

# De Novo and Hybrid Assembly

## On the PacBio® RS

### Introduction

The PacBio® RS utilizes SMRT® technology to generate both Continuous Long Read (“CLR”) and Circular Consensus Read (“CCS”) data. In this document, we describe sequencing the *Escherichia coli* strain MG1655 (“*E. coli*”) on the PacBio RS system and evaluate SMRT Assembly approaches of combining PacBio CLR with either PacBio CCS or Illumina® MiSeq™ data. By leveraging the long read lengths of CLR and the high accuracy of CCS data generated by the PacBio RS, hybrid *de novo* assembly of *E. coli* improved significantly, with fewer resulting contigs and higher N50 values relative to assembly using short-read data alone.

### Definitions

**Continuous Long Read (“CLR”):** Sequence generated by splitting the raw sequence from a ZMW by the adapters. Also referred to interchangeably as *subread*. A *full pass subread* is a subread with two observed adjacent adapters

**Circular Consensus Sequencing (“CCS”):** Sequence generated by collapsing multiple CLRs or subreads from a single ZMW to form a single high-accuracy read. CCS data are generated when  $\geq 2$  full pass subreads are present.

**Double-Cut-Join (“DCJ”) Distance:** Metric representing the number of rearrangement events needed to transform the assembly structure to that of the reference genome. A lower DCJ distance represents a more accurate assembly.

**Locally Collinear Block (“LCB”):** Metric representing a subset of contig alignments that occur in the same order and orientation in the reference genome.

**PacBio corrected Read (“PBcR”):** The result of error correcting PacBio CLR data with high accuracy reads using either PacBio CCS or short-read (e.g., MiSeq) data.

**pacBioToCA:** This utility in Celera Assembler 7.0<sup>1</sup> aligns high accuracy reads to the CLRs, error correcting the CLRs when a minimum coverage is satisfied, and splitting or trimming the CLRs otherwise. Celera Assembler is an open source assembler shown to have good performance with data generated from a variety of sequencing technologies.

### SMRT® Sequencing of *E. coli*

*E. coli* libraries were made with 2 kb and 10 kb shears of gDNA, and libraries were prepared using the standard PacBio sample preparation methods with C2 chemistry specific to each insert size. The 10 kb sample was sequenced on 1 SMRT Cell with a 1X90 min collection protocol. The 2 kb sample was sequenced on 16 SMRT Cells with a 2X45 min collection protocol.

### SMRT® Analysis and Assembly Workflow

The data collected from the PacBio RS instrument were processed and filtered using the SMRT Analysis software suite. The 10 kb CLR data were filtered by read quality ( $>0.75$ ) and read length ( $>1000$  bp), resulting in approximately 98 Mb or 21X coverage of *E. coli*. When processing CLR data, raw reads from the SMRT Cell were split on adapter sequence resulting in  $\geq 1$  subread or CLR per ZMW.


 Experiment Design

Isolate DNA

Template Prep

Sequencing

Analysis

The 2 kb CCS data were filtered by readlength (>500 bp), resulting in 217 Mb or 47X coverage. CCS data is generated by merging the information from all ZMWs with  $\geq 2$  full pass subread per ZMW. Table 1 contains the detailed information about the CLR and CCS yields.

**Table 1: PacBio Post-Filtering Yields for 10 kb CLR and 2 kb CCS Data**

	10 kb CLR	2 kb CCS
# SMRT Cells	1	16
# Reads	50,765	231,629
Total Bases	98,213,822	217,871,193
Mean Readlength	1,934.68 bp	940.60 bp
Max Readlength	14,494 bp	2,627 bp
Coverage	21X	47X

For SMRT hybrid assembly, the filtered CLR data were output to FASTQ format and error corrected separately with either Illumina MiSeq 2X150bp paired-end data or PacBio 2 kb CCS data using the pacBioToCA utility in Celera Assembler. "PBcR" refers to these error corrected reads, and the subsequent designation in parentheses refers to the high accuracy data used for error correction (i.e. 2kb CCS or 2x150bp MiSeq). For comparison, the MiSeq reads were also assembled by themselves with Celera Assembler. The MiSeq data was downsampled to 52X coverage for both hybrid and short-read only assemblies.

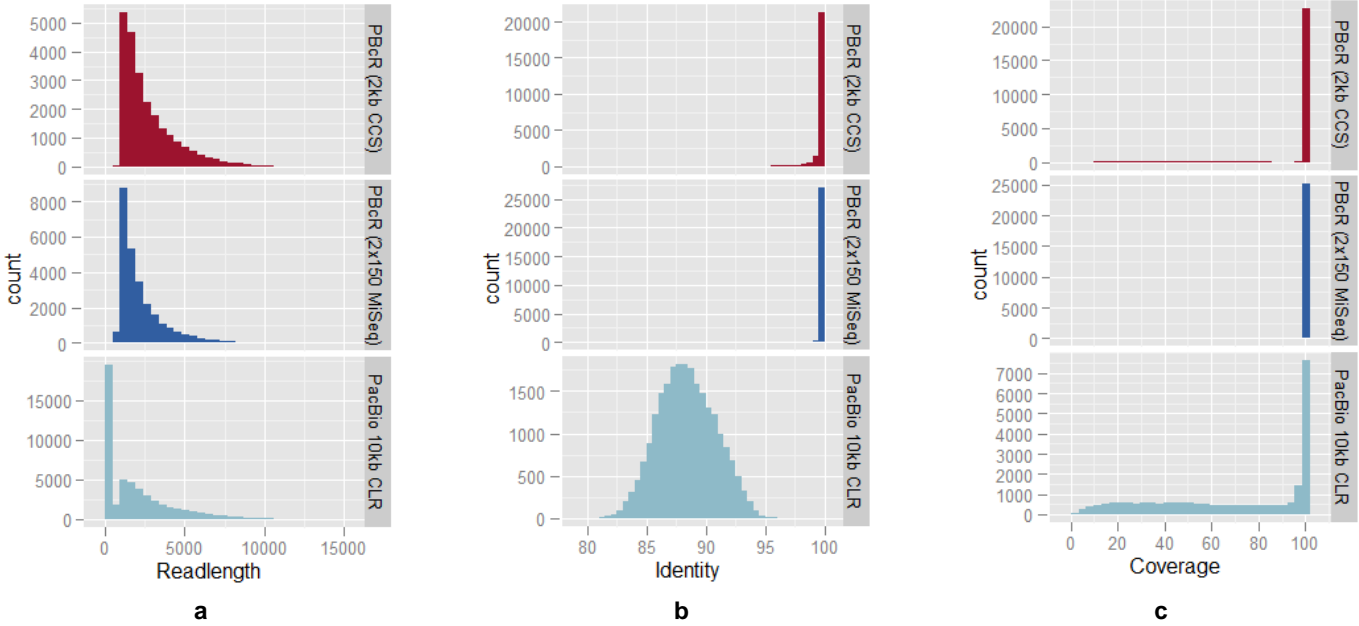
## Results

Both the MiSeq and PacBio CCS data can be used to correct the CLR data with an average 99% identity and coverage when aligned to the reference (see Table 2 and Figure 1). The PBcR (2kb CCS) however, have a higher yield and mean readlength when compared to the PBcR (2x150bp MiSeq) data and, therefore, offer greater error correction utility than the MiSeq short reads. The longer readlengths and lower sequencing bias of the PacBio CCS data result in better mappability and more even genomic coverage.

The error correction step lowers the yield by 35-40% but mean readlength and accuracy increase substantially when compared to the 10kb CLR data (Table 2 and Figure 1).

**Table 2: PBcR Yields After Error Correction of 10kb CLR Data with 2 kb CCS and 2x150 MiSeq Data**

	PBcR (2 kb CCS)	PBcR (2x150 bp MiSeq)
# of Reads	23,440	26,322
Total Bases	63,703,946	59,447,762
Mean Readlength	2,717.75 bp	2,258.48 bp
Max Readlength	11,519 bp	11,527 bp
Coverage	14X	13X



**Figure 1: (a) Readlength, (b) percent identity, and (c) percent coverage.. The top (red) charts were generated from the PBcR (2 kb CCS) data. The middle blue charts were generated from the PBcR (2x150 bp MiSeq) data. The bottom teal charts were generated from the pre-error corrected PacBio 10 kb CLR data. Nucmer 3.1 was used to generate the identity and coverage values.**

To evaluate the assemblies, the resulting contigs were aligned back to the reference with progressiveMauve<sup>2</sup>, and assembly metrics were generated using the Mauve Assembly Metrics<sup>3</sup> software. In addition to the standard number of contigs and N50 values, Mauve Assembly Metrics generates LCB and DCJ metrics. Fewer LCBs in an assembly, aligned to a known reference, correspond to a fewer number of rearrangements and errors in the assembly. Table 3 summarizes these results.

**Table 3: *E. coli* MG1655 Assembly and Hybrid Assembly Metrics**

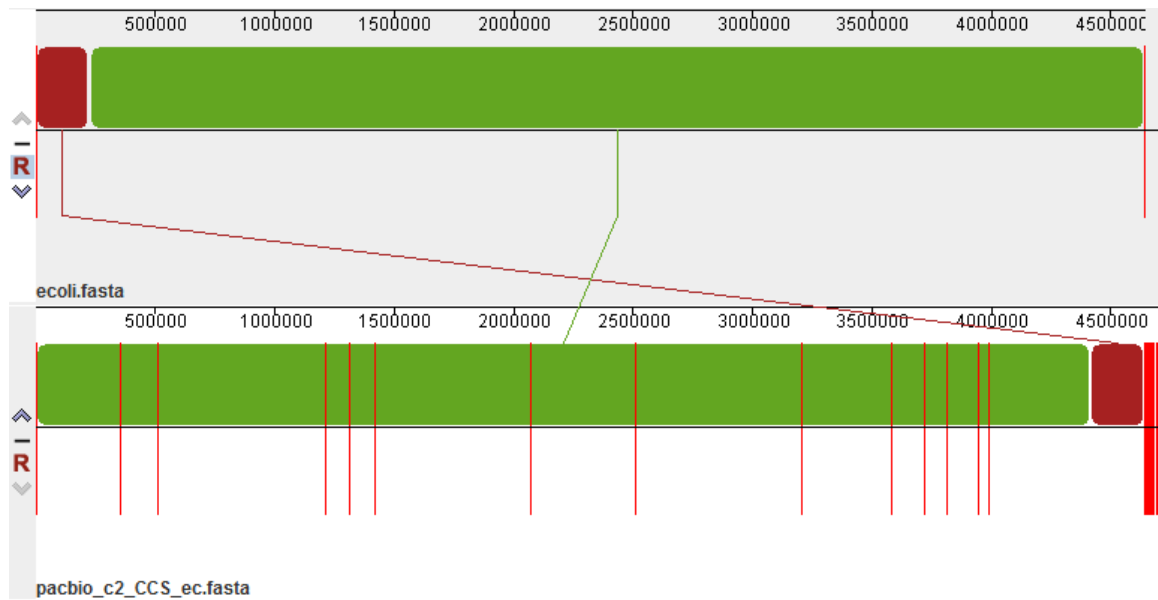
	Short-Read Only Assembly	PacBio Corrected Reads Assembly	
	52X 2x150bp MiSeq	14X PBcR (2kb CCS)	13X PBcR (2x150bp MiSeq)
Number of Contigs	129	30	51
N50	59,830	175,898	143,047
Max Contig Size	188,461	411,562	312,522
Number of LCBs	5	2	9
DCJ Distance	132	16	48
# Mis-calls	31	186	352
Total Bases Missed	284,602	24,642	48,605

The MiSeq only data showed expected results using Celera Assembler with 129 contigs and an N50 of 59,830. When Celera Assembler was given PBcR (2 kb CCS) and PBcR (2x150 bp MiSeq) data, the number of contigs were reduced to 30 and 51, respectively (while the N50 increased to 175,898 and 143,047). The max contig size also increased from 188,461 bp with the short-read only assembly to 411,562 bp with PBcR (2kb CCS) and 312,552 bp with PBcR (2x150 bp MiSeq) data.

Both types of error corrected reads gave a marked improvement to the assembly, reducing the number of contigs and increasing the N50 and max contig size. The Mauve alignments for the PBcR (2 kb CCS – see Figure 2) and PBcR (2x150 bp MiSeq – see Figure 3) assemblies show good coverage across the reference. However, the PBcR (2x150

bp MiSeq) assembly contains contigs that are more fragmented and misassembled. The PBcR (2 kb CCS) assembly yielded fewer contigs that could be arranged in two LCBs (the second LCB in red can be explained by the assembler choosing a different linearization site on the circular genome than the reference sequence). The *E. coli* reference was spanned by only 14 contigs.

Examination of mis-calls and missing bases show that the PBcR (2 kb CCS) assembly had the fewest missing bases of the three. The MiSeq only assembly yielded the fewest mis-calls but had an order of magnitude more missing bases than either of the PBcR assemblies.



**Figure 2: Alignment of *E. coli* reference (top) with PBcR (2 kb CCS) assembly (bottom). The red vertical lines denote contig barriers and the colored blocks represent the LCBs.**

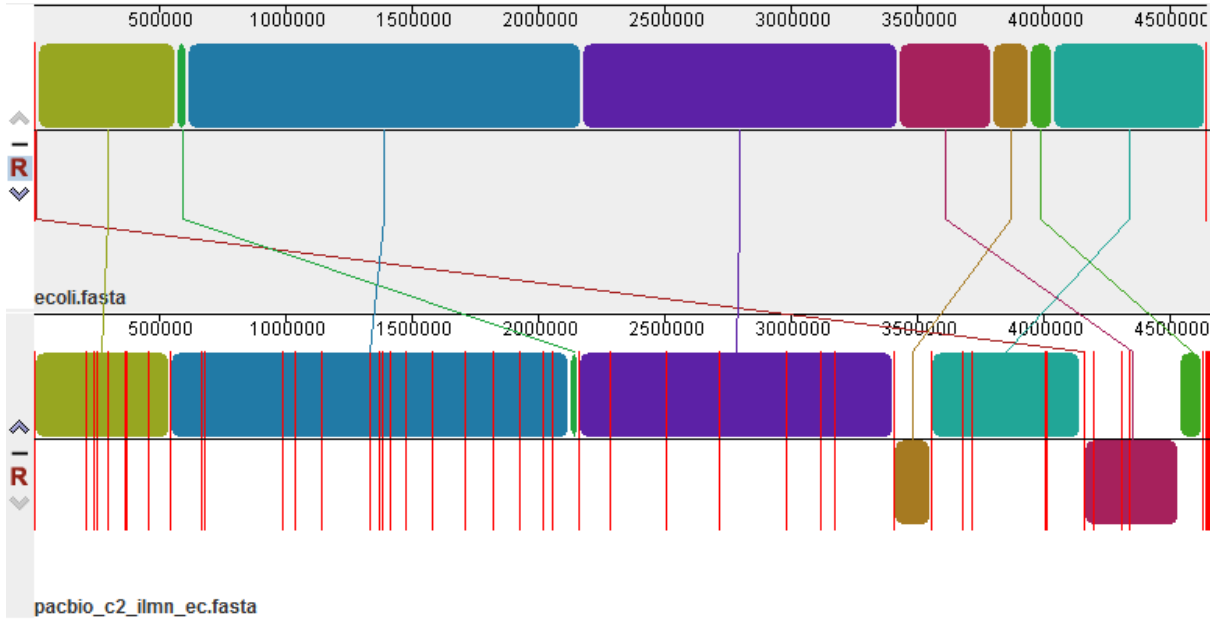


Figure 3: Alignment of *E. coli* reference (top) with PBcR (2x150 bp MiSeq) assembly (bottom). The red vertical lines denote contig barriers and the colored blocks represent the LCBs.

### Conclusion

Using the pacBioToCA utility in the Celera Assembler software package, we were able to error correct ~98Mb of PacBio 10 kb library CLR with both PacBio CCS and Illumina MiSeq reads. The resulting hybrid assemblies of error corrected PacBio reads have fewer contigs, higher N50, and higher max contig sizes than the MiSeq only assembly. When comparing the two hybrid assemblies, the PBcR (2 kb CCS) produced a better assembly than the PBcR (2x150 bp MiSeq) with equivalent coverage of high accuracy reads.

PacBio long reads have high utility in closing the gaps left by short-read only assemblies. With longer readlengths and higher yields expected with future upgrades to the PacBio RS, the addition of as little as one SMRT Cell of PacBio CLR to existing short-read data can improve assemblies significantly. Furthermore, error correcting CLR with PacBio CCS (instead MiSeq data) has the added benefit of longer readlengths and lower sequencing bias, resulting in higher quality, error corrected reads. In both situations, the PacBio RS is a cost-effective solution for finishing high quality genomes.

### References

1. <http://sourceforge.net/apps/mediawiki/wgs-assembler/>
2. Aaron E. Darling, Bob Mau, and Nicole T. Perna. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss, and Rearrangement. PLoS One. (2010) 5(6):e11147.
3. [http://code.google.com/p/ngopt/wiki/How\\_To\\_Score\\_Genome\\_Assemblies\\_with\\_Mauve](http://code.google.com/p/ngopt/wiki/How_To_Score_Genome_Assemblies_with_Mauve)

For Research Use Only. Not for use in diagnostic procedures. © Copyright 2012, Pacific Biosciences of California, Inc. All rights reserved. Information in this document is subject to change without notice. Pacific Biosciences assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of Pacific Biosciences products and/or third party products. Please refer to the applicable Pacific Biosciences Terms and Conditions of Sale and to the applicable license terms at <http://www.pacificbiosciences.com/licenses.html>. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT and SMRTbell are trademarks of Pacific Biosciences in the United States and/or certain other countries. All other trademarks are the sole property of their respective owners. PN 100-092-500-01