D36–D39 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 1 November 2006*
doi:10.1093/nar/gkl778

InSatDb: a microsatellite database of fully sequenced insect genomes

Sunil Archak, Eshwar Meduri, P. Sravana Kumar and J. Nagaraju*

Laboratory of Molecular Genetics, Centre for DNA Fingerprinting and Diagnostics, ECIL Road, Nacharam, Hyderabad 500 076, India

Received August 11, 2006; Revised September 27, 2006; Accepted September 29, 2006

ABSTRACT

InSatDb presents an interactive interface to query information regarding microsatellite characteristics *per se* of five fully sequenced insect genomes (fruit-fly, honeybee, malarial mosquito, red-flour beetle and silkworm). InSatDb allows users to obtain microsatellites annotated with size (in base pairs and repeat units); genomic location (exon, intron, up-stream or transposon); nature (perfect or imperfect); and sequence composition (repeat motif and GC%). One can access microsatellite cluster (compound repeats) information and a list of microsatellites with conserved flanking sequences (microsatellite family or paralogs). InSatDb is complete with the insects information, web links to find details, methodology and a tutorial. A separate 'Analysis' section illustrates the comparative genomic analysis that can be carried out using the output. InSatDb is available at www.cdfd.org.in/insatdb.

INTRODUCTION

Microsatellites are simple sequence repeats (SSRs) that exhibit complex patterns in their frequency of occurrence, genomic distribution, mutability, function and evolution. Apart from being the source of informative genetic markers, microsatellites *per se* have attracted a lot of attention with respect to their origin, distribution, expansion, mutation and disintegration (1–7). Questions are also asked about the functional role of microsatellites in particular and biological significance of the microsatellites in general (4,8–12). Genetic studies and whole genome sequence analysis have established non-random distribution, variability and high mutability as characteristics of microsatellites. Evidences are accruing, which support the role of microsatellites in gene regulation, transcription and protein function (13). Existence of qualitative and quantitative differences between microsatellites of different genomes and their role in adaptive evolution have also been theorized (2,8). However, such studies require information on type (mono to hexa), motif (GC%), abundance (motif preferences), frequency, distribution (linkage group-wise and chromosomal position), location (exon, intron, regulatory element and transposon), nature (perfect, imperfect and compound) and copy number (existence of paralogs) of microsatellites not only on a whole genome basis but also as a comparative analysis of multiple genomes that are related by phylogeny (for instance, fully sequenced primate genomes or fungal genomes or insect genomes) to draw functional conclusions.

Insects have long exhibited the greatest genetic diversity on earth that has puzzled mankind. Biologists have relied on insects to unravel many fundamental tenets of biology. Whole genome sequences of insects have lived up to the reputation of diversity and have thrown immense variability in size and organization of their genomes. Among others, there are five fully sequenced insect genomes: *Drosophila melanogaster* (as a model organism it provides maximum annotated data), *Anopheles gambiae* (another Dipteran but economically highly important as malarial vector), *Tribolium castaneum* (relatively early insect order of Coleoptera), *Apis mellifera* (Hymenoptera, relatively a recent insect order) and *Bombyx mori* (economically important as silk-producing member of Lepidoptera, members of which are crop pests; also significant as a model for insect development). Researchers attempting to understand the biology and evolution of microsatellites are often faced with the following questions: (i) Do microsatellites occur everywhere in the genome? (ii) Does the length of microsatellites have any relationship with their frequency? (iii) Does the flanking sequence composition influence origin of microsatellites? (iv) Does the microsatellite size affect microsatellite disintegration rate? (v) Does the GC content of the motif affect the length, repeat units or mutation rate of microsatellites? (vi) Do genomes possess hotspots and islands of microsatellites?

*To whom correspondence should be addressed. Tel: +91 40 27171427; Fax: +91 40 27155610; Email: jnagaraju@cdfd.org.in

The authors wish to be known that, in their opinion, the second and third authors should be regarded as joint Second Authors

(vii) Is there any favoured association of microsatellites in the compound repeats? (viii) Do microsatellites occur as families of common flanking sequences in the genomes (paralogs)?

InSatDb, unlike many other microsatellite databases that cater to only the needs of microsatellites as markers, allows users to address the above-mentioned questions by accessing qualitative and quantitative genome level microsatellites profile of a single insect or to carry out comparative genomic analysis using all the five genomes.

METHODS

Drosophila melanogaster, *A.mellifera*, *A.gambiae* and *T.castaneum* sequences were downloaded from GenBank (ftp://ftp.ncbi.nlm.nih.gov/genomes) and *Bombyx mori* sequences were downloaded from http://silkworm.genomics.org.cn. Repeats were extracted employing *Tandem Repeat Finder* version 4 (14). To ensure that the extracted repeat sequences were real microsatellites, those with less than five repeat units and shorter than 15 bp in length were excluded. *Tandem Repeat Finder* does not employ minimal alignment score for detecting microsatellites; rather a probabilistic model of random repeat sequences specified by per cent identity and frequency of insertions and deletions. This includes calculation of average per cent identity between the copies (pM) and average percentage of insertions and deletions (pI). The algorithm has a pair of matching probability and indel probability values (pM = 0.80, pI = 0.10) as default to cover most divergent copies at every locus. We used two sets of alignment parameters (match, mismatch, gap), (+2, −3, −5) and (+2, −5, −7) to score the matches. All the microsatellites with a minimum alignment score of 30 are reported in the database, which means that both perfect and imperfect microsatellites are listed. The genome sequences were also analysed using RepeatMasker (A.F.A. Smit, R. Hubley and P. Green, unpublished data; http://www.repeatmasker.org) to obtain indices marking the occurrence of simple repeats, tandem repeats, segmental duplications, interspersed repeats including SINEs, DNA transposons, retrotransposons, LINEs, etc. Further, sequences were analysed for the delineation of exons and introns using GENSCAN (15). Flanking sequences of microsatellites were aligned to catalogue paralogous microsatellites that exhibit identical origin and hence considered belonging to the same *microsatellite family*. Occurrence of two or more microsatellites contiguously with intervening non-repeat sequence of $\leqslant$70 bp were separately categorized as *compound repeats*.

DATABASE ORGANIZATION

InSatDb is developed as a multi-tier relational database (Figure 1). It stores microsatellites from all the five insect genomes separately as well as carries complete annotations of these microsatellites. The database also provides basic information on each of the five insects and important links to obtain further knowledge, and contains a tutorial page and a glossary page. Microsatellite data can be accessed in two formats. End users with adequate computational capabilities can batch download full complement of microsatellites (insect-wise), microsatellite sequences, compound microsatellites and full list of microsatellite loci existing as families. These data are made available as csv files, which are compatible with spreadsheet programmes such as MS Excel. Alternatively, details of the microsatellites with highly

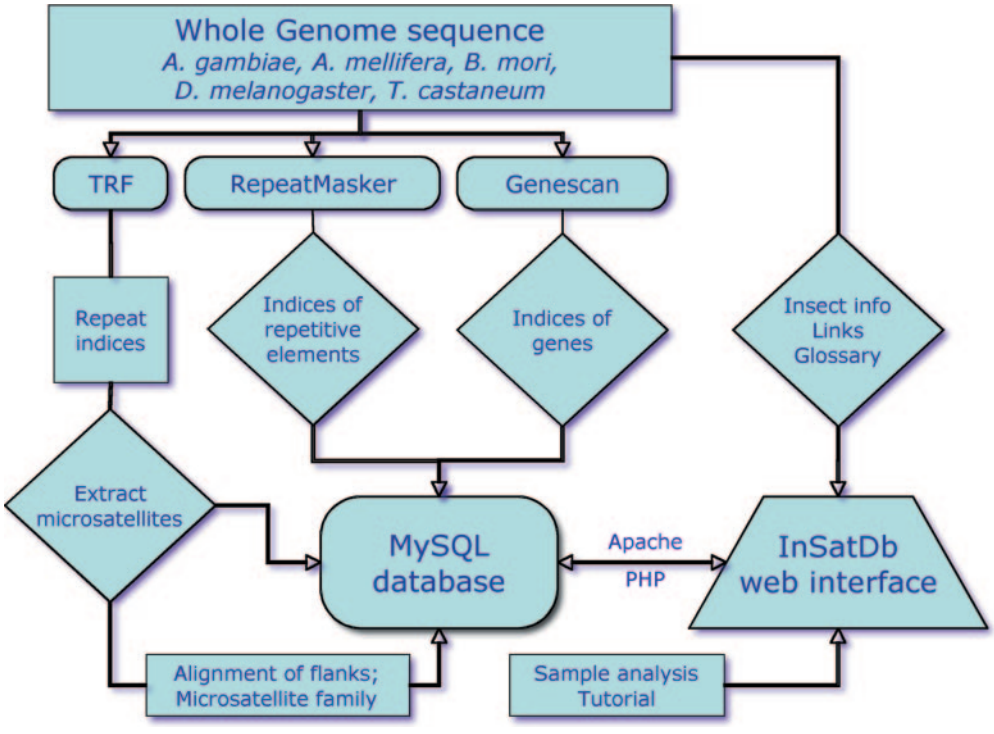

Figure 1. InSatDb organization and implementation.

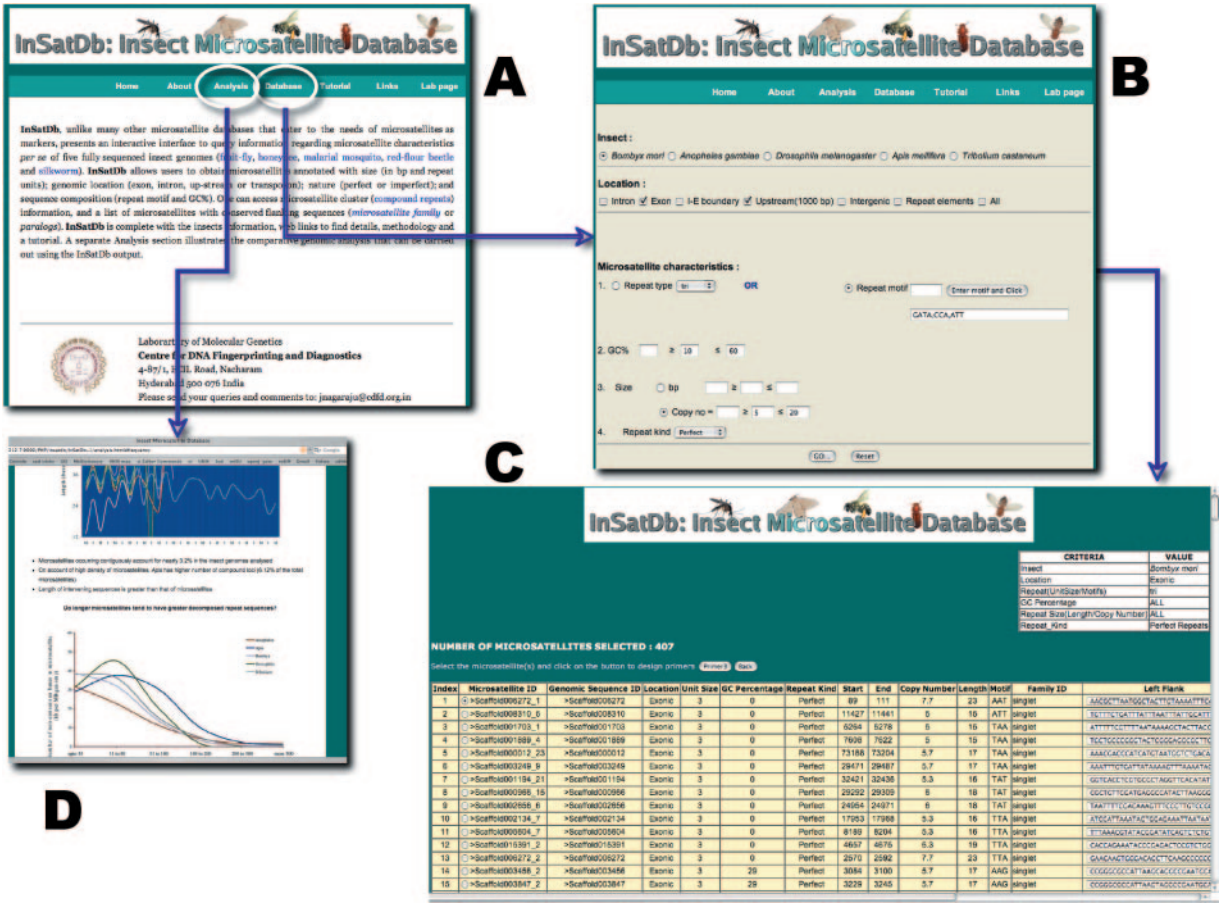

Figure 2. Screen shots of (**A**) InSatDb homepage, (**B**) multi-option query sheet, (**C**) output table and (**D**) analysis page.

specific characteristics may be queried using a multi-option query sheet (Figure 2). The options include insect (one at a time); location (intron, exon, i.e. boundary, upstream, intergenic, repeat elements—single or in combination); repeat type (motif size, mono- to hexa-nucleotide) or actual repeat motif (by essentially entering up to five repeat motifs); GC% (fixed value or range); repeat size in either base pairs or number of units (fixed value or range); repeat kind (perfect or imperfect). Once insect and location options are selected rest of the fields are set at 'ALL' by default. The output is primarily a list of microsatellites annotated for all options of the query sheet and the output table is generated as a hierarchical pre-sorted list. Each microsatellite is given a unique ID that also carries genomic sequence ID and corresponding indices. If the number of microsatellites selected based on the options of the query sheet exceeds 500, the output is split into sets of 500 microsatellites. In addition, a csv file containing total output is also made available for downloading. If the query options do not select any microsatellite, a message indicating zero output is displayed and a back button is provided to refine the options. The table is a 'one-stop' output and gives complete information on microsatellites. SSR motif and 100 bp each of the left and right flanking sequences are given for each microsatellite entry, which allows users to carry out sequence analysis of microsatellite vis-à-vis locus. In addition, users can select individual microsatellites to convert them into locus-specific markers. This is facilitated by automatic uploading of repeat and flanking sequences of the selected microsatellite into Primer3 query form (16).

DATA ANALYSIS

Insect genomes vary greatly in SSR composition, diversity and distribution. Our analysis showed that microsatellite content of five fully sequenced insect genomes is independent of both genome size and GC content (Table 1). The database consists of a dedicated section (*Analysis*) that describes the types of analysis that can be carried out using the data obtained from InSatDb. Some of the quick observations and inferences from a comparative genomic analysis are given in this section.

Preponderance of di- and tri-nucleotide repeats is observed in *Drosophila* and *Anopheles*, whereas tri- and tetra-nucleotide repeats are abundant in *Bombyx* and *Tribolium*. On the whole, shorter microsatellites are abundant in the five insect genomes; as the length of the microsatellite increases their number decreases logarithmically typified by *Bombyx* and *Drosophila* microsatellites (>90% of the microsatellites <50 bp); on the other hand, *Anopheles* and *Tribolium* have longer microsatellites in a relatively high frequency. Shorter microsatellites not only predominate microsatellite population in the five insect genomes, but also seem to possess higher number of imperfect repeat

Table 1. Microsatellite content of insect genomes

Insect	Chr (*n*)	Genome size (Mb)	GC%	Number of repeats	Microsatellite content (% Genome)	Number of microsatellites per Mb genome
Bombyx mori	28	397.71	37.33	111 006	0.72	280
Drosophila melanogaster	4	118.36	42.45	63 637	1.56	538
Anopheles gambiae	3	287.79	40.51	150 936	1.58	525
Apis mellifera	16	228.45	32.28	236 480	3.41	1035
Tribolium castaneum	10	198.06	25.53	24 246	0.41	122

units. On the other hand, microsatellites spanning >100 bp consisted of perfect, rather than imperfect repeats. Imperfect repeat units originate because of substitutions and indels. Interruptions, if at all, occur mainly in the middle region of the repeat sequence and the ends seem to be selected against decomposition. On the whole, most of the microsatellites occur within 20% GC bracket. There is no linear correlation between GC content and the average number of repeat units. Average length of the microsatellite across GC range is 37 ± 9 bp and between 0 and 5% GC content, microsatellites tend to be longer than 60 bp.

Compound microsatellites account for nearly 3.2% in the insect genomes analysed; owing to high density of microsatellites, *Apis* has higher number of compound loci (6.12%). *Anopheles* and *Apis* genomes have as many as 50 and 60% of the total microsatellites in coding region, respectively. *Bombyx* genome has only 10% of the microsatellites in regions spanning exons, introns and their boundary. More than 70% of the microsatellites present in exons are trinucleotide repeats except in *Apis*, where 50% tri- and 25% dinucleotide repeats are present in exonic regions. Microsatellites in insects are AT rich (on an average 23.4% GC); however, they exist within regions that are not always AT rich.

DATABASE ACCESS AND FUTURE PERSPECTIVES

InSatDb is freely available through www.cdfd.org.in/insatdb. Incorporation of microsatellite data from additional insects, query facility for better comparative genomic analysis such as gene-based microsatellite extraction and conservation analysis are planned. Additionally, based on users' feedback, supplementary features will be added to make InSatDb a single window system for insect genome analyses using microsatellite tools.

ACKNOWLEDGEMENTS

J.N. acknowledges the financial assistance from the Department of Biotechnology, Government of India. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Dieringer,D. and Schlotterer,C. (2003) Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. *Genome Res*., **13**, 2242–2251.
2. Karaoglu,H., Lee,C.M. and Meyer,W. (2005) Survey of simple sequence repeats in completed fungal genomes. *Mol. Biol. Evol*., **22**, 639–649.
3. Kruglyak,S., Durrett,R.T., Schug,M.D. and Aquadro,C.F. (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl Acad. Sci. USA*, **95**, 10774–10778.
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. Metzgar,D., Liu,L., Hansen,C., Dybvig,K. and Wills,C. (2002) Domain-level differences in microsatellite distribution and content result from different relative rates of insertion and deletion mutations. *Genome Res*., **12**, 408–413.
6. Nadir,E., Margalit,H., Gallily,T. and Ben-Sasson,S.A. (1996) Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. *Proc. Natl Acad. Sci. USA*, **93**, 6470–6475.
7. Wilder,J. and Hollocher,H. (2001) Mobile elements and the genesis of microsatellites in dipterans. *Mol. Biol. Evol*., **18**, 384–392.
8. Kashi,Y. and King,D.G. (2006) Simple sequence repeats as advantageous mutators in evolution. *Trends Genet*., **22**, 253–259.
9. Boeva,V., Regnier,M., Papatsenko,D. and Makeev,V. (2006) Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics*, **22**, 676–684.
10. Katti,M.V., Ranjekar,P.K. and Gupta,V.S. (2001) Differential distribution of simple sequence repeats in eukaryotic genome sequences. *Mol. Biol. Evol*., **18**, 1161–1167.
11. Morgante,M., Hanafey,M. and Powell,W. (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nature Genet*., **30**, 194–200.
12. Toth,G., Gaspari,Z. and Jurka,J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res*., **10**, 967–981.
13. Borstnik,B. and Pumpernik,D. (2002) Tandem repeats in protein coding regions of primate genes. *Genome Res*., **12**, 909–915.
14. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res*., **27**, 573–580.
15. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol*., **268**, 78–94.
16. Rozen,S. and Skaletsky,H. (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol*., **132**, 365–386.