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 paleovirology, the study of paleoviruses and the way viruses 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
**Author Summary**

Paleovirology is the study of ancient viruses and the way they have shaped the innate immune system of their hosts over millions of years. One way to reconstruct the deep evolution of viruses is to search for viral sequences “fossilized” at different evolutionary time points in the genome of their hosts. Besides retroviruses, few virus families are known to have deposited molecular relics in their host’s genomes. Here we report on the discovery of multiple fragments of viruses belonging to the Hepadnaviridae family (which includes the human hepatitis B viruses) fossilized in the genome of the zebra finch. We show that some of these fragments penetrated the germline genome of passerine birds more than 19 million years ago, which implies that hepadnaviruses are much older than previously thought. Based on this age, we can infer a long-term avian hepadavirus substitution rate, which is a 1,000-fold slower than all short-term substitution rates calculated based on extant hepadnavirus sequences. These results call for a reevaluation of the deep origin of Hepadnaviridae, and indicate that some exogenous hepadnaviruses may still be circulating today in various passerine birds.

Human deaths of acute or chronic liver disease per year [18]. While replication of these viruses does not rely on integration into the host genome, a relatively large number of chromosomal integration events have been characterized in mammalian liver cells sampled from chronically infected individuals [19]. In this study, we show that hepadnaviruses have also infiltrated the germline genome of some of their vertebrate hosts in the distant past. The viral sequences fossilized since these endogenization events offer an unprecedented opportunity to reevaluate the mode and tempo of Hepadnaviridae evolution.

**Results**

**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 73% 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 hepadnaviruses [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* (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).

**Figure 1. Map showing the position of the ezHBVs in the DHBV genome.** Grey rectangles represent the different open reading frames of DHBV (GenBank accession number AY494851, isolated from a puna teal) encoding the precore (preC), core (C), polymerase (P), surface (preS and S), and regulatory HBx-like (X) proteins. ezHBVs (black lines) are labeled with letters as in Table 1. Group 1 and 2 correspond to the two groups of overlapping fragments used for the phylogenetic analyses (Figure 2). Asterisks indicate the occurrence of segmental duplication (see text). The white triangle in ezHBVi* indicates the presence of a 618-bp solo LTR belonging to the TguLTRK1b LTR retrotransposon family [41]. The solo LTR is flanked by a 6-bp target site duplication (GCTCTC).

doi:10.1371/journal.pbio.1000495.g001
Table 1. Characteristics of the endogenous HBV fragments found in the zebra finch genome (eZHBVs).

<table>
<thead>
<tr>
<th>eZHBV</th>
<th>Position in DHBVa</th>
<th>Position in Finch Genome</th>
<th>Length (bp)</th>
<th>NS Mutationsb</th>
<th>Similarity to DHBV</th>
<th>5’ Gene in Finch</th>
<th>3’ Gene in Finch</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2773 27 24</td>
<td>5,223,682 5,223,949</td>
<td>268</td>
<td>0/1</td>
<td>59.8%</td>
<td>Scn3b (690 bp)</td>
<td>Scn3b (50 bp)</td>
</tr>
<tr>
<td>b</td>
<td>85 213 1</td>
<td>50,694,672 50,694,800</td>
<td>129</td>
<td>1/0</td>
<td>63.8%</td>
<td>Fry (2.2 kb)</td>
<td>Fry (1.7 kb)</td>
</tr>
<tr>
<td>c</td>
<td>269 781 12</td>
<td>3,002,393 3,002,898</td>
<td>507</td>
<td>0/1</td>
<td>61.5%</td>
<td>ATP2B2 (10 kb)</td>
<td>ATP2B2 (42 kb)</td>
</tr>
<tr>
<td>d</td>
<td>428 670 26</td>
<td>864,217 864,459</td>
<td>246</td>
<td>1/0</td>
<td>59.3%</td>
<td>Trim33 (1.3 kb)</td>
<td>Trim33 (12.5 kb)</td>
</tr>
<tr>
<td>e</td>
<td>428 556 6</td>
<td>2,324,466 2,324,594</td>
<td>129</td>
<td>0/0</td>
<td>62%</td>
<td>cdh23 (5.5 kb)</td>
<td>cdh23 (500 bp)</td>
</tr>
<tr>
<td>f</td>
<td>476 757 1A</td>
<td>68,942,470 68,942,804</td>
<td>335</td>
<td>4/1</td>
<td>62.2%</td>
<td>LMO3 (13 kb)</td>
<td>MGST1 (34 kb)</td>
</tr>
<tr>
<td>g</td>
<td>476 622 1A</td>
<td>68,944,006 68,944,152</td>
<td>147</td>
<td>1/0</td>
<td>70.5%</td>
<td>LMO3 (15 kb)</td>
<td>MGST1 (32 kb)</td>
</tr>
<tr>
<td>h</td>
<td>530 673 8</td>
<td>25,658,020 25,658,163</td>
<td>144</td>
<td>0/0</td>
<td>69.4%</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>i</td>
<td>590 790 Z</td>
<td>8,647,882 8,648,070</td>
<td>185</td>
<td>1/0</td>
<td>60.8%</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>j</td>
<td>1310 2521 Z</td>
<td>8,648,446 8,649,654</td>
<td>1,209</td>
<td>1/0</td>
<td>65.5%</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>k</td>
<td>1466 2512 1</td>
<td>50,694,887 50,695,943</td>
<td>1,057</td>
<td>4/5</td>
<td>54.8%</td>
<td>Fry (2.5 kb)</td>
<td>Fry (500 bp)</td>
</tr>
<tr>
<td>l</td>
<td>1388 1936 20</td>
<td>1,111,838 1,112,362</td>
<td>537</td>
<td>2/0</td>
<td>72.3%</td>
<td>C20orf4 (14 kb)</td>
<td>C20orf4 (6 kb)</td>
</tr>
<tr>
<td>l*</td>
<td>1388 1936 20</td>
<td>1,111,836 1,112,278</td>
<td>537</td>
<td>1/0</td>
<td>72.7%</td>
<td>Dhx32 (1 kb)</td>
<td>Dhx32 (100 bp)</td>
</tr>
<tr>
<td>m</td>
<td>1334 1780 6</td>
<td>33,336,955 33,337,390</td>
<td>436</td>
<td>2/3</td>
<td>75%</td>
<td>Dhx32 (1 kb)</td>
<td>Dhx32 (100 bp)</td>
</tr>
<tr>
<td>n</td>
<td>1562 1789 19</td>
<td>5,856,561 5,856,833</td>
<td>273</td>
<td>0/0</td>
<td>72.8%</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

eZHBVa-eZHBVb, eZHBVc, and eZHBVd are in an intron; a dash indicates that there is no gene within 50 kb 5’ and/or 3’ of the insertion. eZHBV and eZHBVb derive from post-insertional duplication (see Results for details). 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.

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 eZHBVb occurred, which must postdate the chromosomal integration of the ancestral eZHBVl 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 $2 \times 10^{-9}$ and $3.9 \times 10^{-9}$ substitutions per site per year (sub/site/year) [21], values similar to the range of those estimated for mammals ($2.2 \times 10^{-9}$ to $4.5 \times 10^{-9}$ sub/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 eZHBVl duplicates (0.015) by the bird neutral substitution rates yields a duplication time ranging between 3.8 My (with $3.9 \times 10^{-9}$ sub/site/year) and 7.5 My (with $2 \times 10^{-9}$ sub/site/year). The timing of this duplication provides a minimal estimate for the integration of the ancestral eZHBVl 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 (eZHBVb and eZHBVd) in three other species of estrildid finches (black throated finch [Poephila cincta], scally breasted munia [Lonchura punctulata], and gouldian finch [Chloebia gouldiae]) and in the dark-eyed junco (Junco hyemalis), a non-estridid 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 eZHBVl in the olive sunbird (Cyannomitra 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 eZHBVb and eZHBVd (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 eZHBVl was inserted after the divergence of the
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 eZHBVl prior to 25 My, and that of eZHBVa between 35 and 25 My ago.

Our last estimate of the age of eZHBVs relies on the level of sequence divergence between orthologous eZHBV sequences. The corrected distances inferred for orthologous eZHBV sequences are provided in Datasets S2 and S3.

\[ \text{doi:10.1371/journal.pbio.1000495.g002} \]
Table 2. Corrected distances between avian hepadnaviruses calculated on the region corresponding to group 2 eZHBVs (see Figure 1).

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

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.t002

Discussion

In this study we provide evidence that the germline genome of passerine birds has been infiltrated by several ancient and diverse hepadnaviruses that still show surprisingly high levels of similarity to extant avian hepadnaviruses. Although eZHBVs represent, to our knowledge, the first instance of endogenous DNA viruses reported in animals, several characteristics of hepadnaviruses suggest that endogenization of these viruses may be likely. Hepadnaviruses replicate in the nucleus of their host’s cells via a reverse-transcribed RNA intermediate [32,33]. Part of their life cycle is therefore spent in close proximity to the host DNA, which may facilitate chromosomal integration via various host- or transposable-element-mediated mechanisms that use either DNA or RNA templates (e.g., [15]). Indeed, although integration into the host genome is not required for the replication of the virus, integrated 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
Fossil Hepadnaviruses Uncovered in Bird Genomes

A

B

C

Extant avian hepadnaviruses

Orthologous eZHBVa

Orthologous eZHBVj

Orthologous eZHBVI
Figure 3. Illustration and phylogenetic trees of orthologous eZHBVα, eZHBVβ, and eZHBVγ. The primers (Scn3b-F/R) used to amplify eZHBVα (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 eZHBVβ (B) is located in the region flanking the insertion in 3′ of zebra finch Chromosome Z, while the other (8718F) is anchored in eZHBVγ. One of the primers (1978F) used to amplify eZHBVγ (C) is located in the region flanking the insertion in 5′ of zebra finch Chromosome 20, while the other (Hrf1) is anchored in eZHBVγ. 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 3.2.4/taeGut1).

doi:10.1371/journal.pbio.1000495.g003

Fossil Hepadnaviruses Uncovered in Bird Genomes

...gently bind to many different cell types [35], and there is evidence supporting the presence of HBV DNA in spermatozoa and 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 eZHBVγ (position 2521) and eZHBVκ (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 eZHBVδ and eZHBVε; 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 recombination, 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 eZHBVδ and those between eZHBVδ, eZHBVγ, eZHBVκ, and eZHBVλ 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 eZHBVα and eZHBVβ 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.
Knowledge on the deep evolution of extant viruses remains between short- and long-term substitution rates. Because our possible that saturation may in part hinder our ability to accurately could also explain part of the difference between short- and long-mutational saturation (homoplasy and back mutations), which mutation rates of hepadnaviruses may also result in strong viruses and for cellular organisms. Lastly, the high background rates may be more pronounced for hepadnaviruses than for other the RT, suggests that the effect of purifying selection on long-term genetic organization, combined with the intrinsically low fidelity of faster than overlapping regions [31,62]. This tightly constrained shown that nonoverlapping regions of the HBV genome evolve 60%; see [61] for discussion). About 60% of the HBV genome codes dependency of substitution rates suggested by this study is on them.

Another major process that may be responsible for the time another 3.5% of the HBV genome coding 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 may be 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 (eZHBVg). For eZHBVl, we used a forward primer (1978F) anchored in the 5\' insertion). For eZHBVj, we used a forward primer (8718F) anchored within eZHBVj, at position 712–734 in combination with a reverse primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVg, we designed PCR primers on the flanking regions of three insertions. For eZHBVj, we used a forward primer (hfr1) anchored within eZHBVl, at position 239–257. For eZHBVg, we used a forward primer (Scn3b-R) was anchored in the 3\' downstream of the insertion in T. guttata. The reverse primer (Scn3b-F) was anchored in the 3\' insertion in T. guttata, corresponding to the third exon of the predicted scn3b gene.

The identity of the different bird species used in this study was verified by sequencing a 420-bp fragment of the mitochondrial NADH dehydrogenase subunit 2 (NADH2) gene (Figure S1) using the following primers: Fwd 5\'--AGT CAT TTT GGS AGG AAT CCT G; Rev 5\'--TTC CAY TTC TGA TTY CCA GAA G. Standard PCR conditions were as follows: 2 min at 94\(^\circ\)C; 30 cycles of 1 min at 94\(^\circ\)C, 30 s at 48–62\(^\circ\)C, and 30 s to 2 min at 72\(^\circ\)C. PCR mix was buffer (5x), 5 μl; MgCl\(_2\) (25 mM), 2 μl; dNTP (10 μM), 0.5 μl; primer 1 (10 μM), 1 μl; primer 2 (10 μM), 1 μl; Taq (GoTaq, Promega), 1.25 U; DNA, 30–100 ng; and H\(_2\)O

### Table 3. Corrected distances between avian hepadviruses calculated on the region corresponding to group 1 eZHBVs (see Figure 1).

<table>
<thead>
<tr>
<th></th>
<th>eZHBVc</th>
<th>eZHBVd</th>
<th>eZHBVe</th>
<th>eZHBVh</th>
<th>eZHBVi</th>
<th>eZHBVf</th>
<th>eZHBVg</th>
<th>eZHBVj</th>
<th>Crane</th>
<th>Heron</th>
<th>Ross’s Goose</th>
<th>Snow Goose</th>
<th>White Stork</th>
</tr>
</thead>
<tbody>
<tr>
<td>eZHBVc</td>
<td>0.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eZHBVd</td>
<td>0.55</td>
<td>0.52</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eZHBVe</td>
<td>0.28</td>
<td>0.27</td>
<td>0.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eZHBVh</td>
<td>2.45</td>
<td>2.05</td>
<td>—</td>
<td>1.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eZHBVi</td>
<td>0.61</td>
<td>0.36</td>
<td>0.21</td>
<td>0.32</td>
<td>2.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>eZHBVf</td>
<td>0.30</td>
<td>0.21</td>
<td>0.23</td>
<td>0.14</td>
<td>—</td>
<td>0.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Crane</td>
<td>1.04</td>
<td>1.13</td>
<td>1.46</td>
<td>0.57</td>
<td>1.13</td>
<td>0.95</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heron</td>
<td>1.01</td>
<td>1.42</td>
<td>1.61</td>
<td>0.79</td>
<td>1.89</td>
<td>1.13</td>
<td>0.76</td>
<td>0.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ross’s goose</td>
<td>1.02</td>
<td>1.30</td>
<td>1.37</td>
<td>0.76</td>
<td>0.95</td>
<td>0.96</td>
<td>0.54</td>
<td>0.18</td>
<td>0.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Snow goose</td>
<td>1.03</td>
<td>1.16</td>
<td>1.55</td>
<td>0.70</td>
<td>1.00</td>
<td>1.10</td>
<td>0.71</td>
<td>0.17</td>
<td>0.39</td>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White stork</td>
<td>0.95</td>
<td>1.28</td>
<td>1.69</td>
<td>0.68</td>
<td>1.30</td>
<td>1.00</td>
<td>0.77</td>
<td>0.32</td>
<td>0.20</td>
<td>0.32</td>
<td>0.35</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Duck</td>
<td>1.09</td>
<td>1.34</td>
<td>1.11</td>
<td>0.66</td>
<td>1.20</td>
<td>1.07</td>
<td>0.61</td>
<td>0.18</td>
<td>0.40</td>
<td>0.16</td>
<td>0.09</td>
<td>0.32</td>
<td></td>
</tr>
</tbody>
</table>

Distances were calculated using the TPM2uf+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 0.7. This distance threshold corresponds to twice the average distance between extant avian hepadnaviruses (2×0.27=0.54) 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).

doi:10.1371/journal.pbio.1000495.t003
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 HQ116364–HQ116383).

Analyses of Selection

Analyses of selection were carried out on alignments of each set of orthologous insertions amplified in the various passerine species (eZHBVj, eZHBVa, and eZHBVl; 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 global 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, 0.36) (confidence interval: 0.44, 0.88), and 0.93 (confidence interval: 0.62, 1.24) for eZHBVj, eZHBVa, and eZHBVl 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 eZHBVl, 0.16 for eZHBVa, and 0.81 for eZHBVj), suggesting that eZHBVj, eZHBVa, and eZHBVl 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 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 eZHBVl, 0.29 for eZHBVa, and 0.85 for eZHBVj), suggesting neutral evolution in all branches.

Distances between Avian Hepadnaviruses

All distances were calculated under maximum likelihood settings in PAUP 4.0 [65], using models of nucleotide substitution chosen based on the Akaike Information Criterion in jModeltest [66]: TPM2uf+G for group 1 eZHBVs, TVM+G for group 2 eZHBVs and for the distance between eZHBVj and extant avian hepadnaviruses, TPM1 for the distances between passerine eZHBVj orthologs, and HKY for the distance between passerine eZHBVj orthologs.

In order to estimate whether eZHBVs 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-eZHBV distances to the average distances between extant avian hepadnaviruses and more divergent strains were endogenized at widely different times during the last 19 My ago, we compared inter-eZHBV distances to the average distances between extant avian hepadnaviruses

Distances between eZHBVs

The 0.16 estimate is based on the highest of the distances between two eZHBVs being compared 19 My ago. The 0.16 estimate is based on the highest of the distances between dark-capped junco and zebra finch orthologs (eZHBVj), i.e., 2 ± A in Figure 4.

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]. eZHBVs were aligned at the amino acid level with representative members of extant avian and mammalian hepadnaviruses and analyzed using the rREV (group 1 eZHBVs) and LG+G+F (group 2 eZHBVs) models in PHYML and with a prior setting allowing model jumping between fixed-rate amino acid models in MrBayes. eZHBVj, eZHBVl, and eZHBVl orthologs were analyzed with the TPM2uf+G, TPM2uf+G, and TPM2uf+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 eZHBVs 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 eZHBVs (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 eZHBVs (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.

Found at: doi:10.1371/journal.pbio.1000495.s003 (0.01 MB XLS)

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 TguERVK9_LTR2g element; [41]) inserted within the region orthologous to eZHBVs. The solo LTR is flanked by a 6-bp target site duplication (GACCTT). The alignment includes the sequence of the primer Scn3b-F, 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-R, 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.

Found at: doi:10.1371/journal.pbio.1000495.s004 (0.00 MB XLS)

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 1209 of the T. guttata sequence of the primer 1978F, located in the 5′ flanking region of eZHBVj.

Found at: doi:10.1371/journal.pbio.1000495.s005 (0.01 MB XLS)

Dataset S7 Alignment (in FASTA format) of the NADH2 partial sequences used to construct the tree in Figure S1. Found at: doi:10.1371/journal.pbio.1000495.s006 (0.01 MB XLS)

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. cecalata).

Found at: doi:10.1371/journal.pbio.1000495.s007 (0.01 MB XLS)

Table S1 Tissue samples used in this study. All Estrildid 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.

Found at: doi:10.1371/journal.pbio.1000495.s009 (0.03 MB DOC)

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 Birks and the University of Washington Burke Museum, David Kabelik (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.

References


