Chromosoma (2005) 114: 103–117
DOI 10.1007/s00412-005-0342-8

RESEARCH ARTICLE

Tae-Jin Yang · Seunghee Lee · Song-Bin Chang · Yeisoo Yu · Hans de Jong · Rod A. Wing

In-depth sequence analysis of the tomato chromosome 12 centromeric region: identification of a large CAA block and characterization of pericentromere retrotranposons

Received: 4 February 2005 / Revised: 21 March 2005 / Accepted: 21 March 2005 / Published online: 17 June 2005

Abstract We sequenced a continuous 326-kb DNA stretch of a microscopically defined centromeric region of tomato chromosome 12. A total of 84% of the sequence (270 kb) was composed of a nested complex of repeat sequences including 27 retrotransposons, two transposable elements, three MITEs, two terminal repeat retrotransposons in miniature (TRIMs), ten unclassified repeats and three chloroplast DNA insertions. The retrotransposons were grouped into three families of Ty3-Gypsy type long terminal repeat (LTR) retrotransposons (PCRT1–PCRT3) and one LINE-like retrotransposon (PCRT4). High-resolution fluorescence in situ hybridization analyses on pachytene complements revealed that PCRT1a occurs on the pericentromere heterochromatin blocks. PCRT1 was the prevalent retrotransposon family occupying more than 60% of the 326-kb sequence with 19 members grouped into eight subfamilies (PCRT1a–PCRT1h) based on LTR sequence. The PCRT1a subfamily is a rapidly amplified element occupying tens of megabases. The other PCRT1 subfamilies (PCRT1b–PCRT1h) were highly degenerated and interrupted by insertions of other elements. The PCRT1 family shows identity with a previously identified tomato-specific repeat TGR2 and a CENP-B like sequence. A second previously described genomic repeat, TGR3, was identified as a part of the LTR sequence of an Athila-like PCRT2 element of which four copies were found in the 326-kb stretch. A large block of trinucleotide microsatellite (CAA)*n* occupies the centromere and large portions of the flanking pericentromere heterochromatin blocks of chromosome 12 and most of the other chromosomes. Five putative genes in the remaining 14% of the centromere region were identified, of which one is similar to a transcription regulator (ToCPL1) and a candidate *jointless-2* gene.

Communicated by P. Shaw

The sequence data from this study have been submitted to GenBank under accession no. AY850394

T.-J. Yang
Brassica Genomics Team,
National Institute of Agricultural Biotechnology (NIAB), RDA,
Suwon, 441-707, South Korea

T.-J. Yang · S. Lee · Y. Yu · R. A. Wing (✉)
Arizona Genomics Institute, 303 Forbes building,
University of Arizona,
Tucson, AZ, 85721, USA
e-mail: rwing@ag.arizona.edu
URL: http://www.genome.arizona.edu

S.-B. Chang · H. de Jong
Laboratory of Genetics,
Wageningen University,
Wageningen, The Netherlands

Introduction

The centromere of a eukaryotic chromosome is essential for maintaining faithful segregation and inheritance of genetic information. They are easily distinguishable in the microscope, consist almost exclusively of repetitive DNA sequences, have unique chromatin characteristics and are associated with specific proteins of the kinetochore plate that are required for microtubule attachment during mitotic and meiotic cell divisions.

The best-studied centromere sequence comes from the relatively simple genome of yeast (*Saccharomyces cerevisae*) and consists of less than 220 bp of AT-rich sequences. In the more complex higher eukaryotes, centromere function may be accomplished through specific higher-order DNA repeat arrays that differ even between closely related species (reviewed by Clarke 1998; Lamb and Birchler 2003) and interact with centromeric histone 3 (CenH3) histones (Zhong et al. 2002). Such centromere repeats often extend over several millions of nucleotides, and are often composed of 150- to 180-bp motifs, such as the 180-bp pAL1 satellite in *Arabidopsis* (Kumekawa et al. 2000, 2001; Round et al. 1997; Thompson et al. 1996), the 176-bp tandem repeats in *Brassica* (Harrison and Heslop-Harrison 1995), the 155-bp CentO satellite in rice (Cheng et al. 2002), the 156 bp of CentC satellite in maize (Ananiev et al. 1998), the 169-bp satellite in Medicago (Kulikova et al. 2004) and the 171-bp alphoid satellite re-

peat in humans (Henning et al. 1999). An exception to this list is the 450-kb functional centromeric region of the *Drosophila* minichromosome Dp1187, which contains two penta-nucleotide microsatellite repeat arrays (AATAT and TTCTC), ranging up to approximately 380 kb, interrupted by five transposons and a 39-kb complex of AT-rich repeats (Sun et al. 1997, 2003). Chromatin immunoprecipitation has revealed that part of the *Arabidopsis* 180-bp satellite repeat forms the functional sequences of *Arabidopsis* centromeres (Nagaki et al. 2003b). In contrast, Cheng and Murata (2003) and Nagaki et al. (2003a) observed in rice and maize that centromere-specific retrotransposons can be involved as well. The centromeric retrotransposons of maize (CRM) as well as the CentC arrays were revealed as functional DNA elements in the formation of the kinetochore in maize (Jin et al. 2004; Zhong et al. 2002).

Centromeres are flanked by large pericentromere heterochromatin blocks that may play an important role in maintaining and stabilizing centromere function (Henikoff 2002). These regions, which are microscopically recognized as highly condensed chromatin blocks with aberrant staining properties, display suppression of meiotic recombination and are strikingly gene-poor, have been found to contain various (peri)centromere-specific retrotransposons such as centromeric retrotransposon of rice (CRR) in rice (Cheng et al. 2002; Dong et al. 1998), CRM in maize (Nagaki et al. 2003a; Zhong et al. 2002), CCS in wheat (Cheng and Murata. 2003; Fukui et al. 2001) and cereba in barley (Hudakova et al. 2001). In tomato, the pericentromere regions, which occupy roughly 70% of the total genomic DNA, contain several large repeat families (Frary et al. 1996), including the TGRII and TGRIII interspersed repeats, GATA microsatellite related sequences (Vosman and Arens 1997), telomere repeat-like sequences (Presting et al. 1996) and CENP-B box and human satellite III-like sequences (Weide et al. 1998). However, no distinct tandemly repeated centromeric satellite DNA sequences have yet been identified in tomato centromeres.

Fluorescence in situ hybridization (FISH) experiments used to determine the relationship between genetic and physical distance of BACs with markers tightly linked to the *jointless-2* (*j-2*) locus showed the hybridization signals were located at the border of the long-arm pericentromere block and functional centromere region of tomato chromosome 12 (Budiman et al. 2004). For further identification of the putative sequence of the *j-2* gene, we analyzed a BAC contig and sequenced a tile of four BAC clones resulting in 326 kb of contiguous and unique sequences

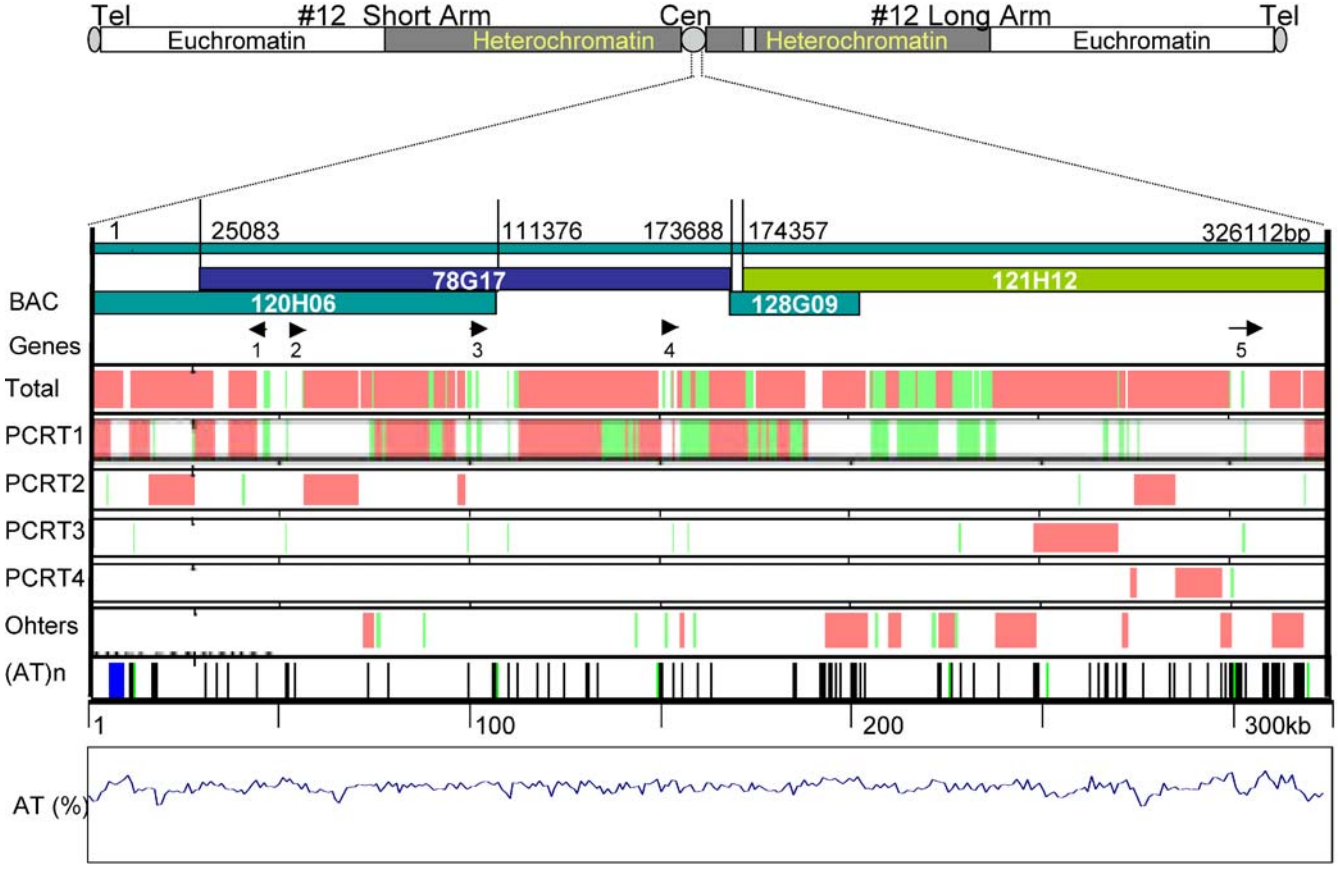

Fig. 1 Overall sequence composition of the 326-kb pericentromeric region of tomato chromosome 12. The LeHBa BAC clones 120H06, 78G17, 128G09 and 121H12 were sequenced and annotated. Five genes are represented as *arrows* based on position and orientation. Centromere repeat elements are represented based on sequence similarity with PCRT1, PCRT2, PCRT3, PCRT4 and other elements including the putative transposable elements, MITE-like, TRIM-like. SSRs including CAA block (*blue*), AT-rich, low complexity regions (*black*) and the other SSRs (*green*) are represented in the (AT)n row. AT composition (%), in 1-kb windows, shows an average of 65% (range 50–80%)

(Fig. 1). Here we present a detailed annotation of this region, as we discover an early glimpse at the molecular organization of a tomato centromeric and pericentromere region and discuss the possible candidate transcription factor for *jointless-2*.

Materials and methods

DNA sequencing

The complete sequences of three tomato BAC clones, 120H06, 78G17 and 121H12 (LeHba00000 in http://www.genome.arizona.edu), were obtained essentially as described by Yu et al. (2003). Shotgun sequencing libraries were constructed in pCUGIblu21 or pCUGIblu31 for average insert sizes of 8 and 3 kb, respectively (Yang et al. 2004). BigDye terminator chemistry v3.0 (ABI) was used for the sequencing reactions. The sequences were analyzed using an ABI3730xl automatic DNA sequencer (ABI). Base-calling was performed automatically using Phred and vector sequences were removed by CROSS_MATCH (Ewing and Green 1998; Ewing et al. 1998). High-quality vector-trimmed sequences were used for the sequence assembly of each BAC clone using Phrap and Consed (Gordon et al. 1998). After sequence assembly and alignment of the three BAC clones, one physical gap remained between BAC clones 78G17 and 121H12 and was closed using a bridging clone derived from an overlapping BAC 128G09. The bridging clone was fully sequenced using transposon-mediated sequencing as described by Yang et al. (2003, 2005).

Sequence analysis

The structure of each element was characterized by pair-wise sequence comparison using PipMaker (Schwartz et

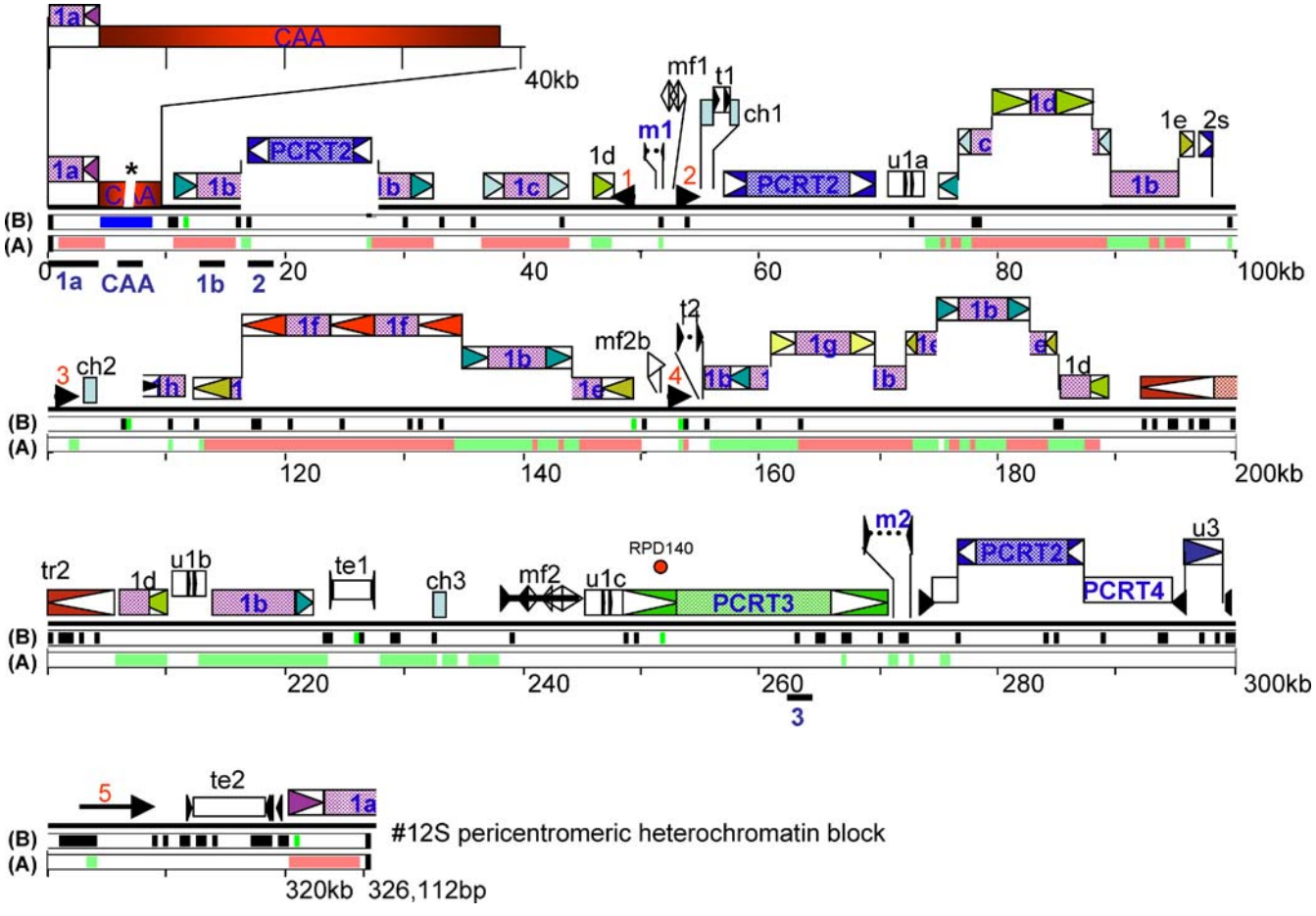

Fig. 2 Detailed representation of the 326-kb tomato sequence. Features of every identified element are shown in two *boxes* (analyzed by Pipmaker): **a** the regions homologous with the PCRT1 (*red* and *green* represent high and medium sequence similarity, respectively); **b** the regions containing AT-rich, low complexity regions. All elements are shaped by their basic structures and distinguished by colors. Three LTR retrotransposon families are represented as *purple* (PCRT1), *blue* (PCRT2), and *green* (PCRT3). Seven PCRT1 subfamilies are distinguished by the *color of the triangle* in the LTR box. LINE like CRT4 is presented as an *empty box*. The other elements are positioned with magnification, in need. Three chloroplast DNAs are represented with *light blue boxes*. The actual name and the nucleotide position of each element are listed in Tables 1 and 3. Five predicted genes are represented with *arrows and red numerals*. Actual position of shotgun clones used for FISH hybridizations is represented as *bold line under* (**a**) *bar*: *1a* (LTR and coding region of PCRT1a); *CAA* (CAA repeat); *2* (LTR region of PCRT2); *3* (rve region of PCRT3)

al. 2000), Miropeat software (Parsons 1995) and Blast2 analysis (http://www.ncbi.nlm.nih.gov/BLAST/). Subsequent Blast-nr, Blast-est, Blast-X, and t-BlastX were used as needed (http://www.ncbi.nlm.nih.gov/BLAST/). Gene annotation was achieved using several web-based gene prediction programs such as FGENE-SH *Arabidopsis* (http://www.softberry.com/berry.phtml), Genescan dicot (http://genes.mit.edu/GENSCAN.html), and GeneMark *Arabidopsis* (http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi). Simple sequence repeats (SSR) were identified using Repeatmasker (http://ftp.genome.washington.edu/RM/webrepeatmaskerhelp.html).

Fluorescence in situ hybridization

Pachytene chromosome spread preparations of 2-mm anthers of tomato *Solanum lycopersicon* cv. VFNT Cherry ($2n=2x=24$) were used in all FISH experiments following the protocols of Zhong et al. (1996) and Budiman et al. (2004). Shotgun clones were used as probes for FISH analysis of each identified element. The clones were chosen from the shotgun libraries based on their end sequences and position (represented as bars in Fig. 2). Each shotgun clone DNA (1–2 μg) was labeled with either biotin-16-dUTP or digoxigenin-11-dUTP by nick translation using the manufacturer's protocol (Roche) and FISH was performed according to Zhong et al. (1996). Chromosomes were counterstained in 5 μg/ml DAPI in Vectashield antifade (Vector Laboratories). Slides were examined under a Zeiss Axioplan 2 Photomicroscope equipped with epifluorescence illumination, filter sets for DAPI, FITC and Texas Red fluorescence. Selected images were captured by a Photometrics Sensys, 1,305×1,024 pixel CCD camera. Image processing and thresholding was performed with the Genus Image Analysis software (Applied Imaging Corporation). DAPI images were separately sharpened with a 7×7 Hi-Gauss high pass spatial filter to accentuate minor details and heterochromatin banding of the chromosomes. All fluorescence images were pseudo-colored and improved for optimal brightness and contrast using Adobe Photoshop.

Southern hybridization and BAC library screening

One BAC filter (LeHba-A, http://www.genome.arizona.edu), representing 2× tomato genome equivalents and southern filters containing enzymes-digested tomato genomic DNA from Heinz 1706 and LA166, was hybridized

Table 1 Nucleotide position, direction and TSD sequence of each PCRT element

Name	Direction	TSD	Start	Nest insertions	End	TSD	Length (bp)	Remarks
PCRT1a_1	Rev.				4,706		4,706	3′ Truncated
PPCRT1a_2	Fow.		319,666		326,088		6,423	3′ Truncated
PPCRT1b_1	Fow.	atata	10,657	//15,433–27,620//	32,505	aaaata	9,664	Nested
PPCRT1b_2	Rev.		73,993	//74,816–89,211//	94,238		5,852	5′ Truncated, nested
PPCRT1b_3	Fow.		134,095		144,489		10,395	5′ Truncated
PPCRT1b_4	Rev.		155,720	//163,337–172,735//	174,787		9,671	5′ Truncated, nested
PPCRT1b_5	Fow.	tagga	177,306		185,939	taggg	8,634	3′ Truncated
PPCRT1b_6	Fow.		214,579		223,002		8,424	5′ Truncated
PPCRT1c_1	Fow.	attaa	36,568		43,592	attac	7,025	
PPCRT1c_2	Rev.	taaca	74,816	//78,101–87,762//	89,211	tatca	4,736	Nested
PPCRT1d_s	Fow.		45,579		47,258		1,681	Solo LTR
PPCRT1d_1	Fow.	atgtt	78,101		87,762	ttgtt	9,662	AF411806, solo LTR
PPCRT1d_2	Rev.		175,563	//177,306–185,939//	188,531		4,336	3′ Truncation, nested
PPCRT1d_3	Rev.		204,276		208,381		4,106	Truncated
PPCRT1d_s	Fow.		208,544		209,613		1,071	Solo LTR
PPCRT1e_s	Fow.		94,238		96,056		1,820	Solo LTR
PCRT1e_1	Rev.	gcaac	113,003	//117,219–144,489//	149,853	aaaac	9,582	BAA95869.1, nested
PCRT1f_1	Rev.	cttgg	117,219		127,536		10,318	Tri LTR
PCRT1f_2	Rev.				134,023	cttgg	6,489	
PCRT1g	Fow.	aaaaa	163,337		172,735	aaaaa	9,399	
PCRT1h	Fow.		109,131		112,494		3,364	Truncated, AC145120
PCRT2_1	Rev.	tgtac	15,433		27,620	aagtac	12,188	
PCRT2_2	Fow.	gaaag	56,562		70,562	ggaag	14,001	
PCRT2_3	Rev.	tattc	275,195		285,721	tttac	10,527	
PCRT2_s	Fow.	tattt	96,842		98,521	tattt	1,681	Solo LTR
PCRT3	Fow.	ctcaa	248,442		270,747	ctcaa	22,306	
PCRT4	Fow.	ccttagtt	274,616	//275,195–285,721//	297,118	ccttagct	11,973	LINE like, nested
Total							210,034	

with probes as noted above to characterize the redundancy of elements listed in this study.

Results

Overall view of a 326-kb centromeric region of tomato chromosome 12

We generated a high-quality (Phred >30) contiguous 326,112-bp sequence from a microscopically defined centromeric region of tomato chromosome 12, except for one putative misassembly in one large CAA repeat block (denoted by asterisk in Fig. 2). To sequence the region, three BAC clones (LeHBa120H06, LeHBa078G17, and LeHBa121H12) were completely sequenced as well as a subclone from a fourth BAC (LeHBa128G09) that was used to fill a small 669-bp gap (Fig. 1). The BAC contig contains the RPD140 marker, which completely cosegregates with *j-2* in a population of 1,122 *jointless-2* homozygous plants (Budiman et al. 2004). FISH analysis demonstrated that this contig is located in the centromeric region of chromosome 12, flanked by large pericentromere heterochromatin blocks of 65.5 Mb, which equals 85% of the total 76.4 Mb of chromosome 12 (Fig. 1; Budiman et al. 2004).

Annotation of the sequence revealed that more than 84% of the sequences were repetitive arrays which were organized in a complex mosaic of retrotransposons and transposable elements (Tables 1, 2 and Fig. 1). We identified 27 complete and truncated retrotransposons, two transposable elements, three MITEs, two TRIMs, ten unclassified repeats and three chloroplast DNA insertions, which altogether represent a total of 270 kb (84% of the 326 kb). The retrotransposons are redundant and located in the pericentromere heterochromatin regions, which we will refer to as PCRT (peri-centromeric retrotransposon of tomato). The most predominant repeat class was the PCRT1 family (PCRT1a thru PCRT1h) representing 60% of the overall sequence (PCRT1 in Fig. 1). Simple sequence repeats (SSR) occurred at 113 positions resulting in a total of 6,034 bp (Figs. 1, 2b). AT-rich, low-complexity regions were identified at 105 sites (1 in every 3 kb). Finally, five genes were identified including one putative transcription regulator and four small computationally derived hypothetical proteins.

CAA microsatellite repeat block

Intensive efforts to obtain a high-quality contiguous sequence were hindered by the presence of a large CAA repeat region, located at position 4,706 bp from the left end of BAC LeHba120H06. By comparing the virtual restriction digestion patterns of the assembled sequence data with the empirical enzyme digestions of BAC 120H06, we deduced that the 4-kb CAA repeat block is likely part of a much larger microsatellite array of approximately 35 kb in size. The CAA block was interrupted by the insertion of a PCRT1a element at the left border of the 326-kb sequence (Fig. 2 top), suggesting that the CAA repeat block could extend much further toward the centromere. All remaining restriction digestion patterns with *Hin*dIII, *Eco*RI and *Bam*HI were coincident with the virtual digestions indicating no other assembly errors in the 326-kb sequence. The CAA blocks appear to occupy large portions of the pericentromere regions of most chromosomes based on pachytene FISH analysis, including a conspicuous site in chromosome 12 (arrow in Fig. 3a.1 and a.2). Chromosome 12 is the shortest chromosome (22.5 μm) of the complement and can be easily identified by its arm ratio (*L*/*S*=1.1)

Table 2 General structure and appearance of four PCRT families found in the 326-kb tomato sequence

Family	Copy no.[a]	Total (bp)	Internal (bp)	LTR (bp)	Homology (%)[b]	Conserved domain[c]	Best hits (GenBank #)	Remarks
PCRT1[d]	19(4?)					gag, gf, rvt, rve	AAN01260	Rice polyprotein
PCRT1a[e]	2	8,953	5,187	1,883	91			
PCRT1b	6 (?)	9,662	7,366	1,051	83			
PCRT1c	2	7,023	2,821	2,131	85			
PCRT1d	4 (2)	9,660	2,197	3,626	80			
PCRT1e	2 (1)	9,478	4,953	2,215	76			
PCRT1f	2	10,316	2,497	3,907	82			
PCRT1g	1	9,397	4,769	2,256	81			
PCRT2	4 (1)	12,186	9,362	1,269	85	gag, gf, rve	AAF24529	*Arabidopsis* polyprotein
PCRT3	1	22,304	14,959	3,630	91	gag, rvt, rve	NP680279.1	*Arabidopsis* polyprotein
PCRT4	1	11,971	11,971	0		rvt, RH	NM130086.1	*Arabidopsis* LINE-like

CRT1h identified by comparison with GenBank accession # AC145120 was not included here because of its degeneracy

[a]No. of truncated element or solo LTR are in parenthesis

[b]Similarity between LTR-L and -R

[c]The domains are aligned based on their order. (gag: pfam03732, Retrotrans_gag, Retrotransposon gag protein, gf: gag finger: smart00343, ZnF_C2HC, zinc finger; pfam00098, zf-CCHC, zinc knuckle, rvt: pfam00078, rvt, reverse transcriptase (RNA-dependent DNA polymerase), rve: pfam00665, rve, Integrase core domain, RH: pfam00075, RNaseH

[d]PCRT1 subfamilies have the same internal sequence but different LTR sequence

[e]The structure is deduced from GenBank accessions # AF411805 and # AF411806

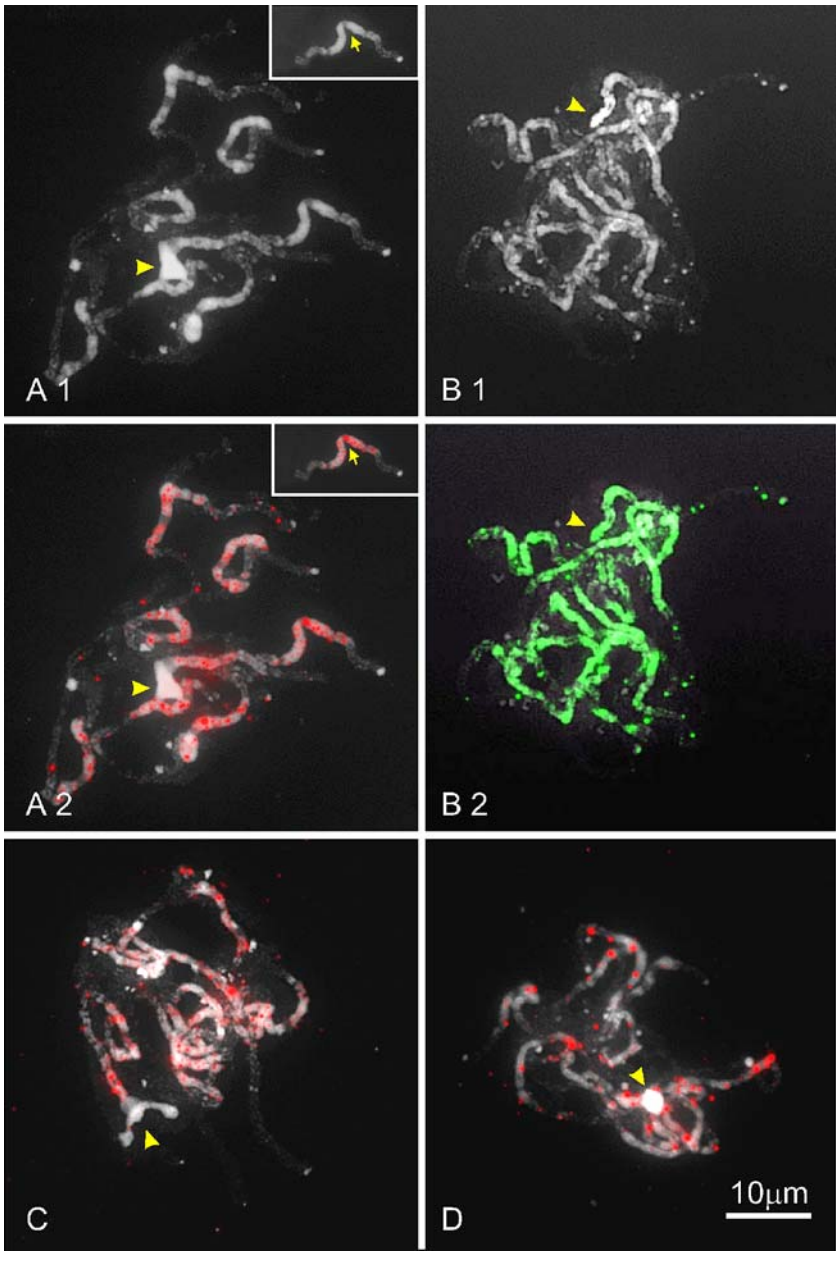

Fig. 3 FISH patterns of the CAA microsatellite, and the PRCT1, PRCT2 and PRCT3 repeats on pachytene complements of tomato (*Solanum esculentum* cv. VFNT Cherry). Chromosomes were counterstained with DAPI and displayed in *gray*. The heterochromatin areas are visible as brightly fluorescing blocks at the distal ends and pericentromere blocks. Centromeres are easily distinguishable as weakly fluorescing gaps flanked by large heterochromatin blocks. Euchromatin areas have very little DAPI fluorescence but are generally more granular than centromeres. The brightest and most conspicuous heterochromatin block is at the satellite and nucleolar organizer region (*arrowheads*) in the distal end of the short arm of chromosome 2. **a**.1 DAPI fluorescence and **a**.2 Texas Red signals of the CAA microsatellite probe. Signals are obvious in all centromeres and pericentromere blocks but not in the distal heterochromatin blocks and the NOR (*arrowheads*). *Inset*: isolated chromosome 12 with the centromere indicated by *arrow*. **b**.1 DAPI fluorescence and **b**.2 FITC detection of the PCRT1 probe. Strong signals were found in all heterochromatin regions, some centromeres and especially at the bright DAPI-stained NORs site (*arrowheads*). **c** PRCT2 was mainly found in the heterochromatin regions and in few centromere and euchromatin spots but was absent in the NOR. **d** PCRT3 also occurred in heterochromatin, and in very few euchromatin spots, but not in the NOR, and was in general far less abundant than PCRT1 and PCRT2. All photomicrographs are at about the same magnification

and symmetric heterochromatin/euchromatin pattern (Barton 1950; Budiman et al. 2004). The centromeres are shown as the faint gray gaps flanked by bright white DAPI-stained heterochromatin blocks (Fig. 3a.1).

PCRTs are major components of the 326-kb sequence

Analyses of the repeat elements were hindered by manifold degenerations, truncations, or interruptions by nested insertions of other elements, and the paucity of annotated

retrotransposon sequence information for the *Solanacea* family. By careful pairwise sequence comparisons, BLAST analysis and manual editing, we characterized the location and features of 27 PCRT elements (Fig. 2). Their positions, features and 5-bp target site duplication (TSD) sequences are represented in Table 1. The overall length of each element and LTR sequences are represented in Table 2.

The PCRT1 family occupies about 60% of the 326-kb sequence. Eight PCRT1 subfamilies, PCRT1a–PCRT1h, were identified based on DNA sequence similarity to coding regions (Table 1). They share a part of the coding region but have different LTR sequences (Fig. 4). High levels of degeneracy in all PCRT1 subfamilies, except PCRT1a, abundant nested insertions by other elements (Fig. 2 and Table 1) and low sequence similarity (around 80%) between the 5′- and 3′-LTR sequences (Fig. 4b and Table 2) suggest that these elements (PCRT1b–PCRT1h) are ancient components of the pericentromere.

The highly amplified retrotransposon subfamily PCRT1a contains intact coding sequences, including conserved domains of Ty3/Gypsy-like retrotransposons (shown on Fig. 4a) and sequence similarity (91%) between two members, PCRT1a-1 and PCRT1a-2 (Table 2), which suggests they are relatively younger than the other PCRT elements. A highly homologous structure with over 90% similarity was found on the additional sequenced tomato BAC clone # 207, which is located in the pericentromere region of chromosome 7 (van der Hoeven et al. 2002). As the partial PCRT1a element exists at both ends of the 326 kb, at the beginning of 120H06 and the end of 121H12, we used the sequence information from BAC # 207 (complementary sequence of 1–3912 bp and direct sequence of 1–5415 bp of GenBank accession # AF411806 and # AF411805, respectively) to reconstruct an intact PCRT1a element (Table 2).

Pachytene FISH of the PCRT1a elements showed high levels of hybridizations signals in all pericentromere heterochromatin regions and in some subtelomeric heterochromatin blocks (Fig. 3b.1 and b.2) including the NOR region of the chromosome 2 short arm (arrowhead in Fig. 3b.1 and b.2), which is easily identified by the presence of 45S rDNA and the absence of the TGR1 repeat (Ganal et al. 1991, 1992; Zhong et al. 1998).

The PCRT1b subfamily, occupying up to almost 90 kb (27% of 326 kb), displays high degrees of degeneracy and truncation. Six elements were identified as almost complete structures after removing the nested insertions. Three of them (PCRT1b-1, -2, -4) were interrupted by other elements (Fig. 2) and five of them (PCRTb-2, -3, -4, -5, -6) were truncated (Fig. 4c and Table 2). The coding sequence shows similarity with PCRT1a (Fig. 4b) even though many stop codons were found in the internal domain.

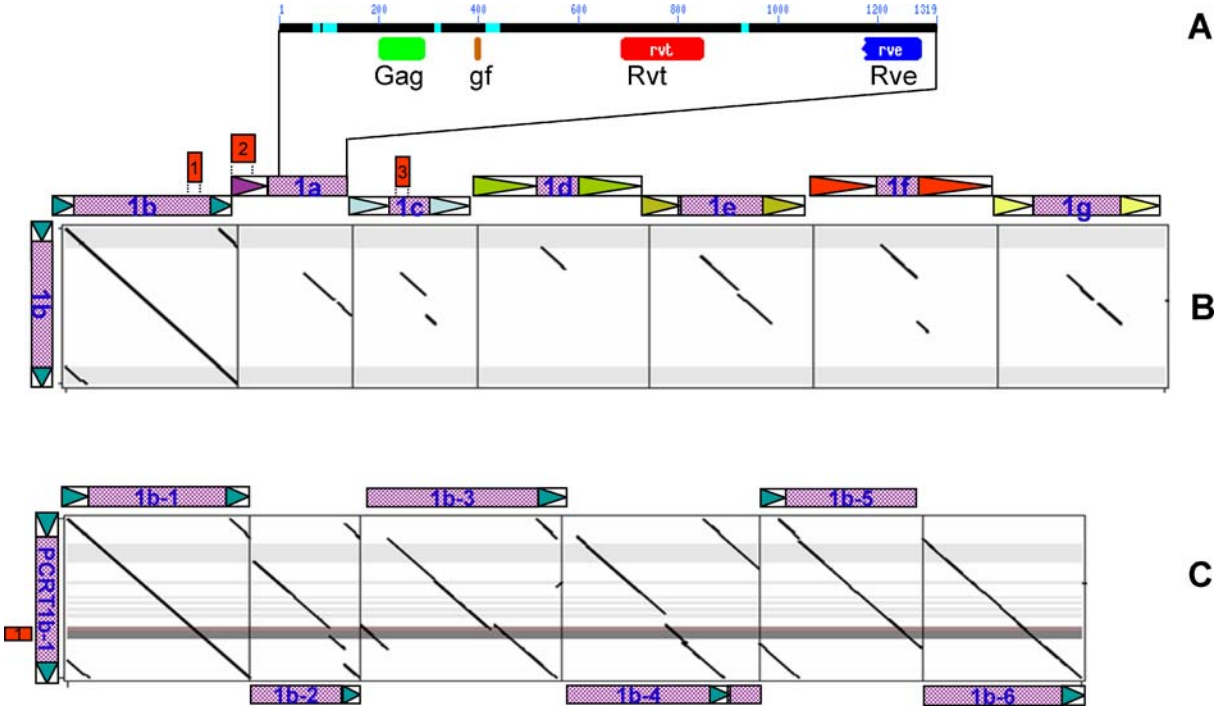

Fig. 4 Features of PCRT1 family and PCRT1b subfamily. **a** Polyprotein deduced from PCRT1a show sequence conservation with Ty3/Gypsy retrotransposons. **b** LTR region of PCRT1b is represented as a *light gray box*, denoting no sequence homology among members. Seven members of PCRT1 family have similar coding sequences but different LTR sequences. Three known centromeric repeat sequences are denoted by *numbers in red boxes*: *1* CENP-B box like sequence (GenBank accession # AF072522), *2* TGR2 (or U30: # X90770), *3* GATA microsatellite-related sequence (# X91108: 207–642). **c** Dotplot comparisons of each member show LTR regions of each PCRT1 subfamily. The LTR region of PCRT1a (denoted as *gray* in PCRT1a dotplot) is deduced from previously sequenced BAC clone # 207 (van der Hoeven et al. 2002). **d** Comparison of six PCRT1b subfamilies. Coding regions and homologous sequence with AA13 harboring CENP-B box-like sequence are shaded as *light gray* and *dark gray boxes* (*red box*), respectively

The PCRT1c–PCRT1h elements are highly truncated and degenerated, and resemble the internal sequences, but different LTR sequences. Each subfamily has LTR sequence ranging from 2131 to 3907 bp with 76–85% sequence similarity between 5′- and 3′-LTRs (Table 2). Their truncated coding regions are similar to parts of the coding regions of PCRT1a and PCRT1b (Fig. 4b). The structure of the polyprotein genes in the elements were not predicted by any gene prediction programs, even though BLAST-X revealed significant sequence similarity with known polyproteins of Ty3/Gypsy type retrotransposon.

PCRT2 family is redundant in heterochromatin regions

A second class of pericentromere repeats, named PCRT2, was detected in four, almost identical elements including one solo LTR (Figs. 2 and 5). It has a 1,269-bp LTR and about 10 kb of internal sequence (Tables 1, 2). The internal coding sequence shows significant sequence similarity with the centromeric retrotransposon Athila in *Arabidopsis* (GenBank accession # AC007534, 1e-75). Based on hybridization to a Heinz 1706 tomato BAC library (Budiman et al. 2004), we estimate the number of copies at more than 1,200. The previously described tomato genome repeat, TGR3, shows significant sequence similarity with a part of the LTR sequence of PCRT2. FISH reveals that the PCRT2 occurs most prominently in the pericentromere heterochromatin and in some minor sites of centromere regions (Fig. 3c).

PCRT3 is a newly inserted large element

One large retrotransposon (named PCRT3) was identified (22,306 bp in length) and contained 3,630 and 3,716 bp of 5′ and 3′-LTRs, respectively (91% homology, Tables 1, 2).

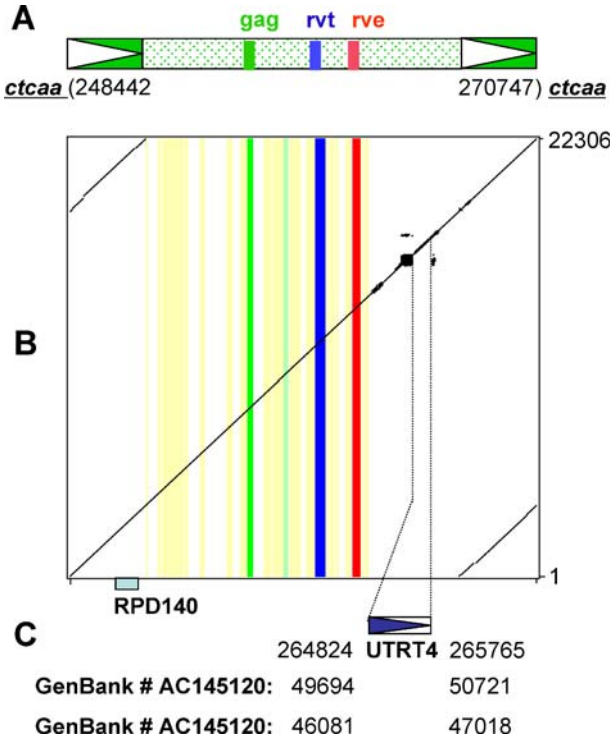

Fig. 6 The structure of a large retrotransposon PCRT3. A total of 22,306 bp (position 248442–270747) shows a complete structure of a retrotransposon. The deduced protein shows conserved gag (*green*), rvt (*blue*) and rve (*red*) domains, respectively (**a**). The regions for exon and LTR of PCRT3 are represented as *black boxes* and *boxed arrowheads*, respectively, in (**b**) based on dotplot, a pair-wise sequence comparison of PCRT3 (**c**). Five nucleotides TSD sequence and a 4-bp inverted terminal repeat in the LTR are represented under the dotplot as *bold italics* and *capital letters*, respectively. A 941-bp solo LTR-like sequence (*UTRT4*) is inserted into short tandem repeat sequence and the complete structure is identified by comparison with a sequence, GenBank accession # AC145120 (**d**). The position of RPD140, an RFLP marker cosegregating with *j*-2, is denoted as *light blue box* under the dotplot (position 251322–251888)

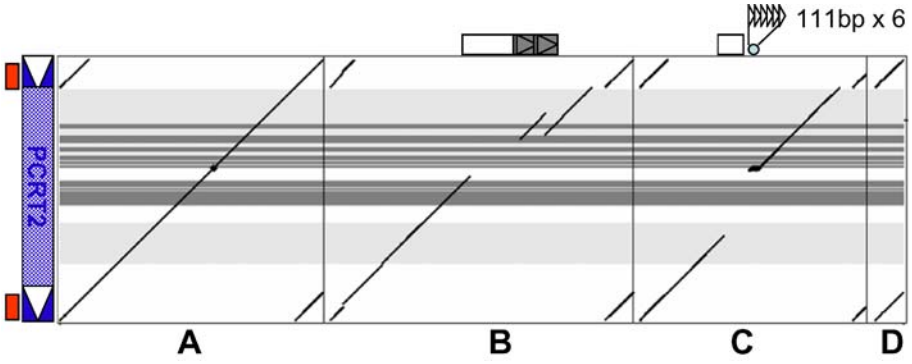

Fig. 5 PCRT2 Elements. Three members and a solo LTR were compared with PCRT2-1: **a** PCRT2-2 (complementary sequence of 12,188 bp, located between 15433 and 27620 bp); **b** PCRT2b (14,001 bp: position 56562–67562 bp); **c** PCRT2-3 (complementary 10,527 bp: position 275195–285721); **d** solo LTR of PCRT2 (96842–98,521). Putative coding regions are shaded with *light gray*. The *red box* represents the homologous region of tomato repeat sequence TGR3. Conserved domains of Ty3/Gypsy type retrotransposons are shaded with *dark gray*: *gag* Athila gag-like, *rvt* reverse transcriptase, *rve* integrase core domain. Duplication of a 944-bp sequence (*black box*) in PCRT2-2 and deletions or substitutions of long segments (*white boxes*) in PCRT2-2 and PCRT2-3 seemed to be mediated by uneven recombination between elements. Six copies of a 111-bp tandem repeat (*circle* in PCRT2-3) may be byproducts of several rounds of uneven recombination

An ORF encoding 1,832 amino acids shows conserved domains with gag, reverse transcriptase (*rvt*) and an integrase core domain (*rve*) (Fig. 6a and b). No significant DNA sequence similarity was detected in GenBank except 941 bp between nucleotides 264824–265765 bp. This 941-bp sequence appears to be an unclassified solo LTR (named UTRT4). The complete retrotransposon structure of the UTRT4 is identified in the sequence of GenBank accession # AC145120 (46081–50721). The UTRT4 remains flanking 5-bp TSD sequence in short tandem repeat (STR) sequences of the PCRT3 (Fig. 6d). It is interesting to note that the RFLP marker RPD140 (Zhang et al. 2000), which cosegregates with *j-2*, is located in the left LTR of PCRT3 (denoted as a bar under Fig. 6c). FISH reveals that PCRT3 is also a component of pericentromere heterochromatin regions of most chromosomes as well as a few sites inside functional centromeres (Fig. 3d).

PCRT4, a LINE-like element, is inserted into a putative MITE element

Figure 7 displays another complex of nested insertions by different elements including the new PCRT4, which was discovered by editing out one PCRT2 element. The PCRT4 encodes a putative polyprotein showing sequence similarity with non-LTR type LINE elements. The flanking target site duplication (TSD) sequence was found behind its poly $(A)_{11}$ tail. The PCRT4 element was flanked on both sites by 209-bp terminal inverted repeats (TIR). The right TIR was interrupted by an unclassified solo LTR-like element (2,914 bp, named UTRT3) showing similarity with an element in Genbank accession # AC139840 (Fig. 7). By trimming out all three elements, PCRT2, PCRT4 and UTRT3, a total of 1,093 bp remained and revealed a MITE-like structure with a 209-bp TIR sequence (named MIT3, Table 3).

Two additional MITE-like elements, MIT1 and MIT2 (Table 3), were identified. The MIT2 element has identical 9-bp TSD sequences, 88 bp of TIR sequence and 762 bp of a degenerate internal sequence. The MIT2 element was found in three sequences of *Solanaceae* reported in GenBank (Fig. 8). Another MITE like element MIT1 is identified together with an unclassified foldback structure (MFT1) and a TRIM element (TRT1) between 51437 and 55803 bp where no retrotransposon were found (Table 3, Fig. 9). The MIT1 has an approximate 100-bp AT-rich internal sequence with flanking 117-bp TIR sequences, showing sequence similarity with a tomato sequence in GenBank accession # AC145120 (Table 3, Fig. 9b). Two unknown complicated foldback structures, MFT1 and MFT2, containing more than 2 units of continuous inverted repeats were identified in the 326-kb sequence (Table 3). The MFT1 consists of two small continuous foldback structures, 150 bp (MFT1-a) and 120 bp (MFT1-b), flanked by a putative 3-bp TSD sequence (Fig. 9c and d). Another large-scale foldback structure, MFT2 (8,048 bp) consists of three continuous foldback units a, b, and c (Table 3). These complicated foldback structures may have arisen from several rounds of nested insertions or juxtaposition of different MITE elements as reported in rice by Jiang and Wessler (2001). However, no similar sequences with these two unclassified structures were detected in GenBank.

Terminal repeat retrotransposon in miniature (TRIM) elements and insertion of chloroplast DNA segments

Two TRIM-like elements (TRT 1 and TRT 2), identified in the 326-kb sequence, show distinct features of the previously described TRIM element with 100–300 bp of internal sequence and about 100 bp of terminal repeat sequence flanked by 5 bp of TSD sequences (Witte et al. 2001; Yang et al., personal communication). The TRT 1 shows sequence similarity with a previously identified TRIM ele-

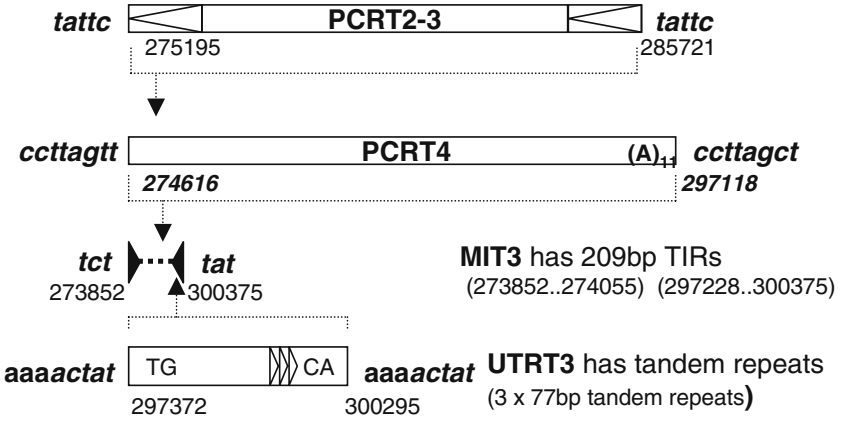

Fig. 7 A complicated set of nested insertions in a MITE-like element, MIT3. A total of 1,093 bp of MIT3 is expanded to 26,523 bp long by interruptions of three other elements, PCRT4, PCRT2 and UTRT3. The PCRT4 appears to be a LINE retrotransposon based on its coding sequence and its structure ending with poly(A) tail, providing 8-bp TSD sequence without LTR sequence. The PCRT4 is subsequently nested by a PCRT2. Another solo LTR-like element UTRT3 is inserted into 3′TIR sequence providing 8 bp (or 5 bp) TSD sequence. ignated as *arrows* and nucleotide positions are numerically represented. Putative TSD sequences are shown as *bold italics* on both sides of each element

Table 3 Other elements identified in the 326 kb

Name	TSD	Position (start)	Position (end)	TSD	Terminal repeat (bp)	Total length (bp)	Homolog and remarks
Putative transposable element in tomato				(TIR)			
TET1	tcatgggttg	223,141	227,059	tcaa-gttg	194	3,918	U81378, BAB93302 Rice Tam3 transposase 5e-27
TET2	agttgtgta	311,797	319,033	aga-gtctt	494–300	7,236	BAB11196, MuDR-like, 4e-70
MITE like elements of tomato				(TIR)			
MIT1		51,439	51,727		144	288	AC145120
MIT2	taaatcatt	271,688	272,626	Taaatcatt	83	938	AC122544, ACC135288, L36364 SI-1 gene
MIT3	tct	273,846	300,381	Tat	209	1,093	Nest insertion by PCRT2, PCRT4, and UTRT3
MITE-like and complicate foldback structure in tomato				(TIR)			
MFT1	ttt	52,373	52,644	Ttt		270	Juxtapose alignment of two different foldbacks
MFT1-a		52,373	52,523		75	150	
MFT1-b		52,524	52,644		60	120	
MFT2		237,977	246,025			8,048	Juxtapose alignment of three large foldback units
MFT2-a	acatatca	238,063	241,869	Atttatca	1,277	3,806	
MFT2-b		241,878	243,995		822	2,117	
MFT2-c	acaatattc	244,013	245,311	acaaaat-c	649	1,298	
MFT2-b		245,318	245,985		667	2,733	
MFT2a		227,060	228,103			1,043	
MFT2b		151,046	151,279			233	
TRIM-like elements of tomato				(TR)			
TRT1	attgt	55,470	55,803	Attac	117	333	AJ276865
TRT2	aatga	155,122	155,720	Aagca	119	598	
Unclassified tandem repeats of tomato: retrotransposon-like				(TR)			
UTRT1a		71,267	73,965			2,698	Three tandem repeats (176 bp × 2, 32 bp × 3, 117 bp × 2)
UTRT1b		210,832	213,134			2,302	
UTRT1c		246,110	248,420			2,310	
UTRT2	gtaca	193,103	204,275	Gaaca	4,928	11,172	Truncated retrotransposon like
UTRT3	aaaactat	297,381	300,295	Aaactat		2,914	AC139840, solo-LTR like
UTRT4		264,824	265,735				AC145120 (46081–50721), solo-LTR like
Chloroplast genome insertion							
Ch1		55,200	55,955			755	Z00044: chloroplast genome (116677–116253)
Ch2		103,550	103,658			108	Z00044: chloroplast genome (114717–114818)
Ch3		231,880	232,280			400	Z00044: chloroplast genome (49092–48673)

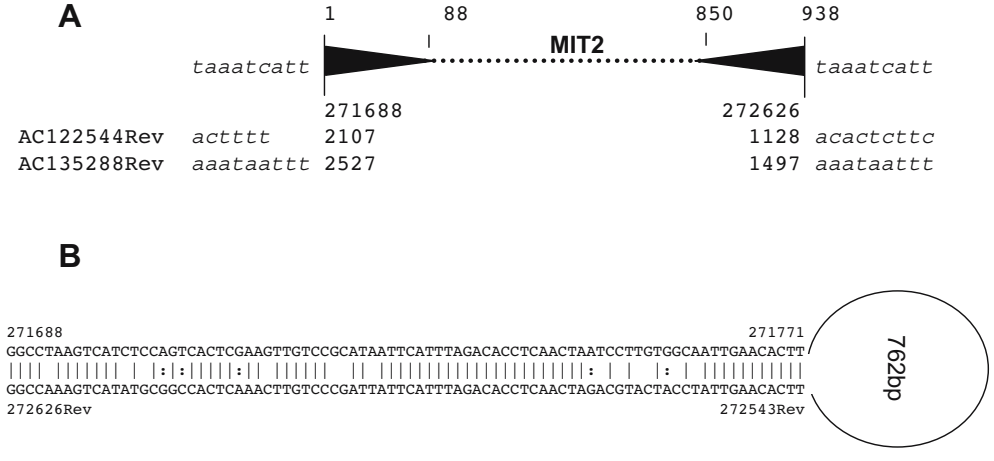

Fig. 8 MIT2-A MITE-like element. **a** MIT2 contains a 762-bp degenerate coding sequence and is flanked by 88 bp of TIR sequence. The 8-bp TSD sequence and the nucleotide positions of two homologous sequences are represented. **b** Hairpin structure of the complementary 88-bp TIR sequences of MIT2

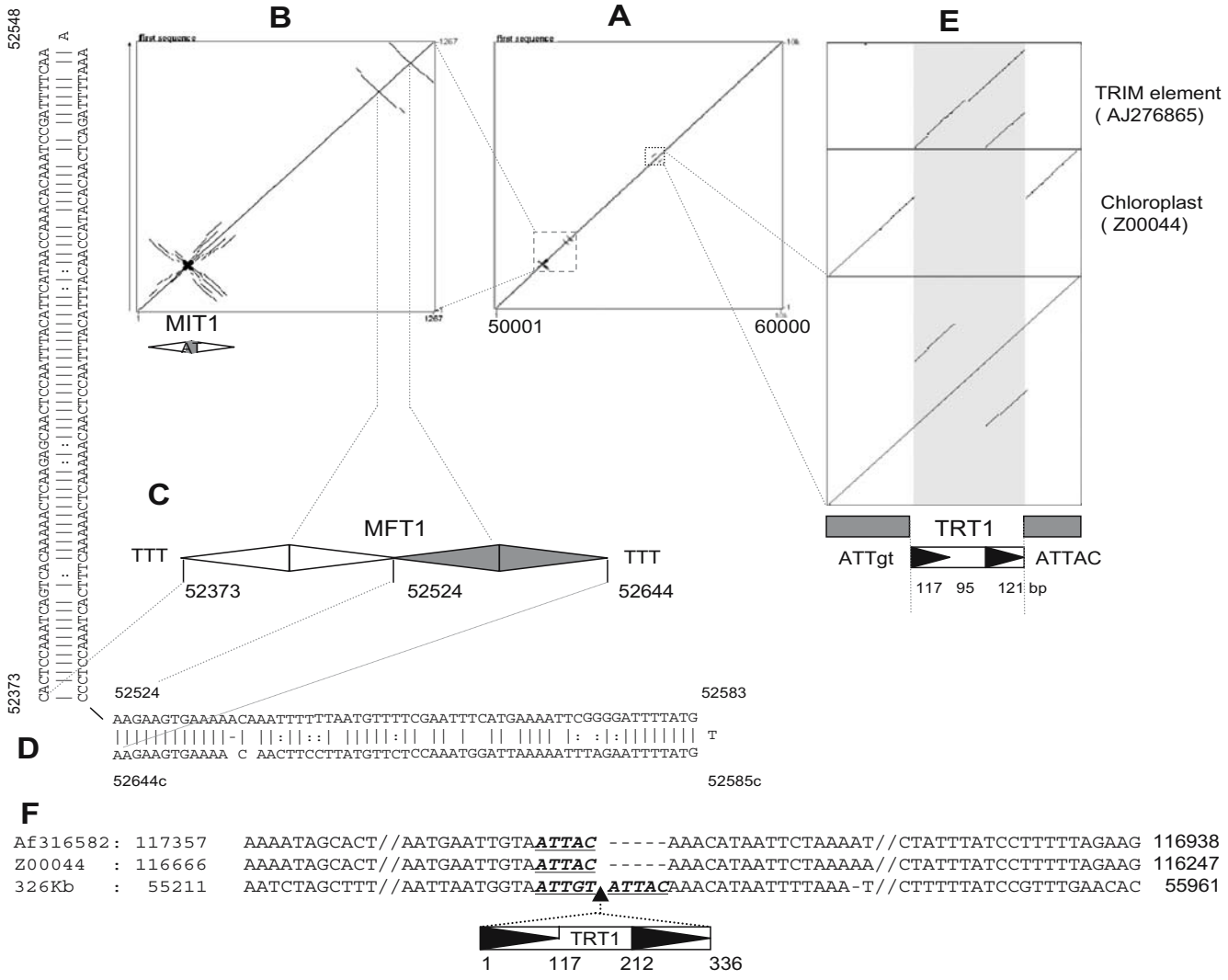

Fig. 9 Two MITE-like foldback structures (*MIT1* and *MFT1*) and one TRIM-like element (*TRT1*). No retrotranposons were identified in the 50- to 60-kb region. Pairwise comparison of the 10-kb sequence (**a**) and BLAST reveal four DNA elements, MIT1, MFT1, TRT1 and a chloroplast DNA fragment. A magnified region (**b**) shows one MITE-like foldback structure MIT1 and one complicated foldback structure MFT1. The MFT1 has a contiguous array of two units of foldback structure (**c** and **d**). The other magnified region (**e**, *bottom*) and *dotplot* with tobacco chloroplast genome (**e**, *middle*) and *dotplot* with TRIM element (**e**, *top* and *gray block*) represent the nested insertion of a TRIM element into the chloroplast DNA insertion. The related empty sites (*RESites*) of the TRT1 are found in the chloroplast genome sequences (**f**). The nucleotide positions are *numerically represented* from the sequences in the GenBank accession numbers AF316582 and Z00044

ment (GenBank accession # AJ276865) and is located within a 755-bp chloroplast DNA fragment (80% similar with the tobacco chloroplast genome) (Fig. 9e). Sequence comparison with the related empty sites (RESites) of the tobacco chloroplast genome reveals flanking 5-bp TSD sequence (Fig. 9f). The TRT2 has a typical TRIM structure but was not similar to any previously identified sequence. Two additional insertions of small chloroplast DNA fragments, 108 and 400 bp, were found in the 326-kb sequence (Table 3) as previously described by Stegemann et al. (2003).

Transposable elements

We also identified two putative DNA transposable elements, TET1 and TET2, with sequence similarity to MuDR and Tam-3, respectively. TET1 contains 194 bp of a TIR sequence, flanked by a degenerate 9-bp TSD (Table 3, Fig. 2). A total of 3,918 bp of TET1 shows significant sequence similarity (91%) with the previously identified Tam3-like putative transposase in tomato (GenBank accession # U81378). However, an ORF structure was not predicted in our sequence, which was found to contain many stop codons. The identification of TET2 was based on the deduced protein coding sequence showing conserved do-

mains with MuDR (pfam03108), ZnF-PMZ (smart 00575) and significant similarity with a MuDR-like transposable element (BAB11196). TET2 also contained degenerate TIR sequences of about 300 bp in length at both sides of the coding region.

Functional genes in the 326-kb centromeric region

A total of 51 genes were predicted in the 326-kb sequence, but 46 of them were associated with transposons. Four of the five remaining gene predictions (genes 1–4) encode short ORFs with no known functions: 154, 30, 181 and 54 amino acids, respectively. Gene 5, however, shows significant sequence similarity with an *Arabidopsis* C-terminal phosphatase-like gene (AtCPL3, GenBank accession # AF486633; Koiwa et al. 2002) and seven tomato ESTs obtained from flower bud or fruit cDNA libraries (AW737931, AW031785, AW616528, AW033495, BF051717, BF050849, BG626651). The deduced protein has four conserved domains: a catalytic domain of ctd-like phophatases (CPDc, smart 00577), a NLI interacting factor (NIF, pfam03031), a BRCA1 C terminus domain (BRCT, pfam 00533) and a TFIIF-interacting CTD phosphatase (FCP1, COG5190). The BRCA1 C terminus domain is known to function for repair of DNA double-strand breakage (Lafarge and Montané 2003; van Gent et al. 2001). The *Arabidopsis* AtCPL gene was shown to have transcriptional regulatory functions by differentiation of the signal output that determines plant responses to stress, and AtCPL1 and AtCPL3 are negative regulators of stress responsive gene transcription and modulators of growth and development (Koiwa et al. 2002). Since this gene 5, named ToCPL1, cosegregates with *jointless-2*, we consider it a likely candidate gene for *jointless-2,* which controls the development of pedicel abscission zones. We confirmed the transcription of the intact ToCPL1 by RT-PCR using total RNA of immature flower bud (data not shown). The gene for *Jointless* was characterized and revealed as a MADS box transcription factor (Mao et al. 2000). The possible function of ToCPL1 as the *jointless-2* gene is currently under investigation.

Discussion

CAA trinucleotide microsatellite repeat block in (peri) centromeric region

Centromeres in most eukaryotic species analyzed so far contain tandem repeats that are AT-rich with motifs of 150–180 bp. In rice centromeres, CentO arrays range from 65 kb to 2 Mb in all 12 chromosomes (Cheng and Presting 2001). The finished sequence of rice chromosome 4, including its centromere, showed the centromeric region contained 379 copies of the CentO repeat (Feng et al. 2002; Zhang et al. 2004). Exceptions to this case have been identified in the functional centromeric region of the *Drosophila*. A minichromosome Dp1187 contains two penta-nucleotide microsatellite repeat arrays (AATAT and TTCTC) and a 39-kb complex of AT-rich repeats, with interruptions of five transposons (Sun et al. 1997, 2003). Also, (AGGGAG)*n* microsatellite arrays have been found in centromeric regions of barley (Hudakova et al. 2001; Jiang et al. 2003). GGAAT satellites and CATTT satellites have been found in pericentromere regions of human chromosome 10 (Guy et al. 2003). In tomato, Broun and Tanksley (1996) and Vosman and Arens (1997) demonstrated the clustering of GAT(C)A-containing microsatellite loci in pericentromere regions by genetic mapping. In beets (*Beta vulgaris*), Gindullis et al. (2001) showed the presence of GATA microsatellite repeats as components of all the centromeres using FISH analysis.

Cheng and Murata (2003) observed short stretches of CAA microsatellite repeats (16 copies) in the 250-bp centromere tandem repeats in wheat and postulated that the tandem repeats originated from centromeric retrotransposons such as *cereba*, and that the CAA microsatellite in the element could be a hotspot for recombination resulting in various copies of CAA repeats. In this report, we identified a large block of CAA repeats in the centromere and pericentromere region of chromosome 12 based on sequence and FISH analysis. The CAA repeats were also found in the pericentromere heterochromatin of most tomato chromosomes, but not all. These findings, as well as others, suggest that microsatellite arrays like the CAA blocks (GATA not be excluded) may be a component of tomato centromeres such as seen in *Drosophila* (Sun et al. 2003).

PCRT1 is the prominent component of the tomato pericentromere heterochromatin

Sequence and FISH analysis showed that the PCRT1 family is a major component of pericentromere heterochromatin blocks of tomato. We also identified that three previously characterized centromeric repeats appear to be components of the PCRT1 elements. Vosman and Arens (1997) mapped and sequenced several small insert clones containing various copies of GATA to tomato centromeric regions. In one GATA repeat clone, they identified a 372-bp sequence that was found to be present in about 4,300 copies in the tomato genome (referred to as U30 or TGR2). We found that one of the prominent tomato repeats, TGR2, is part of the LTR sequence of PCRT1a (denoted as red box no. 2 in Fig. 4b). In another GATA repeat clone, pWVA500 (Vosman and Arens 1997), which contains the GATA repeat, has a 56–80% identity with 400 bp of an internal sequence of the PCRT1c elements (red box no. 3 in Fig. 4b).

Other (peri)centromere sequences were described by Weide et al. (1998), who characterized the sequences of ten RAPD markers that were genetically mapped to the (peri) centromere region of tomato chromosome 6. Three of them show considerable similarity with the conserved 17-bp sequence of the CENP-B binding site and human centromeric satellite III and another two, AA13 and AG12, were shown to be repetitive. Interestingly, the sequence of AA13 (AF072522, 645 bp) shows significant similarity (60%) with the internal sequence of the PCRT1b element (red box 1 and dark gray box in Figs. 4b and d), the most prevalent

element of the sequenced 326-kb centromeric stretch with six members and unidentified truncated sequences (Figs. 1, 2 and 4). All six members of the PCRT1b, like AA13 (Weide et al. 1998), contain a sequence that is similar to a 17-bp conserved human CENP-B binding box (dark gray box of Fig. 4d). The maize centromeric retrotransposon, CRM, interacts with maize CENH3 (Zhong et al. 2002). This suggests that the sequences identified in the PCRT1b and AA13 elements may play a role in the function of the centromere/ kinetochore complex. Various examples of structural heterogeneity were found among the members of the PCRT1b family (Fig. 4d) including: truncation of one LTR (PCRT1b-2, 3, 4, 5, 6 in Fig. 4d), internal deletion (1b–2, 4), and duplication of some segments resembling LTR structures (PCRT1b–3, 5). Internal deletions or duplications found in PCRT2 members (Fig. 5) suggest that unequal crossing over may be the mechanism responsible for sequence variation seen within the ancient PCRT1b-like elements.

The PCRT1a elements are major components of pericentromere heterochromatin

Even though centromere-specific short tandem repeat units are variable and genus- or species-specific, centromeric retrotransposons show common features for all species in any one family (Cheng et al. 2002; Kurata et al. 2002). The large pericentromere heterochromatin blocks flanking the centromere are prerequisite for maintaining regional centromeres (Henikoff et al. 2001). The (peri)centromere organization of the 326-kb sequence shows that the PCRT1a elements expand into both directions of the pericentromere heterochromatin with different orientations (Fig. 2, start and end). Our pachytene FISH experiments revealed that the PCRT1a elements occupy most of the heterochromatin blocks (Fig 3a). So far, the most similar proteins to the PCRT1a polyprotein are identified in monocot plants. More than 100 copies of PCRT1a orthologs were identified in the rice pericentromere heterochromatin. The orthologs in rice, sorghum, and maize (GenBank accession numbers AAN01260, BAB90703 and AF448416, respectively) show over 40 and 60% similarities at the protein and DNA levels, respectively, with 3 kb of internal coding sequence of the PCRT1a. The results suggest that the PCRT1a-like elements are major components of pericentromere heterochromatin blocks in plant genomes.

We performed BLAST searches on the expressed sequence tag (EST) databases using the predicted exon sequence of polyproteins of the PCRT elements in order to the find transcriptional activity of each element. The predicted PCRT1a transcript is significantly similar to 28 various tomato ESTs (such as GenBank accession numbers BI 209546 and BI208878). The PCRT2 transcript was found to be similar to two tomato ESTs, AW033037 and BI205578, which were derived from callus and suspension cell culture, suggesting it might be active under stress conditions.

Genes in centromere regions: the locus of ToCPL1 suggests centromere emergence

Over 50% of tomato chromosome 12 is composed of large pericentromere heterochromatin blocks flanking the much smaller, functional centromere (Fig. 1 and Budiman et al. 2004). Recombination rates are highly suppressed in the whole region (generally more than 100 times less than euchromatin regions) (Budiman et al. 2004) as well as extremely low gene density as shown in this study. Only one putative gene and four hypothetical genes (seems not functional) were predicted in the 326-kb sequence. The gene density is quite low even in comparison with the *Arabidopsis* centromere (five and 12 genes in every 100 kb of centromere of chromosome 2 and 5, respectively; Copenhaver et al. 1999).

Despite the strikingly gene-deficient, recombination-deficient and repeat-rich properties of these chromosomal regions, various genes were discovered that specify important functions (Copenhaver et al. 1999; Entani et al. 1999). This is also true for the putative transcriptional regulator ToCPL1 gene that we describe here in the pericentromeric region of tomato chromosome 12. Studies on microcollinearity between tomato and the *Arabidopsis* genomes (Ku et al. 2000; Mao et al. 2001; Rossberg et al. 2001) revealed synteny between *Arabidopsis* and tomato; however, our analysis of the 326-kb sequence could not establish any relationship with the syntenic region in *Arabidopsis* because of the lack of unique genes/sequences. The coding region of ToCPL1 showed the highest similarity with an *Arabidopsis* gene, AtCPL3 (on BAC clone, F4P9, AC002332 and 60% of DNA sequence similarity; Koiwa et al. 2002), which is located in the euchromatin region of *Arabidopsis* chromosome 2, suggesting new centromere emergence at this region of tomato chromosome12 during evolution after speciation from a common ancestor 112–156 million years ago (Bowers et al. 2003). A more complete sequence of the tomato chromosome 12 and comparative study of the *Arabidopsis* genome will clarify the assumptions of centromere emergence at this position such as the finding of new centromere emergence in maize chromosome 4 (Page et al. 2001) and in the X chromosome of *Lemur catta*, which is perfectly collinear with the human X chromosome (Ventura et al. 2001, 2004; Wong and Choo 2001).

Acknowledgements This work was supported by the NSF Plant Genome Grant # 0116076 to R.A.W.

References

Ananiev EV, Phillips RL, Rines HW (1998) Chromosome-specific molecular organization of maize (*Zea mays* L.) centromeric regions. Proc Natl Acad Sci U S A 95:13073–13078

Barton DW (1950) Pachytene morphology of the tomato chromosome complement. Am J Bot 37:639–643

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature 422:433–438

Broun P, Tanksley SD (1996) Characterization and genetic mapping of simple repeat sequences in the tomato genome. Mol Gen Genet 250:39–49

Budiman MA, Chang SB, Lee S, Yang TJ, Zhang HB, de Jong JH, Wing RA (2004) Localization of jointless-2 gene in the centromeric region of tomato chromosome 12 based on high resolution genetic and physical mapping. Theor Appl Genet 108:190–196

Cheng Z, Murata M (2003) A centromeric tandem repeat family originating from a part of Ty3/gypsy-retroelement in wheat and its relatives. Genetics 164:665–672

Cheng Z, Presting GG, Buell CR, Wing RA, Jiang J (2001) High-resolution pachytene chromosome mapping of bacterial artificial chromosomes anchored by genetic markers reveals the centromere location and the distribution of genetic recombination along chromosome 10 of rice. Genetics 157:1749–1757

Cheng Z, Dong F, Langdon T, Ouyang S, Buell CR, Gu M, Blattner FR, Jiang J (2002) Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. Plant Cell 14:1691–1704

Clarke L (1998) Centromeres: proteins, protein complexes, and repeated domains at centromeres of simple eukaryotes. Curr Opin Genet Dev 8:212–218

Copenhaver GP, Nckel K, Kuromori T, Benito M, Kaul S, Lin X, Bvan M, Murphy G, Harris B, Parnell LD, McCombie WR, Martienssen RA, Marra M, Preuss D (1999) Genetic definition and sequence analysis of *Arabidopsis* centromeres. Science 286:2468–2474

Dong F, Miller JT, Jackson SA, Wang GL, Ronald PC, Jiang J (1998) Rice (*Oryza sativa*) centromeric regions consist of complex DNA. Proc Natl Acad Sci U S A 95:8135–8140

Entani T, Iwano M, Shiba H, Takayama S (1999) Centromeric localization of an S-RNase gene in *Petunia hybrida* Vilm. Theor Appl Genet 99:391–397

Ewing B, Green P (1998) Base-calling of automated sequencer traces using Phred II. Error probabilities. Genome Res 8:186–194

Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using PHRED. I. Accuracy assessment. Genome Res 8:175–185

Feng Q, Zhang Y, Hao P, Wang S, Fu G, Huang Y, Li Y, Zhu J, Liu Y, Hu X et al (2002) Sequence and analysis of rice chromosome 4. Nature 420:316–320

Frary A, Presting GG, Tanksley S (1996) Molecular mapping of the centromeres of tomato chromosomes 7 and 9. Mol Gen Genet 250:295–304

Fukui KN, Suzuki G, Lagudah ES, Rahman S, Appels R, Yamamoto M, Mukai Y (2001) Physical arrangement of retrotransposon-related repeats in centromeric regions of wheat. Plant Cell Physiol 42:189–196

Ganal MW, Lapitan N, Tanksley SD (1991) Macrostructure of the tomato telomeres. Plant Cell 3:87–94

Ganal MW, Broun P, Tanksley SD (1992) Genetic mapping of tandemly repeated telomeric DNA sequences in tomato (*Lycopersicon esculentum*). Genomics 14:444–448

Gindullis F, Desel C, Galasso I, Schdidt T (2001) The large-scale organization of the centromeric region in *Beta* species. Genome Res 11:253–265

Gordon D, Abajian C, Green P (1998) Consed: a graphical tool for sequence finishing. Genome Res 8:195–202

Guy J, Hearn T, Crosier M, Mudge J, Viggiano L, Koczan D, Thiesen HJ, Bailey JA, Horvath JE, Eichler EE, Earthrowl ME, Deloukas P, French L, Rogers J, Benley D, Jackson MS (2003) Genomic sequence and transcriptional profile of the boundary between pericentromere satellites and genes on human chromosome arm 10p. Genome Res 13:159–172

Harrison GE, Heslop-Harrison JS (1995) Centromeric repetitive DNA sequences in the genus *Brassica*. Theor Appl Genet 90: 157–165

Henikoff S (2002) Near the edge of a chromosome's black hole. Trends Genet 18:165–167

Henikoff S, Ahmad K, Malik HS (2001) The centromere paradox: stable inheritance with rapidly evolving DNA. Science 293: 1098–1102

Henning KA, Novotny EA, Compton ST, Guan XY, Liu PP, Ashlock MA (1999) Human artificial chromosomes generated by modification of a yeast artificial chromosome containing both human alpha satellite and single-copy DNA sequences. Proc Natl Acad Sci U S A 96:592–597

Hudakova S, Michalek W, Presting GG, ten Hoopen R, dos Santos K, Jasencakova Z, Schubert I (2001) Sequence organization of barley centromeres. Nucleic Acids Res 29:5029–5035

Jiang N, Wessler SR (2001) Insertion preference of maize and rice miniature inverted repeat transposable elements as revealed by the analysis of nested elements. Plant Cell 13:2553–2564

Jiang J, Birchler JA, Parrott WA, Dawe RK (2003) A molecular view of plant centromeres. Trends Plant Sci 8:570–575

Jin W, Melo JR, Nagaki K, Talbert PB, Henikoff S, Dawe RK, Jiang J (2004) Maize centromeres: organization and functional adaptation in the genetic background of oat. Plant Cell 16:571–581

Koiwa H, Barb AW, Xiong L, Li F, McCully MG, Lee BH, Sokolchik I, Zhu J, Gong Z, Reddy M, Sharkhuu A, Manabe Y, Yokoi S, Zhu JK, Bressan RA, Hasegawa PM (2002) C-terminal domain phosphatase-like family members (AtCPLs) differentially regulate *Arabidopsis thaliana* abiotic stress signaling, growth, and development. Proc Natl Acad Sci U S A 99: 10893–10898

Ku HM, Vision T, Liu J, Tanksley SD (2000) Comparing sequenced segments of the tomato and Arabidopsis genomes: large-scale duplication followed by selective gene loss creates a network of synteny. Proc Natl Acad Sci U S A 97:9121–9126

Kulikova O, Geurts R, Lamine M, Kim DJ, Cook DR, Leunissen J, de Jong JH, Roe BA, Bisseling T (2004) Satellite repeats in the functional centromere and pericentromeric heterochromatin of *Medicago truncatula*. Chromosoma 113:276–283

Kumekawa N, Hosouchi T, Tsuruoka H, Kotani H (2000) The size and sequence organization of the centromeric region of *Arabidopsis thaliana* chromosome 5. DNA Res 7:315–321

Kumekawa N, Hosouchi T, Tsuruoka H, Kotani H (2001) The size and sequence organization of the centromeric region of *Arabidopsis thaliana* chromosome 4. DNA Res 8:285–290

Kurata N, Nonomura KI, Harushima Y (2002) Rice genome organization: the centromere and genome interactions. Ann Bot 90:427–435

Lafarge S, Montané MH (2003) Caracterization of *Arabidopsis thaliana* ortholog of the human breast cancer susceptibility gene 1: AtBRCA1, strongly induced by gamma rays. Nucleic Acids Res 31:1148–1155

Lamb JC, Birchler JA (2003) The role of DNA sequence in centromere formation. Genome Biol 4:214

Mao L, Begum D, Chuang HW, Budiman MA, Szymkowiak EJ, Irish EE, Wing RA (2000) JOINTLESS is a MADS-box gene controlling tomato flower abscission zone development. Nature 406:910–913

Mao L, Begum D, Goff SA, Wing RA (2001) Sequence and analysis of the tomato JOINTLESS locus. Plant Physiol 126:1331–1340

Nagaki K, Song J, Stupar M, Parokonny AS, Yuan Q, Ouyang S, Liu J, Hsiao J, Jones KM, Dawe RK, Buell CR, Jiang J (2003a) Molecular and cytological analysis of large tracks of centromeric DNA reveal the structure and evolutionary dynamics of maize centromeres. Genetics 163:759–770

Nagaki K, Talbert PB, Zhong CX, Dawe RK, Henikoff S, Jiang J (2003b) Chromatin immunoprecipitation reveals that the 180 bp satellite repeat is the key functional DNA element of *Arabidopsis thanlina* centromeres. Genetics 163:1221–1225

Page BT, Wanous MK, Birchler JA (2001) Characterization of a maize chromosome 4 centromeric sequence: evidence for an evolutionary relationship with the B chromosome centromere. Genetics 159:291–302

Parsons JD (1995) Miropeats: graphical DNA sequence comparisons. Comput Appl Biosci 11:615–619

Presting GG, Frary A, Pillen K, Tanksley SD (1996) Telomere-homologous sequences occur near the centromeres of many tomato chromosomes. Mol Gen Genet 251:526–531

Rossberg M, Theres K, Acarkan A, Herrero R, Schmitt T, Schumacher K, Schmitz G, Schmidt R (2001) Comparative sequence analysis reveals extensive microcolinearity in the lateral suppressor regions of the tomato, Arabidopsis, and Capsella genomes. Plant Cell 13:979–988

Round EK, Flowers SK, Richards EJ (1997) Arabidopsis thaliana centromere regions: genetic map positions and repetitive DNA structure. Genome Res 7:1045–1053

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W (2000) PipMaker—a web server for aligning two genomic DNA sequences. Genome Res 10:577–586

Stegemann S, Hartmann S, Ruf S, Bock R (2003) High-frequency gene transfer from the chloroplast genome to the nucleus. Proc Natl Acad Sci U S A 100:8828–8833

Sun X, Wahlstrom J, Karpen G (1997) Molecular structure of a functional Drosophila centromere. Cell 91:1007–1019

Sun X, Le HD, Wahlstrom JM, Karpen GH (2003) Sequence analysis of a functional Drosophila centromere. Genome Res 13:182–194

Thompson H, Schmidt R, Brandes A, Heslop-Harrison JS, Dean C (1996) A novel repetitive sequence associated with the centromeric regions of *Arabidopsis thaliana* chromosomes. Mol Gen Genet 253:247–252

van der Hoeven R, Ronning C, Giovannoni J, Martin G, Tanksley S (2002) Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. Plant Cell 14:1441–1456

van Gent DC, Hoeijmakers JH, Kanaar R (2001) Chromosomal stability and the DNA double-stranded break connection. Nature Rev Genet 2:196–206

Ventura M, Archidiacono N, Rocchi M (2001) Centromere emergence in evolution. Genome Res 11:595–599

Ventura M, Weigl S, Carbone L, Cardone MF, Misceo D, Teti M, D'Addabbo P, Wandall A, Bjorck E, de Jong PJ, She X, Eichler EE, Archidiacono N, Rocchi M (2004) Recurrent sites for new centromere seeding. Genome Res 14:1696–1703

Vosman B, Arens P (1997) Molecular characterization of GATA/GACA microsatellite repeats in tomato. Genome 40:25–33

Weide R, Hontelez J, van Kammen A, Koorneef M, Zabel P (1998) Paracentromeric sequences on tomato chromosome 6 show homology to human satellite III and to the mammalian CENP-B binding box. Mol Gen Genet 259:190–197

Witte CP, Le QH, Bureau T, Kumar A (2001) Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci U S A 98:13778–13783

Wong LH, Choo KHA (2001) Centromere on the move. Genome Res 11:513–516

Yang TJ, Yu Y, Nah GJ, Atkins M, Lee S, Frisch DA, Wing RA (2003) Construction and utilities of 10 kb libraries for efficient clone-gap closure for rice genome sequencing. Theor Appl Genet 107:652–660

Yang TJ, Yu Y, Frisch D, Lee S, Kim HR, Kwon SJ, Park BS, Wing RA (2004) Construction of various copy number plasmid vectors and their utility for genome sequencing. Genomics & Informatics 2:153–158

Yang TJ, Yu Y, Lee S, Chang SB, Ahn SN, de Jong JH, Wing RA (2005) Toward finishing rice telomere gap: mapping and sequencing of rice subtelomere regions. Theor Appl Genet (in press)

Yu Y, Rambo T, Currie J, Saski C, Kim HR, Collura K, Thompson S, Simmons J, Yang TJ, Park GN, Patel AJ et al (2003) In-depth view of structure, activity, and evolution of rice chromosome 10 (The Rice Chromosome 10 Sequencing Consortium). Science 300:1566–1569

Zhang HB, Budiman MA, Wing RA (2000) Genetic mapping of jointless-2 to tomato chromosome 12 using RAPD and RFLP analysis. Theor Appl Genet 100:1183–1189

Zhang Y, Huang Y, Zhang L, Li Y, Lu T, Lu Y, Feng Q, Zhao Q, Cheng Z, Xue Y, Wing RA, Han B (2004) Structural features of the rice chromosome 4 centromere. Nucleic Acids Res 32:2023–2030

Zhong XB, de Jong JH, Zabel P (1996) Preparation of tomato meiotic pachytene and mitotic metaphase chromosomes suitable for fluorescence in situ hybridization (FISH). Chromosome Res 4:24–28

Zhong XB, Fransz PF, Wennekes-Eden J, Ramanna MS, van Kammen A, Zabel P, de Jong JH (1998) FISH studies reveal the molecular and chromosomal organization of individual telomere domains in tomato. Plant J 13:507–517

Zhong CX, Marshall JB, Topp C, Mroczek R, Kato A, Nagaki K, Birchler JA, Jiang J, Dawe RK (2002) Centromeric retroelements and satellites interact with maize kinetochore protein CENH3. Plant Cell 14:2825–2836