JOURNAL OF AGRICULTURAL AND FOOD CHEMISTRY

pubs.acs.org/JAFC Article

Comparative Evaluation of Genome Assemblers from Long-Read Sequencing for Plants and Crops

Hyungtaek Jung,* Min-Seung Jeon, Matthew Hodgett, Peter Waterhouse, and Seong-il Eyun*

Cite This: *J. Agric. Food Chem.* 2020, 68, 7670−7677 Read Online

ACCESS | Metrics & More | Article Recommendations | SI Supporting Information

ABSTRACT: The availability of recent state-of-the-art long-read sequencing technologies has significantly increased the ease and speed of producing high-quality plant genome assemblies. A wide variety of genome-related software tools are now available and they are typically benchmarked using microbial or model eukaryotic genomes such as *Arabidopsis* and rice. However, many plant species have much larger and more complex genomes than these, and the choice of tools, parameters, and/or strategies that can be used is not always obvious. Thus, we have compared the metrics of assemblies generated by various pipelines to discuss how assembly quality can be affected by two different assembly strategies. First, we focused on optimizing read preprocessing and assembler variables using eight different *de novo* assemblers on five different Pacific Biosciences long-read datasets of diploid and tetraploid species. Then, we examined a single scaffolding tool (quickmerge) that has been employed for the postprocessing step. We then merged the outputs from multiple assemblies to produce a higher quality consensus assembly. Then, we benchmarked the assemblies for completeness and accuracy (assembly metrics and BUSCO), computer memory, and CPU times. Two lightweight assemblers, Miniasm/Minimap/Racon and WTDBG, were deemed good for novice users because they involved smaller required learning curves and light computational resources. However, two heavyweight tools, CANU and Flye, should be the first choice when the goal is to achieve accurate and complete assemblies. Our results will provide valuable guidance in future plant genome projects and beyond.

KEYWORDS: *plant genome, next-generation sequencing, Pacific Biosciences, long reads, nanopore, assemblers*

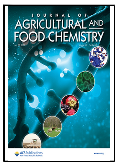
■ INTRODUCTION

The advent of next-generation sequencing (NGS) technologies has initiated a new era in genomics research. Despite dramatic improvements in DNA sequencing technologies and computation tools, assemblies using short reads remain very challenging because of large quantities of repetitive content in large genomes, uneven sequencing coverage, and the presence of (nonuniform) sequencing errors and chimeric reads.[1−3] To overcome the issues of NGS, long-read sequencing (LRS) technologies (Pacific Biosciences [PacBio] and Oxford Nanopore Technology) have been developed and recently actively adopted by the plant genomics community.[3−9] While emerging LRSs offer very long true reads (up to 200 kb in PacBio and 2 Mb in Nanopore), these technologies are still expensive (cost per base) and subject to high sequencing error rates (5−10% for PacBio and 10−15% for nanopore)[3,10] compared to NGS. However, applications of LRS to plant genomes have obvious advantages that are provided by long reads in *de novo* assembly, such as higher contiguity, smaller gaps, and fewer errors.[4,6,8,11−13] Nonetheless, care is required when planning a genome project to maximize assembly quality, cost, extra subchromosome length scaffolding (e.g., BioNano and Hi-C), and choice of assemblers.[3]

Determining which assembler can be used to produce the best quality assembly requires particular attention. The appropriate choice could depend on the size and complexity (repeat content, ploidy, etc.) of the genome to be assembled and the type of sequencing technology used to produce the input reads (e.g., NGSs vs LRSs).[3] While comparative evaluations such as the genome assembly gold-standard evaluation[14] and Assemblathon[15] can provide general guidelines, there is currently no systematic way to determine which assembler and parameter settings would produce the best assembly for a specific genome and/or dataset. Consequently, it is common practice to generate multiple genome assemblies from a few different assemblers, parameters, and algorithms [e.g., de Bruijn Graph (DBG) and overlap-layout-consensus (OLC)]. Then, researchers attempt to predict the best assembly based on assembly statistics, spot-checking, homology analysis, agreement with physical/genetic maps, and so forth. However, what constitutes the best assembly remains undefined. Because no perfect and error-free assembly exists, we must decide whether it is more important to maximize contig and scaffold length or minimize the number of misassemblies.[16,17]

To make matters more difficult, these new technologies and algorithms for long reads are typically benchmarked on microbial genomes or, if they are scaled appropriately, the human genome.[3,18] Unfortunately, the human genome is not the representative of all eukaryotic genomes; in particular,

Received: March 13, 2020
Revised: June 10, 2020
Accepted: June 12, 2020
Published: June 12, 2020



ACS Publications

https://dx.doi.org/10.1021/acs.jafc.0c01647

Table 1. Summary of Input Sequence and Species

species category	*Arabis alpina*	*O. indica*	*Durio zibethinus*	*C. quinoa*	*Z. mays* B73
sequencing platform	RSII	RSII	RSII	RSII	RSII
genome size (Gb)	0.37	0.40	0.74	1.40	2.30
ploidy	diploid	diploid	diploid	allotetraploid	diploid
coverage (*X*)	86	118	153	54	65
accession numbers	ERX1795357, PRJNA241291	PRJCA000313	PRJNA400310	PRJNA306026	SRX1472849
references	Jiao et al. 2017[5]	Du et al. 2017	Teh et al. 2017	Jarvis et al. 2017	Jiao et al. 2017[9]

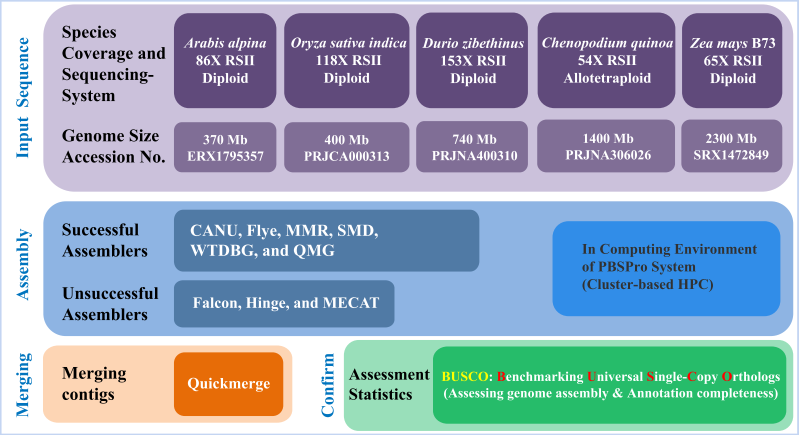

Input Sequence
Species Coverage and Sequencing-System
Arabis alpina 86X RSII Diploid
Oryza sativa indica 118X RSII Diploid
Durio zibethinus 153X RSII Diploid
Chenopodium quinoa 54X RSII Allotetraploid
Zea mays B73 65X RSII Diploid
Genome Size Accession No.
370 Mb ERX1795357
400 Mb PRJCA000313
740 Mb PRJNA400310
1400 Mb PRJNA306026
2300 Mb SRX1472849
Assembly
Successful Assemblers
CANU, Flye, MMR, SMD, WTDBG, and QMG
Unsuccessful Assemblers
Falcon, Hinge, and MECAT
In Computing Environment of PBSPro System (Cluster-based HPC)
Merging
Merging contigs
Quickmerge
Confirm
Assessment Statistics
BUSCO: Benchmarking Universal Single-Copy Orthologs (Assessing genome assembly & Annotation completeness)

Figure 1. Summary of long-read assembly workflow and evaluation.

Table 2. Summary of Computing Resources

		A. alpina	*O. indica*	*D. zibethinus*	*C. quinoa*	*Z. mays*
CANU	wall time[a]	45	47	95	167	288
	MEM usage[b]	1440	1440	1440	1440	1440
Flye	CPU time[a]	626.53	192.57	1037.52	1708.46	6037
	MEM usage[b]	303	136	465	395	1260
MMR	CPU time[a]	252.03	36.32	523.42	1030.65	3224.30
	MEM usage[b]	207	52	215	445	654
SMD	CPU time[a]	392.45	75.32	358.55	78.28	2421
	MEM usage[b]	40	20	45	38	962
WTDBG	CPU time[a]	24.41	47.22	37.47	78.22	393.59
	MEM usage[b]	45	103	49	93	315
QMG	CPU time[a]	1	4	5	13	47
	MEM usage[b]	55	67	118	175	376

[a] All times are in hours. [b] All memory usage (MEM) indicate the maximum memory usage (Gb).

plant genomes are larger and more repetitive than human genomes, and plant biology (e.g., chloroplasts and mitochondrial genomes) makes obtaining high-quality DNA free from contaminants difficult.[3,18] Thus, technologies that work well on vertebrate genomes may not work well with plant genomes[19] because each assembler implements slightly different heuristics to deal with repetitions, uneven coverage, sequencing errors, and chimeric reads while assembling genomes. Furthermore, each sequencing platform comes with its own input and computational requirements, qualities of output and, naturally, labor and material costs. Nonetheless, the final assembly is very rarely finished to feature one solid sequence per chromosome. Instead, typical outputs are presented as unordered/unoriented sets of contiguous regions called contigs. Alternatively, assembly reconciliation (or postprocessing) algorithms have been created to both produce a higher quality consensus assembly by merging two or more draft assemblies and enhance the contiguity of the resultant assembly while avoiding introducing assembly errors.[17]

In this paper, we compare eight different long-read *de novo* genome assemblers (termed preprocessing here) for five different plant PacBio RS II reads and one reconciliation assembler (termed postprocessing here). The genome of the targeted species varies in size from 0.4 to 2.3 Gb and comprises diploids and allotetraploids. Using the PacBio long reads, we carried out a comprehensive evaluation of assemblers by measuring the quality of the consensus assembly. Our results can be used as an initial guide for further sequencing assembly projects and a basis for plant genome studies, as each assembly method has its own biases.

■ MATERIALS AND METHODS

Input Sequence and Species. Five plant genomes were selected to evaluate the performance of multiple genome assemblers because their PacBio RSII reads had more than 50*X* coverage for each genome and were available at the National Center for Biotechnology

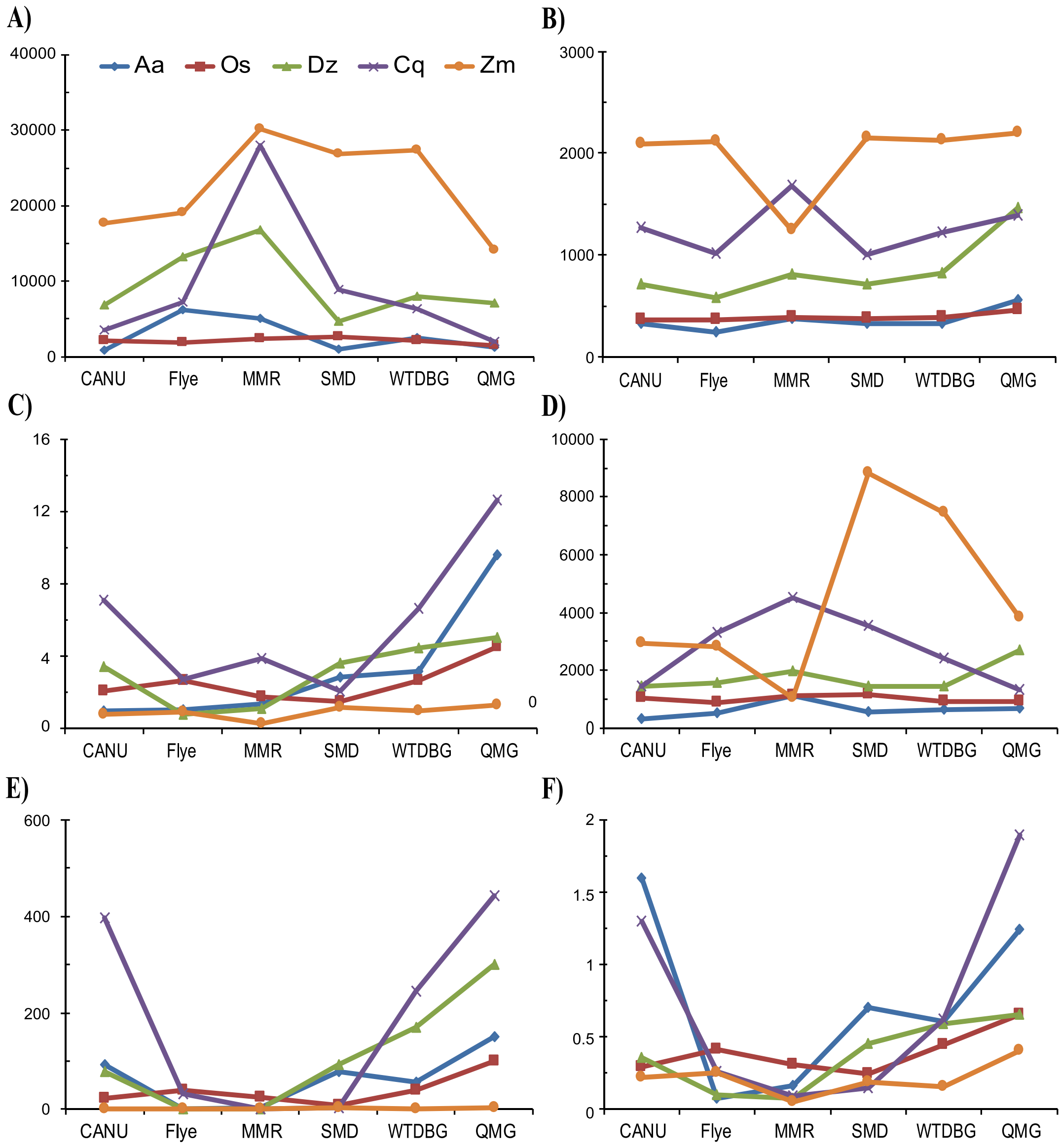


Figure 2. Summary of assembly statistics and metrics. Aa: *Arabis alpina*; Os: *O. indica*; Dz: *Durio zibethinus*; Cq: *C. quinoa*; Zm: *Z. mays* B73. Note that three contigs are longer than 10 Mb in Cq QMG (E). (A) Total number of contigs; (B) total assembled contig size (Mb); (C) longest contig (Mb); (D) number of contigs > 100 kb; (E) number of contigs > 1 Mb; and (F) Nb50 of contigs (Mb).

Information (NCBI) (Table 1). See the previously published papers for more library preparation and sequencing information (Table 1).

Assembly and Evaluation. Eight emerging *de novo* assemblers and pipelines (preprocessing assembly) were tested on a cluster-based high-performance computer (HPC) using PBSpro for job scheduling and workload management at the Queensland University of Technology in Australia and Chung-Ang University in Korea. Because of the specific requirements in computing environments, only five assemblers (with default parameters) were further compared in terms of assembly performance: CANU (ver. 1.7),[20] Flye (ver. 2.3.6),[21] Miniasm (r129)/Minimap/Racon (MMR),[22,23] SMARTdenovo (SMD) (https://github.com/ruanjue/smartdenovo), and WTDBG (ver. 1.2.8) (https://github.com/ruanjue/wtdbg). Three unsuccessful assemblers (Falcon, Hinge, and MECAT) were not included for further comparison (Figure 1). To ensure fairness in the comparisons, the following two criteria were applied to this work: (1) if an error correction and a polishing stage were embedded in the assembler and the pipeline, we processed it until its final assembly output and (2) further polishing such as PacBio (Arrow and Quiver) and/or Illumina (Pilon) was not considered for the final assembly output. However, Racon (mapping with minimap) was used as secondary error correction and consensus calling for the final contigs of WTDBG, and an embedded pipeline was used for MMR.[23] As a postprocessing step in the assembly process, we also employed quickmerge (QMG), a simple, fast, and general meta-assembler that merges assemblies to generate a more contiguous assembly.[24]

For completeness and contiguity, we used the Benchmarking Universal Single-Copy Orthologs (BUSCO) core plant dataset (ver. 3.0.2 and lineage db: embryophyta_odb9) to evaluate the gene contents[25] and calculated a full range of metrics for each assembly (such as a statistical summary of N50 length). To measure the computing time, we used the CPU time and memory usage (MEM) if the assembler could be done in a single-node job (24 cores with 240 Gb RAM or 192 cores with 6 Tb RAM); we used the actual wall-time and total memory usage for CANU if it required a multibranched job (multithreads and nodes) (Table 2).

RESULTS

The results of this study are presented in two parts. First, we compared the five successful long read-based assemblers as the preprocessing portion. Second, we merged all assembled

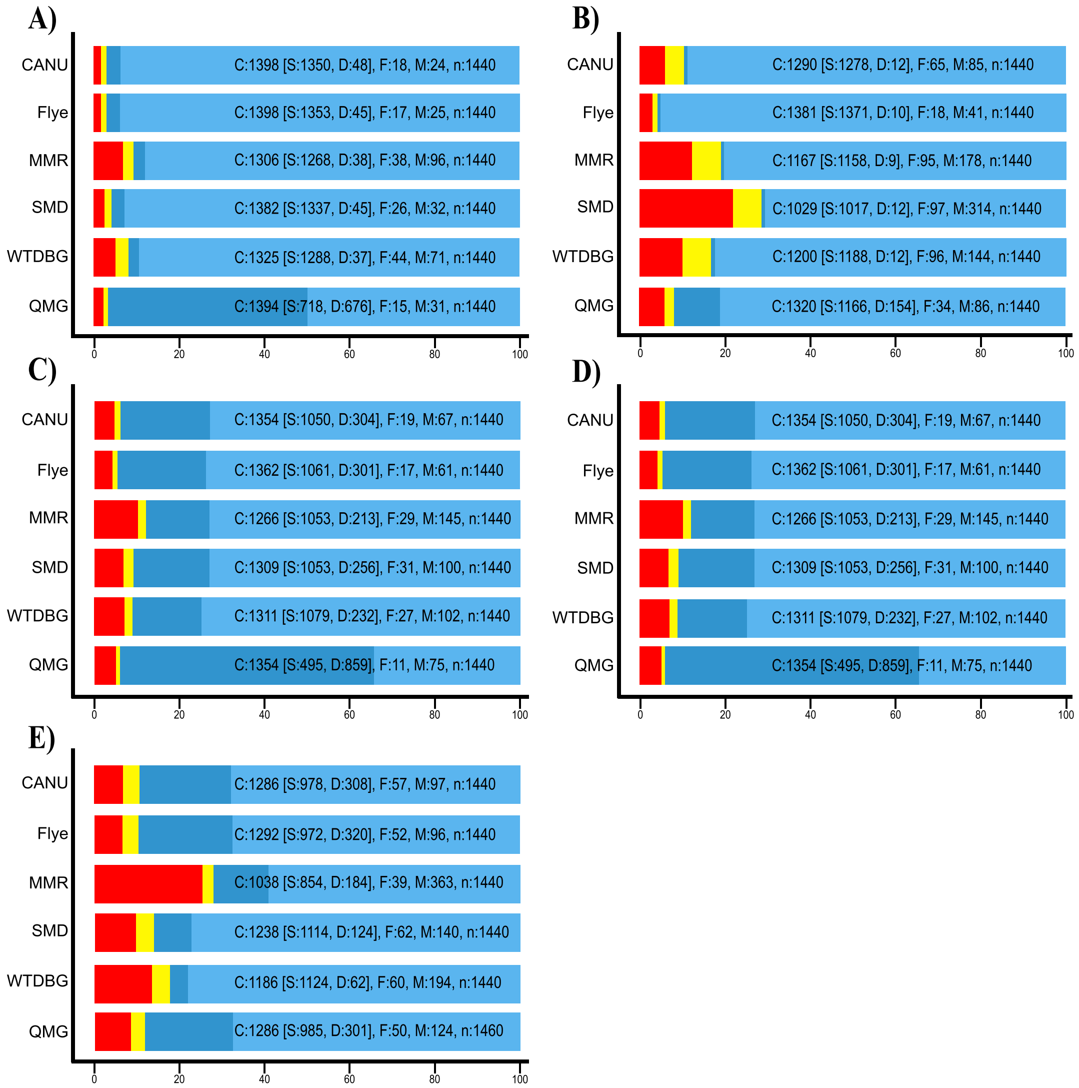


Figure 3. Assessment of genome assembly and annotation completeness with BUSCO. (A) *A. alpina*; (B) *O. indica*; (C) *Durio zibethinus*; (D) *C. quinoa*; and (E) *Z. mays* B73. BUSCO scores indicate the relative levels of completeness and putative gene duplications (see Materials and Methods). In BUSCO graphs, the *X*-axis represents BUSCO scores (%), and the *Y*-axis represents assemblers (CANU, Fly, MMR, SMD, WTDBG, and QMG). The different colors represent the following: dark-blue square: complete (C) and duplicated (D). Light-blue square: complete (C) and single-copy (S). Yellow square: fragmented (F). Red square: missing (M).

contigs in QMG as the postprocessing step. By validating the assemblies for sequence accuracy, we found both strengths and weaknesses and found that different methods resulted in stark differences in DNA sequence complexity, time, computational requirements, and cost.

Preprocessing Assembly Performance. The first stage of an assembly is to piece together long reads to form long contigs. Here, we focus on assessing the various existing pipelines and assemblers and comparing the results obtained from PacBio RSII data. While the long reads provided by PacBio can be used to generate a *de novo* assembly either alone or in conjunction with Illumina data, we show examples of assemblies from pure long reads.

We selected five successful assembly pipelines for LRS data (Figures 1 and 2). CANU, MMR, SMD, and WTDBG are based on the OLC algorithm. Flye is based on a generalized DBG algorithm. While CANU includes a base-error correction step, other assemblers do not require reads that have been error-corrected. After an initial assessment of the BUSCO scores, Miniasm/Minimap and WTDBG showed relatively low values (Figure 3). Thus, a third-party program called Racon was used as secondary error correction and consensus calling for the final contigs of these two assemblers because it aligns raw long reads to the contigs and generates a consensus, thus significantly increasing the initial accuracy.[23] Details on statistics and BUSCO on these assemblers and pipelines can be found in Figures 2 and 3. While no single assembler outperforms all others across all species, the highest assembly quality was observed in CANU/Flye followed by MMR. SMD and WTDBG showed the lowest accuracy in our attempts; as such, they might require a substantial polishing step, as they have no built-in error-correction stage.

Given the computing environment (an HPC with job scheduling and workload management) and dataset, all tested assemblers were relatively user-friendly. Flye, SMD, and WTDBG required fewer than three script lines/commands, but CANU and MMR needed ~20 script lines/commands. All executed scripts are summarized and available in the Supporting Information. Differences were observed in contiguity, completeness, and computing resources (CPU

Table 3. Comparison of Assembly Results between the Current Study and Previous Works[a]

		A. alpina	*O. indica*	*D. zibethinus*	*C. quinoa*	*Z. mays*
			Current Study			
CANU	TNC[b]	797	2156	6894	3521	17,653
	TACS[c]	331	363	707	1265	2093
	N50[d]	1.6	0.3	0.4	1.3	0.2
Flye	TNC[b]	6151	1937	13,238	7251	19,038
	TACS[c]	244	358	584	1011	2118
	N50[d]	0.1	0.4	0.1	0.3	0.2
MMR	TNC[b]	5004	2368	16,740	28,009	30,156
	TACS[c]	379	385	810	1683	1249
	N50[d]	0.2	0.3	0.1	0.1	0.1
SMD	TNC[b]	952	2664	4670	8865	26,875
	TACS[c]	327	378	711	999	2151
	N50[d]	0.1	0.2	0.5	0.2	0.2
WTDBG	TNC[b]	2522	2186	7953	6272	27,328
	TACS[c]	329	383	823	1216	2198
	N50[d]	0.6	0.4	0.6	0.6	0.2
QMG	TNC[b]	1220	1429	7148	1959	14,103
	TACS[c]	550	464	1462	1389	2198
	N50[d]	1.3	0.7	0.7	1.9	0.4
			Previous Works			
CANU	TNC[b]		1226			
	TACS[c]		405			
	N50[d]		0.9			
Falcon	TNC[b]					
	TACS[c]	328				
	N50[d]	0.8				
PBcR[e]	TNC[b]		3822/2045			
	TACS[c]	347	471/436			
	N50[d]	0.9	0.4/1.1			
PBcR-MHAP	TNC[b]					2958
	TACS[c]					2104
	N50[d]					1.2
Celera Assembler	TNC[b]				4232	
	TACS[c]				1325	
	N50[d]				1.7	

[a] Focused on contigs, not scaffolds. Only scaffold information was available for *D. zibethinus*. Empty cells: Data not available. [b] Total number of contigs. [c] Total assembled contig size (Mb). [d] Nb50 of contigs (Mb). [e] PBcR; PacBio corrected reads pipeline.

time, wall time, and memory usage). For Flye, if a target species was a large and complex genome (>2 Gb; *Zea mays*), it required specification in the subset of reads for initial contig assembly (e.g., --asm-coverage 30) because of memory and/or performance optimizations as opposed to the full input of reads. Despite the size and complexity of genomes, CANU and Flye showed reliable contiguity and completeness; however, the results provided by MMR, SMD, and WTDBG fluctuated (Figures 2 and 3). CANU required the most computing power (with larger memory and longer running time) followed by Flye, MMR, SMD, and WTDBG (Table 2). The main reason could be the embedded base−error correction step in CANU that differentiates it from other assemblers. While WTDBG was the fastest and easiest assembler tested, the overall consensus of its final contigs was relatively low (<50% of BUSCO completion). However, after polishing the contigs with Illumina reads (>94% BUSCO completion; verified with a developer) and/or correcting errors with Racon (>83% BUSCO completion), the overall final consensus was good enough to be used in comparisons with other assemblers' outcomes. While all contigs generated from these assemblers required further polishing with long/short reads to increase the overall accuracy, extra cautious steps were required for MMR, SMD, and WTDBG.

Postprocessing Assembly Performance. The second stage of an assembly is to assemble the conserved regions of the genome to reduce the complexity of *de novo* assembly for the nonconserved portions. While the quality of input assemblies is expected to directly affect that of the final merged assembly, we have not explored different input qualities in this paper. QMG was employed as a single reconciliation tool to evaluate postprocessing assembly performance because it allows users to merge an assembly obtained from PacBio reads alone or with another assembly based on second-generation reads.[24]

QMG was run with the default parameters, and BUSCO scores were used to gather extensive assembly statistics and gene content completion.[25] While merging multiple individual assemblies substantially improved both the contig size and number (with general improvements to contiguity) for the majority of species, the BUSCO values maintained similar quality statistics with the single best outcomes in preprocessing. In general, as the number of inputs increased, the contiguity improved, thus resulting in fewer but longer contigs.

Even merging the two best outcomes from CANU and Flye showed improvements (data not shown). Although we did not investigate further misassemblies, Alhakami et al. (2017) proposed that more inputs for postprocessing assembly could decrease misassemblies and improve contiguity.[17] Details on the statistics and BUSCO from QMG can be found in Figures 2 and 3.

DISCUSSION

NGS technologies and sequence data analysis (including *de novo* assembly) have radically transformed the field of plant genomics in recent years. Substantial advancements in LRSs and bioinformatics have also provided the necessary framework to systematically analyze data.[3] However, determining the most effective way to sequence and assemble a large, complex plant genome (>1 Gb) among the increasingly varied sequencing and assembly approaches remains difficult. In particular, the existence of long-read assemblers mainly focused on model species (e.g., humans and bacterial genomes) have made it more difficult to gauge their capability and efficiency in handling plant genome data, which can be larger and more repetitive.[3,18]

Over the last decade, DBG assembly coupled with short reads of NGSs has been the method of choice to sequence and assemble plant and animal genomes. However, OLC assemblers, such as PacBio and Nanopore, are well-suited to *de novo* assembly from long reads.[3,6,10,26] Our PacBio RSII assemblies using five different assemblers (preprocessing) have provided varied results, indicating that different algorithms and/or pipelines can affect assembly quality. In general, estimating the assembly quality requires several statistical evaluations: (1) overall assembly size (match the estimated genome size), (2) measures of assembly contiguity (metrics of N50 from contig numbers, longest contigs, and mean contig size), (3) assembly likelihood scores (calculated by aligning reads against each candidate assembly),[27] (4) accuracy of assembly (aligning the contigs to existing physical maps if available), and (5) completeness of the genome assembly (BUSCO).[2,28] While we have not conducted all statistical evaluations, our outcomes based on the five assemblers studied have achieved unprecedented contiguity compared to that of NGSs, and three assemblers (MMR, SMD, and WTDBG) showed satisfactory results with proper correction with Racon. However, it should be noted that PacBio sequencing and analysis comes with higher sequencing costs, error rates, computational power, and computing time compared to Illumina sequencing and analysis. Thus, obtaining a minimum 50*X* coverage is recommended to produce a high-quality diploid genome (50% more coverage for polyploidy).[3,18] Additionally, extra Illumina reads polished using Pilon (or equivalent tools) would help minimize any residual and/or artefact errors of PacBio. Despite the genome size, if Illumina reads are available for polishing utilizing the two lightweight tools of MMR and WTDBG, this approach would be good for a novice user, as it does not have a steep learning curve and/or heavy computational requirements. However, heavyweight tools, such as CANU and Flye, should be the first choice for an assembler when looking to achieve accurate assemblies. While we did not succeed in getting Falcon to work in our high-performance cluster because it is designed for Sun Grid Engine and not PBSpro, several papers have already proven its capability and efficiency.[18,29]

A comparison of the results between the current study and the previous works has provided another valuable point to consider for selecting a proper assembler (Tables 1 and 3). However, these tables should be interpreted cautiously because the previous works were assembled with early versions of the algorithm and/or pipeline. In the case of *Arabis indica*, for example, the results from two different versions (ver. 1.3 for Du et al., 2017 and ver. 1.7 for current study) show inconsistent assembly outcomes, although the same default parameters are set in CANU. These might be the bubble issues to avoid false breaks (e.g., repeats) and potential improvement of the autoset error rate. This point is already clarified in CANU's GitHub (issue numbers #245 and #852 on https://github.com/marbl/canu/releases). Although it would be problematic to compare the assembly outcomes of *A. indica* with two different CANU versions, it would be a great challenge for software and algorithm developers to improve the high error rate of long-read assemblers. Because there is no single definitive assembler that guarantees to deliver the best result for a given dataset, we highly recommend selecting the best assembly outcome after comparing the latest versions of a minimum of two different assemblers (e.g., CANU and Flye). It is highly likely that different assembly pipelines could generate different results even for the same dataset. The advantages and disadvantages of the five tested assemblers have been summarized to provide a useful guideline for selecting a proper pipeline according to five criteria (memory intensity, running time, BUSCO completion, ease of use, and program update) (Figure 4). According to our experience and tested dataset, if the coverage is more than 70*X* and computational resources are not limited, then selecting CANU and Flye is the optimal choice regardless of repeat content and ploidy issues in plant genome assembly. In addition, particularly in the case of CANU setting, a stringent length filtering option (minOverlapLength >3000 or even higher) after removing all non-nuclear genome data (i.e. chloroplast and mitochondrial DNA) results in more correct assemblies.

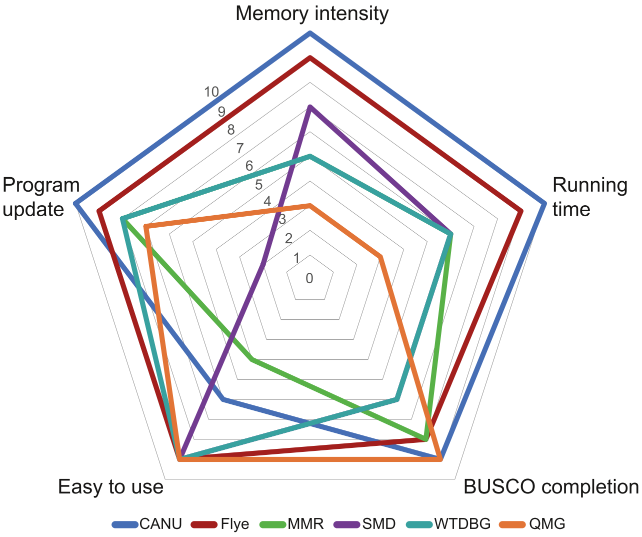


Figure 4. Asset pentagons for five different genome assemblers. Comparison of assembly performance and recommendation for all five assemblies.

Highly contiguous and accurate plant genome assemblies have been shown to generate in a *de novo* manner solely using

PacBio data, but the final assembly is still not entirely finished with results of one solid sequence per chromosome.[3] Given the practical challenges of a *de novo* assembly, the idea of reconciliation (postprocessing) is very appealing because it can allow the merging of all assemblies (generated from multiple assemblies) to obtain a high-quality consensus assembly.[17] Regarding the outcomes, the expectation is that the quality of the merged assembly should be at least as good as the best assembly in the input because one should expect the consensus assembly to inherit good qualities from the given inputs. However, it is difficult to produce a merged assembly that is consistently better than (or at least as good as) the given input assemblies. There were two cases in which the consensus assembly (*Oryza sativa indica* and *Chenopodium quinoa*) was better than the given inputs, but the merged assembly was relatively good for the majority of the inputs, as was demonstrated in previous work.[17] While we do not screen any potential chimeric assemblies from postprocessing, users should be mindful about introducing chimeric/fusion assemblies from prior merging steps into later steps, particularly for polyploid genomes.

Although the current study is limited in scope and use of data, it allows some suggestions to improve future plant and crop genome assemblies: (1) the major availability of the PacBio data is RS II at the moment because we are unable to access a substantial amount of PacBio SEQUEL data to employ various plant species from the public repository. PacBio SEQUEL II has improved over the last 12 months, and this sequencing platform can produce up to 15 times more data per cell (~150 Gb) with higher accuracy in longer reads (reads could be ABI Sanger quality up to 40 kb) and reduced sequencing costs. (2) For analytical tools and assembly algorithms, more sophisticated and computationally efficient assemblers have continually been updated; these include CANU (ver. 2.0),[20] Flye (ver. 2.7.1),[21] and WTDBG (ver. 2.2).[30] It is highly recommended to use the latest versions because these assemblers can achieve more contiguity and accuracy in genome assemblies by fixing many known issues and bugs from previous versions. (3) In eukaryotic contigs, the terminal regions could be scanned using a tandem repeat finder[31] for the presence of telomeres that might be related to the peak computational memory in the form of maximum resident set size and CPU times. While our attempt does not compare the comprehensive tandem repeat, error correction, and polishing stages, a cautious approach should be taken because redundant base pairs in the overlapping terminal regions of fragmented contigs lead to unresolved errors, even after several rounds of consensus polishing. (4) Our attempt highlights how best to use limited genomic resources for effectively evaluating the *de novo* genome assemblies and performances of plant and crop species for novice users (nonexpert in bioinformatics). Furthermore, it provides a minimal advisable requirement of RAM and CPU cores by comparing assembly metrics and BUSCO completion. However, when completed high-quality genome references are available, the most comprehensive genome assembly comparison could be achievable. For example, (i) single-nucleotide variations (SNVs) and indels and structural variations (SVs) could be useful to evaluate assembly correctness and provide a relative measure of assembly errors and (ii) dot plots could be informative to visualize the genomic variations and rearrangements. Despite not having a chance to test new HiFi data, the updated assemblers, and all criteria, according to our outcomes, selecting any of the three *de novo* genome assemblers of CANU, Flye, and MMR would allow novice users to acquire suitable results in plant genome assemblies. Utilizing QMG as a postprocessing assembly step is a good strategy depending on the outcomes of the preprocessing assembly.

Impressive strides have been made in the production of plant genome assemblies, thanks to the availability of high-throughput LRSs and NGS data and improved assembly tools/algorithms.[3] For researchers, selecting the best sequencing platform and analytical approach for genome assembly remains challenging, as each option has pros and cons. Nevertheless, continued advances in both sequencing and bioinformatic technologies increase the likelihood of delivering accurate, contiguous, and eventually entire chromosome sequences at low costs. We hope that the comparison of the long-read assemblers we have tested will aid and encourage researchers to spend less time on genome assembly and focus more on exploring the biology of genomes to achieve their research goals.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jafc.0c01647.

Examples of the executed codes CANU, Flye, MMR, SMD, WTDBG, and QuickMerge (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Hyungtaek Jung − *Centre for Agriculture and Biocommodities, Queensland University of Technology, Brisbane, Queensland 4001, Australia*; Email: h7.jung@qut.edu.au

Seong-il Eyun − *Department of Life Science, Chung-Ang University, Seoul 06974, Korea*; orcid.org/0000-0003-4687-1066; Email: eyun@cau.ac.kr

Authors

Min-Seung Jeon − *Department of Life Science, Chung-Ang University, Seoul 06974, Korea*

Matthew Hodgett − *Information Technology Services, Queensland University of Technology, Brisbane, Queensland 4001, Australia*

Peter Waterhouse − *Centre for Agriculture and Biocommodities, Queensland University of Technology, Brisbane, Queensland 4001, Australia*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jafc.0c01647

Funding

This project was supported by a Chung-Ang University Research Grant in 2017 to SE and an Australian Research Council (ARC) Laureate Fellowship (LF160100155) to PW.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors are grateful to their colleagues/collaborators, the field/technical specialists of each company, and the three anonymous reviewers for their valuable comments. The authors are especially grateful to Michal Lorenc, QUT High-

Performance Computing and Research Support, and the eResearch team for their technical assistance.

■ ABBREVIATIONS

DBGe, Bruijn Graph; HPC, high-performance computer; GAGE, genome assembly gold-standard evaluation; LRS, long-read sequencing; MMR, Miniasm/Minimap/Racon; NGS, next-generation sequencing; OLC, overlap-layout-consensus; PacBio, Pacific Biosciences; SMD, SMARTdenovo; QMG, Quickmerge

■ REFERENCES

(1) Kyriakidou, M.; Tai, H. H.; Anglin, N. L.; Ellis, D.; Stromvik, M. V. Current Strategies of Polyploid Plant Genome Sequence Assembly. *Front. Plant Sci.* **2018**, *9*, 1660.

(2) Sohn, J. I.; Nam, J. W. The present and future of de novo whole-genome assembly. *Briefings Bioinf.* **2018**, *19*, 23−40.

(3) Jung, H.; Winefield, C.; Bombarely, A.; Prentis, P.; Waterhouse, P. Tools and Strategies for Long-Read Sequencing and De Novo Assembly of Plant Genomes. *Trends Plant Sci.* **2019**, *24*, 700−724.

(4) Jiao, W.-B.; Accinelli, G. G.; Hartwig, B.; Kiefer, C.; Baker, D.; Severing, E.; Willing, E.-M.; Piednoel, M.; Woetzel, S.; Madrid-Herrero, E.; et al. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res.* **2017**, *27*, 778−786.

(5) Schmidt, M. H.-W.; Vogel, A.; Denton, A. K.; Istace, B.; Wormit, A.; van de Geest, H.; Bolger, M. E.; Alseekh, S.; Mass, J.; Pfaff, C.; et al. De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *Plant Cell* **2017**, *29*, 2336−2348.

(6) Belser, C.; Istace, B.; Denis, E.; Dubarry, M.; Baurens, F.-C.; Falentin, C.; Genete, M.; Berrabah, W.; Chèvre, A.-M.; Delourme, R.; et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat. Plants* **2018**, *4*, 879−887.

(7) Wee, Y.; Bhyan, S. B.; Liu, Y.; Lu, J.; Li, X.; Zhao, M. The bioinformatics tools for the genome assembly and analysis based on third-generation sequencing. *Briefings Funct. Genomics* **2019**, *18*, 1−12.

(8) Jiao, Y.; Peluso, P.; Shi, J.; Liang, T.; Stitzer, M. C.; Wang, B.; Campbell, M. S.; Stein, J. C.; Wei, X.; Chin, C.-S.; et al. Improved maize reference genome with single-molecule technologies. *Nature* **2017**, *546*, 524−527.

(9) Du, H.; Yu, Y.; Ma, Y.; Gao, Q.; Cao, Y.; Chen, Z.; Ma, B.; Qi, M.; Li, Y.; Zhao, X.; et al. Sequencing and de novo assembly of a near complete indica rice genome. *Nat. Commun.* **2017**, *8*, 15324.

(10) Senol Cali, D.; Kim, J. S.; Ghose, S.; Alkan, C.; Mutlu, O. Nanopore sequencing technology and tools for genome assembly: computational analysis of the current state, bottlenecks and future directions. *Briefings Bioinf.* **2019**, *20*, 1542−1559.

(11) Jarvis, D. E.; Ho, Y. S.; Lightfoot, D. J.; Schmöckel, S. M.; Li, B.; Borm, T. J. A.; Ohyanagi, H.; Mineta, K.; Michell, C. T.; Saber, N.; et al. The genome of Chenopodium quinoa. *Nature* **2017**, *542*, 307−312.

(12) Teh, B. T.; Lim, K.; Yong, C. H.; Ng, C. C. Y.; Rao, S. R.; Rajasegaran, V.; Lim, W. K.; Ong, C. K.; Chan, K.; Cheng, V. K. Y.; et al. The draft genome of tropical fruit durian (Durio zibethinus). *Nat. Genet.* **2017**, *49*, 1633−1641.

(13) Chaw, S.-M.; Liu, Y.-C.; Wu, Y.-W.; Wang, H.-Y.; Lin, C.-Y. I.; Wu, C.-S.; Ke, H.-M.; Chang, L.-Y.; Hsu, C.-Y.; Yang, H.-T.; et al. Stout camphor tree genome fills gaps in understanding of flowering plant genome evolution. *Nat. Plants* **2019**, *5*, 63−73.

(14) Salzberg, S. L.; Phillippy, A. M.; Zimin, A.; Puiu, D.; Magoc, T.; Koren, S.; Treangen, T. J.; Schatz, M. C.; Delcher, A. L.; Roberts, M.; et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* **2012**, *22*, 557−567.

(15) Bradnam, K. R.; Fass, J. N.; Alexandrov, A.; Baranay, P.; Bechner, M.; Birol, I.; Boisvert, S.; Chapman, J. A.; Chapuis, G.; Chikhi, R.; et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* **2013**, *2*, 10.

(16) Soueidan, H.; Maurier, F.; Groppi, A.; Sirand-Pugnet, P.; Tardy, F.; Citti, C.; Dupuy, V.; Nikolski, M. Finishing bacterial genome assemblies with Mix. *BMC Bioinf.* **2013**, *14*, S16.

(17) Alhakami, H.; Mirebrahim, H.; Lonardi, S. A comparative evaluation of genome assembly reconciliation tools. *Genome Biol.* **2017**, *18*, 93.

(18) Paajanen, P.; Kettleborough, G.; López-Girona, E.; Giolai, M.; Heavens, D.; Baker, D.; Lister, A.; Cugliandolo, F.; Wilde, G.; Hein, I.; et al. A critical comparison of technologies for a plant genome sequencing project. *Gigascience* **2019**, *8*, giy163.

(19) Jiao, W.-B.; Schneeberger, K. The impact of third generation genomic technologies on plant genome assembly. *Curr. Opin. Plant Biol.* **2017**, *36*, 64−70.

(20) Koren, S.; Walenz, B. P.; Berlin, K.; Miller, J. R.; Bergman, N. H.; Phillippy, A. M. Canu: scalable and accurate long-read assembly via adaptivek-mer weighting and repeat separation. *Genome Res.* **2017**, *27*, 722−736.

(21) Kolmogorov, M.; Armstrong, J.; Raney, B. J.; Streeter, I.; Dunn, M.; Yang, F.; Odom, D.; Flicek, P.; Keane, T. M.; Thybert, D.; et al. Chromosome assembly of large and complex genomes using multiple references. *Genome Res.* **2018**, *28*, 1720−1732.

(22) Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **2016**, *32*, 2103−2110.

(23) Vaser, R.; Sović, I.; Nagarajan, N.; Šikić, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **2017**, *27*, 737−746.

(24) Chakraborty, M.; Baldwin-Brown, JG; Long, AD; Emerson, JJ Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.* **2016**, *44*, No. e147.

(25) Simão, F. A.; Waterhouse, R. M.; Ioannidis, P.; Kriventseva, E. V.; Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210−3212.

(26) Chu, J.; Mohamadi, H.; Warren, R. L.; Yang, C.; Birol, I. Innovations and challenges in detecting long read overlaps: an evaluation of the state-of-the-art. *Bioinformatics* **2017**, *33*, 1261−1270.

(27) Clark, S. C.; Egan, R.; Frazier, P. I.; Wang, Z. ALE: a generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics* **2013**, *29*, 435−443.

(28) Conte, M. A.; Gammerdinger, W. J.; Bartie, K. L.; Penman, D. J.; Kocher, T. D. A high quality assembly of the Nile Tilapia (Oreochromis niloticus) genome reveals the structure of two sex determination regions. *BMC Genomics* **2017**, *18*, 341.

(29) Girollet, N.; Rubio, B.; Lopez-Roques, C.; Valiere, S.; Ollat, N.; Bert, P. F. De novo phased assembly of the Vitis riparia grape genome. *Sci. Data* **2019**, *6*, 127.

(30) Ruan, J.; Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat. Methods* **2020**, *17*, 155−158.

(31) Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **1999**, *27*, 573−580.

https://dx.doi.org/10.1021/acs.jafc.0c01647