Genomic Fossils Calibrate the Long-Term Evolution of Hepadnaviruses

Clément Gilbert*, Cédric Feschotte*

Department of Biology, University of Texas, Arlington, Texas, United States of America

Abstract

Because most extant viruses mutate rapidly and lack a true fossil record, their deep evolution and long-term substitution rates remain poorly understood. In addition to retroviruses, which rely on chromosomal integration for their replication, many other viruses replicate in the nucleus of their host's cells and are therefore prone to endogenization, a process that involves integration of viral DNA into the host’s germline genome followed by long-term vertical inheritance. Such endogenous viruses are highly valuable as they provide a molecular fossil record of past viral invasions, which may be used to decipher the origins and long-term evolutionary characteristics of modern pathogenic viruses. Hepadnaviruses (Hepadnaviridae) are a family of small, partially double-stranded DNA viruses that include hepatitis B viruses. Here we report the discovery of endogenous hepadnaviruses in the genome of the zebra finch. We used a combination of cross-species analysis, orthologous insertions, molecular dating, and phylogenetic analyses to demonstrate that hepadnaviruses infiltrated repeatedly the germline genome of passerine birds. We provide evidence that some of the avian hepadnavirus integration events are at least 19 My old, which reveals a much deeper ancestry of Hepadnaviridae than could be inferred based on the coalescence times of modern hepadnaviruses. Furthermore, the remarkable sequence similarity between endogenous and extant avian hepadnaviruses (up to 75% identity) suggests that long-term substitution rates for these viruses are on the order of \(10^{-8}\) substitutions per site per year, which is a 1,000-fold slower than short-term rates estimated based on the sequences of circulating hepadnaviruses. Together, these results imply a drastic shift in our understanding of the time scale of hepadnavirus evolution, and suggest that the rapid evolutionary dynamics characterizing modern avian hepadnaviruses do not reflect their mode of evolution on a deep time scale.

Introduction

Most viruses are characterized by high substitution rates, which generally prevent reconstruction of their long-term evolutionary history [1]. Consequently, the origins and age of most extant viruses remain elusive [2]. One solution to this conundrum lies in the advent of paleoviology, the study of paleoviruses and the way they have shaped the antiviral genes of their hosts over millions of years [3]. Although viruses lack a true geological fossil record, some have left footprints of their evolution in their hosts’ genome. For example, vertebrate retroviruses are RNA viruses that normally integrate into the genome of their host’s somatic cells as part of their replication cycle. On occasion, these viruses may integrate into the germline genome of their host, and become inactive and vertically inherited over millions of years. Their molecular relics, called endogenous retroviruses, now make up a substantial fraction of vertebrate genomes (~8% in human; [4]).

While retroviruses account for the major fraction of known viral genomic fossils, various other viruses that do not normally integrate into the genome but replicate in the nucleus of the host cell are susceptible to fortuitous chromosomal integration. For example, pararetroviruses (double-stranded DNA) have deposited numerous endogenous copies in the genome of several plant species [5], and singular integration events have been reported for gemini-like viruses (single-stranded DNA) in tobacco [6], and for non-retroviral RNA viruses such as totovirus-like and M2-killer-like viruses in fungi (double-stranded RNA; [7,8]) and flaviviruses in mosquitoes [9,10].

Genomic fossils closely related to modern viral groups are of particular interest as they have the potential to unveil otherwise inaccessible features of the long-term evolution of viruses. A handful of such precious paleoviruses have recently been unearthed from mammalian genomes. Among these, two ancient lentiviruses (RELIK in rabbit [11] and pSIV in primates [12,13]) and one foamy virus (SloEFV in xenarthrans [14]) revealed that the history of these two retroviral genera can be rooted on a deep time scale, challenging earlier views on retroviral evolution based on comparisons of extant viral genomes. Likewise, the recent discovery of multiple endogenous bornaviruses and filoviruses in diverse mammals showed that these single-stranded RNA viruses were able to infiltrate repeatedly the germline of distant mammalian species over at least the past 40 My [15–17].

Hepadnaviridae (including hepatitis B viruses [HBVs]) are compact (~3,000 bp), partially double-stranded circular DNA viruses infecting various mammal and bird species and responsible for ~600,000
Endogenous Hepadviruses in the Zebra Finch Genome

TBLASTN searches using the duck HBV (DHBV) proteins on all available genomes in GenBank yielded 15 hepadnavirus-like fragments (collectively called endogenous zebra finch HBVs [eZHBVs]). These sequences are interspersed into ten different chromosomes of the zebra finch (Taeniopygia guttata, Estrildidae) and show between 55% and 75% nucleotide similarity to the DHBV genome (Figure 1; Table 1; Dataset S1). Most of these fragments contain one or more mutations compromising their coding capacity, which suggests that they have evolved under no functional constraint since integration. Together, the 15 eZHBV segments cover ~70% of the DHBV genome, which is structurally representative of all hepadviruses [20] (Figure 1). eZHBVs tend to map within two loosely defined regions of DHBV, one encompassing the core and polymerase N-terminal domains (eZHBVc–eZHBVi; group 1), and one overlapping with the preS/S and polymerase C-terminal domains (eZHBVj–eZHBVn; group 2). In addition, two eZHBVs (eZHBVa and eZHBVb) map to other regions of the core domain (Figure 1). eZHBVi and eZHBVi* (both located on Chromosome 20) map to the same region of the DHBV genome and are highly similar (97% over 537 bp). Similar levels of identity are observed between their flanking genomic regions: 96.7% identity over 637 bp in the 5′ flanking region and 97% identity over 534 bp in the 3′ flanking region. These observations suggest that one insertion most likely derives from the other through intrachromosomal duplication of a genomic fragment including the initial eZHBV insertion along with its flanking regions.

In order to assess the phylogenetic relationship among eZHBVs and hepadnaviruses, we conducted phylogenetic analyses of amino acid alignments including extant hepadnaviruses and group 1 (106 amino acids) and group 2 (293 amino acids) eZHBVs. The results show that in both phylogenies (Figure 2A and 2B) hepadnaviruses can be divided into two clusters, one grouping eZHBVs and extant avian hepadnaviruses and the other including all mammalian hepadnaviruses. Within the former cluster, eZHBVs are consistently more distant from extant avian hepadnaviruses than these are from each other. While group 1 eZHBVs form a monophyletic group (Figure 2A), there is no statistically supported clustering of group 2 eZHBVs with each other (Figure 2B). The only exception is the close clustering of eZHBVi and eZHBVi*, which likely reflects their relatively recent origin by duplication rather than as independent insertions (see above).
How Old Are eZHBVs?

A first minimal estimate of the age of eZHBVs can be derived indirectly from the time at which the duplication yielding eZHBV1 and eZHBV1* occurred, which must postdate the chromosomal integration of the ancestral eZHBV1 element. The distance between these duplicates is 0.03 (Table 2). To our knowledge, the most comprehensive estimate of neutral substitution rates available for birds, calculated based on a comparison of multiple intron sequences between chicken and turkey, was found to range between 2 × 10^{-9} and 3.9 × 10^{-9} substitutions per site per year (subs/site/year) [21], values similar to the range of those estimated for mammals (2.2 × 10^{-9} to 4.5 × 10^{-9} subs/site/year; [22,23]). The avian rates are based on a fossil calibration of the split between Anatidae and Anhimidae at 55 My [21,24,25]. Dividing half of the distance between eZHBV1 duplicates (0.015) by the bird neutral substitution rates yields a duplication time ranging between 3.8 My (with 3.9 × 10^{-9} subs/site/year) and 7.5 My (with 2 × 10^{-9} subs/site/year). The timing of this duplication provides a minimal estimate for the integration of the ancestral eZHBV1 fragment.

A more direct way to estimate the age of eZHBVs is to use a phylogenetic approach, reasoning that if an insertion is shared by two species at the same (orthologous) locus, the integration event must be at least as old as the last common ancestor of the two species. It is important to note that the analysis of a large number of chromosomal integrants of HBV in somatic mammalian cells has revealed no preference for insertion in a specific sequence motif (e.g., [19,26]). Thus, the possibility that two identical viral fragments would integrate at the exact same genomic position (i.e., between the same two nucleotides) independently in multiple species is extremely unlikely. Using PCR primers designed on the genomic regions flanking three eZHBVs, we were able to amplify two orthologous insertions (eZHBV1a and eZHBV1b) in three other species of estrildid finches (black throated finch [Poephila cincta], scaly breasted munia [Lonchura punctulata], and gouldian finch [Chloebia gouldiae]) and in the dark-eyed junco (Junco hyemalis), a non-estrid passerine bird belonging to the Emberizidae family (Figure 3). We also obtained a positive PCR product for eZHBVj in the three estrildid finches, and were able to amplify the empty site orthologous to eZHBVj in the olive sunbird (Cyaniornis olivaceus, Nectariniidae) (Figures 3 and 4). The identity of all the eZHBV fragments amplified by PCR was confirmed by DNA sequencing (Datasets S4, S5, S6). This revealed that each orthologous eZHBV is present at the same chromosomal position in all species where it could be amplified. Furthermore, in all three cases, the phylogenetic relationships between orthologous eZHBVs reflect the phylogenetic relationships of the bird species (Figure 3). Together, these data strongly suggest that each of these three insertions descend from an ancestral integration event that occurred prior to the split of the different bird species.

The most recent molecular phylogenetic analyses divide finches and their allies into two major monophyletic clades, one consisting of African and Australasian estrildid finches and weavers, and the other grouping American emberizid sparrows (including the dark-eyed junco) together with fringillid finches and Old World sparrows [27]. Within Estrildidae, the gouldian finch is sister to a clade grouping the scally breasted munia and finches of the genera Poephila (black throated finch) and Taenopygia (zebra finch) (Figures 3A and 4; [28]). The congruence between these relationships and the phylogenies of orthologous eZHBVs and eZHBV1 (Figure 3) indicates that the two eZHBVs result from two independent germline integration events of hepadnavirus-like sequences in a common ancestor of Estrildidae and Emberizidae, and that eZHBV was inserted after the divergence of the

Table 1. Characteristics of the endogenous HBV fragments found in the zebra finch genome (eZHBVs).

| eZHBV | Position in DHBV | Position in Finch Genome | Length (bp) | NS Mutations | Similarity to DHBV | 5’ Gene in Finch | 3’ Gene in Finch |
|-------|------------------|--------------------------|------------|--------------|-------------------|----------------|----------------|
| a     | 2773             | 27                       | 5,223,682  | 5,223,949    | 268               | 0/1            | 58.9% Scn3b (690 bp) |
| b     | 85               | 213                      | 50,694,672 | 50,694,800   | 129               | 1/0            | 63.8% Fry (2.2 kb) |
| c     | 269              | 781                      | 3,002,393  | 3,002,898    | 507               | 1/0            | 61.5% ATP2B2 (10 kb) |
| d     | 428              | 670                      | 864,217    | 864,459      | 246               | 1/0            | 59.3% Trim33 (1.3 kb) |
| e     | 428              | 556                      | 2,324,466  | 2,324,594    | 129               | 0/0            | 62% cdh23 (5.5 kb) |
| f     | 476              | 757                      | 68,942,470 | 68,942,804   | 335               | 4/1            | 62.2% LMO3 (13 kb) |
| g     | 476              | 622                      | 68,944,006 | 68,944,152   | 147               | 1/1            | 70.5% LMO3 (15 kb) |
| h     | 530              | 673                      | 25,658,020 | 25,658,163   | 144               | 0/0            | 69.4% — |
| i     | 590              | 790                      | 8,647,882  | 8,648,070    | 185               | 1/0            | 60.8% — |
| j     | 1310             | 2521                     | 8,648,446  | 8,649,654    | 1,209             | 1/0            | 65.5% — |
| k     | 1466             | 2512                     | 50,694,887 | 50,695,493   | 1,057             | 4/5            | 54.8% Fry (2.5 kb) |
| l     | 1388             | 1936                     | 1,111,838  | 1,112,362    | 537               | 2/0            | 72.3% — |
| l*    | 1388             | 1936                     | 1,119,136  | 1,120,278    | 537               | 1/0            | 72.7% — |
| m     | 1334             | 1780                     | 33,336,955 | 33,337,390   | 436               | 2/1            | 75% Dhx32 (1 kb) |
| n     | 1562             | 1789                     | 5,856,561  | 5,856,833    | 273               | 0/0            | 72.8% — |

eZHBV(1a–eZHBVe, eZHBVk, and eZHBVm are in an intron; a dash indicates that there is no gene within 50 kb 5
upstream and 50 kb 3' downstream. Positions in the zebra finch genome are from the UCSC Genome Bioinformatics browser, based on the assembly WUGSC 3.2.4/taeGut1 (July 2008). See Figure 1 for the mapping of eZHBVs on DHBV. The 15 eZHBV sequences are provided in Dataset S1.

*aGenBank accession number AY494851, isolated from a puna teal.

*bNonsense, stopp, frameshift.

cDistance to the nearest gene or exon.

doi:10.1371/journal.pbio.1000495.t001
Nectariniidae lineage. The divergence time between Estrildidae and Emberizidae has been estimated at 25 My based on relaxed molecular clock analyses of \textit{rag1} and \textit{rag2} nuclear genes using a paleobiogeographical calibration of 82 My for the split between Acanthisittidae and other passerine birds [29,30]. The same analysis yielded an age of 35 My for the most recent common ancestor of Nectariniidae and Estrildidae. These dates would place the origin of \textit{eZHBVl} prior to 25 My, and that of \textit{eZHBVa} between 35 and 25 My ago.

Our last estimate of the age of \textit{eZHBVs} relies on the level of sequence divergence between orthologous \textit{eZHBV} sequences. The corrected distances inferred for orthologous \textit{eZHBV} (222 bp) and
### Table 2. Corrected distances between avian hepadnaviruses calculated on the region corresponding to group 2 eZHBVs (see Figure 1).

|                | eZHBVI | eZHBVI* | eZHBVj | eZHBVk | eZHBVn | eZHBVm | Crane | Ross’s Goose | Heron | Snow Goose | White Stork |
|----------------|--------|---------|--------|--------|--------|--------|-------|--------------|-------|------------|-------------|
| eZHBVI*        | 0.03   |         |        |        |        |        |       |              |       |            |             |
| eZHBVj         | 0.89   | 0.87    |        |        |        |        |       |              |       |            |             |
| eZHBVk         | 1.99   | 1.77    | 1.51   |        |        |        |       |              |       |            |             |
| eZHBVn         | 0.55   | 0.58    | 0.81   | 2.43   |        |        |       |              |       |            |             |
| eZHBVm         | 0.48   | 0.49    | 0.35   | 1.43   | 0.84   |        |       |              |       |            |             |
| Crane          | 0.49   | 0.46    | 0.81   | 1.53   | 0.49   | 0.42   |       |              |       |            |             |
| Ross’s goose   | 0.44   | 0.42    | 0.80   | 1.51   | 0.48   | 0.41   | 0.17  |              |       |            |             |
| Heron          | 0.53   | 0.51    | 0.84   | 1.60   | 0.52   | 0.43   | 0.23  | 0.23         |       |            |             |
| Snow goose     | 0.47   | 0.45    | 0.76   | 1.46   | 0.49   | 0.44   | 0.14  | 0.17         | 0.22  |            |             |
| White stork    | 0.53   | 0.50    | 0.83   | 1.59   | 0.51   | 0.45   | 0.23  | 0.23         | 0.11  | 0.22       |             |
| Duck           | 0.51   | 0.49    | 0.77   | 1.43   | 0.48   | 0.43   | 0.14  | 0.18         | 0.23  | 0.23       | 0.09        | 0.26        |

Distances were calculated using the TVM+G model under maximum likelihood settings (see Materials and Methods) in PAUP [65]. Values in bold correspond to distances between eZHBVs that are larger than 0.54. This distance threshold corresponds to twice the average distance between extant avian hepadnaviruses (2 × 0.19 = 0.38) plus 0.16, which corresponds to a conservatively high estimate of the distance accumulated at the bird genome rate after integration (see Materials and Methods). eZHBVI and eZHBV* derive from post-insertional duplication (see Results for details).

doi:10.1371/journal.pbio.1000495.002

eZHBV (238 bp) are 0.15 and 0.16 respectively between the zebra finch and the dark-eyed junco (see Materials and Methods). Selection analyses on these two fragments did not reveal any sign of positive or purifying selection (see Materials and Methods), suggesting that eZHBVa and eZHBVI have evolved under no functional constraint since their chromosomal integration in the common ancestor of these two birds, thereby accumulating substitutions at the neutral rate of these species. Applying the above-mentioned bird neutral substitution rates to half of the zebra finch/junco distances for eZHBVa and eZHBVI yielded integration times ranging between 40 My (with the eZHBVI distance of 0.08 and a rate of 2 × 10⁻⁹ subs/site/year) and 19.2 My (with the eZHBVI distance of 0.075 and a rate of 3.9 × 10⁻⁹ subs/site/year).

While our estimates of the age of eZHBVs are based on two different calibration points located at distant phylogenetic positions within the avian tree (55 My for the split between Anatidae and Anhimidae, or 82 My for the split between Acantitisidae and other passerine birds), both approaches yield dates that largely overlap (40–19.2 My and 33–25 My). This suggests that eZHBVa and eZHBVI are at least 19 My old (and may be as much as 40 My old), which implies that the origin of avian hepadnaviruses as a whole (including extant and extinct viral populations) is expected to correspond to the sum of (i) the distance accumulated during HBV infection [35], these particles have the capacity to integrate HBV genomic fragments are commonly observed in liver cells of individuals persistently infected, where they tend to be associated with hepatocarcinoma [19]. In addition, while hepadnavirus replication is thought to occur mainly in hepatocytes, its tropism may extend to other tissue and cell types, including germ cells. For example, avian hepadnavirus replication has been shown to occur in the yolk sac of developing duck embryos [34]. Typically, large quantities of viral particles circulate in the blood during HBV infection [35]. These particles have the capacity to...
Figure 3. Illustration and phylogenetic trees of orthologous eZHBVa, eZHBVl, and eZHBVj. The primers (S3n3b-F/R) used to amplify eZHBVa (A) are anchored in exons 3 and 4 of a predicted gene homologous to the human SCN3B gene (blue), on zebra finch Chromosome 24. One of the primers (8718F) used to amplify eZHBVj (B) is located in the region flanking the insertion in 3′ on zebra finch Chromosome 2, while the other (8718F) is anchored in eZHBVl. One of the primers (1978F) used to amplify eZHBVj (C) is located in the region flanking the insertion in 5′ on zebra finch Chromosome 20, while the other (Hfr1) is anchored in eZHBVl. Each orthologous eZHBV tree reflects the bird tree, derived from (28,30) and illustrated in (A). The congruence between orthologous eZHBV trees and the bird tree is in each case consistent with one event of eZHBV integration in a common ancestor of the different birds where the insertion was found. The eZHBVs are rooted using circulating avian hepadnaviruses as an outgroup. Numbers on branches correspond to bootstrap values greater than 70 and posterior probabilities greater than 0.9. The precise position and sequence of the PCR primers for each locus is given in Datasets S4, S5, S6. The chromosomal coordinates are derived from the July 2008 assembly of the zebra finch genome (WUGSC 3.2/taeGut1).

doi:10.1371/journal.pbio.1000495.g003

Fossil Heпадnaviruses Uncovered in Bird Genomes

 secara erat mengikat ke berbagai jenis sel, dan ada bukti bahwa adanya HBV DNA komutatif di dalam mitosis dan ovaries as well as the chromosomal integration of HBV in spermatozoa [36–38]. Based on these data, infiltration of the germline genome by hepadnaviruses followed by long-term vertical inheritance appears largely plausible. Thus, it is likely that other endogenous hepadnaviruses await discovery in other birds and perhaps also in mammalian genomes.

The precise mechanisms underlying the chromosomal integration of HBV remain unclear [19]. One model supported by experimental evidence posits that viral linear double-stranded DNA resulting from aberrant replication can be integrated during repair of double strand breaks via non-homologous end joining [39]. As the 3′ extremity of eZHBVj (position 2521) and eZHBVl (position 2512) map to a region of the DHBV genome that corresponds to the predicted end of a typical linear HBV precursor [40], the structure of these two fragments is potentially consistent with integration via non-homologous end joining. We also note that the extremities of several other fragments map to fairly narrow regions of the viral genome (e.g., same 5′ position for eZHBVd and eZHBVc; Figure 1), which may reflect the presence of breakpoint hotspots in the viral genomes that gave rise to eZHBVs. Finally, while the zebra finch genome contains several families of long terminal repeat (LTR) and non-LTR retrotransposons [41] whose enzymatic machinery could have potentially promoted the chromosomal integration of eZHBVs, none of the insertions examined were terminated by a poly-A tail or flanked by direct repeats, as would be expected if they had occurred through retrotransposition [15].

An intriguing question is whether the multiple eZHBVs result from endogenization events that took place during a short period of time or whether they were assimilated at widely different times over (at least) the past 19 My. Hepadnaviruses do not encode an integrase, and chromosomal integrants generally correspond to truncated genomes (as observed here). Thus, unlike retroviruses, integrated HBV fragments cannot in principle replicate further through intragenomic transposition or reinsertion, and as such they can be considered essentially “dead on arrival.” With this in mind, we contend that eZHBVs are likely to result from multiple independent episodes of germline infiltrations that took place on a deep time scale, possibly spanning several millions of years, and involving distantly related hepadnaviruses. This inference is supported by the large distances observed between eZHBVs (Tables 2 and 3). Specifically, all pairwise distances involving eZHBVi and those between eZHBVj, eZHBVk, and eZHBVn are more than 2-fold higher than the average distance separating extant avian HBVs, even when subtracting an approximate distance accumulated at the bird genome rate since integration (distances in bold in Tables 2 and 3). Together with the long branches leading to eZHBVs in the hepadnavirus tree (Figure 2), these data strongly suggest that diverse hepadnaviruses (at least five based on the distance threshold described above) have been circulating in birds for several million years. More specifically, we believe that the large inter-eZHBV distances likely reflect the fact that eZHBVs’ stem from viruses that were already deeply divergent at the time of integration, and/or that eZHBVs were integrated at time points separated by several million years over at least 19 My. A third, non-mutually exclusive explanation for these large distances is that the evolution of the hepadnavirus genome may be subject to strong mutational saturation (see also below). Considering that these viruses have crossed species boundaries repeatedly over the past 6,000 y [20,31,42], we speculate that a wide range of bird species may have been, and may still be, infected by hepadnaviruses. It would be interesting to explore whether hepadnaviruses are still circulating in extant estrildid finches such as the zebra finch. Such a discovery would provide a powerful system to study the virus and its potential association with hepatocarcinoma in a model bird species with a complete genome sequence [41].

Various calculations of HBV substitution rates based on comparison of extant viruses have produced broadly similar estimates, ranging from 7.72 × 10^-4 to 7.9 × 10^-5 subs/site/year [31,43–47]. Surprisingly, we infer long-term substitution rates that are more than three orders of magnitude slower than these short-term rates. It is important to note that while eZHBVs evolved at the bird genome rate since their integration, this cannot explain the slowdown in long-term rates inferred in this study as the distance accumulated at the bird rate (A in Figure 4) was removed from our calculation of long-term hepadnavirus rates. Our estimates (2.15 × 10^-8 to 6.8 × 10^-8 subs/site/year) therefore represent a range of rates under which avian hepadnaviruses have evolved from the time just preceding the integration of eZHBVa and eZHBVl in the bird genome (~19 My ago) to the time at which circulating avian hepadnavirus genomes were sequenced (the last two decades).

Gibbs et al. [48] recently suggested that viral evolutionary rates may vary dramatically depending on the time scale on which they are measured. The main line of evidence supporting this view was that rates inferred from serially or heterochronously sampled sequences are invariably more than two orders of magnitude higher than those calculated when assuming viruses have co-diverged with, and are therefore as old as, their hosts. In most cases, however, the hypothesis of host/virus co-divergence is only indirectly supported by the seemingly strong host specificity of the virus, and/or the apparent topological congruence (often not formally tested) between host and virus phylogenies. A major pitfall in this reasoning is that processes other than co-divergence may explain congruent phylogenies between hosts and viruses [49–51]. Given the potential caveats associated with the hypothesis of host/virus co-divergence, it is important to emphasize that our results do not rely on this assumption. Rather, they are based on a direct measure of the distance separating extant hepadnaviruses from extinct ones that are at least 19 My old.

How can we explain the apparent major disparity between short- and long-term substitution rates of hepadnaviruses? The rate of nucleotide substitution in any system depends on the
background mutation rate, the rate of replication, and the rate of fixation. Hepadnaviruses replicate their genome via an RNA intermediate using a reverse transcriptase (RT). While to our knowledge there is no precise measure of the fidelity of the hepadnavirus RT, this enzyme lacks a proofreading activity and is known to be highly error prone in all retroviruses and other retroelements for which an error rate has been estimated [52,53]. Up to 20-fold variations in RT error rates have been reported between different families of retroviruses [52]. It is therefore conceivable that variations in the fidelity of the enzyme (i.e., background mutation rate) over time might explain some of the difference between short- and long-term hepadnavirus substitution rates. However, slow long-term substitution rates similar to those reported here have been inferred for mammalian foamy viruses.
(1.7×10^{-8} \text{ subs/site/year}) and human T cell lymphotropic virus type II (1.091×10^{-7} to 7.118×10^{-7} subs/site/year), two mammalian retroviruses that yet replicate via a highly error-prone RT [54,55]. In those cases, it is thought that both viruses evolve slowly because they are non-pathogenic and replicate mainly as integrated proviruses, using the high-fidelity DNA polymerases of their hosts [56,57]. These two examples therefore suggest that even in the presence of a high background mutation rate, viruses can evolve slowly if their replication rate is reduced. By analogy, it could be that hepadnaviruses have been characterized by low levels of pathogenicity and by low rates of replication for most of their evolutionary history. In this context, the high substitution rates and epidemiological dynamics currently associated with circulating hepadnaviruses might reflect recent drastic alterations in the biology of these viruses and of the selective pressures acting on them.

Another major process that may be responsible for the time dependency of substitution rates suggested by this study is purifying selection, as proposed for cellular organisms (e.g., [58–60]; see [61] for discussion). About 60% of the HBV genome codes for at least two overlapping open reading frames and therefore contains very few synonymous sites. Consistent with this, it was shown that nonoverlapping regions of the HBV genome evolve faster than overlapping regions [31,62]. This tightly constrained genetic organization, combined with the intrinsically low fidelity of the RT, suggests that the effect of purifying selection on long-term rates is more pronounced for hepadnaviruses than for other viruses and for cellular organisms. Lastly, the high background mutation rates of hepadnaviruses may also result in strong mutational saturation (homoplasy and back mutations), which could also explain part of the difference between short- and long-term hepadnavirus substitution rates (see also above). While it is possible that saturation may in part hinder our ability to accurately infer the long-term hepadnavirus substitution rates, we believe that this phenomenon alone cannot explain the 1,000-fold difference between short- and long-term substitution rates. Because our knowledge on the deep evolution of extant viruses remains fragmentary and because many factors may influence substitution rates and their variation over time [1,63], it would be necessary to revisit these questions when more fossil and modern hepadnavirus sequences become available.

**Materials and Methods**

**PCR and Sequencing**

In order to screen for the presence or absence of orthologous eZHBVs in several species of passerine birds (Table S1), we designed PCR primers on the flanking regions of three insertions. The sequences produced using these primers were aligned and are provided, together with the sequence of the primers, in Datasets S5 (eZHBVl), S6 (eZHBVj), and S7 (eZHBVa). For eZHBVl, we used a forward primer (1978F) anchored in the 5′ flanking region of three insertions (eZHBVa). For eZHBVa, we used a forward primer (Scn3b-F) anchored in the 3′ flanking region of the insertion. For eZHBVl, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVj, we used a forward primer (Scn3b-F) anchored in the 3′ flanking region of the insertion. For eZHBVa, we used a forward primer (Scn3b-F) anchored in the 3′ flanking region of the insertion. For eZHBVh, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVj, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVh, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVj, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVh, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVj, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVh, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVj, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVh, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVj, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVh, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVj, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVh, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVj, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVh, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVj, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVh, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVj, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVh, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVj, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVh, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVj, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVh, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVj, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVh, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVj, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVh, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257.
up to 25 μl. PCR products were directly sequenced on an ABI 3130XL sequencer (Applied Biosystems). All sequences produced in this study were submitted to GenBank (accession numbers HQ116564–HQ116583).

Analyses of Selection
Analyses of selection were carried out on alignments of each set of orthologous insertions amplified in the various passerine birds (cZHBVl, cZHBVv, and cZHBVv; provided in Datasets S4, S5, and S6, respectively) using HyPhy [64]. We used the trees corresponding to each alignment as inferred in Figure 3. The nucleotide substitution model accomplishing the most accurate fit to the data was determined using the NucModelCompare.bf procedure: HKY85 for each of the three alignments. The MG94xHKY85_3x4 codon substitution model was then fitted to each alignment with local parameters and partition-based equilibrium frequencies. This yielded a global ω (non-synonymous substitutions/synonymous substitutions) ratio of 0.98 (confidence interval: 0.64–2.32, 1.32), 0.66 (confidence interval: 0.44, 0.88), and 0.93 (confidence interval: 0.62, 1.24) for cZHBVl, cZHBVv, and cZHBVv respectively. Using a likelihood ratio test, the likelihood function states for each alignment were then compared to likelihood function states obtained using the same model/alignment/tree but enforcing ω = 1 (neutral evolution). This revealed no significant difference (p = 0.95 for cZHBVl, 0.16 for cZHBVv, and 0.81 for cZHBVv), suggesting that cZHBVl, cZHBVv, and cZHBVv are evolving neutrally. We further tested this by re-optimizing the likelihood function with local parameters (where each branch of the tree has its own parameters) and comparing the likelihood function state obtained when the non-synonymous substitution rate and the synonymous substitution rate can have their own value on each branch with the likelihood function state obtained when the non-synonymous substitution rate is forced to be equal to the synonymous substitution rate on each branch. Again, the likelihood ratio test revealed no significant difference (p = 0.61 for cZHBVl, 0.29 for cZHBVv, and 0.85 for cZHBVv), suggesting neutral evolution in all branches.

Distances between Avian Hepadnaviruses
All distances were calculated under maximum likelihood settings in PAUP 4.0 [63], using models of nucleotide substitution chosen based on the Akaike Information Criterion in jModeltest [66]: TPM2uf+G for group 1 cZHBVs, TVM+G for group 2 cZHBVs and for the distance between cZHBVv and extant avian hepadnaviruses, TPM1 for the distances between passerine cZHBVv orthologs, and HKY for the distance between passerine cZHBVv orthologs.

In order to estimate whether cZHBVs result from multiple integrations of a few very similar viral strains during a narrow time frame or whether more divergent strains were endogenized at widely different times during the last 19 My, we compared inter-cZHBV distances to the average distances between extant hepadnaviruses. In this context, it is important to keep in mind that each pairwise inter-cZHBV distance as we observe them today results from (i) the distance accumulated at the viral rate separating the endogenization of each two sequences being compared (corresponding to B+C or EF+F if cZHBVv and cZHBVv are compared, for example; Figure 4) and (ii) the distance accumulated on each sequence at the bird neutral rate after endogenization (2×A in Figure 4). Several inter-cZHBV distances are more than 2-fold higher than the average distances between extant hepadnaviruses, i.e., more than 2×0.27 = 0.54 for the region corresponding to group 1 cZHBVs, and more than 2×0.19 = 0.38 for the region corresponding to group 2 cZHBVs (Tables 2 and 3). Notably, most of these high-inter-cZHBV distances remain more than 2-fold higher than distances between extant hepadnaviruses even when subtracting a 0.16 distance, which corresponds to a conservatively high estimate of the distance accumulated at the bird genome rate assuming the two cZHBVs being compared were both integrated 19 My ago.

Phylogenetic Analyses
Sequences were aligned by hand using BioEdit 7.0.5.3 [67], and ambiguous regions were removed. Bayesian and maximum likelihood phylogenetic analyses were carried out using MrBayes 3.1.2 [68] and PHYML 3.0 [69], respectively. Nucleotide and amino acid substitution models were chosen based on the Akaike Information Criterion in jModelTest 0.1 [66], MrModeltest 2.3 [70], and ProtTest 2.4 [71]. cZHBVs were aligned at the amino acid level with representative members of extant avian and mammalian hepadnaviruses and analyzed using the rRT (group 1 cZHBVs) and LG+G (group 2 cZHBVs) models in PHYML and with a prior setting allowing model jumping between fixed-rate amino acid models in MrBayes. cZHBVv, cZHBVv, and cZHBVv orthologs were analyzed with the TPM2uf+G, TPM2uf+G, and TIM3+G models of nucleotide substitution, respectively, in PHYML and with the GTR+G, HKY+G, and GTR+G models, respectively, in MrBayes. In order to verify the identity of the bird specimens included in this study, we also analyzed an alignment of a fragment of NADH2 nucleotide sequence produced in this study, as well as GenBank NADH2 sequences available for these species and for representatives of the families Paridae, Corvidae, Pycnonotidae, Turdidae, and Phasianidae (Figure S1). This alignment was analyzed with the TPM2uf+G model in PHYML and with the HKY+G model in MrBayes. For maximum likelihood analyses, the robustness of the branches was evaluated by non-parametric bootstrap analyses involving 1,000 pseudoreplicates of the original matrix. Bayesian analyses were run for at least one million generations, or until the standard deviation of split frequencies between the two parallel runs dropped below 0.01. Then, 25% of the sampled trees were discarded before summarizing the trees. The sequences used for the phylogenetic analyses are provided in Datasets S2, S3, S4, S5, S6, S7.

Supporting Information
Dataset S1  FASTA file containing the 15 cZHBVs found in the July 2008 assembly of the zebra finch genome (see also Table 1).
Found at: doi:10.1371/journal.pbio.1000495.s001 (0.01 MB DOC)

Dataset S2  Amino acid alignment (in FASTA format) of group 1 cZHBVs (see Figure 1) and representatives of known extant hepadnaviruses. Ambiguous regions, stop codons, and frameshifts were removed. The names of the sequences include the GenBank accession numbers.
Found at: doi:10.1371/journal.pbio.1000495.s002 (0.00 MB DOC)

Dataset S3  Amino acid alignment (in FASTA format) of group 2 cZHBVs (see Figure 1) and representatives of known extant hepadnaviruses. Ambiguous regions, stop codons, and frameshifts were removed. The names of the sequences include the GenBank accession numbers.
Dataset S4 Alignment of orthologous eZHBVI and 5’ flanking region (in FASTA format) sequenced in various passerine birds. The 5’ end of eZHBVI corresponds to position 108 of the T. guttata sequence. The alignment includes the sequence of the primer 1978F, located in the 5’ flanking region of eZHBVI, and hfr1, located within eZHBVI.

Dataset S5 Alignment of orthologous eZHBVs and 5’ and 3’ flanking regions (in FASTA format) sequenced in various passerine birds. The 5’ and 3’ ends of eZHBVs correspond to positions 769 and 1012, respectively, of the T. guttata sequence. Positions 504–805 of the J. hyenalis sequence correspond to an endogenous retrovirus solo LTR (closely related to the zebra finch TgulERV99_LTR22g element; [41]) inserted within the region orthologous to eZHBV. The solo LTR is flanked by a 6-bp target site duplication (GACCTT). The alignment includes the sequence of the primer Scn3b-R, located in the 5’ flanking region of eZHBVs, which corresponds to the fourth exon of a predicted gene homologous to human SCN3B, and that of the primer Scn3b-F, located in the 3’ flanking region of eZHBVs, which corresponds to the third exon of the predicted scn3b gene. Positions 1–59 and 1013–1076 of the T. guttata sequence correspond respectively to the partial sequence of the fourth and third exon of the predicted scn3b gene.

Dataset S6 Alignment of orthologous eZHBVj and 3’ flanking region (in FASTA format) sequenced in various passerine birds. The 3’ end of eZHBVj corresponds to position 1209 of the T. guttata sequence. The alignment includes the sequence of the primer 8718F, located within eZHBVj, and 8718R, located in the 3’ flanking region of eZHBVj.

Dataset S7 Alignment (in FASTA format) of the NADH2 partial sequences used to construct the tree in Figure S1.

Figure S1 Phylogenetic tree of NADH2 sequences. Numbers on branches correspond to bootstrap values and posterior probabilities. For most species, there is strong support grouping the sequence produced in this study and a NADH2 sequence of the same species available in GenBank, confirming the identification of the specimens from which the tissues used in this study come. The absence of support for the grouping of our P. cincta and that found in GenBank is due to the fact that the GenBank sequence is partial (Dataset S7). Phylogenetic analysis of a reduced alignment including only the NADH2 portion corresponding to the GenBank P. cincta sequence yields strong support for the grouping of the sequence obtained in this study with that in GenBank (bootstrap = 99, posterior probability = 1; data not shown). There is no NADH2 sequence available for L. punctulata in GenBank. While there is no support for the precise position of our L. punctulata sequence, we note that it tends to group with that of a congeneric species (L. cattalica).

Table S1 Tissue samples used in this study. All Estrildidae species were provided by the University of Washington Burke Museum. The C. olivacea DNA was provided by Drs. Claire Loiseau and Ravinder Sehgal (San Francisco State University). The J. hyenalis tissue was sampled from a dead specimen found in CG’s backyard in Arlington, Texas.

Acknowledgments
We thank Mélanie Debais-Thibaud, Jean-Baptiste Ledoux, Beth Shapiro, and members of the Feschotte lab for critical comments and useful suggestions during the preparation of the manuscript. We are grateful to Sharon Bikis and the University of Washington Burke Museum, David Kabekli (Rhodes College), and Claire Loiseau and Ravinder Sehgal (San Francisco State University) for providing tissue samples. We acknowledge the Genome Sequencing Center at the Washington University in St. Louis School of Medicine for producing the T. guttata sequence data used in this study.

Author Contributions
The author(s) have made the following declarations about their contributions: Conceived and designed the experiments: CG, CF. Performed the experiments: CG. Analyzed the data: CG, CF. Wrote the paper: CG, CF.
33. Nowak MA, Bonhoeffer S, Hill AM, Boehme R, Thomas HC, et al. (1996) Viral evolution of hepatitis B virus over 25 years. J Virol 80: 10307–10314.
34. Fares MA, Holmes EC (2002) A revised evolutionary history of hepatitis B virus (HBV). J Mol Evol 54: 807–814.
35. Hannoun C, Horal P, Lindl M (2000) Long-term mutation rates in the hepatitis B virus genome. J Gen Virol 81: 75–83.
36. Gibbs AJ, Fargette D, Garcia-Arenal F, Gibbs MJ (2010) Time—the emerging dimension of plant virus studies. J Gen Virol 91: 13–22.
37. Charleston MA, Robertson DL (2002) Preferential host-switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. Syst Biol 51: 528–533.
38. de Vries DM, Giraudeau T, Shykoff JA (2007) When can host shifts produce congeneric host and parasite phylogeny? A simulation approach. J Evol Biol 20: 1428–1438.
39. Ramden C, Holmes EC, Charleston MA (2009) Hantavirus evolution in relation to its rodent and invertebrate hosts: no evidence for codivergence. Mol Biol Evol 26: 143–153.
40. Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. Genetics 148: 1667–1686.
41. Srivastava KA, Cheslock SR, Zhai Wei, Hu WS, Pathak V (2003) Retroviral mutation rates and reverse transcriptase fidelity. Front Biosci 8: d117–d134.
42. Lemey P, Pybus OG, Van Dooren S, Vandamme AM (2005) A Bayesian statistical analysis of human T-cell lymphotropic virus evolutionary rates. Infect Genet Evol 5: 291–298.
43. Suzuki WM, Saliem M, Shammugam V, Gao F, Cong M, et al. (2005) Ancient co-speciation of simian foamy viruses and primates. Nature 434: 376–380.
44. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, et al. (2002) Fossil Hepadnaviruses Uncovered in Bird Genomes. PLoS Biology | www.plosbiology.org 12 September 2010 | Volume 8 | Issue 9 | e1000495
45. Axelsson E, Smith NG, Sundstro¨m H, Berlin S, Ellegren H (2004) Male-biased mutation rate and divergence in autosomal, z-linked and w-linked introns of chicken and turkey. Mol Biol Evol 21: 1330–1347.
46. Waddell PJ, Cao Y, Hasegawa M, Minndell DP (1999) Assessing the Cretaceous suprerglacial divergence times among birds and placental mammals by using whole mitochondrial protein sequences and an extended statistical framework. Syst Biol 48: 119–137.
47. Matsukura K, Tokino T (1990) Integration of hepatitis B virus DNA and its implications for hepatocarcinogenesis. Mol Biol Med 7: 243–260.
48. Van der Meij MAA, de Bakker MA, Bout RG (2005) Phylogenetic relationships of finches and allies based on nuclear and mitochondrial DNA. Mol Phylogenet Evol 34: 97–105.
49. Sorensen MD, Balakrishnan CN, Payne R (2004) Clade-limited colonization in brood parasitic finches (Fididae spp.). Syst Biol 53: 140–153.
50. Barker FK, Cibois A, Schikler P, Feinstein J, Cracraft J (2004) Phylogeny and diversification of the largest avian radiation. Proc Natl Acad Sci U S A 101: 11040–11045.
51. Cracraft J, Barker FK (2009) Passerine birds (Passeriformes). In: SBHedges, SKumar, eds. The timetree of life. Oxford: Oxford University Press. pp 423–431.
52. Zhou Y, Holmes EC (2007) Bayesian estimates of the evolutionary rate and age of hepatitis B virus. J Mol Evol 65: 197–205.
53. Beck J, Nossal M (2007) Hepatitis B virus replication. World J Gastroenterol 13: 69–77.
54. Nowak MA, Bonhoeffer S, Hill AM, Boehme R, Thomas HC, et al. (1996) Viral dynamics in hepatitis B virus infection. Proc Natl Acad Sci U S A 93: 214–223.
55. Yang W, Robinson WS, Marion PL (1997) Duck hepatitis B virus in infected liver of the yolk sac of developing embryos. J Virol 61: 2273–2279.
56. Kock J, Thielmann L, Galle P, Schlicht HJ (1996) Hepatitis B virus nucleic acids associated with human peripheral blood mononuclear cells do not originate from replicating virus. Hepatology 23: 405–413.
57. Huang JM, Huang TH, Qiu HY, Chen TS, Bai G, et al. (2006) Presence of HBsAg, HBcAg, and HBVDNA in ovary and ovum of the patients with chronic hepatitis B virus infection. Jpn J Exp Med 57: 231–236.
58. Emerson BC (2007) Alarm bells for the molecular clock? No support for Ho et al.’s model of time-dependent molecular rate estimates. Syst Biol 56: 315–322.
59. Henn BM, Giginoux CR, Feldman MW, Mountain JL (2009) Characterizing the time dependency of human mitochondrial DNA mutation rate estimates. Mol Biol Evol 26: 217–230.
60. Emerson BC (2007) Alarm bells for the molecular clock? No support for Ho et al.’s model of time-dependent molecular rate estimates. Syst Biol 56: 315–322.
61. Mizokami M, Orito E, Ohba K, Ikeo K, Lau JY, Gojobori T (1997) Constrained statistical analysis of human T-cell lymphotropic virus evolutionary rates. Infect Genet Evol 5: 291–298.
62. Axelsson E, Smith NG, Sundstro¨m H, Berlin S, Ellegren H (2004) Male-biased mutation rate and divergence in autosomal, z-linked and w-linked introns of chicken and turkey. Mol Biol Evol 21: 1330–1347.
63. Matsukura K, Tokino T (1990) Integration of hepatitis B virus DNA and its implications for hepatocarcinogenesis. Mol Biol Med 7: 243–260.
64. Waddell PJ, Cao Y, Hasegawa M, Minndell DP (1999) Assessing the Cretaceous suprerglacial divergence times among birds and placental mammals by using whole mitochondrial protein sequences and an extended statistical framework. Syst Biol 48: 119–137.
65. Matsukura K, Tokino T (1990) Integration of hepatitis B virus DNA and its implications for hepatocarcinogenesis. Mol Biol Med 7: 243–260.
66. Waddell PJ, Cao Y, Hasegawa M, Minndell DP (1999) Assessing the Cretaceous suprerglacial divergence times among birds and placental mammals by using whole mitochondrial protein sequences and an extended statistical framework. Syst Biol 48: 119–137.