DOI 10.29042/2017-869-872

Helix Vol. 7(1): 869-872

FungREP 1.0: Online web-repository of microsatellite repeats from fungal genomes

[*1]Suresh B. Mudunuri, [2] Ratna Prabha, [3] Dhananjaya P Singh, [4]Gopala Krishna Murthy Nookala

[1]Centre for Bioinformatics Research (CBR), Sagi Rama Krishnam Raju Engineering College, A.P., India

[2, 3] ICAR-National Bureau of Agriculturally Important Microorganisms (NBAIM), U.P., India

[4]Dept. of Information Technology, Sagi Rama Krishnam Raju Engineering College, A.P., India

**Email: sureshverma@gmail.com*

Received: 10th December 2016, Accepted: 22nd December 2016, Published: 1st January 2017

Abstract

Microsatellites are tandem repetitions of short DNA motifs found in genomes of all organisms. They have been significantly used in DNA fingerprinting, genome mapping, DNA forensics, linkage analysis, genetic disease studies, are known to be involved in gene regulation, bacterial adaptation, genome evolution and are extensively used as genetic markers. Microsatellites have been used as primers for genotyping of fungal strains, to study the evolution of fungal genomes as well as in the epidemiological studies of fungal pathogens. However, the distribution, frequency and abundance of microsatellites in several fungal genomes is not well studied. So, we have constructed an exclusive database of fungal microsatellites and a user-friendly web-interface to analyse the distribution and abundance of these repeats. FungREP currently hosts microsatellite data of 910 sequences of 44 different fungi that includes complete genome sequences, whole genome shotgun sequences and contigs. More than 10 million perfect and imperfect microsatellites and their corresponding information such as the repeat unit, position information, repeat statistics, protein information, etc., can be accessed from the first version of FungREP1.0. The database can be accessed for free from www.mcr.org.in/fungrep

Keywords

Database, Microsatellite, Fungus, Genome, Tandem Repeat, web-interface

I. Introduction

Microsatellites, also called Short Tandem Repeats or Simple Sequence Repeats, are tandem repetitions of DNA motifs of size 1 to 6 bp [1]. Microsatellites are unique elements of DNA that are ubiquitous in nature and are found to be highly polymorphic [2]. The intensity of undergoing mutations in these microsatellite regions is at the rate of 10^{-6} to 10^{-2} per generation which is considered extra ordinary when compared to the mutation rate in normal DNA sequences (10^{-12} to 10^{-10} per generation) [3][4]. The high mutation rates and the polymorphic nature of microsatellites favored them to be used as popular genetic markers and have been widely used in several areas including DNA fingerprinting, strain / species identification, DNA forensics, genome mapping, paternity and evolutionary studies [5]. Additionally, they are known to be responsible for causing certain genetic disorders and cancers, for bacterial / viral adaptation, gene regulation, etc. [6-10].

However, despite of their widespread usage and applications, microsatellites in fungi are not well studied. Few researchers earlier have attempted to analyze the distribution of microsatellites in fungal genomes but these studies are limited to a certain number of genomes and are not comprehensive [2] [11-13]. Microsatellites are widely used in genotyping of fungal strains and in epidemiological studies of various fungal pathogens in humans [14] [15].

The traditional method to develop microsatellite markers is to use laboratory experiments which are time consuming and cost intensive. With the arrival of high throughput sequencing techniques like NGS, the availability of genomic sequences has led to the development of low cost and effective *in silico* detection of microsatellites in genome sequences. Microsatellites can now be extracted easily using different repeat extraction software tools available and the corresponding repeat information can be compiled into a database. Several organism specific microsatellite databases have been developed earlier that include Mouse Microsatellite Data Base of Japan (MMDBJ) [16], Eukaryotic MicroSatellite Database (EuMicroSatdb) [17], Microsatellite Database for Prokaryotes (MICdb) [18][19], Viral Microsatellite Database (VMD) [20], Insect Microsatellite Database (InSatDb) [21], Silkworm Microsatellite Database (SilkSatDb), [22], Chloroplast and Mitochondrial SSR Database (ChloroMitoSSRDB) [23][24], and Fish Microsatellite Database (FishMicrosat) [25]. However, there is no exclusive database of Fungal Microsatellites in public domain which can serve as a valuable resource for researchers to analyze these repeats and to study their role further in various fungal species.

In this paper, we present the details of the first exclusive fungal microsatellite repeats database FungREP version 1.0 that hosts perfect as well as imperfect microsatellites extracted from different fungal genomes.

2. Microsatellite Extraction

The chromosome wise fungal genomes and other related sequences have been downloaded from the NCBI ftp site (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/fungi/).

A total of 44 different species of fungi (900+ complete genomes / chromosomes / contig sequences) have been downloaded for constructing the initial version of

FungREP database. We have considered only the fungal genomes that are annotated in detail and those sequences available in
.fna and .ptt formats. The downloaded sequences have been used for detecting both perfect and imperfect microsatellites using Imperfect Microsatellite Extract or (IMEx) [26] [27]. IMEx uses a sliding window approach that identifies repeat stretches of motifs of size 1 to 6 bp in DNA sequences. IMEx is widely used by researchers worldwide and is an effective tool specifically for microsatellite extraction [28]. The IMEx algorithm allows the user to set the minimum repeat number of each repeat size as well as the imperfection level to detect perfect as well as imperfect microsatellites as per the user's requirement.

The following parameters have been used for IMEx to extract microsatellites for FungREP: *Min. Repeat Numbers*: *Mono*:12, *Di*:6, *Tri*:4, *Tetra:3, Penta:3, Hexa*:3. To detect imperfect microsatellite repeats, the *imperfection %* (*p%*) has been set to 10% and all the remaining parameters have been set to default. A total of 910 sequences have been used for extraction that includes 87 complete genome sequences, 135 complete sequences, 644 whole genome shotgun sequences, and 44 other sequences & contigs. The tool IMEx has detected 1,85,529 perfect microsatellites and 8,48,987 imperfect microsatellites altogether. In-house Python and Perl scripts have been used to automate the microsatellite extraction process.

3. Database Construction

Apart from sequence meta data, the extracted microsatellite information for each repeat such as repeat unit (motif), start and end positions of the microsatellite, nucleotide composition, iteration, tract size, imperfection percentage, protein product information (if the repeat falls in coding region), etc., have been processed using Perl programs and compiled into a relational database. The database has been constructed using MySQL 5.1. The complete fungal microsatellite information has been stored in 3 different MySQL database tables '*fungrepmeta*', '*fungrepperfect*' and '*fungrepimperfect*' that store the genome meta information, perfect microsatellite data and imperfect microsatellite data respectively. Tables 1 and 2 describe the database schema of the FungREP1.0.

The '*fungrepmeta*' holds information pertaining to each fungal sequence including its accession no, sequence name, sequence part and type, overall nucleotide composition etc., and the other two tables store the repeat information of all the sequences extracted using IMEx. FungREP version 1.0 holds more than 10 million perfect and microsatellites altogether.

TABLE1: Database Schema of FungREP Metadata Table '*fungrepmeta*'

Field Name	Sample Data	Max. Length Observed for Entry	Data Type
Accession_ No	NC_001133	12	varchar(12)
Seq_Id	330443391	9	int(11)
Seq_info	>gi|330443391|ref |NC_001133.9| Saccharomyces cerevisiae S288c	75	varchar(100)
Org_name	Saccharomyces cerevisiae S288c	45	varchar(50)
Seq_part	chromosome I	34	varchar(50)
seq_type	complete sequence	30	varchar(50)
seq_length	813184 bp	11	varchar(15)
Seq_A_per	30.33	4	Float
Seq_T_per	30.40	4	Float
Seq_G_per	19.88	4	Float
Seq_C_per	19.39	4	Float
Total_No_ Repeats	331	5	int(11)

TABLE2: **Database Schema of FungREP Repeat Info Tables '*fungrepperfect*' and '*fungrepimperfect*'**

Field Name	Sample Data	Max. Length obs for Entry	Data Type
Accession_No	NC_001133	12	varchar(12)
Consensus/Motif	AAAAG	6	char(6)
SSR_type	Penta	5	varchar(5)
Repeat_type	Perfect/Imperfect	9	varchar(10)
Motif_A_per	80.00	4	Float
Motif_T_per	0.40	4	Float
Motif_G_per	20.00	4	Float
Motif_C_per	0.00	4	Float
Iterations	3	2	int(11)
Tract_size	15	3	int(11)
Start	192480	8	int(11)
End	192494	8	int(11)
Imperfection_per	6	2	int(3)
Coding_type	Coding/Non-coding	10	varchar(10)
Protein_ID	6319300 (0 if non-coding)	9	varchar(15)

* Field 'Imperfection_per' is present in table storing imperfect SSRs.
To facilitate quick querying of repeat information, indexes have been created on the fields Accession_No, Motif, Start and End of the MySQL tables storing perfect and imperfect microsatellites.

4. Web Interface

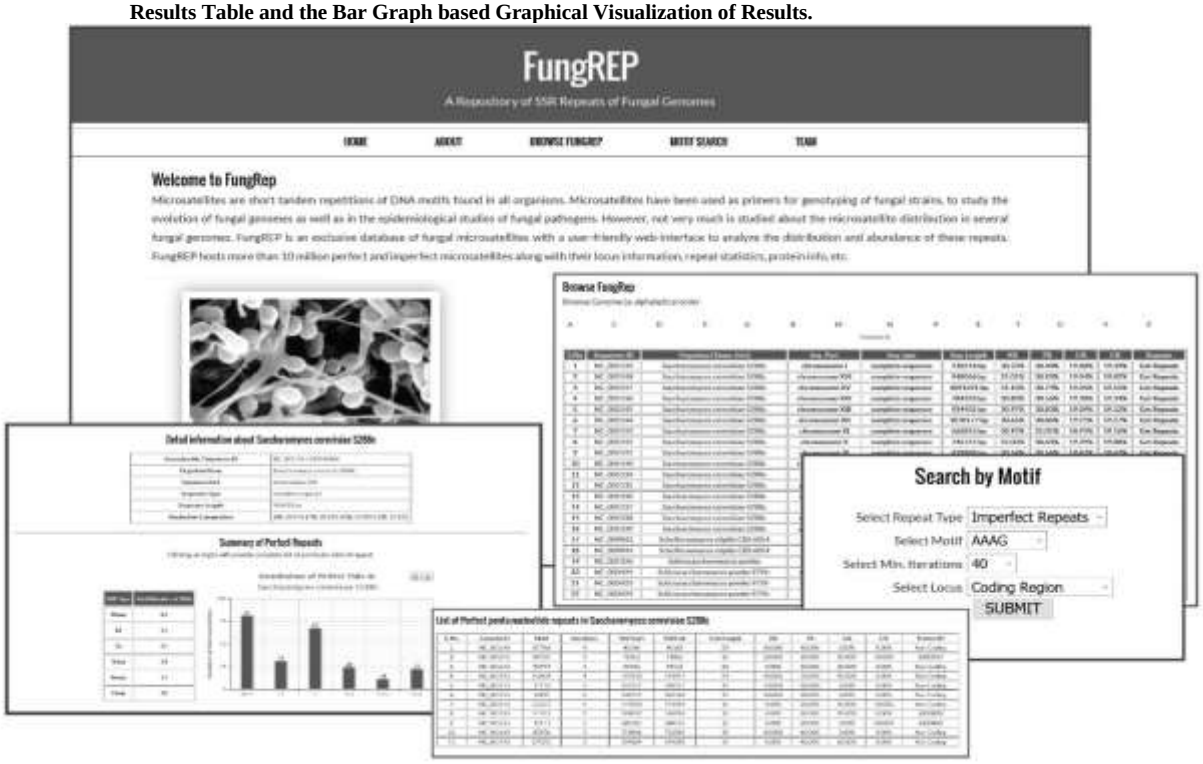
The backend database of FungREP has been connected to a user-friendly web interface using which the users can easily browse through the genomes for their corresponding microsatellite information. The front-end interface has been developed using HTML, Cascading Style Sheets (CSS) & JavaScript and the server-side coding has been implemented using PHP. The web interface is provided with a simple navigation layout with two major functionalities – 'Browse FungREP' and 'Motif Search'. Using 'Browse FungREP', the users can simply browse through the genomes arranged in alphabetical order of organism names and the complete repeat information of an organism is presented in tabular format. Clicking on a genome or a chromosome sequence of interest will take you to a repeat summary page that displays the summary of perfect and imperfect microsatellites in tabular and in graphical format (bar-chart). The bar charts can be very helpful for better visualization of the distribution of microsatellites by repeat size for that sequence. The bar-chart based graphical visualization has been achieved using the advanced JQuery Chart Library called HighCharts (www.highcharts.com/products/highcharts).
To facilitate primer design for researchers based on a motif, a 'Motif Search' option has been designed using which.

users can find microsatellites of a selected motif in all the fungal genomes of FungREP database. User can select a motif (say GACC, AT etc.) of choice, repeat type (perfect/imperfect), and specify the minimum repeat number of the repeat. An option to find repeats in coding or non-coding regions has also been provided. The corresponding repeat information of that motif in all the genomes will be displayed in a tabular format. Server-side scripts written in AJAX have been used to facilitate dynamic and quick display of Motif Search results. A screenshot of the web interface of FungREP has been illustrated in Figure 1.

5. Availability

The FungREP database and web-server have been hosted on a 64-bit Linux server with 8 core Xeon E-5 Processor, 64 GB RAM and 2 TB HDD. The server is pre-installed with MySQL 5.1, PHP 5.3 and Apache 2.2 webserver. The database will be updated by adding microsatellite repeat information of new fungal genomes periodically. Additional modules to facilitate online analysis of fungal microsatellite will be developed and integrated in the future versions of FungREP. FungREP can be accessed for free from http://www.mcr.org.in/fungrep

FIGURE 1: **Screenshot of FungREP web-interface depicting the FungREP Homepage, Genome Browse, Motif Search, Results Table and the Bar Graph based Graphical Visualization of Results.**



6. Conclusion

Microsatellites are widely being used in genotyping of fungal strains, in studies related to fungal pathogenesis and to understand their genome evolution. To provide a valuable resource for researchers working in fungal genomics, we have constructed an online fungal microsatellite database named FungREP that hosts microsatellite repeat information of various fungal genomes. This paper presents the details of FungREP database construction and the different modules integrated to the web interface for microsatellite data analysis and visualization. The database will be updated from time to time by including the repeat information from latest sequenced fungal genomes.

7. Acknowledgments

The authors thank Dr. Gaurav Sablok who has been instrumental in carrying out this work and for providing necessary support during the development of the database. The work is supported by SERB, Department of Science & Technology (DST), Govt. of India (Grant No: /ECR/2016/000346).

8. References

[1] Tautz D, Schlötterer, C. Simple sequences. *Current opinion in genetics & development*, *4*(6), 832-837, 1994.

[2] Tóth G, Gáspári Z, Jurka J. Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Research*, *10* (7), 967-981, 2000.

[3] Schlötterer C, Ritter R, Harr B, Brem G. High mutation rate of a long microsatellite allele in Drosophila melanogaster provides evidence for allele-specific mutation rates. *Molecular Biology and Evolution*, 15(10), 1269-1274, 1998.

[4] Li YC, Korol AB, Fahima T, Beiles A, Nevo E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Molecular Ecology*, 11(12), 2453-65, 2002.

[5] Estoup A, Cornuet JM. Microsatellite evolution: inferences from population data. *Microsatellites: evolution and applications*, 1999, Pg 49-65.

[6] Martin P, Makepeace K, Hill SA, Hood DW, Moxon ER. Microsatellite instability regulates transcription factor binding and gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (10), 3800-4, 2005.

[7] Jarne P, Lagoda PJL. Microsatellites, from molecules to populations and back. *Trends Ecol. Evol.*, 11 (10), 424–429, 1996.

[8] Kashi Y, King DG. Simple sequence repeats as advantageous mutators in evolution. *Trends Genet*, 22(5), 253–9, 2006.

[9] Sreenu VB, Kumar P, Nagaraju J, Nagarajaram HA. Microsatellite polymorphism across the M. tuberculosis and M. bovis genomes: implications on genome evolution and plasticity. *BMC Genomics*, 7(1), 2006.

[10] Davis CL, Field D, Metzgar D, Saiz R, Morin PA, Smith IL, Spector SA, Wills C. Numerous length polymorphisms at short tandem repeats in human cytomegalovirus. *Journal of Virology*, 73(8), 6265-70, 1999.

[11] Karaoglu H, Lee CM, Meyer W. Survey of simple sequence repeats in completed fungal genomes. *Molecular Biology and Evolution*, 22(3), 639-49, 2005.

[12] Field D, Wills C. Long, polymorphic microsatellites in simple organisms. *Proceedings of the Royal Society of London B: Biological Sciences*, 263(1367), 209-15, 1996.

[13] Lim S, Notley-McRobb L, Lim M, Carter DA. A comparison of the nature and abundance of microsatellites in 14 fungal genomes. *Fungal Genetics and Biology*, 41(11), 1025-36, 2004.

[14] Bart-Delabesse E, Humbert JF, Delabesse É, Bretagne S. Microsatellite markers for typing Aspergillus fumigatus isolates. *Journal of Clinical Microbiology*, 36(9), 2413-8, 1998.

[15] Nascimento É, Martinez R, Lopes AR, de Souza Bernardes LA, Barco CP, Goldman MH, Taylor JW, McEwen JG, Nobrega MP, Nobrega FG, Goldman GH. Detection and selection of microsatellites in the genome of Paracoccidioides brasiliensis as molecular markers for clinical and epidemiological studies. *Journal of Clinical Microbiology*, 42(11), 5007-14, 2004.

[16] Sakai T, Miura I, Yamada-Ishibashi S, Wakita Y, Kohara Y, Yamazaki Y, Inoue T, Kominami R, Moriwaki K, Shiroishi T, Yonekawa H. Update of mouse microsatellite database of Japan (MMDBJ). *Experimental Animals*, 53(2), 151-4, 2004.

[17] Aishwarya V, Grover A, Sharma PC. EuMicroSatdb: A database for microsatellites in the sequenced genomes of eukaryotes. *BMC Genomics*, 8(1), 1, 2007.

[18] Sreenu VB, Alevoor V, Nagaraju J, Nagarajaram HA. MICdb: database of prokaryotic microsatellites. *Nucleic Acids Research*, 31(1), 106-8, 2003.

[19] Mudunuri SB, Patnana S, Nagarajaram HA. MICdb3. 0: a comprehensive resource of microsatellite repeats from prokaryotic genomes. *Database*, 2014, bau005, 2014.

[20] Suresh BM, Allam AR, Pallamsetty S, Priyatosh M, Nagarajaram HA. VMD: Viral Microsatellite Database A Comprehensive Resource for All Viral Microsatellites. *Journal of Computer Science & Systems Biology*, 2, 283-286, 2009.

[21] Archak S, Meduri E, Kumar PS, Nagaraju J. InSatDb: a microsatellite database of fully sequenced insect genomes. *Nucleic Acids Research*, 35(1), D36-9, 2007.

[22] Prasad MD, Muthulakshmi M, Arunkumar KP, Madhu M, Sreenu VB, Pavithra V, Bose B, Nagarajaram HA, Mita K, Shimada T, Nagaraju J. SilkSatDb: a microsatellite database of the silkworm, Bombyx mori. *Nucleic Acids Research*, 33(1), D403-6, 2005.

