Genome Biology

BioMed Central
The Open Access Publisher

Low frequency of paleoviral infiltration across the avian phylogeny

Genome Biology 2014, **15**:539 doi:10.1186/s13059-014-0539-3

Jie Cui (jiecui@yahoo.com)
Wei Zhao (zhaowei@genomics.org.cn)
Zhiyong Huang (huangzhiyong@genomics.cn)
Erich D Jarvis (jarvis@neuro.duke.edu)
M Thomas P Gilbert (mtpgilbert@gmail.com)
Peter J Walker (Peter.Walker@csiro.au)
Edward C Holmes (edward.holmes@sydney.edu.au)
Guojie Zhang (zhanggj@genomics.org.cn)

Published online: 11 December 2014

ISSN 1465-6906

Article type Research

Submission date 11 February 2014

Acceptance date 10 November 2014

Article URL http://genomebiology.com/15/11/539

Low frequency of paleoviral infiltration across the avian phylogeny

Jie Cui[1,8,*,†]
Email: jiecui@yahoo.com

Wei Zhao[2,†]
Email: zhaowei@genomics.org.cn

Zhiyong Huang[2]
Email: huangzhiyong@genomics.cn

Erich D Jarvis[3]
Email: jarvis@neuro.duke.edu

M Thomas P Gilbert[4,7]
Email: mtpgilbert@gmail.com

Peter J Walker[5]
Email: Peter.Walker@csiro.au

Edward C Holmes[1]
Email: edward.holmes@sydney.edu.au

Guojie Zhang[2,6,**]
Email: zhanggj@genomics.org.cn

[1] Marie Bashir Institute for Infectious Diseases and Biosecurity, Charles Perkins Centre, School of Biological Sciences and Sydney Medical School, The University of Sydney, Sydney, NSW 2006, Australia

[2] China National GeneBank, BGI-Shenzhen, Shenzhen 518083, China

[3] Howard Hughes Medical Institute, Duke University Medical Center, Department of Neurobiology, Box 3209, Durham, North Carolina 27710, USA

[4] Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, DK-1350 Copenhagen, Denmark

[5] CSIRO Animal, Food and Health Sciences, Australian Animal Health Laboratory, Geelong, Victoria 3220, Australia

[6] Centre for Social Evolution, Department of Biology, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen, Denmark

[7] Trace and Environmental DNA Laboratory, Department of Environment and Agriculture, Curtin University, Perth, Western Australia 6102, Australia

[8] Program in Emerging Infectious Diseases, Duke-NUS Graduate Medical School, Singapore 169857, Singapore

[*] Corresponding author. Marie Bashir Institute for Infectious Diseases and Biosecurity, Charles Perkins Centre, School of Biological Sciences and Sydney Medical School, The University of Sydney, Sydney, NSW 2006, Australia

[**] Corresponding author. China National GeneBank, BGI-Shenzhen, Shenzhen 518083, China

[†] Equal contributors.

Abstract

Background

Mammalian genomes commonly harbor endogenous viral elements. Due to a lack of comparable genome-scale sequence data, far less is known about endogenous viral elements in avian species, even though their small genomes may enable important insights into the patterns and processes of endogenous viral element evolution.

Results

Through a systematic screening of the genomes of 48 species sampled across the avian phylogeny we reveal that birds harbor a limited number of endogenous viral elements compared to mammals, with only five viral families observed: Retroviridae, Hepadnaviridae, Bornaviridae, Circoviridae, and Parvoviridae. Strikingly, only members of the Retroviridae were observed in three nonavian reptile species used as a comparison. All nonretroviral endogenous viral elements are present at low copy numbers and in few species, with only endogenous hepadnaviruses widely distributed, although these have been purged in some cases. We also provide the first evidence for endogenous bornaviruses and circoviruses in avian genomes, although at very low copy numbers. A comparative analysis of vertebrate genomes revealed a simple linear relationship between endogenous viral element abundance and host genome size, such that the occurrence of endogenous viral elements in bird genomes is 6–13 fold less frequent than in mammals.

Conclusions

These results reveal that avian genomes harbor relatively small numbers of endogenous viruses, particularly those derived from RNA viruses, and hence are either less susceptible to viral invasions or purge them more effectively.

Background

Vertebrate genomes commonly harbor retrovirus-like [1] and non-retrovirus-like [2] viral sequences, resulting from past chromosomal integration of viral DNA (or DNA copies of viral RNA) into host germ cells. Tracing the evolutionary histories of these endogenous viral elements (EVEs) can provide important information on the origin of their extant counterparts,

and provide an insight into host genome dynamics [3-7]. Recent studies have shown that these genomic 'fossils' can also influence the biology of their hosts, both beneficially and detrimentally; for example, by introducing novel genomic rearrangements, influencing host gene expression, as well as evolving into new protein-coding genes with cellular functions (i.e., 'gene domestication') [4,6].

Because integration into host genomes is intrinsic to the replication cycle of retroviruses which employ reverse transcriptase (RT), it is no surprise that retroviruses are commonly found to have endogenous forms in a wide range of animal genomes [8]. Indeed, most of the EVEs present in animal genomes are of retroviral origin – endogenous retroviruses (ERVs) – and EVEs representing all retroviral genera, with the exception of *Deltaretrovirus*, have been found to possess endogenous forms. Remarkably, recent studies have revealed the unexpected occurrence of non-retroviral elements in various animal genomes, including RNA viruses that lack a DNA form in their replication cycle [2,6]. Since their initial discovery, EVEs in animal genomes have been documented for families of double-stranded (ds) DNA viruses (virus classification Group I) – Herpesviridae; single-stranded (ss) DNA viruses (Group II) – Circoviridae and Parvoviridae; ssRNA viruses (Group IV) – Bornaviridae and Filoviridae; ssRNA-RT viruses (Group VI) – Retroviridae; and dsDNA-RT viruses (Group VII) – Hepadnaviridae [6].

To date, most studies of animal EVEs have focused on mammals due to their relatively high density of sampling. In contrast, few studies on the EVEs present in avian species have been undertaken. The best-documented avian EVEs are endogenous hepadnaviruses. These virally derived elements were first described in the genome of a passerine bird – the zebra finch [9] – and then in the genome of the budgerigar [10] as well as some other passerines [11], and may have a Mesozoic origin in some cases [11]. Also of note was the discovery of a great diversity of ERVs in the genomes of zebra finch, chicken and turkey, most of which remain transcriptionally active [12]. In contrast, most mammalian ERVs are inert.

In this study, we systematically mined 48 avian genomes for EVEs of all viral families, as one of a body of companion studies on avian genomics [13,14]. Importantly, our data set represents all 32 neognath and two of the five palaeognath orders, and thus represents nearly all major orders of extant birds. Such a large-scale data analysis enabled us to address a number of key questions in EVE evolution, namely (i) what types of viruses have left such genomic fossils across the avian phylogeny and in what frequencies, (ii) what are the respective frequencies of EVE inheritance between species and independent species-specific insertion, and (iii) what is the frequency and pattern of avian EVE infiltration compared to other vertebrates?

Results

Genome scanning for avian EVEs

Our *in silico* genomic mining of the 48 avian genomes (Additional file 1: Table S1; [13,14]) revealed the presence of five families of endogenous viruses – Retroviridae, Hepadnaviridae, Circoviridae, Parvoviridae, and Bornaviridae (Figure 1), almost all of which (>99.99%) were of retroviral origin. Only a single family of RNA viruses (Group IV; the Bornaviridae) was present. The genomes of American alligator, green turtle and anole lizard only contained EVEs of retroviral origin. Notably, three closely related oscine passerine birds – the

American crow, medium ground-finch and zebra finch – possessed greater ERV copy numbers in their genomes than the avian average (Table 1; discussed in detail below), while their suboscine passerine relatives – rifleman and golden-collared manakin – possessed lower ERV numbers close to the avian average (Table 1) and occupied basal positions in the passerine phylogeny (Figure 1). Hence, there appears to have been an expansion of ERVs coincident with the species radiation of the suborder Passeri.

Figure 1 Distribution of EVEs of all virus families across the avian phylogeny. EVEs are colored according to virus family and marked on the species tree. Colors are as follows: red, Hepadnaviridae; black, Retroviridae; blue, Circoviridae; green, Parvoviridae; and yellow, Bornaviridae. The phylogeny is based on the results of our phylogenomics consortium whole genome analyses across all the species shown.

Table 1 EVE copy numbers in avian genomes

Species name	Hepadna-	Borna-	Circo-	Parvo-	Retroviral copy number					
					Total	Alpha-	Beta-	Gamma-	Epsilon-	Others*
Acanthisitta chloris	2	0	0	1	302	8	111	160	9	14
Anas platyrhynchos	4	0	0	0	281	7	54	186	17	17
Antrostomus carolinensis	2	0	0	0	246	15	76	119	16	20
Apaloderma vittatum	2	0	0	0	258	10	97	130	11	10
Aptenodytes forsteri	2	0	0	0	232	11	80	104	12	25
Balearica regulorum	2	0	0	0	244	13	65	113	23	30
Buceros rhinoceros	3	0	0	0	217	9	59	113	12	24
Calypte anna	3	4	0	0	424	27	181	157	17	42
Cariama cristata	3	0	0	0	315	13	78	176	20	28
Cathartes aura	2	0	0	0	199	11	33	115	11	29
Chaetura pelagica	2	1	0	0	383	15	113	213	13	29
Charadrius vociferus	1	0	0	0	467	25	161	221	18	42
Chlamydotis macqueenii	1	0	0	1	216	8	50	127	10	21
Columba livia	2	0	0	0	245	11	81	116	17	20
Colius striatus	1	0	0	0	237	9	94	110	7	17
Corvus brachyrhynchos	1	0	0	2	1,032	13	475	472	22	50
Cuculus canorus	2	0	0	0	191	11	73	95	2	10
Egretta garzetta	2	0	1	1	289	23	95	129	16	26
Eurypyga helias	2	0	0	0	288	6	104	147	12	19
Falco peregrinus	2	0	0	0	336	15	90	196	7	28
Fulmarus glacialis	2	0	0	0	245	10	65	121	11	38
Gallus gallus	0	0	0	0	573	21	146	228	54	124
Gavia stellata	4	0	0	0	207	12	37	125	12	21
Geospiza fortis	10	0	1	0	785	11	340	371	26	37
Haliaeetus albicilla	2	0	0	0	301	11	103	136	15	36
Haliaeetus leucocephalus	2	0	0	0	419	23	134	190	27	45
Leptosomus discolor	3	0	0	0	301	17	96	141	17	30
Manacus vitellinus	4	0	0	1	324	7	142	151	6	18
Meleagris gallopavo	0	0	0	0	303	7	73	140	21	62
Melopsittacus undulatus	38	0	0	0	485	27	117	284	26	31
Merops nubicus	2	0	0	0	418	11	149	191	31	36
Mesitornis unicolor	1	0	0	1	451	10	153	242	21	25
Nestor notabilis	5	0	1	0	223	8	65	116	20	14
Nipponia nippon	3	0	0	0	302	35	79	127	28	33
Opisthocomus hoazin	1	0	0	1	425	10	151	208	21	35
Pelecanus crispus	2	0	0	3	283	13	86	114	22	48

Phalacrocorax carbo	68	0	0	0	305	11	87	153	27	27
Phaethon lepturus	2	0	0	0	480	9	110	312	14	35
Phoenicopterus ruber	2	0	0	0	209	9	54	100	20	26
Picoides pubescens	2	1	0	0	502	9	164	278	20	31
Podiceps cristatus	3	0	0	0	366	7	123	187	23	26
Pterocles gutturalis	1	0	0	1	165	10	43	82	8	22
Pygoscelis adeliae	2	0	0	0	244	12	64	123	21	24
Struthio camelus	2	0	0	0	132	7	30	61	8	26
Taeniopygia guttata	13	0	0	1	725	19	302	322	34	48
Tauraco erythrolophus	1	0	0	0	397	5	168	198	5	21
Tinamus major	3	0	2	0	328	8	148	140	7	25
Tyto alba	5	0	0	0	477	10	169	244	16	38

* Retroviral elements that matched the *Retroviridae* but not to a specific genus.

We next consider each of the EVE families in turn.

EVEs related to the retroviridae

As expected, ERVs were by far the most abundant EVE class in the avian genomes, covering the genera *Alpha-*, *Beta-*, *Gamma-*, and *Epsilonretrovirus*, with total ERV copy numbers ranging from 132 to 1,032. The greatest numbers of ERVs were recorded in the three oscine passerines (American crow, medium ground-finch and zebra finch, respectively) that exhibited EVE expansion (Table 1). ERVs related to beta- and gammaretroviruses were the most abundant in all avian genomes as noted in an important earlier study of three avian genomes [12]. In contrast, ERVs derived from epsilonretroviruses were extremely rare, with very few copies distributed (Additional file 2: Figure S1). We also found that ERVs related to alpharetroviruses were widely distributed in avian phylogeny, although with very low copy numbers [12]. In accord with the overall genetic pattern among the EVEs, the three oscine passerines exhibited greater numbers of ERVs than other taxa (2- to 3-fold higher than the average) (Table 1). This suggests that an ERV expansion occurred in the oscine passerines subsequent to their split from the suboscines. Phylogenetic analysis revealed that this pattern was due to frequent invasions of similar beta- and gammaretroviruses in these species (Additional file 2: Figure S1; Table 1).

Strikingly, the avian and non-avian (American alligator, green turtle and anole lizard) genomes seldom shared orthologous sequences (i.e., only a few avian sequences can be aligned with those of non-avians and without matching flanking regions) and all their ERVs were distantly related (Additional file 2: Figure S1), indicative of a lack of vertical or horizontal transmission among these vertebrates. In addition, no non-retroviral elements were found in the non-avian genomes.

EVEs related to the hepadnaviridae

Hepadnaviruses have very small genomes (~3 kb) of partially double-stranded and partially single-stranded circular DNA. Their replication involves an RNA intermediate that is reverse transcribed in the cytoplasm and transported as cDNA back into the nucleus. Strikingly, we found endogenous hepadnaviral elements in all the avian genomes studied (Additional file 1: Table S2), such that they were the most widely distributed non-retroviral EVEs recorded to date. In this context it is important to note that no mammalian endogenous hepadnaviruses have been described even though primates are major reservoirs for exogenous hepatitis B

viruses [15]. Hepadnaviral EVEs were also absent from the American alligator, green turtle and anole lizard genomes.

Our phylogenetic analysis revealed a number of notable evolutionary patterns in the avian endogenous hepadnaviruses: (i) endogenous hepadnaviruses exhibited a far greater phylogenetic diversity, depicted as diverse clades, than their exogenous relatives (Additional file 3: Figure S2), suggesting they were older, although an acceleration in evolutionary rates among some hepadnaviral EVEs cannot be excluded; (ii) exogenous hepadnaviruses formed a tight monophyletic group compared to the endogenous elements (Additional file 3: Figure S2), indicative of a turnover of exogenous viruses during avian evolution; (iii) there was a marked difference in copy number (from 1 to 68) among avian species (Additional file 1: Table S2), suggestive of the frequent gain and loss of viruses during avian evolution; and (iv) there was a phylogeny-wide incongruence between the virus tree (Additional file 3: Figure S2) and the host tree (P =0.233 using ParaFit method), indicative of multiple independent genomic integration events as well as potential cross-species transmission events.

Despite the evidence for independent integration events, it was also clear that some hepadnavirus EVEs were inherited from a common ancestor of related avian groups, and perhaps over deep evolutionary time-scales. We documented these cases by looking for pairs of endogenous hepadnaviruses from different avian hosts that received strong (>70%) bootstrap support (Additional file 4: Data S1) and which occupied orthologous locations. Specifically: (i) in the genomes of the white-tailed and bald eagles, the 5′ end of an hepadnavirus EVE was flanked by a same unknown gene while the 3′ end was flanked by the *dendritic cell immunoreceptor* (*DCIR*) gene (Additional file 3: Figure S2); (ii) an EVE shared by the emperor penguin and Adelie penguin (Additional file 3: Figure S2) was flanked by a same unknown gene at the 5′ end and the *Krueppel-like factor 8-like* gene at the 3′ end; and (iii) the ostrich and the great tinamou had the same flanking genes, albeit of unknown function, at both ends of an EVE.

We also recorded a rare case of vertical transmission of a hepadnavirus with a complete genome that has seemingly been inherited by 31 species (Additional file 1: Table S2) prior to the diversification of the Neoaves 73 Myr ago [14]. This virus has been previously denoted as eZHBV_C [11], and was flanked by the *furry homolog* (*FRY*) gene at both the 5′ and 3′ ends. Our hepadnavirus phylogeny (Figure 2) showed that this EVE group clustered tightly with extremely short internal branches, although with some topological patterns that were inconsistent with the host topology (Figure 1). A lack of phylogenetic resolution notwithstanding, this mismatch between the virus and host trees could be also in part be due to incomplete lineage sorting, in which there has been insufficient time for allele fixation during the short time period between bird speciation events. Indeed, Neoaves are characterized by a rapid species radiation [16].

Figure 2 Phylogenetic tree of exogenous and endogenous hepadnaviruses generated using complete polymerase (P) protein sequences. Bootstrap values lower than 70% are not shown; one star (*) represents values higher than 70%, while two stars (**) represents values higher than 90%. Branch lengths are drawn to a scale of amino acid substitutions per site (subs/site). The tree is midpoint rooted for purposes of clarity only. The exogenous hepadnaviruses are marked. A cartoon of a virus particle marks the phylogenetic location of an inherited hepadnavirus invasion. Avian host species names are used to denote avian endogenous hepadnaviruses and scaffold numbers are given in Additional file 1: Table S2. All abbreviations are given in Additional file 1: Table S9.

Strikingly, we observed that two Galliformes species, chicken and turkey, have seemingly purged their hepadnaviral EVEs. Specifically, genomic mining revealed no hepadnaviral elements in these galliformes, even though their closest relatives (Anseriformes) contained such elements. In support of this genome purging, we noted that one hepadnaviral element present in the mallard genome has been severely degraded through frequent mutation in the chicken genome (Additional file 5: Figure S3). In addition, remnants of orthologous 5′ and 3′ regions could also be found in the turkey genome, although the rest of the element was deleted (Additional file 5: Figure S3).

EVEs related to the bornaviridae

Bornaviruses (family Bornaviridae) are linear, unsegmented negative-sense single-stranded RNA (ssRNA) viruses with genomes of ~9 kb. They are unusual among animal RNA viruses in their ability to replicate within the host cell nucleus, which in turn assists endogenization. Indeed, orthomyxoviruses and some insect rhabdoviruses also replicate in the nucleus and both have been found to occur as endogenous forms in insect genomes [2]. Endogenous elements of bornaviruses, denoted endogenous bornavirus-like N (EBLN) [2,17,18] and endogenous bornavirus-like L (EBLL) [2,18], have been discovered in mammalian genomes including humans, and those present in primates have been dated to have arisen more than 40 million years ago [17,18]. Although exogenous bornaviruses circulate in both mammals and birds and cause fatal diseases [19,20], endogenous bornaviruses have not yet been documented in avian species.

We report, for the first time, that both EBLN and EBLL are present in several avian genomes (Additional file 6: Figure S4), although in only three species and with very low copy numbers (1 – 4; Additional file 1: Table S3): the Anna's hummingbird, the closely related chimney swift, and the more distantly related woodpecker. Both EBLN and EBLL in the genome of Anna's hummingbird were divergent compared to other avian or mammalian viruses. The chimney swift possessed a copy of EBLN, which was robustly grouped in the phylogenetic tree with the EVE present in Anna's hummingbird (Additional file 6: Figure S4A). However, as these viral copies did not share the same flanking regions in the host genomes, as well as the inconsistent phylogenetic positions of the EBLN (Additional file 6: Figure S4A) and EBLL (Additional file 6: Figure S4C) of Anna's hummingbird, they likely represent independent integration events. In addition, due to the close relationships among some of the viruses in different species, it is possible that cross-species transmission has occurred because of shared geographical distributions (for example, woodpeckers are widely distributed across the United States, with geographic distributions that overlap with those of Anna's hummingbirds). The EBLN in the downy woodpecker was likely to have entered the host genome recently as in the phylogenetic tree it was embedded within the genetic diversity of exogenous viruses; the same pattern was observed in the case of the two viral copies in the genome of Anna's hummingbird (Additional file 6: Figure S4B). Similar to previous studies in mammals [21], we found that more species have incorporated EBLN than EBLL. However, compared to their wide distribution in mammalian genomes, it was striking that only three avian species carried endogenous bornavirus-like elements.

EVEs related to the circoviridae

Circoviruses (family Circoviridae) possess ~2 kb single-stranded DNA (ssDNA), nonenveloped and unsegmented circular genomes, and replicate in the nucleus via a rolling circle mechanism. They are known to infect birds and pigs and can cause a wide range of

severe symptoms such as Psittacine circovirus disease. There are two main open reading frames, usually arranged in an ambisense orientation, that encode the replication (Rep) and capsid (Cap) proteins. Endogenous circoviruses (eCiVs) are rare, and to date have only been reported in four mammalian genomes, with circoviral endogenization in carnivores dating to at least 42 million years [22].

We found circoviruses to be incorporated into only four avian genomes – medium ground finch, kea, egret, and tinamou – and at copy numbers of only 1 – 2 (Additional file 7: Figure S5; Additional file 1: Table S5). No viral copies were found in the American alligator, green turtle and anole lizard genomes analyzed. There were at least two divergent groups of eCiVs in the viral phylogenetic tree, one in the medium ground-finch and great tinamou (Additional file 7: Figure S5A, B, C), which was closely related to exogenous avian circoviruses, and another in the little egret and kea (Additional file 7: Figure S5C, D) which was only distantly related to avian exogenous counterparts. The large phylogenetic distances among these endogenous viruses are suggestive of independent episodes of viral incorporation. In addition, two pieces of evidence strongly suggested that eCiVs in the medium ground-finch and great tinamou (Additional file 7: Figure S5A, B, C) have only recently entered host genomes: (i) they had close relationships with their exogenous counterparts, and (ii) they maintained complete (or nearly complete) open reading frames (ORFs) (Additional file 1: Table S5).

EVEs related to the parvoviridae

The family Parvoviridae comprises two subfamilies – Parvovirinae and Densovirinae – that infect diverse vertebrates and invertebrates, respectively. Parvoviruses typically possess linear, non-segmented single-stranded DNA genomes with an average size of ~5 kb, and replicate in the nucleus. Parvoviruses have been documented in a wide range of hosts including humans and can cause a range of diseases [23]. Recent studies revealed that endogenous parvoviruses (ePaVs) have been broadly distributed in mammalian genomes, with integration events dating back at least 40 million years [22].

We found multiple entries of ePaVs with very low copy numbers (1 – 3; Additional file 1: Table S5) in 10 avian genomes (Additional file 8: Figure S6), and they were not as widely distributed as those parvoviruses present in mammalian genomes [22]. No viral copies were found in the three American alligator, green turtle and anole lizard genomes. All avian ePaVs were phylogenetically close to exogenous avian parvoviruses with the exception of a single one from the brown mesite, which was distantly related to all known animal parvoviruses (Additional file 8: Figure S6). We also found several cases of apparently vertical transmission. For example, one common ePaV in the American crow and rifleman was flanked by the same unknown host gene; the viral copy in the golden-collared manakin and zebra finch was flanked by the *tyrosine-protein phosphatase non-receptor type 13* (*PTPN13*) gene at 5′ end and the same unknown gene at 3′ end; and one viral element in the little egret and Dalmatian pelican was flanked by a same chicken repeat 1 (CR1) at the 5′ end and *collagen alpha 1 gene (COL14A1*) at the 3′ end (Additional file 4: Data S2). These findings suggest both independent integration and vertical transmission (i.e. common avian ancestry) for ePAVs that have seemingly existed in birds for at least 30 Myr (i.e., the separation time of *Corvus* and *Acanthisitta* [14]).

Low frequency of retroviral EVEs in bird genomes

To determine the overall pattern and frequency of infiltration of EVEs in the genomes of birds, American alligator, green turtle, anole lizard, and mammals we documented the phylogeny-wide abundance of LTR-retrotransposons of retrovirus-like origin [24]. As retroviral elements comprise >99.99% of avian EVEs they obviously represent the most meaningful data set to explore patterns of EVE evolution. This analysis revealed that retroviral EVEs are far less common in birds than in mammals: the average retroviral proportion of the genome was 1.12% (range 0.16% – 3.57%) in birds, 2.39% – 11.41% in mammals, and 0.80% – 4.26% in the genomes of American alligator, green turtle and anole lizard (Additional file 1: Table S6, S7). Strikingly, there was also a simple linear relationship between host genome size and EVE proportion (R^2 = 0.787, P =0.007) (Figure 3). Of equal note was the observation that EVE copy numbers in bird genomes were an order of magnitude less frequent than in mammals (Figure 4; Additional file 1: Table S6, S7), and that the relationship between viral copy number and host genome size exhibited a linear trend (R^2 = 0.780, P <0.001). Importantly, in all cases (i.e., genome size versus proportion and genome size versus copy number) we employed phylogenetic regression analyses to account for the inherent phylogenetic non-independence of the data points.

Figure 3 Relationship between the proportion (%) of retrovirus-like elements in each vertebrate genome and host genome size. The y-axis shows the proportion of LTR-retrotransposons in a variety of vertebrate genomes, while the x-axis indicates genome length (Gb, gigabase). The solid line marks the phylogenetic linear regression for host genome size and the EVE proportion of the genome. Hosts are recognized as follows: hollow circles, birds; black, American alligator, green turtle and anole lizard; grey, mammals.

Figure 4 Copy numbers of retroviral EVEs among birds, American alligator, green turtle, anole lizard, and mammals. Different host groups are colored as red (birds), blue (American alligator, green turtle and anole lizard) and green (mammals). A trend of increasing genome size is also noted. Species are listed from bottom to top in accordance with the bird species order given in Additional file 1: Table S6, and the order among the American alligator, green turtle, anole lizard, and mammals given in Additional file 1: Table S7. *Three oscine passerines showing an EVE expansion.

Discussion and conclusions

Although a diverse array of viruses can possess endogenous forms [2], our analysis revealed that they are uncommon in avian genomes, especially those derived from RNA viruses. Indeed, among RNA viruses, we found only bornavirus endogenized forms occurred in avian genomes, and these had a sporadic distribution and very low frequencies. Although bird genomes are approximately one-third to one-half the size of those of mammals [25,26], the proportion of their genomes that comprises EVEs and their EVE copy numbers are six and 13 times less frequent, respectively. It is generally acknowledged that the genome size reduction associated with flying avian species evolved in the asurischian dinosaur lineage [25]. Our broad-scale genomic screening also suggested that a low frequency of EVEs was an ancestral trait in avian lineage, especially in the case of ERVs, such that there has been an expansion of EVE numbers in mammals concomitant with an increase in their genome sizes. Also of note was that although some genomic integration events in birds were vertical, allowing us to estimate an approximate time-scale for their invasion over many millions of years, by far the

most common evolutionary pattern in the avian data was the independent integration of EVEs into different species/genera.

There are a variety of reasons why EVE numbers could be so relatively low in avian genomes. First, it is theoretically possible that birds have been exposed to fewer viral infections than mammals. However, this seems unlikely as, although they are likely to have been examined less intensively than mammals [27], exogenous viruses of various kinds are found in avian species (e.g., Coronaviridae, Flaviviridae, Hepadnaviridae, Orthomyxoviridae, Paramyxoviridae, Poxviridae, Retroviridae). In addition, the most common phylogenetic pattern we noted was that of independent integration, suggesting the presence of diverse exogenous infections. However, it is notable that mammals apparently harbor a more diverse set of exogenous retroviruses than birds, as well as a greater abundance of ERVs, and which is indicative of a deep-seated evolutionary interaction between host and virus [28]. For example, the only gammaretrovirus known in birds is reticuloendotheliosis virus (REV), and a recent study suggested that avian REVs have a mammalian origin [29]. This is consistent with our observation that there are no endogenized forms of REVs among this diverse set of avian genomes.

It is also possible that birds are in some way refractory to EVE integration following viral infection. ERVs can replicate both as retrotransposons, and as viruses via infection as well reinfection. Although bird cells are known to be susceptible to certain retroviruses [1]), the replication of avian ERVs within the host genome could be suppressed, at least in part, by host-encoded factors. However, a general conclusion of our study is that non-retroviral EVEs are seemingly rare in all vertebrates, such that their integration appears to be generically difficult, and the relative abundance of endogenous retroviruses in birds (albeit low compared to mammals) indicates that they are able to enter bird genomes, with some being actively transcribed and translated [12]. Our observation of a lineage-specific ERV expansion in three passerines also argues against a general refractory mechanism.

A third explanation is that birds are particularly efficient at purging EVEs especially for viruses with retroviral origin from their genomes, a process that we effectively ‘caught in the act’ in the case of the galliform hepadnaviruses. Indeed, our observation of a very low frequency of LTR-retrotransposon in avian genomes may reflect the action of a highly efficient removal mechanism, such as a form of homologous recombination. Hence, it is likely that active genome purging must be responsible for some of the relative absence of EVEs in birds, in turn retaining a selectively advantageous genomic compactness [30]. Clearly, additional work is needed to determine which of these, or other mechanisms, explain the low EVE numbers in avian genomes.

Materials and methods

Genome sequencing and assembly

To systematically study endogenous viral elements in birds, we mined the genomes of 48 avian species (Additional file 1: Table S1). Of these, three genomes – chicken [31], zebra finch [32] and turkey [33] – were downloaded from Ensembl [34]. The remaining genomes were acquired as part of our avian comparative genomics and phylogenomics consortium [13,14]. All genomes can be obtained from our two databases: CoGe [35] and Phylogenomics Analysis of Birds [36]. American alligator, green turtle, anole lizard, and 20 mammal

genomes (Additional file 1: Table S7) were downloaded from Ensembl [34] and used for genomic mining and the subsequent comparative analysis.

Genomic mining

Chromosome and whole genome shotgun assembles [13,34-36] of all species (Additional file 1: Table S1) were downloaded and screened *in silico* using tBLASTn and a library of representative viral protein sequences derived from Groups I to VII (dsDNA, ssDNA, dsRNA, +ssRNA, −ssRNA, ssRNA-RT, and dsDNA-RT) of the 2009 ICTV (International Committee on Taxonomy of Viruses) [37] species list (Additional file 9: Table S8). All viral protein sequences were used for genomic mining. Host genome sequences that generated high-identity (E-values $<1e^{-5}$) matches to viral peptides were extracted. Matches to host proteins were filtered and discarded. The sequences were considered virus-related if they were unambiguously matched viral proteins in the NCBI nr (non-redundant) database [38] and the PFAM database [39]. The putative viral gene structures were inferred using GeneWise [40]. The *in silico* mining of LTR-retrotransposons was performed using RepeatMasker [41].

Phylogenetic inference

To establish the phylogenetic positions of the avian EVEs, particularly in comparison with their exogenous counterparts, we collected all relevant reference viral sequences (Additional file 1: Table S9) from GenBank [42]. Protein sequences (both EVEs and exogenous viruses) were aligned using MUSCLE [43] and checked manually. Phylogenetic trees were inferred using the maximum likelihood (ML) method available in PhyML 3.0 [44], incorporating the best-fit amino acid substitution models determined by ProtTest 3 [45]. The robustness of each node in the tree was determined using 1,000 bootstrap replicates. We subdivided our viral data into 16 categories for phylogenetic analysis (and see Results): 1) endogenous hepadnaviruses, using both complete and partial P (polymerase) protein sequences from positions 429 – 641 (reference sequence DHBV, NC_001344); 2) EBLN, using partial N (nucleoprotein) protein sequences, from positions 43 – 224 (BDV, NC_001607); 3) EBLL, using partial L (RNA-dependent RNA polymerase) protein sequences, from positions 121–656; 4) eCiV Cap, using complete Cap (capsid) protein sequences (GooCiV, NC_003054); 5) eCiV Rep data set 1, using complete Rep (replicase) protein sequences; 6) eCiV Rep data set 2, using partial Rep protein sequences, from positions 160 – 228; 7) eCiV Rep data set 3, using partial Rep protein sequences, from positions 8 – 141; 8) ePaV Cap data set 1, using partial Cap protein sequences, from positions 554 – 650 (DucPaV, NC_006147); 9) ePaV Cap data set 2, using partial Cap protein sequences, from positions 406 – 639; 10) ePaV Cap data set 3, using partial Cap protein sequences, from positions 554 – 695; 11) ePaV Cap data set 4, using partial Cap protein sequences, from positions 662 – 725; 12) ePaV Rep data set 1, using partial Rep protein sequences, from positions 104 – 492; 13) ePaV Rep data set 2, using partial Rep protein sequences, from positions 245 – 383; 14) ePaV Rep data set 3, using partial Rep protein sequences, from positions 300 – 426; 15) ePaV Rep data set 4, using partial Rep protein sequences, from positions 1 – 40; and 16) ERVs, using the retroviral motif "DTGA-YMDD" of Pro-Pol sequences. The best-fit models of amino acid substitution in each case were: 1) JTT + Γ; 2) JTT + Γ; 3) LG + Γ; 4) RtREV + Γ; 5) LG + I + Γ; 6) LG + Γ; 7) LG + I + Γ; 8) LG + Γ; 9) WAG + I + Γ; 10) LG + Γ; 11) LG + Γ; 12) LG + Γ; 13) LG + I + Γ; 14) LG + I + Γ; 15) LG + Γ; and 16) JTT + Γ.

Statistical analysis

To account for the phylogenetic relationships of avian taxa when investigating patterns of EVE evolution we employed phylogenetic linear regression as implemented in R [46]. Specifically, using Mesquite [47] we manually created a tree that matched the host vertebrate phylogeny [14,48]. For the subsequent phylogenetic regression analysis we utilized the 'phylolm' package in R [49], which provides a function for fitting phylogenetic linear regression and phylogenetic logistic regression.

The extent of co-divergence between viruses and hosts was tested by using ParaFit [50], as implemented in the COPYCAT package [51]. The significance of the test was derived from 99,999 randomizations of the association matrix.

Data availability

Data can be accessed by GigaDB [52]. Alternatively, the IDs of NCBI BioProject/SRA/study are as follows:

Chaetura pelagica, PRJNA210808 /SRA092327/SRP026688; *Calypte anna*, PRJNA212866/SRA096094/SRP028275; *Charadrius vociferus*, PRJNA212867/SRA096158/SRP028286; *Corvus brachyrhynchos*, PRJNA212869/SRA096200/SRP028317; *Cuculus canorus*, PRJNA212870/SRA096365/SRP028349; *Manacus vitellinus*, PRJNA212872/SRA096507/SRP028393; *Ophisthocomus hoazin*, PRJNA212873/SRA096539/SRP028409; *Picoides pubescens*, PRJNA212874/SRA097131/SRP028625; *Struthio camelus*, PRJNA212875/SRA097407/SRP028745; *Tinamus guttatus*, PRJNA212876/SRA097796/SRP028753; *Acanthisitta chloris*, PRJNA212877/SRA097960/SRP028832; *Apaloderma vittatum*, PRJNA212878/SRA097967/SRP028834; *Balearica regulorum*, PRJNA212879/SRA097970/SRP028839; *Buceros rhinoceros*, PRJNA212887/SRA097991/SRP028845; *Antrostomus carolinensis*, PRJNA212888/SRA098079/SRP028883; *Cariama cristata*, PRJNA212889/SRA098089/SRP028884; *Cathartes aura*, PRJNA212890/SRA098145/SRP028913; *Chlamydotis macqueenii*, PRJNA212891/SRA098203/SRP028950; *Colius striatus*, PRJNA212892/SRA098342/SRP028965; *Eurypyga helias*, PRJNA212893/SRA098749/SRP029147; *Fulmarus glacialis*, PRJNA212894/SRA098806/SRP029180; *Gavia stellata*, PRJNA212895/SRA098829/SRP029187; *Haliaeetus albicilla*, PRJNA212896/SRA098868/SRP029203; *Haliaeetus leucocephalus*, PRJNA237821/SRX475899, SRX475900, SRX475901, SRX475902/SRP038924; *Leptosomus discolor*, PRJNA212897/SRA098894/SRP029206; *Merops nubicus*, PRJNA212898/SRA099305/SRP029278; *Mesitornis unicolor*, PRJNA212899/SRA099409/SRP029309; *Nestor notabilis*, PRJNA212900/SRA099410/SRP029311; *Pelecanus crispus*, PRJNA212901/SRA099411/SRP029331; *Phaethon lepturus*, PRJNA212902/SRA099412/SRP029342; *Phalacrocorax carbo*, PRJNA212903/SRA099413/SRP029344; *Phoenicopterus ruber*, PRJNA212904/SRA099414/SRP029345; *Podiceps cristatus*,

PRJNA212905/SRA099415/SRP029346; *Pterocles gutturalis*,
PRJNA212906/SRA099416/SRP029347; *Tauraco erythrolophus*,
PRJNA212908/SRA099418/SRP029348; *Tyto alba*,
PRJNA212909/SRA099419/SRP029349; *Nipponia nippon*,
PRJNA232572/SRA122361/SRP035852; *Egretta garzetta*,
PRJNA232959/SRA123137/SRP035853. The following IDs are released before this study: *Aptenodytes forsteri*, PRJNA235982/SRA129317/SRP035855; *Pygoscelis adeliae*,
PRJNA235983/SRA129318/SRP035856; *Gallus gallus*,
PRJNA13342/SRA030184/SRP005856; *Taeniopygia guttata*,
PRJNA17289/SRA010067/SRP001389; *Meleagris gallopavo*,
PRJNA42129/Unknown/Unknown; *Melopsittacus undulatus*/PRJEB1588/ERA200248/ERP002324; *Anas platyrhynchos*,
PRJNA46621/SRA010308/SRP001571; *Columba livia*,
PRJNA167554/SRA054954/SRP013894; *Falco peregrinus*,
PRJNA159791/SRA055082/SRP013939; *Geospiza fortis*,
PRJNA156703/SRA051234/SRP011940.

Abbreviations

ds, Double-stranded; EBLL, Endogenous bornavirus-like L; EBLN, Endogenous bornavirus-like N; eCiV, Endogenous circovirus; ePaV, Endogenous parvovirus; ERV, Endogenous retrovirus; EVE, Endogenous viral element; REV, Reticuloendotheliosis virus; RT, Reverse transcriptase; ss, Single-stranded

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JC, ECH and GZ designed research, which was coordinated by EDJ and MTPG; EDJ, MTPG and GZ provided genome data; JC, WZ, ZH analyzed the data; JC and ECH drafted the complete manuscript, with sections of text contributed by EDJ, MTPG, PJW, and GZ. All authors read and approved the final manuscript.

Acknowledgements

We thank the avian comparative genomics and phylogenomics consortium for providing the avian genomes sequenced. We thank Mang Shi, The University of Sydney, and Cai Li, BGI-Shenzhen, for statistical advice. ECH is supported by an NHMRC Australia Fellowship and by NIH grant R01 GM080533. We thank two reviewers for informative comments.

References

1. Weiss RA: **The discovery of endogenous retroviruses.** *Retrovirology* 2006, **3:**67.

2. Katzourakis A, Gifford RJ: **Endogenous viral elements in animal genomes.** *PLoS Genet* 2010, **6:**e1001191.

3. Kazazian HH Jr: **Mobile elements: drivers of genome evolution.** *Science* 2004, **303:**1626–1632.

4. Jern P, Coffin JM: **Effects of retroviruses on host genome function.** *Annu Rev Genet* 2008, **42:**709–732.

5. Emerman M, Malik HS: **Paleovirology–modern consequences of ancient viruses.** *PLoS Biol* 2010, **8:**e1000301.

6. Feschotte C, Gilbert C: **Endogenous viruses: insights into viral evolution and impact on host biology.** *Nat Rev Genet* 2012, **13:**283–296.

7. Stoye JP: **Studies of endogenous retroviruses reveal a continuing evolutionary saga.** *Nat Rev Microbiol* 2012, **10:**395–406.

8. Herniou E, Martin J, Miller K, Cook J, Wilkinson M, Tristem M: **Retroviral diversity and distribution in vertebrates.** *J Virol* 1997, **71:**437–443.

9. Gilbert C, Feschotte C: **Genomic fossils calibrate the long-term evolution of hepadnaviruses.** *PLoS Biol* 2010, **8:**e1000495.

10. Cui J, Holmes EC: **Endogenous Hepadnaviruses in the genome of the budgerigar (*Melopsittacus undulatus*) and the evolution of avian hepadnaviruses.** *J Virol* 2012, **86:**7688–7691.

11. Suh A, Brosius J, Schmitz J, Kriegs JO: **The genome of a Mesozoic paleovirus reveals the evolution of hepatitis B viruses.** *Nat Commun* 2013, **4:**1791.

12. Bolisetty M, Blomberg J, Benachenhou F, Sperber G, Beemon K: **Unexpected diversity and expression of avian endogenous retroviruses.** *mBio* 2012, **3:**e00344-12.

13. Zhang G, Li C, Li Q, Li B, Larkin DM, Lee C, Storz JF, Antunes A, Greenwold MJ, Meredith RW, Ödeen A, Cui J, Zhou Q, Xu L, Pan H, Wang Z, Jin L, Zhang P, Hu H, Yang W, Hu J, Xiao J, Yang Z, Liu Y, Xie Q, Yu H, Lian J, Wen P, Zhang F, Li H, *et al*: **Comparative genomics across modern bird species reveal insights into avian genome evolution and adaptation.** *Science* 2014, Accepted.

14. Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SYW, Faircloth BC, Nabholz B, Howard JT, Suh A, Weber CC, da Fonseca RR, Li J, Zhang F, Li H, Zhou L, Narula N, Liu L, Ganapathy G, Boussau B, Bayzid MS, Zavidovych V, Subramanian S, Gabaldón T, Capella-Gutiérrez S, Huerta-Cepas J, Rekepalli B, Munch K, Schierup M, *et al*: **Whole genome analyses resolve the early branches in the tree of life of modern birds.** *Science* 2014, Accepted.

15. Robertson BH, Margolis HS: **Primate hepatitis B viruses - genetic diversity, geography and evolution.** *Rev Med Virol* 2002, **12:**133–141.

16. Hackett SJ, Kimball RT, Reddy S, Bowie RC, Braun EL, Braun MJ, Chojnowski JL, Cox WA, Han KL, Harshman J, Huddleston CJ, Marks BD, Miglia KJ, Moore WS, Sheldon FH, Steadman DW, Witt CC, Yuri T: **A phylogenomic study of birds reveals their evolutionary history.** *Science* 2008, **320:**1763–1768.

17. Horie M, Honda T, Suzuki Y, Kobayashi Y, Daito T, Oshida T, Ikuta K, Jern P, Gojobori T, Coffin JM, Tomonaga K: **Endogenous non-retroviral RNA virus elements in mammalian genomes.** *Nature* 2010, **463:**84–87.

18. Belyi VA, Levine AJ, Skalka AM: **Unexpected inheritance: multiple integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate genomes.** *PLoS Pathog* 2010, **6:**e1001030.

19. de la Torre JC: **Molecular biology of borna disease virus: prototype of a new group of animal viruses.** *J Virol* 1994, **68:**7669–7675.

20. VandeWoude S, Richt JA, Zink MC, Rott R, Narayan O, Clements JE: **A borna virus cDNA encoding a protein recognized by antibodies in humans with behavioral diseases.** *Science* 1990, **250:**1278–1281.

21. Holmes EC: **The evolution of endogenous viral elements.** *Cell Host Microbe* 2011, **10:**368–377.

22. Belyi VA, Levine AJ, Skalka AM: **Sequences from ancestral single-stranded DNA viruses in vertebrate genomes: the parvoviridae and circoviridae are more than 40 to 50 million years old.** *J Virol* 2010, **84:**12458–12462.

23. Lehmann HW, von Landenberg P, Modrow S: **Parvovirus B19 infection and autoimmune disease.** *Autoimmun Rev* 2003, **2:**218–223.

24. Finnegan DJ: **Retrotransposons.** *Curr Biol* 2012, **22:**R432–R437.

25. Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV: **Origin of avian genome size and structure in non-avian dinosaurs.** *Nature* 2007, **446:**180–184.

26. **Animal genome size database.** http://www.genomesize.com/.

27. Lipkin WI: **The changing face of pathogen discovery and surveillance.** *Nat Rev Microbiol* 2013, **11:**133–141.

28. Cui J, Tachedjian M, Wang L, Tachedjian G, Wang LF, Zhang S: **Discovery of retroviral homologs in bats: implications for the origin of mammalian gammaretroviruses.** *J Virol* 2012, **86:**4288–4293.

29. Niewiadomska AM, Gifford RJ: **The extraordinary evolutionary history of the reticuloendotheliosis viruses.** *PLoS Biol* 2013, **11:**e1001642.

30. Griffin DK, Robertson LB, Tempest HG, Skinner BM: **The evolution of the avian genome as revealed by comparative molecular cytogenetics.** *Cytogenet Genome Res* 2007, **117:**64–77.

31. International Chicken Genome Sequencing Consortium: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432:**695–716.

32. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, Searle S, White S, Vilella AJ, Fairley S, Heger A, Kong L, Ponting CP, Jarvis ED, Mello CV, Minx P, Lovell P, Velho TA, Ferris M, Balakrishnan CN, Sinha S, Blatti C, London SE, Li Y, Lin YC, George J, Sweedler J, Southey B, Gunaratne P, Watson M, *et al*: **The genome of a songbird.** *Nature* 2010, **464:**757–762.

33. Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Blomberg Le A, Bouffard P, Burt DW, Crasta O, Crooijmans RP, Cooper K, Coulombe RA, De S, Delany ME, Dodgson JB, Dong JJ, Evans C, Frederickson KM, Flicek P, Florea L, Folkerts O, Groenen MA, Harkins TT, Herrero J, Hoffmann S, Megens HJ, Jiang A, de Jong P, Kaiser P, Kim H, *et al*: **Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis.** *PLoS Biol* 2010, **8:**e1000475.

34. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, García-Girón C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kähäri AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, *et al*: **Ensembl 2013.** *Nucleic Acids Res* 2013, **41:**D48–D55.

35. **CoGe database.** http://genomevolution.org/CoGe/.

36. **Phylogenomics analysis of birds.** http://phybirds.genomics.org.cn/.

37. King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ: *Virus Taxonomy: Classification and Nomenclature of Viruses: Ninth Report of the International Committee on Taxonomy of Viruses.* San Diego: Elsevier Academic Press; 2012.

38. **RefSeq: NCBI reference sequence database.** http://www.ncbi.nlm.nih.gov/refseq/.

39. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A: **The Pfam protein families database.** *Nucleic Acids Res* 2010, **38:**D211–D222.

40. Birney E, Clamp M, Durbin R: **GeneWise and genomewise.** *Genome Res* 2004, **14:**988–995.

41. Smit AFA, Hubley R, Green P: **RepeatMasker Open-3.0. (1996–2010).** http://www.repeatmasker.org.

42. Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2013, **41:**D36–D42.

43. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32:**1792–1797.

44. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0.** *Syst Biol* 2010, **59:**307–321.

45. Darriba D, Taboada GL, Doallo R, Posada D: **ProtTest 3: fast selection of best-fit models of protein evolution.** *Bioinformatics* 2011, **27:**1164–1165.

46. **The R project.** http://www.r-project.org.

47. Maddison WP, Maddison DR: **Mesquite: a modular system for evolutionary analysis. Version 2.75.** 2011. http://mesquiteproject.org.

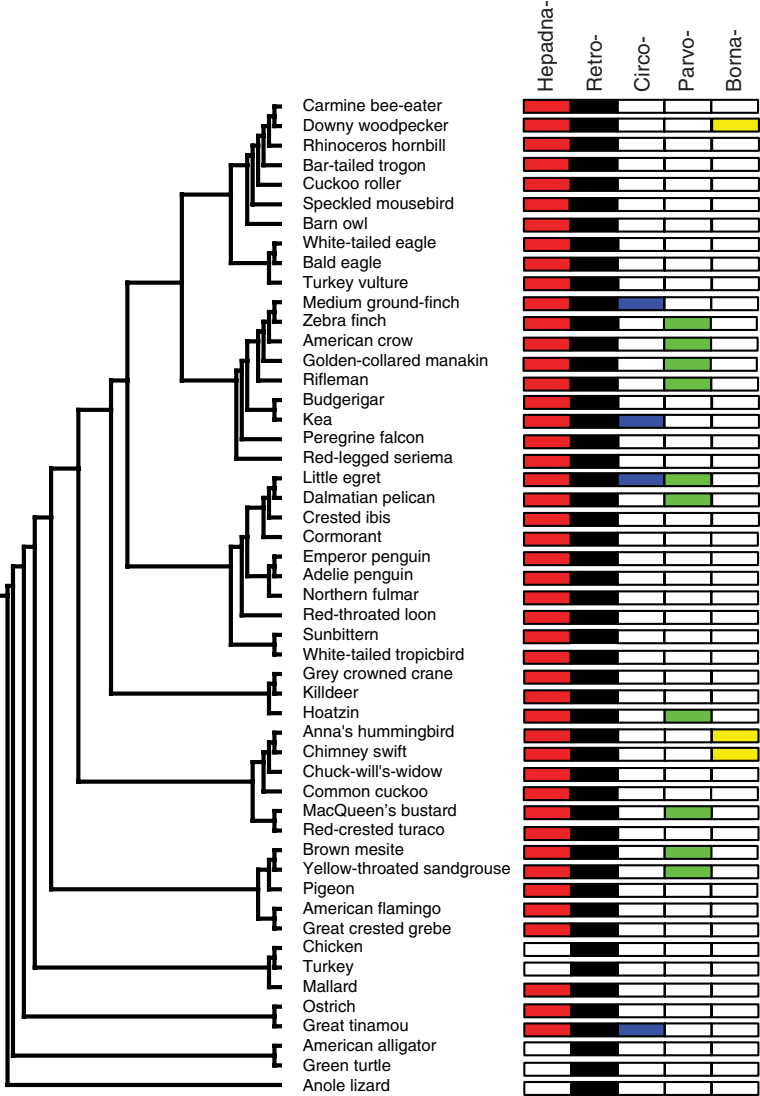
48. Lindblad-Toh K, Garber M, Zuk O, Lin MF, Parker BJ, Washietl S, Kheradpour P, Ernst J, Jordan G, Mauceli E, Ward LD, Lowe CB, Holloway AK, Clamp M, Gnerre S, Alföldi J, Beal K, Chang J, Clawson H, Cuff J, Di Palma F, Fitzgerald S, Flicek P, Guttman M, Hubisz MJ, Jaffe DB, Jungreis I, Kent WJ, Kostka D, Lara M, *et al*: **A high-resolution map of human evolutionary constraint using 29 mammals.** *Nature* 2011, **478:**476–482.

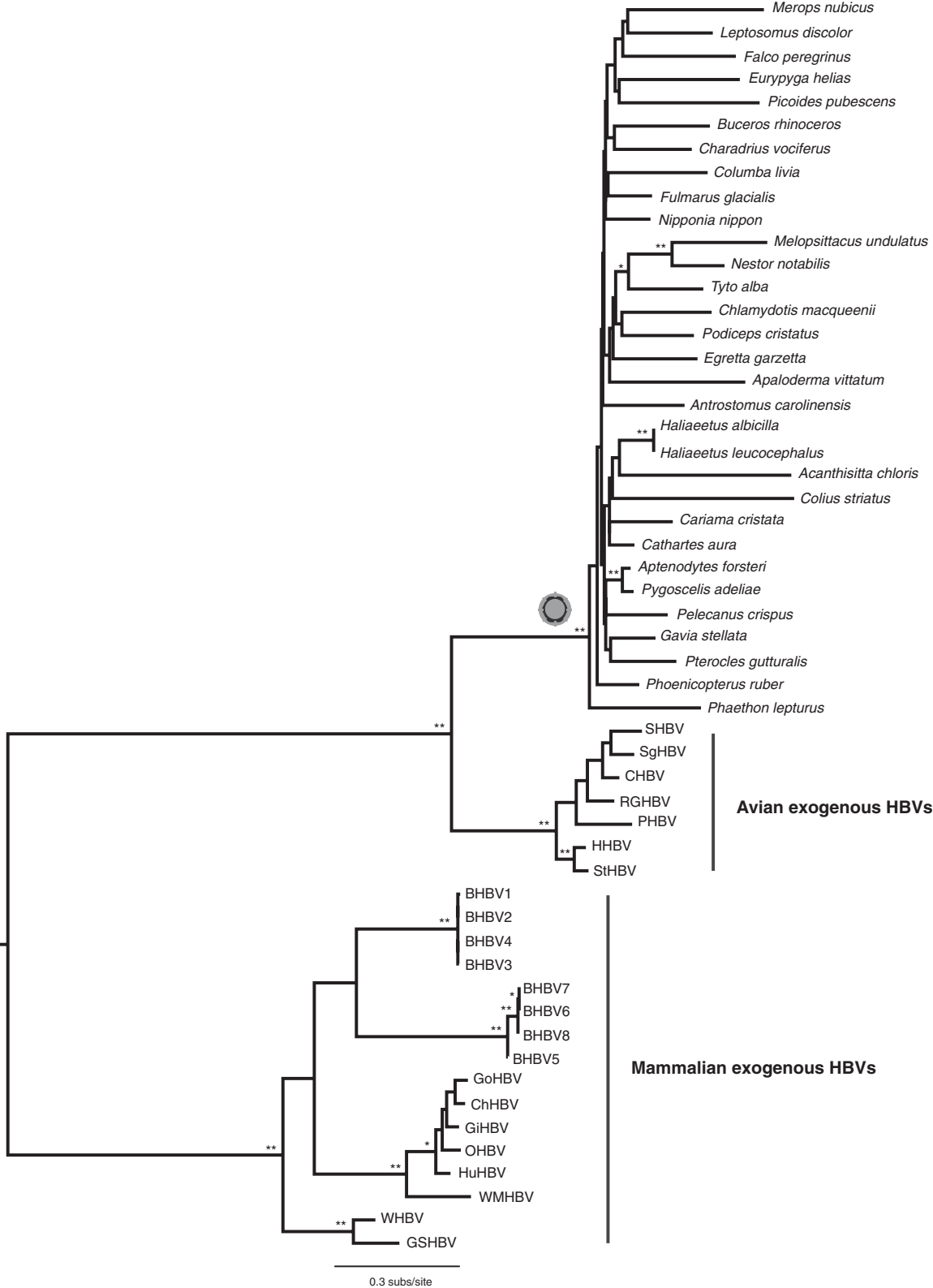
49. Ho LST, Ané C: **A linear-time algorithm for Gaussian and non-Gaussian trait evolution models.** *Syst Biol* 2013. In press.

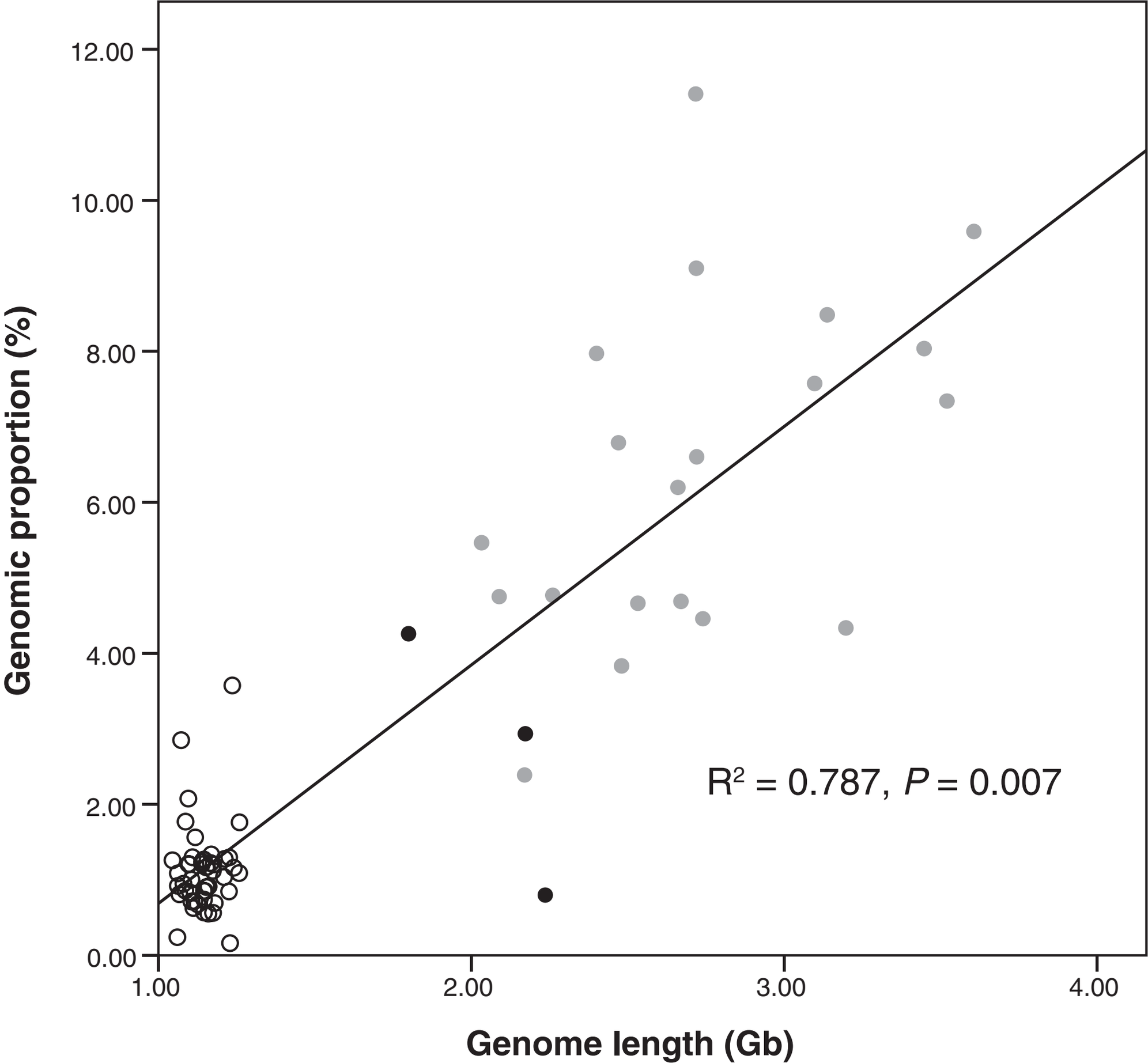
50. Legendre P, Desdevises Y, Bazin E: **A statistical test for host-parasite coevolution.** *Syst Biol* 2002, **51:**217–234.

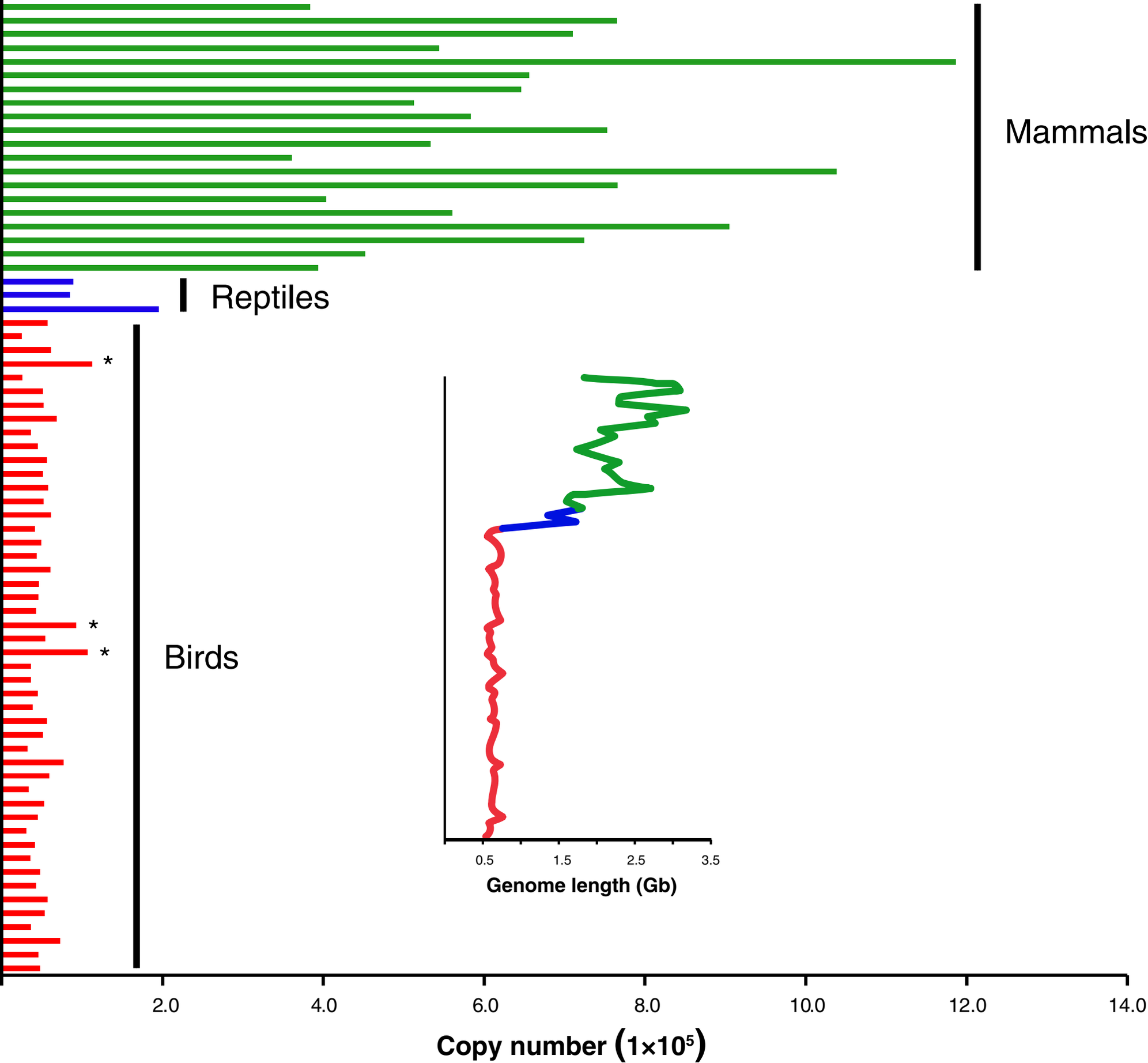
51. Meier-Kolthoff JP, Auch AF, Huson DH, Göker M: **COPYCAT: cophylogenetic analysis tool.** *Bioinformatics* 2007, **23:**898–900.

52. **The avian phylogenomic project data.** http://gigadb.org/dataset/101000.


Hepadna-
Retro-
Circo-
Parvo-
Borna-
Carmine bee-eater
Downy woodpecker
Rhinoceros hornbill
Bar-tailed trogon
Cuckoo roller
Speckled mousebird
Barn owl
White-tailed eagle
Bald eagle
Turkey vulture
Medium ground-finch
Zebra finch
American crow
Golden-collared manakin
Rifleman
Budgerigar
Kea
Peregrine falcon
Red-legged seriema
Little egret
Dalmatian pelican
Crested ibis
Cormorant
Emperor penguin
Adelie penguin
Northern fulmar
Red-throated loon
Sunbittern
White-tailed tropicbird
Grey crowned crane
Killdeer
Hoatzin
Anna's hummingbird
Chimney swift
Chuck-will's-widow
Common cuckoo
MacQueen's bustard
Red-crested turaco
Brown mesite
Yellow-throated sandgrouse
Pigeon
American flamingo
Great crested grebe
Chicken
Turkey
Mallard
Ostrich
Great tinamou
American alligator
Green turtle
Anole lizard


Merops nubicus
Leptosomus discolor
Falco peregrinus
Eurypyga helias
Picoides pubescens
Buceros rhinoceros
Charadrius vociferus
Columba livia
Fulmarus glacialis
Nipponia nippon
Melopsittacus undulatus
Nestor notabilis
Tyto alba
Chlamydotis macqueenii
Podiceps cristatus
Egretta garzetta
Apaloderma vittatum
Antrostomus carolinensis
Haliaeetus albicilla
Haliaeetus leucocephalus
Acanthisitta chloris
Colius striatus
Cariama cristata
Cathartes aura
Aptenodytes forsteri
Pygoscelis adeliae
Pelecanus crispus
Gavia stellata
Pterocles gutturalis
Phoenicopterus ruber
Phaethon lepturus
SHBV
SgHBV
CHBV
RGHBV
PHBV
HHBV
StHBV
Avian exogenous HBVs
BHBV1
BHBV2
BHBV4
BHBV3
BHBV7
BHBV6
BHBV8
BHBV5
GoHBV
ChHBV
GiHBV
OHBV
HuHBV
WMHBV
WHBV
GSHBV
Mammalian exogenous HBVs
0.3 subs/site


12.00
10.00
8.00
6.00
4.00
2.00
0.00
Genomic proportion (%)
1.00
2.00
3.00
4.00
Genome length (Gb)
R² = 0.787, P = 0.007


Mammals
Reptiles
Birds
*
*
*
0.5
1.5
2.5
3.5
Genome length (Gb)
2.0
4.0
6.0
8.0
10.0
12.0
14.0
Copy number (1×10^5)

Addtional files provided with this submission:

Additional file 1. Table S1. Avian genomes used for genomic mining. Table S2. Endogenous hepadnaviruses in avian genomes. Table S3. Endogenous bornaviruses in avian genomes. Table S4. Endogenous circoviruses in avian genomes. Table S5. Endogenous parvoviruses in avian genomes. Table S6. LTR-retrotransposon composition of avian genomes. Table S7. LTR-retrotransposon composition of American alligator, green turtle, anole lizard and mammalian genomes. Table S9. Reference sequences used for phylogenetic analyses (76k)
http://genomebiology.com/supplementary/s13059-014-0539-3-s1.docx

Additional file 2: Figure S1. Phylogenetic tree of endogenous retroviruses (ERVs). The tree was inferred using the conserved motif "DTGA-YMDD" within the Pro-Pol region of retroviruses (~320 amino acids in length, although this differs among retrovirus genera). Bootstrap values lower than 70% are not shown; one star (*) represents values higher than 70%, while two stars (**) represents values higher than 90%. Branch lengths are drawn to a scale of amino acid substitutions per site (subs/site). The tree is midpoint rooted for purposes of clarity only. The host name indicates the species from which the ERV was obtained. Exogenous retroviruses are highlighted using family names. ERVs of alligator, turtle and lizard origin are also highlighted (636k)
http://genomebiology.com/supplementary/s13059-014-0539-3-s2.pdf

Additional file 3: Figure S2. Phylogenetic tree of exogenous and endogenous avian hepadnaviruses. Bootstrap values lower than 70% are not shown; one star (*) represents values higher than 70%, while two stars (**) represents values higher than 90%. Branch lengths are drawn to a scale of amino acid substitutions per site (subs/site). The tree is midpoint rooted for purposes of clarity only. The exogenous hepadnaviruses are highlighted. Avian host species names are used to denote avian endogenous hepadnaviruses, and different EVEs from the same host are numbered. All abbreviations are provided in Additional file 1: Table S9 (416k)
http://genomebiology.com/supplementary/s13059-014-0539-3-s3.pdf

Additional file 4: Data S1. Alignments of the orthologous hepadnaviral scaffolds (179k)
http://genomebiology.com/supplementary/s13059-014-0539-3-s4.docx

Additional file 5: Figure S3. Alignment of a hepadnaviral element in the genome of mallard duck with orthologous (and partial) sequences found in the genomes of chicken and turkey. Note that we found a 94% match to the 5' conserved region (marked as C) in turkey, and a 39% match to the orthologous chicken sequence; 45% of the central 12,042-bp virus-like sequence matched the 5' variable region (marked as V). The relatively conserved nucleotides in chicken showing virus-like characteristics are boxed. Asterisks represent the conserved nucleotides in the alignment, dashes denote deletions (257k)
http://genomebiology.com/supplementary/s13059-014-0539-3-s5.pdf

Additional file 6: Figure S4. Phylogenetic trees of endogenous and exogenous bornaviruses. The phylogenies contain (A) endogenous bornavirus-like N (nucleoprotein) (EBLN) and (B) avian endogenous bornavirus-like L (RNA-dependent RNA polymerase) (EBLL) sequences. Bootstrap values lower than 70% are not shown; one star (*) represents values higher than 70%, while two stars (**) represents values higher than 90%. Branch lengths are drawn to a scale of amino acid substitutions per site (subs/site). The trees are midpoint rooted for purposes of clarity only. Avian host species names for those that harbor EVEs are given in parentheses and different EVEs from the same host are numbered. All abbreviations are provided in Additional file 1: Table S9 (202k)
http://genomebiology.com/supplementary/s13059-014-0539-3-s6.pdf

Additional file 7: Figure S5. Phylogenetic trees of endogenous circoviruses. The phylogenies contain avian endogenous circoviruses (eCiVs) Cap (A) and Rep (B, C and D). Bootstrap values lower than 70% are not shown; one star (*) represents values higher than 70%, while two stars (**) represents values higher than 90%. Branch lengths are drawn to a scale of amino acid substitutions per site (subs/site). The trees are midpoint rooted for purposes of clarity only. Avian host species names for those that harbor EVEs are given in parentheses. All abbreviations are provided in Additional file 1: Table S9 (259k)
http://genomebiology.com/supplementary/s13059-014-0539-3-s7.pdf

Additional file 8: Figure S6. Phylogenetic trees of endogenous and exogenous parvoviruses. The phylogenies contain avian endogenous parvoviruses (ePaVs) Cap (A, B, C, and D) and Rep (E, F, G, and H). Bootstrap values lower than 70% are not shown; one star (*) represents values higher than 70%, while two stars (**) represents values higher than 90%. Branch lengths are drawn to a scale of amino acid substitutions per site (subs/site). The trees are midpoint rooted for purposes of clarity only. Avian host species names for those that harbor EVEs are given in parentheses and different EVEs from the same host are numbered. All abbreviations are provided in Additional file 1: Table S9 (405k)
http://genomebiology.com/supplementary/s13059-014-0539-3-s8.pdf

Additional file 9: Table S8. Reference viral sequences used for genomic searching (263k)
http://genomebiology.com/supplementary/s13059-014-0539-3-s9.xls