Received: 11 June 2019 | Revised: 18 September 2019 | Accepted: 24 September 2019

DOI: 10.1111/1755-0998.13098

RESOURCE ARTICLE

MOLECULAR ECOLOGY RESOURCES WILEY

PSMD: An extensive database for pan-species microsatellite investigation and marker development

Lianming Du[1] | Qin Liu[2,3] | Kelei Zhao[1] | Jie Tang[4] | Xiuyue Zhang[2] | Bisong Yue[2] | Zhenxin Fan[2]

[1]Institute for Advanced Study, Chengdu University, Chengdu, China

[2]Key Laboratory of Bio-resources and Eco-environment, Ministry of Education, College of Life Science, Sichuan University, Chengdu, China

[3]College of Life Sciences and Food Engineering, Yibin University, Yibin, China

[4]School of Pharmacy and Bioengineering, Chengdu University, Chengdu, China

Correspondence
Zhenxin Fan and Bisong Yue, Key Laboratory of Bio-resources and Eco-environment, Ministry of Education, College of Life Science, Sichuan University, Chengdu 610064, China.
Emails: zxfan@scu.edu.cn (ZF); bsyue@scu.edu.cn (BY)

Funding information
Start-up Fund of Chengdu University, Grant/Award Number: 2081916062; National Natural Science Foundation of China, Grant/Award Number: 31970395

Abstract

Microsatellites are widely distributed throughout nearly all genomes which have been extensively exploited as powerful genetic markers for diverse applications due to their high polymorphisms. Their length variations are involved in gene regulation and implicated in numerous genetic diseases even in cancers. Although much effort has been devoted in microsatellite database construction, the existing microsatellite databases still had some drawbacks, such as limited number of species, unfriendly export format, missing marker development, lack of compound microsatellites and absence of gene annotation, which seriously restricted researchers to perform downstream analysis. In order to overcome the above limitations, we developed PSMD (Pan-Species Microsatellite Database, http://big.cdu.edu.cn/psmd/) as a web-based database to facilitate researchers to easily identify microsatellites, exploit reliable molecular markers and compare microsatellite distribution pattern on genome-wide scale. In current release, PSMD comprises 678,106,741 perfect microsatellites and 43,848,943 compound microsatellites from 18,408 organisms, which covered almost all species with available genomic data. In addition to interactive browse interface, PSMD also offers a flexible filter function for users to quickly gain desired microsatellites from large data sets. PSMD allows users to export GFF3 formatted file and CSV formatted statistical file for downstream analysis. We also implemented an online tool for analysing occurrence of microsatellites with user-defined parameters. Furthermore, Primer3 was embedded to help users to design high-quality primers with customizable settings. To our knowledge, PSMD is the most extensive resource which is likely to be adopted by scientists engaged in biological, medical, environmental and agricultural research.

KEYWORDS
database, genetic diversity, genetic marker, microsatellites, tandem repeats

1 | INTRODUCTION

Microsatellites, also referred to as simple repeat sequences (SSRs) or short tandem repeats (STRs), are tandem repeats of short DNA motifs with length of 1–6 bp and flanked by unique sequences that are highly conserved even among different taxa (Buschiazzo & Gemmell, 2010; Ellegren, 2004; Vieira, Santini, Diniz, & Munhoz, 2016). They are ubiquitously dispersed in almost all eukaryotic and prokaryotic genomes (Merkel & Gemmell, 2008), and even present in organellar DNA (Deng, Zhang, He, Wang, & Su, 2017; Sia et al., 2000). SSRs are

highly abundant in intergenic and noncoding regions but also occur in coding regions (Li, Korol, Fahima, & Nevo, 2004; Toth, Gaspari, & Jurka, 2000). Their distribution and density exhibit taxon-specific variations in different regions and vary markedly across genomes (Adams et al., 2016; Srivastava, Avvaru, Sowpati, & Mishra, 2019). There is ample evidence indicating that SSRs can perform diverse biological functions, such as regulating gene expression, modulating transcription factor binding, effecting alternative splicing, altering chromatin structure and regulating meiotic recombination hotspots (Bagshaw, 2017).

Based on their composition, SSRs are generally divided into three main categories namely perfect, imperfect and compound repeats (Weber, 1990). Perfect SSRs are series of a repeat unit without any interruptions (e.g., $(AC)_{10}$) while imperfect SSRs contain incomplete repeat unit or interrupted by nonrepeat nucleotides (e.g., $(AC)_8T(AC)_{10}$), while compound microsatellites consist of two or more adjacent repeats (e.g., $(ATG)_5(CA)_7$) (Oliveira, Pádua, Zucchi, Vencovsky, & Vieira, 2006). The imperfect SSRs are more genetically stable than perfect SSRs as they are less prone to slippage mutations (Brandstrom & Ellegren, 2008; Mudunuri & Nagarajaram, 2007). Perfect SSRs can be classified according to the motif size, being mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats. Although the hexanucleotide repeats were recognized as the dominant SSR type in eukaryotic genomes, their proportion is much lower in animals compared to plants, protists and fungi (Srivastava et al., 2019).Due to replication slippage and unequal crossing-over during meiosis, SSRs possess a high mutation rate (between 10^{-4} and 10^{-3} mutations per locus per generation), which can result in extensive length polymorphism among populations (Ellegren, 2004; Kelkar, Tyekucheva, Chiaromonte, & Makova, 2008). Owing to their highly polymorphic nature, SSRs are increasingly being exploited as codominant genetic markers in a variety of fields such as population and conservation genetics, parentage assignment, gene mapping and forensics (Hodel et al., 2016; Schlötterer, 2004). Moreover, the expansion of SSRs can cause a range of human disorders (Hannan, 2018), and their instability has been causally implicated in numerous cancers (Cortes-Ciriano, Lee, Park, Kim, & Park, 2017; Hause, Pritchard, Shendure, & Salipante, 2016; Kim, Laird, & Park, 2013).

Global server databases have become the most popular and convenient tool for microsatellite investigation and marker development because they are simple to create, maintain and access, relatively inexpensive and are not labour intensive. Considerable worldwide effort has aimed to develop online microsatellite databases for various organisms, such as EuMicroSatdb (Aishwarya, Grover, & Sharma, 2007), InSatDb (Archak, Meduri, Kumar, & Nagaraju, 2007), FishMicrosat (Nagpure et al., 2013), MICdb (Mudunuri, Patnana, & Nagarajaram, 2014) and PMDBase (Yu et al., 2017). However, these databases are specific to a certain taxon and several do not keep up to date with newly released genomic data. Recently, MSDB (Avvaru, Saxena, Sowpati, & Mishra, 2017) and SSRome (Mokhtar & Atia, 2019) were developed as comprehensive databases for SSRs mining across more than 6,500 species. However, these two databases also have limitations that restrict biologists to perform the downstream analysis. For example, MSDB does not provide SSRs identified from viral genomes, compound SSRs, genic locations and primers for PCR amplification. SSRome does not provide concise and intuitive charts for detailed summary statistics. Additionally, neither of them provides a flexible filter function for users to select desired SSRs from huge data sets nor supports for exporting GFF3 formatted file for further analysis.

Therefore, we present PSMD (Pan-Species Microsatellite Database) as a web-based and extensive database that encompasses almost all species with available genomic data in NCBI repository. Additionally, PSMD provides a user-friendly interface for users to browse, identify, filter, extract desired SSRs and develop significant SSR markers.

2 | MATERIALS AND METHODS

2.1 | Data sources

The FASTA formatted complete genome sequences of various species across almost all taxa, together with their GFF formatted gene annotation and assembly reports, were downloaded from the FTP site of NCBI (NCBI Resource Coordinators, 2018) by using aria2 download utility (https://aria2.github.io/). In cases where multiple genome versions were available for the same species, the NCBI reference or representative genome was preferred, or priority was given to the best assembled and annotated genome. Like MSDB, all organisms were also classified and organized based on their taxonomic group information including Kingdom, Group and Subgroup.

2.2 | Data processing

We developed an inhouse Python script that employed tandem repeat search module built from Krait (Du, Zhang, Liu, Zhang, & Yue, 2018) as its repeat search engine to identify perfect and compound SSRs. For eukaryotic genomes, the minimum repeats for mono-, di-, tri-, tetra-, penta- and hexanucleotide SSRs were set to 12, 7, 5, 4, 4 and 4 (respectively) according to previous studies (Huang et al., 2015; Xu et al., 2018). The minimum repeats were set to 6, 3, 3, 3, 3 and 3 for prokaryotic and viral genomes based on empirical criterion (Alam, Singh, Sharfuddin, & Ali, 2014; Wu, Zhou, Zhao, & Tan, 2014) due to their smaller genome size and lower repeat tract content. Maximum distance (d_{max}) allowed between any two adjacent SSRs forming a compound SSR was set to 10 bp, which was estimated as optimal d_{max} by a previous study (Kofler, Schlötterer, Luschützky, & Lelley, 2008). Repeats with unit patterns being circular permutations and/or reverse complements were considered as the same type for statistical analysis (Huang et al., 2015). All identified perfect and compound microsatellites were mapped into genic and nongenic regions by comparing start and stop coordinates of SSRs to feature coordinates within gene annotation files of genomes using NCLS (Alekseyenko & Lee, 2007). Primer3 (Untergasser et al., 2012) was

TABLE 1 Comparison of PSMD database with other active SSR databases in the aspect of (a) number of organisms and (b) database features

	InSatDb	FishMicrosat	MICdb	PMDBase	MSDB	SSRome	PSMD
(a) Number of species							
Animals	5	95	0	0	310	137	1,074
Plants	0	0	0	110	74	98	371
Fungi	0	0	0	0	191	241	1,513
Protists	0	0	0	0	72	78	293
Archaea	0	0	271	0	514	125	632
Bacteria	0	0	4,772	0	5,732	2,828	12,044
Viruses	0	0	0	0	0	1,270	2,466
(b) database features							
Compound SSR	No	Yes	No	Yes	No	No	Yes
Genic SSR	Yes	No	Yes	Yes	No	Yes	Yes
Table browse	No	Yes	Yes	Yes	Yes	No	Yes
SSR filter	No	No	No	No	Yes	No	Yes
Statistical plot	No	Yes	Yes	No	Yes	No	Yes
Comparative analysis	No	No	Yes	No	Yes	Yes	Yes
Data download	Yes	No	No	Yes	Yes	Yes	Yes
Primer design	No	Yes	Yes	Yes	No	Yes	Yes
Export GFF	No	No	No	No	No	No	Yes
Integrate SSR search tool	No	No	No	Yes	No	Yes	Yes

applied to design primers for all identified SSRs with 100 bp flanking sequences.

2.3 | Database construction

In order to build a fast response database, the SSR metadata including sequence name, position, motif, repeats, length, flanking sequence and genic annotation was stored separately into SQLite database for each species. Species' basic information and SSR summary statistical information were stored in a MariaDB relational database (https://mariadb.org/) operated on CentOS server. The interactive web interface was built with Python and the Django web framework (https://www.djangoproject.com/). The Django will automatically switch to a specific SQLite database according to species specified by users. The web pages were constructed using HTML5, CSS3 and rendered using Jinja2 (http://jinja.pocoo.org/) template engine. Jquery library (https://jquery.com/) was used with Bootstrap framework (https://getbootstrap.com/) to establish a responsive user-friendly frontend interface. Highcharts (https://www.highcharts.com/) was adopted to generate statistical charts. Primer3 was seamlessly integrated into database to facilitate users to design primers according to their own specified parameters. Meanwhile, an online version of Krait was implemented to assist users to search for perfect, compound and imperfect microsatellites that meet user requirements. The database construction pipeline was depicted in Figure 1.

3 | RESULTS

3.1 | Database content

We downloaded 18,408 genomes from NCBI assembly database across a wide variety of species ranging from virus to human which covered many more species than other related SSR databases (Table 1a) and presented more than 1,599 gigabases of DNA. In total, we identified 678,106,681 perfect SSRs from this exceptionally large set of genomic data. Among which, 175,311,695 (25.85%) SSRs were presented in genic regions including 87,329,511 (12.88%) SSRs found in coding regions. In addition, we observed that 97,971,818 (14.45%) perfect SSRs comprised 43,848,941 compound SSRs, which contained 9,765,018 (22.27%) genic and 4,210,890 (9.6%) coding compound SSRs. The genome size, GC content, SSR counts, SSR frequency (number of SSRs per Mb of genome, loci/Mb), SSR density (length of SSRs per Mb of genome, bp/Mb), percentage of genome covered by SSRs and percentage of individual SSRs being part of a compound SSRs (cSSRs%) for each genome were calculated and stored into a summary table together with corresponding species name, accession number and taxonomy ID.

3.2 | Database features

The PSMD database allows users to extract both perfect and compound SSRs with their genic locations. The database also offers

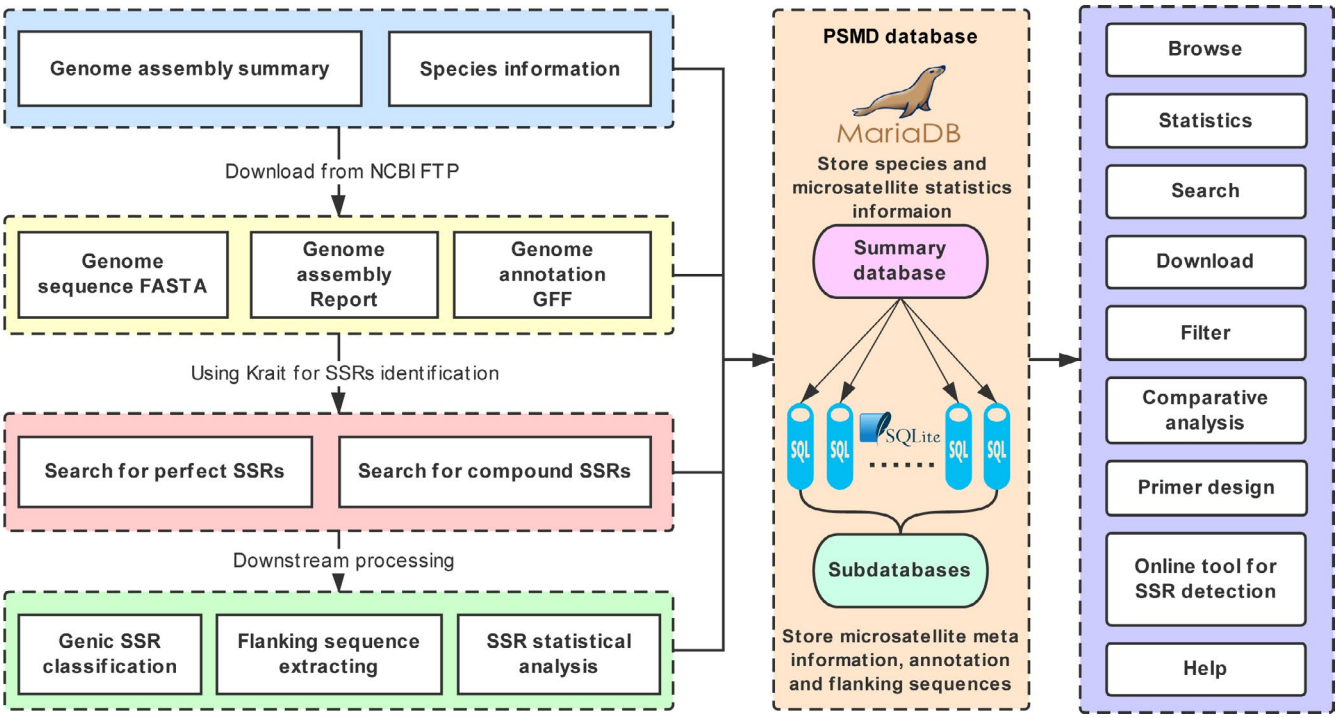

FIGURE 1 The workflow of data processing and PSMD database development [Colour figure can be viewed at wileyonlinelibrary.com]

a flexible filter function unlike any other databases for users to rapidly gain desired SSRs from large data sets. Compared to other SSR databases, PSMD brings some exciting innovative features. For example, PSMD allows users to export GFF3 formatted files for downstream analysis. To date, no consensus has been reached regarding the minimum repeat required to define an SSR. The data sets in the database generated with our preset parameters may not be suitable for some users to perform special analysis. Thus, we have implemented an online tool for mining of SSRs from genomes in the database or user input sequences with custom parameters. Moreover, the primers of each SSR can be designed using embedded Primer3 with user-defined parameters. The full comparison of features between PSMD and other databases was shown in Table 1b.

3.3 | Web interface and usage

The PSMD database provides a user-friendly interactive web interface that facilitates users to browse, search, download and compare microsatellites across all organisms with a modern web browser. The PSMD database offers a navigation bar running across the top of the website to help users quickly access the database features. The search box inside the navigation bar allows users to search a specific genome by species name, assembly accession number, NCBI taxonomy ID or common name (if any). The functionalities of PSMD database can be easily accessed with assistance of six interactive pages including Home, Overview, Browse, Analysis, Tool and Help. On the home page, four dropdown lists for kingdom, group, subgroup and species were provided to help users to easily select or to search a desired species for quick access to species summary information. Further, users can simply click the image links under dropdown lists to immediately jump to human or a model organism. The help page supplies sufficient documentation for users to access the database.

3.3.1 | Overview of statistics

The overview page lists statistical summary information for all species in the PSMD database in a column-wise table. Among the information displayed is genome size, GC content, SSR counts, SSR frequency, SSR density and percentage of genome covered by SSRs as well as corresponding species information including taxonomy, species name and accession number. Users can use taxonomic group dropdown lists on the top of the page to limit the table to show only species in a specific taxon. Users can view detailed summary information for a species by double-clicking a row. Apart from contents displayed in table, users are also allowed to retrieve motifs abundance, noncoding and coding SSRs abundance of all species from a popup dialog displayed by clicking the download button.

3.3.2 | Browsing microsatellites

The PSMD database offers an interactive table on the browse page for users to browse, filter and sort SSRs identified from selected species (Figure 2a). In the table, SSR ID, sequence accession and name, start and end position, motif, standard motif, type, repeats and length are given for each SSR and genic location are given in the last column if SSRs were detected in genic regions. By double-clicking a row, a user can obtain detailed information of SSRs including

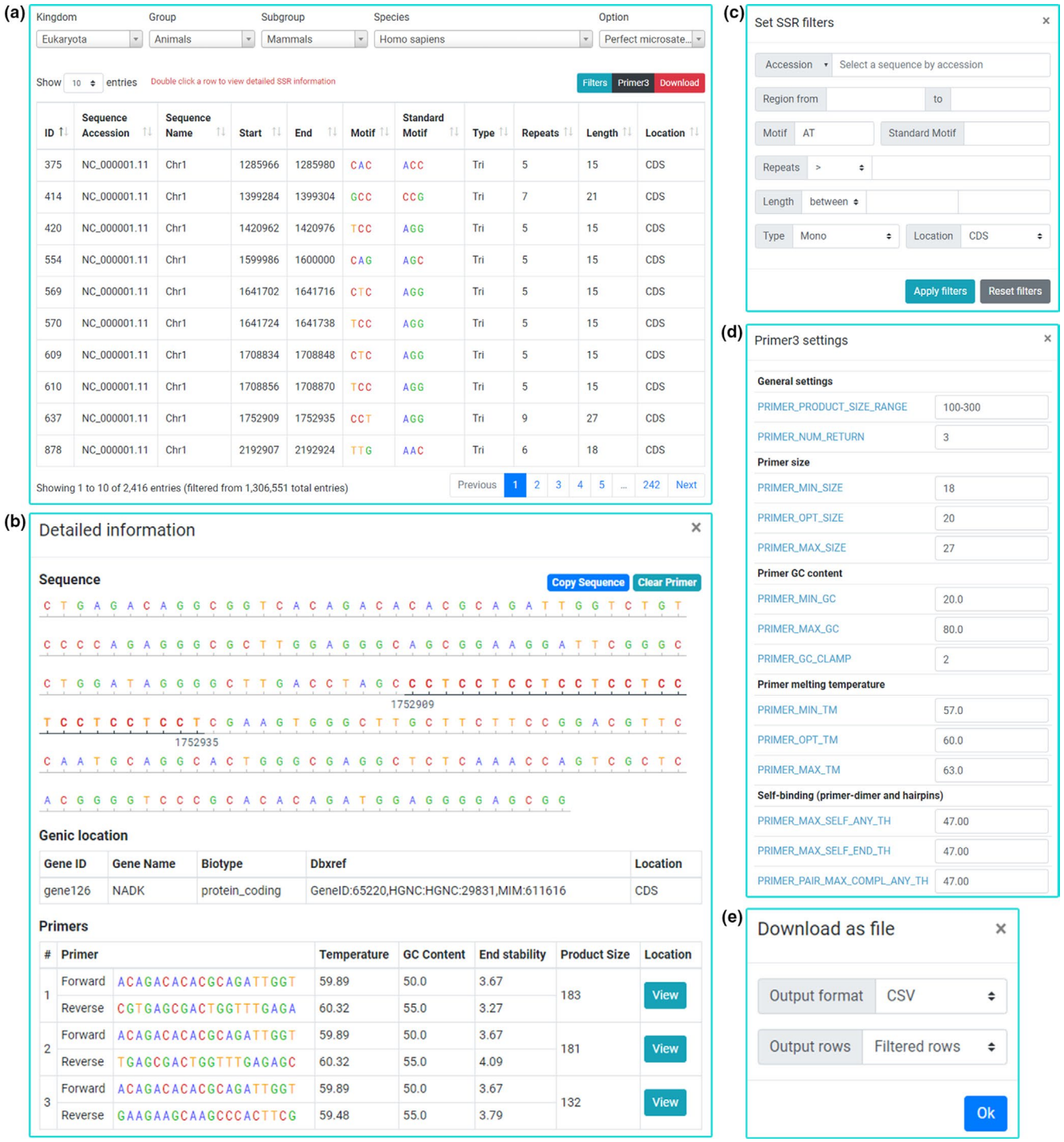

FIGURE 2 Screenshots of SSR browse page. (a) A text formatted table for displaying SSR information. (b) A dialog for displaying SSR sequence, gene location and primers. (c) A dialog for filtering SSRs. (d) A dialog for setting primer3 parameters. (e) A dialog for downloading SSRs [Colour figure can be viewed at wileyonlinelibrary.com]

SSR sequence with flanking sequences, three pairs of primers if successfully designed by primer3 and associated gene information if available (Figure 2b). The current species that is being browsed can be changed by selecting a species in species dropdown lists. In addition, option dropdown lists can facilitate users switching between perfect or compound SSRs browsing page and summary information page. A filter dialog displayed by clicking filters button was designed to enable users to set restricted conditions for filtering SSRs (Figure 2c). Users are allowed to customize the primer3 parameters using the text input boxes provided in a dialog (Figure 2d), which may affect the primers displayed in SSR detailed information dialog. All SSRs or filtered SSRs can be downloaded as TSV, CSV and GFF3 format files for further analysis by clicking the download button (Figure 2e).

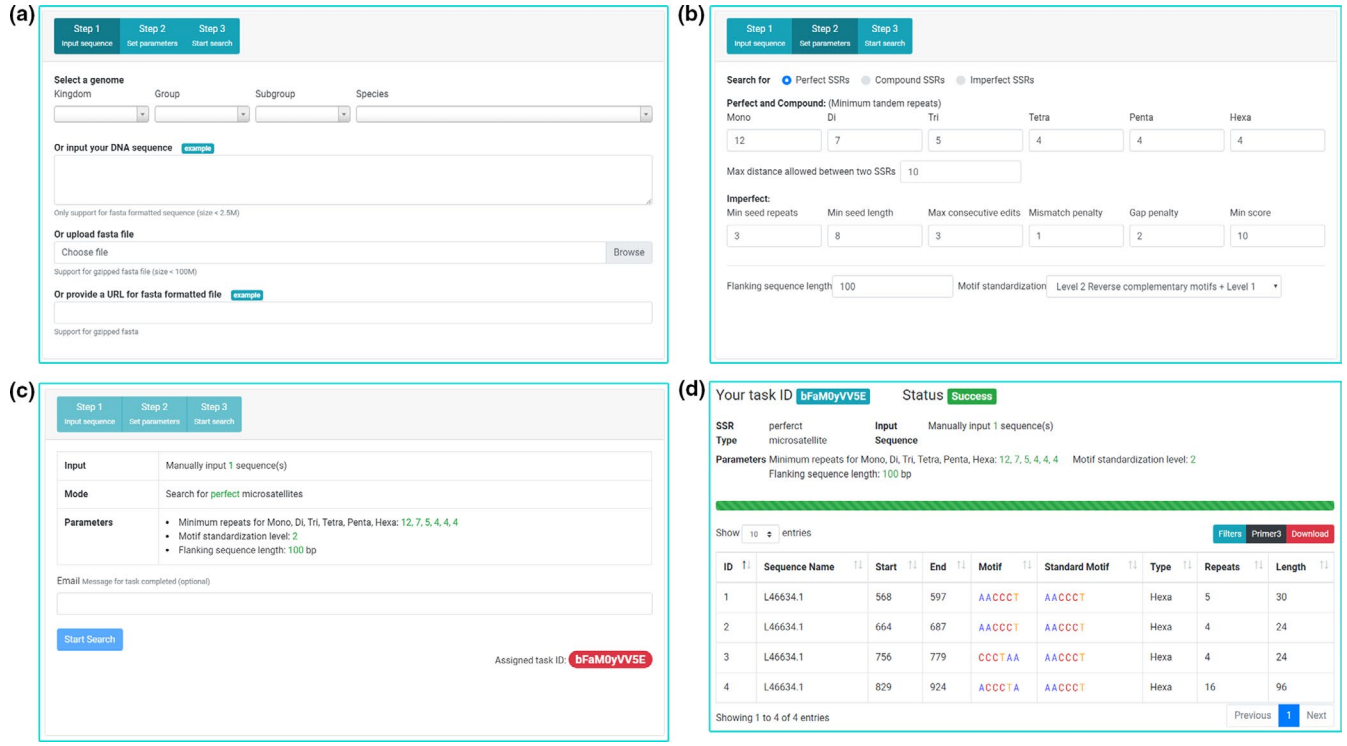

FIGURE 3 Screenshots of online tool for SSR identification. (a) Step one for sequence input and genome selecting. (b) Step two for parameter setting. (c) Step three for starting search task. (d) SSR search result page [Colour figure can be viewed at wileyonlinelibrary.com]

3.3.3 | Comparative analysis

On the analysis page, an easy-to-operate tool was provided to assist users to easily compare the prevalence and distribution of SSRs between several species. In addition to frequency and density, this tool also allows users to compare genome coverage, cSSRs%, motif and repeat abundance as well as dominant SSR types. These analysis results will be presented in concise and interactive plots and provide insight into the evolutionary trends of SSRs in the prevalence of SSR motifs and types. All the generated plots can be downloaded as local files.

3.3.4 | Online tool for microsatellite identification

All SSRs in the PSMD database were identified using predetermined parameters, which may not meet the requirements of users wanting specific analyses. In order to make PSMD more flexible and enable users to customize parameters for SSR identification, we have developed a powerful web service to help users easily identify SSRs from FASTA formatted DNA sequences. The online tool has only three steps for task running. In step one, apart from selecting a genome that exists within the database, this tool provides three ways for users to supply FASTA formatted DNA sequences to identify microsatellites. Users can paste sequences into the text area directly, or click the input box below the text area to choose a local file to upload sequences, or input a URL of a sequence file (Figure 3a). If a URL is provided, the tool will automatically download the sequence file into the server for microsatellite identification. In step two, users can select a type of SSR (perfect, compound or imperfect) for identification and set corresponding parameters including minimum repeats, d_{max}, motif standardization level and flanking sequence length (Figure 3b). In step three, users can click the start search button to submit a task to the server and obtain a unique task ID (Figure 3c). Users can click the hyperlink of task ID to jump to the task page in which they can browse, filter, sort and download detected SSRs as well as design primers (Figure 3d). A more clearly worked example of how to use this online tool to identify SSRs was shown in Figure S1.

3.4 | Application example

The PSMD database provides various statistical results for users to easily download (Figure S2) and perform additional analyses, such as SSR evolutionary origins and history. We used these data sets to explore SSR distribution in eukaryotic genomes and discovered considerable taxon-specific variations in both counts and length of each SSR type (Figure 4a). On the whole, fungi and protists that possess highly abundant trinucleotide repeats presented distinctly different distribution pattern from plants and animals. Trinucleotide repeats are predominant in green algae whose distribution pattern is more like fungi and significantly different from land plants. In animals, birds and fishes showed different expansion bias in SSR type distribution where mononucleotide repeats were more common in birds and dinucleotide repeats were more common in fishes. In addition, fishes have relatively higher SSR abundance than other taxonomic groups

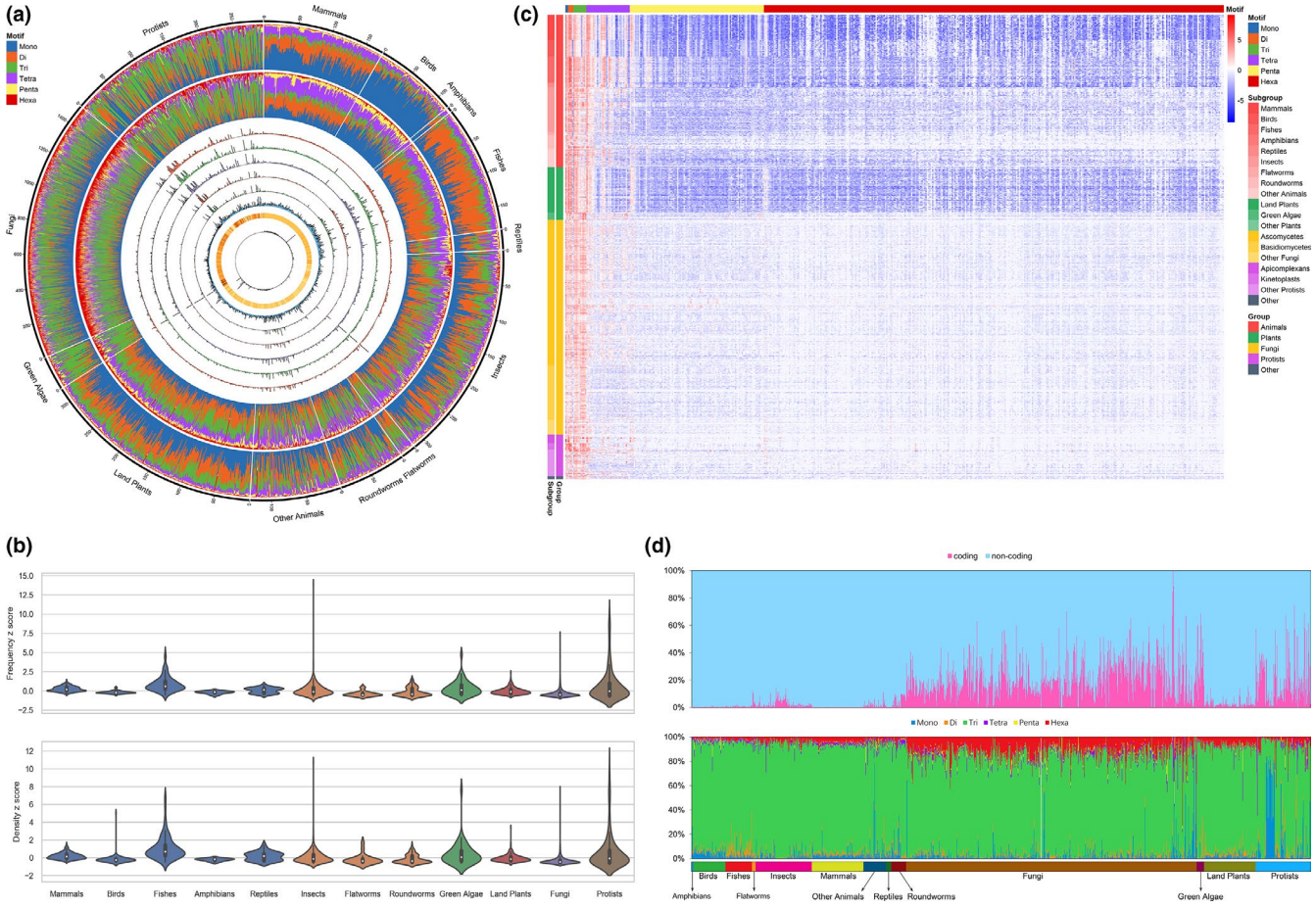

FIGURE 4 Landscape of SSR distribution across taxonomic groups of eukaryota. (a) Circos plot for showing SSR distribution. The outer ideograms represent the taxonomic groups with ticks indicating number of species. The circles from outside to inside represent SSR counts for each type, SSR length for each type, SSR frequency (loci/Mb), SSR density (bp/Mb), genome coverage (%), compound SSR frequency (loci/Mb), compound SSR density (bp/Mb), percentage of individual SSRs being part of a compound SSRs (%), GC content (%), genome size (Gb). (b) Violin plot for showing SSR frequency and density *z* score. (c) Heatmap for showing 501 motifs density in each organism. (d) SSR distribution in coding and noncoding regions, and SSR type distribution in coding region [Colour figure can be viewed at wileyonlinelibrary.com]

(Figure 4a, b). 5,356 possible permutations were grouped into 501 unique motifs by removing cyclically equivalent and reverse complementary permutations for motif abundance analysis. The density trends of the 501 motifs showed obvious differences between different taxonomic groups and even between subgroups (Figure 4c). Large numbers of pentanucleotide and hexanucleotide motifs were absent in fungi and protists and had relatively lower abundance in mammals. Moreover, extremely low percentages of SSRs were distributed in coding regions, except for fungi and protists, and trinucleotide repeats were overwhelming in coding regions across almost all organisms, followed by hexanucleotide repeats (Figure 4d).

4 | DISCUSSION

Microsatellites are acknowledged as important advantageous genetic markers that have immense utility in numerous fields, particularly in genetic diversity and evolution (Abdul-Muneer, 2014; Adams et al., 2016; Sheriff & Alemayehu, 2018). Microsatellites can be used to estimate many parameters of interest to ecologists, such as migration rates, population size, bottlenecks and kinship (Selkoe & Toonen, 2006). The PSMD database represents an extensive resource of microsatellites that facilitate researchers to shed new light on microsatellite evolution and develop reliable molecular markers. Compared to similar databases, PSMD contains data on the most organisms covering almost all species with available genomic data and contains both perfect and compound microsatellites. In addition to biologists engaged in genome organization and evolution studies, population studies, microsatellite abundance and mutability, PSMD with embedded tools is also likely to be adopted by scientists engaged in medical, environmental, agricultural and aquacultural research to understand the functional significance of microsatellites in biological processes.

The PSMD database will continuously be upgraded with newly sequenced genomes and updated draft assemblies in the NCBI genome repository. Each species in the PSMD was designed as a

separate subdatabase to store relevant information of microsatellites for fast data retrieval. This flexible design makes it feasible to increase the size of the database to virtually any size without compromising its accessing performance. Additionally, the web interface will be improved or adjusted and new functions will be added to assist researchers to maximize the utilization of the database. For example, we will integrate more gene functional information including GO function (The Gene Ontology Consortium, 2019) and KEGG pathways (Kanehisa, Furumichi, Tanabe, Sato, & Morishima, 2017) to help users to exploit potential functional microsatellites. Recently, high-throughput sequencing has become an increasingly important tool for microsatellite genotyping in ecology (De Barba et al., 2017); we will integrate next-generation sequencing data and develop a pipeline to help users to perform microsatellite genotyping.

ACKNOWLEDGEMENTS

We thank Haoyang Cai for his suggestions and comments. We are grateful to Megan Price for her critical reading and English editing on the manuscript. This work was supported by the Start-up Fund of Chengdu University (2081916062) and the National Natural Science Foundation of China (31970395).

AUTHOR CONTRIBUTIONS

The study was conceived and devised by D.L., F.Z. and Y.B.; The manuscript was written by D.L.; The manuscript was revised by F.Z., Z.K., T.J. and Z.X.; Sample data were collected by L.Q. and D.L.; The bioinformatics analysis was conducted by D.L.; The database was implemented and designed by D.L. and F.Z.

DATA AVAILABILITY STATEMENT

PSMD database is freely accessible at http://big.cdu.edu.cn/psmd or http://psmd.biosv.com. The scripts used to construct database are freely available at https://github.com/lmdu/psmd.

ORCID

Lianming Du https://orcid.org/0000-0003-0646-0887
Xiuyue Zhang https://orcid.org/0000-0002-9681-2905
Bisong Yue https://orcid.org/0000-0001-6731-538X
Zhenxin Fan https://orcid.org/0000-0003-0422-9497

REFERENCES

Abdul-Muneer, P. M. (2014). Application of microsatellite markers in conservation genetics and fisheries management: Recent advances in population structure analysis and conservation strategies. *Genetics Research International*, *2014*, 691759. https://doi.org/10.1155/2014/691759

Adams, R. H., Blackmon, H., Reyes-Velasco, J., Schield, D. R., Card, D. C., Andrew, A. L., … Castoe, T. A. (2016). Microsatellite landscape evolutionary dynamics across 450 million years of vertebrate genome evolution. *Genome*, *59*(5), 295–310. https://doi.org/10.1139/gen-2015-0124

Aishwarya, V., Grover, A., & Sharma, P. C. (2007). EuMicroSatdb: A database for microsatellites in the sequenced genomes of eukaryotes. *BMC Genomics*, *8*(1), 225. https://doi.org/10.1186/1471-2164-8-225

Alam, C. M., Singh, A. K., Sharfuddin, C., & Ali, S. (2014). Genome-wide scan for analysis of simple and imperfect microsatellites in diverse *carlaviruses*. *Infection, Genetics and Evolution*, *21*, 287–294. https://doi.org/10.1016/j.meegid.2013.11.018

Alekseyenko, A. V., & Lee, C. J. (2007). Nested Containment List (NCList): A new algorithm for accelerating interval query of genome alignment and interval databases. *Bioinformatics*, *23*(11), 1386–1393. https://doi.org/10.1093/bioinformatics/btl647

Archak, S., Meduri, E., Kumar, P. S., & Nagaraju, J. (2007). InSatDb: A microsatellite database of fully sequenced insect genomes. *Nucleic Acids Research*, *35*, D36–D39. https://doi.org/10.1093/nar/gkl778

Avvaru, A. K., Saxena, S., Sowpati, D. T., & Mishra, R. K. (2017). MSDB: A comprehensive database of simple sequence repeats. *Genome Biology and Evolution*, *9*(6), 1797–1802. https://doi.org/10.1093/gbe/evx132

Bagshaw, A. T. M. (2017). Functional mechanisms of microsatellite DNA in eukaryotic genomes. *Genome Biology and Evolution*, *9*(9), 2428–2443. https://doi.org/10.1093/gbe/evx164

Brandstrom, M., & Ellegren, H. (2008). Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. *Genome Research*, *18*(6), 881–887. https://doi.org/10.1101/gr.075242.107

Buschiazzo, E., & Gemmell, N. J. (2010). Conservation of human microsatellites across 450 million years of evolution. *Genome Biology and Evolution*, *2*, 153–165. https://doi.org/10.1093/gbe/evq007

Cortes-Ciriano, I., Lee, S., Park, W. Y., Kim, T. M., & Park, P. J. (2017). A molecular portrait of microsatellite instability across multiple cancers. *Nature Communications*, *8*(1), 15180. https://doi.org/10.1038/ncomms15180

De Barba, M., Miquel, C., Lobréaux, S., Quenette, P. Y., Swenson, J. E., & Taberlet, P. (2017). High-throughput microsatellite genotyping in ecology: Improved accuracy, efficiency, standardization and success with low-quantity and degraded DNA. *Molecular Ecology Resources*, *17*(3), 492–507. https://doi.org/10.1111/1755-0998.12594

Deng, Q., Zhang, H., He, Y., Wang, T., & Su, Y. (2017). Chloroplast microsatellite markers for *Pseudotaxuschienii* developed from the whole chloroplast genome of *Taxuschinensis* var. *mairei* (Taxaceae). *Applications in Plant Sciences*, *5*(3), 1600153.

Du, L., Zhang, C., Liu, Q., Zhang, X., & Yue, B. (2018). Krait: An ultrafast tool for genome-wide survey of microsatellites and primer design. *Bioinformatics*, *34*(4), 681–683. https://doi.org/10.1093/bioinformatics/btx665

Ellegren, H. (2004). Microsatellites: Simple sequences with complex evolution. *Nature Reviews Genetics*, *5*(6), 435–445. https://doi.org/10.1038/nrg1348

Hannan, A. J. (2018). Tandem repeats mediating genetic plasticity in health and disease. *Nature Reviews Genetics*, *19*(5), 286–298. https://doi.org/10.1038/nrg.2017.115

Hause, R. J., Pritchard, C. C., Shendure, J., & Salipante, S. J. (2016). Classification and characterization of microsatellite instability across 18 cancer types. *Nature Medicine*, *22*(11), 1342–1350. https://doi.org/10.1038/nm.4191

Hodel, R. G. J., Segovia-Salcedo, M. C., Landis, J. B., Crowl, A. A., Sun, M., Liu, X., … Soltis, P. S. (2016). The report of my death was an exaggeration: A review for researchers using microsatellites in the 21st century. *Applications in Plant Sciences*, *4*(6), 1600025. https://doi.org/10.3732/apps.1600025

Huang, J., Li, Y.-Z., Du, L.-M., Yang, B. O., Shen, F.-J., Zhang, H.-M., … Yue, B.-S. (2015). Genome-wide survey and analysis of microsatellites in giant panda (*Ailuropodamelanoleuca*), with a focus on the applications of a novel microsatellite marker system. *BMC Genomics*, *16*, 61. https://doi.org/10.1186/s12864-015-1268-z

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research*, *45*(D1), D353–D361. https://doi.org/10.1093/nar/gkw1092

Kelkar, Y. D., Tyekucheva, S., Chiaromonte, F., & Makova, K. D. L. (2008). The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Research*, *18*(1), 30–38. https://doi.org/10.1101/gr.7113408

Kim, T. M., Laird, P. W., & Park, P. J. (2013). Landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell*, *155*(4), 858–868. https://doi.org/10.1016/j.cell.2013.10.015

Kofler, R., Schlötterer, C., Luschützky, E., & Lelley, T. (2008). Survey of microsatellite clustering in eight fully sequenced species sheds light on the origin of compound microsatellites. *BMC Genomics*, *9*, 612. https://doi.org/10.1186/1471-2164-9-612

Li, Y. C., Korol, A. B., Fahima, T., & Nevo, E. (2004). Microsatellites within genes: Structure, function and evolution. *Molecular Biology and Evolution*, *21*(6), 991–1007. https://doi.org/10.1093/molbev/msh073

Merkel, A., & Gemmell, N. (2008). Detecting short tandem repeats from genome data: Opening the software black box. *Briefings in Bioinformatics*, *9*(5), 355–366. https://doi.org/10.1093/bib/bbn028

Mokhtar, M. M., & Atia, M. A. M. (2019). SSRome: An integrated database and pipelines for exploring microsatellites in all organisms. *Nucleic Acids Research*, *47*(D1), D244–D252. https://doi.org/10.1093/nar/gky998

Mudunuri, S. B. & Nagarajaram, H. A. (2007). IMEx: Imperfect microsatellite extractor. *Bioinformatics*, *23*(10), 1181–1187. https://doi.org/10.1093/bioinformatics/btm097

Mudunuri, S. B., Patnana, S., & Nagarajaram, H. A. (2014). MICdb3.0: A comprehensive resource of microsatellite repeats from prokaryotic genomes. *Database*, *2014*, bau005.

Nagpure, N. S., Rashid, I., Pati, R., Pathak, A. K., Singh, M., Singh, S. P., & Sarkar, U. K. (2013). FishMicrosat: A microsatellite database of commercially important fishes and shellfishes of the Indian subcontinent. *BMC Genomics*, *14*, 630. https://doi.org/10.1186/1471-2164-14-630

NCBI Resource Coordinators. (2018). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, *46*(D1), D8–D13.

Oliveira, E. J., Pádua, J. G., Zucchi, M. I., Vencovsky, R., & Vieira, M. L. C. (2006). Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology*, *29*(2), 294–307. https://doi.org/10.1590/S1415-47572006000200018

Schlötterer, C. (2004). The evolution of molecular markers–just a matter of fashion?. *Nature Reviews Genetics*, *5*(1), 63–69. https://doi.org/10.1038/nrg1249

Selkoe, K. A., & Toonen, R. J. (2006). Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers. *Ecology Letters*, *9*(5), 615–629. https://doi.org/10.1111/j.1461-0248.2006.00889.x

Sheriff, O., & Alemayehu, K. (2018). Genetic diversity studies using microsatellite markers and their contribution in supporting sustainable sheep breeding programs: A review. *Cogent Food & Agriculture*, *4*, 1459062. https://doi.org/10.1080/23311932.2018.1459062

Sia, E. A., Butler, C. A., Dominska, M., Greenwell, P., Fox, T. D., & Petes, T. D. (2000). Analysis of microsatellite mutations in the mitochondrial DNA of *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences*, *97*(1), 250–255. https://doi.org/10.1073/pnas.97.1.250

Srivastava, S., Avvaru, A. K., Sowpati, D. T., & Mishra, R. K. (2019). Patterns of microsatellite distribution across eukaryotic genomes. *BMC Genomics*, *20*(1), 153. https://doi.org/10.1186/s12864-019-5516-5

The Gene Ontology Consortium (2019). The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Research*, *47*(D1), D330–D338.

Toth, G., Gaspari, Z., & Jurka, J. (2000). Microsatellites in different eukaryotic genomes: Survey and analysis. *Genome Research*, *10*(7), 967–981. https://doi.org/10.1101/gr.10.7.967

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B. C., Remm, M., & Rozen, S. G. (2012). Primer3–new capabilities and interfaces. *Nucleic Acids Research.*, *40*(15), e115. https://doi.org/10.1093/nar/gks596

Vieira, M. L., Santini, L., Diniz, A. L., & Munhoz, C. (2016). Microsatellite markers: What they mean and why they are so useful. *Genetics and Molecular Biology*, *39*(3), 312–328. https://doi.org/10.1590/1678-4685-GMB-2016-0027

Weber, J. L. (1990). Informativeness of human (dC-dA)n. (dG-dT)n polymorphisms. *Genomics*, *7*(4), 524–530. https://doi.org/10.1016/0888-7543(90)90195-Z

Wu, X., Zhou, L., Zhao, X., & Tan, Z. (2014). The analysis of microsatellites and compound microsatellites in 56 complete genomes of *Herpesvirales*. *Gene*, *551*, 103–109. https://doi.org/10.1016/j.gene.2014.08.054

Xu, Y., Li, W., Hu, Z., Zeng, T., Shen, Y., Liu, S., … Yue, B. (2018). Genome-wide mining of perfect microsatellites and tetranucleotide orthologous microsatellites estimates in six primate species. *Gene*, *643*, 124–132. https://doi.org/10.1016/j.gene.2017.12.008

Yu, J., Dossa, K., Wang, L., Zhang, Y., Wei, X., Liao, B., & Zhang, X. (2017). PMDBase: A database for studying microsatellite DNA and marker development in plants. *Nucleic Acids Research*, *45*(D1), D1046–D1053. https://doi.org/10.1093/nar/gkw906

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Du L, Liu Q, Zhao K, et al. PSMD: An extensive database for pan-species microsatellite investigation and marker development. *Mol Ecol Resour.* 2020;20:283–291. https://doi.org/10.1111/1755-0998.13098