 doi:10.20944/preprints202002.0234.v1

Sequence properties of the MAAP protein and of the VP1 capsid protein of adeno-associated viruses

Running Title: Sequence and evolution of AAV MAAP and VP1

Author: David G Karlin, davidgkarlin@gmail.com
Independent scholar, 13007 Marseille, France
ORCID number: 0000-0002-3033-7013

Keywords: adeno-associated virus, protein sequence analysis, overlapping genes, amino acid depletion, cysteine depletion, tyrosine depletion, capsid design, membrane-binding amphipathic helix.

Abstract

Adeno-associated viruses (AAVs, genus *dependoparvovirus*) are promising gene therapy vectors. In strains AAV1-12, the capsid gene VP1 encodes a recently discovered protein, MAAP, in an overlapping frame. MAAP binds the cell membrane by an unknown mechanism. We discovered that MAAP is also encoded in bovine AAV and in porcine AAVs (which have shown promise for gene transfer into muscle tissues), in which it is probably translated from a non-canonical start codon. MAAP is predicted to be mostly disordered except for a predicted C-terminal, membrane-binding amphipathic α-helix. MAAP has a highly unusual composition. In particular, it lacks internal methionines, and is devoid of tyrosines in most strains. Unexpectedly, we discovered that the N-terminus of VP1 also lacks several amino acids. In all AAVs that encode MAAP, the first 200 aas of VP1 are devoid of internal methionines, probably owing to a selection against ATG codons that could prevent translation of MAAP and of capsid isoforms (VP2, VP3). The N-terminus of VP1 also lacks cysteines, likely to avoid the formation of disulfide bridges when it becomes exposed outside of the capsid during post-endocytic trafficking. Finally, the region common to VP1 and VP2 lacks tyrosine in the vast majority of AAVs that encode MAAP. Avoiding these "forbidden" aas in MAAP and VP1 when creating recombinant AAV capsids might increase the efficiency of capsid design. Conversely, the presence of "forbidden" aas in some rare strains probably indicates that they have unusual properties that could help us understand the viral cycle.

Introduction

Adeno-Associated Viruses (AAVs) are small, non-enveloped viruses that hold great promise as gene therapy vectors [1]. AAVs belong to the genus *dependoparvovirus* ,in the family *Parvovirinae* (for reviews, [2–4]). The model species is *dependoparvovirus A*, of which the prototype strain is AAV2. AAVs encode a replicase protein and a capsid protein, of which 3 isoforms are made: VP1, VP2 and VP3 (Fig 1).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0234.v1

The N-terminus of VP1 (VP1u) contains a domain that has PhosphoLipase A2 ("PLA2") activity [5] and protease activity [6].

AAV2 encodes 2 additional proteins in reading frames overlapping the capsid (Fig 1): AAP (Assembly-Activating Protein) [7], and the recently discovered MAAP (Membrane-Associated Accessory Protein) [8]. MAAP is translated from a non-canonical start codon, CTG and has been reported only in the species *dependoparvovirus A* (strains AAV1-12 except AAV5) and in AAV5, a strain of the species *dependoparvovirus B*. MAAP is associated with the cell membrane and limits the production of other AAV strains through competitive exclusion [8].

Overlapping gene arrangements, such as VP1/MAAP, are thought to originate by a process called "overprinting" [9], in which mutations in an ancestral reading frame enable the expression of a second reading frame, while preserving the expression of the first frame [10]. Consequently, each pair of overlapping frames contains one ancestral frame and one originated *de novo* [11], as opposed to the classical scenario of gene origination by duplication or horizontal gene transfer [12].

Proteins originated *de novo* by overprinting generally have a highly biased composition [10,11], tend to be structurally disordered [11], and evolve faster than the ancestral reading frame [13]. These proteins often play an important role in viral pathogenicity, for instance by neutralizing the host interferon response [14,15] or by inducing apoptosis in host cells [16,17]. Those characterized so far have previously unknown mechanisms of action [18,19], and the minority that are not disordered have previously unknown 3D structural folds [20,21].

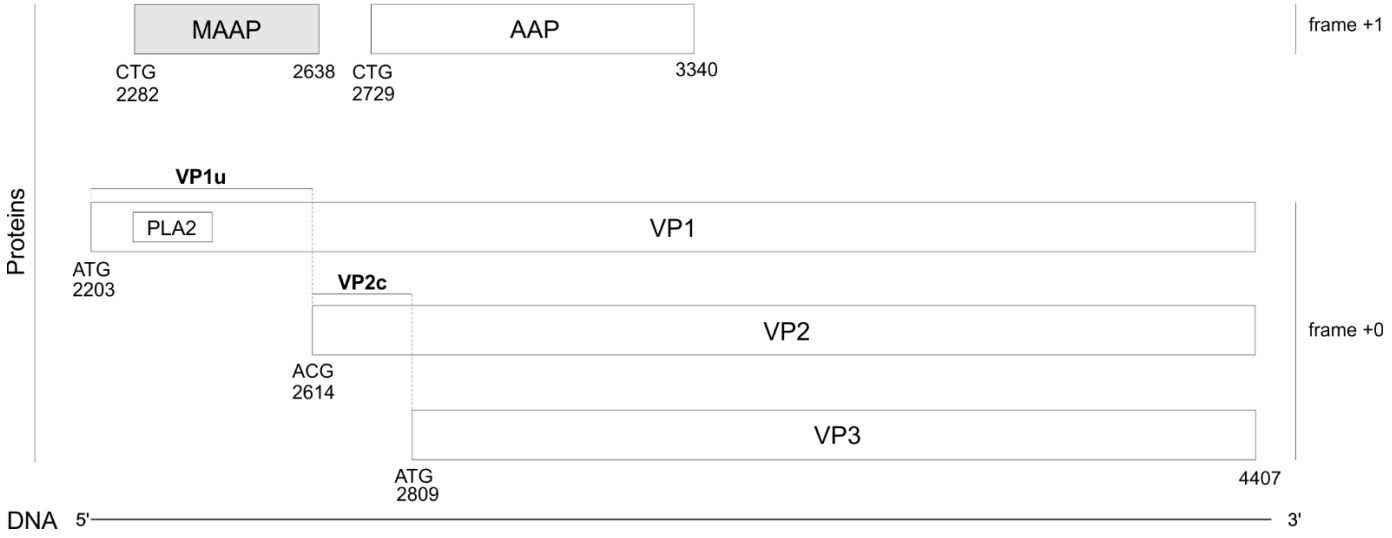

Fig 1: Coding organization of the capsid gene of AAV2

The capsid gene (bottom) encodes 5 ORFs (Open reading frames), represented as boxes, in frame +0 or +1. Numbering corresponds to the genomic coordinates. PLA2: PhosphoLipase A2 domain. VP1u: region unique to VP1. VPc: region common to VP1 and VP2. MAAP: membrane-associated accessory protein. AAP: Assembly-activating protein.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0234.v1

Here, we mapped the taxa in which a MAAP ORF is present; its putative start codon; its sequence properties and in particular its probable membrane-binding region. We also determined the probable evolutionary origin of MAAP. Finally, we discovered amino acids (aas) and a codon that are absent from MAAP and from several regions of VP1.

Results

A MAAP ORF is found in dependoparvoviruses A, B, and in porcine AAVs

Ogden *et al* reported that a MAAP ORF is present in the species *dependoparvovirus A* and in AAV5, a strain of *dependoparvovirus B* [8]. Here we found that MAAP is also present in the other strain of *dependoparvovirus B*, bovine AAV, and in all porcine AAVs, which are related to *dependoparvovirus A* and *B* [22,23] but currently not assigned a taxonomic rank. The MAAP ORF contains no ATG that could act as a start codon (Fig 2) in any of these species, and is thus probably translated by a non-canonical mechanism.

Such mechanisms include initiation of translation at other codons than ATG, which differ from ATG by only one nucleotide (e.g. CTG, GTG, ACG, TTG, and more rarely, ATT, ATC, ATA and AGG) [24]. The efficiency of initiation at these codons is typically much lower [24], and depends in part on the strength of the "Kozak sequence" surrounding them [25]. The most efficient non-ATG start codons were reported to be CTG, ACG, and GTG (in an optimal Kozak context, CTG can drive translation up to 50% of ATG codons). ATT, ATA, ATC, and TTG are less efficient but could still produce translation at levels ranging from 5 to 10% from ATG codons when found with an optimal Kozak sequence [24].

In AAV2, translation of MAAP initiates at a CTG codon (nucleotides 80-82 of the VP1 CDS, see Fig 2); this codon is conserved in all strains of *dependoparvovirus A*, and has a strong Kozak sequence (see Methods), AACCTGG (Fig 2). This CTG is also conserved in porcine AAVs (nucleotide 77-79 in the VP1 CDS of AAVpo1), with the same strong Kozak sequence as in *dependoparvovirus A* (Fig 2), and is thus probably used as the start codon of MAAP in porcine AAVs. This CTG is expected to drive translation of MAAP with relatively high efficiency (see above)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0234.v1

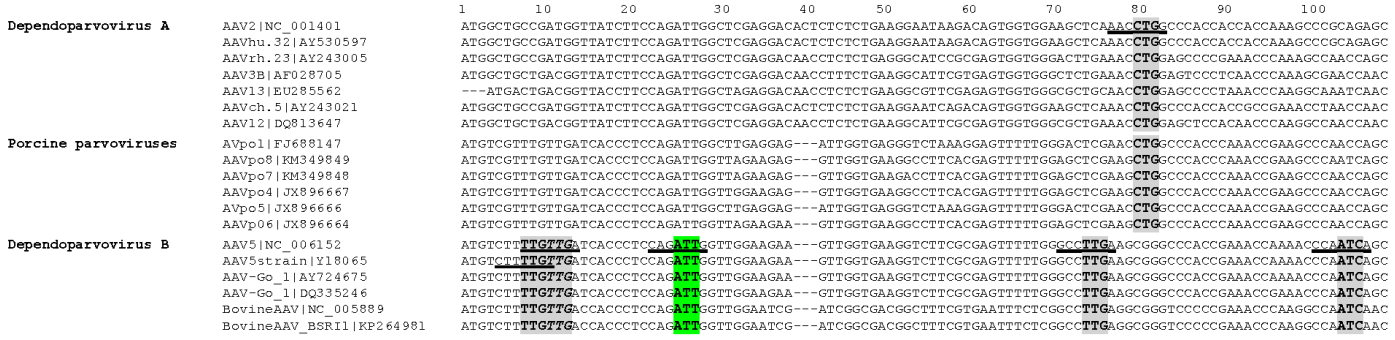

Fig 2: Potential (non-canonical) start codons of the MAAP ORF

5' region of the coding sequence of VP1 in representative species encoding the MAAP ORF, with the Genbank accession number. The numbering above the alignment is according to the AAV2 sequence. The alignment was generated from the VP1 reading frame using TranslatorX (see Methods).

The first codon in each sequence is the ATG start codon of VP1. Potential start codons of MAAP (in frame +1) are highlighted; they are all non-canonical (see text). The Kozak sequence of each is underlined in selected taxa. The ATT codon which has a moderate Kozak sequence is highlighted in green. Two contiguous TTG codons are present in *dependoparvovirus B*; the second one is in italics, for clarity.

In contrast, in *dependoparvovirus* B, there is no CTG codon in the 5' region of the MAAP ORF (Fig 2). If we assume that the MAAP start codon is conserved in dependoparvoviruses B, there are 5 potential non-canonical start codons in the first 80 nucleotides (Fig 2): two contiguous TTGs, an ATT, a third TTG, and an ATC. Only the ATT codon (in green in Fig 2) has a "moderate" Kozak sequence (CAGATTG); all other codons have a weak Kozak sequence. Therefore, ATT is a somewhat more likely candidate start codon for MAAP, since non-canonical translation initiation is very sensitive to the strength of the Kozak sequence, in particular for weak start codons such as TTG, ATT and ATC (see above) [26].

Of course, even with a moderate Kozak sequence, ATT would presumably drive low levels of translation, and thus other mechanisms of expression than initiation at non-canonical start codons should be kept in mind in *dependoparvovirus B* (such as, for example, post-transcriptional nucleotide insertion [27]).

As a note of caution, the first aa of MAAP could be either methionine or leucine if MAAP is translated from a CTG or TTG start codon (which are normally decoded as leucine) [24]. If ATT or ATC (which normally encode isoleucine) are used as a start codon, it is known that they can be decoded as a methionine [24]; but to our knowledge the possibility that they are decoded as isoleucine has not been excluded. Thus, only experimental approaches can reveal what is the first aa of MAAP. In principle incorporation of radioactive methionine would be enough to prove that methionine is the first aa, since MAAP is devoid of internal methionines (see below).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0234.v1

MAAP is predicted to be mostly structurally disordered and to contain short protein-binding regions

Fig 3 presents a multiple sequence alignment of the MAAP protein from representative species. MAAP contains 6 discernable regions (see Methods):

1) a short, disordered N-terminus (aa 1-15 in AAV2), predicted to have the potential to bind other proteins;

2) a short, hydrophobic stretch containing at least one cysteine (C) (aa 16-22), predicted to form a β-strand and to have the potential to bind other proteins;

3) a central, T/S-rich region predicted disordered (aa 23-73), rich in charged aas in all species except bovine AAV. Within this region, T43 and T69 had a high probability (90%) of being phosphorylated. Interestingly, the T/S-rich region closely corresponds to the region of the VP1 coding sequence encoding the PLA2 (PhosphoLipase A2) domain, indicated above the alignment;

4) a region devoid of predicted secondary structure (aa 74-83), predicted to be ordered and to have the potential to bind other proteins;

5) a disordered region predicted to have the potential to form an α-helix (aa 84-94);

6) a C-terminal, amphipathic α-helix predicted to bind membranes (see below), in aa 95-116.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0234.v1

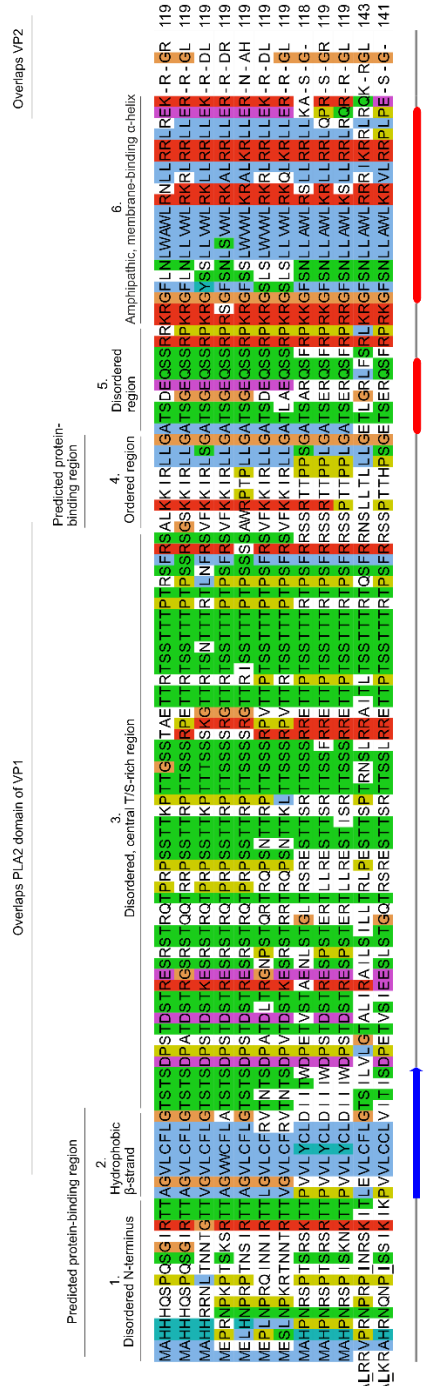

Fig 3: MAAP is mostly disordered and contains a predicted C-terminal, membrane-binding α-helix

Multiple sequence alignment of the MAAP proteins. In *dependoparvovirus A* and porcine AAVs, the first aa of MAAP is represented as a Methionine (M), although it is encoded by CTG, which might also be decoded as a leucine (L) (see text). The potential first aas of MAAP in *dependoparvovirus B* are in bold and underlined (see text and Fig 2); they might also be decoded as methionine. The corresponding alignment in FASTA format is in S1 Alignment.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0234.v1

MAAP contains a predicted amphipathic, membrane-binding α-helix

MAAP binds membranes [8], though its sequence contains no potential transmembrane region (which would require a hydrophobic stretch ≥12aas). Thus, we thought it may bind membranes through an amphipathic, membrane-binding α-helix. These helices are composed of a hydrophobic face, which binds the lipidic membrane, and of a polar face, generally positively charged, which is attracted by the negatively charged membrane [28]. We looked for such helices using Amphipaseek [29] and Heliquest [30] (see Methods).

Both Amphipaseek and Heliquest confidently predict that the C-terminus of AAV2 MAAP contains an amphipathic, membrane-binding α-helix (aa 96-116) (see Fig 3 and Table 1). Fig 4 depicts this region as a helical wheel representation [31]. As expected, it is clearly divided into a hydrophobic face (bottom) and polar face (top), the latter having a high positive net charge (+6). Amphipaseek and Heliquest confidently predicted that in other species, this region also forms an amphipathic, membrane-binding α-helix (Table 1; notice in Fig 3 how the alternance of hydrophobicity and polar aas is conserved in all strains, as well as the high net positive charge).

Table 1: Properties of the predicted membrane-binding, amphipathic α-helix of MAAP

Species	Region predicted by Amphipaseek	Region predicted by Heliquest	Hydrophobic moment (<μH>)	Net charge (z)	Heliquest Discriminating factor (D)[a]	Status
AAV2	101-111	96-116	0.558	+6	2.5	Reliable
Bovine AAV	113-132[b]	110-135	0.63	+9	3.56	Reliable
AAV5	123-132[b]	113-138	0.36	+7	2.65	Reliable
AAV5	5-22[b, c]	1-22	0.439	+3	1.40	Reliable

(a) The Discriminating factor D is equal to 0.944*<μH> +0.33*z. A region is predicted as a “potential” lipid-binding amphipathic α-helix if 0.68<D<1.34, and as "reliable" if D ≥1.34.

(b) Numbering is given assuming that MAAP starts at the first potential non-canonical start, $_{8}$TTG$_{10}$ (see Fig 3). However, the real start codon is unknown in AAV5 and bovine AAV.

(c) Since the start codon is unknown in AAV5, much of this region might in fact not be translated (i.e. MAAP might start downstream and not contain this predicted **amphipathic helix**).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0234.v1

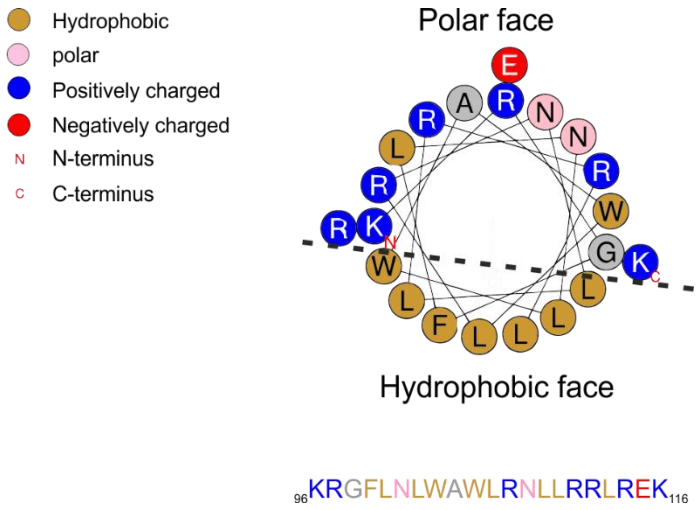

Fig 4: Predicted membrane-binding amphipathic α-helix of AAV2 MAAP

aa 96-116 of AAV2 MAAP are is depicted in helical wheel representation (top) and linear sequence (bottom).

The N-terminus of MAAP is ill-defined in bovine AAV and AAV5 (see above), since we do not know the exact start codon. If we assumed that in these species, MAAP starts at the first potential start codon, bovine AAV and AAV5 would have an N-terminal extension of 23 aas compared to other species (Fig 3). Amphipaseek and Heliquest predict that this region could form a membrane-binding, amphipathic α-helix in AAV5, but not in bovine AAV (Table 1). Of course, its biological relevance is uncertain, since this region might in fact not be translated, depending on the actual start codon of MAAP.

Membrane-binding proteins are sometimes palmitoylated (i.e. a fatty acid chain is added) on cysteine residues, which may affect the membrane binding or conformation of the protein [32]. Since MAAP contains a cysteine in strand β1 (Fig 3), we applied a palmitoylation predictor, MDD-Palm [33]. It consistently predicted no palmitoylation sites in MAAP.

MAAP originated *de novo* by overprinting the PLA2 domain

Overlapping genes, such as VP1/MAAP, are thought to originate by "overprinting", a process in which substitutions in an ancestral reading frame allow the expression of a second reading frame (called "novel"), while preserving the expression of the first frame [9,11]. The ancestral frame can be identified by its phylogenetic distribution (the ORF with the widest distribution is most probably the ancestral one) [11].

VP1 is necessarily the ancestral reading frame, since a PLA2 domain is found not only in most *Parvovirinae* [5], but also in a wide variety of animals and plants [34], whereas MAAP is found only in 3 *dependoparvovirus* species. Therefore, MAAP must have originated by overprinting the region encoding the PLA2 domain of the VP1 frame, in the common ancestor of dependoparvoviruses A, B, and porcine AAVs. This evolutionary scenario, *de novo* origination, as opposed to homologous descent from a pre-existing gene [12], matches the observation that "sequence searches on MAAP identified no homolog" [8].

MAAP has a highly biased sequence composition

Composition Profiler [35] found that MAAP has a highly biased aa composition compared to proteins present in the database Uniprot [36]. In particular, MAAP is significantly (P<0.005) depleted in hydrophobic aas. In addition, MAAP is enriched in the positively charged aa arginine (R) and depleted in the negatively charged aas aspartate (D) and glutamate (E). Consequently, it has a very high positive net charge, and a strikingly high isoelectric point (pI) of 12. MAAP is also highly enriched (P<0.005) in serine (S) and threonine (T), particularly in its central region (Fig 3). Finally, MAAP is highly depleted in Tyrosine (Y) and Methionine (M) (see below).

This compositional bias is similar to the one reported, on average, for proteins originated *de novo* by overprinting (see Figure 6C in [11]), in agreement with our finding above that MAAP originated by overprinting. In the same study that reported the compositional bias of proteins originated by overprinting, about 60% of such proteins or protein regions were disordered [11], again like MAAP.

Given its highly biased composition and predicted structural disorder, there is a possibility that MAAP will migrate above its predicted size in SDS-PAGE electrophoresis [37]; this is not the case in AAV2 [8], but is a point to keep in mind for experimental detection of MAAP in other species.

MAAP lacks tyrosines and internal methionines in most strains

Strikingly, two aas are completely absent from MAAP in most species: tyrosine (Y) and methionine (M). This depletion is highly significant ($P<10^{-6}$).

Tyrosine is absent from MAAP in all strains of *dependoparvovirus B* (S2B alignment), and is found in only 10 *dependoparvovirus A* strains out of 116 (S2A Alignment). Finally, MAAP contains a single tyrosine in all porcine AAV strains (S2C Alignment), in strand β1 (Fig 3).

As we saw above, we do not know whether the first aa of MAAP is a methionine, since MAAP is translated from a non-canonical start codon. However, MAAP lacks *internal* methionines in all dependoparvoviruses B and porcine AAVs, and contains an internal methionine in only 4 dependoparvoviruses A out of 116 (S2 Alignment).

We propose potential explanations for the absence of these aas below and in the Discussion.

The region of the VP1 gene located between the start codons of MAAP and of AAP is devoid of ATG codons

We found above that MAAP contains almost no methionine. This absence might stem from a selection against ATG start codons (which encode methionine), rather than from selection against the methionine amino acid *per se*. Indeed, an ATG codon within MAAP might not only prevent normal initiation of MAAP at its CTG start codon (since CTG is weaker than ATG), but also prevent initiation of AAP and VP2, which are also translated from weak codons (Fig 1), respectively CTG [7] and ACG [38].

To test this hypothesis, we examined whether ATG codons are absent from the region of the VP1 gene located between the MAAP and the AAP start codons (respectively CTG 2282 and CTG 2729, Fig 1). We found that indeed, this region is completely devoid of ATG codons in 97% of dependoparvoviruses A

(160 sequences out of 165), and in all dependoparvoviruses B and porcine AAVs (S3A, S3B and S3C Alignments, respectively).

If the absence of methionine in MAAP resulted only from a selection operating at the protein level, such selection would not result in the absence of ATG codons in all frames of the VP1 gene, contrary to what we observe. Therefore, our findings support the hypothesis of a selection operating against ATG codons (rather than methionine aas *per se*), which could prevent the translation of MAAP, AAP and VP2 from weak, non-canonical start codons.

A consequence of the lack of ATG codons in this region of the VP1 gene is that there is no methionine in the corresponding region of the VP1 protein, roughly corresponding to VP1u (the VP1-unique region) and to the first half of VP2c (the region common to VP1 and VP2) (Fig 5).

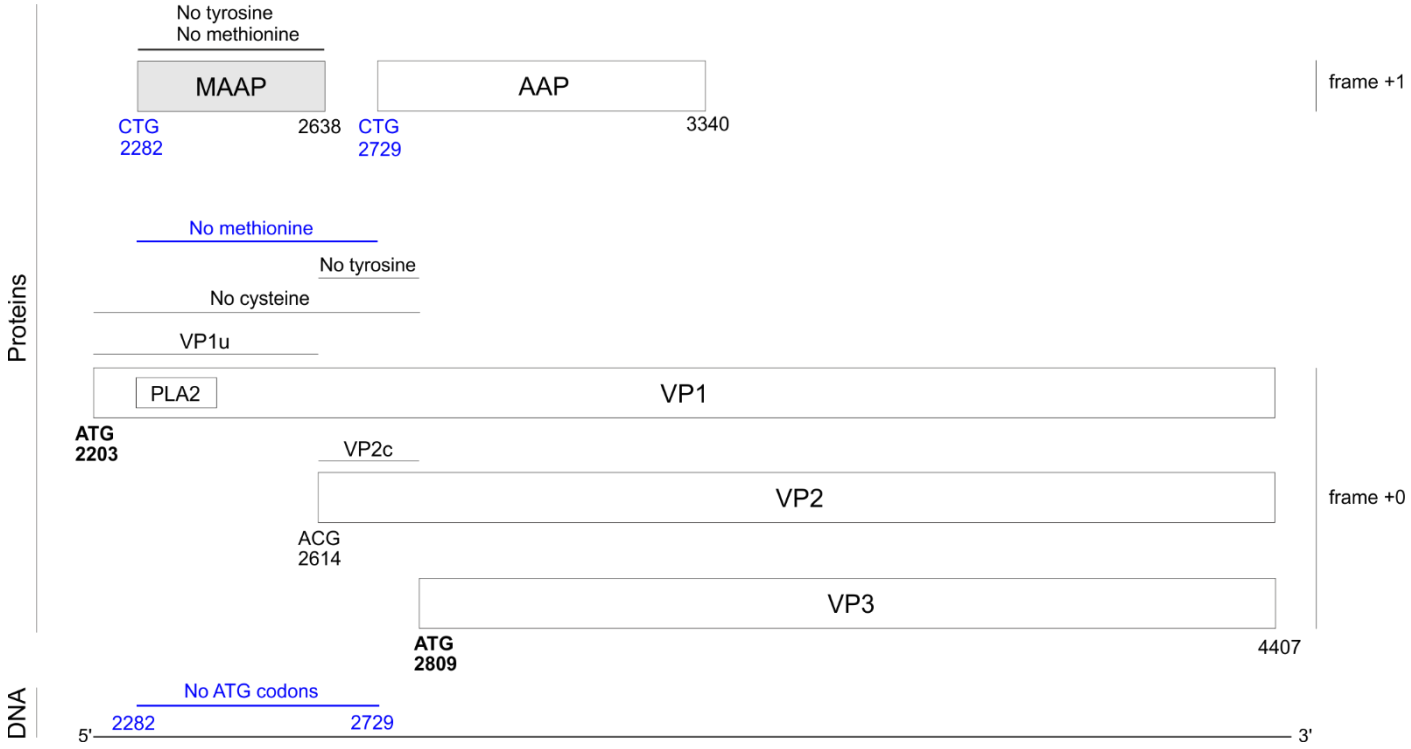

Fig 5: Regions of the VP1 gene that lack certain aas or codons, in dependoparvoviruses A and B and in some porcine AAVs

Conventions are the same as in Fig 1. Numbering corresponds to the genomic coordinates of AAV2.

In contrast, VP3 contains at least 5 ATG codons in all dependoparvoviruses A, B and porcine AAVs (not shown), which are not expected to influence the translation of the upstream coding sequences (VP1, VP2, MAAP and AAP).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0234.v1

The N-terminus of VP1 lacks Cysteines in all dependoparvoviruses

While examining the sequence composition of VP1u and VP2c (S4 and S5 Alignments, respectively), we noticed that they lack cysteine, not only in dependoparvoviruses that encode MAAP, but in all other dependoparvoviruses, except 3 sequences out of 134).

This absence of cysteine is highly significant ($P<10^{-6}$). It is restricted to VP1u and VP2c, since VP3 contains at least 3 cysteines in all dependoparvoviruses. Cysteine is also absent in the VP1u region of most other *Parvovirinae* (not shown). We propose potential explanations to this observation in the Discussion.

Note that a large region of the capsid (aa 2-202 of AAV2 VP1) is devoid of sulfur atoms, as a consequence of lacking methionines and cysteines. Perhaps this finding could be exploited for research or therapeutic purposes.

Finally, VP2c lacks tyrosine in all dependoparvoviruses A and B, and in some porcine AAVs (S5 alignment and Fig 5).

Discussion

MAAP was initially reported in dependoparvoviruses A and AAV5 [8]. We found that it is also encoded in bovine AAV, and in porcine AAVs, which have shown promise for gene transfer into muscle tissues [22,39] or into the retina [23].

We are confident about our prediction that MAAP binds membrane through a C-terminal, amphipathic α-helix (Fig 3 and 4), for 3 reasons: 1) we used complementary software (Amphipaseek [29] and Heliquest [30]); 2) this prediction is conserved in all species (see Fig 3 and Table 1); and 3) the prediction is strong, with a Discriminating factor well above the cutoff (Table 1). By comparison, weaker predictions that we made of a membrane-binding amphipathic α-helix in alphavirus nsp1 [40], in which only points 1) and 2) were applicable, have since been validated experimentally [41,42].

The absence of ATG codons in the N-terminus of VP1 is probably due to regulatory reasons

We discovered that the region of the VP1 gene located between the start codons of MAAP and AAP contains no ATG codon (Fig 5). Ogden *et al* found that substituting most aas in this region by methionine (encoded by ATG) reduced capsid production. They hypothesized that this effect was due to a reduced translation of VP2 and perhaps of VP3 caused by the introduction of ATG codons [8]. Our findings support their hypothesis and go further. We found that there are no ATG codons in the two other reading frames of this region of VP1 either. This observation is compatible with the hypothesis that ATG codons would affect the translation, not only of VP2 and VP3, but also of MAAP and AAP, which use weak, non-canonical start codons.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0234.v1

The absence of cysteines in VP1u and VP2u suggests that these regions are exposed to oxidizing conditions during post-endocytic trafficking

We noticed that all dependoparvoviruses lack cysteine in the region of the capsid located upstream of VP3 (ie VP1u and VP2c, see Fig 5). This observation suggests a strong selection against the presence of cysteines, presumably to avoid the formation of disulfide bridges between capsid subunits. Upon virus entry in cells and endocytosis, VP1u and VP2c are located within the capsid, but are externalized during the next step [43] (i.e. post-endocytic trafficking, prior to release into the cytoplasm [1,44]). Our findings suggest that at some point during this step, VP1u and VP2c are exposed to oxidizing conditions that could create disulfide bridges, which would somehow prevent a normal function.

In a recent study, substituting each aa of VP1u and VP2c to cysteine did not markedly decrease capsid assembly [8]. This observation might at first seem incompatible with the existence of a strong selection against the presence of cysteine. However, this study only assayed the steps occurring after viral gene expression, and thus could not be expected to detect selection occurring in prior steps (including post-endocytosis trafficking).

Other researchers had noticed the absence of cysteine in the N-terminus of VP1 in individual *Parvovirinae* species (e.g. [45]), and their presence in VP3 [46], but we are not sure whether a comparative sequence analysis has ever been published.

Why does MAAP lack Tyrosine?

MAAP lacks tyrosine in the vast majority of dependoparvoviruses A and B, and contains a single tyrosine in porcine AAVs. We see 2 non-exclusive hypotheses to explain this absence: 1) the unusual origin of MAAP, born by overprinting the VP1 reading frame; 2) a selection pressure against tyrosine.

Hypothesis 1) probably contributes to the absence of tyrosine in MAAP, but is unlikely to fully explain it. Although tyrosine is the most depleted aa in proteins originated by overprinting (60% on average) [11], its depletion of MAAP is a lot more pronounced than expected. Since tyrosine has an average abundance of 3% in Uniprot proteins [35], MAAP (119aas) would be expected to contain on average 1.43 tyrosine (=119*0.03*(1-0.60)), instead of 0.08 in *dependoparvovirus A*. By comparison, the other protein originated by overprinting the capsid gene, AAP (Fig 5), is also highly depleted ($P<10^{-6}$) in tyrosine, yet contains 0.80 tyrosine on average in *dependoparvovirus A* (unpublished observations), i.e. MAAP is ten times more depleted in tyrosine than AAP.

In principle, the absence of tyrosine could also result from negative selection (hypothesis 2 above), if tyrosine is deleterious to the function or structure of MAAP. For example, tyrosine phosphorylation of MAAP (but not of other viral proteins) might somehow be recognized by antiviral defenses (a speculative scenario). Or tyrosine might be detrimental because MAAP is mostly disordered (Fig 3), and tyrosine is disfavored in disordered proteins [35]. Interestingly, VP2c, which is fully disordered [47], also lacks tyrosines in all dependoparvoviruses A and B (Fig 5 and S5 alignment).

In summary, the absence of Tyrosine from MAAP probably results at least in part from its origin by overprinting and probably also from another reason, such as negative selection. Testing hypothesis 1) probably requires evolutionary simulations, while hypothesis 2) could be tested by introducing tyrosines in

MAAP without affecting the aa sequence of VP1. We will gladly pay a drink to the first 5 researchers who contact us with a convincing explanation (including another hypothesis persuasively substantiated by observations).

MAAP probably originated independently from the X protein encoded in related genera (*erythroparvovirus* and *tetraparvovirus*)

The genus *erythroparvovirus*, which is related to *dependoparvovirus* [48], encodes an "X ORF" that overlaps the same part of VP1 as MAAP, i.e. the region encoding the PLA2 (PhosphoLipase A2) domain. We recently showed that this X ORF is homologous to the ARF1 ORF encoded in the genus *tetraparvovirus*, closely related to *erythroparvovirus*, and that both X and ARF1 must express functional proteins (submitted). Given that MAAP and X/ARF1 are encoded by similar regions of VP1, they could in principle be homologous, i.e. have a common origin. Yet we think this is unlikely, for two reasons:

1) MAAP and the X protein have extremely different predicted sequence features: the region of MAAP that overlaps PLA2 is disordered and T/S-rich (Fig 3), while in X it contains a transmembrane region;

2) MAAP is found only in 3 dependoparvoviruses, which are not basal to the dependoparvovirus phylogeny [49], making it unlikely that MAAP originated in a common ancestor of dependo-, erythro-, and tetraparvoviruses.

Conclusion

Although this was not our original goal, we discovered that MAAP and certain regions of VP1 completely lack several aas and one codon (Fig 5). This absence has obvious implications for the design of capsid genes of recombinant therapeutic AAVs, but also for fundamental studies of the viral cycle. For example, the presence of aas that are normally "forbidden" suggests that the corresponding strain might have unusual properties. It will be interesting to investigating such strains (6 in total): AAVhu.17 (AY530582.1) [50] contains a cysteine in VP1u (S4 Alignment); CHC2731_AAV.FL.linear (MK139293.1) contains both a cysteine in VP2c (S5 Alignment) and an ATG in the region between the MAAP and AAP start (S3 Alignment). 4 strains contain a methionine in MAAP (S2 Alignment): CHC2320_AAV.FL [51], AAVcy.2 [52], AAV6R2 [53] and AAVhu.67 [50] (respective accession numbers MK139290.1, AY243020.1, EU368911.1, AY530627.1).

The presence of "forbidden" aas might also suggest that the corresponding strain contains a sequencing error. For instance, Duck parvovirus GXN45 (accession MH717783) appears to contain a cysteine in VP2c, which in fact results from a frameshift sequencing error (see S5 Alignment).

Finally, our findings highlight the need for systematic screens of the effect of substitutions in the capsid gene, like the pioneering one recently proposed [8], but which could detect substitutions that are deleterious at *any step* of the whole viral cycle. Indeed, the negative selection against cysteine in VP1u and VP2c, and against tyrosine in VP2c, were not detected in the conditions tested by this screen [8], which assayed genome packaging and capsid assembly.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0234.v1

Materials and Methods

Nucleotide sequence alignment and analysis

We collected the coding sequences of VP1 genes in Genbank [54] (30th July 2019). To generate codon-respecting alignments based on the coding sequence of VP1, we used the program TranslatorX [55] with the "Muscle" option.

Analysis of Kozak consensus sequences of potential ATG start codons

Kozak sequences surrounding an ATG start codon can direct translation from this ATG with varying degrees of strength [25]. The most important factor is the presence of a purine (A or G) 3 nucleotides upstream of the ATG start codon, and of a G (or less favourably an U) immediately after the ATG. For the ORFs considered here, we classified Kozak sequences of potential ATG start codons in 4 categories, as in a recent exhaustive analysis in vertebrates [25]:

1) "optimal" Kozak sequences match the consensus (A/G)CCATGG;
2) "strong" ones match the consensus (A/G)NNATGG, where N is any nucleotide;
3) "moderate" ones match the consensus (A/G)NNATG(A/C/U) or (C/U)NNATGG;
4) "weak" Kozak sequences do not match any of these consensus sequences[25].

Protein sequence alignment and domain identification

All protein multiple sequence alignments are presented using Jalview [56], with the ClustalX coloring scheme [57]. We carried out phylogenetic analyses with phylogeny.fr [58], using default options. S1 Alignment contains the sequence alignment of MAAP proteins from representative strains. We used HHpred [59] to identify domains of VP1.

Prediction of protein structural features

We predicted disordered regions using MetaDisorder [60], in agreement with the principles described in [61]. To predict potential protein-binding regions, we used MoRFchibi_Web [62] and ANCHOR2 [63], called from the IUPred2A web server [63]. We predicted coiled-coil regions using DeepCoil [64]. To detect protein regions of low or medium sequence complexity, we used SEG [65], called from the ANNIE web server [66], set on parameters 45/3.75/3.4.

To predict membrane-binding, amphipathic α-helices, we used Amphipaseek [29] (parameters: high specificity/low sensitivity) and refined its predictions by using Heliquest [30] as follows. For each helix that Amphipaseek predicted, we analyzed the region surrounding it using the “analysis” function of Heliquest. Heliquest makes use of a Discriminating factor (D) to predict membrane-binding helices: $D=0.944*\langle\mu H\rangle +0.33*z$, in which $\langle\mu H\rangle$ is the hydrophobic moment [67] and z is the net charge of the region considered. The helix is predicted as “potential” membrane-binding amphipathic α-helix if $0.68 < D < 1.34$, and as "reliable" if $D \geq 1.34$ [30]. We also used Heliquest to plot helical wheel representations (reviewed in [31]).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 17 February 2020 doi:10.20944/preprints202002.0234.v1

We predicted protein post-translational modifications using Modpred [68], and palmitoylation sites using MDD-Palm [33].

Supporting information

S1 Alignment. Sequence alignment of MAAP proteins from representative strains, in FASTA format
The sequence features of this alignment are presented in detail in Fig 3.

S2 Alignment. MAAP contains no internal methionines and no or few tyrosines

S3 Alignment. The VP1 gene lacks ATG in the region between the start codons of MAAP and AAP

S4 Alignment. The VP1-unique region (VP1u) lacks cysteine in all dependoparvoviruses

S5 Alignment. The VP2c region (common to VP1 and VP2) lacks cysteine in all dependoparvoviruses, and lacks tyrosine in dependoparvoviruses A, B, and in some porcine AAVs

Acknowledgements

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. Wang D, Tai PWL, Gao G. Adeno-associated virus vector as a platform for gene therapy delivery. Nat Rev Drug Discov. 2019;18: 358–378. doi:10.1038/s41573-019-0012-9

2. 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

3. 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

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

5. 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.

6. Kurian JJ, Lakshmanan R, Chmely WM, Hull JA, Yu JC, Bennett A, et al. Adeno-Associated Virus VP1u Exhibits Protease Activity. Viruses. 2019;11: 399. doi:10.3390/v11050399

7. Sonntag F, Schmidt K, Kleinschmidt JA. A viral assembly factor promotes AAV2 capsid formation in the nucleolus. Proc Natl Acad Sci USA. 2010;107: 10220–10225. doi:10.1073/pnas.1001673107

8. Ogden PJ, Kelsic ED, Sinai S, Church GM. Comprehensive AAV capsid fitness landscape reveals a viral gene and enables machine-guided design. Science. 2019;366: 1139–1143. doi:10.1126/science.aaw2900

9. 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

10. 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

11. 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

12. Andersson DI, Jerlström-Hultqvist J, Näsvall J. Evolution of new functions de novo and from preexisting genes. Cold Spring Harb Perspect Biol. 2015;7. doi:10.1101/cshperspect.a017996

13. Pavesi A. Asymmetric evolution in viral overlapping genes is a source of selective protein adaptation. Virology. 2019;532: 39–47. doi:10.1016/j.virol.2019.03.017

14. McFadden N, Bailey D, Carrara G, Benson A, Chaudhry Y, Shortland A, et al. Norovirus regulation of the innate immune response and apoptosis occurs via the product of the alternative open reading frame 4. PLoS Pathog. 2011;7: e1002413. doi:10.1371/journal.ppat.1002413

15. van Knippenberg I, Carlton-Smith C, Elliott RM. The N-terminus of Bunyamwera orthobunyavirus NSs protein is essential for interferon antagonism. J Gen Virol. 2010;91: 2002–2006. doi:10.1099/vir.0.021774-0

16. Noteborn MH, Todd D, Verschueren CA, de Gauw HW, Curran WL, Veldkamp S, et al. A single chicken anemia virus protein induces apoptosis. J Virol. 1994;68: 346–351.

17. Chen W, Calvo PA, Malide D, Gibbs J, Schubert U, Bacik I, et al. A novel influenza A virus mitochondrial protein that induces cell death. Nat Med. 2001;7: 1306–1312. doi:10.1038/nm1201-1306

18. Vargason JM, Szittya G, Burgyán J, Hall TMT. Size selective recognition of siRNA by an RNA silencing suppressor. Cell. 2003;115: 799–811. doi:10.1016/s0092-8674(03)00984-x

19. Lingel A, Simon B, Izaurralde E, Sattler M. The structure of the flock house virus B2 protein, a viral suppressor of RNA interference, shows a novel mode of double-stranded RNA recognition. EMBO Rep. 2005;6: 1149–1155. doi:10.1038/sj.embor.7400583

20. Meier C, Aricescu AR, Assenberg R, Aplin RT, Gilbert RJC, Grimes JM, et al. The crystal structure of ORF-9b, a lipid binding protein from the SARS coronavirus. Structure. 2006;14: 1157–1165. doi:10.1016/j.str.2006.05.012

21. Baulcombe DC, Molnár A. Crystal structure of p19--a universal suppressor of RNA silencing. Trends Biochem Sci. 2004;29: 279–281. doi:10.1016/j.tibs.2004.04.007

22. Bello A, Tran K, Chand A, Doria M, Allocca M, Hildinger M, et al. Isolation and evaluation of novel adeno-associated virus sequences from porcine tissues. Gene Ther. 2009;16: 1320–1328. doi:10.1038/gt.2009.82

23. Puppo A, Bello A, Manfredi A, Cesi G, Marrocco E, Corte MD, et al. Recombinant Vectors Based on Porcine Adeno-Associated Viral Serotypes Transduce the Murine and Pig Retina. Qiu J, editor. PLoS ONE. 2013;8: e59025. doi:10.1371/journal.pone.0059025

24. Kearse MG, Wilusz JE. Non-AUG translation: a new start for protein synthesis in eukaryotes. Genes Dev. 2017;31: 1717–1731. doi:10.1101/gad.305250.117

25. 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

26. Diaz de Arce AJ, Noderer WL, Wang CL. Complete motif analysis of sequence requirements for translation initiation at non-AUG start codons. Nucleic Acids Research. 2018;46: 985–994. doi:10.1093/nar/gkx1114

27. Atkins JF, Loughran G, Bhatt PR, Firth AE, Baranov PV. Ribosomal frameshifting and transcriptional slippage: From genetic steganography and cryptography to adventitious use. Nucleic Acids Res. 2016;44: 7007–7078. doi:10.1093/nar/gkw530

28. Giménez-Andrés M, Čopič A, Antonny B. The Many Faces of Amphipathic Helices. Biomolecules. 2018;8: 45. doi:10.3390/biom8030045

29. Sapay N, Guermeur Y, Deléage G. Prediction of amphipathic in-plane membrane anchors in monotopic proteins using a SVM classifier. BMC Bioinformatics. 2006;7: 255. doi:10.1186/1471-2105-7-255

30. Gautier R, Douguet D, Antonny B, Drin G. HELIQUEST: a web server to screen sequences with specific alpha-helical properties. Bioinformatics. 2008;24: 2101–2102. doi:10.1093/bioinformatics/btn392

31. Mól AR, Castro MS, Fontes W. NetWheels: A web application to create high quality peptide helical wheel and net projections. Bioinformatics; 2018 Sep. doi:10.1101/416347

32. Tom CTMB, Martin BR. Fat Chance! Getting a Grip on a Slippery Modification. ACS Chem Biol. 2013;8: 46–57. doi:10.1021/cb300607e

33. Weng S-L, Kao H-J, Huang C-H, Lee T-Y. MDD-Palm: Identification of protein S-palmitoylation sites with substrate motifs based on maximal dependence decomposition. Xue Y, editor. PLoS ONE. 2017;12: e0179529. doi:10.1371/journal.pone.0179529

34. 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

35. Vacic V, Uversky VN, Dunker AK, Lonardi S. Composition Profiler: a tool for discovery and visualization of amino acid composition differences. BMC Bioinformatics. 2007;8: 211. doi:10.1186/1471-2105-8-211

36. UniProt Consortium T. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2018;46: 2699. doi:10.1093/nar/gky092

37. Iakoucheva LM, Kimzey AL, Masselon CD, Smith RD, Dunker AK, Ackerman EJ. Aberrant mobility phenomena of the DNA repair protein XPA. Protein Science. 2009;10: 1353–1362. doi:10.1110/ps.ps.40101

38. Becerra SP, Rose JA, Hardy M, Baroudy BM, Anderson CW. Direct mapping of adeno-associated virus capsid proteins B and C: a possible ACG initiation codon. Proc Natl Acad Sci USA. 1985;82: 7919–7923. doi:10.1073/pnas.82.23.7919

39. Tulalamba W, Weinmann J, Pham QH, El Andari J, VandenDriessche T, Chuah MK, et al. Distinct transduction of muscle tissue in mice after systemic delivery of AAVpo1 vectors. Gene Ther. 2019 [cited 10 Jan 2020]. doi:10.1038/s41434-019-0106-3

40. Ahola T, Karlin DG. Sequence analysis reveals a conserved extension in the capping enzyme of the alphavirus supergroup, and a homologous domain in nodaviruses. Biol Direct. 2015;10: 16. doi:10.1186/s13062-015-0050-0

41. Moriceau L, Jomat L, Bressanelli S, Alcaide-Loridan C, Jupin I. Identification and Molecular Characterization of the Chloroplast Targeting Domain of Turnip yellow mosaic virus Replication Proteins. Front Plant Sci. 2017;8: 2138. doi:10.3389/fpls.2017.02138

42. Nishikiori M, Ahlquist P. Organelle luminal dependence of (+)strand RNA virus replication reveals a hidden druggable target. Sci Adv. 2018;4: eaap8258. doi:10.1126/sciadv.aap8258

43. Sonntag F, Bleker S, Leuchs B, Fischer R, Kleinschmidt JA. Adeno-associated virus type 2 capsids with externalized VP1/VP2 trafficking domains are generated prior to passage through the cytoplasm and are maintained until uncoating occurs in the nucleus. J Virol. 2006;80: 11040–11054. doi:10.1128/JVI.01056-06

44. Nonnenmacher M, Weber T. Intracellular transport of recombinant adeno-associated virus vectors. Gene Ther. 2012;19: 649–658. doi:10.1038/gt.2012.6

45. Leisi R, Von Nordheim M, Ros C, Kempf C. The VP1u Receptor Restricts Parvovirus B19 Uptake to Permissive Erythroid Cells. Viruses. 2016;8: 265. doi:10.3390/v8100265

46. Pulicherla N, Kota P, Dokholyan NV, Asokan A. Intra- and Inter-Subunit Disulfide Bond Formation Is Nonessential in Adeno-Associated Viral Capsids. Qiu J, editor. PLoS ONE. 2012;7: e32163. doi:10.1371/journal.pone.0032163

47. Venkatakrishnan B, Yarbrough J, Domsic J, Bennett A, Bothner B, Kozyreva OG, et al. Structure and Dynamics of Adeno-Associated Virus Serotype 1 VP1-Unique N-Terminal Domain and Its Role in Capsid Trafficking. Journal of Virology. 2013;87: 4974–4984. doi:10.1128/JVI.02524-12

48. Cotmore SF, Agbandje-McKenna M, Chiorini JA, Mukha DV, Pintel DJ, Qiu J, et al. The family Parvoviridae. Arch Virol. 2014;159: 1239–1247. doi:10.1007/s00705-013-1914-1

49. Lau SKP, Ahmed SS, Tsoi H-W, Yeung HC, Li KSM, Fan RYY, et al. Bats host diverse parvoviruses as possible origin of mammalian dependoparvoviruses and source for bat-swine interspecies transmission. J Gen Virol. 2017;98: 3046–3059. doi:10.1099/jgv.0.000969

50. Gao G, Vandenberghe LH, Alvira MR, Lu Y, Calcedo R, Zhou X, et al. Clades of Adeno-associated viruses are widely disseminated in human tissues. J Virol. 2004;78: 6381–6388. doi:10.1128/JVI.78.12.6381-6388.2004

51. La Bella T, Imbeaud S, Peneau C, Mami I, Datta S, Bayard Q, et al. Adeno-associated virus in the liver: natural history and consequences in tumour development. Gut. 2019. doi:10.1136/gutjnl-2019-318281

52. Gao G, Alvira MR, Somanathan S, Lu Y, Vandenberghe LH, Rux JJ, et al. Adeno-associated viruses undergo substantial evolution in primates during natural infections. Proc Natl Acad Sci USA. 2003;100: 6081–6086. doi:10.1073/pnas.0937739100

53. Vandenberghe LH, Breous E, Nam H-J, Gao G, Xiao R, Sandhu A, et al. Naturally occurring singleton residues in AAV capsid impact vector performance and illustrate structural constraints. Gene Ther. 2009;16: 1416–1428. doi:10.1038/gt.2009.101

54. Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I. GenBank. Nucleic Acids Res. 2020;48: D84–D86. doi:10.1093/nar/gkz956

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

56. 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

57. 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

58. 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

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

60. 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

61. 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

62. Malhis N, Jacobson M, Gsponer J. MoRFchibi SYSTEM: software tools for the identification of MoRFs in protein sequences. Nucleic Acids Res. 2016;44: W488–W493. doi:10.1093/nar/gkw409

63. Mészáros B, Erdős G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Research. 2018;46: W329–W337. doi:10.1093/nar/gky384

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

66. 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

67. Eisenberg D, Weiss RM, Terwilliger TC. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. Nature. 1982;299: 371–374. doi:10.1038/299371a0

68. Pejaver V, Hsu W-L, Xin F, Dunker AK, Uversky VN, Radivojac P. The structural and functional signatures of proteins that undergo multiple events of post-translational modification: Structural and Functional Signatures of PTM Crosstalk. Protein Science. 2014;23: 1077–1093. doi:10.1002/pro.2494