UNIVERSIDADE FEDERAL DE UBERLÂNDIA
INSTITUTO DE CIÊNCIAS AGRÁRIAS
PROGRAMA DE PÓS-GRADUAÇÃO EM AGRONOMIA

LILIAN GUIMARAES VERDOLIN

THE PHYLOGEOGRAPHY OF BEGOMOVIRUSES: MAPPING INFORMATIVE REGIONS IN VIRAL GENOMES

UBERLÂNDIA – MINAS GERAIS
2023

LILIAN GUIMARAES VERDOLIN

THE PHYLOGEOGRAPHY OF BEGOMOVIRUSES: MAPPING INFORMATIVE REGIONS IN VIRAL GENOMES

Dissertação apresentado à Universidade Federal de Uberlândia como requisito parcial para obtenção do título de mestre em Agronomia.
Área de concentração: Produção Vegetal

Orientador: Prof. Dr. Alison Talis Martins Lima

UBERLÂNDIA – MINAS GERAIS
2023

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da UFU, MG, Brasil.

V486p
2023

Verdolin, Lilian Guimarães, 1997-
The phylogeography of begomoviruses [recurso eletrônico] : mapping informative regions in viral genomes / Lilian Guimarães Verdolin. - 2023.

Orientador: Alison Talis Martins Lima.
Dissertação (Mestrado) - Universidade Federal de Uberlândia, Programa de Pós-graduação em Agronomia.
Modo de acesso: Internet.
Disponível em: http://doi.org/10.14393/ufu.di.2024.5053
Inclui bibliografia.
Inclui ilustrações.

1. Agronomia. I. Lima, Alison Talis Martins, 1982-, (Orient.). II. Universidade Federal de Uberlândia. Programa de Pós-graduação em Agronomia. III. Título.

CDU: 631

André Carlos Francisco
Bibliotecário Documentalista - CRB-6/3408

LILIAN GUIMARAES VERDOLIN

THE PHYLOGEOGRAPHY OF BEGOMOVIRUSES: MAPPING INFORMATIVE REGIONS IN VIRAL GENOMES

Dissertação apresentada à Universidade Federal de Uberlândia como requisito parcial para obtenção do título mestre em Agronomia

Área de concentração: Produção Vegetal

Uberlândia, 9 de outubro de 2023

Banca Examinadora:

_____

Prof. Dr. Alison Talis Martins Lima – UFU (Orientador)

_____

Profª. Drª. Nilvanira Donizete Tebaldi - UFU

_____

Prof. Dr. Francisco Murilo Zerbini - UFV

_____

Dr. Fernando Lucas de Melo

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
Secretaria da Coordenação do Programa de Pós-Graduação em Agronomia
Rodovia BR 050, Km 78, Bloco 1CCG, Sala 206 - Bairro Glória, Uberlândia-MG, CEP 38400-902
Telefone: (34) 2512-6715/6716 - www.ppgagro.iciag.ufu.br - posagro@ufu.br

ATA DE DEFESA - PÓS-GRADUAÇÃO

Programa de Pós-Graduação em:	Agronomia				
Defesa de:	Dissertação de Mestrado Acadêmico, 012/2023, PPGAGRO				
Data:	Nove de outubro de dois mil e vinte e três	Hora de início:	09:00	Hora de encerramento:	12:30
Matrícula do Discente:	12122AGR006				
Nome do Discente:	Lilian Guimarães Verdolin				
Título do Trabalho:	THE PHYLOGEOGRAPHY OF BEGOMOVIRUSES: MAPPING INFORMATIVE REGIONS IN VIRAL GENOMES				
Área de concentração:	Produção Vegetal				
Linha de pesquisa:	Controle e Manejo Integrado de Doenças				

Reuniu-se por videoconferência, a Banca Examinadora, designada pelo Colegiado do Programa de Pós-graduação em Agronomia, assim composta: Professores Doutores: Nilvanira Donizete Tebaldi - UFU; Francisco Murilo Zerbini - Universidade Federal de Viçosa; Fernando Lucas de Melo - Onsite Genomics; Alison Talis Martins Lima - UFU orientador(a) do(a) candidato(a).

Iniciando os trabalhos o(a) presidente da mesa, Dr. Alison Talis Martins Lima, apresentou a Comissão Examinadora e o candidato(a), agradeceu a presença do público, e concedeu ao Discente a palavra para a exposição do seu trabalho. A duração da apresentação do Discente e o tempo de arguição e resposta foram conforme as normas do Programa.

A seguir o senhor(a) presidente concedeu a palavra, pela ordem sucessivamente, aos(às) examinadores(as), que passaram a arguir o(a) candidato(a). Ultimada a arguição, que se desenvolveu dentro dos termos regimentais, a Banca, em sessão secreta, atribuiu o resultado final, considerando o(a) candidato(a):

Aprovada.

Esta defesa faz parte dos requisitos necessários à obtenção do título de Mestre.

O competente diploma será expedido após cumprimento dos demais requisitos, conforme as normas do Programa, a legislação pertinente e a regulamentação interna da UFU.

Nada mais havendo a tratar foram encerrados os trabalhos. Foi lavrada a presente ata que após lida e achada conforme foi assinada pela Banca Examinadora.

Documento assinado eletronicamente por **Nilvanira Donizete Tebaldi**, **Professor(a) do Magistério Superior**, em 09/10/2023, às 12:57, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Alison Talis Martins Lima**, **Professor(a) do Magistério Superior**, em 09/10/2023, às 13:24, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Francisco Murilo Zerbini Junior**, **Usuário Externo**, em 09/10/2023, às 15:07, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Fernando Lucas de Melo**, **Usuário Externo**, em 14/11/2023, às 10:34, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site https://www.sei.ufu.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **4845794** e o código CRC **8D2976BB**.

Agradecimentos

Aos meus pais e minha irmã, por todo o carinho, apoio, incentivo e por sempre acreditarem em mim.

Ao Professor Alison, pela amizade, paciência, orientação e ensinamentos proporcionados durante o desenvolvimento deste trabalho.

À Universidade Federal de Uberlândia, pela oportunidade oferecida.

A todos os meus amigos, pelo incentivo, apoio, torcida e momentos de descontração.

A todos do grupo de pesquisa PROVIRUS, pela ajuda e contribuição na realização deste trabalho.

A Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) pela concessão da bolsa de estudos.

Content

Abstract

Begomoviruses are notorious for their ability to infect a wide range of dicotyledonous plant species. These viruses are transmitted by a complex of cryptic whitefly species, collectively known as *Bemisia tabaci*. Begomovirus genomes consist of either one (monopartite) or two (bipartite) single-stranded DNA molecules, and their rapid evolution is primarily driven by mechanisms such as mutation, recombination, and pseudorecombination. The dynamic interplay among these mechanisms significantly contributes to their high genetic variability and capacity to swiftly adapt to new host species. Previous studies on microevolution of begomoviruses have consistently revealed the geographical segregation of their populations, suggesting limited gene flow across distinct regions. While numerous investigations have focused into phylogenetic and population genetic analyses of begomovirus genomes from various continents, sub-regions, or countries, there remains a noticeable research gap concerning the evidence of isolation by distance. This study aimed to investigate the evidence of isolation by distance using a multivariate Procrustean approach applied to datasets containing full-length DNA-A (or DNA-A-like) sequences of begomoviruses. Additionally, a sliding window approach was employed to perform a fine-scale mapping of the geographical signal, utilizing 200-nucleotide segments derived from the segmentation of full-length DNA-A sequences. To achieve this objective, a detailed curation of spatial data associated with each DNA-A sequence was conducted, drawing from GenBank records and related scientific publications. Subsequently, an analysis of genetic divergence among begomovirus isolates was carried out by calculating patristic distances derived from maximum likelihood trees. An extensive correlation analysis between distance matrices, encompassing both spatial and genetic information, was performed using the Procrustean Approach to Cophylogeny (PACo). The study yielded robust evidence of isolation by distance in at least three begomovirus species datasets, comprising sequences from isolates of bean golden mosaic virus, cotton leaf curl Gezira virus and tomato yellow leaf curl virus. Furthermore, the results unequivocally underscored the uneven distribution of the geographical signal across genomes. While population segregation across different geographic regions was discernible in various genomic regions, evidence of isolation by distance tended to be more pronounced in localized segments, often interspersed with regions lacking any isolation by distance signal. Additionally, this study shed light on how recombination-induced variation can obscure evidence of isolation by distance, even in datasets containing a limited number of recombinant DNA-A sequences. Finally, we concluded that recent begomovirus incursions into distant regions from their original sites of origin also contributed to the reduced global congruence between spatial and genetic data.

Keywords: Begomovirus, Phylogeny, Evolution

Resumo

Os begomovírus são notórios pela sua capacidade de infectar uma ampla gama de espécies de plantas dicotiledôneas. Esses vírus são transmitidos por um complexo de espécies crípticas de mosca branca, conhecidas coletivamente como *Bemisia tabaci*. Os genomas do begomovírus consistem em uma (monopartidos) ou duas (bipartidos) moléculas de DNA de fita simples, e sua rápida evolução é impulsionada principalmente por mecanismos como mutação, recombinação e pseudorecombinação. A interação dinâmica entre estes mecanismos contribui significativamente para a sua elevada variabilidade genética e capacidade de adaptação rápida a novas espécies hospedeiras. Estudos anteriores sobre a microevolução de begomovírus revelaram consistentemente a segregação geográfica de suas populações, sugerindo fluxo gênico limitado em regiões distintas. Embora numerosas investigações tenham se concentrado em análises filogenéticas e genéticas populacionais de genomas de begomovírus de vários continentes, sub-regiões ou países, permanece uma lacuna notável na pesquisa relativa à evidência de isolamento por distância. Este estudo teve como objetivo investigar a evidência de isolamento por distância usando uma abordagem de Procrustes multivariada aplicada a conjuntos de dados contendo sequências completas de DNA-A (ou semelhantes a DNA-A) de begomovírus. Além disso, uma abordagem de janela móvel foi empregada para realizar um mapeamento em escala fina do sinal geográfico, utilizando segmentos de 200 nucleotídeos derivados da segmentação de sequências completas de DNA-A. Para atingir este objetivo, foi realizada uma curadoria detalhada de dados espaciais associados a cada sequência de DNA-A, com base em registros do GenBank e publicações científicas relacionadas. Posteriormente, foi realizada uma análise de divergência genética entre isolados de begomovírus, através do cálculo de distâncias patrísticas derivadas de árvores de máxima verossimilhança. Uma extensa análise de correlação entre matrizes de distância, abrangendo informações espaciais e genéticas, foi realizada utilizando a Abordagem de Procrustes para Cofilogenia (PACo). O estudo produziu evidências robustas de isolamento por distância em pelo menos três conjuntos de dados de espécies de begomovírus, compreendendo sequências de isolados de bean golden mosaic virus, cotton leaf curl Gezira virus e tomato yellow leaf curl virus. Além disso, os resultados sublinharam inequivocamente a distribuição desigual do sinal geográfico entre os genomas. Embora a segregação populacional em diferentes regiões geográficas fosse discernível em várias regiões genômicas, a evidência de isolamento por distância tendia a ser mais pronunciada em segmentos localizados, muitas vezes intercalados com regiões sem qualquer isolamento por sinal de distância. Além disso, este estudo esclarece como a variação induzida pela recombinação pode obscurecer as evidências de isolamento por distância, mesmo em conjuntos de dados contendo um número limitado de sequências de DNA-A recombinante. Finalmente, concluímos que as recentes incursões de begomovírus em regiões distantes dos seus locais de origem também contribuíram para a redução da congruência global entre dados espaciais e genéticos.

Palavras-chave: Begomovirus, Filogenia, Evolução

General introduction

The genus *Begomovirus* stands as the most sizeable within the family *Geminiviridae*, encompassing 445 distinct species, as recognized by the International Committee on Taxonomy of Viruses (ICTV) [1]. Begomoviruses target a multitude of dicotyledonous plants and are transmitted by whiteflies belonging to the cryptic species complex referred to as *Bemisia tabaci* [2]. Their infections induce severe symptoms, including mosaic patterns, mottling, yellowing, leaf curling, and dwarfism. Such symptoms can lead to significant yield reductions or even complete crop losses in relevant crops worldwide. Recent research suggests that begomoviruses might also infect monocotyledonous plants [3]. Notably, economically and socially significant diseases, such as cassava mosaic disease in Africa, can be attributed to begomoviruses, resulting in devastating consequences for cassava fields [4]. Another relevant begomovirus is the tomato yellow leaf curl virus, a widespread pathogen that affects tomato crops in various countries across temperate and subtropical regions [5, 6].

Begomoviruses possess genomes composed of one or two circular, single-stranded DNA (ssDNA) molecules known as DNA-A and DNA-B components, each approximately 2600 nucleotides long. The DNA-A component in bipartite begomoviruses encodes essential proteins responsible for replication [7], suppression of gene silencing [8] and encapsidation of the viral progeny [9]. Conversely, DNA-B encodes proteins essential for the virus movement within host plants [10]. A successful systemic infection by a bipartite begomovirus requires the presence of both DNA components within the host plant [1]. In monopartite begomoviruses, the single genomic component closely resembles the DNA-A component found in bipartite counterparts and is referred to as the DNA-A-like component. Monopartite begomoviruses are frequently associated with DNA satellites, which can play a role in inducing disease symptoms [11].

Begomovirus populations exhibit high genetic variability due to high substitution rates [12], frequent occurrence of recombination [13] and pseudo-recombination or reassortment [14–16]. Studies on begomovirus microevolution reveal that viral isolates from distinct geographical locations tend to be genetically differentiated, possibly due to limited gene flow [17]. A relevant feature of their phylogeny is the segregation based on sampling location. Genetic differentiation between begomovirus populations from various sampling locations is evident across different geographical scales [17–20]. An illustrative example is that of euphorbia yellow mosaic virus, an indigenous weed-

infecting begomovirus found in *Euphorbia heterophylla* plants collected in various regions in Brazil [20], exhibited genetic segregation even between viral isolates sampled just about 210 km apart (between the municipalities of Cascavel-PR and Tacuru-MS), similar to the segregation observed between isolates from vastly distant locations (between Chapada-RS and Boqueirão-PE, approximately 2,900 km apart).

Despite the well-documented geographical segregation of begomovirus populations, it remains unclear whether substantial evidence supports isolation by distance. The study of geographic structuring, which assesses whether geographical distance contributes to population isolation, is essential for estimating the degree of population connectivity versus local confinement of individuals. This approach is essential in understanding the behavior of various organisms and is widely applied in biology, including for non-human and human infecting viruses, such as Rabies and Avian influenza viruses, both of which have exhibited isolation by distance [21, 22]. Isolation by distance is the theoretical basis of numerous epidemiological models aimed at evaluating and quantifying population migration dynamics. Recent applications of this concept include studies related to the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus responsible for the COVID-19 pandemic [23, 24]. An in-depth understanding of population structure within a spatio-temporal framework is essential for developing effective management strategies for any pathogen.

In the study conducted by Rocha et al. [19] which involved tomato-infecting begomovirus isolates from various locations in Brazil, analyses to assess the genetic structure and geographic segregation were performed. However, the study did not investigate the evidence of isolation by distance, despite having samples collected at both relatively distant locations (approximately 790 km apart between Paty do Alferes-RJ and Jaíba-MG) and relatively close locations (Florestal-MG and Carandaí-MG, approximately 136 km apart). Therefore, it would be valuable to determine whether there is any evidence of isolation by distance between these populations. Similarly, several other studies involving samples from plants collected in wide geographic ranges observed population segregation but did not explore the existence of isolation by distance [28–32]. An exception is the above-mentioned study involving EuYMV isolates, in which the Mantel's test was employed to investigate the evidence of isolation by distance [20]. In this context, most studies investigating the genetic structure of begomovirus populations have primarily focused on demonstrating geographic segregation. Consequently, there is

a pressing need for more comprehensive population genetics studies, including analyses to assess the existence of geographic isolation.

Review

The geminiviruses

Geminiviruses (family *Geminiviridae*) are among the most destructive plant viruses, causing diseases in major crops worldwide [30–32]. They are transmitted by insects, including various species of leafhoppers, treehoppers, and whiteflies from the cryptic species complex known as *Bemisia tabaci*. Geminiviruses are characterized by their unique twinned icosahedral particle morphology and possess single-stranded circular DNA (ssDNA) genomes ranging from 2500 to 3000 nucleotides in length [33–35]. These viruses can infect both monocotyledonous and dicotyledonous plants, with symptoms varying from mild or asymptomatic infections to severe manifestations such as leaf wrinkling, curling, yellowing, distortion, dwarfing, mosaic patterns, or streaking [36, 37].

The family *Geminiviridae* encompasses approximately 520 species, as listed on the ICTV (International Committee on Taxonomy of Viruses) webpage (https://ictv.global/report/chapter/geminiviridae/geminiviridae). These viruses are further classified into 14 genera, including *Becurtovirus*, *Begomovirus*, *Capulavirus, Citlodavirus*, *Curtovirus*, *Eragrovirus*, *Gablovirus*, *Maldovirus*, *Mastrevirus*, *Mulcrilevirus*, *Opunvirus*, *Topilevirus*, *Topocuvirus*, *Turncurtovirus*. This classification is based on genome organization, host range, phylogenetic relationships, and the specific insect vectors [38, 39]. The development of insecticide resistance and the emergence of new vector biotypes, in particular whiteflies, allowed geminiviruses to invade new geographical regions and assemble new combinations of viruses into disease complexes. These properties allowed such viruses to quickly adapt to other hosts and environments [40]. This has led to a global spread of geminiviruses, posing a major threat to food security in agricultural producing countries [41].

Geminiviruses replicate their compact genomes through double-stranded (ds) DNA intermediates within the nuclei of infected plant cells, employing a rolling circle mechanism [42, 43]. These characteristics set geminiviruses apart from the majority of plant viruses, which typically possess RNA genomes and/or replication intermediates. Geminiviruses encode a limited number of proteins for replication and rely on the host's

DNA replication machinery [44]. The viral particles are introduced into the plant through insect vectors, and upon decapsidation, the viral genetic material is transported to the nucleus. Inside the nucleus, the viral genome needs to be converted from single-stranded DNA (ssDNA) to a double-stranded (ds) DNA intermediate. This conversion is carried out by host-encoded DNA polymerases, which has been identified as DNA polymerases α and δ [45, 46].

The dsDNA replicative intermediate is identified by the virus-encoded replication-associated protein (Rep). Rep binds to a specific sequence characterized by a loop-like structure that contains an invariant nonanucleotide sequence (TAATATT//AC) located within the intergenic region [47, 48]. Once bound, Rep recruits the cellular DNA replication machinery and initiates a strand cleavage, which marks the beginning of rolling circle replication. After the host's DNA polymerases completes several rounds of replication, the Rep protein reconnects the displaced strand, resulting in the release of a new copy of the viral single-stranded DNA (ssDNA) genome [49]. This ssDNA genome can serve as a template for additional rounds of replication, spreading to adjacent cells or being promptly packaged into virions for acquisition by the vector insect [50, 51]. The multifunctional Rep protein is the only virus-encoded protein essential for the replicative cycle. It acts as a helicase, possesses DNA cutting and binding activities, and plays a role in reprogramming the cell cycle to induce the expression of DNA-dependent DNA polymerase [50, 52] .

Until recently, no DNA polymerase activity associated with this viral protein had been identified, despite its essential role in geminivirus replication. Recent studies have demonstrated the DNA polymerases α and δ as essential for the replication of geminiviruses within their host plants. Specifically, polymerase α is responsible for synthesizing the complementary viral strand, while DNA polymerase δ facilitates the production of new copies of the geminiviral single-stranded DNA genome. The involvement of these replicative DNA polymerases aligns with previous findings that treatment with aphidicolin, an inhibitor of DNA polymerases α, δ, and ε, hampers the accumulation of geminiviruses in plants [46]. Interestingly, geminiviruses also utilize an alternative replication mechanism known as recombination-dependent replication (RDR) [53–57].

Within the plant, the infection spreads through the movement of viral DNA out of the nucleus into neighboring cells and into the phloem, facilitated by two viral movement

proteins: NSP (nuclear shuttle protein) and MP (movement protein) [10, 58, 59]. Bipartite begomoviruses require both genomic components (DNA-A and DNA-B) to effectively infect their host and induce systemic symptoms [60].

The begomoviruses

The genus *Begomovirus* stands as the most extensive and diverse group within the family *Geminiviridae*, encompassing a total of 445 distinct species, as of the time of writing this review. Begomoviruses can be categorized into two primary groups, characterized by their geographic distribution and phylogenetic relationships. The first group is native to the 'Old World,' encompassing regions such as Europe, Africa, Asia, and Oceania. The second group includes viruses native from the 'New World' (Americas) [61, 62].

Further classification of begomoviruses comprises that into monopartite genomes, comprising a single-stranded (ss)DNA molecule, and bipartite genomes, composed of two ssDNA molecules. These genomes encode between four and eight overlapping, bidirectional open reading frames (ORFs). While New World begomoviruses predominantly show bipartite genome structures [39], Old World begomoviruses may exhibit monopartite or bipartite genome structures. Monopartite begomoviruses closely resemble the DNA-A component of bipartite begomoviruses and are often associated with virus-like satellite molecules, known as alpha and beta satellites, which play an important role in enhancing symptoms induced by these viruses [63–65].

The DNA-A component of begomoviruses typically encodes from five to seven proteins. The REP protein plays an essential role in replication, while the CP (Coat Protein) multitasks as the viral capsid builder, facilitator of vector transmission, and mediator of nuclear-cytoplasmic movement in monopartite viruses [9, 66]. The V2 (or AV2 in bipartite begomoviruses) protein acts as a suppressor of post-transcriptional gene silencing [67]. The Transcriptional Activator Protein (TrAP) interferes with transcriptional gene silencing (TGS) and post-transcriptional gene silencing (PTGS), acting as a necessary transcription factor for the expression of CP and NSP protein in bipartite begomoviruses [50, 68–70]. The Replication Enhancer Protein (REn), also known as C3 enhances viral DNA accumulation and recruits DNA polymerase δ for the synthesis of new copies of the geminiviral ssDNA genome [46, 68]. The C4 (or AC4) protein acts as a suppressor of RNA silencing [71, 72]. In the DNA-B component, proteins

responsible for cell-to-cell and long-distance movement are found, namely NSP (Nuclear Shuttle Protein) and MP (Movement Protein) [58, 73]. A recent study has found additional small ORFs, including V3, which acts as a gene silencing suppressor [74]. Lastly, the newly discovered C7 protein, encoded by isolates of tomato yellow leaf virus, plays a relevant role in viral infection as a pathogenicity factor, albeit less efficient as an RNA silencing suppressor compared to others [75].

In bipartite begomoviruses, both the DNA-A and DNA-B components share similar segments within their intergenic regions, spanning approximately 200 nucleotides referred to as the Common Region (CR). The CR acts as a relevant hub for sequence elements involved in the replication and transcription processes of the viral genome. The nonanucleotide sequence ('TAATATTAC') is mapped within the CR, which functions as the DNA cleavage site and serves as the initiation point for the replication process [1, 49].

Begomoviruses are transmitted in a persistent, non-propagative, and circulative manner, by a cryptic species complex of whiteflies, referred to as *Bemisia tabaci*. Begomoviruses primarily establish their infection within the phloem of infected plants [40, 41]. Notably, recent research has suggested the possibility of TYLCV replicating within the insect vector, although it remains uncertain whether this phenomenon is exclusive to TYLCV or applicable to all begomoviruses [76].

The widely distributed insect vector has had a relevant role for the successful dissemination of begomoviruses worldwide. Notably, the first reports of begomovirus infections coincided with the global spread of *Bemicia tabaci* species such as Middle East-Asia Minor 1 (MEAM1) and Mediterranean (MED), which introduced these viruses to previously unaffected regions, including Brazil, which is now recognized as a hotspot of begomovirus diversity [40, 77, 78]. It has been demonstrated that begomoviruses possess the ability to manipulate the preference and feeding behavior of whiteflies. Non-viruliferous whiteflies tend to favor virus-infected plants, while viruliferous whiteflies exhibit a higher propensity to feed on uninfected plants [79].

Begomoviruses possess a remarkable capacity for rapid evolution through various mechanisms, including mutation, pseudo-recombination, and recombination [13, 73, 80, 81]. Mutation are alterations in the genetic material of organisms, including viruses, and can occur due to errors during DNA replication. The high nucleotide substitution rates of begomoviruses genomes are similar to those observed in RNA viruses and contribute significantly to the genetic variability observed in their populations [82–84].

Recombination involves the random exchange of DNA or RNA segments between viruses, a process that significantly enhances genetic diversity and adaptive potential of a population [13, 85]. In the context of agriculture, the study of recombination is particularly relevant because it can give rise to recombinant viruses that may pose challenges in resistant crops. These recombinant viruses can potentially overcome previously effective resistance mechanisms, leading to the development of new strains that may threaten agricultural yields and crop sustainability. As a result, understanding and monitoring recombination events in viral populations is a relevant aspect of crop disease management [86–88].

Another mechanism that significantly contributes to genetic diversity in begomoviruses is pseudo-recombination, also known as reassortment. Pseudo-recombination involves the exchange of entire genomic components between viruses, typically within the same species. This process can result in the formation of hybrid viruses, especially when there is a high degree of genetic compatibility, particularly in the common regions of these components [14–16, 80, 89].

The population structure of begomoviruses

Begomoviruses populations show highly structured phylogenies associated with their geographical origin, as highlighted in numerous studies [17, 19, 20, 90]. For instance, a study conducted in Brazil focusing on the genetic structure of begomovirus populations affecting tomato crops and non-cultivated plant species revealed through phylogenetic and population structure analyses that populations comprising isolates of tomato common mosaic virus, tomato chlorotic mottle virus, and tomato severe rugose virus were segregated based on geographic location [19]. Similarly, a population consisting of cleome leaf crumple virus isolates from non-cultivated hosts exhibited geographical structure [19]. However, it is worth noting that despite these comprehensive analyses involving samples collected from diverse locations, spanning distances of up to 800 kilometers between collection sites, the existence of isolation by distance was not assessed.

A comprehensive analysis of begomoviruses in Costa Rica [90], which involved the examination of 651 plant samples infected with tomato yellow mottle virus, tomato leaf curl Sinaloa virus, pepper golden mosaic virus, and tomato yellow leaf curl virus, revealed a strong geographic segregation within their populations. In addition, the study

observed distinct begomovirus distribution depending on the geographical region and found that these viruses exhibited high host specificity.

A comprehensive study on the population genomics of begomoviruses identified the existence of at least seven major global subpopulations [91], further subdividing into as many as thirty-four smaller subpopulations that exhibited significant genetic differentiation and cohesiveness. This research provided evidence of isolation by distance, indicating that geographical barriers, including physical obstacles such as mountains, and reproductive isolation can significantly impact the spread of plant viruses. Furthermore, additional studies conducted on a global scale have found further support to these findings. They have consistently demonstrated the presence of highly differentiated genetic clusters within the *Bemisia tabaci* cryptic species complex, where gene flow between populations ranges from minimal to completely absent. These investigations have provided evidence of robust geographic structuring within this complex [92].

A study conducted in Pakistan has shed light on the potential correlation between the geographic distribution of viruses and vector genotypes. This study found similar phylogenetic relationships between the viral coat protein gene and the mitochondrial gene cytochrome oxidase I (mtCOI) of the insect vector. These findings suggest a complex interplay between the virus and its insect vector, hinting at the possibility of their coevolution or that the insect vector could significantly influence the virus population structure [93].

The concept of geographic isolation in populations is based on a proportional relationship between the geographic distances that separate organisms and their genetic distances [94]. Essentially, it implies that genetic diversity tends to increase as geographic distances become greater. Studies focused on geographic structure are of great importance for gaining insights into evolution of species. They help in understanding the dynamics of natural selection, as well as in estimating gene flow and historical migration patterns between different populations [95]. The multifaceted factors contributing to the genetic structure of populations underscore the need for comprehensive studies to attain a profound understanding of the complete evolutionary history of pathogens like begomoviruses. Therefore, there is a pressing need for further research dedicated to investigate the presence of long-range isolation among begomovirus populations.

References

1. **Fiallo-Olivé E, Lett J-M, Martin DP, Roumagnac P, Varsani A, *et al.*** ICTV Virus Taxonomy Profile: Geminiviridae 2021. *Journal of General Virology* 2021;102:001696. https://doi.org/10.1099/jgv.0.001696
2. **De Barro PJ, Liu S-S, Boykin LM, Dinsdale AB**. Bemisia tabaci: A Statement of Species Status. *Annu Rev Entomol* 2010;56:1–19. https://doi.org/10.1146/annurev-ento-112408-085504
3. **Kil E-J, Byun H-S, Hwang H, Lee K-Y, Choi H-S, *et al.*** Tomato Yellow Leaf Curl Virus Infection in a Monocotyledonous Weed (Eleusine indica). *Plant Pathol J* 2021;37:641–651. https://doi.org/10.5423/PPJ.FT.11.2021.0162
4. **Legg J, Winter S**. Cassava Mosaic Viruses (Geminiviridae). *Encyclopedia of Virology* 2021;301–312. https://doi.org/10.1016/B978-0-12-809633-8.21523-9
5. **Moriones E, Navas-Castillo J**. Tomato yellow leaf curl virus, an emerging virus complex causing epidemics worldwide. *Virus Res* 2000;71:123–134. https://doi.org/10.1016/S0168-1702(00)00193-3
6. **Lefeuvre P, Martin DP, Harkins G, Lemey P, Gray AJA, *et al.*** The Spread of Tomato Yellow Leaf Curl Virus from the Middle East to the World. *PLoS Pathog* 2010;6:e1001164.https://doi.org/10.1371/journal.ppat.1001164
7. **Laufs J, Jupin I, David C, Schumacher S, Heyraud-Nitschke F, *et al.*** Geminivirus replication: Genetic and biochemical characterization of rep protein function, a review. *Biochimie* 1995;77:765–773. https://doi.org/10.1016/0300-9084(96)88194-6
8. **Bisaro DM**. Silencing suppression by geminivirus proteins. *Virology* 2006;344:158–168. https://doi.org/10.1016/j.virol.2005.09.041
9. **Harrison BD, Swanson MM, Fargette D**. Begomovirus coat protein: serology, variation and functions. *Physiol Mol Plant Pathol* 2002;60:257–271. https://doi.org/10.1006/pmpp.2002.0404
10. **Sanderfoot AA, Lazarowitz SG**. Getting it together in plant virus movement: cooperative interactions between bipartite geminivirus movement proteins. *Trends Cell Biol* 1996;6:353–358. https://doi.org/10.1016/0962-8924(96)10031-3
11. **Zhou X**. Advances in Understanding Begomovirus Satellites. *Annu Rev Phytopathol* 2013;51:357–381. https://doi.org/10.1146/annurev-phyto-082712-102234
12. **Duffy S, Holmes EC**. Phylogenetic Evidence for Rapid Rates of Molecular Evolution in the Single-Stranded DNA Begomovirus *Tomato Yellow Leaf Curl Virus*. *J Virol* 2008;82:957–965. https://doi.org/10.1128/JVI.01929-07
13. **Lefeuvre P, Moriones E**. Recombination as a motor of host switches and virus emergence: geminiviruses as case studies. *Curr Opin Virol* 2015;10:14–19. https://doi.org/10.1016/j.coviro.2014.12.005
14. **Andrade EC, Manhani GG, Alfenas PF, Calegario RF, Fontes EPB, *et al.*** Tomato yellow spot virus, a tomato-infecting begomovirus from Brazil with a closer relationship to viruses from Sida sp., forms pseudorecombinants with begomoviruses from tomato but not from Sida. *Journal of General Virology* 2006;87:3687–3696. https://doi.org/10.1099/vir.0.82279-0
15. **Gilbertson RL, Hidayat SH, Paplomatas EJ, Rojas MR, Hou Y-M, *et al.*** Pseudorecombination between infectious cloned DNA components of tomato mottle and bean dwarf mosaic geminiviruses. *Journal of General Virology* 1993;74:23–31. https://doi.org/10.1099/0022-1317-74-1-23
16. **Paplomatas EJ, Patel VP, Yu-Ming Hou, Noueiry AO, Gilbertson RL**. Molecular characterization of a new sap-transmissible bipartite genome geminivirus infecting tomatoes in Mexico. *Phytopathology* 1994;84:1215–1224. https://doi.org/10.1094/Phyto-84-1215
17. **Prasanna HC, Sinha DP, Verma A, Singh M, Singh B, *et al.*** The population genomics of begomoviruses: global scale population structure and gene flow. *Virol J* 2010;7:220. https://doi.org/10.1186/1743-422X-7-220

18. **Bull SE, Briddon RW, Sserubombwe WS, Ngugi K, Markham PG, *et al.*** Infectivity, pseudorecombination and mutagenesis of Kenyan cassava mosaic begomoviruses. *Journal of General Virology* 2007;88:1624–1633. https://doi.org/10.1099/vir.0.82662-0
19. **Rocha CS, Castillo-Urquiza GP, Lima ATM, Silva FN, Xavier CAD, *et al.*** Brazilian Begomovirus Populations Are Highly Recombinant, Rapidly Evolving, and Segregated Based on Geographical Location. *J Virol* 2013;87:5784–5799. https://doi.org/10.1128/JVI.00155-13
20. **Mar TB, Xavier CAD, Lima ATM, Nogueira AM, Silva JCF, *et al.*** Genetic variability and population structure of the New World begomovirus Euphorbia yellow mosaic virus. *Journal of General Virology* 2017;98:1537–1551. https://doi.org/10.1099/jgv.0.000784
21. **Real LA, Henderson JC, Biek R, Snaman J, Jack TL, *et al.*** Unifying the spatial population dynamics and molecular evolution of epidemic rabies virus. *Proceedings of the National Academy of Sciences* 2005;102:12107–12111. https://doi.org/10.1073/pnas.0500057102
22. **Lam TT-Y, Ip HS, Ghedin E, Wentworth DE, Halpin RA, *et al.*** Migratory flyway and geographical distance are barriers to the gene flow of influenza virus among North American birds. *Ecol Lett* 2012;15:24–33. https://doi.org/10.1111/j.1461-0248.2011.01703.x
23. **Ahmad SU, Hafeez Kiani B, Abrar M, Jan Z, Zafar I, *et al.*** A comprehensive genomic study, mutation screening, phylogenetic and statistical analysis of SARS-CoV-2 and its variant omicron among different countries. *J Infect Public Health* 2022;15:878–891. https://doi.org/10.1016/j.jiph.2022.07.002
24. **Latinne A, Hu B, Olival KJ, Zhu G, Zhang L, *et al.*** Origin and cross-species transmission of bat coronaviruses in China. *Nat Commun* 2020;11:4235. https://doi.org/10.1038/s41467-020-17687-3
25. **Farooq T, Umar M, She X, Tang Y, He Z**. Molecular phylogenetics and evolutionary analysis of a highly recombinant begomovirus, Cotton leaf curl Multan virus, and associated satellites. *Virus Evol* 2021;7:veab054. https://doi.org/10.1093/ve/veab054
26. **Kumar RV, Singh AK, Singh AK, Yadav T, Basu S, *et al.*** Complexity of begomovirus and betasatellite populations associated with chilli leaf curl disease in India. *Journal of General Virology* 2015;96:3143–3158. https://doi.org/10.1099/jgv.0.000254
27. **Ilyas M, Qazi J, Mansoor S, Briddon RW**. Genetic diversity and phylogeography of begomoviruses infecting legumes in Pakistan. *Journal of General Virology* 2010;91:2091–2101. https://doi.org/10.1099/vir.0.020404-0
28. **Sánchez-Campos S, Martínez-Ayala A, Márquez-Martín B, Aragón-Caballero L, Navas-Castillo J, *et al.*** Fulfilling Koch's postulates confirms the monopartite nature of tomato leaf deformation virus: A begomovirus native to the New World. *Virus Res* 2013;173:286–293. https://doi.org/10.1016/j.virusres.2013.02.002
29. **Venkataravanappa V, Lakshminarayana Reddy CN, Swaranalatha P, Jalali S, Briddon RW, *et al.*** Diversity and phylogeography of begomovirus-associated beta satellites of okra in India. *Virol J* 2011;8:555. https://doi.org/10.1186/1743-422X-8-555
30. **Patil BL, Fauquet CM**. Cassava mosaic geminiviruses: actual knowledge and perspectives. *Mol Plant Pathol* 2009;10:685–701. https://doi.org/10.1111/j.1364-3703.2009.00559.x
31. **Hema M, Sreenivasulu P, Patil BL, Kumar PL, Reddy DVR**. Chapter Nine - Tropical Food Legumes: Virus Diseases of Economic Importance and Their Control. In: Loebenstein G, Katis N (editors). *Advances in Virus Research*. Academic Press. pp. 431–505. https://doi.org/10.1016/B978-0-12-801246-8.00009-3
32. **Picó B, Díez MJ, Nuez F**. Viral diseases causing the greatest economic losses to the tomato crop. II. The Tomato yellow leaf curl virus — a review. *Sci Hortic* 1996;67:151–196. https://doi.org/10.1016/S0304-4238(96)00945-4
33. **Zhang W, Olson NH, Baker TS, Faulkner L, Agbandje-McKenna M, *et al.*** Structure of the Maize Streak Virus Geminate Particle. *Virology* 2001;279:471–477. https://doi.org/10.1006/viro.2000.0739

34. **Krupovic M, Ravantti JJ, Bamford DH**. Geminiviruses: a tale of a plasmid becoming a virus. *BMC Evol Biol* 2009;9:112. https://doi.org/10.1186/1471-2148-9-112
35. **Bettina B, Sigrid U, Hugo C, B RR, Holger J**. Geminate Structures of African Cassava Mosaic Virus. *J Virol* 2004;78:6758–6765. https://doi.org/10.1128/JVI.78.13.6758-6765.2004
36. **Rojas MR, Macedo MA, Maliano MR, Soto-Aguilar M, Souza JO, *et al.*** World Management of Geminiviruses. *Annu Rev Phytopathol* 2015;56:637–677. https://doi.org/10.1146/annurev-phyto-080615-100327
37. **Inoue-Nagata AK, Lima MF, Gilbertson RL**. A review of geminivirus diseases in vegetables and other crops in Brazil: current status and approaches for management. *Hortic Bras*;34. https://doi.org/10.1590/S0102-053620160000100002
38. **Roumagnac P, Lett J-M, Fiallo-Olivé E, Navas-Castillo J, Zerbini FM, *et al.*** Establishment of five new genera in the family Geminiviridae: Citlodavirus, Maldovirus, Mulcrilevirus, Opunvirus, and Topilevirus. *Arch Virol* 2022;167:695–710. https://doi.org/10.1007/s00705-021-05309-2
39. **Zerbini FM, Briddon RW, Idris A, Martin DP, Moriones E, *et al.*** ICTV Virus Taxonomy Profile: Geminiviridae. *J Gen Virol* 2017;98:131—133. https://doi.org/10.1099/jgv.0.000738
40. **Navas-Castillo J, Fiallo-Olivé E, Sánchez-Campos S**. Emerging Virus Diseases Transmitted by Whiteflies. *Annu Rev Phytopathol* 2011;49:219–248. https://doi.org/10.1146/annurev-phyto-072910-095235
41. **Gilbertson RL, Batuman O, Webster CG, Adkins S**. Role of the Insect Supervectors Bemisia tabaci and Frankliniella occidentalis in the Emergence and Global Spread of Plant Viruses. *Annu Rev Virol* 2015;2:67–93. https://doi.org/10.1146/annurev-virology-031413-085410
42. **Stenger DC, Revington GN, Stevenson MC, Bisaro DM**. Replicational release of geminivirus genomes from tandemly repeated copies: evidence for rolling-circle replication of a plant viral DNA. *Proceedings of the National Academy of Sciences* 1991;88:8029–8033. https://doi.org/10.1073/pnas.88.18.8029
43. **Saunders K, Lucy A, Stanley J**. DNA forms of the geminivirus African cassava mosaic virus consistent with a rolling circle mechanism of replication. *Nucleic Acids Res* 1991;19:2325–2330. https://doi.org/10.1093/nar/19.9.2325
44. **Wu M, Bejarano ER, Castillo AG, Lozano-Durán R**. Chapter 20 - Geminivirus DNA replication in plants. In: Gaur RK, Sharma P, Czosnek H (editors). *Geminivirus : Detection, Diagnosis and Management*. Academic Press. pp. 323–346. https://doi.org/10.1016/B978-0-323-90587-9.00038-9
45. **Saunders K, Lucy A, Stanley J**. RNA-primed complementary-sense DNA synthesis of the geminivirus African cassava mosaic virus. *Nucleic Acids Res* 1992;20:6311–6315. https://doi.org/10.1093/nar/20.23.6311
46. **Wu M, Wei H, Tan H, Pan S, Liu Q, *et al.*** Plant DNA polymerases α and δ mediate replication of geminiviruses. *Nat Commun* 2021;12:2780. https://doi.org/10.1038/s41467-021-23013-2
47. **Fontes EP, Luckow VA, Hanley-Bowdoin L**. A geminivirus replication protein is a sequence-specific DNA binding protein. *Plant Cell* 1992;4:597–608. https://doi.org/10.1105/tpc.4.5.597
48. **Lazarowitz SG, Wu LC, Rogers SG, Elmer JS**. Sequence-specific interaction with the viral AL1 protein identifies a geminivirus DNA replication origin. *Plant Cell* 1992;4:799–809. https://doi.org/10.1105/tpc.4.7.799
49. **Mélia B, Stéphane B, Yannis M**. Replication mechanisms of circular ssDNA plant viruses and their potential implication in viral gene expression regulation. *mBio* 2023;0:e01692-23.
50. **Hanley-Bowdoin L, Settlage SB, Orozco BM, Nagar S, Robertson D**. Geminiviruses: Models for Plant DNA Replication, Transcription, and Cell Cycle Regulation. *CRC Crit Rev Plant Sci* 1999;18:71–106. https://doi.org/10.1080/07352689991309162

51. **Zhao L, Rosario K, Breitbart M, Duffy S**. Chapter Three - Eukaryotic Circular Rep-Encoding Single-Stranded DNA (CRESS DNA) Viruses: Ubiquitous Viruses With Small Genomes and a Diverse Host Range. In: Kielian M, Mettenleiter TC, Roossinck MJ (editors). *Advances in Virus Research*. Academic Press. pp. 71–133. https://doi.org/10.1016/bs.aivir.2018.10.001
52. **Ruhel R, Chakraborty S**. Multifunctional roles of geminivirus encoded replication initiator protein. *Virusdisease* 2019;30:66–73. https://doi.org/10.1007/s13337-018-0458-0
53. **Jovel J, Preiß W, Jeske H**. Characterization of DNA intermediates of an arising geminivirus. *Virus Res* 2007;130:63–70. https://doi.org/10.1016/j.virusres.2007.05.018
54. **Jeske H, Lütgemeier M, Preiß W**. DNA forms indicate rolling circle and recombination-dependent replication of Abutilon mosaic virus. *EMBO J* 2001;20:6158–6167. https://doi.org/10.1093/emboj/20.21.6158
55. **Alberter B, Ali Rezaian M, Jeske H**. Replicative intermediates of Tomato leaf curl virus and its satellite DNAs. *Virology* 2005;331:441–448. https://doi.org/10.1016/j.virol.2004.10.043
56. **Morilla G, Castillo AG, Preiss W, Jeske H, Bejarano ER**. A Versatile Transreplication-Based System To Identify Cellular Proteins Involved in Geminivirus Replication. *J Virol* 2006;80:3624–3633. https://doi.org/10.1128/JVI.80.7.3624-3633.2006
57. **Preiss W, Jeske H**. Multitasking in Replication Is Common among Geminiviruses. *J Virol* 2003;77:2972–2980. https://doi.org/10.1128/JVI.77.5.2972-2980.2003
58. **Hanley-Bowdoin L, Bejarano ER, Robertson D, Mansoor S**. Geminiviruses: masters at redirecting and reprogramming plant processes. *Nat Rev Microbiol* 2013;11:777–788. https://doi.org/10.1038/nrmicro3117
59. **Noueiry AO, Lucas WJ, Gilbertson RL**. Two proteins of a plant DNA virus coordinate nuclear and plasmodesmal transport. *Cell* 1994;76:925–932. https://doi.org/10.1016/0092-8674(94)90366-2
60. **Hamilton WDO, Bisaro DM, Coutts RHA, Buck KW**. Demonstration of the bipartite nature of the genome of a single-stranded DNA plant virus by infection with the cloned DNA components. *Nucleic Acids Res* 1983;11:7387–7396. https://doi.org/10.1093/nar/11.21.7387
61. **Nawaz-ul-Rehman MS, Fauquet CM**. Evolution of geminiviruses and their satellites. *FEBS Lett* 2009;583:1825–1832. https://doi.org/10.1016/j.febslet.2009.05.045
62. **Mondal D, Mandal S, Shil S, Sahana N, Pandit GK, *et al.*** Genome wide molecular evolution analysis of begomoviruses reveals unique diversification pattern in coat protein gene of Old World and New World viruses. *Virusdisease* 2019;30:74–83. https://doi.org/10.1007/s13337-019-00524-7
63. **Gupta N, Reddy K, Bhattacharyya D, Chakraborty S**. Plant responses to geminivirus infection: guardians of the plant immunity. *Virol J* 2021;18:143. https://doi.org/10.1186/s12985-021-01612-1
64. **Paprotka T, Metzler V, Jeske H**. The first DNA 1-like α satellites in association with New World begomoviruses in natural infections. *Virology* 2010;404:148–157. https://doi.org/10.1016/j.virol.2010.05.003
65. **Ferro CG, Silva JP, Xavier CAD, Godinho MT, Lima ATM, *et al.*** The ever increasing diversity of begomoviruses infecting non-cultivated hosts: new species from Sida spp. and Leonurus sibiricus, plus two New World alphasatellites. *Annals of Applied Biology* 2017;170:204–218. https://doi.org/10.1111/aab.12329
66. **Padidam M, Beachy RN, Fauquet CM**. The Role of AV2 ("Precoat") and Coat Protein in Viral Replication and Movement in Tomato Leaf Curl Geminivirus. *Virology* 1996;224:390–404. https://doi.org/10.1006/viro.1996.0546
67. **Chowda-Reddy R V, Achenjang F, Felton C, Etarock MT, Anangfac M-T, *et al.*** Role of a geminivirus AV2 protein putative protein kinase C motif on subcellular localization

and pathogenicity. *Virus Res* 2008;135:115–124. https://doi.org/10.1016/j.virusres.2008.02.014
68. **Sunter G, Bisaro DM**. Regulation of a Geminivirus Coat Protein Promoter by AL2 Protein (TrAP): Evidence for Activation and Derepression Mechanisms. *Virology* 1997;232:269–280. https://doi.org/10.1006/viro.1997.8549
69. **Haley A, Zhan X, Richardson K, Head K, Morris B**. Regulation of the activities of African cassava mosaic virus promoters by the AC1, AC2, and AC3 gene products. *Virology* 1992;188:905–909. https://doi.org/10.1016/0042-6822(92)90551-Y
70. **Chowda-Reddy R V, Dong W, Felton C, Ryman D, Ballard K, *et al.*** Characterization of the cassava geminivirus transcription activation protein putative nuclear localization signal. *Virus Res* 2009;145:270–278. https://doi.org/10.1016/j.virusres.2009.07.022
71. **Fondong VN, Reddy RVC, Lu C, Hankoua B, Felton C, *et al.*** The Consensus N-Myristoylation Motif of a Geminivirus AC4 Protein Is Required for Membrane Binding and Pathogenicity. *Molecular Plant-Microbe Interactions®* 2007;20:380–391. https://doi.org/10.1094/MPMI-20-4-0380
72. **Ramachandran V, Padmanabhan C, S PJ, M FC**. Differential Roles of AC2 and AC4 of Cassava Geminiviruses in Mediating Synergism and Suppression of Posttranscriptional Gene Silencing. *J Virol* 2004;78:9487–9498. https://doi.org/10.1128/JVI.78.17.9487-9498.2004
73. **Rojas MR, Hagen C, Lucas WJ, Gilbertson RL**. Exploiting Chinks in the Plant's Armor: Evolution and Emergence of Geminiviruses. *Annu Rev Phytopathol* 2005;43:361–394. https://doi.org/10.1146/annurev.phyto.43.040204.135939
74. **Gong P, Tan H, Zhao S, Li H, Liu H, *et al.*** Geminiviruses encode additional small proteins with specific subcellular localizations and virulence function. *Nat Commun* 2021;12:4278. https://doi.org/10.1038/s41467-021-24617-4
75. **Liu H, Chang Z, Zhao S, Gong P, Zhang M, *et al.*** Functional identification of a novel C7 protein of tomato yellow leaf curl virus. *Virology* 2023;585:117–126. https://doi.org/10.1016/j.virol.2023.05.011
76. **He Y-Z, Wang Y-M, Yin T-Y, Fiallo-Olivé E, Liu Y-Q, *et al.*** A plant DNA virus replicates in the salivary glands of its insect vector via recruitment of host DNA synthesis machinery. *Proceedings of the National Academy of Sciences* 2020;117:16928–16937. https://doi.org/10.1073/pnas.1820132117
77. **Jones DR**. Plant Viruses Transmitted by Whiteflies. *Eur J Plant Pathol* 2003;109:195–219. https://doi.org/10.1023/A:1022846630513
78. **Ribeiro SG, de Ávila AC, Bezerra IC, Fernandes JJ, Faria JC, *et al.*** Widespread Occurrence of Tomato Geminiviruses in Brazil, Associated with the New Biotype of the Whitefly Vector. *Plant Dis* 1998;82:830. https://doi.org/10.1094/PDIS.1998.82.7.830C
79. **Zhao K, Liu S-S, Wang X-W, Yang J-G, Pan L-L**. Manipulation of Whitefly Behavior by Plant Viruses. *Microorganisms* 2022;10:2410. https://doi.org/10.3390/microorganisms10122410
80. **Pita JS, Fondong VN, Sangaré A, Otim-Nape GW, Ogwal S, *et al.*** Recombination, pseudorecombination and synergism of geminiviruses are determinant keys to the epidemic of severe cassava mosaic disease in Uganda. *Journal of General Virology* 2001;82:655–665. https://doi.org/10.1099/0022-1317-82-3-655
81. **Lima ATM, Sobrinho RR, González-Aguilera J, Rocha CS, Silva SJC, *et al.*** Synonymous site variation due to recombination explains higher genetic variability in begomovirus populations infecting non-cultivated hosts. *Journal of General Virology* 2013;94:418–431. https://doi.org/10.1099/vir.0.047241-0
82. **Acosta-Leal R, Duffy S, Xiong Z, Hammond RW, Elena SF**. Advances in Plant Virus Evolution: Translating Evolutionary Insights into Better Disease Management. *Phytopathology* 2011;101:1136–1148. https://doi.org/10.1094/PHYTO-01-11-0017
83. **Lima ATM, Silva JCF, Silva FN, Castillo-Urquiza GP, Silva FF, *et al.*** The diversification of begomovirus populations is predominantly driven by mutational dynamics. *Virus Evol* 2017;3:vex005. https://doi.org/10.1093/ve/vex005

84. **Duffy S, Holmes EC**. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *Journal of General Virology* 2009;90:1539–1547. https://doi.org/10.1099/vir.0.009266-0
85. **Padidam M, Sawyer S, Fauquet CM**. Possible Emergence of New Geminiviruses by Frequent Recombination. *Virology* 1999;265:218–225. https://doi.org/10.1006/viro.1999.0056
86. **Awadalla P**. The evolutionary genomics of pathogen recombination. *Nat Rev Genet* 2003;4:50–60. https://doi.org/10.1038/nrg964
87. **Sattar MN, Kvarnheden A, Saeed M, Briddon RW**. Cotton leaf curl disease – an emerging threat to cotton production worldwide. *Journal of General Virology* 2013;94:695–710. https://doi.org/10.1099/vir.0.049627-0
88. **Monci F, Sánchez-Campos S, Navas-Castillo J, Moriones E**. A Natural Recombinant between the Geminiviruses Tomato yellow leaf curl Sardinia virus and Tomato yellow leaf curl virus Exhibits a Novel Pathogenic Phenotype and Is Becoming Prevalent in Spanish Populations. *Virology* 2002;303:317–326. https://doi.org/10.1006/viro.2002.1633
89. **Sung YK, Coutts RHA**. Pseudorecombination and complementation between potato yellow mosaic geminivirus and tomato golden mosaic geminivirus. *Journal of General Virology* 1995;76:2809–2815. https://doi.org/10.1099/0022-1317-76-11-2809
90. **Barboza N, Blanco-Meneses M, Esker P, Moriones E, Inoue-Nagata AK**. Distribution and diversity of begomoviruses in tomato and sweet pepper plants in Costa Rica. *Annals of Applied Biology* 2018;172:20–32. https://doi.org/10.1111/aab.12398
91. **De Barro PJ**. Genetic structure of the whitefly Bemisia tabaci in the Asia-Pacific region revealed using microsatellite markers. *Mol Ecol* 2005;14:3695–3718. https://doi.org/10.1111/j.1365-294X.2005.02700.x
92. **Hadjistylli M, Roderick GK, Brown JK**. Global Population Structure of a Worldwide Pest and Virus Vector: Genetic Diversity and Population History of the Bemisia tabaci Sibling Species Group. *PLoS One* 2016;11:e0165105. https://doi.org/10.1371/journal.pone.0165105
93. **Simón B, Cenis JL, Beitia F, Khalid S, Moreno IM, *et al.*** Genetic Structure of Field Populations of Begomoviruses and of Their Vector Bemisia tabaci in Pakistan. *Phytopathology* 2003;93:1422–1429. https://doi.org/10.1094/PHYTO.2003.93.11.1422
94. **Roderick GK**. Geographic Structure of Insect Populations: Gene Flow, Phylogeography, and Their Uses. *Annu Rev Entomol* 1996;41:325–352. https://doi.org/10.1146/annurev.en.41.010196.001545
95. **Bohonak AJ**. Dispersal, Gene Flow, and Population Structure. *Q Rev Biol* 1999;74:21–45. https://doi.org/10.1086/392950

CHAPTER 1

THE PHYLOGEOGRAPHY OF BEGOMOVIRUSES: MAPPING INFORMATIVE REGIONS IN VIRAL GENOMES

Verdolin L.G., Souza G.M., Resende C.L.P., Santos M.M.C., Ribeiro J.P.F., Lima A.T.M. (2023) The phylogeography of begomoviruses: mapping informative regions in viral genomes. *Manuscript in preparation for publication.*

The phylogeography of begomoviruses: mapping informative regions in viral genomes

L. G. Verdolin, G. M. Souza, C. L. P. Resende, M. M. C. Santos, J.P.F. Ribeiro, A. T. M. Lima

Instituto de Ciências Agrárias, Universidade Federal de Uberlândia, Uberlândia – MG, 38405-302, Brazil

Corresponding Author: A.T.M. Lima

E-mail: atmlima@ufu.br

Abstract

Begomoviruses, known for their broad host range among dicotyledonous plants, are transmitted by a complex of cryptic whitefly species collectively referred to as *Bemisia tabaci*. Their genomes consist of single-stranded DNA, either one (monopartite) or two (bipartite) molecules, and their rapid evolution is primarily driven by mechanisms like mutation, recombination, and pseudorecombination, which contribute to their genetic variability and adaptability. Previous studies have consistently shown geographic segregation of begomovirus populations, indicating limited gene flow across regions. Despite numerous investigations into begomovirus genomes from different geographical regions, evidence of isolation by distance remains underexplored. This study employed a multivariate Procrustean approach on full-length DNA-A (or DNA-A-like) sequences, combined with a sliding window method for fine-scale mapping using 200-nucleotide segments. Detailed curation of spatial data associated with each DNA-A sequence was performed, drawing from GenBank and scientific publications. Genetic divergence among begomovirus isolates was analyzed through patristic distances calculated from maximum likelihood trees. A comprehensive correlation analysis of distance matrices, integrating spatial and genetic information, was conducted using the Procrustean Approach to Cophylogeny (PACo). The study provided robust evidence of isolation by distance in at least three begomovirus species datasets, including isolates of bean golden mosaic virus, cotton leaf curl Gezira virus, and tomato yellow leaf curl virus. It also revealed uneven distribution of the geographical signal across genomes, with evidence of isolation by distance more pronounced in localized segments, occasionally interspersed with regions lacking any such signal. Furthermore, the study highlighted how recombination-induced variation can obscure evidence of isolation by distance, even with a limited number of recombinant DNA-A sequences. Finally, we concluded that recent begomovirus incursions into distant regions from their original sites of origin contributed to reduced global congruence between spatial and genetic data.

Keywords: Begomovirus, Evolution, Phylogeny

Introduction

Begomoviruses (genus *Begomovirus*, family *Geminiviridae*) constitute one of the most diverse groups of plant viruses and have a relevant economic impact on global agriculture. These single-stranded DNA viruses are transmitted by a complex of cryptic whitefly species collectively known as *Bemisia tabaci* [1]. Infections by these viruses produce distinctive and severe symptoms like mosaic patterns, mottling, yellowing, leaf curling, and dwarfism, resulting in substantial losses in key crops worldwide [2–4]. The genomes of begomoviruses can exist in either non-segmented (monopartite) or segmented (bipartite) structures. The bipartite genomes consist of two genomic components named DNA-A and DNA-B, each roughly 2,600 nucleotides long. The DNA-A encodes proteins essential for replication, gene silencing suppression, and viral progeny encapsidation [5–7]. In contrast, DNA-B encodes proteins that mediate short distance and systemic movement of the virus within host plants [8]. For a successful infection by a bipartite begomovirus, both genomic components must be present within the host [9].

Begomovirus genomes evolve at rates similar to those observed in RNA viruses [10]. The rapid mutational dynamics combined with their propensity for recombination, provide begomovirus populations with a high degree of adaptability to new hosts [11, 12]. Pseudorecombination among bipartite begomovirus isolates is also a relevant evolutionary mechanism, generating new combinations of genomic components with the potential to produce novel and unique phenotypes [13–15]. Collectively, these mechanisms drive the emergence of new strains capable of overcoming plant resistance mechanisms in economically relevant crops [16–19].

Previous studies, employing phylogenetic and population genetic analytical methods, have revealed the geographical segregation of begomovirus populations [20-23]. However, few investigations have focused on the evidence of isolation by distance, particularly utilizing the Mantel test, a widely-used approach for correlating genetic and spatial distance measurements [24–26].

The concept of isolation by distance is explained by a proportional relationship between genetic similarity and geographic distance among populations. As geographic distance increases, genetic divergence tends to intensify, often attributed to spatial limitations in gene flow or the presence of physical barriers [27, 28]. This concept serves as the theoretical basis for understanding evolutionary and migratory patterns across a number of organisms, including viruses. Its wide-ranging applicability is evident in

studies including both human and non-human infecting viruses, such as the Highly Pathogenic Avian Influenza A (H5N1, hemagglutinin type 5 and neuraminidase type 1). In a comprehensive study conducted with samples obtained from infected birds in North America, the evidence of isolation by distance was confirmed. Migration rates between the most remote flyways, specifically the Pacific and Atlantic flyways, were significantly lower in comparison to other routes. This observation highlights the role of these distant flyways as physical barriers, reducing the spread of the virus [29, 30]. Likewise, in another investigation which analyzed 125 influenza viruses samples collected in Vietnam between 2003 and 2007, compelling evidence of isolation by distance was also documented [31]. This concept also applies to the field of population genetics concerning insect-transmitted viruses. For example, in the case of the dengue virus, transmitted by the *Aedes aegypti* mosquito, larvae samples were collected from various locations across Mexico. These samples were analyzed to assess gene flow and the potential for isolation by distance. Interestingly, no evidence of isolation by distance was found within this specific sample pool [32].

In this context, the primary objective of this study was to determine if there is substantial evidence supporting the isolation by distance among begomovirus populations. Our specific objectives included the evaluation of the strength of geographic signals within full-length DNA sequences and the investigation of whether any evidence of isolation by distance is influenced by the geographic scale or the level of genetic variation in our datasets. To accomplish these objectives, we adapted a multivariate Procrustean analysis to assess the congruency between genetic and spatial data associated with isolates of begomoviruses.

Material and Methods

Begomovirus species data sets

The dataset of this study included sequences of begomovirus isolates belonging to 23 distinct species for which a minimum of 50 DNA-A or DNA-A-like sequences were available on GenBank as of March 7, 2022. Our data curation process involved manual review, in which we only retained sequences with available information regarding the sampling collection site. This information was sourced from either the GenBank database or the associated scientific literature. After excluding sequences that did not meet these

criteria, we retained a total of 3,162 full-length DNA-A nucleotide sequences (Supplementary Table S1).

Sequence alignments and sub-alignments

Multiple sequence alignments were constructed using Muscle5 [33] based on full-length DNA-A sequences of begomovirus isolates from same species. Before the sequence alignment step, we conducted preprocessing of the datasets. If identical sequences were found from the same host, collection location, and date, we retained only one in the final dataset. After constructing the alignment based on the interspecific dataset, we used trimAL [34] to remove columns with 51% or more gaps, particularly in poorly aligned intergenic region. This refinement was necessary due to the wide variation in genome sizes, especially between begomoviruses originating from the New and Old Worlds. This variation resulted in a number of sites predominantly composed of gaps in the interspecific dataset alignment. To minimize data loss, we set the program to preserve a minimum of 95% of the alignment columns. Subsequently, a custom Python3 script was employed to separate the aligned sequences of a same begomovirus species into individual datasets. A second custom Python3 script was used to slice the alignments into sub-alignments. Each sub-alignment was 200 nucleotides long (*i.e.*, the sliding window length) and moved in increments of 20 nucleotides (*i.e.*, step size). The dataset processing yielded 23 alignments based on full-length DNA-A sequences and 3,473 sub-alignments.

Recombination analysis

Alignments based on full-length DNA-A sequences were scanned for recombination events using RDP4 [35]. Any sequences identified as recombinant by at least four of the seven available analytical methods (RDP, Geneconv, Bootscan, Maximum Chi Square, Chimaera, Sister Scan, and 3Seq) were removed. The datasets were then realigned using Muscle5 and submitted to slicing step as described above.

Assessing genetic variability

Nucleotide diversity indices (π, [36]) were calculated for all alignments and sub-alignments using custom Python3 script (available upon request). The 95% bootstrap confidence intervals were derived from 3,000 non-parametric simulations using the boot package [37] in R software [38]. We represented the π values using heatmaps constructed in the R package ComplexHeatmap [39] and clustered similar patterns of genetic variation

171 across DNA-A sequences by means of dendrograms constructed using the R package
172 dendextend [39, 40].
173
174 *Phylogenetic analysis*
175 Maximum Likelihood (ML) phylogenetic trees were constructed using IQ-Tree
176 [41]. The best fit nucleotide substitution models were determined using ModelFinder
177 [42], also implemented in IQ-Tree. Branch support was assessed from 2,000 and 5,000
178 ultrafast bootstrap replicates [43] for sub-alignments and full-length DNA-A sequence
179 alignments, respectively. ML-trees were edited using the R package ggtree [44].
180
181 *Multivariate superimposition of genetic and spatial data*
182 We assessed the extent of superimposition between patristic and geodesic distance
183 matrices by means of a multivariate Procrustean approach [45] implemented in the R
184 package PACo (Procrustean Approach to Cophylogeny [46]). First, the patristic distances
185 separating all tip pairs in our ML trees constructed for full-length DNA-A sequences and
186 their sub-alignments were computed using the "cophenetic" function implemented in the
187 R package ape [47]. We opted to use patristic distances instead of raw genetic distances
188 directly calculated from the sequences since they represent measures optimized via
189 maximum likelihood during the phylogenetic reconstruction. We calculated the geodesic
190 distances separating all pairs of sampling collection sites for a given begomovirus species
191 dataset, whose original spatial information (either, the precise or centroid geographical
192 coordinates) was manually curated from GenBank website or related scientific
193 publications. In all cases, we used the most precise geographical information available.
194 The geodesic distance calculation was performed using the R package geodist [48].
195 Both patristic and geodesic distance matrices were transformed into principal
196 coordinates (PCo) using the package PACo. The individual contributions (*i.e.*, the
197 Procrustes residuals) for every link (the reciprocal projections into the multivariate spaces
198 of patristic and geodesic distances) were estimated using a jackknife method. The global
199 congruency statistic was then calculated by the sum of squared Procrustes residuals
200 ($\sum m^2_{XY}$) and its statistical significance was assessed by means of 1,000 permutations in
201 R software also using the package PACo. This methodology proved to be useful
202 compared to the conventional Mantel's Test, as it quantifies the individual deviations for
203 each taxon in the ML-tree under a prior assumption of isolation-by-distance.

Using the Procrustean Approach to Cophylogeny (PACo, [45]), we were able to assess the robustness of evidence for isolation by distance among the begomovirus datasets. We superimposed projections in the multivariate space of patristic distances calculated from Maximum Likelihood (ML) trees and geodesic distances calculated between all sampling collection sites. The global incongruence between genetic and geodesic distance data is provided by the sum of squared Procrustes residuals (SSPR, $\sum m^2_{XY}$), with values ranging from 0 to 1. A SSPR value of zero indicates a perfect fit between genetic and spatial information and, consequently, robust evidence for isolation by distance. Conversely, a SSPR value of one represents complete absence of evidence for isolation by distance.

We represented graphically patristic distances using heatmaps elaborated in the R package ComplexHeatmap, while geodesic data was visualized using maps elaborated in the R package ggplot2 [49]. Linear regression models were elaborated to correlate the global sum of squared Procrustes residuals with nucleotide diversity indices and geodesic distances using the R package ggpmisc [50]. Standardized geopolitical data, including country names and codes and geographic sub-regions information were sourced from the R package countrycode [51].

Results

The geographical signal in full-length DNA-A sequences

This study involved analyses of datasets of varying sample sizes, with the three largest consisting of full-length DNA-A sequences of isolates belonging to the species *Tomato yellow leaf curl virus* (TYLCV, $N = 450$), *African cassava mosaic virus* (ACMV, $N = 304$), and *Tomato leaf curl New Delhi virus* (ToLCNDV, $N = 209$). Conversely, the smallest dataset encompassed 35 DNA-A sequences of papaya leaf curl China virus isolates (species *Papaya leaf curl China virus*, PaLCuCNV) (Figure 1a). The spatial distribution of sampling collection sites also exhibited considerable variability across these datasets. For example, sweet potato leaf curl virus isolates (species *Sweet potato leaf curl virus*, SPLCV) were collected from sites separated by an average distance of 9,000 kilometers. Some SPLCV isolates were collected from sites geographically distant, such as those in China and Brazil, approximately 16,000 km apart. It also included moderately distant sampling locations, like that between Peru and the United States,

approximately 5,500 km apart, as well as relatively close locations, exemplified by Argentina and Brazil, which were approximately 1,000 km apart. Similarly, sampling sites of TYLCV isolates were, on average, 7,500 kilometers apart, with the most distant sites spanning 19,000 kilometers, between locations in Oceania e North Africa. Additionally, some datasets comprised isolates collected from a more limited geographical area, with distances between sampling sites being less than 1,000 km. This was observed in the datasets comprising isolates of tomato severe rugose virus (species *Tomato severe rugose virus*, ToSRV) and South African mosaic virus (species *South African mosaic virus*, SACMV) (Figure 1a). The wide range of distances between collection sites made these datasets particularly suitable for assessing the extent of isolation by distance at different geographical scales.

We obtained high SSPR values (greater than 0.75) for 11 out of the 23 begomovirus species datasets: AYVV, BYVMV, EACMKV, MYMIV, PaLCuCNV, PepGMV, SACMV, TbCSV, ToLCNDV, ToLCTV and ToSRV, indicating weak evidence of isolation by distance (Figure 1b). Moderate values (from 0.50 to 0.75) were obtained for nine datasets: ACMV, ChiLCV, CLCuGeV, CLCuMuV, EACMV, EuYMV, PepYVMLV, SLCCNV and SLCuV. Values below 0.50 were obtained for three datasets: SPLCV, TYLCV and BGMV (SSPRs of 0.46, 0.31 and 0.24, respectively) suggesting stronger support of isolation by distance.

The wide range of SSPR values led us to investigate whether the geographical coverage of the collection sites influences the evidence of isolation by distance. We conducted a linear regression analysis correlating the SSPR values with the average geodesic distances between collection sites (Figure 1c). We observed that less than 40% ($R^2 = 0.38$) of the variation in the SSPR values could be explained by that of average geodesic distances between collection sites. We also tested whether the genetic variation of each dataset might also influence the extent of isolation by distance (Figure 1d). These analyses indicated a negligible effect of genetic variation on the SSPR values. Therefore, while geographical coverage of the sampling collection sites weakly predicts the extent of isolation by distance, our results indicate that the lack of robust evidence of isolation by distance in most of our datasets was not a consequence of low genetic variation.

The geographic signal is not evenly distributed across DNA-A sequences

We conducted a more detailed investigation of the geographic signal distribution across viral DNA-A sequences using a sliding window approach. We constructed a multiple alignment containing all 3,162 full-length DNA-A sequences of begomovirus isolates analyzed in this study. Subsequently, sequences of isolates belonging to distinct begomovirus species were separated again to compose intraspecific datasets. Then, the individual alignments were sliced into sliding windows with lengths of 200 nucleotides and a step size of 20 nucleotides. This ensured that equivalent sliding windows contained homologous sequences, allowing us to compare the patterns of geographic signal distribution among datasets.

Through a clustering analysis, we observed the existence of three major clusters based on the distribution of SSPR values (Figure 2a). The first cluster consisted of 13 datasets (SACMV, BYVMV, ToSRV, TbCSV, EACMKV, ToLCTV, AYVV, ACMV, EuYMV, MYMIV, PepGMV, ToLCNDV and PepYVMLV) whose sliding windows yielded weak evidence of isolation by distance and exhibited a more even distribution of the geographic signal. A divergent pattern was that of PepYVMLV dataset, in which sliding windows mapped in the central region of their DNA-A sequences showed considerably stronger evidence of isolation by distance, with SSPR values close to 0.50, and some windows even showed values below 0.25.

A second cluster consisting of eight datasets (ChiLCV, EACMV, PaLCuCNV, CLCuMuV, SLCuV, CLCuGeV, SLCCNV and SPLCV) showed stronger evidence of isolation by distance and also exhibited a wider range of SSPR values across their sliding windows, with increased support for isolation by distance at the 5' end and/or central region of the DNA-A sequences. This pattern was particularly evident for the EACMV, SLCuV, CLCuGeV and SLCCNV datasets (Figure 2a and Supplementary Figure S3). Finally, the datasets of BGMV and TYLCV, which showed the lowest SSPR values based on their full-length DNA-A sequences, also exhibited evidence of uneven distribution of the geographic signal, with considerably lower SSPR values, especially at the 5' end and/or central region of their DNA-A sequences. These results strongly suggest that in datasets where more robust evidence of isolation by distance was observed based on full-length DNA-A sequences, certain genomic regions contributed more than others to the overall congruence between genetic and spatial information.

We investigated whether the uneven distribution of the geographic signal might be associated with an uneven distribution of genetic variation in each of the analyzed sliding windows (Figure 2b). Interestingly, genetic variation across DNA-A sequences is also unevenly distributed. We conducted linear regression analyses between the SSPR values and nucleotide diversity values calculated for each of the sliding windows across DNA-A sequences (Figures 2c, 2d, 2e and Supplementary Figure S4). In most cases, the variation in SSPR values explained by genetic variation in the regression models was negligible, once again reinforcing that the variation in SSPR values does not seems to be significantly affected by the content of genetic variation, except in the cases of the BGMV (Figure 2c) and EACMV (Figure 2d) datasets, whose models indicated that 43% and 60% of the variation in SSPR values, respectively, is explained by the variation in nucleotide diversity values. Similar to the other datasets, in the TYLCV dataset, with the largest sample size among all analyzed in this study, we did not observe significant influence (R^2 = 0.04) of the genetic variation on the isolation by distance signal (Figure 2e).

The evidence of isolation by distance at different geographic scales

Given the stronger support of isolation by distance observed for the BGMV and TYLCV datasets, alongside the uneven distribution of the geographic signal across DNA-A sequences, we investigated the contributions of individual begomovirus isolates to the global congruence between genetic and spatial data (Figure 3). Unlike TYLCV, whose isolates were sampled across various sub-geographic regions, mainly in the northern hemisphere, isolates of BGMV has been exclusively sampled in Brazil. A substantial number of isolates were collected from sites in three Brazilian regions: Midwest (State of Goiás and Distrito Federal), Northeast (States of Alagoas and Pernambuco), and Southeast (State of Minas Gerais) (Figure 3a). BGMV isolates were sourced from two main collection sites in the state of Minas Gerais, the first situated near the border with the state of Goiás (municipality of Unaí), while the second site is situated in the central region of the state (municipality of Florestal). Collection sites in Goiás are comparatively close to those in Unaí, Minas Gerais (geodesic distances ranging from 89 to 275 km, with an average of 170 km). However, they are substantially farther from Florestal with an average distance of 557 km. Overall, these collection sites in Goiás and Minas Gerais are significantly more distant from sampling sites in northeastern states (from 1,328 to 1,830 km).

In fact, some of the most substantial patristic distances (PD) in the DNA-A-based ML tree were those between isolates collected in the Southeast and Midwest regions with those from the Northeast region (Figure 3b and Supplementary Figure S6). Moreover, isolates from Minas Gerais collected near the border and those collected within Goiás exhibited closer relationships in the ML tree. Isolates from Florestal were more distantly related to the others within the same state, which is also consistent with some degree of isolation by distance. An exception to this pattern was observed with two BGMV isolates (GenBank accession numbers: KJ939710 and KJ939711) collected in the state of Pernambuco, which clustered with isolates from Southeast and Midwest regions. A noteworthy incongruency is that the greatest patristic distances calculated from the ML tree were the ones separating these isolates from those collected in the state of Alagoas, whose collection sites are separated by comparatively short distances (from 175 to 271 km). Both isolates were relevant contributors to the global incongruence observed between genetic and spatial data (Procrustes residuals of 0.36 and 0.35, respectively).

Given the uneven distribution of the geographic signal, we partitioned the BGMV DNA-A-based alignment by separating genomic regions with stronger support of isolation by distance from those with weaker support. We set an arbitrary threshold of SSPR of 0.3 to partition the viral component into two segments: the first composed of all alignment columns within sliding windows that yielded SSPR values below 0.3 (segment 1), and the second segment encompassed all alignment columns within sliding windows that yielded SSPR values above 0.3 (segment 2) (Figure 3c). We reconstructed the phylogenies based on each segment, separately (Figures 3d and 3e), and re-evaluated the support of isolation by distance by calculating their SSPR values. We obtained SSPR values of 0.2118 and 0.3188 for segments 1 and 2, respectively. Two major clusters were observed for the segment 1-based ML-tree, the first one included all isolates collected in the Midwest and Southeast regions, and two isolates collected in the state of Pernambuco, once again, they were the major individual contributors to the global incongruence (Figure 3d). In contrast, the segment 2-based ML-tree was better resolved, with several long internal branches separating smaller clusters of isolates (Figure 3e). Both BGMV isolates collected in the state of Pernambuco were separated from the other isolates collected in the Northeast region by a long branch ($PD_{[KJ939710:JN419006]}$ = 0.089 substitutions/site, Supplementary Figure S6b), making them considerable contributors to the global incongruence. Another considerable contributor to incongruence was the

isolate collected in the state of Alagoas (accession number KJ939720), which exhibited the shortest patristic distances ($PD_{[KJ939720:FJ665283]}$ = 0.014 substitutions/site, Supplementary Figure S6b) with isolates from the Midwest and Southeast regions.

We also investigated in details the incongruences between the genetic and spatial data in the TYLCV dataset. Given the presence of this begomovirus in a number of countries in both the western and eastern hemispheres, we conducted comparisons in a context where we grouped countries into 12 global sub-regions (Figure 4a). Some of the greatest distances separating any pairs of collection sites for TYLCV isolates were approximately 19,000 km, between Oceania and North Africa, followed by distances between sites in Sub-Saharan Africa and North America (18,250 km), Southern Europe and Oceania (18,039 km). Assuming a scenario of isolation by distance, the greatest expected patristic distances would also be those between isolates sampled from these same regions. A simple visual inspection of the ML-tree and its associated patristic distance matrix (Figure 4b and 4c, respectively) allowed us to observe that the greatest patristic distances were those separating a cluster of isolates with highly diverse geographic origins, including countries from the Middle East, Central America, Northern Europe, Sub-Saharan Africa, and East Asia from all other TYLCV isolates collected around the world. In another perspective of this incongruency, isolates from these same countries and sub-regions could also be found in the second major cluster. For instance, groups of isolates collected in Iran were separated by the greatest patristic distances observed from the ML-tree ($PD_{[GU076441:GU076454]}$ = 0.248 substitutions/site, Figures 4a and 4b). Similarly, isolates collected in the Dominican Republic were separated by considerable patristic distances from isolates collected in Cuba ($PD_{[KJ913683: KM926625]}$ = 0.120 substitutions/site, Figures 4a and 4b), both countries located in close proximity in Central America (850 km). By applying the Procrustean approach, we confirmed all isolates from this cluster as relevant contributors to the global incongruence (Procrustes residuals from 0.0425 to 0.0999, Figure 4a). Isolates from cluster 2 also contributed to the global incongruence; for example, an isolate sampled in the United States (accession number GU322424) was closely related to isolates collected in East Asia, making it one of the largest individual contributors to the global incongruence (Procrustes residual = 0.183). Similarly, three isolates from Costa Rica (accession numbers: KY064016, KF533857, and KF533856) were closely related and grouped with isolates collected in China. It is interesting to note that Procrustes residuals also indicated incongruences when

isolates were separated from others in the same geographical sub-region by smaller patristic distances than expected. This case is well illustrated by the isolate with accession number EF210554, which, despite being related to other isolates also sampled in the United States, also contributed to the global incongruence (Procrustes residual = 0.150, Figure 4a). It is important to note that some of the major incongruences that contributed to obscure the evidence of isolation by distance included TYLCV isolates sampled in New World countries. Contrary to the expectation of being separated by considerable patristic distances, they were closely related to isolates from the Eastern Hemisphere. We hypothesize that recent incursions of begomoviruses from Old World countries into the New World significantly contribute to reducing the support of isolation by distance. Furthermore, additional instances of begomovirus incursions into continents where they are not native, further exacerbating the global incongruence between genetic and spatial data, were also observed for ACMV (Supplementary Figure 1a), CLCuGeV (Supplementary Figure S1e), SLCuV (Supplementary Figure S1p), SPLCV (Supplementary Figure S1q) and PepYVMLV datasets (Supplementary Figure S1m). Another relevant observation is that the most robust support of isolation by distance was observed in datasets that differ significantly in terms of the geographic coverage scales of their sampling sites.

The effect of recombination on the evidence of isolation by distance

The inherent recombination-prone nature of begomovirus genomes has been extensively demonstrated for both monopartite and bipartite begomoviruses [11, 15, 19, 52, 53]. Recombination events often result in increased genetic variability, and in phylogenetic trees, sequences affected by recombination events are frequently associated with long branches [54]. To assess whether recombination events might have influenced the estimates of branch lengths and consequently obscured the geographical signal within the begomovirus datasets, we systematically removed all recombinant DNA-A sequences and re-evaluated the strength of the geographic signal (Figure 5a). Some datasets primarily composed of recombinant sequences were excluded from the reanalyzes such as those of AYVV, BYVMV, CLCuMuV. Three distinct patterns emerged after the removal of recombinant sequences. In the first pattern, there was a noticeable to substantial increase in SSPR values, exemplified by the EACMV dataset (SSPR from 0.65 to 0.70) and PepYVMLV dataset (SSPR from 0.72 to 0.93). The second pattern

included datasets that experienced a drastic reduction in SSPR values, as seen in the BGMV dataset (from 0.24 to 0.15), PepGMV dataset (from 0.81 to 0.55), and CLCuGeV dataset (from 0.63 to 0.09). The third and more widespread pattern encompassed the remaining datasets, where SSPR values remained virtually unchanged.

The removal of recombinant DNA-A sequences reinforced the evidence of isolation by distance in the CLCuGeV dataset. Remarkably, the isolates that contributed significantly to the global incongruence in the CLCuGeV dataset were also recombinant, leading to an unexpected increase in the magnitude of SSPR values. It is important to note that approximately 48 sequences were removed from this dataset following the recombination analysis. Subsequently, a linear regression analysis between SSPR values and geodesic distances revealed a weak influence ($R^2 = 0.27$) of the coverage scale of collection sites on the support for isolation by distance (Figure 5b). We also recalculated nucleotide diversity indices to assess whether the genetic variation content had any impact on the evidence of isolation by distance. The results further support that the variation in SSPR values cannot be attributed to the genetic variation levels in the datasets (Figure 5c).

We subjected the recombinant-free datasets to the same slicing process as previously conducted to examine changes in the distribution patterns of the geographic signal across the full-length DNA-A sequences (Figure 5d). Once again, based on a clustering analysis, we were able to discern three major clusters. The first two clusters closely resembled those observed in the similar analysis conducted on datasets containing all DNA-A sequences, including recombinants. The most notable difference was in the third cluster, where the CLCuGeV dataset was added to the group containing datasets with more robust support of isolation by distance.

It is worth highlighting that even after the removal of recombinant DNA-A sequences, the geographic signal remained unevenly distributed along the sequences. Global congruence levels between genetic and spatial data were particularly enhanced in the sliding windows located at the 5' end of the BGMV DNA-A sequences. In the CLCuGeV dataset, genomic regions including sliding windows yielding SSPR values close to zero were interspersed with others showing SSPR values of 0.50 or higher. We also conducted regression analyses to assess whether SSPR values variation was affected by genetic variation levels along the DNA-A sequences. The removal of recombinant sequences resulted in reduced genetic variation levels in some sliding windows, such as

PaLCuCNV and CLCuGeV (Figure 5d). Nevertheless, we still observed regions with considerably higher genetic variation content, particularly at the 5' ends of DNA-A from SPLCV and BGMV isolates, indicating an uneven distribution of genetic variation levels. In summary, we did not observe any significant impact of genetic variation content on the evidence of isolation by distance along the DNA-A sequences, despite the removal of recombinant sequences. Even after eliminating recombinant sequences, we continued to observe the presence of sequences from distinct continents in ACMV (Supplementary Figure S3a), SLCuV (Supplementary Figure S3n), SPLCV (Supplementary Figure S3o), and TYLCV (Supplementary Figure S3s) datasets. This further supports the hypothesis that these particular isolates may indeed represent migrants.

Discussion

Phylogenetic studies offer a well-established method for analyzing viral populations and understanding their evolutionary patterns. Begomoviruses, highly prone to mutation and recombination, are frequently the focus of such investigations. However, inconsistencies in geminiviral genome phylogenies, particularly in the CP and REP regions, have been noted previously [55]. Our observations underscore the varying levels of variation across begomovirus genomes, emphasizing the need to employ a sliding windows approach when studying DNA-A sequences.

Assessing geographical structure is crucial in pathogen research as it aids in constructing a comprehensive epidemiological picture and tracing the virus's spread to specific locations. This analysis is conducted across different virus types, with some studies revealing clear geographical structuring, as seen in research on Wheat dwarf virus (WDV), a Mastrevirus [53]. In other cases, evidence of the Founder Effect has been found [56]. It's important to note that while phylogenetic trees for begomoviruses often exhibit geographic segregation, this alone doesn't imply geographic structure. For geographic structure to be present, there must be not only segregation but also a correlation between patristic distances and geographic distances among clusters.

Correlations between genetic and geographic distances among populations are frequently attested using the Mantel test, including in plant viruses and their insect vectors [22, 57, 58]. Our study marks the first application of the Procrustes test to investigate isolation by distance in begomoviruses, offering an alternative to the traditional Mantel

498 test. We introduce a likelihood methodology to estimate genetic distances, termed
499 patristic distances, for comparison with geographic distances.
500 Our analysis identifies sequences significantly contributing to geographic
501 structuring and those masking the geographic signal. We experimented with various
502 window sizes and step sizes, finding that 200 nt windows with a 20 nt step yield optimal
503 results. This sliding windows approach proves efficient, revealing substantial variations
504 in π and $\sum m^2_{XY}$ values across the genome in all analyses.
505 A study spanning 7 years (2005-2012) and encompassing BGMV isolates from 5
506 states plus the Federal District in Brazil revealed limited recombination events within this
507 species. It also demonstrated evidence of geographic structuring and significant genetic
508 differentiation among populations [59]. Notably, BGMV exhibited one of the highest
509 levels of geographic structure, with isolates that are on average, less than 1000 km apart,
510 showing evidence of isolation by distance. Recombination had minimal impact on the
511 results, as the only two isolates that showed significant residue values were not considered
512 recombinants. These isolates were from Pernambuco and clustered with MG isolates,
513 indicating migration, as stated by the author in the paper where they were first sequenced
514 [59].
515 Another study in Brazil, focusing on EuYMV isolates, found a significant
516 correlation between genetic divergence and geographic distance, suggesting the existence
517 of subpopulations within specific geographical regions [22]. However, our analysis did
518 not uncover evidence of geographic structure in EuYMV, despite some segregation by
519 collection site in phylogenetic trees, possibly due to the Founder Effect (Supplementary
520 Figure 1i).
521 A study on the geographic spread of TYLCV revealed a reasonably strong spatial
522 structure for the virus, as depicted in the maximum clade credibility (MCC) tree [60].
523 However, incongruities were observed in our TYLCV tree (Figure 4a), where isolates
524 collected at different distances displayed varying residue values. This variability suggests
525 a reasonable level of geographic structuring, with instances of high residue values
526 indicating the presence of recombination, potentially influencing the presence of isolation
527 by distance. Interestingly, some sequences with high residue values lacked evidence of
528 recombination, likely representing migrant isolates.
529 The papers analyzing isolates with the highest residual values, such as the
530 American isolates EF210554 and EF110890 from TYLCV, suggest they were introduced

to their respective regions through transplantation [61, 62], confirming suspicions of migration. Similarly, a study on isolates from Costa Rica (KY064016, KF533857, and KF533856) found them clustering closer to Asian and Mexican isolates, supporting our findings [63]. The Hawaiian isolate appears to be a migrant as well, clustering near East Asian isolates, with no information on its introduction [64]. Likewise, the Guatemalan isolate with high residue values (GU355941) likely originated from the Caribbean Region, consistent with migrant status [65]

The observed incongruities may stem from TYLCV's recent emergence in the New World, initially identified in Israel in 1939 [4] but not reported in the Americas until the 1990s [66–70]. TYLCV continues to spread, with recent reports of isolates in New York likely introduced via transplants, indicating ongoing migration events [71].

Such inconsistencies can also be explained by the strong human intervention in the dispersion of this virus, through the international trade of tomato seedlings, which can effectively affect the distance isolation of these TYLCV isolates, some studies even pointing to the possibility of transmission through the seed [72, 73].

TYLCV serves as a prime example of an Old-World virus introduced to the New World, a phenomenon also observed with CLCuGeV, an African begomovirus introduced to Southern Texas in 2018 [74]. Conversely, SLCuV, initially reported in the USA in 1977 [75], made its way from the New World to the Old World, reaching Israel in 2003 [76]. These cross-hemispheric introductions can distort geographic signals in trees, potentially masking isolation by distance.

Similar events occur on a smaller scale, such as the introduction of African viruses like ACMV, with one of the earliest reports in 1894 in Tanzania [77], to Asia, more specifically, Pakistan in 2008 [78] evidenced by the grouping of Pakistani isolates with those from Sub-Saharan Africa (Supplementary Figure 1a, 3a). PepYVMLV, originating in Africa, was later introduced to China [79], highlighting this phenomenon.

The pattern observed with SPLCV differs from expectations due to its transmission via propagation material, often facilitated by international trade in sweet potato seedlings. While previous studies suggested a lack of geographic structure due to high gene flow [80–82], significant $\sum m^2_{XY}$ values indicate some degree of geographic isolation, particularly evident in trees without recombinant sequences (Supplementary Figure 7o). Nonetheless, inconsistencies in tree topology persist, reflecting human-mediated translocation of the virus across continents.

Lapidot et al. [83] found that SLCuV isolates within the same country exhibit low genetic variability, suggesting frequent virus migration within a country. However, the inclusion of Egyptian sequences, collected far from other countries, heavily influenced this result. When the analysis focused solely on sequences from Israel, Jordan, Lebanon, and Palestine, the correlation between geographic and genetic distance decreased significantly.

Our observation of highly variable $\sum m^2_{XY}$ values across the genome indicates varying degrees of geographic structuring. This variability may be attributed to the recent introduction of the virus into the Old World. Recombination affects patristic distance calculations in trees, leading to altered branch lengths and potentially distorting distance isolation signals [54].

While most regression analyses did not show a predictive relationship between genetic variation and $\sum m^2_{XY}$, EACMV exhibited a notable R^2 value of 0.6, suggesting a link between variation content and $\sum m^2_{XY}$ (Figure 2d). This relationship appears to be influenced by recombination events, as evidenced by π values in affected windows.

Similar observations were made with PepYVMLV, where recombination events involving isolates from China correlated with geographic structuring signals. Removal of these recombinant sequences eliminated the geographic signal in PepYVMLV analyses, indicating their significant contribution to the central region of the genome's geographic signal (Figures 5a and 5b).

Recombination plays a significant role in the genetic variability and evolution of begomoviruses [84]. The phenomena discussed here highlight how recombination can impact tree topology and evidence of geographic structuring, sometimes reinforcing geographic signals and other times obscuring them, depending on the major and minor parents involved.

Robust results from our study indicate the presence of isolation by distance in BGMV and TYLCV cases, independent of geographic scale. However, this evidence was not found in the other 22 begomovirus species examined.

Most isolates showing high $\sum m^2_{XY}$ in BGMV and TYLCV (Figure 3b and 4a) were either recombinant sequences or previously identified migrants, underscoring how these phenomena can distort tree topology and mask isolation by distance signals.

Understanding these factors is crucial for managing virus epidemics and developing effective control strategies. It becomes clear that studying isolation by

distance under conditions of natural dissemination, with minimal human intervention, would provide the most accurate insights.

References

1. **Zerbini FM, Briddon RW, Idris A, Martin DP, Moriones E, *et al.*** ICTV Virus Taxonomy Profile: Geminiviridae. *Journal of General Virology* 2017;98:131–133. https://doi.org/10.1099/jgv.0.000738
2. **Patil BL, Fauquet CM**. Cassava mosaic geminiviruses: actual knowledge and perspectives. *Mol Plant Pathol* 2009;10:685–701. https://doi.org/10.1111/j.1364-3703.2009.00559.x
3. **Hema M, Sreenivasulu P, Patil BL, Kumar PL, Reddy DVR**. Chapter Nine - Tropical Food Legumes: Virus Diseases of Economic Importance and Their Control. In: Loebenstein G, Katis N (editors). *Advances in Virus Research*. Academic Press. pp. 431–505. https://doi.org/10.1016/B978-0-12-801246-8.00009-3
4. **Picó B, Díez MJ, Nuez F**. Viral diseases causing the greatest economic losses to the tomato crop. II. The Tomato yellow leaf curl virus — a review. *Sci Hortic* 1996;67:151–196. https://doi.org/10.1016/S0304-4238(96)00945-4
5. **Harrison BD, Swanson MM, Fargette D**. Begomovirus coat protein: serology, variation and functions. *Physiol Mol Plant Pathol* 2002;60:257–271. https://doi.org/10.1006/pmpp.2002.0404
6. **Bisaro DM**. Silencing suppression by geminivirus proteins. *Virology* 2006;344:158–168. https://doi.org/10.1016/j.virol.2005.09.041
7. **Laufs J, Jupin I, David C, Schumacher S, Heyraud-Nitschke F, *et al.*** Geminivirus replication: Genetic and biochemical characterization of rep protein function, a review. *Biochimie* 1995;77:765–773. https://doi.org/10.1016/0300-9084(96)88194-6
8. **Sanderfoot AA, Lazarowitz SG**. Getting it together in plant virus movement: cooperative interactions between bipartite geminivirus movement proteins. *Trends Cell Biol* 1996;6:353–358. https://doi.org/10.1016/0962-8924(96)10031-3
9. **Fiallo-Olivé E, Lett J-M, Martin DP, Roumagnac P, Varsani A, *et al.*** ICTV Virus Taxonomy Profile: Geminiviridae 2021. *Journal of General Virology* 2021;102:001696. https://doi.org/10.1099/jgv.0.001696
10. **Duffy S, Holmes EC**. Validation of high rates of nucleotide substitution in geminiviruses: phylogenetic evidence from East African cassava mosaic viruses. *Journal of General Virology* 2009;90:1539–1547. https://doi.org/10.1099/vir.0.009266-0
11. **Lefeuvre P, Martin DP, Hoareau M, Naze F, Delatte H, *et al.*** Begomovirus 'melting pot' in the south-west Indian Ocean islands: molecular diversity and evolution through recombination. *Journal of General Virology* 2007;88:3458–3468. https://doi.org/10.1099/vir.0.83252-0
12. **Seal SE, vandenBosch F, Jeger MJ**. Factors Influencing Begomovirus Evolution and Their Increasing Global Significance: Implications for Sustainable Control. *CRC Crit Rev Plant Sci* 2006;25:23–46. https://doi.org/10.1080/07352680500365257
13. **Xavier CAD, Godinho MT, Mar TB, Ferro CG, Sande OFL, *et al.*** Evolutionary dynamics of bipartite begomoviruses revealed by complete genome analysis. *Mol Ecol* 2021;30:3747–3767. https://doi.org/10.1111/mec.15997
14. **Esmaeili M, Heydarnejad J, Massumi H, Varsani A**. Analysis of watermelon chlorotic stunt virus and tomato leaf curl Palampur virus mixed and pseudo-

644 recombination infections. *Virus Genes* 2015;51:408–416.
645 https://doi.org/10.1007/s11262-015-1250-5
646 15. **Idris AM, Brown JK**. Cotton leaf crumple virus Is a Distinct Western
647 Hemisphere Begomovirus Species with Complex Evolutionary Relationships Indicative
648 of Recombination and Reassortment. *Phytopathology* 2004;94:1068–1074.
649 https://doi.org/10.1094/PHYTO.2004.94.10.1068
650 16. **García-Andrés S, Tomás DM, Sánchez-Campos S, Navas-Castillo J,**
651 **Moriones E**. Frequent occurrence of recombinants in mixed infections of tomato yellow
652 leaf curl disease-associated begomoviruses. *Virology* 2007;365:210–219.
653 https://doi.org/10.1016/j.virol.2007.03.045
654 17. **Monci F, Sánchez-Campos S, Navas-Castillo J, Moriones E**. A Natural
655 Recombinant between the Geminiviruses Tomato yellow leaf curl Sardinia virus and
656 Tomato yellow leaf curl virus Exhibits a Novel Pathogenic Phenotype and Is Becoming
657 Prevalent in Spanish Populations. *Virology* 2002;303:317–326.
658 https://doi.org/10.1006/viro.2002.1633
659 18. **García-Andrés S, Monci F, Navas-Castillo J, Moriones E**. Begomovirus
660 genetic diversity in the native plant reservoir Solanum nigrum: evidence for the presence
661 of a new virus species of recombinant nature. *Virology* 2006;350:433–442.
662 https://doi.org/10.1016/j.virol.2006.02.028
663 19. **Lefeuvre P, Moriones E**. Recombination as a motor of host switches and virus
664 emergence: geminiviruses as case studies. *Curr Opin Virol* 2015;10:14–19.
665 https://doi.org/10.1016/j.coviro.2014.12.005
666 20. **Rocha CS, Castillo-Urquiza GP, Lima ATM, Silva FN, Xavier CAD, *et al.***
667 Brazilian Begomovirus Populations Are Highly Recombinant, Rapidly Evolving, and
668 Segregated Based on Geographical Location. *J Virol* 2013;87:5784–5799.
669 https://doi.org/10.1128/JVI.00155-13
670 21. **Prasanna HC, Sinha DP, Verma A, Singh M, Singh B, *et al.*** The population
671 genomics of begomoviruses: global scale population structure and gene flow. *Virol J*
672 2010;7:220. https://doi.org/10.1186/1743-422X-7-220
673 22. **Mar TB, Xavier CAD, Lima ATM, Nogueira AM, Silva JCF, *et al.*** Genetic
674 variability and population structure of the New World begomovirus Euphorbia yellow
675 mosaic virus. *Journal of General Virology* 2017;98:1537–1551.
676 https://doi.org/10.1099/jgv.0.000784
677 23. **Lima ATM, Sobrinho RR, González-Aguilera J, Rocha CS, Silva SJC, *et al.***
678 Synonymous site variation due to recombination explains higher genetic variability in
679 begomovirus populations infecting non-cultivated hosts. *Journal of General Virology*
680 2013;94:418–431. https://doi.org/10.1099/vir.0.047241-0
681 24. **Diniz-Filho JAF, Soares TN, Lima JS, Dobrovolski R, Landeiro VL, *et al.***
682 Mantel test in population genetics. *Genet Mol Biol*;36. https://doi.org/10.1590/S1415-
683 47572013000400002
684 25. **Rodelo-Urrego M, Pagán I, González-Jara P, Betancourt M, Moreno-**
685 **Letelier A, *et al.*** Landscape heterogeneity shapes host-parasite interactions and results in
686 apparent plant–virus codivergence. *Mol Ecol* 2013;22:2325–2340.
687 https://doi.org/10.1111/mec.12232
688 26. **Lapidot M, Gelbart D, Gal-On A, Sela N, Anfoka G, *et al.*** Frequent migration
689 of introduced cucurbit-infecting begomoviruses among Middle Eastern countries. *Virol J*
690 2014;11:181. https://doi.org/10.1186/1743-422X-11-181
691 27. **Wright S**. Isolation by distance. *Genetics* 1943;28:114–138.
692 https://doi.org/10.1093/genetics/28.2.114

28. **Wright S**. Breeding Structure of Populations in Relation to Speciation. *Am Nat* 1940;74:232–248. https://doi.org/10.1086/280891
29. **Lam TT-Y, Ip HS, Ghedin E, Wentworth DE, Halpin RA, *et al.*** Migratory flyway and geographical distance are barriers to the gene flow of influenza virus among North American birds. *Ecol Lett* 2012;15:24–33. https://doi.org/10.1111/j.1461-0248.2011.01703.x
30. **Fourment M, Darling AE, Holmes EC**. The impact of migratory flyways on the spread of avian influenza virus in North America. *BMC Evol Biol* 2017;17:118. https://doi.org/10.1186/s12862-017-0965-4
31. **Carrel MA, Emch M, Jobe RT, Moody A, Wan X-F**. Spatiotemporal Structure of Molecular Evolution of H5N1 Highly Pathogenic Avian Influenza Viruses in Vietnam. *PLoS One* 2010;5:e8631. https://doi.org/10.1371/journal.pone.0008631
32. **Gorrochotegui-Escalante N, Munoz ML, Fernandez-Salas I, Beaty BJ, Black WC**. Genetic isolation by distance among Aedes aegypti populations along the northeastern coast of Mexico. *Am J Trop Med Hyg* 2000;62:200–209. https://doi.org/10.4269/ajtmh.2000.62.200
33. **Edgar RC**. Muscle5: High-accuracy alignment ensembles enable unbiased assessments of sequence homology and phylogeny. *Nat Commun* 2022;13:6968. https://doi.org/10.1038/s41467-022-34630-w
34. **Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T**. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;25:1972–1973. https://doi.org/10.1093/bioinformatics/btp348
35. **Martin DP, Murrell B, Golden M, Khoosal A, Muhire B**. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evol*;1. Epub ahead of print 1 March 2015. https://doi.org/10.1093/ve/vev003
36. **Nei M**. Bibliography. In: *Molecular Evolutionary Genetics*. New York Chichester, West Sussex: Columbia University Press. pp. 433–496.
37. **Canty AJ**. Resampling Methods in {R}: The boot Package. *R News*;2.
38. **R Core Team**. R: A language and environment for statistical computing. 55.
39. **Gu Z, Eils R, Schlesner M**. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* 2016;32:2847–2849. https://doi.org/10.1093/bioinformatics/btw313
40. **Galili T**. dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics* 2015;31:3718–3720. https://doi.org/10.1093/bioinformatics/btv428
41. **Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ**. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* 2015;32:268–274. https://doi.org/10.1093/molbev/msu300
42. **Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS**. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017;14:587–589. https://doi.org/10.1038/nmeth.4285
43. **Peng RD**. simpleboot: Simple Bootstrap Routines. *https://CRANR-project.org/package=simpleboot*.
44. **Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y**. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2017;8:28–36. https://doi.org/10.1111/2041-210X.12628
45. **Balbuena JA, Míguez-Lozano R, Blasco-Costa I**. PACo: A Novel Procrustes Application to Cophylogenetic Analysis. *PLoS One* 2013;8:e61048. https://doi.org/10.1371/journal.pone.0061048

743 46. **Hutchinson MC, Cagua EF, Balbuena JA, Stouffer DB, Poisot T**. paco:
744 implementing Procrustean Approach to Cophylogeny in R. *Methods Ecol Evol*
745 2017;8:932–940. https://doi.org/10.1111/2041-210X.12736
746 47. **Paradis E, Schliep K**. ape 5.0: an environment for modern phylogenetics and
747 evolutionary analyses in R. *Bioinformatics* 2019;35:526–528.
748 https://doi.org/10.1093/bioinformatics/bty633
749 48. **Seong JC, Choi J**. GEODIST: A C++ program for calculating geodesic distances
750 with a shapefile. *Comput Geosci* 2007;33:705–708.
751 https://doi.org/10.1016/j.cageo.2006.09.005
752 49. **Wickham H**. Ggplot2. *WIREs Comput Stat* 2011;3:180–185.
753 https://doi.org/10.1002/wics.147
754 50. **Aphalo PJ**. ggpmisc: An R package.
755 51. **Arel-Bundock V, Enevoldsen N, Yetman CJ**. countrycode: An R package to
756 convert country names and country codes. *The Journal of Open Source Software*
757 2018;3:848. https://doi.org/10.21105/joss.00848
758 52. **Pita JS, Fondong VN, Sangaré A, Otim-Nape GW, Ogwal S, *et al.*** Recombination, pseudorecombination and synergism of geminiviruses are determinant
759 keys to the epidemic of severe cassava mosaic disease in Uganda. *Journal of General*
760 *Virology*;82. Epub ahead of print 2001. https://doi.org/10.1099/0022-1317-82-3-655
761 53. **Wu B, Shang X, Schubert J, Habekuß A, Elena SF, *et al.*** Global-scale
762 computational analysis of genomic sequences reveals the recombination pattern and
763 coevolution dynamics of cereal-infecting geminiviruses. *Sci Rep* 2015;5:8153.
764 https://doi.org/10.1038/srep08153
765 54. **Schierup MH, Hein J**. Consequences of Recombination on Traditional
766 Phylogenetic Analysis. *Genetics* 2000;156:879–891.
767 https://doi.org/10.1093/genetics/156.2.879
768 55. **Claverie S, Bernardo P, Kraberger S, Hartnady P, Lefeuvre P, *et al.*** Chapter
769 Three - From Spatial Metagenomics to Molecular Characterization of Plant Viruses: A
770 Geminivirus Case Study. In: Malmstrom CM (editor). *Advances in Virus Research*.
771 Academic Press. pp. 55–83. https://doi.org/10.1016/bs.aivir.2018.02.003
772 56. **Sánchez-Campos S, Díaz JA, Monci F, Bejarano ER, Reina J, *et al.*** High
773 Genetic Stability of the Begomovirus Tomato yellow leaf curl Sardinia virus in Southern
774 Spain Over an 8-Year Period. *Phytopathology* 2002;92:842–849.
775 https://doi.org/10.1094/PHYTO.2002.92.8.842
776 57. **Maachi A, Donaire L, Hernando Y, Aranda MA**. Genetic Differentiation and
777 Migration Fluxes of Viruses from Melon Crops and Crop Edge Weeds. *J Virol*
778 2022;96:e00421-22. https://doi.org/10.1128/jvi.00421-22
779 58. **T P, Kranthi S, P RK, Kumar R, Suke R, *et al.*** Mitochondrial COI based genetic
780 diversity and phylogeographic structure of whitefly Bemisia tabaci (Gennadius) on cotton
781 in India. *Int J Trop Insect Sci* 2021;41:1543–1554. https://doi.org/10.1007/s42690-020-
782 00354-x
783 59. **Sobrinho RR, Xavier CAD, Pereira HM de B, Lima GS de A, Assunção IP,
784 *et al.*** Contrasting genetic structure between two begomoviruses infecting the same
785 leguminous hosts. *Journal of General Virology* 2014;95:2540–2552.
786 https://doi.org/10.1099/vir.0.067009-0
787 60. **Mabvakure B, Martin DP, Kraberger S, Cloete L, van Brunschot S, *et al.*** Ongoing geographical spread of Tomato yellow leaf curl virus. *Virology* 2016;498:257–
788 264. https://doi.org/10.1016/j.virol.2016.08.033
789
790

61. **Isakeit T, Idris AM, Sunter G, Black MC, Brown JK**. Tomato yellow leaf curl virus in Tomato in Texas, Originating from Transplant Facilities. *Plant Dis* 2007;91:466. https://doi.org/10.1094/PDIS-91-4-0466A
62. **Idris AM, Guerrero JC, Brown JK**. Two Distinct Isolates of Tomato yellow leaf curl virus Threaten Tomato Production in Arizona and Sonora, Mexico. *Plant Dis* 2007;91:910. https://doi.org/10.1094/PDIS-91-7-0910C
63. **Barboza N, Blanco-Meneses M, Esker P, Moriones E, Inoue-Nagata AK**. Distribution and diversity of begomoviruses in tomato and sweet pepper plants in Costa Rica. *Annals of Applied Biology* 2018;172:20–32. https://doi.org/10.1111/aab.12398
64. **Melzer MJ, Ogata DY, Fukuda SK, Shimabuku R, Borth WB, *et al.*** First Report of Tomato yellow leaf curl virus in Hawaii. *Plant Dis* 2010;94:641. https://doi.org/10.1094/PDIS-94-5-0641B
65. **Salati R, Shorey M, Briggs A, Calderon J, Rojas MR, *et al.*** First Report of Tomato yellow leaf curl virus Infecting Tomato, Tomatillo, and Peppers in Guatemala. *Plant Dis* 2010;94:482. https://doi.org/10.1094/PDIS-94-4-0482C
66. **Duffy S, Holmes EC**. Multiple Introductions of the Old World Begomovirus *Tomato yellow leaf curl virus* into the New World. *Appl Environ Microbiol* 2007;73:7114–7117. https://doi.org/10.1128/AEM.01150-07
67. **Zubiaur YM, Zabalgogeazcoa I, De Blas C, Sanchez F, Peralta EL, *et al.*** Geminiviruses Associated with Diseased Tomatoes in Cuba. *Journal of Phytopathology* 1996;144:277–279. https://doi.org/10.1111/j.1439-0434.1996.tb01529.x
68. **Czosnek H, Laterrot H**. A worldwide survey of tomato yellow leaf curl viruses. *Arch Virol* 1997;142:1391–1406. https://doi.org/10.1007/s007050050168
69. **Polston JE, McGovern RJ, Brown LG**. Introduction of Tomato Yellow Leaf Curl Virus in Florida and Implications for the Spread of This and Other Geminiviruses of Tomato. *Plant Dis* 1999;83:984–988. https://doi.org/10.1094/PDIS.1999.83.11.984
70. **Ascencio-Ibáñez JT, Diaz-Plaza R, Méndez-Lozano J, Monsalve-Fonnegra ZI, Argüello-Astorga GR, *et al.*** First Report of Tomato Yellow Leaf Curl Geminivirus in Yucatán, México. *Plant Dis* 1999;83:1178. https://doi.org/10.1094/PDIS.1999.83.12.1178A
71. **Perry KL**. An Anomalous Detection of Tomato Yellow Leaf Curl Virus in Tomato in New York State. *Plant Dis* 2021;105:3312. https://doi.org/10.1094/PDIS-02-21-0356-PDN
72. **Kil E-J, Kim S, Lee Y-J, Byun H-S, Park J, *et al.*** Tomato yellow leaf curl virus (TYLCV-IL): a seed-transmissible geminivirus in tomatoes. *Sci Rep* 2016;6:19013. https://doi.org/10.1038/srep19013
73. **Kil E-J, Park J, Choi E-Y, Byun H-S, Lee K-Y, *et al.*** Seed transmission of Tomato yellow leaf curl virus in sweet pepper (Capsicum annuum). *Eur J Plant Pathol* 2018;150:759–764. https://doi.org/10.1007/s10658-017-1304-8
74. **Villegas C, Ramos-Sobrinho R, Jifon JL, Keith C, Al Rwahnih M, *et al.*** First Report of Cotton leaf curl Gezira virus and Its Associated Alphasatellite and Betasatellite from Disease Affected Okra Plants in the United States. *Plant Dis* 2019;103:3291. https://doi.org/10.1094/PDIS-06-19-1175-PDN
75. **Flock RA, Mayhew DE**. Squash leaf curl, a new disease of cucurbits in California. *Plant Dis* 1981;65:75–76. https://doi.org/10.1094/PD-65-75
76. **Antignus Y, Lachman O, Pearlsman M, Omer S, Yunis H, *et al.*** Squash leaf curl geminivirus – a new illegal immigrant from the Western Hemisphere and a threat to cucurbit crops in Israel. *Phytoparasitica*;31.
77. **Legg JP, Fauquet CM**. Cassava mosaic geminiviruses in Africa. *Plant Mol Biol* 2004;56:585–599. https://doi.org/10.1007/s11103-004-1651-7

78. **Nawaz-ul-Rehman MS, Briddon RW, Fauquet CM**. A Melting Pot of Old World Begomoviruses and Their Satellites Infecting a Collection of Gossypium Species in Pakistan. *PLoS One* 2012;7:e40050. https://doi.org/10.1371/journal.pone.0040050
79. **Zhang J, Jia SP, Yang CX, Liu Z, Wu ZJ**. Detection and molecular characterization of three begomoviruses associated with yellow vein disease of Eclipta prostata in Fujian, China. *Journal of Plant Pathology* 2015;97:161–165.
80. **Zhang C, Sun H, Xie Y, Yang D, Zhang M, *et al.*** The genetic structure and recombination analyses of Sweetpotato leaf curl virus (SPLCV) population in China. *Journal of Plant Diseases and Protection* 2020;127:741–751. https://doi.org/10.1007/s41348-020-00348-4
81. **Aimone CD, Lavington E, Hoyer JS, Deppong DO, Mickelson-Young L, *et al.*** Population diversity of cassava mosaic begomoviruses increases over the course of serial vegetative propagation. *Journal of General Virology*;102. Epub ahead of print 2021. DOI: https://doi.org/10.1099/jgv.0.001622
82. **Ferro CG, Zerbini FM, Navas-Castillo J, Fiallo-Olivé E**. Revealing the Complexity of Sweepovirus-Deltasatellite-Plant Host Interactions: Expanded Natural and Experimental Helper Virus Range and Effect Dependence on Virus-Host Combination. *Microorganisms* 2021;9:1018. https://doi.org/10.3390/microorganisms9051018
83. **Lapidot M, Gelbart D, Gal-On A, Sela N, Anfoka G, *et al.*** Frequent migration of introduced cucurbit-infecting begomoviruses among Middle Eastern countries. *Virol J* 2014;11:181. https://doi.org/10.1186/1743-422X-11-181
84. **Lefeuvre P, Lett J-M, Varsani A, Martin DP**. Widely Conserved Recombination Patterns among Single-Stranded DNA Viruses . *J Virol* 2009;83:2697–2707. https://doi.org/10.1128/JVI.02152-08

Figure legends

Figure 1. (a) Number of DNA-A sequences (depicted as blue bars) from each begomovirus species dataset included in our study and the average distances separating pairs of sampling collection sites (depicted as red bars) calculated using the R package geodist [48]. Non-parametric 95% bootstrap confidence intervals for the mean geodesic distances were estimated using the R package boot. **(b)** Sum of squared Procrustes residuals (SSPR, $\sum m^2_{XY}$) calculated using the PACo package [46] (red bars) and nucleotide diversity indices (blue bars) with their non-parametric 95% bootstrap confidence intervals calculated for each begomovirus species dataset based on full-length DNA-A sequences. **(c)** Linear regression between $\sum m^2_{XY}$ and geodesic distances, and (**d**) between $\sum m^2_{XY}$ and nucleotide diversity indices calculated for each begomovirus species dataset based on full-length DNA-A sequences. The regression models were calculated using the R package ggpmisc [50].

Figure 2. (**a**) Heatmap representing the sum of squares Procrustes residuals (SSPR, $\sum m^2_{XY}$) calculated for each of the 200-nucleotide sliding windows using the R package PACo [46]. All SSPR values were significant at $p < 0.01$. Each column of sliding windows was composed of homologous DNA-A segments sliced from a dataset containing all full-length DNA-A sequences used in this study. Datasets showing similar patterns of SSPR values distribution were grouped by means of a dendrogram constructed from Euclidean distances computed between all pairs of datasets using the dendextend R package [40]. The dendrogram was partitioned into three clusters using the k-means algorithm available in R software. The branches colored in green, blue, and red represent the three clusters determined using k-means. (**b**) Heatmap representing the nucleotide diversity values calculated for each of the sliding windows. The datasets were listed in the same order as the heatmap in (a). Linear regressions were performed between the SSPR values and nucleotide diversity indices calculated for each of the sliding windows obtained from slicing the full-DNA-A sequences of the BGMV (**c**), EACMV (**d**), and TYLCV (**e**) datasets. The regression models were determined using the ggpmisc package in R [50].

Figure 3. (**a**) Map displaying all collection sites of BGMV isolates in three Brazilian geographical regions: Midwest (depicted in orange), Southeast (light blue), and Northeast (green). Curves connecting all pairs of collection sites are shown in colors corresponding to the geographical distance separating the collection sites. Geodesic distances were calculated from geographical coordinates retrieved from GenBank or related scientific publications. (**b**) Maximum Likelihood phylogenetic tree reconstructed for full-length DNA-A sequences of BGMV isolates. (**c**) Line plot presenting the sum of squared Procrustes residuals (SSPR, $\sum m^2_{XY}$) along the sliding windows obtained by slicing the full-length DNA-A sequences of BGMV isolates. The dashed red line is positioned to represent the SSPR value of 0.3. Maximum likelihood trees were constructed from the alignment columns mapped within sliding windows that yielded SSPR lower than 0.3 (**d**), and those columns mapped in sliding windows that yielded values greater than 0.3 (**e**). Tip labels of all ML trees are color-coded according to the regions where the collection sites of BGMV isolates are located. The Procrustes residuals are presented as colored bars for each of the isolates according to the color scale shown in (**b**).

Figure 4. (**a**) Maximum likelihood phylogenetic tree constructed for full-length DNA-A-like sequences of TYLCV isolates. Tip points and labels are color-coded according to the geographical sub-region in which the isolates were collected. Tip labels include, in addition to the GenBank accession number, the standardized 3-letter country retrieved from the R package countrycode (AUS = Australia, AZE = Azerbaijan, CHN = China, CUB = Cuba, DOM = Dominican Republic, EGY = Egypt, ESP = Spain, EST = Estonia, FRA = France, GBR = United Kingdom, GRD = Grenada, GTM = Guatemala, IND = India, IRQ = Iraq, IRN = Iran, ITA = Italy, JPN = Japan, JOR = Jordan, KOR = South Korea, KWT = Kuwait, LBN = Lebanon, MAR = Morocco, MEX = Mexico, MUS = Mauritius, NLD = Netherlands, NCL = New Caledonia, OMN = Oman, PRI = Puerto Rico, PRT = Portugal, REU = Reunion, SAU = Saudi Arabia, SWE = Sweden, TTO = Trinidad and Tobago, TUN = Tunisia, TUR = Turkey, USA = United States of America, VEN = Venezuela). Procrustes residuals are also represented as colored bars according to the provided scale. Due to the large size of the ML tree, two branches containing a large number of isolates collected in the United States and China have been collapsed and are represented as large tip points colored according to the geographical sub-regions where these two countries are located. (**b**) Heatmap representing the patristic distances between all pairs of TYLCV isolates analyzed in this study. A thumbnail of the ML tree presented in (**a**) is positioned to the left of the heatmap to indicate the relative positions of isolate groups from each of the geographical sub-regions. (**c**) Map displaying all collection sites (represented as red crosses) from which TYLCV isolates were obtained. For simplicity (due to the large number of collection sites), we represented as colored points only the centroid coordinates of sub-geographical regions where isolates were sampled. Curves connecting the points are color-coded according to the geodesic distances separating the geographical sub-regions. Note that these distances represented on the map do not necessarily accurately reflect those between the actual collection sites, which may be lower or higher depending on the precise locations of the collection sites.

Figure 5. (**a**) Sum of squared Procrustes residuals (SSPR, ∑m2XY) calculated using the PACo package [46] (red bars) and nucleotide diversity indices (blue bars) with their non-parametric 95% bootstrap confidence intervals calculated for each begomovirus species dataset based on non-recombinant full-length DNA-A sequences. Datasets showing similar patterns of SSPR values distribution were grouped by means of a dendrogram

constructed from Euclidean distances computed between all pairs of datasets using the dendextend R package [40]. The dendrogram was partitioned into three clusters using the k-means algorithm available in R software. The branches colored in green, blue, and red represent the three clusters determined using k-means. **(b)** Linear regression between SSPR ($\sum m^2_{XY}$) values and geodesic distances, and (**c**) between SSPR values and nucleotide diversity indices calculated for each begomovirus species dataset based on non-recombinant full-length DNA-A sequences. The regression models were calculated using the R package ggpmisc [50]. (**d**) Heatmap representing the sum of squares Procrustes residuals (SSPR, $\sum m^2_{XY}$) calculated for each of the 200-nucleotide sliding windows using the R package PACo [46]. All SSPR values were significant at $p < 0.01$. (**b**) Heatmap representing the nucleotide diversity values calculated for each of the sliding windows. The datasets were listed in the same order as the heatmap in (a).

Supplementary Table S1. Sequences of begomoviruses retrieved from GenBank used in this study.

Supplementary Table S2. Recombination events detected by RDP4 [35] in each begomovirus species datasets.

Supplementary Figure S1. Phylogenetic trees of the full length DNA-A of the complete dataset built using iqtree [41] with 5.000 ultrafast bootstrap replications, edited using the ggtree package [44] in addition to the residue values that were calculated using the jackknife method with 1000 replications for each of the isolates (a) ACMV, (b) AYVV, (c) BYVMV, (d) ChiLCV, (e) CLCuGeV, (f) CLCuMuV, (g) EACMKV, (h) EACMV, (i) EuYMV, (j) MYMIV, (k) PaLCuCNV, (l) PepGMV, (m) PepYVMLV, (n) SACMV, (o) SLCCNV, (p) SLCuV, (q) SPLCV, (r) TbCSV, (s) ToLCNDV, (t) ToLCTV, (u) ToSRV.

Supplementary Figure S2. Heatmaps representing the patristic distances for each isolate of the complete dataset, the heatmaps were built using the ComplexHeatmap package [39] in R software [38]. (a) ACMV, (b) AYVV, (c) BGMV, (d) BYVMV, (e) ChiLCV, (f) CLCuGeV, (g) CLCuMuV, (h) EACMKV, (i) EACMV, (j) EuYMV, (k) MYMIV, (l) PaLCuCNV, (m) PepGMV, (n) PepYVMLV, (o) SACMV, (p) SLCCNV, (q) SLCuV, (r) SPLCV, (s) TbCSV, (t) ToLCNDV, (u) ToLCTV, (v) ToSRV.

Supplementary Figure S3. Line plots presenting the nucleotide diversity values and confidence intervals along the full-length DNA-A sequences for each begomovirus species dataset. The graphs were plotted using ggplot in R software.

Supplementary Figure S4. Linear regression between $\sum m^2_{XY}$ and nucleotide diversity indices calculated for each begomovirus species dataset based on full-length DNA-A sequences. The regression models were calculated using the R package ggpmisc [50].

Supplementary Figure S5. Phylogenetic trees of the sliding Windows of 200 nucleotides of the complete dataset built using iqtree [41] with 5.000 bootstrap replications, edited using the ggtree package [44], in addition to the residue values that were calculated using the jackknife method with 1000 replications for each of the isolates (a) SLCuV sliding window starting at 2300 and ending at 2500. (b) SLCuV sliding window starting at 60 and ending at 260. (c) EACMV sliding window starting at 700 and ending at 900. (d) EACMV sliding window starting at 1680 and ending at 1880. (e) SLCCNV sliding window starting at 100 and ending at 300. (f) SLCCNV sliding window starting at 1100 and ending at 1300. (g) EuYMV sliding window starting at 1720 and ending at 1920. (h) SPLCV sliding window starting at 1900 and ending at 2100. (i) PepYVMLV sliding window starting at 1260 and ending at 1460. (j) PepYVMLV sliding window starting at 760 and ending at 960. (k) ACMV sliding window starting at 100 and ending at 300.

Supplementary Figure S6. Heatmaps representing the patristic distances for each isolate of BGMV, the heatmaps were built using the ComplexHeatmap package in R software. (a) Heatmap composed of all alignment columns that yielded SSPR values below 0.3, (b) Heatmap composed of all alignment columns that yielded SSPR values above 0.3.

Supplementary Figure S7. Phylogenetic trees of the full length DNA-A of the data set free of recombination events detectable by RDP4 built using iqtree with 5.000 bootstrap replications, edited using the ggtree package, in addition to the residue values that were calculated using the jackknife method with 1000 replications for each of the isolates (a) ACMV, (b) BGMV, (c) ChiLCV, (d) CLCuGeV, (e) EACMKV, (f) EACMV, (g) EuYMV, (h) MYMIV, (i) PaLCuCNV, (j) PepGMV, (k) PepYVMLV, (l) SACMV, (m)

SLCCNV, (n) SLCuV, (o) SPLCV, (p) ToLCNDV, (q) ToLCTV, (r) ToSRV, (s) TYLCV.

Supplementary Figure S8. Heatmaps representing the patristic distances for each isolate of the complete dataset, the heatmaps were built using the ComplexHeatmap package in R software. (a) ACMV, (b) BGMV, (c) ChiLCV, (d) CLCuGeV, (e) EACMKV, (f) EACMV, (g) EuYMV, (h) MYMIV, (i) PaLCuCNV, (j) PepGMV, (k) PepYVMLV, (l) SACMV, (m) SLCCNV, (n) SLCuV, (o) SPLCV, (p) ToLCNDV, (q) ToLCTV, (r) ToSRV.

Supplementary Figure S9. Linear regression between $\sum m^2_{XY}$ and nucleotide diversity indices calculated along the full-length DNA-A sequences for each begomovirus species dataset free of recombination events detectable by RDP4 [35].The regression models were calculated using the R package ggpmisc [50].

Figure 1

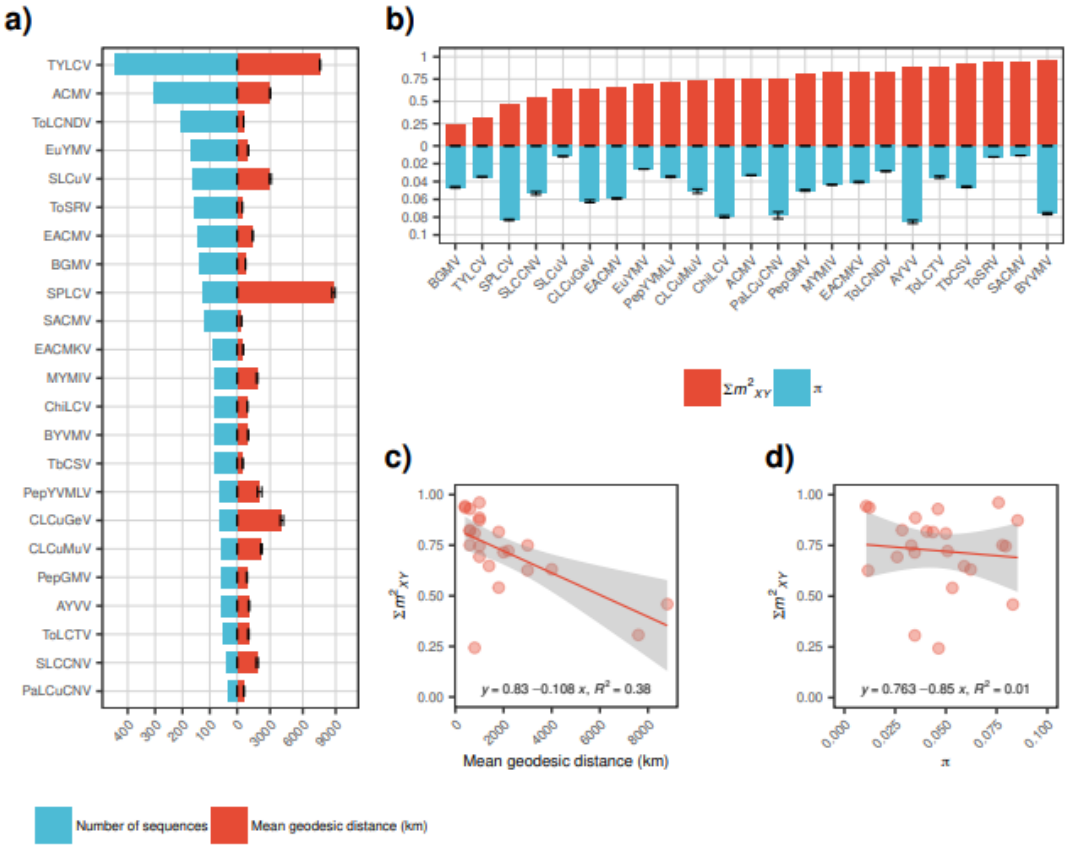
a)
TYLCV
ACMV
ToLCNDV
EuYMV
SLCuV
ToSRV
EACMV
BGMV
SPLCV
SACMV
EACMKV
MYMIV
ChiLCV
BYVMV
TbCSV
PepYVMLV
CLCuGeV
CLCuMuV
PepGMV
AYVV
ToLCTV
SLCCNV
PaLCuCNV
400 300 200 100 0 3000 6000 9000
Number of sequences
Mean geodesic distance (km)
b)
1 0.75 0.5 0.25 0 0.02 0.04 0.06 0.08 0.1
BGMV TYLCV SPLCV SLCCNV SLCuV CLCuGeV EACMV EuYMV PepYVMLV CLCuMuV ChiLCV ACMV PaLCuCNV PepGMV MYMIV EACMKV ToLCNDV AYVV ToLCTV TbCSV ToSRV SACMV BYVMV
Σm^2_{XY}
π
c)
1.00 0.75 0.50 0.25 0.00
Σm^2_{XY}
$y = 0.83 - 0.108\ x,\ R^2 = 0.38$
0 2000 4000 6000 8000
Mean geodesic distance (km)
d)
1.00 0.75 0.50 0.25 0.00
Σm^2_{XY}
$y = 0.763 - 0.85\ x,\ R^2 = 0.01$
0.000 0.025 0.050 0.075 0.100
π

Figure 2

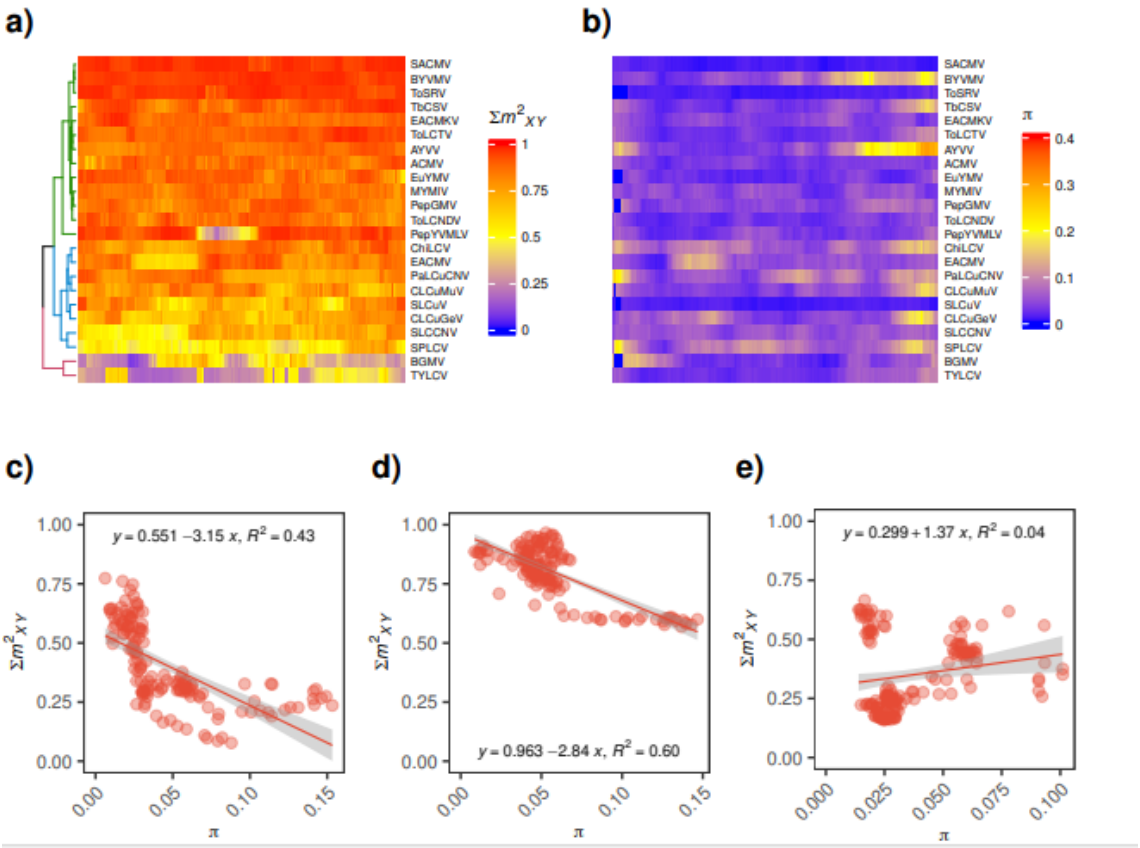
a)
b)
SACMV
BYVMV
ToSRV
TbCSV
EACMKV
ToLCTV
AYVV
ACMV
EuYMV
MYMIV
PepGMV
ToLCNDV
PepYVMLV
ChiLCV
EACMV
PaLCuCNV
CLCuMuV
SLCuV
CLCuGeV
SLCCNV
SPLCV
BGMV
TYLCV
Σm^2_{XY}
1
0.75
0.5
0.25
0
π
0.4
0.3
0.2
0.1
0
c)
d)
e)
$y = 0.551 - 3.15\,x$, $R^2 = 0.43$
$y = 0.963 - 2.84\,x$, $R^2 = 0.60$
$y = 0.299 + 1.37\,x$, $R^2 = 0.04$
1.00
0.75
0.50
0.25
0.00
0.00
0.05
0.10
0.15
0.000
0.025
0.050
0.075
0.100
π

Figure 3

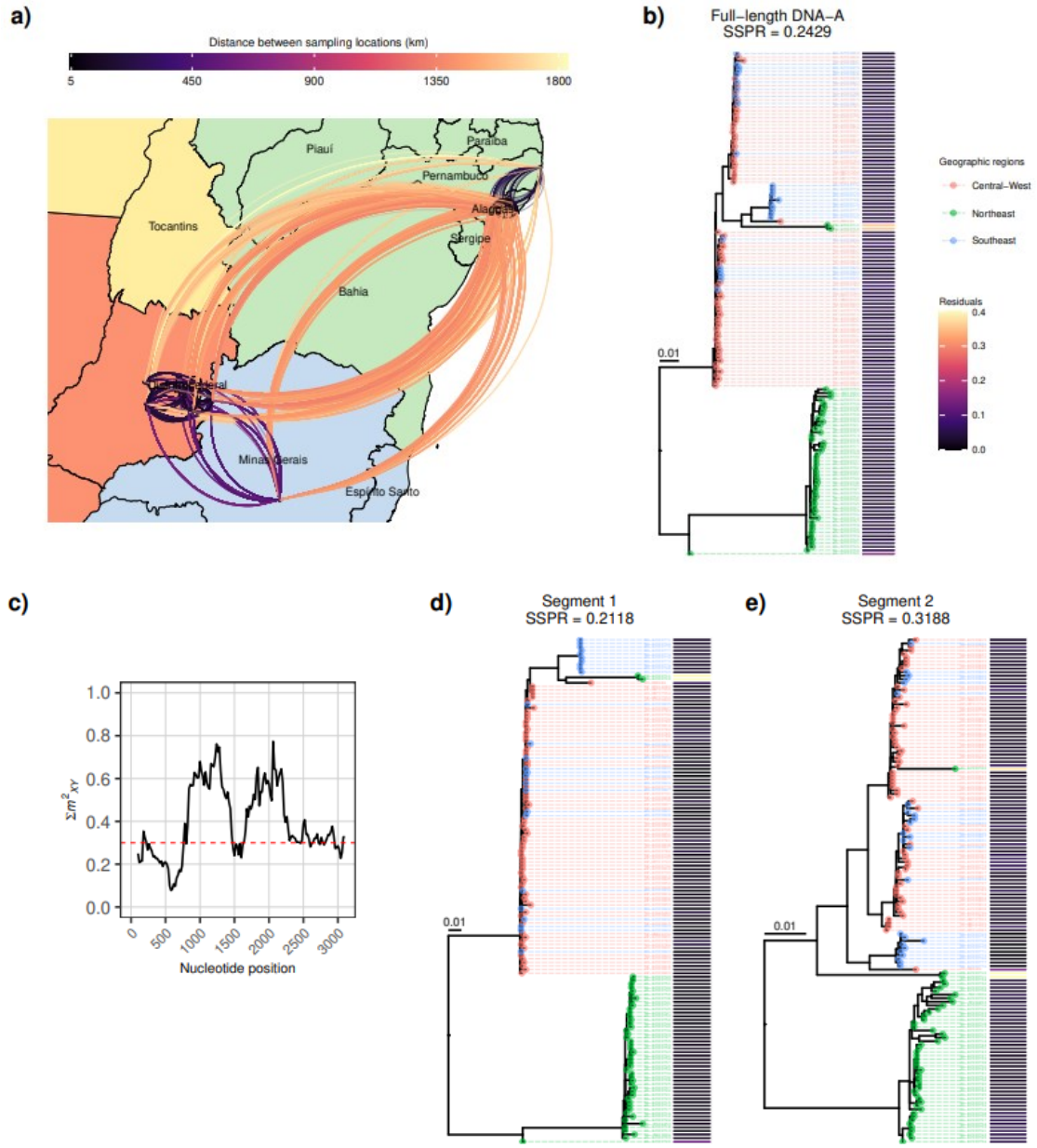
a)
Distance between sampling locations (km)
5
450
900
1350
1800
Piauí
Paraíba
Pernambuco
Alagoas
Sergipe
Tocantins
Bahia
Minas Gerais
Espírito Santo
b)
Full−length DNA−A
SSPR = 0.2429
0.01
Geographic regions
Central−West
Northeast
Southeast
Residuals
0.4
0.3
0.2
0.1
0.0
c)
1.0
0.8
0.6
0.4
0.2
0.0
0
500
1000
1500
2000
2500
3000
Nucleotide position
d)
Segment 1
SSPR = 0.2118
0.01
e)
Segment 2
SSPR = 0.3188
0.01

Figure 4

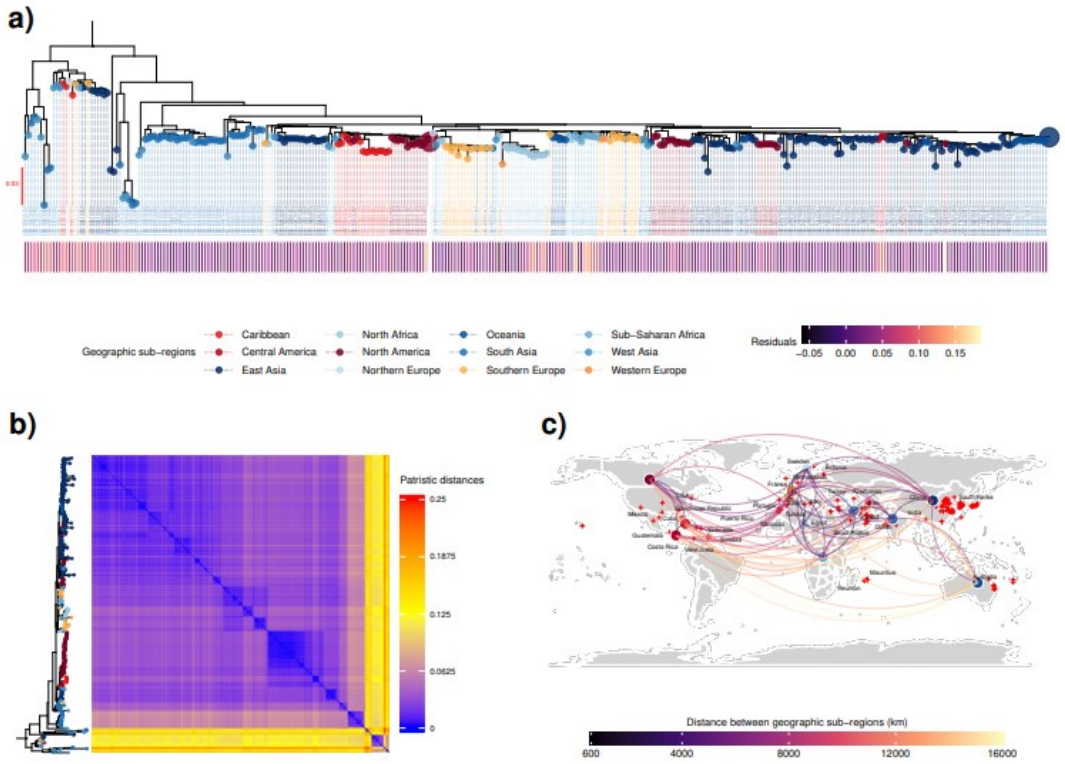

Figure 5