Letter to the Editor

A Mammalian Gene Evolved from the Integrase Domain of an LTR Retrotransposon

Carlos Lloréns and Ignacio Marín

Departamento de Genética and Instituto Cavanilles de Biodiversidad y Biología Evolutiva, Universidad de Valencia, Spain

Ty3/Gypsy long-terminal-repeat (LTR) retrotransposons are among the best-known transposable elements. They inhabit the genomes of many eukaryotic organisms, such as slime molds, plants, fungi, and animals, including vertebrates (Xiong and Eickbush 1990; Malik and Eickbush 1999; Miller et al. 1999; Marín and Lloréns 2000). However, in spite of extensive genomic information, these elements had never been found in mammals. In the process of building a database of integrase domain (IN) sequences, we found an intriguing human sequence very similar to the IN of *Ty3/Gypsy* elements. It was particularly similar to the IN of the *Drosophila melanogaster 412* element (E value = 10^{-27}). The sequence of the human gene, which we called *Gypsy integrase-1*, or *Gin-1*, was reconstructed by combining information from genomic and cDNA sequences present in the National Center for Biotechnology Information databases (online at http://www.ncbi.nlm.nih.gov/; sequences in TIGR and Sanger Center databases did not provide additional information). Partial mouse, rat, and cow orthologous cDNAs were also detected. Moreover, an apparently full-length mouse cDNA sequence (accession number AK015243) was also found. However, the corresponding genomic sequences are not yet available for any of these other mammalian species.

Figure 1 summarizes the structure and describes the protein encoded by the human *Gin-1* gene. It has the characteristic H_2C_2, DDE, and GPY/F motifs found in many retroviral and retrotransposon integrases (Khan et al. 1991; Malik and Eickbush 1999). Homology to IN of *Ty3/Gypsy* elements spans the whole protein sequence (fig. 2), strongly suggesting that GIN-1 is also, and exclusively, an integrase. The similarity between the human and mouse genes is the expected similarity for orthologous genes of these two species. Amino acidic identity is 446/552 = 85%, while a comparative analysis of 1,138 mouse/human orthologs estimated an average identity in their coding regions of 86.4% (Makalowski and Boguski 1998).

Phylogenetic analyses using IN sequences of the known clades of *Ty3/Gypsy* elements (Malik and Eickbush 1999), as well as of representative *Ty1/Copia* elements and retroviruses, confirmed that the putative integrase encoded by *Gin-1* is very similar to Mdg1 clade elements IN (this clade so far includes *D. melanogaster 412* and *Mdg1*; Malik and Eickbush 1999). Their sequences form a strongly supported monophyletic group (fig. 3). It is often impossible to trace the phylogenetic relationships among *Ty3/Gypsy* elements using sequence information (Malik and Eickbush 1999; Marín and Lloréns 2000). Thus, the fact that *Gin-1* can be unambiguously related to *412* elements is striking. These results suggest that *412*-like elements were inhabiting animal genomes at the time of the protostome/deuterostome split and that *Gin-1* evolved from one of those elements in the lineage that gave rise to mammals. However, the occurrence of an ancient horizontal transmission of a *412*-like element from protostomes to deuterostomes, or vice versa, cannot be dismissed.

All available data suggest that *Gin-1* is a single-copy gene, located in our species in chromosome 5q14–5q21. Apart from *Gin-1*, we did not detect any other *412*-related sequences in mammalian genomes. Although still formally possible, it is highly unlikely that *Gin-1* is part of a transposable element. Several data, such as the presence of introns, the presence of a start codon, and the length spanned by the coding region in the genomic DNA (>20 kb; see fig. 1), strongly argue against it being the IN domain of a retrotransposon. *Tdd-4*, a *Dictyostelium discoideum* DNA transposon with 145/146-bp-long inverted repeats (IRs), contains a transposase with similarity to retrotransposon integrases that has several small introns (Wells 1999). To exclude the possibility that *Gin-1* is the transposase of a DNA transposon, we performed searches for IRs. The longest IRs found when 5 kb at both sides of *Gin-1* were analyzed

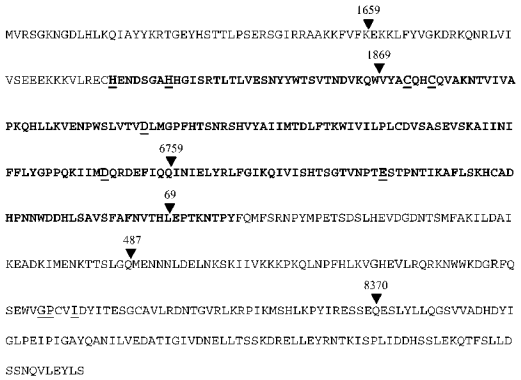
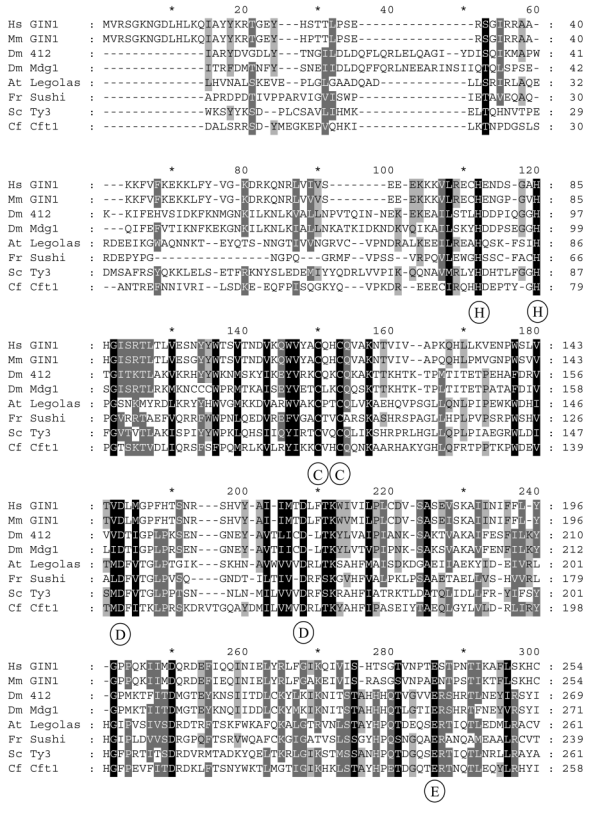


FIG. 1.—Summary of the structure and coding sequence of the human *Gin-1* gene. Sequences of human cDNAs with accession numbers XM_003947.2 (a putative full-length cDNA), BE502574, AW173201.1, AW950418.1, AI631948.1, and AA766836.1 were used to deduce and confirm these data. The full-length protein is 522 amino acids long. The *Gin-1* coding region spans nucleotides 36153–15345 in the genomic clone NT_002663.4. Arrowheads and the numbers above them, respectively, indicate the positions and lengths of introns. Several Alu repeats were detected within the two largest introns. Bold letters indicate the region homologous to the most conserved part of the IN domain, detailed in figure 2 and used to obtain the tree shown in figure 3. Amino acids characteristic of the H_2C_2 and DDE motifs (Khan et al. 1991) and of the most conserved region of the GPY/F module (Malik and Eickbush 1999) are underlined.

Key words: Gypsy, gene emergence, genome protection.

Address for correspondence and reprints: Ignacio Marín, Departamento de Genética, Universidad de Valencia, Calle Doctor Moliner, 50, Burjassot 46100, Valencia, Spain. E-mail: ignacio.marin@uv.es.

Mol. Biol. Evol. 18(8):1597–1600. 2001



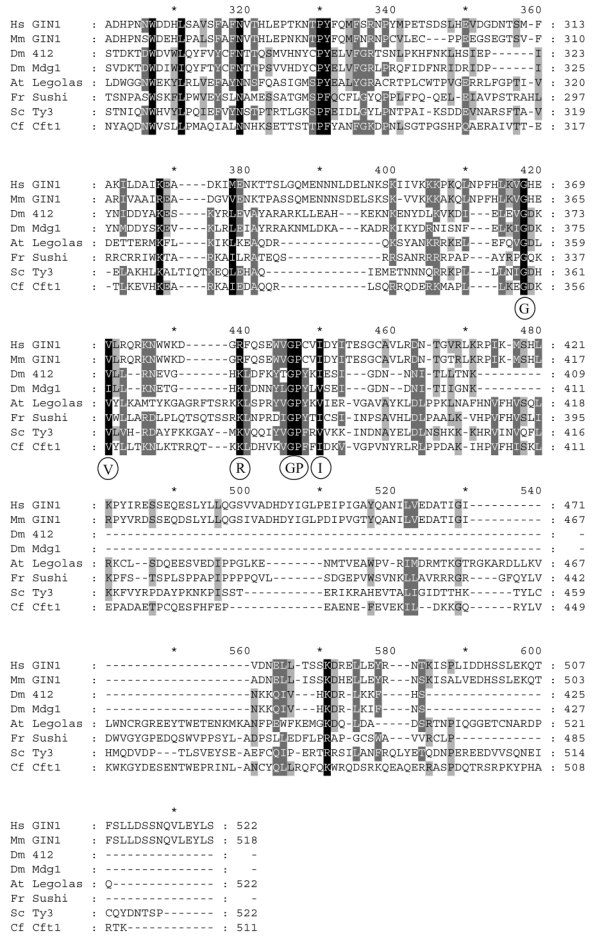


FIG. 2.—Alignments of the integrase sequences in human (Hs) and mouse (Mm) *Gin-1* with those of several *Ty3/Gypsy* retrotransposons, highlighting their similarities. The H_2C_2, DDE and GPY/F motifs are also indicated. The mouse sequence derives from a cDNA with accession number AK015243. The *Ty3/Gypsy* group elements were selected to include different types of IN domain C-terminal ends (see Malik and Eickbush 1999). Thus, apart from the Mdg1 clade elements from *Drosophila melanogaster* (Dm) that are the most similar to GIN-1 (*412* and *Mdg1*), this figure includes chromoviruses from plants (*Legolas* from *Arabidopsis thaliana* [At]), from animals (*Sushi* from the fish *Fugu rubripes*, Fr), and from fungi (*Cft1*, from *Cladiosporium fulvum* [Cf]), as well as the *Saccharomyces cerevisiae Ty3* element, which has a peculiar integrase C terminus (Malik and Eickbush 1999).

were just 9 bp long. If those IRs were the ends of a DNA transposon in which *Gin-1* acted as transposase, such an element would be around 29-kb long. To our knowledge, DNA transposons combining this very large size with very short IRs have never been described. Moreover, the 9-bp putative IR sequence is TTGGCTTGT, while the inverted repeats of DNA transposons generally start with the dinucleotides CA or TA (see examples in Berg and Howe 1989). These elements also generate direct duplications of the genomic DNA upon insertion, which can be detected at both sides of the elements. However, no such duplication was observed at both sides of the 9-bp repeat flanking human *Gin-1*. In summary, there is no evidence that *Gin-1* is part of a *Tdd-4*-like DNA transposon. Actually, our results suggest that *Tdd-4* may have arisen from a *D. discoideum Gin-1*-like gene that acquired the ability to transpose.

Gin-1 is widely expressed. The developmental stages or tissues from which cDNA libraries were obtained that included at least one human *Gin-1* clone were as follows: 12-week embryo (accession number AA328555), kidney (AI631948), brain (H23481), skeletal muscle (BF789980), and placenta (BE929727). *Gin-1* is also expressed in different human tumors, originating from the parathyroid gland (W39050), the colon (AA574153), the stomach (AI933750), the bladder (BE566969), the uterus (AW572091), and the prostate (AA804922). Mouse *Gin-1* cDNAs were found in libraries from embryos (AA162090), the testis (AK015243), the retina (BF463804), the lymph node (AA190067), the thymus (BF658153), the mammary gland (AA542272), and the spleen (AI639767). Finally, a rat *Gin-1* sequence came from an ovary cDNA library (AW144198). The only cow sequence available so far (BF774149) comes from a library of mixed origin.

Some other mammalian genes seem to have derived from transposable elements or retroviruses (reviewed in Smit 1999; International Human Genome Sequencing Consortium 2001). This is, however, the first description of a host gene derived from an LTR retrotransposon protein domain. In a recent work, Volff, Körting, and Schartl (2001) proposed the existence of another human gene (KIAA1051) derived from a *Ty3/Gypsy* element. Human KIAA1051 has similarity to the *gag* protein, as well as to the protease domain and a fragment of the reverse transcriptase domain of the *pol* protein of certain *Ty3/Gypsy* elements that we have called ''chromovirus-

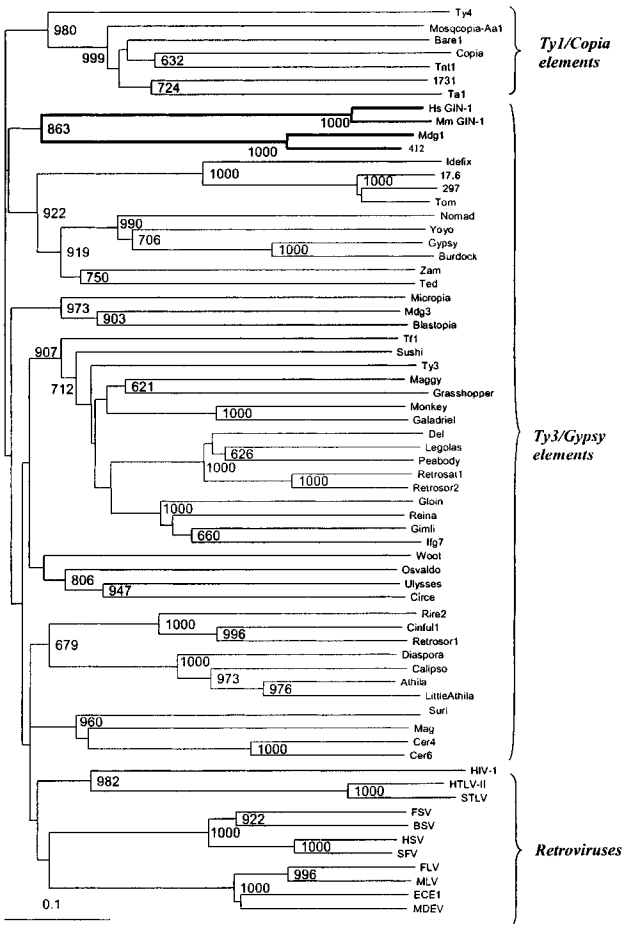


FIG. 3.—Phylogenetic analysis showing the close relationship of the GIN-1 protein to the IN sequences of Mdg1 clade elements. This tree is very similar to those obtained using reverse transcriptase sequences (Marín and Lloréns 2000). However, the inner branches of the tree essentially form a polytomy, and thus some differences in the internal topology arise. A neighbor-joining tree and bootstrap values were obtained using ClustalX (Thompson et al. 1997) as previously described (Marín et al. 1998; Marín and Lloréns 2000). One thousand bootstrap replicates were performed. Branches supported by 600 or more replicates are detailed. *Ty1/copia* elements were used to root the tree (following Xiong and Eickbush 1990). Details for most elements can be found in Marín and Lloréns (2000). The rest of the information is available on request.

es'' (see Marín and Lloréns 2000). It is well known that chromoviruses exist in other vertebrate genomes (Malik and Eickbush 1999; Marín and Lloréns 2000), and therefore the finding of chromovirus-related sequences in mammals is not totally unexpected. In fact, KIAA1051 lacks introns, and its structure resembles that of a truncated retrotransposon, including overlap of the putative *gag* and *pol* genes and an apparent lack of a starting codon for the *pol* open reading frame (ORF). One of the main results suggesting that KIAA1051 may be an active mammalian gene is the finding of very similar, putatively orthologous ORFs in several other mammalian species, of which the best characterized are mouse sequences (see Volff, Körting, and Schartl 2001). However, the similarity between the human and mouse sequences is lower than that found for *Gin-1* (identity in the *gag* region is about 70%, and it is about 75% for the *pol* partial sequence). Moreover, Volff, Körting, and Schartl (2001) detailed that the structures deduced for the putative orthologs in human and mouse are not totally congruent. Also to be considered is that although Volff, Körting, and Schartl (2001) did not find sequences related to KIAA1051 in the human genome, we detected in a recently released human clone (accession number AL117190.5) a KIAA1051-related sequence. Interestingly, in the KIAA1051-like sequence found in the AL117190.5 clone, a 1,358-amino-acid-long ORF is found that has highly significant similarity to the *pol* gene of *Sushi*, a chromovirus of the fish *Fugu rubripes* (E value = 7×10^{-59}). This ORF is transcribed (some related cDNAs are found in the databases), but the functional meaning of the product it encodes is unclear. For example, it lacks some highly conserved regions found in *Ty3/Gypsy* elements, such as the characteristic YXDD signature in the reverse transcriptase domain. Thus, although the evidence obtained by Volff, Körting, and Schartl (2001) is suggestive, it is still an open question whether KIAA1051 and the related sequence in AL117190.5 are bona fide human genes or simply peculiar types of defective, ''pseudogenized'' chromoviruses with nonfunctional ORFs that have retained a background level of transcription.

We can only speculate about the functions of *Gin-1*. Its conservation in different mammalian orders and its transcription in many different tissues argues for its having a significant, perhaps even essential, function. An intriguing possibility is that it is involved in repressing retrotransposon and/or retrovirus activity. It is known that an endogenous retrovirus-derived gene, *Fv1*, is able to confer resistance to murine leukemia viruses in the mouse (reviewed in Stoye 1998). *Gin-1* may be part of an analogous defense system, perhaps having contributed to the virtual absence of *Ty3/Gypsy* elements in mammalian genomes.

LITERATURE CITED

BERG, D. E, and M. M. HOWE, eds. 1989. Mobile DNA. American Society for Microbiology. Washington, D.C.

INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM. 2001. Initial sequencing and analysis of the human genome. Nature **409**:860–921.

KHAN, E., J. P. MACK, R. A. KATZ, J. KULKOSKY, and A. M. SKALKA. 1991. Retroviral integrase domains: DNA binding and the recognition of LTR sequences. Nucleic Acids Res. **19**:851–860.

MAKALOWSKI, W., and M. S. BOGUSKI. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. Proc. Natl. Acad. Sci. USA **95**:9407–9412.

MALIK, H. S., and T. H. EICKBUSH. 1999. Modular evolution of the integrase domain in the *Ty3/Gypsy* class of LTR retrotransposons. J. Virol. **73**:5186–5190.

MARÍN, I., and C. LLORÉNS. 2000. *Ty3/Gypsy* retrotransposons: description of new *Arabidopsis thaliana* elements and evolutionary perspectives derived from comparative genomic data. Mol. Biol. Evol. **17**:1040–1049.

MARÍN, I., P. PLATA-RENGIFO, M. LABRADOR, and A. FONTDEVILA. 1998. Evolutionary relationships among the members of an ancient class of non-LTR retrotransposons found

in the nematode *Caenorhabditis elegans*. Mol. Biol. Evol. **15**:1390–1402.

MILLER, K., C. LYNCH, J. MARTIN, E. HERNIOU, and M. TRISTEM. 1999. Identification of multiple *Gypsy* LTR-retrotransposon lineages in vertebrate genomes. J. Mol. Evol. **49**: 358–366.

SMIT, A. F. A. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr. Opin. Genet. Dev. **9**:657–663.

STOYE, J. P. 1998. Fv1, the mouse retrovirus resistance gene. Rev. Sci. Tech. **17**:269–77.

THOMPSON, J. D., T. J. GIBSON, F. PLEWNIAK, F. JEANMOUGIN, and D. G. HIGGINS. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25**: 4876–4882.

VOLFF, J. N., C. KÖRTING, and M. SCHARTL. 2001. *Ty3/Gypsy* retrotransposon fossils in mammalian genomes: did they evolve into new cellular functions? Mol. Biol. Evol. **18**: 266–270.

WELLS, D. J. 1999. *Tdd-4*, a DNA transposon of *Dictyostelium* that encodes proteins similar to LTR retroelement integrases. Nucleic Acids Res. **27**:2408–2415.

XIONG, Y., and T. H. EICKBUSH. 1990. Origin and evolution of retroelements based upon their reverse transcriptase sequences. EMBO J. **9**:3353–3362.

THOMAS EICKBUSH, reviewing editor

Accepted June 1, 2001