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 Blue Pippin 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 *de facto* 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 CA7.0 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 increased to more than 60% (39 of 64 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 (CA7.0, HGAP, and HGAP2) using a subset of 14 bacterial genomes with a range of sizes and GC%. 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, CA7.0 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. However, these initial data indicate that bacterial genomes assembled using PacBio data alone generate highly accurate consensus sequence.

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

References

DePristo MA, et al. (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491-498.

Hasan NA, et al. (2012) Genomic diversity of 2010 Haitian cholera outbreak strains. Proceedings of the National Academy of Sciences.

Koren S, et al. (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 30(7):693-700.

Li H & Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754-1760.

Magoc T, et al. (2013) GAGE-B: An Evaluation of Genome Assemblers for Bacterial Organisms. Bioinformatics.

Miller JR, et al. (2008) Aggressive assembly of pyrosequencing reads with mates. Bioinformatics 24(24):2818-2824.

Shin SC, et al. (2013) Advantages of Single-Molecule Real-Time Sequencing in High-GC Content Genomes. PLoS ONE 8(7):e68824.

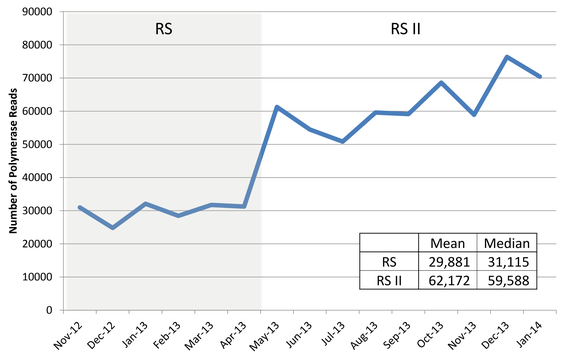
Tettelin H, et al. (2012) Genomic Insights into the Emerging Human Pathogen Mycobacterium massiliense. Journal of Bacteriology 194(19):5450.

Acknowledgements

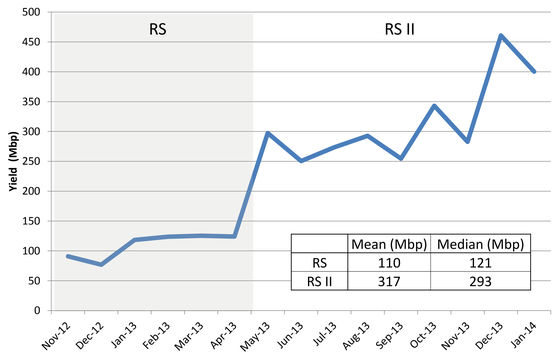
This project has been funded in whole or part with federal funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under contract number HHSN272200900009C.

Improved Total Yield & Read Length

Average SMRT Cell Passed Filter Read Count



Average SMRT Cell Subread Yield (Mbp)



Average SMRT Cell Subread Lengths

	RS		RS II	
	Mean	Median	Mean	Median
Mean subread length	2,347	2,250	3,444	3,386
N50 subread length	3,280	3,139	4,775	4,608
P95 subread length	5,871	5,616	8,188	8,040
Polymerase mean read length (not shown)	3,773	3,899	5,047	4,833
Mean library insert length (not shown)	7,120	7,190	11,773	11,437

Consensus Quality Evaluation

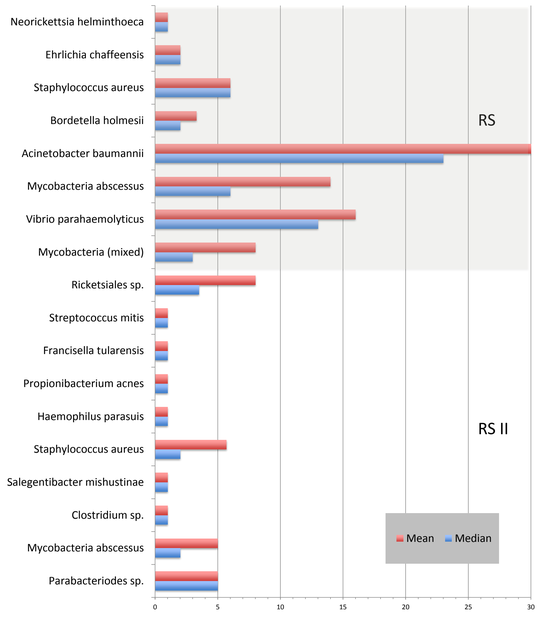
As one measure of genome consensus sequence quality, we used Illumina MiSeq 250bp PE data to align to complete genomes sequenced using PacBio data alone and assembled using one of three genome assemblers. We selected 6 genomes sequenced using RS and 6 using RS II. An average of 50x Illumina coverage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

Species	GenomeSize	%GC	Platform	Assembler	Raw SNPs	PF SNPs
Ehrlichia chaffeensis	1,200,000	30%	RS	CA7.0	2	2
Ehrlichia chaffeensis	1,200,000	30%	RS	CA7.0	16	10
Bordatella holmesii	3,705,000	62%	RS	HGAP	23	0
Bordatella holmesii	3,705,000	62%	RS	HGAP	10	0
Bordatella holmesii	3,705,000	62%	RS	HGAP	5	0
Mycobacterium abscessus	4,800,000	64%	RS	CA7.0	8	5
Mean Per Genome					**10.7**	**2.8**
Median Per Genome					**9.0**	**1.0**
Mean Per Mbp Genome					**3.5**	**0.9**
Neoehrlichia lotoris	1,300,000	28%	RS II	HGAP2	8	8
Francisella tularensis	1,900,000	32%	RS II	CA7.0	0	0
Francisella tularensis	1,900,000	32%	RS II	HGAP	0	0
Staphylococcus aureus	2,850,000	33%	RS II	HGAP2	3	1
Staphylococcus aureus	2,850,000	33%	RS II	HGAP2	12	1
Mycobacterium abscessus	4,800,000	64%	RS II	CA7.0	9	9
Mean Per Genome					**5.3**	**3.2**
Median Per Genome					**5.5**	**1.0**
Mean Per Mbp Genome					**2.1**	**1.2**

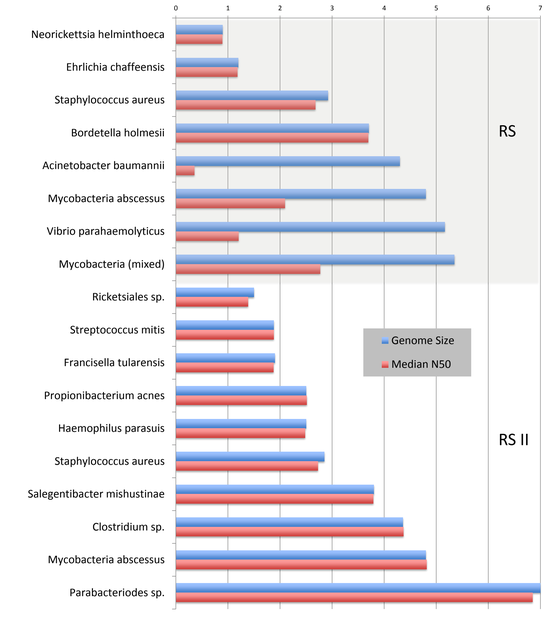
Genome Assembly & Finished Bacterial Genomes

Species	Genome Size	GC%	Genomes	Platform	Median Contigs	Mean Contigs	Median N50	Mean N50	Complete Genomes
Neorickettsia helminthoeca	900,000	42%	1	RS	1	1	894,365	894,365	1
Ehrlichia chaffeensis	1,200,000	30%	8	RS	2	2	1,183,219	1,179,717	6
Staphylococcus aureus	2,920,000	33%	1	RS	6	6	2,677,329	2,677,329	0
Bordetella holmesii	3,705,000	62%	7	RS	2	3.3	3,694,298	3,092,498	3
Acinetobacter baumannii	4,300,000	39%	6	RS	23	30	356,047	327,813	0
Mycobacteria abscessus	4,800,000	64%	12	RS	6	14	2,094,780	2,167,167	2
Vibrio parahaemolyticus	5,165,770	45%	10	RS	13	16	1,206,254	1,704,545	1
Mycobacteria (mixed)	5,350,000	64%	11	RS	3	8	2,768,387	2,800,673	2
RS SUMMARY			**56**		**7.0**	**10.0**	**1,859,335**	**1,855,513**	**15**
Ricketsiales sp.	1,500,000	35%	16	RS II	3.5	8	1,384,902	1,161,052	7
Streptococcus mitis	1,880,000	40%	1	RS II	1	1	1,881,073	1,881,073	1
Francisella tularensis	1,900,000	32%	2	RS II	1	1	1,874,734	1,874,734	2
Propionibacterium acnes	2,500,000	60%	1	RS II	1	1	2,513,768	2,513,768	1
Haemophilus parasuis	2,500,000	40%	1	RS II	1	1	2,478,759	2,478,759	1
Staphylococcus aureus	2,850,000	33%	36	RS II	2	5.7	2,727,458	2,341,993	23
Salegentibacter mishustinae	3,800,000	37%	1	RS II	1	1	3,791,564	3,791,564	1
Clostridium sp.	4,360,000	31%	2	RS II	1	1	4,373,190	4,373,190	2
Mycobacteria abscessus	4,800,000	64%	3	RS II	2	5	4,818,226	4,381,566	1
Parabacteriodes sp.	7,000,000	43%	1	RS II	5	5	6,847,904	6,847,904	0
RS II SUMMARY			**64**		**1.9**	**3.0**	**3,269,158**	**3,164,560**	**39**

Mean & Median Contig Counts



N50 Contig Length vs. Genome Size (Mbp)

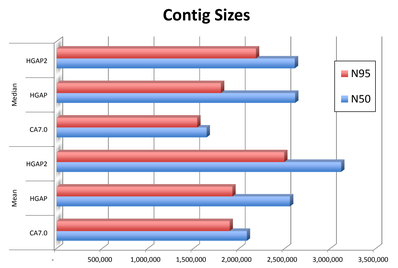


Genome Assembler Comparison

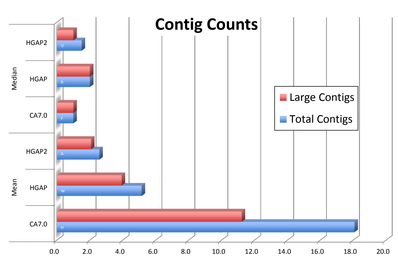
Using 14 genomes with a range of sizes and GC%, and sequenced using both RS and RS II, we evaluated three genome assemblers (CA7.0, HGAP, and HGAP2) for their ability to generate high-quality, complete genomes.

	Mean Total Contigs	Median Total Contigs	Mean Large Contigs	Median Large Contigs	Mean N50	Median N50	Mean N95	Median N95	Complete Genomes
CA7.0	18.1	1	11.2	1	2,081,210	1,640,793	1,892,823	1,538,150	8
HGAP	5.1	2	3.9	2	2,554,490	2,611,204	1,923,204	1,797,944	6
HGAP2	2.6	1.5	2.1	1	3,116,890	2,607,251	2,491,518	2,181,180	7

Contig Sizes



Contig Counts



Species	Assembler	Platform	Corrected Coverage	GC%	Contigs	Contig Bases	N50	N95	Large Contigs*	Large Contig Bases	Min Contig Bases	Max Contig Bases
Erlichia chaffeensis Liberty	CA7.0	RS	26.2	30.26	1	1,177,627	1,177,627	1,177,627	1	1,177,627	1,177,627	1,177,627
	HGAP	RS	30.9	30.22	1	1,183,167	1,183,167	1,183,167	1	1,183,167	1,183,167	1,183,167
	HGAP2	RS	20.7	30.28	1	1,185,549	1,185,549	1,185,549	1	1,185,549	1,185,549	1,185,549
Erlichia chaffeensis Jax	CA7.0	RS	25.8	30.29	1	1,194,801	1,194,801	1,194,801	1	1,194,801	1,194,801	1,194,801
	HGAP	RS	35.7	30.30	1	1,194,333	1,194,333	1,194,333	1	1,194,333	1,194,333	1,194,333
	HGAP2	RS	20.0	30.30	1	1,194,796	1,194,796	1,194,796	1	1,194,796	1,194,796	1,194,796
Streptococcus mitis	CA7.0	RS II	36.3	40.46	1	1,881,498	1,881,498	1,881,498	1	1,881,498	1,881,498	1,881,498
	HGAP	RS II	25.1	40.45	1	1,881,073	1,881,073	1,881,073	1	1,881,073	1,881,073	1,881,073
	HGAP2	RS II	21.6	40.46	1	1,883,601	1,883,601	1,883,601	1	1,883,601	1,883,601	1,883,601
Haemophilus parasuis	CA7.0	RS II	19.9	39.69	1	2,480,212	2,480,212	2,480,212	1	2,480,212	2,480,212	2,480,212
	HGAP	RS II	15.8	39.77	2	2,502,045	2,486,298	2,486,298	2	2,502,045	15,747	2,486,298
	HGAP2	RS II	16.0	39.72	1	2,478,759	2,478,759	2,478,759	1	2,478,759	2,478,759	2,478,759
Propionibacterium acnes	CA7.0	RS II	22.0	60.01	1	2,507,260	2,507,260	2,507,260	1	2,507,260	2,507,260	2,507,260
	HGAP	RS II	30.5	60.04	1	2,513,768	2,513,768	2,513,768	1	2,513,768	2,513,768	2,513,768
	HGAP2	RS II	17.0	60.01	1	2,506,623	2,506,623	2,506,623	1	2,506,623	2,506,623	2,506,623
Staphylococcus aureus 243	CA7.0	RS II	15.8	32.81	2	2,720,678	2,705,850	2,705,850	2	2,720,678	14,828	2,705,850
	HGAP	RS II	16.0	32.77	3	2,750,735	2,708,639	2,708,639	3	2,750,735	12,584	2,708,639
	HGAP2	RS II	15.6	32.77	2	2,739,831	2,707,878	2,707,878	2	2,739,831	31,953	2,707,878
Staphylococcus aureus 221	CA7.0	RS II	21.7	32.81	7	2,830,704	1,048,604	129,521	5	2,815,159	6,147	1,164,165
	HGAP	RS II	20.0	32.82	1	2,822,436	2,822,436	2,822,436	1	2,822,436	2,822,436	2,822,436
	HGAP2	RS II	19.5	32.83	3	2,842,228	2,421,582	407,680	3	2,842,228	12,966	2,421,582
Staphylococcus aureus 241	CA7.0	RS II	15.9	32.84	1	2,820,118	2,820,118	2,820,118	1	2,820,118	2,820,118	2,820,118
	HGAP	RS II	13.4	32.88	2	2,840,510	2,823,566	2,823,566	2	2,840,510	16,944	2,823,566
	HGAP2	RS II	15.0	32.82	1	2,823,087	2,823,087	2,823,087	1	2,823,087	2,823,087	2,823,087
Mycobacterium abscessus 2258	CA7.0	RS	12.3	64.21	14	4,786,270	732,562	116,622	14	4,786,270	18,019	1,128,493
	HGAP	RS	8.6	64.22	6	4,798,287	3,214,483	185,294	5	4,795,950	2,337	3,214,483
	HGAP2	RS	18.5	64.21	3	4,813,404	3,407,250	1,399,956	2	4,807,206	6,198	3,407,250
Mycobacterium abscessus 2446	CA7.0	RS	15.4	64.20	4	4,807,730	1,400,088	1,056,005	3	4,802,889	4,841	2,346,796
	HGAP	RS	12.3	64.20	4	4,832,858	3,187,532	1,630,014	2	4,817,546	5,720	3,187,532
	HGAP2	RS	19.1	64.19	6	4,844,376	1,466,456	948,720	3	4,824,143	6,186	2,408,967
Mycobacterium abscessus 21	CA7.0	RS	24.4	64.17	1	5,030,544	5,030,544	5,030,544	1	5,030,544	5,030,544	5,030,544
	HGAP	RS	8.9	64.17	2	5,024,237	4,888,065	4,888,065	2	5,024,237	136,172	4,888,065
	HGAP2	RS	23.2	64.17	2	5,030,679	5,024,321	5,024,321	1	5,024,321	6,358	5,024,321
Mycobacterium abscessus 1518	CA7.0	RS	19.9	64.03	1	5,252,194	5,252,194	5,252,194	1	5,252,194	5,252,194	5,252,194
	HGAP	RS	14.6	64.03	22	5,394,060	1,162,445	742,287	8	5,301,331	4,815	1,660,474
	HGAP2	RS	22.0	64.04	1	5,252,207	5,252,207	5,252,207	1	5,252,207	5,252,207	5,252,207
Mycobacterium kansasii	CA7.0	RS II	15.9	66.03	197	7,162,511	101,842	22,158	107	6,615,200	1,613	303,398
	HGAP	RS II	14.9	66.11	21	6,801,699	519,935	151,103	21	6,801,699	10,181	1,343,207
	HGAP2	RS II	19.1	66.14	7	6,733,948	4,436,441	220,171	6	6,726,861	7,087	4,436,441
Parabacteroides sp.	CA7.0	RS II	15.2	43.44	21	6,922,820	803,737	125,116	18	6,895,544	8,221	2,901,215
	HGAP	RS II	16.1	43.43	5	6,941,437	5,177,124	1,714,814	5	6,941,437	13,425	5,177,124
	HGAP2	RS II	15.8	43.44	6	6,937,625	6,847,904	6,847,904	5	6,928,550	9,075	6,847,904