Full title :

Understanding evolutionary and functional relationships of RNA polymerases in plant and fungal viruses through structural modeling and divergence date estimations

Short title : Understanding RNA polymerase evolution in plant and fungal viruses

Authors :

Jérôme Gracy[1], Mahan Ghafari[2], Gilles Labesse[1], Denis Fargette[3], Eugénie Hébrard[3*]

[1]Centre de Biochimie Structurale, Montpellier, France

[2]Department of Biology, University of Oxford, UK

[3]PHIM Plant Health Institute, Univ Montpellier, IRD, CIRAD, INRAE, Institut Agro, Montpellier, France

*correspondence : eugenie.hebrard@ird.fr

Abstract

RNA-dependent RNA polymerases (RdRps) are crucial for replication of RNA viruses and serve as key marker genes for defining deep taxonomic ranks and for understanding viral evolutionary history. Despite shared functions and conserved amino acid motifs, the high genetic diversity of RdRps complicates precise sequence comparisons across viral families, hindering accurate taxonomic classification of new species – especially important at the age of metagenomics. When available, three-dimensional (3D) RdRp structures can help address these challenges through structure-based alignments. However, such structures are scarce for understudied viruses infecting fungi and plants, preventing the investigation of their ecological and evolutionary links.

In this study, we focused on the highy divergent order *Sobelivirales*. Using deep-learning structural modeling, we generated highly reliable 3D models on 20 representative viral species. Multiple structural alignment enabled the reconstruction of a robust phylogeny with improved quality and length. Based on this phylogeny, we proposed revisions of existing viral families and reclassified genera. Clade divergence dates were then estimated using the Prisoner of War model, which has previously revealed the ancient origin of the genus *Sobemovirus*. We provided here the first divergence time estimation between these plant and fungal viruses, dating back to 26.6 million years before present - significantly more recent than the divergence between their respective hosts. Our amino acid conservation analysis, validated on 99 other viral species, also identified molecular signatures of sobeliviral families and genera, which could help in future taxonomic assignment and diagnostic tools development. This interdisciplinary approach integrating structure modeling and date estimations offers new insights into the evolutionary divergence between fungus and plant viruses, with potential applications to other viral orders and families.

Author summary

RNA-dependent RNA polymerases (RdRps) are essential for RNA virus replication and serve as important markers for classifying viruses and understanding their evolution. However, with the rising popularity of metagenomics and discovery of viruses with high genetic diversity in RdRps, it is difficult to compare viral families and accurately classify new species. When available, 3D structures of RdRps can help overcome this challenge through structural alignments.

In this study, we focused on the highly divergent order of the Sobelivirales, using deep-learning models to generate reliable 3D structures for 20 representative viral species. These structural alignments allowed us to build a more accurate viral phylogeny. Based on this finding, we proposed updates to existing viral families and genera within the Sobelivirales order. We also estimated divergence dates using a model that previously uncovered the ancient origins of Sobemovirus. Notably, we provided the first estimate of when these plant and fungal viruses diverged - around 26.6 million years ago, which is much more recent than the separation of their hosts. We also identified molecular signatures that are useful for future virus classification and diagnosis, with potential applications to other viral groups.

58 **Introduction**

59 Recent advances in viral metagenomics have revealed how extensive and unexplored
60 viral diversity is, and have contributed significantly to a better understanding of viral evolution.
61 Major adaptations of viral taxonomic classification have been made to incorporate numerous
62 new lineages and their evolutionary links [1,2]. This includes the creation of novel upper ranks
63 and subdivisions such as classes and phyla [3]. The identification of viral hallmark genes
64 (VHG) facilitated the assignment of viral taxa into these novel groups [4,5]. Polymerases
65 especially RNA-dependent RNA polymerases (RdRp) represent one of the major groups of
66 VHG due to their crucial function for realm *Riboviria* (RNA viruses). RdRp comparative
67 analyses revealed unexpected relationships between some viral groups, prompting taxonomic
68 reassessments and modifications. For example, the establishment of the order *Reovirales* and
69 the dissolution of the family *Reoviridae* by the International Committee on Taxonomy of
70 Viruses (ICTV) were informed by detailed RdRp analyses [6]. This new ordering approach
71 which groups viruses with a shared replication strategy and a conserved RdRp core has allowed
72 *Reovirales* to be further subdivised into the families *Sedoreovirinae* and *Spinareovirinae*, based
73 on both major differences on the arrangement of RdRp conserved motifs and on the coat protein
74 (CP) morphology - the other crucial VHG. Similarly, the family *Luteoviridae* had initially
75 grouped the genera *Luteovirus* and *Polerovirus*, primarly due to their similar circulative
76 transmission mode mediated by aphids and the related readthrough domain of their CP.
77 However, based on the low sequence similarity of their RdRps, the family *Luteoviridae* was
78 abolished and the two genera were subsequently reassigned to the families *Tombusviridae* or
79 *Solemoviridae*, respectively [7].

80 Viral RdRps are monomeric proteins responsible for synthesizing complementary RNA
81 strands from a template RNA. They exhibit an overall structure often likened to a cupped human

82 right hand in which the fingers and thumb domains interact to form a characteristic central
83 tunnel (for review [8–11]). The palm domain is formed by a β-sheet with three to six β-strands
84 above two helices, the fingers domain is a mixed α/β structure and the thumb domain is highly
85 variable with a predominantly helical structure. The fingers and thumb domains contribute to
86 the RdRp overall stability and proper positioning of the RNA template and nascent RNA strand.
87 The palm domain is essential for the accurate and efficient RNA chain elongation. It allows the
88 positioning of the primer 3'OH, or the viral protein genome-linked (VPg) depending on the viral
89 taxon, the binding of incoming nucleotides and their incorporation to the RNA chain using two
90 metal ions as cofactors. Comparisons of RdRp sequences of diverse viral groups has identified
91 short linear motifs of highly conserved amino acids [12,13]. Five motifs (A-E) were reported
92 in the palm domain, two motifs (F and G) in the fingers domain and one putative and less
93 conserved motif (H) in the thumb domain. High throughput taxonomic assignment were often
94 based on ABC motifs [14–16]. Analyzing the three-dimensional (3D) structures of RdRps later
95 identified larger protein segments referred to as homomorphs, determined by structural
96 similarities rather than sequence composition [17]. However, for certain viral phyla like those
97 of plant and fungus viruses, lacking 3D structures, pose challenges to compare RdRps and to
98 reveal their phylogenetic relationships. To address this challenge, multiple protein alignments
99 were restricted to the central RdRp palm domain, using classical algorithms combined with
100 manual refinements based on available 3D structures from taxonomically less related viruses,
101 such as animal or human viruses. This approach led to the creation of the kingdom
102 *Orthornavirae* [4].

103 Although the monophyly of viruses in the order *Orthornavirae* has been established
104 based on comparison of their RdRps, reconstruction of their evolutionary history remains
105 difficult due to their strong, probably ancient, evolutionary divergence and the central role of
106 horizontal transfer of other viral genes in plant viruses [18]. In the picorna-like supergroup, now

107 referred as the phylum *Pisuviricota*, viral species share common structural and functional
108 features in RdRp. On the other hand, they also exhibit contrasting biological characteristics in
109 terms of host range (animals, insects, plants, fungi, protists), as well as variations in genomic
110 classification (positive single-stranded RNA or double-stranded RNA) and size (ranging from
111 the largest to the smallest RNA genomes) (see S1 Table). To investigate phylogeny of viral
112 taxons such as Pisiviricota can provide evolutionary hypotheses about the origins of these
113 distinct or shared traits. Viruses of fungi also named mycoviruses have been understudied, but
114 recent discoveries, such as cross-kingdom transmission events, highlight ecological and
115 evolutionary links with plant viruses [18–20]. In this study, we compared the RdRps inside the
116 order *Sobelivirales*, as markers of their common history to gain insights into host range
117 evolution. This order is one of the four orders of positive-sense single-stranded RNA (+ssRNA)
118 mycoviruses. It comprises families *Solemoviridae* and *Barnaviridae*, which contains plant and
119 fungus viruses, respectively. Solemoviridae includes four plant viral genera (*Sobemovirus*,
120 *Polerovirus*, *Enamovirus* and *Polemovirus*), while Barnaviridae consists of a single genus,
121 *Barnavirus*, composed of one approved and two tentative species infecting fungus (Table 1).
122 These families share the same genomic arrangement, chymotrypsine-like protease / VPg / RdRp
123 [21,22], but their sequence conservation is restricted to very short linear motifs. Similarities
124 between RdRp active sites of sobemovirus and polero-/enamovirus were noted as early as 1991
125 [12], leading finally to their taxonomic grouping in 2021 [23]. Close to poleroviruses, the
126 unique polemovirus species results from a recombination event with sobemovirus CP. Although
127 RdRp homologies between sobemoviruses and barnaviruses were proposed in 2008 [10], they
128 have not yet translated into any taxonomic revisions. Furthermore, a newly proposed genus of
129 fungal virus, *Hubsclerovirus*, has recently been assigned to the phytoviral family *Solemoviridae*
130 instead of *Barnaviridae* [24]. Thus, comparing more precisely the RdRps from these two viral
131 families would provide valuable insights into the complex history of plant and fungus viruses.

132 Our focus on Sobelivirales enabled the application of modern methodologies to
133 reconstruct their deep evolutionary histories. Notably, the genus *Sobemovirus* exemplifies a
134 unique case where the impact of time-dependent rate phenomenon (TDRP) on underestimation
135 divergence times using standard molecular clock methods has been considered using the
136 Prisoner of War model (PoW) [25]. Originally developped for reconstructing the evolutionary
137 history of human and animal viruses, the PoW model corrects for heterogeneous and time-
138 dependent evolutionary rates of viruses that may arise due to factors like epistasis, nucleotide
139 biases, and substitution saturation. The PoW model reproduces the empirically observed power-
140 law rate decay in estimated evolutionary rates of many RNA and DNA viruses across deep
141 evolutionary timescales. This method revealed a very ancient origin of sobemovirus genus (over
142 four million years) by contrast to the previous estimation based solely on the extrapolation of
143 short-term substitution rates [26]. The PoW model application needs three key prerequisites -
144 the detection of a temporal signal in time-stamped sequences, identification of non-recombinant
145 genomic regions, and a robust phylogenetic signal in the data. To overcome the last issue at the
146 high taxonomic rank of Sobelivirales, we leveraged both the structural protein conservation and
147 the advancements in deep-learning algorithms. Protein sequences evolve at a faster rate than
148 their corresponding structures. When proteins are distantly related, which is the case in this
149 group of virus sequences, mutational saturation becomes more likely and interpretation of
150 evolutionary signals becomes more challenging. The reliable structural models predicted by
151 Alphafold, which is using deep-learning for protein reconstruction, allows us to build more
152 accurate sequence alignments from highy divergent virus families [27] collected from the vast
153 number of metagenomics studies and to infer phylogenetic relationships more accurately
154 [16,28–30].

155 In this paper, applying AlphaFold on 20 representative species of sobelivirales revealed
156 the highly conserved folding of their RdRps. The subsequent superposition of all generated 3D

157 protein models with the multiple structure alignment algorithm mTM-align resulted in an robust
158 RdRp nucleotidic alignment reaching the quality requirements to estimate the divergence dates
159 with the PoW model at the scale of Sobelivirales. The analysis of amino acid conservation based
160 on the structural alignment also provided molecular signatures of sobeliviral families and
161 genera that will be useful to design taxonomical assignment and diagnostic tools in future. Our
162 interdisciplinary approach integrating structure modeling and date estimations on sobeliviral
163 RdRps provided new insights on the evolutionary divergence between fungal and plant viruses.

164 **Results**

165 Structural modeling

166 A set of 20 viral species was selected to be representative of Sobelivirales infecting
167 plants and fungi (Table 1). The RdRp structural models for these viruses were generated using
168 the Alphafold 2.2 algorithm. Additionally, three outgroup RdRps were identified, modeled and
169 incorporated into the initial dataset (Table 2). These homologues (MyCV, BeSV, XiNV),
170 present in the Uniprot database, were identified through metagenomics analyses from
171 arthropods or soil samples and found through AlphaFold. In total, more than 2,500 related
172 proteins were aligned for each of the 23 target sequences. Predicted 3D models showed local
173 distance difference test (plDDT) scores ranging from 92% to 96%, indicating highly reliable
174 structural models (Fig 1, see S2 Figure & S3 File). We also conducted a preliminary analysis
175 of *Heterocapsa circularisquama RNA virus* (HcRNAV). This virus, which infects
176 dinoflagellates, is currently the only described species of the genus *Dinornavirus* in the family
177 *Alvernaviridae* in the order *Sobelivirales*. While the structural model of HcRNAV RdRp can
178 be easily superimposed on that of RYMV RdRp, the modeling scores were locally low along
179 the protein, with 10 % of the residues having low-confidence plDDT scores (< 60) (see S4
180 Figure). As a result, HcRNAV was excluded from dowstream analyses.

181 The predicted 3D model of RYMV RdRp was initially compared to available 3D
182 structures of poliovirus (PV) and foot-and-mouth disease virus (FMDV), two well-studied
183 species of picornavirales in the phylum *Pisuviricota* [31]. The characteristic fingers-palm-
184 thumb organization was observed in all three structures (Fig 2, see S5 Figure & S6 File).
185 Although conservation between RYMV and PV/FMDV sequences was low (~10% identity)
186 except in the RdRp motifs, the nature and 3D orientation of most secondary structural elements
187 (α-helices and β-sheets) were preserved across all three structures. The central part of the
188 protein, which forms the conserved palm domain, showed high local structural similarity, with
189 root mean square deviations (RMSD) below 1 angstrom (Å). However, in this domain, there
190 was a larger number of β-strand elements which are longer in RYMV compared to the two
191 picornaviruses. In contrast, the thumb domain, formed by α-helices at the C-terminal part of
192 RdRps, was smaller in the predicted RYMV structure. At the N-terminal region, a conserved
193 feature of picornavirales RdRps was observed: the index finger interacted with the top of the
194 thumb, trapping the ring finger beneath it. Additionally, the orientation of the « pinky finger
195 helix », which has been previously described to distinguish two picornavirales groups including
196 FMDV or PV [31], would bring the predicted structure of RYMV closer to that of FMDV.

197 Structural alignment

198 Each RdRp 3D model of the 23 sobelivirales was structurally aligned using the mTM-
199 align algorithm [32]. We firstly analyzed the superimpositions of the models inside the same
200 genus (Fig 3). Given the sequence similarity between the RdRps of polemovirus and
201 poleroviruses (~75%), they were grouped together as « poler/moviruses » for the remainder of
202 the analyses. Intra-genus model superpositions revealed a high degree of structural similarity
203 in the alpha-carbon traces, with maximum RMSD below 1.3 Å. The least divergent structures
204 were found among poler/moviruses (0.68 Å) while hubscleroviruses and enamoviruses showed
205 intermediate values (1.10 Å and 1.12 Å, respectively). Higher RMSD values were observed for

206 sobemoviruses and barnaviruses (1.26 Å and 1.29 Å), primarly due to two main structural
207 variations: the distinct loop lengths at the thumb joint in sobemoviruses and at the edge of the
208 palm in barnaviruses, referred to here as the « orange loop » and the « light blue loop ».
209 However, we noted that unlike the « orange loop », the « light blue loop » exhibited lower
210 plDDT values (<70), indicating that the 3D conformation of this loop was uncertain. The three
211 outgroups exhibited greater structural divergence, with a maximum RMSD of 1.84 Å, which
212 was consistently localized to the external domains of the RdRp.

213 The 3D overlay of the 23 RdRps confirmed the global conservation of the structure of
214 sobeliviral RdRps (1.9 Å < RMSD < 2.5 Å) (Fig 4). While the RdRp core was highly conserved,
215 the main structural differences were observed on the protein surface, particularly in external
216 loops. In addition to the variable loops previously identified at the intra-genus level, the global
217 model superimposition revealed secondary structure elements specific to certain genera. For
218 instance, a longer loop was found in the wrist domain of barnavirus RdRps, which we refer to
219 as the « yellow loop ». Again, low plDDT values were observed in this barnaviral loop. In
220 poler/movirus and enamovirus RdRps, a supplementary helix and a longer loop, named here
221 the « green helix » and « red loop », were identified at the thumb junction and the thumb tip,
222 respectively (Fig 4). These specific secondary elements showed high plDDT scores (>90-80).
223 Additionally, the « light blue loop », found to be variable among barnaviruses but with
224 relatively low plDDT values, was also longer in hubscleroviruses compared to other sobeliviral
225 genera but with high plDDT scores (>90). Notably, the amino-terminal extremity of all
226 sobeliviral RdRps studied forms a β-strand, interacting with other strands of the middle and
227 ring fingers (Fig 4). At this inter-genera taxonomic level, it is the first time that an extra domain
228 has been explicitly included in the sequence alignment.

229 The amino acid alignment of the 23 RdRps from plant and fungal viruses has a mean
230 length of 459 residues across 649 sites (Fig 5). Amino acid conservation was notably low, with

231 only 31 out of 459 residues being strictly conserved (less than 7% of the mean length). Half of
232 these conserved residues are located within well-characterized RdRp motifs, including a third
233 within the ABC motifs. As expected, the solvent accessibility scores and predicted secondary
234 structure elements were well conserved along the linear sequences (see S7 File). The 3D
235 modeling facilitated alignment and optimization of gap positions and lengths. We observed
236 sequence clustering based on their secondary structure elements, which reflected both intra-
237 and inter-genus relationships. For example, positions 56-59 formed an α-helix in both
238 poler/movirus and enamovirus RdRps, while positions 118-130 folded as coil + 3_{10} helix in
239 both hubsclerovirus and sobemovirus polymerases.

240 The insertion/deletion (indels) of gaps to account for longer secondary structure
241 elements is heterogeneously distributed depending on the domains and viral species. These
242 indels corresponded to domains identified through the global superimposition of the 3D models
243 (Fig 4 & 5). Some indels were specific to individual genera, such as those for the « yellow
244 loop » at positions 492-507 in barnaviruses and the « orange loop » at positions 557-565 in
245 sobemoviruses. Other indels were shared between two genera, such as those for the « green
246 helix » and « red loop » at positions 370-380 and 604-609 in both poler/mo- and enamoviruses,
247 and those for the « light blue loop » at positions 138-186 found in barnaviruses and
248 huscleroviruses.

249 Phylogeny and divergence date estimation

250 The structural alignment was translated into a multiple sequence alignment of nucleic
251 acids by replacing each amino acid by its corresponding codon (see S8 File), which was used
252 to construct the maximum-likelihood phylogenetic tree of Sobelivirales using the GTR+G+I
253 substitution model in PhyML (Fig 6). The outgroups formed a well-supported cluster and
254 contributed to reveal two main clades (with high bootstrap supports), referred to as Polena- and

255 Sobarna- throughout the text. Within the Polena-clade, the phylogenetic relationships between
256 poler/moviruses and enamoviruses were confirmed. In the Sobarna-clade, phylogenetic links
257 between plant sobemoviruses and fungal barna-/hubsclerovirus were highlighted, with a
258 bootsptrap value of 90, establishing this monophyletic group for the first time. However, the
259 phylogenetic relationships between hubscleroviruses and sobemoviruses lacked strong support.

260 The time-scaled phylogenetic history for the 20 Sobelivirales was constructed using the
261 PoW model applied to the structural alignment. According to this analysis, the most recent
262 common ancestor (TMRCA) of sobemoviruses dates back to 2.3 million years before present
263 (BP) (95% highest posterior density (HPD): 1.3-3.5 million years BP) (Fig 7), aligning closely
264 with a recent estimate [25]. The divergence time of the Polenamo-clade was estimated to 5.3
265 (95% HPD: 3.4–7.9) million years BP. The findings here provides the first estimate of the
266 TMRCA for fungal (barna)viruses, which dates back to 3 (95% HPD: 1.8-5.0) million years
267 BP, and for the Sobarna-clade to 17.7 (95% HPD: 12.2– 24.8) million years BP. Furthermore,
268 the analysis here offers the first estimation of the TMRCA of the Sobelivirales order, reaching
269 back to 26.6 (95% HPD: 19.2-34.8) million years BP, which is nearly ten times older than the
270 origins of sobemoviruses.

271 Polymerase evolution

272 The RdRp core and active sites within the central tunnel of Sobelivirales were found to
273 be highly conserved across all genera, despite significant divergence in their protein sequences
274 (see S9 Figure). In contrast, genus-specific residues were evenly distributed around the core
275 without being spatially clustered (data not shown). In order to take phylogenic data into account
276 in the study of the evolution of structure/function relationships of sobeliviral polymerases, we
277 analyzed sequence conservation by comparing the Polena- versus Sobarna- clades after
278 exclusion of the three outgroups (see S10 File & S11 Table). We identified 35 clade-specific

residues - i.e. amino acids strictly present (or physico-chemically similar) in all sequences of one clade and strictly absent from sequences of the other clade. These residues were located on 31 polymorphic sites in the RdRps. Of these, 25 residues were strictly conserved : 4/11 in the sobarnaviral sequences and 21/24 in the polenaviral sequences (see S11 Table). Six clade-specific residues were found within the conserved RdRp motifs – three in the palm domain and three in the fingers domains.

To validate these clade-specific residues, 99 additional sobeliviral sequences were incorporated into our aligned dataset, resulting in a total dataset of 119 sequences, which included all ICTV-recognized viral species in the studied genera (see S12 Table & S13 File). Almost 70% (24/35) of the clade-specific polymorphisms were confirmed, including 3/4 and 10/21 strictly conserved residues for the Sobarna- and Polena-clades, respectively (Table 3). Two new Sobarna-specific residues were detected in conserved motifs of the extended dataset : a serine and a histidine within Motifs G and D, respectively, at positions 188 and 486 (referring to the alignment with 20 sequences). In the Sobarna-clade, residues at positions 486 and 489 were particularly notable for their 3D positions within Motif D - residue 489K/R pointed into the central tunnel, while 486H/K was likely stabilized by residue 365E (Fig 8). In the Polena-clade, a strictly conserved lysine at position 263 within Motif F also pointed into the central tunnel, alongside the conserved sobeliviral lysine at position 266 (Fig 8). This analysis revealed molecular signatures of these deeply divergent clades, refining the definitions of the conserved motifs (Fig 9). In addition to clade-specific polymorphic positions such as 354I/C, 468S/A and 188S/G, certain residues were strictly conserved across sobelivirales, including 355-356SG, 261-262PH, 512-514FCS in Motifs A, F and E, respectively. The specificity of these conserved residues to the sobeliviral order have to be tested by analyzing a larger dataset of RdRp sequences from species of other viral orders. Moreover, genera-specific features idenfied on the structural alignment of 20 sequences were confirmed on the expanded corpus. Notably, the

304 polena- « green helix » upstream Motif A and the polena- « red loop » dowstream Motif H
305 provided robust indel signatures. The sobemoviral « orange loop » between Motifs D and E
306 was also validated, appearing less variable than previously observed in the smaller corpus of
307 20 sequences.
308
309

310 **Discussion**

311 Protein structure modeling has proved effective in addressing some of the challenges to
312 making reliable sequence alignments posed by low sequence conservation at high taxonomic
313 levels. In this study, we used AlphaFold models to produce a robust and extensive sequence
314 alignement of sobeliviral RdRps, for the first time at this taxonomic level. Although initially
315 based on a limited set of sequences, the alignment was extensively validated using a full set of
316 known viral species in this order. Minor discrepancies, however, were observed in gap positions
317 at carboxy-terminal extremity of RdRp. In addition, the limited number of known
318 hubscleroviruses and barnaviruses prevented the validation of some features such as the « light
319 blue loop » and « yellow loop », respectively. Protein structure modeling successfully provided
320 high confidence supports that complemented classical functional motifs, serving as anchor
321 points. Notably, the RdRp extremities which are typically trimmed upstream of motif G and
322 downstream of motif E - deleting approximatively 150 residues - were very well predicted by
323 Alphafold, with plDDT exceeding 90%, and showing high structural similarity (RMSD < 2 Å).
324 Here, we considered the amino-terminal extremity of RdRp and its intra-molecular interactions,
325 improving the alignement length and quality.

326 In addition to generating robust structure models, Alphafold helps to identify new
327 homologues in databases. In this study, we selected three outgroups from metagenomic data
328 based on their genetic distance from the 20 initial sequences. Generally closer than
329 alvernaviruses, these outgroups served as a reliable reference for rooting the phylogenetic tree
330 and enhancing the quality of the phylogenetic analysis. We evaluated the accuracy of our
331 structural alignement to traditional alignment algorithms. The residue pairing between our
332 structural alignment and MAFFT, Clustal Omega, and MUSCLE showed differences of 21%,
333 23%, and 30%, respectively. Most of these differences were observed in the loop/coil regions

334 as well as in the amino- and carboxy-terminal extremities of the protein (see S14 File).
335 However, these differences had minimal impact on tree topology as they mainly concerned the
336 outgroup alignment (see S15 Figure). MUSCLE produced results similar to mTM-align, while
337 Clustal Omega and MAFFT failed to detect phylogenetic relationships between
338 hubscleroviruses and sobemovirus. Furthermore, Clustal Omega yielded low internal node
339 support at the sobemovirus / barnavirus split. The lower accuracy of Clustal Omega compared
340 to MUSCLE and MAFFT has been already reported for structural alignments of Flaviviridae
341 glycoproteins [29].

342 The phylogenic analysis led to a new taxonomic proposal : the creation of two new
343 tentative families Polenaviridae and Sobarnaviridae, replacing Solemoviridae and
344 Barnaviridae. By assessing the conservation of the structural alignment, we enriched the
345 common conserved motifs, identified molecular signatures of the new tentative families, and
346 revealed structural diversity (e.g., loop lengths) that can be used as marker of viral genera.
347 Interestingly, newly identified putative sobemo-like and sobeli-like viruses often represent a
348 significant portion of newly discovered plant and fungus taxa within environnemental viromes
349 [16,33–36]. Extending this approach in the near future will accelerate their taxonomic
350 classification, a priority for virus research [37]. Additionally, these RdRp features could be
351 useful for the development of molecular tools for diagnostic and detection.

352 The efficiency of our integrative approach was demonstrated at the taxonomic rank of
353 order/family and could be implemented to other clades, such as fungal viruses in the phylum
354 *Pisuviricota*, such as *Polymycoviridae* and *Yadokarivirales*, currently unassigned at the family
355 and order levels, respectively. However, this approach requires some level of structural
356 similarity below which comparisons are not recommended. To assess this limit, we predicted
357 3D models of the RdRp of *Luteovirus pavhordei* (previously named BYDV-pav) and

358 *Dianthovirus trifolii* (previously named RCNMV), since the former was previously classified
359 in the family *Luteoviridae* with the poleroviruses and is now classified with the latter in
360 *Tombusviridae* (Order *Totivirales*, Class *Tolucaviricetes*, phylum *Kitrinoviricota*) [7]. These
361 3D models were aligned with one representative of each six genera using the same method as
362 described above (see S16 File). Although the core of all three RdRps showed good sequence
363 and structure similarity, the overlay score on the full-length protein was low (3.3 Å > RMSD >
364 3.5Å). As expected, most of the Sobarna-and Polena-specific residues in the motifs, were not
365 shared with the two totivirales analyzed. The unique exception was observed in the motif G for
366 which *Luteovirus pavhordei* showed the same pattern as sobarnavirid sequences (188-189SP).
367 Noteworthy, half of the sobeliviral-conserved residues in Motifs A, F and E were also present
368 in the two totivirales analyzed (355S, 262H, 512-513FC, respectively) suggesting that these
369 residues were not strictly specific. These preliminary analyses should be continued and
370 expanded in the future. By contrast to the core, the external loops of the two totiviral RdRps
371 analyzed here showed very low structural similarity with the sobeliviral ones. In addition,
372 neither the « green helix » and « red loop » nor the « orange loop », the secondary structure
373 elements specific to polenavirids and sobarnavirids, respectively, were observed. Consequently,
374 we considered that entire RdRps of luteoviruses were significantly too distant to RdRps of
375 poleroviruses and sobelivirales to be aligned accurately with our structural approach.

376 Furthermore, structural similarity does not always imply phylogenic relationships. As
377 mentionned earlier, for the picorna-like supergroup, the hypothesis of convergent evolution
378 cannot be excluded [5]. Picornaviruses show both structural and functional similarities with
379 sobelivirales, such as the use of a Viral Protein genome-linked (VPg) to prime RNA replication
380 by the RdRp. However, picornaviral VPg are very small (~20 residues) and are cleaved from
381 the RdRp by a viral protease located downtream of the VPg domain. In contrast, the
382 sobarnavirids and polenavirids share a conserved genomic organization

383 (protease/VPg/polymerase) and have larger VPgs (80-150 residues), which likely result in or
384 induce specific molecular mechanisms. While interactions between VPg and RdRps and their
385 structure-function relationships have been extensively studied in human and animal viruses,
386 very little data exist for plant viruses. Our approach could help to fill this knowledge gap, or at
387 least provide a better understanding of plant RdRp features through *in silico* comparative
388 analyses.

389 For example, the comparative analysis of Motif E, unique to RdRps compared to other
390 polymerases, offers a good starting point for better understanding of some of the plant RdRp
391 features. The amino acid composition of this motif is directly linked to the viral replication
392 strategy, using protein priming (VPg) or not. Described as « primer grip », this motif has been
393 shown to help correctly position the 3' OH end of the primer strand [9,17]. In picornaviral
394 RdRps, the conserved phenylalanine in Motif E, which faces Motif C, is followed by two
395 residues with relatively long side chains (phenylalanine-leucine-lysine, FLK), while in
396 flaviviral RdRps, the central channel is narrower, and these residues have rather short side
397 chains (phenylalanine-cysteine-serine, FCS). Interestingly, plant and fungus sobeliviral RdRps
398 share the same conserved Motif E as flavivirids. Additionaly, the homomorph that contains
399 Motif E forms part of the NTP entry tunnel and can vary in lenght depending on the viral
400 species. Sobemoviruses exhibit a longer « orange loop » at the thumb joint compared to other
401 members of sobelivirales, another unexpected similarity with flaviviral hepatitis virus C.

402 At a different taxonomic scale, Motif D illustrates the common evolutionary history of
403 plant and fungus sobeliviral RdRps. This motif influences the rate of nucleotide addition to the
404 nascent RNA strand. Characterized by a glycine hinge and a positively charged lysine, both
405 located in similar positions across available RdRp crystal structures, the distance inside the
406 linear sequence between these two conserved residues varies across different viral families [9].

407 In the Sobarnaviridae clade, lysine (or its equivalent arginine) is conserved at position N+4, i.e.
408 position 489 (in the reference alignement of 20 sequences) (Fig 9). In the Polenaviridae clade,
409 our initial analysis was too stringent to detect the conserved lysine, as one species out of 89 had
410 a non-conserved hydrophobic isoleucine at this position. However, when *Enamovirus CLEV*
411 was excluded from the dataset, the conserved lysine was found at position N+2 – i.e. position
412 487, with its orientation differing from the sobernaviridae lysine in the 3D structural model.
413 Motif D has been shown to influence not only incorporation speed but also fidelity, both of
414 which are closely linked to viral pathogenesis and virulence. For example, increased fidelity in
415 a poliovirus mutant has been associated with viral attenuation [38].

416 The structural differences and evolutionary distances between RdRps of sobemoviruses
417 and poleroviruses (two phytovirus genera in the order *Sobelivirales*) were found to be greater
418 than those between sobemoviruses and barnaviruses (a mycovirus genus in the family
419 *Barnaviridae*) as well as hubscleroviruses (a mycovirus genus in the family *Solemoviridae*).
420 Given the role of RdRp as a major marker of viral evolutionary history, our results confirmed
421 that horizontal viral transfert(s) (HVT) between plant/fungal viruses occured in the Order
422 *Sobelivirales*. The direction of the HVT(s) cannot yet be determined without further data on the
423 hosts of metagenomic sequences we used as outgroups. Using an interdisciplinary approach,
424 we provided the first divergence time estimation between these plant and fungal viruses, dating
425 back to ~26 million years BP. The TDRP-corrected evolutionary history of sobelivirales was
426 extended five thousand times far beyond previous estimates [26]. However, we note that this
427 estimate may reflect a conservative timeframe based on current methodologies, as our
428 understanding of TDRP and application of the PoW model has primarily been at the genus and
429 family levels. With future advances in modeling and expanded data, these divergence times
430 could potentially be revised further back. Yet, even with our current conservative estimates, the
431 divergence time remains more recent than the divergence between the respective plant and

fungal hosts, suggesting that a strict virus-host co-evolution hypothesis not fully explain these evolutionary relationships. Other lineages of mycoviruses either with single- or double-stranded RNA genomes showed close phylogenetic relationships with plant viruses at different taxonomical ranks [18,20,39], suggesting that taxonomy of these viruses should rather be viewed as a continuum. An interdisciplinary approach at the interface of biophysics, bioinformatics, and virology should be encouraged to better understand evolution, interactions and biological mechanisms of myco- and phytoviruses.

440 **Material and methods**

441 Sequence selection

442 The sobeliviral dataset initially consisted of 20 species from six genera (Table 1). Six
443 sobemoviruses including four recently described (ArtVA, CyCMV, PLYV, XYDV), four
444 poleroviruses, four enamoviruses, one polemovirus, two hubscleroviruses including one
445 candidate species, and three barnaviruses including two candidate species. Complementary
446 analyses were performed on *Heterocapsa circularisquama RNA virus* (HcRNAV, accession
447 NC007518, *Dinornavirus, Alvernaviridae, Sobelivirales*), on *Luteovirus pavhordei* (BYDV-
448 pav, accession OM046620, *Luteovirus, Tombusviridae, Tolivirales*) and on *Dianthovirus trifolii*
449 (RCNMV, accession OL311680, *Dianthovirus, Tombusviridae, Tolivirales*).

450 The validation dataset consisted of 99 sequences of sobelivirales species made of all the
451 ICTV viral species available in the studied genera. In addition to the two hubscleroviruses, the
452 three barnaviruses and the polemovirus already cited, the complete corpus consisted of 25
453 sobemoviruses, 73 poleroviruses and 15 enamoviruses (see S12 Table).

454 3D model prediction

455 The structures of the corresponding RdRps were predicted using a locally installed
456 version of Alphafold 2.2 [40]. For each sequence, 5 models were generated and relaxed using
457 default Alphafold parameters. All homologues detected by Jackhmmer [41] in the Uniprot
458 database (https://www.uniprot.org/) during these 20 Alphafold runs were then clustered using
459 kClust into groups of proteins sharing at least both 80% sequence identity and 50% sequence
460 overlap [42]. The representative sequences of each group were then aligned using MAFFT [43]
461 and a phylogenetic tree was derived from the alignment using FastME [44]. From this tree,
462 three new proteins were selected forming an outlier group (MyCV, BeSV, XiNV) (Table 2).

463 The 3D structure of these new proteins, as well as RdRps of HcRNAV, BYDV-pav and
464 RCNMV, were also predicted using Alphafold.

465 Evaluation of 3D models

466 One representative of each 6 viral genera was selected - RYMV (*Sobemovirus*), PLRV
467 (*Polerovirus*), PeMV (*Enamovirus*), MBV (*Barnavirus*), HuSRV (*Hubsclerovirus*), XyNV
468 (outgroups). The best RdRp 3D models predicted by Alphafold for these viral species were
469 structurally aligned using mTM-align [32]. The pLDDT reliability scores available in the B-
470 factor field of each Alphafold model were coded as integers (0=0-10, 1=10 -20, …, 9=90-100)
471 which were added to the mTM-align structural alignment with sequence identifiers suffixed by
472 .sc.PLDDT. This alignment is displayed in a web page where the aligned sequences and their
473 corresponding pLDDT scores are interfaced with 6 representative 3D models using Jsmol applet
474 and script [45]. Selected Alphafold models can displayed and customized by using the toggle
475 buttons and the popup menus from the right bottom panel and the colors of the amino acids or
476 pLDDT scores of one aligned sequence from the top panel can be projected on the selected 3D
477 structure(s) by clicking on the corresponding protein identifier in the top left panel (see S3 File).

478 The RYMV 3D model with highest Alphafold pLDDT scores was superposed to 1RA6
479 and 1U09 PDB structures of poliovirus (PV) and foot-and-mouth disease virus (FMDV) RdRps,
480 respectively, using mTM-align. The sequence identity between the three proteins was
481 calculated after alignment with Ident and Sim software [46]. The observed Cα deviations, in
482 Angstroms, where added below the obtained structural alignment (identifier suffixed with
483 d3.RMSD_1ra6A an .d3.RMSD_1u09A). The Cα deviation colors can be projected onto one
484 of the displayed structures by cliking one of these identifiers (see S6 File).

485

486 Structural alignment construction

487 All the best 3D models predicted by Alphafold for the 23 virus proteins were structurally
488 aligned using mTM-align. The 23 analyzed proteins were clustered into 6 subgroups
489 (Sobemovirus, Barnavirus, Hubsclerovirus, Poler/movirus, Enamovirus and a subgroup of
490 outlier viruses). The best RdRp 3D model of the 23 sobelivirales were superimposed with the
491 other models inside its group using mTM-align. The RMSD between the aligned carbon alpha
492 atoms was calculated for each viral pair inside each subgroup, then the maximum and average
493 RMSD between subgroup pairs were computed.

494 We carefully delineated the polymerase domain shared by the 23 proteins by inspecting
495 both the positional conservation of the aligned sequences, the superposition of their predicted
496 secondary structures and the solvent accessibility (see S7 File). The secondary structure and the
497 solvent accessibility of each model residue was calculated using STRIDE [47] and the
498 positional conservation was calculated as the frequency of the most represented amino acid in
499 a given alignment column.

500 Phylogenetic analysis and time tree reconstruction with PoW model

501 The mTM-align structural alignment was translated into a multiple sequence alignment
502 of nucleic acids by replacing each amino acid by its corresponding codon (Table 1) (see S8
503 File). JModelTest 2 [48] was used to identify the best nucleotide substitution model for
504 phylogenetic analysis based on the Akaike Information Criterion. The Maximum-Likelihood
505 tree was then constructed with the General Time Reversible Model with an optimised rate
506 variation and class for invariant sites (GTR + G + I) using PhyML implemented in SeaView (v
507 5.05) [49]. The absence of recombination events between species was confirmed with the Phi
508 test [50]. The reliability of the phylogenetic tree nodes was estimated via a bootstrap resampling
509 procedure with 100 replicates. The phylogenetic tree was rooted using the outgroup branch.

510 We have also detected the homologous amino acid pairs which were identically aligned
511 between the mTM-align structural alignment and each multiple sequence alignment built using
512 respectively MAFFT [43], MUSCLE [51] and Clustal Omega [52]. The frequencies of shared
513 amino acid pairs were calculated both over the whole alignments and within each alignment
514 column (see S14 File).

515 We used the Prisoner of War (PoW) model of virus evolution to account for the power-
516 law decay in the substitution rate estimates of sobemoviruses, as described in Ghafari et al. [25],
517 where we also established the presence of a robust and representative molecular clock signal
518 for the RdRp locus for sobemoviruses. Briefly, the short-term substitution rate of RYMV's
519 polymerase gene, inferred from a standard molecular clock model was used as the baseline rate,
520 then ultrametric distance trees were constructed using the substitution rate at the fastest-
521 evolving rate group for our dataset with a strict molecular clock in BEAST (v 1.10.4) [53]. The
522 maximum clade credibility tree was constructed after 100 iterations using TreeAnnotator
523 v.1.10.4 and visualized through a lineages-through-time plot using the phytools library in R
524 [54].

525 Polymerase evolution

526 The analyzed proteins were clustered into 2 subgroups comparing the Polena- versus
527 Sobarna- clades. The three outgroups were not considered in this analysis (see S10 File). At
528 each site of this new mTM-align structural alignment, amino acids strictly present in all
529 sequences of one clade (or that were similar based on BLOSUM score ≥ 0) and strictly absent
530 from sequences of the other clade were considered as clade-specific residue. These conserved
531 residues specific to Sobarna- and Polena-clades were added to the mTm-align alignment as two
532 final consensus sequences suffixed by c1.PID and .c2.PID respectively. The location of these
533 signature residues can be highlighted on the displayed 3D model clicking one of the c1.PID and

.c2.PID identifier and by hovering the mouse pointer onto one of the consensus sequence. In the supplementary table, only strictly conserved or similar in all sequences of one clade were considered. Motif analysis was made with WebLogo v2.8.2 [55].

Acknowledgements

This work benefited from discussions with members of Virus/Cereal Interactions in Tropical Agroecosystems group and colleagues of Plant Health Institute Montpellier. This work was supported by the French Infrastructure for Integrated Structural Biology (FRISBI) ANR-10-INBS-0005.

Data and code availability

The code for the PoW model is available on GitHub https://github.com/mg878/PoW_model

All the data (including accession numbers) are provided within the main text and supporting information.

Competing interest

None.

Funding

None.

Author contribution

Writing (original draft) JG and EH; Writing -- review and editing all authors; Formal Analysis JG MG EH: Conceptualization all authors; Data Collection JG DF EH; Supervision EH. All authors reviewed and edited the manuscript.

Correspondence to Eugénie Hébrard (eugenie.hebrard@ird.fr).

558 **References**

559 1. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, et al. Virus taxonomy
560 in the age of metagenomics. Nat Rev Microbiol. 2017;15: 161–168.
561 doi:10.1038/nrmicro.2016.177

562 2. Gorbalenya AE. Increasing the number of available ranks in virus taxonomy from five to ten and
563 adopting the Baltimore classes as taxa at the basal rank. Arch Virol. 2018;163: 2933–2936.
564 doi:10.1007/s00705-018-3915-6

565 3. Siddell SG, Walker PJ, Lefkowitz EJ, Mushegian AR, Adams MJ, Dutilh BE, et al. Additional
566 changes to taxonomy ratified in a special vote by the International Committee on Taxonomy of
567 Viruses (October 2018). Arch Virol. 2019;164: 943–946. doi:10.1007/s00705-018-04136-2

568 4. Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, et al. Global Organization
569 and Proposed Megataxonomy of the Virus World. Microbiol Mol Biol Rev. 2020;84: e00061-19.
570 doi:10.1128/MMBR.00061-19

571 5. Wolf YI, Kazlauskas D, Iranzo J, Lucía-Sanz A, Kuhn JH, Krupovic M, et al. Origins and
572 Evolution of the Global RNA Virome. Racaniello VR, editor. mBio. 2018;9: e02329-18.
573 doi:10.1128/mBio.02329-18

574 6. Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Adriaenssens EM, Alfenas-Zerbini P, et
575 al. Recent changes to virus taxonomy ratified by the International Committee on Taxonomy of
576 Viruses (2022). Arch Virol. 2022;167: 2429–2440. doi:10.1007/s00705-022-05516-5

577 7. Walker PJ, Siddell SG, Lefkowitz EJ, Mushegian AR, Adriaenssens EM, Alfenas-Zerbini P, et
578 al. Changes to virus taxonomy and to the International Code of Virus Classification and
579 Nomenclature ratified by the International Committee on Taxonomy of Viruses (2021). Arch
580 Virol. 2021;166: 2633–2648. doi:10.1007/s00705-021-05156-1

581 8. Ferrero D, Ferrer-Orta C, Verdaguer N. Viral RNA-Dependent RNA Polymerases: A Structural
582 Overview. In: Harris JR, Bhella D, editors. Virus Protein and Nucleoprotein Complexes.
583 Singapore: Springer Singapore; 2018. pp. 39–71. doi:10.1007/978-981-10-8456-0_3

584 9. Jácome R, Becerra A, Ponce de León S, Lazcano A. Structural Analysis of Monomeric RNA-
585 Dependent Polymerases: Evolutionary and Therapeutic Implications. Kuhn JH, editor. PLOS
586 ONE. 2015;10: e0139001. doi:10.1371/journal.pone.0139001

587 10. Koonin EV, Wolf YI, Nagasaki K, Dolja VV. The Big Bang of picorna-like virus evolution
588 antedates the radiation of eukaryotic supergroups. Nat Rev Microbiol. 2008;6: 925–939.
589 doi:10.1038/nrmicro2030

590 11. te Velthuis AJW. Common and unique features of viral RNA-dependent polymerases. Cell Mol
591 Life Sci. 2014;71: 4403–4420. doi:10.1007/s00018-014-1695-z

592 12. Koonin EV. The phylogeny of RNA-dependent RNA polymerases of positive-strand RNA
593 viruses. J Gen Virol. 1991;72: 2197–2206. doi:10.1099/0022-1317-72-9-2197

594 13. Poch O, Sauvaget I, Delarue M, Tordo N. Identification of four conserved motifs among the
595 RNA-dependent polymerase encoding elements. EMBO J. 1989;8: 3867–3874.
596 doi:10.1002/j.1460-2075.1989.tb08565.x

14. Charon J, Buchmann JP, Sadiq S, Holmes EC. RdRp-scan: A bioinformatic resource to identify and annotate divergent RNA viruses in metagenomic sequence data. Virus Evol. 2022;8: veac082. doi:10.1093/ve/veac082

15. Edgar RC, Taylor B, Lin V, Altman T, Barbera P, Meleshko D, et al. Petabase-scale sequence alignment catalyses viral discovery. Nature. 2022;602: 142–147. doi:10.1038/s41586-021-04332-2

16. Hou X, He Y, Fang P, Mei S-Q, Xu Z, Wu W-C, et al. Using artificial intelligence to document the hidden RNA virosphere. Cell. 2024; S0092867424010857. doi:10.1016/j.cell.2024.09.027

17. Lang DM, Zemla AT, Zhou CLE. Highly similar structural frames link the template tunnel and NTP entry tunnel to the exterior surface in RNA-dependent RNA polymerases. Nucleic Acids Res. 2013;41: 1464–1482. doi:10.1093/nar/gks1251

18. Dolja VV, Krupovic M, Koonin EV. Deep Roots and Splendid Boughs of the Global Plant Virome. Annu Rev Phytopathol. 2020;58: 23–53. doi:10.1146/annurev-phyto-030320-041346

19. Andika IB, Tian M, Bian R, Cao X, Luo M, Kondo H, et al. Cross-Kingdom Interactions Between Plant and Fungal Viruses. Annu Rev Virol. 2023;10: 119–138. doi:10.1146/annurev-virology-111821-122539

20. Roossinck MJ. Evolutionary and ecological links between plant and fungal viruses. New Phytol. 2019;221: 86–92. doi:10.1111/nph.15364

21. Gorbalenya AE, Donchenko AP, Blinov VM, Koonin EV. Cysteine proteases of positive strand RNA viruses and chymotrypsin-like serine proteases: A distinct protein superfamily with a common structural fold. FEBS Lett. 1989;243: 103–114. doi:10.1016/0014-5793(89)80109-7

22. Revill PA, Davidson AD, Wright PJ. Mushroom Bacilliform Virus RNA: The Initiation of Translation at the 5′ End of the Genome and Identification of the VPg. Virology. 1998;249: 231–237. doi:10.1006/viro.1998.9345

23. Sõmera M, Fargette D, Hébrard E, Sarmiento C, ICTV Report Consortium. ICTV Virus Taxonomy Profile: Solemoviridae 2021: This article is part of the ICTV Virus Taxonomy Profiles collection. J Gen Virol. 2021;102. doi:10.1099/jgv.0.001707

24. Simmonds P, Adriaenssens EM, Lefkowitz EJ, Oksanen HM, Siddell SG, Zerbini FM, et al. Changes to virus taxonomy and the ICTV Statutes ratified by the International Committee on Taxonomy of Viruses (2024). Arch Virol. 2024;169: 236. doi:10.1007/s00705-024-06143-y

25. Ghafari M, Sõmera M, Sarmiento C, Niehl A, Hébrard E, Tsoleridis T, et al. Revisiting the origins of the Sobemovirus genus: A case for ancient origins of plant viruses. Nagy PD, editor. PLOS Pathog. 2024;20: e1011911. doi:10.1371/journal.ppat.1011911

26. Fargette D, Pinel A, Rakotomalala M, Sangu E, Traoré O, Sérémé D, et al. *Rice Yellow Mottle Virus* , an RNA Plant Virus, Evolves as Rapidly as Most RNA Animal Viruses. J Virol. 2008;82: 3584–3589. doi:10.1128/JVI.02506-07

27. Carpentier M, Chomilier J. Protein multiple alignments: sequence-based versus structure-based programs. Valencia A, editor. Bioinformatics. 2019;35: 3970–3980. doi:10.1093/bioinformatics/btz236

28. Gutnik D, Evseev P, Miroshnikov K, Shneider M. Using AlphaFold Predictions in Viral Research. Curr Issues Mol Biol. 2023;45: 3705–3732. doi:10.3390/cimb45040240

638 29. Mifsud JCO, Lytras S, Oliver MR, Toon K, Costa VA, Holmes EC, et al. Mapping glycoprotein
639 structure reveals Flaviviridae evolutionary history. Nature. 2024;633: 695–703.
640 doi:10.1038/s41586-024-07899-8

641 30. Nino Barreat JG, Katzourakis A. Deep mining reveals the diversity of endogenous viral elements
642 in vertebrate genomes. Nat Microbiol. 2024;9: 3013–3024. doi:10.1038/s41564-024-01825-4

643 31. Peersen OB. Picornaviral polymerase structure, function, and fidelity modulation. Virus Res.
644 2017;234: 4–20. doi:10.1016/j.virusres.2017.01.026

645 32. Dong R, Peng Z, Zhang Y, Yang J. mTM-align: an algorithm for fast and accurate multiple
646 protein structure alignment. Valencia A, editor. Bioinformatics. 2018;34: 1719–1725.
647 doi:10.1093/bioinformatics/btx828

648 33. Chen Y-M, Sadiq S, Tian J-H, Chen X, Lin X-D, Shen J-J, et al. RNA viromes from terrestrial
649 sites across China expand environmental viral diversity. Nat Microbiol. 2022;7: 1312–1323.
650 doi:10.1038/s41564-022-01180-2

651 34. Duarte MF, De Andrade IA, Silva JMF, De Melo FL, Machado AM, Inoue-Nagata AK, et al.
652 Metagenomic analyses of plant virus sequences in sewage water for plant viruses monitoring.
653 Trop Plant Pathol. 2023;48: 408–416. doi:10.1007/s40858-023-00575-8

654 35. Maksimović O, Bačnik K, Rivarez MPS, Vučurović A, Mehle N, Ravnikar M, et al. Virome
655 analysis of irrigation water sources provides extensive insights into the diversity and distribution
656 of plant viruses in agroecosystems. Water Res. 2024;249: 120712.
657 doi:10.1016/j.watres.2023.120712

658 36. Shi M, Lin X-D, Tian J-H, Chen L-J, Chen X, Li C-X, et al. Redefining the invertebrate RNA
659 virosphere. Nature. 2016;540: 539–543. doi:10.1038/nature20167

660 37. Gorbalenya AE, Siddell SG. Recognizing species as a new focus of virus research. Morrison TE,
661 editor. PLOS Pathog. 2021;17: e1009318. doi:10.1371/journal.ppat.1009318

662 38. Castro C, Smidansky ED, Arnold JJ, Maksimchuk KR, Moustafa I, Uchida A, et al. Nucleic acid
663 polymerases use a general acid for nucleotidyl transfer. Nat Struct Mol Biol. 2009;16: 212–218.
664 doi:10.1038/nsmb.1540

665 39. Ghabrial SA, Castón JR, Jiang D, Nibert ML, Suzuki N. 50-plus years of fungal viruses.
666 Virology. 2015;479–480: 356–368. doi:10.1016/j.virol.2015.02.034

667 40. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate
668 protein structure prediction with AlphaFold. Nature. 2021;596: 583–589. doi:10.1038/s41586-
669 021-03819-2

670 41. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM
671 search procedure. BMC Bioinformatics. 2010;11: 431. doi:10.1186/1471-2105-11-431

672 42. Hauser M, Mayer CE, Söding J. kClust: fast and sensitive clustering of large protein sequence
673 databases. BMC Bioinformatics. 2013;14: 248. doi:10.1186/1471-2105-14-248

674 43. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7:
675 Improvements in Performance and Usability. Mol Biol Evol. 2013;30: 772–780.
676 doi:10.1093/molbev/mst010

677 44. Lefort V, Desper R, Gascuel O. FastME 2.0: A Comprehensive, Accurate, and Fast Distance-
678 Based Phylogeny Inference Program: Table 1. Mol Biol Evol. 2015;32: 2798–2800.
679 doi:10.1093/molbev/msv150

680 45. Hanson RM. *Jmol* – a paradigm shift in crystallographic visualization. J Appl Crystallogr.
681 2010;43: 1250–1260. doi:10.1107/S0021889810030256

682 46. Stothard P. The Sequence Manipulation Suite: JavaScript Programs for Analyzing and
683 Formatting Protein and DNA Sequences. BioTechniques. 2000;28: 1102–1104.
684 doi:10.2144/00286ir01

685 47. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins Struct
686 Funct Bioinforma. 1995;23: 566–579. doi:10.1002/prot.340230412

687 48. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and
688 parallel computing. Nat Methods. 2012;9: 772–772. doi:10.1038/nmeth.2109

689 49. Guindon S, Gascuel O. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies
690 by Maximum Likelihood. Rannala B, editor. Syst Biol. 2003;52: 696–704.
691 doi:10.1080/10635150390235520

692 50. Bruen TC, Philippe H, Bryant D. A Simple and Robust Statistical Test for Detecting the
693 Presence of Recombination. Genetics. 2006;172: 2665–2681. doi:10.1534/genetics.105.048975

694 51. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space
695 complexity. BMC Bioinformatics. 2004;5: 113. doi:10.1186/1471-2105-5-113

696 52. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein
697 sequences. Protein Sci. 2018;27: 135–145. doi:10.1002/pro.3290

698 53. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic
699 and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018;4.
700 doi:10.1093/ve/vey016

701 54. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things).
702 Methods Ecol Evol. 2012;3: 217–223. doi:10.1111/j.2041-210X.2011.00169.x

703 55. Crooks GE, Hon G, Chandonia J-M, Brenner SE. WebLogo: A Sequence Logo Generator:
704 Figure 1. Genome Res. 2004;14: 1188–1190. doi:10.1101/gr.849004

705

706

707 **Tables**

708

709 **Table 1**: Virus name and accession numbers of the 20 sobeliviral species and candidate

710 species or genus (*).

711

Family	Genus (N)	Species	Name	NCBI /uniprot nuc accession number	Uniprot accession nb
Solemoviridae	Sobemovirus (21)	Rice yellow mottle virus	RYMV	AJ608207	Q709L0
		Turnip rosette virus	TRoV	NC004553/AY177608	Q80QX5
		Artemisia virus A	ArtVA	NC017914/JN620802	I1WZY5
		Cymbidium chlorotic mosaic virus	CyCMV	LC381945	A0A2Z6E9A6
		Papaya lethal yellowing virus	PLYV	NC018449/JX123318	J7GBL4
		Xufa yellow dwarf virus	XYDV	ON828429	A0A977KCX8
	Polerovirus (26)	Potato leafroll virus	PLRV	AF453394/X74789	Q84837
		Beet chlorosis virus	BChV	NC002766/MH271171	A0A385L367
		Beet mild yellowing virus	BYMV	NC003491/DQ132996	Q45GB9
		Cereal yellow dwarf virus RPS	CYDV	NC002198/AF235168	Q9IZF5
	Enamovirus (5)	Pea enation mosaic virus 1	PeMV	MN497823	A0A6M8PM86
		Alfalfa enamovirus 1	AlEV	/KU297983	A0A166J1Q2
		Citrus vein enation virus	CVEV	/HF679486	S0EQ55
		Grapevine enamovirus 1	GREnV	/KX645875	A0A1B3B596
	Polemovirus (1)	Poinsettia latent virus	PnLV	AJ867490	Q5NDM9
	Hubsclerovirus (1)	Hubei sclerotinia RNA virus 1	HUSRV	/MK889164	A0A649X039
		Erysiphe necator associated sobemo-like virus 4*	EnAV*	/MN630192	A0A8E3YYS7
Barnaviridae	Barnavirus (1)	Mushroom bacilliform virus	MBV	NC001633/U07551	Q9YPD4
		Rhizoctonia solani barnavirus 1*	RsBV*	/KP900904	A0A0M4KGX8
		Colobanthus quitensis associated barnavirus 1*	CoBV*	MG686618	A0A2P1JMX2

712

713 **Table 2**: Accession numbers of the 5 additional proteins found as homologues of RdRp
714 sobeliviral by AlphaFold

Candidate species	Abbreviation	Uniprot access nb	Source	Reference
Mycroft virus	MyCV	A0A5B8XAH7	arthropods	Harvey et al. 2019
Beihai shrimp virus 5	BeSV	A0A1L3KE75	invertebrates	Shi et al. 2016
Xingshan nematode virus 5	XINV	A0A1L3KF94	invertebrates	Shi et al. 2016

715

716 **Table 3** : Clade-specific residues. Strictly conserved residues and shared sites were indicated
717 (bold and underligned, respectively). Position numbers referred at the structural alignment of
718 the 99 sobelivirales sequences (without the 3 outgroups, see S11 File).

	Sobarnaviruses	Polenamoviruses
Palm		
motifs (A, B, C, D)	I354,S468,K489	**C354**, A468
other sites	**P99,E365**,H486	**A286,F323,D368,N374,T377,R390,W393,**L478
Fingers		
motifs (F,G,H)	S188,**P189**	**G188,K263**
other sites	V205	W25,G26,W205,N529,S561

719

720 **Figure captions**

721 **Figure 1:** RdRp 3D models colored by per-residue model confidence score (local distance
722 difference test, plDDT). The regions modeled with high confidence (plDDT>90) are shown
723 dark blue. The regions colored light blue (plDDT>80) and green (plDDT>70) were modeled
724 with confidence whereas yellow (plDDT>60) and orange (plDDT>50), confidence was low.
725 RYMV (Sobemovirus), PLRV (Polerovirus), PeMV (Enamovirus), MBV (Barnavirus),
726 HuSRV (Hubsclerovirus), XyNV (outgroups).

727 **Figure 2:** Comparison of RdRp 3D model and structures. RdRp domains, palm, fingers and
728 thumb, are colored on RYMV model, grey, purple and pink, respectively. The RdRp
729 picornavirales feature was conserved i.e., index finger structure interacted with the top of the
730 thumb and trapped the ring finger underneath it (black arrows). RdRp structures of
731 picornaviruses (PDB 1RA6, Poliovirus and PDB 1U09, Foot-and-mouth disease virus) are
732 compared to RYMV model and colored based on root mean square deviations (RMSD). The
733 regions with high local structural similarity (RMSD<1Å) are shown dark blue. The regions
734 colored light blue (1<RMSD<2) and green (2<RMSD<3) were similar whereas colored other
735 colors, similarity was low. The α-helix of picornaviruses located at the carboxy-terminal
736 extremity (thumb domain) was lacking in RdRp sobelivirales (shown red). The orientation of
737 the pinky finger in the RYMV predicted structure was closer to FMDV rather than PV (grey
738 arrows).

739 **Figure 3:** Superimposition of RdRp 3D models by genera (a) Sobemovirus (6 ICTV recognized
740 or candidate species), (b) Poler/movirus (5 species), (c) Enamovirus (4 species), (d)
741 Hubsclerovirus (2 recognized or candidate species), (e) Barnavirus (3 recognized or candidate
742 species), (f) outgroups (3 candidate species). Models were colored using standard rainbow code
743 (from N-terminus, dark blue, to C-terminus, red). Major variable segments are pointed out at
744 the intra-genera scale (black arrows) or between outgroups (white arrows).

745 **Figure 4:** Inter-genera superimposition of RdRp 3D models. For clarity reason, only one model
746 per genera was represented (sobemovirus RYMV, poler/movirus PLRV, enamovirus PeMV,
747 barnavirus, hubsclerovirus HuSRV) and only the regions of interest were colored using standard
748 rainbow code (see text). Grey arrows indicated regions with intra-genera variability (as reported
749 in Fig 3), black arrows pointed out regions with inter-genera variability. Highlighted in a circle,
750 the amino-terminal extremity of RdRps forms a conserved β-strand (dark blue) and interacts
751 with other conserved strands of the middle and ring fingers (yellow and green).

752 **Figure 5:** Optimized amino acid alignment of 23 viral RdRps. A-H conserved motifs are
753 framed, strictly conserved residues are indicated in red above. The conservation and RMSD
754 scores were measured at each site and were colored in red and blue to highlight respectively the
755 local similarities and the local differences between both alignments (see Material & Methods).
756 Regions of interest (see text) are figured by arrows, loops or helices.

757 **Figure 6:** Maximum-likelihood phylogenetic tree of sobelivirales based on the structural
758 alignement. The branches are colored according to the five genera and the outgroups. Number
759 at each node indicates bootstrap values based on 100 replicates (only values above 80 are
760 indicated).

761 **Figure 7:** PoW-transformed time tree of the 20 sobelivirales species. The horizontal blue bars
762 show 95% highest posterior density of the node age.

763 **Figure 8:** Clade-specific residues located on 3D RdRp models. Sobarna- and polena-specific
764 residues confirmed on the 99 dataset were figured as sticks on Cα trace of one representative
765 species (RYMV and AlEV). Sobarna- and polena-specific residues which belonged to
766 conserved motifs were colored red & blue, respectively, clade-specific residues of other
767 domains were colored orange & light blue, respectively. Remarkable residues at positions 263,
768 489 or in interaction 365-486 were indicated by arrows or circled in dotted line (see text).

769 **Figure 9:** Figure 9: Optimized motifs of sobeliviral RdRps. (a) For each conserved motif, RdRp
770 conserved residues were colored (black), additional sobeli-specific residues (purple), sobarna-
771 specific residues (red) and polena-specific residues (blue). Common clade-specific positions
772 were highlighted (yellow open boxes). Each sequence logo consists of stacks of symbols, one
773 stack for each position in the sequence. The overall height of the stack indicates the sequence
774 conservation at that position, while the height of symbols within the stack indicates the relative
775 frequency of each amino at that position. Position numbering referred to the alignment with 20
776 sequences. (b) Conserved motifs were indicated on 3D RdRp models of one representative
777 species of each sobeli-clade (sobarna-, RYMV and polena-, AlEV, respectively). Color code
778 was conserved, except clade-specific polymorphic positions colored yellow (arrows).
779 Additional sobeli-specific residues (purple) are circles in dotted lines.

780 **Supporting information**

781 **S1 Table :** Taxonomy and biological characteristics in the phylum Pisuviricota.

782 **S2 Figure :** Back side of RdRp 3D models colored by per-residue model confidence score.
783 Color code is similar to Figure 1. RYMV (Sobemovirus), PLRV (Polerovirus), PeMV
784 (Enamovirus), MBV (Barnavirus), HuSRV (Hubsclerovirus), XyNV (outgroups).

785 **S3 File :** Local pLDDT scores of representative Alphafold models
786 https://pat.cbs.cnrs.fr/sobemo/plddt.html

787 **S4 Figure :** Local Distance Difference Test (pLDDT) data for 5 models of the HcRNAV and
788 RYMV RdRps arranged according to rank by the AlphaFold algorithm.

789 **S5 Figure :** Back side of RYMV RdRp 3D model and picornaviral structures. PDB 1RA6,
790 Poliovirus and PDB 1U09, Foot-and-mouth disease virus. Color code is similar to Figure 2.

791

792 **S6 File :** RMSD between RYMV Alphafold model and PDB templates

793 https://pat.cbs.cnrs.fr/sobemo/rmsd.html

794 **S7 File :** Sequence, secondary structure and solvent accessibility alignments built using

795 mTM-align and STRIDE

796 https://pat.cbs.cnrs.fr/sobemo/align_2d.html

797 **S8 File :** Nucleic acid alignment derived from the amino acid alignment of the 23 Alphafold

798 models

799 https://pat.cbs.cnrs.fr/sobemo/align_dna.fasta

800 **S9 Figure :** Inter-genera superimposition of RdRp 3D models. For clarity reason, only one

801 model per genera was represented (sobemovirus RYMV, poler/movirus PLRV, enamovirus

802 PeMV, barnavirus, hubsclerovirus HuSRV). Major residues of conserved motifs were

803 indicated as black sticks and displayed alone.

804 **S10 File :** Subgroup-conserved amino acids colored and projected on RYMV 3D model

805 https://pat.cbs.cnrs.fr/sobemo/bipart

806 **S11 Table :** Additionnal 79 sequences (genus, name species and accession numbers).

807 **S12 Table :** Clade-specific residues. Strictly conserved residues and shared sites were

808 indicated (bold and underligned, respectively). Position numbers referred at the structural

809 alignment of the 20 sobelivirales sequences (without the 3 outgroups, see S13 File).

810 **S13 File** : MAFFT alignment of 119 sobeliviral RdRps

811 https://pat.cbs.cnrs.fr/sobemo/homologs.html

812

813 **S14 File :** Comparison of mTM-align structural alignment with MAFFT, Clustal Omega and

814 MUSCLE sequence alignments

815 https://pat.cbs.cnrs.fr/sobemo/align_comp.html

816 **S15 Figure :** Phylogenetic tree of sobelivirales based on the structural, MAAFT, Clustalo and

817 Muscle alignements.

818 **S16 File** : RMSD between six sobeliviral and two tombusviral Alphafold RdRp models

819 (BYDV-pav, Luteovirus, and RCNMV, Dianthovirus)

820 https://pat.cbs.cnrs.fr/sobemo/rmsd_bydv_rcnmv

Figure 1

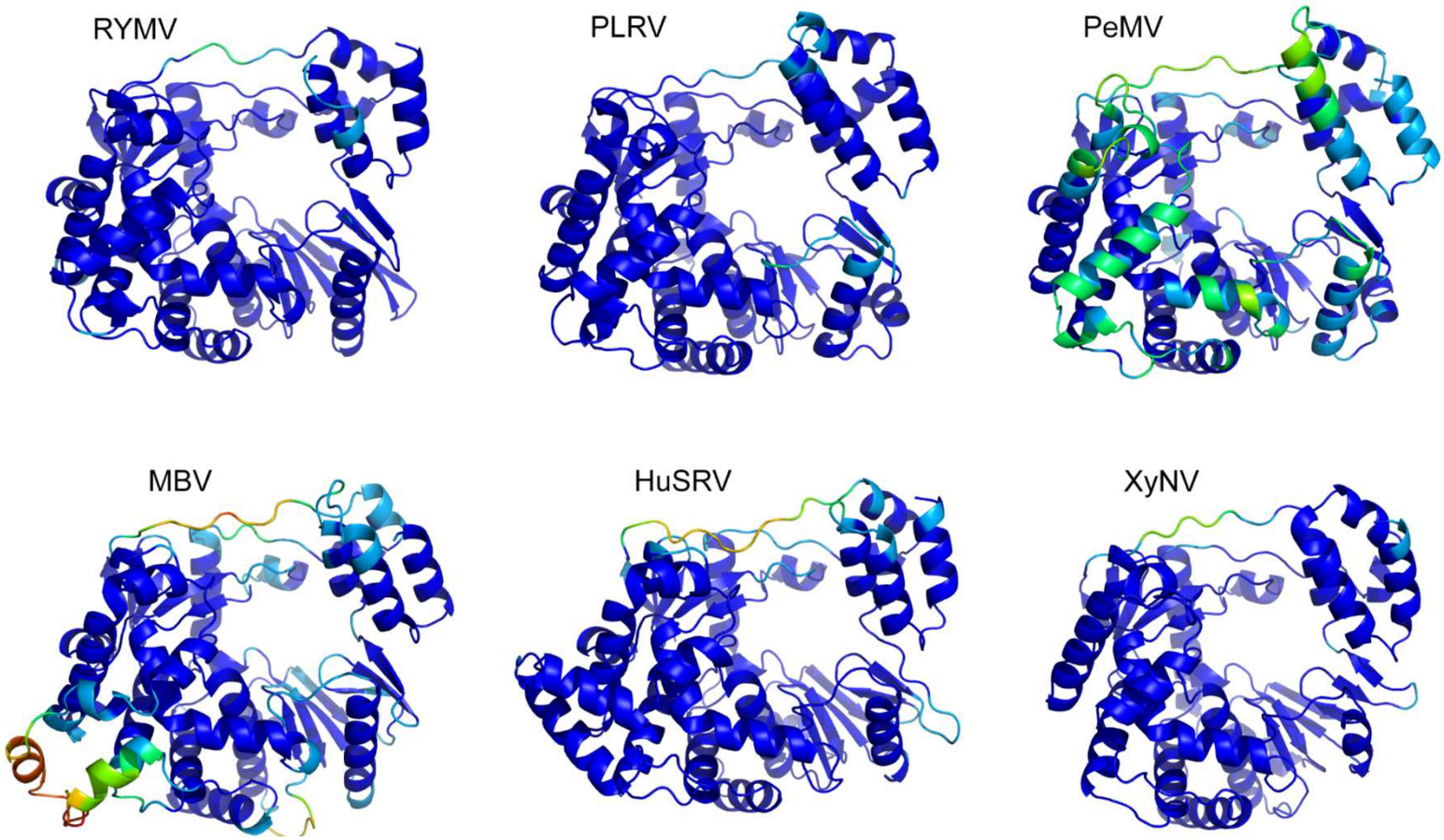


Figure 2

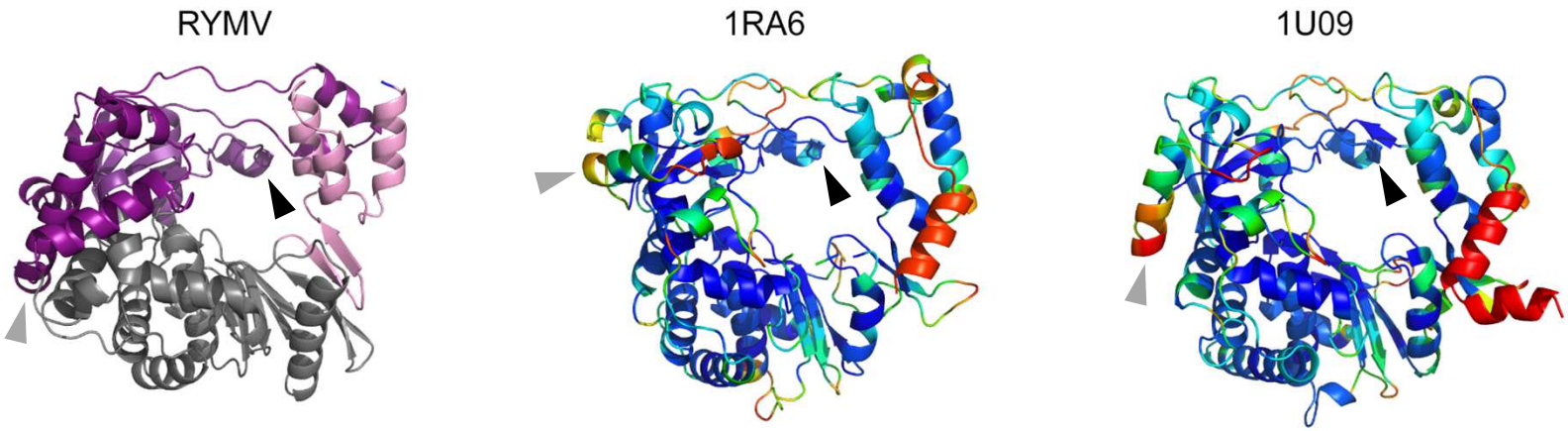


Figure 3

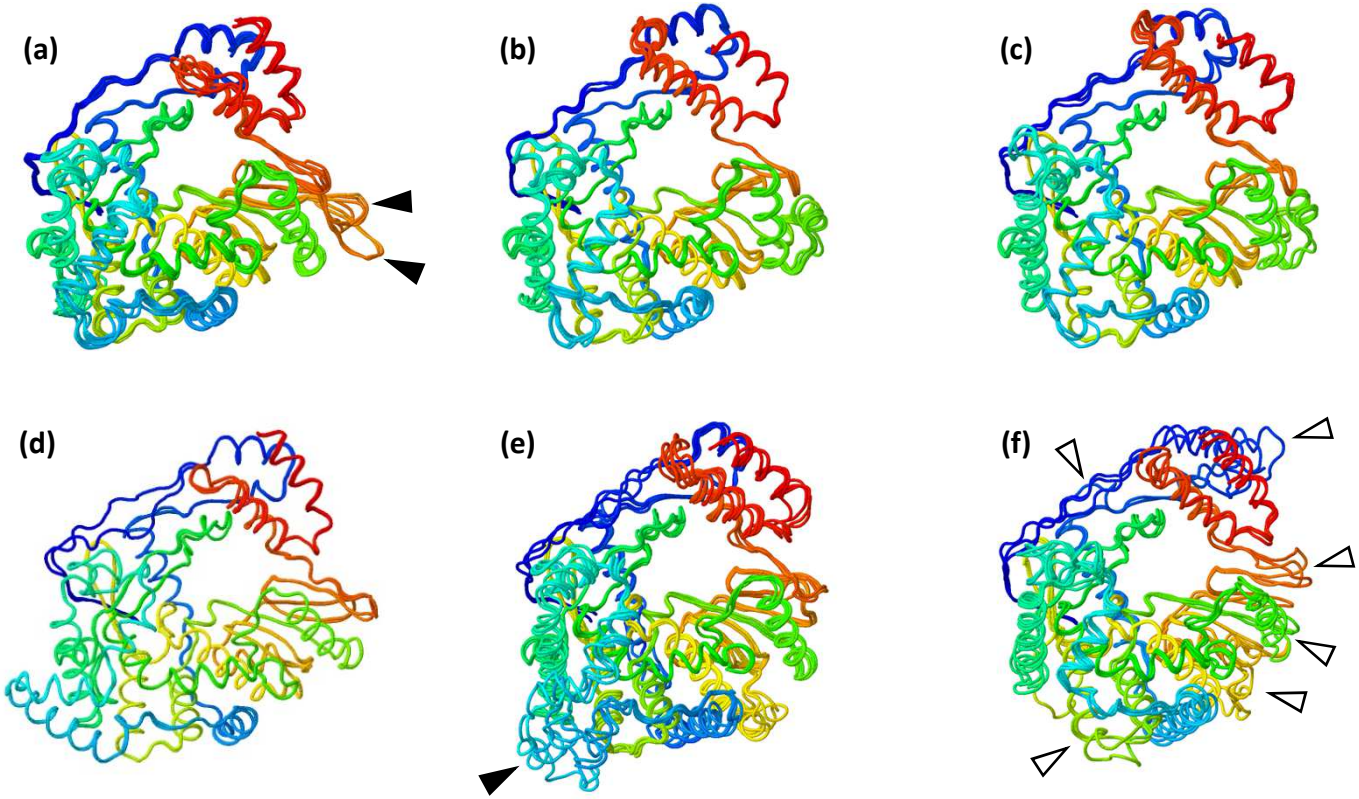


Figure 4

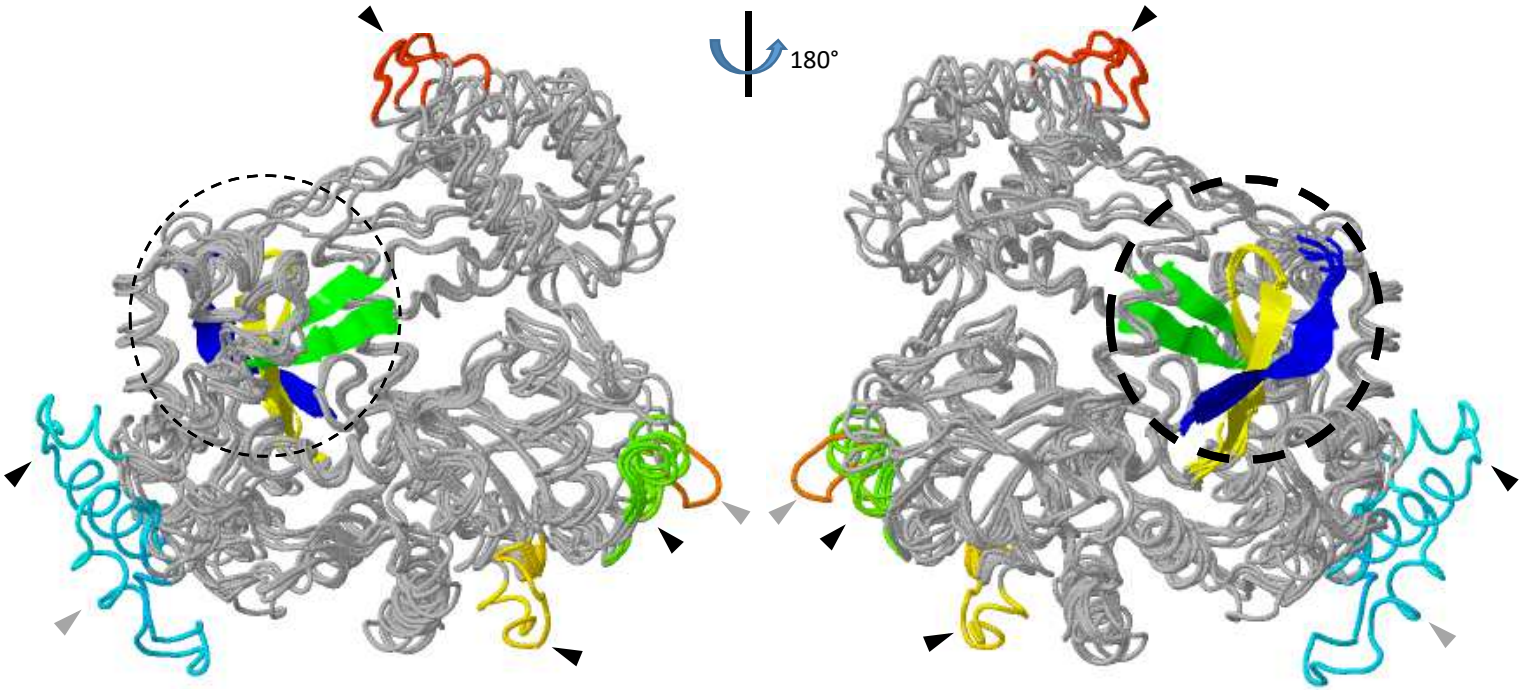


Figure 5 part I

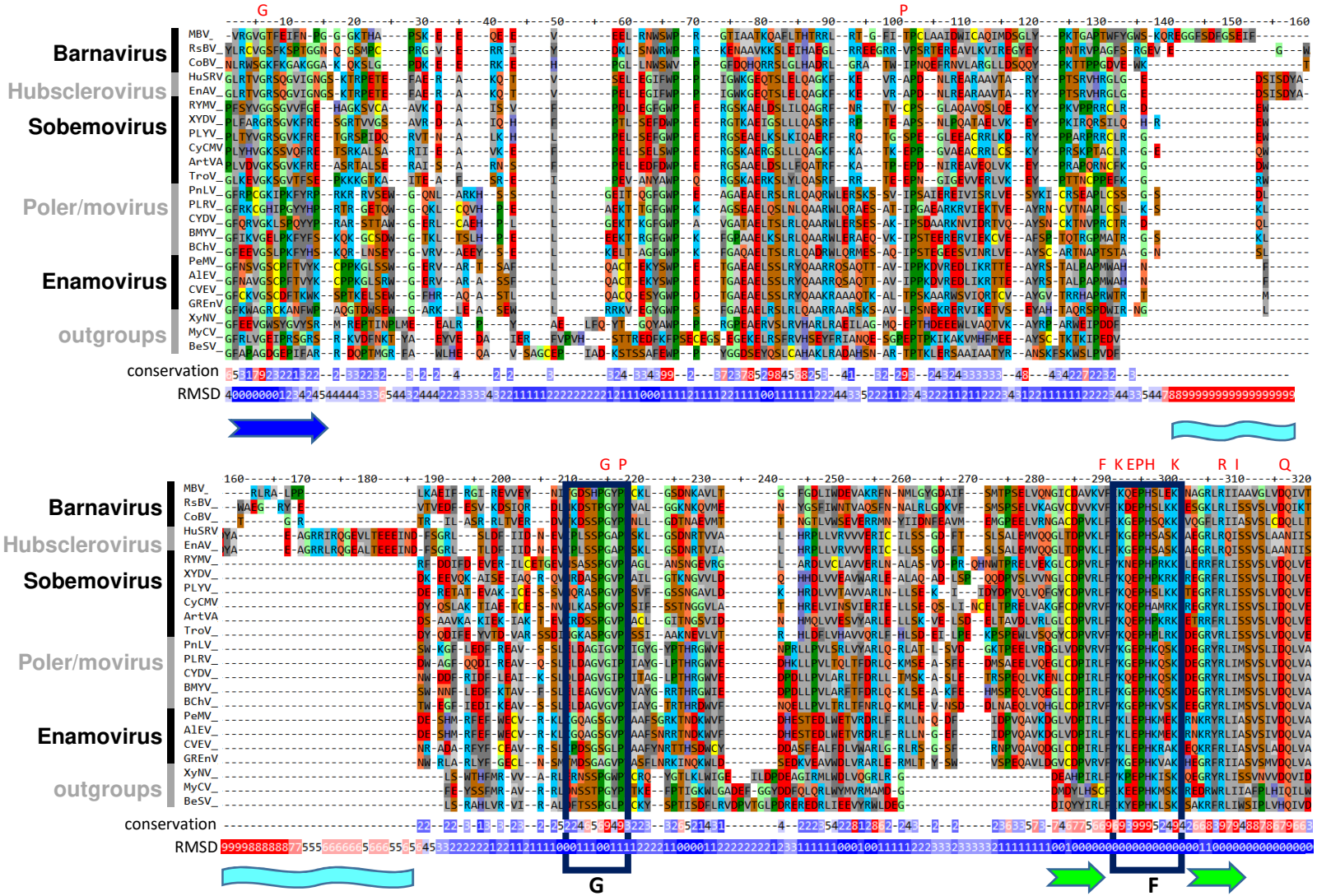
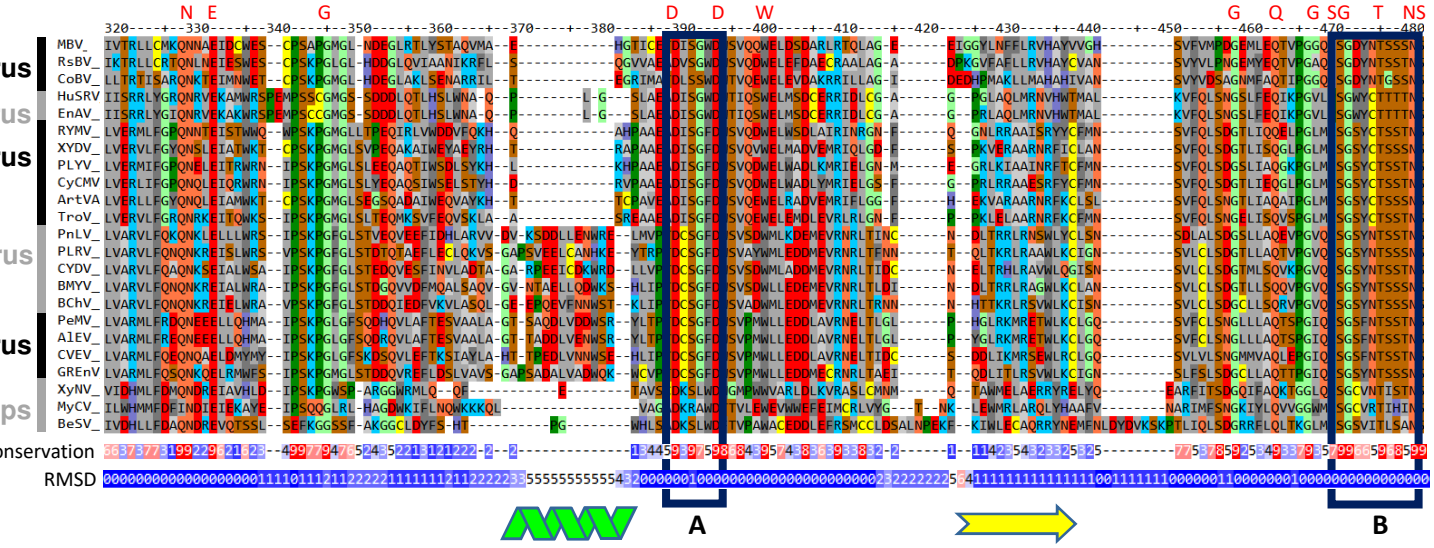


Figure 5 part II



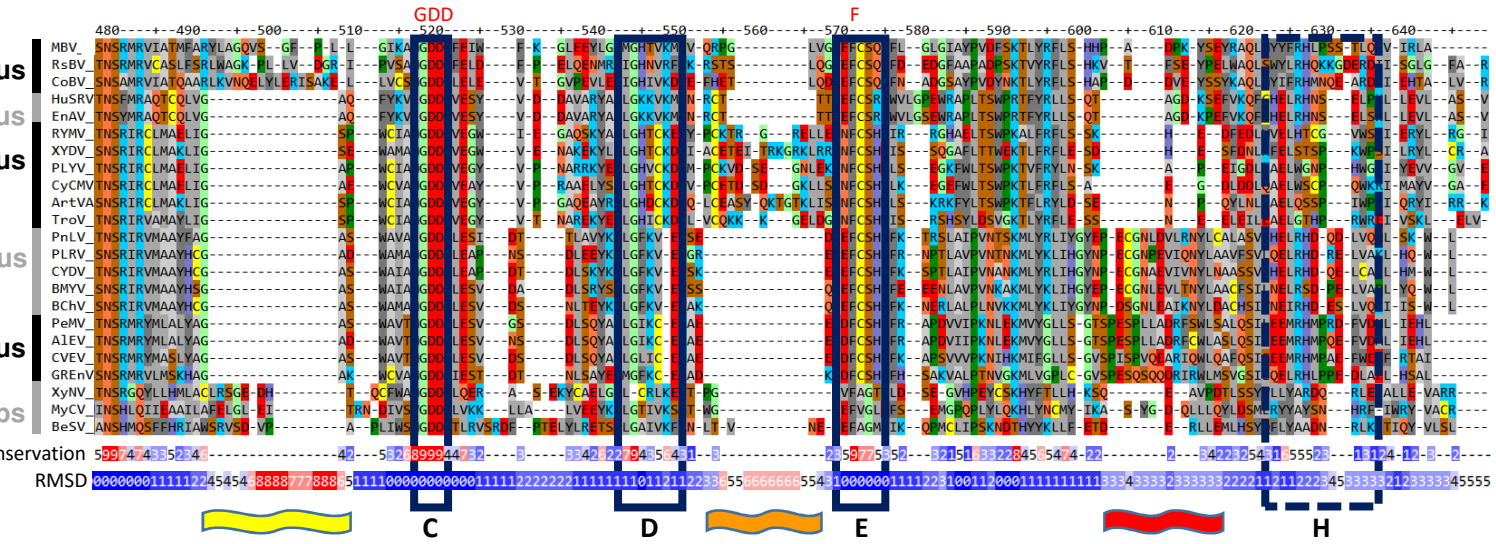


Figure 6

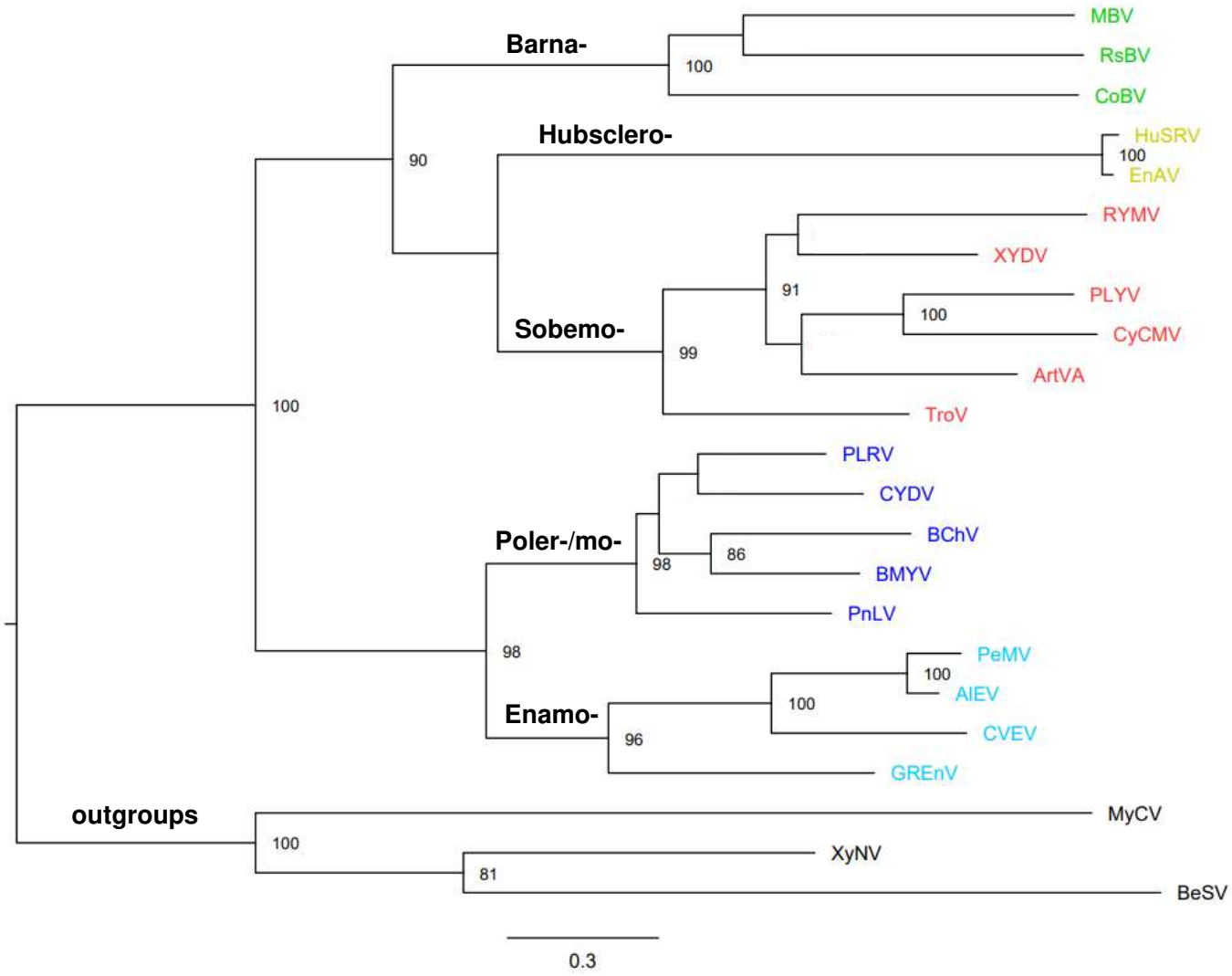

Barna-
100
MBV
RsBV
CoBV
Hubsclero-
90
HuSRV
100
EnAV
RYMV
XYDV
91
PLYV
100
CyCMV
Sobemo-
99
ArtVA
TroV
100
PLRV
CYDV
Poler-/mo-
BChV
86
98
BMYV
PnLV
98
PeMV
100
AlEV
Enamo-
100
CVEV
96
GREnV
outgroups
MyCV
100
XyNV
81
BeSV
0.3

Figure 7

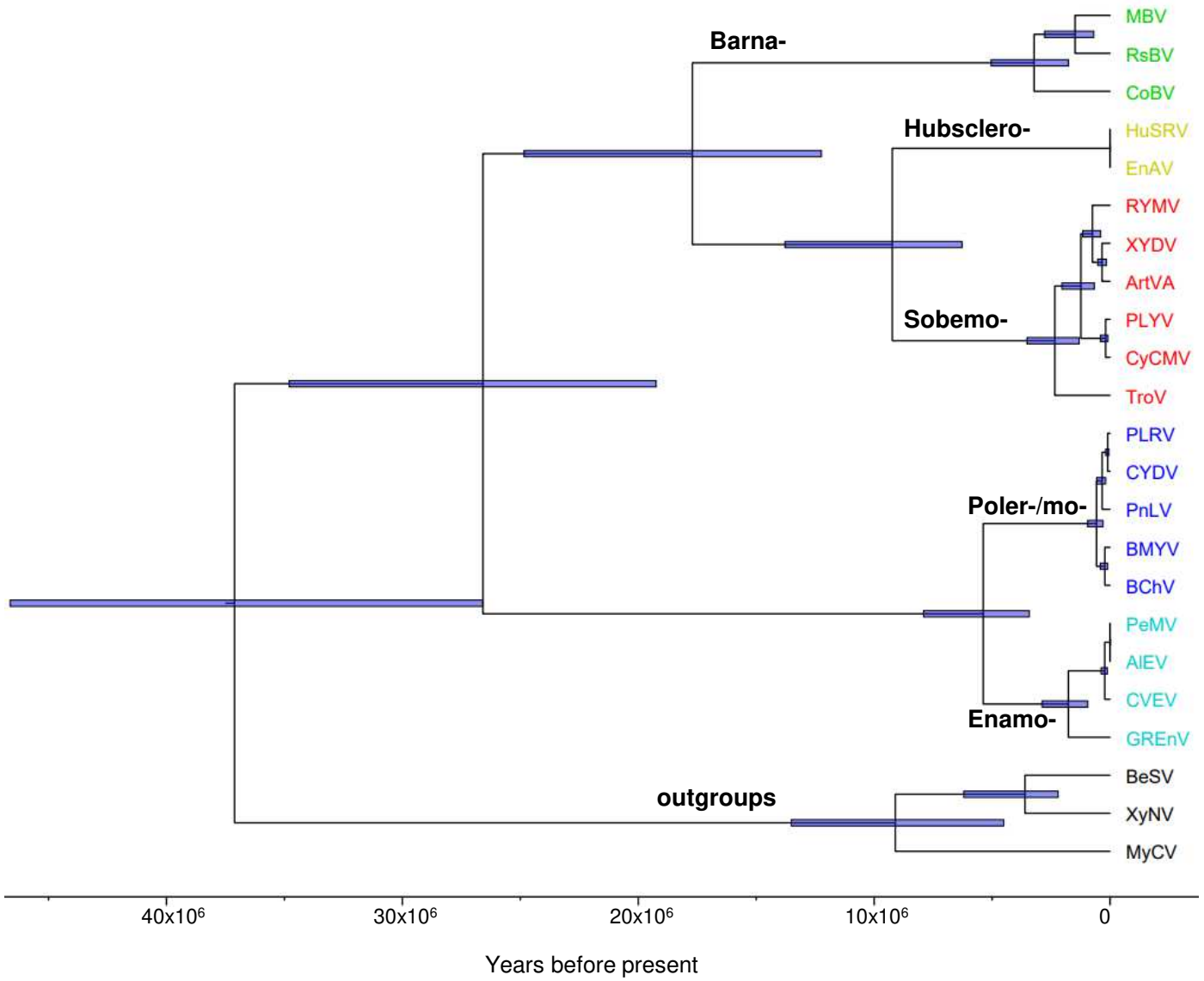

Barna-
Hubsclero-
Sobemo-
Poler-/mo-
Enamo-
outgroups
MBV
RsBV
CoBV
HuSRV
EnAV
RYMV
XYDV
ArtVA
PLYV
CyCMV
TroV
PLRV
CYDV
PnLV
BMYV
BChV
PeMV
AlEV
CVEV
GREnV
BeSV
XyNV
MyCV
40x10⁶
30x10⁶
20x10⁶
10x10⁶
0
Years before present

Figure 8

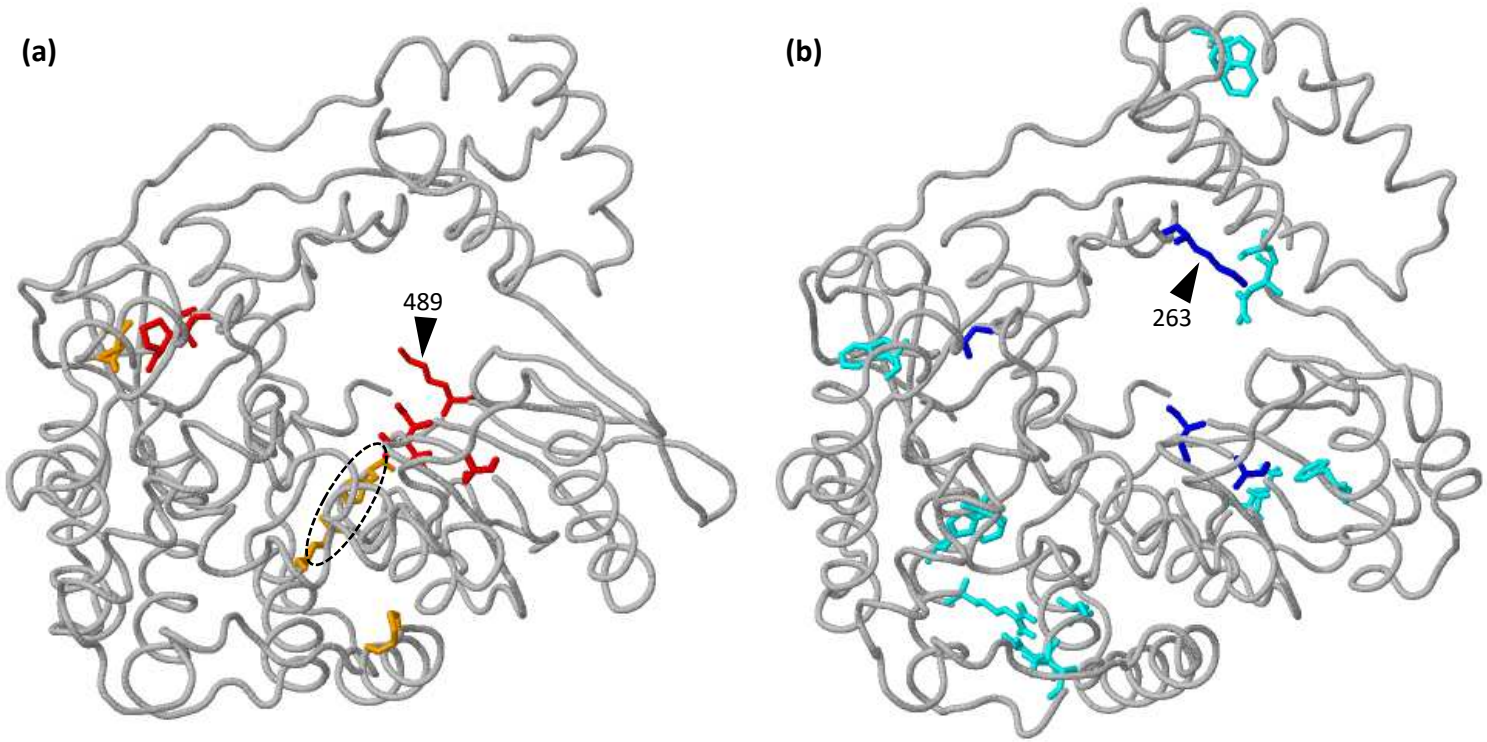
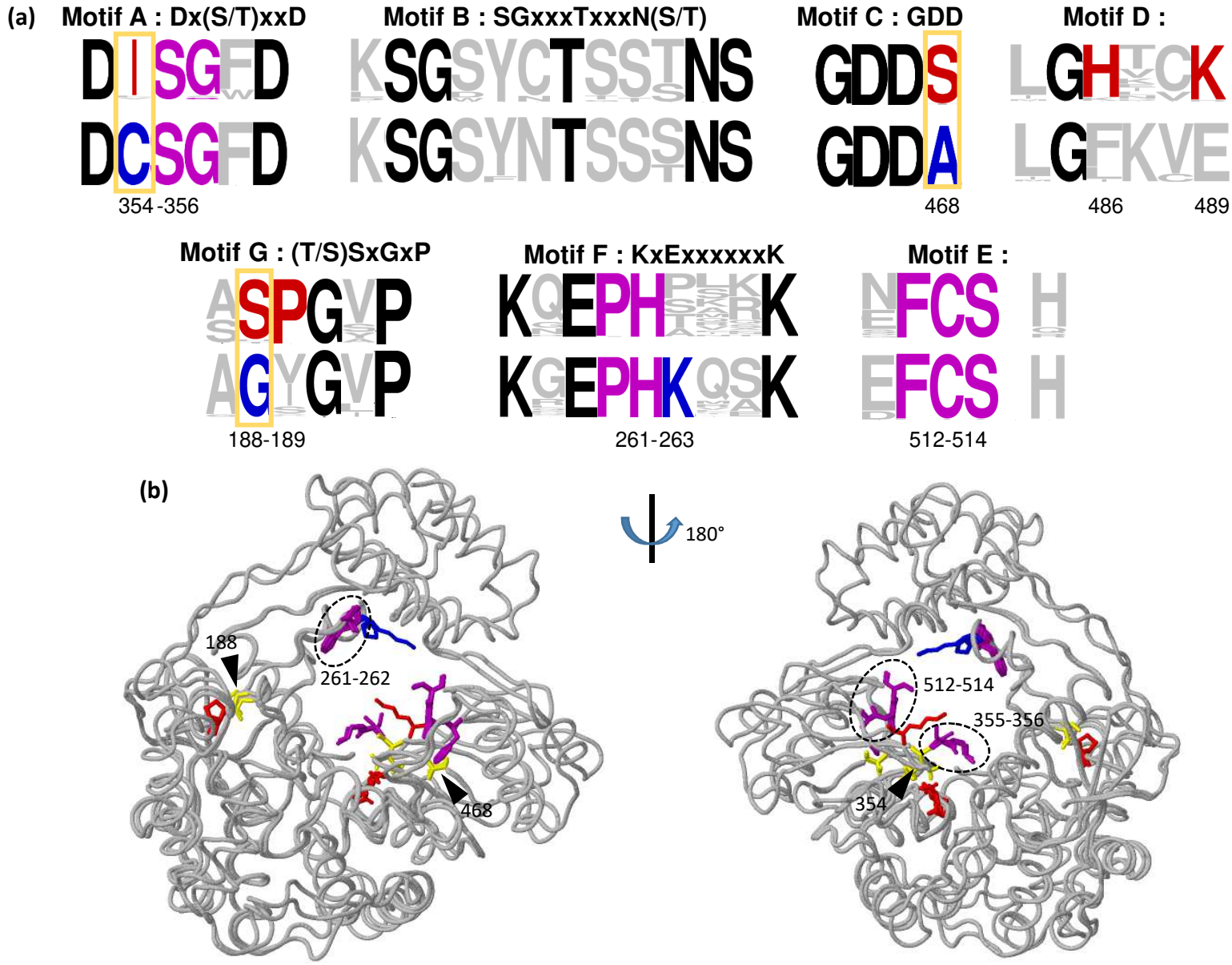


Figure 9


(a)
Motif A : Dx(S/T)xxD
354 -356
Motif B : SGxxxTxxxN(S/T)
Motif C : GDD
468
Motif D :
486
489
Motif G : (T/S)SxGxP
188-189
Motif F : KxExxxxxxK
261-263
Motif E :
512-514
(b)
180°
188
261-262
468
512-514
355-356
354