Unexpected Diversity and Expression of Avian Endogenous Retroviruses

Mohan Bolisetty,[a] Jonas Blomberg,[b] Farid Benachenhou,[b] Göran Sperber,[c] and Karen Beemon[a]

Department of Biology, Johns Hopkins University, Baltimore, Maryland, USA[a]; Section of Virology, Department of Medical Sciences, Uppsala University, Uppsala, Sweden[b]; and Section of Physiology, Department of Neuroscience, Uppsala University, Uppsala, Sweden[c]

ABSTRACT **Endogenous retroviruses (ERVs) were identified and characterized in three avian genomes to gain insight into early retroviral evolution. Using the computer program RetroTector to detect relatively intact ERVs, we identified 500 ERVs in the chicken genome, 150 in the turkey genome, and 1,200 in the zebra finch genome. Previous studies suggested that endogenous alpharetroviruses were present in chicken genomes. In this analysis, a small number of alpharetroviruses were seen in the chicken and turkey genomes; however, these were greatly outnumbered by beta-like, gamma-like, and alphabeta proviruses. While the avian ERVs belonged to the same major groups as mammalian ERVs, they were more heterogeneous. In particular, the beta-like viruses revealed an evolutionary continuum with the gradual acquisition and loss of betaretroviral markers and a transition from beta to alphabeta and then to alpharetroviruses. Thus, it appears that birds may resemble a melting pot for early ERV evolution. Many of the ERVs were integrated in clusters on chromosomes, often near centromeres. About 25% of the chicken ERVs were in or near cellular transcription units; this is nearly random. The majority of these integrations were in the sense orientation in introns. A higher-than-random number of integrations were >100 kb from the nearest gene. Deep-sequencing studies of chicken embryo fibroblasts revealed that about 20% of the 500 ERVs were transcribed and translated. A subset of these were also transcribed *in vivo* in chickens, showing tissue-specific patterns of expression.**

IMPORTANCE **Studies of avian endogenous retroviruses (ERVs) have given us a glimpse of an earlier retroviral world. Three different classes of ERVs were observed with many features of mammalian retroviruses, as well as some important differences. Many avian ERVs were transcribed and translated.**

Received 9 September 2012 **Accepted** 11 September 2012 **Published** 16 October 2012

Citation Bolisetty M, Blomberg J, Benachenhou F, Sperber G, and Beemon K. 2012. Unexpected diversity and expression of avian endogenous retroviruses. mBio 3(5):e00344-12. doi:10.1128/mBio.00344-12.

Editor Stephen Goff, Columbia University

Address correspondence to Karen Beemon, klb@jhu.edu.

M.B. and J.B. contributed equally to this article.

Chromosomal integration, an essential part of the retroviral lifestyle, generates endogenous retroviruses (ERVs) when germ line cells are infected; this leads to their vertical transmission (1, 2). Because of this intimate connection between host and virus, a predominant evolutionary pattern is that the virus follows the evolution of its host. Thus, the expected major avian retroviral repertoire should be based on retroviral variants extant in vertebrates >100 million years ago (3, 4). However, horizontal transfer between highly divergent branches in the vertebrate tree can also occur. Prime examples are the infection of birds by the reticuloendotheliosis virus (5) and the infection of marsupials (koalas) (6) and primates (gibbons) (6, 7) by a murine leukemia virus. Thus, avian retroviral evolution is expected to be a composite of vertical and horizontal transmission.

Endogenous proviruses once contained the full set of regulatory elements present in exogenous viruses. However, over many millions of years of evolution, most of these regulatory elements have mutated and are thought to have little impact on current gene expression or regulation (8). Random integration events create genetic diversity and variance in gene expression that can impact the reproductive fitness of an organism. For example, the expression of salivary amylase in humans is controlled by a retrovirus inserted upstream of the salivary amylase gene (9).

In the present work, we examined the evolution of avian retroviruses on the basis of their fossil remnants in the three avian genomes that have been completely sequenced: two galliform birds, the chicken (*Gallus gallus*) (10) and the turkey (*Meleagris gallopavo*) (11), and a passeriform bird, the zebra finch (*Taeniopygia guttata*) (12). The chicken and turkey are separated by around 40 million years (13), while the zebra finch is separated from the others by around 100 million years (13–15). The chicken and turkey, but not the zebra finch, were found to have a lower ERV density than most vertebrates. Avian ERVs had features of the three major exogenous retroviral classes: gammaretroviruses, betaretroviruses, and spuma-like retroviruses (16–17). However, most of the avian ERVs were distinct from those of other vertebrates. We also saw ERVs with mixed genomes. Retroviral taxonomic markers, developed for mammalian retroviruses, segregated independently, suggesting steps of marker acquisition and loss during retroviral evolution. The well-studied alpharetroviruses (ALVs) seem to have evolved from beta-like precursors late during avian evolution. An alphabeta clade, intermediate between

TABLE 1 Identification and initial classification of ERVs in three avian species[a]

	No. of ERVs initially classified by ReTe as follows:									
Species	α	αβ	β	γ	δ	ε	Lentivirus	Spumavirus	Other	Total
Chicken	31	76	161	208	0	0	0	3	13	492
Turkey	8	4	71	61	0	0	0	0	6	150
Zebra finch	0	9	696	470	0	0	0	18	18	1,221

[a] α, alpharetrovirus; αβ, alphabetaretrovirus; β, betaretrovirus; γ, gammaretrovirus; δ, deltaretrovirus; ε, epsilonretrovirus.

alpha and beta proviruses, included some earlier recognized endogenous avian (EAV) proviruses (18).

We saw a random number of ERV integrations in chicken transcription units but a greater-than-random number of integrations >100 kb away from annotated genes. There were also clustered integrations in the bird chromosomes; some of these were near the centromeres. Interestingly, 20% of the endogenous proviruses identified were transcribed and translated in chicken embryo fibroblasts (CEFs); a subset of these were also transcribed *in vivo* in chickens.

RESULTS

Beta- and gamma-like ERVs are most common in all three avian genomes. An analysis of the three sequenced avian genomes (10–12) was initially carried out using the RetroTector (ReTe) program (19) to detect relatively complete ERVs; the average size was 7 kb. The zebra finch genome contained the most proviruses (1,221), followed by the chicken (492) and turkey (150) genomes (Table 1). For comparison, the human genome is about three times as large as that of these birds (3 billion versus 1 billion bp) and has 3,167 proviruses detected by the same ReTe stringency settings (20). Thus, zebra finch ERVs were comparable in number to human ERVs when adjusted for genome size, while the chicken and turkey genomes had 2 and 7 times fewer ERVs, respectively. In addition to these proviral ERVs, there are many single long terminal repeat (LTR) sequences. These were previously analyzed by RepeatMasker, which found 30,000 LTRs in the chicken genome and 78,000 in the zebra finch genome (12). In each case, the number of solo LTRs was about 60 times the number of nearly full-length ERVs detected by ReTe.

ReTe provides a preliminary genus classification based on motif usage (19); this was used initially to classify the 1,863 avian ERVs (Table 1). Since endogenous alpharetroviruses had previously been observed in chickens (21), we were not surprised that 31 alpha ERVs were detected in chickens (Table 1). However, we saw many other genera, including 208 gamma-like and 161 beta-like retroviruses, as well as 76 alphabeta intermediate ERVs and 3 spuma-like (class III) ERVs. Turkeys had a similar distribution of ERV genera, although fewer total ERVs were observed. The zebra finch genome also had predominantly beta-like (696) and gamma-like (470) ERVs. No alpha ERVs were present, but there were 9 alphabeta and 18 spuma-like ERVs (Table 1). None of the avian genomes analyzed had detectable delta, epsilon, or lentivirus ERVs.

To further analyze the relationships among these 1,863 avian ERVs, we created a phylogenetic tree. The ERVs were first organized into clusters based on similarities in concatenated Gag, Pro, and Pol proteins. This resulted in 480 clusters with 2 to 78 members each. In addition, 558 ERVs were not sufficiently similar to others to be clustered. A second reduction of the data set was obtained by selecting all proviruses with ReTe scores of >1,000, in addition to the best representative of each large cluster with 10 or more members. In this way, 128 proviruses were selected, representing the most intact and most abundant avian ERVs. A phylogenetic tree was constructed from their aligned Pol amino acid sequences, using the minimum-evolution (ME) algorithm (see Fig. S1 in the supplemental material).

To illustrate the major features of this tree, it was manually simplified to depict the major branches, together with reference retroviruses (Fig. 1). The selected avian proviruses could be ascribed to ERV class I (gamma-like), class II (alpha- and beta-like), or class III (distantly spuma-like). An intermediate group also was seen, which we termed "alphabeta." The rationale for this clade was its sequence similarity intermediate between alpha and beta, its intermediate motif usage recorded by ReTe (see Fig. S2 in the supplemental material), its absence of dUTPase in Pro (an alpha-like feature), and its use of the −1, −1 *gag*/*pro* and *pro*/*pol* frameshifts (a beta-like feature).

According to this tree (Fig. 1), the class III ERVs appear to be the oldest (i.e., closest to the root), followed by the beta-like and gamma-like ERVs. The alphabeta and alpha viruses appeared to be the youngest. Approximate proviral ages can also be estimated from the degree of *pol* intactness and the divergence of LTRs from

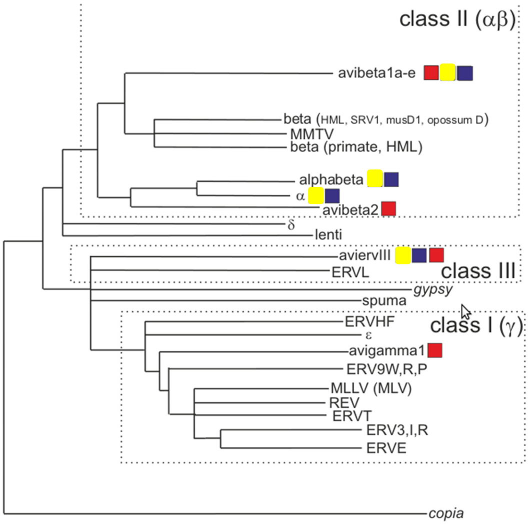

FIG 1 Phylogenetic analysis of avian endogenous proviruses. Shown is a simplified version of the ME tree in Fig. S1 in the supplemental material, resulting from the alignment of selected avian and reference Pol sequences. MLLV, mouse leukemia-like virus sequences (defined in reference 7). Clades containing chicken (yellow box), zebra finch (red box), and turkey (blue box) endogenous proviral chains are shown. MMTV, mouse mammary tumor virus; HML, human MMTV-like; MLV, murine leukemia virus; REV, reticuloendothelial virus.

one another (see Table S1 in the supplemental material). This analysis suggested an evolutionary continuum leading from betaretroviruses to alphabetaretroviruses to alpharetroviruses, where the latter two steps occurred mainly in galliform birds.

Major branches of this tree were used to group the ERVs into six Pol-based clades (Fig. 1) using a custom tree-directed grouping algorithm (J. Blomberg, unpublished data). The clades were named avibeta1a to -e (avibeta1 contains five subclades), avibeta2, alphabeta, alpha, avian ERV III (aviervIII), and avigamma1. The members of each clade had related *pol* sequences and other proviral features, as shown in Table S1. The total of 1,863 proviruses identified were next distributed into six clades and five subclades, where possible, by BLAST analysis of proviral DNA, as well as Gag, Pro, and Pol consensus amino acid sequences. The highest-scoring representative of each clade was compared with each of the 1,863 proviruses. Table S1 shows the number of ERVs in the total set that could be mapped to one of these clades. The BLAST score limits for inclusion in a clade varied according to the sequence compared (see Materials and Methods). The classification was based mainly on proviral DNA similarity, but concatenated Gag, Pro, and Pol sequences were also used and yielded very similar results. Altogether, 46% of the 1,863 proviruses could be classified into one of the clades in this fashion (see Table S1).

We also searched in GenBank for previously published avian ERVs (22–25). All 23 ERVs identified previously were represented in minor clades within the 1,863 sequences (see Table S2 in the supplemental material). In two cases (FET-1 and Chirv1), an overlap with avigamma1 occurred. Otherwise, these clades did not overlap the major clades identified in this paper (see Table S1).

The avian class I sequences included one clade, the avian gamma clade (avigamma1), which was observed in all three avian species but was most common in the zebra finch. The avigamma1 clade was quite homogeneous, and members commonly demonstrated the typical gammaretrovirus 0, 0 frameshift strategy and a gamma *env* gene containing the conserved immunosuppressive domain in the transmembrane protein (26). All members also had a GPY/F domain in integrase, which is common among gammaretroviruses. Their closest relatives were the gammaretrovirus-like human ERV HERV-W, an opossum ERV, and the walleye dermal sarcoma virus, an epsilonretrovirus.

In contrast, the avian class II ERVs were classified into four major clades, alpha-like, alphabeta-like, avian beta-like 1 (avibeta1), and avibeta2. Further, the avibeta1 clade was divided into five subclades (see Table S1 in the supplemental material) based on sequence similarity, the host, and the presence or absence of taxonomic markers. Four of these subclades contained zebra finch ERVs, while avibeta1c was most common in chickens and turkeys. Avibeta1e contained mainly chicken and zebra finch ERVs. The five subclades illustrated that some of the beta taxonomic markers, including Pol similarity and the presence of dUTPase in Pro and two zinc fingers in the Gag nucleocapsid protein, were represented in specific patterns in each of the subclades (see Table S1), while the beta −1, −1 frameshift pattern was common to all. None of the avian ERVs we identified had the mammalian betaretroviral marker Gpatch in protease (data not shown), as previously observed (27).

The avibeta2 clade was more homogeneous than avibeta1. It lacked dUTPase in Pro, had two zinc fingers in Gag, and frequently had the −1, −1 frameshift combination, although 0, −1 frameshifts also occurred. Its most closely related exogenous retrovirus was the lymphoproliferative disease virus of turkeys, which used a −1, −1 frameshift (data not shown). Even closer to the origin of the betaretroviral branch was a *Python molurus* retroviral sequence (see Fig. S1 in the supplemental material), which also used the −1, −1 strategy. This suggests that the frameshift patterns of late offshoots from the beta branch were acquired late in evolution. The primer-binding site of avibeta2 was frequently complementary to the phenylalanine tRNA; in contrast, mammalian betaretroviruses typically use a lysine tRNA primer for reverse transcription.

Interestingly, the avibeta2 envelope proteins were very similar to the envelopes of the avigamma1 clade. Thus, a recombination event between beta and gamma ERVs appears to have occurred during the evolution of the zebra finch. Several of the avibeta2 proviruses were also recombinants with EAV51-like (18, 28) alphabeta proviruses in the distal portion of Pol (data not shown). Cross-clade recombinations were also detected in other clades, including the alphabeta, avibeta1, and aviervIII clades, and usually involved envelope (see Table S1 in the supplemental material).

Envelope genes were delineated by ReTe in only a few of the 128 selected proviruses, belonging to the avibeta1c, avibeta1e, avigamma1, and aviervIII clades. However, transmembrane sequences were detected in the avibeta2 clade. When the whole set of 1,863 proviruses was reclassified by BLASTing with the 10 consensus viral protein sequences derived from the 128 selected proviruses, more envelope sequences were detected in each of the 10 clades (see Table S1).

Alpha- and alphabeta-like retroviruses occurred only in the two galliform birds (chicken and turkey). The likely immediate predecessors of alpharetroviruses, the alphabeta clade, commonly used the −1, −1 frameshift strategy, as did the betaretroviruses. In contrast, 0, −1 is the strategy used by exogenous (8) and endogenous (see Table S1) alpharetroviruses. The alpha clade included two RAV-0-like viruses and five ALV-like viruses, with either an A or an E envelope, as well as nine EAV-HP-like viruses. Two-thirds of the alphabetaretroviruses were highly similar to the previously described EAV ERVs, EAV0 and EAV51 (28) (see Fig. S4 in the supplemental material). The EAV0 provirus sequence was 94% identical to the alphabeta consensus sequence and also uses the −1, −1 frameshift strategy (data not shown). This is of interest because recombinants of exogenous ALVs with envelopes from endogenous alphabeta proviruses have become important pathogenic exogenous retroviruses (ALV subgroup J) (28, 29).

The avian class III sequences contained one clade, aviervIII (Fig. 1), which clustered with the murine ERV class III virus MERV-L and also had motifs similar to those of gamma- and betaretroviruses. The aviervIII proviruses were most similar to HERV-S and HERV-L Pol amino acid sequences and to human spumaretrovirus DNA. The best fit among human Repbase LTR repeats was with HERV18. A previously described turkey provirus, birddawg (30), clustered with aviervIII.

When *pro*/*pol* portions of our avian retroviral clades were analyzed together with 136 *pro*/*pol* sequences (900 to 1,000 bp) from a wide variety of host species (27), nearly all of the avian retroviral sequences clustered into two major groups, beta-like and gamma-like (see Fig. S3 in the supplemental material). The beta-like viruses included alpha and alphabeta sequences. Thus, it seems that the alphabeta, avibeta, and avigamma clades identified here are also present in many other birds and reveal some major general features of avian retroviruses.

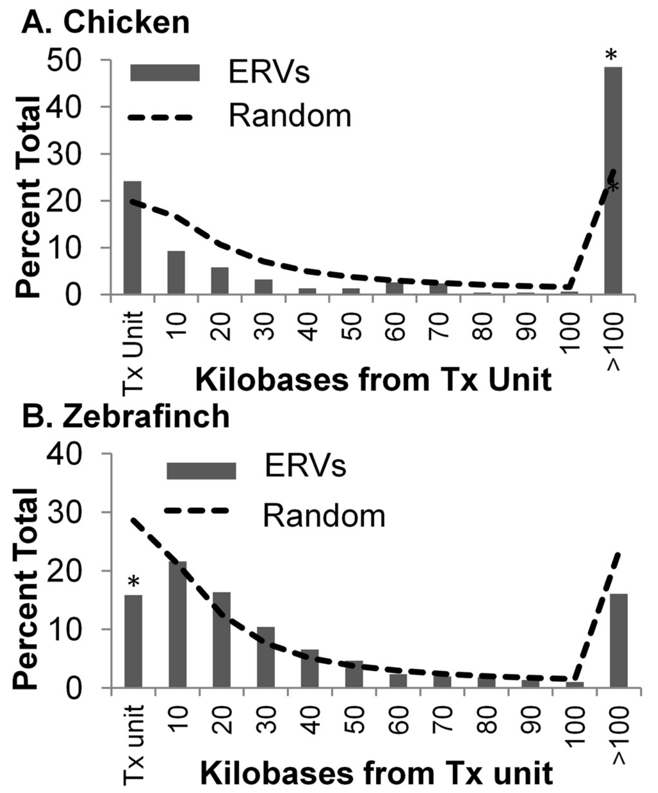

FIG 2 Distribution of endogenous proviruses with respect to transcription units in the chicken (A) and zebra finch (B) genomes. The endogenous proviruses were mapped with respect to distance from the nearest transcription (Tx) unit and binned into 10-kb segments (bar). Ten million random integrations were simulated and were binned on the basis of distance from the nearest gene (dashed line). Differences that were statistically significant (P value of <0.05; χ^2) are indicated by asterisks.

To assist further work with the avian ERVs, a tree based on consensus sequences and the most intact avian ERVs is provided in Fig. S4 in the supplemental material. In addition, diagrams of representative genomes for each clade are shown in Fig. S5 in the supplemental material.

One-fourth of chicken ERVs are within or near cellular genes. The 492 chicken endogenous proviruses identified were individually mapped, using BLAT, to the galGal3.0 genome on the University of California Santa Cruz (UCSC) Genome Browser. For all of the analyses, a transcription unit was defined as either a RefSeq gene or an mRNA or microRNA transcribed from this region. The distance of each provirus from the nearest transcription unit was determined, and the data were catalogued in 10-kb intervals (Fig. 2A, bars). Approximately 25% of these integrations were either in transcription units or within 10 kb of each end, including promoters and 3′ regulatory sequences. These were grouped together and called transcription units because of uncertainties in the annotation of many gene endpoints.

To determine if this distribution had been subject to selective pressure, we compared it to a simulation of 10,000,000 random integrations. These simulated integrations were mapped onto the genome and similarly tabulated and plotted with respect to distance from the nearest transcription unit (Fig. 2A, dotted line). As indicated by the χ^2 analysis, the distribution of the endogenous proviruses was nearly random, except for proviruses that were integrated >100 kb away from the nearest transcription unit. According to our simulation of random integrations, 26% of all integrations should be greater than 100 kb away from transcription units if there was no selection; however, 49% of the observed integrations were >100 kb away. This indicates selection against integrations within 100 kb of transcription units. Surprisingly, the number of integrations within transcription units was slightly higher than random and integrations between 10 and 50 kb from transcription units were slightly less frequent than random. This may, in part, reflect the integration preferences of the different ERVs; for example, exogenous gammaretroviruses, such as murine leukemia virus, prefer to integrate near transcriptional start sites (31).

Next, we extended this analysis to the zebra finch genome. ReTe identified a total of 1,200 proviruses in the zebra finch genome. These integrations were mapped onto the genome, and the distance from the transcription units was catalogued as described above. In contrast to the chicken integrations, only 16% of all zebra finch integrations were found in transcription units (or within 10 kb of either end), in comparison to 28% in the random data set, which is a significant difference ($P < 0.01$ by χ^2 test) (Fig. 2B). This indicated a strong selection against ERVs within transcription units in the zebra finch genome. Further analysis showed that 23% of the integrations in the zebra finch genome were within 10 kb of either end of a transcription unit. Thus, the total number of integrations either within or near genes was comparable for the chicken and the zebra finch. Since the zebra finch was more recently sequenced and annotated, it is possible that the ends of the genes are not as well defined as in the chicken, resulting in an elevated number of integrations near but not in genes. Alternatively, the zebra finch ERVs may have a preference for promoters or 3′ ends of genes, possibly reflecting the different types of ERVs in the two birds. In contrast to the chicken, the zebra finch integrations were somewhat below random at sites >100 kb away from transcription units.

Endogenous proviruses in or near chicken genes are more recent integrations. Since many of these ERVs have been subjected to mutations during evolution, it is important to determine how much these sequences have diverged. The more diverged a sequence, the lower the possibility of retaining functional gene regulatory elements. The age of a proviral integration can be approximated from LTR divergence, intactness of proviral open reading frames (ORFs) (see Table S1 in the supplemental material), and the presence of secondary integrations of other transposable elements. We used LTR divergence to assess age in this study, calculating the differences between the 5′ and 3′ LTRs of each provirus as described previously (19). The calculation is dependent on an exact delineation of LTRs, which ReTe cannot achieve in all cases. Secondary transposon integrations create a further uncertainty. Still, an LTR divergence of 0 to 10% suggests a relatively recent origin of the integration. Around 50% of the avian proviruses detected lacked two discernible LTRs and could not be used for this analysis. These ERVs are probably much older and more evolved, thereby precluding LTR comparison.

Almost 70% of the chicken proviruses with two LTRs showed >90% LTR identity, indicating that they are relatively recent integrations (Fig. 3B). Next, we were interested in investigating the age of proviruses integrated in and near transcription units. Interestingly, 70% of the chicken proviruses with two LTRs that inte-

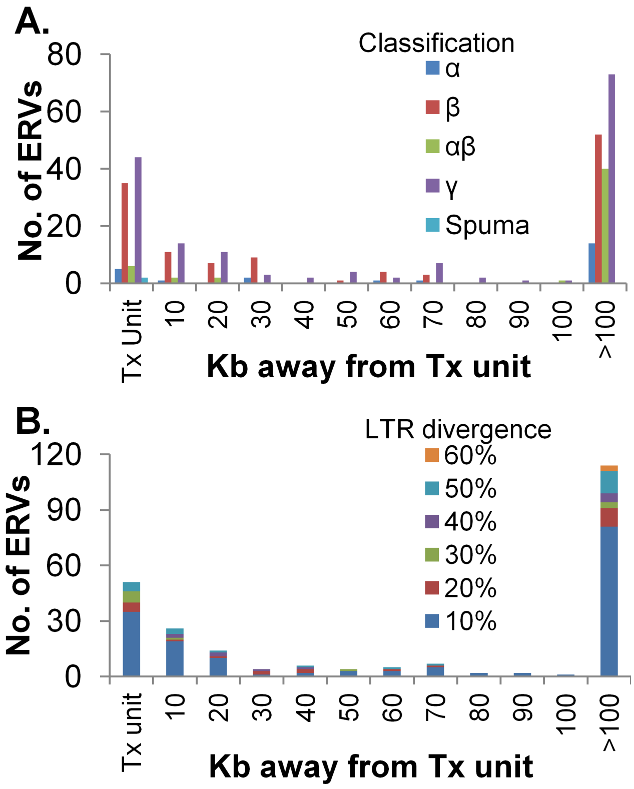

FIG 3 Alphabeta ERVs are excluded from chicken transcription (Tx) units. (A) All endogenous proviruses were classified on the basis of their genera (as in Table 1) and their distances from the nearest transcription unit. (B) The age of each provirus was approximated by calculating the sequence divergence between the two LTRs that flank the provirus (10% divergence denotes a range of 0 to 10%; 20% denotes 11 to 20%, etc.). The endogenous proviruses were classified on the basis of distance from the nearest transcription unit and plotted on the basis of LTR divergence. Older ERVs were more prevalent far away from transcription units.

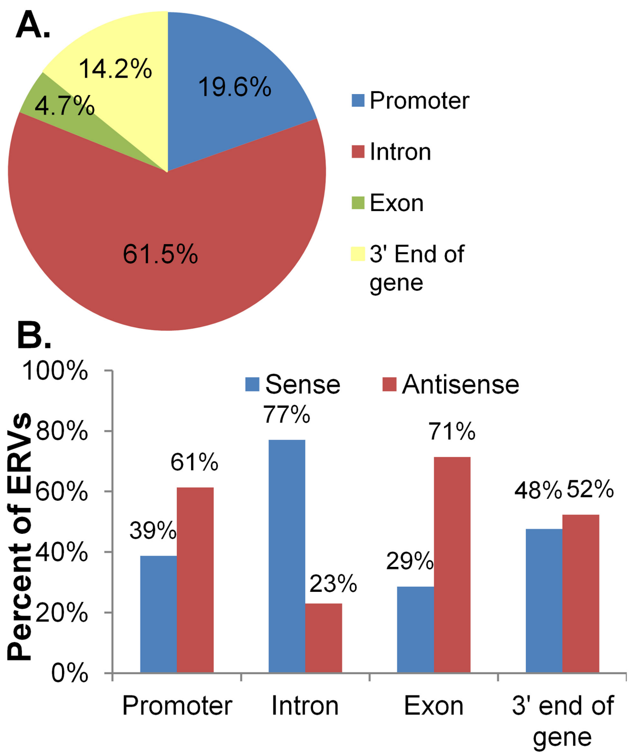

FIG 4 Distribution of chicken endogenous proviruses that could alter gene expression. (A) The positions of all endogenous proviruses that were within a transcription unit were further divided into promoters of transcription (less than 10 kb upstream), introns, exons, and ends of transcription units (less than 10 kb downstream). (B) The orientation of all endogenous proviruses within a transcription unit was calculated with respect to their position in a transcription unit.

grated within a transcription unit have relatively similar LTRs (less than 20% divergent). In contrast, 60% of the more divergent LTRs (greater than 30% divergent) are found farther away from genes.

Next, we looked at the relative age of proviruses integrated >100 kb away from chicken genes. These areas probably have very little selective pressure to maintain LTR sequence. The ERVs with more divergent LTRs were more prevalent >100 kb away from genes. In addition, almost 50% of retroviral chains with less divergent LTRs (less than 20%) are found in this region. Thus, the younger proviruses appeared to be more common within genes and the older proviruses were enriched >100 kb away from genes.

The distribution of these proviruses with respect to transcription units was similar between genera (as defined in Table 1) and random, with the exception of the alphabeta clade. Seventy-eight percent of the alphabeta proviruses were >100 kb away from chicken transcription units (Fig. 3A), which is significantly higher than random (26%). This indicates a strong selection against alphabeta integrations in or close to transcription units.

Many chicken endogenous proviruses are found in promoters and introns. We next characterized the distribution of proviruses in the chicken within and around (+10 kb) transcription units. Some of these integrations could alter gene expression by inserting new promoter elements, as well as splicing, polyadenylation, or other regulatory elements. In addition, the insertions could disrupt genes. We found that 62% of the proviruses within or near a transcription unit in the chicken were in introns (Fig. 4A); introns make up 42% of these transcription units. Surprisingly, 77% of these intronic proviruses are in the same orientation as the transcription unit (Fig. 4B), increasing the possibility that they may provide regulatory sites.

Interestingly, 20% of the ERVs in or near transcription units were in the promoter region (up to 10 kb upstream). Irrespective of the orientation, both of the LTRs have regulatory elements that can affect gene transcription. More than 60% of these proviruses in the promoter regions are in the opposite orientation, raising the possibility that the enhancer elements in the 5′ LTR could regulate gene transcription. In addition, a small number of proviruses actually flanked exons. This raises the interesting possibility that the exon may have evolved from the provirus.

Around 15% of retroviral chains within or near transcription units were within 10 kb of the annotated 3′ end of the gene, which in many cases does not include the 3′ untranslated region (UTR). Recently, it has become increasingly apparent that 3′ UTRs play a big role in gene regulation, since they are targeted by microRNAs (miRNAs) and RNA binding proteins. The 3′ UTRs of many transcriptional units are not tightly defined; therefore, many of these retroviral chains could provide alternative polyadenylation sites and therefore influence gene expression.

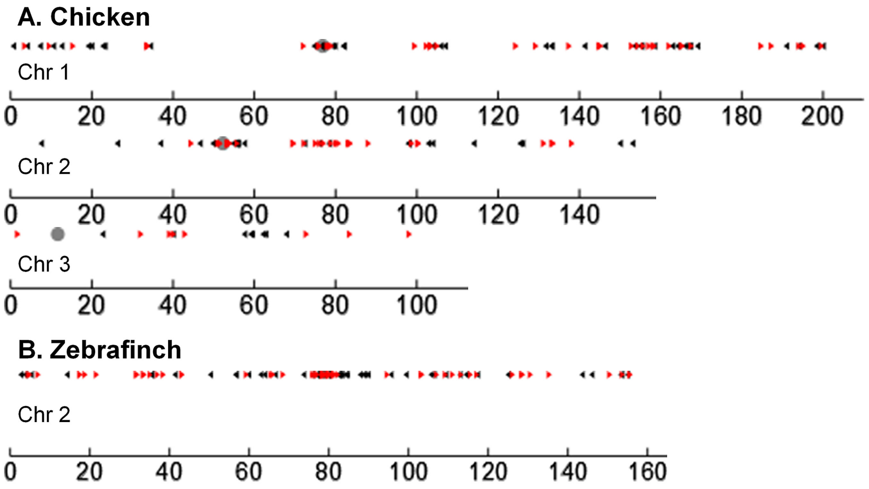

FIG 5 Clustering of endogenous proviruses in the chicken and zebra finch genomes. The positions of endogenous proviruses on chromosomes (Chr) 1, 2, and 3 of the chicken (A) and chromosome 2 of the zebra finch (B) were plotted. The gray circles indicate the centromeres of the chromosome. The red and black arrows indicate the direction of integration of each endogenous provirus.

Many ERVs are clustered on chromosomes. During our analysis of endogenous provirus integrations, we noticed that there were some chicken chromosomal loci with a high concentration of integrations. We decided to map the integrations onto the corresponding chromosomes to identify regions enriched for integrations. We split each chromosome into million-base fragments and mapped the integrations back onto the chromosome. A random distribution of ERVs in the chicken genome would result in one integration per 2 × 10^6 bp. As illustrated in Fig. 5A, we uncovered five different clusters on chromosome 1 of the chicken and three clusters on chromosome 2. We found a centromeric integration cluster on both chromosomes 1 and 2 with 12 and 8 different endogenous proviral integrations, respectively (Fig. 5A). The largest cluster outside the centromere had seven endogenous proviruses in a region of 2 × 10^6 bp. There was very little correlation observed between the members of a cluster and specific genera or the age of the proviruses (data not shown), suggesting that they did not arise by duplication. In general, clusters were found in gene-poor regions of the genome (data not shown). About 40% of the chicken ERVs were found in clusters.

Analysis of the zebra finch genome led to similar conclusions. However, there are many more clusters in the zebra finch genome than in the chicken genome, in part because of the larger number of ERVs identified in the zebra finch. We identified 7 clusters on chromosome 1 (data not shown) and 11 clusters on chromosome 2 (Fig. 5B). One of the clusters on chromosome 2 had 40 integrations within 5 × 10^6 bp, which is significantly higher than random (1 integration in 10^6 bp). Although the centromeres of the zebra finch have not been mapped, it is possible that one of the clusters represents the centromere of this chromosome, by analogy to the chicken.

Transcription of ERVs in CEFs and *in vivo* in chickens. Since many of these endogenous proviruses seem to have reasonably intact LTRs and downstream ORFs with relatively few stop codons (data not shown), we wondered if they were expressed. We conducted mRNA-seq analysis to determine if specific proviruses were transcribed. We performed this analysis using chicken embryo fibroblasts (CEFs). Total poly(A)$^+$ mRNA was isolated from CEFs, and an Illumina high-throughput sequencing library was prepared. The sequence reads were aligned with the chicken genome index using the short-read alignment program Bowtie (32). The alignment output was then parsed to yield reads that mapped only to locations identified to be endogenous proviruses. We defined an alignment as correct if it was a unique alignment with respect to the entire genome and had fewer than two mismatches with the aligned sequence. Using these parameters ensured that a given provirus-related sequence read would be mapped to only one provirus. This yielded a total of 34,564 mRNA-seq reads uniquely aligning with the endogenous proviruses identified in this study. Of the 492 endogenous proviruses, 118 had at least 10 mRNA-seq reads mapping to a predicted proviral chain (Fig. 6A). There was wide variation in expression levels, with a few ERVs having 1,000 or more mRNA-seq reads.

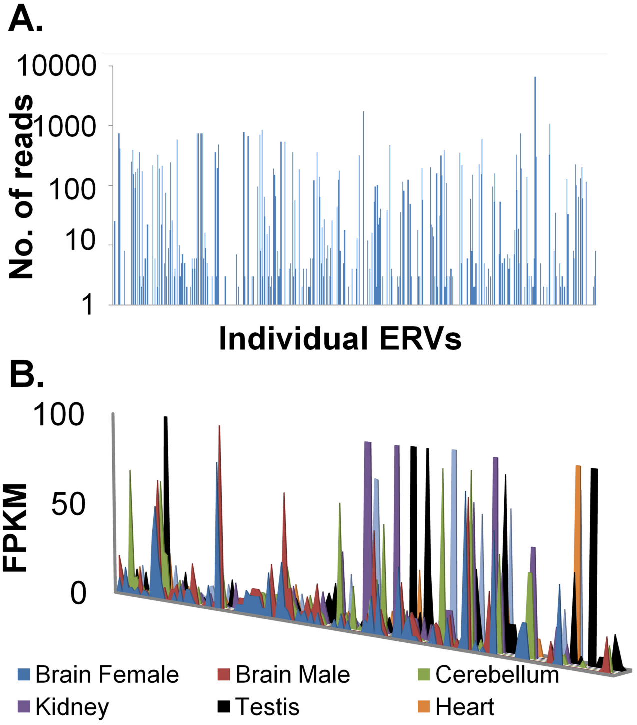

FIG 6 Expression of endogenous proviral chains in CEFs and various chicken tissues. (A) mRNA-seq reads were uniquely aligned with the genome. The mRNA-seq reads mapping to any given endogenous proviral chain were counted and plotted on a log scale. The *x* axis represents unique endogenous proviruses identified by ReTe, with increasing scores from left to right. (B) RNA-seq data from various chicken tissues (33) were aligned using TopHat (47), and transcripts were constructed using Cufflinks (48). Transcripts that overlap ERV coordinates were extracted, and the number of fragments per kilobase per million (FPKM) was plotted against ERV chains. On the *x* axis, from left to right, are unique ERV chains with increasing ReTe chain scores.

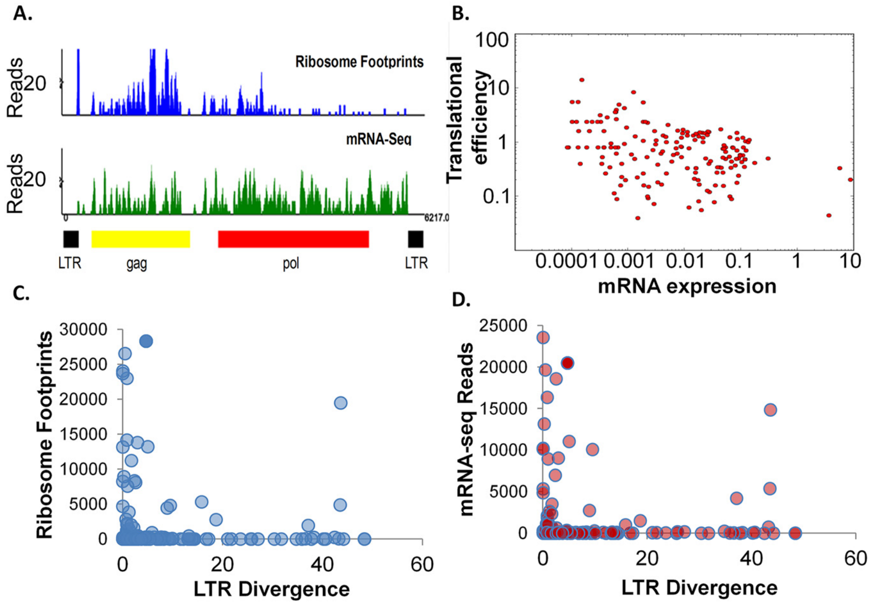

FIG 7 Many ERVs in the chicken genome are transcribed and translated. (A) Most of the ribosome footprints of an endogenous proviral chain are distributed in the predicted *gag* ORF. Ribosome footprints and the corresponding mRNA-seq reads from one experiment are plotted with respect to the predicted proviral chain. ReTe-predicted retroviral features of the proviral chain are shown below. (B) Translational efficiency (ribosome footprints/mRNA-seq reads) of all translated endogenous proviral chains was plotted against normalized mRNA expression from the same data set. The LTR divergence calculated by ReTe is plotted against the number of ribosome footprints (C) or mRNA-seq reads (D) for each proviral chain.

In order to verify that these proviruses were indeed being transcribed in the chicken and not just in cell culture, we analyzed publicly available chicken mRNA-seq data (33) for the expression of these proviruses in various chicken tissues (Fig. 6B). Approximately 50 of these ERVs were expressed in many tissues, and some of them appeared to be expressed in a tissue-specific manner (Fig. 6B).

Many chicken ERVs are translated. We next investigated whether the endogenous proviruses were just transcribed or if they could also be translated to produce protein products. Using ribosome footprinting followed by high-throughput sequencing (34), we analyzed the presence of ribosomes on the transcripts generated from proviruses as an alternative way of studying translation. We analyzed ribosome footprints of proviruses in CEFs from two unrelated samples (data not shown). Sequence reads from both samples were aligned with Bowtie (32), using parameters described in Materials and Methods. Since endogenous proviral chains contain repetitive sequences, only unique alignments with the genome and proviruses were considered for the analysis. In fact, 118 of these endogenous proviruses were translated in both of these data sets. Most of these were translated to similar levels in both of the data sets, as indicated by a Spearman rank correlation of 0.96 (data not shown).

Next we compared the distribution of ribosome footprints across the provirus to ReTe-predicted domains, including the LTR, *gag*, *pro*, *pol*, and *env* domains. Most of the ribosome footprints across all of the translated proviral chains were in the *gag* ORF (Fig. 7A). These predicted *gag* ORFs are smaller than those of exogenous viruses, suggesting that the translation of these ERVs would result in truncated Gag proteins or small peptides. There were also a few instances of ribosome footprints observed in the *env* ORF (data not shown). Although the ribosome footprints are predominantly in the *gag* ORF, the mRNA-seq reads are more evenly distributed across the proviral chains, suggesting that *gag* is preferentially translated (Fig. 7A). Similar studies of the exogenous Rous sarcoma virus also showed a large number of footprints in *gag* and a low number in *pol* (K. Beemon and N. Ingolia, unpublished results).

In order to determine whether translation of the endogenous proviruses was significant, we compared the translational efficiency of annotated genes in the chicken genome with that of ERVs. Translational efficiency is defined as the ratio of the number of ribosome footprints to the number of RNA-seq reads for any given gene. The median translational efficiency of these endogenous proviruses was 0.8, and most ranged from 0.1 to 10.0 (Fig. 7B). In comparison, the median translational efficiency of cellular internal protein coding exons was around 5.0, that of 5′ UTRs was around 0.1, and that of 3′ UTRs was 0.01. As a point of comparison, the chicken actin gene had a translational efficiency of 4.5 in this experiment, while the glyceraldehyde 3-phosphate dehydrogenase gene had a translational efficiency of 8.1. This suggests that the translation of these endogenous proviral chains, although most levels were lower than those of housekeeping genes, was indeed significant and not a result of low-level nonspecific translation.

Furthermore, the majority of the proviruses that are transcribed and translated have a relatively low LTR divergence, indicating that these expressed proviruses are relatively recent integrations (Fig. 7C and D). Mutations in the gag AUG initiation codon could result in decreased ribosomal loading efficiency at alternative initiation sites. Further, mutations generating premature termination codons in *gag* in translated viral RNAs would likely result in partial RNA degradation by nonsense-mediated mRNA decay (35).

DISCUSSION

Analysis of ERVs in the three avian genomes suggests that birds were a melting pot for ERVs hundreds of millions of years ago. The most conspicuous events (gain or loss of taxonomic markers and emergence of the *Alpharetrovirus* genus) were confined to the beta-like lineage. Early studies of chicken ERVs observed only alpharetroviruses, including RAV-0, related to current exogenous avian viruses (8, 21). In contrast, in this study based on complete genome sequences, we observed ERVs in the three bird genomes that belonged to the same major groups as those of mammals, including the gamma-, beta-, and spuma-like groups. However, most of the avian ERVs were distinct from those of nonavian hosts, probably reflecting a separate evolutionary history. Recombinants between different groups were also observed. The zebra finch betaretroviruses were the most like those of other vertebrates.

Importantly, we found that the avian ERVs could be grouped into six clades and five subclades. ERVs in the betaretrovirus group did not have all of the taxonomic markers of existing mammalian betaretroviruses but had one or more beta-like features, suggesting a trial-and-error process of acquisition of beta markers during evolution. This independent segregation of betaretroviral features was most obvious in clade avibeta1. We suggest that some features of archaic betaretroviruses are reflected in the fluctuating use of frameshift strategies, the number of nucleocapsid zinc fingers, and the presence or absence of dUTPase (and the absence of Gpatch) in the protease reading frame. These features have not been recorded in extant nonavian betaretroviruses. A retrovirus which emerges close to the betaretroviral root, like the python retrovirus, uses −1, −1 frameshifts and has two zinc fingers and no dUTPase.

Further, a gradual evolutionary transition was inferred, from betaretroviruses to intermediate alphabetaretroviruses in all three birds and finally to alpharetroviruses only in chickens and turkeys. This differs somewhat from previous analyses (36), where a succession from alpha to alphabeta to beta was presented. However, the present data set is larger than the one used in that study, and its phylogenetic inference result is more logical than the previous one. Most current exogenous avian viruses (avian leukosis viruses [ALVs]) are in the *Alpharetrovirus* genus. Endogenous alpharetroviruses have been reported earlier only in galliform birds, including chickens (*Gallus gallus*) (21, 37, 38) and grouse (*Bonasa umbellus*) (39, 40). We report here that the turkey contains alpharetroviral proviruses while the zebra finch does not (Table 1; see Table S1 in the supplemental material). However, the genus vector indicated a low but persistent alpharetroviral similarity in the avibeta2 clade, which occurs in the zebra finch (see Fig. S2 in the supplemental material). It is likely that some avian betaretroviruses started to evolve toward alpharetroviruses more than 100 million years ago.

The number of ERVs identified in the chicken and turkey genomes was lower than that found in most other vertebrates (19). ReTe is based mostly on mammalian retroviruses; thus, a slight bias toward the detection of mammalian retroviruses can be expected. However, the avian alpha-, beta-, and gamma-like sequences described here got high ReTe scores. The turkey genome had 1/20 of the ERV content of the human genome. Although the avian genomes are about one-third the size of the human genome, the difference is still dramatic. Thus, the two galliform birds both had a light "ERV burden." In contrast, the zebra finch had a number of ERVs similar to those of humans when adjusted for genome size. It will be interesting to determine why these birds vary so much in ERV composition. Some vertebrates seem to efficiently remove repetitive elements by illegitimate recombination (41). Alternatively, they may efficiently restrict the replication of certain retroviruses by other, unknown, mechanisms.

Analysis of retroviral integrations also gives insight into the evolution of the host genome, since ERV integrations can alter host gene expression. Therefore, most endogenous retroviral integrations are thought to be subject to negative selection. However, we found a nearly random number of ERVs in chicken transcription units, with 62% of these in introns. It is interesting that 77% of these intronic integrations were in the sense reading frame, suggesting that they might alter gene expression through the insertion of promoters or splicing or polyadenylation sites. In contrast, an earlier study by Bushman and colleagues (42), which analyzed mainly free retroviral LTRs in the chicken genome, found negative selection within genes and especially in the sense orientation. We did see a significant enrichment of chicken ERVs greater than 100 kb from any mapped transcription unit, where they probably do not influence gene expression.

Endogenous proviral clusters in the genome suggest "dead spots" that do not interfere with host gene expression, although it is possible that they are transcribed (43). Proviruses have accumulated in these regions over many millions of years, suggesting that there might be no negative selection against these integrations. Alternatively, these ERV-rich clusters might undergo positive selection, perhaps because they play a role in generating genetic diversity by promoting recombination. Another possible function could be during cell division; these clusters might be loading points for cohesin or other factors involved in proper chromosome segregation (44).

Surprisingly, we found that many avian ERVs are transcribed and translated, both in CEFs in culture and in many chicken tissues *in vivo*. Further, some of these ERVs are expressed in a tissue-specific fashion. In the future, it will be interesting to study the role, if any, of these ERV RNAs and proteins.

We conclude that avian retroviral evolution differs from that of other vertebrates. Retroviral classes I, II, and III may have been present at the outset of reptile and dinosaur evolution 200 to 300 million years ago. Avian retroviruses seem to have evolved rather independently from the rest of the retroviruses over the last 150 million years, in rare instances complicated by horizontal interchange with nonavian phyla. Taxonomic markers, which segregate together in mammalian retroviruses, do not segregate as clearly in bird retroviruses. It is possible that the selective pressure on retroviral features was more specific in mammals than in birds.

MATERIALS AND METHODS

Identification and classification of avian ERVs. Three avian genomes (red jungle fowl, the ancestor of the domestic chicken, galGal3.0 [10]; zebra finch, taeGut1.0 [12]; and turkey, melGal1.0 [11]; downloaded from the UCSC Genome Browser) were examined for endogenous retroviral sequences using ReTe version 1.01 (19) with default settings and a proviral chain score cutoff of 300. Class III ERVs are not completely covered by ReTe, and this may lead to an underreporting of such sequences (19). To enable an overview of the 1,863 retroviral sequences integrated into the three host genomes, a first reduction of complexity was carried out by clustering into groups of high similarity. A custom algorithm was used to cluster proviruses according to concatenated Gag, Pro, and Pol sequences (as reconstructed by ReTe) at the level of a BLASTP score of 2,100 or higher (Blomberg, unpublished). The resulting clusters were at least 90% identical in this chimeric amino acid sequence. The highest-scoring member of a cluster was used as the source of sequence for phylogenetic inference.

MEGA version 5.05 (45) was used for phylogenetic inference. Trees were based on the polymerase amino acid sequence, and bootstrap analysis was carried out with 500 replicates. The tree in Fig. 1 is a manual simplification of the tree in Fig. S1 in the supplemental material, resulting from alignment of Pol sequences from relatively intact (score of >1,000) or especially prevalent proviruses (the most-intact member of a cluster of >10). Besides the preliminary genus classification inherent to ReTe, which builds on conserved motifs of reference retroviruses, a classification pipeline using data from ReTe was constructed in Visual FoxPro. It was built on the additional features (i) frameshift strategy, (ii) number of zinc fingers in Gag, (iii) presence of Gpatch (beta property), (iv) dUTPase in

Pro (beta), and (v) GPY/F domain in integrase plus the most similar reference genome detected by ReTe. A final classification of the whole 1,863 proviruses was made by BLASTing in successive steps. First the whole proviral DNA was searched with BLASTN using proviral DNA consensus sequences and a cutoff score of 1,000. Next, the predicted Gag, Pro, and Pol proteins were concatenated and searched against concatenated consensus sequences with BLASTP and a cutoff score of 800. Finally, the envelopes were classified using BLASTP and a cutoff score of 200. The consensus sequences for all of the clades and subclades may be found at http://www.bio.jhu.edu/Faculty/Beemon/.

Classifying integrations with respect to transcription units. The RefSeq and all mRNA databases for the galGal3.0 genome were downloaded from the UCSC Genome Browser. A transcription unit was defined as a RefSeq gene or an mRNA expressed from a given locus. We then downloaded the coordinates of all miRNAs from miRBASE (46). These three databases were used to create a database of all of the transcripts in the genome, and all redundant entries were removed. The coordinates of all BLAT alignments were then compared with the transcript database to identify the distance of the endogenous proviral chain from the nearest transcription unit. A similar analysis was conducted using the taeGut (zebra finch) genome.

Simulation of random integrations. Ten million random integrations were simulated using a random-number generator. Each number corresponded to a chromosome and a coordinate in the chromosome. The simulations were mapped with respect to transcription units similar to endogenous proviral chains. The probability that the endogenous proviral chains were different from the simulation was calculated using the χ^2 test. A cluster was defined when more than five endogenous proviral integrations mapped within 10^6 bp of one another. A random distribution of the 500 chicken proviral integrations would yield 1 provirus every 2×10^6 bp.

Library preparation of RNA for mRNA-seq analysis and ribosome footprinting analysis. Poly(A)$^+$ mRNA was purified from CEFs using magnetic oligo(dT) beads (NEB) after heating for 2 min at 80°C. Ribosome footprints were prepared from CEFs as previously described (34). Libraries were prepared for sequencing on the Illumina Hi-Seq 2000.

Analysis of mRNA-seq and ribosome footprinting data. Data generated from the sequencing libraries were aligned with the chicken genome index of Bowtie (32) (galGal3.0, UCSC index) using TopHat with default parameters (47). Custom python scripts were used to identify reads that mapped to internal exons, 5′ UTRs, 3′ UTRs, and endogenous viruses. Only unique alignments were used to calculate translational efficiency and coverage of endogenous proviruses.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00344-12/-/DCSupplemental.

Figure S1, TIF file, 2.4 MB.
Figure S2, TIF file, 0.8 MB.
Figure S3, TIF file, 1.9 MB.
Figure S4, TIF file, 2.3 MB.
Figure S5, TIF file, 2.9 MB.
Table S1, DOCX file, 0.1 MB.
Table S2, DOCX file, 0.1 MB.

ACKNOWLEDGMENTS

This work was supported by Public Health Service grants R01CA048746 and RO1CA124596 from the National Cancer Institute to K.L.B.

We thank Nicholas Ingolia, Carnegie Institute of Embryology, for sequencing ribosome footprints and associated RNA-seq on Hi-Seq 2000 and for helpful discussions. We also thank Naomi Rosenberg and Robin Weiss for review of the manuscript.

REFERENCES

1. **Weiss RA.** 2006. The discovery of endogenous retroviruses. Retrovirology **3**:67.
2. **Stoye JP.** 2012. Studies of endogenous retroviruses reveal a continuing evolutionary saga. Nat. Rev. Microbiol. **10**:395–406.
3. **Padian K, Chiappe LM.** 1998. The origin and early evolution of birds. Biol. Rev. **73**:42.
4. **Xu X, You H, Du K, Han F.** 2011. An archaeopteryx-like theropod from China and the origin of Avialae. Nature **475**:465–470.
5. **Dornburg R.** 1995. Reticuloendotheliosis viruses and derived vectors. Gene Ther. **2**:301–310.
6. **Tarlinton R, Meers J, Young P.** 2008. Biology and evolution of the endogenous koala retrovirus. Cell. Mol. Life Sci. **65**:3413–3421.
7. **Blomberg J, et al.** 2011. Phylogeny-directed search for murine leukemia virus-like retroviruses in vertebrate genomes and in patients suffering from myalgic encephalomyelitis/chronic fatigue syndrome and prostate cancer. Adv. Virol. **2011**:341294.
8. **Boeke JD, Stoye JP.** 1997. Retrotransponsons, endogenous retroviruses and the evolution of retroelements, p 343–435. *In* Coffin JM, Hughes SH, Varmus HE (ed), Retroviruses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
9. **Ting C-N, Rosenberg MP, Snow CM, Samuelson LC, Meisler MH.** 1992. Endogenous retroviral sequences are required for tissue-specific expression of human salivary amylase gene. Genes Dev. **6**:1457–1465.
10. **International Chicken Genome Sequencing Consortium.** 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature **432**:695–777.
11. **Dalloul RA, et al.** 2010. Multi-platform next-generation sequencing of the domestic turkey (*Melearis gallopavo*): genome assembly and analysis. PLoS Biol. **8**:e1000475.
12. **Warren WC, et al.** 2010. The genome of a songbird. Nature **464**:757–762.
13. **van Tuinen M, Hedges SB.** 2001. Calibration of avian molecular clocks. Mol. Biol. Evol. **18**:206–213.
14. **van Tuinen M, Dyke GJ.** 2004. Calibration of galliform molecular clocks using multiple fossils and genetic partitions. Mol. Phylogenet. Evol. **30**: 74–86.
15. **Kriegs JO, et al.** 2007. Waves of genomic hitchhikers shed light on the evolution of gamebirds (Aves: Galliformes). BMC Evol. Biol. **7**:190.
16. **Mayer J, Blomberg J, Seal RL.** 2011. A revised nomenclature for transcribed human endogenous retroviral loci. Mob. DNA **2**:7.
17. **Blomberg J, Benachenhou F, Blikstad V, Sperber G, Mayer J.** 2009. Classification and nomenclature of endogenous retroviral sequences (ERVs): problems and recommendations. Gene **448**:115–123.
18. **Borisenko L, Rynditch AV.** 2004. Complete nucleotide sequences of ALV-related endogenous retroviruses available from the draft chicken genome sequence. Folia Biol. (Praha) **50**:136–141.
19. **Sperber GO, Airola T, Jern P, Blomberg J.** 2007. Automated recognition of retroviral sequences in genomic data—RetroTector. Nucleic Acids Res. **35**:4964–4976.
20. **Blikstad V, Benachenhou F, Sperber GO, Blomberg J.** 2008. Evolution of human endogenous retroviral sequences: a conceptual account. Cell. Mol. Life Sci. **65**:3348–3365.
21. **Astrin SM, et al.** 1980. Ten genetic loci in the chicken that contain structural genes for endogenous avian leukosis viruses. Cold Spring Harb. Symp. Quant. Biol. **44**:1105–1109.
22. **Borysenko L, Stepanets V, Rynditch AV.** 2008. Molecular characterization of full-length MLV-related endogenous retrovirus ChiRV1 from the chicken, Gallus gallus. Virology **376**:199–204.
23. **Acloque H, et al.** 2001. Identification of a new gene family specifically expressed in chicken embryonic stem cells and early embryo. Mech. Dev. **103**:79–91.
24. **Reed KJ, Sinclair AH.** 2002. FET-1: a novel W-linked, female specific gene up-regulated in the embryonic chicken ovary. Mech. Dev. **119**:S87–S90.
25. **Carré-Eusèbe D, Coudouel N, Magre S.** 2009. OVEX1, a novel chicken endogenous retrovirus with sex-specific and left-right asymmetrical expression in gonads. Retrovirology **6**:59.
26. **Mangeney M, Heidmann T.** 1998. Tumor cells expressing a retroviral envelope escape immune rejection *in vivo*. Proc. Natl. Acad. Sci. U. S. A. **95**:14920–14925.
27. **Gifford R, Kabat P, Martin J, Lynch C, Tristem M.** 2005. Evolution and distribution of class II-related endogenous retroviruses. J. Virol. **79**: 6478–6486.
28. **Bai J, Payne LN, Skinner MA.** 1995. HPRS-103 (exogenous avian leukosis virus, subgroup J) has an *env* gene related to those of endogenous elements EAV-0 and E51 and an E element found previously only in sarcoma viruses. J. Virol. **69**:779–784.

29. **Gao Y, et al.** 2012. Molecular epidemiology of avian leukosis virus subgroup J in layer flocks in China. J. Clin. Microbiol. **50**:953–960.
30. **Mitchell RS, et al.** 2004. Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. PLoS Biol. **2**:e234.
31. **Wicker T, et al.** 2005. The repetitive landscape of the chicken genome. Genome Res. **15**:126–136.
32. **Langmead B, Trapnell C, Pop M, Salzberg SL.** 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10**:R25.
33. **Brawand D, et al.** 2011. The evolution of gene expression levels in mammalian organs. Nature **478**:343–347.
34. **Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS.** 2009. Genome-wide analysis of *in vivo* translation with nucleotide resolution using ribosome profiling. Science **324**:218–223.
35. **Barker GF, Beemon K.** 1991. Nonsense codons within the Rous sarcoma virus gag gene decrease the stability of unspliced viral RNA. Mol. Cell. Biol. **11**:2760–2768.
36. **Jern P, Sperber GO, Blomberg J.** 2005. Use of endogenous retroviral sequences (ERVs) and structural markers for retroviral phylogenetic inference and taxonomy. Retrovirology **2**:50.
37. **Fadly AM.** 1997. Avian retroviruses. Vet. Clin. North Am. Food Anim. Pract. **13**:71–85.
38. **Hayman MJ.** 1983. Avian acute leukemia viruses. Curr. Top. Microbiol. Immunol. **103**:109–125.
39. **Dimcheff DE, Krishnan M, Mindell DP.** 2001. Evolution and characterization of tetraonine endogenous retrovirus: a new virus related to avian sarcoma and leukosis viruses. J. Virol. **75**:2002–2009.
40. **Dimcheff DE, Drovetski SV, Krishnan M, Mindell DP.** 2000. Cospeciation and horizontal transmission of avian sarcoma and leukosis virus *gag* genes in galliform birds. J. Virol. **74**:3984–3995.
41. **Tollis M, Boissinot S.** 2011. The transposable element profile of the anolis genome: how a lizard can provide insights into the evolution of vertebrate genome size and structure. Mob. Genet. Elements **1**:107–111.
42. **Barr SD, Leipzig J, Shinn P, Ecker JR, Bushman FD.** 2005. Integration targeting by avian sarcoma-leukosis virus and human immunodeficiency virus in the chicken genome. J. Virol. **79**:12035–12044.
43. **Kapranov P, Willingham AT, Gingeras TR.** 2007. Genome-wide transcription and the implications for genomic organization. Nat. Rev. Genet. **8**:413–423.
44. **Gullerova M, Proudfoot NJ.** 2008. Cohesin complex promotes transcriptional termination between convergent genes in *S. pombe*. Cell **132**:983–995.
45. **Kumar S, Nei M, Dudley J, Tamura K.** 2008. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. Brief. Bioinform. **9**:299–306.
46. **Griffith-Jones S, Saini HK, van Dongen S, Enright AJ.** 2008. miRBase: tools for microRNA genomics. Nucleic Acids Res. **36**:D154–D158.
47. **Trapnell C, Pachter L, Salzberg SL.** 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics **25**:1105–1111.
48. **Trapnell C, et al.** 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. **28**:511–515.