Published online 10 October 2019 *Nucleic Acids Research, 2020, Vol. 48, Database issue* ***D155–D159***
doi: 10.1093/nar/gkz886

MSDB: a comprehensive, annotated database of microsatellites

Akshay Kumar Avvaru[1,2], Deepak Sharma[1,†], Archana Verma[1,†], Rakesh K. Mishra[1] and Divya Tej Sowpati[1,*]

[1]CSIR—Centre for Cellular and Molecular Biology, Hyderabad - 500007, India and [2]Academy of Scientific and Innovative Research (AcSIR), Ghaziabad - 201002, India

Received August 10, 2019; Revised September 28, 2019; Editorial Decision September 30, 2019; Accepted October 01, 2019

ABSTRACT

Microsatellites are short tandem repeats of 1–6 nucleotide motifs, studied for their utility as genome markers and in forensics. Recent evidence points to the role of microsatellites in important regulatory functions, and their length polymorphisms at coding regions are linked to various neurodegenerative disorders in humans. Microsatellites show a taxon-specific enrichment in eukaryotic genomes, and their evolution remains poorly understood. Though other databases of microsatellites exist, they fall short on several fronts. MSDB (MicroSatellite DataBase) is a collection of >4 billion microsatellites from 37 680 genomes presented in a user-friendly web portal for easy, interactive analysis and visualization. This is by far the most comprehensive, annotated, updated database to access and analyze microsatellite data of multiple species. The features of MSDB enable users to explore the data as tables that can be filtered and exported, and also as interactive charts to view and compare the data of multiple species simultaneously. Its modularity and architecture permit seamless updates with new data, making it a powerful tool and useful resource to researchers working on this important class of DNA elements, particularly in context of their evolution and emerging roles in genome organization and gene regulation.

INTRODUCTION

Microsatellites, also known as simple sequence repeats (SSRs), are short tandem DNA repeats of 1–6 nucleotide (nt) motifs comprising a significant proportion of the non-coding genome. The distribution of microsatellites in eukaryotic genomes is nonrandom (1), and a subset of SSRs shows taxon-specific enrichment (2). The variation in the length of microsatellites, particularly their expansion, in protein-coding regions causes various neurodegenerative diseases in humans (3). Historically, microsatellites are studied for their use in marker-assisted selection (4), linkage analysis (5) and DNA fingerprinting (6). In the recent past, several studies have highlighted the role of microsatellites in gene regulatory functions—epigenetic regulation by GAA repeats (7), enhancer-blocker activity of AGAT repeats (8), association of AAGAG repeat transcripts with nuclear matrix (9) and other possible roles in genome organization (10).

Microsatellites show high rates of polymorphisms with a preference for elongation (11). However, their evolution remains poorly understood. A possible reason for this could be the lack of a unified resource and an analytical tool for these elements. Several databases of microsatellites currently exist (Table 1). However, they are taxa-specific (12,13), contain information of only a few species (14,15) or are outdated (14,16,17). In addition, the tabular format used by most databases makes it cumbersome to interpret global trends. Similarly, understanding the distribution of genic/intergenic microsatellites is challenging because most current databases lack the genomic annotation of these elements. A recent database, SSRome, addresses the issue of genomic context, but for only a subset of species and microsatellites (18). Finally, no existing database enables simultaneous comparison of data from multiple species (Table 1B). Here, we created a database of >4 billion microsatellites from more than 37 500 genomes available in various genome repositories. Microsatellites from over 27 500 of these genomes are annotated with their genomic context. MSDB is designed for speed, ease of use and constant updates with new genome data, and also provides functionality to analyze and compare data from multiple species simultaneously as interactive charts and tables.

*To whom correspondence should be addressed. Tel: +91 40 2719 2960; Email: tej@ccmb.res.in
†The authors wish it to be known that, in their opinion, the second and third authors have contributed to the work equally.

Table 1. Comparison of MSDB with existing microsatellite databases based on (**A**) the number of species for which data are available, and (**B**) database features and functionality

(A) Taxonomic group	MTRD	UgMSD	KMD	PMDBase	TRDB	FMS	MICAS	EMSDb	SSRome	MSDB
Bacteria	1109	0	0	0	1	0	4772	0	2828	21 525
Archaea	91	0	0	0	0	0	271	0	125	1397
Plants	0	80	14	110	2	0	0	31	98	604
Fungi	0	80	0	0	1	0	0	31	241	2702
Protozoa	0	80	0	0	0	0	0	31	78	492
Metazoans	0	160	0	0	18	190	0	62	137	1975
Viruses	1463	0	0	0	0	0	0	0	1270	8952

(B) Feature	MTRD	UgMSD	KMD	PMDBase	TRDB	FMS	MICAS	EMSDb	SSRome	MSDB
Interactive tables with column filters	Yes[a]	Yes[a]	No[d]	Yes[e]	Yes	No	Yes[a]	Yes[a]	Yes	Yes
Downstream analysis and plots	No	No	No	No	No	Yes[f]	No	No	No	Yes
Comparison of data from multiple organisms	No[b]	No	No	No	No	No	No	No	No	Yes
Annotation of microsatellites with genomic features	No	No	No	No	No	No	No	No	Yes	Yes
Taxonomic grouping	Yes[c]	Yes	No	Yes	No	Yes	Yes[c]	No	Yes	Yes
Data download	No	Yes	Yes	Yes	Yes	No	No	No	Yes	Yes

[a]Does not support dynamic filtering of the results; the filtering parameters should be selected initially
[b]Comparison only across different strains of same species
[c]Grouping only based on the kingdoms
[d]Only a tabular view of the data without dynamic filters
[e]Filtering only based on the type of repeat
[f]Only pie charts available
MTRD: Microorganism Tandem Repeats Database (http://minisatellites.u-psud.fr); UgMSD: UgMicro SatDb (http://veenuash.info/web1/index.html); KMD: Kazusa Marker Database (http://marker.kazusa.or.jp/); PMDBase: Plant Microsatellite DNAs Database (http://www.sesame-bioinfo.org/PMDBase/index.html); TRDB: Tandem Repeats Database (https://tandem.bu.edu/cgi-bin/trdb/trdb.exe?taskid=0); FMS: FishMicro Sat (http://mail.nbfgr.res.in/fishmicrosat/); MICAS: http://www.mcr.org.in/micas/; EMSDb: EuMicroSat Db (http://www.veenuash.info/); SSRome: http://mggm-lab.easyomics.org/; MSDB: MicroSatellite DataBase (https://data.ccmb.res.in/msdb)

MATERIALS AND METHODS

Data sources

All genomic data was obtained from RefSeq and GenBank databases of NCBI, and the UCSC genome browser (19–21). Sequence information was downloaded in FASTA format, and the genic annotations where available were downloaded in GFF/GTF format. If a species was available in both RefSeq and GenBank, the RefSeq version was chosen as they are better curated and annotated than their GenBank counterparts. Preference was given to assemblies with better contiguity (chromosomes > scaffolds > contigs). If multiple genomes of the same assembly level were available, the most recently released version was chosen. We have verified that the most recent version had the highest N50 in all cases. For genomes from UCSC, the latest build for a species was chosen, and one previous build if available (e.g. hg38 and hg19 builds for humans). In addition to the sequences, other information about the species such as its phylogenetic classification was recorded. All genomes were given a unique MSDB ID before further processing.

Identification of repeats

Microsatellites from genomic sequences were identified using PERF (22). PERF classifies 5356 possible permutations of 1–6 nt long DNA motifs into 501 unique classes of microsatellites based on the cyclical variations and strand of the motif sequence (Supplementary Figure S1 and Table S1). A minimum length cutoff of 12 nt was used for all motif sizes. The output of PERF is a TSV file that follows BED format specification to describe the location and other information of the identified microsatellites. In addition to the FASTA input, gene annotations were provided to PERF in GFF/GTF format (where available), based on which the nearest gene for each microsatellite was identified along with the distance to the nearest TSS (transcription start site). Further, microsatellites were classified as exonic, intronic or intergenic repeats. PERF prints all information in a TSV file, one for each genome. The final output was sorted by chromosome order before it was added to the database.

Database design

The backend of MSDB is powered by MySQL, and queried and accessed using the Python-based Django framework as the middleware. The redesigned database consists of two tables; the genomes table stores all information about the available genomes in the database, and all microsatellite information are stored in a single large repeats table (Supplementary Figure S2). The repeats table is indexed on several keys to optimize retrieval of querysets. Most queries are optimized by the ORM (object relational

model) of Django, which primarily employs lazy querysets to avoid repeated database hits. In addition, a cache is used to store information such as queryset counts to prevent unnecessary/redundant count queries. This improves the speed and responsiveness of the entire web interface and minimizes the computational overhead on the server.

Web interface

The front end of MSDB is an SPA (single-page application) built using VueJS (https://vuejs.org/) and Element UI (https://element.eleme.io/). The entire state of the website is stored in a single JavaScript object, which is valid across the tabs of the website. This ensures that the pages remember user selections and other settings until the website is reloaded again. Basic microsatellite information is precalculated and stored in species-specific JSON files, which are retrieved and used for rendering plots quickly.

DATABASE OVERVIEW AND FUNCTIONALITY

MSDB is a collection of 4 330 912 429 perfect SSRs that are $\geq$12 nt in length, from 37 680 genomes belonging to 37 262 species. The web application of MSDB is designed for interactive exploration and analysis of SSRs across genomes. The home page provides general information about MSDB and quick links to access plots and microsatellite data of commonly studied species. Other features of MSDB are accessed using various tabs of the website, as described below.

Dashboard view

Accessed via the Browse tab, the main page of MSDB is a dashboard style view that summarizes the microsatellite information of the selected species as interactive tables and charts (Figure 1). By default, the page shows microsatellite information of humans (build hg38, UCSC), which can be changed as explained in the next section. Plots and tables available on this page are depicted in Supplementary Figures S3–S8. Detailed explanation of each plot and its customization options are described on the help page of the website.

Species selection

Species can be added or changed via the species selection panel on the left side (Figure 1, left side). Users can either search for the species by their scientific or common names via the search bar or filter species of interest via the Species Table.

Modal view

Most dashboard plots have a button in the header (Figure 1, red arrow) that toggles the modal view of MSDB, which lets the user customize the plot in useful ways. Via the modal, users can access one of the most unique features of MSDB—multispecies comparison of microsatellite data (examples in Supplementary Figures S3–S6). The modal also provides options to normalize the data based on the genome size of the species to facilitate easier comparison of data across genomes of drastically different sizes.

Table view

Clicking on the 'Explore Repeats' button (Figure 1, black arrow) opens a new window with the microsatellite data of the selected species displayed as a table (Supplementary Figure S9). The filter panel at the top of the page allows the users to filter the displayed data on various attributes such as the genomic location, microsatellite motif (repeat class) or length, proximity to specific genes, genomic context or distance to transcription start sites. The table can be sorted by clicking the column header and can be exported as a TSV file using the Export Table button. The flanking sequence of the microsatellites can be obtained by selecting repeats of interest via checkboxes, and clicking the 'Get Sequence' button above the table. This launches a new window with the sequence of all selected microsatellites, with a default flank size of 100 bp on either side. The flank sizes are customizable, and the sequence format can be toggled between tabular and FASTA formats.

Data download

MSDB provides a dedicated download page to quickly retrieve microsatellite data of the desired genome. The download page shows the list of genomes in a layout similar to that of the Species Table of the Browse page (Supplementary Figure S10). For each genome, three links are provided—to launch the Table view of the genome, to download the entire data as a TSV file or to download it as a gzip compressed TSV file.

Help page

The help page of MSDB contains an extensive manual to aid new users in understanding the features and layout of the website. Annotated screenshots guide the users in navigating the site. Various sections of the page describe each chart of MSDB in detail, and also provide information on how the data were obtained or processed.

DISCUSSION

Genomic sequence information in the public domain is growing at an unprecedented scale. It is thus imperative to create resources that stay updated with the new information. MSDB is one such attempt to curate the largest collection of microsatellite data, designed from the ground up for seamless updates. At present, MSDB has data from >37 500 genomes, and we plan to update it with new genome information every 3 months. In addition to being the largest microsatellite collection, MSDB provides a user-friendly web interface for data analysis and visualization. Using its plotting functions, users can understand global microsatellite trends of a genome directly, bypassing the need to manually analyze millions of data points.

With emerging roles of microsatellites in complex regulatory functions such as gene expression and genome organization, understanding the genomic context of these elements becomes crucial. However, this information is lacking in most of the existing databases. In this version of MSDB, data from more than 27 500 genomes have been annotated

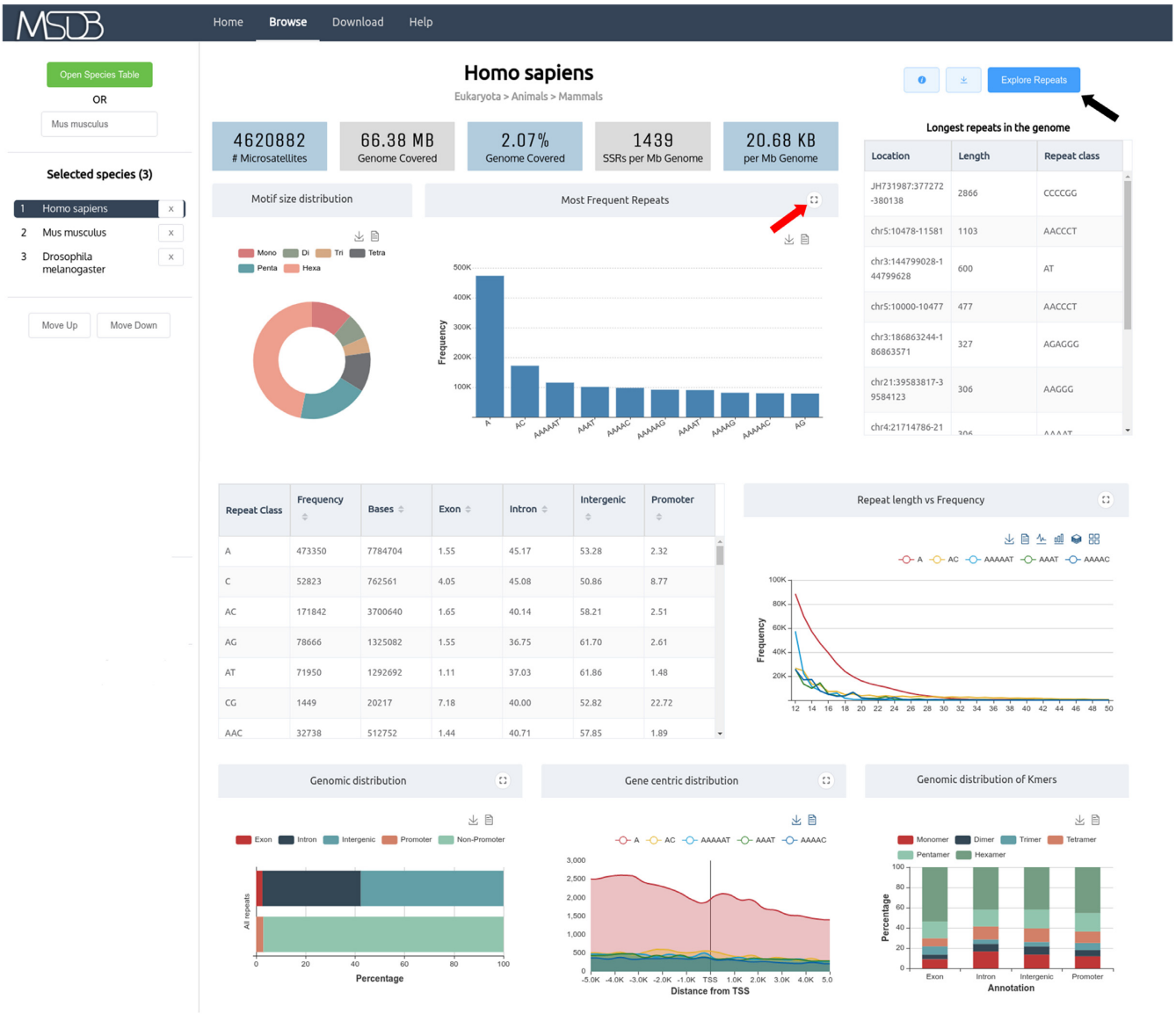

Figure 1. Browse page of MSDB showing microsatellite information of *Homo sapiens* (hg38 build of UCSC). Black arrow indicates the 'Explore Repeats' button that opens the table view. The button to toggle the modal view is indicated with the red arrow.

with genomic context information. This information is visualized in several plots and is also filterable in the table view, thus enabling researchers to get relevant microsatellite information easily. Combined with the unique ability to compare data of multiple species simultaneously, MSDB is a useful and powerful platform to analyze and visualize microsatellite data across species. The utility of MSDB is further highlighted in the following use cases:

Case study 1: length preference of AGAT repeats

Typically, the frequency of repeats reduces as the length of the repeat increases because longer repeats are more likely to accumulate mutations. However, previous studies have shown a preferential enrichment of few microsatellites at longer lengths, and such enriched repeats are likely to have functional roles (2,8). This preference can be observed for AGAT repeats in various species using the 'length versus frequency' line chart of MSDB (Supplementary Figure S11). As seen, *Drosophila melanogaster, Xenopus tropicalis* and *Gallus gallus* do not show any length preference, and the rate of decline in frequency of AGAT repeats is comparable to that of ACAG repeats, whereas *Danio rerio, Anolis carolinensis* and *Homo sapiens* show an increase in the repeat frequency after the initial decline. Furthermore, the length range becomes narrower from *Danio rerio* to *H. sapiens,* peaking at a repeat size of 40–44 bp (10–11 repeating units).

Case study 2: enrichment of AGC repeats in ruminants

We have previously observed enrichment of AGC repeats in ruminant species (unpublished data). This is depicted using the frequency bar plots of MSDB (Supplementary Fig-

ure S12). The top row shows frequency of all trimer classes in three primates (human, chimpanzee and macaques), and the bottom row shows that of three ruminants (cattle, goat and sheep). In addition to the selective enrichment of AGC repeats in ruminants (bottom row), it can be observed that the distribution of all trimer classes is similar in closely related species.

DATA AVAILABILITY

MSDB is developed using the Django framework, VueJS and MySQL. It is freely available at https://data.ccmb.res.in/msdb.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Geetha Thanu, A Padmavathi, and K Ramachary from the IT section of CCMB for help in hosting the server; Phanindhar Kundurthi and Surabhi Srivastava for critical reading of the manuscript and inputs in website design; and all members of the RKM lab for beta-testing the website.

FUNDING

Council of Scientific and Industrial Research (CSIR), India [BSC0208 (BioAge), BSC0118 (EpiHED), BSC0121 (Genesis)]. Funding for open access charge: Institute Funds.
Conflict of interest statement. None declared.

REFERENCES

1. Toth,G., Gaspari,Z. and Jurka,J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.*, **10**, 967–981.
2. Srivastava,S., Avvaru,A.K., Sowpati,D.T. and Mishra,R.K. (2019) Patterns of microsatellite distribution across eukaryotic genomes. *BMC Genomics*, **20**, 153.
3. Usdin,K. (2008) The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. *Genome Res.*, **18**, 1011–1019.
4. Collard,B.C. and Mackill,D.J. (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **363**, 557–572.
5. Hearne,C.M., Ghosh,S. and Todd,J.A. (1992) Microsatellites for linkage analysis of genetic traits. *Trends Genet.*, **8**, 288–294.
6. Zietkiewicz,E., Rafalski,A. and Labuda,D. (1994) Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. *Genomics*, **20**, 176–183.
7. Al-Mahdawi,S., Pinto,R.M., Ismail,O., Varshney,D., Lymperi,S., Sandi,C., Trabzuni,D. and Pook,M. (2008) The Friedreich ataxia GAA repeat expansion mutation induces comparable epigenetic changes in human and transgenic mouse brain and heart tissues. *Hum. Mol. Genet.*, **17**, 735–746.
8. Kumar,R.P., Krishnan,J., Singh,P.N., Singh,L. and Mishra,R.K. (2013) GATA simple sequence repeats function as enhancer blocker boundaries. *Nat. Commun.*,**4**, 1844.
9. Pathak,R.U., Mamillapalli,A., Rangaraj,N., Kumar,R.P., Vasanthi,D., Mishra,K. and Mishra,R.K. (2013) AAGAG repeat RNA is an essential component of nuclear matrix in Drosophila. *RNA Biol.*, **10**, 564–571.
10. Kumar,R.P., Senthilkumar,R., Singh,V. and Mishra,R.K. (2010) Repeat performance: how do genome packaging and regulation depend on simple sequence repeats? *Bioessays*, **32**, 165–174.
11. Ellegren,H. (2004) Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.*, **5**, 435–445.
12. Nagpure,N.S., Rashid,I., Pati,R., Pathak,A.K., Singh,M., Singh,S.P. and Sarkar,U.K. (2013) FishMicrosat: a microsatellite database of commercially important fishes and shellfishes of the Indian subcontinent. *BMC Genomics*, **14**, 630.
13. Yu,J., Dossa,K., Wang,L., Zhang,Y., Wei,X., Liao,B. and Zhang,X. (2017) PMDBase: a database for studying microsatellite DNA and marker development in plants. *Nucleic Acids Res.*, **45**, D1046–D1053.
14. Gelfand,Y., Rodriguez,A. and Benson,G. (2007) TRDB–the tandem repeats database. *Nucleic Acids Res.*, **35**, D80–D87.
15. 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.
16. Aishwarya,V. and Sharma,P.C. (2008) UgMicroSatdb: database for mining microsatellites from unigenes. *Nucleic Acids Res.*, **36**, D53–D56.
17. Aishwarya,V., Grover,A. and Sharma,P.C. (2007) EuMicroSatdb: a database for microsatellites in the sequenced genomes of eukaryotes. *BMC Genomics*, **8**, 225.
18. Mokhtar,M.M. and Atia,M.A.M. (2019) SSRome: an integrated database and pipelines for exploring microsatellites in all organisms. *Nucleic Acids Res.*, **47**, D244–D252.
19. Haft,D.H., DiCuccio,M., Badretdin,A., Brover,V., Chetvernin,V., O'Neill,K., Li,W., Chitsaz,F., Derbyshire,M.K., Gonzales,N.R. *et al.* (2018) RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.*, **46**, D851–D860.
20. Sayers,E.W., Cavanaugh,M., Clark,K., Ostell,J., Pruitt,K.D. and Karsch-Mizrachi,I. (2019) GenBank. *Nucleic Acids Res.*, **47**, D94–D99.
21. Haeussler,M., Zweig,A.S., Tyner,C., Speir,M.L., Rosenbloom,K.R., Raney,B.J., Lee,C.M., Lee,B.T., Hinrichs,A.S., Gonzalez,J.N. *et al.* (2019) The UCSC Genome browser database: 2019 update. *Nucleic Acids Res.*, **47**, D853–D858.
22. Avvaru,A.K., Sowpati,D.T. and Mishra,R.K. (2018) PERF: an exhaustive algorithm for ultra-fast and efficient identification of microsatellites from large DNA sequences. *Bioinformatics*, **34**, 943–948.