Published online 18 October 2007 *Nucleic Acids Research, 2008, Vol. 36, Database issue* **D53–D56**
doi:10.1093/nar/gkm811

UgMicroSat*db*: database for mining microsatellites from unigenes

Veenu Aishwarya and P. C. Sharma*

University School of Biotechnology, Guru Gobind Singh Indraprastha University, Kashmere Gate, Delhi 110 006, India

Received August 14, 2007; Revised September 17, 2007; Accepted September 18, 2007

ABSTRACT

Microsatellites, also known as simple sequence repeats (SSRs) or short tandem repeats (STRs), have extensively been exploited as molecular markers for diverse applications. Recently, their role in gene regulation and genome evolution has also been discussed widely. We have developed UgMicroSat*db* (Unigene MicroSatellite *d*ata*b*ase), a web-based relational database of microsatellites present in unigene sequences covering 80 genomes. UgMicroSat*db* allows microsatellite search using multiple parameters like microsatellite type (simple perfect, compound perfect and imperfect), repeat unit length (mono- to hexa-nucleotide), repeat number, microsatellite length and repeat sequence class. Microsatellites can also be retrieved by specifying EST, cDNA, CDS identity or by using Gene Index, GenBank, UniGene IDs. The database also provides information about trinucleotide repeats encoding various amino acids. Such codon repeats can be searched by specifying characteristics of coded amino acids like charge (basic, acidic or neutral), polarity (polar or non-polar), and their hydrophobic or hydrophilic nature. The nucleotide sequences of the target UniGenes are also provided to facilitate primer designing for PCR amplification of the desired microsatellite. UgMicroSat*db* is available at http://ipu.ac.in/usbt/UgMicroSatdb.htm.

INTRODUCTION

Microsatellites represent arrays of 1–6 bp tandem repeats in DNA. These sequences have proved very useful as molecular markers in diverse areas of genetic research including genome characterization and mapping (1). Recently, microsatellites have also been implicated to play some role in gene regulation and genome evolution (2–4). Association of trinucleotide repeats with many human diseases has been reported over the years (5). Some important examples include Huntington's disease and spinocerebellar ataxia (SCA) caused by expansion of CAG repeats (6,7), and oculopharyngeal muscular dystrophy caused by GCG expansion (8). Some other reports suggest association of trinucleotide repeats with various forms of cancer (9,10). $(A)_n$ repeats too have been assigned both cancer causing (11–13) and tumor-suppressive functions (14). Interestingly, trinucleotide repeats are now also being considered as potential therapeutic agents that act by triggering RNAi pathway as they are highly specific in silencing the mutant transcripts containing complementary repeats (15). In Drosophila, expansion of CAG repeats in homeobox gene *DLX6* leads to cell death (16).

The importance of microsatellites has been appreciated in plant systems also. Microsatellites derived from EST sequences (EST–SSRs) have found immense utility in various research projects in recent years (17). EST–SSR markers have been preferred over genomic-SSR markers for plant improvement programmes owing to their higher interspecific transferability rate. Moreover, EST–SSRs are proposed to be the better candidates for gene tagging. EST based microsatellite markers have been developed for apricot and grape (18), barley (19) and wheat (20), to name a few. However, unlike animal systems, significance of microsatellites in transcriptional activities has not been well documented in plant systems.

Considering vast utility of microsatellites in the fields of medicine and agriculture, many research groups have attempted to characterize their abundance, distribution and genomic localization using *in silico* methods (21–23). Such tools enable research scientists to exploit microsatellites for different applications with greater efficiency and specificity. EST–SSR databases have been developed and released in public domain by different groups. Examples of such databases include PlantSSR database (http://www.genome.clemson.edu/projects/ssr/) developed by Clemson University Genomics Institute (CUGI) and COS–EST–SSRs for cereals and legumes (http://intranet.icrisat.org/gt1/ssr/ESTSSRClustersubmit.asp). CMD (Cotton Microsatellite Database; 24) and

*To whom correspondence should be addressed. Tel: +91 11 23900220; Fax: +91 11 23865941; Email: deansbt@yahoo.co.in

SilkSatDb (25) provide microsatellite data from genomic sequences as well as EST sequences. Satellog (26) catalogs a number of disorders associated with mutations in microsatellite sequences near or within genes in humans. InSatDb (27) provides a compilation of microsatellites in five insect species. Such databases offer useful resources for various research activities aiming towards development of microsatellite markers and also for investigations focusing on deciphering the functional roles of microsatellites.

The databases and resources described above and elsewhere (28) remain specific and limited in their content and purpose. In particular, microsatellites within the transcriptionally active regions of the genome have not received the desired attention over a wide range of taxa. Such information is, however, necessary for undertaking various transcriptome based experimental studies. Therefore, there is a need to develop a platform for mining genic microsatellites to ensure, firstly their better utilization as molecular markers, secondly to understand fundamental questions concerning their abundance, distribution and evolution and thirdly to attribute putative function, if any, to these repeats. Considering this requirement, we have developed UgMicroSat*db* (Unigene MicroSatellite database), a relational database that provides information on microsatellites present within the unigenes across 80 eukaryotic genomes. A classified range of search options facilitate a user friendly and specific extraction interface. The database is so designed that unigene sequences harboring microsatellite(s) of interest can be extracted and further used, for example, in cross amplification experiments. We hope that information retrieved from the database may be helpful in opening new frontiers of basic and applied research on microsatellites.

CONSTRUCTION OF DATABASE

UgMicroSat*db* provides a catalogue of microsatellites occurring in unigene sequences of eukaryotic organisms belonging to a wide range of taxonomic groups. UniGene sequences were downloaded from The National Center for Biotechnology Information (NCBI) and scanned using a simple sequence repeat mining tool called MISA (19) that extracted microsatellite motifs and wrote them in tab delimited text files. This raw microsatellite information was processed using a set of C++ codes and Perl scripts. VMI_PRCS, a C++ code, processed statistics like size and position of microsatellites within the unigene sequences. VUG_PRCS, another C++ code, processed unigene IDs and sequences in the desired format. The data was reassembled using a Perl script VDATA_ASMBL and a data file was created. The data file was further formatted and then imported as a table in a MS-ACCESS database. Similar approach was adopted for all the individual sets of unigene sequences of different species. The overall scheme of database construction is explained in Figure 1. The quick retrieval of information from UgMicroSat*db* has been ensured by creating small, specific sub-databases for different groups of organisms. Furthermore, within each sub-database, individual organism has been represented by a separate table. A parent database that indexes all the sub-databases and the tables therein maintains fast, efficient and precise communication with these sub-databases. The graphical user interface was constructed using Active Server Pages (ASP). The overall architecture of the database has been outlined in Figure 2.

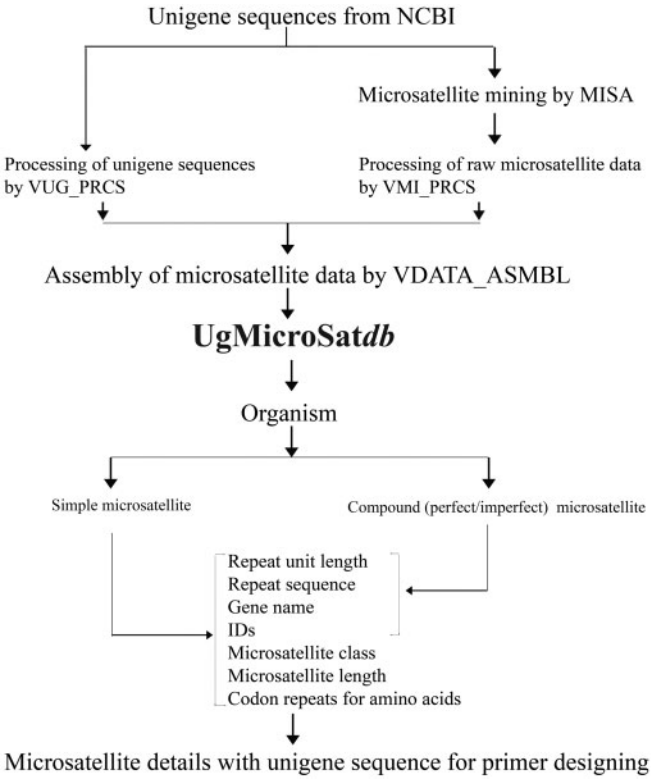

Figure 1. Data exchange flow diagram for UgMicroSat*db*.

ACCESSING DATABASE

UgMicroSat*db* can be accessed to extract simple (perfect) repeats and compound (perfect and imperfect) repeats. Microsatellites can be mined using various search options viz. repeat unit length (mononucleotide to hexanucleotide), repeat sequence (motif search), microsatellite length, host cDNA, CDS, EST name, GenBank ID, gene index number or UniGene ID and microsatellite class search (29). For trinucleotide repeats, the database also gives data for codon repeats i.e. repeats that code for amino acids. The option 'amino acid codon search' allows search for repeats that code for all the 20 amino acids. Further, search can also be made for codon repeats on the basis of characteristics of amino acids like charge (basic, acidic and neutral), polarity (polar and non-polar) and hydrophobicity or hydrophilicity. The database also allows batch download, such that a user can download all the microsatellites mined in response to a particular query in a text file. The database is linked with NCBI for the retrieval of detailed information on unigene

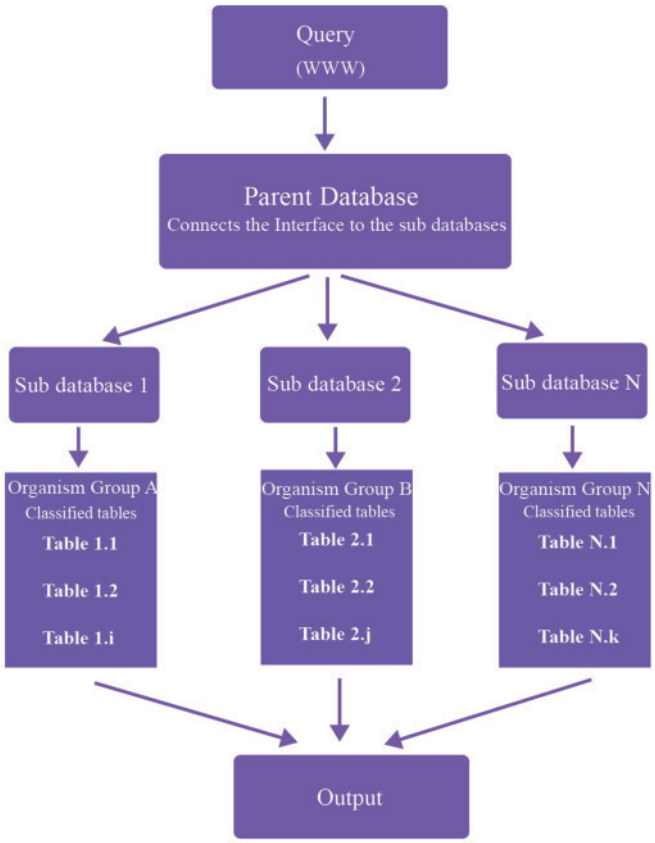

Figure 2. Architecture of UgMicroSat*db*.

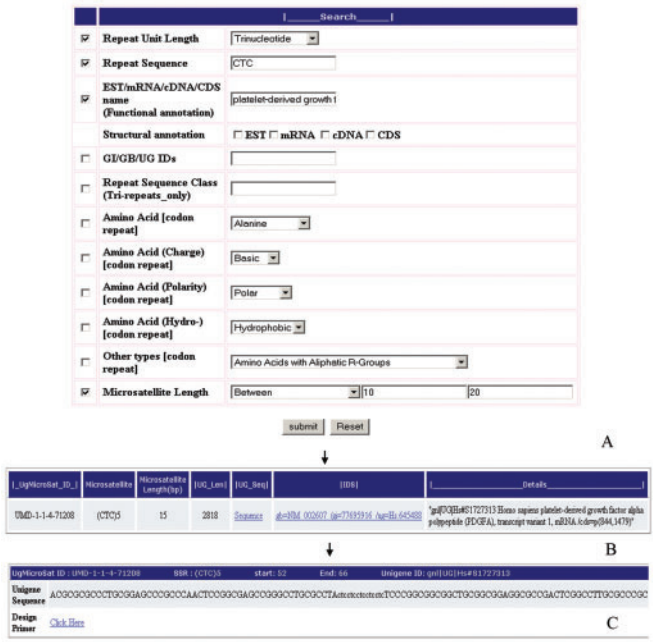

Figure 3. User interface for UgMicroSat*db*, showing (A) various input options, (B) output, and (C) Unigene sequence (only the start sequence, is shown) with microsatellite sequence (CTC)5 and position (52–66).

sequences based on their genbank IDs. Finally, the database allows the user to design primers for PCR amplification of the specific microsatellite locus by providing the selected unigene sequences harboring the particular microsatellite, and is also linked to Primer3, a primer designing tool (30). A quick help mode is provided with examples on how to fill in the search options for easy reference and navigation for the user. The search options are extensively explained with the help of some case studies on the database website. The database can serve as an immense source of information in understanding the microsatellite dynamics in the transcriptionally active regions. For example, information pertaining to a trinucleotide repeat CTC, whose length is between 10 and 20 base pair, present in platelet-derived growth factor alpha polypeptide (PDGFA) of *Homo sapiens* can easily be searched as shown in Figure 3. The output also gives the GenBank IDs and gene index number along with the unigene IDs. Further, the details of the unigene and the localization of the microsatellite (start and end positions) are provided and the microsatellites are marked in lower capitals.

UTILITY OF THE DATABASE

UgMicroSat*db* is likely to be accessed by biologists engaged in research with diverse objectives in both plant and animal systems primarily to develop molecular markers and also to understand the functional significance of microsatellites in regulating gene expression and genome evolution. UgMicroSat*db* offers an important platform for a detailed comparative analysis of microsatellite repeats in genic regions for a wide range of species. The comprehensive options to search for simple and compound microsatellites and to identify the codon repeats in the genic regions allow users to explore new avenues of investigations on these repeats. The availability of unigene sequences for different aspects like designing primers for PCR amplification of desired motifs will facilitate studies on mutability, microsatellite abundance and variability across genomes, etc. Association of these microsatellites with a particular disease or phenotype may be explored by identifying their expansion and contraction possibilities in a given population. Microsatellite data can also be used to investigate various anomalies and disorders using candidate gene approach. Further, such information can be used to design synthetic oligonucleotides representing complementary repeats to be used in RNA interference based silencing to target mutant genes. Such approaches hold much therapeutic value. The database can largely be used to develop EST–SSR markers for various research programmes, particularly on genome mapping and gene tagging. Apart from hosting an extensive form of microsatellite data within the genes, UgMicroSat*db* is unique in a way as compared to the previously developed databases as it hosts microsatellite data covering a large number of organisms including both lower and higher forms of plants and animals. Relative mining of imperfect repeats of such a diverse range of organisms may provide tools to study the dynamics of microsatellites and also their association with similar

or different types of repeats. To conclude, UgMicroSat*db* will serve as an important starting point whereby extracted information serves as an important input in designing experiments in new directions elucidating novel roles and functions of microsatellites in the unexplored transcriptomes.

FUTURE PERSPECTIVES

At present, UgMicroSat*db* hosts data on microsatellites occurring in unigene sequences of 80 genomes. UgMicroSat*db* team aims to update the database commensurating with updation of the NCBI unigene database. The flexible design of the database makes it feasible to increase the size of database to virtually any size without compromising with its fast data retrieval rate.

AVAILABILITY

UgMicroSat*db* can be freely accessed from http://ipu.ac.in/usbt/UgMicroSatdb.htm

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was waived by the Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Schlotterer,C. (2004) The evolution of molecular markers–just a matter of fashion. *Nat. Rev. Genet.*, **5**, 63–69.
2. Kashi,Y. and King,D.G. (2006) Simple sequence repeats as advantageous mutators in evolution. *Trends Genet.*, **22**, 253–259.
3. Li,Y-C., Korol,A.B., Fahima,T., Beiles,A. and Nevo,E. (2002) Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol. Ecol.*, **11**, 2453–2465.
4. Li,Y-C., Korol,A.B., Fahima,T. and Nevo,E. (2004) Microsatellites within genes: Structure, function and evolution. *Mol. Biol. Evol.*, **21**, 991–1007.
5. Cummings,C.J. and Zoghbi,H.Y. (2000) Fourteen and counting: unraveling trinucleotide repeat diseases. *Hum. Mol. Genet.*, **9**, 909–916.
6. Zoghbi,H.Y. and Orr,H.T. (2000) Glutamine repeats and neurodegeneration. *Annu. Rev. Neurosci.*, **23**, 217–237.
7. Nakamura,K., Jeong,S.Y., Uchihara,T., Anno,M., Nagashima,K., Nagashima,T., Ikeda,S., Tsuji,S. and Kanazawa,I. (2001) SCA17, a novel autosomal dominant cerebellar ataxia caused by an expanded polyglutamine in TATA-binding protein. *Hum. Mol. Genet.*, **10**, 1441–1448.
8. Brais,B., Bouchard,J.P., Xie,Y.G., Rochefort,D.L., Chrétien,N., Tomé,F.M., Lafrenière,R.G., Rommens,J.M., Uyama,E. *et al.* (1998) Short GCG expansions in the PABP2 gene cause oculopharyngeal muscular dystrophy. *Nat. Genet.*, **18**, 164–167.
9. Pizzi,C., Di Maio,M., Daniele,S., Mastranzo,P., Spagnoletti,I., Limite,G., Pettinato,G., Monticelli,A., Cocozza,S. *et al.* (2007) Triplet repeat instability correlates with dinucleotide instability in primary breast cancer. *Oncol. Rep.*, **17**, 193–199.
10. De Abreu,F.B., Pirolo,L.J., Canevari Rde,A., Rosa,F.E., Moraes Neto,F.A., Caldeira,J.R., Rainho,C.A. and Rogatto,S.R. (2007) Shorter CAG repeat in the AR gene is associated with atypical hyperplasia and breast carcinoma. *Anticancer Res.*, **27**, 1199–205.
11. Duval,A. and Hamelin,R. (2002) Mutations at coding repeat sequences in mismatch repair-deficient human cancers: toward a new concept of target genes for instability. *Cancer Res.*, **62**, 2447–2454.
12. Vassileva,V., Millar,A., Briollais,L., Chapman,W. and Bapat,B. (2002) Genes involved in DNA repair are mutational targets in endometrial cancers with microsatellite instability. *Cancer Res.*, **62**, 4095–4099.
13. Yamada,T., Koyama,T., Ohwada,S., Tago,K., Sakamoto,I., Yoshimura,S., Hamada,K., Takeyoshi,I. and Morishita,Y. (2002) Frameshift mutations in the MBD4/MED1 gene in primary gastric cancer with high-frequency microsatellite instability. *Cancer Lett.*, **181**, 115–120.
14. Markowitz,S., Wang,J., Myeroff,L., Parsons,R., Sun,L., Lutterbaugh,J., Fan,R.S., Zborowska,E., Kinzler,K.W. *et al.* (1995) Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science*, **268**, 1336–1338.
15. Krol,J., Fiszer,A., Mykowska,A., Sobczak,K., de Mezer,M. and Krzyzosiak,W.J. (2007) Ribonuclease dicer cleaves triplet repeat hairpins into shorter repeats that silence specific targets. *Mol. Cell*, **25**, 575–586.
16. Ferro,P., dell'Eva,R. and Pfeffer,U. (2001) Are there CAG repeat expansion-related disorders outside the central nervous system? *Brain Res. Bull.*, **56**, 259–264.
17. Varshney,R.K., Graner,A. and Sorrells,M.E. (2005) Genic microsatellite markers in plants: features and applications. *Trends Biotechnol.*, **23**, 48–55.
18. Decroocq,V., Favé,M.G., Hagen,L., Bordenave,L. and Decroocq,S. (2003) Development and transferability of apricot and grape EST microsatellite markers across taxa. *Theor. Appl. Genet.*, **106**, 912–922.
19. Thiel,T., Michalek,W., Varshney,R.K. and Graner,A. (2003) Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.*, **106**, 411–422.
20. Eujayl,I., Sorrells,M.E., Baum,M., Wolters,P. and Powell,W. (2002) Isolation of EST-derived microsatellite markers for genotyping the A and B genomes of wheat. *Theor. Appl. Genet.*, **104**, 399–407.
21. La Rota,M., Kantety,R.V., Yu,J-K. and Sorrells,M.E. (2005) Nonrandom distribution and frequencies of genomic and EST-derived microsatellite markers in rice, wheat and barley. *BMC Genomics*, **6**, 23.
22. Garnica,D.P., Pinzoni,A.M., Quesada-Ocampo,L.M., Bernal,A.J., Barreto,E., Grunwald,N.J. and Restrepo,S. (2006) Survey and analysis of microsatellites from transcript sequences in *Phytophthora* species: frequency, distribution and potential as markers for the genus. *BMC Genomics*, **7**, 245.
23. Grover,A. and Sharma,P.C. (2007) Microsatellite motifs with moderate GC content are clustered around genes on *Arabidopsis thaliana* chromosome 2. *In Silico Biol.*, **7**, 0021.
24. Lenda,A., Scheffler,J., Scheffler,B., Palmer,M., Lacape,J-M., Yu,J-Z., Jesudurai,C., Jung,S., Muthukumar,S. *et al.* (2006) CMD: A cotton microsatellite database resource for *Gossypium* genomics. *BMC Genomics*, **7**, 132.
25. Prasad,M.D., Muthulakshmi,M., Arunkumar,K.P., Madhu,M., Sreenu,V.B., Pavithra,V., Bose,B., Nagarajaram,H.A., Mita,K. *et al.* (2005) SilkSatDb: a microsatellite database of the silkworm, *Bombyx mori*. *Nucleic Acids Res.*, **33**, D403–D406.
26. Missirlis,P.I., Mead,C.R., Butland,S.L., Ouellette,B.F., Devon,R.S., Leavitt,B.R. and Holt,R.A. (2005) Satellog: A database for the identification and prioritization of satellite repeats in disease association studies. *BMC Bioinformatics*, **10**, 145.
27. Archak,S., Meduri,E., Kumar,P.S. and Nagaraju,J. (2007) InSatDb: a microsatellite database of fully sequenced insect genomes. *Nucleic Acids Res.*, **35**, D36–D39.
28. Aishwarya,V., Grover,A. and Sharma,P.C. (2007) EuMicroSat*db*: A database for microsatellites in the sequenced genomes of eukaryotes. *BMC Genomics*, **8**, 225.
29. Jurka,J. and Pethiyagoda,C. (1995) Simple repetitive DNA sequences from primates: compilation and analysis. *J. Mol. Evol.*, **40**, 120–126.
30. Rozen,S. and Skaletsky,H.J. (2000) Primer3 on the WWW for general users and for biologist programmers. In Krawetz,S. and Misener,S. (eds), *Bioinformatics Methods and Protocols: Methods in Molecular Biology*, Humana Press, Totowa, NJ, pp. 365–386.