Revisiting the origins of the *Sobemovirus* genus: a case for ancient origins of plant viruses

Authors

Mahan Ghafari[1,*], Merike Somera[2], Cecilia Sarmiento[2], Annette Niehl[3], Eugénie Hébrard[4], Theocharis Tsoleridis[5], Jonathan Ball[5], Benoît Moury[6], Philippe Lemey[7], Aris Katsourakis[1], Denis Fargette[4,*]

[1]Department of Biology, University of Oxford, UK
[2]Department of Chemistry and Biotechnology, Tallinn University of Technology, Akadeemia tee 15, 12618 Tallinn, Estonia
[3]Julius Kühn Institute (JKI) – Federal Research Centre for Cultivated Plants, Institute for Epidemiology and Pathogen Diagnostics, Messeweg 11-12, Braunschweig, Germany
[4] PHIM Plant Health Institute, University of Montpellier, IRD, CIRAD, INRAE, Institut Agro, Montpellier, France
[5]The Wolfson Centre for Global Virus Research and School of Life Sciences, The University of Nottingham, Queen's Medical Centre, Nottingham, NG7 2UH
[6]INRAE, Pathologie Végétale, 84140 Montfavet, France
[7]Department of Microbiology, Immunology and Transplantation, Rega Institute, KU Leuven, Herestraat 49, B3000 Leuven, Belgium

*correspondence: mahan.ghafari@ndm.ox.ac.uk; denis.fargette@ird.fr

Abstract

The discrepancy between short- and long-term rate estimates, known as the time-dependent rate phenomenon (TDRP), poses a challenge to extrapolating evolutionary rates over time and reconstructing evolutionary history of viruses. The TDRP reveals a decline in evolutionary rate estimates with the measurement timescale, explained empirically by a power-law rate decay, notably observed in animal and human viruses. A mechanistic evolutionary model, the Prisoner of War (PoW) model, has been proposed to address TDRP in viruses. Although TDRP has been studied in animal viruses, its impact on plant virus evolutionary history remains largely unexplored. Here, we investigated the consequences of TDRP in plant viruses by applying the PoW model to reconstruct the evolutionary history of sobemoviruses, plant pathogens with significant importance due to their impact on agriculture and plant health. Our analysis showed that the *Sobemovirus* genus dates back over four million years, indicating an ancient origin. We found evidence that supports deep host jumps to Poaceae, Fabaceae, and Solanaceae occurring between tens to hundreds of thousand years ago, followed by specialization. Remarkably, the TDRP-corrected evolutionary history of sobemoviruses was extended far beyond previous estimates that had suggested their emergence during the Neolithic period. By incorporating sequences collected through metagenomic analyses, the resulting phylogenetic tree showcases increased genetic diversity, reflecting a deep history of sobemovirus species with major radiation events taking place during the Neolithic period, suggesting rapid diversification in that period. Our findings make a case for the possibility of deep evolutionary origins of plant viruses.

Introduction

Reconstructing deep evolutionary history of organisms relies on accurate estimation of their evolutionary rates. However, the discrepancy between short- and long-term evolutionary rate estimates in many organisms, in particular viruses, makes any extrapolation of evolutionary rates based on inference from short timescales challenging [1]. This problem, which later became known as the time-dependent rate phenomenon (TDRP) for viruses, describes a ubiquitous decline in the evolutionary rate estimates with the timescale of measurement. The TDRP is best explained empirically by a power-law rate decay and has been extensively documented in animal and human viruses [2–5]. Several factors can contribute to TDRP such as incomplete purifying selection [6,7], accelerated evolution upon introduction to a new host environment [8,9], and site substitution saturation [2,10]. In addition to these biological processes, several other factors such as misspecification of substitution models can also contribute to time-dependent changes in the evolutionary rate estimates [11]. In a recent work [10], a mechanistic evolutionary model, also known as the Prisoner of War (PoW) model, was proposed that readily explains and corrects for time-dependent rate effects across deep evolutionary timescales and reproduces the empirically observed power-law rate decay found for RNA and DNA viruses. This model has so far been implemented to reconstruct evolutionary history of many viruses including human and endemic coronaviruses, foamy viruses, hepaciviruses, human and simian immunodeficiency viruses, monkeypox virus, and variola virus [10,12–17].

TDRP has also been reported in plant viruses [18]. Like many rapidly-evolving animal RNA viruses, a few plant viruses with a robust molecular clock signal based on isolates collected over a few decades have a substitution rate of ca. 10^{-3} substitutions per site per year (s/s/y) [19]. By contrast, their substitution rate estimated over a century and longer time periods and based on calibrations through historical events, are typically an order of magnitude smaller, ca. 10^{-4} s/s/y [20]. There is additional evidence that points to the overall stability of plant viruses over longer time periods based on reconstruction of archaeological virus genomes [21], integration of endogenous viral elements [22], and virus-host co-divergence [18]. However, these estimates are scarce and possibly obscured by contamination with recent isolates, recombination events, unresolved or poorly dated host phylogenies, and controversial or partial co-evolutionary scenarios. Yet they collectively point to much slower rates of ca. 10^{-4}-10^{-8} s/s/y, orders of magnitude slower than their short-term rate estimates and

closer to that of their plant hosts over longer timescales. Despite these independent lines of evidence in support of time-dependent changes in the evolutionary rate estimates, TDRP has been largely ignored in molecular dating studies of plant viruses. A prominent example includes a study of the reconstruction of the rice yellow mottle virus (RYMV) diversification [23], a virus of the *Sobemovirus* genus (solemoviridae family). This study showed that by extrapolating the short-term substitution rate estimate of RYMV (ca. 10^{-3} s/s/y) to much longer timeframes, the diversification of the solemoviridae stretches back to the Neolithic's agricultural expansion around 10,000 years ago. Other studies, taking a similar approach (i.e. not accounting for TDRP), have associated plant virus diversification to historical events and periods within the last centuries or few millennia, such as the Neolithic, crop domestication, Post-Columbian exchanges, and trade along the Silk Road but, crucially, never beyond (see e.g. [20,24]).

Understanding the evolutionary history of sobemoviruses is important as they include plant pathogens with high economic impact, of which RYMV is the most devastating [25]. The *Sobemovirus* genus includes 21 recognised species and 5 candidate species [26], some of which having limited geographical distribution (single continent or country) while others are found worldwide. Many species are transmitted by beetles (order Coleoptera) and a few by aphids (family Aphididae, order Hemiptera) or mirids (family Miridae, order Hemiptera). All sobemoviruses are readily propagated by mechanical wounds. Infections can be asymptomatic or cause severe disease dependent on the virus-host combination and environmental conditions. Symptoms mainly include mosaic and mottle of the infected leaves. While the host range of each virus species in this genus is narrow and confined to a few species of a single plant family, with the exception of sowbane mosaic virus, sobemoviruses collectively infect a wide range of monocot and dicot hosts mainly including plant host species in the Poaceae, Fabaceae, and Solanaceae families [27].

Icosahedral virions of sobemoviruses are 26–34nm in diameter and composed of 180 monomers of viral capsid protein (CP) on a T=3 lattice symmetry. Particles contain a single molecule of positive sense single-stranded RNA, about 4.0–4.5 kb in size. A subgenomic RNA (sgRNA) molecule, co-terminal with the 3′ end of the genomic RNA is synthesised in the virus-infected cells. Both genomic and subgenomic RNAs have a viral genome-linked protein (VPg) covalently bound to their 5′ end. The 3′ terminus is non-polyadenylated and does not contain a tRNA-like structure. Several sobemoviruses encapsidate a circular viroid-

like satellite RNA (220–390 nt). The genome comprises a polycistronic, positive-sense RNA molecule without a 3′-poly(A) tail. The genome 5′-end has a covalently attached VPg. The genome contains five open reading frames (ORFs). The 5′-proximal ORF1 encodes a non-conserved RNA silencing suppressor protein needed for systemic spread, followed by ORFx. Next ORF (ORF2a) encodes a polyprotein that is expressed by ribosomal leaky scanning and cleaved autocatalytically to different functional subunits (membrane anchor, serine protease, VPg and C-terminal domains). Expression of the viral RNA-dependent RNA polymerase (RdRP) as an alternative C-terminal domain of the polyprotein (from ORF2b) is regulated by a −1 programmed ribosomal frameshift. The 3′-proximal ORF3 of sobemoviruses, expressed from sgRNA, encodes CP [26].

Here, we revisited the evolutionary history of sobemoviruses, taking into account the impact of TDRP on estimating diversification events of sobemovirus ancestry using the PoW model. By providing the first short-term substitution rate estimate of cocksfoot mottle sobemovirus (CfMV), we showed that the rate estimates are largely similar between CfMV and RYMV. We then reconstructed the sobemovirus history using the PoW model and showed that they have a deep evolutionary history that stretches back to millions of years. In particular, we found that the CfMV/RYMV split occurred more than five hundred thousand years ago which falls within the Pleistocene epoch, much earlier than the Neolithic period initially predicted based on extrapolation of short-term rate estimates. We then investigated the sobemovirus radiation, the links between virus lineages and plant host families, and estimated speciation dates using sequences of the polymerase gene of 26 sobemovirus species and 42 additional sequences of the phytobiome defined as ‘plants, their environment, and their associated communities of organisms’ [28]. Finally, we discussed the implications of our results for the evolutionary history of plant viruses.

Results

Short-term substitution rates of CfMV and RYMV

We first assessed and confirmed the strength of temporal signal present in the CfMV dataset through the tip cluster randomisation test in root-to-tip regression (see **Methods**). We estimated the short-term substitution rate of the coat protein gene (ORF3) at 7.8×10^{-4} s/s/y (95% highest posterior density (HPD): $6.4 – 9.5\times10^{-4}$ s/s/y) for RYMV and at 5.0×10^{-4} s/s/y (95% HPD: $2.3 – 8.2\times10^{-4}$ s/s/y) for CfMV using the same substitution model in our

144 Bayesian inference (**Figure 1**). The wider variation in the CfMV posterior rate distribution
145 likely reflects the lower number of sequences and weaker clock signal compared to the
146 RYMV dataset. The inferred substitution rate of RYMV for the polymerase gene was 6.5 (5.2
147 – 7.9) $\times 10^{-4}$ s/s/y, similar to and slightly lower than that of the coat protein gene. The
148 similarity between the short-term substitution rates of two sobemovirus species supports the
149 choice of the RYMV substitution rate as the representative short-term rate to reconstruct the
150 evolutionary history of the sobemovirus genus using the PoW model.
151
152 Phylogeny and taxonomy of sobemovirus species
153 We found some degree of clustering between the phylogenetic tree of 26 sobemovirus species
154 (see **Methods**) and that of their respective plant family hosts (**Figure 2a**). The eight
155 sobemovirus species of monocots clustered in two distinct lineages, one consisting of five
156 virus species infecting host plants of the Poaceae family, including RYMV (referred to below
157 as the RYMV lineage), and another consisting of three virus species infecting plants of the
158 Poales order, including ryegrass mottle virus (RGMoV). The seven species infecting plants of
159 the Fabaceae family are split into two lineages, including five and two species. The three
160 species infecting plants of the Solanaceae family belong to a single lineage. The remaining
161 eight virus species infect plants belonging to different and unique families (see **Figure 2a**). In
162 addition to evidence of higher barriers to infection between plant families, we see similar
163 evidence at the within-family rank, as exemplified by the RYMV lineage. Collectively, the
164 species of the RYMV lineage infect plants of several genera in the Poaceae family, but
165 individually, they have a narrow host range that is sometimes constrained to one or a few
166 genera. The sister species CfMV and cynosurus mottle virus (CnMoV), which both originated
167 in Europe, have an overlapping host range, both infecting the temperate crops wheat, oat,
168 barley, and rye but not maize, sorghum or rice. RYMV infects plants belonging to the *Oryza*
169 genus whereas its sister species, Imperata yellow mottle virus (IYMV), found exclusively in
170 Africa, infects only maize, *Imperata cylindrica*, and *Rottboellia exaltata*.
171
172 The distribution of pairwise genetic identity in ORF2b across the sobemovirus species
173 revealed a major mode from 57 to 73% genetic identity, a minor mode from 77 to 84%, and a
174 "shoulder" from 91 to 100% (**Figure 2a**). This showed that nearly every pair of the
175 sobemovirus species have 57-73% genetic identity with the exception of the CnMoV-Poaceae
176 Liege sobemovirus (PLSV) pair and pairs among southern bean mosaic virus (SBMV),

177 sesbania mosaic virus (SeMV) and soybean yellow common mosaic virus (SYCMV) trio.
178 These groups, hereafter referred to as subspecies in the *Sobemovirus* genus, have intermediate
179 levels of genetic identity (77-84%). Finally, the shoulder reflects the intraspecific genetic
180 identity between pairs of isolates, ranging from 90.6 to 98.6%, with RYMV having a 92.1%
181 minimum pairwise intraspecific identity (**Figure 2a**; see also **Supplementary Table 2**).
182
183 Phylogeny and taxonomy of recognized and tentatively new sobemovirus species
184 We next examined a dataset of the 26 sobemovirus species isolated from plants (examined
185 above) merged with 42 additional known or tentatively novel sobemovirus sequences
186 collected from metagenomic analyses (hereafter referred to as metagenomic sequences; see
187 **Methods**) to establish their taxonomy and phylogenetic relationships. Most sequences were
188 obtained from gut metagenomes of herbaceous animals feeding mainly on Poaceae and
189 Fabaceae (**Table 2**). This explains why many isolates were related to sobemovirus species
190 found in Fabaceae and in Poaceae, particularly RGMoV (see **Figure 2b**). In contrast, none of
191 the metagenomic sequences clustered closely with sobemovirus species infecting the
192 Solanaceae (see **Figure 2b** and **Table 2**). We found that although the great majority of the
193 metagenomic sequences originated from China [29], they mostly cluster with isolates found
194 in other continents, reflecting the worldwide distribution of many sobemoviruses.
195 Nevertheless, there was some geographical association between the sequences. For instance,
196 the metagenomic sequences close to CfMV (MN626425 and MW588040) and CnMoV
197 (MW588173) were from Europe.
198
199 The phylogeny for the collection of 68 sequences differed in several respects from that of the
200 26 selected samples representing the recognized and candidate sobemovirus species that we
201 initially investigated. The overall genetic diversity of the merged dataset was noticeably
202 higher as new subclades possibly representing the novel species were revealed, some of
203 which have long independent evolutionary histories (see **Figure 2b**). There were also several
204 sequences that are closely related to known sobemovirus species such as RGMoV. The
205 RGMoV lineage includes strains of RGMoV and related species, likely at the subspecies
206 level. Our analysis of the pairwise genetic identity of the merged dataset revealed a
207 continuous change in pairwise identity of samples from interspecific to intraspecific levels.
208 This contrasts with the earlier findings in the 26 selected samples of sobemovirus species
209 where certain percentages of genetic identity had no corresponding pairs (see **Figure 2a**).

210 This discontinuity is no longer apparent in the merged dataset, confirming the broadened
211 sobemovirus genetic diversity (**Figure 2b**).
212
213 Deep evolutionary history of sobemoviruses

214 We first constructed a time-scaled phylogenetic history for the 26 sobemoviruses using the
215 PoW model (see **Methods**). This analysis dated the most recent common ancestor (TMRCA)
216 of known sobemoviruses back to 3.4 (95% HPD: 1.8 - 5.4) million years before present (BP)
217 (**Figure 3a**), much earlier than previous estimates based on extrapolation of short-term
218 substitution rate of RYMV [23]. The deep evolutionary history of sobemoviruses may
219 indicate a long-term association of these viruses with their host, with host jumps that
220 occurred over long evolutionary timescales. For instance, we estimated the TMRCA of virus
221 species infecting plants of the Poaceae, Fabaceae and Solanaceae at 546,000 (95% HPD:
222 260,000 - 1,490,000), 90,000 (30,000 - 186,000), and 299,000 (109,000 - 616,000) years BP,
223 respectively. Within the RYMV lineage, the RYMV/IYMV split was approximately 117,000
224 years BP while the CfMV/CnMoV split was close to 151,000 years BP. Similarly, the
225 TMRCA of most species was between several dozens of thousands to a few hundred
226 thousands of years BP, with turnip rosette virus (TRoV) and snake melon asteroid mosaic
227 virus (SMAMV) diverging for over a million years before sharing an ancestry with another
228 sobemovirus (**Figure 3a**). The notable exceptions came from the subspecies identified in our
229 taxonomy analysis. The SeMV/SYCMV and CnMoV/PLSV splits were estimated to
230 approximately 2,600 and 3,000 years BP, respectively, while the MRCA of the SBMV, SeMV,
231 and SYCMV existed approximately 9,000 years BP. Over shorter timescales, we found that
232 the TMRCA of the two most divergent RYMV isolates (RYMV-1 and -2 in **Figure 3a**) was
233 156 (95% HPD: 80 - 288) years BP, in agreement with previous estimates obtained from
234 standard Bayesian molecular clock analysis which do not correct for TDRP [30]. This finding
235 further suggests that saturation over such short timescales is unlikely to have a significant
236 impact on TMRCA estimates of sobemoviruses.
237

238 We next constructed the PoW-transformed time tree of the 68 sobemovirus sequences
239 (**Figure 3b**) and found that the root of the tree is 4.3 (95% HPD: 2.0 – 6.6) million years BP,
240 largely similar to the estimates obtained with the 26 sobemovirus dataset. Interestingly, we
241 see that many of the sequences of metagenomic origin share common ancestors with closely
242 related known sobemovirus species (i.e., they belong to the same species but may represent

new subspecies or strains) and diverged less than 10,000 years BP. This suggests a major radiation event during the Neolithic period. To investigate this, we constructed a lineages-through-time plot, which shows a sharp increase in the number of diversification events around 5,000 years ago (see **Figure 3b**). In particular, the TMRCA of the RGMoV lineage (consisting of strains of RGMoV and closely related tentative unrecognized species) was approximately 4,600 years BP. Other radiation events such as those between sobemovirus and metagenomic sequences (RoMoV/MZ218223 split 2,300 years BP) or between metagenomic sequences (MZ218230, MZ218235, and MZ218238 split 3,500 years BP) occurred within the last few thousand years. Several divergence events were estimated to have taken place less than 200 years ago, corresponding to pairwise genetic identity higher than 90% between sequences reflecting divergence among strains.

Discussion

Using an evolutionary model that accounts for time-dependent changes in the evolutionary rates over time, we were able to show that the *Sobemovirus* genus is a few million years old, making a case for ancient origins of plant viruses. Our analysis revealed that the species in this genus became increasingly specialised through time, with likely a few deep host jumps to members of the Poaceae, Fabaceae, and Solanaceae families that occurred tens to hundreds of thousand years ago. The association between the phylogeny of the sobemovirus species and their host range is consistent with the idea that the taxonomic rank of the host family sets a threshold for plant-virus host range, with a higher barrier to infection between plant families [31]. Crucially, our estimates for the evolutionary history of sobemoviruses are several orders of magnitude older than the Neolithic, a result that is in stark contrast with previous estimates that suggested divergence time between sobemoviruses and related viruses is only 9,000 years old [23]. However, we did find evidence of major diversification during the Neolithic, partly supporting the hypothesis that initial radiation of several plant virus groups, including the *Potyvirus* genus, took place at the dawn of agriculture [20,23].

While we found that TMRCA of RYMV using the PoW model is very similar to previous estimates using its short-term substitution rate, extrapolations of this rate to infer timescale of evolutionary events within the *Sobemovirus* genus resulted in a significant underestimation of the evolutionary history of sobemovirus species. The PoW analysis also revealed a clear association between divergence dates and taxonomic levels: sequences with a genetic identity

below 75% (species level) trace back to tens of thousands of years; those between 75% and 90% (subspecies level) emerged within the Neolithic era, thousand years ago. Meanwhile, sequences with over 90% identity (intraspecific level) indicated a recent divergence of one to two hundred years ago (strain level).

Thanks to high particle stability of sobemovirus virions in the environment [27], we were able to include the samples collected through metagenomic analysis of the phytobiome and investigated their evolutionary history along with known and candidate sobemovirus species. This analysis revealed an even wider diversity of sobemovirus species at the interspecific and intraspecific levels. Overall, our findings underscored the critical impact of TDRP in estimating long-term evolutionary history of viruses and offered a new perspective on the possible long-term virus-host associations within the *Sobemovirus* genus. This is exemplified by host range evolution, a pivotal aspect of plant virus emergence that has garnered significant attention in recent years (see [32,33] for recent reviews).

Our findings confirmed that the adaptation of RYMV to the African rice *Oryza glaberrima* occurred during the diversification of the virus in West-Africa, within the past two hundred years [34]. The association between the geographical distribution of virus species and their respective host plant families, such as the relationships observed between CfMV and CnMoV with temperate crops, or RYMV and IYMV with tropical crops, may reflect historical shifts in virus host preferences during the course of agricultural development, as has been postulated for potyviruses [35]. However, adaptation to wild species in the *Oryza* genus may have occurred much earlier, independently from human intervention with rice domestication or rice cultivation, considering that RYMV and IYMV diverged approximately 117,000 years ago. While our findings of the deep history of the *Sobemovirus* genus expands the range of timescales involved in the evolution of the plant virus host range by hundreds of thousand years, it precludes host-virus co-speciation scenarios. For instance, *Oryza sativa* and *Imperata cylindrica*, the respective hosts of RYMV and IYMV, diverged approximately 59 million years ago [36], much earlier than our estimate for the RYMV/IYMV split 117,000 years ago.

Expanding knowledge on the deep history of plant viruses is essential to uncovering the ways in which they have been influenced by the development of agriculture and other environmental perturbations, including climate change [37]. This may also have practical

implications for crop protection strategies and for ensuring the sustainability of agricultural systems in the face of evolving viral challenges.

Acknowledgement

In memoriam: We acknowledge Erkki Truve, who passed away a few years ago. His contribution to the research on sobemoviruses is invaluable and it is thanks to his work that this research continues in Estonia to this day. We remember his dedication and expertise, and his work remains an integral part of this article.

Data and code availability

All data and software code are available on our GitHub repository (https://github.com/mg878/sobemovirus_genus). The code for the PoW model is also available on GitHub (https://github.com/mg878/PoW_model).

Competing interest

None.

Funding

Estonian-French science and technology cooperation programme Parrot travel grant 2021-2022 for the project "Emergence and divergence of plant viruses on cereals: Sobemoviruses - a case study" was granted to EH and CS and contributed to this research.

Author contribution

Writing (original draft): M.G. and D.F.; analysis: M.G., E.H., P.L., and D.F.; study design: M.G. and D.F.; data collection: M.S., C.S., and A.N.; supervision: D.F. All authors reviewed and edited the manuscript.

Correspondence to Mahan Ghafari (mahan.ghafari@ndm.ox.ac.uk) or Denis Fargette (denis.Fargette@ird.fr).

References

1. Ho SYW, Phillips MJ, Cooper A, Drummond AJ. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol Biol Evol. 2005;22: 1561–1568.

2. Duchêne S, Holmes EC, Ho SYW. Analyses of evolutionary dynamics in viruses are hindered by a time-dependent bias in rate estimates. Proc Biol Sci. 2014;281: 20140732.

3. Aiewsakun P, Katzourakis A. Time dependency of foamy virus evolutionary rate estimates. BMC Evol Biol. 2015;15: 119.

4. Aiewsakun P, Katzourakis A. Time-dependent rate phenomenon in viruses. J Virol. 2016;90: 7184–7195.

5. Simmonds P, Aiewsakun P, Katzourakis A. Prisoners of war - host adaptation and its constraints on virus evolution. Nat Rev Microbiol. 2019;17: 321–328.

6. Wertheim JO, Kosakovsky Pond SL. Purifying selection can obscure the ancient age of viral lineages. Mol Biol Evol. 2011;28: 3355–3365.

7. Ghafari M, du Plessis L, Raghwani J, Bhatt S, Xu B, Pybus OG, et al. Purifying selection determines the short-term time dependency of evolutionary rates in SARS-CoV-2 and pH1N1 influenza. Mol Biol Evol. 2022;39. doi:10.1093/molbev/msac009

8. Porter AF, Purcell DFJ, Howden BP, Duchene S. Evolutionary rate of SARS-CoV-2 increases during zoonotic infection of farmed mink. Virus Evol. 2023;9: vead002.

9. Ghafari M, Liu Q, Dhillon A, Katzourakis A, Weissman DB. Investigating the evolutionary origins of the first three SARS-CoV-2 variants of concern. Front Virol. 2022;2. doi:10.3389/fviro.2022.942555

10. Ghafari M, Simmonds P, Pybus OG, Katzourakis A. A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses. Curr Biol. 2021;31: 4689-4696.e5.

11. Ho SYW, editor. The molecular evolutionary clock. 1st ed. Cham, Switzerland: Springer Nature; 2021.

12. Forni D, Cagliani R, Pozzoli U, Mozzi A, Arrigoni F, De Gioia L, et al. Dating the emergence of human endemic coronaviruses. Viruses. 2022;14: 1095.

13. Pekar JE, Lytras S, Ghafari M, Magee AF, Parker E, Havens JL, et al. The recency and geographical origins of the bat viruses ancestral to SARS-CoV and SARS-CoV-2. bioRxivorg. 2023. doi:10.1101/2023.07.12.548617

14. Li YQ, Ghafari M, Holbrook AJ, Boonen I, Amor N, Catalano S, et al. The evolutionary history of hepaciviruses. bioRxivorg. 2023. doi:10.1101/2023.06.30.547218

15. Ghafari M, Simmonds P, Katzourakis A. Reconstructing the evolutionary history of human, simian, and prosimian immunodeficiency viruses. bioRxiv. 2023. doi:10.1101/2023.06.28.546833

16. Forni D, Molteni C, Cagliani R, Sironi M. Geographic structuring and divergence time frame of Monkeypox virus in the endemic region. J Infect Dis. 2023;227: 742–751.

17. Forni D, Molteni C, Cagliani R, Clerici M, Sironi M. Analysis of variola virus molecular evolution suggests an old origin of the virus consistent with historical records. Microb Genom. 2023;9. doi:10.1099/mgen.0.000932

18. Gibbs AJ, Fargette D, Garcia-Arenal F, Gibbs MJ. Time - the emerging dimension of plant virus studies. J Gen Virol. 2010;91: 13–22.

19. 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.

20. Gibbs AJ, Ohshima K, Phillips MJ, Gibbs MJ. The prehistory of potyviruses: their initial radiation was during the dawn of agriculture. PLoS One. 2008;3: e2523.

21. Smith O, Clapham A, Rose P, Liu Y, Wang J, Allaby RG. A complete ancient RNA genome: identification, reconstruction and evolutionary history of archaeological Barley Stripe Mosaic Virus. Sci Rep. 2014;4: 1–6.

22. Chiba S, Kondo H, Tani A, Saisho D, Sakamoto W, Kanematsu S, et al. Widespread endogenization of genome sequences of non-Retroviral RNA viruses into plant genomes. PLoS Pathog. 2011;7: e1002146.

23. Fargette D, Pinel-Galzi A, Sérémé D, Lacombe S, Hébrard E, Traoré O, et al. Diversification of rice yellow mottle virus and related viruses spans the history of agriculture from the neolithic to the present. PLoS Pathog. 2008;4: e1000125.

24. Pagán I. The diversity, evolution and epidemiology of plant viruses: A phylogenetic view. Infect Genet Evol. 2018;65: 187–199.

25. Savary S, Willocquet L, Pethybridge SJ, Esker P, McRoberts N, Nelson A. The global burden of pathogens and pests on major food crops. Nat Ecol Evol. 2019;3: 430–439.

26. Sõmera M, Fargette D, Hébrard E, Sarmiento C, Ictv Report Consortium. ICTV virus Taxonomy profile: Solemoviridae 2021. J Gen Virol. 2021;102. doi:10.1099/jgv.0.001707

27. Sõmera M, Sarmiento C, Truve E. Overview on sobemoviruses and a proposal for the creation of the family Sobemoviridae. Viruses. 2015;7: 3076–3115.

28. Roossinck MJ. Viruses in the phytobiome. Curr Opin Virol. 2019;37: 72–76.

29. Chen Y-M, Sadiq S, Tian J-H, Chen X, Lin X-D, Shen J-J, et al. RNA viromes from terrestrial sites across China expand environmental viral diversity. Nat Microbiol. 2022;7: 1312–1323.

30. Issaka S, Traoré O, Longué RDS, Pinel-Galzi A, Gill MS, Dellicour S, et al. Rivers and landscape ecology of a plant virus, Rice yellow mottle virus along the Niger Valley. Virus Evol. 2021;7. doi:10.1093/ve/veab072

31. Moury B, Fabre F, Hébrard E, Froissart R. Determinants of host species range in plant viruses. J Gen Virol. 2017;98: 862–873.

32. McLeish MJ, Fraile A, García-Arenal F. Evolution of plant–virus interactions: host range and virus emergence. Curr Opin Virol. 2019;34: 50–55.

33. Elena SF, García-Arenal F. Plant virus adaptation to new hosts: A multi-scale approach. Curr Top Microbiol Immunol. 2023;439: 167–196.

34. Poulicard N, Pinel-Galzi A, Traoré O, Vignols F, Ghesquière A, Konaté G, et al. Historical contingencies modulate the adaptability of Rice yellow mottle virus. PLoS Pathog. 2012;8: e1002482.

35. Moury B, Desbiez C. Host range evolution of potyviruses: A global phylogenetic analysis. Viruses. 2020;12: 111.

36. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A resource for timelines, timetrees, and divergence times. Mol Biol Evol. 2017;34: 1812–1819.

37. Malmstrom CM, Martin MD, Gagnevin L. Exploring the emergence and evolution of plant pathogenic microbes using historical and paleontological sources. Annu Rev Phytopathol. 2022;60: 187–209.

38. Murray GGR, Wang F, Harrison EM, Paterson GK, Mather AE, Harris SR, et al. The effect of genetic structure on molecular dating and tests for temporal signal. Methods Ecol Evol. 2016;7: 80–89.

39. Duchene S, Lemey P, Stadler T, Ho SYW, Duchene DA, Dhanasekaran V, et al. Bayesian evaluation of temporal signal in measurably evolving populations. Mol Biol Evol. 2020;37: 3363–3379.

40. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018;4: vey016.

41. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32: 1792–1797.

42. Bruen TC, Philippe H, Bryant D. A simple and robust statistical test for detecting the presence of recombination. Genetics. 2006;172: 2665–2681.

43. Bryant D, Moulton V. Neighbor-net: an agglomerative method for the construction of phylogenetic networks. Mol Biol Evol. 2004;21: 255–265.

44. Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 2010;27: 221–224.

45. Muhire BM, Varsani A, Martin DP. SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. PLoS One. 2014;9: e108277.

46. Simmonds P, Adams MJ, Benkő M, Breitbart M, Brister JR, Carstens EB, et al. Consensus statement: Virus taxonomy in the age of metagenomics. Nat Rev Microbiol. 2017;15: 161–168.

47. Bassi C, Guerriero P, Pierantoni M, Callegari E, Sabbioni S. Novel virus identification through metagenomics: A systematic review. Life (Basel). 2022;12: 2048.

48. Fontdevila Pareta N, Khalili M, Maachi A, Rivarez MPS, Rollin J, Salavert F, et al. Managing the deluge of newly discovered plant viruses and viroids: an optimized scientific and regulatory framework for their characterization and risk analysis. Front Microbiol. 2023;14: 1181562.

49. Revell LJ. phytools: an R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 2012;3: 217–223.

456 **Methods**

457 Short-term substitution rate estimation

458 The sequences of the coat protein gene of 51 isolates of cocksfoot yellow mottle virus
459 (CfMV) collected between 1978 and 2022 (44 years) throughout the world were downloaded
460 from sequence databanks or sequenced in this study (see **Supplementary Table 1**). As the
461 5'end of the sequences of some of the early isolates was not available, a truncated alignment
462 was made of 640 nt long sequences (instead of 765 nt). We evaluated the strength of the
463 temporal signal in the CfMV dataset using a tip cluster-randomisation test in root-to-tip
464 regression, in addition to conducting a Bayesian assessment of temporal signal within the
465 data [38,39]. Similarly, for rice yellow mottle virus (RYMV), we used a previously published
466 dataset of 261 isolates of the coat protein gene (720 nt) collected in West Africa between
467 1975 and 2018 (43 years). We found evidence of sufficient temporal signal for rate inference
468 in both CfMV and RYMV datasets under an HKY85+Gamma substitution model and a
469 constant population size coalescent prior in BEAST (v 1.10.4) [40]. The substitution rate of
470 the polymerase of RYMV was estimated from an alignment of 69 dated ORF2b sequences
471 (1515 nt) by calibrating the root with a normally distributed prior with a mean of 160 years
472 and a standard deviation of 20 years, derived from the analysis of the coat protein gene
473 sequences [30].
474

475 Phylogeny and taxonomy of the sobemovirus species

476 The sobemovirus dataset consists of 26 species, including 21 species recognized by ICTV
477 and five candidate species (see **Table 1**). The sequences were codon-aligned using MUSCLE
478 [41] and the absence of recombination events between species was confirmed with the Phi
479 test [42,43]. We used the ORF2b alignment (encoding the polymerase) as this is the most
480 conserved part of the genome at the genus level to construct the sobemovirus phylogeny
481 under a HKY substitution model with an optimised rate variation using PhyML implemented
482 in SeaView (v 5.05) [44]. We truncated the 3' end of the ORF2b which had poor coverage
483 across many samples to reach a final alignment length of 1140 nt. For 12 (out of 26)
484 sobemovirus species with at least two isolates available, we further assessed the interspecific
485 range of identity within the ORF2b region by taking the two most divergent isolates of each
486 species using the Sequence Demarcation Tool (SDT2) [45]. The number of isolates within
487 each species varied from 2 to 72 (see **Supplementary Table 2**).
488

489 Metagenomic analysis

490 On August 1, 2023, we retrieved a total of 42 sequences of putative sobemoviruses that were
491 collected as components of the phytobiome from gut, sediment, soil, and insect metagenomes
492 from NCBI (see **Table 2**), reflecting the high stability of the sobemoviruses. We referred to
493 them as "metagenomic" sequences since there was no information on the original host plants
494 they may infect, a common problem when dealing with metagenome data (see [46–48]). The
495 sequences have variable lengths, with several of them having near-complete genome lengths.
496 We then constructed an alignment of the ORF2b region for the 42 metagenomic and 26
497 sobemovirus sequences with plant-origin representing the known species (see above) for the
498 phylogenetic analysis. The distribution of pairwise genetic identity between all 68 sequences
499 was assessed using SDT2. Similar to the sobemovirus species dataset, the pool of 68
500 sequences were codon-aligned using MUSCLE, and the absence of recombination events
501 between species was confirmed with a Phi test.
502

503 Time tree reconstruction with PoW model

504 We used the Prisoner of War (PoW) model of virus evolution to account for the power-law
505 decay in the substitution rate estimates of RYMV over time in our estimation of the
506 evolutionary time scale of the *Sobemovirus* genus [10]. To achieve this, we first estimated the
507 short-term substitution rate of RYMV and CfMV using time-stamped datasets (see above).
508 We then constructed the ultrametric distance trees of the 26 sobemovirus species samples as
509 well as the merged dataset of 68 sequences that contained also the sequences of metagenomic
510 origin using the HKY substitution model under a strict clock in BEAST 1. The tree was then
511 transformed into a time tree using the PoW model. Finally, we made a lineages-through-time
512 plot for the full dataset (68 sequences) using the phytools library in R [49].

513 **Tables**

514

515 **Table 1: Virus name, family host, and accession numbers of the 26 sobemovirus species and**
516 **candidate species.**

Virus		Host	Accession number
Artemisia virus A	(ArtVA)	Asteraceae	NC017914
Blueberry shoestring virus	(BSSV)	Ericaceae	LC081344
Cocksfoot mottle virus	(CfMV)	Poaceae	DQ680848
Cymbidium chlorotic mosaic virus	(CyCMV)	Orchidaceae	LC381945
Cynosorus mottle virus*	(CnMoV)	Poaceae	OM323994
Imperata yellow mottle virus	(IYMV)	Poaceae	NC011536
Lucerne transient streak virus	(LTSV)	Fabaceae	NC001696
Mimosa mosaic virus*	(MimMV)	Fabaceae	OP456085
Papaya lethal yellowing virus	(PLYV)	Caricaceae	NC018449
Physalis rugose mosaic virus	(PhyRMV)	Solanaceae	MK681145
Pistacia sobemovirus*	(PisSV)	Anacardiaceae	MT334602
Poaceae Liege sobemovirus*	(PLSV)	Poaceae	ON137710
Rice yellow mottle virus**	(RYMV)	Poaceae	AJ608207
Rottboellia yellow mottle virus	(RoMoV)	Poaceae	KC577469
Ryegrass mottle virus	(RGMoV)	Poaceae	EF091714
Sesbania mosaic virus	(SeMV)	Fabaceae	NC002568
Snake melon asteroid mosaic virus	(SMAMV)	Cucurbitaceae	MT989351
Solanum nodiflorum mottle virus	(SNMoV)	Solanaceae	NC033706
Southern bean mosaic virus	(SBMV)	Fabaceae	AF055887
Southern cowpea mosaic virus	(SCPMV)	Fabaceae	NC001625
Sowbane mosaic virus	(SoMV)	Amaranthaceae	GQ845002
Soybean yellow common mosaic virus	(SYCMV)	Fabaceae	KX096577
Subterranean clover mottle virus	(SCMoV)	Fabaceae	AY376451
Turnip rosette virus	(TRoV)	Brassicaceae	NC004553
Velvet tobacco mottle virus	(VTMoV)	Solanaceae	NC014509
Xufa yellow dwarf virus*	(XYDV)	Cyperaceae	ON828429

517 *candidate species, proposed to ICTV to be recognized as novel species in 2023
518 **the accession number for two most divergent pairs of RYMV sequences selected for the PoW
519 analysis are MF989228 (RYMV-1) and MZ172959 (RYMV-2).
520

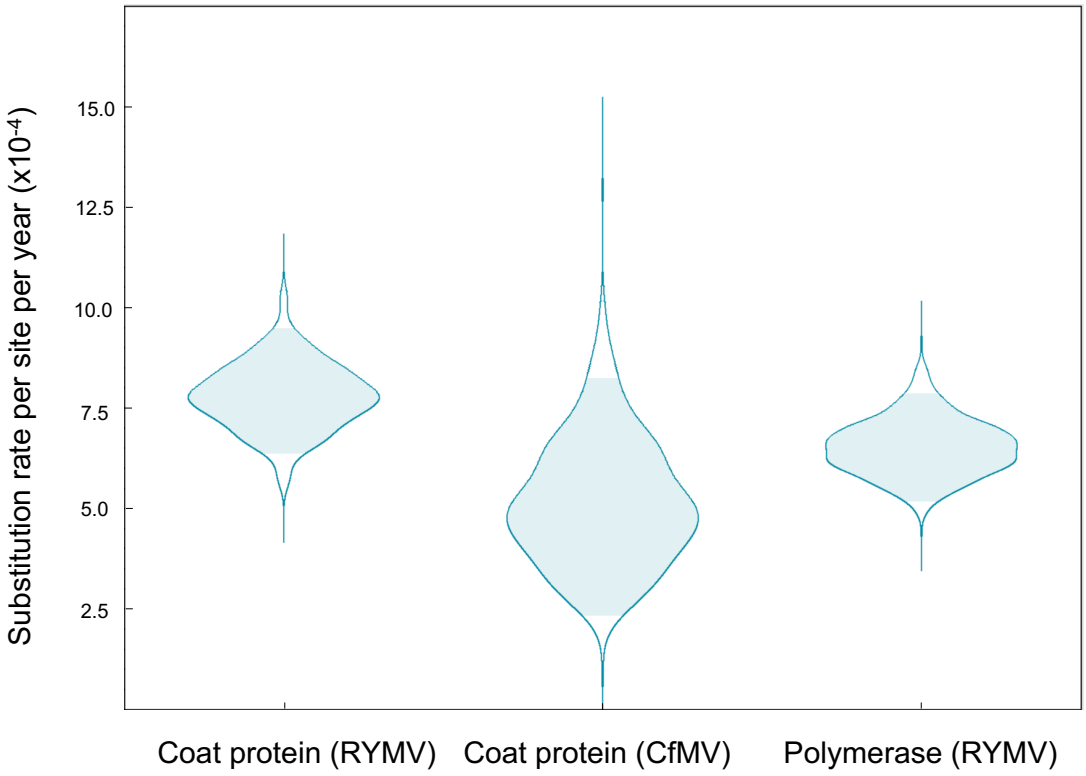
521 **Table 2: Isolation source and Genbank accession numbers of the 42 sequences obtained from**
522 **non-plant metagenomic analysis.**

Gut metagenome	**Accession number**
lizard	MZ375225
bird	MW826411
bird	MW826523
chicken	MZ218230
swan	MW588040
swan	MW588173
swan	MW588174
swan	MW588198
cattle	MZ218223
cattle	MZ218227
cattle	MZ218229
cattle	MZ218234
cattle	MZ218239
cattle	MZ218242
cattle	MZ218246
horse	MZ218231
pig	MZ218144
rabbit	MT129760
rabbit	MN626425
rat	MW826558
rat	MW826559
rat	MW826560
sheep	MZ218226
sheep	MZ218233
sheep	MZ218236
sheep	MZ218237
sheep	MZ218240
sheep	MZ218244
Homo sapiens	OK424596
Sediment metagenome	
semi-desert pond	MZ218228
fishpond	MZ218232
reservoir	MZ218238
sediment	MZ218243
river water	OM954019
river water	OM954025
river water	OM954036
irrigation water	MW411579
Insect metagenome	
soybean thrips	MW039356
soybean thrips	MT224146
Soil metagenome	
old-growth forest	MZ218241
old-growth forest	MZ218245
forest	MZ218235

523

524 **Figures**

525



526

527 **Figure 1: Posterior substitution rates (number of substitutions per site per year) of the rice**
528 **yellow mottle virus (RYMV) and cocksfoot yellow mottle virus (CfMV) datasets.** The RYMV
529 dataset has a stronger clock signal and lower variance compared to the CfMV dataset. The shaded
530 area in light blue shows the 95% highest posterior density region.

531

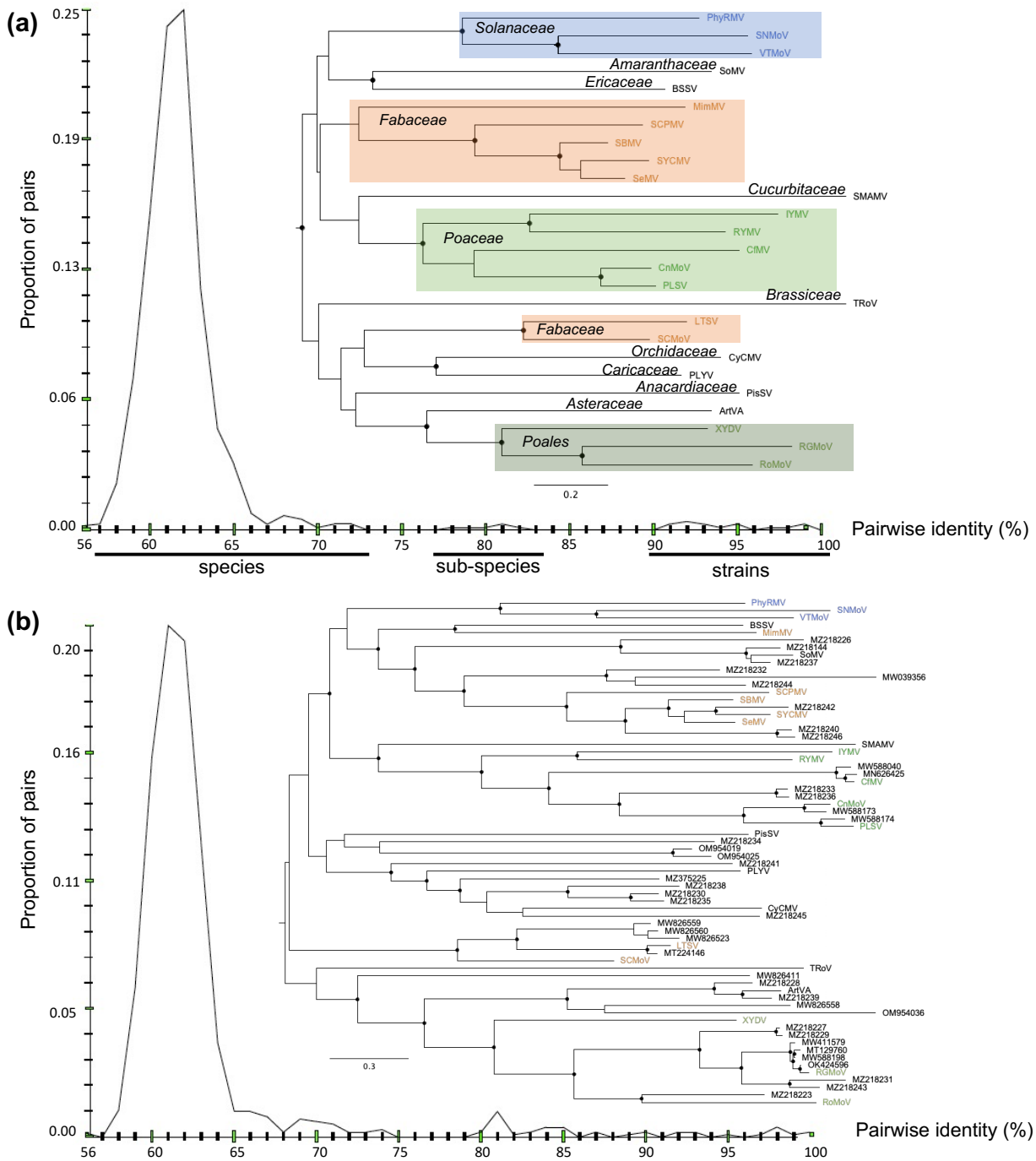
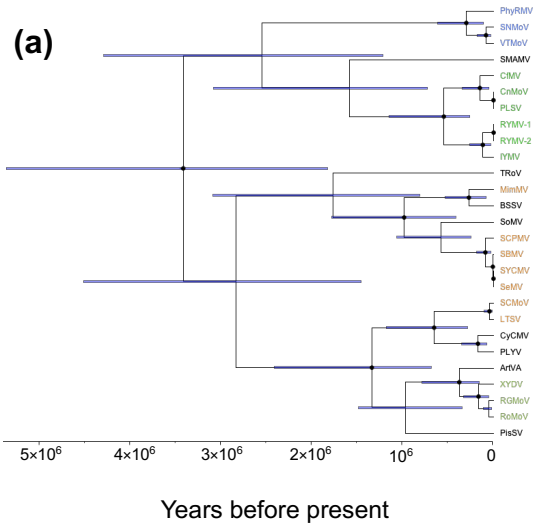
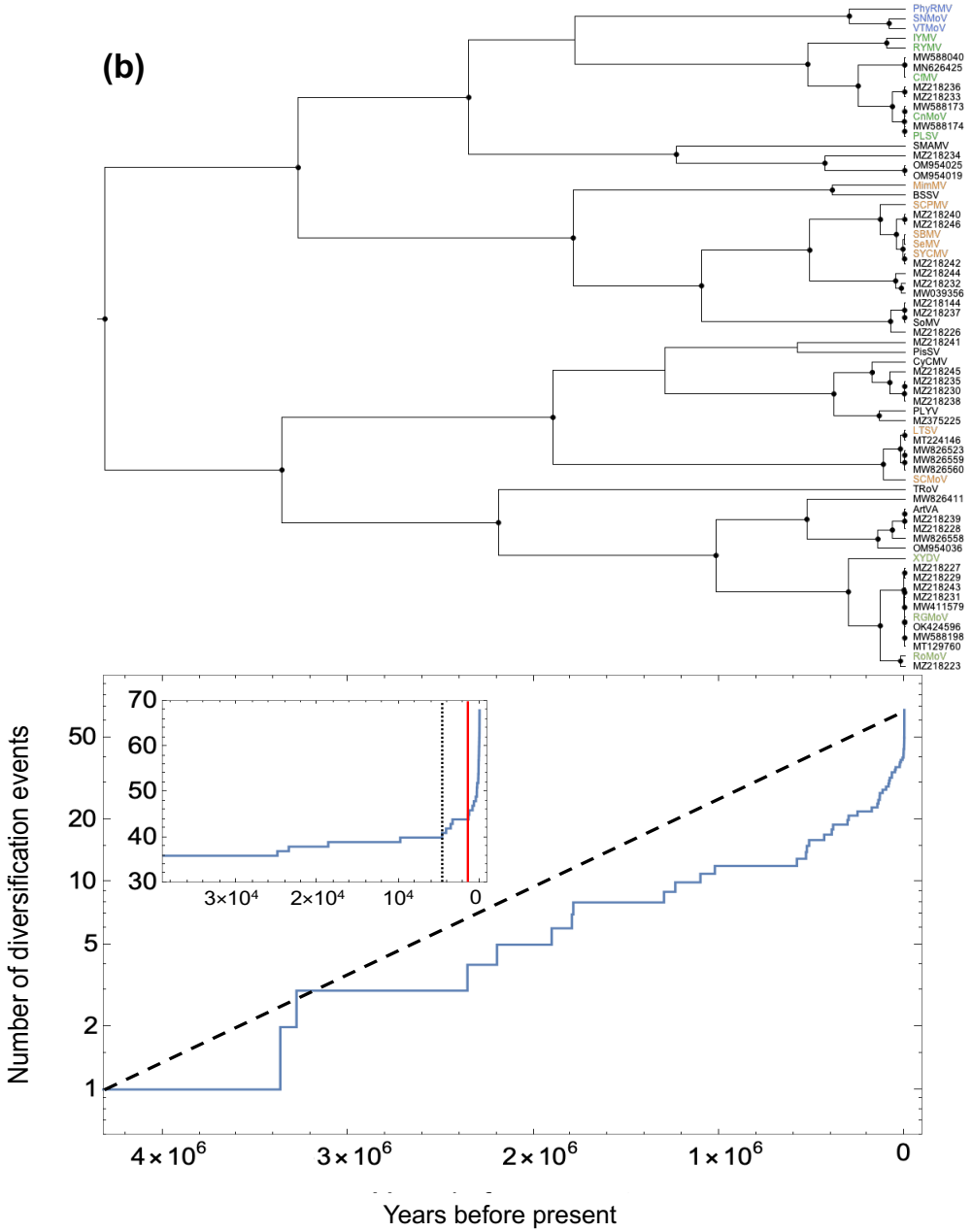


Figure 2: Phylogeny and pairwise distribution of genetic identity of known and tentative sobemovirus sequences identified in metagenomic analysis of the ORF2b (polymerase) region. **(a)** Phylogeny and pairwise identity of the 26 sobemovirus species. The host plant family of each species is highlighted on the tree. For 12 (out of 26) species with at least two isolates (see **Supplementary Table 2**), the two most divergent isolates were compared for pairwise genetic identity. The species, subspecies, and strain tentative taxonomical levels are indicated at the bottom of the figure. **(b)** Phylogeny and pairwise identity of the 26 sequences representing recognized sobemovirus species and 42 sequences of tentative sobemovirus isolates originating from metagenomic studies. The virus species infecting plants belonging to the Solanaceae, Fabaceae, Poaceae families, or the Poales order are indicated in blue, orange, light green, and dark green, respectively. The accession numbers for the 42 additional sequences are shown at the tips of the tree. Black circles show internal nodes with at least 90% bootstrap supports.


(a)
PhyRMV
SNMoV
VTMoV
SMAMV
ClMV
CnMoV
PLSV
RYMV-1
RYMV-2
IYMV
TRoV
MimMV
BSSV
SoMV
SCPMV
SBMV
SYCMV
SeMV
SCMoV
LTSV
CyCMV
PLYV
ArtVA
XYDV
RGMoV
RoMoV
PisSV
5×10^6
4×10^6
3×10^6
2×10^6
10^6
0
Years before present

547


(b)
PhyRMV
SNMoV
VTMoV
IYMV
RYMV
MW588040
MN626425
ClMV
MZ218236
MZ218233
MW588173
CnMoV
MW588174
PLSV
SMAMV
MZ218234
OM954025
OM954019
MimMV
BSSV
SCPMV
MZ218240
MZ218246
SBMV
SeMV
SYCMV
MZ218242
MZ218244
MZ218232
MW039356
MZ218144
MZ218237
SoMV
MZ218226
MZ218241
PisSV
CyCMV
MZ218245
MZ218235
MZ218230
MZ218238
PLYV
MZ375225
LTSV
MT224146
MW826523
MW826559
MW826560
SCMoV
TRoV
MW826411
ArtVA
MZ218239
MZ218228
MW826558
OM954036
XYDV
MZ218227
MZ218229
MZ218243
MZ218231
MW411579
RGMoV
OK424596
MW588198
MT129760
RoMoV
MZ218223
Number of diversification events
70
60
50
40
30
3×10^4
2×10^4
10^4
0
50
20
10
5
2
1
4×10^6
3×10^6
2×10^6
1×10^6
0
Years before present

548

549 **Figure 3: PoW-transformed time tree of the sobemovirus species and number of diversification**
550 **events through time. (a)** PoW-transformed time tree of 26 sobemovirus sequences. RYMV is
551 represented by the sequences of its two most divergent isolates MF989228 (RYMV-1) and MZ172959
552 (RYMV-2). The horizontal blue bars show 95% highest posterior density of the node age. **(b)** PoW-
553 transformed time tree of 68 known and tentative sobemovirus sequences (top panel) and their
554 respective number of diversification events through time (bottom panel). Black dashed line shows a
555 constant rate of speciation event over time with no extinction. The inset shows the number of
556 diversification events over the last 40,000 years. Black vertical dotted line indicates 4,600 years ago, a
557 point beyond which nearly 40% of diversifications occurred until present. Red vertical line indicates
558 2,000 years ago, a point on the graph with the sharpest increase in the number of diversification
559 events. The virus species infecting plants belonging to the Solanaceae, Fabaceae, Poaceae families, or
560 the Poales order are indicated in blue, orange, light green, and dark green, respectively. The accession
561 numbers for the 42 additional sequences are shown at the tips of the tree. Black circles show internal
562 nodes with at least 90% posterior supports.

563 **Supplementary Tables**

564

565 **Supplementary Table 1: Accession numbers and sampling years of the cocksfoot mottle virus**
566 **isolates.**

Accession number	Sample collection date
AB040447	1978
MAFF307042	1978
MAFF104073	1981
AB047688	1982
MAFF104072	1982
DQ680848	1985
NC002618	1985
MAFF307050	1988
MAFF307051	1988
MAFF307049	1989
FJ669143	1990
AJ243469	1995
AJ243470	1995
AJ243471	1995
AJ243472	1995
AJ243473	1995
L40905	1995
GU002336	2001
EF422382	2005
EF422383	2005
EF422384	2005
EF422385	2005
EF422386	2005
EF422387	2005
EF422388	2005
EF422389	2005
EF422390	2005
EF422391	2005
EF422392	2005
EF422393	2005
EF422394	2005
EF422395	2005
EF422396	2005
OR509655	2008
MW147174	2014
KT984653	2015
KT984654	2015
KT984655	2015
KT984656	2015
KT984657	2015
KT984658	2015
KT984659	2015
KT98466	2015
KX880413	2016
MF621330	2016
MW588040	2016
MG765457	2017

MG765458	2017
MG765459	2017
MG765460	2017
OQ974933	2022

568 **Supplementary table 2: Percentage of sequence identity in ORF2b between the most divergent**
569 **pair of isolates within each species.**

Virus name	Number of isolates	Min. genetic identity (%)	Accession numbers of the most divergent pair of isolates	
Cocksfoot mottle virus	7	95.0	L40905	MW197136
Cymbidium chlorotic mosaic virus	6	90.6	KR996515	ON286969
Lucerne transient streak virus	4	98.0	NC001696	OK181163
Physalis rugose mosaic virus	2	97.4	MK681145	MK681145
Pistacia sobemovirus	4	93.0	MT334603	MW548580
Rice yellow mottle virus	72	92.1	MF989228	MZ172959
Ryegrass mottle virus	7	98.6	MT129760	MW411579
Southern bean mosaic virus	20	90.8	AF055888	OK670700
Southern cowpea mosaic virus	2	91.8	MZ355623	NC001625
Sowbane mosaic virus	6	95.0	GQ845002	OK558770
Subterranean clover mottle virus	10	98.6	OM818392	OM818396
Turnip rosette virus	4	91.9	KC778720	KC778721

570