

Article

Insights into the Evolution of a Snake Venom Multi-Gene Family from the Genomic Organization of *Echis ocellatus* SVMP Genes

Libia Sanz * and Juan J. Calvete *

Laboratorio de Venómica Estructural y Funcional, Instituto de Biomedicina de Valencia, Consejo Superior de Investigaciones Científicas, Jaime Roig 11, 46010 València, Spain

* Correspondence: libia.sanz@ibv.csic.es (L.S.); jcalvete@ibv.csic.es (J.J.C.); Tel.: +34-96-339-1760 (L.S.); +34-96-339-1778 (J.J.C.)

Academic Editors: Jay Fox and José María Gutiérrez

Received: 12 June 2016; Accepted: 6 July 2016; Published: 12 July 2016

Abstract: The molecular events underlying the evolution of the Snake Venom Metalloproteinase (SVMP) family from an A Disintegrin And Metalloproteinase (ADAM) ancestor remain poorly understood. Comparative genomics may provide decisive information to reconstruct the evolutionary history of this multi-locus toxin family. Here, we report the genomic organization of *Echis ocellatus* genes encoding SVMPs from the PII and PI classes. Comparisons between them and between these genes and the genomic structures of *Anolis carolinensis* ADAM28 and *E. ocellatus* PIII-SVMP EOC00089 suggest that insertions and deletions of intronic regions played key roles along the evolutionary pathway that shaped the current diversity within the multi-locus SVMP gene family. In particular, our data suggest that emergence of EOC00028-like PI-SVMP from an ancestral PII(e/d)-type SVMP involved splicing site mutations that abolished both the 3' splice AG acceptor site of intron 12* and the 5' splice GT donor site of intron 13*, and resulted in the intronization of exon 13* and the consequent destruction of the structural integrity of the PII-SVMP characteristic disintegrin domain.

Keywords: Snake venom toxin multi-gene family; snake venom metalloproteinase; genomic organization of SVMP genes; PII-SVMP; PI-SVMP; gene duplication; intronic retroelements; intronization

1. Introduction

The ADAM (A Disintegrin-like And Metalloproteinase) family of transmembrane type 1 proteins belongs to the MEROP database M12 family of Zn²⁺-dependent metalloendopeptidases [1] and PFAM family PF01421 [2]. Members of the ADAM family play important roles in cell signaling and in regulating cell-cell and cell-matrix interactions [3,4]. The ADAM family comprises ancient proteins whose origin extends back >750 My [5,6]. To date, close to 40 ADAM genes have been identified in vertebrate and invertebrate bilaterian animals, both in deuterostomes, from the basal chordate, *Ciona intestinalis*, to higher vertebrates, and in protostome, such as arthropods, nematodes, platyhelminths, rotifers, molluscs, and annelids. The evolutionary history of vertebrate ADAM genes is punctuated by gene duplication and retroposition events [7,8], followed by neo- or subfunctionalization [7]. Gene duplications are an essential source of genetic novelty that can lead to evolutionary innovation if the new function has no deleterious effects to its host organism or provides selective advantages. For example, in mammalian species, including marsupials and monotremes, except the platypus, ADAM28, ADAMDEC1 (decysin, a soluble ADAM-like protein), and ADAM7 form a cluster, likely as a result of tandem duplication of ADAM28 [9]. Instead, in most non-mammalian vertebrate genomes investigated, including those of aves, reptiles, and fishes, a single ADAM28 locus is present in this region [7,10]. The data suggest that ADAM7 and ADAMDEC1 were duplicated from

ADAM28, probably only in mammals [7]. On the other hand, as described below in more detail, it is thought that ADAM28 played a starring role in the emergence of toxic metalloproteinases in the superfamily Colubroidea of Caenophidian snakes (viperids, elapids, and colubrids).

The concept that gene duplication plays a major role in evolution has been around for over a century [11]. In his classic and influential book “*Evolution by Gene Duplication*” [12] Susumo Ohno argued that gene duplication is the most important evolutionary force since the emergence of the universal common ancestor. Common sources of gene duplications include ectopic homologous recombination, retrotransposition event, aneuploidy, polyploidy, and replication slippage [13]. Duplication creates genetic redundancy, where the second copy of the gene is often free from selective pressure. Thus, over generations of the organism, duplicate genes accumulate mutations faster than a functional single-copy gene, making it possible for one of the two copies to develop a new and different function. Duplicated genes may switch their transcription to other tissues by localizing closely to, and utilizing the regulatory elements of, a neighboring gene [14–16]. Examples of this are (i) the formation of toxin gene families during the evolution of the venom system of advanced snakes by co-option, multiplication, and weaponization in the venom gland of paralogs of genes encoding for normal body proteins [17–20], and (ii) the finding of 309 distinct widow spider genes exhibiting venom gland biased expression [21], suggesting that the switching of genes to venom gland expression in numerous unrelated gene families has been a dominant mode of evolution [21–23].

Because of its functional importance for prey capture, predator defense, and competitor deterrence, venom represented a key innovation that has underpinned the explosive radiation of toxicofuran reptiles in the Late Jurassic period of the Mesozoic era, ~150 million years before present (MYBP) [24–28]. Toxicofera [18] (Greek for “those who bear toxins”) is the term coined for the clade of squamate reptiles that includes the Serpentes (snakes), Anguimorpha (monitor lizards, gila monster, and alligator lizards) ,and Iguania (iguanas, agamas, and chameleons) lizards. One of the founding families of advanced snake venom comprises the Zn²⁺-dependent metalloendopeptidases (SVMPs) [17–19,29–32]. SVMPs are key enzymes contributing to toxicity of vipers and pitvipers venoms. Hemorrhage is one of the most significant effects in envenomings induced by viperid and crotalid snakebites. Damage to the microvasculature, induced by SVMPs, is the main event responsible for this effect. In addition to hemorrhagic activity, members of the SVMP family also have fibrin(ogen)olytic activity, act as prothrombin activators, activate blood coagulation factor X, possess apoptotic activity, inhibit platelet aggregation, are proinflammatory, and inactivate blood serine proteinase inhibitors [33–36].

The closest non-venom ancestors of SVMPs was likely an ADAM28 precursor gene [37]. The origin of SVMPs has been inferred to have occurred following the split of the Preatidae from the remaining Caenophidians, approximately 60 MYBP around the Cretaceous–Paleocene boundary of the Cenozoic Era [18,19,29,31,38]. SVMPs are found in the venoms of all advanced snakes and are classified into different classes depending upon their domain structure [39–41]. The ancestral multidomain PIII form, which is found in all snake venoms, derives from the extracellular region (metalloproteinase domain with disintegrin-like and cysteine-rich domains at the C-terminus) of a duplicated ADAM28 precursor gene that lost the C-terminal epidermal-growth-factor (EGF)-like, transmembrane, and cytoplasmic domains [31,32,41–43]. On the other hand, the derived PII-SVMPs, comprising the metalloproteinase and C-terminal disintegrin domain, have been only found in venoms of vipers and rattlesnakes (Viperidae). This strongly suggests that they emerged, subsequently to the separation of Viperidae and Elapidae, ~37 million years ago, in the Eocene epoch of the Cenozoic era, but before the separation of the Viperidae subfamilies Viperinae and Crotalinae 12–20 MYBP, from a duplicated PIII-SVMP gene that lost its cysteine-rich domain (see Figures 1 and 8 in [43] and Figure 18.1 in [44]). The disintegrin domain has been lost from the PII-SVMP structure on multiple occasions, resulting in the formation of the PI class of SVMPs [45] made only by the catalytic Zn²⁺-metalloproteinase domain [39–41].

Details on the mechanisms of co-option and the molecular events underlying the transformation of an ADAM28 precursor gene copy into the SVMP multi-gene family of extant snake venoms

remain elusive. In previous works, we described a family of RPTLN genes that exhibit a broad and reptile-specific distribution, for which we hypothesize may have played a key role in the recruitment and restricted expression of SVMP genes in the venom gland of Caenophidian snakes [46]. We have also reported the genomic organization of *Echis ocellatus* PIII-SVMP gene EOC00089, and compared it to those of its closest orthologs from *Homo sapiens* and the lizard, *Anolis carolinensis* [47]. Now, we fit two new pieces in the puzzle: the genomic structures of *E. ocellatus* PII—(EOC00006-like) and PI—(EOC00028-like) SVMP genes. Insights into post-duplication events gained from the structural comparison of the three classes of SVMP genes are discussed.

2. Results and Discussion

2.1. The Genomic Structure of Pre-Pro EOC00006-Like PII-SVMP and Pre-Pro EOC00028-Like Genes

Genomic sequences encoding full-length pre-pro EOC00006-like PII-SVMP (17828 nt) [KX219964] (Figure A1) and EOC00028-like PI-SVMP (21605 nt) [KX219965] (Figure A2) genes were assembled from overlapping PCR-amplified fragments (Appendix A, Figures A1 and A2). The pre-pro PII-SVMP gene consists of 15 exons interrupted by 14 introns (Figure 1A), whereas the pre-pro PI-SVMP gene contains 13 exons and 12 introns (Figure 1B).

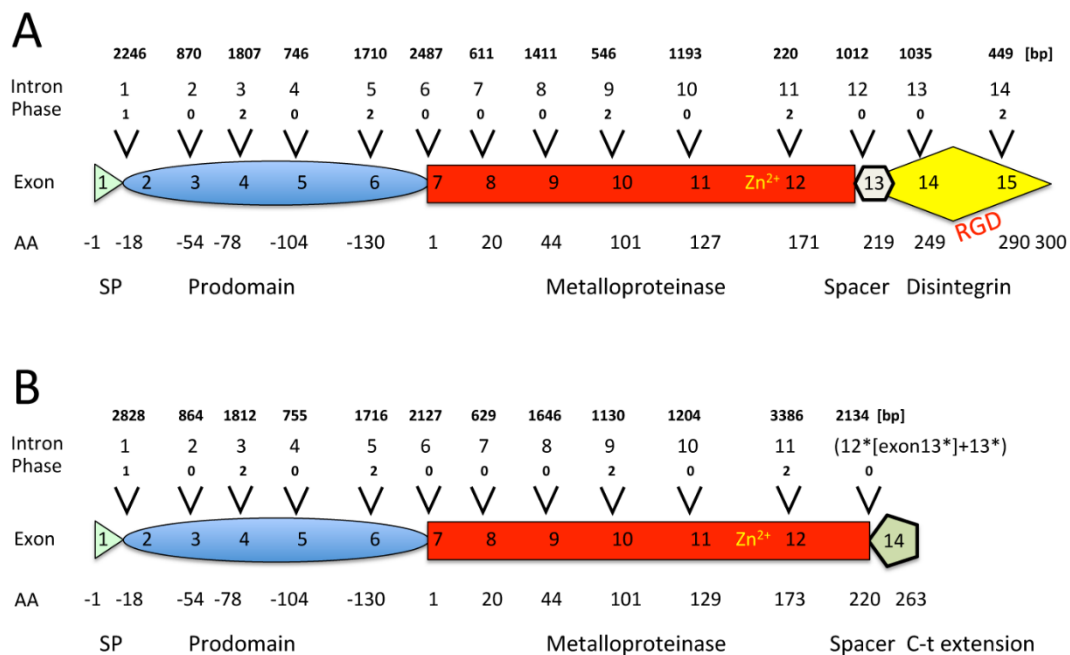


Figure 1. Scheme of the genomic organization of pre-pro EOC00006-like PII-SVMP (A) and pre-pro EOC00028-like PI-SVMP (B) genes. The distribution, phase, and size of the 14 (PII) and 12 (PI) introns and the boundaries of the protein-coding regions are highlighted. SP, signal peptide. Homologous exons and introns have identical numbering. Intron 12 of the PI-SVMP gene corresponds to the fusion of the genomic segment spanning intron12*-exon13*-intron13*. Mature PII- and PI-SVMP amino acid sequences span 299 and 263 amino acid residues, respectively. Zn²⁺, relative location of the catalytic Zn²⁺-binding environment; RGD, integrin-binding arginine-glycine-aspartic acid tripeptide motif.

The translated 494 (PII) and 457 (PI) pre-pro-SVMP amino acid sequences exhibit identical distribution and features (in terms of codon location and phase) for their first 11 introns and 12 exons, which code for the signal peptide (SP), prodomain (PD), metalloproteinase (MP) domain, and the short tetrapeptide (ELLQ) “spacer” sequence (Appendix A, Figures A1 and A2). These 413 (PII)/414 (PI) amino acid sequences show 85% identity, strongly suggesting that both SVMPs have a shared ancestry. It is also worth noting that the protein-coding positions interrupted by each of the introns

of the PII- and PI-SVMP genes are entirely conserved in *Anolis carolinensis* [XP_008118058] (and also in human [NG_029394] ADAM28 gene. Introns are inserted after or between secondary structure elements, supporting the “introns-added-late” model, which proposes that during the evolution of the eukaryotic branch, introns were added at the boundaries of structural modules coded for by ancestral continuous genes [48]. In addition, as will be analyzed in detail below, pairwise alignment of topologically equivalent PII- and PI-SVMP introns show that homologous intronic nucleic acid sequences share 88%–99% identity (Figure 2). This clearly indicates that EOC00006-like PII-SVMP and EOC00028-like PI-SVMP represent paralog genes.

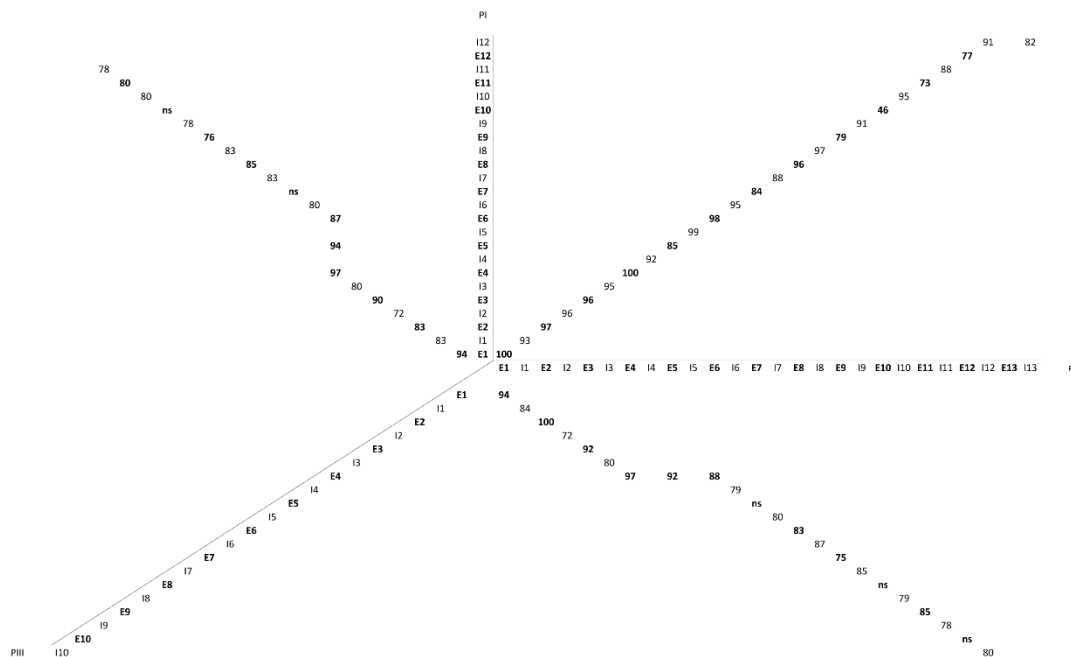


Figure 2. Pairwise comparisons of the sequence identities between the exonic and intronic nucleic acid sequences of pre-pro EOC00089-like PIII-SVMP, EOC00006-like PII-SVMP, and EOC00028-like PI-SVMP genes.

New genes can arise through four mechanisms: gene duplication, retroposition, horizontal gene transfer, and de novo origination from non-coding sequences [49]. Available evidence strongly suggests that gene duplication has played a pivotal role in the origin of venom multi-gene families [20–23,50,51]. Although the fate of many new genes may be to lose their function and become pseudogenes, some can be fixed through evolution of redundancy, subfunctionalization, or neofunctionalization. Several models have been proposed to explain functional divergence following venom toxin gene duplications [52–55]. However, this issue remains controversial and is the subject of vivid debates. The family portrait of SVMPs shows a complicated picture. SVMPs belong to different “generations”, that in the canonical model for the evolutionary expansion of this multi-gene family are hierarchically related, being PIII-SVMPs the most ancient and the PII- and PI-SVMPs the successively most recently derived family members [31,32,42]. However, due to the limited genomic information available, this model can be confounded by high rates of protein amino acid sequence divergence [56], and the occurrence of alternative routes (e.g., PIII > PI) can not be presently ruled out. The only other full-length viperid SVMP gene sequenced to date is *E. ocellatus* EOC00089-like PIII-SVMP [47] [KX219963]. The ORF encoding the pre-pro-metalloproteinase domains of this gene exhibits 63% amino acid sequence identity with the homologous coding regions of the PII- and PI-SVMPs here reported, and 72%–83% nucleotide sequence identity between topologically equivalent PIII-, PII-, and PI-SVMP introns (Figure 2). Although these figures clearly point to a common origin, it is not possible to infer whether they belong to the same or to a different PIII > PII > PI hierarchical lineage. Nonetheless, the

fact that the PIII-SVMP gene has lost introns 5 and 6 (ADAM28 numbering), with the consequence that exons 4, 5, and 6 have merged into a single exon, suggests that either these events occurred after the duplication that gave rise to the PII-SVMP ancestor, or that the PIII-SVMP EOC00089-like gene does not lay in the direct line of descent of the EOC00006-like PII-SVMP and EOC00028-like PI-SVMP genes. Refinement of the family tree of the multi-gene family of *E. ocellatus* SVMPs will surely emerge from future comparative genomic analysis of the carpet viper and other viperid species.

2.2. Role of Introns in the Evolution of the SVMP Multi-Gene Family

Since their discovery in 1977 [57,58], introns have been the subject of considerable debate. It is now generally accepted that introns represent more than merely junk DNA that must be pruned from pre-mRNAs to yield mature, functional mRNAs prior to their translation. Mounting evidence indicates that while introns do not encode protein products, they play essential roles in a wide range of gene expression regulatory functions such as non-sense mediated decay [59], mRNA export [60], and regulation of the amount of recombination between the flanking exons [61], or they serve as locations for nonhomologous recombination that would allow for exon shuffling [62,63]. As discussed below, most of the structural divergence between the EOC00006-like PII-SVMP and EOC00028-like PI-SVMP genes is due to the different size of their topologically equivalent eleven (1–11) introns (Supplementary Figure S1). The role of introns in the evolution of snake venom gene families remains elusive. However, in other biological systems, i.e., *Arabidopsis* and *Drosophila*, intron features, such as sequence and length, have been shown to function in maintaining pre-mRNA secondary structure, thus influencing temporal and spatial patterns of gene expression by modulating transcription efficiency and splicing accuracy [64–67].

Most PII- and PI-SVMP introns belong to phase 0, followed by phase 2; and, in both genes, only intron 1, separating the monoexonic signal peptide from the start of the prodomain, is a phase 1 intron (Figure 1). Analysis of the exon–intron structures of a large number of human genes has revealed a statistically highly significant enrichment of phase 1 introns flanking signal peptide cleavage sites [68]. Phase 1 introns most frequently split the four GGN codons encoding glycine. A plausible explanation for the correlation between signal peptide domains and the intron phase is that the base preferences of proto-splice sites [69,70] mirrors the amino acid preference for glycine in the signal peptidase consensus cleavage site [71].

The signal peptide is the most conserved structural element between pre-pro EOC00006-like PII-SVMP and EOC00028-like PI-SVMP is (Figure 2). In both genes, it is encoded by identical exon 1 amino acid sequences (Figures A1 and A2), which is also highly conserved in present-day SVMPs [46]. These findings support the view that co-option of this signal peptide may have played a role in the restricted expression of SVMP genes in the venom gland of Caenophidian snakes, some 60–50 Mya [46].

Nucleotide sequence comparison of the topologically equivalent introns of the *E. ocellatus* PII- and PI-SVMPs (Supplementary Figure S1) provide insights into the events underlying the conversion of a PII-SVMP into a PI-SVMP gene. In this regard, some introns differ in the number and location of intronic retroelements (Table 1). Thus, insertions in introns PI-SVMP 1 and 9 introduced complete and truncated SINE/Sauria elements in positions 1764–2101 (Figure S1, panel A) and 321–502 (Figure S1, panel I), respectively. The inserted nucleic acid sequence in intron 9 retains the GT-AG splicing sites, indicating that this insertion event created a twintron, an intron within an intron. PII-SVMP intron 6 (Figure S1, panel F) and PI-SVMP introns 11 (Figure S1, panel K) and 12 (Figure S1, panel L) are also twintrons. Compared to its topologically equivalent PII-SVMP intron, a large insertion in intron 11 of the EOC00028-like PI-SVMP gene replaced the first 66 nucleotides for a longer stretch of 3281 nucleotides; region 2461–2561 of the inserted nucleic acid sequence is 97% identical to *Hyla tsinlingensis* Hts-35 [KP204922], a microsatellite sequence that is also partly present in intron 61 of *Podarcis reelin* (RELN) genes [GU181006-13] (positions 554–623) [72]. Microsatellites are simple nucleotide sequence repeats (SSR) ranging in length from two to five base pairs that are tandemly repeated, typically 5–50 times (reviewed in [73]). These non-coding elements are abundant in major

lineages of vertebrates. Mammalian, fish, and squamate reptile genomes appear to be relatively microsatellite rich [74]. However, besides Hts35, RepeatMasker only identified few SSR tracks in introns 1 (5× GTTT; 28× TC) and 2 (13× ATTT; 4× TAA) of the PII-SVMP gene (Figure A1), and introns 1 (11× GTTT; 21× AG) and 2 (9× GTTT; 4× TAA) of the PI-SVMP gene (Figure A2).

Table 1. Comparison of type and location of retroelements identified in introns of *E. ocellatus* PII-SVMP EOC00006-like and PI-SVMP EOC00028-like genes.

Intron	PII-SVMP	PI-SVMP
	Inserted Retroelement	
1	SINE/Sauria	2 SINE/Sauria, LTR/ERV1, DNA/hAT-Ac
3	LINE/L2/CR1	LINE/L2/CR1
5	LINE/L2/CR1	LINE/L2/CR1
6	SINE/Sauria	-
8	LINE/L2/CR1	-
9	-	SINE/Sauria
10	DNA transposon	DNA transposon

Growing evidence supports that repetitive intronic elements, such as the long interspersed elements (LINEs) and the short interspersed elements (SINEs) contained in several introns of both PII- and PI-SVMP genes (Table 1) can influence genome stability and gene expression (reviewed in [75]). Thus, these interspersed repeats may alter genome recombination structure and rates, through a number of mechanisms, including replication slippage and unequal crossover [76,77], potentially impacting regulation of gene expression [78], recombination events leading to tandem duplication of segments of the genome [79,80], gene conversion [81], and chromosomal organization [79]. Moreover, the insertion of interspersed repeats into a new genomic position may introduce promoter or enhancer sequence motifs for transcription of nearby genes [82,83], and alternative splicing sites or polyadenylation sites [84], thereby resulting in a change of overall level of gene expression. Interspersed repetitive elements have also played an important role in expanding the repertoire of transcription factor binding sites in eukaryotic genomes [85]. However, whether these elements have contributed to the genomic context that facilitated the evolution and radiation of venom *loci* in snakes deserves future detailed comparative genomic studies.

2.3. A Fusion Event Led to the Conversion of a PII(e/d)-Type SVMP into EOC00028-like PI-SVMP

PI-SVMP intron 12 is a twintron resulting from the fusion of the genomic region spanning ancestral introns 12* and 13* and exon 13* (homologous to identical numbered elements in the genomic structure PII) (Figures 1 and 3A). Splicing site mutations affecting both the 3' splice AG acceptor site of intron 12* and the 5' splice GT donor site of intron 13* led to the retention, and subsequent intronization, of exon 13* within a fused (12* + 13*) twintron (Figure 3A). Intronization of exon-coding nucleic acid sequences has been proposed as a major contributor to intron creation [86]. Intron 13* encoded part of the N-terminal region of a disintegrin domain, most likely, as discussed below, an eventual subunit of dimeric disintegrin. In addition to the disruption of the structural integrity of the disintegrin domain, a stop codon after exon 14 removed intron 14 and exon 15 from the PII(e/d)-type SVMP (Fox & Serrano's nomenclature [40]) precursor gene structure, thereby completing the conversion of the PII-SVMP into present EOC00028-like PI-SVMP gene (Figure 3A).

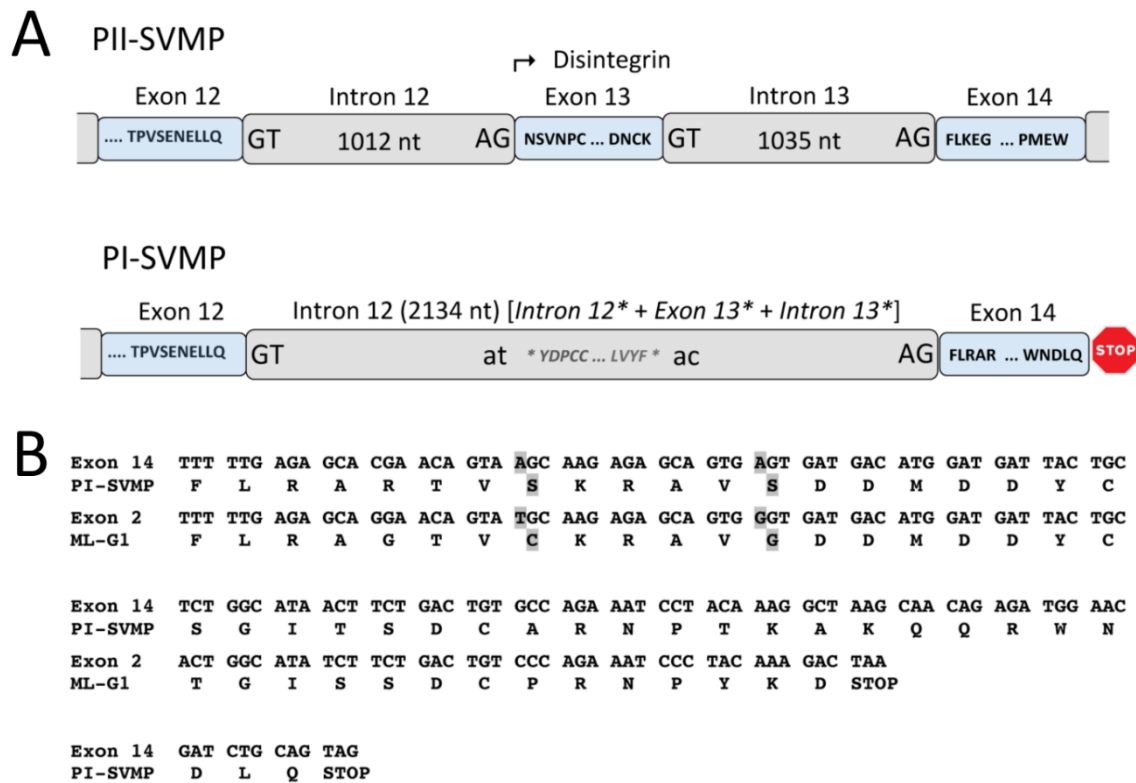


Figure 3. Panel A, cartoon comparing the 3' regions of the PII-SVMP and PI-SVMP genes and highlighting the processes (intronic fusion of ancestral exon 13* inside twintron 12 resulting from the fusion of introns 12* and 13*, and creation of a stop codon after exon 14) that destroyed the integrity of the disintegrin domain, converting an ancestral PII(e/d)-type SVMP into extant EOC00028-like PI-SVMP. Panel B, alignment of the amino acid sequences encoded by exon 14 of EOC00028-like PI-SVMP and exon 2 of the dimeric disintegrin subunit ML-G1 [AM261811] [87]. Degeneration of PI-SVMP's conserved functional and structural amino acid residues in dimeric disintegrins are highlighted in boldface and grey background.

Region 1013–2134 of PI-SVMP intron 12 exhibits 91% nucleotide sequence identity with range 14 to 1135 of *Macrovipera lebetina* gene encoding part of exon 1 and full-length intron 1 of the VGD-containing dimeric disintegrin subunit precursor, ML-G1 [AM261811] [87]. PI-SVMP exon 14 (mature protein amino acid residues 221–263, Figure A2) exhibits strong homology (79% identity) to exon 2 of the same VGD-bearing dimeric disintegrin subunit. The PI-SVMP exon 14 shows the consequences of genetic drift (Figure 3B): the conserved $\alpha_5\beta_1$ integrin-inhibitory VGD tripeptide motif [44] of the PII-SVMP precursor gene has been replaced by a VSD motif (generated by a G > A mutation: GTG AGT GAT > GTG GGT GAT), and the absolutely conserved tenth cysteine residue of dimeric disintegrin subunits has degenerated (TGC) to a serine residue (AGC) (Figure 3B).

3. Concluding Remarks and Perspectives

The event that gave birth to the family of SVMPs was the generation of a STOP codon at the 3' end of exon 16 of a duplicated ADAM28 gene (Figure 4). This mutation produced an ORF truncated at the N-terminal part of the EGF-like domain, which encoded a precursor of an ancestral PIII-SVMP lacking this domain and the C-terminal membrane anchoring and cytoplasmic polypeptides (Figure 4). On the other hand, our results comparing the available genomic structures of SVMP genes, e.g., EOC00089-like PIII-SVMP [47] [KX219963], EOC00006-like PII-SVMP [KX219964], and EOC00028-like PI-SVMP [KX219965] (this work), suggest that the evolutionary history of SVMPs is marked with events of insertions and deletions of intronic regions. This scenario points to introns as key players in

the formation of the multi-locus SVMP gene multifamily. Thus, comparison of the genomic structures of EOC00089-like PIII-SVMP and EOC00006-like PII-SVMP (Figure 5) indicates that replacement of the PIII-specific cysteine-rich domain by a non-homologous region encoding intron 14-exon 15 followed by a STOP codon may represent a step in the conversion of a PIII-SVMP into a PII-SVMP.

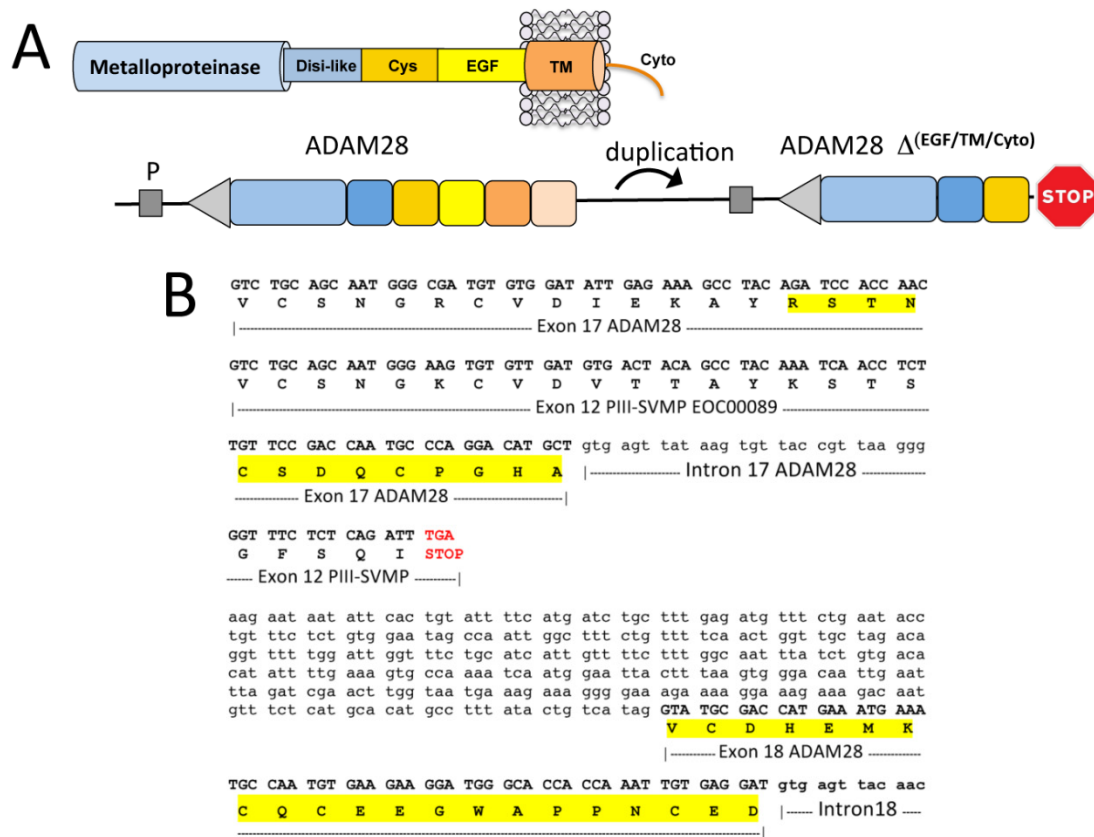


Figure 4. Comparison of the genomic region encompassing exons 17 through 18 of *Anolis carolinensis* ADAM28 [XP_003226913] and the homologous amino acid sequence of *E. ocellatus* SVMP EOC00089 [ADW54351], highlighting the STOP codon after exon 12 of the latter generating a C-terminally truncated molecule, which eventually gave rise to the ancestor of the PIII-SVMPs.

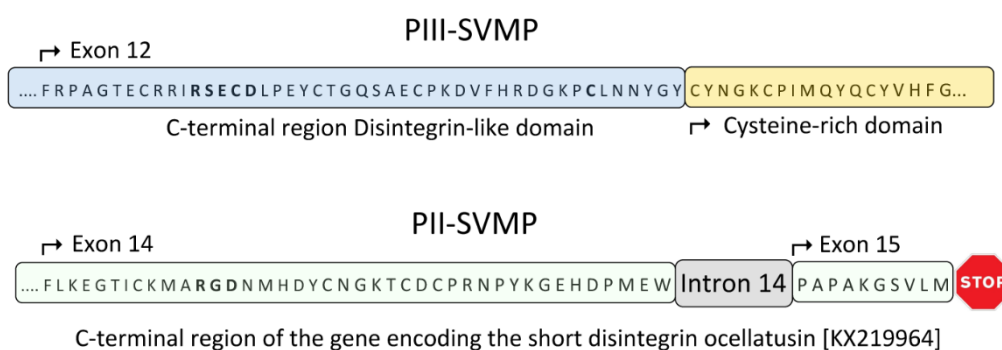


Figure 5. Comparison of the genomic region encoding the C-terminal domains of *E. ocellatus* EOC00089-like PIII-SVMP and EOC00006-like PII-SVMP, suggesting that 3' genomic remodeling represents a seminal step in the generation of PII-SVMPs.

This view is consistent with structural evidence suggesting that the loss of the cysteine-rich domain represents an early seminal event that facilitated the formation of PII class SVMPs [43].

The PII subfamily of SVMPs is characterized by the diversity of disintegrin domains exhibited by different family members [39,40], ranging from the more ancestral long disintegrin domains (~84 amino-acid-residue polypeptide cross-linked by 7 disulfide linkages) to the more recently evolved short disintegrin (41–51 amino-acid-residues crosslinked by 4 disulfide bonds) [42]; for a scheme of the evolutionary path of the disintegrin domains, see Figure 1 in [43]. EOC00006-like is an example of a PII-SVMP with short disintegrin domain. Given the structural diversity of PII-SVMPs, genomic sequences from the different members of the subfamily are required for a more accurate glimpse of the genomic mechanisms operating in the generation and subsequent diversification of PII-SVMPs.

Comparison of the EOC00006-like PII-SVMP and EOC00028-like PI-SVMP gene structures also points to genomic remodeling of the 3' region of a PII(e/d)-type SVMP precursor gene [39,40] as the EOC00028-like PI-SVMP gene generator mechanism. The PII > PI conversion involved the generation of twintron 12 (by fusion of introns 12* and 13*) and the loss, by intronization, of exon 13*, thereby destroying the consistency of the region coding for the disintegrin domain. This elaborated mechanism indicates that the structural diversification of SVMPs is not due to a random mutation generating a STOP codon before the disintegrin domain, but follows a well orchestrated sequence of events imprinted in the genome of snake species sometime after the split of Viperidae and Elapidae, 37 million years ago, but before the separation of the Viperidae subfamilies Viperinae and Crotalinae 12–20 MYBP. The mechanisms underlying loss or gain of spliceosomal introns are still poorly understood. The most widely accepted hypothesis is that intron insertion may occur via a process similar to group II intron retrotransposition [88,89]. According to this view, the spliceosomal components remain transiently associated with a recently excised intron and then attach at a potential splice site of a non-homologous pre-mRNA, where they catalyze the reverse reaction [90,91]. The modified pre-mRNA is reverse-transcribed and the resulting cDNA participates in a recombination with its parent gene, thereby inserting a novel intron into the target gene [90–93]. An attractive feature of this mechanism is that it ensures that the inserted nucleic acid sequence has the full complement of intron signature sequences required for efficient splicing [94].

Studies of multi-gene protein families are crucial for understanding the role of gene duplication and genomic exon-intron organization in generating protein diversity. For example, full-length genomic sequences of Crotalinae group II PLA₂ isogenes from *P. flavoviridis* (Tokunoshima and Amami-Oshima islands, Japan) [95], and *T. gramineus* (Taiwan) [96] have been reported. All these genes exhibit four coding regions and conserved exon-intron structures spanning about 1.9 kb. A cluster of five tandemly arranged PLA₂ genes have been located in a 25 kb 3' segment of a 31 kb fragment of the Amami-Oshima *P. flavoviridis* genome [97], which in addition harbors a PLA₂ pseudogene in its 6 kb 5' region [98]. Genomic sequence comparisons between the pancreatic PLA₂ gene of *P. elegans*, group IB pancreatic PLA₂ gene of *L. semifasciata*, and the *L. semifasciata* group IA venom PLA₂ gene, suggest that Crotalinae group II venom PLA₂ genes emerged before the divergence of Elapinae and Crotalinae, whereas groups of IB and IA PLA₂ genes appeared after Elapinae was established as a taxonomic lineage [99].

Duplicated structures found in eukaryotic genomes may result from complex interplays between different mechanisms [100]. Mitotic and meiotic non-allelic homologous recombination (NAHR) events, resolved as unequal crossing-over, have been traditionally invoked to account for segmental duplications within genomes [101,102]. Duplicated regions can be organized as direct tandems (e.g., the cluster of tandem snake venom PLA₂ genes), but also be separated by hundreds of kb [100]. Our present and previous work [47] inaugurate a line of research that will allow the depiction of a more precise characterization of the genomic context in which the SVMP multi-gene family has emerged. This goal demands populating the current databases with genomic sequences of genes representing the different members of the SVMPs. Although the variety of structural forms comprising the PII family may be considered a challenge for this purpose, this circumstance can be also regarded as a valuable opportunity for the step-by-step description of the molecular pathways that led to the formation of this multi-gene family. Without a doubt, ongoing Viperidae snake genome sequencing projects will mark

the beginning of comparative snake genomics, and will be key to revealing not only the topology and copy number of the genes encoding SVMPS, but also to provide decisive information to reconstruct the evolutionary history of this multilocus gene family.

4. Materials and Methods

4.1. Genomic DNA

Genomic DNA was extracted from the fresh liver of *E. ocellatus* (Kaltungo, Nigeria) maintained at the herpetarium of the Liverpool School of Tropical Medicine. *Echis ocellatus* liver was ground to a fine powder under liquid nitrogen and the genomic DNA extracted using a Roche DNA isolation kit for cells and tissue containing SDS (2% final concentration) and proteinase K (400 µg/mL final concentration). The homogenates were incubated at 55 °C overnight. Thereafter, 300 µL of 6 M NaCl (NaCl-saturated H₂O) was added to each sample, and the mixture was vortexed for 30 s at maximum speed and centrifuged for 30 min at 10,000 g. An equal volume of isopropanol was added to each supernatant, and the sample mixed, incubated at −20 °C for 1 h, and centrifuged for 20 min at 4 °C and 10,000 g. The resulting pellets were washed with 70% ethanol, dried, and, finally, resuspended in 300–500 µL sterile distilled H₂O.

4.2. Strategy for PCR Amplification of Overlapping Genomic DNA Fragments

For sequencing *E. ocellatus* genes encoding PII-SVMP EOC00006 [Q14FJ4] and PI-SVMP EOC00028 [Q2UXQ3] we employed a similar iterative process as described in [47]. Full-length cDNA-deduced amino acid sequences of disintegrin domains [103] and of the genomic organization of dimeric disintegrin domains [AM286800] [87] and PIII-SVMP EOC00089 [47] from the same species were used as templates to design primers for the PCR-amplification of protein-specific genomic sequences (Table 2).

PI-SVMP stretch ⁷²AREILNS....QRWNDLQ²⁶³ was amplified on an Eppendorf Mastercycler[®] epgradient S instrument in a 50 µL reaction mixture containing 17.5 µL of H₂O, 25 µL Master-Mix (Thermo Scientific, Waltham, MA USA) including buffer, dNTPs, and Phusion High-Fidelity DNA polymerase, 2.5 µL of each primer (10 µM) Met1PIRv and Met5PIFw, 1.5 µL of DMSO (100%), and 1 µL of genomic DNA (50 ng/µL). PCR conditions included an initial denaturation step at 98 °C for 30 s followed by 35 cycles of denaturation (20 s at 98 °C), annealing (15 s at 63 °C), extension (300 s at 72 °C), and a final extension for 5 min at 72 °C. All other PCR amplifications were carried out in the same thermocycler using iProof High Fidelity polymerase (BioRad, Hercules, CA, USA). The 50 µL reaction mixture contained 10 µL of 5×buffer, 1 µL of 10 mM (each) dNTPs, 2 µL of MgCl₂ 50 mM, 1.5 µL of DMSO (100%), 1 µL of each Fw and Rv primer (10 µM), 1 µL of genomic DNA (50 ng/µL), and 32.5 µL of water. PCR conditions included an initial denaturation step at 98 °C for 120 s followed by 35 cycles of denaturation (10 s at 98 °C), annealing (15 s at the lower melting temperature of the primers), extension (60 s per Kb at 72 °C), and a final extension for 5 min at 72 °C.

4.3. Purification and Cloning of PCR Products

PCR-amplified DNA fragments were purified from agarose electrophoretic bands using the GENECLEAN Turbo kit (MP Biomedicals). The purified fragments were inserted into pJET_1.2 (Thermo Scientific, Waltham, MA USA) using phage T4 ligase and cloned into *E. coli* DH5α by electroporation at 1700 V. Transformed cells, resuspended in 200 µL LB medium, were incubated at 37 °C for 1 h, and were subsequently plated on LB agar/ampicilline to select positive clones. The presence of the inserted DNA fragments was verified by PCR amplification or digestion of the expression vector with the restriction enzyme Bgl II. The inserted DNA fragments were sequenced in-house on an Applied Biosystems model 377 DNA sequencing system (Foster City, CA, USA) using pJETFw and pJETRv primers.

Table 2. Forward (Fw) and reverse (Rv) primers used to PCR-amplify genomic DNA stretches from *E. ocellatus* PII-SVMP EOC00006-like (right) and PI-SVMP EOC00028-like (left) genes.

Primer	DNA sequence	Primer	DNA sequence
Sp35_Eo Fw	ATGATCCAAGTTCCTCTGGTAACTATATGCTTAGC	5' PS-Disi Fw	ATGATCCAAGTTCCTCTGG
Met14PI Fw	CTATATGCTTAGCAGTTTTTCCATATC	Intr4 Fw	ATGACACTGACCTCTAGAGTTGG
Intr1F1PI Fw	CTAGTCATTCCGGCCATATGAC	IntrB9_4-2 Fw	AAGCTTGCTTGCTAGTAGTTGG
Intr2F1PI Fw	ATCAGTCTGAGAGGATGCATTCC	Intr4 Rv	TGGACATTGTATGGTCCACCTG
Intr3F1PI Fw	GTGACCATGCAATGTCCATATG	Prodom 3 Fw	GGAGCTTTAAGCAGCCAGAG
Met15PI Fw	GTTCCTGTAGGAGCTGTAAAG	Prodom 3 Rv	CTCTGGCTGCTTAAAAGCTCC
Prodom 2 Fw	GACGCTGTGCAATATGAATTTG	Prodom 2 Fw	GACGCTGTGCAATATGAATTTG
Prodom 2 Rv	CAAATTCATATTGCACAGCGTC	Prodom 2 Rv	CAAATTCATATTGCACAGCGTC
Intr3 Rv	GCACCAACTCTGTATCTCAGTC	Intr3 Fw	CACAGGTAATAAGCCACAAACACC
Pro2 Fw	CAGTGAGACTCATTATCCCTGATGGCAG	Intr3 Rv	GCACCAACTCTGTATCTCAGTC
Pro3 Rv	CTGCCATCAGGGGAATAATGAGTCTCACTG	Pro2-SVMP_Fw	CAGAAGATTACAGTGAGACTCATTA TTCCCWGATGG
IntrB13-1 Fw	CTTGCCTCCCTATAGGATCACTGC	Pro3-SVMP_Rv	CTGCCATCAGGGGAATAATGAGTCTCACT
Met16PI Rv	GATGCGTCCATAATAATAGCAGTG	IntrB13-1 Fw	CTTGCCTCCCTATAGGATCACTGC
Prodom 1 Fw	GATGCCAAAAAAGGATGAGG	Prodom 1 Fw	GATGCCAAAAAAGGATGAGG
Prodom 1 Rv	CCTCATCCTTTTTTTGGCATC	Prodom 1 Rv	CCTCATCCTTTTTTTGGCATC
IntronB7PI Fw	TGGAACAACAGCTGTTGTTATGACG	Intr2 Fw	ACAATGGGAAACTGAGGAACAG
IntronB7PI Rv	TGAGAGACATGCTGATGIGGTC	Intr2 Rv	GGGAACTCTGACTTAGAGAAAGTC
Met4 PI Fw	GACCCAAGATAACATTCAGCTTGTG	Met1PII Fw	CAACAGCATTTTCACCCAAGATAC
Met4 PI Rv	GACAAGCTGAATGTATCTTGGGTC	Met1PII Rv	GTATCTGGGTGAAAATGCTGTTG
Met8PI Rv	TATCCATGTTGTTATAGCAGTTAAATC	Met 1-2 Fw	CATGGATACATCAAATTGCAACG
Intron B16 Fw	TGTGCTTACCCAACACTGAGCC	Met 1-3 Rv	TGTACATCTGTCTAGGTGGACATG
Met5 PI Fw	GCACGTGAAAATTTGAACCTA	Met2PII Fw	GCCGTTACCTTGATAACCTTATAGG
Met5PI Rv	GAGTTCAAAATTTACGCTGCTG	Met2PII Rv	CCTATAAGGTTATCAAGTGAACGGC
Met9PI Rv	AGCATTATCATGCGTTATGCG	Met 6 PII Fw	CCACAATCGTCTGTAGCAATTACTGA
Met3 PI Fw	GGAAGAGCTGTACATGGAGAG	Met 6 PII Rv	TCAGTAATTGCTACAGACGATTGTGG
Met3PI Rv	CTCTCCATGTAAGCTCTTCC	Met3 PII Fw	GATCATAGCACAGATCATCTTTGG
Met2PI Rv	GCTCCCCAGACATAACGCATC	Met3PII Rv	CCAAAGATGATCTGTGCTATGATC
IntrB23PI Fw	CTGACTATGACTCACTTAACAACCTGG	Met 4 Fw	ATGATCCAGGTTCTTTGGTAACTATATG
IntrF2PI Fw	GGCCGCTGAATGCATCTGCTTC	Met 4 Rv	TGAACTGATAGGAACGGTATTGTG
Intr2F2PI Fw	GCATCAGTTTGTTCGCACTCAATAAAG	Ww_Ocella NcoI	ATCCATGGTAGACTGTAATCTGGACC
Intr3F2PI Fw	GAGCATAATCTGGAACTAAGATCAAG	IntrDis1 Rv	ATACGGCTAGTATGGAGCAGG
Met7PI Fw	GCACAAGATTCTATCACTTCAG	Dis PII Rv	TCACATCAACACACTGCCTTTTGG
Met13PI Rv	TCCTACCTGCAAAAGTTCATTTTC	-	-
Intron B10PI Rv	CTGACTCAGGGCACCAATCTC	-	-
Met1PI Rv	CTACTGCAGATCGTTCCATCTCTG	-	-

4.4. Sequence Analysis

Exon-intron boundaries were localized by visual inspection and corroborated using Wise2 [104]. Amino acid and nucleotide sequence similarity searches were done using BLAST [105]. Multiple sequence alignments were performed using ClustalW2 [106]. The occurrence of retrotransposable elements and simple nucleotide sequence repeats (SSRs) were assessed using RepeatMasker (version rm-20110920) [107], a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences included in the Repbase database [108].

4.5. Sequence Availability

Pre-pro EOC00006-like PII-SVMP and EOC00028-like PI-SVMP gene sequences have been deposited with the NCBI GeneBank [109] and are accessible under accession codes KX219964 and KX219965, respectively.

Supplementary Materials: The following are available online at www.mdpi.com/2072-6651/8/7/216/s1, Figure S1: Pairwise nucleotide sequence alignments of topologically equivalent paralog introns 1–12 from Pre-pro EOC00006-like PII-SVMP and 1–13 from Pre-pro EOC00028-like PI-SVMP gene sequences.

Acknowledgments: This work has been financed by Grant BFU2013-42833-P from the Ministerio de Economía y Competitividad, Madrid (Spain).

Author Contributions: J.J.C. and L.S. conceived and designed the experiments; L.S. performed the experiments; J.J.C. and L.S. analyzed the data; J.J.C. wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Genomic sequence of *E. ocellatus* EOC00006-like PII-SVMP gene. The locations and identities of the primers used to PCR-amplify genomic sequences (listed in Table 2) are indicated. Protein-coding DNA regions are in upper letters and boldface, and the encoded amino acid sequence is displayed below the DNA sequence. Start of introns are labelled EoPII-X, where “X” corresponds to intron number. The beginning and the signal peptide, propeptide, metalloproteinase and the short-disintegrin domains are specified. Numbers at the right correspond to amino acid numbering of the DNA-deduced pre-pro-PII-SVMP relative to the mature SVMP. The N-terminal glutamine of the metalloproteinase domain has been assigned residue 1. The extended Zn²⁺-binding environment (HEXXHXXGXXH) and the RGD integrin inhibitory motif stand on yellow background. The only two amino acids (–70I/V, and –111H/R) that distinguish this sequence from that of PII-SVMP EOC00006 (Q14FJ4) are shown in bold and red. The remains of a disintegrin-like domain transformed into intron EoPII-12 are underscored in italics and on cyan background. SINE/Sauria, LINE/L2/CR1 and DNA transposon retroelements are highlighted on a gray background. Simple sequence repeats (SSR, microsatellites) are shown in light green background.

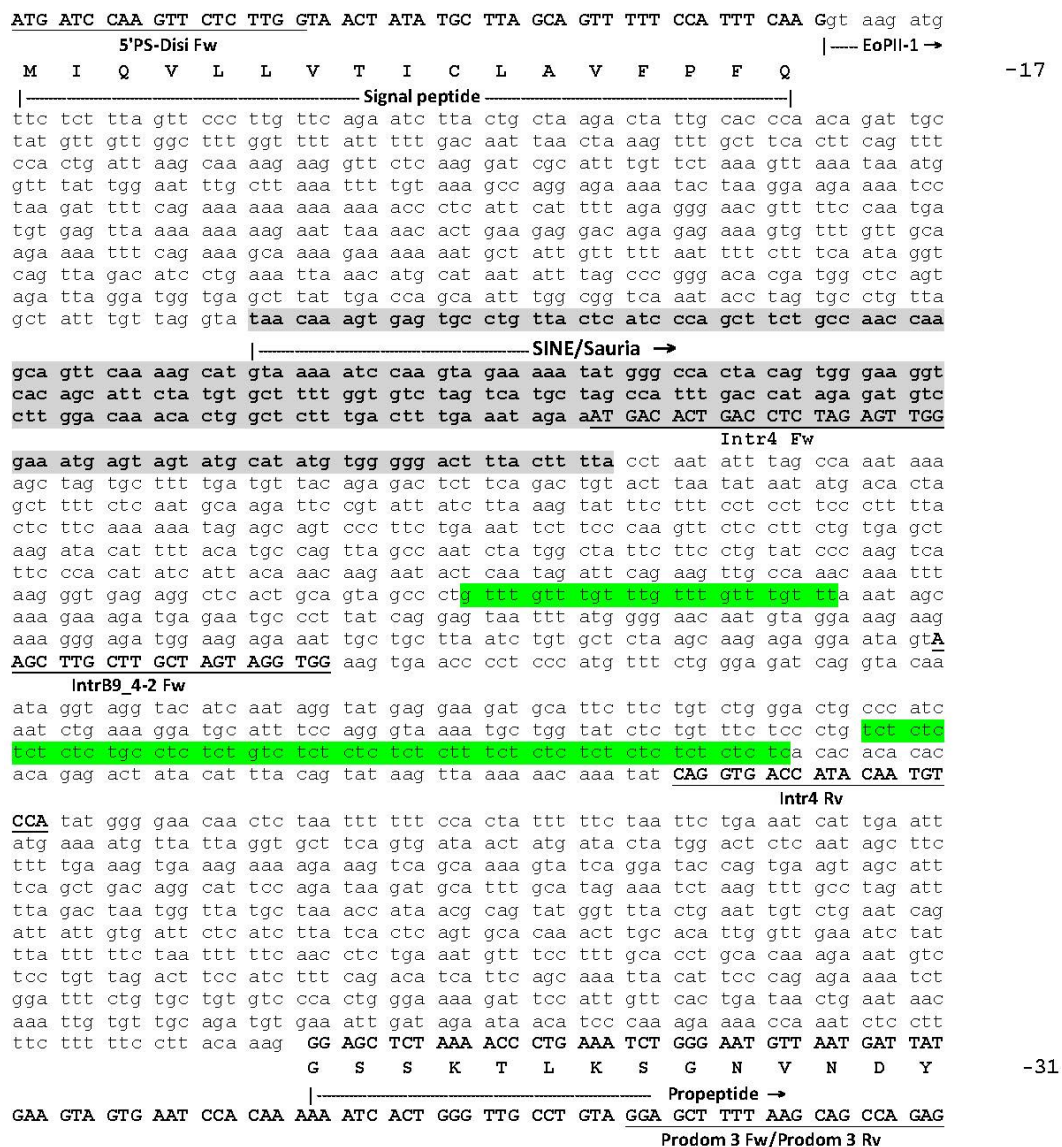


Figure A1. Cont.

E V V N P Q K I T G L P V G A F K Q P E -51
AAA AAG gta aga tat ttc ttt cag caa caa att att ttt gtc agc cca tag aaa gtt tga
K K |-----**EoPII-2**→ -53
 tat tcc ttt cct gct att taa tgg tta ttt gga ttt tgc att gca atc tat gtt cct gtt
 tga aat **att tat tta ttt att tat tta ttt att tat tta ttt att tat tta ttt att tag**
 caa att tac act gcc tcc cag ttt aca caa aat tga gga taa tcc tgg caa ttt ata aac
 aga att tcc taa tat taa tag gtt aat ata acc att agt cat ttc ata ctg caa cag atc
 tac agt gaa gag gaa tca agc agt gaa ata acc att agt cat ttc ata ctg caa cag atc
 cag caa aga cat cct ctc cct cct cca gga aat ccc att aaa gtt tat ctg atc tgc aga
 gaa tgg aag cag tct agg gag agt cct gta tta cag gca aaa aat aat aaa ctg ctt aat
 tta aac ttt gta tat cta cat gcg tca cat aga aat ata aaa gat ata ttt cta gat atg
 ata gta tag tgg tta cag ctc aag gtt agt tct ata tac cag att cat gta tag cgt ggg
 atg act ttt gag tca atc tgt ctg tcc cag atc aac ctc att cag caa tat gtt cag tta
 gag aat gag aat tta aac cta tgc agg aaa aat aaa taa aga tct tat cat tca aag cac
 cag gtg aaa tac tct **aat aat aat aat** cta ata aat aat atg act atc aca aat tat tca
 tta gat tta aat taa tac aga tgt aga gta tta aga aat gac aca ttt tat ttg cct aaa
 ttt tga gaa gta aaa cat aaa ctc ttt gtt ttt cag **TAT GAA GAC GCT GTG CAA TAT GAA**

-61

Prodrom 2 Fw/Prodrom 2 Rv

TTT GAA GTG AAT GGA GAG CCA GTG ATC CTT CAT CTG GAA AAA AAT AA g tat gtt aac
F E V N G E P V I L H L E K N K |-----**EoPII-3**→ -77
 tca gaa att ttc tta act tta cta aac aat gtg gaa aat gta ttt tcc tgg aca caa tct
 gag aga aat aat aat tgc att cct ttg ttt gct ctt aaa tta aat gtt act ata tag gaa
 aag tgg ata tgt taa tta tgt ctt acc ttg gaa ctg aaa ttt ttt tac tgc tgt tct cct
 atg gga cac tgc tga act ata aca att tta gac tca gtt gaa cca ttc aag gct cat aaa
 cgt cta tca ctt gaa tta act agg tta gtt atg aaa tat tgc aga gat ttc aaa atg ttt
 ctt tat gct gct tca tag aca ctc cat caa tct gaa taa cat ttt ctg tga gac ctg tag
 agc tgc tca atg cac atc aaa att tat atg tta aat cat gaa agg cag taa atg tca tct
 cct att aag gcc cga aag ggg aca gga cat aaa taa gcc cag att tca caa gcc ttt ttt
 aaa aga atg cct tgt gag tct tct cag aat tgt tcc att tga act gta aat caa gat att
 aat aaa aat att ata caa ata tga acc aca aaa atg gaa tga aat tat ttt tca gca caa
 aag aca **aac ata gtt gga agg gac ctt gga ggc ctt cta gtc cac ccc ctg ctg aag cag**

|-----**LINE/L2/CR1**→

gag act ata tca ttc cat ggg tgt cca att gtt cct tga aaa cct cca gtg atg gga tac
cca cag ctt ctg aaa gct agc cat tcc acc gat taa ttg ttc tca ttc tca gga aat ttc
ttt tag ttc tag gtt gaa ttt ttc ACA GGT AAA TAA GCC ACA AAC ACC att aaa aga aag

Intr3 Fw

tgt gcc att ttc att acc cta aaa act gta aat gtc ata aca cag ttt aaa tta gtc tgg
 cag tca cta aac tgc aaa cac gta cct cag cct ggt tca gtt taa tac tat agt taa **GAC**
TGA GAT ACA GAG TTG GTG Cat tgg gaa aga ctg cct gcc ttt gat ttt agc atg tat ctt

Intr3 Rv

ggt ttt gat cat tct gca tgt acc ttt aaa agt ata ttg tct tta tta tta aac ata tgc
 tac aag aaa ttg cac tat atg tta gta tct gtg tgt cac ttc ctc cta ttc cac cgg aga
 tat ttc atg cta aat tct aat gtg tgc ttt aga cgt tct agt gtc ccc ttt tgt ttt gtt
 taa ata tga agg cta tac aga aat tca agg gta gca tat gga tgg tct tct ttt gtc ata
 gaa gca aag cag ggc tac agg ggg aag acc agc aag gat ttc tag att ggc aaa aat gga
 aga cga tga caa gtt tat ttt tct gac ttc agg att gga atg tga ctc cat ata ctg gtt
 taa agc ttc aaa ata tat ctt tct gta tga atg ttg aag gca cac att tcc ttc agg ttg
 ttt cat ttt acc ctt tag taa ata gaa cta tac cgt ttt cct tta gag aca aag cag ttt
 cat cag att gtt gca tag aga aga aga ata aaa gga aga cgc aat aca aac tgt cct ctc
 ttt gtc cat cac aat cct ctg agc att gat gaa gag aat tgc aca gct tca ttt ggt gaa
 tct agt atg cac tgc ctt cct ctt aaa aag cca tgg gat aga aat acc tgc tcc att ctt
 ttg ata cca aag aag ata ctt aaa ttg oca tta tac ggt tgt ttt aaa ctt ctt ata aaa
 tta tct ctg tgc tac ata att cct gat ata gat atc ctt tct ttg cat tct ttc cag **A**
GGA CTT TTT TCA GAA GAT TAC AGT GAG ACT CAT TAT TCC CCT GAT GGC AGC GAA ATT ACA

Pro2-SVMP_Fw/Pro3-SVMP_Rv

G L F S E D Y S E T H Y S P D G S E I T -97
ACA AAC CCT CCT GTT GAG gta ggg tct cac ttt tat gag cct ttt ttt tag gaa gta aac
T N P P V E |-----**EoPII-4**→ -103
 tga aac aaa tgt ttg tgc aca ata tta caa ata tac aag aat gag acc agg cta ctc aaa
 caa agt gta tat atg tat gaa gta ttt tat att gat atg tac gta caa gga tgc ctg gat
 tgt taa acc ctg gtt aaa agc caa cat att ttg gag gtg agt ttc aca aat aga ttt att
 atg aga aca tca ggt ttg taa gat tat att ttc att ttt aaa cca gac tac agg gat aaa
 tgc aaa gtc ttt tat aaa tga taa caa ttc act ttg ctc cta tac aga aat cca ttt aat
 atc ttt cat aat aaa atg gtg cca aaa atg gct cta tca gat gta aaa caa tta gag cac
 taa cat gct tca tgt ttg ctc cat gct ccc aaa ttg att taa aag tgc att ctg tgt cta
 ttt ctg gtt tat tat ctt caa gag tca cac aaa tta ctc ttt tgt gtg atc agt ggg agt
 ccc ctg cag tgt gac ttg att tat gga gac ttg cat tta tcc tat gtt cat ttt gca aca
 atc aat att aag aag gtc ttc ggt tct cct gaa tca aaa ttt tct gga aaa att ctg tct
 aaa tat ttc att gat gtt ctg gaa tac att gga act gta ctt ctc ctc atc tca tca
 tac aaa caa gct tta acc agt gta gtc ctc **TTG CCT CCC TAC AG GAT CAC TGC TAT TAT**

IntrB13-1 Fw

CAT GGA CAC GTC CAG AAT GAT GCT GAC TCA ACT GCA AGC ATC AGC ACA TGC AAT GGT TTG

-108

D H C Y Y

Figure A1. Cont.

H G H V Q N D A D S T A S I S T C N G L -128
AA gta aga tag tct cta atc ttt tat ttg ttt att aat aat aat ata gtg ctc ttg gag
K |-----EoPII-5 → -129
 ttc taa ttg tta aaa tga agg aca tcc tca gtt ttt cat gga aat tag ttg ggt gtg atc
 cag gat ttc ggc aga att aag aca tac ttt ggt tga aaa cca aga aga gct gct gcc agc
 cag gag aaa aac tat gga gct aaa tca cat aag tct aaa gga gct tcc aag ccc cgg tct
 cct ttc cca ggg tga ggt gat att aca ggt aga gaa gat tag tag gtt tca aat tgg aga
 cct tgc tag aaa gtg tac agg aag agg caa gaa gtt tca gtt cta ccc aga aac act ttc
 ttg agt cac tct gca cac ttt ctt cag cca act aga tat gtt aac tac ata aag atc cca
 gaa ttc aga agg tcc cta tca ata gta aga atg aac atc acc tca aca tct ttt act gaa
 aaa aga cac tga aac tca cct ttg aac aga gac tgt gtc cat gga gtg gag gaa taa atg
 aaa agc tgg aac aga gca gaa taa caa cag aaa aat aaa gga aaa aca gaa tga cag aat |---
aat agc att gga agg gac ttt gag gtc ttc tag tcc aac ttc ctg ctc aag tag gag acc
 -----LINE/L2/CR1 →
tat atc atc cta gac aaa tag ctg tca atc ttc tct taa aaa gca gta gtg atg gaa cac
cca caa tgt ctg aat agg tta att gtt cca ttt gtg aga aaa tta ctc ctt agt tct aac
tta ttt ctg tct ttg gtt act ttc cca ttg cca ttg ctc ctg cca tca ggt gaa gaa
tag gtt gtc cca cat ttt tta tga cag cct ctt aaa tac tta aag att atc aag tca tct
cta ccc ctt ctt gtc act agc atg agt ata ctc att gtc tgc agc cat tct aac cct cca
 gtt agt atg cat tct tat tcc ttt cat tgt tac tcc ctg ttg ttc tgc att gac ttc tct
 atg aga tgc ttg cca aga atc tat ttc att tat aaa tat cct gtc atc tga ctc tat
 cta aat tgc tat caa act aat ctg att tta ttt cct tga cca cag aca aat att gtt cta
 tac ttg ttt aaa gta aat tgc agt att acc tat aac tct ttt tag ata ttt tag cag tta
 tat ttt tcc ttt ttt atc cta ctt agt tgt gat tct tga gct tta tca qta ata tat atg
 ata aat ata aag tat ttt acc ctt atg aaa taa agt ttt aca caa agc aga atg tta caa
 ttg gct tta gtg ttg tat tta tgt agc tag aaa ctt att ttt tta aca tcc tgg aaa tat
 aca ata ttg ggt tcc atg cca aaa tat ttc caa aca aaa ctg tac acc tat ttt gtg gct
 gca ctg agt ttg tga aat ctc tca tat ctt tct gat cat aac tgc atc tat gaa aag tat
 gag aaa gtg att tga gtg ctg agg aaa gaa tat aaa ata ttc act cat tgt taa gaa gga
 att caa aaa cat gag gtt agt tga aaa tgg gtc tca gag cag agt ttc att acc caa cta
 ggt aac atc atc agt gca gtt ttt ctc tga act aac aat att ctc ttc ttt tgc ttc tcc
 atc tct gat cat cct ttt cac att gtt tta cag **A GGA TTT TTT ACG CTT CGT GGG GAG**
ACG TAC TTA ATT GAA CCC TTG AAG GTT CCC GAC AGT GAA TCC CAT GCA GTC TAC AAA TAT -137
T Y L I E P L K V P D S E S H A V Y K Y -157
GAA GAT GCC AAA AAA AAG GAT GAG GCC CCC AAA ATG TGT GGG GTA ACC CTG ACT AAT TGG
 Prodom 1 Rv /Prodom 1 Fw
E D A K K D E A P K M C G V T L T N W -177
GAA TCA GAT GAG CCC ATC AAA AAG GCT TCT CAT TTA GTT GCT ACT TCT GAA gta agt ctc
E S D E P I K K A S H L V A T S E |----EoPII-6 → -194
 ata gta aac ata gtt taa gat cac ata ctc att tgc ttg ttt aga aaa tat aaa gta aga
 gag aaa ttc ctt ttg gga gag gtg ata gat aga att taa aat gga gaa gcc acc att tct
 ata ttt tta ttg taa agg **taa agg taa agg ttc ccc cac aca tat gtg cta gtc gtt tct**
 |-----SINE/Sauria →
ggt ccg gtg ctc atc tcc gtt tca aag ccg aag ggc cag tgc ttg tct gag gac ata tcc
gtg gtc atg tgg ctg gca tga cta gac acc aaa ggc gta cgg aac gct ttt tcc ttc cca
ccg tag tgg tcc cta ttt ttc tac ttg cat ttt tac atg ctt ttg aat tgc tag ggt ggc
aga agc tgg gac gag taa cag cgg ctc act ctg tta cac agc act cgg gat ttg aac cgc
caa act gct gac ctt cag atc aac aag ctc agt gtc cta acc cac tga gcc act cgc tcc
ctc att ttt att gta gcc aag gta taa aag aaA CAA TGG GAA ACT GAG GAA CAG aaa ata
 Intr2 Fw
 aat ctt cca gtt gtt tga tca aac aaa ctt agt ttg gag att tga atc aaa aat gga ttt
 aaa tga gtt tct aaa tca tct cca ctt tct aag tca att ttg aaa agt aat taa att atc
 aat ttg gat gcc tct ttt atg cat gca gaa aga aat tga gat ggg gga gaa agt ggt ttg
 aaa tat tta atg gtt tta aga tgt ctg ata agg cca tta cat aat tgt gac tcc att ttc
 cag ttt gat ttg aat cat caa gtt gga ttg atg caa tga tgc aat aac tgg atg gaa agt
 ggc aat gtg aac cta gtc aca att ggc cct tat gct ctc aat gtt ttc ctc ctt tac tgg
 agg cac aaa aat tag aaa aca aaa tat tgc atc caa agt gac agt tcc tta cat ctt ttt
 ggt ggc aaa agt tga aac tgg ctc aaa aat ctc aac tgt ttt tat tag aat gtt aaa att
 gac atg gaa caa cag ctg ttg tta tga cgg aat acc aaa aca caa ctg aag agc cca aat
 gaa gtc ggg ttt gtc ttt tgg ctt ctt tca ttc tgg caa ttc aag att ctt tat cct cag
 caa ttt gtg gtt ata cgt tac att taa ctg gtc ctt aat tgg ttc ttt atc ttg aat gtt
 ctt gtc tca aag tca gtc atg atc act tta ttc gga ctc gat ctt gtg cag cag aat taa
 gaa agt ggc tgt gag agg aga agg gaa att gca ttc tag aat tgc aac ttg gct tgg
 tag ttc tgt aca ctc tca tag gga aag aag caa tta atg cac act cac aca ttc agt ctt
 agc ata agt gaa ctt cct gag caa cct ggg ttc tct agt gga gga atc tcc ctt tgg aat
 gca gga gag agg ggc tct ttc cta gag aaa taa gag gca gat cag aaa tga atc ctt gga
 ttg cag aaa gat gcc cgg aag gga tcc tgc cta aag cct tca cag gat tca tct cat ggt
 tac acg gag tgt gct tta gtc tgg caa gat tcc tgt cta cag gag aag agc aat aaa gaa
 act act ctg agt aat taa cca cag ttt ctt gtt cct tgt gct caa gaa gtc tga aac aat
 att tca gta atc att tta aaa ata cat cac tga aaa ggc ata att ctt tgg cca tta aga
 gtg agt tca gtg ggt caa aat gtc tat ttg ctt ctc ctt acc ttg tct ttg atg cat agt
 gtg ata tga ttc cag gcc agg tct aat tgc atg acc aca tca gca tgt ctc tca cgt gat

Figure A1. Cont.

tgg ttg gca tcc ttt gtt tga gaa agg gaa gga aag ttg aga aag tca ttg aag cat cat
 ttt gac agg gtg aaa aac acg tca aag aga aca gtt tcc tca tat gcc tat taa att ctt
 ttc aga gtt agg tat tca tat ata cca tta tct tga caa tcc att gaa taa cgt act ttt
 ttc ttc aaa act tta tca **TGA CTT TCT CTA AGT CAG AGT TCC** Caa acc ttt oca gct ttg

Intr2 Rv

ggg ata ggt gag gga gag ggg atg gtt cca cgt gaa cag tgg ggt cag gtg tgt acc cag
 ctc tat ttg tgt gag cag tgg gca cac ata ccc act cgt gta aac aga gca cac oca cct
 atg ctt gtt cac tgg tca tac aag tag aga tgc agc tgc tca cct gcc att tcc atg gcc
 cag ttc tga agg gct gca ggc oca ggg cta aaa ttt taa caa gct gtc tcc ctg taa ata
 tct tct tga aag aac tga tat ttc tgg aag ttg acc aga gag taa aac aag cat ttt tct
 gat tat ctg agg ttg acc aac gtt ctt gtt ggc tgc tgg gag taa cat gtt taa aca gcc
 att tac ttt tcc gtt tag **CAA CAG CAT TTT CAC CCA AGA TAC** **GTT CAG CTT GTC ATA GTT**

Met1Pll Fw/Met1Pll Rv

Q Q H F H P R Y V Q L V I V 14

|----- Metalloproteinase →

GCA GAC CAC TCA ATG gta agt atc ttg gat atc ttt cta ttt act ttt tgc att gag cgc
A D H S M |----- **EoPll-7 →** 19

agc att tgg ttt tgg cct ttt tta atg tgg gca ttt ttc aag aga tta tcc taa ctg aac
 ttt ctc tta aaa tgc gct tta tca taa act ttg ata ttt ttt gtt att gga ccc aac caa
 aat tta caa agc taa aag tca ttt gta aat ata ttt taa ttt gca cac ttg tta ttc ctg
 gat acc att taa gtt tat ttt tat ccc aca ctg gtt aaa aag ttc tgt agg ttt ctt tga
 aat gtt ccc acg ctc ttt ttt ttg tcc aag ttg gca aca tac aaa gaa aaa agt gaa gaa
 tgc tta tct cac aca tct ctg aaa gag gaa ata ttt tct cta acc agg aaa aag gcc cat
 gta tgt tgc tga aaa gtt aaa aat ctt aat ata tta atg gca caa atg tag att aaa aaa
 aal agc lca aaa gaa lll lll gga ctg cll gga laa aaa lla lla cal caa gaa all caa
 aga tcc ttt cac taa tat att ctt ttt ctc ctc cct tct ttc cct tct tat aat gat caa
 ctt gtc tta act ttt ttt ttt tgt ag **GTC ACG AAA AAC AAC AAT GAT TTA ACT GCT TTA**

V T K N N N D L T A L 30

ACA ACA TGG ATA CAT CAA ATT GTC AAC GAT ATG ATT GTG gta aga aca aat gct tgt tca
Met 1-2 Fw |----- **EoPII-8 →** 43

T T W I H Q I V N D M I V
 ttt taa act tca ctt agg ccc agc cga gat ttt gat tgt gtt aag ata aca aac ata atc
 agg taa ata aag tag atg gat ttc taa atg caa acc tct gct ccg cac **gct gca ttg gct**
 |----- **LINE/L2/CR1 →**

ccc agt tgc tct ccg ggt gag att cag tgc tgg taa tga cct ata aag ccc tac atg gct
tgg gtc cag aat atc tga ggg aac acc tgc agc caa gtt ctc atc gtc cgg tac gct ccc
aca ggg agg ggc tcc tta gag tac cgc cgg caa agg att gcc ggc ggg tga ctc cta gag
aga ggg cct tct ctg gtg cac ccg ccc ttt gga acc agc tcc ccg tgg agt tga gga
ctg ccc ccg acc tgc gtt ttt ttc gga gga acc tga aaa cat ggc tgt tta atc tga ccc
 agg ctg gtg ttt tta gat ttg ggg ttt taa ctt ggt ttt aat ttt gag gat tgt gtt taa

|----- **LINE/L2/CR1 →**

tgt att ttt agc tgt ttt tta att ttt gta ata ttg tct ttt aaa ttc ctg tac acc tcc
ctg agt cct tgc gga aaa ggg tgg ttt aaa aat aga att aaa taa ata aaa ata aat aaa
taa atg aat ggt ttc ttt tgt ggc ttt gag tga tct aaa act atg tta ttt ggc cag ctg
 gtg atc cag cat gtg agg cca tcc cag aaa atg ggt taa ttc agt aac taa att aat aag
 cct cta atg gaa ata gag tca ttg ggt ttg gag gat gca aat oca aaa gtg ttg caa gaa
 ttc agt cag aag tat atc tgt agt tgg att ctt tca ggg ctg tgg tta atg gtt ctg agc
 ttg gag att aaa aaa tga tgg aca gag tca gcc tat tcc aag ctc agt tga tta tga aaa
 tga tct ctg agt aga aac ctt gag aga aga act caa ttt gac ttg gat tgt ggt act tag
 gag gag ata cag tca tat tgc att ctc cga tta gcc ttg ttt aca ttt tca cta ttg atg
 atc agt tag aaa cag gga gag aac agg aag atg gaa gaa tca act ctc ctg tta gtc ctc
 cta ctc ttc tgt tcc ttt tta cag aaa ata atc acc tgc ttt ttt tat cat gtt att tat
 tag agt cct tag tac tgt cta agc ttg gtg gtt acc ttg cag aca ttt cat tat ctg act
 act cta cat gag taa oca aag cag caa gct cat agc acc aag gac ttc aac ctt gag ctt
 cta ttg atg cct agg tct att ttt ttt tca tct aca aaa aga ggc caa tgc cag ttg ctt
 ttc aat att gaa atg ttt ctc aag gtt tac ttt gtt tga gtt tct gac tga tgc **CAT GTC**
CAC CTG ACA G ATG TAC AGA ATT CTG AAT ATT CAT ATA ACA CTG GCT AAC GTA GAA ATT

Met 1-3 Rv

M Y R I L N I H I T L A N V E I 59

TGG TCC AGT GGA GAT TTG ATT GCT GTG ACA TCA TCA GCA CCT ACT ACT TTG AGG TCA TTT
W S S G D L I A V T S S A P T T L R S F 79

GGA GAA TGG AGA GCG AGA AAT TTG GTG AAT CGC ATA ACG CAT GAT AAT GCT CAA TTA ATC
G E W R A R N L V N R I T H D N A Q L I 99

AC gta tgt ctc att gtg ggg aaa ggg agt gag agt ggc tgg gag tgg agg att atg gaa
T |----- **EoPll-9 →** 100

agg tta atg ctt gcc tag agc ttc tgt tct atg ctg tat gct tta aac cat gca tgt agt
 aca ttt ctg ggt caa agt caa cca ctt ata tta tag atg aga cct gcc ttt gag aaa tct
 ttg aat gat tgc agg tga aaa atc cat taa ata aat atg tat ttg ggg ttt gca tag tta
 atg gaa tta aat taa ccc aaa tgg gtt gga ttt gac aac atg cta acc cct ctc ccc aac
 ctc att cat gct aat gcc agc caa aca caa cat ata gtg aaa aat ata tcc tga ata ttt
 tta acg gac ctg ctt aac taa atg gct ttt gag act aga gcc taa aat gaa ttc tag cat
 tct gaa aac tgg tag taa ctg agg aog ggt caa agg gtt tac aga aat oca tat tta tgt
 atg acc taa gac tac ata aca tgc ttc tat ctt cta tca att tta tcc ccc tcc cct tct

Figure A1. Cont.

ttt tta tag A GCC GTT CAC CTT GAT AAC CTT ATA GGA TAC GGT TAC TTA GGT ACT ATG
 Met2Pll Rv/ Met2Pll Fw
 A V H L D N L I G Y G Y L G T M 116
 TGC GAT CCA CAA TCG TCT GTA GCA ATT ACT GAG gtt agt aga aag gat act tta tta tct
 Met6Pll Fw/Met6Pll Rv |----- EoPll-10 →
 C D P Q S S V A I T E 127
 att tgt act caa gtg aaa cct tac ata cag aca aaa cat ctt ttc aaa taa agt ctc ttt
 ctt att ttt gag oca cgt cat ttt cac oca tat tta ttt gca gat ttg aca tct oca ggt
 cct gcg tca act aat ggc att ttg aca cag tgc att cta gaa caa gct ttt tta atg caa
 tga gct ata tgt caa gga tga gaa tat att ata atg ttt atg gtt cag tca aac tgt act
 ctg att ggc aaa tga aca ggt caa agc atg tta caa cac ttc caa ata atg ctt ctg aac
 aat agt ctt agc aat ccc aaa gac aaa cat gaa ttc att oca aga aat tta gtg tct aga
 ttg cat atg att gaa ttc tag tac att gag aaa aca aaa aaa tac taa atc tac tca aaa
 aga aaa aaa acc ctc tag ata tta gtt aag gtg atg cta tgc att tat tga gaa aga gta
 aac tta gct ttt tgt tca cat aga aag aat gga gag aca tgg taa taa aca aaa gtt ata
 caa caa aac tca taa agt ttt gtt tct taa taa gca gag tta gga tcc tgg tag tag tag
 gta ctc ata agc cta ctt gct caa gaa ggt tat ttt att cag aaa gag caa ctc att cta
 agt ctg ttt agg atg ctt acc ttc aat att cag aaa atg caa gat tgt agc aaa gga cac
 tga gta gtt ttt ttc gac tga agt ttc ctg taa gtc agg gct gtc aaa ctc aat ttc att
 |----- DNA transposon →
 gag ggc cac atc agc att gcg gtt gcc ctc aag ggg gtg gtt ggg tgt ggc cag ggt ggg
 cac agc cca cag gca tgg ctg gaa tgt ata tgg cta agt ttt agt aac tga ata agt gca
 gac agc aaa tgg atg cat aca ttt tga tct tat tct gtg ctg tag ctt ctg gct gta aag
 ttt cct tct gga tgt att tgt gta tgt tct gga gtc ttg gtg ggc aca gat act ttc aga
 gga gct aga gga atc ctg aga tgg tat cct caa cct aaa att ggt cac ttg gtc acc agt
 ttt agc cac tta gtg gta ata ata att gga ttc act ttc agt ttc ttg gca gag taa caa
 taa aaa aag tat tct tat ttc ttc aG GAT CAT AGC ACA GAT CAT CTT TGG GTT GCA GCT
 Met3Pll Rv/ Met3 Pll Fw
 D H S T D H L W V A A 138
 ACA ATG GCC CAT GAG ATG GGT CAT AAT CTG GGT ATG AAT CAT GAT GGA AAT CAG TGT AAT
 T M A H E M G H N L G M N H D G N Q C N 158
 TGT GGT GCT GCC GGA TGC ATT ATG TCT GCG ATC ATA TC gta agt att gag gaa tat gct
 C G A A G C I M S A I I S |----- EoPll-11 →
 taa tgg ctt tcc aat caa gtt att ttt aaa tgg ttg caa aaa tga ata aag tat tct ctt
 atc cat tct gtt agc ttt aga aga aaa caa atc att aca ttt ctt cat tag caa ttc ctt
 ttc ctt ata tgt ttt tgc aat gaa att ctg ctc cta gtc caa agt tgg agg atg tca tga
 tct ttt ttc ata tct aca g A CAA TAC CGT TCC TAT CAG TTC AGT GAT TGT AGT ATG
 Met4 Rv / Met4 Fw
 Q Y R S Y Q F S D C S M 183
 AAT GAA TAT CGC AAC TAT ATT ACT ACT CAT AAC CCA CCA TGC ATT CTC AAT CAA GCC CTG
 N E Y R N Y I T T H N P P C I L N Q A L 203
 AGA ACA GAT ACT GTT TCA ACT CCA GTT TCT GAA AAT GAA CTT TTG CAG gta aga gaa gaa
 R T D T V S T P V S E N E L L Q V R E E 219
 |----- Spacer -----||----- EoPll-12 →
 tgt gac tgt ggt tct cct gca tta agt ctt ttt ttt taa tca aca aaa gta att tga aga
 C D C G F P A L S L F F Stop
 ata ttc tca gaa atg aga atc ctt gaa aaa tca tct agc ttt cta agt ggt ttg agc cat
 cca aga ggt tgg ctt gtg aat ggc tga ggt ttg tgc ctt tca tgt aca tgc atg tat gaa
 ggt gtt tct tgg gtt gta gag gaa tgg aga act ggt atc tca cta cta ttt tgg gaa gat
 ggt gaa ttt tta aaa acg ggt gat tga oca ctc caa gaa aat ctt tcc ctc ctg aaa ccc
 cct att ttg ttg ata tag oca cat tat cct gta oca oga ttt tct oga act gct ccc tcc
 cat atc tga tta tct tta atc tat gct ctg atc cta ata ata ttt tta taa gaa cag taa
 tat agt gtt ttt atg ttg tta aat aca cct gtg atg gtc tgt gag aat gtc ctt aag aga
 caa aag aag gac gaa aca tcc agt taa tgg tgg tat aag aag gag att aac ctg cag aag
 caa tgg caa aaa tct caa gat gga cac ttc cca ccc att ctc ttg gtc cgt aaa gat gag
 gtg gta caa ata gat ttt cag tat tga aag att ctg cta ctg taa cct tac aat cat gtc
 gca tta ata ctc aag gtt gct gct tct cta gac taa ctc aaa ggc tgg cat gat gag
 tag aaa atc cat cat gaa taa gaa aag aat ggg gct gta ggt tat gtg ggc ttc aca att
 agg aga tga gga tat ttt tgt ttt att ctt ttc acg tag gaa ata tca gat aag gct ctt
 tcy cag aga aat gcc ttt agc tgt ttt caa taa caa aca att tgt gca tct cct agc atg
 aac tca taa gag gga aca tat cgc aga aat gtt cct ctt caa aat aga cca att aaa aaa
 gaa aat tct atg oca tca ttc gat atg ttt tgg ttt tca g AAT TCT GTA AAT CCA TGC
 N S V N P C 225
 TAT GAT CCT GTA ACA TGT CAA CCA AAA GAA AAG GAA GAC TGT GAA TCT GGA CCA TGT TGT
 Fw_Ocella Ncol
 Y D P V T C Q P K E K E D C E S G P C C 245
 |----- Short Disintegrin Ocellatusin →
 GAT AAC TGC AAA gta aga ctt att tat ttt taa cac caa gag aaa ttt tac cct gct cca
 D N C K |----- EoPll-13 →
 tac tag ccg tat aga aat ata ata ttt ctt ggc tgt tta cta tga taa aaa cat ttc agc
 IntrDis1 Rv
 tct att tcc tat ccc ttc ttc cag ttt att tga ccc tta tga aca taa gca aag gga aga
 taa ttt aac aaa att tct ccc tta ttt caa ttt caa atg cac tct ttc agc atg cta aat

Figure A1. Cont.

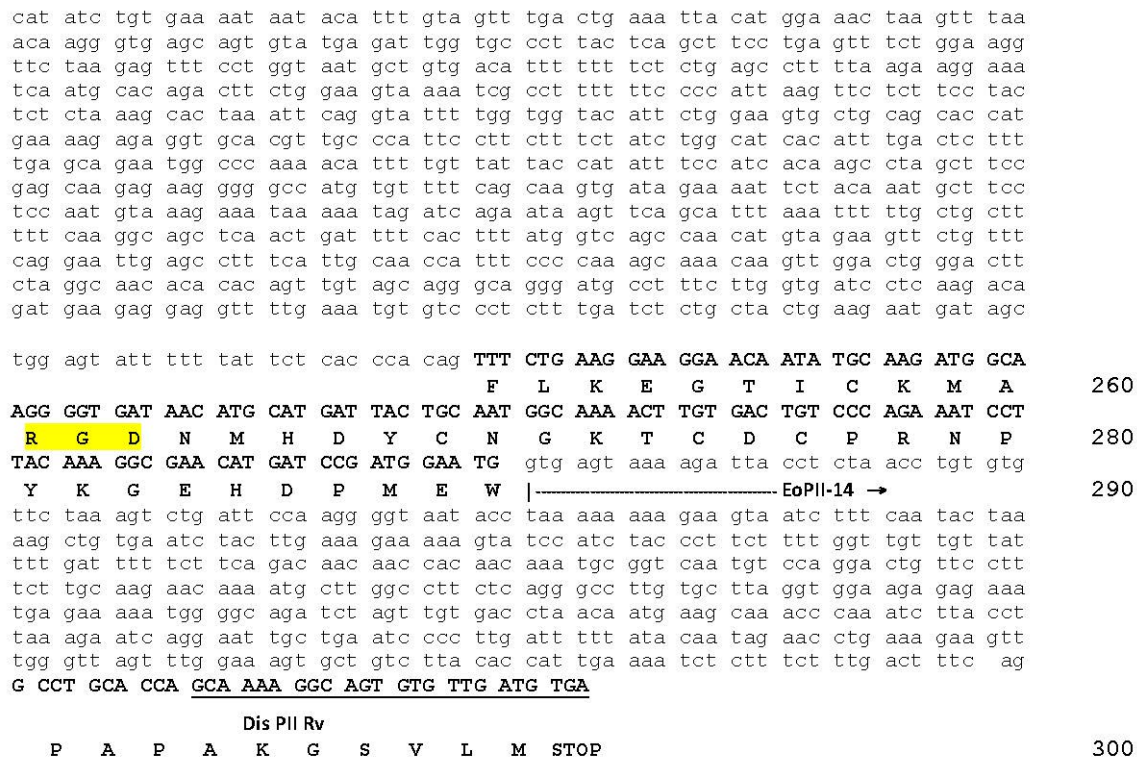


Figure A1. Genomic organization of *E. ocellatus* EOC00006-like PII-SVMP gene.

Genomic sequence of *E. ocellatus* EOC00028-like PI-SVMP gene. The locations and identities of the primers used to PCR-amplify genomic sequences (listed in Table 2) are indicated. Protein-coding DNA regions are in upper letters and boldface, and the encoded amino acid sequence is displayed below the DNA sequence. Start of introns are labelled EoPI-X, where “X” corresponds to intron number. The beginning and the signal peptide, propeptide, metalloproteinase domains and the C-terminal extension are specified. Numbers at the right correspond to amino acid numbering of the DNA-deduced pre-pro-PII-SVMP relative to the mature SVMP. The extended Zn²⁺-binding environment (HEXXHXGXXH) stands on yellow background. The only two amino acids (-124T/A and 15T/A) that distinguish this sequence from that of PI-SVMP EOC00028 (Q2UXQ3) are shown in bold and red. The remains of a disintegrin-like domain and a dimeric disintegrin domain transformed into intron EoPI-12 are underscored in italics and on cyan background. The N-terminal glutamine of the metalloproteinase domain has been assigned residue 1. SINE/Sauria, LINE/L2/CR1, LTR/ERV1, DNA/hAT-Ac and DNA transposon retroelements are highlighted on a gray background. Inserted nucleotide sequences in introns 1 (582 nucleotides between positions 1534–1582, including a SINE/Sauria element); 9 (between nucleotides 194–195 of the topologically equivalent intron of PII); 11 (replacing nucleotides 1-66 of PII intron 11 for a stretch of 3281 nucleotides); and 12 (after nucleotide 999 of the homologous PII intron) are underlined. Simple sequence repeats (SSR, microsatellites) are shown in light green background.

ATG ATC CAA GTT CTC TTG GTA ACT ATA TGC TTA GCA GTT TTT CCA TAT CAA G gta aga tgt tct gtt tag ttc cct tgt
 |-----**Sp35_Eo Fw**-----|-----**EoPI-1**→
M I Q V L L V T I C L A V F P Y Q
 |-----**Signal peptide**-----|
 tca gaa tct tac tgc taa aag act att gca ccc aaa aga ctg cta tgt tgg tag ttt tgg
 ttt tat ttt tga caa tta acc aaa gtt tac ttc act tca gtt tct aaa gat taa gca aaa
 gaa tgt tct caa gga tca cat ttg ttc taa agt tac ata aat ggt tta ttg gta ttt gtt
 taa att ttc tca agt aag gag caa atc cta agg aaa aaa gtc cta aga ttt tca tta aaa
 aag cac att cat gtt gga ggt aat ttt ttc caa tga taa gat taa cac att gaa gag gac
 aga gag aaa gtg ttt gtt gca aaa aaa att cag aaa gca aaa gaa aaa aat gtt ttt att
 ttt aat ttt ctt tca ata ggt caa tta gac atc ctg aaa tta agc atg cat aat att tag
 cca ggg aca caa tgg ctc agt agg ttc gga taa tga att tgt taa cca gat ggt gag cag
 |-----**SINE/Sauria** →
 act ggc ggg tca aat ccc aag tgc cac gta aca gag tga gtg cct gtt act tgt ccc agt
 ttc tgc caa cct agt aat tca aaa gca tgt aaa aat cca agt aga aaa ata agg acc act
 aca gtg gga atg taa cag cat tct atg tgc ttt tgg cat **CTA GTC ATT CCG GCC ATA TGA**
 |-----**Intr1F1PI Fw**-----|
 cca ctg aga tgt cct tgg aca aac act ggc tct ttg act ttg aat gga gat gag cac caa
 ccc cta gag ttg gaa atg agt agt atg cat gtg tgg ggg aac ctt tac ttt tac cta ata
 ttt agc caa ata aaa gct agt gct ttt gat gtt aca gag aat ctt cag act gta ctc aat
 ata atg tga cag tag ctt ttc tca atg caa gat tcc gta tta tct taa agt att tct ttc
 ctc ctt ccc ttt tac tct gca aaa aat aga gca gtc cct tct gaa att ctt ccc aag ttc
 tcc ttc tgt gag cta aga tac att tta cat gcc agt tag cca atc tat ggc tat tct tcc
 tgt atc cca agt cat tcc cac ata tca tta caa aca aga ata ctc aat aga ttc aga agt
 tgc caa aca aat tta agg gtg aga ggc tca ctg cag tag ccc **tgt ttg ttt gtt tgt ttg**
ttt gtt tgt ttg ttt gtt tgt ttg ttt aaa tag caa aga aag atg aga atg ccc tta tca
 gga gta att tat ggg gaa caa tgt agg aaa gaa gaa agg gag atg gaa gag aaa ttg ctg
 ctt aat ctg tgc tct aag caa gag agg aat agt aag ctt gct agt agg tgg aag tga acc
 cct ccc atg ttt ctg gga gaa tca ggt aca aat agg tag gta cat caa tag ata taa gga
 aga tgt att ctt ctg tct ggg gct gcc **cAT CAG TCT GAG AGG ATG CAT TTC** Cag ggt aaa
 |-----**Intr2F1PI Fw**-----|
 atg ctg gtt tct ctg **ttt ctc cct ctc tcc ctc tct ctg ttt ctc ttt cac tct ctc tct**
 |-----**LTR/ERV1 and DNA/hAT-Ac** →
ctc tct cac aca cac aca cac aca aac aca aac aca cac aga cac aga cac aca cag aga
gag aga gag aga gag aga gag aga gag aga gag aga gtc tgt ctc tct agc tcc ttc cct
 ctc ttt ctc tct ctc tct gtc tgt ctc tgt ctc aat tgt tta ggt gtt tgg gta gca gtt
 tct ttc cta taa ggt **aaa ggt aaa gtt tcc ccc ata cac acg tgc tag tca ttt cca act**
 |-----**SINE/Sauria** →
 cta gga gct gat gct cat cgc cgt ttc aaa gct gaa gag cca gtg ctt gtc cat gga cat
 ctc cgg gat cat gtg acc agc atg act aaa tgc cag aga tac atg gaa aac tgt tat ctt

- 17

Figure A2. Cont.

```

ccc act gca gta gtc cct att ttt cta ctt gca ttt tta tgt gct ttc aaa ctg cta ggt
ggg cag aag ctg gga caa gtt aac aag tta act cac tct gtt acg ctg cac tgg gga ttc
aac cca cca atc tgc cga cct tct gac cga caa gct cag tat cct aag cta ccg tgt ctc
ttt tct ttc cta tac cct gtg gtt att gat att tga att tat tgt ttt gtt tct gct tgt
aat tat gag cac aac ata ccc tgt cct ttc cat taa tct gct aac aaa tct tgg tct cag
ttc tga ttg tac atc caa gca tac ctg cag gtg tca gac tgc ctc att tac aat ata agt
taa aca aca aat atc aGT GAC CAT GCA ATG TCC ATA TGG gga coa act cta aat ttt toa

Intr3F1P1 Fw
cta ttg gtc taa ttc tga atc att gaa tta tgg atg tta tta gat aat tat gat gct gtg
aac tct oca tag att att tgg aaa gtg aaa gaa aag aaa gtc agc aaa agt att agg ata
gca gta aag tag cat ttc agc tga aag gaa ttc caa ata atc tgc att tgc ata gaa atc
taa gtt tgc cta ggt ttt aga cta atg gat atg cta aac cat aac aca ata tgg ttt act
gaa ttg tat gaa tca att att gtg att tac agg ttc tca tct tat cac tca ggg cac aaa
cct gca cat tgg ttg aaa tct att tat ttt tct aat tat tca agc tct gaa atg ttt cct
ttg cac ctg cac aaa gaa atg tct cct gtt aga cta oca ggt ctc aga cat cat gca gca
aat tgc att ccc aga gaa aat ctg gat ttc tgt gct gtg tcc cac tgg gaa aag att oca
ttc ttc act gat aac tga ata aca aat tgt gtt gtg gat gtg aaa tcg ata gaa taa cat
ccc aaa gaa aac cga atc ttc ttt tct ttt tcc tta tga ag GG AGC TCT AAA ACC CTG
G S S K T L -23
|----- Propeptide ->
AAA TCT GGG AAT GTT AAT GAT TAT GAA GTA GTG AAT CCA CAA AAA ATC ACT GGG TTG CCT
K S G N V N D Y E V V N P Q K I T G L P -43
GTA GGA GCT GTT AAG CAG CCT GAG AAA AAG gta aga tat ttc ttt cat caa caa att att

Met15PI Fw |----- EoPI-2 ->
V G A V K Q P E K K -53
ttt tgt cag tcc ata gaa ggt ttg ata ttc ctt tcc tgc cat tta atg gtt att tgg att
ttt cat tgc aat cca tgt tcc tgt ttt att tat tta ttt gtt tgt ttg ttt gtt tgt ttg
ttt gtt tgt ttg ttt gtt agc aaa ttt tac tgc ctc oca gtt tac aca aaa ttg agg aga
atc ttg gca act tac aaa cag aat ttc cta ata tta ata gtt tag tat aac taa tca agt
gtt ggg aga aag cca gcc ata cag tga aga gga atc aag cag tga aat aac cat tag tca
ttt cat act gca aca gat oca gca aag aca tcc tct ccc tcc tcc agg aaa tcc cat taa
agt tta tct gat ctg cag aga atg gaa gca gtc tag gga gag tcc tgt att aca ggc aaa
aaa taa taa act gct taa ttt aaa ctt tgt ata tct aca tgc gtc aca tag aaa tat aaa
aga tat att tct aga tat gat agt ata gtg gtt aca gct aaa gat tag ttc tat ata oca
gat tca tgt ctg cgg tgg gac gac ttt tga gtc aat ctc tct gtc oca gat caa cct cat
tca gca ata tgt tca gtt aga gaa tga gat ctt gaa cct atg cag gaa aaa taa ata aag
atc tta tca ttc aaa gca oca ggt gaa ata cct aat aat aat aat cta ata aat aat ttg
act atc aca aat tat tca tta gat tta aat taa tac aga tgt aga gta tta aga aat gac
aca ttt tat ttg cct aaa ttt tga gaa gta aaa cat aaa ctc ttt gtt ttt cag TAT GAA
Y E -55
GAC GCT GTG CAA TAT GAA TTT GAA GTG AAT GGA GAG CCA GTG GTC CTT CAT CTG GAA AAA
Prodom 2 Fw/Prodom 2 Rv
D A V Q Y E F E V N G E P V V L H L E K -75
AAT AA gta tgt taa ctc aga att ttt ttt aac ttt act aaa caa tgt gga aaa tgt ata
N K |----- EoPI-3 -> -77
ttc ctg gac aca atc tga gag aaa taa taa ttg cat tcc ttt gtt tgg aaa tta aaa tta
aat taa atg tta cta tat aga aaa agt gga tat aga tat taa gta tgt taa tta tgt ctt
acc atg aaa ctg aat ttt ttt tac tgc tgt ttt cct atg gaa cat tgc tga acc ata aca
att tta gac tca gtt gaa oca ttc aag gct cat aac ctt cta tca ctt gaa tta act agg
tta gtt atg aaa tat tgc agt gat ttc aaa atg ttt gtt tat gct gct tca tag aca ctc
cat caa tct gaa taa aat ttt cta tga gac ctg tag agc tgt tca acg tac ata aaa aaa
att ata tta aat cat aaa agg cag gaa atg tca tct cct att aag acc oga aag gga cag
gac cta aat aac cct gga ttt cac aag cct ttt tta aaa gaa tgc ctt gtg agt ctt ctc
aga gtt gtt oca ttt aaa ctg taa atc aag ata tta ata aaa ata tta tac aaa tac gaa
cca caa aaa tgg aat gaa att att ttt cag cac aaa aga caa aca tag ttg gaa ggg acc
|----- LINE/L2/CR1 ->
ttg gag gcc ttc tag tcc acc ccc tgc tga agc agg aga cta tat cat tcc atg ggt gtc
caa ttg ttc ctt gaa aac ttc cca cag ctt ctg aaa gct acc cat tcc acc gat taa ttg
ttc tca ttc tca gga att ttt ttt taa ttc tag gtt gaa ttt ttc aca ggt aaa taa gcc
aca aac acc cct aaa aga aag tgt gcc att acc cta aaa act gta aat gtc ata
aca cag ttt aaa tta gtc tgg cag tca cta aac tgc aaa cac gta cct cag cct ggt tca
gtt taa tac tat agt taa GAC TGA GAT ACA GAG TTG GTG Cat tgg gaa aga ctg cct gcc

Intr3 Rv
ttt gat ttt agc atg tat ctt ggt ttt gat cat tct gca tgt acc ttt aaa agt ata ttg
tct tta tta tta aac ata tgc tac aag aaa ttg cac tat atg tta gta tct gtg tgt cac
ttc ctc cta ttc cac cgg aga tat ttc atg cta aat tct aat gtg tgc ttt aga cgt tct
agt gtc ccc ttt tgt ttt gtt taa ata tga agg cta tac aga aat tca agg gta gca tat
gga tgg tct tct ttt gtc ata gaa gca aag cag ggc tac agg ggg aag acc agc aag gat
ttc tag att ggc aaa aat gga aga cga tga caa gtt tat ttt tct gac atc agg act gga
atg tga ctc cat ata ctg gtt taa agc ttc aaa ata tat ctt tct gta tga atg ttg aag
gca taa att tcc ttg agg ttg ctt cat ttt act ctt tag taa ata gaa cta tat cgt tct
cct tta gag aca atg cag ttt cat cag att gtt gca tag act aaa aga ata aaa gga aga
cgc aat aca aac tgt cct ctc ttt gtc cat cac aat cct ctg agc att gat gaa gag aat
tgc aca gct tca ttt ggt gaa tct agt atg cac tgc ctt cct ctt aaa aag oca tgg gat
aga aat acc tgc tcc att ctt ttg ata cca aag aag ata ctt aaa ttg cca tta tac ggt

```

Figure A2. Cont.

tgt ttt aaa ctt ctt ata	aaa tta tct ctg tgc tac ata att cct gat ata gat atc tct	
tct ttg cat tct ttc cag	A GGA CTT TTT TCA GAA GAT TAC AGT GAG ACT CAT TAT TCC	
	Pro2 Fw/Pro3 Rv	
CCT GAT GGC AGC GAA ATT	ACA ACA AAC CCT CCT GTT GAG gta ggg tct cac ttt tat gag	-90
P D G S E I T T N P P V E	 -----EoPI-4-->	-103
cct ttt ttt aag gaa gta aat tga aac aaa tgt ttg tgc act ata tta caa ata tac aag		
aat gag acc agg cta ctc aaa caa agt gta tat aag tat aaa gta tct tat att gat atg		
tac tta caa aga tgc ctg gat tgt taa tcc ttg gtt aaa agc caa cat att tgg gag gtg		
agt ttc aca aat aga ttt att atg aga aca tca ggt ttg taa gat tat att ttc att ttt		
aaa cca gac tac agg gat aaa tgc aaa gtc ttt tat ctg taa tac caa aag tga taa caa		
ttc act ttg ctc cta tac aga aat cca ttt aac atc ttt cat att aaa atg gtg cca aaa		
atg gct cta tca gag gtt aaa aaa tta cag cac taa tat gct tca tgt tgg ctc cat ttc		
ccc aaa ttg att taa aag tgc att ctg tgt cta ttt ctg gtt tag cat ctt cat ggg ttg		
cac aaa tta ctc ctt tgt gcc atc agt ggc act ctc cca tag tgt gac ttg att tat gga		
gac ttg cat tta tcc tat gtt cct ttt gca ata gtc agt att aag aag gtt ttc tgt cct		
cct gaa tca aaa ttt tct gga aaa ctg ctg tct aaa tat ttc att gat gtt atg gaa tac		
att gga act gaa ctt ctg ctc atc aaa tca caa tac aaa gcc tta acc agt gta gtc ctC		
TTG CCT CCC TAT AG GAT CAC TGC TAT TAT TAT GGA CGC ATC CAG AAT GAT GCT GAC TCA		
	IntrB13-1 Fw/Met16PI Rv	
ACT GCA AGC ATC AGC ACA	D H C Y Y G R I Q N D A D S gta aga tag tct cta atc ttt tat ttg	-118
T A S I S T C N G L K	 -----EoPI-5-->	-129
ttt att aat aat aat ata gtg ctc ttg gag ttc taa ttg tta aaa tga agg aca tcc tca		
gtt ttt cat gga aat tag ttg ggt gtg atc cag gat ttc ggc aga att aag aca tac ttt		
ggt tga aaa cca aga gct gct gcc agc cag gag aaa aac tat gga gct aaa tca cat		
aag tct aaa gga gct tcc aag ccc cgg tct cct ttc cca ggg tga ggt gat att aca ggt		
aga gaa gat tag tag gtt tca aat tgg aga cct tgc tag aaa gtg tac agg aag agg caa		
gaa gtt tca gtt cta ccc aga aac act ttc ttg agt cac tct gca cac ttt ctt cag cca		
act aga tat gtt aac tac ata aag atc cca gaa ttc aga agg tcc cta tca ata gta aga		
atg aac atc acc tca aca tct ttt act gaa aaa aga cac tga aac tca cct ttg aac aga		
gac tgt gtc cat gga gtg gag gaa taa atg aaa agc tgg aac aga gca gaa taa caa cag		
aaa aat aaa gga aaa aca gaa tga cag aat aat agc att gga agg gac ttt gag gtc ttc		
	LINE/L2/CR1 -->	
tag tcc aac ttc ctg ctc aag tag gag acc tat atc atc cta gac aaa tag ctg tca atc		
ttc tct taa aaa gca gta gtg atg gaa cac cca caa tgt ctg aat agg tta att gtt cca		
ttt gtg aga aaa tta ctc ctt agt tct aac tta ttt ctc tct ttg gtt act ttc cac gca		
ttg ctt ctt ctc ctg cca tca ggt gaa gaa tag gtt gtc cca cat ttt tta tga cag cct		
ctt aaa tcc tta aag att atc aag tca tct cta ccc ctt ctt gtc act agc atg agt ata		
ctc att gtc tgc agc cat tct aac ctc cag tta gta tgc att ctt att cct tca ttg tta		
ctc ctg ttg ttc tgc att gac ttc tct atg aag atg ctt gcc aag aat tct tat ttt cat		
tat tta tta aat atc ctg gtc atc ctg act ctt atc tta aat tgc tat caa act aat ctg		
att tta ttt cct tga cca cag aca aat att gtt cta tac ttg ttt aaa gta aat tgc agt		
att acc tat aac tct ttt tag ata ttt tag cag tta tat ttt tcc ttt ttt atc cta ctt		
agt tgt gat tct tga gct tta tca gta ata tat atg ata aat ata aag tat ttt acc ctt		
atg aaa taa agt ttt aca caa agc aga atg tta caa ttg gct tta gtg ttg tat tta tgt		
agc tag aaa ctt att ttt tta aca tcc tgg aaa tat aca ata ttg ggt tcc atg cca aaa		
tat ttc caa aca aaa ctg tac acc tat ttt gtg gct gca ctg agt ttg tga aat ctc tca		
tat ctt tct gat cat aac tgc atc tat gaa aag tat gag aaa gtg att tga gtg ctg agg		
aaa gaa tat aaa ata ttc act cat tgt taa gaa gga att caa aaa cat gag gtt agt tga		
aaa tgg gtc tca gag ccg agt ttc att acc caa cta ggt aac atc atc agt gca gtt ttt		
ctc tga act aac aat att ctc ttc ttt tgc ttc ctc gat cct gat cat cct ttt cac att		
gtt tta cag A GGA TTT TTT ACG TTT CGT GGG GAG ACG TAC TTA ATT GAA CCC TTG AAG		
G F F T L R G E T Y L I E P L K		-145
GTT CCC GAC AGT GAA TCC CAT GCA GTC TAC AAA TAT GAA GAT GCC AAA AAA AAG GAT GAG		
	Prodom 1 Rv/Prodom 1 Fw	
V P D S E S H A V Y K Y E D A K K K D E		-165
GCC CCC AAA ATG TGT GGG GTA ACC CTG ACT AAT TGG GAA TCA GAT AAG CCC ATC AAA AAG		
A P K M C G V T L T N W E S D K P I K K		-185
GCT TCT CAT TTA GTT GCT ACT TCT GAA gta agt ctc ata ata aac ata gtt taa gat tac		
A S H L V A T S E	 -----EoPI-6-->	-194
ata cta att tcc ttg tct tga aaa tat aaa gta aga gag aat ttc ctt tgg gaa ggg gtg		
ata gat aga att caa aag gga gaa gcc ccc att tct ata ttt tta ttg tag cca tgg cat		
aaa aga aag aat gga aac ttg agg aac aga aaa tac att ttc cag gct tat agc att ttc		
ttt ggt cat tca aac tta gtt tag aga ttt gaa tca aaa tct att taa atg agt ttc taa		
att atc tct agt ttc taa gtc aat gtt gaa aag taa tta aat tat caa ttt gga ttc ctc		
ttt tat gca tgc aga gat ggg gga caa agt ggt ttg aaa tat taa atg gtt tta aga		
tgt ctg ata agg cca tta cat aat tgt tac tcc att atc caa ttt gat ttg aat cat caa		
gtt gga ttg atg caa tga atg gat gaa aag tga caa tgt gaa cct agt cac aat tga ccc		
tta tgc tct caa tat ttt cct cct tta ttg gac gca caa aaa tta gaa aac aaa ata ttg		
cat cca aag tga cag ttc ctt tcc atc ttt ttg gtg gca aaa gtt gaa act ggc tga aaa		
atc tct act gtt ttt att aga atg tta aaa ttg aca TGG AAC AAC AGC TGT TGT GAT GAC		
	IntronB7PI Fw	
Gga ata cca aaa cac aag tga aga cgc caa atg aag cct ggt ttg tct ttt ggc ttc ttt		
cat tct ggc aat tca aga ttc ttt atc ctc agc aat ttg tgg tta tac gtt aca ttt aac		

Figure A2. Cont.

tgg tcc tta att ggt tct cta tct tga atg ttc ttg tct caa agt cag tca tga tca ctt
tat tcc gac tcc atc ttg tgc agc aga att aag aaa gtg gct gtg aga ggg gaa gga gag
aaa ttg cat tct aga att gca act tgg cct ggt agt tct gta cac ttt cat agg gaa aga
agc aat taa tgc aca ctc aca cat tca gtc tta gca aaa gtg tac ttc ctg agc aac ctg
ggt act cta gtg gag gaa tct ccc ttt gga atg cag gaa gga ggg gct ctt tcc tag aga
aat aag agg cag atc aga aat gaa tcc ttg gat tgc aga aag atg ccc gga agg gat cct
goc taa agc ctt cac agg atg tat ctc atg gtt aca cag agt gtg ctt tag tct ggc aag
gtt cct gtc tac agg aga aga gca ata aag aaa cta ctc taa gta att aac cac agt ttc
ttg ttc ctt gtg ctc aag aag tct gaa aca ata ttt cag taa tca ttt taa aat tac atc
act gaa aag aca taa ttc ttt ggc cat taa gag tga gtt tag tgg gtc aaa atg tct att
tgc ttc tcc tta cct tgt ctt tga tgc ata gtg tga tat gat tcc agg cca ggt cta att
gca **TGA CCA CAT CAG CAT GTC TCT CAC** gtg att ggt tgg cat cct ttg ttt gag aag ggg

IntronB7PI Rv

aag gaa agt tga gaa agt cat tga agc atc att ttg aca ggg tga aaa aca ggt caa aga
gaa cag ttt cct cat atg cct att aaa ttc ttt tca gag tta ggt att cat ata tac cat
tat ctt gac aat cca ttg aat aac gta ctt ttt tct tca aaa ctt tgt cag cac ttt ctc
taa gtc aga gtt ccc aaa cct ttc cag ctt tgg gga tag gtg agg gag agg gga tgg ttc
cac atg aac agt ggg gtc agg tgt gta ccc agc tct att tgt gcg agc agt ggg cac aca
tac cca ctc gtg taa aca gaa cac att cac cta tgc ttg ttc act ggt cat aca agt aga
gat gca gct gct cac ctg cca ttt cca tgg ccc agt tct aaa ggg atc aag gcc cag ggc
taa aac tct aac aaa ctg tct ccc tgt aaa tat ctt ctt gaa aga act gat att tct gga
agt tga cca gag agt aaa aca agc att ttt ctg att atc tga ggt tga cca aca ttc ttg
ttg gct gct ggg agt aac atg ttt aaa cat cca ttt ttt aat tgt tct gtt tag **CAA CAA**

2

Q Q

CAT TTT GAC CCA AGA TAC ATT CAG CTT GTC ATA GTT ACA GAC CAC GCA ATG gta agt atc
Met4 PI Rv/ Met4 PI Fw |--- EoPI-7 →

H F D P R Y I Q L V I V T D H A M 19

Metalloproteinase →

tta aat acc ttt cca ttt act ttc tgc att gag tac agc att ttg ttt ttg aat ttt tta
atg cgg gca ttt ttc aag aga tta ata cac atc cta act gac ttt ctc tta taa tgt gcc
tta tca taa act ttg ata ttt ttt gtt att gat cca acc aat att tac aaa gtt aag agt
cat ttg taa ata tat tct aat ttg cac att ttt tat ttc tgg ata cca tct aag ttt att
ttt atc cca cta tag tta aaa att tct gta ggt ttc ttt gaa atg ttc cca cac tct ttt
ttc atc caa gct aac aca cac aca cac aaa aaa aga tgt aca gtg aag aat gct ttt
ctc aca cat ctc gga aca tgg aaa tat ttt ctc caa cca aaa aaa agg ccc atg tat gct
gct gaa aag tta aaa atc tta ata tat taa cgg cac aaa tgt aga tta aaa aat cag cac
aaa aga att att tgg act gct tgg ata aaa att att aca tca aca aat tca aag atc ctt
tca cta ata tgt tct ttt tct cct ccc ttc ttt ccc ttc tta taa tta tca act tgt ctt
aac ttt ttt ttt ttt tgt ag **GTC ACG AAA AAC AAC AAT GAT TTA ACT GCT ATA ACA ACA**

Met8PI Rv

TGG ATA CAT CAA ATT GTC AAC V T K N N D L T A I T T 32
GAT ATG ATT GTG gta aga aca aat gct tgt tca ttt taa

W I H Q I V N D M I V |--- EoPI-8 → 43

act tca ctt agg ccc agc cga gat ttt gat tgt gtt aag ata aca aac tta atc agg taa
aga aag tag atg gat ttc taa atg aat ggt ttc ttt tgt ggc ttt gag tga tgt aaa act
aag tta ttt gac cag ctg gtg atc cag cat gtg agg cca tcc cag aaa atg ggt taa ttc
agt aac taa att aat aag cct cta atg gaa ata gag tca ttg ggt tgg gag gat gca aat
cca aaa gtg ttg cca gaa ttc agt cag aag tat acc tgt agt tgg att ctg tca ggc ctg
tgg tta atg gtt ctg aac ttg gag att aga aag tga tgg aca gag tca gtg tca ggc gcc
gtg cct gac aca tgc act ggg ggt gga gga gtg ata gcc gcg cgc ccc tcc ttc att gga
ggg gcc ggt aca agg acg cgg agc tcc ctc gca gct gga ttg tgg cgg caa ggc aac tga
cag ctg cag agg gtg gag gcg gcc ttt cca gtg gct ggt gga gac tgc aga gag gag
acc ggc gga caa act aag cca cag aac ttt gga gtg gtt cca gca cag aag gag gag cca
cga ggg cgg **TGT GCT TAC CCA ACA CTG AGC** Ctg gaa aaa agg acg ata tgt tta cgt ttg

Intron B16 Fw

gct gtt tgg agt tca ttc ccc ccc cct tgt ctg cca gat gct gct gcg ggg ctg tat ttt
ggg tgg tgt tag gga ggg gct agg ata ggg gta aag gct aag tgt gga gaa ggt gat ggg
aat aca tac ttg gtg gtg gat tag gaa gaa ggg cgg gtg aaa tga agg gtg gtg att ttg
tgt cac cat gct ggg agg ggt gga gcc cag gcg gaa ggg tgt ggc tgg gtg tag caa tgt
att taa tag tgt ggg atg gat atg taa gcc aac gct ggc ttt ttc cca ctt cta cgt tga
gtt ttg ttg ctg aat aaa gcg tta ttt ctt ttt gga tac ttc ccg tgc ctg tga gac tgc
tca ttg gtg agt aac tgg acg gga ggg act gac agt cag cct att cca agc tca gtt gat
tat gaa aat gat ctc tga gaa gaa acc ttg aga gaa gaa ctc aat ttg act tgg att gtg
gta ctt agg agt aga tac agt cat act gaa ttc tct gat tag cct tat tta cat ttt cac
tat tga tga tca gtt aga aac agg gag aga aca gga aga ttg agg aga taa ctc tcc tgt
tag tcc acc tac tct tct gtt cct gtt tac aga aaa taa tca cct gct tct ttt tat cat
gtt att tat ttc tct tgc tta ata cat tgt tat tat tcc ctg ata tat aga tgc ctt agt
act atc tca gct tgg ttg tta cct tgc aga cgt ttc att att tga gta ctc tac atg atg
aac caa aca gca agc tca gag cac caa gga ctt caa cct tga gct tct att gat acc tag
gtc tat ttt ttt ttt taa atc tac aaa aag agg cca atg tca gtt gct ttt taa tat tga
aat gtt tct caa ggt tta ctt tgt ttg agt ttc tga ttg atg cca tgt cca cct gat ag
ATG TAC ATA GAT TTG AAT ATT CAT ATA ACA CTG GCT GCC GTA GAA ATT TGG TCC AAT GGA
M Y I D L N I H I T L A A V E I W S N G 63
GAT TTG ATT ACT GTG ACA TCA TCA GCA CGT GAA ATT TTG AAC TCA TTT GGA GAA TGG AGA

Met5 PI Rv/Met5 PI Fw

Figure A2. Cont.

D L I T V T S S A R E I L N S F G E W R 83
 CAG AGA GAT TTG GTG AAT CGC ATA ACG CAT GAT AAT GCT CAG TTA CTC AC gta cgt ctc

Met9PI Rv |----- EoPI-9 →

Q R D L V N R I T H D N A Q L L T 100
 act gtg ggg aat ggg agt tag ggt gtc tga gag tgg agg gtt atg gga agg tta ctg ctt
 gca tag agc ttc tgt tct atg ctg tat gct tga aac cat gca tgt act aca ttt ctg ggt
 caa agt cag cca ctt ata gat gag agc tgg ctt tga gaa gtc ctt gaa tga tta cag gtg
 aaa aac ctg gat tta aaa ctc aaa ctt tta tga cta tcc tgc tat act ttt gca gtt ctt
 ata cat aca ttt taa ttc aat taa ata atg ttt tgg ttc cca ttc tgc acc aag taa atg
 aac act tgt att ata caa aag cat aca gaa tgc tat tac ctt ctc tga aat ggt gcc tat

|----- SINE/Sauria →

tta att act tgt gtt tga act gct agg gta gca gga gct gga gca agt aac agg agc tca
 ttg cat cag gca aaa ctt agg tct caa act att cac ttt cca gtc aac tag tcc act gta
 tta act gct cag ttc cac acc cca tca gcc tat ata ttg tca aac aat tta atg agt gag
 aaa ctg cta tgt tta gtg tct tat gtc tgt aac ata ctg cag aaa taa caa ttg tgt tgt
 gag caa tcc tac agg atg ctt gca taa caa gta ggt ctt atg aaa cag gga tta cat tga
 agg cag att taa gtg caa ggg aga ata tta aaa aaa tac cca cag atg gtg gtt tga ctt
 gca att caa tca ctc ttc cca aaa cca tct agt cca ata acc cca caa cta aat tga tag
 aca aaa tct agt aaa taa ata tgt att tgt taa ttt aat gga att aaa tta acc caa atg
 ggt tga att ctg aca aca tgc taa ccc tct occ caa cct cat tca tac taa tgc cat oca
 aac aca act tat agt gaa aaa tag atc ctg aat att ttt aac ggg cct gct tag tta aat
 ggc ttc tga gac tag agc cta aaa tga att cta gca ttc tga aaa ctg gta gta act gag
 gac agg tgc agg agt tta cag aaa tcc ata ctt atg tat gac cta aga cta cat aac atg
 ctt cta tat tca att tta tcc ccc tcc cct tct tct tat ag A GCC GTT AAC CTC AAT

A V N L N 105

GGT GAT ACT ATA GGA AGA GCT TAC ATG GAG AGC ATG TGT GAT CCA AAG AAA TCT GTA GGA

Met3PI Rv/Met3PI Fw

G D T I G R A Y M E S M C D P K K S V G 125
 ATT AAT CAG gtt agt aga aag gat att cta tta tct att tgt act caa gcg aaa cgt ggc

I N Q |----- EoPI-10 → 128
 ata cag aca aaa cat ctt tac caa taa agt ctc ttt ctt att ttt gag cca cgt cat ttt
 cac cca tat tta ttt gca gat ttg aca tct cca ggt cct gcg tca act aat ggc att ttg
 aca cag tgc att cta gaa caa gct ttt tta atg caa tga gct ata tgt caa gga tga gaa
 tat att ata atg ttt atg gtt cag tca aac tgt act ctg att ggc aaa tga aca ggt caa
 agc atg tta caa cac ttc caa ata atg ctt ctg aac aat agt ctt agc aat ccc aaa gac
 aaa cat gaa ttc att cca aga aat tta gtg tct aga ttg cat atg att gaa ttc tag tac
 att gag aaa aca aaa aaa tac taa atc tac tca aaa aga aaa aaa acc ctc tag tat ttt
 aag aaa acc ata tta gtt aag tga tgc tat gca ttt att gag aaa gag taa act tag ctt
 ttt gtt cac ata gaa aga atg gag aga cat ggt aat aaa caa aag tta tac aac aaa act
 cat aaa gtt ttg ttt ctt aat aag cag agt tag gat cct ggt agt agt agg tac tca taa
 gcc tac ttg ctc aag aag gtt att tta ttc aga aag agc agc tca ttc tta agt gtg ttt
 agg atg gct acc ttc att att ctg aaa atg taa gct tgt agc aaa gga cac tga gta gtt

ttt ttc aac tga cgt ttc ctg taa gtc agg gct gtc aaa ctc aat ttc att gag gga cgc

----- DNA Transposon →

atc agc att gcg gtt gcc ctc atg ggg gca gtc ggg tat ggc cag ggt ggg cac agc cca
 cag gca tgg ctg gag tgg gta tgg cta agt ttt agt aac tga ata agt gca cat agc aaa
 tgg atg cat aca ttt tga tct tat cct gtg ctg tag ctt ctg ggc ttt aaa gtt tcc ttc
 tgg atg tat ttg tgt atg ttc tgg agt cat ggt ggg cac aca tac ttt cag agg agc tag
 agg aat cct gag atg gta tcc tca act taa aat tgg tca ctg ggt cac cag ttt agc cag
 tta gtt cta ata att gga ttc act tta cgt ttc tgg cag agt aac aat aaa aaa agt att
 ctt att tct tca g GAT CAT AGC ACA GTA CAT CTT TTG GTT GCA GTT ACA ATG GCC CAT

D H S T V H L L V A V T M A H 143

GAG CTG GGT CAT AAT CTG GGT ATG GAT CAT GAA AAT CAG TGT AAT TGT GGT GCT TCC

E L G H N L G M D H 163
 GGA TGC GTT ATG TCT GGG GAG CTA AG gta agt act gag gaa tat gct taa tgg gtt ttg

Met2PI Rv |----- EoPI-11 →

G C V M S G E L R 172
 aat caa ctt att ttt aaa tgg tta caa aaa tga aaa ggt cag ttt agt tac aaa aaa gag
 gtc att tgg tca ttt gtt tgt cac ttt atc atg tgt agt taa aat gtt tca tat tta aaa
 agg tag aaa tgt ttt cag caa ttg aat agc cat cat tca cag ctt cct cca aaa cac cag
 aat tta aag aaa cag cat agt aag cag agc caa ata tta ata tct gtg atc aat aaa aga
 gga aac caa gct aaa tgt taa aga ttt tcc taa agt ggg aag cag aat ggg gaa aat tat
 tct gaa cca ggc aac acc agg aaa aaa aga taa att cgc atc ctt ggg gtt ttc cta ctt
 tgt cct ccc atc aag gta gaa tat gtg aca tct ctt tca tat act gca gtc ttc tga aat
 aaa cag ata gca atg tac agg aag tca ttg act tac agt tac ttt gaa gtt aca acg tcc
 atg aga act atg act tat gac tgg tcc tgc cag tta caa cca tca tag act ccc ttc agc
 aac aaa atc aaa att ttg gca gtt ggt cac acc caa gtg ggt tca tga tga ctg cat cat
 cct gta ttc atg gga tcc tca ttt gta cat tcc aag gtg tct tcc cat atg caa agc caa
 tag acc aaa ctg ggt acc tta aag act gtg tga ttc act taa caa agt ggc aaa act ggt
 ttt aaa aCT GAC TAT GAC TCA CTT AAC AAC TGG ctt ccc ttg gaa agt gga aat tct ggt

IntrB23PI Fw

cct agt tgt ggc agc gag tca aga act agc aat att ttc ttc gag gtg aaa gtt tag gcc
 tta ctt cat caa gaa gga aaa aga aaa gga aaa agg aag caa aat ctt ctc agc aat tgt
 att ttg caa aac tct cca taa ctt ttg tgt tca gtg ttt tgt tta ggc tca gtg agc atc

Figure A2. Cont.

tgt ttt aaa ctt ctt ata	aaa tta tct ctg tgc tac ata	att cct gat ata gat atc tct	tct ttg cat tct ttc cag	A GGA CTT TTT TCA GAA GAT TAC AGT GAG ACT CAT TAT TCC
Pro2 Fw/Pro3 Rv				
CCT GAT GGC AGC GAA ATT	ACA ACA AAC CCT CCT GTT GAG	gta ggg tct cac ttt tat gag	Y S E T H Y S	-90
P D G S E I T T N P P V E	----- EoPI-4 →			-103
cct ttt ttt aag gaa gta aat tga aac aaa tgt ttg tgc act ata tta caa ata tac aag	aat gag acc agg cta ctc aaa caa agt gta tat aag tat aaa gta tct tat att gat atg	tac tta caa aga tgc ctg gat tgt taa tcc ttg gtt aaa agc caa cat att tgg gag gtg	agt ttc aca aat aga ttt att atg aga aca tca ggt ttg taa gat tat att ttc att ttt	aaa cca gac tac agg gat aaa tgc aaa gtc ttt tat ctg taa tac caa aag tga taa caa
ttc act ttg ctc cta tac aga aat cca ttt aac atc ttt cat att aaa atg gtg cca aaa	atg gct cta tca gag gtt aaa aaa tta cag cac taa tat gct tca tgt tgg ctc cat ttc	ccc aaa ttg att taa aag tgc att ctg tgt cta ttt ctg gtt tag cat ctt cat ggg ttg	cac aaa tta ctc ctt tgt gcc atc agt ggc act ctc cca tag tgt gac ttg att tat gga	gac ttg cat tta tcc tat gtt cct ttt gca ata gtc agt att aag aag gtt ttc tgt cct
cct gaa tca aaa ttt tct gga aaa ctg ctg tct aaa tat ttc att gat gtt atg gaa tac	att gga act gaa ctt ctg ctc atc aaa tca caa tac aaa gcc tta acc agt gta gtc ctC	TTG CCT CCC TAT AG GAT CAC TGC TAT TAT TAT GGA CGC ATC CAG AAT GAT GCT GAC TCA		
IntrB13-1 Fw/Met16PI Rv				
ACT GCA AGC ATC AGC ACA	D H C Y Y G R I Q N D A D S	ACA TGC AAT GGT TTG AA	gta aga tag tct cta atc ttt tat ttg	-118
T A S I S T C N G L K	----- EoPI-5 →			-129
ttt att aat aat aat ata gtg ctc ttg gag ttc taa ttg tta aaa tga agg aca tcc tca	gtt ttt cat gga aat tag ttg ggt gtg atc cag gat ttc ggc aga att aag aca tac ttt	ggt tga aaa cca aga aga gct gct gcc agc cag gag aaa aac tat gga gct aaa tca cat	aag tct aaa gga gct tcc aag ccc cgg tct cct ttc cca ggg tga ggt gat att aca ggt	aga gaa gat tag tag gtt tca aat tgg aga cct tgc tag aaa gtg tac agg aag agg caa
gaa gtt tca gtt cta ccc aga aac act ttc ttg agt cac tct gca cac ttt ctt cag cca	act aga tat gtt aac tac ata aag atc cca gaa ttc aga agg tcc cta tca ata gta aga	atg aac atc acc tca aca tct ttt act gaa aaa aga cac tga aac tca cct ttg aac aga	gac tgt gtc cat gga gtg gag gaa taa atg aaa agc tgg aac aga gca gaa taa caa cag	aaa aat aaa gga aaa aca gaa tga cag aat aat agc att gga agg gac ttt gag gtc ttc
----- LINE/L2/CR1 →				
tag tcc aac ttc ctg ctc aag tag gag acc tat atc atc cta gac aaa tag ctg tca atc	ttc tct taa aaa gca gta gtg atg gaa cac cca caa tgt ctg aat agg tta att gtt cca	ttt gtg aga aaa tta ctc ctt agt tct aac tta ttt ctc tct ttg gtt act ttc cac gca	ttg ctt ctt ctc ctg cca tca ggt gaa gaa tag gtt gtc cca cat ttt tta tga cag cct	ctt aaa tct tta aag att atc aag tca tct cta ccc ctt ctt gtc act agc atg agt ata
ctc att gtc tgc agc cat tct aac ctc cag tta gta tgc att ctt att cct tca ttg tta	ctc ctg ttg ttc tgc att gac ttc tct atg aag atg ctt gcc aag aat tct tat ttt cat	tat tta tta aat atc ctg gtc atc ctg act ctt atc tta aat tgc tat caa act aat ctg	att tta ttt cct tga cca cag aca aat att gtt cta tac ttg ttt aaa gta aat tgc agt	att acc tat aac tct ttt tag ata ttt tag cag tta tat ttt tcc ttt ttt atc cta ctt
agt tgt gat tct tga gct tta tca gta ata tat atg ata aat ata aag tat ttt acc ctt	atg aaa taa agt ttt aca caa agc aga atg tta caa ttg gct tta gtg ttg tat tta tgt	agc tag aaa ctt att ttt tta aca tcc tgg aaa tat aca ata ttg ggt tcc atg cca aaa	tat ttc caa aca aaa ctg tac acc tat ttt gtg gct gca ctg agt ttg tga aat ctc tca	tat ctt tct gat cat aac tgc atc tat gaa aag tat gag aaa gtg att tga gtg ctg agg
aaa gaa tat aaa ata ttc act cat tgt taa gaa gga att caa aaa cat gag gtt agt tga	aaa tgg gtc tca gag ccg agt ttc att acc caa cta ggt aac atc atc agt gca gtt ttt	ctc tga act aac aat att ctc ttc ttt tgc ttc atc tct gat cat cct ttt cac att	gtt tta cag A GGA TTT TTT ACG TTT CGT GGG GAG ACG TAC TTA ATT GAA CCC TTG AAG	
GTT CCC GAC AGT GAA TCC	CAT GCA GTC TAC AAA TAT GAA GAT GCC AAA AAA AAG GAT GAG	Prodrom 1 Rv/Prodrom 1 Fw		
V P D A S E S H A V Y K Y E D A K K K D E	GCC CCC AAA ATG TGT GGG GTA ACC CTG ACT AAT TGG GAA TCA GAT AAG CCC ATC AAA AAG	-165		
A P K M C G V T L T N W E S D K P I K K	GCT TCT CAT TTA GTT GCT ACT TCT GAA	gta agt ctc ata ata aac ata gtt taa gat tac	-185	
A S H L V A T S E	----- EoPI-6 →			-194
ata cta att tcc ttg tct tga aaa tat aaa gta aga gag aat ttc ctt tgg gaa ggg gtg	ata gat aga att caa aag gga gaa gcc ccc att tct ata ttt tta ttg tag cca tgg cat	aaa aga aag aat gga aac ttg agg aac aga aaa tac att ttc cag gct tat agc att ttc	ttt ggt cat tca aac tta gtt tag aga ttt gaa tca aaa tct att taa atg agt ttc taa	att atc tct agt ttc taa gtc aat gtt gaa aag taa tta aat tat caa ttt gga ttc ctc
ttt tat gca tgc aga gag gat ggg gga caa agt ggt ttg aaa tat taa atg gtt tta aga	tgt ctg ata agg cca tta cat aat tgt tac tcc att atc caa ttt gat ttg aat cat caa	gtt gga ttg atg caa tga atg gat gaa aag tga caa tgt gaa cct agt cac aat tga ccc	tta tgc tct caa tat ttt cct cct tta ttg gac gca caa aaa tta gaa aac aaa ata ttg	cat cca aag tga cag ttc ctt tcc atc ttt ttg gtg gca aaa gtt gaa act ggc tga aaa
atc tct act gtt ttt att aga atg tta aaa ttg aca TGG AAC AAC AGC TGT TGT GAT GAC	IntronB7PI Fw			
Gga ata cca aaa cac aag tga aga cgc caa atg aag cct ggt ttg tct ttt ggc ttc ttt	cat tct ggc aat tca aga ttc ttt atc ctc agc aat ttg tgg tta tac gtt aca ttt aac			

Figure A2. Cont.

cca gct atg gat cag ttt tgg att tct tct gct aaa gcc tga aga ctt tgt tgc ctc cta
 ttt cat gca atg aat agg agc cca gta aat atg gag aat atc aca tag cca ttc ctg cag
 tgg cgt agc atg ggg gtg cag ggg ggg cag ccg cac cgg gca caa cat ctg ggg ggg cgc
 gct cgc act cgc agc tct ctg ccc ctg cct ggc tca ctc att ctc tct cca ctg aga aac
 cac gcc gga ttc ccc tca cac gac cac tca ccc ggg aaa gcc gag cga gct cgc ccc acc
 ttt tgg agc ctt ttc tct cat ctc cag cct gtt ggc aac cgc aaa ttg ttt tga gcc ctg
 ttc tct tct ctc ccc ccc cgg ccg tgt taa gcc aag gac aaa ctt tgc aag aaa ttg cag
 ttt tgc ttt ctc ttc ttc ccc ctc ccc tct ccc gta gag tag tgg ggg aaa ggg aat gtg
 gga gat ttg cca gcg gac aca gac ttt cca cta aac tcc ccc cgc ctg gca tct cca cct
 cat ttt **GGC CGC GTG AAT GCA TCT GCT TCT** ttc tct ctt ctc acc cca ccc cac cat cca

IntrF2PI Fw

ctc gtg aaa agg gag ggg gag gtg cta ata cct gga aag aaa cta act ttc att tgc caa
 ttt cat aaa tgg tgg agt taa aga gag ttg ctg gaa aat tat ata tag tta cac gtt cgg
 ttt gtg tga gga aaa caa agt gaa tgc taa ttc ctt caa ggg ggt aat ttc ttt cca gct
 gaa ctg act agc cca tct atg agc cag tgt tgt tta aag aaa act aaa ttt aca aag aaa
 tct gtt gag aaa ttc tat tga ttc tga gca tat ttc atg ggg gca aag agg aga aat tag
 atc tct tga ctc ttt cag aat ctt gct cct ttc tac act tct tta taa cag cac tgg gcc
 tgt gaa agc agg cga gaa gtc cta gag aag cta cca ttt caa tgc aga att cgc gac cca
 tta aac cta ttc ttt tat ggc agc ttc acc aca aac agc agc ttc tcc att ctc tac agg
 cag aaa aaa aat ggg aga ggg gca ttc atg cat cat ttc ctg aaa gaa tct ata cta aag
 aaa gta tgt aaa cgg ttg tct atc aaa aac tta ttc aaa agt tac gta acc aag gga tta
 tga att tgc agc aaa ttt gcc tat gaa ttt gga gta aat ttc ttt gta ggg tag caa aat
 gtg aca cca cta tat aca tat ata cag ggt tgg cca aaa tct gaa aag gaa tat agt cta
 tga aga gtt ata gac aaa tta **AGC ATC AGT TTG TTC GCA CTC AAT AAA** Gtg ttt gaa aat

Intr2F2PI Fw

aaa ctt gta ttt aga tgc att tta ctt taa tta cat cag tat ttt cac aac aaa caa tac
 atg tgc tta ggg ggt aag ggt ttt ttt **aac taa tct agt gga aga gac ttg agt gct aaa**

|----- microsatellite ----->

atc cac ggg tta ggg ggc gca aat tac ttg cct tgc ccc agg tgc tga caa ccc atg cta
cgc cac tgc att cct gtt ctt cct gaa gaa tgt ctg ggc atc acg ctt act ctt taa tag
 ttc tag aca ctt tgt att gta ttg tac aga gga gtt tgc tta gaa aac aat ttt tct caa
 tta acc cca cca aaa ggg tct gct gca act ttg act ttg gaa gaa gac cgc att gtt tgt
 tta atg gcc acc agc aga ttc tct ttg cat tct tgt ttc tac ccc tta agt tga ggg cca
 atg atg ata ttc cta atg att tga att aag tta gaa tgc cat gtt tgc atc tgt tta agt
 att cag ctt cag agt tca gtg gca tgt cgc tga tat tat cac ttt tca acc aaa gaa atc
 aat tag aaa tcc ttt aag gta aat aga aat tta gca tat tta gct ctc tct tca tcc tta
 taa aag gtc cta tcc ctt caa aag aag aaa aga ata act gga tca aaa tta cac agt ttt
 ccc ttc aaa tat att att gtc ctc cca cag aga ggc tga cta ata atg aaa ata aaa ggc
 agc cca ttc aaa cat aag acg cag ctc gat caa cta att tca cat gga ttt aaa tgt aaa
 tgg att tgc taa aaa aag aaa gga aag tta gag aaa tta aaa ctc agg **GAG CAT AAT CTG**
GAA CTA AGA TCA AGg ggt ata aag tat tct ctt atc ttt ttt gtt agc ttt aga aga aaa

Intr3F2PI Fw

aaa aat cac tgc att tct tct tta gca att cct ttt cct tat atg ttt ttg aaa tga att
 ttg ttc cta gtc tga att tgg agg atg tca tga tct ttt ttc cac ttc tac ag **G GCA**

A

CAA GAT TCC TAT CAC TTC AGT GAT TGT AGT AAG AAT GAA TAT CAG AGC TAT ATT GCT ACT 173

Met7PI Fw

Q D S Y H F S D C S K N E Y Q S Y I A T 193
TAT AAC CCA CAG TGC ATT CTC AAT CAA CCC TTG AGA ACA GAT ACT GTT TCA ACT CCA GTT
Y N P Q C I L N Q P L R T D T V S T P V 213
TCT GAA AAT GAA CTT TTG CAG gta gga gaa gaa tqt gac tqt ggc ttt cct gca tta agt

Met13PI Rv

V G E E C D C G F P A L S 220

S E N E L L Q |----- EoPI-12 ----->

|----- Spacer -----|

ctt ttt ttt taa tca aca aaa gta att tga aga ata ttc tca gaa atg aga atc ctt gaa
 L F F **STOP**
 aaa tca tct agc ttt cta agt ggt ttg agc cat cca aga ggt tgg ctt gtg aat ggc tga
 ggt ttg tgc ctt tca tgt aca tgc atg tat gaa atg gtt tct tgg gtt gta gag gaa tag
 aga aat ggt atc tca cta cta ttt ggg gaa gat ggt gaa ttt tta aaa agg ggt gat tga
 cca ttc cat gaa aat ctt tcc ctc ctg aaa acc cct att ttg ttg ata tag cca cat tat
 cct gtc cca caa ttt tct oga act gct cct tcc cat atc tga tta tct tta atc tat gct
 ctg atc cta ata ata ttt tta taa gaa cag taa tat agt gtt ttt atg ttg tta aat aca
 cct gta atg gtc tgt gag aat gtc ctt aag aga caa aag aag gag gaa aca tcc agt cag
 tgg tca tat aaa aag gag att aac ctg cag aaa caa agg ggc ata gca aaa atc tca aga
 ggg aca cct cct acc cat tct ctt ggt cgg taa aga tga ggt ggt aga aat gca ctt tca
 gta ttg aaa gat tct gct act gta act gta caa tca agg tag tgt taa tgc tca tgg ttg
 gtg ctt ctt ctc tgg att acc tca aaa gct ggc atg atg agt aga aaa tct ctc atg aat
 aag aaa gga atg ggg ctg tag gtt atg tgg gct tca caa tta gga gat gag gat att ttt
 gtt tta ttc ttt tca ogt agg aaa tat cag ata agg ctc ttt cgc aga gaa atg oca tta
 tct gtt ttc aat aac aaa caa ttt ttg cat ttg cta gca tga acc cat aaa agg gaa cac
 att gca gaa att tcc ctc ttc aaa ata gac caa tta aaa aag aaa att cta tgc cat cat
 ttg ata tga tat **ccg tgc tgt gat cct ata acg tgt aaa cca aga caa ggg aaa cat**
 * Y D **P C C D P I T C K P R Q G K H**
tgt gta tct gga ctg tgt tgt tgt agc tac aaa gta aga ctt gtt tat ttt taa cac cag
C V S G L C C C S Y K V R L V Y F *
 gag aaa ttt tac cct gct cca tac tag cca tgt aga aat gta ata ttt ctt ggc ttt tta

Figure A2. Cont.

tgt ttt aaa ctt ctt ata	aaa tta tct ctg tgc tac ata	att cct gat ata gat atc tct	tct ttg cat tct ttc cag	A GGA CTT TTT TCA GAA GAT TAC AGT GAG ACT CAT TAT TCC
Pro2 Fw/Pro3 Rv				
CCT GAT GGC AGC GAA ATT	ACA ACA AAC CCT CCT GTT GAG	gta ggg tct cac ttt tat gag	Y S E T H Y S	-90
P D G S E I T T N P P V E	----- EoPI-4 →			-103
cct ttt ttt aag gaa gta aat tga aac aaa tgt ttg tgc act ata tta caa ata tac aag	aat gag acc agg cta ctc aaa caa agt gta tat aag tat aaa gta tct tat att gat atg	tac tta caa aga tgc ctg gat tgt taa tcc ttg gtt aaa agc caa cat att tgg gag gtg	agt ttc aca aat aga ttt att atg aga aca tca ggt ttg taa gat tat att ttc att ttt	aaa cca gac tac agg gat aaa tgc aaa gtc ttt tat ctg taa tac caa aag tga taa caa
ttc act ttg ctc cta tac aga aat cca ttt aac atc ttt cat att aaa atg gtg cca aaa	atg gct cta tca gag gtt aaa aaa tta cag cac taa tat gct tca tgt tgg ctc cat ttc	ccc aaa ttg att taa aag tgc att ctg tgt cta ttt ctg gtt tag cat ctt cat ggg ttg	cac aaa tta ctc ctt tgt gcc atc agt ggc act ctc cca tag tgt gac ttg att tat gga	gac ttg cat tta tcc tat gtt cct ttt gca ata gtc agt att aag aag gtt ttc tgt cct
cct gaa tca aaa ttt tct gga aaa ctg ctg tct aaa tat ttc att gat gtt atg gaa tac	att gga act gaa ctt ctg ctc atc aaa tca caa tac aaa gcc tta acc agt gta gtc ctC	TTG CCT CCC TAT AG GAT CAC TGC TAT TAT TAT GGA CGC ATC CAG AAT GAT GCT GAC TCA		
IntrB13-1 Fw/Met16PI Rv				
ACT GCA AGC ATC AGC ACA	D H C Y Y G R I Q N D A D S	ACA TGC AAT GGT TTG AA	gta aga tag tct cta atc ttt tat ttg	-118
T A S I S T C N G L K	----- EoPI-5 →			-129
ttt att aat aat aat ata gtg ctc ttg gag ttc taa ttg tta aaa tga agg aca tcc tca	gtt ttt cat gga aat tag ttg ggt gtg atc cag gat ttc ggc aga att aag aca tac ttt	ggt tga aaa cca aga aga gct gct gcc agc cag gag aaa aac tat gga gct aaa tca cat	aag tct aaa gga gct tcc aag ccc cgg tct cct ttc cca ggg tga ggt gat att aca ggt	aga gaa gat tag tag gtt tca aat tgg aga cct tgc tag aaa gtg tac agg aag agg caa
gaa gtt tca gtt cta ccc aga aac act ttc ttg agt cac tct gca cac ttt ctt cag cca	act aga tat gtt aac tac ata aag atc cca gaa ttc aga agg tcc cta tca ata gta aga	atg aac atc acc tca aca tct ttt act gaa aaa aga cac tga aac tca cct ttg aac aga	gac tgt gtc cat gga gtg gag gaa taa atg aaa agc tgg aac aga gca gaa taa caa cag	aaa aat aaa gga aaa aca gaa tga cag aat aat agc att gga agg gac ttt gag gtc ttc
----- LINE/L2/CR1 →				
tag tcc aac ttc ctg ctc aag tag gag acc tat atc atc cta gac aaa tag ctg tca atc	ttc tct taa aaa gca gta gtg atg gaa cac cca caa tgt ctg aat agg tta att gtt cca	ttt gtg aga aaa tta ctc ctt agt tct aac tta ttt ctc tct ttg gtt act ttc cac gca	ttg ctt ctt ctc ctg cca tca ggt gaa gaa tag gtt gtc cca cat ttt tta tga cag cct	ctt aaa tcc tta aag att atc aag tca tct cta ccc ctt ctt gtc act agc atg agt ata
ctc att gtc tgc agc cat tct aac ctc cag tta gta tgc att ctt att cct tca ttg tta	ctc ctg ttg ttc tgc att gac ttc tct atg aag atg ctt gcc aag aat tct tat ttt cat	tat tta tta aat atc ctg gtc atc ctg act ctt atc tta aat tgc tat caa act aat ctg	att tta ttt cct tga cca cag aca aat att gtt cta tac ttg ttt aaa gta aat tgc agt	att acc tat aac tct ttt tag ata ttt tag cag tta tat ttt tcc ttt ttt atc cta ctt
agt tgt gat tct tga gct tta tca gta ata tat atg ata aat ata aag tat ttt acc ctt	atg aaa taa agt ttt aca caa agc aga atg tta caa ttg gct tta gtg ttg tat tta tgt	agc tag aaa ctt att ttt tta aca tcc tgg aaa tat aca ata ttg ggt tcc atg cca aaa	tat ttc caa aca aaa ctg tac acc tat ttt gtg gct gca ctg agt ttg tga aat ctc tca	tat ctt tct gat cat aac tgc atc tat gaa aag tat gag aaa gtg att tga gtg ctg agg
aaa gaa tat aaa ata ttc act cat tgt taa gaa gga att caa aaa cat gag gtt agt tga	aaa tgg gtc tca gag ccg agt ttc att acc caa cta ggt aac atc atc agt gca gtt ttt	ctc tga act aac aat att ctc ttc ttt tgc ttc atc tct gat cat cct ttt cac att	gtt tta cag A GGA TTT TTT ACG TTT CGT GGG GAG ACG TAC TTA ATT GAA CCC TTG AAG	
GTT CCC GAC AGT GAA TCC	CAT GCA GTC TAC AAA TAT GAA GAT GCC AAA AAA AAG GAT GAG	Prodrom 1 Rv/Prodrom 1 Fw		
V P D A S E S H A V Y K Y E D A K K K D E	GCC CCC AAA ATG TGT GGG GTA ACC CTG ACT AAT TGG GAA TCA GAT AAG CCC ATC AAA AAG	A P K M C G V T L T N W E S D K P I K K	-165	
GCT TCT CAT TTA GTT GCT ACT TCT GAA	gta agt ctc ata ata aac ata gtt taa gat tac	----- EoPI-6 →		
A S H L V A T S E				-185
ata cta att tcc ttg tct tga aaa tat aaa gta aga gag aat ttc ctt tgg gaa ggg gtg	ata gat aga att caa aag gga gaa gcc ccc att tct ata ttt tta ttg tag cca tgg cat	aaa aga aag aat gga aac ttg agg aac aga aaa tac att ttc cag gct tat agc att ttc	ttt ggt cat tca aac tta gtt tag aga ttt gaa tca aaa tct att taa atg agt ttc taa	att atc tct agt ttc taa gtc aat gtt gaa aag taa tta aat tat caa ttt gga ttc ctc
ttt tat gca tgc aga gat ggg gga caa agt ggt ttg aaa tat taa atg gtt tta aga	tgt ctg ata agg cca tta cat aat tgt tac tcc att atc caa ttt gat ttg aat cat caa	gtt gga ttg atg caa tga atg gat gaa aag tga caa tgt gaa cct agt cac aat tga ccc	tta tgc tct caa tat ttt cct cct tta ttg gac gca caa aaa tta gaa aac aaa ata ttg	cat cca aag tga cag ttc ctt tcc atc ttt ttg gtg gca aaa gtt gaa act ggc tga aaa
atc tct act gtt ttt att aga atg tta aaa ttg aca TGG AAC AAC AGC TGT TGT TAT GAC	IntronB7PI Fw			
Gga	ata cca aaa cac aag tga aga cgc caa atg aag cct ggt ttg tct ttt ggc ttc ttt	cat tct ggc aat tca aga ttc ttt atc ctc agc aat ttg tgg tta tac gtt aca ttt aac		

Figure A2. Cont.

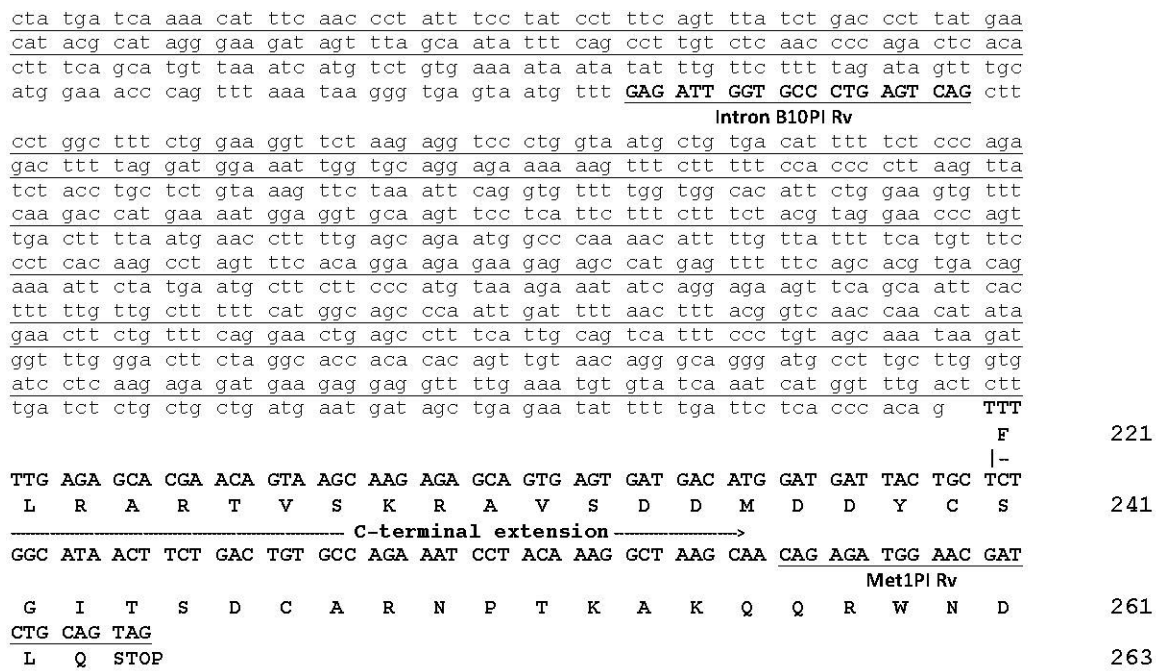


Figure A2. Genomic organization of *E. ocellatus* EOC00028-like PI-SVMP gene.

References

- MEROPS- the Peptidase Database: Family M72. Available online: <https://merops.sanger.ac.uk/cgi-bin/famsum?family=M12> (accessed on 8 May 2016).
- PFAM Family: REprolysin. Available online: <http://pfam.xfam.org/family/PF01421> (accessed on 8 May 2016).
- Seals, D.F.; Courtneidge, S.A. The ADAMs family of metalloproteases: Multidomain proteins with multiple functions. *Genes Dev.* **2003**, *17*, 7–30. [[CrossRef](#)] [[PubMed](#)]
- Giebeler, N.; Zigrino, P. A disintegrin and metalloprotease (ADAM): Historical overview of their functions. *Toxins* **2016**, *8*. [[CrossRef](#)] [[PubMed](#)]
- Arendt, D.; Technau, U.; Wittbrodt, J. Evolution of the bilaterian larval foregut. *Nature* **2001**, *409*, 81–85. [[CrossRef](#)] [[PubMed](#)]
- Tucker, R.P.; Adams, J.C. Adhesion Networks of Cnidarians: A Postgenomic View. *Int. Rev. Cell Mol. Biol.* **2014**, *308*, 323–377. [[PubMed](#)]
- Bahudhanapati, H.; Bhattacharya, S.; Wei, S. Evolution of Vertebrate Adam Genes; Duplication of Testicular Adams from Ancient Adam9/9-like Loci. *PLoS ONE* **2015**, *10*, e0136281. [[CrossRef](#)] [[PubMed](#)]
- Cho, C. Testicular and epididymal ADAMs: Expression and function during fertilization. *Nat. Rev. Urol.* **2012**, *9*, 550–560. [[CrossRef](#)] [[PubMed](#)]
- Bates, E.E.; Fridman, W.H.; Mueller, C.G. The ADAMDEC1 (decysin) gene structure: Evolution by duplication in a metalloprotease gene cluster on chromosome 8p12. *Immunogenetics* **2002**, *54*, 96–105. [[CrossRef](#)] [[PubMed](#)]
- Wei, S.; Whittaker, C.A.; Xu, G.; Bridges, L.C.; Shah, A.; White, J.M.; Desimone, D.W. Conservation and divergence of ADAM family proteins in the *Xenopus* genome. *BMC Evol. Biol.* **2010**, *10*. [[CrossRef](#)] [[PubMed](#)]
- Taylor, J.S.; Raes, J. Duplication and divergence: The evolution of new genes and old ideas. *Annu. Rev. Genet.* **2004**, *38*, 615–643. [[CrossRef](#)] [[PubMed](#)]
- Ohno, S. *Evolution by Gene Duplication*; Springer-Verlag: Berlin-Heidelberg, Germany, 1970.
- Zhang, J. Evolution by gene duplication: An update. *Trends Ecol. Evol.* **2003**, *18*, 292–298. [[CrossRef](#)]
- True, J.R.; Carroll, S.B. Gene co-option in physiological and morphological evolution. *Annu. Rev. Cell Dev. Biol.* **2002**, *18*, 53–80. [[CrossRef](#)] [[PubMed](#)]

15. Kaessmann, H.; Vinckenbosch, N.; Long, M. RNA-based gene duplication: Mechanistic and evolutionary insights. *Nat. Rev. Genet.* **2009**, *10*, 19–31. [[CrossRef](#)] [[PubMed](#)]
16. Vinckenbosch, N.; Dupanloup, I.; Kaessmann, H. Evolutionary fate of retroposed gene copies in the human genome. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 3220–3225. [[CrossRef](#)] [[PubMed](#)]
17. Fry, B.; Wüster, W. Assembling an Arsenal: Origin and evolution of the snake venom proteome inferred from phylogenetic analysis of toxin sequences. *Mol. Biol. Evol.* **2004**, *21*, 870–883. [[CrossRef](#)] [[PubMed](#)]
18. Fry, B.G.; Vidal, N.; Norman, J.A.; Vonk, F.J.; Scheib, H.; Ramjan, S.F.; Kuruppu, S.; Fung, K.; Hedges, S.B.; Richardson, M.K.; et al. Early evolution of the venom system in lizards and snakes. *Nature* **2006**, *439*, 584–588. [[CrossRef](#)] [[PubMed](#)]
19. Fry, B.G.; Casewell, N.R.; Wüster, W.; Vidal, N.; Young, B.; Jackson, N. The structural and functional diversification of the Toxicofera reptile venom system. *Toxicon* **2012**, *60*, 434–448. [[CrossRef](#)] [[PubMed](#)]
20. Casewell, N.R.; Wüster, W.; Vonk, F.J.; Harrison, R.A.; Fry, B.G. Complex cocktails: The evolutionary novelty of venoms. *Trends Ecol. Evol.* **2013**, *28*, 219–229. [[CrossRef](#)] [[PubMed](#)]
21. Haney, R.A.; Clarke, T.H.; Gadgil, R.; Fitzpatrick, R.; Hayashi, C.Y.; Ayoub, N.A.; Garb, J.E. Effects of gene duplication, positive selection, and shifts in gene expression on the evolution of the venom gland transcriptome in widow spiders. *Genome Biol. Evol.* **2016**, *8*, 228–242. [[CrossRef](#)] [[PubMed](#)]
22. Wong, E.S.; Belov, K. Venom evolution through gene duplications. *Gene* **2012**, *496*, 1–7. [[CrossRef](#)] [[PubMed](#)]
23. Vonk, F.J.; Casewell, N.R.; Henkel, C.V.; Heimberg, A.M.; Jansen, H.J.; McCleary, R.J.; Kerckamp, H.M.; Vos, R.A.; Guerreiro, I.; Calvete, J.J.; et al. The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 20651–20656. [[CrossRef](#)] [[PubMed](#)]
24. Hedges, S.B.; Vidal, N. Lizards, snakes, and amphisbaenians (Squamata). In *The Timetree of Life*; Hedges, S.B., Kumar, S., Eds.; Oxford University Press: Oxford, UK, 2009; pp. 383–389.
25. Jones, M.E.; Anderson, C.L.; Hipsley, C.A.; Müller, J.; Evans, S.E.; Schoch, R.R. Integration of molecules and new fossils supports a Triassic origin for Lepidosauria (lizards, snakes, and tuatara). *BMC Evol. Biol.* **2013**, *13*. [[CrossRef](#)] [[PubMed](#)]
26. Pylon, R.A.; Burbrink, F.T.; Wiens, J.J. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. *BMC Evol. Biol.* **2013**, *13*. [[CrossRef](#)] [[PubMed](#)]
27. Reeder, T.W.; Townsend, T.M.; Mulcahy, D.G.; Noonan, B.P.; Wood, P.L., Jr.; Sites, J.W., Jr.; Wiens, J.J. Integrated analyses resolve conflicts over squamate reptile phylogeny and reveal unexpected placements for fossil. *PLoS ONE* **2015**, *10*, e0118199. [[CrossRef](#)] [[PubMed](#)]
28. Hsiang, A.Y.; Field, D.J.; Webster, T.H.; Behlke, A.D.; Davis, M.B.; Racicot, R.A.; Gauthier, J.A. The origin of snakes: Revealing the ecology, behavior, and evolutionary history of early snakes using genomics, phenomics, and the fossil record. *BMC Evol. Biol.* **2015**, *15*. [[CrossRef](#)] [[PubMed](#)]
29. Reeks, T.A.; Fry, B.G.; Alewood, P.F. Privileged frameworks from snake venom. *Cell. Mol. Life Sci.* **2015**, *72*, 1939–1958. [[CrossRef](#)] [[PubMed](#)]
30. Hite, L.A.; Jia, L.G.; Bjarnason, J.B.; Fox, J.W. cDNA sequences for four snake venom metalloproteinases: Structure, classification, and their relationship to mammalian reproductive proteins. *Arch. Biochem. Biophys.* **1994**, *308*, 182–191. [[CrossRef](#)] [[PubMed](#)]
31. Moura-da-Silva, A.M.; Theakston, R.D.; Crampton, J.M. Evolution of disintegrin cysteine-rich and mammalian matrix-degrading metalloproteinases: Gene duplication and divergence of a common ancestor rather than convergent evolution. *J. Mol. Evol.* **1996**, *43*, 263–269. [[CrossRef](#)] [[PubMed](#)]
32. Casewell, N.R. On the ancestral recruitment of metalloproteinases into the venom of snakes. *Toxicon* **2012**, *60*, 449–454. [[CrossRef](#)] [[PubMed](#)]
33. Escalante, T.; Rucavado, A.; Fox, J.W.; Gutiérrez, J.M. Key events in microvascular damage induced by snake venom hemorrhagic metalloproteinases. *J. Proteomics* **2011**, *74*, 1781–1794. [[CrossRef](#)] [[PubMed](#)]
34. Markland, F.S., Jr.; Swenson, S. Snake venom metalloproteinases. *Toxicon* **2013**, *62*, 3–18. [[CrossRef](#)] [[PubMed](#)]
35. Herrera, C.; Escalante, T.; Voisin, M.B.; Rucavado, A.; Morazán, D.; Macêdo, J.K.; Calvete, J.J.; Sanz, L.; Nourshargh, S.; Gutiérrez, J.M.; et al. Tissue localization and extracellular matrix degradation by PI, PII and PIII snake venom metalloproteinases: Clues on the mechanisms of venom-induced hemorrhage. *PLoS Negl. Trop. Dis.* **2015**, *9*, e0003731. [[CrossRef](#)] [[PubMed](#)]
36. Gutiérrez, J.M.; Escalante, T.; Rucavado, A.; Herrera, C. Hemorrhage Caused by Snake Venom Metalloproteinases: A Journey of Discovery and Understanding. *Toxins* **2016**, *8*. [[CrossRef](#)] [[PubMed](#)]

37. Jia, L.G.; Shimokawa, K.; Bjarnason, J.B.; Fox, J.W. Snake venom metalloproteinases: Structure, function and relationship to the ADAMs family of proteins. *Toxicon* **1996**, *34*, 1269–1276. [[CrossRef](#)]
38. Pyron, R.A.; Burnbrink, F.T. Extinction ecological opportunity and the origins of global snake diversity. *Evolution* **2012**, *66*, 163–178. [[CrossRef](#)] [[PubMed](#)]
39. Fox, J.W.; Serrano, S.M. Structural considerations of the snake venom metalloproteinases, key members of the M12 repolysin family of metalloproteinases. *Toxicon* **2005**, *45*, 969–985. [[CrossRef](#)] [[PubMed](#)]
40. Fox, J.W.; Serrano, S.M. Insights into and speculations about snake venom metalloproteinase (SVMP) synthesis, folding and disulfide bond formation and their contribution to venom complexity. *FEBS J.* **2008**, *275*, 3016–3030. [[CrossRef](#)] [[PubMed](#)]
41. Casewell, N.R.; Sunagar, K.; Takacs, Z.; Calvete, J.J.; Jackson, T.N.W.; Fry, B.G. Snake venom metalloprotease enzymes. In *Venomous Reptiles and Their Toxins: Evolution, Pathophysiology and Biodiscovery*; ISBN: 978-0-19-930939-9. Fry, B.G., Ed.; Oxford University Press: Oxford, UK, 2015; Chapter 23; pp. 347–363.
42. Juárez, P.; Comas, I.; González-Candelas, F.; Calvete, J.J. Evolution of snake venom disintegrins by positive Darwinian selection. *Mol. Biol. Evol.* **2008**, *25*, 2391–2407. [[CrossRef](#)] [[PubMed](#)]
43. Carbajo, R.J.; Sanz, L.; Pérez, A.; Calvete, J.J. NMR structure of bitistatin—A missing piece in the evolutionary pathway of snake venom disintegrins. *FEBS J.* **2015**, *282*, 341–360. [[CrossRef](#)] [[PubMed](#)]
44. Calvete, J.J. Brief History and Molecular Determinants of Snake Venom Disintegrin Evolution. In *Toxins and Hemostasis. From Bench to Bedside*; Kini, R.M., Markland, F., McLane, M.A., Morita, T., Eds.; Springer Science+Business Media B.V.: Amsterdam, The Netherlands, 2010; pp. 285–300.
45. Casewell, N.R.; Wagstaff, S.C.; Harrison, R.A.; Renjifo, C.; Wüster, W. Domain loss facilitates accelerated evolution and neofunctionalization of duplicate snake venom metalloproteinase toxin genes. *Mol. Biol. Evol.* **2011**, *28*, 2637–2649. [[CrossRef](#)] [[PubMed](#)]
46. Sanz-Soler, R.; Sanz, L.; Calvete, J.J. Distribution of *RPTLN* genes across Reptilia. Hypothesized role for *RPTLN* in the evolution of SVMPs. *Integr. Compar. Biol.* **2016**. [[CrossRef](#)] [[PubMed](#)]
47. Sanz, L.; Harrison, R.A.; Calvete, J.J. First draft of the genomic organization of a PIII-SVMP gene. *Toxicon* **2012**, *60*, 455–469. [[CrossRef](#)] [[PubMed](#)]
48. Endo, T.; Fedorov, A.; de Souza, S.J.; Gilbert, W. Do Introns Favor or Avoid Regions of Amino Acid Conservation? *Mol. Biol. Evol.* **2002**, *19*, 521–525. [[CrossRef](#)] [[PubMed](#)]
49. Zhou, Q.; Wang, W. On the origin and evolution of new genes—a genomic and experimental perspective. *J. Genet. Genomics* **2008**, *35*, 639–648. [[CrossRef](#)]
50. Kordis, D.; Gubensek, F. Adaptive evolution of animal toxin multigene families. *Gene* **2000**, *261*, 43–52. [[CrossRef](#)]
51. Cao, Z.; Yu, Y.; Wu, Y.; Hao, P.; Di, Z.; He, Y.; Chen, Z.; Yang, W.; Shen, Z.; He, X.; et al. The genome of *Mesobuthus martensii* reveals a unique adaptation model of arthropods. *Nat. Commun.* **2013**, *4*. [[CrossRef](#)] [[PubMed](#)]
52. Fry, B.G.; Wüster, W.; Kini, R.M.; Brusich, V.; Khan, A.; Venkataraman, D.; Rooney, A.P. Molecular evolution and phylogeny of elapid snake venom three-finger toxins. *J. Mol. Evol.* **2003**, *57*, 110–129. [[CrossRef](#)] [[PubMed](#)]
53. Reyes-Velasco, J.; Card, D.C.; Andrew, A.L.; Shaney, K.J.; Adams, R.H.; Schield, D.R.; Casewell, N.R.; Mackessy, S.P.; Castoe, T.A. Expression of venom gene homologs in diverse python tissues suggests a new model for the evolution of snake venom. *Mol. Biol. Evol.* **2015**, *32*, 173–183. [[CrossRef](#)] [[PubMed](#)]
54. Hargreaves, A.D.; Swain, M.T.; Logan, D.W.; Mulley, J.F. Testing the Toxicofera: Comparative transcriptomics casts doubt on the single, early evolution of the reptile venom system. *Toxicon* **2014**, *92*, 140–156. [[CrossRef](#)] [[PubMed](#)]
55. Sunagar, K.; Jackson, T.N.; Undheim, E.A.; Ali, S.A.; Antunes, A.; Fry, B.G. Three-fingered RAVERs: Rapid Accumulation of Variations in Exposed Residues of snake venom toxins. *Toxins* **2013**, *5*, 2172–2208. [[CrossRef](#)] [[PubMed](#)]
56. Chang, D.; Duda, T.F. Extensive and continuous duplication facilitates rapid evolution and diversification of gene families. *Mol. Biol. Evol.* **2012**, *29*, 2019–2029. [[CrossRef](#)] [[PubMed](#)]
57. Chow, L.T.; Gelinis, R.E.; Broker, T.R.; Roberts, R.J. An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* **1977**, *12*, 1–8. [[CrossRef](#)]
58. Berget, S.M.; Moore, C.; Sharp, P.A. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 3171–3175. [[CrossRef](#)] [[PubMed](#)]

59. Bicknell, A.A.; Cenik, C.; Chua, H.N.; Roth, F.P.; Moore, M.J. Introns in UTRs: Why we should stop ignoring them. *BioEssays* **2012**, *34*, 1025–1034. [[CrossRef](#)] [[PubMed](#)]
60. Cenik, C.; Chua, H.N.; Zhang, H.; Tarnawsky, S.P.; Akef, A.; Derti, A.; Tasan, M.; Moore, M.J.; Palazzo, A.F.; Roth, F.P. Genome analysis reveals interplay between 5'-UTR introns and nuclear mRNA export for secretory and mitochondrial genes. *PLoS Genet.* **2011**, *7*, e1001366. [[CrossRef](#)] [[PubMed](#)]
61. Comeron, J.M.; Kreitman, M. The correlation between intron length and recombination in *Drosophila*: Dynamic equilibrium between mutational and selective forces. *Genetics* **2000**, *156*, 1175–1190. [[PubMed](#)]
62. De Souza, S.J.; Long, M.; Gilbert, W. Introns and gene evolution. *Genes Cells* **1996**, *1*, 493–505. [[CrossRef](#)] [[PubMed](#)]
63. Patthy, L. Exon shuffling and other ways of module exchange. *Matrix Biol.* **1996**, *15*, 301–310. [[CrossRef](#)]
64. Hughes, A.L.; Hughes, M.K. Small genomes for better flyers. *Nature* **1995**, *377*, 391. [[CrossRef](#)] [[PubMed](#)]
65. Lynch, M. Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 6118–6123. [[CrossRef](#)] [[PubMed](#)]
66. Haddrill, P.R.; Charlesworth, B.; Halligan, D.L.; Andolfatto, P. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biol.* **2005**, *6*, R67. [[CrossRef](#)] [[PubMed](#)]
67. Zhao, M.; He, L.; Gu, Y.; Wang, Y.; Chen, Q.; He, C. Genome-wide analyses of a plant-specific LIM-domain gene family implicate its evolutionary role in plant diversification. *Genome Biol. Evol.* **2014**, *6*, 1000–1012. [[CrossRef](#)] [[PubMed](#)]
68. Tordai, H.; Patthy, L. Insertion of spliceosomal introns in proto-splice sites: The case of secretory signal peptides. *FEBS Lett.* **2004**, *575*, 109–111. [[CrossRef](#)] [[PubMed](#)]
69. Tomita, M.; Shimizu, N.; Brutlag, D.L. Introns and reading frames: Correlation between splicing sites and their codon positions. *Mol. Biol. Evol.* **1996**, *13*, 1219–1223. [[CrossRef](#)] [[PubMed](#)]
70. Long, M.; de Souza, S.J.; Rosenberg, C.; Gilbert, W. Relationship between proto-splice sites and intron phases: Evidence from dicodon analysis. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 219–223. [[CrossRef](#)] [[PubMed](#)]
71. Von Heijne, G. Patterns of amino acids near signal-sequence cleavage sites. *Eur. J. Biochem.* **1983**, *133*, 17–21. [[CrossRef](#)] [[PubMed](#)]
72. Pinho, C.; Rocha, S.; Carvalho, B.M.; Lopes, S.; Mourão, S.; Vallinoto, M.; Brunes, T.O.; Haddad, C.F.B.; Gonçalves, H.; Sequeira, F.; et al. New primers for the amplification and sequencing of nuclear *loci* in a taxonomically wide set of reptiles and amphibians. *Conserv. Genet. Resour.* **2010**, *2*, 181–185. [[CrossRef](#)]
73. Ellegren, H. Microsatellites: Simple sequences with complex evolution. *Nat. Rev. Genet.* **2004**, *5*, 435–445. [[CrossRef](#)] [[PubMed](#)]
74. Adams, R.H.; Blackmon, H.; Reyes-Velasco, J.; Schield, D.R.; Card, D.C.; Andrew, A.L.; Waynewood, N.; Castoe, T.A. Microsatellite landscape evolutionary dynamics across 450 million years of vertebrate genome evolution. *Genome* **2016**, *59*, 295–310. [[CrossRef](#)] [[PubMed](#)]
75. Liang, K.C.; Tseng, J.T.; Tsai, S.J.; Sun, H.S. Characterization and distribution of repetitive elements in association with genes in the human genome. *Comput. Biol. Chem.* **2015**, *57*, 29–38. [[CrossRef](#)] [[PubMed](#)]
76. Schlötterer, C.; Tautz, D. Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.* **1992**, *20*, 211–215. [[CrossRef](#)] [[PubMed](#)]
77. Charlesworth, B.; Sniegowski, P.; Stephan, W. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* **1994**, *371*, 215–220. [[CrossRef](#)] [[PubMed](#)]
78. Martin, P.; Makepeace, K.; Hill, S.A.; Hood, D.W.; Moxon, E.R. Microsatellite instability regulates transcription factor binding and gene expression. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 3800–3804. [[CrossRef](#)] [[PubMed](#)]
79. Li, Y.C.; Korol, A.B.; Fahima, T.; Beiles, A.; Nevo, E. Microsatellites: Genomic distribution, putative functions and mutational mechanisms: A review. *Mol. Ecol.* **2002**, *11*, 2453–2465. [[CrossRef](#)] [[PubMed](#)]
80. Shaney, K.J.; Schield, D.R.; Card, D.C.; Ruggiero, R.P.; Pollock, D.D.; Mackessy, S.P.; Castoe, T.A. Squamate reptile genomics and evolution. In *Handbook of Toxinology: Venom Genomics and Proteomics*; Gopalakrishnakone, P., Calvete, J.J., Eds.; Springer Science+Business Media: Dordrecht, The Netherlands, 2016; pp. 29–49.
81. Balaesque, P.; King, T.E.; Parkin, E.J.; Heyer, E.; Carvalho-Silva, D.; Kraaijenbrink, T.; de Knijff, P.; Tyler-Smith, C.; Jobling, M.A. Gene conversion violates the stepwise mutation model for microsatellites in Y-chromosomal palindromic repeats. *Hum. Mutat.* **2014**, *35*, 609–617. [[CrossRef](#)] [[PubMed](#)]

82. Eller, C.D.; Regelson, M.; Merriman, B.; Nelson, S.; Horvath, S.; Marahrens, Y. Repetitive sequence environment distinguishes housekeeping genes. *Gene* **2007**, *390*, 153–165. [[CrossRef](#)] [[PubMed](#)]
83. Sverdlov, E.D. Perpetually mobile footprints of ancient infections in human genome. *FEBS Lett.* **1998**, *428*, 1–6. [[CrossRef](#)]
84. Makalowski, W. Genomic scrap yard: How genomes utilize all that junk. *Gene* **2000**, *259*, 61–67. [[CrossRef](#)]
85. Bourque, G.; Leong, B.; Vega, V.B.; Chen, X.; Lee, Y.L.; Srinivasan, K.G.; Chew, J.L.; Ruan, Y.; Wei, C.L.; Ng, H.H.; et al. Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome. Res.* **2008**, *18*, 1752–1762. [[CrossRef](#)] [[PubMed](#)]
86. Irimía, M.; Rukov, J.L.; Penny, D.; Vinther, J.; García-Fernández, J.; Roy, S.W. Origin of introns by ‘intronization’ of exonic sequences. *Trends Genet.* **2008**, *24*, 378–381. [[CrossRef](#)] [[PubMed](#)]
87. Bazaa, A.; Juarez, P.; Marrakchi, N.; Bel Lasfer, Z.; El Ayeb, M.; Harrison, R.A.; Calvete, J.J.; Sanz, L. Loss of introns along the evolutionary diversification pathway of snake venom disintegrins evidenced by sequence analysis of genomic DNA from *Macrovipera lebetina transmediterranea* and *Echis ocellatus*. *J. Mol. Evol.* **2007**, *64*, 261–271. [[CrossRef](#)] [[PubMed](#)]
88. Bonen, L.; Vogel, J. The ins and outs of group II introns. *Trends Genet.* **2001**, *17*, 322–331. [[CrossRef](#)]
89. Dibb, N.J.; Newman, A.J. Evidence that introns arose at proto-splice sites. *EMBO J.* **1989**, *8*, 2015–2021. [[PubMed](#)]
90. Logsdon, J.M., Jr. The recent origins of spliceosomal introns revisited. *Curr. Opin. Genet. Dev.* **1998**, *8*, 637–648. [[CrossRef](#)]
91. Lynch, M.; Richardson, A.O. The evolution of spliceosomal introns. *Curr. Opin. Genet. Dev.* **2002**, *12*, 701–710. [[CrossRef](#)]
92. Patel, A.A.; Steitz, J.A. Splicing double: Insights from the second spliceosome. *Nat. Rev. Mol. Cell Biol.* **2003**, *4*, 960–970. [[CrossRef](#)] [[PubMed](#)]
93. Rodríguez-Trelles, F.; Tarrío, R.; Ayala, F.J. Origins and evolution of spliceosomal introns. *Annu. Rev. Genet.* **2006**, *40*, 47–76. [[CrossRef](#)] [[PubMed](#)]
94. Luo, Y.; Li, C.; Gong, X.; Wang, Y.; Zhang, K.; Cui, Y.; Sun, Y.E.; Li, S. Splicing-related features of introns serve to propel evolution. *PLoS ONE* **2013**, *8*, e58547. [[CrossRef](#)] [[PubMed](#)]
95. Nakashima, K.; Ogawa, T.; Oda, N.; Hattori, M.; Sakaki, Y.; Kihara, H.; Ohno, M. Accelerated evolution of *Trimeresurus flavoviridis* venom gland phospholipase A₂ isozymes. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 5964–5968. [[CrossRef](#)] [[PubMed](#)]
96. Nakashima, K.; Nobuhisa, I.; Deshimaru, M.; Nakai, M.; Ogawa, T.; Shimohigashi, Y.; Fukumaki, Y.; Hattori, M.; Sakaki, Y.; Hattori, S.; et al. Accelerated evolution in the protein-coding regions is universal in crotalinae snake venom gland phospholipase A₂ isozyme genes. *Proc. Natl. Acad. Sci. USA* **1995**, *92*, 5605–5609. [[CrossRef](#)] [[PubMed](#)]
97. Ikeda, N.; Chijiwa, T.; Matsubara, K.; Oda-Ueda, N.; Hattori, S.; Matsuda, Y.; Ohno, M. Unique structural characteristics and evolution of a cluster of venom phospholipase A₂ isozyme genes of *Protobothrops flavoviridis* snake. *Gene* **2010**, *461*, 15–25. [[CrossRef](#)] [[PubMed](#)]
98. Chijiwa, T.; Ikeda, N.; Masuda, H.; Hara, H.; Oda-Ueda, N.; Hattori, S.; Ohno, M. Structural characteristics and evolution of a novel venom phospholipase A₂ gene from *Protobothrops flavoviridis*. *Biosci. Biotechnol. Biochem.* **2012**, *76*, 551–558. [[CrossRef](#)] [[PubMed](#)]
99. Chijiwa, T.; Nakasone, H.; Irie, S.; Ikeda, N.; Tomoda, K.; Oda-Ueda, N.; Hattori, S.; Ohno, M. Structural characteristics and evolution of the *Protobothrops elegans* pancreatic phospholipase A₂ gene in contrast with those of *Protobothrops* genus venom phospholipase A₂ genes. *Biosci. Biotechnol. Biochem.* **2013**, *77*, 97–102. [[CrossRef](#)] [[PubMed](#)]
100. Koszul, R.; Fischer, G. A prominent role for segmental duplications in modeling eukaryotic genomes. *C. R. Biol.* **2009**, *332*, 254–266. [[CrossRef](#)] [[PubMed](#)]
101. Ohta, T. Simple model for treating evolution of multigene families. *Nature* **1976**, *263*, 74–76. [[CrossRef](#)] [[PubMed](#)]
102. Smith, G.P. Evolution of repeated DNA sequences by unequal crossover. *Science* **1976**, *191*, 528–535. [[CrossRef](#)] [[PubMed](#)]
103. Juárez, P.; Wagstaff, S.C.; Sanz, L.; Harrison, R.A.; Calvete, J.J. Molecular cloning of *Echis ocellatus* disintegrins reveals non-venom-secreted proteins and a pathway for the evolution of ocellatusin. *J. Mol. Evol.* **2006**, *63*, 183–193. [[CrossRef](#)] [[PubMed](#)]

104. GeneWise. Available online: <http://www.ebi.ac.uk/Tools/Wise2/index.html> (accessed on 31 May 2016).
105. Basic Local Alignment Search Tool (BLAST). Available online: <http://blast.ncbi.nlm.nih.gov/Blast.cgi> (accessed on 5 June 2016).
106. ClustalW2-Multiple Sequence Alignment. Available online: <http://www.ebi.ac.uk/Tools/msa/clustalw2> (accessed on 5 June 2016).
107. RepeatMasker. Available on: <http://www.repeatmasker.org> (accessed on 30 December 2015).
108. Genetic Information Research Institute. Available on: <http://www.girinst.org> (accessed on 7 January 2016).
109. National Center for Biotechnology Information (NCBI) Database. Available on: <http://www.ncbi.nlm.nih.gov> (accessed on 10 January 2016).



© 2016 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).