DATABASE Open Access

FishMicrosat: a microsatellite database of commercially important fishes and shellfishes of the Indian subcontinent

Naresh Sahebrao Nagpure[1], Iliyas Rashid[1], Rameshwar Pati[1], Ajey Kumar Pathak[2], Mahender Singh[1*], Shri Prakash Singh[2] and Uttam Kumar Sarkar[2]

Abstract

Background: Microsatellite DNA is one of many powerful genetic markers used for the construction of genetic linkage maps and the study of population genetics. The biological databases in public domain hold vast numbers of microsatellite sequences for many organisms including fishes. The microsatellite data available in these data sources were extracted and managed into a database that facilitates sequences analysis and browsing relevant information. The system also helps to design primer sequences for flanking regions of repeat loci for PCR identification of polymorphism within populations.

Description: FishMicrosat is a database of microsatellite sequences of fishes and shellfishes that includes important aquaculture species such as *Lates calcarifer*, *Ctenopharyngodon idella*, *Hypophthalmichthys molitrix, Penaeus monodon*, *Labeo rohita*, *Oreochromis niloticus*, *Fenneropenaeus indicus* and *Macrobrachium rosenbergii*. The database contains 4398 microsatellite sequences of 41 species belonging to 15 families from the Indian subcontinent. GenBank of NCBI was used as a prime data source for developing the database. The database presents information about simple and compound microsatellites, their clusters and locus orientation within sequences. The database has been integrated with different tools in a web interface such as primer designing, locus finding, mapping repeats, detecting similarities among sequences across species, and searching using motifs and keywords. In addition, the database has the ability to browse information on the top 10 families and the top 10 species, through record overview.

Conclusions: FishMicrosat database is a useful resource for fish and shellfish microsatellite analyses and locus identification across species, which has important applications in population genetics, evolutionary studies and genetic relatedness among species. The database can be expanded further to include the microsatellite data of fishes and shellfishes from other regions and available information on genome sequencing project of species of aquaculture importance.

Keyword: Microsatellite loci, Tandem repeats, Polymorphism, Primer, Flanking region, Fish

* Correspondence: mahender@scientist.com
[1]Division of Molecular Biology and Biotechnology, National Bureau of Fish Genetic Resources, Lucknow 226002, India
Full list of author information is available at the end of the article

Background

Microsatellites are observed in almost all known eukaryotic and prokaryotic genomes, present in both coding and non-coding regions. They have a high mutation rate (between 10^{-3} and 10^{-4} mutations per gamete per generation) that generates and maintains extensive length polymorphism [1,2]. This makes microsatellite a powerful genetic marker for a variety of applications like population genetics, genetic linkage mapping, parentage assignment, marker assisted selection, molecular breeding, and allele mining [3,4]. A microsatellite locus generally varies in length between 5 to 40 repeats. Di-, tri- and tetranucleotide repeats are the most common choices for molecular genetic studies. Dinucleotides are an abundant type of microsatellite repeat found in most vertebrates, whereas trinucleotide repeats are most abundant in plants [5-7]. Microsatellites represent ideal molecular markers because they have multiple alleles that are highly polymorphic among individuals and loci that are highly abundant and dispersed evenly throughout eukaryotic genomes. The major drawback of using microsatellite is that for most species they need to be developed *de novo*, a process that is often costly and protracted [8]. Efforts have been made worldwide to compile and develop online and offline microsatellite databases of biological organisms [9-15]. Valuable studies have been done in fishes such as microsatellite genetic linkage maps [16-20], characterization and identification of microsatellites [21-24] and cross-species microsatellite locus identification [25-27]. Despite the importance of microsatellite markers, meagre efforts have been made to develop a microsatellite database of the fishes except *Danio rerio* [28], *Cyprinus carpio* [29] and Fishgen [30].

In this article, we describe the development of a microsatellite database (FishMicrosat) for population genetics and stock management using LAMPP (Linux-Apache-MySQL-PHP-Perl) technology and GenBank of NCBI as a data source to extract the microsatellite data. FishMicrosat is a unique database of microsatellite sequences that covers commercially important fish and shellfish species of the Indian subcontinent. The database currently contains 4398 sequences of 41 species belonging to 15 families and provides information on the type of repeat in terms of mono-, di-, tri-, tetra-, penta and hexanucleotide, simple and compound microsatellite, along with the characteristic of repeats namely size, region, pattern & unit. Additionally, algorithms were implemented for finding loci across species, based on the presence of identical simple sequence repeats (SSRs) with the same or varying frequencies of repeat units but conserved flanking regions. The database is regularly updated based on the release of new records in GenBank for the existing 41 species as well as the addition of new species belonging to the Indian sub continent. It is expected that the database will be a valuable resource in many aspects of fish genetic research of the Indo-Pacific region, Bay of Bengal and Arabian Sea.

Construction and content

Data source

Microsatellite sequences of fish and shellfish species were downloaded from Entrez of NCBI [31] using the keyword search ‘Fish microsatellite’ under nucleotide. Files were downloaded in GenBank and FASTA format for annotation and sequences respectively. Further, a Perl program (SpciesExtractor.pl) was written and used for data extraction for only important species found in the Indian subcontinent from the downloaded files. Other physical information about the species like habitat, distribution, IUCN Red List status was collected from FishBase [32]. Another Perl parsing program (InformationParser.pl) was developed to extract the information from the files according to the database schema and manage the data into the database. These Perl programs are used by the database administrator for updating based on new releases of microsatellite sequences for existing and new species.

Design and development

Database

In order to manage the data, MySQL, a relational database management system, was used for building the database. Tables were designed and relationships among tables were created using unique, primary and foreign keys. Five tables were designed to store the information about microsatellite sequences and species. Table ‘fishinfo’ contains the physical and phenotypical information; ‘satellite_sources’ holds details molecular information about microsatellites; ‘satellite’ works as a bridge between tables ‘fishinfo’ and ‘satellite_sources’; ‘taxonomy’ shows systematic information of the species and acts as a sub table of ‘fishinfo’. And finally the table ‘repeats’ covers the data about repeats of all microsatellites sequences obtained by using the repeat analysis program ‘MISA’ [33] as shown in Figure 1.

Web interface

A web interface integrated with the database was designed and developed to retrieve and access the information of interest using web technologies like PHP, HTML, CSS, JavaScripts, DBI (Database Interface), CGI (Common Gateway Interface), GD (Graphic Design) and Perl. The web interface also incorporates the different tools for searching, viewing and analysing the microsatellite data (Figure 1).

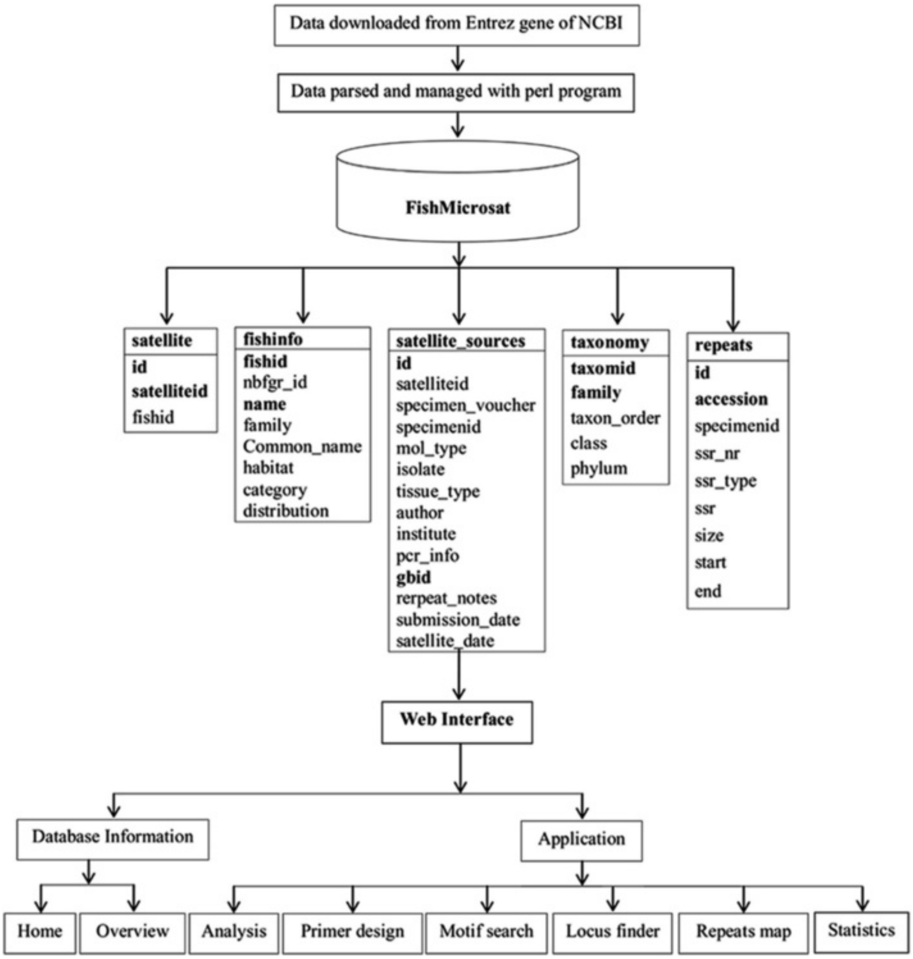


Figure 1 **Architecture and data flow representation in 'FishMicrosat'.**

Identification of microsatellite loci across species

The microsatellite loci among the existing sequences were identified by implementing an algorithm into a program (locusfinder.pl) using Perl. In order to construct the algorithm, a microsatellite sequence of selected species was divided in parts (a) motif of repeats region (b) 25 bp flanking sequence upstream and (c) 25 bp flanking sequence downstream to the repeat region. The repeat region and motifs present in sequences of selected species were fetched from the 'repeats' table of the database sequentially for retrieving identical target motifs and its sequences. Further, the conserved flanking regions were checked in query as well as target sequences. The evolutionary conservation of the flanking region allows hetero specific identification of SSRs [34]. These conserved flanking regions have been used for designing PCR primers for microsatellite amplification and genotyping of individuals of the same species as well as across species [35,36]. Thus, to identify loci across species, an algorithm was designed by considering the approach for example 'ABC' as a repeat pattern and 'L' the number of repeat units in a selected query sequence. The same repeat pattern 'ABC' was used to check its availability and repeat frequency (denoted as P) in the target sequence (Figure 2). Here, because the repeat frequency may be polymorphic, the value of the repeat frequency in the selected query sequence (L) may or may not be equal to the repeat frequency in the target sequence (P) i.e. $L = P$ or $L! = P$. The algorithm uses a 25 bp length of flanking region on either side, which is sufficient for amplification of a microsatellite locus in a PCR reaction for laboratory validation. The loci identification program supports the findings of the previous studies that microsatellite repeats vary within and between different genomes of organisms [37,38].

Search and analysis

Apart from locus finding across species, other search and analysis modules like 'Keyword search', 'Repeat analysis and primer', 'Motifs search', and 'Repeats map' were implemented and integrated in the web interface for browsing information. The 'Keyword search' takes a

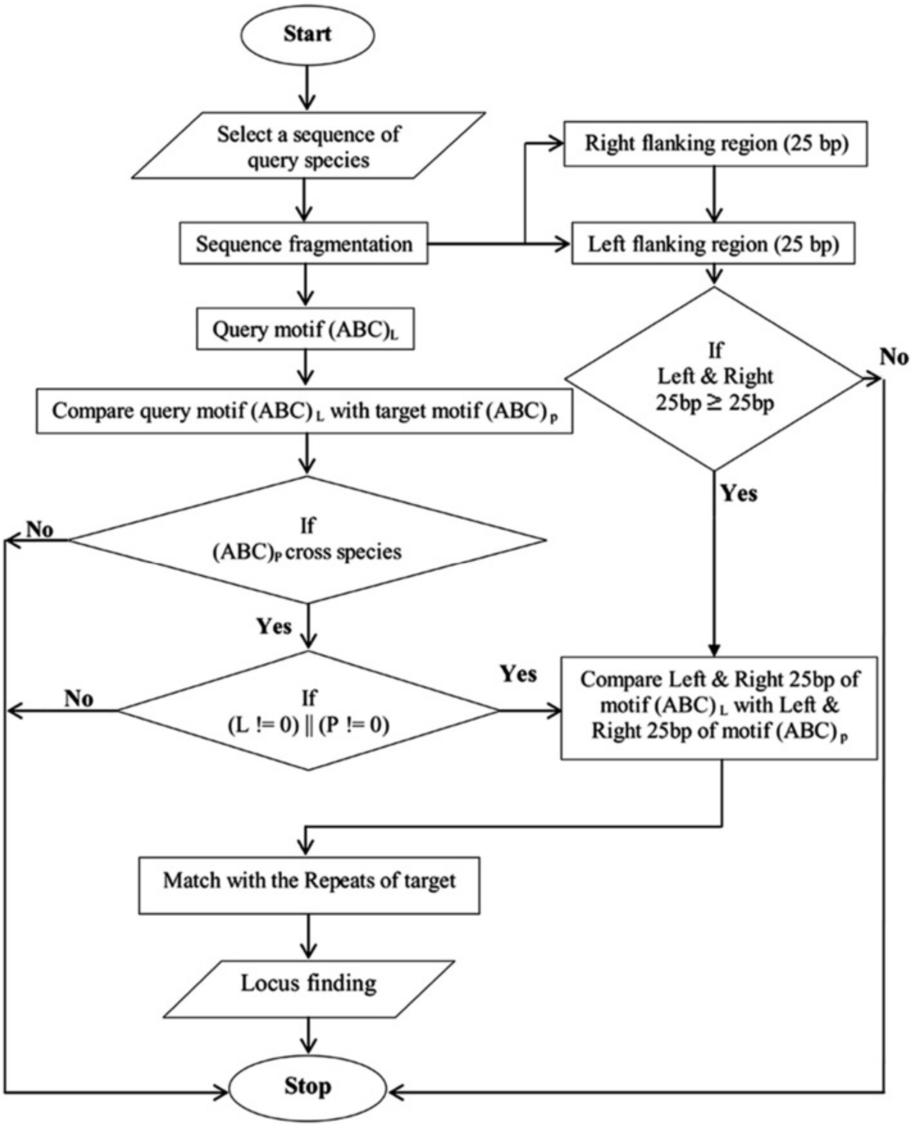

Start
Select a sequence of query species
Sequence fragmentation
Right flanking region (25 bp)
Left flanking region (25 bp)
Query motif (ABC)$_L$
If Left & Right 25bp ≥ 25bp
No
Yes
Compare query motif (ABC)$_L$ with target motif (ABC)$_P$
If (ABC)$_P$ cross species
No
Yes
If (L != 0) || (P != 0)
No
Yes
Compare Left & Right 25bp of motif (ABC)$_L$ with Left & Right 25bp of motif (ABC)$_P$
Match with the Repeats of target
Locus finding
Stop

Figure 2 Flowchart of the locus finder algorithm.

word like 'fish name', 'common name', 'family', 'accession number' and 'author' typed in by the user as input and performs the search. The 'Motif search' takes input parameters like motifs, repeat length and repeat type (simple or composite) and returns the result with the help of regular expression programming and SQL (Structured Query Language) concepts. The 'Repeats map' was developed using Perl, MISA and Blastn [39] programs for identifying similarities among the sequences and mapping the repeats. The Blastn program uses 'blastmsdb' database, which is a blast compatible and created apart from the main database by using 'formatdb' program of blast package. 'Repeats map' analyzes and process the input query sequence through the MISA program to generate the repeats. If repeats are found, it further leads to alignment with other similar sequences; otherwise the program terminates with a warning message. Primer3 program [40,41] was used for primer designing, and a standalone version was downloaded [42] to compute multiple sets of forward and reverse primers for microsatellite loci along with melting temperature (Tm), GC content, start position and product size. These generated primers can be used in PCR reactions for identification of polymorphic loci for genotyping of individuals.

Implementation of 'statistics'

The MISA program was implemented to ascertain the frequently occurring repeat types and repeat information from all the sequences in FishMicrosat. The results obtained from MISA were parsed and stored in the 'repeats' table of the database. The 'GD graph' module was used to design and dynamically display the frequency of different types of repeats (mono to hexa) in the 'Pie diagram'. The 'Pie diagram' presents the frequency of each type of repeat and is revised when the database is updated.

Discussion and utility

FishMicrosat covers 4398 microsatellite records of 41 commercially important aquaculture species belonging to 15 families (Table 1). The Home page of the web interface of FishMicrosat integrated with different analytical modules, presents the numerical statistics of the top 10 families and species in addition to providing information on updating and current status of the database. The 'Top 10 FishMicrosat families' displays ten families which has the largest number of species in FishMicrosat and the 'Top 10 FishMicrosat Species' displays the ten species for which the largest number of specimen records are available in the database. The page also provides an overview of FishMicrosat and its features. The analytical tools like motif, sequence similarity search and repeat mapping, and finding microsatellite loci across species were integrated to increase the utility and scope of the database.

Browsing specimen information

The specimen records of the species of interest can be viewed by using the species instantiation index under the 'Record overview' menu item in the web interface. Hyperlinked navigational indexes by first letter of the generic name have been provided to find the species of interest along with the number of specimen records in square brackets. Further, selection of each species name is hyperlinked and a mouse click over the species name presents information on family, common name, habitat, distribution, microsatellite repeats, its region and size, sequence length, authors and NCBI accession number. The NCBI accession number for each specimen record also has a hyperlink to NCBI. The 'Top 10 FishMicrosat families' and 'Top 10 FishMicrosat Species' on the home page of the web interface provides other means of viewing information about the species and its specimens.

Keyword search

The keyword search works on keywords like species name, common name, family name, accession number, and author for retrieving the information from the database. Different views have been created for all these keywords to present relevant information from the database. For example the species name or common name as an input keyword leads to record overview. The 'author view' displays a list of all the species on which the particular author worked and also displays the specimen records which corresponds to the listed species. Similarly, family name and accession number keywords also lead to respective views (Figure 3).

Repeat analysis and primer design

The menu item 'Analysis & primer' (Figure 4A) detects repeats in the sequences and designs the primer for the selected repeat locus. Thus, to obtain the repeat information and design primers for a specific repeat, the end user selects a species of interest starting with a generic name. Clicking on the species name provides a table that contains information such as accession no., SSR no., SSR type, SSR motif, SSR size, position, sequence length and a link for primer design for each specimen (Figure 4B). For primer designing, the 'Primer3' standalone program computes primers upon user request for microsatellite sequences that have suitable length of flanking regions and ample GC content in that region; otherwise the request is rejected with a warning message (Figure 4C). The program displays a list of multiple primers along with respective values for Tm, GC content, start position and product size (oligo size). The primer sequences will be useful in determining the alleles and finding of loci across species.

Motif search

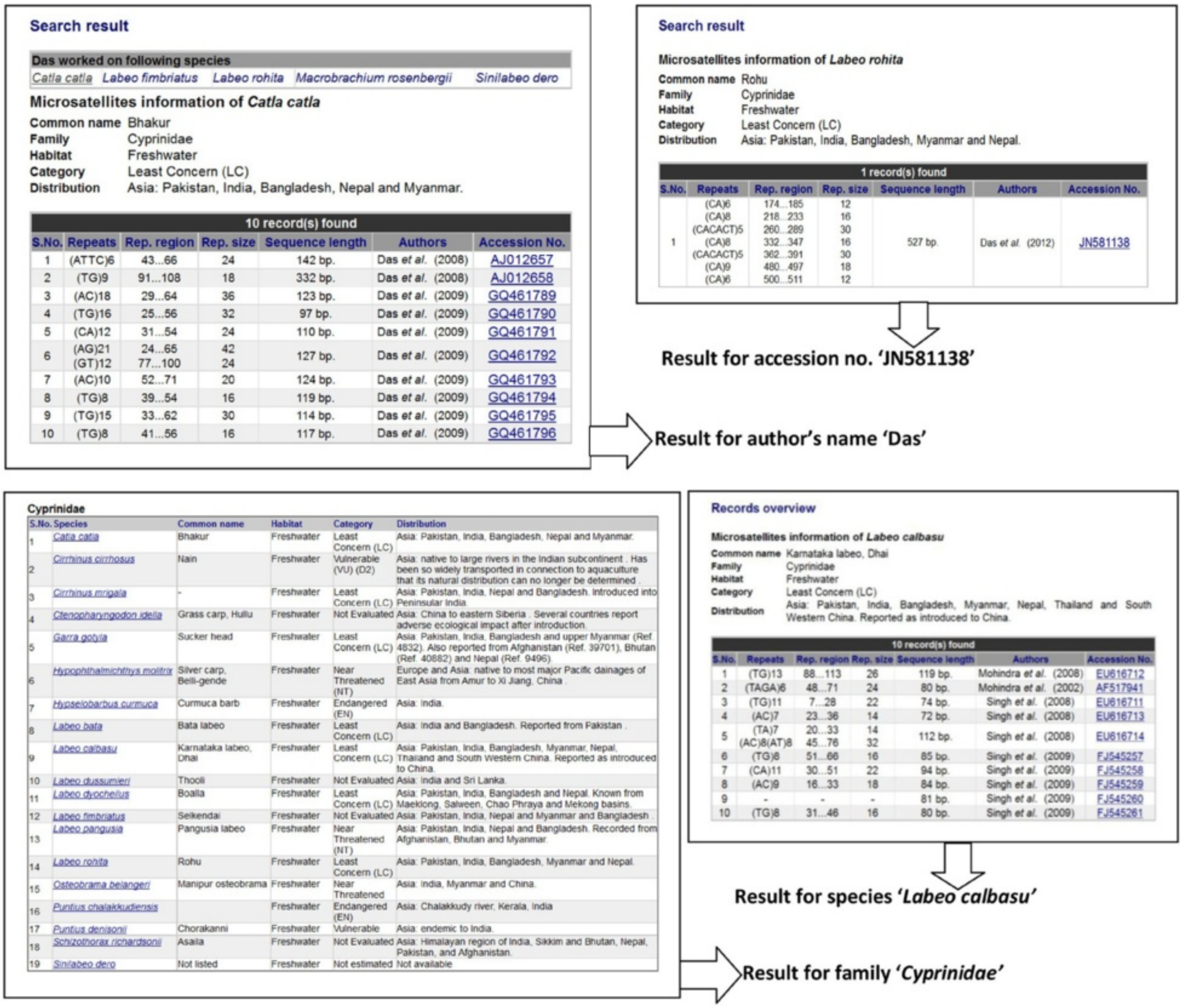
A repeat motif can be searched from the menu item 'Motif search' integrated in the web interface (Figure 5A). It searches repeats in all microsatellite sequences present in the database and fetches information on species name, family, repeats, size, repeat region, NCBI references, and primers for SSRs (Figure 5B). Three input values are required under 'Motif search': 'Motif' for nucleotide pattern (mono-hexa), 'Length' for number of nucleotides (i.e. > 10) and 'Repeats type' (simple or compound). The search results provide a primer link that leads to the design of the primer for the corresponding repeat type.

Locus finder

The Locus finder tab accesses identical microsatellite loci across species based on conserved flanking sequences (approximately 25 bp long) on either side of the polymorphic loci. The program uses two input parameters; length of flanking region and species name. Finding the identical microsatellite locus in other species existing in the FishMicrosat database, is highly useful for cross species amplification of microsatellite loci. For example the sequence of *Labeo rohita* (GenBank accession no. AY291597) and *Catla catla* (GenBank accession no. AJ294957) contain the same motif with conserved flanking regions on the parameter: flanking region '20 bp' and species '*Labeo rohita*'. The sequence alignments are

Table 1 Distribution of SSR's by species

S no.	Species	Simple							Compound
		Sequences	Mono	Di	Tri	Tetra	Penta	Hexa	
1	*Amphiprion sebae*	1	0	1	0	0	0	0	0
2	*Catla catla*	22	1	25	1	4	0	0	0
3	*Channa marulius*	2	0	2	0	0	0	0	0
4	*Chitala chitala*	31	0	31	1	6	0	0	4
5	*Cirrhinus cirrhosus*	7	0	6	0	0	0	0	0
6	*Cirrhinus mrigala*	11	0	9	1	1	0	0	0
7	*Ctenopharyngodon idella*	500	18	523	29	70	0	0	31
8	*Epinephelus fuscoguttatus*	312	43	167	29	11	2	0	5
9	*Fenneropenaeus indicus*	182	11	96	26	86	1	0	18
10	*Garra gotyla*	26	2	13	0	0	0	0	1
11	*Heteropneustes fossilis*	10	5	11	0	0	0	0	0
12	*Hippocampus kuda*	12	1	10	1	2	0	0	0
13	*Hippocampus trimaculatus*	12	1	12	1	2	0	0	0
14	*Horabagrus brachysoma*	11	0	12	2	0	0	0	0
15	*Hypophthalmichthys molitrix*	402	42	425	18	59	1	0	27
16	*Hypselobarbus curmuca*	8	1	7	2	1	0	0	0
17	*Katsuwonus pelamis*	5	1	7	1	0	0	0	0
18	*Labeo bata*	4	0	4	0	0	0	0	0
19	*Labeo calbasu*	10	0	8	0	1	0	0	1
20	*Labeo dussumieri*	7	0	2	2	1	0	0	1
21	*Labeo dyocheilus*	10	0	6	0	1	0	0	1
22	*Labeo fimbriatus*	17	1	28	0	0	0	0	0
23	*Labeo pangusia*	11	0	10	0	0	0	0	1
24	*Labeo rohita*	531	9	689	169	38	0	5	26
25	*Lates calcarifer*	975	43	903	53	157	8	0	51
26	*Macrobrachium malcolmsonii*	5	0	5	0	0	0	0	2
27	*Macrobrachium rosenbergii*	180	20	151	106	3	0	0	16
28	*Mystus oculatus*	2	0	1	0	0	0	0	1
29	*Mystus tengara*	1	0	1	0	0	0	0	0
30	*Notopterus notopterus*	8	0	13	0	1	0	0	0
31	*Oreochromis niloticus*	303	14	399	7	2	0	0	11
32	*Osteobrama belangeri*	3	0	2	0	0	0	0	0
33	*Pampus argenteus*	31	3	45	0	3	0	0	2
34	*Pangasius pangasius*	28	1	37	2	0	0	0	1
35	*Penaeus monodon*	618	59	366	155	104	3	6	68
36	*Puntius chalakkudiensis*	11	0	21	1	0	0	0	1
37	*Puntius denisonii*	11	0	17	0	0	0	0	0
38	*Rita gogra*	2	0	3	0	0	0	0	0
39	*Schizothorax richardsonii*	53	1	119	3	0	0	0	7
40	*Sinilabeo dero*	20	0	15	0	1	0	0	3
41	*Sperata aor*	3	0	5	0	0	0	0	0
Total	**41**	**4398**	**277**	**4207**	**610**	**554**	**15**	**11**	**279**



Search result

Das worked on following species				
Catla catla	*Labeo fimbriatus*	*Labeo rohita*	*Macrobrachium rosenbergii*	*Sinilabeo dero*

Microsatellites information of *Catla catla*

Common name	Bhakur
Family	Cyprinidae
Habitat	Freshwater
Category	Least Concern (LC)
Distribution	Asia: Pakistan, India, Bangladesh, Nepal and Myanmar.

10 record(s) found

S.No.	Repeats	Rep. region	Rep. size	Sequence length	Authors	Accession No.
1	(ATTC)6	43...66	24	142 bp.	Das *et al.* (2008)	AJ012657
2	(TG)9	91...108	18	332 bp.	Das *et al.* (2008)	AJ012658
3	(AC)18	29...64	36	123 bp.	Das *et al.* (2009)	GQ461789
4	(TG)16	25...56	32	97 bp.	Das *et al.* (2009)	GQ461790
5	(CA)12	31...54	24	110 bp.	Das *et al.* (2009)	GQ461791
6	(AG)21 (GT)12	24...65 77...100	42 24	127 bp.	Das *et al.* (2009)	GQ461792
7	(AC)10	52...71	20	124 bp.	Das *et al.* (2009)	GQ461793
8	(TG)8	39...54	16	119 bp.	Das *et al.* (2009)	GQ461794
9	(TG)15	33...62	30	114 bp.	Das *et al.* (2009)	GQ461795
10	(TG)8	41...56	16	117 bp.	Das *et al.* (2009)	GQ461796

Result for author's name 'Das'

Search result

Microsatellites information of *Labeo rohita*

Common name	Rohu
Family	Cyprinidae
Habitat	Freshwater
Category	Least Concern (LC)
Distribution	Asia: Pakistan, India, Bangladesh, Myanmar and Nepal.

1 record(s) found

S.No.	Repeats	Rep. region	Rep. size	Sequence length	Authors	Accession No.
1	(CA)6 (CA)8 (CACACT)5 (CA)8 (CACACT)5 (CA)9 (CA)6	174...185 218...233 260...289 332...347 362...391 480...497 500...511	12 16 30 16 30 18 12	527 bp.	Das *et al.* (2012)	JN581138

Result for accession no. 'JN581138'

Cyprinidae

S.No.	Species	Common name	Habitat	Category	Distribution
1	*Catla catla*	Bhakur	Freshwater	Least Concern (LC)	Asia: Pakistan, India, Bangladesh, Nepal and Myanmar.
2	*Cirrhinus cirrhosus*	Nain	Freshwater	Vulnerable (VU) (D2)	Asia: native to large rivers in the Indian subcontinent . Has been so widely transported in connection to aquaculture that its natural distribution can no longer be determined .
3	*Cirrhinus mrigala*	-	Freshwater	Least Concern (LC)	Asia: Pakistan, India, Nepal and Bangladesh. Introduced into Peninsular India.
4	*Ctenopharyngodon idella*	Grass carp, Hullu	Freshwater	Not Evaluated	Asia: China to eastern Siberia . Several countries report adverse ecological impact after introduction.
5	*Garra gotyla*	Sucker head	Freshwater	Least Concern (LC)	Asia: Pakistan, India, Bangladesh and upper Myanmar (Ref. 4832). Also reported from Afghanistan (Ref. 39701), Bhutan (Ref. 40882) and Nepal (Ref. 9496).
6	*Hypophthalmichthys molitrix*	Silver carp, Belli-gende	Freshwater	Near Threatened (NT)	Europe and Asia: native to most major Pacific dainages of East Asia from Amur to Xi Jiang, China .
7	*Hypselobarbus curmuca*	Curmuca barb	Freshwater	Endangered (EN)	Asia: India.
8	*Labeo bata*	Bata labeo	Freshwater	Least Concern (LC)	Asia: India and Bangladesh. Reported from Pakistan .
9	*Labeo calbasu*	Karnataka labeo, Dhai	Freshwater	Least Concern (LC)	Asia: Pakistan, India, Bangladesh, Myanmar, Nepal, Thailand and South Western China. Reported as introduced to China.
10	*Labeo dussumieri*	Thooli	Freshwater	Not Evaluated	Asia: India and Sri Lanka.
11	*Labeo dyocheilus*	Boalla	Freshwater	Least Concern (LC)	Asia: Pakistan, India, Bangladesh and Nepal. Known from Maeklong, Salween, Chao Phraya and Mekong basins.
12	*Labeo fimbriatus*	Selkendai	Freshwater	Not Evaluated	Asia: Pakistan, India, Nepal and Myanmar and Bangladesh .
13	*Labeo pangusia*	Pangusia labeo	Freshwater	Near Threatened (NT)	Asia: Pakistan, India, Nepal and Bangladesh. Recorded from Afghanistan, Bhutan and Myanmar.
14	*Labeo rohita*	Rohu	Freshwater	Least Concern (LC)	Asia: Pakistan, India, Bangladesh, Myanmar and Nepal.
15	*Osteobrama belangeri*	Manipur osteobrama	Freshwater	Near Threatened	Asia: India, Myanmar and China.
16	*Puntius chalakkudiensis*		Freshwater	Endangered (EN)	Asia: Chalakkudy river, Kerala, India
17	*Puntius denisonii*	Chorakanni	Freshwater	Vulnerable	Asia: endemic to India.
18	*Schizothorax richardsonii*	Asaila	Freshwater	Not Evaluated	Asia: Himalayan region of India, Sikkim and Bhutan, Nepal, Pakistan, and Afghanistan.
19	*Sinilabeo dero*	Not listed	Freshwater	Not estimated	Not available

Result for family '*Cyprinidae*'

Records overview

Microsatellites information of *Labeo calbasu*

Common name	Karnataka labeo, Dhai
Family	Cyprinidae
Habitat	Freshwater
Category	Least Concern (LC)
Distribution	Asia: Pakistan, India, Bangladesh, Myanmar, Nepal, Thailand and South Western China. Reported as introduced to China.

10 record(s) found

S.No.	Repeats	Rep. region	Rep. size	Sequence length	Authors	Accession No.
1	(TG)13	88...113	26	119 bp.	Mohindra *et al.* (2008)	EU616712
2	(TAGA)6	48...71	24	80 bp.	Mohindra *et al.* (2002)	AF517941
3	(TG)11	7...28	22	74 bp.	Singh *et al.* (2008)	EU616711
4	(AC)7	23...36	14	72 bp.	Singh *et al.* (2008)	EU616713
5	(TA)7 (AC)8(AT)8	20...33 45...76	14 32	112 bp.	Singh *et al.* (2008)	EU616714
6	(TG)8	51...66	16	85 bp.	Singh *et al.* (2009)	FJ545257
7	(CA)11	30...51	22	94 bp.	Singh *et al.* (2009)	FJ545258
8	(AC)9	16...33	18	84 bp.	Singh *et al.* (2009)	FJ545259
9	-	-	-	81 bp.	Singh *et al.* (2009)	FJ545260
10	(TG)8	31...46	16	80 bp.	Singh *et al.* (2009)	FJ545261

Result for species '*Labeo calbasu*'

Figure 3 Search results for different keywords in 'FishMicrosat'.

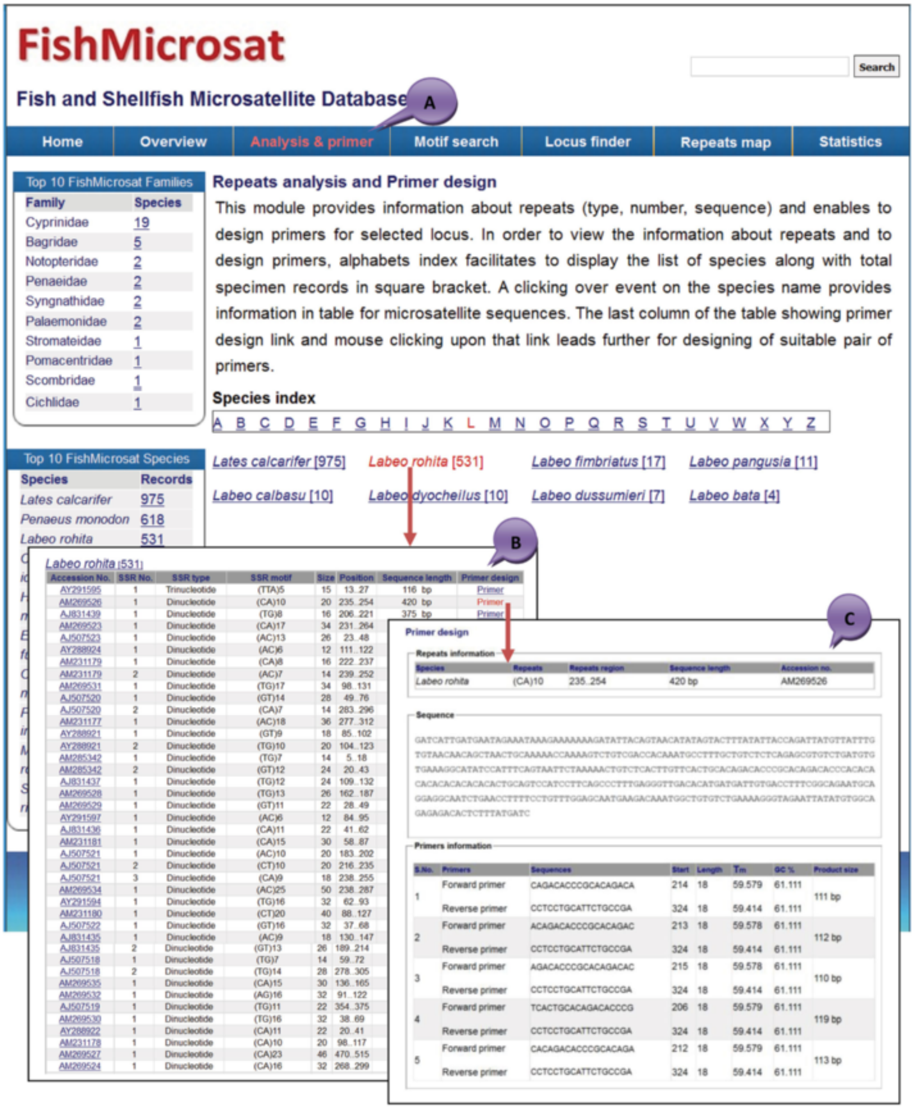
88% identitical, indicating homology between sequences (Figure 5C).

Repeats map

Repeat mapping and sequence similarity searching can be achieved through the menu item 'Repeats map' included in the web interface. The program accepts microsatellite sequences in FASTA format as input in the provided text area. The output presents information on repeats (size of query sequence, presence of compound/ composite repeats, number of identified repeats in query sequence SSRs, SSRs number, repeats location and size of repeats) along with summary on alignment of identical/similar sequences. The alignment summary presents targeted sequences accession no., species name, target length, gaps, matches and identity between query and targeted sequences (Figure 5D). The program initially checks the presence or absence of the repeats in the input sequence and assigns a boolean value. If the value is true the program processes the query sequences by using Blastn program and its compatible 'blastmsdb' database for similar sequence searches. Thus, it helps to find information about repeats orientation and sequence similarity for the newly generated microsatellite sequences.

Repeat statistics

In order to determine the frequency of different types of repeats from the specimen records available in FishMicrosat, the menu item 'Statistics' generates the frequency of each motif found in the database and displays the top three (most common) motifs with the largest frequencies. For example, the statistic view shows that repeats 'AC' was found 998 times, 'TG' 909 and 'CA' 881



Top 10 FishMicrosat Families

Family	Species
Cyprinidae	19
Bagridae	5
Notopteridae	2
Penaeidae	2
Syngnathidae	2
Palaemonidae	2
Stromateidae	1
Pomacentridae	1
Scombridae	1
Cichlidae	1

Repeats analysis and Primer design

This module provides information about repeats (type, number, sequence) and enables to design primers for selected locus. In order to view the information about repeats and to design primers, alphabets index facilitates to display the list of species along with total specimen records in square bracket. A clicking over event on the species name provides information in table for microsatellite sequences. The last column of the table showing primer design link and mouse clicking upon that link leads further for designing of suitable pair of primers.

Species index

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Lates calcarifer [975] *Labeo rohita* [531] *Labeo fimbriatus* [17] *Labeo pangusia* [11]
Labeo calbasu [10] *Labeo dyocheilus* [10] *Labeo dussumieri* [7] *Labeo bata* [4]

Top 10 FishMicrosat Species

Species	Records
Lates calcarifer	975
Penaeus monodon	618
Labeo rohita	531

Labeo rohita [531]

Accession No.	SSR No.	SSR type	SSR motif	Size	Position	Sequence length	Primer design
AY291595	1	Trinucleotide	(TTA)5	15	13..27	116 bp	Primer
AM269526	1	Dinucleotide	(CA)10	20	235..254	420 bp	Primer
AJ831439	1	Dinucleotide	(TG)8	16	206..221	375 bp	Primer
AM269523	1	Dinucleotide	(CA)17	34	231..264		
AJ507523	1	Dinucleotide	(AC)13	26	23..48		
AY288924	1	Dinucleotide	(AC)6	12	111..122		
AM231179	1	Dinucleotide	(CA)8	16	222..237		
AM231179	2	Dinucleotide	(AC)7	14	239..252		
AM269531	1	Dinucleotide	(TG)17	34	98..131		
AJ507520	1	Dinucleotide	(GT)14	28	49..76		
AJ507520	2	Dinucleotide	(CA)7	14	283..296		
AM231177	1	Dinucleotide	(AC)18	36	277..312		
AY288921	1	Dinucleotide	(GT)9	18	85..102		
AY288921	2	Dinucleotide	(TG)10	20	104..123		
AM285342	1	Dinucleotide	(TG)7	14	5..18		
AM285342	2	Dinucleotide	(GT)12	24	20..43		
AJ831437	1	Dinucleotide	(TG)12	24	109..132		
AM269528	1	Dinucleotide	(TG)13	26	162..187		
AM269529	1	Dinucleotide	(GT)11	22	28..49		
AY291597	1	Dinucleotide	(AC)6	12	84..95		
AJ831436	1	Dinucleotide	(CA)11	22	41..62		
AM231181	1	Dinucleotide	(CA)15	30	58..87		
AJ507521	1	Dinucleotide	(AC)10	20	183..202		
AJ507521	2	Dinucleotide	(CT)10	20	216..235		
AJ507521	3	Dinucleotide	(CA)9	18	238..255		
AM269534	1	Dinucleotide	(AC)25	50	238..287		
AY291594	1	Dinucleotide	(TG)16	32	62..93		
AM231180	1	Dinucleotide	(CT)20	40	88..127		
AJ507522	1	Dinucleotide	(GT)16	32	37..68		
AJ831435	1	Dinucleotide	(AC)9	18	130..147		
AJ831435	2	Dinucleotide	(GT)13	26	189..214		
AJ507518	1	Dinucleotide	(TG)7	14	59..72		
AJ507518	2	Dinucleotide	(TG)14	28	278..305		
AM269535	1	Dinucleotide	(CA)15	30	136..165		
AM269532	1	Dinucleotide	(AG)16	32	91..122		
AJ507519	1	Dinucleotide	(TG)11	22	354..375		
AM269530	1	Dinucleotide	(TG)16	32	38..69		
AY288922	1	Dinucleotide	(CA)11	22	20..41		
AM231178	1	Dinucleotide	(CA)10	20	98..117		
AM269527	1	Dinucleotide	(CA)23	46	470..515		
AM269524	1	Dinucleotide	(CA)16	32	268..299		

Primer design

Repeats information

Species	Repeats	Repeats region	Sequence length	Accession no.
Labeo rohita	(CA)10	235..254	420 bp	AM269526

Sequence

```
GATCATTGATGAATAGAAATAAAGAAAAAAAGATATTACAGTAACATATAGTACTTTATATTACCAGATTATGTTATTTG
TGTAACAACAGCTAACTGCAAAAACCAAAAGTCTGTCGACCACAAATGCCTTTGCTGTCTCTCAGAGCGTGTCTGATGTG
TGAAAGGCATATCCATTTCAGTAATTCTAAAAACTGTCTCACTTGTTCACTGCACAGACACCCGCACAGACACCCACACA
CACACACACACACACTGCAGTCCATCCTTCAGCCCTTTGAGGGTTGACACATGATGATTGTGACCTTTCGGCAGAATGCA
GGAGGCAATCTGAACCTTTTCCTGTTTGGAGCAATGAAGACAAATGGCTGTGTCTGAAAAGGGTAGAATTATATGTGGCA
GAGAGACACTCTTTATGATC
```

Primers information

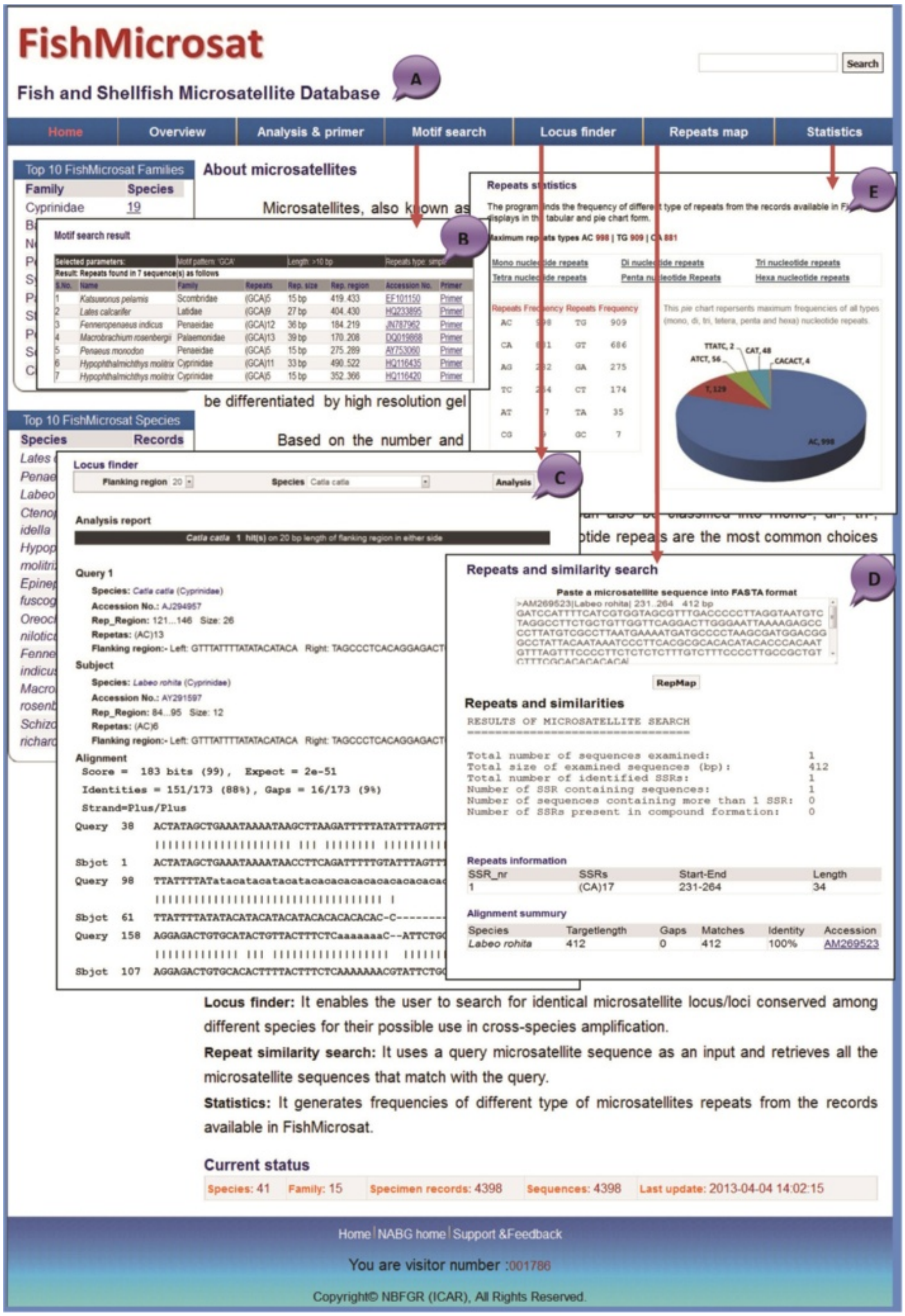
S.No.	Primers	Sequences	Start	Length	Tm	GC %	Product size
1	Forward primer	CAGACACCCGCACAGACA	214	18	59.579	61.111	111 bp
	Reverse primer	CCTCCTGCATTCTGCCGA	324	18	59.414	61.111	
2	Forward primer	ACAGACACCCGCACAGAC	213	18	59.578	61.111	112 bp
	Reverse primer	CCTCCTGCATTCTGCCGA	324	18	59.414	61.111	
3	Forward primer	AGACACCCGCACAGACAC	215	18	59.578	61.111	110 bp
	Reverse primer	CCTCCTGCATTCTGCCGA	324	18	59.414	61.111	
4	Forward primer	TCACTGCACAGACACCCG	206	18	59.579	61.111	119 bp
	Reverse primer	CCTCCTGCATTCTGCCGA	324	18	59.414	61.111	
5	Forward primer	CACAGACACCCGCACAGA	212	18	59.579	61.111	113 bp
	Reverse primer	CCTCCTGCATTCTGCCGA	324	18	59.414	61.111	

Figure 4 The web layout for SSR analysis and primer design of 'FishMicrosat'. (A) Species specific SSR analysis and primer design **(B)** repeat analysis output **(C)** SSR specific primer design output.

throughout all sequences. A repeat type index has also been included to display all the repeats and their frequencies in a table. The dinucleotide repeat type selected as default displays 12 combinations of dinucleotides. The maximum frequency of each type of nucleotide repeats (mono to hexa) can be viewed in the pie diagram (Figure 5E). The largest frequency for a mononucleotide repeat is 'T' with 129 occurrences, dinucleotide repeats 'AC' with 998 occurrences, trinucleotide repeats 'CAT' with 48 occurrences, tetranucleotide 'ATCT' with 56 occurrences, pentanucleotide 'TTATC' with 2 occurrences and hexanucleotide 'CACACT' with 4 occurrences. The database with 4398 sequences of 41 species has 277 mono, 4207 di, 610 tri, 554 tetra, 15 penta, 11 hexa and 279 compound repeats (Table 1). This section also analyzes information on the occurrence of the most frequent and rare nucleotide repeats in the fish genome. The dinucleotide repeats AC|TG (998|909) and CA|GT (881|686) were frequently found while CG|GC (9|7) were rare in fish genome.

Conclusions

FishMicrosat is a database of microsatellite sequences of commercially important fishes including shrimps and currently covers 4398 specimen records for 41 species. The database facilitates mining of SSR motifs, repeat orientations and sequence similarities. The statistics presents the relative abundance of microsatellite repeats that occur frequently in the genomes. Additionally, it



Top 10 FishMicrosat Families

Family	Species
Cyprinidae	19
B	
N	
P	
S	
P	
S	
P	
S	
C	

About microsatellites

Microsatellites, also known as

be differentiated by high resolution gel

Based on the number and

can also be classified into mono-, di-, tri-,

otide repeats are the most common choices

Top 10 FishMicrosat Species

Species	Records
Lates	
Penae	
Labeo	
Cteno *idella*	
Hypop *molitri*	
Epine *fuscog*	
Oreoc *nilotic*	
Fenne *indicu*	
Macro *rosenb*	
Schizo *richar*	

Motif search result

Selected parameters: Motif pattern: 'GCA' Length: >10 bp Repeats type: simple

Result: Repeats found in 7 sequence(s) as follows

S.No.	Name	Family	Repeats	Rep. size	Rep. region	Accession No.	Primer
1	*Katsuwonus pelamis*	Scombridae	(GCA)5	15 bp	419..433	EF101150	Primer
2	*Lates calcarifer*	Latidae	(GCA)9	27 bp	404..430	HQ233895	Primer
3	*Fenneropenaeus indicus*	Penaeidae	(GCA)12	36 bp	184..219	JN787962	Primer
4	*Macrobrachium rosenbergii*	Palaemonidae	(GCA)13	39 bp	170..208	DQ019868	Primer
5	*Penaeus monodon*	Penaeidae	(GCA)5	15 bp	275..289	AY753060	Primer
6	*Hypophthalmichthys molitrix*	Cyprinidae	(GCA)11	33 bp	490..522	HQ116435	Primer
7	*Hypophthalmichthys molitrix*	Cyprinidae	(GCA)5	15 bp	352..366	HQ116420	Primer

Repeats statistics

The program finds the frequency of different type of repeats from the records available in Fi displays in the tabular and pie chart form.

Maximum repeats types AC 998 | TG 909 | CA 881

Mono nucleotide repeats   Di nucleotide repeats   Tri nucleotide repeats
Tetra nucleotide repeats   Penta nucleotide Repeats   Hexa nucleotide repeats

Repeats	Frequency	Repeats	Frequency
AC	998	TG	909
CA	881	GT	686
AG	[illegible]	GA	275
TC	[illegible]	CT	174
AT	[illegible]	TA	35
CG	[illegible]	GC	7

This pie chart repersents maximum frequencies of all types (mono, di, tri, tetera, penta and hexa) nucleotide repeats.

Locus finder

Flanking region 20   Species Catla catla   Analysis

Analysis report

Catla catla 1 hit(s) on 20 bp length of flanking region in either side

Query 1

Species: *Catla catla* (Cyprinidae)
Accession No.: AJ294957
Rep_Region: 121...146 Size: 26
Repetas: (AC)13
Flanking region:- Left: GTTTATTTTATATACATACA Right: TAGCCCTCACAGGAGACT

Subject

Species: *Labeo rohita* (Cyprinidae)
Accession No.: AY291597
Rep_Region: 84...95 Size: 12
Repetas: (AC)6
Flanking region:- Left: GTTTATTTTATATACATACA Right: TAGCCCTCACAGGAGACT

Alignment

```
 Score =  183 bits (99),  Expect = 2e-51
 Identities = 151/173 (88%), Gaps = 16/173 (9%)
 Strand=Plus/Plus
Query  38   ACTATAGCTGAAATAAAATAAGCTTAAGATTTTTATATTTAGTT
            ||||||||||||||||||||| ||| |||||||| |||||||||
Sbjct  1    ACTATAGCTGAAATAAAATAACCTTCAGATTTTTGTATTTAGTT
Query  98   TTATTTTATatacatacatacatacacacacacacacacacac
            |||||||||||||||||||||||||||||||||||| |
Sbjct  61   TTATTTTATATACATACATACATACACACACACAC-C--------
Query  158  AGGAGACTGTGCATACTGTTACTTTCTCaaaaaaaC--ATTCTG
            ||||||||||||| ||| |||||||||||||||||||    ||||||
Sbjct  107  AGGAGACTGTGCACACTTTTACTTTCTCAAAAAAACGTATTCTG
```

Repeats and similarity search

Paste a microsatellite sequence into FASTA format

```
>AM269523|Labeo rohita| 231..264   412 bp
GATCCATTTTCATCGTGGTAGCGTTTGACCCCCTTAGGTAATGTC
TAGGCCTTCTGCTGTTGGTTCAGGACTTGGGAATTAAAAGAGCC
CCTTATGTCGCCTTAATGAAAATGATGCCCCTAAGCGATGGACGG
GCCTATTACAATAAATCCCTTCACGCGCACACATACACCCACAAT
GTTTAGTTTCCCCTTCTCTCTCTTTGTCTTTCCCCTTGCCGCTGT
CTTTCGCACACACACA
```

RepMap

Repeats and similarities

```
RESULTS OF MICROSATELLITE SEARCH
================================

Total number of sequences examined:              1
Total size of examined sequences (bp):           412
Total number of identified SSRs:                 1
Number of SSR containing sequences:              1
Number of sequences containing more than 1 SSR:  0
Number of SSRs present in compound formation:    0
```

Repeats information

SSR_nr	SSRs	Start-End	Length
1	(CA)17	231-264	34

Alignment summury

Species	Targetlength	Gaps	Matches	Identity	Accession
Labeo rohita	412	0	412	100%	AM269523

Locus finder: It enables the user to search for identical microsatellite locus/loci conserved among different species for their possible use in cross-species amplification.

Repeat similarity search: It uses a query microsatellite sequence as an input and retrieves all the microsatellite sequences that match with the query.

Statistics: It generates frequencies of different type of microsatellites repeats from the records available in FishMicrosat.

Current status

Species: 41   Family: 15   Specimen records: 4398   Sequences: 4398   Last update: 2013-04-04 14:02:15

Home | NABG home | Support &Feedback

You are visitor number :001786

Copyright© NBFGR (ICAR), All Rights Reserved.

Figure 5 The different web layouts of 'FishMicrosat' for data retrieval and analyses. (A) Home page **(B)** Motif search **(C)** Locus finder **(D)** Repeats map **(E)** Repeats statistics.

facilitates in identifying polymorphic loci across species and designing primers for repeat loci, thus providing researchers ready to use information from a centralized location, avoiding the cumbersome process of referring to multiple sources of literature and using multiple programmes. This repository with included tools can play a key role in cutting edge areas of research by assisting with marker selection, linkage mapping, population genetics, evolutionary studies, genetic relatedness among the species and genetic improvement programmes of important aquaculture species.

Availability and requirement

FishMicrosat is freely accessible at URL http://mail.nbfgr.res.in/fishmicrosat/ for research and academic use.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

NSN and MS conceived this study. IR, RP & AKP created the work-flow, database, application modules and performed data analyses. SPS and UKS incorporated general information and taxonomy of fishes. IR, RP, NSN, AKP and MS drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Authors are thankful to the National Agricultural Bioinformatics Grid (NABG), under National Agricultural Innovation Project (NAIP), ICAR, New Delhi for providing financial support and the Director, NBFGR, Lucknow for providing necessary facilities to carry out the work.

Author details

[1]Division of Molecular Biology and Biotechnology, National Bureau of Fish Genetic Resources, Lucknow 226002, India. [2]Division of Fish Taxonomy and Resources, National Bureau of Fish Genetic Resources, Lucknow 226002, India.

Received: 16 April 2013 Accepted: 11 September 2013
Published: 18 September 2013

References

1. Weber JL, Wong C: **Mutation of human short tandem repeats.** *Hum Mol Genet* 1993, **2**:1123–1128.
2. Tautz D, Renz M: **Simple sequences are ubiquitous repetitive components of eukaryotic genomes.** *Nucleic Acids Res* 1984, **12**:4127–4138.
3. Jarne P, Lagoda PJ: **Microsatellites, from molecules to populations and back.** *Trends Ecol Evol* 1996, **11**:424–429.
4. Bruford MW, Wayne RK: **Microsatellites and their application to population genetic studies.** *Curr Opin Genet Dev* 1993, **3**:939–943.
5. Beckmann JS, Weber JL: **Survey of human and rat microsatellites.** *Genomics* 1992, **12**:627–631.
6. Chen C, Zhou P, Choi YA, Huang S, Gmitter FG Jr: **Mining and characterizing microsatellites from citrus ESTs.** *Theor Appl Genet* 2006, **112**:1248–1257.
7. Kantety RV, La Rota M, Matthews DE, Sorrells ME: **Data mining for simple sequence repeats in expressed sequence tags from barley, maize, rice, sorghum and wheat.** *Plant Mol Biol* 2002, **48**:501–510.
8. Zane L, Bargelloni L, Patarnello T: **Strategies for microsatellite isolation: a review.** *Mol Ecol* 2002, **11**:1–16.
9. Sakai T, Miura I, Yamada-Ishibashi S, Wakita Y, Kohara Y, Yamazaki Y, Inoue T, Kominami R, Moriwaki K, Shiroishi T, Yonekawa H, Kikkawa Y: **Update of mouse microsatellite database of Japan (MMDBJ).** *Exp Anim* 2004, **53**:151–154.
10. Subramanian S, Madgula VM, George R, Kumar S, Pandit MW, Singh L: **SSRD: simple sequence repeats database of the human genome.** *Comp Funct Genomics* 2003, **4**:342–345.
11. Missirlis PI, Mead CL, Butland SL, Ouellette BF, Devon RS, Leavitt BR, Holt RA: **Satellog: a database for the identification and prioritization of satellite repeats in disease association Studies.** *BMC Bioinforma* 2005, **6**:145.
12. Aishwarya V, Grover A, Sharma PC: **EuMicroSatdb: a database for microsatellites in the sequenced genomes of eukaryotes.** *BMC Genomics* 2007, **8**:225.
13. Boby T, Patch AM, Aves SJ: **TRbase: a database relating tandem repeats to disease genes for the human genome.** *Bioinformatics* 2005, **21**:811–816.
14. Archak S, Meduri E, Kumar PS, Nagaraju J: **InSatDb: a microsatellite database of fully sequenced insect genomes.** *Nucleic Acids Res* 2007, **35**:D36–D39.
15. Gelfand Y, Rodriguez A, Benson G: **TRDB: the tandem repeats database.** *Nucleic Acids Res* 2007, **35**:D80–D87.
16. Knapik EW, Goodman A, Ekker M, Chevrette M, Delgado J, Neuhauss S, Shimoda N, Driever W, Fishman MC, Jacob HJ: **A microsatellite genetic linkage map for zebrafish (Danio rerio).** *Nat Genet* 1998, **18**:338–343.
17. Ardren WR, Borer S, Thrower F, Joyce JE, Kapuscinski AR: **Inheritance of 12 microsatellite loci in Oncorhynchus mykiss.** *J Hered* 1999, **90**:529–536.
18. Gilbey J, Verspoor E, McLay A, Houlihan D: **A microsatellite linkage map for Atlantic salmon (Salmo salar).** *Anim Genet* 2004, **35**:98–105.
19. Liu S, Rexroad CE III, Couch CR, Cordes JF, Reece KS, Sullivan CV: **A microsatellite linkage map of striped bass (Morone saxatilis) reveals conserved synteny with the three-spined stickleback (Gasterosteus aculeatus).** *Mar Biotechnol* 2012, **14**:237–244.
20. Liu F, Sun F, Li J, Xia JH, Lin G, Tu RJ, Yue GH: **A microsatellite-based linkage map of salt tolerant tilapia (Oreochromis mossambicus × Oreochromis spp.) and mapping of sex-determining loci.** *BMC Genomics* 2013, **14**:58–62.
21. Abdul Rahman Z, Choay-Hoong L, Mat Khairuddin R, Ab Razak S, Othman AS: **Isolation and characterization of novel microsatellite loci for Asian sea bass, Lates calcarifer from genome sequence survey database.** *J Genet* 2012, **91**:e82–e85.
22. Nikolic N, Fève K, Chevalet C, Høyheim B, Riquet J: **A set of 37 microsatellite DNA markers for genetic diversity and structure analysis of Atlantic salmon Salmo salar populations.** *J Fish Biol* 2009, **74**:458–466.
23. Sha Z, Wang S, Zhuang Z, Wang Q, Wang Q, Li P, Ding H, Wang N, Liu Z, Chen S: **Generation and analysis of 10 000 ESTs from the half-smooth tongue sole Cynoglossus semilaevis and identification of microsatellite and SNP markers.** *J Fish Biol* 2010, **76**:1190–1204.
24. Patel A, Das P, Swain SK, Meher PK, Jayasankar P, Sarangi N: **Development of 21 new microsatellite markers in Labeo rohita (rohu).** *Anim Genet* 2009, **40**:253–254.
25. Mendonça FF, Ussami LH, Hashimoto DT, Pereira LH, Porto-Foresti F, Oliveira C, Gadig OB, Foresti F: **Identification and characterization of polymorphic microsatellite loci in the blue shark Prionace glauca, and cross-amplification in other shark species.** *J Fish Biol* 2012, **80**:2643–2646.
26. Tong J, Wang Z, Yu X, Wu Q, Chu KH: **Cross-species amplification in silver carp and bighead carp with microsatellite primers of common carp.** *Mol Ecol Notes* 2002, **2**:245–247.
27. Liao M, Yang G, Wang X, Wang D, Zou G, Wei Q: **Development of microsatellite DNA markers of silver carp (Hypophthalmichthys molitrix) and their cross-species application in bighead carp (Aristichthys nobilis).** *Mol Ecol Notes* 2007, **7**:95–99.
28. Rouchka EC: **Database of exact tandem repeats in the Zebrafish genome.** *BMC Genomics* 2010, **11**:347.
29. Mao RX, Liu FJ, Zhang XF, Zhang Y, Cao DC, Lu CY, Liang LQ, Sun XW: **Studies on quantitative trait loci related to activity of lactate dehydrogenase in common carp (Cyprinus carpio).** *Yi Chuan* 2009, **31**:407–411.
30. Imsiridou A, Hardy H, Maudling N, Amoutzias G, Zaldívar Comenges JM: **Web database of molecular genetic data from fish stocks.** *J Hered* 2003, **94**:265–267.
31. Benson DA, Karsch-Mizrachi I, Clark K, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2012, **40**:D48–D53.
32. *FishBase database.* http://www.fishbase.org.
33. *MISA source program.* http://pgrc.ipk-gatersleben.de/misa/.
34. Estoup A, Angers B: **Microsatellites and minisatellites for molecular ecology: theoretical and empirical considerations.** In *Advances in Molecular Ecology.* 2nd edition. Edited by Carvalho GR. IOS press, Ohmsha; Nato Sience Series; 1998:55–79.
35. Rico C, Rico I, Hewitt G: **470 million years of conservation of microsatellite loci among fish species.** *Proc Biol Sci* 1996, **263**:549–557.
36. Primmer CR, Møller AP, Ellegren H: **A wide-range survey of cross-species microsatellite amplification in birds.** *Mol Ecol* 1996, **5**:365–378.
37. Kong J, Gao H: **Analysis of tandem repeats in the genome of Chinese shrimp Fenneropenaeus chinensis.** *Chin Sci Bull* 2005, **50**:1462–1469.
38. Katti MV, Ranjekar PK, Gupta VS: **Differential distribution of simple sequence repeats in eukaryotic genome sequences.** *Mol Biol Evol* 2001, **18**:1161–1167.
39. Altschul SF, Gish W, Miller W, *et al*: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403–410.
40. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** *Methods Mol Biol* 2000, **132**:365–386.
41. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG: **Primer3–new capabilities and interfaces.** *Nucleic Acids Res* 2012, **40**:e115.
42. *Primer3 standalone program.* http://sourceforge.net/projects/primer3/.

doi:10.1186/1471-2164-14-630
Cite this article as: Nagpure *et al.*: FishMicrosat: a microsatellite database of commercially important fishes and shellfishes of the Indian subcontinent. *BMC Genomics* 2013 **14**:630.