

Microbial Pathogen Identification and Characterization with the PacBio *RS* System

Overview

The Pacific Biosciences® *RS* system (“PacBio *RS*”) is a third generation sequencer that provides the unique capability of providing timely results in identifying, characterizing, and ultimately performing surveillance of pathogen outbreaks. Sequence information that provides unparalleled insights into pathogen biology can be available within hours of DNA isolation. This is possible due to rapid, amplification-free sample preparation coupled with short run times, accurate single-molecule resolution, and extremely long reads that enable superior assemblies. These features distinguish the PacBio® technology for research areas such as clinical research, agricultural biotechnology, food safety, and biodefense. In all of these areas, it is important to rapidly generate the most accurate and complete data possible in order to fully characterize a pathogenic organism. Key benefits of the system include:

- **Extremely long read lengths** which allow for characterization of large structural variations as well as more complete full-genome assemblies. This structural knowledge leads to improved *functional* understanding.
- **High accuracy** combined with long read length enables quality *de novo* assemblies.
- **Balanced coverage** and **minimal GC-bias** enable quality assembly on any organism and sequencing through extended repeat regions.
- **Fast time to result** allows for rapid movement to analysis and reaction to observations.
- **More granular sequencing runs** allow for rapid changes to an experimental design based on early sequencing results.
- **Kinetics information** can be used to study strand-specific base modifications, including those potentially related to virulence¹.
- **Simplified library preparation** that does not require amplification of sample nucleic acids.

In this paper, we describe several applications of the PacBio *RS* for microbial pathogen studies. As examples, we explain how a high quality draft of a microbial genome was achieved using standard library preparation protocols consisting of a short- and a long-insert library. We also discuss whole-genome sequencing data, from the PacBio *RS*, and how associated bioinformatics approaches were used for microbial-pathogen characterization.

Background

A typical microbial genome ranges from two to five megabases. A typical SMRT™ Cell from the PacBio *RS* generates 90 Mb of sequencing data from 35,000 reads. The system will cover a microbial genome with extremely long reads -- 2700 bp on average, with 5% of the reads above 6000 bp. Additionally, the coverage is evenly distributed across a genome, so less coverage is necessary to produce a high quality genome assembly. Due to coverage bias in short read technologies, a genome generally needs to be covered at 100X in order to ensure that every subregion has adequate coverage².

Core-Genome Phylogeny Analysis

The first example of microbial research using the PacBio *RS* was to study the evolutionary relationships between different closely related cholera strains that had not been previously sequenced. Such analyses could be done at different levels, for example using a well-known set of SNP sites to build phylogenetic relationships³ or using a set of genes in core-genomes^{4,5,6}. As demonstrated in two New England Journal of Medicine articles, evolutionary relationships can be discovered with a day of sequencing per strain. The long read lengths of the system enable superior full-genome assemblies that help to distinguish closely-related strains.

The bioinformatics work for this type of application generally involves comparative genomic analysis with the data from SNP calls and consensus sequences. The comparative genomic analysis can be also done by directly aligning the long reads to a database, as was done in our work to identify the origin of the Haitian cholera outbreak.

With the PacBio *RS*, it was easy to identify genetic variation such as SNPs, various structural variations, and missing elements of the outbreak strain as compared to known strains. As a result, we were able to quickly characterize the outbreak strain. Additionally, with the extremely long read lengths of the SMRT™ (“Single Molecule Real Time”) sequencing technology, it is possible to characterize more than just SNPs – long insertions and deletions were also identified.

Comparing Closely Related Microbial Strains to Characterize a New Strain

New systems like the PacBio *RS* are capable of generating better assemblies and have spurred new tool development. Long read lengths allow identification of closely-related strains to be performed by simple BLAST reads to a sequence database without any computationally intensive sequence-assembly procedures. Additionally, the PacBio *RS* does not exhibit the coverage bias seen in short-read technologies. It is also able to characterize high-GC or low-GC organisms much more effectively than other sequencing techniques which continue to struggle with GC-bias.

As part of a study to characterize the German *E. coli* outbreak from the summer of 2011, we sequenced three different strains (C227-11, 55989, O42)⁶. The intent was to determine the known strain most closely related to the outbreak strain C227-11, and to subsequently evaluate potential explanations for the new strain’s increased virulence⁵. We took approximately 1000 reads from each and compared them to the NCBI NT database using BLAST. We then used the MEGAN⁷ software to analyze the returned BLAST hits. The results are shown in Figure 1(a),(b),(c).

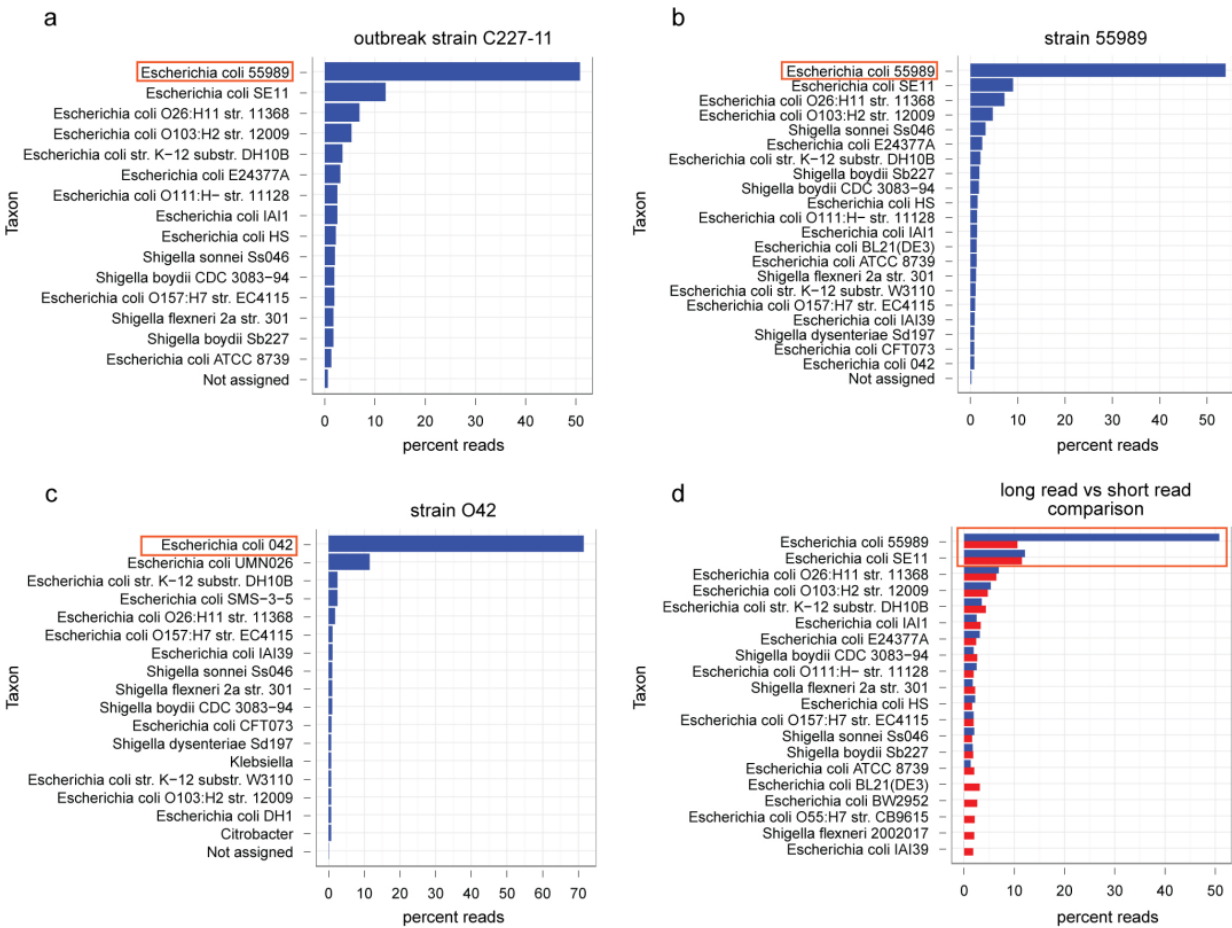


Figure 1: Results of BLAST comparison to NCBI database. (a) Outbreak strain C227-11 is most similar to strain 55989 (b) the 55989 control sequence is most similar to itself as expected (c) the O42 control sequence is most similar to itself as expected, and (d) Comparison of C227-11 long read results (blue) to short read results (red). Short read data⁸ was unable to determine whether the outbreak strain is more similar to strain 55989 or strain SE11. Long read data clearly shows that the outbreak strain is most similar to 55989. Short read data from Ty-2482, a different isolate of strain C227-11.

As expected, the control samples (55989, Figure 1(b) and O42, Figure 1(c) show that approximately 50% of the sequence reads have long blast hits to the 55989 and O42 genomic sequences.

As shown in Figure 1(a), we found that C227-11 was most closely related to the known 55989 strain (longest bar), and that the known strain SE11 was less closely related (second longest bar). With this initial characterization, we proceeded to align the data from the outbreak strain to the 55989 reference sequence to see how well the sequence reads from C227-11 agreed with the 55989 reference sequence.

Read length was a very important factor in obtaining enough specificity for such simple database searches. Longer reads are able to cover more SNP sites or structural variations at once, so we can better resolve homologous regions between two closely related strains. In contrast, if the read length is short, most of the reads might not capture those variants that differentiate the two closely related strains.

To evaluate this hypothesis, we extended our experiment by comparing publicly available short-read data against the NCBI NT database using BLAST. We then compared the short-read species matches with the PacBio *RS* long-read species matches. The results are shown in Figure 1(d), and clearly illustrate the importance of the long reads generated by the PacBio *RS* sequencer as well as the inadequacies of short-read technology. In the short read results, approximately the same percentages of reads were hit for both the 55989 and the SE11 strains. The differences between the SE11 and 55989 strains are identifiable only over longer regions.

By contrast, the short 100 bp reads aligned to SE11 and 55989 equally well. Since the short reads do not provide the proper characterization to resolve which is the closest related strain using our simple BLAST method, it would be necessary to assemble the short-read data to increase specificity which is slow and computationally intensive.

As another example, a BLAST query of extreme long reads can also be submitted to determine whether useful or interesting information can be revealed. Figure 2 shows the BLAST results of a 9 kb read from the NCBI BLAST website. We see that single reads give the most correct and closely related strains in the database. More importantly, this exemplary 9 kb read covers interesting structural variations between different strains (such as large insertions and deletions). Without long reads spanning at least 1500 base pairs, these regions would become very difficult or impossible to identify.



sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	△ E value	Max ident
U928145.2	Escherichia coli 55989 chromosome, complete genome	6447	1.464e+04	99%	0.0	89%
P009240.1	Escherichia coli SE11 DNA, complete genome	6262	1.123e+04	99%	0.0	89%
U928160.2	Escherichia coli IAI1 chromosome, complete genome	6032	1.103e+04	99%	0.0	89%
P002516.1	Escherichia coli KO11, complete genome	3838	9660	99%	0.0	89%
P002185.1	Escherichia coli W, complete genome	3838	9650	99%	0.0	89%
P010958.1	Escherichia coli O103:H2 str. 12009 DNA, complete genome	3836	1.525e+04	99%	0.0	87%
P000038.1	Shigella sonnei Ss046, complete genome	3771	5560	85%	0.0	89%
P001383.1	Shigella flexneri 2002017, complete genome	3757	5427	85%	0.0	89%
E005674.1	Shigella flexneri 2a str. 301, complete genome	3757	5427	85%	0.0	89%
E014073.1	Shigella flexneri 2a str. 2457T, complete genome	3757	5427	85%	0.0	89%
P000266.1	Shigella flexneri 5 str. 8401, complete genome	3748	5418	85%	0.0	89%

Figure 2: BLAST Results of a Single 9 kb showing that a single long read will differentiate between closely related strains.

Identifying Virulence Factors

We applied the same methods of aligning the long SMRT reads to identify virulence factors in the outbreak strain of *E. coli*. In this case, we compared the sequence databases of virulence factors, such as MvirDB (<http://mvirdb.lnl.gov/>) and VFDB (<http://www.mgc.ac.cn/VFs/main.htm>). This allowed us to detect new virulence factors in the sample that were absent in the reference strains. When we aligned our reads from the C227-11 outbreak strain to the VFDB, we found significant alignments to the 960 bp Shiga toxin 2A subunit. Some of the matches covered almost the entire full Shiga toxin 2A, which allowed us to confirm the existence of the Shiga toxin gene in the C227-11 outbreak strains.

Again, this was done by aligning all reads to a sequence database using standard alignment tools, without performing any computationally intensive assembly work. In combination with the rapid sequence-to-results, the rapid analysis enabled quick adjustments to an experimental design by reacting to information as it was discovered. This is an example of a simplified and accelerated method of evaluating pathogen outbreaks.

Assembly of the Bacterial Genome

While a comparative genomic approach can reveal rich information about a sample, a full genome *de novo* assembly provides a more comprehensive view of the organism. With the same continuous long-read data used for comparative genomic analysis, we were able to quickly obtain draft assembly information and to finish the whole genome.

Such assembly can be done in multiple ways:

1. Pure PacBio long read *de novo* assembly,
2. A combined *de novo* assembly of high-accuracy single-molecule PacBio Circular Consensus Sequencing (CCS) reads with PacBio long reads, or
3. A combined assembly of PacBio data with high-throughput short read sequencing for a hybrid *de novo* assembly.

The sequencing of the *E. coli* outbreak strain is an excellent example of the ability to rapidly produce very high quality *de novo* genome sequence using the PacBio RS. Table 1 summarizes examples of resequencing a known strain (55989) as well as a *de novo* sequence of the outbreak strain C227-11 using a combined assembly of long reads and CCS⁶. In both cases, very high quality consensus accuracy was obtained.

Strain	Number of post-filter reads	Single molecule raw % accuracy mean (mode)	Coverage	Mean mapped read length	95th percentile mapped read length	Consensus accuracy (%)
55989	186,794	84.4 (88.3)	63X	2572	6713	99.999
C227-11	648,177	85.0 (88.2)	190X	2900	7811	99.97
C227-11 CCS	419,589	97.8 (99.9)	35X	3076*	8527	(de novo)

Table 1. Examples of high accuracy genome assemblies for both known (55989) and unknown (C227-11) strains of bacteria. Strain 55989 was assembled against a reference strain 55989. Strain C227-11 was assembled *de novo*. CCS mean mapped read length is for the total distance traveled by the DNA polymerase during sequencing*. Each CCS template is ~500 bp long, so the average circular template was sequenced ~6.1 times, leading to the high single-molecule accuracy. The consensus accuracy for the *de novo* combined long-read and CCS read data is shown for strain C227-11. The consensus accuracy for strain 55989 is for long reads mapped to a reference.

Data adapted from Rasko et al, Origins of the *E. coli* Strain Causing an Outbreak of Hemolytic–Uremic Syndrome in Germany⁶.

Below is the outline of the method used to perform the combined long-read and CCS-read *de novo* genome sequence for the *E. coli* outbreak strain. Normal sequencing runs, with each technique, are first combined to improve the accuracy of the long reads by aligning them with the CCS reads. These “corrected” long reads are then aligned to each other, forming a small number of contigs. These contigs can then be re-aligned to the CCS reads, forming a final high-accuracy assembly. The technology also provides even coverage across the genome. Thus, the benefits of the superior structural information in the long reads can be combined with the benefit of the shorter, but high-accuracy CCS reads.

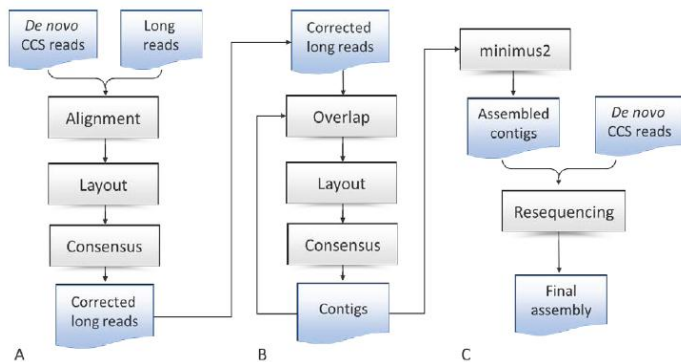


Figure 3. De novo genome assembly using a combination of long reads and high-accuracy CCS reads on a PacBio RS.

- A) The first step of the assembly process uses highly accurate single-molecule CCS reads to correct errors in the single pass long sequence reads
- B) The corrected long reads are then provided as input to ALLORA, an iterative overlap-layout-consensus *de novo* assembly algorithm.
- C) The resulting assembled contigs are then polished with the resequencing pipeline and the original CCS reads, producing the final, high-accuracy assembly.

Summary

The PacBio sequencing system provides unprecedented power to study pathogens and other microbes. We have described the first few examples of the power of rapid, extremely long-read sequencing technology to assist in characterizing several recent pathogen outbreak strains, helping to identify the source of outbreaks as well as sources of virulence.

The high-accuracy assemblies, possible with the PacBio RS, help clarify situations that are challenging to interpret with short-read sequencing technology, such as the identification of the *E. coli* outbreak strain described herein.

Moreover, SMRT® sequencing creates improved genome assemblies: a highly accurate *de novo* sequence was constructed. This type of assembly would typically require a variety of special sample preparation and analysis techniques with other sequencing technologies in order to overcome the inherent limitations of short-read data. However, long-read technologies enable a rapid expansion of *de novo* genome sequencing for pathogens and this capability is necessary to fully understand structural variation even in previously sequenced organisms.

The expansion of relatively easy-and-quick *de novo* genome sequencing will lead us to a deeper understanding of a rich variety of micro-organisms that have been difficult to address with historical techniques.

Bibliography

1. Ratel D, Ravanat J, Berger F, Wion D. N6-methyladenine: the other methylated base of DNA. *BioEssays* 28:309-315 (2006).
2. Gnerre et al, High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* 2011; 106: 1513-1518.
3. Lam C, Octavia S, Reeves P, Wang L, Lan R. Evolution of seventh cholera pandemic and origin of 1991 epidemic, Latin America. *Emerg Infect Dis*;16:1130-2.
4. Chun J, Grim CJ, Hasan NA, et al. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proc Natl Acad Sci USA* 2009;106:15442-7.
5. Chin, et al., The Origin of the Haitian Cholera Outbreak Strain, *The New England Journal of Medicine* 10.1056/NEJMoa1012928.
6. Rasko et al, Origins of the *E. coli* Strain Causing an Outbreak of Hemolytic–Uremic Syndrome in Germany, *The New England Journal of Medicine* 10.1056/NEJMoa1106920.
7. <http://ab.inf.uni-tuebingen.de/software/megan/> D.H. Huson, A.F. Auch, Ji Qi and S.C. Schuster, MEGAN Analysis of Metagenomic Data, *Genome Research*. 17:377-386, 2007.
8. ftp://ftp.genomics.org.cn/pub/Ecoli_TY-2482.

Pacific Biosciences' *E. coli* sequencing data is available with free registration on the web: <http://www.pacbioenvnet.com/share/datasets/ecolioutbreak>