BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 7(7)

Microsatellites in palm (*Arecaceae*) sequences

Manju Kalathil Palliyarakkal[1], Manimekalai Ramaswamy[2]*, Arunachalam Vadivel[3]

[1]Senior Research Fellow, DIT- Agribioinformatics Promotion centre, Central Plantation Crops Research Institute, P. O. Kudlu, Kasaragod-671124, Kerala, India; [2]Senior Scientist, Molecular biology and Biotechnology, Crop Improvement Division, Central Plantation Crops Research Institute, P. O. Kudlu, Kasaragod-671124, Kerala, India; [3]Principal Scientist, Horticulture Section, ICAR Research Complex for Goa, Old Goa 403 402, India; Manimekalai Ramaswamy - E-mail: rmanimekalai@rediffmail.com; *Corresponding author

Received November 16, 2011; Accepted November 16, 2011; Published December 10, 2011

Abstract:
Microsatellites are the most promising co-dominant markers, widely distributed throughout the genome. Identification of these repeating genomic subsets is a tedious and iterative process making computational approaches highly useful for solving this biological problem. Here 38,083 microsatellites were localized in palm sequences. A total of 2, 97,023 sequences retrieved from public domains were used for this study. The sequences were unstained using the tool Seqclean and consequently clustered using CAP3. SSRs are located in the sequences using the microsatellite search tool, MISA. Repeats were detected in 33,309 sequences and more than one SSR had appeared in 3,943 sequences. In the present study, dinucleotide repeats (49%) were found to be more abundant followed by mononucleotide (30%) and trinucleotide (19%). Also among the dinucleotides, AG/GA/TC/CT motifs (55.8%) are predominantly repeating within the palm sequences. Thus in future this study will lead to the development of specific algorithm for mining SSRs exclusively for palms.

Keywords: Seqclean, SSR, MISA, repeats, Arecacea

Background:
Genome, the genetic blue print of the hereditary information of an organism comprises many functional regions, non-coding sections and the vast unexplored areas of the nucleic acids. Determining the nucleotide sequences of the genome has many technical hurdles which could be accomplished by the hand of computational approaches only. Unlocking the knowledge encoded in the genomes can increase our understanding of species, which could lead to substantially increase the productivity of crops. With such enhanced productivity, it could be the sustainable solution for fulfilling the world's demand for the food. Molecular markers are used to provide a link between genotype and phenotype, for the generation of molecular genetic maps and to assess the genetic diversity within and between related species. Molecular markers will provide the relationship between an inherited disease and its genetic cause, and can be used to determine the precise inheritance pattern of the gene that has not yet been exactly localized. Simple Sequence Repeats (SSRs) or microsatellites are becoming the most important molecular markers in both animals and plants, incorporates stretches of tandemly repeated short oligonucleotides **[1]** whose functional and/or structural characteristics distinguish them amidst the general DNA sequence.

Since SSR markers are highly informative, they are widely used for genetics and breeding studies in several plant species. Thus characterization of microsatellites is extremely important for the development of molecular markers. SSRs are highly abundant and exhibit broad levels of polymorphism in eukaryotic **[2, 3]** and prokaryotic **[4, 5]** genomes. Their variability in length is caused by slip-strand mutations and that may affect the local structure of DNA molecule or encoded protein **[6].** SSRs often serves to modify the genes with which they are associated and their influence on gene regulation, transcription and protein function typically depends on the number of repeats **[7].** Thus

SSRs provide a prolific source of quantitative and qualitative variation. Transferability between species and sometimes between genera has often been reported **[8].** The ubiquitous properties of microsatellites made them important targets for genetic diversity studies, genetic and comparative mapping **[9, 10]**. Moreover, the polymorphism revealed by EST–SSR was similar to that exposed by genomic SSR **[11].** Also the SSRs are cross transferable across species and the flanking regions of SSRs are found conserved **[12]**. A high level of polymorphic loci was obtained regardless of the species considered.

Given the interest of the plant genetics community in SSRs as genetic markers, there has been a particular concern in the establishment of methods for rapid identification of robust and informative SSRs linked to genes of agronomic significance. Compared to genome-wide isolation approaches, gene-targeted strategies are more likely to yield SSRs that are relevant to the goals of marker-assisted selection and germplasm assessment **[13].**

The distribution of microsatellites in the palm sequences has practical implications with regard to their use as molecular markers. Cross transferability or cross species amplification of coconut microsatellite markers in rattan (climbing palms) **[14]** and other economic palms (date palm, oil palm, palmyrah and arecanut) **[15]** is already reported. Thus the incorporation of SSR markers in the palm research are extremely valuable and are increasingly becoming popular in comparative genomics where SSR markers developed from one species could be utilized in a related species towards genetic mapping, characterization, gene cloning, diversity and evolutionary studies **[16].** This approach gained momentum in plant genomics during the recent years based on the observation that despite a wide range in genome sizes, plants were found to exhibit extensive conservation of both gene content and gene order **[17].** The availability of a very large set of microsatellite markers distributed throughout the genome in oil palm and date palm would facilitate the development of high resolution maps, which are instrumental for applications like positional gene cloning and detailed comparative mapping to other palms like coconut, arecanut, palmyrah palm etc.

Sequences from many genes and genomes are continuously made freely available in the public databases and mining of these sources using computational approaches permits rapid and economical marker development. But the number of robust, informative and user-friendly markers (e.g., SSRs) publicly available for palms is still insufficient for some applications, particularly considering the low intra-specific polymorphism level observed even with microsatellite markers (10-20%). The main objective of the present work is to locate and analyze the SSRs in the sequences of palms.

Methodology:

Sequence Sources

The substrate sequences, on which we have acted upon, except date palm sequences, were retrieved from the web base National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/) **[18].** The draft date palm (Phoenix dactylifera 'Khalas' variety) genome investigated in the study has been obtained from the site of Weill Cornell Medical College in Qatar (www.qatar-weill.cornell.edu) **[19]**. The predicted size of the date palm genome sequence is ~650Mbp. The sequences of major palms like coconut *(Cocos nucifera),* arecanut *(Areca catechu)*, oil palm *(Elaeis)* and date palm *(Phoenix dactylifera)* were processed individually while rest of the arecaceae members have been underwent treatment in mass.

Sequence Preprocessing

Sequences were then pre-processed to minimize the sequencing errors and to avoid redundant sequences, were grouped into contigs. The sequence flaws were obliterated using the standalone tool SeqClean (Gene Index project) **[20]**, for the trimming of poly A (T) tails, low quality/complexity regions and distal oligoN series. Before execution Seqclean program was customized by configuring the vector contamination database UniVec **[21]** with local Blast **[22]** for the removal of vector, linker and adaptor sequences. The filtered sequences were clustered and then assembled by adapting CAP3 **[23]** tool. The tool uses forward-reverse constraints to correct assembly errors and link contigs and uses base quality values in alignment of sequence reads. Automatic clipping of 5' and 3' poor regions of reads and the generation of assembly results in ace file format are remarkable. The contigs and singletons acquired were blended to engender the substrate file for SSR detection.

Microsatellite mining

Subsequently, the identification and localization of perfect as well as compound microsatellites were done using MISA (http://pgrc.ipk-gatersleben.de/misa) tool **[24],** a repeat searching perl script. Among different programs available in the public domain, the MIcroSAtellite (MISA) search module has improved features that are useful for EST quality control and for designing the primer pairs for EST-SSRs in a batch file **[25].**The categorized results of the microsatellite searches are stored in two files, one to hold the localization and type of identified microsatellite(s) in a table wise manner and a statistics file which summarizes different statistics as the frequency of a specific microsatellite type according to the unit size or individual motifs. The analysis of occurrence and frequency of SSRs among the palm species was carried out by exporting the MISA results to Microsoft Excel spreadsheets **[26].** But to advance for further analysis, it has been found onerous for getting sequences of only SSR containing regions from the bulk palm sequence source file. To resolve this obstacle, the MISA perl script has been configured to get an additional SEQUENCE file holding only SSR containing sequences. This has considerably reduced the manual effort and toil.

Results and Discussion:

Microsatellite density profile it was found that palm sequences are rich source of microsatellites repeats. A total of 46,916 sequences were retrieved from the web base National Center for Biotechnology Information, NCBI (http://www.ncbi.nlm.nih.gov/). Since the sequences in public domain are cumulated exponentially, available sequences as on 14th January, 2011 were undergone for the present analysis. Sequences of coconut (390), arecanut (22), oil palm (41374) and date palm (250107) were collected besides 5130 sequences of remaining 546 palm species. These values show the complexity of different palm sequences, which should be handled carefully. A minor negligence while processing itself will lead to false

positives. The reliable softwares and programs were met the needs at this stage and thus the computational approaches greatly help in getting and manipulating these informations. After pre-processing of these sequences respective size of them were significantly reduced and they were subjected to microsatellite mining. From the 390 *Cocos nucifera* sequences of 205Kb size, 426 SSRs per Kb were collected. Similarly 5, 3024, 34051 and 577 SSRs per Kb were detected in *Areca catechu*, *Elaeis*, *Phoenix dactylifera* and other palms respectively. These repeats were occurred at a frequency of 2.04 (coconut), 0.7 (arecanut), 0.22 (oil palm), 3.5 (date palm genome) and 0.62 (other palms) per Kb of sequences **(Table 1, see supplementary material).** Thus altogether, 38,083 SSRs were detected in the entire palm sequences.

Date palm and coconut were observed to be the richest source of microsatellite repeats. Thus SSR markers developed from date palm and coconut were cross transferable to other species of *Arecaceae* which will be helpful for the generation of genic regions in the less available species of other palms. With an earlier report, of microsatellite repeats in oil palm EST sequences (2413 ESTs), the predicted frequency of 1SSR/4.4Kb (0.227SSR per Kb) is found to be in agreement with the present report (41374) in oil palm sequences of 0.22 SSR per Kb. So even if more genic regions of oil palm are revealed on a daily basis, the occurrence of its repeat frequency is found to be constant. This observation discloses the conservative and ubiquitous character of SSRs over palms. In cereals, including barley, maize, oat, rice, rye and wheat, lower frequencies of SSRs (7-10% of total ESTs) were found from their available genome database. The SSR frequency reported in the present study might change when the more sequences of other palms are made available in the public database.

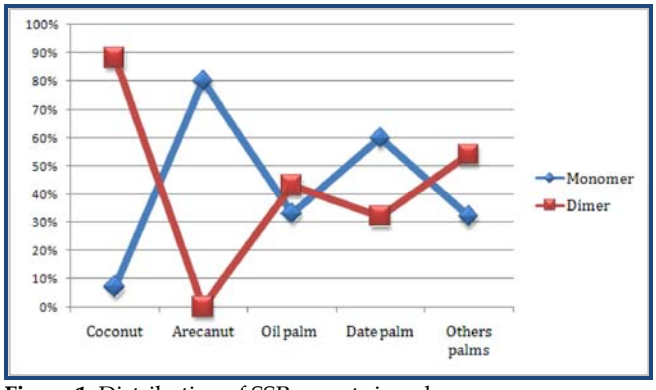

Figure 1: Distribution of SSR repeats in palms

The SSRs were grouped as monomer, dimer, trimer, and tetramer and above based on the size of the repeating unit **(Table 1, see supplementary material).** Mononucleotide repeats were grouped to two as A/T and C/G. Similarly all dinucleotides motifs were grouped into the following four unique classes; (i) AT/TA, (ii) AG/GA/CT/TC, (iii) AC/CA/TG/GT, and (iv) GC /CG. Among all the repeats dinucleotide repeats (49%) were found to be more abundant in palms followed by mononucleotides (30%) and then tri nucleotides (19%) **(Table 2, see supplementary material).** This observation was in exception with arecanut and date palm **(Figure 1),** as these two can be taken concession due to its less sequence availability (arecanut) and non-furnished draft contigs (date palm). Also within dinucleotides, AG/GA/TC/CT motif (55.8%) was observed as dominant repeats.

Conclusion:

By this study, the wide occurrence and conservative nature of microsatellite repeats was accomplished in palms. We have predicted and characterized huge data sets, having all the possible SSRs, their locations, types and classes in palms. The predominant occurrence of GA/CT dimeric and A/T mono type repeats in plants are found to be strengthened in this study. Also the scarcity of AC/GT repeats in plant genomes due to the high frequency of certain amino acids in plants is yet again revealed. The frequency of SSRs in oil palm is found to be preserving even with the addition of more and more updated coding regions. Thus the predicted SSR frequency of different palms can be extended with more upcoming sequence information. This can be targeted towards the genetic mapping, characterization, gene cloning, diversity and evolutionary studies of palms. But, the exploration of genes and genomes will never commit a pause. Magnificent and convoluted amount of data and sequences will be depositing in science each and every day. The development of a specific algorithm which combines the sequence processing and marker extraction will outperform in the near future, which will overwhelm the mining of large data sets. To achieve this target, the currently generated data will be highly useful to design and train algorithm for the extraction of microsatellites exclusively for palms.

Acknowledgments:

This work was supported by grants from Department of Information Technology, India. Our sincere thanks to Dr. George. V. Thomas, Director, Central Plantation Crop Research Institute, Kasaragod, for his broad guidance and support.

References:

[1] Tautz D & Renz M, *Nucleic Acids Res*. 1984 **12**: 4127[PMID: 6328411]
[2] Weber JL. *Genomics* 1990 **7**: 527 [PMID: 1974878]
[3] Katti MV *et al. J Mol Biol Evol*. 2001 **18**: 1161[PMID: 11420357]
[4] Field D & Wills C, *Proc Biol Sci*. 1996 **263**: 209[PMID: 8728984]
[5] Gur-Arie R. *Genome Res*. 2000 **10**: 62[PMID: 10645951]
[6] Mrazek J *et al. J Proc Natl Acad Sci U S A*. 2007 **104:** 8472[PMID: 17485665]
[7] Kashi Y & King DG, Trends in Genetics 2006 **5**: 253[PMID: 16567018]
[8] Dirlewanger E *et al. Theor Appl Genet.* 2002 **105**: 127[PMID: 12582570]
[9] Wu KS & Tanksley SD, *J Mol Gen Genet*. 1993 **241**: 225[PMID: 7901751]
[10] Gonzalo MJ *et al. J Theor Appl Genet*. 2005 **110**: 802[PMID: 15700148]
[11] Saha MC *et al. J Theor Appl Genet*. 2005 **110**: 323[PMID: 15558229]
[12] Poncet V *et al. J Mol Genet Genomics*. 2006 276: 436[PMID: 16924545]
[13] Edenilson R *et al.* Abstract: *Genetics and Molecular Biology*. 2005 **28**: 582
[14] NageswaraRao M *et al. J Silvae Genetica*. 2007 **56**: 282

[15] Anitha N *et al. Indian Journal of Horticulture*. 2008 **3**: 65
[16] Cordeiro GM *et al. J Plant Sci.* 2001 **6**: 1115 [PMID: 11337068]
[17] Bennetzen JL & Freeling M, *Trends Genet.* 1993 **9:** 259 [PMID: 8379002]
[18] National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/)
[19] Weill Cornell Medical College in Qatar (www.qatar-weill.cornell.edu)
[20] Seqclean (http://compbio.dfci.harvard.edu/tgi/software/)[PMID: 17997864]
[21] Univec (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html)[PMID- 11935017]
[22] Altschul SF *et al. J Nucleic Acid Research.* 1997 **17**: 3389[PMID: 9254694]
[23] Huang X & Madan A, *J Genome Res*. 1999 **9**: 868[PMID: 10508846]
[24] Thiel T *et al. Theor Appl Genet*. 2003 **106**: 411[PMID: 12589540]
[25] Riju A & Arunachalam V, *Nature Preceedings*. 2009 **3581**: 1
[26] Varshney RK et *al. Trends Biotecnol.* 2005 **23**: 48[PMID: 15629858]

Edited by P Kangueane
Citation: Palliyarakkal *et al.* Bioinformation 7(7): 347- 351 (2011)

BIOINFORMATION

open access

Supplementary material:

Table 1: Summary of sequences and SSRs in palms:

Quantity Features	Coconut	Arecanut	Oil palm	Date palm	Others palms
No of sequences retrieved	390	22	41374	250107	5130
Size of seq (Kb)	209	11	21322	321278	4396
After polyA trimming size(Kb)	205	7.6	14022	97361	936.3
Total number of SSRs	426	5	3024	34051	577
SSR frequency per Kb	2.04	0.7	0.22	3.5	0.62
No of sequences with SSR	273	4	2453	30160	419
No of sequences with > 1 SSR	124	1	400	3301	117

Table 2: SSR types & no. of repeats in different palms:

SSR Type	Coconut	Arecanut	Oil palm	Date palm	Others palms
Monomers	29	4	1007	20504	185
Dimers	373	0	1294	10838	311
Trimers	16	1	684	2493	62
Tetramers	6	0	31	204	18
Penta and above	0	0	8	12	1
Compound	130	0	396	1457	115

Table 3: Different types of mono, di and tri motifs in palms

Motif Type	Coconut	Arecanut	Oil palm	Date palm	Others palms
A/T	28	2	859	18,337	149
G/C	3	2	148	885	36
AT/TA	10	0	179	3997	41
GC/CG	4	0	1	41	1
AG/GA/CT/TC	269	0	969	5615	202
CA/AC/GT/TG	90	0	139	1013	67
CTC	2	0	26	93	0
AAC/ACA/CAA/TTG/TGT/GTT	0	0	14	97	6
AAG/AGA/GAA/TTC/TCT/CTT	0	0	126	655	21
AAT/ATA/TAA/ATT/TTA/TAT	1	2	37	706	49
ACC/CCA/CAC/TGG/GGT/GTG	0	0	63	119	0
CAT/ATC/TCA/GTC/TCG	9	0	21	156	15
CTG/TGC/GCT/AGC/CGA/CGT	2	0	134	106	0
GCA/CAG/GAC/GTC/TCG/ACG	2	0	64	83	0
AGG/GAG/GGA/ATG/GAT/TAG	0	0	104	347	3
CCG/CGC/GCC/CCT/TCC	0	0	122	139	1
CGG/GGC/GCG/TAC/CTA/ACT	0	0	71	21	0
AGT/TGA	0	0	0	26	0