Tree Genetics & Genomes (2015) 11: 50
DOI 10.1007/s11295-015-0859-9

ORIGINAL PAPER

Analyses of random BAC clone sequences of Japanese cedar, *Cryptomeria japonica*

Miho Tamura[1] · Yosuke Hisataka[1] · Etsuko Moritsuka[2] · Atsushi Watanabe[3] · Kentaro Uchiyama[4] · Norihiro Futamura[5] · Kenji Shinohara[5] · Yoshihiko Tsumura[4,6] · Hidenori Tachida[2]

Received: 8 August 2014 / Revised: 2 March 2015 / Accepted: 10 March 2015 / Published online: 5 May 2015

Abstract Conifers have larger genomes than most angiosperms, long generation times, and undergone relatively few chromosome duplications during their evolution. Thus, conifers are interesting targets for molecular evolutionary studies. Despite this, there have been few studies regarding their genome structure, and these studies are mostly limited to the Pinaceae. Our target species, *Cryptomeria japonica*, belongs to the Cupressaceae family, which is phylogenetically separated from the Pinaceae family by a few hundred million years, and is the most important timber tree in Japan, making investigation of its genome structure both interesting and worthwhile. We analyzed the sequences of eight random bacterial artificial chromosome (BAC) clones from *C. japonica* and compared them with sequences of comparable size from eight other model plants, including *Arabidopsis thaliana* and *Pinus taeda*. From this analysis, we identified several features of the *C. japonica* genome. First, the genome of *C. japonica* has many divergent repetitive sequences, similar to those of *Physcomitrella patens* and *P. taeda*. Additionally, some *C. japonica* transposable elements (TEs) seem to have been active until recently, and some might be unidentified novel TEs. We also found a putative protein-coding gene with a very long intron (approximately 70 kb). The three Pinaceae species whose genome sequences have been determined share these features, despite the few hundred million years of independent evolution separating the Pinaceae species from *C. japonica*.

Keywords BAC clone · Cupressaceae · Repeat elements · Intron length

Communicated by P. Ingvarsson

Topical Collection on *Genome Biology*

Electronic supplementary material The online version of this article (doi:10.1007/s11295-015-0859-9) contains supplementary material, which is available to authorized users.

✉ Hidenori Tachida
htachida@kyudai.jp

[1] Graduate School of Systems Life Sciences, Kyushu University, Fukuoka 812-8581, Fukuoka, Japan

[2] Department of Biology, Faculty of Sciences, Kyushu University, Fukuoka 812-8581, Fukuoka, Japan

[3] Department of Forest Environmental Science, Faculty of Agriculture, Kyushu University, Fukuoka 812-8581, Fukuoka, Japan

[4] Department of Forest Genetics, Forestry and Forest Products Research Institute, Tsukuba 305-8687, Ibaraki, Japan

[5] Department of Molecular and Cell Biology, Forestry and Forest Products Research Institute, Tsukuba 305-8687, Ibaraki, Japan

[6] Faculty of Life & Environmental Sciences, University of Tsukuba, Tsukuba 305-8572, Ibaraki, Japan

Introduction

The whole genome sequencing of many model plant species, including *Arabidopsis thaliana* (The Arabidopsis Genome Initiative 2000), *Oryza sativa* (Goff et al. 2002; Yu et al. 2002), *Glycine max* (Schmutz et al. 2010), *Populus trichocarpa* (Tuskan et al. 2006), and *Physcomitrella patens* (Rensing et al. 2008), has been completed. These studies have revealed the diversity of the plant genome structure in terms of total size, nucleotide composition, and the number and type of coding genes and transposable elements (TEs). These features are considered products of evolution and provide insight into how these genomes have formed.

Although the list of sequenced genomes includes a wide variety of land plants, including trees, a moss (*P. patens*; Rensing et al. 2008), and a lycophyte (*Selaginella moellendorffii*; Banks et al. 2011), no conifer whole genome

sequence has been reported until recently. This apparent delay in conifer sequencing is due to the large size of conifer genomes—usually in excess of 10 Gb (Ohri and Khoshoo 1986; Ahuja and Neale 2005)—and the large amount of repetitive sequence in their genomes (Miksche and Hotta 1973). Because conifers have their own evolutionary histories of hundreds of millions years, their genome features are expected to be different from those of the previously sequenced plants. Considering the ecological importance and utility of conifers as timber, their genome structure is important and interesting.

Recently, draft assemblies of the genomes of conifers, *Picea abies*, *Picea glauca*, and *Pinus taeda*, belonging to the Pinaceae family have been reported (Nystedt et al. 2013; Birol et al. 2013; Neal et al. 2014; Wegrzyn et al. 2014). These genome assemblies revealed many interesting features, including diverse sets of TEs and long introns. Although whole genome sequencing is the best strategy to determine genome structure, whole genome sequencing of conifer genomes is very time- and resource-consuming, even with next generation sequencing. Two alternative approaches have therefore been used to investigate conifer genome structure. In the first approach, bacterial artificial chromosome (BAC) libraries were developed, and the sequences of selected BAC clones containing certain target genes were preliminarily analyzed (Keeling et al. 2010; Hamberger et al. 2009; Kovach et al. 2010). However, because the clones were not chosen randomly and the regions that were investigated were limited to those surrounding the genes, the general features of the genome could not be determined using this method. In the second approach, the extent of sequence repetition in the genome was estimated by generating numerous short reads by next generation sequencing (Kovach et al. 2010) or by Cot analysis (Liu et al. 2011), revealing high proportions of divergent repetitive sequences in conifer genomes. However, we still cannot infer the structure, such as how repeated sequences are distributed within a region or among regions, of conifer genomes by this approach.

In this study, we analyzed the sequence of eight randomly chosen BAC clones of *Cryptomeria japonica* by comparing their features with those of other plant genomes. These eight BAC clones were randomly chosen from a BAC library that was recently developed by the Forestry and Forest Products Research Institute in Japan. *C. japonica* belongs to the Taxodioideae of Cupressaceae sensu lato (Gadek et al. 2000; Kusumi et al. 2000). The Taxodioideae includes *Taxodium distichum*, for which a BAC library has recently been developed (Liu et al. 2011). The genome size of *C. japonica* is estimated to be 22.09 pg, corresponding to approximately 11 Gb; it is slightly larger than the 19.90 pg *T. distichum* genome (Hizume et al. 2001). Because *C. japonica* is the most important timber tree in Japan, extensive genetic and population genetic studies have been performed: its total genetic map is 1405.2 cM (Moriguchi et al. 2012), the average silent nucleotide diversity is 0.38 % (Kado et al. 2003), and populations of the two main varieties of this species have clearly diverged from each other (Tsumura et al. 2012). The average divergence at silent sites between *C. japonica* and *T. distichum* is estimated to be 8.98 % (Kado et al. 2003).

Here, using the sequences of the eight random BAC clones, we investigated features of the *C. japonica* genome, including the nucleotide composition and the number and types of coding genes and TEs. To assess these features from a comparative perspective, we compared sequences of the BAC clones with fragments of genome sequences of other plants by sampling similar amounts of sequence from their genomes and by carrying out the same analyses as those performed for *C. japonica*. Some of the results presented here have been cited briefly in Moritsuka et al. (2012) as "M. Tamura, A. Watanabe, K. Uchiyama, N. Futamura, K. Shinohara, Y. Tsumura, H. Tachida, unpublished results," but here we describe the analyses in detail.

Methods

Sequences

Eight BAC clones (Table 1) were randomly selected from a *C. japonica* BAC library. The library consists of approximately 368,000 BAC clones covering approximately 4.4 times the total length of the genome and was developed by the Forestry and Forest Products Research Institute in Japan. The eight BAC clones were named BAC1–BAC8. They were sequenced using a Roche-454 Genome Sequencer FLX Titanium (Roche, Indianapolis, IN, USA) by TaKaRa Bio Inc. (Ohtsu, Japan). All of the reads from each BAC clone were assembled using the GS De Novo Assembler version 2.0 (Roche, Indianapolis, IN, USA) with its default setting. We first removed sequences of bacterial origin from the contigs of the assembled sequences of each BAC clone and then selected all contigs larger than 4 kb for the present analysis. The total length of the sequences used for this analysis was 1,329,495 bp (Table 1, Supplementary Table S1).

To place the features of the *C. japonica* genome in a comparative perspective, we compared them to those of other previously sequenced plant genomes: *A. thaliana*, *P. trichocarpa*, *O. sativa*, *Zea mays*, *S. moellendorffii*, *P. patens*, *P. abies*, and *P. taeda*. For *A. thaliana*, *P. trichocarpa*, *O. sativa*, *Z. mays*, *S. moellendorffii*, and *P. patens*, eight genomic regions, each with a length of 165 kb, were randomly chosen from the whole genome sequence (respectively: TAIR, http://www.arabidopsis.org/; the Rice Genome Annotation Project, http://rice.plantbiology.msu.edu/; the JGI Populus Genome database, http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html; Phytozome database version 9, http://ftp.jgi-psf.org/

Table 1 Summary of *C. japonica* BAC assemblies

	Clone ID	No. of contigs	No. of reads	Sum of the reads (bp)	No. of contigs (>4 kb)	Sum of the lengths of the contigs (>4 kb)	Average depth of the contigs
BAC1	CJ-012B01	137	37,677	13,606,461	5	186,284	65.4
BAC2	CJ-012C01	23	41,463	14,939,070	3	201,114	69.1
BAC3	CJ-012D01	86	18,875	6,770,323	14	255,146	22.3
BAC4	CJ-012E01	122	39,030	13,859,796	5	207,302	60.8
BAC5	CJ-012F01	226	17,232	6,032,122	9	138,861	35.5
BAC6	CJ-012G01	38	6126	2,200,838	2	139,805	13
BAC7	CJ-012H01	16	25,820	9,343,926	2	109,566	77.8
BAC8	CJ-012I01	7	26,189	9,079,365	2	91,417	85.4

pub/compgen/phytozome/v9.0/Smoellendorffii/ and http://ftp.jgi-psf.org/pub/compgen/phytozome/v9.0/Ppatens_v1.6/), so that the total length (1320 kb) of these regions was comparable to that of the sequences of the eight random *C. japonica* BAC clones. We repeated this procedure five times for each of the six species and calculated the averages of statistics characterizing the respective genomes. For *P. abies*, we used the eight longest scaffolds excluding putative contamination from the genome assembly *P. abies* 1.0 (ftp://congenie.org/congenie//fasta/GenomeAssemblies/Pabies1.0-genome.fa.gz). The total length was 1,388,513 bp. For *P. taeda*, we chose eight scaffolds from *P. taeda* assembly v1.01 (http://loblolly.ucdavis.edu/bipod/ftp/Genome_Data/genome/pinerefseq/Pita/v1.01/), whose lengths were more than 165 kb and whose proportions of undetermined nucleotides "N" were smallest. Their total length was 929,539 bp excluding "N"s. The scaffolds used in the analyses for *P. abies* and *P. taeda* are listed in Supplementary Table S2.

The BAC clone sequences of *C. japonica* obtained for this study have been deposited with the DNA Data Bank of Japan (DDBJ) under the accession numbers DDBJ: AP014614–AP014621 (see Supplementary Table S1).

Nucleotide composition

First, we examined the guanine-cytosine (GC) content in the sequence of each plant genome. We then computed the frequencies of dinucleotides (CpG dinucleotides, etc.) and the ratio of the observed dinucleotide frequency to that expected from the GC content using a custom perl script written in our laboratory. The script is available on request from MT.

Identification of transposable elements and other repeats

To identify TEs and other repeated sequences, we used the following four approaches. First, the TEs were identified based on sequence similarity to known transposable elements in Repbase (Jurka et al. 2005; Jurka 1998, 2000), a database of repetitive DNA elements (Jurka 2000), using CENSOR (Kohany et al. 2006) with its default parameters. Second, to identify unknown repetitive sequences, we searched using BLASTN with its default parameters for segments of more than 50 bp in length that were highly similar (more than 80 %) to other segments in the BAC clone sequences (self comparison of sequences within a species); the BAC clone sequences of a species were used as both subjects and queries. If such segments were found, we regarded them as "repeats." Third, consensus sequences of repetitive elements were identified using RepeatScout (Price et al. 2005) with its default parameters. RepeatScout extracts consensus sequences of repeat element families from the target sequences and then RepeatMasker (2013) is used to identify the repeat elements that are similar to either one of these consensus sequences within the target sequences. Fourth, we used Tandem Repeats Finder with its default parameters to identify simple tandem repeats (Benson 1999). We performed all of these analyses on each of the nine species.

For *C. japonica*, we conducted further detailed analyses of TEs and repetitive sequences. First, we identified consensus sequences and associated repeat families using RepeatScout and RepeatMasker. Some of the consensus sequences had homology with each other. If this was the case, we chose the shortest one in the group of homologous consensus sequences as the representative consensus sequence and discarded the longer ones. In most such cases, the shortest consensus sequence had similarity to almost all members of the group. Therefore, most of the members of the repeat families originally identified by RepeatScout were included in either one of the families associated with the shortest consensus sequences. Then, we selected repeat elements containing more than 50 % of the consensus sequence in each repeat family. Those repeat elements were aligned, and the average pairwise divergence within each family was computed. In this analysis, we classified the repeat element families into a TE-like group and an unknown group. The repeat element families in the TE-like group contained at least one member that has been identified

as a TE by CENSOR. The other repeat element families belonged to the unknown group. Second, long terminal repeat (LTR) retrotransposons were identified using LTR_FINDER (Xu and Wang 2007) with its default parameters. In the present paper, we use acronyms LTR-RT and LTR to indicate the entire long terminal repeat retrotransposon identified by LTR_FINDER and the single long terminal repeat, respectively. If an LTR-RT was found by LTR_FINDER, coding sequences were searched for in the region flanked by its LTRs using BLSTX and nonredundant protein sequences database (nr) of NCBI.

Gene identification

We searched for coding genes in *C. japonica*. First, to identify the transcribed regions we ran BLASTN with its default parameters, using the ForestGEN *C. japonica* cDNA database (http://forestgen.ffpri.affrc.go.jp: Futamura et al. 2008) developed by the Forestry and Forest Products Research Institute in Japan as a query and the eight BAC clone sequences as subjects. Then, among the identified hits, we searched for complementary DNA (cDNA) sequences with more than 95 % of their sequences showing more than 98 % similarity to some part of the BAC clone sequences.

Molecular analyses

We found several putative LTR-RTs with unknown internal sequences using LTR_FINDER (see "Identification of transposable elements and other repeats"). However, their ability to retrotranspose in the genome was not known. Therefore, we chose two of them (BAC4_LTR1 and BAC7_LTR1) and investigated their multiplicity in the genome and their divergence by sequencing their PCR products. For the PCR, we used haploid DNA that was extracted from a megagametophyte from a seed of *C. japonica* as template. First, pairs of primers were designed from the LTR pair sequence of each putative LTR-RT to amplify the region between the LTR pair. The lengths of the PCR fragments were 3.5 and 6 kb for BAC4_LTR1 and BAC7_LTR1, respectively. Note that because LTRs are direct repeat sequences, both primers could potentially match the LTR pair. Therefore, the primers were designed so that only the region flanked by the LTR pair could be amplified (see Supplementary Fig. S1). The primer sequences are listed in Supplementary Table S3. The PCR was performed on 1~10 ng/μl DNA using TaKaRa Ex Taq (TaKaRa, Ohtsu, Japan). The PCR conditions were 3 min at 95 °C; 30 cycles of 30 s at 95 °C, 30 s at 50 °C, and 3 or 5 min at 72 °C; and a final extension of 7 min at 72 °C. Because the amplified fragments of each putative LTR-RT were of various sizes, those with sizes of approximately 3.5 and 6 kb were extracted from agarose gels using the Wizard® SV Gel and PCR Clean-Up System (Promega, Madison, USA) so that fragments of similar size to BAC4_LTR1 and BAC7_LTR1, respectively, could be obtained. The extracted fragments were cloned into a pGEM-T easy vector (Promega, Madison, WI, USA), amplified by colony PCR and directly sequenced using an Applied Biosystems 3730 DNA Analyzer and the BigDye Terminator v3.1 Cycle Sequencing Kit (Life Technologies, CA, USA). The two ends of each amplified fragment were sequenced using M13 primers (M13F 5′-CGCCAGGGTT TTCCCAGTCACGAC-3′ and M13R 5′-TCACACAGGA AACAGCTATGAC-3′). The resulting sequences were aligned using ClustalW (Thompson et al. 1994) with its default setting and manual adjustments, and neighbor-joining trees (Saitou and Nei 1987) that were based on the Kimura 2-parameter distance (Kimura 1980) were constructed using MEGA 5.0 (Tamura et al. 2011).

As described in the "Results", we found a large putative intron in the BAC clone sequences. We suspected that the large size of the intron might be an artifact caused by the misassembly of the reads. To confirm its length, several primers were designed to bind within the intron and used to amplify segments of DNA from four individuals by long range PCR with LA Taq (TaKaRa, Ohtsu, Japan). The approximate lengths of the segments were estimated using gel electrophoresis. Additionally, to determine whether there are any paralogs of the gene, we amplified the exons. A pair of closely positioned exons of the gene was found on each side of the intron. We amplified each pair of exons from *C. japonica* haploid DNA (megagametophyte) using primers designed to bind in the exons and directly sequenced the products. Furthermore, to determine whether the putative gene was a pseudogene, we measured the ratio of the nonsynonymous to synonymous divergences by determining the corresponding regions in *T. distichum*. For sequencing, we used an Applied Biosystems 3730 DNA Analyzer with a BigDye Terminator v3.1 Cycle Sequencing Kit (Life Technologies, CA, USA). All sequences obtained for this study from *C. japonica* and *T. distichum* have been deposited with the DDBJ under the accession numbers DDBJ: AB849633–AB849692.

Dot plot and graphical summary

To summarize the obtained results graphically, a gff3 file of the longest contig of each BAC clone was created, and Apollo (Lee et al. 2009) was used to draw graphical summaries of the results. A dot plot of all contigs was drawn using GenomeMatcher (Ohtsubo et al. 2008) with its default parameters.

Results

C. japonica BAC sequences

We chose contigs whose sizes were more than 4 kb and which did not contain sequences of bacterial origin from each

C. japonica BAC clone and used these contigs for the analyses (Supplementary Table S1). The total length of the contigs for each BAC clone ranged from 91,417 to 255,146 bp, and the average read depths of the selected contigs ranged from 13.0 to 85.4 (Table 1).

Summaries of the results are shown graphically in Fig. 1 and Supplementary Fig. S2, and the dot plots of all the contigs used are shown in Supplementary Fig. S3. The details of the analyses are described below.

Nucleotide composition

The GC contents of the approximately 1320-kb sampled sequences of *A. thaliana*, *P. trichocarpa*, *O. sativa*, *Z. mays*, *S. moellendorffii*, *P. patens*, *P. abies*, *P. taeda*, and *C. japonica* were 35.64, 33.45, 43.58, 46.94, 45.73, 33.32, 37.12, 34.35, and 36.98 %, respectively (Table 2). The GC contents of the three conifers were quite similar to each other and were within the range spanned by the other plants. Moreover, the variation in the GC content among BACs within a species as expressed by the standard deviation (SD) was small and similar among the plants investigated here (0.67–1.81 %).

Next, we examined the dinucleotide frequencies. If neighboring nucleotide substitutions occur independently, we would expect the dinucleotide frequency to be the product of the single nucleotide frequencies. We computed the ratio of the observed dinucleotide frequency to that expected from the GC content of each species; the results are shown in Supplementary Fig. S4. Although the well-known CpG dinucleotide deficiency (which is partially compensated for by increases of TpG and CpA dinucleotides) was found in all nine species, the deficiency was higher in the three conifers, especially in *C. japonica* (Fig. 2).

Repeated sequences

Repeated sequences in the nine species were identified by CENSOR, BLASTN (self comparisons), RepeatScout, and Tandem Repeats Finder and their proportions are given in Table 2. The proportion of sequences that are homologous to known transposable elements was smaller in conifers (*P. abies*, 35.71 %; *P. taeda*, 13.68 %; *C. japonica*, 22.95 %) than in the other plants (26.75–83.43 %). While *A. thaliana*, *P. trichocarpa*, and *O. sativa*, respectively, contained comparable amounts of LTR-RTs and DNA transposons, the other plants, including *C. japonica* and *P. abies*, contained a much larger proportion of sequences homologous to LTR-RTs. In *S. moellendorffii* and *P. patens*, *gypsy*-like elements were at least tenfold more prevalent than *copia*-like elements; however, their prevalence was only up to fourfold greater in the seed plants.

Although the proportions of repeated sequences that were identified by BLASTN (self comparison) varied greatly among angiosperms (7.89–66.37 %), the ranges were not so large among the other plants (ca. 20–40 %). Note that each sequence was identified as a repeated sequence when there was at least one other similar fragment of no less than 50 bp that shared no less than 80 % similarity within the 1320-kb region found in each species. Although the total proportion of repeated sequences was similar among non-angiosperms, the level of similarity among the repeats was different. Figure 3 shows the level of similarity between the repeated sequences that were identified by BLASTN (self comparison). Among the repeated sequences with similarities of 80 % or more, the ratio of divergent repeats (80–90 %) to similar repeats (90–100 %) was higher in *P. patens*, *P. abies*, and *C. japonica* than in the other plants. Note that because our BLASTN (self comparison) search may miss some of the divergent repeats, this trend might be even stronger. Thus, the amplification of repeated sequences in these three species seemed to have occurred earlier than in the other plants if we assume the evolutionary rates of nucleotide sequences among species to be constant.

The proportion of the repeated sequences identified by RepeatScout and BLASTN (self comparison) but not by CENSOR was large in *C. japonica* and *P. taeda*. However, it was generally small in the other seven species (Supplementary Fig. S5).

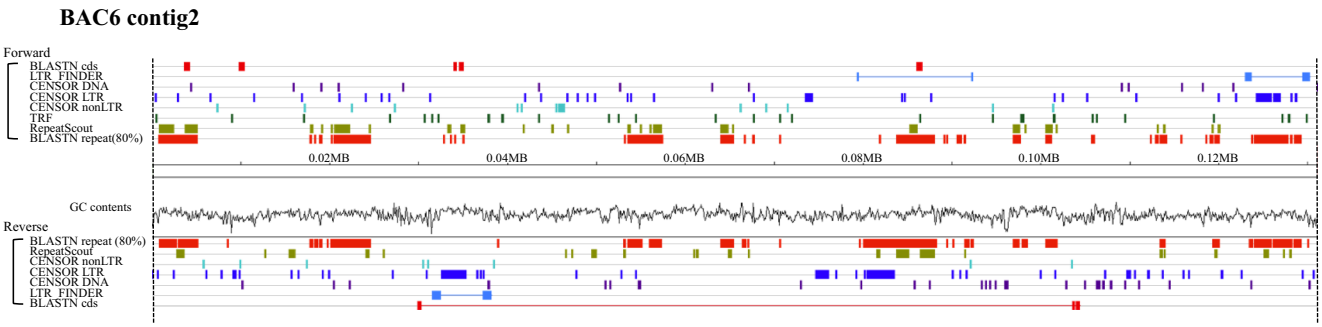

Fig. 1 Annotated sequence of the longest contig of BAC6. The position in the contig is shown in the *middle*. Annotations are given in the forward and reverse directions. The content of the gff3 file used to draw this figure is found in the Supplementary document

Table 2 Proportions of repetitive sequences

	A. thaliana	*P.trichocarpa*	*O. sativa*	*Z. mays*	*S.moellendorffii*	*P. patens*	*P. abies*	*P. taeda*	*C.japonica*
Genome size (Mb)	115	480	430	2300	100	511	20,000	22,000	11,000
GC contents[a] (%)	35.64 (1.61)	33.45 (1.00)	43.58 (1.81)	46.94 (1.76)	45.73 (1.53)	33.32 (1.53)	37.12 (0.67)	34.35 (1.51)	36.98 (1.14)
Transposons[b] (%)	26.75	27.95	42.42	83.43	36.68	51.85	35.7	13.68	22.95
Class II DNA transposon	14.01	11.08	18.49	5.31	6.13	5.05	2.76	5.00	3.61
Class I LTR retrotransposon	9.08	14.37	21.36	76.65	28.95	45.18	26.76	6.01	16.96
Copia	2.86	3.57	3.67	27.92	1.25	3.57	7.2	1.52	8.79
Gypsy	5.9	9.38	13.77	48.29	26.54	41.78	18.75	2.99	7.3
Non-LTR retrotransposon	2.98	2.04	2.06	1.18	1.3	0.88	1.88	1.36	1.77
Repeat by BLASTN (self comparison) contents (%)	7.89	16.61	17.9	66.37	36.82	44.31	21.34	39.58	44.93
80–90 %	2.4	5.81	7.66	17.69	12.18	28.42	15.1	18.47	27.38
90–100 %	5.49	10.8	10.24	48.68	24.64	15.89	6.24	21.12	17.55
RepeatScout contents (%)	6.01	11.02	11.94	55.9	27.7	44.65	5.04	12.47	29.38
Total repeat contents[c] (%)	31.28	38.87	45.12	86.54	47.33	61.51	47.98	50.11	59.38
Tandem repeats contents (%)	2.25	2.78	2.72	2.72	2.08	2.14	5.34	3.85	2.08

[a] Standard errors are indicated in parentheses

[b] Transposons identified by CENSOR

[c] Total repeat contents identified by CENSOR, BLASTN (self comparison), or RepeatScout

The repeated sequences that were identified by CENSOR, BLASTN (self comparison), and RepeatScout overlap. To determine the total proportion of repeated sequences, we counted the total number of repeats considering these overlapping sequences (Table 2). Because the proportions of uncharacterized repeated sequences identified by BLASTN (self comparison) or RepeatScout but not by CENSOR were high, the total number of repeats was higher in the three conifers than in *A. thaliana* and *P. trichocarpa*. However, the number of repeats in the three conifers was lower than in *Z. mays*.

The proportions of the tandem repeats that were identified by the Tandem Repeats Finder in *P. abies* and *P. taeda* were higher than in the other plants, but the proportion in *C. japonica* was similar to those in angiosperms (Table 2).

To investigate the characteristics of repeated sequences in *C. japonica* in more detail, we carried out two additional analyses. First, we examined the variation in the proportion of repeated sequences among the eight BACs (Fig. 4). We classified the repeated sequences into three groups based on how they were identified: only by CENSOR, only by homology (BLASTN (self comparison) or RepeatScout), or by both methods. Although the average proportion was 59.38 %, the variation in the proportion of repeated sequences among the BACs was great (31.52–82.75 %). Note that this variation can be mainly attributed to the variation in the repeated sequences identified only by homology (9.23–54.46 %). The proportion of repeated sequences identified only by homology was not correlated with the size of the BAC clone ($P<0.320$,

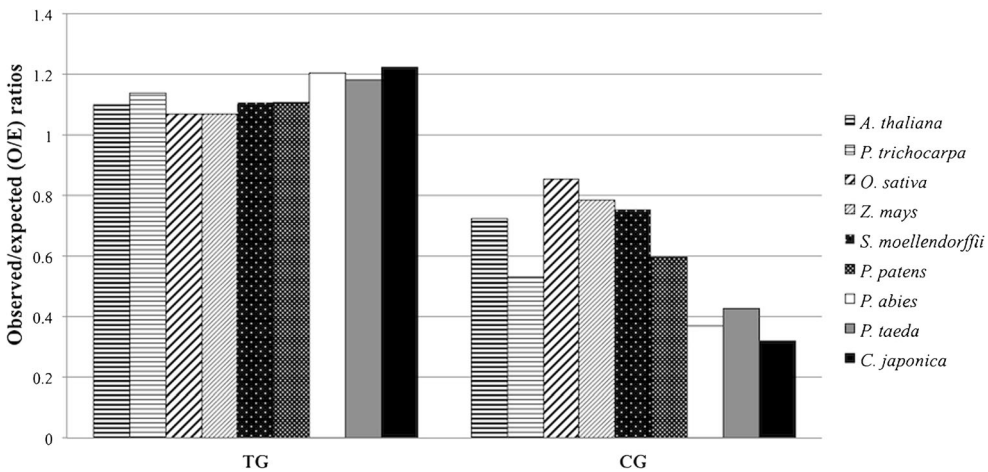

Fig. 2 The ratio of the observed to expected dinucleotide frequencies for each species

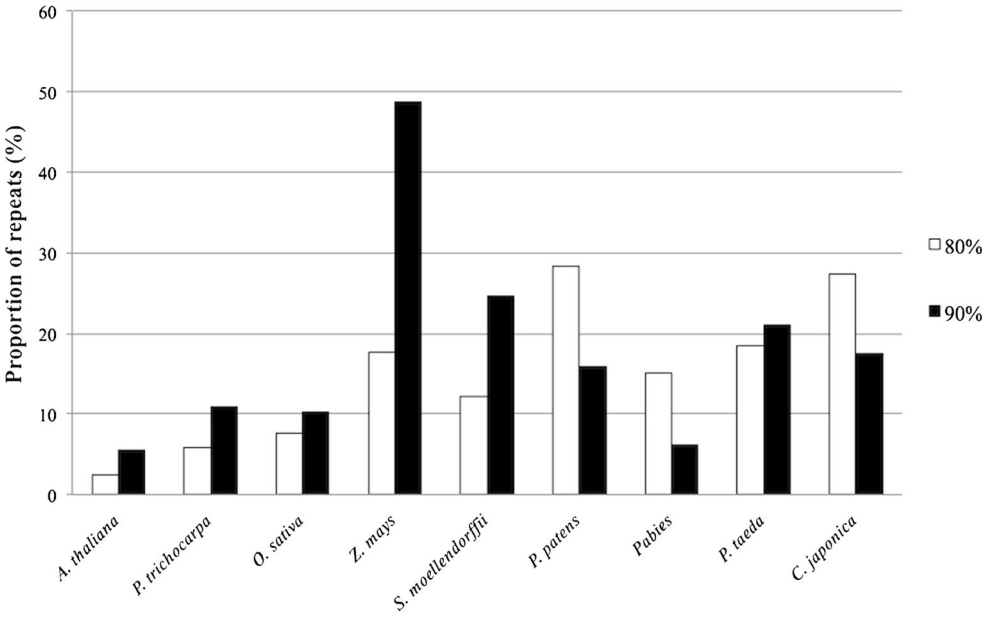

Fig. 3 The proportion of repeated sequence families classified according to divergence within families

Pearson correlation) or with the number of protein-coding genes contained in the BAC clone.

Next, we investigated the evolutionary timing of the amplification of repeated sequences in *C. japonica*. RepeatScout was useful in this investigation because it groups repeated sequences into families by identifying consensus sequences. RepeatScout identified 130 different repeat element families in *C. japonica*, but 9 of them were homologous to one of the other repeat element families; thus, we regarded that there were 121 different repeat element families. Furthermore, nine of these families consisted only of elements with lengths less than 50 % of that of the consensus sequences. We eliminated these families due to difficulties in estimating the divergence within families and calculated the average distance of each repetitive element family for the remaining 112 families. Figure 5 shows the distribution of the average divergences. The mean of the average divergence across families was 0.233. Some families had a low divergence, indicating a recent amplification if the substitution rates were the same among the families, but generally, the divergence was more than 0.10. Among the 112 families, 96 had homology to known TEs and are referred to as the TE-like group. Their mean divergence was 0.252. In contrast, the other 16 families had no homology to known TEs and are referred to as the unknown group. The mean of the average divergence in the unknown group was 0.115 and was significantly lower ($P=7.65\times 10^{-7}$) than that of the TE-like group. We further divided the TE-like group into three subgroups based on

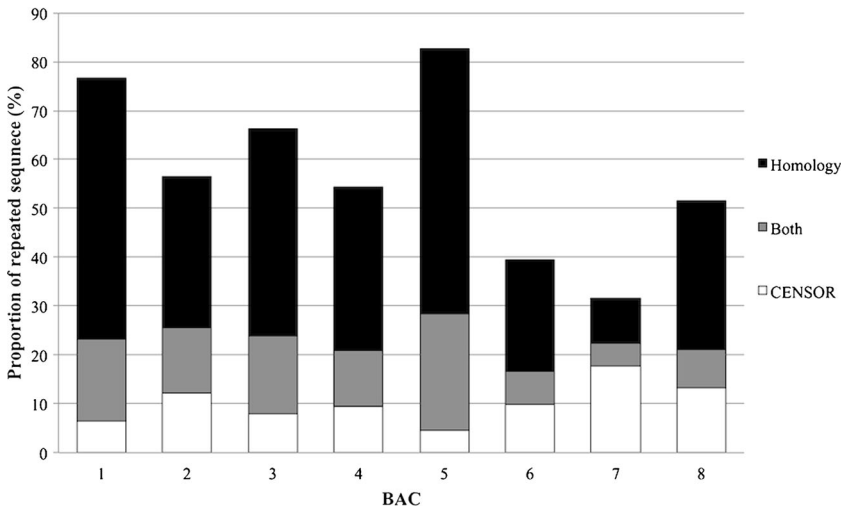

Fig. 4 The proportion of repeated sequences among the eight BAC clones in *C. japonica*. The repeated sequences were classified into three groups based on the method of identification. *White*, only by CENSOR; *black*, only by homology (BLASTN (self comparison) or RepeatScout); *gray*, by both methods

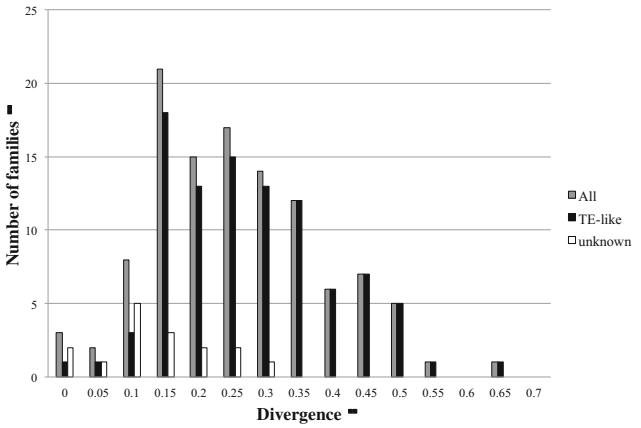

Fig. 5 The distribution of the average divergence within each repetitive element family in *C. japonica. Gray*, all repetitive element families; *black*, the TE-like group; *white*, the unknown group

homology to DNA transposons, LTR-RTs, and non-LTR retrotransposons. The average divergences of the three subgroups were nearly the same (0.264–0.273).

Finally, we used LTR_FINDER, which searches for LTR-RTs based on structure rather than on similarity to known transposon sequences. In total, 19 putative LTR-RTs were identified (Supplementary Table S4). Sequences homologous to known retrotransposons were identified by CENSOR in 14 out of the 19 putative LTR-RTs. However, in the remaining five putative LTR-RTs, no retrotransposon-like sequences were identified using CENSOR or BLASTx. Multiple copies of one of these sequences (BAC4_LTR1) were found in BAC4, and a partial fragment of the sequence was found in BAC3.

To examine whether those putative LTR-RTs without known flanked retrotransposon-like sequences were transposable, two pairs were selected (BAC4_LTR1 and BAC7_LTR1), and each of the regions flanked by them was amplified from the genomic DNA of an individual of *C. japonica*. The fragments that were amplified by PCR were cloned; 13 (BAC4_LTR1) and 15 (BAC7_LTR1) different sequences were obtained (see Supplementary Figs. S6 and S7) and were found in multiple locations within the *C. japonica* genome. We refer to these repetitive sequences as BAC4_LTR1 and BAC7_LTR1, respectively. Note that we sequenced only approximately 500 bp of each end of the intervening sequences flanked by the LTRs. We first examined the correlation of the divergence between the sequences of the forward end with those of the reverse end using the Mantel test. The correlations were highly significant ($P<0.0001$) for both BAC4_LTR1 and BAC7_LTR1, indicating that the forward and reverse sequences were transposed together. Therefore, we combined the sequences from both ends and constructed a neighbor-joining tree (Fig. 6). If we assume the substitution rates were the same in BAC4_LTR1 and BAC7_LTR1, the tree suggested that many of the multiplications of BAC7_LTR1, whose terminal branch lengths were shorter than those of BAC4_LTR1, occurred more recently than those of BAC4_LTR1.

Some other features of interest were found in the 19 putative LTR-RTs. One LTR-RT (BAC4_LTR2) identified in BAC4 was located between another LTR pair (BAC4_LTR1), indicating that insertion of the former occurred after that of the latter. Such nested structure of LTR-RTs has been reported in other eukaryotic genomes (Gao et al. 2012). BAC6_LTR2 inserted recently because its LTRs were identical. The LTR was duplicated when it was inserted. The LTR had the canonical LTR terminal TG-CA, and the region flanked by the LTRs contained a copia-like element (see Supplementary Table S4).

Coding regions

To identify currently transcribed genes, we searched against a cDNA database using BLASTN with a strict criterion: 95 % or more of a cDNA sequence had to exhibit >98 % similarity to a region or regions of one of the BAC clone sequences. Eight cDNAs met these criteria and eight corresponding BAC sequence regions were identified. The first cDNA had a corresponding exon on BAC7 and showed a strong similarity to a hAT family DNA transposon. The second and third cDNAs were homologous to fragments on BAC6 and showed strong similarity to a gypsy family retrotransposon. We think that they are actively transcribed, although the possibility of contamination of the cDNA library with genomic DNA cannot be excluded until we carry out additional experiments.

The fourth cDNA had homology to four segments, possibly corresponding to four exons, in the sequence of BAC6. This cDNA showed a high level of similarity to a calcium-dependent protein kinase gene (accession no. XP002511506); thus, these four exons seem to encode part of a gene. To find out whether the gene containing these exons has a paralog, we amplified and sequenced the exons from haploid DNA. Only one sequence for each exon was obtained, and the sequences had more than 98 % similarity to the BAC clone sequences. Thus, there seems no closely related paralog of this gene in the genome of *C. japonica*.

Next, we examined a possibility of this gene being a pseudogene. First, no stop codon was located within those four exons. Second, the fragments corresponding to the four segments in *T. distichum* were sequenced, and the ratio of nonsynonymous to synonymous divergences was estimated. The ratio of nonsynonymous to synonymous divergences was 0.086, indicating selective constraints on the nonsynonymous substitutions. Thus, this gene seemed alive at least until recently.

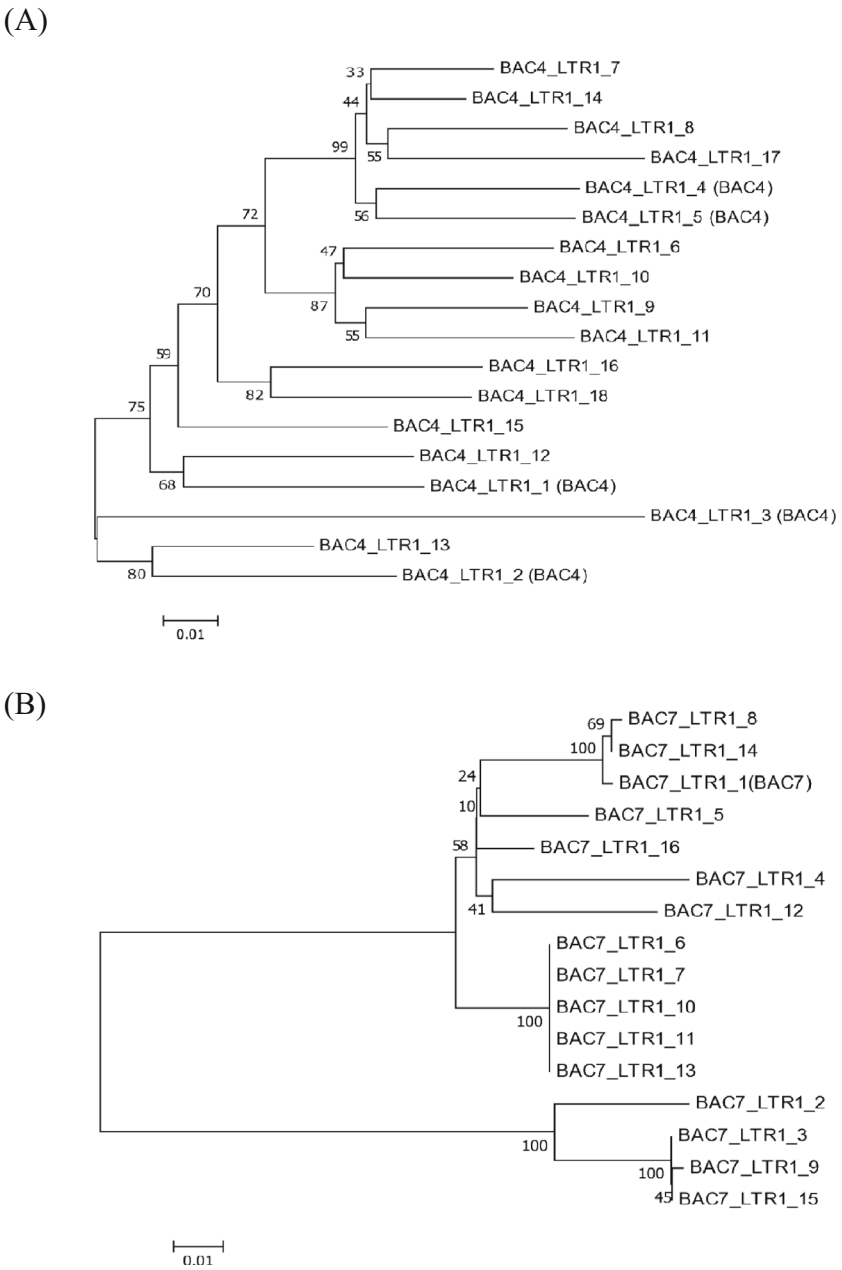

Fig. 6 Phylogenetic relationships of repetitive sequences, BAC4_LTR1 and BAC7_LTR1. Neighbor-joining trees of **a** BAC4_LTR1 and **b** BAC7_LTR1 are shown

Interestingly, the length of the sequence separating the second and third segments, a putative intron, was approximately 70 kb and very long. Its length was verified using gel electrophoresis of PCR fragments of the intron from four individuals (data not shown). This putative intron contained two LTR-RTs, one of which was homologous to the second and third cDNAs described above. The electrophoresis to confirm the length of the intron also showed that there was an insertion polymorphism with regard to one of the LTR-RTs. This result suggests that the insertion into this site of the LTR-RT was recent. No homology to known proteins in the nonredundant protein sequences database (nr) of NCBI was found in the remaining four cDNAs.

Discussion

Sequences of BAC clones and those extracted from genome assemblies

Here, we compared sequences of eight random BAC clones from *C. japonica* with genomic sequences of comparable size from eight model plants. This strategy was taken because we

wanted to know the effects of sampling small portions of the genome on our inferences concerning the general features of the *C. japonica* genome. However, there may be some systemic differences between sequences of random BAC clones and sequences extracted from genome sequences. If the genome sequence is fairly complete, like in *Arabidopsis* or rice, these differences might be small. However, if the genome sequence is a rough draft and mostly assembled from short reads, such as that of *P. abies*, regions with repeated sequences might not be assembled successfully and their contig sizes might be short. Thus, these repeated sequences might be omitted from or underrepresented in our samples, because we sampled sequences of size 165 kb or more from each genome sequence. This may have resulted in an underestimation of the proportion of repeated sequences in species for which only a draft sequence has been obtained.

CpG dinucleotide deficiency

Stronger CpG dinucleotide deficiencies were observed in conifers than in angiosperms. In general, the C of a CpG dinucleotide is prone to change to T by DNA methylation (Coulondre et al. 1978; Bird 1980). In conifers, increases in the frequencies of TpG and CpA dinucleotide pairs were also observed. The deficiency of CpG dinucleotides in conifers seems to indicate that DNA methylation occurs. In plants, DNA methylation has been found at CpG, CpNpG, and CpNpN sites (Gruenbaum et al. 1981; Meyer et al. 1994). DNA methylation contributes to chromatin structure (Razin 1998) and affects the activity of transposons (Miura et al. 2001). In addition, inactive transposons are densely methylated in plants (Chandler and Walbot 1986; Flavell 1984; Martienssen 1998). Because conifer genomes seem to contain many TEs, DNA methylation might have been used to suppress their activities as in angiosperms (Suzuki and Bird 2008), and CpG deficiencies may be a consequence of this process.

The CpG deficiency was stronger in conifers than in other plants (Fig. 2). Two factors may affect the level of CpG deficiency: one is the proportion of methylated regions in the genome and the other is the length of time elapsed since the genome methylation began. The methylated domains are interspersed in the genome of *A. thaliana*, whereas they are distributed throughout the genome of *Z. mays*; thus, the size of the methylated regions is larger in *Z. mays*, suggesting that the proportion of methylated regions should be larger (Suzuki and Bird 2008). However, Fig. 2 shows that the CpG deficiency is not stronger in *Z. mays* than in *A. thaliana.* This could be explained by recent amplification of TEs in *Z. mays* (see Table 2). Although we do not know whether the methylation pattern is localized in conifers, an ancient amplification of repeats might explain the high level of CpG deficiency in conifers.

Repeated sequences in conifers

CENSOR detected a smaller proportion of the genome occupied by known TEs in *C. japonica*. This is because TEs in *C. japonica* are less studied than those in the other plant species used in this study and because there may be many unknown or highly diverged TEs in conifers. Indeed, our BLASTN (self comparison) and RepeatScout searches identified more repeats in the BAC clone sequences of *C. japonica* than in those of the angiosperms, excluding *Z. mays*. Note that these sequence comparisons were made using only a few BAC clone sequences spanning approximately 1320 kb in each species. Because the genome sizes of conifers are much larger than those of the angiosperms analyzed in this study, the ratios of the sequences used here to the whole genomes are much smaller in conifers than in angiosperms. Thus, if we used the whole genome sequence of *C. japonica*, the proportion of repeats defined by BLASTN (self comparison) and RepeatScout is expected to be much higher. Indeed, in *P. abies* and *P. taeda*, whose genome sequences have been recently determined, the repeated sequences comprised 70 −80 % of the genome (Nystedt et al. 2013; Wegrzyn et al. 2014). This is consistent with *P. taeda* studies using whole genome shotgun sequencing (Kovach et al. 2010) and with studies of *T. distichum*, a species related to *C. japonica*, performed using the Cot analysis (Liu et al. 2011).

Interestingly, the proportion of repeated sequences identified by the homology-based method varied substantially among the BAC clones (Fig. 4). Therefore, the distribution of repeated sequences seems to not be uniform along the genome of *C. japonica*. If a BAC clone contains an abundance of repeated sequence, the assembly of reads may become more difficult and we would expect the length of the BAC clone sequence to be shorter. However, we did not find such a tendency. This may be because the repeated sequences are divergent in this species. In any case, nonuniformity of the distribution of repeated sequences among genomic regions is an interesting feature and should be studied further by increasing the number of BAC clones analyzed to obtain a more detailed understanding of the *C. japonica* genome structure.

Among the repeated sequences that were identified by BLASTN (self comparison), the proportion of those with 90 % or more similarity was much smaller than those with 80–90 % similarity among *C. japonica*, *P. abies*, and *P. patens*. This result suggests that the expansion of repeated sequences occurred earlier in these species than in angiosperms, assuming that the substitution rates are the same among the species. In fact, the substitution rates of conifers are known to be lower than those of angiosperms (Willyard et al. 2006). Thus, the evolutionary time of the expansion of repeated sequences in conifers might be even earlier than that suggested by the divergence data. Although the ancient expansion of repetitive elements has also been suggested in

P. taeda (Kovach et al. 2010; Wegrzyn et al. 2014), the extent of ancient expansion is much more extensive in *C. japonica*, *P. abies*, and *P. patens* (Fig. 3). Because *C. japonica* and *T. distichum* are separated by approximately 100 Ma (Leslie et al. 2012), and the silent divergence between them is estimated to be 8.98 % (Kado et al. 2003), major transposition events in *C. japonica* seem to have occurred more than 100 Ma ago, although the expansion might be more recent if TEs exhibit higher mutation rates than nuclear genes.

The RepeatScout analysis showed that the distribution of the average divergence within each repeat element family differed significantly between the TE-like group and the unknown group in *C. japonica* (0.252 and 0.115, respectively; Welch's two sample *t* test, $P=7.65\times10^{-7}$; see also Fig. 5). This result demonstrates that unknown repeat element families expanded more recently than did the TE-like repeat element families or that they were more conserved than the TE-like repeat element families. It is difficult to explain why this difference in the divergence occurred in the *C. japonica* genome. If the unknown repeat element families expanded more recently, one possible explanation would be as follows: because the TE-like families identified in *C. japonica* seemed to have existed for a long time, their transpositions might have been repressed by some defensive mechanism of the host genome, such that their expansion rate became low. However, if the unknown element families were more recent in origin, the *C. japonica* defensive mechanisms might not be effective against them, such that the rates of expansion might be still high in those element families. This difference in the rates of expansion in recent years may explain the difference in divergence between the two groups. Alternatively, unknown repeat element families may contain more conserved regions of unknown TEs than the TE-like repeat element families. However, because the TE-like repeat element families were mostly recognized by their protein-coding domains, we may expect their sequences to be equally or more conserved. However, the TE-like repeat element families were actually more divergent, and thus, this explanation seems unlikely.

Some of the putative LTR-RTs identified by LTR_FINDER did not contain LTR-RT-like regions between the LTR pairs. The PCR experiment showed that these LTR-RTs had multiple copies in the *C. japonica* genome and were possibly dispersed. These LTR-RTs may constitute the "large retrotransposon derivatives" (LARDS) that have been classified as nonautonomous LTR-RTs (Havecker et al. 2004). Alternatively, they might be new unknown types of LTR-RTs. In any case, because multiple copies exist in the *C. japonica* genome, the putative LTR-RTs that were identified computationally have been involved in the multiplication of the segments that they flank.

The average divergence within BAC4_LTR1 and BAC7_LTR1 (0.19 and 0.25) was similar to that of the repeat families identified by RepeatScout (0.233). Therefore, on average, these putative LTR-RTs seemed to have expanded around the same time as those found by RepeatScout if the substitution rates were the same among them.

Finally, although most expansions of repeated sequences seemed to have occurred in ancient times, we found some indications of recent multiplications. First, some of the LTR-RTs, belonging to BAC7_LTR1, were duplicated very recently (Fig. 6). Second, an LTR-RT located in the large intron of the putative calcium-dependent protein kinase gene showed the presence or absence of a polymorphism in *C. japonica*. Third, one LTR-RT (BAC6_LTR2) contained an identical pair of LTRs (a pair of LTRs was considered identical when the LTR-RT was inserted; Perlman and Boeke 2004). Thus, although most *C. japonica* TEs expanded in ancient times like those in *P. taeda* and *T. distichum*, some TEs might still have been active until recently and might transpose to other locations of the genome, although their rates of transposition may be low.

Protein-coding gene with a large intron

We found a putative protein-coding gene in BAC6. The gene seems to have no closely related paralog in the genome of *C. japonica.* It had a very large intron of 70 kb. Introns may become larger if the gene is inactivated because constraints on intron size disappear following inactivation. However, because the nonsynonymous to synonymous divergence ratio in this putative protein-coding gene in *C. japonica* and *T. distichum* was much smaller than 1, this gene seems to still be functional or had been functional until recently. Moreover, this gene is currently transcribed, as evidenced by the presence of a transcript in the cDNA library. Therefore, the large intron size does not seem to be a result of gene inactivation.

Such large introns have also been found in *P. abies* (Nystedt et al. 2013) and *P. taeda* (Wegrzyn et al. 2014), whose genome sequences have been determined recently. In the case of *P. taeda*, the largest intron was 319 kb in length. On the other hand, Yu et al. (2002) have reported that there are few introns of more than 10 kb in *A. thaliana* or in *O. sativa*, but there are many introns larger than 10 kb in the human genome (Yu et al. 2002), with the largest intron being approximately 500 kb (Sakharkar et al. 2004). It has been suggested that intron size is correlated with genome size across the broad phylogenetic groups (Deutsch and Long 1999; Vinogradov 1999; McLysaght et al. 2000). The large introns found in *C. japonica*, *P. abies*, and *P. taeda* indicate that the genome size and intron size are positively correlated in plants as well, and we may find more large introns in plants such as conifers that have large genome sizes.

Conclusion

The *C. japonica* genome consists mostly of diverse and divergent repetitive elements, similar to the genomes of *P. abies*, *P. taeda*, and *T. distichum*, indicating the ancient expansion of TEs. Additionally, there are some potentially active TEs, including some that are previously undescribed. Another feature of interest is the presence of a gene with a huge intron. Indeed, recent evidence suggests that conifer genomes may have multiple genes with very long introns, which differentiate them from other plant clades known to be dominated by small intron genes. Therefore, *C. japonica* genome shows features that are distinct from those of angiosperms. Although full-scale genome sequencing is still time- and resource-consuming in conifers, there is much more to learn by studying the structure of conifer genomes.

Acknowledgments We would like to thank Alfred E. Szmidt, Junko Kusumi, and two anonymous referees for constructive comments on earlier drafts of the manuscript. This study was partially supported by the Program for Promotion of Basic and Applied Researches for Innovations in Bio-oriented Industry and by Grants-in-Aid for Scientific Research from the Japan Society for the Promotion of Science (nos. 22370083 and 26291082).

Conflict of interest The authors declare that they have no conflict of interest.

Data archiving statement Nucleotide sequences were deposited with the DNA Data Bank of Japan (DDBJ).

References

Ahuja MR, Neale DB (2005) Evolution of genome size in conifers. Silvae Genet 54(3):126–137

Banks JA, Nishiyama T, Hasebe M, Bowman JL, Gribskov M et al (2011) The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. Science 20:960–963

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

Bird AP (1980) DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids Res 8:1499–1504

Birol I, Raymond A, Jackman SD, Pleasance S, Coope R et al (2013) Assembling the 20 Gb white spruce (*Picea glauca*) genome from whole-genome shotgun sequencing data. Bioinformatics 29:1492–1497

Chandler VL, Walbot V (1986) DNA modification of a maize transposable element correlates with loss of activity. Proc Natl Acad Sci 83: 1767–1771

Coulondre C, Miller JH, Farabaugh PJ, Gilbert W (1978) Molecular basis of base substitution hotspots in Escherichia coli. Nature 274:775–780

Deutsch M, Long M (1999) Intron-exon structures of eukaryotic model organisms. Nucleic Acids Res 27(15):3219–3228

Flavell RB (1984) Inactivation of gene expression in plants as a consequence of specific sequence duplication. Proc Natl Acad Sci 91: 3490–3496

Futamura N, Totoki Y, Toyoda A, Igasaki T, Nanjo T et al (2008) Characterization of expressed sequence tags from a full-length enriched cDNA library of *Cryptomeria japonica* male strobili. BMC Genomics 9:383

Gadek PA, Alpers DL, Heslewood MM, Quinn C (2000) Relationships within Cupressaceae sensu lato: a combined morphological and molecular approach. Am J Bot 87(10):1480–1488

Gao C, Xiao M, Ren X, Hayward A, Yin J, Wu L, Fu D, Li J (2012) Characterization and functional annotation of nested transposable elements in eukaryotic genomes. Genomics 100:222–230

Goff SA, Ricke D, Lan TH, Presting G, Wang R et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296:92–100

Gruenbaum Y, Naveh-Many T, Cedar H, Razin A (1981) Sequence specificity of methylation in higher plant DNA. Nature 292:860–862

Hamberger B, Hall D, Yuen M, Oddy C, Hamberger B et al (2009) Targeted isolation, sequence assembly and characterization of two white spruce (*Picea glauca*) BAC clones for terpenoid synthase and cytochrome P450 genes involved in conifer defense reveal insights into a conifer genome. BMC Plant Biol 9:106

Havecker ER, Gao X, Voytas DF (2004) The diversity of LTR retrotransposons. Genome Biol 5:225

Hizume M, Shibata F, Matsusaki Y, Garajova Z (2001) Chromosome identification and comparative karyotypic analyses of four Pinus species. Theor Appl Genet 105:491–497

Jurka J (1998) Repeats in genomic DNA: mining and meaning. Curr Opin Struct Biol 8:333–337

Jurka J (2000) Repbase update: a database and an electronic journal of repetitive elements. Trends Genet 9:418–420

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O et al (2005) Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110:462–467

Kado T, Yoshimaru H, Tsumura Y, Tachida H (2003) DNA variation in a conifer, *Cryptomeria japonica* (Cupressaceae sensu lato). Genetics 164(4):1547–1559

Keeling CI, Dullat HK, Yuen M, Ralph SG, Jancsik S et al (2010) Identification and functional characterization of monofunctional ent-copalyl diphosphate and ent-kaurene synthases in white spruce reveal different patterns for diterpene synthase evolution for primary and secondary metabolism in gymnosperms. Plant Physiol 152: 1197–1208

Kimura M (1980) A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kohany O, Gentles AJ, Hankus L, Jurka J (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinf 7:474

Kovach A, Wegrzyn JL, Parra G, Holt C, Bruening GE (2010) The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences. BMC Genomics 11:420

Kusumi J, Tsumura Y, Yoshimaru H, Tachida H (2000) Phylogenetic relationships in Taxodiaceae and Cupressaceae sensu stricto based on *matK* gene, *chlL* gene, *trnL-trnF* IGS region, and *trnL* intron sequences. Am J Bot 87(10):1480–1488

Lee E, Harris N, Gibson M, Chetty R, Lewis S (2009) Apollo: a community resource for genome annotation editing. Bioinformatics 25(14): 1836–1837

Leslie AB, Beaulieu JM, Rai HS, Crane PR, Donoghue MJ, Mathews S (2012) Hemisphere-scale differences in conifer evolutionary dynamics. Proc Natl Acad Sci U S A 109:1217–16221

Liu W, Thummasuwan S, Sehgal SK, Chouvarine P, Peterson DG (2011) Characterization of the genome of bald cypress. BMC Genomics 12: 553

Martienssen R (1998) Transposons, DNA methylation and gene control. Trends Genet 14:263–264

McLysaght A, Enright L, Skrabanek L, Wolfe KH (2000) Estimation of synteny conservation and genome compaction between pufferfish (Fugu) and human. Yeast 17:22–36

Meyer P, Niedenhof I, ten Lohuis M (1994) Evidence for cytosine methylation of non-symmetrical sequences in transgenic *Petunia hybrida*. EMBO 13:2084–2088

Misksche JP, Hotta Y (1973) DNA base composition and repetitious DNA in several conifers. Chromosoma 41:29–36

Miura A, Yonebayashio S, Watanabe K, Toyama T, Shimada H et al (2001) Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. Nature 411:212–214

Moriguchi Y, Ujino-Ihara T, Uchiyama K, Futamura N, Saitou M et al (2012) The construction of a high-density linkage map for identifying SNP markers that are tightly linked to a nuclear-recessive major gene for male sterility in *Cryptomeria japonica* D. Don. BMC Genomics 13:95

Moritsuka E, Hisataka Y, Tamura M, Uchiyama K, Watanabe A et al (2012) Extended linkage disequilibrium in noncoding regions in a conifer, *Cryptomeria japonica*. Genetics 190:1145–1148

Neal DB, Wegrzyn JL, Stevens KA, Zimi AV, Puiu D et al (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol 15:R59

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC et al (2013) The Norway spruce genome sequence and conifer genome evolution. Nature 497:579–584

Ohri D, Khoshoo TN (1986) Genome size in gymnosperms. Plant Syst Evol 153:119–132

Ohtsubo Y, Ikeda-Ohtsubo W, Nagata Y, Tsuda M (2008) GenomeMatcher: a graphical user interface for DNA sequence comparison. BMC Bioinf 9:376

Perlman PS, Boeke JD (2004) Ring around the retroelement. Science 2004(9):182–184

Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. Bioinformatics 21:I351–I358

Razin A (1998) CpG methylation, chromatin structure and gene silencing—a three-way connection. EMBO 17:4905–4908

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A et al (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science 319:64

RepeatMasker (2013) Available at: http://www.repeatmasker.org/

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4(4):406–425

Sakharkar MK, Chow VT, Kangueane P (2004) Distributions of exons and introns in the human genome. In Silico Biol 4:387–393

Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T (2010) Genome sequence of the palaeopolyploid soybean. Nature 463:178–183

Suzuki MM, Bird A (2008) DNA methylation landscapes: provocative insights from epigenomics. Nat Rev Genet 9:465–476

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance and maximum parsimony methods. Mol Biol Evol 28:2731–2739

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815

Thompson JD, Higgins DG, Gibson TJ (1994) ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Tsumura Y, Uchiyama K, Moriguchi Y, Ueno S, Ihara-Ujino T (2012) Genome scanning for detecting adaptive genes along environmental gradients in the Japanese conifer, Cryptomeria japonica. Heredity 109:349–360

Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604

Vinogradov AE (1999) Intron-genome size relationship on a large evolutionary scale. J Mol Evol 49:376–384

Wegrzyn J, Liechty J, Stevens K, Wu L-S, Loopstra C, Vasquez-Gross H, Dougherty W, Lin B, Zieve J, Martínez-García P, et al. (2014) Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. Genetics 196:891–909

Willyard A, Syring J, Gernandt DS, Liston A, Cronn R (2006) Fossil calibration of molecular divergence infers a moderate mutation rate and recent radiations for *Pinus*. Mol Biol Evol 24:90–101

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35:W265–W268

Yu J, Hu S, Wang J, Wong GKS, Li S et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 2002(296):79–92