PMDBase: a database for studying microsatellite DNA and marker development in plants

Jingyin Yu[1], Komivi Dossa[1,2], Linhai Wang[1], Yanxin Zhang[1], Xin Wei[1], Boshou Liao[1] and Xiurong Zhang[1,*]

[1]Key Laboratory of Biology and Genetic Improvement of Oil Crops, Ministry of Agriculture; Oil Crops Research Institute, the Chinese Academy of Agricultural Sciences, Wuhan 430062, China and [2]Centre d' Etudes Régional pour l'Amélioration de l'Adaptation à la Sécheresse (CERAAS), BP 3320 Route de Khombole, Thiès, Sénégal

ABSTRACT

Microsatellite DNAs (or SSRs) are important genomic components involved in many important biological functions. SSRs have been extensively exploited as molecular markers for diverse applications including genetic diversity, linkage/association mapping of gene/QTL, marker-assisted selection, variety identification and evolution analysis. However, a comprehensive database or web service for studying microsatellite DNAs and marker development in plants is lacking. Here, we developed a database, PMDBase, which integrates large amounts of microsatellite DNAs from genome sequenced plant species and includes a web service for microsatellite DNAs identification. In PMDBase, 26 230 099 microsatellite DNAs were identified spanning 110 plant species. Up to three pairs of primers were supplied for every microsatellite DNA. For 81 species, genomic features of the microsatellite DNAs (genic or non-genic) were supplied with the corresponding genes or transcripts from public databases. Microsatellite DNAs can be explored through browsing and searching modules with a user-friendly web interface and customized software. Furthermore, we developed MISAweb and embedded Primer3web to help users to identify microsatellite DNAs and design corresponding primers in their own genomic sequences online. All datasets of microsatellite DNAs can be downloaded conveniently. PMDBase will be updated regularly with new available genome data and can be accessed freely via the address http://www.sesame-bioinfo.org/PMDBase.

INTRODUCTION

Microsatellites, also known as Simple Sequence Repeats (SSRs), are sequence blocks containing 1 to 6 nucleotide units repeated in tandem and flanked by sequences that are generally unique in the genome, but conserved in organisms (1,2). They are abundantly and distributed throughout the eukaryotic nuclear genome in both coding and non-coding regions (3) and also appear in prokaryotic and eukaryotic organellar genomes (4) and mitochondria (5). SSRs are more abundant in non-coding regions than in exons (6). Their distribution is a function of the dynamics and history of genome evolution and of selective constraints (7). Commonly presumed to be largely neutral, SSR diversity influences many biological characters: biochemical, morphological, physiological and behavioral. In recent years, microsatellites present in genic regions have been demonstrated to have many important biological functions including regulation of chromatin organization, DNA metabolic processes, gene activity, RNA structure, cell cycle, mismatch repair system, among many other functions (8,9), thereby providing a molecular basis for fast adaptation to environmental stresses and changes in both prokaryotes and eukaryotes (10).

According to the type of repeated sequence, SSRs are classified as 'perfect type repeats' (when displaying only perfect repetitions), 'imperfect type repeats' (when the repeated sequence is interrupted by different nucleotides that are not repeated), and 'composite type' (when there are two or more different motifs in tandem) (11). Different taxa exhibit different preferences for SSR types. For example, in plants AG/CT repeats are most abundant (7), while in mammals, A and AC repeats are the most common motifs (12). Moreover, mammal genomes show a higher density of SSRs among eukaryotic genomes. Microsatellites are characterized by a high co-dominant inheritance, reproducibility, multi-allelic variation (13). Due to their high mutation rate ranging from 10^{-3} to 10^{-6} per generation (14), which generally increases with the length of the repeat unit (15), they are highly polymorphic and potentially

the most informative molecular marker with the advantage of easy and low-cost detection by high-throughput PCR-based platforms(16). These features put them as the preferential choice among the available genetic marker systems (e.g. RFLP, RAPD, SSR, AFLP, SRAP and SNP) and provide the foundation for their successful applications in a wide range of fundamental and applicable fields, such as genetic diversity, linkage/association mapping of gene/QTL, marker-assisted selection, variety identification and evolution analysis (17).

Until now, different microsatellite DNA databases have been developed, including UgMicroSatdb (http://veenuash.info/web1/index.htm) (18), EuMicroSatdb (http://veenuash.info/) (19), Kazusa Marker DataBase (http://marker.kazusa.or.jp/) (20), Gramene Markers Database (http://archive.gramene.org/markers/) (21) et al. Although extensive studies have been conducted in identifying microsatellites in different plant species, any of these databases provide comprehensive information about SSRs in plant taxon, because data are heavily fragmented within the small and single-species databases existing. In addition, more and more plant genomes have been sequenced with the rapid development of genome sequencing technology. In March 2016, genome sequencing projects have been completed for more than 120 plant species. The accessibility and data analysis of microsatellite content in whole genome sequence of different species would facilitate comprehensive studies on the direct role of microsatellites in genome organization, recombination, gene regulation, quantitative genetic variation and the evolution of genes (22). In this project, PMDBase is presented as a comprehensive database which systematically integrates large amounts of microsatellites from majority of cultivated or model plant species in genome-wide fashion and includes a user-friendly web service for identifying microsatellite DNAs in plant genomes.

MATERIALS AND METHODS

Data sources

At present, PMDBase has collected 110 plant genome data from species-specific or public comprehensive genomic databases (Supplementary Table S1). These species span not only important model organisms for research community but also grain and economic crops in the world. Out of the 110 plant species, genome data of 53 species were downloaded from species-specific genomic databases, such as the Arabidopsis Information Resource (http://www.arabidopsis.org/) (23) and the *Brassica oleracea* Genome Database (http://ocrigenomics.org/bolbase/) (24) et al. The cotton diploid ancestral species and tetraploid cultivars of cotton were downloaded from Cotton Genome Project (CGP, http://cgp.genomics.org.cn/). *Vitis vinifera* and *Brassica napus* were downloaded from Genoscope (http://www.genoscope.cns.fr/). The two peanut diploid ancestral species were downloaded from PeanutBase (http://www.peanutbase.org/) (25). The three cultivated tobaccos were downloaded from Sol Genomics Network (https://www.solgenomics.net/) (26). Genomic data of the cultivated pepper Zunla-1 (*C. annuum* L.) and its wild progenitor Chiltepin (*C. annuum* var. glabriusculum) were downloaded from The Pepper Genome Database (http://peppersequence.genomics.cn) (27). *Dianthus caryophyllus*, *Solanum melongena*, *Jatropha curcas*, *Lotus japonicas* and *Fragaria ananassa* were downloaded from Kazusa DNA Res. Inst. (KDRI, http://www.kazusa.or.jp). The remaining eukaryotic species were downloaded from public integrated genomic database, such as Ensembl Genomes (http://ensemblgenomes.org/) (28), JGI Genome Portal (http://genome.jgi.doe.gov/) (29) and NCBI Genomes (http://www.ncbi.nlm.nih.gov/genome) (30). Among the remaining species, 5, 15 and 21 species were downloaded from Ensembl Genomes, JGI Genome Portal and NCBI Genomes, respectively. PMDBase extracted the latest or complete version of these available genomic data to identify microsatellite DNAs.

Database implementation

The PMDBase implementation was divided into three steps: (i) data mining of microsatellite DNA and corresponding marker development; (ii) integration and consolidation of microsatellites data information for PMDBase; (iii) PMDBase development (including embedding useful software and developing MISAweb). Firstly, the MISA (MIcroSAtellite identification tool, http://pgrc.ipk-gatersleben.de/misa/) was used to identify normal microsatellite DNAs, as well as compound microsatellite DNAs which were interrupted by a certain number of bases, and offer supplemental Perl tools for designing flanking markers of the microsatellite loci (31). In this study, Perl scripts from MISA were modified for reducing program complexity to identify microsatellite DNAs and designed microsatellite markers with Primer3 software. For normal microsatellite DNAs, the parameters were set for detecting mono, di, tri, tetra, penta and hexa -nucleotide (nt) motifs with a minimum of 10, 6, 5, 5, 5 and 5 repeats, respectively. For compound microsatellite DNAs, the parameters were defined as ≥ 2 repeats interrupted by ≤ 100 bp. Using in-house Perl and Python scripts and modified Perl scripts from MISA, a pipeline was developed for identifying microsatellite DNAs among the 110 genomic sequences. The in-house Perl and Python scripts were used to batch convert all genome data, normalize input/output files, and generate intermediate data for developing PMDBase. At first, we downloaded the 110 genome data from public databases and converted them into consistent formats. Then, modified Perl scripts from MISA were employed along with in-house Perl and Python scripts to obtain localization and types of microsatellite DNAs, statistics of microsatellite DNAs and total microsatellite DNAs with primers from the 110 eukaryotic species. After curation, we obtained 26,230,099 microsatellite DNA entries spanning the 110 eukaryotic species (Supplementary Table S2). For each species, the results of microsatellite DNAs search were summarized as well as the frequency of identified microsatellite motif-types (Supplementary Table S3). For all microsatellite DNAs, up to three pairs of primers were supplied using Primer3 software. In order to determine whether microsatellite DNAs were located in genic region or not, general feature format (GFF) files of genes or transcripts were used together with the positions of microsatellite DNAs on chromosomes or linkage groups in the plant genomes. Due to incomplete

or irregular genomic data for some plant species, only 81 species were analysed for the identification of different types of microsatellite DNAs (genic or non-genic) and the corresponding genes or transcripts, along with the biological functions from public species-specific genomic databases were supplied (Supplementary Table S4) (Figure 1A). Secondly, the datasets of microsatellite DNAs were curated to create a logic relationship among the different types of microsatellite data for their integration in PMDBase. Using in-house Perl and Python scripts, we batch processed intermediate data into different types of data, and then converted them into consistent formats fitting for conformance requirements of relational database systems (Figure 1B). Thirdly, PMDBase was implemented using LMAP (Linux + Apache + Mysql + Perl/PHP/Python) web application program platform. Hypertext Markup Language (HTML) and JavaScript script language were also used to design a user-friendly web interface. BLAST and GBrowse were embedded for users to browse and search microsatellite DNAs conveniently (32,33). Furthermore, the pipeline for identifying microsatellite DNAs was employed to develop a web service, MISAweb, for identifying microsatellite DNAs online. Combining with embedded Primer3web (34), users could design primers of microsatellite DNAs from their interested genomic regions or sequences online directly in PMDBase (Figure 1C).

RESULTS

Database organization

In PMDBase, six functional units were designed including 'Home', 'Browse', 'Tools', 'Document', 'Help' and 'Community'. From the 'Browse' functional unit, users can browse the whole information of plant microsatellite DNAs and the corresponding primers in each plant species by alphabetical order of species Latin names or by plant taxonomy. By using table and GBrowse formats, users can also browse all microsatellite DNAs in corresponding plant species and further obtain detail information for each microsatellite DNA. In 'Tools' functional unit, we embedded several customized open resource software, such as BLAST and GBrowse were embedded, which can be used to detect orthologous genomic regions and display microsatellite DNAs in target genomic regions (32,33). Moreover, the user-friendly MISAweb could help users to identify microsatellite DNAs in preferred genomic sequences online while the embedded Primer3web will assist in designing primers for their microsatellite DNAs online. In the 'Help' functional unit, PMDBase offered a user-friendly interface for users to download three types of microsatellite data conveniently, including basic information of microsatellite DNAs, summary of microsatellite DNAs, and microsatellite DNAs with primers. Finally, the 'Community' functional unit supplied important publications and public microsatellite DNA resources (Figure 2A).

Web interface

Browse module to show overall view of microsatellite DNAs. PMDBase supplied multi-layers browse function for users

to browse all information of microsatellite DNAs and corresponding primers. The 110 plant species were organized by two ways, including alphabetical order and taxonomy order. Alphabetical order was mainly based on Latin names of the 110 eukaryotic species. Each species was then displayed in alphabetical order following the first character of its Latin name (Figure 2B). The taxonomy order among the 110 species was implemented based on NCBI taxonomy database (35) and phylogenetic tree of plant species in Phytozome (36). Each species was displayed in taxonomy order following its phylogenetic relationship (Figure 2C). Out of the 110 plant species, only 48 plant genomes were assembled into complete chromosomes or linkage groups and the visual map of their chromosomes or linkage groups were thus provided in PMDBase. In addition, PMDBase offered a simple table format for users to browse all microsatellite DNAs in the 110 plant species, with their locations on chromosomes or scaffolds directly. Hyperlinks were added to the microsatellite DNAs to display detailed information of their key attributes (Figure 2D). GBrowse format is another way to browse all microsatellite DNAs in target genomic regions and then to get detailed information of microsatellite DNAs (33). If microsatellite DNAs were detected in genic regions, users could browse corresponding genes or transcripts in GBrowse software. For each gene or transcript, we added hyperlink to existing genome databases, which supplied biological functions of the gene or transcript and established the connection between microsatellite DNAs and biological functions (Figure 2E). Detailed information of microsatellite DNA contains basic information of microsatellite DNA such as location, type, size and sequence, and primer sequences designed by Primer3 software. In addition, corresponding genes or transcripts and associated biological functions are also provided if applicable (34) (Figure 2F).

Search modules for microsatellite DNAs in PMDBase. Search module is an important branch of PMDBase, which supplied multiple function search modules including keyword, BLAST and GBrowse. For keyword search module, users can get detailed information of microsatellite DNA by its identifier, microsatellite DNAs in genic region by gene or transcript name, and also get microsatellite DNA(s) by species name, different SSR motifs and their location on chromosomes/linkage groups or scaffolds. For BLAST search module, users could use their own genomic or peptide sequence(s) to blast with the corresponding plant genome in PMDBase, and obtain identical or orthologous genomic regions. For GBrowse search module, it could be used to display microsatellite DNAs in genome, and also search microsatellite DNAs in target genomic regions.

MISAweb—a web service for microsatellite identification tool. Dr. Thomas Thiel has developed MISA—MIcroSAtellite identification tool and released several Perl scripts for identifying microsatellite DNAs (31). But the MISA users are supposed to be familiar with Perl programming language. It is difficult for users, who have no background on Perl programming foundation, to run these Perl scripts. So, a web service for identifying microsatellite DNAs in genomic sequences is lacking. In order to overcome this shortage and bring convenience for
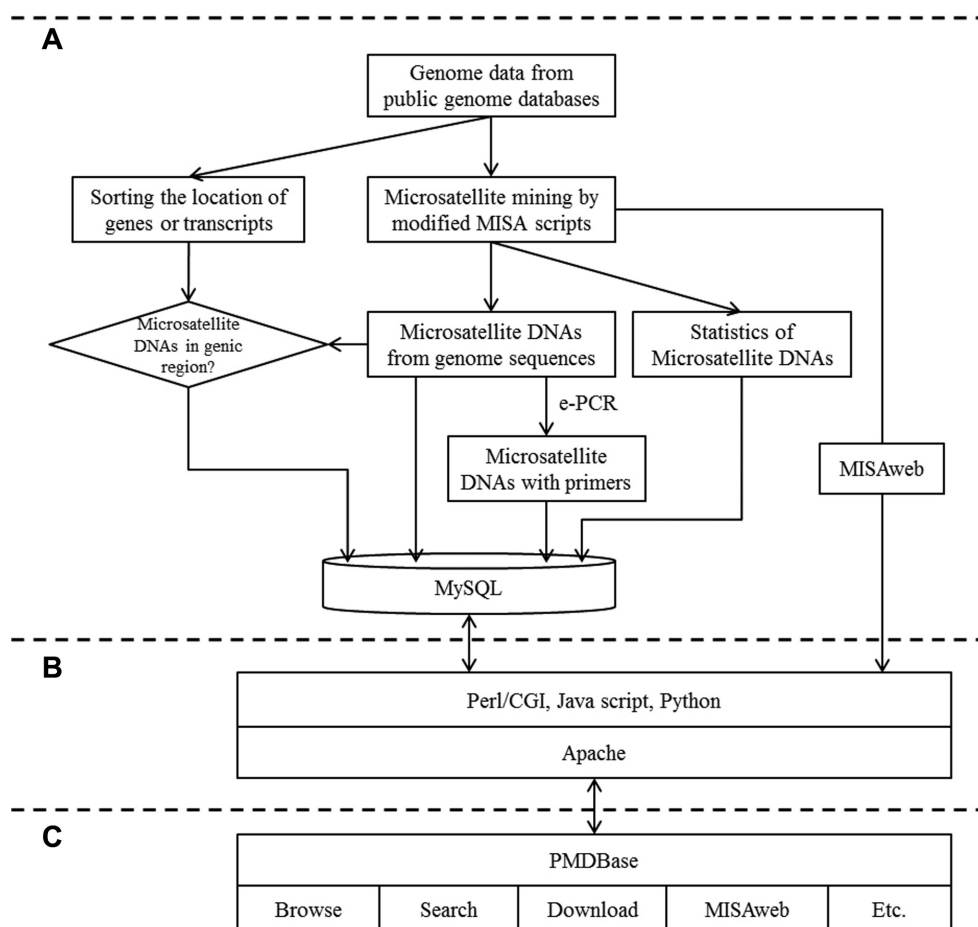
Figure 1. The workflow of PMDBase development. (**A**) Analysis pipeline for generating microsatellite DNAs. (**B**) Implementation of PMDBase. (**C**) Structure of PMDBase.

community, we have developed a user-friendly MISAweb in PMDBase, a web service for microsatellite identification tool, which can help users to easily and directly identify microsatellite DNAs in genomic sequences (FASTA format) online (Figure 3A). In employing embedded Primer3web, users can design primers of identified microsatellite DNAs in interested genomic sequences. And when users submit genomic sequence(s) into MISAweb, PMDBase would give a unique job *id* from server to their submission. Users could then download their microsatellite results directly from MISAweb result page after their job has finished (Figure 3B). Previous MISAweb results would be retained in the server of PMDBase for one week. MISAweb offered a functional module for users to download their previous MISAweb results by supplying the job *id* (Figure 3C).

Usage

Browsing. We took '*Arabidopsis thaliana*' as an example to show how PMDBase can be used. Firstly, users can browse summary of microsatellite DNAs in *Arabidopsis thaliana* through alphabetical order or taxonomy order. By clicking '+/-' (expanded/collapsed) or '<SSRs Details>' tag, result page shows that the genome of *Arabidopsis thaliana* is 119 146 348 bp including five chromosomes, which contains

45 552 microsatellite DNAs. Secondly, users can browse all microsatellite DNAs in *Arabidopsis thaliana* by table or GBrowse formats. In table format, the 45 552 microsatellite DNAs were displayed in 152 pages with 300 entries per page. By clicking identifier of microsatellite DNA, users get detail information of target microsatellite DNA, as well as information related to their presence in genic region or not. In GBrowse format, users will browse all genes and microsatellite DNAs located on Chr1 from 15 000 to 60 000 as defaults. The hyperlink of each gene will connect to the Arabidopsis Information Resource (http://www.arabidopsis.org/) displaying its biological functions. Based on the relative position between genes and microsatellite DNAs displayed in GBrowse, and users can finally judge whether genes contain microsatellite DNAs or not and further retrieve the associated biological functions.

Searching. As mentioned in search modules, we supplied several search options, including keyword search and embedded open-resource tools within PMDBase that can be used to obtain the available data about microsatellite DNAs. For keyword search, we designated *Arabidopsis thaliana* as model species to search useful information about microsatellite DNAs by clicking 'submit' button directly. Users can select their interested plant species instead

Figure 2. Screenshots of navigation and browse modules in PMDBase. (**A**) Navigation bar of PMDBase. (**B**) Organization of plant species by alphabetical order and summary of microsatellite DNAs for corresponding plant species. (**C**) Organization of plant species by taxonomy order. (**D**) Example of microsatellite DNAs as table format. (**E**) Example of microsatellite DNAs following positions on chromosomes or scaffolds in GBrowse. (**F**) Detailed information of microsatellite DNA with primers.

Figure 3. Screenshots of main functional web interface in MISAweb. (**A**) Sequence input web interface. (**B**) Microsatellite DNAs result web interface. (**C**) Previous result retrieving module.

of *Arabidopsis thaliana* to obtain useful information about microsatellite DNAs. For embedded open-resource tools, BLAST can be used to obtain identical or orthologous genomic or peptide sequences by pasting users' interested genomic or peptide sequence(s). Thereafter, users can get microsatellite DNAs in their interested genomic sequences collected in PMDBase. GBrowse can be used to search microsatellite DNAs in target genomic region after inputting its position on chromosomes or scaffolds.

Using MISAweb. PMDBase supplied a simple usage for MISAweb, which will bring convenience for users to get useful information for plant microsatellite DNAs. Users can paste DNA sequence into the text-area directly or upload demo sequence below text-area as shown in Figure 3A. After clicking submit button, the result page will supply a random job id for submission and also present the microsatellite DNA results including total microsatellite DNAs and their statistics as shown in Figure 3B. Users can also download results for their genomic sequence directly. Moreover, since all MISAweb results are retained on the server, users can supply their job id and download their previous results for microsatellite DNA identification as shown in Figure 3C.

DISCUSSION

In PMDBase, we gathered microsatellite DNAs from the majority of cultivated species as well as model plants with the aim of providing a comprehensive platform for microsatellite DNA analysis at single species, taxon or plant kingdom level. The designed microsatellite DNA markers are expected to assist germplasm management, genetic studies and molecular breeding applications leading to increased crop productivity (37). Moreover, the customization for microsatellites DNAs search based on user's preferential locations in the genome may be helpful for gene fine mapping. Based on their genomic features (genic or nongenic), microsatellite DNAs can provide opportunity for biologists to understand their functional significance in regulating gene expression (8,9) and also in targeting some useful genes involved in important traits in plants (38,39). This database is suitable for plant biologists engaged in whole-genome level research like genome organization and evolution studies, microsatellite abundance and its impact on genome size, mutability etc. (22). For plant species with limited genomic resources and not yet included in PMDBase, users can exploit microsatellites DNA resources from relative plant species present in this database with the homologous searching modules or employ MISAweb to identify microsatellite DNAs in their interested regions.

CONCLUSIONS AND PERSPECTIVES

PMDBase is a database for studying microsatellite DNAs and marker development in plants. It is not only a comprehensive database, which has a collection of 26 230 099 microsatellite DNAs from 110 plant species, but also a user-friendly web service for identifying microsatellite DNAs online. Thus, PMDBase can help to show microsatellite DNAs across selected species in a genome-wide fashion and

also identify all putative microsatellite DNAs within interested genomic regions through online web service, MISAweb. PMDBase provides effective searching modules including keyword, BLAST and GBrowse, which can help users to get useful microsatellite DNAs and markers, as well as functional microsatellite DNAs present in genic regions. The database will continuously be updated with newly detected microsatellite DNAs from available plant genomes and its function will be improved and refined with new technologies of database design and development. For example, we are currently developing a genome-wide enhancer map in PMDBase based largely on the microsatellite information, which will be more convenient for displaying microsatellites in PMDBase. We hope to provide systematic data resources and an integrative analytical web service to improve the study of microsatellite DNAs for genome evolution and structure variants, gene regulation, as well as molecular breeding in plants.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

REFERENCES

1. Powell,W., Machray,G.C. and Provan,J. (1996) Polymorphism revealed by simple sequence repeats. *Trends Plant Sci.*, **1**, 215–222.
2. Gupta,P.K. and Varshney,R.K. (2000) The development and use of microsatellite markers for genetic analysis and plant breeding with emphasis on bread wheat. *Euphytica*, **113**, 163–185.
3. Jarne,P. and Lagoda,P.J. (1996) Microsatellites, from molecules to populations and back. *Trends Ecol. Evol.*, **11**, 424–429.
4. Powell,W., Morgante,M., Andre,C., McNicol,J.W., Machray,G.C., Doyle,J.J., Tingey,S.V. and Rafalski,J.A. (1995) Hypervariable microsatellites provide a general source of polymorphic DNA markers for the chloroplast genome. *Curr. Biol.*, **5**, 1023–1029.
5. Soranzo,N., Provan,J. and Powell,W. (1999) An example of microsatellite length variation in the mitochondrial genome of conifers. *Genome*, **42**, 158–161.
6. Hancock,J.M. (1995) The contribution of slippage-like processes to genome evolution. *J. Mol. Evol.*, **41**, 1038–1047.
7. Morgante,M., Hanafey,M. and Powell,W. (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.*, **30**, 194–200.
8. Li,Y.C., Korol,A.B., Fahima,T. and Nevo,E. (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol. Ecol.*, **11**, 2453–2465.
9. Li,Y.C., Korol,A.B., Fahima,T. and Nevo,E. (2004) Microsatellites within genes: structure, function, and evolution. *Mol. Biol. Evol.*, **21**, 991–1007.
10. Kashi,Y. and King,D.G. (2006) Has simple sequence repeat mutability been selected to facilitate evolution? *Israel J. Ecol. Evol.*, **52**, 331–342.
11. Selkoe,K.A. and Toonen,R.J. (2006) Microsatellites for ecologists: a practical guide to using and evaluating microsatellite markers. *Ecol. Lett.*, **9**, 615–629.
12. Toth,G., Gaspari,Z. and Jurka,J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.*, **10**, 967–981.

13. Kalia,R.K., Rai,M.K., Kalia,S., Singh,R. and Dhawan,A. K. (2011) Microsatellite markers: an overview of the recent progress in plants. *Euphytica*, **177**, 309–334.

14. Xu,X., Peng,M. and Fang,Z. (2000) The direction of microsatellite mutations is dependent upon allele length. *Nat. Genet.*, **24**, 396–399.

15. Wierdl,M., Dominska,M. and Petes,T.D. (1997) Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics*, **146**, 769–779.

16. Hoshino,A.A., Bravo,J.P., Nobile,P.M. and Morelli,K.A. (2012) Microsatellites as tools for genetic diversity analysis. In: Caliskan,M (ed). *Microsatellites as Tools for Genetic Diversity Analysis, Genetic Diversity in Microorganisms*. InTech, doi:10.5772/35363.

17. Jiao,Y., Jia,H.-M., Li,X.-W., Chai,M.-L., Jia,H.-J., Chen,Z., Wang,G.-Y., Chai,C.-Y., van de Weg,E. and Gao,Z.-S. (2012) Development of simple sequence repeat (SSR) markers from a genome survey of Chinese bayberry (Myrica rubra). *BMC Genomics*, **13**, 1–16.

18. Aishwarya,V. and Sharma,P.C. (2008) UgMicroSatdb: database for mining microsatellites from unigenes. *Nucleic Acids Res.*, **36**, D53–D56.

19. Aishwarya,V., Grover,A. and Sharma,P.C. (2007) EuMicroSatdb: a database for microsatellites in the sequenced genomes of eukaryotes. *BMC Genomics*, **8**, 1–8.

20. Shirasawa,K., Isobe,S., Tabata,S. and Hirakawa,H. (2014) Kazusa Marker DataBase: a database for genomics, genetics, and molecular breeding in plants. *Breed Sci.*, **64**, 264–271.

21. Tello-Ruiz,M.K., Stein,J., Wei,S., Preece,J., Olson,A., Naithani,S., Amarasinghe,V., Dharmawardhana,P., Jiao,Y., Mulvaney,J. *et al.* (2016) Gramene 2016: comparative plant genomics and pathway resources. *Nucleic Acids Res.*, **44**, D1133–D1140.

22. Katti,M.V., Ranjekar,P.K. and Gupta,V.S. (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol.*, **18**, 1161–1167.

23. Huala,E., Dickerman,A.W., Garciahernandez,M., Weems,D., Reiser,L., Lafond,F., Hanley,D., Kiphart,D., Zhuang,M. and Huang,W. (2001) The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.*, **29**, 102–105.

24. Yu,J., Zhao,M., Wang,X., Tong,C., Huang,S., Tehrim,S., Liu,Y., Hua,W. and Liu,S. (2013) Bolbase: a comprehensive genomics database for Brassica oleracea. *BMC Genomics*, **14**, 1–7.

25. Dash,S., Cannon,E.K.S., Kalberer,S. R., Farmer,A.D. and Cannon,S.B. (2016) PeanutBase and other bioinformatic resources for peanut (Chapter 8). In: Stalker,TH and Wilson,RF (eds).*Peanuts Genetics, Processing, and Utilization(link is external)*. AOCS Press, pp. 241–252.

26. Fernandez-Pozo,N., Menda,N., Edwards,J.D., Saha,S., Tecle,I.Y., Strickler,S.R., Bombarely,A., Fisher-York,T., Pujar,A., Foerster,H. *et al.* (2015) The Sol Genomics Network (SGN)–from genotype to phenotype to breeding. *Nucleic Acids Res.*, **43**, D1036–D1041.

27. Qin,C., Yu,C., Shen,Y., Fang,X., Chen,L., Min,J., Cheng,J., Zhao,S., Xu,M., Luo,Y. *et al.* (2014) Whole-genome sequencing of cultivated and wild peppers provides insights into Capsicum domestication and specialization. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 5135–5140.

28. Kersey,P.J., Allen,J.E., Armean,I., Boddu,S., Bolt,B.J., Carvalho-Silva,D., Christensen,M., Davis,P., Falin,L.J., Grabmueller,C. *et al.* (2016) Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res.*, **44**, D574–D580.

29. Nordberg,H., Cantor,M., Dusheyko,S., Hua,S., Poliakov,A., Shabalov,I., Smirnova,T., Grigoriev,I.V. and Dubchak,I. (2014) The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res.*, **42**, D26–D31.

30. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D. *et al.* (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.

31. Thiel,T., Michalek,W., Varshney,R.K. and Graner,A. (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). *Theor. Appl. Genet.*, **106**, 411–422.

32. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

33. Stein,L.D. (2013) Using GBrowse 2.0 to visualize and share next-generation sequence data. *Brief. Bioinform.*, **14**, 162–171.

34. Untergasser,A., Cutcutache,I., Koressaar,T., Ye,J., Faircloth,B.C., Remm,M. and Rozen,S.G. (2012) Primer3–new capabilities and interfaces. *Nucleic Acids Res.*, **40**, e115.

35. Federhen,S. (2012) The NCBI Taxonomy database. *Nucleic Acids Res.*, **40**, D136–D143.

36. Goodstein,D.M., Shu,S., Howson,R., Neupane,R., Hayes,R.D., Fazo,J., Mitros,T., Dirks,W., Hellsten,U., Putnam,N. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–D1186.

37. Iquebal,M.A., Sarika, Arora,V., Verma,N., Rai,A. and Kumar,D. (2013) First whole genome based microsatellite DNA marker database of tomato for mapping and variety identification. *BMC Plant Biol.*, **2013**, 1–7.

38. Lopez,C.G., Banowetz,G.M., Peterson,C.J. and Kronstad,W.E. (2003) Dehydrin Expression and Drought Tolerance in Seven Wheat Cultivars. *Crop Sci.*, **43**, 577–582.

39. Lata,C. and Prasad,M. (2013) Association of an allele-specific marker with dehydration stress tolerance in foxtail millet suggests SiDREB2 to be an important QTL. *J. Plant Biochem. Biotechnol.*, **23**, 1–4.