VIRUS EVOLUTION

Virus Evolution, 2023, **9(2)**, 1–10

DOI: https://doi.org/10.1093/ve/vead062
Advance Access Publication 2 November 2023
Research Article

Diving deep into fish bornaviruses: Uncovering hidden diversity and transcriptional strategies through comprehensive data mining

Mirette I. Y. Eshak,[†] Dennis Rubbenstroth,[‡] Martin Beer,[§] and Florian Pfaff[*,**]

Friedrich-Loeffler-Institut, Institute of Diagnostic Virology, Südufer 10, Greifswald—Insel Riems 17493, Germany
[†]https://orcid.org/0009-0004-3254-8422
[‡]https://orcid.org/0000-0002-8209-6274
[§]https://orcid.org/0000-0002-0598-5254
[**]https://orcid.org/0000-0003-0178-6183
*Corresponding author: E-mail: florian.pfaff@fli.de

Abstract

Recently, we discovered two novel orthobornaviruses in colubrid and viperid snakes using an *in silico* data-mining approach. Here, we present the results of a screening of more than 100,000 nucleic acid sequence datasets of fish samples from the Sequence Read Archive (SRA) for potential bornaviral sequences. We discovered the potentially complete genomes of seven bornavirids in datasets from osteichthyans and chondrichthyans. Four of these are likely to represent novel species within the genus *Cultervirus*, and we propose that one genome represents a novel genus within the family of *Bornaviridae*. Specifically, we identified sequences of Wǔhàn sharpbelly bornavirus in sequence data from the widely used grass carp liver and kidney cell lines L8824 and CIK, respectively. A complete genome of Murray–Darling carp bornavirus was identified in sequence data from a goldfish (*Carassius auratus*). The newly discovered little skate bornavirus, identified in the little skate (*Leucoraja erinacea*) dataset, contained a novel and unusual genomic architecture (N-Vp1-Vp2-X-P-G-M-L), as compared to other bornavirids. Its genome is thought to encode two additional open reading frames (tentatively named Vp1 and Vp2), which appear to represent ancient duplications of the gene encoding the viral glycoprotein (G). The datasets also provided insights into the possible transcriptional gradients of these bornavirids and revealed previously unknown splicing mechanisms.

Keywords: *Bornaviridae*; *Cultervirus*; fish; data mining; transcription; gene duplication; phylogeny.

Introduction

The family *Bornaviridae* belongs to the order *Mononegavirales* and includes viruses that are considered zoonotic and can cause severe disease in humans, such as Borna disease virus 1 (Niller et al. 2020) and the variegated squirrel bornavirus 1 (Hoffmann et al. 2015). Other members are of veterinary interest because they can cause severe disorders in birds, such as parrots (Rubbenstroth 2022). Taxonomically, the family *Bornaviridae* currently consists of the three genera *Orthobornavirus, Carbovirus* and *Cultervirus* (Kuhn et al. 2015). Of these, the orthobornaviruses have the widest so far known host spectrum and have been identified in birds, reptiles, and mammals (Rubbenstroth et al. 2021). So far, carbo- and culterviruses have only been identified in reptiles and fish (Hyndman et al. 2018; Shi et al. 2018; Costa et al. 2021; Rubbenstroth et al. 2021). The genus *Cultervirus* currently comprises a single virus that has been discovered in fish (Wǔhàn sharpbelly bornavirus [WhSBV], species *Cultervirus hemicultri*) (Shi et al. 2018; Rubbenstroth et al. 2021). Partial genome sequences of another cultervirus, Murray–Darling carp bornavirus (MDCBV), have recently been published, but its classification is still pending (Costa et al. 2021).

The genome of bornavirids consists of an ∼9 kb non-segmented and single-stranded RNA molecule of negative polarity (–ssRNA) (Briese et al. 1994). Typically, six viral proteins are encoded by the viral genome: nucleoprotein (N), accessory protein X, phosphoprotein (P), matrix protein (M), glycoprotein (G), and the large protein (L) containing an RNA-directed RNA polymerase domain (Briese et al. 1994; Rubbenstrothet al. 2021). The open reading frames (ORF) encoding these viral proteins are arranged in two known genomic architectures: (i) 3'-N-X/P-M-G-L-5' (genus *Orthobornavirus*) and (ii) 3'-N-X/P-G-M-L-5' (genera *Carbovirus* and *Cultervirus*). Bornaviral replication and transcription occur in the nucleus of infected cells (Briese et al. 1992) and multiple viral transcripts are produced using conserved transcription initiation and termination sites (Schneemann et al. 1994). Atypically for mononegavirals, bornavirids use alternative splicing in order to control and diversify their transcriptional capacity (Schneider, Schneemann, and Lipkin 1994; Tomonaga et al. 2000).

Recently, we used an *in silico* data-mining approach based on '*Serratus*' (Edgar et al. 2022) in order to screen for traces of potential bornavirids hidden in archived sequence data from public nucleic acid sequences databases, such as the SRA

(Pfaff and Rubbenstroth 2021). The SRA stores raw nucleic acid sequence reads from next-generation sequencing runs from multidisciplinary research experiments, along with extensive metadata. In these archived sequencing reads, we identified and characterised two potential novel orthobornaviruses of colubrid and viperid snakes: Caribbean watersnake bornavirus and Mexican black-tailed rattlesnake bornavirus, in datasets from a Caribbean watersnake (*Tretanorhinus variabilis* [Duméril, Bibron, and Duméril 1854]) and a Mexican black-tailed rattlesnake (*Crotalus molossus nigrescens* [Gloyd 1936]), respectively (Pfaff and Rubbenstroth 2021).

In the present study, we extended the search for previously undetected bornavirids by screening 116,082 transcriptomic datasets from fish samples from the orders Osteichthyes and Chondrichthyes and identified seven bornavirid genomes.

Material and methods

Selection of datasets

We generated a list of datasets using the European Nucleotide Archive (ENA) Browser advanced search portal (https://www.ebi.ac.uk/ena/browser/advanced-search) and selected the data type 'raw reads' using the search query:

```
(tax_tree(1476529) OR tax_tree(7777) OR tax_tree
(7898) OR tax_tree(7878)) AND (library_source=
"METATRANSCRIPTOMIC" OR library_source=
"TRANSCRIPTOMIC SINGLE CELL" OR library_source=
"VIRAL RNA" OR library_source="TRANSCRIPTOMIC")
```

Specifically, this search included the taxonomic units of jawless vertebrates (Cyclostomata; NCBI:txid1476529), cartilaginous fishes (Chondrichthyes; NCBI:txid7777), ray-finned fishes (Actinopterygii; NCBI:txid7898) and lungfish (Dipnomorpha; NCBI:txid7878). We further restricted the search to RNA-derived datasets from (meta-) transcriptomic or viral RNA sequencing experiments.

Data mining of raw reads

In order to identify even single reads within the selected datasets that may be related to bornavirids, we developed the bioinformatics pipeline 'SRAminer'. The 'SRAminer' pipeline is based on *snakemake* (Mölder et al. 2021), is multi-threading and can be run in most Linux-like environments. The code for 'SRAminer' and detailed instructions on how to use it can be found at: https://gitlab.com/FPfaff/sraminer (Pfaff et al. 2023).

A simplified workflow of the pipeline includes the steps (i) download, (ii) blastx, and (iii) report: (i) A subset of reads from each dataset is downloaded using *fastq-dump* (v3.0.3; SRA Toolkit). Typically, a subset of 100,000–1,000,000 reads is sufficient to identify datasets containing sequence reads of interest. (ii) Using *diamond blastx* (v2.0.15; [Buchfink, Reuter, and Drost 2021]), the subset of reads is then searched against a user-provided protein database. In this case, we selected and obtained the protein sequences from all available members of the family *Bornaviridae* from NCBI. (iii) If at least a single read matches the search criteria, additional metadata for this dataset is obtained using *ffq* (0.0.4; [Gálvez-Merchán et al. 2023]) and the results are summarised into individual reports using *R* (R Core Team 2022).

Further raw read processing

After an initial screening of subsets of each 100,000 reads using SRAminer, all datasets that had at least a single blastx match were selected for further analysis. We then downloaded the full datasets of these SRA entries using *parallel-fastq-dump* (v0.6.7; [Valieris 2021]) and trimmed them for low-quality regions and adapter contamination using *TrimGalore!* (v0.6.10; [Krueger 2019; Martin 2011]) running in automatic mode. The trimmed reads were then used for *de novo* assembly with *SPAdes* genome assembler (v3.15.5; [Bushmanova et al. 2019]) running in RNA mode. The resulting contigs were then searched against the representative bornavirid protein database using *diamond blastx* (v2.0.15; [Buchfink, Reuter, and Drost 2021]). Contigs matching the search criteria were selected and imported into Geneious Prime (v2021.0.1) for further characterisation.

Genomic characterisation

Potential ORFs were predicted using the Geneious Prime (v2021.0.1) *Find ORFs* function, and for identification the deduced amino acid sequences were searched against the non-redundant blast database (nr) using the NCBI *blastp* suite. Contigs that represented only a single bornavirid gene did not have an intact full-length ORF and contained flanking host-derived sequences were considered to be potential endogenous bornavirus-like elements (EBLs). Contigs containing multiple intact ORFs that resembled different bornavirid genes, together with flanking untranslated regions, were considered to be bornavirid genomes. Final genomes were additionally screened and trimmed for any vector or adapter contamination using the NCBI *VecScreen* suite (https://www.ncbi.nlm.nih.gov/tools/vecscreen). The trimmed raw reads were mapped back to the respective potential viral genome using the Geneious Prime (v2021.0.1) generic mapper (options: medium sensitivity; find structural variants, short insertions, and deletions of any size) in order to visualise transcriptional profiles and potential splice junctions. Potential transcription initiation and termination sites were predicted based on sequence similarity to known bornaviral signal sequences (Schneemann et al. 1994). They were further verified by manual inspection of the read coverage at these positions (e.g. transition to poly(A) at the termination sites). In addition to visual inspection of the potential transcription start and termination sites, we used MEME (v5.5.2) to discover conserved motifs.

Genomic classification

For the phylogenetic characterisation of potential bornavirid genomes, we used amino acid alignments based on the predicted and translated N, G, and L genes. The amino acid sequences of these genes were individually aligned with 19 reference sequences using Muscle (v3.8.425 [Edgar RC 2004]). The reference viruses were selected to represent all species of the family *Bornaviridae* accepted by the International Committee on Taxonomy of Viruses (ICTV) (n = 12), as well as viruses below the species level (n = 7). The individual alignments were then concatenated into a single alignment and IQ-TREE (v2.2.2.6 [Minh et al. 2020]) was used to infer the phylogenetic relationships. Specifically, a partitioning model (–Q [Chernomor, Haeseler, and Minh 2016]) was used that allowed for individual substitution models and evolutionary rates in each partition. The substitution model was selected automatically (–m MFP + MERGE [Kalyaanamoorthy et al. 2017]) and branch support was assessed using the ultrafast bootstrap (–bb [Hoang, Chernomor et al. 2018]) and SH-aLRT tests (–alrt) with 1,000,000 replicates each.

In addition, the Pairwise Sequence Comparison (PASC) (Bao, Kapustin, and Tatusova 2008; Bao, Chetvernin, and Tatusova 2014) was used to classify the potential bornaviral genomes within the family *Bornaviridae*. PASC is based on pairwise global nucleotide

Table 1. Summary of SRA datasets that were selected for *de novo* assembly of complete bornavirid genomes

SRA accession	Sampled organism	Sampled material	*de novo* assembled virus	Reads matching viral genome
SRR10323915	Grass carp *Ctenopharyngodon idella* (Valenciennes, 1844)	Permanent kidney cell line (CIK) (Wengong et al. 1986, Chen et al. 2018)	Wǔhàn sharpbelly bornavirus WhSBV BK063520	311,433 (0.515%)
SRR6207428	Goldfish *Carassius auratus* (Linnaeus, 1758)	Tissue pool of adult male (Shan, Liu, and Yang et al. 2021)	Murray-Darling carp bornavirus MDCBV BK063521	90,150 (0.123%)
SRR1299086	Electric eel *Electrophorus electricus* (Linnaeus, 1766)	Ampullae of Lorenzini tissue of an adult female	Electric eel bornavirus EEBV BK063519	132,721 (0.046%)
SRR13236436	Finepatterned puffer *Takifugu poecilonotus* (Temminck and Schlegel, 1850)	Radial glial cells from the brain of an adult female (Da Fonte, Martyniuk, and Xing et al. 2017)	Finepatterned puffer bornavirus FPBV BK063517	9469 (0.022%)
SRR9592747	Little skate *Leucoraja erinacea* (Mitchill, 1825)	Kidney tissue (Gallant, Traeger, and Volkening et al. 2014)	Little skate bornavirus LSBV BK063518	37,573 (0.217%)
SRR17661348	Pará molly *Poecilia parae* (Eigenmann, 1894)	Head of an adult female	Pará molly bornavirus PMBV BK063657	615,854 (0.38%)
SRR17441645	Bombay duck fish *Harpadon nehereus* (Hamilton, 1822)	Gill tissue	Bombay duck fish bornavirus BFBV BK063658	65,661 (0.154%)

sequence alignments along the entire viral genome using a blast-based approach.

For the prediction of the potential transmembrane domains (TM) along with the signal peptide (SP) and cleavage sites (CS) within the G protein of the new genomes, we used DeepTMHMM (pybiolib, version 1.1.944 [Hallgren, Tsirigos, and Pedersen et al. 2022]) and ProP-1.0 (Duckert, Brunak, and Blom 2004), respectively.

Sample species assignment

To corroborate the species assignment of the sampled organism, we selected all contigs from the assembly that matched the mitochondrial cytochrome B gene (*MT-CYB*) and submitted them to the NCBI *blastn* suite (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

Results

Data mining

During data mining, we analysed subsets of 116,078 raw transcriptomic SRA datasets from fish (jawless vertebrates, cartilaginous fish, ray-finned fish, and lungfish; see Supplementary Table S1). In 72 of the 116,078 SRA datasets, we found at least one single read that matched one of the bornavirid protein references. For all of these 72 datasets, a *de novo* assembly of all available data was performed and the resulting contigs were scored (see Supplementary Table S2). In eight of the *de novo* assembled datasets, we identified potential EBLs, that were not further analysed. In four datasets, we identified complete genomes from members of the viral family *Chuviridae* (for SRA accessions see Supplementary Table S2). In a further 15 datasets, none of the resulting contigs showed any sequence similarity to the bornavirid reference database. In 44 *de novo* assembled SRA datasets, full or nearly full-length bornaviral genomes were identified. As some of these SRA datasets represented either different organ samples from the same animal or multiple replicates belonging to a single study or were based on the very same cell line, we selected only representative genomes for further characterisation.

As a result, seven complete and unique bornaviral genomes were assembled from SRA datasets SRR10323,915, SRR6207428, SRR1299086, SRR13236436, SRR9592747, SRR17661348, and SRR17441645 (Table 1). The *MT-CYB* sequences assembled from each of these datasets matched those of the specified sampled organisms (Supplementary Table S3).

Taxonomic relationship and classification

Phylogenetic analysis of the predicted viral proteins N, G, and L revealed that the potential bornavirids clustered with viruses of the genus *Cultervirus*, represented by WhSBV (NC_055169) and MDCBV (MW645025-7), rather than carbo- or orthobornaviruses (Fig. 1).

Specifically, the full genome derived from a grass carp kidney cell line dataset (CIK; SRR103 23915) had 87.9 per cent nucleotide identity to WhSBV (NC_055169) and was therefore considered to be a variant of WhSBV. We identified the nearly identical WhSBV genome sequence in 36 SRA datasets, all derived from RNA sequencing of either grass carp kidney (CIK; $n = 10$) or liver (L8824; $n = 26$) cell lines (Supplementary Table S4).

In contrast, the full bornavirid genome from a goldfish tissue pool dataset (SRR6207428) showed 99.5 per cent nucleotide identity to partial sequences of MDCBV (MW645025-7). This sequence can therefore be considered to be the first complete genome of MDCBV.

Additional bornaviral sequences from Bombay duck fish (SRR17441645: *Harpadon nehereus* [Hamilton, 1822]), electric eel (SRR1299086: *Electrophorus electricus* [Linnaeus, 1766]), Pará molly (SRR17661348: *Poecilia parae* [Eigenmann, 1894]), finepatterned puffer (SRR13236436: *Takifugu poecilonotus* [Temminck & Schlegel, 1850]), and little skate (SRR9592747: *Leucoraja erinacea* [Mitchill, 1825]) formed distinct taxonomic units. Hence, we tentatively named these potential viruses based on the origin of the underlying sampling material: Bombay duck fish bornavirus (BFBV; BK063658), electric eel bornavirus (EEBV; BK063519), Pará molly bornavirus (PMBV; BK063657), finepatterned puffer bornavirus (FPBV; BK063517), and little skate bornavirus (LSBV; BK063518). BFBV, EEBV, PMBV, and FPBV maintained between 42 per cent and 66 per cent PASC identity to the

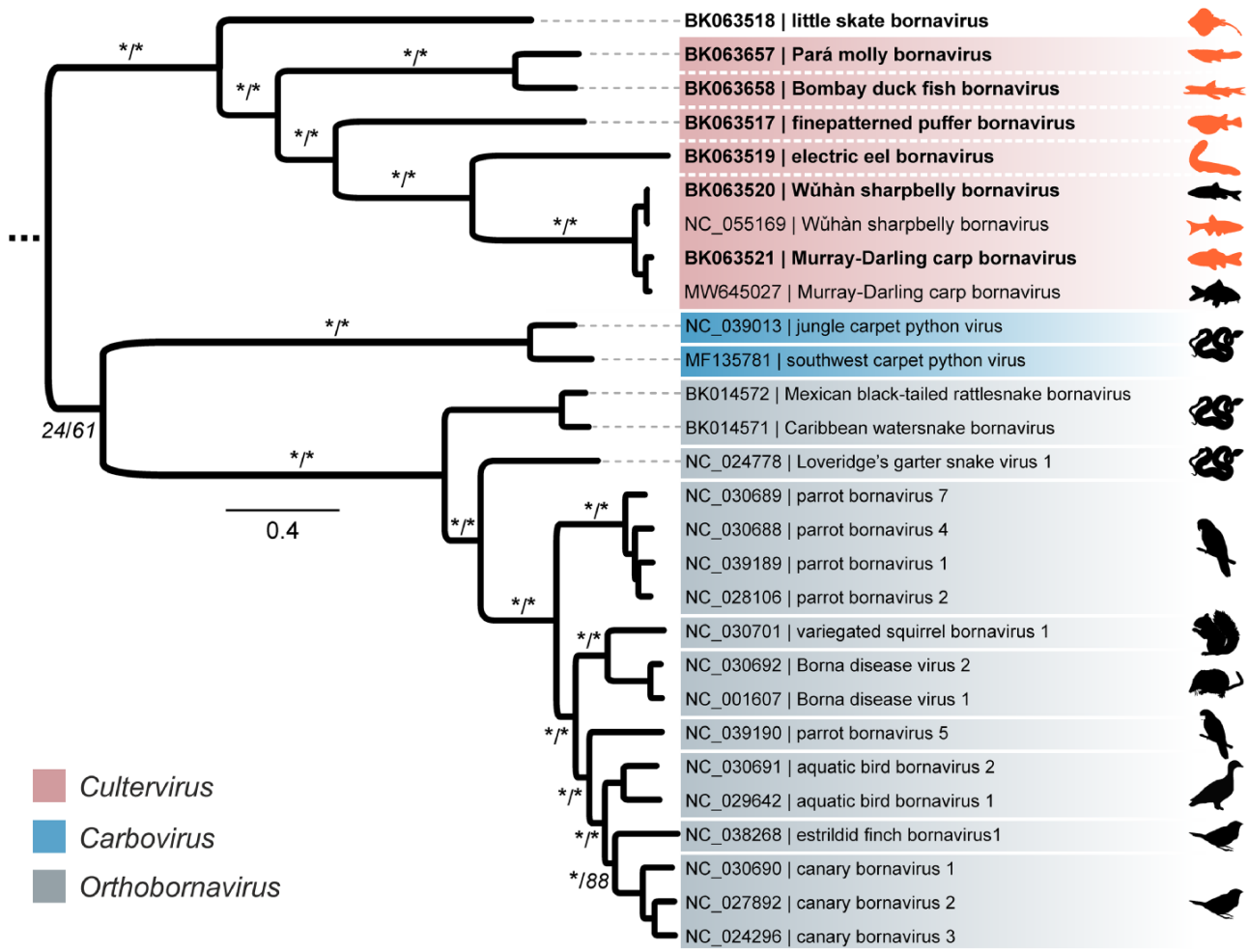

Figure 1. Phylogenetic relationships within the family *Bornaviridae*.

The maximum-likelihood tree was based on the concatenated amino acid sequence alignments of the viral proteins N, G, and L of the newly identified potential fish bornavirids (bold) together with representative members of the genera *Cultervirus*, *Carbovirus*, and *Orthobornavirus*. Two viruses of the family *Nyamiviridae* (Soybean cyst nematode virus 1 (NC_024702.1) and Nyamanini virus (NC_012703.1)) were used as an outgroup to root the tree (data not shown). White lines indicate separate virus species. The silhouettes represent typical host organisms of previously published bornaviruses or the reported sampling source (highlighted) of the viral genomes identified in this study. The tree was constructed using IQ-TREE (version 2.2.2.3), an optimal partition model and statistical support with 1 million replicates each for ultrafast bootstrap and SH-aLRT test. Statistical support is shown for main branches using the format [ultrafast bootstrap/SH-aLRT]. Asterisks indicate statistical support ≥ 90 per cent and ≥ 90 per cent for ultrafast bootstrap and SH-aLRT, respectively.

known culterviruses WhSBV and MDCBV and to each other (Supplementary Figure S1). At 65.8 per cent, BFBV and PMBV were more closely related to each other than to any other virus. LSBV showed the greatest genetic divergence, with PASC identities ranging from 38.2 per cent to 39.9 per cent relative to all other viruses.

Genome architecture

The genome architecture of MDCBV, BFBV, EEBV, PMBV, and FPBV was analogous to that of known culter- and carboviruses, characterised by the arrangement of genes as 3′-N-X/P-G-M-L-5′ (Fig. 2). The identified grass carp WhSBV variant, as well as the goldfish MDCBV variant, closely resembled the WhSBV reference NC_055169 in structure and length (8,989–8,990 nt). In contrast, BFBV, EEBV, PMBV, and FPBV had genome lengths of 9,110, 9,148, 9,324, and 9,397 nt, respectively. Notably, the genome structure of LSBV differed from the other bornaviral genomes in that it was significantly longer, spanning 11,090 nt, and contained two additional ORFs designated viral proteins 1 and 2 (Vp1 and Vp2): 3′-N-Vp1-Vp2-X/P-G-M-L-5′.

Transcriptional profiles, motifs, and alternative splicing

The transcriptional profiles and splice sites of the discovered bornavirid genomes were investigated by aligning/mapping the corresponding raw sequence data to the *de novo* assembled genomes (Fig. 2). The observed sequence coverage was not uniform across the genomes and abrupt increases or decreases were observed within some of the potential intergenic regions. These changes in genome coverage colocalised with predicted transcription start and termination motifs. Specifically, the predicted start sites were characterised by a large increase in read coverage, whereas the termination sites correlated with decrease in read coverage and the presence of reads transitioning to poly(A) at the respective termination site. The respective positions of these predicted regulatory sites were highly conserved between the different viruses. In detail, start sites were present immediately upstream of the N, X/P, and M ORFs. The potential termination sites were located downstream of the N, G, and L ORFs. An additional termination site T3 was present within the L ORF (Fig. 3).

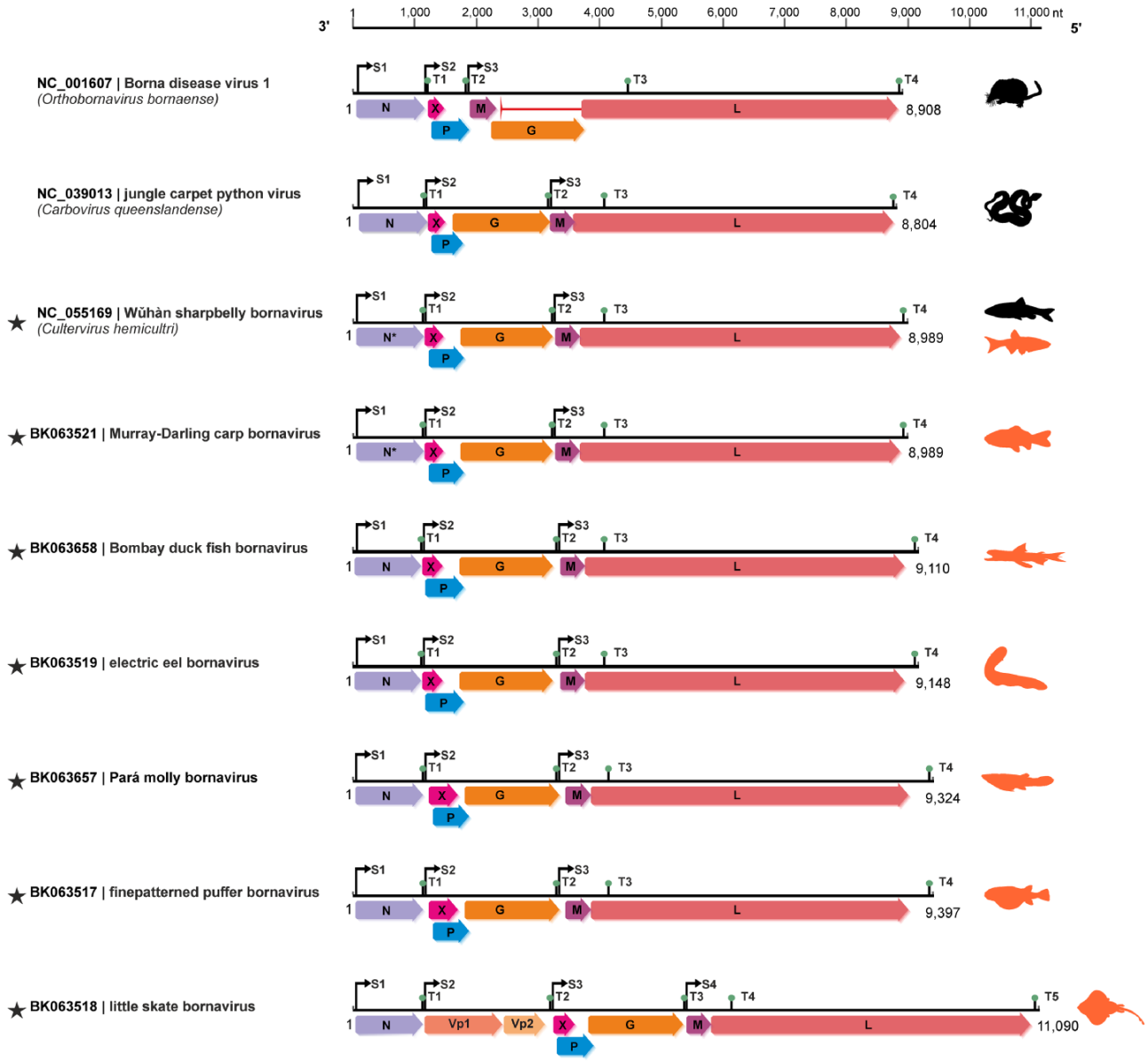

Figure 2. Genome architectures of current and novel bornavirids.

Representative overall genome organisations are shown for representative viruses along with the potential novel viruses (star). The (predicted) ORF are shown as arrows together with the predicted transcription start (S) and transcription termination (T) sites. For each of the genomes, the potential hosts/sources for each virus are shown. Note the different genomic arrangements: 3′-N-X-P-M-G-L-5′ (genus *Orthobornavirus*) and 3′-N-X-P-G-M-L-5′ (genera *Carbovirus* and *Cultervirus*). The little skate bornavirus shares the genomic structure of carbo- and culterviruses, but encodes two additional predicted ORFs: 3′-N-Vp1-Vp2-X-P-M-G-L-5′.

Genomic regions that showed homogeneous coverage and were flanked by adjacent start and termination sites were interpreted as belonging to the same viral RNA transcripts or mRNA (Fig. 3). The overall pattern of viral transcription was highly conserved among all fish bornavirids analysed. In detail, the N protein appeared to be expressed from a monocistronic mRNA, whereas X/P and G were expressed from a polycistronic mRNA. The M and L transcripts appeared to share a single transcription start site (S3), but their expression levels were very different, with L being expressed at low levels and M at relatively high levels.

Interestingly, LSBV showed an additional start and termination site, that were located adjacent to the hypothetical ORFs of Vp1 and Vp2, suggesting that both proteins may be expressed from a bicistronic mRNA. An additional intron was identified between the Vp1 and Vp2 ORFs at nucleotide positions 1,869–2,475, which would result in an in-frame hybrid of the Vp1 and Vp2 ORFs, tentatively named Vp3 (see 'Results' section below).

In addition, we identified an alternative splice site at the beginning of the L ORF, which was present in all the viruses that were analysed. The identified splice site was supported by multiple reads missing the intronic sequence. The intron had a size of 110–176 nt and was located 23–53 nt downstream of the M ORF stop codon. The coverage depth of the unspliced RNA was comparable to that of the L ORF, while the spliced RNA had a coverage comparable to that of the M ORF (Fig. 3). It could be speculated that M is expressed from an RNA that undergoes alternative splicing and uses the T3 transcription termination site located within the L ORF (Fig. 4A). The viral RNA for L on the other hand is expressed from the same S3 transcription start as M but does not undergo splicing and uses a the T4 termination site. The intronic sequences of all viruses analysed showed the

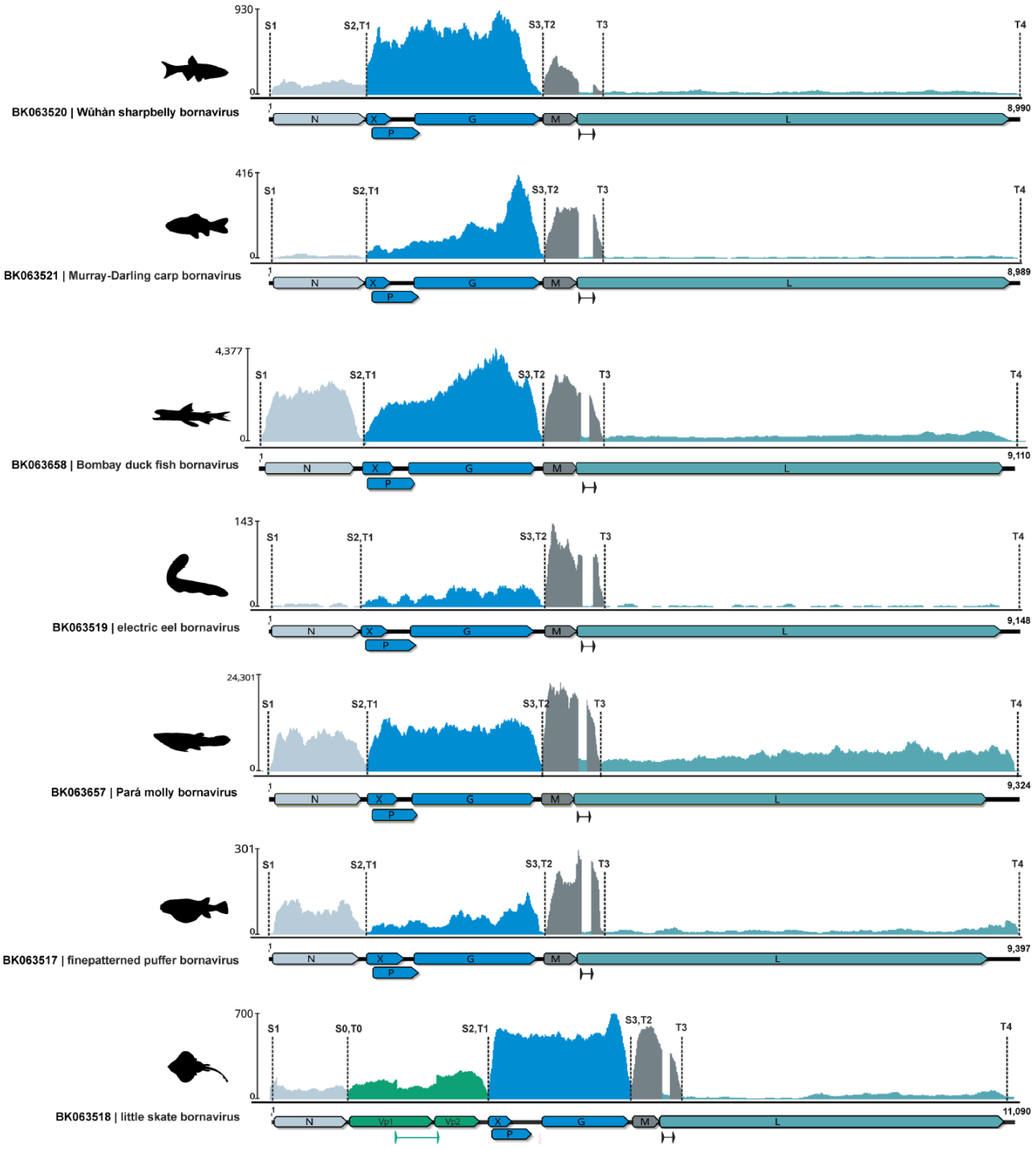

Figure 3. Transcriptional profiles of novel bornavirids.

Raw reads were mapped to the novel viral genomes and the coverage was plotted. ORFs are shown as arrows and predicted transcription start (S) and transcription termination (T) motifs are indicted as dashed lines. S and T sites collocate with large increases and decreases in coverage, respectively. Regions with similar coverage and bordered by S and T sites were considered to represent individual RNA transcripts. These viral transcripts and their corresponding ORFs are highlighted in different colours. Alternative splicing was detected within all viruses for the potential M transcript (intron shown as a line arrow). In addition, a potential intron was identified in the bicistronic transcript encoding Vp1 and Vp2 of little skate bornavirus.

canonical dinucleotides GU and AG for donor and acceptor sites, respectively (Fig. 4A).

Motif prediction revealed conserved sequence patterns for transcription termination and start sites (Fig. 4B). The termination sites T1–4 shared the conserved nucleotide sequence pattern 'AYUUWAKAAAAACAU', whereas the start sites S1, S2, and S3 shared the conserved nucleotide sequence pattern 'GAM'. S2 and S3 were immediately adjacent to T1 and T2, respectively.

LSBV Vp1 and Vp2 are homologs of the glycoprotein G

When analysed by pairwise alignment, the hypothetical viral proteins Vp1 and Vp2 of LSBV shared amino acid similarity with the glycoprotein G of LSBV (Fig. 5A). In detail, the pairwise amino acid identity between the G and Vp1 was 28 per cent, between the G and Vp2 it was 21 per cent and between Vp1 and Vp2 it was 44 per cent (Fig. 5B). While Vp1 shares the N-terminus of

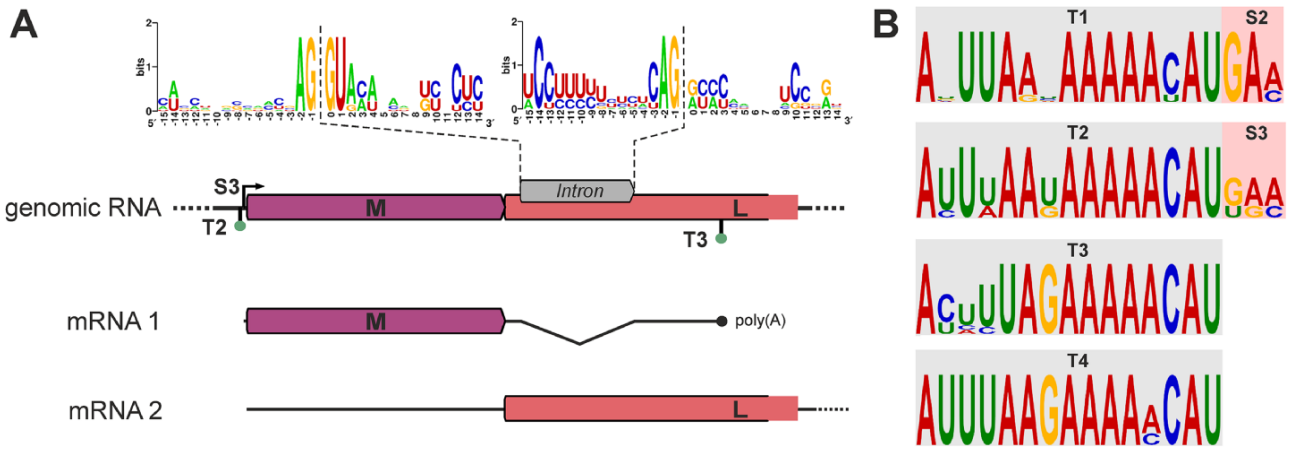

Figure 4. Conserved splicing mechanisms and transcriptional motifs among fish bornavirids.

(**A**) Alternative splicing was detected in the M/L ORF region for all viruses analysed. The genomic arrangement in this region is shown with ORFs indicated by arrows. The potential intron is shown as an arrow within the L ORF region. The sequence motifs of the splice acceptor and donor sites from all analysed viruses are shown and the dashed lines indicate the position of the splicing. The canonical GU/AG splice site is present in all viruses analysed. Two possible mRNAs are shown: The spliced mRNA contains only the M ORF and is terminated at the T3 site, while the unspliced mRNA contains the full L ORF. (B) The conserved motifs of the transcription termination and start sites of the analysed viruses are shown. Note that T1/S2 and T2/S3 are directly adjacent to each other.

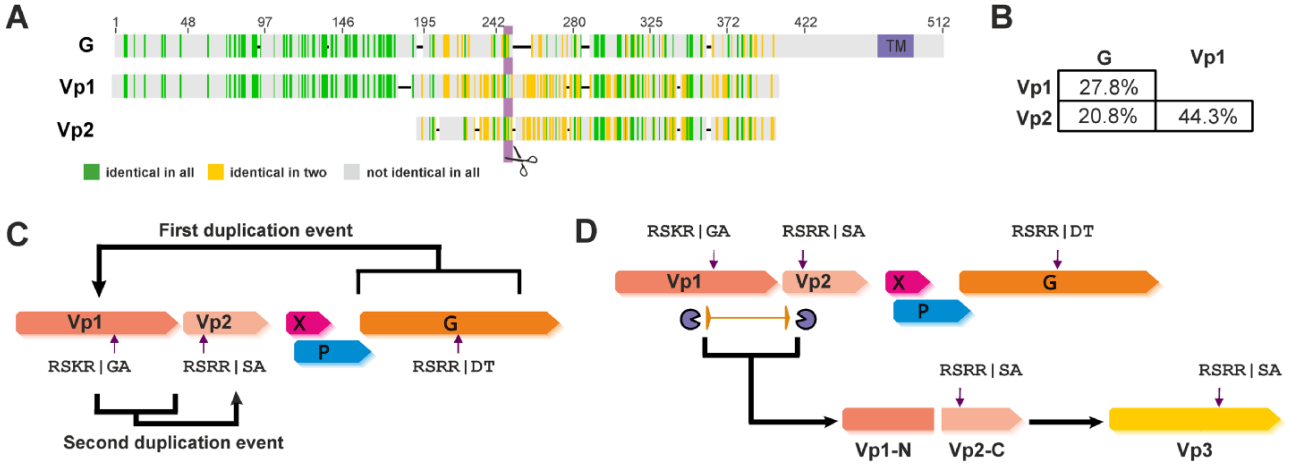

Figure 5. Little skate bornavirus encodes two proteins that may be the result of an ancient duplication event of the glycoprotein.

(**A**) The amino acid alignment of LSBV viral proteins 1 and 2 (Vp1 and Vp2) together with the glycoprotein (G) shows, that they share similarity. Vp1 and Vp2 lack the corresponding transmembrane domain (TM) of G, but each contain a predicted furin protease cleavage site (highlighted by scissors symbol). (**B**) Pairwise amino acid sequence identities indicate, that Vp1 and Vp2 are more closely related to each other than to G. Therefore, in (**C**), supported by phylogenetic analysis (see also the Supplementary Figure S3), we hypothesised that Vp1 was first duplicated from G, followed by a second duplication of Vp1, which gave rise to Vp2. Predicted cleavage sites are indicated by arrows. (**D**) Transcriptional profiling suggested the possibility of alternative splicing of Vp1 and Vp2, resulting in a hybrid of the Vp1 C-terminus, including its cleavage site, and the Vp2 N-terminus, tentatively named Vp3.

the G protein, it lacks the C terminus. Vp2 shares only the central region of the G protein and lacks both, the respective N- and C-terminal regions of G. Both, Vp1 and Vp2, have no detectable transmembrane domain, as they lack the respective C-terminal part of the G protein (470–491 aa; Fig. 5A and Supplementary Table S5).

A phylogenetic tree was constructed based on an amino acid alignment of glycoproteins from selected members of the *Bornaviridae* family, supplemented by LSBV Vp1 and Vp2 (see Supplementary Figure S2). The tree provided evidence for the occurrence of a duplication event of the LSBV G gene, with Vp1 sharing the last common ancestor with G and Vp2 sharing the last common ancestor with Vp1. One possible scenario could be that initially a large part of the glycoprotein gene G was duplicated to form Vp1 and later only the part encoding the C terminus of Vp1 (representing the central part of the G) was duplicated to form Vp2 (Fig. 5C). We also predicted potential furin endoprotease cleavage sites within Vp1, Vp2, and G, following the amino acid consensus motif 'RS(K/R)R' (Fig. 5C and Supplementary Table S5).

As noted above, the predicted mRNA encoding both Vp1 and Vp2, may also undergo splicing, resulting in a hybrid ORF, tentatively named Vp3 (Fig. 5D). The potential Vp3 protein would consist of the N-terminal portion of Vp1 and the C-terminal portion of Vp2, including the protease cleavage site. Similar to Vp1 and Vp2, the potential Vp3 would lack a transmembrane domain (Supplementary Table S5).

Discussion

Knowledge on fish bornavirids has been limited to a single full-length genome of WhSBV (Shi, Lin, and Chen et al. 2018) and a partial genome of MDCBV (Costa, Mifsud, and Gilligan et al. 2021). To identify additional and more diverse fish bornaviruses, we used an *in silico* data-mining approach that screened publicly

available SRA raw sequence datasets from fish (Osteichthyes and Chondrichthyes) samples. Using a similar approach, we had previously successfully identified and characterised two novel snake orthobornaviruses, CWBV and MRBV, as well as novel EBLs in reptile datasets (Pfaff and Rubbenstroth 2021). Here, the screening combined with *de novo* assembly led to the identification of five putative complete bornavirid genomes from different samples.

We found additional sequences of WhSBV (87.9 per cent nt identity to the previously published sequence) and MDCBV (99.5 per cent nt identity) in fish other than the originally reported host species. The first full-length genome sequence of MDCBV presented here matched that of WhSBV in overall structure, sequence identity, and length, indicating that MDCBV and WhSBV are closely related. According to the criteria defined by the ICTV *Bornaviridae* Study Group (Rubbenstroth et al. 2021), they are thought to be viruses of the same virus species (*Cultervirus hemicultri*). WhSBV was previously identified by RNA sequencing of the gut, liver, and gill tissues from a sharpbelly or wild carp (*Hemiculter leucisculus* [Basilewsky, 1855], family Cyprinidae) from China (Shi et al. 2018), whereas MDCBV was discovered in a liver and gill tissue pool of a common carp (*Cyprinus carpio* [Linnaeus, 1758], family Cyprinidae) during a meta-transcriptomic survey of freshwater species in the Murray–Darling Basin in Australia (Costa, Mifsud, and Gilligan et al. 2021). Here, we identified WhSBV in multiple datasets from cell lines derived from the kidney and liver of a grass carp (*Ctenopharyngodon idella* [Valenciennes, 1844], family Cyprinidae) and MDCBV in a dataset from goldfish (*Carassius auratus* [Linnaeus, 1758], family Cyprinidae) brain samples. Both, WhSBV and MDCBV thus appear to be members of a group of bornavirids that are particularly common in fishes of the family Cyprinidae. Cyprinidae includes a wide range of carp and is an ancient evolutionary lineage (Cavender 1991). With a global production of ~30 million tonnes (FAO 2021), carps are of great economic interest and are often cultivated in large-scale aquaculture farms Therefore, the impact of these bornavirids on animal health needs to be carefully assessed and the genome sequences identified in this study may provide valuable information to further investigate the distribution and variability of these viruses.

Using the data-mining approach, identical WhSBV genomes were identified in datasets from the grass carp cell lines CIK (kidney) and L8824 (liver). Both cell lines originate from the Freshwater Fisheries Research Center of Chinese Academy of Fishery Sciences (formerly the Yangtze River Fisheries Research Institute) (Wengong et al. 1986). The CIK and L8824 cell lines have been repeatedly used to study viral transcriptional changes during infection, e.g. with grass carp reovirus (GCRV), and immune regulation. The presence of WhSBV in samples labelled 'mock infection' or 'cell control' (see Supplementary Table S4) indicates that both cell lines may be persistently infected and all experimental results using these cells should be interpreted with caution. It remains unclear whether the WhSBV found in these cell lines originated from the individual(s) from which the two cell lines were derived, or whether both cell lines may have been subsequently contaminated.

We also identified four additional bornaviral genomes in non-cyprinid ray-finned fishes, and one in a cartilaginous fish. These viruses were related to WhSBV and MDCBV, but formed clearly separate taxonomic entities based on a phylogenetic analysis of N, G, and L protein sequences. Despite clear differences at the nucleotide and amino acid level, four of these viruses shared the same overall genomic structure with the culteriviruses WhSBV and MDCBV, and with the viruses of the genus *Carbovirus* (Hyndman et al. 2018). The genome arrangement of reptilian carboviruses and these novel fish bornavirids is peculiar in that it does not follow the standard N-X/P-M-G-L pattern of mononegavirals in general and of orthobornaviruses in particular. This could indicate that the N-X/P-G-M-L genome arrangement evolved independently in reptile and fish bornavirids, or that they share an ancient common ancestor that already had this genome architecture.

The rearrangement of G and M may have resulted in a favourable regulation of gene expression for these viruses. By analysing the transcriptional profiles of the novel fish bornavirids, we found that X/P and G are most likely co-expressed from the same polycistronic mRNA and M is transcribed from a spliced mRNA. In contrast, orthobornaviruses express X and P from a bicistronic mRNA starting from transcription start site S2, whereas M and G are expressed from different splice variants of mRNAs starting from S3 (Schneider, Schneemann, and Lipkin 1994; Rubbenstroth et al. 2021).

Genomic rearrangements do not seem to be an isolated event in bornavirids, as illustrated by the unique genome architecture of LSBV, which encoded two more possible ORFs (Vp1 and Vp2). Both appeared to be the result of at least two independent duplication events: First, a large part of the G gene appears to have been copied into the intergenic region between N and X/P, forming Vp1. Subsequently, a part of Vp1 was duplicated into the intergenic region between Vp1 and X/P, forming Vp2 (Fig. 5). Comparable duplication events in RNA viruses are considered very rare (Simon-Loriere and Holmes 2013), but have been reported for other mononegavirals, such as rhabdovirids (Wang and Walker 1993; Gubala et al. 2008; Gubala et al. 2010; Simon-Loriere and Holmes 2013). Exceptionally long branches in the phylogenetic analysis indicated an accelerated evolution for Vp1 and Vp2 after the duplication events, possibly as a result of changing evolutionary context and selection pressure (Lynch and Conery 2000). However, an alternative hypothesis is that Vp1 and/or Vp2 may have been acquired from another closely related, currently unknown or extinct, virus that shares a common ancestor with LSBV.

In addition, the Vp1 and Vp2 genes may produce a hybrid gene product Vp3 by alternative splicing, extending the coding potential of LSBV even further. The function of Vp1, Vp2, and the splice hybrid Vp3 is currently unknown, but conserved furin cleavage sites suggest that these proteins undergo some form of post-translational modification, similar to the glycoprotein of other bornavirids (Richt et al. 1998). As Vp1, Vp2, and Vp3 lack a detectable transmembrane domain, it can be speculated that they could function as soluble glycoproteins, similar to that of vesicular stomatitis virus (Graeve et al. 1986). The predicted cleavage site within the Vp1, Vp2, and Vp3 sequences may have functional significance for the virus, and future experimental investigations are needed to gain deeper insights into the unique genome architecture of this bornavirid. It would be very interesting to investigate whether other bornavirids from cartilaginous fish share this unique genome structure, or whether LSBV is the result of an isolated evolutionary event.

Based on the phylogenetic analysis and PASC, we propose that LSBV does not belong to any of the existing genera within the family *Bornaviridae*. We have therefore submitted a taxonomic proposal to the ICTV to establish a new genus and new species within this family.

Although the combination of gene arrangement, expression profile and potential hosts was plausible for these potential viruses, it cannot be excluded that these genomes were based

on contaminated samples or inaccurate datasets and therefore did not originate from the reported host species. However, the identification of known viruses such as WhSBV and MDCBV in fish datasets related to the originally reported host may support the credibility of our findings. Confirmation by standard methods, such as PCR and virus isolation, using independent samples from the same species would nevertheless be required to fully confirm the existence of these interesting new viruses in the reported hosts.

Conclusion

The study demonstrates the power of *in silico* SRA data screening and its ability to advance the knowledge of viral diversity and evolution. The screening can easily be applied to the discovery of novel viruses from other viral families or to the identification of known viruses in datasets from previously unknown potential host species.

Data availability

Nucleotide sequence data reported are available in the Third Party Annotation Section of the DDBJ/ENA/GenBank databases under the TPA accession numbers: BK063517-BK063521, BK063657 and BK063658.

Supplementary data

Supplementary data is available at *Virus Evolution* online.

Acknowledgements

We are grateful to the global scientific community for their invaluable contribution in sharing raw sequencing data. These datasets, originally intended for specific research purposes, contain valuable information that goes beyond their original purpose. We would also like to thank Robin Garcia Victoria, Jörg Linde, and Michael Weber for their help with the 'SRAminer' pipeline.

Funding

The investigations were supported by a Friedrich-Loeffler-Institut internal PhD program, grant number FLI-IVD-XX-2021-83, granted to F.P.

Conflict of interest: The authors declare no conflict of interest. The funders had no role in the design of the study, in the collection, analyses, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

References

Bao, Y., Chetvernin, V., and Tatusova, T. (2014) 'Improvements to Pairwise Sequence Comparison (PASC): A Genome-based Web Tool for Virus Classification', *Archives of Virology*, 159: 3293–304.

Bao, Y., Kapustin, Y., and Tatusova, T. (2008) 'Virus Classification by Pairwise Sequence Comparison (PASC)', in B. W. J. Mahy and M. H. V. Van Regenmortel (eds) *Encyclopedia of Virology*, pp. 342–8. Oxford: Elsevier.

Briese, T. et al. (1992) 'Borna Disease Virus, a Negative-strand RNA Virus, Transcribes in the Nucleus of Infected Cells.', *Proceedings of the National Academy of Sciences of the United States of America*, 89: 11486–9.

Briese, T. et al. (1994) 'Genomic Organization of Borna Disease Virus', *Proceedings of the National Academy of Sciences of the United States of America*, 91: 4362–6.

Buchfink, B., Reuter, K., and Drost, H.-G. (2021) 'Sensitive Protein Alignments at Tree-of-life Scale Using DIAMOND', *Nature Methods*, 18: 366–8.

Bushmanova, E. et al. (2019) 'rnaSPAdes: A de Novo Transcriptome Assembler and Its Application to RNA-Seq Data', *Gigascience*, 8: giz100.

Cavender, T. M. (1991) 'The Fossil Record of the Cyprinidae', in *Winfield, Nelson (Hg.) 1991 – Cyprinid Fishes*, pp. 34–54. Springer Science & Business Media.

Chen, G. et al. (2018) 'Transcriptomics Sequencing Provides Insights into Understanding the Mechanism of Grass Carp Reovirus Infection', *International Journal of Molecular Sciences*, 19: 488.

Chernomor, O., Haeseler, A. V., and Minh, B. Q. (2016) 'Terrace Aware Data Structure for Phylogenomic Inference from Supermatrices', *Systematic Biology*, 65: 997–1008.

Costa, V. A. et al. (2021) 'Metagenomic Sequencing Reveals a Lack of Virus Exchange between Native and Invasive Freshwater Fish across the Murray-Darling Basin, Australia', *Virus Evolution*, 7: veab034.

Da Fonte, D. F. et al. (2017) 'Secretoneurin A Regulates Neurogenic and Inflammatory Transcriptional Networks in Goldfish (*Carassius Auratus*) Radial Glia', *Scientific Reports*, 7: 14930.

Duckert, P., Brunak, S., and Blom, N. (2004) 'Prediction of Proprotein Convertase Cleavage Sites', *Protein Engineering, Design & Selection: PEDS*, 17: 107–12.

Edgar, R. C. et al. (2022) 'Petabase-scale Sequence Alignment Catalyses Viral Discovery', *Nature*, 602: 142–7.

Edgar RC. (2004) 'MUSCLE: Multiple Sequence Alignment with High Accuracy and High Throughput', *Nucleic Acids Research*, 32: 1792–7.

FAO. (ed) (2021) *FAO Yearbook: Fishery and Aquaculture Statistics 2019/FAO Annuaire. Statistiques Des Pêches Et de L'aquaculture 2019/FAO Anuario. Estadísticas de Pesca Y Acuicultura 2019*. Rome/Roma: FAO.

Gallant, J. R. et al. (2014) 'Nonhuman Genetics. Genomic Basis for the Convergent Evolution of Electric Organs', *Science*, 344: 1522–5.

Gálvez-Merchán, Á. et al. (2023) 'Metadata Retrieval from Sequence Databases with Ffq', *Bioinformatics*, 39: btac667.

Graeve, L. et al. (1986) 'The Soluble Glycoprotein of Vesicular Stomatitis Virus Is Formed during or Shortly after the Translation Process', *Journal of Virology*, 57: 968–75.

Gubala, A. J. et al. (2008) 'Genomic Characterisation of Wongabel Virus Reveals Novel Genes within the Rhabdoviridae', *Virology*, 376: 13–23.

Gubala, A. et al. (2010) 'Ngaingan Virus, a Macropod-associated Rhabdovirus, Contains a Second Glycoprotein Gene and Seven Novel Open Reading Frames', *Virology*, 399: 98–108.

Hallgren, J. et al. (2022) 'DeepTMHMM Predicts Alpha and Beta Transmembrane Proteins Using Deep Neural Networks', *BioRxiv*: 2022. 10.1101/2022.04.08.487609.

Hoang, D. T. et al. (2018) 'UFBoot2: Improving the Ultrafast Bootstrap Approximation', *Molecular Biology and Evolution*, 35: 518–22.

Hoffmann, B. et al. (2015) 'A Variegated Squirrel Bornavirus Associated with Fatal Human Encephalitis', *New England Journal of Medicine*, 373: 154–62.

Hyndman, T. H. et al. (2018) 'Divergent Bornaviruses from Australian Carpet Pythons with Neurological Disease Date the Origin of Extant *Bornaviridae* Prior to the end-Cretaceous Extinction', *PLoS Pathogens*, 14: e1006881.

Kalyaanamoorthy, S. et al. (2017) 'ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates', *Nature Methods*, 14: 587–9.

Krueger, F. (2019), *Trim Galore* <https://github.com/FelixKrueger/TrimGalore> accessed 8 Jun 2023.

Kuhn, J. H. et al. (2015) 'Taxonomic Reorganization of the Family *Bornaviridae*', *Archives of Virology*, 160: 621–32.

Lynch, M., and Conery, J. S. (2000) 'The Evolutionary Fate and Consequences of Duplicate Genes', *Science*, 290: 1151–5.

Martin, M. (2011) 'Cutadapt Removes Adapter Sequences from High-throughput Sequencing Reads', *EMBnet J*, 17: 10–12.

Minh, B. Q. et al. (2020) 'IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era', *Molecular Biology and Evolution*, 37: 1530–4.

Mölder, F. et al. (2021) 'Sustainable Data Analysis with Snakemake', *F1000Res*, 10: 33.

Niller, H. H. et al. (2020) 'Zoonotic Spillover Infections with Borna Disease Virus 1 Leading to Fatal Human Encephalitis, 1999-2019: An Epidemiological Investigation', *The Lancet Infectious Diseases*, 20: 467–77.

Pfaff, F. et al. (2023) SRAminer. 10.5281/ZENODO.8385126

Pfaff, F., and Rubbenstroth, D. (2021) 'Two Novel Bornaviruses Identified in Colubrid and Viperid Snakes', *Archives of Virology*, 166: 2611–4.

R Core Team (2022) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Richt, J. A. et al. (1998) 'Processing of the Borna Disease Virus Glycoprotein Gp94 by the Subtilisin-like Endoprotease Furin', *Journal of Virology*, 72: 4528–33.

Rubbenstroth, D. (2022) 'Avian Bornavirus Research-A Comprehensive Review', Viruses, 14: 1513.

Rubbenstroth, D. et al. (2021) 'ICTV Virus Taxonomy Profile: Bornaviridae', *The Journal of General Virology*, 102: 001613.

Schneemann, A. et al. (1994) 'Identification of Signal Sequences that Control Transcription of Borna Disease Virus, a Nonsegmented, Negative-strand RNA Virus', *Journal of Virology*, 68: 6514–22.

Schneider, P. A., Schneemann, A., and Lipkin, W. I. (1994) 'RNA Splicing in Borna Disease Virus, a Nonsegmented, Negative-strand RNA Virus', *Journal of Virology*, 68: 5007–12.

Shan, B. et al. (2021) 'Comparative Transcriptome Analysis of Female and Male Fine-Patterned Puffer: Identification of Candidate Genes Associated with Growth and Sex Differentiation', *Fishes*, 6: 79.

Shi, M. et al. (2018) 'The Evolutionary History of Vertebrate RNA Viruses', *Nature*, 556: 197–202.

Simon-Loriere, E., and Holmes, E. C. (2013) 'Gene Duplication is Infrequent in the Recent Evolutionary History of RNA Viruses', *Molecular Biology and Evolution*, 30: 1263–9.

Tomonaga, K. et al. (2000) 'Identification of Alternative Splicing and Negative Splicing Activity of a Nonsegmented Negative-strand RNA Virus, Borna Disease Virus', *Proceedings of the National Academy of Sciences of the United States of America*, 97: 12788–93.

Valieris, R. (2021) *Parallel-Fastq-Dump* <https://github.com/rvalieris/parallel-fastq-dump> accessed 9 Jun 2023.

Wang, Y., and Walker, P. J. (1993) 'Adelaide River Rhabdovirus Expresses Consecutive Glycoprotein Genes as Polycistronic mRNAs: New Evidence of Gene Duplication as an Evolutionary Process', *Virology*, 195: 719–31.

Wengong, Z. et al. (1986) 'A Cell Line Derived from the Kidney of Grass Carp (*Ctenopharyngodon Idellus*)', *Journal of Fisheries of China*, 10: 10–7.