MSDB: A Comprehensive Database of Simple Sequence Repeats

Akshay Kumar Avvaru†, Saketh Saxena†, Divya Tej Sowpati*, and Rakesh Kumar Mishra

CSIR – Centre for Cellular and Molecular Biology, Hyderabad, India

Accepted: June 12, 2017

†These authors contributed equally to this work.

*Corresponding author: E-mail: tej@ccmb.res.in.

Abstract

Microsatellites, also known as Simple Sequence Repeats (SSRs), are short tandem repeats of 1–6 nt motifs present in all genomes, particularly eukaryotes. Besides their usefulness as genome markers, SSRs have been shown to perform important regulatory functions, and variations in their length at coding regions are linked to several disorders in humans. Microsatellites show a taxon-specific enrichment in eukaryotic genomes, and some may be functional. MSDB (Microsatellite Database) is a collection of >650 million SSRs from 6,893 species including Bacteria, Archaea, Fungi, Plants, and Animals. This database is by far the most exhaustive resource to access and analyze SSR data of multiple species. In addition to exploring data in a customizable tabular format, users can view and compare the data of multiple species simultaneously using our interactive plotting system. MSDB is developed using the Django framework and MySQL. It is freely available at http://tdb.ccmb.res.in/msdb.

Key words: microsatellites, simple sequence repeats, database, genomics, Django, JavaScript.

Introduction

Simple Sequence Repeats (SSRs) or microsatellites are sequences of 1–6 nt motifs repeated in tandem, present in all genomes. They comprise a significant proportion of the noncoding genome in complex organisms. Microsatellites are distributed nonrandomly in eukaryotic genomes, and a subset of SSRs show genomic enrichment in a taxon-specific manner (Toth et al. 2000). SSRs constitute 3% of the human genome (Lander et al. 2001), and their abnormal expansion in protein-coding regions causes various neurodegenerative diseases (Usdin 2008). In addition, microsatellites have been used in linkage analysis (Hearne et al. 1992), marker-assisted selection (Collard and Mackill 2008), and DNA fingerprinting (Zietkiewicz et al. 1994). Some SSRs have a role in epigenetic regulation (Al-Mahdawi et al. 2008), as enhancer blocker elements in multiple species (Kumar et al. 2013), and as important constituents of nuclear matrix (Pathak et al. 2013). SSRs evolve faster than point mutations with a bias for their elongation rather than shortening (Ellegren 2004).

Despite being an important class of regulatory elements, there exists no comprehensive database of microsatellites across various organisms (table 1A). The existing databases are either specific to a certain taxon (e.g., FishMicroSat [Nagpure et al. 2013]), Plant microsatellite database (Yu et al. 2017), or have information for a limited number of species. In addition, many of the existing databases were released several years ago, and have not been updated with current sequence information. More importantly, the data is generally provided in a static tabular format (table 1B), which makes data exploration and observation of global trends cumbersome. Finally, no existing database provides a direct means to compare the microsatellite data of multiple species, making evolutionary analysis of these elements laborious and challenging. Using the available genome sequence information from NCBI, we created a database of SSRs from almost 7000 species that is thorough, up-to-date, and enables users to easily query and explore microsatellite data of several species simultaneously.

Materials and Methods

Data Generation

Genomic sequences of various species from bacteria to humans were downloaded in FASTA format using wget from the FTP site of NCBI (NCBI Resource Coordinators 2016). In case multiple files were available for the same

Table 1

Comparison of MSDB with Other SSR Databases: (A) Number of Species and (B) Database Features.

(A) Number of Species

Kingdom/ Group	Micro Organism 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
Bacteria	1,109	0	0	0	1	0	85	4,772	0	5732
Archaea	91	0	0	0	0	0	0	217	0	514
Plants	0	80	14	110	2	0	0	0	31	74
Fungi	0	80	0	0	1	0	0	0	31	191
Protozoa	0	80	0	0	0	0	0	0	31	72
Invertebrates	0	80	0	0	9	0	0	0	31	112
Vertebrates	0	80	0	0	9	190	0	0	31	198
Viruses	1,463	0	0	0	0	0	0	0	0	0

(B) Database Features

Feature	Micro Organism Tandem Repeats Database	UgMicro SatDb	Kazusa Marker Database	Plant Microsatellite DNAs database	Tandem Repeats Database	FishMicro Sat	Polymorphic Simple Sequence Repeats Database	MICAS	EuMicroSatDb	MSDB
Interactive tables with column filters	Yes[a]	Yes[a]	No[d]	Yes[e]	Yes	No	No[d]	Yes[a]	Yes[a]	Yes
Downstream analysis 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
Taxonomic Grouping	Yes[c]	Yes	No	Yes	No	Yes	No	Yes[c]	No	Yes
Data Download	No	Yes	Yes	Yes	Yes	No	Yes	No	No	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.

species, the most recently modified file was downloaded. The species were organized in a hierarchical phylogenetic order based on the Kingdom, Group, and Subgroup they belong to.

Identification of Repeats

In total, 5356 possible permutations of 1–6 nt long DNA motifs were grouped into 501 unique classes of repeats based on the cyclical variations and strand of the motif sequence. A motif that is a palindrome or a cyclical variation of a palindrome was counted only on the "+" strand (table 2). The FASTA files downloaded from NCBI were used as input for identification of SSRs using a custom Python script which scans the sequences chromosome wise for all 501 possible repeat classes. A minimum length cut-off of 12 nt was used for all repeats. For each repeat, the chromosome name, start and stop coordinates, repeat class, actual repeat, motif length, strand, and the total length of repeat were recorded (terms explained in fig. 1). NCBI uses unique IDs to name each chromosome/scaffold in FASTA files. For example, the ID of human chromosome 1 in the current genome build is NC_000001.11. As these IDs are not directly meaningful to the user exploring the data, we developed a Python script that parses the sequence description in the FASTA file and extracts the chromosome/scaffold number wherever possible. This information was stored as the "colloquial" name of the chromosome, and is displayed to the user by default. The final output file was sorted by the chromosome order before it was appended to the database table.

Database Design

The database backend of MSDB is primarily written in and managed via MySQL using the Python-based Django

Table 2

Examples of Repeat Motif Classification Shown for a Normal Motif (ACT), a Palindrome (ACGT), and Cyclical Variation of a Palindrome (AATTCG, Variation of GAATTC)

Repeat Class	Cyclical Variations ("+" Strand)	Reverse Complement ("−" Strand)	Number of Motifs in Class
ACT	ACT, CTA, TAC	AGT, GTA, TAG	6
ACGT	ACGT, CGTA, GTAC, TACG	ACGT, CGTA, GTAC, TACG	4
AATTCG	AATTCG, ATTCGA, TTCGAA, TCGAAT, CGAATT, GAATTC	CGAATT, GAATTC, AATTCG, ATTCGA, TTCGAA, TCGAAT	6

framework in the frontend. The database consists of 16 tables in total; 10 of these are implicitly created and managed by Django for authorization and session management, while the remaining six tables are organized in a hierarchical architecture to store the relevant repeats information for all species. These six tables are normalized to reduce redundancy, lowering the size of the actual database from around 180 GB to 64 GB. Of the six tables, five store the kingdom, group, subgroup, species, and sequence/chromosome information of the genomes present in MSDB and the repeats information for all genomes is stored in one major repeats table (supplementary fig. S1, Supplementary Material online).

The Object Relational Model (ORM) provided by Django is used for querying the database. ORM uses lazy query sets to reduce the number of database hits that are required when the user navigates through the website. Query retrieval is optimized by using foreign key relations and primary key lookups with the sequences and species tables while querying the main repeats table. This increases the speed and responsiveness of the web interface as a whole and reduces the computational overhead on the server. The interactive plotting system is developed in JavaScript using d3.js (Bostock et al. 2011) and the nvd3 helper library (http://nvd3.org).

Results

Using a custom Python script, we identified a total of 650,613,391 SSRs that are ≥12 nt in length, from genomic sequence data of 6,893 species. All repeat information is stored in six backend tables, and is accessed using the front-end web application of MSDB. The MSDB web application is designed for interactive exploration and analysis of SSRs across genomes. It provides multiple features and charts for plotting, browsing and downloading the repeats data. The repeat data can be accessed via four interactive pages—View, Compare, Explore, and Download. All the four pages have a similar interface to choose one or more species of interest via dropdown menus or a search bar. All species are taxonomically grouped to facilitate easier selection and analysis of data across the evolutionary landscape.

The detailed description and usage of various interactive pages in MSDB are described below:

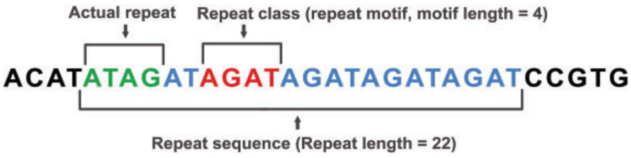

Actual repeat
Repeat class (repeat motif, motif length = 4)
ACATATAGATAGATAGATAGATAGATCCGTG
Repeat sequence (Repeat length = 22)

Fig. 1.—An example of AGAT repeat illustrating the details that were recorded by the custom repeat identification Python script.

View

The View page provides several interactive plots to analyze the SSR data of a single species.

Repeat Frequency

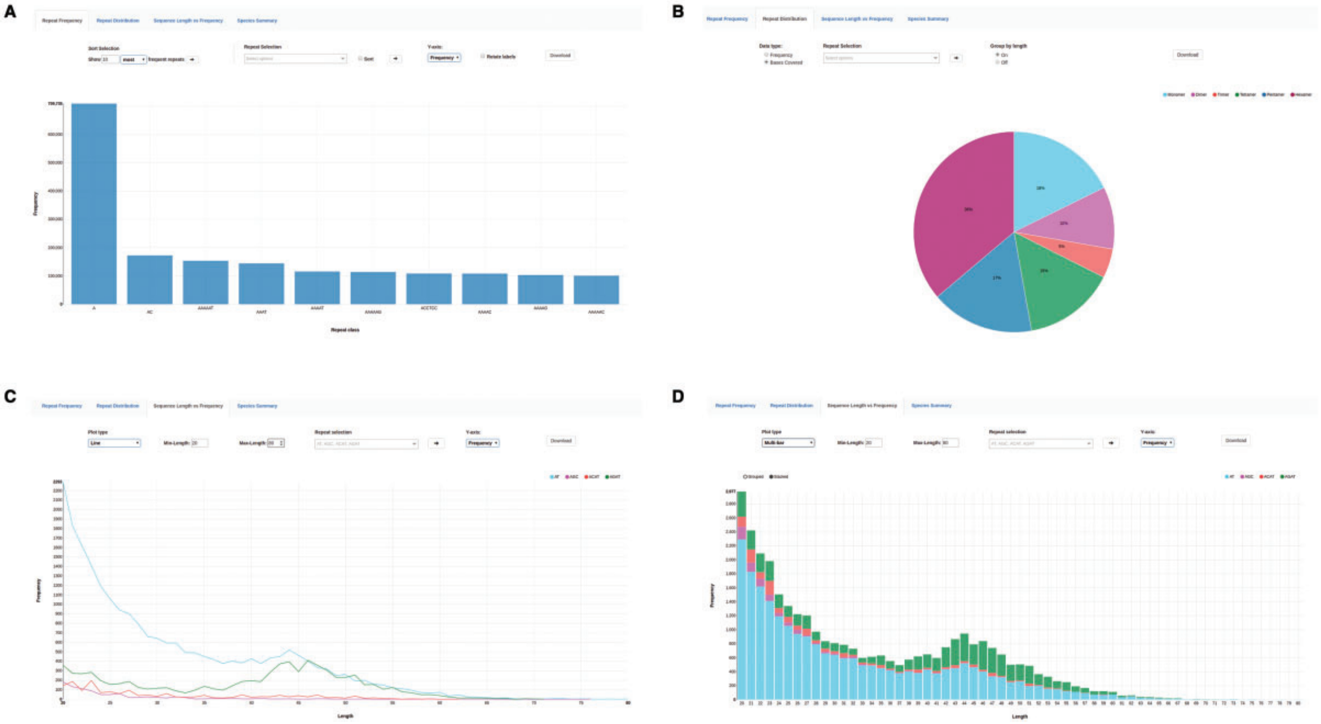
Repeat frequency represents the number of times a repeat has been found in the genome. The default bar plot shows the 10 most frequently occurring repeats in a genome, which can be changed to display any number of most or least frequent repeats (fig. 2*A*). A multi-select dropdown allows users to select repeats of their choice to be plotted on the bar chart. The plot can be further configured to be sorted either in alphabetical order of repeat classes (default) or in descending order of the repeat frequency.

Repeat Distribution

Repeat distribution plots the relative percentage of repeat frequency in the form of a pie chart. By default, the pie chart shows the percentage frequency of all repeats grouped together by the motif length (fig. 2*B*). Instead of frequency, users can choose to plot the total number of bases covered by each repeat or group of repeats. The plot can be customized to show only some repeats of interest by selection via the multi-select drop-down. Each pie slice can be toggled from the legend, upon which the percentages get recalculated based on the new total and the plot is updated accordingly.

Sequence Length versus Frequency

Sequence length refers to the length (in bp) of the repeat sequence (as explained in fig. 1). This tab displays the frequency of selected repeat(s) at each length as line plots by default, with the repeat length plotted on *X*-axis and the respective frequency plotted on *Y*-axis (fig. 2*C*). The minimum



Show 10 most frequent repeats
Y-axis: Frequency
Data type: ● Bases Covered
Group by length: ● On
Plot type: Line
Min-Length: 20
Max-Length: 80
Plot type: Multi-bar
Min-Length: 20
Max-Length: 80
Y-axis: Frequency

Fig. 2.—View page of MSDB showing various interactive plots from the repeat data of *Homo sapiens*. (*A*) Bar plot of the 10 most frequent repeats. (*B*) Pie chart showing distribution of repeat classes grouped by motif length (mono-, di-, tri-, tetra-, penta-, and hexamers). (*C* and *D*) Relation between the frequency and length of AT, AGC, ACAT, and AGAT repeats depicted as line and stacked-bar charts respectively.

and maximum lengths to be plotted are customizable by the user. The display of repeats can be toggled from the legend. The relative frequency of the selected repeats at each length can also be plotted as stacked or grouped bar charts (fig. 2*D*).

In addition to the above interactive plots, the View page provides a general summary of microsatellites in the selected species, including the total number of SSRs, most frequent repeat classes, and the longest repeats.

Compare

As a unique feature of MSDB, the compare page is designed for simultaneous and direct comparison of SSR data across several species. In the compare page, data of multiple species and/or repeats is plotted in separate panels next to each other. The plot options and controls provided are similar to that of the View page. In addition, plots can be grouped either by organism (all data of a species is plotted together in a single chart) or by repeat (data of each repeat across all selected species is plotted together). Data type can be chosen either as frequency (total number of repeats in each species) or as density (number of repeats per MB of genome). Plotting density instead of frequency is especially useful when comparing the data of species with drastically different genome sizes. The utility of this page in understanding the evolutionary trends of various SSRs is highlighted using the examples below.

The relative distribution of individual mono- and dimer repeat classes in *Homo sapiens* (human), *Mus musculus* (mouse), and *Danio rerio* (zebrafish) can be studied using repeat distribution pie charts. Instead of frequency, the data type to be plotted was chosen as the total bases covered by each repeat class. This will normalize differences that may arise due to the preference of any repeat class to be present as longer repeats. As can be observed from the pie charts (supplementary fig. S2, Supplementary Material online), A repeats are the most predominant in humans (64%) whereas AC (46%) and AT (43%) repeats take the highest share among the chosen repeats in mouse and zebrafish, respectively.

Usually, the frequency of a repeat is inversely proportional to the length of the repeat sequence. However, previous studies have shown enrichment of some repeat classes at specific lengths (Ramamoorthy et al. 2014), and have hinted towards the role of such repeats in important cellular functions. This can be observed in the line charts shown in supplementary figure S3, Supplementary Material online. AC, AG, and AT repeats of human, mouse, zebrafish and fruit fly were chosen, and their frequency at lengths 14–100 nt was plotted. The plots are grouped by repeats, and the *Y*-axis shows the density of each repeat class at a given length. AC shows a clear length preference peaking at ~40 bases in mouse, and a mild preference peaking at the same length in humans, whereas the length preference of AG is seen only in mouse peaking at ~45 bases. AT shows an interesting trend: the length preference peaks at ~45 bases and ~70 bases in mouse and zebrafish respectively.

Explore

Users can use this page to explore the repeat data of the selected species in an interactive tabular format. By default, 25 repeats and six columns (Chromosome, Start, Stop, Repeat Class, Repeat Length, and Strand) are displayed. Columns can be added or removed using the "Show/Hide columns" button. The repeats can be displayed in ascending or descending order based on their length by clicking on the column header. In addition, users can filter the data using the text input boxes provided underneath column headers. As a convenience feature for primer design or further downstream analysis of a microsatellite, clicking on any repeat entry opens a modal with the repeat and its flanking genomic sequence. The flanks are 100 bp on either side of the repeat by default, and are customizable. To easily distinguish the SSR from its flanks, the repeat sequence is displayed in red font, while the flanks are in black font. Data can be exported as a TSV file using the "Export Data button" whereas the "Reset all filters" button clears the sorting order and any filters applied.

Download

Species-wise repeat data can be downloaded from the download page of MSDB, either in a plain text format or as a gzip-compressed file. Users begin by selecting the species of their interest via the familiar interface of dropdown menus and search bar. A checkbox next to the selected species can be used to specify whether the download should be in plain text or a compressed format. The downloaded file is in a tab-separated format, and all the columns that can be toggled in "Explore" table are included in the file.

Discussion

Microsatellites comprise a discernible portion of the noncoding genome, and some of them are thought to play important functional roles. The MSDB database was developed with the aim of being an exhaustive resource of microsatellites that enables researchers to query and derive insights easily.

The entire web application and backend of MSDB was developed with performance in mind. Written using modern web technologies such as Django and MySQL, MSDB is designed to be fast, responsive, and capable of handling thousands of concurrent users in a session. The database uses normalization to eliminate redundancy of stored data, thereby reducing the computational overhead of the server. The plotting system is developed using JavaScript, which runs natively on web browsers of users' computers. Instead of querying the database for every request, the plotting system uses pre-computed data served as flat files. By isolating the frontend from backend wherever feasible, MSDB remains responsive and snappy, giving the user a lag-free experience while browsing the data.

Most biological databases in general and existing microsatellite databases in particular present the data in simple static tabular format. As the number of SSRs in each species can be in millions, understanding species-wide data trends becomes tedious, error-prone and difficult. MSDB solves this issue via its plotting system which depicts genome wide trends as clear, concise and intuitive charts. At the same time, the interactive controls provided with each chart allow a fine tuning of the plotted data to assist researchers narrow-down on their interests. Combined with the ability to plot data of multiple species together, MSDB is a unique platform to perform comprehensive analysis of simple sequence repeats across species.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

The authors thank Nitin Rani for the initial prototype of MSDB, Surabhi Srivastava for critical reading of the manuscript, and all members of the RKM lab for beta-testing the website. This work has been supported by the network projects of Council of Scientific and Industrial Research (CSIR), India—BSC0208 (BioAge), BSC0118 (EpiHED), and BSC0121 (Genesis).

Literature Cited

Al-Mahdawi S, et al. 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(5):735–746.

Bostock M, et al. 2011. D(3): data-driven documents. IEEE Trans Vis Comput Graph. 17(12):2301–2309.

Collard BC, Mackill DJ. 2008. Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos Trans R Soc Lond B Biol Sci. 363(1491):557–572.

Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 5(6):435–445.

Hearne CM, et al. 1992. Microsatellites for linkage analysis of genetic traits. Trends Genet. 8(8):288–294.

Kumar RP, et al. 2013. GATA simple sequence repeats function as enhancer blocker boundaries. Nat Commun. 4:1844.

Lander ES, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):860–921.

Nagpure NS, et al. 2013. FishMicrosat: a microsatellite database of commercially important fishes and shellfishes of the Indian subcontinent. BMC Genomics 14:630.

NCBI Resource Coordinators. 2016. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 44(D1): D7–D19.

Pathak RU, et al. 2013. AAGAG repeat RNA is an essential component of nuclear matrix in *Drosophila*. RNA Biol. 10(4):564–571.

Ramamoorthy S, et al. 2014. Length and sequence dependent accumulation of simple sequence repeats in vertebrates: potential role in genome organization and regulation. Gene 551(2):167–175.

Toth G, et al. 2000. Microsatellites in different eukaryotic genomes: survey and analysis. Genome Res. 10(7):967–981.

Usdin K. 2008. The biological effects of simple tandem repeats: lessons from the repeat expansion diseases. Genome Res. 18(7):1011–1019.

Yu J, et al. 2017. PMDBase: a database for studying microsatellite DNA and marker development in plants. Nucleic Acids Res. 45(D1):D1046–D1053.

Zietkiewicz E, et al. 1994. Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. Genomics 20(2):176–183.

Associate editor: Dan Graur