RESEARCH ARTICLE

Plant virus movement proteins originated from jelly-roll capsid proteins

Anamarija Butkovic[1], Valerian V. Dolja[2], Eugene V. Koonin[3], Mart Krupovic[1]*

1 Institut Pasteur, Université Paris Cité, CNRS UMR6047, Archaeal Virology Unit, Paris, France,
2 Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon, United States of America, **3** National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland, United States of America

* mart.krupovic@pasteur.fr

Abstract

Numerous, diverse plant viruses encode movement proteins (MPs) that aid the virus movement through plasmodesmata, the plant intercellular channels. MPs are essential for virus spread and propagation in distal tissues, and several unrelated MPs have been identified. The 30K superfamily of MPs (named after the molecular mass of tobacco mosaic virus MP, the classical model of plant virology) is the largest and most diverse MP variety, represented in 16 virus families, but its evolutionary origin remained obscure. Here, we show that the core structural domain of the 30K MPs is homologous to the jelly-roll domain of the capsid proteins (CPs) of small RNA and DNA viruses, in particular, those infecting plants. The closest similarity was observed between the 30K MPs and the CPs of the viruses in the families *Bromoviridae* and *Geminiviridae*. We hypothesize that the MPs evolved via duplication or horizontal acquisition of the CP gene in a virus that infected an ancestor of vascular plants, followed by neofunctionalization of one of the paralogous CPs, potentially through the acquisition of unique N- and C-terminal regions. During the subsequent coevolution of viruses with diversifying vascular plants, the 30K MP genes underwent explosive horizontal spread among emergent RNA and DNA viruses, likely permitting viruses of insects and fungi that coinfected plants to expand their host ranges, molding the contemporary plant virome.

Introduction

Viruses are ubiquitous, obligate intracellular parasites that infect (nearly) all life forms and show enormous diversity with respect to the routes of genome replication and expression, genome size, and gene composition [1–5]. This immense variability notwithstanding, virus proteins can be divided into 3 broad classes involved in distinct functions: (1) genome replication and expression; (2) virion assembly and structure; and (3) virus–host interactions [6,7]. The evolutionary trajectories of the proteins in the first 2 classes are drastically different from those in the third class. Proteins involved in replication and virion structure formation apparently were captured by viruses from hosts at early stages of evolution, and certain replication system components might even descend from primordial replicators antedating the emergence of modern type cells. Some of these proteins are virus hallmarks shared by many groups of

viruses spanning the boundaries of the virus realms and infecting widely diverse hosts including prokaryotes and eukaryotes. In a sharp contrast, proteins involved in virus–host interactions are typically host-specific and hence are restricted to relatively narrow groups of viruses, in most cases, within a virus family or order. Comparatively recent acquisition from the host is demonstrable for many of these proteins. A common route of evolution in this class of virus proteins is exaptation whereby a host or a virus protein is repurposed for a new function in virus–host interaction [8–10].

However, exceptions are known when homologous proteins mediate the interactions of diverse viruses with a particular group of hosts. A quintessential example are the movement proteins (MPs) of plant viruses that help the viruses move through plasmodesmata, membranous channels in plant cell walls [11,12]. The plasmodesmata are permeable for small molecules but have a size exclusion limit that precludes free passage of larger molecules, such as most proteins and RNA, and macromolecular complexes, such as virus particles [13]. Although the properties of the plasmodesmata vary widely across different plant cell types and species, typically, active transport mechanisms are required for the passage of large molecules and particles. Therefore, most plant viruses, to the exclusion of capsid-less *Endornaviridae*, *Narnaviridae*, and *Mitoviridae* which are vertically transmissible RNA replicons [14,15], encompass genes or gene blocks encoding dedicated MPs that mediate virus passage across plasmodesmata. The MPs have been shown to bind the virus genome RNA or DNA and increase the size exclusion limit of the plasmodesmata, providing channels for passage of virions or virus genomes [16,17].

The most common MPs by far belong to the so called 30K superfamily that spans 2 realms of viruses, *Riboviria* including its both kingdoms, *Orthornavirae* and *Pararnavirae*, and *Monodnaviria*. More specifically, the 30K superfamily MPs are encoded by numerous families of RNA viruses within *Orthornavirae* (*Alphaflexiviridae*, *Secoviridae*, *Betaflexiviridae*, *Tombusviridae*, *Bromoviridae*, *Virgaviridae*, *Tospoviridae*, *Botourmiaviridae*, *Fimoviridae*, *Phenuiviridae*, *Aspiviridae*, *Kitaviridae*, *Mayoviridae*, and *Rhabdoviridae*) and the expansive families *Caulimoviridae* within *Pararnavirae* and *Geminiviridae* within *Monodnaviria* [18,19]. Although viruses encoding 30K MPs were previously only described in "higher" vascular plants, 30K MP-like sequences closely related to those of nepoviruses [20] and ophioviruses [21] were recently found in moss (*Bryopsida*, *Selaginellaceae*, *Lycopodiaceae* [nonvascular plants]), liverworts (*Lepidoziaceae* and *Anastrophyllaceae* [nonvascular plants]), and fern (*Vittaria lineata*, *Cyrtomium fortunei*, and *Lonchitis hirsute* [vascular plants]), basal plant lineages that were pivotal to the land plant evolution [22–24].

The prototype of the 30K superfamily is the MP of tobacco mosaic virus (TMV), a positive-sense RNA virus of the *Virgaviridae* family, the classical experimental model of virology (the name of the 30K superfamily comes from the molecular mass of the TMV MP, 30 kDa) [18,25]. The TMV MP is an RNA-binding protein that forms a ribonucleoprotein with the virus genomic RNA that is transported through the plasmodesmata by increasing the size exclusion limit [26–28], a mechanism likely used by most members of the 30K superfamily. However, some members of the 30K MP superfamily have been shown to change the size exclusion limit of the plasmodesmata via a different mechanism, namely, by forming tubular structures that mediate virion trafficking [19,29–31] or through interaction with virus capsid proteins (CPs) [29,32].

The broad conservation of the 30K MPs across diverse families of plant viruses including unrelated ones belonging to different realms implies the spread of the MP genes via horizontal gene transfer (HGT). However, the ultimate origin of the 30K MPs remains enigmatic because no homologs of these proteins have been detected by sequence similarity searches, even using the most sensitive of the available methods, whereas tertiary structures of the MPs have not

been determined. Comparison of the predicted secondary structure elements suggested that the 30K MPs share a common core that consists of 7 to 8 β-strands [18,19,25], and it has been noted that this core domain might be related to the single jelly-roll (SJR) fold found in CPs of numerous small viruses with icosahedral capsids [19]. However, because these predictions were not statistically significant, it remained unclear whether the similarities between CPs and MPs reflected homology [19].

Recently, protein structure prediction has been revolutionized by high performance machine learning-based methods, AlphaFold2 (AF2) and RoseTTAFold [33,34]. These methods consistently yield accurate structure predictions for globular proteins with many diverse homologs. We took advantage of these tools to probe the origin of the 30K superfamily of MPs. Comparisons of the AF2 models of the MPs to the available protein structures unequivocally demonstrated close structural similarities between the MPs and virus jelly-roll CPs. We therefore conclude that the 30K MPs evolved via ancient duplication of the SJR CP gene followed by exaptation for the movement function.

Methods

Representative protein sequences of 30K MPs, clustering and phylogenetic analysis

Sequences of 18 representative 30K MP superfamily proteins [19] (S1 Table) were downloaded from GenBank and used as queries in sequence similarity searches performed with blastp [35] against the nr_vir70_1_Nov database (E-value cutoff of 0.001) [36]. The retrieved sequences of MP homologs were clustered to 90% minimum sequence identity using MMseqs2 [37]. The resulting dataset was used for clustering analysis using CLANS [38] and maximum-likelihood phylogenetic analysis with IQ-TREE 1.6.12 using the options -m TEST -bb 1000 -alrt 1000 [39]. CLANS analysis, where the sequences are positioned in a multidimensional space based on the strength of their pairwise similarities, was performed with the PSI-BLAST option and e-value of 10^{-3}. The clusters were identified at *P-value* = 10^{-15}. For the maximum-likelihood analysis and the identification of the D motif in SJR CPs, the sequences were aligned with PROMALS3D [40] with default parameters. In the alignment used for the maximum-likelihood analysis, poorly aligned (low information content) positions were removed using trimal with the -gt 0.2 option [41]. Phylogenetic analysis was performed using IQ-TREE [39], with the protein substitution model detection. The tree was rooted at midpoint and visualized in iTOL [42]. Sensitive profile–profile comparisons for remote sequence similarity detection were performed using HHsearch [43] against the Protein Data Bank (PDB) database.

The local charge distribution plots of the selected MPs and SJR CPs were obtained with the help of the "*chargeCalculationLocal*" option in the "*idpr*" package in R, using window size option 21 [44].

Three-dimensional structure prediction and analysis

The 18 MP sequences selected to represent different virus families and the 8 MP sequences from viruses associated with mosses, liverworts, and ferns [20,21] were used as inputs for AF2 (version 2.1.1, [33]) and RoseTTAFold [34] structure prediction. In particular, we used RoseTTAFold when MP structures modeled by AF2 were of poor quality, as estimated using the local distance difference test (lDDT) [45]. The quality of the RoseTTAFold models was assessed using residue-wise CA-lDDT implemented in the end-to-end version of RoseTTAFold.

Structure-based searches were performed with the DALI [46] server, and structural similarities between the MPs and their homologs were evaluated based on the DALI Z scores. The Z

score measures the quality of the structural alignment, with scores above 2 generally considered significant. The structural matches were further evaluated by superimposition of the structures using the MatchMaker algorithm implemented in University of California, San Francisco (UCSF) Chimera [47], followed by visual inspection. The top 20 hits against the PDB50 database in DALI searches were extracted and used for all-against-all structure comparisons on the DALI server. As an additional structural alignment tool, we used MUSTANG [48] to generate a pairwise similarity matrix based on root-mean-square deviation (RMSD) values between all modeled MP structures and related CPs. Similarity matrices that were generated from the DALI and MUSTANG comparisons were used in "pvclust" R package version 2.2.-0 [49] to generate a dendrogram with bootstrap supports (approximately unbiased (AU) *p*-value was computed by multiscale bootstrap resampling) from a similarity matrix by average linkage clustering. The heatmaps were plotted using the "pheatmap" R package version 1.0.12 [50]. Different clustering methods were tested (average, complete, ward.D, single, mcquitty linkage methods) and the cophenetic correlation coefficient was calculated for all to determine the clustering method that best represented the data. The complete linkage clustering method proved to be the best choice with respect to the correlation coefficient values and biological interpretation of the clusters.

Results and discussion

Horizontal spread of the 30K superfamily movement proteins across the plant virome

The representative MPs (*n* = 18; S1 Table) that belong to the 30K superfamily were used as queries in one iteration of protein BLAST search against the virus database filtered to 70% identity (nr_vir70). The MP sequences detected during the parallel searches were dereplicated and clustered at 90% sequence identity, yielding 389 clusters of related sequences, representing 16 virus families. Representatives from each cluster were then subjected to CLANS analysis (S2 Table) to identify more coarse-grained clusters (Fig 1A). CLANS detected 16 clusters of MPs that mostly corresponded to virus families and grouped into 5 superclusters, including: (1) *Geminiviridae* (realm *Monodnaviria*); (2) *Aspiviridae*, *Fimoviridae*, and *Phenuiviridae* (phylum *Negarnaviricota* within *Orthornavirae*); (3) *Kitaviridae*, *Bromoviridae*, *Mayoviridae*, *Alphaflexiviridae*, *Virgaviridae (Furovirus)*, *Tombusviridae (Umbravirus)* (phylum *Kitrinoviricota* within *Orthornavirae*), and *Tospoviridae* (*Negarnaviricota*); (4) *Rhabdoviridae* (*Negarnaviricota*), *Caulimoviridae* (*Pararnavirae*), *Virgaviridae (Tobamovirus)*, *Betaflexiviridae*, and *Secoviridae*; (5) *Tombusviridae* (*Tombusvirus* and *Aureusvirus*) (*Kitrinoviricota*), and *Botourmiaviridae* (*Lenarviricota* within *Orthornavirae*) families (Fig 1A). Superclusters 1, 2 and 5 were homogeneous, each including 1 or several related virus families, but superclusters 3 and 4 each included highly diverse, distantly related viruses, implying multiple HGT events. Notably, different genera of the families *Virgaviridae* and *Tombusviridae* did not cluster together, unlike other virus families, and were represented in both superclusters 3, 4 and 5 suggestive of relatively recent HGT and non-orthologous (although homologous) MP gene replacements. Furthermore, among the virgaviruses, some encode a 30K MP, whereas others encompass the so-called triple gene block MPs [51], demonstrating exchangeability of unrelated movement machineries.

    To further explore the relationships among the 30K MPs, the 389 MP sequences representing the 90% identity clusters were aligned using PROMALS3D [40], and a maximum-likelihood phylogenetic tree was constructed (Fig 1B and S1 Data). Overall, the tree topology recapitulated the results of CLANS analysis, with clusters and superclusters forming clades, mostly, with high bootstrap support (Fig 1B).
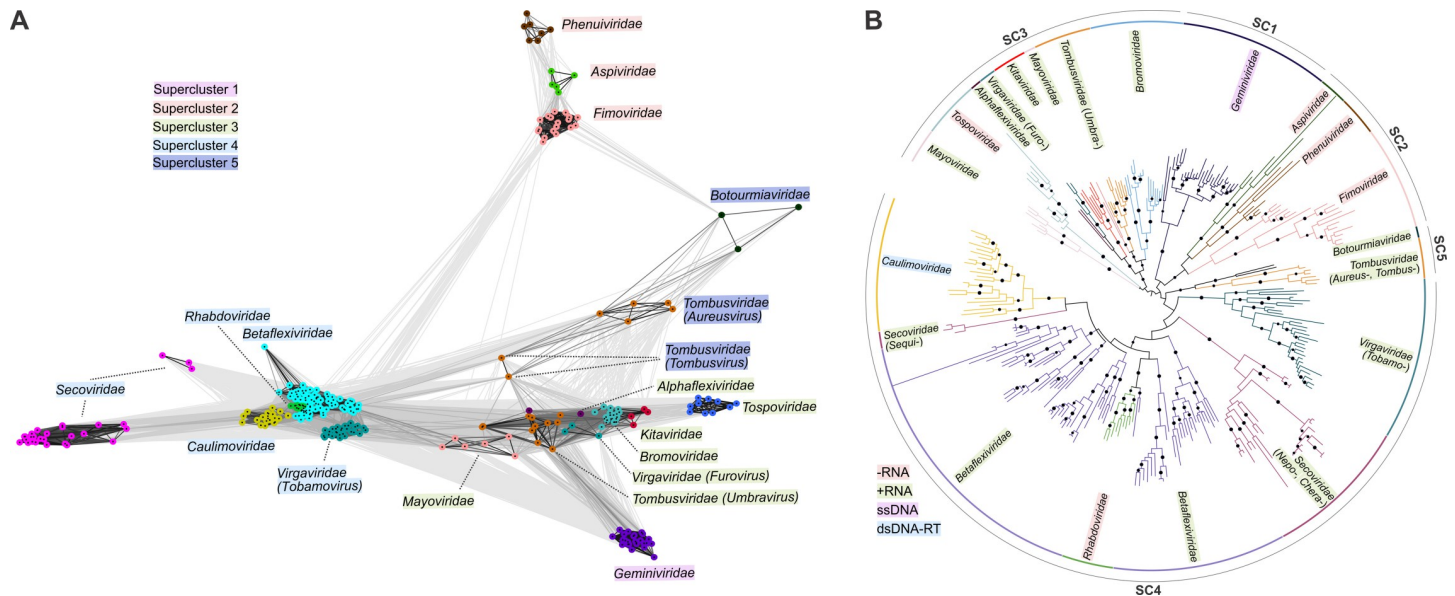
Fig 1. Sequence similarity and phylogeny of the 30K MPs. (**A**) Clustering of 30K MP sequences by pairwise sequence similarity (CLANS P-value $\leq 1 \times 10^{-15}$). The clusters are colored and named by virus families, while the outline boxes indicate if the virus family is part of superclusters 1–4. The lines represent sequence relationships, darker colors indicate closer sequence similarity. The HSP values used for clustering can be found in S2 Table. (**B**) Maximum-likelihood phylogenetic tree of 30K MP sequences obtained by IQ-TREE. SC, supercluster. The circles at the nodes indicate bootstrap branch support values $\geq$90. Superclusters 1–5 are also indicated. The tree in newick format can be found in S1 Data. HSP, high scoring pair; MP, movement protein.

https://doi.org/10.1371/journal.pbio.3002157.g001

Phylogenetic analysis further confirmed multiple horizontal exchanges of the MP genes during the evolution of plant viruses. For example, reverse-transcribing caulimoviruses, negative-sense (-)RNA viruses of the family *Rhabdoviridae* (genus *Cytorhabdovirus*) and positive-sense (+)RNA viruses of the family *Secoviridae* (genus *Sequivirus*) are nested among MPs of *Betaflexiviridae*, another family of (+)RNA viruses, suggesting that the latter virus group acted as a superspreader of the MP genes during the evolution of plant viruses. By contrast, (-)RNA *Tospoviridae* cluster with different families of (+)RNA viruses in the supercluster 3, with the rest of (-)RNA viruses forming a disconnected clade corresponding to supercluster 1 (Fig 1B), suggesting at least 3 independent MP introduction events into (-)RNA viruses. An extensive shuffling of the MP genes is also observed in the families *Tombusviridae* and *Virgaviridae*, where viruses from different genera form paraphyletic groups in the phylogeny.

The 30K MPs are homologous to the single jelly-roll capsid proteins

No high-resolution structure is available for any member of the 30K MP superfamily. Thus, to gain insights into the deeper evolutionary history of the 30K MPs through structure-based homology searches, we leveraged the state-of-the-art high performance structural modeling methods AF2 and RoseTTAFold [33,34]. The quality of the obtained 30K MP structural models, assessed using the lDDT [45], was found to be generally high in the conserved central region of the proteins, whereas the variable terminal regions were often unstructured and therefore modeled with a lower quality (S1 Fig and S3 Table and S2 Data). The well-structured central region was found to adopt the jelly-roll fold (Fig 2A) consisting of 8 antiparallel β-strands, typically denoted B through I, that form 2 juxtaposed β-sheets, composed of BIDG and CHEF strands, respectively (Fig 2A and 2B) [52,53]. The jelly-roll domain was readily identifiable in MPs encoded by viruses from all analyzed virus families (Fig 2B), with the molecular mass of the core jelly-roll domain varying between 14.6 kDa for tomato spotted wilt
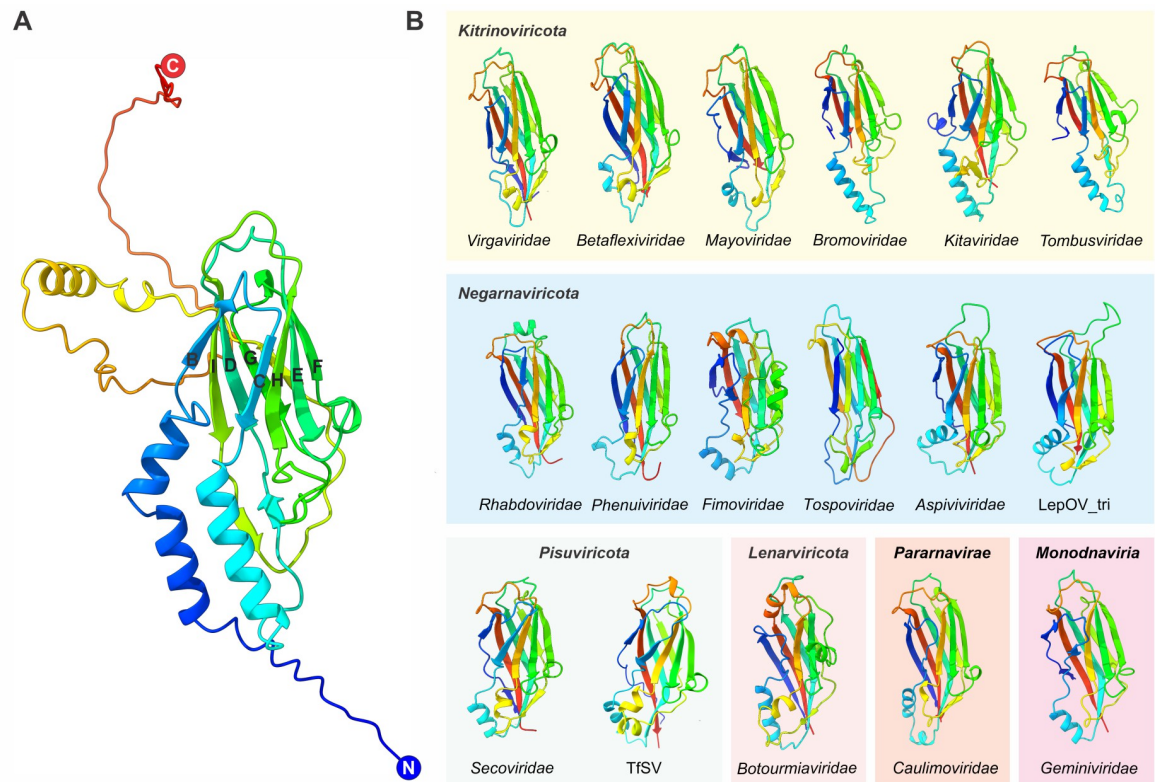
Fig 2. Structural modeling of 30K MPs. (**A**) Structural model of a representative full-length MP of the cabbage leaf curl virus (family *Geminiviridae*). The structure is colored using the rainbow scheme from blue (N-terminus) to red (C-terminus). The β-strands of the jelly-roll domain are indicated with Roman letters. (**B**) Structural models of the 30K MPs representing different virus families. The variable terminal ends were trimmed for the convenience of presentation. The structures are colored using the rainbow scheme from blue (N-terminus) to red (C-terminus). The structures are grouped according to established virus taxonomy. In the case of *Orthornavirae*, the corresponding phyla are indicated. Phylum *Kitrinoviricota*: *Virgaviridae* is represented by TMV, *Betaflexiviridae* by actinidia virus, *Mayoviridae* by raspberry bushy dwarf virus, *Bromoviridae* by cucumber mosaic virus, *Kitaviridae* by citrus leprosis virus C, *Tombusviridae* by carrot mottle virus; phylum *Negarnaviricota*: family *Rhabdoviridae* is represented by lettuce necrotic yellows virus, *Phenuiviridae* by rice stripe virus, *Fimoviridae* by rose rosette virus, *Tospoviridae* by tomato spotted wilt virus, *Aspiviridae* is represented by citrus psorosis virus and lepidozia ophiovirus tri (LepOV_tri) associated with hairy liverwort; phylum *Pisuviricota*: *Secoviridae* is represented by cherry rasp leaf virus and tomato fern seco-like virus (TfSV); phylum *Lenarviricota*: family Botourmiaviridae is represented by ourmia melon virus. Family *Caulimoviridae* (kingdom *Pararnaviae*) is represented by cauliflower mosaic virus, whereas family *Geminiviridae* (realm *Monodnaviria*) is represented by cabbage leaf curl virus. The PDB structure files for the modeled MPs can be found in S2 Data. MP, movement protein; PDB, Protein Data Bank; TMV, tobacco mosaic virus; TfSV, tomato fern seco-like virus.

https://doi.org/10.1371/journal.pbio.3002157.g002

virus (*Tospoviridae*) and 19 kDa for parsnip yellow fleck virus (*Secoviridae*), accounting for about half of the entire mass of the corresponding MPs.

The structural models of representative MPs were used as queries in DALI searches of the PDB database of protein structures. These searches retrieved as best hits the SJR CPs from diverse icosahedral viruses of eukaryotes, with significant Z scores ranging from 6.2 to 9.9 (S1 Table). The majority of the best hits were to CPs of the family *Tombusviridae* (S1 Table). However, the MPs of the viruses in the families *Caulimoviridae*, *Betaflexiviridae* (*Vitivirus*) and *Virgaviridae* produced the same highest scoring hit to the CP of satellite panicum mosaic virus (SPMV, *Papanivirus*; S1 Table). The rest of the hits were to CPs of viruses from other families, largely associated with plant hosts, but also including some animal viruses, such as those of the families *Astroviridae* and *Hepeviridae* (S1 Table).

Structural comparison of the 30K MPs and SJR CPs revealed closely similar jelly-roll topologies (Figs 3A and S2). The α-helix between G and H β-strands (not part of the canonical jelly-roll fold)
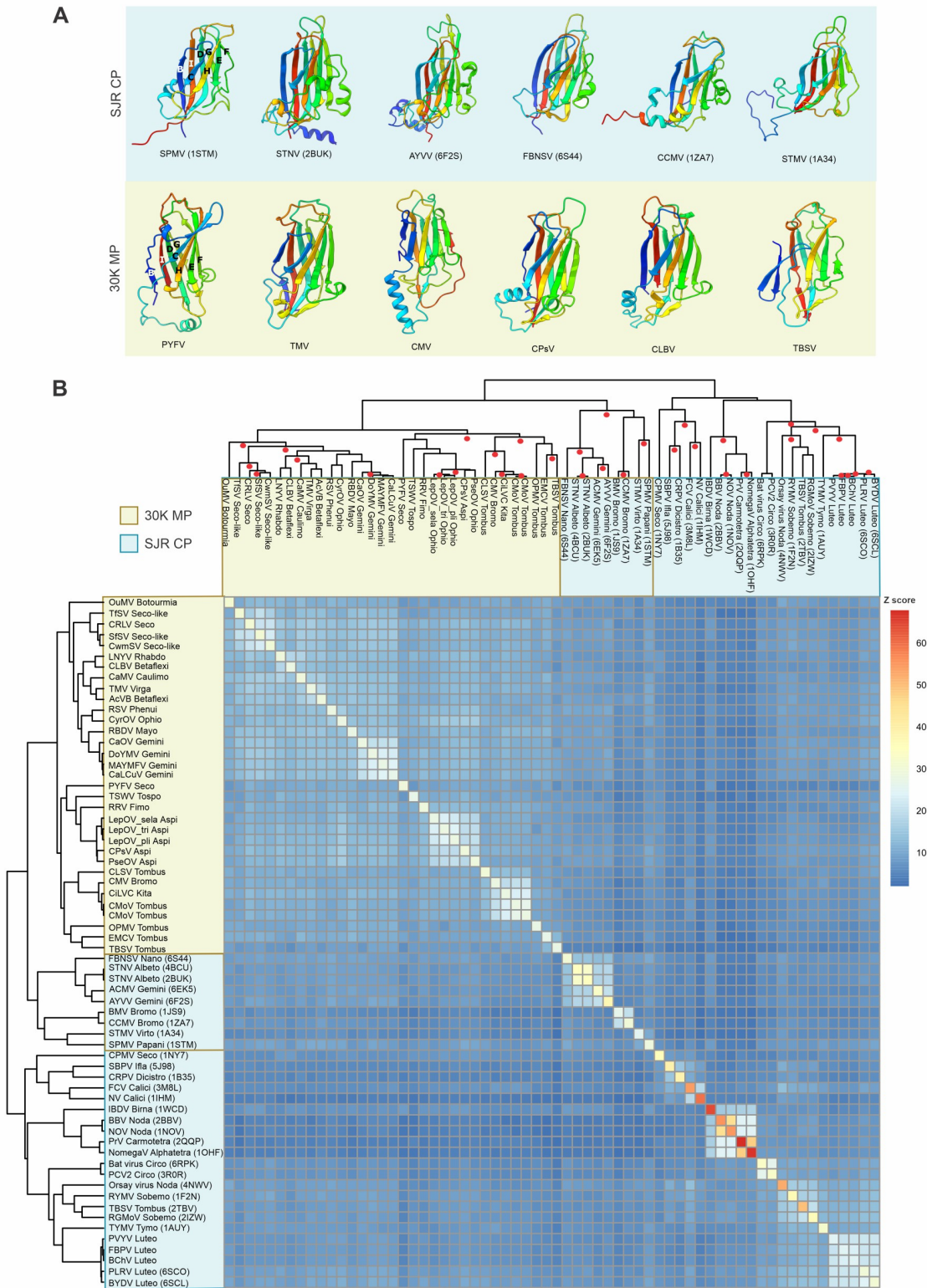
Fig 3. Structural similarity between SJR CPs and 30K MPs. (**A**) Structures of the SJR CPs homologous to 30K MPs obtained after a DALI search of PDB database, in the upper row highlighted with a blue background. The bottom row shows the jelly-roll region for the selected structures of 30K MP representatives, highlighted with a yellow background. The first structures on the utmost left in the upper and bottom row have the BIDG-CHEF β-strands annotated. The structures are colored using the rainbow scheme from blue (N-

terminus) to red (C-terminus). (**B**) Dendrogram and heatmap of complete linkage clustering of 30K representatives and SJR CPs. The red circles indicated in the top dendrogram, represent bootstrap values ≥90 obtained with R package "pvclust." The CPs and MPs are indicated in blue and yellow, respectively. Structures of 30K MPs and SJR CPs belong to: BMV, CCMV, FBNSV, STNV, ACMV, AYVV, STMV, SPMV, IPNV, IBDV, BBV, NoV, PrV, NomegaV, BFDV, PCV2, PhMV, TYMV, BYDV, BChV, PVYV, FBPV, RGMoV, RYMV, SBMV, TNV, BPMV, CPMV, FCV, NV, HRV16, SBPV, and CrPV. The newick format of the dendrogram obtained in DALI can be found in S3 Data. ACMV, African cassava mosaic virus; AYVV, ageratum yellow vein virus; BBV, black beetle virus; BChV, beet chlorosis virus; BFDV, beak and feather disease virus; BMV, brome mosaic virus; BPMV, bean pod mottle virus; BYDV, barley yellow dwarf virus; CCMV, cowpea chlorotic mottle virus; CP, capsid protein; CPMV, cowpea mosaic virus; CrPV, cricket paralysis virus; FBNSV, faba bean necrotic stunt virus; FBPV, faba bean polerovirus 1; FCV, feline calicivirus; HRV16, human rhinovirus; IBDV, infectious bursal disease virus; IPNV, infectious pancreatic necrosis virus; MP, movement protein; NomegaV, nudaurelia capensis omega virus; NoV, nodamura virus; NV, Norwalk virus; PCV2, porcine circovirus 2; PDB, Protein Data Bank; PhMV, physalis mottle virus; PrV, providence virus; PVYV, pepper vein yellows virus; RGMoV, ryegrass mottle virus; RYMV, rice yellow mottle virus; SBMV, southern bean mosaic virus; SBPV, slow bee paralysis virus; SJR, single jelly-roll; SPMV, satellite panicum mosaic virus; STMV, satellite tobacco mosaic virus; STNV, satellite tobacco necrosis virus; TNV, tobacco necrosis virus; TYMV, turnip yellow mosaic virus.

https://doi.org/10.1371/journal.pbio.3002157.g003

found in many MPs is also present in the SJR CPs of bromoviruses and solemoviruses as well as geminiviruses (e.g., ageratum yellow vein virus, PDB: 6F2S) and satellite tobacco necrosis virus (STNV, *Albetovirus*, PDB: 4BCU), suggesting a closer evolutionary relationship between the MPs and the CPs of these plant viruses. The consistent, significant structural similarity between the MPs and SJR CPs, and in particular, the same topology of the jelly-roll domains indicate that the 2 groups of proteins are indeed homologous. The SJR CPs are ubiquitous among the numerous groups of riboviruses and monodnaviruses with icosahedral capsids that infect diverse unicellular and multicellular eukaryotes from at least 9 eukaryotic kingdoms [52,54]. By contrast, the 30K MPs show a broad but scattered spread among viruses that primarily infect plants, i.e., restricted to a single eukaryotic kingdom, Chloroplastida, or in some cases, plants and their vectoring organisms. Thus, it appears highly likely that the 30K MPs evolved from the CPs.

To further analyze the evolutionary relationships between 30K MPs and SJR CPs, we performed an all-against-all comparison of the 30K MP structural models and SJR CP structures identified through the DALI searches. To avoid potential artifacts caused by the variable terminal regions of the MPs that have no counterparts in the CPs, for this analysis, only the jelly-roll domains of the MPs were considered. In the dendrogram obtained from the DALI Z scores, all MPs formed a single clade that was lodged within the diversity of the CPs (Fig 3B and S3 Data), suggesting monophyly of the 30K MP superfamily. The MP clade clustered with a distinct CP subclade that includes plant satellite RNA viruses, *Geminiviridae*, *Nanoviridae*, and *Bromoviridae* (Fig 3B). All these viruses infect plants and have highly compact SJR CP structures [55–57]. Given that satellite viruses are relatively rare in plant infections, the CPs of *Geminiviridae*, *Nanoviridae*, and *Bromoviridae* families seem to be the more likely ancestors of the 30K MPs. Furthermore, the CPs of bromoviruses and geminiviruses share with the 30K MPs the characteristic α-helical insertions within the jelly-roll domain. To corroborate these results, we used MUSTANG, an algorithm that aligns residues on the basis of similarity in patterns of both residue–residue contacts and local structural topology, creating a multiple structural alignment [48]. The dendrogram resulting from hierarchical clustering of the structural similarity values obtained with MUSTANG was largely congruent with that produced by DALI (S3 Fig and S4 Data). In this dendrogram, the MPs were nested within the CP diversity and formed a sister group to the CPs of the same assemblage of plant RNA and DNA viruses (geminiviruses, nanoviruses, bromoviruses, satellite viruses) as in the Z-score-based dendrogram, with the only notable difference being that the CP of SPMV was placed among the MPs. The latter placement is likely due to poor representation of the SPMV CP group (only 1 structure with no homologs identifiable at the sequence level) as well as a genuine high structural similarity to the MPs (S1 Table). Although we consider it unlikely that the SPMV CP evolved from an MP, this possibility cannot be formally ruled out.

Fig 4. Validation of the homology between SJR CPs and 30K MPs by sensitive sequence analysis. (**A**) Homologous regions between the CP of PLRV (PDB ID: 6SCO) and Camellia oleifera geminivirus (CaOV) 30K MP (accession number: QIE08114) obtained with HHsearch analysis against the PDB database. Secondary structure prediction is indicated by arrows for beta strands in yellow. (**B**) The structural model of CaOV 30K MP and PLRV CP. The homology region between the 2 proteins found in HHsearch against the PDB database is shown in red. The superposition of the conserved jelly-roll regions of CaOV 30K MP and PLRV CP is shown in the middle. The PLRV CP is colored light purple, and the CaOV 30K MP is colored light gray. CP, capsid protein; MP, movement protein; PDB, Protein Data Bank; PLRV, potato leaf roll virus; SJR, single jelly-roll.

https://doi.org/10.1371/journal.pbio.3002157.g004

Initial sequence similarity searches using BLASTP queried with 30K MP sequences yielded no significant matches outside the 30K superfamily, consistent with previous analyses [19]. However, in retrospect, after discovering the structural similarities between the MPs and SJR CPs, we reexamined this relationship using more sensitive comparisons of profile hidden Markov models (HMMs). Searches queried with the profile HMMs of 30K MPs against the profile HMMs of the PDB database yielded matches between the MPs of geminiviruses and SJR CP of potato leaf roll virus (PLRV, *Solemoviridae*; PDB ID: 6SCO), with significant probability scores (>90%, Fig 4A). The aligned regions mapped within the jelly-roll domains of the 2 proteins (Fig 4A). Consistently, the corresponding regions of the PLRV CP and geminivirus MP structural models, including the α-helix between β-strands C and D, could be superposed (Fig 4B). Thus, geminivirus MPs appear to more closely resemble the ancestral state of the 30K MP superfamily, with the relationship between the MPs and CPs still detectable at the sequence level. Phylogenetic and clustering analyses suggest that, following the divergence from the ancestral SJR CP, geminivirus MPs largely evolved vertically, without interfamilial horizontal exchange with other plant virus families (Fig 1), which conceivably contributed to the conservation of the ancestral features. We note, however, that the potentially archaic features of the geminivirus MPs do not necessarily imply that these proteins are ancestral to the 30K MPs of other viruses. Indeed, the vast virome of the vascular, particularly flowering plants, is dominated by RNA viruses of the kingdom *Orthornavirae* [58], suggestive of their rapid co-diversification and long coevolution with their hosts. Thus, a scenario under which the ancestral 30K MP gene was hosted by RNA viruses appears more parsimonious.

The MPs of multicellular algae and nonvascular plants

Ultimately, identification of the origins of the 30K MPs requires understanding the coevolution of contemporary plant virome with its plant hosts. It is generally recognized that

emergence and diversification of the plant virome occurred during terrestrialization of plants that apparently started with subaerial *Zygnematophyceae* freshwater algae followed by nonvascular terrestrial mosses and vascular plants [58,59]. The closest relatives of *Zygnematophyceae* algae for which viruses are known are algae of the genus *Chara*. The Charavirus canadiensis (CV-Can) and Charavirus australis (CV-Aus) viruses are 2 closely related, presumably rod-shaped (+)RNA viruses that encode TMV-like CPs along with the genes of unknown function that occupy the same genomic location as the 30K MP gene of TMV [60,61]. However, the proteins encoded by these genes exhibit no sequence similarity to any of the known MPs or other proteins. Notably, our AF2 modeling showed that the core structure of these *Chara* virus proteins was closely similar to that of the CPs of flexible filamentous viruses, such as alphaflexiviruses (S4 Fig). It seems likely that these proteins of *Chara* viruses evolved from capsid proteins of filamentous viruses to facilitate virus movement between *Chara* cells through the distinct algal plasmodesmata unrelated to those of vascular plants [11]. This evolutionary scenario parallels the exaptation of SJR CPs for the movement function of the 30K MPs, an analogy further strengthened by the acquisition of N-terminal extension observed in both cases (see below). Whether this protein functions in virus movement in *Chara* algae, remains to be validated experimentally.

As established previously, viruses encoding 30K MPs are present in lower vascular plants (ferns and lycophytes) and are ubiquitous in gymnosperms and angiosperms [19,62]. Recently, such 30K MP-encoding viruses were not only confirmed in ferns, but also found in nonvascular plants, namely, mosses and liverworts [20,21]. To address the possibility that the 30K MPs from moss, liverwort, and fern viruses resemble the ancestral state, we modeled their structures from secoviruses associated with common water moss (*Fontinalis antipyretica*), shoestring fern (*Vittaria lineata*), and tomato fern (*Lonchitis hirsuta*) [17] as well as from ophioviruses associated with hairy liverwort (*Lepidozia trichodes*) and basket liverwort (*Plicanthus hirtellus*), holly fern (*Cyrtomium fortunei*), Krauss' spike moss (*Selaginella kraussiana*), and Slender bog club-moss (*Pseudolycopodiella caroliniana*) [21]. All secovirus MPs grouped tightly with the MPs of the viruses of angiosperms in the genus *Nepovirus*, family *Secoviridae* (Fig 2B). Similarly, MPs of ophioviruses associated with nonvascular and lower plants clustered with the MP of angiosperm-infecting citrus psorosis ophiovirus (family *Aspiviridae*). Notably, besides the SJR domain, all ophiovirus MPs shared a characteristic C-terminal domain (PF11330; 30K_MP_C_Ter; HHpred probability = 98.3%) that is exclusive to ophiovirus 30K MPs. These observations, consistent with the previous phylogenetic analysis [20], suggest horizontal virus transfer (HVT) to lower vascular and nonvascular plants following the diversification of the *Secoviridae* and *Aspiviridae* families in angiosperms rather than emergence of 30K MPs in nonvascular mosses or liverworts that lack PDs.

Possible functions of the core and terminal regions of the 30K MPs in virus movement

The N- and C-terminal regions of the 30K MPs are predicted to be largely disordered, without recognizable folded domains, and vary dramatically both within and between different virus families (Fig 5A and S4 Table), which can explain the lower quality of the structural models in these regions (S1 Fig). The length of the N-terminal extensions (relative to the jelly-roll domain) varies from 9 amino acid residues (aa) in geminiviruses to 130 aa in mayoviruses, whereas the C-terminal extensions vary from 12 to 289 aa, in betaflexiviruses alone. The N-terminal region has been implicated in tubule polymerization and plasmodesmatal targeting of the MP [25,63], whereas the C-terminal region appears to be predominantly responsible for the interactions with CPs, virions, virion packaging into tubules, and long-distance movement
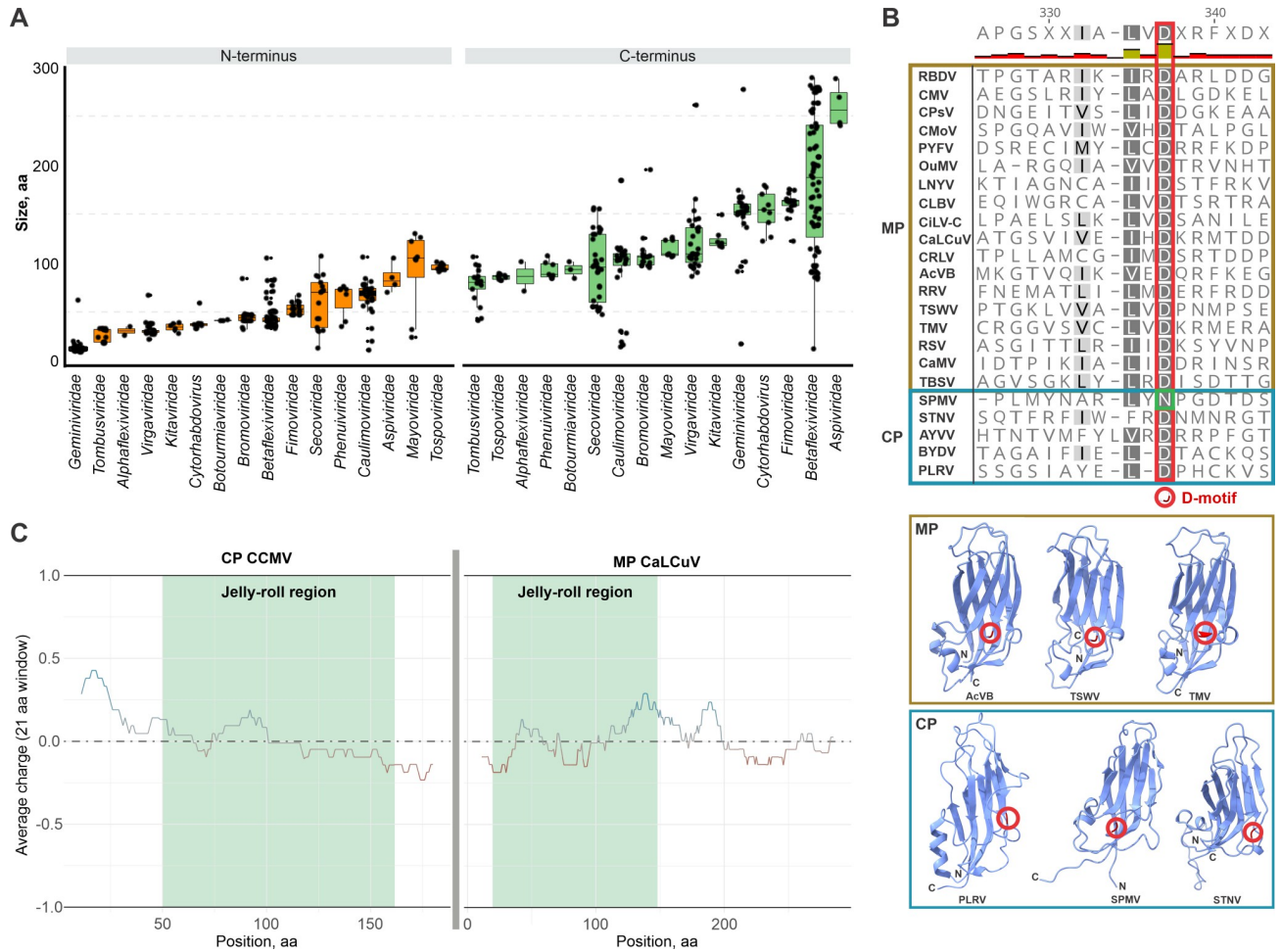
Fig 5. Length variation of the terminal regions of the 30K MPs, D motif conservation and charge distribution in the 30K MPs and SJR CPs. (**A**) Boxplot of the lengths of the N and C-terminal regions of 30K MPs. Orange boxes indicate values for N-terminal sizes and the green boxes indicate the C-terminal sizes. The x-axis denotes virus families and the y-axis the size of terminal ends by the number of amino acids. All the values are ordered by size from the smallest to the largest. The numeric values corresponding to the lengths of the N- and C-termini used for the boxplot can be found in S4 Table. (**B**) Top: the D motif region in the alignment of representative 30K MPs and SJR CPs. Note that in SPMV, the aspartate (D) is conservatively substituted with an asparagine (N). Bottom: the position of the D motif mapped on the MP and CP protein structures. The D motif is marked with a red circle. (**C**) Local charge distribution for CaLCuV 30K MP and CCMV SJR CP (PDB: 1ZA7) sequence by amino acid residue position (window size 21). The jelly-roll region is represented by a light green box. The height of the line above the gray threshold (0.0) indicates the value of the positive charges. The numerical values used to plot the charge distributions can be found in S5 Table. CCMV, cowpea chlorotic mottle virus; CP, capsid protein; MP, movement protein; SJR, single jelly-roll; SPMV, satellite panicum mosaic virus.

https://doi.org/10.1371/journal.pbio.3002157.g005

[32,64–69]. Overall, viruses in the families *Aspiviridae*, *Betaflexiviridae*, *Fimoviridae*, *Rhabdoviridae*, and *Geminiviridae* have longer N-terminal regions compared to the rest of the MPs, but this does not seem to correlate with the tubule formation (Fig 5A). The C-terminal MP regions are equally variable in size (Fig 5A), but again, there is no obvious correlation between the size of the extensions and the reported interactions between the C-termini of MPs with the respective virus CPs.

The most conserved feature of the 30K MPs is the D-motif [19] that includes a conserved aspartate residue located between β-strands E and F (marked dark red in Fig 5B), consistent with previous predictions on the position of the D-motif between 2 β-strands [18,19,25]. Alignment of the representative 30K MPs and SJR CPs reveals a degree of conservation of the D-motif in SJR CPs, particularly in CPs with the closest structural similarity to the MPs,

including geminiviruses, bromoviruses, and some satellite viruses (Figs 5B and S5). The sporadic presence of the D-motif in SJR CPs is consistent with a scenario under which MPs evolved from a specific group of CPs that contain this motif, rather than from a more ancient common ancestor with CPs.

Positively charged N-terminal regions of SJR CPs, commonly known as R-arms, bind viral RNA, or DNA genomes, promoting virion formation [70–72]. Similarly, positive charges have been shown to be required for nucleic acid binding by the 30K MPs [70,73–75]. However, whereas in SJR CPs, the positive charges involved in nucleic acid binding concentrate in the unstructured R-arms preceding the jelly-roll domain, in the 30K MPs, positively charged patches are distributed across the jelly-roll domain itself or the C-terminal extensions with no counterparts in the CPs (Figs 5C and S6 and S5 Table). We hypothesize that positive charge redistribution played an important role in the evolution of the CP into MP, facilitating the formation of distinct virus genome-MP complexes capable of passing through the plasmodesmata.

Evolution of plant virus movement proteins

Our results suggest that the 30K MPs originated from a distinct group of the SJR CPs (Fig 6). The viruses that encoded the ancestral SJR CP at the origin of the 30K MPs might no longer be part of the contemporary virome. Thus, it might not be possible to pinpoint with confidence the actual ancestor. Regardless of the exact identity of the ancestral virus, we hypothesize that the 30K MPs emerged in a virus that infected multicellular freshwater algae during their evolution on the route to nonvascular and later vascular land plants. After a chance duplication of the original SJR CP gene, exaptation of one of the copies for the movement function provided a strong fitness advantage by facilitating efficient spread of the virus through evolving plasmodesmata (Fig 6). The following rapid horizontal spread of the MP gene among emerging plant viruses with different genome types drove the diversification of the 30K MP superfamily and the dramatic expansion of the global plant virome.
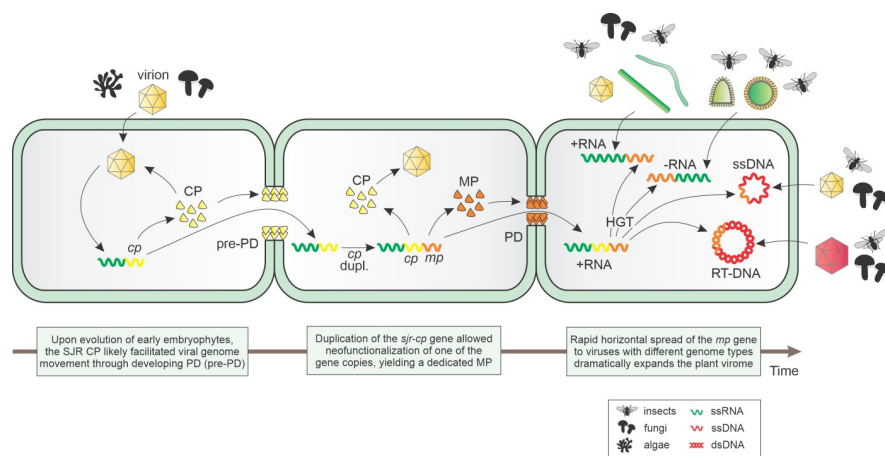

Fig 6. An evolutionary scenario for the origin of the 30K MP superfamily from SJR CPs. The ancestral virus is predicted to have an RNA genome (green wavy line) and encode an SJR CP, which was responsible for capsid formation and promoted intercellular movement through developing plasmodesmata. Duplication and neofunctionalization of the *cp* gene (yellow wavy line) led to the emergence of a dedicated *mp* gene (orange wavy line). Subsequently, the *mp* gene was horizontally transferred to other RNA viruses and viruses with DNA genomes (red wavy lines). Abbreviations: CP, capsid protein; (pre-)PD, (developing) plasmodesmata; MP, 30K movement protein; dupl., gene duplication; SJR, single jelly-roll.

https://doi.org/10.1371/journal.pbio.3002157.g006

The diversity of the contemporary plant virome that is dominated by RNA viruses remains to be a subset of the invertebrate RNA virome diversity [76,77]. Therefore, it appears most likely that the invertebrate virome seeded the plant virome through HVT enabled by plant-feeding nematodes and arthropods that currently serve as vectors for plant viruses. The expansion of the plant virome was contingent on the acquisition of MP, putting the horizontal spread of 30K MP among diverse virus families into the same timeframe. In support of this perspective, it was shown that the transgenic expression of the TMV MP in *Nicotiana benthamiana* enabled cell-to-cell and systemic movement of flock house virus, a single-stranded RNA insect virus not known to otherwise infect plants [78], providing experimental illustration of the critical role of MPs in the adaptation of insect viruses to plant hosts. Notably, the horizontal spread of the 30K MP gene placed it into widely different genome contexts including (+)RNA and (-)RNA viruses, reverse-transcribing viruses and single-strand DNA viruses. Furthermore, 30K MPs were combined with diverse virion architectures formed by the SJR CPs and several other, unrelated CPs as in the classic case of rod-shaped TMV or enveloped (-)RNA viruses. Conceivably, this diversity of the genomic contexts drove the functional and evolutionary diversification of the 30K MPs that remains to be explored in detail.

The route of 30K MP evolution represents a remarkable case of "intramural" exaptation, whereby a preexisting virus protein dramatically changed its function, providing strong selective advantage to the virus [10]. Notably, 3 divergent copies of non-jelly-roll CP of filamentous closteroviruses were exapted along a parallel route for distinct functions in virus capsid formation and transport [79]. One of the components of the triple gene block movement machinery, which represents an alternative to the 30K MPs [80], is a specialized superfamily 1 helicase, providing an additional example of functional exaptation of a preexisting virus protein for the function in virus movement. The exaptation of both the 30K MP and the helicase for enabling virus movement apparently involved addition of an extended unstructured N-terminal region that is important for the formation and transport of the virus nucleoprotein [51,81]. Finally, the putative MP of Chara viruses with an extended N-terminal domain and a core alphaflexivirus-like CP domain (S3 Fig) might represent yet another, independent case of CP exaptation for virus movement along a route similar to that of 30K MPs. These examples further emphasize exaptation as a key mechanism that shaped the virosphere ever since its inception and continues to contribute to virus diversification and evolution [6,10].

To conclude, this work demonstrates the potential of the new generation of protein structure prediction and analysis methods to illuminate key evolutionary events that remained out of reach of protein sequence-based analyses. Such findings, in turn, can be expected to inform further experimental studies.

Supporting information

S1 Table. Table of top single jelly-roll capsid protein hits in structural homology search with DALI using 30K movement proteins.
(XLSX)

S2 Table. High scoring pairs (HSP) values obtained by running psi-blast via CLANS and used for plotting the clustering network.
(XLSX)

S3 Table. plDDT values for all AlphaFold2 structural models. Each excel sheet corresponds to a virus MP.
(XLSX)

S4 Table. Sizes of N and C terminal MP ends per virus family used for the barplot.
(XLSX)

S5 Table. Distribution of local charge values for selected MPs in a window size 21, calculated with the "*chargeCalculationLocal*" option in the "*idpr*" package in R.
(XLSX)

S1 Data. The maximum-likelihood phylogenetic tree of the 30K MPs in newick format.
(NWK)

S2 Data. All MP AlphaFold models generated in this study.
(ZIP)

S3 Data. The dendrogram tree of 30K MPs and SJR CP hits obtained with DALI in newick format.
(NWK)

S4 Data. The dendrogram tree of 30K MPs and SJR CP hits obtained with MUSTANG in newick format.
(NWK)

S1 Fig. The per-residue confidence scores for AlphaFold2 (plDDT) and RoseTTAFold (Cα-lDDT) structural models. Regions with lDDT > 90 are expected to be modeled to high accuracy, whereas regions with lDDT between 70 and 90 are expected to be modeled well (a generally good backbone prediction). Abbreviated virus names are explained in the legend of Fig 2. Numerical data used to generate the plDDT plots can be found in S3 Table.
(TIF)

S2 Fig. The superimposition of the 30K MP of TMV (NP_597748) and the SJR CP from satellite tobacco mosaic virus (STMV, PDB: 1A34).
(TIF)

S3 Fig. Dendrogram and heatmap of complete linkage clustering of representative 30K MP and SJR CP based on the pairwise comparisons of the RMSD values calculated by MUSTANG. The red circles indicated in the top dendrogram, represent bootstrap values ≥90 obtained with R package "pvclust." The CPs and MPs are indicated in blue and yellow, respectively.
(TIF)

S4 Fig. Structural comparison of the pepino mosaic virus (PepMV) CP (PDB: 5FN1) and the putative MP of Charavirus canadiensis (QBG78689). The structures are colored using the rainbow scheme from blue (N-terminus) to red (C-terminus) and α-helices equivalent between the 2 proteins are numbered. For the charavirus protein, only the region corresponding to the PepMV CP is shown.
(TIF)

S5 Fig. The conservation of the D-motif in 30K MPs and SJR CPs. The alignment was made using PROMALS3D. Only the region encompassing the D-motif is shown.
(TIF)

S6 Fig. Plots of local charges in 21 amino acid sliding window for four 30K MP and four SJR CP representatives. The jelly-roll region is marked in light green.
(TIF)

Author Contributions

Conceptualization: Valerian V. Dolja, Eugene V. Koonin, Mart Krupovic.

Formal analysis: Anamarija Butkovic, Mart Krupovic.

Investigation: Anamarija Butkovic, Valerian V. Dolja, Eugene V. Koonin, Mart Krupovic.

Supervision: Mart Krupovic.

Visualization: Anamarija Butkovic.

Writing – original draft: Anamarija Butkovic, Eugene V. Koonin.

Writing – review & editing: Valerian V. Dolja, Mart Krupovic.

References

1. Zhang YZ, Chen YM, Wang W, Qin XC, Holmes EC. Expanding the RNA Virosphere by Unbiased Meta-genomics. Annu Rev Virol. 2019 Sep 29; 6(1):119–39. https://doi.org/10.1146/annurev-virology-092818-015851 PMID: 31100994

2. Dion MB, Oechslin F, Moineau S. Phage diversity, genomics and phylogeny. Nat Rev Microbiol. 2020 Mar; 18(3):125–38. https://doi.org/10.1038/s41579-019-0311-5 PMID: 32015529

3. Schulz F, Abergel C, Woyke T. Giant virus biology and diversity in the era of genome-resolved metage-nomics. Nat Rev Microbiol. 2022 Dec; 20(12):721–36. https://doi.org/10.1038/s41579-022-00754-5 PMID: 35902763

4. Krupovic M, Cvirkaite-Krupovic V, Iranzo J, Prangishvili D, Koonin EV. Viruses of archaea: Structural, functional, environmental and evolutionary genomics. Virus Res. 2018 Jan 15; 244:181–93. https://doi.org/10.1016/j.virusres.2017.11.025 PMID: 29175107

5. Koonin EV, Dolja VV, Krupovic M, Varsani A, Wolf YI, Yutin N, et al. Global Organization and Proposed Megataxonomy of the Virus World. Microbiol Mol Biol Rev. 2020 Mar 4; 84(2):e00061–19. https://doi.org/10.1128/MMBR.00061-19 PMID: 32132243

6. Krupovic M, Dolja VV, Koonin EV. Origin of viruses: primordial replicators recruiting capsids from hosts. Nat Rev Microbiol. 2019 Jul; 17(7):449–58. https://doi.org/10.1038/s41579-019-0205-6 PMID: 31142823

7. Koonin EV, Senkevich TG, Dolja VV. The ancient Virus World and evolution of cells. Biol Direct. 2006; 1 (1):29. https://doi.org/10.1186/1745-6150-1-29 PMID: 16984643

8. Krupovic M, Makarova KS, Koonin EV. Cellular homologs of the double jelly-roll major capsid proteins clarify the origins of an ancient virus kingdom. Proc Natl Acad Sci U S A. 2022 Feb; 119(5): e2120620119. https://doi.org/10.1073/pnas.2120620119 PMID: 35078938

9. Koonin EV, Krupovic M. The depths of virus exaptation. Curr Opin Virol. 2018 Aug; 31:1–8. https://doi.org/10.1016/j.coviro.2018.07.011 PMID: 30071360

10. Koonin EV, Dolja VV, Krupovic M. The logic of virus evolution. Cell Host Microbe. 2022 Jul 13; 30 (7):917–29. https://doi.org/10.1016/j.chom.2022.06.008 PMID: 35834963

11. Brunkard JO, Zambryski PC. Plasmodesmata enable multicellularity: new insights into their evolution, biogenesis, and functions in development and immunity. Curr Opin Plant Biol. 2017 Feb; 35:76–83. https://doi.org/10.1016/j.pbi.2016.11.007 PMID: 27889635

12. Cilia ML, Jackson D. Plasmodesmata form and function. Curr Opin Cell Biol. 2004 Oct; 16(5):500–6. https://doi.org/10.1016/j.ceb.2004.08.002 PMID: 15363799

13. Brunkard JO, Runkel AM, Zambryski PC. The cytosol must flow: intercellular transport through plasmo-desmata. Curr Opin Cell Biol. 2015 Aug; 35:13–20. https://doi.org/10.1016/j.ceb.2015.03.003 PMID: 25847870

14. Hillman BI, Cai G. The Family Narnaviridae. Adv Virus Res. 2013; 86:149–76. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780123943156000064.

15. Fukuhara T. Endornaviruses: persistent dsRNA viruses with symbiotic properties in diverse eukaryotes. Virus Genes. 2019 Apr; 55(2):165–73. https://doi.org/10.1007/s11262-019-01635-5 PMID: 30644058

16. Navarro JA, Sanchez-Navarro JA, Pallas V. Key checkpoints in the movement of plant viruses through the host. Adv Virus Res. 2019; 104:1–64. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0065352719300119.

17. Wu X, Cheng X. Intercellular movement of plant RNA viruses: Targeting replication complexes to the plasmodesma for both accuracy and efficiency. Traffic. 2020 Dec; 21(12):725–36. https://doi.org/10.1111/tra.12768 PMID: 33090653

18. Mushegian AR, Koonin EV. Cell-to-cell movement of plant viruses: Insights from amino acid sequence comparisons of movement proteins and from analogies with cellular transport systems. Arch Virol. 1993 Sep; 133(3–4):239–57.

19. Mushegian AR, Elena SF. Evolution of plant virus movement proteins from the 30K superfamily and of their homologs integrated in plant genomes. Virology. 2015 Feb; 476:304–15. https://doi.org/10.1016/j.virol.2014.12.012 PMID: 25576984

20. Mifsud JCO, Gallagher RV, Holmes EC, Geoghegan JL. Transcriptome Mining Expands Knowledge of RNA Viruses across the Plant Kingdom. Simon AE, editor. J Virol. 2022 May; 31:e00260–e00222.

21. Debat H, Garcia ML, Bejerman N. Expanding the Repertoire of the Plant-Infecting Ophioviruses through Metatranscriptomics Data. Viruses. 2023 Mar 25; 15(4):840. https://doi.org/10.3390/v15040840 PMID: 37112821

22. de Vries J, Archibald JM. Plant evolution: landmarks on the path to terrestrial life. New Phytol. 2018 Mar; 217(4):1428–34. https://doi.org/10.1111/nph.14975 PMID: 29318635

23. Soltis PS, Folk RA, Soltis DE. Darwin review: angiosperm phylogeny and evolutionary radiations. Proc R Soc B. 2019 Mar 27; 286(1899):20190099.

24. One Thousand Plant Transcriptomes Initiative. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 2019 Oct 31; 574(7780):679–85.

25. Melcher U. The '30K' superfamily of viral movement proteins. Microbiology. 2000 Jan 1; 81(1):257–66. https://doi.org/10.1099/0022-1317-81-1-257 PMID: 10640565

26. Citovsky V, Wong ML, Shaw AL, Prasad BV, Zambryski P. Visualization and characterization of tobacco mosaic virus movement protein binding to single-stranded nucleic acids. Plant Cell. 1992 Apr; 4 (4):397–411. https://doi.org/10.1105/tpc.4.4.397 PMID: 1379865

27. Kiselyova OI, Yaminsky IV, Karger EM, Yu Frolova O, Dorokhov YL, Atabekov JG. Visualization by atomic force microscopy of tobacco mosaic virus movement protein–RNA complexes formed in vitro. J Gen Virol. 2001 Jun 1; 82(6):1503–8. https://doi.org/10.1099/0022-1317-82-6-1503 PMID: 11369897

28. Waigmann E, Lucas WJ, Citovsky V, Zambryski P. Direct functional assay for tobacco mosaic virus cell-to-cell movement protein and identification of a domain involved in increasing plasmodesmal permeability. Proc Natl Acad Sci U S A. 1994 Feb 15; 91(4):1433–7. https://doi.org/10.1073/pnas.91.4.1433 PMID: 8108427

29. Kumar G, Dasgupta I. Variability, Functions and Interactions of Plant Virus Movement Proteins: What Do We Know So Far? Microorganisms. 2021 Mar 27; 9(4):695. https://doi.org/10.3390/microorganisms9040695 PMID: 33801711

30. van Lent J, Storms M, van der Meer F, Wellink J, Goldbach R. Tubular structures involved in movement of cowpea mosaic virus are also formed in infected cowpea protoplasts. J Gen Virol. 1991 Nov 1; 72 (11):2615–23. https://doi.org/10.1099/0022-1317-72-11-2615 PMID: 1940857

31. Tilsner J, Taliansky ME, Torrance L. Plant Virus Movement. In: John Wiley & Sons, Ltd, editor. eLS [Internet]. 1st ed. Wiley; 2014. Available from: https://onlinelibrary.wiley.com/doi/10.1002/9780470015902.a0020711.pub2.

32. Takeda A, Kaido M, Okuno T, Mise K. The C terminus of the movement protein of Brome mosaic virus controls the requirement for coat protein in cell-to-cell movement and plays a role in long-distance movement. J Gen Virol. 2004 Jun 1; 85(6):1751–61.

33. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021 Aug 26; 596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844

34. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021 Aug 20; 373 (6557):871–6. https://doi.org/10.1126/science.abj8754 PMID: 34282049

35. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5; 215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712

36. Gabler F, Nam SZ, Till S, Mirdita M, Steinegger M, Söding J, et al. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. Curr Protoc Bioinformatics. 2020 Dec; 72(1):e108. https://doi.org/10.1002/cpbi.108 PMID: 33315308

37. Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nat Commun. 2018 Dec; 9(1):2542. https://doi.org/10.1038/s41467-018-04964-5 PMID: 29959318

38. Frickey T, Lupas A. CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics. 2004 Dec 12; 20(18):3702–4. https://doi.org/10.1093/bioinformatics/bth444 PMID: 15284097

39. Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res. 2016 Jul 8; 44(W1):W232–5. https://doi.org/10.1093/nar/gkw256 PMID: 27084950

40. Pei J, Kim BH, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 2008 Apr; 36(7):2295–300. https://doi.org/10.1093/nar/gkn072 PMID: 18287115

41. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009 Aug 1; 25(15):1972–3. https://doi.org/10.1093/bioinformatics/btp348 PMID: 19505945

42. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 2021 Jul 2; 49(W1):W293–6. https://doi.org/10.1093/nar/gkab301 PMID: 33885785

43. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. BMC Bioinformatics. 2019 Dec; 20(1):473. https://doi.org/10.1186/s12859-019-3019-7 PMID: 31521110

44. McFadden WM, Yanowitz JL. idpr: A package for profiling and analyzing Intrinsically Disordered Proteins in R. Permyakov EA, editor. PLoS ONE. 2022 Apr 18; 17(4):e0266929. https://doi.org/10.1371/journal.pone.0266929 PMID: 35436286

45. Mariani V, Biasini M, Barbato A, Schwede T. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics. 2013 Nov 1; 29(21):2722–8. https://doi.org/10.1093/bioinformatics/btt473 PMID: 23986568

46. Holm L. Using Dali for Protein Structure Comparison. In: Gáspári Z, editor. Structural Bioinformatics [Internet]. New York, NY: Springer US; 2020 [cited 2022 May 20]. p. 29–42. (Methods in Molecular Biology; vol. 2112). Available from: http://link.springer.com/10.1007/978-1-0716-0270-6_3.

47. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera?A visualization system for exploratory research and analysis. J Comput Chem. 2004 Oct; 25(13):1605–12. https://doi.org/10.1002/jcc.20084 PMID: 15264254

48. Konagurthu AS, Whisstock JC, Stuckey PJ, Lesk AM. MUSTANG: A multiple structural alignment algorithm. Proteins. 2006 May 30; 64(3):559–74. https://doi.org/10.1002/prot.20921 PMID: 16736488

49. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006 Jun 15; 22(12):1540–2. https://doi.org/10.1093/bioinformatics/btl117 PMID: 16595560

50. Kolde R. pheatmap: Pretty Heatmaps. R package version 1.0. 12. R Packag version 10. 2019;8.

51. Verchot-Lubicz J, Torrance L, Solovyev AG, Morozov SY, Jackson AO, Gilmer D. Varied Movement Strategies Employed by Triple Gene Block–Encoding Viruses. Mol Plant Microbe Interact. 2010 Oct; 23(10):1231–47. https://doi.org/10.1094/MPMI-04-10-0086 PMID: 20831404

52. Krupovic M, Koonin EV. Multiple origins of viral capsid proteins from cellular ancestors. Proc Natl Acad Sci U S A. 2017 Mar 21; 114(12):E2401–E2410. Available from: https://pnas.org/doi/full/10.1073/pnas.1621061114. PMID: 28265094

53. Rossmann MG, Johnson JE. Icosahedral RNA virus structure. Annu Rev Biochem. 1989; 58:533–573. https://doi.org/10.1146/annurev.bi.58.070189.002533 PMID: 2673017

54. Krupovic M, Dolja VV, Koonin EV. The virome of the last eukaryotic common ancestor and eukaryogenesis. Nat Microbiol. 2023 Jun; 8(6):1008–1017. https://doi.org/10.1038/s41564-023-01378-y PMID: 37127702

55. Ban N, Larson SB, McPherson A. Structural comparison of the plant satellite viruses. Virology. 1995 Dec 20; 214(2):571–83. https://doi.org/10.1006/viro.1995.0068 PMID: 8553559

56. Bennett A, Agbandje-McKenna M. Geminivirus structure and assembly. Adv Virus Res. 2020; 108:1–32. https://doi.org/10.1016/bs.aivir.2020.09.005 PMID: 33837714

57. Lucas RW, Larson SB, McPherson A. The crystallographic structure of brome mosaic virus. J Mol Biol. 2002 Mar 15; 317(1):95–108. https://doi.org/10.1006/jmbi.2001.5389 PMID: 11916381

58. Dolja VV, Krupovic M, Koonin EV. Deep Roots and Splendid Boughs of the Global Plant Virome. Annu Rev Phytopathol. 2020 Aug 25; 58(1):23–53. https://doi.org/10.1146/annurev-phyto-030320-041346 PMID: 32459570

59. Cheng S, Xian W, Fu Y, Marin B, Keller J, Wu T, et al. Genomes of Subaerial Zygnematophyceae Provide Insights into Land Plant Evolution. Cell. 2019 Nov 14; 179(5):1057–1067.e14. https://doi.org/10.1016/j.cell.2019.10.019 PMID: 31730849

60. Gibbs AJ, Torronen M, Mackenzie AM, Wood JT, Armstrong JS, Kondo H, et al. The enigmatic genome of Chara australis virus. J Gen Virol. 2011; 92(Pt 11):2679–2690. https://doi.org/10.1099/vir.0.033852-0 PMID: 21733884

61. Vlok M, Gibbs AJ, Suttle CA. Metagenomes of a Freshwater Charavirus from British Columbia Provide a Window into Ancient Lineages of Viruses. Viruses. 2019 Mar 25; 11(3):299. https://doi.org/10.3390/v11030299 PMID: 30934644

62. Mushegian A, Shipunov A, Elena SF. Changes in the composition of the RNA virome mark evolutionary transitions in green plants. BMC Biol. 2016 Dec; 14(1):68. https://doi.org/10.1186/s12915-016-0288-8 PMID: 27524491

63. Ding B, Haudenshield JS, Hull RJ, Wolf S, Beachy RN, Lucas WJ. Secondary plasmodesmata are specific sites of localization of the tobacco mosaic virus movement protein in transgenic tobacco plants. Plant Cell. 1992 Aug; 4(8):915–28. https://doi.org/10.1105/tpc.4.8.915 PMID: 1392601

64. Bertens P, Heijne W, Van der Wel N, Wellink J, Van Kammen A. Studies on the C-terminus of the Cowpea mosaic virus movement protein. Arch Virol. 2003 Jan 1; 148(2):265–79. https://doi.org/10.1007/s00705-002-0918-z PMID: 12556992

65. Aparicio F, Pallas V, Sanchez-Navarro J. Implication of the C terminus of the Prunus necrotic ringspot virus movement protein in cell-to-cell transport and in its interaction with the coat protein. J Gen Virol. 2010 Jul 1; 91(7):1865–70.

66. Brill LM, Nunn RS, Kahn TW, Yeager M, Beachy RN. Recombinant tobacco mosaic virus movement protein is an RNA-binding, α-helical membrane protein. Proc Natl Acad Sci U S A. 2000 Jun 20; 97 (13):7112–7.

67. Gafny R, Lapidot M, Berna A, Holt CA, Deom CM, Beachy RN. Effects of terminal deletion mutations on function of the movement protein of tobacco mosaic virus. Virology. 1992 Apr; 187(2):499–507. https://doi.org/10.1016/0042-6822(92)90452-u PMID: 1546450

68. Lekkerkerker A, Wellink J, Yuan P, van Lent J, Goldbach R, van Kammen AB. Distinct functional domains in the cowpea mosaic virus movement protein. J Virol. 1996 Aug; 70(8):5658–61. https://doi.org/10.1128/JVI.70.8.5658-5661.1996 PMID: 8764083

69. Bertens P, Wellink J, Goldbach R, van Kammen A. Mutational Analysis of the Cowpea Mosaic Virus Movement Protein. Virology. 2000 Feb; 267(2):199–208. https://doi.org/10.1006/viro.1999.0087 PMID: 10662615

70. Requião RD, Carneiro RL, Moreira MH, Ribeiro-Alves M, Rossetto S, Palhano FL, et al. Viruses with different genome types adopt a similar strategy to pack nucleic acids based on positively charged protein domains. Sci Rep. 2020 Mar 25; 10(1):5470. https://doi.org/10.1038/s41598-020-62328-w PMID: 32214181

71. Twarock R, Stockley PG. RNA-Mediated Virus Assembly: Mechanisms and Consequences for Viral Evolution and Therapy. Annu Rev Biophys. 2019 May 6; 48:495–514. https://doi.org/10.1146/annurev-biophys-052118-115611 PMID: 30951648

72. Patel N, Wroblewski E, Leonov G, Phillips SEV, Tuma R, Twarock R, et al. Rewriting nature's assembly manual for a ssRNA virus. Proc Natl Acad Sci U S A. 2017 Nov 14; 114(46):12255–60. https://doi.org/10.1073/pnas.1706951114 PMID: 29087310

73. Carmen Herranz M, Sanchez-Navarro JA, Saurí A, Mingarro I, Pallás V. Mutational analysis of the RNA-binding domain of the Prunus necrotic ringspot virus (PNRSV) movement protein reveals its requirement for cell-to-cell movement. Virology. 2005 Aug; 339(1):31–41. https://doi.org/10.1016/j.virol.2005.05.020 PMID: 15963545

74. Herranz MC, Pallás V. RNA-binding properties and mapping of the RNA-binding domain from the movement protein of Prunus necrotic ringspot virus. J Gen Virol. 2004 Mar 1; 85(3):761–8. https://doi.org/10.1099/vir.0.19534-0 PMID: 14993662

75. Dong Y, Li S, Zandi R. Effect of the charge distribution of virus coat proteins on the length of packaged RNAs. Phys Rev E. 2020 Dec 28; 102(6):062423. https://doi.org/10.1103/PhysRevE.102.062423 PMID: 33466113

76. Shi M, Lin XD, Tian JH, Chen LJ, Chen X, Li CX, et al. Redefining the invertebrate RNA virosphere. Nature. 2016 Dec; 540(7634):539–43. https://doi.org/10.1038/nature20167 PMID: 27880757

77. Dolja VV, Koonin EV. Metagenomics reshapes the concepts of RNA virus evolution by revealing extensive horizontal virus transfer. Virus Res. 2018 Jan 15; 244:36–52. https://doi.org/10.1016/j.virusres.2017.10.020 PMID: 29103997

78. Dasgupta R, Garcia BH, Goodman RM. Systemic spread of an RNA insect virus in plants expressing plant viral movement protein genes. Proc Natl Acad Sci U S A. 2001 Apr 24; 98(9):4910–5. https://doi.org/10.1073/pnas.081288198 PMID: 11296259

79. Dolja VV, Kreuze JF, Valkonen JPT. Comparative and functional genomics of closteroviruses. Virus Res. 2006 Apr; 117(1):38–51. https://doi.org/10.1016/j.virusres.2006.02.002 PMID: 16529837

80. Solovyev AG, Kalinina NO, Morozov SY. Recent Advances in Research of Plant Virus Movement Mediated by Triple Gene Block. Front Plant Sci. 2012; 3:276. Available from: http://journal.frontiersin.org/article/10.3389/fpls.2012.00276/abstract.

81. Makarov VV, Rybakova EN, Efimov AV, Dobrov EN, Serebryakova MV, Solovyev AG, et al. Domain organization of the N-terminal portion of hordeivirus movement protein TGBp1. J Gen Virol. 2009; 90(Pt 12):3022–32. https://doi.org/10.1099/vir.0.013862-0 PMID: 19675186