[Fly 1:5, e1-e3, EPUB Ahead of Print: http://www.landesbioscience.com/journals/fly/article/5250; September/October 2007]; ©2007 Landes Bioscience

Extra View

InSatDb

A Genomic Tool for Insect Geneticists

Sunil Archak

Javaregowda Nagaraju*

Centre of Excellence for Genetics and Genomics of Silkmoths; Laboratory of Molecular Genetics; Centre for DNA Fingerprinting and Diagnostics; Hyderabad, India

*Correspondence to: Javaregowda Nagaraju; Centre of Excellence for Genetics and Genomics of Silkmoths; Laboratory of Molecular Genetics; Centre for DNA Fingerprinting and Diagnostics; ECIL Road, Nacharam; Hyderabad 500 076 India; Tel.: 40.27171427; Fax: 40.27155610; Email: jnagaraju@cdfd.org.in

Original manuscript submitted: 09/28/07
Manuscript accepted: 11/04/07

This manuscript has been published online, prior to printing for Fly, Volume 1, Issue 5. Definitive page numbers have not been assigned. The current citation is: Fly 2007; 1(5):
http://www.landesbioscience.com/journals/fly/article/5250
Once the issue is complete and page numbers have been assigned, the citation will change accordingly.

KEYWORDS

microsatellite, interactive database, insect, GC content, repeat length, tri-nucleotide repeats, exon, intron

ACKNOWLEDGEMENTS

Javaregowda Nagaraju acknowledges the Department of Biotechnology, Government of India for the financial support in the form of Centre of Excellence grant.

Addendum to:

InSatDb: A Microsatellite Database of Fully Sequenced Insect Genomes

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

Nucleic Acids Research 2007; 35: D36-9

ABSTRACT

Microsatellites show tremendous variation between genomes in terms of their occurrence and composition. Availability of whole genome sequences allows us to study microsatellite characteristics of fully sequenced insect genomes to understand the evolution and biological significance of microsatellites. InSatDb is an insect microsatellite database that provides an interactive interface to query information on 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%). Here we present a snapshot of the distribution and composition of microsatellites in introns and exons of insect genomes. The data present interesting observations regarding the microsatellite life-cycle and genome flux.

Microsatellites have never ceased to attract the attention of geneticists with respect to their origin, distribution, expansion, mutation and disintegration.[1-7] Evidence points to the biological significance of microsatellites in general and their regulatory role in particular.[4,8-13] Studies about the role of microsatellites in adaptive evolution require information on the existence of qualitative and quantitative differences between microsatellites of different genomes.[2,10] The availability of a whole genome sequence can refine our perception of microsatellite characteristics (occurrence, length, repeat units, mutation rate) and the relationships between length and frequency, sequence composition and origin, length and disintegration rate, etc. However, many geneticists may not be interested because such studies require an infrastructure to locally store massive quantities of data and, more importantly, the computational expertise to process various types of genomic data. An effective alternative is to develop an interactive database.

In a recent paper, we presented a database that facilitates easy and interactive analysis of microsatellites of insect genomes.[14] The database, InSatDb, unlike many other microsatellite databases that cater to the needs of microsatellites as markers, allows users to address questions regarding occurrence, composition, organization, origin etc. of microsatellites in five insect genomes (fruit-fly, honeybee, malarial mosquito, red-flour beetle and silkworm). This is achieved by accessing the qualitative and quantitative genome-level microsatellite profile of a single insect species or by carrying out comparative genomic analysis using all the five genomes. In this article, we have attempted to present an account of the database in simple terms and to use an example to demonstrate the potential of the database to produce information of biological significance.

InSatDb was developed as a multi-tier relational database with open access (www.cdfd.org.in/insatdb). 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 information (compound repeats), and a list of microsatellites with conserved flanking sequences (microsatellite family or paralogs). Microsatellite data can be accessed in two formats. End users with adequate computational capabilities can batch download the full complement of microsatellites as CSV files. Alternatively, as a main feature of the database details of the microsatellites with highly specific characteristics may be retrieved using a multi-option query sheet. The options include insect (one at a time), location (intron, exon, intron-exon boundary, upstream, intergenic, repeat elements-single or in combination), repeat type (motif size, mono- to hexa-nucleotide) or actual repeat motif (by essentially typing the sequence of repeat motifs), GC% (fixed value or range), repeat size in either base-pairs or number of units (fixed value or range) and repeat kind

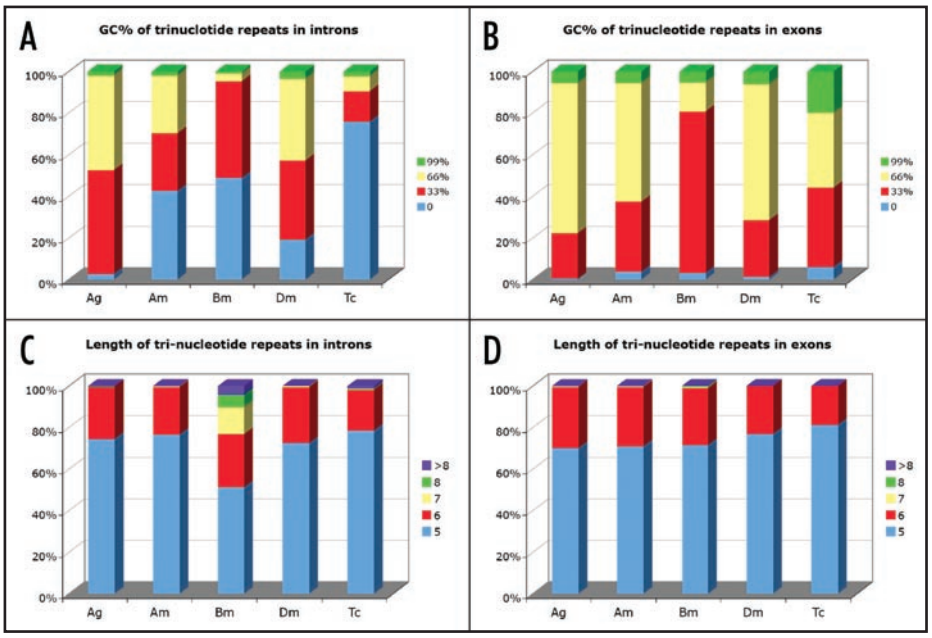

Figure 1. Distribution of microsatellites in the coding regions of the five insect species (Top panels: sequence composition; Bottom panels: number of repeat units).

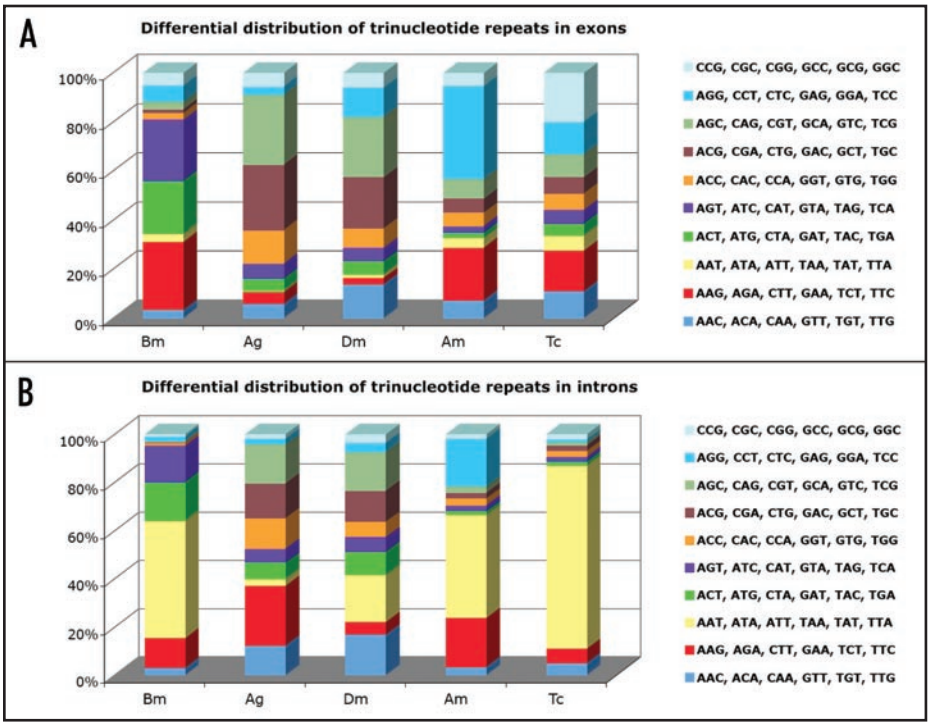

Figure 2. Distribution of microsatellite repeat types in the coding regions of the five insect species.

(perfect or imperfect). Once insect and location options are selected, the 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. The table is a "one-stop" output and gives complete information on microsatellites. In addition, users can select individual microsatellites to convert into locus specific markers. This is facilitated by automatic uploading of repeat and flanking sequences of the selected microsatellite into the query form of Primer3 software.

Using a query-based retrieval facility we downloaded perfect tri-nucleotide microsatellites occurring in the introns and exons from all five insects. The objective was to find qualitative and quantitative differences between insect genomes and to draw possible conclusions upon the origin and maintenance of microsatellites in the coding regions.

The distribution pattern of intronic microsatellites compared to exonic ones is dramatically different with respect to GC-content (Fig. 1A). *Bombyx* and *Tribolium* introns possess repeats in excess of 90% that are AT-rich; *Tribolium* introns particularly contrast completely with their exonic counterparts with three-fourth of the repeats being 100% AT rich. In the dipteran insects on the other hand, *Drosophila* (~43%) and *Anopheles* (~47%) introns possess repeats that are GC-rich.

A quick look at the sequence composition of the exonic tri-nucleotide repeats indicates that both dipteran genomes have 70% of the tri-nucleotide repeats with 66% GC-content (Fig. 1B). *Tribolium* exons show that as much as one-fifth of the microsatellites have 100% GC-content. Contrastingly, the majority of the *Bombyx* exons consisted of AT-rich repeats (~80%).

Among such divergent distribution of sequence compositions of microsatellites, lengths of tri-nucleotide repeats, as an exception, do not show variation at all between insects as well as between locations (intron or exon, Fig. 1C and D). On an average, three out of every four microsatellites occurring in coding regions exhibit the minimal length (five repeat units). There are three significant observations to be made: (i) coding regions exert selection pressure to restrict the microsatellite expansion, (ii) such selection operates with equal stringency in exons and introns, and (iii) the length of the microsatellite is independent of the sequence composition (compare exonic repeats of *Anopheles* with others). The only exception is the tri-nucleotide repeats in the introns of *Bombyx* genome where only half of the intronic repeats show minimum size of five units. Repeats of six, seven and eight units occur 25.6%, 13% and 6% of the total repeat loci. Microsatellite loci having greater than eight repeat units are also seen 4.14% of the time. This kind of microsatellite distribution may be attributed to the holocentric nature of chromosomes, abundance of transposable elements and large gene size owing to longer introns.

Although GC composition of the repeats shows a differential distribution among insects, it does not reveal specific propensities of insects or locations for particular repeats. Exact repeat types occurring in exons and introns indeed present a kaleidoscopic distribution (Fig. 2A

and B). The graphical interpretation of the proclivities for repeat types threw open several interesting observations as given below:

i) *Bombyx* coding regions typically possess ACT/ATG (green) and AGT/ATC (violet) repeats that are not particularly observed in other insects.
ii) *Tribolium* and *Bombyx* introns are especially rich in AAT/ATT (pale yellow) repeats that are almost absent in *Anopheles*.
iii) Dipteran coding regions (both introns and exons) possess AGC/GTC (off-grey) and ACG/CTG (brown) repeats that are mostly absent in other genomes.
iv) Remarkably, coding regions of two dipterans not only show a nearly balanced distribution of different repeat types (all colors in the bar), but such a distribution remains the same in both introns and exons.

A simple extraction of microsatellites using a query-based approach provided us with many clues about the fact that microsatellites are not randomly distributed in the expressed regions of insect genomes and that the selection process operates upon the repeat sequences. InSatDb, therefore, is a handy resource for obtaining ready-made genomic annotations using microsatellites as a benchmark. A completely different set of analyses is also possible where an individual microsatellite is tracked in known gene/s across insects for answering evolutionary questions.

InSatDb is gearing up for an up gradation based on additional data (16 insect genomes) and tools (comparative analysis). We wish to obtain as much feedback as possible from the users to incorporate suggested improvements before an upgraded version is released for public use.

References

1. Dieringer D, Schlotterer C. Two distinct modes of microsatellite mutation processes: Evidence from the complete genomic sequences of nine species. Genome Res 2003; 13:2242-51.
2. Karaoglu H, Lee CM, Meyer W. Survey of simple sequence repeats in completed fungal genomes. Mol Biol Evol 2005; 22:639-49.
3. Kruglyak S, Durrett RT, Schug MD, Aquadro CF. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci USA 1998; 95:10774-8.
4. Li YC, Korol AB, Fahima T, Nevo E. Microsatellites within genes: Structure, function, and evolution. Mol Biol Evol 2004; 21:991-1007.
5. Metzgar D, Liu L, Hansen C, Dybvig K, Wills C. Domain-level differences in microsatellite distribution and content result from different relative rates of insertion and deletion mutations. Genome Res 2002; 12:408-13.
6. Nadir E, Margalit H, Gallily T, Ben-Sasson SA. Microsatellite spreading in the human genome: Evolutionary mechanisms and structural implications. Proc Natl Acad Sci USA 1996; 93:6470-5.
7. Wilder J, Hollocher H. Mobile elements and the genesis of microsatellites in dipterans. Mol Biol Evol 2001; 18:384-92.
8. Boeva V, Regnier M, Papatsenko D, Makeev V. Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. Bioinformatics 2006; 22:676-84.
9. Borstnik B, Pumpernik D. Tandem repeats in protein coding regions of primate genes. Genome Res 2002; 12:909-15.
10. Kashi Y, King DG. Simple sequence repeats as advantageous mutators in evolution. Trends Genet 2006; 22:253-9.
11. Katti MV, Ranjekar PK, Gupta VS. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Mol Biol Evol 2001; 18:1161-7.
12. Morgante M, Hanafey M, Powell W. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nat Genet 2002; 30:194-200.
13. Toth G, Gaspari Z, Jurka J. Microsatellites in different eukaryotic genomes: Survey and analysis. Genome Res 2000; 10:967-81.
14. Archak S, Meduri E, Kumar PS, Nagaraju J. InSatDb: A microsatellite database of fully sequenced insect genomes. Nucleic Acids Res 2007; 35:D36-9.