Mol Genet Genomics (2009) 281:579–590
DOI 10.1007/s00438-009-0433-y

ORIGINAL PAPER

# Patterns of tandem repetition in plant whole genome assemblies

**Rafael Navajas-Pérez · Andrew H. Paterson**

Received: 4 November 2008 / Accepted: 3 February 2009 / Published online: 26 February 2009


**Abstract** Tandem repeats often confound large genome assemblies. A survey of tandemly arrayed repetitive sequences was carried out in whole genome sequences of the green alga *Chlamydomonas reinhardtii*, the moss *Physcomitrella patens*, the monocots rice and sorghum, and the dicots *Arabidopsis thaliana*, poplar, grapevine, and papaya, in order to test how these assemblies deal with this fraction of DNA. Our results suggest that plant genome assemblies preferentially include tandem repeats composed of shorter monomeric units (especially dinucleotide and 9–30-bp repeats), while higher repetitive units pose more difficulties to assemble. Nevertheless, notwithstanding that currently available sequencing technologies struggle with higher arrays of repeated DNA, major well-known repetitive elements including centromeric and telomeric repeats as well as high copy-number genes, were found to be reasonably well represented. A database including all tandem repeat sequences characterized here was created to benefit future comparative genomic analyses.

**Keywords** Tandem repeats · Whole genome assemblies

## Introduction

Plant genomes contain large quantities of repetitive DNA sequences, reaching in some cases up to 97% of total nuclear DNA content (Flavell et al. 1974; Murray et al. 1981). Due to technical difficulties this portion is often not fully reflected or even largely ignored in genomic assemblies. Although the cost of generating a genomic sequence continues to decrease, refining that sequence by the process of "sequence finishing" remains expensive. Near-perfect finished sequence is an appropriate goal for a small set of reference genomes; however, such a high-quality product cannot be cost-justified for large numbers of additional genomes, at least for the foreseeable future (Blakesley et al. 2004). Also in many cases, coding regions are specifically sequenced to the detriment of the repetitive fraction, as is the case in 454 technology projects (Margulies et al. 2005) or because the large amount of repetitive DNA complicates assembly. In some cases, genome-filtration techniques are used in order to avoid repetitive DNA and to enrich gene sequences in genomic libraries. These methods have been suggested to provide a low-cost alternative to whole genome sequencing for complex genomes such as maize (Okagaki and Phillips 2004) and many others (Peterson et al. 2002). Here, by analyzing the genomes of eight representative plants, we aim to test how different sequencing and assembly approaches affect the representation of TR sequences in large genomic datasets.

The vast majority of the plant repeatome is comprised of transposable elements, mainly LTR retrotransposons (Bennetzen 2002; Bergman and Quesneville 2007) that constitute a high percentage of total genome size, as in maize (58%, Messing et al. 2004), papaya (52%, Nagarajan et al. 2008), rice (35%, International Rice Genome Sequencing Project 2005) or *Arabidopsis thaliana* (14%, *Arabidopsis thaliana* Genome Initiative 2001). However, despite the predominance of transposable sequences, the portion of the repetitive fraction that poses more problems to assemble is by far tandemly arrayed repeats (TR). TR are normally categorized into micro-, mini- or satellite-DNA

Communicated by Y. Van de Peer.

R. Navajas-Pérez (✉) · A. H. Paterson
Plant Genome Mapping Laboratory,
University of Georgia, Athens, GA 30602, USA
e-mail: rnavajas@ugr.es

sequences and appear preferentially in the centromeric, telomeric, and subtelomeric regions of many eukaryotes, comprising hundreds or thousands of repeats (Kubis et al. 1998; Ugarković and Plohl 2002). These sequences are also found at interspersed positions and may play an important role in sex-chromosome (Navajas-Pérez et al. 2006) and B-chromosome evolution (Camacho et al. 2000), as major constituents of heterochromatin (Elder and Turner 1995). Also, significant proportions of plant protein coding genes occur in tandem arrays (e.g. *A. thaliana*, 16%, Rizzon et al. 2006; *O. sativa*, 14%, International Rice Genome Sequencing Project 2005; *P. trichocarpa*, 11%, Tuskan et al. 2006).

The study of repetitive sequence elements has been essential to our understanding of the nature and consequences of genome size variation between different species, and for studying the large-scale organization and evolution of plant genomes. In this context, the present paper describes a survey of TR performed in eight plant genome sequences, broadly sampling the plant evolutionary tree (green alga *Chlamydomonas reinhardtii*, moss *Physcomytrella patens*, monocots rice and sorghum, and dicots *Arabidopsis thaliana*, poplar, grapevine and papaya). For this task, we have used common computing tools (Tandem Repeats Finder, BLAST, CD-HIT, PlantSat database) in order to clarify the impact of different whole genome project approaches on TR characterization and possible underlying biological reasons and technical issues, as well as to test their usability for the study of this particular region of the genome.

## Materials and methods

### Tandem repeat detection

Eight completed plant whole genome sequences, including large main scaffolds or pseudo-chromosomes when possible (Table 1) were explored for TR by using the Tandem Repeats Finder software (Benson 1999), according to the following parameters: 2, 7, 7, 80, 10, 50, 2,000 match, mismatch, indels, matching probability, indel probability, minimum alignment score, maximum period size. Repeats were classified into micro- (1–6 bp), mini- (7–100 bp) and satellite (>100 bp) tandemly arrayed sequences. The program cd-hit-est, as implemented in the package CD-HIT (Li and Godzik 2006), was used to construct non-redundant sets of sequences at 85% of identity and 5 of word-length. For annotation, sequences were BLASTed using BLASTn against the CDS sequences of *A. thaliana* (TAIR 7 release, Poole 2007) and the hits classified according to the MIPS functional catalogue database (http://mips.gsf.de), and against PlantSat DB (http://w3lamc.umbr.cas.cz/PlantSat/—February 2008 version—Macas et al. 2002) to investigate the presence of satellite-DNA families. Best hits at 0.01 e-value cutoff were considered.

### Data access and retrieval

The TR sequences described in this paper are available at http://www.plantgenome.uga.edu/tandemrepeats/. Redundant and non-redundant databases are available in multi-fasta

**Table 1** List of genomes analyzed, tandem repeat composition, and AT content

| Species | Affiliation | Size (Mb) | References | Microsatellites (%) | Minisatellites (%) | Satellites (%) | Total (%) | AT content (%) |
|---|---|---|---|---|---|---|---|---|
| *Arabidopsis thaliana* | Magnoliopsida, Brassicales, dicot | 120 | *Arabidopsis thaliana* Genome Initiative (2001) | 0.18 | 0.98 | 0.56 | 1.72 | 70.14 |
| *Carica papaya* | Magnoliopsida, Brassicales, dicot | 372 | Ming et al. (2008) | 0.19 | 0.68 | 0.43 | 1.3 | 72 |
| *Populus trichocarpa* | Magnoliopsida, Malpighiales, dicot | 550 | Tuskan et al. (2006) | 0.16 | 0.71 | 0.35 | 1.22 | 76.5 |
| *Vitis vinifera* | Magnoliopsida, Vitales, dicot | 475–505 | Jaillon et al. (2007) | 0.19 | 0.93 | 0.45 | 1.57 | 76 |
| *Oryza sativa* | Liliopsida, Poales, monocot | 420 | International Rice Genome Sequencing Project (2005) | 0.21 | 1.57 | 0.53 | 2.31 | 58.45 |
| *Sorghum bicolor* | Liliopsida, Poales, monocot | 735 | Paterson et al. (2009) | 0.19 | 0.87 | 2.66 | 3.72 | 59.2 |
| *Physcomitrella patens* | Bryopsida, Funariales, moss | 454 | Rensing et al. (2008) | 0.67 | 1.32 | 0.08 | 2.07 | 78.37 |
| *Chlamydomonas reinhardtii* | Chlorophyceae, Volvocales, green alga | 100 | Merchant et al. (2007) | 1.76 | 2.27 | 0.45 | 4.48 | 31.3 |

files, a format convenient for use with RepeatMasker. Further details can be found in the README file accompanying the database.

## Results

The genomes of six angiosperms (2 monocots and 4 dicots), a moss and a green alga (see Table 1), were scanned for tandemly repetitive elements using the Tandem Repeats Finder software (Benson 1999). TR amount was variable among the species analyzed, ranging from an average of 1.45 and 3% in dicots and monocots, respectively, to 2.07 and 4.48% of the moss and the green alga (Table 1). The average AT richness was high in dicots and moss (~74 and 78.4%, respectively). Monocots and the green alga showed a lower AT content, with an average of ~59 and 31.3%, respectively (Table 1).

According to the repeat-unit size, tandem repeats were assigned to one of three classes: microsatellites (1–6 bp), minisatellites (7–100 bp), and satellites (>100 bp). Figure 1 illustrates the sizes and abundance of tandem repeats in all eight analyzed plant genomes. In general, most represented families are those with smaller monomeric repetitive units, especially dinucleotides and repeats ranging from 5 to 7 and 9–30 bp long. In all cases the number of microsatellite repeats was higher, but due to their greater monomeric repeat length, mini- and satellite-DNAs comprised a higher percentage of the respective genomes (Table 1). Microsatellites constitute ~0.19% of angiosperm assemblies, versus 0.67 and 1.76% of moss and green alga genomes, respectively. Minisatellites are abundant in general with 0.83, 1.22, 1.32 and 2.27% of average for dicots, monocots, moss, and green alga, respectively. Satellites constitute in all cases ~0.5% of the assemblies except the extreme situations of sorghum and *Chlamydomonas*, 2.66 and 0.08%, respectively (Table 1). Nevertheless, in all cases except sorghum (Fig. 1f), repeats with monomeric units >100 bp seem underrepresented with respect to the other repeats (Fig. 1). As described below, the predominance of satellite-DNA sequences in sorghum is mainly due to the large number of copies of a centromeric 137-bp repeat.

As for microsatellites, in all studied dicots and the moss *Physcomitrella*, stretches of (A/T)n were fairly common, while small quantities of mononucleotide repeats were detected in monocots and the green alga *Chlamydomonas* (Fig. 1). In all species analyzed, dinucleotides were the most numerous microsatellite class with AT/TA accounting in all species for more than 40% of total microsatellite repeats. (AG/TC/CT/GA)n were also common in papaya, and (GT/CA/AC/TG)n were predominant in *Chlamydomonas*. The presence of the most common triplets was also investigated (Fig. 2). Repeats of AAT/TAA/ATT/TAA and TAT/ATA are abundant in poplar, grapevine and sorghum, being the former one the most represented in papaya, while AAG/TTC/CTT/GAA, TCT/AGA and TAC/ATG/GTA/CAT are commonly found in *A. thaliana*, and papaya genomes (Fig. 2). Interestingly, stretches of GGC/CCG/GCC/CGG and CGC/GCG are well represented in sorghum, rice, and *Chlamydomonas*, being almost absent in the other species analyzed. Additionally, *Chlamydomonas* has predominance of GCA/CGT/TGC/ACG, GCT/CGA/AGC/TCG and CAG/GTC/CTG/GAC repeats, found in low-copy number in the other species. Trinucleotide repeats are relatively infrequent in *Physcomitrella*. However, when they appear they tend to be those most represented in dicots (Fig. 2).

Mini- and satellite TR were screened by BLAST against PlantSat DB (Macas et al. 2002), that includes an updated list of satellite-DNA sequences isolated in plants. When the species were represented in the PlantSat database, the detection of all reported satellite-DNA families was possible (Table 2). This is the case for *A. thaliana*, for which we detected the presence of the centromeric AluI, 180 and AR3 satellites and telomere-like 500 repeat (Simoens et al. 1988), as well as the pericentromeric 1360, and IID2_8 families (Tutois et al. 1999). In poplar we detected the presence of the centromeric-like 145-bp repeat (Rajagopal et al. 1999). For rice, the centromeric repeat CentO (Dong et al. 1998) and the interspersed 150, TrsA and 880 repeats (Wu and Wu 1987; Wu et al. 1991) were detected. In sorghum, the centromeric family CEN38 (Miller et al. 1998) was abundant. As for species not represented in the PlantSat database; in papaya we detected some repeats showing homology with telomere-like *A. thaliana* 500 repeat, while the moss *Physcomitrella patens* repeats showed homology to telomere-like *A. thaliana* 500 and *Sinapis arvensis* 700 repeats, and also to rDNA-like *Anemone blanda* AbS1 family (Hagemann et al. 1993; Kapila et al. 1996). No significant matches were found under our experimental conditions for grapevine and the green alga *C. reinhardtii* (see Table 2 for details).

We also BLASTed mini- and satellite repeats against the TAIR 7 *A. thaliana* annotation release, in order to annotate them when possible. Sequences analyzed fell mainly into four categories according to the MIPS functional catalogue database (Table 3). The most represented categories corresponded to unclassified proteins, transposable elements, genes involved in metabolism (especially binding elements and transcription factors), and defense and virulence related processes (Table 3).

Finally, for all angiosperms and the moss *P. patens*, we found a moderate number of copies of the plant telomeric sequence (TTTAGGG)n, while the (TTTTAGGG)n motif was found in *C. reinhardtii*. In all cases analyzed, we also

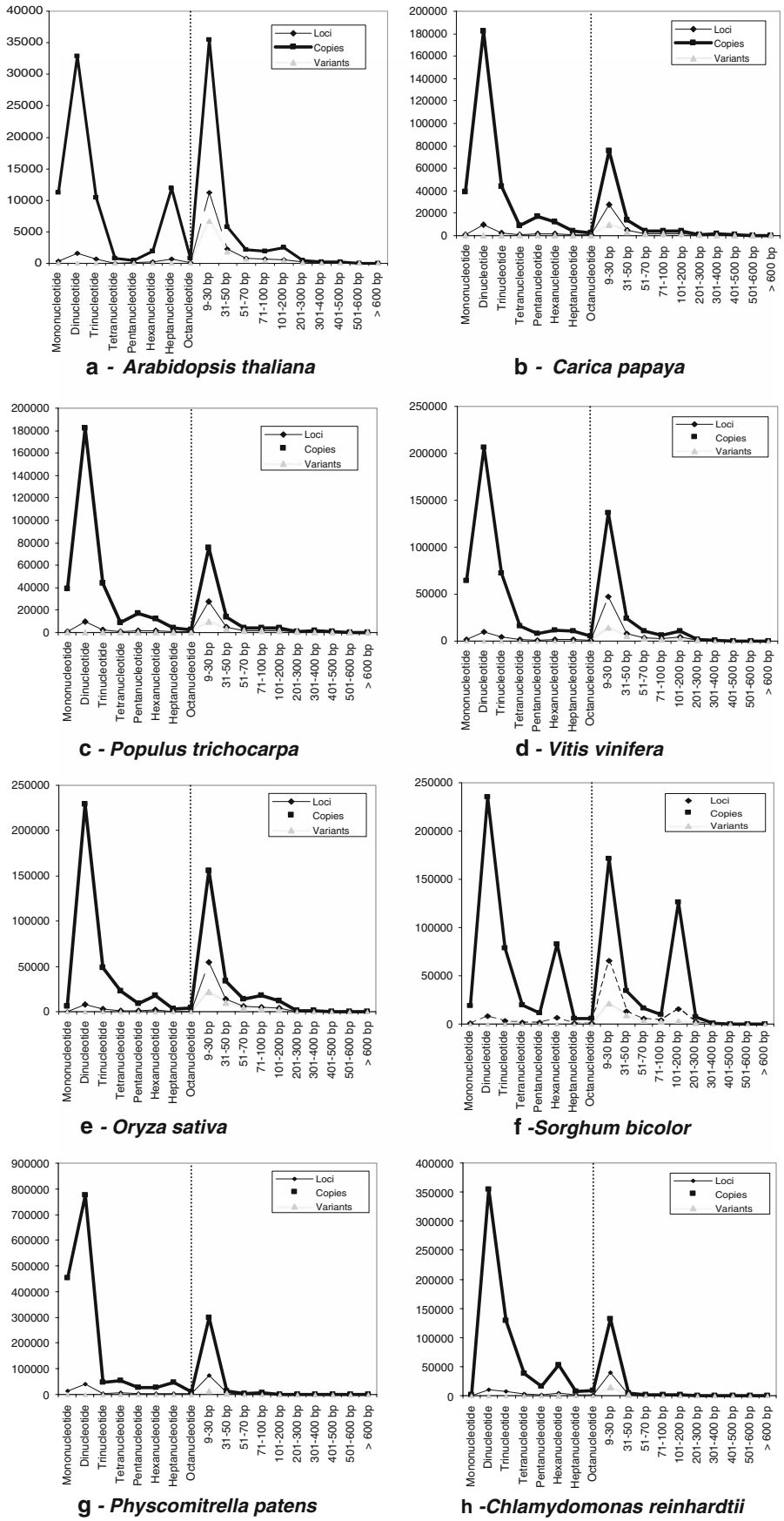


**Fig. 1** Distribution of size and abundance of tandem repeats in the eight genomes analyzed. *Vertical broken lines* reflect a scale change in the graphic. Notes: variants being each of non-redundant monomeric types. Scale change from 1-unit window size to 20/100-units window size

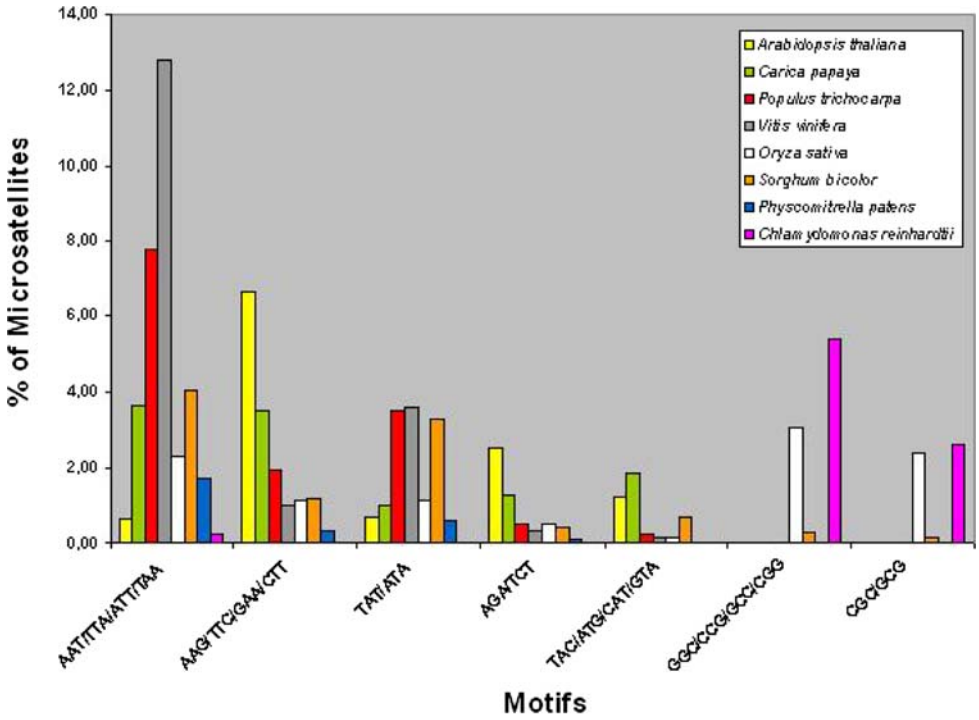


**Fig. 2** Percentage of the most represented triplets in the eight genomes analyzed

detected several telomeric-like hexa- and heptameric variants, which are sufficiently abundant to be discerned as visible peaks in *A. thaliana* and sorghum repeat distributions (Fig. 1a, f).

## Discussion

There are many examples in the literature of estimations for plants TR-DNA. Punctual quantifications of satellite-DNA sequences reveal families comprising just a few copies, but more frequently cases in which these sequences constitute a large portion of the genome; for example, the centromeric satellites of *Brassica rapa* encompass about 30% of the total chromosomes (Lim et al. 2005), while satellite-DNAs comprise up to 20% of citrus plant genome (Fann et al. 2001). Other estimations reveal that micro- and minisatellites can be copious; for example, microsatellites constitute 2% of the soybean genome (Saini et al. 2008), while the family OPG9-130 made by 15-bp minisatellite repeats, represents 1.5% of bean genome (Métais et al. 1998). All these data together indicate that TR may be underrepresented in the assemblies analyzed here (ranging from 1.22% for poplar to 4.48% for *C. reinhardtii*, Table 1). However, we demonstrate that most known repetitive elements and high copy-number genes are found to be reasonably well represented and further analyses can be performed.

In the present paper we have studied microsatellites belonging to the class I. This class is especially interesting for our purposes due to their size (>20 bp) and their hypervariability, two factors that may complicate the assembly. Although previous analyses use slightly different experimental conditions and datasets, it is possible to draw some common conclusions by contrasting them with our results. As for the repeat size, we have found that in all analyzed species dinucleotide is the most common repetition unit, trimer being the second one. This has been also observed by other authors; in rice, 72% of the microsatellites longer than 30 bp were dinucleotides (La Rota et al. 2005). Class-I dinucleotides were also the most common type in papaya (Nagarajan et al. 2008; Wang et al. 2008). This is also the general case of *Physcomitrella* and other algal and plant genomes (von Stackelberg et al. 2006).

Some monomeric repetitions have been proved to be more frequent than other. We found AT the most frequent dinucleotide repetition in the genomes analyzed. Previous reports have demonstrated the same in rice, *Mesostigma*, *Ginkgo*, *Picea*, *Pinus*, *Gossypium*, *Solanum*, *Allium* or papaya (La Rota et al. 2005; von Stackelberg et al. 2006; Nagarajan et al. 2008). Figure 2 shows the abundance of the most common triplets in the analyzed genomes; repeats of AAT/TAA/ATT/TAA are abundant in poplar, grapevine, papaya and sorghum. These repeats are also predominant in wheat (Song et al. 2002), tomato (Smulders et al. 1997), soybean (Akkaya et al. 1995) and *A. thaliana* (Loridon et al. 1998). AAG/TTC/CTT/GAA, TCT/AGA and TAC/ATG/GTA/CAT are commonly found in *A. thaliana* and papaya genomes. This was also found by other authors (Depeiges et al. 1995; Nagarajan et al. 2008).

**Table 2** Mini- and satellite-DNA BLAST hits summary in PlantSat database (Macas et al. 2002)

| Family | Hits | Estimated % | Location | Monomeric length (bp) | References |
|---|---|---|---|---|---|
| *Arabidopsis thaliana* | | | | | |
| Arabidopsis_thaliana_180 | 78 | 0.8–1.4 | Centromeric | 178 | Simoens et al. (1988) |
| Arabidopsis_thaliana_AR3 | 47 | 0.2–0.4 | Unknown | 159 | Simoens et al. (1988) |
| Arabidopsis_thaliana_1360 | 26 | Unknown | Pericentromeric | 1,364 | Tutois et al. (1999) |
| Arabidopsis_thaliana_IID2_8 | 7 | Unknown | Pericentromeric | 675 | Tutois et al. (1999) |
| Arabidopsis_thaliana_500 | 6 | 0.2–0.4 | Telomere-like | 500 | Simoens et al. (1988) |
| Arabidopsis_arenosa_180 | 5 | 0.8–1.4 | Centromeric | 178 | Kamm et al. (1995) |
| Arabidopsis_thaliana_AluI | 4 | Unknown | Centromeric | 147 | Brandes (unpublished) |
| Zea_mays_MBsC216 | 1 | Unknown | Centromeric | 216 | Zhang et al. (unpublished) |
| Grand total | 174 | | | | |
| *Carica papaya* | | | | | |
| Family | | | | | |
| Arabidopsis_thaliana_500 | 5 | 0.2–0.4 | Telomere-like | 500 | Simoens et al. (1988) |
| Grand total | 5 | | | | |
| *Populus trichocarpa* | | | | | |
| Family | | | | | |
| Populus_145 | 132 | 1.50 | Unknown (centromere-like) | 145 | Rajagopal et al. (1999) |
| Zea_mays_MBsC216 | 1 | Unknown | Centromeric | 216 | Zhang et al. (unpublished) |
| Grand total | 133 | | | | |
| *Vitis vinifera* | | | | | |
| Hitless | | | | | |
| *Oryza sativa* | | | | | |
| Family | | | | | |
| Oryza_sativa_CentO | 210 | Unknown | Centromeric | 159 | Dong et al. (1998) |
| Oryza_150 | 22 | 0.01–1.97 | Unknown | 150 | Liang et al. (unpublished) |
| Oryza_TrsA | 21 | Unknown | Interspersed (5S rDNA and tRNA-like) | 354 | Wu and Wu (1987) |
| Oryza_sativa_880 | 5 | Unknown | Interspersed | 878 | Wu et al. (1991) |
| Hordeum_vulgare_Crep2 | 2 | Unknown | Centromeric | 238 | Aragón-Alcaide et al. (1996) |
| Grand total | 260 | | | | |
| *Sorghum bicolor* | | | | | |
| Family | | | | | |
| Sorghum_CEN38_Sau3A10 | 6829 | 1.6–1.9 | Centromeric | 137 | Miller et al. (1998) |
| Saccharum_SCEN | 147 | 0.60 | Centromeric | 135 | Nagaki et al. (1998) |
| Grand Total | 6976 | | | | |
| *Physcomitrella patens* | | | | | |
| Family | | | | | |
| Anemone_blanda_AbS1 | 2 | 2 | Interspersed (25S rDNA-like) | 1639 | Hagemann et al. (1993) |
| Arabidopsis_thaliana_500 | 2 | 0.2–0.4 | Telomere-like | 500 | Simoens et al. (1988) |
| Sinapis_arvensis_700 | 1 | Unknown | Interspersed (telomere-like) | 697 | Kapila et al. (1996) |
| Grand total | 5 | | | | |
| *Chlamydomonas reinhardtii* | | | | | |
| Hitless | | | | | |

TAT/ATA is almost equally represented in grapevine, poplar and sorghum. Interestingly, stretches of GGC/CCG/GCC/CGG and CGC/GCG are well represented in sorghum, rice, and *Chlamydomonas*, being almost absent in the other species analyzed. This agrees with previous analyses in different grasses genomes (Morgante et al. 2002; La Rota et al. 2005; von Stackelberg et al. 2006). Additionally, *Chlamydomonas* has predominance of GCA/CGT, GCT/

CGA, and CAG/GTC repeats, found in low-copy number in the other species. In contrast, (AAC/TTG)n and (ACC/TGG)n that account for 84.5% of eggplant microsatellites (Nunome et al. 2003), seem to be underrepresented in the species analyzed here, with only a few copies.

According to this survey, AT-rich microsatellites seem to be frequent in dicots (*A. thaliana*, papaya, grapevine, poplar) and *Physcomitrella*, while GC-rich microsatellites are more predominant in monocots (i.e. rice/sorghum) and *Chlamydomonas*. AT richness has been suggested for dicots, based on analyses on *Arabidopsis*, papaya, and legumes TR (Cardle et al. 2000; Mun et al. 2006; Nagarajan et al. 2008). On the other hand, GC-richness has been suggested for monocots and *Chlamydomonas* TR (Morgante et al. 2002; von Stackelberg et al. 2006). This GC bias has been suggested to result from genomic repeat presence, interspecific genome size variation, chromosome size evolution, and a long-term cycle of GC-rich retrotransposon proliferation and removal (Delseny 2003).

Interestingly, according to microsatellite data the green alga *Chlamydomonas* more closely resembles monocots while the moss *Physcomitrella* more closely resembles dicots. These data do not imply relatedness of the organisms but of the feature analyzed—indeed, species belonging to related phylogenetic groups did not reveal consistent clustering according to microsatellite types (Fig. 2).

For longer tandem repeats (mini- and satellites), inferring relationships with other groups of plants is more difficult because these sequences are normally specific to a related group of species (Rajagopal et al. 1999; Navajas-Pérez et al. 2006) and also undergo rapid evolutionary changes (Miklos 1985). The massive generation of genomic information for a wide range of taxa is promoting comparative analyses. By using the PlantSat DB (Macas et al. 2002), that compiles all existing satellite-DNA sequences in plants to date, and the *Arabidopsis* TAIR 7 annotation release (Poole 2007), we have been able to characterize the TR isolated here.

Generally speaking, there is a tendency in the assemblies analyzed to better detect satellite-DNA with shorter monomeric units ($\sim$150 bp). Additionally, in all cases, satellites belonging to the analyzed species or to a related taxon were detected and detection of centromeric, telomeric or rDNA-related sequences is most abundant (Table 2). In eukaryotes, centromeres and telomeres are often composed of cytologically distinctive heterochromatin and are associated with long arrays of satellite-DNA (Kipling 1995; Henikoff et al. 2001). This highly repetitive nature makes centromeres and telomeres difficult for sequencing and fine-scale genetic mapping. However, according to our survey and to other authors (Yan and Jiang 2007), new sequencing projects are succeeding in the characterization of such regions. In this respect, it is worth mentioning the sorghum assembly, for which we have detected here a 137-bp repeat that corresponds to the centromeric repeat CEN38 (Miller et al. 1998; Table 2).

One additional question we wanted to address was the ability of genomic assemblies to detect big blocks of heterochromatin. For that task, we used the CEN38 centromeric sequences that seem to be massively represented in the sorghum assembly (Fig. 3). Our 137 bp sequences together with the CEN38 repeats were comparatively mapped onto sorghum chromosomes. Both TR mapped to the centromeric regions of most chromosomes except for chromosome I, and interestingly the 137-bp sequences covered the majority of CEN38 repeats, suggesting that our method is able to detect most of the sequences of this satellite-DNA (Fig. 3). We also detected that CEN38 repeats would be constituted by 137 and 274-bp repeats, as previously suggested (Zwick et al. 2000). This case contrasts with the other genomes analyzed. The whole genome shotgun sequence and assembly strategy is now widely practiced. However, slightly different conditions are used to generate and process the data that could be reflected in different amount of TR detected. There are several varying parameters that could explain why the heterochromatic regions are differentially reflected in the genomic assemblies studied; (1) differences in heterochromatin content: for example, the papaya genome is mostly euchromatic (Ming et al. 2008), agreeing with the inability of our method to detect large blocks of TR in that species. In fact, TR in papaya are more-or-less randomly distributed, with their numbers positively correlated with supercontig length (this paper; Nagarajan et al. 2008). *A. thaliana*, whose chromosomes contain roughly 93% euchromatin (Koornneef et al. 2003), resembles papaya in the inability of our method to detect large blocks of TR. (2) Sequencing coverage: sequence finishing is normally expensive, tedious, and not affordable for many sequencing projects. Low coverage sequencing normally leads to a great amount of unassembled shotgun sequences, mainly comprising highly repetitive genomic DNA. Then, draft-quality assemblies may largely omit big blocks of heterochromatin. This could be another factor in the inability of our method to detect large blocks of TR in papaya (Ming et al. 2008) or *P. patens* (Rensing et al. 2008). (3) Sequencing approach: in BAC-based assemblies those BACs that are rich in repetitive DNA often do not contain a sufficient number of non-redundant bands to form contigs by fingerprinting. Therefore, TR may tend to get left out of BAC-based physical mapping efforts. Other approaches, as is the case of sorghum (Paterson et al. 2009), could offer a solution to deal better with TR as our results suggest.

As for the telomeric sequences, we have found the motif (TTTAGGG)n in all species analyzed except for

**Table 3** Mini- and satellite-DNA BLAST hits summary in *Arabidopsis thaliana* TAIR7 release (Poole 2007)

| | *Arabidopsis* | Papaya | Poplar | Grapevine | Rice | Sorghum | *Physcomitrella* | *Chlamydomonas* |
|---|---|---|---|---|---|---|---|---|
| Unclassified proteins | 4,256 | 63 | 71 | 56 | 50 | 98 | 60 | 84 |
| Transposable elements, viral and plasmid proteins | 1,047 | 426 | 14 | 9 | 6 | 22 | 3 | 1 |
| Metabolism | 283 | 14 | 5 | 16 | 12 | 4 | 0 | 1 |
| Cell rescue, defense and virulence | 180 | 1 | 6 | 3 | 4 | 6 | 3 | 0 |
| Classification not yet clear-cut | 179 | 1 | 3 | 6 | 2 | 3 | 0 | 6 |
| Protein synthesis | 88 | 0 | 2 | 8 | 5 | 2 | 0 | 3 |
| Cellular transport, transport facilitation and transport routes | 80 | 1 | 1 | 2 | 2 | 3 | 0 | 0 |
| Transcription | 75 | 7 | 5 | 6 | 2 | 5 | 6 | 4 |
| Cellular communication/signal transduction mechanism | 58 | 0 | 2 | 1 | 0 | 1 | 1 | 0 |
| Protein fate | 51 | 2 | 4 | 2 | 3 | 5 | 0 | 0 |
| Subcellular localization | 50 | 1 | 2 | 0 | 3 | 1 | 2 | 0 |
| Biogenesis of cellular components | 44 | 2 | 0 | 0 | 0 | 1 | 0 | 1 |
| Cell cycle and DNA processing | 13 | 2 | 0 | 1 | 1 | 2 | 0 | 0 |
| Energy | 12 | 1 | 1 | 1 | 0 | 0 | 0 | 3 |
| Development (systemic) | 10 | 0 | 0 | 1 | 0 | 6 | 0 | 6 |
| Protein with binding function or cofactor requirement (structural or catalytic) | 8 | 0 | 0 | 0 | 2 | 0 | 0 | 2 |
| Cell fate | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Systemic interaction with the environment | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Interaction with the environment | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Storage protein | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| Regulation of metabolism and protein function | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | 6,449 | 521 | 117 | 113 | 98 | 159 | 75 | 111 |

*C. reinhardtii*, for which the (TTTTAGGG)n motif was found. This suggests the predominance of the *A. thaliana*-type motif in all embryo plants. The (TTTTAGGG)n motif was found to be the green-alga-specific telomeric sequence (Petracek et al. 1990). In all cases, we found hexa- and heptameric variants of the telomeric motif. The presence of such derived telomeric motifs has also been reported in lilies (de la Herrán et al. 2005) and may be due to the high change rate of this type of sequences.

Most TR characterized here showed homology with unclassified proteins. This fact may be due to the massive number of hits to pseudogenes, and would suggest TR involvement in gene disruption and degeneration. However, additional categories included hits to genes involved in metabolism (especially binding elements and transcription factors), and defense and virulence related processes (Table 3). These types of genes have been found to be highly duplicated in poplar and *A. thaliana* (Tuskan et al. 2006), and then our data would suggest that a small percentage of detected TR correspond to uncharacterized tandemly arrayed genes (Table 3). The second most represented category corresponds to transposable elements. Forces governing satellite-DNA appearance and amplification are not well understood. The most accepted hypothesis suggests the continuous evolution of satellites from pre-existing satellites, through replication slippage and unequal crossing-over mechanisms (Ugarković and Plohl 2002). However, our data would support the possible origin of TR as transposable elements. The relation between TR and transposable elements has been demonstrated by different studies (Batistoni et al. 1995; Kapitonov et al. 1998; Zhao et al. 1998; López-Flores et al. 2004). Such TR DNAs may have originated from interspersed retrotransposons by unequal crossing-over (Kapitonov et al. 1998), although alternative mechanisms might be operating.

We found a predictable drop in the number of BLAST hits with phylogenetic distance along the evolutionary line, the higher number of hits being found in *A. thaliana*. It is also reflected in the variety of elements detected; for example in the transposable elements category, we found that *A. thaliana* repeats are related to several DNA transposons (*Athila*, *hAT*, *Mutator*), retrotransposons (*copia*, *gypsy*), and non-LTR transposable elements. This would suggest that our approach could be under-representing some species-specific sequences. However, this is a valid method for TR annotation and it will benefit from information and genomic resources

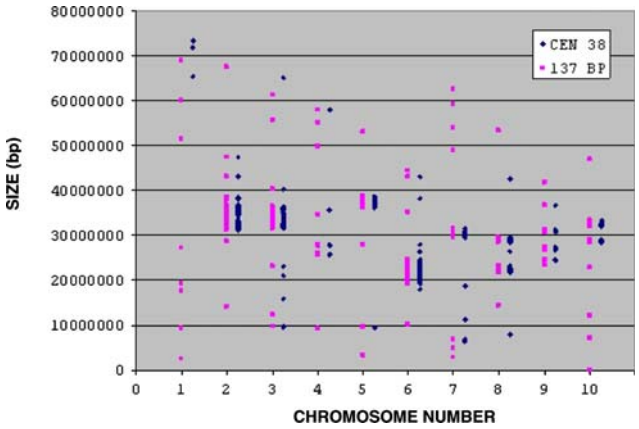


**Fig. 3** Comparative plotting of CEN38 (Miller et al. 1998) and the 137-bp repeats (this analysis) in sorghum genome. Chromosome number and size (in bp) are shown in *X* and *Y* axis, respectively

on their way from many other plant species. To contribute to that aim, we have created a database with all sequences characterized in the present paper (http://www.plantgenome.uga.edu/tandemrepeats/), expecting that it will be a useful resource for the scientific community.

Finally, we have detected in all species analyzed a bias in the distribution of repeat-unit sizes. It appears that sequences between 9 and 30 bp account for a high number of copies as well as for the maximum number of loci (Fig. 1). Interestingly, for the 9–30 bp range we also found the higher number of monomeric variants (Fig. 1). Stephan (1989) found similar results performing *in silico* simulations under certain experimental conditions (i.e. low rates of recombination, unequal crossing-over and replication slippage). This would be in accordance with the "library hypothesis", according to which related species share a pool of conserved TR; these simple-sequence DNAs would act as hot spots of recombination and some of them could be amplified into a major satellite (Mestrović et al. 1998). All these evidences together could indicate not only that tandem repeat units in this range are preferred in plant genomes, but also that the preservation of longer repetitive stretches would be somehow related to some features of the nucleotide sequence. Also, some authors have argued that structural features such as monomer length, AT content, short sequence motifs or secondary and tertiary structures may be important factors for tandem repeat preservation and evolution (Fitzgerald et al. 1994; Plohl et al. 1998; Ugarković and Plohl 2002). It has been proposed that these structural constraints could be important for tight packing of DNA and proteins in heterochromatin, and are consequently under selective pressure (Ugarković and Plohl 2002) and this could be an important area for future studies. However, since large numbers of repeats, especially those that are long and highly conserved, are particularly difficult to handle in sequence assembly, this affirmation should be taken with caution and data from other sources taken into consideration.

As sequencing technologies advance, large collections of genomic sequences, EST databases, and whole genome sequences are becoming increasingly available. Here we aimed to test if these large-scale DNA sequences are useful resources for investigating the frequency, distribution, and organization of TR. According to our data, current genomic assemblies include the vast majority of the TR fraction, including micro-, mini-, satellite-DNAs, and tandemly arrayed genes. However, factors such as the amount of heterochromatin in the source genomes, depth of sequence coverage, or parameters used in sequence processing, may influence the quantity of repeats assembled, and as a consequence most of the analyzed genomes may largely omit big clusters of heterochromatin. Collectively, these data demonstrate that although currently available sequencing technologies can be overwhelmed by megabase-sized satellite-DNA arrays, high-quality genomic sequences can be a good source for TR analysis. In this context, we have generated here a DB containing all the sequences analyzed with the aim to facilitate future analyses in this field.

**Acknowledgments** R. N.-P. is a Fulbright postdoctoral scholar (FU-2006-0675) supported by Spanish MEC.

## References

Akkaya MS, Shoemaker RC, Specht JE, Bhagwat AA, Cregan PB (1995) Integration of simple sequence repeat DNA markers into a soybean linkage map. Crop Sci 35:1439–1445

*Arabidopsis thaliana* Genome Initiative (2001) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Aragón-Alcaide L, Miller T, Schwarzacher T, Reader S, Moore G (1996) A cereal centromeric sequence. Chromosoma 105:261–268

Batistoni R, Pesole G, Marracci S, Nardi I (1995) A tandemly repeated DNA family originated from SINE-related elements in the European plethodontid salamanders (*Amphibia*, Urodela). J Mol Evol 40:608–615

Bennetzen JL (2002) Mechanisms and rates of genome expansion and contraction in flowering plants. Genetica 115:29–36

Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucl Acids Res 2:573–580

Bergman CM, Quesneville H (2007) Discovering and detecting transposable elements in genome sequences. Brief Bioinform 8:382–392

Blakesley RW, Hansen NF, Mullikin JC, Thomas PJ, McDowell JC, Maskeri B, Young AC, Benjamin B, Brooks SY, Coleman BI, Gupta J, Ho SL, Karlins EM, Maduro QL, Stantripop S, Tsurgeon C, Vogt JL, Walker MA, Masiello CA, Guan X, NISC Comparative Sequencing Program, Bouffard GG, Green ED (2004) An intermediate grade of finished genomic sequence suitable for comparative analyses. Genome Res 14:2235–2244

Camacho JP, Sharbel TF, Beukeboom LW (2000) B-chromosome evolution. Philos Trans R Soc Lond B Biol Sci 355:163–178

Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R (2000) Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 156:847–854

de la Herrán R, Cuñado N, Navajas-Pérez N, Santos JL, Ruiz Rejón C, Garrido-Ramos MA, Ruiz Rejón M (2005) The controversial telomeres of lily plants. Cytogenet Genome Res 109:144–147

Delseny M (2003) Towards an accurate sequence of the rice genome. Curr Opin Plant Biol 6:101–105

Depeiges A, Goubely C, Lenoir A, Cocherel S, Picard G, Raynal M, Grellet F, Delseny M (1995) Identification of the most represented repeated motifs in *Arabidopsis thaliana* microsatellite loci. Theor Appl Genetics 91:160–168

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 USA 95:8135–8140

Elder JR, Turner BJ (1995) Concerted evolution of repetitive DNA sequences in eukaryotes. Q Rev Biol 70:297–320

Fann J-Y, Kovarik A, Hemleben V, Tsirekidze NI, Beridze TG (2001) Molecular and structural evolution of *Citrus* satellite DNA. Theor Appl Genet 103:1068–1073

Fitzgerald DJ, Dryden GL, Bronson EC, Williams JS, Anderson JN (1994) Conserved patterns of bending in satellite and nucleosome positioning DNA. J Biol Chem 269:21303–21314

Flavell RB, Bennett MD, Smith JB, Smith DB (1974) Genome size and proportion of repeated nucleotide-sequence DNA in plants. Biochem Genet 12:257–269

Hagemann S, Scheer B, Schweizer D (1993) Repetitive sequences in the genome of *Anemone blanda*: identification of tandem arrays and of dispersed repeats. Chromosoma 102:312–324

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

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793–800

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyère C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pè ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quétier F, Wincker P, French–Italian Public Consortium for Grapevine Genome Characterization (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–467

Kamm A, Galasso I, Schmidt T, Heslop-Harrison JS (1995) Analysis of a repetitive DNA family from *Arabidopsis arenosa* and relationships between *Arabidopsis* species. Plant Mol Biol 27:853–862

Kapila R, Das S, Srivastava PS, Lakshmikumaran M (1996) A novel species-specific tandem repeat DNA family from *Sinapis arvensis*: detection of telomere-like sequences. Genome 39:758–766

Kapitonov VV, Holmquist GP, Jurka J (1998) L1 repeat is a basic unit of heterochromatin satellites in cetaceans. Mol Biol Evol 15:611–612

Kipling D (1995) The telomere. Oxford University Press, Oxford

Koornneef M, Fransz P, de Jong H (2003) Cytogenetic tools for *Arabidopsis thaliana*. Chromosom Res 11:183–194

Kubis SE, Schmidt T, Heslop-Harrison JS (1998) Repetitive DNA elements as a major component of plant genomes. Ann Botany 82:45–55

La Rota M, Kantety RV, Yu JK, Sorrells ME (2005) Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. BMC Genomics 6:23

Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658–1659

Lim KB, de Jong H, Yang TJ, Park JY, Kwon SJ, Kim JS, Lim MH, Kim JA, Jin M, Jin YM, Kim SH, Lim YP, Bang JW, Kim HI, Park BS (2005) Characterization of rDNAs and tandem repeats in the heterochromatin of *Brassica rapa*. Mol Cells 19:436–444

López-Flores I, de la Herrán R, Garrido-Ramos MA, Boudry P, Ruiz-Rejón C, Ruiz-Rejón M (2004) The molecular phylogeny of oysters based on a satellite DNA related to transposons. Gene 339:181–188

Loridon K, Cournoyer B, Goubely C, Depeiges A, Picard G (1998) Length polymorphism and allele structure of trinucleotide microsatellites in natural accessions of *Arabidopsis thaliana*. Theor Appl Genet 97:591–604

Macas J, Mészáros T, Nouzová M (2002) PlantSat: a specialized database for plant satellite repeats. Bioinformatics 18:28–35

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376–380

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, Terry A, Salamov A, Fritz-Laylin LK, Maréchal-Drouard L, Marshall WF, Qu LH, Nelson DR, Sanderfoot AA, Spalding MH, Kapitonov VV, Ren Q, Ferris P, Lindquist E, Shapiro H, Lucas SM, Grimwood J, Schmutz J, Cardol P, Cerutti H, Chanfreau G, Chen CL, Cognat V, Croft MT, Dent R, Dutcher S, Fernández E, Fukuzawa H, González-Ballester D, González-Halphen D, Hallmann A, Hanikenne M, Hippler M, Inwood W, Jabbari K, Kalanon M, Kuras R, Lefebvre PA, Lemaire SD, Lobanov AV, Lohr M, Manuell A, Meier I, Mets L, Mittag M, Mittelmeier T, Moroney JV, Moseley J, Napoli C, Nedelcu AM, Niyogi K, Novoselov SV, Paulsen IT, Pazour G, Purton S, Ral JP, Riaño-Pachón DM, Riekhof W, Rymarquis L, Schroda M, Stern D, Umen J, Willows R, Wilson N, Zimmer SL, Allmer J, Balk J, Bisova K, Chen CJ, Elias M, Gendler K, Hauser C, Lamb MR, Ledford H, Long JC, Minagawa J, Page MD, Pan J, Pootakham W, Roje S, Rose A, Stahlberg E, Terauchi AM, Yang P, Ball S, Bowler C, Dieckmann CL, Gladyshev VN, Green P, Jorgensen R, Mayfield S, Mueller-Roeber B, Rajamani S, Sayre RT, Brokstein P, Dubchak I, Goodstein D, Hornick L, Huang YW, Jhaveri J, Luo Y, Martínez D, Ngau WC, Otillar B, Poliakov A, Porter A, Szajkowski L, Werner G, Zhou K, Grigoriev IV, Rokhsar DS, Grossman AR (2007) The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science 318:245–250

Messing J, Bharti AK, Karlowski WM, Gundlach H, Kim HR, Yu Y, Wei F, Fuks G, Soderlund CA, Mayer KF, Wing RA (2004) Sequence composition and genome organization of maize. Proc Natl Acad Sci USA 101:14349–14354

Mestrović N, Plohl M, Mravinac B, Ugarković D (1998) Evolution of satellite DNAs from the genus Palorus—experimental evidence for the "library" hypothesis. Mol Biol Evol 15:1062–1068

Métais I, Aubry C, Hamon B, Peltier D, Jalouzot R (1998) Cloning, quantification and characterization of a minisatellite DNA sequence from common bean *Phaseolus vulgaris* L. TAG 97:232–237

Miklos GL (1985) Localited highly repetitive DNA sequences in vertebrate and invertebrate genomes. In: McIntryre JR (ed) Molecular evolutionary genetics. Plenum Press, New York, pp 231–241

Miller JT, Jackson SA, Nasuda S, Gill BS, Wing RA, Jiang J (1998) Cloning and characterization of a centromere specific DNA element from *Sorghum bicolor*. Theor Appl Genet 96:832–839

Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang ML, Zhu YJ, Schatz M, Nagarajan N, Acob RA, Guan P, Blas A, Wai CM, Ackerman CM, Ren Y, Liu C, Wang J, Wang J, Na JK, Shakirov EV, Haas B, Thimmapuram J, Nelson D, Wang X, Bowers JE, Gschwend AR, Delcher AL, Singh R, Suzuki JY, Tripathi S, Neupane K, Wei H, Irikura B, Paidi M, Jiang N, Zhang W, Presting G, Windsor A, Navajas-Pérez R, Torres MJ, Feltus FA, Porter B, Li Y, Burroughs AM, Luo MC, Liu L, Christopher DA, Mount SM, Moore PH, Sugimura T, Jiang J, Schuler MA, Friedman V, Mitchell-Olds T, Shippen DE, de Pamphilis CW, Palmer JD, Freeling M, Paterson AH, Gonsalves D, Wang L, Alam M (2008) The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus). Nature 452:991–996

Morgante M, Hanafey M, Powell W (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 30:194–200

Mun JH, Kim DJ, Choi HK, Gish J, Debellé F, Mudge J, Denny R, Endré G, Saurat O, Dudez AM, Kiss GB, Roe B, Young ND, Cook DR (2006) Distribution of microsatellites in the genome of *Medicago truncatula*: a resource of genetic markers that integrate genetic and physical maps. Genetics 172:2541–2555

Murray MG, Peters DL, Thompson WF (1981) Ancient repeated sequences in the pea and mung bean genomes and implications for genome evolution. J Mol Evol 17:31–42

Nagaki K, Tsujimoto H, Sasakuma T (1998) A novel repetitive sequence of sugar cane, SCEN family, locating on centromeric regions. Chromosom Res 6:295–302

Nagarajan N, Navajas-Pérez R, Pop M, Alam M, Ming R, Paterson AH, Salzberg SL (2008) Genome-wide analysis of repetitive elements in papaya. Tropical Plant Biol 3–4:191–201

Navajas-Pérez R, Schwarzacher T, de la Herrán R, Ruiz Rejón C, Ruiz Rejón M, Garrido-Ramos MA (2006) The origin and evolution of the variability in a Y-specific satellite-DNA of *Rumex acetosa* and its relatives. Gene 368:61–71

Nunome T, Suwabe K, Ohyama A, Fukuoka H (2003) Characterization of trinucleotide microsatellites in eggplant. Breed Sci 53:77–83

Okagaki RJ, Phillips RL (2004) Maize DNA-sequencing strategies and genome organization. Genome Biol 5:223

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Lyons E, Maher C, Narechania A, Penning B, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein PE, Kresovich S, McCann MC, Ming R, Peterson DG, Ware D, Westhoff P, Mayer KFX, Messing J, Rokhsar DS (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature 457:551–556

Peterson DG, Wessler SR, Paterson AH (2002) Efficient capture of unique sequences from eukaryotic genomes. Trends Genet 18:547–550

Petracek ME, Lefebvre PA, Silflow CD, Berman J (1990) *Chlamydomonas* telomere sequences are A+T-rich but contain three consecutive G-C base pairs. Proc Natl Acad Sci USA 87:8222–8226

Plohl M, Mestrović N, Bruvo B, Ugarković D (1998) Similarity of structural features and evolution of satellite DNAs from *Palorus subdepressus* (Coleoptera) and related species. J Mol Evol 46:234–239

Poole RL (2007) The TAIR database. Methods Mol Biol 406:179–212

Rajagopal J, Das S, Khurana DK, Srivastava PS, Lakshmikumaran M (1999) Molecular characterization and distribution of a 145-bp tandem repeat family in the genus *Populus*. Genome 42:909–918

Rensing SA, Lang D, Zimmer AD, Terry A, Salamov A, Shapiro H, Nishiyama T, Perroud PF, Lindquist EA, Kamisugi Y, Tanahashi T, Sakakibara K, Fujita T, Oishi K, Shin-I T, Kuroki Y, Toyoda A, Suzuki Y, Hashimoto S, Yamaguchi K, Sugano S, Kohara Y, Fujiyama A, Anterola A, Aoki S, Ashton N, Barbazuk WB, Barker E, Bennetzen JL, Blankenship R, Cho SH, Dutcher SK, Estelle M, Fawcett JA, Gundlach H, Hanada K, Heyl A, Hicks KA, Hughes J, Lohr M, Mayer K, Melkozernov A, Murata T, Nelson DR, Pils B, Prigge M, Reiss B, Renner T, Rombauts S, Rushton PJ, Sanderfoot A, Schween G, Shiu SH, Stueber K, Theodoulou FL, Tu H, Van de Peer Y, Verrier PJ, Waters E, Wood A, Yang L, Cove D, Cuming AC, Hasebe M, Lucas S, Mishler BD, Reski R, Grigoriev IV, Quatrano RS, Boore JL (2008) The *Physcomitrella* genome reveals evolutionary insights into the conquest of land by plants. Science 319:64–69

Rizzon C, Ponger L, Gaut BS (2006) Striking similarities in the genomic distribution of tandemly arrayed genes in *Arabidopsis* and rice. PLoS Comput Biol 2(9):e115

Saini N, Shultz J, Lightfoot DA (2008) Re-annotation of the physical map of Glycine max for polyploid-like regions by BAC end sequence driven whole genome shotgun read assembly. BMC Genomics 9:323

Simoens CR, Gielen J, Van Montagu M, Inzé D (1988) Characterization of highly repetitive sequences of *Arabidopsis thaliana*. Nucleic Acids Res 16:6753–6766

Smulders MJM, Bredemeijer G, Rus-Kortekaas W, Arens P, Vosman B (1997) Use of short microsatellites from database sequences to generate polymorphisms among *Lycopersicon esculentum* cultivars and accessions of other *Lycopersicon* species. Theor Appl Genet 97:264–272

Song QJ, Fickus EW, Cregan PB (2002) Characterization of trinucleotide SSR motifs in wheat. Theor Appl Genet 104:286–293

Stephan W (1989) Tandem-repetitive noncoding DNA: forms and forces. Mol Biol Evol 6:198–212

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Déjardin A, Depamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M, Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjärvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leplé JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouzé P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai CJ, Uberbacher E, Unneberg P, Vahala J, Wall

K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604

Tutois S, Cloix C, Cuvillier C, Espagnol MC, Lafleuriel J, Picard G, Tourmente S (1999) Structural analysis and physical mapping of a pericentromeric region of chromosome 5 of *Arabidopsis thaliana*. Chromosome Res 7:143–156

Ugarković D, Plohl M (2002) Variation in satellite DNA profiles, causes and effects. EMBO J 21:5955–5959

von Stackelberg M, Rensing SA, Reski R (2006) Identification of genic moss SSR markers and a comparative analysis of twenty-four algal and plant gene indices reveal species-specific rather than group-specific characteristics of microsatellites. BMC Plant Biol 6:9

Wang J, Chen C, Na J-K, Yu Q, Hou S, Paull RE, Moore PH, Alam M, Ming R (2008) Genome-wide comparative analyses of microsatellites in papaya. Tropical Plant Biol 3–4:278–292

Wu TY, Wu R (1987) A new rice repetitive DNA shows sequence homology to both 5S RNA and tRNA. Nucleic Acids Res 15:5913–5923

Wu HK, Chung MC, Wu TY, Ning CN, Wu R (1991) Localization of specific repetitive DNA sequences in individual rice chromosomes. Chromosoma 100:330–338

Yan H, Jiang J (2007) Rice as a model for centromere and heterochromatin research. Chromosom Res 15:77–84

Zhao X, Ding X, Ji Y, Stelly D, Paterson AH (1998) Macromolecular organization and genetic mapping of a recently-amplified tandem repeat family (B77) in cotton. Plant Mol Biol 38:1031–1042

Zwick MS, Islam-Faridi MN, Zhang HB, Hodnett GL, Gomez MI, Kim JS, Price HJ, Stelly DM (2000) Distribution and sequence analysis of the centromere-associated repetitive element CEN38 of *Sorghum bicolor* (Poaceae). Am J Bot 12:1757–1764