D244–D252 *Nucleic Acids Research, 2019, Vol. 47, Database issue* *Published online 26 October 2018*
doi: 10.1093/nar/gky998

SSRome: an integrated database and pipelines for exploring microsatellites in all organisms

Morad M. Mokhtar and Mohamed A. M. Atia *

Molecular Genetics and Genome Mapping Laboratory, Genome Mapping Department, Agricultural Genetic Engineering Research Institute (AGERI), ARC, Giza, 12619, Egypt

Received August 14, 2018; Revised September 28, 2018; Editorial Decision October 08, 2018; Accepted October 14, 2018

ABSTRACT

Over the past decade, many databases focusing on microsatellite mining on a genomic scale were released online with at least one of the following major deficiencies: (i) lacking the classification of microsatellites as genic or non-genic, (ii) not comparing microsatellite motifs at both genic and non-genic levels in order to identify unique motifs for each class or (iii) missing SSR marker development. In this study, we have developed 'SSRome' as a web-based, user-friendly, comprehensive and dynamic database with pipelines for exploring microsatellites in 6533 organisms. In the SSRome database, 158 million microsatellite motifs are identified across all taxa, in addition to all the mitochondrial and chloroplast genomes and expressed sequence tags available from NCBI. Moreover, 45.1 million microsatellite markers were developed and classified as genic or non-genic. All the stored motif and marker datasets can be downloaded freely. In addition, SSRome provides three user-friendly tools to identify, classify and compare motifs on either a genome- or transcriptome-wide scale. With the implementation of PHP, HTML and JavaScript, users can upload their data for analysis via a user-friendly GUI. SSRome represents a powerful database and mega-tool that will assist researchers in developing and dissecting microsatellite markers on a high-throughput scale.

INTRODUCTION

Microsatellites, also referred to as simple sequence repeats (SSRs) or short tandem repeats (STRs), are sequence motifs containing 1–6 nucleotide units repeated in tandem patterns and classified into perfect repeats, imperfect repeats or composite repeats based on the composition of the repeats (1). Microsatellites are co-dominant markers. SSRs are inherited in a Mendelian fashion and are universally scattered within all organisms' genomes. SSR frequencies and distributions vary among prokaryotic (2) and eukaryotic genomes (3), as well as between coding (genic) and non-coding (non-genic) regions (4,5). Their flanking sequences are often conserved within the same species and even across species (6). SSRs are involved in many important biological functions, including the modulation of transcription factor binding and enhancer function (7), the adjustment of gene expression as 'tuning knobs' (8), chromatin organization (9), nucleosome positioning (10), the modulation of mRNA stability (11), acting as meiotic recombination hot spots (12), cytosine methylation, alternative splicing and unusual structural conformations (13). SSRs thereby provide a molecular basis for genome evolution and rapid adaptation to environmental stresses in both eukaryotes and prokaryotes (14).

Due to their characteristics, which include high abundance, hyper-variability (with a high mutation rate ranging from 10^{-3} to 10^{-6} per generation), multi-allelic nature, high reproducibility and detection simplicity, microsatellites have become the preferred choice among all other genetic marker techniques, such as AFLP, SCoT, RAPD, ISSR and SNP (15). Over the past decades, they have been extensively used in a variety of fundamental and applied biological sciences for prokaryote, plant and animal studies. Microsatellites have also been extensively exploited as genetic markers for diverse applications including genetic diversity assessment, genome mapping, genetic material characterization/identification, population genetics and genome evolution studies (16).

With the development of genome sequencing technologies, hundreds of completed or drafted genomes and transcriptomes have been decoded and released to date. These published genomes/transcriptomes, in addition to the available bioinformatic tools, provide an important opportunity for scientists to identify microsatellite motifs that could be used to develop useful SSR and EST-SSR markers (17).

Over the last decade, many bioinformatic tools were developed for *in silico* SSR mining, such as TRF (18), MISA (19), SciRoko (20), GMATo (21), IMEx (22), mreps (23), TROLL (24) and MsDetector (25), as well as for marker development, such as SSRLocator (26), SSRPoly (27) and CandiSSR (28). However, these tools/pipelines

*To whom correspondence should be addressed. Tel: +20 100 016 4922; Fax: +20 235 731 574; Email: matia@ageri.sci.eg, matia_ageri@yahoo.com

usually have at least one of the following major limitations: (i) insufficient processing capability when analyzing complete genomes, (ii) platform dependencies, (iii) time-consuming processes, especially pipelines that possess multiple functions and integrate different software, (iv) the lack of a graphical interface or (iv) missing marker design. Therefore, the use of these tools represents an impediment for non-bioinformatician users. Although recent tools/pipelines such as GMATA (29) were developed to overcome previous limitations and provide multiple functions such as SSR mining, characterization of SSR distribution and SSR marker design on a genome scale, these methods still have some drawbacks. These drawbacks include the lack of SSR classifications as genic or non-genic and the lack of microsatellite motifs comparison on both the genic and non-genic level in order to identify unique motifs for each class.

To date, with the aid of the above-mentioned bioinformatic tools/pipelines, different microsatellite databases have been developed and publicly released, including UgMicroSatdb (30), EuMicroSatdb (31), the Kazusa marker database (32) and the Gramene markers database (33) (Supplementary Table S1). However, most of these databases either are specific to a certain species/taxon or have out-of-date content due to irregular updating. Recently, MSDB (34) was developed as a comprehensive database for SSR mining across a large number of species, providing interactive controls and plots that depict genome-wide trends in SSRs via concise and intuitive charts. Despite all of the above-mentioned advantages of MSDB, it still has critical drawbacks, including the following: (i) it does not classify the identified SSRs for each organism/species into genic and non-genic sets, (ii) it is missing comparisons of SSR motifs in both genic and non-genic regions to identify unique motifs for each class, (iii) it does not provide SSR mining for sequenced mitochondrial and chloroplast genomes and (iv) it does not provide user support for any pipelines for genome/transcriptome-scale SSR mining, comparative analysis or marker development.

In this project, SSRome is presented as a web-based, user-friendly, integrated, comprehensive and dynamic database that includes interactive pipelines for identifying, classifying, comparing and developing microsatellite markers on the genome- or transcriptome-wide scale.

DATA COLLECTION, ANALYSIS AND DATABASE CONSTRUCTION

Data collection

Nearly, all of the completely annotated and sequenced genomes of various species across prokaryotes, plants, metazoans and human as well as mitochondrial and chloroplast genomes were downloaded in both FASTA and GenBank formats from the FTP site of NCBI (NCBI Resource Coordinators 2018). In cases where many versions were available for the same genome, the most recently modified version was downloaded. All organisms were then classified and organized based on their taxonomic group. In addition, by March 2018, all expressed sequence tag (EST) records available on the NCBI dbEST (76 644 005 sequences) were downloaded as batch files that included EST sequences in FASTA format.

SSRome data analysis and database construction

To construct the SSRome core database, we developed a set of 'Perl and Shell' scripts to string all standalone bioinformatic routines together into three main pipelines. These pipelines were implemented to construct genomic, transcriptomic and organelle sub-databases. To establish both the genomic and organelle sub-databases, six major steps were designed as follows: (i) mining microsatellites on the genome scale; (ii) classifying SSR motifs into genic or non-genic according to their corresponding location within the genome; (iii) sub-classifying these motifs into unique genic, unique non-genic and shared SSR motifs; (iv) designing SSR primers and developing markers; (v) preparing and integrating the developed SSR markers and their associated information into the SSRome database and (vi) implementing all generated datasets into the SSRome web interface. For transcriptomic sub-databases, the same workflow was used with the exception of the steps classifying identified SSRs into genic/non-genic and sub-classifying the unique genic/non-genic/shared SSR motifs (Figure 1).

In this study, the MISA (MIcroSAtellite identification) tool (19) was initially used for the identification and localization of perfect and compound SSR motifs according to the following parameters: mono- ($\geq$10), di- ($\geq$6), tri- ($\geq$5), tetra- ($\geq$4), penta- ($\geq$3) and hexa-nucleotide ($\geq$3). For compound SSR motifs, the parameters were set as $\geq$2 repeats interrupted by $\leq$100 bp. The generated results were handled using in-house-developed Perl scripts, and the generated results were compared against genome annotations to classify the identified SSR motifs into genic and non-genic regions.

For efficient marker development, we used a simple strategy to enhance the speed of data analysis and results generation. Before we began, we developed in-house Perl scripts that were used to convert all curated genome/transcriptome/EST data and unify the input/output files, which were used for SSR mining and SSRome database development.

First, we indexed all identified SSR motifs after classification into genic and non-genic regions. This classification was done with the aid of general feature format (GFF) files of the genes/transcripts in order to determine the positions of microsatellites within each genome. Then, based on the start and stop coordinates within the analyzed sequence, a short flanking sequence of each SSR locus was extracted at a user-controlled length (default 200 bp) instead of using the full-length DNA/RNA sequence. Thereafter, SSR primers were designed for all sub-classified SSR motifs using in-house-developed Perl scripts combined with Primer3 software (35), and a unique marker ID was then assigned. In the case of genome-scale SSR mining, another level of sub-classification was achieved by comparing the identified genic and non-genic SSR motifs in order to classify them into unique motifs for each class and into motifs shared between the two classes. The complete information for each motif and its developed primers were saved in tabulated file format. The file contained the primer ID, region (genic or non-genic), NCBI accession number of the sequence con-

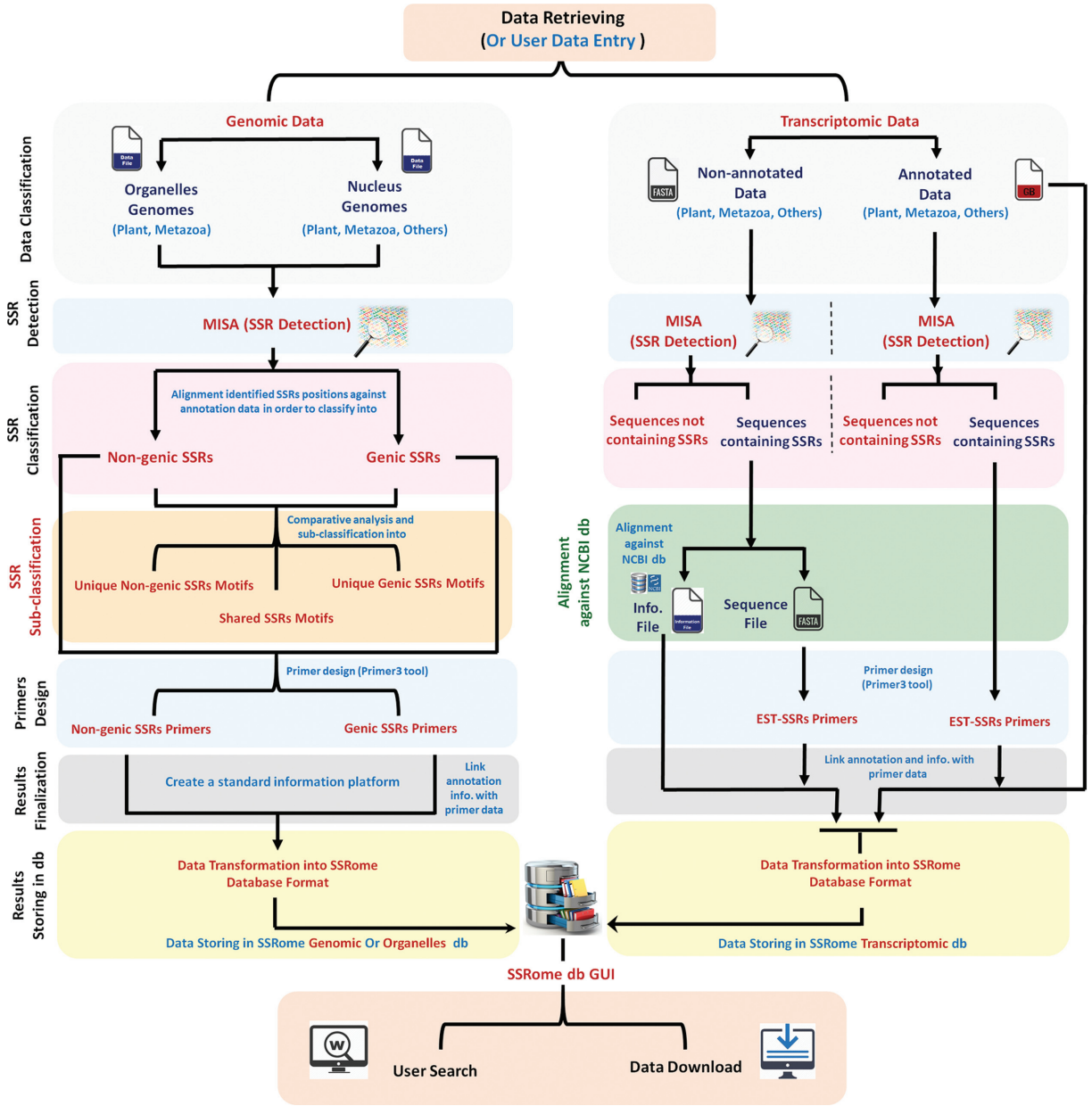

Figure 1. The workflow of SSRome database development: Genomic data analysis and Transcriptomic data analysis pipelines.

taining this motif, repeat type, repeat sequence, primer sequences, primer Tm and product size. Second, all generated datasets of SSR markers were curated to create a standard information platform among all classes and sub-classes of SSRs for integration into the SSRome SQL database. With the aid of in-house-developed Perl scripts, we batch processed all intermediate data generated from the different types of data and then converted the data into a consistent format. Third, SSRome was implemented using the LPPM (Linux + Perl + PHP + MySQL) web application platform. In addition, the JavaScript script language, CSS and Hypertext Markup Language (HTML) were also used to design a user-friendly web interface.

The aforementioned developed SSRome pipelines are integrated as free online tools on the SSRome website and are divided into: the SSRome Genomic Pipeline, the SSRome Transcriptomic Pipeline and the SSRome Comparative Analysis Pipeline, supported with web user interface, local processing capability and database storage.

SSROME FEATURES

SSRome database interface

The SSRome database offers a user-friendly interactive web interface with multiple features to explore, search, download and compare microsatellites across all organisms. The SSRome website provides a top navigation bar designed to help users access the different sections of the SSRome database and tools in a very convenient and responsive way. The SSRome data can be easily accessed and retrieved

via five interactive pages: Homepage (SSRome-db Quick Access), Statistics, Download, Search and SSRome Comparisons. On all five pages, organisms are taxonomically grouped to make the selection and exploration easier and convenient. Users can simply browse any section of the SSRome database by making a selection from the main navigation bar drop-down menus.

The Homepage introduces SSRome as an integrated database along with its pipelines by giving highlights of all of the sections, as well as offering an 'SSRome-db Quick Access' option for all analyzed genomes and ESTs. In the SSRome-db Quick Access, organisms are categorized first according to type (nuclear genome, organellar genome or EST) and then according to genera (metazoa, plant, archaea, bacteria, virus, fungus or protozoa) in an interactive manner, which enables direct access and searches within each organism independently. The Statistics' page was developed initially to provide users with a preliminary indication of the number of developed SSR markers stored in the SSRome database for each organism (Figure 2).

On the Download page, users can retrieve all downloadable files for each organism/genome (separately or in batch) in a very simple way. The downloadable files, which are compressed separately, include genic SSR motifs, genic SSR statistics, unique genic SSRs, genic SSR primers, non-genic SSR motifs, non-genic SSR statistics, unique non-genic SSRs, non-genic SSR primers and SSR shared motifs.

The Search page/hub of SSRome allows users to easily access and retrieve all data stored in the SSRome database. This page gives users the possibility of searching within three main categories: (i) genomic database search utilities, (ii) transcriptomic database search utilities and (iii) organelle database search utilities. Under each one of the above-mentioned search utilities, searches are divided into separate pages according to the taxonomic group (Metazoa, Plant and Others). The 'Others' search utility involves different sub-taxonomic groups classified as archaea, bacteria, viruses, fungi and protozoa. In all search utilities, users can easily customize the available options according to their interest. One of the most effective and powerful features of the SSRome search hub is the possibility to search and explore microsatellites independently within genic or non-genic regions for each organism. Under each search utility, users can easily obtain the results by inputting one of the following types of interest keywords: repeat sequence (e.g. 'AGA_6'), gene symbol (e.g. SOS4), locus tag (e.g. 'LOC109706734') or primer ID (e.g. 'VigSSR_8'). Keywords are not case-sensitive, but they are sensitive to any spelling mistakes. Simultaneously, users are required to select an organism of interest (e.g. '*Homo sapiens*') and an SSR microsatellites repeat type (e.g. 'Tetra') from the available drop-down menus. For each parameter, an example is given inside the text box below.

The search results of all the main categories are presented in a user-friendly tabulated view containing the related important information for each SSR motif/marker (e.g. Primer ID, Organism Name, NCBI Accession No., Repeat Type, Repeat Sequence, Primer Sequence, Primer annealing temp., Primer position within genome, Product Length, Gene symbol or Locus tag etc.).

When comparing the SSRome database to previously developed biological databases focused on SSR mining (30–34), we can clearly see that SSRome has some unique features that are not present/available anywhere else. The SSRome Motif Comparisons page presents one of these unique features of the SSRome database. This page allows users to compare the prevalence and distribution of SSR motifs between genic and non-genic regions within any organism belonging to the plants or metazoans. This comparative analysis is presented in different customizable sections. Under each search section, users can easily obtain the results by selecting the 'organism name' and 'repeat type' of interest from the available drop-down menus under any of the following sections: (i) detection of shared SSR motifs between genic and non-genic regions, (ii) show the genic motifs statistics, (iii) show the non-genic motifs statistics, (iv) identify the unique motifs of genic regions and (v) identify the unique motifs of non-genic regions. The SSRome Motif Comparisons utility page can effectively eliminate the ambiguity of why some motifs only appear within genic regions while other motifs are unique to non-genic. This page will also provide a deep understanding of the evolutionary trends of SSR motifs within genomes. All the above presented features and other features are summarized in Table 1.

SSRome tools and utilities

The classical schemes of generating SSR markers from genomic libraries (36) are being replaced rapidly by modern *in silico* SSR mining approaches. Interestingly, the distribution and frequency of SSRs within genomes are greatly variable, in part because the effectiveness of SSR mining relies on many factors, such as the SSR mining tool utilized, the mining settings and the size of the analyzed genome (1,4,5). Any alteration in the mining criteria usually results in highly significant deviations in the identified number of SSR motifs in a given organism using the same genome/dataset. Despite the many modern and sophisticated SSR mining tools/pipelines that have been developed and released, a commonly observed drawback is that none of these tools provide a deep classification of SSRs into genic and non-genic regions or compare the microsatellite motifs between genic and non-genic regions in order to identify unique motifs for each class.

To overcome the above-mentioned drawbacks, SSRome provides a set of developed tools (genomic, transcriptomic and comparative) as free pipelines implemented into the SSRome website, which comprises a graphical-user interface, local processing and database storage availability. These tools collectively facilitate *in silico* microsatellite mining on the genome/transcriptome scale, SSR marker development and comparative SSR motif analysis. Each tool provides fully independent and easy-to-use control parameters according to the user's preference.

On the familiar and interactive 'SSRome Upload Files' page, users must first upload their raw data (genomic or transcriptomic) using the form in the upload page. In cases where a user's raw data are small in size and/or divided into many files (each chromosome in a separate file or genomes in separate contig files), users should compress all

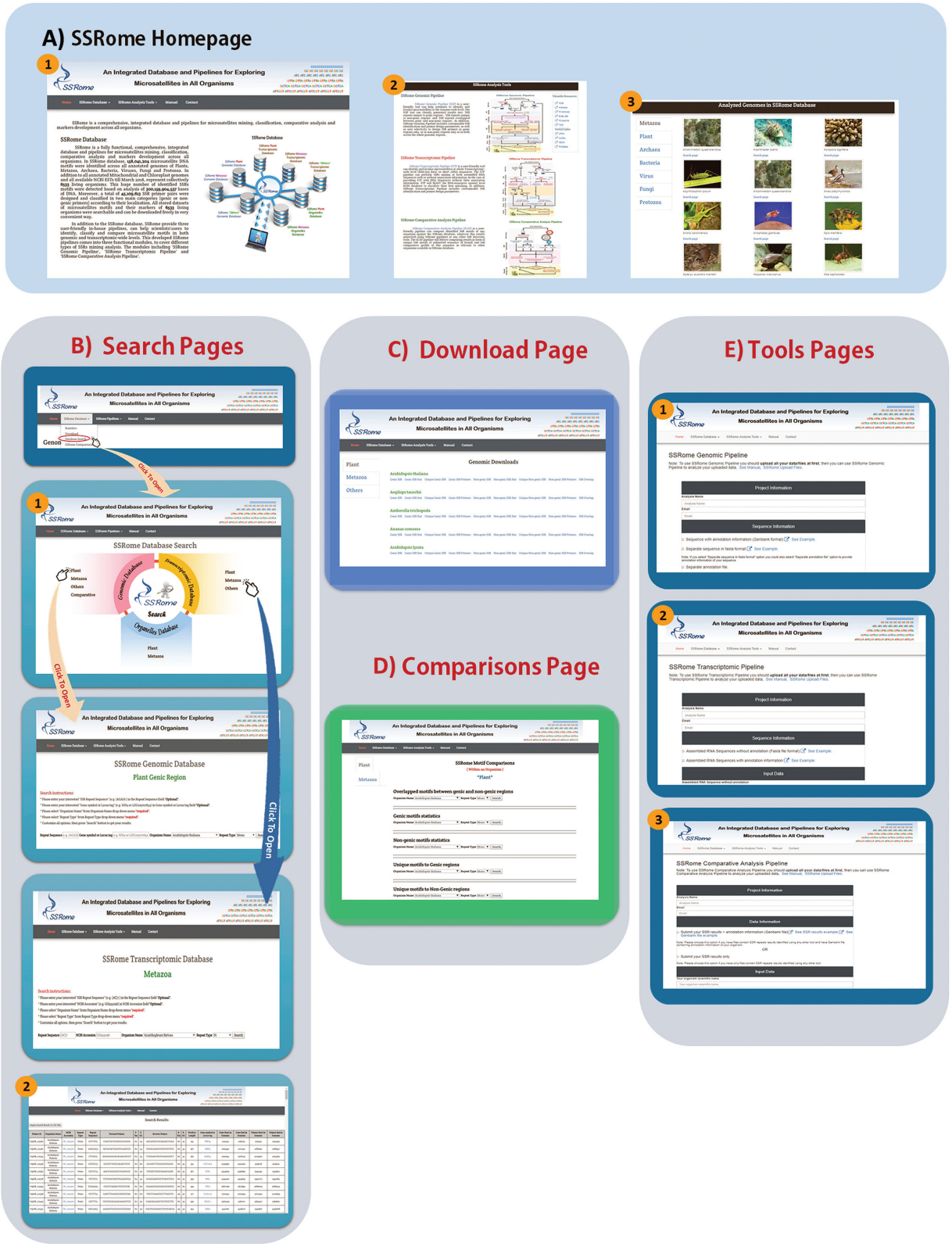

Figure 2. Screenshots of the SSRome database (**A**) SSRome homepage; (**B**) (1) SSRome search pages and (2) example of search results; (**C**) Download page; (**D**) Comparisons page and (**E**) SSRome Tools pages (1) SSRome Genomic Pipeline page, (2) SSRome Transcriptomic Pipeline page and (3) SSRome Comparative Analysis Pipeline page.

Table 1. Comparison of SSRome database with other SSR Databases in term of (i) number of species and (ii) database features

	Microorganism tandem repeats database	UgMicro SatDb	Kazusa marker database	Plant Microsatellite DNAs database	Tandem repeats database	FishMicro Sat	Polymorphic simple sequence repeats database	MICAS	EuMicroSat Db	MSDB	SSRome*
(A) Number of species											
Plant - (Nucleus Genome)	0	80	14	110	2	0	0	0	31	74	98
Plant - (Chloroplast Genome)	0	0	0	0	0	0	0	0	0	0	1034
Plant - (Mitochondrial Genome)	0	0	0	0	0	0	0	0	0	0	120
Metazoa - (Nucleus Genome)	0	160	0	0	18	190	0	0	62	310	137
Metazoa - (Mitochondrial Genome)	0	0	0	0	0	0	0	0	0	0	2669
Fungi	0	80	0	0	1	0	0	0	31	191	241
Archaea	91	0	0	0	0	0	0	217	0	514	125
Bacteria	1109	0	0	0	1	0	85	4772	0	5732	2828
Viruses	1463	0	0	0	0	0	0	0	0	0	1270
Protozoa	0	80	0	0	0	0	0	0	31	72	78
ESTs (Metazoa, Plant and Others)	0	0	0	0	0	0	0	0	0	0	1637
(B) Database features											
Active Database (Availability)	No	No	Yes	Yes	Yes	Yes	No	Yes	No	Yes	Yes
Database Link	http://minisatellites.u-psud.fr/	http://veenuash.info/web1/index.htm	http://marker.kazusa.or.jp/	http://www.sesame-bioinfo.org/PMDBase/index.html	https://tandem.bu.edu/cgi-bin/trdb/trdb.exe?taskid=0	http://mail.nbfgr.res.in/fishmicrosat/	http://www.cdfd.org.in/PSSRdb/	http://www.mcr.org.in/micas/	http://www.veenuash.info/	http://tdb.ccmb.res.in/msdb/	http://mggm-lab.easyomics.org/
Compare SSR motifs at both genic and non-genic levels within the same genome	No	No	No	No	No	No	No	No	No	No	Yes
Identify unique motifs for each genic and non-genic class	No	No	No	No	No	No	No	No	No	No	Yes
Data Download	No	Yes	Yes	Yes	Yes	No	Yes	No	No	Yes	Yes

*The number of genomes in the SSRome database represents all annotated genomes present in RefSeq. For the other databases, this number may represent annotated or draft genomes.

files into one archive before uploading the data. Additionally, in cases where the upload process cannot be successfully completed, users can utilize the SSRome FTP account (available for free) to upload their raw data. Otherwise, if the raw data are stored in large files, users can use one of the three recommended tools (Cyberduck, FileZilla and CoffeeCup FTP) available on the upload page to upload their files.

Once the raw data have been uploaded, a notification email will be sent to indicate that the upload was successful. Upon completion of data uploading, users can effectively proceed to the configuration page of any one of the SSRome tools (genomic, transcriptomic or comparative). SSRome provides users with various customizable parameters for every step within all SSRome tools. In particular, the 'SSRome genomic pipeline' page powerfully supports user scanning of SSRs on a genome-scale level, SSR classification into genic and non-genic region, unique and shared motifs determination, genome-wide SSR marker development and comparative SSR motif analysis. Once the pipeline configuration is done, users can launch the *in silico* analysis by clicking the 'Submit' button. After the completion of the analysis process, users will again be notified by email. Moreover, SSRome gives users the availability to store and integrate their generated/analysed results (in the case of a genome/transcriptome analysis of a new organism) into the SSRome database in a very convenient way.

The 'SSRome transcriptomic pipeline' page allows users to mine microsatellites on a transcriptome-wide scale, develop SSR markers on a transcriptome-wide scale and annotate sequences with successfully designed SSR primers. Once the pipeline parameters (sequence information, SSR mining, SSR primer design, RNA annotation and results availability and storage) are configured, users can initiate the analysis cascade by clicking the 'Submit' button, and upon analysis completion, users will be notified by email. The pipeline also offers users the availability to store or in-

Table 2. Summarization of the whole analyzed data in SSRome database across all organisms

	Number of organisms	Total size of examined genomes (bp)	Total number of identified SSRs (motif)	Total number of primers
Plant - (Nucleus Genome)	98	60 916 941 772	26 191 694	9 269 346
Plant - (Chloroplast Genome)	1034	338 174 520	125 548	18 169
Plant - (Mitochondrial Genome)	120	47 769 383	8746	26 045
Metazoa - (Nucleus Genome)	137	181 365 452 299	106 547 578	30 916 347
Metazoa - (Mitochondrial Genome)	2669	92 247 604	18 887	11 076
Archaea	125	862 936 874	15 867	5523
Fungi	241	7 030 771 349	1 660 819	1 577 469
Bacteria	2828	10 066 003 314	98 876	67 050
Virus	1270	194 822 180	13 960	8119
Protozoa	78	3 334 296 655	2 795 751	1 134 578
ESTs (Metazoa, Plant and Others)	1637	35 910 488 587	20 563 478	2 075 891
Total	10 237*	300 159 904 537	158 041 204	45 109 613

*The real number after removing duplication among the nucleus, mitochondrial, chloroplast genomes and ESTs was 6533 organisms. The removal of duplication aims at representing each organism/species once.

tegrate their results into the SSRome database in a user-friendly way.

Finally, the SSRome Comparative Analysis Pipeline page offers users a powerful deep-mining tool to compare the prevalence, motif uniqueness and motif sharing between the user's genome of interest and the genome of any other organism available in the SSRome database. Likewise, upon completion of the comparative analysis, users will be notified via email. With all tools/pipelines, users can download their results retained on the SSRome server in a very easy way.

STATISTICS AND SUMMARY

As of March 2018, the SSRome database consists of 158 million microsatellite motifs that were identified across all annotated genomes of plants, metazoa, archaea, bacteria, viruses, fungi and protozoa. This dataset is in addition to all annotated mitochondrial and chloroplast genomes, as well as all NCBI ESTs available to date, which collectively represent 6533 living organisms. This large number of identified SSR motifs was detected based on the analysis of 300.16 gigabases of DNA. Moreover, a total of 45.1 million SSR primers were designed and classified as genic or non-genic according to their localization. This large number of developed SSR primers has been supplemented with all related important information (e.g. marker ID, repeat seq., repeat type, PCR conditions, position within genome and cross-reference with NCBI) to provide a modern version of SSR markers called 'Annotated SSR markers'. All generated microsatellite motif datasets and their developed markers for all 6533 living organisms are stored in different backend tables (sub-databases) based on the taxonomic group. All of these data are accessible, searchable and downloadable freely in a very convenient way via the SSRome web interface. The statistics of all the data categories analyzed within the SSRome database (e.g. number of organisms, total size of examined genomes in bp, total number of identified SSRs and total number of developed primers/markers) are summarized in Table 2.

In summary, we present 'SSRome' as a comprehensive, integrated database and set of pipelines for microsatellite mining, classification, comparative analysis and marker development across all organisms. To our knowledge, SSRome is the first repository to provide and offer unique features not present in any previous database. SSRome has the advantage of classifying and comparing the SSR motifs within the same genome and between genic and non-genic regions in order to determine the unique motifs for each class (genic or non-genic), in addition to defining the motifs shared between classes. This functionality can allow researchers to focus on deeper levels of SSR mining and the functional roles of SSRs within genomes and may provide the answers to many unanswered questions in the future. In addition to the SSRome database, SSRome provides three user-friendly in-house tools/pipelines that can help scientists to identify, classify and compare microsatellite motifs on both the genomic and transcriptomic levels. These developed tools come in three functional modules to cover different types of SSR mining analyses. These modules include the 'SSRome Genomic Pipeline', the 'SSRome Transcriptomic Pipeline' and the 'SSRome Comparative Analysis Pipeline'. SSRome is freely available to all users via the web and does not require any login or registration. Moreover, the SSRome website provides sufficient help material for users to ensure ease of use by first-time visitors (Supplementary Data S2).

INSIGHTS AND PROSPECTIVE WORKS

Few publications on the origins of some scientific terms such as 'genome' and 'microsatellites' and some of their implications are available. The word 'microsatellite' was first coined in the late 1980s by the American scientists Litt and Luty (37). Microsatellites, which account for ∼3% of the human genome (two times the percentage used for coding proteins), were reported to play a vital functional role in various biological processes (16,38,39). The appearance or disappearance of certain SSR motifs in the genome is apparently governed by one or more kinds of effects, such as point mutations, DNA polymerase slippage and any other activities involved in chromatin reorganization or nucleosome positioning (40).

For many decades, microsatellites represented the most common source of genetic markers, as well as serving as a cornerstone for different kinds of genetic studies due to their unique and powerful features. SSRs have proven

their effectiveness in a wide variety of applications, including genetic diversity, forensics, genome mapping, parentage identification, population genetics and molecular phylogeny (41,42). Recently, there have been numerous studies aimed at exploring the origin, distribution, functions and dynamics of microsatellites in both prokaryotic and eukaryotic genomes. *In silico* SSR mining on the genome scale has clearly advanced our understanding of the evolutionary mechanisms leading to the formation of microsatellite motifs in the genome (43). Simultaneously, the availability of genome-wide functional data (transcriptomic or ESTs) and modern high-throughput technologies will help scientists to find further evidence and elucidate the functional mechanisms of microsatellites (e.g. the identification of molecular networks causing the alterations in expression).

Due to the above-mentioned ultimate importance of microsatellites and considering the concept that anything can be 'OMEfied', the term 'SSRome' was coined as a new member in the '-OME' family 2 years ago by the corresponding author (Dr. Mohamed Atia) to supplement the current biology-related OMEs (genome, transcriptome, exome, proteome etc.). The 'SSRome' can be defined as the complete set or profile of SSR motifs within a certain genome and the quantity, classification, distribution and function of these motifs. Due to the fast mutation rate and hyper-variable features of microsatellites, the SSRome is dynamic and can give an in-depth representation of the evolutionary state within the genome.

From this point of view, it is worthwhile to clarify how our developed SSRome database is different from and substantially better than similar available resources. SSRome provides new insight with a focus on high-resolution SSR mining through the sub-categorization and comparative analysis of SSR motifs between coding and non-coding parts of the genome.

The SSRome database will continuously be updated with SSR mining of newly released genomes, transcriptomes and ESTs. Furthermore, the database design and functions will be regularly refined, improved and supported with new technologies. For example, currently, we are developing a new section within our SSRome database called 'Pub-SSR', which aims to link and list all original publications completed or related to the SSRs for each organism. This tool is expected to represent a hub, facilitating research in the microsatellite field.

Overall, we believe that the SSRome database will represent a cornerstone or starting point for SSRome research. We also believe that it is of great interest to a wide range of scientists (e.g. medical, environmental, agricultural, industrial etc.) in different disciplines, including population genetics, genetic diversity, genome evolution, genome mapping, genome-wide analysis and species identification, particularly to researchers working on SSR mining on the genomic level, molecular marker development or targeted trait improvement.

DATA AVAILABILITY

SSRome is an online free access database initiative available in the following link http://mggm-lab.easyomics.org.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank the administration of Agricultural Genetic Engineering Research Institute (AGERI) as well as the administration of the Agricultural Research Center (ARC) for their continued support.

FUNDING

The author(s) received no specific funding for this work. Funding for open access charge: Dr. Mohamed Atia, Principal Investigator.
Conflict of interest statement. None declared.

REFERENCES

1. Ellegren,H. (2004) Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.*, **5**, 435–445.
2. Kassai-Jáger,E., Ortutay,C., Tóth,G., Vellai,T. and Gáspári,Z. (2008) Distribution and evolution of short tandem repeats in closely related bacterial genomes. *Gene*, **410**, 18–25.
3. Tóth,G., Gáspári,Z. and Jurka,J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.*, **10**, 967–981.
4. Jarne,P. and Lagoda,P.J. (1996) Microsatellites, from molecules to populations and back. *Trends Ecol. Evol.*, **11**, 424–429.
5. Morgante,M., Hanafey,M. and Powell,W. (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.*, **30**, 194–200
6. Varshney,R.K., Sigmund,R., Börner,A., Korzun,V., Stein,N., Sorrells,M.E., Langridge,P. and Graner,A. (2005) Interspecific transferability and comparative mapping of barley EST-SSR markers in wheat, rye and rice. *Plant Sci.*, **168**, 195–202.
7. Martin,P., Makepeace,K., Hill,S.A., Hood,D.W. and Moxon,E.R. (2005) Microsatellite instability regulates transcription factor binding and gene expression. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 3800–3804.
8. Rockman,M.V. and Wray,G.A. (2002) Abundant raw material for cis-regulatory evolution in humans. *Mol. Biol. Evol.*, **19**, 1991–2004.
9. Makova,K.D. and Hardison,R.C. (2015) The effects of chromatin organization on variation in mutation rates in the genome. *Nat. Rev. Genet.*, **16**, 213–223.
10. Gymrek,M., Willems,T., Guilmatre,A., Zeng,H., Markus,B., Georgiev,S., Daly,M.J., Price,A.L., Pritchard,J.K., Sharp,A.J. and Erlich,Y. (2016) Abundant contribution of short tandem repeats to gene expression variation in humans. *Nat. Genet.*, **48**, 22–29.
11. Chen,T.M., Kuo,P.L., Hsu,C.H., Tsai,S.J., Chen,M.J., Lin,C.W. and Sun,H.S. (2007) Microsatellite in the 3′ untranslated region of human fibroblast growth factor 9 (FGF9) gene exhibits pleiotropic effect on modulating FGF9 protein expression. *Hum. Mutat.*, **28**, 98–98.
12. Choi,K., Zhao,X., Kelly,K.A., Venn,O., Higgins,J.D., Yelina,N.E., Hardcastle,T.J., Ziolkowski,P.A., Copenhaver,G.P., Franklin,F.C.H. and McVean,G. (2013) Arabidopsis meiotic crossover hot spots overlap with H2A. Z nucleosomes at gene promoters. *Nat. Genet.*, **45**, 1327–1336.
13. Ribeiro,M.M., Teixeira,G.S., Martins,L., Marques,M.R., de Souza,A.P. and Line,S.R.P. (2015) G-quadruplex formation enhances splicing efficiency of PAX9 intron 1. *Hum. Genet.*, **134**, 37–44.
14. 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.
15. Vieira,M.L.C., Santini,L., Diniz,A.L. and Munhoz,C.D.F. (2016) Microsatellite markers: what they mean and why they are so useful. *Genet. Mol. Biol.*, **39**, 312–328.
16. Bagshaw,A.T. (2017) Functional mechanisms of microsatellite DNA in eukaryotic genomes. *Genome Biol. Evol.*, **9**, 2428–2443.
17. Sharma,P.C., Grover,A. and Kahl,G. (2007) Mining microsatellites in eukaryotic genomes. *Trends Biotechnol.*, **25**, 490–498.

18. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
19. 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.
20. Kofler,R., Schlötterer,C. and Lelley,T. (2007) SciRoKo: a new tool for whole genome microsatellite search and investigation. *Bioinformatics*, **23**, 1683–1685.
21. Wang,X., Lu,P. and Luo,Z. (2013) GMATo: a novel tool for the identification and analysis of microsatellites in large genomes. *Bioinformation*, **9**, 541–544.
22. Mudunuri,S.B. and Nagarajaram,H.A. (2007) IMEx: imperfect microsatellite extractor. *Bioinformatics*, **23**, 1181–1187.
23. Kolpakov,R., Bana,G. and Kucherov,G. (2003) mreps: efficient and flexible detection of tandem repeats in DNA. *Nucleic Acids Res.*, **31**, 3672–3678.
24. Castelo,T., Martins,W. and Gao,R. (2002) TROLL—tandem repeat occurrence locator. *Bioinformatics*, **18**, 634–636.
25. Girgis,H.Z. and Sheetlin,S.L. (2012) MsDetector: toward a standard computational tool for DNA microsatellites detection. *Nucleic Acids Res.*, **41**, e22.
26. Da Maia,L.C., Palmieri,D.A., De Souza,V.Q., Kopp,M.M., de Carvalho,F.I.F. and Costa de Oliveira,A. (2008) SSR locator: tool for simple sequence repeat discovery integrated with primer design and PCR simulation. *Int. J. Plant Genomics*, **2008**, 412696.
27. Duran,C., Singhania,R., Raman,H., Batley,J. and Edwards,D. (2013) Predicting polymorphic EST-SSRs in silico. *Mol. Ecol. Resour.*, **13**, 538–545.
28. Xia,E.H., Yao,Q.Y., Zhang,H.B., Jiang,J.J., Zhang,L.P. and Gao,L.Z. (2016) CandiSSR: an efficient pipeline used for identifying candidate polymorphic SSRs based on multiple assembled sequences. *Front. Plant Sci.*, **6**, 1171.
29. 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.
30. Aishwarya,V. and Sharma,P.C. (2007) UgMicroSat db: database for mining microsatellites from unigenes. *Nucleic Acids Res.*, **36**, D53–D56.
31. Aishwarya,V., Grover,A. and Sharma,P.C. (2007) EuMicroSatdb: a database for microsatellites in the sequenced genomes of eukaryotes. *BMC Genomics*, **8**, 225.
32. 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.
33. 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.
34. Avvaru,A.K., Saxena,S., Sowpati,D.T. and Mishra,R.K. (2017) MSDB: a comprehensive database of simple sequence repeats. *Genome Biol. Evol.*, **9**, 1797–1802.
35. 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.
36. Weising,K. and Gardner,R.C. (1999) A set of conserved PCR primers for the analysis of simple sequence repeat polymorphisms in chloroplast genomes of dicotyledonous angiosperms. *Genome*, **42**, 9–19.
37. Litt,M. and Luty,J.A. (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am. J. Hum. Genet.*, **44**, 397–401.
38. Atia,M.A., Osman,G.H. and Elmenofy,W.H. (2016) Genome-wide in silico analysis, characterization and identification of microsatellites in Spodoptera littoralis multiple nucleopolyhedrovirus (SpliMNPV). *Sci. Rep.*, **6**, 33741.
39. Mokhtar,M.M., Adawy,S.S., El-Assal,S.E.D.S. and Hussein,E.H. (2016) Genic and intergenic SSR database generation, SNPs determination and pathway annotations, in date palm (*Phoenix dactylifera* L.). *PLOS one*, **11**, e0159268.
40. Kruglyak,S., Durrett,R.T., Schug,M.D. and Aquadro,C.F. (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl. Acad. Sci. U.S.A.*, **95**, 10774–10778.
41. Esselink,G.D., Nybom,H. and Vosman,B. (2004) Assignment of allelic configuration in polyploids using the MAC-PR (microsatellite DNA allele counting—peak ratios) method. *Theor. Appl. Genet.*, **109**, 402–408.
42. 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.
43. Quilez,J., Guilmatre,A., Garg,P., Highnam,G., Gymrek,M., Erlich,Y., Joshi,R.S., Mittelman,D. and Sharp,A.J. (2016) Polymorphic tandem repeats within gene promoters act as modifiers of gene expression and DNA methylation in humans. *Nucleic Acids Res.*, **44**, 3750–3762.