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.

Breakpoint structure and inter-clade exchange in recombinant duck circovirus genomes

Sharadindu Shil

dr.sharadindu@gmail.com

Genomics & Molecular Biology Laboratory (GMBL), Paschimbanga Go Sampad Bikash Sanstha (PBGSBS)

Anusuya Debnath

Brainware University

Research Article

Keywords: duck circovirus, recombination, breakpoint, clade, purifying selection

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11441655/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

Recombination shapes the evolution of circoviruses, yet its extent and its breakpoints in duck circovirus (DuCV) are known only from regional datasets analysed in different alignment coordinates. We screened 444 non-redundant complete DuCV genomes from GenBank with the pairwise homoplasy index, a genetic algorithm for recombination detection (GARD) and a phylogenetic discordance score, and confirmed the 14 most discordant genomes with three independent breakpoint analyses, all reported in the numbering of one reference genome. Recombination signal was significant ($p = 0.001$), and GARD placed four breakpoints at positions 585, 971, 1141 and 1697. The 14 extreme genomes fell into four breakpoint signatures and, after grouping by shared breakpoints and parents, represent at most nine events, and the ten group 1 genomes are not one lineage. Breakpoints lay in the replicase gene (*rep*), the capsid gene (*cap*) and the intergenic regions. In 13 of the 14 genomes the two parental arcs came from different major clades, DuCV-1 and DuCV-2, so exchange crosses the deepest split in the species, and because breakpoints cut through *rep*, a clade call from *rep* was ambiguous in 12 of the 14 genomes and in none of the others. The data carry no clock-like signal ($R^2 = 0.01$), so the events cannot be dated. Selection on recombination-free parts of *rep* and *cap* was mostly purifying, with five candidate positively selected sites that survived stricter filtering. All 14 genomes are Sanger sequences without public raw reads, so the events should be confirmed from reads.

Introduction

Duck circovirus (DuCV) is a small, non-enveloped virus of the species *Circovirus duck* (genus *Circovirus*, family *Circoviridae*), with a circular single-stranded DNA genome of 1,987–1,996 nt. The genome carries three main open reading frames: *rep*, which encodes the replicase, *cap*, which encodes the capsid protein, and a third frame encoding an apoptosis-related protein, with non-coding regions between *rep* and *cap* (Zhang et al. 2023). DuCV was first described in ducks in Germany in 2003. It mainly impairs the immune system, and the resulting immunosuppression can lead to secondary infections and poor growth (Zhang et al. 2023).

Recombination is reported repeatedly in DuCV, but each study has examined a different, mostly regional, sample. Liu et al. (2020) sequenced 19 genomes from southern and southwestern China and placed putative breakpoints at nt 989 and 1122 in three Yunnan strains. Li et al. (2025) found 23 recombination events among 317 sequences, counting an event when three or more of seven RDP-package methods supported it. Dong et al. (2025) reported recombination in the *cap* gene of three Vietnamese strains, with a breakpoint at site 619. Wang et al. (2026) predicted 17 events among 13 new and 104 reference sequences, 10 of them involving goose-derived strains.

These studies differ in sample, tool and coordinate system. Li et al. (2025), for example, give breakpoints as aligned positions, not original sequence positions. Most rely on one family of detection methods. It is therefore unclear how often recombination occurs across all public complete genomes, whether breakpoints recur at the same places, and whether they lie in *rep*, *cap* or the regions between. These are

questions about how recombination acts on the genome of the species, and they matter because exchange between lineages breaks the correspondence between a gene tree and the history of the whole genome.

We therefore screened all 444 non-redundant complete DuCV genomes in GenBank and asked four questions about how recombination shapes the species. How many independent events do the recombinant genomes represent? Do breakpoints recur at the same positions, and do they fall in *rep*, *cap* or the regions between? Do the two parental arcs come from the same clade or from different clades, and what does this do to clade assignment from a single gene? And what selection acts on the parts of the genome that are free of recombination? We localised breakpoints with a genetic algorithm for recombination detection (GARD), ranked genomes by phylogenetic discordance between genome segments, and confirmed the 14 most discordant genomes with three independent analyses, adding GENECONV as a further check. We also tested for temporal signal before considering whether the events could be dated. All breakpoints are reported in the numbering of one reference genome (HM162346.1).

Methods

Data and alignment: Duck circovirus genomes (1,500–2,500 bp) were retrieved from NCBI nuccore through E-utilities on 18 October 2026 as the union of three queries: the organism subtree txid324685; titles containing "duck circovirus" or "DuCV" in any taxon; and the same title terms within the Circovirus subtree, to catch mis-assigned taxonomy. Of 528 records fetched, identical sequences and one non-waterfowl host were removed, leaving 498; 444 were complete (≥ 1,960 nt) and formed the analysis set. Genomes were oriented and rotated to the origin of HM162346.1 by shared 14-mer voting (161 rotated, none reverse-complemented), aligned with MAFFT v7.526 (Katoh and Standley 2013) (--auto), and columns with < 95% non-gap occupancy were removed (2,032 to 1,980 columns). Coordinates are HM162346.1 numbering unless stated.

Phylogenetics

Maximum-likelihood trees were built with IQ-TREE 3.1.4 (-m MFP; TIM3 + F+R4 chosen by BIC; 1,000 ultrafast bootstrap and 1,000 SH-aLRT replicates).

Overall recombination signal

The pairwise homoplasy index test (Bruen et al. 2006) was implemented in Python. It uses four-gamete incompatibility between the 370 biallelic informative sites, the mean incompatibility within 100-site windows, and a one-sided null from 1,000 permutations of site order. Positive controls, in which 5%, 10%, 25% or 50% of taxa were replaced by manufactured recombinants, were all detected ($p = 0.001$, $z = -5.3$ to -12.5).

Negative control for the PHI test: To measure the false-positive rate of the permutation test, we simulated 200 recombination-free alignments of 444 sequences and 1,980 sites on the maximum-likelihood tree under the fitted TIM3 + F+R4 model, in two ways: with each site drawing its rate category independently, and with sites keeping the rank of their variability in the real alignment, so that fast and slow sites sit at the same positions. The gaps and ambiguous bases of the real alignment were imposed on each replicate, and each replicate was analysed with 199 permutations (script phi_negctl.py in the archive).

Breakpoint detection (GARD): GARD (Kosakovsky Pond et al. 2006) in HyPhy 2.5.103 (Kosakovsky Pond et al. 2020; nucleotide data, Faster mode) was run on 60 genomes chosen by greedy farthest-point sampling on pairwise differences. Two runs on this set differed only in rate variation: run B used none and run C used gamma-distributed rates. Run C is reported. A partial earlier run on a 100-genome set (gamma) placed the first breakpoint at 676. Run B's five-breakpoint model (alignment columns 664, 970, 1127, 1687 and 1855) defined six segments for the discordance screen.

Discordance screen

For each genome, patristic distances to all other genomes were taken from the IQ-TREE consensus tree of each segment. Discordance was 1 minus the mean Pearson correlation of these distance profiles between pairs of segment trees (segments 1, 2 and 4 for the main score; all six for a secondary score). Genomes above median + 3 × MAD-sigma of the main score were flagged, and those with a main score above 0.5 were called extreme. Candidate parents were nearest neighbours in the segment trees.

Confirmation of the extreme genomes

Three analyses were run on the 14 extreme genomes

1. A MaxChi-type triplet scan (Smith 1992) with 2,000 permutations.
2. 3SEQ v1.8.0 (Lam et al. 2018) (-full, default settings, packaged 700-site P-value table) on 71 sequences: the 14 extreme genomes, their candidate parents, and up to eight random genomes per genotype (seed 7).
3. An informative-site scan. For each genome's top 3SEQ parent pair, at sites where the parents differ, we found the one- or two-breakpoint partition of the circular genome that best explained the genome's bases. This scan is descriptive and has no null model.

Follow-up of marginal flagged genomes: The 45 genomes flagged by the discordance screen but not among the 14 extreme genomes (score ≤ 0.5) were examined with an informative-site mosaic scan. For each genome, candidate parents were its ten nearest genomes by whole-alignment p-distance plus its nearest neighbors in the six segment trees. For every candidate pair (P, Q), we took the sites at which P and Q differ and the query matches one of them, and scored the best two-breakpoint partition of the circular genome (an interior arc plus its complement) against the best single-parent score; the difference is the "gain", and the fit is the mosaic score divided by the number of informative sites. The 385

unflagged genomes were scanned identically as a reference. A site-order permutation null was tried and rejected: 43 of 60 unflagged control genomes came out significant, because rate variation and phylogenetic structure produce blocks of similar sites in real genomes. We therefore report the scan descriptively and compare flagged genomes with unflagged ones empirically, without p-values. Because parents are drawn from nearest relatives, the scan cannot separate a donor from a recombinant.

Genotype context

Each genome was assigned to the nearest reference sequence for DuCV-1 (subgenotypes 1a–1d) and DuCV-2 (2a–2c), for the whole genome and for each of the six segments. The 2a reference was not used as an anchor for subgenotype calls because it is itself a recombinant in our analysis. Subgenotype labels were therefore treated as unstable and the analysis relied on the DuCV-1 versus DuCV-2 (clade) call. For the 14 extreme genomes, the arc between the two breakpoints and the complementary arc were assigned to a clade separately, using the informative columns in each arc.

***rep* versus *cap* trees**

Clade calls from the *rep* region (HM162346.1 nt 48–927, + strand) and the *cap* region (nt 1156–1930, − strand) were compared for every genome, and the number of genomes with different calls was tested against the extreme and non-extreme groups with Fisher's exact test.

Selection: Codon alignments of *rep* and *cap* were split into recombination-free partitions (*rep*: before and after CDS nt 630; *cap*: before and after CDS nt 234) after removing the 14 extreme genomes and collapsing identical sequences; a separate maximum-likelihood tree was built for each partition with IQ-TREE. We ran HyPhy 2.5 (Kosakovsky Pond et al. 2020) FUBAR (Murrell et al. 2013; posterior probability ≥ 0.9), FEL (Kosakovsky Pond and Frost 2005) and MEME (Murrell et al. 2012; $p \leq 0.1$) for site-level selection, and BUSTED (Murrell et al. 2015) with synonymous rate variation for gene-wide evidence of episodic positive selection on each partition. A site was called supported when at least two of the three site-level methods agreed on positive selection. Site numbers are reported as codons of the HM162346.1 coding sequence. As a sensitivity analysis the whole analysis was repeated after removing all 59 flagged genomes (the 14 extreme and 45 marginal genomes), with a GTR + F+R4 tree for each partition (script sel_noflag.sh in the archive).

GENECONV

As a further, independent detection method, GENECONV (Sawyer 1989; version from the author's Unix source) was run on the 444-genome alignment, rotated by 1000 columns so that the origin does not fall at the alignment ends, with nucleotide mode, 1000 permutations and seed 123, and global inner fragments with simulated $p \leq 0.01$ were retained. Fragment endpoints were converted back to original alignment columns. A genome was counted as supported if any fragment involving it started and ended within 30 nt of its own two breakpoints. The same criterion was applied to the non-extreme genomes against the breakpoint pairs of any of the nine events.

Temporal signal

Root-to-tip regression was run in TreeTime (Sagulenko et al. 2018) on the whole-genome maximum-likelihood tree after removing the 14 extreme genomes and the genomes with no collection year (335 tips, sampled 2006–2026).

Breakpoint recurrence

For the nine event clusters we took the midpoint of each breakpoint interval in alignment columns and used as statistic the largest number of the 18 breakpoints inside any window of 100 or 150 nt on the circular genome. The null distributions (200,000 permutations each) placed breakpoints uniformly on the circle, drew them from the columns of the 370 biallelic informative sites, or rotated each event's pair of breakpoints as a rigid arc to a uniform position (script hotspot_test.py in the archive).

Monophyly of group 1

To ask whether the ten group 1 genomes descend from one lineage, we built maximum-likelihood trees (IQ-TREE, model chosen by ModelFinder, 1,000 ultrafast bootstraps, identical sequences kept) for the segment between the two group 1 breakpoints (alignment columns 80–570; TIM2 + R3) and for the rest of the genome (columns 730–1940; TIM3 + F+R4). The ten genomes are EU344802, EU344803, EU344805 and EU499311 (Fujian) and MN928792, MN928802, MN928803, MN928804, MN928806 and MN928811 (Henan). On each unrooted tree we asked whether the ten, the four Fujian and the six Henan genomes each formed a clade. We then searched for the best tree with the ten genomes constrained to be monophyletic and compared it with the unconstrained tree by the approximately unbiased (AU) test with 10,000 RELL resamplings (scripts g1_monophyly.sh and g1_post.sh in the archive). The same constrained-tree comparison was applied to the six Henan genomes, the four Fujian genomes and the pair EU344803 and EU499311 (script g1_subgroups.sh) and, separately, to cluster A (four Henan genomes) and cluster B (two) (script g1_clusters_AB.sh).

Replication origin

We searched HM162346.1 on both strands for the degenerate circovirus origin nonanucleotide NANTATTAC and for stem-loops (stems of at least 6 bp with G–T pairs allowed, loops of 3–12 nt), and measured the distance from each breakpoint interval midpoint to the nearest nonamer match, comparing the mean with 20,000 random placements of the same number of breakpoints (script origin_motif.py in the archive).

Use of a large language model

A large language model (Claude, Anthropic) assisted with debugging the scripts. The authors ran all analyses, checked the results against the outputs and the references against their sources, and take full responsibility for the content.

Results

Global signal and GARD: The alignment showed significant recombination signal (PHI 0.339 against a permuted null of 0.358 ± 0.005; z = − 3.77; p = 0.001, the minimum possible with 1,000 permutations). GARD's best model (run C) had four breakpoints at 585, 971, 1141 and 1697 (c-AIC 27107.7); a fifth breakpoint did not improve it. The first breakpoint varied between runs: 577–583 in run C, 664–670 in run B and 676 in the partial 100-genome run.

Negative control for the PHI test

In recombination-free simulations the test called 5.0% of independent-rate replicates and 2.0% of position-preserving replicates significant at $p < 0.05$ (200 replicates each; 1.5% and 0.5% at $p \leq 0.01$). The lowest simulated z was − 2.47, against − 3.77 for the real alignment with 1,000 permutations and − 4.04 in the same 199-permutation pipeline, and no simulated replicate reached the real value. The simulations produced fewer informative sites than the real alignment (mean 268–269 against 370), so they do not reproduce every feature of the data.

Discordance screen

Of 444 genomes, 59 were flagged and 14 were extreme (main score 0.5–1.30). The 59 flagged genomes had a median score of 0.142 (minimum 0.14), so the other 45 lay close to the threshold.

Confirmation of the extreme genomes

MaxChi gave p = 0.0005, the minimum for 2,000 permutations, for 13 of the 14. KP943594 gave p = 0.0235 from 34 informative sites, about 0.33 after Bonferroni correction for 14 tests. 3SEQ reported significant recombinant triplets for all 71 subset sequences (53,877 triplets, p below printed precision), so its p-values did not separate genomes and we used its parent pairs and breakpoint intervals. A two-parent mosaic explained 96.9–100% of informative sites in 13 genomes against 53–86% for the best single parent (40–110 more sites), and 69% against 51% for KP943594.

Four breakpoint signatures

Breakpoints are 3SEQ intervals in HM162346.1 numbering, with the mosaic fit shown against the best single parent (Table 1).

Table 1
Breakpoint signatures of the 14 extreme genomes. Breakpoints are 3SEQ intervals in HM162346.1 numbering; mosaic fit is the percentage of informative sites explained by the two-parent mosaic against the best single parent

Group	Genomes	Breakpoint 1	Breakpoint 2	Mosaic fit vs best single parent
1	EU344802, EU344803, EU344805, EU499311, MN928792, MN928802, MN928803, MN928804, MN928806, MN928811	1–25 at the origin junction (EU344802 and the six MN928xxx genomes) or 46–72 (EU344803, EU344805, EU499311)	within *rep*, 579–682 (579–726 for MN928792 and MN928811)	97.2–100% vs 66.2–84.2%
2	OR134497	1110–1125 (intergenic)	1649–1723 (*cap*)	96.9% vs 55.1%
2	OR134498	871–975 (spans the 3' end of *rep*)	1655–1664 (*cap*)	98.3% vs 52.9%
3	OK094644	198–222 (*rep*)	582–672 (*rep*)	97.7% vs 86.0%
4	KP943594	180–192 (*rep*)	1125–1138 (intergenic)	69.0% vs 51.3%

In the six MN928xxx genomes, informative sites matched the GQ4237xx parents in alignment columns 1–577 (54 of 55 sites) and the MN9288xx parents in columns 725–1950 (154 of 154), with the switch between columns 578 and 681. OR134497 and OR134498 use each other as parents, so they probably represent one event.

Agreement across methods

GARD's 585/586 falls inside group 1's breakpoint interval, 971/972 inside OR134498's and 1697/1698 inside OR134497's. GARD's 1141/1142 lies 3 nt beyond KP943594's interval. MaxChi's best cuts fall inside the corresponding intervals for group 1 (674–681), OR134497 (1115) and OR134498 (894). The exception is OK094644, whose cut (681) lies about 10 nt beyond its interval end (672).

Independent events: The 14 extreme genomes represent fewer than 14 recombination events. Grouping genomes whose two breakpoints agree within a tolerance and that share a parent gives 9 clusters at tolerances of 20 and 60 nt, and 10 at 5 nt. Four genomes (MN928802, MN928803, MN928804, MN928806) share breakpoints at 1–25 and 579–675 and the same parents, and so are one event; MN928792 and MN928811 share a second event (breakpoints 1–25 and 579–726); and EU344803 and EU499311 share a third (breakpoints near 59 and 579–682). EU344802, EU344805, OR134497, OR134498, OK094644 and KP943594 are each their own cluster. The clustering is descriptive: parents are nearest relatives, and the clusters that share the left breakpoint at 1–25 may descend from a single ancestral event, which would lower the count further.

Recurrence of breakpoints

Treating the nine event clusters as independent, six of the 18 breakpoints fall in one 100-nt window of *rep* (alignment columns 626–676, interval midpoints). The probability of this many in one window is 0.018 when breakpoints are placed uniformly on the circular genome, 0.029 when they are drawn from the 370 informative-site columns, and 0.008 when each event's pair of breakpoints is rotated as a rigid arc; with a 150-nt window the probabilities are 0.09, 0.12 and 0.04. These tests are exploratory, and they rest on counting the five group 1 clusters as independent events. If the two Henan clusters (A and B) were merged into one event, five of 16 breakpoints would share the window and the probabilities would rise to 0.07, 0.10 and 0.04 for a 100-nt window and 0.24, 0.29 and 0.16 for a 150-nt window; the next paragraph tests whether that merger is justified.

Group 1 is not one lineage

Forcing the ten group 1 genomes to be monophyletic lowered the log-likelihood by 172 units in the segment between the breakpoints (columns 80–570) and by 1,107 units in the rest of the genome, and the AU test rejected the constrained trees in both ($p = 8.5\times10^{-34}$ and 1.5×10^{-49}). On neither unrooted tree did the ten genomes or the four Fujian genomes form a clade. The six Henan genomes formed a clade in the segment between the breakpoints, with only 19% bootstrap support, and forcing them together there was not rejected (ΔlnL 20.7, AU p = 0.19), but forcing them together in the rest of the genome was rejected (ΔlnL 405, $p = 3.4\times10^{-64}$). The Henan genomes therefore share a similar segment on different backgrounds, which is what separate acquisitions of the segment would give and not descent from one recombinant. The four Fujian genomes were rejected as a clade in both arcs (ΔlnL 158 and 523; $p = 1.8\times10^{-53}$ and 7.9×10^{-66}), whereas the two genomes of cluster C (EU344803 and EU499311) were not rejected in either (AU p = 0.47 and 1.0; ΔlnL 2.8 and 0). The ten genomes therefore do not descend from one recombinant ancestor. Forcing the four genomes of cluster A, or the two of cluster B, to be monophyletic was not rejected in either arc (AU p = 0.41 and 0.41 for A; 0.30 and 0.39 for B), so the nine-cluster grouping is consistent with the trees.

Replication origin

HM162346.1 contains one NANTATTAC match, AAGTATTAC at position 600, and it lies inside the right-hand breakpoint interval of group 1 (579–726). The stem-loop search found stem-loops at 630–659 and 657–677 in the same interval. The 18 breakpoint midpoints lay on average 393 nt from the nonamer match, against 495 nt for random positions (P = 0.068).

Provenance of the extreme genomes: All 14 were Sanger-sequenced and none has a BioProject, BioSample or SRA record, so the original reads cannot be re-examined. The group 1 genomes come from two unrelated submissions about a decade apart: EU344802, EU344803, EU344805 and EU499311 (Fujian Academy of Agricultural Sciences, deposited around 2008, unpublished), and MN928792, MN928802–MN928806 and MN928811 (a Henan laboratory, samples from 2018–2019). Two laboratories finding the same right-hand breakpoint argues against a single-batch artefact. OR134497

and OR134498 are from one Guangxi submission, OK094644 is a single unpublished Sichuan study, and KP943594 is the only genome from a wild host (velvet scoter, Poland) and the only one with a published description (Matczuk et al. 2015). The terminal 1–25 and 1951–1980 sequences of the group 1 genomes are close to those of their parents and carry no shared primer-like signature.

GENECONV as a supporting method: GENECONV reported 2000 significant global inner fragments, the same number as in a 100-permutation run, which suggests an internal limit on the list, and these involved 426 of the 430 non-extreme genomes, so it is not specific in this near-clonal dataset. Even so, a fragment with both ends within 30 nt of the genome's own breakpoints was found for 5 of the 14 extreme genomes (EU344802, EU344805, OK094644, OR134497, OR134498), against 52 of 426 (12%) of the other genomes when they were allowed to match any of the nine events (Fisher's exact test p = 0.024, unadjusted and exploratory). The fragments for EU344802 (origin to about 667) and EU344805 (46–667) also named GQ423742, the inside parent in our scan, as the partner. We treat this as modest independent support: it is enriched in the extreme genomes, but it is not a confirmation of each event.

Temporal signal: The data carry almost no clock-like signal: root-to-tip distance explained 1% of the variance in sampling year ($R^2 = 0.01$) and the implied root date (about 1866) is not plausible, so we did not attempt Bayesian dating of the recombinant lineages. The age of the events relative to their parents is therefore unknown; the only bound is that the group 1 genomes EU344802, EU344803, EU344805 and EU499311 (collection years not recorded) were deposited in GenBank around 2008 (EU499311 was submitted in February 2008).

Marginal flagged genomes

Most of the 45 marginal genomes are not clearly more mosaic than unflagged genomes. At a strict threshold (gain ≥ 30 informative sites and mosaic fit ≥ 0.90), 5 of 45 marginal genomes qualified, against 8 of 385 unflagged genomes (11% versus 2%), a modest enrichment. Four of the five name one or more extreme genomes as a parent, so they are probably descendants or relatives of the main events; ON227540 is the only one with independent parents. At a looser threshold the marginal and unflagged groups were indistinguishable. We therefore treat the discordance screen as reliable for the 14 extreme genomes and as only suggestive for the rest, and we do not claim that the 45 marginal genomes are recombinants.

Genotype context: Subgenotype labels from nearest-reference assignment proved unstable, and the 2a reference is itself a recombinant, so we did not interpret events in terms of subgenotype pairs. The clade level was stable. Assigning the two arcs of each extreme genome separately, 13 of the 14 were inter-clade: one arc fell in DuCV-1 and the other in DuCV-2. Examples are EU344802 (617 columns DuCV-1 inside the arc, 1363 DuCV-2 outside), the group 1 genomes (the arc DuCV-2, the rest DuCV-1), OR134498 and OK094644 (arc DuCV-1, rest DuCV-2), and KP943594 (arc DuCV-2, rest DuCV-1, with a poor overall fit). The outside arc of OR134497 was ambiguous. Flagged genomes were not enriched in any country: 49 of the 59 flagged genomes came from China, which supplies most of the dataset, and none of the South Korean genomes were flagged.

rep versus cap incongruence: Only 2 of 444 genomes (both extreme) were placed in different clades by the *rep* and *cap* regions. This is a conservative count, because the *rep* clade call was ambiguous in 12 of the 14 extreme genomes against none of the 430 others: for most extreme genomes the *rep* region itself straddles a breakpoint, so a single clade call is not meaningful. The recombinants are therefore not simple *rep*/*cap* swaps; their breakpoints cut through *rep* (Fig. 1), which is also why a whole-gene tree comparison understates them.

Selection: Selection was analyzed on recombination-free partitions (331, 223, 170 and 310 unique sequences for the first and second *rep* and *cap* partitions). BUSTED found gene-wide evidence of episodic positive selection in all four ($p = 1.3\times10^{-6}$, 1.2×10^{-3}, 4.4×10^{-7} and 1.1×10^{-5}), but purifying selection dominated site by site: FEL found 1, 0, 5 and 1 positively selected sites against 81, 52, 34 and 131 negatively selected sites. Seven sites were supported by at least two methods: *rep* codon 71 (FEL and MEME), and in *cap* codons 5, 12, 18, 23 and 106 (FEL, MEME and FUBAR) plus codon 49 (FEL and MEME). We did not map these sites to protein structure or test them experimentally, so we regard them as candidates only. After removing all 59 flagged genomes (291, 191, 145 and 273 unique sequences), BUSTED remained significant in all four partitions ($p = 1.1\times10^{-8}$, 9.4×10^{-3}, 7.5×10^{-8} and 2.8×10^{-5}), FEL found 1, 0, 3 and 1 positively and 83, 46, 31 and 128 negatively selected sites, and five of the seven sites remained supported (*rep* codon 71 and *cap* codons 5, 12, 23 and 106), while *cap* codons 18 and 49 did not. The smaller alignments also have less power, so we cannot say whether those two sites depended on the marginal genomes.

Discussion

Recombination in public DuCV genomes is concentrated, not diffuse. Of 444 complete genomes, 14 were extreme discordance outliers. They fell into four breakpoint signatures and up to nine independent events, and several methods placed the signatures in the same intervals. GARD's four breakpoints (585, 971, 1141 and 1697) each lie inside or within 3 nt of the interval of at least one signature. This agreement matters because the methods use different information: GARD scores alignment-wide likelihood, MaxChi compares sites within triplets, 3SEQ searches parent pairs, and the informative-site scan fits a mosaic to the sites where two parents differ. GENECONV added only weak support, matching the breakpoints of 5 of the 14 genomes and flagging nearly every other genome, so the other methods carry the evidence.

The breakpoints are not confined to one gene. Groups 1 and 3 break within *rep* (group 1 at 579–682; group 3 at 198–222 and 582–672), group 2 breaks in the *cap* region (1649–1723 and 1655–1664) and in the intergenic region or the 3' end of *rep* (1110–1125 and 871–975), and KP943594 breaks in *rep* and the intergenic region (180–192 and 1125–1138). Earlier studies reported breakpoints in the *cap* gene (Dong et al. 2025) and in other positions (Liu et al. 2020; Li et al. 2025), but in their own alignment coordinates, so their positions cannot be compared directly with ours. Reporting breakpoints in the numbering of one reference genome, as here, is what makes such comparisons possible. Breakpoints in circular single-stranded DNA viruses are generally non-random, and they fall less often inside structural protein genes

than inside other genes, a pattern attributed in part to selection against recombinant proteins (Lefeuvre et al. 2009; Martin et al. 2011a). Only one of our four signatures, group 2, places a breakpoint inside *cap*, which is in line with that tendency, but with at most nine events we cannot test it.

Breakpoints that recur at the same positions are what a sequence-linked or replication-linked process would produce, and they are also what descent from one recombinant ancestor would produce. Groups 1 and 3 both break in *rep* within roughly 579–682, and GARD's first breakpoint falls in the same interval, which is the strongest sign of recurrence in these data. Because group 1 supplies ten of the 14 extreme genomes, however, one group drives the pattern, and the permutation test above counts its subclusters as independent events. The ten genomes do not form a clade in either arc (AU $p < 10^{-33}$), the Henan six do not in the background ($p = 3.4\times10^{-64}$) and the Fujian four do not in either arc, so the shared interval is not explained by descent from one recombinant. Clusters A and B are each consistent with a single event (AU $p \geq 0.30$ in both arcs) while the Henan six together are not, and if the two Henan clusters were merged the excess would not be significant at 100 nt for two of the three nulls ($p = 0.07$ and 0.10; 0.04 for the third). Recurrence at the same interval is therefore suggested, not shown. If it is real, it need not reflect a mechanical hotspot alone: in experimental coinfections of begomoviruses, breakpoint positions were influenced by local sequence similarity but not by genome secondary structure, and the overall pattern was shaped by selection against recombinants with disrupted intra-genome interactions (Martin et al. 2011b), so a shared interval could also mark junctions that tolerate recombination. Our data cannot separate these possibilities. The only NANTATTAC match in the genome (position 600), a candidate for the circovirus replication-origin nonamer (the conserved sequence at the apex of the origin stem-loop, which the Rep proteins of porcine circovirus 1 and beak and feather disease virus cleave in vitro; Steinfeldt et al. 2006; Chen et al. 2019), lies inside the right-hand breakpoint interval of group 1, and the breakpoints lie on average 393 nt from it against 495 nt for random positions ($P = 0.068$), a weak test given a single match. We did not verify that this is the functional origin, so we cannot say whether the junctions near the origin or in the intergenic region reflect rolling-circle replication, the mechanism by which circular single-stranded DNA viruses copy their genomes. Testing that would need an experimentally annotated origin and the original sequencing reads.

The practical consequence is for classification. For a recombinant, the two sides of a breakpoint are inherited from different parents, so a genotype assigned from one gene or region can differ from the one assigned from the rest of the genome. Studies that genotype from *cap* or a *rep* fragment alone, or that build a single tree from whole genomes, may therefore misplace the genomes in groups 1–3. The concern is not specific to DuCV. Recombination is a pivotal process in the evolution of single-stranded DNA viruses (Martin et al. 2011a), and the replication-associated protein (Rep), widely used as a marker to classify circular single-stranded DNA viruses, appeared chimeric in 71% of the unclassified members analysed (Kazlauskas et al. 2018). We suggest screening a dataset for recombination before assigning genotypes or building the final phylogeny, and reporting breakpoints in a common coordinate system.

Most of these events join the two major lineages. When the two arcs of each extreme genome were assigned separately, 13 of 14 fell in different clades, one in DuCV-1 and the other in DuCV-2, so these are

not recombinants between close relatives. This is why subgenotype labels were unstable for these genomes. Only 2 of 444 genomes gave different clade calls from *rep* and *cap*, because the breakpoints cut through *rep* itself: the *rep* call was ambiguous in 12 of the 14 extreme genomes and in none of the other 430. The clade assignment rests on nearest-reference calls, and one of the references, 2a, is itself a recombinant in our analysis, so we draw conclusions only at the clade level.

Selection acts on both genes, but mostly as constraint. BUSTED found episodic positive selection in all four recombination-free partitions, while FEL found far more negatively than positively selected sites (for example 131 against 1 in the second *cap* partition). Seven sites were supported by two or more methods, five of which (*rep* codon 71 and *cap* codons 5, 12, 23 and 106) remained supported after all 59 flagged genomes were removed: *rep* codon 71 and *cap* codons 5, 12, 18, 23, 49 and 106. Site-level tests can be misled by recombination unless alignments are screened and split at breakpoints (Kosakovsky Pond et al. 2006), which is why the partitions were analysed separately. BUSTED gives gene-wide evidence that some site on some branch has $\omega > 1$ rather than testing individual sites (Murrell et al. 2015), so the support for individual sites rests on FUBAR, FEL and MEME. Partition boundaries are approximate, so residual recombination may inflate these calls, and we treat the sites as candidates for follow-up.

The 45 marginal flagged genomes were not clearly more mosaic than unflagged genomes (5 of 45 against 8 of 385 at a strict threshold), so we do not call them recombinants. All 14 extreme genomes are Sanger-sequenced and none has raw reads in public archives, so PCR chimeras or mixed infections cannot be excluded directly, although the independent origin of the group 1 genomes argues against a single-batch artefact. The strongest next step is to confirm the events from original reads or fresh sequencing of the isolates.

Group 1 raises a question these data only partly settle. Its ten genomes form five clusters of shared breakpoints and parents (four genomes in one, two in each of two others, and two singletons), and they come from two unrelated laboratories about a decade apart. That fits one lineage that was inherited and spread, and it also fits repeated recombination at a favoured junction. The right-hand breakpoint in *rep* is well supported, but the left breakpoint at the origin is not resolved. Telling the two explanations apart needs dated, geolocated genomes and an explicit test of descent from a common recombinant, and the absence of temporal signal ($R^2 = 0.01$) means the present data cannot be dated. Recombination can obscure root-to-tip signal, and fast-evolving single-stranded DNA circoviruses can be dated when datasets are suitable (a rate of about 1.2×10^{-3} substitutions/site/year was estimated for porcine circovirus 2; Firth et al. 2009), so a dated analysis of the recombination-free partitions is a feasible next step. Because 13 of the 14 extreme genomes are from China, the pattern may also reflect where DuCV has been sequenced most. Complete genomes from other regions would show whether these breakpoints recur elsewhere.

Limitations

The breakpoint evidence agrees across methods, but the parent sets, sampling and the origin-side junction limit how far it can be pushed.

Sampling and parent choice limit the inference. Thirteen of the 14 extreme genomes are from China, so recurring breakpoints may partly reflect where DuCV has been sequenced. Parents were drawn from a 71-genome subset and from nearest neighbours in segment trees, so a true parent missing from the subset would leave a genome poorly fitted, as for KP943594, and the mosaic scan cannot tell a donor from a recombinant because several parents of the extreme genomes look mosaic themselves. Collection years were available for 78% of genomes and sampling is concentrated in 2018–2025, but the whole-genome tree showed no temporal signal ($R^2 = 0.01$), so the events cannot be dated. Recombination detection methods generally capture the presence of recombination but are not very powerful (Posada and Crandall 2001), so failing to flag a genome does not show that it is non-recombinant.

The statistics are mostly descriptive. The 14 extreme genomes were chosen because they were the most discordant, so the confirmation tests were applied to a pre-selected set, and their p-values, such as MaxChi $p = 0.0005$, describe the fit of the chosen genomes and are not corrected for the 444 genomes screened. The positive controls for the pairwise homoplasy index show that the test detects manufactured recombinants, and in recombination-free simulations it gave false-positive rates of 5% and 2% at $p < 0.05$, although the simulated alignments had fewer informative sites than the real one (about 270 against 370), so the control is imperfect; a site-order permutation null did prove invalid for the mosaic scan, so the overall $p = 0.001$ should be read with some caution. The informative-site scan chooses breakpoints from the data and has no null model, and the thresholds for the marginal genomes (gain ≥ 30, fit ≥ 0.90) were set after looking at the data. 3SEQ flagged all 71 sequences, so its p-values did not discriminate, and GARD's first breakpoint differed between runs B and C, which differ only in rate variation, so it is reported as a region (577–676). KP943594 is unresolved: its fit is poor and its MaxChi $p = 0.0235$ does not survive correction.

The left breakpoint of the group 1 genomes (1–25, the origin junction) lies in a highly conserved stretch, is defined by a single informative site at position 25, and is not resolved because GARD treats the alignment as linear; only the right-hand breakpoint (about 579–682 in *rep*) is well supported. Group 1 shares breakpoints across several parent pairs; the ten genomes are not one lineage (AU test), the Henan six share a similar segment on different backgrounds, and the tree of the 491-nt segment is weakly supported (19% for the Henan clade); clusters A and B were each consistent with a single event, but these tests have limited power on short arcs. No raw reads are available for any of the 14 extreme genomes, so PCR chimeras or mixed infections cannot be excluded directly. Subgenotype labels were unstable and the 2a reference is itself a recombinant, so only clade-level (DuCV-1 versus DuCV-2) conclusions are drawn. Finally, the selection partition boundaries (*rep* nt 630, *cap* nt 234) are approximate, so residual recombination may remain inside partitions and can inflate BUSTED and positive-site calls; removing all 59 flagged genomes left BUSTED significant and kept five of the seven sites but also reduced power.

Declarations

Competing interests:

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

Ethics approval:

This study used only publicly available nucleotide sequences from GenBank. No animals were handled and no human participants were involved, so no ethics approval was required.

Funding:

No funds, grants, or other support was received.

Author Contribution

Sharadindu Shil: conceptualization, data curation, formal analysis, investigation, methodology, software, visualization, writing – original draft. Anusuya Debnath: supervision, writing – review and editing. Both authors read and approved the final manuscript.

Data Availability

This study used only publicly available nucleotide sequences from GenBank. No animals were handled and no human participants were involved, so no ethics approval was required.

References

1. Bruen T, Philippe H, Bryant D (2006) A simple and robust statistical test for detecting the presence of recombination. Genetics 172:2665–2681. https://doi.org/10.1534/genetics.105.048975
2. Chen JK, Hsiao C, Wu JS, Lin SY, Wang CY (2019) Characterization of the endonuclease activity of the replication-associated protein of beak and feather disease virus. Arch Virol 164:2091–2106. https://doi.org/10.1007/s00705-019-04292-z
3. Dong HV, Trinh DQ, Tran GHT et al (2025) Characterization of an emerging recombinant duck circovirus in northern Vietnam, 2023–2024. Viruses 17(5):732. https://doi.org/10.3390/v17050732
4. Firth C, Charleston MA, Duffy S, Shapiro B, Holmes EC (2009) Insights into the evolutionary history of an emerging livestock pathogen: porcine circovirus 2. J Virol 83:12813–12821. https://doi.org/10.1128/JVI.01719-09
5. Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780.

https://doi.org/10.1093/molbev/mst010

6. Kazlauskas D, Varsani A, Krupovic M (2018) Pervasive chimerism in the replication-associated proteins of uncultured single-stranded DNA viruses. Viruses 10:187. https://doi.org/10.3390/v10040187
7. Kosakovsky Pond SL, Frost SDW (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22:1208–1222. https://doi.org/10.1093/molbev/msi105
8. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SDW (2006) Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 23:1891–1901. https://doi.org/10.1093/molbev/msl051
9. Kosakovsky Pond SL, Poon AFY, Velazquez R et al (2020) HyPhy 2.5: a customizable platform for evolutionary hypothesis testing using phylogenies. Mol Biol Evol 37:295–299. https://doi.org/10.1093/molbev/msz197
10. Lam HM, Ratmann O, Boni MF (2018) Improved algorithmic complexity for the 3SEQ recombination detection algorithm. Mol Biol Evol 35:247–251. https://doi.org/10.1093/molbev/msx263
11. Lefeuvre P, Lett JM, Varsani A, Martin DP (2009) Widely conserved recombination patterns among single-stranded DNA viruses. J Virol 83:2697–2707. https://doi.org/10.1128/JVI.02152-08
12. Li P, Zhang F, Bao C et al (2025) Epidemiological investigation and analysis of the genetic evolution of duck circovirus in China, 2022. PLoS ONE 20(5):e0323282. https://doi.org/10.1371/journal.pone.0323282
13. Liu H, Li LX, Sun WC et al (2020) Molecular survey of duck circovirus infection in poultry in southern and southwestern China during 2018 and 2019. BMC Vet Res 16:80. https://doi.org/10.1186/s12917-020-02301-x
14. Martin DP, Biagini P, Lefeuvre P, Golden M, Roumagnac P, Varsani A (2011a) Recombination in eukaryotic single stranded DNA viruses. Viruses 3:1699–1738. https://doi.org/10.3390/v3091699
15. Martin DP, Lefeuvre P, Varsani A et al (2011b) Complex recombination patterns arising during geminivirus coinfections preserve and demarcate biologically important intra-genome interaction networks. PLoS Pathog 7:e1002203. https://doi.org/10.1371/journal.ppat.1002203
16. Matczuk AK, Krawiec M, Wieliczko A (2015) A new duck circovirus sequence, detected in velvet scoter (Melanitta fusca) supports great diversity among this species of virus. Virol J 12:121. https://doi.org/10.1186/s12985-015-0352-y
17. Murrell B, Moola S, Mabona A et al (2013) FUBAR: a fast, unconstrained Bayesian approximation for inferring selection. Mol Biol Evol 30:1196–1205. https://doi.org/10.1093/molbev/mst030
18. Murrell B, Weaver S, Smith MD et al (2015) Gene-wide identification of episodic selection. Mol Biol Evol 32:1365–1371. https://doi.org/10.1093/molbev/msv035
19. Murrell B, Wertheim JO, Moola S et al (2012) Detecting individual sites subject to episodic diversifying selection. PLoS Genet 8:e1002764. https://doi.org/10.1371/journal.pgen.1002764

20. Posada D, Crandall KA (2001) Evaluation of methods for detecting recombination from DNA sequences: computer simulations. Proc Natl Acad Sci USA 98:13757–13762. https://doi.org/10.1073/pnas.241370698
21. Sagulenko P, Puller V, Neher RA (2018) TreeTime: maximum-likelihood phylodynamic analysis. Virus Evol 4:vex042. https://doi.org/10.1093/ve/vex042
22. Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6:526–538. https://doi.org/10.1093/oxfordjournals.molbev.a040567
23. Smith JM (1992) Analyzing the mosaic structure of genes. J Mol Evol 34:126–129. https://doi.org/10.1007/BF00182389
24. Steinfeldt T, Finsterbusch T, Mankertz A (2006) Demonstration of nicking/joining activity at the origin of DNA replication associated with the rep and rep' proteins of porcine circovirus type 1. J Virol 80:6225–6234. https://doi.org/10.1128/JVI.02506-05
25. Wang H, Dong Y, Hu X et al (2026) Genetic variability and intra-genotype recombination of DuCV from ducks and geese in central and north China. Front Vet Sci 13:1775879. https://doi.org/10.3389/fvets.2026.1775879
26. Zhang T, Liu N, Zhang L et al (2023) Research note: Complete genome cloning and genetic evolution analysis of four Cherry Valley duck circovirus strains in China in 2022. Poult Sci. https://doi.org/10.1016/j.psj.2023.102920

Figures

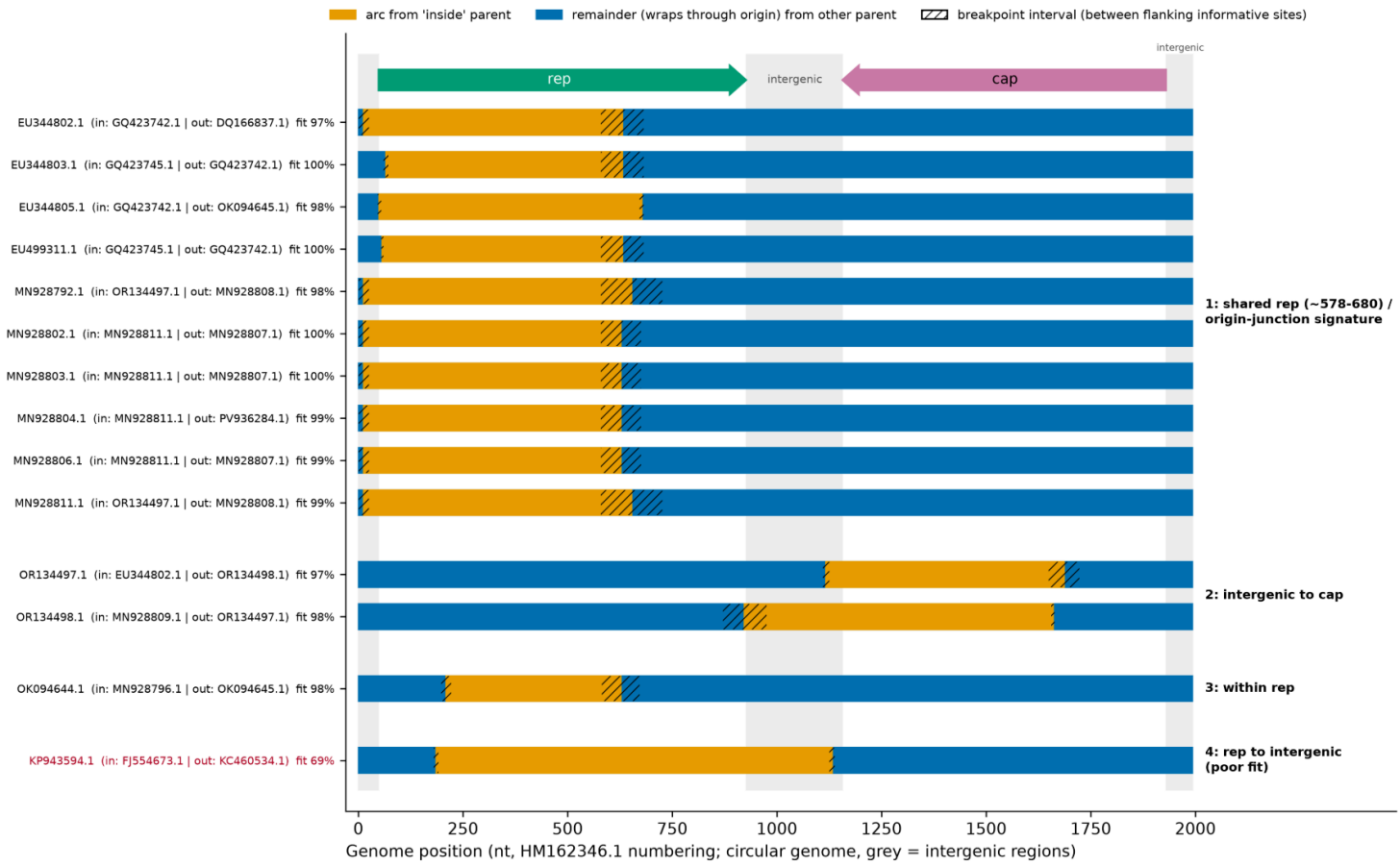


Figure 1

Mosaic structure of the 14 extreme genomes on the circular duck circovirus genome (HM162346.1 numbering). Orange arcs come from the "inside" parent and blue arcs, which wrap through the origin, from the other parent. Hatching marks breakpoint intervals between flanking informative sites. Parents are named per genome, so colours show arc role, not lineage. Red labels mark a mosaic fit below 80% of informative sites. The group 1 label gives approximate bounds (579–682 in Table 1)