SMRT Sequencing of the `Alalā Genome

Jill Muehling[1], Richard Hall[1], Primo Baybayan[1], Emily Hatas[1], Jenny Gu[1],Leona Chemnick[2], Marisa Korody[2], Bryce Masuda[2], Cynthia Steiner[2], Jolene Sutton[3], Oliver Ryder[2],
[1]Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025
[2]San Diego Zoo Institute for Conservation Research
[3]University of Hawaii at Hilo – Department of Biology

Abstract

Single Molecule Real-Time (SMRT) Sequencing was used to generate long reads for whole genome shotgun sequencing of the genome of the`alalā (Hawaiian crow). The ‘alala is endemic to Hawaii, and the only surviving lineage of the crow family, Corvidae, in the Hawaiian Islands. The population declined to less than 20 individuals in the 1990s, and today this charismatic species is extinct in the wild. Currently existing in only two captive breeding facilities, reintroduction of the ‘alalā is scheduled to begin in the Fall of 2016. Reintroduction efforts will be assisted by information from the ‘alalā genome generated and assembled by SMRT Technology, which will allow detailed analysis of genes associated with immunity, behavior, and learning.

Using SMRT Sequencing, we present here best practices for achieving long reads for whole genome shotgun sequencing for complex plant and animal genomes such as the ‘alalā genome. With recent advances in SMRTbell library preparation, P6-C4 chemistry and 6-hour movies, the number of useable bases now exceeds 1 Gb per SMRT Cell. Read lengths averaging 10 – 15 kb can be routinely achieved, with the longest reads approaching 70 kb. Furthermore, > 25% of useable bases are in reads greater than 30 kb, advantageous for generating contiguous draft assemblies of contig N_{50} up to 5 Mb. *De novo* assemblies of large genomes are now more tractable using SMRT Sequencing as the standalone technology. We also present guidelines for planning out projects for the *de novo* assembly of large genomes.

Workflow for Ultra Large Insert Libraries

Fig. 1. Recommended library construction workflow for ultra-large-insert SMRTbell libraries. Thorough QC of the genomic DNA, shearing optimization, gentle handling of the DNA and an additional DNA damage repair are necessary for generating long reads in the PacBio RS II.

PFGE and Shearing

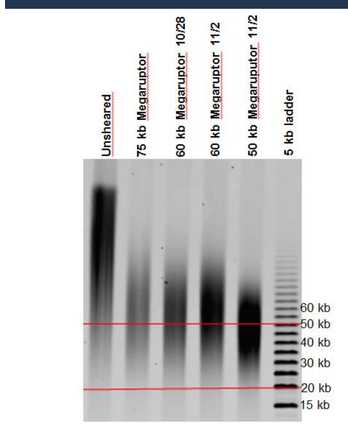

Fig. 2. Determine the best shearing condition by performing small volume test shears using the Megaruptor instrument (Diagenode). The ‘alalā gDNA was sheared using 75, 60 and 50 kb conditions. The best shearing condition was determined to be 50 kb with a 30-80 kb fragment distribution (ran on PFGE on Bio-Rad CHEF mapper system).

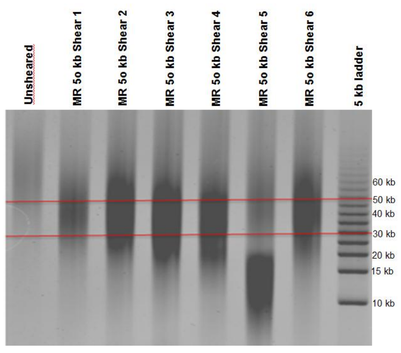

Fig. 3. Large-scale shears of the ‘alalā gDNA using the optimal 50 kb shear condition (86 µg of gDNA was recovered from shearing). Shear #5 was oversheared and was not included in the pool.

SMRTbell Library Construction

The sheared ‘alalā genome was constructed into SMRTbell template using the > 30 kb library construction procedure available on the PacBio website (http://www.pacb.com/wp-content/uploads/2015/09/Unsupported-Preparing-Greater-than-30kb-SMRTbell-Libraries-Megaruptor-Shearing.pdf), with minor modifications.

- 1x AMPure PB purification
- Room Temperature rotation instead of vortexing
- Two step elution process during AMPure PB elution to maximize recovery

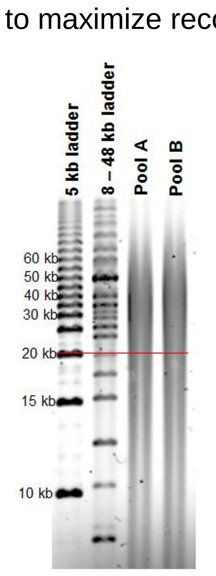

Fig. 4. Distribution of the final SMRTbell libraries were evaluated using PFGE (Pippin Pulse, Sage Science) to determine optimal size selection cutoff.

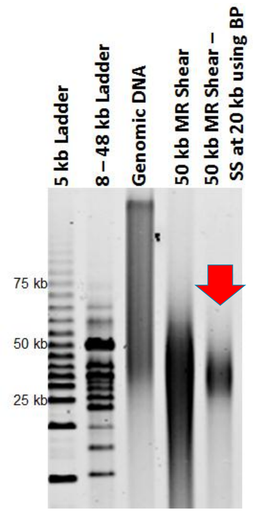

Fig. 5. Size-selected ‘alalā SMRTbell library using the BluePippin system. Size selection eliminates small-insert fragments < 20 kb (0.75% DF Marker S1 high-pass 15-20kb cassette definition) that preferentially load during sequencing. The size-selected library shows a mode of ~30 kb.

	Pre library construction Mass (ng)	Post library construction Mass (ng)	Post Size Selection Mass (ng)	Total % Yield	# of SMRT Cells Run	Loading Conc on Plate (pM)
Pool A	25380	7874	2310	9.1	22	100
					40	150
Pool B	27720	12400	2935	10.5	71	150

Table 1: ‘alalā SMRTbell library yields. The library yield after size selection depends on the desired size cutoff and fragment distribution. Prior to size selection, it is highly recommended to QC the SMRTbell library on PFGE to determine the appropriate size selection cutoff. For the ‘alalā library, yield after size selection using a cutoff of 20 kb is approximately 10% yielding enough library to sequence >130 SMRT Cells

Single Molecule, Real-Time Sequencing

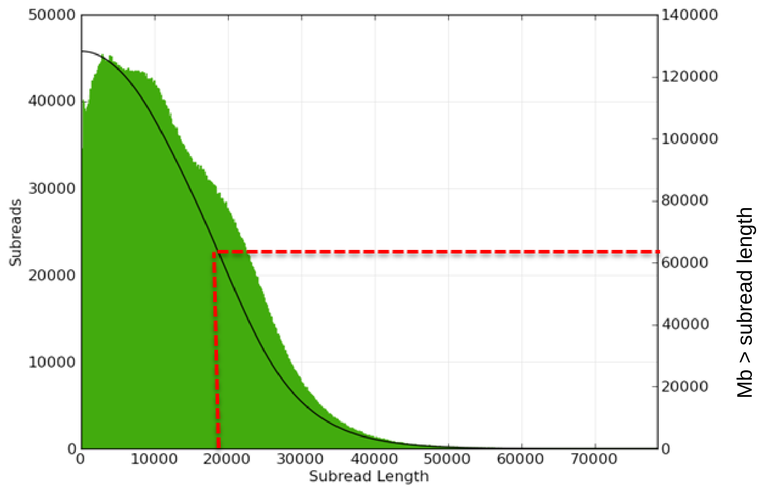

Fig. 6. N50 Subread Plot of 133 SMRT Cells used in the construction of the ‘alalā genome . N50 is the read length at which 50% of the bases are in subreads greater than or equal to this value.
Post filtering, the subread N50 is equal to 18,661 bp. Red line indicating 55-fold coverage of the expected genome size is covered by reads ≥ 20 kb.

PacBio *de novo* Assembly

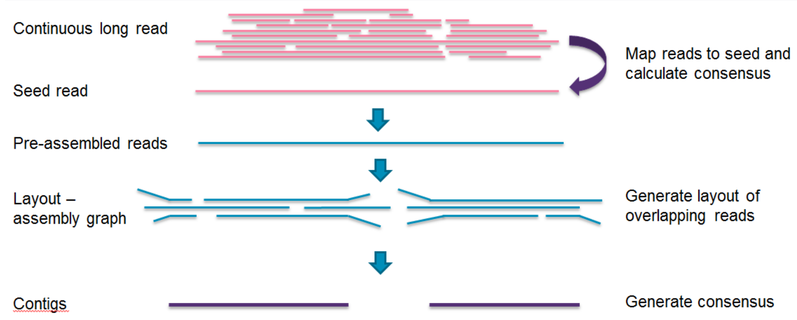

Fig. 7. Workflow for *de novo* assembly of SMRT Sequencing data. The hierarchical assembly approach is common to many assembly algorithms designed for long read data, including HGAP, MHAP and Falcon.

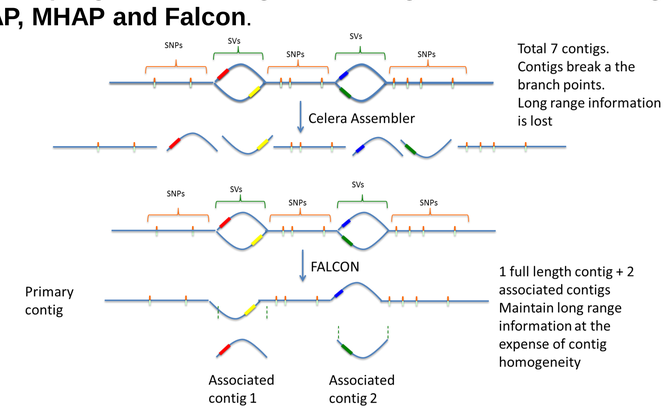

Fig. 8. FALCON assembler process. The FALCON assembler has been specifically designed as diploid aware. A set of alternative contigs are generated representing two haplotypes. https://github.com/PacificBiosciences/FALCON

	# Bases	N50	# Contigs	Largest Contig
‘alalā Primary Contigs	**1.09 Gb**	**11 Mb**	**1,026**	**53 Mb**
‘alalā Alternative Contigs	66 Mb	46 kb	1,435	281 kb
Corvus brachyrhynchos (scaffolds)	**1.05 Gb (1.09 Gb)**	**24 kb (6.9 Mb)**	**125,019 (33,301)**	**253 kb (26 Mb)**

Table 2: FALCON Assembly of the ‘alalā genome. When compared to the assembly of the American Crow (Corvus brachyrhynchos) the PacBio long reads generated contig N50 of 11 Mb and reduced the number of contigs by a factor of 10. A set of 229 core eukaryotic genes (CEGMA) from chicken were aligned against the assembly, resulting in 94.7% fully intact orthologs.

Conclusions

- A large insert library (30 kb) sequenced with PacBio P6-C4 chemistry allowed *de novo* assembly of the ‘alalā genome with contig N50 of 11 Mb.
- Best practices must be employed to successfully construct ultra-large insert libraries including:
  - QC of gDNA, test shears, scale-up shears and SMRTbell templates using PFGE
  - Shearing gDNA for uniform sample distribution, employing test shears prior to bulk sample shearing
  - Size selection to eliminate small inserts
  - Eliminate vortexing. use gentle handling of the DNA throughout the library process
  - DNA damage repair after size selection
- New technology advances with PacBio’s Sequel System will reduce cost by reducing the number of SMRT Cells required to assemble a large genome.
- To learn how the San Diego Zoo Institute for Conservation Research is going to use the ‘alalā genome assembly to assist in population management and reintroduction, attend the PacBio SMRT Sequencing workshop on Jan 12, 2016.

Acknowledgements

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