Genome-wide analysis of tandem repeats in plants and green algae

Zhixin Zhao¹, Cheng Guo¹, Sreeshandarajan Sutharzan¹, Pei Li¹,², Craig S Echt³, Jie Zhang⁴ and Chun Liang¹*

¹ Department of Botany, Miami University, Oxford, Ohio 45056, USA
² Department of Automation, Xiamen University, Xiamen, 361005, China
³ Southern Research Station, USDA Forest Service, Saucier, MS 39574, USA
⁴ State Key Laboratory for Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agricultural Science, Beijing 100193, China

*Corresponding author

Email addresses:
ZZ: zhaoz@miamioh.edu
CG: guoc2@miamioh.edu
SS: sreesks@miamioh.edu
PL: peil@miamioh.edu
CSE: cecht@fs.fed.us
JZ: zhangjie05@caas.cn
CL: liangc@miamioh.edu
ABSTRACT

Tandem repeats (TRs) extensively exist in the genomes of prokaryotes and eukaryotes. Based on the sequenced genomes and gene annotations of 31 plant and algal species in Phytozome v8.0 (http://www.phytozome.net/), we examined TRs in a genome-wide scale, characterized their distributions and motif features, and explored their putative biological functions. Among the 31 species, no significant correlation was detected between the TR density and genome size. Interestingly, green alga *Chlamydomonas reinhardtii* (42,059 bp/Mbp) and castorbean *Ricinus communis* (55,454 bp/Mbp) showed much higher TR densities than all other species (13,209 bp/Mbp averagely). In the 29 land plants, including 22 dicots, 5 monocots and 2 bryophytes, 5'-UTR and upstream intergenic 200 nt (UI200) regions had the first and second highest TR densities, whereas in the two green algae (*C. reinhardtii* and *Volvox carteri*), the first and second highest densities were found in intron and CDS regions respectively. In CDS regions, tri- and hexa-nucleotide motifs were those most frequently represented in all species. In intron regions, especially in the two green algae, significantly more TRs were detected near the intron-exon junctions. Within intergenic regions in dicots and monocots, more TRs were found near both the 5' and 3' ends of genes. GO annotation in two green algae revealed that the genes with TRs in introns are significantly involved in transcriptional and translational processing. As the first systematic examination of TRs in plant and green algal genomes, our study showed that TRs displayed non-random distribution for both intragenic and intergenic regions, suggesting that they have potential roles in transcriptional or translational regulation in plants and green algae.
Tandem repeats (TRs) are DNA sequence motifs that contain at least two adjacent repeating units. They extensively exist in prokaryotes and eukaryotes (Tautz and Renz 1984; Tóth et al. 2000; Sharma et al. 2007; Sureshkumar et al. 2009; Christians and Watt 2009; Orsi et al. 2010; Roorkiwal and Sharma 2011). Generally, two categories are given to distinguish TRs based on different repeat unit size: microsatellites (unit size: 1–6 or 1–10 bp, also known as simple sequence repeats (SSR)) and minisatellites (unit size: 10–60 or 10–100 bp) (Mayer et al. 2010; Gemayel et al. 2012). In plants and animals, SSRs are widely detected in both mRNAs (cDNA/ESTs) and genomes (Tautz and Renz 1984; Jurka and Pethiyagoda 1995; Tóth et al. 2000; Subramanian et al. 2003; Fujimori et al. 2003; Sharma et al. 2007; Gemayel et al. 2010). For example, through investigating SSRs (repeat unit size: 1–6 bp) using EST databases in 11 plant and green algal species, Victoria et al. find that dimer motifs have higher frequencies in green algae, bryophytes and ferns; while trimer motifs are more frequent in flowering plants (Victoria et al. 2011). Different from nuclear genomes, mitochondrial genomes appear to prefer mononucleotide repeats (A/T) first and dinucleotide repeats (AT) next in 16 investigated plant species (Kuntal and Sharma 2011). Although most published research papers focus on SSRs, Mayer et al. find that in coding regions, densities of longer TRs (unit size: 7–50 bp) in arthropoda Daphnia pulex are much higher than shorter TRs (unit size: 1–6 bp), and suggest the importance of including longer TRs in comparative analyses (Mayer et al. 2010).

TRs are extremely mutable with mutation rates that are much higher than other parts of the genome (Gemayel et al. 2010). Most mutations in TRs are caused by the changes in the number of the repeating units, not by point mutations (Verstrepen et al. 2005; Gemayel et al. 2010, 2012). In humans, such repeat number variants are related to some serious diseases or defects, such as Fragile X Syndrome (FRAXA) (Verkerk et al. 1991), Spinobulbar Muscular Atrophy (SBMA) (La Spada et al. 1991) and Huntington’s disease (HD) (Walker 2007). In plants, the well-known Bur-0 IIL1 defect in Arabidopsis thaliana that generates a detrimental phenotype is caused by the expansion of triplet TTC/GAA in the intron of IIL1 gene (Sureshkumar et al. 2009).

Through investigating TR density variation in a few plant and animal species, it is concluded that there is no significant relationship between genome size and TR density in plants and animals (Da Maia et al. 2009; Mayer et al. 2010). Based on EST data from two green algae, two mosses, a fern, a fern palm, the ginkgo tree, two conifers, ten dicots and five monocots, SSRs are found to have highly variable abundances among different species (Von Stackelberg et al. 2006). Recently, a comparative analysis for 282 species including plants and animals shows no sequence conservation in centromere TRs (Melters et al. 2013). Moreover, TRs show a non-random distribution in many genomes and are often located within genes and regulatory regions (Streelman and Kocher 2002; Rockman and Wray 2002; Li et al. 2002; Martin et al. 2005; Legendre et al. 2007; Vinces et al. 2009). Variable TRs are abundant in genes that are involved in transcriptional regulation and morphogenesis in humans (Legendre et al. 2007). 5’-UTRs have higher TR density among different genic regions in plants (Morgante et al. 2002; Fujimori et al. 2003; Zhang et al. 2006). In A. thaliana, for example, 5’-UTRs have the highest TR
density and the abundant motifs are dinucleotide CT/GA and trinucleotide CTT/GAA (Zhang et al. 2006). In the yeast *Saccharomyces cerevisiae*, about 25% genes possess TRs in their promoters, and the variations of repeat unit number can cause changes in gene expression and local nucleosome positioning (Vinces et al. 2009). Among coding sequences (CDS), the dominant repeat unit sizes are three-fold nucleotides (*e.g.* tri- and hexa-nucleotides), because it is assumed that such motifs are selected to avoid frame shift mutations that would affect translation (Legendre et al. 2007; Metzgar et al. 2002). In 42 fully sequenced prokaryotic genomes, the TR distributions in CDS are biased toward CDS termini, yielding U-shaped TR density curves across the span of the CDS (Lin and Kussell 2011).

So far, no systematic research on TR variation and characterization has been conducted on a genome-wide scale in plants. The rapid advance of sequencing technologies has made a number of plant and algal genomes available to investigate the characteristics and distributions of TRs in both intragenic (*i.e.*, 5'-UTR, CDS, intron and 3'-UTR) and intergenic regions. Using genome sequence data from 31 species (*i.e.*, 29 land plants and 2 green algae) released in Phytozome v8.0 (http://www.phytozome.net/), we detected and characterized TRs and examined their distributions and variations in intragenic and intergenic regions. This research will facilitate our understanding of TRs and their potential biological functions in transcription or translation in land plants and green algae.

**MATERIALS AND METHODS**

**Collecting genomes and annotation data**

The assembled genome sequences (including chromosomes, mitochondria and chloroplasts) and gene annotations of the 31 species were downloaded from Phytozome v8.0 (http://www.phytozome.com/) (see Figure 1 for the species list). Only valid nucleotides (A, T, G and C) were counted when analyzing the sequences. For each species, the nucleotide sequences from whole genome were used for genome-wide TR detection and density calculation. According to the data extraction schema shown in Figure 2, individual intergenic and intragenic regions were also extracted and used for TR analysis. In Phytozome v8.0, UTR annotations, including 5'-UTRs and 3'-UTRs, were not available for *Carica papaya, Brassica rapa, Linum usitatissimum* and *Malus domestica*. Therefore, the UTR regions were not examined individually for these four species. However, the upstream and downstream intergenic regions (*e.g.*, UI1000, DI1000) were still examined based on the relevant gene start and end positions annotated for these four species (see Figure 2). Perl (Practical Extraction and Report Language) was used to write codes to extract sequences, initiate TR detection and parse results for downstream data analysis.

**TR detection and analysis**

For both perfect and imperfect TR detection, we utilized a tandem repeat search tool for complete genomes - Phobos (Mayer et al. 2010) (version 3.3.12). Considering
the computational resource and execution time required for processing all 31 genomes, we adopted 1-50 bp as the repeat unit size, similar to what has been utilized by Mayer et al. The minimum length of the detected repeats needed to be at least 12 nt, and the minimum repeat alignment score for imperfect repeats was set to be 12. As for the recursive TRs, only one motif was selected based on alphabetical ordering to be representative (Jurka and Pethiyagoda 1995). For example, AAG, AGA and GAA were the repeat units of (AAG)_n, but only AAG was selected to represent the repeat motif. Moreover, the TR motifs and their corresponding reverse complement motifs (e.g., AAG and CTT motifs) were investigated separately. This was because (1) genes are annotated in different strand (i.e., + versus -), (2) there are plenty of sense and anti-sense transcripts reported recently for many genes (Gu et al. 2009; Kerin et al. 2012), emphasizing the importance of gene orientation in genome annotations, and (3) a similar strategy had been adopted by others (Zhang et al. 2006; Kuntal and Sharma 2011).

TR density was defined by base pairs per mega base pairs (bp/Mbp), namely the length of detected TRs out of the total length of the sequences for detection. To enable comparison among different species or different regions (e.g., intragenic vs. intergenic regions) within the same species, we normalized the TR densities and computed the relative density: for each species, the whole genome density was defined as 100, and then the relative density for a specific region was computed by (TR density for a given region)/(the whole genome density). To investigate the TR distribution profiles within a given region, the sequence length of a specific region was firstly normalized to 0–99 scale that contained 10 intervals of same size (e.g. 0–9, 10–19, …, 90–99). The motif percentages were then calculated for the 10 intervals based on their occurrences. In this way, the same intergenic or intragenic regions with different sequence lengths can be compared.

The 198 experimentally verified plant promoter sequences, which were extracted from -499 to 100 around the transcription start site (0 position in the coordinate), were downloaded from EPD (Eukaryotic Promoter Database, http://epd.vital-it.ch/seq_download.php#) (Périer et al. 2000). These promoter sequences were also scanned for perfect and imperfect TRs.

Based on *Chlamydomonas reinhardtii* GO annotation (v4.0) from JGI (http://genome.jgi-psf.org/Chlre4/Chlre4.download.ftp.html), GOEAST (Gene Ontology Enrichment Analysis Software Toolkit) was used to detect the significance of GO terms for the genes with TRs in introns. GO annotations in *Volvox carteri* were analyzed by annot8r (Schmid and Blaxter 2008) and ranked based on E-value. Pearson correlation (r) test statistics were conducted using Minitab 16 (www.minitab.com). The figures of box-plot were drawn using R (http://www.r-project.org/). Also through R, both ANOVA F-test and Tukey's Honestly Significant Difference (HSD) test (Yandell 1997) were performed for significance tests.
RESULTS

The TR density variation among different genome sizes

The species that we examined span a large evolutionary distance, including two green algae, two mosses, five monocots, and twenty-two dicots (Figure 1). As shown in Figure 3, there was no correlation between genome sizes and TR densities ($r = 0.010$, $p = 0.957$). The mean TR density at the whole-genome level was 13,209 bp/Mbp ($sd = 10,309$) among all tested species except in *C. reinhardtii* (42,059 bp/Mbp) and *Ricinus communis* (55,454 bp/Mbp), which showed dramatically higher TR densities than the other species. Excluding these two outliers (*C. reinhardtii* and *R. communis*), we still cannot find a significant correlation between genome sizes and TR densities among the remaining species ($r = 0.311$, $p = 0.101$) (see Figure 3).

The TR density variation in intragenic and intergenic regions

Sequences from functionally different intragenic regions (*i.e.*, 5'-UTR, CDS, intron and 3'-UTR) and progressively flanking upstream (*i.e.*, UI200, UI 500, UI1000) and downstream (*i.e.*, DI200, DI500 and DI1000) intergenic regions were analyzed for TRs (Figure 2). UTR annotations were not available from Phytozome v8.0 for four species (*C. papaya*, *B. rapa*, *L. usitatissimum* and *M. domestica*), therefore 5'-UTR and 3'-UTR were analyzed only for the remaining 27 species.

We found that TRs showed clearly localization preferences among different intragenic and intergenic regions. In the two green algae *C. reinhardtii* and *V. carteri* (Figure 4 A and Table S1 and S2), intron regions have the highest relative TR densities of 162 and 120 respectively, which are 1.62 and 1.20 times of the relevant whole-genome TR densities (the whole-genome relative TR density is defined as 100 for each species). Based on F-test from ANOVA, the null hypothesis that all tested intergenic and intragenic regions have the equal mean relative TR densities can be rejected ($p < 0.001$), and Tukey’s Honestly Significant Difference (HSD) test (Yandell 1997) also showed a significant difference between the intron and each of the other regions ($p < 0.05$). In contrast, CDS regions had the second highest relative TR densities in the genic regions (86 and 65), whereas 5'-UTRs have the lowest (17 and 22). Interestingly, TRs in intergenic regions increased their relative densities away from the genes in these two green algae (Figure 4A and Table S1 and S2).

In the 29 land plants we examined, 5'-UTRs had the most significant and highest relative densities ($p < 0.001$, F-test from ANOVA, $p < 0.01$, HSD test; Figure 4 B and C, Table S1 and S2) among different intragenic and intergenic regions. In the dicots and bryophytes, CDS regions had the significantly lowest relative TR densities ($p < 0.01$, F-test from ANOVA, $p < 0.03$, HSD test; Figure 4 B, Table S1) among different regions. In monocots, the relative densities of CDS, intron and 3'-UTRs are similarly low (Figure 4 C, Table S1 and S2). Different from the two green algae, the intergenic regions in land plants generally show higher TR densities than their genomes average (Figure 4 B, C and D, Table S1 and S2). Comparing all intergenic regions (Figure 4 B and C), promoter regions closing to 5'-UTR appear to have more TR occurrences in land plants. In particular, the UI200 (upstream intergenic 200nt) regions display a strong positive
correlation with 5'-UTR in term of relative TR densities for land plants ($r = 0.755$, $p = 1.998e-05$).

The nucleotide content of the most abundant TR motifs are influenced by GC content

As shown in Table S3, all 22 dicots have GC contents ranged from 32.40% to 39.56%, five monocots from 43.57% to 46.14%, and two green algae from 55.70% to 63.45%. For the two bryophyte species, the GC content of moss *Physcomitrella patens* (33.60%) is within the range of dicots whereas spikemoss *Selaginella moellendorffii* (45.25%) is within monocots. GC contents in genome wide scale seem to follow the pattern of green algae > monocots > dicots. On the other hand, GC contents vary greatly among different intragenic and intergenic regions. In intragenic regions of dicots (see Table 1 and Table S3), the highest and lowest GC contents are detected in CDS (44.34%) and intron (33.19%) respectively, and 5'-UTR has the second highest (39.89%). In intergenic regions of dicots, GC contents vary from 31.75% to 35.56%. In intragenic regions of monocots and bryophytes, the highest and lowest GC contents are detected in 5'-UTR (54.79%) and intron (39.15%) respectively, CDS has the second highest content (53.32%). In intergenic regions of monocots and bryophytes, GC contents change from 42.39% to 49.20%. In green algae, the highest GC content is detected in CDS (66.17%), lowest in 5'-UTR (52.82%) and its adjacent intergenic region (UI200, 52.32%). The 3'-UTR and other intergenic regions in green algae have GC contents from 53.75% to 57.38%. Differently from both monocots and dicots, introns in green algae show the second highest GC content (58.38%).

Our data suggests a clear relationship between GC contents and nucleotide content of the most frequent TR motifs detected within either a whole genome or individual intragenic or intergenic regions. If a high GC content is detected within a given region, the abundant TR motifs will preferably be GC-rich. In dicots, the most abundant TRs have repeat unit sizes of mono-, di- and tri-nucleotides, except CDS where tri-nucleotide TR motifs are the most frequent and then tri-fold TR motifs (e.g., hexa- and 9- nucleotide motifs) the second (see Table 1). As shown in Table S4, the top TR motifs in dicots are dinucleotide motifs (16.87%, e.g., AT), mononucleotide motifs (14.48%, e.g., A/T) and trinucleotide motifs (9.17%, e.g., ATT/AAT and AAG/CTT). Also, 4-7 bp motifs still show high frequencies (>3%) whereas other longer motifs have low frequencies (<2%) except 39-nucleotide motifs (3.92%). This exception is caused by the dramatically high frequency of 39-nucleotide motifs detected in *R. communis* (69.57%, see Table S4). Moreover, only AT-motifs (e.g., T, AT and ATT) are detected in introns in dicots where the lowest GC content is evident in comparison with other intragenic regions (see Table 1). Differently from dicots, mononucleotide motifs are lower in frequency (8.73%) whereas tri-nucleotide motifs (14.12%) and di-nucleotide motifs (13.05%) are obviously preferred in monocots (>13%, Table S4). Meanwhile, GC-rich motifs like CGG/GCC are more frequently found in monocots than in dicots due to their higher GC contents. Although tri- and tri-fold nucleotide motifs are still dominant in CDS regions in monocots, those are essentially GC-rich motifs (e.g., CGG/GCC). Interestingly, dinucleotide (16.69%) and 12-nucleotide (17.48%) motifs have higher frequencies in two bryophytes (Table S4), because dramatically high
dinucleotide (26.79%) and 12-nucleotide (31.11%) motifs are detected in *P. patens* and *S. moellendorffii* respectively. In green algae, mononucleotide TR motifs show an extremely low frequency (0.94% only, see Table S4). In green alga *C. reinhardtii*, di-nucleotide GT/AC motifs are dominantly used in all intra- and inter-genic regions except 5'-UTR and CDS regions where tri-nucleotide AGC and CGG/CGG motifs are frequently used. In green alga *V. carteri*, the long 17-nucleotide motifs are frequently found in all intragenic regions except CDS regions where tri-nucleotide CGG and AGC are frequent. Interestingly, the top three frequent motifs in *V. carteri* are 17-nucleotide motifs (8.18%), tri-nucleotide (7.44%) and 50-nucleotide (7.39%) (see Table S4).

Based on the analysis of 198 experimentally verified plant promoter sequences downloaded from EPD (Eukaryotic Promoter Database), the abundant TR motif units are mono-, di- and tetra-nucleotide, and the top-ranked frequent motifs are A- and AT-rich (e.g., A/T, AT, ATGC, CTTT and ATTT), which is very similar to our results in the intergenic region adjacent to 5'-UTR (i.e., UI200 regions) in the dicots and monocots.

**TR distribution and frequency profiles in intragenic (5'-UTR, CDS, intron and 3'-UTR) and intergenic regions**

As shown in Figure 5 A and Table S5, within the upstream intergenic UI200 regions of both dicots and monocots, the highest and lowest relative TR motif contents are found in the 80−89 (p < 0.001, F-test from ANOVA; p < 0.01, HSD test) and 0−9 intervals (p < 0.001, F-test from ANOVA; p < 0.01, HSD test) respectively. This suggests that the distribution of TR motifs is significantly toward the 3' ends of UI200, closer to 5' ends of genes. In contrast, within the downstream intergenic DI200 regions (see Figure 5 B and Table S5), TR motifs are shown to distribute significantly toward the 5' ends of DI200, closer to 3' ends of genes, considering that the highest and lowest relative motif contents are detected in 10−19 (p < 0.001, F-test from ANOVA; p < 0.05, HSD test) and 90−100 (p < 0.001, F-test from ANOVA; p < 0.01, HSD test) intervals respectively. Interestingly, the progressively increasing trend of motif frequency toward gene ends does not keep in the sub-regions that are immediately adjacent to gene ends (e.g., the interval 90−99 in Figure 5 A and the interval 0-9 in Figure 5 B). Meanwhile, TR motif frequencies appear to be relatively consistent within 5'-UTR, 3'-UTR and CDS regions (Figure S1), except near their ends (i.e., the interval 0-9 and the interval 90−99) where lower frequencies are detected. Within the introns of all 31 species, more motifs are significantly detected in 10-19 (p < 0.001, F-test from ANOVA; p < 0.05, HSD test) and 80−89 intervals (p < 0.001, F-test from ANOVA; p < 0.01 except comparing with 20−29 where p is 0.13, HSD test) intervals. So, it is deduced that TR motifs are more frequently distributed towards intron ends forming U-shape (Figure 5 C), which has a different trend from intergenic UI200 and DI200 regions.

Different from dicots and monocots, bryophytes and green algae show special trends in TR distribution profiles in both intergenic UI200 and DI200 regions: more motifs are detected in the middle intervals (see Figure 5 D, E and Table S5), and no progressive increase or decrease trend is observed. On the other hand, the TR distribution profiles within intragenic regions are similar to dicots and monocots (see Figure S1 and Table S5).
As shown in Figure 6 and Table S6, we have determined TR occurrences among all annotated genes, their intragenic regions and adjacent promoter regions for 4 groups of all 31 species (bryophytes, monocots, dicots and green algae). First, ~84% of all annotated mRNAs (or genes, some genes have more than one mRNA annotated) possess TRs (see Figure 6 A and Table S6), and no significant difference is detected among 4 groups by F-test and HSD test. Interestingly, monocots show less TR frequency than other three groups, but the difference is not statistically significant. In UI200 and 5'-UTR regions (see Figure 6 B and C), about 4% and 10% of the annotated mRNAs have TRs respectively, except for green algae, in which only ~2% are found with TRs in both regions. The difference in TR frequencies in UI200 and 5'-UTR regions of all annotated mRNAs is not significant among 4 groups. However, as shown in Figure 6 D, E and F, green algae display significantly higher TR frequencies in 3'-UTR (~17%), CDS (~9%) and intron (~23%) regions for all annotated mRNAs in comparison with other three groups: bryophytes, monocots and dicots ($p < 0.01$, F-test from ANOVA; $p < 0.01$, HSD test).

Utilizing GOEAST (Gene Ontology Enrichment Analysis Software Toolkit) (Zheng and Wang 2008), the GO terms of C. reinhardtii genes with TRs in introns were analyzed. As shown in Table 2, the most highly enriched GO terms involve catalytic activity ($p = 5.384e-35$) and hydrolase activity ($p = 1.344e-10$). In green alga V. carteri, the most significant GO functions mainly involve ribosomal proteins and heat shock proteins in the genes with TRs in introns (see Table S7). Such results suggest that the TRs could involve RNA and/or protein activity in intron processing.

DISCUSSION

The variation of TR densities in different genomes

In genome-wide study of TRs using 12 species including two fungi (S. cerevisiae and Neurospora crassa), one green alga (Ostreococcus lucimarinus), one plant (A. thalina), three vertebrates (Homo sapiens, Mus musculus, Gallus gallus), one nematode (Caenorhabditis elegans) and three arthropods (Daphnia pulex, Drosophila melanogaster, Apis mellifera), Mayer et al (Mayer et al. 2010) detected weak but no significant correlation between the genome sizes and TR densities ($r = 0.483$, $p = 0.111$). In three plant families Brassicaceae, Solanaceae and Poaceae (da Maia et al. 2009), the association between genome sizes and TR densities detected in mRNA/cDNA data was also not found. In a recent study of 257 virus genomes, the relative SSR densities (i.e., SSRs sequence base pairs per kilo genomic base pairs) showed quite weak correlation with genome size (Zhao et al. 2012). Our analysis shows no significant relationship detected between TR density and genome size in green algae and plants (see Figure 3). Furthermore, it is obviously showed that TR densities have species-specific features rather than group-based features, like the two green algae; such result is coincided with the SSR density variation detected in 25 algae and plants (von Stackelberg et al. 2006). It seems that there is a week positive, but not significant, correlation detected between genome sizes and TR densities for both compact genomes (like viruses) and genomes...
with lot of intergenic regions (like plants). This suggests that TRs might have not contributed significantly to the genome size expansion in evolution.

The variation of TR densities in different intragenic and intergenic regions

In Arabidopsis and rice, TRs are significantly enriched within 5'-UTRs (Fujimori et al. 2003; Zhang et al. 2004; Lawson and Zhang 2006). In our study, both dicot and monocot plants possess the first and second highest TR densities in 5'-UTRs and their immediate upstream intergenic regions (i.e., UI200), which belong to the promoter regions where core promoter elements are often represented (Kokulapalan 2011) (see Figure 4 B and C). 5'-UTRs are thought to be the hotspots for TRs in eukaryotes. Previous studies on genes for light and salicylic acid responses (Li et al. 2004; Zhang et al. 2006) suggested that TRs in 5'-UTRs might be involved in the transcription and/or translation regulation. It is reported that as many as 25% genes in yeast \textit{S. cerevisiae} have TRs in the promoter regions (Vinces et al. 2009). Our study also demonstrates that ~4–25% of genes in dicots and monocots possess TR in both 5'-UTR and promoter UI200 regions (see Figure 6 B and C). In both dicots and monocots, TR abundance is the least in CDS region, indicating that low TR abundance may decrease the evolvability of proteins. This is reasonable because it is demonstrated that the mutations of CDS could cause protein functional changes, loss of function and protein truncation (Li et al. 2004). Interestingly, intron and 3'-UTR regions have much lower TR densities in monocots than in dicots. Such TR differences between dicots and monocots are still not clear in their biological meanings.

In the two green algae we examined, the first and second highest TR densities are detected in intron and CDS regions respectively (Figure 4 A), which is completely different from all other land plants. Our data show that green algae have significantly more intron sequences (32.85% and 37.03% in the whole genome in \textit{C. reinhardti} and \textit{V. carteri}) comparing with land plants (15.73% averagely). This may imply that in green algae the TRs in intron and CDS regions are not randomly expanded and could involve in intron- or CDS-related activities and in RNA processing (e.g. exon splicing). In fact, our GO analysis for \textit{C. reinhardti} genes with TRs in intons shows that the most significant GO functions are catalytic activity and hydrolase activity (Table 2). Those functions indicate that the genes with rich TR motifs in their introns could involve in protein synthesis and degradation.

The top TR motifs are influenced by GC content

In our study, the top ranked TR motifs are CT/AG and CTT/AAG in 5'-UTR in dicots. This is consistent with the results from Zhang \textit{et al}, in which the motifs (CT/AG and CTT/AAG) were preferred in 5'-UTR in Arabidopsis and acted as regulatory elements for genes involved in light and salicylic acid responses (Zhang et al. 2006). Our result also shows that CDS regions are preferentially associated with tri- and hexa-nucleotides motifs, which has been reported previously by other researchers (Subramanian et al. 2003; Fujimori et al. 2003; Li et al. 2004; Zhang et al. 2006; Mayer et al. 2010). It is suggested that there is a strong evolutionary pressure against TR expansion in CDS than in introns to keep stable protein products (Dokholyan et al. 2000). Such feature can help explain why tri-fold nucleotide motifs (e.g. tri- and hexa-
nucleotides motifs) are more frequent than others to reduce potential translational frame shifting. Two green algae have the highest TR densities in introns and the relevant abundant motifs are dinucleotide GT/AC in our study. As known, canonical splicing signals GT and AG are located at the 5’ and 3’ ends of the intron respectively. The abundant GT/AC dinucleotide TRs in introns might suggest that such repeats may be involved in exon splicing or alternative splicing in green algae (Gemayel et al. 2012).

In dicots, most TR motifs contain A and/or T nucleotide(s); while both A/T-rich motifs and CCG/CGG motifs are often used in monocots. However, A/T-rich motifs are rarely detected in the two green algae. So it is clear that the top TR motifs have strong relationship with the GC content (Table 1). If there is a high GC content, the top frequent TR motifs prefer to be GC-rich instead of AT-rich. A similar relationship also has been demonstrated in other 11 species (including green algae, bryophytes, ferns, gymnosperms and angiosperms) (Victoria et al. 2011) and AAR (Amino Acid Repeats) in 10 angiosperms (Zhou et al. 2011).

In term of repeat unit size length distribution (Table S4), mono-nucleotide motifs are not the most frequent TR motifs in all 31 investigated species. It is known that longer repeats (>6bp) have high densities in D. pulex (Mayer et al. 2010). In our study, some longer repeats also show higher frequencies than many short TRs: 39-nucleotide motifs in dicot R. communis, 17-nucleotide and 50-nucleotide motifs in green alga V. carteri, and 12-nucleotide motifs in bryophyte S. moellendorffii. So, this suggests that TRs are not generated randomly in genome and longer TRs may play some roles in gene expression and regulation.

The distribution and frequency of TRs in intragenic and intergenic regions

It is clear that the distribution of TR motifs in intergenic regions is significantly biased towards both the 5’ and 3’ ends of genes in dicots and monocots (Figure 5 A and B). TRs have been shown to locate within genes and regulatory regions and participate in transcriptional and translational regulation (Streelman and Kocher 2002; Rockman and Wray 2002; Li et al. 2002; Martin et al. 2005; Legendre et al. 2007; Vinces et al. 2009). In our study, the biased TR motif distribution in intergenic regions further supports this notion.

In introns of all the 31 species, especially in the two green algae, more abundant TR motifs are significantly detected toward the ends of introns. Interestingly, SSR densities in CDS regions of 42 prokaryote genomes also show a similar U-shaped profile (Lin and Kussell 2011). Because introns contain important regulatory motifs for many biological processing including splicing (Matlin et al. 2005; Barbazuk et al. 2008), our results suggest that the TRs in introns might have localization preference in their regulatory roles. Considering exon splicing that utilizes the canonical splicing signals (GT and AG) at the 5’ and 3’ end of introns and the GO functions of genes with TRs in introns (see Table 2), we believe that the highly abundant TRs in introns, especially in the two green algae, may involve with both constitutive and alternative splicing activities.

The frequencies of TRs are consistent with the TR density variations in the four different groups. It is showed that 5’-UTR and UI200 have much higher TR densities in
dicots, monocots and bryophytes, whereas higher TR densities are found in intron and CDS regions in green algae (Figure 4 and Figure 6). Comparatively speaking, there are more TRs (densities and frequencies) in 5’-UTR and promoter (UI200) regions in land plants (dicots, monocots and bryophytes), while green algae have more TRs in intron and CDS regions.

In this study, the genome assemblies and gene annotations are obtained from Phytozome v8.0. Within this release, some species (e.g., Arabidopsis and rice) apparently have better, high-quality gene annotation than other species (e.g., papaya and apple without UTR annotation). We also noticed that many genome assemblies have unfinished gaps (e.g., ...NNN...). Perhaps, this is due to the highly repetitive nature of the sequences and the limitation of current sequencing technologies. On other hand, our data analysis is obviously biased towards dicot plants because the species number available in Phytozome v8.0 is not balanced for all four groups: 2 species in green algae, 2 in bryophytes, 5 in monocots, and 22 in dicots. Another limitation in our data analysis is the repeat unit size selection. Ideally, we should have examined all TRs with the repeat unit size of 1-100 (i.e., covering all microsatellites and minisatellites) or even longer. Unfortunately, we have decided to examine the TR motifs of 1-50 bp due to the constraints in both current bioinformatics tools and the demanding computational resources required for processing all 31 genomes (i.e., CPU, memory and execution periods). Clearly, these limitations will affect the quality of our data analysis results presented in this paper to some extent. With the rapid advances in sequencing and computational technologies and with the rapid increase of transcriptomics data, we can expect more high-quality, accurate genome assemblies and gene annotations available for in-depth TR analyses from more plant and green algal species. This will definitely help us improve our understanding of the evolution of TRs and their roles in gene expression regulation.

CONCLUSIONS

It is known that TRs involve plenty of roles in gene expression and genome evolution. In this study, as the first systematic examination of TRs in plant and green alga genomes, we found that TRs density has no significantly discernible relationship with genome size, and TRs display non-random distribution within both intragenic and intergenic regions, suggesting that they might have been involved in transcriptional or translational regulation in plants and green algae. Obviously, more research work is needed to facilitate our understanding of TRs in terms of their motif features and potential biological functions, as well as their evolutionary trends in land plants and green algae.

ACKNOWLEDGMENTS

This project was funded partially by the NIH-AREA (1R15GM94732-1 A1 to CL), and Botany Department and Office for the Advancement of Research and Scholarship (OARS) in Miami University. CL managed and coordinated the project. ZZ carried out data collection and data analysis. CG and SS helped with statistical analyses. GO analysis in V. carteri was implemented by PL. All authors participated in manuscript
writing and editing. We thank Qingshun Quinn Li and two anonymous reviewers for their constructive comments to improve the manuscripts.
Table 1. The top frequent TR motifs and GC contents in different genomic regions

<table>
<thead>
<tr>
<th>Group name (genome GC content range)</th>
<th>Region</th>
<th>Average GC content (%)</th>
<th>Top motifs</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Dicots</strong></td>
<td>Whole genome</td>
<td>35.61</td>
<td>A/T; AT; ATT/AAT, AAG/CTT</td>
</tr>
<tr>
<td>(33%, 40%)</td>
<td>UI1000</td>
<td>32.63</td>
<td>A/T; AT; ATT/AAT, AAG/CTT</td>
</tr>
<tr>
<td></td>
<td>UI500</td>
<td>31.75</td>
<td>A/T; ATT/AAT</td>
</tr>
<tr>
<td></td>
<td>UI200</td>
<td>35.56</td>
<td>A/T; AT, CT; CTT/AAG</td>
</tr>
<tr>
<td></td>
<td>5'UTR</td>
<td>39.89</td>
<td>CT/AG; CTT/AAG</td>
</tr>
<tr>
<td></td>
<td>CDS</td>
<td>44.34</td>
<td>AAG/CTT</td>
</tr>
<tr>
<td></td>
<td>Intron</td>
<td>33.19</td>
<td>T; AT; ATT</td>
</tr>
<tr>
<td></td>
<td>3'UTR</td>
<td>35.41</td>
<td>T; AT; ATT/AAT</td>
</tr>
<tr>
<td></td>
<td>DI200</td>
<td>34.58</td>
<td>T/A; AT, AG; AAT/ATT, AAG/CTT</td>
</tr>
<tr>
<td></td>
<td>DI500</td>
<td>32.78</td>
<td>T/A; AT; AAT/ATT, AAG/CTT</td>
</tr>
<tr>
<td></td>
<td>DI1000</td>
<td>33.50</td>
<td>T/A; AT; AAT/ATT, AAG/CTT</td>
</tr>
<tr>
<td><strong>Monocots and bryophytes</strong></td>
<td>Whole genome</td>
<td>43.29</td>
<td>AT; AAT/ATT, CCG/CGG</td>
</tr>
<tr>
<td>(43%, 55%)</td>
<td>UI1000</td>
<td>42.39</td>
<td>AT; AAT/ATT, CCG/CGG</td>
</tr>
<tr>
<td></td>
<td>UI500</td>
<td>43.38</td>
<td>AT; ATT, CCG</td>
</tr>
<tr>
<td></td>
<td>UI200</td>
<td>49.20</td>
<td>AT; CCG</td>
</tr>
<tr>
<td></td>
<td>5'UTR</td>
<td>54.79</td>
<td>AG/CT; CCG</td>
</tr>
<tr>
<td></td>
<td>CDS</td>
<td>53.32</td>
<td>CCG/CGG</td>
</tr>
<tr>
<td></td>
<td>Intron</td>
<td>39.15</td>
<td>C; CT</td>
</tr>
<tr>
<td></td>
<td>3'UTR</td>
<td>42.34</td>
<td>CTT/AAG, CCG, GT</td>
</tr>
<tr>
<td></td>
<td>DI200</td>
<td>45.27</td>
<td>AT; CCG/CGG</td>
</tr>
<tr>
<td></td>
<td>DI500</td>
<td>42.67</td>
<td>AT; CCG/CGG</td>
</tr>
<tr>
<td></td>
<td>DI1000</td>
<td>42.56</td>
<td>AT; CCG/CGG</td>
</tr>
<tr>
<td><strong>Green algae</strong></td>
<td>Whole genome</td>
<td>59.58</td>
<td>AC/GT; CCG/CGG; AAGCATATGCGATCTGC</td>
</tr>
<tr>
<td>(&gt;55%)</td>
<td>UI1000</td>
<td>57.38</td>
<td>AC/GT; CCG/CGG; AAGCATATGCGATCTGC</td>
</tr>
<tr>
<td></td>
<td>UI500</td>
<td>55.84</td>
<td>GT/AC</td>
</tr>
<tr>
<td></td>
<td>UI200</td>
<td>52.32</td>
<td>GT/AC</td>
</tr>
<tr>
<td></td>
<td>5'UTR</td>
<td>52.82</td>
<td>AGC</td>
</tr>
<tr>
<td></td>
<td>CDS</td>
<td>66.17</td>
<td>CCG/CGG, AGC</td>
</tr>
<tr>
<td></td>
<td>Intron</td>
<td>58.38</td>
<td>GT/AC</td>
</tr>
<tr>
<td></td>
<td>3'UTR</td>
<td>55.42</td>
<td>GCT/AGC; GT</td>
</tr>
<tr>
<td></td>
<td>DI200</td>
<td>53.75</td>
<td>GT/AC</td>
</tr>
<tr>
<td></td>
<td>DI500</td>
<td>56.32</td>
<td>GT/AC</td>
</tr>
<tr>
<td></td>
<td>DI1000</td>
<td>57.16</td>
<td>GT/AC; AAGCATATGCGATCTGC</td>
</tr>
</tbody>
</table>
Table 2. The most significant GO functions of genes with TRs in introns in *C. reinhardtii*

<table>
<thead>
<tr>
<th>Term</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>catalytic activity</td>
<td>5.384e-35</td>
</tr>
<tr>
<td>hydrolase activity</td>
<td>1.344e-10</td>
</tr>
<tr>
<td>oxidoreductase activity</td>
<td>7.792e-6</td>
</tr>
<tr>
<td>peptidase activity</td>
<td>1.109e-4</td>
</tr>
<tr>
<td>peptidase activity, acting on L-amino acid peptides</td>
<td>1.766e-4</td>
</tr>
<tr>
<td>endopeptidase activity</td>
<td>5.793e-4</td>
</tr>
</tbody>
</table>
LITERATURE CITED


Unstable Tandem Repeats in Promoters Confer Transcriptional Evolvability.


Yandell, B. S., 1997 *Practical data analysis for designed experiments*. Chapman & Hall,
London; New York.

Zhang, L., D. Yuan, S. Yu, Z. Li, Y. Cao *et al.*, 2004 Preference of simple sequence
repeats in coding and non-coding regions of Arabidopsis thaliana. Bioinformatics

Zhang, L., K. Zuo, F. Zhang, Y. Cao, J. Wang *et al.*, 2006 Conservation of noncoding

Zhao, X., Y. Tian, R. Yang, H. Feng, Q. Ouyang *et al.*, 2012 Coevolution between simple
sequence repeats (SSRs) and virus genome size. BMC genomics 13: 435.

Zheng, Q., and X.-J. Wang, 2008 GOEAST: a web-based software toolkit for Gene

Zhou, Y., J. Liu, L. Han, Z.-G. Li, and Z. Zhang, 2011 Comprehensive analysis of
tandem amino acid repeats from ten angiosperm genomes. BMC genomics 12:
632.
Figures

Figure 1. The phylogenetic tree of 31 species showed in Phytozome v8.0 (http://www.phytozome.com/). The abbreviated names (the first two letters from both genus and species name are combined) and common names are listed in the parenthesis.

Figure 2. The schematic intragenic and intergenic regions used for TR analysis. UI200: 1-200 nt upstream of 5’UTR; UI500: 201-700 nt upstream of 5’UTR; UI1000: 701-1700 nt upstream of 5’UTR; DI200: 1-200 nt downstream of 3’UTR; DI500: 201-700 nt downstream of 3’UTR; DI1000: 701-1700 nt downstream of 3’UTR.

Figure 3. Genome size versus genomic TR density in 31 land plants and green algae. The abbreviation name is based on the first two letters from both genus and species name (See Figure 1). The linear trendline is made by Excel.

Figure 4. The relative TR densities in different intragenic and intergenic regions. (A) 2 green algae; (B) 20 species including dicots and bryophytes; (C) 5 monocot land plants; (D) 4 land plant species without UTR annotations.

Figure 5. The relative distribution position of TRs in the intron and intergenic regions. (A) upstream 200 nt region in dicots and monocots; (B) downstream 200 nt region in dicots and monocots; (C) intron region in the 31 investigated species; (D) upstream 200 nt region in bryophytes and green algae; (E) downstream 200 nt region in bryophytes and green algae.

Figure 6. The percentage of TRs in different intragenic and intergenic regions. (A) mRNAs; (B) UI200 region; (C) 5’-UTR region; (D) 3’-UTR region; (E) CDS region; (F) Intron region.
Additional files

**Supplemental File 1: Figure S1.** The relative distribution position of TRs in the 3 intragenic regions. (A) 5’-UTR regions in the 27 investigated species; (B) CDS regions in the 31 investigated species; (C) 3’-UTR regions in the 27 investigated species.

**Supplemental File 2: Table S1.** The relative TR densities in the 31 investigated species.

**Supplemental File 3: Table S2.** The means and SD values of relative TR densities shown in Figures 4-6.

**Supplemental File 4: Table S3.** GC contents and top frequent TR motifs in the 31 investigated species.

**Supplemental File 5: Table S4.** 1-50 bp TR motif length distribution (%) in all 31 investigated species.

**Supplemental File 6: Table S5.** TR distribution in intragenic and intergenic regions in the 31 investigated species.

**Supplemental File 7: Table S6.** The TR numbers and percentages in intragenic and intergenic regions in the 31 investigated species.

**Supplemental File 8: Table S7.** The list of GO functions from the genes with TRs in intron regions in green alga *V.carteri*. 


21 / 21
Land plants

Dicots

Monocots

Bryophytes

Green algae
Upstream intergenic region

Gene

Downstream intergenic region

UI1000  UI1500  UI200  Exon  Intron  Exon  Intron  Exon  DI200  DI500  DI1000

5'UTR     CDS     CDS     CDS     3'UTR

Figure 2
Figure 4

(A) Relative density of UI1000, UI500, UI200, 5'UTR, CDS, Intron, 3'UTR, DI200, DI500, and DI1000.

(B) ...
Figure 5

Motif relative content (%)