Highly Conserved Core Residues Define Old-World Alphaviruses and Trace Early Evolutionary Divergence

Wendy Carolina Piña-Ruiz,[1,2,#], Luis Rubén Jaime-Rocha[1,2,#], Andrea Castorena-Robles[1,3], Roger Castells-Graells[4], Roya Zandi [5], Juan Carlos Muñoz-Escalante[6,7*], Mauricio Comas-Garcia[1,2,9*].

[1]High-Resolution Section, Research Center for Health Sciences and Biomedicine, Autonomous University of San Luis Potosi, Mexico

[2]Department of Sciences, Autonomous University of San Luis Potosi, Mexico

[3]Life Sciences Graduate Program, Department of Sciences, Autonomous University of San Luis Potosi, Mexico

[4]Structural Biology Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

[5]Department of Physics, University of California, Riverside

[6]Bioinformatic Unit, Research Center for Health Sciences and Biomedicine, Autonomous University of San Luis Potosi, Mexico

[7]Infectious Diseases Unit, Research Center for Health Sciences and Biomedicine, Autonomous University of San Luis Potosi, Mexico

[8]Translational and Molecular Medicine Section, Research Center for Health Sciences and Biomedicine, Autonomous University of San Luis Potosi, Mexico

[#]These authors contributed equally

[*]To whom send any correspondence: carlos.escalante@uaslp.mx and mauricio.comas@uaslp.mx

Abstract

Alphaviruses are positive-sense, single-stranded RNA viruses that assemble into striking double-icosahedral particles. During budding, their nucleocapsid core forms in the cytoplasm and adopts a T=4 icosahedral symmetry, a hallmark of Alphaviruses. Here, we combine structural and phylogenetic analyses to identify the amino acids most likely to govern capsomer formation (pentamers and hexamers) and core organization. We find that these residues are highly conserved in present-day Old-World alphaviruses but are divergent in New-World lineages. This suggests that the common ancestor of both groups likely assembled cores using the same molecular interactions seen in present-day Old-World

viruses. We propose that early divergence in these interactions weakened core assembly efficiency, potentially contributing to the attenuation observed in encephalitic New-World alphaviruses. This attenuation may reflect an adaptive trade-off, in which reduced assembly efficiency lowers viral replication and virulence, supporting long-term persistence in enzootic cycles. Revealing how specific residues control capsid architecture and tracing their evolutionary history, this study provides key insights into alphavirus assembly mechanisms, opening new avenues for antiviral strategies and rational vaccine design.

Keywords

Alphaviruses; core evolution; capsid protein-capsid protein interactions; viral attenuation

Introduction

Alphaviruses belong to the *Togaviridae* family, have a positive-sense single-stranded RNA genome of around 11-12 kilobases, and two open reading frames (ORFs) that code for two polyproteins (1). The ORF1 and ORF2 code the non-structural and structural proteins, respectively. Alphaviruses infect vertebrates and are transmitted by invertebrate vectors (2). These viruses have a double-layered virion consisting of two T = 4 icosahedral layers separated by a lipid membrane; the outer layer is composed of 80 trimeric spikes, each trimer composed of heterodimers of the glycoproteins E1 and E2, and the inner layer or core is made of 240 copies of the capsid protein (3). These spikes are not all placed at the same position in the icosahedron; 20 and 60 spikes at the 3- and quasi-3-fold symmetry axes, respectively. The core contains the full-length viral genome and assembles into an amorphous shell in the cytoplasm (4-7). At the plasma membrane, the core interacts with the heterodimer E1/E2, leading to the assembly and budding of the viral particle (4). Without the core, E1 and E2 bud as amorphous particles (8).

The Alphavirus virion has a unique architecture; the viral particle consists of two concentric T = 4 icosahedral layers separated by a lipid bilayer. Both layers interact through

56 hydrophobic amino acids in the cytoplasmic domain of E2 and the capsid protein (CP) (3).
57 These virions have structural features unique for an icosahedral viral particle(9-15): i) there
58 are large "holes" at the center of the pentamer and hexamers (capsomers) of the core, ii) the
59 capsomers do not establish large contact surface areas between them, and iii) the torsion
60 angles and distances between capsomers can significantly differ between viral species.

61 Alphaviruses can be grouped by antigenic complexes (*e.g.,* Semliki Forest and
62 Western equine encephalitis complexes) (16), by the location where they are endemic or first
63 isolated (*i.e.,* Old- and New-World Alphaviruses) (17), or by some characteristic of the host
64 (*e.g.,* if they infect aquatic animals or have no known vertebrate hosts) (18). Most
65 Alphaviruses are classified as either Old- or New-World; the latter is restricted only to the
66 American continent, and the former to the rest of the world. On the one hand, Old-World
67 Alphaviruses mostly cause a febrile and arthrogenic dengue-like illness with a low mortality
68 rate (19). On the other hand, New-World Alphaviruses affect the central nervous system,
69 resulting in an encephalitic disease that, in some cases, has high mortality rates (up to 90%
70 depending on the virus and the age of the patient) (20). Forrester and coworkers proposed
71 that Alphaviruses originated from a marine host that jumped to terrestrial hosts, with birds
72 being key to their global spread (18). Interestingly, some Alphaviruses are endemic to the
73 New World (*i.e.,* Mayaro and Una viruses) but are phylogenetically part of the Old-World
74 Alphaviruses. The opposite is true for the Sindbis virus, which is endemic to the Old World
75 but is genetically related to New-World Alphaviruses (17, 18).

76 We first focused on identifying some structural differences in the cores of
77 Alphaviruses using known structures to determine putative interactions that regulate CP-CP
78 and capsomer-capsomer interactions. Then, we predicted the structure of the CP for all
79 Alphaviruses and performed bioinformatic analysis of the consensus and representative
80 amino acid sequences of the CPs. We found at least eight highly conserved amino acids in
81 Old-World Alphaviruses, which are likely to mediate core assembly through electrostatic
82 interactions (salt bridges). These amino acids are not conserved in other Alphaviruses.
83 Finally, through a series of evolutionary analyses, we found that the core of the common
84 ancestor most likely had almost the same electrostatic interactions as in Old-World

Alphaviruses. These results suggest that a strong selective pressure resulted in the divergence of Old- and New-World Alphaviruses and altered the interactions of the core's CP.

Materials and Methods

Viral sequences for the prediction of the capsid protein structure.

The CP sequences for 42 *Alphavirus* species were obtained from the NCBI Taxonomy browser (Viruses; *Riboviria, Orthonovirae; Kitrinoviricota; Alsuviricetes; Martellivirales; Togaviridae*). For the E2 protein sequences, only 29 species had sufficient data to obtain representative sequences. The consensus sequences were generated using BioEdit v7.2.5.

Structural data and structure modeling.

The cryo-electron microscopy (cryo-EM) structures Sindbis (SINV) (9), Getah (GETV) (10), Chikungunya (CHIKV) (11), Mayaro (MAYV) (12), Western equine encephalitis (WEEV) (13), Venezuelan equine encephalitis (VEEV) (14), Eastern equine encephalitis (EEEV) (15), Barmah forest (BFV) (22), and Semliki forest (SFV) (23) viruses were obtained from the RCSB PDB website (with PDB IDs 3J0F, 7FD2, 3J2W, 7KO8, 8DEC, 7SFV, 6XO4, 2YEW, and 1DYl, respectively). The experimental resolution of these structures is 7.0, 2.8, 5.0, 4.4, 4.7, 4.0, 4.2, 5.0, and 9.0 Å, respectively. Models for the structures of the capsid protein and E2 were generated using AlphaFold2 (24); some sequences were randomly selected to be predicted using AlphaFold3 (25), and we obtained the same results in all cases as with AlphaFold2. The structures were analyzed and visualized using ChimeraX (26). All CP and E2 structures were compared to CHIKV (PDB ID: 3J2W) (11); in cases where multiple structures were obtained (*i.e.,* when using AlphaFold3), we reported the structure with the lowest RMSD when overlapped to CHIKV PDB ID 3J2W. All distances between side chains were calculated by measuring the centroid-to-centroid distances by generating spheres with 2 Å radius.

Alphavirus sequences datasets curation for phylogenetic analysis.

The complete nucleotide sequences corresponding to the 32 *Alphavirus* species were downloaded from GenBank using their taxonomy IDs. We used a different number of viruses for the structural prediction because the number of available sequences for some was not ideal for phylogenetic analyses. Sequences for each species were individually aligned using MAFFT v7.450 (27). The respective annotated RefSeq sequence for each species was used to identify the core coding sequence (CDS) region and to trim the multiple sequence alignments accordingly using BioEdit v7.2.5. To improve the quality of subsequent analyses, sequences were removed if they lacked more than 5% of the total length of the core CDS, contained more than 5% ambiguous nucleotides, or exhibited atypical indels not present in other sequences from the same species.

Consensus and representative sequences for each *Alphavirus* species were generated using CD-HIT-EST at a 90% similarity threshold. These representative core CDS sequences for each species were included in a single dataset and aligned using MEGA v10.2.6 with the MUSCLE algorithm (21). Amino acid sequences were inferred from the aligned nucleotide sequences, and manual curation of the Multiple Sequence Alignment (MSA) was performed using BioEdit v7.2.5.

Phylogenetic analysis and reconstruction of ancestral sequences

Phylogenetic inference was conducted in MEGA v10.2.6 using the best-fitting model, and tree robustness was assessed with 1,000 bootstrap iterations. The most probable ancestral sequences for each internal node of the resulting phylogenetic tree were reconstructed using the ExtAncSeqMEGA extension.

137 **Selective pressure, Entropy, and Recombination analysis**

138 To assess potential recombination events within *Alphavirus* CP coding sequences,
139 alignments were analyzed using RDP v5.66, using RDP, GENECONV, BootScan, MaxChi,
140 Chimaera, SiScan, and 3Seq as assessment methods. Potential events supported by at least
141 three different methods with statistically significant *p*-values (<0.05) were considered true
142 recombination events. Selective pressure acting on codon sites was evaluated on either all
143 *Alphavirus* or only Old-World Alphaviruses nucleotide MSA using the Datamonkey Adaptive
144 Evolution server employing SLAC, FEL, FUBAR and MEME methods; sites were classified as
145 being under pervasive or episodic positive selection if detected by MEME ($p<0.1$), FUBAR
146 (posterior probability >0.9), FEL ($p<0.1$), or SLAC ($p<0.1$), respectively.

147 Site-wise variability (either all *Alphavirus* or only Old-World Alphaviruses) was further
148 examined using Shannon entropy calculation performed in BioEdit v7.2.5 on the MSA.
149 Entropy scores were used to identify highly variable positions across the core protein
150 sequences, supporting the detection of sites associated with adaptive or host-specific
151 changes.

152

153 **Results**

154 **The interactions between capsid proteins in Old-World viruses are highly conserved.**

155 First, we analyzed the cryo-electron microscopy (cryo-EM) structures of the core of Sindbis
156 (SINV) (9), Getah (GETV) (10), Chikungunya (CHIKV) (11), Mayaro (MAYV) (12), Western
157 equine encephalitis (WEEV) (13), Venezuelan equine encephalitis (VEEV) (14), Eastern
158 equine encephalitis (EEEV) (15), Barmah forest (BFV) (22), and Semliki forest (SFV) (23)
159 viruses (see Figure 1). In Figure 1, the different colors within a structure represent the four
160 proteins that constitute the core asymmetric unit. These cores have a T = 4 symmetry; they
161 are composed of 240 capsid proteins arranged in 12 pentamers and 30 hexamers. Despite
162 having conserved overall architecture, close inspection of the hexamers, pentamers, and
163 the three- and quasi-threefold symmetry axes shows interesting species-specific features

164 (*e.g,* the number of amino acids establishing inter- and intra-capsomere interactions, the
165 distance between capsomers and interacting amino acids, and the torsion angle between
166 them).

167 Figure 2A shows no significant differences in the pentamers of these three Old-World
168 Alphaviruses (*i.e.,* CHIKV, MAYV, and GETV) compared to those in the three New-World
169 Alphaviruses in Figure 2B; in EEEV, the capsid proteins (CP) in the pentamer establish fewer
170 inter-subunit contacts than in VEEV and WEEV (see the asterisks). The structures of the VEEV
171 and WEEV pentamers exhibit no noticeable differences compared to those of CHIKV, MAYV,
172 and GETV. However, the quasi-threefold symmetry axis (q3) shows some changes
173 (supplementary figure 1); the space between two adjacent hexamers and one pentamer is
174 approximately 23 Å, 19 Å, and 25 Å wide for CHIKV, MAYV, and GETV, respectively. These
175 distances were calculated by generating 2 Å centroids for the amino acids K122, K119, and
176 K129 in CHIKV, MAYV, and GETV, respectively. Then we measured the centroid-to-centroid
177 distance between these amino acids in the three proteins forming this symmetry axis. Amino
178 acids Q121, K125, and Q119 were used for EEEV, VEEV, and WEEV, respectively (in the
179 following sections we will explain why we used those amino acids as references).
180 Interestingly, while the general features of the VEEV pentamer are like those of the Old-World
181 Alphaviruses, the space at the q3 is much smaller (≈ 13 Å) than for the other five viruses in
182 this figure (≥ 15 Å) (see supplementary figure 1). The three-fold symmetry axis (*i.e.,* between
183 three hexamers) shows a pattern (data not shown).

184 These results suggest that the physicochemical properties of the amino acids, which
185 should control capsomer assembly and capsomer-capsomer interactions, may be similar
186 for these three Old-World viruses and vary in the New-World Alphaviruses. To explore this
187 idea, we focused on the CP contact surfaces, using the CHIKV core cryo-EM structure (PDB
188 ID 3J2W) as a reference. The assembly of Alphaviruses occurs at neutral pH at the plasma
189 membrane, while disassembly occurs in the endosome under acidic conditions (3).
190 Therefore, we focused on pH-dependent interactions in multiple CP-CP interfaces using
191 CHIKV as our reference system.

192 Figure 3A suggests electrostatic interactions (salt bridges) between CHIKV CP K172
193 and E234 (outer), and K177 and E184 (inner). It should be noted that the resolution of these
194 structures does not permit precise localization of side-chain positions. Nonetheless, the
195 distances of these amino acids and their chemical environment suggest electrostatic
196 interactions between the CPs in the pentamers. These putative interactions are also present
197 in the hexamers (data not shown); however, because of the asymmetry of the hexamer, the
198 distances between K177 and E184 vary between adjacent CPs and are greater than in the
199 pentamers (data not shown). Figure 3B shows the three-fold symmetry (i3) axis of CHIKV,
200 suggesting that the capsomer-capsomer interactions depend on two sets of electrostatic
201 interactions: E120/K151 and K122/D148. The position of these amino acids resembles a
202 dipole-dipole interaction; thus, the strength of these interactions should depend on the
203 distances and angles between capsomers. To further understand the structures shown in
204 Figure 2, we overlapped the structure of the monomeric CHIKV CP (Figure 3C) with those of
205 Alphaviruses, MAYV and EEEV, with a known cryo-EM structure. Figures 3D and 3E show the
206 overlap of CHIKV CP with MAYV and EEEV CPs, respectively. On the one hand, Figure 3D
207 shows that MAYV CP maintains perfect charge conservation and identity at the positions of
208 the eight amino acids of interest. The only difference is that CHIKV CP K151 has an arginine
209 in MAYV (R148). On the other hand, Figure 3E shows key differences in the EEEV CP; the
210 amino acids that should be in the same position as in CHIKV CP K122, K151, D148, K177,
211 and E184 are instead Q, Q, A, L, and P, respectively. The overlap of the structures of GETV,
212 SFV, and BFV with CHIKV shows that the amino acids in these eight positions are 100%
213 conserved in charge, while the degree of conservation for SINV, WEEV, and VEEV is much
214 lower (data not shown).

215 Figure 4 shows a closer look at the interfaces in a pentamer; Figure 4A (top row)
216 contains only Old-World Alphaviruses, which exhibit the same salt bridges as CHIKV. Figure
217 4B shows the equivalent interface for three New-World Alphaviruses; the amino acids in
218 positions equivalent to CHIKV CP K177 and E184 (inner interactions) are very different
219 between EEEV, VEEV, and WEEV. Nonetheless, the charge of the amino acids equivalent to
220 CHIKV CP K172 and E234 (outer interactions) is the same in all cases. The “outer

221 interactions" in the pentamer seem to be conserved in Old- and New-World Alphaviruses,
222 while the "inner interactions" are only conserved in Old-World Alphaviruses. Interestingly,
223 while there are large differences in the CP conformation and the CP-CP surfaces in the
224 pentamers of Old-World Alphaviruses, the "inner" and "outer" electrostatic interactions
225 seem to be the same. In the New-World Alphaviruses, the CP conformation, the CP-CP
226 surfaces, and the putative interacting amino acids that are equivalent in position to those in
227 Old-World Alphaviruses are not conserved.

228 To better understand the interactions between capsomers (Figure 3B), we analyzed
229 the i3 axis (see Figure 5). The leftmost panel in Figure 5A shows the centroid-to-centroid
230 distance (2-Å centroid) and torsion angle between three CHIKV CP K122 amino acids. The
231 letters in parentheses indicate the chains being compared in the Cryo-EM structure. The
232 distances between equivalent amino acids are asymmetric, although the exact positions of
233 the side chains are unknown. This analysis was also conducted for MAYV, GETV, EEEV, VEEV,
234 and WEEV. The interactions in CHIKV, MAYV, and GETV (Figure 5A) are the same in these
235 viruses, and the distances and torsion angles between equivalent amino acids do not differ
236 between these viruses. Figure 5B shows that at the i3, the interactions in EEEV and WEEV are
237 between polar uncharged amino acids (Q and N) instead of charged amino acids.

238 Additionally, the middle panel in Figure 5B shows that in VEEV, a cluster of charged
239 amino acids (K135 and E135 from three capsid proteins) does not appear to have a
240 corresponding pair of amino acids that could establish a dipole-dipole-like interaction, as
241 observed in CHIKV, MAYV, and GETV (see Figure 5A). The distances between Q121 and Q119
242 in EEEV and WEEV are similar to the basic amino acids in the i3 of CHIKV, MAYV, and GETV.
243 The only significant difference in WEEV is that the torsion angle of the amino acid is greater
244 than that of Old-World Alphaviruses. The interactions in VEEV are very different from those
245 of the other viruses in this Figure; for example, the charged amino acids are about half the
246 distance as in CHIKV (≈ 12.8-15.8 Å vs ≈ 27.2-30.6 Å, respectively). There is no pair of amino
247 acids with an opposite charge that could help to establish a dipole-dipole-like interaction.
248 The torsion angles between these amino acids are indicated in the center of the i3 site; these
249 angles are very similar for these three Old-World Alphaviruses and EEEV (between 51-57 °).

250 However, EEEV and WEEV have different torsion angles, albeit having similar distances and
251 amino acids at the sites of interest. This suggests that other amino acids contribute to the
252 CP-CP in these interfaces.

253

254 **The electrostatic interactions between capsid proteins are a unique feature of Old-**
255 **World viruses.**

256 The results in the previous section were limited to Alphaviruses with known Cryo-EM
257 structures; however, we wanted to identify these interactions for all Alphaviruses with known
258 CP sequences. First, we generated a consensus sequence for the CP of all Alphaviruses
259 (including those in the previous section) and predicted the structure of the monomeric
260 capsid protein using Alphafold2 (24). We also used Alphafold3 (25) for a few randomly
261 selected structures, and the results were consistent with those obtained with Alphafold2.
262 The goal of this prediction was to identify amino acids for Alphaviruses without structural
263 data that are in positions equivalent to those of interest in the monomeric form of the CHIKV
264 CP and not to infer any data from a possible multimeric context. In fact, we only analyzed the
265 predicted monomeric structures of the capsid proteins because Alphafold2 and 3 failed to
266 assemble them into pentamers or hexamers. For this analysis, we used consensus
267 sequences. Then, we overlapped the predicted structures with the CHIKV CP (PDB 3J2W) as
268 was done in Figures 3D and 3E to identify the amino acids that are in equivalent positions to
269 those of CHIKV CP, specifically E120, K122, D148, K151, K172, K177, E184, and E234 (see
270 supplementary files). From here on, we will use this numbering to indicate the amino acids
271 of reference, and CHIKV will be the virus to which we will compare all interactions. The
272 results of these analyses are presented in Figures 6-9. The numbers in green indicate that
273 the percentage of amino acid identity and/or charge conservation is greater than or equal to
274 80%.

275 Additionally, we performed a multiple-sequence alignment (MUSCLE) of the
276 consensus sequences (see Supplementary Figures 2 and 3). The comparison between the
277 two methods revealed a few differences. The lavender, turquoise, and maroon amino acids

278 in Figures 7-9 indicate that the positions of the amino acids predicted through Alphafold2
279 and MUSCLE differ. The lavender color indicates that the amino acid overlapping with the
280 reference structures is due to an insertion. The amino acids in turquoise overlap at a -1
281 position in the aligned sequence, and the maroon color refers to amino acids that overlap
282 with the reference CHIKV amino acids because of a deletion.

283 Figure 6 shows almost 100% charge conservation of these amino acids in all Old-
284 World Alphaviruses; the Ndumu and Middelburg viruses have a charge inversion in K122 and
285 a Q in K177, respectively. Also, except for the positions equivalent to E120 and K172, the
286 identities of the other six amino acids are conserved between 80% and 100%. These results
287 suggest that for Old-World Alphaviruses, there is strong evolutionary pressure on the charge
288 of amino acids that might control the intra- and inter-capsomer interactions and the length
289 of the side chains of these amino acids. These findings are consistent with the results
290 presented in the previous section.

291 In New-World Alphaviruses, except for the amino acids equivalent to K172/E234 (*i.e.,*
292 the "outer interactions"), the charge is not conserved in the rest of the amino acids of interest
293 (Figure 7). However, there seems to be some degree of conservation within antigenic
294 complexes. In most cases, the charged amino acids are replaced by polar uncharged amino
295 acids, except in the positions equivalent to D148 and K151, where there is a significant
296 amount of hydrophobic side chains. Interestingly, E120 and K122 are conserved in the VEEV
297 antigenic complex. In contrast, in the WEEV complex, E120 can be a polar uncharged, or
298 acidic side chain, and K122 is either an acidic or a polar uncharged amino acid. In WEEV and
299 VEEV, D148 is mostly substituted for Q or V. In the WEEV complex, K177 and E184 are Q and
300 P, respectively, while in the VEEV complex, the charge of K177 is conserved, and E184 is
301 almost always a Q. The identities of five of the eight amino acids of interest within the VEEV
302 complex are more conserved than within viruses of the WEEV complex. These results
303 suggest a common ancestor for the CP of Aura virus, SINV, and Whataroa virus and a distinct
304 ancestor for the other viruses within this antigenic complex. Furthermore, based on the
305 results of Table 2, it is likely that the CP of the Madariaga virus may have had a common
306 ancestor with the cores of WEEV, Fort Morgan Virus, and Highland J virus.

307 The lack of conservation of intra- and inter-capsomeric interactions is more evident
308 for Alphaviruses found in mosquitoes with no known vertebrate host that are not included in
309 the Old-/New-World classification and on aquatic Alphaviruses (see Figures 8 and 9). Figure
310 8 shows that, on the one hand, like in New-World Alphaviruses, a hydrophobic side chain
311 replaces D148. On the other hand, K151 is replaced by a polar uncharged amino acid, which
312 is unique for this group. Interestingly, the charge conservation of the viruses in Table 3 is
313 closer to that of Old-World than New-World Alphaviruses.

314 The degree of charge conservation of the amino acids of interest in aquatic
315 Alphaviruses is low compared to CHIKV. However, in aquatic Alphaviruses that infect
316 mammals, six of the eight amino acids of interest are conserved in these viruses (see shaded
317 rows in Figure 9). While the number of viruses in this category is too small to make a
318 conjecture, it could suggest that, at least based on the core and/or CPs, these two viruses
319 might be related to Old-World Alphaviruses. In aquatic Alphaviruses that infect fish, there
320 are two positions with charge inversion, K151 and K177. This charge inversion should result
321 in intra- and inter-capsomeric interactions that differ from the rest of the Alphaviruses.

322 The data in Figures 6-9 show a wide range of interactions for non-Old-World
323 Alphaviruses. Nonetheless, it is clear that the "outer intracapsomeric interactions"
324 (E234/K172) are highly conserved in almost all Alphaviruses; only in the Trocara virus and the
325 Wenling fish alphavirus, the amino acids in the equivalent positions to CHIKV CP E234 and
326 K172 are a Q and an H, respectively. These results suggest that there is a strong evolutionary
327 pressure on the salt bridge between CHIKV CP E234 and K172, and that this interaction has
328 been conserved during evolution, possibly playing a crucial role in core assembly.

329

330 **The electrostatic interactions between capsid proteins at the core might be related to**
331 **glycosylation patterns in E2.**

332 The results presented so far suggest that the organization and type of interactions in the core
333 of Alphaviruses could have changed when Old- and New-World Alphaviruses diverged. But
334 these are not the only changes. The glycosylation profiles of E2 and E1 for Old-World

335 Alphaviruses differ from New-World ones (28). For example, CHIKV E2 has only one
336 glycosylation site in N623. A bioinformatic and structural analysis identical to the one
337 described in the previous section, aiming at finding an N-X-S/T motif at the same position as
338 CHIKV E2 N263, shows that this glycosylation site is conserved in almost all Old-World
339 Alphaviruses (see Figure 10A). Only the Bebaru virus appears to have an N-to-G substitution
340 at this site. We could not find an N-X-T/S motif in or around the E2 region of the Bebaru virus
341 that is structurally equivalent to the CHIKV motif. Figure 10B shows the Cryo-EM density of
342 the E2 glycosylation site for MAYV, which has been proposed to act as a "molecular
343 handshake" between trimers that stabilize the viral particle (12). It should be noted that
344 analyses E2 and CP do not have the same number of species, as there are fewer available
345 sequences in E2 than in CP. Therefore, the number of viruses in Figures 6-9 and Figures 10A,
346 10B, and 10C differs, and thus, this could limit our study.

347 Figure 10C shows no N-glycosylation motif on or around CHIKV E2 N263 in New-
348 World Alphaviruses. This figure also suggests that the Fort Morgan virus may have an N-
349 glycosylation site (N260) in a position close to CHIKV E2 (N263). However, E2 N260 in the
350 Fort Morgan virus is not equivalent to CHIKV E2 N263. In the Old-World Alphaviruses, the
351 glycosylation site equivalent to CHIKV N263 is in a β-sheet, while the Fort Morgan N260 (N-
352 V-T motif) is in a loop before this β-sheet (Figure 10D). The position of this possible
353 glycosylation site makes it unlikely to result in the same trimer-trimer interactions as the
354 "molecular handshake" proposed for MAYV (12).

355 Finally, the data in Figure 10E suggest that Harbor porpoise alphavirus, which infects
356 marine mammals, has the same glycosylation site as CHIKV. Interestingly, this marine
357 Alphavirus has seven of the eight amino acids studied in the previous sections conserved in
358 charge compared to CHIKV. Therefore, there may be a relationship between the identity of
359 the amino acids that control assembly and the position of the glycosylation sites in E2. It is
360 possible that in non-Old-World Alphavirus cores, the interactions between CPs are weak,
361 and glycosylation in E2 N263 could introduce rigidity that inhibits the assembly of the
362 icosahedral core during budding.

363

364 **The common ancestor of Old- and New-World Alphaviruses may have had intra- and**
365 **inter-capsomer interactions like those of CHIKV.**

366 Phylogenetic inference analysis of Alphaviruses is usually done using the structural viral
367 protein E1, E2, and TF, or a non-structural protein (17, 18, 29). However, using E1 and E2
368 provides information about the evolution of viruses in the context of their interactions with
369 hosts. Therefore, using these proteins is extremely useful for understanding the evolution of
370 a pathogen from a perspective that incorporates contributions from both the host and
371 vectors. Nonetheless, we wanted to understand the evolution of the core and determine
372 what kind of core was first: was the core with strong electrostatic interactions (Old-World
373 Alphaviruses) first, and then the core with weak interactions appeared (New-World
374 Alphaviruses), or vice versa? First, we generated representative viral sequences with
375 sufficient CP data and performed phylogenetic inference using MEGA v10.2.6 and
376 ExtAncSeqMEGA (see Methods and Materials).

377 Figure 11 shows the result of the phylogenetic analysis using CP. The structural
378 analysis was conducted using consensus sequences; thus, we studied 42 viral species.
379 However, to obtain a robust phylogenetic analysis, we used representative sequences,
380 which limited us to 32 species. As expected from the results shown in Figures 6 and 7, there
381 is a clear separation between Old- and New-World Alphaviruses; however, this analysis
382 shows that the CP of Sindbis, Aura, and Whataroa viruses diverged separately from the rest
383 of the New-World clades. This separation is similar but not identical to those reported using
384 the structural proteins E1, E2, and TF (17, 18, 29). Figures 9 and 10D suggest that Harbor
385 porpoise alphavirus did not jump directly from fish to a porpoise, but it appeared when the
386 Old-World group separated from the Aura virus clade (blue clade in Figure 11). As shown by
387 Weaver and Forrester (17, 18), Middelburg, Bebaru, Semliki, Getah, and Ross River viruses
388 belong to a distinct clade, separate from CHIKV, O'nyong nyong, Una, and Barmah Forest
389 viruses. However, our results differ in some respects. By looking only at the CP, the Una and
390 Mayaro viruses are not part of the same clade as proposed based on E1, E2, and TF (17).

391 Analysis of the CP results reveals that within the WEE antigenic complex, Aura, SINV,
392 and Whataroa viruses belong to a different clade than the Highlands J and WEE viruses.
393 Interestingly, this analysis suggests that the evolution of the CP of Fort Morgan virus differs
394 significantly from that observed for other structural proteins. New-World Alphaviruses
395 appear to have given rise to three main clades of the viral core. One clade includes Aura,
396 SINV, and Whataroa viruses, which appear to be closer ancestors of Old-World
397 Alphaviruses than the rest of the New-World Alphaviruses. This is consistent with SINV and
398 Whataroa being found in the Old World. The second clade includes Fort Morgan, Madariaga,
399 Highlands J, EEEV, and WEE viruses, while the third clade encompasses the rest.

400 Additionally, we performed an entropic and selective pressure analysis of Old- and
401 New-World Alphavirus CPs to understand how variable these amino acids are and the types
402 of selective pressure they may be subject to during evolution. When these sequences were
403 analyzed, we did not detect strong selective pressure to maintain the identities of CP E120,
404 K122, D148, K151, K172, K177, E184, and E234 (supplementary files). However, when this
405 analysis was conducted with only Old-World Alphavirus CP, 31 sites within the CP MSA were
406 under purifying selection, as detected by FEL or FUBAR analyses (supplementary files).
407 Interestingly, 16 amino acids also exhibited low entropy values, indicating high evolutionary
408 conservation. A Shannon entropy value less than 0.9 indicates high conservation; the
409 maximum possible Shannon entropy if there are 20 possible amino acids with the same
410 probability to appear is 4.32 (*i.e.,* random distribution), and a value of 0 refers to complete
411 conservation. Therefore, a value of 0.9 indicates that there are only two or less possible
412 amino acids for that position. These conserved residues are likely to be functionally or
413 structurally important, potentially contributing to the proper functioning of these proteins.
414 The overlap between strong negative selective pressure and low sequence variability
415 supports the idea that these sites are subject to evolutionary constraints due to their
416 essential roles in the viral cycle.

417 Notably, a general trend of low entropy was observed starting at position 130 of the
418 295-amino-acid alignment (of the Old-World Alphavirus consensus sequence), suggesting
419 a region of sustained conservation across multiple species, potentially associated with a

420 structurally constrained domain (the first 100 to 130 amino acids are disordered and interact
421 with the genomic RNA). Furthermore, most of the sites subjected to purifying selection were
422 located between positions 96-145 and 187-206 (supplementary files), highlighting two
423 discrete regions of potential functional importance that may be critical for the structural
424 integrity of the viral capsid.

425 A closer look at CHIKV CP in positions E120, K122, D148, K151, K172, K177, E184,
426 and E234 showed that the Shannon entropy of these amino acids is 0.7, 0.3, 0.3, 0.3, 0.3, 0,
427 and 0.3, respectively (see Figure 12). These results are consistent with the hypothesis that
428 these eight amino acids are key in Old-World Alphavirus CP. The consequences of these
429 observations will be discussed later, but they may be related to the pathogenesis and type
430 of disease they cause.

431 Then we used the ExtAncSeqMEGA to infer a few most probable ancestor sequences
432 (see the three asterisks and letters A-G in the insert in Figure 11). This analysis strongly
433 suggests that the common ancestor of Alphaviruses that infect mammals but not fish had
434 seven of the eight amino acids we propose mediate CP-CP and capsomer-capsomer
435 interactions in Old-World viruses. This data implies that, after the appearance of the Old-
436 World Alphavirus ancestor in node C, selective pressure on the eight amino acids of interest
437 became so strong that they became invariant. The only difference between this common
438 ancestor's amino acids of interest and Old-World Alphaviruses is that the equivalent amino
439 acid to CHIKV D148 is V158.

440 The event that led to the divergence of these two main groups resulted in a significant
441 diversification of the New-World Alphavirus CP. However, on the one hand, the common
442 ancestor of the Sindbis clade has the same eight amino acids as the common ancestor for
443 all Alphaviruses that do not infect fish (*** in Figure 11). On the other hand, the common
444 ancestor for the VEE and EEE complexes already had three mutations in the CP: one
445 mutation in the inter- and two mutations in the intracapsomeric interactions, respectively.
446 The most divergent node is the one that gave rise to the EEE complex.

447

448 **Discussion**

449 The results presented here suggest that the CP-CP interactions that mediate core assembly
450 of New- and Old-World Alphaviruses are distinct. On the one hand, based on the analysis of
451 cryo-EM structures and the predicted structure of the CPs, it is likely that the core of CHIKV
452 and all other Old-World Alphaviruses is controlled through highly conserved electrostatic
453 interactions. Most likely, salt bridges control the assembly of capsomers (*i.e.,* pentamers
454 and hexamers), and dipole-dipole-like interactions at the i3 and q3 axes are responsible for
455 capsomer-capsomer interactions. On the other hand, the amino acids that control the
456 assembly of capsomers and capsomer-capsomer interactions in New-World and other
457 Alphaviruses are not conserved, and their interactions should be weaker than in the Old-
458 World cores. Based on the proposed amino acids, it is likely that the assembly of the core of
459 Old-World Alphaviruses is mediated by highly conserved, strong electrostatic interactions
460 unique to this group of viruses.

461 Kaelber *et al*. showed that CHIKV, EEEV, and VEEV have T = 4 virions, with a minority
462 population of T = 3 and abnormal particles (30). However, the prevalence of T = 3 and
463 abnormal virions is higher in VEEV and EEEV than in CHIKV. While they did not propose why
464 EEEV and VEEV virions have a higher population of misassembled virions than CHIKV, our
465 results are consistent with a scenario where the assembly of EEEV and VEEV virions results
466 in a smaller population of (correct) T = 4 virions than CHIKV, because the intra- and
467 intercapsomere interactions of EEEV and VEEV CP are weaker than those of CHIKV and
468 hence are more prone to making mistakes during assembly. Unpublished results from our
469 group show that point mutations in the CHIKV CP at any of these eight amino acids proposed
470 here to control the intra- and intercapsomere interactions increase the percentage of
471 abnormal virions. While the biological implications of Kaelber and co-workers' observations
472 are unclear, the fact that EEEV and VEEV have a higher rate of misassembled virions could
473 be interpreted as an attenuation mechanism. It is essential to note that EEEV and VEEV are
474 lethal viruses, whereas CHIKV is rarely deadly (31). Therefore, there is the possibility that
475 encephalitic Alphaviruses may be more likely to make more mistakes during assembly to

476 achieve attenuation, thereby increasing the probability of infecting new hosts before the
477 original host dies. However, this hypothesis is yet to be tested.

478 Mukhopadhyay and coworkers cryo-EM reconstructed Ross River virus and WEEV *in*
479 *vitro* assembled core-like particles (CLPs) and could not achieve a resolution higher than 30
480 Å (7). They also attempted to reconstruct cores produced in mammalian cells, in which the
481 lipid bilayer and the E1/E2 layer were removed, obtaining similar results to those with the
482 CLPs. Therefore, they proposed that their inability to achieve the same resolution as intact
483 virions suggested that CLPs deviate from an icosahedron and/or are heterogeneous (7).
484 Recently, Chmielewski and co-workers used cryo-ET to reconstruct CHIKV particles as they
485 budded from cultured cells (4). In this work, they demonstrated that the cores in the
486 cytoplasm are not icosahedral, and the double icosahedral symmetry of the virion is
487 assembled during budding. Although Chmielewski and coworkers did not investigate the
488 molecular assembly mechanism of CHIKV virions, their results can be correlated with the
489 hypothesis presented here. The transition from a disordered to an icosahedral core can be
490 viewed as an annealing process in which the cytoplasmic core is not in its lowest free-energy
491 state. It is upon interaction with the glycoproteins at the plasma membrane that multiple
492 protein-protein interactions overcome the loss of entropy (because of the transition from
493 disordered to icosahedral core). Based on data from Chmielewski and co-workers (4), it is
494 reasonable to propose that weak interactions would facilitate this mechanism; for example,
495 the proposed dipole-dipole-like interactions between capsomers depend on the distance
496 and angle between them. An angle dependency would facilitate the reordering of a
497 heterogeneous core into an icosahedral structure as observed by Cryo-ET (4).

498 We investigated the evolution of the CP, and by only focusing on the CP, we could infer
499 that the common ancestor of non-fish-infecting Alphaviruses had almost the same putative
500 CP-CP and capsomer-capsomer interactions as Old-World Alphaviruses. On the one hand,
501 New-World Alphaviruses are mostly restricted to the American continent, yet there is
502 significant variability in these interactions. On the other hand, Old-World Alphaviruses can
503 be found in the rest of the world, and these interactions are highly conserved.

504 Our analysis of Shannon entropy and selective pressure clearly shows that the CPs
505 of these two groups of Alphaviruses differ significantly; the amino acids proposed to control
506 core assembly are invariant in Old-World viruses. Altogether, our results, along with those
507 obtained by studying the evolution of Alphaviruses using the glycoproteins (which are
508 responsible for viral tropism) and the virulence factor TF (17, 18), which resulted in the
509 generation of arthrogenic and encephalitic Alphaviruses, could suggest a relationship
510 between the CP-CP that control core assembly and pathogenicity. The event that resulted in
511 a rapid evolution of the CP on the American continent, and allowed for weak CP-CP
512 interactions and possibly core defects (as observed in VEEV and WEEV (30)) could have
513 resulted in a viral attenuation mechanism. This hypothesis is consistent with the fact that
514 the glycosylation profiles of E2 and E1 differ between these two groups of Alphaviruses.
515 Therefore, it is possible that glycosylation changes not only affected tropism but also
516 contributed to stabilizing cores with weak interactions (unpublished data). In other words, it
517 is possible that the "molecular handshake" proposed by Ribeiro-Filho for MAYV (12) had to
518 be removed to allow for assembly of the core when the interactions became weaker, and that
519 this event might have happened when the divergence of Old- and New-World viruses
520 occurred.

521 It has been proposed that Alphaviruses have a marine origin (18). However, a
522 phylogenetic analysis of the CP alone suggests that the marine alphaviruses infecting
523 mammals evolved later than those that infect fish. Our analysis also indicates that Harbor
524 porpoise virus was reintroduced to marine animals after the divergence of Old- and New-
525 World viruses. We are unclear when the Southern elephant seal virus (SEAV) was introduced.
526 However, given that SEAV CP has six of the eight amino acids of interest conserved with
527 CHIKV, this suggests it may have been reintroduced into a marine environment at the same
528 time as Harbor porpoise virus.

529 The inference of the CP common ancestor suggests it had seven of the eight
530 interactions we propose to control CHIKV assembly. This indicates that the event that
531 separated Old- and New-World Alphaviruses resulted in homogeneous, and most likely
532 more stable, Old-World cores. In contrast, this event favored increased heterogeneity in

533 New-World Alphaviruses. The increase in variability in the intra- and intercapsomere
534 interactions could likely be linked to the development of a lethal encephalitic disease.

535

536 **Conclusions**

537 The interactions that control core assembly in Old-World Alphaviruses differ significantly
538 from those of other Alphaviruses. The amino acids studied here are likely to play a crucial
539 role in the assembly of capsomers and the core in the cytoplasm, so that they favor the
540 transition from an irregular to icosahedral core during assembly. In Old-World Alphaviruses,
541 these interactions are highly conserved and likely emerged early in the evolution of these
542 viruses. The divergence of Old- and New-World Alphaviruses resulted not only in differences
543 in tropism and disease but in structural differences. These differences could have resulted
544 in a novel attenuation mechanism in which lethal Alphaviruses are more prone to make
545 mistakes during assembly than non-lethal Alphaviruses. However, the results presented
546 here require experimental confirmation, primarily because the cryo-EM data were not fully
547 resolved and because the AlphaFold2 and 3 models were generated from the monomeric
548 capsid protein; therefore, we may be missing protein surfaces that appear once the core is
549 assembled. Nonetheless, combining structural prediction and MSA analyses proved to be
550 crucial as we were able to find that only 12 out of 293 amino acids (4.1%) showed
551 discordance in position. This suggests that combining both approaches is likely to yield a
552 highly accurate prediction of these interactions of interest. Our group is currently testing
553 these scenarios to investigate the contribution of these amino acids by performing site-
554 directed mutagenesis in our CHIKV system (32) with a reporter gene to validate so we can
555 further validate these hypothesis on core assembly, as well as the correlation between the
556 E2 N263 glycosylation site characteristic of Old-World Alphaviruses and the identity of the
557 amino acids that control core assembly.

558

559 **Funding**

M.C-G. thanks the Mexican Government for the SEIHCITI CBF-2023-2024-1145 grant, the San Luis Potosi State Government for the COPOCYT 2024-03-M07 grant, and the CICSaB/UASLP for intramural support. A.R-C. thanks the SEIHCITI for the Ph.D. fellowship 4032286.

Acknowledgments

We also thank Jose Alberto Campillo-Balderas for his helpful comments during the writing of the manuscript.

Competing interests

The authors do not have any competing financial or non-financial interests directly or indirectly related to this work.

References

1. Ou JH, Rice CM, Dalgarno L, Strauss EG, Strauss JH. 1982. Sequence studies of several alphavirus genomic RNAs in the region containing the start of the subgenomic RNA. Proc Natl Acad Sci U S A 79:5235-9.
2. Lambrechts L, Scott TW. 2009. Mode of transmission and the evolution of arbovirus virulence in mosquito vectors. Proc Biol Sci 276:1369-78.
3. Button JM, Qazi SA, Wang JC, Mukhopadhyay S. 2020. Revisiting an old friend: new findings in alphavirus structure and assembly. Curr Opin Virol 45:25-33.
4. Chmielewski D, Schmid MF, Simmons G, Jin J, Chiu W. 2022. Chikungunya virus assembly and budding visualized in situ using cryogenic electron tomography. Nat Microbiol 7:1270-1279.
5. Cheng F, Mukhopadhyay S. 2011. Generating enveloped virus-like particles with in vitro assembled cores. Virology 413:153-60.
6. Cheng F, Tsvetkova IB, Khuong YL, Moore AW, Arnold RJ, Goicochea NL, Dragnea B, Mukhopadhyay S. 2013. The packaging of different cargo into enveloped viral nanoparticles. Mol Pharm 10:51-8.
7. Mukhopadhyay S, Chipman PR, Hong EM, Kuhn RJ, Rossmann MG. 2002. In vitro-assembled alphavirus core-like particles maintain a structure similar to that of nucleocapsid cores in mature virus. J Virol 76:11128-32.

592 8. Zhang YN, Deng CL, Li JQ, Li N, Zhang QY, Ye HQ, Yuan ZM, Zhang B. 2019. Infectious
593 Chikungunya Virus (CHIKV) with a Complete Capsid Deletion: a New Approach for a
594 CHIKV Vaccine. J Virol 93.
595 9. Tang J, Jose J, Chipman P, Zhang W, Kuhn RJ, Baker TS. 2011. Molecular links between
596 the E2 envelope glycoprotein and nucleocapsid core in Sindbis virus. J Mol Biol
597 414:442-59.
598 10. Wang A, Zhou F, Liu C, Gao D, Qi R, Yin Y, Liu S, Gao Y, Fu L, Xia Y, Xu Y, Wang C, Liu Z.
599 2022. Structure of infective Getah virus at 2.8 A resolution determined by cryo-
600 electron microscopy. Cell Discov 8:12.
601 11. Sun S, Xiang Y, Akahata W, Holdaway H, Pal P, Zhang X, Diamond MS, Nabel GJ,
602 Rossmann MG. 2013. Structural analyses at pseudo atomic resolution of
603 Chikungunya virus and antibodies show mechanisms of neutralization. Elife
604 2:e00435.
605 12. Ribeiro-Filho HV, Coimbra LD, Cassago A, Rocha RPF, Guerra J, de Felicio R, Carnieli
606 CM, Leme L, Padilha AC, Paes Leme AF, Trivella DBB, Portugal RV, Lopes-de-Oliveira
607 PS, Marques RE. 2021. Cryo-EM structure of the mature and infective Mayaro virus at
608 4.4 A resolution reveals features of arthritogenic alphaviruses. Nat Commun 12:3038.
609 13. Sutton MS, Pletnev S, Callahan V, Ko S, Tsybovsky Y, Bylund T, Casner RG, Cerutti G,
610 Gardner CL, Guirguis V, Verardi R, Zhang B, Ambrozak D, Beddall M, Lei H, Yang ES,
611 Liu T, Henry AR, Rawi R, Schon A, Schramm CA, Shen CH, Shi W, Stephens T, Yang Y,
612 Florez MB, Ledgerwood JE, Burke CW, Shapiro L, Fox JM, Kwong PD, Roederer M.
613 2023. Vaccine elicitation and structural basis for antibody protection against
614 alphaviruses. Cell 186:2672-2689 e25.
615 14. Kafai NM, Williamson LE, Binshtein E, Sukupolvi-Petty S, Gardner CL, Liu J, Mackin S,
616 Kim AS, Kose N, Carnahan RH, Jung A, Droit L, Reed DS, Handley SA, Klimstra WB,
617 Crowe JE, Diamond MS. 2022. Neutralizing antibodies protect mice against
618 Venezuelan equine encephalitis virus aerosol challenge. J Exp Med 219.
619 15. Williamson LE, Gilliland T, Jr., Yadav PK, Binshtein E, Bombardi R, Kose N, Nargi RS,
620 Sutton RE, Durie CL, Armstrong E, Carnahan RH, Walker LM, Kim AS, Fox JM, Diamond
621 MS, Ohi MD, Klimstra WB, Crowe JE, Jr. 2020. Human Antibodies Protect against
622 Aerosolized Eastern Equine Encephalitis Virus Infection. Cell 183:1884-1900 e23.
623 16. Calisher CH, Karabatsos N. 1988. Arbovirus serogroups: definition and geographic
624 distribution. CRC Press,, Boca Raton, FL.
625 17. Weaver SC, Winegar R, Manger ID, Forrester NL. 2012. Alphaviruses: population
626 genetics and determinants of emergence. Antiviral Res 94:242-57.
627 18. Forrester NL, Palacios G, Tesh RB, Savji N, Guzman H, Sherman M, Weaver SC, Lipkin
628 WI. 2012. Genome-scale phylogeny of the alphavirus genus suggests a marine origin.
629 J Virol 86:2729-38.
630 19. Suhrbier A, Jaffar-Bandjee MC, Gasque P. 2012. Arthritogenic alphaviruses--an
631 overview. Nat Rev Rheumatol 8:420-9.
632 20. Zacks MA, Paessler S. 2010. Encephalitic alphaviruses. Vet Microbiol 140:281-6.
633 21. Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time
634 and space complexity. BMC Bioinformatics 5:113.

22. Kostyuchenko VA, Jakana J, Liu X, Haddow AD, Aung M, Weaver SC, Chiu W, Lok SM. 2011. The structure of barmah forest virus as revealed by cryo-electron microscopy at a 6-angstrom resolution has detailed transmembrane protein architecture and interactions. J Virol 85:9327-33.
23. Mancini EJ, Clarke M, Gowen BE, Rutten T, Fuller SD. 2000. Cryo-electron microscopy reveals the functional organization of an enveloped virus, Semliki Forest virus. Mol Cell 5:255-66.
24. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Zidek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596:583-589.
25. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, Bodenstein SW, Evans DA, Hung CC, O'Neill M, Reiman D, Tunyasuvunakool K, Wu Z, Zemgulyte A, Arvaniti E, Beattie C, Bertolli O, Bridgland A, Cherepanov A, Congreve M, Cowen-Rivers AI, Cowie A, Figurnov M, Fuchs FB, Gladman H, Jain R, Khan YA, Low CMR, Perlin K, Potapenko A, Savy P, Singh S, Stecula A, Thillaisundaram A, Tong C, Yakneen S, Zhong ED, Zielinski M, Zidek A, Bapst V, Kohli P, Jaderberg M, Hassabis D, Jumper JM. 2024. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature 630:493-500.
26. Meng EC, Goddard TD, Pettersen EF, Couch GS, Pearson ZJ, Morris JH, Ferrin TE. 2023. UCSF ChimeraX: Tools for structure building and analysis. Protein Sci 32:e4792.
27. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772-80.
28. Pletnev SV, Zhang W, Mukhopadhyay S, Fisher BR, Hernandez R, Brown DT, Baker TS, Rossmann MG, Kuhn RJ. 2001. Locations of carbohydrate sites on alphavirus glycoproteins show that E1 forms an icosahedral scaffold. Cell 105:127-136.
29. Tsetsarkin KA, Chen R, Sherman MB, Weaver SC. 2011. Chikungunya virus: evolution and genetic determinants of emergence. Curr Opin Virol 1:310-7.
30. Kaelber JT, Chmielewski D, Chiu W, Auguste AJ. 2022. Alphavirus Particles Can Assemble with an Alternate Triangulation Number. Viruses 14.
31. Baxter VK, Heise MT. 2018. Genetic control of alphavirus pathogenesis. Mamm Genome 29:408-424.
32. Colunga-Saucedo M, Rubio-Hernandez EI, Coronado-Ipina MA, Rosales-Mendoza S, Castillo CG, Comas-Garcia M. 2022. Construction of a Chikungunya Virus, Replicon, and Helper Plasmids for Transfection of Mammalian Cells. Viruses 15.

676 **Figure legends**

677 **Figure 1**. A) Comparison of the cryo-EM structures of the cores of CHIKV (PDB ID: 3J2W),
678 MAYV (7KO8), and GETV (7FD2). B) Comparison of cryo-EM structures of the cores EEEV
679 (6XO4), VEEV (7SFV), and WEEV (8DEC). Each structure has four different colors,
680 representing that in a T = 4 icosahedron, the capsid protein (CP) in the asymmetric unit has
681 four different chemical environments.

682

683 **Figure 2.** Comparison of the structure of the pentamers with the structures in Figure 1. The
684 proteins in the pentamer have the same color because they share the same chemical
685 environment, unlike the proteins in the hexamer, which have three distinct chemical
686 environments. A) The structures of the pentamers of these three Old-World Alphaviruses are
687 very similar. B) The structure of the EEEV pentamer has fewer contact angles (asterisks) than
688 the other New-World Alphaviruses. The contacts between CPs in VEEV and WEEV are similar
689 to those of CHIKV, MAYV, and GETV.

690

691 **Figure 3.** A) Putative electrostatic interactions at the surface between three proteins in the
692 CHIKV pentamer, between E184 and E234 of one protein and K172 and K177 of the other
693 protein. B) The interactions at the quasi-threefold symmetry axis resemble dipole-dipole
694 interactions between the pair of amino acids E120/K122 and D148/K151. C) CHIKV CP (PDB
695 ID: 3J2W) monomer displaying E120, K122, D148, K151, K172, K177, E184, E234. D) Overlap
696 of CHIKV CP with MAYV CP (PDB ID: 7KO8), where, although the numbering and chemical
697 identity of some amino acids differ, the charges are conserved in all amino acids. E) Overlap
698 of CHIKV CP with EEEV CP (PDB ID: 8DEC). This overlap indicates that some amino acids are
699 not conserved in EEEV, suggesting that the intra- and inter-capsomere interactions in EEEV
700 differ significantly from those in MAYV and CHIVK.

701

Figure 4. Comparison of the interfaces between the three proteins at i5 (pentamer). A) The "outer interactions" for CHIKV are between K172 and E234, and the "inner interactions" are between K177 and E184. It also shows that the possible interactions between amino acids at equivalent positions at the same interfaces for MAYV and GETV are the same as in CHIKV. B) The "outer interactions" for EEEV, VEEV, and WEEV are the same as in the Old-World Alphaviruses from the top row. However, the "inner interactions" are different. The inner interactions for EEEV involve a P and L, which could explain the large holes observed in Figure 2. In VEEV, the amino acids at the "inner position" are Gln and Lys, while for WEEV, they are P and Q.

Figure 5. A) The leftmost panel shows the geometry of the i3 axis for CHIVK, MAYV, and GETV, which was determined by measuring the centroid-to-centroid distance (2-Å centroid) and torsion angle between K122 of the three proteins at this axis. B) The distances between these three viruses and EEEV and WEEV, as well as those in the top row, are very similar. Only VEEV has a distance that is much smaller than that of the other viruses. Also, the distances for EEEV and WEEV were calculated by using Q121 and Q119, respectively, which are in equivalent positions to CHIKV CP K122. Interestingly, in VEEV, K134 and E133 do not have a pair of charged amino acids, as seen in the top row, that could help establish a dipole-dipole-like interaction. The torsion angles between three equivalent amino acids are indicated at the center of the i3 axes. The torsion angles for CHIKV, MAYV, and EEEV are very similar; however, VEEV and WEEV have larger torsion angles than the other four viruses.

Figure 6. Comparison of the identity of amino acids in equivalent positions to CHIKV CP E120, K122, D148, K151, K177, E184, and E234 for Old-World Alphaviruses. These amino acids were identified by predicting the structure of the consensus CP sequence of all viruses using AlphaFold2. The green color indicates that the charge and/or identity of the amino acids are conserved at least 80%.

730 **Figure 7.** Comparison of the identity of amino acids in equivalent positions to CHIKV CP
731 E120, K122, D148, K151, K177, E184, and E234 for New-World Alphaviruses. These amino
732 acids were identified, as explained in Table 1. The lavender and turquoise colors indicate that
733 these amino acids are at equivalent positions to those in CHIKV due to an insertion or a -1
734 shift, respectively. Only the positions equivalent to K172 and E234 are conserved at 94% and
735 100%, respectively.

736

737 **Figure 8.** Comparison of the identity of amino acids in equivalent positions to CHIKV CP
738 E120, K122, D148, K151, K177, E184, and E234 for Alphaviruses with no known host is not
739 part of the Old-/New-World classifications. These amino acids were identified, as explained
740 in Table 1. The lavender color indicates that those amino acids are at equivalent positions to
741 CHIKV due to the insertion. Only the positions equivalent to K172 and E234 are conserved at
742 94% and 100%, respectively. In this group, positions K172 and E234 are conserved, as well
743 as E120 and K172.

744

745 **Figure 9.** Comparison of the identity of amino acids in equivalent positions to CHIKV CP
746 E120, K122, D148, K151, K177, E184, and E234 for aquatic Alphaviruses. These amino acids
747 were identified, as explained in Table 1. The lavender, turquoise, and maroon colors indicate
748 that those amino acids are at equivalent positions to CHIKV due to an insertion, a -1 shift, or
749 a deletion, respectively. Positions K172, E234, E120, and K172 are conserved. The gray color
750 indicates that those viruses infect mammals. The rest infect fish.

751

752 **Figure 10.** A) Except for the Bebaru virus, all other Old-World Alphaviruses in this table have
753 a predicted N-glycosylation site at a position that is equivalent to CHIKV N263. B) Cryo-EM
754 structure of MAYV virus (PDB ID: 7KO8) showing the "molecular handshake" between
755 glycosylation sites. This interaction occurs between two trimers of heterodimers. C) New-
756 World Alphaviruses do not have a predicted N-glycosylation at an equivalent position to

757 CHIKV N263. The only exception is the Fort Morgan virus, which has a possible glycosylation
758 site at N260, which is not in the same position as CHIKV E2 N263. D) Overlap of CHIKV E2
759 (green) with the predicted structure for the Fort Morgan virus E2 protein (salmon). CHIKV E2
760 N263 is in red and indicated by a red # and is in a β-sheet. The amino acid in the Fort Morgan
761 virus E2 protein equivalent to CHIKV E2 N263 is in black and indicated by a black asterisk
762 (Proline), while the Asp in an N-L-T motif is in a loop and is indicated with a red asterisk. If
763 this amino acid were glycosylated, it is unlikely to be able to establish the same interactions
764 as the glycosylation site shown in panel B. E) From the Alphaviruses that infect aquatic
765 animals with available E2 sequences, only the Harbor Alphavirus has the same predicted E2
766 glycosylation site as CHIKVK. Interestingly, the core of this virus has seven of the eight
767 interactions conserved in the CPs of Old-World Alphaviruses, as shown in Figures 4, 5, and
768 6 and Tables 1 and 4. This small correlation could suggest that the nature of the core and the
769 interactions between trimers are related to each other.

770

771 **Figure 11.** Phylogenetic tree for the Alphaviruses in Tables 1-4. The predictions of possible
772 common ancestors made from the data in Tables 1-4 are supported by this tree. The letters
773 refer to the nodes of interest: *** is the common ancestor of the CP for Old- and New-World
774 Alphaviruses (OWA and NWA, respectively). The table shows the identity of the amino acids
775 of interest in nodes *** and A-G; the green and red colors indicate whether the charge of
776 these amino acids is conserved or not with respect to CHIVK, respectively. The colors of the
777 branches indicate the different clades. The prediction of the sequence of this node suggests
778 that the seven amino acids of interest, which are highly conserved in Old-World
779 Alphaviruses, were present in this ancestor; however, the origin of the amino acids observed
780 in Old-World Alphaviruses today arose early in their evolution (see B). The node that gave
781 origin to Sindbis is shared with Old-World Alphaviruses. At the same time, the early
782 beginning of the rest of the New-World Alphaviruses started with a fast diversification of the
783 intra- and intracapsomeric interactions. As these viruses evolved, the diversity of these
784 interactions increased significantly.

785 **Table 5.** Comparison of the Shannon entropy and selective pressure of amino acids in
786 equivalent positions to CHIKV CP E120, K122, D148, K151, K177, E184, and E234 Old-World
787 Alphaviruses. This analysis was performed using SLAC, FEL, FUBAR, and MEME methods;
788 sites were classified as under pervasive or episodic positive selection if detected by MEME
789 (p<0.1), FUBAR (posterior probability >0.9), FEL (p<0.1), or SLAC (p<0.1).

790

791

Figures

Figure 1

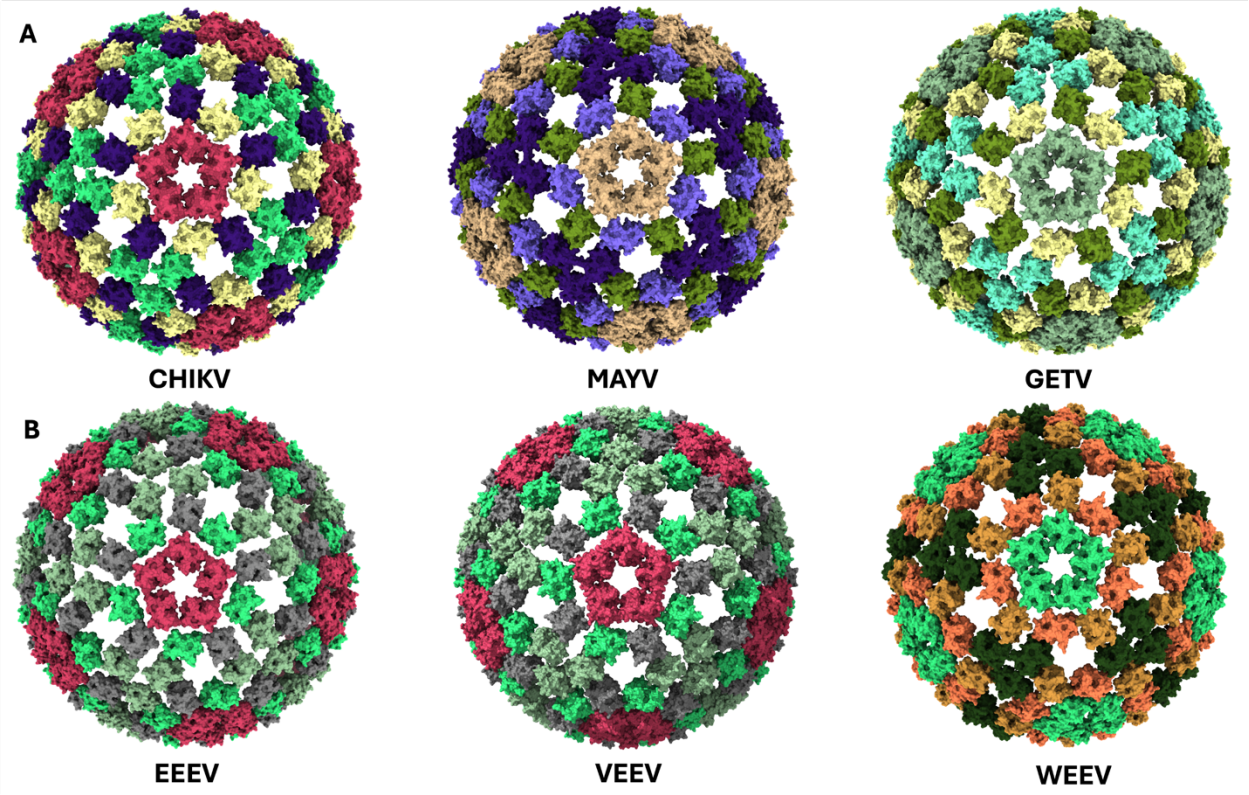


Figure 2

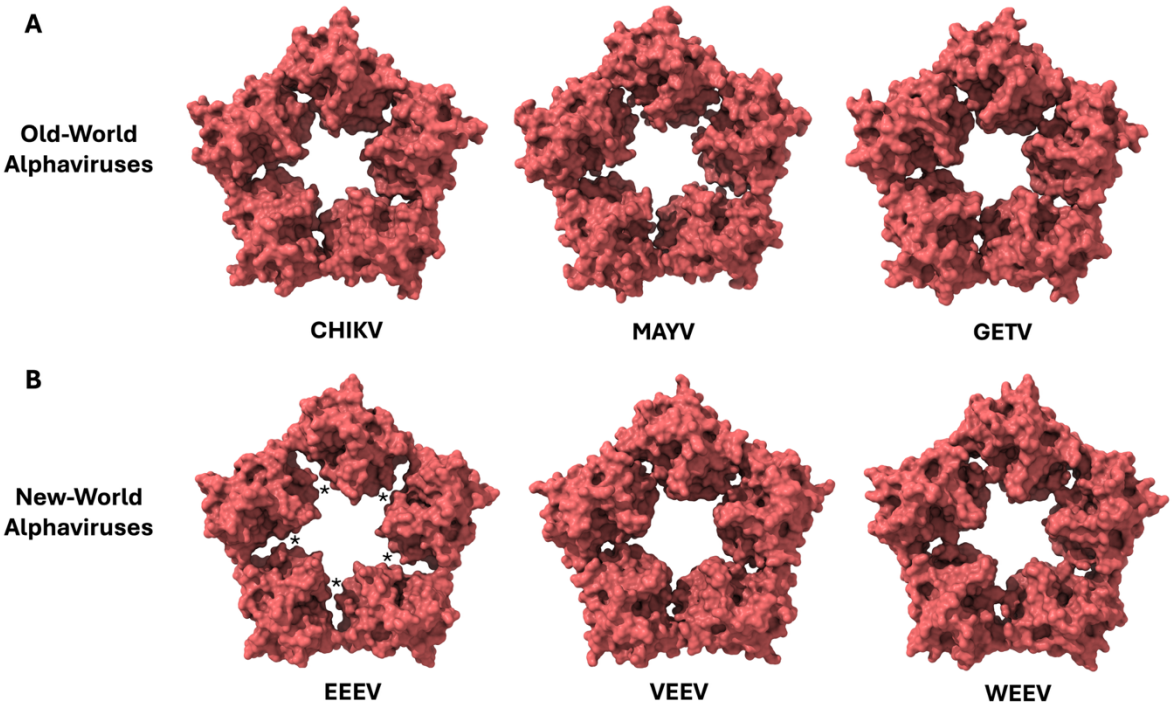

A
Old-World Alphaviruses
CHIKV
MAYV
GETV
B
New-World Alphaviruses
EEEV
VEEV
WEEV

Figure 3

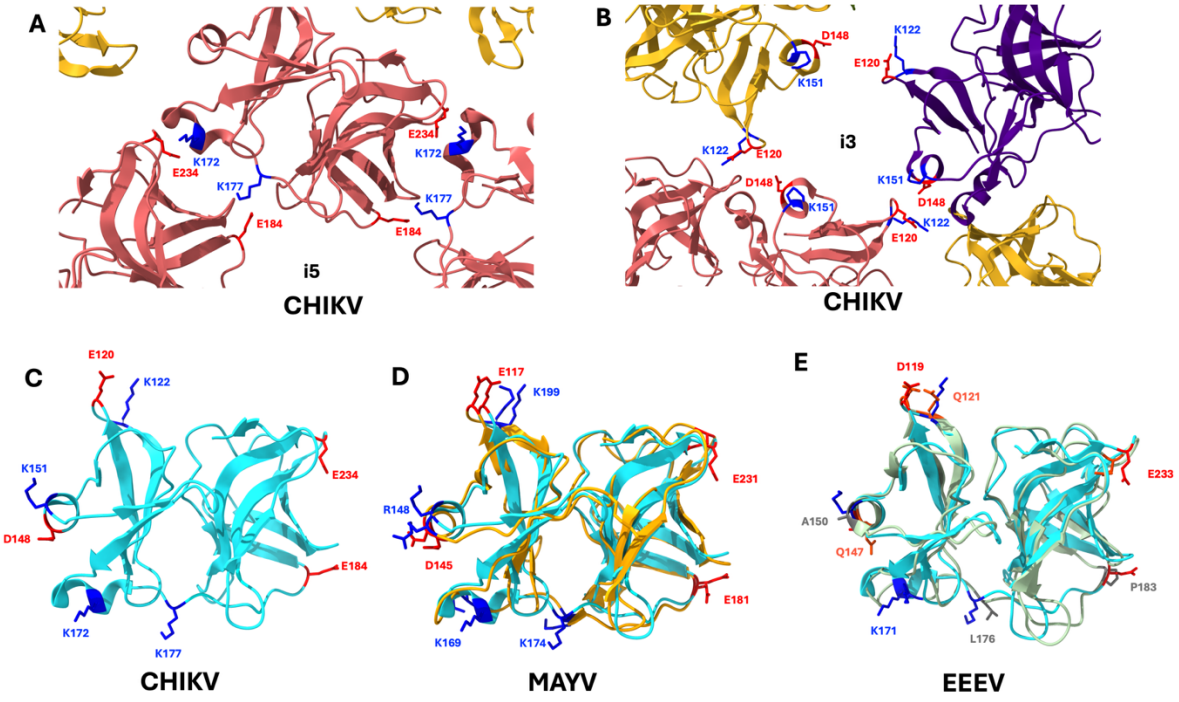

A
K172
E234
K177
E184
E234
K172
K177
E184
i5
CHIKV
B
D148
K122
E120
K151
i3
K122
E120
D148
K151
K151
D148
K122
E120
CHIKV
C
E120
K122
K151
E234
D148
E184
K172
K177
CHIKV
D
E117
K199
E231
R148
D145
E181
K169
K174
MAYV
E
D119
Q121
E233
A150
Q147
P183
K171
L176
EEEV

Figure 4

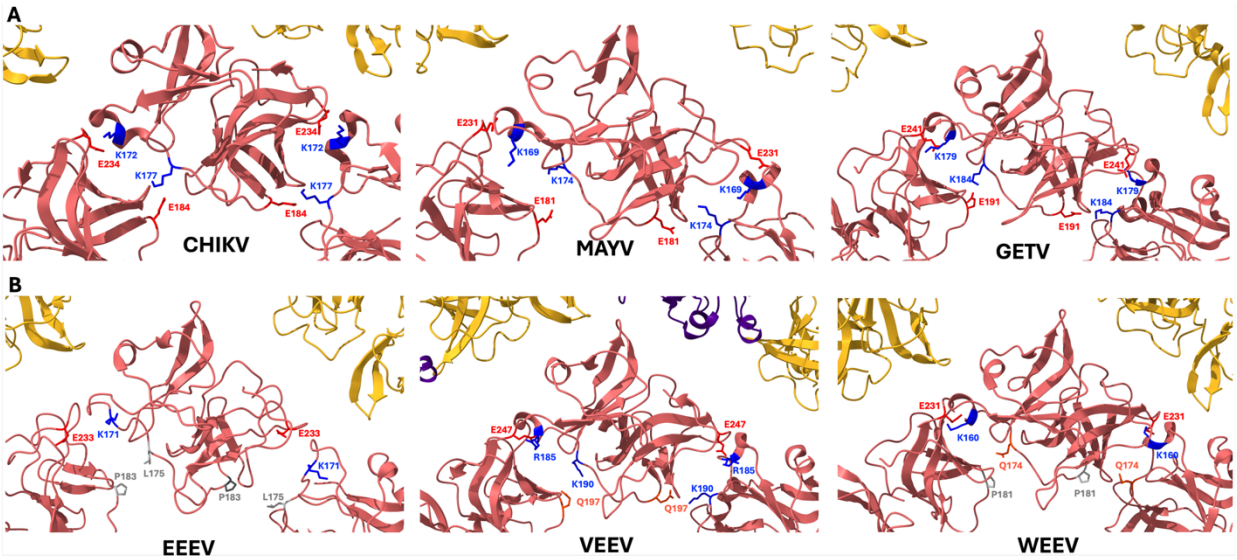
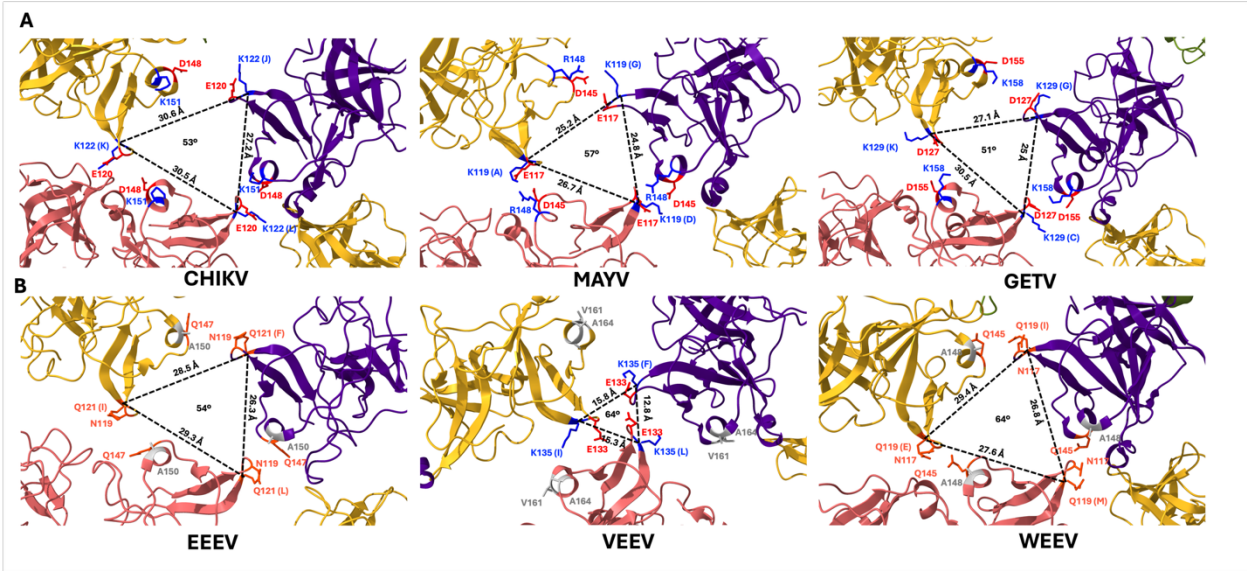

A
K172
E234
K177
E184
E234
K172
K177
E184
CHIKV
E231
K169
K174
E181
E231
K169
K174
E181
MAYV
E241
K179
K184
E191
E241
K179
K184
E191
GETV
B
E233
K171
P183
L175
E233
K171
P183
L175
EEEV
E247
R185
K190
Q197
E247
R185
K190
Q197
VEEV
E231
K160
Q174
P181
E231
K160
Q174
P181
WEEV

Figure 5



819 **Figure 6**

820

	Virus	Conservation of the amino acids of interest in the Old World Alphaviruses with respect to CHIKV							
	CHIKV (Reference)	E120	K122	D148	K151	K172	K177	E184	E234
SF complex	CHIKV	E	K	D	K	K	K	E	E
	Getah virus	D	K	D	K	K	K	E	E
	Semliki Forest virus	E	K	D	K	R	K	E	E
	Mayaro virus	E	K	D	R	K	K	E	E
	Ross River	D	K	D	K	K	K	E	E
	Bebaru virus	D	K	D	K	R	K	E	E
	Onyong-nyong virus	E	K	D	K	K	K	E	E
	Una virus	D	K	D	K	R	K	E	E
BF complex	Barmah forest virus	D	K	E	K	K	K	E	E
MIID	Middelburg virus	E	K	D	K	K	Q	E	E
Ndumu complex	Ndumu virus	D	E	E	K	R	K	E	E
	% ID conservation with respect to CHIKV	45	91	82	91	64	91	100	100
	% Charge conservation with respect to CHIKV	100	91	100	100	100	91	100	100

821

822

823 **Figure 7**

	Virus	Conservation of the amino acids of interest in the New World Alphaviruses with respect to CHIKV							
	CHIKV (Reference)	E120	K122	D148	K151	K172	K177	E184	E234
WEE complex	Western equine encephalitis virus	N	Q	Q	A	K	Q	P	E
	Sindbis virus	D	D	V	K	R	T	E	E
	Fort Morgan virus	N	Q	E	A	K	Q	P	E
	Aura virus	D	K	A	K	K	G	E	E
	Highlands J virus	N	Q	Q	A	K	Q	P	E
	Whataroa virus	E	D	V	K	K	N	E	E
EEE complex	Eastern equine encephalitis virus	N	Q	Q	A	K	L	P	E
VEE complex	Venezuelan equine encephalitis virus	E	K	V	A	R	K	Q	E
	Everglades virus	E	K	V	A	R	K	Q	E
	Cabassou virus	D	K	T	A	R	R	Q	E
	Mucambo virus	D	K	T	S	R	K	Q	E
	Pixuna virus	E	R	A	A	R	K	Q	E
	RioNegro virus	D	R	V	T	R	K	Q	E
	Tonate virus	D	K	T	A	R	K	Q	E
Caaingua complex	Caaingua virus	D	K	E	N	I	K	M	E
Madariaga complex	Madariaga virus	N	Q	Q	A	K	Q	P	E
MdP complex	Mosso das Pedras virus	D	K	I	A	R	K	R	E
	% ID conservation with respect to CHIKV	24	47	0	18	41	47	18	100
	% Charge conservation with respect to CHIKV	71	59	12	18	94	53	18	100

824

825

826

827

828

829

830 **Figure 8**

	Virus	Conservation of the amino acids of interest in Alphaviruses found in Mosquitoes with no vertebrate host with respect to CHIKV							
	CHIKV (Reference)	E120	K122	D148	K151	K172	K177	E184	E234
Eilat complex	Eilat virus	K	V	L	S	R	A	D	E
Trocara complex	Trocara virus	E	K	A	K	R	K	E	Q
Unclassified	AguaSalud alphavirus	D	Q	I	K	K	R	E	E
Unclassified	Mwinilunga alphavirus	D	A	L	T	R	N	D	E
Unclassified	Pirahy virus	D	K	T	S	R	K	Q	E
Unclassified	Tai Forest alphavirus	D	V	L	S	R	S	D	E
Unclassified	Yada yada virus	E	R	T	S	R	A	D	E
	% ID conservation with respect to CHIKV	29	29	0	29	14	29	29	86
	% Charge conservation with respect to CHIKV	86	43	0	29	100	43	86	86

831

832

833 **Figure 9**

834

	Virus	Conservation of the amino acids of interest in aquatic Alphaviruses with respect to CHIKV							
	CHIKV (Reference)	E120	K122	D148	K151	K172	K177	E184	E234
SPDV	Salmon pancreatic disease virus	D	A	E	D	R	E	D	D
SES complex	Southern Elephant Seal virus	D	L	I	K	R	K	E	E
Trocara complex	Comber alphavirus	E	T	E	K	K	K	D	D
Unclassified	Harbor alphavirus	D	K	V	K	K	K	D	E
Unclassified	Norwegian salmonid	D	A	E	D	R	E	D	D
Unclassified	Wenling fish alphavirus	K	T	L	E	H	A	E	E
Unclassified	Wenling hagfish alphavirus	E	G	D	N	R	K	D	D
	% ID conservation with respect to CHIKV	29	14	14	43	29	57	29	43
	% Charge conservation with respect to CHIKV	86	14	57	43	86	57	100	100

835

836

837

838

839

840

841

842

843

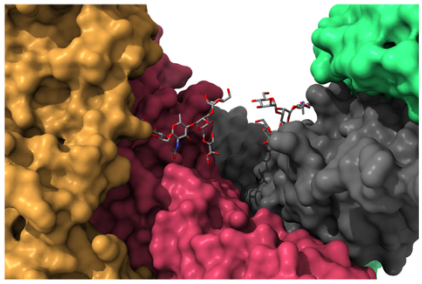
844

845 **Figure 10**

A

Virus	Table 5. Conservation of the amino acids of interest in the Old World Alphaviruses with respect to CHIKV			
CHIKV (Reference)	N263	+1	+2	Possible N-Gly
CHIKV	N	V	T	Yes
Getah virus	N	S	T	Yes
Semliki Forest virus	N	I	T	Yes
Mayaro virus	N	V	T	Yes
Ross River	N	T	T	Yes
Bebaru virus	G	V	T	No
Onyong-nyong virus	N	T	T	Yes
Una virus	N	S	T	Yes
Middelburg virus	N	S	T	Yes
Ndumu virus	N	T	T	Yes
% ID conservation with respect to CHIKV	90		100	

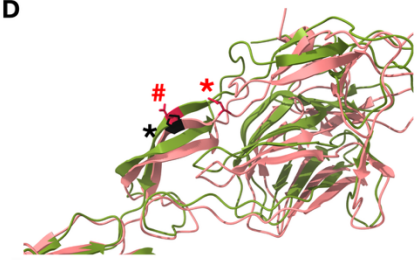
B



C

Virus	Table 6. Conservation of the amino acids of interest in the New World Alphaviruses with respect to CHIKV			
CHIKV (Reference)	N263	+1	+2	Possible N-Gly
Sindbis virus	P	S	T	No
Fort Morgan virus	P	T	T	No
	N	L	T	Yes
Aura virus	Q	A	E	No
Highlands J virus	A	T	A	No
Western equine encephalitis virus	P	T	V	No
Whataroa virus	P	G	Q	No
Eastern equine encephalitis virus	K	A	K	No
Venezuelan equine encephalitis virus	D	G	K	No
Everglades virus	D	G	K	No
Cabassou virus	E	A	K	No
Mucambo virus	E	A	K	No
Pixuna virus	D	A	R	No
RioNegro virus	G	A	K	No
Tonate virus	D	A	K	No
Madariaga virus	T	S	K	No
Mosso das Pedras virus	E	A	K	No
% ID conservation with respect to CHIKV	6		18	

D



E

Virus	Table 7. Conservation of the amino acids of interest in aquatic Alphaviruses with respect to CHIKV			
CHIKV (Reference)	N263	+1	+2	Possible N-Gly
Salmon pancreatic disease virus	A	A	T	No
Harbor porpoise alphavirus	N	V	T	Yes
Norwegian salmonid	T	A	T	No
% ID conservation with respect to CHIKV	33		33	

846

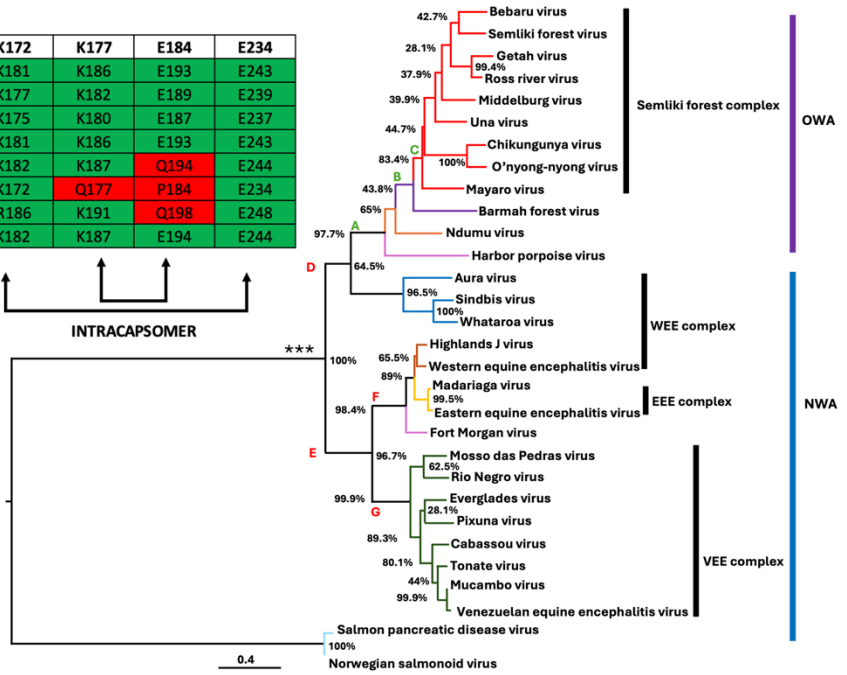
847

Figure 11

	E120	K122	D148	K151	K172	K177	E184	E234
A	D129	K131	V157	K160	K181	K186	E193	E243
B	D125	K127	D153	K156	K177	K182	E189	E239
C	D123	K125	D151	K154	K175	K180	E187	E237
D	D129	K131	V157	K160	K181	K186	E193	E243
E	D130	K132	V158	A161	K182	K187	Q194	E244
F	N120	Q122	Q148	A151	K172	Q177	P184	E234
G	D134	K136	V162	A165	R186	K191	Q198	E248
***	D130	K132	V158	K161	K182	K187	E194	E244

INTERCAPSOMER

INTRACAPSOMER



861

862 **Figure 12**

V irus	Selection						
	-	-	-	-	Purifying*	-	
	Entropy						
	0.689	0.305	0.305	0.305	0.655	0.000	0.305
Chikungunya Virus/MK286896	120	122	148	151	172	184	234
	E	K	D	K	K	E	E
Barmah Forest Virus/MW556197	112	114	140	143	164	176	226
	D	K	E	K	K	E	E
Bebaru Virus/NC 016962	122	124	150	153	174	186	236
	D	K	D	K	R	E	E
Semliki Virus/KF283988	126	128	154	157	178	190	240
	E	K	D	K	R	E	E
Getah Virus/PP236766	127	129	155	158	179	191	241
	D	K	D	K	K	E	E
Ross River Virus/NC 001544	129	131	157	160	181	193	243
	D	K	D	K	K	E	E
Una Virus/AF339481	138	140	166	169	190	202	252
	D	K	D	K	R	E	E
Middleburg Virus/AF398374	130	132	158	161	182	194	244
	E	K	D	K	K	E	D
Mayaro Virus/KP842808	117	119	145	148	169	181	231
	E	K	D	R	K	E	E
O nyongnyong Virus/NC 075006	119	121	147	150	171	183	233
	E	K	D	K	K	E	E
Ndumu Virus/MZ091531	128	130	156	159	180	192	242
	D	E	D	K	R	E	E

*Detected by two different algorithms

*Detected by one or no algorithms

863

864

865

866

867

868

869

870