A Comparison of Microbial De Novo Assemblies Utilizing Long Read Sequencing Data

Lawrence Hon, Lawrence Lee, John Beaulaurier, Jason Chin, Aaron Klammer, Khai Luong, Jonas Korlach
Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025

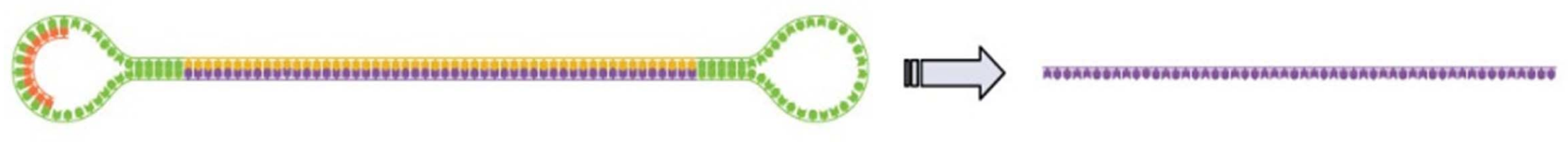
Introduction

Current *de novo* assembly tools have generally been designed to target shorter read data. The PacBio® *RS* platform is unique in that it can generate reads much longer than 1 kb and can be run in different ways to optimize for longer reads or lower error depending on the needs of the application. To provide the best performance on this type of data, assembler tools need to be modified to account for these characteristics. Here, we examine several assemblers, including ALLORA, Celera® Assembler, ALLPATHS-LG, and MIRA, which are able to handle PacBio data. In particular, we focused on the error correction hybrid assembly approach enabled by the PacBioToCA module within Celera Assembler. By analyzing several microbial data sets, we suggest how to incorporate PacBio data to achieve the best possible assemblies.

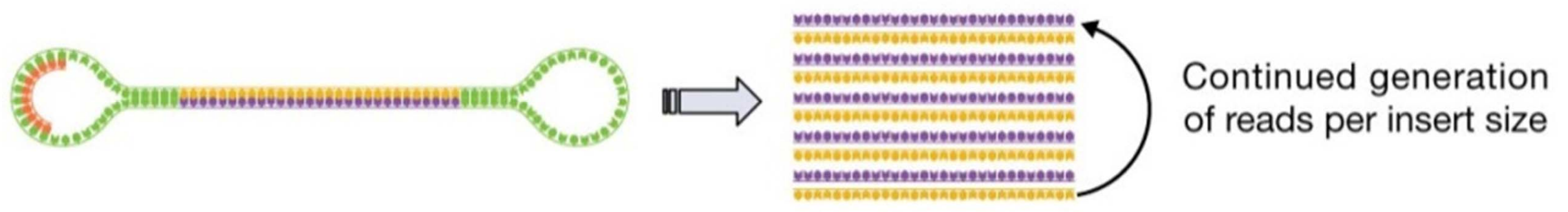
PacBio® Data Characteristics

Depending on the insert size of the library, the PacBio® *RS* can be optimized to generate longer sequences or shorter but higher quality sequences.

Continuous Long Reads (CLR). A large insert library (e.g. 6-10kb) results in long CLR reads up to 10 kb:



Circular Consensus Sequence (CCS). A short insert library (e.g. 500-1000 bp) favors multiple passes around each circular SMRTbell™ construct. The sequence generated by multiple observations of a single DNA molecule can be summarized as a higher quality (>99% accuracy) consensus sequence.



SMRT® Assembly Overview

Assembly methods incorporating PacBio single molecule, real time (SMRT®) data can be generalized into three categories (Table 1):

- *SMRT de novo*: assembly of PacBio CLR reads only
- *SMRT Hybrid*: hybrid *de novo* assembly of PacBio CLR and a second high accuracy data type (either PacBio CCS reads or second generation short-read data)
- *SMRT Scaffolding*: use PacBio CLR to scaffold existing contigs generated from short-read data

Table 1. Assembly algorithms that can incorporate PacBio data.

Description	SMRT de novo	SMRT Hybrid	SMRT Scaffolding
AHA (SMRT Analysis). Assemble short reads into high-confidence contigs and scaffold with PacBio CLR.			✓
ALLORA (SMRT Analysis). Assemble PacBio CLR and short read or CCS data. The P_ErrorCorrection module has to be run manually to error correct the CLR reads prior to assembly with ALLORA.	✓	✓	
ALLPATHS-LG. Error correct and scaffold PacBio CLR in a multistage process using different types of short read data, optimized for single node high-memory computation.		✓	
Celera Assembler (via PacBioToCA). Error correct PacBio CLR with accurate short reads and assemble, optimized for cluster computation.		✓	
MIRA. Assemble error corrected PacBio CLR generated by another error correction pipeline, e.g. Celera Assembler.		✓	

Results

The SMRT® Hybrid approach performs better than a short read only assembly on E. coli strain MG1655. Short read data is often limited by GC bias and an inability to span larger repeats. To overcome these limitations, it is possible to incorporate PacBio long read data (CLR) to perform SMRT Hybrid assembly. We used Celera® Assembler to *de novo* assemble three combinations of data. When PacBio CLR data was combined with CCS or MiSeq® data, we first ran the error correction step from Celera ® Assembler (PacBioToCA) to generate PacBio corrected reads (PBcR) (see Table 2), followed by running Celera® Assembler proper. The results show that hybrid assembly of PacBio CLR and CCS reads gave the best assembly metrics (see Table 3, Figure 1).

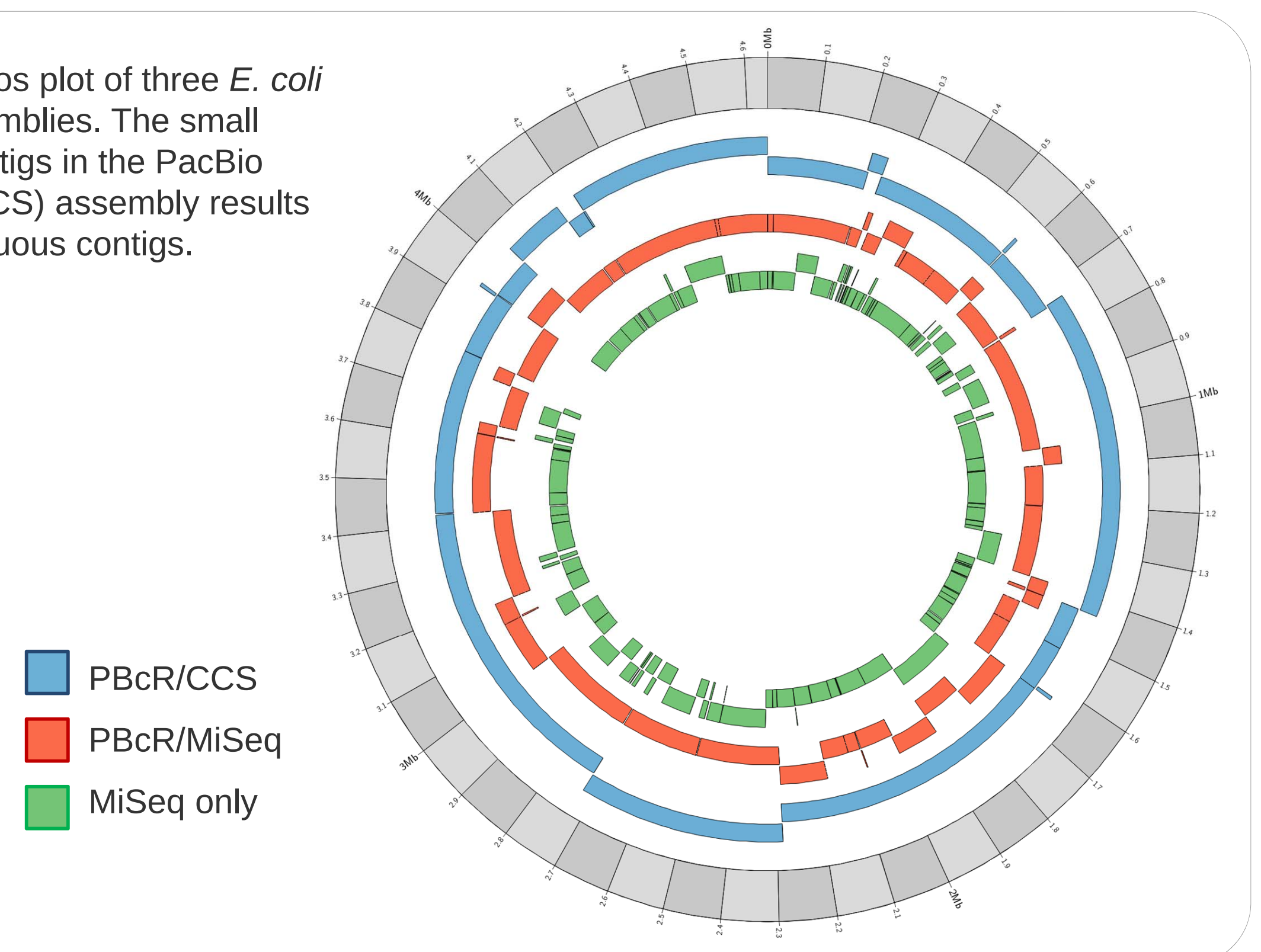
	Before Error Correction	After Error Correction	
Data Used	PacBio 10kb CLR	PBcR (CCS)	PBcR (MiSeq)
Number of Reads	50,765	23,440	26,322
Total Bases	98,213,822	63,703,946	59,447,762
Mean Readlength	1934.68 bp	2717.75 bp	2258.48 bp
Max Readlength	14,494 bp	11,519 bp	11,527 bp
Coverage	**21X**	**14X**	**13X**

Table 2. PacBio corrected read (PBcR) yields after error correction of 10 kb CLR reads with 2 kb CCS and 50X 2x150 MiSeq data

	Short Read Only Assembly	PacBio Corrected Reads (PBcR) Assemblies	
Data Used	52X 2x150 MiSeq	14X PBcR (43X CCS)	13X PBcR (MiSeq)
Number of Contigs	129	**30**	51
N50	59,830	**175,898**	143,047
Max Contig Size	188,461	**411,562**	312,522

Table 3. *E. coli* MG1655 Assembly and Hybrid Assembly Metrics. The best values for a given metric are in bold.

Figure 1. Circos plot of three *E. coli* MG1655 assemblies. The small number of contigs in the PacBio only (PBcR/CCS) assembly results in long, continuous contigs.



Microbial genomes are readily de novo assembled using the SMRT® Hybrid assembly approach. We obtained data for two additional strains of *E. coli*, comprised of PacBio CLR and CCS data. The assemblies generated by Celera® Assembler with error correction had favorable assembly metrics (see Table 4).

	E. Coli C227 German Outbreak strain	*E. Coli* strain #2
Data used	27X PBcR (61X CCS)	54X PBcR (24X CCS)
Libraries used	8 kb for CLR data 800 bp for CCS data	8-10 kb for CLR data 800 bp for CCS data
Sample source	Rasko et al.	NEB
Number of contigs	44	12
Number of contigs >10 kb	28	10
N50	911,711	1,208,623
Max Contig size	1,075,876	1,326,347

Table 4. Assembly metrics for two microbial genomes using PacBio SMRT hybrid assembly.

SMRT® Hybrid assembly can be run using a single library, yielding high quality assemblies. Sample preparation can be a laborious step during DNA sequencing. By limiting to a single insert size, this can significantly reduce the time and expense of sample preparation. Here, we used data from 16 SMRT Cells of *E. coli* strain MG1655 based on a single 2 kb library, and extracted both CLR and CCS reads (see Figure 2). These two datasets were then used in SMRT hybrid assembly via Celera® Assembler, yielding high quality assemblies, with N50s as high as 600 kb and maximum contig size over one megabase (see Figure 3).

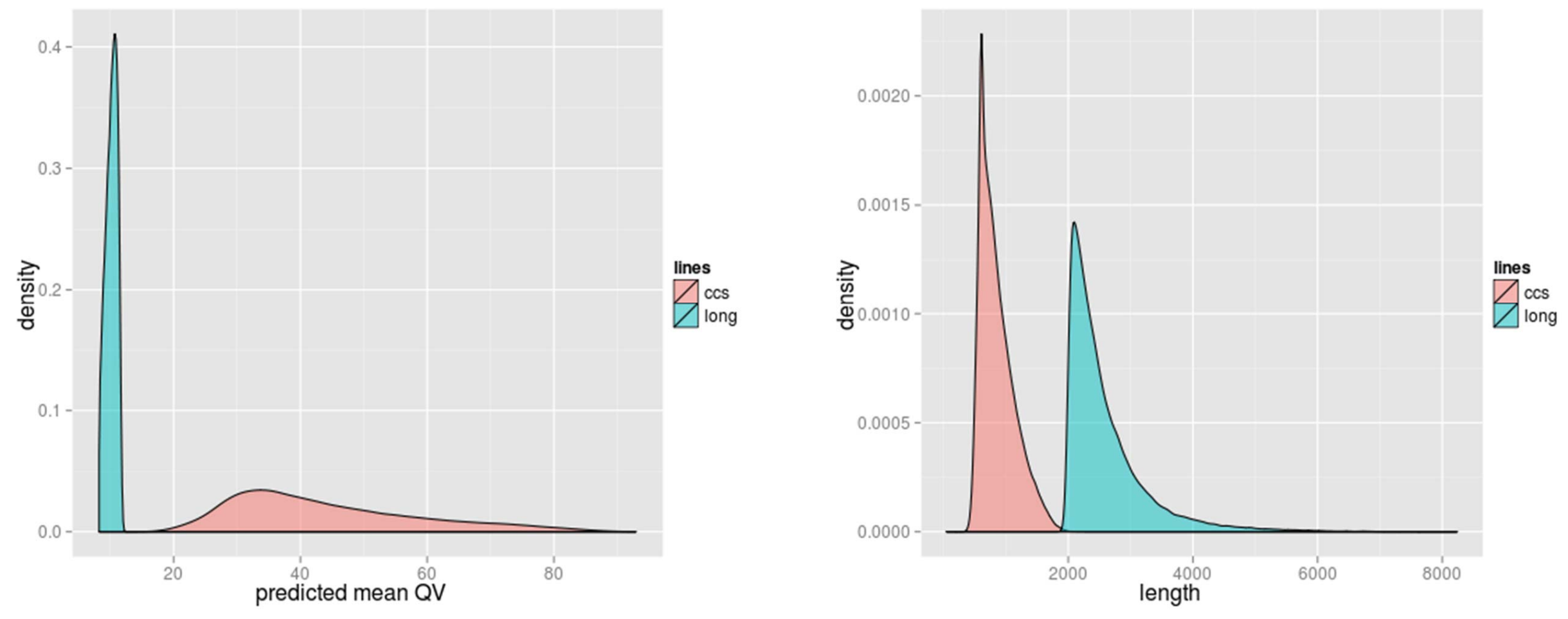


Figure 2. Predicted mean QV and length distributions of both CLR and CCS reads post-filtering (>0.85 mean QV, >2000 bp, and longest subread per ZMW for CLR reads; 3 pass minimum for CCS reads). CLR and CCS reads can be extracted from the same library, with substantially different read characteristics that are useful for SMRT Hybrid assembly.

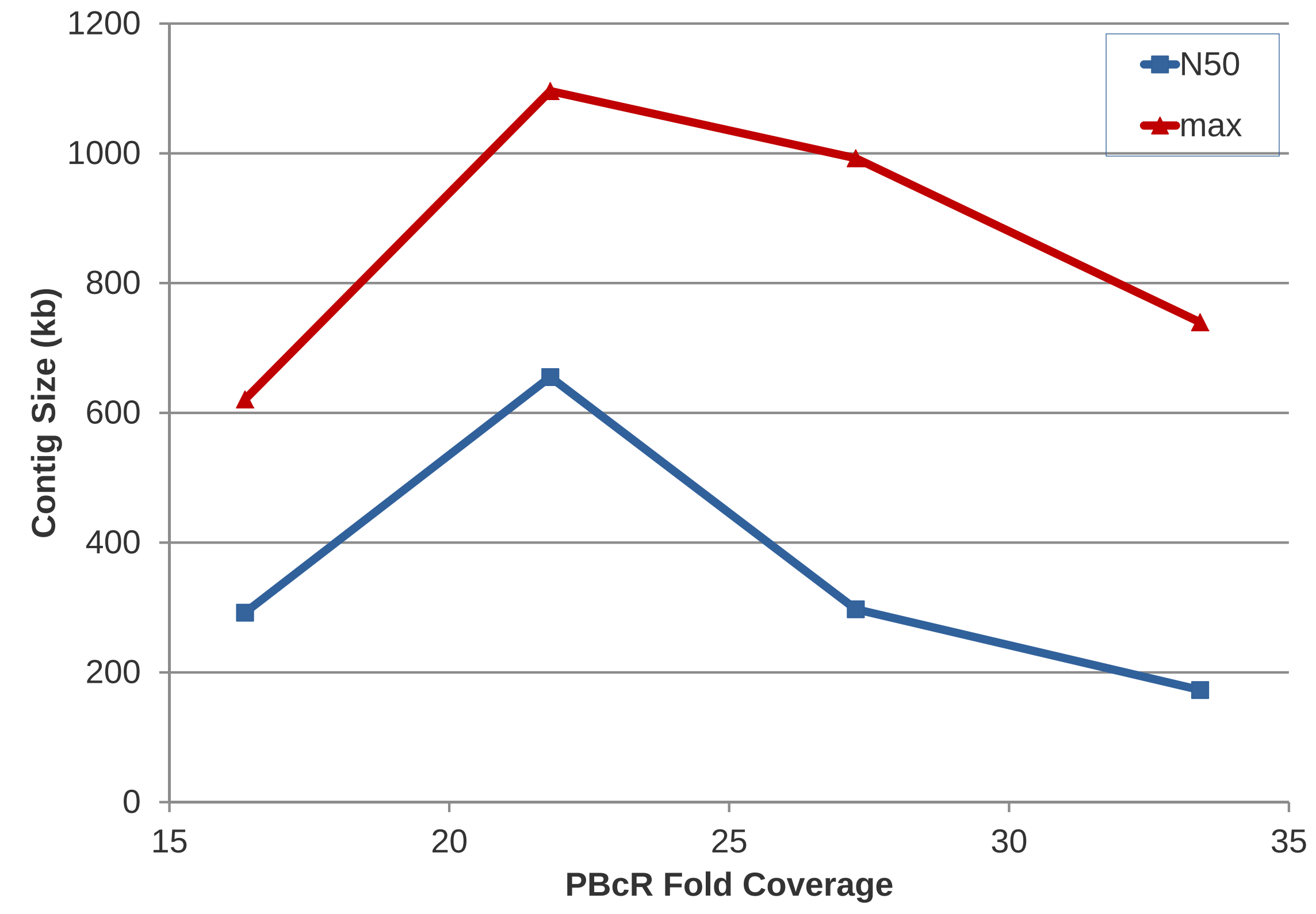


Figure 3. Assembly metrics vary depending on PBcR fold coverage of data passed to Celera Assembler. Data was downsampled to coverage levels between 15-35X, keeping the longest reads. Celera Assembler performed best in the 20-25X coverage range, possibly because it is sensitive to errors introduced by the additional coverage.

SMRT® hybrid assembly quality can be greatly affected by data filtering strategies. We have found the following to improve assembly quality when using Celera® Assembler:

- Choosing 25-50X coverage filtering for the longest reads and highest QV
- Requiring at least three passes if using PacBio CCS reads
- Using 20-25X coverage of PacBio corrected reads, filtering using the longest sequences (see Figure 2)
- Including more data, as higher coverage allows more stringent filtering
- Constructing larger insert libraries to span repeats and other complex regions

References

Rasko D, *et al*. "Origins of the E. coli Strain Causing an Outbreak of Hemolytic–Uremic Syndrome in Germany." *N Engl J Med* 2011; 365:709-717

Koren S, *et al*. "Hybrid error correction and de novo assembly of single-molecule sequencing reads." 2012, (Under review).

http://sourceforge.net/apps/mediawiki/wgs-assembler/index.php?title=PacBioToCA

Acknowledgements

We would like to thank Richard Roberts and Iain Murray at NEB for providing the *E. coli* sample.