

GABRIEL BONFIM DA SILVA MACHADO

UMA ABORDAGEM EM SÍLICO PARA VALIDAÇÃO DA
ARQUITETURA DE CAPSÍDEO DE NOVOS VÍRUS
ICOSAÉDRICOS: *GEMINIVIRIDAE* COMO CASO DE ESTUDO



**Plano de Monografia
apresentado ao Instituto
de Microbiologia Paulo
de Góes, da
Universidade Federal do
Rio de Janeiro, como pré-requisito
para a obtenção do grau de
bacharelado em ciências biológicas:
Microbiologia e imunologia.**

INSTITUTO DE MICROBIOLOGIA PAULO DE GÓES
UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
RIO DE JANEIRO 2020

**Trabalho a ser realizado no
Departamento de Virologia do Instituto
de Microbiologia Paulo de Góes, UFRJ,
sob a orientação do(a) Professor(a)
Tatiana Domitrovic.**

FICHA CARTOGRÁFICA

Machado, Gabriel Bonfim

Uma abordagem em sílico para a validação da arquitetura de capsídeo de novos vírus icosaédricos: *Geminiviridae* como caso de estudo / Gabriel Bonfim da Silva Machado – Rio de Janeiro: UFRJ, 2020.

xi; 45f. : il ; 30cm

Orientador: Tatiana Domitrovic

Trabalho de Conclusão de Curso (Bacharelado em Ciências Biológicas: Microbiologia e Imunologia) - Universidade Federal do Rio de Janeiro, Instituto de Microbiologia Paulo de Góes, 2020.

Bibliografia: f.41-45

1. Capsídeo
 2. Domínios de arginina (R)
 3. *Geminiviridae*
 4. Bioinformática
- I. Domitrovic, Tatiana II. UFRJ. Instituto de Microbiologia Paulo de Góes. III. Título

**INSTITUTO DE MICROBIOLOGIA PAULO DE GÓES /
UFRJ
COORDENAÇÃO DE ENSINO DE GRADUAÇÃO**

**ATA DA APRESENTAÇÃO DE MONOGRAFIA PARA APROVAÇÃO
NO RCS DE TRABALHO DE CONCLUSÃO DE CURSO,
BACHARELADO EM CIÊNCIAS BIOLÓGICAS: MICROBIOLOGIA
E IMUNOLOGIA**

ALUNO: **Gabriel Bonfim da Silva Machado**

DRE: 117061204

BANCA EXAMINADORA: Prof. Davis Fernandes Ferreira (Presidente)

Dra. Simone da Graça Ribeiro

Prof. Fernando Luz de Castro

Profa. Renata Campos Azevedo (Suplente)

Título da Monografia: **“Uma abordagem sílico para validação da arquitetura de capsídeo de novos vírus icosaédricos: *Geminiviridae* como caso de estudo”**

Local: Sala virtual <https://meet.google.com/nbm-azpy-oay>.

Data e hora de início: **09 de novembro de 2020 às 14:00h**

Em sessão pública, após exposição de cerca de 50 minutos, o aluno foi argüido pelos membros da Banca Examinadora, demonstrando suficiência de conhecimentos e capacidade de sistematização no tema de sua Monografia, tendo, então, obtido nota 10,0 neste requisito do RCS de **TRABALHO DE CONCLUSÃO DE CURSO**. Na forma regulamentar, foi lavrada a presente ata que é assinada pelo presidente da banca examinadora, aluno, orientador (ou coorientador) e pelo coordenador do RCS.

Rio de Janeiro, 9 de novembro de 2020.

NOTA

___10,0__

___10,0__

___10,0__

Banca Examinadora:

Prof. Davis Fernandes Ferreira

Dra. Simone da Graça Ribeiro

Prof. Fernando Luz de Castro

Profa. Renata Campos Azevedo

Presidente da banca



Prof. Davis Fernandes Ferreira

Aluno:



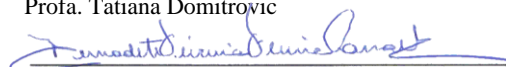
Gabriel Bonfim da Silva Machado

Orientador:



Profa. Tatiana Domitrovic

Coordenador
de TCC



Profa. Bernadete Teixeira Ferreira Carvalho

AGRADECIMENTOS

À Deus, por ser meu alicerce e guia na minha jornada, não me permitindo desistir e por quem sou grato as oportunidades que se apresentam durante o caminho.

À minha família por ser minha base e apoio, por sempre acreditarem no meu potencial e por estarem sempre dispostos a ajudar em tudo que é possível.

À professora Tatiana Domitrovic pela oportunidade, orientação e ajuda, sem elas não seria possível chegar até aqui. Serei sempre grato.

À Luca Cestari, que me auxiliou no começo dessa jornada, compartilhando o seu conhecimento.

À professora Maite Vaslin com quem compartilhamos o laboratório, pela disponibilidade, ajuda e suporte.

À Andreia Santino e Fernanda Barreiro pelos momentos de descontração, ajuda e acolhimento, deixando o dia a dia do laboratório mais leve. E todos os membros do laboratório, pela ajuda, trocas de conhecimento e pelas pausas para o café.

Aos amigos, Talita Almeida pelo apoio e incentivo, alguém com que posso contar para todos os momentos. Rafaela Almeida, Júlia Maria e Felipe Matheus com quem compartilhei alegrias e tristezas ao longo de toda a graduação. Obrigado por toda ajuda e apoio!

À todos os professores do instituto de Microbiologia e Imunologia, que foram peças fundamentais na minha formação, durante toda a graduação.

Ao CNPQ pela bolsa de iniciação científica concedida.

E a todos que deixaram a sua marca durante essa jornada e contribuíram para a conclusão desse trabalho.

RESUMO

GABRIEL BONFIM DA SILVA MACHADO

UMA ABORDAGEM EM SÍLICO PARA A VALIDAÇÃO DA ARQUITETURA DE CAPSÍDEO DE NOVOS VÍRUS ICOSAÉDRICOS: *GEMINIVIRIDAE* COMO CASO DE ESTUDO

Orientadora: Tatiana Domitrovic

Resumo da Monografia apresentada no Instituto de Microbiologia Paulo de Góes da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para obtenção do título de Bacharel em Ciências Biológicas: Microbiologia e Imunologia e aprovação no RCS Trabalho de Conclusão de Curso.

Dados da metagenômica revelaram muitas novas espécies, vírus conhecidos apenas por suas sequências. A classificação taxonômica baseada apenas nas sequências dessas entidades virais é um desafio, por isso são necessárias mais ferramentas para entender e prever os aspectos biológicos delas. Várias proteínas de capsídeo dos vírus icosaédricos apresentam um domínio carregado positivamente (R-arm). Sabe-se que esses domínios são importantes para o empacotamento e a estabilidade do genoma. Recentemente, nosso grupo desenvolveu uma abordagem computacional baseada no cálculo da carga líquida de segmentos de proteínas, conseguindo assim identificar o R-arm de forma automática e calcular a carga líquida eletrostática. Com essa análise, foram identificados vírus de diversas famílias, com genomas de ssDNA, ssRNA, e dsDNA, em que a capacidade de empacotamento do genoma está relacionada com o número total de subunidades do capsídeo (arquitetura do capsídeo). Portanto, propomos que sabendo a sequência da proteína do capsídeo e o tamanho total do genoma, é possível aplicar essa análise para checar se um membro putativo de uma dada família viral apresenta a morfologia de partícula esperada. Neste trabalho, testamos essa hipótese com os geminivírus (*Geminiviridae*), que estão entre os vírus que usam os R-arms para estabilizar seus capsídeos. São vírus de planta de genoma ssDNA, divididos em nove gêneros, sendo dois principais, *Begomovirus* e *Mastrevirus*, e outros 7 menores. Esses vírus apresentam uma estrutura de capsídeo tridimensional conhecida por capsídeo geminado, formado por dois capsídeos icosaédricos $T = 1$ unidos por um vértice pentamérico, totalizando 110 subunidades repetidas. Aplicamos o nosso programa para calcular a carga líquida do R-arm para todas as sequências da família *Geminiviridae* incluídas no 10th relatório International Committee on Taxonomy of Viruses, ICTV (n=442). Além de sequências de geminivírus não classificados e altamente divergentes do GenBank do NCBI. Assim, conseguimos observar uma correlação positiva entre a carga líquida linear e o tamanho do genoma para a maioria das sequências do nosso banco de dados. Todos os gêneros menores têm uma proporção genoma/capsídeo similar, corroborando a hipótese de que eles compartilham a mesma arquitetura de capsídeo geminado. Também foi possível observar que alguns vírus da família estão deslocados, nossa hipótese é que essas espécies podem ter uma arquitetura alternativa para o seu capsídeo. Por fim, fomos capazes de identificar que uma família viral intimamente relacionada, *Genomoviridae* (T=1, 60 subunidades) não obedece a arquitetura geminada. Os resultados apontam para o potencial que a nossa metodologia apresenta na determinação da morfologia capsídica dessas sequências virais e por consequência no auxílio da caracterização de novos vírus vindos da metagenômica.

Palavras-chaves: Capsídeo, Domínios de arginina (R), *Geminiviridae*, Bioinformática

ABSTRACT**GABRIEL BONFIM DA SILVA MACHADO****AN IN-SILICO APPROACH TO VALIDATE THE CAPSID ARCHITECTURE OF
NEW PUTATIVE ICOSAHEDRAL VIRUSES: GEMINIVIRIDAE AS CASE
STUDY****Orientadora: Tatiana Domitrovic**

Abstract da Monografia apresentada no Instituto de Microbiologia Paulo de Góes da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários para obtenção do título de Bacharel em Ciências Biológicas: Microbiologia e Imunologia e aprovação no RCS Trabalho de Conclusão de Curso.

With the advent of metagenomics approaches, a diversity of known and unknown viruses has been identified in various types of samples. The sequence-based only taxonomic classification of these viruses is a challenge, and more tools are needed to understand and predict biological aspects of this viral-dark matter. Many capsid proteins from icosahedral viruses have a positively charged domain (R-arm) that is important for genome packaging and particle stability but that have poor sequence conservation. Recently, our group developed a computational approach based on the net charge calculation of discrete protein segments that can automatically identify R-arms and calculate its electrostatic net-charge. With this analysis, we identified various virus families with ssDNA, ssRNA, and dsDNA genomes, for each the genome packaging capacity is highly correlated with the total number of capsid subunits (capsid architecture). Therefore, we propose that by knowing the capsid protein sequence and the total genome size, we could apply this analysis to check if a putative new member of a given viral family complies with the particle morphology expected for that group. In this work, we tested this hypothesis with geminiviruses (*Geminiviridae*), that were among the viruses that use capsid R-arms to stabilize their capsids. These ssDNA plant viruses are divided into two main genera, *Begomovirus* and *Mastrevirus*, which contain the viruses with known tridimensional capsid structure: a geminated capsid, formed by two T=1 icosahedral capsids joined by a missing pentameric vertex, totalizing 110 repeated subunits. The family also comprises other 7 minor genera, 2 unsigned species and several new species. We applied our program to calculate the R-arm net charge for all *Geminiviridae* sequences included in the 10th ICTV report (n=442) and other putative isolates. We observed a linear correlation between the R-arm net charge and the genome size for most of the data-set. All minor genera had similar genome/capsid charge ratio, corroborating the assumption that they share the same geminated capsid architecture. Importantly, our plot was able to predict that a virus closely related *Genomoviridae* family (T=1, 60 subunits) did not comply with the geminated architecture. Moreover, our analysis indicated that mulberry mosaic dwarf associated virus that is yet unassigned to a genus in the family *Geminiviridae*, could have an alternative T=1 virus architecture based on the R-arm charge and genome size.

Key-words: Bioinformatics, Capsid, *Geminiviridae*, R-arm

LISTA DE ABREVIATURAS E SIGLAS

ACMV	<i>African cassava mosaic virus</i>
AGmV	<i>Apple geminivirus</i>
ATP	Adenosina trifosfato
BCTV	<i>Beet curly top virus</i>
CCMV	<i>Cowpea chlorotic mottle virus</i>
CP	Proteína de capsídeo
CR	Região comum
CRESS DNA	Vírus DNA de fita simples codifica proteína associada a replicação celular
Cryo-EM	Microscopia crioeletrônica
DI DNA	DNA Interferente defeituoso
dsDNA	Fita dupla de DNA
FHV	<i>Flock house Vírus</i>
ICTV	International Committee on Taxonomy of Viruses
IR	Região intergênica
JmaV	<i>Juncus maritimus associated virus</i>
Kb	Kilobases
LaaV	<i>Limeum africanum associated vírus</i>
LIR	Longa região intergênica
MLS	Listra mestre de espécies
MMDaV	<i>Mulberry mosaic dwaft associated vírus</i>
Mp	Proteína de movimento
MSV	<i>Maize streak virus</i>
NCBI	National Center for Biotechnology Information
NimiV	Niminivirus
Nm	Nanômetro
NSP	Proteína de transporte nuclear
N ω V	<i>Nudaurelia capensis omega virus</i>
OpV1	<i>Opuntia virus 1</i>
PcMoV	<i>Passion fruit chlorotic mottle virus</i>
PCV2	<i>Porcine circovirus 2</i>
Qc	Concentração de carga
Qgenome	Carga líquida do genoma
Qmax	Carga máxima
Qmax30res	Carga máxima em um trecho de 30 resíduos
RCA	Amplificação do círculo rolante
RdRp	RNA polimerase dependente de RNA
RefSeq	Sequências de referência
REn	Proteína intensificadora da replicação
Rep	Proteína associada a replicação
ssDNA	Fita simples de DNA
ssRNA	Fita simples de RNA
ssRNA+	Fita simples de RNA de sentido positivo
TaGV1	<i>Tomate associated geminivirus</i>
TALCV	<i>Tomato apical leaf curl vírus</i>
TrAP	Proteína ativadora da transcrição
VP	Proteína de capsídeo viral

ÍNDICE

FICHA CARTOGRÁFICA	iii
RESUMO	vi
ABSTRACT	vii
LISTA DE ABREVIATURAS E SIGLAS	viii
1. INTRODUÇÃO	10
1.1 Estrutura de capsídeos virais e mecanismos de automontagem	10
1.1.1 Caracterização dos domínios positivos em vírus	13
1.2 Os desafios da taxonomia de vírus	16
1.3 <i>Geminiviridae</i> como caso de estudo:	19
1.3.1 Relação entre os <i>Genomoviridae</i> e os <i>Geminiviridae</i>	23
2. JUSTIFICATIVA	25
3. OBJETIVOS	25
3.1 Objetivo principal	25
3.2 Objetivos específicos.	25
4. MATERIAIS E MÉTODOS	26
4.1 Banco de dados	26
4.2 Cálculo da carga líquida	26
4.3 Correlação entre a carga do genoma e do domínio da CP	28
5. RESULTADOS	29
5.1 <i>Begomovirus</i> e <i>Mastrevirus</i> apresentam uma correlação positiva entre a carga do domínio positivo e a carga do genoma.	29
5.2 <i>Gemonoviridae</i> : família recém criada, intimamente relacionada aos <i>Geminiviridae</i>	33
6. DISCUSSÃO	36
7. CONCLUSÃO	39
8. REFERÊNCIAS BIBLIOGRÁFICAS	40

1. INTRODUÇÃO

1.1 Estrutura de capsídeos virais e mecanismos de automontagem

Uma das características fundamentais dos vírus é a existência do capsídeo, um dos principais componentes da partícula viral infecciosa. O capsídeo viral é uma cápsula proteica com função protetora, que interage com o genoma. Sua formação é essencial para a replicação viral, protegendo o material genético de virar alvo de degradação (Santos, Romanos e Wigg., 2015). O capsídeo também pode estar envolvido na interação com os receptores da célula hospedeira, levando a penetração na membrana celular do hospedeiro.

Além disso, a proteína de capsídeo (CP) também pode auxiliar no direcionamento do genoma para o local da replicação, como observado nas proteínas de capsídeo dos parvovírus (VP1) (Berns e Parrish., 2013). O capsídeo também participa da liberação do material genético viral. Essa liberação pode ser realizada, por meio da desmontagem do capsídeo no citoplasma, ou diretamente no núcleo da célula hospedeira, como o caso do vírus da hepatite B e baculovírus (Cohen, Au e Panté, 2011).

O capsídeo viral é composto por uma associação de moléculas proteicas, muitas vezes idênticas, em um arranjo repetitivo. Os capsídeos apresentam duas conformações principais. A helicoidal, que têm por característica a aparência filamentosa, tubular (estrutura fina e alongada) onde as subunidades proteicas que formam o capsídeo interagem com o genoma e se associam com simetria helicoidal de forma a manter o genoma protegido no interior do filamento (Perlmutter e Hagan, 2015). Essa estrutura pode ser mais rígida ou frouxa, dependendo do arranjo que as CPs adquirem no espaço.

A outra conformação, a mais comum entre os vírus, é a esférica, com as subunidades interagindo de acordo com os princípios da simetria icosaédrica. A forma icosaédrica mais simples é composta por 60 subunidades idênticas que formam as faces triangulares de um polígono de 20 lados (Figura 1). Essa estrutura permite a organização mais eficiente das subunidades repetidas com maior aproveitamento do volume interno. Assim, os menores capsídeos são formados por 60 subunidades proteicas (Perlmutter e Hagan, 2015).

Existem maneiras de se produzir capsídeos maiores seguindo os princípios da simetria icosaédrica. Muitos capsídeos icosaédricos, são compostos por um número de subunidades proteicas múltiplas de 60. Para isso, um leve ajuste estrutural deve acontecer nos domínios de interação entre os capsômeros, permitindo a formação de pentâmeros nos

12 vértices do icosaedro, e hexâmeros nas faces. Esse princípio chama-se *quasi-equivalência* (Caspar e Klug, 1962). Assim, de acordo com a simetria icosaédrica, o número de triangulação (T), corresponde ao número de proteínas na unidade assimétrica, ou seja, o número de estruturas únicas que irão se repetir 60 vezes para formar o capsídeo (Santos, Romanos e Wigg, 2015). Portanto, o número total de subunidades de um capsídeo icosaédrico será $60 \times T$ (Caspar e Klug, 1962), assim um capsídeo com $T=3$, tem 180 subunidades que assumem até 3 tipos alternativos de conformação tridimensional (Figura 1).

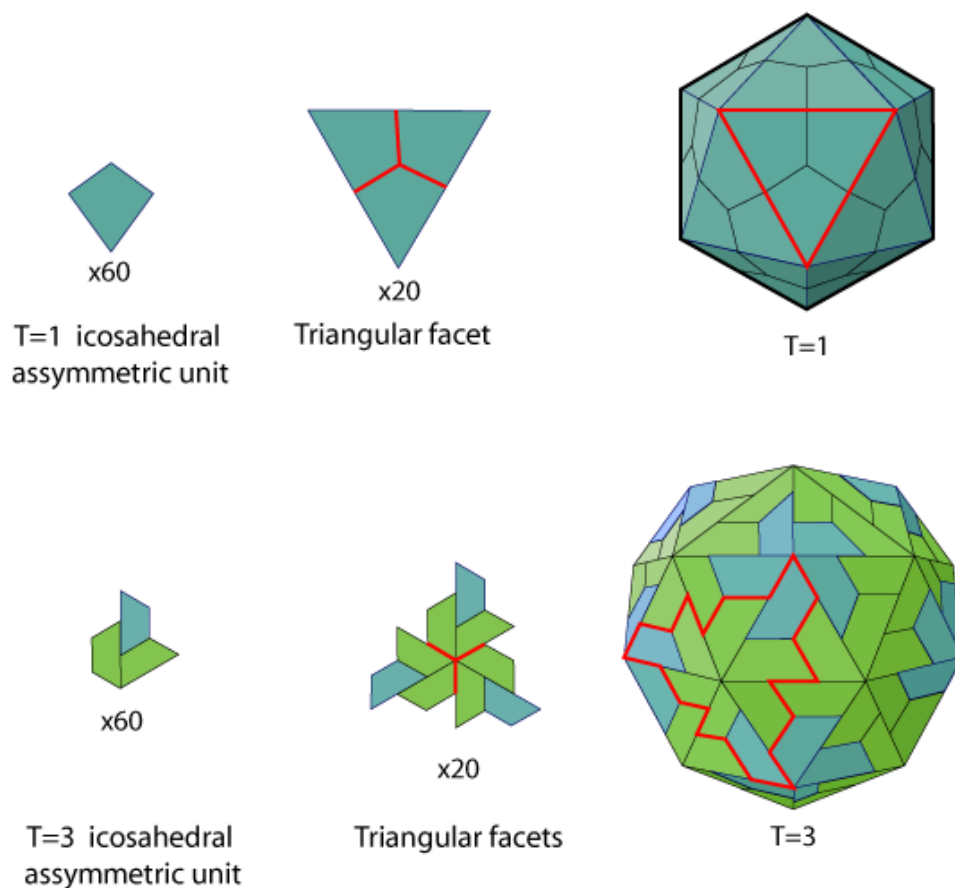


Figura 1. Modelos da estrutura do capsídeo icosaédrico. O capsídeo $T = 1$ é composto por 60 unidades assimétricas formadas por uma proteína. As faces do icosaedro, compostas por 3 dessas unidades assimétricas, vão se repetir 20 vezes para a formação do polígono de 60 subunidades proteicas. O capsídeo $T = 3$ é composto por 60 unidades assimétricas formadas por três proteínas, totalizando 180 subunidades proteicas. (Adaptado da fonte: Virolzone).

A capacidade de automontagem dos vírus tem sido bastante estudada. Ela se refere à capacidade que as subunidades proteicas e as cadeias de ácidos nucleicos têm de se juntar para a formar a partícula viral. Em um trabalho com o vírus icosaédrico cowpea chlorotic

mottle virus (CCMV), foi avaliado o papel das interações entre proteínas de capsídeo vizinhas (CP-CP), e entre a proteína de capsídeo e o ácido nucleico (CP- RNA) no processo de montagem do vírus. Foi possível monitorar o mecanismo de montagem *in vitro* do CCMV, controlando as forças de interações entre CP-CP por ajuste no pH, e as interações CP-RNA, modificando a força iônica. Como resultado, ficou evidenciado que uma montagem bem sucedida depende das forças de atração CP-CP em relação à atração CP-RNA. Caso essas atrações sejam muito fracas, o capsídeo não consegue se formar, em contrapartida se essas atrações forem muito fortes a estrutura é formada irregularmente (Garmann, *et al.*, 2014).

São diversas as possíveis vias de montagem do capsídeo, incluindo diferentes detalhes. Os bacteriófagos, por exemplo, empacotam seus genomas por meio de motores moleculares compostos principalmente por ATPases, responsáveis por bombear o DNA viral para dentro do capsídeo (Hilbert *et al.*, 2015). Já os vírus de DNA de fita dupla (dsDNA), pertencentes ao gênero *Poliomavirus*, têm o genoma empacotado por histonas (Hurdiss, *et al.*, 2016).

Muitos vírus icosaédricos apresentam uma alta concentração de aminoácidos de carga positiva, localizados nas extremidades da proteína de capsídeo (Requião R.D, *et al.*, 2019). Essas sequências têm como principal característica uma alta concentração de arginina, sendo denominadas, domínios poliarginina ou R-arm (braço de arginina) e ficam voltadas para o interior do capsídeo em contato com o ácido nucleico (figura 2) (Perlmutter e Hagan,, 2015). A importância dos domínios de carga positiva localizados nas regiões N/C terminais para a estabilização do capsídeo foi comprovada para várias espécies de vírus RNA fita simples positiva (Speir, J. A *et al.*, 2006; Khayat, R *et al.*, 2011; Hassani-Mehraban, A. *et al.*, 2015; Rayaprolu, V. *et al.*, 2017). Um exemplo típico é o vírus icosaédrico flock house vírus (FHV), da família *Nodaviridae* (Figura 1). FHV's mutantes, que apresentavam deleção na região N terminal positivamente carregada da proteína do capsídeo, passavam a formar partículas defectivas, que diferiam no tamanho e forma esperados para o FHV (Dong, et al., 1998). Mais tarde, outro estudo, evidenciou a importância que a região N terminal da CP dos FVHs, para o empacotamento e reconhecimento do genoma (Marshall e Schneemann, 2001).

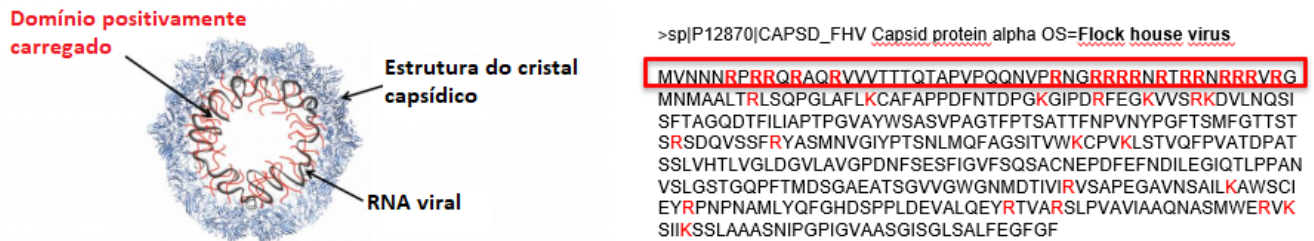


Figura 2- Representação dos domínios positivamente carregados em proteínas de capsídeo viral (esquerda). As coordenadas tridimensionais X-ray do capsídeo do *Flock House virus* (FHV) foram usadas para gerar uma representação em corte do capsídeo icosaédrico T = 3. O RNA (representado pelas linhas pretas) é em sua maioria dinâmico e está entrelaçado com o domínio da proteína rica em arginina (representado pelas linhas vermelhas). Sequência da proteína de capsídeo do FHV (direita), em destaque dentro do retângulo em vermelho o domínio positivo (Adaptado da fonte: Requião R.D, *et al.*, 2019).

Estudos utilizando modelos *in vitro*, permitiram verificar de forma ainda mais clara a influência da carga do domínio positivo na capacidade de encapsulamento do genoma. Nesse estudo foram utilizados modelos compostos por monômeros da proteína de capsídeo do CCMV modificadas para conter mais ou menos carga positiva no seu braço de arginina e moléculas de ssRNA de diferentes tamanhos. A reação da montagem foi controlada através do pH, utilizado para ajustar as forças intramoleculares. Os resultados indicaram que o empacotamento dos RNA testados, demandam uma proporção correspondente entre a carga majoritariamente negativa do genoma, e a carga positiva das extremidades da proteína de capsídeo (Garmann, *et al.*, 2015).

Em análise teórica, com vírus ssRNA+, foi realizada uma análise de comparação quantitativa entre a carga do genoma, e a carga do R-arm de vírus de diferentes famílias. Os resultados demonstraram que a razão entre o número de bases do genoma e a carga do domínio positivo das CPs apresenta um coeficiente linear de 1.61 (± 0.03) conservado em todos os vírus ssRNA analisados (Vladimir e Muthukumar, 2006). Ficando evidente, portanto, que a carga do domínio positivo apresenta uma correlação positiva com o tamanho do genoma encapsulado. Acredita-se que esses motivos de poliarginina das CPs podem estar envolvidos na compensação das forças repulsivas eletrostáticas, resultantes da alta concentração dos ácidos nucleicos de carga negativa que constituem o material genético, condensado dentro do capsídeo (Perlmutter, *et al.*, 2013; Garmann, *et al.*, 2016).

1.1.1 Caracterização dos domínios positivos em vírus

Apesar dos R-arms estarem presentes em diversos vírus, alguns fatores dificultam a identificação automática desses motivos positivamente carregados. Por consequência, esses domínios nunca foram formalmente anotados em bancos de dados como o Pfam ou InterPro

(Richardson, L. J. *et al* 2019). Entre esses fatores, está a baixa conservação da sequência de aminoácidos dessas sequências, o que prejudica o método empregado pelos bancos de dados de proteína como o Pfam, que é baseado no alinhamento com sequências de referência e identificação de padrões repetitivos (Finn *et al.*, 2015). Além disso, a natureza desordenada ou flexível dos domínios positivos, impede a sua visualização em modelos estruturais. Todas essas condições reforçam o caráter único da estrutura dos segmentos positivamente carregados das proteínas de capsídeo viral. Essa singularidade dificulta o uso de abordagens tradicionais, baseadas apenas na comparação de sequência, para a identificação dos R-arms em vírus não relacionados, e até mesmo, em vírus da mesma família.

Portanto, para que fosse possível determinar a ocorrência dos domínios de carga positiva em CPs de diferentes vírus, nosso grupo desenvolveu um programa, capaz de calcular a carga eletrostática líquida em trechos consecutivos de um número pré determinado de aminoácidos (Requião R.D, *et al.*, 2019). Assim, é possível localizar a região com maior carga (Q_{max}).

Baseando-se no cálculo da carga líquida de segmentos da CP ($Q_{max30res}$), nosso grupo caracterizou a distribuição desses domínios em proteínas de diversos organismos. Os resultados das análises demonstraram que as proteínas de capsídeo viral estão entre as que possuem os segmentos mais positivamente carregados, apresentando trechos com pelo menos, quatro vezes mais arginina que lisina, uma característica pouco comum nas proteínas celulares. Os vírus icosaédricos, especialmente, concentram os trechos mais positivamente carregados (Requião, *et al.*, 2019).

A partir das evidências da correlação positiva entre a carga total dos R-arms presentes nos capsídeos e a carga do genoma de um grupo de vírus ssRNA icosaédrico, o trabalho do nosso grupo, expandiu essa hipótese para vírus de diferentes genomas. Esse estudo analisou 179 vírus icosaédricos, incluindo bacteriófagos e vírus de eucariotos de 29 famílias diferentes e diversos genomas, com exceção dos ssRNA- (vírus helicoidais) e ssRNA-RT. As cargas dos domínios positivos das sequências de CPs analisadas foram calculadas, por meio do programa desenvolvido no laboratório (detalhado acima). Em seguida, a carga total do capsídeo foi calculada multiplicando a $Q_{max30res}$ achada, pelo número de subunidades que formam o capsídeo de cada vírus estudado (Total $Q_{max30res}$). A carga do genoma (Q_{genome}) foi calculada considerando que cada resíduo de nucleotídeo contribui com -1 de carga. A partir desses dados, foi montado um gráfico de correlação entre a carga do genoma e carga total dos domínios positivamente carregados (Figura 3).

Nessa análise, foram identificadas 17 famílias (Figura 3) que apresentam forte correlação positiva entre carga do genoma e carga total dos domínios positivos, (Pearson, $r = 0.91$, p -valor < 0.0001) (Requião, *et al.*, 2019). Esses resultados, indicam que a capacidade de empacotamento desses vírus está altamente relacionada com a carga líquida interna do capsídeo. Estão entre eles, vírus +RNA, como os membros das famílias *Nodaviridae* e *Alphatetraviridae*, cujos experimentos demonstram que são extremamente dependentes dos R-arms para a formação do capsídeo (Venter, Marshall e Schneemann, 2009). Além disso, também estão presentes dentro do ajuste linear, vírus ssDNA, que estão entre aqueles que apresentam os segmentos de proteína com maior carga líquida, como por exemplo, os vírus que compõe a família *Circoviridae* (Requião, *et al.*, 2019). Dentro do ajuste linear, também foi possível identificar, representantes de dsDNA-RT (*Hepadnaviridae* e *Caulimoviridae*) e dsDNA (*Papillomaviridae* e *Polyomaviridae*). Juntos esses vírus representam 40% das famílias virais de conformação icosaédrica (Requião, *et al.*, 2019). Portanto, os resultados do estudo indicam que vírus, não relacionados, com diferentes tipos de genoma (ssDNA, ssRNA e dsDNA), tem em comum o R arm como o principal mecanismo para a montagem do capsídeo.

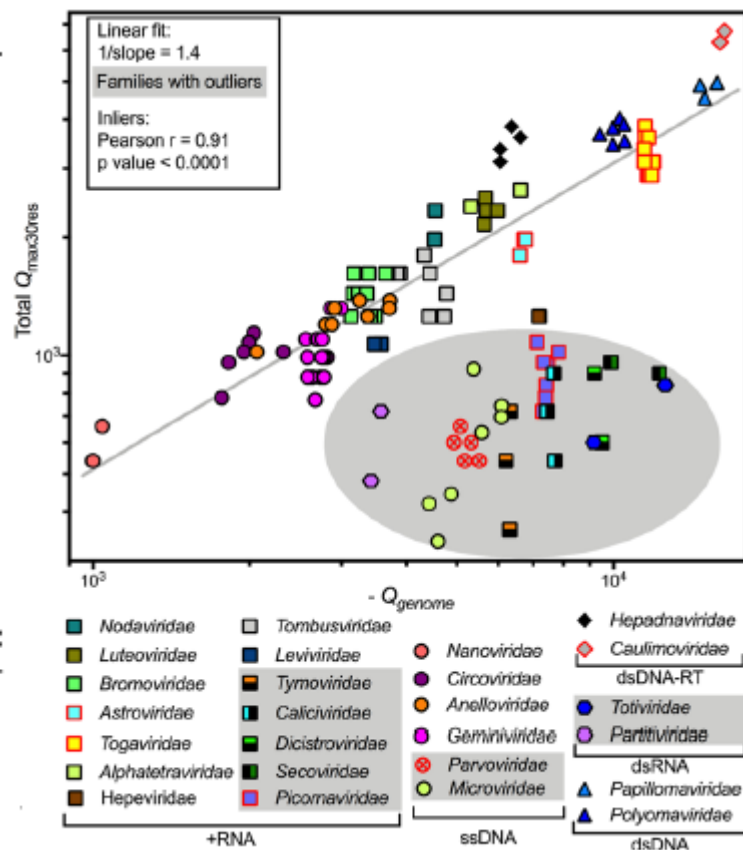


Figura 3 - Correlação entre a carga líquida interna do capsídeo, calculado a partir do segmento mais positivamente carregado da proteína de capsídeo, e a capacidade de empacotamento do genoma. O valor máximo da carga líquida em um trecho de 30 aminoácidos foi multiplicado pelo número de subunidades que formam o capsídeo ($Q_{\text{max30res total}}$) em um grupo de 179 vírus de 29 famílias diferentes. A carga líquida do genoma foi calculada pelo número de aminoácidos no genoma (Q_{genoma}). (B) Gráfico de correlação entre a carga do genoma e a $Q_{\text{max30res total}}$. Círculos e quadrados representam vírus eucarióticos e bacteriófagos, respectivamente. Coloridos de acordo com o número de triangulação. Os vírus eucarióticos e os bacteriófagos *Leviviridae* e *Microviridae* foram utilizados para calcular o ajuste em linha reta ($n = 133$). A área sombreada indica famílias com outliers. Os resultados da correlação de Pearson obtidos a partir dos inliers (103) são mostrados na inserção. Os pontos de dados contornados em vermelho representam vírus que têm mais Lys que Arg em seus segmentos carregados positivamente. (Adaptado da fonte: Requião R.D, *et al.*, 2019).

Apesar do método de rastreamento das sequências positivas utilizando um número fixo de resíduos de amino ácidos ter se mostrado rápido e eficiente em identificar as regiões de R-arm, nosso grupo desenvolveu um novo programa capaz de aperfeiçoar a análise, permitindo a variação do número de resíduos de aminoácidos que formam o domínio positivamente carregado. Para tanto, assumimos que a região de R-arm será a sequência com maior carga líquida e com a maior concentração de carga da proteína ($Q_c = Q / \text{número de resíduos}$). Assim, o novo programa, denominado CargaFlex funciona inicialmente como o algoritmo anterior, salvando o trecho de carga líquida mais alta em uma janela pré determinada (ex, 8 resíduos). Em seguida, ele reinicia a pesquisa, em uma janela maior (ex, 9 resíduos) e irá substituir o trecho salvo, caso encontre outro segmento com uma carga líquida mais alta e com uma concentração maior ou igual ao limite Q_c determinado. O programa continuará a procura até que sejam eliminadas as possibilidades de extensão de janelas, limitadas pelo tamanho da sequência.

O CargaFlex foi testado e os resultados comparados com o programa da forma fixa. As análises comprovaram que o uso da janela variável introduz ajustes discretos de carga e tamanho do R-arm sem distorcer os dados (Requião, *et al.*, 2019). Com isso, conclui-se que apesar da busca do R-arm através do quadro fixo de 30 resíduos ser eficiente, a utilização da janela flexível permite refinar a análise e pode ser útil na análise de um conjunto de dados com proteínas muito similares entre si.

1.2 Os desafios da taxonomia de vírus

O estudo da taxonomia lida com a identificação, descrição e classificação dos organismos. Assim, é construída uma estrutura que permite aprimorar nosso entendimento, inclusive sobre a relação evolutiva entre essas espécies. Além disso, a taxonomia pode ter o

papel de auxiliar na comunicação entre os virologistas, e deles com outros setores da sociedade (agricultores, produtores, financiadores...) (Simmonds, *et al.* 2017).

O International Committee on Taxonomy of Viruses (ICTV) é o órgão responsável pela classificação e nomenclatura dos vírus, divulgando as informações taxonômicas através de uma lista mestra de espécies (MLS), atualizada a cada ano (Lefkowitz.E.J., *et al.*, 2018). Os vírus podem ser classificados nas categorias hierárquicas de ordem, família, gênero e espécies, em que cada grupo taxonômico possui um único nome definido e regulamentado. De acordo com a última MLS, o ICTV reconhece 55 ordens, 168 famílias, 1421 gêneros e 6590 espécies virais (Virus Taxonomy: 2019 Release).

Em geral, a taxonomia dos vírus difere da de organismos celulares, principalmente pela falta de um gene universalmente conservado, como o gene do RNA ribossomal. Antigamente, a classificação de um novo vírus pelo ICTV, dependia de uma série de informações como hospedeiro, ciclo de replicação ou a estrutura da partícula viral (Simmonds, *et al.* 2017), com o tempo, esse padrão teve que ser modificado. O avanço nos estudos da metagenômica, provocado por métodos de sequenciamento de alto rendimento, se apresentou como um grande desafio para a classificação de novos vírus. Atualmente, são reveladas diversas novas espécies virais não classificadas, conhecidas apenas por suas sequências. Assim, são necessárias mais ferramentas para entender e prever os aspectos biológicos e estruturais desses vírus putativos.

Uma das formas mais tradicionais de se extrair mais informações a partir das sequência genômica é buscar similaridades com sequências de vírus já caracterizados e classificados. Mas essa metodologia nem sempre se mostra eficiente para responder perguntas sobre o hospedeiro, a patogênese ou mesmo a estrutura desses novos vírus. Por exemplo, vírus que apresentam proteínas capsídicas com similaridade de sequências e até de estrutura tridimensional, podem formar capsídeos com número de subunidades diferentes. Esse é o caso do nudaurelia capensis omega virus (N ω V) que monta um capsídeo T=4 e do FHV que monta um capsídeo T=3, apesar de apresentarem proteínas do capsídeo com enovelamento (Figura 4).

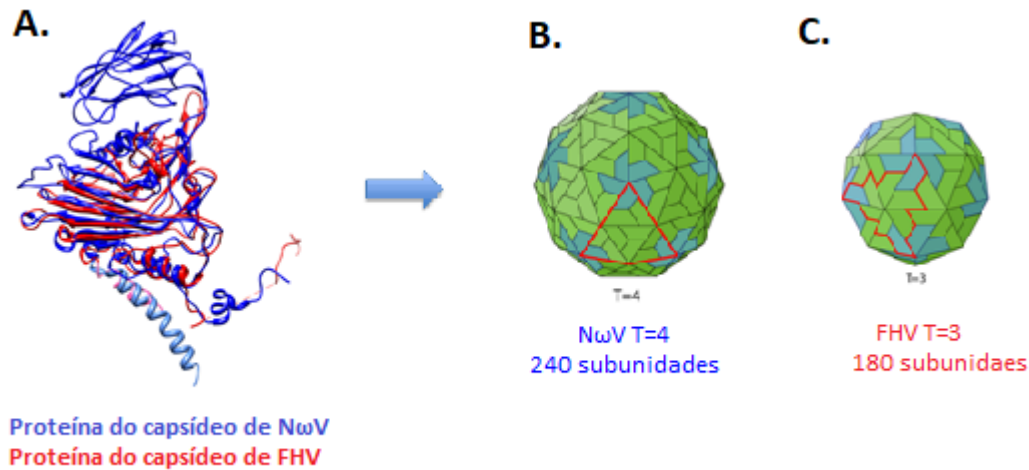


Figura 4- Comparação entre a estrutura dos capsídeos montados por vírus, que apresentam proteínas capsídicas (CP) com similaridade de seqüências e de estrutura tridimensional. (A) Sobreposição das estruturas tridimensionais similares, das seqüências de CP do *Nudaurelia capensis omega virus* (NωV) (azul), e do Flock house virus (FHV) (vermelho). (B e C) Esquema da estrutura do capsídeo icosaédrico T=4 montado pelo vírus NωV e do capsídeo T=3 montado vírus FHV.

Além disso, uma grande quantidade de seqüências recém descobertas são caracterizadas como matéria escura. São consideradas matéria escura, qualquer seqüência de nucleotídeos ou proteína que não possui homologia com nenhuma outra seqüência conhecida (Krishnamurthy e Wang, 2017).

Apesar da escassez de informação sobre essas seqüências, acredita-se que grande parte dessa matéria escura seja composta por organismos virais (Tisza *et al.*, 2020). Isso porque, os vírus são um grupo que apresenta uma vasta diversidade de seqüências. Além de explorar diversas formas de organização do material genético, mecanismos de replicação com a alta taxa de mutação.

Os vírus de ssDNA que contém proteínas associadas à replicação circular (ou CRESS- DNA de *Circular rep-encoding single-strand DNA*), são um exemplo de grupo viral para o qual se tem dificuldade em estabelecer classificação taxonômica. Isso devido a baixa conservação dos genes estruturais e não estruturais, e da dificuldade no cultivo desses pequenos vírus ssDNA (Kazlauskas, D., *et al.*, 2019).

A replicação circular, se refere ao processo capaz de sintetizar rapidamente várias cópias de moléculas circulares de DNA. A proteína associada a replicação circular (Rep) é responsável por dar início mecanismo, ao reconhecer e clivar a origem da replicação. A mesma Rep também é responsável, ao final do processo pela ligação da extremidade das novas fitas formadas, concluindo a formação dos novos genomas de ssDNA circular (Martin *et al.*, 2011).

A maioria dos CRESS DNA tem apenas dois genes, um codifica para a proteína de capsídeo (CP) e outra para a Rep. São conhecidos por serem extremamente difundidos e comuns na natureza, e por apresentarem alta diversidade (Zhao *et al.*, 2019). Além disso, esse grupo costuma apresentar altas taxas de mutação e recombinação. Atualmente se encaixam nessa classificação vírus de sete famílias, *Bacilladnaviridae*, *Circoviridae*, *Geminiviridae*, *Genomoviridae*, *Nanoviridae*, *Smacoviridae* e *Redonoviridae* (Abbas, *et al.*, 2019). Mas a maioria dos vírus CRESS DNA seguem sem classificação. Isso ocorre porque são espécimes sem hospedeiro conhecido e que não apresentam similaridade com nenhum vírus conhecido. Análises filogenéticas das sequências de aminoácidos da Rep de vírus CRESS DNA, revelaram a formação de vários clados correspondente a vírus não classificados, indicando grupos coerentes, que podem representar novas famílias (Kazlauskas, Varsani e Krupovic, 2018).

Como visto, para muitas famílias de vírus, a relação entre carga total do domínio positivo e tamanho do genoma reflete a arquitetura do capsídeo. Assim, hipotetizamos que a determinação desses parâmetros poderia auxiliar na validação de novas espécies putativas para essas famílias. Além disso, a localização de domínios positivamente carregados em proteínas de matéria escura, poderia servir como indício de funcional como proteína capsídica, auxiliando na anotação dessas sequências.

1.3 *Geminiviridae* como caso de estudo:

Para validar nossa hipótese, utilizamos como modelo para esse estudo os vírus da família *Geminiviridae*. A família pertence ao grupo dos vírus que dependem dos braços R para estabilizar seus capsídeos, assim como demonstrado no plot, presente no estudo do nosso grupo (Requião R.D, *et al.*, 2019), e experimentalmente em outros trabalhos (Krupovic, Ravanti e Bamford, 2009; Hesketh *et al.*, 2018). A família *Geminiviridae* é composta por vírus não envelopados, de DNA circular de fita simples, pequenos, variando entre 2.5 a 3.0 kilobases (kb). São transmitidos por insetos da ordem Hemiptera, tendo na maioria dos casos, uma ou algumas espécies vetoriais intimamente relacionadas com um único gênero da família. Tem a capacidade de infectar plantas monocotiledôneas ou dicotiledôneas de diversas culturas. Assim, os vírus da família representam importantes patógenos vegetais, causando doenças que são consideradas uma séria ameaça à segurança alimentar dos países em desenvolvimento, principalmente aqueles localizados nas regiões tropicais e subtropicais do mundo (Moffat, 1999).

O Grupo de estudo *Geminiviridae* do ICTV, propõe uma série de diretrizes sobre como novas sequências de geminivírus devem ser classificadas e nomeadas taxonomicamente. No nível de gênero, os geminivírus são classificados com base na variedade de hospedeiros, vetor de insetos, organização do material genético e identidades de sequências pareadas em todo o genoma (Brown JK *et al.*, 2012). Projetos com base em tecnologias de sequenciamento de alto rendimento tiveram alto impacto aos geminivírus, sendo responsáveis por estabelecer novos gêneros a família, além de identificar uma série de novos geminivírus de linhagens divergentes (Fontenele R.S, *et al.*, 2020).

De acordo com a edição mais atual do ICTV (ICTV, Julho 2019 Release), a família é composta por um total de 485 espécies divididos em nove gêneros, sendo dois principais, *Begomovirus* e *Mastrevirus*, dos quais já se conhece a estrutura, formada por um capsídeo geminado, com diâmetro de aproximadamente 20 a 30 nm de largura (Figura 5). Essa estrutura característica dos geminivírus é composta por dois capsídeos icosaédricos gêmeos incompletos de T=1. As metades são unidas por um vértice pentamérico, totalizando 110 subunidades repetitivas (Zhang, *et al.*, 2001; Xiongbiao Xu, *et al.*, 2019).

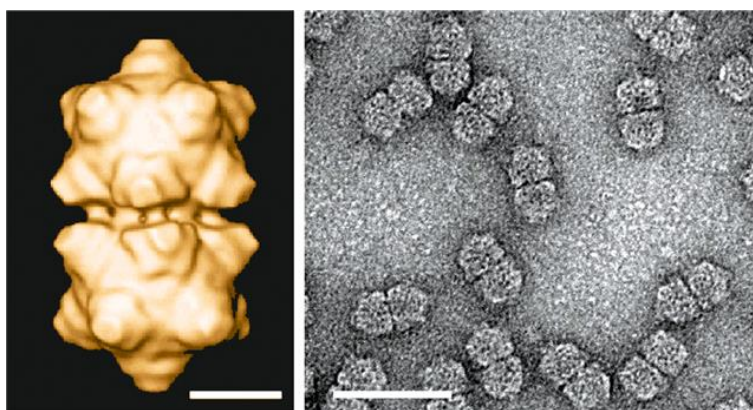


Figura 5 - Estrutura do capsídeo dos *Geminiviridae*. (Esquerda) Reconstrução microscopia crioelétrica do *maize streak virus* (MSV) vista por um eixo de dupla simetria. A barra representa 10 nm. (Direita) Partículas purificadas de MSV coradas com acetato de urânio, mostrando subunidades quase isométricas. A barra representa 50 nm. (Fonte: Zhang et al., 2001)

Estudos com o *African cassava mosaic virus* (ACMV) sugerem a multiplicidade das partículas do capsídeo de geminivírus. Nesses estudos, o ACMV foi inoculado, em linhagens de *Nicotiana benthamiana* não transformadas, e linhagens transgênicas, que continham DNA interferente defeituoso (DI). As partículas virais foram purificadas, analisadas em Southern Blot, e em seguida as frações foram visualizadas por microscopia

eletrônica. Nas frações contendo o DI, foram encontradas pequenas partículas esféricas, com metade do tamanho de uma partícula geminada (Fig. 6B). Também foi possível identificar frações de DNA maiores associadas a partículas provavelmente constituídas por três icosaedros incompletos (Fig 6C). Esses resultados indicam que a estrutura de capsídeo dos geminivírus, apresenta certa plasticidade, sendo capaz de formar partículas de diferentes formas e estruturas (Figura 6), de acordo com o tamanho do DNA empacotado (Frischmuth, Ringel e Kocher, 2001).

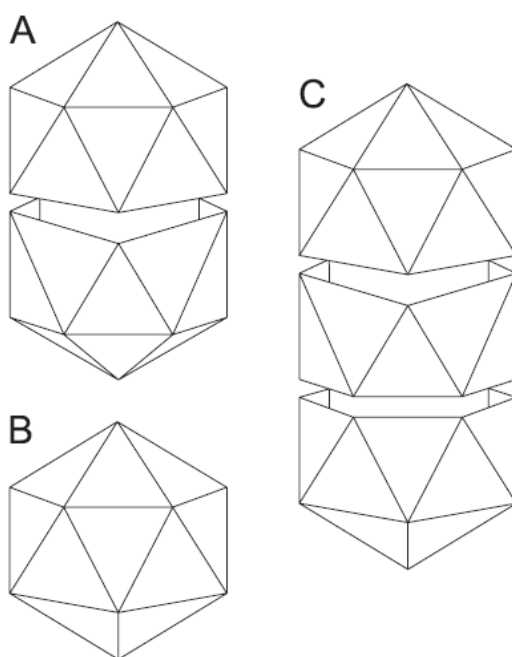


Figura 6 - Representação em esquema das partículas virais que podem ser construídas a partir das proteínas de capsídeo dos geminivírus. (A) Partícula geminada do tipo selvagem. (B) Partícula icosaédrica isométrica. (C) Partícula composta por três icosaédricos incompletos. Essa representação é altamente simplificada para uma maior clareza (Fonte: Krupovic, Ravantii e Bamford, 2009)

Além dos dois gêneros principais, a família ainda contém outros sete gêneros menores (*Becurtovirus*, *Curtovirus*, *Capulovirus*, *Eragrovirus*, *Grablovirus*, *Topocuvirus*, *Turncurtovirus*), e várias novas espécies propostas.

O *Begomovirus* é o maior gênero da família, composto por espécies, capazes de infectar plantas dicotiledôneas, tendo uma ampla gama de hospedeiros. Têm por característica a transmissão mediada por moscas brancas. Os begomovírus podem apresentar genoma monopartido ou bipartido, contendo dois componentes genômicos, denominados DNA-A e DNA-B (Figura 7). Nesse caso, os dois componentes são essenciais para a replicação viral, se diferenciando tanto no número quanto na função dos genomas que codificam. O componente A é responsável por codificar genes, envolvidos no

empacotamento do vírus, e na replicação viral, além da AC4, gene que funciona como um determinante para o desenvolvimento dos sintomas ou como gene supressor do silenciamento. Enquanto isso, o componente B é responsável por codificar dois genes, envolvidos no movimento célula-célula do vírus na planta. Em comum os dois elementos (DNA-A e DNA-B) apresentam uma região intergênica (IR), também chamada de região comum (CR). Essa região é essencial para a replicação e transcrição do genoma, composta por uma estrutura em forma de grampo (“hairpin”) que inclui um motivo de nonanuclotídeo, TAATATTAC, não variável, encontrado em quase todos os geminivírus, agindo como uma área de clivagem para que a Rep inicie a replicação viral (Zerbini, *et al.*, 2017).

O segundo maior gênero, em número de espécies é o *Mastrevirus*, são constituídos por vírus de genoma monopartido, incluindo mais de 40 espécies capazes de infectar hospedeiros monocotiledônios e dicotiledônios, tendo transmissão realizada por diferentes espécies de cigarrinhas. São responsáveis por impactar culturas de milho e trigo, nas regiões da África, norte da Europa, Oriente Médio e Ásia (Kvarnheden *et al.*, 2002).

Apenas os genes da CP e Rep são conservados em todos os nove gêneros de geminivírus. Os membros dos gêneros *Begomovirus*, *Curtovirus*, *Topocovirus* e *Turncurtovirus*, têm organização genômica similares (Figura 7). Apesar de todos os gêneros já estabelecidos apresentarem a proteína de movimento (mp), que nos vírus monopartidos, fica localizado “upstream” a CP, não há homologia detectável entre as mp’s dos diferentes gêneros (Varsani, *et al.*, 2017).

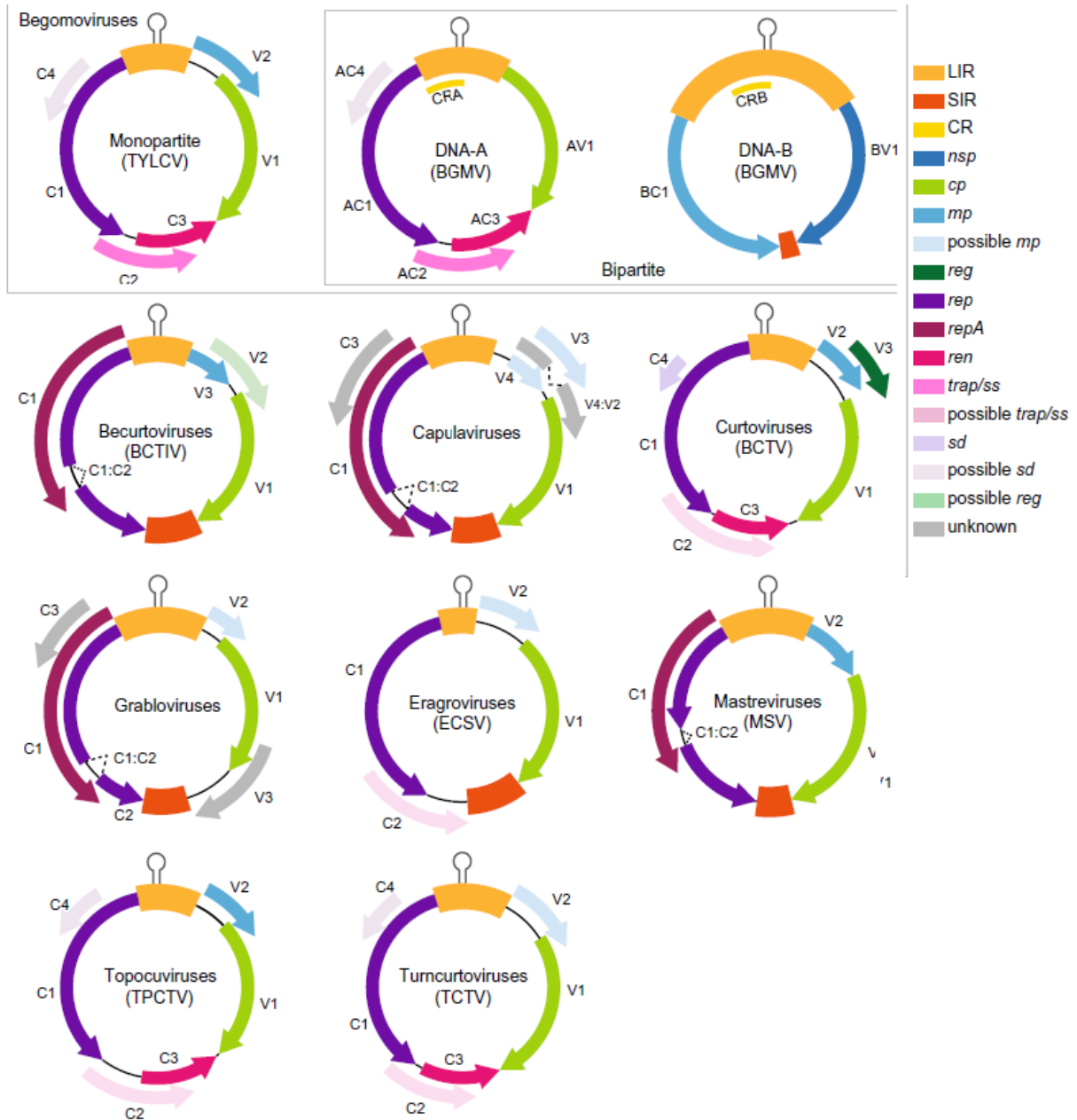


Figura 7 - Organização do genoma dos diferentes gêneros da família *Geminiviridae*. As ORFs(V1, V2, V3, C1, etc.) estão coloridas de acordo com as proteínas que codificam (rep, proteína associada a replicação; ren, protein intensificadora da replicação; trap, proteína ativadora da transcrição; cp, proteína de capsídeo; mp, proteína de movimento; nsp, proteína de transporte nuclear). As regiões IR incluem a estrutura em grampo com a origem de replicação (Adaptado da fonte: Varsani, A *et al.*, 2017).

1.3.1 Relação entre os *Genomoviridae* e os *Geminiviridae*

Para comprovar a nossa hipótese, também utilizaremos como modelo uma família intimamente relacionada aos geminivírus. Anteriormente, esses vírus pertenciam à família *Geminiviridae*, mas em 2016 o ICTV fez uma reclassificação, baseando-se em análises de homologia de sequências, realocando-os na nova família *Genomoviridae* (Martin Krupovic,

et al, 2016). São compostos por vírus não envelopados de simetria icosaédrica, de capsídeo T=1 e com cerca de 20nm de diâmetro. O genoma pequeno, tendo por volta de 2.17 kb, contém apenas dois genes, sendo um correspondente a CP e o outro responsável por codificar a Rep (Figura 8). Duas regiões intergênicas separam as ORFs, e assim como vários outros vírus ssDNA com genoma circular, a região LIR contém uma estrutura em grampo (Figura 8), com a presença de um motivo nonanucleotídico (TAATATTAT), importante para a replicação de círculo rolante. A transcrição ocorre de forma bidirecional à LIR, para codificação dos dois promotores divergentes.

Assim como os geminivírus, eles apresentam a Rep intimamente relacionada com dois domínios conservados (os domínios catalítico e central da Rep), além da presença de motivos conservados para replicação. Em contrapartida, a CP dos genomovírus exibe baixa similaridade de sequência com a proteína dos geminivírus, além disso o tamanho do genoma e o número de genes que codificam diferem consideravelmente. Por fim, os genomovírus carecem da proteína do movimento, que é essencial para que os geminivírus possam estabelecer uma infecção sistêmica nas plantas. Atualmente, a família é composta por nove gêneros, tendo a filogenia baseada na análise da Rep.

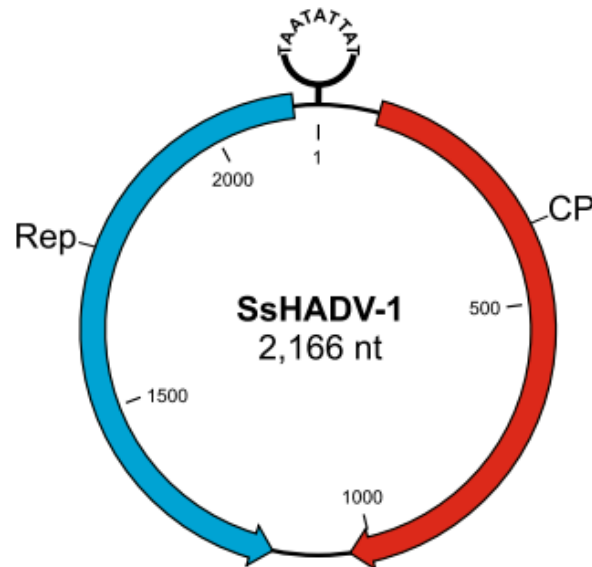


Figura 8 - Mapa genômico do SsHADV-1. O genoma codifica a proteína de início da replicação(rep) e da proteína de capsídeo(cp), representadas pelas setas azul e vermelha, respectivamente. A posição do nonanucleotídeo (TAATATTAT), também está indicada no ápice da estrutura em haste (Fonte: Martin Krupovic, *et al*, 2016).

Através desses modelos virais, pretendemos analisar se a carga dos domínios positivamente carregados, pode ser utilizada para validar a conformação predita para um novo vírus, e dessa forma, contribuir para a anotação e classificação de novas espécies virais.

2. JUSTIFICATIVA

O avanço nos estudos da metagenômica revelaram uma série de novos vírus conhecidos apenas por suas sequências. Por isso, são necessárias mais ferramentas para entender e prever os aspectos biológicos desses vírus. Essas informações podem auxiliar na classificação taxonômica de novas espécies virais, na detecção de sequências de vírus em trabalhos de metagenômica e, conseqüentemente, no diagnóstico e prevenção de doenças virais.

Atualmente, a classificação de novos vírus é baseada em métodos de comparação de sequência e análise de similaridade. Mas essa metodologia apresenta limitações, sendo incapaz de prever, por exemplo, a estrutura dos vírus. Nesse projeto testamos uma nova metodologia *in silico* capaz de fornecer informação sobre a arquitetura do capsídeo viral e identificação de novas proteínas de capsídeo ainda sem relação de similaridade com outras sequências virais já conhecidas.

3. OBJETIVOS

3.1 Objetivo principal

Verificar se a relação entre carga do domínio positivo da proteína do capsídeo e o tamanho do genoma funciona como possível indicador da morfologia de capsídeos virais, permitindo checar se um membro putativo de uma dada família viral apresenta a morfologia de partícula esperada.

3.2 Objetivos específicos.

- Analisar a correlação entre a carga líquida do genoma e do domínio positivo na família *Geminiviridae*;
- Verificar se a análise de correlação entre carga positivas do capsídeo e cargas negativas do genoma consegue separar vírus pertencentes às famílias *Genomoviridae* e *Geminiviridae*;

- Analisar geminivírus putativos derivados de metagenômica e avaliar sua conformidade na relação entre cargas positivas do capsídeo e negativas do genoma, de acordo com o esperado para a família *Geminiviridae*.

4. MATERIAIS E MÉTODOS

4.1 Banco de dados

A construção do banco de dados, incluiu, na maioria dos casos, a pesquisa de todas as sequências de geminivírus ($n = 442$) e genomovírus ($n = 73$) incluídas no 10º relatório do International Committee on Taxonomy of Viruses (ICTV, 2018). As sequências de geminivírus não classificadas foram obtidas no GenBank, banco de dados do National Center for Biotechnology Information (NCBI, 2020). As sequências de isolados da nova espécie OpV1 ($n = 79$), foram obtidas no NCBI nucleotide. O genoma do OpV1 foi sequenciado em estudo recente, onde foram analisadas amostras de cactos na região da América do Norte (Fontenele R.S, *et al.*, 2020). O conjunto de dados foi selecionado com cuidado, para que os dados correspondessem a vírus com sequência completa do genoma. Portanto utilizamos sempre que disponível para consulta do genoma o NCBI RefSeq.

Utilizamos o Uniprot.org, a fim de filtrar as sequências de proteína do capsídeo, correspondentes as espécies analisadas. Portanto, a partir dos identificadores das “RefSeq nucleotide”, utilizamos as ferramentas de pesquisa disponíveis no Uniprot, para obter as proteínas relacionadas ao organismo referente aquele código. Através da opção de buscas avançadas, fomos capazes de filtrar a pesquisa para a obtenção exclusiva das proteínas de capsídeo.

4.2 Cálculo da carga líquida

O programa desenvolvido em laboratório, CargaFlex, foi utilizado para a identificação e cálculo dos trechos positivamente carregados (Q_{max}), presentes nas CPs, das sequências virais. O programa realiza a identificação por meio da análise da sequência primária de uma determinada proteína, calculando, a carga líquida dessa sequência, em quadros consecutivos. Para determinação da carga líquida os resíduos de arginina (R) e lisina (K) são considerados tendo carga +1; ácido glutâmico e aspártico, são considerados tendo carga -1, e os demais resíduos 0. Esses parâmetros simplificados são equivalentes a um cálculo usando cargas parciais de aminoácidos em pH 7,4 de acordo com o valor pKa usado na equação de Henderson-Hasselbach (Requião, *et al.*, 2017). É utilizado como

“input” para o algoritmo, um arquivo fasta contendo as sequências de aminoácidos das CPs dos vírus analisados. O algoritmo estabelece uma janela fixa, com um número pré determinado de aminoácidos para realizar a análise, como representado no esquema abaixo pelo retângulo tracejado em azul (Figura 9). Normalmente nas análises, utilizamos uma janela de 30 resíduos, pois é considerado o tamanho médio da região N terminal não estruturada para diversos vírus, logo, um bom ponto de partida para as nossas análises. A fim de simplificar, nesse exemplo, utilizamos uma janela de 8 aminoácidos. O valor da carga e a posição do aminoácido são temporariamente salvos na memória (Fig. 9B). Em seguida, o programa avança para o próximo trecho (ex, n°1 a n°8), realizando a mesma análise. Como o valor da carga desse segundo trecho é menor do que o anteriormente salvo, o programa não substitui os dados memorizados (Fig. 9C). O programa continua avançando, calculando a carga nos trechos consecutivos de 8 aminoácidos, até atingir a quantidade total de aminoácidos da proteína em análise. (Fig. 9D). O processo então é reiniciado, em janelas maiores de aminoácidos (ex, 9 resíduos) (Fig. 9E). A expansão no tamanho da janela analisada permite a identificação da concentração de carga (Q_c), fator adicional na identificação do domínio positivamente carregado. A análise continua até que sejam calculados trechos de todos os tamanhos possíveis. Caso encontre um valor de Q_{max} mais alto, com Q_c maior ou igual do que o salvo anteriormente na memória, os valores da carga e a posição do aminoácido serão substituídos (Requião R.D, *et al.*, 2019). Ao final, o programa retorna à informação da janela com maior carga líquida na cp e com maior concentração de carga.

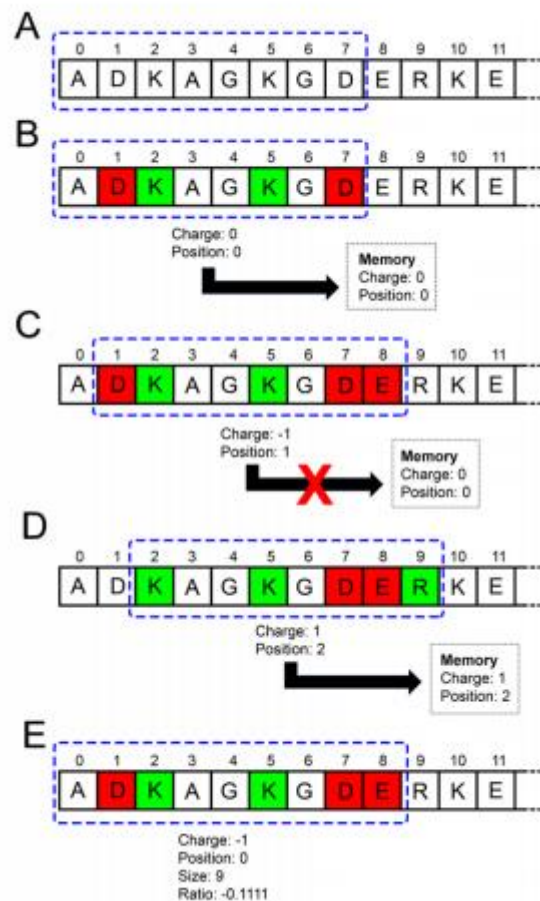


Figura 9 - Cálculo da carga líquida. O trecho analisado é destacado como um retângulo azul tracejado. (B) A carga líquida é então calculada. Para cada lisina (K) ou arginina (R), é contado +1, e para cada ácido aspártico (D) ou ácido glutâmico (E), é contado -1. O valor da carga e a posição do primeiro aminoácido são salvos temporariamente na memória. (C e D) Então, nosso algoritmo avança para o próximo trecho, do aminoácido nº 1 ao aminoácido nº 8. Ele continua avançando através da proteína até atingir a quantidade total de aminoácidos dessa proteína. Se o valor da carga de um desses outros segmentos for maior que o anteriormente salvo na memória, o valor e a posição atuais serão substituídos. (E) Um segundo algoritmo expande as possibilidades. Quando termina a busca pela carga mais alta em um quadro de 8 aminoácidos em uma proteína, ele reinicia o mesmo processo, mas com um quadro de 9 aminoácidos. Isso é repetido até que todos os tamanhos possíveis sejam calculados. (Fonte: Requião R.D, *et al.*, 2019)

4.3 Correlação entre a carga do genoma e do domínio da CP

Inicialmente, o cálculo da carga líquida do interior do capsídeo foi feito multiplicando o valor da carga do domínio positivo (determinada pelo programa CargaFlex), pelo número de subunidades que formam o capsídeo (ex, *Geminiviridae* = 110; *Genomoviridae* = 60). Já a determinação da carga do genoma, foi feita pelo cálculo em que cada grupo fosfato é considerado tendo -1 de carga. As análises estatísticas, como a linha de regressão linear e os gráficos exibidos foram gerados no Graph-pad Prism 7.0 software. A margem de verificação nos gráficos foi determinada de acordo com a distribuição dos gêneros *Begomovirus* e *Mastrevirus*, não sendo um dado estatístico.

5. RESULTADOS

5.1 *Begomovirus* e *Mastrevirus* apresentam uma correlação positiva entre a carga do domínio positivo e a carga do genoma.

Para analisar a correlação entre a carga líquida do genoma e do domínio positivo na família *Geminiviridae*, nós selecionamos cuidadosamente nosso conjunto de dados, selecionando entradas referentes a geminivírus com sequência genômica completa. Esses dados permitem a determinação da carga do genoma (Q_{genome}). As sequências de proteína do capsídeo foram obtidas no banco de dados do total UniProt KB/Swiss-Prot reviewed. Para determinar o R-arm, nós utilizamos o programa desenvolvido em laboratório, CargaFlex, já detalhado acima. Em seguida, a carga total do capsídeo foi calculada multiplicando a Q_{max} da sequência da CP pelo número de subunidades que formam o capsídeo (Total Q_{max}).

Como ponto de partida, para nossas análises, recolhemos as sequências de proteínas do capsídeo dos maiores gêneros da família em número de espécies, *Begomovirus* e *Mastrevirus*. Sendo 381 delas referentes aos *Begomovirus* e 37 referentes aos *Mastrevirus* ($n = 418$). Nós conseguimos determinar uma correlação positiva entre a carga do genoma e o Total Q_{max} (Pearson $r = 0,54$) para o conjunto de dados dos dois gêneros (Figura 9). Por serem gêneros dos quais já se conhece a estrutura, o resultado da análise dessas sequências serve como referência de como se comporta a partícula geminada em relação a proporção, carga do genoma e do domínio positivo. Portanto, em cima desses dados calculamos uma linha de regressão linear e determinamos, por meio da distribuição de *Begomovirus* e *Mastrevirus*, o nosso intervalo de verificação, indicado pelas linhas pontilhadas. (Figura 9) Assim é possível prever se as sequências que serão analisadas posteriormente se comportam como esperado para o capsídeo geminado da família *Geminiviridae*.

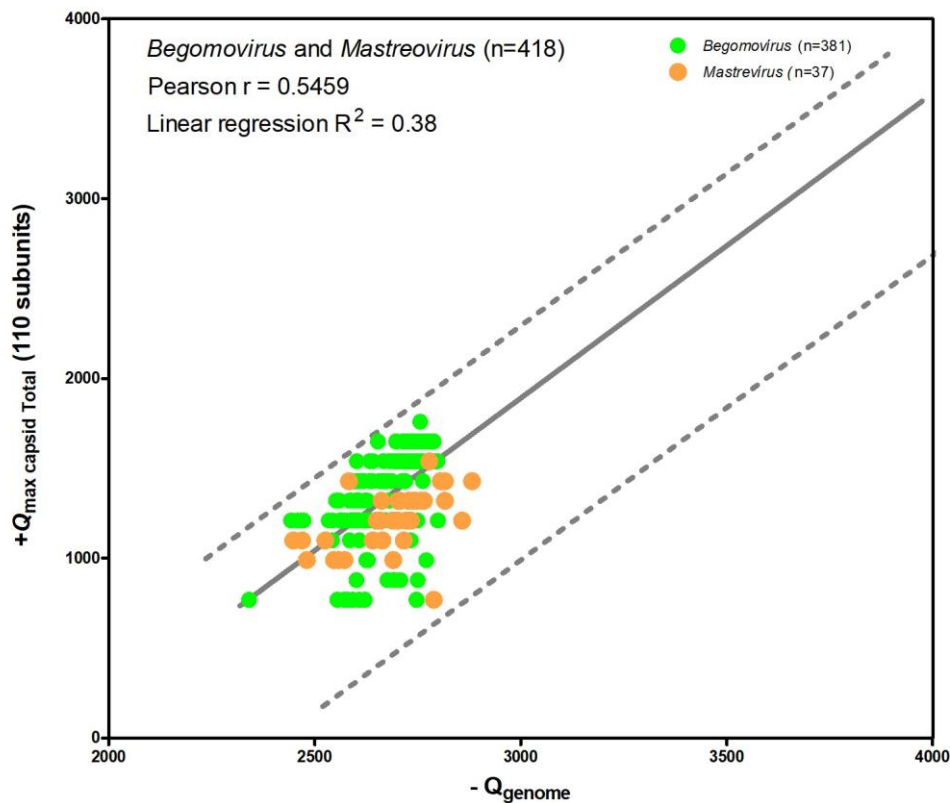


Figura 10. *Begomovirus* e *Mastrevirus* apresentam uma correlação positiva entre a carga do domínio positivo e a carga do genoma. O valor da carga do domínio positivo foi calculado como descrito em métodos. Para o cálculo da Q_{genome} cada nucleotídeo foi considerado como -1 de carga. A carga do domínio positivo da proteína do capsídeo foi multiplicada pelo número de subunidades que formam o capsídeo (110) da família *Geminiviridae*. A linha cinza representa o ajuste linear calculado a partir do banco de dados de *Begomovirus* (verde) e *Mastrevirus* (laranja). A linha pontilhada representa o intervalo de confiança inferido pela distribuição dos dois gêneros. Foram analisadas 381 sequências de cp do gênero *Begomovirus* e 37 sequências do gênero *Mastrevirus*, totalizando 418 sequências. Os resultados de correlação Pearson obtidos a partir das 418 sequências estão mostrados no gráfico.

Em seguida, adicionamos à análise membros dos sete gêneros menores da família *Geminiviridae*, que incluem, *Becurtovirus* (n=1), *Capulavirus* (n=1), *Curtovirus* (n=2), *Eragrovirus* (n=2), *Grablovirus* (n=2), *Topocurvirus* (n=1) e *Turnucovirus* (n=3). Além disso, adicionamos sequências de geminivírus altamente divergentes (OpV1, n=79) (Fontenele R.S, *et al.*, 2020), e não classificados (n=12). Portanto, ao final foram analisadas 521 sequências de CPs da família *Geminiviridae* (Figura 10).

A partir do nosso intervalo de confiança, podemos determinar, que os vírus pertencentes aos outros gêneros menores de geminivírus apresentam uma proporção esperada entre a carga líquida do capsídeo e a carga do genoma. Analisamos também sequências do geminivírus altamente divergentes, algumas como *Opuntia virus 1* (OpV1), *Juncus maritimus associated virus* (JmaV), *Limeum africanum associated virus* (LaaV), *Apple geminivirus* (AGmV) e *Tomato apical leaf curl virus* (TALCV), que apresentaram o

comportamento esperado para o capsídeo geminado, ficando próximos à linha de ajuste. Outras sequências de vírus não classificados, como Citrus chlorotic dwarf associated virus (CCDaV) e Temperate fruit decay-associated vírus (TFDaV) também permaneceram dentro da linha de ajuste, se comportando como o esperado para um capsídeo de 110 subunidades. Esses resultados confirmam a classificação dessas espécies, ainda que não tenham um gênero definido dentro da família.

Enquanto isso, outros vírus não classificados, se localizam fora da nossa margem de erro, estão entre eles o Baminivirus, Niminivirus (NimiV) e Passion fruit chlorotic mottle virus (PCMoV). Consideramos essas sequências que desviaram da nossa linha de ajuste, como outliers. Nossa hipótese é de que essas espécies de vírus não classificados sejam capazes de montar estruturas alternativas do capsídeo.

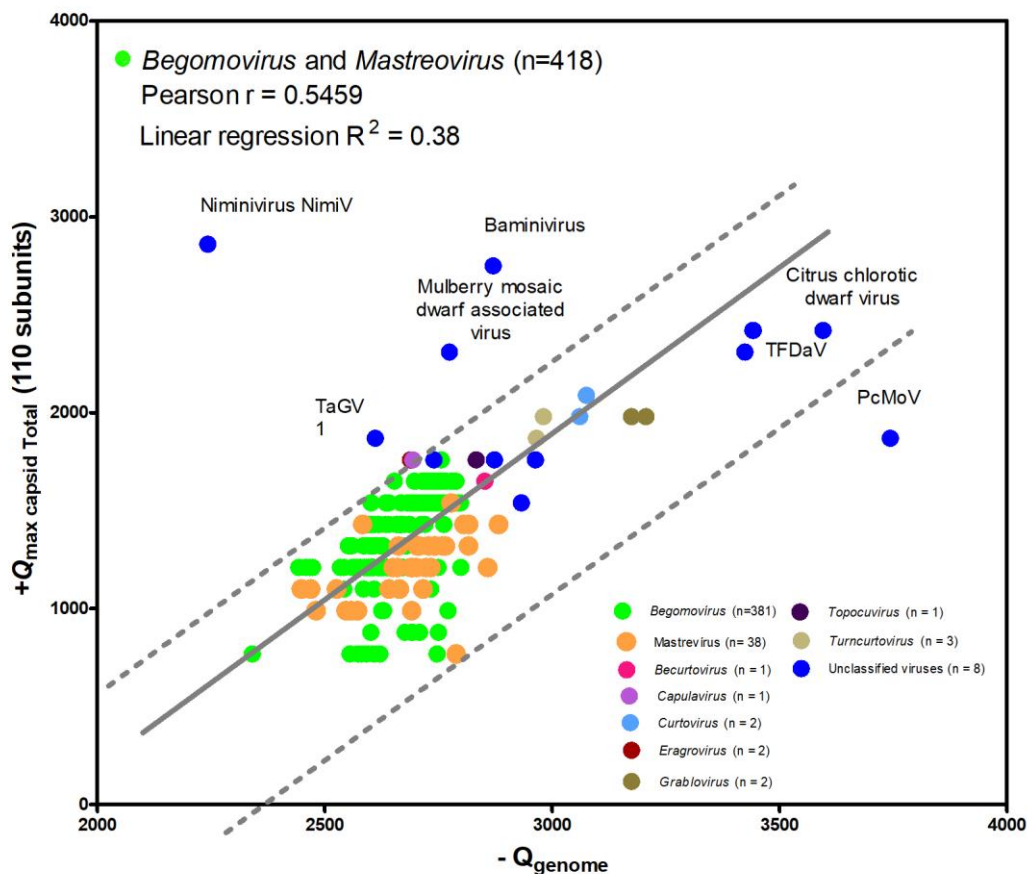


Figura 11. Nem todos os *Geminiviridae* se comportaram como o esperado para um capsídeo geminado. O valor da carga do domínio positivo foi calculado como descrito em métodos. Para o cálculo da Q_{genome} cada nucleotídeo foi considerado como -1 de carga. A carga do domínio positivo da proteína do capsídeo foi multiplicada pelo número de subunidades que formam o capsídeo (110) da família *Geminiviridae*. A linha cinza representa o ajuste linear calculado a partir do banco de dados de *Begomovirus* e *Mastrevirus* (verde). A linha pontilhada representa o intervalo de confiança inferido pela distribuição dos dois gêneros. No total

foram analisadas 518 sequências de cp dos vírus da família *Geminiviridae*, pertencentes aos gêneros *Begomovirus* (verde), *Mastrevirus* (laranja), *Becurtovirus* (rosa), *Curtovirus* (azul), *Eragrovirus* (vinho), *Capulavirus* (roxo claro), *Grablovirus* (marrom escuro), *Topocovirus* (roxo escuro), *Turnucurtovirus* (cinza), além dos vírus não classificados (azul) e o geminivírus divergente OpV1 (preto). Algumas espécies foram consideradas outliers.

Para os vírus que se distribuíram acima da nossa margem de erro, propomos que elas montam um capsídeo de 60 subunidades, $T = 1$. Enquanto, sugerimos que o PCMoV, localizada abaixo da margem de erro, possa montar um capsídeo de 150 subunidades (Figura 11). Para testar essa hipótese, resolvemos fazer um ajuste no cálculo da carga interna do capsídeo dessas espécies consideradas como outliers. Desse modo, para os vírus localizados acima da margem de erro, NimiV, Banimivírus, mulberry mosaic dwaft associated vírus (MMDaV) e tomato associated geminivírus (TaGV1), multiplicamos a carga do domínio positivo dada pelo programa por 60 subunidades. Já para PCMoV, localizado abaixo da margem de erro, multiplicamos a carga do domínio por 150 subunidades (Figura 11A).

Ao fazer o ajuste, pode-se observar que na maioria dos casos, os vírus se distribuíram melhor dentro do nosso gráfico (Figura 11B). Dos vírus propostos para montar um capsídeo mais simples, de $T = 1$, *Banimivírus*, MMDaV e TaGV1, se localizaram dentro da nossa margem de erro, após o ajuste de cargas. Entretanto, a espécie *Nimivírus* mesmo após o ajuste, permaneceu fora da margem de erro, ainda que tenha ficado mais próximo quando comparado com a posição anterior. No caso do PcMoV, para a qual propomos a montagem de um capsídeo de 150 subunidades, após a mudança, ele passou a se localizar dentro da linha de ajuste. Nossos resultados indicam que provavelmente essas espécies adotam uma estrutura alternativa do capsídeo.

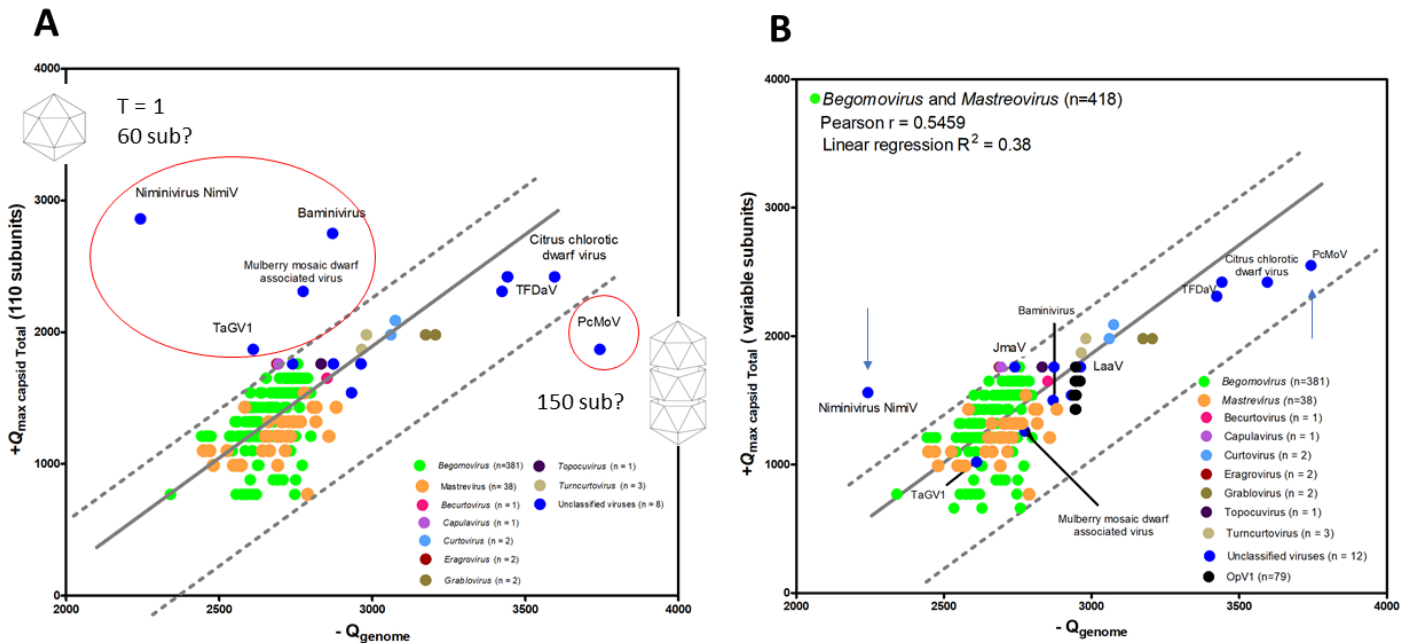


Figura 12 A. O valor da carga do domínio positivo foi calculado como descrito em métodos. Para o cálculo da Q_{genome} cada nucleotídeo foi considerado como -1 de carga. A carga do domínio positivo da proteína do capsídeo foi multiplicada pelo número de subunidades que formam o capsídeo (110) da família *Geminiviridae*. A linha cinza representa o ajuste linear calculado a partir do banco de dados de *Begomovirus* e *Mastrevirus* (verde). A linha pontilhada representa o intervalo de confiança inferido pela distribuição das duas famílias. No total foram analisadas 518 sequências de cp dos vírus da família *Geminiviridae*, pertencentes aos *Begomovirus* + *Mastrevirus* (verde), *Becurtovirus* (rosa), *Curtovirus* (azul), *Eragrovirus* (vinho), *Capulavirus* (roxo claro), *Grablovirus* (marrom escuro), *Topocovirus* (roxo escuro), *Turnucurtovirus* (laranja), além dos vírus não classificados (azul) e o geminivírus divergente OpV1 (preto). Nem todos os *Geminiviridae* se comportaram como o esperado para um capsídeo geminado, essas espécies outliers podem apresentar uma arquitetura alternativa do capsídeo como mostrado no gráfico. **B.** Propomos para os vírus que se localizaram fora da margem de erro, um ajuste no cálculo da carga do capsídeo para uma conformação alternativa.

5.2 *Gemonoviridae*: família recém criada, intimamente relacionada aos *Geminiviridae*

A fim de validar nossa metodologia, adicionamos à nossa análise a família *Genomoviridae* (Figura 12). As espécies da nova família eram classificadas como sendo pertencentes a família *Genimiviridae*. Recentemente esses vírus foram reclassificados com base em análises de homologia de sequências (Martin Krupovic, *et al.*, 2016). Além disso, partir da microscopia da única espécie isolada da família, *Sclerotinia sclerotiorum hypovirulence-associated DNA virus 1* (SsHADV-1), foi possível observar que o vírus não monta um capsídeo geminado, como o dos geminivírus, mas um de $T = 1$ (60 subunidades) (Xiao Yu *et al.*, 2010). Utilizamos essa relação entre as duas famílias, fazendo dos geminivírus um controle para nossa análise, com o objetivo de verificar se nossa

metodologia seria capaz de identificar que a família *Genomoviridae* não obedece a arquitetura geminada.

Para realizar essa análise montamos um novo banco de dados com as sequências da proteína de capsídeo dos vírus da família *Genomoviridae*, totalizando ao final, 73 espécies. As sequências foram obtidas por meio de consulta do UniProt KB/Swiss-Prot reviewed. Determinamos o R arm das cps por meio do nosso programa desenvolvido em laboratório, CargaFlex. Ao calcular a carga total do capsídeo dos *Genomoviridae*, multiplicamos a Q_{max} da sequência da CP por 110, número de subunidades que formam o capsídeo geminado, para que fosse possível fazer a comparação entre as análises de correlação das duas famílias. Assim emulando, o que se acreditava anteriormente, que todas essas espécies pertenciam a uma mesma família. Em seguida, , adicionamos à análise, os vírus , pertencentes a família *Geminiviridae*.

Comparando o comportamento das sequências na análise de correlação das duas famílias, foi possível observar que as espécies referentes aos *Genomoviridae* (marrom), se distribuíram de forma deslocada quando comparada a posição dos *Geminiviridae* (verde), se localizando acima da nossa margem de erro (Figura 12). Esse resultado deixa clara a diferença conformacional entre as duas famílias, em que os genomovírus apresentaram um padrão de genoma e carga do domínio diferente. Logo, podemos determinar que nossa metodologia foi capaz de identificar corretamente que os vírus da família *Genomoviridae* não tem um comportamento esperado para um capsídeo geminado, obedecendo, portanto, outra arquitetura.

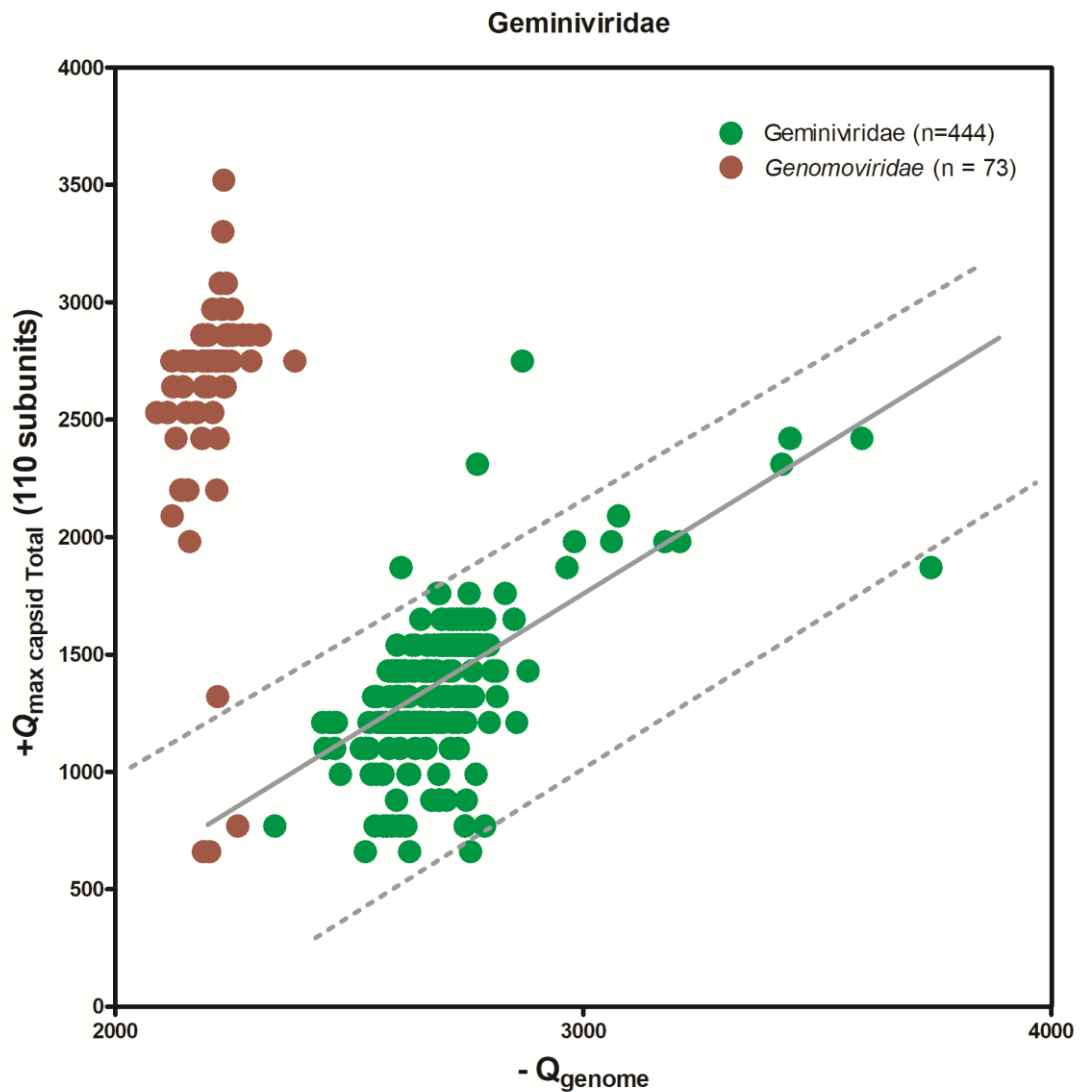


Figura 13. *Gemonoviridae*: família recém criada, intimamente relacionada aos *Geminiviridae*. Foram analisadas O valor da carga do domínio positivo foi calculado como descrito em métodos. Para o cálculo da Q_{genome} cada nucleotídeo foi considerado como -1 de carga. A carga do domínio positivo da proteína do capsídeo foi multiplicada pelo número de subunidades que formam o capsídeo (110) da família *Geminiviridae*. A linha cinza representa o ajuste linear calculado a partir do banco de dados de *Begomovirus* e *Mastrevirus*. Foram analisados, para fim de comparação, 466 seqüências das CPs de Geminiviridae (verde) e 117 seqüências de proteína do capsídeo dos vírus da família *Genomoviridae* (marrom).

6. DISCUSSÃO

A descoberta de uma série de novos geminivírus tem sido facilitada pelos recentes avanços nas metodologias de sequenciamento de nucleotídeos, além de técnicas de enriquecimento do genoma viral, como a amplificação por círculo rolante (RCA) (Fontenele, R S., *et al.*, 2018). O grande número de novas sequências vindas da metagenômica, aponta para a necessidade da caracterização e classificação desses vírus. Essa caracterização pode auxiliar inclusive na detecção de possíveis novos patógenos, já que a família *Geminiviridae* é conhecida por causar doenças severas em diversos cultivos de importância econômica (Rojas, M. R., *et al.*, 2018). Nosso trabalho apresenta uma metodologia desenvolvida em laboratório, que propõe usar a relação entre carga total do domínio positivo e tamanho do genoma, que para algumas famílias reflete a arquitetura do capsídeo, tendo assim, potencial para auxiliar na classificação de espécies putativas da família.

No cenário atual são indicados para classificação de novas espécies de geminivírus, uma análise padronizada que inclui, consultas ao BLASTn, a fim de identificar os membros que tenham mais semelhança com a nova espécie. Outra etapa é a utilização de algoritmos de alinhamento, com exclusão de gaps, para calcular a identidade entre cada par de sequência do conjunto de dados (Brown, J. K., *et al.*, 2015). São considerados diferentes “cut- off’s” (limiares de identidade) para cada gênero da família. Estão entre algumas críticas feitas quanto a taxonomia dos *Geminiviridae*, o grande número de espécies dentro do gênero begomovírus e o reconhecimento de novas espécies baseadas somente na comparação de sequências (Van Regenmortel MH, *et al.*, 2013). Isso dificulta, por exemplo, a caracterização de espécies com alta taxa de recombinação, ou muito divergentes (Brown, J. K., *et al.*, 2015). A nossa metodologia se apresenta como um auxílio na classificação desses vírus permitindo a adição da consideração de propriedades biológicas dessas espécies no momento da classificação.

Nossas análises identificaram uma correlação positiva entre a carga do domínio positivo e a carga do genoma dos vírus da família *Geminiviridae*. A partir desses resultados, somos capazes de confirmar que os geminivírus utilizam o domínio positivamente carregado localizado nas extremidades da proteína de capsídeo para estabilizar o capsídeo viral, corroborando com resultados mostrados anteriormente pelo trabalho do nosso grupo (Requião R.D, *et al.*, 2019) e experimentalmente ((Krupovic, Ravantti e Bamford, 2009; Hesketh *et al.*, 2018)

Para que fosse possível avaliar a conformação do capsídeo do nosso modelo viral, utilizamos os gêneros *Begomvirus* e *Mastrevirus* dos quais a estrutura já é conhecida (Zhang, *et al.*, 2001; Xiongbiao Xu, *et al.*, 2019), como base para o nosso cálculo de ajuste linear. Utilizamos esses dados para analisar se outras espécies da família e geminivírus putativos da metagenômica, têm a conformação predita para a família. Na nossa análise todos os sete gêneros menores (*Becurtovirus*, *Capulavirus*, *Curtovirus*, *Eragrovirus*, *Grablovirus*, *Topocuvirus*, *Turncurtovirus*) já estabelecidos da família *Geminiviridae*, apresentaram a proporção de carga do genoma e do domínio similares, corroborando a hipótese de que eles compartilham a mesma arquitetura de capsídeo geminado (Varsani A., *et al.*, 2017).

Adicionamos a análise sequências de geminivírus divergentes, revelados pela metagenômica como os OpV1, nesse caso, também foi possível determinar que o vírus apresenta a confirmação esperada para um capsídeo geminado, esses dados reforçam a caracterização dessa nova espécie como pertencentes à família *Geminiviridae* (Fontenele R.S *et al.*, 2020). Da mesma forma na análise vírus não classificados como, AGmV, CCDaV, JmaV, LaaV, TALCV e TDaFV, nossos resultados auxiliam na caracterização dessas sequências como geminivírus (Loconsole, G. *et al.*, 2012; Liang, P. *et al.*, 2015; Vaghi Medina, C.G. *et al.*, 2017; Claverie, S. *et al.*, 2018.). Contribuem para a descoberta de novas espécies divergentes, a grande frequência de recombinação entre as espécies dos geminivírus, além de taxas de substituição tão altas quanto a de vírus ssRNA (Duffy and Holmes, 2008; Monjane *et al.*, 2011).

A descoberta de novos geminivírus divergentes podem revelar mais sobre a ecologia e evolução dos *Geminiviridae* (Claverie, S., *et al.*, 2018). Algumas das espécies recém-descobertas, caracterizadas pela nossa metodologia, como JmaV e LaaV, foram inicialmente isolados em hospedeiros não cultivados, onde não causavam sintomas óbvios. Acredita-se que essas espécies de plantas não cultivadas tenham um papel no surgimento de

novas doenças, com a introdução de uma planta exótica em seu ecossistema (Susi *et al.*, 2017; Claverie, S., *et al.*, 2018).

Também fomos capazes de identificar algumas espécies não classificadas que ficaram deslocadas na nossa análise. Estudos já publicados demonstraram que os geminivírus são capazes de formar estruturas do capsídeo de diferentes formas e estruturas, apresentando capsídeos de triangulações diferentes da geminada (Frischmuth, Ringel e Kocher, 2001). Com base nesses dados, nossa hipótese é de que algumas espécies indicadas como outliers na nossa análise, podem ter uma arquitetura alternativa de vírus. Os resultados seguintes obtidos pela nossa análise, com a correção do cálculo da carga líquida interna do capsídeo das espécies que desviam do padrão, corrobora com esses estudos. Por exemplo, os *Banimivírus*, que antes apareciam acima da nossa margem de erro, ao considerar para eles a montagem de um capsídeo mais simples, de 60 subunidades, a sequência passou a se localizar dentro da linha de ajuste linear. Do mesmo modo, o vírus PCMoV, que se localizava abaixo da margem de erro, ao se sugerir a montagem de um capsídeo de 150 subunidades, pode-se observar uma melhora na distribuição do vírus, ficando mais próximo da linha de ajuste linear.

Alguns desafios podem diminuir a precisão da análise na avaliação de algumas sequências, principalmente aquelas consideradas como altamente divergentes ou matéria escura, para as quais existe uma escassez de informações mais exatas. Nossa metodologia foi capaz de identificar três das espécies de begomovírus que se localizaram no limite da nossa margem de erro, apresentando os valores de carga do domínio mais baixo da família, são eles, *Duranta leaf curl virus*, *Whitefly-associated begomovirus 3* e *Cabbage leaf curl virus* ($Q_{\text{máx}} = 660$). Ao comparar esses vírus com outros begomovírus, podemos perceber que os três apresentam sequências de proteína do capsídeo mais curtas. Isso pode indicar uma falha na anotação dessas sequências, com possíveis lacunas, o que leva a identificação pouco precisa dos domínios. Da mesma forma, sequências com falta de exatidão no tamanho do genoma também podem levar a uma correlação imprecisa, já que o cálculo da carga do genoma, nesse caso, é afetado. Para transpor essas dificuldades, outros critérios de caracterização devem ser considerados em conjunto à análise de correlação. Esses fatores também reforçam a necessidade de se continuar investindo em melhorias para aumentar a confiabilidade da análise.

Anteriormente, as espécies de genomovírus eram considerados pertencentes à família *Geminiviridae*. Somente em 2016, por meio de uma reclassificação do ICTV, baseados em análises mais detalhadas de homologias das sequências da Rep e da proteína

do capsídeo, os vírus foram reclassificados à família *Genomoviridae* (Martin Krupovic, *et al.*, 2016). Considerando essa relação entre as duas famílias resolvemos utilizar as sequências de genomovírus como um controle para validar as nossas análises. Portanto, ao analisar as sequências das duas famílias, para efeito de comparação, no cálculo da carga líquida interna do genoma, consideramos a montagem de capsídeos de 110 subunidades tanto para as sequências de geminivírus como a de genomovírus. Nossa análise foi capaz de identificar corretamente, que os *Genomoviridae* (T=1, 60 subunidades) não obedecem a arquitetura geminada. Como exceção três sequências de genomovírus, parecem se comportar de acordo com a partícula geminada, segundo a nossa análise. Quando comparada com as outras sequências da família, se destacam por ter a carga do domínio mais baixa. Isso pode estar relacionado a montagem de um capsídeo com triangulação maior, diminuindo a necessidade de uma carga mais elevada. Apesar disso, análises filogenéticas dos genes Reps e Cps dessas sequências, demonstram uma maior aproximação com os geminivírus (Krupovic, *et al.*, 2016). Outra hipótese é a identificação de um domínio truncado, que como dito anteriormente pode estar relacionado com uma má anotação da sequência. Essas desconfianças levam a um olhar mais detalhado para caracterização desses vírus.

De forma geral esse resultado confirma a confiabilidade da nossa metodologia, além de reforçar o potencial da análise em caracterizar e classificar novos vírus vindos de análise da metagenômica.

7. CONCLUSÃO

A correlação positiva entre a carga do R arm e a carga do genoma achada nos *Geminiviridae*, indica que os vírus da família utilizam o domínio positivamente carregado para estabilizar seus capsídeos.

Nossas análises conseguiram corroborar com a caracterização de novas espécies de geminivírus recentemente descobertos em projetos de sequenciamento em larga escala. Mas nem todos os vírus que são atualmente atribuídos a família *Geminiviridae* apresentam um padrão de correlação da carga do domínio e do genoma compatível com a de uma partícula geminada. Essas observações indicam que essas espécies que desviam do padrão podem apresentar uma estrutura de capsídeo diferente.

A arquitetura de capsídeo inferida pela nossa análise de correlação entre o genoma e a carga do domínio positivo, identificou corretamente que os vírus da família *Genomoviridae* não apresentam uma partícula geminada. Essa análise demonstra que nossa metodologia é capaz de diferenciar padrões de montagem da partícula capsídica.

Os resultados demonstram o potencial da análise no auxílio das caracterizações de novas sequências identificadas, sendo necessário que se continue investindo no refinamento da ferramenta, a fim de torná-la ainda mais confiável.

8. REFERÊNCIAS BIBLIOGRÁFICAS

1. Abbas AA, Taylor LJ, Dothard MI, e. Redondovirida. A Family of Small, Circular DNA Viruses of the Human Oro-Respiratory Tract Associated with Periodontitis and Critical Illness [published correction appears in *Cell Host Microbe*. 2019 Aug 14;26(2):297]. *Cell Host Microbe*. 2019;25(5):719-729.e4. doi:10.1016/j.chom.2019.04.001
2. Arvind Varsani, Jesu´s Navas-Castillo, Enrique Moriones, Cecilia Hernandez-Zepeda, Ali Idris, Judith K. Brown, F. Murilo Zerbini e Darren P. Martin (2014) Establishmen, identificou corr of three new genera in the family *Geminiviridae*: Becurtovirus, Eragrovirus and Turncurtovirus. *Archives of Virology*, 159(8), 2193–2203. Disponível em: DOI 10.1007/s00705-014-2050-2
3. Bahder, B. W., Zalom, F. G., Jayanth, M., e Sudarshana, M. R. (2016). Phylogeny of Geminivirus Coat Protein Sequences and Digital PCR Aid in Identifying *Spissistilus festinus* as a Vector of Grapevine red blotch-associated virus. *Phytopathology*, 106(10), 1223–1230. doi:10.1094/phyto-03-16-0125-fi
4. Berns K, Parrish CR. Parvoviridae. In: Knipe D M, Howley P M, Cohen J I et al. (eds.). *Fields Virology*, 6th ed., vol II. Philadelphia: Wolters Kluwer/Lippincott Williams & Wilkins, 2013. p. 1768-1791.
5. Berns K e Parrish CR. Parvoviridae. In: Knipe D M, Howley P M, Cohn J I et al. (eds.). *Fields Virology*, 6th ed., vol II. Philadelphia : Wolters Kluwer/Lippincott Williams & Wilkins, 2013. p. 1768-1791
6. Briddon, R. W., Pinner, M. S., Stanley, J., & Markham, P. G. (1990). Geminivirus coat protein gene replacement alters insect specificity. *Virology*, 177(1), 85–94. doi:10.1016/0042-6822(90)90462-z
7. Briddon, R.W.; Bedford, I.D.; Tsai, J.H. e Markham, P.G.(1996). Analysis of the nucleotide sequence of the treehopper-transmitted geminivirus, tomato pseudo-curly top virus, suggests a recombinant origin. *Virology* 1996, 219, 387–394.
8. Brown JK, Fauquet CM, Briddon RW, Zerbini M, Moriones e, Navas-Castillo J (2012) Virus taxonomy: Ninth report of the International Committee on Taxonomy of Viruses. Academic press, London
9. Caspar, D. L. D., e Klug, A. (1962). Physical Principles in the Construction of Regular Viruses. *Cold Spring Harbor Symposia on Quantitative Biology*, 27(0), 1–24. doi:10.1101/sqb.1962.027.001.005

10. Claverie, S., Bernardo, P., Kraberger, S., Hartnady, P., Lefeuvre, P., Lett, J.-M., ... Roumagnac, P. (2018). *From Spatial Metagenomics to Molecular Characterization of Plant Viruses: A Geminivirus Case Study*. *Advances in Virus Research*, 55–83. doi:10.1016/bs.aivir.2018.02.003
11. Claverie, S.; Bernardo, P.; Kraberger, S.; Hartnady, P.; Lefeuvre, P.; Lett, J.-M.; Galzi, S.; Filloux, D.; Harkins, G.W.; Varsani, A.; et al. From spatial metagenomics to molecular characterization of plant viruses: A geminivirus case study. *Adv. Virus Res.* 2018, in press
12. Dong, X. F., Natarajan, P., Tihova, M., Johnson, J. E., e Schneemann, A. (1998). Particle polymorphism caused by deletion of a peptide molecular switch in a quasi-equivalent virus. *J. Virol.* 72, 6024–6033.
13. Duffy, S., & Holmes, E. C. (2007). *Phylogenetic Evidence for Rapid Rates of Molecular Evolution in the Single-Stranded DNA Begomovirus Tomato Yellow Leaf Curl Virus*. *Journal of Virology*, 82(2), 957–965. doi:10.1128/jvi.01929-07
14. Finn, R. D., Coghill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Simon C. Potter, Punta ., Qureshi M., Sangrador-Vegas A., Gustavo A. Salazar, Tate J., e Bateman, A. (2015). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, 44(D1), D279–D285. doi:10.1093/nar/gkv1344
15. Fontenele, R., Abreu, R., Lamas, N., Alves-Freitas, D., Vidal, A., Poppiel, R., ... Ribeiro, S. (2018). *Passion Fruit Chlorotic Mottle Virus: Molecular Characterization of a New Divergent Geminivirus in Brazil*. *Viruses*, 10(4), 169. doi:10.3390/v10040169
16. Fontenele, R. S., Salywon, A. M., Majure, L. C., Cobb, I. N., Bhaskara, A., Avalos-Calleros, J. A., Gerardo R., Argüello-Astorga., Schmidlin K., Khalifeh A., Smith K., Schreck J., Michael C. Lund, Köhler M., Martin F. Wojciechowski, Wendy C. Hodgson, Puente-Martinez R., Van Doors K., Kumari S., Vernière C., Filloux F., Roumagnac P., Lefeuvre P., Ribeiro, S. G., Kraberger S., Martin, D.P e Varsani, A. (2020). A Novel Divergent Geminivirus Identified in Asymptomatic New World Cactaceae Plants. *Viruses*, 12(4), 398. doi:10.3390/v12040398
17. Frischmuth T, Ringel M e Kocher C. The size of encapsidated single-stranded DNA determines the multiplicity of African cassava mosaic virus particles. *J Gen Virol.* 2001, 82 (Pt 3): 673-676.
18. Garmann, R. F., Comas-Garcia, M., Knobler, C. M., & Gelbart, W. M. (2015). Physical Principles in the Self-Assembly of a Simple Spherical Virus. *Accounts of Chemical Research*, 49(1), 48–55.
19. Garmann, R. F., Comas-Garcia, M., Knobler, C. M. & Gelbart, W. M. Physical Principles in the Self-Assembly of a Simple Spherical Virus. *Acc. Chem. Res.* 49, 48–55 (2016).
20. Garmann RF, Comas-Garcia M, Gopal A, Knobler CM, & Gelbart WM (2014) The assembly pathway of an icosahedral single-stranded RNA virus depends on the strength of intersubunit attractions. *Journal of molecular biology* 426(5):1050-1060.
21. Halder, S., Nam, H.-J., Govindasamy, L., Vogel, M., Dinsart, C., Salome, N., McKenna R., e Agbandje-McKenna, M. (2013). Structural Characterization of H-1 Parvovirus: Comparison of Infectious Virions to Empty Capsids. *Journal of Virology*, 87(9), 5128–5140. doi:10.1128/jvi.0341612
22. Hanley-Bowdoin, L., Settlage, S. B., Orozco, B. M., Nagar, S., & Robertson, D. (1999). Geminiviruses: Models for Plant DNA Replication, Transcription, and Cell Cycle Regulation. *Critical Reviews in Plant Sciences*, 18(1), 71–106. doi:10.1080/07352689991309162
23. Hassani-Mehraban, A., Creutzburg, S., van Heereveld, L., & Kormelink, R. (2015). *Feasibility of Cowpea chlorotic mottle virus-like particles as scaffold for epitope presentations*. *BMC Biotechnology*, 15(1). doi:10.1186/s12896-015-0180-6

24. Hesketh, E. L., Saunders, K., Fisher, C., Potze, J., Stanley, J., Lomonosoff, G. P., & Ranson, N. A. (2018). *The 3.3 Å structure of a plant geminivirus using cryo-EM*. *Nature Communications*, 9(1). doi:10.1038/s41467-018-04793-6
25. Hilbert, B. J., Hayes, J. A., Stone, N. P., Duffy, C. M., Sankaran, B., & Kelch, B. A. (2015). Structure and mechanism of the ATPase that powers viral genome packaging. *Proceedings of the National Academy of Sciences*, 112(29), E3792–E3799. doi:10.1073/pnas.1506951112
26. Hurdiss, D. L., Morgan, E. L., Thompson, R. F., Prescott, E. L., Panou, M. M., Macdonald, A., & Ranson, N. A. (2016). *New Structural Insights into the Genome and Minor Capsid Proteins of BK Polyomavirus using Cryo-Electron Microscopy*. *Structure*, 24(4), 528–536. doi:10.1016/j.str.2016.02.008
27. Kazlauskas, D., Varsani, A., & Krupovic, M. (2018). Pervasive Chimerism in the Replication-Associated Proteins of Uncultured Single-Stranded DNA Viruses. *Viruses*, 10(4), 187. doi:10.3390/v10040187
28. Kazlauskas, D., Varsani, A., Koonin, E.V. *et al.* Multiple origins of prokaryotic and eukaryotic single-stranded DNA viruses from bacterial and archaeal plasmids. *Nat Commun* 10, 3425 (2019). <https://doi.org/10.1038/s41467-019-11433-0>
29. Khayat, R., Brunn, N., Speir, J. A., Hardham, J. M., Ankenbauer, R. G., Schneemann, A., & Johnson, J. E. (2011). *The 2.3-Angstrom Structure of Porcine Circovirus 2*. *Journal of Virology*, 85(21), 11542–11542. doi:10.1128/jvi.05863-11
30. Krishnamurthy, S. R., & Wang, D. (2017). Origins and challenges of viral dark matter. *Virus Research*, 239, 136–142. doi:10.1016/j.virusres.2017.02.002
31. Krupovic, M., Ghabrial, S. A., Jiang, D., & Varsani, A. (2016). *Genomoviridae: a new family of widespread single-stranded DNA viruses*. *Archives of Virology*, 161(9), 2633–2643. doi:10.1007/s00705-016-2943-3
32. Krupovic, M., Ravantti, J. J., & Bamford, D. H. (2009). *Geminiviruses: a tale of a plasmid becoming a virus*. *BMC Evolutionary Biology*, 9(1), 112. doi:10.1186/1471-2148-9-112
33. Krupovic, M., Ravantti, J.J. & Bamford, D.H. Geminiviruses: a tale of a plasmid becoming a virus. *BMC Evol Biol* 9, 112 (2009). <https://doi.org/10.1186/1471-2148-9-112>
34. Kunik T., Palanichelvam K., Czosnek H., Citovsky V., Gafni Y. (1998); Nuclear import of the capsid protein of tomato yellow leaf curl virus (TYLCV) in plant and insect cells. *Plant Journal*13:393–399
35. Lefkowitz, E. J., Dempsey, D. M., Hendrickson, R. C., Orton, R. J., Siddell, S. G., & Smith, D. B. (2018). Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic acids research*, 46(D1), D708–D717. <https://doi.org/10.1093/nar/gkx932>
36. Liang, P.; Navarro, B.; Zhang, Z.; Wang, H.; Lu, M.; Xiao, H.; Wu, Q.; Zhou, X.; Di Serio, F.; Li, S. Identification and characterization of a novel geminivirus with a monopartite genome infecting apple trees. *J. Gen. Virol.* 2015, 96, 2411–2420.
37. Loconsole, G.; Saldarelli, P.; Doddapaneni, H.; Savino, V.; Martelli, G.P.; Saponari, M. Identification of a single-stranded DNA virus associated with citrus chlorotic dwarf disease, a new member in the family geminiviridae. *Virology* 2012, 432, 162–172
38. Lozano, G., Trenado, H. P., Fiallo-Olivé, E., Chirinos, D., Geraud-Pouey, F., Briddon, R. W., & Navas-Castillo, J. (2016). Characterization of Non-coding DNA Satellites Associated with

- Sweepoviruses (Genus *Begomovirus*, *Geminiviridae*) – Definition of a Distinct Class of Begomovirus-Associated Satellites. *Frontiers in Microbiology*, 7. doi:10.3389/fmicb.2016.00162
39. Luque, D., Mata, C., Suzuki, N., Ghabrial, S., & Castón, J. (2018). Capsid Structure of dsRNA Fungal Viruses. *Viruses*, 10(9), 481. doi:10.3390/v10090481
 40. L Zhao, K Rosario, M Breitbart e S Duffy (2019). Eukaryotic circular Rep-Encoding Single-Stranded DNA (CRESS DNA) Viruses: ubiquitous viruses with small genomes and a diverse host range. *Advances in Virus Research* 103:71–133. <https://doi.org/10.1016/bs.aivir.2018.10.001>
 41. Martin, Darren P.; Biagini, Philippe; Lefevre, Pierre; Golden, Michael; Roumagnac, Philippe; Varsani, Arvind (2011). *Recombination in Eukaryotic Single Stranded DNA Viruses*. *Viruses*, 3(9), 1699–1738. doi:10.3390/v3091699
 42. Marshall, D., & Schneemann, A. (2001). Specific Packaging of Nodaviral RNA2 Requires the N-Terminus of the Capsid Protein. *Virology*, 285(1), 165–175. doi:10.1006/viro.2001.0951
 43. Moffat AS (1999) Plant pathology—geminiviruses emerge as a serious crop threat. *Science* 286:1835. Disponível: <https://doi.org/10.1126/science.286.5446.1835>.
 44. Monjane, A.L., Harkins, G.W., Martin, D.P., Lemey, P., Lefevre, P., Shepherd, D.N., Oluwafemi, S., Simuyandi, M., Zinga, I., Komba, E.K., Lakoutene, D.P., Mandakombo, N., Mboukoulida, J., Semballa, S., Tagne, A., Tiendrebeogo, F., Erdmann, J.B., Van Antwerpen, T., Owor, B.E., Flett, B., Ramusi, M., Windram, O.P., Syed, R., Lett, J.M., Briddon, R.W., Markham, P.G., Rybicki, E.P., Varsani, A., 2011. Reconstructing the history of maize streak virus strain a dispersa to reveal diversification hot spots and its origin in southern Africa. *J. Virol.* 85, 9623–9636.
 45. Perlmutter, J. D., Qiao, C. & Hagan, M. F. Viral genome structures are optimal for capsid assembly. *Elife*. 2, e00632 (2013).
 46. Perlmutter J.D. & Hagan M.F. Mechanisms of virus assembly. *Annual review of physical chemistr* (2015) 66:217-239.
 47. Rayaprolu, V., Moore, A., Wang, J. C.-Y., Goh, B. C., Perilla, J. R., Zlotnick, A., & Mukhopadhyay, S. (2017). Length of encapsidated cargo impacts stability and structure of in vitro assembled alphavirus core-like particles. *Journal of Physics: Condensed Matter*, 29(48), 484003. doi:10.1088/1361-648x/aa90d0
 48. Requião, R.D., Carneiro, R.L., Moreira, M.H. *et al.* Viruses with different genome types adopt a similar strategy to pack nucleic acids based on positively charged protein domains. *Sci Rep* 10, 5470 (2020). <https://doi.org/10.1038/s41598-020-62328-w>
 49. Richardson, L. J. *et al.* Genome properties in 2019: a new companion database to InterPro for the inference of complete functional attributes. *Nucleic Acids Res.* 47, D564–D572 (2018).
 50. Rojas, M. R., Macedo, M. A., Maliano, M. R., Soto-Aguilar, M., Souza, J. O., Briddon, R. W., ... Gilbertson, R. L. (2018). *World Management of Geminiviruses*. *Annual Review of Phytopathology*, 56(1), 637–677. doi:10.1146/annurev-phyto-080615-100327
 51. Roumagnac, P., Granier, M., Bernardo, P., Deshoux, M., Ferdinand, R., Galzi, S., Fernandez E., Julian C., Abtl I., Fillouxl D., Mesléard F., Varsani A., Blanc S., Darren P. Martin Peterschmitt, M. (2015). Alfalfa Leaf Curl Virus: an Aphid-Transmitted Geminivirus. *Journal of Virology*, 89(18), 9683–9688. doi:10.1128/jvi.00453-15
 52. Santos, Romanos e Wigg (2015). Norma Suely de Oliveira Santos, Maria Teresa Villela Romanos, Marcia Dutra Wigg. In: *Virologia Humana*. 3.ed. (Rio de Janeiro: Guanabara Koogan), pp. 78-91.

53. Sarah Cohen, Shelly Au, Nelly Panté (2011). How viruses access the nucleus, *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, Volume 1813, Issue 9, 2011, Pages 1634-1645, ISSN 0167-4889, <https://doi.org/10.1016/j.bbamcr.2010.12.009>.
54. Simmonds, P., Adams, M., Benkő, M. et al. Virus taxonomy in the age of metagenomics. *Nat Rev Microbiol* 15, 161–168 (2017). <https://doi.org/10.1038/nrmicro.2016.177>
55. Speir, J. A., Bothner, B., Qu, C., Willits, D. A., Young, M. J., & Johnson, J. E. (2006). *Enhanced Local Symmetry Interactions Globally Stabilize a Mutant Virus Capsid That Maintains Infectivity and Capsid Dynamics*. *Journal of Virology*, 80(7), 3582–3591. doi:10.1128/jvi.80.7.3582-3591.2006
56. Tisza M J et al., (2020). Discovery of several thousand highly diverse circular DNA viruses. *eLife* 2020;9:e51971. DOI: [10.7554/eLife.51971](https://doi.org/10.7554/eLife.51971)
57. Vaghi Medina, C.G.; Teppa, E.; Bornancini, V.A.; Flores, C.R.; Marino-Buslje, C.; Lopez Lambertini, P.M. Tomato apical leaf curl virus: A novel, monopartite geminivirus detected in tomatoes in Argentina. *Front. Microbiol.* 2017, 8, 2665.
58. Van Regenmortel MH, Ackermann HW, Calisher CH, Dietzgen RG, Horzinek MC, Keil GM, Mahy BW, Martelli GP, Murphy FA, Pringle C, Rima BK, Skern T, Vetten HJ, Weaver SC (2013). Virus species polemics: 14 senior virologists oppose a proposed change to the ICTV definition of virus species. *Arch Virol* 158:1115–1119
59. Varsani, A., Roumagnac, P., Fuchs, M., Navas-Castillo, J., Moriones, E., Idris, A., Rob W. Briddon R.W., Rivera-Bustamante R., Zerbini F.M., e Martin, D. P. (2017). *Capulavirus* and *Grablovirus*: two new genera in the family *Geminiviridae* . *Archives of Virology*, 162(6), 1819–1831. doi:10.1007/s00705- 017-3268-6
60. Varsani A., Krupovic M. (2017). Sequence-based taxonomic framework for the classification of uncultured single-stranded DNA viruses of the family *Genomoviridae*. *Virus Evol.* 2, 3(1):vew037.
61. Venter, P. A., Marshall, D. & Schneemann, A. Dual roles for an arginine-rich motif in specific genome recognition and localization of viral coat protein to RNA replication sites in flock house virus-infected cells. *J. Virol.* 83, 2872–2882 (2009).
62. Vladimir A. Belyi e M. Muthukumar (2006) Electrostatic origin of the genome packing in viruses, *PNAS* November 14, 2006 103 (46) 17174-17178; <https://doi.org/10.1073/pnas.0608311103>
63. Wolf, Y. I., Kazlauskas, D., Iranzo, J., Lucía-Sanz, A., Kuhn, J. H., Krupovic, M., Dolja V.V., e Koonin, E. V. (2018). Origins and Evolution of the Global RNA Virome. *mBio*, 9(6). doi:10.1128/mbio.02329-18
64. Xu, X., Zhang, Q., Hong, J., Li, Z., Zhang, X., & Zhou, X. (2019). Cryo-EM Structure of a Begomovirus Geminiate Particle. *International Journal of Molecular Sciences*, 20(7), 1738. doi:10.3390/ijms20071738
65. Yu, Xiao., Li, Bo., Fu, Yanping., Jiang, Daohong., Ghabrial, Said A., Li, Guoqing., Peng, Youliang., Xie, Jiatao., Cheng, Jiasen.,Huang, Junbin e Yi, Xianhong (2010). A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. *Proceedings of the National Academy of Sciences* May 2010, 107 (18) 8387-8392; DOI: 10.1073/pnas.0913535107
66. Zerbini, F.M.; Briddon, R.W.; Idris, A.; Martin, D.P.; Moriones, E.; Navas-Castillo, J.; Rivera-Bustamante, R.;Roumagnac, P.; Varsani, A.; Ictv Report, C. ICTV Virus Taxonomy Profile: *Geminiviridae*. *J. Gen. Virol.* 2017,98, 131–133.Zhang, W., Olson, N. H., Baker, T. S., Faulkner, L., Agbandje-McKenna, M., Boulton, M. I., Davies, J.W., McKenna, R. (2001). Structure of the Maize Streak Virus Geminiate Particle. *Virology*, 279(2), 471–

