Improved Yield and Diverse Finished Bacterial Genomes using Pacific Biosciences RS II SMRT Sequencing

Lisa D. Sadzewicz, Naomi Sengamalay, Xinyue Liu, Sushma Nagaraj, Qi Su, Ivette Santana-Cruz, Alvaro Godinez, Luke J. Tallon

Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD

Abstract

Recent studies have shown that SMRT sequencing by Pacific Biosciences is a rapid, effective, and highly accurate platform for generation of complete microbial genome sequences. As early-adopters of the RS II sequencer upgrade, we conducted an extensive and broad comparison to evaluate the new platform and chemistries for simultaneous generation of complete or nearly complete microbial genome sequences and analysis of epigenetic base modifications. Comparing more than 120 bacterial genomes from more than 16 species ranging in genome size from 900 Kbp to 7 Mbp and in GC-content from 30.2% - 64.3%, we generated complete genome sequences at twice the rate for isolates sequenced on RS II compared to isolates sequenced on RS. Overall, when combined with longer insert libraries and rigid size-selection using the BluePippin by Sage Science, the RS II upgrade yielded an increase in mean read length and tripling of total per-SMRTcell yield. This significant increase in read length and throughput has enabled more rapid and efficient generation of finished microbial genomes and has rendered this approach the def fact standard for small genome sequencing in our center. Further, using comparative Illumina sequencing, we found a median of one putative consensus basecall error per finished genome. Here, we present our experiences with RS II sequencing, a comparison of SMRT sequencing based generation of complete genomes of diverse microbial species using RS and RS II, and a comparison of available genome assemblers for these data.

Discussion

The improvement in both sequencing and assembly results using RS II is significant. Comparing PacBio bacterial genome sequencing data metrics from our last six months (11/2012 – 4/2013) using RS with our first nine months using RS II, we achieved more than a doubling in passed filter reads per SMRT Cell, a nearly 200% increase in total base pair yield, and a 50% increase in subread lengths. Increased polymerase read length and longer insert libraries enabled by BluePippin size selection both contributed to the observed increased subread lengths.

Genome assemblies using RS II data demonstrated significant improvements as well. We compared bacterial genome assembly metrics from the same time periods as above. Each genome was assembled both with CAT 0.7 and HGAP and the best assembly was selected. Of the 56 isolates sequenced on RS, 15 genomes (27%) assembled into complete genomes. The rate of genomes assembling into complete sequences on RS II was increased to more than 60% (39 of 54 isolates). The RS II data also yielded increased contig N50s with an average N50 equal to 98% of genome size compared to an average N50 equal to 65% of genome size for RS.

In addition to comparing sequence and assembly metrics between RS and RS II data, we evaluated three genome assemblers (CAT 0.7, HGAP, and HGAP2) using a subset of 14 bacterial genomes with a range of sizes and GC content. Based upon contig count and N50, each assembler produced the best assembly for some of the isolates. Though these metrics can be limited in their utility, they provide a reasonable assessment of assembler performance on aggregate. Overall, CAT 0.7 produced the largest number of complete genome assemblies, while HGAP2 generated the lowest mean contig count and longest mean contig N50.

In our evaluation of assembly consensus quality, we found an average of 3 passed-filter (PF) SNPs and a median of 1 PF SNP per genome for both RS and RS II sequenced genomes. When taking genome size into account, we find just over 1 PF SNP per million bases of genome sequence. Validations of these discrepancies are underway to determine which are PacBio consensus errors and which are due to Illumina sequencing or alignment errors. In general, these initial data indicate that bacterial genomes assembled using PacBio data alone generate highly accurate consensus sequences.

Ongoing studies are extending these comparisons to larger genomes and metagenomes, new sequencing chemistries and run lengths, and new assembly methods.

References


Acknowledgements

This project has been funded in whole or in part with federal funds from the National Institutes of Health (NIH) under contract HHSN261200900010C. We would like to thank the Department of Health and Human Services under contract number HHSN261200900010C.

Improved Total Yield & Read Length

<table>
<thead>
<tr>
<th>Genome Size (Mb)</th>
<th>GC%</th>
<th>Species</th>
<th>Total Reads</th>
<th>Mean Reads</th>
<th>Median Contigs</th>
<th>Mean N50</th>
<th>Mean N95</th>
<th>Complete Genomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.20</td>
<td>50%</td>
<td>Brachyspira aalborgi</td>
<td>2,000,000</td>
<td>2,000</td>
<td>120</td>
<td>1,500</td>
<td>1,000</td>
<td>1</td>
</tr>
<tr>
<td>2.96</td>
<td>60%</td>
<td>Brachyspira aalborgi</td>
<td>5,000,000</td>
<td>120</td>
<td>2,000</td>
<td>1,500</td>
<td>1,000</td>
<td>1</td>
</tr>
<tr>
<td>3.75</td>
<td>55%</td>
<td>Escherichia coli</td>
<td>2,000,000</td>
<td>2,000</td>
<td>120</td>
<td>1,500</td>
<td>1,000</td>
<td>1</td>
</tr>
<tr>
<td>4.80</td>
<td>65%</td>
<td>Mycobacterium avium</td>
<td>8,000,000</td>
<td>4,000</td>
<td>200</td>
<td>2,500</td>
<td>1,000</td>
<td>2</td>
</tr>
<tr>
<td>5.35</td>
<td>70%</td>
<td>Mycobacterium avium</td>
<td>10,000,000</td>
<td>5,000</td>
<td>250</td>
<td>3,000</td>
<td>1,500</td>
<td>2</td>
</tr>
</tbody>
</table>

Consensus Quality Evaluation

As one measure of genome consensus quality sequence, we used Ilumina MiSeq 250bp paired-end data to align to complete genomes sequenced using PacBio data alone and assembled using one of three genome assemblers. We selected 9 genomes sequenced using RS and 6 using RS II. An average of 50x Illumina coverage was aligned to the contig consensus sequence using BWA and variants were called using GATK.

Genome Assembler Comparison

Using 14 genomes with a range of sizes and GC content and sequenced using both RS and RS II, we evaluated three genome assemblers (CAT 0.7, HGAP, and HGAP2) for their ability to generate high-quality, complete genomes.