LA-UR-11-10745
Approved for public release; distribution is unlimited.

Title:	Exploring Applications for the PacBio RS in the Sequencing Workflow at LANL
Author(s):	Reitenga, Krista G.
Intended for:	Sequencing, Finishing, Analysis in the Future, 2011-06-01/2011-06-03 (Santa Fe, New Mexico, United States)

Exploring Applications for the PacBio RS in the Sequencing Workflow at LANL

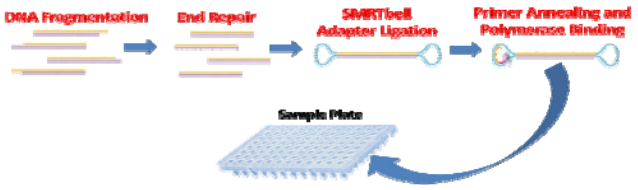
Krista Reitenga and Genome Science Group (B-6), Bioscience Division

Los Alamos NATIONAL LABORATORY

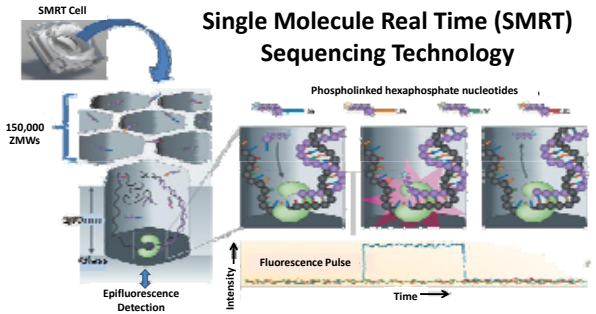
The Genome Science Group in the Bioscience Division at Los Alamos National Laboratory (LANL) works with a variety of LANL-internal and external collaborators and sponsors to address a variety of genomic challenges, ranging from sequencing microbial and eukaryotic genomes, to single cell genomics, to RNAseq and metagenomics. Depending on the specifics of the project, we utilize capillary Sanger sequencing, 454 pyrosequencing, Illumina (GA or HiSeq) sequencing, or more recently, PacBio single molecule sequencing. Since the installation of our PacBio RS instrument in March, we have worked to evaluate strategies to take advantage of the PacBio's unique technology and integrate this new data type to improve current multi-next-gen platform workflows. One of LANL's sequencing strengths is the closure of microbial genomes, a process for which we hope to evaluate and capitalize on methods including "strobe" sequencing and generation of long reads to both fill and scaffold gaps and repeats between contigs in sequence assemblies. Our first attempt at using PacBio long reads on repeat-rich genomes for the resolution of repetitive gaps suggests that this strategy may indeed help close gaps, although improvement of both chemistry and informatic processing will be required. In a DTRA-sponsored exercise designed to simulate a potential biothreat outbreak, we have also tested the capability of PacBio to help identify and characterize target pathogens present at low-levels within complex samples (air filter and blood). Our experience using PacBio to sequence these metagenomic samples suggests that the greatest advantage of the PacBio over other next-gen platforms is the speed with which sequence data can be produced to rapidly help identify target organisms. In order to make precise strain determinations of targets present at very low abundances within a sample with the PacBio, the trade-off in speed for throughput and readlength for accuracy may require optimization.

PacBio SequencingTechnology

DNA Sample Preparation



Single Molecule Real Time (SMRT) Sequencing Technology



Sequencing Strategies

Protocol	Template	Sequencing Output
Standard	•Large insert, 2kb •2-3 µg sample input	•Generates one pass on each molecule
Strobe	•Very large insert, 10kb •20 µg sample input	•Alternating periods of laser on/off generate distributed reads on each molecule
Circular Consensus	•Small insert, 250bp •1 µg sample input	•Generates multiple passes for each molecule

Biothreat Outbreak Simulation

DNA Samples

Low level spikes of unknown potential pathogens in complex backgrounds

Human blood sample + spike
Air filtrate + spike

Preparing 2kb Insert Libraries for Sequencing

Due to poor quality and low quantity of DNA, we processed whole genome amplification products in parallel with unamplified samples.

Standard Sequencing on PacBio RS

Blood Sample DNA
Unamplified 2 cells, 30 x 2 min.
Amplified 2 cells, 30 x 2 min.
2 cells, 45 x 2 min.

Air Filtrate DNA
Unamplified **Library Failed**
Amplified 2 cells, 30 x 2 min.
2 cells, 45 x 2 min.

Blood Sample Data, 2 chips Unamplified

	Pre-Filter	Post-Filter
# of Bases (bp)	203876568	7067101
# of ZMWs / # of Reads	300584	3521
Mean Readlength (bp)	28	1678
Mean Read Quality Score	0.014	0.801

Within 22 hours of sample receipt, presence of Hepatitis B virus in Blood Sample was identified by mapping Unamplified PacBio reads to NCBI Viral Genome Database .

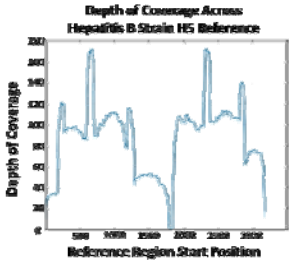
Blood Sample Data, 4 chips Amplified

	Pre-Filter	Post-Filter
# of Bases (bp)	856708296	556410211
# of ZMWs / # of Reads	526029	223266
Mean Readlength (bp)	875	1810 *(1528 / 2003)
Mean Read Quality Score * (30 x 2min./45 x 2min.)	0.392	0.818

Blood Sample Combined Data Mapped to Hepatitis B virus, strain H5 Reference Sequence

# Post-filter reads	226787
# Mapped bases (bp)	517182
Maximum mapped readlength (bp)	4658
# Mapped subreads	759
Mean mapped subread accuracy (%)	83.62
# Mapped reads	363
95th Percentile mapped readlength (bp)	3419
Mean mapped readlength (bp)	1425
Mean mapped subread readlength (bp)	412
Mean depth of coverage	88.56
Missing bases (%)	1.32

Combining unamplified and amplified Blood Sample data gave 88x coverage. Mapping to the Hepatitis B strain H5 reference identified 1 potential SNP.



Air Sample Data, 4 chips Amplified

	Pre-Filter	Post-Filter
# of Bases (bp)	777877885	504912538
# of ZMWs / # of Reads	601168	246681
Mean Readlength (bp)	723	1533 *(1392 / 1670)
Mean Read Quality Score	0.375	0.812

* (30 x 2min./45 x 2min.)

While PacBio data alone did not did not identify the *Francisella* strain present in extremely low abundance in the Air Sample, hits to multiple *Francisella* reference genomes confirmed the presence of *Francisella* as determined by the Illumina platform.

Combined Air Sample Data Mapped to 10 *Francisella* completed Chromosome/Plasmid Sequences

- 135 of filtered reads mapped
- Returned hits to 7 *Francisella* references:
  1. *Francisella philomiragia* subsp. *philomiragia* ATCC 25017 plasmid pFPHI01
  2. *Francisella philomiragia* subsp. *philomiragia* ATCC 25017
  3. *Francisella tularensis* subsp. *tularensis* FSC198
  4. *Francisella tularensis* subsp. *holarctica* FTNF002-00
  5. *Francisella tularensis* subsp. *mediasiatica* FSC147
  6. *Francisella tularensis* subsp. *novicida* U112
  7. *Francisella tularensis* subsp. *tularensis* WY96-3418

Closing Gaps with PacBio Long Reads

Test Genome:

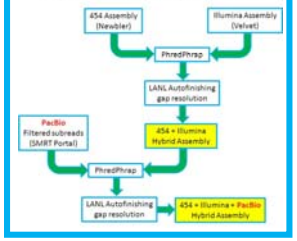
Yersinia pestis strain 1670 Georgian

- Genome size: ≈4.6 Mbp
- G + C content: 47%
- Rich in repeats and rearrangements
- Genomic DNA was amplified for library preparation

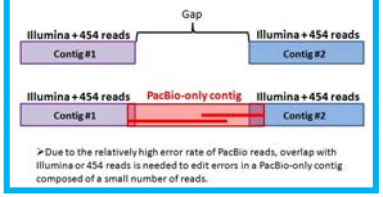
3 Sequence Datasets

1. **454 Standard**
   Average readlength: 387bp
   Average estimated genome coverage: 22x
2. **Illumina GAIIx**
   Readlength: 50bp
   Average estimated genome coverage: 92x
3. **PacBio RS**
   4 SMRT cells sequenced
   Average post-filter readlength: 1,778 bp
   Average post-filter quality: 0.870
   Average estimated genome coverage: 16x

Hybrid Assembly Method

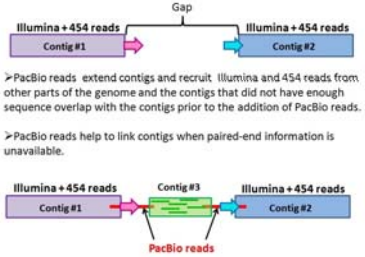


Overlapping reads from PacBio can close a gap



➢Due to the relatively high error rate of PacBio reads, overlap with Illumina or 454 reads is needed to edit errors in a PacBio-only contig composed of a small number of reads.

PacBio reads span repeats within a gap region



➢PacBio reads extend contigs and recruit Illumina and 454 reads from other parts of the genome and the contigs that did not have enough sequence overlap with the contigs prior to the addition of PacBio reads.

➢PacBio reads help to link contigs when paired-end information is unavailable.

PacBio reads yield a modest improvement in *Y. pestis* genome assembly

454 + Illumina		454 + Illumina + PacBio	
Contigs	133	Contigs	107
N50 (bp)	98168	N50 (bp)	102437
N90 (bp)	28664	N90 (bp)	29619
Max contig length (bp)	370999	Max contig length (bp)	370999
Min contig length (bp)	207	Min contig length (bp)	343
Total bases	4659423	Total bases	4645874