High-quality *Arabidopsis thaliana* genome assembly with Nanopore and HiFi long reads

Authors:

Bo Wang[1], Yanyan Jia[1], Peng Jia[1,2], Quanbin Dong[3], Xiaofei Yang[4], Kai Ye[1,3,†]

Affiliations:

[1] MOE Key Laboratory for Intelligent Networks & Network Security, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China

[2] School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China.

[3] School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, China

[4] School of Computer Science and Technology, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, Shaanxi, China.

† **Corresponding authors:** kaiye@xjtu.edu.cn

Abstract:

Here, we report a high-quality (HQ) and almost complete genome assembly with a single gap and quality value (QV) larger than 60 of the model plant *Arabidopsis thaliana* ecotype Columbia (Col-0), generated using combination of Oxford Nanopore Technology (ONT) ultra-long reads, high fidelity (HiFi) reads and Hi-C data. The total genome assembly size is 133,877,291 bp (chr1: 32,659,241 bp, chr2: 22,712,559 bp, chr3: 26,161,332 bp, chr4: 22,250,686 bp and chr5: 30,093,473 bp), and introduces 14.73 Mb (96% belong to centromere) novel sequences compared to TAIR10.1 reference genome. All five chromosomes of our HQ assembly are highly accurate with QV larger than 60, ranging from QV62 to QV68, which is significantly higher than TAIR10.1 referecne (44-51) and a recent published genome (41-43). We have completely resolved chr3 and chr5 from telomere-to-telomere. For chr2 and chr4, we have completely resolved apart from the nucleolar organizing regions, which are composed of highly long-repetitive DNA fragments. It has been reported that the length of centromere 1 is about 9 Mb and it is hard to assembly since tens of thousands of CEN180 satellite repeats. Based on the cutting-edge sequencing data, we assembled about 4Mb continuous sequence of centromere 1. We found different identity patterns across five centromeres, and all centromeres were significantly

enriched with CENH3 ChIP-seq signals, confirming the accuracy of the assembly. We obtained four clusters of CEN180 repeats, and found CENH3 presented a strong preference for a cluster 3. Moreover, we observed hypomethylation patterns in CENH3 enriched regions. This high-quality assembly genome will be a valuable reference to assist us in the understanding of global pattern of centromeric polymorphism, genetic and epigenetic in naturally inbred lines of *Arabidopsis thaliana.*

Keywords: centromere, genome assembly, bacterial artificial chromosome (BAC), telomere-to-telomere (T2T)

Introduction

The *Arabidopsis thaliana* Col-0 genome was first released in 2000 (Arabidopsis Genome Initiative, 2000). After decades of work, this reference genome has become a "gold standard" for this model plant. However, the centromeres, telomeres, and nucleolar organizing regions (NORs) have remained as gaps or misassembled due to the enrichment of highly repetitive elements (Long et al. 2013; Kawakatsu et al. 2016). Long read sequencing technologies, e.g. ONT sequencing, PacBio SMRT sequencing, is able to generate single molecular reads with length more than 10 kb, which exceeds most of the simple repeats in many genomes, driving the highly contiguous genome assemblies (Michael et al. 2018). Highly repetitive regions, however, still remain mostly unassembled due to limitations of read length and error rate of long read. ONT sequencing has overcome at least one of these challenges by producing ultra-long (longest >4 Mb) (https://nanoporetech.com/products/promethion), but these still have 5-15% per-base error rates (Istace et al. 2017), may lead to misassembly. Naish et al. (2021) used ONT ultra-long reads assembly a high contiguity *A. thaliana* Col-0 genome, but the QV of all the five chromosomes was ranging from QV41 to QV43, lower than the "gold standard" reference TAIR10.1 (QV44-QV51). In contrast, HiFi data generated from a circular consensus strategy, in which DNA is to circularized and sequenced multiple times to create high quality consensus reads (Wenger et al. 2019). This new technology is extremely promising for repeat characterization and centromeric satellite assembly, promoting the T2T assembly of human chrX and chr8 (Miga et al. 2020; Logsdon et al. 2021).

In this paper, we used ONT ultra-long reads to build contigs, and then used Hi-C data

to scaffold. After scaffolding, we used HiFi assembly contigs to replace the scaffolds based on BAC anchors. Finally, we completely resolved centromere of chr2 to chr5 and 4Mb continues centromere sequences of chr1. The genome is high accuracy with QV > 60, significantly larger than TAIR10.1 and a recent published genome. We observed various epigenetic patterns in centromere regions based on our high-quality *Arabidopsis thaliana* genome assembly.

T2T assembly using a combination of ONT ultra-long reads, HiFi reads and Hi-C data

We assembled *A. thaliana* genome from the ONT ultra-long reads using the long-read assembler NextDenovo v2.0, and we obtained 14 contigs (N50: 15.39 Mb) (Figure 1A). The Hi-C sequencing data were used to anchor 14 contigs using Juicer + 3D-DNA scaffolding pipeline. Telomeric repeat unit $(TTAGGG)_{(n=369\text{-}461)}$ was present on one end of eight contigs, and 45S rDNA units were present on one end of two contigs (Figure 1B). We obtained five scaffolds with seven gaps located at centromere region (Figure 1C). To improve the assembly, we further assembled HiFi reads by Hicanu and hifiasm v. 0.14-r312 and did gap closing based on centromeric BAC sequences as anchors (Kumekawa et al. 2000; Kumekawa et al. 2000; Hosouchi et al. 2002) (Figure 1C). We filled all the gaps on chr2, chr3, chr4 and chr5 (Figure 1C). Then, we applied BAC anchor method to replace the low accuracy ONT whole genome assembly with the high accuracy contigs assemblied by hifiasm (Figure 1D). The final genome assembly is composed of 96% HiFi assembly, and only 5,931,675 bp comes from ONT assembly. Per-base consensus accuracy was estimated using Merqury (Rhie et al. 2020) with Illumina sequencing data (101× coverage). The total

genome assembly size is 133,877,291 bp (chr1: 32,659,241 bp, chr2: 22,712,559 bp, chr3: 26,161,332 bp, chr4: 22,250,686 bp and chr5: 30,093,473 bp). The QV of all five chromosomes larger than 60 (from 62 to 68), significantly larger than than the QV (44 to 51) of TAIR10.1 reference genome and the QV (41-43) of a recent published genome (Naish et al, 2021) (Figure 1E). The dot plot between our genome and TAIR10.1 assemblies show that our genome is highly concordant with TAIR10.1 and we completely resolved most of the centromere region (Figure 1F). The assembly size of centromere 1 (CEN1) centromere 2 (CEN2), centromere 3 (CEN3), centromere 4 (CEN4) and centromere 5 (CEN5) is 3.8 Mb, 3.6Mb, 4.0 Mb, 5.5 Mb and 4.9 Mb, respectively. The assembly sizes of CEN2, CEN3, CEN4 and CEN5 were consistent with physical map-based centromeric sizes (Kumekawa et al. 2000; Kumekawa et al. 2000; Hosouchi et al. 2002). We assembled about 4Mb continues sequence of CEN1, smaller than the physical map indicated size of approximately 9 Mb (Hosouchi et al. 2002). Moreover, the recent published *A. thaliana* genome with ~5Mb CEN1 sequences, still smaller than the physical map size (Hosouchi et al. 2002), indicating it is not a complete genome (Naish et al. 2021).

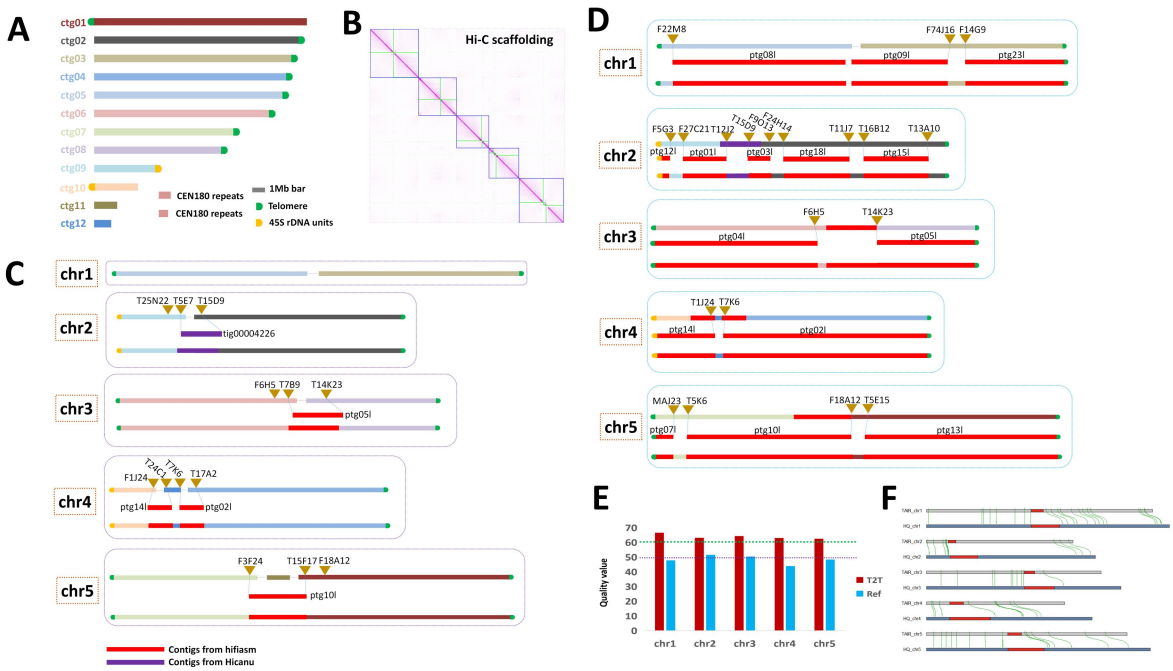

Figure 1. High-quality T2T genome assembly pipeline. **(A)** ONT ultra-long reads assembly. We obtained 14 contigs, and two of them are consisted entirely of CEN180 repeats sequences. **(B)** The Hi-C scaffolding results. The 12 contigs were ordered and oriented to five scaffolds using Hi-C data. After scaffolding, the assembly was represented in five pseudomolecules corresponding to the five chromosomes of the TAIR10.1 genome. **(C)** Gaps were filled using HiFi assembly contigs based on BAC anchors. **(D)** Whole genome assembly replaced with HiFi congtis based on BAC anchors. **(E)** QV comparison between our genome and TAIR10.1 assembly. **(F)** Linear plot between our genome and TAIR10.1 assembly. The red bars represent the centromere regions, and links represent the region corrected in our genome assembly.

Structure and epigenetic of the *Arabidopsis* centromeres

We observed the that CEN1, CEN2, CEN3 and CEN4 highly private similarity of CEN180 arrays (Figure 2A), but that CEN5 shows reduced interior similarity and was heavily disrupted by LRT/Gypsy elements (Figure 2B). The CEN180 arrays share more than 90% similarity among five centromeres apart from downstream part of CEN4. Interestingly, CEN4 consists of two distinct CEN180 arrays with the downstream ones showing low similarity (< 90%) with other 4 centromeric CEN180

arrays (Figure 2B). We identified 14,015 CEN180 monomers in CEN4, and applied a phylogenetic clustering to detect the patterns within CEN180 arrays. We obtained four CEN180 clusters, which presents four obviously specific patterns by plotting segregating sites (Figure 2C). Most of the downstream CEN180 arrays of CEN4 belong to cluster2, and the upstream CEN180 arrays were classified to three clusters labeled as cluster1, cluster3, and cluster4. Interestingly, cluster2 lost CENH3 binding signals, and we found that CENH3-containing nucleosomes exhibits a strong preference for cluster3 centromeric repeats (Figure D). The CEN3, CEN4, CEN5 assembly includes the 5S rDNA arrays. The 5S rDNA in CEN4 and CEN5 show high similarity with 95% identity. Moreover, we observed a hypomethylation pattern in CENH3 ChIP-seq signal enrichment regions (Figure 2B).

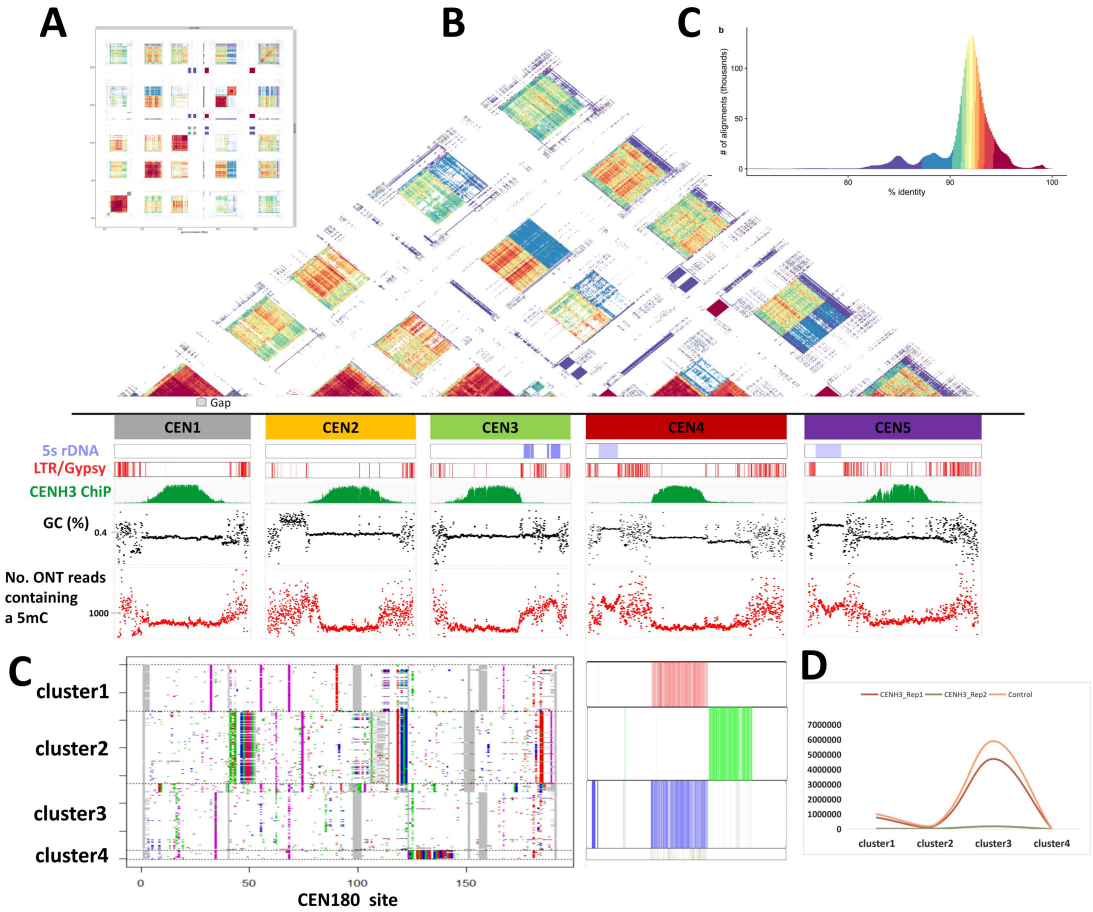

Figure 2. Structure and epigenetic map of the five chromosome centromeric regions. (A) Dot plot of the five centromeres. **(B)** Schematic showing the patterns of 5S rDNA, LTR/Gypsy, GC contents, CENH3 ChIP-seq binding map, and CpG methylation of the five centromeric regions. **(C)** An alignment plot of the14,015 CEN180 repeats of centromere 4, where each sequence is a row and each site is represented as a column. Only segregating sites with respect to the consensus sequence are colored with green, blue, purple, red and grey representing A, G, C, T and ‘ – ’ respectively. **(D)** Distribution of CENH3 ChIP signal across the four clusters.

Materials and Methods

Plant source and growth

The *A. thaliana* accession Col-0 was obtained from Shandong Agricultural University as a gift. The illumina short reads data of our Col-0, public wild type Col-0 (JGI Project ID: 1119135) and another accession AT1741 (SRA: ERR2173372) were mapped to reference genome TAIR10.1 (RefSeq: GCF_000001735.4). We used combined software including bwa v. 0.7.17-r1188, biobambam v. 2.0.87, samtools v. 1.9, varscan v. 2.4.4, bcftools v.1.9 and tabix v. 1.9 to call SNP, and the detail pipeline refer to (https://github.com/xjtu-omics/ArabidopsisAssm). The SNP calling results showed our Col-0 is the same as the public wild type.

The *A. thaliana* seeds were grown in potting soil, and they were put in the growth chamber at 22 ℃ with a 16 h light/8 h dark photoperiod with a light intensity of 100~120 μmol m-2 s-1. Young true leaves taken from 4-week-old and healthy seedlings were used in the subsequent sequencing.

Genomic DNA Sample Preparation

DNA was extracted using the Qiagen® Genomic DNA Kit following manufacturer

guidelines (Cat#13323, Qiagen). Quality and quantity of total DNA were evaluated using a NanoDrop™ One UV-Vis spectrophotometer (Thermo Fisher Scientific, USA) and Qubit® 3.0 Fluorometer (Invitrogen, USA), respectively. The Blue Pippin system (Sage Science, USA) was used to retrieve large fragments by gel cutting.

Oxford Nanopore PromethION library preparation and sequencing

For each ultra-long Nanopore library, approximately 8-10 μg of gDNA was sizeselected (>50 kb) with SageHLS HMW library system (Sage Science, USA), and processed using the Ligation sequencing 1D kit (SQK-LSK109, Oxford Nanopore Technologies, UK) according the manufacturer′s instructions. About 800ng DNA libraries were constructed and sequenced on the Promethion (Oxford Nanopore Technologies, UK) at the Genome Center of Grandomics (Wuhan, China). A total of 56.54 Gb ONT long reads with ~377× coverage were generated, and including ~177× coverage of ultra-long (>50 kbp) reads. ONT Read length N50 is 46,452 bp, and the longest reads reached 495,032 bp.

Oxford Nanopore assembly and correction

The long-read assembler NextDenovo v2.0 (https://github.com/Nextomics/NextDenovo) was used to assemble the ONT ultra-long reads data set with parameters 'read_cutoff = 5k' and 'seed_cutoff = 108,967'. Nextpolish v. 1.3.0 (Hu et al. 2020) with hifi_options '-min_read_len 10k -max_read_len 45k -max_depth 150' was used to polish the ONT contigs assembly.

HiFi library preparation and sequencing

SMRTbell target size libraries were constructed for sequencing according to PacBio′s standard protocol (Pacific Biosciences, CA, USA) using 15 kb preparation solutions.

The main steps for library preparation are: (1) gDNA shearing; (2) DNA damage repair, end repair and A-tailing; (3) ligation with hairpin adapters from the SMRTbell Express Template Prep Kit 2.0 (Pacific Biosciences); (4) nuclease treatment of SMRTbell library with SMRTbell Enzyme Cleanup Kit; (5) size selection, and binding to polymerase. A total amount of 15 μg DNA per sample was used for the DNA library preparations. The genomic DNA sample was sheared by gTUBEs (Covaris, USA) according to the expected size of the fragments for the library. Single-strand overhangs were then removed, and DNA fragments were damage repaired, end repaired and A-tailing. Then the fragments ligated with the hairpin adaptor for PacBio sequencing. And the library was treated by nuclease with SMRTbell Enzyme Cleanup Kit and purified by AMPure PB Beads. Target fragments were screened by the BluePippin (Sage Science, USA). The SMRTbell library was then purified by AMPure PB beads, and Agilent 2100 Bioanalyzer (Agilent technologies, USA) was used to detect the size of library fragments. Sequencing was performed on a PacBio Sequel II instrument with Sequencing Primer V2 and Sequel II Binding Kit 2.0 in Grandomics. A total of 22.90 Gb HiFi reads with ~153× coverage were generated, and read length N50 is 15,424 bp.

Hi-C sequencing and scaffolding

Basically, a single Hi-C library prepared from cross-linked chromatins of fungal cells using a standard Hi-C protocol (Belton and Dekker 2015) were sequenced by NovaSeq 6000 to yield Illumina PE150 reads. The Hi-C sequencing data were used to anchor all contigs using Juicer v. 1.5 (Durand et al. 2016a), followed by using a 3D-DNA scaffolding pipeline (Dudchenko et al. 2017) and manually checked and refined with Juicebox v. 1.11.08 (Durand et al. 2016b).

Centromeric satellite DNA cluster analysis

LASTZ (http://www.bx.psu.edu/~rsharris/lastz/) (Harris 2007) with parameters '--coverage=90 --format=general: score, name1, strand1, size1, start1, end1, name2, strand2, identity, length1, align1' was used to identify CEN180 (AAAAGCCTAAGTATTGTTTCCTTGTTAGAAGATACAAAGACAAAGACTCATATGGACTTCGGCTACACCATCAAAGCTTTGAGAAGCAAGAAGAAGCTTGGTTAGTGTTTTGGAGTCAAATATGACTTGATGTCATGTGTATGATTGAGTATAACAACTTAAACCGCAACCGGATCTT) (Maheshwari et al. 2017) repeats within the complete centromere. The repeats were further filtered to exclude sequences with >14 gapped positions and lengths outside the 165-to 185-bp range. Then, the CEN180 repeats were aligned using Clustal Omega v. 1.2.4 (Sievers et al. 2011) with default parameters, and the alignment was refined by "msa_refiner.py" (Maheshwari et al. 2017). Clustering was performed on this alignment using find.best() function in phyclust (Chen 2011). We evaluated a range of clusters (2-10) and using the BIC (Bayesian Information Criterion) inflection point approach to choose which K was an optimal solution.

ChIP-seq read-mapping and peak-calling

The ChIP-seq paired-end reads (SRR4430537 for Rep1 and SRR4430537 for Rep2) were mapped to the our genome using the mem algorithm of BWA (Li and Durbin 2009), and the mapping results of two repeats were merged. We performed peak calling using MACS2 (Zhang et al. 2008) with the parameters "-t merged.bam -c control.bam -f BAM --outdir ATChipseq -n ATChipseq -B --nomodel --extsize 165 --keep-dup all" and used the appropriate CENH3 ChIP negative controls (SRR4430541) as input.

Assigned ChIP-seq reads to clusters

We assigned ChIP-seq reads to clusters using a k-mer-based strategy (Maheshwari et al. 2017). We found that the specificity of cluster assignment increased with k-mer size and therefore chose 35-bp as the default k-mer length. If a read contained signature k-mers, we annotated it as a CEN180 sequence. If k-mers from a read were restricted to a single cluster, we assigned the read to that cluster.

Methylation analysis

Nanopolish v. 0.13.2 with the parameters "call-methylation --methylation cpg" was used to measure the frequency of CpG methylation from raw ONT reads. The ONT reads were aligned to whole-genome assemblies via Winnowmap v. 2.0 (Jain et al. 2020). CpG sites with log-likelihood ratios greater than 2.5 (methylated) or less than -2.5 (unmethylated) are considered high quality and included in the further analysis (Logsdon et al. 2021).

Data availability

The genome project has been deposited at GenBank under the BioProject ID PRJNA734474. The five chromosomal sequences are indexed at GenBank with accession numbers CP076438-CP076442. PacBio HiFi reads, ONT long-reads, Illumina short reads and Hi-C Illumina data are uploaded to the NCBI Sequenced Read Archive under the accession numbers SRR14728885, SRR14739179, SRR14715607, and SRR14723051, respectively.

Competing interests

The authors declare that they have no competing interests.

Literature Cited

Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering

plant *Arabidopsis thaliana*. Nature. 2000, 408(6814):796-815.

Chen WC. Overlapping codon model, phylogenetic clustering, and alternative partial expectation conditional maximization algorithm. PhD thesis, Iowa State University, Ames, IA. 2011.

Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021, 18(2):170-175.

Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016a, 3(1):95-98.

Durand NC, Robinson JT, Shamim MS, Machol I, Mesirov JP, Lander ES, Aiden EL. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. Cell Syst. 2016b, 3(1):99-101.

Harris RS. Improved pairwise alignment of genomic DNA. PhD thesis, Pennsylvania State University, State College, PA. 2007.

Hosouchi T, Kumekawa N, Tsuruoka H, Kotani H. Physical map-based sizes of the centromeric regions of *Arabidopsis thaliana* chromosomes 1, 2, and 3. DNA Res. 2002, 9(4):117-121.

Hu J, Fan J, Sun Z, Liu S. NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics. 2020, 36(7):2253-2255.

Istace B, Friedrich A, d'Agata L, Faye S, Payen E, Beluche O, Caradec C, Davidas S, Cruaud C, Liti G, Lemainque A, Engelen S, Wincker P, Schacherer J, Aury JM. de novo assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. Gigascience. 2017, 6(2):1-13.

Jain C, Rhie A, Zhang H, Chu C, Walenz BP, Koren S, Phillippy AM. Weighted minimizer sampling improves long read mapping. Bioinformatics. 2020, 36(Suppl_1):i111-i118.

Kawakatsu T, Huang SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, Castanon R, Nery JR, Barragan C, He Y, Chen H, Dubin M, Lee CR, Wang C, Bemm F, Becker C, O'Neil R, O'Malley RC, Quarless DX; 1001 Genomes Consortium, Schork NJ, Weigel D, Nordborg M, Ecker JR. Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. Cell. 2016, 166(2):492-505.

Kumekawa N, Hosouchi T, Tsuruoka H, Kotani H. The size and sequence organization of the centromeric region of *Arabidopsis thaliana* chromosome 5. DNA Res. 2000, 7(6):315-21.

Kumekawa N, Hosouchi T, Tsuruoka H, Kotani H. The size and sequence organization of the centromeric region of *Arabidopsis thaliana* chromosome 4. DNA Res. 2001, 8(6):285-90.

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

Logsdon GA, Vollger MR, Hsieh P, Mao Y, Liskovykh MA, Koren S, Nurk S, Mercuri L, Dishuck PC, Rhie A, de Lima LG, Dvorkina T, Porubsky D, Harvey WT, Mikheenko A, Bzikadze AV, Kremitzki M, Graves-Lindsay TA, Jain C, Hoekzema K, Murali SC, Munson KM, Baker C, Sorensen M, Lewis AM, Surti U, Gerton JL, Larionov V, Ventura M, Miga KH, Phillippy AM, Eichler EE. The structure, function and evolution of a complete human chromosome 8. Nature. 2021, 593(7857):101-107.

Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, Zhang Q,

Vilhjálmsson BJ, Korte A, Nizhynska V, Voronin V, Korte P, Sedman L, Mandáková T, Lysak MA, Seren Ü, Hellmann I, Nordborg M. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. Nat Genet. 2013, 45(8):884-890.

Maheshwari S, Ishii T, Brown CT, Houben A, Comai L. Centromere location in *Arabidopsis* is unaltered by extreme divergence in CENH3 protein sequence. Genome Res. 2017, 27(3):471-478.

Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker JR. High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. Nat Commun. 2018, 9(1):541.

Miga KH, Koren S, Rhie A, Vollger MR, Gershman A, Bzikadze A, Brooks S, Howe E, Porubsky D, Logsdon GA, Schneider VA, Potapova T, Wood J, Chow W, Armstrong J, Fredrickson J, Pak E, Tigyi K, Kremitzki M, Markovic C, Maduro V, Dutra A, Bouffard GG, Chang AM, Hansen NF, Wilfert AB, Thibaud-Nissen F, Schmitt AD, Belton JM, Selvaraj S, Dennis MY, Soto DC, Sahasrabudhe R, Kaya G, Quick J, Loman NJ, Holmes N, Loose M, Surti U, Risques RA, Graves Lindsay TA, Fulton R, Hall I, Paten B, Howe K, Timp W, Young A, Mullikin JC, Pevzner PA, Gerton JL, Sullivan BA, Eichler EE, Phillippy AM. Telomere-to-telomere assembly of a complete human X chromosome. Nature. 2020, 585(7823):79-84.

Naish et al. The genetic and epigenetic landscape of the *Arabidopsis* centromeres. bioRxiv. 2021.05.30.446350; doi: https://doi.org/10.1101/2021.05.30.446350.

Nurk S, Walenz BP, Rhie A, Vollger MR, Logsdon GA, Grothe R, Miga KH, Eichler EE, Phillippy AM, Koren S. HiCanu: accurate assembly of segmental

duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res. 2020, 30(9):1291-1305.

Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020, 21(1):245.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011, 7:539.

Wenger AM, Peluso P, Rowell WJ, Chang PC, Hall RJ, Concepcion GT, Ebler J, Fungtammasan A, Kolesnikov A, Olson ND, Töpfer A, Alonge M, Mahmoud M, Qian Y, Chin CS, Phillippy AM, Schatz MC, Myers G, DePristo MA, Ruan J, Marschall T, Sedlazeck FJ, Zook JM, Li H, Koren S, Carroll A, Rank DR, Hunkapiller MW. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. Nat Biotechnol. 2019, 37(10):1155-1162.

Zhang W, Lee HR, Koo DH, Jiang J. Epigenetic modification of centromeric chromatin: hypomethylation of DNA sequences in the CENH3-associated chromatin in *Arabidopsis thaliana* and maize. Plant Cell, 20(1):25-34.