Gene 504 (2012) 226–232

Contents lists available at SciVerse ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Genome-wide comparative analysis of simple sequence coding repeats among 25 insect species

Susanta K. Behura *, David W. Severson

Eck Institute for Global Health, Department of Biological Sciences, University of Notre Dame, Notre Dame, IN 46556, USA

ARTICLE INFO

Article history:
Accepted 12 May 2012
Available online 23 May 2012

Keywords:
Simple sequence repeats
Codon bias
Codon pair repeats
Insect
Comparative genomics

ABSTRACT

We present a detailed genome-scale comparative analysis of simple sequence repeats within protein coding regions among 25 insect genomes. The repetitive sequences in the coding regions primarily represented single codon repeats and codon pair repeats. The CAG triplet is highly repetitive in the coding regions of insect genomes. It is frequently paired with the synonymous codon CAA to code for polyglutamine repeats. The codon pairs that are least repetitive code for polyalanine repeats. The frequency of hexanucleotide and dinucleotide motifs of codon pair repeats is significantly ($p<0.001$) different in the *Drosophila* species compared to the non-*Drosophila* species. However, the frequency of synonymous and non-synonymous codon pair repeats varies in a correlated manner ($r^2=0.79$) among all the species. Results further show that perfect and imperfect repeats have significant association with the trinucleotide and hexanucleotide coding repeats in most of these insects. However, only select species show significant association between the numbers of perfect/imperfect hexamers and repeat coding for single amino acid/amino acid pair runs. Our data further suggests that genes containing simple sequence coding repeats may be under negative selection as they tend to be poorly conserved across species. The sequences of coding repeats of orthologous genes vary according to the known phylogeny among the species. In conclusion, the study shows that simple sequence coding repeats are important features of genome diversity among insects.

1. Introduction

The simple sequence repeats, also well known as microsatellites, are repetitions of sequence motifs of generally 1–6 bp that are ubiquitously found in all genomes (Tautz et al., 1986). Because of wide distribution in the genomes, the simple sequence repeat sequences are also found as the most commonly shared features among eukaryotic proteins (Golding, 1999; Huntley and Golding, 2005; Marcotte et al., 1999). It is well known that microsatellite loci undergo rapid expansion and contraction in length (Kruglyak et al., 1998; Lai and Sun, 2003; Tautz, 1989; Weber and Wong, 1993). Presence of microsatellite repeats within coding sequences is known to promote rapid variation of eukaryotic proteins (Kashi and King, 2006).

Although unequal crossing-over during meiosis often generates variation of repeat length of simple sequence repeats, replication slippage is considered as the major force in the evolution of these repeat sequences (Levinson and Gutman, 1987; Richard and Pâques, 2000; Schlotterer and Tautz, 1992). When a slippage error occurs within a microsatellite, it creates a loop in one of strands that gives rise to an insertion or a deletion in the subsequent replications depending upon if the loop is formed in the replicating strand or in the template strand respectively. This leads to increase or decrease of repeat length of microsatellites.

Apart from slippage, selection also plays a role in the variation or maintenance of repeats in the protein coding sequences (Huntley and Golding, 2006). The rate of mutation of dinucleotide repeats is generally higher than the rate of mutation of trinucleotide repeats (Schlotterer and Tautz, 1992). The same study (Schlotterer and Tautz, 1992) also suggests that repeats containing A/T are prone to higher mutation rate than repeats containing G/C. It has also been found that longer microsatellites have a higher mutation rate than small size microsatellites (Schlötterer, 1998; Wierdl et al., 1997) indicating that longer microsatellites are relatively more susceptible to potential slippage errors than short sequences. According to the proportional slippage model, microsatellite length variation is dependent on the mutation rate of the loci (Di Rienzo et al., 1994) whereas the step-wise mutation model (Ohta and Kimura, 1973) proposes that repeat sequences increase or decrease by one motif at a time. Furthermore, mutation bias has also been shown to affect microsatellite evolution both in

Abbreviations: SCR, single codon repeats; CPR, codon pair repeats; SAR, single amino acid repeats; APR, amino acid pair repeats; RSCU, relative synonymous codon usage; Syn, synonymous; Non-syn, non-synonymous; Dmel, *Drosophila melanogaster*; Dsim, *Drosophila simulans*; Dsec, *Drosophila sechellia*; Dyak, *Drosophila yakuba*; Dere, *Drosophila erecta*; Dana, *Drosophila ananassae*; Dpse, *Drosophila pseudoobscura*; Dper, *Drosophila persimilis*; Dwil, *Drosophila willistoni*; Dgri, *Drosophila grimshawi*; Dvil, *Drosophila virilis*; Dmoj, *Drosophila mojavensis*; Aaeg, *Aedes aegypti*; Agam, *Anopheles gambiae*; Cqui, *Culex quinquefasciatus*; Acep, *Atta cephalotes*; Cflo, *Camponotus floridanus*; Lhum, *Linepithema humile*; Hsal, *Harpegnathos saltator*; Pbar, *Pogonomyrmex barbatus*; Nvit, *Nasonia vitripennis*; Amel, *Apis mellifera*; Phum, *Pediculus humanus*; Bmor, *Bombyx mori*; Apis, *Acyrthosiphon pisum*.

* Corresponding author. Tel.: +1 574 904 2794; fax: +1 574 631 7413.
E-mail address: sbehura@nd.edu (S.K. Behura).

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2012.05.020

prokaryotes and eukaryotes (Metzgar et al., 2002; Rubinsztein et al., 1999). Collectively, these studies have suggested that evolution of simple sequence repeats is a complex process (Ellegren, 2004; Wu and Drummond, 2011).

In insects, although simple sequence repeats have been extensively exploited as molecular markers in ecology and population studies (Behura, 2006), the coding features of simple sequence repeats have not been well studied. Although numerous studies have been conducted in discovering microsatellites either experimentally or computationally from whole genome sequences or expressed sequence tags (ESTs) (Sharma et al., 2007; Vasemägi et al., 2005; Zane et al., 2002), distribution of simple sequence repeats representing codon repeats is not well understood. Previously, a comparative analysis was performed to study the amino acid repeats among the sequenced genomes of twelve *Drosophila* species (Huntley and Clark, 2007). But, this investigation was not oriented to address the said objectives of the present study. Moreover, genome sequences of a number of insect species are now available where no information on codon repeats is available. In this study, we present a detailed investigation on simple sequence repeats within protein coding sequences in genome-scale manner among 25 insect species.

2. Materials and methods

2.1. Sequence data

A total of 25 insect genomes were investigated in this study. They included twelve *Drosophila* species [*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*, *D. pseudoobscura*, *D. persimilis*, *D. willistoni*, *D. grimshawi*, *D. virilis*, *D. mojavensis*], three mosquito species [*Aedes aegypti* (*A. aegypti*), *Anopheles gambiae* (*A. gambiae*), *Culex quinquefasciatus* (*C. quinquefasciatus*)], five ant species [leaf cutter ant (*Atta cephalotes*), carpenter ant (*Camponotus floridanus*), Argentine ant (*Linepithema humile*), jumping ant (*Harpegnathos saltator*) and red harvester ant (*Pogonomyrmex barbatus*)] and the wasp (*Nasonia vitripennis*), the honey bee (*Apis mellifera*), the body louse (*Pediculus humanus*), the silk worm (*Bombyx mori*) and the pea aphid (*Acyrthosiphon pisum*). The insect names have been abbreviated as the first letter of the genus followed by three letters of the species names throughout the text and the illustrations. The annotated coding sequences (CDS) of the twelve *Drosophila* genes were downloaded from FlyBase (www.flybase.org). They were r1.3 version for each *Drosophila* (except r5.27 for *D. melanogaster*, r2.10 for *D. pseudoobscura* and r1.2 for *D. virilis*). The coding sequences of the three mosquitoes and the body louse were downloaded from VectorBase (http://www.vectorbase.org). The CDS of *A. mellifera* genes (pre-release 2), and the four ant species (Acep OGS1.2, Cflo v3.3, Hsal v3.3, Lhum OGS1.2 and Pbar OGS 1.2) were downloaded from http://hymenopteragenome.org/. The Nasonia (*N. vitripennis*) coding sequences (*N. vitripennis*_OGS_v1.2) were obtained from http://www.hgsc.bcm.tmc.edu. The aphid CDS and protein sequences were obtained from the AphidBase (http://www.aphidbase.com/aphidbase/). The silkworm CDS and protein sequences were obtained from the SilkDB (http://www.silkdb.org/silkdb/). The protein fasta files of each genome were also obtained from the respective sources.

2.2. Identification of simple sequence coding repeats

The simple sequence coding repeats were identified using a method as shown in Fig. 1. First, the CDS sequences were aligned with the protein sequences using the RevTrans software (Wernersson and Pedersen, 2003) to extract the codon sequences of genes in each genome. The codon sequences (5′–3′) were then subjected to SciRoKo, a simple sequence repeat (SSR) identification program (Kofler et al., 2007) to identify SSRs in the protein coding sequences. The coding motifs repeated more than 3 times were considered as repetitive in each case. The genes where one or more coding sites were ambiguous nucleotide (such as 'N's) were excluded from the analysis. The mono-, di-, tri- and tetra- and hexa-nucleotide SSRs were searched comprehensively to extract both perfect and imperfect repeat sequences by SciRoKo. The SciRoKo program was set to the default parameters (mismatch, fixed penalty = 5). The repeats with more than 3 consecutive mismatch sites were not allowed to report.

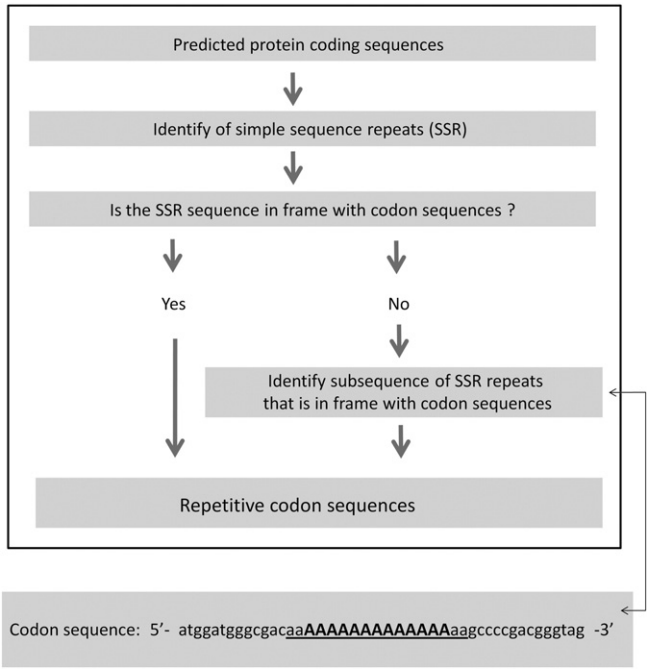

Fig. 1. Schematic description of method that was used to identify simple sequence coding repeats from whole-genome sequences. An example is provided to explain how subsequences of SSRs were determined wherein the extracted sequences were in frame with the codon sequences (bold and underlined) of the genes.

Table 1
Count statistics of simple sequence coding repeats in the genome of 25 species. The average length shown is in basepairs. Density is expressed as number of repeats per Mbp of coding sequences. Percentage is expressed as amount (in bp) of repeats to the total amount of coding sequences in the genome.

Species	Counts	Avr. length	Density	Percentage
Aaeg	818	19.94	34.49	0.069
Acep	1918	41.12	98.18	0.404
Agam	5582	23.75	246.28	0.585
Amel	1817	27.36	99.7	0.273
Apis	4283	23.14	120.08	0.278
Bmor	694	24.86	38.76	0.096
Cflo	1344	39.54	64.77	0.256
Cqui	2976	23.07	120.1	0.277
Dana	1540	23.44	136.26	0.319
Dere	4184	26.94	190.98	0.515
Dgri	9392	25.99	421.01	1.094
Dmel	6923	25.04	156.66	0.392
Dmoj	9056	30.7	418.61	1.285
Dper	6344	24.96	292.58	0.730
Dpse	7010	24.99	294.97	0.737
Dsec	2811	24.79	130.69	0.324
Dsim	2586	24.53	135.89	0.333
Dvir	8942	27.51	411.18	1.131
Dwil	7303	23.67	321.81	0.762
Dyak	3983	26.51	175.82	0.466
Hsal	3832	63.62	187.71	1.194
Lhum	2468	33.58	120.3	0.404
Nvit	2836	25.21	96.19	0.242
Pbar	3042	37.9	148.04	0.561
Phum	4232	22.31	254.26	0.567