1 **One hundred million years history of bornavirus infections hidden in vertebrate**
2 **genomes**

3
4 Junna Kawasaki^{1,2}, Shohei Kojima¹, Yahiro Mukai^{1,2}, Keizo Tomonaga^{1,2,3}, Masayuki
5 Horie^{1,4*}

6
7 ¹ Laboratory of RNA Viruses, Department of Virus Research, Institute for Frontier Life
8 and Medical Sciences, Kyoto University, Kyoto, Japan

9 ² Laboratory of RNA Viruses, Department of Mammalian Regulatory Network,
10 Graduate School of Biostudies, Kyoto University, Kyoto, Japan

11 ³ Department of Molecular Virology, Graduate School of Medicine, Kyoto University,
12 Kyoto, Japan

13 ⁴ Hakubi Center for Advanced Research, Kyoto University, Kyoto, Japan

14
15 *** Corresponding Author**

16 Masayuki Horie, DVM, PhD

17 Hakubi Center for Advanced Research, Kyoto University

18 53 Kawahara-cho, Shogo-in, Sakyo, Kyoto 606-8507, Japan

19 Email: horie.masayuki.3m@kyoto-u.ac.jp

20

21 **Abstract**

22 Although viruses have threatened our ancestors for millions of years, prehistoric
 23 epidemics of viruses are largely unknown. Endogenous bornavirus-like elements
 24 (EBLs) are ancient viral sequences that have been integrated into animal genomes.
 25 These elements can be used as molecular fossil records to trace past bornaviral
 26 infections. In this study, we systematically identified EBLs in vertebrate genomes and
 27 revealed the history of bornavirus infections over nearly 100 million years. We found
 28 that ancient bornaviral infections have occurred in diverse vertebrate lineages,
 29 especially in primate ancestors. Phylogenetic analyses indicated that primate ancestors
 30 had been infected with various bornaviral lineages during evolution. Moreover, EBLs in
 31 primate genomes formed clades according to their integration ages, suggesting that
 32 epidemic lineages of bornaviruses had changed chronologically. However, we found
 33 that some bornaviral lineages coexisted with primate ancestors and underwent repeated
 34 endogenizations for tens of millions of years. Furthermore, this viral lineage that
 35 coexisted with primate ancestors was also endogenized in some ancestral bats. Notably,
 36 the geographic distributions of these bat ancestors have been reported to overlap with
 37 the migration route of primate ancestors, suggesting that long-term virus-host
 38 coexistence could have expanded the geographic distributions of the viral lineage and
 39 might have spread their infections to new hosts. Thus, our findings describe hidden
 40 virus-host co-evolutionary history over geological timescales, including chronological
 41 change in epidemic bornaviral lineages, long-term virus-host coexistence, and
 42 expansion of viral infections.

43

44 **Introduction**

45 Viral infectious diseases profoundly affect human health, livestock productivity,
46 and ecosystem diversity. Similar to recent viral outbreaks [1], our ancestors were also
47 probably challenged by viral epidemics. Investigations of past viruses using historical
48 specimens have provided insights into the origin and spread of viral infections [2-6].
49 However, the epidemic history of viruses across hundreds of millions of years is largely
50 unclear.

51 Endogenous viral elements (EVEs) are formed by the occasional integration of
52 ancient viral sequences into the host germline genomes [7]. EVEs are millions of years
53 old and provide critical information on ancient viruses, such as their host ranges [8, 9],
54 evolutionary timescales [10], or geographical distributions [11]. Endogenous
55 bornavirus-like elements (EBLs) are the most abundant viral fossils of RNA viruses
56 found in vertebrate genomes [12, 13]. Therefore, EBLs could help us trace the epidemic
57 history of bornaviral infections on geological timescales and are good model systems to
58 study long-term virus-host co-evolutionary history.

59 The family of *Bornaviridae* consists of three genera, *Orthobornavirus*,
60 *Carbovirus*, and *Cultervirus* [14]. Until 2018, only the genus *Bornavirus* (today
61 *Orthobornavirus*), which includes viruses that cause immune-mediated neurological
62 diseases in mammals and birds, constituted the family *Bornaviridae* [15]. However,
63 two new genera, *Carbovirus* and *Cultervirus*, were established upon discovering novel
64 bornaviral species in carpet pythons and sharpbelly fish samples, respectively [16, 17].
65 Furthermore, these discoveries also led to the identification of novel EBLs classified
66 into these bornaviral genera [16]. Since most previous studies have exclusively used
67 orthobornaviruses for EBL detection [10, 12, 13, 18-20], numerous EBLs may remain

68 to be detected and analyzed. Therefore, current understanding of the epidemic history of
69 ancient bornaviruses is probably incomplete.

70 In this study, we searched for EBLs derived from three bornaviral genera and
71 characterized the evolutionary timescale, host range, geographical distribution, and
72 genetic diversity of ancient bornaviruses to reconstruct the long-term history of their
73 infections. Large-scale dating analysis revealed that ancient bornaviral infections have
74 occurred in diverse vertebrate lineages for nearly 100 million years. Primate ancestors,
75 in particular, had been repeatedly infected with ancient bornaviruses. Phylogenetic
76 analyses of EBLs in primate genomes showed clustering according to their integration
77 ages, suggesting that epidemic lineages of bornaviruses in primate ancestors have
78 changed chronologically. Furthermore, some bornaviral lineages may have coexisted
79 with primate ancestors for tens of millions of years. Interestingly, we found that
80 long-term virus-host coexistence could have expanded the geographic distributions of
81 the virus and generated new infections in some bat ancestors. Thus, our findings
82 describe virus-host coevolutionary history over geological timescales, which cannot be
83 deduced from research using extant viruses.

84

85 **Results**

86 **Systematic identification of EBLs in the host genomes**

87 To systematically identify EBLs, we searched for bornavirus-like sequences in
88 the genomic data of 969 eukaryotic species by tBLASTn using bornaviral protein
89 sequences from all genera in the family *Bornaviridae* as queries (**Fig 1A**). Next, we
90 concatenated these sequences based on their location in the host genome and their
91 alignment positions to extant bornaviral proteins because most bornavirus-like
92 sequences were fragmented due to mutation after endogenization.

93 The bornaviral genome encodes 6 viral proteins: nucleoprotein (N),
94 phosphoprotein (P), matrix protein (M), envelope glycoprotein (G), large
95 RNA-dependent RNA polymerase (L), and accessory protein (X) [21]. EBLs are
96 considered to have been formed by integrations of bornaviral mRNA into the host
97 genome; EBLs derived from N, M, G, and L genes, designated EBLN, EBLM, EBLG,
98 and EBLL, respectively, have been reported so far [13]. Our EBL search identified
99 1,465 EBLs in 131 vertebrate species, including 1,079 EBLNs, 30 EBLPs, 46 EBLMs,
100 195 EBLGs, and 115 EBLLs (**Fig 1B and 1C**). Notably, we identified EBLPs in animal
101 genomes for the first time, although these were considered difficult to detect due to
102 methodological limitations, such as low sequence conservation of P genes among
103 bornaviruses [22].

104 We next classified ancient bornaviruses from which EBLs originated. Since
105 some EBLs were too short to construct reliable phylogenetic trees, we sought to classify
106 all the EBLs based on sequence similarity scores in the tBLASTn search. However, if
107 there is an unknown genus consisting exclusively of ancient bornaviruses, the method
108 based on sequence similarity with modern viruses may lead to misclassification. To

109 assess this possibility, we first performed phylogenetic analyses using relatively long
 110 EBLs and their gene counterparts in modern bornaviruses. **Fig 1D** shows that EBLs
 111 were clearly divided into three clades corresponding to the current bornaviral
 112 classification [14]. Therefore, we decided to apply the sequence similarity-based
 113 method for classifying all the EBLs into current bornaviral genera (**details in Materials**
 114 **and Methods**). Based on the similarity scores, the EBLs were classified into 364
 115 orthobornaviral, 729 carboviral, and 372 culterviral EBLs (**Fig 1C**). Among 1465 EBLs,
 116 870 loci were undetectable by a search using orthobornaviral sequences that have been
 117 almost exclusively used for EBL detection as queries. Therefore, the EBL search found
 118 numerous previously undetected loci and created a comprehensive dataset for
 119 reconstructing the history of bornavirus infections.

120

121 **Large-scale dating analysis for bornaviral integration ages**

122 EVE integration ages can be estimated based on the gene orthology [7] , and
 123 thus we sought to determine the presence and absence patterns of orthologous EBLs.
 124 Here, we developed a network-based method to handle the large datasets (**S1 Fig**).
 125 Briefly, we first constructed an all-against-all matrix of alignment coverages by
 126 pairwise sequence comparison among EBL integration sites. Next, we constructed a
 127 sequence similarity network using the matrix and extracted community structures from
 128 the network in order to divide the EBLs into groups based on their orthologous
 129 relationships. Finally, we manually checked the groupings to avoid inaccurate estimates
 130 (**details in Materials and Method and S2 Fig**).

131 We divided 1,465 EBLs into 281 groups by our network-based dating method
 132 (**S1 Table**). These groupings reflected the alignment coverage among EBL integration

133 sites (**S1 Fig**). We divided these groups into two categories: 113 groups of "EBLs with
134 orthologs" and 168 groups of "EBLs without orthologs" (**Fig 2A**). "EBLs with
135 orthologs" share the same bornaviral integration in each group, while each group of
136 "EBLs without orthologs" consists of a single EBL locus. Such "EBLs without
137 orthologs" might be young integrations that occurred after the divergence of hosts from
138 their sister species. Alternatively, the lack of orthologs may simply be a methodological
139 limitation due to the inaccessibility of the genomic data of sister species. For example,
140 only three distantly related species of Eulipotyphla, in which no EBL orthologous
141 relationships could be detected, were present in our database (**Fig 2A**). This suggests
142 that accumulating genomic data could help estimate integration ages with higher
143 accuracy.

144

145 **Bornaviral infections have occurred since the Mesozoic era**

146 Our dating analysis allowed us to trace the history of bornavirus infections
147 back to 100 million years ago (MYA) (**Fig 2A**). The oldest records of bornaviral
148 infection have been reported in ancestral afrotherians at least 83.3 MYA [10, 20]. Here,
149 we found six EBLs that were orthologous among species of Boreoeutheria, suggesting
150 that the oldest bornavirus infections had occurred at least 96.5 MYA. Additionally, we
151 identified 18 bornaviral integration events in the Mesozoic era, which occurred in the
152 ancestors of Afrotheria, Tethytheria, Metatheria, Primates, and Rodentia. Besides, we
153 found the first record of bornavirus infection in Mesozoic birds in the ancestor of
154 Passeriformes at least 66.6 MYA. Thus, these results provide strong evidence that
155 bornaviral infections had already occurred in multiple vertebrate lineages in the
156 Mesozoic era.

157

158 **Ancient bornaviral infections in various vertebrate lineages**

159 We found that ancient bornaviruses infected much broader vertebrate lineages
160 than modern bornaviruses are known to infect (**Fig 2A**). Modern orthobornavirus
161 infections have been reported in ungulate animals, shrews, squirrels, humans, a wide
162 range of birds, and garter snakes [15]. However, we identified multiple endogenizations
163 of ancient orthobornaviruses in mice, afrotherians, and marsupials, which have not been
164 reported as host species of modern orthobornaviruses (**Fig 2A**). In particular, a previous
165 survey of bornaviral reservoirs did not detect orthobornavirus infections in mice [23].
166 Furthermore, the extant carboviruses and cultervirus were detected only in carpet
167 pythons and sharpbelly fish, respectively [16, 17]. In contrast, ancient viruses belonging
168 to these genera endogenized in various host lineages, including mammals and birds (**Fig**
169 **2A**). These results indicate that ancient bornaviruses infected a wider range of
170 vertebrate lineages than known extant bornaviruses.

171

172 **The geographical distributions of ancient bornaviral infections**

173 We performed integrative analysis of bornaviral endogenizations and
174 mammalian biogeography dynamics to infer the geographical distributions of ancient
175 bornavirus infections (**S2 Table**). Our results suggest that ancient bornavirus infections
176 occurred in different continents: Laurasia and Africa in the Mesozoic era, Antarctica or
177 Australia around the K-Pg boundary, and possibly Eurasia, Africa, or South America in
178 the Cenozoic era (**Fig 2B**).

179 First, we identified EBLs in animals that inhabited Laurasia and Africa in the
180 Mesozoic era (**N1, N2, N4, and N6 in Fig 2B**). It has been reported that ancestors of

181 Boreoeutheria and Primates were distributed in Laurasia [24-27], while those of
 182 Afrotheria and Tethytheria were found in Africa [24, 25, 28]. These results suggest that
 183 bornaviral infections might have spread in Laurasia and Africa during the Cretaceous
 184 period. Second, we identified EBLs integrated into the genome of Australidelphia
 185 ancestors, but not in other marsupials in South America (**N8 in Fig 2**). Since ancestral
 186 Australian marsupials are considered to have moved from South America to Australia
 187 via Antarctica [29, 30], bornavirus infections are likely to have occurred in Antarctica
 188 or Australia.

189 Furthermore, we identified bornaviral endogenizations that occurred in
 190 ancestral primates in the Cenozoic era. First, we found EBLs integrated into the genome
 191 of the ancestor of the Madagascar lemur (**N10 in Fig 2**); however, EBLs were not
 192 identified in African galagos, suggesting that bornavirus infections occurred in
 193 Madagascar Island. On the other hand, EBL integration age was estimated at 37.8-59.3
 194 MYA, which overlapped with the migration of lemur ancestors from Africa to
 195 Madagascar Island around 50-60 MYA [31]. This overlap presents an alternate
 196 possibility that bornaviral endogenization had occurred in African animals before they
 197 migrated to Madagascar. Second, we identified several EBLs integrated into the
 198 ancestral Platyrrhini genomes (**N11 in Fig 2**). Ancestors of Platyrrhini are presumed to
 199 have migrated from Africa to South America during their divergence from Simiiformes
 200 [26, 32], thus providing evidence of bornavirus infections in these continents (**see**
 201 **Discussion**). Taken together, these results suggest worldwide occurrence of ancient
 202 bornavirus infections.

203

204 **The complex history of bornaviral infections during primate evolution: epidemics** 205 **of distinct bornaviral lineages in each age**

206 We found that bornaviruses in the three genera repeatedly endogenized during
207 primate evolution (**Fig 2A**). We inferred phylogenetic relationships among ancient
208 bornaviruses in each genus using EBLNs that are the most abundant records among
209 EBLs (**Fig 3A-C and S3 Fig**) in order to understand the origin of endogenizations of
210 bornaviral lineages into primate ancestor genomes.

211 Consequently, we found several distinct lineages of bornaviruses that had
212 sequentially endogenized during primate evolution, rather than a single bornaviral
213 lineage that had repeatedly endogenized. For example, the carboviral EBLNs in the
214 primate genome were clearly divided into two viral lineages: the clade 1 viral lineages
215 endogenized in Boreoeutherian ancestors and the clade 2 viral lineages endogenized in
216 Simiiformes and Catarrhini ancestors (**Fig 3A and 3D**). Furthermore, orthobornaviral
217 EBLNs formed three different clades according to their integration ages (**clades 3 to 5**
218 **in Fig 3B and 3D**). These results suggest that different bornaviral lineages were
219 prevalent during primate evolution across different eras.

220 Next, to infer how diverse bornaviruses have endogenized during primate
221 evolution, we calculated the genetic distances between these ancient viral lineages
222 (clades 1 to 5) in our phylogenetic tree (**S3 Table**). Using genetic distance as a
223 comparative standard for classifying extant species of bornaviruses, we found that the
224 genetic diversity among these ancient bornaviral lineages was higher than that among
225 extant bornaviral species (**Fig 3A-C and S3 Table**). Thus, we infer that recurrent
226 bornaviral endogenizations during primate evolution occurred due to infections of
227 multiple bornaviral lineages comparable to different viral species.

228

229 **The complex history of bornaviral infections during primate evolution: long-term**
 230 **virus-host coexistence**

231 In addition to the sequential infections of primate ancestors by distinct
 232 bornavirus lineages (**Fig 3**), we found that some lineages might have established
 233 long-term coexistence with the hosts. For example, the clade 2 carboviral lineage has
 234 repeatedly endogenized in Simiiformes and Catarrhini ancestors between 29.4 and 67.1
 235 MYA (**Fig 3A and 3D**). Furthermore, endogenizations of the clade 5 orthobornaviral
 236 lineage have recurred in Simiiformes and Platyrrhini ancestors between 19.7 and 67.1
 237 MYA (**Fig 3B and 3D**). These results suggested that these bornaviral lineages have
 238 coexisted with primate ancestors for tens of millions of years. In summary, we
 239 described the complex history of recurrent bornaviral endogenizations during primate
 240 evolution, including sequential infections of diverse bornaviral lineages and long-term
 241 virus-host coexistence.

242

243 Discussion

244 Snapshots of ancient bornaviral infections have been reported since the
245 discovery of EBLs [10, 12, 13, 16, 18-20]; however, the long-term history of bornavirus
246 infections has remained unclear. Here, we systematically identified EBLs in 131
247 vertebrate species (**Fig 1**) and reconstructed the epidemic history of bornaviral
248 infections for approximately 100 million years (**Fig 2**). To our knowledge, this is the
249 first report to comprehensively trace the history of RNA virus infections over geological
250 timescales. Furthermore, phylogenetic analyses suggested differences in the epidemic
251 lineages of bornaviruses during primate evolution across different geological ages as
252 well as coexistence of some lineages with ancestral primate ancestors for tens of
253 millions of years (**Fig 3**). Virus-host co-divergence alone, which is thought to be as the
254 background of viral evolutionary history [17], is insufficient to explain this mixed
255 pattern. Therefore, our findings suggested that the virus-host coevolutionary
256 relationships had been dramatically changed over geological timescales, which
257 complicated the viral evolutionary history.

258 We also found that various host lineages might have been infected by
259 phylogenetically related bornaviruses in each geological age (**Fig 3**). Although EBLs
260 are the most abundant RNA virus fossils, it is not easy to trace the details of viral
261 transmission because EBLs have only rarely been fossilized. Nonetheless, our
262 phylogenetic analyses showed that bornaviruses closely related to the lineage that
263 caused epidemics in ancestral primates had almost contemporaneously endogenized in
264 the other animals (**Fig 3 and S3 Fig**). For example, carboviruses similar to the clade 1
265 viral lineage had also endogenized in ancestral afrotherians around the late Mesozoic
266 era (**EBLN41, 49, 63, and 66 in Fig 3A**). Additionally, carboviruses similar to the

clade 2 lineage endogenized in ancestors of Yangochiroptera bats from the late Mesozoic to Cenozoic era (**EBLN59 in Fig 3A**). A similar tendency was observed in the orthobornaviral phylogenetic tree and extant viruses as well: genetically similar orthobornaviruses infect various host species in mammals, birds, and reptiles at present (**Fig 3B**). These results suggest that bornaviral lineages have spread to various hosts in each era, and epidemic lineages of bornaviruses have changed over time.

By integrating information on host geographical distributions and phylogeny of bornaviruses, we found long-term coexistence of ancient bornaviruses with primate ancestors that could have expanded the viral lineage to other continents. **Fig 3B** shows that the clade 5 orthobornaviral lineage has repeatedly endogenized in ancestors of Simiiformes and Platyrrhini. Ancestors of Simiiformes were reportedly distributed in Eurasia or Africa, while Platyrrhini ancestors likely migrated from Africa to South America during their divergence from Simiiformes (**Fig 2B**) [26, 32]. These results suggested that the clade 5 viral lineage could have moved between the continents along with host migrations. Interestingly, viruses in the clade 5 lineage also endogenized in several bat genomes (**Fig 3B**), such as *Rhinolophus* bats (EBLN61), *Desmodus* bats (EBLN106 and EBLN124), and *Miniopterus* bats (EBLN122 and EBLN127). Since *Rhinolophus* and *Desmodus* bats have reportedly originated in Eurasia and South America, respectively [33], the clade 5 viral lineage may have moved across continents along with primate migrations and been further transmitted to these bat ancestors. Thus, long-term virus-host coexistence could have expanded the viral geographic distributions and generated new infections in other hosts.

We found that the number of bornaviral integration events varied according to the host lineage (**Fig 2A**). In non-mammalian vertebrates, we observed low frequencies

291 of bornaviral integrations, consistent with previous studies [34, 35]. Furthermore, the
 292 numbers of bornaviral endogenizations differ among descendant lineages that diverged
 293 from boreoeutherian ancestors. Remarkably, bornaviral endogenizations rarely occurred
 294 in most of the laurasiatherian lineages, but repeatedly occurred in some other host
 295 lineages, including Primates, Chiroptera, and Rodentia. Factors such as viral infection,
 296 germ-line integration, and inheritance may be responsible for these differences. Further
 297 studies on the impact of these factors on EBL fixations are necessary to explain such
 298 differences.

299 This study also proposed that a large number of EVEs derived from unknown
 300 ancient viruses may be hidden in the host genomic data. One of the problems with
 301 paleovirological research is the methodological limitation for identifying viral fossil
 302 records comprehensively. Because the method to identify EVEs involves searching for
 303 virus-like sequences in the host genome using modern viral sequences as queries, the
 304 detectability of EVEs depends on sequence similarities with extant viruses. Here, we
 305 demonstrated that the EBL search using genetically diverse extant bornaviruses
 306 provided a large dataset, including previously undetectable loci (**Fig 1**). Therefore,
 307 further elucidation of extant viral diversity could help in clarifying ancient viral
 308 diversity.

309 Furthermore, the sequence data of EBLs could be a useful resource for
 310 exploring extant viral diversity because metagenomic analyses to detect viral infections
 311 also rely on sequence similarities with known viral sequences. Our phylogenetic
 312 analyses indicated that EBLNs originated from diverse ancient bornaviruses, and almost
 313 all EBLNs formed clades completely different from modern ones (**Fig 3 and S3 Fig**).
 314 Hence, ancient bornaviruses appear highly divergent and phylogenetically distinct from

315 known modern viruses. Furthermore, these results raise a fascinating question regarding
 316 whether viruses genetically similar to EBLs are extinct or just yet to be discovered.
 317 Future viral metagenomic analyses using EBL sequence data may address this question.
 318 Therefore, reusing data between metagenomic analyses for extant viruses and
 319 paleovirological investigations could elucidate viral diversity and connect modern with
 320 ancient viral evolution.

321 In conclusion, we traced bornaviral infections over geological timescales and
 322 depicted the epidemic history of bornaviruses during vertebrate evolution. Our findings
 323 provide novel insights into coevolutionary history between viruses and hosts, which
 324 cannot be deduced from research using extant viruses.

325

326 **Materials and Methods**

327 *Identification of EBLs in vertebrate genomic data*

328 EBLs were identified by: (1) searching for bornavirus-like sequences in
329 genomes of 969 eukaryotic species, (2) reconstructing EBL sequences, and (3)
330 validating whether these EBL sequences were derived from ancient bornaviruses or not.

331 First, bornavirus-like sequences were screened in Refseq genomic database
332 (version: 20190329) provided from NCBI [36] by tBLASTn (version 2.6.0+) [37] with
333 the option “-evalue 0.1” using sequences in all genus of Bornaviridae as queries (**S4**
334 **Table**). Second, because most EBLs were detected as fragmented sequences due to
335 mutations occurring after integration, we reconstructed EBL sequences by
336 concatenating sequences if the following conditions were met: (1) detected
337 bornavirus-like sequences were located within 1000 bp (EBLN, EBLP, EBLM, and
338 EBLG) or 2000 bp (EBLL), and (2) the order of sequences in the alignment with extant
339 bornaviral proteins were consistent with those in the host genome (**Fig 1A**). When more
340 than two bornavirus-like sequences were detected in the same genomic position, we
341 preferentially used the sequence with higher reliability (low E-value in the tBLASTn
342 search). The alignment of bornavirus-like sequences and modern bornaviral proteins
343 was conducted using MAFFT (version 7.427) with options “--addfragment” and
344 “--keeplengths” [38]. Finally, we checked the origin of the EBL candidates based on the
345 bit score obtained from BLASTP (version 2.9.0+) using the Refseq protein database
346 (version: 20200313) and a database consisting of bornaviral protein sequences listed in
347 **S4 Table**. If the candidate was more similar to the host proteins than published EBLs or
348 other viral proteins, we considered the sequence a false positive and removed it from the
349 analysis. After this process, only one EBLL candidate was identified in the insect

genome, but we excluded this sequence in subsequent analyses. The concatenation of
bornavirus-like sequences yielded over 800 EBL loci equivalent to more than half the
length of the intact bornaviral proteins (**Fig 1B**).

Dating analysis for the integrated age of EBLs

To determine orthologous relationships among EBLs, we sought to cluster loci
based on the alignment coverages in pairwise sequence comparison between EBL
integration sites (**S1 Fig**). First, we extracted the upstream and downstream sequences
of EBLs with lengths of 15 kbp for EBLs and 10 kbp for other EBLs. These sequences
were trimmed by removing repetitive elements using RepeatMasker (version
open-4.0.9) (<http://repeatmasker.org>) with option “-q xsmall -a -species” and RepBase
RepeatMasker libraries (version 20181026) [39]. Second, we performed the pairwise
alignment of these sequences using BLASTN (version 2.9.0+) and constructed an
all-against-all matrix for alignment coverage among EBL integration sites. The
sequence similarity network was constructed by connecting nodes when their sequence
alignment coverage was over 9.0% of the flanking sequence length. Selection of the
best criteria to construct a sequence network is described in the next section. The groups
were extracted by detecting a community structure using Louvain heuristics. These
network analyses were performed using NetworkX [40].

We simultaneously checked the phylogenetic relationships of host species with
sequence alignment coverage to correctly estimate EBL ages. The contamination of
sequences unrelated to true orthologous relationships leads to overestimation of
integration ages, as shown in **example 4 in S4 Fig**. To avoid such issues, when multiple
EBL loci were present in the same species genome and alignment coverage was lower

374 than 50%, we considered these loci as located in different genomic sites and divided
375 them into different groups. Furthermore, the integration ages of older elements tend to
376 be underestimated because the alignment quality among their integration sites may
377 deteriorate due to the accumulation of sequence changes, such as genomic
378 rearrangement (**example 3 in S4 Fig**). Thus, by checking the phylogenetic relationships
379 of host species and sequence alignment coverage, we combined some groups into
380 EBLG2, EBLL2, EBLL35, or EBLL36 (**S1 Fig and S2 Fig**). For example, the EBLG2
381 group was previously reported to have endogenized into the genome of laurasiatherian
382 ancestors at least 77.0 MYA [16]. This group was divided into two groups in the initial
383 analysis, including primate and laurasiatherian loci. However, these groups were
384 connected by low alignment coverage, which led to another hypothesis that these
385 sequences are descendants of the same integration event in the boreoeutherian ancestor
386 (**S1 Fig**). To test this hypothesis, we confirmed the alignment quality among the EBL
387 integration sites using AliTV [41-43]. We found that over 70% sequence similarity
388 covered more than 40% of the alignment by lastz (version 1.04.00) [42] with options
389 “--noytrim, --gapped, and --strand=both” (**S2 Fig**). Therefore, we combined these
390 groups into the same group. The cases of EBLL2 and EBLL35 were similar to that of
391 EBLG2 (**S2 Fig**). EBLL36 contained tandemly repeated loci at close genomic locations
392 (**S1 Table**) because we could not distinguish whether these loci were derived from
393 independent integration events or gene duplications post-integration. Presently, we
394 considered these loci as descendants from the same integration event to avoid
395 overestimating the number of EBL integration events.

396 After curation, the dates of EBL integration events were assigned according to
397 a vertebrate evolutionary tree provided from the TimeTree database [44]. Each EBL

locus was named according to the nomenclature for endogenous retroviruses [45] (**S1 Table**). It should be noted that the number of bornaviral integration events was less than that of EBL loci shown in **Fig 1C** because redundant sequences in the genomic database used for the EBL search were grouped as the same integration event.

Validation of the network-based dating method using human transposable elements

To validate our dating method, we compared the integration ages of human transposable elements (TEs) estimated based on genomic alignment and those estimated based on network analysis (**S4 Fig**). The genomic positions of all human TEs were obtained from the RepeatMasker database (<http://www.repeatmasker.org>). First, the orthologs of all human TEs were determined in 18 mammalian genomes in LiftOver (version 357) with the option “-minMatch=0.5” using genomic alignments provided by the University of California Santa Cruz (UCSC) genome browser [46]. Their integration ages were determined by the presence and absence patterns of orthologs. Second, to examine the ortholog detection rate by the network-based dating method for each timescale, we prepared test datasets by random sampling of 100 loci for each timescale based on the dating results using genomic alignment (**S4 Fig**). The details of the network-based dating method are described in the previous section. The estimation of human TE integration ages in the test datasets followed the same strategy, except that: (1) the flanked sequence of human TEs were extracted with lengths of 10 kbp from the soft-masked assembly sequence provided by the UCSC genome browser, and (2) the UCSC genome browser procedure detected repetitive elements. Finally, we compared the results between the two methods by checking the following points: predicted ages and detected orthologs (**examples are shown in S4 Fig**).

Furthermore, we tried nine different criteria to connect nodes in the sequence similarity network (**S4 Fig**) and decided to connect network edges if their sequence alignment coverage was over 9.0% of the flanking sequence length (**S4 Fig**). The concordant rates between the two methods in predicting ages of Cenozoic TEs were 88.0-100.0% according to this criterion, and those of older TEs integrated in the Cretaceous period were 43.0-66.0% (**S4 Fig**). The chain files used for LiftOver and the genome assembly sequences are listed in **S5 Table**.

Phylogenetic analysis

We used amino acid sequences of EBLs with lengths longer than 200 amino acids for EBLN and 100 for EBLG. Multiple sequence alignments (MSAs) were constructed by MAFFT with options "--addfragment" and "--keeplengths." MSAs for EBL classification (**Fig 1D**) were trimmed by excluding sites where over 30% of sequences were gaps, subsequently removing sequences with less than 70% of the total alignment sites. MSA for the EBLN tree (**Fig 3 and S3 Fig**) was trimmed by excluding sites where over 20% of sequences were gaps, and subsequently removing sequences with less than 80% of the total alignment sites. Phylogenetic trees were constructed by the maximum likelihood method using IQTREE (version 1.6.12) [47]. The substitution models were selected based on the Bayesian information criterion score provided by ModelFinder [48]: VT+F+G4 for EBLNs, VT+F+G4 for EBLGs, VT+F+R3 for EBLs (**Fig 1D**), and JTT+F+G4 for EBLNs (**Fig 3 and S3 Fig**). The branch supports were measured as the ultrafast bootstrap values given by UFBoot2 [49] with 1000 replicates. The extant viral sequences used for the phylogenetic analyses are listed in **S6 Table**. We used ggtree [50] and ete3 packages [51] for the visualization of trees.

446

447 ***EBL classification according to current bornaviral genera***

448 EBLs were classified into current bornaviral genera based on the query
449 bornaviral sequence with the lowest E-value in the tBLASTn search. The results of the
450 phylogenetic analysis-based method and the similarity score-based method were highly
451 concordant (EBLN: 99.7%, EBLG: 100.0%, and EBL: 100.0%). Thus, we applied the
452 classification method for all EBL loci (**Fig 1C**). We could not create reliable
453 phylogenetic trees using EBLP or EBLM due to the small number of sites available for
454 phylogenetic analysis.

455

456 ***Assessment of genetic diversity of ancient bornaviral sequences***

457 We compared the genetic diversity of ancient and extant bornaviruses to infer
458 how diverse bornaviruses have endogenized during primate evolution. First, we used the
459 most recent common ancestor of the EBLN orthologs as the ancestral bornaviral N gene
460 to avoid overestimating the sequence diversity of ancient bornaviruses. Next, to provide
461 a comparison standard for interpreting the ancient bornaviral genetic diversity, we
462 calculated the genetic distance for classifying extant bornaviral species in our
463 phylogenetic tree (0.06 substitutions per site) (**S3 Table**). The genetic distances between
464 nodes in the phylogenetic tree were calculated using the ete3 toolkit. It should be noted
465 that this is an alternative method, and ICTV classification for extant bornaviral species
466 is based on the sequence similarity among intact viral genomes, differences in their host
467 ranges, and phylogenetic analysis using viral proteins.

468

469 **Data Availability**

470 The codes are available at https://github.com/Junna-Kawasaki/EBL_2020. The
471 versions of bioinformatics tools are listed in **S7 Table**.

472

473 **Acknowledgments**

474 We thank Dr. Keiko Takemoto (Institute for Virus Research, Kyoto University,
475 Japan) for technical support. We are grateful to Jumpei Ito (Institute of Medical Science,
476 the University of Tokyo, Japan), Bea Clarise Garcia, Lin Hsien Hen, Koichi Kitao, and
477 Michiko Iwata (Institute for Frontier Life and Medical Sciences, Kyoto University) for
478 helpful discussions.

479 This study was supported by JSPS KAKENHI JP19J2224 (JK); and
480 JP18K19443 (MH); MEXT KAKENHI JP17H05821 (MH) and JP19H04833 (MH);
481 Hakubi project at Kyoto University (MH). Computations were partially performed on
482 the supercomputing systems SHIROKANE (Human Genome Center, the Institute of
483 Medical Science, the University of Tokyo) and the NIG supercomputer (ROIS National
484 Institute of Genetics).

485

486 **Author contributions**

487 MH conceived the study; JK mainly performed bioinformatics analyses; MH,
488 SK, and YM supported bioinformatics analyses; JK prepared the figures and wrote the
489 initial draft of the manuscript; all authors contributed designed the study, interpreted
490 data, revised the paper, and approved the final manuscript.

491

492 **Competing interests**

493 The authors declare that they have no competing interests.

494

495 **References**

- 496 1. Grubaugh ND, Ladner JT, Lemey P, Pybus OG, Rambaut A, Holmes EC, et al.
497 Tracking virus outbreaks in the twenty-first century. *Nat Microbiol.* 2019;4(1):10-9.
498 Epub 2018/12/14. doi: 10.1038/s41564-018-0296-2. PubMed PMID: 30546099;
499 PubMed Central PMCID: PMCPMC6345516.
- 500 2. Dx A, Lequime S, Patrono LV, Vrancken B, Boral S, Gogarten JF, et al.
501 Measles virus and rinderpest virus divergence dated to the sixth century BCE. *Science.*
502 2020;368(6497):1367-70. doi: 10.1126/science.aba9411.
- 503 3. Taubenberger JK. Initial Genetic Characterization of the 1918 "Spanish"
504 Influenza Virus. *Science.* 1997;275(5307):1793-6. doi: 10.1126/science.275.5307.1793.
- 505 4. Mhlemann B, Margaryan A, Damgaard PDB, Allentoft ME, Vinner L, Hansen
506 AJ, et al. Ancient human parvovirus B19 in Eurasia reveals its long-term association
507 with humans. *Proceedings of the National Academy of Sciences.* 2018;115(29):7557-62.
508 doi: 10.1073/pnas.1804921115.
- 509 5. Duggan AT, Perdomo MF, Piombino-Mascoli D, Marciniak S, Poinar D, Emery
510 MV, et al. 17 th Century Variola Virus Reveals the Recent History of Smallpox. *Current*
511 *Biology.* 2016;26(24):3407-12. doi: 10.1016/j.cub.2016.10.061.
- 512 6. Mhlemann B, Jones TC, Damgaard PDB, Allentoft ME, Shevnina I, Logvin A,
513 et al. Ancient hepatitis B viruses from the Bronze Age to the Medieval period. *Nature.*
514 2018;557(7705):418-23. doi: 10.1038/s41586-018-0097-z.
- 515 7. Aiewsakun P, Katzourakis A. Endogenous viruses: Connecting recent and
516 ancient viral evolution. *Virology.* 2015;479-480:26-37. doi: 10.1016/j.virol.2015.02.011.

- 517 8. Kryukov K, Ueda MT, Imanishi T, Nakagawa S. Systematic survey of
518 non-retroviral virus-like elements in eukaryotic genomes. *Virus Res.* 2019;262:30-6.
519 Epub 2018/02/10. doi: 10.1016/j.virusres.2018.02.002. PubMed PMID: 29425804.
- 520 9. Hayward A, Cornwallis CK, Jern P. Pan-vertebrate comparative genomics
521 unmasks retrovirus macroevolution. *Proc Natl Acad Sci U S A.* 2015;112(2):464-9.
522 Epub 2014/12/24. doi: 10.1073/pnas.1414980112. PubMed PMID: 25535393; PubMed
523 Central PMCID: PMC4299219.
- 524 10. Katzourakis A, Gifford RJ. Endogenous viral elements in animal genomes.
525 *PLoS Genet.* 2010;6(11):e1001191. Epub 2010/12/03. doi:
526 10.1371/journal.pgen.1001191. PubMed PMID: 21124940; PubMed Central PMCID:
527 PMCPMC2987831.
- 528 11. Gifford RJ, Katzourakis A, Tristem M, Pybus OG, Winters M, Shafer RW. A
529 transitional endogenous lentivirus from the genome of a basal primate and implications
530 for lentivirus evolution. *Proceedings of the National Academy of Sciences.*
531 2008;105(51):20362-7. doi: 10.1073/pnas.0807873105.
- 532 12. Horie M, Honda T, Suzuki Y, Kobayashi Y, Daito T, Oshida T, et al.
533 Endogenous non-retroviral RNA virus elements in mammalian genomes. *Nature.*
534 2010;463(7277):84-7. doi: 10.1038/nature08695.
- 535 13. Horie M, Kobayashi Y, Suzuki Y, Tomonaga K. Comprehensive analysis of
536 endogenous bornavirus-like elements in eukaryote genomes. *Philosophical Transactions*
537 *of the Royal Society B: Biological Sciences.* 2013;368(1626):20120499. doi:
538 10.1098/rstb.2012.0499.

- 539 14. Amarasinghe GK, Ayllón MA, Bào Y, Basler CF, Bavari S, Blasdel KR, et al.
540 Taxonomy of the order Mononegavirales: update 2019. Archives of Virology.
541 2019;164(7):1967-80. doi: 10.1007/s00705-019-04247-4.
- 542 15. Kuhn JH, Durrwald R, Bao Y, Briese T, Carbone K, Clawson AN, et al.
543 Taxonomic reorganization of the family Bornaviridae. Arch Virol. 2015;160(2):621-32.
544 Epub 2014/12/03. doi: 10.1007/s00705-014-2276-z. PubMed PMID: 25449305;
545 PubMed Central PMCID: PMC4315759.
- 546 16. Hyndman TH, Shilton CM, Stenglein MD, Wellehan JFX. Divergent
547 bornaviruses from Australian carpet pythons with neurological disease date the origin of
548 extant Bornaviridae prior to the end-Cretaceous extinction. PLOS Pathogens.
549 2018;14(2):e1006881. doi: 10.1371/journal.ppat.1006881.
- 550 17. Shi M, Lin X-D, Chen X, Tian J-H, Chen L-J, Li K, et al. The evolutionary
551 history of vertebrate RNA viruses. Nature. 2018;556(7700):197-202. doi:
552 10.1038/s41586-018-0012-7.
- 553 18. Belyi VA, Levine AJ, Skalka AM. Unexpected inheritance: multiple
554 integrations of ancient bornavirus and ebolavirus/marburgvirus sequences in vertebrate
555 genomes. PLoS Pathog. 2010;6(7):e1001030. Epub 2010/08/06. doi:
556 10.1371/journal.ppat.1001030. PubMed PMID: 20686665; PubMed Central PMCID:
557 PMC2912400.
- 558 19. Mukai Y, Horie M, Tomonaga K. Systematic estimation of insertion dates of
559 endogenous bornavirus-like elements in vesper bats. Journal of Veterinary Medical
560 Science. 2018;80(8):1356-63. doi: 10.1292/jvms.18-0211.

- 561 20. Kobayashi Y, Horie M, Nakano A, Murata K, Itou T, Suzuki Y. Exaptation of
562 Bornavirus-Like Nucleoprotein Elements in Afrotherians. PLOS Pathogens.
563 2016;12(8):e1005785. doi: 10.1371/journal.ppat.1005785.
- 564 21. Briese T, Schneemann A, Lewis AJ, Park YS, Kim S, Ludwig H, et al.
565 Genomic organization of Borna disease virus. 1994;91(10):4362-6. doi:
566 10.1073/pnas.91.10.4362.
- 567 22. Horie M, Tomonaga K. Paleovirology of bornaviruses: What can be learned
568 from molecular fossils of bornaviruses. Virus Research. 2019;262:2-9. doi:
569 10.1016/j.virusres.2018.04.006.
- 570 23. Hilbe M, Herrsche R, Kolodziejek J, Nowotny N, Zlinszky K, Ehrensperger F.
571 Shrews as Reservoir Hosts of Borna Disease Virus. Emerging Infectious Diseases.
572 2006;12(4):675-7. doi: 10.3201/eid1204.051418.
- 573 24. Springer MS, Meredith RW, Janecka JE, Murphy WJ. The historical
574 biogeography of Mammalia. Philos Trans R Soc Lond B Biol Sci.
575 2011;366(1577):2478-502. Epub 2011/08/03. doi: 10.1098/rstb.2011.0023. PubMed
576 PMID: 21807730; PubMed Central PMCID: PMC3138613.
- 577 25. Nishihara H, Maruyama S, Okada N. Retroposon analysis and recent
578 geological data suggest near-simultaneous divergence of the three superorders of
579 mammals. Proceedings of the National Academy of Sciences. 2009;106(13):5235-40.
580 doi: 10.1073/pnas.0809297106.
- 581 26. Springer MS, Meredith RW, Gatesy J, Emerling CA, Park J, Rabosky DL, et al.
582 Macroevolutionary Dynamics and Historical Biogeography of Primate Diversification
583 Inferred from a Species Supermatrix. PLoS ONE. 2012;7(11):e49521. doi:
584 10.1371/journal.pone.0049521.

- 585 27. Bloch JJ, Silcox MT, Boyer DM, Sargis EJ. New Paleocene skeletons and the
586 relationship of plesiadapiforms to crown-clade primates. *Proceedings of the National*
587 *Academy of Sciences*. 2007;104(4):1159-64. doi: 10.1073/pnas.0610579104.
- 588 28. Gheerbrant E, Schmitt A, Kocsis L. Early African Fossils Elucidate the Origin
589 of Embryothod Mammals. *Current Biology*. 2018;28(13):2167-73.e2. doi:
590 10.1016/j.cub.2018.05.032.
- 591 29. Nilsson MA, Churakov G, Sommer M, Tran NV, Zemmann A, Brosius J, et al.
592 Tracking Marsupial Evolution Using Archaic Genomic Retroposon Insertions. *PLoS*
593 *Biology*. 2010;8(7):e1000436. doi: 10.1371/journal.pbio.1000436.
- 594 30. Eldridge MDB, Beck RMD, Croft DA, Travouillon KJ, Fox BJ. An emerging
595 consensus in the evolution, phylogeny, and systematics of marsupials and their fossil
596 relatives (Metatheria). *Journal of Mammalogy*. 2019;100(3):802-37. doi:
597 10.1093/jmammal/gyz018.
- 598 31. Poux C, Madsen O, Marquard E, Vieites D, De Jong W, Vences M.
599 Asynchronous Colonization of Madagascar by the Four Endemic Clades of Primates,
600 Tenrecs, Carnivores, and Rodents as Inferred from Nuclear Genes. 2005;54(5):719-30.
601 doi: 10.1080/10635150500234534.
- 602 32. Jaeger JJ. PALEONTOLOGY: Shaking the Earliest Branches of Anthropoid
603 Primate Evolution. *Science*. 2005;310(5746):244-5. doi: 10.1126/science.1118124.
- 604 33. Teeling EC. A Molecular Phylogeny for Bats Illuminates Biogeography and the
605 Fossil Record. *Science*. 2005;307(5709):580-4. doi: 10.1126/science.1105113.
- 606 34. Cui J, Zhao W, Huang Z, Jarvis ED, Gilbert MTP, Walker PJ, et al. Low
607 frequency of paleoviral infiltration across the avian phylogeny. *Genome Biology*.
608 2014;15(12). doi: 10.1186/s13059-014-0539-3.

- 609 35. Gilbert C, Meik JM, Dashevsky D, Card DC, Castoe TA, Schaack S.
610 Endogenous hepadnaviruses, bornaviruses and circoviruses in snakes. *Proc Biol Sci.*
611 2014;281(1791):20141122. Epub 2014/08/01. doi: 10.1098/rspb.2014.1122. PubMed
612 PMID: 25080342; PubMed Central PMCID: PMC4132678.
- 613 36. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al.
614 Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion,
615 and functional annotation. *Nucleic Acids Research.* 2016;44(D1):D733-D45. doi:
616 10.1093/nar/gkv1189.
- 617 37. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al.
618 BLAST+: architecture and applications. *BMC Bioinformatics.* 2009;10(1):421. doi:
619 10.1186/1471-2105-10-421.
- 620 38. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software
621 Version 7: Improvements in Performance and Usability. *Molecular Biology and*
622 *Evolution.* 2013;30(4):772-80. doi: 10.1093/molbev/mst010.
- 623 39. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive
624 elements in eukaryotic genomes. *Mobile DNA.* 2015;6(1). doi:
625 10.1186/s13100-015-0041-9.
- 626 40. Aric A. Hagberg DASaPJS, editor Exploring Network Structure, Dynamics,
627 and Function using NetworkX. *Proceedings of the 7th Python in Science Conference*
628 *(SciPy2008)*; 2008 Aug; Pasadena, CA USA.
- 629 41. Ankenbrand MJ, Hohlfield S, Hackl T, Förster F. AliTV—interactive
630 visualization of whole genome comparisons. *PeerJ Computer Science.* 2017;3. doi:
631 10.7717/peerj-cs.116.

632 42. Harris RS. Improved pairwise alignment of genomic dna: Pennsylvania State
633 University; 2007.

634 43. Stajich JE. The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome*
635 *Research*. 2002;12(10):1611-8. doi: 10.1101/gr.361602.

636 44. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for
637 Timelines, Timetrees, and Divergence Times. *Mol Biol Evol*. 2017;34(7):1812-9. Epub
638 2017/04/08. doi: 10.1093/molbev/msx116. PubMed PMID: 28387841.

639 45. Gifford RJ, Blomberg J, Coffin JM, Fan H, Heidmann T, Mayer J, et al.
640 Nomenclature for endogenous retrovirus (ERV) loci. *Retrovirology*. 2018;15(1):59.
641 Epub 2018/08/30. doi: 10.1186/s12977-018-0442-1. PubMed PMID: 30153831;
642 PubMed Central PMCID: PMC6114882.

643 46. Haeussler M, Zweig AS, Tyner C, Speir ML, Rosenbloom KR, Raney BJ, et al.
644 The UCSC Genome Browser database: 2019 update. *Nucleic Acids Research*.
645 2019;47(D1):D853-D8. doi: 10.1093/nar/gky1095.

646 47. Nguyen L-T, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A Fast and
647 Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.
648 *Molecular Biology and Evolution*. 2015;32(1):268-74. doi: 10.1093/molbev/msu300.

649 48. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermini LS.
650 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*.
651 2017;14(6):587-9. doi: 10.1038/nmeth.4285.

652 49. Hoang DT, Chernomor O, Von Haeseler A, Minh BQ, Vinh LS. UFBoot2:
653 Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*.
654 2018;35(2):518-22. doi: 10.1093/molbev/msx281.

655 50. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. ggtree : an r package for
656 visualization and annotation of phylogenetic trees with their covariates and other
657 associated data. *Methods in Ecology and Evolution*. 2017;8(1):28-36. doi:
658 10.1111/2041-210x.12628.

659 51. Huerta-Cepas J, Serra F, Bork P. ETE 3: Reconstruction, Analysis, and
660 Visualization of Phylogenomic Data. *Mol Biol Evol*. 2016;33(6):1635-8. Epub
661 2016/02/28. doi: 10.1093/molbev/msw046. PubMed PMID: 26921390; PubMed Central
662 PMCID: PMC4868116.

663

664 **Figure captions**

665 **Fig. 1. Identification of EBLs in the vertebrate genomes. (A)** Schematic diagram for
666 the procedure to identify EBLs. First, bornavirus-like sequences were detected from
667 host genomes by tBLASTn using extant bornaviral sequences as queries. Second, the
668 detected bornavirus-like sequences were aligned with corresponding proteins of extant
669 bornaviruses. Third, the bornavirus-like sequences were concatenated based on the host
670 genomic locations and alignment positions with the bornaviral proteins. When several
671 bornavirus-like sequences were detected in the same genomic positions, the sequence
672 with higher reliability (low E-value score in the tBLASTn search) was used for EBL
673 sequence reconstruction. **(B)** Alignment coverage plot of EBLs. The scales on the x-axis
674 in the alignment are marked at intervals of 100 amino acids. The y-axis indicates the
675 number of EBLs identified in this study. **(C)** Numbers of EBLs in the host genomes.
676 The x-axis indicates the vertebrate species and the y-axis indicates the number of EBLs
677 identified in the species genome. The bar color shows the bornaviral gene (upper panel)
678 or the genus from which the EBL originated (lower panel). **(D)** Phylogenetic trees of

EBLs and extant bornaviruses. These trees were constructed by the maximum likelihood method using the amino acid sequences of EBLs and extant bornaviral proteins. The branch colors indicate the sequence groups: EBLs (gray), extant nyamivirus used as outgroup (black), extant orthobornaviruses (red), extant carboviruses (blue), and extant cultervirus (green). Colored arrows mark extant bornaviruses. The highlights correspond to the current bornaviral classifications: genus Orthobornavirus (light red), genus Carbovirus (light blue), and genus Cultervirus (light green). Representative supporting values (%) are shown on branches. The scale bars indicate genetic distances (substitutions per site).

Fig. 2. The history of bornaviral integration events for approximately 100 million years. (A) Bornaviral integration events during vertebrate evolution. The evolutionary tree of vertebrates was obtained from the TimeTree database. The positions of pie-charts on the tree indicate the lower limit ages of bornaviral integration events, and their size shows the number of events in each period determined based on the gene orthology. Annotations in the internal nodes on the tree indicate the common ancestors of Boreoeutheria (N1), Afrotheria (N2), Metatheria (N3), Primates (N4), Rodentia (N5), Tethytheria (N6), Passeriformes (N7), Australidelphia (N8), Simiiformes (N9), Lemuroidea (N10), Platyrrhini (N11), and Catarrhini (N12). The representative hosts of extant bornaviruses are shown by animal silhouettes on the left side of the tree. The right panel shows the numbers of EBLs categorized into EBLs with orthologs or EBLs without orthologs in each species. The definitions of these categories are described in the section titled Large-scale dating analysis for bornaviral integration ages. The bar colors show the viral genus as indicated on the left side of the tree. **(B)** A schematic

703 diagram of the geographical distributions of ancient bornaviruses and their hosts in each
704 era. The colored continents, except for yellow, indicate the continents where bornaviral
705 endogenization may have occurred: Laurasia or Eurasia (blue), Africa (green),
706 Antarctica (beige), Australia (dark brown), and South America (brown). The
707 biogeography of hosts during their evolution was cited from previous reports (**S2**
708 **Table**). The plate tectonic maps were downloaded from ODSN Plate Tectonic
709 Reconstruction Service (<http://www.odsn.de/odsn/services/paleomap/paleomap.html>).

710

711 **Fig. 3. Phylogenetic relationships of ancient bornaviruses that infected primate**
712 **ancestors. (A-C)** Phylogenetic analyses of ancient and modern bornaviral N genes.
713 These trees were constructed by the maximum likelihood method using the amino acid
714 sequences of EBLNs and extant bornaviral N proteins of genus Carbovirus (A),
715 Orthobornavirus (B), or Cultervirus (C). Colored arrows mark extant bornaviruses. The
716 square and triangle nodes indicate collapsed clades containing all and over half of the
717 orthologs used in the phylogenetic analyses, respectively. Phylogenetic trees with all
718 expanding nodes are available in S3 Fig. The node colors indicate the host lineages of
719 ancient bornaviruses or extant bornaviral genera as indicated in the lower right panel.
720 The colored boxes highlight the bornaviral lineages endogenized during primate
721 evolution. The number on the branches are bootstrap values (%) based on 1000
722 replications. The scale bars show genetic distances (substitutions per site). The genetic
723 distance to distinguish extant bornaviral species is shown as the comparative standard
724 for estimating the genetic diversity of ancient bornaviruses. **(D)** EBLN integration
725 events during primate evolution. Arrowheads indicate the occurrence of ancient
726 bornaviral integrations: orthobornaviral EBLN (red), carboviral EBLN (blue), and

727 culterviral EBLN (green). The colors of highlighted boxes correspond to ancient
728 bornaviral lineages shown in (A-C).

729

730 **Supporting information**

731 **S1 Fig. Dating analysis for bornaviral integration events.** (A) Procedure to
732 determine presence and absence patterns of orthologous EBLs. First, we performed
733 pairwise alignment among EBL integration sites using BLASTN and made an
734 all-against-all matrix of their alignment coverages. Second, we constructed a network
735 using the matrix and grouped EBL loci by extracting community structures. Finally, the
736 ages of bornavirus integration events were assigned from the divergence times of host
737 species with orthologous EBLs. (B) The all-against-all matrix of alignment coverages
738 among EBL integration sites. In the heatmap, the blue color palette shows the alignment
739 coverage between EBL integration sites (%) and yellow indicates that sequence
740 similarity was not detected (ND). The column colors indicate EBL groups; in particular,
741 the white shows manually modified groups (EBLG2, EBLL2, EBLL35, and EBLL36)
742 (details in Materials and Methods). The row colors show host lineages of each EBL
743 locus.

744

745 **S2 Fig. Alignment quality between EBL integration sites.** (A-C) Schematic images
746 of alignments between EBL integration sites. The sequence alignments of EBLG2 (A),
747 EBLL2 (B), and EBLL35 (C) were visualized using AliTV. Blue lines indicate host
748 chromosomal DNA, and the location of EBLs are shown as white colored portions of
749 the lines. The black vertical lines are shown for every 1000 bp. The color palette from
750 red to green indicates identity scores obtained from lastz. The representative host

species are shown as silhouettes to the left of the alignments. **(D)** Dot plot between laurasiatherian and primate EBLG2 integration sites. The line colors except for gray correspond to (A), and gray lines indicate short fragments that were aligned by lastz. The white portions within the blue thick lines indicate the positions of EBLG2 in the genomes.

S3 Fig. Phylogenetic tree of EBLNs and modern bornaviral N proteins. These trees were constructed by the maximum likelihood method using amino acid sequences of EBLN and modern bornaviral N genes of the genus Carbovirus (A), Orthobornavirus (B), or Cultervirus (C). Colored arrows indicate extant bornaviruses. The color of external nodes indicates the extant bornaviral genus or the host species in which the EBLN was identified, as shown in the lower right corner. The square or triangle labels on the internal nodes correspond to the collapsed nodes in **Fig 3A-C**. The colored boxes highlight the bornaviral lineages endogenized during primate evolution. The numbers on the branches are bootstrap values (%) based on 1000 replications. The scale bars show genetic distances (substitutions per site). The genetic distance to distinguish extant bornaviral species is shown as the comparative standard for estimating the genetic diversity of ancient bornaviruses.

S4 Fig. Validation of the network-based method for detection of orthologs using human transposable elements. (A) Strategy for evaluating the detection rate of orthologs using our network-based method. To validate our network-based method, we compared it with the method of detecting orthologs using genomic alignments. First, we estimated the integration age of all human transposable elements (TEs) by LiftOver

775 using the genomic alignment among 18 mammalian species shown in (B). Second, we
776 randomly sampled 100 loci for each of the nine timescales, shown as a to i in (B), from
777 the dating results of the genomic alignment-based method. Using these test datasets, we
778 performed dating analysis by our network-based method. Third, we compared the
779 results between the two methods by checking the predicted ages and detected orthologs.
780 Example 1: integration ages coincided between two methods, and our method could
781 detect all orthologs defined by the genomic alignment-based method. Example 2:
782 integration ages coincided between two methods, but our method detected an
783 incomplete set of orthologs. Example 3: the integration ages were mismatched between
784 the two methods. Example 4: estimation ages were matched between two methods, but
785 there was a contamination of sequence unrelated to true orthologous relationships.
786 Furthermore, to select the best criteria for our network-based dating analysis, we tried
787 nine different criteria shown in (C) (**details in Materials and Methods**). (B)
788 Phylogenetic tree of mammalian species used to detect orthologs of human TEs. These
789 18 mammalian species were used to detect orthologs of human TEs based on the
790 genomic alignment obtained from the UCSC genome browser. To validate the
791 network-based dating method for each timescale, we randomly sampled 100 TE loci
792 from nine different timescales (a to i). (C) Concordant rates between two methods for
793 estimating integration ages. Each panel shows the result using different criteria for
794 network construction (**details in Materials and Methods**). The x-axis indicates the
795 timescales shown in (B). The y-axis indicates the concordant rate (%) between
796 estimations using the genomic alignment-based method and our network analysis-based
797 method. The blue labels indicate the concordant rates (%) between the two methods at
798 the criteria used for the dating analysis for EBLs.

799

800 **S1 Table. The genomic position of EBLs**

801 **S2 Table. Reference list for mammalian biogeography related to ancient**
802 **bornaviral infections**

803 **S3 Table. Genetic distances between ancient and extant bornaviral N genes**

804 **S4 Table. Accession numbers of bornaviral sequences used for the tBLASTn**
805 **search**

806 **S5 Table. Chain files and genome assemblies used for validation for our dating**
807 **method**

808 **S6 Table. Extant viral sequences used for phylogenetic analyses**

809 **S7 Table. Bioinformatics tools used in this study**

810

A

1. Search bornavirus-like sequences

2. Align to bornaviral protein

3. Concatenate bornavirus-like sequences

B

C

D

Figure 1

No. of EBL integrations

Bornaviral genus

Extant viral hosts

B

Figure 2

Figure 3