Journal of Applied Biology & Biotechnology Vol. 9(2), pp. 36-59, Mar-Apr, 2021
Available online at http://www.jabonline.in
DOI: 10.7324/JABB.2021.9204

Plasticity of tandem repeats in expressed sequence tags of angiospermic and non-angiospermic species: Insight into cladistic, phenetic, and elementary explorations

Shamshad Ul Haq[1,2], Prerna Dhingra[2], Meenakshi Sharma[2], S. L. Kothari[3], Sumita Kachhwaha[2]*

[1]Interdisciplinary programme of life Science for Advance Research and Education (IPLS), University of Rajasthan, Jaipur-302004, India
[2]Department of Botany, University of Rajasthan, Jaipur-302004, India
[3]Amity Institute of Biotechnology, Amity University Rajasthan, Jaipur 302006, India

ARTICLE INFO

Article history:
Received on: August 02, 2020
Accepted on: December 05, 2020
Available online: March 10, 2021

Key words:
Angiospermic species,
Non-angiospermic species,
Expressed sequence tags,
Tandem repeats,
Simple sequence repeats,
Evolutionary clades or phylogenetic clades.

ABSTRACT

Angiospermic and non-angiospermic groups comprise plant species representing short and long range of discrepancies in their morphological, physiological, biochemical, molecular, and developmental processes. Analysis at molecular level plays crucial role to ascertain the heterogeneity within and across the species. The tandem repetitive DNA elements are one of the most important elements which play a significant role in various genetic and genomic applications. Therefore, the plasticity of tandem repetitive DNA element especially simple sequence repeats (SSRs) was analyzed in the expressed sequenced tags (ESTs) of both angiospermic and non-angiospermic species comprising 75 plant species belonging to different evolutionary clades such as algae, fungi, bryophytes, pteridophytes, gymnosperms, dicots, and monocots. Significantly, angiospermic and non-angiospermic species represented distinctiveness at GC content, SSR incidence and SSR motif distributions in their EST sequences. Notably, non-angiosperms revealed more GC-content compared to angiosperms but angiosperms depicted enhanced tandem repetitions (EST-SSRs) compared to non-angiosperms. Among different types of SSRs, mononucleotide SSRs represented widespread distribution followed by trinucleotide SSRs distribution in both angiosperms and non-angiosperms. In general, SSR motifs such as A/T, AG/CT, AAG/CTT, and CCG/CGG were found to be more repeated but highly complex motifs patterns were observed within hexa, penta, and tetranucleotide SSRs, respectively. Thus, a quantity of nexus and diversification were observed within and across the species as well as evolutionary clades. To infer, differential patterns of DNA tandem identified within ESTs can unfold the genetic polymorphism, diversification, conservation, and genome evolution within and across species.

1. INTRODUCTION

Angiospermic and non-angiospermic groups encompass an enormous diversity of plant species which represent homogenous as well as heterogeneous relationships at morphological, physiological, biochemical, and molecular levels. These kinds of relationships among species, allows to strengthen the adaptability, flexibility, and survivability of species or populations against different ecological conditions or environmental fluctuations. Last few decades, a swift in genetic and cytogenetic explorations were observed which provided thorough details of genome organization, genetic diversity, and genome evolution through the analysis of nuclear DNA, organelle DNA, expressed sequence tag (EST) and chromosomal aberration, etc. While, repetitive DNA elements-based studies were found to be more in practice due to their major portion in nuclear genome as well as expressed region of genome (EST) among eukaryotic organisms. Repetitive DNA elements are present in the form of tandem repeats (microsatellite or simple sequence repeat, minisatellite, etc.) and interspersed repeats (transposons, retrotransposons, etc.). These tandems can repeat massive times and might be responsible for structural and functional participations in the genome. In several studies, DNA element is observed to be very important for its involvement in genome size, genetic diversity, genome organization, conservation, and evolution within and across the species and taxa [1-3].

Especially, expressed sequence tags (EST) are the most important genomic resources owing to their functional role in the genome and can serve as a connection between genomics and molecular ecology [4]. Last few decades, ESTs have gained momentum in extensive and rapid applications for gene discovery, gene annotation, genetic polymorphism, transcriptomics profiling, and proteomic exploration [5,6]. ESTs are randomly selected, unedited, and single pass sequencing of clones from cDNA libraries, ranging from 200 to 800 nucleotide bases. These sequences have gained advantages over whole genome sequencing because of their direct association in the gene function. Besides this, it

**Corresponding Author:*
Sumita Kachhwaha, Department of Botany, University of Rajasthan, Jaipur-302004, India. E-mail: kachhwahasumita@rediffmail.com

is a rapid approach, less expensive, easy handling, and consuming less time [7]. Astonishing involvement of ESTs has been confirmed in identification of miRNA precursors and targets [8-10], transcriptome analysis using cDNA microarrays [11-13], and gene discovery and gene expression analysis [8,14-17].

Moreover, EST sequences are also very important resource for tandem repetitive DNA elements especially simple sequence repeats (SSRs) which serve as molecular markers and are very useful for variety of genetic or genomic applications. Microsatellites or SSRs are tandemly repeated DNA sequences generally ranging from 1 to 6 nucleotides long which are dispersed randomly and ubiquitously throughout the genomes in both prokaryotic and eukaryotic organisms [18-20]. They are frequently present in both coding and non-coding regions of genome [21]. Thus, EST-SSRs based studies are found to be more implemented in various plant genetic applications, namely, genetic diversity, ecological, evolutionary, phylogeny, taxonomical, and comparative genomic studies [22,23]. All these genetic applications became possible due to the multi-allelic nature, co-dominancy, and high reproducibility of microsatellite (SSRs) [24]. SSRs markers also allow the identification of prototype of gene content, generation of genetic relatedness, and frequency of genetic drift which are very crucial factors in the population for recognizing the conservation units [25]. In addition, the use of publicly available EST libraries has shown an alternative way for EST-SSRs resource which has proved to be a powerful and promising tool for variety of applications, namely, population genetics, biodiversity, genetic drift, high resolution genetic maps, gene mapping, QTL (quantitative trait locus), germplasm characterization, cultivar identification, paternity analyses, and marker assisted breeding [8,26-32].

The present study provides the information about the distribution dynamic of DNA tandem repeats in the ESTs of angiospermic and non-angiospermic plant species. For the analysis, a total of 75 species were selected under different phylogenetic lineage such as, algae, fungi, bryophytes, pteridophytes, gymnosperms dicots, and monocots. Furthermore, ESTs of selected species were used for the analysis of SSRs distribution within and across different species and imperative of EST-SSRs were discussed according to their origin, distribution, conservation, and evolution.

2. MATERIALS AND METHODS

2.1. Plant Materials

The 75 different plant species belonging to six distinct evolutionary clades were used for the tandem repetitive DNA elements (EST-SSRs) analysis. Out of 75 species, 30 species were non-angiosperms which included 10 species of algae, 10 species of fungi, 3 species of bryophytes, 2 species of pteridophytes, and 5 species of gymnosperms. Among angiosperms, 34 species were dicots and 11 species were monocots, as shown in Table 1.

2.2. Expressed Sequence Tags Sequences Retrieval

A total of 43,52,515 partial EST transcripts were examined from National Center for Biotechnology Institute (NCBI), a public database which provides easy accessibility and user-friendly platform for the

Table 1: Details of non-angiospermic species and angiospermic species used for tandem repeat analysis.

Non-angiospermic species				
Algae	***Fungi***	**Bryophytes**	**Pteridophytes**	**Gymnosperms**
Chaetosphaeridium globosum	*Albugo candida*	*Marchantia polymorpha*	*Adiantum capillus-veneris*	*Ginkgo biloba*
Chlamydomonas reinhardtii	*Aspergillus niger*	*Physcomitrella patens*	*Selaginella moellendorffii*	*Gnetum gnemon*
Chlorella variabilis	*Cercospora zeae-maydis*	*Syntrichia ruralis*		*Cycas rumphii*
Chlorokybus atmophyticus	*Fusarium graminearum*			*Pinus pinaster*
Ectocarpus siliculosus	*Mucor circinelloides*			*Welwitschia mirabilis*
Klebsormidium flaccidum	*Neurospora crassa*			
Mesotigma viride	*Phytophthora infestans*			
Nitella hyalina	*Puccinia triticina*			
Porphyra yezoensis	*Saccharomyces cerevisiae*			
Volvox carteri	*Ustilago maydis*			
Angiospermic species				
Dicots			**Monocots**	
Cantharanthus roseus	*Euphorbia esula*	*Pisum sativum*	*Avena barbata*	
Ocimum basilicum	*Hevea brasiliensis*	*Fragaria vesca*	*Avena sativa*	
Capsicum annuum	*Manihot esculenta*	*Malus domestica*	*Cenchrus ciliaris*	
Nicotiana tabacum	*Ricinus communis*	*Prunus persica*	*Hordium vulgare*	
Solanum lycopersicum	*Arachis hypogaea*	*Vitis vinifera*	*Oryza sativa*	
Daucus carota	*Cajanus cajan*	*Arabidopsis thaliana*	*Secale cereale*	
Panax ginseng	*Cicer arietinum*	*Brassica napus*	*Sorghum bicolor*	
Artemisia annua	*Glycine max*	*Raphanus sativus*	*Sorghum propinquum*	
Helianthus annuus	*Lotus japonicus*	*Carica papaya*	*Triticum aestivum*	
Citrullus lanatus	*Medicago truncatula*	*Gossypium hirsutum*	*Zea mays*	
Cucumis melo	*Trifolium pratense*	*Theobroma cacao*	*Musa acuminata*	
Liriodendron tulipifera				

analysis. The batch files of EST sequences were retrieved as FASTA format for the selected plant species and range was fixed between the limit: 10 thousand to 100 thousand sequences, according to the availability of sequence information for the selected species at NCBI as well as system competency.

2.3. EST Sequences Assembling and Computational Analysis

For the analysis, all the retrieved EST sequences were subjected to sequence assembling program for minimization of sequences redundancy through CAP3 platform using default parameters. The CAP3 assembly program has a capability to clip 5′ and 3′ low-quality regions of reads. As well, it uses base quality values in computation of overlaps between reads, construction of multiple sequence alignments of reads, and generation of consensus sequences [33]. Furthermore, some basic computational analyses were performed for all the assembled EST sequences using Perl script from the internet bioinformatics resources.

2.4. Simple Sequence Repeats (SSRs) or Microsatellites Screening

To study the distribution dynamics of SSRs, all the assembled EST sequences of 75 species were subjected to MIcroSAtellite identification tool (MISA) (http://pgrc.ipk-gatersleben.de/misa/). It is Perl command line exercise for identifications and characterizations of different types of SSRs. It produces separate output text files with the following information such as sequence name, number of SSRs, type of SSR, types of SSR motif, SSR position, repeat length, and repeat number. Moreover, only mono to hexa nucleotide SSRs were considered and limitation for SSRs detection were 10, 6, 5, 5, and 5 repeat units for mono, di, tri, tetra, penta, and hexa nucleotides repeats, respectively.

3. RESULTS AND DISCUSSION

3.1. EST Sequences Characterization

The comparative analysis of EST-SSRs was performed among 75 different plant species belonging to diverse phylogenetic lineage such as algae, fungi, bryophytes, pteridophytes, gymnosperms, dicots, and monocots. A total of 4352515 (4.35 millions) EST transcripts were examined and 1306939 non-redundant ESTs (NR-ESTs) sequences were obtained after assembling [Figures 1 and 2]. A set of 528211 contigs were obtained with higher N50 value compare to N25 and N75 and N50 value was ranged from 500 bp to 1200 bp with an average of 900bp. Similarly, a total of 778728 singlets were obtained and sequence lengths ranged from 500bp to 1600bp with an average of 800bp in size. The overall average length of NR-ESTs sequence was 717.69 bp long ranging from 513.56 bp to 1033.83 bp long which is quite comparable with previous studies in the different plant species [34,35]. It was observed that there were deviations in the number of reads among contigs and singlets. This variation may be explained by related or distal part of the sequencing and inadequacy of the sequencing data of the species and used parameter in the assembling pipeline. Regarding to mean values of sequence length, non-angiosperms showed high average sequence length as compared to angiosperms. Among phylogenetic clade, bryophytes and pteridophytes revealed high average sequence length coverage and lowest was observed in gymnosperms [Figure 3]. Among species, high average sequence length was reported in *Albugo candida* (1033.83bp) followed by *Selaginella moellendorffii* (991.30bp), and *Chlorokybus atmophyticus* (953.14bp). Similarly, lowest average length was seen as 513.56 bp and 524.43 bp in *Lotus japonicas* and *Theobroma cacao,* respectively [Additional file 1].

3.2 Distribution of GC-content in ESTs

Comparative distribution of GC-content was examined in NR-ESTs belonging to 75 different species. In general, the average GC-content was 46.61%, ranging from 38.61% to 65.16% which is in wake of earlier observations within various plant species [36,37]. Significantly, higher GC-content was found commonly in non-angiosperms compared to angiosperms. Within evolutionary clades, algae showed relatively increased GC-content followed by fungi, bryophytes,

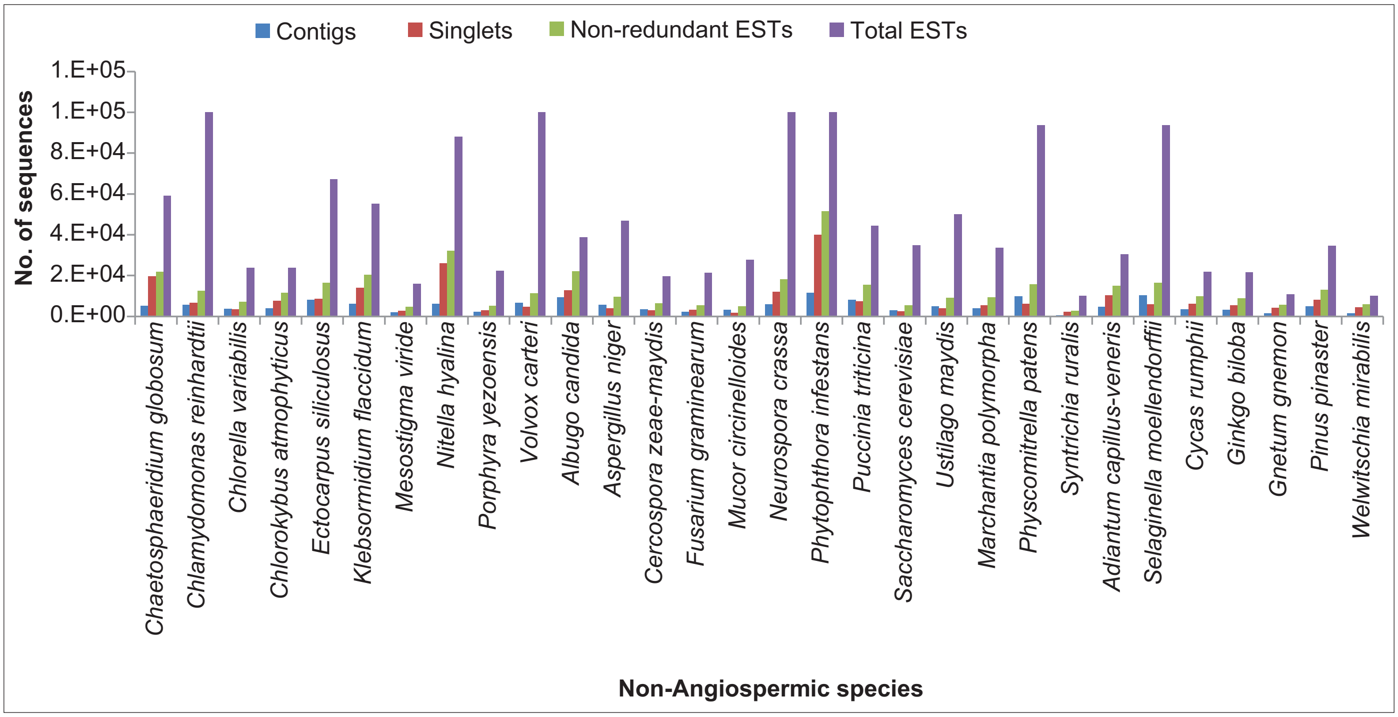


Figure 1: Comparative details of EST characterizations among 30 non-angiospermic species.

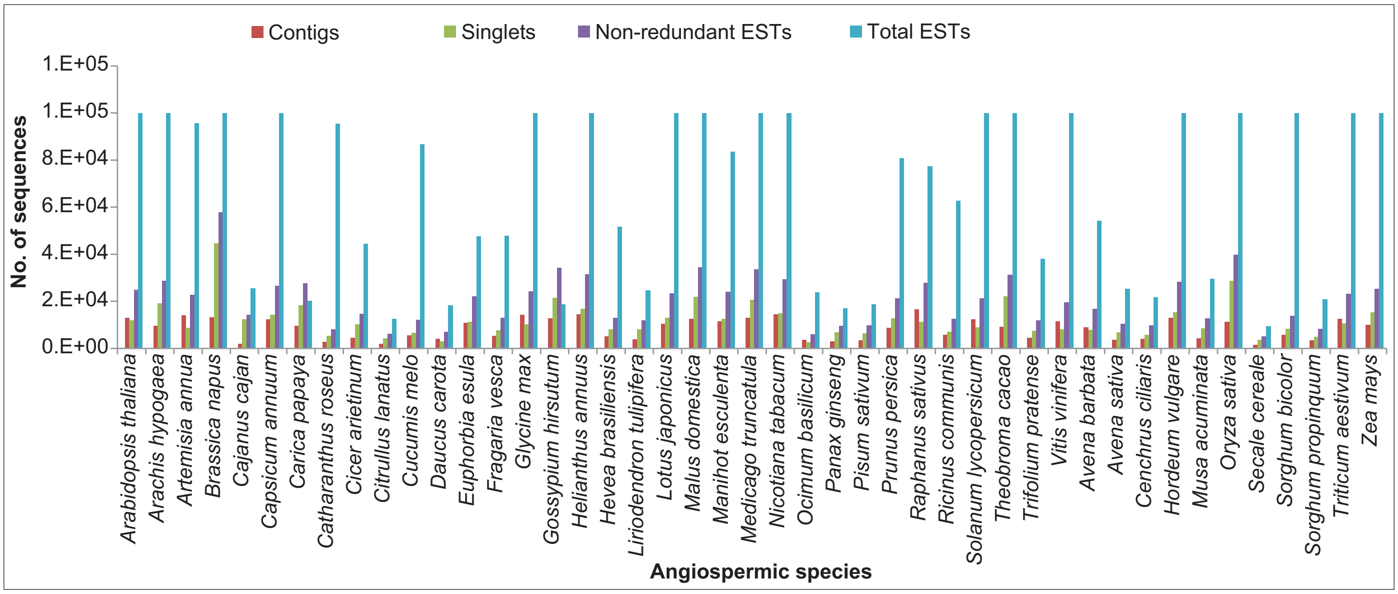


Figure 2: Comparative details of EST characterizations among 45 angiospermic species.

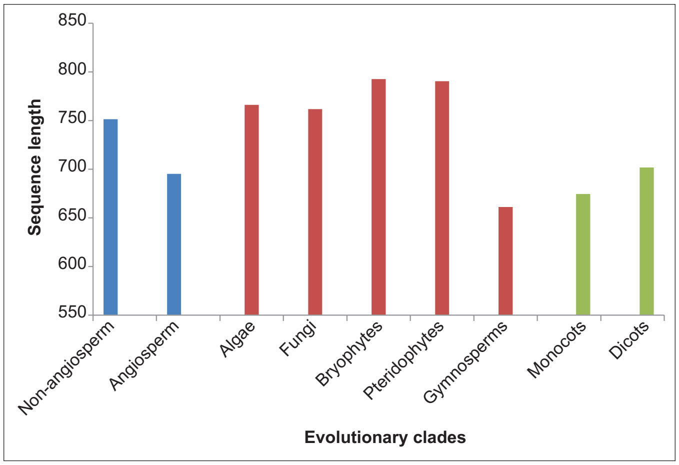


Figure 3: Average sequence length (nucleotides) distribution in non-redundant EST sequences among different evolutionary clades.

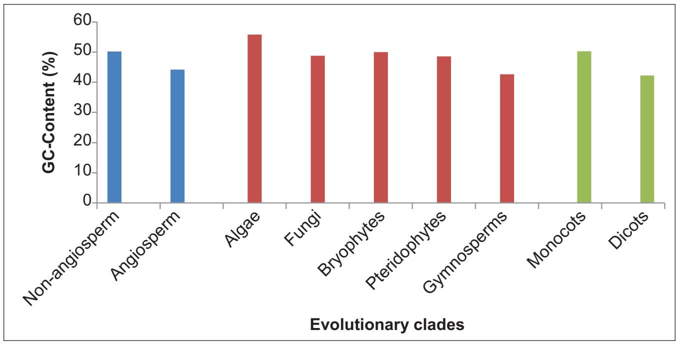


Figure 4: Average GC-content distributions in non-redundant EST sequences among different evolutionary clades.

pteridophytes, and gymnosperms, respectively [Figure 4]. Among non-angiospermic species, significantly increased GC value was observed in algae, *Chlorella variabilis* and *Klebsormidium flaccidum*; in fungi, *Ustilago maydis* and *Cercospora zeae-maydis*; in bryophyte, *Syntrichia ruralis*; in pteridophyte, *Selaginella moellendorffii,* and in gymnosperm, *Gnetum gnemon* [Additional file 2]. Among angiosperms, an increased GC-content was identified in monocots compared to dicots which are in agreement with previous study [38]. While, within dicot species, rosid species showed relatively enhanced GC-content related to asteroid plant species but no skewness was observed within asteroid and rosid species. For angiospermic species, the rise of GC value was seen in dicot species namely; *Brassica napus*, *Fragaria vesca,* and *Ocimum basilicum* while *Zea mays* and *Sorghum propinquum* represented high GC value in monocots [Additional file 3]. Notably, GC-content is considered as very important parameter reflecting the information about gene structure (intron size and number), thermostability, gene regulation, and evolution [39,40]. While, more GC-content is indicative of high gene density and their compactness [41,42], display earlier replication timing [43], influences rates of recombination [44], and determining of physical and physiological properties of DNA [45].

3.3. Frequency Distribution of SSRs in ESTs

The circulation of SSRs was examined among ESTs of selected species and mainly mono to hexanucleotide SSRs were considered. A total of 678260 SSRs were identified and an average frequency distribution was 9.65%, ranged from 1% to 24.81% excluding mononucleotide SSRs. The range of SSR distribution in the present study is found to be exhibit similarity with previous studies reported in various plant species [27,46-50]. Twisting in SSR frequency can be explained by various factors used such as, types of SSR mining tool, parameters used for mining, and wealth of sequences which may develop significant differences in the SSRs frequency distributions. Comparatively increased SSR incidence was observed in angiosperms with 10.50% frequency distribution in comparison to non-angiosperms with 8.42% frequency distribution [Figure 5]. Among non-angiosperms, increased SSRs distribution was identified in pteridophytes and algae while lowest was seen in gymnosperms. For angiosperms, monocots showed more SSRs incidence than dicots and this increased SSR incidence can be explained by highly dynamic nature of angiosperm genomes, large genome size, and their structure [51] as well as rise of polyploidy in higher plants may also be responsible for changing of SSR incidences. It appears that SSR incidence was inversely proportional to GC-content as angiosperms revealed a reduced GC-content (44.21%) with high SSRs occurrence and non-angiosperms showed high GC-content (50.22%) with low SSRs occurrence. Therefore, the nature of divergence in the SSR incidence, SSR length, motif structure, and GC-content are very important influencing factors for conservation and evolutionary action [52].

Moreover, the randomness in the average value with extremely reduced SSR frequency was observed in alga, *Klebsormidium flaccidum* (1.37%); fungi, *Albugo candida* (1.0%), *Phytophthora infestans* (1.57%); and gymnosperms, *Pinus pinaster* (2.76%) while extremely increased SSR frequency was observed in *Volvox carteri* (20.77%), *Chlorokybus atmophyticus* (20.02%), and *Chlorella variabilis* (17.24%) among non-angiospermic species [Figure 6]. Among angiosperms, *Pisum sativum* (3.48%), *Cajanus cajan* (3.55%), and *Daucus carota* (4.24%) showed decreased SSR frequency distribution from average while, *Oryza sativa*, *Trifolium pratense*, *Ricinus communis*, *Cucumis melo*, and *Raphanus sativus* significantly deviated from the average value with an extremely increased SSR frequency of 24.81, 20.20, 19.19, 17.90, and 17.23, respectively [Figure 7]. Our observation of ascended SSR frequency is in accordance with the earlier reports of comparative genomic analysis by various workers [46,48,49,53-56].

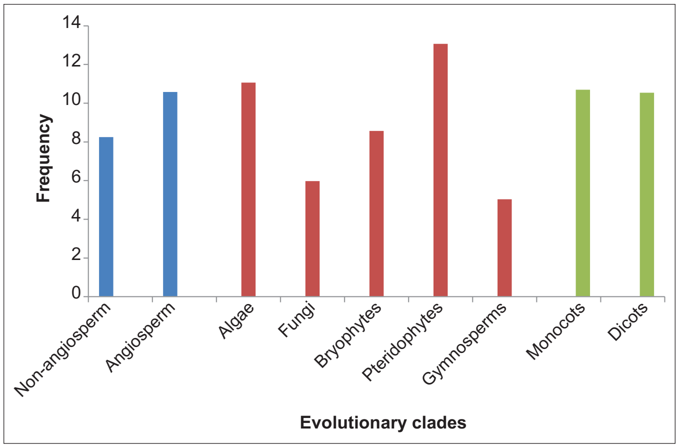


Figure 5: Comparative details of EST-SSRs frequency (%) among different phylogenetic clades.

3.4. Frequency Distribution of Different Type of SSRs in ESTs

A comparison in the distribution of different types of SSRs was analyzed within ESTs of selected species belonging to different clades. Overall, the occurrence of mononucleotide repeats was found to be with 80.95% frequency distribution while 19.05% frequency distribution belonged to other types of SSRs (di to hexa nucleotide SSRs). Mononucleotide SSRs were observed to be highly repetitive with uniform distribution and few fluctuations. It has been seen that mononucleotide SSRs might be responsible to play a vital role in filling the gaps in linkage maps and their applications have been successfully established in some populations [47]. Among mononucleotide SSRs distribution, the non-angiosperms (70.67%) showed increased incidence compared to angiosperms (67.67%). Usually among non-angiosperms, algae (53.91%), bryophytes (67.84%), and pteridophytes (69.48%) displayed reduced mononucleotide SSRs incidence while, gymnosperms (84.47%) and fungi (77.63%) showed significantly increased mononucleotide SSRs incidence, respectively. Similarly, for angiosperms, the increased mononucleotide SSRs incidence was observed in dicots (70.96%) as compared to monocots (64.37%).

Excluding mononucleotide SSRs, trinucleotide SSRs were found to be in major (51.28%) repetition, followed by dinucleotide SSRs (39.32%), hexa nucleotide SSRs (3.43%), tetra nucleotide SSRs (3.01%), and penta nucleotide SSRs (2.96%) in general analysis [Figure 8]. The increased trinucleotide SSRs incidence is in agreement with previous genomics studies done in various species [57-59] and relatively high accountability of our tri and hexa nucleotide repeats is also in accordance with previous reports [20,60]. Increased frequency of trinucleotide SSRs has also been reported in coding and noncoding genome of viruses, organelles, plasmids, prokaryotes, fungi, protists, and humans [61,62]. High recurrence of tri and hexa nucleotide SSRs has also been observed more than other types of SSRs in genomic and EST sequences [63,64].

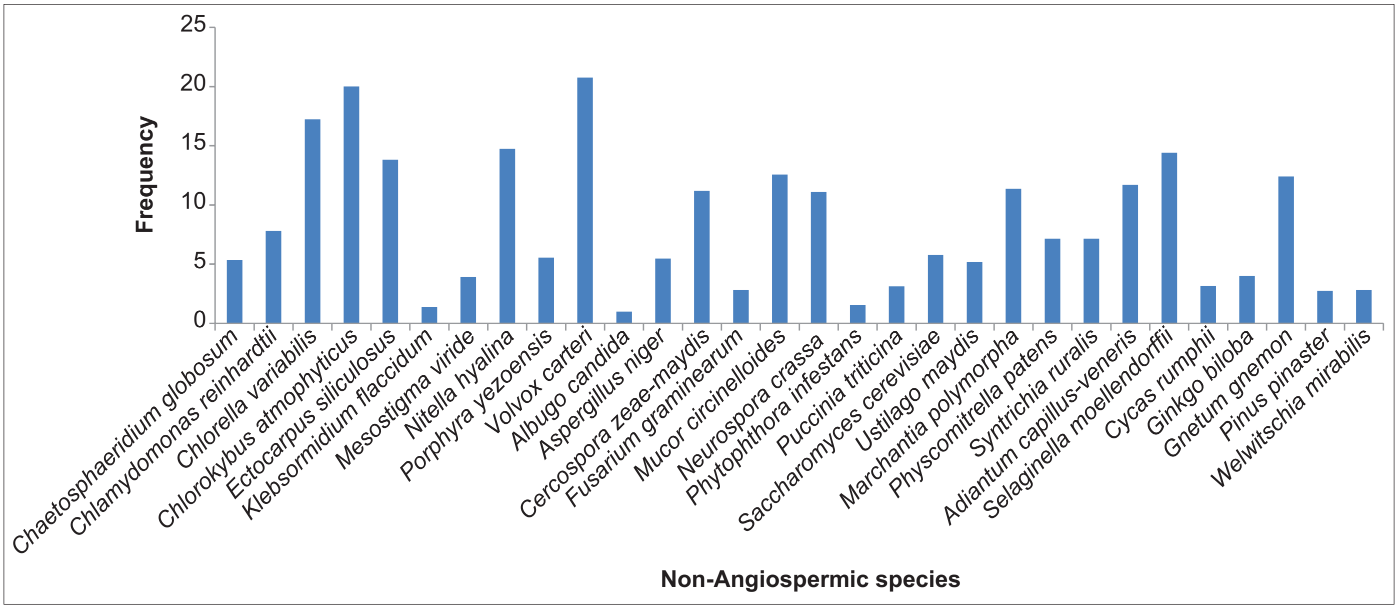


Figure 6: Percentage of SSR incidence within the species belonging to non-angiosperm.

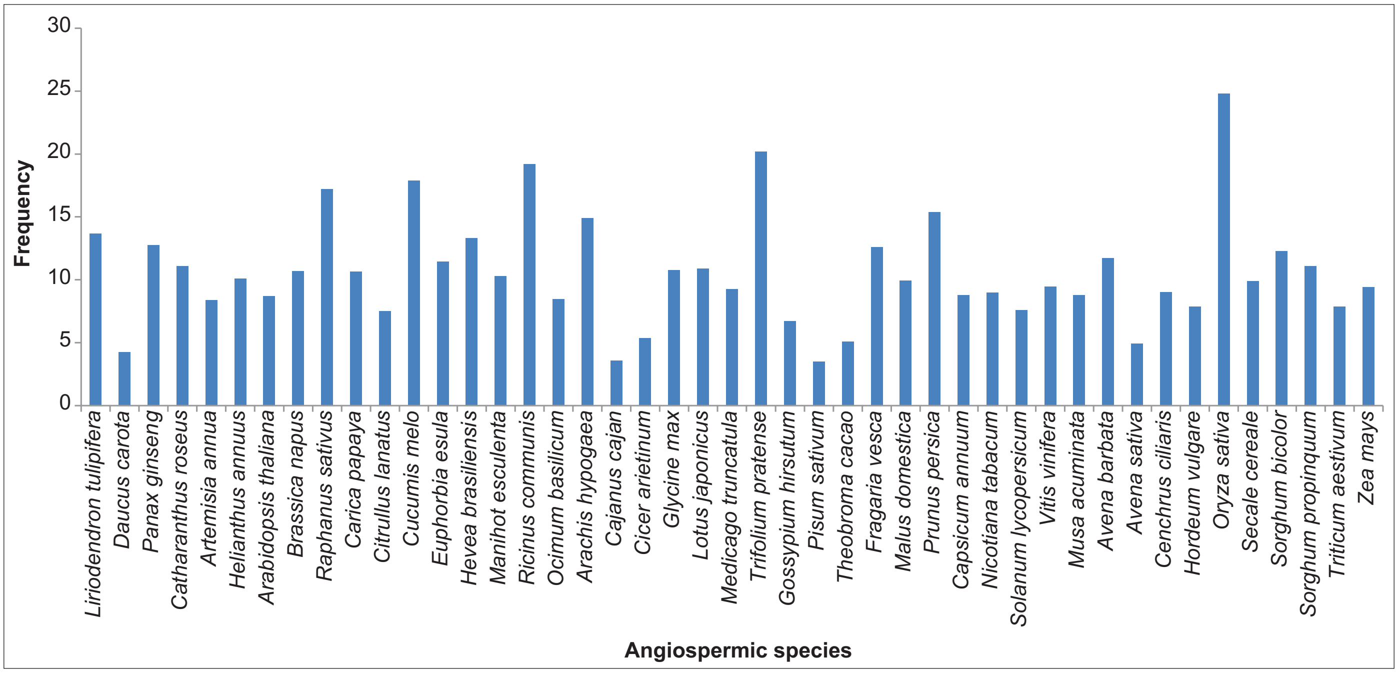


Figure 7: Percent of SSR incidence within the species belonging to angiosperm.

Significantly, the common pattern was observed for different types of SSR in both non-angiosperms and angiosperms but some fluctuations in tetra, penta, and hexa nucleotide SSRs were observed from the general trend among different evolutionary clades [Figure 8]. Significantly, distinguish species revealed a deviation from the average value of SSRs such as, *Adiantum capillus-veneris* (82.00%), *Daucus carota* (66.44%), and *Liriodendron tulipifera* (65.74%) showed deviation in dinucleotide SSR, *Chlorella variabilis* (91.45%), *Chlorokybus atmophyticus* (83.49%), and *Porphyra yezoensis* (79.17%) showed in tri nucleotide SSR, *Mesostigma viride* (38.92%) in tetra nucleotide SSR, *Mesostigma viride* (23.78%) in penta nucleotide SSR, and *Fusarium graminearum* (10.90%) in hexa nucleotide SSR [Additional file 4]. Earlier observations gave similar view of uneven distribution of average frequency among distinct plant species [48,55,56,58,65-67].

Moreover, mono, tri, and dinucleotide SSRs have shown increased distribution in comparison to hexa, tetra, and penta nucleotide SSRs, respectively. However, the existence of different types of SSRs and their complete molecular mechanism, distribution and dominant behavior of SSRs are unstated but it may have possibly risen from selection pressures applied on that specific motif during evolution in the plant genome. While, the replication slippage mechanism is also very important factor that affects a process involving addition or removal of one or more motif repeats and nucleotide substitutions, or duplication events, besides that unequal crossing over have been also seen to influence microsatellite variations [68-70].

3.5. SSRs Motif Length and Categorization

The motif length in different types of SSRs was examined in the ESTs of selected plant species. In total, the average SSRs motif length was found to be 21.12 bp long which is slightly deviated from the earlier reports [67,71]. In general, hexa nucleotide motif (26.60bp) showed high average motif length, followed by tetra nucleotide (22.30 bp), penta nucleotide (22.26 bp), dinucleotide (19.12 bp), mononucleotide (18.51 bp), and trinucleotide motif (17.94 bp). This trend of motif length was found to be common in both non-angiosperms and angiosperm but few deviations were seen among non-angiospermic clades. The motif length strengthening or shortening within particular types of SSRs have an influential role on biological complexity which can be correlated with genetic evolution and regulation of evolutionary mechanism while their existence in protein-coding regions can be involved in gain or loss of gene function [69,72-74]. The uniformity in the basic style of SSRs motif length was observed in angiosperms compared to non-angiosperms which represented some divergence [Figure 9]. However, some skewness in motif length was also observed among different evolutionary groups and species [Additional file 5].

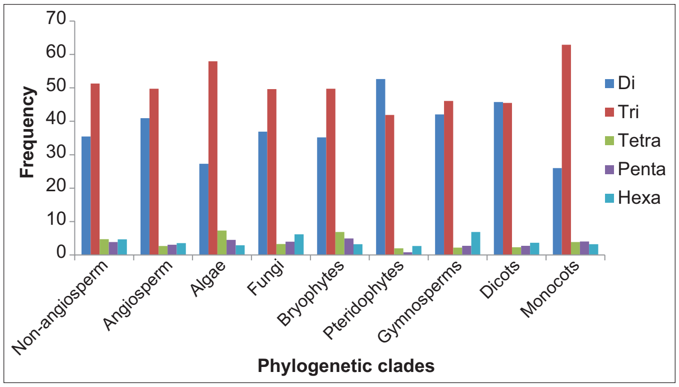


Figure 8: Comparative distribution of different types of SSRs comprising Di to Hexa nucleotide repeats among different phylogenetic clades.

On the basis of motif length, microsatellites or SSRs can be categorized into class I and class II perfect microsatellites. A total of 26.43% SSRs were recognized as class I (≥20bp) type perfect microsatellites and rest (73.56%) were belong to class II (12-20bp) type perfect microsatellites, excluding both mono and compound SSRs which is in compliance with earlier report [75]. The class II type of microsatellites was found to be widespread than class I which is in consensus with previous observations [76,77]. Microsatellites which acquire the length between 20 nucleotides or 12 and 19 nucleotides are reported to be highly mutable [74,78]. The class II type of microsatellites was observed to be more prevalent in angiosperms as compared to non-angiosperms. While, class II type SSRs revealed more regularity in monocots, dicots, and fungi but class I SSRs were widespread in pteridophytes, gymnosperms, and bryophytes [Figure 10]. Consequently, class II type SSR was found to be more frequent than class I types SSR within selected species whether belonging to any evolutionary clade [Additional file 6 and 7].

3.6. Annotation of Most Frequent SSRs Motifs

The enormous diversity in the SSR motifs was obtained within mono to hexa nucleotide SSRs. For example, two motifs (A/T and G/C) with complementarity were identified in mononucleotide SSR followed by four motifs (AC/GT, AG/CT, AT/AT, and CG/CG) in dinucleotide SSR and ten motifs (AAC/GTT, AAG/CTT, AAT/ATT, ACC/GGT, ACG/CGT, ACT/AGT, AGC/CTG, AGG/CCT, ATC/ATG, and CCG/CGG) in trinucleotide SSR. While, the complexed or unfashionable motifs pattern were found onward from tetra to hexa nucleotide SSRs and this nature might be explained by more combinations and permutations of four bases of nucleotides within the motifs. For mononucleotide SSRs, motif A/T was found to be dominant over G/C motif and this rise of A/T motif pattern was almost widespread within and across all species. In general, non-angiosperms represented more A/T motif circulation than angiosperms. Among evolutionary clades, highest A/T incidence was observed in gymnosperms (94.12%) followed by dicots (88.82%) but relatively lower occurrence was seen in algae (68.79%) and monocots (70.17%) with high G/C motif incidence inversely [Figure 11]. The presence of mononucleotide repeats along with their base composition (A/T and G/C) is known to have vital impact on stability of gene and gene functions due to their highly capricious nature which might be responsible for the frameshift mutation in the coding region [79]. The distribution of mononucleotide motifs was noted to be irregular within number of species, for instance A/T motif was found to be more frequent in *Triticum aestivum* (99.88%), followed by *Saccharomyces cerevisiae* (99.82%), *Pisum sativum* (99.76%), and *Raphanus sativus* (99.56%). Similarly, the G/C motif found to be more circulated with 48.51%, 44.43%, 43.52%, 37.18%, and 37.18% in *Ectocarpus siliculosus*, *Volvox carteri*, *Porphyra yezoensis*, *Ustilago maydis*, and *Oryza sativa,* respectively [Additional file 8].

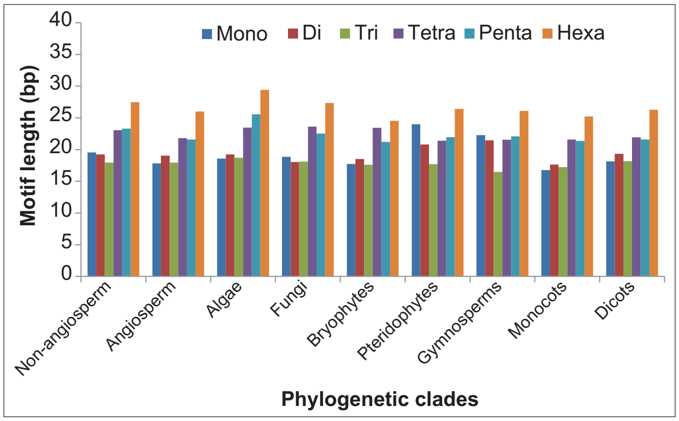


Figure 9: Comparative details of SSR motif length distributions among evolutionary clades.

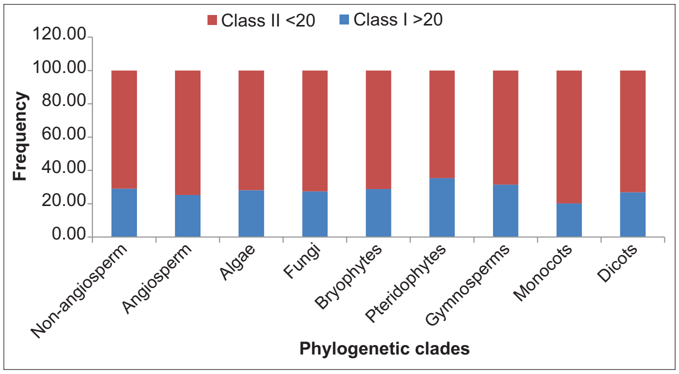


Figure 10: Comparative distribution of Class I and Class II perfect microsatellites among distinct phylogenetic clades.

Furthermore, the skewness was observed in the frequency distribution of dinucleotide SSR motifs among species. Commonly, motif AG/CT was identified in major circulations (56.03%) followed by AC/GT (21.22%) and AT/AT (19.41%) but motif CG/CG was in least repetition (3.33%). These patterns of motifs distribution were uniform in phylogenetic clades except algae in which, motif AC/GT was frequent over AG/CT motif and motif CG/CG was dominant over AT/AT motif but motif AT/AT was dominant over AC/GT in gymnosperm. The most frequent AG/CT motif in present study is in compliance with earlier reports followed by either AC/GT or AT/AT and least reported was CG/CG motif in various comparative genomic analysis [52,70]. Accordingly, the abundance of homopurine-homopyrimidine stretches may be explained due to their more commonness in transcribe region and their useful role in the DNA structures modification, regulation of gene expression, and methylation of CpG [69]. Remarkable divergence was seen to emerge from the average value of dimer motifs. For example, motifs AC/GT (45.68%) and CG/CG (15.09%) were found to be common in algae followed by motif AG/CT (65.15%) in bryophytes then motifs AG/CT (63.17%) and AT/AT (5.83%) in pteridophytes while motif AT/AT (36.82%) was in gymnosperms. Similarly, motifs CG/CG (5.81%) and AG/CT (65.66%) were identified to be more reiterated in monocot and dicots, respectively [Figure 11]. Some extreme deviation in the frequency of dimer motifs was also seen in some species, namely, motif AC/GT was frequent in *Volvox carteri* (83.62%) and *Chlamydomonas reinhardtii* (72.97%), followed by AG/CT motif found to be widespread in *Marchantia polymorpha* (89.62%), *Fragaria vesca* (85.85%), and *Malus domestica* (82.28%). Further, motif AT/AT was common in *Saccharomyces cerevisiae* (87.03%) and *Albugo candida* (72.06%) then motif CG/CG was also frequent in *Klebsormidium flaccidum* (59.79%), *Porphyra yezoensis* (30.61%), and *Mesostigma viride* (21.43%) [Additional file 8].

For trinucleotide SSR motifs, ten distinct motifs were identified in the ESTs of selected plant species. Overall, motif AAG/CTT found to be most dominant, followed by AGC/CTG, CCG/CGG, AGG/CCT, ATC/ATG, ACC/GGT, AAC/GTT, AAT/ATT, ACG/CGT, and ACT/AGT, respectively. Motif AAG/CTT appeared to be widespread among non-angiosperms and angiosperms. Among the non-angiosperm clades, some trinucleotide SSRs motifs showed more repetition such as, motif AGC/CTG was consistently more common within fungi, bryophytes, pteridophytes, and gymnosperms with frequency distribution of 23.91%, 38.51%, 39.69%, and 31.63%, respectively. Motif AAC/GTT (14.09%) was also common in fungi and motifs AAG/CTT (22.04%) and CCG/CGG (12.39%) were common in gymnosperms. Moreover, few motifs seemed to be common in different evolutionary clades such as, motif CCG/CGG in algae, AAC/GTT in fungi, AGG/CCT in bryophytes, ACC/GGT in pteridophytes, and ATC/ATG in gymnosperms [Figure 11]. The commonness of tri nucleotide motifs in the present study is in the wake of accordance with earlier studies [20,46,49,66].

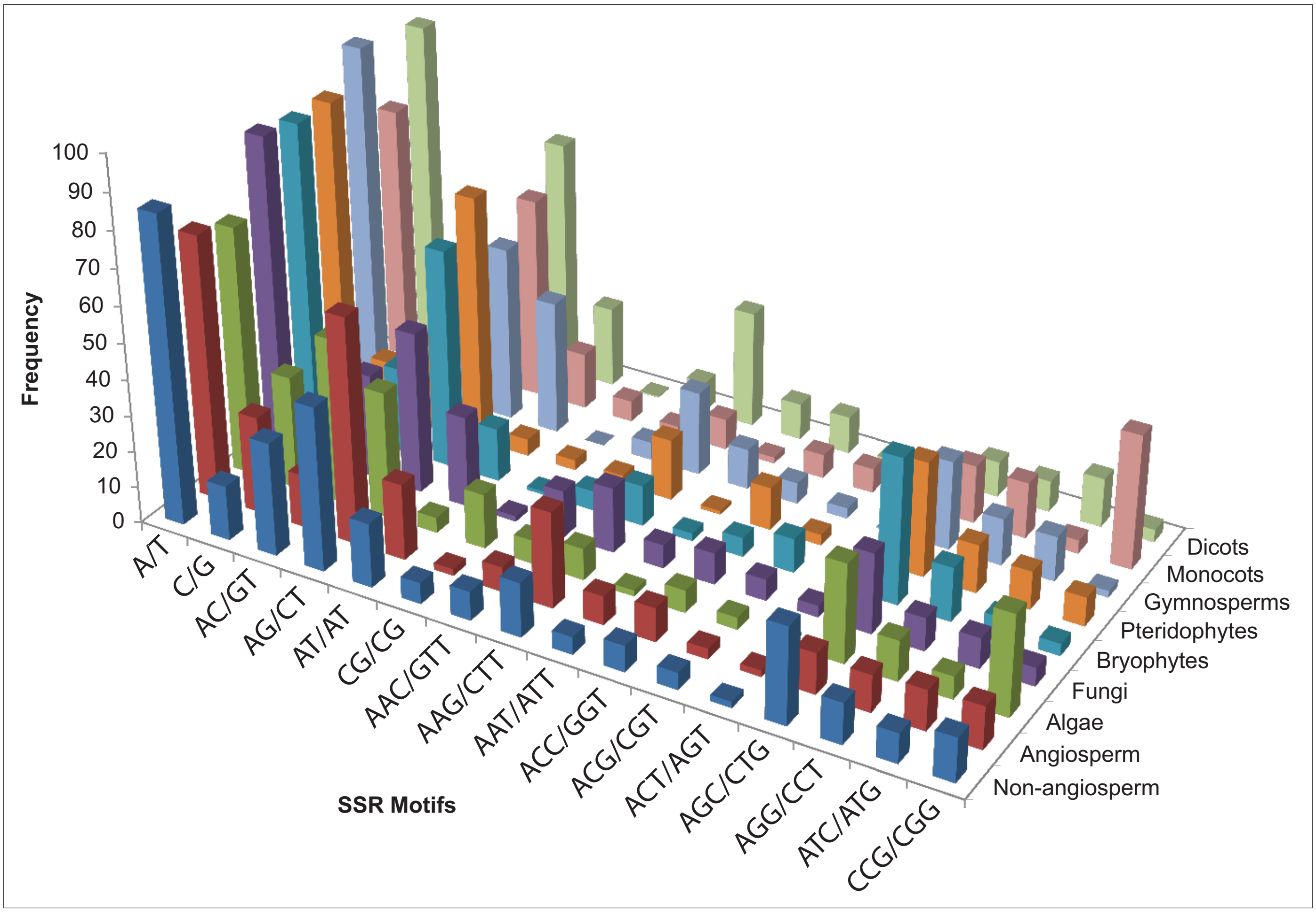


Figure 11: Comparative analysis of different SSR motif distributions amongst mono, di and tri nucleotide repeat motifs amongst phylogenetic clades. Motifs, A/T, AG/CT, AC/GT, AAG CCG/CGG, and AGC/CTG were showed more repetitions.

Among dicots, the trinucleotide motifs such as AAG/CTT, ATC/ATG, and ACC/GGT were identified in more repetition but ACG/CGT, ACT/AGT, and CCG/CGG motifs were seen in least circulation. Analysis revealed that motif AAG/CTT was found to be most dominant in *Cucumis melo* and *Citrullus lanatus* with 61.71% and 40.85%, respectively. This motif also revealed more repetition in few species such as, *Carica papaya*, *Arabidopsis thaliana*, *Nicotiana tabacum*, *Euphorbia esula*, and *Arachis hypogaea* and this repetition is in accordance with various earlier studies [57,80-84]. Individually, few motifs also seemed to be highly duplicated among various species such as, motif ATC/ATG was commonly rich in *Daucus carota*, *Artemisia annua*, and *Gossypium hirsutum* followed by motif ACC/GGT which appeared to be widespread in *Trifolium pratense*, *Helianthus annuus*, and *Lotus japonicas*. While motif AAC/GTT was highly repeated in *Pisum sativum*, *Artemisia annua*, and *Capsicum annuum*, motif AAT/ATT was common in *Cajanus cajan*, *Cicer arietinum,* and *Hevea brasiliensis* [Additional file 9]. All of these common tri nucleotide motifs which appeared in the present study have been reported in various dicot plant species [47,85-89].

Among monocots, the trinucleotide motif like CCG/CGG was more prevalent and this motif incidence was uniformly followed by AGG/CCT, AGC/CTG, ACG/CGT, and AAG/CTT motifs, respectively [Figure 11]. Significantly, motif CCG/CGG found to be widespread among species of Poaceae family wherein, *Cenchrus ciliaris*, *Oryza sativa*, *Zea mays,* and *Sorghum propinquum* showed highly repeated nature of this motif except *Musa acuminate*. The predominance of CCG/CGG motif in the present study is in agreement with previous observations in various plant species [26,27,54,58,90]. In the present study, increased repetition of CCG/CGG motif was observed as unique feature for algae and monocots species and this rise of CCG/CGG distribution could be related to increase of GC-content [18,48, 91]. Further, motif AGC/CTG and AGG/CCT were also evenly distributed in grass family except *Oryza sativa* and *Zea mays*. The dominancy of different motifs was also detected over average value in certain species, namely, motif AAG/CTT was widespread in *Musa acuminate* and *Avena sativa* then motif ACG/CGT was common in *Sorghum bicolor* and *Secale cereale*. Some motifs, AGC/CTG and AGG/CCT were found to be frequent in monocot species such as *Avena barbata*, *Avena sativa*, *Hordeum vulgare,* and *Triticum aestivum* [Additional file 9]. Distinctive more repeated type of trimer motifs were also observed in the present study which are in resemblance with earlier studies reported in some monocot species [27,46,48,49,53,54,91-93].

At present, the asymmetrical incidence of trinucleotide motifs was observed in monocots and dicots and their distribution was found to be almost inversely proportional to the each other. For example, motif CCG/CGG revealed dominancy in monocots compared to dicots whereas in dicot, motifs AAG/CTT seemed to be highly repeated than monocots. However, the common motif AGC/CTG found to be least distribution in both monocots and dicots. In addition, some motifs namely; CCT/AGG, CCG/GGC, GGA/TTC, and GAA/TTC were also identified which are responsible for making unusual DNA folding structures including hairpin form, bipartite triplex form, and simple loop folding. These motifs may also be responsible for having an impact on gene expression and their regulation mechanism. Moreover, the presence of trinucleotide repeats in the coding region encodes distinct type of amino acid tracts within the peptide or protein which might play an important role in various metabolic activities [48-50,94].

In addition, trinucleotide SSRs motifs are known to have influential role at proteome level because they have direct relation with exons level and can generate amino acids stretch in protein. Therefore, various types of predicted amino acids are identified in the first frame translation for different types of tri nucleotide SSRs motifs. In general, serine (Ser), arginine (Arg), leucine (Leu), alanine (Ala), and proline (Pro) amino acids appeared in huge account in the present analysis. For non-angiosperms, Ala found to be more frequent followed by Ser, Gln, and Leu, whereas, Arg, Ser, Ala, and Leu showed more distribution in angiosperms [Figure 12]. Among non-angiospermic clades, frequent distribution of few amino acids was observed such as, Ala was seen commonly in algae and pteridophytes with frequency 19.08% and 15.34% respectively, followed by Leu (11.73%) in fungi then Ser was more widespread in both bryophytes (14.83%) and gymnosperms (13.95%). Among angiosperms, increased level of Ala and Arg was identified in monocots whereas Ser and Leu were commonly identified in dicots [Figure 13]. This finding is in accordance with earlier genomic studies reported in different species [20,46,49,66]. It is obvious that long stretch of amino acid is responsible for increasing protein size which can create a transition in protein activity. Certain types of single amino acid repetitions have potential to regulate transcriptional activities and contribute in protein-protein interactions. These kinds of amino acids distribution at protein level are involved in the various molecular activities such as ubiquitin activity, structural activity, and receptor activity. While, single amino acid stretch may also provide assistance as spacer elements and also help in distinguishing protein domains [95]. Furthermore, numbers of amino acids were observed majorly within different species, namely, Ala was found to be frequent in *Chlorella variabilis*, *Ectocarpus siliculosus*, *Chlorokybus atmophyticus*, *Neurospora crassa*, *Marchantia polymorpha*, and *Selaginella moellendorffii*. Then, Ser frequently was identified in *Gnetum gnemon* and *Arachis hypogaea* and Arg was familiar in *Oryza sativa*. It was also observed that some amino acids were in moderate amount but amino acids also such as methionine (Met), tryptophan (Trp), and tyrosine (Tyr) were shown their repetitions in very diminutive amount. The stop codons such as Amber (Am*), Ochre (Oc*), and Opal (Op*) were also detected but among them, Op* was more frequently distributed than Oc and Am. Moreover, dicots, monocots, and algal species showed high frequency of Op* codon in comparison to Am* and Oc*. While, high frequency of the Op stop codon was also seen in *Nitella hyaline*, *Brassica napus,* and *Raphanus sativus* with 7.44%, 3.08%, and 2.83% distribution separately [Additional file 10].

Due to combination and permutation of nucleotides in SSRs motif, an immense diversity was observed in the SSR motifs belonging to tetra, penta, and hexa nucleotide SSRs with lack of relation which was identified in the frequency of motifs and type of motifs within and across species. Therefore, the complexed incidence of different types of motifs was observed in the present study and their distributions were immense. For tetranucleotide SSRs, few numbers of specific SSR motifs were observed comparatively within species, namely, motifs AATC/ATTG, ACAT/ATGT, and AATT/AATT were more duplicated in *Nitella hyaline*, *Volvox carteri*, and *Mesostigma viride,* respectively. Further, motif AGGC/CCTG was found to be highly repeated in *Neurospora crassa*, followed by motif AGGC/CCTG in *Marchantia polymorpha* and motif AGCG/CGCT in *Selaginella moellendorffii*. In monocots, motif ATCC/ATGG was found to be highly repeated in *Oryza sativa* while in dicot, motif AAAT/ATTT was widespread in *Artemisia annua* and *Prunus persica*. Furthermore, motif AAAG/CTTT was more frequent in *Arachis hypogaea*, *Cucumis melo*, *Ricinus communis,* and *Theobroma cacao*. The prevalence of these types of tetramer motifs is in concurrence with earlier observations reported in various species [47,50,52,58,65].

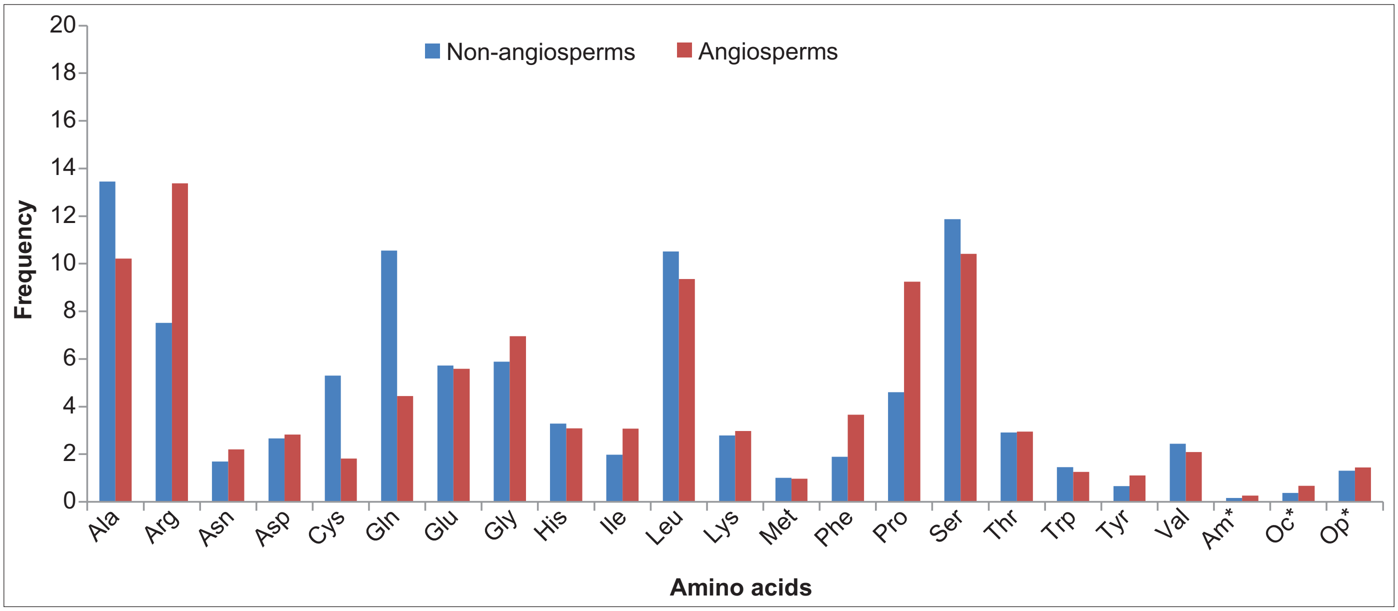


Figure 12: Relative amino acids distribution between non-angiosperms and angiosperms. In general, amino acids, namely, alanine (Ala), arginine (Arg), leucine (Leu), serine (Ser), and proline (Pro) were found to be widespread.

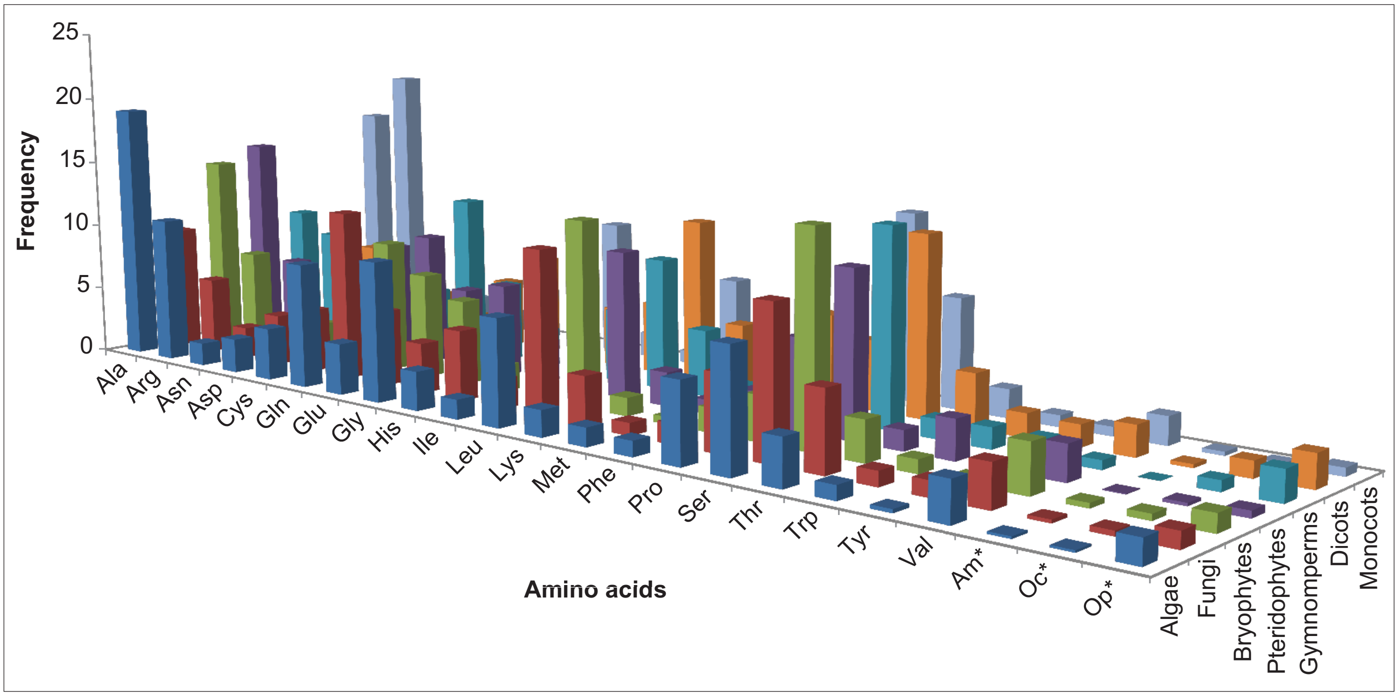


Figure 13: Comparative distribution of predicted amino acids encoded by trinucleotide repeat motifs amongst different evolutionary clades..

Similarly, a complexed trend was identified in pentanucleotide SSRs but few motifs seemed to more common than other within the species such as, motifs AAATT/AATTT, AGCCT/AGGCT, AAAAT/ATTTT, and AGAGG/CCTCT were found to be more frequent in non-angiospermic species especially in *Mesostigma viride*, *Neurospora crassa,* and *Physcomitrella patens,* respectively. In monocot, motifs AGAGG/CCTCT, AAGAG/CTCTT, and AGGGG/CCCCT were common in *Oryza sativa* followed by motifs AGAGG/CCTCT and AGGGG/CCCCT in *Hordeum vulgare* and motifs AGCTC/AGCTG and AGAGG/CCTCT were in *Zea mays.* In dicot species, the reiteration of motif like AAAAG/CTTTT was found to be common in *Manihot esculenta*, *Theobroma cacao*, *Cucumis melo,* and *Arachis hypogaea*. Motif AAAAT/ATTTT was more common among *Artemisia annua*, *Prunus persica*, and *Hevea brasiliensis* and this observation is in agreement with previous studies among different plant species [65,82]. Significantly, the hexanucleotide SSRs seemed to be more dominant over tetramer and penta nucleotide SSRs which is in compliance with earlier analysis in various plant species [57,96]. Surprisingly, massive diversity was identified in hexanucleotide SSRs motif patterns and limitless array of different types of motifs was seen with diminutive repetition. Besides, few hexa nucleotide motifs showed comparatively enhanced repetitions in distinct plant species, namely, motif ATCGCC/ATGGCG was found to be common in *Nitella hyaline* and *Selaginella moellendorffii* followed by motif ACAGAT/ATCTGT in *Neurospora crassa*. Motifs AGGCGG/CCGCCT, AGCCTG/AGGCTC, and AACCCT/AGGGTT observed in *Oryza sativa, Gossypium hirsutum,* and *Artemisia annua,* respectively, are in compliance with previous reports in different species [48,66,97,98].

4. CONCLUSION

The present study aimed to explore the plasticity of tandem repeated DNA elements, especially SSRs analysis in expressed sequence tags (ESTs). In general, mononucleotide to hexa nucleotide SSRs were annotated at large scale ESTs of 75 different species belonging to diverge evolutionary clades such as algae, fungi, bryophytes, pteridophytes, gymnosperms, dicots, and monocots. Approximately, 4.35 million EST sequences were examined for SSRs exploration which resulted in identification of huge diversity in SSRs distributions in ESTs of selected species. Mononucleotide SSRs were identified as utmost in circulation in the ESTs uniformly followed by trinucleotides, dinucleotides, hexanucleotides, tetra nucleotides, and penta nucleotides SSR, respectively. An immense diversity in the SSR frequencies and their motifs distribution were identified within and across the species belonging to angiosperms and non-angiosperms. According to SSR motifs incidence, mononucleotide to trinucleotide SSR motifs showed remarkable distribution in the ESTs and their categorization was found to be explicit. Conversely, more complex pattern of motifs distribution was identified within hexanucleotide SSRs and pentanucleotide SSRs in comparison to tetranucleotide SSRs which showed slightly less diversity in motifs relatively. Therefore, a number of distinctive attributes were revealed which enhanced our understanding about the SSRs variation, distribution, expansion, and divergence within and across angiospermic and non-angiospermic species or different evolutionary clades.

5. CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest regarding the publication of this paper.

6. ACKNOWLEDGMENT

Authors are thankful to Council of Scientific and Industrial Research (CSIR) for the fellowship (CSIR-RA). Authors are grateful to DBT-Bioinformatics Infrastructure Facility, UGC-UPE and Department of Botany, University of Rajasthan for providing necessary facilities.

7. AUTHOR CONTRIBUTIONS

All authors made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; took part in drafting the article or revising it critically for important intellectual content; agreed to submit to the current journal; gave final approval of the version to be published; and agree to be accountable for all aspects of the work. All the authors are eligible to be an author as per the international committee of medical journal editors (ICMJE) requirements/guidelines.

8. ETHICAL APPROVALS

This study does not involve experiments on animals or human subjects.

9. PUBLISHER'S NOTE

This journal remains neutral with regard to jurisdictional claims in published institutional affiliation.

REFERENCES

1. Kubis S, Schmidt T, Heslop-Harrison JS. Repetitive DNA elements as a major component of plant genomes. Ann Bot 1998;82:45-55.
2. Shapiro JA, Von Sternberg R. Why repetitive DNA is essential to genome function. Biol Rev 2005;80:227-50.
3. Biscotti MA, Olmo E, Heslop-Harrison JP. Repetitive DNA in Eukaryotic Genomes. Berlin, Germany: Springer; 2015.
4. Bouck A, Vision T. The molecular ecologist's guide to expressed sequence tags. Mol Ecol 2007;16:907-24.
5. Edwards NJ. Novel peptide identification from tandem mass spectra using ESTs and sequence database compression. Mol Syst Biol 2007;3:102.
6. Parkinson J, Blaxter M. Expressed sequence tags: An overview. In: Expressed Sequence Tags (ESTs). Berlin, Germany: Springer; 2009. p. 1-12.
7. Nagaraj SH, Gasser RB, Ranganathan S. A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief Bioinform 2007;8:6-21.
8. Ewing RM, Kahla AB, Poirot O, Lopez F, Audic S, Claverie JM. Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. Genome Res 1999;9:950-9.
9. Prabu G, Mandal A. Computational identification of miRNAs and their target genes from expressed sequence tags of tea (*Camellia sinensis*). Genom Proteom Bioinform 2010;8:113-21.
10. Zhang Y, Zhu X, Chen X, Song C, Zou Z, Wang Y, *et al*. Identification and characterization of cold-responsive microRNAs in tea plant (*Camellia sinensis*) and their targets using high-throughput sequencing and degradome analysis. BMC Plant Biol 2014;14:271.
11. Alba R, Payton P, Fei Z, McQuinn R, Debbie P, Martin GB, *et al*. Transcriptome and selected metabolite analyses reveal multiple points of ethylene control during tomato fruit development. The Plant Cell 2005;17:2954-65.
12. Cui G, Huang L, Tang X, Zhao J. Candidate genes involved in tanshinone biosynthesis in hairy roots of *Salvia miltiorrhiza* revealed by cDNA microarray. Mol Biol Rep 2011;38:2471-8.
13. Zhou GF, Liu YZ, Sheng O, Wei QJ, Yang CQ, Peng SA. Transcription profiles of boron-deficiency-responsive genes in citrus rootstock root by suppression subtractive hybridization and cDNA microarray. Front Plant Sci 2015;5:795.
14. Baxevanis AD, Ouellette BF. Bioinformatics: A Practical Guide to the Analysis of Genes and Proteins. Vol. 43. United States: John Wiley & Sons; 2004.
15. Hampton M, Xu WW, Kram BW, Chambers EM, Ehrnriter JS, Gralewski JH, *et al*. Identification of differential gene expression in *Brassica rapa* nectaries through expressed sequence tag analysis. PLoS One 2010;5:e8782.
16. Sui S, Luo J, Ma J, Zhu Q, Lei X, Li M. Generation and analysis of expressed sequence tags from *Chimonanthus praecox* (Wintersweet) flowers for discovering stress-responsive and floral development-related genes. Comp Funct Genom 2012;2012:134596.
17. Sasaki K, Mitsuda N, Nashima K, Kishimoto K, Katayose Y, Kanamori H, *et al*. Generation of expressed sequence tags for discovery of genes responsible for floral traits of *Chrysanthemum morifolium* by next-generation sequencing technology. BMC Genom 2017;18:683.
18. Morgante M, Olivieri A. PCR-amplified microsatellites as markers in plant genetics. The Plant J 1993;3:175-82.
19. Jurka J, Pethiyagoda C. Simple repetitive DNA sequences from primates: Compilation and analysis. J Mol Evol 1995;40:120-6.
20. Tóth G, Gáspári Z, Jurka J. Microsatellites in different eukaryotic genomes: Survey and analysis. Genome Res 2000;10:967-81.
21. Ellegren H. Microsatellites: Simple sequences with complex evolution. Nat Rev Genet 2004;5:435.
22. Agarwal M, Shrivastava N, Padh H. Advances in molecular marker techniques and their applications in plant sciences. Plant Cell Rep 2008;27:617-31.
23. Masouleh AK, Waters DL, Reinke RF, Henry RJ. A high-throughput assay for rapid and simultaneous analysis of perfect markers for important quality and agronomic traits in rice using multiplexed MALDI-TOF mass spectrometry. Plant Biotechnol J 2009;7:355-63.
24. Oliveira EJ, Pádua JG, Zucchi MI, Vencovsky R, Vieira ML. Origin, evolution and genome distribution of microsatellites. Genet Mol Biol 2006;29:294-307.
25. Heywood VH, Iriondo JM. Plant conservation: Old problems, new perspectives. Biol Conserv 2003;113:321-35.
26. Cordeiro GM, Casu R, McIntyre CL, Manners JM, Henry RJ. Microsatellite markers from sugarcane (*Saccharum* spp.) ESTs cross transferable to erianthus and sorghum. Plant Sci 2001;160:1115-23.
27. Kantety RV, La Rota M, Matthews DE, Sorrells ME. Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat. Plant Mol Biol 2002;48:501-10.
28. Eujayl I, Sledge M, Wang L, May G, Chekhovskiy K, Zwonitzer J, *et al*. Medicago truncatula EST-SSRs reveal cross-species genetic markers for *Medicago* spp. Theor Appl Genet 2004;108:414-22.
29. Varshney RK, Chabane K, Hendre PS, Aggarwal RK, Graner A. Comparative assessment of EST-SSR, EST-SNP and AFLP markers for evaluation of genetic diversity and conservation of genetic resources using wild, cultivated and elite barleys. Plant Sci 2007;173:638-49.
30. Simko I. Development of EST-SSR markers for the study of population structure in lettuce (*Lactuca sativa* L.). J Hered 2009;100:256-62.
31. Fu N, Wang PY, Liu XD, Shen HL. Use of EST-SSR markers for evaluating genetic diversity and fingerprinting Celery (*Apium graveolens* L.) cultivars. Molecules 2014;19:1939-55.
32. Ukoskit K, Posudsavang G, Pongsiripat N, Chatwachirawong P, Klomsa-Ard P, Poomipant P, *et al*. Detection and validation of EST-SSR markers associated with sugar-related traits in sugarcane using linkage and association mapping. Genomics 2018;111:1-9.
33. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res 1999;9:868-77.
34. Liu M, Shi J, Lu C. Identification of stress-responsive genes in *Ammopiptanthus mongolicus* using ESTs generated from cold-and drought-stressed seedlings. BMC Plant Biol 2013;13:88.
35. Silva CC, Mantello CC, Campos T, Souza LM, Gonçalves PS, Souza AP. Leaf-, panel-and latex-expressed sequenced tags from the rubber tree (*Hevea brasiliensis*) under cold-stressed and suboptimal growing conditions: The development of gene-targeted functional markers for stress response. Mol Breed 2014;34:1035-53.
36. Ronning CM, Stegalkina SS, Ascenzi RA, Bougri O, Hart AL, Utterbach TR, *et al*. Comparative analyses of potato expressed sequence tag libraries. Plant Physiol 2003;131:419-29.
37. Garg R, Patel RK, Tyagi AK, Jain M. *De novo* assembly of chickpea

transcriptome using short reads for gene discovery and marker identification. DNA Res 2011;18:53-63.

38. Šmarda P, Bureš P, Horová L. The Evolution of Base Composition in Monocots. Brno: Muni Press; 2010.

39. Vinogradov AE. DNA helix: The importance of being GC-rich. Nucleic Acids Res 2003;31:1838-44.

40. Li XQ, Du D. Variation, evolution, and correlation analysis of C+ G content and genome or chromosome size in different kingdoms and phyla. PLoS One 2014;9:e88339.

41. Mouchiroud D, D'Onofrio G, Aïssani B, Macaya G, Gautier C, Bernardi G. The distribution of genes in the human genome. Gene 1991;100:181-7.

42. Amit M, Donyo M, Hollander D, Goren A, Kim E, Gelfman S, *et al*. Differential GC content between exons and introns establishes distinct strategies of splice-site recognition. Cell Rep 2012;1:543-56.

43. Costantini M, Bernardi G. Replication timing, chromosomal bands, and isochores. Proc Natl Acad Sci 2008;105:3433-7.

44. Duret L, Arndt PF. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet 2008;4:e1000071.

45. Šmarda P, Bureš P. The variation of base composition in plant genomes. In: Plant Genome Diversity. Vol. 1. Berlin, Germany: Springer; 2012. p. 209-35.

46. Varshney RK, Thiel T, Stein N, Langridge P, Graner A. *In silico* analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol Lett 2002;7:537-46.

47. Kumpatla SP, Mukhopadhyay S. Mining and survey of simple sequence repeats in expressed sequence tags of dicotyledonous species. Genome 2005;48:985-98.

48. Victoria FC, da Maia LC, de Oliveira AC. *In silico* comparative analysis of SSR markers in plants. BMC Plant Biol 2011;11:15.

49. Haq SU, Kumar P, Singh R, Verma KS, Bhatt R, Sharma M, *et al*. Assessment of functional EST-SSR markers (Sugarcane) in cross-species transferability, genetic diversity among poaceae plants, and bulk segregation analysis. Genet Res Int 2016;2016:16.

50. Singh RB, Singh B, Singh RK. Development of potential dbEST-derived microsatellite markers for genetic evaluation of sugarcane and related cereal grasses. Ind Crops Prod 2019;128:38-47.

51. Kejnovsky E, Leitch IJ, Leitch AR. Contrasting evolutionary dynamics between angiosperm and mammalian genomes. Trends Ecol Evol 2009;24:572-82.

52. Sonah H, Deshmukh RK, Sharma A, Singh VP, Gupta DK, Gacche RN, *et al*. Genome-wide distribution and organization of microsatellites in plants: An insight into marker development in *Brachypodium*. PLoS One 2011;6:e21298.

53. Cardle L, Ramsay L, Milbourne D, Macaulay M, Marshall D, Waugh R. Computational and experimental characterization of physically clustered simple sequence repeats in plants. Genetics 2000;156:847-54.

54. Yu JK, La Rota M, Kantety R, Sorrells M. EST derived SSR markers for comparative mapping in wheat and rice. Mol Genet Genom 2004;271:742-51.

55. Cai K, Zhu L, Zhang K, Li L, Zhao Z, Zeng W, *et al*. Development and characterization of EST-SSR markers from RNA-Seq data in *Phyllostachys violascens*. Front Plant Sci 2019;10:50.

56. Sharma H, Kumar P, Singh A, Aggarwal K, Roy J, Sharma V, *et al*. Development of polymorphic EST-SSR markers and their applicability in genetic diversity evaluation in *Rhododendron arboreum*. Mol Biol Rep 2020;47:2447-57.

57. Lawson MJ, Zhang L. Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. Genome Biol 2006;7:R14.

58. Shi J, Huang S, Fu D, Yu J, Wang X, Hua W, *et al*. Evolutionary dynamics of microsatellite distribution in plants: Insight from the comparison of sequenced brassica, *Arabidopsis* and other angiosperm species. PLoS One 2013;8:e59988.

59. Haq S, Jain R, Sharma M, Kachhwaha S, Kothari S. Identification and characterization of microsatellites in expressed sequence tags and their cross transferability in different plants. Int J Genom 2014;2014:863948.

60. Metzgar D, Bytof J, Wills C. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Res 2000;10:72-80.

61. Field D, Wills C. Long, polymorphic microsatellites in simple organisms. Proc R Soc Lond B 1996;263:209-15.

62. Wren JD, Forgacs E, Fondon JW 3rd, Pertsemlidis A, Cheng SY, Gallardo T, *et al*. Repeat polymorphisms within gene regions: Phenotypic and evolutionary implications. Am J Hum Genet 2000;67:345-56.

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

64. Liu G, Xie Y, Zhang D, Chen H. Analysis of SSR loci and development of SSR primers in *Eucalyptus*. J Forestry Res 2018;29:273-82.

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

66. Maia LC, Souza VQ, Kopp MM, Carvalho FI, Oliveira AC. Tandem repeat distribution of gene transcripts in three plant families. Genet Mol Biol 2009;32:822-33.

67. Ranade SS, Lin YC, Zuccolo A, Van de Peer Y, García-Gil MR. Comparative *in silico* analysis of EST-SSRs in angiosperm and gymnosperm tree genera. BMC Plant Biol 2014;14:220.

68. Schlötterer C, Tautz D. Slippage synthesis of simple sequence DNA. Nucleic Acids Res 1992;20:211-5.

69. Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: Structure, function, and evolution. Mol Biol Evol 2004;21:991-1007.

70. Hosseinzadeh-Colagar A, Haghighatnia MJ, Amiri Z, Mohadjerani M, Tafrihi M. Microsatellite (SSR) amplification by PCR usually led to polymorphic bands: Evidence which shows replication slippage occurs in extend or nascent DNA strands. Mol Biol Res Commun 2016;5:167.

71. Tang S, Okashah RA, Cordonnier-Pratt MM, Pratt LH, Johnson VE, Taylor CA, *et al*. EST and EST-SSR marker resources for Iris. BMC Plant Biol 2009;9:72.

72. Kashi Y, King DG. Simple sequence repeats as advantageous mutators in evolution. Trends Genet 2006;22:253-9.

73. Sathishkumar R, Lakshmi P, Annamalai A, Arunachalam V. Mining of simple sequence repeats in the genome of gentianaceae. Pharmacogn Res 2011;3:19.

74. Vieira ML, Santini L, Diniz AL, Munhoz CF. Microsatellite markers: What they mean and why they are so useful. Genet Mol Biol 2016;39:312-28.

75. Jia H, Yang H, Sun P, Li J, Zhang J, Guo Y, *et al*. *De novo* transcriptome assembly, development of EST-SSR markers and population genetic analyses for the desert biomass willow, *Salix psammophila*. Sci Rep 2016;6:39591.

76. Mun JH, Kim DJ, Choi HK, Gish J, Debellé F, Mudge J, *et al*. Distribution of microsatellites in the genome of *Medicago truncatula*: A resource of genetic markers that integrate genetic and physical maps. Genetics 2006;172:2541-55.

77. Pandey G, Misra G, Kumari K, Gupta S, Parida SK, Chattopadhyay D, *et al*. Genome-wide development and use of microsatellite markers for large-scale genotyping applications in foxtail millet (*Setaria italica* (L.)). DNA Res 2013;20:197-207.

78. Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations, and genetic marker potential. Genome Res

2001;11:1441-52.

79. Gu T, Tan S, Gou X, Araki H, Tian D. Avoidance of long mononucleotide repeats in codon pair usage. Genetics 2010;186:1077-84.
80. Kong Q, Xiang C, Yu Z, Zhang C, Liu F, Peng C, *et al*. Mining and charactering microsatellites in *Cucumis melo* expressed sequence tags from sequence database. Mol Ecol Notes 2007;7:281-3.
81. Verma M, Arya L. Development of EST-SSRs in watermelon (*Citrullus lanatus* var. lanatus) and their transferability to *Cucumis* spp. J Horticult Sci Biotechnol 2008;83:732.
82. Liang X, Chen X, Hong Y, Liu H, Zhou G, Li S, *et al*. Utility of EST-derived SSR in cultivated peanut (*Arachis hypogaea* L.) and *Arachis* wild species. BMC Plant Biol 2009;9:35.
83. Qiu L, Yang C, Tian B, Yang JB, Liu A. Exploiting EST databases for the development and characterization of EST-SSR markers in castor bean (*Ricinus communis* L.). BMC Plant Biol 2010;10:278.
84. Tong Z, Yang Z, Chen X, Jiao F, Li X, Wu X, *et al*. Large-scale development of microsatellite markers in *Nicotiana tabacum* and construction of a genetic map of flue-cured tobacco. Plant Breeding 2012;131:674-80.
85. Tuskan G, DiFazio S, Teichmann T. Poplar genomics is getting popular: The impact of the poplar genome project on tree research. Plant Biol 2004;6:2-4.
86. Nagy I, Stágel A, Sasvári Z, Röder M, Ganal M. Development, characterization, and transferability to other *Solanaceae* of microsatellite markers in pepper (*Capsicum annuum* L.). Genome 2007;50:668-88.
87. Schwarzacher T, Zhang Y, Lin Z, Xia Q, Zhang M, Zhang X. Characteristics and analysis of simple sequence repeats in the cotton genome based on a linkage map constructed from a BC1 population between *Gossypium hirsutum* and *G. barbadense*. Genome 2008;51:534-46.
88. Cavagnaro PF, Chung SM, Manin S, Yildiz M, Ali A, Alessandro MS, *et al*. Microsatellite isolation and marker development in carrot-genomic distribution, linkage mapping, genetic diversity analysis and marker transferability across *Apiaceae*. BMC Genomics 2011;12:386.
89. Yang T, Jiang J, Burlyaeva M, Hu J, Coyne CJ, Kumar S, *et al*. Large-scale microsatellite development in grasspea (*Lathyrus sativus* L.), an orphan legume of the arid areas. BMC Plant Biol 2014;14:65.
90. Thiel T, Michalek W, Varshney R, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theor Appl Genet 2003;106:411-22.
91. Rota M, Kantety RV, Yu JK, Sorrells ME. Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat, and barley. BMC Genomics 2005;6:23.
92. Gupta P, Rustgi S, Sharma S, Singh R, Kumar N, Balyan H. Transferable EST-SSR markers for the study of polymorphism and genetic diversity in bread wheat. Mol Genet Genomics 2003;270:315-23.
93. Ebrahimi A, Mathur S, Lawson SS, LaBonte NR, Lorch A, Coggeshall MV, *et al*. Microsatellite borders and micro-sequence conservation in *Juglans*. Sci Rep 2019;9:1-10.
94. Li YC, Korol AB, Fahima T, Beiles A, Nevo E. Microsatellites: Genomic distribution, putative functions and mutational mechanisms: A review. Mol Ecol 2002;11:2453-65.
95. Kumar AS, Sowpati DT, Mishra RK. Single amino acid repeats in the proteome world: Structural, functional, and evolutionary insights. PLoS One 2016;11:e0166854.
96. Jiang D, Zhong GY, Qi-Bing H. Analysis of microsatellites in citrus unigenes. Acta Genet Sin 2006;33:345-53.
97. Wang Y, Chen M, Wang H, Wang JF, Bao D. Microsatellites in the genome of the edible mushroom, *Volvariella volvacea*. BioMed Res Int 2014;2014:281912.
98. Fu L, Ding Z, Kumpeangkeaw A, Tan D, Han B, Sun X, *et al*. *De novo* assembly, transcriptome characterization, and simple sequence repeat marker development in duckweed *Lemna gibba*. Physiol Mol Biol Plants 2020;26:133-42.

How to cite this article:
Ul Haq S, Dhingra P, Sharma M, Kothari SL, Kachhwaha S. Plasticity of tandem repeats in expressed sequence tags of angiospermic and non-angiospermic species: Insight into cladistic, phenetic and elementary explorations. J App Biol Biotech. 2021;9(2):36-59.
DOI: 10.7324/JABB.2021.9204

ADDITIONAL FILES

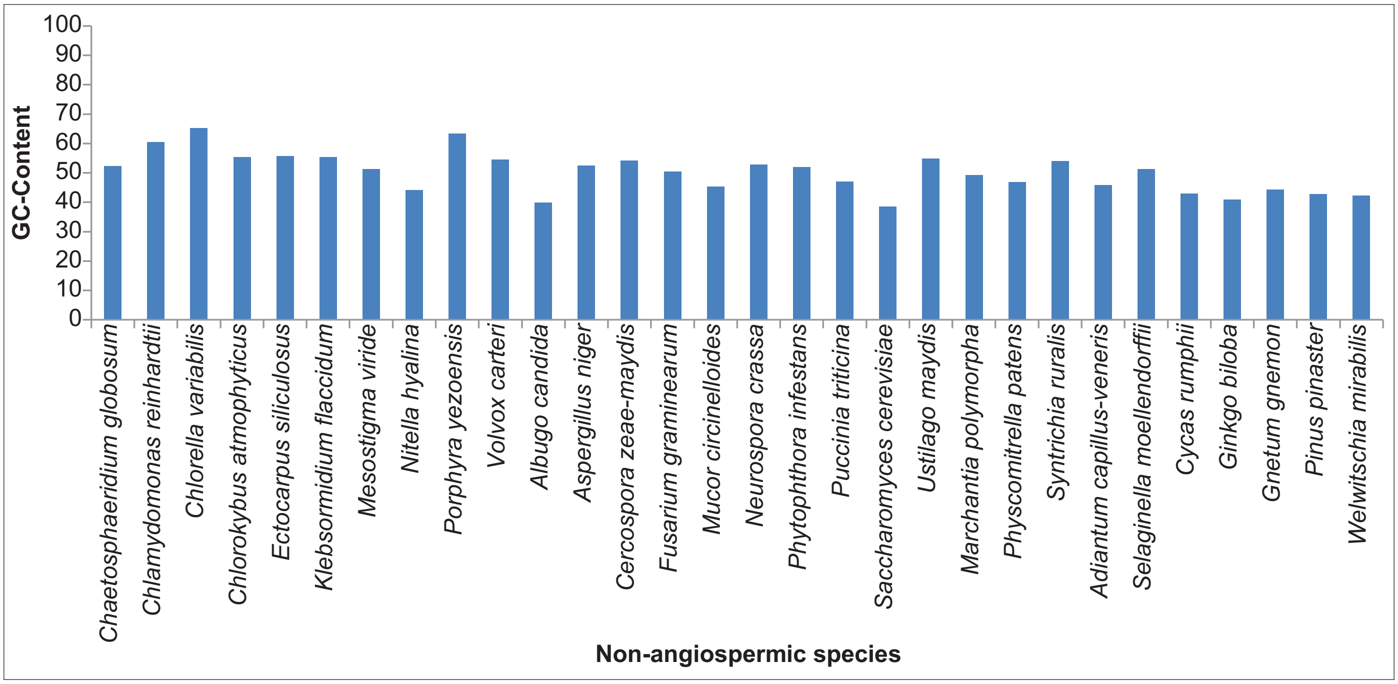
Additional file 1: Details of EST sequence characterizations amongst seventy five different species belonging to distinct evolutionary clades.

Serial no.	Plant Species	Average sequence length (nucleotides)
1	*Chaetosphaeridium globosum*	692.01
2	*Chlamydomonas reinhardtii*	683.48
3	*Chlorella variabilis*	754.86
4	*Chlorokybus atmophyticus*	953.14
5	*Ectocarpus siliculosus*	735.69
6	*Klebsormidium flaccidum*	947.35
7	*Mesostigma viride*	702.30
8	*Nitella hyalina*	730.02
9	*Porphyra yezoensis*	592.17
10	*Volvox carteri*	870.89
11	*Albugo candida*	1033.83
12	*Aspergillus niger*	912.83
13	*Cercospora zeae-maydis*	774.42
14	*Fusarium graminearum*	680.33
15	*Mucor circinelloides*	760.17
16	*Neurospora crassa*	895.66
17	*Phytophthora infestans*	649.40
18	*Puccinia triticina*	535.82
19	*Saccharomyces cerevisiae*	585.32
20	*Ustilago maydis*	789.02
21	*Marchantia polymorpha*	802.07
22	*Physcomitrella patens*	947.22
23	*Syntrichia ruralis*	628.52
24	*Adiantum capillus-veneris*	589.69
25	*Selaginella moellendorffii*	991.30
26	*Cycas rumphii*	683.22
27	*Ginkgo biloba*	669.07
28	*Gnetum gnemon*	573.55
29	*Pinus pinaster*	626.47
30	*Welwitschia mirabilis*	753.65
31	*Liriodendron tulipifera*	636.13
32	*Daucus carota*	529.22
33	*Panax ginseng*	805.12
34	*Catharanthus roseus*	569.59
35	*Artemisia annua*	827.60
36	*Helianthus annuus*	764.56
37	*Arabidopsis thaliana*	675.70

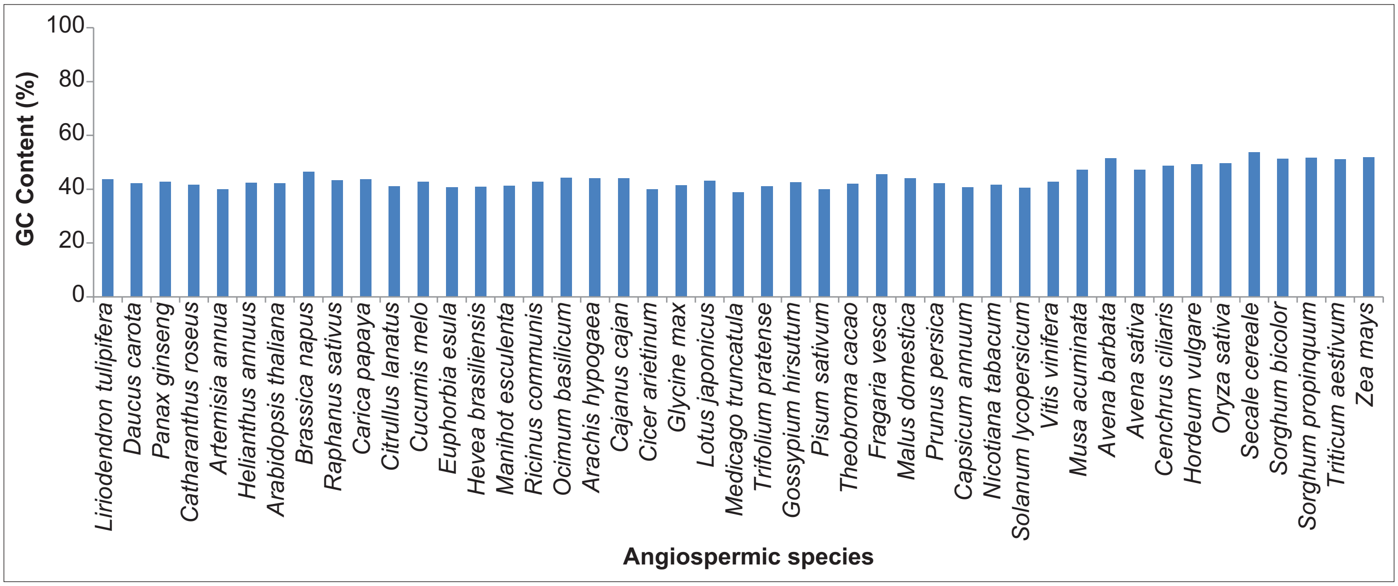
(*Contd...*)

Additional file 1: (*Continued*).

Serial no.	Plant Species	Average sequence length (nucleotides)
38	*Brassica napus*	823.45
39	*Raphanus sativus*	857.84
40	*Carica papaya*	914.18
41	*Citrullus lanatus*	535.15
42	*Cucumis melo*	719.85
43	*Euphorbia esula*	803.22
44	*Hevea brasiliensis*	643.11
45	*Manihot esculenta*	657.09
46	*Ricinus communis*	848.37
47	*Ocimum basilicum*	799.63
48	*Arachis hypogaea*	566.18
49	*Cajanus cajan*	580.99
50	*Cicer arietinum*	591.61
51	*Glycine max*	796.04
52	*Lotus japonicus*	513.56
53	*Medicago truncatula*	750.16
54	*Trifolium pratense*	673.01
55	*Gossypium hirsutum*	890.06
56	*Pisum sativum*	555.84
57	*Theobroma cacao*	524.43
58	*Musa acuminata*	647.10
59	*Avena barbata*	855.34
60	*Avena sativa*	634.58
61	*Cenchrus ciliaris*	803.52
62	*Hordeum vulgare*	641.13
63	*Oryza sativa*	837.47
64	*Secale cereale*	530.14
65	*Sorghum bicolor*	673.97
66	*Sorghum propinquum*	561.64
67	*Triticum aestivum*	644.58
68	*Zea mays*	590.59
69	*Fragaria vesca*	815.68
70	*Malus domestica*	625.90
71	*Prunus persica*	803.77
72	*Capsicum annuum*	726.52
73	*Nicotiana tabacum*	654.87
74	*Solanum lycopersicum*	654.77
75	*Vitis vinifera*	730.04



Additional file 2: Details of GC-content (%) within 30 species belonging non-angiospermic group.



Additional file 3: Details of GC-content (%) within 45 species belonging to angiospermic group.

Additional file 4: Comparative details of di, tri, tetra, penta and hexa nucleotide SSRs frequency (%) distribution amongst seventy five different species belonging to different phylogenetic groups.

Plant species	Nucleotide Repeat				
	Di	Tri	Tetra	Penta	Hexa
Chaetosphaeridium globosum	62.04	29.05	3.26	1.97	3.68
Chlamydomonas reinhardtii	34.05	59.61	1.84	2.04	2.45
Chlorella variabilis	4.11	91.45	1.45	1.37	1.61
Chlorokybus atmophyticus	6.98	83.49	2.25	2.30	4.98
Ectocarpus siliculosus	27.78	64.74	2.22	2.26	3.00
Klebsormidium flaccidum	34.89	53.60	3.24	3.96	4.32
Mesostigma viride	7.57	28.11	38.92	23.78	1.62
Nitella hyalina	46.02	38.72	8.16	4.59	2.50
Porphyra yezoensis	17.01	79.17	0.69	0.35	2.78
Volvox carteri	32.31	51.93	10.92	2.61	2.23
Albugo candida	61.26	26.58	0.45	3.15	8.56
Aspergillus niger	31.11	50.38	7.44	5.92	5.15
Cercospora zeae-maydis	19.64	66.94	5.12	5.39	2.90
Fusarium graminearum	27.56	51.92	1.28	8.33	10.90
Mucor circinelloides	32.97	64.99	0.47	0.16	1.41
Neurospora crassa	20.61	60.03	10.20	4.13	5.03
Phytophthora infestans	39.51	55.19	1.73	0.74	2.84
Puccinia triticina	59.63	31.47	2.28	4.14	2.48
Saccharomyces cerevisiae	58.73	33.33	1.90	2.22	3.81
Ustilago maydis	18.24	55.58	1.72	5.58	18.88
Marchantia polymorpha	24.41	57.37	7.98	7.98	2.25
Physcomitrella patens	48.99	38.54	5.36	4.83	2.28
Syntrichia ruralis	32.12	53.37	7.25	2.07	5.18
Adiantum capillus-veneris	82.00	14.48	0.91	0.34	2.27
Selaginella moellendorffii	23.30	69.31	3.08	1.22	3.08
Cycas rumphii	55.05	37.13	2.61	1.63	3.58
Ginkgo biloba	59.83	32.76	1.14	1.99	4.27
Gnetum gnemon	15.89	66.38	2.84	3.40	11.49
Pinus pinaster	45.18	42.15	1.38	3.86	7.44
Welwitschia mirabilis	34.50	52.05	2.92	2.92	7.60
Liriodendron tulipifera	65.74	27.49	1.85	1.72	3.20
Daucus carota	66.44	30.87	0.00	1.01	1.68
Panax ginseng	55.23	31.37	3.02	4.49	5.88
Catharanthus roseus	49.10	44.24	1.47	2.37	2.82
Artemisia annua	30.88	54.50	5.63	3.73	5.26
Helianthus annuus	39.23	50.50	3.43	2.93	3.90

(*Contd*...)

Additional file 4: (*Continued*).

Plant species	Nucleotide Repeat				
	Di	Tri	Tetra	Penta	Hexa
Arabidopsis thaliana	38.12	59.23	0.70	0.60	1.35
Brassica napus	52.73	44.08	0.76	0.94	1.49
Raphanus sativus	43.69	51.99	1.14	1.62	1.56
Carica papaya	58.53	34.35	2.04	1.91	3.17
Citrullus lanatus	42.17	46.30	4.13	3.48	3.91
Cucumis melo	38.94	50.12	2.75	3.73	4.47
Euphorbia esula	22.75	66.25	3.67	4.06	3.27
Hevea brasiliensis	60.96	31.26	1.67	2.71	3.40
Manihot esculenta	52.94	36.74	2.44	3.69	4.18
Ricinus communis	39.59	52.11	2.15	2.23	3.93
Ocimum basilicum	47.61	45.02	2.79	2.39	2.19
Arachis hypogaea	38.05	51.17	3.23	3.23	4.31
Cajanus cajan	57.06	31.21	4.97	3.18	3.58
Cicer arietinum	42.08	46.51	1.77	4.69	4.94
Glycine max	40.34	50.46	2.10	3.02	4.08
Lotus japonicus	35.00	53.06	1.81	3.30	6.83
Medicago truncatula	36.44	51.05	3.26	4.13	5.13
Trifolium pratense	24.04	69.51	2.31	1.64	2.52
Gossypium hirsutum	41.06	46.74	3.19	3.10	5.90
Pisum sativum	19.94	69.50	0.88	3.52	6.16
Theobroma cacao	51.14	38.19	2.97	4.04	3.66
Fragaria vesca	45.77	49.85	0.99	1.30	2.10
Malus domestica	63.44	29.74	1.73	2.14	2.96
Prunus persica	65.21	26.34	2.10	3.72	2.62
Capsicum annuum	54.65	39.05	1.97	1.71	2.61
Nicotiana tabacum	54.34	39.18	1.59	1.86	3.03
Solanum lycopersicum	33.97	59.64	1.05	1.92	3.41
Vitis vinifera	49.97	39.03	2.75	3.40	4.85
Musa acuminata	47.43	47.43	1.77	1.95	1.42
Avena barbata	22.35	64.56	4.48	3.97	4.63
Avena sativa	31.05	57.23	6.05	2.54	3.13
Cenchrus ciliaris	21.74	69.79	2.17	4.00	2.29
Hordeum vulgare	27.15	57.87	5.23	5.73	4.01
Oryza sativa	22.72	70.23	1.69	3.05	2.31
Secale cereale	15.38	71.05	6.48	4.25	2.83
Sorghum bicolor	24.76	63.47	2.88	5.00	3.88
Sorghum propinquum	22.25	64.36	4.00	5.18	4.21
Triticum aestivum	21.25	66.25	5.19	4.42	2.89
Zea mays	29.95	59.82	2.31	3.94	3.98

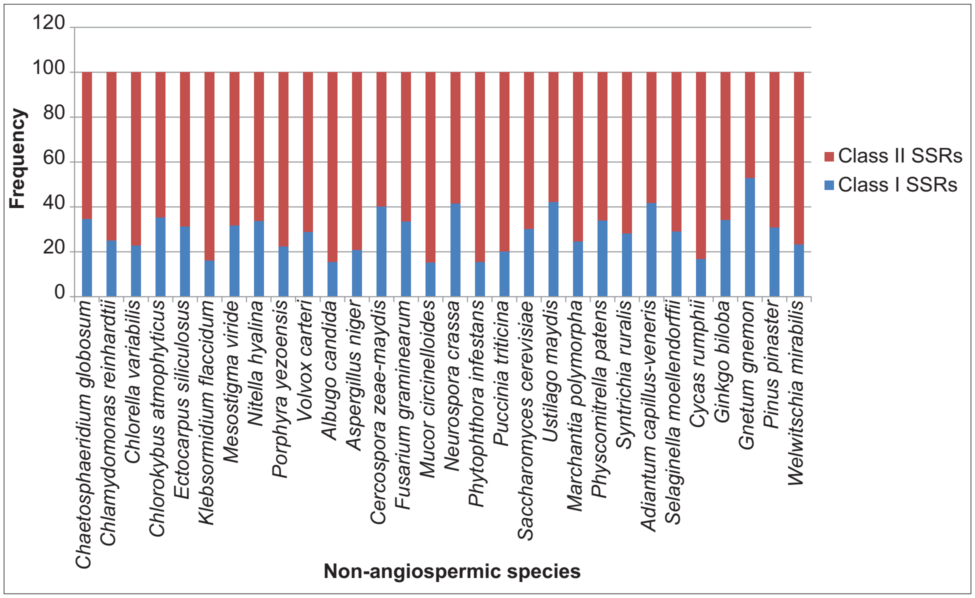
Additional file 5: Details of average SSRs or microsatellites motif length distributions amongst seventy-five different plant species.

Plant species	Motifs length					
	Mono	Di	Tri	Tetra	Penta	Hexa
Chaetosphaeridium globosum	22.8	24.02	18.02	22.75	39.09	26.75
Chlamydomonas reinhardtii	17.33	24.26	17.67	23.2	22.22	31
Chlorella variabilis	15.72	14.95	17.61	22.13	25	28
Chlorokybus atmophyticus	18.05	19.56	18.74	22.79	23	27.08
Ectocarpus siliculosus	12.02	16.6	19.81	23.23	25.2	26.94
Klebsormidium flaccidum	13.04	14.7	16.92	24	23.18	33
Mesostigma viride	23.03	16	21.83	33.47	23.88	30
Nitella hyalina	14.37	21.29	19.19	30.49	25.41	26.3
Porphyra yezoensis	15.78	22.9	18.22	0	24	38
Volvox carteri	33.5	18.04	19.06	32.6	24.58	27.2
Albugo candida	19.12	13.64	16.36	20	20	30.66
Aspergillus niger	23.23	15.66	17.65	21.87	22.1	25.75
Cercospora zeae-maydis	19.1	19.01	20.76	24.53	22.27	29.25
Fusarium graminearum	21.85	22.57	17.48	36	23.07	24.42
Mucor circinelloides	13.83	15.17	17.02	21.33	20	24.85
Neurospora crassa	17.71	26.05	18.65	24.2	23.87	27.58
Phytophthora infestans	26.53	13.97	16.8	20	20	26.66
Puccinia triticina	14.13	19.28	19.5	20	27.33	26
Saccharomyces cerevisiae	12.31	17.5	18.38	22	24.28	26.25
Ustilago maydis	20.65	17.52	18.41	26.28	22.27	31.94
Marchantia polymorpha	12.12	18.11	17.31	21.79	21.83	24.3
Physcomitrella patens	23.84	20.42	17.75	23.83	21.78	24
Syntrichia ruralis	17.22	17.02	17.8	24.66	20	25.33
Adiantum capillus-veneris	24.59	24.03	17.24	21.2	22	26.72
Selaginella moellendorffii	23.38	17.6	18.19	21.61	21.94	26.11
Cycas rumphii	22.26	18.58	16.47	20	20	24.75
Ginkgo biloba	26.12	21.3	16.3	22.66	22	26.3
Gnetum gnemon	20.65	16.53	17.34	21.06	20.58	25.52
Pinus pinaster	23.92	30.47	16.06	20	20.41	25.3
Welwitschia mirabilis	18.47	20.47	16.3	24	27.5	28.66
Liriodendron tulipifera	19.22	24.78	18.56	21.06	23.57	26.36
Daucus carota	17.35	16.67	16.75	0	21.66	25.5
Panax ginseng	16.23	19.03	18.31	24.88	22.09	27.04
Catharanthus roseus	20.32	21.71	18.39	21.45	20.76	26.18
Artemisia annua	19.48	16.22	17.09	23.73	22.59	26.15
Helianthus annuus	14.6	15.94	17.05	22.97	21.66	26.09
Arabidopsis thaliana	21.87	21.17	17.29	21.53	20	27.36
Brassica napus	21.3	17.34	17.07	21.11	20.04	26.59
Raphanus sativus	18.32	16.13	17.31	21.18	21.16	28.52
Carica papaya	17.94	18.12	19.02	23.41	22.43	26.67
Citrullus lanatus	16.07	17.16	19.83	22.9	20.55	26.57
Cucumis melo	13.12	19.71	21.63	26.53	22.53	28.2
Euphorbia esula	17.11	21	19.83	23.29	21.23	26.4
Hevea brasiliensis	19.23	32.16	19.68	22.3	21.62	27.44
Manihot esculenta	16.49	18.75	18.25	22.84	22.39	26.48
Ricinus communis	18.34	18.96	18.16	23.03	20.87	25.74
Ocimum basilicum	18.14	19.55	17.97	22.22	20.45	25.8

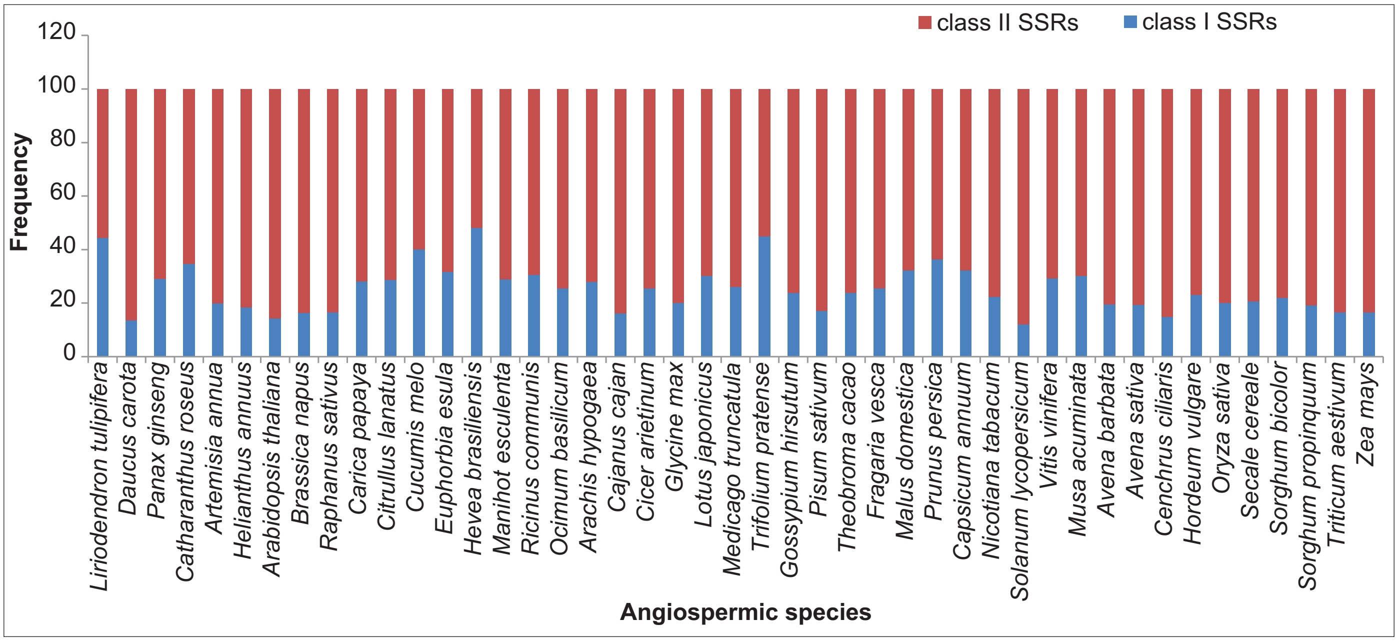
(*Contd...*)

Additional file 5: (*Continued*).

Plant species	Motifs length					
	Mono	Di	Tri	Tetra	Penta	Hexa
Arachis hypogaea	13.07	19.39	18.05	23.19	21.51	25.73
Cajanus cajan	22.9	14.07	17.66	22.28	21	24.85
Cicer arietinum	24.31	18.29	17.62	23.2	28.83	26.06
Glycine max	18.21	15.78	17.48	21.14	20.79	25.09
Lotus japonicus	13.62	19.83	18.07	21.67	21.53	26.1
Medicago truncatula	12.36	19.1	17.76	22.02	21.33	25.88
Trifolium pratense	12.4	25.19	22.98	22.87	21.29	26.19
Gossypium hirsutum	16.69	15.55	18	22.86	22.11	27.15
Pisum sativum	16.96	14.37	16.88	25.33	20	26.33
Theobroma cacao	25.28	16.64	18	22.51	23.37	25.14
Fragaria vesca	18.32	20.11	17.37	22.54	20	24.38
Malus domestica	17.47	22.09	17.38	23.17	21.59	26.24
Prunus persica	19.54	22.33	18	21.52	22.04	26.75
Capsicum annuum	19.8	30.53	17.22	20.05	22.42	26.72
Nicotiana tabacum	24.55	18	18.36	22.77	20.97	26.24
Solanum lycopersicum	12.87	14.19	16.73	20.3	20.71	25
Vitis vinifera	24.08	23.19	18.02	22.4	21.09	26.458
Musa acuminata	16.46	18.97	18.57	21.64	21.5	26.8
Avena barbata	17.61	16.84	16.95	23.02	21.58	26.08
Avena sativa	18	17.83	17.57	21.27	20.45	26
Cenchrus ciliaris	22.42	18.02	16.53	20.57	21.85	24
Hordeum vulgare	16	18.63	17.31	21.76	21.1	24.52
Oryza sativa	19.5	17.75	17.29	20.91	21.13	24.88
Secale cereale	20.9	17.82	17.15	20.5	21	24
Sorghum bicolor	11.75	18.11	17.37	23.33	22.15	24.65
Sorghum propinquum	16.6	19.93	16.84	20.72	20.93	24.9
Triticum aestivum	10.06	14.02	17.46	21.9	21.89	26.6
Zea mays	15.15	15.97	16.58	21.65	21.5	24.87



Additional file 6: Detail of class I and class II perfect SSRs within 30 species belonging to non-angiospermic group.



Additional file 7: Detail of class I and class II perfect SSRs within 45 species belonging to angiospermic group.

Additional file 8: Frequency distribution (%) of mono and di nucleotide SSR motifs in 75 different species.

Plant Species	A/T	C/G	AC/GT	AG/CT	AT/AT	CG/CG
Chaetosphaeridium globosum	75.31	24.69	34.94	59.81	2.49	2.76
Chlamydomonas reinhardtii	76.16	23.84	72.97	15.92	3.60	7.51
Chlorella variabilis	77.30	22.70	58.82	25.49	1.96	13.73
Chlorokybus atmophyticus	91.21	8.79	58.39	31.06	1.86	8.70
Ectocarpus siliculosus	51.49	48.51	36.68	56.58	3.61	3.13
Klebsormidium flaccidum	74.53	25.47	22.68	15.46	2.06	59.79
Mesostigma viride	95.90	4.10	35.71	28.57	14.29	21.43
Nitella hyalina	85.81	14.19	26.47	64.29	7.68	1.55
Porphyra yezoensis	56.48	43.52	26.53	40.82	2.04	30.61
Volvox carteri	55.57	44.43	83.62	11.96	2.73	1.69
Albugo candida	89.79	10.21	10.29	16.91	72.06	0.74
Aspergillus niger	89.20	10.80	28.83	57.06	13.50	0.61
Cercospora zeae-maydis	88.36	11.64	44.37	44.37	7.04	4.23
Fusarium graminearum	95.59	4.41	30.23	55.81	9.30	4.65
Mucor circinelloides	90.76	9.24	41.43	41.43	17.14	0.00
Neurospora crassa	81.59	18.41	27.78	58.94	11.35	1.93
Phytophthora infestans	91.53	8.47	28.44	50.00	19.06	2.50
Puccinia triticina	81.04	18.96	21.88	64.58	12.85	0.69
Saccharomyces cerevisiae	99.82	0.18	9.19	3.78	87.03	0.00
Ustilago maydis	62.82	37.18	45.88	51.76	2.35	0.00
Marchantia polymorpha	99.09	0.91	5.77	89.62	2.31	2.31
Physcomitrella patens	78.61	21.39	24.73	40.68	34.05	0.54
Syntrichia ruralis	93.41	6.59	40.32	51.61	8.06	0.00
Adiantum capillus-veneris	94.08	5.92	29.78	67.24	0.55	2.42
Selaginella moellendorffii	82.02	17.98	17.21	70.65	8.88	3.26
Cycas rumphii	92.75	7.25	19.53	34.32	46.15	0.00
Ginkgo biloba	89.85	10.15	13.81	35.71	50.48	0.00
Gnetum gnemon	97.78	2.22	20.54	55.36	24.11	0.00

(*Contd...*)

Additional file 8: (*Continued*).

Plant Species	A/T	C/G	AC/GT	AG/CT	AT/AT	CG/CG
Pinus pinaster	99.53	0.47	6.71	53.66	39.63	0.00
Welwitschia mirabilis	95.21	4.79	13.56	62.71	23.73	0.00
Daucus carota	99.13	0.87	17.17	78.79	4.04	0.00
Panax ginseng	85.95	14.05	7.10	41.27	51.18	0.44
Catharanthus roseus	98.75	1.25	3.22	63.22	33.56	0.00
Artemisia annua	87.29	12.71	55.88	22.66	20.78	0.68
Helianthus annuus	84.00	16.00	20.22	65.73	14.04	0.00
Arabidopsis thaliana	92.42	7.58	10.12	72.20	17.32	0.37
Brassica napus	85.36	14.64	10.07	73.89	15.39	0.64
Raphanus sativus	99.56	0.44	10.84	80.17	8.94	0.05
Carica papaya	77.23	22.77	11.17	55.26	32.81	0.76
Citrullus lanatus	99.35	0.65	8.25	46.91	44.85	0.00
Cucumis melo	92.50	7.50	8.61	69.26	21.89	0.24
Euphorbia esula	98.47	1.53	7.11	62.22	30.50	0.17
Hevea brasiliensis	98.71	1.29	3.31	79.28	17.22	0.19
Manihot esculenta	97.66	2.34	6.06	73.39	20.55	0.00
Ricinus communis	97.19	2.81	5.74	70.88	23.28	0.10
Ocimum basilicum	80.06	19.94	21.76	53.14	25.10	0.00
Arachis hypogaea	92.11	7.89	8.13	81.95	9.86	0.06
Cajanus cajan	89.27	10.73	21.60	42.86	30.66	4.88
Cicer arietinum	91.76	8.24	10.84	58.13	30.42	0.60
Glycine max	98.19	1.81	11.64	72.94	15.33	0.09
Lotus japonicus	92.34	7.66	12.68	76.32	10.77	0.22
Medicago truncatula	94.88	5.12	11.24	66.73	21.86	0.18
Trifolium pratense	88.89	11.11	9.42	78.18	12.39	0.00
Gossypium hirsutum	76.73	23.27	13.10	40.68	45.58	0.64
Pisum sativum	99.76	0.24	16.18	55.88	27.94	0.00
Theobroma cacao	90.05	9.95	5.56	69.88	24.57	0.00
Fragaria vesca	90.98	9.02	5.66	85.85	8.36	0.13
Malus domestica	98.72	1.28	6.00	82.28	11.72	0.00
Prunus persica	87.93	12.07	4.82	75.18	19.92	0.09
Capsicum annuum	81.99	18.01	10.12	65.41	24.47	0.00
Nicotiana tabacum	92.70	7.30	24.76	61.09	13.81	0.35
Solanum lycopersicum	77.45	22.55	12.41	54.01	33.39	0.18
Vitis vinifera	99.31	0.69	4.53	70.23	25.13	0.11
Liriodendron tulipifera	99.12	0.88	8.33	86.44	5.05	0.19
Musa acuminata	87.62	12.38	9.16	77.38	13.27	0.19
Avena barbata	89.20	10.80	27.33	44.87	19.13	8.66
Avena sativa	99.07	0.93	29.56	48.43	20.75	1.26
Cenchrus ciliaris	92.21	7.79	16.84	65.26	14.21	3.68
Hordeum vulgare	83.86	16.14	27.57	51.00	16.45	4.98
Oryza sativa	62.94	37.06	13.95	65.86	12.17	8.02
Secale cereale	98.14	1.86	26.32	48.68	19.74	5.26
Sorghum bicolor	95.83	4.17	23.04	47.98	20.67	8.31
Sorghum propinquum	97.82	2.18	24.27	49.51	16.50	9.71
Triticum aestivum	99.88	0.12	30.59	56.04	6.43	6.94
Zea mays	65.41	34.59	22.27	58.96	11.90	6.86

Additional file 9: Comparative details of different types of tri nucleotide SSRs motifs distribution amongst seventy five distinct species.

Plant species	AAC/GTT	AAG/CTT	AAT/ATT	ACC/GGT	ACG/CGT	ACT/AGT	AGC/CTG	AGG/CCT	ATC/ATG	CCG/CGG
Chaetosphaeridium globosum	6.49	13.86	1.18	10.03	4.72	0.29	29.20	16.22	5.60	12.39
Chlamydomonas reinhardtii	1.72	0.51	0.00	6.86	1.03	0.00	40.65	5.49	0.69	43.05
Chlorella variabilis	0.09	1.41	0.18	2.38	0.35	0.00	38.45	8.11	0.44	48.59
Chlorokybus atmophyticus	13.29	1.77	0.10	4.15	8.77	0.42	17.96	10.38	3.12	40.03
Ectocarpus siliculosus	8.20	2.69	0.13	6.52	3.77	0.27	42.43	8.34	2.35	25.29
Klebsormidium flaccidum	2.01	26.17	0.00	2.68	4.70	0.00	21.48	20.81	3.36	18.79
Mesostigma viride	3.85	15.38	9.62	3.85	1.92	1.92	21.15	21.15	9.62	11.54
Nitella hyalina	14.18	21.90	4.40	6.52	3.15	2.23	9.18	8.91	27.50	2.01
Porphyra yezoensis	10.09	2.19	0.00	5.70	3.95	1.75	18.86	3.95	0.44	53.07
Volvox carteri	5.58	1.86	1.05	13.75	1.13	0.89	33.82	8.58	10.03	23.30
Albugo candida	6.78	32.20	30.51	5.08	5.08	1.69	6.78	5.08	3.39	3.39
Aspergillus niger	10.23	16.29	3.41	13.26	1.89	9.09	19.70	9.47	9.85	6.82
Cercospora zeae-maydis	12.19	9.50	0.62	14.05	6.82	4.13	23.14	9.09	13.43	7.02
Fusarium graminearum	16.05	22.22	0.00	6.17	7.41	2.47	12.35	16.05	9.88	7.41
Mucor circinelloides	18.60	6.52	1.21	5.80	2.66	6.04	46.86	6.28	4.11	1.93
Neurospora crassa	18.41	13.27	0.58	11.44	5.14	1.66	16.92	15.26	8.04	9.29
Phytophthora infestans	4.92	25.06	8.50	12.30	5.59	0.67	24.16	9.40	3.13	6.26
Puccinia triticina	11.18	23.03	1.32	9.21	4.61	4.61	11.18	15.13	15.13	4.61
Saccharomyces cerevisiae	20.00	22.86	19.05	0.95	4.76	1.90	18.10	1.90	10.48	0.00
Ustilago maydis	10.42	8.88	0.00	8.49	15.06	0.00	42.47	3.86	6.56	4.25
Marchantia polymorpha	3.44	11.13	0.00	4.58	5.89	1.15	46.15	23.24	4.26	0.16
Physcomitrella patens	10.02	18.68	5.47	9.34	7.74	2.05	25.28	13.90	6.38	1.14
Syntrichia ruralis	2.91	3.88	0.97	1.94	14.56	3.88	49.51	7.77	5.83	8.74
Adiantum capillus-veneris	3.14	22.35	0.39	10.20	3.14	1.18	20.78	16.47	20.39	1.96
Selaginella moellendorffii	2.13	11.39	1.58	13.52	2.80	0.85	42.63	10.29	0.79	14.01
Cycas rumphii	4.39	22.81	21.05	2.63	3.51	0.00	18.42	16.67	10.53	0.00
Ginkgo biloba	4.35	21.74	16.52	9.57	4.35	0.00	14.78	12.17	15.65	0.87
Gnetum gnemon	6.62	21.15	3.85	1.71	2.99	0.00	44.44	8.33	9.40	1.50
Pinus pinaster	2.61	19.61	9.80	11.76	2.61	1.96	16.99	15.03	15.03	4.58
Welwitschia mirabilis	6.74	30.34	4.49	3.37	1.12	0.00	28.09	12.36	11.24	2.25
Liriodendron tulipifera	4.92	41.39	6.71	4.70	2.68	0.45	22.82	2.68	12.53	1.12
Daucus carota	5.43	26.09	1.09	8.70	1.09	2.17	15.22	6.52	25.00	8.70
Panax ginseng	3.13	26.56	11.46	12.50	1.30	2.60	15.36	12.24	11.20	3.65
Catharanthus roseus	4.59	35.46	12.50	5.36	2.04	1.53	12.24	9.44	13.27	3.57
Artemisia annua	19.40	13.13	15.93	17.08	0.29	1.64	7.72	3.09	20.66	1.06
Helianthus annuus	7.92	19.45	12.47	23.50	0.87	1.25	8.42	4.55	19.20	2.37

(*Contd...*)

Additional file 9: (*Continued*).

Plant species	**AAC/GTT**	**AAG/CTT**	**AAT/ATT**	**ACC/GGT**	**ACG/CGT**	**ACT/AGT**	**AGC/CTG**	**AGG/CCT**	**ATC/ATG**	**CCG/CGG**
Arabidopsis thaliana	12.56	46.47	1.65	5.97	1.18	1.65	5.02	7.77	16.41	1.33
Brassica napus	9.70	31.48	3.56	8.34	1.91	4.30	6.54	13.89	16.57	3.71
Raphanus sativus	11.51	33.36	2.92	7.79	1.72	1.92	6.79	14.74	17.02	2.24
Carica papaya	3.87	50.55	6.74	5.25	1.78	0.69	8.62	8.13	13.18	1.19
Citrullus lanatus	3.76	40.85	15.02	5.63	1.88	1.41	10.33	5.16	11.27	4.69
Cucumis melo	4.37	61.71	4.55	3.53	1.49	0.28	4.65	9.01	7.06	3.35
Euphorbia esula	3.45	40.65	13.51	4.82	0.89	3.33	4.23	11.01	16.19	1.90
Hevea brasiliensis	2.40	38.75	17.16	6.27	1.29	0.74	12.18	9.23	10.89	1.11
Manihot esculenta	2.10	37.90	13.26	8.62	0.99	0.55	12.38	7.85	13.59	2.76
Ricinus communis	5.39	31.01	11.50	11.50	1.82	2.06	16.10	6.90	8.88	4.84
Ocimum basilicum	3.54	13.72	11.95	11.95	2.21	2.21	13.72	15.04	10.62	15.04
Arachis hypogaea	8.75	40.86	9.80	8.84	1.51	1.79	5.36	7.47	12.05	3.57
Cajanus cajan	5.10	22.29	26.11	6.37	3.18	2.55	9.55	3.82	17.83	3.18
Cicer arietinum	13.62	25.61	19.89	12.26	1.09	2.45	4.90	3.81	14.99	1.36
Glycine max	15.20	29.43	9.98	9.30	3.40	2.50	7.34	6.58	12.56	3.71
Lotus japonicus	12.73	34.27	2.37	18.58	1.70	1.63	5.70	8.14	10.95	3.92
Medicago truncatula	12.76	36.13	10.68	7.39	1.83	3.92	5.69	4.93	15.86	0.82
Trifolium pratense	12.25	14.54	4.16	32.59	0.06	2.17	3.20	17.44	13.52	0.06
Gossypium hirsutum	7.39	26.38	8.79	13.75	1.96	1.50	11.13	5.43	20.21	3.46
Pisum sativum	19.41	22.36	15.19	11.81	0.42	1.27	5.49	2.11	19.83	2.11
Theobroma cacao	6.61	37.19	12.89	7.93	1.16	0.99	9.92	9.26	12.07	1.98
Fragaria vesca	7.30	31.44	1.86	10.89	3.59	1.36	9.41	20.05	7.18	6.93
Malus domestica	7.58	24.21	4.13	12.01	4.72	1.38	15.85	16.63	8.07	5.41
Prunus persica	6.37	28.94	8.68	8.56	2.78	0.58	18.75	10.30	12.73	2.31
Capsicum annuum	17.45	25.25	13.83	10.43	0.55	3.18	7.90	6.70	10.87	3.84
Nicotiana tabacum	10.83	43.04	8.90	6.87	1.93	3.48	8.12	7.06	5.51	4.26
Solanum lycopersicum	8.52	36.69	11.33	9.15	1.77	2.08	9.25	5.20	10.19	5.82
Vitis vinifera	3.31	29.56	11.33	12.57	2.21	0.97	13.81	9.67	12.71	3.87
Musa acuminata	2.06	27.10	2.99	7.29	5.23	0.93	14.39	23.18	7.10	9.72
Avena barbata	3.39	7.18	2.13	7.65	5.44	3.15	18.93	16.88	4.50	30.76
Avena sativa	5.80	10.92	2.05	8.53	3.41	3.41	18.43	17.41	4.10	25.94
Cenchrus ciliaris	1.48	3.61	0.82	5.08	5.74	0.16	16.89	14.26	1.80	50.16
Hordeum vulgare	3.12	9.35	1.33	6.00	6.39	1.56	17.93	14.19	5.92	34.22
Oryza sativa	1.66	5.97	1.10	5.81	8.09	0.85	8.40	16.49	2.21	49.43
Secale cereale	1.99	6.84	1.71	5.98	9.40	1.14	14.53	14.81	3.99	39.60
Sorghum bicolor	1.85	6.02	1.48	5.93	11.68	0.93	16.77	12.79	2.69	39.85
Sorghum propinquum	1.85	4.53	2.35	6.21	6.54	1.51	16.28	13.26	2.35	45.13
Triticum aestivum	1.65	7.67	1.24	7.91	4.20	1.90	17.56	15.42	4.78	37.68
Zea mays	1.96	4.70	1.61	6.31	8.06	1.26	15.99	11.92	1.82	46.35

Additional file 10: Comparative details of amino acids frequency (%) distribution amongst 75 different species.

Species/Amino acids	Ala	Am*	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Oc*	Op*	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Chaetosphaeridium globosum	15.93	0.29	12.39	1.18	3.83	4.13	9.14	6.49	6.49	2.95	1.47	6.78	5.90	1.18	0.00	1.18	0.00	8.26	5.60	2.95	2.06	0.00	1.77
Chlamydomonas reinhardtii	16.41	0.30	12.77	1.22	3.95	4.26	9.42	3.65	6.69	3.04	1.52	6.99	6.08	1.22	0.00	1.22	0.00	8.51	5.78	3.04	2.13	0.00	1.82
Chlorella variabilis	36.42	0.00	12.96	0.00	0.09	3.00	8.64	2.82	12.79	0.44	0.00	7.94	0.53	0.18	0.00	0.09	0.09	7.94	5.20	0.26	0.53	0.00	0.09
Chlorokybus atmophyticus	17.91	0.16	16.15	1.19	1.66	3.17	10.33	2.86	15.68	1.45	0.10	4.15	0.73	0.21	0.00	0.52	0.21	7.11	6.44	5.19	0.42	0.00	4.36
Ectocarpus siliculosus	24.75	0.00	9.62	0.67	1.68	5.31	12.17	3.56	12.24	1.75	0.13	7.53	1.08	0.27	0.00	0.61	0.00	3.23	9.08	2.08	1.01	0.07	3.16
Klebsormidium flaccidum	10.74	0.00	16.11	0.00	0.67	1.34	6.71	11.41	6.71	0.00	0.00	11.41	12.75	1.34	0.00	1.34	2.68	4.03	8.72	2.01	0.67	0.00	1.34
Mesostigma viride	9.62	0.00	7.69	1.92	0.00	0.00	7.69	17.31	3.85	1.92	3.85	13.46	7.69	0.00	1.92	1.92	1.92	7.69	5.77	1.92	0.00	1.92	1.92
Nitella hyalina	3.26	0.43	5.71	4.13	6.14	2.45	6.90	5.71	2.83	4.84	4.35	10.49	3.91	4.18	0.43	5.27	4.24	2.55	13.42	4.40	0.92	0.87	2.55
porphyra yezoensis	26.32	0.44	15.35	4.39	1.32	0.88	8.33	0.88	17.54	0.44	0.00	4.82	0.88	0.00	0.00	0.00	0.00	5.70	3.95	3.07	1.75	0.00	3.95
volvox carteri	22.49	0.00	7.44	1.29	1.70	6.55	7.20	2.02	7.36	5.99	2.10	8.25	0.40	1.46	0.24	1.05	0.24	8.09	7.93	3.16	2.02	0.24	2.75
Albugo candida	0.00	0.00	11.86	5.08	0.00	1.69	5.08	6.78	0.00	5.08	11.86	20.34	1.69	0.00	1.69	0.00	5.08	5.08	10.17	5.08	0.00	3.39	0.00
Aspergillus niger	8.33	0.38	3.79	2.27	2.65	5.30	7.58	5.30	3.41	6.44	2.27	12.50	4.17	2.27	0.00	1.52	1.14	7.20	8.33	7.95	0.38	3.79	3.03
Cercospora zeae-maydis	9.09	0.21	6.20	2.27	4.75	4.55	8.47	2.69	4.34	7.44	2.27	11.57	1.45	2.07	0.21	1.65	1.45	7.02	9.92	8.47	0.62	0.83	2.48
Fusarium graminearum	7.41	1.23	11.11	4.94	6.17	6.17	4.94	13.58	7.41	4.94	2.47	9.88	4.94	0.00	0.00	1.23	0.00	0.00	7.41	1.23	0.00	0.00	4.94
Mucor circinelloides	9.18	0.72	2.66	1.93	2.17	6.76	29.47	3.62	2.42	2.42	0.97	13.04	2.42	0.48	0.24	0.00	0.00	3.14	8.70	5.07	0.72	1.21	2.66
Neurospora crassa	9.62	0.08	5.56	1.74	3.23	5.39	8.13	5.22	3.57	4.56	1.24	13.85	2.49	0.66	0.17	1.24	1.41	7.21	13.10	5.64	1.58	0.33	3.98
Phytophthora infestans	9.17	0.00	6.71	2.68	2.68	2.68	11.19	7.38	4.70	7.16	3.36	9.17	7.61	0.22	0.89	0.67	3.80	4.03	8.50	2.68	1.12	1.57	2.01
Puccinia triticina	0.66	0.66	9.21	2.63	2.63	3.95	11.84	10.53	6.58	4.61	3.95	7.24	5.92	0.66	0.00	2.63	1.97	3.29	11.84	3.29	1.32	1.32	3.29
Saccharomyces cerevisiae	5.71	0.00	3.81	6.67	5.71	2.86	20.95	9.52	0.00	0.00	6.67	5.71	7.62	2.86	2.86	0.95	0.00	0.95	6.67	4.76	0.00	3.81	1.90
Ustilago maydis	17.76	0.00	5.02	1.16	9.27	1.93	20.08	4.63	2.32	4.63	0.39	7.34	3.09	0.00	0.00	0.77	0.00	2.70	10.04	6.56	1.16	0.00	1.16
Marchantia polymorpha	18.17	0.33	7.53	0.49	2.62	3.27	12.11	10.64	6.87	1.31	0.33	10.31	1.31	0.49	0.00	0.33	1.15	2.95	15.06	1.64	0.82	0.16	2.13
Physcomitrella patens	7.74	0.23	6.15	0.91	2.73	8.20	5.24	5.24	5.47	2.96	0.91	17.08	1.14	0.23	1.14	2.51	2.96	3.42	14.58	4.56	1.37	0.68	4.56
Syntrichia ruralis	18.45	0.97	9.71	0.00	0.97	1.94	13.59	0.00	4.85	0.97	0.97	14.56	1.94	1.94	0.00	0.97	0.00	3.88	14.56	2.91	0.00	0.97	5.83
Adiantum capillus-veneris	8.63	0.39	7.06	1.18	5.88	1.96	7.45	8.63	3.92	5.49	0.39	7.45	5.10	6.67	0.39	2.35	2.35	3.14	12.55	2.35	3.14	0.00	3.53
Selaginella moellendorffii	16.38	0.00	5.97	0.79	0.79	9.56	9.99	5.54	7.19	2.62	0.55	11.02	2.01	0.12	0.18	0.18	2.19	6.94	11.45	1.28	2.74	0.18	2.31
Cycas rumphii	6.14	0.00	7.02	5.26	4.39	2.63	7.89	12.28	4.39	1.75	10.53	7.89	5.26	0.88	1.75	2.63	0.88	1.75	10.53	0.00	1.75	3.51	0.88
Ginkgo biloba	3.48	0.00	9.57	1.74	5.22	1.74	9.57	4.35	3.48	4.35	9.57	9.57	6.96	3.48	1.74	3.48	0.87	5.22	7.83	0.87	2.61	4.35	0.00
Gnetum gnemon	12.39	0.00	7.26	1.71	1.71	6.84	14.96	3.63	1.07	3.42	3.85	10.90	2.99	0.21	0.64	0.43	3.42	1.50	20.51	1.71	0.00	0.21	0.64
Pinus pinaster	7.19	0.00	6.54	5.88	5.23	1.31	5.88	9.15	6.54	5.88	3.92	7.19	7.84	2.61	0.00	3.92	1.96	3.27	3.92	2.61	5.88	1.96	1.31
Welwitschia mirabilis	8.99	0.00	12.36	4.49	2.25	3.37	15.73	5.62	2.25	2.25	2.25	7.87	5.62	2.25	0.00	5.62	7.87	2.25	8.99	0.00	0.00	0.00	0.00
Arabidopsis thaliana	1.49	0.08	8.56	3.06	3.53	3.14	3.45	9.11	3.45	3.53	3.85	14.13	3.92	1.18	0.08	2.59	8.08	2.83	16.64	3.38	0.78	0.63	2.51
Arachis hypogaea	1.83	0.18	6.00	5.36	2.34	1.37	4.99	7.51	3.34	4.49	4.67	13.06	5.22	1.56	1.37	1.92	8.06	5.36	13.93	3.99	0.78	1.42	1.24
Artemisia annua	1.93	0.19	2.32	8.59	4.25	4.05	6.76	3.86	3.09	6.56	8.59	10.91	2.03	2.32	2.03	3.38	2.22	3.96	8.11	4.83	3.38	2.03	4.63
Brassica napus	3.01	0.26	7.60	2.53	3.97	2.50	3.09	7.49	4.59	3.34	3.71	11.94	5.62	1.91	0.66	3.09	4.11	5.03	15.65	3.89	1.80	1.40	2.79
Cajanus cajan	2.55	0.00	9.55	4.46	1.91	3.82	1.91	5.73	0.64	1.91	11.46	10.83	2.55	1.91	1.91	4.46	2.55	3.82	13.38	3.18	1.27	6.37	3.82
Capsicum annuum	2.96	0.77	5.71	6.59	1.54	3.07	7.68	6.59	2.52	4.72	5.49	10.76	3.84	1.87	2.41	1.98	5.27	5.16	9.55	5.60	1.76	2.31	1.87
Carica papaya	2.87	0.20	7.04	1.29	3.87	0.59	3.57	9.81	2.78	3.67	6.44	13.48	6.05	0.79	0.89	1.88	10.90	2.08	17.54	0.89	0.89	1.49	0.99

(*Contd...*)

Additional file 10: (*Continued*).

Species/Amino acids	Ala	Am*	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Oc*	Op*	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Catharanthus roseus	5.39	0.00	8.33	4.41	4.17	3.68	8.09	3.19	2.21	5.64	10.05	6.13	4.90	0.98	1.47	2.21	7.35	4.66	10.29	1.72	0.98	1.96	2.21
Cicer arietinum	1.63	0.27	3.00	5.45	3.27	2.72	6.27	5.45	0.54	5.18	8.72	12.53	3.27	1.91	2.18	2.72	7.36	3.81	8.99	4.63	1.91	5.45	2.72
Citrullus lanatus	4.69	0.47	3.76	6.10	3.29	1.41	5.63	9.39	4.69	2.35	7.98	15.02	5.16	1.41	1.88	2.35	7.98	3.29	8.92	0.47	0.47	1.88	1.41
Cucumis melo	2.60	0.09	6.88	2.70	2.51	0.93	2.04	7.81	2.51	2.14	3.35	17.75	5.58	0.65	0.56	0.28	15.80	3.53	18.87	1.95	0.56	0.46	0.46
Daucus carota	7.61	0.00	6.52	1.09	7.61	1.09	6.52	8.70	2.17	8.70	4.35	8.70	7.61	2.17	0.00	3.26	0.00	3.26	10.87	6.52	1.09	0.00	2.17
Euphorbia esula	2.14	0.36	6.73	3.10	4.46	0.89	1.96	10.18	3.27	2.44	7.20	13.45	4.76	1.79	2.74	2.56	7.98	3.10	15.30	1.90	0.65	2.20	0.83
Fragaria vesca	4.95	0.12	7.55	1.98	3.22	1.11	6.56	9.03	7.80	4.33	1.86	12.13	4.58	1.11	0.00	1.49	4.70	7.55	13.86	3.34	0.62	0.37	1.73
Glycine max	3.63	0.53	6.73	5.90	4.61	2.57	5.90	6.20	1.51	4.61	5.30	9.61	4.77	1.44	0.98	1.66	6.96	5.45	10.36	5.22	1.66	2.19	2.19
Gossypium hirsutum	4.21	0.19	3.37	3.74	4.86	1.87	6.55	6.17	2.81	6.74	6.17	10.76	3.93	1.59	0.47	4.30	5.71	5.24	11.97	2.71	3.27	1.59	1.78
Helianthus annuus	2.68	0.19	3.87	5.36	3.55	1.93	5.17	6.42	5.55	6.48	8.60	7.11	2.93	2.74	1.75	3.62	3.37	7.04	7.61	5.42	3.24	2.18	3.18
Hevea brasiliensis	4.24	0.37	6.27	3.51	3.69	3.14	3.69	7.93	1.66	3.14	7.38	11.99	6.83	0.55	2.03	1.66	6.83	5.35	13.10	1.48	1.29	3.32	0.55
Liriodendron tulipifera	8.50	0.00	8.50	1.79	3.58	1.12	11.19	4.70	0.89	3.36	4.70	12.08	10.74	1.79	0.00	1.57	7.83	1.12	10.96	2.46	1.34	0.67	1.12
Lotus japonicus	3.03	0.07	5.85	3.85	2.59	1.33	5.63	6.29	2.89	6.44	3.11	12.81	2.96	0.44	0.15	1.11	7.25	9.10	14.36	7.18	0.74	0.30	2.52
Malus domestica	6.00	0.30	8.07	2.46	3.44	2.76	6.69	7.58	4.43	4.33	3.54	12.60	4.82	0.59	0.30	1.48	3.25	8.37	11.22	4.04	1.38	0.59	1.77
Manihot esculenta	5.41	0.11	6.41	3.43	3.31	0.99	4.53	7.62	3.31	3.09	6.74	14.03	5.08	1.44	1.22	2.54	7.51	4.75	12.82	1.88	0.99	1.22	1.55
Medicago truncatula	2.02	0.38	4.49	5.12	3.16	1.52	5.43	7.14	2.27	4.55	5.62	11.37	4.80	2.40	1.83	1.26	10.17	2.84	12.32	4.04	1.20	2.72	3.35
Nicotiana tabacum	3.38	0.48	6.77	3.48	1.64	1.26	5.90	8.70	2.51	1.84	3.58	16.05	5.71	1.16	1.45	1.16	8.22	3.58	12.86	3.38	1.45	2.32	3.09
Ocimum basilicum	11.50	0.44	8.85	1.77	2.65	2.65	3.54	4.42	9.73	2.65	9.29	5.75	2.21	0.88	0.00	1.77	2.21	10.18	9.73	2.65	1.77	2.65	2.65
Panax ginseng	6.51	0.26	9.38	2.34	3.39	2.08	4.95	6.51	3.65	3.65	6.77	10.94	3.91	1.30	0.52	1.82	3.91	7.55	11.20	2.60	1.30	2.60	2.86
Pisum sativum	3.38	0.00	3.38	6.33	4.64	1.69	8.02	8.86	2.53	2.53	7.59	9.28	2.11	3.80	1.27	3.80	3.80	2.53	8.02	5.91	2.11	3.80	4.64
Prunus persica	6.94	0.12	7.64	3.24	3.13	1.97	7.87	5.56	2.31	4.75	3.24	14.58	4.98	1.74	0.69	3.01	4.63	5.44	10.88	3.13	1.50	1.16	1.50
Raphanus sativus	3.08	0.28	8.23	3.04	4.67	1.80	4.59	10.19	4.71	3.76	3.96	11.55	4.79	1.60	0.52	2.84	4.04	4.75	13.46	3.92	1.24	0.52	2.48
Ricinus communis	5.47	0.63	4.84	3.49	2.14	2.93	7.38	6.50	4.28	4.28	4.92	13.88	4.44	1.67	2.78	2.06	5.87	4.36	10.86	3.09	1.82	1.59	0.71
Solanum lycopersicum	3.64	0.21	7.69	4.37	1.77	1.56	6.55	9.25	3.64	2.18	5.72	9.36	6.86	1.35	1.87	3.33	6.34	3.22	12.06	1.98	3.12	2.18	1.77
Theobroma cacao	2.98	0.17	7.27	5.45	1.82	2.31	3.80	7.93	3.64	2.98	6.28	12.07	5.12	1.32	1.65	1.98	7.27	4.96	14.55	2.81	0.17	1.98	1.49
Trifolium pratense	1.21	0.24	3.62	3.74	1.21	2.41	4.47	3.26	11.04	9.54	4.16	6.52	1.51	1.75	0.54	1.69	3.56	7.54	13.70	8.57	5.49	0.91	3.32
Vitis vinifera	6.22	0.00	7.60	2.90	3.59	2.21	5.66	7.04	3.87	3.18	4.83	10.77	4.01	1.38	1.24	3.18	5.94	4.97	12.29	3.18	2.07	1.24	2.62
Avena barbata	15.38	0.71	15.77	1.26	2.05	2.13	5.68	3.79	8.60	1.66	1.18	7.81	1.89	0.55	0.32	0.95	0.87	12.78	9.78	2.21	0.95	1.66	2.05
Avena sativa	15.36	0.34	7.51	0.00	2.39	4.10	7.17	5.12	10.92	1.71	1.71	8.19	3.07	0.00	0.34	1.71	3.07	9.56	8.19	2.39	1.71	1.02	4.44
Cenchrus ciliaris	21.97	0.16	19.51	0.49	1.64	2.62	4.43	2.79	12.30	1.15	0.66	6.89	1.15	0.33	0.00	0.16	0.66	13.44	6.89	1.97	0.66	0.16	0.00
Hordeum vulgare	16.21	0.39	16.29	0.70	2.65	2.26	5.92	4.29	9.20	2.18	1.40	6.24	2.96	1.01	0.23	1.09	1.25	12.16	8.57	1.71	0.94	0.47	1.87
Musa acuminata	6.17	0.19	12.71	0.56	5.61	1.68	6.36	10.09	6.92	2.62	1.31	14.02	2.24	0.56	0.56	1.12	3.93	7.66	10.28	2.06	1.50	0.75	1.12
Oryza sativa	16.09	0.13	23.89	0.37	1.90	1.04	2.38	3.56	10.36	1.54	0.72	6.86	0.88	0.26	0.06	0.36	0.81	16.16	7.68	2.02	0.68	0.55	1.70
Secale cereale	12.82	0.57	20.23	0.28	3.42	0.85	3.42	5.70	11.11	2.56	0.85	8.55	1.99	0.57	0.85	1.14	0.85	10.54	6.84	2.85	0.85	0.85	2.28
Sorghum bicolor	18.35	0.28	20.30	0.37	3.43	1.48	5.28	2.59	9.73	1.67	0.65	6.39	1.48	0.46	0.28	0.93	1.11	12.51	6.86	2.13	0.65	0.56	2.50
Sorghum propinquum	22.65	0.34	18.96	0.17	3.02	1.34	3.69	3.52	8.72	1.51	1.01	6.04	1.34	0.00	0.17	0.50	1.01	11.74	6.88	1.17	1.68	1.34	3.19
Triticum aestivum	16.32	0.66	15.66	0.25	1.57	2.47	3.54	5.52	9.73	2.39	1.48	8.66	1.48	0.33	0.33	0.58	1.15	12.20	9.48	1.81	1.24	0.49	2.64
Zea mays	20.97	0.14	19.78	0.56	2.38	1.82	4.00	2.66	9.75	2.24	0.84	6.73	1.12	0.14	0.14	0.35	0.77	11.43	8.13	2.45	0.63	0.56	2.38