LETTERS

CGIN1: A Retroviral Contribution to Mammalian Genomes

Antonio Marco* and Ignacio Marín†

*Center for Evolutionary Functional Genomics, The Biodesign Institute, Arizona State University; and †Instituto de Biomedicina de Valencia, Consejo Superior de Investigaciones Científicas (IBV-CSIC), Spain

This study describes the origin and structural features of a mammalian gene, *CGIN1* (*Cousin of GIN1*). CGIN1 proteins contain an NYN domain, retroviral RNase H and integrase domains, and a domain of unknown function (CGIN1 domain) that is also present in two other genes (*N4BP1* and *KIAA323*). We suggest that *CGIN1* derives from the fusion of a *KIAA323*-like gene with retroviral sequences, which occurred prior to the marsupial–eutherian split. Sequence and structural analyses indicate that the CGIN1 integrase domain is inactive but still retains the 3D folding observed in retroviral integrases. We hypothesize that *CGIN1* may contribute to retroviral resistance in mammals by regulating the ubiquitination of viral proteins.

Mammalian genomes are full of sequences that derive from retroviruses and retrotransposons, some of which have been recruited to perform cellular functions (Smit 1999; Makałowski 2000; Nekrutenko and Li 2001; Britten 2006). Not only sequences derived from retroviral long terminal repeats (LTRs) act as promoters of some cellular genes (Stavenhagen and Robins 1988; Ting et al. 1992; Ling et al. 2002), but also some coding sequences from retrotransposons and retroviruses have been coopted to perform functions for the host. Among the best-known cases are those of the primate *syncitin* gene, essential for placentation, and the murine *Fv1* and *Fv4* genes, involved in resistance against retrovirus infection (Ikeda et al. 1985; Best et al. 1996; Kozak and Chakraborti 1996; Qi et al. 1998; Mi et al. 2000; Goff 2004; Bonnaud et al. 2005). Protection against infection was also hypothesized to be the function of *GIN1* (*Gypsy integrase 1*), a cellular gene derived from the integrase of an LTR retrotransposon (Lloréns and Marín 2001). More recently, several other genes of unknown functions derived from retroviral or retrotransposon sequences have been characterized in vertebrates (Zdobnov et al. 2005; Campillos et al. 2006).

When we recently performed a search for genes related to *GIN1*, we detected another mammalian gene with a similar integrase domain, which we have called *Cousin of GIN1* (*CGIN1*; formerly *KIAA1305*). In our species, it is located in chromosome 14q11.2 and encodes for a 1,898-amino-acid-long protein. Human *CGIN1* is widely expressed, according to the data compiled in UniGene (http://www.ncbi.nlm.nih.gov/UniGene/). *CGIN1* genes, very similar in sequence and structure, were found to be restricted to mammals, including the marsupial *Monodelphis domestica* (opossum; see supplementary results and supplementary fig. 1, Supplementary Material online). However, we did not detect any *CGIN1* gene in monotremes, such as the platypus, *Ornithorhynchus anatinus*. These results suggest that *CGIN1* emerged after the monotreme split from the rest of mammals, but before the marsupial–eutherian split, that is, 125–180 Ma.

Key words: retrovirus, integrase, cooption, viral resistance.

E-mail: imarin@ibv.csic.es.

In phylogenetic analyses using the sequences of integrase domains (see supplementary methods, Supplementary Material online), CGIN1 integrase domains appeared as a monophyletic group in an intermediate position between the integrases of retroviruses and gypsy retrotransposons (fig. 1). The sequences most similar to CGIN1 integrase domains were a few integrases detected in fishes and birds (fig. 1). Our findings refute a previous description of the gene *CGIN1* as being related to Sushi retrotransposons (Youngson et al. 2005, which called the gene "Sushi-14C1"). Figure 1 shows that the integrases of Sushi elements and the CGIN1 integrase domains are totally unrelated. We found that these *CGIN1*-like sequences corresponded to endogenous retroviruses (ERVs; see supplementary results, Supplementary Material online). Additional analyses using reverse transcriptase sequences confirmed that the *CGIN1*-like sequences group with retroviruses and not with gypsy retrotransposons (supplementary fig. 2, Supplementary Material online). The simplest hypothesis to explain these results is therefore that part of *CGIN1* has a retroviral origin. The structure of the protein encoded by *CGIN1* is complex. Combining Blast, Prosite, and InterProScan analyses (see supplementary methods, Supplementary Material online), we determined that the gene contains four regions related to domains found in other proteins (amino acids 24–196, 790–926, 1308–1446, and 1609–1730, respectively, in human CGIN1 protein). The first conserved domain, so far undescribed and that we have called "CGIN1 domain," is present in two other human proteins, encoded by the genes *N4BP1* and *KIAA323*, as well as in the proteins encoded by the orthologs of those two genes in other species. The second conserved region corresponds to an NYN domain, a domain of unknown function described by Anantharaman and Aravind (2006) in multiple eukaryotic and prokaryotic proteins. Experimental data for NYN domain functions are not yet available. The third and fourth domains in CGIN1 contain an RNase H fold. The third domain may correspond to a highly divergent RNase H. The fourth corresponds to the integrase, already mentioned.

Figure 2 shows the structures deduced for all the human proteins that contain NYN domains. Phylogenetic analyses with NYN domain sequences indicate that *CGIN1* and *KIAA323* are closely related (see supplementary fig. 3, Supplementary Material online). This is confirmed by the structures of the two genes, which only differ significantly
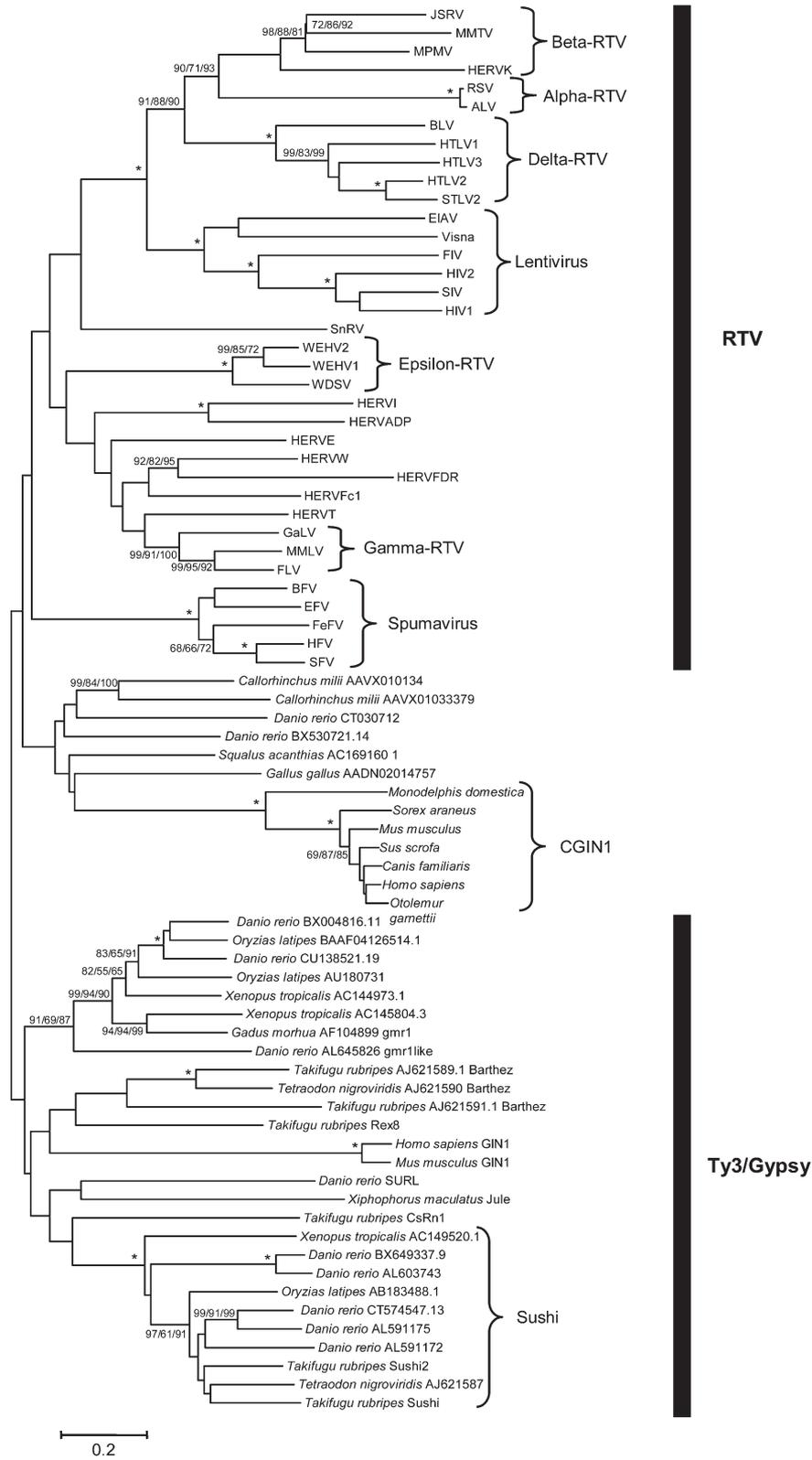
FIG. 1.—Neighbor-Joining (NJ) tree derived from integrase sequences, which includes also the Maximum Parsimony (MP) and Maximum Likelihood (ML) results (see supplementary methods, Supplementary Material online). RTV: retroviruses. Ty3/Gypsy: Gypsy retrotransposons. Numbers refer to bootstrap support for the branches, in percentage, for the three methods of phylogenetic reconstruction (NJ/MP/ML). Asterisks indicate a 100% support in the three independent analyses. Values are shown only for branches for which the three bootstrap percentages were greater than 50%.
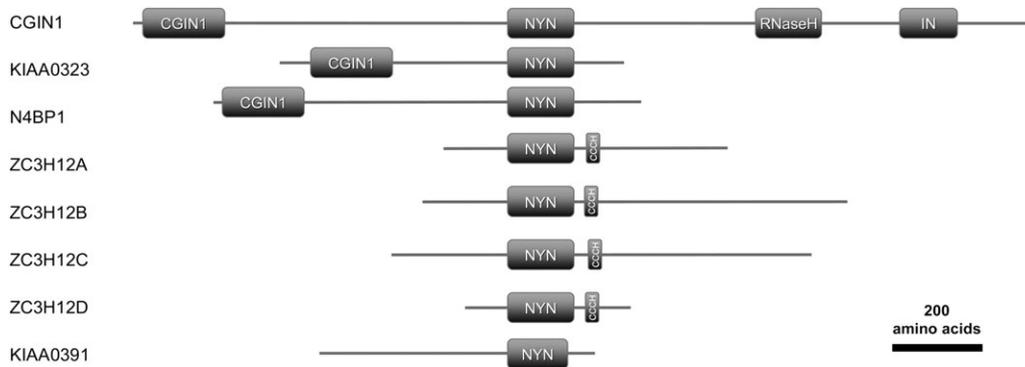
FIG. 2.—Structures of human NYN domain–containing proteins. CGIN1: CGIN1 domain; NYN: NYN domain; RNaseH: Ribonuclease H domain; IN: Integrase domain; and CCCH: C₃H zinc finger.

in their final exons. The last exon of *CGIN1* contains the sequences of retroviral origin (i.e., both the putative RNAse H domain–encoding sequences and the integrase domain–encoding sequences), whereas the last exon of *KIAA0323* lacks those sequences (see supplementary fig. 1, Supplementary Material online). *KIAA0323* is also mammalian specific. However, it is present not only in marsupials and eutherians, as *CGIN1*, but also in monotremes. It is therefore older than *CGIN1*. Significantly, *KIAA0323* is found adjacent to *CGIN1* in the human genome, in the same strand and orientation. These results indicate that *CGIN1* is a *KIAA0323* duplicate that suffered the substitution of its last exon by a fragment of an ERV. The precise way of *CGIN1* emergence, as the product of a duplication plus a recombination event leading to the fusion of sequences of different origin, is identical to the one that we described some time ago for the *PARC* gene (Marín and Ferrús 2002; Marín et al. 2004). However, in the case of

PARC, recombination merged two genes that encoded potentially interacting proteins. That made reasonable to postulate that such fusion was a secondary event that provided the advantage of avoiding the independent regulation of two genes whose products could be needed in the same tissues and potentially at the same levels (Marín et al. 2004). In the case of *CGIN1*, such interpretation cannot be proposed: It is a novel addition to the repertoire of mammalian genes and may thus provide an innovative function.

Figure 3 shows an alignment of the integrase domain encoded by *CGIN1* and the sequences of several other integrases. In figure 1, we demonstrated that the integrase domain of CGIN1 has a sequence that is quite dissimilar to that of other integrases. Data in figure 3 show that such dissimilarity has functional implications. One of the characteristic features of the catalytic core of active integrases, the DDE motif, which is present not only in retroviral
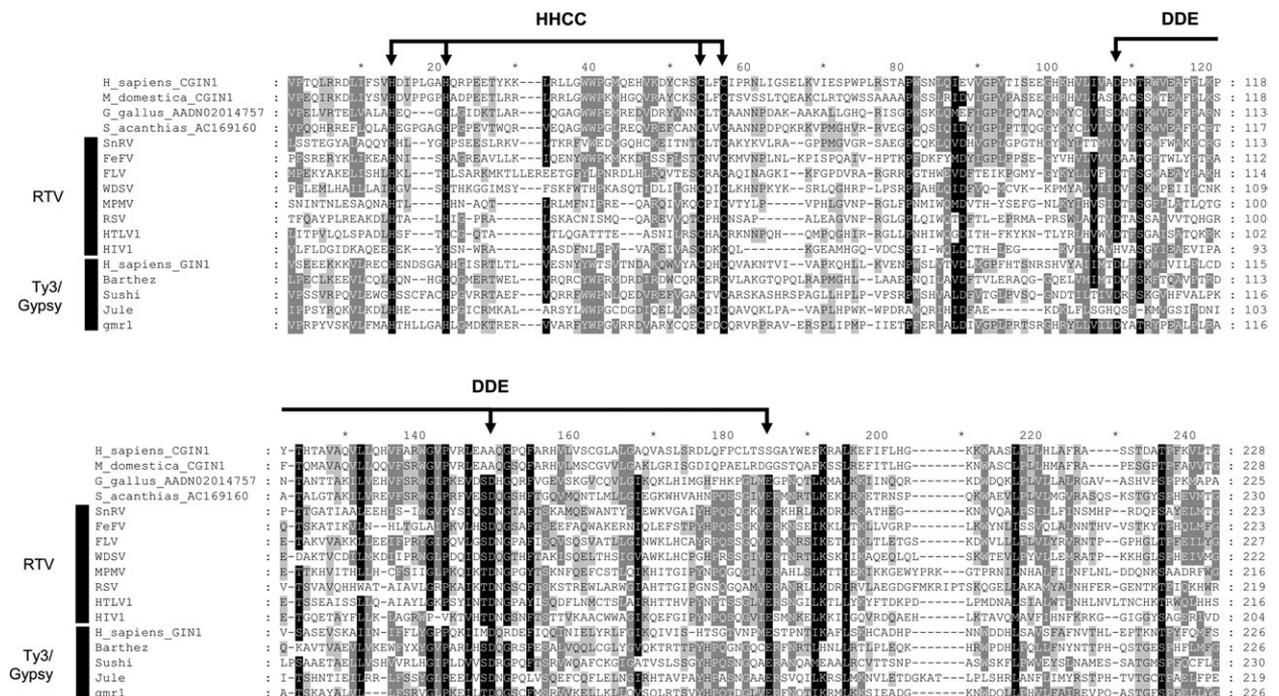

FIG. 3.—Sequences of representative CGIN1, CGIN1-like, retroviral and retrotransposon integrase sequences. The locations of the HHCC and DDE domains are indicated. Arrows point to the critical residues that gave name to those domains. Notice that CGIN1 proteins lack the two last acidic residues of the DDE motif.

integrases but also in eukaryotic and prokaryotic transposases, and is required for integrase activity (Haren et al. 1999), is missing in CGIN1. Two of the key amino acids have suffered nonconservative substitutions. This means that CGIN1 protein most probably lacks integrase activity. However, the critical residues in the HHCC domain, involved in integrase multimerization (see again the review by Haren et al. 1999), are intact. A model of the 3D structure of the integrase domain of CGIN1 suggests that it folds as a typical integrase, except in the DDE motif (supplementary fig. 4, Supplementary Material online).

We may ask which could be the function of *CGIN1* based on what is known of related genes. Some functional data exist for the N4BP1 protein, involved in the regulation of ubiquitination through its interaction with the ubiquitin ligase Itch. Oberst et al. (2007) showed that N4BP1 physically interacts with Itch, inhibiting further interactions with Itch substrates. We hypothesize that *CGIN1* function may also be linked to the ubiquitination machinery, leading to a role in retroviral control. The enzymatically inactive integrase domain of CGIN1 could be incorporated into multimeric integrase complexes. After that (and given the inhibitory role described for N4BP1), CGIN1 might interfere with integrase complex ubiquitination and degradation. This may lead to repression of viral expression. It has been shown that ubiquitination and degradation of HIV1 integrase is essential for transcription of viral genes after provirus integration (Mousnier et al. 2007). Interestingly, we suggested a related mechanism for *GIN1*, which may explain the paucity of active Gypsy elements in mammals (Lloréns and Marín 2001). Future experimental work may establish whether this hypothesis for CGIN1 protein function is correct.

Supplementary Material

Supplementary results, including four supplementary figures, and supplementary methods are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

Literature Cited

Anantharaman V, Aravind L. 2006. The NYN domains: novel predicted RNases with a PIN domain-like fold. RNA Biol. 3:18–27.

Best S, Le Tissier P, Towers G, Stoye JP. 1996. Positional cloning of the mouse retrovirus restriction gene Fv1. Nature. 382:826–829.

Bonnaud B, Beliaeff O, Bouton J, Oriol G, Duret L, Mallet F. 2005. Natural history of the ERVWE1 endogenous retroviral locus. Retrovirology. 2:57.

Britten R. 2006. Transposable elements have contributed to thousands of human proteins. Proc Natl Acad Sci USA. 103:1798–1803.

Campillos M, Doerks T, Shah PK, Bork P. 2006. Computational characterization of multiple Gag-like human proteins. Trends Genet. 22:585–589.

Goff SP. 2004. Retrovirus restriction factors. Mol Cell. 16:849–859.

Haren L, Ton-Hoang B, Chandler M. 1999. Integrating DNA: transposases and retroviral integrases. Annu Rev Microbiol. 53:245–281.

Ikeda H, Laigret F, Martin MA, Repaske R. 1985. Characterization of a molecularly cloned retroviral sequence associated with Fv-4 resistance. J Virol. 55:768–777.

Kozak CA, Chakraborti A. 1996. Single amino acid changes in the murine leukemia virus capsid protein gene define the target of Fv1 resistance. Virology. 225:300–305.

Ling J, Pi W, Bollag R, Zeng S, Keskintepe M, Saliman H, Krantz S, Whitney B, Tuan D. 2002. The solitary long terminal repeats of ERV-9 endogenous retrovirus are conserved during primate evolution and possess enhancer activities in embryonic and hematopoietic cells. J Virol. 76:2410–2423.

Lloréns C, Marín I. 2001. A mammalian gene evolved from the integrase domain of an LTR retrotransposon. Mol Biol Evol. 18:1597–1600.

Makałowski W. 2000. Genomic scrap yard: how genomes utilize all that junk. Gene. 259:61–67.

Marín I, Ferrús A. 2002. Comparative genomics of the RBR family, including the Parkinson's disease-related gene parkin and the genes of the ariadne subfamily. Mol Biol Evol. 19:2039–2050.

Marín I, Lucas JI, Gradilla AC, Ferrús A. 2004. Parkin and relatives: the RBR family of ubiquitin ligases. Physiol Genomics. 17:253–263.

Mi S, Lee X, Li X, et al. (12 co-authors). 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. Nature. 403:785–789.

Mousnier A, Kubat N, Massias-Simon A, Ségéral E, Rain JC, Benarous R, Emiliani S, Dargemont C. 2007. von Hippel Lindau binding protein 1-mediated degradation of integrase affects HIV-1 gene expression at a postintegration step. Proc Natl Acad Sci USA. 104:13615–13620.

Nekrutenko A, Li W-H. 2001. Transposable elements are found in a large number of human protein-coding genes. Trends Genet. 17:619–621.

Oberst A, Malatesta M, Aqeilan MRI, et al. (12 co-authors). 2007. The Nedd4-binding partner 1 (N4BP1) protein is an inhibitor of the E3 ligase Itch. Proc Natl Acad Sci USA. 104:11280–11285.

Qi CF, Bonhomme F, Buckler-White A, Buckler C, Orth A, Lander MR, Chattopadhyay SK, Morse 3rd HC. 1998. Molecular phylogeny of Fv1.. Mamm Genome. 9:1049–1055.

Smit AF. 1999. Interspersed repeats and other mementos of transposable elements in mammalian genomes. Curr Opin Genet Dev. 9:657–663.

Stavenhagen JB, Robins DM. 1988. An ancient provirus has imposed androgen regulation on the adjacent mouse sex-limited protein gene. Cell. 55:247–254.

Ting CN, Rosenberg MP, Snow CM, Samuelson LC, Meisler MH. 1992. Endogenous retroviral sequences are required for tissue-specific expression of a human salivary amylase gene. Genes Dev. 6:1457–1465.

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, Campillos M, Harrington ED, Torrents D, Bork P. 2005. Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. Nucleic Acids Res. 33:946–954.