University of Illinois at Urbana-Champaign

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, 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 sequencing enables functional profiling as well, with the ultimate goal of complete genome assemblies.

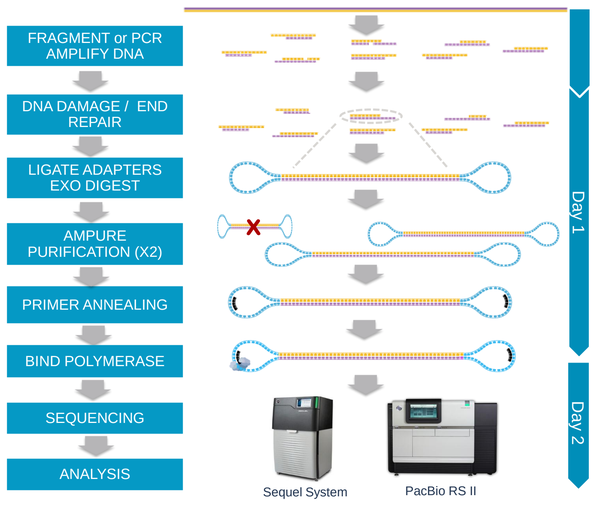
Here we compare the complex microbiomes in 5 cow rumen samples, for which Illumina WGS sequence data was also available. To maximize the PacBio single-molecule sequence accuracy, libraries of 2 to 3 kb were generated, allowing many polymerase passes per molecule. The resulting reads were filtered at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

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.



Long 16S Sequencing

Shared Protocol: Full-Length 16S PCR, Library Prep and Sequencing

- Includes tips for minimizing chimeras
- Requires high-fidelity PCR polymerase

ZymoBIOMICS™ Microbial Community DNA Standard

- Clean amplification (Bioanalyzer, right)
- 44% SMRTbell library prep yield
- Sequencing results shown below:

Figure 2. Bioanalyzer trace of full-length 16S amplicon, Zymo Microbial Community

Number of Primary Reads	Minimum Predicted Accuracy	Number of CCS reads	Number of CCS Bases	CCS Read Score (mean)	Number of Passes (mean)
412,577	0.9	221,526	348,494,608	0.9925	10
	0.99	173,898	271,385,368	0.9983	12
	0.999	105,082	162,659,830	0.9997	15
	0.9999	42,247	65,107,469	1	17

Table 1. Sequencing results, 1.5 kb full-length 16S amplicon, 1 SMRT Cell 1M, Sequel System

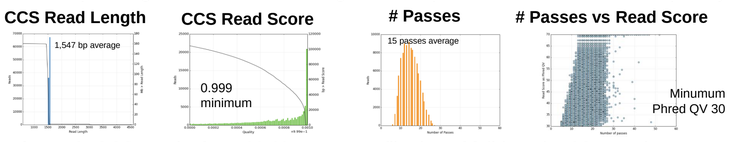


Figure 3. CCS results from 99.9% accuracy filtering of full-length 16S sample

Extremely accurate sequences obtained from a shorter region: 625 bp V3-V5 hypervariable region (data from a different sample)

Number of Primary Reads	Minimum Predicted Accuracy	Number of CCS reads	Number of CCS Bases	CCS Read Score (mean)	Number of Passes (mean)
797,532	0.9	386,440	244,657,437	0.9941	20
	0.99	324,929	204,283,315	0.9988	23
	0.999	232,238	145,592,356	0.9999	27
	0.9999	160,661	100,556,067	1	31
	0.99999	128,619	80,420,437	1	33

Table 2. Sequencing results, v3-v5 16S amplicon, 1 SMRT Cell 1M, Sequel System

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 includes many passes, which are used to generate highly accurate sequence from a single molecule:

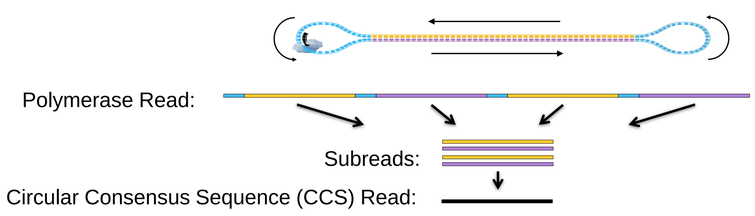
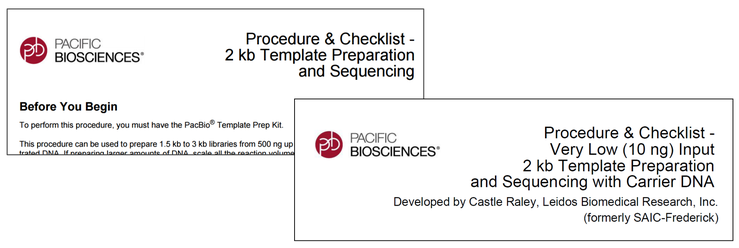


Figure 4. Multiple reads generated from a single molecule.

- 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



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 sequencing.
- For each sample, 1 µg of DNA was sheared to ~3 kb for SMRTbell library prep.

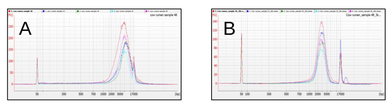


Figure 5. **Bioanalyzer electropherograms** of input sample **(A)** and samples after shearing to 3kb **(B)**. Shearing was done using the Covaris® S2 Focused-ultrasonicator according to the manufacturer's instructions.

Sequencing

- Samples were run on the Sequel System (v1.2.1 chemistry)

Sample		Cell	Gbases	# of Primary (P1) Reads	Polymerase Read Length	Insert Read Length
CR43	calf	1	5.13	449,658	11,411	2,781
		2	2.33	183,776	12,656	2,488
CR44	adult cow	1	4.68	435,439	10,743	2,952
		2	4.24	451,406	9,389	2,957
CR45	calf	1	3.90	339,470	11,494	2,764
		2	2.48	254,463	9,740	2,753
CR46	adult cow	1	4.50	420,208	10,704	2,798
		2	4.50	406,313	11,086	2,789
CR48	adult cow	1	5.86	513,653	11,405	2,880
		2	4.29	439,459	9,760	2,886

Table 3. Sequencing results from 2 SMRT Cells 1M per sample, Sequel System

CCS Analysis

- CCS sequences were generated and filtered at several different levels of predicted accuracy.

Sample		CCS Filtering Criteria (# of Reads @ Minimum Predicted Accuracy)		
		90% Accurate 2 passes	99% Accurate 3 passes	99.9% Accurate 3 passes
CR43	calf	207,078 / 98.22%	116,675 / 99.70%	36,416 / 99.96%
CR44	adult	182,177 / 98.05%	95,724 / 99.69%	28,122 / 99.96%
CR45	calf	155,634 / 98.35%	93,129 / 99.72%	31,006 / 99.96%
CR46	adult	179, 261 / 98.46%	110, 343 / 99.74%	43,593 / 99.96%
CR48	adult	227,382 / 98.49%	141,526 / 99.75%	59,104 / 99.96%

Table 4. CCS results from 1 SMRT Cell 1M per sample, Sequel System

Gene Prediction

- Predicted genes were determined using Prodigal (**Pro**karyotic **D**ynamic **P**rogramming **G**enefinding **Al**gorithm)[1] in the consensus sequence and the amino acid sequence are calculated. 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 5. 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]

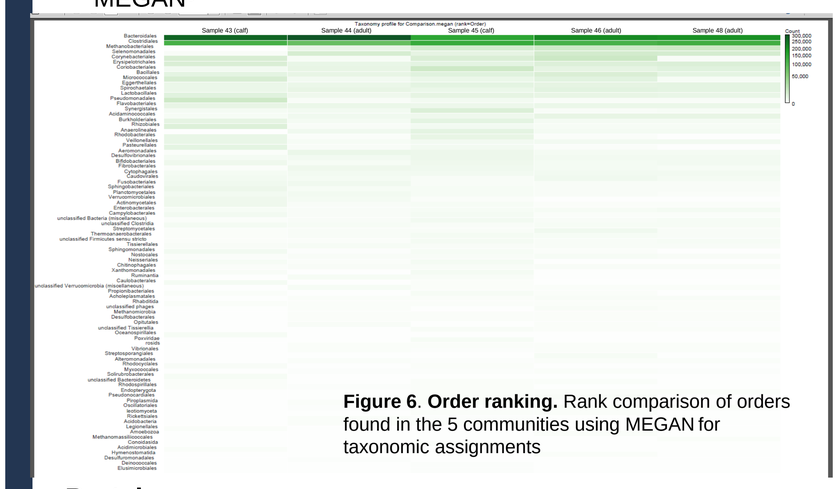


Figure 6. **Order ranking.** Rank comparison of orders found in the 5 communities using MEGAN for taxonomic assignments

Protein sequences

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

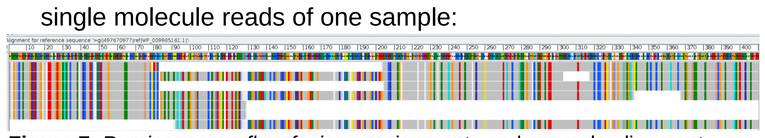


Figure 7. *Ruminococcus flavefaciens* assignments and example alignment

Comparison with Short-read Data

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

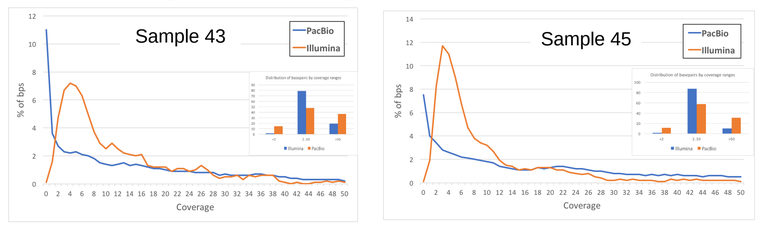


Figure 8. Distribution of bases according to coverage of assembled contigs (>500 bp) Inset: Distribution of base pairs by coverage ranges (<2-fold, 2- to 50-fold, >50-fold)

Assembly from PacBio CCS reads* generates long contigs

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

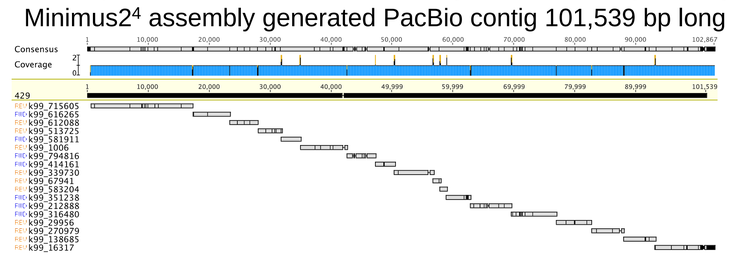


Figure 9. Illumina assembly is consistent with 101 kb PacBio contig (99.9% identity)

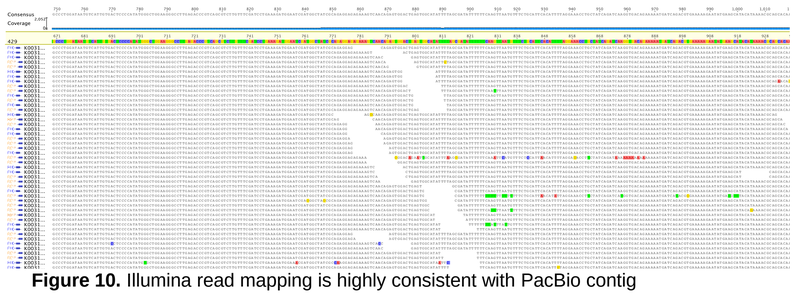


Figure 10. Illumina read mapping is highly consistent with PacBio contig

- Canu[5] assembly generated 675 kb PacBio contig, also consistent with Illumina

PacBio assembly from CCS reads* does not require high coverage

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

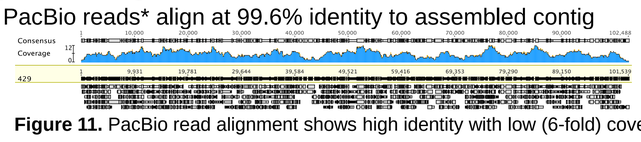


Figure 11. PacBio read alignment shows high identity with low (6-fold) coverage

*99% predicted accuracy CCS reads

Conclusions and References

- We demonstrate that PacBio Systems can generate highly accurate single-molecule sequences from templates up to several kb in length, providing important information for analysis of populations of genomes which may be difficult to obtain from short-read data.
- Single molecule CCS sequences ≥1 kb are often sufficient for identifying community members, providing high, unbiased coverage of low abundance community members not found in short-read WGS assemblies.
- Microbiome assemblies using PacBio CCS sequences generate contigs >100,000 kb with 6-fold coverage.
- PacBio contig sequences and assemblies were highly consistent with Illumina data.

[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 D.H. et al., (2011) Integrative analysis of environmental sequences using MEGAN *Genome Research*. 2011. 21(9),1552-1560
[4] http://amos.sourceforge.net/wiki/index.php/Minimus2
[5] http://biorxiv.org/content/biorxiv/early/2016/08/24/071282.full.pdf