Maintenance in the Chicken Genome of the Retroviral-like *cENS* Gene Family Specifically Expressed in Early Embryos

Emmanuelle Lerat · Anne-Marie Birot · Jacques Samarut · Anne Mey

Received: 19 December 2006 / Accepted: 18 May 2007 / Published online: 2 August 2007

Abstract Embryonic stem (ES) cells are important developmental cells that appear very early during development and subsequently give rise to all the cell lineages of the future adult organism. In these cells a limited subset of transcription factors is expressed that are well conserved among species and essential for the fate of the stem cell. The transcriptome analysis of ES cells from chicken has revealed a gene family, *cENS*, that is specifically expressed in ES cells and in early embryos and is repressed during the differentiation process. This family is characterized by displaying retroviral structures and shares no homology with other species' genes. These characteristics are probably not restricted to the chicken genome and raise the question of whether similar genes are present and have been maintained in other species. We have examined the different copies of this gene in the sequenced chicken genome to investigate its dynamics and its evolution. We have distinguished two groups of *cENS*-related copies. The first group, resulting from recent transposition events, contains the transcribed *ENS-1* and *ENS-3* plus copies subjected to negative selection pressures. The second group contains degenerate copies that were integrated into the genome earlier. Comparison with copies previously isolated from three Galliformes showed that they are also subjected to selection pressures. We also detected numerous solo-LTRs containing the *ENS-1* promoter that may control the expression of host genes. Taken together, these findings suggest a function sustained by a neogene of retroviral origin during the early stages of chicken development.

Keywords Retrovirus · Embryonic stem cell · Chicken genome · Sequence analysis

Reviewing Editor: Dr. Nicolas Galtier

Electronic supplementary material The online version of this article (doi: 10.1007/s00239-007-9001-4) contains supplementary material, which is available to authorized users.

E. Lerat
Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS, UMR 5558, 43 boulevard du 11 novembre 1918, Villeurbanne F-69622, France

A.-M. Birot · J. Samarut · A. Mey
Institut Fédératif Biosciences Gerland Lyon Sud, Université de Lyon, Lyon, F-69003, France, and Institut de Génomique Fonctionnelle de Lyon, Université Lyon 1, CNRS, INRA, Ecole Normale Supérieure de Lyon, F-69364 Lyon, France

E. Lerat (✉)
Laboratoire Biométrie et Biologie Evolutive, Université Claude Bernard - Lyon 1, UMR-CNRS 5558, Bat. Mendel,
69622 Villeurbanne, cedex, France
e-mail: lerat@biomserv.univ-lyon1.fr

A. Mey (✉)
Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure de Lyon, UMR CNRS 5242 - UMR INRA 1288,
46 Allée d'Italie, 69364 Lyon cedex 07, France
e-mail: Anne.Mey@ens-lyon.fr

Introduction

Endogenous retroviruses in birds have been divided into three groups: the *ev* loci close to avian sarcoma/leukosis viruses (ASLVs) and restricted to domestic chicken and its wild relatives, the endogenous avian retrovirus family (AEV) that is found only in the four *Gallus* species, and the human endogenous retrovirus type I-related retroviruses found in almost all vertebrates (Borisenko 2003). Recently, Wicker et al. (2005) estimated the repetitive sequences that can be found in the chicken genome using a technique of Cot-based cloning and sequencing. Their work has made it possible to identify other retrovirus-like elements and LTR retrotransposons. In particular, they identified an element they called *Soprano*, which was already known as *cENS* (for chicken Embryonic Normal Stem cell gene) (Acloque et al. 2001). The expression of this gene, also described as *erni* in the neural plate (Streit et al. 2000), is restricted to the early stages of chick embryonic development and to the undifferentiated embryonic stem (ES) cells isolated from chicken epiblasts (Acloque et al. 2001). This gene appeared to belong to a multigene family. The characterization of three copies cloned from a cDNA expression bank established that the *cENS* family might have some links with retroviral sequences (Acloque et al. 2001). The first determined copy, *ENS-1*, which encodes the protein known as ENS-1/ERNI, corresponds to a single open reading frame (ORF) (the *ENS* ORF), which is devoid of introns and surrounded by direct repeats corresponding to retroviral long terminal repeat (LTR). We have characterized the distribution of the protein ENS-1 in chicken ES cells using specific antibodies. ENS-1 has been found in the nucleus where it binds to a protein controlling gene expression. The functional consequences of this interaction are under investigation (unpublished data). Another sequence, known as *ENS-3*, displayed potentially active *pol*- and *env*-like sequences in addition to the *ENS* ORF, while the third copy, *ENS-2*, was a truncated sequence of the *ENS* ORF. It has been demonstrated that the LTR sequences of *ENS-1* contain the promoter that gives the gene its highly specific transcription pattern in the undifferentiated chicken ES cells (Acloque et al. 2004). Another puzzling characteristic of *cENS* is that the *ENS* ORF has no homology with any other known gene, even though it has been detected in various species in the Galliforme group to which the chicken belongs (Acloque et al. 2001).

Numerous copies of the *Soprano* element have been identified, including 75 containing internal domains and one with a potentially intact *ENS* ORF (Wicker et al. 2005). These findings have confirmed the retroviral or retrotransposable origin of the various members of the *cENS* gene family. Several examples have shown that transposable elements or retroviral sequences are able to promote the creation of new genes (for review see Long 2001). The origin of the *ENS* ORF is therefore not clear, and it could result from the fusion of various host protein domains, it could be a gene created *de novo* and subsequently integrated into a viral sequence, or it could be a viral gene that has evolved to fulfill an important function within its host and is no longer recognizable.

The comparison between human and murine ES cells at the transcription level has revealed that both species express a limited subset of common genes thought to be involved in the self-renewal of ES cells and which are repressed during differentiation (Sato et al. 2003; Wei et al. 2005). However, a large subset of genes, strictly associated with the undifferentiated state of ES cells, appeared to be species-specific. The specific expression pattern of the neogene *cENS* during early development raises the question of how such genes originally appeared and were maintained in the Galliformes. Indeed, ES cells are unique, self-renewing cells that can give rise to all cell lineages, including germline cells, during embryogenesis. This means that these cells sustain the existence of all descendants. The expression in ES cells of mobile elements or retroviruses has already been demonstrated in the mouse (Maksakova and Mager 2005; Peaston et al. 2004). In addition to investigating the functionality of the ENS-1 protein in chicken ES cells, understanding the origin, the evolution, and the maintenance of *ENS-1* in the chicken genome and among the Galliformes in general can be expected to lead to new insights into neogene formation in vertebrate species and into the acquisition of major functions as a result of retroviral combinations. However, the possibility remains that this gene has been conserved for its potentiality to replicate itself. In particular, it would be a good strategy for a mobile element to be expressed in the ES cells that give rise to the germline in order for new insertions to be transmitted to the descendants.

To clarify the situation, we identified and analyzed all copies of *cENS* present in the sequenced chicken genome (International Chicken Genome Sequencing Consortium 2004) to pinpoint where it occurs within the genome and to clarify its evolutionary history. The findings reported here demonstrate that this gene family is quite an ancient component of the Galliforme genomes, a group that emerged about 100 million years ago (van Tuinen and Hedges 2001). In the chicken, the copy of the full-length *ENS* ORF seems to result from a rather recent insertion event, which indicates that the retroviral system was still active in the recent past. This study demonstrates that this gene is subjected to negative selection pressure to be maintained in the chicken genome and suggests that counterparts could well exist in other Galliformes.

Materials and Methods

Identification and Analysis of the *cENS* Copies in the Chicken Genome

We used the genome sequence of the chicken *Gallus gallus* (version WASHUC 1, March 2004, retrieved from Ensembl at http://www.ensembl.org/index.html) (International Chicken Genome Sequencing Consortium 2004). The complete sequences of *Soprano* (obtained from the supplementary online data of Wicker et al. 2005), *ENS-1* (GenBank accession No. AF327879), and *ENS-3* (GenBank accession No. AF329451) were used as queries to search for the copies in the chicken genome using BLASTN (Altschul et al. 1997). The different copies were aligned with each of the query sequences to determine the various features of the sequences (ORFs, LTRs). The positions of the copy with regard to genes were determined using the web interface of Ensembl (http://www.ensembl.org/) (Hubbard et al. 2005). The nomenclature we used to name the copies detected was of the GGX_Y type, where GG stands for *Gallus gallus* and X corresponds to the chromosome number. Y corresponds to the section on the chromosome after the chromosomes had been split into 100-Mb portions by Ensembl.

The identity percentages of *ENS* ORF, *pol*, and *env* were computed by comparison to the *ENS* ORF of *ENS-1* and to the *ENS-3* retroviral ORFs using the *dnadist* program from the PHYLIP package version 3.6 (Felsenstein 2002). The insertion date estimation was computed using the method described by Bowen and McDonald (2001). When an LTR retrotransposon is inserted into the genome, the mechanism of insertion implies that both LTRs are identical. Then after the insertion mutations can accumulate in both LTRs and it is thus possible to estimate the date of insertion of a given copy by comparing its two LTRs. The more divergent the two LTRs, the more ancient is the insertion. Intraelement LTR divergence was calculated using the Kimura-2 parameter method in the *dnadist* program from the PHYLIP package version 3.6 (Felsenstein 2002). Ages were computed using the formula $T = K/(2r)$, where T is the time of divergence, K is the divergence, and r is the substitution rate (Li 1997). We used 0.0036 substitutions per site per million years as the value of the substitution rate for the autosomal chromosomes; this was estimated from gene comparisons between the chicken and the turkey (*Meleagris galopavo*), taking the divergence time to be 28 million years before present (Axelsson et al. 2004).

Estimation of the Synonymous and Nonsynonymous Rates

We used the program PAML (Yang 1997) to perform a phylogenetic analysis of the synonymous and nonsynonymous rates using the five potentially complete ORFs (GG1_186 and GG5_48 and the cloned sequences of the grey partridge, the quail, and the helmeted guineafowl). We tested several models based on the phylogeny presented in Fig. 3. Model M0 considers the *dN/dS* ratio to be the same for all lineages. Model M1 discriminates on the one hand the two chicken sequences and on the other hand the three sequences of the other galliformes compared to the two groups of sequences in each case. The two models were compared by likelihood ratio tests.

Phylogenetic Reconstruction of the *ENS* Copies in Chicken and Other Galliformes

The DNA sequences of the *ENS* ORFs of the different copies were aligned using Clustalw version 1.83 (Thompson et al. 1994) and were manually corrected using the sequence editor SEAVIEW (Galtier et al. 1996). Some copies were too short to be used in the tree phylogeny reconstruction so we eliminated them.

To determine the relative position of each of the copies in the Gallinaceae where *cENS* had previously been detected, we included in the alignment the DNA sequences of the cloned *cENS* from various galliform species obtained from Acloque et al. (2001). These were *Gallus gallus* (the same chicken species as the sequenced genome), *Coturnix coturnix* (the common quail), *Alectoris rufa* (the red-legged partridge), *Perdix perdix* (the grey partridge), *Numida meleagris* (the helmeted guineafowl), *Meleagris gallopavo* (the wild turkey), and *Phasianus colchicus* (the common pheasant). The species phylogeny of the Gallinaceae used in the study was reconstructed using the nucleic acid sequences of the *cytB* gene, as proposed by Kimball et al. (1999).

We obtained each tree using a maximum-likelihood method with the HKY model of substitution (Hasegawa et al. 1985) and the gamma correction with 500 bootstrap replicates implemented in the PHYML program (Guindon and Gascuel 2003).

Phylogenetic Tree of *pol* Sequences

To reconstruct the phylogenetic relationships of the *pol* genes, we retrieved from GenBank amino acid retroviral and retrotransposable *pol* sequences of the following retroelements: HCML-ARV (AF499232), HERV-E (M10976), IPHA (P04026), RV-koala (AF151794), ERV3 (M12140), HERV-W (AY101585), AtSV (DQ174103), *Python-molurus*_ERV (AAN77283), IPMAI (X04120), HERVH-RGH2 (D11078), *412* (X04132), MuERV-L (Y12713), ALV (M37980), MMTVB (AF033807), MPMV (AF033815), SRV2 (M16605), JSRV (M80216), HTLV-2 (M10060), HTL1C (AF033817), BLVJ (K02120), SnRV (U26458), Xen1 (AJ506107), WDSV (AF033822), *297* (X03431), *gypsy* (AF033821), *17.6* (X01472), SMRVH (M23385), MLVRK (M93052), Reticuloendotheliosis virus (Reticul ABC26820), MLVFF (Z11128), GALV (M26927), FLV (AF052723), BAEVM (D10032), PERV (AJ293656), FIVPE (M25381), BIV06 (M32690), Visna

(S55323, AY101611), HIV-2 (M30502), OMVVS (M34193), OvRV (AF479638), CAEVC (M33677), EIAV9 (M16575), HIV-1 (K03455), SIVCZ (L40990, L06042, AJ580407), SIVVT (X07805), SFV1 (X54482), HFV (Y07725), HSRV (AF033816), BFV (U94514), and FFV (AJ564745). Other sequences were retrieved from the supplementary data of the study by Jern et al. (2005) (HERV-ADP and HERV-T).

The alignment of the amino acid sequences was created using the T-coffee program version 3.27 (Notredame et al. 2000), and the conserved blocks in the alignment were then selected using Gblocks version 0.91b (Castresana 2000). Tree reconstruction was done using the maximum-likelihood method with the JTT model of substitution (Jones et al. 1992) and the gamma correction with 500 bootstrap replicates implemented in the PHYML program (Guindon and Gascuel 2003).

Phylogenetic Tree of *env* Sequences

The *env* protein sequences used for the alignments and the tree reconstruction were retrieved from GenBank: HCML-ARV (AAP06678), ERV-R (*Homo sapiens*, Q14264; *Hylobates moloch*, CAI15392; *Pan troglodytes*, CAI15390; *Pongo pygmaeus*, CAI15391), ERV3 (NP_001007254), HERV-FRD (*Gorilla gorilla*, CAE12263; *P. troglodytes*, CAE12264; *H. moloch*, CAE12265; *Macaca fascicularis* CAE12266; *Callithrix jacchus*, CAE12267; *H. sapiens*, P60508), Avian Leukosis Virus (AAX18665), and syncitin A (*Clethrionomys glareolus*, AAW62448; *Mesocricetus auratus*, AAW62449; *Rattus norvegicus*, AAW62447; *Mus musculus*, AAW62446). Another hit was obtained with *FET-1* (AAM52407), which corresponds to a novel gene known as Female Expressed Transcript 1, which is expressed only in females and is upregulated in the cortex of the left gonad during the sex-determining period (Reed and Sinclair 2002). Blocks of similarities were determined using the MEME program (Bailey and Elkan 1994). The resulting alignment was used to perform a tree reconstruction using the maximum-likelihood method with the JTT model of substitution (Jones et al. 1992) and the gamma correction with 500 bootstrap replicates implemented in the PHYML program (Guindon and Gascuel 2003).

Results

Copies of *cENS* in the Chicken Genome

A total of 46 sequences corresponding to *cENS* were detected, independent of the solo-LTRs that correspond to 874 insertions. Table 1 summarizes the sequence characteristics and the position of the 46 sequences with internal domains. The discrepancy with the copy number detected by Wicker et al. (2005) may be explained by insertion polymorphism between the sequenced genome and the biological material they used, but also by differences in the match selection criteria. Wicker et al. (2005) did indeed use less stringent criteria. Furthermore, some gaps in the present sequenced genome assembly may also have prevented us from detecting other insertions. We defined different kinds of copies by comparing their features to those of the *ENS-1* and *ENS-3* ORFs.

ENS-1 and ENS-3-like copies

These copies correspond to the first ten rows in Table 1. Their structures are shown in Fig. 1. They all contain the *ENS* ORF with a nucleic sequence identity to the ORF of *Soprano/ENS-1* ranging from 87.51% to 100%. Most of the sequences are interrupted by different indels (insertions and deletions) that disrupt the reading frames of the ORF. These are the most highly conserved copies of *cENS*.

The GG1_186 copy (see nomenclature in the Materials and Methods section), located on chromosome 1, corresponds to the complete sequence of *Soprano*. Its coding sequence is intact and displays 98.37% identity with the cloned *ENS* ORF. It is flanked by two 920-bp LTRs with 99.78% identity to each other. The divergence between the two LTRs of this copy indicates that the insertion is recent, with an age estimation of about 0.3 million year (MYR), which signifies that a specific transposition event has occurred in the chicken lineage. Analysis of the sequence between the *ENS* ORF and the 3′ LTR identified remains of the *pol* and *env* genes. They correspond to a 50-bp sequence with 94% identity to the *pol* and to a 57-bp sequence with 86% identity to the *env*, both of which are found in *ENS-3*.

Another copy containing a complete *ENS* ORF has been found on chromosome 5. The ORF of GG5_48 displays 98.98% identity to the *ENS* ORF of GG1_186. The identity between the two LTRs however is lower (98.90%), indicating a more ancient insertion date that is estimated as 1.5 MYR. This copy is inserted into the eighth intron of gene ENSGALT00000011858, which codes for a member of voltage-gated potassium channel subfamily H. This copy has the same *pol* and *env* relics as GG1_186 with high identity percentages. The same relics are also found in four other copies, GG3_98, GG1_55, GG2_77, and GG1_28, which are deleted forms of *ENS-1*. These four copies display 96.49%–100% identity between their *env* relics and that of GG1_186, whereas the mean identity to the complete *env* gene of *ENS-3* is only 86.50%. Two other copies,

Table 1 *ENS* copies in the chicken genome

	Name	Chr	Start	Stop	Strand	% identity *ENS* ORF	Length (bp)	% identity *pol*	Length (bp)	% identity *env*	Length (bp)	% identity LTR5′–LTR3′	Age (Myr)
ENS-like	GG1_186[c]	1	95459527	95464200	+	100.00	1470	94.00	50	85.96	57	99.78	0.3
	GG5_48[c]	5	50372942	50377518	–	98.98	1470	98.00	50	86.21	58	98.90	1.5
	GG2_118[a]	2	71452745	71460210	+	95.98	906	100.00	3072	100.00	1460	98.68	1.9
	GG3_98	3	88080492	88083383	–	97.47	522	96.00	50	87.93	58	99.74	0.4
	GG1_63	1	155175037	155178510	–	96.19	1465	98.00	50	92.70	178	N/A	N/A
	GG1_55	1	147918766	147923235	+	87.51	1111	N/A	N/A	86.21	58	97.84	3.0
	GG2_77	2	35364285	35367058	–	98.10	260	96.00	50	86.21	58	98.30	2.4
	GG1_28	1	123659390	123663798	–	97.67	1459	98.00	50	87.93	58	98.32	2.4
	GG2_142	2	93122707	93124496	+	96.42	589	97.80	45	N/A	N/A	ND	ND
	GG2_36	2	131746322	131748499	+	97.63	653	N/A	N/A	N/A	N/A	1 LTR	N/A
Degenerate *ENS1/ENS3*	GG15_4	15	12403564	12410486	–	80.09	1362	89.20	1966	86.00	1435	89.63	15.4
	GGZRD_6	ZRD	14038977	14041654	+	73.45	1021	79.00	1902	N/A	N/A	N/A	N/A
	GG3_1a	3	587	4017	+	82.91	239	78.70	95	78.30	73	79.78	34.8
	GG4_16	4	23098583	23099931	+	81.00	1385	N/A	N/A	85.20	177	N/A	N/A
	GG1_49	1	141875338	141878120	+	79.76	250	72.30	85	82.20	169	N/A	N/A
	GG1_46	1	140293331	140297159	+	80.47	1467	N/A	N/A	81.80	372	N/A	N/A
	GG2_97	2	53295178	53298276	–	81.68	1165	91.10	45	76.50	102	N/A	N/A
	GG8_30	8	8895981	8901922	+	73.05	505	84.20	2567	N/A	N/A	88.1	17.8
	GG4_51	4	54925516	54927933	–	80.72	947	76.80	151	N/A	N/A	N/A	N/A
	GG4_21[b]	4	28453568	28456619	–	N/A	N/A	N/A	N/A	N/A	N/A	99.65	0.5
	GG1_179	1	89081086	89082355	+	95.73	1270	N/A	N/A	N/A	N/A	N/A	N/A
	GG1_56a	1	148554621	148555062	–	97.74	442	N/A	N/A	N/A	N/A	N/A	N/A
	GG1_56b	1	149034543	149035073	–	82.7	531	N/A	N/A	N/A	N/A	N/A	N/A
	GG1_56c	1	149248756	149249428	+	79.67	673	N/A	N/A	N/A	N/A	N/A	N/A
	GG5_51	5	53430443	53434029	–	83.40	1447	83.00	83	80.00	56	N/A	N/A
	GG1_64	1	155436066	155438241	+	83.16	1468	N/A	N/A	N/A	N/A	1 LTR	N/A
	GG1_80	1	17522086	17524319	–	82.86	1410	N/A	N/A	N/A	N/A	ND	ND
	GG1_40	1	134449688	134452994	–	82.11	1060	N/A	N/A	N/A	N/A	1 LTR	N/A
	GG2_145	2	96105457	96106208	+	82.22	752	N/A	N/A	N/A	N/A	N/A	N/A
	GG1_178	1	88487852	88495267	+	78.94	1935	N/A	N/A	N/A	N/A	1 LTR	N/A
	GGZ_20	Z	27715599	27718385	–	81.03	505	N/A	N/A	N/A	N/A	1 LTR	N/A
	GG2_72	2	30282551	30283790	+	73.18	1059	N/A	N/A	N/A	N/A	1 LTR	N/A
	GG1_156	1	68721440	68721687	–	87.90	248	N/A	N/A	N/A	N/A	N/A	N/A
	GG6_8a	6	1658371	16583900	+	85.16	182	N/A	N/A	N/A	N/A	N/A	N/A
	GG4_12	4	20358944	20359141	+	87.88	198	N/A	N/A	N/A	N/A	N/A	N/A
	GGZRD_10	ZRD	4920215	4922205	–	N/A	N/A	80.40	793	76.50	1168	N/A	N/A
	GG2_99a	2	54810664	54812555	+	N/A	N/A	82.10	263	74.10	817	N/A	N/A
	GG3_1b	3	293355	294181	+	N/A	N/A	76.00	798	N/A	N/A	N/A	N/A
	GG1_56d	1	149194817	149196923	–	N/A	N/A	77.40	178	70.10	77	1 LTR	N/A
	GG1_62	1	154234476	154231483	+	N/A	N/A	81.10	2997	N/A	N/A	N/A	N/A
	GG2_99b	2	54480053	54480496	+	N/A	N/A	74.60	691	N/A	N/A	N/A	N/A
	GG1_100	1	188225204	188225378	+	N/A	N/A	N/A	N/A	81.80	175	N/A	N/A

Table 1 continued

Name	Chr	Start	Stop	Strand	% identity *ENS* ORF	Length (bp)	% identity *pol*	Length (bp)	% identity *env*	Length (bp)	% identity LTR5′–LTR3′	Age (Myr)
GG28_2	28	2088036	2088709	–	N/A	N/A	N/A	N/A	84.40	186	1 LTR	N/A
GG1_58	1	15369309	15371615	–	N/A	N/A	N/A	N/A	83.20	125	89.04	16.5
GG6_8b	6	16593748	16595134	+	N/A	N/A	N/A	N/A	83.70	342	1 LTR	N/A
GG3_93	3	84158774	84159827	–	N/A	N/A	N/A	N/A	87.90	58	1 LTR	N/A

N/A = no corresponding sequences detected; ND = not possible to compute the % identity or the age because the sequences were too short.

[a] Copy corresponding to *ENS-3*

[b] Copy with no internal domains

[c] Copies with intact *ENS* ORF

GG1_63 and GG2_142, also display similar parts of *pol*. This could indicate that all these sequences are derived from a common ancestral copy, as it seems unlikely that exactly the same deletion could have occurred at the same position on several occasions in *pol* and *env*. Estimates for their insertion dates range from 3.0 to 0.4 MYR, which makes them quite recent insertions into the chicken genome. A partial copy, GG2_36, displays a high percentage identity to *ENS* ORF, but no viral sequences are detectable and only one LTR is present.

ENS-3 was identified as copy GG2_118, a sequence located on chromosome 2. It contains the ORFs corresponding to *pol* and *env* that appear to be potentially active, because they are complete with no interrupting mutations. The *ENS* ORF is not complete, as it displays several small insertions and deletions. It is not possible to determine whether the 5′ LTR is complete, as a portion of its sequence has not been entirely determined and corresponds to Ns. However, the 3′ LTR contains a big, 98-bp deletion. It has still been possible to compute the identity between the two LTRs, which is high (98.68%) and corresponds to an insertion date of 1.9 MYR.

We did not find any copy corresponding to *ENS-2*. It could be either that this copy does not exist in the sequenced genome or that the gaps in the present sequenced genome assembly may have prevented us from detecting this insertion. However, when we compared the nucleic sequence of the *ENS* ORF of *ENS-2* with those of the various *ENS-1*-like copies, we found very high percentage identities: 96.61% with GG1_186 and 100% with GG5_48. This indicates that the internal deletion present in *ENS-2* is very recent.

Degenerate ENS-1/ENS-3 copies

Thirty-six copies correspond to degenerate forms of *ENS-1* or *ENS-3* sequences (Table 1). Some copies are close to identified genes (see Supplementary Table 1 for details). The copies can be divided into different categories according to the sequence features they display. Two copies, GG15_4 and GG3_1a, display the three ORFs of *ENS-3*, but they are all severely degraded. The two *ENS* ORF of those copies show identity levels to the complete *ENS* ORF of 80.09% and 82.91%, respectively. The *pol* and *env* sequences of GG15_4 display 86.00% and 89.63% identity to their respective complete genes, whereas those of GG3_1a display only 78.30% and 79.78% identity. Their insertion dates can be estimated as 15.4 MYR and 37.8 MYR, respectively. Seven copies correspond to more degraded sequences of *cENS* (on average 80.01% ± 3.28% identity to the cloned *ENS* ORF). These copies reveal the presence of *pol*, *env*, or LTR, but in different ways. Eight copies correspond only to the *ENS* ORF (average percentage identity of 87.39% ± 6.44%) and do not display any viral characteristics. Six copies contain the *ENS* ORF (mean percentage identity to the complete *ENS* ORF is 80.21% ± 3.77%) and generally one LTR, but no viral genes. Eleven copies do not display an *ENS* ORF. They do, however, contain remnant sequences of the *pol* and/or *env* genes. Some of these sequences also reveal the presence of at least one LTR.

The phylogenetic relationship of the copies

We have reconstructed a phylogenetic tree based on the nucleic alignment of the *ENS* ORFs of different copies (Fig. 2). It was not possible to use all the copies containing an *ENS* ORF because some were relatively too short to be used. We thus used the sequences that were at least one third of the complete *ENS* ORF and removed the sequences of the selected size that were not alignable with the majority of the other sequences. The tree shows three significant groupings of the copies. One well-supported group contains all the copies in the ENS-like category defined in Table 1 plus two copies, GG1_56a and GG1_179, which correspond to copies displaying only the *ENS* ORF. The small branches within this group confirm that these copies have been inserted into the chicken genome recently and/or have evolved slowly. The two other groups, which have longer branches, correspond to degenerate copies that are likely to have been mobilized a long time ago and are no longer actively transposed.

The population of solo-LTRs

Comparing the complete sequence of *Soprano* with the cloned sequence of *ENS-1* revealed that the LTRs were longer than previously described (Acloque et al. 2001). They actually correspond to a sequence of 920 bp. Using the complete sequence of the 5′ LTR of the complete copy GG1_186, we searched for solo-LTRs in the genome. A solo-LTR corresponds to a remnant of a copy that has been eliminated as a result of recombination between its two LTRs. We found 874 solo-LTRs scattered throughout the genome, ranging in size from 101 to 1656 bp, the latter size resulting from internal duplication. One hundred fifty-five of the 874 solo-LTRs are inserted less than 20 kb from genes. We determined whether there were any solo-LTRs located within various maximum distances from genes (20 kb, 10 kb, and 5 kb) and, if so, determined their position relative to the gene(s) (in the 5′ region, inside the gene, in the 3′ region) and their relative orientation (antisense or sense relative to the neighboring gene). For this analysis we did not take into account solo-LTRs close to several genes at the same time. Our findings are shown in Table 2. We see that the solo-LTRs tend to be inserted inside genes in antisense orientation (χ^2 test, p = 1.321e-06) rather than upstream or downstream of the genes. However, when inserted in the flanking regions of the genes, the solo-LTRs do not exhibit any significant bias of orientation.

Some solo-LTRs, depending on the maximum distance from the genes used, were actually located in the vicinity of several genes. When we looked at insertions within 20 kb of genes, 17 of the solo-LTRs were in fact in the neighborhood of several genes. It can also be seen that some genes are close to or inserted by more than one solo-LTR.

Phylogenetic position of the pol and env genes of ENS-3 among the retroviruses

To find out to which class of retroviruses *cENS* is related, we determined the phylogenetic relationship of the *pol* and *env* genes of *ENS-3* with homologous genes from various retroviruses. The *pol* gene is the most highly conserved retroviral gene. We therefore retrieved the amino acid sequences of *pol* from different classes of retroviruses. We also included four *pol* sequences from *Drosophila* transposable elements that are known to be closely related to retroviruses. The phylogenetic tree that we obtained is shown in Fig. 3. The retroviral sequences are grouped

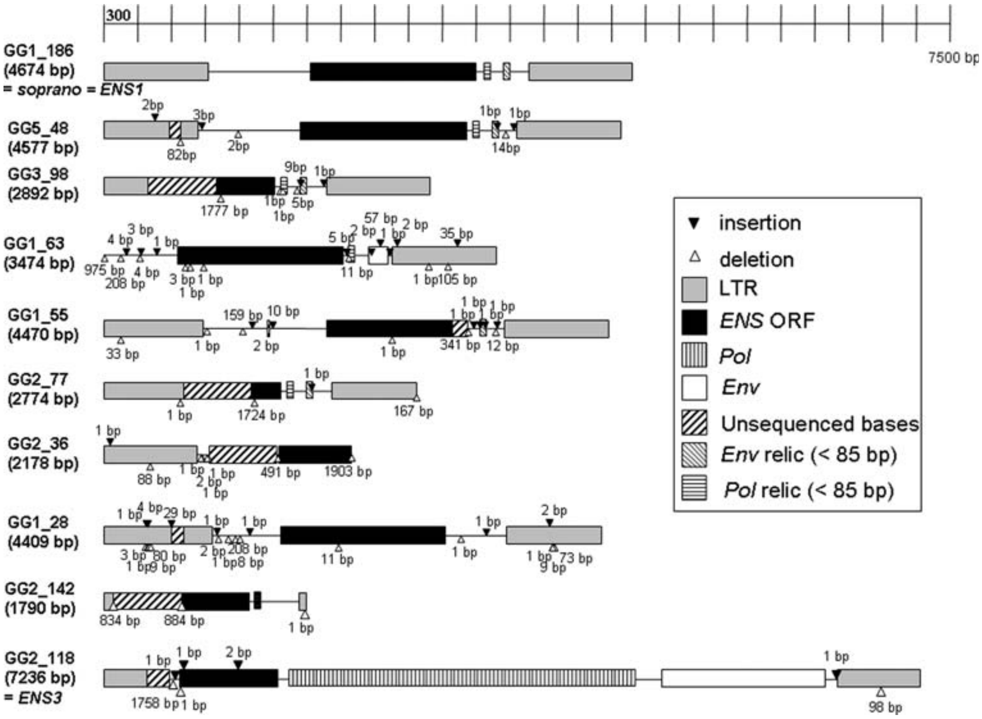

Fig. 1 Structures of the *ENS*-like copies. Insertions and deletions are given according to the pairwise comparison with *Soprano* or *ENS-3*

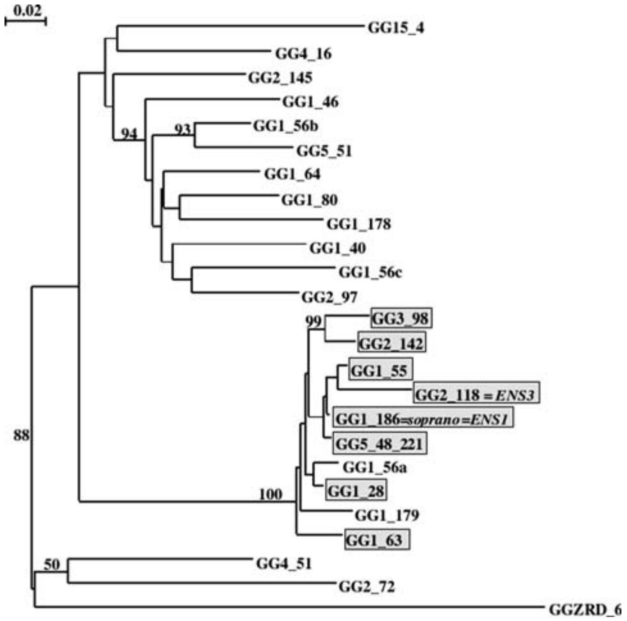

Fig. 2 Maximum-likelihood tree based on the *ENS* ORF nucleic sequences of the copies in the chicken genome. Only bootstrap values greater than 50% are given. Gray boxes correspond to the copies of the *ENS*-like category in Table 1

according to class, and we can see that the sequence of *ENS3* is close to type L endogenous retroviruses (ERV) from the mouse.

The *env* gene is less highly conserved among retroviruses. This is why we did not attempt the same approach to this gene. We made a BLAST search against the GenBank databases to produce a list of hits, and then we retrieved the amino acid sequences of all significant hits (E value < 10e-7). Because only particular regions of *env* may be conserved in retroviruses (Lerat and Capy 1999), we searched for blocks of similarities. The blocks identified were then used to produce an alignment, which we later used to perform a tree reconstruction. The result is shown in Fig. 4. The *env* gene of *ENS-3* is grouped with ERV-R and ERV3, which are found in primate genomes, and with HCML-ARV, an endogenous retrovirus associated with human chronic myeloid leukemia. ERV-R and ERV-L correspond to different classes of retroviruses (de Parseval and Heidmann 2005).

Relationships of *ENS* Among the Galliformes

We aligned the various DNA sequences of the chicken copies of the *ENS* ORF with the DNA sequences cloned from different Galliforme species (Acloque et al. 2001). The alignment was used to perform a phylogenetic reconstruction using the maximum-likelihood method (Fig. 5). We found that the grouping of chicken *ENS*-like copies shown in Fig. 2 is also found with good statistical support. The sequence cloned from the chicken is among them. However, the other sequences cloned from different species are found within the degenerate copies of the chicken genome. This suggests that the complete copy found in the chicken genome is part of a recent retrotransposition event of *cENS*, and this is confirmed by the LTR-based estimation of the insertion date. The position of the sequences compared to the species tree based on the cytochrome b gene (Fig. 6) reveals that *cENS* has probably been transmitted vertically for a long time and displays a pattern of duplication and loss that is typical of any multigene family. There is therefore no evidence that lateral transfer events between species have occurred, indicating that *cENS* displayed no infectivity that could be derived from its viral genes. We analyzed the cloned Galliforme sequences to estimate their potential activity. With the exception of the sequences from the quail, the helmeted guineafowl, and sequence 1 from the grey partridge, all the sequences displayed internal stop codon or frameshifts, indicating that they are no longer functional and so do not produce full-length *ENS-1* transcripts.

To find out whether the five sequences (two in the chicken and three from the quail, the helmeted guineafowl, and sequence 1 from the grey partridge) with no such disruptions are still potentially functional, we performed a phylogenetic *dN/dS* analysis. In the first analysis we determined whether the three Galliforme sequences had a *dN/dS* ratio significantly different from the rest of the sequences in the tree represented in Fig. 5. For that we considered two groups in the tree: group 1 corresponds to the group containing GG1_186 and GG5_48, and the group 2 corresponds to the one containing the other Galliforme sequences. The *dN/dS* ratio for the three Galliforme sequences, estimated to be 0.16, was significantly different

Table 2 Occurrence of solo-LTR inserted near genes

		Distance from gene					
		<20 kb		<10 kb		<5 kb	
		sense	antisense	sense	antisense	sense	antisense
Number of solo-LTR	In 5′ region	16	20	8	13	1	6
	Inside gene	18	61	18	61	18	61
	In 3′ region	8	15	6	10	4	5

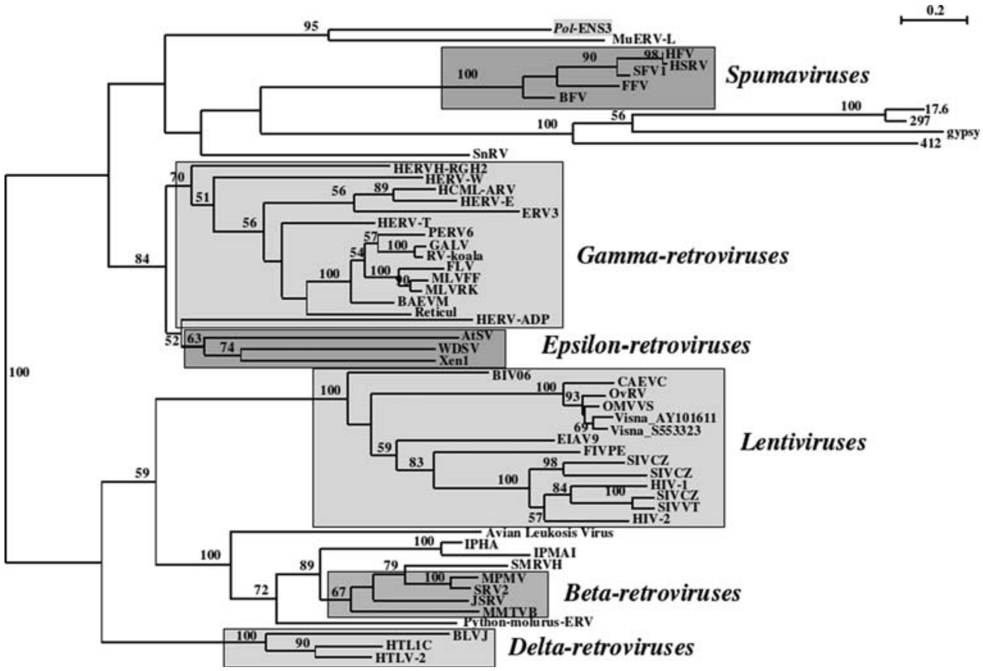

Fig. 3 Maximum-likelihood tree based on the amino acid sequences of the *pol* gene. Only bootstrap values greater than 50% are given

from the rest of the tree (*dN/dS* = 0.23 for the group 1 and *dN/dS* = 0.41 for the group 2) using a likelihood ratio test (LRT). In the case of the chicken sequences GG1_186 and GG5_48, the *dN/dS* ratio was quite low (0.13), but the LRT was not significant even if the *dN/dS* ratios of group 1 (0.46) and of group 2 (0.33) were higher. This may be explained by the fact that the sequences in group 1 are not very divergent in sequences. Moreover, the inactivated sequences are likely to have been active very recently. This analysis, however, shows that the five copies that do not display any disrupting mutations are under negative selection pressure.

Discussion

Embryonic stem (ES) cells are unique cells that appear very early during embryo development, and they soon disappear after giving rise to all the cell lineages that subsequently constitute the entire organism. Their major role in sustaining the descendants explains the current interest in genes that are expressed only during the undifferentiated state. The *cENS* gene family (Acloque et al. 2001) belongs to this group of ES-specific genes in chicken (Pain et al. 1996). Another characteristic of this gene family is its lack of homology with the sequences of genomes from species other than Galliformes, which is probably a consequence of its retroviral origin. In an attempt to clarify the mechanism of formation of this neogene and the role of retroviruses in this process, we analyzed the *cENS* sequences in the chicken genome and our findings led us to two major conclusions. The first is that selection pressure must exist to maintain some *ENS* ORFs in the chicken genome. The second concerns some aspects of the origin of this gene family and suggests that counterparts may be found in other bird species.

The Maintenance of *cENS* Genes

We found 46 copies of *ENS*-like structures, most of which were degraded, and 874 solo-LTRs within the genome. Among the 46 copies that contain internal domains, only two copies revealed a complete *ENS* ORF. One of the copies, GG1_186, was the copy initially identified using molecular methods (Acloque et al. 2001). The other copy, GG5_48, is inserted inside the intron of a gene. These two copies are not completely identical but display 98.98% identity. The *dN/dS* ratio analysis indicates that both GG5_48 and GG1_186 sequences are subject to negative selection pressure to be maintained. This suggests that of all the copies of *cENS* present in the genome, these two sequences, known as the complete sequences, are the only ones likely to promote their function, which could only be to allow the family to be maintained by transposition.

The remaining copies are all degraded sequences. The degree of degradation may be more or less pronounced. This indicates that some copies are very recent, as has been confirmed by the LTR estimation of the insertion date, whereas others are very ancient. The recent copies are grouped with the two complete copies (GG5_45 and GG1_186) in the phylogenetic tree based on the *ENS* ORF (Fig. 2), which indicates that the retrotransposable activity of the expressed copies of *ENS* ORF in ES cells must be

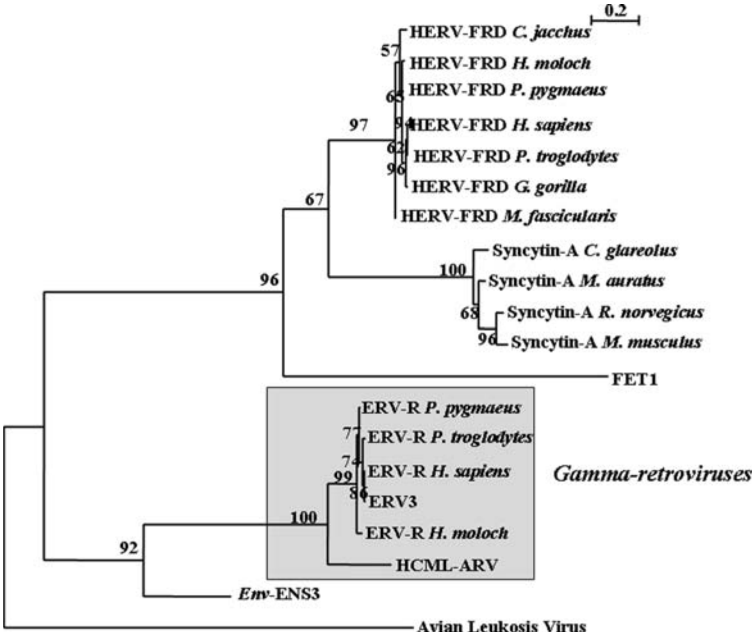

Fig. 4 Maximum-likelihood tree based on the amino acid sequences of *env* gene. Only bootstrap values greater than 50% are given

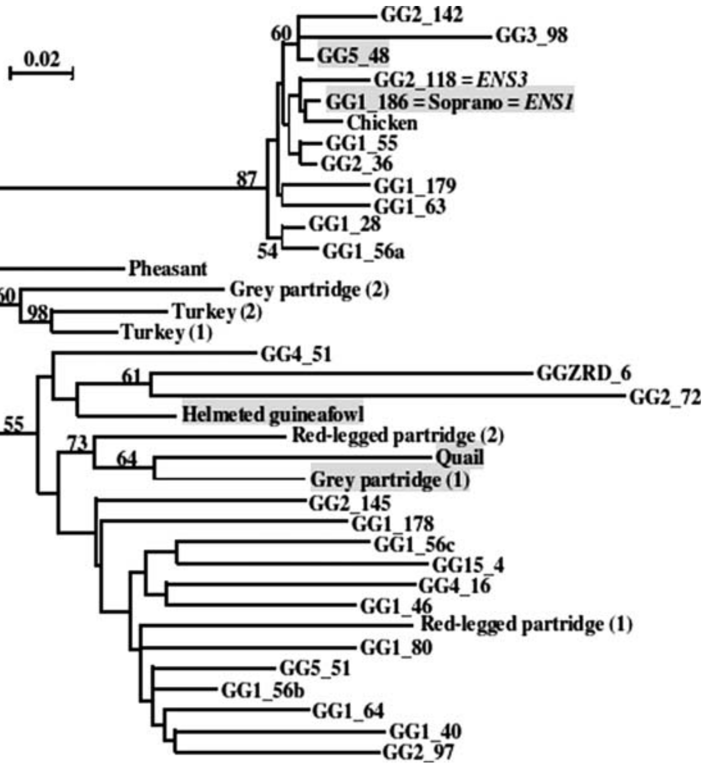

Fig. 5 Maximum-likelihood tree of the *ENS* ORF nucleic sequences. Only bootstrap values greater than 50% are given. The sequences in gray correspond to potentially active *ENS* ORFs

very recent. It is indeed important for the gene to conserve intact LTRs in order to be expressed, as assessed by functional studies (Acloque et al. 2004). LTRs are also important in the replication of all retroviruses because they have to be specifically recognized by the reverse transcriptase and by the integrase (Hindmarsh and Leis 1999; Wilhelm and Wilhelm 2001). The other *cENS* sequences in the tree represent degraded remnants that have been present in the genome for a long time, with the exception of the sequences from three Galliformes, namely, the quail, the helmeted guineafowl, and sequence 1 of the grey partridge. The *dN/dS* ratios of these sequences indicate that they are subjected to negative selection pressure. This means that complete counterparts probably remain to be identified in the species in which only degraded forms of *cENS* have so far been detected. The presence of complete *cENS* in other

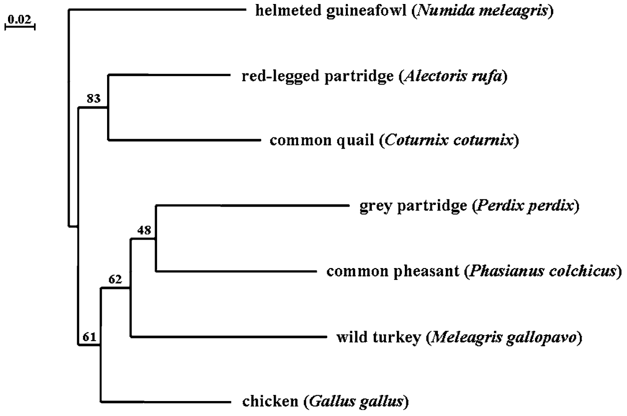
0.02
helmeted guineafowl (Numida meleagris)
red-legged partridge (Alectoris rufa)
83
common quail (Coturnix coturnix)
grey partridge (Perdix perdix)
48
common pheasant (Phasianus colchicus)
62
wild turkey (Meleagris gallopavo)
61
chicken (Gallus gallus)

Fig. 6 Maximum-likelihood tree of the Galliforme species based on the *cytb* gene. Only bootstrap values greater than 50% are given

Galliforme species would suggest that the gene could have an important function in these species. However, in human it has been shown that a particular endogenous retrovirus for which negative selection was detected on the envelope gene presented no particular function for the host (de Parseval and Heidmann 1998).

A very interesting finding is that there is only one copy of the *ENS-3* sequence presenting the complete retroviral structures of *pol* and *env*. This sequence is also grouped with the more recent insertions. As complete sequencing has not been done for this copy within which some bases remain to be identified, it is difficult to know whether the *ENS* ORF is complete; however, the 5′ LTR is sufficiently conserved to promote transcription in chicken ES cells (Acloque et al. 2001). However, the fact that no other copy of *ENS-3* has been found would seem to indicate either that it is no longer able to transpose or that a host mechanism prevents it from replicating by postintegration selection against new insertions, for example. It seems unlikely that the copy has recently been inactivated for transposition because if that were the case, we should detect other conserved copies of this kind, as in the case of retrotransposons in *D. melanogaster* (Lerat et al. 2003). It is also possible that the gaps in the sequenced genome prevent us from detecting other complete copies of *ENS-3*. The activity of the *pol* and *env* ORFs may be necessary for the transposition of the *ENS* ORF of the complete copy (GG1_186) or, more generally, for the host genome. Roles of some endogenous retroviruses for the host genome have been described in mammals. In the human genome, for example, a particular endogenous retrovirus, *ERV-W*, has been shown to possess an active *env* gene selectively constrained that seems to be involved in trophoblast differentiation (Mallet et al. 2004).

Apart from *ENS* ORFs, it is of interest to note that almost 900 solo-LTRs have been identified within the chicken genome. Some of them have been found inserted near genes. It has been shown that the 5′ LTR of *cENS* is responsible for the specific expression pattern of the *ENS* ORF (Acloque et al. 2004). Consequently, it is likely that some of the solo-LTRs detected may have retained their promoter activity and could, therefore, influence the expression of neighboring genes. There are examples in mammals of particular LTRs that are able to contribute to gene expression (van de Lagemaat et al. 2003). It is therefore possible that some genes with a *cENS* LTR in their vicinity may display either the same expression pattern as *cENS* or an alternative expression pattern under certain conditions. It is also possible that particular genes expressed in chicken ES may contain remnants of LTR that have been domesticated and are now their constitutive promoters.

These findings are in accordance with a function associated with *cENS* expression during early development in the chicken. This could happen in two ways: The function could be sustained by the protein produced by the *ENS* ORF and/or through the *cENS* solo-LTRs that may specifically control the expression of some host genes .

The Origin of *cENS*

The expression pattern of the *cENS* gene clearly indicates that it is strictly controlled during chicken development (Acloque et al. 2001, 2004; Streit et al. 2000). Its structural features suggest a retroviral origin. However, the intrinsic origin of the *ENS* ORF remains to be determined. Because no obvious homology with any other known protein has been established, the emergence of this gene remains a mystery. However, several hypotheses can be proposed.

One mechanism that allows a host sequence to be acquired by a transposable element is the transduction. The first such phenomenon to be reported involved the 3′ transduction of LINE elements (Moran et al. 1999). Recently, a primate gene has been shown to result from fusion between a host gene and a transposase gene (Cordaux et al. 2006). More interesting, a human gene, *FAM8A1*, seems to have been captured by an endogenous retrovirus during primate evolution by means of a process resembling oncogene transduction, and this was followed by multiple retrotransposition events (Jamain et al. 2001). The mechanism proposed by Jamain et al. (2001) is an illegitimate recombination between the mRNAs of the active gene and of a retrovirus during reverse transcription. The result is a mosaic mRNA, in which a portion of *FAM8A1* replaces part of the retrovirus.

Another possibility is that the *ENS* ORF results from the fusion of different protein domains from different gene products. A domain gene fusion event can be promoted by exon shuffling via the transduction of a LINE element (Moran et al. 1999). More recently, another mechanism has been proposed, that of transcription-induced chimerism, where several genes may be cotranscribed in the same pre-mRNA molecule (Akiva et al. 2006). If this is the case, it should be possible to identify the donor proteins if the domains have been conserved throughout evolution. However, to date, no obvious donors have been detected.

The last and most likely hypothesis is that the *ENS* ORF is in fact a *gag* gene that has evolved beyond recognition and possibly acquired a functional role for the host genome. The location of the *ENS* ORF in the *ENS-3* sequence just before the *pol* gene seems to strengthen this hypothesis. Moreover, the sequences of gag proteins are not well conserved among retroviruses. Only particular domains are generally found that make it possible to identify the protein, but some retroviruses such as the spuma retroviruses do not display these conserved domains such as the Major Homology Region (MHR). A possible scenario could therefore be the infection of an ancestor of modern birds by an exogenous retrovirus, which subsequently infected the germinal line to become an endogenous retrovirus. We can suppose that the *gag* gene later developed an important function for the host.

Whatever the mechanism of formation of *cENS*, the question remains about when this gene appeared. It has already been detected in several different species of Gallinaceae (Acloque et al. 2001), a group thought to have emerged between 90 and 100 MYR ago (van Tuinen and Hedges 2001). *In silico* searches in complete or ongoing sequencing genome projects have not detected *cENS* in mammals. However, a significant match has been detected in the draft sequences of the emu, *Dromaius novaeholandiae* (data not shown). The emu is a member of the Palaeognathae, a group that diverged around 120 MYR ago within the bird clade (van Tuinen and Hedges 2001), which would indicate a much more ancient origin for *cENS*. Currently, the sequenced genomes are biased because only one bird has been completely sequenced and no reptiles are yet available; thus, most vertebrate diversity is still not accessible. Further studies will therefore be necessary to look for *cENS* in other bird species and also in reptiles, which are closely related to birds.

Conclusion

Our findings have shown that the expressed copies of the *ENS* ORF are subjected to negative selection pressure in the chicken genome. They result from a rather recent insertion event by activation of the transposition of a quite ancient component of the Galliforme genomes, a group that emerged about 100 MYR ago. A similar process has probably occurred in three other Galliformes, suggesting that homologous counterparts may well remain to be discovered in other species.

Acknowledgments We thank Laurent Duret for his critical reading of the manuscript, Gabriel Marais for his help in the use of the PAML program, and Monika Ghosh for editing the English. This work has been supported by the ANR Genanimal**.**

References

Acloque H, Risson V, Birot A-M, Kunita R, Pain B, Samarut J (2001) Identification of a new gene family specifically expressed in chicken embryonic stem cells and early embryo. Mech Dev 103:79–91

Acloque H, Mey A, Birot A-M, Gruffat H, Pain B, Samarut J (2004) Transcription factor cCP2 controls gene expression in chicken embryonic stem cells. Nucleic Acids Res 32:2259–2271

Akiva P, Toporik A, Edelheit S, Peretz Y, Diver A, Shemesh R, Novik A, Sorek R (2006) Transcription-mediated gene fusion in the human genome. Genome Res 16:30–36

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

Axelsson E, Smith NGC, Sundström H, Berlin S, Ellegren H (2004) Male-biased mutation rate and divergence in autosomal, Z-linked and W-linked introns of chicken and turkey. Mol Biol Evol 21:1538–1547

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In: Altman R (ed) Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology. AAAI Press, Menlo Park, CA, pp 28–36

Borisenko L (2003) Avian endogenous retroviruses. Folia Biol (Praha) 49:177–182

Bowen NJ, McDonald JF (2001) *Drosophila* euchromatic LTR retrotransposons are much younger than the host species in which they reside. Genome Res 11:1527–1540

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552

Cordaux R, Udit S, Batzer MA, Feschotte C (2006) Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. Proc Natl Acad Sci U S A 103:8101–8106

de Parseval N, Heidmann T (1998) Physiological knockout of the envelop gene of the single-copy ERV-3 human endogenous retrovirus in a fraction of the Caucasian population. J Virol 72:3442–3445

de Parseval N, Heidmann T (2005) Human endogenous retroviruses: from infectious elements to human genes. Cytogenet Genome Res 110:318–332

Felsenstein J (2002) PHYLIP (phylogeny inference package), version 3.6. Distributed by the author, Department of Genetics, University of Washington, Seattle

Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. Comput Appl Biosci 12:543–548

Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Hindmarsh P, Leis J (1999) Retroviral DNA integration. Microbiol Mol Biol 63:836–843

Hubbard T, Andrews D, Caccamo M, Cameron G, Chen Y, Clamp M, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Gilbert J, Hammond M, Herrero J, Hotz H, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Keenan S, Kokocinsci F, London D, Longden I, McVicker G, Melsopp C, Meidl P, Potter S, Proctor G, Rae M, Rios D, Schuster M, Searle S, Severin J, Slater G, Smedley D, Smith J, Spooner W, Stabenau A, Stalker J, Storey R, Trevanion S, Ureta-Vidal A, Vogel J, White S, Woodwark C, Birney E (2005) Ensembl 2005. Nucleic Acids Res 33(Database issue):D447–D453

International Chicken Genome Sequencing Consortium (2004) Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–777

Jamain S, Girondot M, Leroy P, Clergue M, Quach H, Fellous M, Bourgeron T (2001) Transduction of the human gene FAM8A1 by endogenous retrovirus during primate evolution. Genomics 78:38–45

Jern P, Sperber GO, Blomberg J (2005) Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. Retrovirology 2:550

Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

Kimball RT, Braun EL, Zwartjes PW, Crowe TM, Ligon JD (1999) A molecular phylogeny of the pheasants and partridges suggests that these lineages are not monophyletic. Mol Phylogenet Evol 11:38–54

Lerat E, Capy P (1999) Retrotransposons and retroviruses: analysis of the envelope gene. Mol Biol Evol 16:1198–1207

Lerat E, Rizzon C, Biémont C (2003) Sequence divergence within transposable element families in the *Drosophila melanogaster* genome. Genome Res 13:1889–1896

Li W (1997) Molecular Evolution. Sinauer, Sunderland, MA

Long M (2001) Evolution of novel genes. Curr Opin Genet Dev 11:673–680

Mallet F, Bouton O, Prudhomme S, Cheynet V, Oriol G, Bonnaud B, Lucotte G, Duret L, Mandrand B (2004) The endogenous retroviral locus ERVWE1 is a bona fide gene involved in hominoid placental physiology. Proc Natl Acad Sci U S A 101:1731–1736

Maksakova IA, Mager DL (2005) Transcriptional regulation of early transposon elements, an active family of mouse long terminal repeat retrotransposon. J Virol 79:13865–13874

Moran JV, DeBerardinis RJ, Kazazian HH Jr (1999) Exon shuffling by L1 retrotransposition. Science 283:1530–1534

Notredame C, Higgins D, Heringa J (2000) T-Coffee: a novel method for multiple sequence alignments. J Mol Biol 302:205–217

Pain B, Clark ME, Shen M, Nakazawa H, Sakurai M, Samarut J, Etches RJ (1996) Long-term *in vitro* culture and characterisation of avian embryonic stem cells with multiple morphogenetic potentialities. Development 122:2339–2348

Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB (2004) Retrotransposons regulate host genes in mouse oocytes and preimplantation embryos. Dev Cell 7:597–606

Reed KJ, Sinclair AH (2002) FET-1: a novel W-linked, female specific gene up-regulated in the embryonic chicken ovary. Gene Expr Patterns 2:83–86

Sato N, Sanjuan IM, Heke M, Uchida M, Naef F, Brivanlou AH (2003) Molecular signature of human embryonic stem cells and its comparison with the mouse. Dev Biol 260:404–413

Streit A, Berliner AJ, Papanayotou C, Sirulnik A, Stern CD (2000) Initiation of neural induction by FGF signaling before gastrulation. Nature 406:74–78

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

van de Lagemaat LN, Landry J-R, Mager DL, Medstrand P (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 19:530–536

van Tuinen M, Hedges SB (2001) Calibration of avian molecular clocks. Mol Biol Evol 18:206–213

Wei CL, Miura T, Robson P, Lim SK, Xu XQ, Lee MY, Gupta S, Stanton L, Luo Y, Schmitt J, Thies S, Wang W, Khrebtukova I, Zhou D, Liu ET, Ruan YJ, Rao M, Lim B (2005) Transcriptome profiling of human and murine ESCs identifies divergent paths required to maintain the stem cell state. Stem Cells 23:166–185

Wicker T, Robertson JS, Schulze SR, Feltus FA, Magrini V, Morrison JA, Mardis ER, Wilson RK, Peterson DG, Paterson AH, Ivarie R (2005) The repetitive landscape of the chicken genome. Genome Res 15:126–136

Wilhelm M, Wilhelm F-X (2001) Reverse transcription of retroviruses and LTR retrotransposons. Cell Mol Life Sci 58:1246–1262

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555–556