Prototype endogenous avian retroviruses of the genus *Gallus*

Melanie Ann Sacco\(^1\) and Venugopal K. Nair\(^2\)

\(^1\)Center for Applied Biotechnology Studies, Department of Biological Science, California State University Fullerton, Fullerton, CA 92834-6850, USA
\(^2\)Pirbright Institute, Compton Laboratory, Newbury, Berkshire RG20 7NN, UK

Ancient endogenous retroviruses (ERVs), designated endogenous avian retrovirus (EAVs), are present in all *Gallus* spp. including the chicken, and resemble the modern avian sarcoma and leukemia viruses (ASLVs). The EAVs comprise several distinct retroviruses, including EAV-0, EAV-E51 and EAV-HP, as well as a putative member previously named the avian retrotransposon of chickens (ART-CH). Thus far, only the EAV-HP elements have been well characterized. Here, we determined sequences of representative EAV-0 and EAV-E51 proviruses by cloning and data mining of the 2011 assembly of the *Gallus gallus* genome. Although the EAV-0 elements are primarily deleted in the *env* region, we identified two complete EAV-0 *env* genes within the *G. gallus* genome and prototype elements sharing identity with an EAV-E51-related clone previously designated EAV-E33. Prototype EAV-0, EAV-E51 and EAV-E33 *gag*, *pol* and *env* gene sequences used for phylogenetic analysis of deduced proteins showed that the EAVs formed three distinct clades, with EAV-0 sharing the last common ancestor with the ASLVs. The EAV-E51 clade showed the greatest level of divergence compared with other EAVs or ASLVs, suggesting that these ERVs represented exogenous retroviruses that evolved and integrated into the germline over a long period of time. Moreover, the degree of divergence between the chicken and red jungle fowl EAV-E51 sequences suggested that they were more ancient than the other EAVs and may have diverged through mutations that accumulated post-integration. Finally, we showed that the ART-CH elements were chimeric defective ERVs comprising portions of EAV-E51 and EAV-HP rather than authentic retrotransposons.

INTRODUCTION

Endogenous retroviruses (ERVs) that exist as stable genetic elements have been shown to be distributed widely in all vertebrate species examined, including the class *Aves* (Bénit et al., 1999; Gifford et al., 2005; Herniou et al., 1998; Martin et al., 1999). The genome structures of the avian ERVs determined to date primarily resemble avian exogenous C-type retroviruses of the avian sarcoma and leukemia viruses (ASLVs) in the genus *Alpharetrovirus* (Boyce-Jacino et al., 1992; Coffin et al., 1983; Dunwiddie et al., 1986; Hughes et al., 1981; Ruis et al., 1999; Sacco et al., 2000). The first chicken ERVs identified were the subgroup E avian leukemia viruses (ALV-E), previously termed *ev* loci, which exist at various segregating loci in the chicken (*Gallus gallus* subsp. *domesticus*) genome as defective elements or intact infectious proviruses (reviewed by Crittenden, 1991). These appear to be the most recently integrated of the ERVs based on their restricted distribution in the species *G. gallus* (Frisby et al., 1979), which includes the chicken and its undomesticated progenitor, the red jungle fowl (*G. gallus* subsp. *gallus*), for which the genome sequence has been obtained (Hillier et al., 2004). Other chicken ERVs, designated endogenous avian retroviruses (EAVs), were described subsequently (Boyce-Jacino et al., 1989; Dunwiddie & Faras, 1985; Dunwiddie et al., 1986). EAVs appear to be more ancient than the ALV-E elements, as suggested by their distribution in all species of the genus *Gallus* (Boyce-Jacino et al., 1989; Resnick et al., 1990; Sacco et al., 2000) and the presence of shared loci between different *Gallus* spp., because germline integrations are rare and identical proviral insertions likely indicate integration occurred before divergence of the two species (Resnick et al., 1990; Sacco et al., 2001).

The EAVs comprise several distinct retrovirus elements designated EAV-0, EAV-E51 and EAV-HP. EAV-0 was first identified following hybridization with Rous sarcoma virus (RSV) probes to DNA from line 0 white leghorn chickens bred to be free of ALV-E loci (Dunwiddie & Faras, 1985).
The EAV-E51 sequences were identified subsequently in line 0 genomic DNA (Boyce-Jacino et al., 1992). Only partial provirus elements were sequenced in initial published reports; however, both EAVs appeared to be defective, primarily as a consequence of deletions in their genomes of various sizes (Boyce-Jacino et al., 1992). The EAV-HP proviruses (also designated cv/J) have been examined in the most detail due to their similarity with and likely contribution of env sequences to the exogenous ALV subgroup J (ALV-J) by recombination (Bai et al., 1995; Ruis et al., 1999). Whilst most of the EAV-HP provirus sequences have large deletions, including the entire pol region, intact and expressed proviral sequences have been identified that segregate within the chicken population (Sacco et al., 2004). A final member ascribed to the EAV family is an endogenous retroelement named the avian retrotransposon of chickens (ART-CH) due to the absence of env sequences (Gudkov et al., 1992). ART-CH was found subsequently to share a high degree of identity with EAV-HP and EAV-E51, suggesting it is a defective EAV chimeric retrovirus with the env sequences lost, rather than a bona fide retrotransposon; however, a portion of the ART-CH genome did not match any known ERV sequences (Sacco et al., 2000).

Further interest in avian ERVs also developed from the concerns for human health arising from detection of reverse transcriptase (RT) activity in ALV-free chicken embryos or embryo fibroblast cells used as substrates for human vaccine production (Robertson et al., 1997; Weissmahr et al., 1997). Whilst the EAV-HP elements in line 0 chickens appear to be completely devoid of pol gene sequences (Sacco et al., 2004), the EAV-0 elements appear to have a complete pol region and the structures of EAV-E51 elements are unknown, suggesting that these elements may be the source of RT activity (Boyce-Jacino et al., 1992). The study presented here was aimed at obtaining full-length representative sequences to define the structures of the EAV-0 and EAV-E51 elements. We also aimed to determine the relationships among the different EAV family members and recombinant proviruses, such as ART-CH elements, and to ascertain the ability of these elements to produce infectious virus.

**RESULTS**

**Characterization of prototype EAV-0 elements and retrieval of novel env sequences**

We initiated a study to determine prototype sequences of the EAV-0- and EAV-E51-type ERVs from the chicken initially using molecular biology approaches, and continued this endeavour by capitalizing on the release of the *G. gallus* genome sequence (Hillier et al., 2004). PCR amplification of EAV-0 proviral DNA using long terminal repeat (LTR)-specific primers produced two products of ~2.8 and 5.8 kbp (data not shown). The 5.8 kbp PCR product was the expected size based on previously described EAV-0 element transcripts amplified by PCR (Weissmahr et al., 1997) or mapped by restriction enzyme digestion and hybridization (Resnick et al., 1990). A representative clone (pEAV5) was found to be 5811 bp in length, including the virus-derived sequences in the oligonucleotide primers with NotI restriction sites, and showed 98 % identity to the corresponding regions in previously reported partial sequences. The EAV-0 provirus had a large deletion of the env gene, as previously described for these ERVs, and was grossly intact for the gag and pol genes, although clone pEAV5 had a single-base-pair deletion resulting in a −1 frameshift within the protease gene that would render it defective in those genes. A BLAST search using the deduced amino acid sequence of the EAV-0 pol gene demonstrated 66 % identity and 80 % similarity to the translated ASLV pol genes, and 66 % identity and 79 % similarity to the translated EAV-HP pol gene. The less-conserved gag genes showed 45–47 % identity and 55–57 % similarity between the deduced amino acid sequences of the EAV-0 prototype and the ASLVs; the same region demonstrated only 37 % identity and 48 % similarity to EAV-HP.

Using the full-length EAV-0 sequence, a search of the red jungle fowl genome database produced two hits with additional env coding sequences (Table 1), both from contigs that had not yet been mapped and for which only partial provirus sequences were available at the time of submission of this manuscript. These sequences were designated EAV-0Un1 and EAV-0Un2 (EAV-0 unmapped sequences 1 and 2, respectively). The sequences of EAV-0Un1 and EAV-0Un2 spanned from the integrate-coding region of the pol gene corresponding to nt 5061 and 4652 of the pEAV5 clone, respectively. Alignment of the EAV-0 clone pEAV5 with EAV-0Un1 allowed the size of the EAV-0 env gene deletion in pEAV5 to be determined as 1189 bp. The deduced translated env sequences demonstrated an intact env gene in EAV-0Un1 and a premature stop codon in the transmembrane (TM) protein sequence of the partial EAV-0Un2. Comparison of this EAV-0 deduced amino acid sequence with other avian retroviral sequences demonstrated divergence within the envelope protein peptide leader sequence (Fig. S1, available in the online Supplementary Material). When the more-conserved envelope surface (SU) and TM glycoprotein-coding regions were examined, EAV-0 shared 43–44 % identity and 57 % similarity with the ALV-J and EAV-HP sequences, and had the closest similarity (46 % identity and 73 % similarity) with the previously described EAV-E51 endogenous element.

**Intact EAV-E51 elements exist at few segregating genomic loci**

PCR amplification using LTR-specific primers for EAV-E51 was performed to obtain a full-length proviral clone as described for the EAV-0 prototype provirus above; however, repeated attempts amplified products of only ~3 kbp in size (data not shown). A clone of the PCR products (pEAV27) was sequenced and the small size (2959 bp) of the provirus was found to be mainly due to a large deletion of the pol...
### Table 1. EAV endogenous retrovirus loci from the red jungle fowl genome used in this study

<table>
<thead>
<tr>
<th>EAV species and locus name</th>
<th>NCBI reference sequence</th>
<th>Contig location</th>
<th>LTR identity (%)</th>
<th>Provirus sequence size (bp)</th>
<th>Retrovirus structure spanned*</th>
<th>Flanking direct repeat</th>
</tr>
</thead>
<tbody>
<tr>
<td>EAV-0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EAV-0Chr1</td>
<td>NW_003763650.1</td>
<td>21 841 561–21 846 150</td>
<td>95</td>
<td>4590</td>
<td>5’-LTR–Δgag–Δpol–Δenv–LTR-3’</td>
<td>GCTGAC</td>
</tr>
<tr>
<td>EAV-0Chr2</td>
<td>NW_001471639.2</td>
<td>8 095 311–8 101 488</td>
<td>100</td>
<td>6178</td>
<td>5’-LTR–Δgag–Δpol–Δenv–LTR-3’</td>
<td>none</td>
</tr>
<tr>
<td>EAV-0Chr2A</td>
<td>NW_001471639.2</td>
<td>21 006 029–21 012 227</td>
<td>100</td>
<td>6200</td>
<td>5’-LTR–Δgag–Δpol–Δenv–LTR-3’</td>
<td>CCTTT</td>
</tr>
<tr>
<td>EAV-0Chr2B</td>
<td>NW_003763661.1</td>
<td>6 587 549–6 593 731</td>
<td>100</td>
<td>6183</td>
<td>5’-LTR–Δgag–Δpol–Δenv–LTR-3’</td>
<td>TAAAC</td>
</tr>
<tr>
<td>EAV-0Un1</td>
<td>NW_003779871.1</td>
<td>1–2161†</td>
<td>ND</td>
<td>2161</td>
<td>5’-env-LTR-3’</td>
<td>ND</td>
</tr>
<tr>
<td>EAV-0Un2</td>
<td>NW_003778217.1</td>
<td>1–1790§</td>
<td>ND</td>
<td>1790</td>
<td>5’-env-LTR-3’</td>
<td>ND</td>
</tr>
<tr>
<td>EAV-E51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EAV-E51Chr2</td>
<td>NW_001471637.2</td>
<td>3 060 739–3 068 550</td>
<td>97</td>
<td>7812</td>
<td>5’-LTR–gag–pol–env–LTR-3’</td>
<td>C(T)(A)/GA</td>
</tr>
<tr>
<td>EAV-E51Chr28</td>
<td>NW_003764302.1</td>
<td>160 868–168 457</td>
<td>96</td>
<td>7590</td>
<td>5’-LTR–gag–pol–env–LTR-3’</td>
<td>TGTAC</td>
</tr>
<tr>
<td>EAV-E51ChrZ</td>
<td>NW_003764319.1</td>
<td>4 190 360–4 183 955</td>
<td>99</td>
<td>6406</td>
<td>5’-LTR–gag–pol–env–LTR-3’</td>
<td>none</td>
</tr>
<tr>
<td>EAV-E51ChrZB</td>
<td>NC_006127.3</td>
<td>49 054 924–49 061 458</td>
<td>98</td>
<td>6534</td>
<td>5’-LTR–gag–pol–env–LTR-3’</td>
<td>GTTGT</td>
</tr>
<tr>
<td>EAV-E51Un</td>
<td>NW_003773815.1</td>
<td>2747–9716∥</td>
<td>ND</td>
<td>6969</td>
<td>5’-LTR–gag–pol–env–ΔLTR-3’</td>
<td>ND</td>
</tr>
<tr>
<td>EAV-E33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EAV-E33Chr1</td>
<td>NW_003763647.1</td>
<td>859 363–867 090</td>
<td>98</td>
<td>7728</td>
<td>5’-LTR–gag–pol–env–LTR-3’</td>
<td>C(T/G)TGA</td>
</tr>
<tr>
<td>EAV-E33Chr4</td>
<td>NW_003763739.1</td>
<td>3 009 329–3 013 895</td>
<td>ND</td>
<td>4567</td>
<td>5’-pol–env-LTR-3’</td>
<td>ND</td>
</tr>
<tr>
<td>EAV-E33Chr6</td>
<td>NW_003763812.1</td>
<td>10 867 263–10 861 262</td>
<td>ND</td>
<td>6002</td>
<td>5’-Δgag–pol–env–LTR-3’</td>
<td>ND</td>
</tr>
</tbody>
</table>

ND, LTR and flanking sequence analysis not determined due to lack of sequence for one side.

*Retroviral gene sequence present in locus with a partial deletion is indicated by Δ.

†Beginning of contig in env gene.

‡Contig location in the full contig length for the accession.

∥Right LTR incomplete due to poor-quality sequence before position 2747 of contig.

¶Sequence size was determined after subtracting a region of duplicated and uncalled sequence in the contig.
gene, producing a gag–env fusion. Additional smaller deletions were present in the remaining provirus sequence, including several single-base-pair deletions disrupting the reading frame of the grossly intact gag gene that rendered it non-functional.

As a first step to obtain more complete EAV-E51 sequence data, an E51 probe obtained by PCR amplification of broiler chicken line 21 DNA was used to screen by hybridization a lambda genomic DNA library of the same chicken line. A positive clone (pEAV-E51) was isolated and subcloned for sequencing, but was found to be only a partial clone, comprising 4625 bp of the 3′-end sequence. The pEAV-E51 sequence was used to search the G. gallus genome database to obtain the complete provirus sequence for this locus, which was found on the Z chromosome and designated EAV-E51LZ (EAV-E51-like provirus from red jungle fowl genome chromosome Z; Table 1). Comparison of the original EAV-E51 sequence, our pEAV-E51 clone from line 21 and the red jungle fowl EAV-E51LZ sequence indicated that they were 98–99 % identical to each other with identical 3′ integration junctions.

Hybridizations of chicken genomic DNA with probes for the EAV-E51 env gene have shown previously that these ERVs exist as moderately repeating elements within the chicken genome (Boyce-Jacino et al., 1992). The detection of only the truncated 3 kbp EAV-E51 proviruses by LTR-primed PCR suggested that these deleted structures may be the predominant form of these elements within the chicken genome, although preferential amplification of smaller amplicons could prevent amplification of more intact proviral templates. In order to examine the distribution of EAV-E51 proviruses with intact pol genes, a 275 bp RT region, which was deleted from the 3 kbp structures, was hybridized to EcoRI-digested G. gallus spp. genomic DNA using high-stringency conditions. Only two hybridizing fragments from line 0 and brown leghorn chickens were detected, whilst the broiler chicken line 21 and two jungle fowl species showed additional hybridizing bands (Fig. 1), indicating that at least some of these EAV-E51 loci segregate within the chicken population. The red jungle fowl fragments were poorly resolved; however, a search of the red jungle fowl genome with the 275 bp probe sequence identified 11 potentially hybridizing proviruses from this subspecies.

Fig. 1. EAV-E51 elements segregate in G. gallus. The distribution of EAV-E51 RT sequences in the G. gallus genome were observed by Southern blot hybridization performed on EcoRI-digested genomic DNA from three chicken lines [line 0 (L0), line 21 (L21) and brown leghorn (BRL)] and two jungle fowl species [red jungle fowl (RJF) and G. sonneratii (Sonnerat’s jungle fowl; SJF)] using a 275 bp EAV-E51-specific 32P-labelled RT probe under high-stringency conditions. Molecular mass size markers are indicated. Segregating proviruses are apparent within the chicken genomes. Although red jungle fowl bands were poorly resolved, the G. gallus genome database was used to count 11 potentially hybridizing proviruses from this subspecies.

was also highly defective, with multiple insertions/deletions throughout the gag coding region causing frameshifts in the deduced translated protein. A search of the G. gallus database also localized the pEAV21 ERV sequence to the Z chromosome, with a 100 % identical match that was designated EAV-E51LChrZB (Table 1).

Phylogenetic analysis of EAV family member relationships

In order to examine the relationships among the EAVs and ASLVs, sequences of the novel EAV-0- and EAV-E51-like clones were used to obtain additional G. gallus proviral sequences with complete pol and/or env genes for alignment of their deduced translated sequences; some of these more complete proviruses also contained gag genes that could be included for analysis (Table 1). An overall intact provirus on chromosome 1 most closely resembled the previously described clone designated EAV-E33 (Boyce-Jacino et al., 1992) and represented the full-length prototype sequence for EAV-E33, designated EAV-E33L1.

Comparison of the deduced amino acid sequences of the EAV-E51 gag genes demonstrated a region of divergence
encompassing the C-terminal portion of the matrix or membrane-associated (MA) protein (p19) to the start of the capsid (CA) protein, which was shown previously to be dispensable for ASLV budding and infectivity in vitro (Nelle & Wills, 1996). Overall there was little similarity in deduced sequences in these regions and the size of peptides between MA and CA varied, making it difficult to predict borders of protease cleavage products to determine if the polyprotein structure was the same as delineated for the ASLV Gag. However, several important elements required for retroviral budding were apparent, including the first 87 aa of the MA (the conserved N-terminal M domain), and the presence of the virus late (L) domain motif PPPY found in the ASLV p2 and common to most retroviruses except the lentiviruses (Bowles et al., 1994; Pepinsky et al., 1996) (Fig. S2). Similarly, EAV sequences also demonstrated the presence of other polyprotein signature motifs, including the conserved three-residue TDW sequence seen in the ASLV p10, two identically spaced zinc-finger motifs (CX2CX4HX4C) in the nucleocapsid (NC; p12) and the protease (PR; p15) DSG active-site motif (Fig. S2) (Arad et al., 1995; Méric & Spahr, 1986).

As a result of the variability in length of the Gag polyprotein sequences, the high degree of divergence between the membrane-binding domain and CA, and the presence of point deletions/insertions causing frameshifts in many of the proviruses, the deduced CA sequence alone was used to generate an alignment for phylogenetic analysis of the EAV and ASLV Gag proteins (Fig. S3), particularly as CA has strongly conserved boundaries for recognition by the viral protease during processing of the polyprotein and contains the major homology region that is conserved in nearly all avian and mammalian retroviruses (Mammano et al., 1994). The CA of a betaretrovirus, mouse mammary tumour virus (MMTV), was used as the phylogenetically distant outgroup. The three different EAV-type elements formed distinct clades with strong bootstrap support, with EAV-E33 and EAV-E51 placed as sister taxa (Fig. 2a). Similar to the ALV clade, branch lengths were very short within the EAV-HP and EAV-0 clades, indicating that these clade members were very closely related. The EAV-0 sequences were also observed to share the last common ancestor with the modern ALVs. When considered together with the high degree of LTR sequence identity within individual EAV-0 proviruses, the clustering of the EAV-0 and ALV clades might hint at the EAV-0 proviruses being the most recently integrated EAVs.

Alignment of the RT sequences of the EAVs and representative ALVs was also performed (Fig. S4) to generate a phylogenetic tree of these viruses (Fig. 2b). Two additional EAV-0 proviruses in the red jungle fowl genome with deletions in the gag gene as well as two additional EAV-E51-like elements that showed more sequence similarity to EAV-E33 (E33L4 and E33L6 for EAV-E33-like chromosome 4 and 6, respectively) were included in the analysis, and the lymphoproliferative disease virus (LPDV) of turkeys, a phylogenetically distant alpharetrovirus, was used as the outgroup. The neighbour-joining tree based on the sequence tree showed congruence with the CA tree. The inclusion of additional EAV-E51-related RT sequences demonstrated that the EAV-E51-related proviruses formed two distinct subclades of EAV-E51- or EAV-E33-like sequences. The branch lengths of the EAV-E51 clade indicated a greater degree of divergence than within the EAV-0, EAV-HP or ALV clades. This degree of divergence may reflect the antiquity of these sequences, which is further supported by the divergence between the chicken and red jungle fowl Z chromosome EAV-E51 locus, the increased LTR sequence identity (Table 1) and the accumulation of one or multiple stop codons in the pol gene within this clade.

A final neighbour-joining tree was constructed using an alignment of the complete deduced sequence of the SU proteins (Fig. S5). The relationships among the ALV and EAV clades based on this SU/TM phylogenetic tree (Fig. 2c) were again congruent with the data obtained using the gag- or pol-encoded sequences. The EAV-HP clade included the recently emerged ALV-J envelope sequence, which likely arose in meat-type chicken flocks by recombination between an unknown ALV and EAV-HP genomic RNA transcribed from a locus identified previously in line 151 white leghorn chickens (Sacco et al., 2004). Members of the EAV-0 and EAV-HP clades again showed short branch lengths, indicating their close relationships, whilst the EAV-E51 clade had branch lengths similar to or longer than those in the ALV clade. The divergent sequences in the ASLV env were focused primarily in the variable and hypervariable regions defined in the SU protein that determined receptor binding and consequently the virus subgroup (Fig. S1). Conversely, the EAV-E51 residue substitutions appeared to be random in SU and TM, most notably for the E33L4 sequence that appeared to be the most divergent of the clade; the sequences included one or two nonsense mutations in several of the proviruses, as well as different frameshifts in both the chicken and red jungle fowl EAV-E51 proviruses and in the EAV-E33 provirus. Taken together, these data suggested the EAV-E51 proviruses may be more ancient than the EAV-0 and EAV-HP proviruses.
Chimeric EAVs, including ART-CH, have resulted from recombination with a deleted EAV-E51

Previous reports on EAVs have indicated that recombination events have resulted in multiple ERV elements that share regions of identity (Sacco et al., 2000). PCR amplification of EAV-0 sequences produced an ~2.8 kbp product (pEAV3), which was found upon pair-wise sequence comparison with the EAV-0 prototype and the 3 kbp EAV-E51 clone (pEAV27) to represent a chimeric provirus: the 550 bp 5' end derived from EAV-0 and the remainder of the sequence derived from the EAV-E51 provirus.

The full-length published sequence of the ERV designated ART-CH is ~3 kbp long and appears to be a chimeric provirus (Gudkov et al., 1992; Sacco et al., 2000). Previous sequence comparisons indicated that the ART-CH 5' region was derived from EAV-HP, including the U5 up to nt 793 in the gag gene, whilst the 3' sequences corresponding to the partial env gene to the U3 region were derived from EAV-E51; the remainder of the gag region had no close nucleotide identity with any published sequence (Sacco et al., 2000). When we compared the defective EAV-E51 provirus of clone pEAV27 and ART-CH, the sequences were found to have 90% sequence identity in the region of the provirus 3' to the putative recombination point shown in Fig. 3(a). Moreover, the ART-CH provirus structure was found to have the same pol deletions as this 3 kbp EAV-E51 clone, suggesting ART-CH arose by recombination between an already defective EAV-E51 and an EAV-HP. Interestingly, the recombination points for the EAV-0/E51 chimera and the EAV-HP/E51 chimera differed, but were in close proximity within a highly conserved nucleotide block (Fig. 3a). Phylogenetic analysis further illustrated the relationship of the gag gene sequences deriving from EAV-HP 5' to the recombination site (Fig. 3b) and deriving from EAV-E51 on the 3' side of the recombination point (Fig. 3c).

Using the full-length ART-CH sequence to search the G. gallus genome database, nine complete matches were obtained (Table 2), although many more elements with further deletions or that matched the 3 kbp EAV-E51 sequence were also observed. These nine G. gallus EAV-HP/E51 proviruses shared 98–99% sequence identity to each other and at least 97% overall identity to the sequence of the chicken ART-CH clone. A comparison of the left and right LTR sequences of the individual proviruses revealed that seven loci had 100% identical LTRs, suggesting recent integration activity of these elements.

**DISCUSSION**

The presence of ERVs in the genome has been postulated to have some importance for host evolution, accounting for the retention of the loci within the genome (Sverdlov, 2000). This is exemplified by studies showing the involvement of human ERV (HERV) loci in genomic rearrangements (Hughes, 2001) and the adaptation of a HERV Env glycoprotein to placental development (Mi et al., 2000).

**Fig. 3.** Multiple chimaeric EAV proviruses have likely arisen through recombination in a conserved region. (a) Sequence alignment of representative sequences demonstrating the recombination point between EAV-HP and EAV-E51 giving rise to the recombinant ART-CH provirus, and the recombination point between EAV-0 and EAV-E51 giving rise to the recombinant provirus represented by clone pEAV3. Sequences identical between the parent genomes where strand transfers likely occurred during replication are indicated by a line above or below the alignment. The deduced amino acid sequence is shown below the alignment for the block of nucleotide sequence that is highly conserved among all EAVs that includes both recombination points. (b, c) Maximum-likelihood analysis of EAV sequences 5' (b) or 3' (c) to the putative recombination point of chimeric proviruses. Bar, 0.1 expected nucleotide replacements per site. Bootstrap values are shown as per cent occurrence in 500 random samplings.
Recently, an EAV-HP provirus was shown to determine blue eggshell colour phenotype through overexpression of the SLC01B3 gene activated by the proviral LTR (Wang et al., 2013; Wragg et al., 2013). However, endogenous ALV (ev loci) activity has been shown to correlate with a reduction in poultry productivity traits, including fecundation, hatchability and feed conversion rates (Bacon et al., 2000). Similarly, HERVs have been implicated in human disease (reviewed by Ryan, 2004). The maintenance of ERVs, including the EAVs, in most cases likely demonstrates a lack of selection pressure to lose elements that have become non-infectious as a result of coding defects or transcriptional silencing by host methylation (Groudine et al., 1981). Previous examination of the G. gallus genome suggested that purifying selection has operated to eliminate ERVs that disrupt gene function as both ASLVs and HIV favour integration into active transcription units (Barr et al., 2005). In our analysis of the G. gallus genome, no EAV proviruses that represent complete and intact retroviruses capable of producing infectious virus were identified. This suggests that similar purifying selection has likely operated on the EAVs to eliminate viruses capable of producing infectious particles that would reduce bird fitness. In addition to the numerous frameshift and stop codon mutations of the ancient EAVs, most EAV-E51-related viruses had large deletions within the pol gene, whilst only two EAV-0 proviral env sequences could be retrieved.

Phylogenetic analysis of the deduced EAV Env proteins demonstrated that the EAV-0 and EAV-HP sequences form clades with short branch lengths to a common ancestor compared with the ASLV Envs with different host range determinants (subgroups A, B, C and E), suggesting each clade comprises members with specificity of a single subgroup. However, the E51-related proviruses appear to radiate from a common ancestor to a similar or greater degree than the modern ASLV clade. These proviruses may represent different ancient EAV-E51 subgroups with different host range determinants as seen for the ASLVs. Alternatively, the EAV-E51 proviruses may have resulted from repeated germline integration of an evolving quasispecies with further divergence accumulating over time by random mutation since integration. This latter possibility is more likely for at least some of the members of this clade as most of the divergent residues appear to be randomly distributed through SU and TM sequences rather than concentrated in regions that would correspond to hypervariable or variable regions defined for SU that are under host selection pressure (Bai et al., 1995; Bova et al., 1988; Dorner et al., 1985; Rainey et al., 2003). The phylogenetic analysis may also provide a clue to the chronology of EAV integration into the Gallus genome, as the clustering of the EAV-0 and ASLV clades suggests that the EAV-0 proviruses integrated most recently of the EAVs as they are more closely related to the modern alpharetroviruses.

The spread of defective ERVs has been postulated to occur during exogenous retrovirus infections, whereby exogenous virus proteins mediate packaging of the defective ERV genomes, and their subsequent reverse transcription and integration at another genomic site. The apparent emergence of the ALV-J subgroup by recombination between an EAV-HP and an exogenous ALV provided support for this idea, showing recent activity of an EAV and its mobilization by exogenous virus-mediated packaging (Sacco et al., 2004). The data presented in this paper provide support for mobilization of defective EAVs by exogenous retrovirus infection. As retroviral replication generates identical LTRs that diverge in ERVs over the course of time through the accumulation of random mutations, the degree of divergence correlates with the time since integration (Cantrell et al., 2005). EAV-0 elements examined on chromosome 2 possessed identical LTRs, suggesting recent integration of these EAVs; however, the chromosome 1 locus (EAV-0Chr1) LTR sequences showed divergence of

Table 2. EAV-HP/E51 (ART-CH) from the red jungle fowl genome identified in this study

<table>
<thead>
<tr>
<th>Chromosome</th>
<th>NCBI reference sequence</th>
<th>Contig location</th>
<th>LTR identity (%)</th>
<th>Provirus sequence size (bp)</th>
<th>Flanking direct repeat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NW_001471551.2</td>
<td>3 924 891–3 928 244</td>
<td>99</td>
<td>3354</td>
<td>ND</td>
</tr>
<tr>
<td>1</td>
<td>NW_003763650.1</td>
<td>2 803 238–2 806 585</td>
<td>100</td>
<td>3348</td>
<td>CTCCT</td>
</tr>
<tr>
<td>1</td>
<td>NW_003763461.1</td>
<td>503 544–506 922</td>
<td>100</td>
<td>3379*</td>
<td>CACTCT</td>
</tr>
<tr>
<td>2</td>
<td>NW_003763690.1</td>
<td>348 872–352 219</td>
<td>100</td>
<td>3350</td>
<td>GGCTG</td>
</tr>
<tr>
<td>2</td>
<td>NW_003763661.1</td>
<td>20 918 248–20 921 603</td>
<td>99</td>
<td>3356</td>
<td>GGGCCAG</td>
</tr>
<tr>
<td>3</td>
<td>NW_003763720.1</td>
<td>28 501 302–28 504 648</td>
<td>100</td>
<td>3347</td>
<td>CAGAC</td>
</tr>
<tr>
<td>3</td>
<td>NW_001471673.2</td>
<td>2 099 517–2 102 903</td>
<td>100</td>
<td>3387*</td>
<td>GTTTC</td>
</tr>
<tr>
<td>3</td>
<td>NW_001471668.2</td>
<td>42 668 903–42 672 289</td>
<td>100</td>
<td>3387*</td>
<td>AGGGAC</td>
</tr>
<tr>
<td>5</td>
<td>NW_003763748.1</td>
<td>4 216 429–4 219 813</td>
<td>100</td>
<td>3385*</td>
<td>CTGA</td>
</tr>
</tbody>
</table>

*Sequence includes a 30 bp duplication in the env gene portion of the provirus.
~5%. Using the autosomal mutation rate of $3.6 \times 10^{-9}$ substitutions per site per year as calculated by Axelsson et al. (2004), the interval between integration of the chromosome 1 and chromosome 2 elements is >4 million years. This interval is consistent with the observed distribution of EAVs in the genus *Gallus*, indicative of integration before separation of species from a common ancestor (Resnick et al., 1990), and the recent estimate that *G. gallus* and *Gallus varius*, the green jungle fowl, diverged from a common ancestor 3.6 million years ago (Sawai et al., 2010). The presence of nine chimeric EAV-HP/E51 proviruses in the red jungle fowl genome sharing 99% overall sequence identity suggests that these elements may still be actively mobilized by infection with modern alpharetroviruses. We have shown previously that EAV-HP elements segregate within the chicken population (Sacco et al., 2001); the recent activity of the EAVs and the different EAV-E51 RT hybridization patterns observed for different chicken lines reported here support that other EAV elements segregate within the chicken population as well.

Previously, low levels of RT activity have been found associated with virus-like particles (VLPs) in live attenuated virus vaccines derived from chicken cell substrates or from line 0 ALV-E-free chicken cells through the use of a sensitive PCR-based method for detection of the reverse-transcribed product of a heterologous RNA substrate (Böni et al., 1996; Robertson & Minor, 1996). Detection of RT activity in chicken cells raised concerns for human safety; however, this RT activity did not appear to be associated with any replication-competent retrovirus (Khan et al., 1998). EAV-0 transcripts were found to be associated with VLPs obtained from live attenuated vaccines produced from chicken cell substrates (Weissmahr et al., 1997) and EAV-HP transcripts were detected in line 0 chicken VLPs (Sacco, 2001). The presence of EAV transcripts in VLPs suggests that ERVs capable of producing functional RT and Gag protein enable packaging of such transcripts; our findings in this paper suggest that only EAV-0 elements could provide the RT activity as EAV-E51-like elements appear to be highly defective and EAV-HP elements with intact pol genes were absent from line 0 (Sacco et al., 2001). A recent study of ERVs in the *G. gallus* genome determined that the ERVs present are outnumbered by retroviruses of the beta and gamma classes, and of an ancestral intermediate alphbeta-like class (Bolissetty et al., 2012), so ERVs that are yet to be studied could contribute to this RT activity.

Our analysis of the distribution of EAV elements in the *G. gallus* genome has not been exhaustive as our purpose was to define prototype sequences for complete EAV proviruses or, where that was not possible, for complete genes of the various elements. Taken together with previous reports on EAV proviruses (Sacco et al., 2004), EAV proviruses that can produce functional Gag–Pro–Pol proteins are rare. The existence of an intact EAV-0 env gene devoid of mutations that would terminate Env translation prematurely may provide a reservoir for emergence of a new ALV subgroup in the future through recombination, in the same way that ALV-J is proposed to have emerged through recombination with an EAV-HP (Smith et al., 1999). It could be of interest to determine whether EAV-0 proviruses carrying the intact env sequences are present within breeder flocks in order to remove this potential threat.

**METHODS**

**Cloning of EAV-E51.** To create an E51 hybridization probe, line 0 genomic DNA was amplified by PCR with primers E51For (5'-TCTTGGAGGTCCATGTAC-3') and E51Rev (5'-CCTACATACCTGTCG-3') designed from the published E51 clone env and host flanking sequences, respectively. PCR products were cloned into the pGEM-T cloning vector (Promega), and the E51-specific DNA probe was radio labelled and hybridized to a line 21 chicken genomic DNA lambda library as described previously (Sacco, 2001). Filters were washed once with 2 x SSC (30 mM, sodium citrate pH 7.0, 300 mM NaCl) and three times with 0.1% (w/v) SDS for 15 min at 55°C before exposure to Kodak BioMax MR film overnight at ~70°C with intensifying screens. Positive plaques were rescreened and an ~6 kb EcoRI fragment hybridizing to the E51 probe from a positive lambda library clone was subcloned into vector pGEM-3Z (Promega) and sequenced by MWG Biotech.

**PCR amplification of EAV sequences.** For primary reactions, PCR was performed in 50 µl reaction mixtures on 100 ng genomic DNA template using 2.5 U Herculase Hotstart DNA polymerase (Stratagene), 0.25 µM primers, 0.2 mM dNTPs and 4% DMSO in 1 x Herculase reaction buffer. Thermocycling was performed with 1 cycle of 95°C for 2 min, and 35 cycles of 95°C for 15 s, 50°C for 30 s and 72°C for 5 min, followed by one cycle of 72°C for 5 min in a Mastercycler gradient thermocycler (Eppendorf). Nested PCR was performed similarly using 1 µl primary reaction as template. For EAV-0, nested PCR was carried out with primers EAV0ForNot (5’-ATTGGCGGCGCGCCATTTACCGTTACCACGTC-3’) and EAV0Rev (5’-CTTACATCCCGAGACCCC-3’), followed by primers EAV0ForNot and EAV0Rev2Not (5’-CTTACGCGCGGACGCCGTATTACAAACATC-3’). Nested amplification of the EAV-E51 3 kbp provirus was performed with primers E51U5For (5’-ATTGGCGGCGCGCCATTTACAGTACCAACGTC-3’) and E51U3Rev1 (5’-CTTACATCCCGAGACCCC-3’), followed by primers E51U5For and E51U3Rev2Not (5’-CTTACGCGCGGACGCCGTATTACAAACATC-3’). Amplification of the EAV-E51-related gag-pol fragments was carried out using a single PCR with primers E51U5For and L4E9 (5’-CGTAAGGAGAGACGCAAGG-3’), agarose gel-purified PCR products were A-tailed and cloned into the pGEM-T Easy vector (Promega), and EAV clones (pEAV3, pEAV5 and pEAV21) were sequenced.

**Southern blot analysis of the E51 RT region.** A 275 bp E51 RT region was PCR amplified from line 21 chicken DNA with primers RT1 (5’-CTTTCGGTTGCTGATCAGAC-3’) and RT3 (5’-TGCGCATATTACCGCTAC-3’), and cloned into the pGEM-T Easy vector to produce clone pGEM-RT6. For preparation of the probe, the insert was amplified with SP6 and T7 primers, radiolabelled, and hybridized to membrane blots of EcoRI-digested genomic DNA as described previously (Sacco et al., 2000). High stringency washes were performed as detailed above for hybridization screening of the phage lambda library.

**Sequence and phylogenetic analysis.** Sequence identity to published retroviral sequences available in the National Center for Biotechnology Information database and *Gallus gallus*-4.0 assembly (released November 2011) were performed using the BLAST search tools (www.ncbi.nlm.nih.gov/BLAST). For phylogenetic analysis of
deduced protein sequences, DNA sequences were translated to protein using the Translator in the JustBio suite (http://www.justbio.com/) and aligned using the CLUSTALW2 algorithm-based aligner program at the European Bioinformatics Institute (Goujon et al., 2010; Larkin et al., 2007). Phylogenetic trees were generated using SeaView version 4 (Gouy et al., 2010) using the Bio neighbour-joining (BioNJ) method with Poisson distances and confidence limits of branching points from 1000 bootstrapped datasets.

**Nucleotide sequence accession numbers.** New nucleotide sequences presented in this paper have been submitted to Genbank/EMBL/DDBJ with accession numbers as follows: pEAV5 (AM418554), pEAV3 (AM418555), pEAV21 (AM418557), pEAV27 (AM4185560) and pEAV-E51 (AM418553). Sequences retrieved from the *G. gallus* genome database are indicated in Tables 1 and 2. Published sequences used in alignments for comparative analyses are as follows (accession numbers in parentheses): chicken transmembrane protein, 3′ end, clone EAV-E51 (M95189), chicken transmembrane protein, clone EAV-E33 (M95190), *G. gallus* (clones 14, 15) complete retrotransposon ART-CH (L25626), layer-type chicken line 151 EAV-HP clone EAV-151 (AJ622389), grey jungle fowl EAV-HP clone EAV-JF1 (AJ279966), locus ev-1 ALV-E (AY013303), subigroup J ALV (ALV-J) prototype HPRS-103 (Z46390), RSV-B Schmidt-Ruppin B strain (AF052428), RSV-C Prague strain; (J02342), ALV subgroup A cloning vector RCAS-A (AF484679), MMTV p27 CA (NP 955569), Rous-associated virus RAV-1 (AAA67094), Myeloblastosis-associated virus MAV-1 (AAA64305) and LPDV of *Meleagris gallopavo* (turkey) Pol precursor (AAA62195).

**ACKNOWLEDGEMENTS**

This work was partly funded by the British Biotechnology and Biological Sciences Research Council (V. N. K.) and start-up funds from California State University, Fullerton (M. A. S.).

**REFERENCES**


