 doi:10.20944/preprints202004.0064.v1

Parvovirus B19 and human parvovirus 4 encode a homologous "X protein" in a reading frame overlapping the VP1 capsid gene

Running Title: a VP1/X overlap in parvovirus B19 and PARV4

David G Karlin[1,*]

1. Independent scholar.
Marseille, France.

* Corresponding author
E-mail: davidgkarlin@gmail.com (DK)

Abstract

30 years ago, researchers noticed that the capsid (VP1) gene of B19 parvovirus might encode a second protein, called "X", in an overlapping reading frame. Since then, experimental approaches failed to detect it. In contrast, sequence analyses can reliably predict whether a protein is expressed from an overlapping frame, provided that it is beneficial to the virus and thus under selection pressure. We used a dedicated software, Synplot2, to identify regions of VP1 likely to encode functional proteins in overlapping frames. Synplot2 detected the X open reading frame and confirmed it is under highly significant selection pressure. We discovered that the X protein is homologous to the ARF1 protein of human parvovirus 4, another suspected protein encoded in a frame overlapping VP1. These findings provide compelling evidence that the X protein must be expressed and functional. We predict that it contains a predicted transmembrane region. We found that the X frame contains a potential AUG start codon in parvovirus B19 and in all related species. Yet no currently known viral transcript has the potential to encode the X protein in a monocistronic fashion. Therefore, the X protein is probably expressed either from an unmapped monocistronic mRNA, or translated by a non-canonical mechanism from the VP1 mRNA or from a short transcript, R3, which has no currently known function. Finally, Synplot2 also detected proteins likely to be expressed from a frame overlapping VP1 in species distantly related to parvovirus B19: porcine parvovirus 2 and bovine parvovirus 3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

Introduction

Parvoviruses are small, non-enveloped viruses (for reviews, see [1–3]). We will focus on two in particular: human parvovirus B19 (B19V) and human parvovirus 4 (PARV4). B19V causes several diseases in humans, such as fifth disease in children, cardiomyopathy, and persistent anemia in immunocompromised persons [4]. PARV4 is not formally associated to any disease, despite suspicions that it may cause encephalitis or accelerate HIV progression [5]. B19V and PARV4 respectively belong to the genera *erythroparvovirus* and *tetraparvovirus*, which are closely related [2]; other species in these genera infect a variety of mammals (see Fig 1).

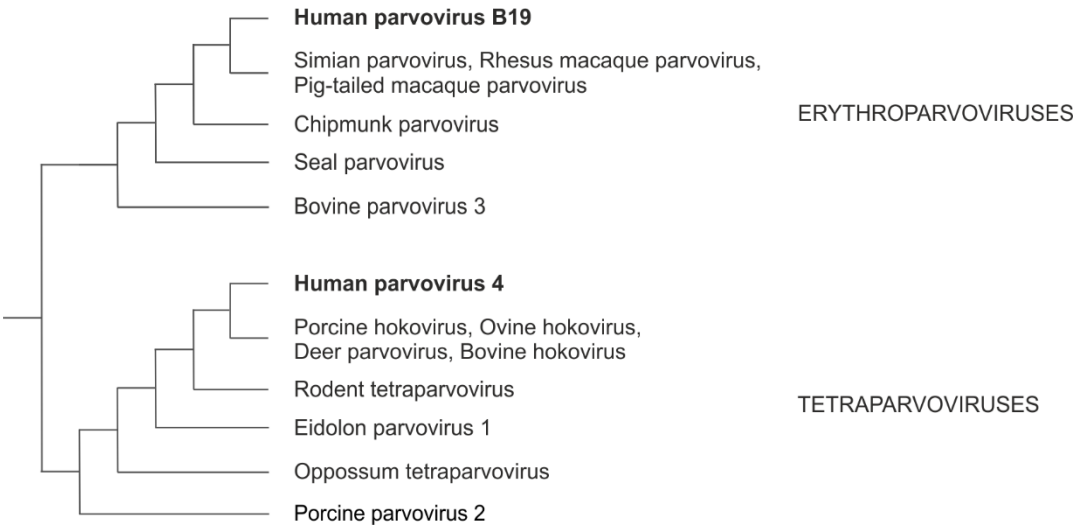

Fig 1: Cladogram of the VP1 proteins of erythro- and tetraparvoviruses

The genome of every erythro- and tetraparvovirus encodes at least two proteins: the replicase NS1 and the capsid protein, of which at least two isoforms are made: VP1 and VP2 (Fig 2). In B19V, three additional ORFs (open reading frames) have been reported (Fig 2A): the 7.5 kDa ORF, which overlaps the NS1 ORF; the X ORF (which has the potential to code for a 9 kDa protein), which overlaps the VP1 ORF; and the 11 kDa ORF, which partially overlaps the 3' region of the VP1 ORF. The expression of the 7.5 kDa protein [6] and of the 11 kDa protein [7,8] have been proven experimentally. In contrast, the expression of the X protein has never been confirmed in infected cells. A substitution meant to knock out the expression of the X ORF caused no discernable change

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

in viral replication or infectivity [9], raising doubts on the expression or functionality of the X protein. Likewise, in PARV4, two ORFs overlapping the VP1 ORF have been noticed, but never confirmed experimentally [10]: ARF1 and ARF2 (ARF stands for "Alternative Reading Frame) (Fig 2B).

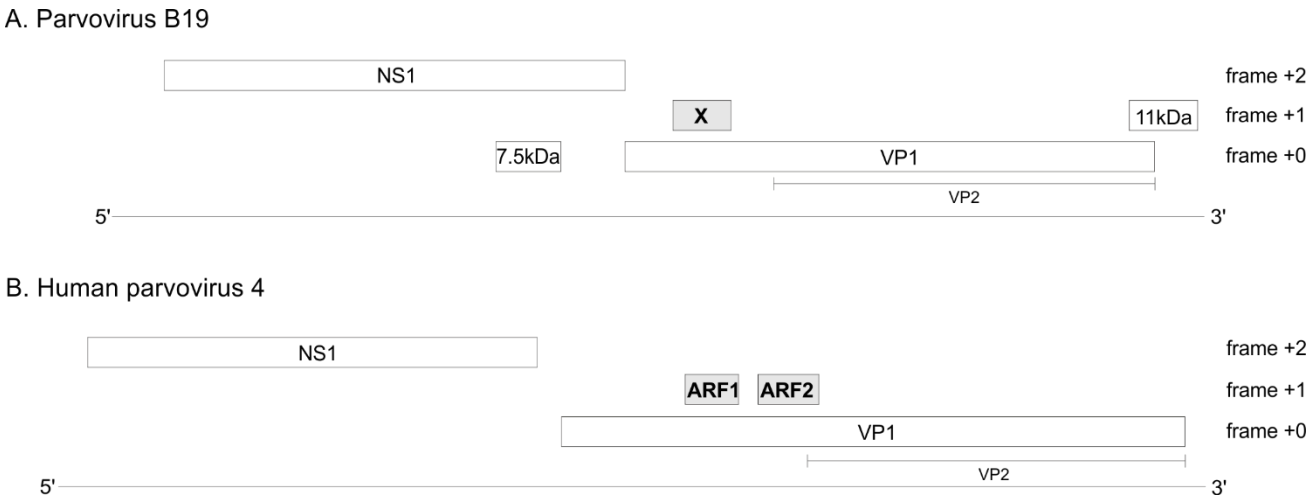

Fig 2: B19V and PARV4 encode three suspected protein-coding ORFs

Long, horizontal lines represent the viral genomes. Boxes represent ORFs (Open reading frames). The three ORFs suspected to code for a protein are in grey. The VP2 isoform of VP1 is represented under VP1.

Overlapping ORFs are frequently overlooked in viral genomes [11]. It is possible, in principle, to predict merely from sequence analyses whether a protein is expressed from overlapping ORFs, provided that the protein confers a beneficial function to the virus. In that case, the additional selection pressure that it causes on the sequence of the reading frame that it overlaps results in a lower rate of synonymous codon substitution in that second frame [12,13]. Surveys of the B19V and PARV4 genomes detected such a lower rate in the region of VP1 corresponding to the X ORF [14], as well as in the region corresponding to ARF1 and ARF2 [10], but did not provide an estimate of the statistical significance of this reduction. In contrast, the software synplot2 [15] can quantify the probability that an ORF with a reduced synonymous codon substitution rate is expressed and functional. Synplot2 has been successfully used to detect over 15 overlapping ORFs later been confirmed experimentally (e.g. [16–18]).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

We thus chose to use Synplot2 to analyze the VP1 coding sequences of B19V and PARV4. Synplot2 detected several regions which correspond either to protein-coding ORFs (including that of the X protein and of ARF1) or to potential functional RNA elements. We compared the sequence properties of the erythroparvovirus X protein with that of tetraparvovirus ARF1 and determined that they were homologous. Finally, we examined the known transcription profiles of erythro- and parvoviruses and identified the most likely expression mechanisms of X and ARF1.

Results

The VP1 coding sequence of B19V and PARV4 contains regions with reduced synonymous variability

The VP1 gene of B19V contains 3 regions with significantly reduced variability at synonymous substitution sites

Table 1 lists the accession numbers of all GenBank reference genome sequences used in this work. We collected the coding sequences (CDS) of all genotypes of B19V VP1 available in GenBank, translated them, aligned their amino acid sequences, and back-translated them to yield a nucleotide sequence alignment. Next, we determined whether the alignment contains regions with a reduced variability at synonymous sites, using Synplot2 [15] (see Methods).

Table 1. Nucleotide sequences of virus species analyzed in this work.

Genus	Species	Common name(s) [Abbreviation]	Genbank genome accession number	Boundaries of the X ORF in the genome sequence (in nucleotides)
Erythroparvovirus	Primate erythroparvovirus 1	**Parvovirus B19 [B19V]**	NC_000883	2874-3119
Erythroparvovirus	Primate erythroparvovirus 2	Simian parvovirus	U26342.1	2718-2963

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

Erythroparvovirus	Primate erythroparvovirus 3	Rhesus macaque parvovirus	AF221122.1	2841-3080
Erythroparvovirus	Primate erythroparvovirus 4	Pig-tailed macaque parvovirus	AF221123.1	2563-2802
Erythroparvovirus	Rodent erythroparvovirus 1	Chipmunk parvovirus	GQ200736.1	3031-3228
Erythroparvovirus	Seal parvovirus	Seal parvovirus	KF373759.1	2789-3100
Erythroparvovirus (*)	Ungulate erythroparvovirus 1	**Bovine parvovirus 3 [bPARV3]**	NC_037053	2627-2926
Tetraparvovirus	Chiropteran tetraparvovirus 1	Eidolon helvum parvovirus	NC_016744.1	2829-3062
Tetraparvovirus	Primate tetraparvovirus 1	**Human parvovirus 4 [PARV4]**	NC_007018.1	2937-3140
Tetraparvovirus	Ungulate tetraparvovirus 1	Bovine hokovirus 1	NC_028136	2857-3111
Tetraparvovirus	Ungulate tetraparvovirus 2	Porcine hokovirus	EU200677.1	2808-3062
Tetraparvovirus	Ungulate tetraparvovirus 5	Deer tetraparvovirus	NC_031670.1	2766-3020
Tetraparvovirus (*)	Ungulate tetraparvovirus 3	**Porcine parvovirus 2 [pPARV2]**; Porcine cnvirus; Parvovirus YX	NC_035180	No X ORF; boundaries of the Z ORF are 2817-3098
Tetraparvovirus	Ungulate tetraparvovirus 4	Ovine hokovirus	JF504699.1	2855-3112
Tetraparvovirus	-	Opossum parvovirus	MG745671.1	2862-3092
Tetraparvovirus	-	Rodent parvovirus	MG745669.1	2960-3217

The main species analyzed here are in bold.

(*) The taxonomic classification of these species might need a revision in view of our analyses.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

Synplot2 identified three regions with a statistically significant increase in the conservation of synonymous sites (Fig 3B):

1) The first region spans codons 58-163 of VP1 (see Table 2), and corresponds to the hypothetical X ORF. In all B19V sequences, this ORF is devoid of stop codons in frame +1 relative to VP1 (Fig 1C). A potential AUG start codon overlaps codon 84 of VP1 and is conserved in all B19V sequences, confirming that the X ORF has the potential to code for a protein. As Fig 3A shows, the X ORF is entirely embedded within the region encoding VP1u (the N-terminus of the capsid protein, found in VP1 but not in VP2), and partially overlaps the region encoding the Phospholipase A2 (PLA2) domain of VP1 [19,20]. An ORF similar to the X ORF is found in all other erythroparvoviruses (see below for the special case of bovine parvovirus 3). We discuss potential expression mechanisms of the X ORF later.

2) The second region detected by Synplot2 spans codons 185-239 of VP1 (Fig 3B and Table 2), and has not been described yet, to our knowledge. We called it "Y region". It is devoid of stop codons in frame +2 relative to VP1 in all B19V sequences (Fig 3C). However, it lacks a potential AUG start codon. It might thus either be translated through a non-canonical mechanism, or correspond to a functional RNA, rather than a protein-coding frame. RNAz [21,22] could detect no secondary structure in the Y region to support the hypothesis of a functional RNA. The Y region overlaps the region of VP1 located downstream of the PLA2 domain and extends slightly into VP2 (Fig 3A). Other erythroparvoviruses do not contain an equivalent region devoid of stop codons.

3) The third region detected by Synplot2 is located at the very C-terminus of the VP1 CDS (codons 771-782) (Fig 3B). It corresponds to the N-terminus of the 11 kDa protein (Fig 3A), known to be expressed in the +1 frame relative to VP1 from an AUG that overlaps codon 756 of VP1 [7,8]. As expected, the region downstream of this AUG is devoid of stop codons in frame +1 relative to VP1 in all B19V sequences except one (accession number KF724386) (Fig 3C).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

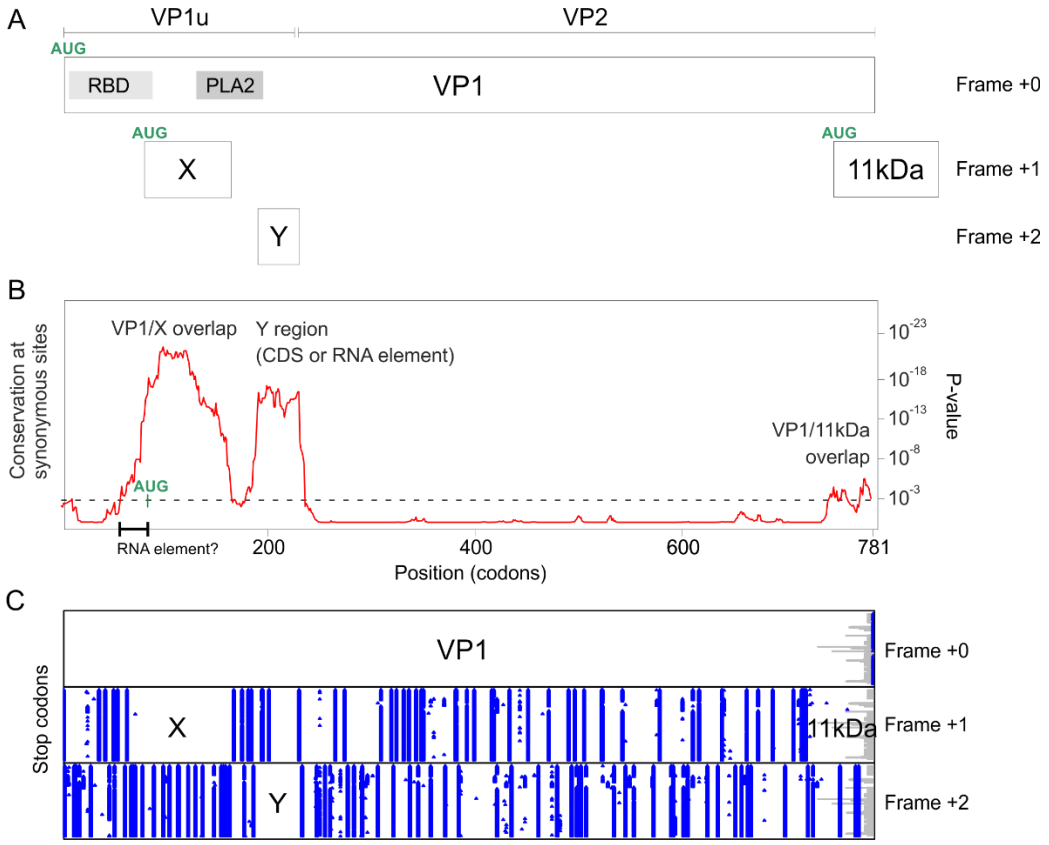

Fig 3. Synplot2 detects 3 regions with significantly lower synonymous-site variability in the VP1 coding sequence of B19V

A. Representation, to scale, of the VP1 gene and of its overlapping protein-coding sequences (CDS) or functional RNA elements. The potential AUG start codon of the X ORF is shown. PLA2: Phospholipase A2 domain. RBD: receptor-binding domain [23]. VP1u: Vp1-unique region.

B. Sequence conservation at synonymous sites in an alignment of coding sequences of B19V VP1 (121 non-redundant sequences ranging from 87% to 99% nucleotide identity), using a 25-codon sliding window. The plot corresponds to the P-value calculated by Synplot2 based on the number of substitutions observed and the number expected under a null model (in which synonymous sites evolve neutrally). Regions in which synonymous substitutions are significantly decreased are indicated. The horizontal dotted line shows the significance cut-off value (10^{-3}). Notice that the first region with a reduced synonymous variability starts markedly before the potential AUG start codon of the X protein (in green). This region is indicated by a thick line. It might correspond to a functional RNA element, which perhaps facilitates the translation of the X protein or the splicing of an X-specific RNA transcript (see text).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

C. Position of stop codons (blue) in the 3 potential frames, and gaps in alignment (gray) in the 121 B19V sequences.

Table 2. Boundaries of the regions of VP1 with significantly lower synonymous codon variability identified by Synplot2 and encompassing potential protein-coding ORFs.

Virus name	Region	Boundaries of the region with lower synonymous codon variability in the VP1 CDS	Boundaries of the corresponding ORF in the VP1 CDS
Parvovirus B19	X ORF	Codons 58-163 (nucleotides 172-489)	Codons 84-166 (Nucleotides 251-496)
Parvovirus B19	Y region(*)	Codons 185-239 (nucleotides 553-715)	Codons 185-230(*) (nucleotides 553-715)
Human parvovirus 4	X ORF (=ARF1)	Codons 180-263 (nucleotides 538-789)	Codons 187-255 (nucleotides 560-763)
Human parvovirus 4	ARF2	Codons 294-397 (nucleotides 880-1189)	Codons 295-379 (nucleotides 884-1135)
Bovine parvovirus 3	X-like ORF	Codons 225-289 (nucleotides 673-867)	Codons 215-315 (nucleotides 644-943)
Porcine parvovirus 2	Z ORF	Codons 193-309 (nucleotides 577-927)	Codons 193-285 (nucleotides 578-854)

(*): this region contains an ORF devoid of stop codon, but lacks a potential AUG start codon, and might not code for a protein.

The VP1 gene of PARV4 contains 2 regions with significantly reduced synonymous variability, corresponding to ARF1 and ARF2

We analyzed the VP1 coding sequence of all strains of PARV4 by using Synplot2, as described above for B19V. Fig 4B shows that two regions have a highly significant increase in the conservation of synonymous sites (Table 2):

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

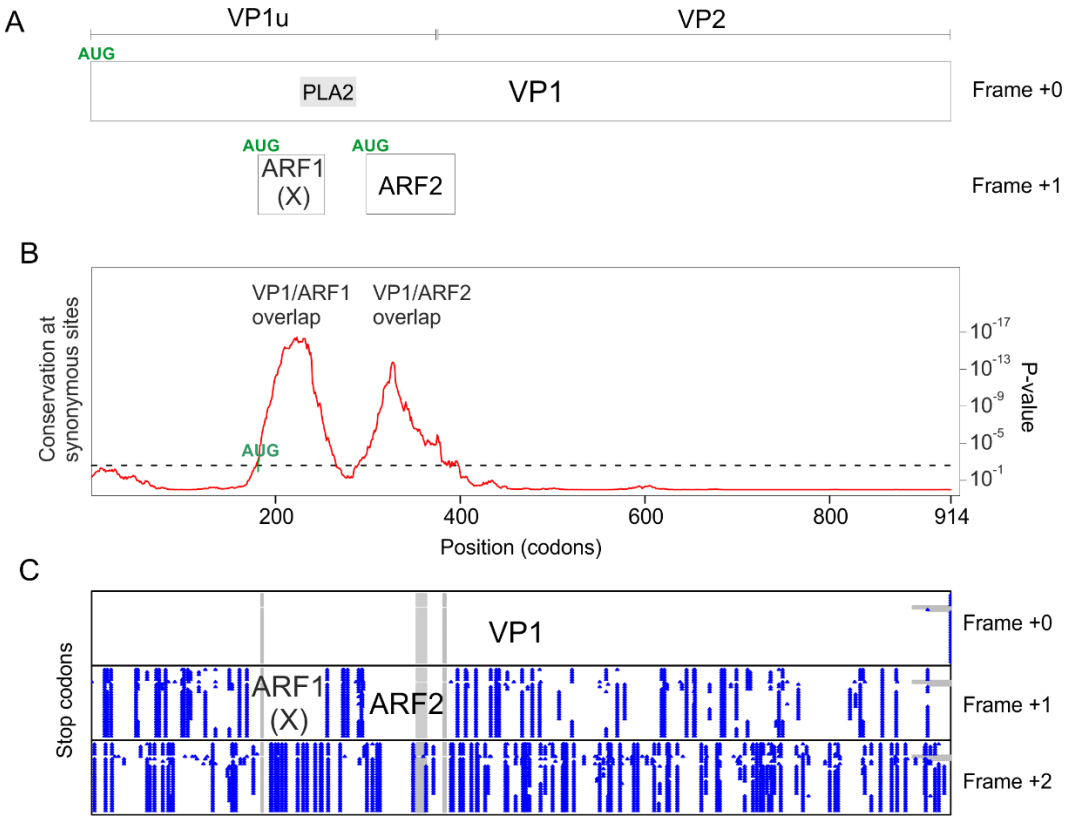

Fig 4. Synplot2 detects 2 regions with significantly lower synonymous-site variability in the VP1 coding sequence of B19V

A. Conventions are the same as in Fig 3. The potential AUG start codon of the X ORF is shown.

B. Conservation at synonymous sites in an alignment of coding sequences of PARV4 VP1 (21 non-redundant sequences ranging from 93% to 99% identity), using a 25-codon sliding window in Synplot2.

C. Position of stop codons (blue) in the 3 potential frames, and gaps in alignment (gray) in the 21 sequences.

1) The first region spans codons 180-263 of VP1 (Table 2), which corresponds to the hypothetical ARF1 protein [10] (see Introduction). In all PARV4 sequences, ARF1 is devoid of stop codons in frame +1 relative to VP1 (Fig 4B). It has a potential AUG start codon conserved in all PARV4 sequences, overlapping codon 187 of VP1. ARF1 is embedded within the VP1u region, and partially overlaps the PLA2 domain (Fig 4A). An ORF similar to ARF1 was found in all other tetraparvoviruses, with the exception of porcine parvovirus 2 (see below).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

2) The second region detected by Synplot2 spans spanning codons 294-397, and corresponds to the hypothetical ARF2 protein [10] (see Introduction). ARF2 is devoid of stop codons in frame +1 relative to VP1 (Fig 4C). It has a potential AUG start codon conserved in all PARV4 sequences, overlapping codon 294 of VP1. The ARF2 frame overlaps the region of VP1 located immediately downstream of the PLA2 domain, and extends slightly into VP2 (Fig 4A). Note that PARV4 ARF2 and the putative Y protein of B19V cannot be homologous, because they are encoded in different frames relative to VP1 (respectively +1 and 2, compare Fig 4A and Fig 3A).

An ORF similar to ARF2 is found only in tetraparvoviruses closely related to PARV4: hokoviruses (porcine, bovine and ovine), and deer tetraparvovirus. We present their aa sequence in S1 Fig. ARF2 has a predicted transmembrane segment near its N-terminus. We discuss potential expression mechanisms of ARF2 later.

The X protein and ARF1 are homologous

The B19V X protein and PARV4 ARF1 protein have similar predicted features, in particular a central transmembrane segment

Fig 5 presents multiple sequence alignments of the erythroparvovirus X protein (Fig 5A) and of tetraparvovirus ARF1 (Fig 5B). The erythroparvovirus X protein contains a predicted central transmembrane segment (Fig 5A). It is followed by a positively charged region, predicted to be inside the cytosol ("positive-inside rule" [24]). Therefore, the N-terminus of X, which must be on the other side of the transmembrane segment, is necessarily extra-cytosolic (Fig 5A). In B19V and the three closely related erythroparvoviruses infecting monkeys, the C-terminus of the X protein is predicted to form a second transmembrane segment (boxed in Fig 5A).

Tetraparvovirus ARF1 has a size and predicted organization similar to that of the X protein (compare Fig 5B and 5A), composed of an extra-cytosolic N-terminus, a central transmembrane segment, and a positively charged, intra-cytosolic region.

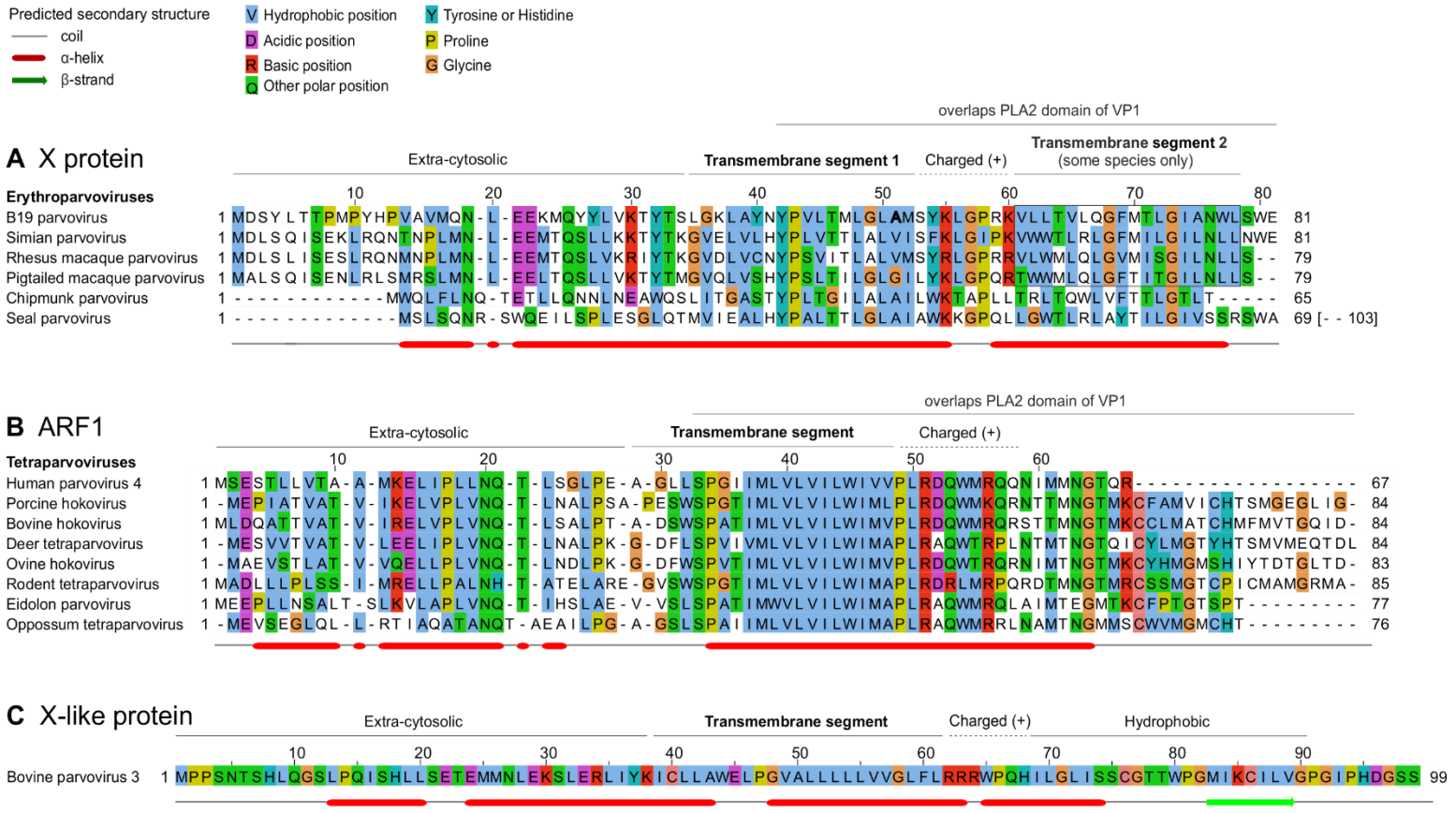

Fig 5. Similar organization of the erythroparvovirus X protein, tetraparvovirus ARF1, and bovine parvovirus 3 X-like protein

A. Multiple sequence alignment of the erythroparvovirus X proteins. Numbering corresponds to B19V. The sequences presented assume that the first AUG of each X ORF is used to initiate translation. PLA2: Phospholipase A2 domain.

B. Alignment of the tetraparvovirus ARF1. Numbering corresponds to PARV4

C. Sequence of the X-like protein of bovine parvovirus 3.

doi:10.20944/preprints202004.0064.v1

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

The X protein of erythroparvoviruses and the ARF1 protein of tetraparvoviruses are homologous

3 lines of evidence suggest that the *erythroparvovirus* X protein of and the *tetraparvovirus* ARF1 protein might be homologous, i.e. share a common origin: 1) they overlap a similar region of the VP1 gene (encoding the PLA2 domain, indicated above the alignments in Fig 5); 2) they are both in the +1 frame relative to VP1 (see Fig 3A and 4A); 3) they have similar sequence features, as shown above. However, the presence of a transmembrane segment could be explained by convergent evolution [25]. Therefore, to check whether X and ARF1 are homologous, we examined how their sequences align when based on the much more reliable alignment of VP1, and in particular of its PLA2 domain. Indeed, PLA2 contains numerous strictly conserved amino acids (aas) [19,20], which makes its sequence alignment highly reliable.

We followed two steps to generate the alignment of erythroparvovirus X proteins and tetraparvovirus ARF1 based on VP1: 1) we converted the aa alignment of the VP1 proteins into an alignment of nucleotide sequences by using TranslatorX [26]; 2) we translated this alignment in the reading frame of X and ARF1, i.e. the +1 frame relative to VP1. This procedure is also described graphically in a previous article [27].

The resulting alignment of X and ARF1 is shown in Fig 6A, while the reference alignment of VP1 is shown below, in Fig 6B. (We only show the PLA2 domain of VP1 because the region upstream is not well conserved). As Fig 6 A shows, the transmembrane segments of X and ARF1 align together perfectly. Three aa positions are strictly conserved between X and ARF1, and one position is semi-conserved (aromatic: Y, W or F). They are indicated above the alignment in Fig 6A. This high degree of conservation, coupled to the fact that erythro- and tetraparvoviruses are closely related genera [2], indicates that X and ARF1 are most probably homologous.

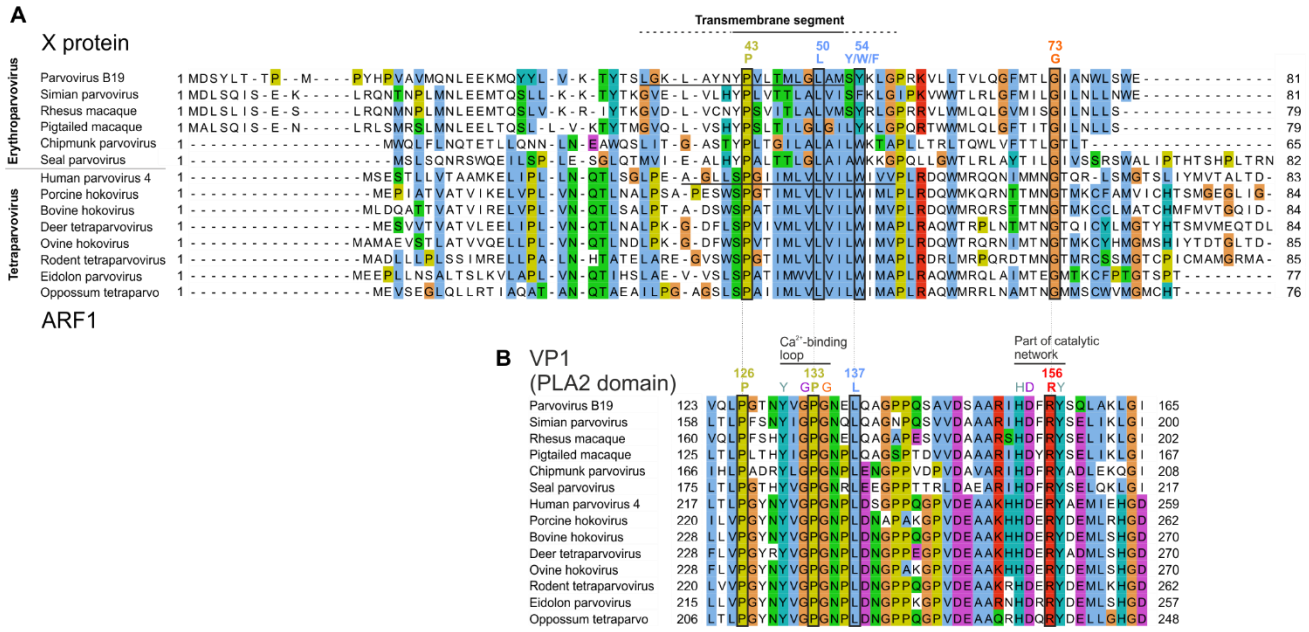

Fig 6 Alignment of all X proteins based on the reliable alignment of the PLA2 domain of VP1

Conventions are the same as in Fig 5. Numbering corresponds to B19V.

A. Alignment of the X protein of erythro- and tetraparvoviruses, derived from the reference alignment of VP1 presented in panel B. The X alignment was generated from the VP1 alignment by using TranslatorX [26] (see text). Strictly- or semi-conserved aas are boxed and indicated above the alignment. Predicted transmembrane regions are underlined in the sequence of B19V X and PARV4 ARF1. The region that forms a transmembrane segment in both B19V X and PARV4 ARF1 is indicated above the alignment by a thick line; the region that forms a transmembrane segment only in either of these proteins is indicated by a dotted line.

B. Alignment of VP1 on which is based the alignment of the X protein in panel A. Only the reliably aligned region of VP1 that overlaps X is shown; it encompasses the N-terminal part of the PLA2 domain. Thin vertical lines show the correspondence between aas encoded by overlapping codons in the X frame (panel A) and in the VP1 frame (panel B). Aas that overlap conserved positions of the X protein are boxed and indicated above the alignment. Other conserved aas involved in functional elements of PLA2 are also indicated.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

Conserved features of the X protein mostly correspond to conserved motifs of the Phospholipase A2 domain of VP1

We next asked whether conserved sequence features of the X protein correspond to conserved sequence motifs of the PLA2 domain that it overlaps. As Fig 6B shows, the region of PLA2 overlapped by the X protein contains two conserved features: 1) the putative calcium (Ca^{2+})-binding loop (aa 130-134 in B19V); and 2) a region involved in the catalytic network, containing strictly conserved aas H153, D154 and Y157 in B19V numbering [19,20]. The conserved features of the X protein correspond to these conserved features of PLA2. First, the transmembrane segment of the X protein overlaps the Ca^{2+}-binding loop. Second, strictly conserved positions of the X protein (corresponding, in B19V, to aa P43, L50, G73, boxed in Fig 6A) overlap strictly conserved positions of PLA2, boxed in Fig 6B: P126 and P133 (both within the Ca^{2+}-binding loop), and R156, close to conserved aas of the catalytic network. Likewise, the semi-conserved position of the X protein (Y54 in B19V) corresponds to a strictly conserved position of VP1 (L137 in B19V).

Clearly the PLA2 enzyme is under stringent selection pressure to conserve aas responsible for its catalytic activity. Therefore, one might assume that the sequence conservation within the X protein is dictated by PLA2. However, the sequence of strictly conserved aas of X is not *completely* imposed by PLA2. For instance, consider the strictly conserved P133 and G134 in PLA2, which overlap the strictly conserved aa L50 in the X frame (Fig 6). The strict conservation of this Leucine in the X frame is *not* imposed by the conservation of P133 and G134, since the dipeptide PG (Proline-Glycine) can be encoded by the nucleotides CCNGGN, in which N is any nucleotide. The first corresponding codon in the +1 frame relative to PLA2 is therefore CNG, which can encode not only Leucine (CTG), but also 3 other aas: Proline (CCG), Glutamine (CAG), or Arginine (CGG). Likewise, none of the conserved positions of the X protein are completely imposed by conservation of PLA2.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

The VP1 gene of Bovine parvovirus 3 and porcine parvovirus 2 differs from that of other erythro- and tetraparvoviruses

Bovine parvovirus 3 may encode a homolog of the X ORF, despite not encoding a Phospholipase A2 domain

We noticed that one *erythroparvovirus* species completely lacks the signature of a PLA2 domain in VP1 (as seen using HHpred [28]), unlike all other erythroparvoviruses: *ungulate erythroparvovirus 1*, also called bovine parvovirus 3 (bPARV3) [29], which is basal to the *erythroparvovirus* phylogeny [29] (Fig 1).

Synplot2 detects in the VP1 CDS of bPARV3 a region with reduced synonymous variability, in a location similar to the X ORF of erythro- and tetraparvoviruses, i.e. slightly upstream of the VP1/VP2 boundary (Fig 7B). This region corresponds to an ORF conserved in all 4 strains of bPARV3, in frame +1 relative to VP1 (Fig 7C). The reduction in synonymous variability in this region is moderate compared to other erythroparvoviruses (compare Fig 7B with Figs 3B and 4B), but could not be expected to be high, owing to the limited number of nucleotide sequences available (4) and to their limited divergence (they share over 93% sequence identity). Therefore, the signal detected by synplot2 corresponds to that expected for a protein-coding ORF, which we called "X-like" protein.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

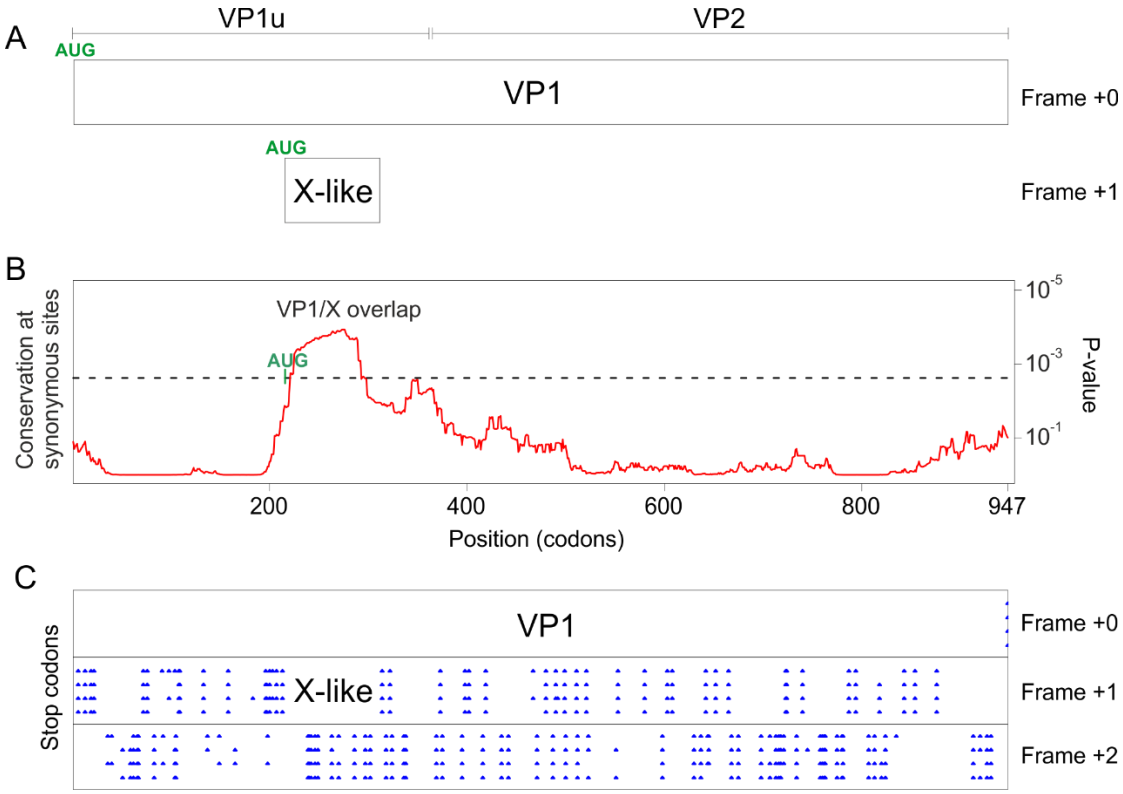

Fig 7. Synonymous-site variability in the VP1 coding sequence of bovine parvovirus 3

A. Conventions are the same as in Fig 3. The position of the VP1/VP2 boundary is approximate. Bovine parvovirus 3 VP1 does not contain a PLA2 domain, unlike all other erythro- and tetraparvoviruses (see text).

B. Conservation at synonymous sites in an alignment of the coding sequences of bPARV3 VP1 (4 sequences ranging from 93% to 99% identity), using a 45-codon sliding window in Synplot2.

C. Position of stop codons (blue) in the 3 potential frames, and gaps in alignment (gray) in the 4 sequences.

The sequence of the X-like protein is shown in Fig 5C. Its sequence features are strikingly similar to those of the X protein of erythro- and tetraparvoviruses, such has a similar length (99aas) and organization (central transmembrane segment), suggesting that they might be homologous, i.e. have a common origin. However, because bPARV3 VP1 lacks a PLA2 domain, it is not possible to examine this hypothesis by using the same approach as above, using PLA2 as an anchor to align the X-like protein of bPARV3 with the X proteins. Instead, using MAFFT-add [30], we aligned the sequence of the X-like protein of bPARV3 with the reference alignment of the X proteins of erythro-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

and tetraparvoviruses given in Fig 5. The resulting alignment, presented in S2 Fig, indicates that 2 of the 3 aas strictly conserved in erythro- and tetraparvovirus X proteins (P and L, both within the transmembrane segment) are also conserved in the X-like protein of bPARV3.

Thus, the X-like protein of bPARV3 might be homologous to the X protein of erythro-and tetraparvoviruses, given their similarity in overall organization and in sequence features. However, it is not yet possible to be certain of this homology in the absence of a PLA2 domain and of sequences intermediate between bPARV3 and other erythroparvoviruses (see Discussion).

Porcine parvovirus 2 does not encode an X ORF, but encodes a "Z ORF" overlapping VP1

As mentioned above, there is no X-like ORF in porcine parvovirus 2 (pPARV2) (also called cnvirus [31]), which belongs to the species *Ungulate tetraparvovirus 3*, and is basal to the *tetraparvovirus* phylogeny [31] (Fig 1). We examined its VP1 coding sequence with Synplot2. Three regions have a significant increase in the conservation of synonymous sites (Fig 8B):

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

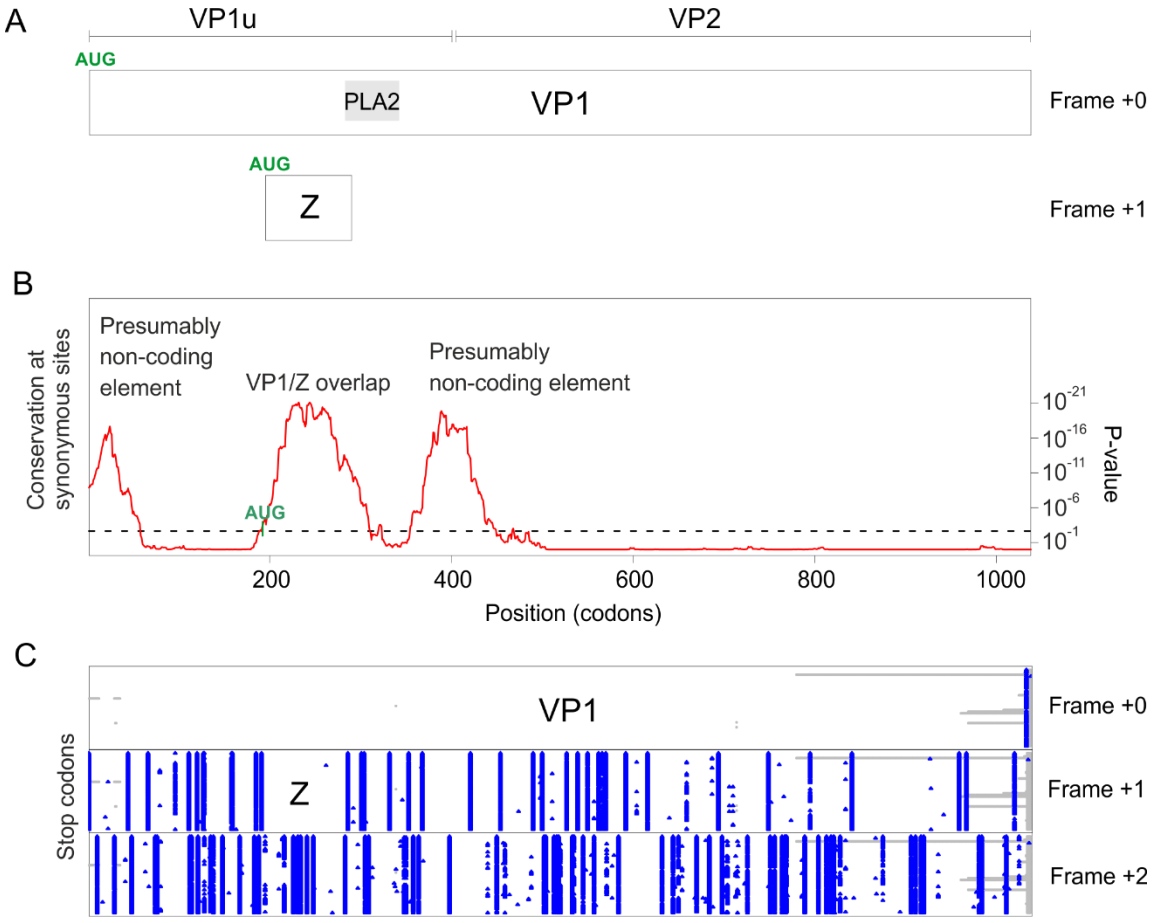

Fig 8. Synonymous-site variability in the VP1 coding sequence of porcine parvovirus 2

A. Conventions are the same as in Fig 3. The position of the VP1/VP2 boundary is approximate.

B. Conservation at synonymous sites in an alignment of the coding sequences of pPARV2 VP1 (90 sequences ranging from 93% to 99% identity), using a 45-codon sliding window in Synplot2.

C. Position of stop codons (blue) in the 3 potential frames, and gaps in alignment (gray) in the 90 sequences.

1) The first region spans codons 1-57. It is interrupted by stop codons both in +1 and +2 frames relative to VP1 (Fig 8C) and is thus unlikely to encode a protein. It may correspond to an RNA element. RNAz [21,22] could detect no secondary structure in this region.

2) The second region spans codons 193-309. It is devoid of stop codons in frame +1 relative to VP1 (Fig 8C) in all sequences of pPARV2, except one (accession number MK378188). It contains a potential AUG start codon overlapping codon 193 of VP1, conserved in all sequences. Thus, this region probably encodes a protein, which we called "Z protein". The Z ORF overlaps the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

region of VP1 upstream of the PLA2 domain and slightly extends into the N-terminus of PLA2 (Fig 8A). The sequence of the Z protein is shown in S3 Fig. It has a rather low sequence complexity, as estimated by SEG [32], and its N-and C-termini are predicted to be structurally disordered.

3) The third region spans codons 355-449. It is interrupted by stop codons both in frames +1 and +2 relative to VP1 (Fig 8C). Thus, it probably corresponds to an RNA element. RNAz [21,22] could detect no secondary structure in this region.

The X protein could either be translated by a non-conventional mechanism or expressed from an overlooked mRNA

We think that the X protein is probably translated from a standard AUG start codon, but that either this AUG start codon is accessed by a non-canonical mechanism, or the X protein is translated from a currently unmapped mRNA (presumably thanks to an overlooked splice site). Our reasoning is based on 3 observations:

1) An AUG is found near the beginning of the X ORF in absolutely all erythro- and tetraparvoviruses;
2) No known viral mRNA could encode the X ORF in a monocistronic fashion;
3) The putative AUG start codon at the start of the X ORF is not located in a position favorable to canonical translation.

We detail these observations and our reasoning below.

The X ORF contains a potential AUG start codon in all erythro- and tetraparvoviruses

In all erythro- and tetraparvoviruses, a potential AUG start codon is found at the beginning of the X ORF (see S6 Alignment). This AUG is conserved in all isolates within a given species (not shown). This observation strongly suggests that the X ORF is translated from an AUG start codon. From which viral mRNA (messenger RNA) is it likely to be translated? We discuss this point in the next paragraph.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

No known viral RNA transcript could encode the X ORF in a monocistronic fashion

A transcription profile is available only in 4 species: B19V, PARV4, simian parvovirus, and chipmunk parvovirus. In these species, there is no monocistronic mRNA that could encode the X protein. We describe their cases below.

1) B19V produces 12 known transcripts by a combination of alternative splicing and alternative polyadenylation [33,34] (for a review, see [4]). Only the transcripts that could encode the X protein are presented in Fig 9A. 5 transcripts could in principle express the X protein, although these transcripts would be polycistronic (i.e. have the capacity to express at least another protein); they are called R1', R2', R3', R4, and R5 in [4]. As Fig 9A shows, R1' could in principle express 3 proteins in addition to X: NS1, 7.5 kDa, and a truncated N-terminal version of VP1. R2' could also express 2 proteins other than X: 7.5 kDa and a truncated N-terminus of VP1. R3', also called the "small" mRNA [35] could encode a truncated N-terminus of VP1, in addition to X. R4 could express 3 proteins other than X: 7.5 kDa, VP1, and the 11kDa protein. Finally, R5 could in principle express VP1 and the 11kDa protein in addition to the X protein.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

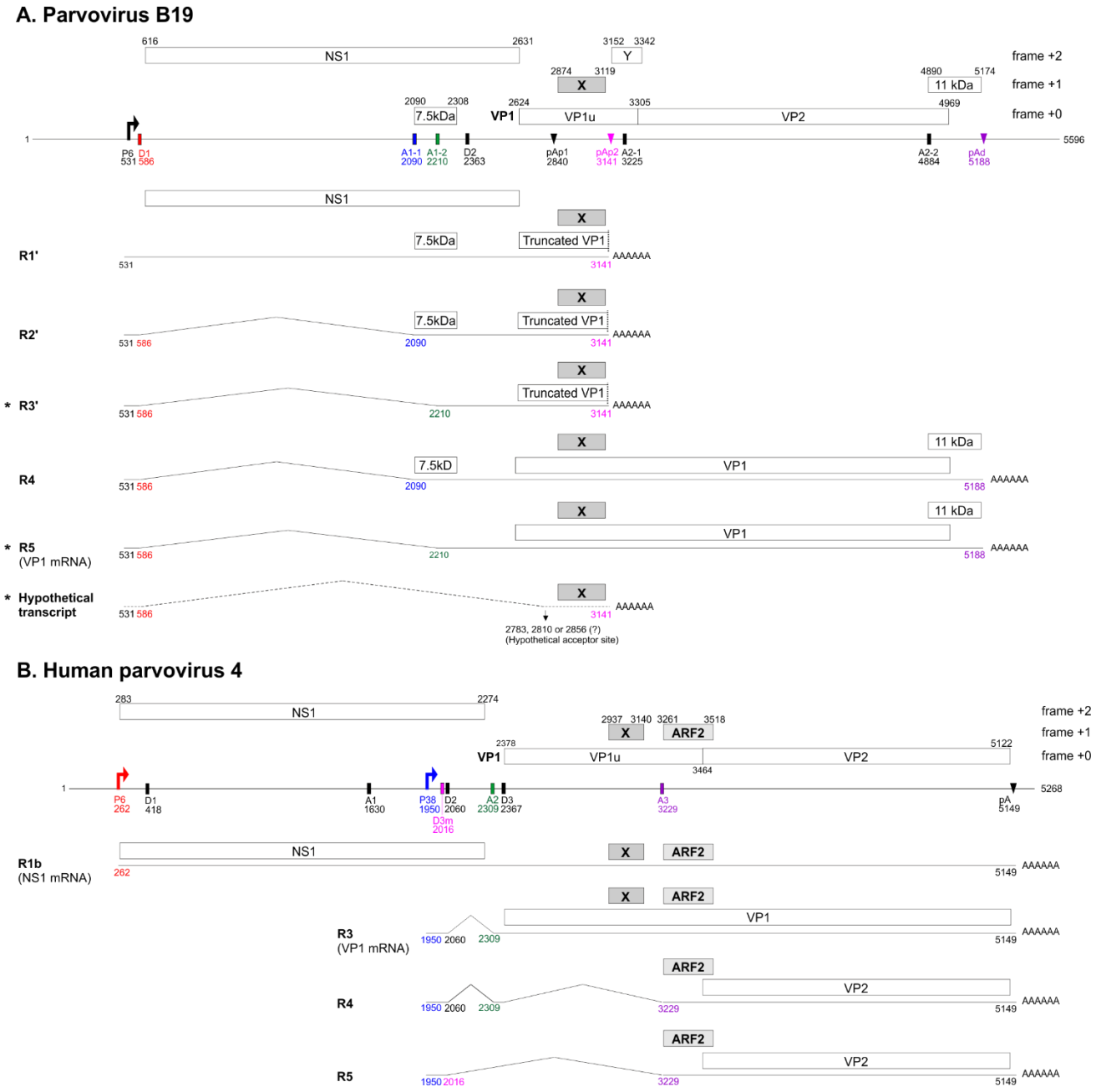

Fig 9. All currently known transcripts that could in principle express the X and ARF2 proteins are polycistronic

A. Splicing profile of B19V. Numbering refers to the B19V reference genome. Abbreviations: A1-1, A1-2, A2-1, A2-2: splicing acceptor sites. D1 and D2: splicing donor sites. pAp: proximal poly-adenylation sites. pAd: distal poly-adenylation site. P6: viral promoter. Transcripts that are most likely to encode the X protein are marked by an asterisk (*).

B. Splicing profile of PARV4. Numbering refers to the PARV4 reference genome (Table 1). Color coding is not the same as in panel A. Abbreviations: A1, A2, A3: splicing acceptor sites. D1, D2, D3:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

splicing donor sites. pA: poly-adenylation site. P6 and P38: viral promoters. Note that unlike B19V, PARV4 uses only one poly-adenylation site, but two promoters.

2) PARV4 produces 7 known transcripts by a combination of alternative splicing and alternative promoters [36]. Only the transcripts that could encode the X protein are presented in Fig 9B. Two transcripts could in principle express the X protein: the NS1mRNA and the VP1 mRNA, respectively called R1b and R3 in [36] (Fig 9B). Again, these transcripts would be polycistronic: both could in principle also express ARF1 and ARF2.

3) In simian parvovirus, the X ORF encompasses nt 2718-2963, and could be in principle expressed from at least 4 transcripts (not shown), called R4 to R7 in [37]. Again, none of these transcripts would be monocistronic: they also have the potential to encode full-length or truncated VP1, sometimes fused with other accessory proteins (10 kDa and 14 kDa, which are homologous to the B19V 7.5 kDa and 11 kDa proteins, respectively).

4) In chipmunk parvovirus, the X ORF encompasses nt 3031-3228, and could in principle be expressed from at least 3 transcripts (not shown), called R2, R3, and R5 [38]. Again, none of these transcripts would be monocistronic: they are thought to respectively encode NS1, VP1, and a putative protein unique to chipmunk parvovirus called NS2, encoded in a frame overlapping NS1.

In summary, no monocistronic mRNA could encode the X protein in the 4 species for which a transcription profile is available. Canonical translation relies on a monocistronic transcript in which the first AUG located in an optimal context is translated (see below; for a review, see [39]). Therefore, we think 3 hypotheses are likely (Fig 10): 1) the X protein is expressed from an unmapped monocistronic transcript, presumably thanks to an overlooked splice acceptor site; 2) the X protein is translated through a non-canonical mechanism from the VP1 mRNA (transcript R5); 3) the X protein is translated by a non- canonical mechanism from transcript R3', not currently known to encode a protein. We have marked the corresponding transcripts by an asterisk to the left of Fig 9A. Below we present the arguments that support each of these three hypotheses, focusing on B19V. The hypotheses are presented in the order that seemed most logical to us, and we make no claim regarding the most probable one.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

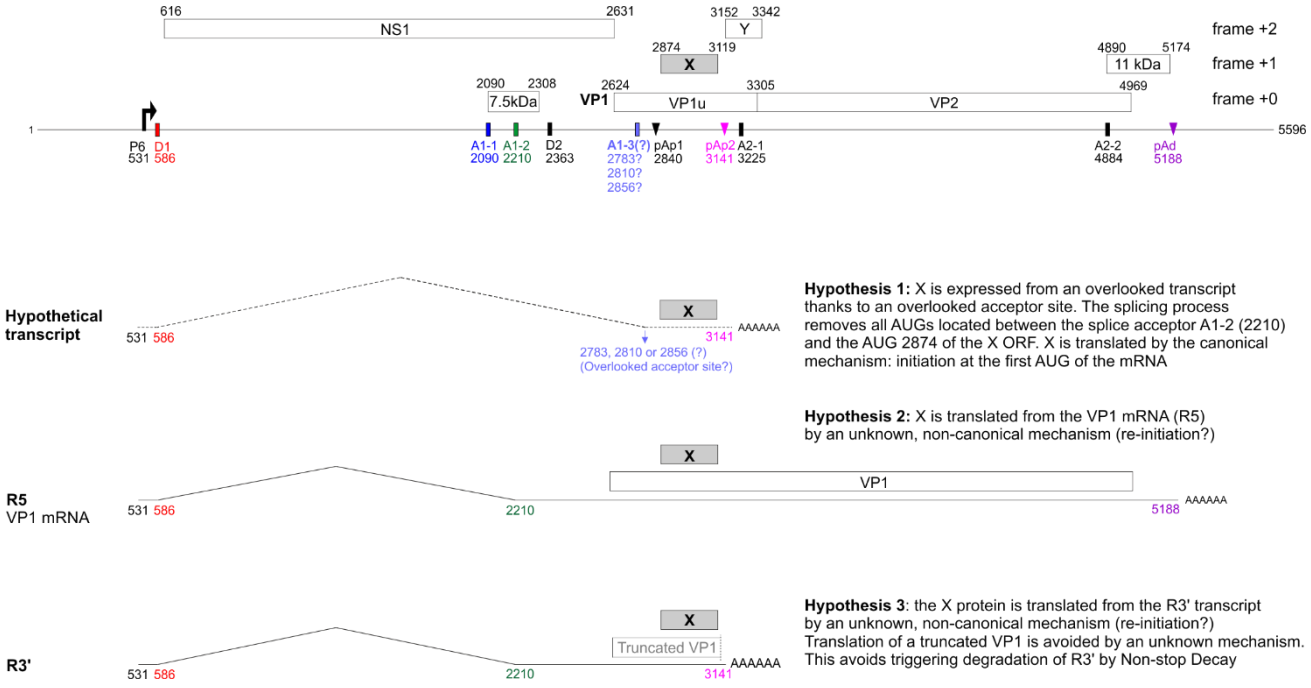

Fig 10. Three hypotheses about the mechanism by which the B19V X protein is expressed

Conventions are the same as in Fig 9.

First hypothesis: an overlooked splice acceptor site yields a monocistronic transcript that expresses the X protein

Two conditions would be required for a splice acceptor site to generate a monocistronic transcript that encodes the X protein: 1) this site must be conserved in all isolates of B19V: 2) it must be located in the region between the VP1 start codon and the presumed start codon of the X protein (nt 251-253 of the VP1 CDS).

Canonical splice acceptor sites have the sequence (C/U)AG preceded by a region rich in pyrimidines (C/U) [40]. We found 3 such potential sites, at nucleotides 158-160, 185-187, and 231-233 of the VP1 CDS. (The respective coordinates of the acceptor G in the genomic sequence are 2783, 2810 and 2856, see Figs 9A and 10). Each acceptor site would yield a monocistronic transcript that could encode the X protein, since they would splice out both the VP1 AUG start codon and the 4 following AUG codons located upstream of the presumed AUG start codon of the X protein (in red in Fig 11).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

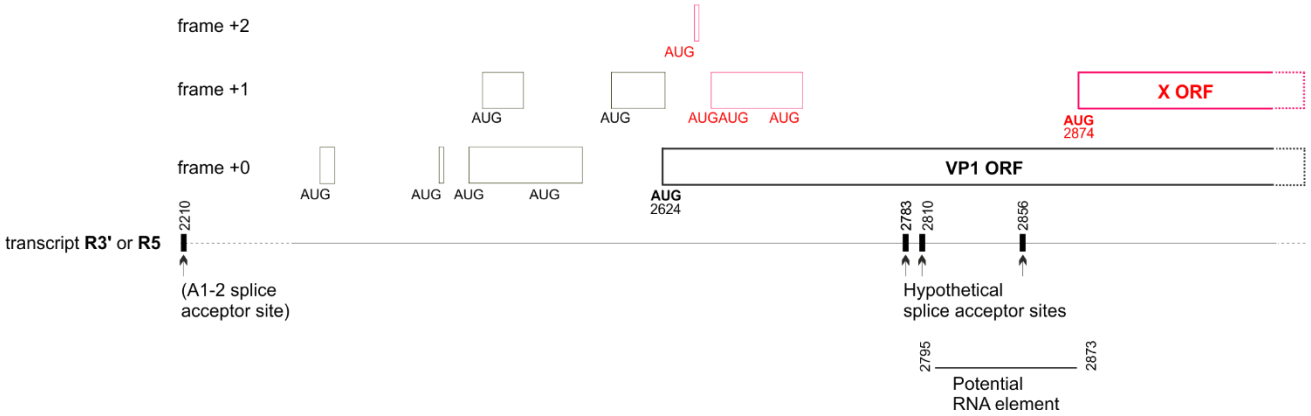

Fig 11. Elements that could influence translation of the VP1 and X ORFs: upstream mini-ORFs, potential splice acceptors, and RNA element.

Thin boxes represent mini-ORFs (that may play a role in regulating the translation of VP1 and X ORFs). The mini-ORFs in black are known to influence the translation of VP1 [41], and might also influence that of the X protein. The mini-ORFs in red are expected to influence the translation of the X protein but presumably not that of VP1. The potential RNA element corresponds to the region with a decreased variability at synonymous sites upstream of the X ORF (see text and Fig 3B).

Interestingly, these potential splice acceptor sites are located near, or in the region that has a decreased synonymous variability immediately upstream of the X ORF (Fig 3B), in nt 172-250 of the VP1 CDS (see Table 2). This region might play a role in facilitating splicing at one of these sites, which would explain its decreased variability. We have represented it as a "potential RNA element" in Fig 11. Such RNA elements sometimes have a peculiar secondary structure, but we could not detect an RNA structure in this region using RNAz [21,22].

We made no further effort to look for potential overlooked splice acceptors in other erythro- and tetraparvoviruses, since we present these observations on B19V as a starting point to guide experimental approaches.

Second hypothesis: the X protein is translated from the VP1 mRNA by a non-canonical mechanism, such as re-initiation

In vertebrates, two main factors influence canonical translation from an AUG codon: 1) the strength of the "Kozak sequence" surrounding it [42] (we present Kozak sequences and their

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

degrees of strength in the Methods); and 2) the position of the AUG codon in the mRNA. In general, the first AUG with an optimal Kozak sequence is used to initiate translation, but many exceptions are known. For instance, a downstream AUG can also initiate translation thanks to a mechanism called "leaky scanning", particularly if the first AUG has a weak Kozak sequence and the downstream AUG has an optimal Kozak sequence (for a review, see [39]).

In certain cases, a downstream AUG can initiate translation even if it is separated from the first optimal AUG by intervening AUGs, thanks to a mechanism called "re-initiation" (for a review, [41]). For instance, in B19V, the VP1 AUG codon is preceded by 7 upstream AUG codons that form mini-ORFs (Fig 11) and is accessed by re-initiation after having first initiated translation at some of these mini-ORFs [43]. Note that the presence of these 7 upstream AUGs severely decreases the translation level of VP1 [43].

In principle, the B19V X ORF might likewise be translated from the VP1 mRNA by re-initiation, since it is separated from the VP1 AUG start codon by 4 AUGs (Fig 11). However, the efficiency of translation would presumably be very low [41]. Translation of the X ORF might be facilitated in B19V by the fact that the AUG start codon of the X ORF has a strong Kozak sequence (see Methods), GUCAUGG, contrary to that of VP1, which has a weak Kozak sequence, AUUAUGA. Interestingly, in B19V, the 77 nucleotides upstream of the presumed AUG start codon of the X ORF (nt 172-250 of the VP1 CDS, see Table 2) have a significantly reduced variability in synonymous codons (see Fig 3B). This region with reduced variability might be a regulatory RNA element that would enhance the translation of the X protein.

For all erythro- and tetraparvoviruses, a similar scenario is possible (translation of the X protein from the VP1 mRNA by a non-canonical mechanism such as re-initiation). Indeed, in all species, the potential AUG start codon of the X ORF is separated from the VP1 AUG start codon by intervening AUG codons. We detail briefly the case of PARV4: both the AUG start codon of VP1 and the potential AUG start codon of the X ORF (nt 560-562 of the VP1 CDS) have a weak Kozak sequence (GCAAUGC and CAGAUGU, respectively). They are separated by 9 AUG codons, i.e. much more than in B19V (4 AUGs). In contrast to B19V, the position of the potential AUG start

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

494 codon of the X ORF of PARV4 corresponds almost exactly to the start of the region with decreased
495 synonymous variability (see Table 2 and Fig 4B).
496
497 ## Third hypothesis: the X protein is translated from the small RNA, made
498 ## monocistronic by a mechanism preventing translation of a truncated VP1
499 In B19V, translation of a truncated form of VP1 from the R1', R2' or R3' transcripts would
500 probably trigger their degradation by a mechanism of "Non-stop decay" [44], since they are devoid
501 of a stop codon for VP1 (Fig 9A). It is thus reasonable to think that translation of a truncated VP1 is
502 probably prevented somehow in the cell. In transcripts R1' and R2', this translation might be
503 naturally prevented by the fact that the VP1 ORF is located downstream of other translated ORFs.
504 However, in R3', translation of the VP1 ORF is presumably prevented by a specific mechanism.
505 This would make the R3' transcript monocistronic in practice, encoding only the X ORF. Translation
506 of X from this transcript would still require a non-canonical mechanism, such as re-initiation, since
507 the putative AUG start codon of the X ORF is preceded by 12 AUGs in the R3' transcript (Fig 11).
508
509 ## PARV4 ARF2 might be expressed by leaky scanning from the
510 ## VP2 mRNA
511
512 A methionine that corresponds to a potential AUG start codon is found immediately at the
513 beginning of the ARF2 ORF in all isolates of PARV4, porcine hokovirus, ovine hokovirus, and deer
514 tetraparvovirus (S1 Fig). In bovine hokovirus however, this methionine codon is immediately
515 followed by a stop codon (S1 Fig). A potential start AUG codon, conserved in all bovine hokovirus
516 isolates, is found 36 nucleotides downstream, but if it were used to initiate translation, bovine
517 hokovirus would encode an ARF2 amputated of 13 aas that are well conserved in other species. In
518 summary, the first AUG codon is probably used to translate ARF2, except in bovine hokovirus, in
519 which ARF2 might be translated by another mechanism, might be translated in a shorter version by
520 a downstream AUG, or not be translated.
521 From which transcript is ARF2 expressed? In PARV4, 4 transcripts could in principle express
522 it. Following the nomenclature of [36], they are called R1b (the NS1mRNA), R3 (the VP1 mRNA),

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

R4 and R5 (which both have the capacity to express VP2 in addition to ARF2) (Fig 9B). Interestingly, in the R4 and R5 transcripts, the potential AUG start codon of ARF2 is located upstream of the VP2 AUG start codon. It is thus possible that ARF2 be translated as the "primary" product of the R4 and R5 transcripts, whereas VP2 would be expressed by leaky scanning [39]. Both the ARF2 and VP2 AUG have a weak Kozak sequence, making it hard to predict their relative expression levels in this scenario.

Discussion

Sequence analyses provide evidence that the X protein must be expressed and have a crucial function

The X ORF was noticed as early as 1986 [45], but has truly lived to its name, since no experimental support has ever been provided for its translation or function in infected cells. Indeed, substituting its presumed start codon by a stop codon had no effect on replication, infectivity, or capsid production in cells permissive for B19 [9].

In contrast, earlier sequence analyses provided support for the translation of a functional product of the X ORF, by detecting a decrease in synonymous codon variability in the region of VP1 that it overlaps [14]. Here we quantify this reduction, using Synplot2, and show that it is highly significant. In addition, we show that the X ORF is conserved not only in all erythroparvoviruses but also in the closely related tetraparvoviruses (in which it is called ARF1 [10]). Given the high rate of evolution of viruses, the conservation of the X ORF in two genera provides additional evidence, altogether compelling, that it must be expressed and play a crucial function.

Why would the X protein have escaped detection for so long? A first hypothesis is that it could be produced only at low levels. This hypothesis fits well with our observations about its potential mechanism of expression: on the one hand, if the X protein is translated from an overlooked transcript, this transcript must be expressed at low levels to have escaped detection. On the other hand, if the X protein is translated by re-initiation, its translation would be expected to occur at low levels [41].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

A second hypothesis is that the X protein could be expressed only in certain conditions or cell types (B19V being extraordinarily narrow in the range of cells it infects [4]). However, a study showed that it can be expressed in a wide variety of cells (permissive, semi-permissive, or non-permissive) from a plasmid [46]. Therefore, its absence of detection so far might be caused by its expression being restricted to a certain time period and/or certain conditions of infection, rather than to a certain type of cells.

Finally, the low expected size of the X protein (9 kDa) could have prevented its detection in standard protein detection experiments.

Experimental studies of the X protein provide very few clues

Although there are no data regarding the X protein in infected cells, two experimental studies provide some hints about this protein. The first relies on indirect evidence. A genomic clone of B19V, pB19-FL, does not produce infectious virus [47]. A comparison with other infectious genomic clones flagged 3 substitutions which were unique to pB19-FL, and might thus be responsible for its lack of infectivity. One of these, A51V (in bold in Fig 5), occurs in the X protein, within its predicted transmembrane segment. The 2 other substitutions occur within NS1 (F526L) and VP1 (E176K, located in the C-terminus of the phospholipase A2 (PLA2) domain, not visible in Fig 6). The substitution within VP1 is only in part responsible for the lack of viral infectivity, and thus it is possible that the substitution A51V is also in part responsible for it; this was not tested in the study [47].

The second study [48] reported that the X protein transactivated the P6 viral promoter (which controls the expression of all B19V RNA transcripts, see Fig 9A), when transfected in HeLa cells. The authors hypothesized that this effect was indirect, since the promoter is localized in the nucleus. The study also reported that expression of the X protein into HeLa cells resulted in no visible change.

The X protein is not homologous to the protoparvovirus SAT protein

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

An earlier work on PARV4 [10] hypothesized that the ARF1/X protein was homologous to the SAT protein, another short, transmembrane protein encoded in the +1 frame of the VP1 gene in the genus *protoparvovirus* [49]. However, SAT and X cannot be homologous (i.e. have a common origin), since SAT is encoded by the N-terminus of VP2, immediately downstream of the region encoding the PLA2 domain (our observations), unlike the X protein, which overlaps the N-terminus of PLA2 (see Figs 3 and 4).

The X ORF most probably originated by overprinting the VP1 ORF

Most overlapping gene pairs originate by overprinting, a process in which substitutions in an ancestral reading frame enable the expression of a second reading frame (the novel frame), while preserving the expression of the first frame [50,51]. The ancestral frame can be identified by its phylogenetic distribution (the ORF with the widest distribution is most probably the ancestral one) [50,52], or by their codon usage [53] if both frames have the same phylogenetic distribution.

The phylogenetic distribution of X and of VP1 indicates that VP1 is necessarily the ancestral reading frame, since a PLA2 domain is found not only in most *Parvoviridae*, but also in a wide variety of metazoans and plants [54], whereas the X protein is found only in erythro- and tetraparvoviruses. Therefore, the X protein must have originated by overprinting the region encoding the PLA2 domain in the VP1 frame, in the putative common ancestor of erythro- and tetraparvoviruses.

Convergent or divergent evolution in bPARV3 and pPARV2?

Two species differ from other erythro- and tetraparvoviruses in the coding strategy in their VP1 gene: bovine parvovirus 3 (bPARV3) and porcine parvovirus 2 (pPARV2).

bARV3, currently classified as *erythroparvovirus*, does not encode a PLA2 domain, yet encodes an X-like protein in a location similar to that of other erythroparvoviruses, i.e. upstream of the VP1/VP2 boundary (Fig 7). Assuming that the ancestor of bPARV3 had a PLA2 domain like all other erythroparvoviruses, the presence of an X-like protein in bPARV3 suggests two hypotheses:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

either 1) the bPARV3 X-like ORF is unrelated to the X ORF, and originated in bPARV3 by overprinting VP1 after it had lost the PLA2 domain ("convergent evolution"); or 2) the X-like ORF is descended from the X ORF, and persisted in the viral genome even when substitutions accumulated in the region encoding the PLA2 domain to the point of erasing its sequence signature ("divergent evolution"). In the second scenario, constraints imposed by PLA2 on the X-like ORF would have disappeared, which would explain why the X-like protein is divergent in sequence.

pPARV2 is currently classified as a *tetraparvovirus*, though some authors have noticed it forms a separate sublineage [31]. pPARV2 encodes a PLA2 domain but no X protein. However, it may encode a "Z protein" immediately upstream of PLA2 (Fig 8). Again, this observation suggests two hypotheses: 1) either the Z ORF is unrelated to the X ORF; or 2) it is descended from the X ORF but lost the 3' region that encodes the transmembrane region and overlaps PLA2.

Conclusion

Like most research, our work raises more questions than it answers. One that we find of particular interest is whether, and how, the R3' transcript of B19V (Fig 9A) avoids translation of a truncated form of VP1, which would presumably trigger Non-stop decay [55,56] (for a review, see [44]), and degradation of R3'. We are not sure whether this question has been raised before.

On another note, our findings suggest that numerous proteins encoded by overlapping genes remain to be discovered in single-stranded DNA viruses (we know of at least one potential such case already flagged by sequence analyses, in human bocavirus [57]). Indeed, while a systematic effort has been made to discover overlapping genes in RNA viruses [15], this has not yet been the case in DNA viruses. We therefore recommend that readers analyze their own genome of interest using the tools and strategies presented here. This is perfectly feasible for bench virologists lacking computing skills (like the author), since the present work required no programming; all analyses were done using web-based, relatively user-friendly programs (see Methods) on a standard laptop computer. In addition, no virologist was harmed during the work.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

Materials and Methods

Sequence collection

We collected the coding sequences of VP1 for all isolates of viral species investigated here by using Blastn [58] against Genbank (30th July 2019) on the reference sequence of each species. We retained sequences with >75% nucleotide similarity over 90% of the length of the query (i.e. 90% coverage). We removed duplicate sequences, sequences containing insertions or deletions longer than 50 nucleotides with respect to the reference sequence, or those marked as "synthetic" sequences.

Nucleotide sequence alignment and analysis

To generate codon-respecting alignments based on the coding sequence of VP1, we used the program TranslatorX [26] with the "Muscle" option. The resulting codon-based alignments are in the S1-S4 Alignments.

Analysis of Kozak consensus sequences of potential AUG start codons

Kozak sequences surrounding an AUG start codon can direct translation from this AUG with varying degrees of strength [42]. The most important factor is the presence of a purine (A or G) 3 nucleotides upstream of the AUG start codon, and of a G (or less favourably an U) immediately after the AUG. For the ORFs considered here, we classified Kozak sequences of potential AUG start codons in 4 categories, as in a recent exhaustive analysis in vertebrates [42]: 1) "optimal" Kozak sequences match the consensus (A/G)CCAUGG. 2) "strong" ones match the consensus (A/G)NNAUGG, where N is any nucleotide; 3) "moderate" match the consensus (A/G)(A/C)(A/C)AUG(G/U); finally 4) "weak" Kozak sequences do not match any of these consensus sequences [42].

Detection of regions with lower synonymous substitution rate

We used Synplot2 [15] to identify overlapping functional elements, with two sizes of sliding window: 25 and 45 codons. A window of 25 codons provides better specificity, which helped us identify *how many* regions have a decreased synonymous substitution rate; whereas a window of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

45 codons provides better sensitivity, which helped us map the *precise boundaries* of the regions identified. We present Synplot2 plots computed with a window of either 25 or 45 codons, depending on which window size better shows the regions identified. The boundaries of these regions were always mapped with a window of 45 codons.

Protein sequence alignment and domain identification

All protein sequence alignments are presented using Jalview [59] with the ClustalX colouring scheme [60]. We carried out phylogenetic analyses using phylogeny.fr [61] with default options. To add unaligned sequences into a reference alignment, we used MAFFT with the --add option [30]. The S5 alignment contains the sequence alignment of all X and X-like proteins. We used HHpred [28] to identify protein domains.

Prediction of protein structural features

We used MetaDisorder [62] to predict disordered regions, in accordance with the principles described in [63], and DeepCoil [64] to predict coiled-coil regions. We used SEG [32], called via the ANNIE web server [65], to detect protein regions of low or medium sequence complexity, with parameters 45/3.75/3.4.

We used two complementary methods to detect reliably predicted transmembrane segments, as explained in [66]. First, we compared the predictions of several transmembrane prediction programs on a single protein, for each protein (“vertical approach”), by using ANNIE [65]. Second, we compared the prediction of a single program (TM-Coffee [67]) on several homologs (“horizontal” approach).

Supporting information captions

S1 Fig. Multiple sequence alignment of the tetraparvovirus ARF2 ORF.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 April 2020 doi:10.20944/preprints202004.0064.v1

Conventions are the same as in Fig 5. N-terminal Methionines that could correspond to an AUG start codon are indicated in bold. In other tetraparvoviruses more distant from PARV4 (not shown here) the ARF2 ORF is interrupted by stop codons.

S2 Fig. Alignment of the X-like protein of bovine parvovirus 3 with the reference alignment of the X protein of erythro- and tetraparvoviruses.

The corresponding alignment in text format is provided in S5 Alignment.

We used MAFFT-add to align the X-like protein of bovine parvovirus 3 with the reference alignment of the X protein of erythro- and tetraparvoviruses, derived from the alignment of the PLA2 domain, and presented in Fig 6 (see main text). The two positions strictly conserved in all X proteins and in the X-like protein are indicated. Notice that a third position, towards the C-terminus, containing a Glycine (G73 in B19V), appears to be also conserved; however this region of the alignment is not reliable, owing to the presence of gaps and to its high variability. The corresponding alignment in text format is in S5 Alignment.

S3 Fig. Sequence of the Z protein of porcine parvovirus 2.

Conventions are the same as in Fig 5.

S1 Alignment. Codon alignment of all B19V VP1 coding sequences

S2 Alignment. Codon alignment of all PARV4 VP1 coding sequences

S3 Alignment. Codon alignment of all bPARV3 VP1 coding sequences

S4 Alignment. Codon alignment of all pPARV2 VP1 coding sequences

S5 Alignment. Alignment of the X-like protein of bPARV3 with the reference alignment of the X proteins of erythroparvoviruses and tetraparvoviruses, in text format.

The corresponding alignment in Jalview format is shown in S2 Fig.

S6 Alignment. The X ORF has a potential AUG start codon in all erythro- and tetraparvoviruses

Acknowledgements

We gratefully acknowledge AE Firth for useful advice in using Synplot2 and for help with preparing the Synplot2 Figs, and S. Courtès, J Qiu and G. Gallinella for commenting on the manuscript. We thank all the authors of the user-friendly, web-based software without whom this work would not have been possible. The author would like to thank the Marie Skłodowska-Curie European programme for not funding his research project and thereby allowing him to lead a fulfilling life, doing research as a rewarding hobby.

References

1. Söderlund-Venermo M. Emerging Human Parvoviruses: The Rocky Road to Fame. Annu Rev Virol. 2019;6: annurev-virology-092818-015803. doi:10.1146/annurev-virology-092818-015803

2. Kailasan S, Agbandje-McKenna M, Parrish CR. Parvovirus Family Conundrum: What Makes a Killer? Annu Rev Virol. 2015;2: 425–450. doi:10.1146/annurev-virology-100114-055150

3. Cotmore SF, Tattersall P. Parvoviruses: Small Does Not Mean Simple. Annu Rev Virol. 2014;1: 517–537. doi:10.1146/annurev-virology-031413-085444

4. Ganaie SS, Qiu J. Recent Advances in Replication and Infection of Human Parvovirus B19. Front Cell Infect Microbiol. 2018;8: 166. doi:10.3389/fcimb.2018.00166

5. Matthews PC, Sharp C, Simmonds P, Klenerman P. Human parvovirus 4 'PARV4' remains elusive despite a decade of study. F1000Res. 2017;6: 82. doi:10.12688/f1000research.9828.1

6. Luo W, Astell CR. A Novel Protein Encoded by Small RNAs of Parvovirus B19. Virology. 1993;195: 448–455. doi:10.1006/viro.1993.1395

7. St Amand J, Beard C, Humphries K, Astell CR. Analysis of splice junctions and in vitro and in vivo translation potential of the small, abundant B19 parvovirus RNAs. Virology. 1991;183: 133–142. doi:10.1016/0042-6822(91)90126-v

8. St Amand J, Astell CR. Identification and characterization of a family of 11-kDa proteins encoded by the human parvovirus B19. Virology. 1993;192: 121–131. doi:10.1006/viro.1993.1014

9. Zhi N, Mills IP, Lu J, Wong S, Filippone C, Brown KE. Molecular and functional analyses of a human parvovirus B19 infectious clone demonstrates essential roles for NS1, VP1, and the 11-kilodalton protein in virus replication and infectivity. J Virol. 2006;80: 5941–5950. doi:10.1128/JVI.02430-05

10. Simmonds P, Douglas J, Bestetti G, Longhi E, Antinori S, Parravicini C, et al. A third genotype of the human parvovirus PARV4 in sub-Saharan Africa. J Gen Virol. 2008;89: 2299–2302. doi:10.1099/vir.0.2008/001180-0

11. Pavesi A, Vianelli A, Chirico N, Bao Y, Blinkova O, Belshaw R, et al. Overlapping genes and the proteins they encode differ significantly in their sequence composition from non-overlapping genes. PLoS ONE. 2018;13: e0202513. doi:10.1371/journal.pone.0202513

12. Firth AE, Brown CM. Detecting overlapping coding sequences with pairwise alignments. Bioinformatics. 2005;21: 282–292. doi:10.1093/bioinformatics/bti007

13. Sabath N, Landan G, Graur D. A method for the simultaneous estimation of selection intensities in overlapping genes. PLoS ONE. 2008;3: e3996. doi:10.1371/journal.pone.0003996

14. Norja P, Eis-Hübinger AM, Söderlund-Venermo M, Hedman K, Simmonds P. Rapid sequence change and geographical spread of human parvovirus B19: comparison of B19 virus evolution in acute and persistent infections. J Virol. 2008;82: 6427–6433. doi:10.1128/JVI.00471-08

15. Firth AE. Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. Nucleic Acids Res. 2014;42: 12425–12439. doi:10.1093/nar/gku981

16. Chung BY-W, Miller WA, Atkins JF, Firth AE. An overlapping essential gene in the Potyviridae. Proc Natl Acad Sci USA. 2008;105: 5897–5902. doi:10.1073/pnas.0800468105

17. Jagger BW, Wise HM, Kash JC, Walters K-A, Wills NM, Xiao Y-L, et al. An overlapping protein-coding region in influenza A virus segment 3 modulates the host response. Science. 2012;337: 199–204. doi:10.1126/science.1222213

18. Ratinier M, Caporale M, Golder M, Franzoni G, Allan K, Nunes SF, et al. Identification and characterization of a novel non-structural protein of bluetongue virus. PLoS Pathog. 2011;7: e1002477. doi:10.1371/journal.ppat.1002477

19. Zádori Z, Szelei J, Lacoste MC, Li Y, Gariépy S, Raymond P, et al. A viral phospholipase A2 is required for parvovirus infectivity. Dev Cell. 2001;1: 291–302.

20. Dorsch S, Liebisch G, Kaufmann B, von Landenberg P, Hoffmann JH, Drobnik W, et al. The VP1 unique region of parvovirus B19 and its constituent phospholipase A2-like activity. J Virol. 2002;76: 2014–2018. doi:10.1128/jvi.76.4.2014-2018.2002

21. Gruber AR, Neubock R, Hofacker IL, Washietl S. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. Nucleic Acids Research. 2007;35: W335–W338. doi:10.1093/nar/gkm222

22. Washietl S, L. Hofacker I. Identifying Structural Noncoding RNAs Using RNAz. In: Baxevanis AD, Davison DB, Page RDM, Petsko GA, Stein LD, Stormo GD, editors. Current Protocols in Bioinformatics. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2007. p. bi1207s19. doi:10.1002/0471250953.bi1207s19

23. Leisi R, Di Tommaso C, Kempf C, Ros C. The Receptor-Binding Domain in the VP1u Region of Parvovirus B19. Viruses. 2016;8: 61. doi:10.3390/v8030061

24. Baker JA, Wong W-C, Eisenhaber B, Warwicker J, Eisenhaber F. Charged residues next to transmembrane regions revisited: “Positive-inside rule” is complemented by the “negative inside depletion/outside enrichment rule.” BMC Biol. 2017;15: 66. doi:10.1186/s12915-017-0404-4

787 25. Wong W-C, Maurer-Stroh S, Eisenhaber F. Not all transmembrane helices are born equal:
788 Towards the extension of the sequence homology concept to membrane proteins. Biol Direct.
789 2011;6: 57. doi:10.1186/1745-6150-6-57

790 26. Abascal F, Zardoya R, Telford MJ. TranslatorX: multiple alignment of nucleotide sequences
791 guided by amino acid translations. Nucleic Acids Res. 2010;38: W7-13.
792 doi:10.1093/nar/gkq291

793 27. Lo MK, Søgaard TM, Karlin DG. Evolution and structural organization of the C proteins of
794 paramyxovirinae. PLoS ONE. 2014;9: e90003. doi:10.1371/journal.pone.0090003

795 28. Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection
796 and structure prediction. Nucleic Acids Res. 2005;33: W244-248. doi:10.1093/nar/gki408

797 29. Allander T, Emerson SU, Engle RE, Purcell RH, Bukh J. A virus discovery method
798 incorporating DNase treatment and its application to the identification of two bovine parvovirus
799 species. Proc Natl Acad Sci USA. 2001;98: 11609–11614. doi:10.1073/pnas.211424698

800 30. Katoh K, Frith MC. Adding unaligned sequences into an existing alignment using MAFFT and
801 LAST. Bioinformatics. 2012;28: 3144–3146. doi:10.1093/bioinformatics/bts578

802 31. Wang F, Wei Y, Zhu C, Huang X, Xu Y, Yu L, et al. Novel parvovirus sublineage in the family
803 of Parvoviridae. Virus Genes. 2010;41: 305–308. doi:10.1007/s11262-010-0506-3

804 32. Wootton JC. Non-globular domains in protein sequences: automated segmentation using
805 complexity measures. Comput Chem. 1994;18: 269–285.

806 33. Yoto Y, Qiu J, Pintel DJ. Identification and characterization of two internal cleavage and
807 polyadenylation sites of parvovirus B19 RNA. J Virol. 2006;80: 1604–1609.
808 doi:10.1128/JVI.80.3.1604-1609.2006

809 34. Ozawa K, Ayub J, Hao YS, Kurtzman G, Shimada T, Young N. Novel transcription map for the
810 B19 (human) pathogenic parvovirus. J Virol. 1987;61: 2395–2406.

811 35. Bua G, Manaresi E, Bonvicini F, Gallinella G. Parvovirus B19 Replication and Expression in
812 Differentiating Erythroid Progenitor Cells. Qiu J, editor. PLoS ONE. 2016;11: e0148547.
813 doi:10.1371/journal.pone.0148547

814 36. Lou S, Xu B, Huang Q, Zhi N, Cheng F, Wong S, et al. Molecular characterization of the newly
815 identified human parvovirus 4 in the family Parvoviridae. Virology. 2012;422: 59–69.
816 doi:10.1016/j.virol.2011.09.033

817 37. Liu Z, Qiu J, Cheng F, Chu Y, Yoto Y, O'Sullivan MG, et al. Comparison of the transcription
818 profile of simian parvovirus with that of the human erythrovirus B19 reveals a number of unique
819 features. J Virol. 2004;78: 12929–12939. doi:10.1128/JVI.78.23.12929-12939.2004

820 38. Chen Z, Chen AY, Cheng F, Qiu J. Chipmunk parvovirus is distinct from members in the genus
821 Erythrovirus of the family Parvoviridae. PLoS ONE. 2010;5: e15113.
822 doi:10.1371/journal.pone.0015113

823 39. Firth AE, Brierley I. Non-canonical translation in RNA viruses. Journal of General Virology.
824 2012;93: 1385–1409. doi:10.1099/vir.0.042499-0

825 40. Baralle M, Baralle FE. The splicing code. Biosystems. 2018;164: 39–48.
826 doi:10.1016/j.biosystems.2017.11.002

41. Gupta A, Bansal M. RNA-mediated translation regulation in viral genomes: computational advances in the recognition of sequences and structures. Briefings in Bioinformatics. 2019; bbz054. doi:10.1093/bib/bbz054

42. Hernández G, Osnaya VG, Pérez-Martínez X. Conservation and Variability of the AUG Initiation Codon Context in Eukaryotes. Trends in Biochemical Sciences. 2019; S096800041930146X. doi:10.1016/j.tibs.2019.07.001

43. Ozawa K, Ayub J, Young N. Translational regulation of B19 parvovirus capsid protein production by multiple upstream AUG triplets. J Biol Chem. 1988;263: 10922–10926.

44. Karamyshev AL, Karamysheva ZN. Lost in Translation: Ribosome-Associated mRNA and Protein Quality Controls. Front Genet. 2018;9: 431. doi:10.3389/fgene.2018.00431

45. Shade RO, Blundell MC, Cotmore SF, Tattersall P, Astell CR. Nucleotide sequence and genome organization of human parvovirus B19 isolated from the serum of a child during aplastic crisis. J Virol. 1986;58: 921–936.

46. Zhi N, Wan Z, Liu X, Wong S, Kim DJ, Young NS, et al. Codon optimization of human parvovirus B19 capsid genes greatly increases their expression in nonpermissive cells. J Virol. 2010;84: 13059–13062. doi:10.1128/JVI.00912-10

47. Filippone C, Zhi N, Wong S, Lu J, Kajigaya S, Gallinella G, et al. VP1u phospholipase activity is critical for infectivity of full-length parvovirus B19 genomic clones. Virology. 2008;374: 444–452. doi:10.1016/j.virol.2008.01.002

48. Dong Y, Huang Y, Wang Y, Xu P, Yang Y, Liu K, et al. The effects of the 11 kDa protein and the putative X protein on the p6 promoter activity of Parvovirus B19 in Hela cells. Virus Genes. 2013;46: 167–169. doi:10.1007/s11262-012-0839-1

49. Zádori Z, Szelei J, Tijssen P. SAT: a late NS protein of porcine parvovirus. J Virol. 2005;79: 13129–13138. doi:10.1128/JVI.79.20.13129-13138.2005

50. Rancurel C, Khosravi M, Dunker AK, Romero PR, Karlin D. Overlapping genes produce proteins with unusual sequence properties and offer insight into de novo protein creation. J Virol. 2009;83: 10719–10736. doi:10.1128/JVI.00595-09

51. Keese PK, Gibbs A. Origins of genes: “big bang” or continuous creation? Proc Natl Acad Sci USA. 1992;89: 9489–9493. doi:10.1073/pnas.89.20.9489

52. Sabath N, Wagner A, Karlin D. Evolution of viral proteins originated de novo by overprinting. Mol Biol Evol. 2012;29: 3767–3780. doi:10.1093/molbev/mss179

53. Pavesi A, Magiorkinis G, Karlin DG. Viral proteins originated de novo by overprinting can be identified by codon usage: application to the “gene nursery” of Deltaretroviruses. PLoS Comput Biol. 2013;9: e1003162. doi:10.1371/journal.pcbi.1003162

54. Schaloske RH, Dennis EA. The phospholipase A2 superfamily and its group numbering system. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids. 2006;1761: 1246–1259. doi:10.1016/j.bbalip.2006.07.011

55. Frischmeyer PA, van Hoof A, O’Donnell K, Guerrerio AL, Parker R, Dietz HC. An mRNA surveillance mechanism that eliminates transcripts lacking termination codons. Science. 2002;295: 2258–2261. doi:10.1126/science.1067338

56. van Hoof A, Frischmeyer PA, Dietz HC, Parker R. Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. Science. 2002;295: 2262–2264. doi:10.1126/science.1067272

57. Sealfon RS, Lin MF, Jungreis I, Wolf MY, Kellis M, Sabeti PC. FRESCo: finding regions of excess synonymous constraint in diverse viruses. Genome Biol. 2015;16: 38. doi:10.1186/s13059-015-0603-7

58. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25: 3389–3402. doi:10.1093/nar/25.17.3389

59. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25: 1189–1191. doi:10.1093/bioinformatics/btp033

60. Procter JB, Thompson J, Letunic I, Creevey C, Jossinet F, Barton GJ. Visualization of multiple alignments, phylogenies and gene family evolution. Nat Methods. 2010;7: S16-25. doi:10.1038/nmeth.1434

61. Dereeper A, Guignon V, Blanc G, Audic S, Buffet S, Chevenet F, et al. Phylogeny.fr: robust phylogenetic analysis for the non-specialist. Nucleic Acids Res. 2008;36: W465-469. doi:10.1093/nar/gkn180

62. Kozlowski LP, Bujnicki JM. MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. BMC Bioinformatics. 2012;13: 111. doi:10.1186/1471-2105-13-111

63. Ferron F, Longhi S, Canard B, Karlin D. A practical overview of protein disorder prediction methods. Proteins. 2006;65: 1–14. doi:10.1002/prot.21075

64. Ludwiczak J, Winski A, Szczepaniak K, Alva V, Dunin-Horkawicz S. DeepCoil-a fast and accurate prediction of coiled-coil domains in protein sequences. Bioinformatics. 2019;35: 2790–2795. doi:10.1093/bioinformatics/bty1062

65. Ooi HS, Kwo CY, Wildpaner M, Sirota FL, Eisenhaber B, Maurer-Stroh S, et al. ANNIE: integrated de novo protein sequence annotation. Nucleic Acids Res. 2009;37: W435-440. doi:10.1093/nar/gkp254

66. Kuchibhatla DB, Sherman WA, Chung BYW, Cook S, Schneider G, Eisenhaber B, et al. Powerful sequence similarity search methods and in-depth manual analyses can identify remote homologs in many apparently “orphan” viral proteins. J Virol. 2014;88: 10–20. doi:10.1128/JVI.02595-13

67. Floden EW, Tommaso PD, Chatzou M, Magis C, Notredame C, Chang J-M. PSI/TM-Coffee: a web server for fast and accurate multiple sequence alignments of regular and transmembrane proteins using homology extension on reduced databases. Nucleic Acids Res. 2016;44: W339-343. doi:10.1093/nar/gkw300