Romanian Biotechnological Letters *Vol. 16, No. 3, 2011*

ORIGINAL PAPER

Insertional hotspots of artificial P transposons are tagged by consensus motifs in various genomic sequences of *Drosophila melanogaster*

Received for publication, December 08, 2010
Accepted, May 11, 2011

A. A. ECOVOIU[1], A. C. RATIU[1], M. GRAUR[1], L. SAVU[2]
[1]Department of Genetics, University of Bucharest, Aleea Portocalelor 1-3, Sect. 6, Bucharest, Romania.
[2]Genetic Lab, Bd. Ghencea 43 B, Ghencea Business Center – 3rd floor, Sect. 6, Bucharest, Romania.
Corresponding author: Ratiu Attila Cristian, Department of Genetics, University of Bucharest, Intrarea Portocalelor 1-3, 060101, phone number: 0040722250366, Fax: 0213181565, email: ratiuattila@botanic.unibuc.ro

Abstract

In an attempt to reveal the biological significance of the P transposon insertional hotspots found in Drosophila melanogaster genome, we performed a bioinformatics analysis upon an original collection of mutant lines harboring reinsertions of P{lacW} and P{EP} artificial derivatives. Consensus motif sequences were identified in various arbitrary 200 bp sequences, centered around the insertion sites, suggesting that the hotspots are determined by the local genomic landscape. The consistent hotspots are placed mainly in introns and such an insertional preference may reflect an interdependent regulatory behavior of both introns and transposons. The consensus motif KMGAAD defining a HSE (heat-shock element) unit and BGAGHGV consensus closely resembling the GAGA-factor binding site are well represented in the arbitrary sequences hosting hotspots. The presence of both consensus motifs in noncoding regions could be an evidence that candidate molecular relics of a heat-shock-like regulatory strategy are located throughout the genome of D. melanogaster.

Keywords: *P{lacW}*, *P{EP}*, *insertional hotspots, consensus motif sequence, Drosophila melanogaster*.

Introduction

Understanding the molecular mechanisms responsible for the biases of insertions/reinsertions of P mobile elements is a critical aspect for revealing the physiologically roles of insertional hotspots in *D. melanogaster* genome, but also for refining the use of the artificial transposon derivatives in experimental strategies. The insertional hotspots are present in *5'UTR*s, upstream of *5'UTR*s, in introns, in exons and even in intergenic regions. If hotspots located in *5'UTR*s may reflect a regulatory role of P family transposons, questions are raised by the existence at relative high-frequency of such hotspots in introns. Our data suggest that hotspots located in big introns are alluring targets for *P{lacW}* reinsertions when the original insertion is repaired/conserved in the same genetic background. It makes sense to look for consensus sequences/motifs in the close vicinity of the hotspots in order to decipher how and why such insertional preferences are determined. Finding attractor sequences for insertion is a classical effort in the field of P element biology, but discrepant data were obtained during the past few decades [1], [2]. The reported results reflect a focus on the very close vicinity of insertion sites, or target sites. We consider that narrowing of the searching area around the very close vicinity of the insertion situs may impede on revealing of potential attractor sequence motifs present in a more extended genomic landscape. Starting from such an assumption, we defined an arbitrarily sequence of

200 bp encompassing each of the insertion sites, represented by two 100 bp windows on each side of the insertion point. The arbitrary sequences were searched with two motif finder programs, *SCOPE* and *Sequence Manipulation Suite* (*SMS*) respectively, [3], [4].

We analyzed a battery of 29 such arbitrary sequences, harboring either a *P{lacW}* or a *P{EP}* insertion in the original mutant line. The transgenic lines were generated by insertional mutagenesis, the sequence of each allele was sent to *GenBank* [5] and the accession numbers are reported in Table 1. The insertion points were determined with *Genome ARTIST* bioinformatics tool [6], the fragments pertaining to the artificial P elements were then removed, and 100 nucleotides genomic fragments from both sides of each insertion point were defined with *BLAT* application [7], generating the arbitrary sequences. The arbitrary sequences were converted to FASTA format, compiled in a single document and comparatively screened for hosting of potential consensus motifs with *SCOPE*. The results generated by *SCOPE* were further checked for conformity with *DNA Pattern Find* application from *SMS*. The specific motifs/consensus sequences were also manually annotated and evaluated with *Tomtom* application [8] and statistically significant sequences were chosen for further interpretation. Our work revealed two main consensus motifs residing in almost each of the 29 arbitrary sequences, in the close vicinity of insertional hotspots described in *FlyBase* [9].

Materials and methods

The mutant lines harboring reinsertions were obtained by mobilization of $P\{lacW\}^{gammaCop057302}$ artificial mobile element from region 100C in the genetic background of *l(3)S057302* transgenic line (details about genetics strategy are described elsewhere [10]) or by mobilization of *P{EP}EP3313* artificial transposon close to CG6199 from 68B1 [11]. In almost all of the new lines, the original insertion was apparently repaired. The insertion sites were detected by a combined strategy of inverse PCR and sequencing [12] and the specific sequences were submitted to *GenBank* and are available under accession numbers reported in the text.

The exact insertion sites were detected with *Genome ARTIST* software [6], the arbitrary 200 bp sequences around each insertion were defined with *BLAT* browser, and the search for consensus sequences was performed with *SCOPE* motif finder and double checked with *SMS*. Insertional hotspots were identified in *D. melanogaster* genome *R5.26* with *GBrowse* application from *FlyBase* [9].

Comparative analysis of the putative consensus motifs was performed with *Tomtom* application [8] from *MEME* suite searching the *flyreg.v2.meme* database (http://www.danielpollard.com/bergman2004matrices.html).

Results and Discussions

In an attempt to identify consensus motifs in the close vicinity of different insertional hotspots, we scrutinized a total of 29 arbitrary sequences, representing symmetric sequences of 200 nucleotides centered around the coordinates of each insertional mutation. The insertions are different regarding to the nature of the artificial transposon (*P{lacW}* or *P{EP}*) employed for mutagenesis, the germline type affected by the reinsertion of the transposon (namely if transposition was induced in the male or female germline) and to the genic region affected by mutation; the details are presented in Table 1. We evaluated 20 insertions of *P{lacW}* and 9 insertions of *P{EP}*; two of them were generated outside of our laboratory, particularly the lines *l(3)S057302* [13] and *EP3313* [14] harboring the original insertions

located in *5'UTR* of *gammaCop* and downstream of *CG6199* respectively. Regarding to the germline nature, 14 insertions originated in the male germline (including *l(3)S057302* and *P{EP}EP3313*) and 15 were induced in the female germline.

The exact sites of transposon insertions defining each mutant allele were detected with *Genome ARTIST* [6], as exemplified in Figure 1.

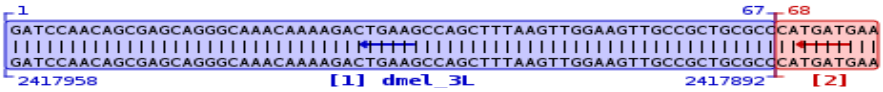

Figure 1. Detection of the exact insertional locus defining a mutant allele of *CG42669* gene (the complete sequence is available in *GenBank*/NCBI under the accession number *HM210948*). The fragment labeled as **[1] dmel_3L** belongs to *CG42669* gene and the sequence fragment tagged as **[2]** belongs to *P{lacW}* artificial transposon.

When the affected genic region is considered, the most abundant insertions are located in introns (12) and in *5'UTR* (7) respectively. Two insertions affect exons, whereas eight insertions are located in intergenic regions (the closest gene was considered as the affected one and some insertions are located just upstream of *5'UTR* sequences). Simultaneous comparative analysis of the arbitrary sequences with *SCOPE* revealed the presence of 63 repeats of KMGAAD (5'[GT][CA]GAA[ATG]3') consensus motifs, distributed on 24 of the arbitrary sequences. Only *CG14709*, *CG1500*, *CG42669*, *CG5222* and *CG9805* genes do not host a KMGAAD consensus motif in their respective arbitrary sequence. Comparative study with *Tomtom* application revealed that KMGAAD is closely related with HSE (HSF Sequence-binding Elements), with a high statistical significance (p = 2.84442e-07).

A second relatively abundant consensus motif detected in the arbitrary sequences group is BGAGHGV (5'[TCG]GAG[CTG]G[AGC]3'), detected in 37 copies distributed on 17 genes described herein. Bioinformatics investigation performed with *Tomtom* algorithm unveils that BGAGHGV is similar with both *z* (*zeste*) and Trithorax-like/GAGA (*Trl*) consensus motifs, with a high statistical significance (p = 1.25461e-05 and p = 0.00214451 respectively). Some of the arbitrary sequences contain the equivalent GAGCG motif instead of BGAGHGV.

It is interesting to note that three genes, namely *CG1500*, *CG42699* and *CG5222*, do not contain any of the two candidate consensus motifs described above. The 200 bp arbitrary sequences pertaining to *CG42699* and *CG5222* host only mini hotspots of 2 insertions each, suggesting that an absence of such consensus motifs might be related with a low abundance of the insertions in the hotspots. On the contrary, *CG1500* has a very consistent hotspot of 29 insertions placed in the arbitrary sequence, raising questions about such a special case. Supplemental bioinformatics analysis revealed that the arbitrary sequence of *CG1500* contains 8 copies of TCGGNTBVRYTCGG (5'TCGGNT[TGC][GCA][GA][TC]TCGG3') motif located just next to the insertion point. The exotic consensus motif is also present in one copy in the arbitrary sequence of *CG1007*, hosting a very consistent hotspot of 40 insertions. Although with poor statistical significance (p = 0.119439), TCGGNTBVRYTCGG consensus motif is partially related to the *z* binding motif. It is possible that the tandem repeated copies located in *CG1500* are able to surmount the absence of consensus sequences KMGAAD and BGAGHGV.

Table 1. A comprehensive description of the insertional mutations and consensus motifs associated with insertional hotspots. Consensus motifs located on the sense strand of each affected gene are explicitly mentioned from the total repeats of the consensus (ex: 3/2 notation means that there are 3 consensus motifs in the arbitrary sequence, but only 2 are located on the sense strand of the gene). A virtual sense strand matching the sense strand of each affected gene was arbitrarily considered for insertions in intergenic regions. Any hotspot defined by at

least 3 insertions described in *GBrowse* of *FlyBase* was considered a consistent hotspot, otherwise it was considered a mini hotspot. Our insertions were not counted for such a classification, so where there is no insertion reported in *GBrowse* for an arbitrary 200 bp sequence, actually at least one *P{lacW}* or *P{EP}* insertion placed in that region should be considered.

Affected gene/cytological map location	Insertion site/GenBank accession number	Affected region	Original insertion fate	Consensus Motif Sequences	Affected germline	Observations
CG10640 **(*Uev-1A*)** 3L; 64C13	*5358453* ***FJ603691***	Intron	*P{lacW}, repaired*	KMGAAD (3/2) BGAGHGV (1/1)	♂	The insertion is located in a big intron and there is no other reported insertion placed in the arbitrary target sequence (5358354-5358553).
CG42803 **(*grappa*)** 3R; 83E6-E7	2250049 ***FJ603693***	Intron	*P{lacW}, repaired*	KMGAAD (2/2) BGAGHGV (2/2)	♂	The insertion is located in a big intron and there is only one other but close insertion placed in the arbitrary target sequence (2250149-2249950), at nucleotide 2250084.
CG42551 **(*larp*)** 3R; 98C3-4	24152042 ***HM210954***	Intron	*P{lacW}, lost*	KMGAAD (6/4) BGAGHGV (3/3)	♂	There are 14 insertions placed in the arbitrary sequence (24151939-24152138), defining a **consistent hotspot** in a big intron; 4 different insertions are located also at 24152042**.**
CG9381 **(*mura*)** 3R;85D22-24	5374735 ***HM210955***	Intron	*P{lacW}, repaired*	KMGAAD (2/0) GAGCG (1/0)	♂	The insertion is located in a big intron. No other reported insertion is placed in the arbitrary target sequence (5374839-5374640), but *mura* contains three hotspots just upstream of the target sequence.
CG13895 3L; 61C8	707809 ***FJ603694***	Intron	*P{lacW}, repaired*	KMGAAD (1/1)	♂	The insertion is located in a short intron, in a **consistent hotspot** represented by 9 insertions placed in the arbitrary target sequence (707709-707908); two insertions are located in the genomic range 707798-707816.
CG11172 **(*NFAT*)** X; 12A9-B2	13536112 ***HM210949***	Intron	*P{lacW}, repaired*	KMGAAD (5/1) BGAGHGV(3/2)	♀	The insertion is located in a big intron, in a **consistent hotspot** represented by 22 insertions placed in the arbitrary target sequence (13536210-13536011). Most of them (13) are located quite close to our insertion, in the genomic range 13536109-13536120. There is no other insertion matching nucleotide 13536112, but 3 of them hit 13536111 nucleotide.
CG18135 3L; 75F2	18989855 ***HQ695001***	Intron	*P{lacW}, lost*	KMGAAD (2/1) BGAGHGV (1/1)	♀	The insertion is located in a big intron, in a **consistent hotspot** represented by 3 insertions placed in the arbitrary target sequence (18989754-18989953). There is no other insertion matching nucleotide 18989855.
CG7702 3R; 91B8-C1	14495944 ***HM210944***	Intron	*P{lacW}, partially repaired*	KMGAAD (5/0) BGAGHGV (1/0)	♀	Reinsertion is located in a big intron; there is only one close insertion placed in the arbitrary target sequence (14495844-14496043), at 14495896.
CG9364 (*Treh*) 2R; 57B16-19	16963565 ***HM210950***	Intron	*P{lacW}, repaired*	KMGAAD (4/3) BGAGHGV (3/2)	♀	The insertion is located in a big intron, in a **consistent hotspot** represented by 12 insertions placed in the arbitrary target sequence (16963465-16963664). Most of them (5) are located quite close to our insertion, in the genomic range 16963560-16963580.
CG42669 3L; 62E2	2417892 ***HM210948***	Intron	*P{lacW}, repaired*	-	♀	The insertion is located in a big intron, in a mini hotspot located in the arbitrary sequence (2417991-2417792), consisting of 2 insertions at nucleotides 2417863 and 2417924.
CG4029 **(*jumu*)** 3R; 86B1	6176448 ***GQ401363***	Intron	*P{EP}, repaired*	KMGAAD (3/1) BGAGHGV (1/1)	♂	The insertion is located in a big intron, and there are two other very close insertions in the arbitrary target sequence (6176546-6176347), located at nucleotides 6176438 and 6176446.
CG1677 X; 7A8-B2	7179081 ***GU134146***	Intron	*P{EP}, repaired*	KMGAAD (3/2)	♀	The insertion is located in a big intron and there is only one other reported insertion placed in the arbitrary target sequence (7178981-7179180) at nucleotide 7179129.
CG1528 **(*γCop*)** 3R; 100C6	27397926 ***AJ492220***	5'UTR	*P{lacW}*	KMGAAD (1/1) BGAGHGV (1/1)	♂	There are two insertions located in the arbitrary sequence (27397826-27398025), at coordinates 27397805 and 27397901.

***CG1433* (*Atu*)** 3R; 83B6	1439101 ***HM210951***	5'UTR	*P{lacW}, repaired*	KMGAAD (2/2)	♂	The arbitrary sequence (1439003-1439202) contains a **consistent hotspot** of 6 insertions and one of them is located right at 1439101.
CG6568 2R; 54B7	13294911 ***HM210952***	5'UTR	*P{lacW}, repaired*	KMGAAD (4/1) GAGCG (1/0)	♂	The arbitrary sequence (13294813-13295012) contains only one insertion.
CG9805 (*eIF3-S10*) 3R; 82B1	259590 ***GU814269***	5'UTR	*P{lacW}, repaired*	GAGCG (1/0)	♀	The arbitrary sequence (259411-259690) contains 9 insertions in a **consistent hotspot**, the closest one is located at 259600.
***CG8846* (*Thor*)** 2L; 22F3-F6	3478544 ***GU814268***	5'UTR	*P{lacW}, repaired*	KMGAAD (2/1)	♀	There are two insertions bordering the arbitrary sequence (3478643-3478444); just upstream of the arbitrary sequence there is a **consistent hotspot**.
CG*17342* (*Lk6*) 3R; 86E18	7590179 ***GU134144***	5'UTR	*P{EP}, repaired*	KMGAAD (2/1)	♂	The arbitrary sequence (7590081-7590280) contains a **consistent hotspot** of 38 insertions and two of them are located at 75901178.
***CG42396* (*wech*)** 2R; 43C5	3377332 ***GU134145***	5'UTR	*P{EP}, repaired*	KMGAAD (3/2) BGAGHGV (3/3)	♀	There is a **consistent hotspot** of 8 insertions in the arbitrary sequence (3377227-3377426), one of them is placed right at 3377332 nucleotide.
***CG12410* (*cv*)** X; 5A13	5584012 ***HM210956***	Intergenic	*P{EP}, repaired*	KMGAAD (1/0) BGAGHGV (2/0)	♀	The arbitrary sequence (5583912-5584111) contains a **consistent hotspot** of 5 insertions, the closest of them is located at 5584024.
CG1500 **(*furrowed*)** X; 11A1	11898010 ***HM216795***	Intergenic	*P{EP}, repaired*	TCGGNTBVRYTCGG (8/0)	♀	The insertion is located in a **consistent hotspot** represented by 29 insertions placed in the arbitrary target sequence (11897913-11898112); 12 insertions are at 11898010-11898011, revealing an **absolute hotspot**. There are 7 contiguous special sequences on the sense strand, jus next to the insertion site.
CG14709 3R; 86E11	7394886 ***HM210946***	Intergenic	*P{lacW}, repaired*	BGAGHGV (4/4)	♀	The insertion is located upstream of *CG14709* gene, at coordinates 7394886, in a **consistent hotspot** represented by about 50 insertions placed in the arbitrary target sequence (7394985-7394786). Almost all of them are located quite close to our insertion and, remarkably, 12 of them are placed in the genomic range 7394886-7394887.
CG34460 2R;53D11	12716578 ***HM210947***	Intergenic	*P{lacW}, repaired*	KMGAAD (2/1) GAGCG (1/0)	♀	The insertion is located at coordinates 12716578, in a **consistent hotspot** represented by 21 insertions placed in the arbitrary target sequence (12716677-12716478); 20 are located quite close to our insertion, in the genomic range 12716547-12716585, 3 insertions located right at 12716578 and one at 12716577.
***CG31349* (*pyd*)** 3R;85B2-7	4757618 ***FJ603690***	Intergenic	*P{lacW}, repaired*	KMGAAD (3/1) BGAGHGV (2/1)	♂	The insertion is located upstream of *CG31349* gene, in a **consistent hotspot** represented by 10 insertions placed in the arbitrary target sequence (4757718-4757519). Almost all of them are located quite close to our insertion and 3 of them are placed in the genomic range 4757610-4757621.
CG5222 3L; 72D6	16098780 ***FJ603692***	Intergenic	*P{lacW}, repaired*	-	♂	The insertion is quite upstream of *CG5222* gene, in a mini hotspot represented by 2 insertions placed in the arbitrary target sequence (16098681-16098880), one of them being placed in close vicinity, at nucleotide 16098775.
***CG1007* (*emc*)** 3L; 61C9	749348 ***GU134147***	Intergenic	*P{EP}, lost*	KMGAAD (1/1) BGAGHGV (5/2) TCGGNTBVRYTCGG (1/1)	♀	Reinsertion located upstream of *CG1007*, in a **consistent hotspot** represented by about 40 insertions placed in the arbitrary target sequence (749427-749228), where 9 of them are located in the interval 749340-749346. There is one special consensus on the antisense strand, just next to the insertion.
CG6199 3L; 68B1	11190673 ***AQ073905***	Intergenic	*P{EP}*	KMGAAD (2/1) BGAGHGV(1/0)	♂	There is a **consistent hotspot** in the arbitrary sequence (11190597-11190796) represented by 5 insertions; one of them is located right at 11190673.

CG30115 2R; 55D3-E1	14498718 ***GQ401364***	Exon	*P{EP}, lost*	KMGAAD (2/0) BGAGHGV(1/1)	♂	There is a small hotspot in the arbitrary sequence (14498618-14498817) represented by only one insertions one of them is located right at 14498796.
CG11033 3R; 85C3-4	4883677 ***HM210945***	Exon	*P{lacW}, repaired*	KMGAAD (2/2) BGAGHGV (3/2)	♀	A **consistent hotspot** in the artificial sequence (4883577-4883776) represented by 5 insertions, one of them is located right at 4883677.

Regarding to the distribution of the hotspots among the genic regions, it results from Table 2 that most of the consistent hotspots are located in *5'UTR*s and in introns, but the highest density of hits in such hotspots occurs in intergenic regions. When a specific germline is taken into consideration, the consistent hotspots were more abundantly hit in the germline of transgenic females (11/17).

Table 2. From a total of 17 consistent hotspots, 11 were hit by reinsertions underwent in the female germline. Almost all of the intergenic reinsertions hit consistent hotspots (7/8), and 5 of them were detected in strains derived from transpositions events occurring in the female germline.

Location of the consistent hotspots	5'UTR	Introns	Exons	Intergenic
Frequency	4/7	5/12	1/2	7/8
Male germline	2/4	2/5	-	2/7
Female germline	2/4	3/5	1/1	5/7

Our bioinformatics work revealed that most of the arbitrary sequences considered herein harbor 1-9 copies of the described consensus motifs (KMGAAD, BGAGHGV, GAGCG or TCGGNTBVRYTCGG). There are 10 arbitrary sequences containing at least 5 copies of the consensus motifs (see Table 1) and 7 out of them contains consistent hotspots (defined as such if they harbor at least 3 insertions). It is interesting to point out that 6 of these hotspots are remarkable since each of them hosts at least 10 insertions (see the cases of *CG42551*, *CG11172*, *CG9364*, *CG1500, CG31349* and *CG1007* genes) as revealed by *GBrowse* application from *FlyBase*. For 6/7 of them, the hit of the consistent hotspots is associated with repairing of the original insertion in our experiments. Such data point to the fact that a high density of the consensus motifs around the potential insertional target contributes to defining of consistent hotspots.

Tian *et al*., 2010 revealed that HSEs evolve in size and sequence inside of the promotor sequences. It worth to consider that the plasticity of the genome relocated former promotor sequence modules into intronic or intergenic region during its evolution, prior to P element invasion [15]. If P mobile element used to have a HSE-dependent regulatory role in the genome of its original host, it is possible to exhibit a similar tendency in *D. melanogaster* also. The HSE reminiscent modules may behave as molecular attractors for P transposon and their artificial derivatives, leading to situations where insertional hotspots do not necessarily reflect an evident regulatory role. On the other hand, at least for genes subjected to alternative splicing, some of their introns may be involved in gene expression regulation. If such introns harbor gapped or complete HSE consensus motifs they are prone to be preferentially hit by P element derivatives. KMGAAD consensus motif is different but more stringent than the NGAAN subunit sequence found in the canonical HSE [15], but it was not found in the classical contiguous alternating pattern in our arbitrary sequences. The canonical HSEs are contributing to an "opening" of the adjacent genomic sequence for insertional phenomena [16]. By a similar mechanism BGAGHGV consensus motifs close to KMGAAD units may increase the incidence of insertional hotspots in the arbitrary sequences. Our data are consistent with a special case described by Shilova *et al*., 2006 in a study about the features of heat-shock genes biasing them to insertional mutagenesis. They revealed that HSE and GAGA binding sites are abundant in the promoter of *Hsp70* gene, predisposing it to insertions of P-

derived artificial transposons. Most of our arbitrary sequences exhibit a similar but less stringent molecular landscape, since 19/29 of the 200 bp sequences host both KMGAAD and BGAGHGV (the alternative GAGCG motif was considered for arbitrary sequences missing BGAGHGV) consensus motifs, paralleling HSE and GAGA binding sites described by Shilova *et al*., 2006. Most of them (14/19) are located in intronic or intergenic regions, partially explaining why they do not perfectly match the canonical motif sequences described for *5'UTR* of *Hsp70*. Since putative regulatory intronic or intergenic sequences do not perfectly fit the characteristics of *5'UTR*, the presence of only one of the consensus motifs described herein may still predispose to a high local rate of insertions able to trigger physiological effects.

The consensus motifs KMGAAD and BGAGHGV may represent molecular relics of the canonical sequences of HSE and *Trl/z*, witnessing genome's restructuration during the evolutionary process of *D. melanogaster*. The genomic landscape is rapidly reorganized by transposition events [17] therefore, revealing the biological significance of the insertional patterns of P elements contributes to understanding of such dramatic changes.

It may be assumed that the genes harboring consistent hotspots in the respective arbitrary sequence (Table 1) are prone to a more rapid evolutionary process mediated by transposition. P elements may create variation in the regulatory sequences of different genes, a condition for the evolutive process. Genes enriched with consensus motifs that make them more amenable to be preferentially hit by P elements offer a vast material for evolution. It makes sense to reinforce that the majority of the insertional alleles described in our paper are actually reinsertions of an often conserved/repaired original insertion (either *P{lacW}* or *P{EP}*), a condition posing questions about the biological significance of targeting a new sequence. It is tempting to assume that the secondary target shares similar consensus motifs with the one hosting the original, repaired insertion. As described in Table 1, the *5'UTR* of *gammaCop* gene is the host of the original $P\{lacW\}^{gammaCop057302}$ insertion, which is repaired in almost all of the mutant lines harboring reinsertions in different genes. As expected, the *5'UTR* of *gammaCop* hosts both KMGAAD and BGAGHGV consensus motifs, suggesting that the excised original element hits related sequences, at least when the original insertion is repaired. A similar situation may be observed for the original insertion *P{EP}EP3313* located close to *CG6199* gene, where again KMGAAD and BGAGHGV consensus motifs are found in the arbitrary sequence encompassing the original insertion.

We detected 17 consistent hotspots and 11 out of them were hit in our experiments by reinsertions induced in the transgenic female germline. The data suggest that such hotspots may be more abundant in female genome, where many genes are maternal-effect or are involved in the adaptive response. Such genes may have a more rapid pace of evolution if they are preferred targets for insertional mutagenesis via P derivative transposons. The evolutionary aspects concerning the physiological role of P element insertional tendencies and the biological significance of the insertional hotspots are still elusive but the combined efforts of bioinformatics and genomics are about to solve some aspects of the issue. We presume that the presence of KMGAAD and BGAGHGV consensus motifs in some noncoding sequences of *D. melanogaster* genome that host insertional hotspots may represent molecular relics of an ancient heat-shock-like gene regulatory strategy.

Conclusions

We detected the consensus motifs KMGAAD and BGAGHGV in the sequences encompassing insertional hotspots of *P{lacW}* and *P{EP}* artificial transposons. Our data suggest that the short sequences of 8-14 bp defining the insertion sites are not exhaustively informative for understanding the insertional patterns. The study of more extended regions

centered around the insertion sites may reveal interesting molecular characteristics, including sequences that predispose some genomic regions to be hit by multiple insertions. A heat-shock like regulatory strategy involving P transposons insertions in intronic and intergenic regions could be considered as a way to induce adaptive biological responses. The presence in the noncoding regions of consensus sequences specific for *5'UTR* of *Hsp70* may also represent an example of molecular relics generated during the genome evolution.

Acknowledgments

The sequencing work was kindly performed by Adriana Maria Stan from Genetic Lab, Bucharest, Romania. **This research was financially supported by the research grant PNII/IDEI no.147/2007, CNCSIS, Romania.**

References

1. K. O'HARE, G.M. RUBIN, *Structures of P transposable elements and their sites of insertion and excision in the Drosophila melanogaster genome*, Cell 34, 25-35 (1983).
2. G.C. LIAO, E.J. REHM, G.M. RUBIN, *Insertion site preferences of the P transposable element in Drosophila melanogaster,* Proc. Natl. Acad. Sci. USA 97, 3347-3351 (2000).
3. J.M. CARLSON, A. CHAKRAVARTY, C.E. DEZIEL, R.H. GROSS, *SCOPE: a web server for practical de novo motif discovery*, Nucleic Acids Research 35, W259–W264 (2007). http://genie.dartmouth.edu/scope/
4. P. STOTHARD, *The Sequence Manipulation Suite: JavaScript programs for analyzing and formatting protein and DNA sequences*, Biotechniques 28, 1102-1104 (2000). http://www.bioinformatics.org/sms
5. D.A. BENSON, I. KARSCH-MIZRACHI, D.J. LIPMAN, J. OSTELL, E.W. SAYERS, *GenBank*, Nucleic Acids Research 38 (suppl 1), D46-D51 (2010). http://www.ncbi.nlm.nih.gov/genbank
6. A.A. ECOVOIU, I.C. GHIONOIU, M.A. CIUCA, A.C. RATIU, M. GRAUR, *Genome ARTIST (Genome Artificial Transposon Insertion Site Tracker), a bioinformatics tool for rapid detection of insertional mutations*, (2010). http://www.bio.unibuc.ro
7. W.J. KENT, *BLAT - the BLAST-like alignment tool*, Genome Res. 12 (4), 656-64 (2002).
8. S. GUPTA, J.A. STAMATOYANNOPOULOS, T.L. BAILEY, W.S. NOBLE, *Quantifying similarity between motifs,* Genome Biology 8 (2), R24 (2007).
9. S. TWEEDIE, M. ASHBURNER, K. FALLS, P. LEYLAND, P. MCQUILTON, S. MARYGOLD, G. MILLBURN, D. OSUMI-SUTHERLAND, A. SCHROEDER, R. SEAL, H. ZHANG, THE FLYBASE CONSORTIUM, *FlyBase: enhancing Drosophila Gene Ontology annotations*, Nucleic Acids Research 37, D555-D559 (2009). www.flybase.org
10. A.C. RATIU, A.A. ECOVOIU, M. GRAUR, L. SAVU, L. GAVRILA, *Transposition of P{lacW}gammaCop057302 into the germline of Drosophila melanogaster correlates with retaining of the original insertion*, Roumanian Biotechnological Letters, vol 13, no 5, 3891-3900 (2008).
11. M. GRAUR, A.C. RATIU, A.A. ECOVOIU, L. SAVU, *Mobilization of P{EP}EP3313 atificial transposon in the germline of Drosophila melanogaster***,** Bulletin USAMV/Animal Sci. and Biotech. 66 (1-2), 441-446 (2009).
12. A.C. RATIU, A.A. ECOVOIU, M. GRAUR, L. SAVU, *Mapping of multiple P{lacW} insertions into the germline of Drosophila melanogaster*, Bulletin USAMV/Animal Sci. and Biotech. 66 (1-2), 424-429 (2009).
13. P. DEAK, M.M. OMAR, R.D.C. SAUNDERS, M. PAL, O. KOMONYI, J. SZIDONYA , P. MARÓY, Y. ZHANG, M. ASHBURNER, P. BENOS, C. SAVAKIS,I. SIDEN-KIAMOS, C. LOUIS, V.N. BOLSHAKOV, F.C KAFATOS, E. MADUENO, J. MODOLELL, D.M. GLOVER, *P element insertion alleles of essential genes on the third chromosome of Drosophila melanogaster: correlation of physical and cytogenetic maps in chromosomal region 86E87F,* Genetics 147, 1697-1722 (1997).
14. P. RORTH, *A modular misexpression screen in Drosophila detecting tissue-specific phenotypes,* Proc.Natl. Acad.Sci.USA, 93, 12418-12422 (1996).
15. S. TIAN, R.A. HANEY, M.E. FEDER, *Phylogeny Disambiguates the Evolution of Heat-Shock cis-Regulatory Elements in Drosophila,* PLOS ONE 5 (5), 1-15, e10669 (2010).
16. V.Y. SHILOVA, D.G. GARBUZ, E.N. MYASYANKINA, B. CHEN, M.B. EVGEN'EV, M.E. FEDER, O.G. ZATSEPINA, *Remarkable site specificity of local transposition into the Hsp70 promoter of Drosophila melanogaster,* Genetics 173, 809-820 (2006).
17. C.M. BERGMAN, H. QUESNEVILLE, D. ANXOLABEHERE, M. ASHBURNER, *Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome*, Genome Biology 7 (11), R112 (2006).