Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/10510
DOI URL: http://dx.doi.org/10.21474/IJAR01/10510

RESEARCH ARTICLE

INTEGRATED STANDALONE TOOL FOR IDENTIFICATION OF MICROSATELLITES

Umang[1] and Surya Pratap Singh[2]

1. Department of Computer Science, Banasthali University, Bansathali, 304022, Rajasthan, India.
2. Department of Bioinformatics, Banasthali University, Bansathali, 304022, Rajasthan, India.

Manuscript Info

Manuscript History
Received: 15 December 2019
Final Accepted: 18 January 2020
Published: February 2020

Key words:-
Simple Sequence Repeats, Perfect, Compound, MISA, Standalone Tool, Data Mining

Abstract

Microsatellites or simple sequence repeats are short tandem repeats of 1 to 6 nucleotides in length. These are widely used as molecular markers in plants, animals and human genome sequences. These Play important role in studying DNA variations, differentiate organisms at the molecular level, help in studying the evolutionary path of various genes, forensics studies, disease identification, paternity researches, etc. Laboratory identification of microsatellites is costly and time-consuming. Many computing tools, both web-based and standalone have been developed to detect simple sequence repeats. Presently available web-based microsatellite mining tools are costly to deploy and have technical complications of usages. Whereas standalone tools available are not user-friendly and easily accessible. Due to processing inadequateness in available tools, an effort was made to develop a new integrated tool with a rich graphic user interface in Java Net beans 8.0.2 and an algorithm was implemented in java to call Perl scripts, primer3 software and BlastN in the pipeline.

Introduction:-

Simple sequence repeats (SSRs) or “microsatellites” as coined by Litt and Luty in 1989 are short tandem repeats (motifs) of length 1–6 nucleotides (Toth et al., 2000) and are found in genomes of both prokaryotes and eukaryotes (Field and Wills, 1996). As per repeating unit pattern SSRs can be categorized into mono- (C)n, di- (TA)n, tri- (TAT)n, tetra- (AGAC)n, penta- (ATTAA)n and hexa- (TTTAAC)n nucleotide repeats where n is the number of repeating motif within the SSR locus. Further SSR can be categorized as perfect [without interruptions or we can say continuous repeat of single motif; (AGA) 15], imperfect [with interruptions by non-repeat nucleotide or with a base pair disruption between repeats; (AGA) 7A (AGA) 8] and compound [two or more SSRs are found adjacent to one another; (GTG)8(AT)16] (Bachmann and Bare, 2004). In addition to this, a compound SSR can also be categorized as a perfect compound [(AT)n(GA)n] and overlapping compound [overlap of few bases of previous SSRs with next SSR; (CCA)n(CT)n]. Further, Microsatellites can be genomic, if developed from genomic DNAs (gSSRs) or can be expressed, referred to as EST-SSRs if developed from an expressed sequence database, (Vieira et al., 2016, Ellis and Burke 2007). EST- SSRs have high power because of their associations with expressed genes, directly contributing to a phenotype (Varshney et al., 2005). In plants, SSRs can also be classified as nuclear SSRs if they occurred in nuclear DNA (nuSSRs) and chloroplast SSRs (cpSSRs) if they occurred in chloroplast DNA.SSRs are present in both coding and non-coding regions of the genome (Tauz and Renz, 1984; Gupta et al., 1996; Shanker et al., 2007a). The SSRs found in the coding region affect gene activation, resulting in the expression of protein and reflects lesser

Corresponding Author:- Umang
Address:- Department of Computer Science, Banasthali University, Bansathali, 304022, Rajasthan, India.

polymorphism in the coding part (Hancock, 1995) and SSRs are present in the non-coding region affect gene regulation (Lawson and Zhang, 2006). Tautz and Schlotterer (1994) reported that these repeats may be generated as a result of the slippage mechanism during replication. These microsatellites promote the development of markers that are widely used by researchers in DNA-based genetic analyses for the past 25 years, which show locus specificity, high reproducibility, co-dominance inheritance and hypervariability (Squirrell et al., 2003). The flanking sequences of SSRs help in selecting PCR primers that amplify the repeat sequence (Botstein et al., 1980). SSRs play important role in studying genetic variation, gene tagging, linkage mapping (Gupta et al., 2003; McCouch et al., 1997; Ramsay et al., 2000) and evolutionary studies (Buchanan et al., 1994). Many researchers have reported the involvement of SSRs in transcription, translation, regulation of promoters (Martin et al., 2005; Vinces et al., 2009) and certain neurodegenerative diseases (Ashley and Warren, 1995). These are widely used in paternity testing, mapping locations within the genome; Researchers use microsatellites in population genetics and species conservation projects. Plant geneticists have proposed the use of microsatellites for marker-assisted selection of desirable traits in plant breeding. Microsatellites are used for assessing chromosomal DNA deletions in cancer diagnosis. Microsatellites are widely used for DNA profiling, also known as "genetic fingerprinting", of crime stains (in forensics) and tissues (in transplant patients).

Considering the importance of microsatellites in studying genetic similarity/dissimilarity, various studies have been made to identify and characterize them in the laboratory. The development of SSR markers in the laboratory is intensive and time-consuming (Zane et al., 2002). The increasing availability of genome sequences of various organisms in biological databases proved to be a fast and inexpensive way for in silico mining of SSRs (Shanker et al., 2007b). Therefore with the advancement in technology and the easy availability of genome sequences at NCBI, many bioinformatics tools have been developed to detect microsatellites. In a survey broadly three types of tools were found like web-based, species-specific databases and standalone software. The web-based tools such as Tandem Repeat Finder by Benson (1999), MISA-web (Beier et al., 2017), SSRFinder, PALFinder (Hodel et al., 2016 and Vieira et al., 2016) and SciRoko (Kofler et al., 2007) conduct SSR mining, whereas other tools, such as SSRLocator (Maia et al., 2008) and SSRPoly (Duran et al., 2013) also design primers. Genome specific SSR identification databases such as Cotton Marker Database (Blenda et al., 2006), EuMicroSatdb (Aishwarya et al., 2007), PIPEMicroDB (Sarika et al., 2013), Setaria (Pandey et al., 2013), MitoSatPlant (Kumar et al., 2014), ChloroSSRdb (Kapil et al., 2014), cotton (Wang et al., 2015) and CyanoSat (Kabra et al., 2016) were developed.

These may detect perfect, imperfect or both microsatellites. Our study revealed that for many web-based SSR and EST mining tools like ParPEST (Chiusano et al., 2005) and ESAP plus (Ponyard et al., 2016) literature is available but access to them is difficult. Similarly, WebSat: software for marker development restricts the user to upload 150,000 characters only. The offline tools available do not provide batch processing like GMATo (Wang et al., 2013) and GMATA (Wang and Wang, 2016). Krait tool (Lianming Du et al., 2017) performs batch primer design. All these tools have different features that cater to different needs as per study or objectives. In the present work, an effort has been made to develop robust standalone software that provides batch processing of files for primers design of Perfect motifs to identify common, unique and polymorphic microsatellites to show length polymorphism in other genomes for a specific genetic study.

Material and Methods:-

Input files:

Genome or nucleotide sequences in Genbank and FASTA format were downloaded from NCBI Genbank. (https://www.ncbi.nlm.nih.gov/genbank/).

Technology used:

Input files were processed for microsatellites detection using the in-house built standalone tool with an interactive user-friendly graphic user interface that has been designed using Java Net Beans IDE 8.0.2; it is robust and platform-independent technology. Strawberry Perl is used for implementing Perl script misa.ini which is a configuration file to set the number of interruptions and repeat size and misa.pl (MISA, http://pgrc.ipk-gatersleben.de/misa/download/misa.pl), a Perl script that is used for mining perfect and compound simple sequence repeats. The entire algorithm has been written using Java programming language that performs a call to misa.ini, misa.pl and Primer3 software (http://primer3.org/releases.html) with default parameters. Microsatellites considering the flanking region of 200 nucleotides are used in the pipeline to design batch primers of all the detected microsatellites. The workflow is demonstrated in figure 1.

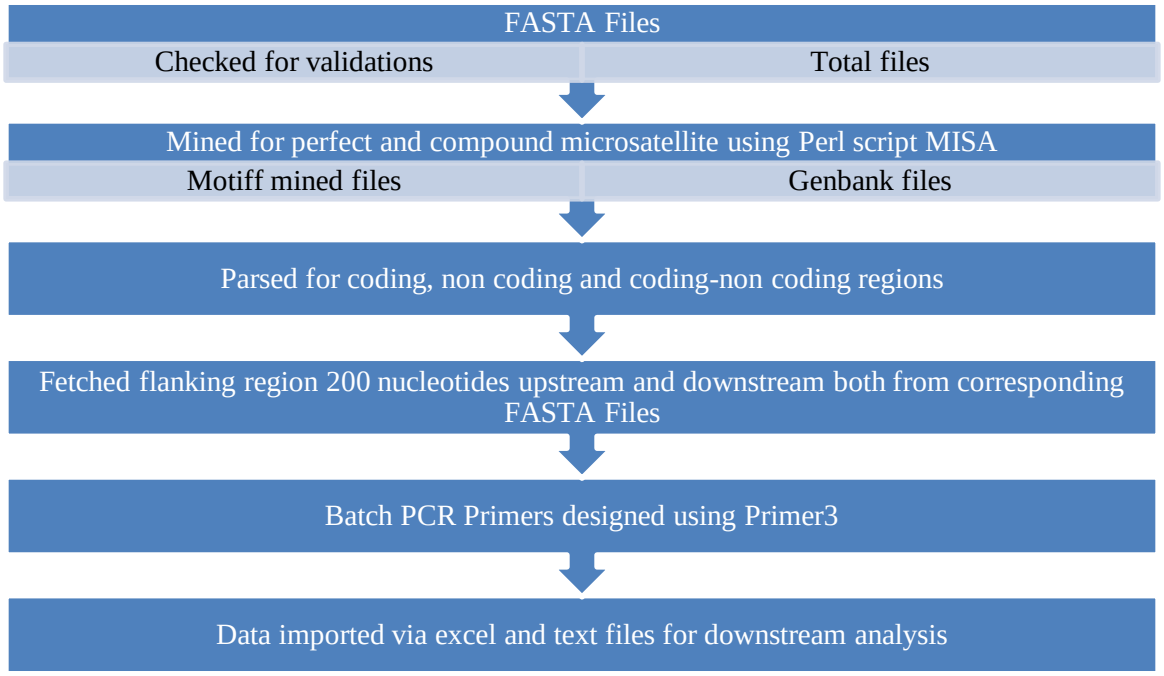

Figure 1:- Workflow of the Microsatellite Identification/Primer design and related statistics.

Results and Discussion:-

This tool has been successfully tested and run on various size complete genome sequences for mining and analyzing perfect and compound microsatellites. Batch processing of FASTA format files results into information such as SSR type (perfect or compound), size, start position, end position, coding, and non-coding region, Flanking sequences of 200 nucleotides for both upstream and downstream region of SSRs are provided with left and right primers, left and right primer length, primer TM, primer GC %, product size. These primers are important for developing microsatellite-based markers and to identify common, unique and polymorphic simple sequence repeats. Details of null primers i.e. microsatellites for which primers are not formed due to insufficient flanking regions or poor melting temperatures are stored in separate files. Statistics details is displayed in separate files mentioning total number of sequences examined, total size of examined sequences (base pair), total number of identified SSRs, number of SSR containing sequences, number of sequences containing more than 1 SSR, Number of SSRs present in compound formation, distribution to different repeat type classes; having unit size with corresponding number of SSR and frequency of identified SSR motifs. The data is automatically saved to text and MS excel formats in designated folders.

Unique features of the tool:

1. The minimum number of nucleotides and interruptions can be reset and saved in the configuration file.
2. Proper validations are applied to check valid file types and extensions.
3. It checks the FASTA format files. For batch processing the number, names and location of files send for mining is displayed (figure2).
4. Multiple files can be processed at a time with a single click, no need to attach or upload files again and again.
5. The files after mining (simple sequence repeats and statistics) are saved to properly designated folders (figure 3).
6. No file size restrictions.
7. The users can either simply detect microsatellites and related statistics or by using corresponding GenBank file they can detect coding, non-coding regions with flanking sequence and primer design.
8. Output formats are text and excel files. Results can be used with BlastN for further analysis.

Batch processing of sequences with accession number BK010421.1 (364 kb), EU999004.1 (2.0 kb), EU999005.1(3.0 kb), EU999006.1(2.0 kb), EU999007.1(1.0 kb), EU999008.1 (2.0 kb) and FJ156734.1(1.0 kb) having total file size 375 kb was performed in 6.0 seconds.

.

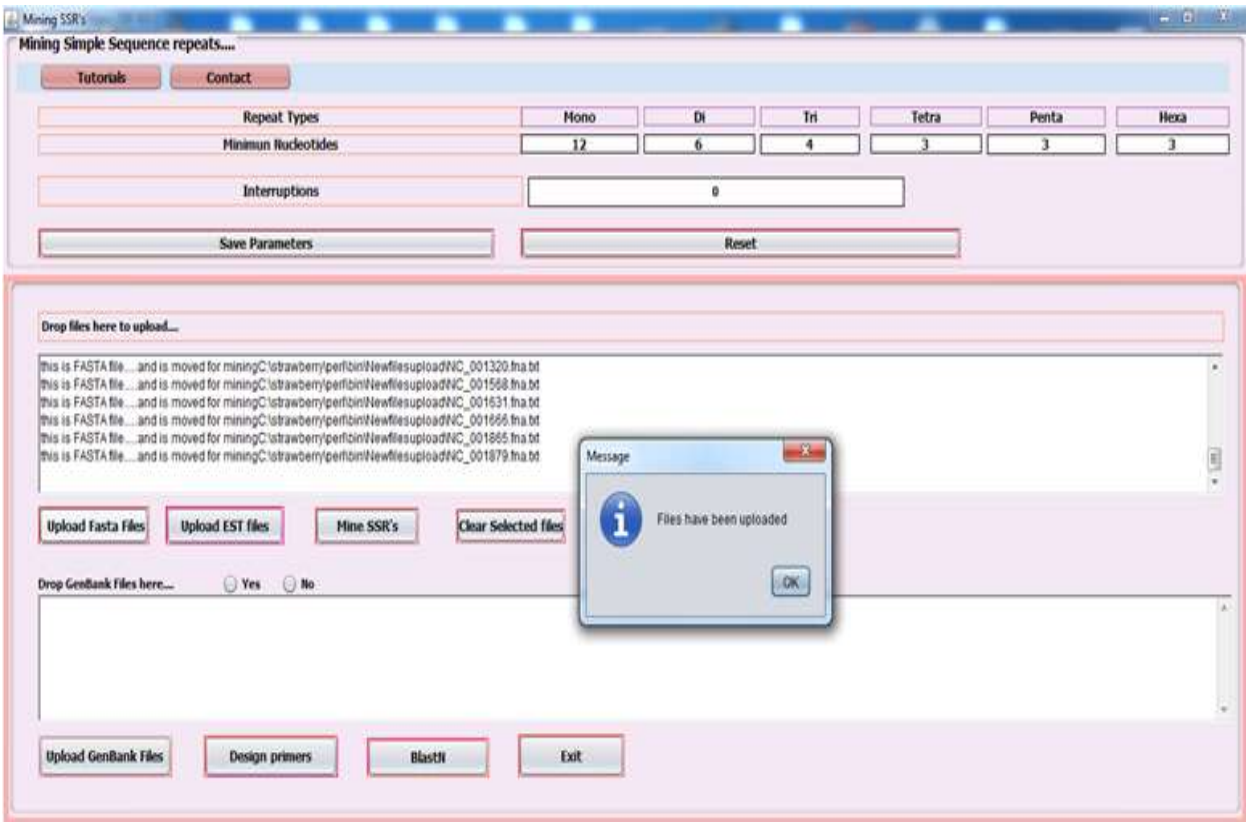

Figure 2:- Tool displaying batch submission of files for processing.

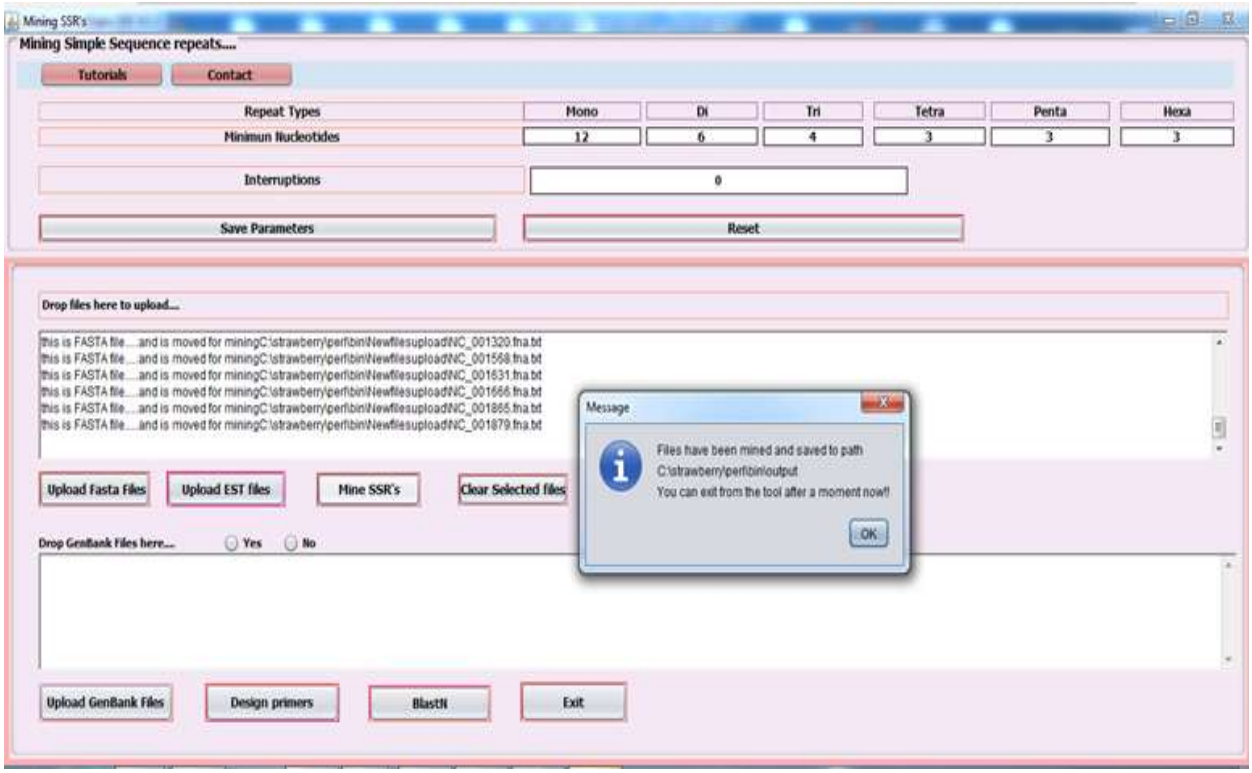

Figure 3:- Tool displaying batch files processed and saved at respective paths.

Conclusion:-

This integrated tool is a standalone application of MISA that detects perfect and compound simple sequence repeats, design primers and can perform similarity search of flanking regions using BlastN for finding common, unique and putative polymorphic simple sequence repeats. This tool contains rich features that save time, cost and efforts for identifying microsatellites. The output files are properly formatted and can be used as input to other pipelines. The tool is developed and tested on AMD E-350 processor 1.60 GHz with 2.0 GB RAM and 32-bit operating system. This tool can be used with upgraded computers and can give much better performance.

References:-

1. Aishwarya, V., Grover, A., Sharma, P.C., 2007. EuMicroSatdb: a database for microsatellites in the sequenced genomes of eukaryotes. BMC Genomics 8, 225.
2. Ashley, C.T., Warren, S.T., 1995. Trinucleotide repeats expansion and human disease. Annu.Rev. Genet. 29, 703–728.
3. Bachmann, L., Bare, P.T.J., 2004. Allelic variation, fragment length analyses and population genetic model: a case study on Drosophila microsatellites. Zool. Syst. Evol. Res. 42, 215–222.
4. Beier et al.(2017) MISA-web:aweb server for microsatellite prediction.Bioinformatics,33(16),2017,2583-2585.
5. Benson. (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res., 27, 573-580.
6. Blenda, A., Scheffler, J., Scheffler, B., Palmer, M., Lacape, J.M., Yu, J.Z., Jesudurai, C., Jung, S., Muthukumar, S., Yellambalase, P., Ficklin, S., Staton, M., Eshelman, R., Ulloa, M., Saha, S., Burr, B., Liu, S., Zhang, T., Fang, D., Pepper, A., Kumpatla, S., Jacobs, J., Tomkins, J., Cantrell, R., Main, D., 2006. CMD: a cotton microsatellite database resource for Gossypium genomics. BMC Genomics 7, 132.
7. Botstein, D., White, R.L., Skolnick, M., Davis, R.W., 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32, 314–331.
8. Buchanan, F.C., Adams, L.J., Littlejohn, R.P., Maddox, J.F., Crawford, A.M., 1994. Determination of evolutionary relationships among sheep breeds using microsatellites. Genomics 22, 397–403.
9. Chiusano et al. (2005) ParPEST: a pipeline for EST data analysis based on parallel computingBMC Bioinformatics 2005, 6(Suppl 4):s9 doi 10.1186/1471-2105-6-s4-s9.
10. Duran,C.,Singhania,R.,Raman,H.,Batley,J.,andEdwards,D.(2013). Predicting polymorphic EST-SSRs *insilico*. *Mol.Ecol.Resour*. 13,538–545.doi: 10.1111/1755-0998.12078
11. Ellis JR, Burke JM. EST-SSRs as a resource for population genetic analyses. Heredity (Edinb). 2007; 99:125-132.
12. Field, D., Wills, C., 1996. Long, polymorphic microsatellites in simple organisms. Proc. Biol.Sci. 263, 209—215.
13. Gupta, P.K., Balyan, H.S., Sharma, P.C., Ramesh, B., 1996. Microsatellites in plants: a new class of molecular markers. Curr. Sci. 70, 45—54.
14. Gupta, P.K., Rustgi, S., Sharma, S., Singh, R., Kumar, N., Balyan, H.S., 2003. Transferable EST- SSR markers for the study of polymorphism and genetic diversity in bread wheat. Mol. Genet. Genomics 270, 315–323.
15. Hancock, J.M., 1995. The contribution of slippage-like processes to genome evolution.
16. Hodel RG, Segovia-Salcedo MC, Landis JB, Crowl AA, Sun M, Liu X, Gitzendanner MA, Douglas NA, Germain-Aubrey CC, Chen S, Soltis DE, Soltis PS. The report of my death was an exaggeration: A review for researchers using microsatellites in the 21st century. Appl Plant Sci. 2016; 4:1600025. DOI:10.3732/apps.1600025.J. Mol. Evol. 41, 1038—1047.
17. Kabra R, Kapil A, Kherunnisa A,Rai PK, Shanker A (2016) Identification of common, unique and polymorphic microsatellites among 73 cyanobacterial genomes. World J Microbiol Biotechnol (2016) 32:71 DOI 10.1007/s11274-016-2061-0.
18. Kapil A, Rai PK, Shanker A (2014) ChloroSSRdb: a repository of perfect and imperfect chloroplastic simple sequence repeats (cpSSRs) of green plants. Database 2014:1–5.
19. Kofler,R. et al. (2007) SciRoKo: a new tool for whole genome microsatellite search and investigation. Bioinformatics, 23, 1683–1685. doi:10.1093/bio-informatics/btm157.
20. Kumar M, Kapil A, Shanker A (2014) MitoSatPlant: mitochondrial microsatellites database of Viridiplantae. Mitochondrion 19:334–337.
21. Lawson, M.J. and L. Zhang (2006). Distinct patterns of SSR distribution in the Arabidopsis thaliana and rice genomes. Genome Biol., 7: R14.1-11.

22. Lianming Du, Chi Zhang, Qin Liu, Xiuyue Zhang, Bisong Yue, Krait: an ultrafast tool for genome-wide survey of microsatellites and primer design, Bioinformatics, Volume 34, Issue 4, 15 February 2018, Pages 681–683, https://doi.org/10.1093/bioinformatics/btx665.
23. Litt M, Luty JA. A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet. 1989; 44:397–401.
24. Luciano Carlos da Maia, Dario Abel Palmieri, Velci Queiroz de Souza, Mauricio Marini Kopp, Fernando Irajá Félix de Carvalho, and Antonio Costa de Oliveira, "SSR Locator: Tool for Simple Sequence Repeat Discovery Integrated with Primer Design and PCR Simulation," International Journal of Plant Genomics, vol. 2008, Article ID 412696, 9 pages, 2008. https://doi.org/10.1155/2008/412696.
25. Martin, P., Makepeace, K., Hill, S.A., Hood, D.W., Moxon, R., 2005. Microsatellite instability regulates transcription factor binding and gene expression. Proc. Natl. Acad. Sci. U. S. A. 102, 3800—3804.
26. Matthies,I.E. et al. (2012) Population structure revealed by different marker types (SSR or DArT) has an impact on the results of genome-wide association mapping in European barley cultivars. Mol. Breed., 30, 951–966.
27. McCouch, S.R., Chen, X., Panaud, O., Temnykh, S., Xu, Y., Cho, Y.G., Huang, N., Ishii, T., Blair, M., 1997. Microsatellite marker development, mapping and application in rice genet- ics and breeding. Plant Mol. Biol. 35, 89—99.
28. Pandey,G.,Misra,G.,Kumari,K.,Gupta,S.,Parida,S.K.,Chattopadhyay,D.,etal.(2013).Genome wide development and useof microsatellite markers for large- scale genotyping applications in foxtailmillet *Setariaitalica* (L.). *DNARes.* 20, 197–207.doi:10.1093/dnares/dst002.
29. Ponyared et al. 2016 ESAP plus: a web based server for EST-SSR marker development.BMC Genimics 2016,17(Suppl 13):1035 DOI 10.1186/s12864-016-3328-4.
30. Ramsay, L., Macaulay, M., degli Ivannissevich, S., MacLean, K., Cardle, L., Fuller, J., Edwards, K.J., Tuvesson, S., Morgante, M., Massari, A., Maestri, E., Marmiroli, N., Sjakste, T., Ganal, M., Powell, W., Waugh, R., 2000. Simple sequence repeat-based linkage map of barley. Genetics 156, 1997–2005.
31. Sarika, Arora, V., Iquebal, M.A., Rai, A., Kumar, D., 2013. PIPEMicroDB: microsatellite data- base and primer generation tool for pigeonpea genome. Database (Oxford) 2013. http://dx.doi.org/10.1093/database/bas054.
32. Shanker, A., Bhargava, A., Bajpai, R., Singh, S., Srivastava, S., Sharma, V., 2007b. Bioinformatically mined simple sequence repeats in expressed sequences of *Citrus sinensis*. Sci. Hortic. 113, 353—361
33. Shanker, A., Singh, A., Sharma, V., 2007a. *In silico* mining in expressed sequences of *Neu- rospora crassa* for identification and abundance of microsatellites. Microbiol. Res. 162, 250–256.
34. Squirrell, J., Hollingsworth, P.M., Woodhead, M., Russell, J., Low, A.J., Gibby, M., Powell, W., 2003. How much effort is required to isolate nuclear microsatellites from plants? Mol. Ecol. 12, 1339—1348.
35. Tautz D. and M. Renz (1984). Simple sequence repeats are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Res.,12:4127-4138.
36. Tautz,D. and Schlotterer, (1994) Simple sequences. Curr. Opin. Genet. Dev., 4, 832–837.
37. Toth, G., Gaspari, Z., Zurka, J., 2000. Microsatellites in different eukaryotic genome, survey and analysis. Genome Res. 10, 1967—1981.
38. Varshney RK, Graner A, Sorrells ME. Genic microsatellite markers in plants: features and applications. Trends Biotechnol. 2005; 23:48–55.
39. Vieira ML, Santini L, Diniz AL, Munhoz CF. Microsatellite markers: what they mean and why they are so useful. Genet Mol Biol. 2016; 39:12–328. DOI: 10.1590/1678-4685- GMB-2016-0027.
40. Vinces, M.D., Legendre, M., Caldara, M., Hagihara, M., Verstrepen, K.J., 2009. Unstable tan- dem repeats in promoters confer transcriptional evolvability. Science 324, 1213—1216.
41. Wang X and Wang L (2016) GMATA: An Integrated Software Package for Genome-Scale SSR Mining, Marker Development and Viewing. Front. Plant Sci. 7:1350. doi: 10.3389/fpls.2016.01350.
42. Wang, X. et al. (2013) GMATo: a novel tool for the identification and analysis of microsatellites in large genomes. Bioinformation, 9, 541–544. Doi: 10.6026/97320630009541.
43. Wang,Q.,Fang,L.,Chen,J.,Hu,Y.,Si,Z.,Wang,S.,etal.(2015).Genome-wide mining,characterization,and development ofmicrosatellite markers in Gossypium species. *Sci.Rep.* 5:10638.doi:10.1038/srep10638.
44. Zane, L., Bargelloni, L., Patarnello, T., 2002. Strategies for microsatellite isolation: a review. Mol. Ecol. 11, 1—16.