3.47 Mb	0.025 Mb
Alternate Haplotigs	Total Length	89.2 Mb	unresolved
	No. contigs	830	N/A
	Contig N50	0.199 Mb	N/A
BUSCO (diptera N=2799)	Complete	98.1 %	87.5 %
	Duplicate	2.4 %	0.1 %
	Fragmented	0.9 %	6.8 %
	Missing	1.0 %	5.7 %

Table 2. Assembly statistics of the PacBio *Anopheles coluzzii de novo* assembly. Stats compared with the previous Sanger-sequence based assembly for this species from [1] (GCA_000150765.1). BUSCO [2] was run on the PacBio primary contigs after curation with Purge Haplotigs [3].

Example: *Schistosoma mansoni*

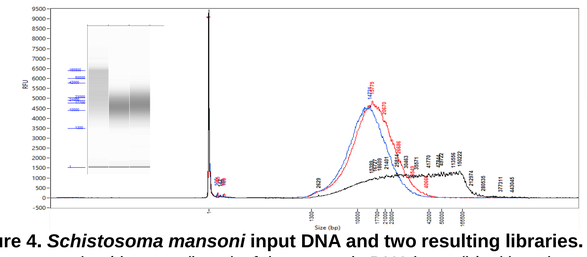


Figure 4. *Schistosoma mansoni* input DNA and two resulting libraries. FEMTO *Pulse* QC traces and gel images (inset) of the genomic DNA input (black) and two final libraries (blue: 100 ng input, yield 50%; red: 45 ng input, yield 55%). The second library (red) was used for sequencing.

Loading Conc.	Total Yield (Gb)	Unique Mol. Yield (Gb)	N50 Polymerase Read Length	N50 Subread Length	P0	P1	P2
4 pM	32.3	3.8	161,677	13,437	47.3%	44.3%	8.5%
4 pM	26.9	3.2	157,064	13,351	54.3	36.3%	9.4%
4 pM	32.5	3.8	168,417	13,586	50.7%	39.9%	9.4%
4 pM*	20.6	3.8	78,715	12,121	34.1%	52.7%	13.2%

Table 3. Run statistics for the four Sequel SMRT Cells run from the library. (Chemistry 3.0, SW 6.0, 20 hour runs/Cell; *10 hour run for the last Cell). A total of 41-fold UMC was generated for assembly.

	PacBio	2012 reference
Total Length	382 Mb	364 Mb
No. contigs	327	9,516
Contig N50	3.8 Mb	0.077 Mb
No. scaffolds	--	885
Scaffold N50	--	32 Mb
Gap length	0	2 Mb
BUSCO (eukaryota, N=303)		
Complete	86.1 %	84.8 %
Duplicate	17.8 %	3.0 %
Fragmented	4.0 %	5.3 %
Missing	9.9 %	9.9 %

Table 4. Assembly statistics of the PacBio *Schistosoma mansoni de novo* assembly. Stats compared with the previous assembly for this species (GCA_000237925.2). BUSCO was run on the combined primary contig and haplotigs PacBio assembly.

Conclusion

- The single-tube addition-only workflow with SMRTbell Express Template Prep 2.0 minimizes DNA loss enabling library construction and high-quality *de novo* genome assemblies.
- While we have successfully assembled a genome with 45 ng DNA, PacBio recommends starting with 150 ng.
- Size-selection is not required if majority of gDNA is >20 kb.
- Use 10-hour movies with 120 min pre-extension.
- Enables sequencing single individuals with no need for inbreeding or pooling multiple samples.
- For more information, please see *An. coluzzii* preprint: https://www.biorxiv.org/content/early/2018/12/19/499954

References

1. Lawniczak MKN et al. Widespread divergence between incipient *Anopheles gambiae* species revealed by whole genome sequences. *Science*. 2010;330: 512–514.
2. Simão FA et al. (2015) BUSCO: assessing genome assembly and annotation completeness with single- copy orthologs. *Bioinformatics*. 31(19), 3210-3212.
3. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*. 2018;19: 460.
4. https://www.vectorbase.org/organisms/anopheles-gambiae/pest/agamp4 [cited 7 Aug 2018]
5. Kingan et al. (2019). A High-Quality *De novo* Genome Assembly from a Single Mosquito Using PacBio Sequencing. *Genes*, *10*(1), 62 https://doi.org/10.3390/genes10010062

The authors would like to thank everyone who helped generate data for the poster.