PNAS
www.pnas.org

Supplementary Information for
One hundred million years history of bornavirus infections hidden in vertebrate genomes

Junna Kawasaki[a,b], Shohei Kojima[a,1], Yahiro Mukai[a,b], Keizo Tomonaga[a,b,c*], Masayuki Horie[a,d*]

***Corresponding authors**

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

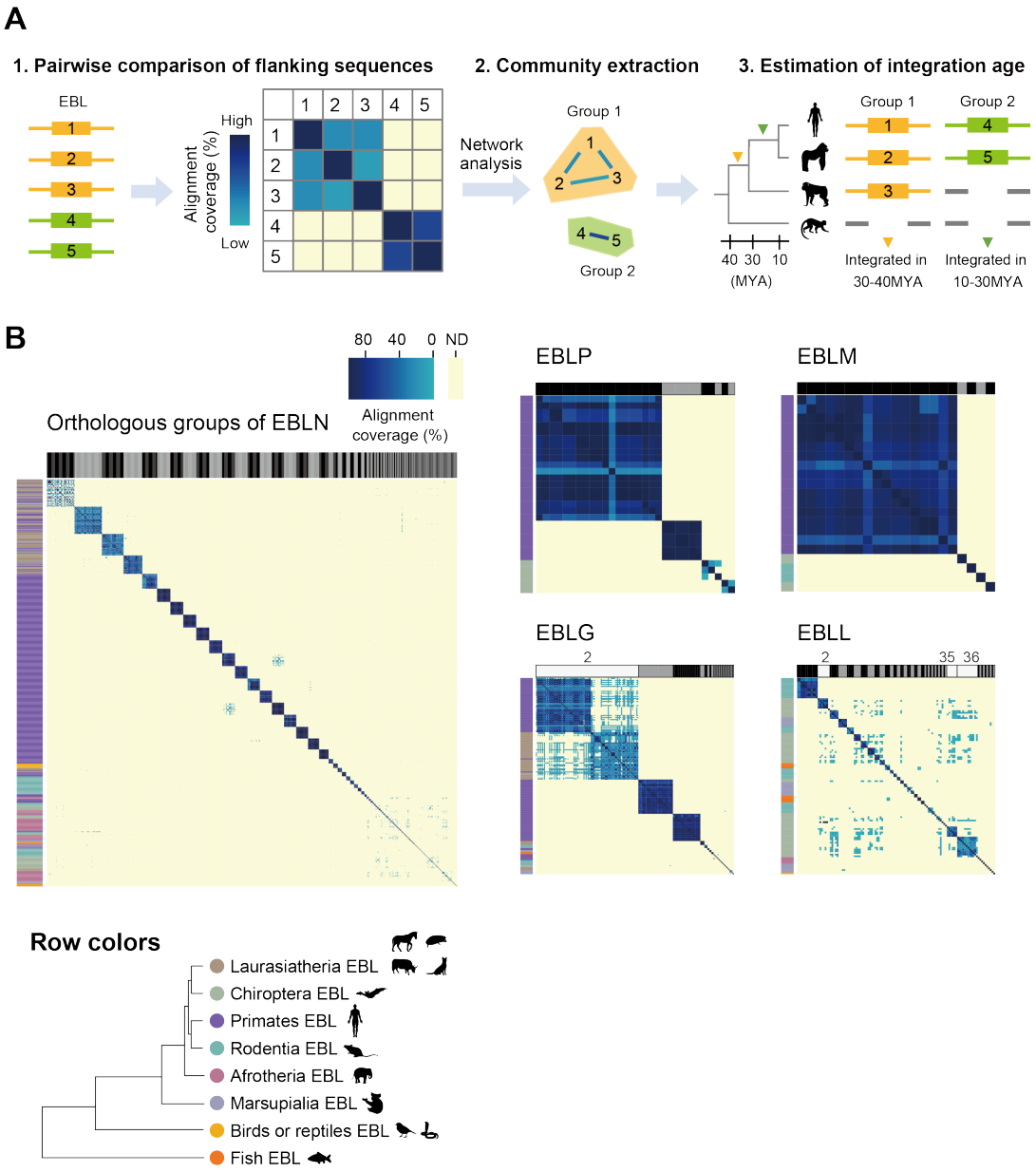
Keizo Tomonaga
Email: tomonaga.keizo.5r@kyoto-u.ac.jp

This PDF file includes:

Figures S1 to S4

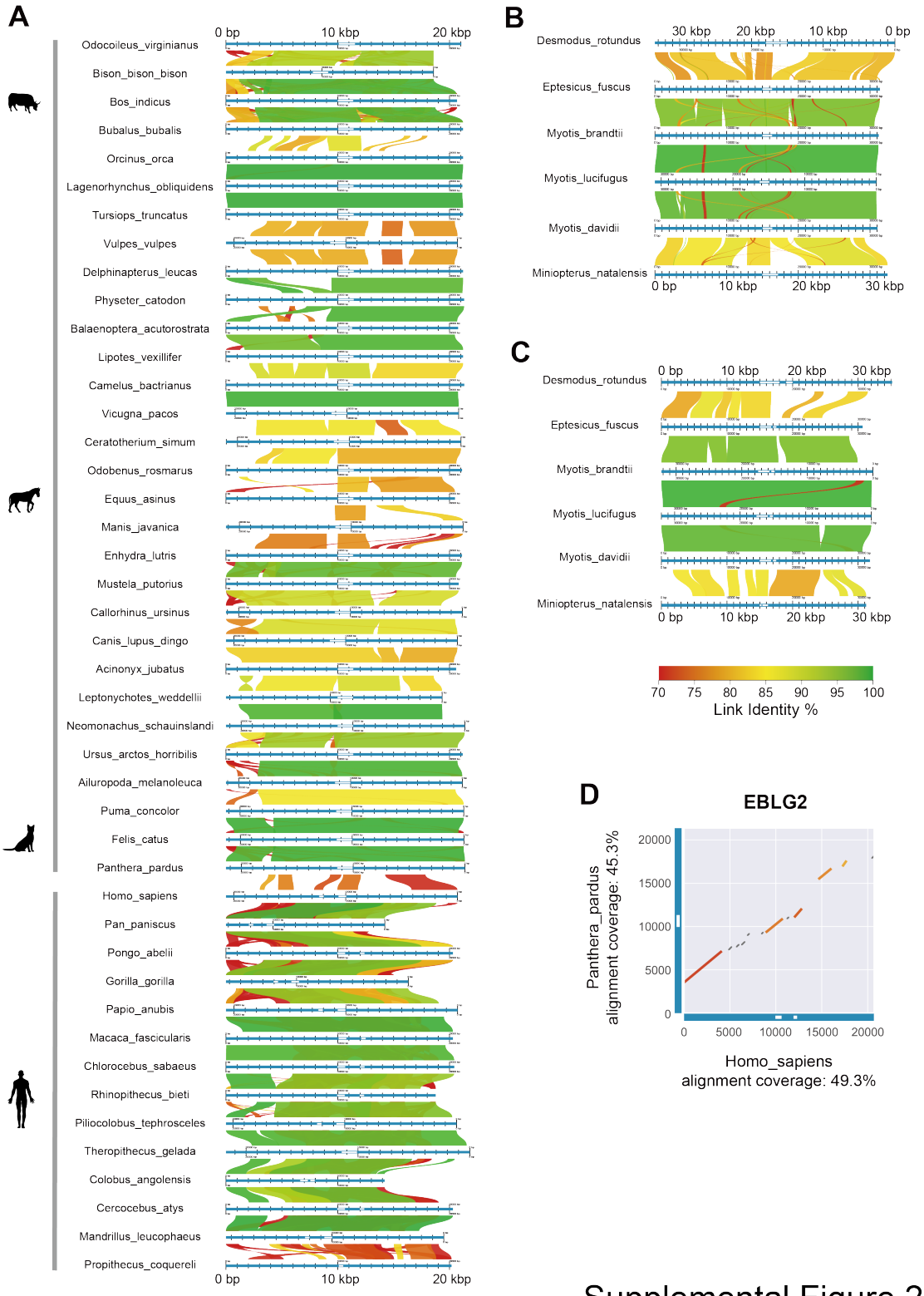
Other supplementary materials for this manuscript include the following:

Datasets S1 to S7



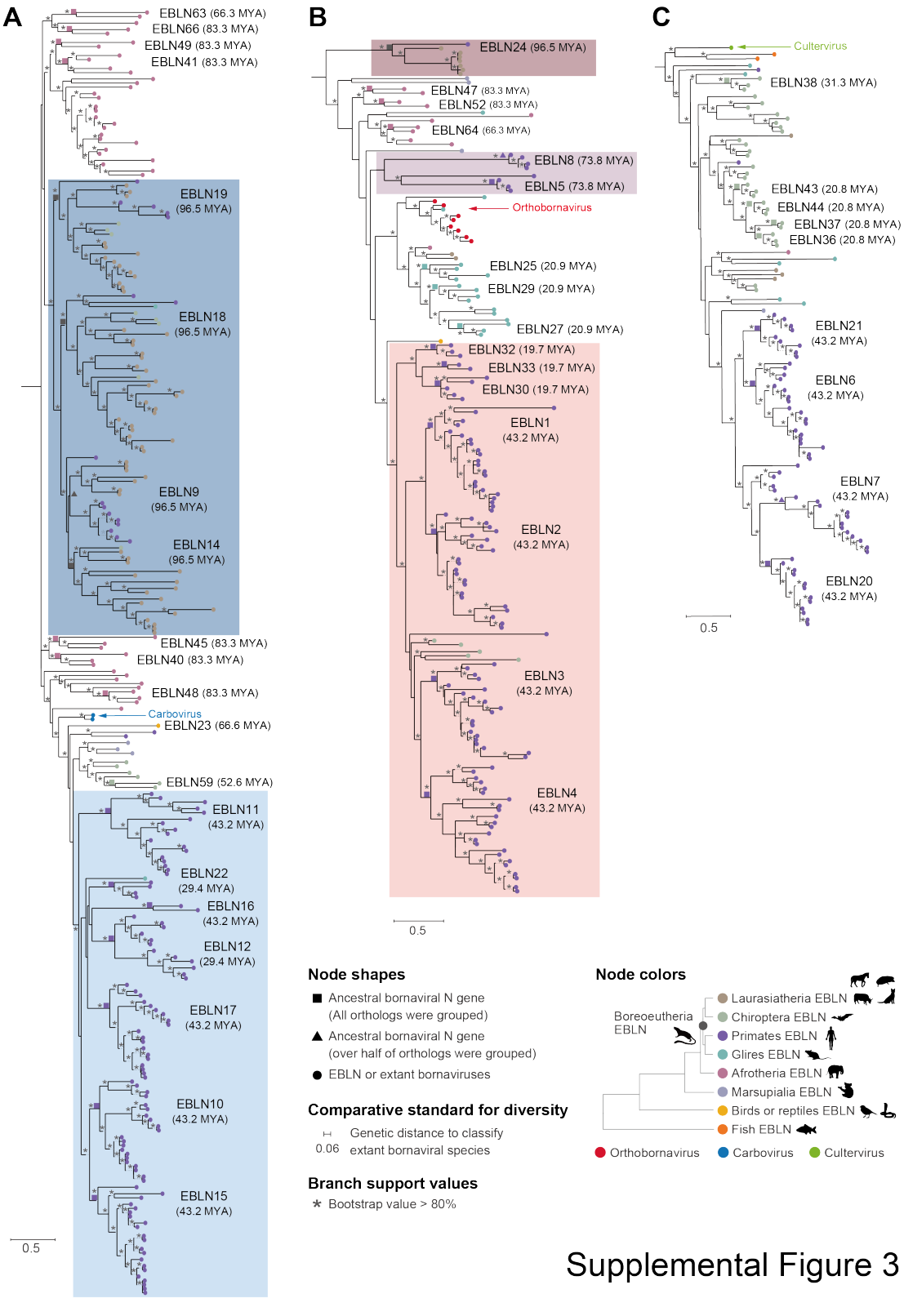
Supplementary Figure 1

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



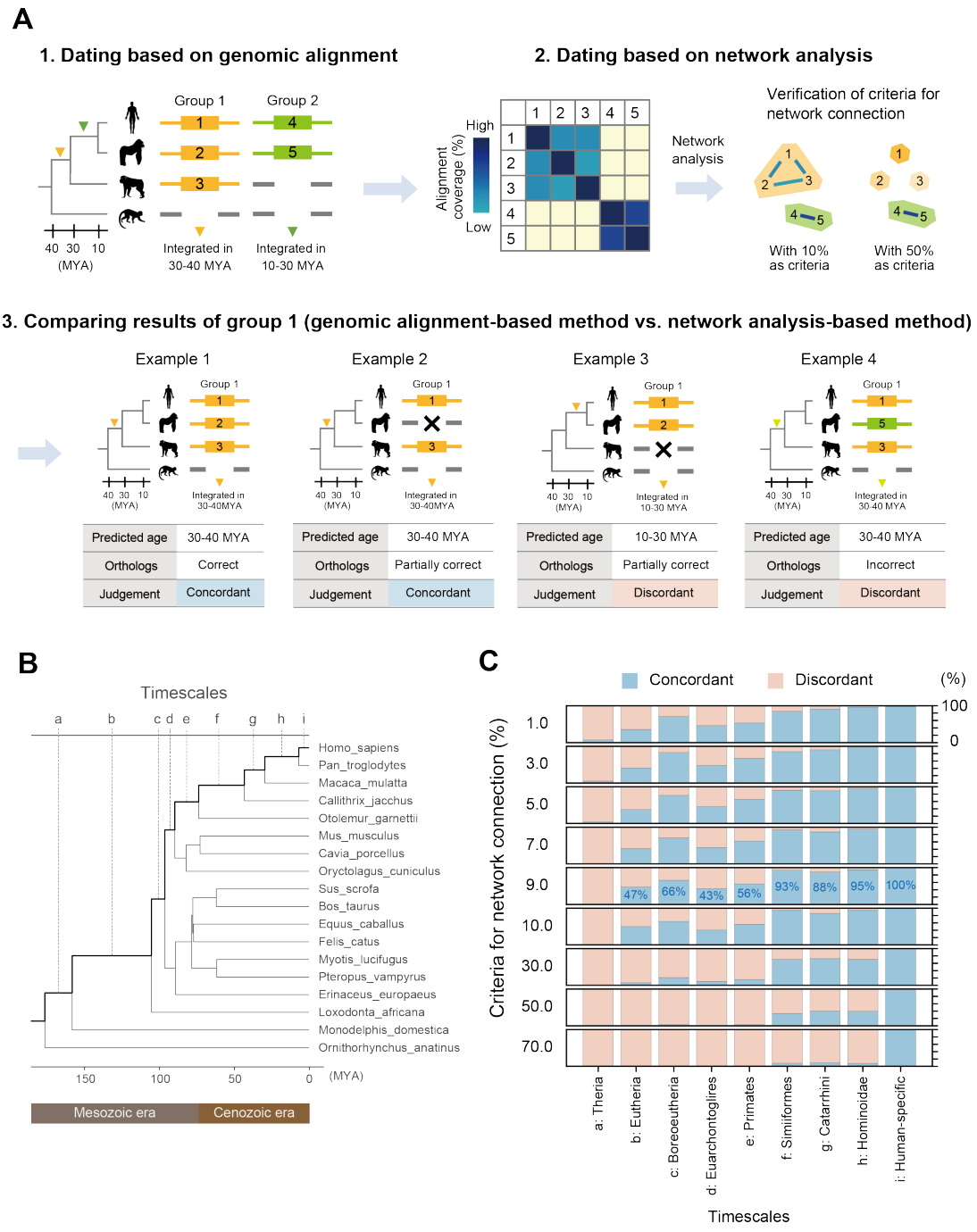
Supplemental Figure 2

Fig. S2. Alignment quality between EBL integration sites. (A-C) Schematic images of alignments between EBL integration sites. The sequence alignments of EBLG2 (A), EBLL2 (B), and EBLL35 (C) were visualized using AliTV. Blue lines indicate host chromosomal DNA, and the location of EBLs are shown as white colored portions of the lines. The black vertical lines are shown for every 1,000 bp. The color palette from 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. Line colors except for gray correspond to (A), and gray lines indicate short fragments aligned by lastz. White portions within the thick blue lines indicate the positions of EBLG2 in the genomes.



Supplemental Figure 3

Fig. S3. 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. 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. Square or triangle labels on the internal nodes correspond to the collapsed nodes in **Figs. 3A-C**. Colored boxes highlight the bornaviral lineages endogenized during primate evolution. Asterisks on the branches indicate that the bootstrap value based on 1,000 replications is more than 80%. 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.



Supplemental Figure 4

Fig. S4. Validation of the network-based method for orthologs detection 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 using the genomic alignment among 18 mammalian species shown in (B). Second, we randomly sampled 100 loci for each of the nine timescales, shown as a to i in (B), from the dating results of the genomic alignment-based method. Using these test datasets, we performed dating analysis by our network-based method. Third, we compared the results between the two methods by checking the predicted ages and detected orthologs. Example 1: integration ages coincided between two methods, and our method detected all orthologs defined by the genomic alignment-based method. Example 2: integration ages coincided between two methods, but our method detected an incomplete set of orthologs. Example 3: integration ages were mismatched between the two methods. Example 4: estimation ages were matched between two methods, but there was a contamination of sequence unrelated to true orthologous relationships. Furthermore, to select the best criteria for our network-based dating analysis, we evaluated the nine different criteria shown in (C) (**details in Materials and Methods**). (B) Phylogenetic tree of mammalian species used to detect orthologs of human TEs. Genomic alignments among these 18 species were obtained from the UCSC genome browser. To validate the network-based dating method for each timescale, we randomly sampled 100 TE loci from nine different timescales (a to i). (C) Concordant rates between the genomic alignment-based and network-based methods for estimating integration ages. Each panel shows the result using different criteria for network construction (**details in Materials and Methods**). The x-axis indicates the timescales shown in (B). The y-axis indicates the concordant rate (%) between two methods. Blue labels indicate concordant rates (%) at the criteria used for the dating analysis for EBLs.

Dataset S1 (separate file). Genomic position of EBLs

Dataset S2 (separate file). Reference list for mammalian biogeography related to ancient bornaviral infections

Dataset S3 (separate file). Genetic distances between ancient and extant bornaviral N genes

Dataset S4 (separate file). Accession numbers of bornaviral sequences used for the tBLASTn search

Dataset S5 (separate file). Chain files and genome assemblies used to validate for our dating method

Dataset S6 (separate file). Extant viral sequences used for phylogenetic analyses

Dataset S7 (separate file). Bioinformatics tools used in this study