RSV G selection analyses support constraint of the CX3C/cystine-noose core and diversification in mucin-like regions

Zahid Ayomide Nassoro-Ally[1,*]

[1]Independent Researcher, Chicago, IL, USA

ORCID: 0009-0002-0550-5115

[*]Correspondence: nassorozahid@gmail.com

Running title: RSV G CX3C core constraint

Keywords: respiratory syncytial virus; G glycoprotein; central conserved region; CX3C motif; purifying selection; d_N/d_S; molecular evolution; FEL; MEME

Abstract

The RSV attachment (G) glycoprotein is a highly variable surface antigen and an important target of humoral immunity, yet it contains a short central conserved region (CCR; RSV-A2 residues 157–198) with a CX3C chemokine-mimic motif and disulfide-bonded cystine noose. A pre-specified hypothesis was tested: FEL-positive diversifying sites are enriched in the CCR and/or the CX3C motif. Using a uniform align → maximum-likelihood phylogeny → site-wise selection pipeline based on FEL, validated against influenza H3N2 haemagglutinin (positive selection significantly enriched in antigenic sites A–E; $p = 0.0081$), the FEL enrichment hypothesis was not supported: no mappable RSV G FEL-positive sites fell in the CCR or CX3C motif in RSV-A, RSV-B, or a descriptive pooled summary. MEME, used as an episodic-selection companion analysis, identified one CCR site in RSV-A and one in RSV-B, but zero CX3C sites in either subtype; most RSV MEME-positive sites were in the mucin-like C-terminal domain or ON1/unmappable sequence. Because the FEL site-count test has low power with only 3–6 significant sites, a threshold-free companion analysis compared the per-codon $d_N - d_S$ distribution across regions. Region-level distributions support stronger CCR purifying constraint in RSV-A (Kruskal–Wallis $p = 0.0275$; CCR vs. mucin Mann–Whitney $p = 0.0066$, rank-biserial $= -0.26$, small-to-moderate shift) and a concordant but non-significant trend in RSV-B ($p = 0.074$). Mapping per-codon $d_N - d_S$ onto the ordered CX3C/cystine-noose core in PDB 5WN9 is consistent with purifying selection over the functional motif. These evolutionary data support, but do not by themselves establish, the rationale for monitoring and targeting the CX3C/cystine-noose core while recognizing sparse episodic-selection signals in the broader CCR.

1 Introduction

Respiratory syncytial virus (RSV) is a leading cause of lower-respiratory-tract infection in infants, older adults, and immunocompromised individuals. Two antigenic subgroups, RSV-A and RSV-B, co-circulate and are distinguished largely by their attachment (G) glycoprotein [10, 11]. G is a heavily O-glycosylated type-II membrane protein whose ectodomain is dominated by two mucin-like, hypervariable regions flanking a short central conserved region (CCR). The CCR (residues approximately 157–198 in the RSV-A2 reference, GenBank M11486) is the least variable segment of an otherwise highly variable protein and contains two functionally critical features: a CX3C chemokine motif (residues 182–186, Cys-X-X-X-Cys) that mimics the chemokine fractalkine (CX3CL1) and engages its receptor CX3CR1, and a cystine noose formed by two nested disulfide bonds that structurally locks the motif [8, 14, 11].

Here, CCR denotes RSV-A2 G residues 157–198, CX3C denotes residues 182–186, and the structural core denotes the ordered CX3C/cystine-noose segment visualized from PDB 5WN9 (residues 169–189). The term central conserved domain (CCD) is used only where it matches the terminology of the structural or vaccine literature.

RSV F remains the leading target of licensed and advanced vaccine and immunoprophylaxis approaches because it is more conserved and contains potent neutralizing epitopes. RSV F and G are both major targets of the humoral immune response, but vaccine and therapeutic development has largely focused on F [9, 10]. G is more variable, yet its CCR is antibody-accessible and has re-emerged as a vaccine and monoclonal-antibody target [10, 13]. This creates a clear, testable evolutionary prediction. If antibody pressure drives adaptive change in G, one of two patterns should hold: either (i) the CCR is a hotspot of episodic positive, or diversifying, selection, as is seen in the antigenic sites of influenza haemagglutinin [15]; or (ii) functional and structural constraint on the CX3C motif and cystine noose overrides immune pressure, confining diversifying selection largely to the mucin-like flanks and leaving the core under purifying selection.

To discriminate between these hypotheses, a uniform molecular-evolution pipeline is used. Selection is quantified at each codon by comparing the non-synonymous (amino-acid-changing) substitution rate with the synonymous (silent) rate: an excess of amino-acid change indicates diversifying selection, whereas a deficit indicates purifying (conserving) selection. Two complementary site-wise tests are applied—FEL, which detects *pervasive* selection acting consistently across the tree, and MEME, which additionally detects *episodic* selection acting on only a subset of lineages. Influenza H3N2 haemagglutinin is used as a positive control because its antibody-binding antigenic sites are already known to be under diversifying selection, so a correct pipeline should recover that signal. The primary hypothesis—enrichment of positively selected sites in the CCR and CX3C motif—and its region definitions were specified before RSV selected-site positions were inspected. The pipeline is first validated on influenza H3N2 haemagglutinin, where the expected answer is known, and is then applied to RSV-A and RSV-B G. Recognising that the site-count test is statistically weak when few sites reach significance, the analysis is complemented with a threshold-free comparison of the selection-pressure distribution across regions, and the result is anchored to the experimental structure of the CX3C core.

2 Methods

2.1 Sequence data

Coding sequences of the RSV G gene were obtained for RSV-A ($n = 108$ sequences) and RSV-B ($n = 115$ sequences); influenza A/H3N2 haemagglutinin (HA) coding sequences were obtained as a positive-control dataset. Accession lists are provided in `notes/rsv_a_nt_accessions.tsv`, `notes/rsv_b_nt_accessions.tsv`, and `notes/flu_h3_ha_nt_accessions.tsv`. The exact RSV database query, download date, and random seed used to create the frozen RSV FASTA inputs could not be reconstructed from the current workspace snapshot; the accession lists and checksums therefore define the reproducible input set for this manuscript. All coordinates for RSV G are reported in RSV-A2 G protein residue numbering, using GenBank M11486 as the 298-aa reference; H3 HA sites are reported in mature-HA1 (H3) numbering.

2.2 Alignment and phylogenetics

Nucleotide coding sequences were codon-aware aligned with MAFFT v7.526 using a translation-guided alignment that was back-translated to a codon alignment [1]. Maximum-likelihood phylogenies were inferred with IQ-TREE v3.1.3 with automated model selection using ModelFinder [2, 3]. The best-fit substitution models were TVM+F+I+R2 for H3N2 HA and TN+F+I+R2 for both RSV-A and RSV-B G. Trees were used unrooted for the selection analysis (Supplementary Figures 5–7).

2.3 Site-wise selection

Per-codon selection was estimated with FEL (Fixed Effects Likelihood) in HyPhy v2.5.100, run in Docker using image `quay.io/biocontainers/hyphy:2.5.100` [4, 5]. FEL estimates a synonymous rate ($\alpha = d_S$) and a non-synonymous rate ($\beta = d_N$) at each codon and tests for pervasive site-wise selection. A site was called positively, or diversifyingly, selected when $\beta > \alpha$ at $p \leq 0.05$, and purifying when $\beta < \alpha$ at $p \leq 0.05$. Site-wise p-values are unadjusted and are used as screening thresholds for region-enrichment and summary analyses; individual selected sites are therefore interpreted cautiously. The per-codon selection metric used throughout is $d_N - d_S$ $(\beta - \alpha)$.

As a companion analysis for episodic diversifying selection, MEME (Mixed Effects Model of Evolution) was run on the same codon alignments and phylogenies [6]. A site was called MEME-positive at an unadjusted $p \leq 0.05$, and the estimated number of branches under selection was retained from the MEME output.

2.4 Pre-specified functional regions

The RSV-A2 G coordinate definitions were specified before RSV selected-site positions were inspected: N-terminal region, residues 1–156; CCR, residues 157–198; CX3C motif, residues 182–186 nested within the CCR; and mucin-like C-terminal domain, residues 199–298. For the RSV-A ON1

genotype, the 24-amino-acid C-terminal duplication has no A2-equivalent residue and was labelled as an insertion [12]. In RSV-B, A2 mapping covered all CCR/CX3C residues but omitted four distal C-terminal A2 residues; analyses involving the mucin tail should therefore be interpreted with this minor mapping loss in mind. H3 HA antigenic sites A–E were taken from the canonical Wiley/Wilson definitions [15].

2.5 Primary test: enrichment of selected sites

FEL-significant site positions were mapped to reference coordinates via the reference sequence added to each protein alignment with MAFFT `-add -keeplength`. Enrichment of positively selected sites within a target region was assessed by a one-sided permutation test (10,000 permutations, seed 42): the observed fraction of selected sites falling in the region was compared to a null distribution generated by placing the same number of sites at random over the mapped protein length. This one-sided test evaluates enrichment only, not formal depletion. Tests were run for the CCR and the CX3C motif in RSV-A and RSV-B. As a descriptive pooled summary, the six A2-mappable FEL-positive sites from the two subtype-specific analyses were also combined; this is not a joint evolutionary model.

2.6 Threshold-free companion analysis: region-level selection distribution

Because the primary test has low power when few sites are significant, the full per-codon $d_N - d_S$ distribution was compared across the pre-specified regions using all mapped codons. This threshold-free analysis uses approximately 300 codons per subtype and treats per-codon FEL estimates as comparable site-level summaries; it does not propagate uncertainty in individual rate estimates. An omnibus Kruskal–Wallis test across the N-terminal, CCR, and mucin regions was followed by directional Mann–Whitney U tests specified a priori: CCR more purifying than mucin; CCR more purifying than the rest of the protein; and CX3C more purifying than mucin. Each test is reported with a rank-biserial effect size ($|r_{\mathrm{bc}}|$: 0.1 small, 0.3 medium, 0.5 large). A secondary Fisher exact test compared the fraction of significantly purifying codons between CCR and mucin. Nonparametric tests were chosen a priori given the heavy-tailed, non-normal $d_N - d_S$ distributions.

2.7 Structural mapping

The experimental structure of the RSV-A2 G CCD peptide (PDB 5WN9, ordered chain-A coordinate residues 169–189, bound to scFv 2D10) was retrieved from the RCSB PDB. Identity and residue range were verified programmatically: chain A source organism *Human respiratory syncytial virus A2*, ordered coordinate sequence `NFVPCSICSNNPTCWAICKRI`; disulfides 173–186 and 176–182 were confirmed from coordinates at S–S $\approx$ 2.0 Å. The deposited peptide includes additional flanking residues; the colouring here uses the ordered coordinate residues available for this core segment. Per-codon $d_N - d_S$ values from the RSV-A FEL analysis were written into the B-factor column and rendered with open-source PyMOL v3.1.0 using `code/render_ccd_pymol.pml`, with a Matplotlib-composited colorbar and annotation using `code/compose_structure_figure.py` [16].

2.8 Software and reproducibility

Analyses used Python 3.11 with Biopython 1.87, NumPy, SciPy, and Matplotlib [7]. Summary CSVs, available HyPhy JSON outputs, figures, and SHA256 checksums are in the project repository. Every table is regenerated from stable files in `results/`; no values are hand-entered.

3 Results

3.1 The pipeline recovers known positive selection in influenza HA

Applied to influenza H3N2 HA, the FEL output contained six positively selected alignment codons, of which five were mappable to mature-HA1 antigenic-site coordinates. All five mappable sites (100%) fall in antigenic sites A–E (observed 100% vs. 39.4% expected by chance; enrichment 2.5×; permutation $p = 0.0081$; Figure 1; Table 1). These sites map to HA1 antigenic sites C (residue 50), A (135), B (157), B (193), and E (261). This supports that the align → tree → FEL → enrichment workflow can detect immune-driven diversifying selection where it is known to occur.

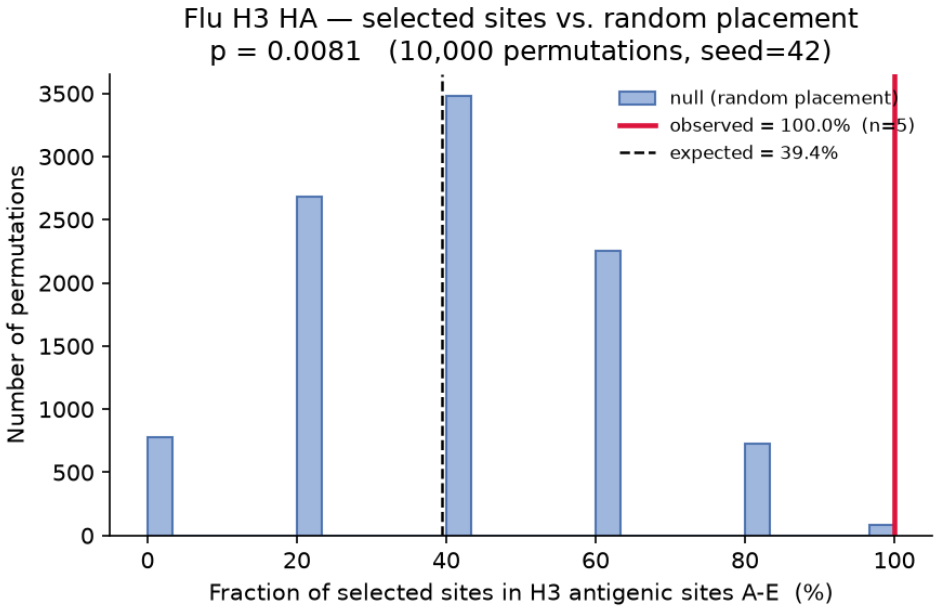


Figure 1: **Pipeline validation on influenza H3N2 HA.** Permutation-null distribution of the fraction of positively selected HA sites expected in antigenic sites A–E. The observed fraction, shown by the red line, is 100% and lies far in the upper tail ($p = 0.0081$).

3.2 FEL-positive selection is not enriched in the RSV G CCR or CX3C motif

In contrast to influenza, no FEL-positive site fell within the RSV G CCR or CX3C motif in either subtype or in the descriptive pooled summary (Table 1; Figure 2). RSV-A had 0/3 mappable FEL-

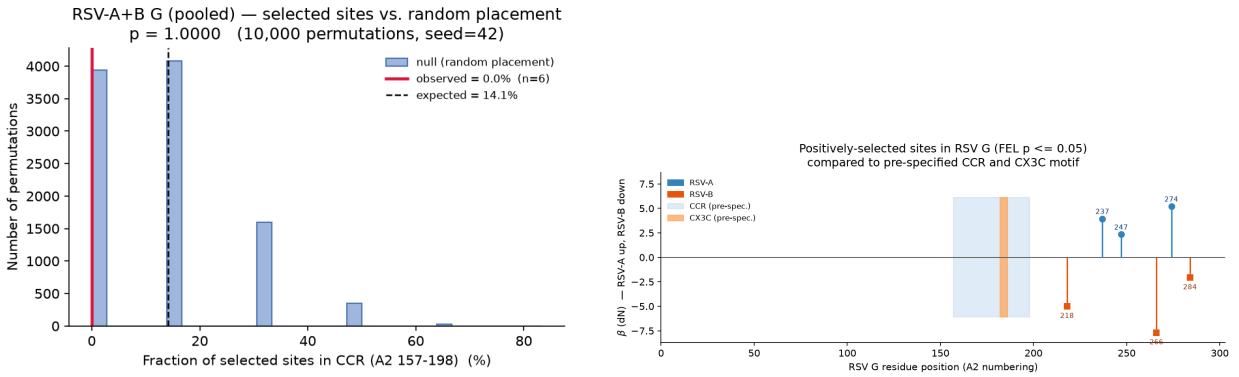
positive sites in the CCR (observed 0% vs. 13.9% expected; $p = 1.0$) and 0/3 in the CX3C motif ($p = 1.0$). RSV-B likewise had 0/3 in the CCR ($p = 1.0$) and 0/3 in the CX3C motif ($p = 1.0$); the per-subtype permutation-null distributions are shown in Supplementary Figure 8. In the descriptive pooled data, 0/6 FEL-positive sites fell in the CCR (0% vs. 14.1% expected; $p = 1.0$) and 0/6 fell in the CX3C motif ($p = 1.0$; pooled CX3C null in Supplementary Figure 9). The pre-specified FEL enrichment hypothesis was not supported: no mappable FEL-positive RSV G sites fell in the CCR or CX3C motif. Because this site-count test has low power with only 3–6 significant sites, this result should be interpreted as absence of enrichment evidence, not proof that the region never experiences positive selection. Section 3.5 addresses the result with a threshold-free companion analysis.

Table 1: **FEL-positive site enrichment in pre-specified regions.** Permutation tests used 10,000 permutations, seed 42, and a one-sided enrichment alternative. Source data: `results/summary_enrichment.csv`.

Target	Region	n	Hits	Obs. %	Exp. %	Enrichment	p-value	Sig.
Flu	H3 antigenic sites A–E	5	5	100.0	39.43	2.54×	0.0081	**YES**
RSV-A	CCR 157–198	3	0	0.0	13.9	0×	1.0	no
RSV-A	CX3C 182–186	3	0	0.0	1.58	0×	1.0	no
RSV-B	CCR 157–198	3	0	0.0	13.9	0×	1.0	no
RSV-B	CX3C 182–186	3	0	0.0	1.58	0×	1.0	no
RSV-A+B	CCR 157–198	6	0	0.0	14.1	0×	1.0	no
RSV-A+B	CX3C 182–186	6	0	0.0	1.66	0×	1.0	no

3.3 FEL-positive RSV G sites localize to the mucin-like C-terminal domain

Every mappable FEL-positive site—six across both subtypes (RSV-A2 positions 237, 247, and 274 from RSV-A; 218, 266, and 284 from RSV-B)—lies in the C-terminal mucin-like hypervariable domain (residues 199–298), i.e., 6/6 versus approximately 34% expected (Figure 2; the complete FEL-positive site list with coordinate mapping is given in Supplementary Table S1). An additional strongly FEL-positive RSV-A site ($\beta = 8.9$, $p = 0.009$) falls within the ON1-genotype C-terminal duplication itself, indicating that the duplicated segment is also a locus of diversification. Because the mucin-domain boundary was examined after observing site positions, this localization is reported as exploratory and post hoc.



(a) Pooled CCR enrichment null.

(b) Selected-site lollipop plot.

Figure 2: **FEL-positive sites in RSV G are not enriched in the conserved core.** The descriptive pooled CCR permutation-null distribution shows observed enrichment of zero ($p = 1.0$). The lollipop plot places every FEL-positive site along the RSV-A2 G coordinate axis, with the CCR and CX3C motif shaded; all mappable sites fall in the C-terminal mucin-like domain.

3.4 MEME detects sparse episodic selection in the CCR but not the CX3C motif

MEME identified more RSV G positive sites than FEL, consistent with its sensitivity to episodic branch-specific selection (Table 2; Supplementary Table S2). In RSV-A, MEME found 8 positive sites, 6 of which were A2-mappable: 1 in the CCR, 0 in the CX3C motif, 4 in the mucin-like C-terminal domain, and 1 in another mapped region. Two additional RSV-A MEME sites were ON1 insertion or otherwise unmappable sites. In RSV-B, MEME found 7 positive sites, all A2-mappable: 1 in the CCR, 0 in the CX3C motif, 5 in the mucin-like C-terminal domain, and 1 in another mapped region. Thus MEME softens the strict FEL-only statement for the broader CCR, but the CX3C motif remains free of positive-site calls in both subtypes, and most RSV MEME signal is still concentrated in the mucin-like/ON1 portion of G.

Table 2: **FEL and MEME positive-site counts by region.** Counts are based on positive sites at unadjusted $p \leq 0.05$ with RSV coordinates mapped to the RSV-A2 G reference where possible. Source data: `results/meme/summary_meme_region_counts.csv`.

Dataset	Method	n positive	n A2-mappable	CCR	CX3C	Mucin	ON1/unmappable	Other mapped	H3 antigenic
RSV-A	FEL	4	3	0	0	3	1	0	–
RSV-A	MEME	8	6	1	0	4	2	1	–
RSV-B	FEL	3	3	0	0	3	0	0	–
RSV-B	MEME	7	7	1	0	5	0	1	–
H3N2 HA	FEL	6	–	–	–	–	–	–	5
H3N2 HA	MEME	11	–	–	–	–	–	–	6

3.5 The CCR is under stronger purifying selection than the mucin domain

The threshold-free, region-level analysis uses all codons rather than only the handful that reach FEL or MEME significance. This companion analysis shows that the three regions differ in RSV-A $d_N - d_S$ (Kruskal–Wallis $H = 7.19$, $p = 0.0275$; Figure 3; Table 3). The CCR is significantly more purifying than the mucin domain (Mann–Whitney $p = 0.0066$, rank-biserial $= -0.26$, consistent with a small-to-moderate shift) and more purifying than the rest of the protein ($p = 0.0094$, $r_{\mathrm{bc}} = -0.22$). In RSV-B, the same directional pattern holds but does not reach significance (omnibus $p = 0.25$; CCR vs. mucin $p = 0.074$, $r_{\mathrm{bc}} = -0.15$; CCR vs. rest $p = 0.056$). Thus, despite sparse MEME-positive CCR sites, the conserved core is overall constrained by purifying selection relative to the flanks, significantly so in RSV-A and with RSV-B showing a concordant but weaker trend.

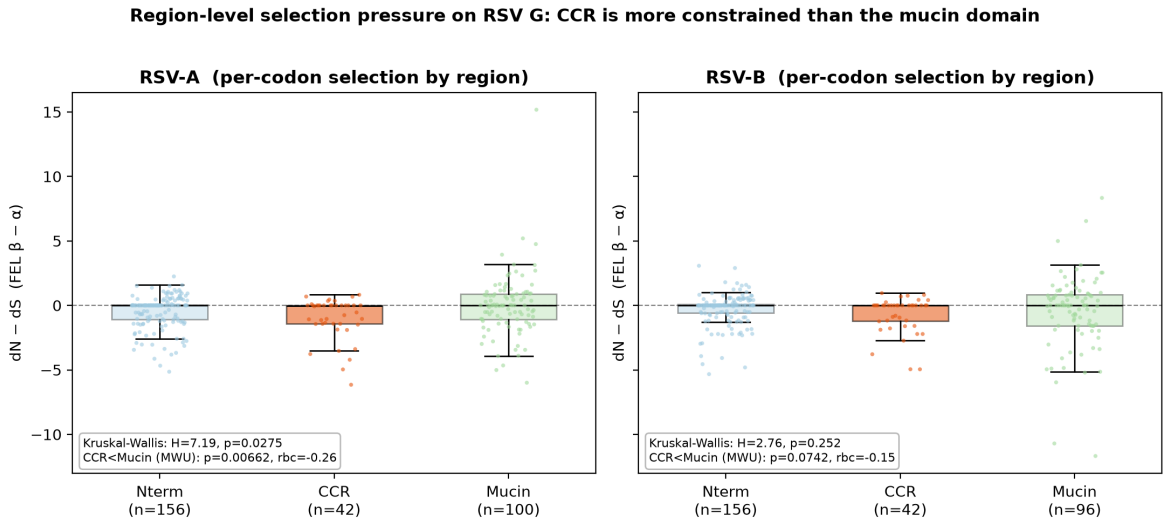


Figure 3: **Region-level selection pressure.** Per-codon $d_N - d_S$ distributions by region (N-terminal, CCR, and mucin) for RSV-A and RSV-B, with Kruskal–Wallis and CCR-vs.-mucin Mann–Whitney statistics annotated. The CCR distribution is shifted toward purifying selection, significantly so in RSV-A.

3.6 The purifying signature maps onto the CX3C/cystine-noose structure

Colouring the ordered experimental CCD coordinates (PDB 5WN9, residues 169–189) by per-codon $d_N - d_S$ shows the CX3C motif and cystine noose sitting in a predominantly purifying core, with only two mildly non-negative residues (Figure 4). The two nested disulfides (173–186 and 176–182) that lock the fractalkine-mimic motif fall entirely within this constrained segment, providing a structural rationale for the constraint quantified in Section 3.5.

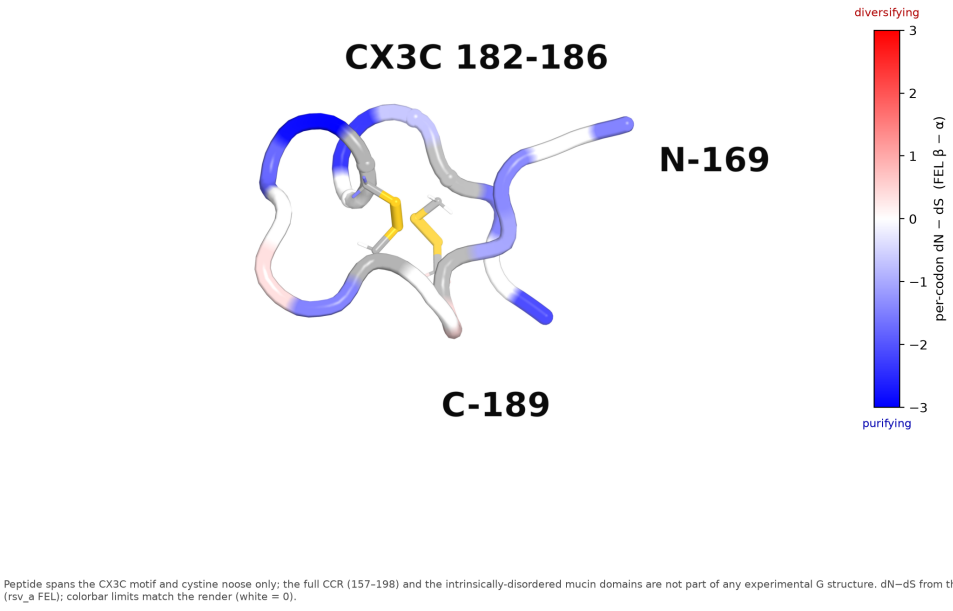


Figure 4: **The CX3C/cystine-noose core is predominantly purifying.** Ray-traced structure of the ordered RSV-A2 G CCD coordinates from PDB 5WN9 (residues 169–189) with the backbone coloured by per-codon $d_N - d_S$ (blue = purifying, red = diversifying). The two nested disulfides (173–186 and 176–182) are shown in gold and the CX3C motif is labelled.

Table 3: **Region-level $d_N - d_S$ comparison across RSV G regions.** Per-codon $d_N - d_S$ was compared across regions as a threshold-free companion analysis. Negative rank-biserial values indicate stronger purifying selection in the CCR. Source data: `results/summary_region_selection.csv`.

Target	Comparison	Test	Statistic	Effect size	p-value	Sig.
RSV-A	N-term/CCR/mucin	Kruskal–Wallis	$H = 7.19$	$\epsilon^2 = 0.018$	0.0275	**YES**
RSV-A	CCR < mucin	Mann–Whitney U	$U = 1546$	$r_{\mathrm{bc}} = -0.264$	0.0066	**YES**
RSV-A	CCR < rest	Mann–Whitney U	$U = 4172$	$r_{\mathrm{bc}} = -0.224$	0.0094	**YES**
RSV-B	N-term/CCR/mucin	Kruskal–Wallis	$H = 2.76$	$\epsilon^2 = 0.003$	0.252	no
RSV-B	CCR < mucin	Mann–Whitney U	$U = 1705$	$r_{\mathrm{bc}} = -0.154$	0.0742	no
RSV-B	CCR < rest	Mann–Whitney U	$U = 4494$	$r_{\mathrm{bc}} = -0.151$	0.0561	no

4 Discussion

The central finding is that adaptive diversification of the RSV G glycoprotein is spatially partitioned: the strongest and most consistent selected-site signals are in the mucin-like flanking domains, while the CX3C chemokine-mimic motif and its cystine noose show no FEL- or MEME-positive selected-site calls in these data. The primary, pre-specified FEL enrichment test gives a clear negative result for both the CCR and CX3C motif, but this is an absence of enrichment evidence rather than proof of absence. MEME detects sparse episodic selection in the broader CCR, but still detects no CX3C sites and places most RSV sites in the mucin-like or ON1/unmappable portions of G. The threshold-free companion analysis supports this spatial pattern: the CCR is under purifying constraint overall in RSV-A, with a concordant but non-significant trend in RSV-B, so it is unlikely to behave as a broad hotspot of positive selection in the sampled data.

This pattern is mechanistically coherent. The CX3C motif mimics fractalkine and engages CX3CR1, a function that plausibly imposes strong purifying selection on the motif and the disulfide scaffold that presents it [8, 14, 11]. A diversifying substitution here would risk abolishing receptor engagement and disrupting the noose. By contrast, substitutions and indels may be more readily tolerated in the mucin-like domains, which are long, O-glycosylated, structurally flexible, and include duplication-prone genotype-defining segments such as ON1 [12, 11]. The influenza HA positive control demonstrates that the pipeline detects immune-driven diversifying selection where a known antigenic region is not equivalently constrained.

The subtype asymmetry—significant constraint in RSV-A and a concordant trend in RSV-B—should be stated plainly rather than glossed over. It may reflect genuinely weaker constraint, a smaller effective sample of informative substitutions in the RSV-B CCR, or the minor mapping loss described below. This subtype difference is not over-interpreted.

Together with structural and antibody studies of the RSV G CCD, these evolutionary results support the rationale for monitoring and targeting the CX3C/cystine-noose core, while recognizing that vaccine or therapeutic efficacy must be established experimentally [14, 10, 13]. Surveillance for antigenic change in G, conversely, should focus on the mucin-like domains while still tracking the rare CCR sites detected by MEME.

5 Limitations

1. **Power of the primary test.** With only 3–6 FEL-significant sites, the pre-specified site-count enrichment test has limited power; a null result from it alone would be weak. This is why the region-level distribution analysis, which uses all codons and no significance threshold, is treated as the primary supporting analysis.

2. **Subtype dependence.** The purifying-constraint result is statistically clear only for RSV-A; RSV-B shows the same direction as a non-significant trend.

3. **Cross-subtype coordinate mapping.** RSV-B G was mapped onto RSV-A2 numbering because the CCR is conserved across subtypes. Adding the A2 reference via MAFFT `-keeplength` dropped four C-terminal A2 residues in the RSV-B alignment (`a2_ref_len` 294

vs. 298), affecting 4 of approximately 96 mucin codons; this does not alter the RSV-B conclusion but is disclosed for completeness.

4. **Structural coverage.** No full-length or mucin-domain G structure exists; the crystallized peptide (PDB 5WN9) spans only the CX3C/cystine-noose core (169–189). The structural figure therefore illustrates constraint on the functional core, not the entire CCR or the flanks.

5. **Post-hoc localization.** The concentration of selected sites in the mucin domain used a boundary chosen after observing site positions and is reported as exploratory.

6. **Sequence sampling metadata.** The current workspace preserves accession lists and checksums but not the exact original RSV database query, download date, or random seed used to create the frozen input FASTA files.

7. **Site-wise multiplicity.** FEL and MEME calls use unadjusted $p \leq 0.05$ thresholds as screening-level selected-site definitions. False-discovery control or threshold-sensitivity analyses would be useful before treating individual sites as definitive.

8. **Distributional-analysis assumptions.** The threshold-free region analysis treats per-codon FEL point estimates as comparable site-level summaries and does not propagate uncertainty in those estimates or account for residual dependence among codons.

9. **Recombination and tree uncertainty.** No formal recombination screen or tree-uncertainty sensitivity analysis was performed; alignment, recombination, or temporal clustering artifacts could affect site-wise selection calls.

10. **Method scope.** FEL detects pervasive site-wise selection, while MEME detects episodic site-wise selection; branch-level or lineage-restricted selection, for example with aBSREL, was not modelled and could refine the picture.

6 Conclusion

Diversifying-selection signals on the RSV G glycoprotein are strongest in the mucin-like flanks and ON1 insertion or other unmappable sequence, with sparse MEME-positive sites in the broader CCR but none in the CX3C motif. The CCR is nevertheless under purifying constraint overall: significant in RSV-A and trending in RSV-B. The CX3C chemokine-mimic motif and its cystine noose lie within this constrained, purifying core. These results support the rationale for continued monitoring and experimental evaluation of the RSV G conserved region, especially the CX3C/cystine-noose motif, while antigenic surveillance should prioritize the variable mucin-like domains.

Data and code availability

The project repository at `https://github.com/zahid-bio/rsv-g-selection-analysis` contains the manuscript source, analysis scripts, accession lists, archived input sequence files, codon and protein alignments, maximum-likelihood trees and IQ-TREE reports, FEL and MEME outputs, summary CSV files, publication figures, and SHA256 checksums required to reproduce the reported

analyses from the archived input files forward. Structure coordinates are from the RCSB PDB, accession 5WN9 [14]. The RSV FASTA and accession files are treated as frozen inputs because the exact original RSV Entrez query text and random seed were not captured; the RSV download step is therefore not byte-for-byte reproducible, but the analysis from the archived FASTA files forward is reproducible from repository files. Analysis code is released under the MIT License, and the manuscript text, figures, tables, and derived data are released under the Creative Commons Attribution 4.0 (CC BY 4.0) License; third-party structure coordinates (PDB 5WN9) remain under their original RCSB PDB terms.

Author contributions

I conceived the study, performed the analyses, interpreted the results, prepared the figures and tables, and wrote the manuscript.

Funding

No specific funding was received.

Acknowledgements

None.

Conflicts of interest

No competing interests are declared.

Supplementary tables

Supplementary Table S1: **FEL-positive sites with reference-coordinate mapping.** Source data: `results/summary_selected_sites.csv`.

Target	Aln. pos.	Ref. residue	Reference	Region/site	β	p-value
flu (H3N2)	66	50	H3 mature HA1	Site C	2.3248	0.03510
flu (H3N2)	161	135	H3 mature HA1	Site A	4.2112	0.00193
flu (H3N2)	183	157	H3 mature HA1	Site B	4.1714	0.04621
flu (H3N2)	219	193	H3 mature HA1	Site B	2.5898	0.01495
flu (H3N2)	287	261	H3 mature HA1	Site E	4.3934	0.00167
RSV-A	237	237	RSV-A2 (M11486)	outside	3.9430	0.03408
RSV-A	247	247	RSV-A2 (M11486)	outside	2.3345	0.00385
RSV-A	274	n/a (ON1 insertion)	RSV-A2 (M11486)	–	8.9442	0.00876
RSV-A	298	274	RSV-A2 (M11486)	outside	5.2034	0.00260
RSV-B	236	218	RSV-A2 (M11486)	outside	5.0000	0.00205
RSV-B	304	266	RSV-A2 (M11486)	outside	7.6912	0.01965
RSV-B	322	284	RSV-A2 (M11486)	outside	2.0747	0.02770

References

[1] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*. 2013;30(4):772–780. doi:10.1093/molbev/mst010.

[2] Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Molecular Biology and Evolution*. 2020;37(5):1530–1534. doi:10.1093/molbev/msaa015.

[3] Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*. 2017;14(6):587–589. doi:10.1038/nmeth.4285.

[4] Kosakovsky Pond SL, Frost SDW. Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Molecular Biology and Evolution*. 2005;22(5):1208–1222. doi:10.1093/molbev/msi105.

[5] Kosakovsky Pond SL, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. *Bioinformatics*. 2005;21(5):676–679. doi:10.1093/bioinformatics/bti079.

[6] Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. Detecting individual sites subject to episodic diversifying selection. *PLoS Genetics*. 2012;8(7):e1002764. doi:10.1371/journal.pgen.1002764.

[7] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJL. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. 2009;25(11):1422–1423. doi:10.1093/bioinformatics/btp163.

[8] Tripp RA, Jones LP, Haynes LM, Zheng H, Murphy PM, Anderson LJ. CX3C chemokine mimicry by respiratory syncytial virus G glycoprotein. *Nature Immunology*. 2001;2(8):732–738. doi:10.1038/90675.

[9] Tang A, Chen Z, Cox KS, Su H-P, Callahan C, Fridman A, Zhang L, Patel SB, Cejas PJ, Swoyer R, Touch S, Citron MP, Govindarajan D, Luo B, Eddins M, Reid JC, Soisson SM, Galli J, Wang D, Wen Z, Heidecker GJ, Casimiro DR, DiStefano DJ, Vora KA. A potent broadly neutralizing human RSV antibody targets conserved site IV of the fusion glycoprotein. *Nature Communications*. 2019;10:4153. doi:10.1038/s41467-019-12137-1.

[10] Rainho-Tomko JN, Pavot V, Kishko M, et al. Immunogenicity and protective efficacy of RSV G central conserved domain vaccine with a prefusion nanoparticle. *npj Vaccines*. 2022;7:74. doi:10.1038/s41541-022-00487-9.

[11] Anderson LJ, Jadhao SJ, Paden CR, Tong S. Functional features of the respiratory syncytial virus G protein. *Viruses*. 2021;13(7):1214. doi:10.3390/v13071214.

[12] Eshaghi A, Duvvuri VR, Lai R, Nadarajah JT, Li A, Patel SN, Low DE, Gubbay JB. Genetic variability of human respiratory syncytial virus A strains circulating in Ontario: a novel genotype with a 72 nucleotide G gene duplication. *PLoS ONE*. 2012;7(3):e32807. doi:10.1371/journal.pone.0032807.

[13] Lee J, Klenow L, Coyle EM, Golding H, Khurana S. Protective antigenic sites in respiratory syncytial virus G attachment protein outside the central conserved and cysteine noose domains. *PLoS Pathogens*. 2018;14(8):e1007262. doi:10.1371/journal.ppat.1007262.

[14] Fedechkin SO, George NL, Wolff JT, Kauvar LM, DuBois RM. Structures of respiratory syncytial virus G antigen bound to broadly neutralizing antibodies. *Science Immunology*. 2018;3(21):eaar3534. doi:10.1126/sciimmunol.aar3534. PDB 5WN9: 10.2210/pdb5wn9/pdb.

[15] Wiley DC, Wilson IA, Skehel JJ. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. *Nature*. 1981;289(5796):373–378. doi:10.1038/289373a0.

[16] Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 3.1.0. 2010. `https://pymol.org/`.

A Supplementary figures

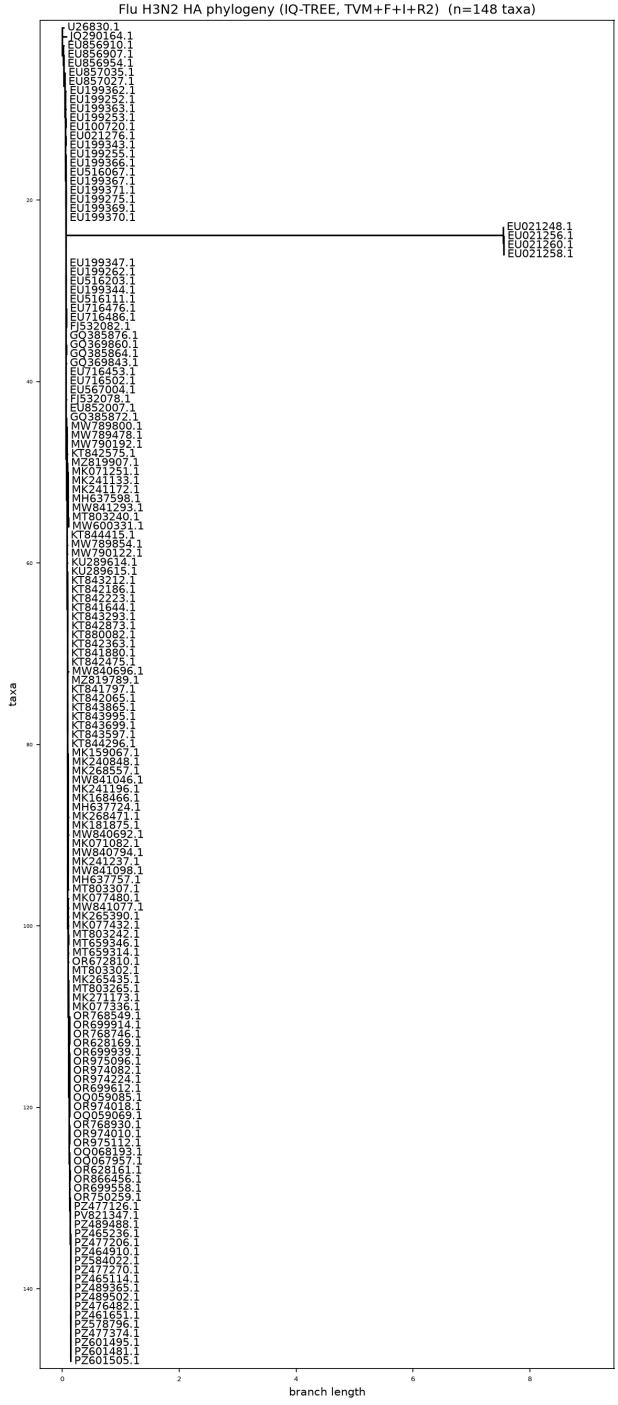


Figure 5: **Supplementary Figure S1.** Maximum-likelihood phylogeny for influenza H3N2 HA.

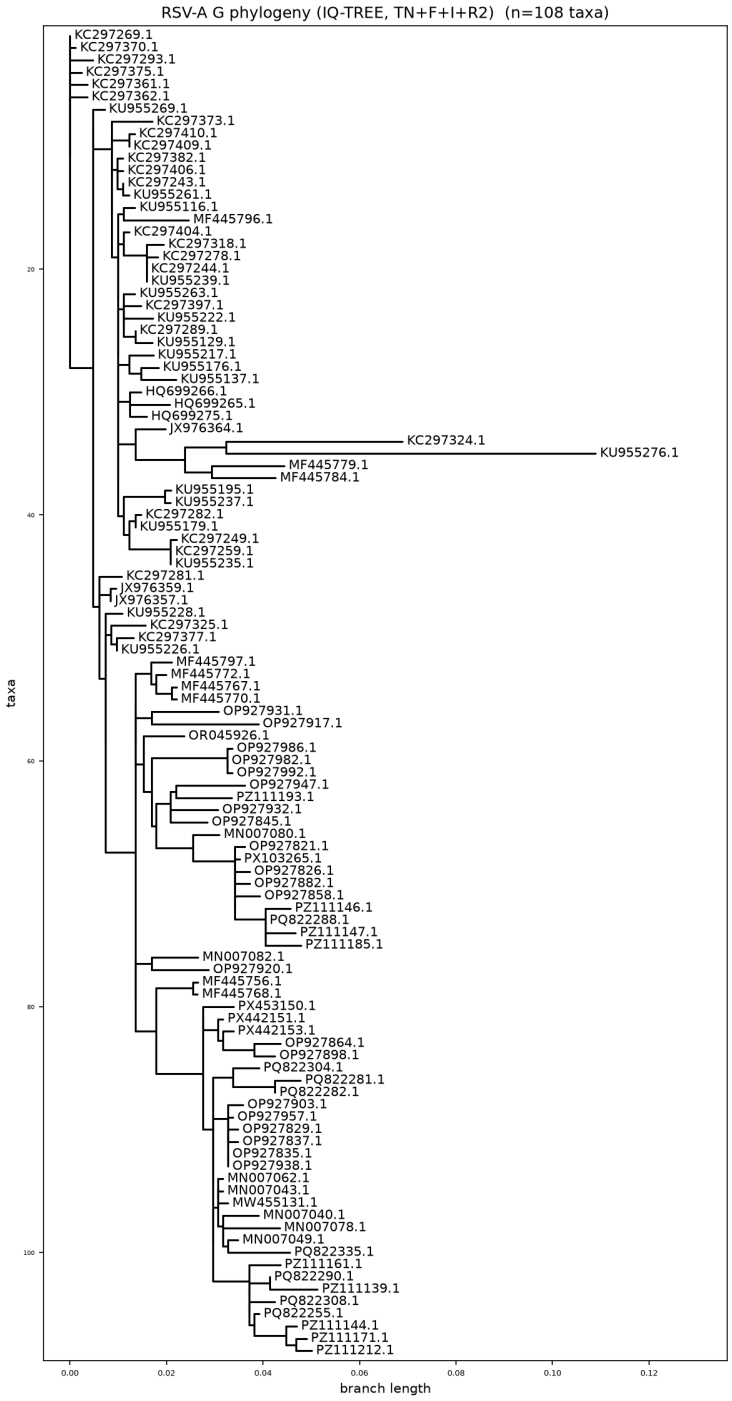


Figure 6: **Supplementary Figure S2.** Maximum-likelihood phylogeny for RSV-A G.

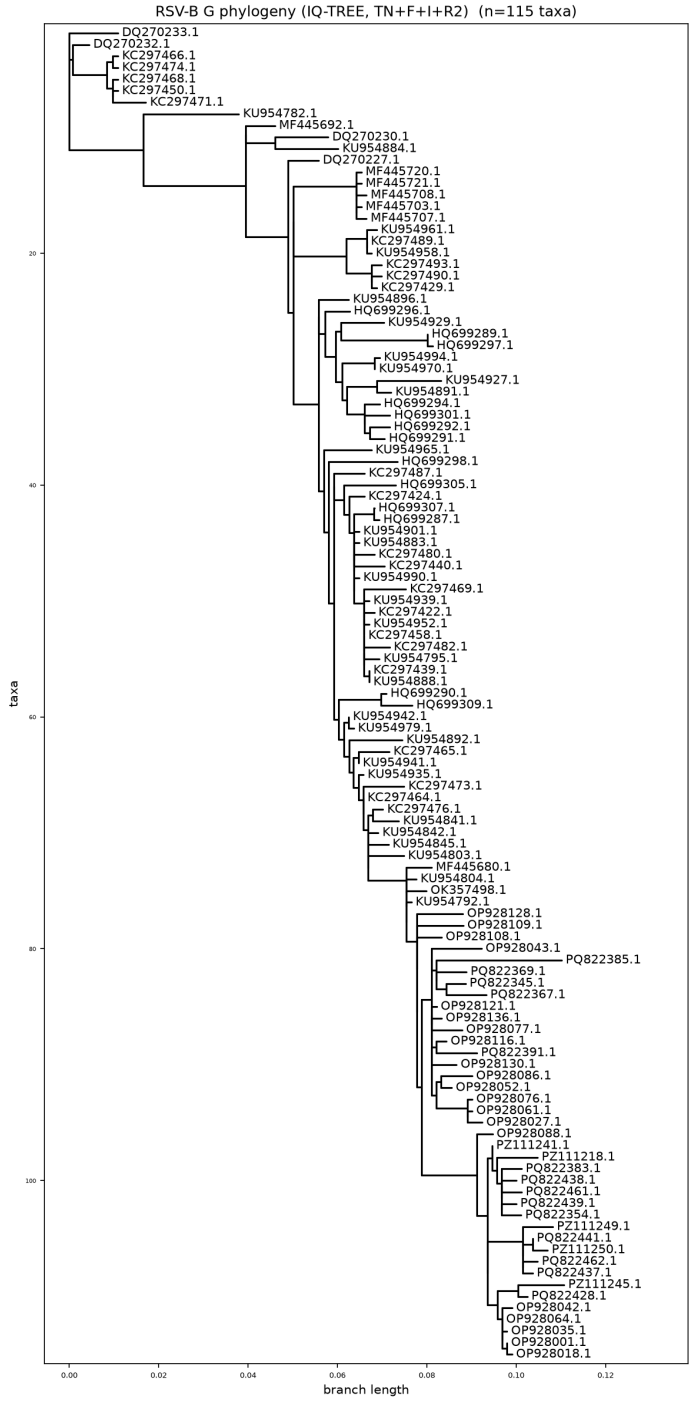
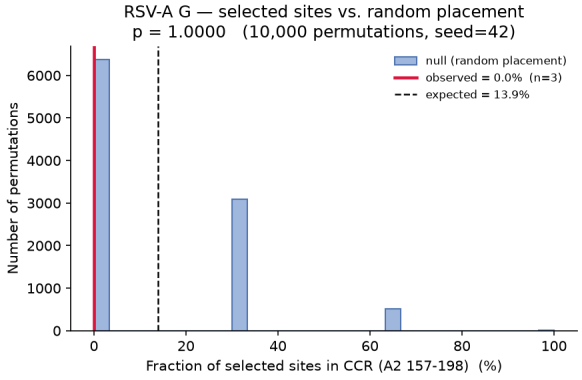


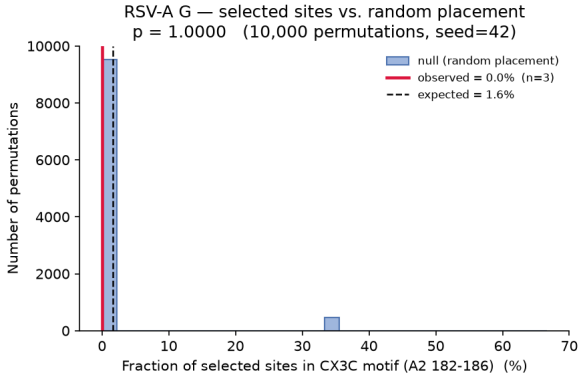
Figure 7: **Supplementary Figure S3.** Maximum-likelihood phylogeny for RSV-B G.

Supplementary Table S2: **FEL-positive and MEME-positive sites with coordinate mapping.** Source data: `results/meme/summary_fel_meme_sites.csv`.

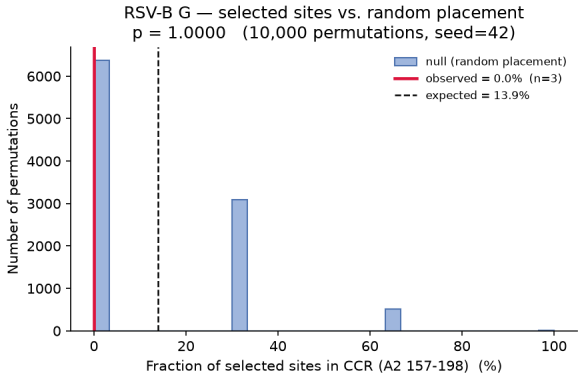
Dataset	Method	Alignment codon	A2/reference residue	Region	A2-mappable	FEL p-value	MEME p-value	MEME branches under selection
RSV-A	MEME	142	142	N-terminal region	yes	–	0.0174	3
RSV-A	MEME	178	178	CCR	yes	–	0.0438	1
RSV-A	FEL+MEME	237	237	C-terminal mucin-like domain	yes	0.0341	0.0487	3
RSV-A	FEL+MEME	247	247	C-terminal mucin-like domain	yes	0.0039	0.0066	6
RSV-A	MEME	267	–	ON1/unmappable insertion	no	–	0.0073	1
RSV-A	FEL+MEME	274	–	ON1/unmappable insertion	no	0.0088	0.0145	5
RSV-A	FEL+MEME	298	274	C-terminal mucin-like domain	yes	0.0026	0.0047	4
RSV-A	MEME	321	297	C-terminal mucin-like domain	yes	–	0.0067	2
RSV-B	MEME	94	77	N-terminal region	yes	–	0.0255	1
RSV-B	MEME	181	164	CCR	yes	–	5.27e-04	1
RSV-B	MEME	235	217	C-terminal mucin-like domain	yes	–	0.0241	2
RSV-B	FEL+MEME	236	218	C-terminal mucin-like domain	yes	0.0021	0.0037	5
RSV-B	FEL+MEME	304	266	C-terminal mucin-like domain	yes	0.0197	0.0295	8
RSV-B	MEME	311	273	C-terminal mucin-like domain	yes	–	0.0396	2
RSV-B	FEL+MEME	322	284	C-terminal mucin-like domain	yes	0.0277	0.0402	5
H3N2 HA	MEME	2	–	outside mature HA1	–	–	0.0334	3
H3N2 HA	FEL	9	–	outside mature HA1	–	0.0459	–	–
H3N2 HA	MEME	16	–	outside mature HA1	–	–	0.0205	2
H3N2 HA	FEL	66	50	H3 antigenic site C	–	0.0351	–	–
H3N2 HA	MEME	150	124	H3 antigenic site A	–	–	0.0016	4
H3N2 HA	MEME	157	131	H3 antigenic site A	–	–	1.23e-05	2
H3N2 HA	FEL+MEME	161	135	H3 antigenic site A	–	0.0019	0.0036	8
H3N2 HA	FEL	183	157	H3 antigenic site B	–	0.0462	–	–
H3N2 HA	MEME	216	190	H3 antigenic site B	–	–	0.0440	2
H3N2 HA	FEL+MEME	219	193	H3 antigenic site B	–	0.0150	0.0239	5
H3N2 HA	MEME	249	223	non-antigenic HA1	–	–	0.0053	4
H3N2 HA	FEL+MEME	287	261	H3 antigenic site E	–	0.0017	0.0031	6
H3N2 HA	MEME	374	–	outside mature HA1	–	–	0.0033	4
H3N2 HA	MEME	557	–	outside mature HA1	–	–	0.0264	4



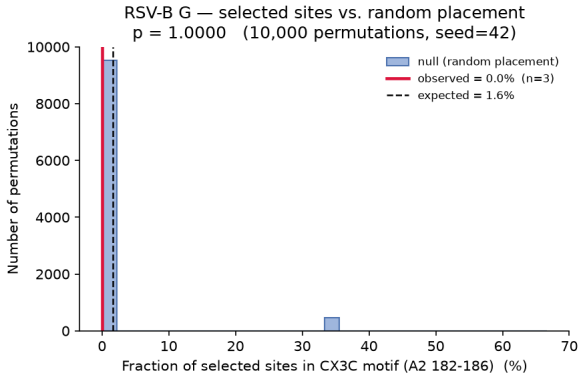
(a) RSV-A CCR.



(b) RSV-A CX3C.



(c) RSV-B CCR.



(d) RSV-B CX3C.

Figure 8: **Supplementary Figure S4.** Per-subtype permutation-null histograms for CCR and CX3C enrichment tests.

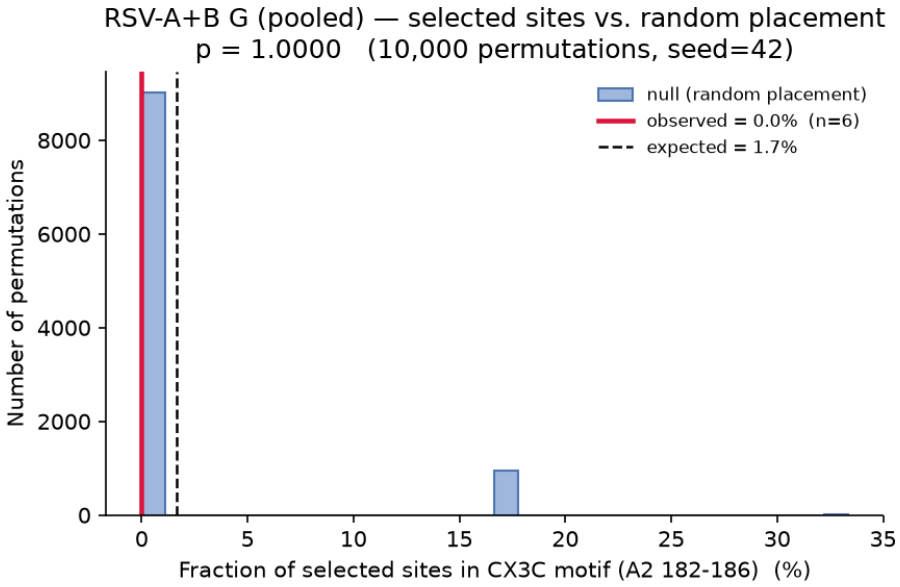


Figure 9: **Supplementary Figure S5.** Pooled RSV-A+B permutation-null histogram for the CX3C enrichment test.