scientific data

OPEN

DATA DESCRIPTOR

Two long read-based genome assembly and annotation of polyploidy woody plants, *Hibiscus syriacus* L. using PacBio and Nanopore platforms

Hyunjin Koo[1,6], Gir-Won Lee[2,6], Seo-Rin Ko[1,3], Sangjin Go[1,3], Suk-Yoon Kwon[1,3], Yong-Min Kim[1,4,5] ✉ & Ah-Young Shin[1,4] ✉

Improvements in long read DNA sequencing and related techniques facilitated the generation of complex eukaryotic genomes. Despite these advances, the quality of constructed plant reference genomes remains relatively poor due to the large size of genomes, high content of repetitive sequences, and wide variety of ploidy. Here, we developed the *de novo* sequencing and assembly of high polyploid plant genome, *Hibiscus syriacus*, a flowering plant species of the Malvaceae family, using the Oxford Nanopore Technologies and Pacific Biosciences Sequel sequencing platforms. We investigated an efficient combination of high-quality and high-molecular-weight DNA isolation procedure and suitable assembler to achieve optimal results using long read sequencing data. We found that abundant ultra-long reads allow for large and complex polyploid plant genome assemblies with great recovery of repetitive sequences and error correction even at relatively low depth Nanopore sequencing data and polishing compared to previous studies. Collectively, our combination provides cost effective methods to improve genome continuity and quality compared to the previously reported reference genome by accessing highly repetitive regions. The application of this combination may enable genetic research and breeding of polyploid crops, thus leading to improvements in crop production.

Background & Summary

Recent genome sequencing approaches, such as Pacific Biosciences Sequel (PacBio Sequel) and Oxford Nanopore Technologies (ONT), featuring long reads provide several advantages, which ultimately reduce additional sequencing costs and simplify preparation[1–4]. Such approaches hold promise for solving challenges associated with sequencing and assembling large, repetitive, and complex plant genomes through the production of large quantities of long reads to help bridge difficult regions in the genome[5,6]. Indeed, with notable improvements in long-read sequencing and related techniques, over 800 species of land plant genomes have been assembled recently[7,8]. Among these technologies, ONT led to a substantial improvement in plant genome contiguity and contig reduction. The wild tomato species *Solanum punnellii* (1.0 Gb) genome was assembled with a high contig N50 of 2.5 Mb using 135 Gb of ONT long-read data generated from 31 flow cells[9]. The highly heterozygous *Eucalyptus pauciflora* was sequenced and assembled using a combination of ONT long-read data (174×) and short-read Illumina data (228×) with five different assemblers. MaSuRCA generated the best assembly, which is 594.87 Mb in size, with a contig N50 of 3.23 Mb, and an estimated error rate of $\sim$0.006 errors per base[10].

[1]Plant Systems Engineering Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon, 34141, Republic of Korea. [2]SML Genetree Co. Ltd., Seoul, 05855, Republic of Korea. [3]Biosystems and Bioengineering Program, University of Science and Technology, Daejeon, 34113, Korea. [4]Department of Bioinformatics, KRIBB School of Bioscience, Korea University of Science and Technology (UST), Daejeon, 34141, Republic of Korea. [5]Digital Biotech Innovation Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon, 34141, Republic of Korea. [6]These authors contributed equally: Hyunjin Koo, Gir-Won Lee. ✉e-mail: ymkim@kribb.re.kr; shinay@kribb.re.kr

	Illumina[35]	ONT
Raw read coverage	122.8X (233.3 Gb)	63.4X (120.5 Gb)
Number of contigs	77,492	406
N50 (bp)	139,814	8,098,919
Average contig size	22,560	4,623,639
Number of Contigs >1 kb	34,103	405
Number of Contigs >100 kb	5,558	376
Number of Contigs >1 M	6	320
Number of Contigs >10 M	0	48
Total (Mb)	1,748	1,877

Table 1. Statistics of Illumina and NECAT genome assembly of *H. syriacus* cv. Gangneung.

Upon the advantages offered by long-read sequencing platforms, diverse plant genomes, including those of radish, oat, cotton, have been assembled at the chromosome level[11–13].

Despite these advances, the quality of published sequences remains relatively poor[14]. The major barriers to plant genome sequencing and assembly are that plant genomes vary widely in size, have a high content of repetitive sequences, and exhibit a wide variety of ploidy[15–17]. Furthermore, isolation of high-quality, high-molecular-weight (HMW) DNA from plants poses a unique challenge due to rigid cell walls, co-purification of mitochondrial and chloroplast genomes, polysaccharides, and phenolic compounds that directly damage DNA, reducing sequencing yields[9,18–24]. Although various DNA extraction and library preparation protocols have been developed for different organisms or tissues of interest, obtaining reliable, high-quality yields from highly repetitive and polyploid plant genomes can be challenging. Furthermore, choosing among the many genome assemblers – such as Miniasm[25], canu[26], Flye[27], wtdbg2[28], SMARTdenovo[29], Shasta[30], NECAT[31], and nextDenovo[32] – can pose a barrier. NECAT relays a novel progressive, two-step error-correction algorithm with adaptive candidate-read selection for ONT raw reads[31]. NextDenovo is a string graph-based *de novo* assembler for long reads[32]. Flye relies on a repeat graph data structure that also tolerates more sequencing errors[27]. To date, few sequencing examples used more than three assemblers to construct large plant genomes, and available information on how to select *de novo* assembly tools or evaluate the quality of an assembled genome using ONT data is limited[33,34].

To address this limitation, we report the *de novo* sequencing and assembly of representative hexaploid plant genome *Hibiscus syriacus* – a flowering plant species of the Malvaceae family – using two long read sequencing platforms: PacBio Sequel for *H. syriacus* cultivars cv. Baekdansim and ONT for cv. Gangneung. Due to the physically tough leaf tissues, which contain high levels of polysaccharides and phenolic compounds, it was challenging to isolate pure HMW DNA from *H. syriacus*. Therefore, we tested and optimized an efficient combination of intact HMW DNA isolation and sequencing using the ONT and PacBio. Additionally, we combined multiple assemblers to select the best genome assembly for the successful sequencing. To date, one *H. syriacus* genome was sequenced using short-read sequencing covering 92% of the genome with 1.7% gap sequences[35]. In the current study, the genome size (from 1.75 Gb in Illumina to 1.87 Gb in Nanopore) and contig N50 (from 140 kb to 8.1 Mb) were remarkably increased using ONT sequencing data (Table 1). Sanger sequencing evaluation revealed that the tandem repeat sequences missing from the Illumina-generated genome were successfully and accurately assembled into the new genomes with error correction by polishing. This *de novo* genome assembly strategy based on long-read ONT sequencing allows for construction of contiguous, improved-quality genomes. To date, genome assembly has been carried out for various species within the *Gossypium*, *Hibiscus*, and *Corchorus* genera of the Malvaceae family, ranging from scaffold level to chromosome level[11,36–39]. Our high-quality *H. syriacus* genomes provide an essential model to develop an effective strategy for polyploid plant genome assembly. These resources will provide valuable insight into functional genomics and evolutionary studies within the Malvaceae family.

Methods

DNA preparation.

High-molecular-weight (HMW) genomic DNA was extracted from leaf tissues of *H. syriacus* plants. We first isolate nuclei from *H. syriacus* plant cells using NIBM (10 mM Tris-HCl pH8.0, 10 mM EDTA pH8.0, 100 mM KCL, 0.5 M sucrose, 4 mM spermidine, 1 mM spermine, 0.15% ß-mercaptoethanol) buffer. From these intact nuclei, we successfully obtained high-quality genomic DNA using lysis buffer (50 mM Tris-HCl pH 7.5, 1.4 M NaCl, 20 mM EDTA pH 8.0, 0.5% SDS). Absorbance ratios to determine DNA quality ranged from 1.8 to 2.0 at A260/280 nm and from 2.0 to 2.2 at A260/230 nm using a Nanodrop spectrophotometer (Thermo Scientific). DNA size was assessed using a TapeStation system (Agilent). Most genomic DNA (gDNA) fragments were distributed between 10 and 100 kb.

Library construction and sequencing.

To investigate optimal conditions for constructing a reference genome using long reads, we sequenced two *H. syriacus* cv. Baekdansim and Gangneung, using two sequencing platforms: PacBio Sequel and ONT. *H. syriacus* cv. Baekdansim, which has more complex genome structure, was sequenced using PacBio Sequel and cv. Gangneung was sequenced using ONT (Fig. 1). While the sequencing protocol using PacBio Sequel is well-defined, ONT sequencing for large and complex plant genomes requires optimization[36,40]. Therefore, using R9.4 MinION (M) or PromethION (P), we compared three library construction methods for ONT sequencing to optimize the final distribution of read lengths and total throughput: non-sheared

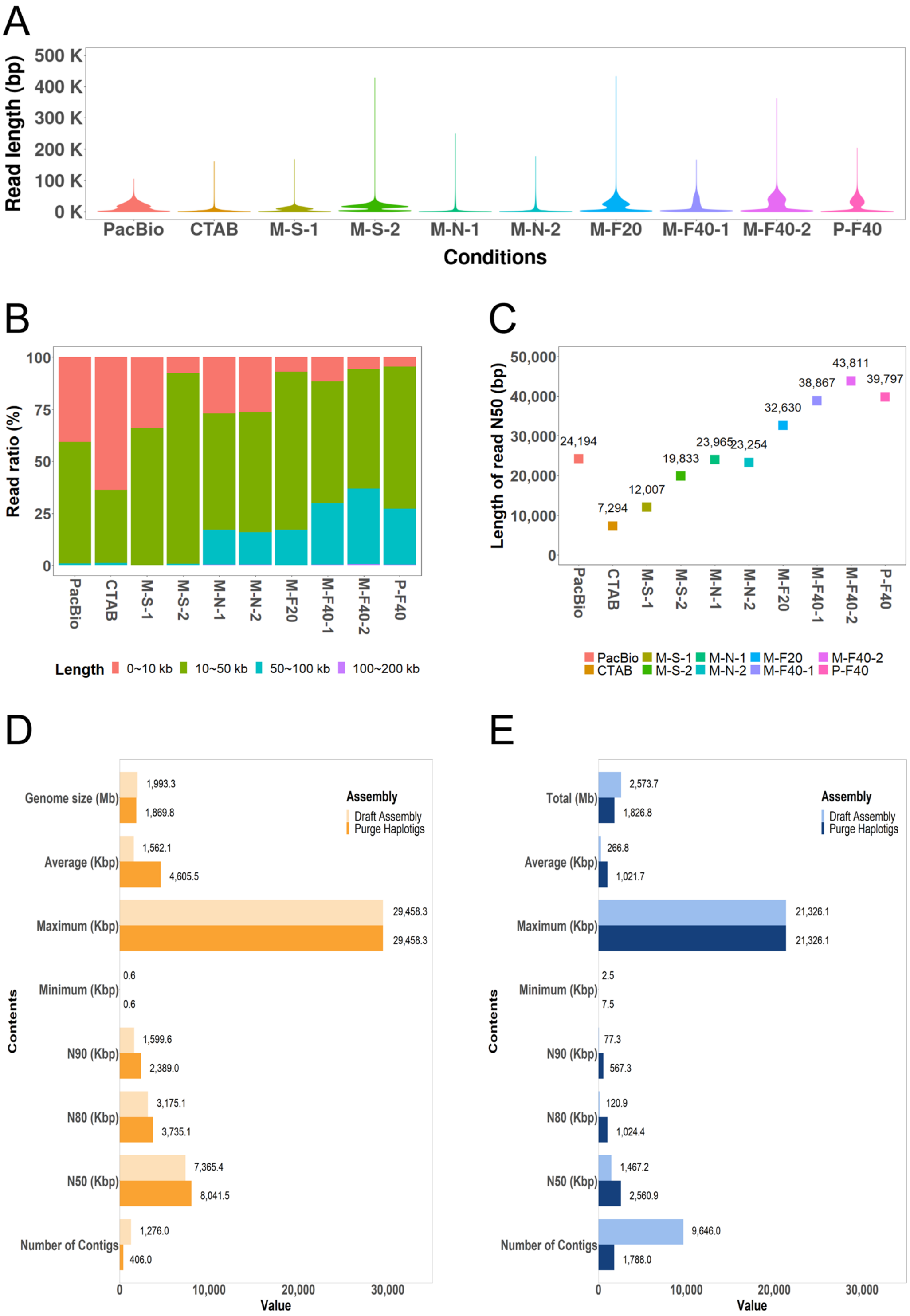


Fig. 1 Distribution of raw reads from long read sequencing and genome assembly. (**A**) Read distributions from each sequencing batch. The *x*-axis indicates Nanopore sequencing batch, and the *y*-axis indicates read length. M, MinION; P, PromethION; S, sheared; N, non-sheared; F20, 20 kb size fractionation; F40, 40 kb size fractionation. (**B**) Read length ratio for each sequencing batch. The *x*-axis indicates ratio of read length, and the *y*-axis indicates sequencing batch. (**C**) Read N50 values for each sequencing batch. (**D,E**) Genome assembly improvements by Purge Haplotigs with ONT (**D**) and PacBio (**E**). The x-axis indicates value of parameters, and the y-axis indicates genome quality parameters.

gDNA (N), sheared gDNA (S), and size fractionation of non-sheared gDNA (F20 and F40). For shearing, 10 µg of pure HMW DNA was processed through a g-TUBE (Covaris). For size-selective batches, a BluePippin system

	Baekdansim (PacBio)	Gangneung (ONT)
Assembler	MECAT	NECAT
Raw read coverage	74X (148.1 Gb)	63.4X (120.5 Gb)
Number of Contigs	1,788	406
N50 (bp)	2,560,905	8,098,919
N80 (bp)	1,024,377	3,747,792
N90 (bp)	567,339	2,397,205
Minimum (bp)	7,536	594
Maximum (bp)	21,326,069	29,519,541
Average (bp)	1,021,683.3	4,623,639
Protein-coding genes	88,414	88,573
Total (bp)	1,826,769,743	1,877,197,537

Table 2. Genome assembly statistics for the multiple reference genomes of *Hibiscus syriacus*.

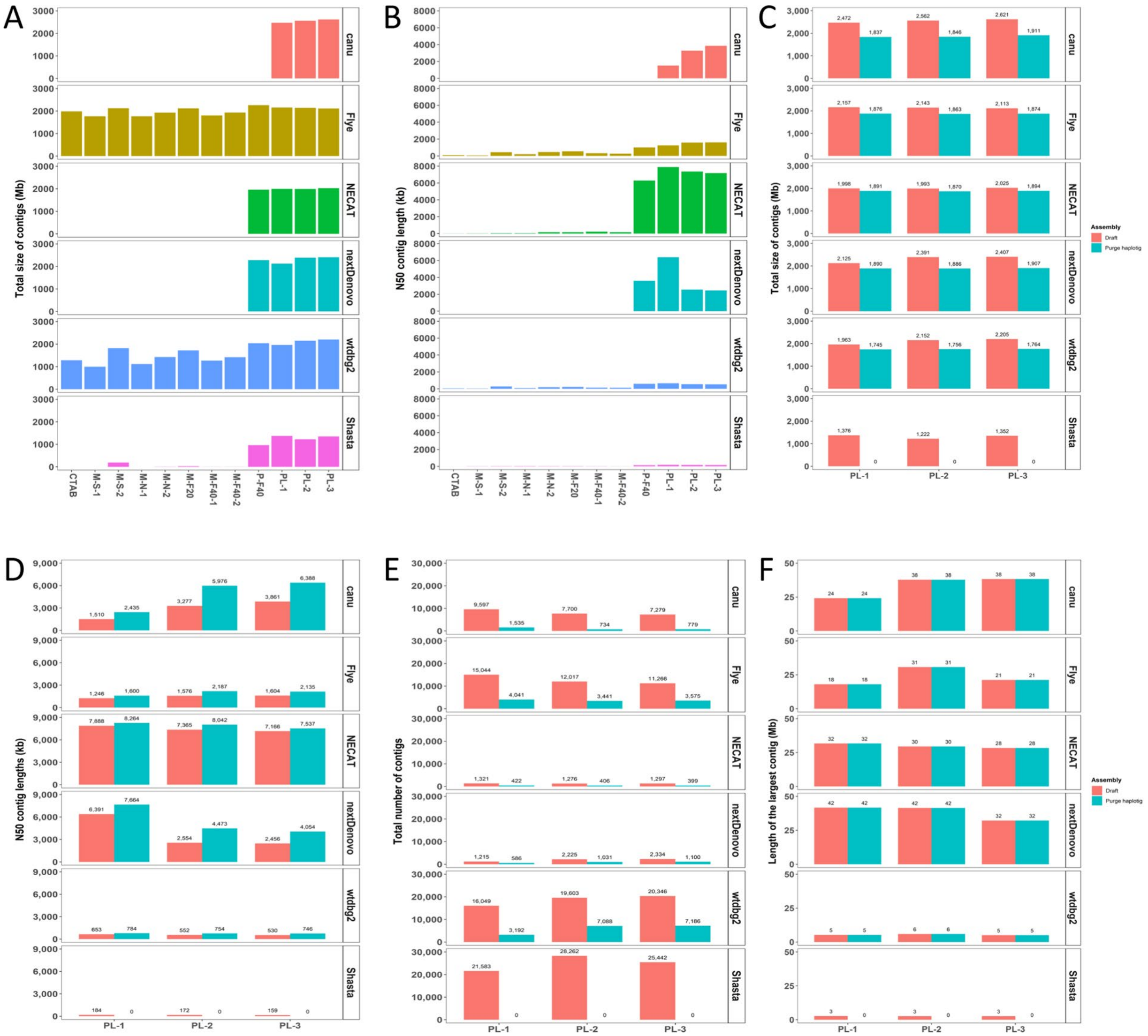


Fig. 2 Assembly statistics for Nanopore sequencing. (**A**) Assembled genome size of each genome assembler. The x-axis indicates nanopore sequencing batch and the y-axis indicates total contig size. (**B**) Contig N50 of assembled genomes. The x-axis indicates nanopore sequencing batch and the y-axis indicates contig N50 length of the genome. (**C**) Assembled genome size for Draft and Purge Haplotigs-processed genomes. (**D**) Read N50 for Draft and Purge Haplotigs-processed genomes. (**E**) Number of contigs in the Draft and Purge Haplotigs-processed genomes. (**F**) Comparison of the longest contig lengths for Draft and Purge Haplotigs-processed genomes.

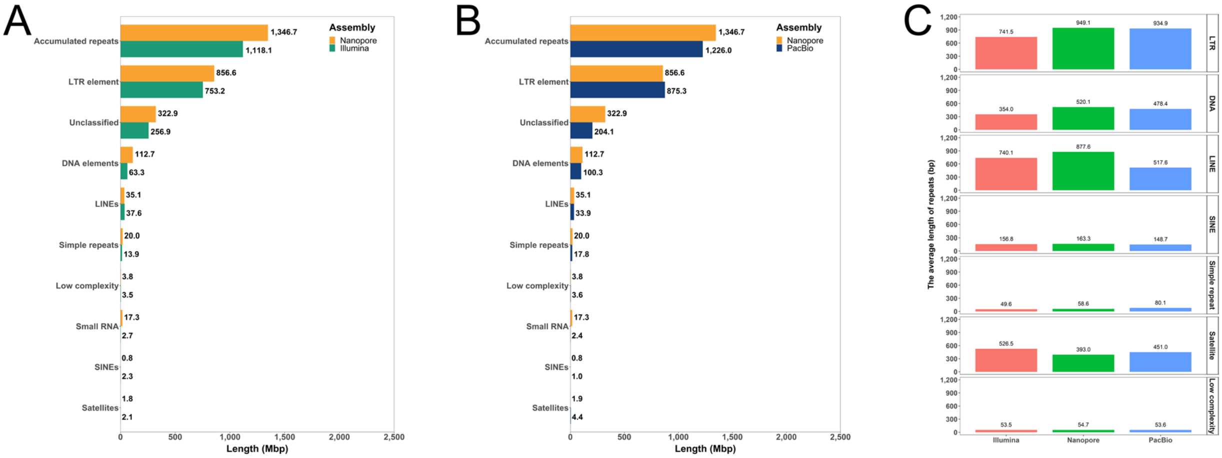


Fig. 3 Repetitive element annotation. (**A**) Identification of repetitive sequences in Nanopore (orange) and Illumina (green). (**B**) Identification of repetitive sequences in Nanopore (orange) and PacBio (indigo). (**c**) The average repeat length distribution of each repetitive element types.

Organelle	Species	Accession ID	Genome size (bp)	Number of mapped read(%)
Chloroplast	*Gossypium arboreum*	NC_016712.1	160,230	0.3193
Chloroplast	*Gossypium hirsutum*	NC_007944.1	160,301	0.3193
Chloroplast	*Gossypium raimondii*	NC_016668.1	160,161	0.3193
Chloroplast	*Hibiscus syriacus*	NC_026909.1	161,019	0.3251
Mitochondrion	*Gossypium arboreum*	NC_035073.1	687,482	0.0039
Mitochondrion	*Gossypium hirsutum*	NC_027406.1	668,584	0.004
Mitochondrion	*Gossypium raimondii*	NC_029998.1	676,078	0.004
Mitochondrion	*Hibiscus cannabinus*	NC_035549.1	569,915	0.0052

Table 3. Investigation of organellar DNA contamination.

(Sage Science) was used with 10 μg of DNA without a shear step, followed by the selection of fragments >20 kb or >40 kb, and DNA recovery from the elution well (Fig. 1A,B). Sequencing libraries were prepared according to recommendations by ONT. The cetyltrimethylammonium bromide (CTAB) precipitation method, which eliminates polysaccharides from plant tissues[41], was used as a control to evaluate the improvement of sequencing quality. The genomic DNA extracted using CTAB buffer underwent MinION sequencing after a DNA shearing step. MinION sequencing was performed as per the manufacturer's guidelines using R9.4 SpotON Flow Cell (FLO-MIN106) and controlled using ONT MinKOW software. The final library batch was sequenced using PromethION (FLO-PRO002). Concerning high-quality DNA yields, longer N50 of read lengths were obtained with non-sheared DNA samples than with sheared DNA samples. DNA fractionation for library construction remarkably increased N50 of read lengths, with a maximum value of 43 kb (Fig. 1C).

Genome assembly. *De novo* assembly of *H. syriacus* cv. Baekdansim was performed with MECAT2, a time-efficient genome assembler known as the fast mapping, error-correction, and *de novo* assembly tool[42] (Table 2). For cv. Gangneung genome assembly, we compared six *de novo* genome assemblers, canu (v.2.0), Flye (v.2.8.3), NECAT (v.0.0.1), nextDenovo, wtdbg2 (v2.5), and Shasta (v.0.1.0)) using individual sequencing data and combining datasets of MinION data (PL-1), size fraction data (PL-2), and whole data (PL-3) (Fig. 2A). Of the assemblers we tested, NECAT and nextDenovo performed significantly better than the others. Specifically, both showed a high level of contig N50 length over 8 Mb, whereas the other assemblers had lower contig N50 lengths (Fig. 2B). The comparison between PacBio and ONT showed that genome assembly using ONT generated a longer contig N50 and fewer contig numbers compared to PacBio (Table 2).

Regional duplication due to regional heterogeneity is one of the major barriers to genome assembly using long reads, especially in polyploid genomes that contain native duplicated chromosomes. These haplotype-fused contigs lead to larger genome assemblies and can be problematic for downstream analysis[43,44]. In our case, where the regional heterogeneity of sequenced reads was very high, and five genome assemblers assembled these regions into separate contigs. Canu and nextDenovo generated more haplotype-fused contigs compared to other genome assemblers (Fig. 2C,D). To remove heterozygous sequences from the assembled genome, reassignment of redundant contigs in the primary assembly was performed using Purge Haplotigs[44] with read-depth cutoffs of 10 (low), 53 (mid), and 110 (high). After processing with Purge Haplotigs, the final *H. syriacus* cv. Gangneung genome size was smaller than predicted, and contig N50 lengths were improved with a remarkably fewer number of contigs (Fig. 2C–F). Remarkably higher content of haplotype-fused contigs was detected in

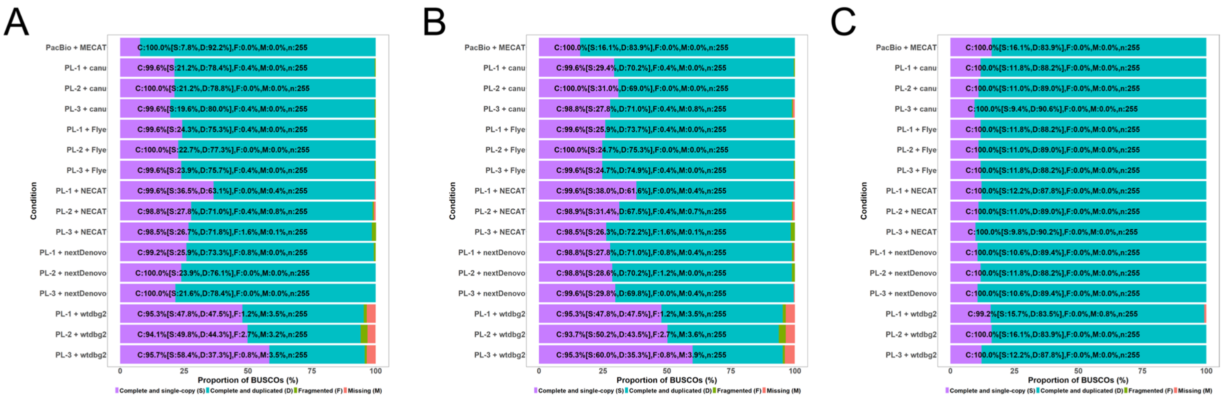


Fig. 4 Quality assessment of the genome assembly. (**A**) Outputs from assemblies without further correction. (**B**) Assemblies corrected using Purge Haplotigs. (**C**) Assemblies corrected using NextPolish. The x-axis indicates the proportion of BUSCOs and the y-axis indicates individual genome assembly and their modified version. Purple shows the percentage of complete and single copy genes, the turquoise shows percentage of complete and duplicated genes, green shows the percentage of fragmented genes, and coral shows the percentage of of missing genes in the assemblies.

cv. Baekdansim assembled using PacBio Sequel sequencing data compared to cv. Gangneung using ONT data (Fig. 1D,E). Then, the contigs of Gangneung were polished with ONT raw data thrice using nextPolish[45] v.1.01 and two times with filtered Illumina reads used in the previous study[35].

Repetitive element annotation. In general, repetitive sequences lead to genome assembly errors and automated gene annotation caused by frame shift of genes containing microsatellite sequences[46]. These barriers to genome assembly can be overcome using long-read sequencing technology. Thus, we compared the repeat content of previous and newly assembled genomes using repeat annotation analysis (Fig. 3). Repeat annotation was implemented using RepeatModeler and RepeatMasker as described previously[35]. After RepeatModeler was used to construct a repeat library with the assembled genomes, repeat annotation was performed using RepeatMasker (http://www.repeatmasker.org). We found that more repetitive sequences (about 215 Mb) of almost all categories were detected in the newly assembled genome using long-read sequencing (Fig. 3A). Further comparison of PacBio and ONT revealed higher repetitive contents and genome coverage in genome assembly using ONT (Fig. 3B). The average length of repetitive contents was also longer in ONT compared to Illumina sequencing (Fig. 3C).

Genome annotation. Annotation was performed using the KOBIC annotation pipeline[35] and consisted of repeat masking, mapping of different protein sequence sets, and *ab initio* gene prediction performed by AUGUSTUS v3.2.3[47]. Transcript assembly was performed by reference-based algorithm using HISAT2[48] and StringTie[49] with the assembled genome and RNA-Seq data in previous study[35]. The protein sequences of *Arabidopsis thaliana* (TAIR10, http://www.arabidopsis.org), *Theobroma cacao*[50], *Gossypium raimondii*[36], and *H. syriacus*[35] were mapped using GeneWise v2.1[51] to generate protein-based gene models for consensus modeling. AUGUSTUS was used for gene prediction. Subsequently, the predicted gene models from AUGUSTUS were validated using BLASTp with protein sequences from the four genomes (*T. cacao, G. raimondii, H. syriacus* and *A. thaliana*) as queries, and erratic gene models were filtered with a BLASTp cut-off value of query coverage ≥ 0.3. The assembled transcripts were validated using tBLASTn against the above-listed four protein sets and were filtered with query coverage ≥ 0.5 and subject coverage ≥ 0.3. The predicted gene models from GeneWise were also filtered using query coverage ≥ 0.3. The remaining gene models from GeneWise were reformatted to GFF3 format and were used to determine the consensus gene model via EVidenceModeler (EVM)[52]. EVM combines *ab initio* gene predictions with protein alignments into weighted consensus gene structures.

Data Records

The assembled genome sequence of *H. syriacus* cv. Baekdansim has been deposited at NCBI, GenBank, under accession number VEPZ02000000[53]. All PacBio and Illumina raw read files are available through the NCBI Sequence Read Archive (SRA) with the identifier SRP193812[54]. The assembled genome sequence of *H. syriacus* cv. Gangneung has been deposited at NCBI, GenBank, under accession number JAUEMI000000000[55]. The Nanopore raw data are available through the NCBI Sequence Read Archive with the identifier SRP087036[56]. Both genome assembly and gene annotation results for *H. syriacus* cv. Gangneung and *H. syriacus* cv. Baekdansim are available at the online open access repository figshare database[57,58].

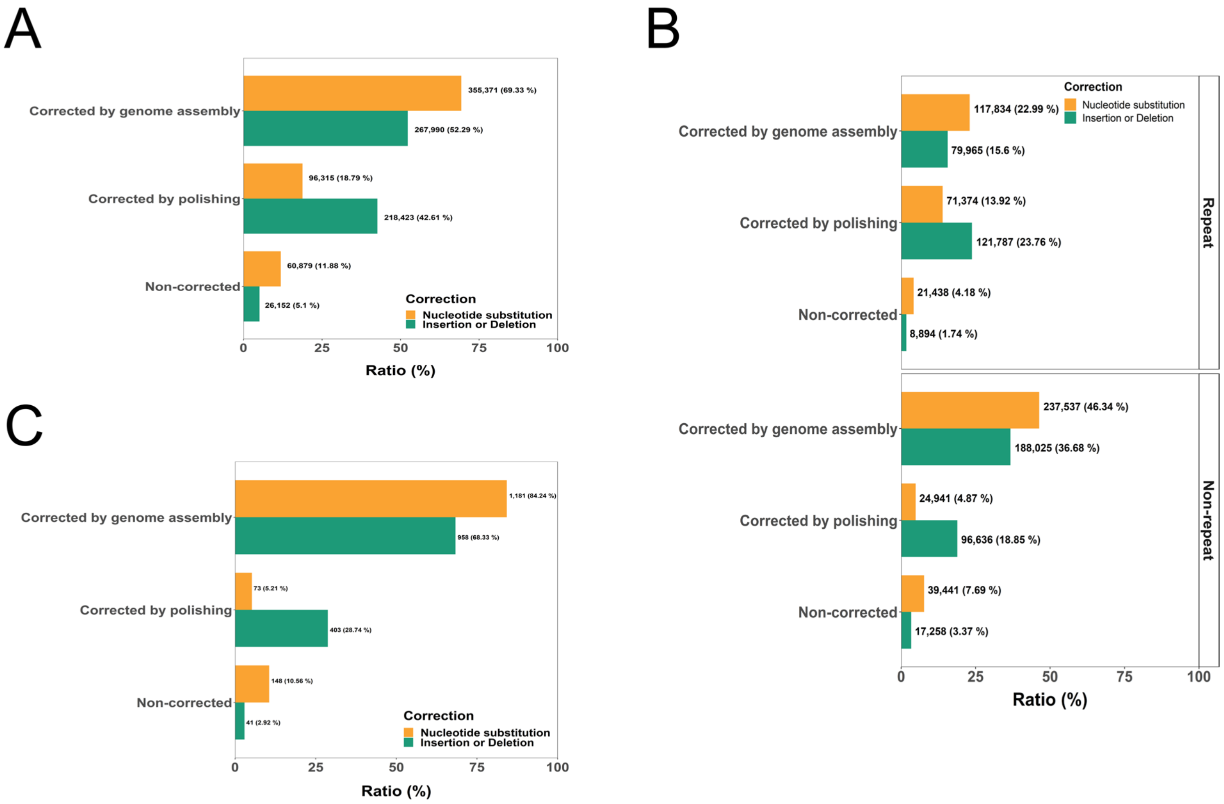


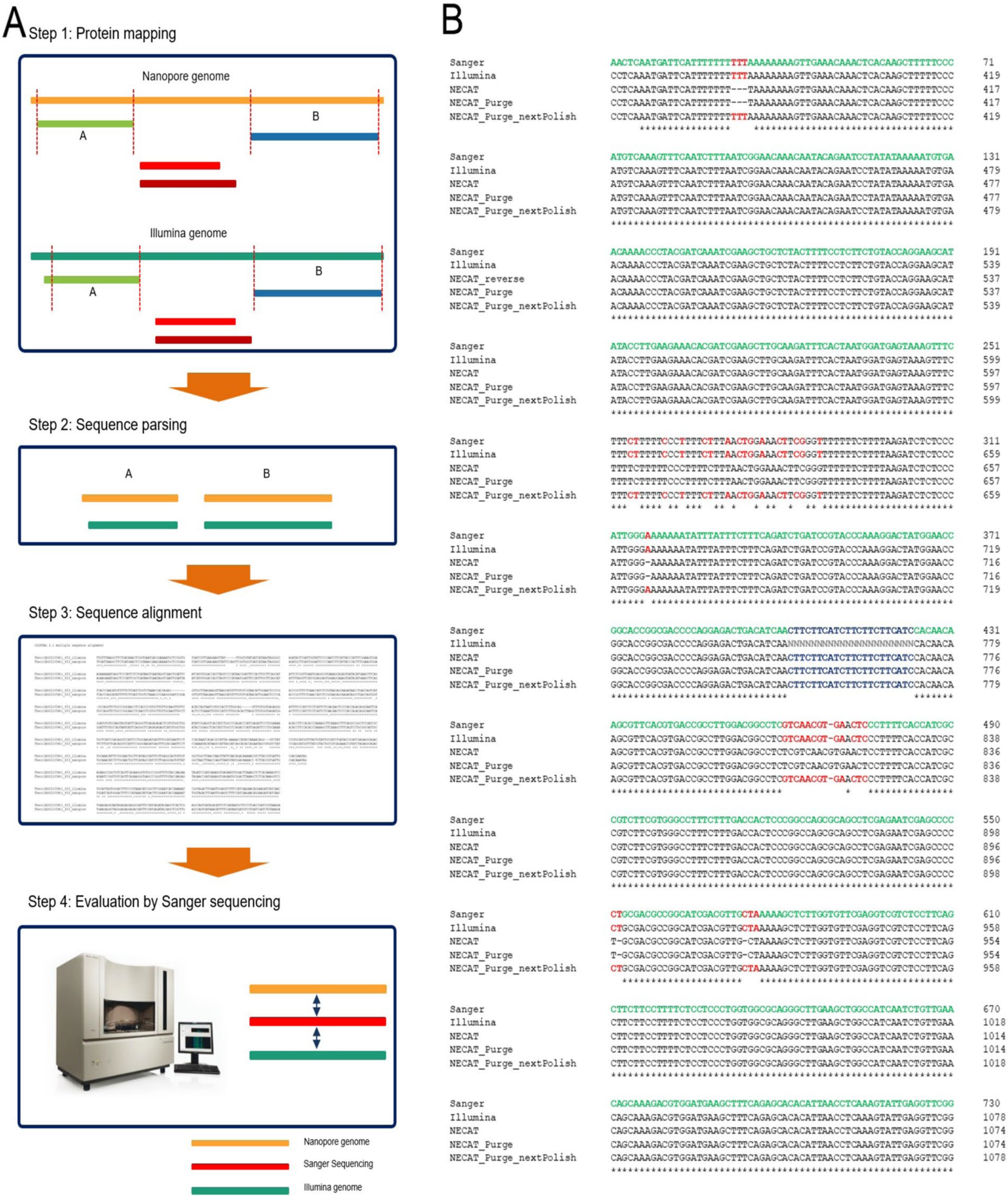
Fig. 5 Genome assembly improvements by nanopore sequencing. (**A**) Identification of homopolymeric regions in draft genome assembly using NECAT, polished genome assembly, and Illumina genome assembly. (**B**) Homopolymer distribution in repetitive and non-repetitive sequences. (**C**) Identification of homopolymers in putative genic regions. Orange represents nucleotide substitutions and green represents nucleotide insertions or deletions.

Technical Validation

Distribution of raw reads and detection of potentially contaminated sequences. The distribution of reads indicated that the ratio of ultra-long reads (longer than 50 kb) in non-sheared pure HMW DNA with size fractionation conditions accounted for up to 36% of the total, which was significantly higher than other conditions, particularly compared to the general CTAB method (Fig. 1). Raw data read distributions showed that ONT generated a remarkably higher content of ultra-long reads, whereas PacBio Sequel generated a high content of relatively small long-read sizes ranging from 10 to 20 kb (Fig. 1A–C). Collectively, the high content of ultra-long reads might contribute to a more precise assembly of the cv. Gangneung genome by reducing the content of haplotype-fused contigs. The contig N50 read length and average contig length of both genome assemblies were improved by Purge Haplotigs (Fig. 1D,E).

In conventional DNA extraction methods, the step of removing organelles was absent, leading to relatively high contamination of organelle genomes as reported in the previous study[59]. To assess the proportion of organelle genomes, we mapped the ONT raw read data to the previously reported organelle genomes of *H. syriacus* and its related species. These results revealed that the prepared nuclear DNA had a notably small amount of unwanted organelle DNA contamination (Table 3). In plants, input data usually consist of 5–20% of unwanted organelle DNA reads, such as chloroplast and mitochondrial sequences[59]. These contaminant DNA reads led to over estimation of sequence coverage for genome assembly. Therefore, additional sequence data may be necessary to increase the quality of complex genome assembly. In our case, high-quality sequencing results were obtained with non-sheared, pure HMW DNA with an appropriate size fraction, resulting in a remarkably increased N50 of read length, maximum value of 43 kb (Fig. 1A,B). As unwanted organelle fragments were already filtered out, the *H. syriacus* genome could be fully completed even at the apparent 63.4× (120.5 Gb) coverage, which is relatively low depth compared to previous studies (Table 1).

Quality assessment of the genome assembly. To assess genome assembly completeness, we performed BUSCO (Benchmarking Universal Single-Copy Orthologs) v5.3.2[60] assessment of the assembled genomes using PacBio Sequel data and the combined ONT datasets (Fig. 4). Genome completeness ranged from 94% to 100% and revealed that wtdbg2 yielded the lowest completeness values (Fig. 4A). Relatively high levels of genome completeness were observed in canu, Flye, NECAT, and nextDenovo. This pattern was also seen in haplotype-fused contigs-removed genomes (Fig. 4B). The ratio of single- and duplicated-complete copy genes was decreased in



```
Sanger                   AACTCAATGATTCATTTTTTTTTTAAAAAAAAGTTGAAACAAACTCACAAGCTTTTTCCC   71
Illumina                 CCTCAAATGATTCATTTTTTTTTTAAAAAAAAGTTGAAACAAACTCACAAGCTTTTTCCC   419
NECAT                    CCTCAAATGATTCATTTTTTT---TAAAAAAAGTTGAAACAAACTCACAAGCTTTTTCCC   417
NECAT_Purge              CCTCAAATGATTCATTTTTTT---TAAAAAAAGTTGAAACAAACTCACAAGCTTTTTCCC   417
NECAT_Purge_nextPolish   CCTCAAATGATTCATTTTTTTTTTAAAAAAAAGTTGAAACAAACTCACAAGCTTTTTCCC   419
                              ****************    ***********************************

Sanger                   ATGTCAAAGTTTCAATCTTTAATCGGAACAAACAATACAGAATCCTATATAAAAATGTGA   131
Illumina                 ATGTCAAAGTTTCAATCTTTAATCGGAACAAACAATACAGAATCCTATATAAAAATGTGA   479
NECAT                    ATGTCAAAGTTTCAATCTTTAATCGGAACAAACAATACAGAATCCTATATAAAAATGTGA   477
NECAT_Purge              ATGTCAAAGTTTCAATCTTTAATCGGAACAAACAATACAGAATCCTATATAAAAATGTGA   477
NECAT_Purge_nextPolish   ATGTCAAAGTTTCAATCTTTAATCGGAACAAACAATACAGAATCCTATATAAAAATGTGA   479
                         ************************************************************

Sanger                   ACAAAACCCTACGATCAAATCGAAGCTGCTCTACTTTTCCTCTTCTGTACCAGGAAGCAT   191
Illumina                 ACAAAACCCTACGATCAAATCGAAGCTGCTCTACTTTTCCTCTTCTGTACCAGGAAGCAT   539
NECAT_reverse            ACAAAACCCTACGATCAAATCGAAGCTGCTCTACTTTTCCTCTTCTGTACCAGGAAGCAT   537
NECAT_Purge              ACAAAACCCTACGATCAAATCGAAGCTGCTCTACTTTTCCTCTTCTGTACCAGGAAGCAT   537
NECAT_Purge_nextPolish   ACAAAACCCTACGATCAAATCGAAGCTGCTCTACTTTTCCTCTTCTGTACCAGGAAGCAT   539
                         ************************************************************

Sanger                   ATACCTTGAAGAAACACGATCGAAGCTTGCAAGATTTCACTAATGGATGAGTAAAGTTTC   251
Illumina                 ATACCTTGAAGAAACACGATCGAAGCTTGCAAGATTTCACTAATGGATGAGTAAAGTTTC   599
NECAT                    ATACCTTGAAGAAACACGATCGAAGCTTGCAAGATTTCACTAATGGATGAGTAAAGTTTC   597
NECAT_Purge              ATACCTTGAAGAAACACGATCGAAGCTTGCAAGATTTCACTAATGGATGAGTAAAGTTTC   597
NECAT_Purge_nextPolish   ATACCTTGAAGAAACACGATCGAAGCTTGCAAGATTTCACTAATGGATGAGTAAAGTTTC   599
                         ************************************************************

Sanger                   TTTCTTTTTCCCTTTTCTTTAACTGGAAACTTCGGGTTTTTTTCTTTTAAGATCTCTCCC   311
Illumina                 TTTCTTTTTCCCTTTTCTTTAACTGGAAACTTCGGGTTTTTTTCTTTTAAGATCTCTCCC   659
NECAT                    TTTTCTTTTTCCCTTTTCTTTAACTGGAAACTTCGGGTTTTTTCTTTTAAGATCTCTCCC   657
NECAT_Purge              TTTTCTTTTTCCCTTTTCTTTAACTGGAAACTTCGGGTTTTTTCTTTTAAGATCTCTCCC   657
NECAT_Purge_nextPolish   TTTCTTTTTCCCTTTTCTTTAACTGGAAACTTCGGGTTTTTTTCTTTTAAGATCTCTCCC   659
                         ***  **** ** ***  ** *   * **  *  ** ***********************

Sanger                   ATTGGGAAAAAAATATTTATTTCTTTCAGATCTGATCCGTACCCAAAGGACTATGGAACC   371
Illumina                 ATTGGGAAAAAAATATTTATTTCTTTCAGATCTGATCCGTACCCAAAGGACTATGGAACC   719
NECAT                    ATTGGG-AAAAAATATTTATTTCTTTCAGATCTGATCCGTACCCAAAGGACTATGGAACC   716
NECAT_Purge              ATTGGG-AAAAAATATTTATTTCTTTCAGATCTGATCCGTACCCAAAGGACTATGGAACC   716
NECAT_Purge_nextPolish   ATTGGGAAAAAAATATTTATTTCTTTCAGATCTGATCCGTACCCAAAGGACTATGGAACC   719
                         ****** *****************************************************

Sanger                   GGCACCGGCGACCCCAGGAGACTGACATCAACTTCTTCATCTTCTTCTTCATCCACAACA   431
Illumina                 GGCACCGGCGACCCCAGGAGACTGACATCAANNNNNNNNNNNNNNNNNNNNNNCACAACA   779
NECAT                    GGCACCGGCGACCCCAGGAGACTGACATCAACTTCTTCATCTTCTTCTTCATCCACAACA   776
NECAT_Purge              GGCACCGGCGACCCCAGGAGACTGACATCAACTTCTTCATCTTCTTCTTCATCCACAACA   776
NECAT_Purge_nextPolish   GGCACCGGCGACCCCAGGAGACTGACATCAACTTCTTCATCTTCTTCTTCATCCACAACA   779
                         *******************************                      *******

Sanger                   AGCGTTCACGTGACCGCCTTGGACGGCCTCGTCAACGT-GAACTCCCTTTTCACCATCGC   490
Illumina                 AGCGTTCACGTGACCGCCTTGGACGGCCTCGTCAACGT-GAACTCCCTTTTCACCATCGC   838
NECAT                    AGCGTTCACGTGACCGCCTTGGACGGCCTCTCGTCAACGTGAACTCCTTTTCACCATCGC   836
NECAT_Purge              AGCGTTCACGTGACCGCCTTGGACGGCCTCTCGTCAACGTGAACTCCTTTTCACCATCGC   836
NECAT_Purge_nextPolish   AGCGTTCACGTGACCGCCTTGGACGGCCTCGTCAACGT-GAACTCCCTTTTCACCATCGC   838
                         ******************************           *   ***************

Sanger                   CGTCTTCGTGGGCCTTTCTTTGACCACTCCCGGCCAGCGCAGCCTCGAGAATCGAGCCCC   550
Illumina                 CGTCTTCGTGGGCCTTTCTTTGACCACTCCCGGCCAGCGCAGCCTCGAGAATCGAGCCCC   898
NECAT                    CGTCTTCGTGGGCCTTTCTTTGACCACTCCCGGCCAGCGCAGCCTCGAGAATCGAGCCCC   896
NECAT_Purge              CGTCTTCGTGGGCCTTTCTTTGACCACTCCCGGCCAGCGCAGCCTCGAGAATCGAGCCCC   896
NECAT_Purge_nextPolish   CGTCTTCGTGGGCCTTTCTTTGACCACTCCCGGCCAGCGCAGCCTCGAGAATCGAGCCCC   898
                         ************************************************************

Sanger                   CTGCGACGCCGGCATCGACGTTGCTAAAAAGCTCTTGGTGTTCGAGGTCGTCTCCTTCAG   610
Illumina                 CTGCGACGCCGGCATCGACGTTGCTAAAAAGCTCTTGGTGTTCGAGGTCGTCTCCTTCAG   958
NECAT                    T-GCGACGCCGGCATCGACGTTG-CTAAAAGCTCTTGGTGTTCGAGGTCGTCTCCTTCAG   954
NECAT_Purge              T-GCGACGCCGGCATCGACGTTG-CTAAAAGCTCTTGGTGTTCGAGGTCGTCTCCTTCAG   954
NECAT_Purge_nextPolish   CTGCGACGCCGGCATCGACGTTGCTAAAAAGCTCTTGGTGTTCGAGGTCGTCTCCTTCAG   958
                           *********************   **********************************

Sanger                   CTTCTTCCTTTTCTCCTCCCTGGTGGCGCAGGGCTTGAAGCTGGCCATCAATCTGTTGAA   670
Illumina                 CTTCTTCCTTTTCTCCTCCCTGGTGGCGCAGGGCTTGAAGCTGGCCATCAATCTGTTGAA   1018
NECAT                    CTTCTTCCTTTTCTCCTCCCTGGTGGCGCAGGGCTTGAAGCTGGCCATCAATCTGTTGAA   1014
NECAT_Purge              CTTCTTCCTTTTCTCCTCCCTGGTGGCGCAGGGCTTGAAGCTGGCCATCAATCTGTTGAA   1014
NECAT_Purge_nextPolish   CTTCTTCCTTTTCTCCTCCCTGGTGGCGCAGGGCTTGAAGCTGGCCATCAATCTGTTGAA   1018
                         ************************************************************

Sanger                   CAGCAAAGACGTGGATGAAGCTTTCAGAGCACACATTAACCTCAAAGTATTGAGGTTCGG   730
Illumina                 CAGCAAAGACGTGGATGAAGCTTTCAGAGCACACATTAACCTCAAAGTATTGAGGTTCGG   1078
NECAT                    CAGCAAAGACGTGGATGAAGCTTTCAGAGCACACATTAACCTCAAAGTATTGAGGTTCGG   1074
NECAT_Purge              CAGCAAAGACGTGGATGAAGCTTTCAGAGCACACATTAACCTCAAAGTATTGAGGTTCGG   1074
NECAT_Purge_nextPolish   CAGCAAAGACGTGGATGAAGCTTTCAGAGCACACATTAACCTCAAAGTATTGAGGTTCGG   1078
                         ************************************************************
```

Fig. 6 Identification of common genic regions between Illumina and nanopore assembly by NECAT. (**A**) Protein mapping were carried out using proteins of *Theobroma cacao* and *Gossypium raimondii* in step 1. In step 2, extraction of genome sequences from both genomes and sequence alignment were performed. Then, assembly errors by homopolymers were identified. In step 4, Sanger sequencing of sequences containing assembly errors were performed and multiple sequence alignments were carried out using Sanger sequencing results and genic sequences from Illumina and nanopore genomes. (**B**) Identification of assembly errors by homopolymers in genic regions. Sanger sequencing was used to identify the common genic region and multiple sequence alignment was performed to identify assembly errors by homopolymers in each version of genome assembly. Green, red, and blue represent Sanger sequencing results, assembly errors by homopolymers, and repetitive sequences missing in the Illumina genome, respectively.

polished genomes (Fig. 4C) as a result of recovering the collinearity of the duplicated gene by correcting the sequencing error using a polishing program. This correction step reduced fragmented and missing BUSCO genes and greatly facilitated genome completeness for all assemblies. Among them, genome assembly by NECAT using size fraction data showed a high contig N50 length with fewer contigs and high levels of genome assembly completeness, and was therefore selected for the next repeat sequence comparison.

Genome assembly improvements by nanopore sequencing. Despite the development of an improved version of MinION[61], long-read sequencing in previous studies showed a relatively low accuracy of 85–95%[62–64]. A common sequencing error is due to homopolymeric regions, or short repeat regions, which account for about half of all sequencing errors[61]. Despite vast improvements in raw error rate, assembled sequences still contain homopolymers, which are known to cause frameshift errors during gene annotation[65]. Polishing is one solution to correct these errors. To detect homopolymers and investigate their correction by polishing, each raw read batch was aligned to each version of draft genomes using Minimap2[66] with default parameters for ONT sequencing and options (–secondary = no–sam-hit-only) to discard unmapped reads and perform secondary alignment as previously reported[61] (Fig. 5). In total, 512,565 of 1,000 kb fragments containing homopolymeric regions were identified, and these homopolymers were mainly caused by sequencing errors in the newly assembled genome (Fig. 5A). Importantly, most sequencing errors were corrected during either the genome assembly or polishing step. A large proportion of nucleotide substitutions (69.33%) were corrected during genome assembly, whereas about half of the insertions or deletions were corrected by polishing. These homopolymers may arise due to low sequencing quality in the later stages of sequencing[67]. A higher proportion of homopolymers was detected in the non-repetitive sequences, including both genic and untranslated regions (UTR) (Fig. 5B,C).

We further investigated this type of error by evaluating common genic regions using tBlastN analysis with default parameters. For this analysis, we used proteins from the *Theobroma cacao* and *Gossypium raimondii* and Sanger sequencing (Fig. 6). The evaluation revealed tandem repeat sequences, which were missing from the previous genome, were successfully and accurately assembled in the new genomes (Fig. 6). Although nucleotide substitutions, deletions, or insertions were confirmed by Sanger sequencing, assembly errors due to this type of homopolymer were corrected during a polishing step (Figs. 5, 6). Collectively, these data suggest the application of Nanopore long-read sequencing technology has improved the construction of a reference plant genome, *H. syriacus*. Notably, about 215 Mb of repetitive sequences were incorporated into the newly assembled genome (Figs. 3A,B, 6).

Code availability

All software used for data processing were executed following the manual of the bioinformatic software cited above, and all commands used to assemble the genome are available in figshare[68]. If no detailed parameters are described for the software, the default parameters were used. Additionally, R codes used for figure construction are also available in figshare[68].

Received: 16 June 2023; Accepted: 11 October 2023;
Published online: 18 October 2023

References

1. Aury, J.-M. *et al.* Long-read and chromosome-scale assembly of the hexaploid wheat genome achieves high resolution for research and breeding. *GigaScience* **11**, giac034 (2022).
2. Faulk, C. De novo sequencing, diploid assembly, and annotation of the black carpenter ant, Camponotus pennsylvanicus, and its symbionts by one person for $1000, using nanopore sequencing. *Nucleic acids research* **51**, 17–28 (2023).
3. Kress, W. J. *et al.* Green plant genomes: What we know in an era of rapidly expanding opportunities. *Proceedings of the National Academy of Sciences* **119**, e2115640118 (2022).
4. Pucker, B., Irisarri, I., de Vries, J. & Xu, B. Plant genome sequence assembly in the era of long reads: Progress, challenges and future directions. *Quantitative Plant Biology* **3**, e5 (2022).
5. Huddleston, J. *et al.* Reconstructing complex regions of genomes using long-read sequencing technology. *Genome research* **24**, 688–696 (2014).
6. Michael, T. P. & VanBuren, R. Building near-complete plant genomes. *Current Opinion in Plant Biology* **54**, 26–33 (2020).
7. Marks, R. A., Hotaling, S., Frandsen, P. B. & VanBuren, R. Representation and participation across 20 years of plant genome sequencing. *Nature plants* **7**, 1571–1578 (2021).
8. *Plabipd*. http://www.plabipd.de/timeline_view.ep (2014).
9. Schmidt, M. H.-W. *et al.* De novo assembly of a new Solanum pennellii accession using nanopore sequencing. *The Plant Cell* **29**, 2336–2348 (2017).
10. Wang, W. *et al.* The draft nuclear genome assembly of Eucalyptus pauciflora: a pipeline for comparing de novo assemblies. *Gigascience* **9**, giz160 (2020).
11. Udall, J. A. *et al.* De novo genome sequence assemblies of Gossypium raimondii and Gossypium turneri. *G3: Genes, Genomes, Genetics* **9**, 3079–3085 (2019).
12. Xu, L. *et al.* A chromosome-level genome assembly of radish (Raphanus sativus L.) reveals insights into genome adaptation and differential bolting regulation. *Plant Biotechnology Journal* **21**, 990–1004 (2023).
13. Yuanying, P. *et al.* Reference genome assemblies reveal the origin and evolution of allohexaploid oat. (2021).
14. Kersey, P. J. Plant genome sequences: past, present, future. *Current opinion in plant biology* **48**, 1–8 (2019).
15. Jiao, W.-B. & Schneeberger, K. The impact of third generation genomic technologies on plant genome assembly. *Current opinion in plant biology* **36**, 64–70 (2017).
16. McCann, J. *et al.* Differential genome size and repetitive DNA evolution in diploid species of Melampodium sect. Melampodium (Asteraceae). *Frontiers in Plant Science* **11**, 362 (2020).
17. Pellicer, J., Fay, M. F. & Leitch, I. J. The largest eukaryotic genome of them all? *Botanical Journal of the Linnean Society* **164**, 10–15 (2010).
18. Friar, E.A. Isolation of DNA from plants with large amounts of secondary metabolites. in *Methods in enzymology*, **Vol. 395** 1–12 (Elsevier, 2005).
19. Healey, A., Furtado, A., Cooper, T. & Henry, R. J. Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant methods* **10**, 1–8 (2014).

20. Inglis, P. W., Pappas, M. D. C. R., Resende, L. V. & Grattapaglia, D. Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. *PloS one* **13**, e0206085 (2018).
21. Mayjonade, B. *et al.* Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *Biotechniques* **61**, 203–205 (2016).
22. Schalamun, M. *et al.* Harnessing the MinION: An example of how to establish long-read sequencing in a laboratory using challenging plant tissue from Eucalyptus pauciflora. *Molecular ecology resources* **19**, 77–89 (2019).
23. Varma, A., Padh, H. & Shrivastava, N. Plant genomic DNA isolation: an art or a science. *Biotechnology Journal: Healthcare Nutrition Technology* **2**, 386–392 (2007).
24. Zhang, M. *et al.* Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. *nature protocols* **7**, 467–478 (2012).
25. Li, H. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**, 2103–2110 (2016).
26. Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research* **27**, 722–736 (2017).
27. Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology* **37**, 540–546 (2019).
28. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nature methods* **17**, 155–158 (2020).
29. Liu, H., Wu, S., Li, A. & Ruan, J. SMARTdenovo: a de novo assembler using long noisy reads. *Gigabyte* **2021** (2021).
30. Shafin, K. *et al.* Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nature biotechnology* **38**, 1044–1053 (2020).
31. Chen, Y. *et al.* Efficient assembly of nanopore reads via highly accurate and intact error correction. *Nature Communications* **12**, 60 (2021).
32. NextDeNovo. NextDeNovo. (2019).
33. Nagy, I. *et al.* Chromosome-scale assembly and annotation of the perennial ryegrass genome. *BMC genomics* **23**, 505 (2022).
34. Shearman, J. R. *et al.* A draft chromosome-scale genome assembly of a commercial sugarcane. *Scientific reports* **12**, 20474 (2022).
35. Kim, Y.-M. *et al.* Genome analysis of Hibiscus syriacus provides insights of polyploidization and indeterminate flowering in woody plants. *Dna Research* **24**, 71–80 (2017).
36. Chen, Z. J. *et al.* Genomic diversifications of five Gossypium allopolyploid species and their impact on cotton improvement. *Nature genetics* **52**, 525–533 (2020).
37. Sarkar, D. *et al.* The draft genome of Corchorus olitorius cv. JRO-524 (Navin). *Genomics Data* **12**, 151–154 (2017).
38. Sheng, K. *et al.* A reference-grade genome assembly for Gossypium bickii and insights into its genome evolution and formation of pigment glands and gossypol. *Plant Communications* **4** (2023).
39. Zhang, L. *et al.* The genome of kenaf (Hibiscus cannabinus L.) provides insights into bast fibre and leaf shape biogenesis. *Plant Biotechnology Journal* **18**, 1796–1809 (2020).
40. Marchant, D. B. *et al.* Dynamic genome evolution in a model fern. *Nature Plants* **8**, 1038–1051 (2022).
41. Murray, M. & Thompson, W. Rapid isolation of high molecular weight plant DNA. *Nucleic acids research* **8**, 4321–4326 (1980).
42. Xiao, C.-L. *et al.* MECAT: fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. *nature methods* **14**, 1072–1074 (2017).
43. Gan, H. M. *et al.* Best foot forward: nanopore long reads, hybrid meta-assembly, and haplotig purging optimizes the first genome assembly for the southern hemisphere blacklip abalone (Haliotis rubra). *Frontiers in genetics* **10**, 889 (2019).
44. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC bioinformatics* **19**, 1–10 (2018).
45. Hu, J., Fan, J., Sun, Z. & Liu, S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. *Bioinformatics* **36**, 2253–2255 (2020).
46. Tørresen, O. K. *et al.* Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic acids research* **47**, 10994–11006 (2019).
47. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research* **34**, W435–W439 (2006).
48. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* **37**, 907–915 (2019).
49. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature biotechnology* **33**, 290–295 (2015).
50. Argout, X. *et al.* The genome of Theobroma cacao. *Nature genetics* **43**, 101–108 (2011).
51. Birney, E., Clamp, M. & Durbin, R. GeneWise and genomewise. *Genome research* **14**, 988–995 (2004).
52. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* **9**, 1–22 (2008).
53. Kim, Y.-M. *Hibiscus syriacus* cultivar Baekdansim isolate YM2019G1, whole genome shotgun sequencing project. *GenBank* https://identifiers.org/ncbi/insdc:VEPZ00000000 (2019).
54. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP193812 (2019).
55. Koo, H. *et al. Hibiscus syriacus* isolate Gangneung, whole genome shotgun sequencing project. *GenBank* https://identifiers.org/ncbi/insdc:JAUEMI000000000 (2023).
56. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP087036 (2022).
57. Kim, Y.-M. *Hibiscus syriacus* cv. Gangneung Draft Genome. *figshare.* https://doi.org/10.6084/m9.figshare.23041847 (2023).
58. Kim, Y.-M. Hibiscus syriacus cv. Baekdansim Draft Genome. *figshare.* https://doi.org/10.6084/m9.figshare.23041751 (2023).
59. Soorni, A., Haak, D., Zaitlin, D. & Bombarely, A. Organelle_PBA, a pipeline for assembling chloroplast and mitochondrial genomes from PacBio DNA sequencing data. *BMC genomics* **18**, 1–8 (2017).
60. Manni, M., Berkeley, M. R., Seppey, M. & Zdobnov, E. M. BUSCO: assessing genomic data quality and beyond. *Current Protocols* **1**, e323 (2021).
61. Delahaye, C. & Nicolas, J. Sequencing DNA with nanopores: Troubles and biases. *PloS one* **16**, e0257521 (2021).
62. Giordano, F. *et al.* De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms. *Scientific reports* **7**, 3935 (2017).
63. Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology* **36**, 338–345 (2018).
64. Jain, M. *et al.* MinION Analysis and Reference Consortium: Phase 2 data release and analysis of R9. 0 chemistry. *F1000Research* **6** (2017).
65. Watson, M. & Warr, A. Errors in long-read assemblies can critically affect protein prediction. *Nature biotechnology* **37**, 124–126 (2019).
66. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
67. Amarasinghe, S. L. *et al.* Opportunities and challenges in long-read sequencing data analysis. *Genome biology* **21**, 1–16 (2020).
68. Kim, Y.-M. Code availability. *figshare.* https://doi.org/10.6084/m9.figshare.24105303 (2023).

Acknowledgements

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1I1A2044678), and the Korea Forest Service of the Korean government through the R&D Program for Forestry Technology (Project No. 2014071H10-2122-AA04) to Y.M.K and the National Research Foundation of Korea(NRF) grant funded by the Korea government (MSIT) (2023R1A2C1006404) to A.Y.S.

Author contributions

A.Y.S. and Y.M.K. conceived the project, designed the analysis, and organized the manuscript. A.Y.S. and S.Y.K. generated Nanopore raw data, H.K. and G.W.L. performed genome assembly and annotation. S.R.K., S.G. and Y.M.K. performed genome assembly evaluation. A.Y.S. and Y.M.K. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.-M.K. or A.-Y.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.