Deposited research article

MRD: a microsatellite repeats database for prokaryotic and eukaryotic genomes

Subbaya Subramanian, Vamsi M Madgula#, Ranjan George#, Rakesh K Mishra, Madhusudhan W Pandit, Chandrashekar S Kumar# and Lalji Singh

Addresses: Centre for Cellular and Molecular Biology, Uppal Road, Hyderabad 500 007, India. #Ingenovis, ilabs ltd., 97, Road No.3, Banjara Hills, Hyderabad, 500 034, India

Correspondence: Lalji Singh. E-mail: lalji@ccmb.res.in

Posted: 13 November 2002

Genome ***Biology*** 2002, **3(12)**:preprint0011.1-0011.13

Received: 8 November 2002

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/12/preprint/0011

This is the first version of this article to be made available publicly and no other version is available at present.

MRD: A Microsatellite Repeats Database for prokaryotic and eukaryotic Genomes

Subbaya Subramanian, Vamsi M Madgula[#], Ranjan George[#], Rakesh K Mishra, Madhusudhan W Pandit, Chandrashekar S Kumar[#] and Lalji Singh

Centre for Cellular and Molecular Biology, Uppal Road, Hyderabad 500 007, INDIA
and [#]Ingenovis, ilabs ltd., 97, Road No.3, Banjara Hills, Hyderabad, 500 034, INDIA

Running Title

Database for microsatellite repeats

Correspondence to

Dr Lalji Singh
Centre for Cellular and Molecular Biology
Uppal Road, Hyderabad 500 007
INDIA

Tel: +91-40-7160789

Fax: +91-40-7160252

Email: lalji@ccmb.res.in

ABSTRACT

MRD is a database system to access the microsatellite repeats information of genomes such as archea, eubacteria, and other eukaryotic genomes whose sequence information is available in public domains. MRD stores information about simple tandemly repeated k-mer sequences where k= 1 to 6, i.e. monomer to hexamer. The web interface allows the users to search for the repeat of their interest and to know about the association of the repeat with genes and genomic regions in the specific organism. The data contains the abundance and distribution of microsatellites in the coding and non-coding regions of the genome. The exact location of repeats with respect to genomic regions of interest (such as UTR, exon, intron or intergenic regions) whichever is applicable to organism is highlighted. MRD is available on the World Wide Web at http://www.ccmb.res.in/mrd and/or http://www.ingenovis.com/mrd. The database is designed as an open-ended system to accommodate the microsatellite repeats information of other genomes whose complete sequences will be available in future through public domain.

INTRODUCTION

Microsatellites are tandemly repeated sequence motifs of 1 to 6 base pairs [1] found in abundance in the genomes of prokaryotes [2] and eukaryotes [3]. These repeats are found in both coding and non-coding regions of the genome. The presence of microsatellites in the coding region and in the regulatory region of the genome can directly influence the gene expression. Studies have indicated that microsatellites are predominantly present in the non-coding part of the genome and play a significant role in the genome evolution and possibly in gene regulation [4]. While, there is no direct correlation of the microsatellite content with the genome size, it is generally believed that microsatellite content of a genome depends on the genome size [5].

Microsatellites show a high degree of length polymorphisms and are extremely useful in human genetic studies. Many markers have been developed from the known sequences containing these repeats available from databases as well as derived from screening genomic libraries. In spite of the recognition of simple repeats as markers, mechanisms underlying the microsatellite allelic diversity are still poorly understood. Strand slippage during replication has been suggested to be the most likely mechanism in generation of mutation and polymorphism in the microsatellites [6]. Questions such as why certain repeats motifs are common than others and why there exists a variation of such repeats among taxa are important from evolutionary point of view. Though there are extensive studies on the microsatellite repeats in the human and other genomes, a complete inventory of the microsatellite repeats in the human genome and other genomes, as a single resource is still not available. However with the completion of many prokaryotic and eukaryotic genomes this analysis has become possible. Realising the importance of

microsatellites in the genome, we under took to analyse in detail the repeat distribution and genes associated with them.

IMPLEMENTATION

We have created a microsatellite repeats database of all repeat combinations from mono- to hexanucleotide repeats. For the analysis we have used the sequence information that is available on the Genbank genomes FTP site. The build number and the release date for the genomes are given in the database. All 501 theoretically possible non-overlapping repeat types were searched [7]. We have analysed the distribution of perfect repeats of >=12 base pairs. The rationale for choosing the small cutoff value was that the microsatellites are often disrupted by single base substitutions. We have also included an extensive analysis of distribution of these repeats and their association with coding and non-coding regions of the genome, such as exon, intron and intergenic regions wherever applicable. MRD provides a comprehensive resource for studying various aspects of microsatellite repeats in prokaryotic and eukaryotic genomes which may help in understanding their probable function and evolutionary significance. The FTP sites of the database from where various genome sequences are obtained are given in the MRD database. The entire data is stored in an Oracle Database.

DATA STRUCTURE

The database is presented as various views that tabulate the details about a particular repeat type and the repeat. In the current form we have analysed 77 prokaryotic genomes and 7 eukaryotic genomes such as human, mouse, *Drosophila,* Yeast *(S cerevesiae and S pombe), C elegans,* and *Arabidopsis thaliana*. For each organism the microsatellite repeats were analysed and the density and distribution has been tabulated. The repeat associations with specific genomic

regions are tabulated for each repeat. Both the strands in the sequence were searched for the microsatellite repeats. The microsatellites were searched for a min cut-off of 12bp in both prokaryotes and eukaryotes.

The database is organized so as to provide summary as well as detailed views of the repeat regions and their associations with genomic regions, genes. The complete data is organized into six tables. "Size" of a chromosome refers to the cumulative size of those regions for which the sequence is known and analyzed. All densities mentioned in the tables are with respect to this size. All "numbers" mentioned are a sum of the occurrences of the particular pattern and its reverse complement.

The first page of the database provides a brief description of the database and a link to the page that enables the user to select the genome and repeat class of interest. In its current shape, the database deals only with microsatellites. The database has been designed in an open-ended fashion so that it is possible to add other types of repeats in the future. Once the user has chosen the genome of interest, he/she may view information organized within five views.

View 1: Abundance and distribution of monomer to hexamer repeat types

This table presents the abundance and density of each of the six repeat types, i.e. monomer to hexamer, across each chromosome of the organism selected. The total for all repeats types across a particular chromosome and for all chromosomes for a particular repeat type are also presented. Density of each repeat type in terms of repeats per mb (mega base pairs) of chromosome is given.

View 2: Abundance and distribution of all 501 repeats across the genome selected

This table gives the same information as above but for all repeats in a particular repeat type. One is thus able to view details of occurrence and density for each of the 501 repeats across all chromosomes or the whole genome in case of prokaryotes.

View 3: Abundance and distribution of monomer to hexamer repeat types across genomic regions for a selected genome

It is essential and interesting to know the distribution of microsatellites across genomic regions, i.e. exon, intron and intergenic regions. The sizes of exon, intron and intergenic regions (in terms of base pairs) for each chromosome have been calculated from data given in the annotation lists of Genbank entries. For each repeat found, the genomic region it belongs to is captured. Thus for a given chromosome, the density of repeat types on exon, intron and intergenic regions is presented. However, only those repeats that start and end in the same region have been considered for density calculations. Repeats which span regions, say, start in an exon and end in an intron, have not been considered; the occurrence of such repeats is, in any case, rare. In the case of prokaryotic genomes the repeat abundance and density is analysed based on the coding and non-coding regions of the genome.

View 4: Detailed view of each repeat, association with proximal genes and STS markers

This table gives complete details of each repeat found in a given genome. For a given repeat, the repeat number and the length of the total repeated sequence are given. The start position of the repeat, both with respect to the contig sequence and the original Genbank entry (denoted by its accession id) where the repeat is found, is given. If the repeat is found on a particular gene, the

name of the gene and the exact regions of the gene where the repeat is found (exon number or intron number) are displayed. (In the case of prokaryotic genomes, currently the database refers to the coding sequences as exons). Wherever applicable if a repeat is found to lie in the UTR region of an mRNA, the region is mentioned as UTR1 or UTR2, as the case may be. UTR1 refers to the distance between the first exon on the mRNA and its corresponding coding sequence. UTR2 refers to the distance between the stop codon of the mRNA and the last base pair of the transcript. However if more than one mRNA is present for a given gene, the mRNA with the largest exon regions (sum of all exon lengths) and/or starting the earliest is considered. In the event that a repeat lies in an intergenic region, the nearest downstream gene and the distance between the repeat and the gene (in terms of base pairs) is given. A distinction is also made between intergenic regions, which are upstream and downstream of a gene. The terms "upstream" and "downstream" are used with reference to the sequence as given in the Genbank entries.

In the case of the human genome, association of microsatellites with Sequence Tagged Sites (STS) is also said to be important and revealing. Thus, if a repeat lies on a STS marker, we have therefore included the "Standard Name" of the STS, as mentioned in the annotation. Otherwise, the nearest STS marker and the distance between the STS marker and the repeat are given. The start and stop positions of the repeat vis-à-vis the STS, i.e. whether they are upstream or downstream of the STS, are also given. This however, is not applicable in the case of other genomes.

Detailed view of repeats contained within genes

A separate option is provided in View 4 which facilitates the user to view details of only those repeats that are contained within genes or which span a gene and the nearing intergenic region.

View 5: Details of each repeat for a specific gene

This table allows the user to specify the gene of interest and view details of all repeats associated with the gene, i.e. those that are either contained within the gene or are proximal to it. The names of the genes have been taken from the contig files of Genbank. The database therefore currently accepts gene names written in Genbank nomenclature only.

AVAILABILITY

MRD is available on the World Wide Web at http://www.ccmb.res.in/mrd and http://www.ingenovis.com/mrd. A user-friendly interactive interface is in place that provides researchers the facility to view details of microsatellites of their interest. The database also includes detailed instructions on how to access and utilize the resource. Technical concerns and queries may be directed to subree@ccmb.res.in or vamsi.madhav@ingenovis.com

CONCLUSION

The high lights of MRD database are as follows

1. The database includes more than eighty different genome information for the microsatellite repeats and abundance, density and distribution.

2. Comprehensive tables which details about the repeat association with specific genomic region. In case of human and mouse the STS marker association is also included which will help in researches to identify more STRs.

3. Using MRD it is possible to compare different repeats and investigate the evolution of associated genomic regions. For example, looking at the triplet repeat containing genes in different organisms can reveal its potential for repeat expansion associated disease.

4. The database is expected provide an excellent platform for the analysis of microsatellite evolution and their possible role in genome organization.

5. One of the possible roles of such repeat is suggested to be in gene regulation. Study of the association of these repeats in the flanking sequences of the gene may be the first step to understand role of microsatellites in gene regulation. We have provided an option of searching repeats in the flanking sequence of specific gene.

6. Using this database it is possible to list out most abundant or rare repeats in different organisms.

We hope that this database will be a comprehensive resource for studying various aspects of microsatellite repeats in the human and other genomes and that it will be helpful in identifying their probable function and evolutionary significance. Information on the abundance of microsatellites, coupled with the distribution patterns in the coding as well as non-coding regions of the genome and their associations with genes and STS markers will shed more light on the function of microsatellites in gene regulation. As and when genome sequences of new species become available through the public domain the list of genomes in the MRD database will be expanded.

Acknowledgements

The authors are thankful to Sreedhar, Siva Prasad, Saritha and Kavitha for their support in developing the MRD database. We are grateful to Thangaraj and Ramesh Aggarwal for providing helpful discussions. Financial support from CSIR and DBT is duly acknowledged.

REFERENCES

1. Vogt P: **Potential genetic functions of tandem repeated DNA sequence blocks in the human genome are based on a highly conserved "chromatin folding code.** *Hum. Genet* 1990, **84**:301- 336

2. Gur-Arie R, Cohen C, Eitan Y, Shelef L, Hallerman EM, Kashi Y: **Simple sequence repeats in *Escherichia coli*: abundance, distribution, composition, and polymorphism.** *Genome Res* 2000, **10**: 62 –71

3. Toth G, Gaspari Z, Jurka J: **Microsatellites in different eukaryotic genomes: survey and analysis**. *Genome Res* 2000, **10**:967- 981

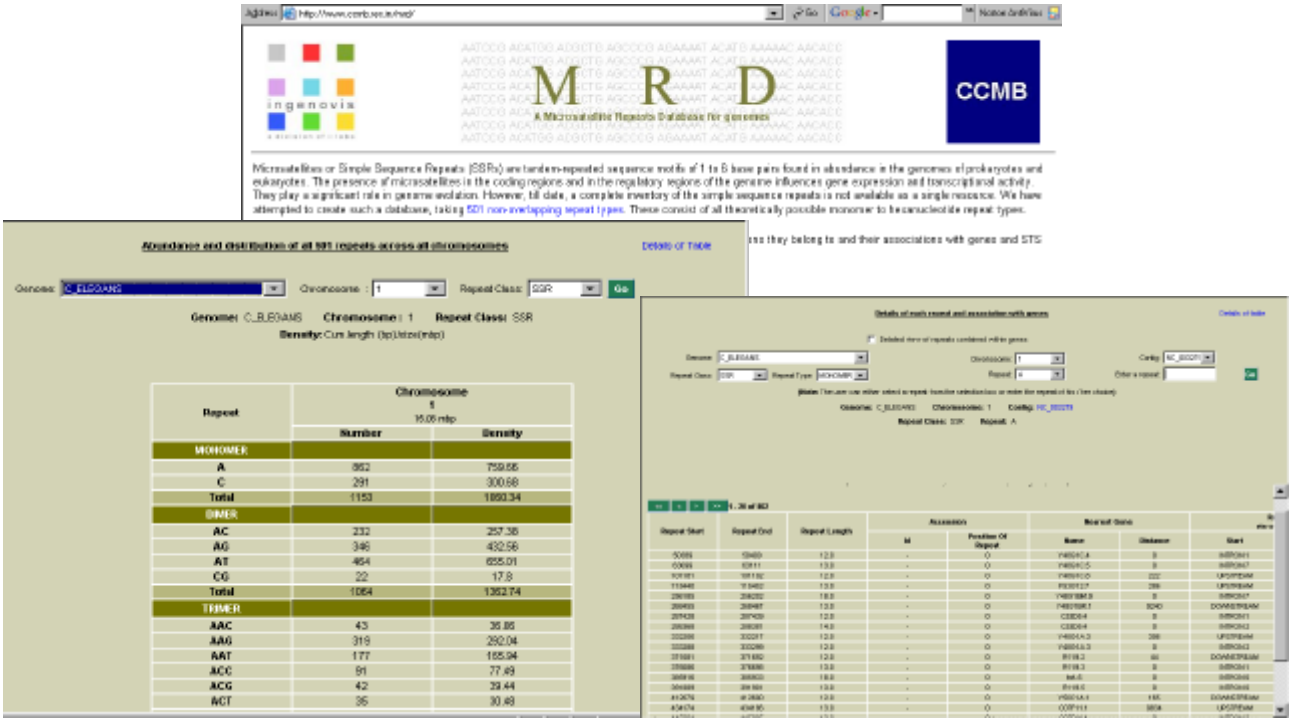
4. Kashi Y, King D, Soller M: () **Simple sequence repeats as a source of quantitative genetic variation.** *Trends Genet* 1997, **13**:74-78

5. Primmer CR, Raudsepp T, Chowdhary BP, Moller AP, Ellegren H: **Low frequency of microsatellites in the avian genome**. *Genome Res* 1997, **7**:471-482

6. Pearson CE, Sinden RR: **Trinucleotide repeat DNA structures: dynamic mutations from dynanic DNA** *Curr Opin Struct Biol* 1998, **8**:321 –330

7. Jurka, J., and Pethiyagoda, C. **Simple repetitive DNA sequences from primates: compilation and analysis.** *J Mol Evol* 1995, **40**:120 –126

Figure 1



Chromosome	Exon			Intron			Intergenic			Total		
	Number	Size	Density	Number	Size	Density	Number	Size	Density	Number	Size	Density
1	24	3.96	87.37	611	5.04	1648.92	516	6.06	1200.16	1151	15.06	1057.68
2	15	4.24	55.0	475	4.28	1494.01	462	6.66	960.85	952	15.17	858.31
3	16	3.56	61.87	627	4.6	1853.42	508	5.7	1247.93	1151	13.86	1144.7
4	9	4.0	29.49	435	4.73	1239.98	597	8.75	953.34	1041	17.48	819.4
5	18	5.77	40.24	479	5.36	1222.03	592	9.79	857.01	1089	20.92	725.5
10	21	3.55	85.85	286	4.16	952.89	748	10.03	1075.09	1055	17.75	848.4
Total	103	25.07	57.99	2913	28.18	1406.04	3423	46.99	1027.89	6439	100.24	891.59