Profiling Complex Population Genomes with Highly Accurate Single Molecule Reads: Cow Rumen Microbiomes

Cheryl Heiner[1], Itai Sharon[2], Steve Oh[1], Alvaro G. Hernandez[3], Itzhak Mizrahi[4] and Richard Hall[1]

[1]PacBio, 1305 O'Brien Drive, Menlo Park, CA; [2]Tei-Hai College, Upper Galilee, and MIGAL Galilee Research Institute, Israel; [3]University of Illinois at Urbana-Champaign; [4]Ben-Gurion University of the Negev, Israel

Abstract

Determining compositions and functional capabilities of complex populations is often challenging, especially for sequencing technologies with short reads that do not uniquely identify organisms or genes. Long-read sequencing improves the resolution of these mixed communities, but adoption for this application has been limited due to concerns about throughput, cost and accuracy.

The recently introduced PacBio Sequel System generates hundreds of thousands of long and highly accurate single-molecule reads per SMRT Cell.

We investigated how the Sequel System might increase understanding of metagenomic communities. In the past, focus was largely on taxonomic classification with 16S rRNA sequencing. Recent expansion to WGS enables functional profiling as well, with the ultimate goal of complete genome assemblies.

Here we compare the complex microbiomes in 5 cow rumen samples, for which Illumina WGS sequence data was also available. To maximize the PacBio single-molecule sequence accuracy, libraries of 2 to 3 kb were generated, allowing many polymerase passes per molecule. The resulting reads were filtered at predicted single-molecule accuracy levels up to 99.99%.

Community compositions of the 5 samples were compared with Illumina WGS assemblies from the same set of samples, indicating rare organisms were often missed with Illumina. Assembly from PacBio CCS reads yielded a contig >100 kb in length with 6-fold coverage. Mapping of Illumina reads to the 101 kb contig verified the PacBio assembly and contig sequence. Scaffolding with reads from a PacBio unsheared library produced a complete genome of 2.4 Mb.

These results illustrate ways in which long accurate reads benefit analysis of complex communities.

Workflow: Library Prep to Analysis

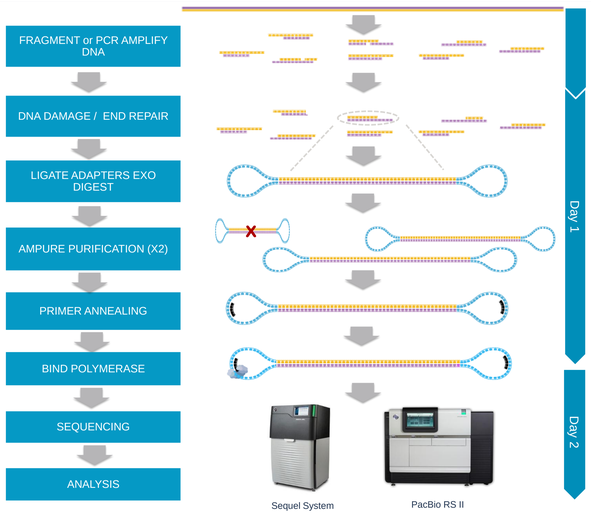


Figure 1. Workflow for long read metagenomic profiling on PacBio Systems. The entire process from shearing and library prep through sequencing and CCS analysis can be completed in under 48 hours.

Profiling Populations from Sheared Genomic DNA

2 to 3 kb reads from sheared metagenomic DNA can be utilized to determine taxonomic composition and profile community functions; this size has many advantages:

- 2 to 3 kb reads include many passes, which are used to generate highly accurate sequence from a single molecule:

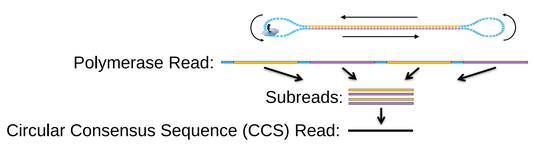
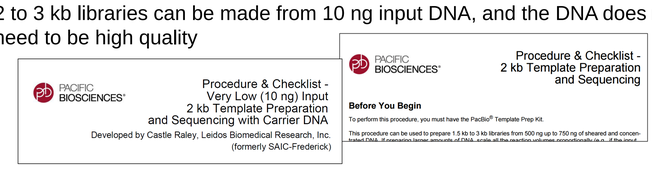


Figure 2. Multiple reads generated from a single molecule.

- 2 to 3 kb reads often span 1 or more entire gene sequences
- Abundance of community members (relative to genome size) are maintained in the data, since there is no amplification step, and minimal bias in PacBio sequencing
- A single long read with a unique match to a published sequence is sufficient to determine presence
- 2 to 3 kb libraries can be made from 10 ng input DNA, and the DNA does not need to be high quality


PACIFIC BIOSCIENCES®

Procedure & Checklist - Very Low (10 ng) Input 2 kb Template Preparation and Sequencing with Carrier DNA

Developed by Castle Raley, Leidos Biomedical Research, Inc. (formerly SAIC-Frederick)

PACIFIC BIOSCIENCES®

Procedure & Checklist - 2 kb Template Preparation and Sequencing

Before You Begin

To perform this procedure, you must have the PacBio® Template Prep Kit.

This procedure can be used to prepare 1.5 kb to 3 kb libraries from 500 ng up to 750 ng of sheared and concentrated DNA. If preparing larger amounts of DNA, scale all the reaction volumes proportionally (e.g., if the input

For all PacBio library prep and sequencing protocols, visit http://www.pacb.com/support/documentation/

WGS of Cow Rumen Microbiomes

Samples and library prep

- Cow rumen microbiomes from 5 samples were compared using WGS
- For each sample, 1 µg of DNA was sheared to ~3 kb for SMRTbell library prep

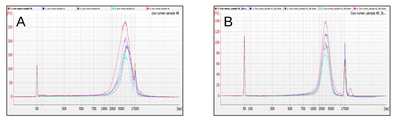


Figure 3. Bioanalyzer electropherograms of input sample **(A)** and samples after shearing to 3kb **(B)**. Shearing was done using the Covaris® S2 Focused-ultrasonicator according to the manufacturer's instructions.

- Longer insert libraries were also made for 2 of the samples, 43 and 44. As is typical for metagenomic samples, input DNA was not high molecular weight (Figure 3A, above). To generate the longest reads possible, no shearing was done for these libraries; 1 µg of DNA / sample was taken directly into SMRTbell library prep.

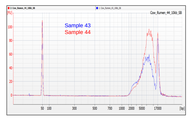


Figure 4. Bioanalyzer electropherograms of final SMRTbell libraries generated from unsheared DNA, samples 43 and 44.

Sequencing and CCS analysis

- **~3 kb SMRTbell libraries** were run on the Sequel System
- CCS sequences were generated and filtered at several different levels of predicted accuracy. Approximately 100,000 reads of >99% predicted accuracy were produced from 1 Sequel SMRT Cell 1M for every sample.

Sample, ~3 kb library		Primary Analysis Results, 1 SMRT Cell 1 M				CCS Filtering Criteria (# of Reads @ Minimum Predicted Accuracy)		
		Gb	# of Primary (P1) Reads	Polymerase Read Length	Insert Read Length	90% Accurate 2 passes	99% Accurate 3 passes	99.9% Accurate 3 passes
CR43	calf	**5.13**	449,658	11,411	2,781	207,078 / 98.2%	**116,675** / 99.7%	36,416 / 99.96%
CR44	adult	**4.68**	435,439	10,743	2,952	182,177 / 98.1%	**95,724** / 99.7%	28,122 / 99.96%
CR45	calf	**3.90**	339,470	11,494	2,764	155,634 / 98.4%	**93,129** / 99.77%	31,006 / 99.96%
CR46	adult	**4.50**	420,208	10,704	2,798	179, 261 / 98.5%	**110, 343** / 99.77%	43,593 / 99.96%
CR48	adult	**5.86**	513,653	11,405	2,880	227,382 / 98.5%	**141,526** / 99.8%	59,104 / 99.96%

Table 1. Primary analysis and CCS results from ~3 kb sheared libraries, 1 SMRT Cell 1M per sample, Sequel System, v1.2.1 sequencing chemistry

- **Unsheared libraries** were also run on the Sequel System under the same conditions. The data from the larger libraries was used directly without CCS analysis.

Sample		Gbases	# of Primary (P1) Reads	Polymerase Read Length	Insert Read Length
CR43	calf	4.38	487,837	9,363	**4,243**
CR44	adult	4.37	458,715	9,541	**4,759**

Table 2. Primary analysis results from unsheared libraries, 1 SMRT Cell 1M per sample, Sequel System, v1.2.1 chemistry

Gene prediction

- Predicted genes were determined using Prodigal (Prokaryotic Dynamic Programming Genefinding Algorithm)[1] in the consensus sequence and the amino acid sequence are calculated. Each read contained an average of nearly 3 full-length genes. Diamond[2] was used to align the putative protein sequences to the RefSeq protein database.

Sample		# of Sequel Cells	CCS Reads (≥99% Accuracy)	CCS N50 Read Length	# of Predicted Genes	Predicted Genes / Read	# of Full-Length Genes	Full-length Genes / Read
CR43	calf	2	180,849	2,518	736,199	4.07	486,669	**2.69**
CR44	adult	3	226,244	2,731	1,037,382	4.59	727,738	**3.22**
CR45	calf	2	147,971	2,667	635,924	4.30	432,048	**2.92**
CR46	adult	3	283,198	2,652	1,215,590	4.29	817,121	**2.89**
CR48	adult	2	239,282	2,603	1,011,589	4.23	669,335	**2.80**

Table 3. Predicted genes from protein alignments using calculated amino acid sequences

Cow Rumen Microbiome Communities

Community composition

- Composition by order was determined for each sample using MEGAN[3]

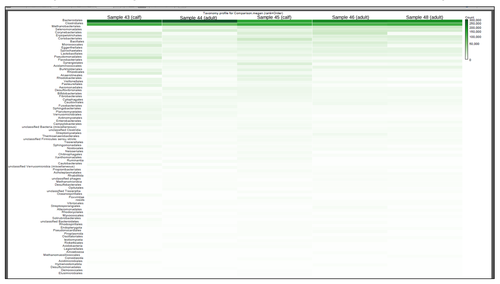


Figure 5. Order ranking. Rank comparison of orders found in the 5 communities using MEGAN for taxonomic assignments

Protein sequences

- Consistent variants from the reference were found in several single molecule reads of one sample:

Figure 6. *Ruminococcus flavefaciens* assignments and example alignment

Comparison with Short-read Data

Compared to Illumina assemblies, PacBio data* has a higher fraction of rare and most abundant organisms

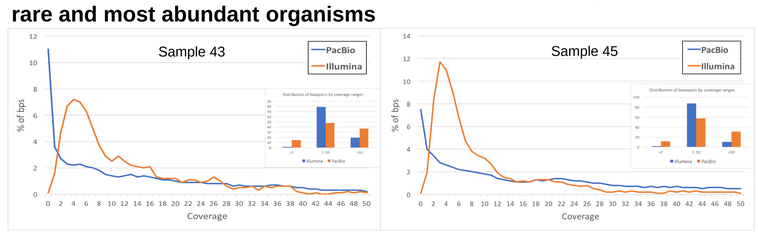


Figure 7. Distribution of bases according to coverage of assembled contigs (>500 bp) Inset: Distribution of base pairs by coverage ranges (<2-fold, 2- to 50-fold, >50-fold)

Assembly from PacBio CCS reads* generates long contigs

- Minimus2[4] assembly generated PacBio contig 101,539 bp long

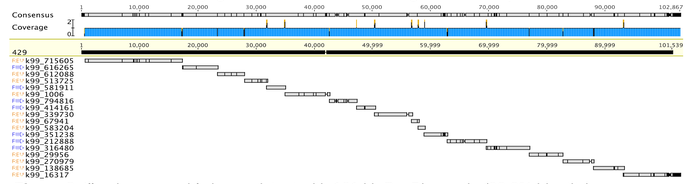


Figure 8. Illumina assembly is consistent with 101 kb PacBio contig (99.9% identity)

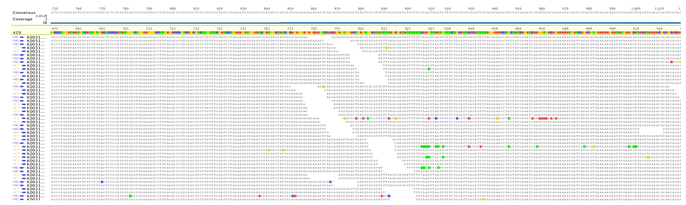


Figure 9. Illumina read mapping is highly consistent with PacBio contig

PacBio assembly from CCS reads* does not require high coverage

- PacBio reads* align at 99.6% identity to assembled contig

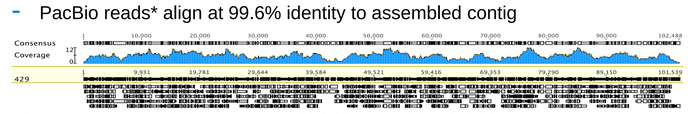


Figure 10. PacBio read alignment shows high identity with low (6-fold) coverage

*99% predicted accuracy CCS reads

Complete Genome Assembly

- The 3 kb CCS reads from sample 43 were assembled using Canu[5]
- Sequences from unsheared libraries were then added to the assembly to scaffold contigs from CCS reads
- The assembly produced one complete 2.4 Mb circular genome and a 4.0 Mb genome in a handful of contigs
- The 2.4 Mb genome showed 92% identity with *Porphyromonas endodontalis*

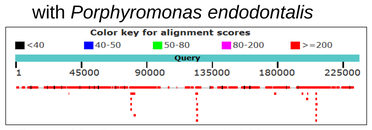


Figure 11. Blast results from 2.4 Mb genome

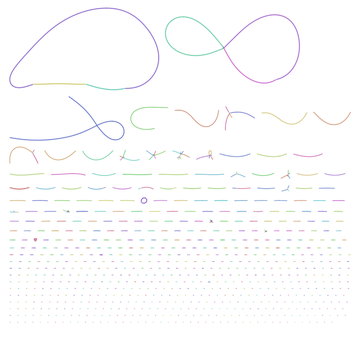


Figure 12. Assembly graph - Sample 43

Conclusions and References

- Microbiome CCS sequences 2 to 3 kb in length often contain multiple genes, which may be sufficient for identifying community members
  - This approach provides better coverage of low abundance community members compared to short-read WGS assemblies
- Microbiome assemblies using PacBio CCS sequences can generate contigs >100,000 kb with low (6-fold) coverage
  - Complete or near-complete genomes may be obtained using longer reads from unsheared libraries for scaffolding
- PacBio contig sequences and assemblies from microbiome CCS reads are highly consistent with Illumina data.

References

[1] Hyatt D, et al. (2012). Gene and translation initiation site prediction in metagenomic sequences. *Bioinformatics*. 28(17), 2223-2230

[2] https://omictools.com/diamond-tool

[3] Huson DH, et al. (2011) Integrative analysis of environmental sequences using MEGAN 4. *Genome Research*. 2011. 21(9),1552-1560.

[4] http://amos.sourceforge.net/wiki/index.php/Minimus2

[5] Koren S, et al. (2016) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *bioArxiv.* doi:10.1101/071282.