bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

A chromosome-level genome assembly of a model conifer plant, the Japanese cedar, *Cryptomeria japonica* D. Don

Takeshi Fujino[1$], Yamaguchi Katsushi[2$], Toshiyuki T Yokoyama[1], Toshiya Hamanaka[1], Yoritaka Harazono[1], Hiroaki Kamada[1], Wataru Kobayashi[1], Tokuko Ujino-Ihara[3], Kentaro Uchiyama[3], Asako Matsumoto[3], Ayako Izuno[3], Yoshihiko Tsumura[4], Atsushi Toyoda[5], Shuji Shigenobu[2*], Yoshinari Moriguchi[6*], Saneyoshi Ueno[3*], Masahiro Kasahara[1*]

(1) Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa, Japan, (2) Trans-Scale Biology Center, National Institute for Basic Biology, Okazaki, Japan, (3) Department of Forest Molecular Genetics and Biotechnology, Forestry and Forest Products Research Institute, Tsukuba, Japan, (4) Faculty of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan, (5) Comparative Genomics Laboratory, National Institute of Genetics, Mishima, Japan, (6) Graduate School of Science and Technology, Niigata University, Niigata, Japan

[$] These two authors contributed equally to this work.

[*]Corresponding authors:

Masahiro Kasahara
Graduate School of Frontier Sciences, The University of Tokyo
5-1-5 Kashiwanoha, Kashiwa, Chiba 277-8561, Japan
E-mail: mkasa@k.u-tokyo.ac.jp

Saneyoshi Ueno
Department of Forest Molecular Genetics and Biotechnology, Forestry and Forest Products Research Institute, Forest Research and Management Organization,
1 Matsunosato, Tsukuba, Ibaraki 305-8687, Japan
E-mail: saueno@ffpri.affrc.go.jp

Shuji Shigenobu
Trans-Scale Biology Center, National Institute for Basic Biology,
38 Nishigonaka, Myodaiji, Okazaki, Aichi, 444-8585 Japan
E-mail: shige@nibb.ac.jp

Yoshinari Moriguchi
Graduate School of Science and Technology, Niigata University,
8050 Ikarashi 2-no-cho, Nishi-ku, Niigata 950-2181, Japan
E-mail: chimori@agr.niigata-u.ac.jp

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Abstract

Japanese cedar (*Cryptomeria japonica* D. Don) is the most important Japanese forest tree, occupying about 44% of artificial forests in Japan, and planted in East Asia, Azores Archipelago, and some islands in the Indian Ocean. Although the huge genome of the species (ca. 11 Gb) with abundant repeat elements might have been an obstacle for genetic analysis, the species is easily propagated by cutting, flowered by plant hormones like gibberellic acid, transformed by agrobacterium, and edited by CRISPR/Cas9. These characteristics of *C. japonica* are preferable to make the species a model conifer for which reference genome sequences are necessary. In this study, we report the first chromosome-level assembly for *C. japonica* (2n = 22) using a third generation selfed progeny with an estimated homozygosity of 0.96. Young leaf tissue was used to extract high-molecular-weight DNA (>50 kb) for HiFi PacBio long read sequencing and to construct Hi-C/Omni-C library for Illumina short read sequencing. Using the 29× and 26× genome coverage of HiFi and Illumina reads, respectively, de novo assembly resulted in 2,650 contigs (9.1 Gb in total) with N50 contig size of 12.0 Mb. The Hi-C analysis mapped 97% of the nucleotides on the 11 chromosomes. The assembly was verified by comparing with a consensus linkage map of 7,785 markers. The BUSCO analysis confirmed ~91% of conserved genes. Annotations of genes, repeat elements and synteny with other Cupressaceae and Pinaceae species were performed, providing fundamental resources for genomic research of conifers.

Introduction

Conifers are known to possess large genomes (6.5 to 37 Gbp) and an abundance of repetitive sequences in comparison to the genomes of animals such as humans [1–6]. As a result, the cost of determining the genome sequence of a conifer used to be quite high. In addition, creating genetic maps for these species takes a long time due to the long generation cycle of conifer, leading to a limited application of genetic methods (e.g., genetic markers, transgenesis, gene editing) in conifers; efforts at breeding via molecular biological techniques based on both genetic maps and genome sequences are limited only to a few species. Therefore, there is a pressing need for constructing a chromosome-level genome assembly of a conifer species for efficient molecular breeding, although assembling a conifer genome with abundant repeat sequences was challenging. Such a genome assembly would facilitate breeding elite trees with a variety of favorable quantitative and qualitative traits, including rapid growth rate, resistance to pathogens, and desired timber quality [7,8].

With the advent of accurate long read sequencing methods such as PacBio HiFi reads or Oxford Nanopore kit v14 has significantly decreased the cost of determining large genomes [4,9,10], leading us to believe that now is an opportune time to establish a high-accuracy, high-contiguity model conifer genome at the chromosome scale with comprehensive gene annotation. The chromosome-level assembly of a conifer genome and its associated high-density genetic map and gene annotation would facilitate the rapid identification of the relationship between various traits and genes (or mutations) on a molecular basis in a systematic manner.

Japanese cedar or sugi (*Cryptomeria japonica* D. Don), which has a large genome size (about 11 Gb estimated from DNA content (22.09pg/2C; [11]), is an allogamous and monoecious conifer, naturally

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

distributed from Aomori Prefecture to Yakushima Island in Japan [12]. *C. japonica* is a forestry species that covers the largest area of any planted trees in Japan, making it the most important species in Japan's tree industry. In particular, it was extensively planted for timber production in the 1960s to support the country's economic growth. Currently, the species occupies nearly 4.5 million hectares, accounting for 44% of all artificial forests in Japan [13]. The proliferation of artificial forests of *C. japonica* has led to a widespread prevalence of pollinosis in Japan, with a reported prevalence of 38.8% and corresponding major social issues [14]. There is therefore a pressing need to reduce the amount of *C. japonica* pollen dispersed by renewing them with male-sterile trees of *C. japonica*. Indeed, sequence-based markers for male sterility were established in this line of research [15]. These sequence-based markers allow for the easier identification of wild trees (for breeding) with a male-sterility allele; the estimated frequency of trees with male-sterile alleles in breeding materials is approximately 1% [16,17]. Had there been a chromosome-level assembly, developing genetic markers for marker-assisted selection would have been significantly facilitated. Given the industrial and social needs for the genome of *C. japonica*, we initiated PacBio HiFi sequencing of the genome of *C. japonica* with the goal of establishing a chromosome-level assembly covering over 95% of the nucleotides on the eleven chromosomes of the *C. japonica*.

Here we present a highly contiguous, chromosome-level assembly of *C. japonica* using a combination of HiFi sequencing, Hi-C scaffolding and a high-density genetic map. Our chromosome-level assembly of *C. japonica* genome is the first genome sequence in a conifer species with high sequence contiguity (N50 contig >12 Mb), with the majority of nucleotides mapped to the chromosomes (>97% assigned to the 11 chromosomes (2n=22)), and with validation through high-density genetic map (7,781 sequence-based markers; Supplementary Table 6). Genes are annotated using existing and newly sequenced RNA-seq data, ESTs in public databases, Iso-seq previously reported in literature, and homology to proteins in other species, resulting in an unprecedented high completeness score (91.4%) of BUSCO [18]. This newly sequenced genome will serve as a reliable foundation for breeding and studying other conifer species as well; for instance, it is valuable for identifying loci associated with tolerance to pathogens, faster growth, adaptation to various environments, and the evolution of conifer genomes and retrotransposons.

Results

Breeding Homozygous Trees

Assembling highly heterozygous genomes had long been known to be challenging [3,19–21], especially when we initiated our sequencing project of *C. Japonica* prior to the advent of PacBio HiFi sequencing. To reduce the heterozygous regions of the genome as much as possible, we attempted to use inbred *C. japonica* trees. However, we were unable to obtain a fully homozygous genome like inbred lines seen for rice or maize. As the self-fertilization of more than four times was not accomplished, we decided to use `Kunisaki-3 S3-3`, a third-generation inbred tree with the lowest heterozygosity among the selfed progenies, for genome sequencing. The heterozygosity rate of the selected tree was confirmed to be approximately 0.96 through a SNPType assay (Fluidigm) with 158 SNP markers (Supplementary Table 1), consistent with the results of GenomeScope analysis, which will be presented later.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Genome Sequencing

Genome DNA preparation from Young Leaves

Young leaves in April and May were sampled from `Kunisaki-3 S3-3`. To prepare high molecular weight (HMW) DNA from *C. japonica*, the sample ground into powder in liquid nitrogen was mixed with 50 ml of a pre-wash buffer [50 mM Tris-HCl (pH 8.0), 10 mM EDTA, 0.2% 2-ME, 0.3% TritonX] and incubated at room temperature for 1 hour with gentle agitation. This pre-wash step serves to reduce the viscosity of the extracted DNA solution. Following centrifugation at 9,000 × *g* for 20 minutes at 4°C, the pellet sample was snap-frozen. Fifty ml of frozen powdered QIAGEN G2 buffer (Qiagen, Netherlands, Cat #1014636), which was produced by spraying the buffer into liquid nitrogen in a glass beaker, was added to the sample and mixed quickly. Upon allowing the mixture to thaw in a tube, 30 ul of RNaseA (QIAGEN, Cat #19101) and 2.5 ml of Proteinase K (QIAGEN, Cat #1019499) were added, and the sample was incubated at room temperature for 2 hours without agitation. The sample was centrifuged at 9,000 × *g* for 20 minutes at 4°C, and the supernatant was subjected to DNA extraction using two QIAGEN Genomic-tip 100/G columns (Cat # 10243). The genomic DNA was eluted with 5 ml of Buffer QF, precipitated with isopropyl alcohol, and dissolved with 600 ul of TE at 4°C for a week, which yielded 247 µg of DNA (412 ng/µl). We repeated this procedure to obtain another batch of HMW genome DNA from the young leaves of the same tree, yielding 356 µg of DNA (594 ng/µl). The size distribution of the HMW DNA was evaluated by pulsed-field gel electrophoresis (PFGE). Briefly, 20 ng of HMW DNA was run on a 1% agarose gel (Seakem Gold Agarose, Lonza, Rockland, ME, USA, Cat #50150) in 0.5x TBE with the BioRad CHEF Mapper system (BioRad, Hercules, CA, USA, Cat #M1703650) for 15 hours, and the gel was stained with SYBR Gold dye (Thermo Fisher Scientific, Waltham, MA, USA, Cat #11494). Lambda ladder, *Saccharomyces cerevisiae* genome, and 5 kbp ladder (BioRad,Cat #170–3624) were used as standards. These results demonstrated that the HMW DNA had an approximate mean size of 50 kb. DNA concentration was measured using a Qubit fluorometer (Thermo Fisher Scientific, Cat #Q32866) and the Qubit™ dsDNA BR Assay Kit (Thermo Fisher Scientific, Cat #Q32850). The quality of the extracted DNA was evaluated by analyzing the absorbance spectra with Nanodrop ND-2000C (Thermo Fisher Scientific).

NGS Library preparation and sequencing

PCR-free Illumina Read Sequencing

For PCR-free Illumina library preparation, we used 6 ug of the genomic DNA (gDNA) extracted using the QIAGEN Genomic-tip 100/G columns following the method described in the previous section. The gDNA was fragmented into 200-800 bp fragments with a peak at 550 bp using the Covaris Focused-ultrasonicator S2 system with the following settings: intensity=4; duty cycle=10%; cycles per burst=200; time=55 sec; temperature=7°C. Subsequently, ~400-bp fragments were size-selected by using BluePippin and 2% EF gel in the tight mode setting, followed by concentration with AMpure XP (1.8x vol.), yielding 226 ug of DNA with a sharp peak at 410 bp as evaluated with Agilent 2100 Bioanalyzer. A library for whole genome sequencing was prepared using the TruSeq DNA PCR-Free Library Prep Kit (Illumina) according to the manufacturer's instructions. The concentration of the library was quantified by KAPA Library Quantification Kit and Applied Biosystems 7500 Real-Time PCR Systems (Applied Biosystems). The Illumina library was sequenced

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

using the HiSeq X platform (Illumina) at Macrogen Japan (Tokyo, Japan) with the 2x151 bp paired-end sequencing protocol. The total number of the raw Illumina paired-end reads was 734439281.

HiFi Long Read Sequencing

HMW genome DNA obtained through the procedures described above was shared to an average target size of 20-30 kbp via centrifugation in a Covaris g-Tube (cat# 520079) at 2800 - 4800 rpm for 2 min using a microcentrifuge (himac CG15RX). SMRTbell libraries for sequencing on the PacBio Sequel platform were constructed according to the manufacturer's recommended protocol (Preparing HiFi SMRTbell Libraries using SMRTbell Express Template Prep Kit2.0 PN101-853-100 Version 03 (January 2020) or Version 04 (April 2021)) with minor modifications. We altered the condition of genome shearing with the g-Tube and the size selection of the library with the BluePippin system (Sage Science, .MA,USA.) with the aim of constructing libraries with longer inserts. Specifically, we constructed seven SMRTbell libraries with various settings of centrifugation speed, including 4800 rpm, 3800 rpm, 4300 rpm, 4800 rpm, 3300 rpm, 2800 rpm, and 2800 rpm, yielding libraries with average sizes of 20 kb, 20 kb, 18 kb, 16 kb, 22 kb, 25 kb, and 25 kb, respectively. These insert size distributions were determined by PFGE with the BioRad CHEF Mapper system as previously performed for HMW genome evaluation (see above). These SMRTbell libraries were sequenced in 1 SMRT cell (movie time of 20 hours) on a Sequel instrument at the NIBB Functional Genomics Facility and 10 cells (movie time of 30 hours) on a Sequel II instrument at the National Institute of Genetics, ROIS, Japan (Supplementary Table 3).

Hi-C Sequencing

The Omni-C library was prepared using the Dovetail Omni-C Kit (Dovetail cat# 21005) according to the manufacturer's protocol (Omni-C Protocol Non-mammal v1.0 for Plants) with modifications. Due to our inability to obtain chromatin through the standard in-place fixation procedure from leaves, we employed a third-party protocol [High Molecular Weight DNA Extraction from Recalcitrant Plant Species for Third Generation Sequencing by Workman; dx.doi.org/10.17504/protocols.io.4vbgw2n], to isolate nuclei from leaf samples, after which we proceeded to nuclease digestion. The ends of chromatins were polished and ligated to a biotinylated bridge adapter prior to proximity ligation of adapter-containing ends. Following proximity ligation, crosslinks were reversed, and DNA was purified from proteins and size-selected. Purified DNA was subjected to end-repair, ligated to an adapter, and treated with USER Enzyme. Four sequencing libraries were generated using Illumina-compatible adapters, and biotin-containing fragments were isolated using streptavidin beads before PCR enrichment of the library. These libraries were sequenced on two lanes of the Illumina HiSeq X platform, yielding 470 million 2 × 151 bp read pairs.

Transcriptome Sequencing

In order to perform gene annotation for *C. japonica*, we carried out RNA-Seq for several tissues: root, corn, female flower bud, and seeds in corn (Supplementary Table 2). Each tissue sample (50 to 100 mg) was disrupted under liquid nitrogen, and the tissue powder was emulsified by 600 µL of CTAB extraction buffer (Supplementary Table 2) and incubated for 10 min. at 65 °C. After centrifugation for 5 min. at 15,000 rpm at 4 °C, the supernatant (400 µL) was mixed with 200 µL of Lysis Buffer of the Maxwell® RSC Plant RNA Kit (Promega). RNA was then extracted by the Maxwell® instrument (Promega) according to the manufacturer's instructions. The extracted RNA was sent to a sequencing company (Macrogen) to construct sequencing libraries and sequenced on NOVASeq6000 (Illumina)

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

in 2 x 101 bp paired-end form. Publicized RNA-Seq data (Supplementary Table 2) [22,23] were also used in the current study.

Genome Assembly

Assembling PacBio HiFi reads

The collective statistics of the obtained PacBio HiFi reads are displayed in Table 1. Through the use of FastQC [24], we discovered that the first 12+ bases occasionally contained adapter sequences that were not filtered by the official PacBio pipeline (Supplementary Figure 1). After cutting the first 20 base pairs, the base frequency bias due to the adapters disappeared (Supplementary Figure 2). To see how the remaining adapters affect the assembly statistics, we generated three sets of the HiFi reads by trimming the first 0, 20, 30 base pairs using fastp ver 0.20.0 [25]. Then we assembled the three sets of the PacBio HiFi reads using hifiasm 0.15.5-r350 [26] and Flye 2.8.3 [27]. However, the Flye assembly jobs did not complete due to out of memory (> 300 GB), and we estimated that even with infinite memory the job would take more than 62 days, which is the maximum computation time for our shared computing environments. Hifiasm completed in two days, so we used the result of the Hifiasm (Supplementary Table 2). We selected the assembly with the trim size = 20 bp because it has the longest N50 contig size and the other metrics are largely similar to the other assemblies (Supplementary Table 4). The final contigs had a N50 contig size of 12 Mb, which is significantly larger than those for previously published conifer genomes such as [4]. The total contig size is 9.07 Gbp, which is close to the estimated genome size. The statistics for the final contigs are presented in Table 2 below.

Table 1: Sequencing Statistics.

Sequence Type	# Reads	Total # bp	Genome Fold	Average Len (bp)
PacBio HiFi	18,877,201	316,631,828,055	28.8	16,960

Table 2: Assembly Statistics.

Assembler	N50 Contig Size (bp)	L50	Total Count	Total Length (bp)	Max Contig Size (bp)
HiFiasm	12,036,660	217	2,651	9,068,249,119	88,939,144

Hi-C Scaffolding

We utilized Juicer v1.6 [28] and BWA v0.7.17-r1188 [29,30] to align Hi-C reads to the contigs, subsequently utilizing the 3D-DNA pipeline v180922 [28] to correct misjoins, order, and orient the contigs into scaffolds. We subsequently employed Juicebox Assembly Tools v2.13.07 for manual review of the scaffolds, making a few necessary corrections for obvious misassemblies. The 3D-DNA pipeline was subsequently utilized to rescaffold the final chromosome-length scaffolds, with all tools being executed using default parameters. The largest 11 scaffolds, which correspond to the 11 chromosomes of *C. japonica*, contained 97.6% of the nucleotides (Figure 1; Table 3; Supplementary Figure 3).

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

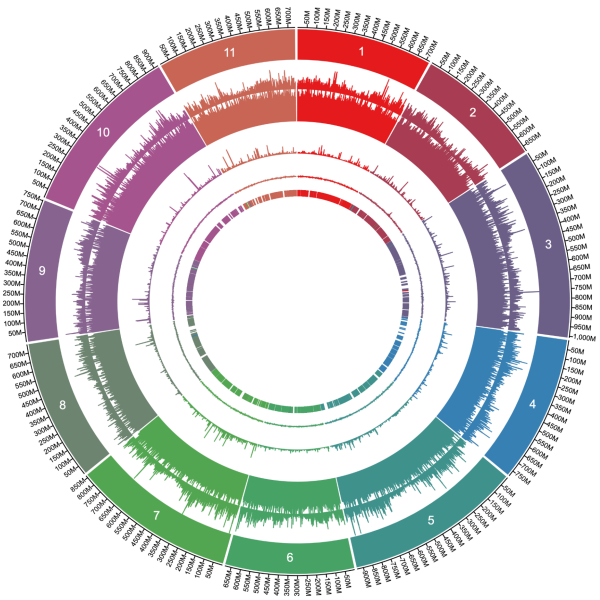


Figure 1: Overview of the 11 chromosomes of the *C. japonica* genome. Rings represent (from outside) chromosomes, gene density, repeat sequence %, N% (gaps between contigs), GC%, and genetic markers (described later). The genome is highly repetitive across the genome.

Post-processing of the Hi-C scaffolds

Utilizing NCBI-BLAST [31] with the parameters specified in VecScreen [32], we screened for contaminants such as PacBio adapters within the scaffolds and subsequently masked them as 'N's. Any scaffold whose entire sequence was masked was subsequently removed.

To identify scaffolds corresponding to the chloroplast, we employed minimap2 [33] with default parameters to align an existing reference chloroplast sequence (acc: AP009377) against the scaffolds. Subsequent manual review removed scaffolds that largely aligned with the entire reference sequence. A single representative chloroplast sequence was then selected from the genome, reverse complemented and rotated to better align with the existing reference sequence, the annotation of which was transferred to the new sequence using GeSeq [34]. While we attempted to locate mitochondrial sequences as well, we were unable to do so reliably due to the nature of complex mitochondrial genome structure of conifers [35] and thus, potential mitochondrial sequences remain as ordinary scaffolds. The statistics of the final scaffolds are presented below.

Table 3: Assembly Statistics after Hi-C Scaffolding.

Top 11 Scaffolds' shows the percentage of the nucleotides on the 11 chromosomes of the genome.

N50 Scaffold Size (bp)	# Scaffolds	Total Length (bp)	Max Scaffold (bp)	Top 11 Scaffolds (%)
754,841,075	2,699	9,046,663,964	1,007,622,168	97.6

Genome Assembly Validation by Genetic Map

Linkage maps were constructed from four existing mapping families (F1O7, S1-2, S5HK7, and S8HK5) [36] using the LPmerge program [37] (Supplementary Table 6). The sequence-based markers

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

in the merged genetic map are aligned to the scaffolds using minimap2 [33] with the options -c -k 9. The majority of the markers were consistent with the scaffolds (Supplementary Figure 4). As the genetic map was developed independently of the genome assembly, the high level of consistency between the Hi-C scaffolds and the genetic map attests to the high accuracy of the Hi-C scaffolds in terms of minimizing global misjoins. Markers that mapped multiple times may not be consistent and minor inconsistencies resulting from ambiguities during the merging of the four independent genetic maps were also noted (Supplementary Figure 4). The mapped markers are well scattered so any single locus trait would be easily mapped (Figure 1).

It is worth noting that there are long regions where recombination is suppressed in the middle of the chromosomes. We presume that these regions correspond to centromeres. The gene density decreases towards these putative centromeres (Figure 1).

Identifying Repeat Elements

We ran RepeatModeler 2.0.3 [38] to identify repeat elements in the contigs, yielding a repeat library (Table 4) that was subsequently input to RepeatMasker 4.1.2-p1 [2] for identifying repeat elements in the genome. We used the most sensitive preset, -s option. We used Cross_match [39] for the search engine of RepeatMasker because it is the most sensitive engine. As is common in other conifer species, the *C. japonica* genome is abundant in repeat elements, with 83.6% of the genome identified as such (Table 5). These repeat elements are evenly scattered across the genome, showing no observable bias on the chromosome location (Figure 1). The majority of 'Unknown' and 'LTR/Unknown' (included in 'Other LTR' in Table 5) presumably consists of fragmented copies of *Gypsy* or Copia superfamilies not precisely identified by RepeatModeler2.

Table 4: Statistics of the Output of RepeatModeler2

# sequences	Total Length	Min. Length	Avg. Length	Max. Length	N50
4,089	10,847,836	34	2,653	39,182	5,201

Table 5: Breakdown of Repeat Elements

Unknown and LTR/Unknown are presumably fragmented LTR/Gypsy and LTR/Copia which RepeatModeler failed to identify the element category.

class	# bp	% to the contigs
LTR/Gypsy	2,706,211,341	29.8%
LTR/Copia	1,787,198,891	19.7%
LTR/Pao	914,583	0.0%
LTR/ERVK	492,659	0.0%
LTR/Caulimovirus	278,675	0.0%
LTR/ERV1	97,436	0.0%
Other LTR	726,514,838	8.0%
Total LTR	5,221,708,423	57.6%
LINE/L1	266,585,601	2.9%

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

LINE/L1-Tx1	40,379,039	0.4%
LINE/I	14,549,997	0.2%
LINE/RTE-X	8,624,746	0.1%
LINE/Penelope	2,966,778	0.0%
Total LINE	333,106,161	3.7%
DNA/MULE-MuDR	94,429,850	1.0%
DNA/Sola-2	56,686,823	0.6%
DNA/hAT-Tip100	43,403,967	0.5%
DNA/CMC-EnSpm	32,307,893	0.4%
DNA/hAT-Tag1	27,462,207	0.3%
DNA/TcMar-Fot1	3,788,044	0.0%
DNA/hAT-Ac	2,762,813	0.0%
DNA/Maverick	660,006	0.0%
DNA/Dada	658,721	0.0%
Other DNA	713,765	0.0%
Total DNA	262,874,089	2.9%
RC/Helitron	14,193,395	0.2%
Unknown	1,666,554,795	18.4%
Total interspersed	7,498,436,863	82.7%
Simple repeat	46,895,194	0.5%
rRNA	26,423,944	0.3%
Low complexity	7,191,750	0.1%
Satellite	3,133,394	0.0%
tRNA	562,090	0.0%
Total	7,582,643,235	83.6%

Gene and Functional Annotation

We used BRAKER 2.1.6 [40] to predict two sets of genes, one based on RNA-seq data and the other based on homology with protein databases. For the former prediction set, the RNA-seq alignments were used for training the gene model, while the homology-based prediction was conducted using the proteins at the Viridiplantae level of OrthoDB v10 [41–43]. We assembled the RNA-seq data *de novo* using Trinity 2.13.2 [44]. In addition, we performed reference-based transcriptome assembly using StringTie 2.1.7 [45] with the RNA-seq data, ESTs in GenBank/EMBL/DDBJ [22] and Iso-seq reads previously reported in [23]. Subsequently, the *de novo* transcriptome assembly, the ESTs, the Iso-seq reads and the gene structures produced by StringTie were input into the alignment assembly pipeline of PASA 2.5.2 [46,47] to generate an integrated gene structure on the contigs. Finally, EVidenceModeler 1.1.1 [47] was employed to integrate all the gene sets from the BRAKER pipeline

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

and the PASA alignment assembly pipeline. EVidenceModeler output a total of 152,527 genes, which we refer to as *a permissive gene set*, as it contains many false positive genes.

As the permissive gene set contains many false positives, we created another gene set, referred to as *the standard gene set*, for general analysis and databases. To do so, we first removed genes from the permissive gene set that overlap 95% of their length with repeat elements, as these genes are typically proteins almost entirely within transposable elements. We then eliminated genes that lack both at least 5% overlapping length with the PASA transcripts and homology with the proteins in at least one of SwissProt [48], NCBI RefSeq [49], and Viridiplantae of OrthoDB v10 [42] (defined by any statistically significant hits using the default parameter of DIAMOND [50]) are eliminated. In other words, we eliminated predicted genes that lack both expression evidence and homology to other species. In all, the remaining gene set, which we call the standard gene set, contained 55,246 genes. We confirmed that throughout these filtering the BUSCO scores do not change significantly (data not shown). Supplementary Table 7 shows for each gene if it is in the standard gene set, if it has an overlap with the PASA transcripts (RNA-seq evidence), if it has homology to proteins in other species (homology evidence), if it is almost entirely contained in the repeat elements, and if it contains a part of transposable elements (identified by TEsorter ver 1.3 [51] with an options, -db rexdb-plant -dp2 and with Viridiplantae v3.0 of REXdb).

EnTAP 0.10.8 [52] was used for functional annotation with default parameters and following databases: UniProtKB release 2022_05 [48] and NCBI RefSeq plant proteins release 215 [49]. The descriptions of the top protein hit were transferred to the gene descriptions when EnTAP found it informative. When the hit was not informative, we transferred the EggNOG [53,54] description when available.

Evaluation of Gene Annotation

The gene prediction and annotation for the genome was first done on the contigs and then were lifted over to the chromosome-level assembly. We here evaluated the quality and completeness of our gene annotation using BUSCO v5.3.0 [18,55,56]. As expected from the high contiguity and high accuracy of the HiFi assembly, the BUSCO score for the standard gene set showed very high (> 90%) completeness for the conserved gene set (embryophyta_odb10; see Table 6). Approximately 5% of the duplicated genes (Complete duplicated) may account for the heterozygosity between the two haplotypes (4%) that remained even through three rounds of selfing (see the *Breeding Homozygous Trees* section).

To our knowledge, the complete gene rate (91.4%) of the *C. japonica* genome is the highest among conifer species reported to have their genomes sequenced, demonstrating the potential of *C. japonica* as a model conifer species.

Table 6: BUSCO Assessment of the Assembly

Complete (%)	(Complete single) (%)	(Complete duplicated) (%)	Fragmented (%)	Missing (%)	Total (n)
91.4	86.1	5.3	3.2	5.4	1614

Evolution of Repeat Elements

Assembling genomes with abundant repeat elements used to yield contigs with less contiguity (i.e., orders of magnitude smaller N50 contig size) when PacBio HiFi reads were not used for genome

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

assembly. Such genome assemblies tended to selectively miss repeat elements, likely having led to inaccurate analysis of repeat elements. The N50 contig size of our *C. japonica* genome assembly is over 10 Mb; the assembly should contain more repeat elements then ever. This gives us an opportunity to analyze more accurately how repeat elements evolved.

With this in mind, we analyzed the ages of repeat elements in the *C. japonica* genome. We largely followed Ma et al [57]. Briefly, (1) we identified the locations of LTR elements by running EDTA [58], (2) aligned the two LTR copies for each identified LTR retrotransposon; the two copies have exactly the same sequence when it was inserted, and therefore the discrepancies between the two LTR copies can serve as a molecular clock, (3) using the mutation rate of 1.38×10^{-8}/site/year previously reported for *C. japonica* [59], the number of mutations between the two LTR copies is converted to the age of the inserted LTR, and (4) we created the histogram of the age distribution of the LTR retrotransposons for each of *Gypsy* and *Copia* superfamilies (Figure 2). To our knowledge, this is the first description of the age distribution of LTR retrotransposons for conifer genomes with N50 contig size of over a megabase.

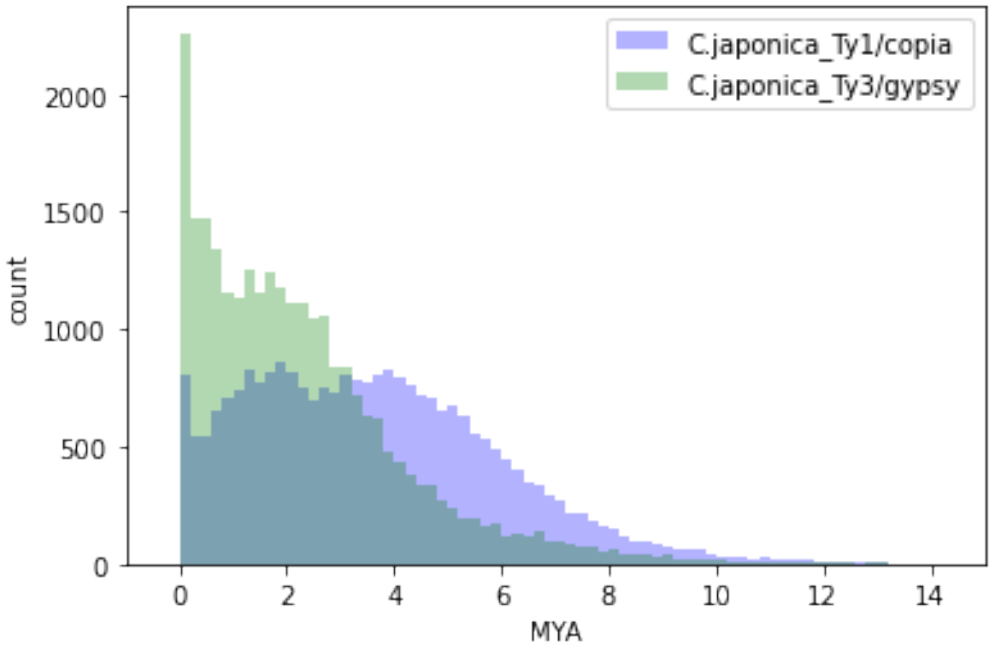


Figure 2: Age distribution of Copia and Gypsy in the *C. japonica* genome. The x-axis is million years old and the y-axis is the number of their occurrences in the genome.

Genome Browser at ForestGEN

The genome sequence of *C. japonica* can be explored from ForestGEN (FORest EST and GENome database), where JBrowse [60] displays the genomic sequences and gene annotations (https://forestgen.ffpri.go.jp/en/index.html). ForestGEN also has database functions which enable users to query the functional annotation of the predicted genes.

Discussions

We assembled the *C. japonica* genome sequence with a very high degree of continuity (N50 contig: 12.0 Mb) and were able to map the majority (> 97%) of the nucleotide sequences to all of the 11

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

chromosomes using the Hi-C scaffolding and the genetic map. We were also able to validate the correctness of the chromosome-level assembly by comparing it with the high-density genetic mapping. This allowed us to establish a genetic information foundation that can be used to develop selecting markers for *C. japonica* breeding in a shorter period of time using a systematic method. In addition, we annotated about 50,000 genes in the *C. japonica* genome sequence and achieved the highest BUSCO score (91.4% complete) among conifers with chromosome-level assembly, to our knowledge. These genome resources thereby reinforce the position of *C. japonica* as a model conifer tree.

C. japonica can be used as a model conifer for several reasons. The species is easily propagated by cutting [[61,62], flowered by plant hormones like gibberellic acid [63], transformed by agrobacterium [64], and can be edited by CRISPR/Cas9 [65]. This study aims at providing the high-quality, chromosome-level genome sequence of the *C. japonica*.

One of such breeding efforts is the development of male-steriile trees of *C. japonica* for ameliorating the prevalence of pollinosis in Japan. Since the discovery of male-sterile tree in 1992 [66], breeding of male-sterile trees was initiated including screening of new male-sterile trees in the field, because such variety is expected to reduce airborne pollen dramatically. Four Mendelian inherited male sterility loci (*MS1–MS4*) are known [36], but their exact sequences were not known and the candidate genomic regions for them were too huge to sequence by using tiling clones such as Bacterial Artificial Chromosomes due to the huge genome size.

Sequencing the whole genome of *C. japonica* and narrowing the candidate genomic regions of each of the *MS1–MS4* loci using genetic markers was considered to be the fastest way to obtain the sequence-based markers of *MS1* to *MS4*. Indeed, our chromosome-level assembly of the *C. japonica* genome, which was enabled by the power of long read DNA sequencers and the Hi-C scaffolding, contributed to successfully identify the candidate genes of *MS1* and *MS4* [15,67], having demonstrated the usefulness of our chromosome-level genome sequences.

Data Availability

The raw PCR-free Illumina reads were deposited in the Sequence Read Archive database of the DNA Data Bank of Japan (DDBJ) under accession number DRX412221, the Omni-C Illumina reads under DRX412222, the PacBio HiFi reads under DRX409629, the assembly and gene annotation under BSEH01000001-BSEH01002698. The assembly, gene annotation, genetic map, and BLAST search are available in the ForestGEN database (https://forestgen.ffpri.go.jp/en/index.html).

Abbreviations

BAC Bacterial Artificial Chromosomes
DDBJ DNA Data Bank of Japan
FFPRI Forestry and Forest Products Research Institute
HMW High Molecular Weight
PCR Polymerase Chain Reaction
ROIS Research Organization of Information and Systems

Competing Interests

The authors declare no competing interests.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Acknowledgements

This work has been supported in part by FFPRI Grant (#201406, #201421, #201906), JSPS KAKENHI Grant Number JP16H06279 (PAGS), and JP20H03239, Bio-oriented Technology Research Advancement Institution (BRAIN) Grant (JPJ007097; Project ID 28013B), and NIBB Collaborative Research Program (15-829, 16-403, 17-405, 18-408, 19-420, 20-428, 21-302, and 22NIBB402). The super-computing resource was provided in part by Human Genome Center (the University of Tokyo), in part by NIG supercomputer at ROIS National Institute of Genetics, in part by the SuperComputer System, Institute for Chemical Research, Kyoto University, in part by the Data Integration and Analysis Facility, National Institute for Basic Biology, and in part by the supercomputer of AFFRIT, MAFF, Japan. The Arboretum and Nursery Office, FFPRI was involved in the maintenance of the *C. japonica* nursery. Ms. Furusawa, FFPRI, performed RNA experiments. Dr. Murai (a retired researcher at FFPRI) conducted artificial crossings, including initiation of inbred lines of *C. japonica*. The sequenced *C. japonica* (`Kunisaki-3 S3-3`) specimens were housed in the xylarium (TWTw) of FFPRI with voucher (TWTw-29147). We appreciate Miwako Matsumoto, Hisayo Asao, and Dr. Asaka Akita (NIBB) for next generation sequencing.

References

1. Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin Y-C, Scofield DG, et al.. The Norway spruce genome sequence and conifer genome evolution. *Nature*. 497:579–842013;

2. Wang X-Q, Ran J-H. Evolution and biogeography of gymnosperms. *Mol Phylogenet Evol*. 75:24–402014;

3. Mackay J, Dean JFD, Plomion C, Peterson DG, Cánovas FM, Pavy N, et al.. Towards decoding the conifer giga-genome. *Plant Mol Biol*. 80:555–692012;

4. Niu S, Li J, Bo W, Yang W, Zuccolo A, Giacomello S, et al.. The Chinese pine genome and methylome unveil key features of conifer evolution. *Cell*. 185:204–17.e142022;

5. Lorenz WW, Ayyampalayam S, Bordeaux JM, Howe GT, Jermstad KD, Neale DB, et al.. Conifer DBMagic: a database housing multiple de novo transcriptome assemblies for 12 diverse conifer species. *Tree Genet Genomes*. 8:1477–852012;

6. Raj Ahuja M, Neale DB. Evolution of Genome Size in Conifers. *Silvae Genet*. 54:126–372005;

7. Neale DB, Savolainen O. Association genetics of complex traits in conifers. *Trends Plant Sci*. 9:325–302004;

8. Beaulieu J, Doerksen T, Boyle B, Clément S, Deslauriers M, Beauseigle S, et al.. Association genetics of wood physical traits in the conifer white spruce and relationships with gene expression. *Genetics*. 188:197–2142011;

9. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al.. The complete sequence of a human genome. *Science*. 376:44–532022;

10. Schloissnig S, Kawaguchi A, Nowoshilow S, Falcon F, Otsuki L, Tardivo P, et al.. The giant axolotl genome uncovers the evolution, scaling, and transcriptional control of complex gene loci. *Proc Natl Acad Sci U S A*. 2021; doi: 10.1073/pnas.2017176118.

11. Hizume M, Kondo T, Shibata F, Ishizuka R. Flow Cytometric Determination of Genome Size in

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

the Taxodiaceae, Cupressaceae sensu stricto and Sciadopityaceae. *Cytologia* . 66:307–112001;

12. Tsumura Y. Cryptomeria. In: Kole C, editor. *Wild Crop Relatives: Genomic and Breeding Resources*. Springer Berlin Heidelberg; p. 49–64.

13. Forestry Agency, Japan. State of Japan's Forests and Forest Management. Forestry Agency, Japan; 2019 Jul.

14. Matsubara A, Sakashita M, Gotoh M, Kawashima K, Matsuoka T, Kondo S, et al.. Epidemiological Survey of Allergic Rhinitis in Japan 2019. *Nippon Jibiinkoka Gakkai Kaiho*. 123:485–902020;

15. Hasegawa Y, Ueno S, Wei F-J, Matsumoto A, Uchiyama K, Ujino-Ihara T, et al.. Identification and genetic diversity analysis of a male-sterile gene (MS1) in Japanese cedar (Cryptomeria japonica D. Don). *Sci Rep*. 11:14962021;

16. Moriguchi Y, Ueno S, Hasegawa Y, Tadama T, Watanabe M, Saito R, et al.. Marker-Assisted Selection of Trees with MALE STERILITY 1 in Cryptomeria japonica D. Don. *For Trees Livelihoods*. Multidisciplinary Digital Publishing Institute; 11:7342020;

17. Watanabe M, Ueno S, Hasegawa Y, Moriguchi Y. Efficient low-cost marker-assisted selection of trees with MALE STERILITY 1 (MS1) in Japanese cedar (Cryptomeria japonica D. Don) using bulk DNA samples. *Tree Genet Genomes*. 18:292022;

18. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. *Mol Biol Evol*. 38:4647–542021;

19. Pryszcz LP, Gabaldón T. Redundans: an assembly pipeline for highly heterozygous genomes. *Nucleic Acids Res*. 44:e1132016;

20. Claros MG, Bautista R, Guerrero-Fernández D, Benzerki H, Seoane P, Fernández-Pozo N. Why assembling plant genome sequences is so challenging. *Biology* . 1:439–592012;

21. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, et al.. The zebrafish reference genome sequence and its relationship to the human genome. *Nature*. 496:498–5032013;

22. Futamura N, Totoki Y, Toyoda A, Igasaki T, Nanjo T, Seki M, et al.. Characterization of expressed sequence tags from a full-length enriched cDNA library of Cryptomeria japonica male strobili. *BMC Genomics*. 9:3832008;

23. Wei F-J, Ueno S, Ujino-Ihara T, Saito M, Tsumura Y, Higuchi Y, et al.. Construction of a reference transcriptome for the analysis of male sterility in sugi (Cryptomeria japonica D. Don) focusing on MALE STERILITY 1 (MS1). *PLoS One*. 16:e02471802021;

24. Andrews S. FastQC: A Quality Control Tool for High Throughput Sequence Data.

25. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 34:i884–902018;

26. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*. 18:170–52021;

27. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol*. 37:540–62019;

28. Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, et al.. De novo assembly

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science*. American Association for the Advancement of Science; 356:92–52017;

29. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 25:1754–602009;

30. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv.org*. q-bio.GN2013;

31. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.. BLAST+: Architecture and applications. *BMC Bioinformatics*. 2009; doi: 10.1186/1471-2105-10-421.

32. NCBI. NCBI VecScreen (https://www.ncbi.nlm.nih.gov/tools/vecscreen/).

33. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. *Bioinformatics*. 34:3094–1002018;

34. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, et al.. GeSeq - versatile and accurate annotation of organelle genomes. *Nucleic Acids Res*. 45:W6–112017;

35. Robicheau BM, Young AP, LaButti K, Grigoriev IV, Walker AK. The complete mitochondrial genome of the conifer needle endophyte, Phialocephala scopiformis DAOMC 229536 confirms evolutionary division within the fungal Phialocephala fortinii s.l. - Acephala appalanata species complex. *Fungal Biol*. 121:212–212017;

36. Hasegawa Y, Ueno S, Matsumoto A, Ujino-Ihara T, Uchiyama K, Totsuka S, et al.. Fine mapping of the male-sterile genes (MS1, MS2, MS3, and MS4) and development of SNP markers for marker-assisted selection in Japanese cedar (Cryptomeria japonica D. Don). *PLoS One*. 13:e02066952018;

37. Endelman JB, Plomion C. LPmerge: an R package for merging genetic maps by linear programming. *Bioinformatics*. 30:1623–42014;

38. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al.. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A*. 117:9451–72020;

39. Green P: cross_match. http://www.phrap.org/

40. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform*. 3:lqaa1082021;

41. Kriventseva EV, Tegenfeldt F, Petty TJ, Waterhouse RM, Simão FA, Pozdnyakov IA, et al.. OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Res*. 43:D250–62015;

42. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, et al.. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res*. 47:D807–112019;

43. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, et al.. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res*. 45:D744–92017;

44. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al.. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*. 29:644–522011;

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

45. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biol*. 20:2782019;

46. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, et al.. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res*. 31:5654–662003;

47. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al.. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol*. 9:R72008;

48. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res*. 2022; doi: 10.1093/nar/gkac1052.

49. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al.. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 44:D733–452016;

50. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods*. 18:366–82021;

51. Zhang R-G, Li G-Y, Wang X-L, Dainat J, Wang Z-X, Ou S, et al.. TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes. *Hortic Res*. 2022; doi: 10.1093/hr/uhac017.

52. Hart AJ, Ginzburg S, Xu MS, Fisher CR, Rahmatpour N, Mitton JB, et al.. EnTAP: Bringing faster and smarter functional annotation to non-model eukaryotic transcriptomes. *Mol Ecol Resour*. 20:591–6042020;

53. Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol*. 38:5825–92021;

54. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al.. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res*. 47:D309–142019;

55. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. Oxford University Press; 31:3210–22015;

56. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al.. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol Biol Evol*. 35:543–82018;

57. Natali L, Cossu RM, Mascagni F, Giordani T, Cavallini A. A survey of Gypsy and Copia LTR-retrotransposon superfamilies and lineages and their distinct dynamics in the Populus trichocarpa (L.) genome. *Tree Genet Genomes*. 11:1072015;

58. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al.. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol*. 20:2752019;

59. Kusumi J, Tsumura Y, Tachida H. Evolutionary rate variation in two conifer species, *Taxodium distichum* (L.) Rich. var. *distichum* (baldcypress) and *Cryptomeria japonica* (Thunb. ex L.f.) D. Don (Sugi, Japanese cedar). *Genes Genet Syst*. 90:305–152015;

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

60. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al.. JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biol*. 17:662016;

61. Toda R. VEGETATIVE PROPAGATION IN RELATION TO JAPANESE FOREST TREE IMPROVEMENT. *New Zealand journal of forestry science*. 4:410–71973;

62. Dirr MA, Heuser CW Jr. The Reference Manual of Woody Plant Propagation: From Seed to Tissue Culture, Second Edition. 2nd ed. Timber Press;

63. Kato Y, Miyake I, Ishikawa H. Stimulation of gibberellin on differentation of "Sugi." *JOURNAL OF THE JAPANESE FORESTRY SOCIETY*. 40:35–61958;

64. Taniguchi T, Ohmiya Y, Kurita M, Tsubomura M, Kondo T. Regeneration of transgenic Cryptomeria japonica D. Don after Agrobacterium tumefaciens-mediated transformation of embryogenic tissue. *Plant Cell Rep*. 27:1461–62008;

65. Nanasato Y, Mikami M, Futamura N, Endo M, Nishiguchi M, Ohmiya Y, et al.. CRISPR/Cas9-mediated targeted mutagenesis in Japanese cedar (Cryptomeria japonica D. Don). *Sci Rep*. 11:161862021;

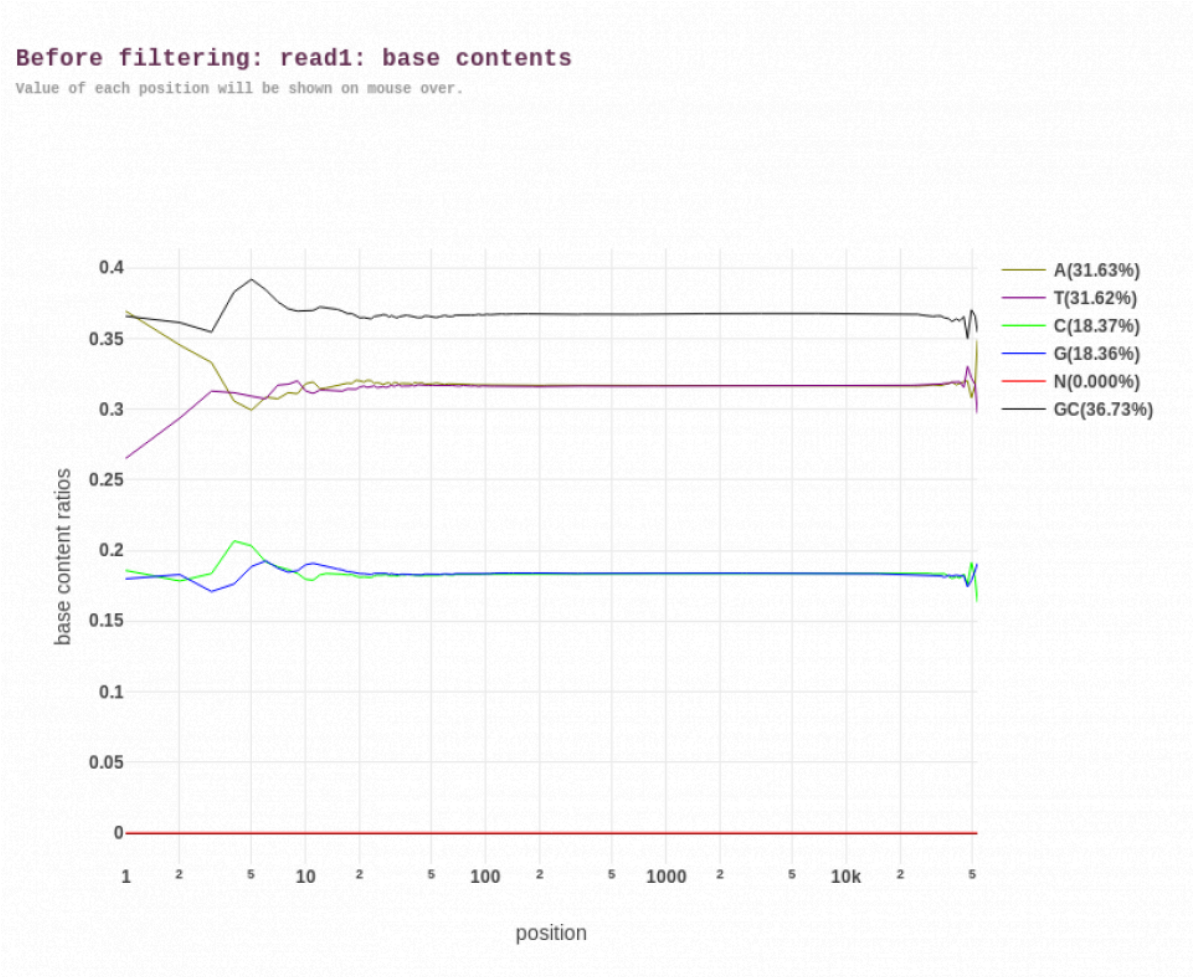
66. Taira H, Teranishi H, Kenda Y. A case study of male sterility in sugi (*Cryptomeria japonica*). *JOURNAL OF THE JAPANESE FORESTRY SOCIETY*. 75:377–91993;

67. Kakui H, Ujino-Ihara T, Hasegawa Y, Tsurisaki E, Futamura N, Iwai J, et al.. Single-nucleotide substitution determines pollen production in Japanese cedar. bioRxiv.

Supplementary Materials

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

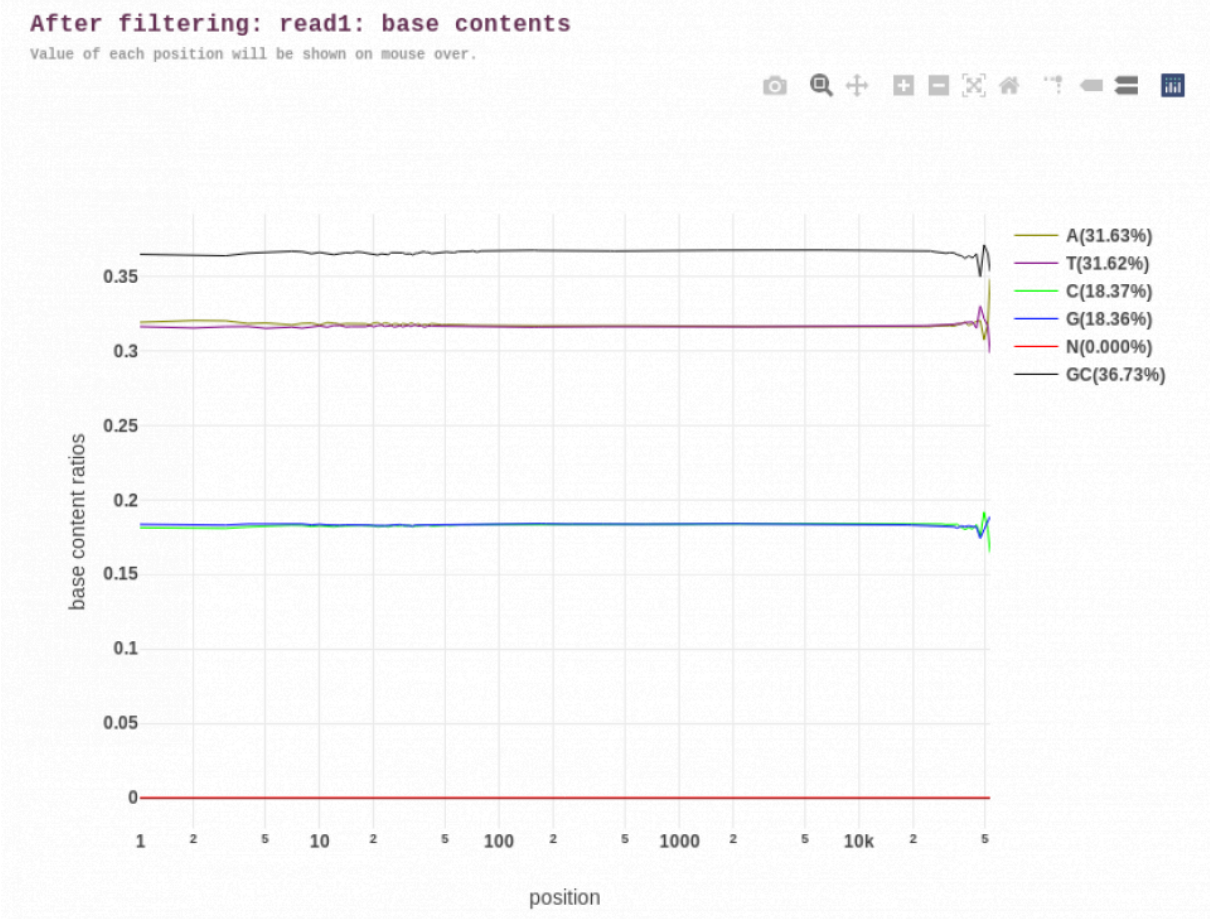
Supplementary Figure 1



Supplementary Figure 1: Base Composition Report of the HiFi reads by FastQC (before trimming)

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

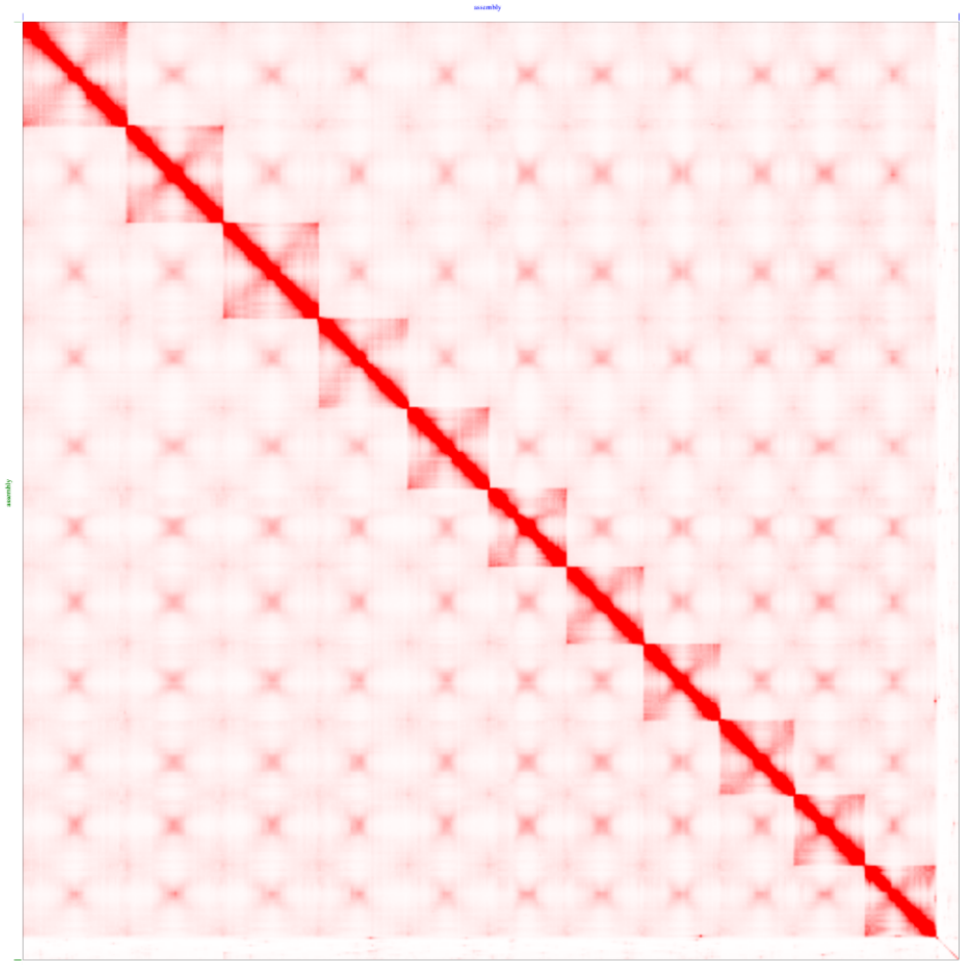
Supplementary Figure 2



Supplementary Figure 2: Base Composition Report of the HiFi reads by FastQC (after trimming)

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

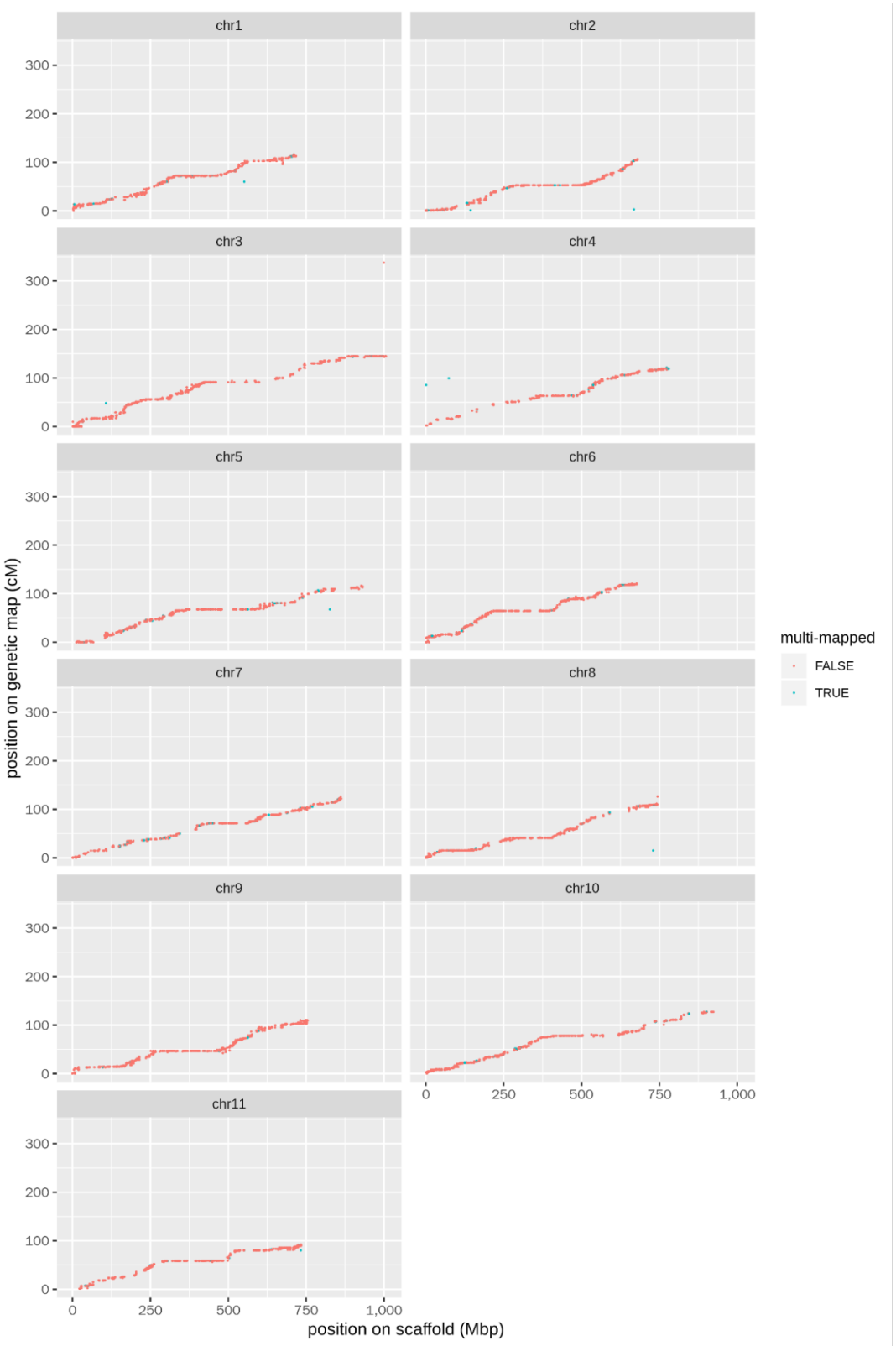
Supplementary Figure 3



Supplementary Figure 3: Final Contact Map of the Hi-C analysis (after the manual correction). Eleven chromosomes are clearly seen.

bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Supplementary Figure 4



bioRxiv preprint doi: https://doi.org/10.1101/2023.02.24.529822; this version posted February 24, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Supplementary Figure 4: Relationship between the marker positions (cM) and the physical position (Mbp) of the markers. The genetic map and the scaffolds are highly consistent except for the multi-mapped markers where the sequence aligns to multiple locations on the genome.