Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Gene- and capsid-region-resolved codon usage and CpG/UpA ordering in canine astrovirus

Jun-Wei Yang

or0706555@gmail.com

Independent Researcher

Cheng-Hsun Lee

Kuan Mei Clinic

Hsuan-Wei Huang

Kuan Mei Clinic

Wei-Mei Hsieh

Xiang Le Zhu Yi

Su-Lan Hsu

Xiang Le Zhu Yi

Research Article

Keywords: canine astrovirus, codon usage, CpG, UpA, synonymous evolution, host correspondence

Posted Date: August 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10561846/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Canine astrovirus contains three principal coding regions whose synonymous RNA evolution may differ by gene and within the capsid. We analysed all ORF1a, ORF1b and ORF2 sequences from 31 exact-deduplicated complete genomes sampled from domestic dogs. ORF2 showed the greatest amino-acid divergence and a distinct codon-usage profile. All three genes had fewer codon-boundary CpG and UpA motifs than expected under 9,999 randomizations preserving the protein and exact synonymous-codon inventory, with the strongest gene-wide depletion in ORF1b. Within ORF2, however, the rapidly evolving central capsid region was more depleted than its conserved N-terminal region. Identity-representative and non-overlap analyses retained the main directions. ORF2 also showed the greatest relative correspondence to pooled dog coding preferences across three descriptive indices. These patterns reveal hierarchical protein- and RNA-level evolution and suggest possible host-associated constraints that require experimental validation.

Introduction

Astroviruses are small, non-enveloped, positive-sense RNA viruses infecting a broad range of vertebrates. Their compact genomes usually encode a non-structural polyprotein in ORF1a, an RNA-dependent RNA polymerase in ORF1b and a capsid precursor in ORF2; ORF1b is accessed through a ribosomal frameshift and ORF2 partially overlaps the replicase region (Arias and DuBois 2017). Canine astrovirus (CaAstV; taxon species *Mamastrovirus canis*, formerly Mamastrovirus 5) has been detected in domestic dogs with and without enteric disease and occurs in genetically heterogeneous lineages distributed across several countries (Martella et al. 2009; Caddy and Goodfellow 2015; Mihalov-Kovács et al. 2017; Zhang et al. 2020; Qiu et al. 2026). This diversity makes CaAstV useful for asking whether different coding functions experience distinct layers of protein and synonymous RNA evolution.

Codon usage in RNA viruses reflects mutational composition, protein constraints, genome architecture and possible selection on translation or RNA-level functions (Belalov and Lukashev 2013; Di Giallonardo et al. 2017; Gaunt and Digard 2022). Conventional measures such as relative synonymous codon usage (RSCU), effective number of codons (ENc), codon adaptation index (CAI) and relative codon deoptimization index (RCDI) summarize important global properties, but they do not distinguish which synonymous codons are present from how those codons are ordered. That distinction is relevant to dinucleotides crossing codon junctions. CpG and UpA are often underrepresented in vertebrate RNA viruses, and experimental recoding has shown that increasing their frequencies can alter replication in some virus–cell systems (Atkinson et al. 2014; Tulloch et al. 2014; Takata et al. 2017). Nevertheless, depletion in a natural viral sequence cannot by itself establish a specific immune, translational or fitness mechanism.

An exact-codon conditional null can isolate one component of this problem. If the amino acid at every position and the total count of every synonymous codon in a coding sequence (CDS) are held fixed, rearranging codons only among positions encoding the same amino acid changes codon order without

changing the protein, nucleotide composition or codon inventory. A departure of observed codon-boundary CpG or UpA from this null therefore identifies positional organization beyond simple avoidance of particular codons. This framework is especially informative for CaAstV because its genes have different roles and its ORF2 capsid contains a conserved N-terminal region and a more variable central region (Caddy and Goodfellow 2015). Human astrovirus structures similarly distinguish a conserved shell from a surface-exposed projection, although human coordinates should not be imposed directly on canine sequences (Dong et al. 2011; Toh et al. 2016).

Here, we integrated all three principal CaAstV genes with a predefined within-ORF2 comparison. We asked whether ORF1a, ORF1b and ORF2 differ in protein divergence, synonymous codon use, dog-relative codon correspondence and codon-order-dependent CpG/UpA representation; whether the gene-wide pattern is stable after sequence identity reduction; and whether local capsid evolution reveals constraints that whole-gene averages may obscure. The design treats the 31 genomes as paired observations across genes and uses the host-relative indices only as hypothesis-generating measures of correspondence, not as proof of adaptation.

Materials and methods

Sequence audit and coding-sequence definition

NCBI Nucleotide records available in July 2026 were audited for complete CaAstV genomes whose host field explicitly reported dog or *Canis lupus familiaris*. Vaccine strains, artificial constructs, non-canine or ambiguous hosts, incomplete genomes, records containing ambiguous nucleotides and records lacking reliable principal CDS annotations were excluded. Exact whole-genome duplicates were removed by nucleotide-sequence hash. The retained dataset comprised 31 genomes from seven countries, spanning records collected between 2012 and 2023 where dates were available. Every genome contained one complete ORF1a, ORF1b and ORF2, giving 93 principal CDS. Small putative accessory ORFs were not analysed because annotation was not sufficiently consistent across the complete set. Accession-level metadata and inclusion decisions are provided in Online Resource 1.

ORF2 regional analyses used KX599350.1 (834 amino acids) as the coordinate reference and canine-specific boundaries reported by Caddy and Goodfellow (2015): N-terminal amino acids 1–446, central 447–730 and C-terminal 731–834. The primary regional statistical comparison was central versus N-terminal. A second N-terminal comparator, amino acids 70–446, removed the portion overlapping the preceding coding region. A 50-amino-acid sliding profile was retained as a descriptive localization analysis and was not used to choose the tested region boundary.

Protein divergence and synonymous substitution context

Translated ORF1a, ORF1b and ORF2 sequences were separately aligned with FAMSA (Deorowicz et al. 2016). Pairwise amino-acid p-distance and alignment-column Shannon entropy were summarized for

each gene. Codons were projected through the corresponding protein alignment, and pairwise nonsynonymous and synonymous distances were estimated with the Nei–Gojobori method (Nei and Gojobori 1986). Median dN/dS was used only as descriptive gene-level context because recombination, substitution saturation and short local regions can make pairwise ratios unstable. The same protein-alignment and codon-projection procedure was applied to the predefined ORF2 regions.

Codon usage and domestic-dog reference

For every CDS we calculated GC content, GC at synonymous third positions (GC3), ENc (Wright 1990) and RSCU. Gene separation in standardized RSCU space was summarized by principal component analysis. The fraction of standardized RSCU variation attributable to gene identity was tested with 9,999 within-genome restricted permutations, so each genome retained one label for each of its three genes.

A domestic-dog codon reference was built from NCBI RefSeq *Canis lupus familiaris* Annotation Release 106, assembly ROS_Cfam_1.0 (GCF_014441545.1). Of 66,992 input CDS records, the longest clean CDS per GeneID was retained, yielding 20,737 genes and 11,518,934 pooled sense codons. CAI (Sharp and Li 1987), RCDI (Puigbò et al. 2010) and the RSCU cosine-distance index D (Silverj and Rota-Stabelli 2020) were then calculated for each viral CDS. Higher CAI, lower RCDI and lower D indicate greater relative correspondence under their respective definitions. These metrics use a pooled coding reference rather than tissue expression or tRNA availability and were interpreted descriptively.

Exact-codon conditional randomization

For each of the 93 CDS, 9,999 deterministic randomizations fixed the amino acid at every translated position and preserved the exact count of every synonymous codon. Codons were permuted only among positions encoding the same amino acid. Consequently, protein sequence, CDS length, nucleotide composition, GC3, global codon usage and the complete synonymous-codon inventory were unchanged; only codon order varied. A master seed of 20260801 was combined with record identifiers to derive reproducible CDS-specific seeds.

All 16 dinucleotides spanning the third nucleotide of one codon and the first nucleotide of the next were counted. CpG and UpA were the prespecified primary motifs; UpA is represented computationally as TpA. The conditional effect was log2[(observed + 0.5)/(mean randomized + 0.5)], with observed/null ratios also reported for readability. Negative values indicate fewer observed boundaries than expected from the fixed codon inventory. The same null model was applied to the predefined ORF2 regions. Empirical tail probabilities used a plus-one correction (Phipson and Smyth 2010).

Statistical analysis and sensitivity checks

Gene differences were tested across matched genomes with Friedman tests followed by two-sided paired Wilcoxon signed-rank tests. The three pairwise P values within each metric were adjusted by the Benjamini–Hochberg method (Benjamini and Hochberg 1995). CpG and UpA constituted the primary dinucleotide analyses; the 16-boundary screen used a separate 16-test adjustment. Regional ORF2 contrasts were paired by genome and the two primary CpG/UpA P values were adjusted together.

To reduce dependence among closely related sequences, deterministic representatives were selected after complete-genome clustering at 99%, 98% and 95% nucleotide identity. Representatives were chosen by a fixed quality-first ordering (complete three-gene annotation, no ambiguity, then stable accession order). Gene contrasts were repeated in each representative set. The regional analysis additionally used the non-overlap comparator and leave-one-95%-cluster-out equal-cluster summaries. Analyses were conducted in Python using NumPy, pandas, SciPy and Biopython (Cock et al. 2009). GPT-5.6 assisted initial drafting of analysis scripts and text organization; Jun-Wei Yang specified the biological design, executed and inspected the analyses, and verified source records and outputs.

Results

Audited genome set and gene-scale protein evolution

The final dataset contained 31 exact-deduplicated complete genomes and 93 principal CDS. Median ORF1a, ORF1b and ORF2 lengths were 2,667, 1,533 and 2,493 nucleotides, respectively (Table 1). ORF2 was the most divergent protein: mean pairwise amino-acid p-distance was 0.125, compared with 0.053 for ORF1b and 0.030 for ORF1a (Fig. 1C). Mean alignment-column entropy followed the same ordering (0.318, 0.149 and 0.083 bits). Median pairwise NG86 dN/dS was 0.113 for ORF2, 0.029 for ORF1b and 0.027 for ORF1a. These ratios were below one and are compatible with broad protein-level constraint, while the relative ranking highlights greater capsid diversification.

Gene-resolved codon usage and dog-relative correspondence

Codon composition differed reproducibly among the three genes. Median GC3 was 0.434 in ORF1a, 0.399 in ORF1b and 0.422 in ORF2 (Friedman P = 1.50 × 10−7). Median ENc values were 54.76, 54.64 and 53.44, respectively (P = 9.82 × 10−10), indicating generally weak-to-moderate codon bias with a modestly stronger bias in ORF2. RSCU principal components 1 and 2 explained 30.5% and 25.9% of variance. Gene identity explained 52.0% of standardized RSCU variation under the matched-genome design (9,999 restricted permutations, P = 0.0001; Fig. 1D).

ORF2 showed the greatest correspondence to the pooled dog coding reference across all three secondary indices (Fig. 3). Median CAI was 0.686 in ORF2, 0.679 in ORF1b and 0.669 in ORF1a (omnibus P = 4.44 × 10−11). Median RCDI was lower for ORF2 (1.282) than ORF1a (1.325) or ORF1b (1.329; P =

2.57 × 10−6), and median RSCU distance D was likewise lowest for ORF2 (0.062 versus 0.071 and 0.075; P = 7.34 × 10−10). The agreement of the three metrics suggests gene-specific host-relative codon correspondence, although it does not establish enhanced translation or adaptation in canine tissues.

CpG and UpA ordering differed among genes

Every ORF1a, ORF1b and ORF2 sequence had a negative conditional CpG effect and a negative conditional UpA effect. Thus, all three genes placed their existing synonymous codons in orders that formed fewer CpG and UpA codon junctions than expected under the fixed-inventory null (Fig. 1A, B). Median CpG observed/null ratios were 0.387 for ORF1a, 0.367 for ORF1b and 0.401 for ORF2; corresponding log2 effects were −1.370, −1.447 and −1.319. The gene effect was significant (Friedman P = 0.0137). ORF1b was more depleted than ORF2 (median paired ORF1b-minus-ORF2 difference −0.087, q = 0.00294), and ORF1b was also more depleted than ORF1a (ORF1a-minus-ORF1b difference 0.062, q = 0.0297); ORF1a and ORF2 did not differ significantly.

UpA separation was stronger (Friedman P = 1.44 × 10−6). Median observed/null ratios were 0.735 in ORF1a, 0.644 in ORF1b and 0.686 in ORF2, with log2 effects of −0.445, −0.635 and −0.544. All three paired contrasts were significant after within-metric correction: ORF1b was more depleted than ORF1a (q = 3.33 × 10−5) and ORF2 (q = 0.0233), while ORF2 was more depleted than ORF1a (q = 3.33 × 10−5). CpG and UpA gene effects also remained significant in the all-16 boundary screen (q16 = 0.0137 and 1.77 × 10−6). Several other boundaries differed among genes, indicating that CpG/UpA are prominent prespecified signals within a broader pattern of gene-specific RNA organization rather than the only changing motifs.

Identity reduction retained the principal directions

Whole-genome identity reduction retained 29, 27 and 12 representatives at 99%, 98% and 95%, respectively. The direction of stronger ORF1b CpG depletion than ORF2 was preserved at all thresholds, although its adjusted P value was 0.0959 in the smallest 95% set. Stronger ORF1b than ORF1a CpG depletion remained supported at 95% identity (q = 0.0366). UpA directions were also stable for ORF1a versus ORF1b and ORF1a versus ORF2, while the smaller ORF1b-versus-ORF2 contrast was not resolved after 98% or 95% reduction. The reduced precision is expected when 31 genomes collapse to 12 representatives, but the main rank ordering was not reversed.

A local ORF2 pattern differed from the whole-gene average

The gene-wide results identified ORF1b as the strongest whole-gene CpG/UpA-depleted CDS, but ORF2 contained a sharper local pattern (Fig. 2). ORF2 proteins ranged from 823 to 842 amino acids. Mean pairwise amino-acid p-distance was 0.254 in the predefined central region, compared with 0.068 in the

N-terminal region and 0.026 in the C-terminal region. Mean alignment-column entropy was 0.645, 0.156 and 0.071 bits, respectively. The descriptive 50-amino-acid scan peaked at reference positions 621–670 (mean p-distance 0.304), within the central region. Median pairwise NG86 dN/dS was 0.123 centrally and 0.062 in the N-terminal region.

The central region was also more strongly depleted under the exact-codon null. Median CpG observed/null was 0.259 centrally and 0.431 in the N-terminal region; the paired central-minus-N log2 difference was −0.778 (P = q = 4.99 × 10−7). Median UpA observed/null was 0.602 centrally and 0.797 N-terminally, with a paired difference of −0.365 (P = 1.77 × 10−8; q = 3.54 × 10−8). All 31 genomes had negative CpG and UpA effects in both regions, so this test compared degree rather than presence of depletion. The non-overlapping N-terminal comparator retained both contrasts (CpG difference −0.805, P = 7.10 × 10−7; UpA −0.369, P = 3.07 × 10−8). At 95% identity, 12 representatives remained and both regional effects were still supported (CpG q = 0.00684; UpA q = 0.00488).

Table 1 Gene-resolved evolutionary, compositional and conditional codon-order summaries. Values are medians across 31 genomes unless otherwise indicated.

Measure	**ORF1a**	**ORF1b**	**ORF2**
CDS length (nt)	2667	1533	2493
Mean aa p-distance	0.030	0.053	0.125
Mean entropy (bits)	0.083	0.149	0.318
Median NG86 dN/dS	0.027	0.029	0.113
GC3	0.434	0.399	0.422
ENc	54.76	54.64	53.44
CpG observed/null	0.387	0.367	0.401
CpG log2 effect	-1.370	-1.447	-1.319
UpA observed/null	0.735	0.644	0.686
UpA log2 effect	-0.445	-0.635	-0.544
Dog-relative CAI	0.669	0.679	0.686
Dog-relative RCDI	1.325	1.329	1.282
Dog-relative D	0.071	0.075	0.062

ENc, effective number of codons; CAI, codon adaptation index; RCDI, relative codon deoptimization index; D, RSCU cosine distance. Negative conditional effects indicate depletion beyond the fixed codon inventory.

Table 2 Paired gene contrasts for the prespecified CpG and UpA conditional effects.

Motif	Contrast	n	Median paired difference	P	q
CpG	ORF1a – ORF1b	31	0.062	0.0198	0.0297
CpG	ORF1a – ORF2	31	-0.054	0.152	0.152
CpG	ORF1b – ORF2	31	-0.087	0.000982	0.00294
UpA	ORF1a – ORF1b	31	0.147	1.35e-05	3.33e-05
UpA	ORF1a – ORF2	31	0.092	2.22e-05	3.33e-05
UpA	ORF1b – ORF2	31	-0.077	0.0233	0.0233

Differences are left minus right in log2 observed/null effect; q values are adjusted within motif.

Discussion

The combined analysis reveals a hierarchical organization of CaAstV evolution. At the scale of complete genes, ORF2 is the most protein-divergent and codon-distinct gene, whereas ORF1b shows the strongest overall codon-boundary CpG and UpA depletion. At a finer scale, the rapidly evolving central portion of ORF2 carries a much stronger conditional depletion than the conserved N-terminal region. Whole-gene and subgene summaries therefore answer different questions: the former identify functional-genome compartments, while the latter reveal localized RNA constraints that can be masked by averaging across a long CDS.

The exact-codon null is central to this interpretation. It does not generate arbitrary synonymous sequences and does not change the encoded protein or the total usage of any codon. It asks whether the same set of synonymous codons could form more or fewer third-to-first boundary motifs if placed differently among equivalent amino-acid positions. The negative effects in every sequence and every principal gene therefore cannot be attributed only to amino-acid composition, GC content or reduced use of codons that themselves contain CpG or UpA. Instead, the observed codons are non-randomly arranged at the tested boundary scale. This codon-order component adds an RNA organizational layer to conventional codon-bias measures.

ORF1b encodes the viral polymerase and is embedded in the ORF1a/ORF1b frameshift architecture. Its strong whole-gene depletion may be compatible with stringent RNA-level constraints linked to replication, translation or the organization of the replicase region. ORF1a was intermediate for CpG and least depleted for UpA, while ORF2 occupied an intermediate whole-gene position. These differences need not imply one shared mechanism. Local RNA structure, polymerase-associated mutation, transcript abundance, packaging and linked selection can all influence synonymous placement (Di Giallonardo et al. 2017; Gaunt and Digard 2022). Overlapping reading frames are also relevant: randomizing a focal CDS preserves its own protein but does not preserve an alternative protein encoded in another frame. The analysis accordingly identifies natural ordering relative to a focal-CDS null rather than a complete mechanistic model of the compact genome.

The ORF2 result suggests that protein diversification and synonymous RNA organization can evolve together. The central capsid region tolerates or favours substantially more amino-acid change than the N-terminal region, yet its unchanged codon inventory is arranged to produce especially few boundary CpG and UpA motifs. This co-localization could provide an evolutionary substrate through which surface-associated capsid change proceeds while maintaining local RNA properties. Structural work in human astroviruses places major exposed and variable features in the capsid projection (Dong et al. 2011; Toh et al. 2016), and canine sequence studies have documented central ORF2 variability (Caddy and Goodfellow 2015). Directly assigning antigenic or receptor-binding functions to the canine coordinates, however, will require canine-specific structural and experimental evidence.

CpG and UpA are biologically plausible motifs because experimental recoding can influence replication and innate recognition in some vertebrate virus systems (Atkinson et al. 2014; Tulloch et al. 2014; Takata et al. 2017). The present natural-sequence pattern may therefore be consistent with host-associated pressures, but it does not identify ZAP activity, immune escape or altered virulence. CaAstV replication occurs in a canine cellular environment where translation, RNA turnover and innate sensing may interact with viral RNA composition. Other processes—such as structure, packaging and replication fidelity—could produce the same statistical signature. The most direct follow-up would use synonymous recoding that changes boundary CpG or UpA while preserving the protein and overall codon inventory, followed by RNA-structure, expression and infection measurements.

The three dog-relative indices add a complementary but distinct evolutionary view. ORF2 had higher CAI and lower RCDI and D than the two replicase genes. Concordance across metrics suggests that the capsid coding sequence is relatively closer to pooled domestic-dog coding preferences. One possible interpretation is that repeated expression of ORF2 during circulation in dogs has allowed some gene-specific host correspondence to accumulate. However, CAI, RCDI and D quantify different aspects of codon usage, and the reference pools genes across tissues and expression levels. They cannot establish translation efficiency, tissue tropism or host adaptation without canine-cell expression and tRNA data. The stronger local CpG/UpA depletion in central ORF2, together with its global codon correspondence, instead provides a focused hypothesis that host-associated evolution may operate at more than one synonymous scale.

Lineage history and recombination provide further context. Global CaAstV analyses describe multiple groups and possible recombination, including events involving ORF2 (Mihalov-Kovács et al. 2017; Qiu et al. 2026). Exact duplicate removal alone cannot make all genomes phylogenetically independent. Nevertheless, reducing the dataset to 99%, 98% and 95% representatives retained the central gene rankings and the regional ORF2 directions, showing that the principal patterns are not confined to one set of identical sequences. The weakest pairwise gene contrasts lost statistical resolution in the 12-representative set, appropriately limiting claims of universal lineage-independent differences. Recombination-aware phylogenetic models and denser balanced sampling could distinguish inherited ordering from repeated convergence.

Several additional limitations define the interpretation. Public metadata and annotations are heterogeneous even after accession-level auditing. Small accessory ORFs could not be compared consistently, so 'all genes' here means all three principal, reliably annotated coding regions. Pairwise NG86 ratios are descriptive and cannot localize selected sites. The conditional shuffle does not preserve long-range RNA structure, transcriptional signals or constraints in an overlapping frame. Finally, no infection, expression, tRNA, RNA-structure or phenotypic experiment was performed. These limitations do not erase the observed ordering; they specify which mechanistic and evolutionary explanations remain open.

In conclusion, CaAstV combines gene-scale and subgene-scale organization of synonymous evolution. ORF1b has the strongest whole-gene CpG/UpA codon-order depletion, ORF2 is the most protein-divergent and dog-correspondent gene, and the rapidly evolving central ORF2 region carries an especially strong local depletion. The findings suggest that viral protein change, global codon usage and local dinucleotide ordering are related but non-equivalent layers of genome evolution. This hierarchical framework provides experimentally testable candidates for future studies of RNA organization and possible host-associated adaptation in canine astrovirus.

Declarations

Competing interests:

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval:

Not applicable. The study used only publicly available nucleotide sequences and involved no animal sampling or intervention.

Consent to participate:

Not applicable.

Consent for publication:

Not applicable.

Funding:

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Author Contribution

J.-W.Y. conceived the study, curated the public data, designed and executed the analyses, verified the source records and outputs, interpreted the results, prepared the figures and drafted the manuscript. C.-H.L. and H.-W.H. contributed to study-design discussions, interpretation, supervision and critical manuscript revision. W.-M.H. and S.-L.H. contributed to research discussions, interpretation, research-environment support and review of the final manuscript. All authors approved the submitted version.

Acknowledgement

The authors thank Mr. Lee, Jen-Chung Yang, Lai-You Yang and Kuei-Ying Hsu for encouragement, inspiration and support.

Data Availability

All viral accession numbers, source-record audit decisions, CDS-level CpG/UpA results, gene summaries, identity-representative sensitivities, ORF2 regional results and deterministic seeds are supplied in Online Resource 1. The domestic-dog codon reference was derived from NCBI RefSeq Canis lupus familiaris Annotation Release 106, assembly ROS_Cfam_1.0 (GCF_014441545.1). No new biological sequences were generated.

References

1. Arias CF, DuBois RM (2017) The astrovirus capsid: a review. Viruses 9:15. https://doi.org/10.3390/v9010015
2. Atkinson NJ, Witteveldt J, Evans DJ, Simmonds P (2014) The influence of CpG and UpA dinucleotide frequencies on RNA virus replication and characterization of the innate cellular pathways underlying virus attenuation and enhanced replication. Nucleic Acids Res 42:4527–4545. https://doi.org/10.1093/nar/gku075
3. Belalov IS, Lukashev AN (2013) Causes and implications of codon usage bias in RNA viruses. PLoS One 8:e56642. https://doi.org/10.1371/journal.pone.0056642
4. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
5. Caddy SL, Goodfellow I (2015) Complete genome sequence of canine astrovirus with molecular and epidemiological characterisation of UK strains. Vet Microbiol 177:206–213.

https://doi.org/10.1016/j.vetmic.2015.03.011

6. Cock PJA, Antao T, Chang JT et al (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25:1422–1423. https://doi.org/10.1093/bioinformatics/btp163
7. Deorowicz S, Debudaj-Grabysz A, Gudyś A (2016) FAMSA: fast and accurate multiple sequence alignment of huge protein families. Sci Rep 6:33964. https://doi.org/10.1038/srep33964
8. Di Giallonardo F, Schlub TE, Shi M, Holmes EC (2017) Dinucleotide composition in animal RNA viruses is shaped more by virus family than by host species. J Virol 91:e02381-16. https://doi.org/10.1128/JVI.02381-16
9. Dong J, Dong L, Méndez E, Tao Y (2011) Crystal structure of the human astrovirus capsid spike. Proc Natl Acad Sci USA 108:12681–12686. https://doi.org/10.1073/pnas.1104834108
10. Gaunt ER, Digard P (2022) Compositional biases in RNA viruses: causes, consequences and applications. WIREs RNA 13:e1679. https://doi.org/10.1002/wrna.1679
11. Martella V, Moschidou P, Lorusso E et al (2009) Genetic characterization of a new astrovirus detected in dogs suffering from diarrhoea. Vet Microbiol 139:147–152. https://doi.org/10.1016/j.vetmic.2009.04.031
12. Mihalov-Kovács E, Martella V, Lanave G et al (2017) Genome analysis of canine astroviruses reveals genetic heterogeneity and suggests possible inter-species transmission. Virus Res 232:162–170. https://doi.org/10.1016/j.virusres.2016.12.005
13. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426. https://doi.org/10.1093/oxfordjournals.molbev.a040410
14. Phipson B, Smyth GK (2010) Permutation P-values should never be zero: calculating exact P-values when permutations are randomly drawn. Stat Appl Genet Mol Biol 9:Article 39. https://doi.org/10.2202/1544-6115.1585
15. Puigbò P, Aragonès L, Garcia-Vallvé S (2010) RCDI/eRCDI: a web-server to estimate codon usage deoptimization. BMC Res Notes 3:87. https://doi.org/10.1186/1756-0500-3-87
16. Qiu Y, Lai L, Wang M et al (2026) Evolutionary and phylogenetic insights from global canine astrovirus genomes. Arch Virol 171:154. https://doi.org/10.1007/s00705-026-06620-6
17. Sharp PM, Li WH (1987) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295. https://doi.org/10.1093/nar/15.3.1281
18. Silverj A, Rota-Stabelli O (2020) On the correct interpretation of similarity index in codon usage studies: comparison with four other metrics and implications for Zika and West Nile virus. Virus Res 286:198097. https://doi.org/10.1016/j.virusres.2020.198097
19. Takata MA, Gonçalves-Carneiro D, Zang TM et al (2017) CG dinucleotide suppression enables antiviral defence targeting non-self RNA. Nature 550:124–127. https://doi.org/10.1038/nature24039

20. Toh Y, Harper J, Dryden KA et al (2016) Crystal structure of the human astrovirus capsid protein. J Virol 90:9008–9017. https://doi.org/10.1128/JVI.00694-16
21. Tulloch F, Atkinson NJ, Evans DJ, Ryan MD, Simmonds P (2014) RNA virus attenuation by codon pair deoptimisation is an artefact of increases in CpG/UpA dinucleotide frequencies. eLife 3:e04531. https://doi.org/10.7554/eLife.04531
22. Wright F (1990) The 'effective number of codons' used in a gene. Gene 87:23–29. https://doi.org/10.1016/0378-1119(90)90491-9
23. Zhang W, Wang R, Liang J et al (2020) Epidemiology, genetic diversity and evolution of canine astrovirus. Transbound Emerg Dis 67:2901–2910. https://doi.org/10.1111/tbed.13663

Figures

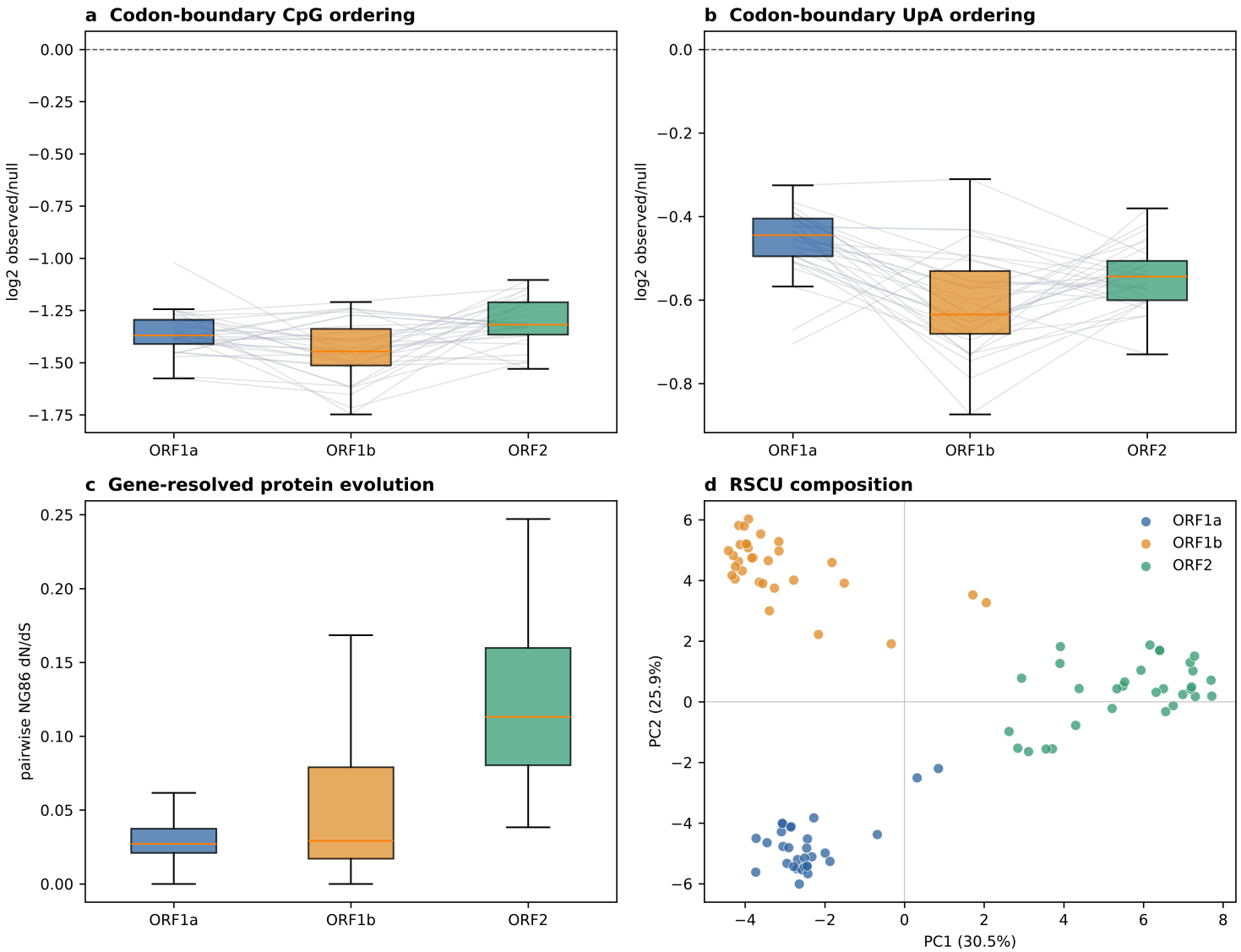


Figure 1

Gene-resolved protein, codon-usage and conditional ordering evolution. a and b Paired CDS-level log2 observed/null effects for codon-boundary CpG and UpA after 9,999 exact-codon randomizations. Lines join the three genes from one genome. c Pairwise NG86 dN/dS distributions, shown as descriptive context. d RSCU principal component analysis; PC1 and PC2 explain 30.5% and 25.9% of variance

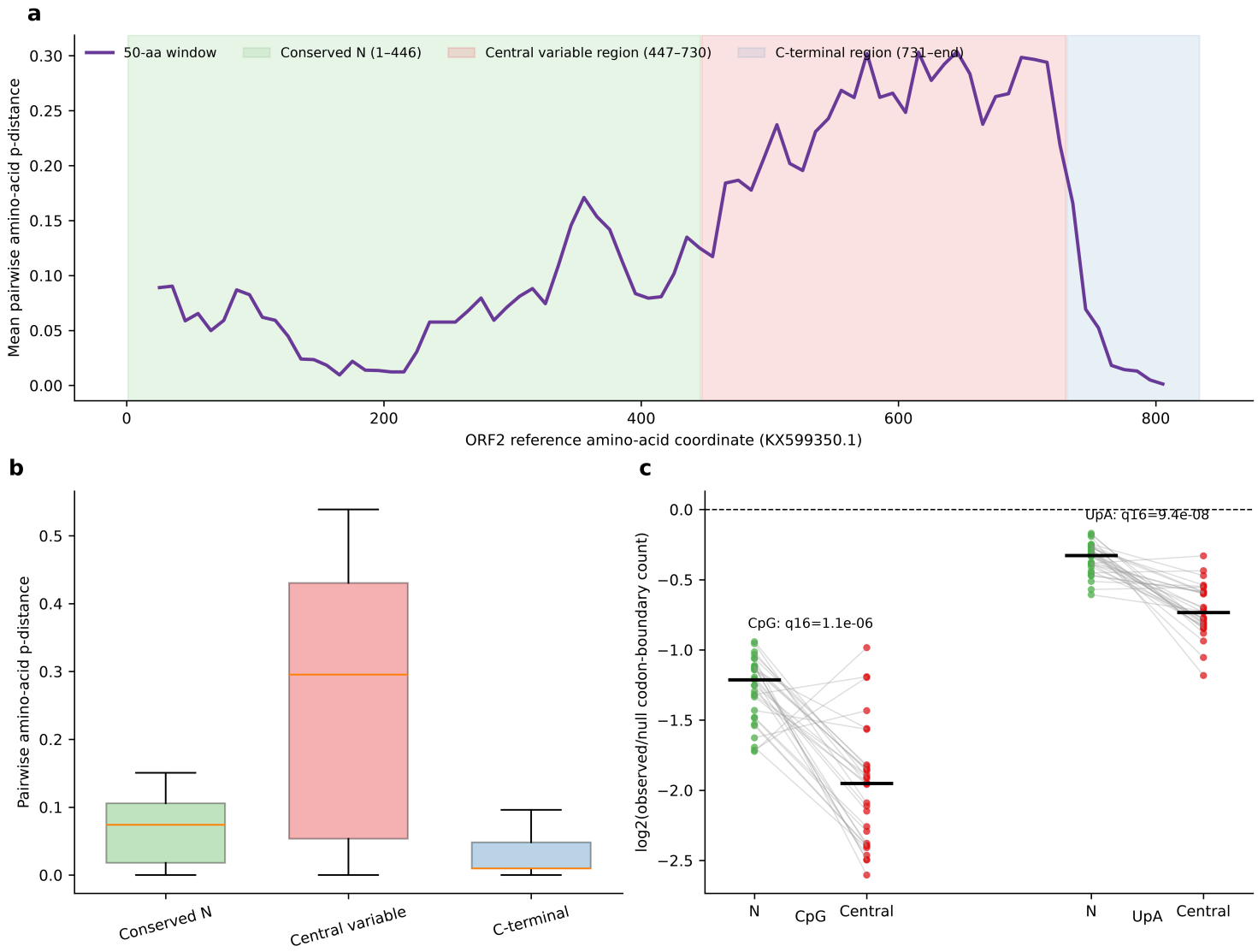


Figure 2

Regional evolution within canine astrovirus ORF2. a Descriptive 50-amino-acid sliding profile of mean pairwise amino-acid p-distance across KX599350.1. Backgrounds mark the predefined N-terminal (1–446), central (447–730) and C-terminal (731–834) regions. b Regional pairwise amino-acid p-distances. c Paired CpG and UpA conditional effects in the N-terminal and central regions

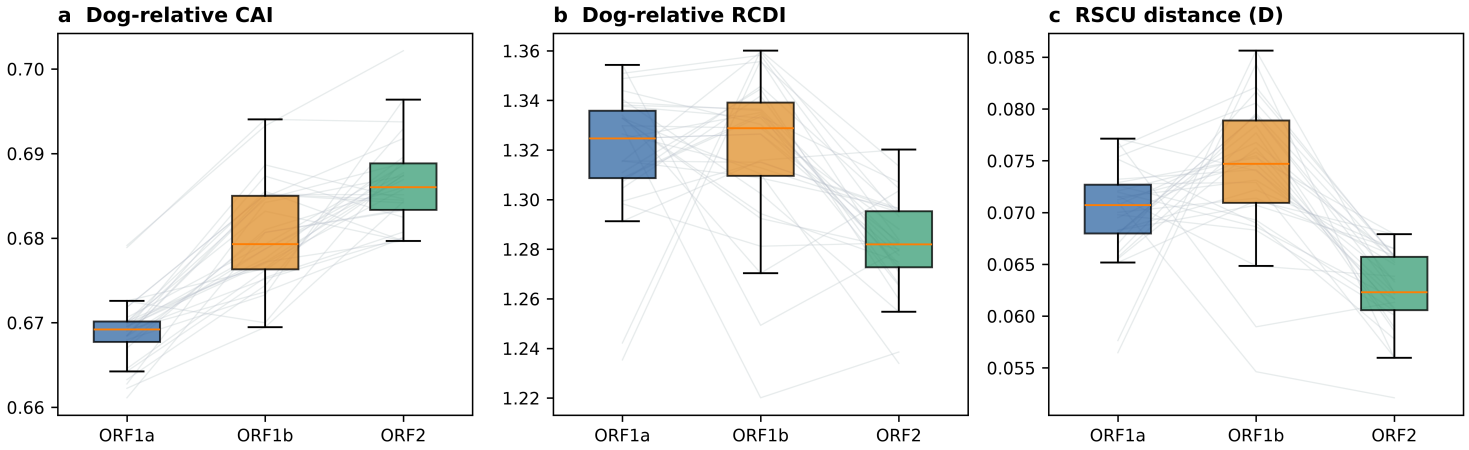


Figure 3

Dog-relative codon correspondence by gene. a CAI, b RCDI and c RSCU cosine distance D. Each line joins the three CDS from one genome. These indices describe relative correspondence to pooled dog CDS and do not by themselves establish host adaptation

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 03OnlineResource1CaAstVAllGenes.xlsx