PacBio Only Assembly with Low Genomic DNA Input

Zhiying Zhao[1*], Yu-Chih Tsai[2], Alicia Clum[1], Katherine Munson[1], Chris Daum[1, 3], Stephen W. Turner[2], Jonas Korlach[2], Len A. Pennacchio[1], Feng Chen[1]

[1]Department of Energy Joint Genome Institute // LBNL - Walnut Creek, CA
[2]Pacific Biosciences – Menlo Park, CA
[3]Lawrence Livermore National Laboratory – Livermore, CA

[a]To whom correspondence may be addressed. E-mail: ZYZhao@lbl.gov

March 25, 2013

ACKNOWLEDGMENTS:

The work conducted by the US Department of Energy (DOE) Joint Genome Institute is supported by the Office of Science of the DOE under Contract Number DE-AC02-05CH11231. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government, or any agency thereof, or the Regents of the University of California.

DISCLAIMER:

[**LBNL**] This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

[**LLNL**] This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

PacBio Only Assembly with Low Genomic DNA Input

Zhiying Zhao[1*], Yu-Chih Tsai[2], Alicia Clum[1], Katherine Munson[1], Chris Daum[1], Stephen W. Turner[2], Jonas Korlach[2], Len A. Pennacchio[1], Feng Chen[1]

1. Department of Energy, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, 94598. 2. Pacific Biosciences, 1380 Willow Rd., Menlo Park, CA, 94025

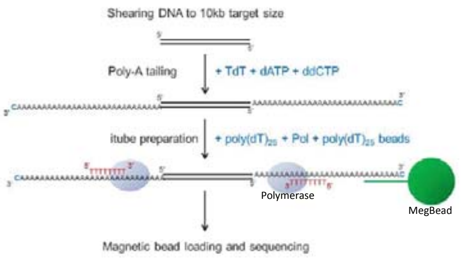
INTRODUCTION

The assembly and analysis of microbial species on earth remains a largely unexplored area of life. This is partially due to their inability to be cultured but also based on the large historic cost of drafting and finishing individual microbial species genomes.

The single-molecule real-time (SMRT^{TM}) sequencing platform developed by Pacific Biosciences (PacBio) offers several benefits including Single Molecule real-time analysis, longer read length at fast speed, low sequencing redundancy and bias. Thus, it was used at JGI as a quick-turnaround and cost-effective solution for finishing microbial genomes.

Construction of PacBio library by traditional protocol still requires micrograms of genomic DNA. In many cases, getting high quantity of genomic DNA remains as a major challenge. Recently, PacBio developed a more efficient library construction method using terminal deoxynucleotidyl transferase (TdT), which makes it possible to obtain sufficient sequencing data for assembly from significantly smaller amount of genomic DNA. We have tested and validated this newly developed method. Preliminary analysis results suggested that this technology can be used for microbial genome assembly with PacBio only data.

PRINCIPLE OF THE METHOD


Shearing DNA to 10kb target size
5'
5'
Poly-A tailing
+ TdT + dATP + ddCTP
3' CAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAC 3'
itube preparation
+ poly(dT)25 + Pol + poly(dT)25 beads
5' TTTTTTTT 3'
Polymerase
MegBead
Magnetic bead loading and sequencing

Standard SMRTbell 10kb lib

5-10ug

Shear DNA to 10kb 1.5h
(Qubit & Bioanlyzer)
↓
Damage & end Repair 1h
↓
Blunt end ligation & Exo treatment 3.5h
(Qubit & Bioanlyzer)

Total Prep Time 6h

Lib Yield 15-25%

TdT 10kb lib

100-200ng

Shear DNA to 10kb 1h
(Qubit)
↓
Damage Repair 1h
↓
Exo treatment & Poly-A tailing 2h
(Qubit)

Total Prep Time 4h

Lib Yield 70-80%

LIBRARY CONSTRUCTION

Ten bacterial samples (various GC% and genome size) are selected for validation. The library creation process begins with fragmenting genomic DNA (100-200ng) to 10kb using Covaris G-tube, followed by damage repair, quick exonuclease treatment and PolyA tailing. Ampure SPRI beads are used throughout library preparation process to select and purify sample DNA. The total preparation time is shorter then standard SMRTbell library construction processes. The library could be constructed within 4hours.

Sample Info			Input	Library Info	
Sample Name	GC%	Genome Size (MB)	ng	ng	Yield %
Tolumonas sp. BRL6-1	47	4.1	100	81	**81%**
Gillisia sp. JM1	34	5.4	100	72	**72%**
Teredinibacter sp. strain 991H.S.0a.06	50	7.2	200	160	**80%**
Geopsychrobacter electrodiphilus DSM 16401	53	5.0	200	136	**68%**
Hippea medeae KM1	43	4.7	181	123	**68%**
Desulfospira joergensenii DSM 10085	50	6.3	70	45	**65%**
Streptomyces sp. WmmB714	72	6.6	200	152	**76%**
Nocardia sp. BMG111209	69	9.1	200	127	**63%**
Nocardia sp. BMG51109	68	8.8	116	94	**81%**
Meiothermus ruber DSM 1279	63	3.1	100	73	**73%**

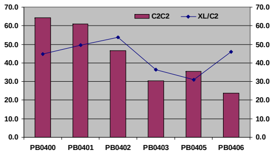
SEQUENCING RUN AND RESULTS

Sequencing run was done with Magbead loading, stage start, and 120min movies on either V2 or V3 chips, targeting 100x coverage per genome.

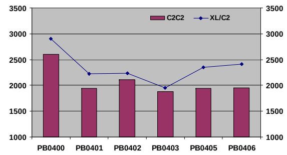
Sample Info			Sequencing Chemistry (# of Cells)		Sequencing Results			
Sample Name	GC%	Genome Size (MB)	C2/C2	XL/C2	PF # of Reads/cell	PF Mean RL (bp)	PF # of Mb/cell	PF Mean RQ
Tolumonas sp. BRL6-1	47	4.1	12	4	10225	2950	29.9	82.5%
Gillisia sp. JM1	34	5.4	12	4	11714	2848	32.9	81.2%
Teredinibacter sp. strain 991H.S.0a.06	50	7.2	6	6	20000	2755	54.4	83.1%
Geopsychrobacter electrodiphilus DSM 16401	53	5.0	4	5	26325	2101	54.7	83.0%
Hippea medeae KM1	43	4.7	6	2	22597	2156	48.4	79.8%
Desulfospira joergensenii DSM 10085	50	6.3	6	5	17268	1911	33.0	80.1%
Nocardia sp. BMG111209	69	9.1	8	8	15729	2146	33.3	82.3%
Nocardia sp. BMG51109	68	8.8	4	8	16762	2263	38.6	82.3%
Meiothermus ruber DSM 1279	63	3.1	6	0	29513	1899	55.7	83.3%

Differences between sequencing chemistries: XL/C2 tends to give longer read length. There is no clear trend for per cell output in terms number of bases.

PF # of Mb/cell



PF Mean RL (bp)

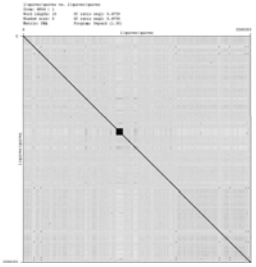


DATA ANALYSIS AND RESULTS

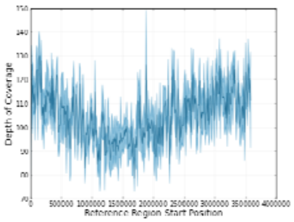
Data analysis with HGAP (de novo assembly using TdT read data only) and subsequent SMRT analysis for base methylation detection.

Two examples are presented here:

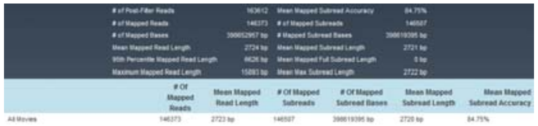
Tolumonas sp. BRL6-1: HGAP produced 1 contig with 3,598,394 bases.



Remapping Coverage Examination



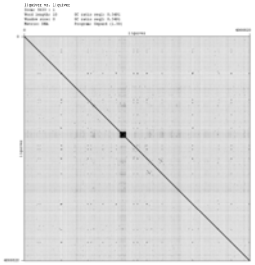
Remapping Quality Examination



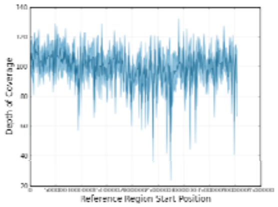
# of Post-Filter Reads	163612	Mean Mapped Subread Accuracy	84.75%
# of Mapped Reads	146373	# of Mapped Subreads	146507
# of Mapped Bases	398652957 bp	# Mapped Subread Bases	398619395 bp
Mean Mapped Read Length	2724 bp	Mean Mapped Subread Length	2721 bp
95th Percentile Mapped Read Length	6626 bp	Mean Mapped Full Subread Length	0 bp
Maximum Mapped Read Length	15093 bp	Mean Max Subread Length	2722 bp

	# Of Mapped Reads	Mean Mapped Read Length	# Of Mapped Subreads	# Of Mapped Subread Bases	Mean Mapped Subread Length	Mean Mapped Subread Accuracy
All Movies	146373	2723 bp	146507	398619395 bp	2720 bp	84.75%

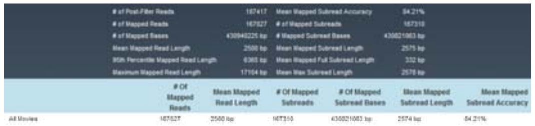
Gillisia sp. JM1: HGAP produced 1 contig with 4,066,858 bases.



Remapping Coverage Examination



Remapping Quality Examination



# of Post-Filter Reads	187417	Mean Mapped Subread Accuracy	84.21%
# of Mapped Reads	167027	# of Mapped Subreads	167310
# of Mapped Bases	430940225 bp	# Mapped Subread Bases	430821063 bp
Mean Mapped Read Length	2580 bp	Mean Mapped Subread Length	2575 bp
95th Percentile Mapped Read Length	6365 bp	Mean Mapped Full Subread Length	332 bp
Maximum Mapped Read Length	17104 bp	Mean Max Subread Length	2578 bp

	# Of Mapped Reads	Mean Mapped Read Length	# Of Mapped Subreads	# Of Mapped Subread Bases	Mean Mapped Subread Length	Mean Mapped Subread Accuracy
All Movies	167027	2580 bp	167310	430821063 bp	2574 bp	84.21%

m6A methylation motifs were also detected in both genomes.

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.