DNA Research **20**, 127–133, (2013)
Advance Access publication on 2 January 2013
doi:10.1093/dnares/dss038

ChloroMitoSSRDB: Open Source Repository of Perfect and Imperfect Repeats in Organelle Genomes for Evolutionary Genomics

Gaurav Sablok[1,*], Suresh B. Mudunuri[2], Sujan Patnana[3], Martina Popova[4], Mario A. Fares[5,6], and Nicola La Porta[1]

Sustainable Agro-ecosystems and Bioresources Department, IASMA Research and Innovation Centre, Fondazione Edmund Mach, Via E. Mach 1, San Michele all'Adige, Trentino 38010, Italy[1]; Department of Computer Science & Engineering, Vishnu Institute of Technology, Vishnupur, Bhimavaram, Andhra Pradesh, India[2]; TalentSprint Edu. Services, International Institute of Information Technology (IIIT) Campus, Gachibowli, Hyderabad, Andhra Pradesh, India[3]; Department of Plant Physiology and Molecular Biology, University of Plovdiv, 24 Tsar Assen St., Plovdiv 4000, Bulgaria[4]; Department of Genetics, Trinity College Dublin, University of Dublin, Dublin 2, Dublin, Ireland[5] and Department of Abiotic Stress, Instituto de Biología Molecular y Celular de Plantas, CSIC-UPV, Ingeniero Fausto Elio s/n, València 46022, Spain[6]

*To whom correspondence should be addressed. Tel. +39 3270484732. Email: sablokg@gmail.com

Edited by Prof. Kenta Nakai
(Received 3 September 2012; accepted 26 November 2012)

Abstract

Microsatellites or simple sequence repeats (SSRs) are repetitive stretches of nucleotides (A, T, G, C) that are distributed either as single base pair stretches or as a combination of two- to six-nucleotides units that are non-randomly distributed within coding and in non-coding regions of the genome. ChloroMitoSSRDB is a complete curated web-oriented relational database of perfect and imperfect repeats in organelle genomes. The present version of the database contains perfect and imperfect SSRs of 2161 organelle genomes (1982 mitochondrial and 179 chloroplast genomes). We detected a total of 5838 chloroplast perfect SSRs, 37 297 chloroplast imperfect SSRs, 5898 mitochondrial perfect SSRs and 50 355 mitochondrial imperfect SSRs across these genomes. The repeats have been further hyperlinked to the annotated gene regions (coding or non-coding) and a link to the corresponding gene record in National Center for Biotechnology Information(www.ncbi.nlm.nih.gov/) to identify and understand the positional relationship of the repetitive tracts. ChloroMitoSSRDB is connected to a user-friendly web interface that provides useful information associated with the location of the repeats (coding and non-coding), size of repeat, motif and length polymorphism, etc. ChloroMitoSSRDB will serve as a repository for developing functional markers for molecular phylogenetics, estimating molecular variation across species. Database URL: ChloroMitoSSRDB can be accessed as an open source repository at www.mcr.org.in/chloromitossrdb.

Key words: chloroplast; database; mitochondria; microsatellites; simple sequence repeats; web interface

1. Introduction

Microsatellites, or simple sequence repeats (SSRs), are repetitive stretches of a tandemly repeated motif of one to six base pairs, which has evolved and expanded owing to the replication slippage mechanism that is supposed to be the cause of their high polymorphic rates.[1] Recently, using a genome-wide alignment of two *Orzya* species var. indica and japonica, it has been demonstrated that the distribution of microsatellites is also influenced by the motif sequence and the sequence characteristics of the adjoining regions possessing the microsatellites, in addition to the replication slippage and point mutation model.[2] These repetitive stretches may occur in coding and in non-coding regions of the genome. SSRs have been potentially designated as a class of co-dominant markers for evaluating germplasm,

establishing phylogenetic and evolutionary relationships. It has been observed that clusters of microsatellite motifs with moderate GC are abundant on chromosome number 2 in the model plant *Arabidopsis thaliana*, which suggests that repetitive stretches may be biased towards the accumulation in a certain regions.[3] Microsatellites have been associated with various functional roles such as their possible role in the regulation of promoters, transcription and translation, and these sequence repeats have been credited with evolutionary importance.[4–6] The positioning of microsatellites in the genome seems to play an important role in their regulatory activity; hence, studying the distribution and understanding the possible reasons of microsatellites expansions across genomes have currently been the focus of current intense research.

Organelle genomes, plant chloroplast and animal mitochondrial genomes have been referred to as natural counterparts.[7,8] Features such as conserved gene order, lack of heteroplasmy (occurrence of more than one type of organelle genome), low recombination rates and their relative small size are making these organelle genomes the widely used tools for phylogenetic studies. However, lack of heteroplasmy has not been universally observed in all the mitochondrial genomes and has been earlier potentially reviewed with the occurrence and factors affecting the stoichiometry of heteroplasmy in mitochondrial genomes of plants and animals.[9] The uniparental inheritance of the organelle markers provides a means to elucidate the genetic flow and genetic structure of the population and the organelle markers have been widely used in population studies (for a review see Provan et al.).[8] *In silico* development of SSRs of organelle genomes has brought them up as potential markers for transferability among the species, ease of development and as key players in genome length variation. They have been widely demonstrated as potential markers for establishing molecular evolutionary histories, demographic diversity and resolving phylogeny in a wide variety of species from *Pinus* (forest species) to *Oryza sativa* (Monocots).[10–12] There have been recent reports on the identification of perfect repeats in organelle genomes of various organisms.[11–16] However, previous studies have only been focused on a relatively small number of genomes and only perfect repeats have been identified. A proper characterization system that would allow researchers to search for the association of these repeats with the coding or non-coding regions has been lacking in these reports.

In the past few years, systematic curated web repositories have been developed for the organelle genomes, which includes FUGOID that displays the curated distribution of introns in organelle genomes with functional and structural data.[17] A database of universally published primer sequences of chloroplast genomes has been developed, providing a platform for studying molecular variations and evolution in chloroplasts.[18] These organelle genomes have been exploited further for the mining of genes, exons, introns, gene products, taxonomy, RNA editing sites, SNPs and haplotype information, all of which are displayed as curated information in GOBASE.[19] A comprehensive repository of unique proteins expressed in chloroplast proteome using liquid chromatography-mass spectrometry/mass spectrometry has been developed (AT_CHLORO), serving as a knowledge base to explore the envelope proteins.[20] However, a complete curated web-oriented integrated repository of repeat pattern is still lacking. This has motivated us to undertake a genome-wide study and to develop a web-enabled interface to analyse the perfect and the imperfect repeats in organelle genomes.

We propose ChloroMitoSSRDB that offers a wide visualization of perfect and imperfect repeats across the chloroplast and mitochondrial genomes with corresponding genomic coordinates. The aim of ChloroMitoSSRDB is to constitute a platform to access the utility of SSRs as markers for phylogenetic classification across species. To our knowledge, this is the first updated integrated repository of the genomic repeats in chloroplast and mitochondrial genomes accessible via web interface.

2. Material and methods

2.1. *Genome data retrieval and pattern search*

All the studied chloroplast (179) and mitochondrial (1982) genomes were retrieved from the National Center for Biotechnology Information (NCBI) RefSeq database (www.ncbi.nlm.nih.gov/). The required files such as gbk, fna, faa, gff and ptt were downloaded for the studied chloroplast and mitochondrial genomes and were stored as flat files sorted for each genome. For the identification of the perfect and imperfect repeats, the software tool Imperfect Microsatellite Extractor (IMEx)[21] has been used, which uses a sliding window algorithm to identify the regions with a repetitive stretch of a particular nucleotide motif either stretched perfectly or with levels of imperfection.

The algorithm allows the user to specify the minimal length of the consecutive nucleotide stretch and reports the SSR motif, motif repeat counts, coordinates of the SSRs tract in the genome and its location relative to coding and non-coding regions. The association of the repeats in coding and intercoding regions was determined based on the sequence annotation information available in GenBank database (NCBI, www.ncbi.nlm.nih.gov). We applied the following length

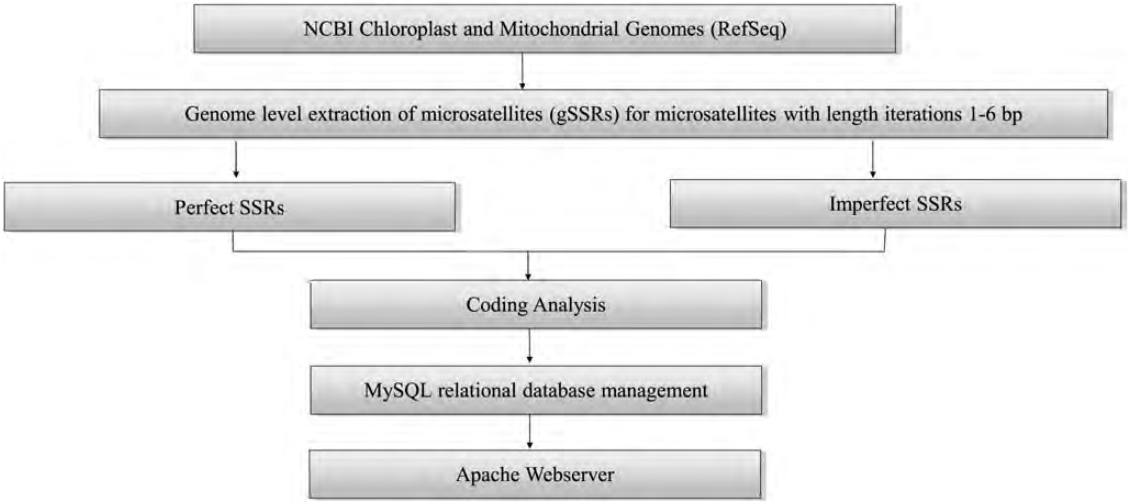

Figure 1. Schematic illustration showing the flow of the organization of the data in ChloroMitoSSRDB.

criteria (Mono-, 12; Di-, 6; Tri-, 4; and for Tetra- to Hexa repeats, a minimum stretch of three minimum repetitions) to define each SSRs as a true repeat. In case of imperfect repeats, the parameter for imperfection percentage (p%) is set to 10% indicating the level of imperfection allowed in each repeat tract.

3. Results and discussions

3.1. *Structure of ChloroMitoSSRDB database*

ChloroMitoSSRDB is hosted on a 32-bit Linux server pre-installed with MySQL (http://www.mysql.com/), Apache (http://www.apache.org/) and PHP (http://www.php.net/) commonly called as LAMP. A flow chart explaining the organization and the work flow of the ChloroMitoSSRDB has been presented (Fig. 1). ChloroMitoSSRDB is based on a simple comprehensive relational database management system, MySQL, that is sufficient for organizing, storing and retrieving the data with a single query. The details of the relational MySQL tables used in the construction of the ChloroMitoSSRDB database are explained in Tables 1 and 2. Table 1 shows the metadata for each genome, whereas the structure of the MySQL relational tables depicting the repeat information stored for the coding and the non-coding regions is given in Table 2. Each query has been split into hierarchical levels of information that displays information on each Genome (e.g. accession, sequence length and nucleotide composition) (Table 1).

The information for the genome composition (A-, T-, G- and C- counts, etc.) has been computed from the flat files obtained from the NCBI RefSeq database (Table 1). The complete repeat information of the database is stored in two different tables (refer Table 2), storing the perfect and imperfect repeats of all chloroplast and mitochondrial genomes. The repeat information includes the details of individual repeats such as the sequence ID, start and end coordinates of the repeat, the repeating motif, number of

Table 1. Structure of the table 'chloromitometa' that stores the meta-information of all the mitochondrial and chloroplast genomes

Information	Field	Data type	Key	Example
Accession number	acc_no	int(11)		5881414, 110189662
Sequence ID	seq_id	varchar(11)	PRI	NC_000834, AC_000022
Sequence name	seq_name	varchar(500)		*Rattus norvegicus* strain Wistar mitochondrion, *Porphyra purpurea* chloroplast
Sequence type	seq_type	varchar(50)		Complete genome, complete sequence
Sequence length	seq_length	int(11)		16 613 bp, 7686 bp
Nucleotide composition of A	a_per	Float		33.06%
Nucleotide composition of T	t_per	Float		41.87%
Nucleotide composition of G	g_per	Float		13.58%
Nucleotide composition of C	c_per	Float		11.49%
Organelle type	organelle	Char(1)		M (for Mitochondrion), C (Chloroplast)
Taxon ID	taxon	Int		263 995

Table 2. Structure of the tables 'chloromitoperfectmicrosatellite' and 'chloromitoimperfectmicrosatellite' that store the repeat information of all perfect and imperfect microsatellites of mitochondrial and chloroplast genomes

Information	Field	Data type	Key	Example
Sequence ID	index_no	varchar(11)	PRI	NC_000834, AC_000022
Starting co-ordinate of SSR	start	int(11)	PRI	172, 12843
Ending co-ordinate of SSR	end	int(11)	PRI	182, 12885
motif (repeating unit)	motif	varchar(10)		AT, G, CAAC
Number of repetitions	iterations	int(5)		3, 7
Length of repeat tract	tract_length	int(11)		12 bp, 18 bp
Nucleotide composition of A	a_per	Float		50.00%
Nucleotide composition of T	t_per	Float		0.00%
Nucleotide composition of G	g_per	Float		33.33%
Nucleotide composition of C	c_per	Float		16.67%
Repeat position information	coding_info	varchar(50)		Coding (if repeat is in the coding region) or Null (if outside)
Protein ID (if repeat in coding region)	protein_id	int(11)		110189664 (if repeat is in the coding region) or 0 (if non-coding)
[a]Imperfection percentage of the tract	imperfection	Float		9%, 0%
[a]Alignment line 1	alignment_line1	Text		TTAA-TAATTAA
[a]Alignment line 2	alignment_line2	Text		**** *******
[a]Alignment line 3	alignment_line3	Text		TTAATTAATTAA

[a]The last four columns (imperfection, alignment_line1, alignment_line2 and alignment_line3) are present only in the table that stores imperfect microsatellites (chloromitoimperfectmicrosatellite).

iterations, total tract length, nucleotide composition of the repeat, protein information of coding repeats. In addition, the table displaying the imperfect repeats also stores the imperfection percentage and alignment information that can be used to study the evolution of these repeats.

3.2. *Web visualization of ChloroMitoSSRDB*

The front end of the database is integrated via web accessible PHP scripts. The web interface allows various patterns of search for the repeats in organelle genomes. The complete browsing outlay of the ChloroMitoSSRDB is displayed (Fig. 2). The curated information is organized into several search patterns, and proper navigation pages have been provided. The curated information from the IMEx has been processed further according to gene IDs, organism name, and the SSRs were sorted according to the coding or non-coding regions. The position of the coding regions has been determined using the annotated ptt files of each chloroplast and mitochondrial genome as downloaded from the NCBI Refseq database.

ChloroMitoSSRDB interface provides information on several repeat statistics, including the distribution of the repeat types, length of the motifs and their positions (coding or non-coding repeats). The querying of ChloroMitoSSRDB through the web interface is organized into three search patterns that accomplish all interface functionalities: query page, result page and report page: (i) the first search pattern is according to the organelle classification and it has been classified into chloroplast and mitochondrial genomes, (ii) the second search pattern has been classified according to the type of repeat pattern (perfect or imperfect) and (iii). the last search pattern allows the user to select the repeat size. With the appropriate selection pattern, the user will be directed to the organelle-specific page (chloroplast and mitochondrial) containing the list of the organism for which the SSRs have been identified, which are further linked to the organism-specific repeat pages for further information on the distribution of the repetitive tracts.

To ease the access of the database and to enhance the user functionality, we also provide chloroplast and mitochondrial repeat-specific pages alphabetically ordered according to the organism name. An advanced search option has been provided to filter the repeats based on the user-specific criteria allowing the user to search for a repeat region of a specific length. An option to export the search results and the repeat information in excel format has been provided, so that the user can save and analyse the repeats, design primers and can utilize the information for further downstream processing of the observed repeats.

A query page for every organism is directed to a ChloroMitoSSRDB repeat summary page for organism-specific summary page that gives a detailed illustration of the distribution of the perfect and the imperfect repeats distribution and the genome

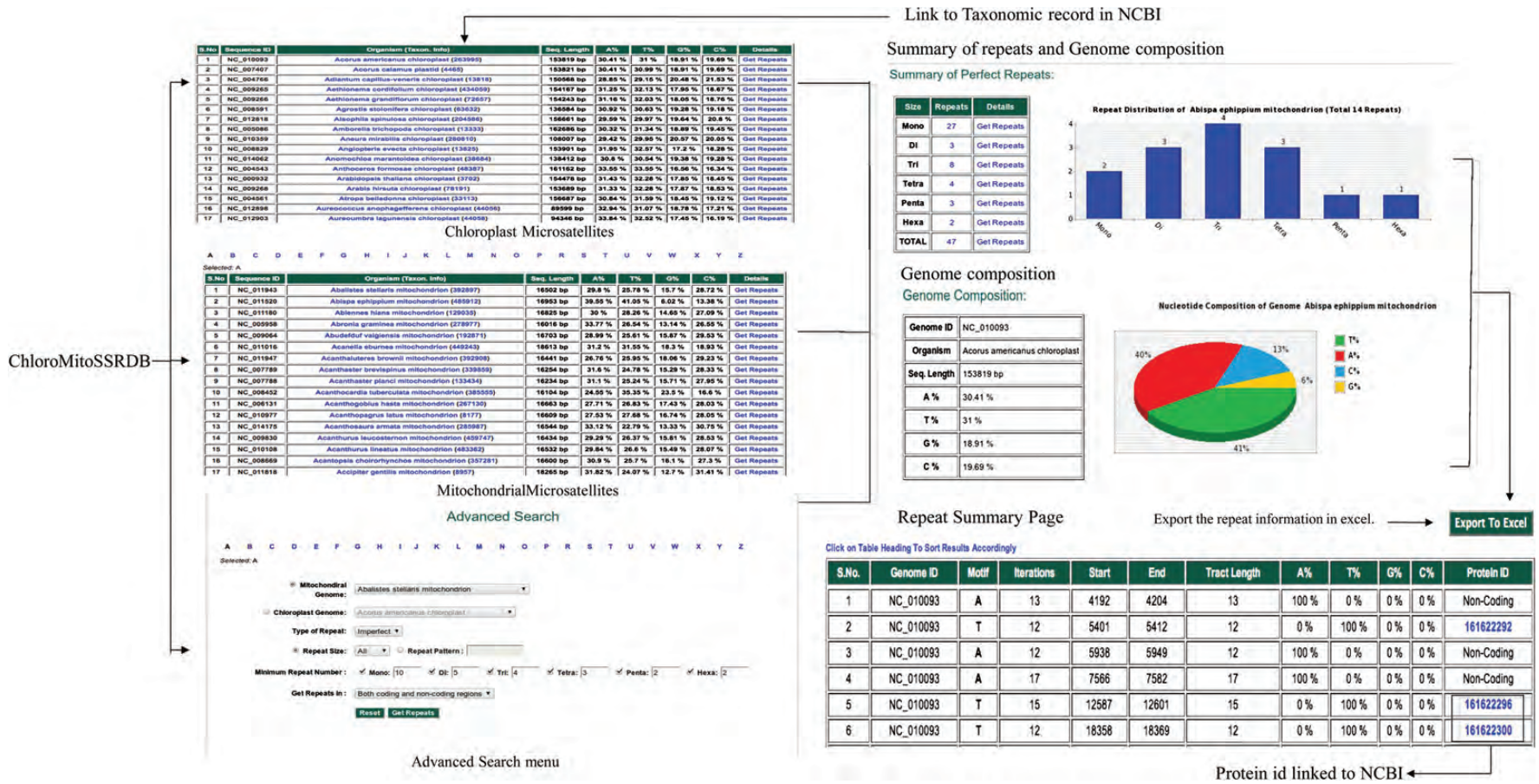

Figure 2. How to browse: schematic browsing of ChloroMitoSSRDB.

composition using bar and pie charts. The genome composition and the repeat occurrence graphs were generated dynamically based on the repeat information using Libchart, a PHP chart drawing library (http://naku.dohcrew.com/libchart/). The repeat pattern summary displayed on the organism specific page are clickable links, which redirects and give further information on the start and end of the SSR repeat containing tract, Motif and the occurrence of the respective repeat pattern across the genomes.

Mutations in the SSR stretches prevailing in the coding region may affect the subsequent transcription and translation of the gene harbouring the repetitive stretches of SSRs.[22] Mutations in chloroplast SSRs (mutation rates at cpSSR loci as between 3.2×10^{-5} and 7.9×10^{-5}) have been described as low when compared with substitution rates.[23] Recently, it was observed that the plant mitochondrial substitution rates are relatively lower when compared with the invertebrates and mammalian mitochondrial genomes.[24,25] To evaluate the distribution of the SSRs in the coding regions, the repeat-rich regions on the organism page have been linked to the corresponding protein IDs (NCBI, www.ncbi.nlm.nih.gov/), in case of coding repeats, which can shed light on the evolution of these repeated regions either through mutational bias or through selective forces in further ongoing work.

4. Conclusion

We have consecutively constructed a database ChloroMitoSSRDB that displays curated information of wide spread occurrences of genomic repeats in chloroplast and mitochondrial genomes available so far, and we will be constantly updating ChloroMitoSSRDB with the new chloroplast and mitochondrial genomes as and when they are released. The repeats in the coding regions of the genes may prove to be candidate markers to study the functional role of repeats associated with the genes, as possible markers for species delimitation, evolutionary analyses and also for evaluating the germplasm and to hypothesize conservation strategies for endangered species. In future release, we will make efforts to upgrade the primer pair information for the repeat-rich regions and will also upgrade the database with the systematic visualization of imperfect alignments through the availability of hyperlinked pages in case of imperfect repeats. We believe that ChloroMitoSSRDB will serve as a standard database for exploring and understanding genomic repeats in organelle genomes, and the data represented in ChloroMitoSSRDB make a good starting point for further exploratory investigations on SSR polymorphism, large comparative genome comparison and provide a platform to understand the repetitive nature of organelle genomes.

Acknowledgements: G.S. thanks Ivan Milev for reading the manuscript. S.M. gratefully acknowledges the support of Vishnu Educational Society in providing necessary infrastructure and resources to carry out the database development.

Funding

This work was supported by BIOMASFOR (Z0912003I, Italy) and EC FP7 (BIOSUPPORT, Bulgaria). M.A.F. was supported by a grant from the Spanish Ministerio de Ciencia e Inovación (BFU2009-12022).

References

1. Levinson, G. and Gutman, G.A. 1987, Slipped-strand mispairing: a major mechanism for DNA sequence evolution, *Mol. Biol. Evol.*, **4**, 203–21.
2. Roorkiwal, M., Grover, A. and Sharma, P.C. 2009, Genome-wide analysis of conservation and divergence of microsatellites in rice, *Mol. Genet. Genomics*, **282**, 205–15.
3. Grover, A. and Sharma, P.C. 2007, Microsatellite motifs with moderate GC content are clustered around genes on Arabidopsis thaliana chromosome 2, *In Silico Biol.*, **7**, 201–13.
4. Field, D. and Wills, C. 1996, Long, polymorphic microsatellites in simple organisms, *Proc. Biol. Sci.*, **263**, 209–15.
5. Vinces, M.D., Legendre, M., Caldara, M., Hagihara, M. and Verstrepen, K.J. 2009, Unstable tandem repeats in promoters confer transcriptional evolvability, *Science*, **324**, 1213–6.
6. Martin, P., Makepeace, K., Hill, S.A., Hood, D.W. and Moxon, R. 2005, Microsatellite instability regulates transcription factor binding and gene expression, *Proc. Natl. Acad. Sci. USA*, **102**, 3800–4.
7. Olmstead, R.G. and Palmer, J.D. 1994, Chloroplast DNA systematics: a review of methods and data analysis, *Amer. J. Bot.*, **81**, 1205–24.
8. Provan, J., Powell, W. and Hollingsworth, P.M. 2001, Chloroplast microsatellites: new tools for studies in plant ecology and evolution, *Trends Ecol. Evol.*, **16**, 142–7.
9. Kmiec, B., Woloszynska, M. and Janska, H. 2006, Heteroplasmy as a common state of mitochondrial genetic information in plants and animals, *Curr. Genet.*, **50**, 149–59.
10. Powell, W., Morgante, M., McDevitt, R., Vendramin, G.G. and Rafalski, J.A. 1995, Polymorphic simple sequence repeat regions in chloroplast genomes: applications to the population genetics of pines, *Proc. Natl. Acad. Sci. USA*, **92**, 7759–63.
11. Rajendrakumar, P., Biswal, A.K., Balachandran, S.M., Srinivasarao, K. and Sundaram, R.M. 2007, Simple sequence repeats in organellar genomes of rice: frequency

and distribution in genic and intergenic regions, *Bioinformatics*, **23**, 1–4.
12. Rajendrakumar, P., Biswal, A.K., Balachandran, S.M. and Sundaram, R.M. 2008, In silico analysis of microsatellites in organellar genomes of major cereals for understanding their phylogenetic relationships, *In Silico Biol.*, **8**, 87–104.
13. Kuntal, H. and Sharma, V. 2011, In silico analysis of SSRs in mitochondrial genomes of plants, *OMICS*, **15**, 783–9.
14. Kuntal, H., Sharma, V. and Daniell, H., 2012, Microsatellite analysis in organelle genomes of Chlorophyta, *Bioinformation*, **8**, 255–9.
15. Lim, K.G., Kwoh, C.K., Hsu, L.Y. and Wirawan, A. 2012, Review of tandem repeat search tools: a systematic approach to evaluating algorithmic performance, *Brief Bioinform.*, doi: 10.1093/bib/bbs023.
16. Filiz, E. and Koc, I. 2012, In Silico chloroplast SSRs mining of Olea species, *Biodiversitas*, **13**, 114–7.
17. Li, F. and Herrin, D.L. 2002, FUGOID: functional genomics of organellar introns database, *Nucleic Acids Res.*, **30**, 385–6.
18. Heinze, B. 2007, A database of PCR primers for the chloroplast genomes of higher plants, *Plant Methods*, **3**, 4.
19. O'Brien, E.A., Zhang, Y., Wang, E., et al. 2009, GOBASE: an organelle genome database, *Nucleic Acids Res.*, **37**, D946–50.
20. Ferro, M., Brugière, S., Salvi, D., et al. 2010, AT_CHLORO, a comprehensive chloroplast proteome database with subplastidial localization and curated information on envelope proteins, *Mol. Cell Proteomics*, **9**, 1063–84.
21. Mudunuri, S.B. and Nagarajaram, H.A. 2007, IMEx: imperfect microsatellite extractor, *Bioinformatics*, **23**, 1181–7.
22. Kumar, P., Chaitanya, P.S. and Nagarajaram, H.A. 2011, PSSRdb: a relational database of polymorphic simple sequence repeats extracted from prokaryotic genomes, *Nucleic Acids Res.*, **39**, D601–5.
23. Provan, J., Soranzo, N., Wilson, N.J., Goldstein, D.B. and Powell, W. 1999, A low mutation rate for chloroplast microsatellites, *Genetics*, **153**, 943–7.
24. Lynch, M., Koskella, B. and Schaack, S. 2006, Mutation pressure and the evolution of organelle genomic architecture, *Science*, **311**, 1727–30.
25. Sloan, D.B., Oxelman, B., Rautenberg, A. and Taylor, D.R. 2009, Phylogenetic analysis of mitochondrial substitution rate variation in the angiosperm tribe Sileneae, *BMC Evol. Biol.*, **9**, 260.