Novel repetitive sequences decipher the evolution and phylogeny in Carthamus L

CURRENT STATUS: POSTED

Shweta Mehrotra  SHWETAMEHROTRAGOYAL@GMAIL.COM
University of Delhi
Corresponding Author

Vinod Goyal
University of Delhi

DOI:
10.21203/rs.2.9839/v1

SUBJECT AREAS
Plant Molecular Biology and Genetics
Abstract

Repetitive sequences are ubiquitous features of eukaryotic genomes, which contribute up to 70-80% of the nuclear genomic DNA. They are known to impact genome evolution and organization and play important role in genome remodelling. The widespread distribution and sufficient conservation of repeats reinforce the value of repetitive DNA sequences as markers of evolutionary processes. The repetitive DNA-based phylogeny reconstruction method is consistent in resolving expected phylogenetic and evolutionary relationships. In the present study, we address the isolation and characterization of four novel repetitive sequences (pCtHaeIII-I, pCtHaeIII-II, pCtHaeIII-III and pCtTaqI-I) from Carthamus tinctorius. Detailed phylogenetic analysis of 18 taxa belonging to 7 species of Carthamus has also been done with pCtHaeIII-I, and pCtHaeIII-II which clearly indicated concerted evolution while delineating phylogenetic relationships among the 18 taxa studied. The above understanding can assist in the marker assisted genetic improvement/ enhancement programmes in this crop species.

Background

The genus Carthamus (Asteraceae) includes about 25 species and subspecies distributed from Spain and North Africa along Middle East to Northern India. Carthamus tinctorius is grown commercially for the purpose of edible seed oil, orange-red dye, carthamine, and also has applications in herbal medicine.

The genus Carthamus consists of taxa with five gametic numbers (n=10, 11, 12, 22, and 32) including diploids as well as polyploid species. Most of the wild species of the genus including the cultivated safflower are diploids with 2n=2x=20, 22 and 24 whereas the polyploid taxa exhibit 2n=4x=44 and 2n=6x=64 chromosomes. The taxonomic classification of the genus has seen many revisions (refer to Table 1, Mehrotra et al.,
Comprehensive studies using molecular tools like RAPD, ISSR, AFLP, ITS-RFLP, nuclear SACPD, ITS, ETS sequence and microsatellite data, and cytogenetic approaches have been used to resolve the phylogenetic and taxonomic relationships between different taxa of the genus (Sehgal et al., 2009; Bowles et al., 2010).

Repetitive sequences are ubiquitous features of eukaryotic genomes, which contribute up to 70-80% of the nuclear genomic DNA. They may be arranged as tandem arrays of monomers in continuous clusters in a ‘head to tail’ organization or may be dispersed widely throughout the genome. They evolve in a concerted fashion such that the intraspecific sequence similarity of repeating units exceeds the interspecific sequence similarity (Elder and Turner 1995). They are known to impact genome evolution and organization and play important role in genome remodelling (Fedoroff, 2012; Beritoli et al., 2013). They are also reported to be involved in the evolution of plant sex chromosomes by triggering heterochromatization and causing recombination suppression, leading to structural and morphological differentiation of sex chromosomes and X chromosome dosage compensation (Li et al., 2016). Repetitive sequences can accumulate variations in sequence and copy number during evolution, and are therefore considered important tools for taxonomic and phylogenetic/phylogenomics studies (Mehrotra and Goyal, 2014). Intraspecific changes in genome size may play a role in environmental adaptation and can affect developmental dynamics and phenotypic characteristics at the cellular and organismal level.

The widespread distribution and sufficient conservation of repeats reinforce the value of repetitive DNA sequences as markers of evolutionary processes. The repetitive DNA-based phylogeny reconstruction method is consistent in resolving expected phylogenetic and evolutionary relationships. Repetitive sequences have been used for the analysis of phylogenetic relationships in various plant species like *Cucumis, Hordeum, Citrus, Beta,*...
Crocus, Vicia and Centaurea (Zentgraf et al., 1992; Svitashev et al., 1994; De Felice et al., 2004; Dechyeva and Schmidt, 2006; Frello et al., 2000; Frediani et al., 2004; Suarez-Santiago et al., 2007). However, repetitive sequences as markers relating to phylogeny in Carthamus have not been well explored. In our previous paper, we reported a detailed phylogenetic analysis in Carthamus employing two repetitive sequences, pCtkpnI-I and pCtkpnI-II (Mehrotra et al., 2013; Raina et al., 2005).

In the present study, we address the isolation and characterization of four novel repetitive sequences (pCtHaeIII-I, pCtHaeIII-II, pCtHaeIII-III and pCtTaqI-I) from Carthamus tinctorius. Detailed phylogenetic analysis of 18 taxa belonging to 7 species of Carthamus has also been done with pCtHaeIII-I, and pCtHaeIII-II. The reported data allowed further clarification of the origin of cultivated safflower and the phylogenetic relationships among the Carthamus taxa.

Methods

Plant Material

The germplasm of different taxa of Carthamus was obtained from United States Department of Agriculture (USDA), Beltsville, USA; Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany; all India Coordinated Research Projects on Oilseeds, Solapur and Directorate of Oilseeds Research (DOR), Hyderabad, India. The details of germplasm used in the present study is mentioned in our previous paper (Refer to Table 2, Mehrotra et al., 2013).

The leaf samples of different taxa of angiosperms (monocots and dicots) including those of family Asteraceae were collected from various regions of Delhi and are listed in Table 7.

Restriction Digestion of Total Genomic DNA
Total genomic DNA of 18 taxa of *Carthamus* was extracted from young, tender leaves by modified CTAB method as detailed by Porebski et al. (1997). The quality and quantity of DNA was determined through agarose gel electrophoresis (AGE) using 0.8 % agarose gel. Total genomic DNA (100 mg/sample) of *Carthamus tinctorius* was separately digested with *Hae* III and *Taq* I restriction endonucleases, at 37° C (*Hae* III) and 65° C (*Taq* I). The digested DNA was fractionated overnight on 0.85% AGE prepared in 1XTAE (0.04M Tris acetate, 1mM Na₂EDTA) buffer at 15V. Flourescein labelled *Hind*III digested Lambda DNA served as a molecular size marker. After completion of AGE, the gel was photographed under UV light (Fig. 1).

**Colony Hybridization, Southern Hybridization and Dot-Blot**

The *Hae* III and *Taq* I digested DNA fragments of size <1000bp were separately eluted from the gel and purified using agarose gel extraction kit (Qiagen) according to manufacturers’ instructions. The purified DNA fragments were then cloned and transformed using Ecoli DH5α strain and plated on 2% Luria broth (1.5% agar) medium with 100 mg/ml ampicillin. The DNA library thus obtained was used for colony hybridization purpose. For preparing colony lifts, petridishes with bacterial colonies were precooled for at least 30 min at 4° C before taking a lift. Correct size of nylon membrane disc matching the size of petriplate having colonies was allowed to sit on the surface. The disc position was marked at several positions with a pin to ensure correct orientation of colonies for subsequent manipulations. Membrane was removed from the petridish after 1 min, in one continuous movement using blunt ended forceps. Membrane disc was placed on Whatman paper No. 1 for drying, with the side having colonies facing up. DNA was liberated from the colonies, denatured and fixed to the membrane by placing the
membrane discs with colonies uppermost on a series of solution saturated 3 MM paper pads. Membrane discs were initially placed in denaturation buffer for 2-5 min and then in neutralization buffer for 3 min and again in neutralization buffer for another 3 min. Finally, membrane was vigourously washed in 2X SSC to remove proteinaceous debris. The membrane disc, with DNA side up, was finally transferred to a pad of 3 MM papers and air dried. The DNA was fixed to the membrane by baking it for 2 h at 80°C, and the membrane was stored in saran wrap until further use. Total genomic DNA of *Carthamus tinctorius* was used as a probe for colony hybridization. Labelling of the probe and the subsequent hybridization reactions were performed according to Mehrotra et al. (2013). The colonies with bright signals were expected to harbor repetitive sequences and were sequenced. The identified repetitive sequences were designated as pCtHaeIII-I, pCtHaeIII-II, pCtHaeIII-III and pCtTaqI-I.

Total genomic DNAs (1µg/sample) of 18 taxa of *Carthamus* were separately digested with *Hae* III and *Taq* I restriction endonucleases, separated in 0.8 % agarose gels, and then transferred onto a nylon membrane (Hybond N+, Amersham, UK) by alkaline transfer method (Reed and Mann, 1985). Total genomic DNAs (50ng/sample) were also dot-blotted onto a nylon membrane (Hybond N+, Amersham, UK) and allowed to dry at room temperature. The membrane was then baked at 80 °C for 1 h. Finally, the membrane was rinsed in 2x SSC and was wrapped in saran wrap and stored wet at 4 °C until further use. Hybridizations were performed using the four identified repetitive sequences separately as probes according to Mehrotra et al., 2013.

**PCR Amplification, Cloning and Sequencing**

Amplification of the four repetitive sequences, pCtHaeIII-I, pCtHaeIII-II, pCtHaeIII-III and
pCtTaqI-I in each of the 18 taxa of *Carthamus* was done by PCR using primers designed from the sequences of *C. tinctorius* by the authors using Primer 3 software. Primer sequences have been detailed in Table 6. Amplification and cloning of repetitive sequences were performed according to Mehrotra *et al.* (2013). The amplified products were cloned into pGEMT-Easy vector (Promega Co., USA) in *E.coli* strain DH5α and positive clones (4 clones of each sequence in each taxa) were sequenced at the DNA Sequencing Facility, University of Delhi (South Campus), India.

**DNA Sequence Analysis**

Cloned sequences were analyzed for homology to known nucleotide sequences from the database (GeneBank, EMBL) using BLAST from NCBI and PlantSat database (http://w3lamc.umbr.cas.cz/PlantSat). Sequences have been submitted to GenBank under the accession numbers KX986356-KX986359.

Dot-matrix analyses of self comparisons of repetitive sequences was done using MegAlign application from Lazergene ‘99 software package at different matrix stringencies. The repetitive sequences were analyzed using a predictive model of sequence-dependent DNA bending. The bendability/curvature propensity plot was calculated according to Goodsell and Dickerson (1994), Burkner *et al.* (1995) and the consensus bendability scale. The values of the curvature are presented as the deflection angle per 10.5 residue helical turn (1°/bp). The maximum curvature peak is localized within the monomer satellite DNA consensus sequence.

DNA Motif search was performed from Prosite documentation using MOTIF tool of Genome Net database. Retroposon finder was used to search for any retroposons in the sequences.

**Phylogenetic Analysis**
For sequence data, alignment was done with Clustal X program (Saitou and Nei, 1987; Thompson et al., 1997) using default settings with a fixed gap penalty of 6.66, and DNA transition weight of 0.5 in the multiple alignment parameter option. The presence of phylogenetic signal was assessed by likelihood mapping analysis (LMA) using TreePuzzle-5.0 software based on quartet analysis (Strimmer and Haesler, 1997). Neighbor joining and maximum parsimony methods were used to create phylogenetic trees from the aligned sequence data matrix using PAUP*4.0 b 8 (Swofford, 2002). Gaps were treated as missing data. Given a large size of the data set, heuristic searches used the Tree Bisection Reconstruction (TBR) option with MULPARS and ACCTRAN optimization. The amount of support for the branches was assessed using 100 bootstrap replicates with 10 random additions per replicate using TBR and MULPARS. A 50 % majority rule consensus tree was calculated from the most parsimonious trees using the CONTREE command in PAUP.

Results

Restriction digestion of total genomic DNA of Carthamus tinctorius with enzymes, HaeIII and TaqI, separately, showed a smear with some prominent bands between 1000-500bp (Fig. 1). Colony hybridization of HaeIII and TaqI libraries of C. tinctorius with total genomic DNA of C. tinctorius revealed colonies with faint and bright signals. Among these colonies, four colonies with bright signals were considered as repeat sequences and were selected for further analysis. These colonies were designated as pCtHaeIII-I, pCtHaeIII-II, pCtHaeIII-III and pCtTaqI-I (Fig. 2). Analysis of each repetitive sequence was done separately.

Restriction Analysis of Genomic DNA

Hybridization of HaeIII digested total genomic DNA of 18 taxa of Carthamus with pCtHaeIII-
revealed regular periodicity of the hybridization bands in a typical ladder pattern with the smallest visible band of 340 bp (Fig. 3a). Hybridization profile of pCtHaelll-II for 18 taxa of *Carthamus* showed homogenous presence of bands at 284 bp, and 568 bp positions (Fig. 4a). Hybridization with pCtHaelll-III in 18 taxa of *Carthamus* showed homogenous presence of a prominent band at 158 bp and a faint band at 316 bp position (Fig. 5a). Hybridization signals of longer DNA fragments were gradually less pronounced. Southern hybridization with pCtTaqI-I in 18 taxa of *Carthamus* showed homogenous presence of a prominent single band at 362 bp (Fig.6a). The hybridization pattern did not change with increasing amount of enzyme or with longer incubation times in each case suggesting that digestions were complete and multiple bands were due to alterations in restriction site. The dot-blot of total genomic DNA of 18 taxa of *Carthamus*, probed with pCtHaelll-I, pCtHaelll-II, pCtHaelll-III and pCtTaqI-I separately, showed strong signals (Fig.3b, 4b, 5b, 6b) suggesting repetitive nature of these sequences in all the taxa.

**PCR Screening of Repetitive Sequences in 18 Taxa of Carthamus**

Sequence analysis of the four repetitive sequences enabled the design of primers (Table 6) which could amplify 340bp pCtHaelll-I, 284bp pCtHaelll-II, 158bp pCtHaelll-III and 362bp pCtTaqI-I sequences separately, representing the monomer repeat units from 18 taxa of *Carthamus* (Fig. 3c, 4c, 5c, 6c).

**Sequence Analysis and Characterization**

The repetitive sequences pCtHaelll-I, pCtHaelll-II, pCtHaelll-III and pCtTaqI-I were found to be 340bp, 284bp, 158bp and 362bp in length respectively (Table 1). The sequences did not show any significant similarity to the previously reported sequences when subjected to
homology searches in GenBank, EMBL, DDBJ and PDB databases using BLAST.

pCtHaeIII-I, pCtHaeIII-II, pCtHaeIII-III and pCtTaqI-I had around 28%, 45%, 51% and 39% GC content respectively. Base changes were analyzed within the clones of each taxa separately in case of each sequence. Almost all the changes were single base pair substitutions in which transitions and transversions occurred evenly. The base changes did not seem to be clustered in any restricted regions. Microsatellites were also evident in the four sequences (Table 5). The sequences showed the presence of GG, GA, and AG nearest neighbours. Sequences analysis also revealed a frequent occurrence of GGT and GTT trinucleotides and presence of poly-A tracts and a pentanucleotide CAAAA (or its inverse complementary TTTTG). A perfect polyadenylation signal, AATAAA was present in pCtHaeIII-I and pCtHaeIII-III.

The curvature propensity plot of pCtHaeIII-I, pCtHaeIII-II, pCtHaeIII-III and pCtTaqI-I revealed a fair value of curvature propensity of around 9.5, 11, 7 and 12 respectively (Fig.3e, 4e, 5e, 6e). The dot matrix for each of the four repetitive sequences was analyzed at different matrix stringencies (Fig.3d, 4d, 5d, 6d). 100% matrix stringency revealed that the pCtHaeIII-I sequence contained 8 bp and 10 bp direct repeats; 8 bp and 9 bp inverted repeats; 8 bp mirror repeat; and 8 bp complementary repeat. The other three repetitive sequences contained 6-10 bp direct repeats (Table 5).

DNA motif search revealed the presence of Thiolases active site domain (PS00099) in pCtHaeIII-I; 2Fe-2S ferredoxin-type iron-sulfur binding region (PS00197) in pCtHaeIII-II and Anaphylatoxin domain (PS01177) in pCtTaqI-I repetitive sequence.

Phylogenetic Analysis

Detailed phylogenetic analysis of the genus Carthamus was carried out with pCtHaeIII-I
and pCtHaeIII-II repetitive sequences using *Carthamus arborescens* (2n=24) as outgroup. The monomer units of pCtHaeIII-I and pCtHaeIII-II of 18 taxa of *Carthamus* were separately cloned and sequenced. Four randomly selected clones for pCtHaeIII-I and pCtHaeIII-II were sequenced for each of the 18 diploid (2n = 20, 24) and polyploid (2n = 44, 64) taxa.

Interclonal sequence variation of each sequence ranged from 2-5% within each taxa. All phylogenetic reconstructions showed that repeat types in each taxon were more closely related to one another than to repeat types of the other taxa. Therefore, a consensus sequence was obtained for each of them, in all the 18 taxa separately, which was then used for further phylogenetic analysis. The two sequences were analyzed separately for phylogeny.

**pCtHaeIII-I**

The length of amplified pCtHaeIII-I repetitive sequence in the taxa surveyed, varied from 306 bp to 320bp. Intertaxa sequence divergence in *Carthamus* averaged 8.59%. The average sequence divergence within the lanatus complex was 6.48%. The consensus tree and NJ tree shared similar topologies (Fig. 3i,j). The parsimony analysis of pCtHaeIII-I resulted in the strict consensus tree (Fig. 3j) having a length of 145 steps, with a consistency index of 0.8069, CI excluding uninformative characters of 0.7455, homoplasy index of 0.1931, HI excluding uninformative characters of 0.2545, and a retention index of 0.8503. There were 62 Parsimony Informative Sites (Table 2). Within the ingroup, 287 indels were present ranging from 1 to 15. One indel, a 1bp deletion separated the polyploid taxa from the diploid taxa. Another indel, a 1bp deletion was present in 6 *Carthamus* taxa (*C. tinctorius tinctorius*, *C. tinctorius inermis*, *C. oxyacantha*, *C.palaestinus*, *C.glaucus* and *C. arborescens*). There were 5 synonymous substitutions
within the ingroup. Likelihood mapping analysis of pCtHaeIII-I sequence data revealed that 81.8% of all quartets were within the three regions representing a well resolved phylogeny, 4.9% were unresolved and 13.4% showed star like evolution. The per cent of well resolved was much high in this sequence data (Fig. 3h).

The inter taxa genetic similarity indices ranged from 0.8492 between *C. tinctorius inermis* and *C. lanatus creticus* to 0.9965 between *C. oxyacantha* and *C. tinctorius inermis*, with a mean value of 0.854 (Table 3). Based on pCtHaeIII-I repetitive sequence of the 18 taxa of *Carthamus*, neighbour joining (NJ) tree yielded two distinct clades separating the diploid (2n=24) taxa (tinctorius complex) and the taxa with 2n=20 from polyploid taxa (lanatus complex) (Fig.3i). *C.palaestinus* and *C. tinctorius tinctorius* (2n=24) formed a sister group with *C. tinctorius inermis* and *C. oxyacantha* with a bootstrap value of 65, supported by 56% bootstrap confidence. *C. glaucus, C. glaucus anatolicus* and *C. boisserii* weakly allied with lanatus species complex. Lanatus complex formed a distinct clade where *C. lanatus montanus* (2n = 44) grouped with *C. lanatus lanatus* (2n=44), with a bootstrap value of 95. *C.lanatus creticus* (2n = 64) grouped with *C.lanatus* (2n = 44) and *C. lanatus turkestanicus* (2n = 64) with a bootstrap confidence of 69%. This further formed a sister group with *C. species 3* and *C. species 2* (2n = 64) with a bootstrap confidence of 72%. All the taxa of lanatus complex and the unverified polyploid *Carthamus* taxa intermingled with each other.

**pCtHaeIII-II**

The length of amplified pCtHaeIII-II repetitive sequence in the taxa surveyed, varied from 281 bp to 286 bp. Intertaxa sequence divergence in *Carthamus* averaged 14.57% in
pCtHaeIII-II. *C. arborescens* showed highest divergence. The average sequence divergence within the lanatus complex was 16.59%. The consensus tree had topology almost similar to NJ tree (4i,j). The parsimony analysis of pCtHaeIII-II resulted in the strict consensus tree (Fig. 4j) having a length of 220 steps, with a consistency index of 0.7000, CI excluding uninformative characters of 0.6489, homoplasy index of 0.3000, HI excluding uninformative characters of 0.3511, and a retention index of 0.8226. There were 100 Parsimony Informative Sites (Table 2). Within the ingroup, 44 indels were present ranging from 1 to 5. One indel, a 1bp deletion in all polyploid taxa separated them from diploid taxa. There were 4 synonymous substitutions within the ingroup.

Likelihood mapping analysis of pCtHaeIII-II sequence data revealed that 93.6% of all quartets were within the three regions representing a well-resolved phylogeny, 3.0% were unresolved and 3.4% showed star like evolution (Fig. 4h). The per cent of well resolved was higher in pCtHaeIII-II sequence data as compared to pCtHaeIII-I.

The inter taxa genetic similarity indices ranged from 0.713 between *C. lanatus turkestanicus* and *C. species 5* to 0.9965 between *C. tinctorius inermis* and *C. glaucus*; *C. palaestinus* and *C. glaucus*; and *C. species 4* and *C. lanatus creticus* with a mean value of 0.8543 (Table 4). Based on pCtHaeIII-II repetitive sequence analysis, neighbour joining (NJ) tree yielded two distinct clades (Fig.4i). First clade included all the polyploid taxa with 2n = 44 and 64 (lanatus complex). The second clade resolved into two sub clades. One of the subclade included *C. glaucus anatolicus* (2n=20), *C. boisserii* (2n=20) and *C. species 5* (2n=64) and the other included the taxa with 2n=24 (tinctorius complex). *C. glaucus anatolicus* and *C. boisserii* with a bootstrap value of 77, grouped with *C. species 5* with a bootstrap confidence of 58%. The diploid taxa grouped with 100% bootstrap confidence. *C. glaucus* and *C. tinctorius inermis* with a bootstrap value of 61, strongly allied with *C. palaestinus* with 95% bootstrap confidence. All the taxa of lanatus complex and the
unverified polyploid *Carthamus* taxa intermingled with each other.

**Screening of the Four Repetitive Sequence in Various Angiosperms**

Dot-blots of taxa other than *Carthamus* did not show any signals with pCtHaeIII-I, pCtHaeIII-I, pCtHaeIII-I and pCtTaqI-I (Fig. 3f, 4f, 5f, 6f). Moreover, no amplification product was obtained for primers designed from any of the four sequences (Fig. 3g, 4g, 5g, 6g).

**Discussion**

Repetitive sequences have proven successful in resolving species relationships and understanding genome evolution in various angiosperms (Dodsworth et al., 2015). Repetitive sequences have not been extensively studied in the family Asteraceae. There are only two major reports of repetitive sequences in Asteraceae. Subtribe Centaureinae (of tribe Cardueae) which comprises of the genus *Centaurea* and the *Carthamus* complex has proved an excellent model group to analyze evolution in satellite repetitive DNA (Bosque et al., 2013, 2014). Detailed phylogenetic and evolutionary studies have been reported in *Hinf*I satellite DNA of *Centaurea* and related species (Suarez-Santiago et al., 2007; Bosque et al., 2013). Phylogenetic studies have also been reported in genus *Carthamus* with the *KpnI* satellite repeats (Mehrotra et al., 2013).

The present study reports four novel repetitive sequences, pCtHaeIII-I, pCtHaeIII-II, pCtHaeIII-III and pCtTaqI-I which were isolated by screening of DNA libraries of *HaeIII* and *TaqI* digested DNA with total genomic DNA of *C. tinctorius*. Plasmids which gave strong hybridization signals were expected to harbor repetitive DNA sequences and served as
sources of DNA probes in further hybridization experiments after sequencing.

**Organization of Repetitive Sequences**

To understand the organization of the identified repetitive sequences, Southern hybridization was carried out using the repetitive DNA clones using the total genomic DNA of *C. tinctorius* and the corresponding restriction endonuclease (*Hae*III/*Taq*I). Various patterns of molecular hybridization were obtained with different probes. pCt*Hae*III-I formed a ladder of monomers, dimers, trimers and so on with the strongest hybridization signals corresponding to 340 bp (Figs. 3a) in *C. tinctorius* confirming its tandem nature. pCt*Hae*III-II and pCt*Hae*III-III showed a restriction profile of two bands of 285 bp and 158 bp monomer units; and 570 bp and 316 bp of dimers, respectively in a tandem fashion (Figs. 4a and 5a); and pCt*Taq*I-I showed a single band of 362 bp (Fig. 6a) implying its disperse repeat nature. Dispersed repeats have been reported in various plant species like *Vicia faba* (VfB) (Frediani et al., 1999), *Brassica nigra* (pBN-4 and pBNE8) (Kapila et al., 1996) and *Gossypium* (Zhao et al., 1998). It has been suggested that dispersed repeats also can contribute to alterations in the amount of nuclear DNA. Dispersed repeats have been known to be involved in recruitment of genes, repair of chromosomal, and induction of favorable mutants (Martignetti and Brosius 1993, Teng et al. 1996, Zeyl et al. 1996).

Similar analysis of 18 taxa of *Carthamus* with the repetitive sequences as probes revealed that these sequences are present homogenously in all the taxa studied and produced a similar pattern as in *C. tinctorius* (Figs. 3a, 4a, 5a, 6a). The observed tandem repetitive pattern of pCt*Hae*III-I, pCt*Hae*III-II, pCt*Hae*III-III is probably the result of either mutation or of methylation or both, which might have altered the restriction endonuclease recognition sequence.
The typical unit sizes of plant satellite repeats are 150-180 bp or 300-360 bp (Hemleben et al., 1982; Lin et al., 1999; Heslop-Harrison, 2000). Dimerization and formation of complex higher order repeats is a molecular feature typical for satellite DNA and has been observed in many plant species, such as Pennisetum (Ingham et al., 1993), Avena (Grebenstein et al., 1996), and Arabidopsis thaliana (Simoens et al., 1988). Such repetitive unit sizes could be favored by evolution because they might correspond to the length of the DNA strand wrapped around the nucleosome core (Fischer et al., 1994; Vershinin & Heslop-Harrison, 1998). The presence of multimers of repeat units may be due to loss of restriction enzyme sites, due to mutation events or methylation (Kulikova et al., 2004). Loss (or alteration) of a restriction sites resulting in a ladder pattern, has also been reported in two dimer sequences of radish satellite DNA and canrep sequences in Brassica (Grellet et al., 1986; Xia et al., 1993).

Repetitive sequences may be species, genus, or family specific or may even be widespread among a taxonomic class or kingdom (Mehrotra et al., 2014). However the repetitive families analyzed in the present study show specificity to a single genus (Figs. 3f,g; 4f,g; 5f,g; 6f,g), indicating that new or diverged sequences have appeared and amplified during speciation. This evolution of tandem repeats during speciation is a characteristic of many tandem array families in plants (Heslop-Harrison, 2000), and the rapid amplification of homogeneous repeat units is followed sequentially by mutation and independent amplification of coexisting sequence variants (Nijman and Lenstra, 2001).

**Characteristics of Repetitive Sequences**

The repetitive DNA sequences pCtHaeIII-I, pCtHaeIII-II, pCtHaeIII-III and pCtTaqI-I, reported in the present study shared some common features which are common to most of the
repetitive sequence families in plants. The sequences showed the presence of repetitions of poly A and T tracts scattered randomly in the sequences which are reported to be typical structures of bent DNA which may cause intrinsic binding of DNA molecules and may possibly form the heterochromatin (Koo and Crothers, 1988; Macas et al., 2000; Mehrotra et al., 2013). The AT content of pCtHaeIII-I, pCtHaeIII-II, pCtHaeIII-III and pCtTaqI-I was 72%, 55%, 49% and 61%, respectively.

Sequence analysis revealed the presence of direct, inverted, mirror, complementary repeats and microsatellites within pCtHaeIII-I sequence. The other three sequences showed only some direct repeats and microsatellites (Table 5). Presence of these internal repeats is a characteristic feature of diverse plant satellite families suggesting that the repetitive units are formed by amplification of smaller repeats (Nagaki et al., 1999). These regions are reported to be preferential sites for DNA alterations and potential substrates for homologous recombination (Gordenin et al., 2003; Linares et al., 1998; Vershinin et al., 1994, 1995, 2001).

The four repetitive sequences showed high frequencies of GG, AG and GA nearest neighbours which are characteristic of repetitive DNA families (Blake et al., 1997) and are involved in repair of heteroduplex products of unequal cross-over (Smith, 1976; Friedberg et al., 1995). Presence of GGT and GTT trinucleotides in the monomers of the four repetitive sequences is reported to aid in de novo synthesis of telomere (Tsujimoto, 1993, 1999). The pentanucleotide CAAAA in pCtHaeIII-I and pCtTaqI-I, which is supposed to be involved in a breakage-reunion mechanism of repeat sequences during arrays evolution (Appels et al., 1986; Katsiotis et al., 1998; Macas et al., 2002) may provide specific structural properties required for the amplification and maintenance of satellite DNA in the genome and may also act as a hotspot for transposon insertions (Appels and Peacock, 1971; Appels et al., 1986; Katsiotis et al., 1998; Macas et al., 2000, Ansari et al., 2004).
The polyadenylation signal, AATAAA, present in pCtHaeIII-I and pCtHaeIII-III is known to influence the transmission rate of the chromosome to descendants (Murphy and Karpen, 1995). The curvature-propensity values of the four repetitive sequences ranged between 7 and 12 (Figs. 3e, 4e, 5e, 6e) implying that the repeats are possibly curved and are responsible for tight compacting of heterochromatin (Mehrotra et al., 2013).

**Phylogenetic Analysis**

A detailed phylogenetic analysis was carried out with two tandem repetitive sequences (pCtHaeIII-I, and pCtHaeIII-II) in all the 18 taxa of *Carthamus*. Likelihood mapping analysis of the two sequences revealed that pCtHaeIII-II sequence data shows higher percentage (93.6%) of quartets within the three regions representing a well resolved phylogeny (Figs. 3h, 4h) and has a higher value of Parsimony Informative Characters. A high value of consistency index excluding uninformative characters of 0.7455, and a low value of homoplasy index of 0.1931 suggests that pCtHaeIII-II sequence is phylogenetically more informative. All phylogenetic reconstructions showed that repeat types in each taxon were more closely related to one another than to repeat types of the other taxa supporting their concerted evolution.

The present sequence assays indicated that *C. palaestinus*, *C. oxyacantha*, *C. tinctorius tinctorius* and *C. tinctorius inermis* are closely related. The grouping of *C. oxyacantha*, *C. palaestinus* with the two varieties of *C. tinctorius* in the repetitive sequence based dendrograms (Figs. 3i,j; 4i,j) strengthens the conclusion that these species are closely related, hence supporting the earlier views (Sehgal et al., 2009; Mehrotra et al., 2013).

Moreover, pCtHaeIII-I based phylogeny also suggests that *C. palaestinus* is involved in the ancestry of *C. tinctorius inermis* and *C. oxyacantha* is the probable ancestor of *C.*
tinctorius inermis. According to Imrie and Knowles (1970), C. tinctorius and C. oxyacantha have evolved concurrently from C. palaestinus through adaptive radiation. C. tinctorius is the product of selection by man in an agricultural environment whereas C. oxyacantha is a weed of disturbed areas.

The cladograms of pCtHaeIII-I and pCtHaeIII-II studied in detail (Figs. 3i,j; 4i,j), showed two major evolutionary lines in the genus Carthamus, while considering C. arborescens as the third lineage as supported by Sehgal et al., 2009. The first lineage included the diploid taxa with 2n=24 and taxa with 2n=20 (C. glaucus anatolicus and C. boisserii); and the other included the polyploid taxa with 2n=44 and 64. The repetitive sequence, pCtHaeIII-II showed distinct and better resolution between the diploid taxa with 2n=24 and taxa with 2n=20 as compared to pCtHaeIII-I. The present analysis revealed that none of the x=12 taxa grouped with polyploids. However, according to previous reports based on molecular markers like RAPD, ISSR, PCR-RFLP of chloroplast DNA, ITS and ETS sequence data and nuclear SACPD and chloroplast trnL-trnF IGS region and also the repetitive sequences, pCtKpnI-I and pCtKpnI-II, one of the lineages included all the diploid taxa with 2n=24 and the other included the taxa with 2n=20 and polyploid taxa with 2n=44 and 64 (Sasanuma et al., 2008; Sehgal et al., 2009; Mehrotra et al., 2013) which suggest that C. glaucus anatolicus and C. boisserii are likely to be involved in the ancestry of polyploids. However, the present analysis with pCtHaeIII-I, and pCtHaeIII-II, indicated grouping of taxa with 2n=20 with diploid taxa of Carthamus (2n=24) which could be due to the coevolution of these two repetitive sequences in diploid taxa of Carthamus (2n=24) and taxa with 2n=20. These two repetitive sequences seem to have originated and evolved before speciation.

According to previous studies, the first lineage comprised species of section I consisting of diploid taxa according to Ashri and Knowles (1960) or Carthamus of Hanelt (1961). Second lineage comprised species from sections II, III and IV of Ashri and Knowles (1960) or
sections Lepidopappus and Atractylis of Hanelt (1961). According to Sehgal et al. (2009),
the genus *Carthamus* should be divided into two sections, i.e. Section *Carthamus* and a
combined section of Lepidopappus and Atractylis sections, taking *C. arborescens* as the
outgroup. *C. arborescens* has been placed as the most divergent taxa and comprises the
third section Thamnacanthus (Sehgal et al., 2009).

Our study showed separate clades for diploid and polyploid taxa of *Carthamus*. Our study
has also been successful in assigning the unverified taxa sent by USDA to different
phylogenetic groups and in resolving several taxonomic considerations. Five unverified
taxa, not given any name by USDA were included in the present study out of which four
had 2n=6x=64 and the remaining one had 2n=2x=24. Four of the taxa with 2n=64 (C.
species 2, C. species 3, C. species 4 and C. species 5) clustered along with polyploid taxa
and the remaining one taxon with 2n=24 (C. species 1) clustered with diploid taxa of
*Carthamus*.

The phylogeny constructed on the basis of the repetitive sequences, pCtHaeIII-I, and
pCtHaeIII-II is more or less consistent with the evolutionary tree reconstructed from
molecular markers (Sehgal et al., 2009; Sasanuma et al., 2008) and repetitive sequences
(Mehrotra et al., 2013) except for the placement of taxa with 2n=20 with diploid taxa
(2n=24). The results presented here indicate that analysis of the distribution and
sequences of repetitive DNA is a valuable part of genome analysis and evolution.

The repetitive sequences, pCtHaeIII-I, pCtHaeIII-II analyzed in *Carthamus* species clearly
indicated concerted evolution while delineating phylogenetic relationships among the 18
taxa studied. The above understanding can assist in the marker assisted genetic
improvement/ enhancement programmes in this crop species. These novel repetitive
sequences could further be analyzed using *insitu* hybridization technique to elucidate
genome evolution of the various taxa of genus *Carthamus*. 

20
References

Ansari, H.A., Ellison, N.W., Griffiths, A.G. and Williams W.M. 2004. A lineage-specific centromeric satellite sequence in the genus *Trifolium*. Chr. Res. 12: 357-367.

Appels, R., Moran, L.B. and Gustafson, J.P. 1986. Rye heterochromatin. I. Studies on clusters of the major repeating sequence and the identification of a new dispersed repetitive sequence element. Can. J. Genet. Cytol. 28: 645-657.

Ashri, A. and Knowles, P.F. 1960. Cytogenetics of safflower (*Carthamus* L.) species and their hybrids. Agron. J. 52: 11-17.

Bertioli D.J., Vidigal B., Nielen S., Ratnaparkhe M.B., Lee T.-H., Leal-Bertioli S.C.M., Kim C., Guimarães P.M., Seijo G., Schwarzacher T., Paterson 256 A.H., Heslop-Harrison P. & Araujo A.C.G. 2013. The repetitive component of the A genome of peanut (*Arachis hypogaea*) and its role in remodelling intergenic sequence space since its evolutionary divergence from the B genome. Ann Bot 112, 545-59.

Blake, R.D., Wang, J.Z. and Beauregard, L. 1997. Repetitive sequence families in *Alces alces americana*. J. Mol. Evol. 445: 509-520.

Bosque M.E.Q, López-Flores I., Suárez-Santiago V.N., Garrido-Ramos M.A. 2013. Differential spreading of *Hinf* satellite DNA variants during radiation in Centaureinae. Ann. Bot. 112 (9): 1793-1802.

Bosque M.E.Q, López-Flores I., Suárez-Santiago V.N., Garrido-Ramos M.A. 2014. Satellite-DNA diversification and the evolution of major lineages in Cardueae (Carduoideae Asteraceae). J. Pl. Res.127(5): 575-583

Brukner, I., Sanchez, R., Suck, D. and Pongor, S. (1995) "Sequence dependent bending propensity of DNA as revealed by DNAsase I: Parameters for trinucleotides" EMBO J. 14, 1812-1818.
Dodsworth S, Chase MW, Kelly LJ, Leitch IJ, Macas J, Novák P, et al. Genomic repeat abundances contain phylogenetic signal. Systematic Biology. 2014. doi:10.1093/sysbio/syu08.

Dechyeva, D. and Schmidt, T. 2006. Molecular organization of the terminal repetitive DNA in Beta species. Chr. Res. 14: 881-897.

Elder JF Jr, Turner BJ. 1995. Concerted evolution of repetitive DNA sequences in eukaryotes. Q Rev Biol.;70(3):297-320.

Fedoroff N.V. 2012. Transposable elements, epigenetics, and genome evolution. Science 338:758-767.

Felice, B.D., Wilson, R.R., Ciarmiello, L., and Conicella, C. 2004. A novel repetitive DNA sequence in lemon (Citrus limon (L.) Burm.) and related species. J. Appl. Gen. 35(3): 315-320.

Fischer T.C., Groner, S., Zentgraf, U. and Hemleben, V. 1994. Evidence for nucleosomal phasing and a novel protein specifically binding to cucumber satellite DNA. Z Naturforsch 49c: 79-86.

Frediani, M., Gelati, M.T., Maggini, F., Galasso, I., Minelli, S., Ceccarelli, M. and Cionini, P.G., 1999. A family of dispersed repeats in the genome of Vicia faba: structure, chromosomal organization, redundancy modulation, and evolution. Chromosoma 108: 317-324.

Frello, S. and Heslop-Harrison, J.S. 2000. Repetitive DNA sequences in Crocus vernus Hill. (Iridiaceae): the genomic organization and distribution of dispersed elements in the genus Crocus and allies. Genome 43: 902-909.

Friedberg, E.C, Walker, G.C. and Siede, W. 1995. DNA Repair and Mutagenesis, ASM Press, Washington DC.

Goodsell, D.S., and Dickerson, R.E. 1994. Bending and curvature calculations in B-DNA.
Gordenin, D.A., Lobachev, K.S., Degtyareva, N.P., Malkova, A.L., Perkins, E. and Resnick, M.A. 2003. Inverted DNA repeats: a source of eukaryotic genomic instability. Trends Plant Sci. 8: 570-575.

Grebenstein, B., Grebenstein, O., Sauer, W., and Hemleben, V. 1996. Distribution and complex organization of satellite DNA sequences in Aveneae species. Genome 39: 1045-1050.

Grellet, F., Delcasso, D., Panabieres, F. and Delseny, M. 1986. Organization and evolution of a higher plant alphoid-like satellite DNA sequence. J Mol Biol. 187(4):495-507.

Hanelt, P. 1961. Zur Kenntnis von Carthamus tinctorius L. Kulturpflanze (in German). 9: 114-145.

Hemleben, V., Lewecke, B., Roth, A. and Stadler, J. 1982. Organization of highly repetitive satellite DNA of two Cucurbitaceae species (Cucumis melo and Cucumis sativus). Nucl. Acids Res. 10: 631 -643.

Heslop-Harrison, J.S. 2000. Comparative genome organization in plants: from sequence and markers to chromatin and chromosomes. Plant Cell 12: 617-635.

Imrie, B.C. and Knowles, P.F. 1970 Inheritance studies in interspecific hybrids between Carthamus flavescens and C. tinctorius, Crop Sci. 10: 349-352.

Ingham, L.D., Hanna, W.W., Baier, J.W. and Hannah L.C. 1993. Origin of the main class of repetitive DNA within selected Pennisetum species. Mol. Gen. Genet. 238: 350-356.

Katsiotis, A., Hagidimitriou, M., Douka, A. and Hatzopoulos P. 1998. Genomic organization, sequence interrelationship, and physical localization using in situ hybridization of two tandemly repeated DNA sequences in the genus Olea. Genome 41: 527-534.

Koo, H.S. and Crothers, D.M. 1988. Calibration of DNA curvature and a unified description of sequence-directed bending. Proc. Natl. Acad. Sci. U S A. 85(6): 1763-1767.
Kulikova, O., Geurts, R., Lamine, M., Kim, D.J., Cook, D.R., Leunissen, J., Jong, H.D. and Roe, B.A. and Bisseling T. 2004. Satellite repeats in the functional centromere and and pericentromeric heterochromatin of *Medicago truncatula*. Chromosoma. 113(6):276-83.

Li S.F., Zhang G.J., Yuan J.H., Deng C.L., Gao W.J. 2016. Repetitive sequences and epigenetic modification: inseparable partners play important roles in the evolution of plant sex chromosomes. Planta, 243: 1083-1095.

Lin, X.Y., Kaul S. *et al.* 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature 402: 761-768.

Linares, C., Ferrer, E. and Fominaya, A. 1998. Discrimination of the closely related A and D genomes of the hexaploid oat *Avena sativa* L. Proc. Natl. Acad. Sci. USA. 95: 12450-12455.

Macas, J., Meszaros, T. and Nouzova, M. 2002. PlantSat: a specialized database for plant satellite repeats. Bioinformatics 18: 28-35.

Macas, J., Pozarkova, D., Navratilova, A., Nouzova, M., Neumann, P. 2000. Two new families of tandem repeats isolated from genus *Vicia* using genomic self-priming PCR. Mol. Gen. Genet. 263:741-751.

Mehrotra S, Goel S, Sharma S, Raina SN, Rajpal VR. 2013. Sequence analysis of KpnI repeat sequences to revisit the phylogeny of the genus *Carthamus*L..Applied Biochemistry and Biotechnology 169 (4): 1109-1125.

Mehrotra S, Goyal V. 2014. Repetitive sequences in plant nuclear DNA-types, distribution, evolution and function. Genomics, Proteomics and Bioinformatics 12(4): 164-171.

Murphy, T.D. and Karpen, G.H. 1995. Localization of centromere function in a *Drosophila* minichromosome. Cell 82: 599-609.

Nagaki, K., Kishii M., Tsujimoto, H. and Sasakuma, T. 1999. Tandem repetitive *Afa*-family sequences from *Leymus racemosus* and *Psathyrostachys juncea* (Poaceae). Genome 42: 1258-1260.
Nijman IJ, Lenstra, JA. 2001. Mutation and recombination in cattle satellite DNA: a feedback model for the evolution of satellite DNA repeats. J. Mol. Evol. 52: 361-371.

Porebski, S., Bailey, L.G. and Baum, B.R. 1997. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Mol. Biol. Rep. 15: 8-15.

Raina, S.N., Sharma, S., Sasakuma, T., Kishii, M. and Vaishnavi, S., 2005. Novel repeated DNA sequences in safflower (Carthamus tinctorius L.) (Asteraceae): cloning, sequencing and physical mapping by fluorescent in situ hybridization. J. Hered. 96: 424-429.

Sasanuma, T., Sehgal, D., Sasakuma, T. and Raina, S.N., 2008. Phylogenetic analysis of Carthamus species based on nucleotide sequence of nuclear-encoded SACPD gene and chloroplast trnLF-IGS region. Genome 51: 721-727.

Sehgal, D., Raina, S. N., Devarumath, R. M., Sasanuma, T., & Sasakuma, T. (2009). Nuclear DNA assay in solving issues related to ancestry of the domesticated diploid safflower (Carthamus tinctorius L.) and the polyploid (Carthamus) taxa, and phylogenetic and genomic relationships in the genus Carthamus L. (Asteraceae). Molecular Phylogenetics and Evolution, 53, 631-644.

Simoens, C.R., Gielen, J., Van Montagu, M., and Inzé, D. 1988. Characterization of highly repetitive sequences of Arabidopsis thaliana. Nuc. Acids Res. 16 (14): 6753-6766.

Smith, G.P. 1976. Evolution of repeated DNA sequences by unequal crossover. Science 191:528-535.

Strimmer, K. and Haesler, A.V. 1997 Quartet puzzling: a quartet maximum-likelihood method for reconstruction tree topologies. Mol. Biol. Evol. 13: 964-969.

Suárez-Santiago, V.N., Blanca, G., Rejón, M.R. and Garrido-Ramos, M.A. 2007b. Satellite-DNA evolutionary patterns under a complex evolutionary scenario: The case of Acrolophus subgroup (Centaurea L., Compositae) from the western Mediterranean. Gene 392: 283-
Svitashev, T., Bryngelsson, A., Vershinin, C., Pedersen T. S. and Bothmer, R.V. 1994. Phylogenetic analysis of the genus *Hordeum* using repetitive DNA sequences. Theor. Appl. Genet. 89: 801-810.

Swofford, D.L., 2002. PAUP. Phylogenetic Analysis Using Parsimony (and other methods), version 4. Sinauer Associates, Sunderland, Massachusetts.

Thompson J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25: 4876-4882.

Tsujimoto, H. 1993. Molecular cytological evidence for gradual telomere synthesis at the broken chromosome ends in wheat. J. Pl. Res. 106: 239-244.

Tsujimoto H., Usami N., Hasegawa K., Yamada T., Nagaki K., Sasakuma T. 1999. De novo synthesis of telomere sequences at the healed breakpoints of wheat deletion chromosomes. Mol Gen Genet 262: 851-856.

Vershinin, A.V., Alkhimova, A.G., Heslop-Harrison J.S., Potapova, T.A. and Omelianchuk, N. 2001. Different patterns in molecular evolution of the Triticeae. Hereditas 135:153-160.

Vershinin, A.V. and Heslop-Harrison, J.S. 1998. Comparative analysis of the nucleosomal structure of rye, wheat and their relatives. Plant Mol. Biol. 36 (1): 149-161.

Vershinin, A.V., Schwarzacher, T. and Heslop-Harrison, J.S. 1995. The large-scale genomic organization of repetitive DNA families at the telomeres of rye chromosomes. Plant Cell 7: 1823-1833.

Vershinin, A.V., Svitashev, S., Gummesson, P.O., Salomon, B., Bothmer, R.V. and Bryngelsson, T. 1994. Characterization of a family of tandemly repeated DNA sequences in Triticeae. Theor. Appl. Genet. 89: 217-225.

Xia, X., Selvaraj, G. and Bertrand, H. 1993. Structure and evolution of a highly repetitive
DNA sequence from *Brassica napus*. Pl. Mol. Biol. 21: 213-224.

Zentgraf, U., King, K. and Hemleben, V. 1992. Repetitive sequences are valuable as molecular markers in studies of phylogenetic relationships within the genus *Cucumis*. Acta Botanica Neerlandica, 41: 397-406.

Tables

**Table 1:** Repeat sequences isolated from *Carthamus tinctorius*. The pentanucleotides CAAAA and TTTTG are in bold face.

**pCtHaeIII-I**

TCCCCAACTAATCTAATATAAAATTACAATAGGGTTGCAAAATGGCAGCAGTGCCTCCCTTATCAAAAAGA

**pCtHaeIII-II**

TCAACTATAGCAGCTTTGAGTTTCTCATGACACTCTTGCAGTCGTGTAAGGTCTAAGGCTCAAGATTGACCTTTTC

**pCtHaeIII-III**

AATCGCGTTAGAAAGGAAGACCAAGACCTTTGCGGAATAAAAAGAATTGGGACTAGTAAGGCTCAAGATTGACCTTTTC

**pCtTaqI-I**

ATGATGATCAATCAACCATC

**Table 7:** The angiosperm taxa used in the present study

| S. No. | Genus and species | Family                  |
|--------|-------------------|-------------------------|
| 1.     | *Centaurea cyanus* L. | Asteraceae (Cynareae)   |
| 2.     | *Chrysanthemum coronarium* L. | Asteraceae (Anthemidae) |
| 3.     | *Eclipta prostrata* L. | Asteraceae (Heliantheae) |
| 4.     | *Ageratum conizoides* L. | Asteraceae (Eupatorieae) |
| 5.     | *Brassica campestris* Czern and Cross | Cruciferae            |
| 6.     | *Lycopersicon esculentum* Mill. | Solanaceae           |
| 7.     | *Lathyrus odoratus* L. | Fabaceae               |
| 8.     | *Eleusine coracana* Gaertn. | Poaceae               |
| 9.     | *Ranunculus sceleratus* L. | Ranunculaceae         |
| 10.    | *Morus alba* L.     | Moraceae               |
| 11.    | *Mangifera indica* L. | Anacardiaceae         |

Table 6: Primers designed for amplification of the four repetitive sequences

| Primer | Primer sequence | Length |
|--------|-----------------|--------|
|        |                 |        |
(bp)

|          | Sequence                                | Length |
|----------|-----------------------------------------|--------|
| pCtHaeIII-IF | 5’AATAAAATTACAATAGGGTTGCAAATG3’         | 27     |
| pCtHaeIII-IR | 3’T TTG GACCAAAAGTTTTTAATTG5’           | 23     |
| pCtHaeIII-IIIF | 5’CCTCAACTATAGCGAGCTCTTTT3’            | 23     |
| pCtHaeIII-IIIR | 3’CCTGTCTGATGGCTATCATCG5’              | 21     |
| pCtHaeIII-IIIIF | 5’ATCGCGGTGTAGAAGAATG3’            | 20     |
| pCtHaeIII-IIIIR | 3’CGCTCAAGCTTTGGAACAAT5’            | 20     |
| pCtTaqI-IF  | 5’ATGATGATCAATCAACCATCTTTT3’           | 25     |
| pCtTaqI-IR  | 3’CCAATTATCATCTATCTCAATCCCCTTA5’       | 27     |

**Table 5:** Overall homologies between pCtKpnI-I, pCtKpnI-II, pCtHaeIII-I, pCtHaeIII-II, pCtHaeIII-III and pTaqI-I
|          | pCtKpnI-I | pCtKpnI-II | pCtHaeIII-I | pCtHaeIII-II | pCtHaeIII-III |
|----------|-----------|------------|-------------|---------------|----------------|
| pCtKpnI-I | 100%      | 49%        | 45%         | 46%           | 46%            |
| pCtKpnI-II| 100%      | 47%        | 48%         | 47%           |                |
| pCtHaeIII-I|          |            | 100%        | 46%           | 46%            |
| pCtHaeIII-II|         |            |             | 100%          | 48%            |
| pCtHaeIII-III|        |            |             |               | 100%           |

**Table 2: Nucleotide sequence characteristics of the amplified pCtHaeIII-I and pCtHaeIII-II**
| Sequence characteristics                                      | pCtHaeIII-I | pCtHaeIII-II |
|---------------------------------------------------------------|-------------|--------------|
| Length in bp                                                  | 340         |              |
| Aligned length                                               | 310         |              |
| Average sequence divergence in percentage                    | 8.59        | 1            |
| Size of indels in bp                                         | 1-15        |              |
| Number of constant sites                                     | 215         |              |
| Number of parsimony informative sites                        | 62          |              |
| Expected Transition/ Tranversion ratio                       | 0.84        | 1            |
| Expected pyridine Transition/ Purine transition ratio        | 0.88        | 1            |
| Species                       | Similarity 1 | Similarity 2 | Similarity 3 | Similarity 4 | Similarity 5 |
|-------------------------------|--------------|--------------|--------------|--------------|--------------|
| *C. tinctorius* var. *tinctorius* | 1.0000       |              |              |              |              |
| *C. species* 1                | 0.9787       | 1.0000       |              |              |              |
| *C. glaucus*                  | 0.9608       | 0.9578       | 1.0000       |              |              |
| *C. tinctorius* var. *inermis* | 0.9644       | 0.9613       | 0.9960       | 1.0000       |              |
| *C. palaestinus*              | 0.9678       | 0.9613       | 0.9960       | 0.9920       | 1.0000       |
| *C. oxyacantha*               | 0.9750       | 0.9609       | 0.9750       | 0.9780       | 1.0000       |
| *C. glaucus ssp. anatolicus*  | 0.8672       | 0.8762       | 0.8722       | 0.8710       | 0.8744       |
| *C. boisserii*                | 0.8751       | 0.8733       | 0.8760       | 0.8750       | 0.8821       |
| *C. species* 5                | 0.8427       | 0.8491       | 0.8414       | 0.8480       | 0.8500       |
| *C. arborescens*              | 0.8645       | 0.8772       | 0.8661       | 0.8690       | 0.8714       |
| *C. lanatus ssp. creticus*    | 0.8249       | 0.8308       | 0.8197       | 0.8230       | 0.8247       |
| *C. species* 4                | 0.8213       | 0.8273       | 0.8161       | 0.8190       | 0.8211       |
| *C. lanatus*                  | 0.7928       | 0.7956       | 0.7917       | 0.7950       | 0.7966       |
| *C. species* 3                | 0.8337       | 0.8397       | 0.8316       | 0.8350       | 0.8407       |
| *C. species* 2                | 0.8264       | 0.8325       | 0.8245       | 0.8280       | 0.8335       |
| *C. lanatus ssp. lanatus*     | 0.7322       | 0.7392       | 0.7278       | 0.7310       | 0.7320       |
| *C. lanatus ssp. monatnus*    | 0.7321       | 0.7392       | 0.7277       | 0.7310       | 0.7320       |
| *C. lanatus ssp. turkestanicus* | 0.7320     | 0.7390       | 0.7277       | 0.7270       | 0.7354       |

**Table 4:** Similarity matrix of pCtHaeIII-II in 18 taxa of *Carthamus*
| Species          | Similarity 1 | Similarity 2 | Similarity 3 | Similarity 4 | Similarity 5 |
|------------------|--------------|--------------|--------------|--------------|--------------|
| *C. lanatus*     | 1.0000       |              |              |              |              |
| ssp. montanus    | 0.9727       | 1.0000       |              |              |              |
| ssp. lanatus     | 0.9162       | 0.9097       | 1.0000       |              |              |
| ssp. turkestanicus | 0.9025      | 0.8961       | 0.9656       | 0.9760       | 1.0000       |
| ssp. creticus    | 0.9166       | 0.9101       | 0.9657       | 0.9795       | 0.9795       |
| C. species 3     | 1.0000       |              |              |              |              |
| C. species 2     | 0.9134       | 0.9068       | 0.9624       | 0.9761       | 0.9760       |
| C. species 4     | 0.9297       | 0.9230       | 0.9720       | 0.9756       | 0.9756       |
| C. species 5     | 0.9232       | 0.9199       | 0.9515       | 0.9548       | 0.9549       |
| *C. glaucus*     | 0.9379       | 0.9242       | 0.9129       | 0.9236       | 0.9203       |
| ssp. anatolicus  |              |              |              |              |              |
| C. boisserii     | 0.9416       | 0.9380       | 0.9032       | 0.9138       | 0.9103       |
| *C. glaucus*     | 0.9314       | 0.9241       | 0.8895       | 0.8931       | 0.8896       |
| C. palaestinus   | 0.9072       | 0.9069       | 0.8617       | 0.8617       | 0.8582       |
| C. tinctorius    | 0.9097       | 0.9095       | 0.8708       | 0.8744       | 0.8708       |
| var. tinctorius  |              |              |              |              |              |
| *C. tinctorius*  | 0.8823       | 0.8823       | 0.8532       | 0.8565       | 0.8492       |
| var. inermis     |              |              |              |              |              |
| *C. oxyacantha*  | 0.8857       | 0.8857       | 0.8568       | 0.8602       | 0.8528       |
| *C. arborescens* | 0.9261       | 0.9120       | 0.8899       | 0.8929       | 0.8858       |

**Table 3:** Similarity matrix of pCTHaeIII-I in 18 taxa of *Carthamus*

Figures
Figure 1: Restriction digestion of genomic DNA of *Carthamus tinctorius* with 1) *Hae* III and 2) *Taq* I

Figure 2: Colony Hybridization of i) *Hae*III and ii) *Taq*I library of *Carthamus tinctorius* with total genomic DNA of *Carthamus tinctorius* as probe. a) pCtHaeIII-I, b) pCtHaeIII-II, c) pCtHaeIII-III, d) pCtTaqI-I and e) control
Figure 3
Figure 5a: Southern hybridization of total genomic DNA of 10 taxa of Caricaceae probed with pC3H3-II-II.

Figure 5b: Dot plots of genomic DNA from 10 taxa of Caricaceae probed using pC3H3-II-II repeats.

Figure 5c: PCR amplification of 108 bp pC3H3-II-II in 10 taxa of Caricaceae.

Figure 5d: Dot-matrix analysis of all comparisons of pC3H3-II-II repeats using MegaBlot application from Lasergene™ software package. The matrix stringency was set to: a) 8/10, b) 4/10 and c) 2/10 identical nucleotides.

Figure 5e: Curvature-proporty plot of pC3H3-II-II showing the presence of a curvature propensity in the region between nucleotides 110 and 120.

Figure 5f: Dot-blots of genomic DNAs from genera of 12 different families probed using pC3H3-II-II-II repeat.

Figure 5g: PCR amplification of pC3H3-II-II-II repetitive sequences in genera of different families of angiosperms including Asteraceae.
Figure 6a: Southern hybridization of total genomic DNA of 18 taxa of Carthamus probed with pCItaq-l-I.

Figure 6b: Dot-blot analysis of genomic DNAs from 18 taxa of Carthamus probed using pCItaq-l-I repeats.

Figure 6c: PCR amplification of 362 bp pCItaq-l-I in 18 taxa of Carthamus.

Figure 6d: Dot-matrix analysis of self comparisons of pCItaq-l-I repeats using Megalign application from Lasergene 9.0 software package. The matrix stringency was set to a) 3/100, b) 6/100, c) 8/100, d) 15/100, and e) 20/70 identical nucleotides.

Figure 6e: Curvature-propensity plot of pCItaq-l-I showing the presence of a curvature propensity in the region between nucleotides 8 and 58.

Figure 6f: Dot-blot analysis of genomic DNAs from genera of 12 different families probed using pCItaq-l-I repeats.

Figure 6g: PCR amplification of pCItaq-l-I repetitive sequences in genera of different families of angiosperms including Asteraceae.