BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 7(5)

Mining functional microsatellites in legume unigenes

Manish Roorkiwal[1,2] & Prakash Chand Sharma[1]*

[1]University School of Biotechnology, Guru Gobind Singh Indraprastha University, Dwarka, Sector 16C, New Delhi-110075, India. [2]International Crops Research Institute for the Semi-Arid Tropics, Patancheru-502324, India. Prakash Chand Sharma - Phone: +91-11-25302306 (Direct), +91-11-25302303 (Office), Fax: +91-11-25302305, Email: prof.pcsharma@gmail.com; *Corresponding author

Received October 12, 2011; Accepted October 21, 2011; Published October 31, 2011

Abstract
Highly polymorphic and transferable microsatellites (SSRs) are important for comparative genomics, genome analysis and phylogenetic studies. Development of novel species-specific microsatellite markers remains a costly and labor-intensive project. Therefore, interest has been shifted from genomic to genic markers owing to their high inter-species transferability as they are developed from conserved coding regions of the genome. This study concentrates on comparative analysis of genic microsatellites in nine important legume (*Arachis hypogaea*, *Cajanus cajan*, *Cicer arietinum*, *Glycine max*, *Lotus japonicus*, *Medicago truncatula*, *Phaseolus vulgaris*, *Pisum sativum* and *Vigna unguiculata*) and two model plant species (*Oryza sativa* and *Arabidopsis thaliana*). Screening of a total of 228090 putative unique sequences spanning 219610522 bp using a microsatellite search tool, MISA, identified 12.18% of the unigenes containing 36248 microsatellite motifs excluding mononucleotide repeats. Frequency of legume unigene-derived SSRs was one SSR in every 6.0 kb of analyzed sequences. The trinucleotide repeats were predominant in all the unigenes with the exception of *C. cajan*, which showed prevalence of dinucleotide repeats over trinucleotide repeats. Dinucleotide repeats along with trinucleotides counted for more than 90% of the total microsatellites. Among dinucleotide and trinucleotide repeats, AG and AAG motifs, respectively, were the most frequent. Microsatellite positive chickpea unigenes were assigned Gene Ontology (GO) terms to identify the possible role of unigenes in various molecular and biological functions. These unigene based microsatellite markers will prove valuable for recording allelic variance across germplasm collections, gene tagging and searching for putative candidate genes.

Keywords: Microsatellites, SSRs, Unigenes, Legumes, Functional annotation

Background:
Comparative genomics is a proven and established tool for genome analysis, annotation and evolutionary studies **[1].** Coding regions, in particular, can be exploited for developing DNA markers, already proved very useful in comparative studies. Microsatellites or Simple Sequence Repeats (SSRs) are ubiquitous in eukaryotic genomes, with non-random distribution in the genomic regions. Microsatellites provide a rich source of hypervariable co-dominant markers owing to their high mutation rates that generate allelic variation in array length **[2].** Microsatellites have been implicated in genome evolution, gene regulation or functional evolution of the genes **[3].** Microsatellites are important tools for comparative mapping because of their high polymorphism and transferability. Genic Microsatellite Markers (GMMs) have been extensively used in different areas such as genome characterization, genome mapping, comparative genomics, phylogenetic studies, population genetics and molecular breeding **[4].**

In general, microsatellites are identified from both non-coding and coding regions of the genome. Standard methods for the development of microsatellite markers require considerable amount of time, money and labour **[5].** Moreover, microsatellites developed by these standard methods show an

 © 2011 Biomedical Informatics

element of biasness depending upon the method or probe used for their development. Recently, researchers have shifted their attention from genomic markers to genic markers that represent coding sequences or transcriptome **[4].** The advancements in the field of genomics have resulted in the accumulation of huge amount of sequence data in the public domain including vast collection of expressed sequence tags (ESTs) and unigenes. This huge sequence data has provided an alternative approach for the identification and development of molecular markers. ESTs have provided a potentially rich source of GMMs **[4].** GMMs are widely favoured as molecular markers owing to their inexpensive development, representation of transcribed genes/coding regions and a putative function can often be deduced by a homology search. Development of EST-derived microsatellite markers, however, suffers with a limitation of high redundancy prevailing in EST sequences yielding multiple markers at the same locus. To overcome this limitation of ESTs, unigenes are derived by clustering ESTs into singletons and contigs. Microsatellite markers developed from these unigenes can be used to detect variation in the functional genome with unique identity and position **[4].** Parida and co-workers **[6]** identified and characterized microsatellite motifs in the unigenes available in five cereal crops (rice, wheat, maize, sorghum, barley) and *Arabidopsis*. These unigene derived microsatellite (UGMS) markers have high inter-specific transferability as they are developed from conserved coding regions of the genome. Moreover, they also serve as a potential tool to study functional diversity and genome evolution patterns more accurately. Therefore, these unigene derived microsatellite markers would be of great use for comparative mapping and phylogenetic analysis and to facilitate development of syntenic networks for understanding the evolution of genes and genomes.

Fabaceae (earlier included in Leguminosae) is the third largest and economically important family of flowering plants. Members of Fabaceae include *Arachis hypogaea*, *Cajanus cajan*, *Cicer arietinum*, *Glycine max*, *Pisum sativum* and many other important legumes. These crops serve as a source of staple, essential food for supplementing dietary proteins for vegetarian people. Recent progress towards accumulation of various genomic resources (ESTs, unigenes) for legumes have facilitated comparative mapping in these plant species. Although a lot has been achieved towards development of genic microsatellite markers in plants, yet only few studies have been undertaken to develop such resources in case of legumes. This study concentrates on identifying genic microsatellite repeats in legumes and comparative analysis of genic microsatellites in legumes, which can further be used as a valuable tool for future studies in legumes related to genome evolution, gene tagging and genetic diversity.

Methodology:

Sequence resources

Unigene collections of 11 plant species, including 9 legumes namely *Arachis hypogaea* (*Aha*), *Cajanus cajan* (*Cca*), *Cicer arietinum* (*Car*), *Glycine max* (*Gma*), *Lotus japonicus* (*Lja*), *Medicago truncatula* (*Mtr*), *Phaseolus vulgaris* (*Pvu*), *Pisum sativum* (*Psa*) and *Vigna unguiculata* (*Vun*), and two model plants, *Oryza sativa* (*Osa*) and *Arabidopsis thaliana* (*Ath*) were subjected to *in silico* mining of microsatellites **(Supplementary table 1).** Unigene sequences for all these species were downloaded from NCBI Unigene database (ftp://ftp.ncbi.nih.gov/repository/UniGene/) except for *C. cajan*, *C. arietinum* and *P. sativum*, as unigene data for these legumes were not available in the public domain. For these species, large numbers of ESTs are available in NCBI dbEST (http://www.ncbi.nlm.nih.gov/dbEST/). To nullify redundancy prevalent in ESTs, we used the sequence assembly program CAP3 **[7]** to cluster ESTs into contigs/singletons and generated non-redundant unigenes for each of these legume species. The non-redundant unigene sequences were used to identify microsatellites and perform gene ontology annotation.

Microsatellite mining and variability prediction

A perl script, MISA (MIcroSAtelitte) was used to identify microsatellites in all these unigene sequences (http://pgrc.ipk-gatersleben.de/misa/) **[8].** A simple sequence repeat with repeat motif length varying between 1 and 6 bp was identified as a microsatellite. Mononucleotide repeats were excluded from the analysis because of the abundance of poly A/T repeats mostly resulting from sequencing artifacts and poly A tails. Repeat-motifs like AG, GA, TC and CT were considered in the same class considering complementary sequences and/or different reading frames. Compound microsatellites were considered with at least two different repeat-motifs without any interruption. The analysis of mined microsatellites was done on the basis of their motif length (di- to hexa-nucleotide), number and type of repeats, relative frequency of occurrence and length as class I (≥20 nucleotides) and class II (12 to 20 nucleotides) types **[9].** GC content was also calculated. Trinucleotide repeats were examined for the possible encoded amino acid motif and codon biasness.

Assessment of functional relevance of unigenes having SSRs

Unigene sequences containing microsatellites were used for similarity search using Blast2GO **[10]** to identify their putative function. Unigene sequences not showing any match were considered as unique to that particular species. The *C. arietinum* microsatellite positive unigenes were run through a Gene Ontology (GO) assignment database in order to assess associations between SSR loci and biological processes, cellular components and molecular function of known genes.

Discussion:

The availability of large unigene collections for some legumes in public domain allowed us to explore these resources for the presence and functional relevance of different microsatellite repeats. The unigenes being longer and without redundancy offer advantages over the EST sequences for the development of microsatellite markers. However, development of microsatellites for the species with no sequence information is an expensive and time-consuming task. To overcome this limitation, microsatellite markers developed in closely related species can be utilized **[4].** EST-SSRs representing the coding regions of the genome, are expected to be conserved with a high rate of cross species transferability in comparison to genome derived SSRs **[11].** Success of EST derived SSR markers across diverse taxonomic groups has been reported **[12].**

© 2011 Biomedical Informatics

BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 7(5)

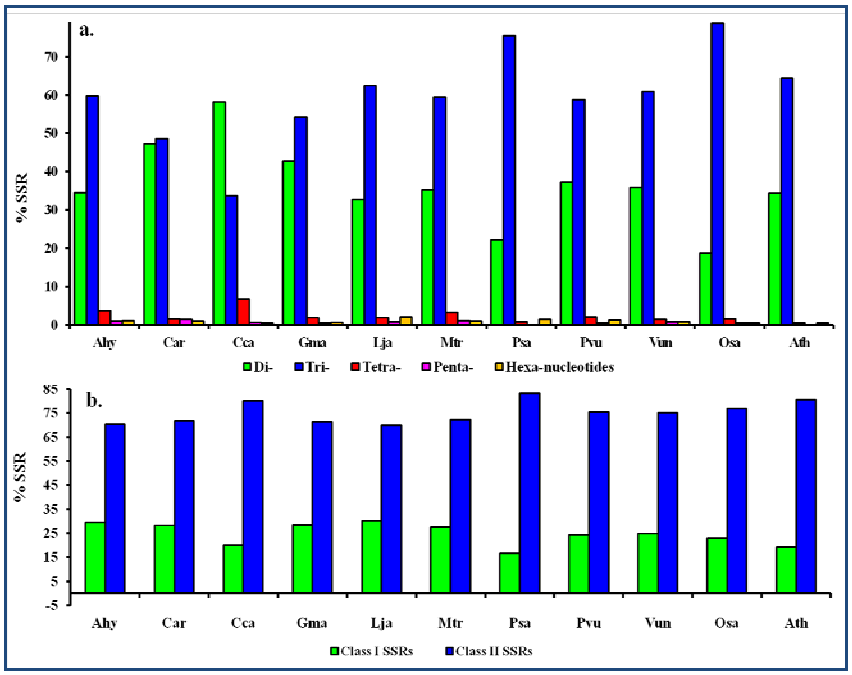


Figure 1a: Distribution of microsatellite repeats in various legumes, rice and Arabidopsis. **b)** Distribution of Class I and Class II microsatellite repeats

A total of 228090 putative unique sequences were screened for the presence of microsatellites, of which 12.18% (27791) contained specified repeat motifs excluding mononucleotide repeats, yielding 36248 unique SSRs. In legumes, a total of 156013 unique sequences were used for microsatellite search, of which only 6.85% (10688) contained microsatellites representing 12220 unique SSRs **(Table 1 see supplementary material).** This is a relatively higher abundance of SSRs for plant unique coding sequences, compared to the previous reports for some cereals **[13]** and wild *Arachis* species **[14].** The variable abundance of SSRs is known to be dependent on the SSR search criteria, the size of the dataset, the database-mining tools and the species concerned **[4].** The frequency of occurrence for unigene-derived SSRs was one SSR in every 6.0 kb. In previous reports, this frequency ranged from 3.4 kb in rice to 20.0 kb in cotton **[15].**

In earlier reports, trinucleotide repeats generally formed the most common motif in various plant species **[15],** regardless of the EST-SSR search criteria. However, abundance of dinucleotide repeats has also been reported in many of the dicot species **[16].** We also found trinucleotide repeats to be the most abundant followed by dinucleotides with the sole exception of *C. cajan*, where the situation was reversed as dinucleotide repeats were more abundant followed by trinucleotide repeats (**Figure 1a**). Dinucleotide and trinucleotide repeats together counted for more than 90% of all the microsatellites (**Figure 1a**). Tetranucleotide (~2.0%), pentanucleotide (<1.0%) and hexanucleotide (~1.0%) repeats showed very low abundance among all the species. In terms of single SSR motif, the dinucleotide motif AG/CT was most frequent **[9, 13, 17].** The two most dominant motif types recorded in our search were AG and AT in agreement with a study on cultivated peanut and wild *Arachis* species **[14].** Low abundance of "CG" repeats may be attributed to their tendency of forming secondary structures (hairpins), leading to a selective pressure against 'CG' accumulation in genomes. Microsatellites were also classified into two classes on length basis. Firstly, Class I microsatellites, which include microsatellites more than equal to 20 nucleotides in length, and *secondly*, Class II microsatellites including microsatellites of less than 20 nucleotides. Class II microsatellites are more abundant (>70%) among all the species **(Figure 1b).**

Among trinucleotide repeat motifs, AAG motif was the most abundant, which is the second most abundant motif in *Arabidopsis* **[6].** In other plant species, the most frequent trinucleotide repeat motifs were AAC/TTG in wheat, AAG/TTC in soybean, and CCG/GGC in barley, rice, maize and sorghum **[13, 8, 18].** The previous studies on *Arabidopsis* and soybean [15] also reported abundance of trinucleotide motif AAG, contrasting to the abundance of CCG motif in cereal species **[6].** The trinucleotide repeats code for 21 amino acids

 © 2011 Biomedical Informatics

and stop codon. The predicted amino acid pattern for the trinucleotide motifs detected is shown in **supplementary figure 1**. CTA/CTC/CTG/CTT/TTA/TTG motifs coding for leucine were most common followed by AGC/AGT/TCA/TCC/TCG/TCT coding for serine and glutamic acid (GAA/GAG). Abundance of small/hydrophilic amino acid repeat motifs like that of serine in the unigenes of cereals and *Arabidopsis* is explainable since these repeats are tolerated in many proteins, while strong selection pressure possibly eliminates codon repeats encoding for hydrophobic/other amino acids **[19].** Trinucleotide repeats try to maintain codon biasness and thus vary their frequency significantly to manage frameshift mutations in coding regions **[20].**

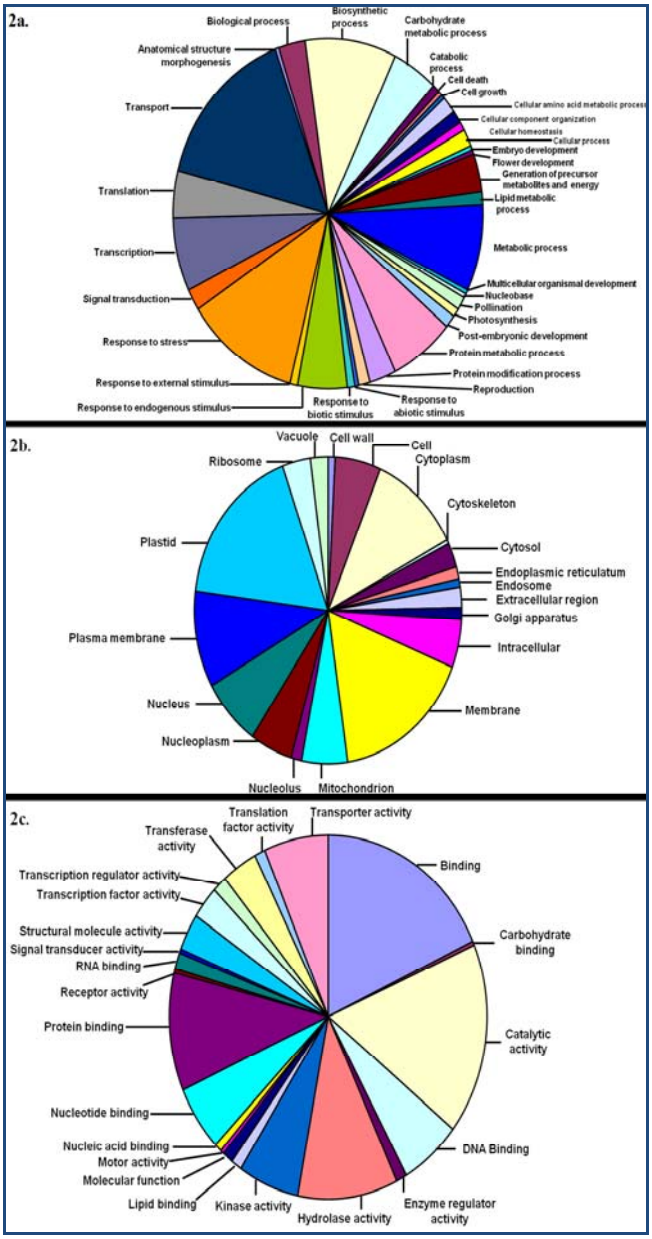


Figure 2 (a-c): Gene Ontology annotation of SSR positive *Cicer arietinum* unigenes. **a)** Biological process; **b)** Cellular component; and **c)** Molecular function

In silico identification of SSRs from various sequence resources like genomic sequences, ESTs or unigenes is a low cost and easy method for development of microsatellite markers. Such markers can be used for understanding the nature and possible biological functions. EST-SSRs have been of great interest to researchers and there are many recent reports about development of EST-SSR markers, in plant materials such as soybean **[21],** potato **[22],** seabuckthorn **[23]** and many more.

To characterize unigene sequences harboring SSRs, we performed sequence similarity search against non redundant NCBI protein database. On an average, more than 70% of unigenes with SSRs showed homology to genes having known function for each species under study, with an exception of *C. cajan*, which had only 30% of such unigenes. The remaining unigenes showing no hit during similarity search were considered as organism specific. Most of the unigene sequences represented enzymes of general metabolism as reported earlier **[17].** On the basis of GO annotation, microsatellite positive chickpea unigenes were assigned GO terms associated with biological process, cellular component and molecular function**.** In case of biological processes, *C. arietinum* unigenes were assigned to thirty three different categories **(Figure 2a).** Majority of unigenes were assigned to the "transport" category (15.5%). For the cellular components, unigenes were assigned to nineteen different categories with majority of them participated in "plastid" category (17.45%) **(Figure 2b).** When concentrating on molecular functions, the unigenes were assigned to twenty two categories with majority covering "binding" category (18.18%) **(Figure 2c).** In general, microsatellite containing *C. arietinum* unigene sequences matched to proteins having distinct molecular functions such as, binding, catalytic, transport, enzyme regulators, and structural activities in different biological processes, and cellular and sub-cellular organization. Unigenes related to biological process such as response to abiotic and biotic stresses should be explored as candidates for studying their role in response to that particular stress or trait. One of the favorable approaches to use them could be to assign marker trait association study based on the phenotypic data and allele variance across diverse collections.

Conclusion:

The present study has focused on the *in silico* mining/identification of microsatellites from the unique coding sequences of nine members of Fabaceae family and two model plant species. Microsatellite markers developed from conserved coding region of genome show a higher transferability through cross-amplifications in related species than microsatellites developed using genomic regions. Development of microsatellite markers from coding region using computational approach has reduced the cost significantly and allowed their use for related species with less sequence information. Microsatellite dynamics with regard to frequency and types of microsatellites showed marked variability in the legume unigenes. The trinucleotide repeats were predominant in all the unigenes analysed except in *C. cajan*. Unigene sequences are derived from the expressed portion of the genome, therefore, markers developed from these resources can be assayed as gene based functional marker for diversity assessment, and gene mapping and marker assisted selection. To characterize unigene sequences with SSRs, we performed sequence similarity search

© 2011 Biomedical Informatics

against non-redundant NCBI protein database. Unigene derived markers may be implicated in biological, cellular and molecular functions and provide opportunity to investigate the possible role of microsatellites in various gene functions.

Acknowledgement:
Authors are thankful to GGSIP University, New Delhi, India for providing processing charges for this article. Manish Roorkiwal acknowledges research fellowship from University Grants Commission (UGC).

References:

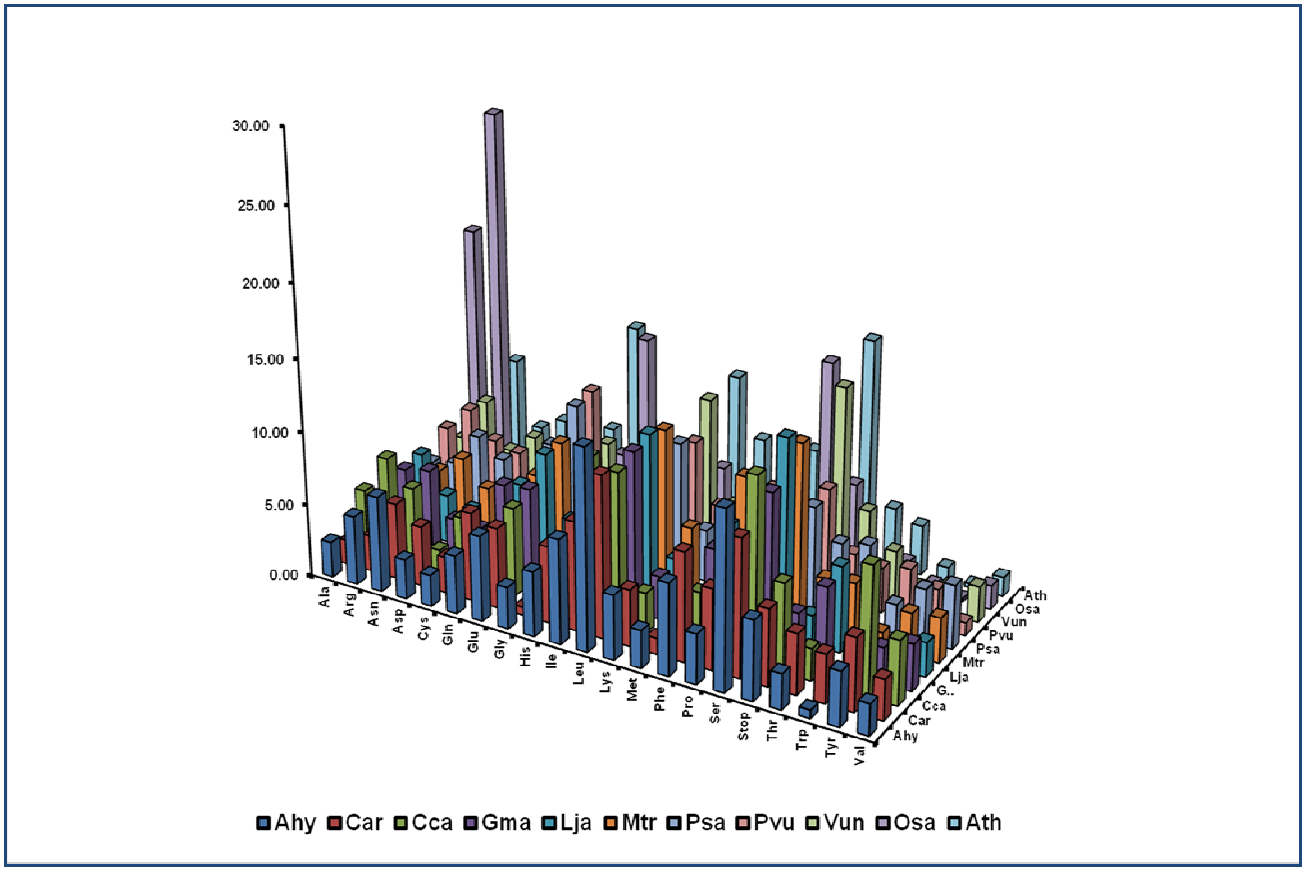
[1] Li X *et al. BMC Evol. Biol*. 2010 **10**: 190 [PMID: 20565927]
[2] Powell W *et al. Trends Plant Sci*. 1996 **1**: 215
[3] Grover A & Sharma PC, *Curr. Sci*. 2011 **100**: 6
[4] Varshney RK *et al. Trends Biotechnol*. 2005 **23**: 48 [PMID: 15629858]
[5] Squirrell J *et al. Mol. Ecol*. 2003 **12**: 1339 [PMID: 12755865]
[6] Parida SK *et al. Theor. Appl. Genet*. 2006 **112**: 808 [PMID: 16429310]
[7] Huang X & Madan A. *Genome Res*. 1999 **9**: 868 [PMID: 10508846]
[8] Thiel T *et al. Theor. Appl. Genet*. 2003 **106**: 411 [PMID: 12589540]
[9] Temnykh S *et al. Genome Res*. 2001 **11**: 1441 [PMID: 11483586]
[10] Conesa A *et al. Bioinformatics* 2005 **21**: 3674 [PMID: 16081474]
[11] Moccia MD *et al. BMC Genomics* 2009 **10**: 243 [PMID: 19467153]
[12] Vendramin E *et al. Mol. Ecol. Notes* 2007 **7**: 307
[13] Kantety RV *et al. Plant Mol. Biol*. 2002 **48**: 501 [PMID: 11999831]
[14] Proite K *et al. BMC Plant Biol*. 2007 **7**: 7 [PMID: 17302987]
[15] Cardle L *et al. Genetics* 2000 **156**: 847 [PMID: 11014830]
[16] Kumpatla SP & Mukhopadhyay S. *Genome* 2005 **48**:985 [PMID: 16391668]
[17] Newcomb RD *et al. Plant Physiol*. 2006 **141**: 147 [PMID: 9069178]
[18] Roorkiwal M *et al. Mol. Genet. Genomics* 2009 **282**: 205 [PMID: 19484264]
[19] Katti MV *et al. Mol. Biol. Evol*. 2001 **18**: 1161 [PMID: 11420357]
[20] Metzgar D *et al. Genome Res*. 2000 **10**: 72 [PMID: 10645952]
[21] Hisano H *et al. DNA Res*. 2007 **14**: 271 [PMID: 18192281]
[22] Grover A *et al. Physiol. Mol. Biol. Plants* 2009 **15**: 343
[23] Jain A *et al. Physiol. Mol. Biol. Plants* 2010 **16**: 375

Edited by P Kangueane
Citation: Roorkiwal & Sharma. Bioinformation 7(5): 264-270 (2011)
License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited.

BIOINFORMATION

open access

Supplementary material:



Supplementary Figure 1: Distribution of amino acids encoded by trinucleotide repeats in various species.

Table 1: Overall occurrence and distribution of microsatellites in the unigene sequences of legumes, rice and Arabidopsis

	Ahy	*Car*	*Cca*	*Gma*	*Lja*	*Mtr*	*Psa*	*Pvu*	*Vun*	*Osa*	*Ath*
Total number of Unigenes examined	11909	16062	17760	36295	22342	18790	10393	6686	15776	41407	30670
Total size of examined unigenes (bp)	7900811	9195242	9617817	29057670	18019715	14493627	5698163	5298852	10923442	65609946	43795237
Total number of identified SSRs	11128	15186	17580	18656	11912	9192	2992	1465	8580	38046	15656
Number of SSR positive unigenes	8635	7762	7805	14151	9817	7609	2754	1192	7257	20396	11415
Number of unigenes (Containing >1 SSR)	1860	2319	2689	3505	1712	1306	215	170	1125	8563	3147
Number of compound SSRs	76	284	269	99	29	26	3	9	42	533	76
Number of perfect class I microsatellites	345	209	108	1123	763	424	56	94	265	4239	1089
Number of perfect class II microsatellites	823	533	430	2803	1772	1111	278	290	804	14167	4533

© 2011 Biomedical Informatics

Average size of unigene (bp)	663	572	542	801	807	771	548	793	692	1585	1428
SSR frequency (No of SSR per unigene)	0.93	0.95	0.99	0.51	0.53	0.49	0.29	0.22	0.54	0.92	0.51
Total number of identified SSRs (except mononucleotides)	1168	742	527	3926	2535	1535	334	384	1069	18406	5622
Unigenes containing SSRs (except mononucleotides)	1017	685	486	3375	2145	1369	302	340	969	12422	4681

Supplementary Table 1: Unigene database size, average length and GC content available in public domain

Species	Unigene count	Size (bp)	Average bp count per unigene	GC content (%)
Arachis hypogaea	11909	7900811	663	39.4
Cicer arietinum	16062	9195242	572	39.9
Cajanus cajan	17760	9617817	542	44.2
Glycine max	36295	29057670	801	41.5
Lotus japonicus	22342	18019715	807	42.4
Medicago truncatula	18790	14493627	771	39.3
Pisum sativum	10393	5698163	548	39.9
Phaseolus vulgaris	6686	5298852	793	42.9
Vigna unguiculata	15776	10923442	692	42.0
Oryza sativa	41407	65609946	1585	50.2
Arabidopsis thaliana	30670	43795237	1428	42.1

© 2011 Biomedical Informatics