3 Biotech (2020) 10:480
https://doi.org/10.1007/s13205-020-02474-0

ORIGINAL ARTICLE

Evaluation of assembly methods combining long-reads and short-reads to obtain *Paenibacillus* sp. R4 high-quality complete genome

Seung Chul Shin[1] · Woong Choi[2] · Junhyuck Lee[2,3] · Hyo Jin Kim[4,5] · Han-Woo Kim[2,3]

Received: 6 August 2020 / Accepted: 7 October 2020 / Published online: 19 October 2020

Abstract

We sequenced the *Paenibacillus* sp. R4 using Oxford Nanopore Technology (ONT), single molecule real-time (SMRT) technology from Pacific Biosciences (PacBio), and Illumina technologies to investigate the application of nanopore reads in de novo sequencing of bacterial genomes. We compared the differences in both genome sequences between genome assemblies using nanopore and PacBio reads and focused on the difference in the prediction of coding sequences. The results indicated that for more accurate predictions of open reading frames, contigs in the assemblies using only PacBio reads also needed to be corrected using short reads with high-quality bases, and repeat regions in genomes did not affect the increase of mispredicted coding sequences via genome polishing significantly. In assemblies using only nanopore reads, genome polishing was essential, but many repeat regions in genomes might increase the number of mispredicted coding sequences via genome polishing. The hybrid assembly combining the long reads and short reads represents the best result for coding sequence predictions in genome assemblies using nanopore reads.

Keywords Hybrid assembly · Long-read sequencing · Oxford Nanopore technology · *Paenibacillus* sp.

Introduction

The development of long-read (LR) sequencing or third-generation sequencing methods is overcoming the early limitations of short-read sequencing accelerating their application in microbial genomics. Single molecule real-time (SMRT) technology from Pacific Bioscience (PacBio) is the representative sequencing technology used in LR sequencing (Eid et al. 2009) and has been used for complete genome sequencing of many bacterial strains (Chin et al. 2013). The accuracy of each base in raw sequencing reads is known to be nearly 85% (Ross et al. 2013). Recently, another LR technology, Oxford Nanopore Technology (ONT), emerged as a sequencing service and research tool (Deschamps et al. 2016; Eccles et al. 2018; Giordano et al. 2017; Jain et al. 2018, 2016). The SMRT technology detects the signals departed from elongated bases by polymerases (Eid et al. 2009), whereas the ONT detects the differences in the electric signals when nucleotides pass through pore proteins (Clarke et al. 2009). Earlier researches have shown the potential of ONT to generate longer reads and produce more sequencing reads than that of SMRT at a cheaper cost (Giordano et al. 2017). A disadvantage of nanopore reads

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s13205-020-02474-0) contains supplementary material, which is available to authorized users.

✉ Seung Chul Shin
ssc@kopri.re.kr

✉ Han-Woo Kim
hwkim@kopri.re.kr

1 Division of Life Sciences, Korea Polar Research Institute (KOPRI), Inchon 21990, Republic of Korea

2 Unit of Polar Genomics, Korea Polar Research Institute (KOPRI), Inchon 21990, Republic of Korea

3 Department of Polar Sciences, University of Science and Technology, Inchon 21990, Republic of Korea

4 Graduate School of International Agricultural Technology, Seoul National University, Pyeongchang 25354, Republic of Korea

5 Institutes of Green Bio Science and Technology, Seoul National University, Pyeongchang 25354, Republic of Korea

is the lower base quality compared to that of PacBio reads generated using SMRT technology (Ashton et al. 2015; Lu et al. 2016). However, nanopore reads are sufficient to generate scaffolds and unravel complicated structures, similar to PacBio reads. ONT can also generate ultra-long reads, with lengths of up to 882 kb (Jain et al. 2018). The lower base quality of the assembled contigs than that of the assembled contigs from the PacBio reads (Jain et al. 2018) can be curated through polishing of the assembled contigs with the signal level data of the nanopore reads using 'nanopolish' (Loman et al. 2015) or with high quality reads such as Illumina reads using Pilon or Racon (Michael et al. 2018; Walker et al. 2014). The hybrid assemblies that combine the short-read and long-read sequencing datasets have emerged as one of the promising approaches to generating accurate bacterial genome assemblies. The tools like 'Unicycler' and 'SPAdes' have been reported to be the emerging hybrid assemblers used in the assembly of the complete bacterial genome (Antipov et al. 2016; Wick et al. 2017). In eukaryotes, the prediction of the coding sequence (CDS) from the polished genome assembled with the nanopore reads showed that the genome completeness and gene set completeness in BUSCO analysis were greatly increased (Shin et al. 2019). In bacterial genome assemblies with the nanopore reads, a few studies have shown that genome polishing and hybrid assembly could also increase genome completeness and correct the predictions of the CDS with the group of bacterial genome (De Maio et al. 2019; Goldstein et al. 2019; Passera et al. 2018). In this study, focusing on the coding sequence in assemblies using the nanopore reads, we tried to decipher the best assembly method for a single bacterial genome. We sequenced the *Paenibacillus* sp. R4, a strain isolated from the soils of the Arctic region that has been recently used for crystallization of a recombinant form of PsEst3, a psychrophilic esterase (Kim et al. 2018). For the purpose, we have used the SMRT technology, ONT, and Illumina sequencing technology, and compared the annotated genomes generated from the various methods.

Methods

Sample and DNA preparation

Paenibacillus sp. R4 was isolated from the soil of the active layer in Council, Alaska. The soil sample was preserved at – 80 °C until the use. For bacterial isolation, a serially diluted aliquot (100 µl) of the soil sample in 0.85% NaCl (w/v) was spread on Reasoner's 2A (Difco, Sparks, MD, USA) plates and incubated at 10 °C for 20 days. *Paenibacillus* sp. R4 was one of the bacterial isolates and maintained routinely on trytone soy broth (TSB; Himedia, Mumbai, India). Genomic DNA was isolated using DNeasy tissue and blood kit (Qiagen, Valencia, CA, USA) according to the manufacturer's protocol.

ONT library preparation and 1D sequencing

A genomic library for ONT sequencing was constructed using the ONT 1D ligation sequencing kit (SQK-LSK108), according to the manufacturer's instructions. To reduce the damage to the DNA that causes the sequencing errors, 2.0 µg of genomic DNA was repaired using a NEBNext FFPE repair mix (NEB cat no. M6630) and purified using AMPure XP beads. For end repair and dA-tailing, the resulting DNA was treated using NEBNext Ultra II End-Repair/dA-tailing module (NEB cat no. E7546) and purified using AMPure XP beads. An adapter was ligated for sequencing to the purified DNA using an adapter mix 1D in an SQK-LSK108 kit and an NEB Blunt/TA ligase Master Mix (NEB cat no. M0367). Finally, the adaptor-ligated DNA was cleaned using AMPure XP beads, an ABB buffer, and an elution buffer.

Sequencing was carried out using a GridION X5 sequencer, and a single 1D flow cell (FLO-MIN106) with protein pores R9.4 1D chemistry for 48 h, according to the manufacturer's instructions. Live base-calling was performed using Guppy software (ver. 0.5.1), and the FAST5 files were generated during sequencing. All sequencing procedures were performed by Phyzen Co. Ltd. (Seongnam, Korea). Nanopore reads that were similar to the number of bases in the PacBio reads were compared with each other and were randomly selected using seqtk (v. 1.3) (seqtk. https://github.com/lh3/seqtk Accessed 26 August 2019.).

PacBio library preparation and sequencing

A total of 5 µg of genomic DNA was used for the construction of a 20-kb insert library. The SMRTbell library for the PacBio RS II (Pacific Biosciences) was constructed with SMRTbell™ Template Prep Kit 1.0 (PN 100-259-100) following the manufacturer's instructions (Pacific Biosciences). The fragments smaller than 20 kb of the SMRTbell template were removed using the Blue Pippin Size selection system. The 20-kb insert library was sequenced using 1 SMRT cells (Pacific Biosciences) using C4 chemistry (DNA sequencing Reagent 4.0) at LabGenomics (Seongnam, Korea).

Illumina sequencing

The Illumina-compatible sequencing library had a fragment size range of 500 bp. It was constructed and sequenced using the Illumina HiSeq (IH) platform with 150 paired-end chemistry by Phyzen Co. Ltd. (Seongnam, Korea). Trimmomatic (v. 0.36) was used to remove Illumina adapters and trim

low-quality regions with an average Phred score < 15 over a four-bp window (Bolger et al. 2014).

De novo genome assembly and genome polishing

Three assemblers were used in this study for constructing de novo genome sequences. Four sets of sequencing reads were used in the assemblies: PacBio reads (LR^{PB}), Illumina short reads (SR^{IH}), Nanopore reads (LR^{ONTt}), and randomly selected Nanopore reads (LR^{ONTs}). For hybrid assembly, Unicycler (v. 0.4.3) was used (Wick et al. 2017). Each of the three sets of long reads was assembled with SR^{IH} and generated three assemblies: $LR^{ONTt} + SR^{IH}{}_{(Unicycler)}$, $LR^{ONTs} + SR^{IH}{}_{(Unicycler)}$, and $LR^{PB} + SR^{IH}{}_{(Unicycler)}$. For long read only assemblies, Canu (ver. 1.1.1) and SMARTdenovo were used (Koren et al. 2017; SMARTdenovo. https://github.com/ruanjue/smartdenovo. Accessed 19 November 2018.). In the Canu assembly, corrections, trimmings, and assembly phases were performed with default parameters and with "genome size = 8.9 m". "-pacbio-raw" and "-nanopore-raw" options were used for each read set and three assemblies were generated: $LR^{PB}{}_{(canu)}$, $LR^{ONTt}{}_{(canu)}$ and $LR^{ONTs}{}_{(canu)}$. In SMARTdenovo assembly, "-p pac" for PacBio reads and "-p ont" for nanopore read were used, and three assemblies were generated: $LR^{PB}{}_{(SMART)}$, $LR^{ONTt}{}_{(SMART)}$, $LR^{ONTs}{}_{(SMART)}$. canuSMART was the assembly method that assembled initial corrected reads using Canu into contigs using the SMARTdenovo assembler (Schmidt et al. 2017): $LR^{PB}{}_{(canuSMART)}$, $LR^{ONTt}{}_{(canuSMART)}$, and $LR^{ONTs}{}_{(canuSMART)}$. For long read assemblies, genome polishing was performed. SR^{IH} was aligned using Minimap2 (Li 2018) and Burrows-Wheeler Aligner (BWA; ver. 0.7.17) (Li 2013), and the obtained information was used for genome polishing using Racon and Pilon (ver. 1.22) with default parameters.

Assembly evaluation and the identity values of the draft genome sequences

REAPR (recognition of errors in assemblies using paired reads) (Hunt et al. 2013) was used for assembly evaluation. REAPR is a tool that precisely identifies errors in genome assemblies and provides a warning for less serious inconsistencies in the assembly through paired-end read mapping without a reference genome (Hunt et al. 2013). The identity values of the assemblies were computed based on the $LR^{PB} + SR^{IH}{}_{(Unicycler)}$ using the nucmer command in the MUMmer tool (ver. 3.0.) (Delcher et al. 2002). The resulting delta file was processed with the dnadiff script in the MUMmer tool, and an average 1-to-1 alignment identity, total indel, and total substitution were used. The genome completeness of assemblies was also validated using benchmarking universal single-copy orthologs (BUSCO; ver. 3) (Simão et al. 2015). Because *Paenibacillus* is a species of bacteria within the order *Bacillales*, we conducted BUSCO analyses against *Bacillales* datasets containing 526 genes. CheckM was also used for estimation of genome completeness and contamination using their domain-specific markers (bacteria: 104 markers) (Parks et al. 2015).

Gene annotation and quality assessments

We carried out gene annotations using a prokaryotic genome annotation pipeline, DFAST (Tanizawa et al. 2017), and Clusters of orthologous groups of proteins (COGs) were predicted by COGnitor (Tatusov et al. 2000). To set the coding sequence before comparison between the coding sequence (CDS), ORFs of $LR^{PB} + SR^{IH}{}_{(Unicycler)}$ were considered as CDSs of *Paenibacillus* sp. R4 and CDSs of $LR^{PB} + SR^{IH}{}_{(Unicycler)}$ were compared using BLAST against ORFs of $LR^{PB}{}_{(canu)}$, and vice versa. CDSs, which were considered as a misprediction, were confirmed by comparison against the nr databases. If single ORFs of other assemblies showed perfect or nearly exact matches (above 99% identity and below 5% difference in length) with single CDS of $LR^{PB} + SR^{IH}{}_{(Unicycler)}$, it was thought to be the same CDS. If multiple ORFs of other assemblies showed nearly exact matches with single CDS of $LR^{PB} + SR^{IH}{}_{(Unicycler)}$ and are located next to each other in their genome, they were thought to be split by the error in assemblies. If there are no matched CDS against CDS of $LR^{PB} + SR^{IH}{}_{(Unicycler)}$, CDS was thought to be missing in the assemblies.

Repeat analysis

To identify repeat regions in the genome sequences, self-BLAST was performed using the BLASTN program. Sequences with the similarity over 99% and the length over 500 bp in BLASTN were considered as repeat sequences, and these regions were compared to the region of the split for the CDS against the repeat regions of the genome.

Results

Sequencing and read correction

We sequenced the genome of *Paenibacillus* sp. R4 using GridION X5, PacBio RS II, and Illumina HiSeq. A GridION X5 sequencer generated 2 262 281 reads (LR^{ONTt}) bearing 15 330 789 967 bases, using a single 1D flow cell (Table 1). The LR^{ONTt} that were longer than 1 kb comprised 98.4% of the total read sum, and the reads longer than 10 kb comprised 60.46%. PacBio RS II generated 149 031 reads (LR^{PB}) bearing 1 438 734 509 bases using one SMRT cell (Pacific Biosciences). LR^{PB} longer than 1 kb comprised 99.7% of total read sum and reads longer than 10 kb, comprised 70.4%.

Table 1 Characteristics of sequencing reads

	LR^{ONTt}	LR^{ONTs}	LR^{PB}	SR^{IH}
Total reads	2,262,281	215,000	157,471	10,314,098
Total bases	15,330,789,967	1,456,295,612	1,438,734,509	1,557,428,798
Read length N50 (bp)	12,625	12,589	13,760	151
Max read length (bp)	121,377	113,620	45,888	151
Number above 1 kbp	1,903,181 (98.64%)	180,756 (98.64%)	149,031 (99.67%)	
Number above 5 kbp	990,595 (82.83%)	93,969 (82.78%)	101,928 (90.63%)	
Number above 10 kbp	513,149 (60.46%)	48,713 (60.42%)	62,688 (70.39%)	
Number above 25 kbp	81,947 (17.38%)	7771 (17.44%)	4501 (9.12%)	

bp base pairs. LR^{ONTt} denotes long reads generated from Oxford Nanopore Technology (ONT), LR^{ONTs} denotes randomly selected reads among LR^{ONTt}, LR^{PB} denotes long reads generated from single molecule real-time (SMRT) technology, and SR^{IH} denotes short reads generated from Illumina HiSeq 2000. The percentage of read sum is shown in parentheses

10 314 098 short reads bearing 1 557 428 798 bases (SR^{IH}) were obtained using HiSeq 2000. To compare the assembled genome sequences using LR^{ONT} against those using LR^{PB}, 215 000 nanopore reads (LR^{ONTs}) comprising 1 456 295 612 bases were randomly selected using seqtk (ver. 1.3).

Before assembly, the long reads were corrected using Canu (ver. 1.1.1) (Table 2) (Koren et al. 2017). Canu was set to select the longest 40 × subset and generate 40 × corrected reads by default. Therefore, 13 427 reads comprising 363 678 567 bases in LR^{ONTs} were generated after correction. In LR^{PB}, 22 626 corrected reads comprising 349 109 761 bases were generated, and the percentage of reads over 10 kbp increased for both corrected LR^{PB} and LR^{ONTs}. However, a large number of reads were selected from LR^{ONTt} and corrected by Canu. The number of corrected reads in LR^{ONTt} was 646 153, comprising 5 402 055 529 bases.

De novo genome assembly and comparison

To obtain accurate genome sequences from the LRs (LR^{ONTt}, LR^{ONTs}, and LR^{PB}), hybrid assembly (Wick et al. 2017), long read only assembly (Koren et al. 2017; Schmidt et al. 2017; SMARTdenovo. https://github.com/ruanjue/smartdenovo. Accessed 19 November 2018.), and genome polishing were used in this study. In all assemblies, one contig comprising about 8.9 Mbp with 46.5% G + C content was generated (Table 3). In the assemblies using LR^{PB}, four assemblies were generated: $LR^{PB} + SR^{IH}{}_{(Unicycler)}$, $LR^{PB}{}_{(canu)}$, $LR^{PB}{}_{(SMART)}$, and $LR^{PB}{}_{(canuSMART)}$. $LR^{PB}{}_{(canu)}$, $LR^{PB}{}_{(SMART)}$, and $LR^{PB}{}_{(canuSMART)}$ among assemblies were polished with SR^{IH} by using the programs Pilon (ver. 1.22) and Racon (ver. 1.4.3) (Table 3). The resulting genome size of *Paenibacillus* sp. R4 was 8,989,550–8,992,906 bases in length. For LR^{ONTs}, assemblies and genome polishing were performed similarly to the LR^{PB} assemblies: $LR^{ONTs} + SR^{IH}{}_{(Unicycler)}$, $LR^{ONTs}{}_{(canu)}$, $LR^{ONTs}{}_{(SMART)}$, and $LR^{ONTs}{}_{(canuSMART)}$. For LR^{ONTt}, the following four assemblies were performed: $LR^{ONTt} + SR^{IH}{}_{(Unicycler)}$, $LR^{ONTt}{}_{(canu)}$, $LR^{ONTt}{}_{(SMART)}$, and $LR^{ONTt}{}_{(canuSMART)}$ (Table 3). The resulting genome sizes were 8,966,533–8,989,554 bases. To evaluate the accuracy of the genome assemblies, we used REAPR (Table 3); (Hunt et al. 2013). REAPR reports a warning for less serious inconsistencies in the assembly through paired-end read mapping (Hunt et al. 2013). A small deletion or insertion error, incorrect assembly of a repetitive sequence, or a region with low coverage of read pairs were included in the warning. Genome sequences using a hybrid assembler with SR^{IH} and long read assemblies with LR^{PB} showed a low number of warnings (below 331) without genome polishing. In contrast, assemblies with only LR^{ONT} showed a high

Table 2 Characteristics of corrected reads

	Corrected LR^{ONTt}	Corrected LR^{ONTs}	Corrected LR^{PB}
Total reads	646,153	13,427	22,626
Total bases	5,402,055,529	363,678,567	349,109,761
Read length N50 (bp)	12,896	27,688	18,207
Max read length (bp)	114,802	80,640	41,647
Number above 1 kbp	641,746 (99.92%)	13,411 (99.96%)	22,450 (99.96%)
Number above 5 kbp	372,271 (86.26%)	12,870 (99.70%)	18,290 (97.69%)
Number above 10 kbp	199,085 (63.05%)	12,869 (99.70%)	18,083 (97.22%)
Number above 25 kbp	13,664 (9.67%)	7341 (65.48%)	1820 (14.96%)

The percentage of read sum is shown in parentheses

Table 3 Evaluation of the genome sequence in each assembly

Long read	Assembly	Genome size	GC Content (%)	Error-free bases (%)	Warnings							
					Total	Low score regions	Links	Soft clip	Collapsed repeats	Low read coverage	Low perfect coverage	Wrong read orientation
LR^{PB}	$LR^{PB} + SR^{IH}$ (Unicycler)	8,989,550	45.6	96.43	159	2	0	0	0	2	154	1
	LR^{PB} (canu)	8,989,890	45.6	96.54	194	1	0	0	0	6	183	4
	LR^{PB} (canu+Pilon×3)	8,990,111	45.6	96.56	186	1	0	0	0	3	177	5
	LR^{PB} (canu+Racon+Pilon×3)	8,992,552	45.6	97.43	462	98	0	42	2	95	212	13
	LR^{PB} (SMART)	8,989,802	45.6	96.33	331	8	0	12	0	22	280	9
	LR^{PB} (SMART+Pilon×3)	8,990,096	45.6	96.55	202	4	0	0	0	15	179	4
	LR^{PB} (SMART+Racon+Pilon×3)	8,992,407	45.6	97.44	467	104	0	39	2	94	214	14
	LR^{PB} (canuSMART)	8,990,095	45.6	96.50	203	4	0	11	0	5	181	2
	LR^{PB} (canuSMART+Pilon×3)	8,990,152	45.6	96.56	181	0	0	2	0	2	177	0
	LR^{PB} (canuSMART+Racon+Pilon×3)	8,992,906	45.6	97.42	499	110	3	56	1	109	210	10
LR^{ONT}	$LR^{ONTt} + SR^{IH}$ (Unicycler)	8,988,896	45.6	96.42	157	0	0	0	0	1	151	5
	$LR^{ONTs} + SR^{IH}$ (Unicycler)	8,989,550	45.6	96.41	156	4	0	0	0	2	149	1
	LR^{ONTt} (canu)	8,967,236	45.6	90.89	4037	35	1	13	15	92	3861	20
	LR^{ONTs} (canu)	8,966,533	45.6	90.64	4238	39	3	30	13	94	4023	36
	LR^{ONTs} (canu+Pilon×3)	8,989,264	45.6	96.53	318	18	0	15	3	63	204	15
	LR^{ONTs} (canu+Racon+Pilon×3)	8,989,554	45.6	97.47	470	94	1	32	3	98	221	21
	LR^{ONTt} (SMART)	8,977,120	45.6	93.85	2226	23	1	13	13	97	2061	18
	LR^{ONTs} (SMART)	8,973,338	45.6	92.90	2682	46	3	30	23	105	2434	41
	LR^{ONTs} (SMART+Pilon×3)	8,987,851	45.6	96.38	355	18	0	15	13	84	210	15
	LR^{ONTs} (SMART+Racon+Pilon×3)	8,989,554	45.6	97.47	470	94	1	32	3	98	221	21
	LR^{ONTt} (canuSMART)	8,971,414	45.6	91.98	3437	34	1	27	7	92	3260	16
	LR^{ONTs} (canuSMART)	8,967,702	45.6	91.38	3817	31	3	6	15	99	3636	27
	LR^{ONTs} (canuSMART+Pilon×3)	8,988,158	45.6	96.54	305	13	0	2	5	74	206	5
	LR^{ONTs} (canuSMART+Racon+Pilon×3)	8,989,326	45.6	97.45	481	96	1	38	8	107	211	20

Summary of REAPR (recognition of errors in assemblies using paired reads) results. REAPR outputs a warning for less serious inconsistencies in the assembly; Low score regions denote the region having a score below 0.5, a score from zero to one is assigned to every base of the assembly by REAPR. Links indicate that a significant proportion of the reads in this region mapped elsewhere in the assembly. Low perfect coverage denotes a low coverage (under 5) of perfect uniquely mapping reads in this region. Genome sequences without genome polishing are shown in black, those polished using Pilon in blue, and those polished with Racon + Pilon × 3 in red

number of warnings (above 2 226) (Table 3). More nanopore reads slightly decreased the number of warning for the $LR^{ONTt}_{(canu)}$, $LR^{ONTt}_{(SMART)}$, and $LR^{ONTt}_{(canuSMART)}$, than in the $LR^{ONTs}_{(canu)}$, $LR^{ONTs}_{(SMART)}$, and $LR^{ONTs}_{(canuSMART)}$, respectively. In most assemblies using only LR, the number of warnings decreased after genome polishing (Table 3 and Supplementary Table S1), and the polished genome sequences with Pilon × 3 showed the lowest number of warnings among the long read only assembly (Table 3). However, in polished genome sequences using both Pilon × 3 and Racon, the number of warnings in assemblies using LR increased compared to genome polishing with only Pilon × 3.

To evaluate the usages of the nanopore reads for bacterial genome assembly and further genome annotations, we compared the assemblies using LR^{ONT} against the assemblies with those using LR^{PB}. $LR^{PB} + SR^{IH}_{(Unicycler)}$was selected as the base assembly because it had the lowest number of warnings without genome polishing among assemblies using LR^{PB} (Table 3). Through dnadiff (Delcher et al. 2002), the assemblies were compared against $LR^{PB} + SR^{IH}_{(Unicycler)}$ (Table 4). Identities of the assemblies using the nanopore reads ranged from 99.732% to 99.990%. The hybrid assemblies ($LR^{ONTt} + SR^{IH}_{(Unicycler)}$ and $LR^{ONTs} + SR^{IH}_{(Unicycler)}$) were highly similar to $LR^{PB} + SR^{IH}_{(Unicycler)}$, and assemblies using only LR^{ONT} showed lower identities (99.732–99.829%) than those of the hybrid assemblies (99.988–99.990%), without genome polishing. The identities of the assemblies using LR^{PB} ranged from 99.962% to 99.989%, and those of assemblies without genome polishing ranged from 99.984 ~ 99.987%. After genome polishing using Pilon × 3, the identities of most of the assemblies increased to over 99.983% (Table 4). However, the identities of the genome sequence polished using Racon and Pilon × 3 increased up to 99.962% in the assemblies using nanopore reads. Rather, in the assemblies using the PacBio reads, the identities decreased to 99.955% after genome polishing via Racon and Pilon × 3. Total substitutions in the results of dnadiff ranged from 460 to 1 011 bases in the assemblies using LR^{ONT} and

Table 4 Assembly quality assessment using dnadiff and CheckM

Long read	Assembly	Identity (%)	Total substitutions	Total indels	Completeness (%)	Contamination (%)
LR^{PB}	$LR^{PB} + SR^{IH}_{(Unicycler)}$				98.28 (101)	3.45 (2)
	$LR^{PB}_{(canu)}$	99.987	168	270	98.28 (101)	3.45 (2)
	$LR^{PB}_{(canu+Pilon\times3)}$	99.987	167	41	98.28 (101)	3.45 (2)
	$LR^{PB}_{(canu+Racon+Pilon\times3)}$	99.962	396	2446	91.93(95)	3.45 (2)
	$LR^{PB}_{(SMART)}$	99.984	174	1078	98.28 (101)	3.45 (2)
	$LR^{PB}_{(SMART+Pilon\times3)}$	99.988	133	105	98.28 (101)	3.45 (2)
	$LR^{PB}_{(SMART+Racon+Pilon\times3)}$	99.955	408	2804	90.52 (96)	1.88 (2)
	$LR^{PB}_{(canuSMART)}$	99.986	244	543	97.93 (100)	3.45 (2)
	$LR^{PB}_{(canuSMART+Pilon\times3)}$	99.989	179	52	98.28 (101)	3.45 (2)
	$LR^{PB}_{(canuSMART+Racon+Pilon\times3)}$	99.989	333	2398	90.49 (93)	3.61 (3)
LR^{ONT}	$LR^{ONTt} + SR^{IH}_{(Unicycler)}$	99.988	56	6	98.28 (101)	3.45 (2)
	$LR^{ONTs} + SR^{IH}_{(Unicycler)}$	99.99	44	6	98.28 (101)	3.45 (2)
	$LR^{ONTt}_{(canu)}$	99.734	460	22,908	90.75 (91)	1.72 (1)
	$LR^{ONTs}_{(canu)}$	99.732	508	23,293	89.34 (94)	1.72 (1)
	$LR^{ONTs}_{(canu+Pilon\times3)}$	99.983	181	475	98.28 (101)	3.45 (2)
	$LR^{ONTs}_{(canu+Racon+Pilon\times3)}$	99.962	453	2080	92.24 (98)	1.72 (1)
	$LR^{ONTt}_{(SMART)}$	99.829	592	14,198	90.88 (92)	1.72 (1)
	$LR^{ONTs}_{(SMART)}$	99.798	1011	16,867	95.53 (99)	1.72 (1)
	$LR^{ONTs}_{(SMART+Pilon\times3)}$	99.984	349	563	95.69 (99)	3.45 (2)
	$LR^{ONTs}_{(SMART+Racon+Pilon\times3)}$	99.964	535	2453	95.69 (99)	3.45 (2)
	$LR^{ONTt}_{(canuSMART)}$	99.761	566	19,556	90.44 (91)	1.72 (1)
	$LR^{ONTs}_{(canuSMART)}$	99.757	523	20,782	90.36 (97)	0 (0)
	$LR^{ONTs}_{(canuSMART+Pilon\times3)}$	99.986	165	247	98.28 (101)	3.45 (2)
	$LR^{ONTs}_{(canuSMART+Racon+Pilon\times3)}$	99.965	446	2085	91.07 (95)	1.72 (1)

Identity, total substitutions, total indels, translocations, and inversions were calculated using dnadiff. Total indels denote the sum of single nucleotide insertions and deletions in the aligned region. Unaligned bases were not included in this table, and rough insertions and deletions calculated with dnadiff are not shown in the table. CheckM was used to estimate the completeness and level of contamination. One hundred four markers were tested, and the numbers of the markers are indicated in parentheses. Genome sequences without genome polishing are shown in black, those polished using Pilon in blue, and those polished with Racon + Pilon × 3 in red

ranged from 168 to 244 bases in the assemblies using LR^{PB}, before genome polishing. With three times the genome polishing using Pilon, the total substitutions decreased to 165–349 bases in the assemblies using the nanopore reads and decreased to 133–179 bases in the assemblies using the PacBio reads. In genome polishing with Racon and Pilon × 3, the total substitutions slightly decreased to 446–535 bases in the assemblies using LR^{ONT}, but it increased to 333–408 bases in the assemblies using LR^{PB}. Total indels, denoting single inserted bases and deleted bases in the dnadiff results, decreased from 14 198 ~ 23 293 to 247 ~ 563 bases in the assemblies using LR^{ONT} and decreased from 270 ~ 1 078 to 41 ~ 105 bases in the assemblies using LR^{PB} after genome polishing using Pilon × 3. However, genome polishing using both Racon and Pilon × 3 increased the total indels up to 2398–2804. In genome polishing of assemblies using LR^{ONT}, Racon increased the number of both total substitutions and indels compared to those of genome polishing using only Pilon × 3. Total bases, identities, total substitutions, and total indels of the hybrid assemblies with LR^{ONT} were most similar to $LR^{PB} + SR^{IH}_{(Unicycler)}$.

The genome completeness of assemblies was also validated using benchmarking universal single-copy orthologs (BUSCO; ver. 3) (Simão et al. 2015), and CheckM (Parks et al. 2015). We conducted BUSCO analyses against *Bacillales* datasets containing 526 genes (Fig. 1) and assessed genome quality with CheckM against 104 maker genes (Table 4). The BUSCO genome completeness of hybrid assemblies was the same at 98.48%. Assemblies using LR^{PB} without genome polishing showed 97.91–98.29% genome completeness, and the BUSCO genome completeness increased up to 98.29–98.48% after genome polishing. In the case of assemblies using LR^{ONT} without genome polishing, the BUSCO genome completeness ranged from 35.74 to 49.62%, which increased to 98.29–98.48% after genome polishing. No difference was observed between assemblies using Pilon and assemblies using both Racon and Pilon in the BUSCO analysis. The CheckM analysis also showed similar results of the genome completeness as of the hybrid assemblies, which was 98.28%, the highest among the assemblies (Table 4). In assemblies using LR^{PB}, genome polishing with Racon decreased the genome completeness, whereas, in assemblies using LR^{ONT} without genome polishing, the genome completeness ranged from 89.34% to 95.53%, which increased to 91.07–98.28% after genome polishing.

ORF predictions and comparisons

The bacterial genome annotation was performed using DFAST (Tanizawa et al. 2017). The open reading frame (ORF) predicted in the other assemblies were compared with the CDSs of the $LR^{PB} + SR^{IH}_{(Unicycler)}$, which was selected as the base assembly in the sequence comparison. There were 8

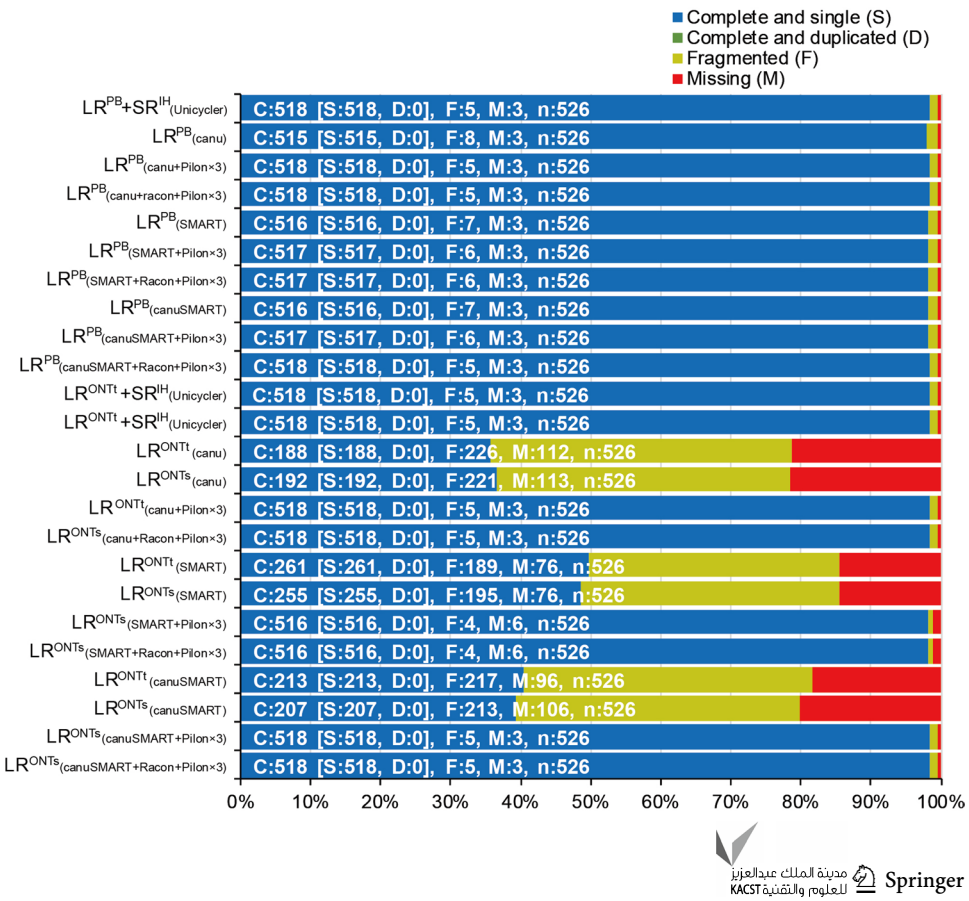

Fig. 1 Benchmarking Universal Single-Copy Orthologs (BUSCO) analyses against *Bacillales* datasets. *Paenibacillus* sp. R4 is a species of bacteria within the order *Bacillales*. The *Bacillales odb9* dataset, comprising 526 genes, was used in the BUSCO analysis

186 ORFs, 31 rRNAs (11 copies of 5S rRNA gene, 10 copies of 16S rRNA gene, and 10 copies of 23S rRNA gene), and 83 tRNAs that were predicted in the $LR^{PB} + SR^{IH}_{(Unicycler)}$ (Fig. 2 and Table 5). In the $LR^{PB}_{(canu)}$, 108 more ORFs were predicted than the $LR^{PB} + SR^{IH}_{(Unicycler)}$. To compare the differences between the two assemblies, ORFs of the $LR^{PB}_{(canu)}$ were searched against the ORF of the $LR^{PB} + SR^{IH}_{(Unicycler)}$ using the BLASTN program. To identify the ORF with the correct CDS, the ORFs showing differences in their sequences were searched against the nr database using the BLAST program, and we found that one CDS of $LR^{PB} + SR^{IH}_{(Unicycler)}$ was split into two ORFs. There were 105 CDS of $LR^{PB} + SR^{IH}_{(Unicycler)}$ that were split into 213 ORFs in $LR^{PB}_{(canu)}$, comprising 2.57% of the total CDS and 13 CDS of the $LR^{PB} + SR^{IH}_{(Unicycler)}$ were missing from the ORF of the $LR^{PB}_{(canu)}$ (Figs. 3 and 4). In the $LR^{PB}_{(SMART)}$, there were 362 more ORFs predicted, 684 of the ORFs were matched to 307 CDSs, and 26 CDS were missing. In the $LR^{PB}_{(canuSMART)}$, there were 73 more ORF predicted, 152 ORF were matched to 76 CDSs, and 8 CDSs of $LR^{PB} + SR^{IH}_{(Unicycler)}$ showed no match with the ORFs of $LR^{PB}_{(canuSMART)}$. Differently predicted ORF numbers, compared with the CDS of the $LR^{PB} + SR^{IH}_{(Unicycler)}$, showed similar patterns to the number of warnings from REAPR (Table 3). After genome polishing with Pilon, the differences in the number of ORFs between the $LR^{PB} + SR^{IH}_{(Unicycler)}$ and the other assemblies using LR^{PB} were diminished below 25. There were 21 ORFs of $LR^{PB}_{(canu+Pilon\times3)}$ that were matched to the 10 CDSs of the $LR^{PB} + SR^{IH}_{(Unicycler)}$, 31 ORFs of $LR^{PB}_{(SMART+Pilon\times3)}$ were matched to 15 CDSs, and 18 ORFs of $LR^{PB}_{(canuSMART+Pilon\times3)}$ were matched to 10 CDSs. The ORFs that were not matched to CDSs of $LR^{PB} + SR^{IH}_{(Unicycler)}$ were 31 in the $LR^{PB}_{(canu+Pilon\times3)}$, 14 in the $LR^{PB}_{(SMART+Pilon\times3)}$, and 20 in the $LR^{PB}_{(canuSMART+Pilon\times3)}$

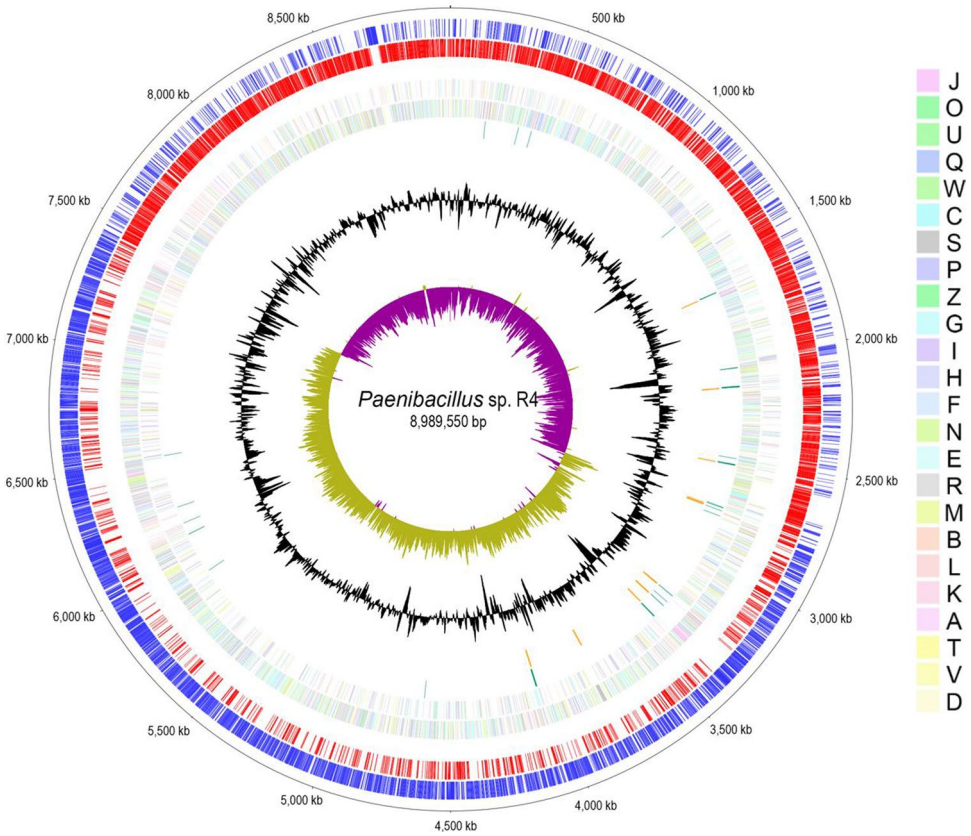

Fig. 2 Circular map of the *Paenibacillus* sp. R4 genome. Labeling from the outside to the center: Genes on the forward strand, genes on the reverse strand, RNA genes (rRNAs in orange, tRNAs in green), GC content (black), and GC skew (olive/purple). Individual genes are colored by COG categories: J (translation, ribosomal structure, and biogenesis), A (RNA processing and modification), K (transcription), L (replication, recombination, and repair), B (chromatin structure and dynamics), D (cell cycle control, cell division, and chromosome partitioning), Y (nuclear structure), V (defense mechanisms), T (signal transduction mechanisms), M (cell wall/membrane/envelop biogenesis), N (cell motility), Z (cytoskeleton), W (extracellular structures), U (intracellular trafficking, secretion, and vesicular transport), O (posttranslational modification, protein turnover, and chaperones), X (mobilome: prophages and transposons), C (energy production and conversion), G (carbohydrate transport and metabolism), E (amino acid transport and metabolism), F (nucleotide transport and metabolism), H (coenzyme transport and metabolism), I (lipid transport and metabolism), P (inorganic ion transport and metabolism), Q (secondary metabolites biosynthesis, transport, and catabolism), R (general functional prediction only), and S (function unknown)

Table 5 Characteristics of genome sequences in each assembly

Long read	Assembly	ORF	Average protein length	Coding ratio (%)	Number of rRNAs	Number of tRNAs
LR^{PB}	$LR^{PB}+SR^{IH}_{(Unicycler)}$	8186	318.2	87.1	31	83
	$LR^{PB}_{(canu)}$	8294	314.5	87.1	31	83
	$LR^{PB}_{(canu+Pilon\times1)}$	8170	319.4	87.1	31	83
	$LR^{PB}_{(canu+Pilon\times2)}$	8170	319.4	87.1	31	83
	$LR^{PB}_{(canu+Pilon\times3)}$	8170	319.4	87.1	31	83
	$LR^{PB}_{(canu+Racon+Pilon\times3)}$	8262	315.8	87.1	31	83
	$LR^{PB}_{(SMART)}$	8548	304.2	86.8	31	83
	$LR^{PB}_{(SMART+Pilon\times1)}$	8210	317.8	87.1	31	83
	$LR^{PB}_{(SMART+Pilon\times2)}$	8180	318.9	87.1	31	83
	$LR^{PB}_{(SMART+Pilon\times3)}$	8199	318.3	87.1	31	83
	$LR^{PB}_{(SMART+Racon+Pilon\times3)}$	8251	316.1	87	30	83
	$LR^{PB}_{(canuSMART)}$	8259	315.9	87.1	31	83
	$LR^{PB}_{(canuSMART+Pilon\times1)}$	8193	318.5	87.1	31	83
	$LR^{PB}_{(canuSMART+Pilon\times2)}$	8194	318.5	87.1	31	83
	$LR^{PB}_{(canuSMART+Pilon\times3)}$	8191	318.6	87.1	31	83
	$LR^{PB}_{(canuSMART+Racon+Pilon\times3)}$	8209	317.7	87	31	83
LR^{ONT}	$LR^{ONTt}+SR^{IH}_{(Unicycler)}$	8190	318.6	87.1	31	83
	$LR^{ONTs}+SR^{IH}_{(Unicycler)}$	8199	318.3	87.1	31	83
	$LR^{ONTt}_{(canu)}$	13,864	171.4	79.5	31	81
	$LR^{ONTs}_{(canu)}$	13,953	170.5	79.6	31	82
	$LR^{ONTs}_{(canu+Pilon\times1)}$	8267	315.3	87	31	83
	$LR^{ONTs}_{(canu+Pilon\times2)}$	8257	315.8	87	31	83
	$LR^{ONTs}_{(canu+Pilon\times3)}$	8249	316	87	31	83
	$LR^{ONTs}_{(canu+Racon+Pilon\times3)}$	8206	317.6	87	31	83
	$LR^{ONTt}_{(SMART)}$	12,181	202.6	82.5	31	81
	$LR^{ONTs}_{(SMART)}$	12,691	193.3	82	31	80
	$LR^{ONTs}_{(SMART+Pilon\times1)}$	8269	314.7	86.8	31	83
	$LR^{ONTs}_{(SMART+Pilon\times2)}$	8189	317.6	86.8	31	83
	$LR^{ONTs}_{(SMART+Pilon\times3)}$	8191	317.5	86.8	31	83
	$LR^{ONTs}_{(SMART+Racon+Pilon\times3)}$	8258	315.5	87	30	83
	$LR^{ONTt}_{(canuSMART)}$	13,404	180.1	80.7	31	81
	$LR^{ONTs}_{(canuSMART)}$	13,732	174.2	80	31	81
	$LR^{ONTs}_{(canuSMART+Pilon\times1)}$	8240	316.2	87	31	83
	$LR^{ONTs}_{(canuSMART+Pilon\times2)}$	8242	316.4	87	31	83
	$LR^{ONTs}_{(canuSMART+Pilon\times3)}$	8237	316.6	87	31	83
	$LR^{ONTs}_{(canuSMART+Racon+Pilon\times3)}$	8261	315.6	87	30	83

Bacterial genome annotation was performed using DFAST, and the statistics were identified. In assemblies using only nanopore reads, the increase of the number of ORFs was observed. Genome sequences without genome polishing are shown in black, those polished using Pilon in blue, and those polished with Racon + Pilon × 3 in red

(Fig. 3b). In the genome sequences that were polished with both Racon and Pilon × 3, compared with those polished with Pilon × 3, the number of mis-predicted CDSs increased to 60 in the $LR^{PB}_{(canu+Racon+Pilon\times3)}$, 60 in $LR^{PB}_{(SMART+Racon+Pilon\times3)}$, and 59 in the $LR^{PB}_{(canuSMART+Racon+Pilon\times3)}$ (Figs. 3a and 4).

ORFs of the assemblies using LR^{ONT} were also compared against those of $LR^{PB}+SR^{IH}_{(Unicycler)}$. ORFs of the $LR^{ONTt}+SR^{IH}_{(Unicycler)}$ and $LR^{ONTs}+SR^{IH}_{(Unicycler)}$ were similar to the number of CDS (Fig. 3a). Only 3 CDS and 6 CDS of $LR^{PB}+SR^{IH}_{(Unicycler)}$ were split into 6 in $LR^{ONTt}+SR^{IH}_{(Unicycler)}$ and 12 in $LR^{ONTs}+SR^{IH}_{(Unicycler)}$, respectively. Only one CDS was missing in $LR^{ONTs}+SR^{IH}_{(Unicycler)}$, and there were no missing CDSs in the $LR^{ONTt}+SR^{IH}_{(Unicycler)}$ (Figs. 3 and 4). However, the number of predicted ORF were largely increased in other assemblies using LR^{ONT}: $LR^{ONTs}_{(canu)}$, LR^{ONTs}

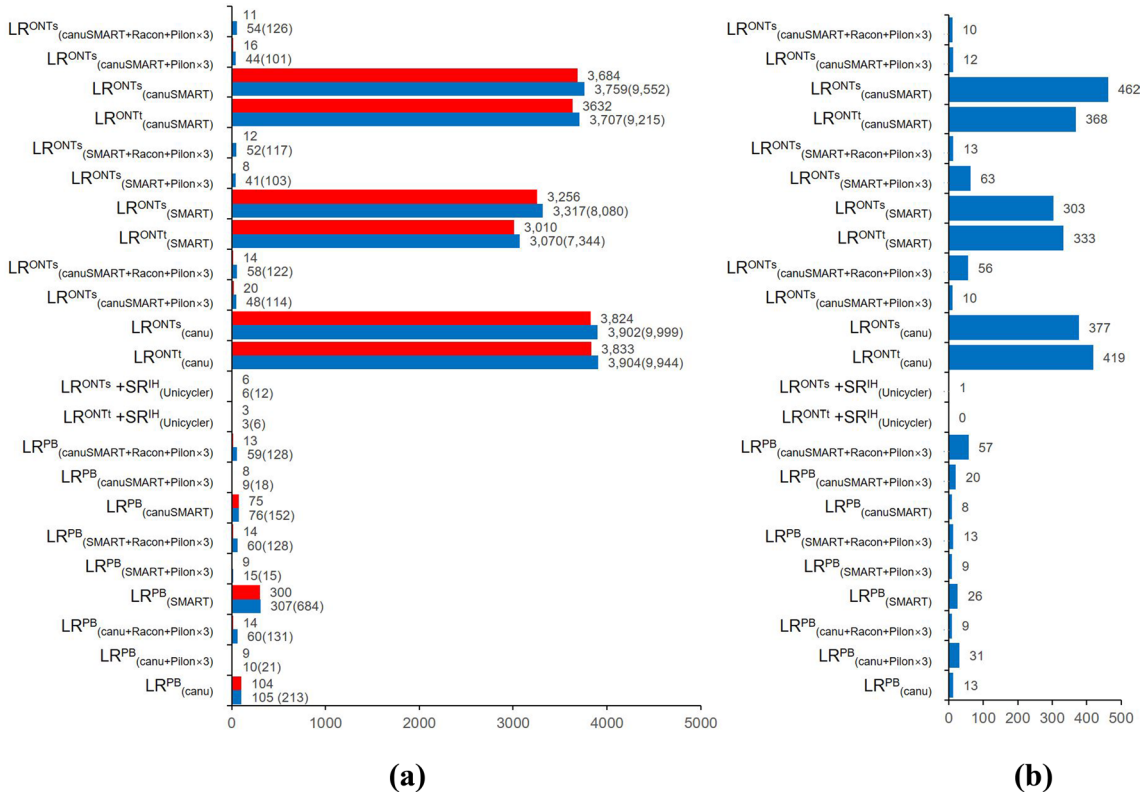

Fig. 3 The number of mispredicted and missing coding sequences (CDS) in the assemblies compared with CDS in $LR^{PB}+SR^{IH}$ (Unicycler). **a** Numbers of mispredicted CDS are shown in blue, and the number of mispredicted CDS not in the repeat regions in red. Racon polishing increased the number of mispredicted CDS in the repeat regions. The mispredicted CDS in the assemblies using LR^{PB} did not seem to be related to repeat regions. The number in parentheses is the number of split ORFs corresponding to the CDS. **b** The number of missing CDS in the assemblies

(SMART), LR^{ONTs} (canuSMART), LR^{ONTt} (canu), LR^{ONTt} (SMART), and LR^{ONTt} (canuSMART). The increasing read number of LR^{ONT} could not diminish the differences of the CDSs when comparing between the assemblies using LR^{ONTt} and assemblies using LR^{ONTs}. In LR^{ONTs} (canu), 9 999 ORFs were matched with 3 902 CDS of $LR^{PB}+SR^{IH}$ (Unicycler), and 9 939 ORFs matched with 3 904 CDS of the $LR^{PB}+SR^{IH}$ (Unicycler) in LR^{ONTt} (canu). There were 377 and 419 CDS of $LR^{PB}+SR^{IH}$ (Unicycler) that were not found in the ORF of LR^{ONTt} (canu) and LR^{ONTs} (canu), respectively. After genome polishing with Pilon × 3 against the assemblies using LR^{ONT} with SR^{IH}, the mis-predicted ORFs were remarkably diminished. The statistics of the predicted ORFs were more similar than those of the assemblies using only the LR^{PB}. There were 114 ORF in the LR^{ONTs} (canu+Pilon×3) that were partially matched to the 48 CDS, 101 ORFs in LR^{ONTs} (canuSMART+Pilon×3) were matched to 44 CDS, and 84 ORFs in LR^{ONTs} (SMART+Pilon×3) were matched to 41 CDS of $LR^{PB}+SR^{IH}$ (Unicycler). When Pilon polishing was performed three times after Racon polishing, in assemblies using LR^{ONT}, 58, 52, and 54 CDS of $LR^{PB}+SR^{IH}$ (Unicycler) were split in the LR^{ONTs} (canu+Racon+Pilon×3), LR^{ONTs} (SMART+Racon+Pilon×3), and LR^{ONTs} (canuSMART+Racon+Pilon×3), respectively.

Error-correction and ORF predictions in the repeat regions

To identify whether genome polishing in repeat regions of this genome affected the mis-prediction of CDSs in these assemblies, we identified repeat regions using BLASTN in $LR^{PB}+SR^{IH}$ (Unicycler) and compared the regions in the split ORF with the repeat regions of the genome. Sequences with similarities of over 99% and lengths of over 500 bp in the self BLASTN were considered as repeat sequences (Table 6). In the assemblies using LR^{PB}, these repeat regions did not largely affect the mispredictions of the CDSs in the genome sequences polished using Pilon (Figs. 3a and 4). One of the ten split CDSs in the LR^{PB} (canu+Pilon×3) and one of the nine split CDS in the LR^{PB} (canuSMART+Pilon×3), and six of the fifteen split CDSs in the LR^{PB} (SMART+Pilon×3), resided in the repeat region. However, 28 of the 48 split CDSs in the LR^{ONTs} (canu+Pilon×3), 33 of the 41 split CDSs in LR^{ONTs} (SMART+Pilon×3), and 28 of the

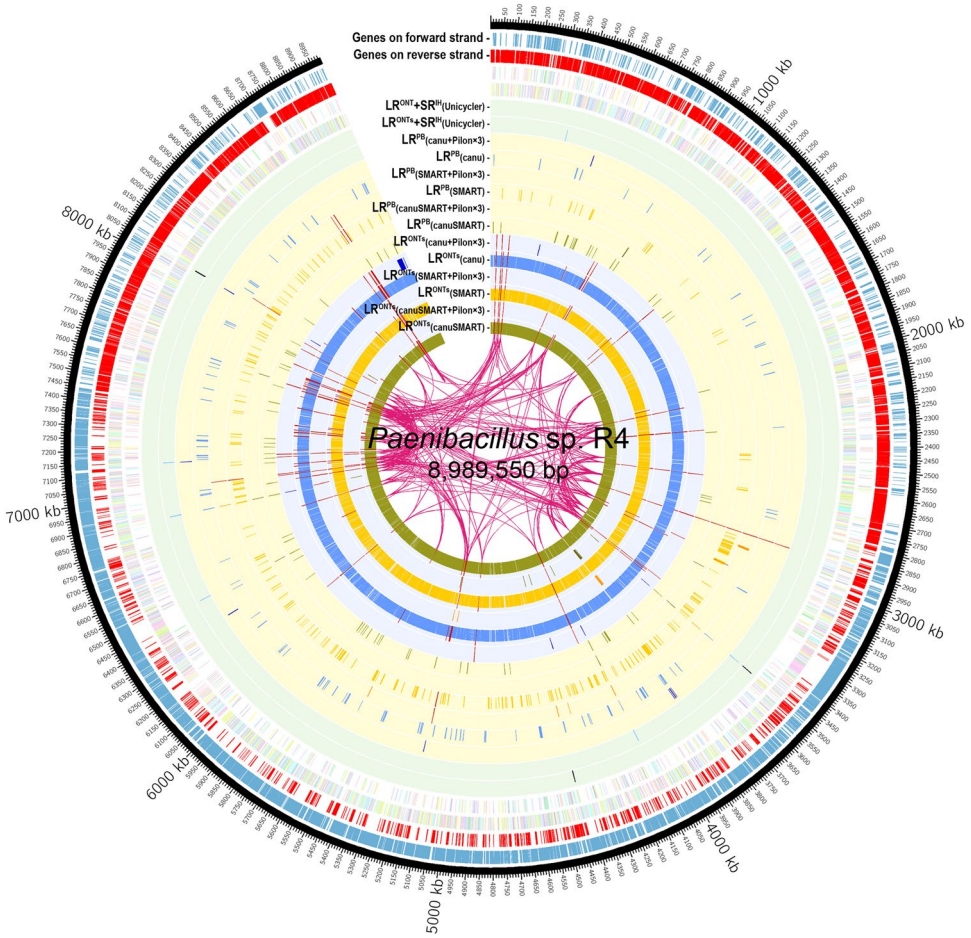

Fig. 4 Circular map of the *Paenibacillus* sp. R4 genome. Labeling from outside to the center: genes on the forward strand, genes on the reverse strand, genes on the forward strand are colored by Clusters of orthologous groups of proteins (COG) categories as in Fig. 2, genes on the reverse strand are colored by COG, split CDS in $LR^{ONTt}+SR^{IH}{}_{(Unicycler)}$ (black), split CDS in $LR^{ONTs}+SR^{IH}{}_{(Unicycler)}$ (blue) in green background, split CDS in $LR^{PB}{}_{(canu+Pilon\times3)}$ (dark blue), split CDS in $LR^{PB}{}_{(canu)}$ (blue) in the light yellow background, split CDS in $LR^{PB}{}_{(SMART+Pilon\times3)}$ (dark yellow), split CDS in the $LR^{PB}{}_{(SMART)}$ (yellow) in the light yellow background, split CDS in the $LR^{PB}{}_{(canuSMART+Pilon\times3)}$ (dark olive), split CDS in the $LR^{PB}{}_{(canuSMART)}$ (olive) in the light yellow background, split CDS in assemblies using LR^{ONTs} were drawn similar to split CDS in the assemblies using LR^{PB} in the light blue background, All split CDSs in repeat region are marked in pink, and repeat regions corresponding to each other are indicated with a ribbon (pink). The number of split CDS was reduced through genome polishing (Pilon × 3) in the assemblies using LR^{ONTs} and LR^{PB}, but the number of split CDS was not affected by repeat sequences in the assemblies using LR^{PB}

Table 6 Repeat sequences in $LR^{PB}+SR^{IH}{}_{(Unicycler)}$

Length range of repeat sequence	Number of repeat regions	Number of bases in the repeat sequence
1 kb >	6	3860
1–2 kb	464	715,734
2–3 kb	2	4204
3–4 kb	52	165,820
4–5 kb	10	49,011
5–6 kb	8	41,334
10–11 kb	2	21,728
21–22 kb	2	43,418
Sum	546	1,045,109

44 split CDS in the $LR^{ONTs}{}_{(canuSMART+Pilon\times3)}$, were positioned in repeat regions after the genome polishing and the effect of the repeat regions to mispredict CDSs was largely found in the polished assemblies using LR^{ONT}. When Racon was performed prior to polishing using Pilon × 3, mis-predicted ORFs in the repeat regions increased in the assemblies using LR^{PB}, 45, 40, and 45 more CDS in the repeat regions of $LR^{PB}+SR^{IH}{}_{(Unicycler)}$ were split in the $LR^{PB}{}_{(canu+Racon+Pilon\times3)}$, $LR^{PB}{}_{(SMART+Racon+Pilon\times3)}$, and $LR^{PB}{}_{(canuSMART+Racon+Pilon\times3)}$, respectively. In the assemblies using LR^{ONT}, the number of mis-predicted ORFs did not increase significantly via Racon, as in the assemblies using LR^{PB}, but 28, 33, and 28 mis-predicted CDSs increased to 44 in the $LR^{PB}{}_{(canu+Racon+Pilon\times3)}$,

40 in the $LR^{PB}_{(SMART+Racon+Pilon\times3)}$, and 43 in the $LR^{PB}_{(canuSMART+Racon+Pilon\times3)}$, respectively (Figs. 3 and 4).

Discussion

In this study, the long reads were assembled using Canu and SMARTdenovo (Koren et al. 2017; SMARTdenovo. https://github.com/ruanjue/smartdenovo. Accessed 19 November 2018.); both programs could assemble LR^{PB} and LR^{ONT} into contigs. In SMARTdenovo, different algorithms, called dot matrix alignment, were used that reduced the time of assembly compared with those using canu, which used MHAP using the MinHash algorithm (Broder 1997) to detect overlaps. Corrected reads using Canu were also assembled using SMARTdenovo called canuSMART (Schmidt et al. 2017). In assemblies using LR^{PB}, SMARTdenovo generated a higher number of total indels in dnadiff analysis and warnings in REAPR analysis among the three assembly methods (Tables 3 and 4). Conversely, the genome sequence assembled by SMARTdenovo showed the best result in the number of total indels and warnings among assemblies using LR^{ONT} (Tables 3 and 4). BUSCO analysis also showed that SMARTdenovo was the best assembly method with only LRs (Fig. 1). However, genome sequences assembled with nanopore reads showed a much lower quality in BUSCO analysis than those with PacBio reads. Compared to hybrid assemblies with high-quality reads, the genome sequence assembled using LR^{ONT} showed considerable differences.

Genome polishing using Pilon could reduce the differences between the hybrid assemblies with the SR^{IH} and other assemblies using only LRs (Table 4). Genome polishing reduced the number of total indels in the assemblies using LR^{PB} but did not significantly reduce the number of total substitutions in the dnadiff analysis. In assemblies using LR^{ONT}, genome polishing using Pilon reduced the number of both the total indels and total substitutions. The quality of the polished assemblies using LR^{ONT} seemed better than the unpolished assemblies using LR^{PB} compared with $LR^{PB} + SR^{IH}_{(Unicycler)}$. Genome polishing using Pilon was performed thrice. The first round of Pilon showed the best efficiency in correction (Supplementary Table S1), and the second round of Pilon further reduced the difference between $LR^{PB} + SR^{IH}_{(Unicycler)}$ and the other assemblies. However, the efficiency of the corrections was very low during the third round of Pilon. Thus, we did not proceed with further genome polishing using Pilon. Since Racon is also known to polish genome sequences with high-quality reads, we applied it to genome polishing (Supplementary Table S1). However, the efficiency of genome polishing was lower than that of Pilon in assemblies using LR^{ONT}. Rather, Racon reduced the quality of the bases in assemblies using LR^{PB}. In all assemblies polished using Racon and Pilon × 3, the number of total indels and total substitution of the polished genome increased significantly than those of polished genome using Pilon × 3 (Table 4 and Supplementary Table S1).

These differences in sequence similarity might affect the predictions of the CDS, and the exact prediction of the CDS was important for the majority of researchers. Though, several studies have evaluated the assemblies and contigs with identities against the reference sequences (Giordano et al. 2017; Goldstein et al. 2019; Loman et al. 2015; Schmidt et al. 2017), the reports on the accuracy of the prediction of the CDS in assemblies using nanopore reads (De Maio et al. 2019; Goldstein et al. 2019) are limited. In hybrid assemblies, it has been reported that the average gene length was higher, and the number of genes was smaller than those of the assemblies obtained using nanopore reads in several bacterial genomes. These studies have also reported that the fragmentation of the biosynthetic gene cluster residing in the repetitive genomic region was reduced (De Maio et al. 2019; Goldstein et al. 2019). Here, we focused on the accuracy of the prediction of the total CDS among the assemblies to evaluate the assemblies using nanopore reads obtained for the *Paenibacillus* sp. R4 genome. In assemblies using LR^{ONT}, the predicted ORF of $LR^{ONTt} + SR^{IH}_{(Unicycler)}$ and $LR^{ONTs} + SR^{IH}_{(Unicycler)}$ were almost the same. Six CDS and three CDS split, respectively, and only one CDS of $LR^{ONTs} + SR^{IH}_{(Unicycler)}$ was not found in the CDS of the $LR^{PB} + SR^{IH}_{(Unicycler)}$. In the hybrid assemblies using Unicyler, the low base qualities of the LR^{ONT} were compensated with high quality bases of the SR^{IH}. Without SR^{IH}, the assemblies using LR^{ONT} showed many differences in their ORFs, compared to the CDS of the $LR^{PB} + SR^{IH}_{(Unicycler)}$. Approximately 37.5–47.7% in total CDS were split because of the low base quality of the contigs; however, genome polishing with SR^{IH} corrected most of the errors in the assemblies using only LR^{ONT}, which reduced the split CDSs to 0.50–0.58%. Though the number of split CDS was larger than the polished assemblies using LR^{PB} and the hybrid assemblies with SR^{IH} ($LR^{ONTt} + SR^{IH}_{(Unicycler)}$, $LR^{ONTs} + SR^{IH}_{(Unicycler)}$, and $LR^{PB} + SR^{IH}_{(Unicycler)}$), it was smaller than the assemblies using LR^{PB} without polishing (0.92–3.75%), which were usually submitted to the GenBank databases. Repeat sequences in *Paenibacillus* sp. R4 could be the major reason for these differences. In the polished assemblies using LR^{ONT}, more than 58% of split CDSs were present within repeat sequences. Outside of repeat sequences, only 8–20 CDS were mis-predicted in the polished assemblies using LR^{ONT}. In the polished assemblies using LR^{PB}, the relatively higher base quality of repeat sequences generated from the LR^{PB} seemed to be properly discerned and corrected using Pilon. The number of split CDS in repeat regions was much smaller than those

in the polished assemblies using LR^{ONT} (Fig. 3). However, contigs generated from LR^{ONT} did not seem to be discerned by the genome polishing with SR^{IH} efficiently, because the higher similarity of repeat sequence might be greater than the accuracy of the bases in the assemblies using LR^{ONT}. Though SR^{IH} showed high-quality bases, repeat sequences were not properly corrected in the assemblies using LR^{ONT}. Genome polishing using Racon rather increased the error in the repeat regions of the assemblies using LR^{PB} and LR^{ONT}. If genome sequences had many repeat sequences, Racon could increase the rate of error in the contigs in the *Paenibacillus* sp. R4 genome.

Conclusion

In this study, we assembled the nanopore reads into a single contig with various assembly methods and evaluated the quality of the genome sequence and the prediction of CDS. Nanopore reads were sufficient to construct a single contig, but the low base quality of the assemblies resulted in fragmentation of many of the CDS requiring high-quality reads. Here we have shown that for accurate prediction of ORFs, high quality reads could be used in hybrid assemblies with long reads or with genome polishing. However, a genome sequence with repeat sequences may result in errors in the genome polishing process. The findings revealed that the hybrid assembly with high-quality reads is the best assembly method for the *Paenibacillus* sp. R4 genome using the nanopore reads. Moreover, it was also observed that the assemblies using only PacBio reads showed some CDS prediction errors, thus for a more accurate prediction for CDS, a genome sequence assembled using PacBio reads also needs high quality reads for genome.

Author contribution Conceptualization: HWK, JHL, SCS; methodology: HJK, SCS; software: SCS; formal analysis: WC, SCS; investigation: WC; resources: SCS; data curation: SCS; writing—original draft preparation: All authors; writing—review and editing: All authors; visualization: SCS, HWK; supervision: SCS, HWK; project administration: SCS, HWK; funding: HWK.

Funding This research was supported by a National Research Foundation of Korea Grant from the Korean Government (MSIT; the Ministry of Science and ICT) (NRF-2017M1A5A1013568) (KOPRI-PN20082) (Title: application study on the Arctic cold-active enzyme degrading organic carbon compounds).

Availability of data and material The raw data have been deposited at the National Center for Biotechnology Information (NCBI) BioProject repository PRJNA564035 (SRX6807868-SRX6807870). This strain is available from the Polar and Alpine Microbial Collection (PAMC) of Korea Polar Research Institute with the accession number PAMC 29622.

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing financial and non-financial interests.

Ethics approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Antipov D, Korobeynikov A, McLean JS, Pevzner PA (2016) HybridSPAdes: an algorithm for hybrid assembly of short and long reads. Bioinformatics 32:1009–1015

Ashton PM et al (2015) MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat Biotechnol 33:296

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120

Broder AZ (1997) On the resemblance and containment of documents. In: Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171). IEEE, pp 21–29

Chin C-S et al (2013) Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10:563

Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H (2009) Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol 4:265

De Maio N et al. (2019) Comparison of long-read sequencing technologies in the hybrid assembly of complex bacterial genomes. BioRxiv:530824

Delcher AL, Phillippy A, Carlton J, Salzberg SL (2002) Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res 30:2478–2483

Deschamps S et al (2016) Characterization, correction and de novo assembly of an Oxford Nanopore genomic dataset from *Agrobacterium tumefaciens*. Sci Rep 6:28625

Eccles D, Chandler J, Camberis M, Henrissat B, Koren S, Le Gros G, Ewbank JJ (2018) De novo assembly of the complex genome of Nippostrongylus brasiliensis using MinION long reads. BMC Biol 16:6

Eid J et al (2009) Real-time DNA sequencing from single polymerase molecules. Science 323:133–138

Giordano F et al (2017) De novo yeast genome assemblies from MinION PacBio and MiSeq platforms. Sci Rep 7:3935

Goldstein S, Beka L, Graf J, Klassen JL (2019) Evaluation of strategies for the assembly of diverse bacterial genomes using MinION long-read sequencing. BMC Genomics 20:23

Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD (2013) REAPR: a universal tool for genome assembly evaluation. Genome Biol 14:R47

Jain M et al (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol 36:338

Jain M, Olsen HE, Paten B, Akeson M (2016) The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. Genome Biol 17:239

Kim H, Park AK, Lee JH, Shin SC, Park H, Kim HW (2018) PsEst3, a new psychrophilic esterase from the Arctic bacterium Paenibacillus sp. R4: crystallization and X-ray crystallographic analysis. Acta Crystallogr F Struct Biol Commun 74:367–372. https://doi.org/10.1107/S2053230X18007525

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27:722–736

Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997
Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100
Loman NJ, Quick J, Simpson JT (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 12:733
Lu H, Giordano F, Ning Z (2016) Oxford Nanopore MinION sequencing and genome assembly. Genomics Proteomics Bioinformatics 14:265–279
Michael TP et al (2018) High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. Nat Commun 9:541
Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055
Passera A, Marcolungo L, Casati P, Brasca M, Quaglino F, Cantaloni C, Delledonne M (2018) Hybrid genome assembly and annotation of Paenibacillus pasadenensis strain R16 reveals insights on endophytic life style and antifungal activity. PLoS ONE 13:e0189993
Ross MG et al (2013) Characterizing and measuring bias in sequence data. Genome Biol 14:R51
Schmidt MH-W et al (2017) De novo assembly of a new *Solanum pennellii* accession using nanopore sequencing. Plant Cell 29:2336–2348
seqtk. https://github.com/lh3/seqtk Accessed 26 Aug 2019
Shin SC et al (2019) Nanopore sequencing reads improve assembly and gene annotation of the Parochlus steinenii genome. Sci Rep 9:5095
Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212
SMARTdenovo. https://github.com/ruanjue/smartdenovo. Accessed 19 Nov 2018
Tanizawa Y, Fujisawa T, Nakamura Y (2017) DFAST: a flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics 34:1037–1039
Tatusov RL, Galperin MY, Natale DA, Koonin EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res 28:33–36
Walker BJ et al (2014) Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9:e112963
Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595