J. Plant Biol. (2009) 52:616–624
DOI 10.1007/s12374-009-9080-4

ORIGINAL RESEARCH

Evolutionary Process of the Genomic Sequence Around the 100 Map Unit of Chromosome 1 in *Arabidopsis thaliana*

Atsushi Kato · Hiroaki Kato · Takuhiro Shida · Tamao Saito · Yoshibumi Komeda

Received: 30 July 2009 / Revised: 9 October 2009 / Accepted: 14 October 2009 / Published online: 31 October 2009

Abstract Comparative analysis of nucleotide sequences of the genomic region located around 100 map unit of chromosome 1 using two accessions, Columbia (Col) and Landsberg *erecta* (L*er*), of *Arabidopsis thaliana* was performed. High divergence was detected between them, and the length of the L*er* sequence was half of corresponding sequence of Col. This divergence occurred by tandem duplication, deletion of large regions, and insertion of unrelated sequences. These events led to the high polymorphism of plant disease resistant genes, which are located in the analyzed region. It is highly probable that two-round duplication occurred, and the insertion sequences are transposable elements. The data suggest that the analyzed region had been evolving until quite recently.

Keywords *Arabidopsis* · Gene duplication · Repeated sequence · *R* gene · Transposable element

A. Kato (✉) · H. Kato · T. Shida
Department of Biological Sciences,
Graduate School of Science, Hokkaido University,
Sapporo 060-0810, Japan
e-mail: atsushi@sci.hokudai.ac.jp

T. Saito
Department of Materials and Life Sciences,
Faculty of Science and Technology, Sophia University,
Chiyoda-ku, Tokyo 102-8554, Japan

Y. Komeda
Department of Biological Sciences,
Graduate School of Science, The University of Tokyo,
Hongo, Tokyo 113-0033, Japan

Introduction

The structure of genetic polymorphism in a genome is influenced by different evolutionary processes. Natural variation within the species has been the important subject of evolutionary genetics. The model plant *Arabidopsis thaliana* is a useful species for investigating natural variation because of its small genome, short generation, and adaptation to various climates. Hundreds of accessions have been collected from diverse worldwide locations and placed in public stock centers (Koornneef et al. 2004). Recently, many studies on natural variation of *A. thaliana* have been carried out (Maloof 2003; Weigel and Nordborg 2005; Mitchell-Olds and Schmitt 2006; Shindo et al. 2007), and the influence of genetic variation on phenotypes has been analyzed (Ungerer et al. 2003; Juenger et al. 2005; de Meaux et al. 2005; Shindo et al. 2005; DeCook et al. 2006; Balasubramanian et al. 2006; Briggs et al. 2006).

Within species, the level of genetic variation is generally low due to frequent genetic exchange among individuals and relatively short time to the common ancestor between alleles. However, strikingly high levels of polymorphism within species at some loci have been reported. In plants, plant disease resistance (*R*) genes are typical examples (Ellis et al. 2000; Meyers et al. 2005). Complete determination of the nucleotide sequence in the *A. thaliana* accession Columbia (Col) genome revealed that this genome contains about 200 genes that encode proteins with similarity to the domains characteristic of plant disease resistant proteins. Their genes are arranged as single genes and as clustered loci. Analyses of the evolution of *R* genes in *A. thaliana* have been performed by comparison of intra-accessions and inter-accessions. For example, results of phylogenic analyses and genome distribution of 149 NBS-LRR-encoding genes in *A. thaliana* Col have been reported

(Meyers et al. 2003). Comparative analysis of the *RPP5* locus was performed using two accessions, Col and Landsberg *erecta* (L*er*; Noël et al. 1999), and the same analysis was performed in the *RPP8* locus (McDowell et al. 1998). Recently, many accessions have been used to analyze the evolution of *R* genes (Xiao et al. 2004; Shen et al. 2006; Bakker et al. 2006).

Mechanistic investigation of genomic change is necessary for the understanding of the genome evolution. Such investigation will also lead to a better understanding of environmental adaptation and evolution of plants. It is thought that dynamic change in genomes occurs largely by gene duplication followed by recombinational events (Blanc et al. 2000; Leister 2004; Sampedro et al. 2005; Kong et al. 2007) and transposable elements, especially retrotransposon, related to the genome evolution (Zhang and Wessler 2004; Bennetzen et al. 2005; Wang et al. 2006).

We previously analyzed the genome organization of the *A. thaliana* Col genomic region located around the 100-map unit of chromosome 1 (Kato et al. 1999). Determination of the nucleotide sequence showed that repeated sequences exist in this region and that their sequences are highly conserved among repeated units. One repeated unit contains two disease resistant protein genes, one retrotransposon (Kato et al. 1999; Kuwahara et al. 2000), and five other genes. Southern hybridization analyses of Col DNA and L*er* DNA using the inner-sequence and outer-sequence of this repeated region as probes suggested that the outer-sequence of the repeated region was conserved in both accessions, but the major part of this repeated sequence was deleted in the L*er* genome. This result suggests that this region has a striking difference in genome organization within *A. thaliana* species. We compared the genome sequences of this region using two accessions, Col and L*er*, as the first step to analyze genome evolution in *A. thaliana*. These accessions may be commonly used to cross for gene mapping. We therefore chose these accessions.

In this study, we cloned the L*er* sequence corresponding to the repeated region and its flanking regions in the Col genome and compared their sequences. The result reveals that there are large deleted regions in the L*er* genome, which correspond to the repeated region and non-repeated region in the Col genome. These changes have resulted in high polymorphism of *R* genes. Transposable element-like sequences are also related to the genome variation.

Materials and Methods

Construction and Screening of a Genomic Library

Total DNA was extracted from seedlings of *A. thaliana* accession Landsberg *erecta* according to the method of (Bedbrook et al. 1980). Genomic libraries were produced using Lambda GEM 12 arms (Promega) as suggested by the manufacturer. GigaPack XL (Stratagene) was used for packaging reaction. Screening of a genomic library was performed by plaque hybridization according to standard procedures (Sambrook et al. 1989). The inserted DNAs of cloned phages were amplified by PCR reaction. These amplified DNA fragments were used to construct a physical map and determined the end sequences.

PCR Reaction

PCR reactions were performed using LA-Taq (TAKARA). The reaction conditions were in accordance with the manufacturer's recommendations. PCR products were purified by electrophoresis in agarose gel and subjected to further analyses.

Sequence Analyses

Cloned lambda phages were propagated by the plate lysate method (Sambrook et al. 1989). Phage DNAs were isolated from plate lysate using a Wizard Lambda Preps DNA Purification System (Promega) as suggested by the manufacturer. Phage DNAs were digested with *Eco*RI or *Bam*HI, and then they were recloned to plasmid vectors. These plasmids were subjected to DNA sequencing. The primer-walking method was used to decide the whole sequences of inserted DNAs in plasmids. The DNA sequences were determined using an ABI PRISM 377 DNA sequencer (Applied Biosystems). Nucleotide sequence data have been reported in the DDBJ/EMBL/GenBank databases under the accession numbers AB425270, AB425271, AB425272, AB425273, AB425274, and AB425275. The sequences were analyzed using a sequence analysis tool on the WWW sites of DDBJ (http://www.ddbj.nig.ac.jp) and TAIR (http://www.Arabidopsis.org). DNA sequence analysis software, DNASIS (TAKARA), was also used for detailed analysis.

The procedure to detect repeated regions in the Col genome and L*er* genome was as follows. First, we sequentially divided the determined sequences (approximately 212.3 kb in Col and approximately 77.6 kb in L*er*) into 10-kb sequences, and a search was carried out for homologous sequences with each 10-kb sequence using BLAST search. We used the WWW site of TAIR to analyze the Col sequence and the WWW site of DDBJ to analyze the L*er* sequence. When a homologous region was detected in a certain 10-kb sequence, a search was carried out again for a homologous sequence with the neighboring sequence of this homologous region. This process was repeated to determine the boundary between the homologous region and non-homologous region. After determination of the homologous region, we calculated the sequence identity

using Smith–Waterman search in DNASIS. When the homologous region was long, we calculated the sequence identity after dividing the homologous region into 3-kb sequences.

To identify the putative genes in the L*er* genome, we searched for homologous sequences to annotated genes in the Col genome by BLAST search using DNASIS. When a homologous region was found, a search was carried out for open reading frames in this region and its neighboring region using DNASIS. After the putative amino acid sequence homology was calculated by Smith–Waterman search using DNASIS, the putative coding region was determined. We also used the GENSCAN program in the WWW site of MIT (http://genes.mit.edu/GENSCAN.html).

To find orthologous regions from the Col and L*er* genome, we used BLAST search and Smith–Waterman search in the WWW site of TAIR and DNASIS. We fundamentally performed the same procedure as that used for searching for repeated regions.

Results

Repeated Sequence in Col Genome

Previously, we determined the nucleotide sequence of the Col genome around the 100 map unit of chromosome 1 using P1 clones X7J (accession No. AB077822) and R18I (accession No. AB078516). These results and other sequence data of BAC clone T4M14 (accession No. AC027036) reported by Town et al. revealed that large repeated sequences exist in this region (Fig. 1a). Table 1 represents the detailed information of this repeated region. This repeated sequence was 89206 nucleotides in length and constructed from two complete repeated units (unit 2 and 3) and one truncated unit (unit 1). The sequence of a repeated unit was able to divide into five sequences, named A, B, C, D, and E. Two complete units, repeated units 2 and 3, have 100% sequence identity in their whole regions. The truncated unit, repeated unit 1, contained sequences B, D, and E. Sequence E in repeated unit 1 was also 100% identical to that of repeated unit 2. However, the sequences of B and D were divergent between repeated units 1 and 2.

Sequence Determination of L*er* DNA

In order to analyze the variation of genome organization around the 100-map unit of chromosome 1 between L*er* and Col, a contiguous DNA fragment from L*er* was generated. It was established with cloned DNAs using λ-phage and PCR fragment (Fig. 1b). First, the genomic library of L*er* was screened with the Col sequences located outside of the repeated region as probes (Fig. 1a). Physical mapping and end sequence determination showed that one phage clone (No. 14) has a sequence corresponding to the upstream sequence of the repeated region, and two phage clones (No. 1 and No. 18) have a sequence corresponding to the downstream sequence of the repeated region. Next, the same library was screened with the end sequences of phage clone No. 14 and No. 1 as probes. The same analyses showed that one phage clone (No. 2–9) overlapped phage No. 14, and another phage clone (No. 2–7) overlapped phage No. 1. The inserted DNAs of these five phage clones were recloned in a plasmid vector, and their sequences were determined. However, the sequences of phage No. 2–9 and No. 2–7 were not overlapped. In order to obtain the DNA fragment of this gapped region, PCR was performed, and the sequence of the PCR product was determined. Finally, as shown in Fig. 1b, contiguous DNA was generated. This sequence was 70,716 bp and corresponded to the sequence

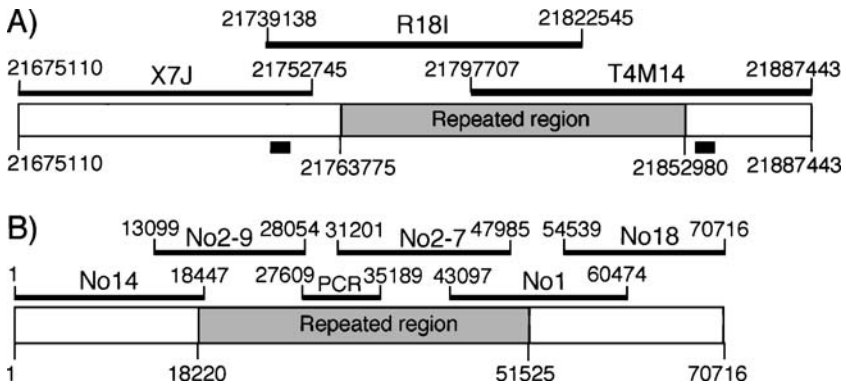

Fig. 1 **a** Contiguous DNA of Col generated by using two P1 clones and one BAC clone. Accession numbers of clones are as follows: X7J, AB077822; R18I, AB078516; and T4M14, AC027036. *Black boxes* indicate the regions using probes for phage screening. One is from 21743236 to 21746767 and the other is from 21853002 to 21856517. Nucleotide numbers are relative to whole sequence of chromosome 1. **b** Contiguous DNA of L*er* generated by using five λ-phage clones and one PCR product. Accession numbers of them are as follows: No14, AB425270; No2-9, AB425271; No2-7, AB425273; No1, AB425274; No18, AB425275; and PCR product, AB425272. Nucleotide numbers are relative to one end of the contiguous DNA

Table 1 Repeated sequences found in the Col genome

Sequence name	Location of repeat unit 1	Identity between repeated units 1 and 2 (%)	Location of repeat unit 2	Identity between repeated units 2 and 3 (%)	Location of repeat unit 3
A			52430–67706	100	88688–103964
B	35740–41031	80	67707–72565	100	103965–108823
C			72566–77384	100	108824–113642
D	41032–42163	87	77385–78421	100	113643–114679
E	42164–52429	100	78422–88687	100	114680–124945

Locations represent by nucleotide numbers used in Fig. 2

from 21728036 to 21864589 of chromosome 1 in the Col genome.

Sequence Organization of L*er* Genome

In order to characterize the L*er* sequence, putative genes and repeated sequences were examined. Detection of sequence similarity to annotated genes in the corresponding region of the Col genome revealed that 11 putative genes exist in the L*er* sequence (Fig. 2 and Table 2). We also searched for putative genes using the GENSCAN program. No additional genes were predicted, although different exon–intron structures were revealed in some genes. In these cases, we selected the exon–intron structure to obtain the best matching between amino acid sequences of the Col gene and L*er* gene.

Six genes (ORFs 1, 2, 3, 9, 10, and 11) were orthologous genes located outside the repeated region of the Col genome, since a part of ORF 9 was located in the repeated region. They were highly conserved between the two accessions. Three genes (ORFs 6, 7, and 8) existed as single-copy genes in the L*er* genome, but their orthologous genes were located in the repeated region of the Col genome. These genes were also highly conserved between the two accessions. Two genes (ORFs 4 and 5) coded putative disease resistance proteins, which were similar to At1g58848 of Col. The identities of amino acid sequences of ORF 4 and ORF 5 to At1g58848 were 87% and 69%,

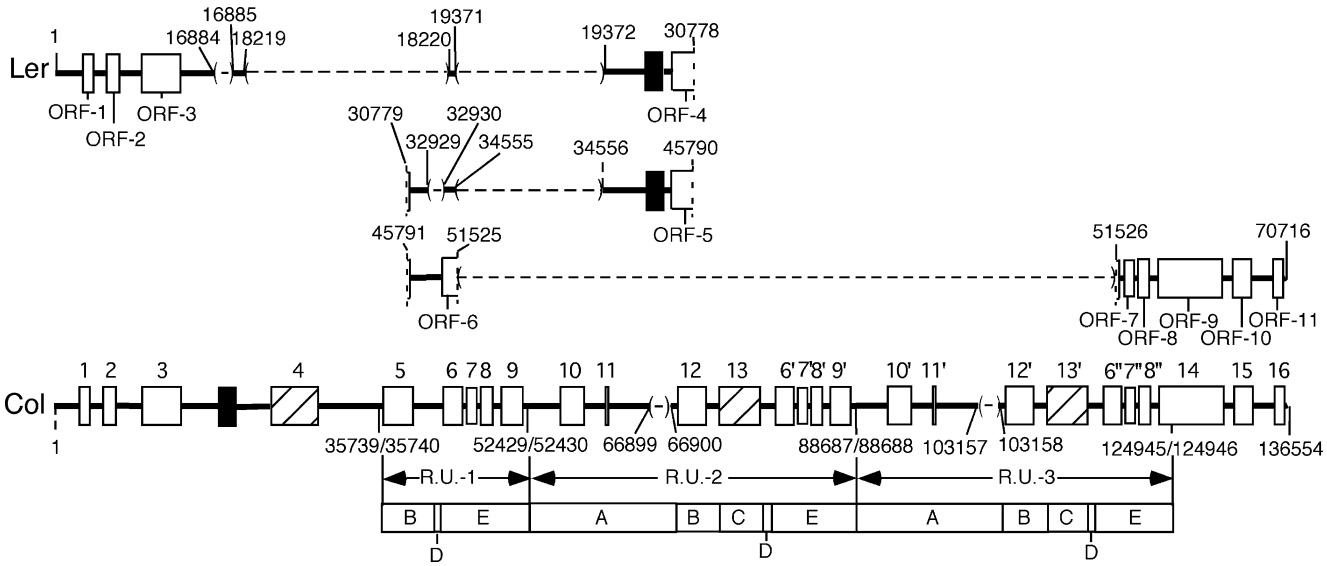

Fig. 2 Schematic overview of comparison of L*er* sequence and Col sequence. Similar sequences are arranged lengthwise. *Dotted lines in brackets* indicated the deleted regions for arrangement. The Col sequence corresponds to the region from 21728036 to 21864589 in chromosome 1. *White boxes* and *black boxes* in the L*er* sequence indicate putative genes and transposable element-like sequence, respectively. ORF numbers correspond to these in Table 2. *White boxes with numbers*, *hatched boxes with numbers*, and a *black box* in the Col genome indicate the annotated genes, *copia*-like elements, and transposable element-like sequence, respectively. Locus names of annotated genes and *copia*-like elements are as follows: *1*, At1g58470; *2*, At1g58480; *3*, At1g58520, *4*, At1g58561; *5*, At1g58602; *6*, At1g58643; *7*, At1g58684; *8*, At1g58725; *9*, At1g58766; *10*, At1g58807; *11*, identical to At1g59171, though this is not annotated in this region; *12*, At1g58848; *13*, At1g58889; *6′*, At1g58936; *7′*, At1g58983; *8′*, At1g59030; *9′*, At1g59077; *10′*, At1g59124; *11′*, At1g59171; *12′*, At1g59218; *13′*, At1g59265; *6″*, At1g59312; *7″*, At1g59359; *8″*, At1g59406; *14*, At1g59453; *15*, At1g59500; and *16*, At1g59510. *R.U.-1*, *2*, and *3* indicate repeated units of the Col sequence, and *A*, *B*, *C*, *D*, and *E* correspond to sequence names in Table 1

Table 2 A list of putative genes in the L*er* genome

	Location (initiation-stop codon)	CDS size (bp)	ORF size (a.a)	Description	Locus name in Col genome	Nucleotide sequence identity with Col gene (%)	Amino acid sequence identity with Col protein (%)
ORF1	3086–4248	1,080	359	RNA binding protein	At1g58470	98	97
ORF2	5619–7049	1,017	338	GDSL-motif lipase	At1g58480	93	93
ORF3	8923–13278	1,974	657	Unknown protein	At1g58520	99	99
ORF4	27698–31022	3,159	1,052	Disease resistance protein	At1g58848	92	87
ORF5	42906–46031	2,439	812	Disease resistance protein	At1g58848	73	69
ORF6	49648–51730	1,428	475	Unknown protein	At1g59312	98	98
OFR7	52304–53500	855	284	Ribosomal protein S2	At1g59359	97	100
ORF8	54223–55537	936	311	Carboxylic ester hydrolase	At1g59406	98	100
ORF9	56613–63113	5,088	1,695	Transcriptional factor	At1g59453	96	95
ORF10	64268–66394	1,794	597	IAA-amido synthetase	At1g59500	97	97
ORF11	69560–70705	1,146	381	Unknown protein	At1g59510	99	99

Locations represent by nucleotide numbers used in Fig. 2

respectively. These values were lower than those of other ORFs.

A repeated structure was also found in the L*er* genome (Fig. 1b and Table 3). This region was 33306 nucleotides in length and constructed from three repeated units. The first unit was 1,152 bp from 18220 to 19371, the second unit was 15,184 bp from 19372 to 34555, and the third unit was 16,970 bp from 34556 to 51525. The first unit has only a small portion of the 3′ sequence of the repeated unit. An extra sequence from 47935 to 49902, the length of which is 1,968 bp, existed within only the third unit. The first unit has 91% sequence identity to the second and third units. The whole sequence identity between the second and third units except the 1,968-bp inserted sequence was 97%.

Sequence Comparison Between L*er* and Col

The determined sequence of L*er* in this study was compared to the Col sequence in the corresponding region. The regions of the Col sequence most similar to L*er* sequence are listed in Table 4. Their sequence identities were from 90% to 99%. Figure 2 shows a diagrammatic representation of the arrangement of both sequences. Repeated sequences B and D in the Col genome were divergent between repeated units 1 and 2. L*er* sequences from 30779 to 34555 and from 45791 to 51525 were more similar to the Col sequences in repeated unit 1 than those in repeated unit 2.

The length of the L*er* sequence was about half of that of the Col sequence. One reason for the difference in the sequence lengths of the two accessions is the difference in lengths of repeated units. The lengths of repeated units were about 15 kb in the L*er* genome and about 36 kb in the Col genome. An approximately 9.3-kb sequence at the 5′-side and a 4.8-kb internal sequence of the Col repeated unit were deleted in the L*er* sequence. An approximately 7.5-kb sequence at the 3′-side of the Col repeated unit exists as a single-copy sequence in L*er*. Another reason for the difference is a large deletion of the L*er* sequence, the regions of which corresponded to the Col sequences from 17413 to19833 and from 21361 to 35738. It seemed that one of these deletions was related to transposable elements. The Col sequence from 17413 to 19824 was not annotated, and a significant open reading frame was not observed. However, this 2,412-bp sequence was flanked by an identical sequence, TTATTTTAA, and only one TTATTTTAA

Table 3 Repeated sequences found in the L*er* genome

Location of repeat unit 1	Identity between repeated units 1 and 2 (%)	Location of repeat unit 2	Identity between repeated units 2 and 3 (%)	Location of repeat unit 3
		19372–32929	97	34556–47934
				47935–49902
		32930–33416	100	49903–50386
18220–19371	91	33417–34555	99	50387–51525

Locations represent by nucleotide numbers used in Fig. 2

Table 4 Comparison of orthologous regions from the Col and L*er* genome

Sequences of Ler genome	Identity (%)	Sequences of Col genome
1–16884	97	1–17412
16885–18219	94	19834–21360
18220–19371	90	43740–44881 (79998–81139, 116256–117397)
19372–24520	99	61742–66899 (98000–103157)
24521–26832		
26833–30778	94	66900–70844 (103158–107102)
30779–32929	95	39155–41310
32930–34555	98	43256–44881 (79514–81139, 115722–117397)
34556–39712	99	61742–66899 (98000–103157)
39713–42040		
42041–45790	92	66900–70709 (103158–106967)
45791–51525	96	39155–44881
51526–59048	94	44882–52429 (117398–124945)
59049–70716	97	124946–136554

Sequences represent by nucleotide numbers used in Fig. 2

sequence was found in the corresponding region in the L*er* sequence (Fig. 3a). These structures suggested that this 2,412-bp sequence was inserted like a transposable element and that the sequence of TTATTTTAA was a target-duplicated sequence.

As shown in Fig. 2, another sequence like a transposable element was found in the repeated sequence of L*er*. The sequences from 24521 to 26823 and from 39713 to 42031 were flanked by an identical sequence, TATTTTATT, and only one TATTTTATT sequence was found in the corresponding region in the Col sequence (Fig. 3b). A similar sequence was found in another locus in the Col genome. The sequence from 18640728 to 18643049 of chromosome 3 in Col has 96% and 97% identity to the sequences from 24521 to 26823 and from 39713 to 42031 of L*er*, respectively. The sequence in Col was flanked by a completely identical sequence, TTTTATTTA (Fig. 3b). These results suggested that these sequences were also transposable elements.

Insertion of a *copia*-like element was also found in the Col genome. Sequence C in the repeated sequence in the Col genome has characters of a *copia*-like retrotransposon, and we named it *AtRE1* (Kuwahara et al. 2000). This sequence was found in repeated units 2 and 3 in the Col sequence, but it was not found in the corresponding regions in the L*er* sequence and repeated unit 1 in the Col sequence. *AtRE1* was flanked by an identical sequence, AATAC, and only one similar sequence (AAGAC) was found in the L*er* sequence and repeated unit 1 in the Col sequence. These structures suggested that *AtRE1* was

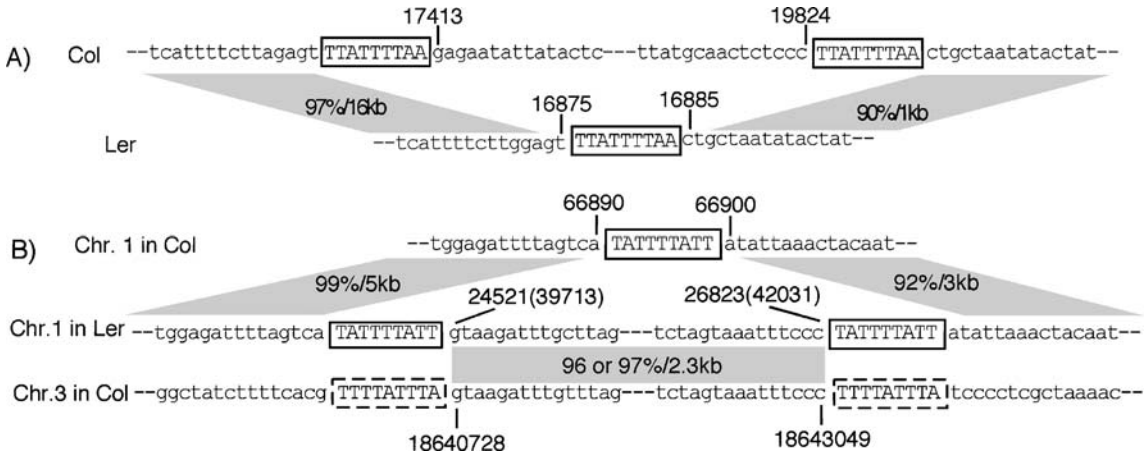

Fig. 3 **a** Sequence from 17413 to 19824 in Col was flanked by identical sequences, TTATTTTAA. This sequence and one TTATTTTAA sequence were deleted in L*er*. **b** Sequences from 24521 to 26823 and from 39713 to 42031 in L*er* were flanked by identical sequences, TATTTATT. This sequence and one TATTTTATT sequence were deleted in the corresponding region in Col chromosome 1. The sequence from 18640728 to 18643049 in Col chromosome 3 has 96% and 97% identities to those from 24521 to 26823 and from 39713 to 42031 in L*er*, respectively, and is flanked by the sequences TTTTATTTA. *Shadows* represent the regions found sequence similarities, and the length of comparing sequence (kb) and the homology (percent) are indicated

inserted in repeated units 2 and 3 and that the sequence of AATAC was a target-duplicated sequence.

Discussion

In this study, we compared the genome organizations of two accessions of *A. thaliana*, Columbia, and Landsberg *erecta*. The analyzed regions were about 135 kb in Col and 70 kb in L*er*, and dynamic change in genome organization was revealed. Figure 4 shows one of the possible processes of genome evolution in Col and L*er*. The common ancestral sequence may contain A, B, D, and E sequences. The evolutionary process of the Col sequence was thought to be as follows. First, the regions containing sequences A, B, D, and E were duplicated, and repeated units 1 and 2 were generated. Then, sequence A in repeated unit 1 was deleted, and sequence C was inserted between sequence B and sequence D in repeated unit 2. Next, the duplication of repeated unit 2 generated repeated unit 3, and a transposable element-like sequence was inserted outside of the repeating region. On the other hand, the evolutionary process of the L*er* sequence was thought to be as follows. First, one third of the 3′-portion in sequence A (3′A in Fig. 4), entire sequences B and D, and one fourth of the 5′-portion of sequence E (5′E in Fig. 4) were duplicated, and a transposable element-like sequence was inserted in sequence 3′A. Next, the region from sequence 5′E to sequence D was duplicated. Finally, a large region containing the flanking region of sequence A, sequences A, B and D and a part of sequence 5'E was deleted, and a part of sequences D and E in repeated unit 2 was also deleted.

These processes revealed that three mechanisms were involved in sequence evolution. One mechanism was tandem duplication, and two-round duplication occurred. It seemed that different regions were duplicated in Col and L*er* from the common ancestral sequence. It is possible that the first duplication occurred in the same region in Col and L*er*, and a part of the middle region of the generated sequence was deleted in L*er*. In either case, the resultant repeated unit in L*er* was different from that in Col, and the sequences of the second duplication were different from each other. The difference in duplicated regions caused high divergence of sequences between Col and L*er*. For example, it caused the difference in number of *R* genes found in this region. The Col sequence contains five *R* genes (5, 10, 12, 10′, and 12′ in Fig. 2), and the L*er* sequence contains two *R* genes (ORF-4 and ORF-5 in Fig. 2). It was reported that the tandem duplication and following recombination caused polymorphisms of several loci of the *R* gene (Mitchell-Olds and Schmitt 2006; Shindo et al. 2007). A difference between gene numbers in the Col genome and L*er* genome has been reported in the *RPP8* locus. In this case, the L*er* genome contains the functional gene and a nonfunctional homolog. The Col genome contains only one gene, which is a chimera, related to the two L*er* genes. The Col gene appears to be derived from an unequal crossover within the two L*er* genes (McDowell et al. 1998). However, it is unlikely that the same process occurred in our case, because the internal sequence of the

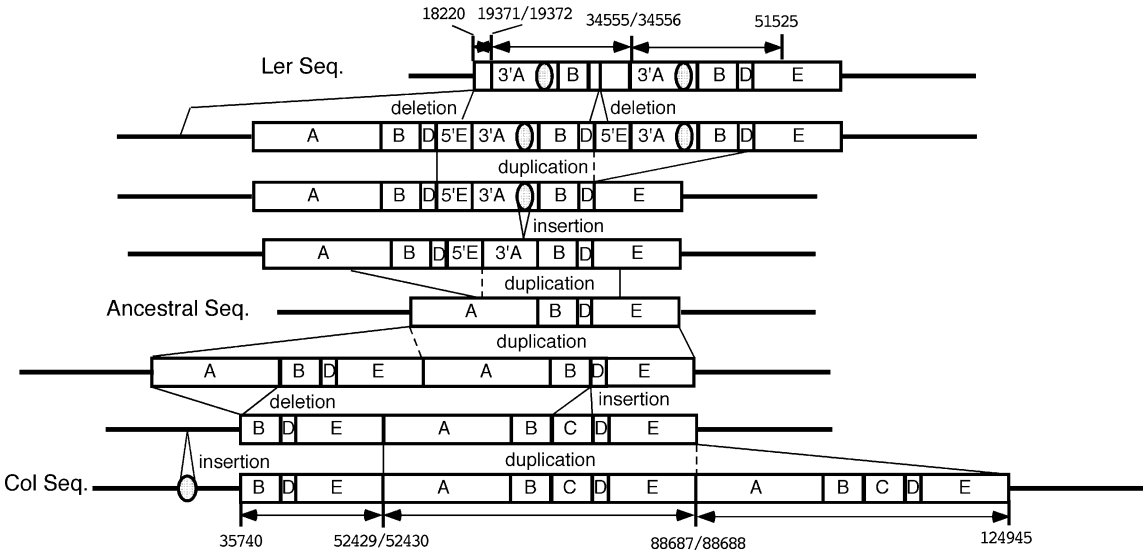

Fig. 4 Possible process of sequence evolution in the Col and L*er* genomes. Ancestral Seq., Col Seq., and Ler Seq. show the common ancestral genome organization, present genome organization in Col and that in L*er*, respectively. *A*, *B*, *C*, *D*, and *E* indicate the sequence names in the repeated unit. 3′A and 5′E indicate the 3′-portion of sequence A and 5′-portion of sequence E, respectively. *Shadowed circles* indicate transposable-like elements. *Arrows* and *numerals above Ler Seq. and under Col Seq.* indicate the repeated units found to be present in the L*er* and Col genomes, respectively. *Numerals* correspond to the nucleotide numbers in Table 3 (Ler Seq.) and Table 1 (Col Seq.)

two *R* genes located in the Col genome remained in the L*er* sequence.

The second mechanism involved in sequence evolution is deletion of large regions. It seemed that an approximately 22-kb sequence (from 21361 to 43739 in the Col sequence) was deleted in L*er* from a comparison of the present sequences. However, the putative evolution process suggests that an approximately 15-kb sequence was deleted in Col and an approximately 37-kb sequence was deleted in L*er*. Another deleted sequence with a length of about 2-kb was also detected in L*er* genome. The mechanism of these deletions is unknown. Evidence for homologous recombination was not found in the deleted sequence and its flanking sequence.

The third mechanism involved in sequence evolution is insertion of several elements. The putative evolution process suggests that an approximately 5-kb sequence corresponding sequence C and an approximately 2.4-kb sequence were inserted in Col genome, and an approximately 2.3-kb sequence was inserted in L*er* genome. The 5-kb sequence corresponds to a *copia*-like element, named *AtRE1*, found in another region of the L*er* genome (Kuwahara et al. 2000). The *AtRE1* sequence was found in only repeated units 2 and 3 in the Col genome, and its orthologous gene was found in only one locus in the L*er* genome. Moreover, comparison of flanking sequences of *AtRE1* in the L*er* genome with the whole sequence of the Col genome in the TAIR database suggested that *AtRE1* is located in chromosome 4 in the L*er* genome. Generally, retrotransposons including *copia*-type elements are inserted at a second site as a copy, and the original sequence remains. However, the fact that *AtRE1* did not exist at the same locus in the two accessions suggests that excision, which may be a result of homologous recombination in its flanking sequence, occurred. Similar events occurred in *AtRE2*, which is another *copia*-like element. The locations of *AtRE2* in the Col and L*er* genomes were different (Kuwahara et al. 2000). Deletion of a retrotransposon may frequently occur in *A. thaliana*. Although the original locus of *AtRE1* has not been identified, it is highly probable that *AtRE1* located in the repeated region in Col was inserted after Col and L*er* diverged, since both sequences of long terminal repeats and target-duplicated sequences were completely conserved (Kuwahara et al. 2000). On the other hand, the short tandem duplication found on both sides of the 2.4- and 2.3-kb inserted sequences suggests that they are transposable elements. Although the 2.4-kb inserted sequence in Col has no similarity with reported transposable elements, the 2.3-kb inserted sequence found in L*er* is partially similar to At1g41930, which was reported as one of the CACTA family transposons (Miura et al. 2001). One third of the 3′ portion of the inserted sequence has 90% identity with the sequence of the middle region of At1g41930. The relation between the inserted sequences found in this study and CACTA family transposons is unknown. However, this sequence similarity also suggested that the 2.3-kb insertion sequence identified in this study is a transposable element. Transposable elements can play an important role in the evolution of a plant genome (Zhang and Wessler 2004; Bennetzen et al. 2005; Wang et al. 2006). Comparison of the *RPP5* locus between Col and L*er* revealed that this highly diverged region contains many transposable elements (Noël et al. 1999). Analyses of the *SKP1* gene family showed that retroposition plays an important role in the evolution of a plant gene family (Kong et al. 2007). Sequences similar to CACTA1 copies were distributed among 19 accessions and showed high degrees of polymorphism in genomic localization (Miura et al. 2004).

Many studies have revealed that gene or segment duplication is the main force in genome evolution. It has been reported that duplication regions cover 70% of the *Arabidopsis* genome (Blanc et al. 2003) and that the genome has undergone multiple rounds of entire genome duplication (Simillion et al. 2002). After a duplication event, rearrangements by deletion, insertion, inversion, and translocation occurred, and the present genome was generated (Blanc et al. 2000). These results were obtained from analyses of the whole genome of one accession, Columbia. On the other hand, results of studies in which a part of the genome containing specific genes was analyzed in various accessions have also been reported (McDowell et al. 1998; Noël et al. 1999; Xiao et al. 2004; Bakker et al. 2006; Shen et al. 2006; Kong et al. 2007). These results showed that dynamic change except for point mutations play an important role in sequence divergence among many accessions.

In this study, repeated units in which sequences were completely conserved were detected, and it was found that a short duplicated sequence generated by insertion was also completely conserved. These findings suggest that the region analyzed in this study had been evolving until quite recently. Further analyses of this region in various accessions will clarify the evolutionary process of *A. thaliana* genome and the relationships of various *A. thaliana* accessions.

Acknowledgement A part of this work was supported by grants-in-aid for scientific research from the Japanese Ministry of Education, Science, and Culture.

References

Bakker EG, Toomajian C, Kreitman M, Bergelson J (2006) A genome-wide survey of *R* gene polymorphisms in *Arabidopsis*. Plant Cell 18:1803–1818

Balasubramanian S, Sureshkumar S, Agrawal M, Michael TP, Wessinger C, Maloof JN, Clark R, Warthmann N, Chory J, Weigel D (2006) The PHYTOCHROME C photoreceptor gene mediates natural variation in flowering and growth responses of *Arabidopsis thaliana*. Nat genet 38:711–715

Bedbrook JR, Jhones J, O'Dell M, Thompson RD, Flavell RB (1980) A molecular description of telomeric heterochromatin in *Secale* sp. Cell 19:545–560

Bennetzen JL, Ma J, Devos KM (2005) Mechanisms of recent genome size variation in flowering plants. Ann Bot 95:127–132

Blanc G, Barakat A, Guyot R, Cooke R, Delseny M (2000) Extensive duplication and reshuffling in the *Arabidopsis* genome. Plant Cell 12:1093–1101

Blanc G, Hokamp K, Wolfe KH (2003) A recent polyploidy superimposed on older large-scale duplications in the *Arabidopsis* genome. Genome Res 13:137–144

Briggs GC, Osmont KS, Shindo C, Sibout R, Hardtke CS (2006) Unequal genetic redundancies in *Arabidopsis*—a neglected phenomenon? Trends Plant Sci 11:492–498

de Meaux J, Goebel U, Pop A, Michell-Olds T (2005) Allele-specific assay reveals functional variation in the *Chalcone synthase* promoter of *Arabidopsis thaliana* that is compatible with neutral evolution. Plant Cell 17:676–690

DeCook R, Lall S, Nettleton D, Howell SH (2006) Genetic regulation of gene expression during shoot development in *Arabidopsis*. Genetics 172:1155–1164

Ellis J, Dodds P, Pryor T (2000) Structure, function and evolution of plant disease resistance genes. Curr Opin Plant Biol 3:278–284

Juenger TE, Sen S, Stowe KA, Simms EL (2005) Epistasis and genotype-environment interaction for quantitative trait loci affecting flowering time in *Arabidopsis thaliana*. Genetica 123:87–105

Kato A, Suzuki M, Kuwahara A, Ooe H, Higano-Inaba K, Komeda Y (1999) Isolation and analysis of cDNA within a 300 kb *Arabidopsis thaliana* genomic region located around the 100 map unit of chromosome 1. Gene 239:309–316

Kong H, Landherr LL, Frohlich MW, Leebens-Mack J, Ma H, dePamphilis CW (2007) Patterns of gene duplication in the plant *SKP1* gene family in angiosperms: evidence for multiple mechanisms of rapid gene birth. Plant J 50:873–885

Koornneef M, Alonso-Blanco C, Vreugdenhil D (2004) Naturally occurring genetic variation in *Arabidopsis thaliana*. Annu Rev Plant Biol 55:141–172

Kuwahara A, Kato A, Komeda Y (2000) Isolation and characterization of *copia*-type retrotransposons in *Arabidopsis thaliana*. Gene 244:127–136

Leister D (2004) Tandem and segmental gene duplication and recombination in the evolution of plant disease resistance genes. Trends Genet 20:116–122

Maloof JN (2003) Genomic approaches to analyzing natural variation in *Arabidopsis thaliana*. Curr Opin Genet Dev 13:576–582

McDowell JM, Dhandaydham M, Long TA, Aarts MGM, Goff S, Holub EB, Dangl JL (1998) Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the *RPP8* locus of *Arabidopsis*. Plant Cell 10:1861–1874

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW (2003) Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. Plant Cell 15:809–834

Meyers BC, Kaushik S, Nandety RS (2005) Evolving disease resistance genes. Curr Opin Plant Biol 8:129–134

Mitchell-Olds T, Schmitt J (2006) Genetic mechanisms and evolutionary significance of natural variation in *Arabidopsis*. Nature 441:947–952

Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, Kakutani T (2001) Mobillization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. Nature 411:212–214

Miura A, Kato M, Watanabe K, Kawabe A, Kotani H, Kakutani T (2004) Genomic localization of endogenous mobile CACTA family transposons in natural variants of *Arabidopsis thaliana*. Mol Genet Genomics 270:524–532

Noël L, Moores TL, van der Biezen EA, Parniske M, Daniels MJ, Parker JE, Jones JDG (1999) Pronounced intraspecific haplotype divergence at the *RPP5* complex disease resistance locus of *Arabidopsis*. Plant Cell 11:2009–2111

Sambrook J, Fritsch EF, Maniatis T (1989) Identification and analysis of recombinants. In: Nolan C (ed) Molecular cloning, 2nd edn. Cold Spring Harbor Laboratory Press, New York, pp 2.108–2.120

Sampedro J, Lee Y, Carey RE, dePamphilis C, Cosgrove DJ (2005) Use of genomic history to improve phylogeny and understanding of births and deaths in a gene family. Plant J 44:409–419

Shen J, Araki H, Chen L, Chen J-Q, Tian D (2006) Unique evolutionary mechanism in *R*-genes under the presence/absence polymorphism in *Arabidopsis thaliana*. Genetics 172:1243–1250

Shindo C, Aranzana MJ, Lister C, Baxter C, Nicholls C, Nordborg M, Dean C (2005) Role of *FRIGIDA* and *FLOWERING LOCUS C* in determining variation in flowering time of *Arabidopsis*. Plant Physiol 138:1163–1173

Shindo C, Bernasconi G, Hardtke CS (2007) Natural genetic variation in *Arabidopsis*: tools, traits and prospects for evolutionary ecology. Ann Bot 99:1043–1054

Simillion C, Vandepoele K, Van Montagu MCE, Zabeau M, Van de Peer Y (2002) The hidden duplication past of *Arabidopsis thaliana*. Proc Natl Acad Sci USA 99:13627–13632

Ungerer MC, Halldorsdottir SS, Purugganan MD, Mackay TFC (2003) Genotype-environment interactions at quantitative trait loci affecting inflorescence development in *Arabidopsis thaliana*. Genetics 165:353–365

Wang W, Zheng H, Fan C, Li J, Shi J, Cai Z, Zhang G, Liu D, Zhang J, Vang S, Lu Z, Wong GK-S, Long M, Wang J (2006) High rate of chimeric gene organization by retroposition in plant genomes. Plant Cell 18:1791–1802

Weigel D, Nordborg M (2005) Natural variation in *Arabidopsis*. How do we find the casual genes? Plant Physiol 138:567–568

Xiao S, Emerson B, Ratanasut K, Patrick E, O'Neill C, Bancroft I, Turner JG (2004) Origin and maintenance of a broad-spectrum disease resistance locus in *Arabidopsis*. Mol Biol Evol 21:1661–1672

Zhang X, Wessler SR (2004) Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*. Proc Natl Acad Sci USA 101:5589–5594