J. Microbiol. Biotechnol. (2016), 26(1), 207–212
http://dx.doi.org/10.4014/jmb.1507.07055

jmb

Toward Complete Bacterial Genome Sequencing Through the Combined Use of Multiple Next-Generation Sequencing Platforms S

Haeyoung Jeong[1,3]*, Dae-Hee Lee[2,3], Choong-Min Ryu[1,3], and Seung-Hwan Park[1,3]

[1]*Super-Bacteria Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 34141, Republic of Korea*
[2]*Synthetic Biology and Bioengineering Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 34141, Republic of Korea*
[3]*Biosystems and Bioengineering Program, Korea University of Science and Technology (UST), Daejeon 34113, Republic of Korea*

Received: July 15, 2015
Revised: September 15, 2015
Accepted: October 14, 2015

First published online
October 14, 2015

**Corresponding author*
Phone: +82-42-860-4237;
Fax: +82-42-860-4488;
E-mail: hyjeong@kribb.re.kr

Supplementary data for this paper are available on-line only at http://jmb.or.kr.

pISSN 1017-7825, eISSN 1738-8872

PacBio's long-read sequencing technologies can be successfully used for a complete bacterial genome assembly using recently developed non-hybrid assemblers in the absence of second-generation, high-quality short reads. However, standardized procedures that take into account multiple pre-existing second-generation sequencing platforms are scarce. In addition to Illumina HiSeq and Ion Torrent PGM-based genome sequencing results derived from previous studies, we generated further sequencing data, including from the PacBio RS II platform, and applied various bioinformatics tools to obtain complete genome assemblies for five bacterial strains. Our approach revealed that the hierarchical genome assembly process (HGAP) non-hybrid assembler resulted in nearly complete assemblies at a moderate coverage of ~75x, but that different versions produced non-compatible results requiring post processing. The other two platforms further improved the PacBio assembly through scaffolding and a final error correction.

Keywords: Next-generation sequencing, complete genome sequencing, PacBio, non-hybrid assembly

Emerging long-read-based sequencing technologies, such as PacBio RS II and Illumina synthetic long reads, have been changing the way in which complete, high-quality microbial genomes are created [13]. Owing to the high error rate of PacBio reads, data from the initial version can only be used with other high-quality data [7, 14, 21]. However, the introduction of a non-hybrid assembly approach that depends on self-correction [5] greatly facilitates the complete sequencing of small genomes. Although long-read-based approaches are promising for bacterial genome assemblies, a post-processing step is still required, and conventional low-cost short reads generated from Illumina or other platforms are often helpful [12]. To present a standard procedure for bacterial genome assemblies, a recent study reported a performance comparison among five popular assemblers using publicly available data [16], and the results showed that one single-molecule real-time (SMRT) cell is adequate for completing the bacterial genome sequencing. In this study, we describe the best practice for achieving complete bacterial genome assemblies using multiple next-generation sequencing platforms that include both short (short-insert shotgun libraries and mate-pair libraries) and long reads.

We utilized short reads from three plant pathogens, which were produced through previously published studies [11, 20]. For this study, two additional bacterial genomes, namely, a human enteric pathogen, *Shigella boydii* ATCC 9210, obtained from the American Type Culture Collection, and a newly isolated plant growth-promoting rhizobacterium, *Paenibacillus* sp. HS311 [19], were sequenced. Cells were grown in a tryptic soy broth (Difco, MI, USA) at 37°C (ATCC 9210) or 30°C (HS311), harvested, and resuspended in 50 mM EDTA (pH 8.0) before lysing with lysozyme (2 mg/ml). The genomic DNA was isolated using a Wizard genomic DNA purification kit according to the manufacturer's instructions (Promega, WI, USA). Ion Torrent and PacBio

Table 1. Bacterial strains and datasets used in this study.

Bacterial strain and reference	Fragment (Illumina HiSeq)	3 kb Mate-pair[b] (Ion Torrent PGM)	10 kb Long reads[c] (PacBio RS II)	DDBJ/EMBL/NCBI assembly accession
Shigella boydii ATCC 9210 (human enteric pathogen)	2 × 101; 393 bp insert[a] 2,968,831,774 bp	417,515 reads 70,104,774 bp (168 bp avg.)	3 SMRT cells CLR: 234,935 reads, 601,459,803 bp (2,560 bp avg.) Pre-assembly: 14,955 reads, 81,405,114 bp (5,443 bp avg.)	CP011511 (this study)
Paenibacillus sp. HS311 (plant growth-promoting rhizobacterium) [19]	2 × 101; 406 bp insert[a] 3,181,910,868 bp		3 SMRT cells CLR: 211,758 reads, 753,392,799 bp (3,558 bp avg.) Pre-assembly: 17,355 reads, 114,072,371 bp (6,573 bp avg.)	CP011512-3 (this study)
Pseudomonas syringae pv. *syringae* KCTC 12500^T (plantpathogen) [20]	2 × 101; 329 bp insert[a] 3,473,679,668 bp	302,283 reads 50,880,397 bp (168 bp avg.)	2 SMRT cells CLR: 164,324 reads, 588,640,241 bp (3,582 bp avg.) Pre-assembly: 15,109 reads 97,523,882 bp (6,455 bp avg.)	AYTM00200000 (updated by this study)
Pseudomonas amygdali pv. *tabaci* ATCC 11528 (plant pathogen) [11]	2 × 101; 385 bp insert[a] 2,992,727,364 bp			LCWS01000000[d]
Pseudomonas amygdali pv. *lachrymans* 98A-744 (plant pathogen) [11]	2 × 101; 377 bp insert[a] 3,159,709,250 bp	313,102 reads 51,217,854 bp (164 bp avg.)		LCWT01000000[d]

[a]Calculated from the de novo assembly results by CLC Genomics Workbench.

[b]One 314 chip was used for each sample. The read numbers and lengths are based on the raw SFF files that were not yet split into di-tags. Thus, the reads contain a 35 bp internal linker (5′-CTGCTGTACCGTACATCCGCCTTGGCCGTACAGCAG-3′).

[c]Results of SMRT Analysis 2.1 (SMRT Pipe 1.79).

[d]Assembly results obtained through this study are available from http://wiki.genoglobe.kr/kribb/Pseudomonas_amygdali. The first versions (NCBI), which exhibit better assembly statistics than the recent assemblies, were not replaced.

sequencing were carried out for differently selected strains on the basis of research relevance and assembly statistics. The details of each sequencing procedure are given below. All sequencing data are summarized in Table 1.

First, Illumina reads from all five strains were used for an evaluation of various de Bruijn graph-based short-read assemblers. Sequencing was carried out using the Illumina HiSeq 2000 system by the National Instrumentation Center for Environmental Management at Seoul National University (Seoul, Korea). Short insert libraries with an average insert size of 500 bp were constructed using a TruSeq DNA sample preparation v2 kit, and produced 101 cycle paired-end reads (>500x coverage) (Table 1). De novo assemblies were achieved using A5-miseq v20141120 [6], SPAdes v3.5.0 [1], CLC Genomics Workbench v8.0 (CLC bio), and Velvet v1.2.10 [22]. The assembly results were compared using Quast v2.3 [9], a software tool for evaluating genome assemblies (Table 2). Overall, A5-miseq produces the best assembly results in terms of the contig numbers and N_{50}. Adapter removal and error correction seem to be dispensable if quality trimming and read-length filtering are carried out, as shown through the CLC Genomics Workbench example. *Shigella boydii* yielded the most fragmentary assemblies owing to hundreds of insertion sequences that cannot be spanned through short reads.

We then applied mate-pair library sequencing to scaffold the Illumina contigs for *S. boydii*, *Pseudomonas syringae* pv. *syringae*, and *P. amygdali* pv. *lachrymans*, which generated too many contigs. Although mate-pair libraries providing long-range "jumping" sequences are considered mandatory for a de novo assembly or detecting structural variations for higher organisms, they are often omitted in the production of bacterial draft genomes. We used a SOLiD 5500 mate-pair library kit because the Illumina mate-pair library kit is prone to producing undesirable "inward-facing" read pairs and chimeric reads [18]. Ion Torrent PGM sequencing from 3 kb mate-pair libraries, using one

Table 2. Comparison of assembly results.

		Short reads (Illumina HiSeq)					Long reads (PacBio RS II)		
							Non-hybrid (HGAP)		Hybrid
		A5-miseq	SPAdes	CLC[a]	CLC-ec[b]	Velvet	V2.1	V2.3	SPAdes
Shigella boydii ATCC 9210	No. of contigs	**369**	461	417	418	484	3	15	174
	Largest contig	**175652**	104872	104473	104473	104469	4587019	1976317	461617
	Total length	4336574	**4495998**	4319423	4318423	4320507	4604476	4648679	4709393
	N_{50}	**23065**	21407	22487	22487	22901	4587019	1243574	104879
Paenibacillus sp. HS311	No. of contigs	**30**	45	37	37	69	2	3	14
	Largest contig	1484161	1545235	1351120	1482938	**1712449**	6006274	5907378	1914721
	Total length	**6163388**	6162685	6159762	6159478	6162742	6226787	6235253	6208146
	N_{50}	748138	746736	676231	746420	**1454902**	6006274	5907378	1459016
Pseudomonas syringae pv. syringae KCTC 12500	No. of contigs	78	602	88	85	**64**	5	2	495
	Largest contig	**1128838**	667357	667266	667266	679163	382564	6123363	4515141
	Total length	6114261	**6291756**	6092023	6091748	6080556	6177218	6137722	6305168
	N_{50}	**485402**	324026	393565	381880	382564	2652685	6123363	4515141
Pseudomonas amygdali pv. tabaci ATCC 11528	No. of contigs	**18**	56	41	41	58			
	Largest contig	**900244**	595932	893651	893661	868410			
	Total length	6129081	**6129459**	6122066	6122098	6124482			
	N_{50}	**542503**	318569	344242	381935	325117			
Pseudomonas amygdali pv. lachrymans 98A-744	No. of contigs	**193**	292	260	260	463			
	Largest contig	**460476**	266847	266682	266682	216051			
	Total length	6175433	**6257853**	6154729	6153490	6154393			
	N_{50}	**85995**	75956	83802	78970	79601			

All statistics are based on contigs of size ≥ 200 bp. The bold and underlined figures represent the best and worst results, respectively (shown only for short-read assemblies). The *k*-mer values were automatically chosen by the assemblers within the range of 20–100 (A5-miseq) or 21–99 (SPAdes and Velvet). A word size of 64 was chosen for CLC Genomics Workbench.

[a]The reads were quality trimmed and filtered (quality limit of 0.01, max. of one ambiguous base per read, and min. read length of 70).

[b]Error-corrected reads (using the A5-miseq pipeline) were used.

314 chip run per sample, was carried out by GenoTech Corporation (Daejeon, Korea). After converting the reads into di-tags (reverse-reverse direction), SSPACE-premium v2.3 (BaseClear) was applied along with CLC assemblies to produce the scaffolds. The number of scaffolds decreased by 21.3% to 65.6%, whereas the N_{50} values increased by 386% to 1,058% (Table S1). The most dramatic improvement was observed for *S. boydii*.

Next, we employed the PacBio sequencing platform to obtain nearly finished genome assemblies for *S. boydii*, *Paenibacillus* sp. HS311, and *Pseudomonas syringae* pv. *syringae*. Two of the strains were chosen based on their research relevance; the other (*S. boydii*), although improved through the use of Ion Torrent PGM-based scaffolding, was chosen because it is still too far from completion. We also evaluated the utility of other platform-based data for an improvement of the PacBio assemblies.

The 10 kb libraries constructed from the three strains were sequenced with modest coverage (64.4x to 77.7x) using the C2-P4 Chemistry by DNA Link (Seoul, Korea). Subread filtering, a non-hybrid assembly, and subsequent polishing using SMRT Analysis v2.1 (RS_HGAP.1 protocol, the initial hierarchical genome assembly process (HGAP) production implementation) [5], resulted in nearly complete assemblies; that is, two to five contigs (Table 2). At the time of writing, a new version of SMRT Analysis, v2.3 has been made available. From v2.2, the assembly protocol was replaced with RS_HGAP_Assembly.2, with improvements in both the correction step and the overlap detection in the preassembly process. Expecting to obtain the highest level of performance through the use of the latest version of the SMRT Analysis program, we also applied v2.3 to the same dataset, and compared the two results using either LASTZ [10] or MUMmer [15]. For *S. boydii* and *Paenibacillus* sp. HS311, the two sets of assemblies were shown to be incompatible with each other (Figs. 1A and 1B). A cumulative GC skew analysis [17] and whole-genome comparisons against the available complete genomes (*S. boydii* Sb227 and CDC 3083-94, *Paenibacillus polymyxa* Sb3-1, and *Pseudomonas syringe* B728a) support the idea that the earlier version of

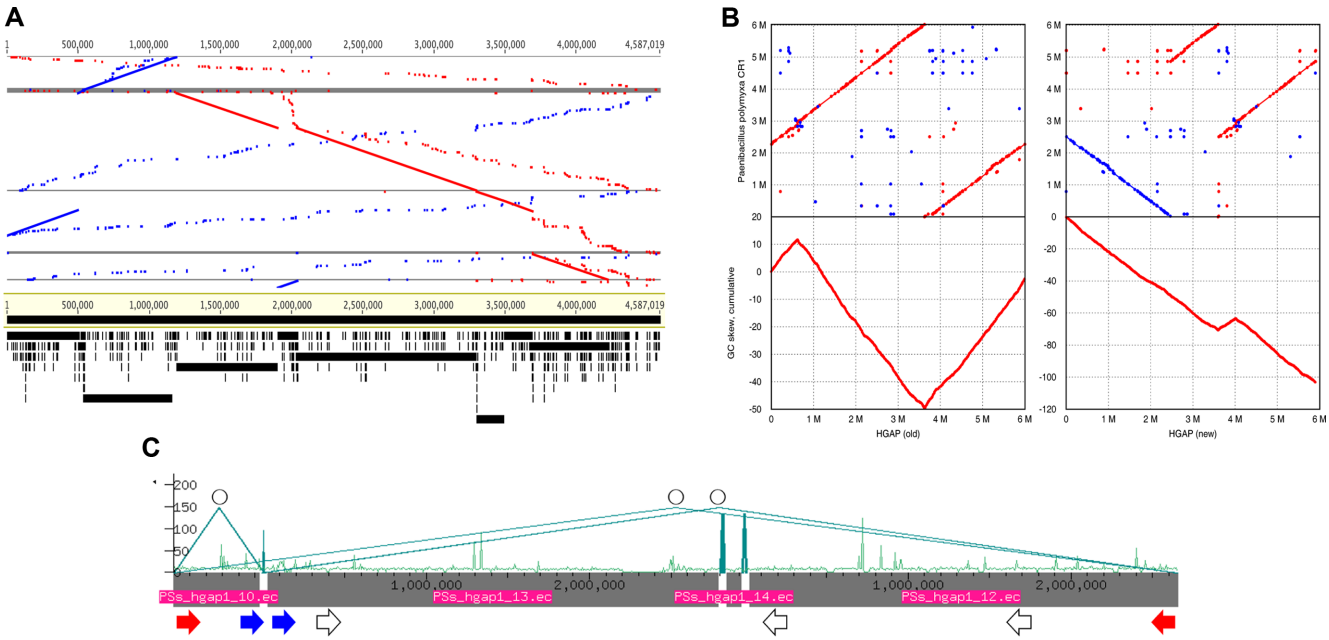

Fig. 1. Validation of HGAP assemblies.

(**A**) LASTZ alignment [10] between two versions of *Shigella boydii* HGAP assemblies. The upper panel shows a dot plot; and the lower panel, alignment blocks. The major contig from the old version of HGAP is shown in the horizontal axis. The plots were generated using Geneious Pro R8 (http://www.geneious.com). (**B**) MUMmer whole-genome alignments [15] of two versions of *Paenibacillus* sp. HS311 HGAP assemblies (left, old version; right, new version) with the complete genome sequence of *P. polymyxa* CR1 (upper panel) and cumulative GC skew plots as calculated by (G-C)/(G+C) with a window size of 5 kb (lower panel). (**C**) Ion Torrent PGM mate-pair reads on *Pseudomonas syringa*e pv. *syringae* HGAP contigs were mapped and visualized using Consed software [8], the results indicating that the four contigs are arranged in a single scaffold. The light-green plot designates the read depth. Multiple copies of ribosomal RNA genes, designated by the thick arrows at the bottom, induced mate reads to align at a longer span (○). RNA genes at the end of the adjacent contigs, represented through filled-in arrows of the same color, were used to join them, resulting in two contigs.

the HGAP assembler yields more accurate results (Fig. 1B, lower panel). For HS311, atypically strong base skews in the Firmicutes [4] helped identify the proper orientation and order of the contigs. A SPAdes hybrid assembly using Illumina reads and PacBio filtered subreads (processed from SMRT Analysis v2.1) did not yield results comparable to those of a non-hybrid assembler.

We finally proceeded with a post-processing of the PacBio assemblies (derived from SMART Analysis v2.1), which can make use of data from other sequencing data types. Illumina short reads were re-mapped to the contigs using the CLC Genomics Workbench, and 8 to 16 additional nucleotide-level differences were identified (Table S2). Consensus extraction and re-mapping, followed by variant detection, were iterated until no further variants were found. We did not carry out a final error correction for *S. boydii* because the random mapping of reads originating from repetitive regions caused "new" variants after each round of mapping. From the error-corrected PacBio assemblies, small contigs that were already contained in the other contigs were regarded as untrustworthy and thus discarded (Fig. S1), resulting in a single contig in both *S. boydii* and HS311, and four contigs in *P. syringae*. The contigs from *P. syringae* were arranged in a single scaffold by comparing them with other complete genomes and mapping the mate-pair reads onto them (Fig. 1C). Both ends of the single contigs were inspected to determine whether they had any overlapping sequences generating circular chromosomes. Redundant sequences at the end of the *S. boydii* contig were trimmed, and a small gap (~260 bp) at the end of the HS311 contig was filled in using GapFiller [3]. The final finished genome sequences were obtained by adjusting the starting nucleotide position based on the putative replication origin of the reference genome sequences. The remaining 221 kb contig in HS311 was assigned a putative plasmid because its gene contents and alignment with the contigs resulting from the short-read assemblers strongly support a circular replicon structure (Fig. S2).

Using the three finalized HGAP assemblies as references, we ran Quast again to assess the results of the short-read

Table 3. Comparison of assembly quality using the final HGAP assemblies (based on SMRT Analysis v2.1) as the reference sequences. Misassemblies are defined as described by Plantagora [2]. The bold and underlined figures represent the best and worst results, respectively (shown only for short-read assemblies).

		Short reads (Illumina HiSeq)					Long reads (PacBio RS II)	
		A5-miseq	SPAdes	CLC	CLC-ec[a]	Velvet	Non-hybrid (HGAP) V2.3	Hybrid SPAdes
Shigella boydii ATCC 9210	**Misassemblies**							
	# misassemblies	21	**7**	8	8	4	4	12
	Misassembled contig length	475648	**5662**	43860	63353	101141	3229510	232527
	Mismatches							
	# mismatches per 100 kb	1.08	4.22	**0.92**	2.050	2.63	0	10.1
	# indels per 100 kb	1.25	0.44	**0.36**	0.51	1.41	1.2	0.17
	# Ns per 100 kb	28.19	**6.41**	85.67	99.95	224.73	0	0
Paenibacillus sp. HS311	**Misassemblies**							
	# misassemblies	4	**1**	**1**	3	7	2	4
	Misassembled contig length	51114	27845	**500**	496131	3533478	5907378	1278
	Mismatches							
	# mismatches per 100 kb	0.8	2.1	**0.45**	0.91	1.37	0.43	5.43
	# indels per 100 kb	0.8	0.59	**0.31**	**0.31**	0.33	0.47	1.130
	# Ns per 100 kb	15.48	**3.54**	12.93	11.91	56.94	0	0
Pseudomonas syringae pv. syringae KCTC 12500	**Misassemblies**							
	# misassemblies	10	**2**	3	5	9	5	4
	Misassembled contig length	2635861	548817	**547148**	1182337	1744776	6137722	6121097
	Mismatches							
	# mismatches per 100 kb	2.74	1.49	**1.33**	2.77	4.72	0.47	7.34
	# indels per 100 kb	2.490	**1.05**	1.25	1.33	1.78	1.25	1.82
	# Ns per 100 kb	9.27	**0.83**	4.21	18.850	170.09	0	2.62

[a]Error-corrected reads by the A5-miseq pipeline were used.

assemblers in terms of the misassemblies and base mismatches. In contrast to the assembly metric measurement, A5-miseq exhibited the lowest assembly quality, whereas CLC Genomics Workbench showed the highest level of performance (Table 3). We also compared the mapping statistics of Ion Torrent mate-pair reads on A5-miseq and CLC assemblies. The mapping rate, number of consistent pairs, and quality of the scaffolding results were higher for the CLC assemblies (Table S3).

In summary, a HGAP non-hybrid assembly using PacBio long reads (~100x) is currently the most efficient method for obtaining bacterial genome sequences at the finishing levels. Despite the reported improvement in performance, we unexpectedly found that the latest SMRT Analysis program did not produce better results for the same dataset than the earlier version. During the post-processing phase, Illumina reads were particularly useful for gap filling and the correction of local errors. The Ion Torrent mate-pair reads were shown to be helpful for the scaffolding of short-read-based assemblies and for the ordering of the HGAP contigs. We believe that our results can be used as general guidelines for the completion of bacterial genome sequencing when datasets from multiple sequencing platforms are made available.

Acknowledgments

This work was supported by the KRIBB Research Initiative Program, Ministry of Science, ICT, and Future Planning, and by the Next-Generation BioGreen 21 Program (SSAC Grant No. PJ009524) funded by the RDA (to C.M.R), Republic of Korea.

References

1. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, *et al.* 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19:** 455-477.
2. Barthelson R, McFarlin AJ, Rounsley SD, Young S. 2011.

Plantagora: modeling whole genome sequencing and assembly of plant genomes. *PLoS One* **6:** e28436.
3. Boetzer M, Pirovano W. 2012. Toward almost closed genomes with GapFiller. *Genome Biol.* **13:** R56.
4. Charneski CA, Honti F, Bryant JM, Hurst LD, Feil EJ. 2011. Atypical at skew in Firmicute genomes results from selection and not from mutation. *PLoS Genet.* **7:** e1002283.
5. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, *et al.* 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10:** 563-569.
6. Coil D, Jospin G, Darling AE. 2015. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. *Bioinformatics* **31:** 587-589.
7. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, *et al.* 2012. Mind the gap: upgrading genomes with Pacific Biosciences RS long-read sequencing technology. *PLoS One* **7:** e47768.
8. Gordon D, Green P. 2013. Consed: a graphical editor for next-generation sequencing. *Bioinformatics* **29:** 2936-2937.
9. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29:** 1072-1075.
10. Harris RS. 2007. Improved pairwise alignment of genomic DNA. PhD thesis. Pennsylvania State University.
11. Jeong H, Kloepper JW, Ryu C-M. 2015. Genome sequences of *Pseudomonas amygdali* pv. *tabaci* strain ATCC 11528 and pv. *lachrymans* strain 98A-744. *Genome Announc.* **3:** e00683-00615.
12. Kamada M, Hase S, Sato K, Toyoda A, Fujiyama A, Sakakibara Y. 2014. Whole genome complete resequencing of *Bacillus subtilis natto* by combining long reads with high-quality short reads. *PLoS One* **9:** e109999.
13. Koren S, Phillippy AM. 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Curr. Opin. Microbiol.* **23:** 110-120.
14. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, *et al.* 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat. Biotechnol.* **30:** 693-700.
15. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* **5:** R12.
16. Liao YC, Lin SH, Lin HH. 2015. Completing bacterial genome assemblies: strategy and performance comparisons. *Sci. Rep.* **5:** 8747.
17. Lobry JR, Louarn JM. 2003. Polarisation of prokaryotic chromosomes. *Curr. Opin. Microbiol.* **6:** 101-108.
18. Park N, Shirley L, Gu Y, Keane TM, Swerdlow H, Quail MA. 2013. An improved approach to mate-paired library preparation for Illumina sequencing. *Methods Next Gener. Seq.* **1:** 10-20.
19. Park S-H, Choi S-K, Park S-Y, Jeon JH, Kim HR, Jeong J, Kim YT. 2015. Novel *Paenibacillus* sp. and the method for yield increase of potato using the same. Republic of Korea patent application 10-1498155.
20. Park YS, Jeong H, Sim YM, Yi HS, Ryu CM. 2014. Genome sequence and comparative genome analysis of *Pseudomonas syringae* pv. *syringae* type strain ATCC 19310. *J. Microbiol. Biotechnol.* **24:** 563-567.
21. Ribeiro FJ, Przybylski D, Yin S, Sharpe T, Gnerre S, Abouelleil A, *et al.* 2012. Finished bacterial genomes from shotgun sequence data. *Genome Res.* **22:** 2270-2277.
22. Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18:** 821-829.