

# GIN Transposons: Genetic Elements Linking Retrotransposons and Genes

Ignacio Marín<sup>\*1</sup>

<sup>1</sup>Instituto de Biomedicina de Valencia, Consejo Superior de Investigaciones Científicas, Valencia, Spain

**\*Corresponding author:** E-mail: imarin@ibv.csic.es.

**Associate editor:** Norihiro Okada

## Abstract

In a previous work, we characterized a gene, called *Gypsy Integrase 1 (GIN1)*, which encodes a protein very similar to the integrase domains present in *Gypsy/Ty3* retrotransposons. I describe here a paralog of *GIN1* and *GIN2* and show that both genes are present in multiple vertebrates and that a likely homolog is found in urochordates. Surprisingly, phylogenetic and structural analyses support the counterintuitive idea that the *GIN* genes did not directly derive from retrotransposons but from a novel type of animal-specific DNA transposons, the *GIN* elements. These elements, described for the first time in this study, are characterized by containing a gene that encodes a protein that is also very similar to *Gypsy/Ty3* integrases. It turns out that the sequences of the integrases encoded by *GIN1* and *GIN2* are more similar to those found in *GIN* elements than to those detected in retrotransposons. Moreover, several introns are in the same positions in the integrase-encoding genes of some *GIN* elements, *GIN1* and *GIN2*. The simplest explanation for these results is that *GIN* elements appeared early in animal evolution by co-option of the integrase of a retrotransposon, they later expanded in multiple animal lineages, and, eventually, gave rise to the *GIN* genes. In summary, *GIN* transposons may be the “missing link” that explain how *GIN* genes evolved from retrotransposons. *GIN1* and *GIN2* may have contributed to control the expansion of *GIN* elements and *Gypsy/Ty3* retrotransposons in chordates.

**Key words:** integrase, retrotransposon, DNA transposon, selfish DNA, cysteine protease, otubain.

## Introduction

Once considered strictly selfish sequences, it is today accepted that mobile elements are in fact subtly coevolving with the genome of the hosts in which they thrive (see, e.g., the recent reviews by Feschotte 2008; Venner et al. 2009). Particularly, it has been extensively documented that new, often essential, genes of the hosts derive from different classes of mobile sequences (reviewed by Volff 2006; Dooner and Weil 2007; Feschotte and Pritham 2007; Jurka et al. 2007). The emergence of many of these novel genes follows simple patterns. For example, the insertion of mobile sequences may contribute novel exons to a gene. In another typical scenario, recombination events put together coding sequences of a mobile element and a gene. However, other cases are far less evident. The *Gypsy Integrase 1 (GIN1)* gene is one of these more complex examples. Some years ago, we found that gene in several eutherian mammals, including humans. It encoded an integrase that was closely related to the integrase domains included in the pol polyproteins of *Gypsy/Ty3* retrotransposons (Lloréns and Marín 2001). However, how *GIN1* emerged was difficult to envisage. On one hand, active *Gypsy/Ty3* retrotransposons are absent in mammals. Moreover, although the mammalian genomes contain, in addition to *GIN1*, a substantial group of genes derived from *Gypsy/Ty3* retrotransposons, none of those genes encodes for an integrase (Volff et al. 2001; Lynch and Tristem 2003; Brandt et al. 2005; Youngson et al. 2005; Campillos et al. 2006; Marco and Marín 2009). Finally, the process that allowed an integrase do-

main, part of a polyprotein, to become an independent gene was not obvious. Recently, I found sequences very similar to *GIN1* in multiple animal species. The analyses of those sequences provide new clues about the evolutionary past of the *Gypsy Integrase* genes. I show here that *GIN1* is much older than previously thought and also describe a second *GIN* gene in vertebrates, *GIN2*, and a potential *GIN* gene in urochordates. However, the most significant finding derives from the characterization of a novel type of DNA transposons, the *GIN* elements. The evidence obtained indicates that *GIN1* and *GIN2* evolved from these transposons and not, as it was assumed so far, from *Gypsy/Ty3* retrotransposons.

## Materials and Methods

*GIN1*-related sequences were obtained from the nr, est, gss, wgs, and htgs databases available at the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>). Multiple searches were performed using TBlastN, BlastP, or TBlastX (<http://blast.ncbi.nlm.nih.gov/>) until no additional sequences were recovered. Thus, all the full-length sequences closely related to *GIN1* available at the end of October 2009 were detected. The number of new significant sequences was striking. After eliminating duplicates and very similar sequences ( $\geq 99\%$  identical), 68 new animal sequences were detected that were much more similar to human and mouse *GIN1* than the integrase domains of *Mdg1* retrotransposons, previously characterized as the closest relatives of *GIN1* by Lloréns and Marín

(2001). For example, some sequences of the cnidarian *Hydra magnipapillata* were 31% identical and 51% similar to human *GIN1* along 386 amino acids (*E* value, TBLASTN against nr database:  $10^{-49}$ ). When *GIN1* was compared with Mdg1 retrotransposons, the most similar sequences had just 30% identity and 50% similarity along 202 amino acids (corresponding *E* value:  $10^{-24}$ ).

To sort out these new sequences, phylogenetic analyses were performed following methods similar to those recently described in other recent papers of my group (e.g., Marco and Marín 2009). First, protein sequences were aligned using ClustalX 2.07 (Larkin et al. 2007). The alignments were manually corrected, when needed, with the GeneDoc sequence editor (Nicholas KB and Nicholas HB 1997). Dendrograms were then built using data extracted from that alignment, following three different procedures: neighbor joining (NJ), maximum parsimony (MP), and maximum likelihood (ML). The NJ trees were obtained using the routine in MEGA 4 (Tamura et al. 2007), whereas MP analyses were performed using PAUP\* 4.0 beta 10 version (Swofford 2002), and ML reconstructions were established using PhyML 3.0 (Guindon and Gascuel 2003). For NJ, the pairwise deletion option was used (as recommended by Dwivedi and Gadagkar 2009) and Kimura's correction implemented. Parameters for MP were as follows: 1) all sites included, gaps treated as unknown characters; 2) randomly generated trees used as seeds; 3) maximum number of trees saved equal to 200; and 4) heuristic search using the tree bisection–reconnection algorithm. Finally, for ML analyses, ProtTest (Abascal et al. 2005) was used to determine the best model of sequence evolution. The best ProtTest results were obtained with the LG + I + G + F model (i.e., Le and Gascuel matrix of amino acidic substitutions, presence of invariable sites, multiple rates of change and frequencies at equilibrium estimated from the alignment). Therefore, this model was used in the PhyML analyses. ML searches were started from the BioNJ tree and gaps were also treated as unknown characters. Reliability of the topologies was tested by bootstrap analyses. One thousand bootstrap replicates were performed for the NJ and MP analyses and 100 for the more computer-intensive ML analyses. MEGA 4 was used to edit and draw the final trees.

Gene and transposon structures were determined by combining the results of analyses performed with TBLASTN and TBLASTX, ORF Finder (<http://www.ncbi.nlm.nih.gov/gorf/>), InterProScan (<http://www.ebi.ac.uk/Tools/InterProScan/>; Zdobnov and Apweiler 2001), and GenomeScan (<http://genes.mit.edu/genomescan.html>; Yeh et al. 2001). The combination of these analyses allowed establishing the most likely beginning and end of the genes and mobile elements, the intron–exon structures of their coding regions, and the protein domains present in their products. Finally, the characterizations of the current locations of the *GIN* genes in different genomes were performed at the Ensembl Genome Browser Web page (<http://www.ensembl.org/index.html>; Hubbard et al. 2009). Blast analyses against Ensembl data were performed in order to determine the location of the *GIN1* and *GIN2* sequences in multiple genomes and, when required, additional

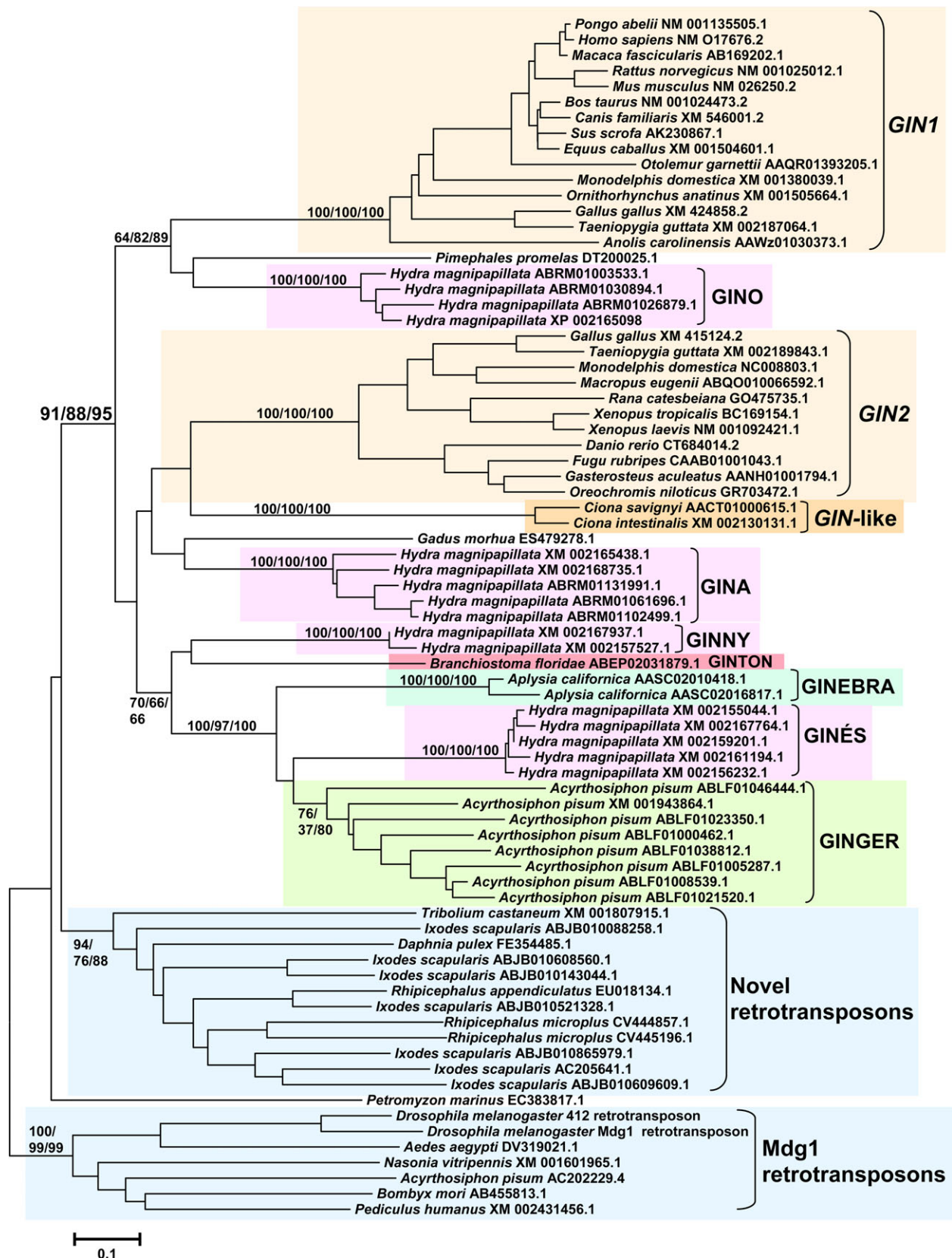
TBLASTN analyses against the NCBI databases to confirm orthologies between particular genes, adjacent to either *GIN1* or *GIN2*, in different species.

## Results

This work started when I observed, after an update of a database of Gypsy retrotransposon and retroviral integrases generated to characterize the *CGIN1* gene (Marco and Marín 2009), that there were many novel sequences that had a striking similarity to *GIN1*. Preliminary phylogenetic analyses (see [supplementary file 1](#), Supplementary Material online) indicated that many of those new sequences were indeed substantially more similar to *GIN1* than the sequences previously described as its closest relatives, derived from retrotransposons of the Mdg1 clade (Lloréns and Marín 2001). Therefore, all the animal sequences that were potentially interesting were selected and phylogenetic trees were built including them and also sequences from Mdg1 retrotransposons, which were to be used as outgroups. The results are shown in [figure 1](#), and the aligned sequences can be found in [supplementary file 2](#) (Supplementary Material online). The detailed analysis of these new sequences totally changes our views of how the *GIN1* gene originated.

A first result that can be deduced from [figure 1](#) is that the suggestion that *GIN1* is a mammalian-specific gene (Lloréns and Marín 2001) was incorrect. The phylogenetic data indicated the presence of a *GIN1* gene in the birds *Gallus gallus* and *Taeniopygia guttata* and in the lizard *Anolis carolinensis*. This result was confirmed by Blast searches at the Ensembl Genome Browser. In those three species, the genes that are located at both sides of the *GIN1*-like sequences are the same that are found in mammals around the *GIN1* gene: *HISPPD1* and *PAM*. These findings, together with the apparent lack of *GIN1* sequences in amphibians and fishes (but see below), suggest that *GIN1* emerged early in amniote evolution, perhaps 300 millions of years ago (Ponting 2008). In birds, *GIN1* resides in the Z chromosome. The available information of the *Anolis* genome does not allow to establish in which chromosome is located *GIN1* in that species.

A second notable result that emerged when analyzing these sequences is that there is a second related gene in vertebrates, which I have logically called *Gypsy Integrase 2* (*GIN2*). The first hint that a second gene existed derived again from the phylogenetic analyses. It became clear that a large set of unknown sequences from a variety of vertebrates (just one sequence per species) formed a monophyletic group ([fig. 1](#)). Detailed analyses of the structure and chromosomal positions of these sequences indeed indicated that they correspond to a novel gene, a *GIN1* paralog. First, no structural or sequence evidence for *GIN2* integrases being contained in a retrotransposon or any other type of repetitive element, as a DNA transposon, was found. Second, *GIN1* and *GIN2* have similar intron–exon structures, as will be detailed below. Finally, respect to their chromosomal positions, and as described above for *GIN1*, the same two genes (*OGFOD2* and *ABCB9*) are found adjacent to *GIN2* in the birds *G. gallus* and *T. guttata* and in



**Fig. 1.** Summary of the phylogenetic analyses. *GIN1*, *GIN2*, and possibly the *GIN*-like sequences found in *Ciona* correspond to genes, whereas *Gino*, *Gina*, *Ginny*, *Ginton*, *Ginebra*, *Ginés*, and probably *Ginger* are *GIN* DNA transposons (see main text). Results from NJ, MP, and ML analyses were congruent enough as to be shown together in a single tree (in the figure, the NJ tree obtained). Numbers refer to bootstrap support (NJ/MP/ML), in percentages. For simplicity, only the values obtained for internal branches that were supported by the three methods of phylogenetic reconstruction are shown. Notice the high level of support for the branch containing *GIN* genes and *GIN* elements. The accession number of each sequence is detailed after the name of the corresponding species.

**Table 1.** Features of Canonical GIN Elements.

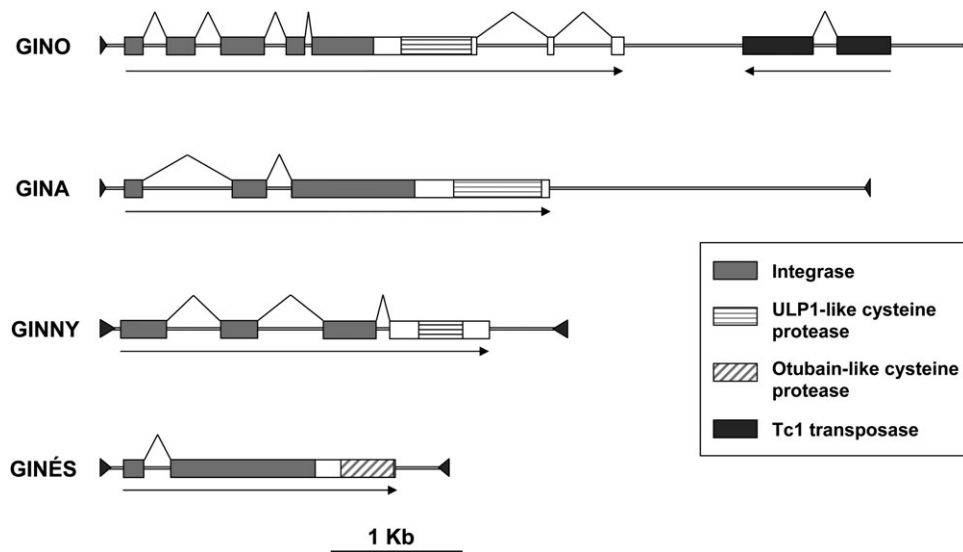
Element Name	Species	Accession Number for Canonical Element (start, end of the element)	Size (kb)	Size of Inverted Repeats (bp)	Coding Potential	Approximate Number of Sequences Related to the Canonical Element in the Current Databases	Site Duplications (bp)
GINO	<i>Hydra magnipapillata</i>	ABRM01009625.1 (16028, 9445)	6.6	45	ORF1: integrase/ULP1-like cysteine protease; ORF2: Tc1-like transposase	>100	4
GINA	<i>H. magnipapillata</i>	ABRM01027140.1 (3432, 9270)	5.8	40	Integrase/ULP1-like cysteine protease	45	4
GINNY	<i>H. magnipapillata</i>	ABRM01018334.1 (15616, 12073)	3.5	111	Integrase/ULP1-like cysteine protease	30	4
GINÉS	<i>H. magnipapillata</i>	ABRM01000799.1 (23816, 26452)	2.6	82	Integrase/otubain-like cysteine protease	>100	4
GINEBRA	<i>Aplysia californica</i>	AASC02016817.1 (36379, 31560)	4.8	119	Integrase	25	4
GINTON	<i>Branchiostoma floridae</i>	ABEP02034879.1	4.7	75	Integrase	4	4?
GINGER (putative element)	<i>Acyrtosiphon pisum</i>	Most complete elements: ABLF01044789.1 (2736, 5123), ABLF01050592.1 (2972, 605)	≥2.3	Not determined	Integrase	>100	Not determined

the fishes *Danio rerio* and *Fugu rubripes*, a strong evidence against *GIN2* being part of a mobile sequence. For the rest of species that contain *GIN2*, it was not possible to confirm this result due to lack of data. In any case, the combined evidence indicates that *GIN2* is an ancient gene, perhaps even more ancient than *GIN1* given its presence in fishes, and that *GIN1* and *GIN2* are paralogs. The fact that *GIN2* was not discovered before is due in part to the fact that, although it is present in some marsupials (*Monodelphis domestica* and *Macropus eugenii*) whose genomes only recently have been sequenced in detail, *GIN2* has been lost in eutherian mammals. Clearly, no *GIN2* sequences were available when we found *GIN1*, given that they would have been impossible to miss.

The third main result derived from the phylogenetic analyses was the finding of several ensembles of closely related sequences, each ensemble belonging to a single species. Four groups of sequences from the cnidarian *H. magnipapillata*, a group from the insect *Acyrtosiphon pisum*, and some sequences from the mollusk *Aplysia californica* were detected. Given the similarity of *GIN1* protein and retrotransposon integrases, I first thought that they corresponded to Gypsy/Ty3 retrotransposons. However, a close inspection of the data showed that these sequences did not actually derive from retrotransposons. In several cases, it was unambiguously determined that the integrase-encoding region is included in sequences with the typical structure of DNA transposons, a novel type of elements that I have called GIN. [Supplementary file 3](#) (Supplementary Material online) contains the sequences of canonical copies of the full-length elements described below (summarized in [table 1](#)).

The *Hydra* sequences were the first analyzed, given that it would be important to find potential progenitors of *GIN* genes in such animals, which belong to a basal lineage of the metazoan tree. By combining TBlastN, BlastX, TBlastX, and ORF Finder analyses, it was determined that all the *Hydra* sequences corresponded to copies of just four different DNA transposons, which I have called Gino, Gina, Ginny, and Ginés. The structural data are conclusive (summarized in [table 1](#), in which the features of canonical elements are detailed; [fig. 2](#)). First, multiple similar copies were found, and when the longest ones were compared, it was possible to establish that they ended in characteristic 40- to 111-bp-long terminal inverted repeats. These inverted repeats are element specific; their sizes and sequences are different for each of the four elements ([table 1](#)). Second, as is also typical in DNA transposons, direct duplications caused by the insertion of the elements, in this case of four nucleotides (often CCGG), were in most cases found at both sides of the terminal repeats. Finally, retrotransposon proteins or protein domains, such as reverse transcriptases, were never detected in close proximity to the *GIN*-like integrases that characterize these elements.

By combining the results of different types of Blast analyses and ORF Finder results with data obtained with InterProScan and GenomeScan (see Materials and Methods), the most likely structure of the coding regions of these elements was determined ([fig. 2](#)). The four elements are similar. First, in all cases, the open reading frame (ORF) encoding the *GIN1*-like integrase sequences starts very close to one of the extremes of the element (shown as 5' end in [fig. 2](#)). Second, also in all elements, introns were deduced to exist. However, some significant differences



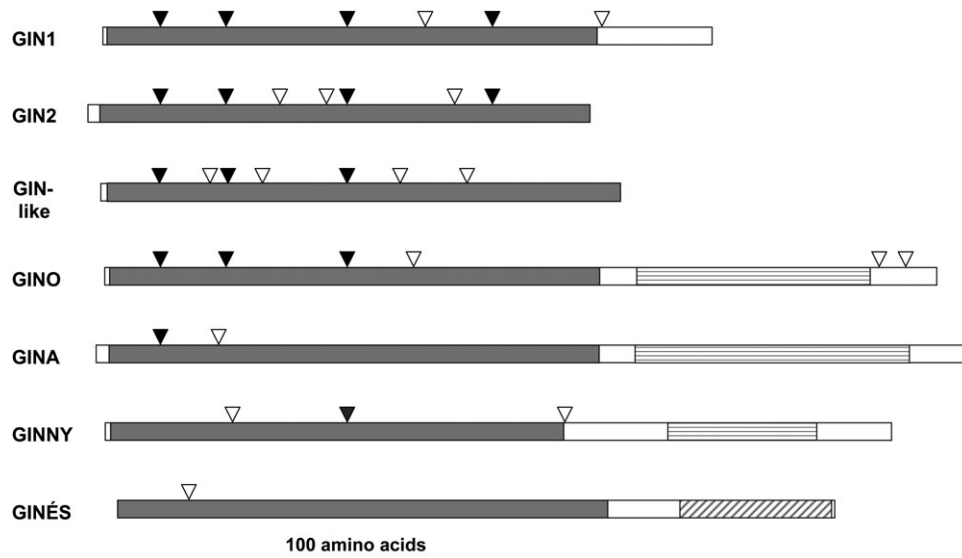
**Fig. 2.** Structures of the *Hydra* GIN elements. Triangles at both sides indicate the terminal inverted repeats. Arrows indicate the direction in which the proteins are encoded.

were also detected. Surprisingly, the longest element, Gino, contains a second ORF (ORF2) that may encode a transposase, very similar to those found in elements of the Tc1-mariner superfamily. This finding is discussed in detail below. On the other hand, in the four elements, a protein domain, related to cysteine proteases, was detected, C-terminal to the integrase sequences. However, although Gino, Gina, and Ginny contain a domain clearly related to ubiquitin-like cysteine proteases (ULPs; Hay 2007), it turns out that Ginés contains a totally different domain, this time obviously similar to otubain cysteine proteases, a type of deubiquitinating enzymes (Kim et al. 2003). Current classifications establish that ULPs and otubain proteases are only distantly related (see, e.g., the MEROPS classification at <http://merops.sanger.ac.uk/>; Rawlings et al. 2008), and indeed, the protease sequences deduced from Gino, Gina, and Ginny are similar and very different from the ones in Ginés. Critical amino acids typical of those types of cysteine proteases, such as the His–Asp–Cys catalytic triad of ULPs (Hay 2007) and the Cys–His catalytic couple of otubains (Nanao et al. 2004), were found intact in the protein domains deduced for the GIN elements.

The apparent presence of two ORFs in Gino elements was puzzling. Three potential explanations for this fact were examined. First, the elements could actually contain both ORFs. Second, it could be an artifact caused by the close proximity of Tc1-like and Gino elements in the canonical copies examined. Third, it was possible for a Tc1 element to be preferentially inserted within Gino sequences giving rise to an apparent ORF2. The second potential explanation was quite easily refuted: in no less than 20 cases, the sequences corresponding to the Gino integrases and the sequences corresponding to the Tc1 transposases were found to be adjacent. Moreover, it was possible to reconstruct 12 complete or almost complete Gino elements that contained both integrase and transposase sequences (supplementary file 4, Supplementary Material

online). Therefore, the presence of the two ORFs together was not a casual finding. The third potential explanation is more difficult to refute, but it is unlikely, given that Tc1 sequences as the ones detected within Gino elements were not found isolated and no inverted repeats around the Tc1 sequences were detected within Gino elements. This impossibility of characterizing the putative Tc1 element associated with Gino contrasts with how easy was to establish the complete structure of other Tc1 elements present in *H. magnipapillata*. For example, two elements that encoded transposases similar to those in Gino were found (Type A: accession numbers ABRM01018993.1 and XP\_002158263.1; identity with Gino Tc1 transposase sequences: 44%; similarity: 65%. Type B: accession numbers ACZU01091993.1 and XP\_002170130.1; identity and similarity with Gino Tc1 sequences: 55% and 73%, respectively). In both cases, their sizes (1,841 and 1,739 bp, respectively) and their inverted repeats (27 and 32 bp long, respectively) were characterized without difficulty. In summary, all these results indicate that Gino elements may contain as an integral part of their structure an ORF2 encoding a Tc1-like transposase.

The analyses of the *Ap. californica* sequences led to the characterization of another DNA transposon, Ginebra. Ginebra elements have 119-bp-long inverted repeats and encode a single protein, which contains just an integrase domain. They also generate 4-bp-long direct duplications that can be observed at both sides of the inverted repeats. This time, introns were not detected. Details are summarized in table 1. Also, the *Ac. pisum* sequences detected most likely belong also to DNA transposons (Ginger elements). Unfortunately, no full-length copies are currently available so it was impossible to characterize the ends of the elements. However, the presence of a large amount of copies and the lack of any other retrotransposon-related sequences argue against them being either host genes or retrotransposons. Finally, a single intact full-length



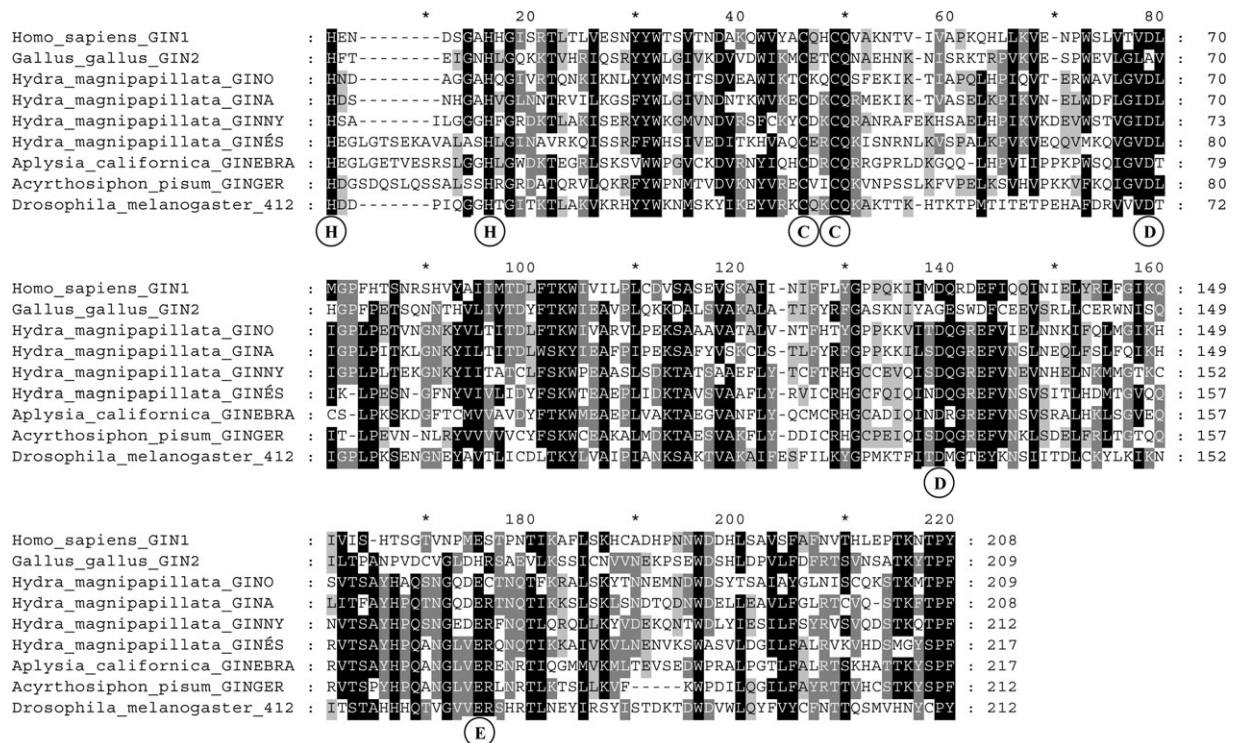
**Fig. 3.** Intron–exon structures of *GIN* genes (*GIN1*, *GIN2*, and the *GIN*-like gene in *Ciona intestinalis*) and *Hydra* GIN elements. Triangles indicate the positions of the introns along the coding region. Black triangles indicate that the position is the same found in *GIN1*. The similarity of *GIN1*, *GIN2*, and the GIN transposons, especially Gino, is clear.

sequence was found in *Branchiostoma floridae* that may also correspond to a GIN transposon (Ginton element). Inverted sequences 75 bp long and potential 4-bp direct repeats were found around this sequence, which encodes for an intronless integrase. Moreover, in the databases, three additional partial copies that encoded fragments (61–113 amino acids) of very similar integrases were detected. The available information for these putative elements is also summarized in table 1.

Most sequences in figure 1 corresponded to one of the classes already described: *GIN1* genes, *GIN2* genes, or GIN elements in different species. However, a few more were typical of Gypsy retrotransposons. In fact, a lineage of arthropod retroelements more similar to *GIN1* than Mdg1 retrotransposons was found and that is the reason why it has been included in figure 1. Finally, a few additional sequences are difficult to classify. Two intriguing *GIN*-like ORFs were found in the urochordates *Ciona savignyi* and *C. intestinalis*. The available evidence suggests that they may also correspond to *GIN* genes. They are single-copy sequences, and no obvious terminal or direct repeats, or other types of transposon- or retrotransposon-related sequences, were detected around them. Additional evidence against transposition is the fact that one of the genes adjacent to these *GIN*-like sequences is common in *C. intestinalis* and *C. savignyi*. Moreover, the location of several introns is the same in these *GIN*-like sequences, *GIN1* and *GIN2* (fig. 3). However, no genes obviously related to those found around *GIN1* or *GIN2* in vertebrates were found adjacent to the place where these *Ciona* sequences are located. Also, their relationship with either *GIN1* or *GIN2* is not significantly supported by bootstrap analyses (fig. 1). Therefore, whether they are bona fide *GIN* genes or not is still an open question. The other three orphan sequences (from the teleost fishes *Pimephales promelas* and *Gadus morhua* and the lamprey *Petromyzon marinus*; fig. 1)

were derived from cDNAs and the corresponding genomic sequences are not available, so no further characterization was possible. The *Pimephales* sequence is particularly interesting, given that it is quite similar to *GIN1* genes. Therefore, it is not impossible that *GIN1* sequences, corresponding to either active genes or pseudogenes, are present in some fish species. If this is confirmed, it would mean that *GIN1* is even older than it was deduced from the data presented above.

Figure 1 shows that the phylogenetic analyses put together *GIN1* and *GIN2* genes and GIN transposons or likely transposons in *Hydra*, *Aplysia*, and *Acyrtosiphon* in a highly supported monophyletic group, suggesting that *GIN1*, *GIN2*, and the GIN DNA transposons have a common origin. Further data supporting the close evolutionary link between these genes and transposons are presented in figure 3, which describes the position of the introns in the different GIN sequences. As it may be expected for paralogs, *GIN1* and *GIN2* have similar intron–exon structures. Identical positions for four introns were detected in both genes (black triangles in fig. 3). The same applies for the *GIN*-like sequences in *Ciona* (fig. 3). In addition, and most significantly, similar positions were also detected in several cases in *Hydra* GIN transposons. Particularly, Gino shares the position of three introns with both *GIN1* and *GIN2* (see also fig. 3). These results further suggest the existence of a common ancestor of both the current GIN elements and the GIN genes. That ancestor must have included an integrase-coding region with introns. Figure 4 shows in detail the great similarity among GIN integrases of genes and transposons and also that the integrases of *Hydra* GIN elements may be active, given that they contain typical C<sub>2</sub>H<sub>2</sub> and DDE signatures critical for integrase function (Haren et al. 1999). However, it is unlikely that the *GIN2* protein may act as an integrase, given that it lacks the three critical acidic residues of the DDE signature.



**Fig. 4.** Core integrase domains of representative sequences for *GIN1*, *GIN2*, and *GIN* transposons. The integrase of the *Drosophila melanogaster* 412 element (an Mdg1 retrotransposon) is also shown. The critical  $C_2$ ,  $H_2$  and DDE residues are indicated. Notice that *GIN2* integrase lacks the DDE motif.

## Discussion

The main result of this work is the characterization of the close similarity between *GIN* genes and a novel type of DNA transposons, the *GIN* elements. We may ask now what is the most likely evolutionary history that explains their similarities. The facts to consider are as follows: 1) no known retrotransposons are similar enough to *GIN1* or *GIN2* to explain their origin; 2) on the contrary, *GIN* transposons are very similar in structure (related intron positions; fig. 3) and sequence (figs. 1 and 4) to *GIN* genes; and 3) their current phylogenetic range suggests that *GIN* DNA transposons, present in both cnidarians and protostomes, have an older origin than either *GIN1* and *GIN2*, which are restricted to chordates, perhaps even to vertebrates. It is therefore very logical to hypothesize a single common ancestor that provided the integrase sequences found today both in *GIN* elements and in *GIN* genes and that *GIN* genes, recently evolved, derive from the more ancient *GIN* DNA transposons. The possibility of *GIN* elements to have a broad phylogenetic range due in part to horizontal transmission must be taken into account in this context. However, in my opinion, it does not change the fact that, with the current information, the other possible evolutionary histories are difficult to envisage. For example, an alternative hypothesis would be that *GIN* elements and *GIN* genes emerged independently from different retrotransposons. However, in that case, neither their sequence similarity nor their related intron–exon structures are easy to explain. A third potential explanation is that *GIN* elements derive

from *GIN* genes. However, we should then postulate either 1) that the genes derived directly from a retrotransposon, were present in ancient animals and later were independently lost several times while novel DNA transposons emerged from these genes and persisted in those same lineages or 2) that *GIN* elements originated recently in vertebrates/chordates and later they spread by independent horizontal transmissions to several, very different, lineages of animals. Although formally possible, the first option seems very unlikely. The second possibility, although more logical, is also unlikely. This is showed by the fact that the four different elements described in *Hydra* cannot be aligned along their lengths and even within a particular type of element, the similarity among copies is quite low (e.g., Gino copies in supplementary file 1, Supplementary Material online, are just 57–86% identical). This suggests an ancient origin and argues against recent horizontal transfers. In summary, the simplest hypothesis is that *GIN* DNA transposons are genetic elements that emerged long ago by an evolutionary accident leading a retrotransposon integrase to be used as a transposase, a singular event that probably occurred in early animal evolution. Later, they contributed to the emergence of *GIN* genes by multiplying in animal genomes such integrase sequences until one of them by chance co-opted to work as a common gene in the genome of a chordate. According to this view, *GIN* transposons would be an evolutionary “missing link” between retrotransposons and *GIN* genes.

In summary, I postulate that *GIN* genes derive from the “domestication” of *GIN* transposons. Whether this

domestication occurred just once, generating an ancestral *GIN* gene that later become duplicated, or, alternatively, occurred twice, independently originating the two genes that we found today in vertebrates, is not totally clear. The first option seems more likely, especially due to the similarity of the intron–exon structures of both *GIN* genes. However, the data are still too fragmentary to be certain. For example, the fact that the integrase of some *GIN* transposons is more similar to those in *GIN1*, whereas others are more similar to *GIN2* (fig. 1), argues in favor of two independent domestications. True, that evidence is quite weak.

It is most interesting that previous studies suggest that the evolutionary path described here may have occurred in parallel in a totally unrelated case. Wells (1999) described the Tdd-4 DNA transposon of *Dictyostelium discoideum*, a 3.8-kb element encoding an integrase with clear relationships with retrotransposon integrase domains. In later works, Gao and Voytas (2005) and Feschotte and Pritham (2005) established that a group of proteins called *c*-integrases were very similar to the integrase of Tdd-4 but just distantly related to those in Gypsy/Ty3 retrotransposons and retroviruses. *c*-Integrases were found in multiple species, and it was concluded after structural analyses that many of them were actually included in giant DNA transposons called Mavericks or Polintons (Feschotte and Pritham 2005; Kapitonov and Jurka 2006; Pritham et al. 2007). The parallelism between the origin of *GIN* transposons and the origin of the Tdd-4 and the Maverick/Polinton elements is obvious: both types of elements emerged from the co-option of similar integrases. The question is whether this parallelism can be even more exceptional, given that there are DNA sequences in mammalian genomes that encode for *c*-integrases but so far have not been found to be included in any DNA transposon (Feschotte and Pritham 2005). These authors suggested that they could correspond to host genes derived from Mavericks/Polintons. If this is indeed the case, it would be a second independent case of a common evolutionary route for domestication of an integrase: from retrotransposons/retroviruses to host genes through an intermediate phase as part of DNA transposons. It is also interesting in this context the suggestion by Kapitonov and Jurka (2006) that the integrases of Mavericks/Polintons may have been co-opted from a Tdd-4-like DNA transposon.

*GIN* transposons are quite peculiar. In some ways, they are examples of a novel type of element, given that they have structural features and encode types of proteins that have never been described so far in other DNA transposons. However, they have clear relationships to other, already known, types. First, they are structurally related to Tdd-4, which also contains an integrase-coding region with introns (Wells 1999). Second, the presence of ULP cysteine protease domains in the proteins of several *Hydra* *GIN* elements also have an interesting precedent: DNA transposons encoding ULPs (as independent proteins) have been detected in plants (Hoen et al. 2006; van Leeuwen et al. 2007). The potential functions of the ULP activity (or the otubain protease activity in the case of the *Ginés* ele-

ment) are obscure. How and why *Gino* elements incorporated a second ORF, encoding a Tc1 transposase, and how the integrase/cysteine protease and the Tc1 transposase activities may collaborate for *Gino* replication are also puzzling questions that deserve further study. In the next years, we can expect to increase our collection of these curious elements in other animal species, so perhaps several of these questions will be soon solved. Finally, the main mystery, the function of *GIN* genes in vertebrates, persists. In our original report, we suggested that *GIN* integrases could be part of a defense mechanism against retrotransposons and retroviruses, perhaps contributing to the elimination of Gypsy/Ty3 elements in mammals (Lloréns and Marín 2001). Now, it is possible to postulate that they may have contributed in the past to the control and elimination not only of retroelements but also of *GIN* DNA transposons, which are apparently absent in all species with *GIN* genes. Perhaps they are still involved in some specific type of repetitive element control in modern genomes. However, our knowledge of domesticated genes is so incomplete that indeed it would not be a surprise to find them performing totally new host-specific functions.

## Supplementary Material

Supplementary files 1–4 are available at *Molecular Biology and Evolution* online (<http://www.mbe.oxfordjournals.org/>).

## Acknowledgments

This project was supported by grant BIO2008-05067 (Programa Nacional de Biotecnología; Ministerio de Ciencia e Innovación, Spain).

## References

- Abascal F, Zardoya R, Posada D. 2005. ProtTest: selection of best-fit models of protein evolution. *Bioinformatics* 21:2104–2105.
- Brandt J, Schrauth S, Veith AM, Froschauer A, Haneke T, Schultheis C, Gessler M, Leimeister C, Volff JN. 2005. Transposable elements as a source of genetic innovation: expression and evolution of a family of retrotransposon-derived neogenes in mammals. *Gene* 345:101–111.
- Campillos M, Doerks T, Shah PK, Bork P. 2006. Computational characterization of multiple Gag-like human proteins. *Trends Genet.* 22:585–589.
- Dooner HK, Weil CF. 2007. Give-and-take: interactions between DNA transposons and their host plant genomes. *Curr Opin Genet Dev.* 17:486–492.
- Dwivedi B, Gadagkar SR. 2009. Phylogenetic inference under varying proportions of indel-induced alignment gaps. *BMC Evol Biol.* 9:211.
- Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet.* 9:397–405.
- Feschotte C, Pritham EJ. 2005. Non-mammalian *c*-integrases are encoded by giant transposable elements. *Trends Genet.* 21:551–552.
- Feschotte C, Pritham EJ. 2007. DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet.* 41:331–368.
- Gao X, Voytas DF. 2005. A eukaryotic gene family related to retroelement integrases. *Trends Genet.* 21:133–137.
- Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol.* 52:696–704.

- Haren L, Ton-Hoang B, Chandler M. 1999. Integrating DNA: transposases and retroviral integrases. *Annu Rev Microbiol.* 53: 245–281.
- Hay RT. 2007. SUMO-specific proteases: a twist in the tail. *Trends Cell Biol.* 17:370–376.
- Hoen DR, Park KC, Elrouby N, Yu Z, Mohabir N, Cowan RK, Bureau TE. 2006. Transposon-mediated expansion and diversification of a family of ULP-like genes. *Mol Biol Evol.* 23:1254–1268.
- Hubbard TJP, Aken BL, Ayling S, et al. (58 co-authors). 2009. Ensembl 2009. *Nucl Acids Res.* 37:D690–D697.
- Jurka J, Kapitonov VV, Kohany O, Jurka MV. 2007. Repetitive sequences in complex genomes: structure and evolution. *Annu Rev Genomics Hum Genet.* 8:241–259.
- Kapitonov VV, Jurka J. 2006. Self-synthesizing DNA transposons in eukaryotes. *Proc Natl Acad Sci U S A.* 103:4540–4545.
- Kim JH, Park KC, Chung SS, Band O, Chung CH. 2003. Deubiquitinating enzymes as cellular regulators. *J Biochem.* 134:9–18.
- Larkin MA, Blackshields G, Brown NP, et al. (13 co-authors). 2007. Clustal W and Clustal X version 2.0. *Bioinformatics* 23: 2947–2948.
- Lloréns C, Marín I. 2001. A mammalian gene evolved from the integrase domain of an LTR retrotransposon. *Mol Biol Evol.* 18:1597–1600.
- Lynch C, Tristem M. 2003. A co-opted gypsy-type LTR-retrotransposon is conserved in the genomes of humans, sheep, mice and rats. *Curr Biol.* 13:1518–1523.
- Marco A, Marín I. 2009. *CGIN1*: a retroviral contribution to mammalian genomes. *Mol Biol Evol.* 26:2167–2170.
- Nanao MH, Tcherniuk SO, Chroboczek J, Dideberg O, Dessen A, Balakirev MY. 2004. Crystal structure of human otubain 2. *EMBO Rep.* 5:783–788.
- Nicholas KB, Nicholas HB Jr. 1997. GeneDoc: a tool for editing and annotating multiple sequence alignments. [cited 2010 Jun 25] Available from: <http://www.nrbcs.org/gfx/genedoc/index.html>.
- Ponting CP. 2008. The functional repertoires of metazoan genomes. *Nat Rev Genet.* 9:689–698.
- Pritham EJ, Putliwala T, Feschotte C. 2007. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. *Gene* 390:3–17.
- Rawlings ND, Morton FR, Kok CY, Kong J, Barrett AJ. 2008. MEROPS: the peptidase database. *Nucl Acids Res.* 36:D320–D325.
- Swofford DL. 2002. PAUP\*: Phylogenetic Analysis Using Parsimony (\*and other methods). Version 4. Sunderland (MA): Sinauer Associates.
- Tamura K, Dudley J, Nei M, Kumar S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol.* 24:1596–1599.
- Van Leeuwen H, Monfort A, Puigdomenech P. 2007. Mutator-like elements identified in melon, *Arabidopsis* and rice contain ULP1 protease domains. *Mol Genet Genom.* 277:357–364.
- Venner S, Feschotte C, Biémont C. 2009. Dynamics of transposable elements: towards a community ecology of the genome. *Trends Genet.* 25:317–323.
- Volff JN. 2006. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. *Bioessays* 28:913–922.
- Volff JN, Körting C, Schartl M. 2001. Ty3/Gypsy retrotransposon fossils in mammalian genomes: did they evolve into new cellular functions? *Mol Biol Evol.* 18:266–270.
- Wells DJ. 1999. Tdd-4, a DNA transposon of *Dictyostelium* that encodes proteins similar to LTR retroelement integrases. *Nucl Acids Res.* 27:2408–2415.
- Yeh RF, Lim LP, Burge CB. 2001. Computational inference of homologous gene structures in the human genome. *Genome Res.* 11:803–816.
- Youngson NA, Kocialkowski S, Peel N, Ferguson-Smith AC. 2005. A small family of Sushi-class retrotransposon-derived genes in mammals and their relation to genomic imprinting. *J Mol Evol.* 61:481–490.
- Zdobnov EM, Apweiler R. 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17:847–848.