A bioactive phlebovirus-like envelope protein in a hookworm endogenous virus

Monique Merchant[1,2,†], Carlos P. Mata[1,2,†,‡], Yangci Liu[1,2], Haoming Zhai[1,2], Anna V. Protasio[3,4], Yorgo Modis[1,2*]

[1]Molecular Immunity Unit, Department of Medicine, University of Cambridge, MRC Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge, CB2 0QH, United Kingdom

[2]Cambridge Institute of Therapeutic Immunology & Infectious Disease (CITIID), University of Cambridge School of Clinical Medicine, Cambridge CB2 0AW, United Kingdom

[3]Department of Pathology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QP, United Kingdom

[4]Christ's College, University of Cambridge, St Andrew's Street, Cambridge, CB2 3BU, United Kingdom

[†]These authors contributed equally to this work

[‡]Present address: Electron and Confocal Microscopy Unit, National Centre for Microbiology, 28220, Majadahonda, Madrid, Spain

*Corresponding author: ymodis@mrc-lmb.cam.ac.uk

Keywords

Transposable element; retrotransposon; endogenous retrovirus (ERV); envelope protein; membrane fusion; horizontal gene transfer (HGT); virus evolution; gene exaptation; nematode

Short title (50 characters)

A bioactive phlebovirus-like Env in a hookworm EVE

Abstract

Endogenous viral elements (EVEs), accounting for 15% of our genome, serve as a genetic reservoir from which new genes can emerge. Nematode EVEs are particularly diverse and informative of virus evolution. We identify Atlas virus – an intact retrovirus-like EVE in the human hookworm Ancylostoma ceylanicum, with an envelope protein genetically related to GN-GC glycoproteins from phleboviruses. A cryo-EM structure of Atlas GC reveals a class II viral membrane fusion protein fold not previously seen in retroviruses. Atlas GC has the structural hallmarks of an active fusogen. Atlas GC trimers insert into membranes with endosomal lipid compositions and low pH. When expressed on the plasma membrane, Atlas GC has cell-cell fusion activity. RNA-Seq data analysis detected transcripts mapping to Atlas virus at different stages of hookworm development. With its preserved biological activities, Atlas GC has the potential to acquire a cellular function. Our work reveals structural plasticity in reverse-transcribing RNA viruses.

Introduction

Retroviruses and other reverse-transcribing RNA viruses can frequently integrate their genome, reverse-transcribed from RNA into DNA, into the host-cell genome. Viral genomes integrated into germline cells are inherited by future generations as endogenous viral elements (EVEs). Most EVEs are endogenous retroviruses (ERVs) and account for approximately 15% of the human genome, seven times more than protein-coding genes (*1, 2*). EVEs and other transposons were initially viewed as parasitic DNA. It is now evident that they serve as a genetic reservoir, from which new genes and regulatory elements can emerge. Sequences of retroviral origin help control gene expression by serving as promoters, enhancers and other regulatory elements (*3, 4*). Genes coopted from EVEs have evolved to fulfill vital cellular functions (*1*). For example, syncytins, which drive cell-cell fusion of trophoblasts during placental development, are encoded by ERV envelope glycoprotein (*env*) genes (*3, 5*). Another recent example is the

50 Gag capsid protein encoded by the *Caenorhabditis elegans* Cer1 retrotransposon, which
51 encapsidates small non-coding RNAs into nucleocapsids that can transfer RNAs conferring
52 learned pathogen avoidance behavior from infected parents to naïve progeny (*6*).

53 The reduced mutation rate of host versus retrovirus genomes (10^{-9} vs. 10^{-3} mutations
54 per site per year (*7*)) means that EVEs are windows to ancestral retroviral sequences –
55 evolutionary fossils preserved from the time of integration (*2*). Some EVE genes are expressed
56 in human tissues and retain their biological activities, such as membrane fusion activity in the
57 case of Env proteins (Envs) (*3, 5, 8*). Aberrant expression of Envs is associated with disease (*9,*
58 *10*). With the biology of EVEs still largely uncharted, it is likely that many cellular functions of
59 EVEs in health and disease remain undiscovered. Studying EVE genes with novel properties
60 could therefore provide insights on the evolutionary history of reverse-transcribing RNA viruses
61 and identify fundamental principles in host-virus coevolution.

62 Nematode EVEs are particularly diverse and informative of virus evolution. EVEs from
63 the *Belpaoviridae* (BEL/Pao) family (*11*), related to retroviruses and widespread across
64 metazoa, have revealing genetic features in nematodes (*12*). The presence in *C. elegans* EVEs
65 of overlapping ORFs, otherwise unique to complex vertebrate retroviruses, suggests
66 retroviruses originated in early metazoa with a common ancestor resembling belpaoviruses
67 (*13*). Furthermore, nematode endogenous belpaoviruses encode Envs that are genetically
68 unrelated to retrovirus Envs (*14*). Instead of a class I viral membrane fusion protein with a core
69 fold of three bundled α-helices (*15-18*) – a defining feature of modern retroviruses (*19*) –
70 belpaovirus Envs have sequence similarity to G_C (G2) envelope glycoproteins from
71 phleboviruses (family *Phenuiviridae*) (*14*). G_C proteins are class II membrane fusion proteins,
72 with a three-domain β-strand architecture (*20, 21*) also found in alphaviruses (*22*), flaviviruses
73 (*23, 24*) and Rubella virus (*25*) but structurally unrelated to class I fusion proteins. A series of
74 conformational changes in class II fusion proteins, triggered by endosomal acidification,

75 catalyses fusion of the viral and endosomal membranes to deliver the viral genome into the
76 cytosol (*26-28*). A hydrophobic fusion loop first inserts into the endosomal membrane. The
77 proteins then form trimers and fold back on themselves, pulling the cell membrane (held by the
78 fusion loop) and the viral membrane (held by a transmembrane anchor) together so they fuse
79 (*19, 21*).

80 Class II fusion proteins are not limited to viruses: they also drive cell-cell fusion events of
81 fundamental importance, including syncytial epithelia formation in *C. elegans* and other
82 nematodes (*29-32*), and gamete fusion in protozoa, plants, algae and invertebrates (*33-38*). The
83 identical topology and overall arrangement of the three domains of viral and eukaryotic class II
84 fusion proteins, along with similarities in their membrane fusion mechanisms, makes it all but
85 certain they evolved from a common ancestor (*33*). Although the evolutionary origin of the
86 ancestral class II fusion protein remains unknown, the presence of class II fusion proteins in
87 EVEs raises the provocative prospect that a gene transfer from a virus to a cell led to the advent
88 of sexual reproduction (*33, 39*).

89 Here, we identify a novel, intact endogenous belpaovirus in the human hookworm
90 *Ancylostoma ceylanicum* (a parasitic nematode) with an Env more similar than any other
91 eukaryotic sequence to phlebovirus G_C protein sequences. We expressed and purified the G_C-
92 homologous fragment from this EVE, henceforth Atlas virus. A cryo-electron microscopy (cryo-
93 EM) structure of Atlas G_C reveals a class II viral fusion protein fold remarkably similar to
94 phlebovirus G_C proteins and not seen in retroviruses, as predicted 20 years ago (*12, 14*). We
95 show that Atlas G_C has all the hallmarks of an active class II membrane fusion protein. It
96 undergoes a monomer-to-trimer transition and inserts into lipid membranes with a specific lipid
97 composition in response to a low pH trigger. Our work provides biochemical validation for the
98 hypothesis that acquisition of a fusion protein from an infectious virus, as exemplified by Atlas
99 virus, represents a general paradigm for how retrotransposons can become retroviruses (*14*),

100 and indeed how ancestral reverse-transcribing viruses may have originated (*11*). The preserved
101 biological activities of Atlas G_C, including membrane fusion activity, raise the question of
102 whether these activities, and those of EVE gene products more broadly, have cellular functions
103 or cause disease.

104 **Results**

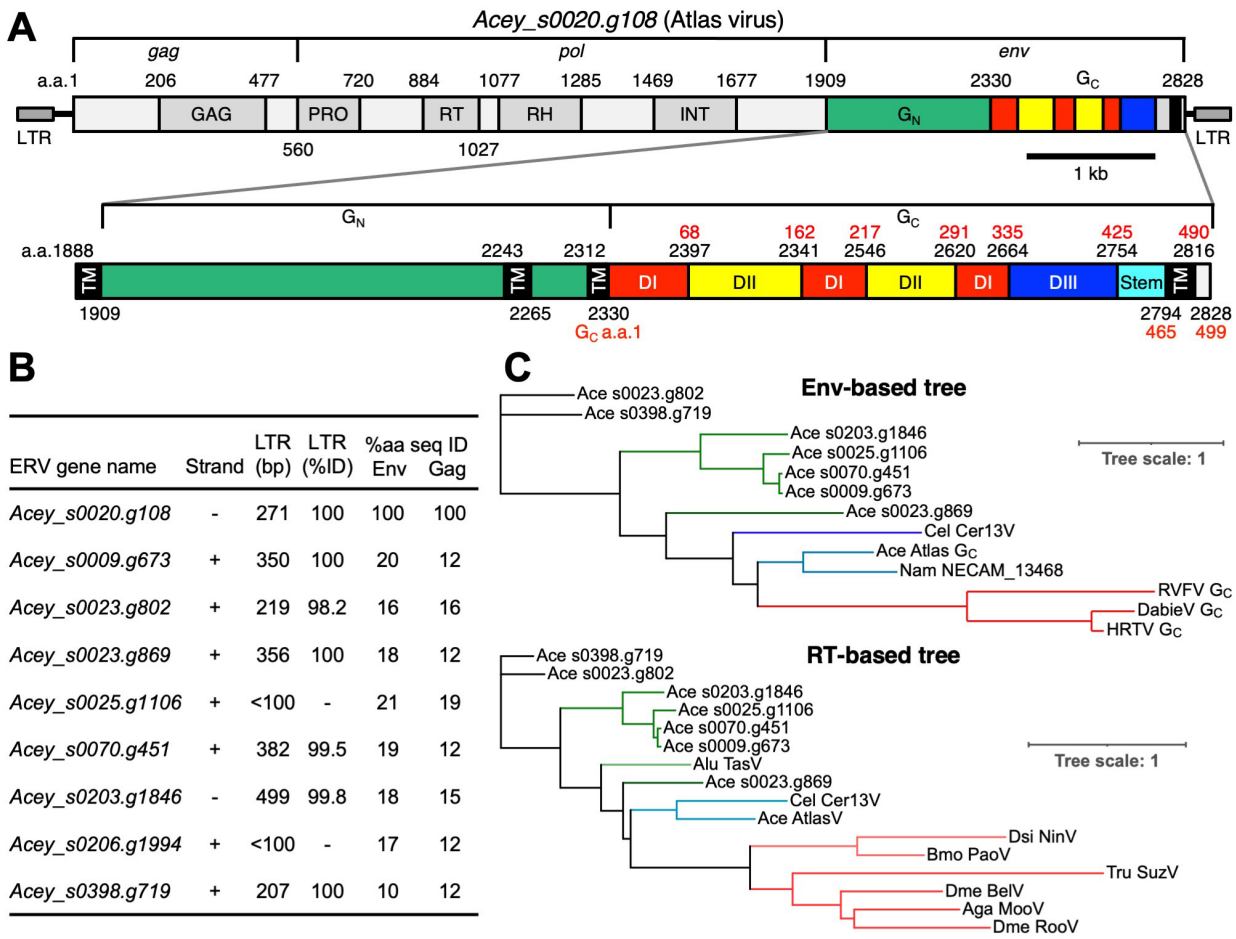
105 **An intact EVE with a phlebovirus-like Env in the hookworm *A. ceylanicum***

106 The bioinformatic discovery of nematode EVEs with phlebovirus G_C-like Env sequences not
107 seen in retroviruses (*12, 14*), requires biochemical validation. To identify phlebovirus-like EVE
108 Envs suitable for biochemical analysis, we performed PSI-BLAST searches (*40*) for protein
109 sequences similar to biochemically characterized phlebovirus G_C proteins. A search with Rift
110 Valley Fever virus (RVFV) G_C as the query identified the gene *Acey_s0020.g108* (UniProt
111 A0A016UZK2) in the human hookworm *A. ceylanicum* as containing the most similar sequence
112 outside infectious virus taxa (Expected value 10^{-20}). The homologous sequence lies within a
113 9,204-nucleotide element possessing all the features of an intact EVE, including 100% identical
114 271-nt LTRs and a coding sequence encoding a single 2,828-residue Gag-Pol-Env polyprotein
115 without any stop codons or introns (**Fig. 1A and fig. S1A**). We refer to this element as *A.*
116 *ceylanicum* Atlas virus. It is one of nine *A. ceylanicum* EVEs that encode complete Gag-Pol-Env
117 polyproteins (**Fig. 1, B and C**). These EVEs have the distinguishing genomic features of
118 belpaoviruses from other nematode species, including an atypical Env and an aspartate to
119 asparagine substitution (Y[X]D<u>D</u> → YVD<u>N</u>) in the most conserved reverse transcriptase motif,
120 the polymerase site (*13, 41, 42*) (**fig. S1B**). Phylogenetic classification of Atlas virus based on
121 reverse-transcriptase (RT) sequences confirms the *Belpaoviridae* phylogeny of Atlas virus, with
122 Cer13 from *C. elegans* (*12*) as the closest neighbor in the RT tree (**Fig. 1D**). With a Gag
123 sequence less than 20% identical to its closest homolog (*Acey_s0020.g1106*), the Atlas EVE is

124 a candidate for classification as a member of the *Belpaoviridae* family (which contains a single
125 genus, *Semotivirus*) (**fig. S1C**) (*43*).

126 The phlebovirus G_C-like sequence spans the last 498 residues of the Atlas virus
127 polyprotein (residues 2330-2828). It contains a single predicted C-terminal transmembrane
128 helix, like phlebovirus G_C proteins. Phleboviruses and other *Phenuiviridae* family members
129 express a glycoprotein precursor that is cleaved by cellular proteases into two envelope
130 glycoproteins, G_N and G_C (or G1 and G2). G_N is the receptor-binding protein required for cellular
131 attachment and G_C is the membrane fusion protein required for cell entry. G_N is highly antigenic
132 and more variable in sequence than G_C. Our analysis of the Atlas virus Env sequence detected
133 a slight but statistically significant similarity in the 421 residues preceding the G_C-like sequence
134 (residues 1909-2329) to G_N glycoproteins from *Phenuiviridae* family viruses (*E*-value $\geq 10^{-7}$).
135 Moreover, the distribution of predicted transmembrane helices and proteolytic cleavage sites in
136 and adjacent to the Atlas G_N- and G_C-like sequences is the same as in phlebovirus
137 glycoproteins (**Fig. 1A**). Together, these sequence features suggest the Atlas virus Env
138 contains tandem phlebovirus-like G_N and G_C glycoproteins instead of a retrovirus-like
139 glycoprotein. With all the features of a recently active EVE and an apparently intact set of
140 phlebovirus-like glycoproteins, Atlas virus is an excellent candidate for biochemical analysis.

141



ERV gene name	Strand	LTR (bp)	LTR (%ID)	%aa seq ID Env	%aa seq ID Gag
Acey_s0020.g108	-	271	100	100	100
Acey_s0009.g673	+	350	100	20	12
Acey_s0023.g802	+	219	98.2	16	16
Acey_s0023.g869	+	356	100	18	12
Acey_s0025.g1106	+	<100	-	21	19
Acey_s0070.g451	+	382	99.5	19	12
Acey_s0203.g1846	-	499	99.8	18	15
Acey_s0206.g1994	+	<100	-	17	12
Acey_s0398.g719	+	207	100	10	12

Fig. 1. Atlas is an endogenous belpaovirus with a phlebovirus-like Env. (**A**) Gene architecture of the Atlas endogenous viral element (EVE), gene *Acey_s0020.g108* from the *Ancylostoma ceylanicum* human hookworm parasite. Inset: annotated closeup of the Atlas Env region encoding phlebovirus G_N- and G_C-like glycoproteins. Red residue number refer to the G_C sequence alone. (**B**) List of *A. ceylanicum* EVEs encoding complete Gag-Pol-Env polyproteins with phlebovirus-like Envs. LTR sequence identities for two of the EVEs could not be calculated: *Acey_s0206.g1994* had degenerate LTRs, and one LTR of *Acey_s0025.g1106* was truncated in the genome assembly. (**C**) Phylogenetic tree of G_C proteins from *Phenuiviridae* and G_C-like (G2-like) sequences from EVEs. Ace, *A.* ceylanicum; Cel, *C. elegans*; Nam, *Necator Americanus*, RVFV, Rift Valley Fever virus; DabieV, Dabie bandavirus (SFTSV); HRTV, Heartland virus. (**D**) Phylogenetic classification of Atlas based on reverse-transcriptase (RT) sequences. Alu TasV,

154 *Ascaris lumbricoides* Tas virus; Dsi NinV, *Drosophila simulans* Ninja virus; Bmo PaoV, *Bombyx*
155 *mori* Pao virus; Tru SuzV, *Takifugu rubripes* Suzu virus; Dme BelV, *D. melanogaster* Bel virus;
156 Aga MooV, *Anopheles gambiae* Moose virus; RooV, Roo virus. Trees drawn with iTOL v6 (*44*).

157 **Atlas G_C has a class II membrane fusion protein fold not seen in retroviruses**

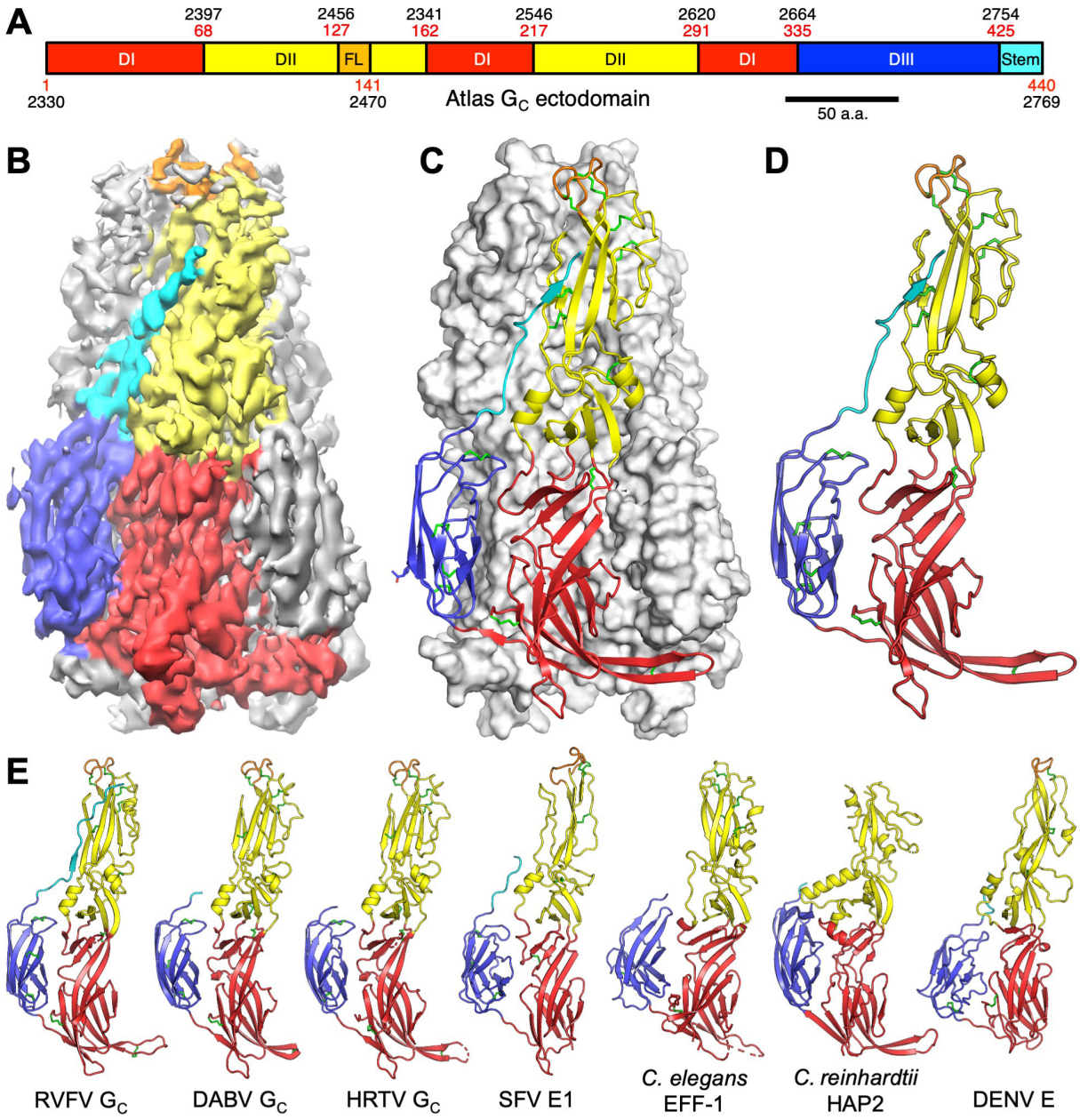
158 As the molecular structure of phlebovirus G_C proteins, and how they drive fusion of the viral and
159 cellular membranes, are well established from previous studies (*20, 28*), we focused our
160 biochemical analyses on the G_C-like sequence from Atlas virus. A recombinant ectodomain
161 fragment of Atlas G_C (polyprotein residues 2330-2772) was expressed in *Drosophila*
162 *melanogaster* D.mel-2 cells as a secreted protein. The purified protein was a soluble, folded
163 homotrimer (**fig. S2**). The structure of the G_C trimer was determined by single-particle cryo-EM
164 image reconstruction at an overall resolution of 3.76 Å (**Fig. 2, A and B; table S1; fig. S3**). The
165 map was sufficiently detailed for an atomic model to be built and refined for Atlas G_C residues
166 2330-2769 using the crystal structure of RVFV G_C (*28*) as a starting model (see Methods; **fig.**
167 **S4**). The structure reveals a three-domain class II membrane fusion protein fold (**Fig. 2C**). All
168 previously described retroviral Env structures have a helical coiled-coil class I fusion protein fold
169 (*15-18*), exemplified by the hemagglutinin of influenza virus (*19*). The structure of the Atlas G_C
170 ectodomain fragment is remarkably similar to phlebovirus G_C structures, specifically the trimeric
171 postfusion G_C structures from RVFV (*28*), Dabie bandavirus (DABV, formerly SFTS phlebovirus)
172 (*45*) and Heartland virus (HRTV) (*46*) (**Fig. 2, D and E**). Domain I, a 10-stranded β-barrel
173 augmented by a 3-stranded sheet, organizes the structure. Two insertions in domain I form the
174 elongated, mostly β-stranded domain II. Domain III has the 7-stranded β-sandwich topology of
175 fibronectin type III (FN3) domains also found in macroglobulin domains (*35, 47*), but the
176 hydrophobic core and disulfide bonding pattern of domain III differ from these and other
177 annotated domains from non-viral species. A 16-amino acid portion of the stem region, which
178 links domain III to the C-terminal transmembrane anchor in class II fusion proteins, could be

179 modeled, spanning 5 nm from the end of domain III to within approximately 1 nm of the tip of
180 domain II. The stem forms trimer contacts, adding a β-strand to domain II of a different subunit,
181 as seen in RVFV $G_{C(28)}$. The overall configuration bears strong similarity to other viral and
182 cellular class II fusion proteins including, in order of decreasing similarity: alphavirus E1 proteins
183 (*22*); EFF-1/AFF-1 cell-cell fusion proteins from *C. elegans* and other animals *(29, 30)*; HAP2
184 gamete fusion proteins from protozoa (*33-35, 48*) and plants (*34, 36*); and flavivirus E proteins
185 (*49*) (**Fig. 2E**). Despite these structural similarities, the only proteins or domains of known
186 structure with detectable amino acid sequence similarity to Atlas G_C (*E*-value <1 in PSI-BLAST
187 (*40*)) are the phlebovirus G_C proteins (22-24% sequence identity; **fig. S5**).

188 Atlas G_C has the same structural features that distinguish phlebovirus glycoproteins from
189 other class II fusion proteins: a larger number of disulfide bonds, ten of which are conserved in
190 phlebovirus G_C sequences but not in other class II proteins; N-linked glycosylation in domain III;
191 and a more extensive and rigid interface between domains I and II (**Fig. 2, D and E, and fig.**
192 **S5**). The most notable differences between Atlas virus and phlebovirus G_C structures are
193 differences in the disulfide bonding pattern and in the composition of side chains lining the
194 glycerophospholipid (GPL) headgroup binding pocket conserved in arboviral class II fusion
195 proteins (*28*). We discuss these differences and their potential functional implications below.
196 Atlas G_C also has a different glycosylation pattern, with a single predicted N-linked glycosylation
197 site at Asn414 in domain III with a weak corresponding feature in the density map (**fig. S6**). In
198 contrast, phlebovirus G_C proteins contain two N-linked glycans in domain III, at two different
199 sites. One of these, Asn1035 in RVFV, covers the fusion loop in the prefusion conformation of
200 RVFV G_C and stabilizes the prefusion dimer by forming contacts across the dimer interface (*20*),
201 as also seen in flavivirus E proteins (*23, 24*). This glycosylation site is conserved in DABV and
202 HRTV G_C but absent in Atlas G_C. Despite these minor differences, the striking overall structural
203 similarity of Atlas G_C to phlebovirus G_C proteins in the postfusion conformation experimentally

204 validates the evolutionary link between nematode EVEs from the *Belpaoviridae* family and the

205 fusion proteins of phleboviruses postulated based on previous genomic analyses (*12, 14*).



206

207 **Fig. 2. Cryo-EM structure of a phlebovirus G_C-like Env fragment from Atlas virus.** (**A**)

208 Domain organization of Atlas G_C. (**B**) Cryo-EM image reconstruction of a soluble ectodomain

209 fragment of Atlas G_C at 3.76 Å overall resolution. The purified protein was a homotrimer (**fig. S2**)

210 and threefold (C3) symmetry was imposed. The map is colored by domain as in (**A**)**.** A

211 representative cryo-EM micrograph is shown in **fig. S4A.** (**C**) Overview of the refined atomic

212 model of the Atlas G_C trimer. Disulfide bonds (green) and an N-linked glycan (blue) are shown
213 as sticks. (**D**) Protein fold of an Atlas G_C protomer. (**E**) The structures most similar to Atlas G_C
214 based on structure comparison with DALI (*50*), G_C glycoproteins from Rift Valley Fever virus
215 (RVFV, Rmsd 2.6 Å, Z-score 29, PDB:6EGU (*28*)), Dabie bandavirus (DABV, formerly SFTS
216 phlebovirus; Rmsd 2.6 Å, Z-score 28, PDB:5G47 (*45*)) and Heartland virus (HRTV, Rmsd 2.6 Å,
217 Z-score 28, PDB:5YOW (*46*)). Other representative class II fusion proteins are shown for
218 comparison: Semliki Forest virus (SFV) E1 (Z-score 19, PDB:1RER (*27*)), *C. elegans* EFF-1 (Z-
219 score 19, PDB:4OJD (*29*)), *Chlamydomonas reinhardtii* HAP2 (Z-score 16; PDB:5MF1 (*33*)) and
220 dengue virus (DENV) E (Z-score 15, PDB:4GSX (*51*)).

221 **Structure of the putative lipid membrane anchor of Atlas G_C**

222 Viral fusion proteins insert a membrane anchor – the fusion loop, in class II proteins – into the
223 host cell membrane to initiate virus-cell membrane fusion. The putative fusion loop of Atlas G_C
224 can be identified by analogy to phlebovirus G_C proteins as spanning residues 127-140. The local
225 resolution of the cryo-EM density for this region is lower than for the rest of the map but the
226 large number of structural constraints imposed by the positions of disulfide-bonded cysteines
227 and other residues conserved in phleboviruses G_C proteins allowed an atomic model to be built
228 unambiguously (**Fig. 3A and fig. S6, B and C**). Specifically, the fusion loop is constrained by
229 four disulfide bonds conserved in phleboviruses, a fifth disulfide specific to Atlas G_C (Cys129-
230 Cys138), a phenylalanine (Phe136) required in phleboviruses at the apex of the fusion loop for
231 membrane binding and fusion (*28, 45, 52*), and two conserved glycines (Gly128 and Gly134)
232 that provide the torsional flexibility necessary for the fusion loop's tightly folded conformation
233 (**Fig. 3, A to C**).

234 The structure and chemical properties of the Atlas G_C fusion loop resemble phlebovirus
235 fusion loops. Ile130, Phe136, Phe137 and the Atlas-specific disulfide (Cys129-Cys138) form a
236 hydrophobic surface similar to phlebovirus and flavivirus fusion proteins (*26, 28*) (**Fig. 3C**). The

237 area of this surface is greater than in most viral class II fusion proteins. By analogy with other
238 viral class II fusion proteins, the location and extent of the hydrophobic surface formed by the
239 Atlas G_C fusion loop suggest it could function as a membrane anchor.

240 In addition to inserting nonpolar side chains into the hydrophobic region of the
241 membrane, viral class II fusion proteins form polar contacts with lipid headgroups via the fusion
242 loop and an adjacent glycerophospholipid (GPL) binding pocket (*28*). By selecting for
243 headgroups with complementary electrostatic potential, polar contacts confer a degree of
244 specificity to lipid binding. In phleboviruses, a set of conserved polar residues in the GPL
245 binding pocket bind selectively to zwitterionic GPLs (*28*). The Atlas G_C structure reveals a
246 putative GPL binding pocket with both conserved and novel features (**Fig. 3**). The arginine that
247 forms bidentate hydrogen bonds with the GPL phosphate moiety in phleboviruses (*28*) is
248 conserved in Atlas virus (Arg86). The disulfide bond and short-chain hydrophobic residue that
249 bind the GPL glycerol moiety are also conserved (Cys87-Cys135, Ala90). However, an
250 aspartate-arginine pair that binds choline and ethanolamine GPL moieties in phleboviruses is
251 replaced in Atlas G_C by two histidines (His271 and His85). Moreover, Atlas virus has an extra
252 residue in the fusion loop, Asp133, compared to phlebovirus G_C proteins. The Asp133 side
253 chain points into the putative GPL binding pocket and is located near the position of the choline
254 GPL moiety in the superimposed structure of RVFV G_C bound to a phosphatidylcholine ligand
255 (*28*), suggesting that Asp133 could compensate for the lack of a conserved aspartate at position
256 271 (**Fig. 3, A to C**). Hence the putative GPL binding pocket of Atlas G_C appears to have the
257 necessary physicochemical attributes to support GPL binding, with Arg86 binding the phosphate
258 moiety, Cys87/Cys135/Ala90 binding the glycerol moiety, and His85/Asp133/His271
259 coordinating the end of the headgroup. We note the presence in the cryo-EM reconstruction of a
260 bulge in the density around the GPL binding pocket that is unaccounted for by the atomic model
261 (**fig. S6C**). Additionally, the absorbance at 260 nm of purified Atlas G_C was higher than

expected despite treatment with nucleases during purification (**fig. S2A**). These two observations would be consistent with lipid molecules with unsaturated acyl chains copurifying with Atlas G_C, but the local resolution of the map is insufficient to ascertain whether the GPL binding pocket contains a ligand.

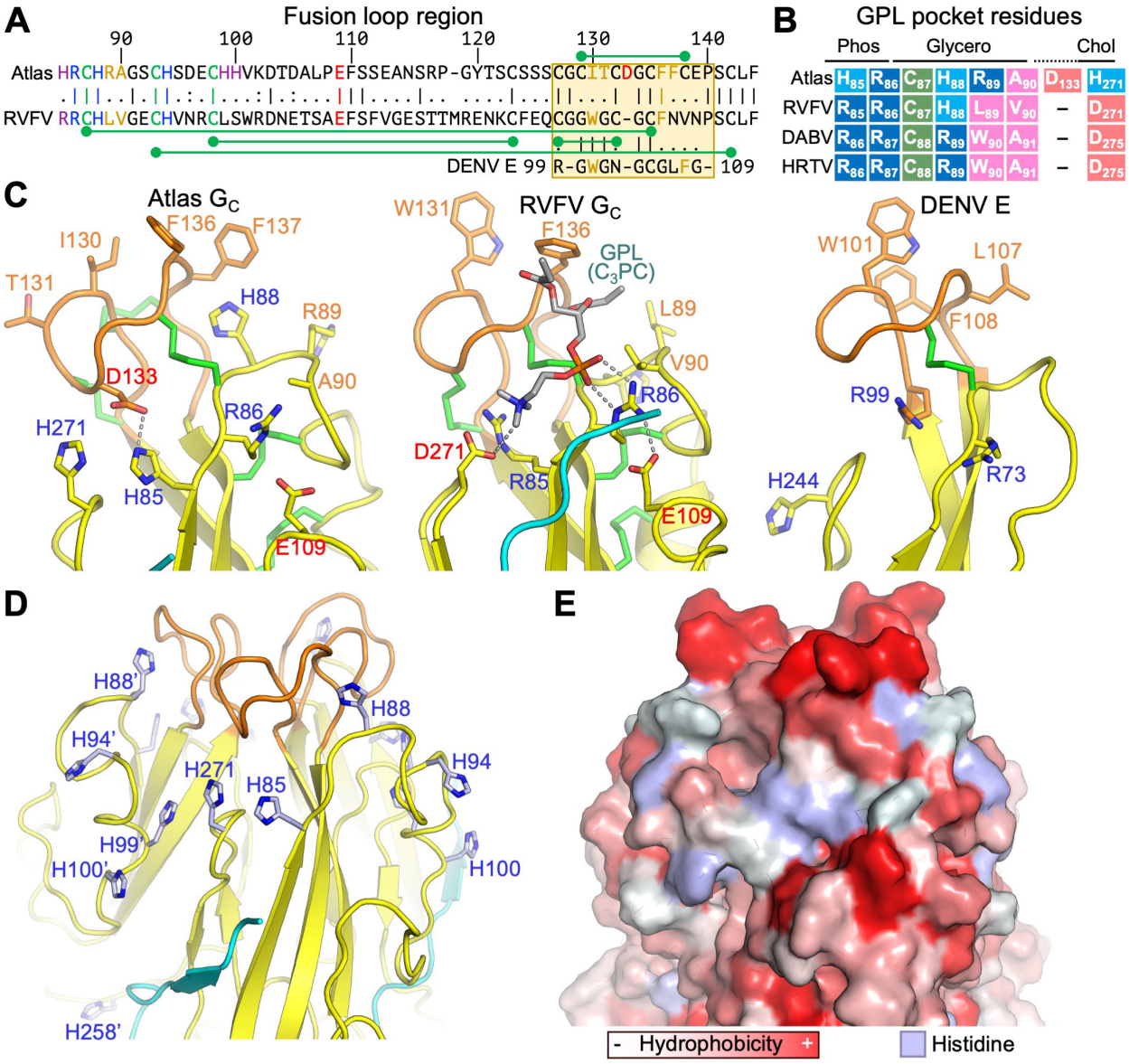


Fig. 3. Structure of the putative fusion loop and lipid binding pocket of Atlas G_C. (**A**) Local sequence alignment of the fusion loops (orange boxes) of Atlas G_C, RVFV G_C, and DENV E. Green dumbbells indicate disulfides. Disulfides shown below the RVFV sequence are conserved in Atlas and RVFV. (**B**) Residues contributing to binding of the phosphate (Phos),

271 glycerol and choline (Chol) moieties of glycerophospholipid (GPL) molecules by RVFV G_C (*28*).
272 Colors indicate side chain properties: blue, positive charge; pink, negative charge; magenta,
273 hydrophobic; green, sulfhydryl; light blue, positive charge at endosomal pH. (**C**) Closeup of the
274 fusion loop and GPL binding pockets of Atlas G_C, RVFV G_C and dengue E. A single protomer
275 from the postfusion trimer is shown, with key residues shown in stick representation. Label
276 colors indicate side chain properties: blue, positive charge; red, negative charge; orange,
277 hydrophobic. (**D**) Closeup of the fusion loop and GPL binding pocket of the Atlas G_C trimer
278 showing histidine residues in or near the pocket. Prime symbols following residue numbers
279 denote the protomer to which the residue belongs. (**E**) Surface representation of the same view
280 as in (**D**), colored by side chain hydrophobicity, except for histidine residues shown in light blue.

281 **Atlas G_C binds membranes with endosome-like lipid composition at low pH**

282 A key step in viral membrane fusion is binding of the fusion protein to the host cell membrane.
283 We assessed binding of Atlas G_C ectodomain to liposomes in density gradient centrifugation
284 and dynamic light scattering (DLS) experiments. Viruses containing class II fusion proteins, like
285 many retroviruses, undergo membrane insertion and fusion in endosomal compartments where
286 the pH is acidic (*53-56*). We therefore assayed liposome binding at a range of pH values. In
287 contrast to RVFV G_C ectodomain, Atlas G_C ectodomain did not bind liposomes containing
288 phosphatidylcholine (PC), phosphatidylethanolamine (PE), cholesterol and sphingomyelin (SM)
289 at neutral or acidic pH (pH 4-8; **Fig. 4, A and B**). At neutral pH (pH 7.8), Atlas G_C ectodomain
290 also failed to bind liposomes containing anionic lipids enriched in early or late endosomes:
291 phosphatidylserine (PS) or bis(monoacylglycerol)phosphate (BMP, also known as
292 lysobisphosphatidic acid or LBPA), respectively. At pH 4, however, Atlas G_C ectodomain bound
293 tightly to liposomes containing PS or BMP, with weaker binding observed at pH 4.6 (**Fig. 4, A**
294 **and B, and fig. S7**). Remarkably, Atlas G_C ectodomain bound only weakly to liposomes
295 containing phosphatidylglycerol (PG) instead of BMP even though PG and BMP are

296 regioisomers with identical chemical composition and electrostatic charge (of -1). PG and BMP
297 differ only in the position of the second acylglycerol linkage, resulting in a linear configuration for
298 BMP instead of the usual branched configuration for PG. Our liposome binding data show that
299 Atlas G_C binds to membranes containing specific GPLs that are enriched in the endosomal
300 pathway in a pH-dependent manner. No other class II fusion protein has been reported to
301 require low pH, PS or BMP for membrane insertion. However, phleboviruses, which require only
302 PE or PC and cholesterol for membrane insertion (*28*), subsequently require BMP and low pH
303 for fusion (*57*). Similarly, flaviviruses require BMP, PS or other anionic lipids and low pH for
304 efficient fusion (*53, 55, 58*).

305 The optimal pH for membrane binding of Atlas G_C (pH 4-4.5) is similar to the optimal pH
306 of hemifusion of Uukuniemi virus (*57*), a model phlebovirus, and would be consistent with
307 membrane insertion in late endosomes, as is the case for phleboviruses (*54*). The side chains of
308 His85 and His271, in the GPL binding pocket of Atlas G_C, would be fully protonated at pH 4-4.5.
309 The resulting net positive charge of the His85/Asp133/His271 triad (+1/-1/+1), analogous to the
310 Arg/Asp pair that coordinates the end of GPL headgroups in phleboviruses, mirrors the charge
311 of the phosphoserine headgroup of PS (-1/+1/-1). Moreover, Atlas G_C contains four additional
312 solvent-exposed histidines (residues 88, 94, 99 and 100) in the vicinity of the GPL binding
313 pocket (**Fig. 3, D and E**). Protonation of these histidines at low pH may promote further
314 interactions with anionic lipid headgroups. The presence of six histidines in and around the GPL
315 binding pocket provides a possible explanation for the observed pH-dependent insertion of Atlas
316 G_C into membranes containing PS and BMP. Consistent with a conserved role for the GPL
317 binding pocket in determining lipid specificity of class II fusion proteins, mutations in
318 alphaviruses at a position equivalent to His271, in the *ij* loop, determine the extent to which
319 alphaviruses depend on cholesterol for membrane binding (*59*).

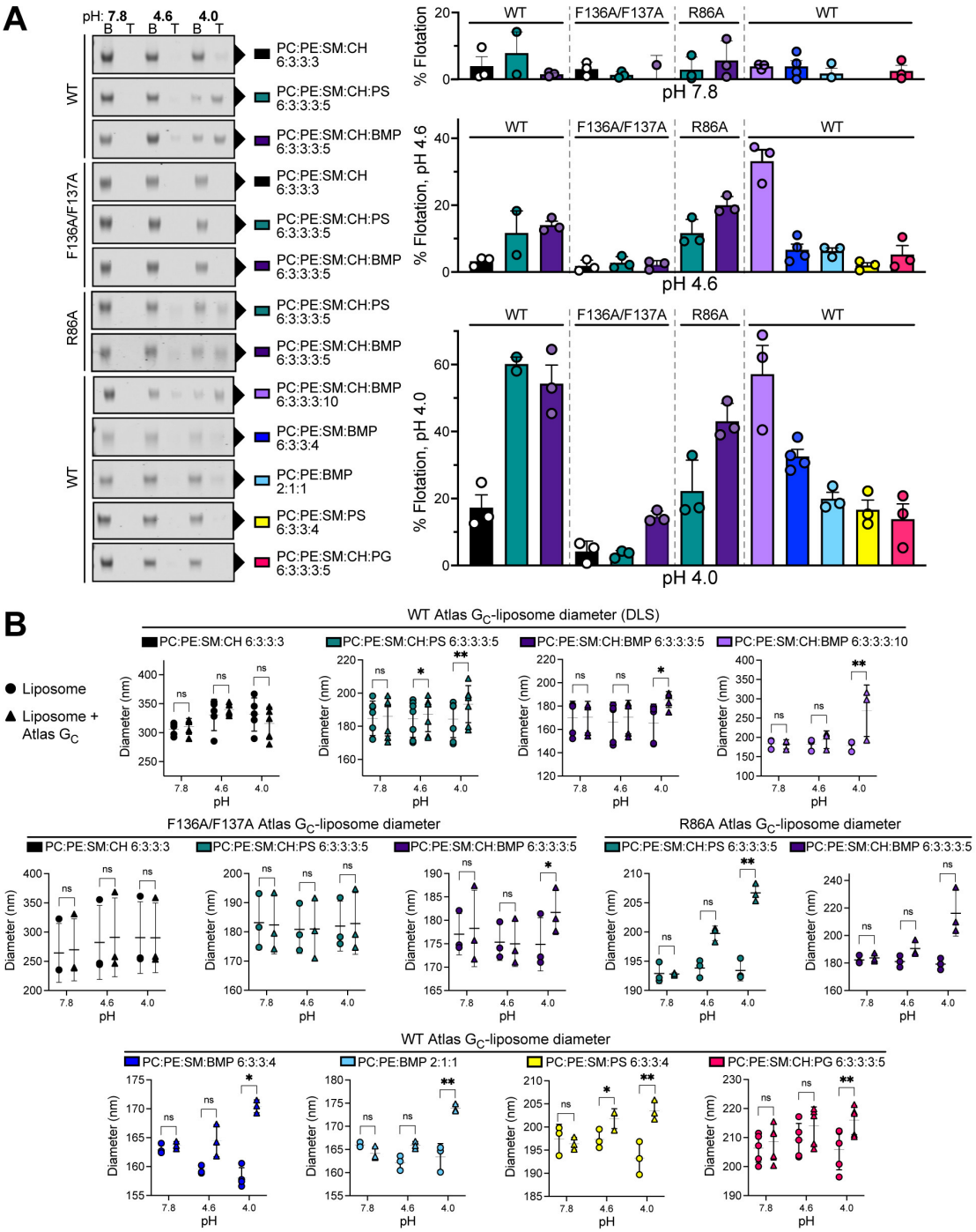


Fig. 4. Atlas G_C binds liposomes in a lipid- and pH-dependent manner. (**A**) Liposome co-flotation lipid binding assay. A liposome-G_C ectodomain mixture in 40% OptiPrep density gradient medium was overlaid with a 30% OptiPrep cushion and centrifuged at 100,000 g. Flotation was defined as the amount of G_C ectodomain co-floating with liposomes in the top-half fraction (T) divided by the total amount of G_C ectodomain in the top- and bottom-half (B)

326 fractions. G_C ectodomain was quantified by Coomassie-stained SDS-PAGE. Error bars show
327 the standard deviation of three replicates except WT+PC:PE:SM:CH:PS (two replicates) and
328 WT+PC:PE:SM:BMP (four replicates). See **fig. S8** for uncropped gels for all replicates. (**B**)
329 Binding of G_C ectodomain (WT, F136A/F137A or R86A) to liposomes measured by dynamic
330 light scattering (DLS) as differences in liposome diameter in the presence and absence of Atlas
331 G_C ectodomain. Error bars show the standard deviation of three to seven replicates.
332 Significance was determined by 2-way ANOVA of the mean change in liposome diameter, using
333 Sidak's multiple comparisons test with a 95% confidence interval in GraphPad Prism 8 (see **fig.**
334 **S7**). **, $0.001 < p < 0.01$; *, $p < 0.05$; ns, not significant. See **dataset S1** for source data.

335 **Atlas G_C membrane binding does not strictly require cholesterol and occurs via the**
336 **fusion loop**

337 In addition to GPLs, phleboviruses and alphaviruses (but not flaviviruses) require cholesterol for
338 efficient membrane binding and subsequent fusion (*28, 57, 59*). Alphaviruses additionally
339 require sphingolipids (such as SM) for efficient fusion (*60, 61*). We found that neither cholesterol
340 nor SM were required for Atlas G_C to bind liposomes (**Fig. 4, A and B**). Removal of cholesterol
341 reduced the fraction of G_C bound by 50% in the liposome flotation assay (**Fig. 4A**) though
342 binding was still detected in the DLS assay (**Fig. 4B**). Hence, although cholesterol and SM are
343 not strictly required for binding, they enhance binding, possibly by increasing membrane fluidity.
344 Notably, the concentration of cholesterol in nematode cell membranes is approximately 20 times
345 lower than in vertebrates (*62, 63*). This is insufficient for cholesterol to regulate the structure or
346 fluidity of nematode membranes, in which cholesterol is thought to be instead a precursor for
347 low-abundance metabolites (*62-64*). Likewise, *Drosophila* can grow indefinitely with only trace
348 amounts of exogenous sterols (*62*), suggesting that arthropods, which are obligate vectors of
349 the vast majority of viruses containing class II fusion proteins, rely on lipids other than
350 cholesterol to regulate membrane fluidity.

351 To determine whether Atlas G_C binds membranes in a manner analogous to other viral
352 class II fusion proteins – via the fusion loop and GPL binding pocket – we purified Atlas G_C
353 ectodomain variants with mutations in the two phenylalanine residues in the fusion loop
354 (F136A/F137A), or in the arginine predicted to bind the GPL phosphate moiety (R86A). The
355 F136A/F137A mutant failed to bind liposomes containing PS and co-flotation with liposomes
356 containing BMP at pH 4 and pH 4.6 was reduced to approximately one third of wild-type (**Fig.**
357 **4C**). The R86A mutation reduced binding to liposomes containing PS or BMP at pH 4, to one
358 third and 75% of wild-type, respectively (**Fig. 4D**). For both variants, preparations contained
359 trimers as the major species but a small monomeric fraction was also present (**fig. S2**),
360 suggesting that the mutated lipid binding residues are required for efficient trimer assembly. We
361 conclude that Atlas G_C binds to lipid membranes through insertion of hydrophobic fusion-loop
362 residues and coordination of lipid headgroups in the GPL binding pocket, as in other viral class
363 II fusion proteins.

364 **Evidence for monomeric and trimeric states of Atlas G_C**

365 Class II fusion protein ectodomains can be monomeric, dimeric or form icosahedral shells in
366 their prefusion conformation, but the fusogenic conformational change is always accompanied
367 by reorganization into trimers (*26, 57, 65, 66*). Fusion proteins from classes I and III, including
368 retrovirus fusogens, remain trimeric throughout the fusion reaction, but no class II fusion
369 proteins are known to be trimeric in their prefusion conformation. Having established that Atlas
370 G_C can insert into membranes as a trimer with a postfusion-like conformation, we set out to
371 determine whether it could undergo a conformational change as seen in the fusion reaction of
372 class II proteins from infectious viruses. The Atlas G_C construct described above was expressed
373 as a trimer with no trace of monomers or dimers (**Fig. 5A and fig. S2A**). However, we found
374 that a construct with the stem region truncated, G_C(DI-III), spanning residues 2330-2751, was
375 expressed as a mixture of monomers, trimers, and higher order oligomers (**Fig. 5B and fig.**

376 **S2B**). In contrast to G_C trimers, which were stable at different protein concentrations and pH
377 values, G_C(DI-III) monomers were unstable over time. As noted above, monomeric fractions
378 were also present in preparations of the fusion loop mutant (F136A/F137A) and GPL binding
379 pocket mutant (R86A) (**fig. S2, C and D**). Whether these monomeric species are in a prefusion
380 conformation remains to be determined, but the presence of metastable monomers and stable
381 trimers recapitulates a key property of class II fusogens from infectious viruses.

382 **pH-dependent stabilization of the Atlas G_C trimer by protonated histidine residues**

383 The increase in positive charge resulting from histidine protonation is an important part of the
384 pH sensing mechanism of viral class II fusion proteins. Protonation of conserved histidines at
385 the domain I-domain III interface of alphavirus, flavivirus and phlebovirus glycoproteins
386 promotes the fusogenic conformational change by destabilizing the prefusion conformation and
387 stabilizing the postfusion conformation (*20, 28, 67-70*). For example, histidines in domain III of
388 phlebovirus G_C proteins form interprotomer salt bridges with negatively charged side chains in
389 the postfusion trimer and mutation of one such histidine in RVFV, His1087, renders the virus
390 uninfectious (*71*). Similarly, in Atlas G_C His258, His333 and His336 form interprotomer or
391 interdomain polar contacts (**Fig. 5C**). These histidine-dependent contacts would stabilize the
392 trimeric postfusion-like conformation of Atlas G_C specifically in acidic endosomal compartments.
393 The parallels in how Atlas G_C and class II fusogens from infectious viruses respond at the
394 ultrastructural level to environmental cues support the hypothesis that Atlas G_C would have
395 membrane fusion activity in late endosomes, like phleboviruses and many retroviruses.

396 **Fifteen disulfide bonds stabilize Atlas G_C in its postfusion-like conformation**

397 With 30 cysteine residues forming 15 disulfide bonds, Atlas G_C contains twice the average
398 abundance of cysteines, more than has been found in any other class II protein. 20 of these
399 cysteines form disulfides that are structurally conserved in phleboviruses (including RVFV,
400 DABV and HRTV) but not in other class II proteins (**fig. S5**). An eleventh disulfide, in domain III,

is conserved in Atlas virus, DABV and HRTV but not RVFV. However, Atlas G_C contains four additional disulfides not conserved in phleboviruses: one in the fusion loop, two in domain II in the β-hairpin containing the *ij* loop (one of the cysteines forming these disulfides is conserved in phleboviruses but forms a disulfide with a cysteine in a different β-strand in domain II), and one in domain III (**Figs. 2C and 3A, and fig. S5**). As discussed above, the Atlas-specific disulfide in the fusion loop (Cys129-Cys138) extends the hydrophobic surface formed by conserved residues in the fusion loop that are required for membrane insertion (**Fig. 5D**). We note that due to the location of Cys129 and Cys138 close to the threefold symmetry axis of the trimer, the side chains of the two residues can be rearranged by torsional rotation, to form intermolecular disulfides across the trimer interface, thereby crosslinking all three protomers in the trimer.

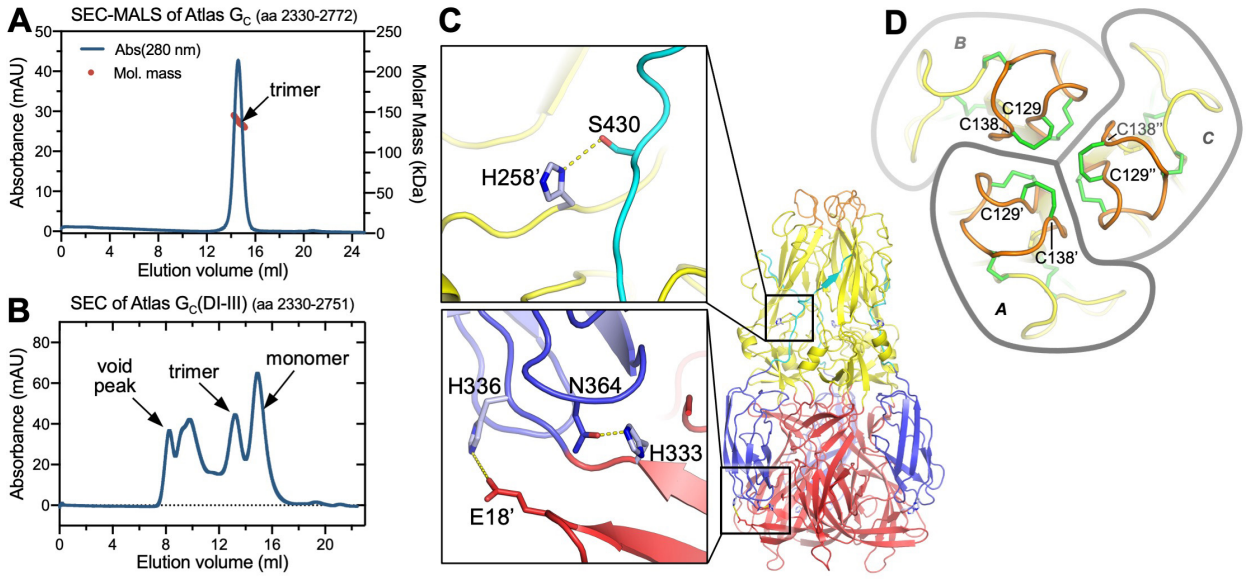


Fig. 5. Oligomeric states and disulfide bonding pattern of Atlas G_C. (**A**) SEC-MALS of Atlas G_C ectodomain (residues 2330-2772). The protein formed trimers. (**B**) SEC of Atlas G_C(DI-III) (residues 2330-2751) expressed in a mixture of oligomeric states, including monomers, trimers, and larger aggregates (void peak). G_C(DI-III) monomers were unstable**.** (**C**) His258, His333 and His336 form interprotomer or interdomain polar contacts predicted to stabilize the G_C trimer specifically at pH <6, when the histidine side chains are charged. Residues from different

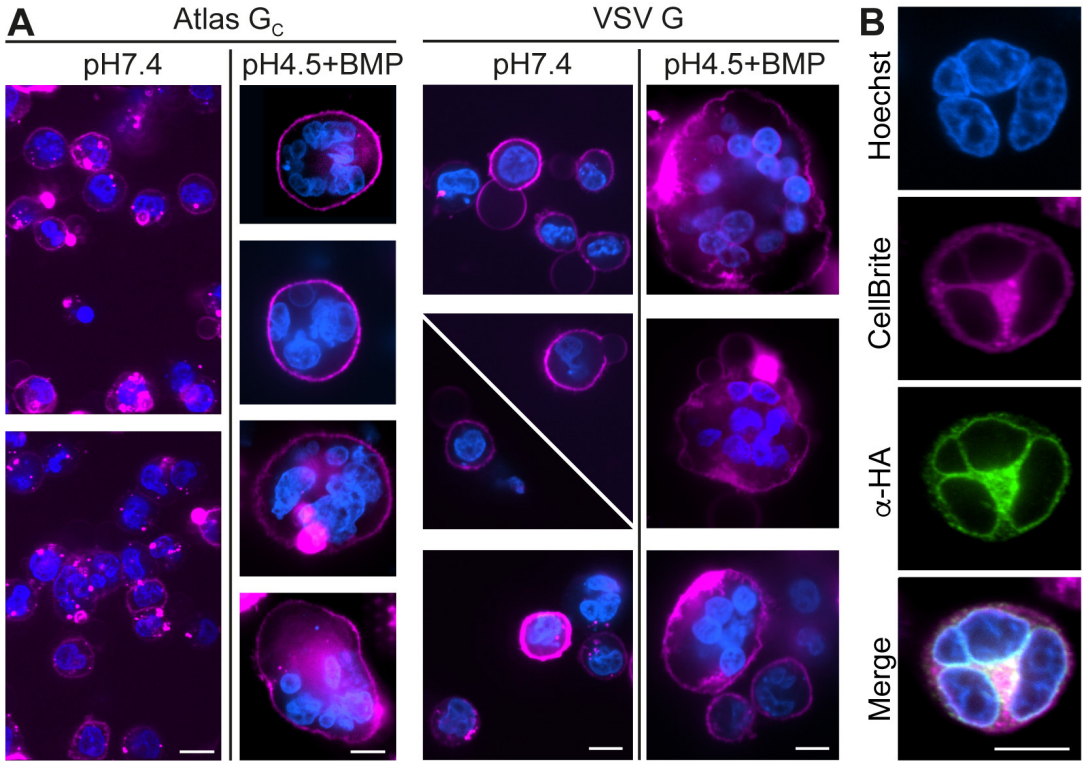
418 protomers are denoted with a prime symbol. (**D**) View along the threefold axis of the G_C trimer
419 with intramolecular disulfides between the cysteines most proximal to the axis, Cys129 and
420 Cys138.

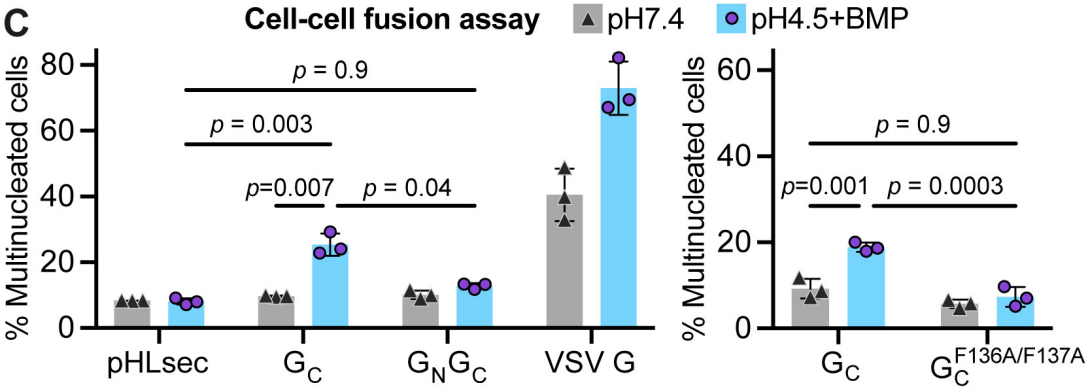
421 **Atlas G_C retains membrane fusion activity**

422 Having identified the hallmarks of a fusion protein in Atlas G_C, we measured its membrane
423 fusion activity in a cell-cell fusion assay. CHO cells were transfected with plasmids encoding
424 Atlas G_C or Atlas G_N-G_C. To promote plasma membrane localization and minimize ER retention,
425 we replaced the predicted transmembrane domain and cytosolic tail of G_C with the C-terminal
426 transmembrane anchor and cytosolic tail from HLA-A2, known to localize to the plasma
427 membrane (*72*). Plasmids encoding vesicular stomatitis virus (VSV) G, or no protein, were used
428 as positive and negative controls, respectively. Since Atlas G_C trimers require low pH and BMP
429 or PS to efficiently bind membranes, we treated transfected cells with exogenous BMP and then
430 transferred them to pH 4.5 buffered medium to trigger fusion. Similar treatments have been
431 used previously to measure cell-cell fusion activity of flavi- and alphaviruses (*53*). Confocal light
432 microscopy with nuclear and plasma membrane stains showed that cells with three or more
433 nuclei were common in cells expressing Atlas G_C following treatment with BMP and pH 4.5,
434 though less abundant than in cells expressing VSV G, with or without treatment (**Fig. 6, A and**
435 **B, and movie S1**). Cells with more than three nuclei were rare in cells expressing Atlas G_N-G_C.
436 To quantify cell-cell fusion we counted nuclei and multinuclear cells in micrographs. We found
437 that in cells expressing Atlas G_C, the fraction of multinuclear cells, defined as cells containing
438 two or more nuclei, was 25 ± 3 % following treatment with BMP and pH 4.5, versus 10 ± 0.2 %
439 without treatment (**Fig. 6C**; 95% confidence intervals). Transfection of the G_C fusion loop mutant
440 F136A/F137A resulted in essentially identical multinucleation fractions as with empty vector,
441 with or without treatment with BMP and pH 4.5 (**Fig. 6C**), underpinning the importance of the
442 fusion loop in fusion activity. By comparison, in cells expressing VSV G, 73 ± 8 % of treated

443 cells were multinuclear. 8 ± 1 % of cells transfected with empty vector were binuclear with or
444 without treatment, but none contained more than two nuclei (**fig. S9**). Incubation of transfected
445 cells with CDK4, which arrests the cell cycle at G1 phase, had no measurable effect on the
446 fraction of multinuclear cells (**fig. S9D**), indicating that cytokinesis effects do not significantly
447 contribute to this cell-cell fusion assay readout. We conclude that Atlas G_C has approximately
448 one third of the membrane fusion activity of VSV G under the treatment conditions tested, which
449 is significant given that VSV G is considered highly fusogenic and widely used as a model
450 fusogen.

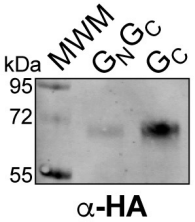
451 The Atlas G_N-G_C construct tested had little fusion activity (**Fig. 6C**). To investigate this,
452 we measured G_C expression on the surface of cells transfected with the G_N-G_C plasmid. Cell-
453 surface proteins were biotinylated, affinity-purified and analyzed by immunoblot and liquid
454 chromatography with tandem mass spectrometry (LC-MS/MS). When expressed as part of a
455 G_N-G_C construct, only trace levels of cell-surface G_C were detected by Western blot (**Fig. 6D**).
456 Moreover, no G_N peptides were observed by LC-MS/MS, and the LC-MS/MS signal for G_C was
457 significantly weaker than in cells transfected with G_C alone (**Fig. 6E**), consistent with the lack of
458 fusion activity in cells transfected with G_N-G_C. We conclude that the G_N was degraded or failed
459 to be secreted in CHO cells, which in turn prevented efficient secretion of G_C.





D

Cell-surface biotinylation Western blot



E **Cell-surface biotinylation and mass spectrometry**

Sample	Biotin-ylation	Norm. E signal	Intensity (10^6)	G_C peptides detected	Score	MS/MS count	%G_C seq. coverage
G_C	-	1.138	4.58	DHQIVSESVHR	78.7	1	4.6
				NLCTCTNALHK	69.7	1	
G_C	+	3.858	11.5	DHQIVSESVHR	102	1	10.5
				IISQSYSATDVER	94.3	1	
				NFLVFEK	64.6	2	
				QFECHFSR	71.4	1	
				NLCTCTNALHK	106	1	
G_NG_C	-	N/A	0		0	0	0
G_NG_C	+	0.709	3.72	DHQIVSESVHR	119	1	7.8
				NFLVFEK	71.3	1	
				QFECHFSR	49.1	1	
				NLCTCTNALHK	68.5	1	

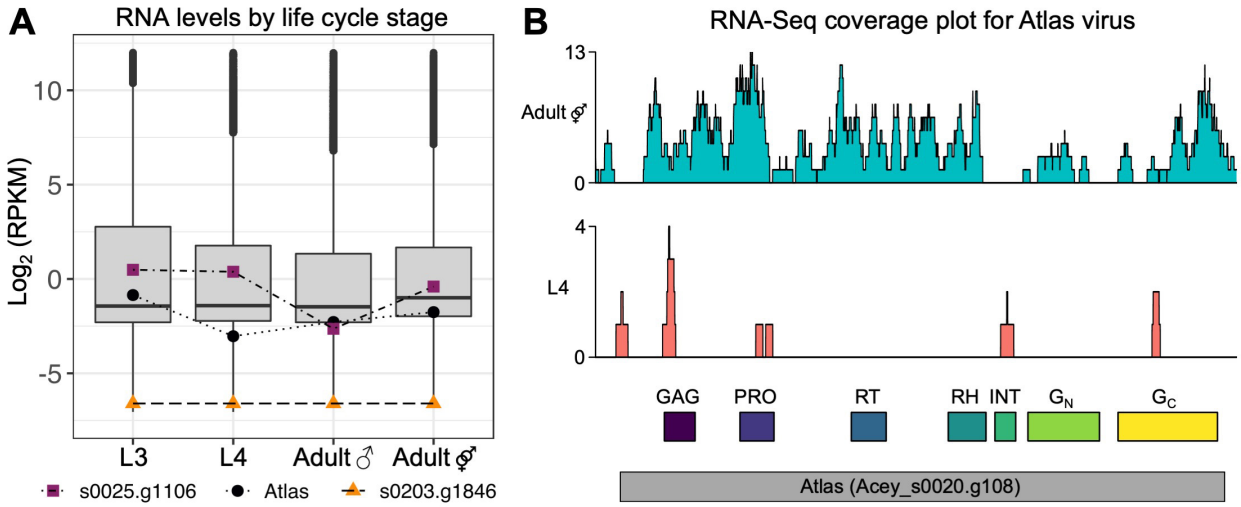
Fig. 6. Cell-cell fusion assay. (**A**) Confocal micrographs of CHO cells expressing Atlas G_C fused to the transmembrane anchor from HLA-A2, or VSV G. Cells were treated with pH 4.5

463 buffer containing BMP, or pH 7.4 buffer without lipid. Blue, Hoechst 33342 nuclear stain;
464 Magenta, CellBrite Red plasma membrane dye. Micrographs were selected to show fusion
465 events – see **fig. S9** for representative raw micrographs. Scale bars, 10 µm. (**B**) Confocal
466 micrograph of a multinucleated CHO cell expressing Atlas G_C following treatment with pH 4.5
467 and BMP. Blue, Hoechst 33342; magenta, CellBrite Red; green, a-HA antibody for Atlas G_C
468 detection. Scale bar, 10 µm. A composite Z-stack of this cell is shown in **movie S1**. (**C**) The
469 fraction of multinuclear cells transfected with plasmids encoding Atlas G_C, Atlas G_N-G_C, VSV G
470 or pHLsec empty vector was calculated by counting mono- and multi-nucleated cells using the
471 Hoechst and CellBrite stains. Error bars represent SD between measurements from three
472 distinct experiments, with 135-407 nuclei counted per replicate. *P*-values were calculated by 2-
473 way ANOVA (Tukey's multiple comparisons test, GraphPad Prism 9). (**D** and **E**) G_C cell-surface
474 expression analysis. Cells were biotinylated, cell-surface proteins were affinity-purified with
475 NeutrAvidin agarose, and the cell-surface G_C fraction was quantified by Western blot, (**D**), and
476 LC-MS/MS mass spectrometry, (**E**). See **dataset S2** for source data.

477 **Transcriptional activity of Atlas virus based on RNA-Seq data**

478 Transcription of endogenous viruses is tightly regulated to prevent harmful gene expression and
479 genome damage from transposition events. However, some EVEs escape repression and are
480 highly transcribed, including a subset of belpaoviruses from the parasitic trematode
481 *Schistosoma mansoni* (blood fluke) (*73*). We analyzed transcript levels of Atlas and other
482 complete belpaoviruses in previously published RNA-Seq data for *A. ceylanicum* (*74*).
483 Transcript abundance of the belpaovirus EVEs encoding complete polyproteins varied, ranging
484 from undetectable to above the median value for all annotated genes (**Fig. 7 and fig. S10**). For
485 Atlas virus, transcript levels varied across developmental stages: expression was above the
486 median for annotated genes at the larval L3 stage, below the first quartile at stage L4, and near
487 the median in adult hookworms (**Fig. 7A**). Notably, Atlas virus transcript expression increased

488 15-fold from the L4 to adult (mixed sex) stages (adjusted *p*-value = 0.0021, see **dataset S3** for
489 differential expression calculations).



490

491 **Fig. 7. RNA-Seq signal from Atlas virus and other *A. ceylanicum* genes.** (**A**) Box plots of
492 RNA-Seq Reads Per Kilobase of transcript per Million mapped reads (RPKM) for all annotated
493 *A. ceylanicum* genes, at different developmental stages (L3 and L4 larva, adult male and adult
494 mixed sex). RNA-Seq data are from Bernot *et al. (74).* Bars in each box represent the median.
495 Box boundaries represent the first and third quartiles. RPKM values are shown for Atlas virus
496 and the complete belpaovirus EVEs with the highest and lowest RNA-Seq signals, s0025.g1106
497 and s0203.g1846, respectively. (**B**) Distribution of RNA-Seq reads over the Atlas virus coding
498 sequence for the adult (mixed sex) and L4 developmental stages. Read coverage was greater
499 in adults. See **fig. S10** for a snapshot of this region from Integrative Genomics Viewer with
500 unmerged BAM files.

501 **Discussion**

502 Here, we identify Atlas virus as an endogenous belpaovirus containing intact *gag*, *pol* and *env*
503 genes with novel features in the human hookworm *A. ceylanicum*. The cryo-EM structure of the
504 Atlas Env reveals a class II viral fusion protein fold remarkably similar to that of the G_C
505 glycoprotein from Rift Valley Fever virus. Atlas G_C has the hallmarks of an active class II

506 membrane fusion protein: a stable trimeric assembly, a putative fusion loop, membrane
507 insertion triggered by low pH with specificity for late endosomal lipid composition, and
508 membrane fusion activity. Moreover, Atlas virus mRNA is expressed at specific stages of
509 hookworm development.

510 Our work supports the model first proposed based on phylogenetic studies that the
511 nematode belpaoviruses acquired their Env by horizontal gene transfer from a virus from the
512 *Phenuiviridae* family or a phlebovirus-like ancestor (*14*). It remains unclear whether RNA
513 encoding the phlebovirus-like glycoprotein integrated into the belpaovirus ancestor as mRNA in
514 a splicing event, or by first becoming a substrate for the reverse transcriptase with subsequent
515 genomic integration as dsDNA. The envelope proteins from retroviruses, including ERVs, that
516 have been biochemically characterized were all found to be class I fusion proteins with an α-
517 helical coiled-coil as the core fold. Viral class II fusion proteins have so far been found only in
518 non-integrating RNA viruses. Our discovery of an endogenous belpaovirus with a potentially
519 functional, phlebovirus-like class II fusion protein that is structurally unrelated to retrovirus Envs
520 reveals an unexpected degree of structural and genetic plasticity in reverse-transcribing RNA
521 viruses. More generally, the presence of the class II fusion protein fold in EVEs and across
522 many phyla suggests this fold is derived from a common ancestor that could have been viral or
523 cellular in origin (*33-36, 75*).
524

525 While rare, horizontal gene transfer of atypical fusogens into retroelements is not unique
526 to the belpaoviruses. The Tas element from the nematode *Ascaris lumbricoides*, a common
527 parasitic worm in humans, has an *env* gene with weak genetic similarity to herpesvirus gB
528 proteins (*14*), which have a class III fusion protein fold (also found in *Rhabdoviridae* and
529 *Baculoviridae*). Taken together, these findings lead us to hypothesize that acquisition of a fusion
530 protein from an infectious virus represents a general paradigm for how retrotransposons can

531 become retroviruses, and indeed how ancestral reverse-transcribing viruses may have
532 originated.

533 The *env* gene is often the first element to be lost in EVEs, as it is not required for
534 intracellular proliferation, so it is notable that the G_N-G_C *env* module is intact in Atlas virus. With
535 its identical LTRs, and no stop codons or frameshift mutations, Atlas virus shows all the signs of
536 being intact and recently active. This supports the notion that the envelope may be functional
537 (*12*). The preserved biological activities of Atlas G_C suggest these activities could have cellular
538 functions in health and disease, as reported for a small but increasing number of ERV *env* and
539 *gag* gene products (*1, 3, 5, 76*). Further studies are required to determine the full extent to
540 which protein expression from transposable elements – and its dysregulation – contribute to
541 basic cellular functions, embryonic development, and disease outcomes. This work provides a
542 blueprint for such efforts.

543 **Materials and Methods**

544 **Genomic analyses of *A. ceylanicum* Atlas virus.** A PSI-BLAST search (*40*) for protein
545 sequences similar to biochemically characterized phlebovirus fusion proteins identified the gene
546 *Acey_s0020.g108* (UniProt A0A016UZK2; Genbank JARK01001356.1, genomic translation
547 EYC20859.1) in the human hookworm *Ancylostoma ceylanicum* as the most similar sequence
548 outside infectious virus taxa, with an Expected value (E-value) of 10^{-20} against the Rift Valley
549 Fever virus (RVFV) G_C sequence (UniProt P03518). A second iteration performed using
550 position-specific scoring matrix based on an alignment of sequences identified in the first
551 iteration gave an E-value of 10^{-144}. The *Acey_s0020.g108* gene is referred to here as Atlas
552 virus.

553 Phylogenetic analysis of the phlebovirus G_C-like protein from Atlas virus and other G_C or
554 G_C-like proteins with similar protein sequences was performed as follows. A reference multiple
555 sequence alignment was initially generated by aligning the Atlas G_C protein sequence with the

556 sequences of reference protein RVFV G_C and other viral G_C proteins identified as having
557 structural similarity in structure comparisons with DALI (*50*), namely the G_C proteins from Dabie
558 bandavirus (DABV, formerly SFTSV phlebovirus, GenBank:AGM33042.1) and Heartland virus
559 (HRTV, GenBank:AFP33394.1). For each G_C protein, the sequence range present in the
560 corresponding Protein Data Bank (PDB) entry (7A4A, 6EGU, 5G47, or 5YOW, respectively) was
561 defined as the G_C ectodomain sequence and used to generate the reference alignment with
562 MAFFT in SnapGene v5.1.7. The sequences of the following EVEs with detectable Env protein
563 sequence similarity were then added to the alignment: *Necator americanus* NECAM_13468
564 (GenBank:XM_013440329.1)*, C. elegans* Cer13 virus (GenBank: Z81510.2; WormBase
565 WBTransposon00000728, gene Y75D11A.5), and seven *A. ceylanicum* EVEs encoding
566 complete Gag-Pol-Env polyproteins with predicted phlebovirus-like Envs (**Fig. 1B**; GenBank
567 EYC27361.1, EYC19962.1, EYC20099.1, EYC18998.1, EYC07469.1, EYB91703.1,
568 EYB80846.1). The tree was calculated with IQ-TREE v1.6.12 and drawn with iTOL v6 (*44*).

569 The reverse transcriptase (RT) phylogenetic tree was calculated based on the protein
570 sequences of Atlas virus, Cer13 virus, the seven *A. ceylanicum* EVEs listed above, and seven
571 other belpaoviruses (GenBank AF060859.1, Z29712.1, L09635.1, U23420.1, AY180917.1,
572 D83207.1, AF537216.1). RT domains (Pfam:PF00078, RVT_1) were identified with PfamScan
573 in five of the sequences (GenBank Z81510.2, EYC27361.1, EYC19962.1, EYC07469.1,
574 EYB80846.1). The protein sequences of these five RTs were extracted and aligned with MAFFT
575 and Clustal Omega v1.2.4. The tree was generated with IQ-TREE v1.6.12 and iTOL v6 (*44*).

576 The Atlas virus LTRs were identified with REPuter (*77*). Sequence identity matrices (**fig.**
577 **S1**) were calculated with SDT (*78*).

578 **Protein expression and purification.** Synthetic genes encoding soluble ectodomain fragments
579 of the Env of endogenous retrovirus Y032_0020g108 from *A. ceylanicum* were subcloned into

580 the pMT/BiP/V5-His vector (ThermoFisher) in frame with the BiP signal sequence and the C-
581 terminal V5 and six-histidine tags. The constructs referred to here as Atlas G_C and Atlas G_C(DI-
582 III) span amino acids 2330-2772 and 2330-2751 from UniProt A0A016UZK2, respectively. Atlas
583 G_C mutants were generated by Dpn I-based site-directed mutagenesis. D.mel-2 insect cells
584 (ThermoFisher) were co-transfected with the expression construct and blasticidin resistance
585 marker pCoBlast (ThermoFisher) at a 20:1 molar ration and cultured for 6 weeks in 0.5 µg ml^{-1}
586 blasticidin to obtain a population of expressor cells. Expression was induced in a shaking cell
587 suspension at 27˚C with 0.5 mM $CuSO_4$ at a cell density of 5 x 10^6 cells ml^{-1}. The cell culture
588 medium was harvested 4 to 5 days after induction, centrifuged to remove cells (2,000 g) and cell
589 debris (17,000 g), filtered with a 0.2 µm filter, concentrated by tangential-flow filtration, and
590 buffer exchanged into 20 mM Tris pH 7.8, 0.3 M NaCl, 5% glycerol, 20 mM imidazole, 0.5 mM
591 TCEP (tris(2-carboxyethyl)phosphine). Atlas G_C was purified by nickel-affinity chromatography
592 with a HisTrap Excel column (Cytiva), followed by anion-exchange chromatography with a
593 MonoQ or Resource Q column (Cytiva) using 20 mM Tris pH 8.0, 50 mM NaCl, 5% glycerol, 0.5
594 mM TCEP as the binding buffer and binding buffer plus 1 M NaCl as the elution buffer. Peak
595 fractions were concentrated and further purified by size-exclusion chromatography (SEC) with a
596 Superdex 200 Increase (10/300) column (Cytiva) in 20 mM Tris pH 7.8-8.0, 0.15 M NaCl, 5%
597 glycerol and 0.5 mM TCEP. The C-terminal V5 and histidine tags were optionally cleaved by
598 incubation with carboxypeptidase A (CPA) for 3 h at 4˚C (1:500 CPA:G_C molar ratio).

599 **Liposome binding assay.** 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (PC), 1-palmitoyl-
600 2-oleoyl-snglycero-3-phosphoethanolamine (PE), egg sphingomyelin (SM), 1-palmitoyl-2-oleoyl-
601 sn-glycero-3-phospho-L-serine (PS), 1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-(1'-rac-glycerol)
602 (PG), (*S*,*R*) Bis(Monoacylglycero)Phosphate (BMP) (Avanti Polar Lipids), and 1-cholesterol
603 (Sigma-Aldrich) were dissolved in chloroform. 25 mM lipid solutions were mixed at various molar
604 ratios and dried under nitrogen gas for over 4 h. The lipid film was resuspended in Liposome

605 Buffer (20 mM Tris pH 7.8, 0.15 M NaCl, 5% glycerol, 0.5 mM TCEP, 2 mM $MgCl_2$, 2 mM $CaCl_2$,
606 2 mM KCl) and subjected to five cycles of freeze-thawing in liquid nitrogen, followed by 25
607 cycles of extrusion through two 0.2 µm polycarbonate filter membranes (Whatman). Purified
608 Atlas G_C ectodomain was added in a 1:771 protein:lipid molar ratio and incubated at 37˚C for 5
609 min. The pH was reduced by adding a 2 M stock solution of sodium acetate pH 4.6 or 4.0 to a
610 final concentration of 0.2 M. Following a 2 h incubation at 37˚C the pH of the suspension was
611 neutralized with 1 M Tris pH 8. OptiPrep density gradient medium (Sigma-Aldrich) was added to
612 a concentration of 40%, maintaining 0.15 M NaCl throughout. Approximately 0.5 ml of the
613 liposome suspension was placed in a centrifuge tube, overlayed with a 2.5 ml cushion of 30%
614 OptiPrep solution and centrifuged at 100,000 g for 1 h at 4˚C in a TLA100.3 rotor (Beckman
615 Coulter). Top (T) and bottom (B) fractions (approximately 1.5 ml each) were collected from the
616 top meniscus with a micropipette. Atlas G_C was quantified by densitometry of the absorbance at
617 700 nm of bands in Coomassie-stained SDS-PAGE gels with an Odyssey scanner (LI-COR).
618 Flotation was defined as the amount of Atlas G_C ectodomain in the top fraction divided by the
619 total amount of Atlas G_C ectodomain in both fractions.

620 For measurement of liposome diameter by dynamic light scattering (DLS), the liposome
621 suspensions were diluted ten-fold in Liposome Buffer prior to the addition of 40% OptiPrep
622 solution. Following centrifugation liposome diameters were measured in 384-well clear-
623 bottomed optical imaging plates (Corning) with a DynaPro Plate Reader III (Wyatt
624 Technologies). The mean diameter was calculated as the average of three independent
625 measurements, each consisting of fifteen two-second acquisitions. Protein-free acidified
626 liposome controls were treated and measured in parallel, with Liposome Buffer instead of Atlas
627 G_C solution.

628 **Size-exclusion chromatography and multi-angle scattering (SEC-MALS) analysis.** 100 µl
629 samples containing 1.6 – 2.5 mg ml^{-1} Atlas G_C ectodomain were analysed by size-exclusion

630 chromatography (SEC) at 293 K on a Superdex 200 (10/300) column (Cytiva) in 20 mM Tris pH
631 7.8, 0.15 M NaCl, 5% glycerol and 0.5 mM TCEP with a flow rate of 0.5 ml min^{-1}. The SEC
632 system was coupled to both multi-angle light scattering (MALS) and quasi-elastic light scattering
633 (QELS) modules (DAWN-8+, Wyatt Technology). The protein was also detected as it eluted
634 from the column with a differential refractometer (Optilab T-rEX, Wyatt Technology) and a UV
635 detector at 280 nm (Agilent 1260 UV, Agilent Technology). Molar masses of peaks in the elution
636 profile were calculated from the light scattering and protein concentration, quantified using the
637 differential refractive index of the peak assuming a dn/dc of 0.1860, with ASTRA6 (Wyatt
638 Technology).

639 **Cryo-EM sample preparation and data collection.** Purified Atlas G_C ectodomain trimer (3 µl
640 at a concentration of 0.025 mg ml^{-1}) in 20 mM Tris pH 7.8, 0.15 M NaCl, 5% glycerol and 0.5
641 mM TCEP, was applied onto glow-discharged R1.2/1.3 400 mesh copper grids (Quantifoil Micro
642 Tools, Germany). The grids were blotted for 4 s and plunge-frozen in liquid ethane with a
643 Vitrobot Mark IV (ThermoFisher) at 4˚C and 100% humidity. Preliminary sample screening and
644 initial datasets were acquired on a FEI Tecnai F20 microscope operated at 200 kV equipped
645 with Falcon II direct electron detector (ThermoFisher) at -4 µm defocus. High-resolution cryo-EM
646 dataset collection was performed on a Titan Krios microscope (ThermoFisher) operated at
647 300 kV equipped with a 20 eV slit-width GIF quantum energy-filtered Gatan K2-Summit direct
648 electron detector in counting mode. A total of 3,027 movies were recorded at a calibrated
649 magnification of 130,000x leading to a magnified pixel size of 1.047 Å on the specimen. Each
650 movie comprised 36 frames with an exposure rate of 1.28 e^{-} $Å^{-2}$ per frame, with a total exposure
651 time of 8 s and an accumulated exposure of 46.18 e^{-} $Å^{-2}$. Data acquisition was performed with
652 EPU Automated Data Acquisition Software for Single Particle Analysis (ThermoFisher) with
653 three shots per hole at -1.3 µm to -3.5 µm defocus.

654 **Image processing.** Micrographs from initial datasets allowed us to obtain a consistent model at
655 ~19 Å resolution from 3,790 particles selected after 2D and 3D classification, and consequent
656 auto-refinement. All movies from high-resolution datasets were motion-corrected and dose-
657 weighted with MOTIONCOR2 (*79*). Aligned, non-dose-weighted micrographs were then used to
658 estimate the contrast transfer function (CTF) with GCTF (*80*). All subsequent image processing
659 steps were performed using RELION 3.0 (*81, 82*). 2D references from initial datasets were used
660 to auto-pick the micrographs. One round of reference-free 2D classification was performed to
661 produce templates for better reference-dependent auto-picking, resulting in a total of 987,570
662 particles. After a first round of 2D classification, 595,011 particles were selected to perform a
663 second 2D classification, resulting in a final number of 320,041 selected particles. Then, a 3D
664 classification imposing C3 symmetry was performed using the model from the initial datasets
665 filtered at 40 Å resolution as initial model. The best class, containing 197,145 particles, was
666 selected, and subjected to 3D auto-refinement imposing C3 symmetry, yielding a map with an
667 overall resolution at 4.11 Å based on the gold-standard (FSC = 0.143) criterion. After
668 refinement, the CTF refinement (per-particle defocus fitting and beam tilt estimation) and
669 Bayesian polishing (*83*) routines implemented in RELION 3.0 were performed, yielding a final
670 map with an overall resolution at 3.76 Å. Local resolution was estimated with RELION.

671 **Model building and refinement.** The most similar sequence to Atlas G_C with a structure
672 available was glycoprotein G_C from Rift Valley Fever virus (RVFV). The crystal structure of
673 RVFV G_C in the postfusion conformation (PDB:6EGU (*28*)) was used as template to build a
674 homology model with the sequence of Atlas G_C using the Swiss-Model server
675 (https://swissmodel.expasy.org). The output model was docked as a rigid body into the density
676 with UCSF Chimera (*84*). Initial docking was performed manually and was followed by real
677 space fitting with the Fit in Map routine. A preliminary step of real space refinement was
678 performed on the three-subunit model, with Phenix 1.13 (*85*), with global minimization, atomic

displacement parameter (ADP), simulated annealing and morphing options selected. The model was then rebuilt in Coot (*86*) to optimize the fit to the density. Due to low resolution information in the fusion loop region, the density was converted to .mtz file using CCP-EM software package tools and blurring of the density allowed us to localize bulky residues and disulphide bonds, and thus use them as a guide to build the entire fusion loop. A final step of real space refinement was performed with Phenix 1.15, with global minimization and ADP options selected. The following restraints were used in the real space refinement steps: secondary structure restraints, non-crystallographic symmetry (NCS) restraints between the protein subunits, side chain rotamer restraints, and Ramachandran restraints. Key refinement statistics are listed in **table S1**.

Model validation and analysis. The FSC curve between the final model and full map after post-processing in RELION, Model vs Map, is shown in **fig. S3**. Cross-validation against overfitting was performed as described (*87*). The atoms in the final atomic model were displaced by 0.5 Å in random directions with Phenix. The shifted coordinates were then refined against one of the half-maps generated in RELION, the “work set”. This test refinement was performed in Phenix using the same procedure as for the refinement of the final model (see above). The other half-map, the “test set” was not used in refinement for cross-validation. FSC curves of the refined shifted model against the work set, FSCwork, and against the test set, FSCtest, are shown in **fig. S3**. The FSCwork and FSCtest curves are not significantly different, consistent with the absence of overfitting in our final models.

The quality of the atomic models, including basic protein geometry, Ramachandran plots, clash analysis, was assessed and validated with Coot, MolProbity (*88*) as implemented in Phenix 1.15, and with the Worldwide PDB (wwPDB) OneDep System (https://deposit-pdbe.wwpdb.org/deposition).

703 **Cell-cell fusion assay.** Chinese hamster ovary (CHO) Lec3.2.8.1 cells were transfected with
704 pHL-Sec plasmids encoding ectodomain fragments of Atlas G_N-G_C (polyprotein residues 1909-
705 2795) or Atlas G_C (residues 2330-2795) fused to the C-terminal transmembrane domain from
706 HLA-A2 (residues 288-345) and cloned in frame with the vector's secretion signal and a C-
707 terminal HA tag. Empty pHL-Sec plasmid, and pcDNA encoding vesicular stomatitis virus (VSV)
708 G, were used as negative and positive controls, respectively. 16-20 h post-transfection, cells
709 were transferred to PBS supplemented with 2.5 mM BMP (18:1(*S,S*)
710 Bis(Monoacylglycero)Phosphate (BMP); Avanti Polar Lipids). To obtain a homogeneous BMP
711 suspension, the mixture was freeze-thawed five times using liquid nitrogen and a water bath,
712 followed by a 3 min incubation in a sonicating water bath. Cells were incubated in the BMP
713 suspension (or PBS for the untreated control) at 37˚C for 5 min, shown previously to be
714 sufficient for anionic lipid incorporation into the plasma membrane (*53*). Cell were transferred to
715 pH 4.5 complete medium (DMEM adjusted to pH 4.5 with HCl supplemented with 10% foetal
716 bovine serum), or pH 7.4 complete medium for the untreated control, and centrifuged at 2,500 g
717 at 37˚C for 2 min. Cells were immediately resuspended in complete media and plated out.
718 Following reattachment (4-6 h after treatment), cells were stained with Hoechst 33342 (Bio-Rad)
719 and CellBrite Red cytoplasmic membrane dye (Biotium, cat. no. 30023) and imaged on a Nikon
720 iSIM Swept Field inverted confocal microscope with a 60x/1.2NA water objective. Cluster
721 analysis of the Hoechst channel with Fiji (*89*) was used to count single nuclei and identify
722 polynuclear clusters. Nuclei within polynuclear clusters were counted by visual inspection. The
723 plasma membrane stain was used to confirm polynuclear clusters and count the number of
724 multinuclear cells, defined as cells with two or more nuclei, by visual inspection. The fraction of
725 multinucleated cells (*F*) was calculated using the formula:

726 $$F = 1 - (n.mononucleated\ cells)\ /\ (total\ n.nuclei)$$

727 For testing the effect of inhibiting the cell cycle on the fraction of multinucleated cells,
728 cells were transfected with plasmid pEGFP-N1, and treated on the following day with 10 μM
729 CDK4 (Cayman Chemical, Cat. No. 17648), or PBS in the negative control, for 6 h at pH 7.4.

730 **Cell surface biotinylation and mass spectrometry.** Proteins on the cell surface were
731 biotinylated and isolated with the Pierce™ Cell Surface Protein Biotinylation and Isolation Kit
732 (ThermoFisher, A44390). HEK293T cells were transfected with 3 μg pHLsec-G_NG_C, or pHLsec-
733 G_C. One day later, cells were washed with PBS and incubated with 0.25 g/l Sulfo-NHS-SS-Biotin
734 on ice for 10 min. Then cells were washed with PBS and suspended in 0.5 ml of ThermoFisher
735 Lysis Buffer supplemented with cOmplete EDTA-free Protease Inhibitor Cocktail (Merck). Cells
736 were lysed on ice for 30 min (with 5s vortexing before and after lysis) and centrifuged at 15,000
737 x g for 5 min at 4˚C. The lysate supernatant was incubated with 0.25 ml NeutrAvidin Agarose
738 resin for 30 min at 20˚C on an end-over-end rotator. The resin was washed with Wash Buffer,
739 and bound proteins eluted Elution Buffer from the kit

740 Samples were prepared for mass spectrometry with the EasyPep Mini MS Sample Prep
741 Kit (ThermoFisher, A40006) following the manufacturer's instructions. Briefly, proteins were
742 extracted, reduced, alkylated, and digested with trypsin/Lys-C protease. Hydrophilic and
743 hydrophobic contaminants were removed with a peptide clean-up step. LC-MS/MS analysis of
744 peptide samples was performed on an Ultimate 3000 rapid separation liquid chromatography
745 (RSLC) system connected to a Q-Exactive plus mass spectrometer (ThermoFisher). The
746 acquired raw files were processed with MaxQuant v1.6.6.0.

747 **Western blotting.** The cell surface biotinylated protein samples purified from HEK293T cells as
748 described above were heated to 95˚C in SDS sample buffer for 10 min. 15 μl of each sample
749 was run on a polyacrylamide gel. Gels were blotted onto PVDF membranes (Merck). Blots were
750 blocked in 5% milk in PBS, 0.2% Tween-20 and incubated overnight with anti-HA rabbit

751 monoclonal antibody (Cell Signaling Technology, 3724, RRID:AB_1549585) diluted 1:1,000 in
752 blocking solution. Blots were imaged with the 800 nm channel of a LI-COR Odyssey fluorescent
753 scanner after incubation with anti-rabbit DyLight 800-conjugated secondary antibody (Cell
754 Signaling Technology, 5151, RRID:AB_10697505) at 1:5,000 dilution for 30 min at room
755 temperature.

756 **RNA-Seq analysis.** We analysed data published by Hawdon and colleagues (*74*).
757 ArrayExpress run accession numbers and corresponding life cycle stages were: SRR6359160,
758 L3; SRR6359161 and SRR6359163, L4; SRR6359164 and SRR6359165, Adult mixed pooled
759 worms; SRR6359162 and SRR6359166, Adult male pooled worms. Our mapping and counting
760 strategy was modified as recommended for repetitive genomic features (*90*). Reads were
761 mapped against the reference *A. ceylanicum* genome (*91*) with STAR v2.7.5a (*92*) with
762 parameters: "--outSAMtype BAM SortedByCoordinate --runMode alignReads --
763 outFilterMultimapNmax 1000 --outSAMmultNmax 1 --outFilterMismatchNmax 3 --
764 outMultimapperOrder Random --winAnchorMultimapNmax 1000 --alignEndsType EndToEnd --
765 alignIntronMax 1 --alignMatesGapMax 350". FeatureCounts v2.0.1 (*93*) with parameters “-M -F
766 GFF -s 0 -p -t exon -g gene_id” was used to count reads over a modified annotation file based
767 on the original annotation (*91*) and downloaded from Wormbase Parasite (*94*). The original
768 annotation file in GFF format was slightly modified to contain a “gene_id” feature in the ninth
769 column and hence facilitate the calculation of aggregate reads by gene (for more details see
770 header of the GFF file, **dataset S4**). The resulting table of counts was imported into R (*95*) and
771 processed for differential expression analysis using DESeq2 v1.26.0 (*96*). Plots were generated
772 using ggplot2 (*97*) and karyoploteR (*98*). RNA-Seq analysis scripts and instructions for how to
773 use them are provided in the Supplementary Materials (**dataset S5**) and are additionally
774 available on Github [https://github.com/annaprotasio/Merchant_et_al_2020].

775 **Statistics.** Error bars represent the standard deviation or standard error –as indicated in the
776 respective figure legend– of 2-7 replicates conducted across at least two independent
777 experiments. SDS-PAGE gels and DLS data shown are representative of at least two
778 independent experiments. Significance and *P*-values were determined by 2-way ANOVA. For
779 DLS data, Sidak's multiple comparisons test was used with a 95% confidence interval, in Prism
780 8 (GraphPad). For cell-cell fusion data, Tukey's multiple comparisons test was used in Prism 9
781 (GraphPad). Source data are provided in **datasets S1 and S2.** No statistical methods were
782 used to predetermine sample size, experiments were not randomized, and the investigators
783 were not blinded to experimental outcomes.

784 **Data availability**

785 All data needed to evaluate the conclusions in the paper are present in the paper and/or the
786 Supplementary Materials. The atomic coordinates were deposited in the Protein Data Bank with
787 accession code 7A4A [https://doi.org/10.2210/pdb7A4A/pdb]. The cryo-EM density was
788 deposited in the EM Data Bank with code EMD-11630
789 [https://www.ebi.ac.uk/pdbe/entry/emdb/EMD-11630]. The raw electron micrographs were
790 deposited in EMPIAR with code EMPIAR-10266 [https://doi.org/10.6019/EMPIAR-10266]. RNA-
791 Seq analysis scripts and instructions for how to use them are provided in the Supplementary
792 Materials (**dataset S5**) and are additionally available on Github
793 [https://github.com/annaprotasio/Merchant_et_al_2020].

794 **References**

795 1. J. A. Frank, C. Feschotte, Co-option of endogenous viral sequences for host cell function.
796 *Curr Opin Virol* **25**, 81-89 (2017).
797 2. A. Katzourakis, R. J. Gifford, Endogenous viral elements in animal genomes. *PLoS genetics*
798 **6**, e1001191 (2010).

799 3. M. Friedli, D. Trono, The developmental control of transposable elements and the evolution
800 of higher species. *Annu Rev Cell Dev Biol* **31**, 429-451 (2015).
801 4. E. B. Chuong, N. C. Elde, C. Feschotte, Regulatory evolution of innate immunity through
802 co-option of endogenous retroviruses. *Science* **351**, 1083-1087 (2016).
803 5. A. Dupressoir, C. Lavialle, T. Heidmann, From ancestral infectious retroviruses to bona fide
804 cellular genes: role of the captured syncytins in placentation. *Placenta* **33**, 663-671 (2012).
805 6. R. S. Moore *et al.*, The role of the Cer1 transposon in horizontal transfer of
806 transgenerational memory. *Cell* **184**, 4697-4712 e4618 (2021).
807 7. T. Gojobori, S. Yokoyama, Rates of evolution of the retroviral oncogene of Moloney murine
808 sarcoma virus and of its cellular homologues. *Proc Natl Acad Sci U S A* **82**, 4198-4201
809 (1985).
810 8. K. R. McCarthy *et al.*, The ancient fusogen EnvP(b)1 is expressed in human tissues and its
811 structure informs the evolution of gammaretrovirus envelope proteins. *bioRxiv*,
812 2020.2004.2022.056234 (2020).
813 9. D. Kremer *et al.*, Human endogenous retrovirus type W envelope protein inhibits
814 oligodendroglial precursor cell differentiation. *Ann Neurol* **74**, 721-732 (2013).
815 10. W. Li *et al.*, Human endogenous retrovirus-K contributes to motor neuron disease. *Sci*
816 *Transl Med* **7**, 307ra153 (2015).
817 11. M. Krupovic *et al.*, Ortervirales: New Virus Order Unifying Five Families of Reverse-
818 Transcribing Viruses. *J Virol* **92**, (2018).
819 12. I. G. Frame, J. F. Cutfield, R. T. Poulter, New BEL-like LTR-retrotransposons in Fugu
820 rubripes, Caenorhabditis elegans, and Drosophila melanogaster. *Gene* **263**, 219-230
821 (2001).
822 13. N. J. Bowen, J. F. McDonald, Genomic analysis of Caenorhabditis elegans reveals ancient
823 families of retroviral-like elements. *Genome Res* **9**, 924-935 (1999).

824 14. H. S. Malik, S. Henikoff, T. H. Eickbush, Poised for contagion: evolutionary origins of the
825 infectious abilities of invertebrate retroviruses. *Genome Res* **10**, 1307-1318 (2000).
826 15. D. Fass, S. C. Harrison, P. S. Kim, Retrovirus envelope domain at 1.7 angstrom resolution.
827 *Nat Struct Biol* **3**, 465-469 (1996).
828 16. D. C. Chan, D. Fass, J. M. Berger, P. S. Kim, Core structure of gp41 from the HIV envelope
829 glycoprotein. *Cell* **89**, 263-273 (1997).
830 17. W. Weissenhorn, A. Dessen, S. C. Harrison, J. J. Skehel, D. C. Wiley, Atomic structure of
831 the ectodomain from HIV-1 gp41. *Nature* **387**, 426-430 (1997).
832 18. B. Kobe, R. J. Center, B. E. Kemp, P. Poumbourios, Crystal structure of human T cell
833 leukemia virus type 1 gp21 ectodomain crystallized as a maltose-binding protein chimera
834 reveals structural evolution of retroviral transmembrane proteins. *Proc Natl Acad Sci U S A*
835 **96**, 4319-4324 (1999).
836 19. S. C. Harrison, Viral membrane fusion. *Virology* **479-480**, 498-507 (2015).
837 20. M. Dessau, Y. Modis, Crystal structure of glycoprotein C from Rift Valley fever virus. *Proc.*
838 *Natl. Acad. Sci. U. S. A.* **110**, 1696-1701 (2013).
839 21. Y. Modis, Relating Structure to Evolution in Class II Viral Membrane Fusion Proteins. *Curr*
840 *Opin Virol*, DOI:10.1016/j.coviro.2014.1001.1009 (2014).
841 22. J. Lescar *et al.*, The Fusion glycoprotein shell of Semliki Forest virus: an icosahedral
842 assembly primed for fusogenic activation at endosomal pH. *Cell* **105**, 137-148 (2001).
843 23. F. A. Rey, F. X. Heinz, C. Mandl, C. Kunz, S. C. Harrison, The envelope glycoprotein from
844 tick-borne encephalitis virus at 2 A resolution. *Nature* **375**, 291-298 (1995).
845 24. Y. Modis, S. Ogata, D. Clements, S. C. Harrison, A ligand-binding pocket in the dengue
846 virus envelope glycoprotein. *Proc Natl Acad Sci U S A* **100**, 6986-6991 (2003).
847 25. R. M. DuBois *et al.*, Functional and evolutionary insight from the crystal structure of rubella
848 virus protein E1. *Nature* **493**, 552-556 (2013).

849 26. Y. Modis, S. Ogata, D. Clements, S. C. Harrison, Structure of the dengue virus envelope
850 protein after membrane fusion. *Nature* **427**, 313-319 (2004).
851 27. D. L. Gibbons *et al.*, Conformational change and protein-protein interactions of the fusion
852 protein of Semliki Forest virus. *Nature* **427**, 320-325 (2004).
853 28. P. Guardado-Calvo *et al.*, A glycerophospholipid-specific pocket in the RVFV class II fusion
854 protein drives target membrane insertion. *Science* **358**, 663-667 (2017).
855 29. J. Perez-Vargas *et al.*, Structural basis of eukaryotic cell-cell fusion. *Cell* **157**, 407-419
856 (2014).
857 30. O. Avinoam *et al.*, Conserved eukaryotic fusogens can fuse viral envelopes to cells.
858 *Science* **332**, 589-592 (2011).
859 31. A. Sapir *et al.*, AFF-1, a FOS-1-regulated fusogen, mediates fusion of the anchor cell in C.
860 elegans. *Dev Cell* **12**, 683-698 (2007).
861 32. B. Podbilewicz *et al.*, The C. elegans developmental fusogen EFF-1 mediates homotypic
862 fusion in heterologous cells and in vivo. *Dev Cell* **11**, 471-481 (2006).
863 33. J. Fedry *et al.*, The Ancient Gamete Fusogen HAP2 Is a Eukaryotic Class II Fusion Protein.
864 *Cell* **168**, 904-915 e910 (2017).
865 34. J. Fedry *et al.*, Evolutionary diversification of the HAP2 membrane insertion motifs to drive
866 gamete fusion across eukaryotes. *PLoS Biol* **16**, e2006357 (2018).
867 35. J. Feng *et al.*, Fusion surface structure, function, and dynamics of gamete fusogen HAP2.
868 *eLife* **7**, (2018).
869 36. C. Valansi *et al.*, Arabidopsis HAP2/GCS1 is a gamete fusion protein homologous to
870 somatic and viral fusogens. *J Cell Biol* **216**, 571-581 (2017).
871 37. T. Mori, H. Kuroiwa, T. Higashiyama, T. Kuroiwa, GENERATIVE CELL SPECIFIC 1 is
872 essential for angiosperm fertilization. *Nat Cell Biol* **8**, 64-71 (2006).

873 38. K. von Besser, A. C. Frank, M. A. Johnson, D. Preuss, Arabidopsis HAP2 (GCS1) is a
874 sperm-specific gene required for pollen tube guidance and fertilization. *Development* **133**,
875 4761-4769 (2006).
876 39. T. Clark, HAP2/GCS1: Mounting evidence of our true biological EVE? *PLoS Biol* **16**,
877 e3000007 (2018).
878 40. S. F. Altschul *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database
879 search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).
880 41. O. Poch, I. Sauvaget, M. Delarue, N. Tordo, Identification of four conserved motifs among
881 the RNA-dependent polymerase encoding elements. *EMBO J* **8**, 3867-3874 (1989).
882 42. H. Felder *et al.*, Tas, a retrotransposon from the parasitic nematode Ascaris lumbricoides.
883 *Gene* **149**, 219-225 (1994).
884 43. M. Krupovic, E. V. Koonin, Homologous Capsid Proteins Testify to the Common Ancestry of
885 Retroviruses, Caulimoviruses, Pseudoviruses, and Metaviruses. *J Virol* **91**, (2017).
886 44. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL) v4: recent updates and new
887 developments. *Nucleic Acids Res* **47**, W256-W259 (2019).
888 45. S. Halldorsson *et al.*, Structure of a phleboviral envelope glycoprotein reveals a
889 consolidated model of membrane fusion. *Proc Natl Acad Sci U S A* **113**, 7154-7159 (2016).
890 46. Y. Zhu *et al.*, The Postfusion Structure of the Heartland Virus Gc Glycoprotein Supports
891 Taxonomic Separation of the Bunyaviral Families Phenuiviridae and Hantaviridae. *J Virol*
892 **92**, (2018).
893 47. D. J. Leahy, W. A. Hendrickson, I. Aukhil, H. P. Erickson, Structure of a fibronectin type III
894 domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science*
895 **258**, 987-991 (1992).
896 48. J. F. Pinello *et al.*, Structure-Function Studies Link Class II Viral Fusogens with the
897 Ancestral Gamete Fusion Protein HAP2. *Curr Biol* **27**, 651-660 (2017).

898 49. M. Kielian, F. A. Rey, Virus membrane-fusion proteins: more than one way to make a
899 hairpin. *Nat. Rev. Microbiol.* **4**, 67-76 (2006).
900 50. L. Holm, P. Rosenstrom, Dali server: conservation mapping in 3D. *Nucleic Acids Res* **38**,
901 W545-549 (2010).
902 51. D. E. Klein, J. L. Choi, S. C. Harrison, Structure of a dengue virus envelope protein late-
903 stage fusion intermediate. *J Virol* **87**, 2287-2293 (2013).
904 52. S. Murakami, K. Terasaki, S. I. Ramirez, J. C. Morrill, S. Makino, Development of a novel,
905 single-cycle replicable rift valley Fever vaccine. *PLoS neglected tropical diseases* **8**, e2746
906 (2014).
907 53. E. Zaitseva, S. T. Yang, K. Melikov, S. Pourmal, L. V. Chernomordik, Dengue virus ensures
908 its fusion in late endosomes using compartment-specific lipids. *PLOS Pathog.* **6**, e1001131
909 (2010).
910 54. P. Y. Lozach *et al.*, Entry of bunyaviruses into mammalian cells. *Cell Host Microbe* **7**, 488-
911 499 (2010).
912 55. A. M. Nour, Y. Li, J. Wolenski, Y. Modis, Viral membrane fusion and nucleocapsid delivery
913 into the cytoplasm are distinct events in some flaviviruses. *PLoS Pathog* **9**, e1003585
914 (2013).
915 56. Y. Kubo, H. Hayashi, T. Matsuyama, H. Sato, N. Yamamoto, Retrovirus entry by
916 endocytosis and cathepsin proteases. *Adv Virol* **2012**, 640894 (2012).
917 57. D. Bitto, S. Halldorsson, A. Caputo, J. T. Huiskonen, Low pH and Anionic Lipid-dependent
918 Fusion of Uukuniemi Phlebovirus to Liposomes. *J Biol Chem* **291**, 6412-6422 (2016).
919 58. D. L. Esposito, J. B. Nguyen, D. C. DeWitt, E. Rhoades, Y. Modis, Physico-chemical
920 requirements and kinetics of membrane fusion of flavivirus-like particles. *J Gen Virol* **96**,
921 1702-1711 (2015).
922 59. M. Umashankar *et al.*, Differential cholesterol binding by class II fusion proteins determines
923 membrane fusion properties. *J Virol* **82**, 9245-9253 (2008).

924 60. P. K. Chatterjee, C. H. Eng, M. Kielian, Novel mutations that control the sphingolipid and
925 cholesterol dependence of the Semliki Forest virus fusion protein. *J Virol* **76**, 12712-12722
926 (2002).
927 61. J. M. Smit, R. Bittman, J. Wilschut, Low-pH-dependent fusion of Sindbis virus with receptor-
928 free cholesterol- and sphingolipid-containing liposomes. *J Virol* **73**, 8476-8484 (1999).
929 62. T. V. Kurzchalia, S. Ward, Why do worms need cholesterol? *Nat Cell Biol* **5**, 684-688
930 (2003).
931 63. M. Merris, J. Kraeft, G. S. Tint, J. Lenard, Long-term effects of sterol depletion in C.
932 elegans: sterol content of synchronized wild-type and mutant populations. *J Lipid Res* **45**,
933 2044-2051 (2004).
934 64. M. Ruiz *et al.*, Membrane fluidity is regulated by the C. elegans transmembrane protein
935 FLD-1 and its human homologs TLCD1/2. *eLife* **7**, (2018).
936 65. K. Stiasny, S. L. Allison, J. Schalich, F. X. Heinz, Membrane interactions of the tick-borne
937 encephalitis virus fusion protein E at low pH. *J. Virol.* **76**, 3784-3790 (2002).
938 66. D. L. Gibbons, A. Ahn, P. K. Chatterjee, M. Kielian, Formation and characterization of the
939 trimeric form of the fusion protein of Semliki Forest Virus. *J Virol* **74**, 7772-7780 (2000).
940 67. R. Fritz, K. Stiasny, F. X. Heinz, Identification of specific histidines as pH sensors in
941 flavivirus membrane fusion. *J Cell Biol* **183**, 353-361 (2008).
942 68. V. Nayak *et al.*, Crystal structure of dengue virus type 1 envelope protein in the postfusion
943 conformation and its implications for membrane fusion. *J. Virol.* **83**, 4338-4344 (2009).
944 69. Z. L. Qin, Y. Zheng, M. Kielian, Role of conserved histidine residues in the low-pH
945 dependence of the Semliki Forest virus fusion protein. *J Virol* **83**, 4670-4677 (2009).
946 70. Y. Zheng, C. Sanchez-San Martin, Z. L. Qin, M. Kielian, The domain I-domain III linker
947 plays an important role in the fusogenic conformational change of the alphavirus membrane
948 fusion protein. *J Virol* **85**, 6334-6342 (2011).

949 71. S. M. de Boer *et al.*, Acid-activated structural reorganization of the Rift Valley fever virus Gc
950 fusion protein. *J Virol* **86**, 13642-13652 (2012).
951 72. M. A. Vega, J. L. Strominger, Constitutive endocytosis of HLA class I antigens requires a
952 specific portion of the intracytoplasmic tail that shares structural features with other
953 endocytosed molecules. *Proc Natl Acad Sci U S A* **86**, 2688-2692 (1989).
954 73. R. DeMarco *et al.*, Saci-1, -2, and -3 and Perere, four novel retrotransposons with high
955 transcriptional activities from the human parasite Schistosoma mansoni. *J Virol* **78**, 2967-
956 2978 (2004).
957 74. J. P. Bernot *et al.*, Transcriptomic analysis of hookworm Ancylostoma ceylanicum life cycle
958 stages reveals changes in G-protein coupled receptor diversity associated with the onset of
959 parasitism. *Int J Parasitol* **50**, 603-610 (2020).
960 75. D. Moi *et al.*, Archaeal origins of gamete fusion. *bioRxiv*, 2021.2010.2013.464100 (2021).
961 76. R. S. Moore *et al.*, Horizontal and vertical transmission of transgenerational memories via
962 the *Cer1* transposon. *bioRxiv*, 2020.2012.2028.424563 (2020).
963 77. S. Kurtz *et al.*, REPuter: the manifold applications of repeat analysis on a genomic scale.
964 *Nucleic Acids Res* **29**, 4633-4642 (2001).
965 78. B. M. Muhire, A. Varsani, D. P. Martin, SDT: a virus classification tool based on pairwise
966 sequence alignment and identity calculation. *PLoS One* **9**, e108277 (2014).
967 79. S. Q. Zheng *et al.*, MotionCor2: anisotropic correction of beam-induced motion for improved
968 cryo-electron microscopy. *Nat Methods* **14**, 331-332 (2017).
969 80. K. Zhang, Gctf: Real-time CTF determination and correction. *J Struct Biol* **193**, 1-12 (2016).
970 81. S. He, S. H. W. Scheres, Helical reconstruction in RELION. *J Struct Biol* **198**, 163-176
971 (2017).
972 82. S. H. Scheres, RELION: implementation of a Bayesian approach to cryo-EM structure
973 determination. *J Struct Biol* **180**, 519-530 (2012).

974 83. J. Zivanov, T. Nakane, S. H. W. Scheres, A Bayesian approach to beam-induced motion
975 correction in cryo-EM single-particle analysis. *IUCrJ* **6**, 5-17 (2019).
976 84. E. F. Pettersen *et al.*, UCSF Chimera--a visualization system for exploratory research and
977 analysis. *J Comput Chem* **25**, 1605-1612 (2004).
978 85. P. D. Adams *et al.*, PHENIX: a comprehensive Python-based system for macromolecular
979 structure solution. *Acta Crystallogr D Biol Crystallogr* **66**, 213-221 (2010).
980 86. P. Emsley, K. Cowtan, Coot: model-building tools for molecular graphics. *Acta Crystallogr D*
981 *Biol Crystallogr* **60**, 2126-2132 (2004).
982 87. A. Amunts *et al.*, Structure of the yeast mitochondrial large ribosomal subunit. *Science* **343**,
983 1485-1489 (2014).
984 88. V. B. Chen *et al.*, MolProbity: all-atom structure validation for macromolecular
985 crystallography. *Acta Crystallogr D Biol Crystallogr* **66**, 12-21 (2010).
986 89. J. Schindelin *et al.*, Fiji: an open-source platform for biological-image analysis. *Nat Methods*
987 **9**, 676-682 (2012).
988 90. A. Teissandier, N. Servant, E. Barillot, D. Bourc'his, Tools and best practices for
989 retrotransposon analysis using high-throughput sequencing data. *Mob DNA* **10**, 52 (2019).
990 91. E. M. Schwarz *et al.*, The genome and transcriptome of the zoonotic hookworm
991 Ancylostoma ceylanicum identify infection-specific gene families. *Nat Genet* **47**, 416-422
992 (2015).
993 92. A. Dobin, T. R. Gingeras, Mapping RNA-seq Reads with STAR. *Curr Protoc Bioinformatics*
994 **51**, 11.14.11-11.14.19 (2015).
995 93. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for
996 assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).
997 94. K. L. Howe, B. J. Bolt, M. Shafie, P. Kersey, M. Berriman, WormBase ParaSite - a
998 comprehensive resource for helminth genomics. *Molecular and biochemical parasitology*
999 **215**, 2-10 (2017).

1000 95. R. C. Team. (R Foundation for Statistical Computing, Vienna, Austria, 2019).
1001 96. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for
1002 RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
1003 97. H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag, New York,
1004 2016).
1005 98. B. Gel, E. Serra, karyoploteR: an R/Bioconductor package to plot customizable genomes
1006 displaying arbitrary data. *Bioinformatics* **33**, 3088-3090 (2017).

1007 **Acknowledgements**

1008 We thank Sew-Yeu Peak-Chew at the MRC Laboratory of Molecular Biology (LMB) Mass
1009 Spectrometry Facility for assistance with mass spectrometry data collection and analysis. We
1010 acknowledge the MRC-LMB Electron Microscopy (EM) Facility and UK national electron Bio-
1011 Imaging Centre (eBIC) at Diamond Light Source (DLS) for access and support of EM sample
1012 preparation and data collection. We acknowledge DLS for access and support of eBIC under
1013 proposals EM17434 and EM20121 funded by the Wellcome Trust, MRC and BBSRC. We are
1014 grateful to Kyle Dent for aiding in using the microscopes at eBIC. We are grateful for access to
1015 the MRC-LMB Scientific Computing facilities. This work was supported by Senior Research
1016 Fellowships 101908/Z/13/Z and 217191/Z/19/Z from the Wellcome Trust and NIH grant R01
1017 GM102869-01 to Y.M.

1018 **Author contributions**

1019 MM and YM conceived the experiments. MM purified the proteins and performed the
1020 biochemical assays. CPM collected and processed the cryo-EM data and performed the image
1021 reconstruction. CPM and YM built and refined the atomic model. YL and HZ carried out the cell-
1022 cell fusion assay. AVP performed the RNA-Seq analysis. All authors contributed to the figures.
1023 YM wrote the manuscript. Project supervision, administration, and funding acquisition, YM.

1024 **Competing interests**

1025 YM is a consultant for Related Sciences LLC and has profits interests in Danger Bio LLC. The
1026 authors declare no other competing interests.

1027 **Materials & Correspondence**

1028 Correspondence and material requests should be addressed to Y.M. (ymodis@mrc-
1029 lmb.cam.ac.uk).

1030 # A bioactive phlebovirus-like envelope protein in a hookworm

1031 # endogenous virus

1032 Monique Merchant, Carlos P. Mata, Yangci Liu, Haoming Zhai, Anna V. Protasio,

1033 Yorgo Modis*

1034 *Corresponding author. Email: ymodis@mrc-lmb.cam.ac.uk

1035

1036 **Supplementary Materials**

1037 **Figures S1 to S10**

1038 **Table S1**

1039 **Movie S1**

1040 **Datasets S1 to S5**

1041

1042

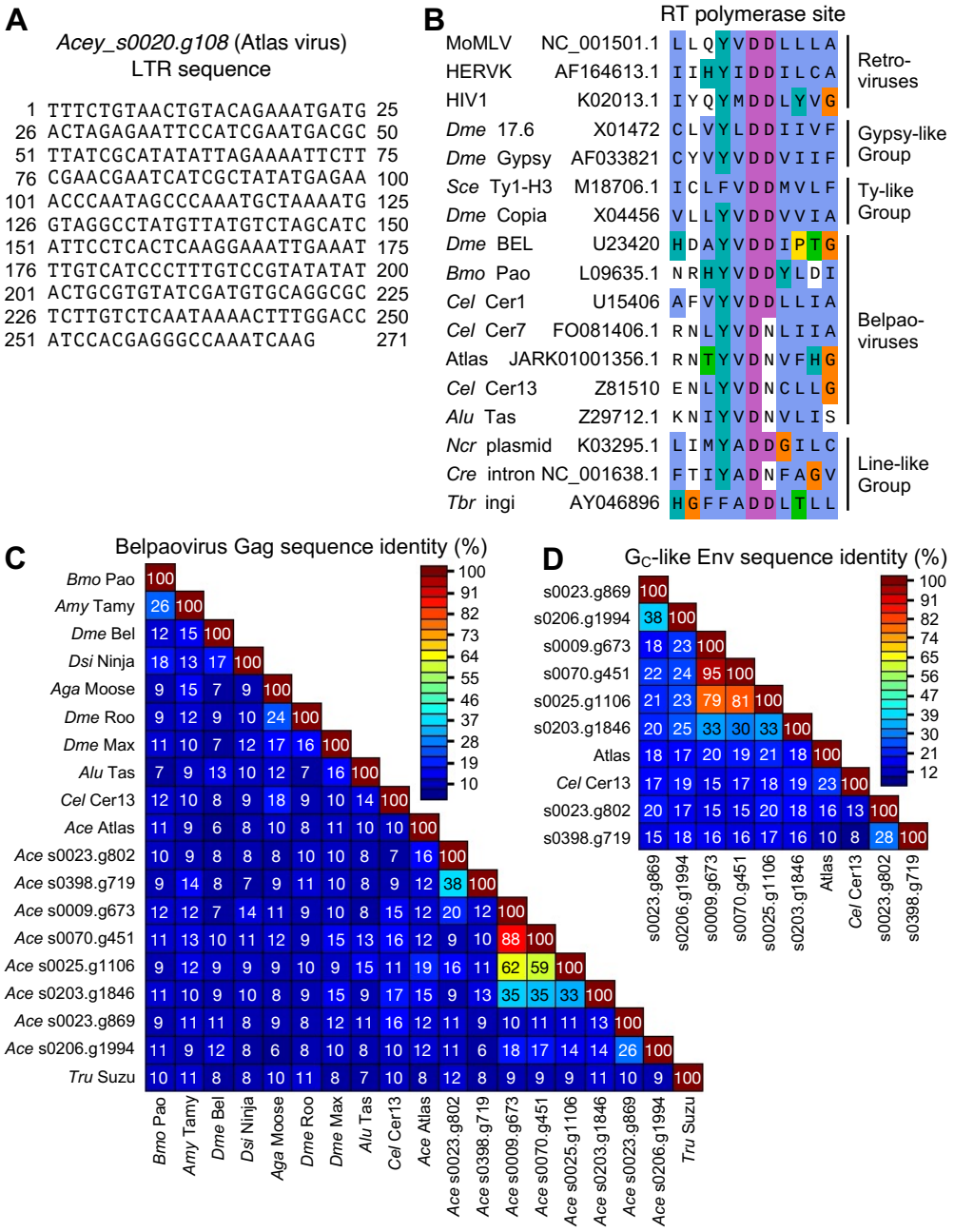


Fig. S1. Atlas is an endogenous belpaovirus with a phlebovirus-like Env. (**A**) Nucleotide sequence of the Atlas LTRs, predicted by REPuter (*76*). The LTRs are 100% identical. (**B**) Aligned protein sequences of reverse transcriptase (RT) polymerase site from Atlas and other representative belpao- and retroviruses. *A. ceylanicum* belpaoviruses have an Asp to Asn substitution (Y[X]DD to YVDN) in the polymerase site, the most conserved RT motif (Motif V, or Motif C). GenBank accession numbers are listed after virus names. MoMLV, Moloney murine leukemia virus; HERVK, human ERV-K; *Dme*, *Drosophila melanogaster*; *Sce, Saccharomyces cerevisiae*; *Bmo, Bombyx mori*; *Cel, C. elegans*; *Ace, A. ceylanicum*; *Alu, Ascaris lumbricoides*; *Ncr, Neurospora crassa*; *Cre, C. reinhardtii*; *Tbr, Trypanosoma brucei*. (**C**) Sequence identity matrix for Gag protein sequences from representative belpaoviruses and *A. ceylanicum* belpaoviruses with intact protein coding regions. *Amy, Antheraea mylitta; Dsi, Drosophila simulans; Aga, Anopheles gambiae; Tru, Takifugu rubripes.* (**D**) Sequence identity matrix for phlebovirus-like Env protein sequences from nematode belpaoviruses. Identity matrices were calculated with SDT (77).

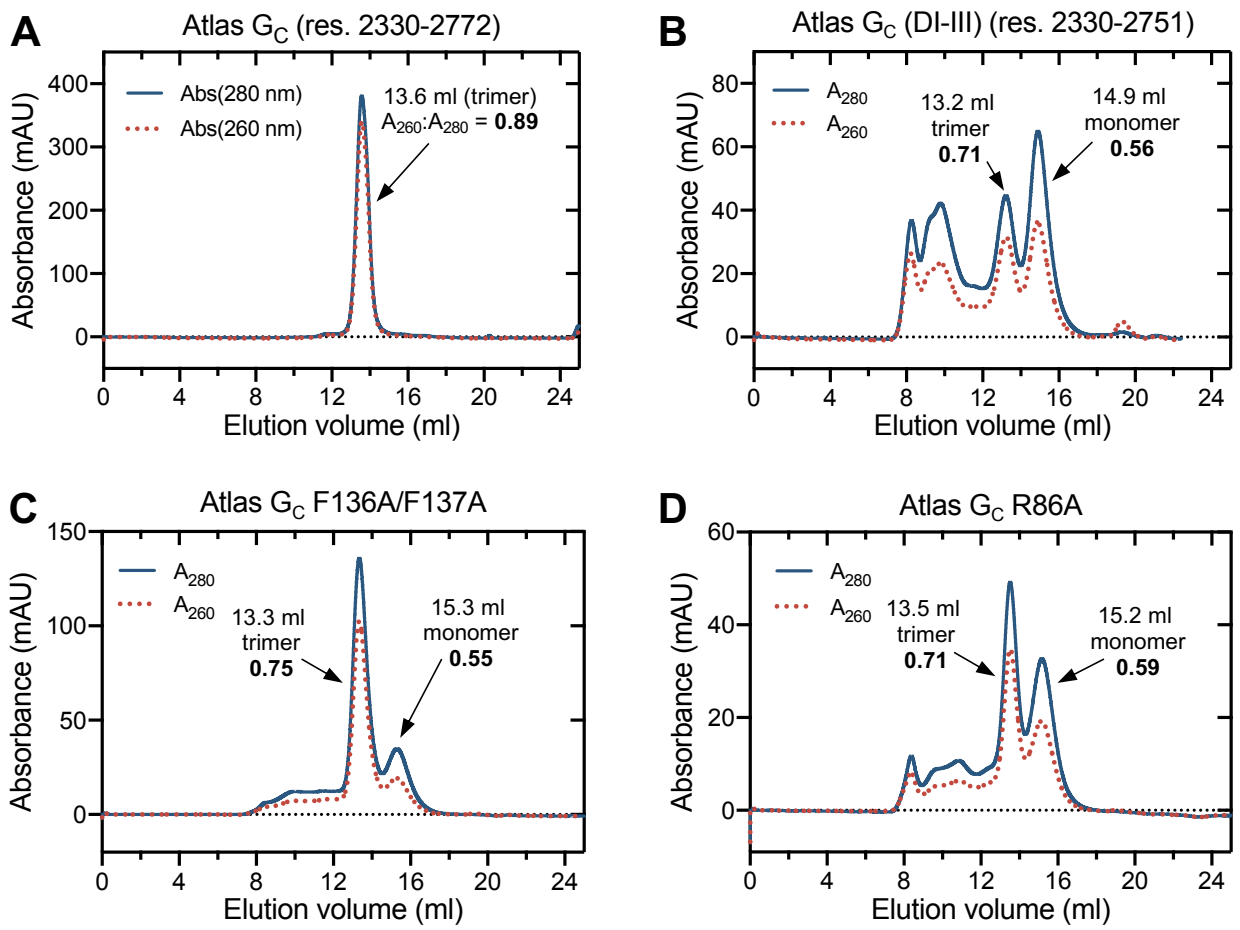


Fig. S2. Size-exclusion chromatography of Atlas G_C following ion-exchange chromatography. Samples were analysed on a Superdex 200 Increase (10/300) column (Cytiva). The A_{260}:A_{280} ratios for selected peaks are shown in bold. (**A**) The Atlas G_C fragment used for cryo-EM image reconstruction (residues 2330-2772). The protein had higher absorbance at 260 nm (A_{260}) than expected for pure protein. The elution volume was consistent with a homotrimer. (**B**) Atlas G_C(DI-III), maintained at pH > 7 throughout expression and purification. Multiple oligomeric states were present including peaks with elution volumes consistent with monomers, trimers, and higher order oligomers. (**C**) Atlas G_C fusion loop mutant F136A/F137A. (**D**) Atlas G_C GPL binding pocket mutant R86A.

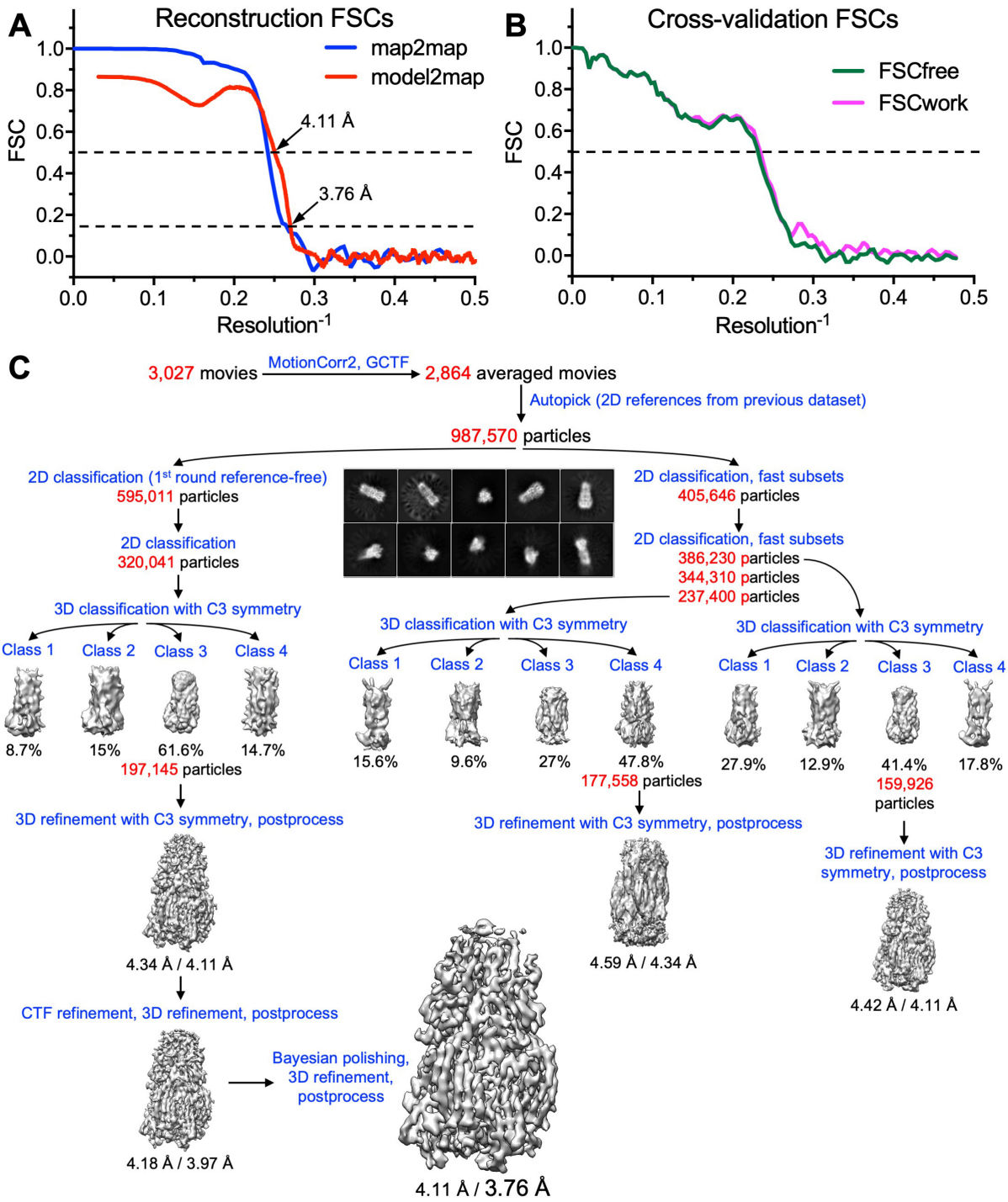


Fig. S3. Image reconstruction quality assessment and workflow. (**A**) Fourier Shell Correlations (FSC) of the reconstruction of Atlas G_C from two independently refined half-maps (map2map, in blue); and of the reconstruction from the whole dataset versus a map calculated from the refined atomic model (model2map, in red). The gold-standard cutoff (FSC = 0.143) and the FSC = 0.5 level are marked with dashed lines. The resolution values of each curve at these levels are indicated. (**B**) FSC plots for cross-validation as described by Amunts *et al.* (*86*). FSCwork (magenta), FSC of refined test model versus work set (half-map used in test refinement). FSCfree (green), FSC of refined test model versus test set (half-map not used in test refinement). The FSC = 0.5 level is indicated by a dashed line. (**C**) Flow chart showing the workflow pipeline for cryo-EM image processing, classification, and model refinement, as described in the Materials and Methods.

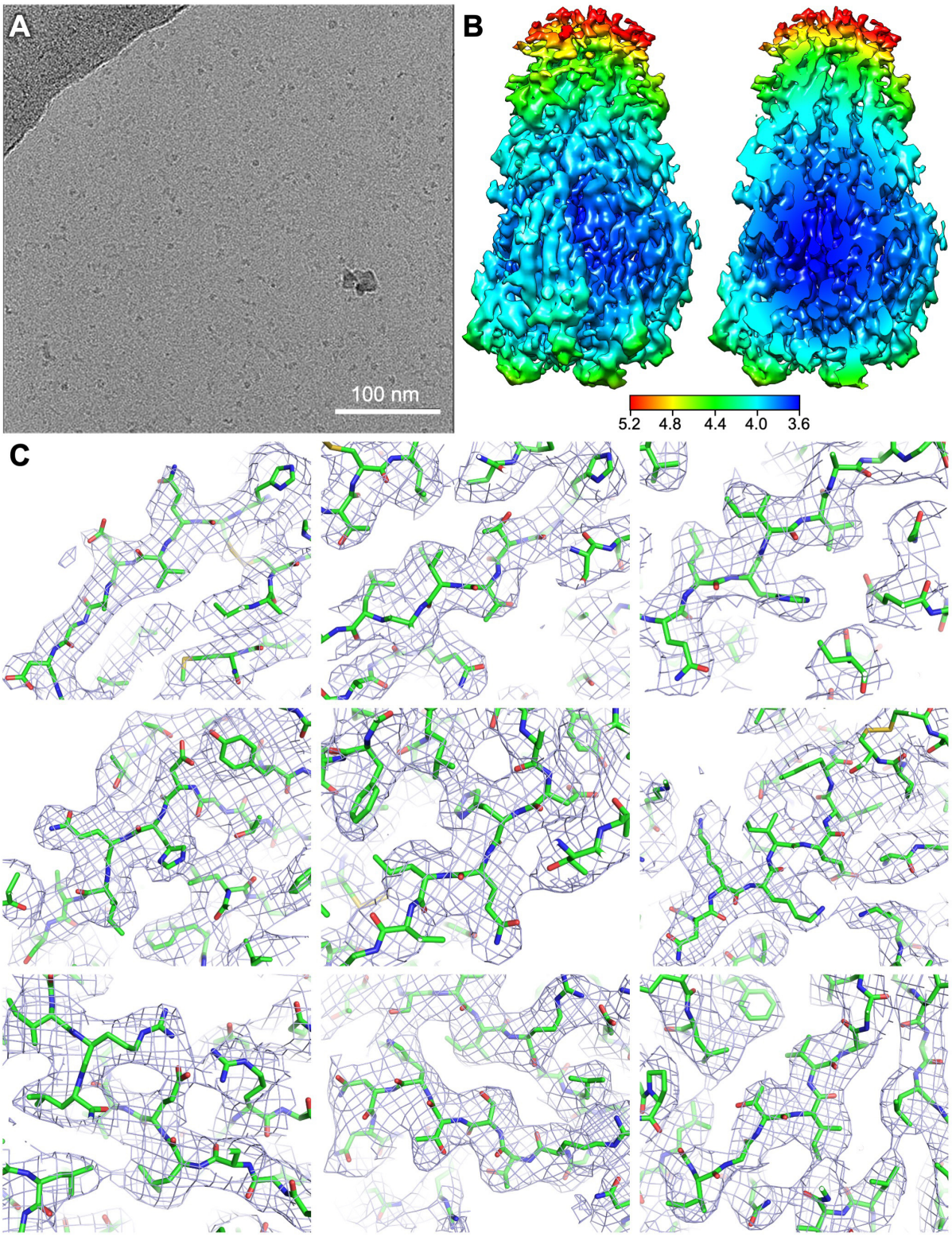


Fig. S4. Representative Cryo-EM micrograph and density with local resolution estimation (**A**) Representative cryo-EM micrograph of Atlas G_C trimers. (**B**) Local resolution estimation for the cryo-EM volume, calculated in RELION 3.0 (*82*). (**C**) Representative samples of local cryo-EM density from the structure with fitted and refined atomic models. The deposited density map was contoured at 3 σ in Pymol (Schrodinger, LLC).

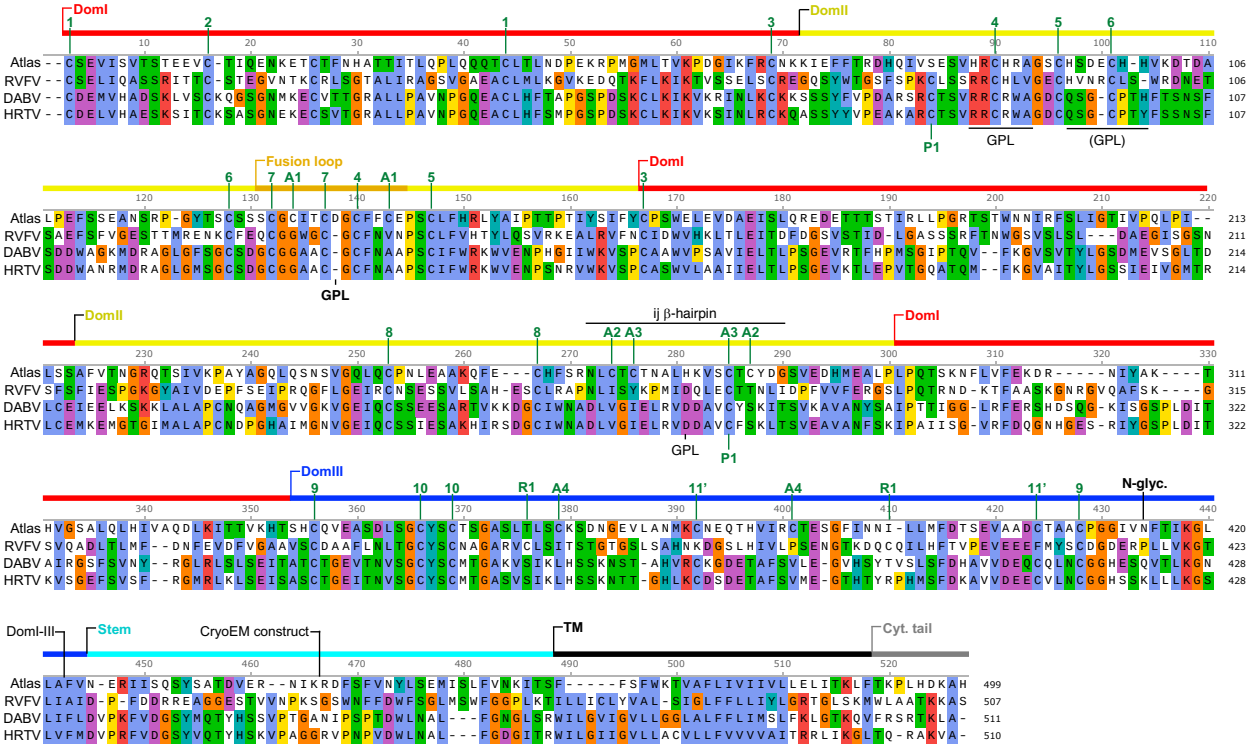


Fig. S5. Protein sequence alignment of Atlas G_C and its closest orthologs

Amino acid sequence alignment of Atlas G_C and the most similar sequences from infectious viruses: Rift Valley Fever virus (RVFV), Dabie bandavirus (DABV, formerly SFTSV phlebovirus) and Heartland virus (HRTV). Domains and structural features are marked with lines above the alignment. GPL, residues in or near the glycerophospholipid headgroup binding pocket. N-glyc., N-glycosylation site. Cyt. tail, cytoplasmic tail. Conserved disulfide bonds are numbered in green. Disulfide bonds specific to Atlas virus are denoted with an “A”. An RVFV-specific disulfide is denoted with an “R”. A phlebovirus-specific disulfide is denoted with a “P”. The disulfide marked 11’ is conserved in Atlas, DABV and HRTV but not RVFV.

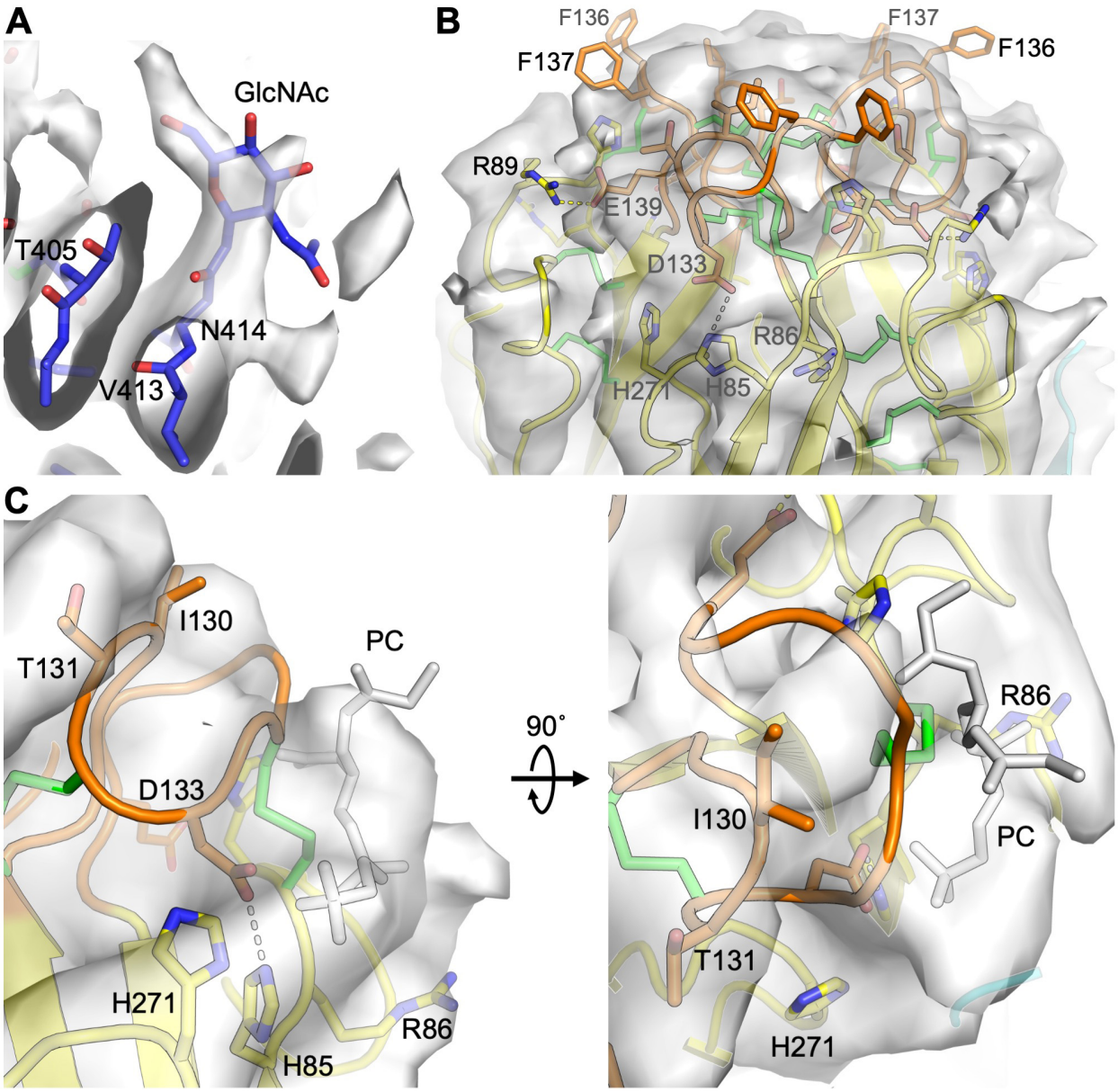


Fig. S6. Cryo-EM density for N-linked glycan and GPL pocket of Atlas G_C. (**A**) Cryo-EM density and atomic model of the N-linked glycan at Asn414. The deposited density map was contoured at 2.2 σ in Pymol (Schrodinger, LLC). (**B**) Cryo-EM density and atomic model for the fusion loop region of the Atlas G_C trimer. A B-factor blurring correction of +150 $Å^2$ was applied to the deposited density map. The resulting map was contoured at 1 σ in Pymol. (**C**) Cryo-EM density in the glycerophospholipid (GPL) headgroup binding pocket unaccounted for by the atomic model. Shown in grey stick representation is the phosphatidylcholine (PC) headgroup bound to RVFV G_C structure (*28*) superimposed on the Atlas G_C structure. A B-factor blurring correction of +163 $Å^2$ was applied to the deposited density map. The resulting map was contoured at 1.2 σ in Pymol.

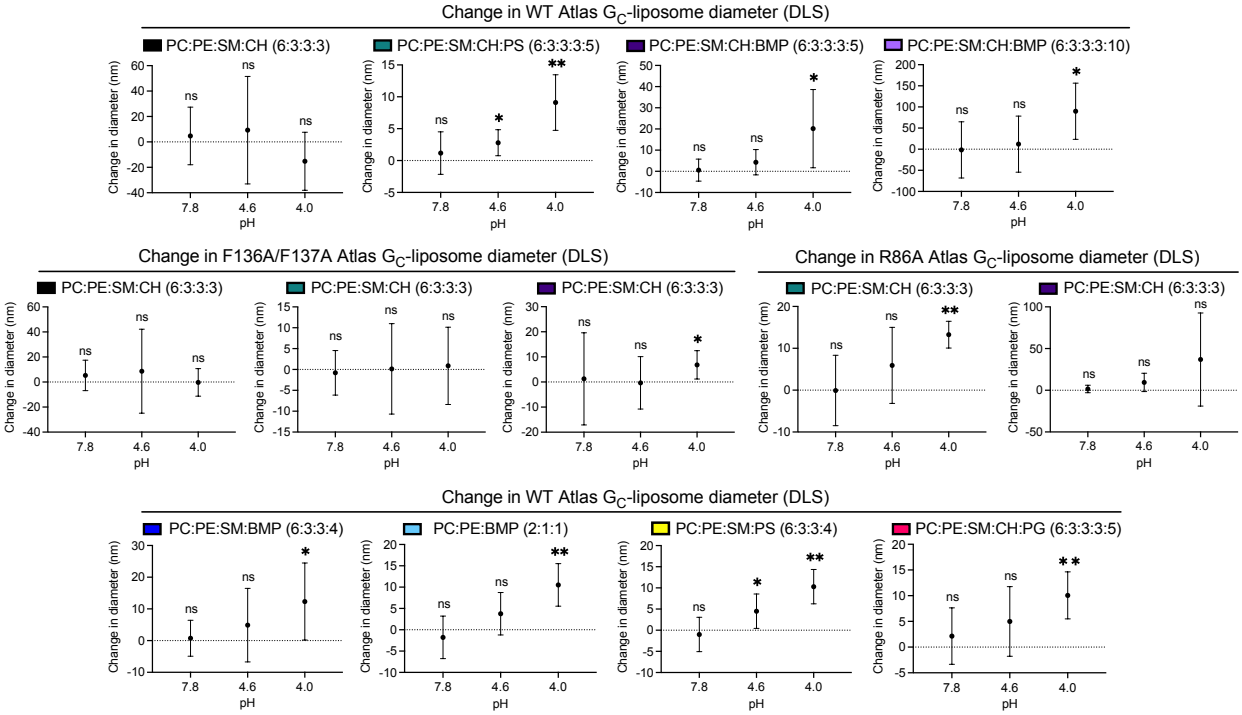


Fig. S7. Dynamic light scattering of Atlas G_C-liposome complexes. Liposome diameter dynamic light scattering (DLS) data from **Fig. 4B** are shown as the mean change in liposome diameter upon binding Atlas G_C ectodomain (WT, F136A/F137A or R86A). Error bars show the standard error of the mean (s.e.m.) of three to seven replicates (see **dataset S1** for source data). Significance was determined by 2-way ANOVA analysis using Sidak's multiple comparisons test with a 95% confidence interval, in Prism 8 (GraphPad). **, $0.001 < p < 0.01$; *, $p < 0.05$; ns, not significant.

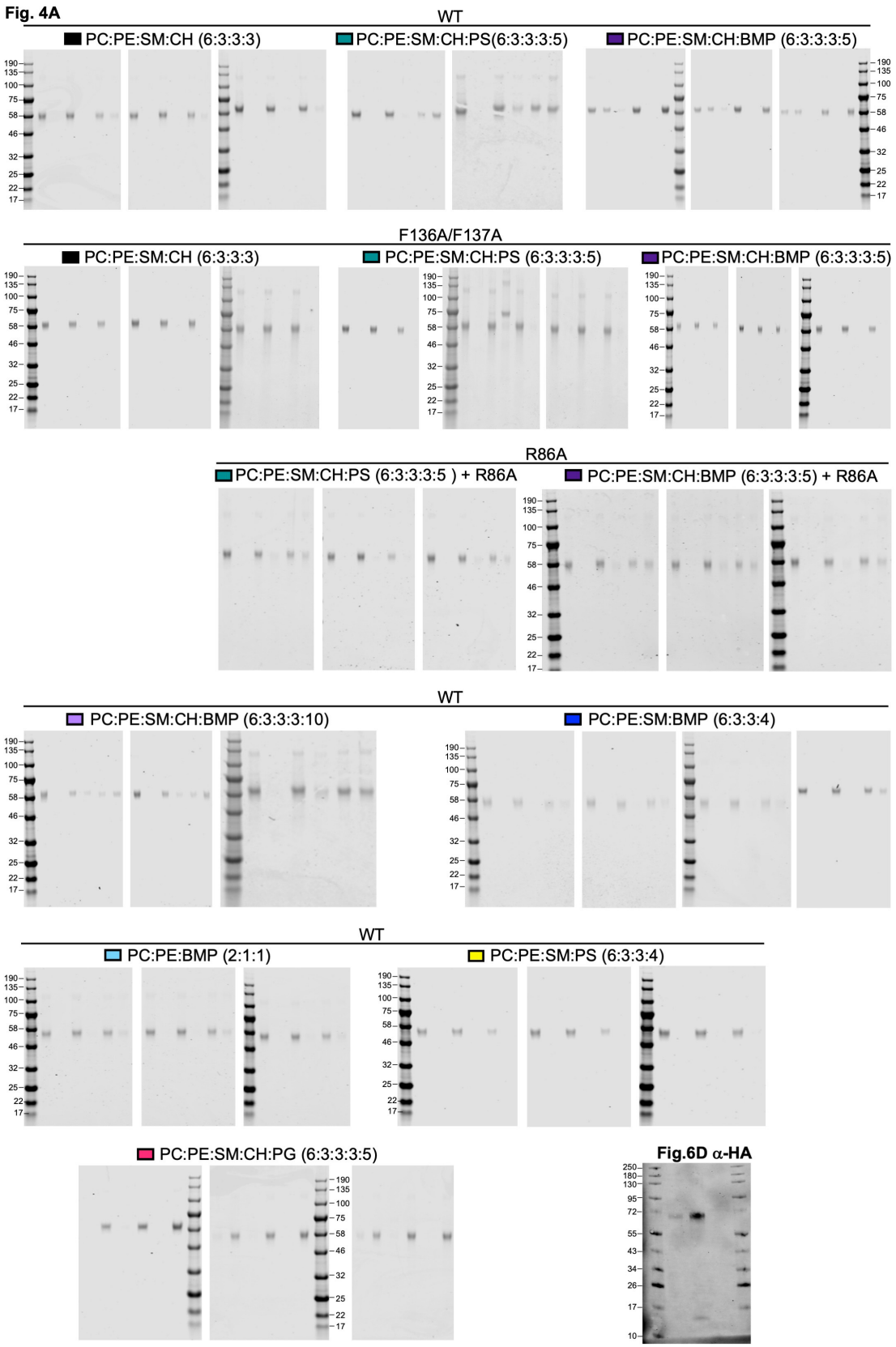


Fig. S8. Uncropped gels. Gels shown for all liposome flotation replicates (**Fig. 4A**), along with immunoblot from **Fig. 6D**.

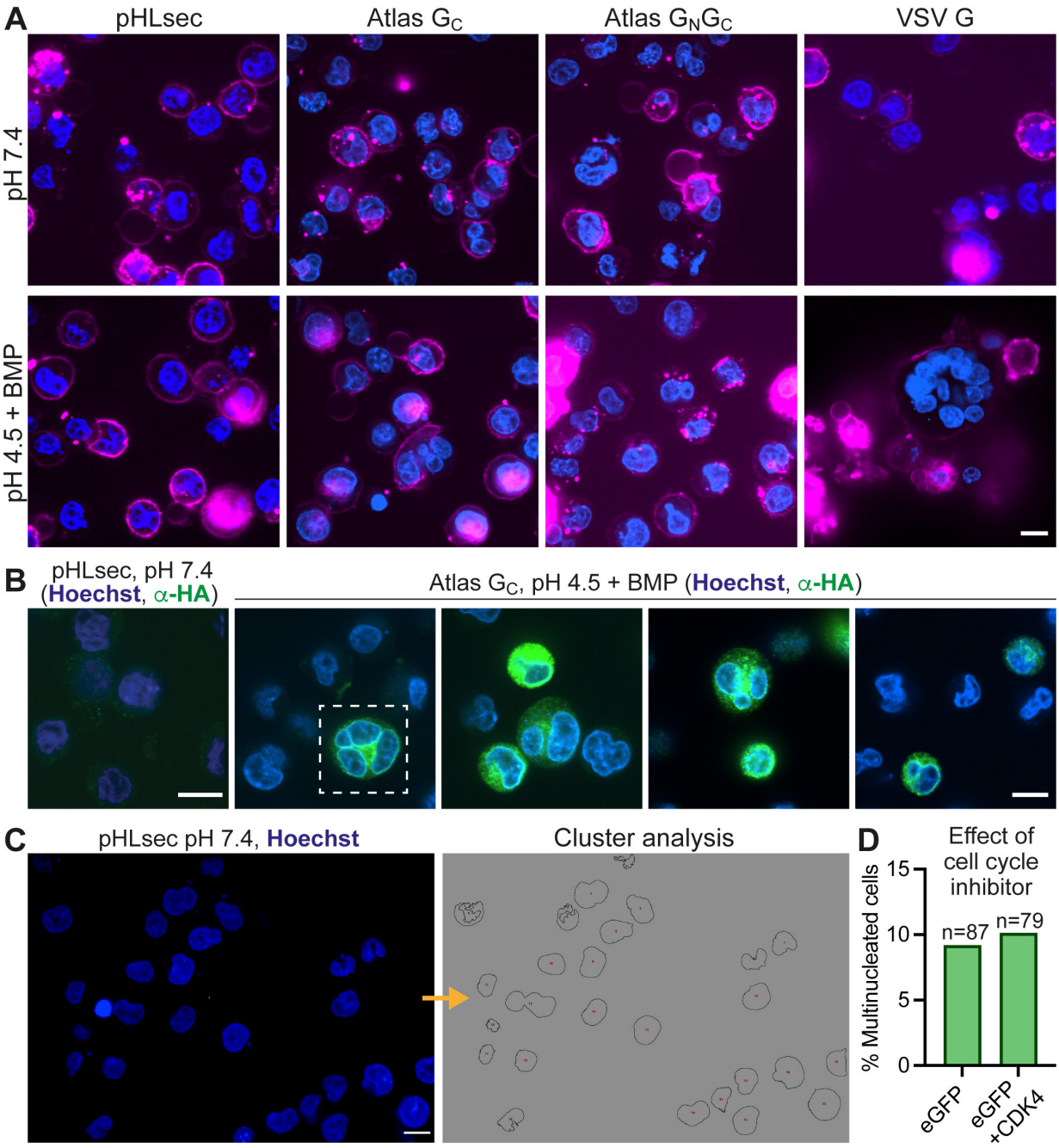


Fig. S9. Cell-cell fusion assay with CHO cells. (**A**) Representative confocal micrographs of CHO cells transfected with plasmid encoding Atlas G_C, Atlas G_N-G_C, VSV G, or no protein (pHLsec empty vector). Cells were treated with pH 4.5 buffer containing BMP, or pH 7.4 buffer without exogenous lipid. Blue, Hoechst 33342 nuclear stain; Magenta, CellBrite Red plasma membrane dye. (**B**) Permeabilized and fixed cells transfected with plasmid encoding G_C stained with anti-HA antibody, to detect the C-terminal HA tag on G_C. Blue, Hoechst 33342; green, anti-HA antibody. The boxed region is also shown in **Fig. 6B** and **Movie S1**. (**C**) Left, confocal micrograph of cells transfected with pHLsec plasmid stained with Hoechst 33342. Right, masks from cluster analysis of the Hoechst channel with Fiji (*88*). All scale bars are 10 μm. (**D**) Effect of cell cycle inhibitor CDK4 on the fraction of di- or multinucleated cells. Cell were transfected with a plasmid encoding eGFP (pEGFP-N1) and incubated with 10 μM CDK4 (right) or with PBS (left) for 6 h at pH 7.4. *n*, number of nuclei counted. See **dataset S2** for source data.

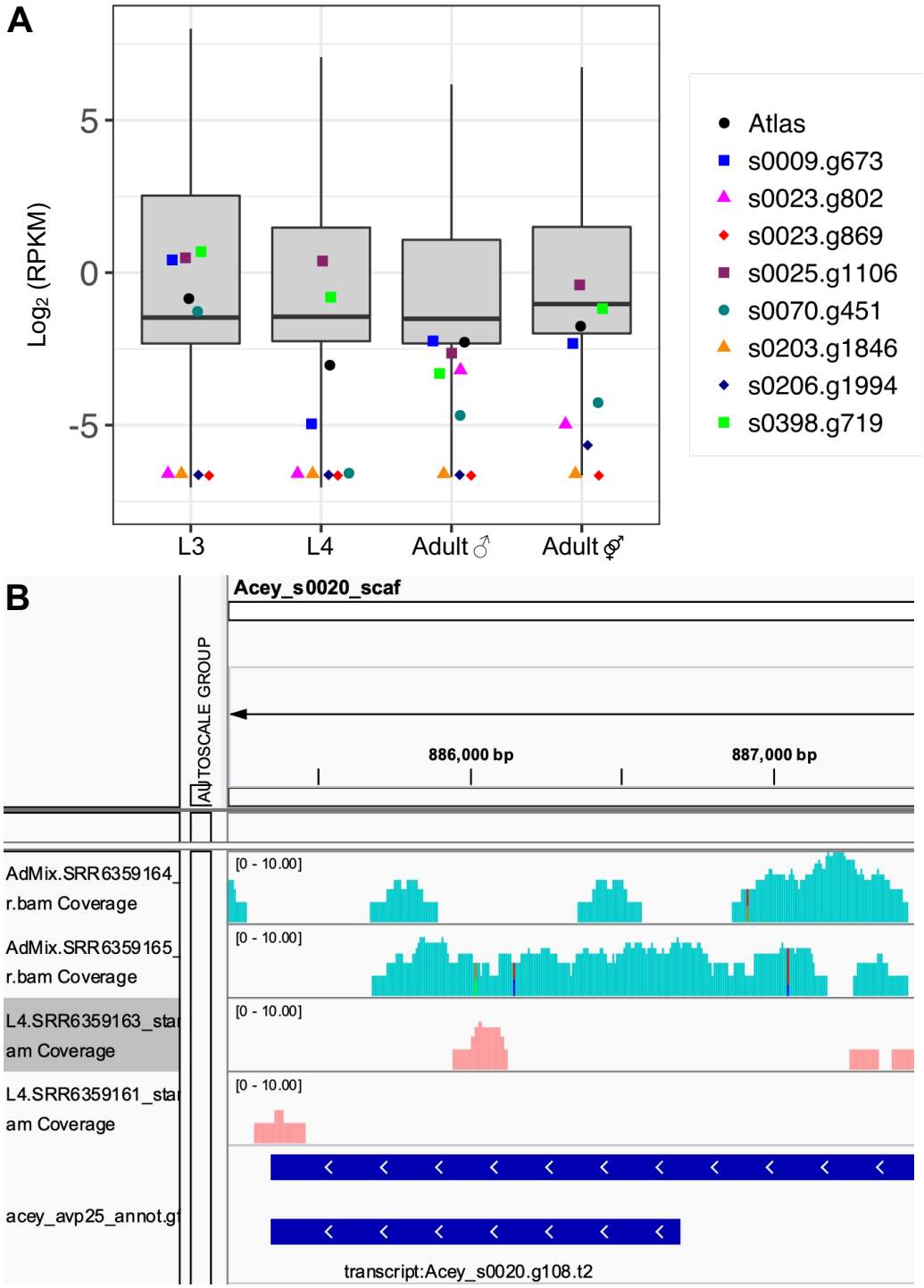


Fig. S10. RNA-Seq analysis of intact *A. ceylanicum* belpaovirus EVEs. (**A**) Box plots of RNA-Seq expression in Reads Per Kilobase of transcript per Million mapped reads (RPKM) for all annotated genes, and for nine intact *A. ceylanicum* belpaovirus EVEs including Atlas virus, at the L3, L4 and adult (male and mixed sex) developmental stages. RNA-Seq data are from Bernot *et al.* (*71*). (**B**) Integrative Genomics Viewer snapshot showing the distribution of raw RNA-Seq reads over the Atlas virus (*Acey_s0020.g108*) coding sequence in mixed sex adult worms and L4 larvae with individual unmerged BAM files.

1141 **Table S1. Cryo-EM data collection, structure determination, model building and**
1142 **refinement parameters and statistics.**

	Atlas G_C
Data Collection and Processing	
Microscope	FEI Titan Krios
Voltage (kV)	300
Electron exposure (electrons $Å^{-2}$)	46.12
Exposure per frame (electrons $Å^{-2}$)	1.28
Defocus range (μm)	-1.3 to -3.5
Pixel size (Å)	1.047
N. initial particles	987,570
N. final particles	197,145
Map resolution (Å)	3.76
Map resolution range (Å)	3.50 - 5.20
FSC threshold for resolution limit	0.143
Model refinement	
Map sharpening B factor ($Å^2$)	-180
Symmetry imposed	C3
Mask correlation coefficient	0.78
Model composition	
N. non-hydrogen atoms	10,159
Protein residues (chains A, B, C)	1,320
R.m.s. deviations	
Bond lengths (Å)	0.005
Bond angles (˚)	0.694
Planarity (Å)	0.004
B*-factors/ADPs	
Minimum	35
Maximum	176
Mean	82
Validation (Phenix 1.15)	
MolProbity overall score	2.23
MolProbity all-atom clashscore	5.86
Rotamer outliers (%)	2.28
Ramachandran plot	
% favored	86.2
% allowed	13.8
% outliers	0.0
PDB code	7A4A
EMDB code	EMD-11630
EMPIAR code	10266

1143 ADPs, atomic displacement parameters.

1144

1145 **Movie S1. Confocal Z-stack of the multinucleated cell shown in Fig. 6B (separate file).**

1146 CHO cell expressing Atlas G_C following treatment with pH 4.5 and BMP. Blue, Hoechst 33342;

1147 magenta, CellBrite Red; green, a-HA antibody for Atlas G_C detection. Scale bar, 10 μm.

1148

1149 **Dataset S1. (Separate file)**

1150 Source data file for Fig. 4 – liposome binding assays.

1151

1152 **Dataset S2. (Separate file)**

1153 Source data file for Fig. 6 and Fig. S9 – cell-cell fusion assays.

1154

1155 **Dataset S3. (Separate file)**

1156 Source data file for Fig. 7 – differential RNA-Seq expression calculations.

1157

1158 **Dataset S4. (Separate file)**

1159 Source data file for Fig. 7 – original RNA-Seq annotation (GFF) file.

1160

1161 **Dataset S5. (Separate file)**

1162 RNA-Seq analysis scripts and instructions (also available on Github

1163 [https://github.com/annaprotasio/Merchant_et_al_2020]).