UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOLOGIA CELULAR
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA MOLECULAR

Dinâmica e evolução viral em diferentes escalas: da emergência à interação com o hospedeiro

JOÃO MARCOS FAGUNDES SILVA

Orientador: Dr. Tatsuya Nagata
Coorientador: Fernando Lucas Melo

Brasília, 2022

UNIVERSIDADE DE BRASÍLIA
INSTITUTO DE CIÊNCIAS BIOLÓGICAS
DEPARTAMENTO DE BIOLOGIA CELULAR
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA MOLECULAR

Dinâmica e evolução viral em diferentes escalas: da emergência à interação com o hospedeiro

JOÃO MARCOS FAGUNDES SILVA

Orientador: Dr. Tatsuya Nagata
Coorientador: Fernando Lucas Melo

Tese apresentada ao Programa de Pós-Graduação em Ciências Biológicas – Biologia Molecular, do Departamento de Biologia Celular, do Instituto de Ciências Biológicas da Universidade de Brasília como parte dos requisitos para obtenção do título de Doutor em Biologia Molecular.

Brasília, 2022

JOÃO MARCOS FAGUNDES SILVA

Dinâmica e evolução viral em diferentes escalas: da emergência à interação com o hospedeiro

Tese apresentada ao Programa de Pós-Graduação em Ciências Biológicas – Biologia Molecular, do Departamento de Biologia Celular, do Instituto de Ciências Biológicas da Universidade de Brasília como parte dos requisitos para obtenção do título de Doutor em Biologia Molecular.

Banca examinadora:

Tatsuya Nagata (Orientador) (CEL - UnB)

Ricardo Henrique Kruger (CEL - UnB)

Rosana Blawid (Fitossanidade - UFRPE)

Alison Talis Martins Lima (Instituto de Ciências Agrárias - UFU)

Suplente:

Erich Yukio Tempel Nakasu (Lab. Virologia e Biologia Molecular - Embrapa Hortaliças)

Agradecimentos

Os diversos trabalhos apresentados nesta tesa não poderiam ter sido realizados sem a ajuda e o suporte dos meus orientadores, colegas, amigos e familiares. Em especial, agradeço pela orientação do dr. Tatsuya Nagata, pela coorientação do dr. Fernando Lucas Melo e supervisão do dr. Santiago Elena durante minha estadia em seu laboratório, e pelo apoio e suporte dos meus pais Tacito Furtado Silva e Irene Fagundes Silva. Tive o privilégio de trabalhar com pessoas incríveis durante o meu doutorado, e agradeço pelo suporte dos grupos de pesquisa dos drs. Tatsuya Nagata, Santiago Elena, Bergmann Morais Ribeiro e Renato Oliveira Resende. Sou grato aos grupos de pesquisa dos drs(as). Jorge Alberto Marques Rezende, Bergmann Morais Ribeiro, Renato Oliveira Resende, Thor Vinícius Martins Fajardo, Fabrício Souza Campos, Alice Kazuko Inoue-Nagata e Ricardo Henrique Kruger pelas várias oportunidades de colaborações. Agradeço também pelo apoio da UnB, do Programa de Pós-Graduação em Biologia Molecular, da Universidade de Valência e dos órgãos de fomento à pesquisa, CNPq e CAPES e pelo aceite do convite da banca examinadora em sua participação.

Índice

Resumo

Vírus de RNA evoluem em escalas temporais curtas, os fazendo capazes de se adaptar a condições adversas rapidamente. De uma perspectiva temporal mais ampla, o histórico evolutivo dos vírus de RNA é marcado por uma origem primordial e extensa transferência horizontal de genes. Essas características são responsáveis pela vasta diversidade encontrada nesse grupo que permanece ainda pouco explorada. A integração do conhecimento de diversas áreas da ciência é necessária para melhor entendermos e melhor avaliarmos processos evolutivos importantes como mudança de hospedeiro, emergência de zoonoses e evasão da resposta imune; assim como para unificar diversidade viral e evolução viral em micro e macro escalas temporais. As pressões seletivas impostas pelo hospedeiro durante a replicação viral e eventos de transferência horizontal de genes são eventos intracelulares de alta relevância para a evolução e diversidade dos vírus de RNA. Nesta tese, estudamos a resposta celular de células do trato gastrointestinal de *Drosophila melanogaster* à infecção com dois vírus, thika virus (TV) e *Drosophila melanogaster* Nora virus (DMelNV), onde mostramos que esses vírus usam estratégias distintas para se replicar, e devido à resposta celular à infecção ser dependente de vírus e tipo celular, esses vírus estão sujeitos a pressões seletivas dependente de tipo celular; caracterizamos eventos de rearranjo entre dois tospovírus, tomato chlorotic spot virus (TCSV) e groundnut ringspot virus (GRSV), e mostramos que o segmento M do TCSV foi extinto ou permanece não sequenciado; e propomos novos critérios taxonômicos para a família *Betaflexiviridae* com base em análises evolutivas, onde demostramos a influência que eventos de recombinação exercem sobre a evolução e diversidade da família.

Abstract

RNA viruses are fast-evolving entities able to adapt quickly to challenging environments. Their long scale evolutionary history is marked by a primordial origin and extensive horizontal gene transfer. These characteristics are responsible for the vast diversity found in this group which still remains underexplored. The integration of knowledge from various areas of science is necessary to better understand and evaluate important evolutionary processes such as jumps to different hosts, zoonosis emergence and immune evasion; but also to unify viral diversity and evolution at small and large time scales. In this thesis, a series of bioinformatic analyses are conducted to study the evolution and diversity of RNA viruses, as well as to study virus-host interactions given their influence on the evolution of both pathogen and host. Selective pressures acting on viruses during viral replication and horizontal gene transfer events are intracellular phenomena that are extremely relevant to the evolution and diversity of RNA viruses. In this thesis, we study the cellular response of *Drosophila melanogaster* midgut cells to infection of two viruses, thika virus (TV) and *Drosophila melanogaster* Nora virus (DMelNV), where we show that these viruses employ different replication strategies, and given that cellular response was dependent on cell type and specific to each virus, these viruses are subjected to different selective pressures that is influenced by cell type; we characterize reassortment events between two tospoviruses, tomato chlorotic spot virus (TCSV) e groundnut ringspot virus (GRSV), and show that the M segment of TCSV was extinct or has not been sequenced yet; and we propose novel taxonomic criteria for the family *Betaflexiviridae* based on evolutionary analyses, where we show the impact of recombination on the evolution and diversity of this family.

Capítulo1. Introdução

Por que estudar evolução? O quão previsíveis são os processos evolutivos? A predição da trajetória evolutiva de um organismo é uma das grandes metas da biologia evolutiva, entretanto, a nossa capacidade de tanto analisar eventos passados quanto predizer processos evolutivos depende da assimilação do conhecimento de diversas áreas da biologia (Reznick & Travis, 2018). Nessa tese, apresentamos análises evolutivas, descrição de novos vírus e estudos sobre aspectos biológicos relevantes para a evolução de vírus com genoma de RNA sem intermediário de DNA (referidos aqui somente por vírus de RNA). Primeiro, introduziremos o leitor aos diversos tópicos abordados, enfatizando a sua relevância para a evolução dos vírus de RNA. Nos Capítulos 2, 3 e 4, escritos em inglês e em formato de artigo, iremos abordar, respectivamente, interação vírus-hospedeiro a nível celular em células do trato gastrointestinal da mosca da fruta; evolução de tospovírus; e evolução e taxonomia da família *Betaflexiviridae*. No Capítulo 5 serão listadas as publicações obtidas por meio de colaborações. Por último, os resultados obtidos serão compilados e discutiremos o impacto dos mesmos no cenário evolutivo a curto e longo prazo dos vírus de RNA.

1. Evolução de vírus de RNA

A exata origem dos vírus de RNA permanece elusiva. Considerando que moléculas de RNA são capazes de armazenar informação e catalisar reações químicas, acredita-se que estejam envolvidas nos primeiros estágios da evolução da vida (Gilbert, 1986). Possivelmente, vírus de RNA são resquícios do primevo Mundo de RNA, os fazendo mais antigos que o Último Ancestral Comum Universal (*Last Universal Common Ancestral*; LUCA) (Wolf et al., 2018). A origem ancestral dos vírus de RNA é corroborada pela ausência de homólogos da enzima viral polimerase dependente de RNA (*RNA-dependent RNA polymerase*; RdRp) em organismos celulares (Koonin et al., 2006) e aquisição de uma proteína de capsídeo *single jellyroll* (SJR) nos estágios iniciais da evolução celular (Krupovic & Koonin, 2017). Independentemente da exata emergência dos vírus de RNA, a origem primordial dessas entidades é refletida na sua vasta diversidade.

Devido à falta de atividade de correção, a RdRp dos vírus de RNA é caracterizada por uma alta taxa de erro (entre 10^{-3} e 10^{-5} erros por sítio) dando aos vírus de RNA uma alta capacidade de adaptação (Barr & Fearns, 2010). Consequentemente, vírus de RNA acumulam mutações rapidamente, possibilitando acompanhar sua evolução a curto prazo (Biek et al., 2015). Em contrapartida, a alta divergência a nível genômico entre vírus de RNA distantemente relacionados dificulta a reconstrução do histórico evolutivo completo desses vírus. Somente devido à alta quantidade de novos genomas sequenciados recentemente, especialmente em organismos não-modelos (Shi et al., 2016; Shi et al., 2018), foi possível a reconstrução da filogenia global dos vírus de RNA (Wolf et al., 2018), e ainda assim, esses resultados devem ser interpretados com cautela (Holmes & Duchêne, 2019). Outra característica importante da RdRp viral é a atividade de troca de fita molde (*template switch*; TS), possibilitando que vírus de RNA recombinem (Simon-Loriere & Holmes, 2011). Rearranjos de segmentos genômicos em vírus segmentados e eventos de recombinação são responsáveis pela extensa transferência horizontal de genes em vírus de RNA (Zhang et al., 2019). A integração de análises evolutivas em curtos e

longos períodos de tempo e um melhor conhecimento da diversidade viral são necessários para melhor entendermos os processos evolutivos que levam a emergência de novos vírus, troca de hospedeiro e escape da resposta imune (Sanjuán et al., 2021).

2. Metagenômica e a virosfera

Estima-se que existam 10^{31} partículas virais em ambientes marinhos, fazendo dos vírus as entidades biológicas mais abundantes, e provavelmente as mais diversas (Hendrix et al., 1999). Estudos virológicos tem sido, até recentemente, enviesados para vírus humanos ou de interesse pecuário e agrícola, e consequentemente, a nossa compreensão da extensão da virosfera até então extremamente limitada. As tecnologias de sequenciamento em larga escala (*High-Throughput Sequencing*; HTS) nos permitiram caracterizar um alto número de genomas de vírus de RNA muitas vezes distantemente relacionados aos genomas até então descritos, revelando um cenário mais amplo da diversidade e evolução dessas entidades (Shi et al., 2016; Shi et al., 2018; Wolf et al., 2018; Zhang et al., 2019). A reconstrução global da filogenia dos vírus de RNA, possível somente devido ao sequenciamento de genomas virais divergentes, salienta a importância de se estudar o viroma de organismos não-modelos para melhor compreender o histórico evolutivo global desses vírus. Em adição, a caracterização do viroma de animais silvestres, combinada a ensaios experimentais e novas ferramentas computacionais, é essencial para melhor predizermos a emergência de zoonoses no futuro (Sanjuán et al., 2021).

2.1. Novos desafios para a taxonomia viral

Ao contrário dos organismos celulares, vírus não possuem um único ancestral comum (Koonin et al 2006; Krupovic & Koonin, 2017, Koonin et al., 2020), fato que dificulta a utilização do histórico evolutivo de genes conservados para a demarcação taxonômica viral (Koonin et al., 2020). Consequentemente, os critérios para demarcação taxonômica de vírus são dinâmicos e específicos para cada grupo particular de vírus, tendo em vista características biológicas e evolutivas. O Comitê Internacional de Taxonomia de Vírus (*International Committee on Taxonomy of Viruses*; ICTV) é a entidade responsável pela classificação taxonômica dos vírus, onde grupos de estudo são montados para avaliar a taxonomia vigente e recomendar atualizações.

O novo cenário da diversidade viral proporcionado pelo sequenciamento massivo de novos vírus ocasionou em uma revisão da taxonomia viral e a proposta de uma megataxonomia baseada em informação genômica (Koonin et al., 2020). Apesar das relações evolutivas globais entre os vírus de RNA terem sido elucidadas com a ajuda da informação genômica de novos vírus, características biológicas importantes de vírus descritos somente pelo seu genoma permanecem desconhecidas. Esse fato é particularmente problemático nos casos onde as características biológicas dos vírus, como gama de hospedeiro e sintomatologia, são utilizadas para demarcação taxonômica (Simmonds et al., 2017). Apesar de ser uma construção artificial, a taxonomia viral, em especial a demarcação de espécies, é de extrema importância não somente para comunidade científica, mas também para o público geral e órgãos reguladores que monitoram a importação de vírus quarentenários (Hull & Rima, 2020).

3. Sequenciamento de RNA de células únicas (single cell RNA-sequencing; scRNA-seq) aplicado à virologia

A heterogeneidade celular em amostras subjugadas a sequenciamento de RNA (*RNA-sequencing*; RNA-seq) é mascarada em ensaios convencionais. Com as tecnologias de sequenciamento de RNA de células únicas (*single cell RNA-sequencing*; scRNA-seq) é possível recuperar a perfil de expressão de RNA de células individuais em paralelo. Contudo, devido à baixa eficiência na captura de mRNA, dados gerados por scRNA-seq são ruidosos, o que levou ao desenvolvimento de diversas ferramentas e algoritmos focados nas características desses dados (matrizes de expressão esparsas, alta variabilidade biológica e técnica, etc.) (Chen et al., 2019).

Tanto a resposta e a susceptibilidade celular a infecções virais são heterogêneas e dependente de tipo celular (Cristinelli & Ciuffi, 2018). Utilizando técnicas com resolução celular como o scRNA-seq, é possível capturar a expressão de determinantes de tropismo celular (Xu et al., 2020; Qi et al., 2021; Zou et al., 2020; Ravindra et al., 2021; Chua et al., 2020), a resposta celular à infecção dependente de tipo celular e resposta de células expostas mas não-infectadas (células *bystanders*) (Ren et al., 2021; Steuerman et al., 2018; Kotliar et al., 2020; Ravindra et al., 2021; Chua et al., 2020), acumulo de RNA viral (Zanini et al., 2018; Steuerman et al., 2018; Kotliar et al., 2020; Ravindra et al., 2021; Chua et al., 2020), mudanças nas proporções de diferentes tipos celulares e rotas de diferenciação celular em resposta à infecção (Ren et al., 2021; Kotliar et al., 2020; Chua et al., 2020), interação entre diferentes tipos celulares (Chua et al., 2020) e trajetórias de progressão de células infectadas (Zanini et al., 2018; Hein & Weissman, 2021). Essas possibilidades fazem com que scRNA-seq seja uma ferramenta poderosa para estudos de interações vírus-hospedeiro (Cristinelli & Ciuffi, 2018). Interações entre patógenos e hospedeiro são umas das principais forças que moldam a evolução e a diversidade genética de ambos (Sironi et al., 2015). Nesse contexto, scRNA-seq e tecnologias relacionadas são capazes de oferecer informações valiosas para se estudar evolução e coevolução de patógenos e hospedeiro.

4. Objetivos gerais

Temos como objetivo geral realizar diversas análises bioinformáticas abrangendo múltiplas áreas da virologia, com o intuito de estudar de vários ângulos aspectos importantes para a evolução dos vírus de RNA, como eventos de rearranjo e recombinação, diversidade viral, interação vírus-hospedeiro e tropismo celular. Através da integração das análises, pretendemos discutir a importância dos resultados obtidos para a compreensão da evolução global dos vírus de RNA, assim como estudar aspectos biológicos de vírus específicos que influenciem a evolução dos mesmos.

5. Objetivos específicos

- Estudar a resposta celular genérica em células infectadas com os vírus thika virus (TV) e *Drosophila melanogaster* nora virus (DMelNV) em células do trato gastrointestinal da mosca da fruta.

- Estudar a resposta celular dependente de tipo celular a infecções com TV e DMelNV em células do trato gastrointestinal da mosca da fruta.
- Determinar a susceptibilidade de diferentes tipos e subtipos celulares de células do trato gastrointestinal da mosca da fruta à infecção com TV e DMelNV, avaliando também o acumulo do RNA genômico viral.
- Comparar características da resposta celular e resposta sistêmica à infecção com DMelNV sob uma perspectiva de biologia de sistemas.
- Determinar eventos de rearranjos entre os tospovírus tomato chlorotic spot virus (TCSV) e groundnut ringspot virus (GRSV).
- Estimar a partir do uso de relógio molecular o período de ocorrência de rearranjos entre TCSV e GRSV.
- A partir de análises genômicas e filogenéticas dos vírus pertencentes à família *Betaflexiviridae*, propor para a família critérios taxonômicos focados em informação genômica.

Capítulo 2. Heterogeneity in the response of different subtypes of *Drosophila melanogaster* midgut cells to viral infections

Este Capítulo foi publicado na revista *Viruses* (ISSN: 1999-4915).

Silva, J.M.F.; Nagata, T.; Melo, F.L.; Elena, S.F. Heterogeneity in the Response of Different Subtypes of Drosophila melanogaster Midgut Cells to Viral Infections. *Viruses* **2021**, 13, 2284.

A versão presente nesta tese apresenta mudanças na formatação e pequenas alterações.

Abstract: Single-cell RNA sequencing (scRNA-seq) offers the possibility to monitor both host and pathogens transcriptomes at the cellular level. Here, public scRNA-seq datasets from Drosophila melanogaster midgut cells were used to compare the differences in replication strategy and cellular responses between two fly picorna-like viruses, Thika virus (TV) and D. melanogaster Nora virus (DMelNV). TV exhibited lower levels of viral RNA accumulation but infected a higher number of cells compared to DMelNV. In both cases, viral RNA accumulation varied according to cell subtype. The cellular heat shock response to TV and DMelNV infection was cell-subtype- and virus-specific. Disruption of bottleneck genes at later stages of infection in the systemic response, as well as of translation-related genes in the cellular response to DMelNV in two cell subtypes, may affect the virus replication.

Keywords: cell-type-specific gene expression; Drosophila viruses; dual RNA-seq; single-cell genomics; single-cell RNA-seq; virus-host interaction; antiviral heat shock response

1. Introduction

Multi-cellular organisms respond to viral infections at both cellular and systemic levels. How different cell types and how infected and uninfected bystander cells respond to viral infections are questions that are being recently addressed using single-cell RNA sequencing (scRNA-seq) and other single-cell techniques. As an example, single-cell profiling of Ebola virus (EBOV)-infected immune cells from rhesus macaques revealed that interferon-stimulated genes (ISGs) are down-regulated in infected cells compared to bystanders (Kotliar et al., 2020), shedding light into previous seemingly contradictory results from studies of EBOV infection in culture and in vivo, where ISGs and downstream signaling genes were, respectively, down- and up-regulated compared to their healthy counterparts (Basler et al., 2006; Gupta et al., 2001; Harcourt et al., 1999; Kash et al., 2006; Caballero et al., 2016; Liu et al., 2017). Similarly, studies using scRNA-seq found that bystander cells from mice infected with influenza virus A (IAV), and bystander cells from patients positive for severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) show an over-expression of ISGs compared to cells from healthy individuals (Steuerman et al., 2018; Ravindra et al., 2021), stressing the importance of a systemic response to these respiratory viruses. Yet, this powerful technique has been largely applied to viral infections in mammalian cells, and studies in non-mammalian hosts are currently missing.

The fruit fly (*Drosophila melanogaster* Meigen) is an attractive invertebrate model for studying virus-host interactions (Huszar & Imler, 2008) in which RNA interference (RNAi)

plays a major antiviral role (Kemp & Imler, 2009; van Mierlo et al., 2012). Heat shock response to viral infections has also been shown to contribute to antiviral defense by restricting viral replication (Merkling et al., 2015). Overexpression of the heat shock factor (Hsf) and Hsp70 induce resistance to viral infections in transgenic flies. In particular, overexpression of Hsf diminished viral loads to undetectable levels in some instances, suggesting that these transformed flies were cleared from viral infections (Merkling et al., 2015).

Novel sequencing technologies are allowing the discovery of novel RNA viruses in both wild and stock flies (Webster et al., 2015b) that may serve as models for studying virus-host interactions. However, the biology of only a small subset of these viruses has been investigated in-depth, such as *D. melanogaster* sigma virus (DMelSV; genus *Sigmavirus*, family *Rhabdoviridae*), Drosophila C virus (DCV; genus *Cripavirus*, family *Dicistroviridae*), and, more recently, *D. melanogaster* Nora virus (DMelNV; unclassified picorna-like virus), an enteric virus that is known to cause persistent infections on both wild and stock flies with no obvious pathological outcome (Kemp & Imler, 2009; Webster et al., 2015b; Habayeb et al., 2009; Habayeb et al., 2006).

The gastrointestinal tract of the fruit fly is composed of diverse cell types that perform essential tasks, such as nutrient uptake and secretion of neuropeptide hormones. Enterocytes (ECs) are responsible for the secretion of digestive enzymes and nutrient uptake. These cells possess specialized gene expression profiles depending on regionalization and can be divided into ECs from the anterior (aEC), middle (mEC), and posterior (pEC) portions of the fly's midgut as well as into a few subtypes (Hung et al., 2020). The middle region, also known as gastric region, of the fly's midgut, resembles the mammalian stomach due to acidification by copper cells (Dubreuil, 2004). Enteroendocrine cells (EEs) secrete a variety of neuropeptide hormones that play an important role in controlling many physiological processes. They are scattered throughout the gastrointestinal tract and produce more than 20 neuropeptide hormones, including allatostatines A, B, and C (AstA, AstB, and AstC, respectively); tachykinin (Tk); neuropeptide F (NPF); diuretic hormone 31 (DH31); and CCHamide-1 and -2 (CCHa1 and CCHa2, respectively) (Rehfeld, 2013; Furness et al., 2013; Gribble & Reimann, 2019; Beehler-Evans & Micchelli, 2015). Recently, ten EE subtypes were identified in the fruit fly. These subtypes can be divided into two major classes: class I is composed of cells expressing AstC, and class II is composed of cells that express Tk (Beehler-Evans & Micchelli, 2015; Guo et al., 2019). In addition to these classes, a subtype dubbed III that does not express neither AstC nor Tk has been also identified. Further classification of EE subtypes is based on whether they are located in the anterior, medium (gastric region), or posterior regions of the gastrointestinal tract (-a, -m, and -p suffixes; Figure 1a) and on their gene expression (Guo et al., 2019).

Here, publicly available Drosophila scRNA-seq datasets from the midgut epithelium cells were used to investigate the infection dynamics of two viruses, DMelNV and the recently discovered Thika virus (TV; unclassified picorna-like virus). By analyzing both viral RNA accumulation and host transcription levels in a manner that is analogous to dual RNA-seq (Wesolowska-Andersen et al., 2017), we show that in the presence of multiple infections, it is possible to use scRNA-seq to analyze the transcription of the host and multiple pathogens

simultaneously. Not only we have been able to monitor both virus replication and host transcriptional response to infection, we have been also able to simultaneously compare the replication strategies of two viruses and some of the cellular responses to these viruses in vivo, revealing more of the uniqueness and similarities of viral infections at the single-cell level and providing possible new models for invertebrate viruses.

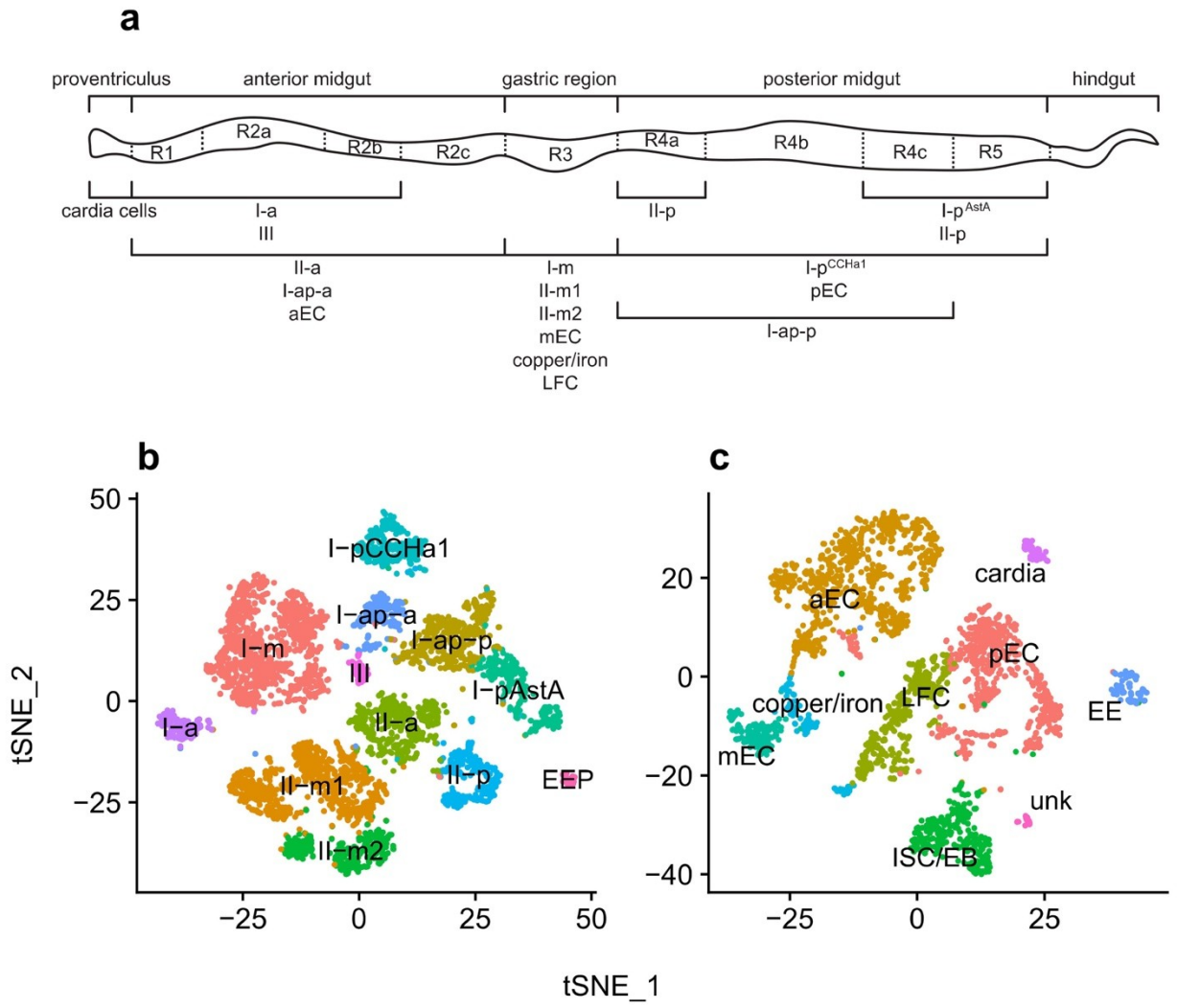


Figure 1. Spatial distribution of EE cells and of viral infections. (**a**) Schematic representation of the fly gastrointestinal tract showing the spatial distribution of EEs (adapted from Guo et al., 2019; published under creative commons license). (**b**) t-SNE (t-distributed stochastic neighbor embedding) reduction plots of EEs with identified clusters. (**c**) t-SNE reduction plots of cells from the entire fly midgut with identified clusters. In both t-SNE plots, each point represents a single cell, and colors indicate labeled cell types.

2. Materials and methods

2.1. Data collection

Raw sequencing data were downloaded from NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra last accessed 15 November 2021; BioProject accession: PRJNA547484). Briefly, these data were generated by following 10X Genomics GemCode protocol (Zheng et al., 2017) using ~8000 EEs harvested from the midgut of ~200 female fruit flies (CG32547-GAL4 > GFP line) aged between five and seven days (Guo et al., 2019). GFP was used to sort cells with a FACS Aria III sorter (BD Biosciences,) and sequencing was performed at an Illumina X10 platform. Additionally, a fruit fly scRNA-seq dataset from the entire midgut, hereafter referred to as midgut atlas (MA) dataset, was also obtained through SRA (BioProject accession: PRJNA493298). Only data generated by 10X technology were analyzed. Briefly, guts from seven days old female flies (esg-sfGFP/+, pros-GAL4 > RFP/+ line) were dissected, and libraries were prepared following 10X Genomics GemCode protocol (Hung et al., 2020). Two technical replicates for two samples were prepared for this data, resulting in a total of four libraries.

2.2. Identification of viruses in scRNA-seq datasets

For each dataset, the FASTQ files corresponding to the transcripts were trimmed with BBDuk v38.87 (BBMap package; sourceforge.net/projects/bbmap/; last accessed 15 November 2021), with parameters ktrim = r ref = adapters k = 21 qtrim = r trimq = 10. Trimmed reads were concatenated and aligned to the *D. melanogaster* reference genome (accession GCA_000001215.4) with BWA v0.7.15 (Li, 2013) using default parameters. Next, mapped reads were filtered out with samtools v1.12 (Li et al., 2009). The remaining reads were assembled with MEGAHIT v1.2.9 (Li et al., 2016), and the resulting contigs were queried against the nr database (download in June 2021) with DIAMOND BLASTX (Buchfink et al., 2021). False positives were determined based on further BLAST searches. Three picorna-like viruses were found in the EE dataset: DCV, DMelNV, and TV; and two viruses were found in the MA dataset: DMelNV and Drosophila A virus (DAV).

2.3. Filtering and cluster generation

Barcode processing and gene counts were performed with CellRanger v3.1.0 (10X Genomics, CA, USA) using an edited *D. melanogaster* reference genome that included the viral sequences from DCV (accession AF014388), DMelNV (accession GQ257737), and TV (accession KP714072). Downstream analysis was performed with Seurat v3.1.4 (Stuart et al., 2019). A total of 4994 cells containing between 200 and 3000 detected genes and < 5% mitochondrial genes were retained. Scaling and normalization were performed with the SCTransform function (Hafemeister & Satija, 2019), with mitochondrial genes and virus counts regressed out. Principal component analysis (PCA) and t-distributed stochastic neighbor embedding (t-SNE) reductions were performed with top 20 PCs, and clusters were generated at a 0.4 resolution. Clusters were identified using marker genes (Guo et al., 2019), as shown in Supplementary Figure S1a.

The same analysis was performed for the MA dataset, albeit with some minor differences. The DAV genome (accession FJ150422) was added to the GTF and FASTA files; however, no counts from this virus were found in cells after quality filtering. Possibly, the lack

of a poly(A) tail hampered the accurate detection of this virus in this dataset. CellRanger v3.1.0 was run separately for each library. Technical replicates counts were merged, and the two samples were integrated using SCTransform normalized counts with FindIntegrationAnchors and IntegrateData functions in Seurat v3.1.4. After filtering, 2375 cells containing between 200 and 3000 detected genes and < 25% mitochondrial genes remained. All integration steps, PCA, and t-SNE reduction were performed with top 50 PCs, and clusters were generated at a 1.0 resolution. Clusters were identified with a set of marker genes (Hung et al., 2020), as shown in Supplementary Figure S1b.

After identification of cell types in these datasets, ambient RNA contamination was removed with CellBender (Fleming et al., 2019). Ambient RNA contamination is ubiquitous in droplet-based scRNA-seq protocols, and filtering off background RNA contamination can ameliorate downstream analysis, such as differential expression (Fleming et al., 2019). This analysis was performed on each library independently. CellBender was run on the raw output count matrices produced by CellRanger in which viral counts were removed.

2.4. Determination of infected cells

Cells were determined to be infected based on the estimated fraction of ambient RNA in each cell and the probability of finding viral unique molecular identifiers (UMI) in empty droplets. The profile of ambient RNA estimated from cell-free empty droplets is a compelling approach to call infected cells since it accounts for the possible contamination of viral particles in cell-containing droplets (Kotliar et al., 2020). First, for each cell i, we estimated the fraction of ambient RNA, $A(i)$, by dividing the number of UMIs filtered out with CellBender by the total number of non-viral UMIs before the removal of background RNA. Next, the probability of finding viral UMIs in ambient RNA, $V(a)$, was calculated from the proportion of viral UMIs in empty droplets. Cell barcodes not called by CellRanger containing <100 non-viral UMIs were considered empty droplets. Then, for each cell i, the probability of infection $P(i)$ was calculated based on the following Binomial survival, or reliability, function with N trials and probability of success p:

$$P(i) = S[V(i) \mid p, N] = 1 - F[V(i) \mid p = A(i)V(a), N = UMI(i)], \quad (1)$$

where $F[\cdot]$ is the Binomial CDF. $V(i)$ and $UMI(i)$ are, respectively, the number of viral and total UMIs in cell i, and p is the probability of finding $V(a)$ viral UMIs derived from ambient RNA contamination $A(i)$. Cells with $P(i) < 0.01$ were determined to be infected. In addition, a second criterion in which infected cells need to have at least two viral UMIs was applied. Notably, no TV-derived reads were found in the ambient RNA, and cells were determined to be infected based solely on the second criterion.

2.5. Virus Infection/Replication Analyses

The proportion of viral counts in each cell was multiplied by a factor of 10,000 and log(pseudocount + 1)-transformed. One-way ANOVA tests were performed to investigate whether the accumulation of DMelNV and TV are influenced by cell subtype using R v4.0.3. Significant results were further investigated by performing Tukey–Kramer post-hoc tests using

the agricolae R package v1.3-2 (https://CRAN.R-project.org/package=agricolae; last accessed 15 November 2021) in R v4.0.3.

2.6. Gene expression analyses

Differential gene expression analysis was performed with glmGamPoi (Ahlmann-Eltze & Huber, 2020). A oneway layout with one level for each cell subtype/infection status was used. An intercept term was not included in this analysis. Only uninfected cells with no viral UMIs were included as uninfected cells, and the unk cell type from the MA dataset was not included in this analysis. Generic cellular response differentially expressed genes (DEGs) were identified by testing for the effect of infection in all cell subtypes. The generic response to DMelNV was tested for the EE and MA datasets separately. Cell-subtype-specific DEGs were identified by testing for the effect of the interaction between infection and cell subtype for each subtype separately. For TV, the response to cells located in the posterior region of the midgut was also tested with the corresponding contrast. For each virus, cell-subtypespecific response DEGs were pulled together in lists of up- and down-regulated genes. To find genes which expression correlates to viral accumulation, we ran glmGamPoi with expression matrices containing only infected cells using cell subtype as a factor and the percentage of viral UMIs as a covariate, and then, tests were conducted for DEGs by setting only the percentage of viral UMIs as a contrast. For every list of up- or down-regulated DEGs composed by at least three genes, pathway enrichment analysis was conducted with ReactomePA (Yu et al., 2016).

2.7. Gene regulatory network activity analysis

SCENIC (Aibar et al., 2017) R package v1.2.4 was used to infer regulon (a regulatory network composed of a transcription factor and its target genes) activity on both EE and MA datasets separately. CellBender-corrected counts were used as input to this analysis. Genes expressed in at least 1% of the cells and having at least 3 × total number of cells × 0.01 (which corresponds to the amount of UMI counts a gene would have if it had 3 counts in 1% of the cells) were retained. Network inference based on co-expression was performed with GENIE3 (Huynh-Thu et al., 2010), and the area under the curve (AUC) activity values for each regulon was then obtained with AUCell. Differential regulon activity of the AUC values was conducted via the Seurat FindAllMarkers function with the default Mann–Whitney U-test (Appendix File S3).

2.8. Gene network analyses

A high quality predicted interactome of *D. melanogaster* was downloaded from http://drosophila.biomedtzc.cn v2018_01 (last accessed 15 November 2021) (Ding et al., 2020). Gene interaction networks were analyzed as undirected graphs with the igraph R package v1.2.5 (Csardi & Nepusz, 2006). The degree probability distribution and the betweenness centrality of each node was computed for the interactome network. Then, a linear regression on the degree probability distribution in the log-log space was computed to obtain the critical exponent, γ, of the power-law fit. For each list of DEGs containing more than three genes, the corresponding subnetwork was extracted from the complete interactome, and its critical exponent was obtained as described above. Next, we performed *t*-tests between the critical exponent of the full interactome and each subnetwork to test for significant differences in the degree distribution

between the subnetwork and complete interactome. The betweenness centrality of DEGs were computed using the complete network. One-tailed Man–Whitney U-tests were then performed to compare the betweenness centrality of the DEGs with that obtained from all genes in the network, considering the upper tail of the distribution. All *p*-values were adjusted by the Benjamini–Hochberg method. Articulation points were determined with the articulation_points function.

2.9. DMelNV infection bulk RNA-seq data analysis

A list of DEGs from DMelNV-infected female flies was obtained from Lopez et al., 2018. This data contains DEGs detected at 2, 10, 20, and 30 days post-eclosion. Briefly, these data were generated by stablishing DMelNV-infected white-eyed flies (w^{1118}; Vienna Drosophila Resource Center, Vienna, Austria) stocks via fecal-oral infection. RNA extraction was performed with TRIzol reagent (ThermoFisher Scientific, Waltham, MA, USA) on triplicates for each time point, and samples were sequenced at an Illumina HiSeq system platform. Fragments per kilobase of transcript per million mapped reads (FPKM) values were used to determine DEGs. Gene network analysis was performed for up- and down-regulated genes as described above for each time point separately.

2.10. Analysis of the effects of infection on the I-p^{AstA} subtype

Cluster generation analysis of the I-p^{AstA} subtype was done as described previously with some minor differences. First, normalization was performed on CellBender-generated counts based on cell size with the NormalizeData function. The top 20 PCs were used for cluster generation and t-SNE reductions, and clusters were generated at a 0.5 resolution.

3. Results

3.1. Detection of RNA viruses in public D. melanogaster scRNA-seq datasets

A publicly available EE scRNA-seq dataset from *D. melanogaster* (BioProject accession: PRJNA547484) was obtained from the Sequence Read Archive (SRA; https://www.ncbi.nlm.nih.gov/sra; last accessed 15 November 2021) and investigated for the presence of viruses. A total of 2,672,571 out of 166,308,891 reads (1.6%) remained after filtering Drosophila-derived reads. Forty contigs were assembled and queried against the non-redundant (nr) GenBank database, leading to the identification of three RNA viruses in this dataset: DCV, DMelNV, and TV (Table 1). We also identified one contig with similarity to the drosophila endosymbiont *Wolbachia pipientis*, indicating that some flies might be infected by this bacterium. Hits to *Saccharomyces cerevisiae* and *Muntiacus reevesi* are most likely false positives or derived from sample contamination. Next, reads were aligned to the *D. melanogaster* and virus genomes to obtain gene and viral counts for each cell, which were used for clustering purposes. All clusters were identified based on a set of marker genes (Guo et al., 2019) (Supplementary Figure S1a). In total, 6, 55, and 766 cells infected with DCV, DMelNV, and TV, respectively, were found. Figure 1a shows a schematic representation of *D. melanogaster* gastrointestinal tract, along with the approximate spatial distribution of EEs subtypes. t-SNE reduction plots of the cell clusters are shown in Figure 1b. No infected EE

progenitor (EEP) could be found, and no cell from the I-a and III subtypes was found to be infected with DMelNV. Only eight cells were coinfected with DMelNV and TV, which is in perfect agreement with the hypothesis of independent infection based on the proportion of cells singly infected with DMelNV and TV (probabilities of infection with DMelNV 0.0110 and with TV 0.1534; hence, the probability of coinfection is 0.0017 and the expected number of coinfected cells $0.0017 \times 4994 = 8.4898$; Binomial test $p = 0.4902$).

Table 1. DIAMOND BLASTX results. Hits to Drosophila and synthetic constructs were omitted.

Dataset	Contig	Contig Length	*E*-Value	Lowest Taxonomic Rank
EE	k99_30	813	7.1×10^{-156}	Thika virus
	k99_1	541	1.2×10^{-95}	*Saccharomyces cerevisiae*
	k99_22	402	9.5×10^{-61}	*Saccharomyces cerevisiae*
	k99_3	306	1.8×10^{-51}	Thika virus
	k99_24	543	2.2×10^{-84}	Thika virus
	k99_7	328	2.0×10^{-37}	*Muntiacus muntjak*
	k99_10	2627	0	Thika virus
	k99_9	490	1.5×10^{-81}	Thika virus
	k99_28	319	1.1×10^{-24}	*Wolbachia pipientis*
	k99_29	488	2.2×10^{-56}	*Saccharomyces cerevisiae* YJM1401
	k99_14	349	4.8×10^{-61}	Thika virus
	k99_16	1027	1.2×10^{-189}	Thika virus
	k99_17	1694	0	Thika virus
	k99_39	12,357	0	Nora virus
	k99_21	337	3.0×10^{-52}	Drosophila C virus
	k99_12	656	2.2×10^{-54}	Drosophila C virus
	k99_36	3353	0	Drosophila C virus
	k99_38	551	1.3×10^{-97}	Drosophila C virus
MA	k59_25	207	6.5×10^{-13}	Nora virus
	k59_27	228	1.7×10^{-19}	*Muntiacus reevesi*
	k59_45	984	1.2×10^{-144}	Nora virus
	k59_62	2013	0	Nora virus
	k59_64	328	1.9×10^{-43}	*Saccharomyces cerevisiae*
	k59_35	485	3.4×10^{-41}	*Staphylococcus aureus*
	k59_72	754	3.8×10^{-127}	Nora virus

	k59_77	281	5.3×10^{-26}	*Macaca mulatta* polyomavirus 1
	k59_54	7298	0	Nora virus
	k59_58	344	4.6×10^{-16}	Nora virus
	k59_29	313	1.3×10^{-46}	*Drosophila melanogaster* tetravirus SW-2009a
	k59_21	443	4.2×10^{-78}	Drosophila A virus
	k59_44	325	2.2×10^{-52}	Drosophila A virus
	k59_32	328	1.2×10^{-53}	Drosophila A virus
	k59_11	700	1.3×10^{-86}	Drosophila A virus
	k59_36	414	6.3×10^{-68}	Drosophila A virus
	k59_76	854	1.9×10^{-167}	Drosophila A virus

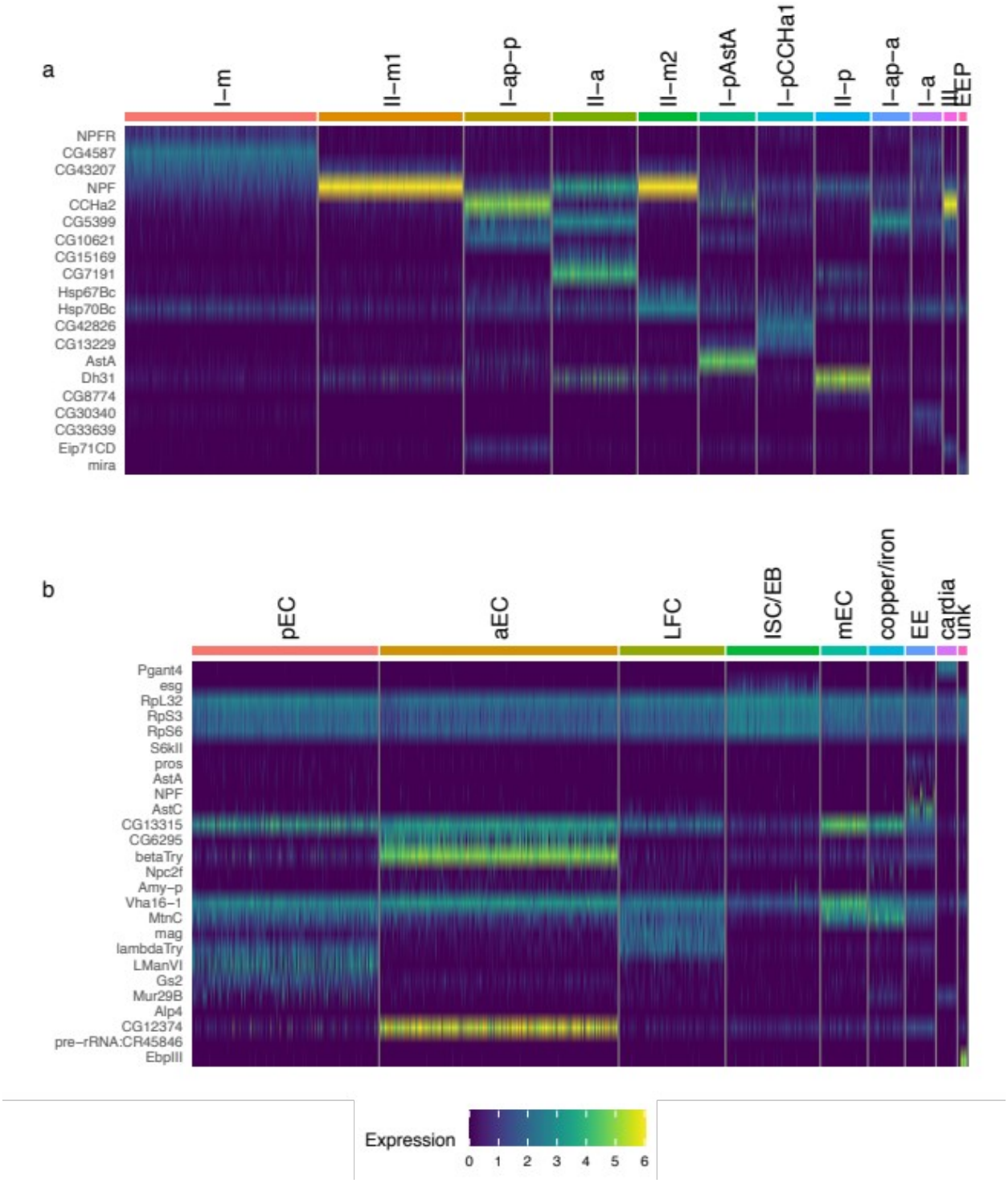


Figure S1. Marker genes used to identify clusters on the (**a**) EE dataset and (**b**) midgut atlas dataset

In addition to the EE scRNA-seq dataset, a publicly available scRNA-seq dataset from the entire midgut of fruit flies was also investigated for the presence of viruses (BioProject accession: PRJNA493298). In these data, to which we will refer as midgut atlas (MA) dataset, 173,810,419 out of 467,222,107 reads (37.2%) were retained after filtering. Queries of 60 contigs against the nr database led to the detection of DMelNV and DAV, which is likely a non-polyadenylated virus (Ambrose et al., 2009) (Table 1). However, no counts from DAV were found in cells after quality control, suggesting contamination from viral particles. Likely, the lack of a poly(A) tail hindered a precise detection of DAV in this dataset. Further inspection of one contig with hit to *D. melanogaster* tetravirus SW-2009a revealed that it consists of a false positive. Contigs with hits to *Muntiacus reevesi*, *S. cerevisae*, and *Staphylococcus aureus* are likely false positives or are derived from sample contamination. Like the EE dataset, cell clusters were generated and identified based on a set of marker genes (Hung et al., 2020) (Supplementary Figure S1b). Figure 1c shows t-SNE plots of major cell types that were annotated in this dataset. Clusters composed of aECs, mECs, and pECs were identified, along with clusters composed by cardia cells, located in the proventriculus region; intestinal stem cells and enteroblasts (ISC/EBs); large flat cells (LFCs), which are located in the posterior gastric region; copper/iron cells, located in the gastric region; and one cluster composed by unknown cell types (hereafter referred as "unk"). EEs were also identified in this dataset; however, their subtypes could not be reliably annotated. Clusters composed by differentiating ECs and EC-like cells (Hung et al., 2020) were not identified, possibly due to removal of these cells in our pre-processing step and/or due to them clustering to their most similar ECs. A higher prevalence of DMelNV was noted in this dataset, with 131 cells infected with this virus. However, some cell types had only a few DMelNV-infected cells, such as EEs (4), cardia (2), and unk (1), which likely hampered downstream analysis in these cell types.

3.2. DMelNV and TV exhibit different patterns of replication

The replication level of both DMelNV and TV was evaluated by analyzing both the percentage of viral RNA in the cells and log-normalized expression values (Figure 2a,b). A dramatic contrast in the replication levels between the two viruses was found. TV infected a higher number of cells compared to DMelNV but exhibited lower replication levels per cell. The percentage of viral RNA in TV-infected cells was always below 5% with only one exception (a single-cell I-ap-p), whereas for DMelNV, the percentage of viral RNA was up to ~80% of total mRNA. The percentage of infected cells on each cluster also varied drastically between these viruses (Figure 2c). TV infected a very low proportion of cells from subtypes located in the gastric region and was found in 66% of the cells of the I-p^{AstA} cell subtype located in the posterior region of the gastrointestinal tract that is characterized by the expression of AstA, AstC, and CCHa1.

One-way ANOVA tests showed a significant influence of EE subtype on the expression level of TV ($F_{10,755}$ = 1.9289, p = 0.0384) but not of DMelNV ($F_{8,46}$ = 1.5465, p = 0.1677)

although a significant influence of cell type was found for DMelNV when analyzing the MA dataset ($F_{8,122} = 2.2179$, $p = 0.03046$), as shown in Figure 2b.

3.3. Generic and cell-subtype-specific transcriptional response to TV and DMelNV

Differential expression (DE) analyses were conducted to uncover the generic and cellsubtype-specific cellular responses to TV and DMelNV (Appendix File S1). Seventeen, four, and one differentially expressed genes (DEGs) were found in the generic response to TV and DMelNV on the EE dataset and to DMelNV on the MA dataset, respectively (Figure 3a). Notably, in the generic response to TV, all DEGs were up-regulated. The generic response to TV of subtypes from the midgut posterior region was also investigated, with 17 and 39 genes found to be up- and down-regulated, respectively. Pathway enrichment analyses were not significant for these small subsets of genes, with the exception of upregulated genes in response to DMelNV on the EE dataset. In this list, genes related to regulation of heat shock factor 1 (HSF1)-mediated heat shock response, cellular response to stress and heat stress, and GABA synthesis, among a few others, were enriched (Figure 3b; Appendix File S2). Despite the fact that only eight cells coinfected with DMelNV and TV were annotated, we were able to detect an up-regulation of Hsc70-3 and a downregulation of the neuroendocrine protein 7B2 in these cells (Appendix File S1).

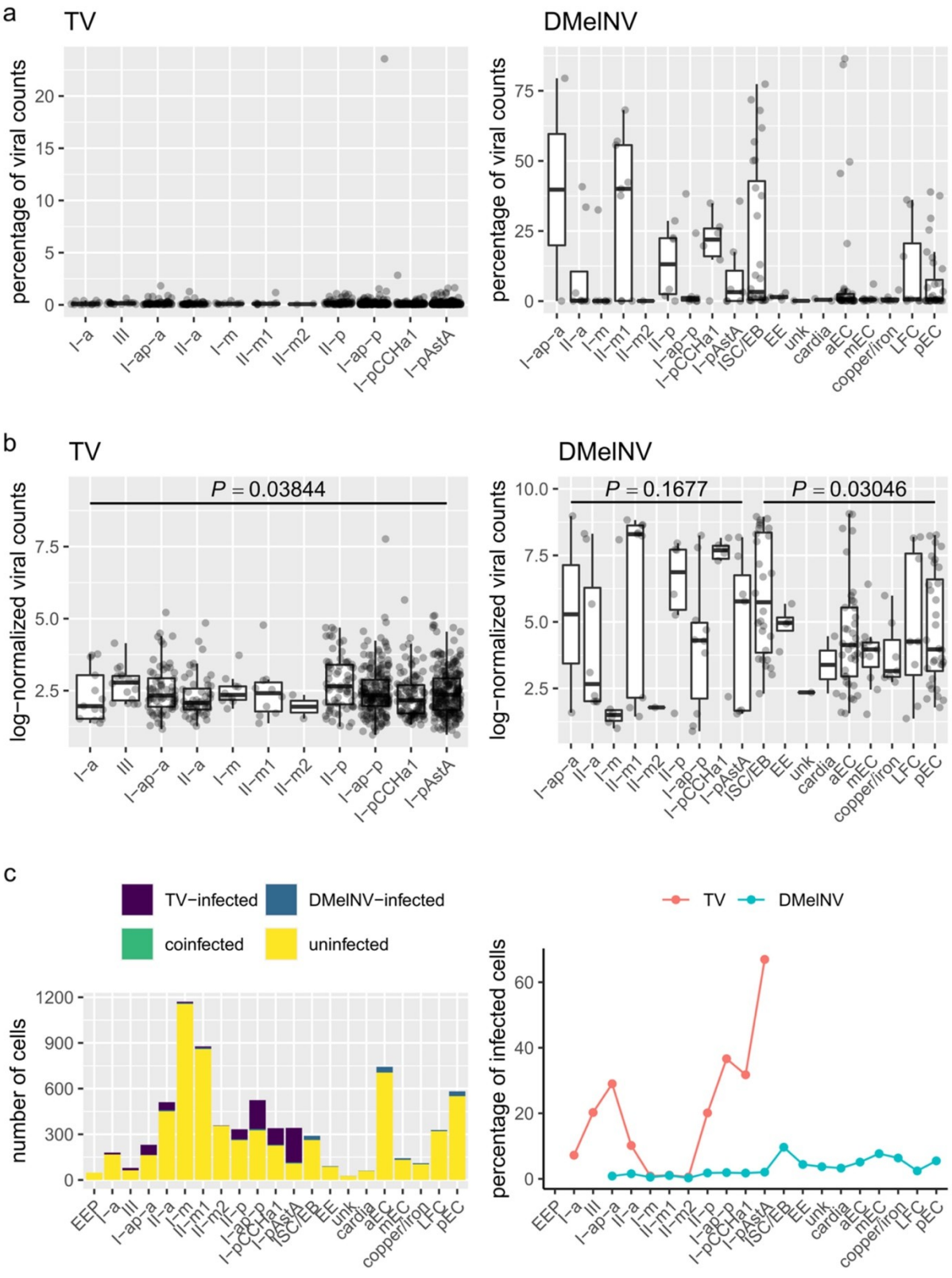


Figure 2. Statistic describing the spatial heterogeneity in the infection of TV and DMelNV along D. melanogaster gastrointestinal tract. (**a**) Boxplots showing the percentage of viral RNA from

TV and DMelNV for each cell subtype. Each point represents an individual cell. (**b**) Boxplots showing log-expression values of TV and DMelNV for each cell subtype. Each point represents an individual cell. (**c**) Number (left) and percentage (right) of cells from each subtype infected with TV and DMelNV.

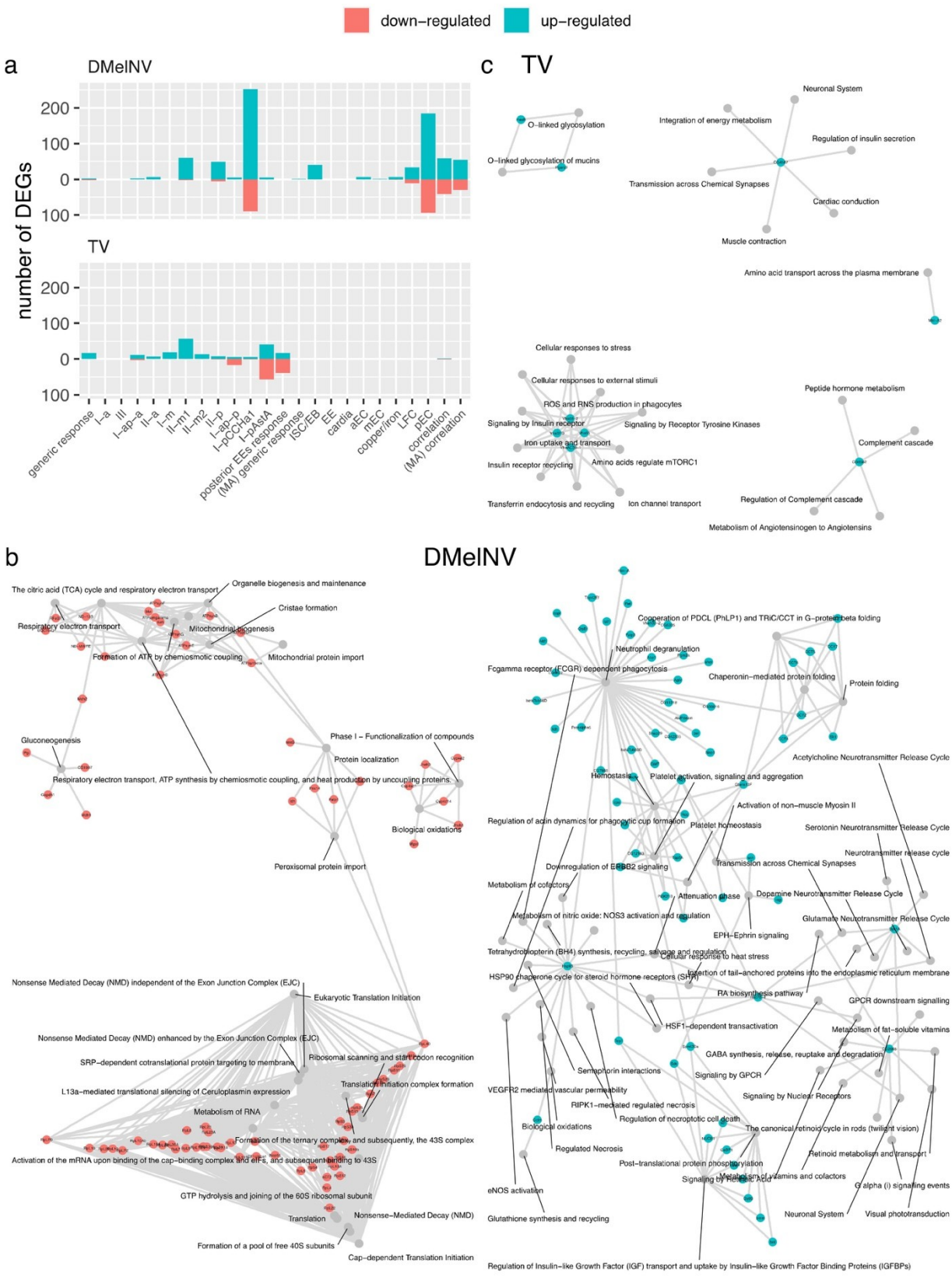

down−regulated
up−regulated
a
DMelNV
TV
number of DEGs
200
100
0
100
generic response
l−a
lll
l−ap−a
ll−a
l−m
ll−m1
ll−m2
ll−p
l−ap−p
l−pCCHa1
l−pAstA
posterior EEs response
(MA) generic response
ISC/EB
EE
cardia
aEC
mEC
copper/iron
LFC
pEC
correlation
(MA) correlation
c
TV
O−linked glycosylation
O−linked glycosylation of mucins
Neuronal System
Integration of energy metabolism
Regulation of insulin secretion
Transmission across Chemical Synapses
Cardiac conduction
Muscle contraction
Amino acid transport across the plasma membrane
Cellular responses to stress
Cellular responses to external stimuli
ROS and RNS production in phagocytes
Signaling by Insulin receptor
Signaling by Receptor Tyrosine Kinases
Iron uptake and transport
Insulin receptor recycling
Amino acids regulate mTORC1
Transferrin endocytosis and recycling
Ion channel transport
Peptide hormone metabolism
Complement cascade
Regulation of Complement cascade
Metabolism of Angiotensinogen to Angiotensins
b
DMelNV
The citric acid (TCA) cycle and respiratory electron transport
Organelle biogenesis and maintenance
Cristae formation
Respiratory electron transport
Mitochondrial biogenesis
Formation of ATP by chemiosmotic coupling
Mitochondrial protein import
Gluconeogenesis
Phase I − Functionalization of compounds
Protein localization
Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins
Biological oxidations
Peroxisomal protein import
Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)
Eukaryotic Translation Initiation
Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)
SRP−dependent cotranslational protein targeting to membrane
Ribosomal scanning and start codon recognition
L13a−mediated translational silencing of Ceruloplasmin expression
Translation initiation complex formation
Metabolism of RNA
Formation of the ternary complex, and subsequently, the 43S complex
Activation of the mRNA upon binding of the cap−binding complex and eIFs, and subsequent binding to 43S
GTP hydrolysis and joining of the 60S ribosomal subunit
Translation
Nonsense−Mediated Decay (NMD)
Formation of a pool of free 40S subunits
Cap−dependent Translation Initiation
Cooperation of PDCL (PhLP1) and TRiC/CCT in G−protein beta folding
Neutrophil degranulation
Fcgamma receptor (FCGR) dependent phagocytosis
Chaperonin−mediated protein folding
Protein folding
Acetylcholine Neurotransmitter Release Cycle
Hemostasis
Platelet activation, signaling and aggregation
Activation of non−muscle Myosin II
Regulation of actin dynamics for phagocytic cup formation
Platelet homeostasis
Serotonin Neurotransmitter Release Cycle
Neurotransmitter release cycle
Downregulation of ERBB2 signaling
Transmission across Chemical Synapses
Metabolism of cofactors
Attenuation phase
Dopamine Neurotransmitter Release Cycle
Metabolism of nitric oxide: NOS3 activation and regulation
Glutamate Neurotransmitter Release Cycle
EPH−Ephrin signaling
Tetrahydrobiopterin (BH4) synthesis, recycling, salvage and regulation
Cellular response to heat stress
HSP90 chaperone cycle for steroid hormone receptors (SHR)
Insertion of tail−anchored proteins into the endoplasmic reticulum membrane
RA biosynthesis pathway
GPCR downstream signalling
HSF1−dependent transactivation
Metabolism of fat−soluble vitamins
GABA synthesis, release, reuptake and degradation
Semaphorin interactions
Signaling by GPCR
VEGFR2 mediated vascular permeability
Signaling by Nuclear Receptors
RIPK1−mediated regulated necrosis
Regulation of necroptotic cell death
Biological oxidations
The canonical retinoid cycle in rods (twilight vision)
Regulated Necrosis
Retinoid metabolism and transport
Post−translational protein phosphorylation
Metabolism of vitamins and cofactors
Signaling by Retinoic Acid
G alpha (i) signalling events
eNOS activation
Neuronal System
Visual phototransduction
Glutathione synthesis and recycling
Regulation of Insulin−like Growth Factor (IGF) transport and uptake by Insulin−like Growth Factor Binding Proteins (IGFBPs)

Figure 3. Differential expression and pathway enrichment analyses. (**a**) Number of differentially expressed genes for each generic and cell-subtype-specific test. (**b,c**) Summary of all enriched pathways and its related genes found in the cellular response to TV and DMelNV infection, respectively. Each pathway is represented by a grey node connected to genes found to be differentially expressed.

When testing for cell-subtype-specific responses, the number of detected DEGs varied greatly between cell subtypes, with a few instances where no DEGs were found (Figure 3a; Appendix File S1). The highest number of detected DEGs was in the DMelNV-infected I-p^{CCHa1} and pEC subtypes. Up-regulated cell-subtype-specific DEGs in response to TV infection were enriched in genes related to glycosylation, insulin receptor recycling and insulin secretion, and cellular response to stress as well as in genes related to the immune system, such as complement cascade, and ROS and RNS production in phagocytes, in addition to others (Figure 3c; Appendix File S2). It is worth noticing that these insulin- and immune-related reactome pathways in Drosophila were inferred based on human orthologs (see https://reactome.org/documentation/inferred-events; last accessed 15 November 2021) (Jassal et al., 2020), and as such, genes in these categories were manually inspected. These pathways were enriched due to the up-regulation of the vacuolar H^+ ATPses subunits SFD, 55, 39-1, and 68-2 (VhaSFD, Vha55, Vha39-1, and Vha68-2, respectively), which are multirole proton pumps often expressed in a region-specific manner in the fruit fly's midgut (Allan et al ,2020; Miguel-Aliaga et al., 2018). Up-regulated cell-subtype-specific DEGs in DMelNV-infected cells showed an enrichment of genes related to protein folding and cellular response to heat stress and neutrophil degranulation, among others. Down-regulated cell-type-specific DEGs in response to DMelNV showed an enrichment of genes mostly related to translation, nonsense-mediated decay, gluconeogenesis, and respiratory chain, among others (Figure 3b; Appendix File S2).

In addition to testing for the effect of infection and the interaction between infection and cell subtype in DE analysis, the effect of viral percentage on each infected cell was also tested to find genes whose expression correlates to viral RNA accumulation. As viral RNA accumulation is expected to increase with time, we hypothesized that these genes may serve as indicatives of the time a cell has been infected. The expression of 3, 100, and 84 genes were found to be correlated to the accumulation of TV and DMelNV on the EE dataset and of DMelNV on the MA dataset, respectively. Possibly, the narrow range of variation of the accumulation level of TV hampered the detection of host-correlated genes. Henceforth, only genes in which expression values were correlated to the accumulation of DMelNV were further analyzed. Pathway enrichment analyses results for these genes were similar to the results obtained above with lists of DEGs. Pathways related to protein folding, response to heat stress, retinoid metabolism, and transport and glutathione synthesis and recycling were enriched in genes that correlated positively to DMelNV accumulation on the EE dataset. In both EE and MA datasets, an enrichment of pathways mostly related to respiratory chain was found for genes that correlated negatively to DMelNV accumulation. Pathways related to translation were also found for genes that correlated negatively to DMelNV accumulation on the MA dataset only (Appendix File S2).

3.3.1. Heat shock response to infection is cell-subtype- and virus-Specific

We focused on analyzing the heat shock response to DMelNV and TV given its wellknown role in antiviral defenses (Merkling et al., 2015) and the enrichment of heat stress-related pathways in lists of DEGs in response to DMelNV infection. Differential expression of genes associated with cellular response to heat stress (Reactome pathway: R-DME-3371556) was detected in one and eight cell subtypes as a response to TV and DMelNV infection, respectively (Figure 4). In EEs, Hsc70-3 and -4 were up-regulated as a generic response to DMelNV, and the expression of four heat shock proteins correlated positively to DMelNV accumulation (Figure 4b,d,e). Nevertheless, the number of differentially expressed heat shock proteins varied substantially between EE subtypes, where seven genes associated with heat stress were found to be up-regulated in the I-p^{CCHa1} subtype (Figure 4j). In the MA dataset, fewer differentially expressed heat stress-associated genes were detected. The expression of Hsc70Cb was found to correlate positively to DMelNV accumulation in this dataset, while Hsc70-3, Hsp23, starvin, and the elongation factor eEF1a1 were differentially expressed in a cell-subtype-dependent manner (Figure 4). In mammalian cells, the eEF1a1 homolog recruits HSF1 to induce a heat shock response (Vera et al., 2014). Therefore, its down-regulation in the II-p, pEC, and LFC cell types may be associated with a less robust antiviral defense. The differences between the EE and MA datasets might reflect differences in the fruit fly's genotype from each experiment. Regardless, these results suggest that heat shock response to DMelNV is cell-subtype-specific.

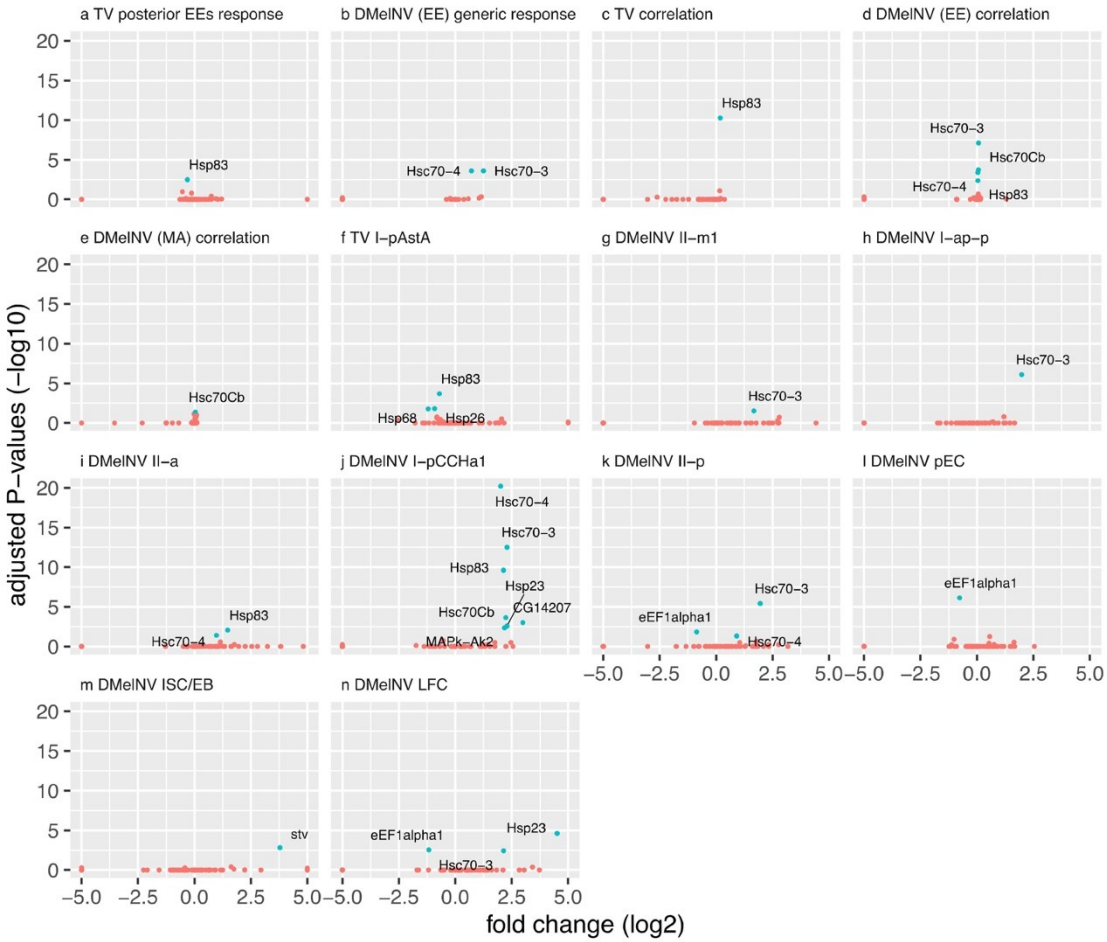


Figure 4. (**a–n**) Volcano plots of genes associated with heat shock response. Each panel correspond to a differential expression test where at least one gene associated with cellular response to heat stress (Reactome pathway: R-DME-3371556) was differentially expressed (adjusted $p < 0.05$). These genes are highlighted in blue. Log_2 -fold changes of genes that expression correlate to viral RNA accumulation correspond to the expected increase in gene expression when the viral percentage increases by one unit.

Interestingly, heat shock proteins were down-regulated in response to TV (Figure 4a,f). Hsp83 was down-regulated in TV-infected EEs subtypes from the posterior region of the fly's midgut, but surprisingly, its expression correlated positively to TV RNA accumulation (Figure 4a,c). Proteins Hsp26, Hsp83, and Hsp68 were down-regulated in the I-p^{AstA} subtype in response to TV infection (Figure 4f). Down-regulation of eEF1a1 and heat shock proteins in response to DMelNV and TV infection, respectively, may indicate that these viruses employ different strategies to target and suppress antiviral heat shock response.

3.3.2. Hsf regulon activity is higher in EEs and variable between cell subtypes

The activity of the Hsf regulon was investigated to explore whether heat shock response varies according to cell subtype. By analyzing cells from the whole midgut, we found that Hsf activity is higher in EEs (Figure 5A; Mann–Whitney U-test; adjusted $p < 0.0001$; average log_2-

fold change = 0.0133). Interestingly, Hsf activity was highly variable both between and within EEs subtypes (Figure 5b). Higher levels of Hsf regulon activity between and within cell types/subtypes may translate to higher intrinsic antiviral immunity, and may at least partially explain heterogenic response to viral infections.

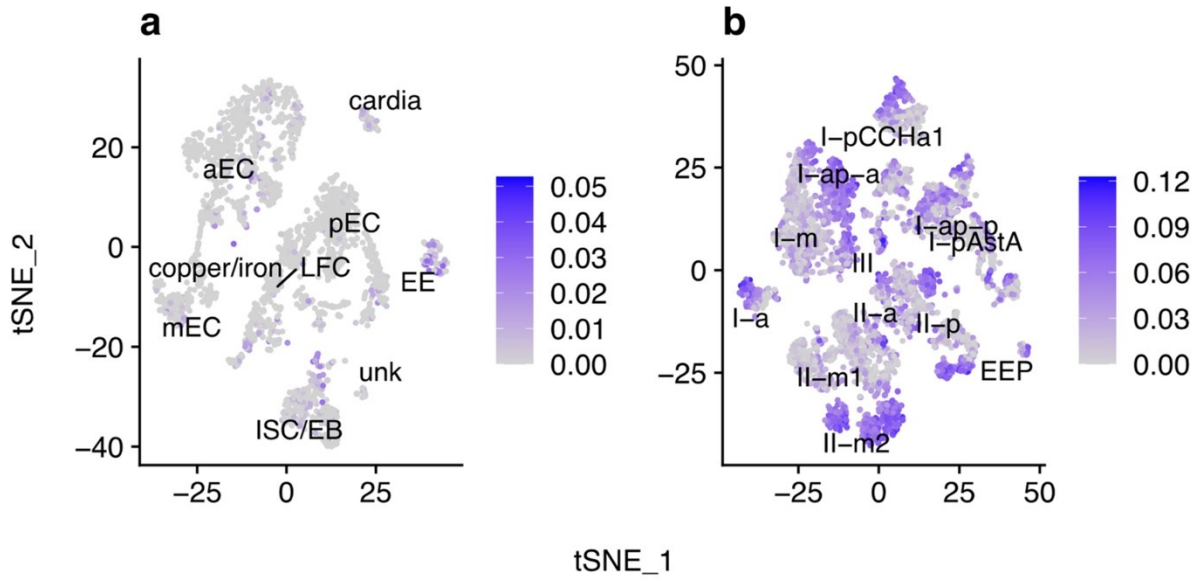


Figure 5. Hsf regulon activity in cells from the entire fly's midgut (**a**) and EEs (**b**) showing AUC values for each cell. Each point represents a single cell.

3.4. Time-Course Analysis of the Systemic Response to DMelNV

Bulk RNA-seq data from DMelNV-infected flies (Lopez et al., 2006) were obtained to analyze the systemic response to this virus. Most DEGs were found to be up-regulated, especially at earlier time points (Figure 6a). An increase in the expression of immune genes overtime has been previously found in this dataset (Lopez et al., 2006). Accordingly, an enrichment of genes related to complement cascade was found in up-regulated genes only at 20 and 30 days posteclosion (Figure 6b). Inspection of these genes revealed the presence of known Drosophila immune-related genes, such as thioester-containing protein 2, 3, and 4 (Tep2, 3, and 4, respectively) (Shokal & Eleftherianos, 2017). The *Aedes aegypti* Teps (*Ae*Teps), in particular *Ae*Tep1 and *Ae*Tep 2, were shown to regulate flavivirus infection (Cheng et al., 2011). A total of 270 genes were present in both cellular and systemic response to DMelNV, in addition to 68 genes whose expression was correlated to DMelNV accumulation that were also present in the systemic response to this virus at any time point (Appendix File S1). Up- and down-regulated genes in the cellular and systemic responses were generally consistent although in some cases, genes that were up-regulated in the cellular response were down-regulated in the systemic response and vice versa. The most interesting cases were from genes that were down-regulated only at later stages of infection (20 and 30 days post-eclosion) in the systemic response to DMelNV, but their expression was positively correlated to the virus' RNA accumulation in the cellular response (15 genes) or were up-regulated in the cellular response (84 genes). An enrichment of genes related to cellular response to stress, heat stress and external stimuli, as well

as related to HSP90 chaperone cycle for steroid hormone receptors and regulation of HSF1-mediated heat shock response was found in genes whose expression correlated positively to DMelNV accumulation but are down-regulated in the systemic response at 20 or 30 days post-eclosion (Figure 6c).

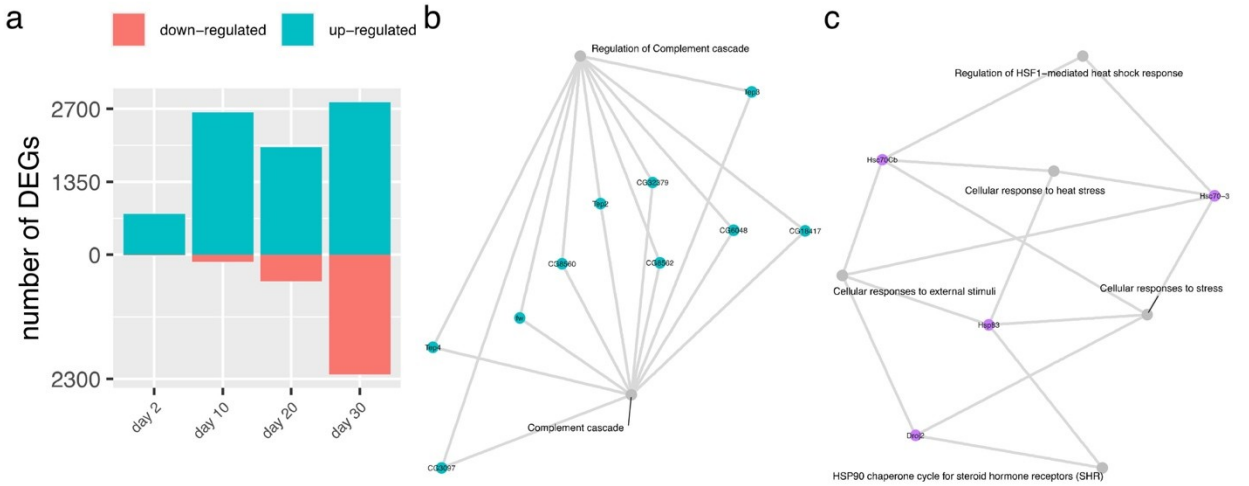


Figure 6. Time course analysis of the systemic response to DMelNV. (**a**) Number of differentially expressed genes at each time point. (**b**) Enriched pathways related to immune defense in up-regulated genes at 20 and 30 days post-eclosion. (**c**) Enriched pathways in genes which expression correlate to DMelNV RNA accumulation but are down-regulated at 20 and/or 30 days post-eclosion. Each pathway is represented by a grey node connected to genes found to be differentially expressed.

3.5. Mapping DEGs into D. melanogaster interactome

A network biology approach was taken to study the effects of the DEGs on the host transcriptome. In this approach, the *D. melanogaster* interactome is represented by a network where genes (nodes) are connected to each other by edges to represent their coordinated expression. Topological parameters of a predicted D. melanogaster interactome network were computed to investigate whether DEGs constitute essential nodes in the network. Enrichment of hubs and bottlenecks in lists of DEGs were investigated by computing two parameters: degree, i.e., the number of interactions of a gene and betweenness centrality, i.e., the number of shortest paths that pass through a gene in the network. Essential genes are very likely defined by either a high degree (hubs) or high betweenness centrality (bottlenecks) (Cho et al., 2012; Ahmed et al., 2018). We also sought to determine articulation points in the fruit fly's interactome, which are nodes that increase the number of connected components in a graph when they are removed, essentially disconnecting the network. The values of the critical exponent, mean degree, and mean betweenness centrality for each DEG list as well as a list of all articulation points are shown in Appendix File S2.

3.5.1. Regulation of essential genes in response to DMelNV and TV

While only a few DEGs in response to TV infection were found to constitute articulation points in the fly's interactome, a higher disruption of articulation points was found in response to DMelNV infection (Figure 7a). In accordance, by comparing the betweenness centrality of DEGs to that of the whole interactome, a wide down-regulation of bottleneck genes in the systemic response to DMelNV infection was found in 20- and 30-day old flies (one-tailed Mann–Whitney U-test; adjusted $p = 0.0027$ and $p < 0.0001$, respectively). Given that bottlenecks bridge different parts of a graph, this result suggests that the flow of information to some components of the fly's interactome is limited at later stages of infection in the systemic response to DMelNV.

Some genes related to heat shock response were found to comprise articulation points in the fly's interactome, such as Hsc70-3 and Hsc70-4, Hsp83, and the MAPK-AK2 kinase, which suggests that heat shock response may play a role in activating antiviral defenses. Surprisingly, Hsc70-3 and Hsp83 are not up-regulated in the systemic response to DMelNV but rather down-regulated at later stages of infection (Figure 6c; Appendix File S1).

Owning to the enrichment of translation-related genes that were down-regulated in the cell-type-specific response to DMelNV in the I-p^{CCHa1} EE and pECs subtypes, this subset of genes was mapped into the fruit fly's interactome. As expected, given the essential role of translation, these subsets of genes represent hubs in the fly's interactome (adjusted $p = 0.0221$ and $p = 0.0493$ for I-p^{CCHa1} EE and pEC, respectively; Figure 7b). In contrast, neither down-regulation of hub genes nor an enrichment of translational genes in down-regulated genes was found on the systemic response to DMelNV at any time point (Appendix File S2).

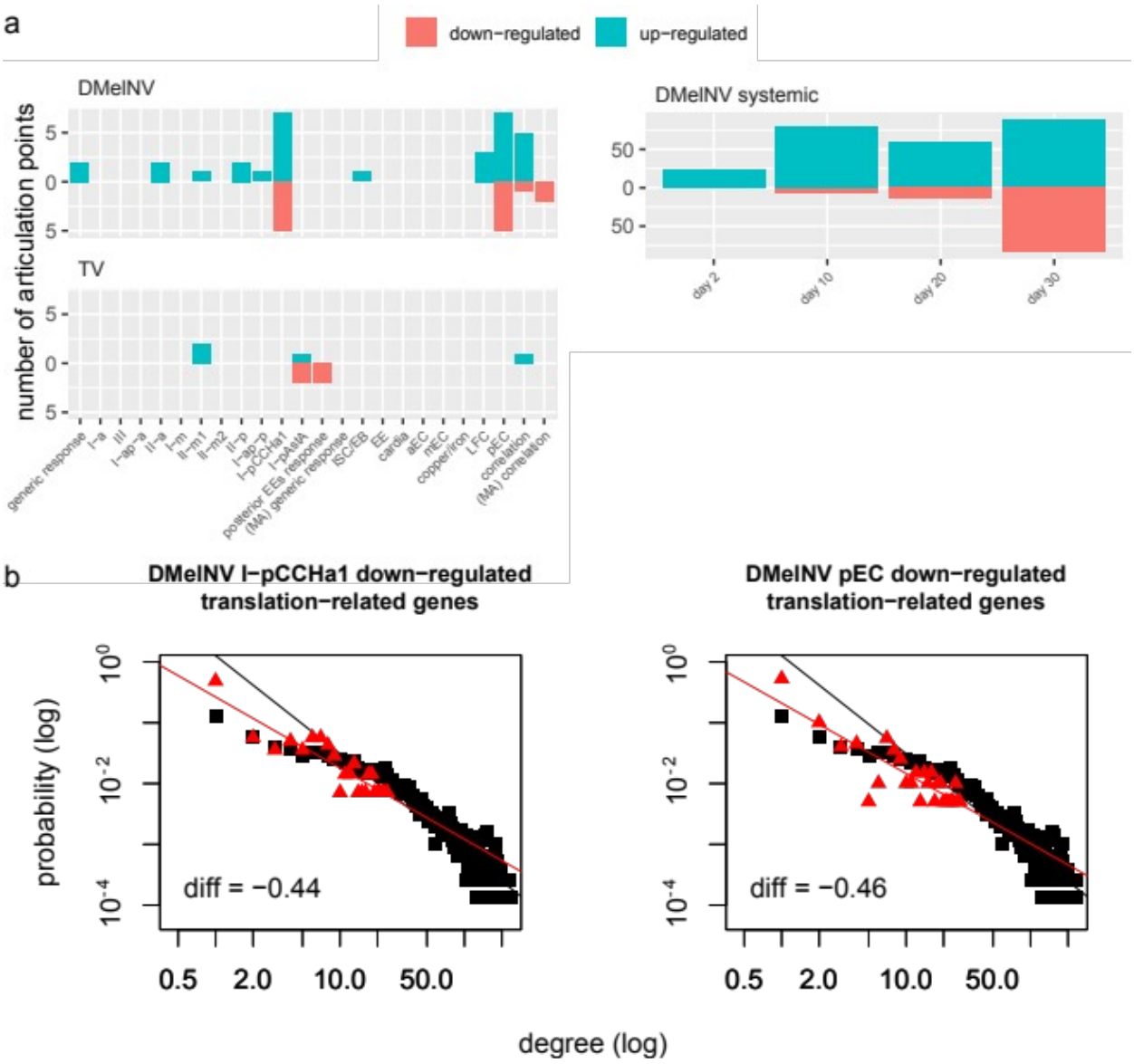


Figure 7. Perturbation of bottlenecks and hubs in the Drosophila interactome. (**a**) Number of articulation points found in each list of DEGs. (**b**) Down-regulated translation-related genes in response to DMelNV infection compose hubs in the drosophila interactome. Log-degree distributions of the whole *D. melanogaster* interactome (black) and subnetworks constructed with translation-related genes that are down-regulated in DMelNV-infected I-p^{CCHa1} and pEC cell subtypes (red). The slope of the regression lines represents the critical exponent of the power-law, γ.

3.6. Effects of virus infection on cell clustering and cell-type annotation

We investigated whether virus infection can jeopardize cell clustering and cell-type annotation analyses given that unacknowledged viruses may be confounding factors in single-cell data. Cluster annotation on the EE and MA datasets appear to not be influenced by cell infection status although some infected cells seemed to be closer to each other on the t-SNE

maps (Figure 8a). Overall, the EE and MA datasets are composed by heterogeneous sets of cells, and we asked whether the effects of infection might have a higher influence on clustering when analyzing a particular cell type/subtype. To investigate if this is the case, cells from the I-p^{AstA} subtype were subset, and we repeated the cluster generation analysis. Three clusters can be observed for the I-p^{AstA} subtype, of which cluster two appears to be composed mainly by uninfected cells (Figure 8b). These results indicate that the effects of infection are more noticeable when analyzing more homogeneous group of cells.

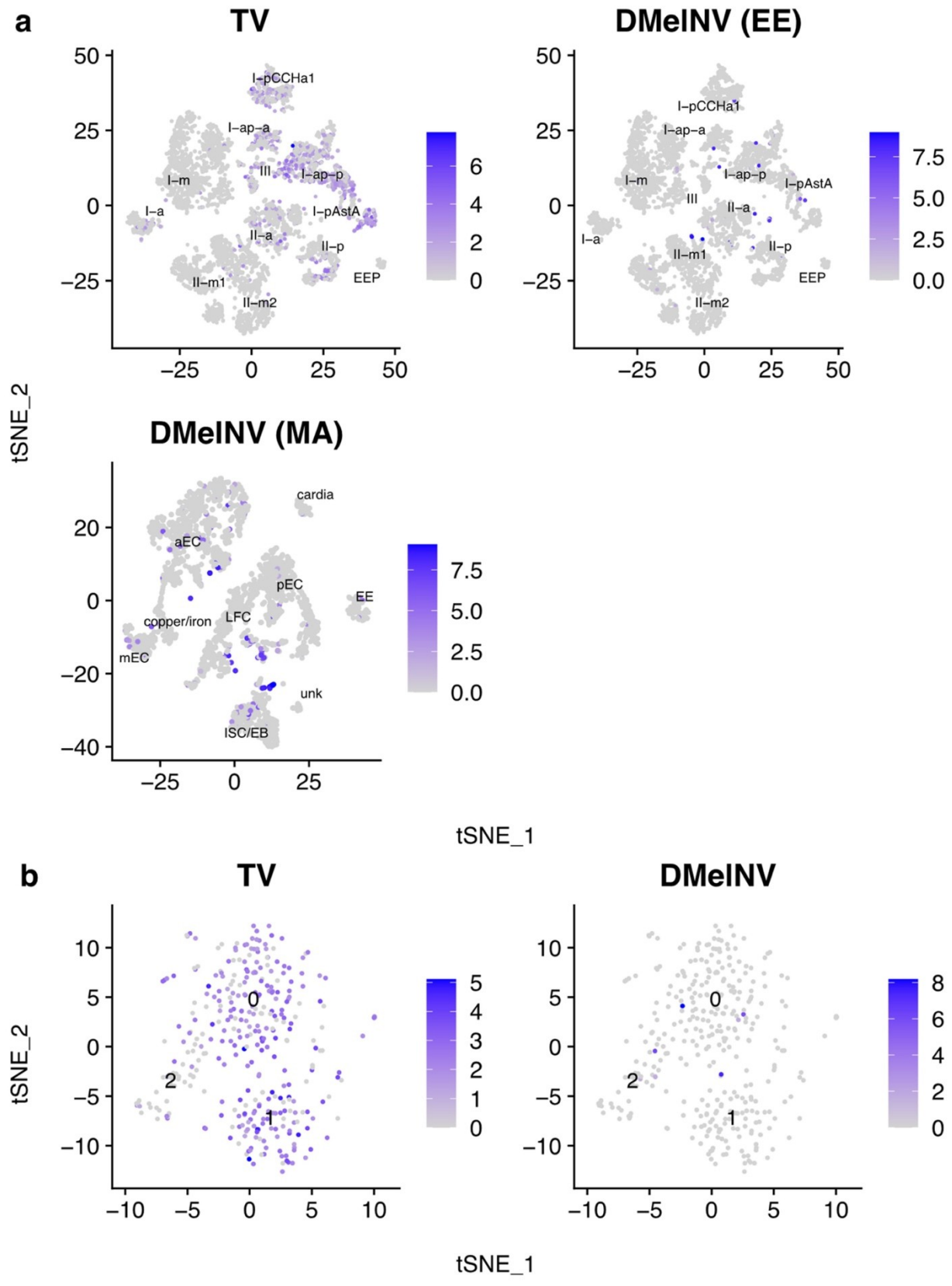

a
TV
DMelNV (EE)
DMelNV (MA)
tSNE_2
tSNE_1
I–pCCHa1
I–ap–a
III
I–ap–p
I–m
I–a
I–pAstA
II–a
II–p
II–m1
II–m2
EEP
cardia
aEC
pEC
EE
copper/iron
LFC
mEC
unk
ISC/EB
b
TV
DMelNV
0
1
2

Figure 8. Confounding effects of viral infection on cell clustering and cell type annotation. t-SNE reduction plots of the (**a**) EE and MA datasets and (**b**) I-p^{AstA} EE subtype. Each point represents a single cell. The log(pseudocount + 1)-transformed proportion of viral counts (multiplied by a factor of 10,000) is shown for each cell.

4. Discussion

Cellular heterogeneity and complexity are masked by whole tissue and other population-averaged transcriptomic methods. These limitations can be overcome by single-cell technologies, which are able to measure the RNA expression profile of thousands of individual cells in a single batch (Zheng et al., 2017). Viral RNA kinetics in individual cells, especially in the case of polyadenylated viruses, can also be captured simultaneously with transcriptional changes to infection, making this a powerful tool for virologists. By analyzing scRNA-seq data from *D. melanogaster*, we showed that two picorna-like viruses, DMelNV and TV, employ different strategies regarding replication and alteration of host transcription on EEs. DMelNV was also found on a variety of cell types in a second scRNA-seq data of the whole fly midgut, and possibly, both DMelNV and TV infect other cell types not included in this study. Whereas DMelNV is known to be a non-pathological enteric virus, the biology of TV remains unknown, as this virus has been recently described by a metagenomic approach (Webster et al., 2015b). Here, we showed that TV's capacity to infect and its replication level on EEs depended on the particular cell subtype being infected, and also, its accumulation was generally low compared to DMelNV (Figure 2). Up to 66% of the I-p^{AstA} subtype, located in the posterior region of the gastrointestinal tract, was infected with TV, while a low percentage of EEs from the gastric region was infected with this virus. On the other hand, DMelNV RNA accumulation was highly variable, sometimes exceeding 50% and up to nearly 80% of the total transcripts of a cell, and far less cells infected with this virus were detected on the EE dataset. Similar to DMelNV, a wide range in viral load/transcripts in infected cells was observed for other RNA viruses, such as vesicular stomatitis virus (Zhu et al., 2009), dengue and zika viruses (Zanini et al., 2018), poliovirus (Schulte & Andino, 2014), foot-and-mouth disease virus (Xin et al., 2018), and influenza A virus (IAV) (Steuerman et al., 2018; Heldt et al., 2015; Russell et al., 2018). Cell subtype had a significant influence on viral RNA accumulation level. While TV accumulation was influenced by EE subtype, we failed to detect a significant influence of EE cell subtype on DMelNV accumulation and only detected a significant impact of cell type on its accumulation when analyzing cells from the entire midgut.

EEs response to TV and DMelNV infection was diverse, and different sets of genes were generally altered between TV- and DMelNV-infected cells. For TV, genes mostly related to glycosylation, insulin receptor recycling and insulin secretion, and cellular response to stress were enriched in cell-subtype-specific up-regulated genes lists in addition to pathways that suggest activation of the immune system (Figure 3b; Appendix File S2). For DMelNV, genes mostly related to protein folding, cellular response to heat stress, and immune system (albeit distinct pathways to TV) were enriched in up-regulated genes, whereas genes related to translation and respiratory chain, among others, were enriched in down-regulated genes lists (Figure 3c; Appendix File S2).

Viral RNA accumulation is expected to increase as a function of time; therefore, genes whose expression correlates to viral RNA accumulation may be indicatory of time of infection. Pathway enrichment analysis results for genes whose expression correlated to DMelNV accumulation were similar to the results obtained from DEGs lists, suggesting that overall, genes that respond to virus infection can also be used to estimate time of infection.

Genes associated with heat shock response were differentially expressed in both DMelNV- and TV-infected cells and the late systemic response to DMelNV. A previous study found that Hsp23, Hsp26, and Hsp70 were up-regulated in S2 cells challenged with DCV or invertebrate iridescent virus 6 (IIV-6), whereas only Hsp23 was up-regulated in response to cricket paralysis virus (CrPV). However, in whole adult flies, only Hsp23 and Hsp70 were up-regulated upon infection with DCV, and no up-regulation of heat shock proteins were detected upon IIV-6 infection. In contrast, Hsp70 was found to be up-regulated in adult flies infected with CrPV in addition to Hsp23 (Merkling et al., 2015). These results suggest differences between cellular and systemic heat shock response to viral infections and also specificity to particular viruses. Similarly, while we found Hsc70-3 and Hsc70-4 to be up-regulated as a generic response of EEs to DMelNV, and several heat shock proteins, such as Hsc70Cb, Hsp23, and Hsp83, were found to be up-regulated in a cell-subtype-specific manner, no similar up-regulation of heat shock proteins was found in the systemic response to this virus. Surprisingly, Hsc70-3, Hsc70Cb, and Hsp83 were down-regulated at 20 and 30 days post-eclosion in the systemic response to DMelNV.

Down-regulation of genes related to heat shock response was also observed in the cellular response to DMelNV and TV and might be attributed to viral mechanisms to suppress antiviral defenses. The elongation factor eEF1a1 was down-regulated in DMelNVinfected cells from the II-p EE subtype and in pECs and LFCs. Hsp26, 68, and 83 were down-regulated in TV-infected cells, whereas Hsp26 and 63 were down-regulated only in the I-p^{AstA} EE subtype (Figure 4).

Hsf regulon activity was higher in EEs and variable both between and within its subtypes (Figure 5). Given that overexpression of Hsf in transgenic fruit flies was accompanied by restriction of viral replication, sometimes even to undetectable levels (Merkling et al., 2015), our results suggests that some EEs may have higher intrinsic immunity to viral infections.

Two cell subtypes showed a down-regulation of various proteins related to translation in response to DMelNV infection (Figure 3c; Appendix File S2). Analysis of this subset of genes showed that they compose hubs in the fly's interactome (Figure 4). This indicates a widespread shutoff of the translational machinery in these cell subtypes, possibly to tamper DMelNV replication. Alternatively, DMelNV targets only specific ribosomal genes as an attempt to suppress the translation of host RNA while enhancing the translation of its own proteins and therefore enhancing its own replication. Dysregulation of the translational machinery is a core function of ISGs in mammals (Kotlier et al., 2020; Li et al., 2015), which corroborates with the first hypothesis. Previous microarray and bulk RNA-seq studies were unable to find a similar down-regulation of translation-related genes in the systemic response of DMelNV-infected flies (Lopez et al., 2018; Cordes et al., 2013).

Expression of bottleneck genes in the systemic response to DMelNV decreases at late stages of infection as the expression of immune genes increases (Appendix File S2), suggesting a shift to a streamlined antiviral state in which resources are directed to the immune response. Further studies are necessary to test whether a late systemic downregulation of bottlenecks in response to viral infections is a conserved phenomenon in drosophila and other models. Both RNA and DNA viruses were shown to target hubs and bottlenecks in their host's interactome to regulate a wide range of cellular processes (Meyniel-Schicklin et al., 2012; Rodrigo et al., 2012; Martínez et al., 2021). While we were not able to measure any direct protein–protein interaction in our analysis, a perturbation of both hubs and bottleneck genes in DMelNV-infected flies was found. According to the centrality-lethality rule, disruption of essential genes in the interactome might disentangle and dismantle it (Ahmed et al., 2018; Jeong et al., 2001), which, on the one hand, might favor the virus infection process, while, on the other hand, it might diminish it. The down-regulation of translation-related genes in two cell subtypes infected with DMelNV and the downregulation of bottleneck genes in the late systemic response to this virus are, more likely, related to antiviral response.

The heat stress-associated proteins Hsc70-3, Hsp83, and MAP-AK2 constitute articulation points in the fly's interactome, suggesting they may play a role in activating antiviral pathways. If this is the case, it would be interesting to investigate whether a systemic antiviral response can be induced by cellular heat shock response. In contrast, we failed to detect any perturbation of hub genes in TV-infected cells, and a small number of articulation points were differentially expressed. As the time of infection is an important factor in regulation of the host transcriptome, and the scRNA-seq data used here are from 5–7 days old flies, we may have failed to detect a more substantial change to the host transcriptome at the cellular level. Given that this virus exhibited low RNA accumulation and elicited only moderate transcriptional response on infected cells where no significant modulation of essential genes was found, it is tantalizing to hypothesize that EEs are secondary targets of this virus.

We must acknowledge several drawbacks in our study. In addition to DMelNV, TV, and DCV, we also detected a non-polyadenylated virus, DAV, in the MA dataset in addition to the endosymbiontic bacteria *W. pipientis* in the EE dataset. The inability to precisely determine cells infected by non-polyadenylated viruses and intracellular bacteria means that some analyses, in particular differential expression, could be confounded by the presence of these pathogens. We also show that intrinsic heat shock response is highly variable, but the reasons for this variability is unknown. While some differences in the heat shock response between the EE and MA datasets might also be attributed to differences in the genetic background of the fruit fly or environmental factors, the heterogeneity in the cellular response to viral infections is already observed even within these datasets. Differences in the genetic background and environmental factors could also explain some differences between the cellular and systemic response to DMelNV, especially the lack of an up-regulation of the heat shock genes in the systemic response. Regardless of the motives, we show that heat shock response to viral infections is highly heterogeneous and cell-subtype- and virus-specific.

Lastly, given the presence of unacknowledged viruses in public scRNA-seq data, we hypothesize that viruses may be confounding factor in these kinds of experiments. In the datasets analyzed here, cells did not seem to cluster based on infection status when analyzing the EE and MA datasets as a whole (Figure 8a). However, when analyzing solely the I-p^{AstA} subtype, the effects of virus infection on cell clustering became more apparent (Figure 8b). Additionally, we have no information of which uninfected cells are responding to virus infections as bystanders, adding more hidden confounding factor to these analyses. One possibility to mitigate this problem would be to remove from the analysis all immune-related genes or genes that are correlated to viral load. A similar approach, where genes correlated to the top two principal components (PCs) composed by antiviral and inflammatory genes were omitted from the analysis, was performed to cluster pulmonary cells from IAV-infected and uninfected mice without the possible confounding effects of antiviral genes (Steuerman et al., 2018).

5. Conclusions

In this work, through analysis of in-vivo scRNA-seq data, we show the similarities and differences in the replication and infection strategies of two *D. melanogaster* viruses. We found a drastic contrast in the replication pattern between these two viruses. On the one hand, DMelNV only infected a few cells but exhibited high expression levels that sometimes exceeded 50% of the total mRNA of the cell, while, on the other hand, TV was able to infect a higher number of cells, but its replication level was generally low. Cells infected with either DMelNV or TV exhibited both generic and cell-subtype-specific transcriptional responses, and most importantly, heat shock response to viral infection was cell-subtype- and virus-specific. We detected a wide down-regulation of translation related genes in two cell subtypes in response to DMelNV infection. Further inspection of these translation-related genes revealed that they compose hubs in the host's interactome. By analyzing publicly available bulk RNA-seq data from DMelNV-infected flies (Lopez et al., 2018), we found that in this systemic response to DMelNV, bottleneck genes were down-regulated at 20 and 30 days post-eclosion. In contrast, no significant perturbation of hubs nor bottleneck genes were detected for TV.

Author Contributions: J.M.F.S. performed the analyses and wrote the draft. T.N., F.L.M., and S.F.E. supervised the study. All authors have read and agreed to the published version of the manuscript.

Capítulo 3. Tomato chlorotic spot virus (TCSV) putatively incorporated a genomic segment of groundnut ringspot virus (GRSV) upon a reassortment event

Este Capítulo foi publicado na revista *Viruses* (ISSN: 1999-4915).

Silva, J.M.F.; de Oliveira, A.S.; de Almeida, M.M.S.; Kormelink, R.; Nagata, T.; Resende, R.O. Tomato chlorotic spot virus (TCSV) putatively incorporated a genomic segment of groundnut ringspot virus (GRSV) upon a reassortment event. *Viruses* **2019**, 11, 187.

A versão presente nesta tese apresenta mudanças na formatação e pequenas alterações.

Abstract: Tomato chlorotic spot virus (TCSV) and groundnut ringspot virus (GRSV) share several genetic and biological traits. Both of them belong to the genus *Tospovirus* (family *Peribunyaviridae*), which is composed by viruses with tripartite RNA genome that infect plants and are transmitted by thrips (order Thysanoptera). Previous studies have suggested several reassortment events between these two viruses, and some speculated that they may share one of their genomic segments. To better understand the intimate evolutionary history of these two viruses, we sequenced the genomes of the first TCSV and GRSV isolates ever reported. Our analyses show that TCSV and GRSV isolates indeed share one of their genomic segments, suggesting that one of those viruses may have emerged upon a reassortment event. Based on a series of phylogenetic and nucleotide diversity analyses, we conclude that the parental genotype of the M segment of TCSV was either eliminated due to a reassortment with GRSV or it still remains to be identified.

Keywords: tospovirus; tomato chlorotic spot virus; groundnut ringspot virus; virus evolution; reassortment

1. Introduction

The tospoviruses have a great impact on agriculture since they can cause from mild to severe symptoms in their plant hosts (Pappu et al., 2009). These viruses have recently been reclassified within the *Tospovirus* (family *Peribunyaviridae*; order *Bunyavirales*), a genus that solely encompasses the "bunyaviruses" that have plants as hosts and are propagatively transmitted by thrips (order Thysanoptera) (Maes et al., 2019; Rotenberg et al., 2015). Since they have a tripartite single-stranded RNA genome, each segment is named according to its size as small (S), medium (M), or large (L) RNAs (Turina et al., 2016; Plyusnin et al., 2012). While the L RNA is of negative polarity, both the S and M segments contain an ambisense gene arrangement.

The main criterion for determination of new tospovirus species resides in the amino acid (aa) sequence of the nucleocapsid (N) protein that is encoded in the S RNA (Plyusnin et al., 2012). New isolates can only be recognized as belonging to a new species when their N protein shares less than 90% amino acid (aa) sequence identity with members of established species. Besides the N protein, the S RNA codes for a nonstructural protein (NSs) with RNA silencing suppression activity (Takeda et al., 2002). The M RNA codes for a cell-to-cell movement protein (NSm) and the precursor (GP) to the glycoproteins (Gn and Gc), while the L RNA codes for an

RNA-dependent RNA polymerase (RdRp) (Kormelink et al., 1994; Adkins et al., 1995). Although the N protein demarcates new species, phylogenetic trees based on any of the abovementioned proteins usually tell the evolutionary history of tospoviruses since the taxa tend to cluster similarly (de Oliveira et al., 2012). In case this is not observed, a reassortment event may likely be the cause for this incongruence.

Tomato chlorotic spot virus (TCSV) and groundnut ringspot virus (GRSV), although serologically related, were first suggested as members of different tospovirus species in the early 1990s after sequencing the *N* genes of isolates BR-03 (TCSV) and SA-05 (GRSV). Both isolates were initially identified from infected tomato in Brazil and groundnut in South Africa, respectively (de Avila et al., 1993a). Sequence analysis revealed 81% N protein identity between these isolates, while they could additionally be distinguished based on their biology (host range) and serology (de Avila et al., 1993a). Such observations helped to establish the 90% threshold for species demarcation. In 2004 (Lovato et al., 2004), the *Gn* and *Gc* genes of BR-03 and SA-05 were sequenced and revealed 92% aa identity. However, when looking in further detail, the phylogenetic distance between BR-03 and SA-05 based on the glycoprotein sequences was similar to that observed between tomato spotted wilt virus (TSWV) isolates, which was unexpected and questioned the N protein-based species demarcation (Lovato et al., 2004). The real meaning of this higher identity between the glycoproteins of TCSV and GRSV had been overlooked until a recent report on a proposed first interspecific reassortant tospovirus collected in the United States (U.S.) (Webster et al., 2011). This isolate contained the S and L segments from GRSV and was proposed to contain the M segment from TCSV. Since all viruses/isolates presented this genomic configuration and their parental genotypes have not been found, the authors suggested that GRSV/TCSV isolates were introduced in the U.S. as reassortants. To support the assumption that the M segment of these reassortants originated from TCSV, the authors used the glycoproteins of BR-03 and SA-05 for comparison, considering them as parental genotypes (Webster et al., 2011).

From all these results abovementioned, we started wondering whether all GRSV and TCSV isolates sequenced so far have been sharing a highly conserved M segment and, more importantly, whether the BR-03 and SA-05 isolates could be considered as parental genotypes of TCSV and GRSV, respectively. To increase the reliability of our analyses, the complete genome of the original BR-03 and SA-05 isolates were high-throughput sequenced in this work. After several analyses, our results suggested that TCSV may have incorporated the M segment of GRSV and that the parental genotype of TCSV was eliminated or remains to be identified.

2. Materials and methods

2.1. Sample preparation and sequencing

Nicotiana benthamiana leaf material infected with isolates BR-03 (TCSV) or SA-05 (GRSV) were kept frozen (−80°C) as a virus inoculum stock since the early 1990s at the Wageningen University, Netherlands. These frozen leaves were ground and used for mechanical inoculation of wild type N. benthamiana plants as previously described (de Avila et al., 1993b). Inoculated plants were kept in greenhouse until the onset of symptoms (two weeks). Total RNA

of symptomatic (systemic) leaves was extracted with RNeasy Plant Mini Kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. High-throughput sequencing was performed on a HiSeq™ 2000 platform (2 × 100 bp read length) at Macrogen (Seoul, South Korea).

2.2. De novo assembly of virus genomes

Reads generated by the Illumina platform were trimmed and de novo assembled using the software CLC Genome Workbench 6.5.2 (CLC bio, Aarhus, Denmark). Contigs corresponding to virus sequences were identified using Blastx (Altschul et al., 1990) against a Refseq virus database. Alignments with other tospoviruses were then performed using the program Geneious R8 (Biomatters, Auckland, New Zealand) to evaluate if the assembled contigs corresponded to the complete genomes of BR-03 and SA-05.

2.3. Phylogenetic analyses

A preliminary neighbor-joining tree was constructed using fragments of NSm genes from TCSV, GRSV, and TSWV isolates obtained from the NCBI portal using 1000 bootstrap replicates. TSWV sequences were included in this analysis for comparison. Maximum likelihood (ML) trees were constructed using the amino acid sequences coded by the S, M, and L segments of TCSV, GRSV, and TSWV available at the NCBI portal. For the S and M segments that code for more than one protein, the amino acid sequences were concatenated. Only tospovirus isolates with all open reading frames (ORFs) sequenced were included in this analysis. As done previously, TSWV isolates were included for comparison since they display an intraspecific genetic distance between each other. Multiple alignments were done using the program MUSCLE (Edgar, 2004) and the phylogenetic trees were built using the software PhyML v3.2 (Guindon et al., 2010) (bootstrap = 1000 replicates), both implemented in Geneious R8. The program ProtTest v.3.4.2 (Darriba et al., 2011) was used to estimate the best substitution model for all ML phylogenetic trees, which were then visualized and edited using the program FigTree v1.3.1. Heat maps for pairwise amino acid identities were built using the program SDT (Muhire et al., 2014).

2.4. Evaluation of synonymous-site variability and nucleotide diversity

To evaluate the accumulation of silent mutations along the S, M, and L segments of TCSV and GRSV, the synonymous-site variability was estimated with the program SynPlot2 (Firth, 2014) using the concatenated ORFs sequences. The program DnaSP (Librado & Rozas, 2009) was used to estimate nucleotide diversity parameters between TCSV and GRSV. Due to the low number of available sequences from the L segment of TCSV and GRSV, only fragments of N and NSm genes were included in this analysis.

2.5. TMRCA calculation

The time to most recent common ancestor (TMRCA) between TCSV and GRSV was estimated by Markov chain Monte Carlo (MCMC) Bayesian analysis for each segment using BEAST v2.5.0 (Bouckaert et al., 2014). Due to the low number of TCSV and GRSV sequences, TSWV was also included in this analysis to estimate the substitution rates using only coding

genomic regions. The temporal structure of the sequences was investigated with TempEst v1.6.0 (Rambaut et al., 2016) prior to the MCMC runs. Using only *NSm* genes for the M segment resulted in a better association between genetic distance and sampling dates, possibly due to the high number of these sequences on public databases. Trees were inferred under an uncorrelated lognormal relaxed molecular clock and a Bayesian Skyline tree prior (Drummond et al., 2006; Drummond et al., 2005). The package bModelTest (Bouckaert et al., 2017) was used as a site model to average over substitution models during the MCMC runs. Convergence of the parameters was determined by its effective sample size (ESS) with the program Tracer v1.6.0 and 10% of the samples of each run was discarded as burn-in.

3. Results

The majority of sequences available on public databases that match the M segment of TCSV and GRSV isolates correspond to fragments of *NSm* genes. Thus, these sequences were used to evaluate whether TCSV and GRSV isolates indeed share a highly identical M segment. Regarding genetic distance, *NSm* genes of TSWV isolates (most abundant on public databases) were included for comparison. Interestingly, the trees generated on the *NSm* genes showed that the interspecific diversity between TCSV and GRSV isolates is comparable to the intraspecific diversity between TSWV isolates (Figure 1). Additionally, the GRSV isolates segregated in two groups, one containing only GRSV isolates and another in which GRSV isolates intercalated with TCSV isolates.

By this work, there have been only partial sequences of the isolates BR-03 (TCSV) and SA-05 (GRSV) available on public databases. To circumvent this problem and further substantiate our findings, we have sequenced the complete genome of these isolates from infected *N. benthamiana* leaf material kept frozen at −80 ◦C. After de novo assembly, both genomes presented the standard tripartite single-stranded RNA pattern as seen in Figure 2. The S, M, and L RNA segments of each isolate were deposited in GenBank, respectively; (i) MH742961, MH742960, and MH742959 for BR-03 (TCSV) and (ii) MH742958, MH742957, and MH742956 for SA-05 (GRSV).

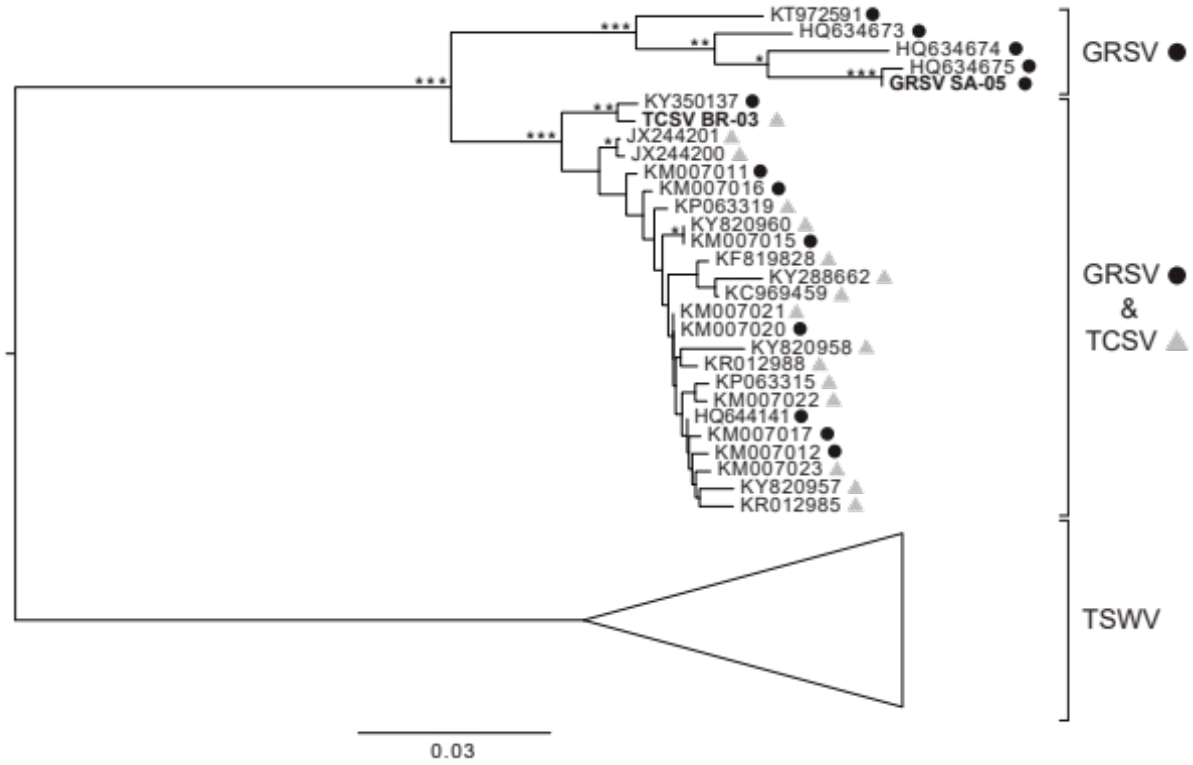


Figure 1. Phylogenetic tree based on *Nsm* gene fragments from groundnut ringspot virus (GRSV), tomato chlorotic spot virus (TCSV) and tomato spotted wilt virus (TSWV). TCSV sequences are representes by gray triangles next to their accession numbers, while GRSV are representes by black circles. Bootstrap values above 50%, 70% and 90% are represented by one, two and three asterisks, respectively.

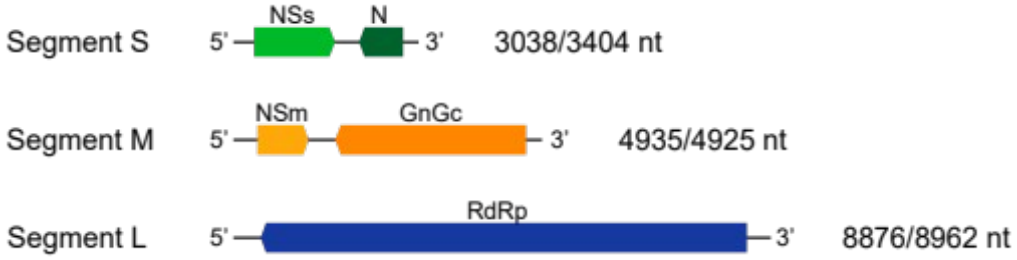


Figure 2. Genomic organization of TCSV and GRSV. The numbers on the right indicate the length of each segment of isolates SA-05 (GRSV) and BR-03 (TCSV), respectively.

With the availability of BR-03 and SA-05 genomes, phylogenetic trees were built with protein sequences derived from the three RNA segments. In the trees based on protein sequences from the S and L segments, TCSV and GRSV isolates clustered separately, forming two groups (Figure 3a). In contrast, in the phylogenetic tree based on proteins from M segments, TCSV isolates intercalated in a single group together with a previously assumed reassortant (Webster et al., 2011) and a GRSV isolate recently identified in peanut in Brazil (Figure 3a). The highest identity between a TCSV isolate and a GRSV isolate was 89.2 % for the S segment (N and NSs

proteins), 93.8% for the L segment (RdRp protein), and 99.8% for the M segment (Gn, Gc, and NSm proteins).

To investigate whether the M segments from TCSV and GRSV isolates have accumulated less silent mutations in comparison with the S and L segments, and are, therefore, more closely related, we evaluated the suppression of synonymous mutations of the coding regions in the whole genome of both TCSV and GRSV isolates. Synonymous site variability of *L* (L segment) and of *NSs* and *N* (S segment) was within the expected value with an observed/expected ratio ~ 1 or > 1. Differently, the *Gn*, *Gc*, and *NSm* genes (M segment) showed an observed/expected ratio < 1 with high significance as seen in the *p*-value graphic (Figure 3b). To verify whether TCSV isolates accumulated less diversity overtime than GRSV isolates, nucleotide diversity parameters were estimated using the program DnaSP (Librado & Rozas, 2009). These analyses revealed that the M segments of TCSV isolates presented lower nucleotide diversity than those from GRSV isolates (Table 1).

Table 1. Nucleotide diversity analysis of TCSV and GRSV.

Protein	Virus	Number of Isolates	Number of Segregating Sites (S)	Average Number of Differences (K)	Nucleotide Diversity (π)	Nucleotide Diversity with Jukes Cantor Correction (π JC)
N (212 nt)	TCSV	28	43	5.26190	0.02482	0.02554
	GRSV	25	57	9.64333	0.04549	0.04742
	TCSV and GRSV	53	96	22.98621	0.10843	0.12246
NSm (381 nt)	TCSV	16	27	4.28333	0.01124	0.01135
	GRSV	13	48	16.69231	0.04381	0.04581
	TCSV and GRSV	29	66	11.20936	0.02942	0.03066

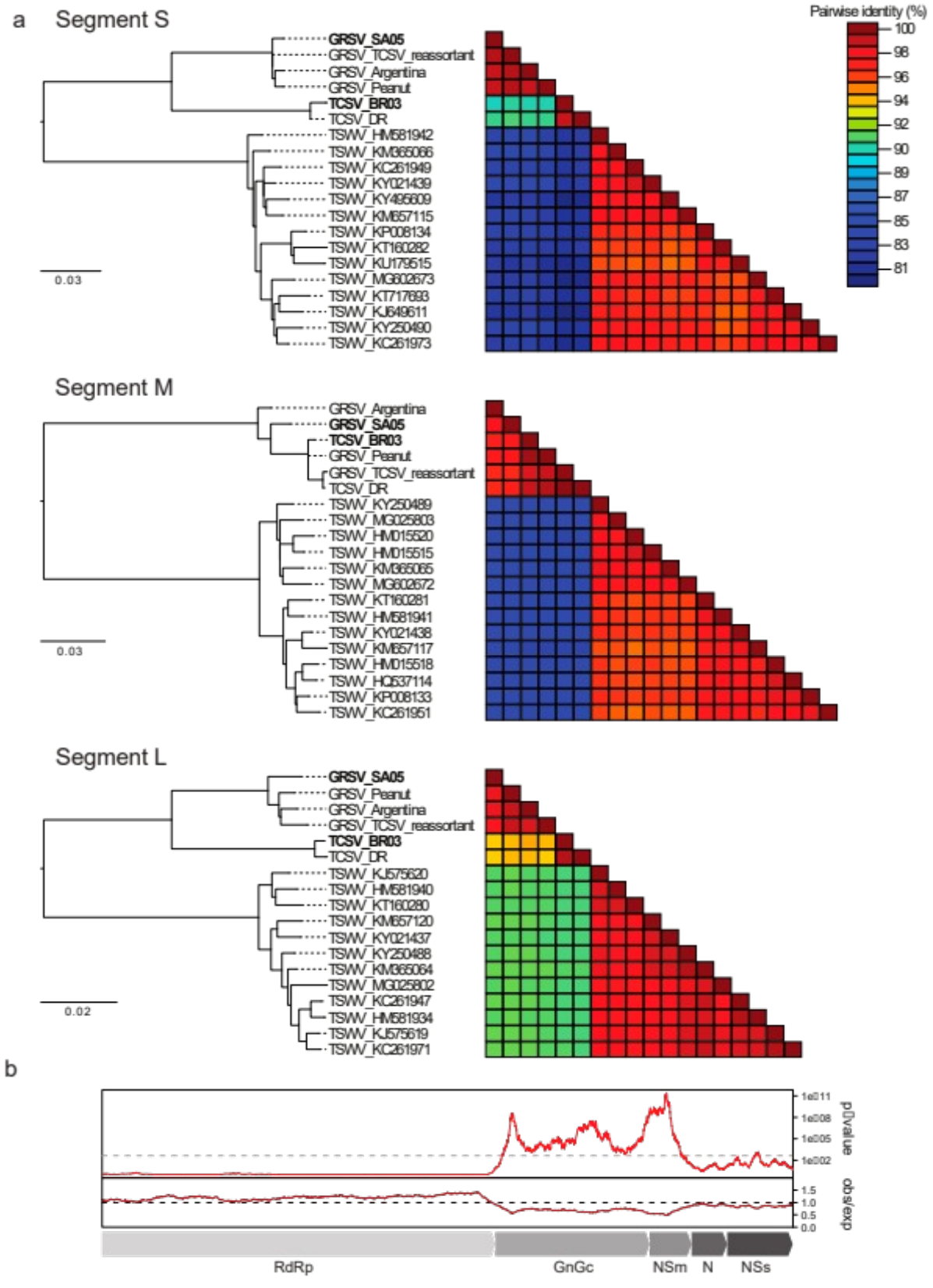
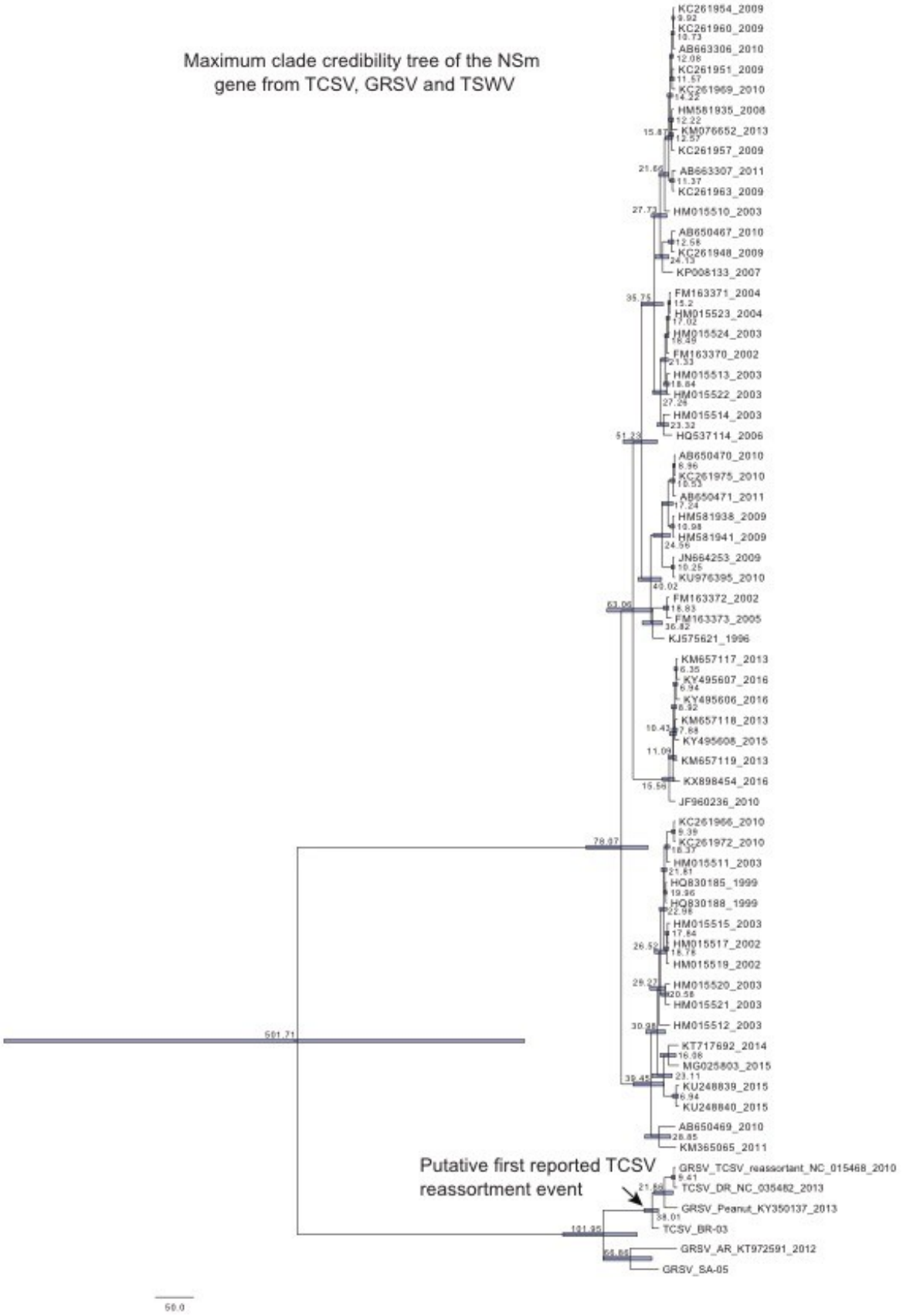

a
Segment S
Pairwise identity (%)
100
98
96
94
92
90
89
87
85
83
81
GRSV_SA05
GRSV_TCSV_reassortant
GRSV_Argentina
GRSV_Peanut
TCSV_BR03
TCSV_DR
TSWV_HM581942
TSWV_KM365066
TSWV_KC261949
TSWV_KY021439
TSWV_KY495609
TSWV_KM657115
TSWV_KP008134
TSWV_KT160282
TSWV_KU179515
TSWV_MG602673
TSWV_KT717693
TSWV_KJ649611
TSWV_KY250490
TSWV_KC261973
0.03
Segment M
GRSV_Argentina
GRSV_SA05
TCSV_BR03
GRSV_Peanut
GRSV_TCSV_reassortant
TCSV_DR
TSWV_KY250489
TSWV_MG025803
TSWV_HM015520
TSWV_HM015515
TSWV_KM365065
TSWV_MG602672
TSWV_KT160281
TSWV_HM581941
TSWV_KY021438
TSWV_KM657117
TSWV_HM015518
TSWV_HQ537114
TSWV_KP008133
TSWV_KC261951
0.03
Segment L
GRSV_SA05
GRSV_Peanut
GRSV_Argentina
GRSV_TCSV_reassortant
TCSV_BR03
TCSV_DR
TSWV_KJ575620
TSWV_HM581940
TSWV_KT160280
TSWV_KM657120
TSWV_KY021437
TSWV_KY250488
TSWV_KM365064
TSWV_MG025802
TSWV_KC261947
TSWV_HM581934
TSWV_KJ575619
TSWV_KC261971
0.02
b
1e011
1e008
1e005
1e002
pValue
1.5
1.0
0.5
0.0
obs/exp
RdRp
GnGc
NSm
N
NSs

Figure 3. Phylogenetic trees based on concatenated protein sequences encoded in the S, M, and L RNAs of TCSV, GRSV, and TSWV isolates, and the observed/expected synonymous mutations in the coding regions of TCSV and GRSV isolates. (**a**) Maximum likelihood trees and protein identity plots of the S, M, and L segments of TCSV, GRSV, and TSWV and (**b**) suppression of synonymous mutation variability in the concatenated ORFs of TCSV and GRSV using a sliding window of 250 codons.

To test whether the M segment from TCSV and GRSV are indeed more closely related than the S and L segments, the TMRCA was estimated by Bayesian phylogenetic analysis. The mean substitution rates were similar, with 2.4392E-4, 2.9163E-4, and 2.1783E-4 substitutions/site/year for the S, M (*NSm*), and L segments, respectively. Additionally, they presented narrow 95% high posterior density (95% HPD) intervals: 7.8493E-5 to 4.0956E-4, 1.5095E-4 to 4.4141E-4, and 1.1819E-4 to 3.2171E-4 for the S, M (*NSm*), and L segments, respectively. The mean TMRCA estimated between TCSV and GRSV was 548.52 years (95% HPD interval: 212.54 to 1065.5) for the S segment and 571.68 years (95% HPD interval: 305.03 to 933.48) for the L segment. Interestingly, but somewhat expected, the mean TMRCA of the M segment was much more recent, being 101.95 years (95% HPD interval: 57.71 to 154.47). The fact that the 95% HPD interval of the TMRCA between GRSV and TCSV for the M segment do not overlap with those from the S and L segments states that the TMRCA of the M segment is more recent. Focusing on the cluster containing intercalated GRSV and TCSV isolates (M segment), the first reported TCSV reassortment event may have happened about 38.01 years ago (95% HPD interval 29.71 to 48.43) as seen in Supplementary Materials Figure S1. The mean TMRCA of TSWV was 136.99 (95% HPD interval: 56.82 to 258.87), 78.07 (95% HPD interval: 43.53 to 124), and 121.15 (95% HPD interval: 69 to 193.24) years for the S, M, and L segments, respectively. Note that although the mean TMRCA for the M segment of TSWV is also more recent than for S and L segments, their 95% HPD interval overlap.

Maximum clade credibility tree of the S segment from TCSV, GRSV and TSWV


Maximum clade credibility tree of the NSm gene from TCSV, GRSV and TSWV
KC261954_2009
KC261960_2009
AB663306_2010
KC261951_2009
KC261969_2010
HM581935_2008
KM076652_2013
KC261957_2009
AB663307_2011
KC261963_2009
HM015510_2003
AB650467_2010
KC261948_2009
KP008133_2007
FM163371_2004
HM015523_2004
HM015524_2003
FM163370_2002
HM015513_2003
HM015522_2003
HM015514_2003
HQ537114_2006
AB650470_2010
KC261975_2010
AB650471_2011
HM581938_2009
HM581941_2009
JN664253_2009
KU976395_2010
FM163372_2002
FM163373_2005
KJ575621_1996
KM657117_2013
KY495607_2016
KY495606_2016
KM657118_2013
KY495608_2015
KM657119_2013
KX898454_2016
JF960236_2010
KC261966_2010
KC261972_2010
HM015511_2003
HQ830185_1999
HQ830188_1999
HM015515_2003
HM015517_2002
HM015519_2002
HM015520_2003
HM015521_2003
HM015512_2003
KT717692_2014
MG025803_2015
KU248839_2015
KU248840_2015
AB650469_2010
KM365065_2011
Putative first reported TCSV reassortment event
GRSV_TCSV_reassortant_NC_015468_2010
TCSV_DR_NC_035462_2013
GRSV_Peanut_KY350137_2013
TCSV_BR-03
GRSV_AR_KT972591_2012
GRSV_SA-05
50.0

Maximum clade credibility tree of the L segment from TCSV, GRSV and TSWV

Figure S1. Bayesian phylogenetic analysis of the S, M (*NSm*) and L segments of TCSV, GRSV and TSWV. Maximum clade credibility trees of the S, M (*NSm*) and L RNAs of TCSV, GRSV and TSWV. Node heights represent the mean heights of each MCMC run. Blue bars represent the 95% HPD interval of the tree heights. The TMRCA between TCSV and GRSV is shown on the S and L segment trees, and the putative first reassortment event between TCSV and GRSV is shown on the NSm tree.

4. Discussion

The evolutionary history of TCSV and GRSV isolates has been inferred in this work. For our analyses, we have sequenced the complete genomes of the first isolates (BR-03 and SA-05) of these viruses as they were in the 1990s. The focus was to understand why they share a highly conserved M segment, while their S and L genomic segments are more genetically distant. Previous works have observed this trait between these two tospoviruses (Webster et al., 2011; de Breuil et al., 2016). While these previous studies drew their conclusions from protein identity analysis, we examined the diversity accumulation in the coding regions of the three viral genomic segments and provided a refined phylogenetic analysis comparing GRSV, TCSV, and TSWV isolates.

Our analyses revealed that the TCSV and GRSV M segments exhibit the lowest diversity accumulation in comparison with the S and L segments. There are two possibilities that could explain this variation. Either the coding sequences of the M segment are highly functional at the RNA level, constraining the introduction of mutations, or a reassortment event took place. The first possibility seems very unlikely due to the fact that the substitution rate estimated is similar for the three RNA segments. Based on the nucleotide diversity analysis, the M segment of TCSV isolates presented less diversity accumulation than GRSV isolates, suggesting that the TCSV isolates are reassorted tospoviruses. In previous works, GRSV isolates found in the U.S were reported as reassortants, containing the M segment of TCSV (Webster et al., 2011; Webster et al., 2015a). To come up with their conclusions, the authors considered both BR-03 and SA-05 isolates as parental genotypes. Our analyses, however, suggest that the parental genotype of TCSV still remains to be identified or it was eliminated since the M segment of BR-03 is not genetically distant enough to be regarded as such.

Reassortment events appear to be frequent between TCSV and GRSV, given that TCSV sequences are scattered throughout the TCSV and GRSV cluster for *NSm* gene/M segment trees. The reason why this reassorted M segment was fixed in all known TCSV isolates remains to be investigated. In any case, it may have increased the virus adaptation to both plant and invertebrate hosts since the parental genotypes of TCSV have not been reported so far. It is worthy to notice that in recent years, TCSV have significantly increased its spread to other regions of the Americas (Webster et al., 2015a; Almeida et al., 2014; Martínez et al., 2018; Webster et al., 2013; Sui et al., 2018; Londoño et al., 2012) as a possible biological advantage of these recurrent interspecific reassortment events.

Author Contributions: A.S.d.O. and R.O.R. designed and supervised the study. R.K. and T.N. supervised the study. M.M.S.d.A. performed sample preparation. J.M.F.S. and A.S.d.O. performed data analyses and wrote the manuscript. All authors revised the manuscript.

Capítulo 4. Revamping the classification of the *Betaflexiviridae* based upon in-depth sequence analyses and proposal for new demarcation criteria

Este capítulo está em preparação para submissão.

João Marcos Fagundes Silva, Fernando Lucas Melo, Santiago F. Elena, Thierry Candresse, Sead Sabanadzovic, Ioannis E. Tzanetakis, ICTV Betaflexiviridae study group, Tatsuya Nagata. Revamping the classification of the *Betaflexiviridae* based upon in-depth sequence analyses and proposal for new demarcation criteria.

Abstract: The family *Betaflexiviridae* is composed by monopartite, positive strand RNA viruses. Currently, these viruses are classified based on their genome organization and nucleotide/amino acid identities of their replication and capsid proteins. Although biological traits such as vector specificities and host range are useful in taxonomy, this information is scarce for the majority of recently identified viruses which were characterized only for molecular traits. Accordingly, genomic information is being frequently used as the major, if not sole, criteria for virus classification. Herein, we propose an update on the current demarcation criteria for the family *Betaflexiviridae*, based on phylogenetic and stepwise pairwise identity analyses that should streamline classification of viruses in the family.

1. Introduction

Members of the family *Betaflexiviridae* (order *Tymovirales*) have monopartite, polyadenylated positive strand RNA genomes and form flexuous and filamentous particles. They encode an alpha-like replication protein (Rep) with methyltransferase (Met), papain-like protease (Pro), helicase (Hel) and RNA-dependent RNA polymerase (RdRp) motifs (Adams et al., 2012). The family is currently composed of two subfamilies; members of the *Quinvirinae* have a triple gene block (TGB) module that facilitates cell-to-cell movement (Morozov & Solovyev, 2003) whereas viruses in the *Trivirinae* encode a 30K-like movement protein (Melcher, 2000).

Virologists have traditionally relied on comparisons of multiple virus properties for their classification. These included morphological, biological, serological and epidemiological traits. However, with the advancement of genome sequencing procedures and, in particular, high throughput sequencing technologies, scientists more frequently rely primarily, if not only, on genomic information to classify novel viruses (Simmonds et al., 2017). A species is defined as "a monophyletic group of mobile genetic elements (MGEs) whose properties can be distinguished from those of other species by multiple criteria" (https://talk.ictvonline.org/information/w/ictv-information/383/ictv-code), and thus, it may be defined by a combination of properties derived from genomic sequences.

According to the currently valid species demarcation criteria for species and genera in this family, viruses belonging to different species should have less than about 72% nucleotide (nt) identity in the coat protein (CP) or replicase (Rep) genes or 80% amino acid (aa) identity in

the respective encoded proteins, while those classified in different genera usually have less than about 45% nucleotide identity in these genes (Adams et al., 2012). This approach worked well at the time it was proposed. However, with the increasing pace of virus discovery, it has encountered some difficulties when trying to classify some recently characterized viruses (Reynard et al., 2020; Marais et al., 2016; Maree et al., 2020; Liu et al., 2019; Cao et al., 2018). For example, it is unclear which of the two genes/proteins is more representative for taxonomic demarcation so that it is difficult to reach an unambiguous conclusion in situations where one of the genes/proteins shows identity values below the species threshold while the contrary case is observed for the second gene/protein. Furthermore, double thresholds using nt and aa sequences for each of the two genes/proteins may provide ambiguous messages (Reynard et al., 2020; Marais et al., 2016; Maree et al., 2020; Liu et al., 2019; Cao et al., 2018). To avoid this ambiguity, some families such as *Secoviridae* make use of dual molecular thresholds for two distinct proteins, namely the Pro-Pol and CP (Thompson et al., 2017).

In order to try to streamline the process of taxonomic assignation of novel viruses in the *Betaflexiviridae* family and, if needed, revise the sequence-based species discrimination criteria, we have re-analyzed pairwise genetic distances for almost all *Betaflexiviridae* isolates present in the databanks. The results obtained show that the Rep and CP genes/proteins do not diverge at the same rate in the family so that different species threshold should likely be used for the two genes/proteins. We further propose to take into account these novel criteria within a modified decision framework aimed to limit the ambiguities emerging in the current system as a consequence of the absence of prioritization between the Rep-based and CP-based criteria.

2. Materials and methods

2.1. Data collection

All complete and near complete genomes containing at least the complete Rep and CP sequences from the *Betaflexiviridae* available in GenBank in May 2021 were accessed and used for analyses. Taxonomic relationship for each accession was annotated based on currently accepted species. For accession that correspond to non-recognized species, taxonomic relationship was annotated based on relevant publications (Reynard et al., 2020; Marais et al., 2016; Zhao et al., 2020; Goh et al., 2021; Thekke-Veetil & Ho, 2019; Peracchio et al., 2020; Brewer et al., 2020; Wu et al., 2019; Marais et al., 2020; Wang et al., 2018; Liu et al., 2020; de la Torre-Almaráz et al., 2020; Goh et al., 2018; Park et al., 2019; Mumo et al., 2020; Diaz-Lara et al., 2020; Diaz-Lara et al., 2021; Luo et al., 2021; Li et al., 2020; Gazel et al., 2020; Goh et al., 2019; Marais et al., 2019; Thekke-Veetil et al., 2021; Silva et al., 2019; Maachi et al., 2020; da Silva et al., 2019; Zhou et al., 2018; Alabi et al., 2019), GenBank/EMBL/DDBJ description and BLAST (Altschul et al., 1990) analysis. Complete Rep and CP aa sequences were extracted using Geneious R8.1.9. Prior to phylogenetic and pairwise identity analysis, redundant sequences in which both Rep and CP had 100% aa identity with any other sequence were removed with CD-HIT (Fu et al., 2012). This resulted in two datasets, one for each protein, of 1230 sequences each.

2.2. Phylogenetic, recombination, pairwise identity and selection analyses

Alignments of the aa sequences of the Rep and CP were performed with MAFFT v7.110 (Katoh & Standley, 2013). The N- and C-terminal portions of the alignments, which are often poor aligned, as well as columns with more than 50% gap were removed with CIAlign (Tumescheit et al., 2020), resulting in two alignments containing 1976 (Rep) and 435 (CP) sites. Phylogenetic inference was then conducted with FastTree v2.1.11 (Price et al., 2012). A tanglegram showing the position of the same virus on Rep and CP-based trees was constructed to investigate incongruences between the phylogenies of the two proteins. Intragenus recombinant sequences were detected using the concatenated Rep and CP sequences with RDP5 command line tool (Martin et al., 2020). For the analysis of sequences that must be reassigned, the relevant clade was subset from the Rep tree with the R package treeio v1.15.7 (Wang et al., 2020). Pairwise identities of the Rep and CP sequences were obtained with SDT v1.2 (Muhire et al., 2014) using MUSCLE (Edgar, 2004). Selection analysis of species that exhibited a great disparity between the Rep and CP identities was performed with aBSREL (Smith et al., 2015) using only sequences flagged as non-recombinant. Downstream analyses were performed using the R programming language 3.6.3 with the treeio, ggtree v2.0.4 (Yu et al., 2017), ape v5.5 (Paradis & Schliep, 2019) and phytools 0.7-70 (Revell, 2012) packages. Tree topology tests were performed with IQ-TREE (Nguyen et al., 2015).

2.3. Accuracy of taxonomic criteria analysis

The accuracy of the proposed criteria with varying CP thresholds was investigated for selected species, which were selected due to their high number of available isolates or because they would be affected by the choice of different CP thresholds. For each species, true positives (TP) and false negatives (FN) were determined based on the ability of the applied criteria to maintain the current species, while true negatives (TN) and false positives (FP) were determined based on the capability of the applied criteria to distinguish randomly selected species. For Cherry rusty mottle-associated virus (CRMaV), cherry twisted leaf-associated virus (CTLaV) and cherry necrotic rusty mottle virus (CNRMV), since we are interested in the capability of the species demarcation in separating these species, TN and FP were calculated based on randomly selected interspecies comparisons of these viruses only. This same strategy was applied to cherry mottle leaf virus (CMLV) and peach mosaic virus (PMV). This analysis was also performed with R programming language.

3. Results and discussion

3.1. The Rep and CP exhibit different evolutionary histories

Phylogenetic trees of the family *Betaflexiviridae* were constructed for the Rep and CP amino acid sequences (Figure 1). Two major groups corresponding to the subfamilies *Trivirinae* and *Quinvirinae* are present in the Rep tree, however, in the CP tree, the *Quinvirinae* clade also included viruses from two genera, *Citrivirus* and *Wamavirus* belonging to the *Trivirinae*, indicating that they may have arisen through recombination, which is known to be a major force that drives the evolution of RNA viruses (Wolf et al., 2018), being no exception with the family *Betaflexiviridae* (Cao et al., 2018; Goh & Hahn, 2019; da Silva et al., 2019; Marais et al., 2015; Singh et al., 2012; Villamor & Eastwell, 2013; Yoon et al., 2014; Alabi et al., 2014; Zanardo et

al., 2014). To visualize incongruences between the Rep- and CP-based phylogenies, both trees were represented as a tanglegram (Figure 1). Some sequences did not maintain the same position in the Rep and CP trees, indicating that recombination occurred amongst *Betaflexiviridae* members. Even though some incongruences may be artifacts caused by low support branches on the CP tree, the positioning of the genera *Citrivirus* and *Wamavirus* as part of the *Quinvirinae* in the CP tree is well supported (Figure 1). To further test for the accuracy of the CP tree topology showing unprecise subfamily separation, unconstrained and Rep-constrained CP trees were compared by the approximately unbiased (AU) test with IQ-TREE. The Rep-constrained CP tree was rejected ($p = 0$), thus confirming that the Rep and CP possess distinct evolutionary histories. Given that the CP-based phylogeny does not clearly separates the two subfamilies, this analysis provides a further argument to support the notion that the Rep aa sequences should be used as the main demarcation criterion in the family *Betaflexiviridae*.

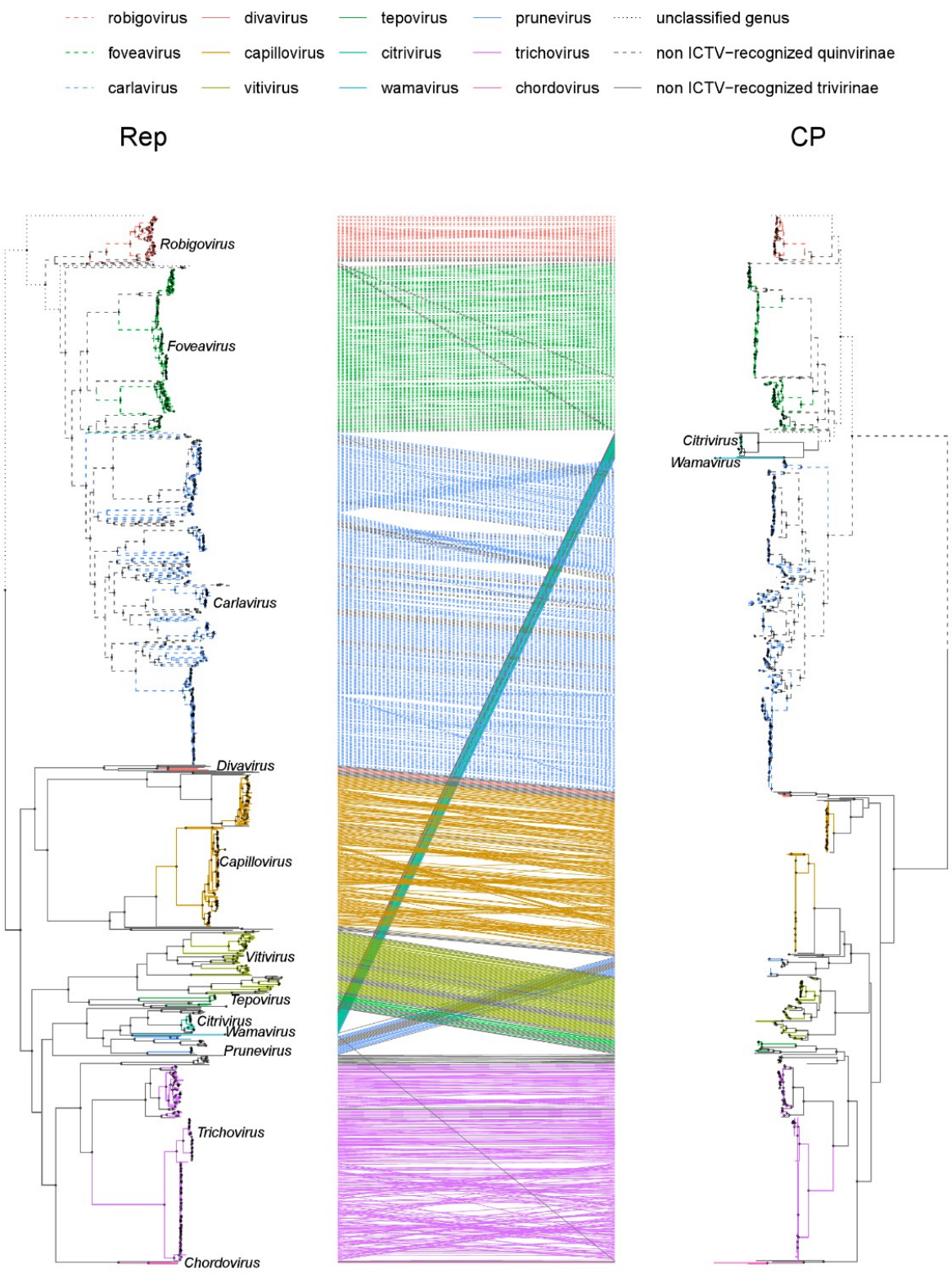

robigovirus
divavirus
tepovirus
prunevirus
unclassified genus
foveavirus
capillovirus
citrivirus
trichovirus
non ICTV−recognized quinvirinae
carlavirus
vitivirus
wamavirus
chordovirus
non ICTV−recognized trivirinae
Rep
CP
Robigovirus
Foveavirus
Carlavirus
Divavirus
Capillovirus
Vitivirus
Tepovirus
Citrivirus
Wamavirus
Prunevirus
Trichovirus
Chordovirus
Citrivirus
Wamavirus

Figure 1. Tanglegram of the replicase (Rep) and coat protein (CP) phylogenies of the family *Betaflexiviridae*. Trees were constructed by approximately maximum-likelihood (AML) using FastTree with alignments of the aa sequences of the Rep and CP from 1230 genomes. The Rep tree was rooted using Botrytis virus F (BotVF; accession: AF238884) as an outgroup, whereas the CP tree was midpoint-rooted. Support values above 70% are represented by diamonds.

3.2. Non-recombinant sequences exhibit a narrower pairwise distribution range for the CP

Given that the discrepancies between the Rep and CP phylogenies, we hypothesized that recombination in the *Betaflexiviridae* would have a significant impact on pairwise identities distributions. Therefore, the Rep and CP aa identities between each sequence pair was determined for all sequences (Figure 2a) and non-recombinant ones (Figure 2b). At the species level, pairwise distributions of non-recombinant sequences were more concentrated at the space containing identities above 80% for the Rep and ~85% for the CP (Figure 2b). Additionally, the correlation between the Rep and CP identity was higher for non-recombinant sequences at both genus and species levels, indicating that recombination has a significant impact on pairwise identity distributions. Based on these results, all subsequent pairwise identity analyses were performed using only non-recombinant sequences.

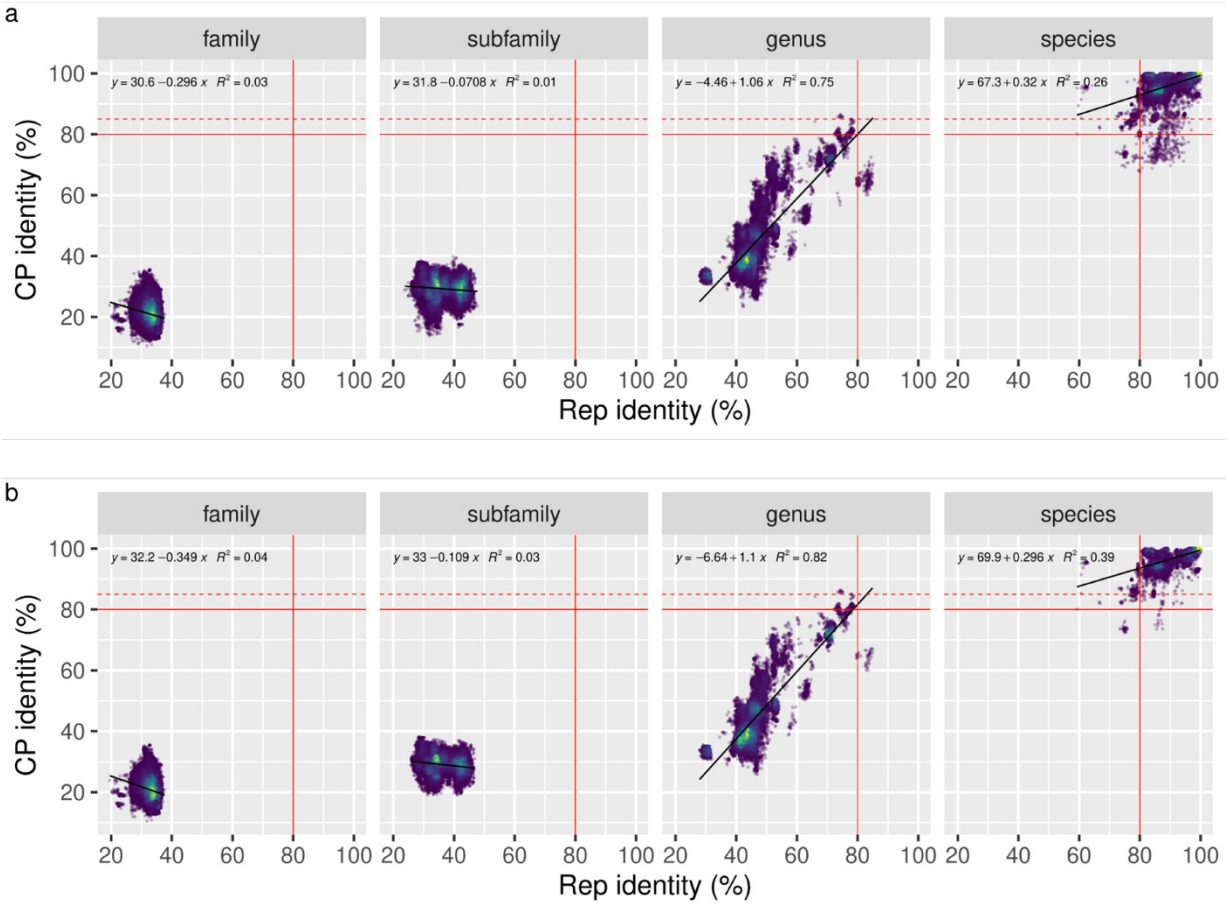


Figure 2. Effects of recombination on pairwise identities distributions. Rep and CP aa identities between two GenBank/EMBL/DDBJ accessions are represented as dots, warm and cold colors represent high and low density of data, respectively. Linear regressions were conducted for each panel to investigate the relationship between the identities of the two proteins. Red lines represent the current identity thresholds for species in the family and the red dashed line represent the proposed threshold for the CP (see Figure 3). (**a**) Dotplot generated using all available sequences from the family *Betaflexiviridae*; (**b**) pairwise comparisons between non-recombinant sequences.

3.3. Pairwise identities of the Rep and CP do not follow the same distribution

In addition to showing inconsistent phylogenetic topologies, the distribution of pairwise aa identities of the Rep and CP follow different distribution patterns (Figure 3a and b). At the lowest identity values (< 50%), the Rep distribution showed two well-defined peaks at 33% and 41%, respectively; whereas for the CP the lowest identity values are more spread and had less defined peaks at 20% and 28%, but with a lower number of pairwise comparisons falling at these percentages. On the other hand, at the higher identity values within the currently accepted species boundary (> 80%), the Rep distribution showed two peaks at 85% and 98%; while the CP peaks were at 93% and 99% (Figure 3a and b). These results indicates that at the species level the CP is more conserved than the Rep (Figure 3c and d). Linear regression and Pearson's correlation analyses of the Rep and CP pairwise identities indicate a higher conservation of the CP at the species levels; while at the genus level, an almost linear relationship between the pairwise identities of these proteins was found, in addition to the highest R^2 of 0.82 (Figure 2b). The shift from high conservation of the CP at the species level to an evolutionary rate comparable to that of the Rep, may be related to evolutionary pressures including vector specificities, novel host adaptations or other possible functions, such as suppression of pattern-triggered immunity (PTI) (Nicaise & Candresse, 2017). These differences may be also at least partially attributed to undetected recombination events. The distribution of three types of comparisons (family, subfamily and genus) shows clear overlaps and are, therefore, not precisely separated.

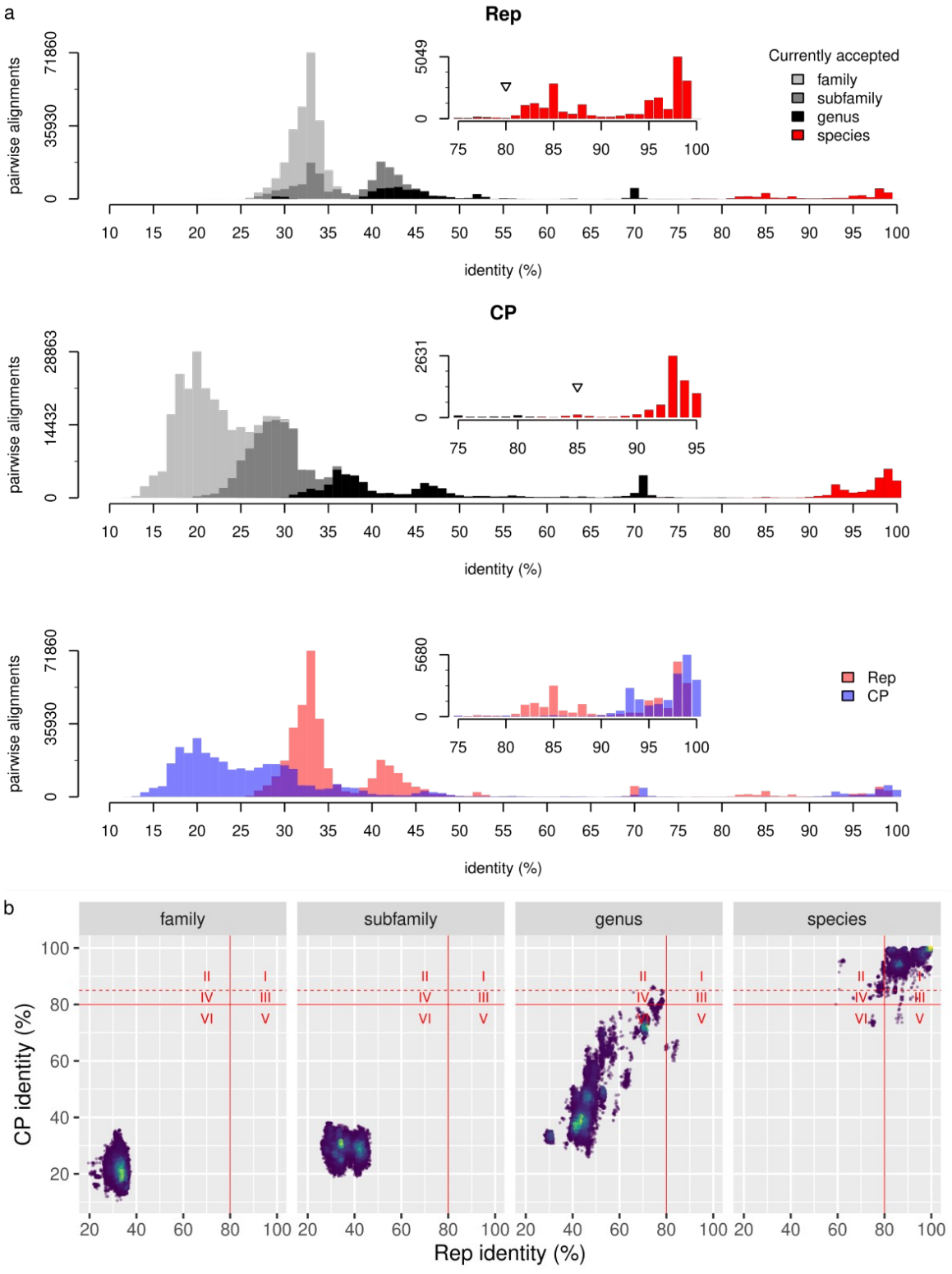


Figure 3. Distribution of aa pairwise identities of non-recombinant sequences. (**a**) Distributions of the pairwise identities of the Rep (up), CP (middle) and overlap of Rep and CP (bottom). Zoomed-in graphs of identities above 75% are shown for better visualization of within-species

identities. An open triangle is shown at the proposed Rep and CP identities threshold for species. (**b**) Dotplot of Rep and CP identities of non-recombinant betaflexiviruses with six categories by the proposed threshold of 85% for the CP. Red solid lines represent the current identity thresholds for the family and the red dashed line represent the proposed threshold of 85% for CP. Warm and cold colors represent high and low density of data, respectively.

When analyzing the Rep and CP identities between each sequence pair and comparing it with the current and proposed taxonomic criteria, six categories arose (Figure 3c; quadrants I-VI). Quadrant I contains bona-fide within-species comparisons. Quadrant II contains both within-genus and within-species comparisons, in which the identity of the Rep in some cases is much lower than threshold. Species with comparisons within this quadrant must be split in two separated species, if they meet the Rep monophyly criterion, since priority is given to the Rep identity. Quadrant III contains within-species comparisons in which the identities of the CP are lower than the proposed threshold (85%, see section below), but since the identities of the Rep are above its threshold, these species should remain unaltered. Quadrant IV contains both within-genus and within-species comparisons. Species within this quadrant must be split in two separated species in accordance with the proposed criteria. Quadrant V also contains both within-genus and within-species comparisons. Within-genus comparisons above the 80% identity threshold of the Rep will be analyzed below to assess whether these species should be merged into one (see Figure 4). Lastly, species with comparisons within quadrant VI should be clearly split in different species, based on either the current or proposed criteria.

Based on the Rep and CP evolutionary history and their pairwise identities distributions, we first conclude that recombination within the *Betaflexiviridae* complicates the use of both Rep and CP parameters simultaneously since these proteins present different evolutionary histories. Given that the evolution of RNA viruses is better understood based on the evolution of their polymerase, we propose that the phylogeny of the Rep should be used for the demarcation of monophyletic species. Secondly, at the species level the CP is more conserved than the Rep, and thus, the same threshold of 80% for both proteins is inappropriate. We propose that the Rep identity should be given priority over the CP identity, and that a higher threshold for CP should be used. A threshold for Rep of 80% is maintained and a new threshold for the CP of 85% is proposed as determined below.

3.4. Accuracy of taxonomic criteria to recapture accurate species demarcation

We then sought to evaluate the propriety of the current identity criteria in recapturing accurate species demarcation. For this purpose, possible conflicts between the virus identification based on the genome sequences and biological properties for species demarcation were discussed. Some interspecific comparisons have their Rep or CP aa identities above the current threshold (Figure 3b, quadrants II, IV and V at the genus panel).

The CP pairwise identity between grapevine virus D (GVD) and grapevine virus J (GVJ) is above 84.2%, implying they should be merged into one species. Cherry rusty mottle-associated virus (CRMaV), cherry twisted leaf-associated virus (CTLaV) and cherry necrotic rusty mottle virus (CNRMV) can be distinguished based on their symptomatology, although the association

of CNRMV to necrotic rusty mottle disease is limited (Villamor & Eastwell, 2013; Villamor et al., 2015); however, based on the current criteria they should also be members of the same species due to their CP identities. This disagreement of the identities criteria to recapture distinct biological properties of these robigoviruses has been previously noted (Villamor & Eastwell). Similarly, cherry mottle leaf virus (CMLV) and peach mosaic virus (PMV) can be differentiated based on host range, vector and symptomatology (James et al., 2006), but pairwise CP identities between these viruses are above the current threshold (Figure 4a). The pairwise Rep identities between Asian prunus virus 1 and 2 (APV1 and APV2, respectively) are at the ~80% threshold, while the Rep identities between apricot latent virus (ApLV) and apple stem pitting virus (ASPV) are above 80% (Figure 4b). ApLV and ASPV can be discerned based on host range (Nemchinov et al., 2000), and thus, the 80% identity threshold of the Rep is not able the recapture their distinct features.

Next, we sought to determine a new threshold for the CP that can recapture adequate species demarcation while causing minimum changes to other species of the family. We propose that priority is given to the Rep, meaning that if the Rep phylogeny and identities are not sufficient for a clear demarcation (for example, if Rep identities are at the borderline ranging of 78-82%), CP identities must be evaluated. By giving preference to the Rep, a clear separation is seen between CRMaV, CTLaV and CNRMV, as well as between GVD and GVJ. However, the separation between CMLV and PMV, and between APV1 and APV2 is still ambiguous, and thus, the CP criteria is applied. We found that an 85% CP identity is sufficient to distinguish these species (Figure 4). Only the separation between ApLV and ASPV could not be recaptured by sequence identities alone in our proposed criteria. Since the Rep and CP identities between APV1 and APV2 and between ASPV and ApLV do not follow the same trend for the family (CP more conserved than the Rep), branch-specific selection analyses were performed to investigate whether positive selection can account for this variation in Rep and CP identities trend. Evidence of positive selection was found in the Rep of ApLV.

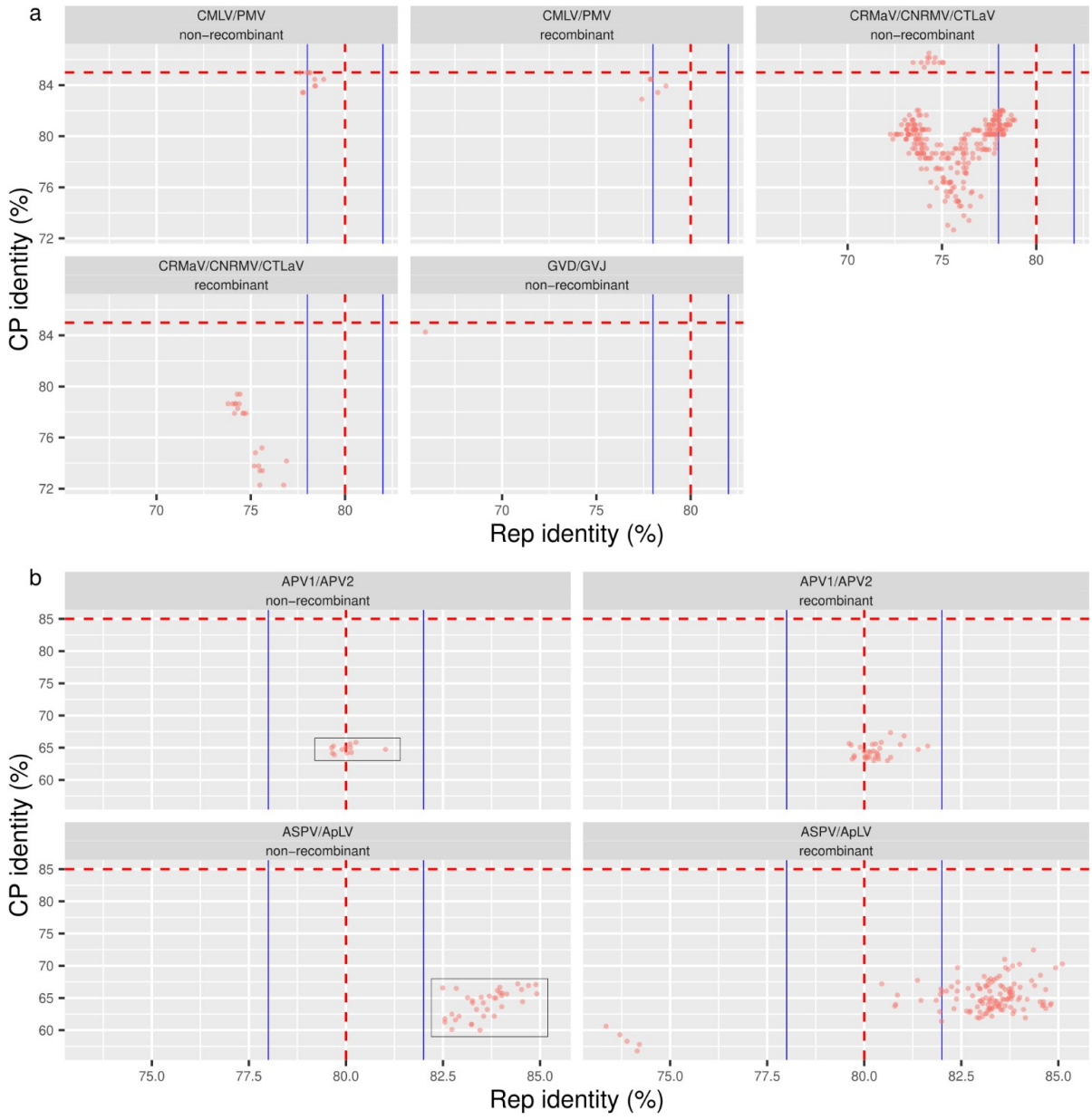


Figure 4. Analysis of within-genus comparisons above the species threshold of the CP and Rep. (**a**) Pairwise comparisons in which the CP threshold is above the current 80% threshold; (**b**) Pairwise comparisons in which the Rep threshold is above the current 80% threshold. Red dashed lines represent the proposed thresholds for the Rep and CP, and blue lines represent the borderline range of the Rep. Comparisons between Asian prunus virus 1 and 2 (APV1 and 2, respectively) are within the borderline range of the Rep (~78-82%) but the identity of the CP is much below its threshold. Although the identities of the Rep between apple stem pitting virus (ASPV) and apricot latent virus (ApLV) are above the 80% identity threshold, differentiation between these species is supported by their distinct host ranges. Rectangles show cases where the discrepancy between the identities of the Rep and CP cannot be explained by recombination.

The accuracy of the proposed criteria with varying CP thresholds was investigated for selected species. We found that the 85% CP threshold provides the best cost/benefit in maintaining the species demarcation while distinguishing CRMaV, CNRMV and CTLaV and CMLV and PMV (Figure 5a). However, at 85% CRMaV should be split into two species, where the 82% CP threshold is the optimal CP threshold for the separation of CRMaV, CNRMV and CTLaV. These results suggests that optimal results can be achieved by a flexible CP threshold which can be obtained by accuracy statistics analysis. On the downside, accuracy is better calculated when bona fide species can be determined with the aid of biological properties, which might not be available for newly described sequenced.

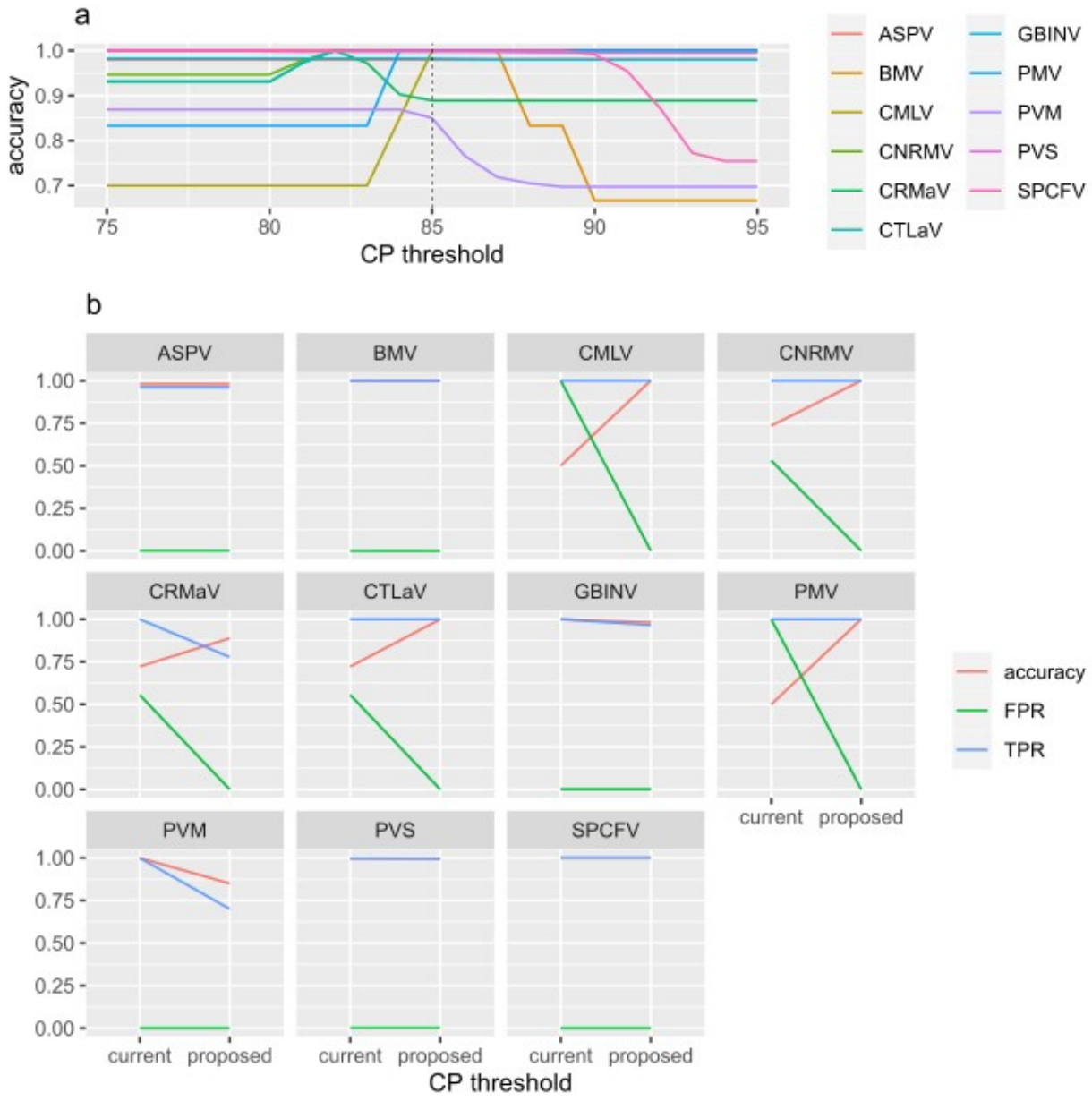
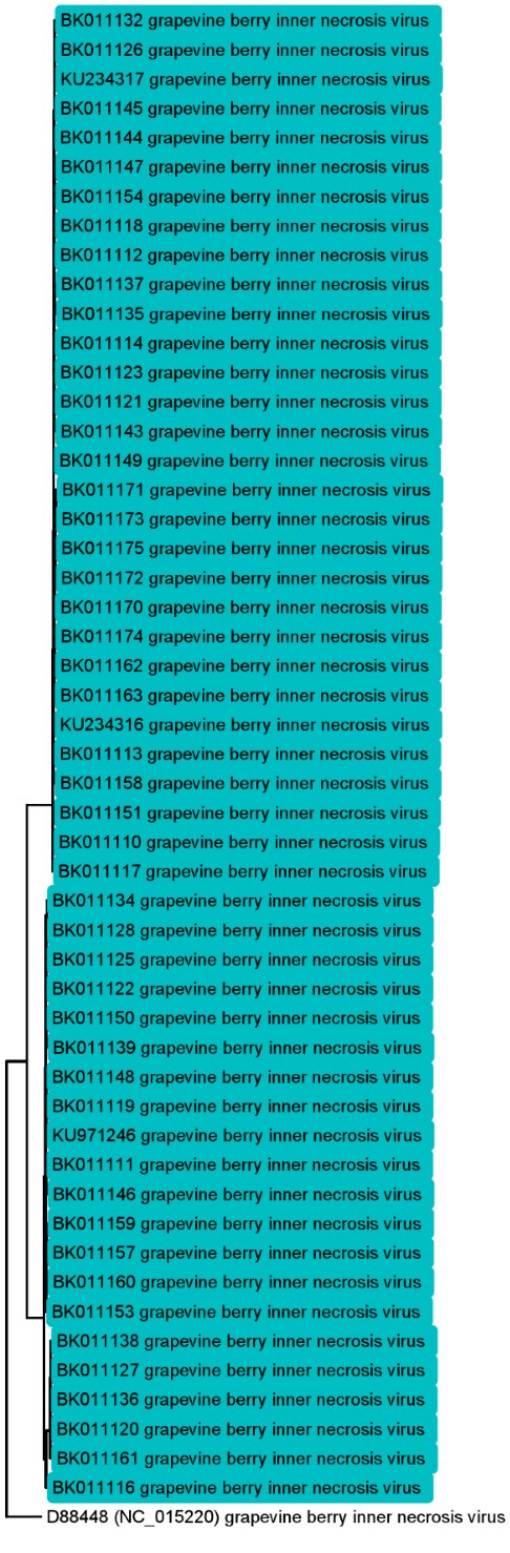
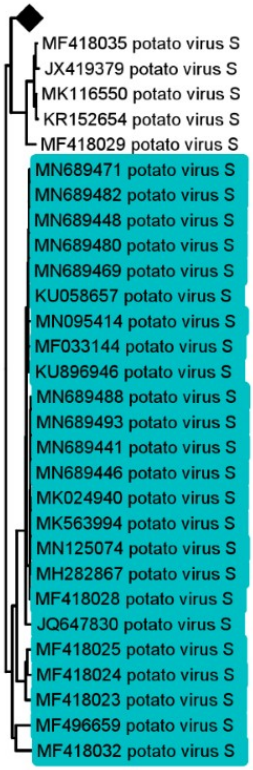


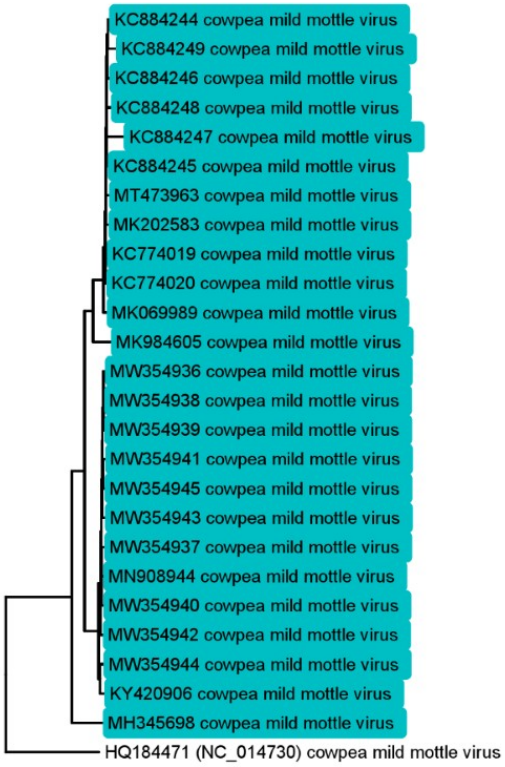
Figure 5. Statistics describing the accuracy of the different criteria of CP identity in distinguishing bona fide species. (**a**) Accuracy of the proposed criteria with varying CP

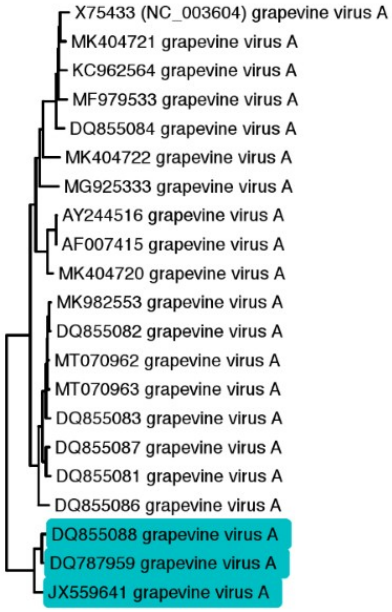
thresholds for selected species. (**b**) Accuracy, false positive rate (FPR) and true positive rate (TPR) of the current (80%) and proposed (85%) criteria of CP threshold.

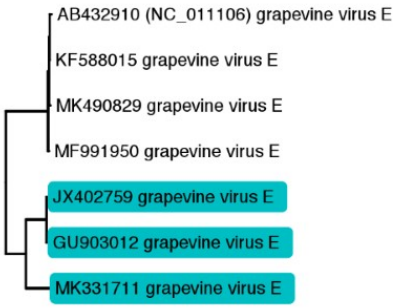
We also compared the current CP threshold (80%) with the proposed one with 85% (Figure 5b). A rise in accuracy and drop in FPR was observed for CMRaV, CNRMV, CTLaV, CMLV and PMV. A drop in TPR was noted for PVM, GBINV and CRMaV, indicating that some isolates of these species are below the species demarcation threshold. Noticeably, PVM cannot be split into monophyletic species that will only contain pairwise comparisons with each other below the proposed threshold, and thus, this species must be conserved (Figure S3).

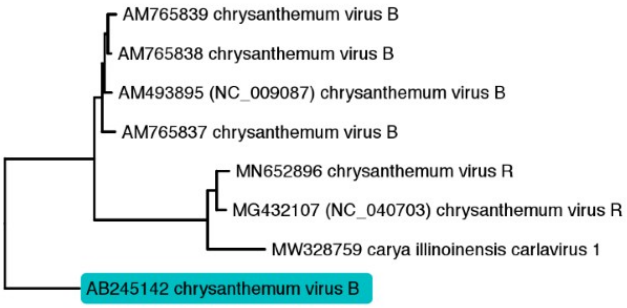

BK011132 grapevine berry inner necrosis virus
BK011126 grapevine berry inner necrosis virus
KU234317 grapevine berry inner necrosis virus
BK011145 grapevine berry inner necrosis virus
BK011144 grapevine berry inner necrosis virus
BK011147 grapevine berry inner necrosis virus
BK011154 grapevine berry inner necrosis virus
BK011118 grapevine berry inner necrosis virus
BK011112 grapevine berry inner necrosis virus
BK011137 grapevine berry inner necrosis virus
BK011135 grapevine berry inner necrosis virus
BK011114 grapevine berry inner necrosis virus
BK011123 grapevine berry inner necrosis virus
BK011121 grapevine berry inner necrosis virus
BK011143 grapevine berry inner necrosis virus
BK011149 grapevine berry inner necrosis virus
BK011171 grapevine berry inner necrosis virus
BK011173 grapevine berry inner necrosis virus
BK011175 grapevine berry inner necrosis virus
BK011172 grapevine berry inner necrosis virus
BK011170 grapevine berry inner necrosis virus
BK011174 grapevine berry inner necrosis virus
BK011162 grapevine berry inner necrosis virus
BK011163 grapevine berry inner necrosis virus
KU234316 grapevine berry inner necrosis virus
BK011113 grapevine berry inner necrosis virus
BK011158 grapevine berry inner necrosis virus
BK011151 grapevine berry inner necrosis virus
BK011110 grapevine berry inner necrosis virus
BK011117 grapevine berry inner necrosis virus
BK011134 grapevine berry inner necrosis virus
BK011128 grapevine berry inner necrosis virus
BK011125 grapevine berry inner necrosis virus
BK011122 grapevine berry inner necrosis virus
BK011150 grapevine berry inner necrosis virus
BK011139 grapevine berry inner necrosis virus
BK011148 grapevine berry inner necrosis virus
BK011119 grapevine berry inner necrosis virus
KU971246 grapevine berry inner necrosis virus
BK011111 grapevine berry inner necrosis virus
BK011146 grapevine berry inner necrosis virus
BK011159 grapevine berry inner necrosis virus
BK011157 grapevine berry inner necrosis virus
BK011160 grapevine berry inner necrosis virus
BK011153 grapevine berry inner necrosis virus
BK011138 grapevine berry inner necrosis virus
BK011127 grapevine berry inner necrosis virus
BK011136 grapevine berry inner necrosis virus
BK011120 grapevine berry inner necrosis virus
BK011161 grapevine berry inner necrosis virus
BK011116 grapevine berry inner necrosis virus
D88448 (NC_015220) grapevine berry inner necrosis virus

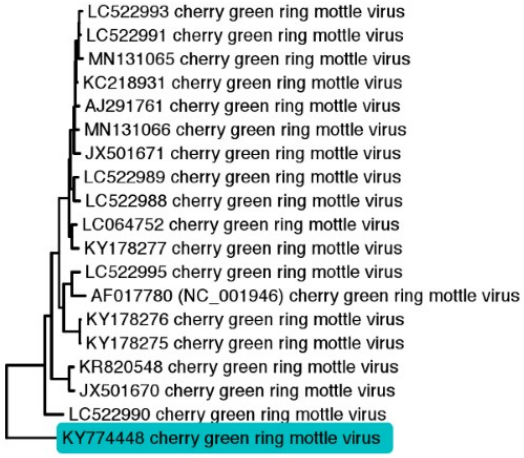

MF418035 potato virus S
JX419379 potato virus S
MK116550 potato virus S
KR152654 potato virus S
MF418029 potato virus S
MN689471 potato virus S
MN689482 potato virus S
MN689448 potato virus S
MN689480 potato virus S
MN689469 potato virus S
KU058657 potato virus S
MN095414 potato virus S
MF033144 potato virus S
KU896946 potato virus S
MN689488 potato virus S
MN689493 potato virus S
MN689441 potato virus S
MN689446 potato virus S
MK024940 potato virus S
MK563994 potato virus S
MN125074 potato virus S
MH282867 potato virus S
MF418028 potato virus S
JQ647830 potato virus S
MF418025 potato virus S
MF418024 potato virus S
MF418023 potato virus S
MF496659 potato virus S
MF418032 potato virus S

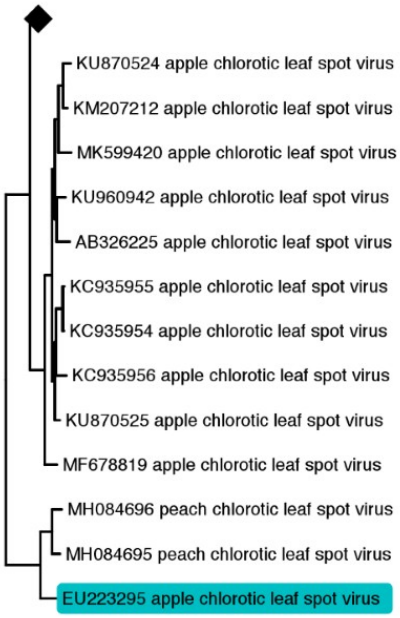

KC884244 cowpea mild mottle virus
KC884249 cowpea mild mottle virus
KC884246 cowpea mild mottle virus
KC884248 cowpea mild mottle virus
KC884247 cowpea mild mottle virus
KC884245 cowpea mild mottle virus
MT473963 cowpea mild mottle virus
MK202583 cowpea mild mottle virus
KC774019 cowpea mild mottle virus
KC774020 cowpea mild mottle virus
MK069989 cowpea mild mottle virus
MK984605 cowpea mild mottle virus
MW354936 cowpea mild mottle virus
MW354938 cowpea mild mottle virus
MW354939 cowpea mild mottle virus
MW354941 cowpea mild mottle virus
MW354945 cowpea mild mottle virus
MW354943 cowpea mild mottle virus
MW354937 cowpea mild mottle virus
MN908944 cowpea mild mottle virus
MW354940 cowpea mild mottle virus
MW354942 cowpea mild mottle virus
MW354944 cowpea mild mottle virus
KY420906 cowpea mild mottle virus
MH345698 cowpea mild mottle virus
HQ184471 (NC_014730) cowpea mild mottle virus

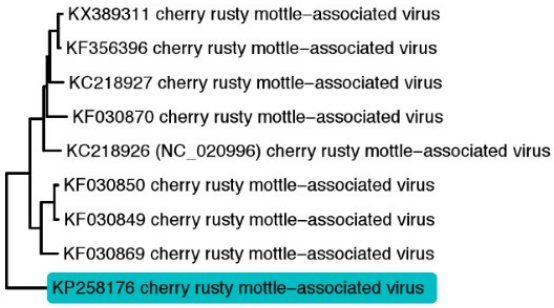

X75433 (NC_003604) grapevine virus A
MK404721 grapevine virus A
KC962564 grapevine virus A
MF979533 grapevine virus A
DQ855084 grapevine virus A
MK404722 grapevine virus A
MG925333 grapevine virus A
AY244516 grapevine virus A
AF007415 grapevine virus A
MK404720 grapevine virus A
MK982553 grapevine virus A
DQ855082 grapevine virus A
MT070962 grapevine virus A
MT070963 grapevine virus A
DQ855083 grapevine virus A
DQ855087 grapevine virus A
DQ855081 grapevine virus A
DQ855086 grapevine virus A
DQ855088 grapevine virus A
DQ787959 grapevine virus A
JX559641 grapevine virus A

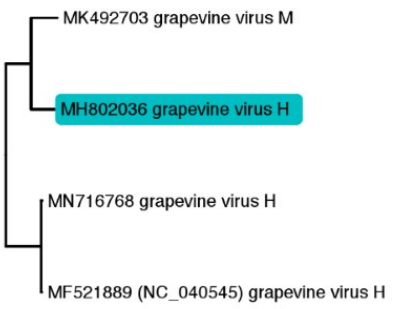

AB432910 (NC_011106) grapevine virus E
KF588015 grapevine virus E
MK490829 grapevine virus E
MF991950 grapevine virus E
JX402759 grapevine virus E
GU903012 grapevine virus E
MK331711 grapevine virus E

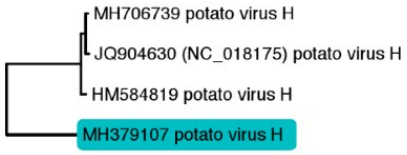

AM765839 chrysanthemum virus B
AM765838 chrysanthemum virus B
AM493895 (NC_009087) chrysanthemum virus B
AM765837 chrysanthemum virus B
MN652896 chrysanthemum virus R
MG432107 (NC_040703) chrysanthemum virus R
MW328759 carya illinoinensis carlavirus 1
AB245142 chrysanthemum virus B

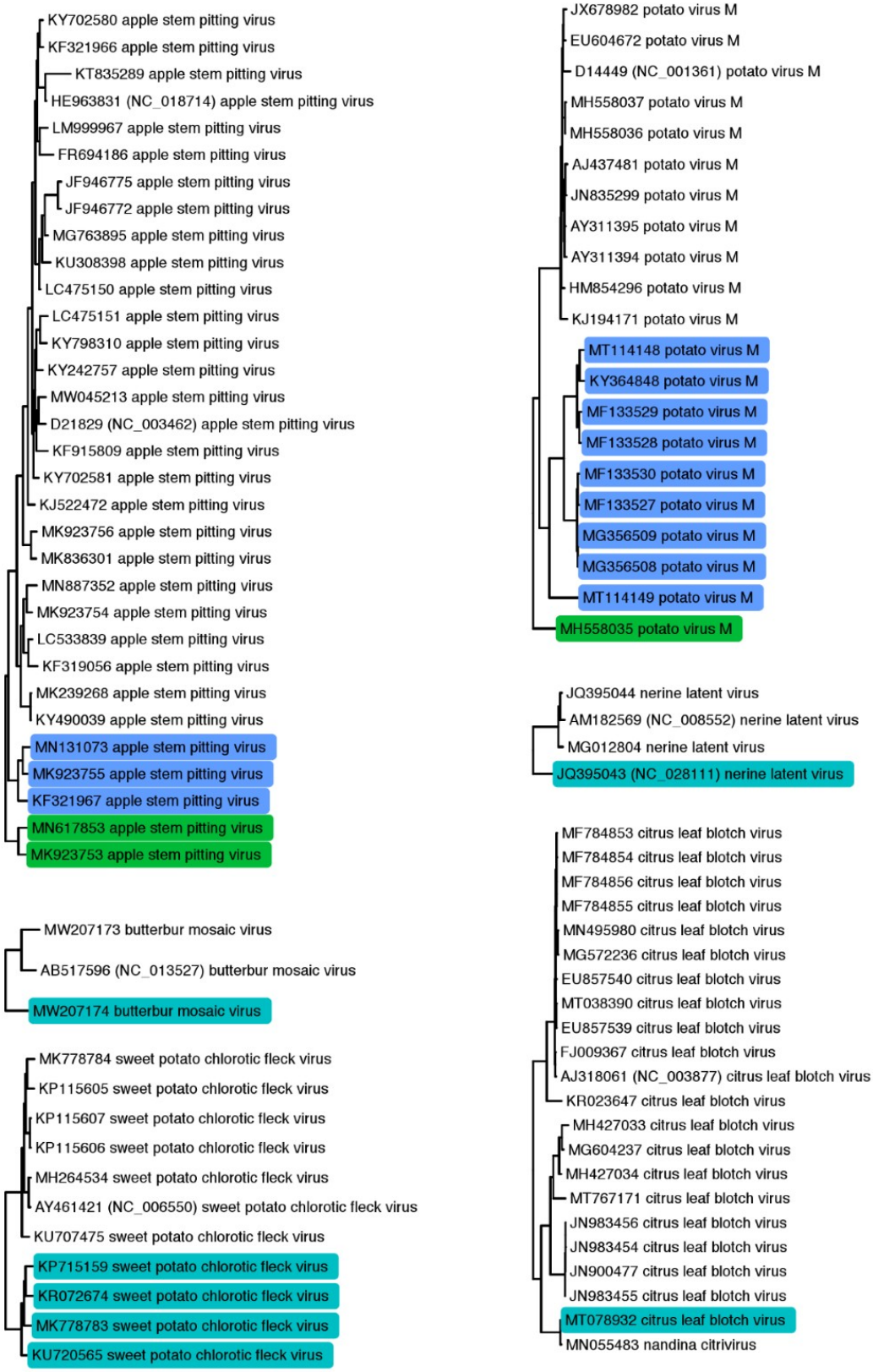

LC522993 cherry green ring mottle virus
LC522991 cherry green ring mottle virus
MN131065 cherry green ring mottle virus
KC218931 cherry green ring mottle virus
AJ291761 cherry green ring mottle virus
MN131066 cherry green ring mottle virus
JX501671 cherry green ring mottle virus
LC522989 cherry green ring mottle virus
LC522988 cherry green ring mottle virus
LC064752 cherry green ring mottle virus
KY178277 cherry green ring mottle virus
LC522995 cherry green ring mottle virus
AF017780 (NC_001946) cherry green ring mottle virus
KY178276 cherry green ring mottle virus
KY178275 cherry green ring mottle virus
KR820548 cherry green ring mottle virus
JX501670 cherry green ring mottle virus
LC522990 cherry green ring mottle virus
KY774448 cherry green ring mottle virus


KU870524 apple chlorotic leaf spot virus
KM207212 apple chlorotic leaf spot virus
MK599420 apple chlorotic leaf spot virus
KU960942 apple chlorotic leaf spot virus
AB326225 apple chlorotic leaf spot virus
KC935955 apple chlorotic leaf spot virus
KC935954 apple chlorotic leaf spot virus
KC935956 apple chlorotic leaf spot virus
KU870525 apple chlorotic leaf spot virus
MF678819 apple chlorotic leaf spot virus
MH084696 peach chlorotic leaf spot virus
MH084695 peach chlorotic leaf spot virus
EU223295 apple chlorotic leaf spot virus


KX389311 cherry rusty mottle–associated virus
KF356396 cherry rusty mottle–associated virus
KC218927 cherry rusty mottle–associated virus
KF030870 cherry rusty mottle–associated virus
KC218926 (NC_020996) cherry rusty mottle–associated virus
KF030850 cherry rusty mottle–associated virus
KF030849 cherry rusty mottle–associated virus
KF030869 cherry rusty mottle–associated virus
KP258176 cherry rusty mottle–associated virus


MK492703 grapevine virus M
MH802036 grapevine virus H
MN716768 grapevine virus H
MF521889 (NC_040545) grapevine virus H


MH706739 potato virus H
JQ904630 (NC_018175) potato virus H
HM584819 potato virus H
MH379107 potato virus H


KY702580 apple stem pitting virus
KF321966 apple stem pitting virus
KT835289 apple stem pitting virus
HE963831 (NC_018714) apple stem pitting virus
LM999967 apple stem pitting virus
FR694186 apple stem pitting virus
JF946775 apple stem pitting virus
JF946772 apple stem pitting virus
MG763895 apple stem pitting virus
KU308398 apple stem pitting virus
LC475150 apple stem pitting virus
LC475151 apple stem pitting virus
KY798310 apple stem pitting virus
KY242757 apple stem pitting virus
MW045213 apple stem pitting virus
D21829 (NC_003462) apple stem pitting virus
KF915809 apple stem pitting virus
KY702581 apple stem pitting virus
KJ522472 apple stem pitting virus
MK923756 apple stem pitting virus
MK836301 apple stem pitting virus
MN887352 apple stem pitting virus
MK923754 apple stem pitting virus
LC533839 apple stem pitting virus
KF319056 apple stem pitting virus
MK239268 apple stem pitting virus
KY490039 apple stem pitting virus
MN131073 apple stem pitting virus
MK923755 apple stem pitting virus
KF321967 apple stem pitting virus
MN617853 apple stem pitting virus
MK923753 apple stem pitting virus
JX678982 potato virus M
EU604672 potato virus M
D14449 (NC_001361) potato virus M
MH558037 potato virus M
MH558036 potato virus M
AJ437481 potato virus M
JN835299 potato virus M
AY311395 potato virus M
AY311394 potato virus M
HM854296 potato virus M
KJ194171 potato virus M
MT114148 potato virus M
KY364848 potato virus M
MF133529 potato virus M
MF133528 potato virus M
MF133530 potato virus M
MF133527 potato virus M
MG356509 potato virus M
MG356508 potato virus M
MT114149 potato virus M
MH558035 potato virus M
JQ395044 nerine latent virus
AM182569 (NC_008552) nerine latent virus
MG012804 nerine latent virus
JQ395043 (NC_028111) nerine latent virus
MW207173 butterbur mosaic virus
AB517596 (NC_013527) butterbur mosaic virus
MW207174 butterbur mosaic virus
MK778784 sweet potato chlorotic fleck virus
KP115605 sweet potato chlorotic fleck virus
KP115607 sweet potato chlorotic fleck virus
KP115606 sweet potato chlorotic fleck virus
MH264534 sweet potato chlorotic fleck virus
AY461421 (NC_006550) sweet potato chlorotic fleck virus
KU707475 sweet potato chlorotic fleck virus
KP715159 sweet potato chlorotic fleck virus
KR072674 sweet potato chlorotic fleck virus
MK778783 sweet potato chlorotic fleck virus
KU720565 sweet potato chlorotic fleck virus
MF784853 citrus leaf blotch virus
MF784854 citrus leaf blotch virus
MF784856 citrus leaf blotch virus
MF784855 citrus leaf blotch virus
MN495980 citrus leaf blotch virus
MG572236 citrus leaf blotch virus
EU857540 citrus leaf blotch virus
MT038390 citrus leaf blotch virus
EU857539 citrus leaf blotch virus
FJ009367 citrus leaf blotch virus
AJ318061 (NC_003877) citrus leaf blotch virus
KR023647 citrus leaf blotch virus
MH427033 citrus leaf blotch virus
MG604237 citrus leaf blotch virus
MH427034 citrus leaf blotch virus
MT767171 citrus leaf blotch virus
JN983456 citrus leaf blotch virus
JN983454 citrus leaf blotch virus
JN900477 citrus leaf blotch virus
JN983455 citrus leaf blotch virus
MT078932 citrus leaf blotch virus
MN055483 nandina citrivirus

Figs S1-3. Rep phylogenies of species that contains pairwise comparisons below the proposed threshold. Trees are subsets of the Rep phylogeny built previously. Sequences that would be split into a new species are highlighted. Collapsed nodes are shown as black diamonds.

3.5. Implementation of the proposed taxonomic criteria to the family Betaflexiviridae

We applied the proposed criteria to the rest of the members of *Betaflexiviridae* (Figure 6). One isolate of apple chlorotic leaf spot virus (ACLSV) should be reassigned as peach chlorotic leaf spot virus (PCLSV), and one isolate of grapevine virus H (GVH) should be reassigned as grapevine virus M (GVM); CGRMV, cowpea mild mosaic virus (CpMMV), CRMaV, chrysanthemum virus B (CVB), grapevine berry inner necrosis virus (GBINV), grapevine virus A (GVA), grapevine virus E (GVE) and potato virus H (PVH) should be split in two species. Citrus leaf blotch virus (CLBV) should be split in two species based on the proposed criteria, however, doing so, it would make CLBV paraphyletic (Figure S3). Likewise, potato virus M can be divided into three major monophyletic groups (Figure S3), however, the identity distributions of the group (represented by blue in Figure S3) extends to ~75% for the Rep. Thus, PVM cannot be divided while preserving monophyly. Evidence of positive selection in the Rep was found in four branches of CpMMV and in one isolate of PVS. The CP of one isolate of ASPV was subjected to positive selection. Evidence of positive selection was found in both Rep and CP of CVB.

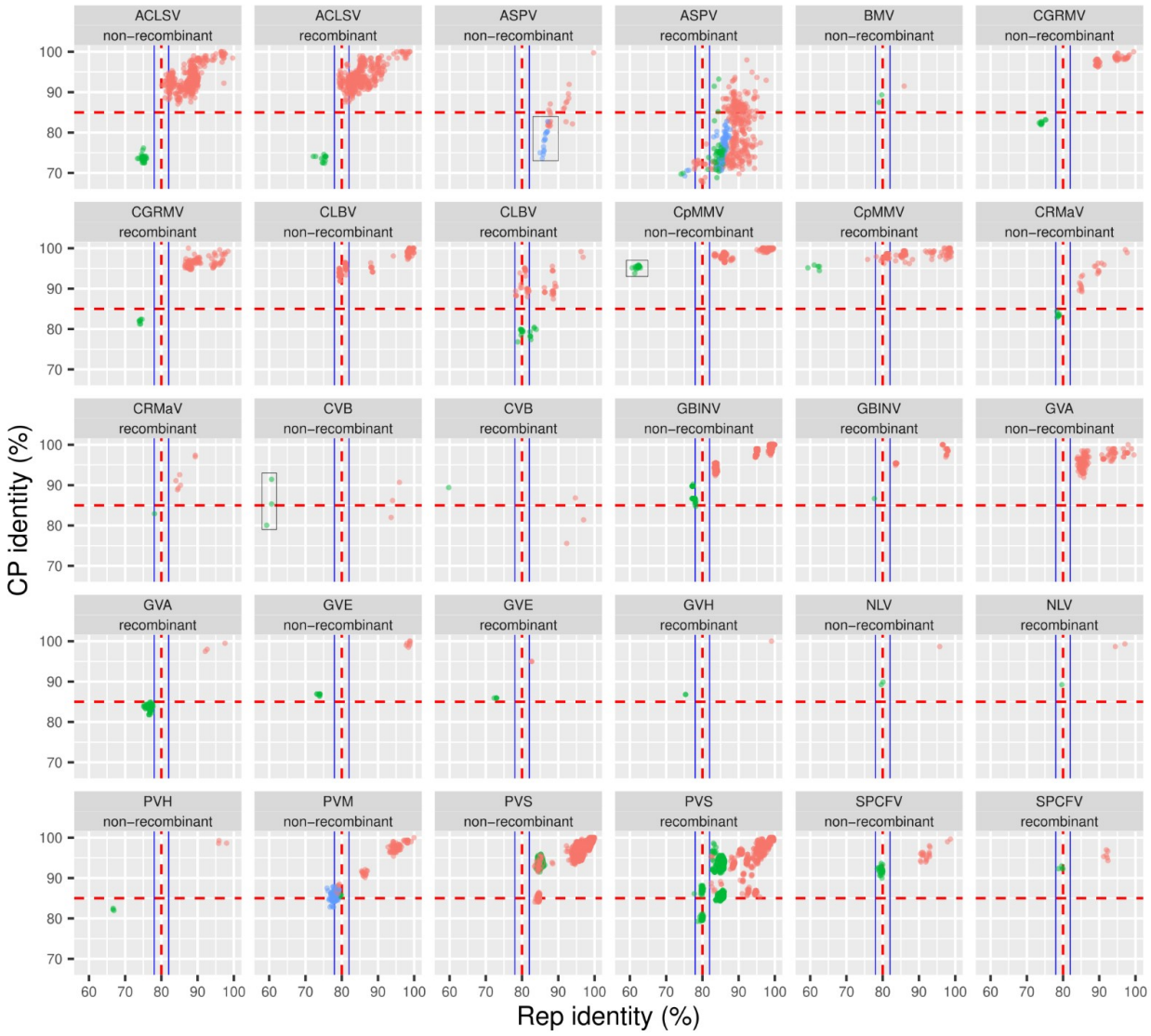


Figure 6. Analysis of species with pairwise identities below the proposed threshold. Species containing pairwise identities within quadrants II, IV and VI of Figure 3 were individually analyzed. Green and blue dots represent monophyletic groups that could be assigned to a new species (see Figure S1-3). Red dashed lines represent the proposed thresholds for the Rep and CP, and blue lines represent the borderline range of the Rep. Rectangles show cases where the discrepancy between the identities of the Rep and CP cannot be explained by recombination. These species were subjected to episodic selection analysis. ACSLV, apple chlorotic leaf spot virus; ASPV, apple stem-pitting virus; BMV, butterbur mosaic virus; CGRMV, cherry green ring mottle virus; CLBV, citrus leaf blotch virus; CpMMV, cowpea mild mosaic virus; CRMaV, cherry rusty mottle-associated virus; CVB, chrysanthemum virus B; GBINV, grapevine berry inner necrosis virus; GVA, grapevine virus A; GVE, grapevine virus E; GVH, grapevine virus H; NLV, nerine latent virus; PVH, potato virus H; PVM, potato virus M; PVS, potato virus S; SPCFC, sweet potato chlorotic fleck virus.

We also analyzed cases where the pairwise identities of the Rep are in the upper borderline range of 80-82%. CTLaV and ligustrum virus A (LVA) could be divided into two species since they contain monophyletic groups which have their Rep identities at the borderline range and their CP identities mostly below the CP threshold. We argue that if the Rep identity distribution is at the borderline range but exclusively at the upper limit of 80-82%, these sequences should be classified as one species. Some non-recombinant sequences showed Rep identities far to lower than expected (Fig 6, sequences within rectangles), and were subjected to further selection analyses. Evidence of positive selection was found for both the Rep and CP of CTLaV. The CP gene of CTLaV overlaps with ORF 5a, and thus, positive selection of this protein is most likely a false positive.

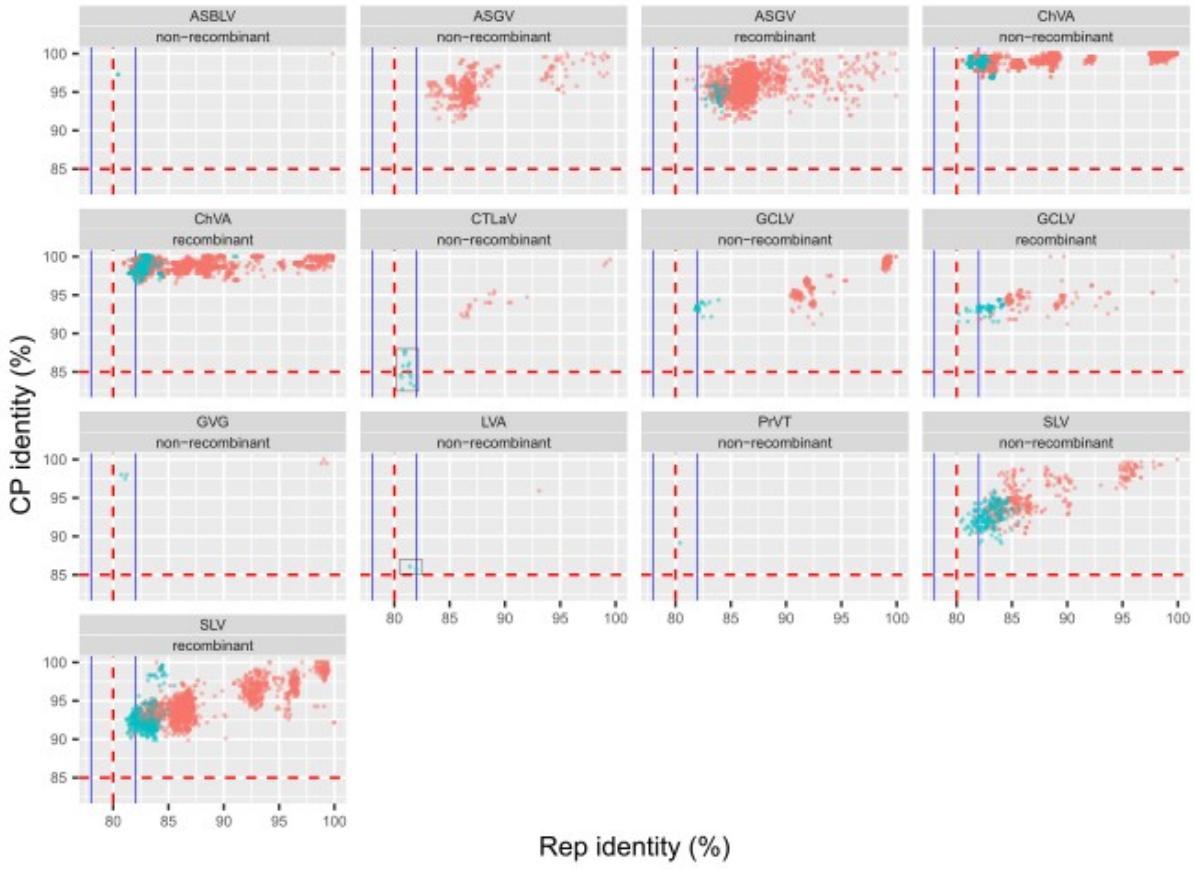


Figure 7. Analysis of species with pairwise identities of the Rep at the upper borderline range. Red dashed lines represent the proposed thresholds for the Rep and CP, and blue lines represent the borderline range of the Rep. Blue dots represent monophyletic groups that could be assigned to a new species. Rectangles show cases where the discrepancy between the identities of the Rep and CP cannot be explained by recombination. ASBLV, Actinidia seed-born latent virus; ASGV, apple stem grooving virus; ChVA, cherry virus A; CTLaV, cherry twisted leaf-associated virus; GCLV, garlic common latent virus; GVG, grapevine virus G; LVA, ligustrum virus A; PrVT, prunus virus T; SLV, shallot latent virus.

4. Conclusion

In this study, we propose an update on the current taxonomic criteria of the family *Betaflexiviridae* based on phylogenetic and phased identity analyses that should facilitate the assignment of novel species. A schematic representation of the proposed criteria and workflow is shown in Figure 8. Notably, these criteria are aimed at classifying viruses based on complete genomic sequences alone, although biological properties such as host range, vector specificity and mode of transmission should be used whenever this information is available. In this communication, we propose new taxonomic criteria that simplify the assignment of viruses in the family summarized below:

1. Members of the same species must be monophyletic in regard to the Rep phylogeny.
2. Recombination in betaflexiviruses makes the evolutionary history of CP different from Rep. Because of this characteristic, the demarcation criterium by Rep protein identity should be taken a priority.
3. The demarcation criterium for betaflexivirus species is 80% of aa identity of Rep protein. If the aa identity is in borderline (78-82), aa identity of CP would be applied. In this case, the threshold is 85%. If all pairwise comparisons of the Rep are at the borderline range but above the threshold (80-82%), we advise not to create a new species. Alternatively, a flexible threshold can be applied when accuracy statistics can be calculated, preferentially with the aid of biological properties to determine bona fide species.
4. Recombinant sequences which are below the thresholds of demarcation criteria can be new species if these isolates are established in nature. To conclude this, more recombinant sequences should be revealed.
5. Biological characteristics, if they are available, can be considered to differentiate as distinct species, especially for borderline cases.

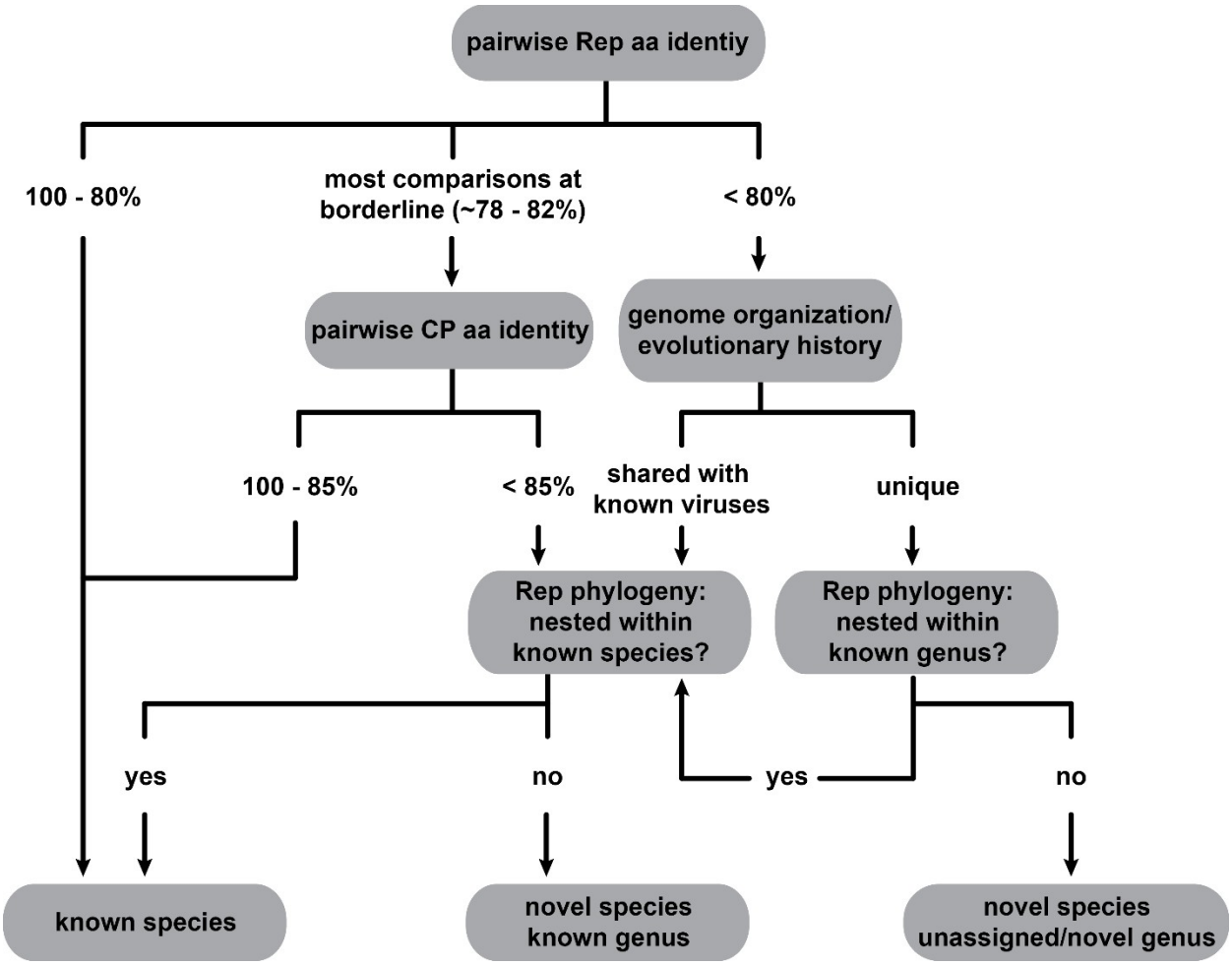


Figure 8. Proposed taxonomic criteria of the family Betaflexiviridae schematized as a decision tree to classify novel viruses based on genomic information.

Capítulo 5. Produção durante o doutorado

Neste Capítulo são listadas as colaborações e outros trabalhos que não fizeram parte do texto principal. Esses trabalhos apresentam a descrição e caracterização de novos vírus, detecção e monitoramento de vírus humanos e estudos sobre diversidade genética viral.

Souza, L.D.; Blawid, R.; Silva, J.M.F. Nagata, T. Human virome in nasopharynx and tracheal secretion samples. *Mem. Inst. Oswaldo Cruz* **2019**, 114.

Vasconcellos, A.F.; Silva, J.M.F. de Oliveira A.S.; Prado, P.S.; Nagata, T.; Resende, R.O. Genome sequences of chikungunya virus isolates circulating in midwestern Brazil. *Arch. Virol.* **2019**, 164, 1205-1208.

Camelo-García, V.M.; Molina, J.P.E.; Nagata, T.; Rezende, J.A.M.; Silva, J.M.F. Effect of rhizomania on red table-beet biomass production and molecular characterization of an isolate of Beet necrotic yellow vein virus from Brazil. *Eur. J. Plant Pathol.* **2019**, 154, 1021-1028.

Duarte, M.A.; Silva, J.M.F.; Brito, C.R.; Teixeira, D.S.; Melo, F.L.; Ribeiro, B.M.; Nagata, T.; Campos, F.S. Faecal virome analysis of wild animals from Brazil. *Viruses* **2019**, 11, 803.

Fariña, A.E.; Gorayeb, E.S.; Camelo-García, V.M.; Bonin, J.; Nagata, T.; Silva, J.M.F.; Bogo, A.; Rezende, J.A.M.; da Silva, F.N.; Kitajima, E.W. Molecular and biological characterization of a putative new sobemovirus infecting Physalis peruviana. *Arch. Virol.* **2019**, 164, 2805-2810.

Favara, G.M.; Camelo-García, V.M.; Nagata, T.; Silva, J.M.F.; Saito, M.; Rezende, J.A.M.; Salaroli, R.B.; Kitajima, E.W. Tobacco mild green mosaic virus found naturally infecting Nicotiana glauca in Brazil. *Australas. Plant Dis. Notes* **2019**, 14, 13.

Martínez, R.T.; de Almeida, M.M.S.; Rodriguez, R.; Cayetano, X.; de Oliveira, A.S.; Silva, J.M.F.; Melo, F.L.; Resende, R.O. Analyses of orthotospovirus populations and dispersion under different environmental conditions in Brazil and in the Dominican Republic. *Trop. Plant Pathol.* **2019**, 44, 511-8.

Favara, G.M.; Camelo-García, V.M.; Spadotti, D.M.A.; Silva, J.M.F.; Nagata, T.; Kitajima, E.W.; Rezende, J.A.M. First report of lettuce chlorosis virus infecting periwinkle in Brazil. *Plant Dis.* **2020**, 104, 1263.

Alves, T.M.; de Novaes, Q.S.; de Paula, A.; Camelo-García, V.M.; Nagata, T.; Silva, J.M.F.; Rezende, J.A.M.; Kitajima, E.W. Near-complete genome sequence and biological properties of an allexivirus found in Senna rizzinii in Brazil. *Arch. Virol.* **2020**, 165, 1463-1467.

Favara, G.M.; Camelo-García, V.M.; Silva, J.M.F., Silva, T.N.Z.; Mituti, T.; Nagata, T.; Kitajima, E.W.; Rezende, J.A.M. Biological and molecular characterization of isolates of catharanthus mosaic virus infecting *Mandevilla* sp. *Trop. Plant Pathol.* **2020**, 45, 461-465.

Souza, T.A.; Silva, J.M.; Nagata, T.; Martins, T.P.; Nakasu, E.Y.T.; Inoue-Nagata, A.K. A temporal diversity analysis of brazilian begomoviruses in tomato reveals a decrease in species richness between 2003 and 2016. *Front. Plant Sci.* **2020**, 11, 1201.

Maachi, A.; Nagata, T.; Silva, J.M.F. Date palm virus A: first plant virus found in date palm trees. *Virus Genes* **2020**, 56, 792-795.

Rego-Machado, C.M.; Nakasu, E.Y.T.; Silva, J.M.F.; Lucinda, N.; Nagata, T.; Inoue-Nagata, A.K. siRNA biogenesis and advances in topically applied dsRNA for controlling virus infections in tomato plants. *Sci. Rep.* **2020**, 10, 22277.

Silva, J.M.F.; Fajardo, T.V.M.; Al Rwahnih, M.; Nagata, T. First report of grapevine associated jivivirus 1 infecting grapevines in Brazil. *Plant Dis.* **2021**, 105, 514.

Capítulo 6. Discussão geral

Evolução viral é uma disciplina multifacetada que une várias áreas da biologia em diferentes escalas. Nesta tese, a evolução, diversidade e a interação com o hospedeiro a nível celular de vírus de RNA pertencentes a grupos distintos foram exploradas. Pressões seletivas agindo sobre os vírus à nível celular moldam a sua evolução e são ultimamente responsáveis pela vasta diversidade viral encontrada nesse grupo. Ademais, ao infectarem a mesma célula, vírus distintos interagem entre si, possibilitando que ocorra transferência horizontal de genes entre eles. Destacamos de cada capítulo pontos de extrema importância para o tema desta tese. Esses pontos nos permitem enxergar como eventos intracelulares são responsáveis pela evolução global dos vírus de RNA.

No capitulo 2, analisamos a resposta celular de células entéricas da mosca da fruta à infecção por dois vírus de RNA, *Drosophila melanogaster* Nora virus (DMelNV) e Thika virus (TV), e também analisamos os padrões de acumulação do RNA viral desses dois vírus em diferentes tipos e subtipos celulares. Células simultaneamente infectadas pelos dois vírus foram detectadas, sem que houvesse interferência de um vírus na probabilidade de infecção pelo outro. A via de reposta a choque térmico antiviral é ativada em células infectadas pelo DMelNV ou TV de maneira dependente de tipo celular e vírus. Interessantemente, alguns genes relacionados à resposta a choque térmico foram menos expressos em células infectadas, e podem representar mecanismos que evoluíram em ambos DMelNV e TV para inibir a resposta antiviral da mosca. Esses resultados sugerem que os TV e DMelNV estejam sob diferentes pressões seletivas que é dependente de tipo celular, fato relevante para a adaptação, diversificação e evolução desses vírus.

Demostramos também que o DMelNV e TV apresentam padrões distintos de acumulação do RNA viral, e que em ambos os casos há influência do tipo ou subtipo celular tanto no acumulo do RNA viral quanto na susceptibilidade à infecção. A diferente susceptibilidade de diversos tipos celulares à ambas infecção e replicação viral podem influenciar a evolução e adaptação viral devido à diversidade genômica associada ao tamanho populacional, onde variantes de maior aptidão teriam maior chance de emergir em tipos celulares mais suscetíveis. Similarmente, a taxa de mutação do vírus da estomatite vesicular (*vesicular stomatitis virus*; VSV) é menor em células de insetos do que em células de mamíferos, e consequentemente, o vírus evolui mais devagar no inseto vetor (Combe & Sanjuan, 2014). Entretanto, a diversidade viral associada a diferentes tipos celulares ainda deve ser estudada a fundo para melhor entendermos o grau em que tipos celulares influenciam a evolução viral. Caso essa influência seja grande, é de se esperar que a distribuição de mutações da população viral seja significantemente distinta entre os tipos celulares.

No Capítulo 3 caracterizamos eventos de rearranjo a nível de espécies em dois tospovírus proximamente relacionados, e mostramos que o segmento M do tomato chlorotic spot virus (TCSV) foi substituído pelo segmento M do groundnut ringspot virus (GRSV). Além da diversidade genética do segmento M compartilhado pelos TCSV e GRSV ser menor que a dos segmentos S e L desses vírus, esperamos que o segmento M esteja sujeito a diferentes pressões evolutivas dependendo do vírus no qual ele se encontra. Análises filogenéticas bayesianas

sugerem que o tomato spotted wilt virus (TSWV) tenha se diversificado recentemente dentro dos últimos 260 anos. A estimativa da diversificação dos TCSV e GRSV foi prejudicada pelo baixo número de isolados datados disponíveis para a análise, entretanto, foi possível estimar o tempo para o ancestral comum mais recente (*time to most recent common ancestor*; TMRCA) dos segmentos S, M e L. O TMRCA do segmento M entre os dois vírus é significantemente mais recente do que o dos segmentos S e L, corroborando com a hipótese de que o segmento M do TCSV tenha sido perdido ou permanece não sequenciado.

No capítulo 4, eventos de recombinação na família *Betaflexiviridae* (ordem *Tymovirales*) foram analisados, onde a emergência de novas espécies e de gêneros na família estão ligadas à recombinação. Entretanto, apenas eventos de recombinação envolvendo a replicase (Rep) e capa proteíca (CP) foram analisados. O impacto que eventos de recombinação exercem na família também são refletidos na estrutura genômica dos vírus que a compõem. As subfamílias *Tri-* e *Quinvirinae* possuem proteínas de movimento distintas, e alguns gêneros codificam uma proteína de ligação a ácido nucleico (*nucleic acid-binding protein*; NBP). Ademais, somente vírus pertencentes ao gênero *Vitivirus* codificam uma proteína 20k de função desconhecida, e possivelmente, alguns vitivírus adquiriram um domínio de alquilação B (*alkilation B*; AlkB) de closterovírus (Maree et al., 2020; Dolja et al., 2017). A CP dos vírus pertencentes aos gêneros *Citrivirus* e *Wamavirus*, da subfamília *Trivirinae*, foi obtida por recombinação de vírus da subfamília *Quinvirinae*, mais provavelmente pertencentes ao gênero *Foveavirus*. Adicionalmente, hipotetizamos que pressões evolutivas distintas agindo sobre a Rep e CP são responsáveis pela discordância nas identidades aos pares entre essas proteínas em diferentes níveis taxonômicos. A nível de espécies a CP é mais conservada do que a Rep, mas passa a ter identidades aos pares comparáveis com aqueles da Rep a nível de gênero.

Em suma, fenômenos intracelulares moldam a evolução dos vírus de RNA, ocasionando na alta diversidade global dessas entidades. A evolução em pequena escala dos TV e DMelNV é diretamente ligada às interações desses vírus com o hospedeiro a nível celular, e ao longo do tempo, essas interações serão responsáveis pela continua diversificação e evolução desses vírus. A diversidade e evolução a nível de espécies do TCSV e GRSV e a nível de família dos vírus pertencentes à família *Betaflexiviridae* é visivelmente influenciada pela transferência horizontal de genes, fenômenos intracelulares, salientando como a interação entre vírus que infectam a mesma célula influencia a evolução dos mesmos.

Referência bibliográfica

Adams, M.J.; Candresse, T.; Hammond, J.; Kreuze, J.F.; Martelli, G.P.; Namba, S.; Pearson, M.N.; Ryu, K.H.; Saldarelli, P.; Yoshikawa, N. Betafexiviridae. In *Virus taxonomy, ninth report of the international committee on taxonomy of viruses*; King, A.M.Q., Adams, M.J., Carstens, E.B., Lefkowitz, E.J., Eds.; Elsevier: San Diego, CA, USA, 2012; pp. 920–941.

Adkins, S.; Quadt, R.; Choi, T.J.; Ahlquist, P.; German, T. An RNA-dependent RNA polymerase activity associated with virions of tomato spotted wilt virus, a plant- and insect-infecting bunyavirus. *Virology* **1995**, 207, 308–311.

Ahlmann-Eltze, C.; Huber, W. glmGamPoi: Fitting Gamma-Poisson generalized linear models on single cell count data. *Bioinformatics* **2020**, 36, 5701–5702.

Ahmed, H.; Howton, T.C.; Sun, Y.; Weinberger, N.; Belkhadir, Y.; Mukhtar, M.S. Network biology discovers pathogen contact points in host protein-protein interactomes. *Nat. Commun.* **2018**, 9, 2312.

Aibar, S.; González-Blas, C.B.; Moerman, T.; Imrichova, H.; Hulselmans, G.; Rambow, F.; Marine, J.C.; Geurts, P.; Aerts, J.; van den Oord, J.; et al. SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* **2017**, 14, 1083–1086.

Alabi, O.J.; McBride, S.; Appel, D.N.; Al Rwahnih, M.; Pontasch, F.M. Grapevine virus M, a novel vitivirus discovered in the American hybrid bunch grape cultivar Blanc du Bois in Texas. *Arch. Virol.* **2019**, 164, 1739-1741.

Alabi, O.J.; Al Rwahnih, M.; Mekuria, T.A.; Naidu, R.A. Genetic diversity of Grapevine virus A in Washington and California vineyards. *Phytopathology* **2014**, 104, 548-560.

Allan, A.K.; Du, J.; Davies, S.A.; Dow, J.A. Genome-wide survey of V-ATPase genes in Drosophila reveals a conserved renal phenotype for lethal alleles. *Physiol. Genom.* **2005**, 22, 128–138.

Almeida, M.M.S.; Orilio, A.F.; Melo, F.L.; Rodriguez, R.; Feliz, A.; Cayetano, X.; Martinez, R.T.; Resende, R.O. The First Report of Tomato chlorotic spot virus (TCSV) Infecting Long Beans and Chili Peppers in the Dominican Republic. *Plant Dis.* **2014**, 98, 1285.

Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, 215, 403–410.

Ambrose, R.L.; Lander, G.C.; Maaty, W.S.; Bothner, B.; Johnson, J.E.; Johnson, K.N. Drosophila A virus is an unusual RNA virus with a T = 3 icosahedral core and permuted RNA-dependent RNA polymerase. *J. Gen. Virol.* **2009**, 90, 2191–2200.

de Avila, A.C.; de Haan, P.; Kormelink, R.; Resende, R.O.; Goldbach, R.W.; Peters, D. Classification of tospoviruses based on phylogeny of nucleoprotein gene sequences. *J. Gen. Virol.* **1993**, 74, 153–159.

de Avila, A.C.; de Haan, P.; Smeets, M.L.L.; Resende, R.D.; Kormelink, R.; Kitajima, E.W.; Goldbach, R.W.; Peters, D. Distinct levels of relationships between tospovirus isolates. *Arch. Virol.* **1993**, 128, 211–227.

Barr, J.N.; Fearns, R. How RNA viruses maintain their genome integrity. *J. Gen. Virol.* **2010**, 91, 1373-1387.

Basler, C.F.; Mikulasova, A.; Martinez-Sobrido, L.; Paragas, J.; Mühlberger, E.; Bray, M.; Klenk, H.D.; Palese, P.; García-Sastre, A. The Ebola virus VP35 protein inhibits activation of interferon regulatory factor 3. *J. Virol.* **2003**, 77, 7945–7956.

Beehler-Evans, R.; Micchelli, C.A. Generation of enteroendocrine cell diversity in midgut stem cell lineages. *Development* **2015**, 142, 654–664.

Biek, R.; Pybus, O.G.; Lloyd-Smith, J.O.; Didelot, X. Measurably evolving pathogens in the genomic era. *Trends Ecol. Evol.* **2015**, 30, 306-313.

Bouckaert, R.; Heled, J.; Kühnert, D.; Vaughan, T.; Wu, C.H.; Xie, D.; Suchard, M.A.; Rambaut, A.; Drummond, A.J. BEAST 2: A software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **2014**, 10, e1003537.

Bouckaert, R.R.; Drummond, A.J. bModelTest: Bayesian phylogenetic site model averaging and model comparison. *BMC Evol. Biol.* **2017**, 17, 42.

de Breuil, S.; Cañizares, J.; Blanca, J.M.; Bejerman, N.; Trucco, V.; Giolitti, F.; Ziarsolo, P.; Lenardon, S. Analysis of the coding-complete genomic sequence of groundnut ringspot virus suggests a common ancestor with tomato chlorotic spot virus. *Arch. Virol.* **2016**, 161, 2311–2316.

Brewer, E.; Cao, M.; Gutierrez, B.; Bateman, M.; Li, R. Discovery and molecular characterization of a novel trichovirus infecting sweet cherry. Virus genes 2020, 56, 380–385.

Buchfink, B.; Xie, C.; Huson, D.H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **2015**, 12, 59-60.

Caballero, I.S.; Honko, A.N.; Gire, S.K.; Winnicki, S.M.; Melé, M.; Gerhardinger, C.; Lin, A.E.; Rinn, J.L.; Sabeti, P.C.; Hensley, L.E.; et al. In vivo Ebola virus infection leads to a strong innate response in circulating immune cells. *BMC Genom.* **2016**, 17, 707.

Cao, M.; Li, P.; Zhang, S.; Yang, F.; Zhou, Y.; Wang, X.; Li, R.; Li, Z. Molecular characterization of a novel citrivirus from citrus using next-generation sequencing. *Arch. Virol.* **2018**, 163, 3479-3482.

Chen, G.; Ning, B.; Shi, T. Single-cell RNA-seq technologies and related computational data analysis. *Front. Genet.* **2019**, 10, 317.

Cheng, G.; Liu, L.; Wang, P.; Zhang, Y.; Zhao, Y.O.; Colpitts, T.M.; Feitosa, F.; Anderson, J.F.; Fikrig, E. An in vivo transfection approach elucidates a role for Aedes aegypti thioester-containing proteins in flaviviral infection. *PLoS ONE* **2011**, 6, e22786.

Cho, D.Y.; Kim, Y.A.; Przytycka, T.M. Network biology approach to complex diseases. *PLoS Comput. Biol.* **2012**, 8, e1002820.

Chua, R.L.; Lukassen, S.; Trump, S.; Hennig, B.P.; Wendisch, D.; Pott, F.; Debnath, O.; Thürmann, L.; Kurth, F.; Völker, M.T.; Kazmierski, J. COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis. *Nat. Biotechnol.* **2020**, 38, 970-979.

Combe, M.; Sanjuan, R. Variation in RNA virus mutation rates across host cells. *PLoS Pathog.* **2014**, 10, e1003855.

Cordes, E.J.; Licking-Murray, K.D.; Carlson, K.A. Differential gene expression related to Nora virus infection of *Drosophila melanogaster*. *Virus Res.* **2013**, 175, 95–100.

Cristinelli, S.; Ciuffi, A. The use of single-cell RNA-Seq to understand virus–host interactions. *Curr. Opin. Virol.* **2018**, 29, 39-50.

Csardi, G.; Nepusz, T. The igraph software package for complex network research. *Int. J. Complex Syst.* **2006**, 1695, 1–9.

Darriba, D.; Taboada, G.L.; Doallo, R.; Posada, D. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **2011**, 27, 1164–1165.

Diaz-Lara, A.; Mollov, D.; Golino, D.; Al Rwahnih, M.; Complete genome sequence of rose virus A, the first carlavirus identified in rose. *Arch. Virol.* **2020**, 165, 241-244.

Diaz-Lara, A.; Mollov, D.; Golino, D.; Al Rwahnih, M. Detection and characterization of a second carlavirus in Rosa sp. *Arch. Virol.* **2021**, 166, 321-323.

Ding, X.B.; Jin, J.; Tao, Y.T.; Guo, W.P.; Ruan, L.; Yang, Q.L.; Chen, P.C.; Yao, H.; Zhang, H.B.; Chen, X. Predicted Drosophila interactome resource and web tool for functional interpretation of differentially expressed genes. *Database* **2020**, 2020, baaa005.

Dolja, V.V.; Meng, B.; Martelli, G.P. Evolutionary aspects of grapevine virology. In *Grapevine viruses: molecular biology, diagnostics and management*; Meng, B., Martelli, G.P., Golino D.A., Fuchs, M., Eds.; Springer, Cham, 2017, pp. 659-688.

Drummond, A.J.; Ho, S.Y.; Phillips, M.J.; Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **2006**, 4, e88.

Drummond, A.J.; Rambaut, A.; Shapiro, B.; Pybus, O.G. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol.* **2005**, 22, 1185–1192.

Dubreuil, R.R. Copper cells and stomach acid secretion in the Drosophila midgut. Int. J. Biochem. *Cell Biol.* **2004**, 36, 742–752.

Edgar R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids Res.* **2004**, 32, 1792-1797.

Firth, A.E. Mapping overlapping functional elements embedded within the protein-coding regions of RNA viruses. *Nucleic Acids Res.* **2014**, 42, 12425–12439.

Fleming, S.J.; Marioni, J.C.; Babadi, M. CellBender remove-background: A deep generative model for unsupervised removal of background noise from scRNA-seq datasets. *bioRxiv* **2019**, 791699.

Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, 28, 3150-3152.

Furness, J.B.; Rivera, L.R.; Cho, H.J.; Bravo, D.M.; Callaghan, B. The gut as a sensory organ. *Nat. Rev. Gastroenterol. Hepatol.* **2013**, 10, 729–740.

Gazel, M.; Roumi, V.; Ördek, K.; Maclot, F.; Massart, S.; Çağlayan, K. Identification and molecular characterization of a novel foveavirus from Rubus spp. in Turkey. *Virus Res.* **2020**, 286, 198078.

Gilbert, W. Origin of life: The RNA world. *Nature* **1986**, 319, 618.

Gribble, F.M.; Reimann, F. Function and mechanisms of enteroendocrine cells and gut hormones in metabolism. *Nat. Rev. Endocrinol.* **2019**, 15, 226–237.

Goh, C.J.; Hahn, Y. Identification of a novel member of the family Betaflexiviridae from the hallucinogenic plant Salvia divinorum. *Acta Virol.* **2019**, 63, 373-379.

Goh, C.J.; Park, D.; Hahn, Y. A novel tepovirus, Agave virus T, identified by the analysis of the transcriptome data of blue agave (Agave tequilana). *Acta Virol.* **2021**, 65, 68-71.

Goh, C.J.; Park, D.; Kim, H.; Sebastiani, F.; Hahn, Y. Novel Divavirus (the family *Betaflexiviridae*) and Mitovirus (the family *Narnaviridae*) species identified in basil (Ocimum basilicum). *Acta Virol.* **2018**, 62, 304-309.

Guindon, S.; Dufayard, J.F.; Lefort, V.; Anisimova, M.; Hordijk, W.; Gascuel, O. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **2010**, 59, 307–321.

Guo, X.; Yin, C.; Yang, F.; Zhang, Y.; Huang, H.; Wang, J.; Deng, B.; Cai, T.; Rao, Y.; Xi, R. The cellular diversity and transcription factor code of Drosophila enteroendocrine cells. *Cell Rep.* **2019**, 29, 4172–4185.

Gupta, M.; Mahanty, S.; Ahmed, R.; Rollin, P.E. Monocyte-derived human macrophages and peripheral blood mononuclear cells infected with Ebola virus secrete MIP-1α and TNF-α and inhibit poly-IC-induced IFN-α in vitro. *Virology* **2001**, 284, 20–25.

Habayeb, M.S.; Cantera, R.; Casanova, G.; Ekström, J.O.; Albright, S.; Hultmark, D. The Drosophila Nora virus is an enteric virus, transmitted via feces. *J. Invertebr. Pathol.* **2009**, 101, 29–33.

Habayeb, M.S.; Ekengren, S.K.; Hultmark, D. Nora virus, a persistent virus in Drosophila, defines a new picorna-like virus family. *J. Gen. Virol.* **2006**, 87, 3045–3051.

Hafemeister, C.; Satija, R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. *Genome Biol.* **2019**, 20, 296.

Harcourt, B.H.; Sanchez, A.; Offermann, M.K. Ebola virus selectively inhibits responses to interferons, but not to interleukin-1β, in endothelial cells. *J. Virol.* **1999**, 73, 3491–3496

Hein, M.Y.; Weissman, J.S. Functional single-cell genomics of human cytomegalovirus infection. *Nat. Biotechnol.* **2021**, https://doi.org/10.1038/s41587-021-01059-3.

Heldt, F.S.; Kupke, S.Y.; Dorl, S.; Reichl, U.; Frensing, T. Single-cell analysis and stochastic modelling unveil large cell-to-cell variability in influenza A virus infection. *Nat. Commun.* **2015**, 6, 8938.

Hendrix, R.W.; Smith, M.C.M.; Burns, R.N.; Ford, M.E.; Hatfull, G.F. Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proc. Natl. Acad. Sci. USA* **1999**, 96, 2192–97.

Holmes, E.C.; Duchêne, S. Can sequence phylogenies safely infer the origin of the global virome? *mBio* **2019**, 10, e00289-19.

Hull, R.; Rima, B. Virus taxonomy and classification: naming of virus species. *Arch. Virol.* **2020**, 165, 2733-2736.

Hung, R.J.; Hu, Y.; Kirchner, R.; Liu, Y.; Xu, C.; Comjean, A.; Tattikota, S.G.; Li, F.; Song, W.; Sui, S.H.; et al. A cell atlas of the adult Drosophila midgut. *Proc. Natl. Acad. Sci. USA* **2020**, 21, 1514–1523.

Huszar, T.; Imler, J.L. Drosophila viruses and the study of antiviral host-defense. *Adv. Virus Res.* **2008**, 72, 227–265.

Huynh-Thu, V.A.; Irrthum, A.; Wehenkel, L.; Geurts, P. Inferring regulatory networks from expression data using tree-based methods. *PLoS ONE* **2010**, 5, e12776.

James, D.; Varga, A.; Croft, H.; Rast, H.; Thompson, D.; Hayes, S. Molecular characterization, phylogenetic relationships, and specific detection of Peach mosaic virus. *Phytopathology* **2006**, 96, 137-44.

Jassal, B.; Matthews, L.; Viteri, G.; Gong, C.; Lorente, P.; Fabregat, A.; Sidiropoulos, K.; Cook, J.; Gillespie, M.; Haw, R.; et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **2020**, 48, D498–D503.

Jeong, H.; Mason, S.P.; Barabási, A.L.; Oltvai, Z.N. Lethality and centrality in protein networks. *Nature* **2001**, 411, 41–42.

Jones, P.; Binns, D.; Chang, H.Y.; Fraser, M.; Li, W.; McAnulla, C.; McWilliam, H.; Maslen, J.; Mitchell, A.; Nuka, G.; Pesseat, S.; et al. InterProScan 5: genome-scale protein function classification. *Bioinformatics.* **2014**, 30, 1236-1240.

Kash, J.C.; Mühlberger, E.; Carter, V.; Grosch, M.; Perwitasari, O.; Proll, S.C.; Thomas, M.J.; Weber, F.; Klenk, H.D.; Katze, M.G. Global suppression of the host antiviral response by Ebola-

and Marburgviruses: Increased antagonism of the type I interferon response is associated with enhanced virulence. *J. Virol.* **2006**, 80, 3009–3020.

Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **2013**, 30, 772-780.

Kemp, C.; Imler, J.L. Antiviral immunity in Drosophila. *Curr. Opin. Immunol.* **2009**, 21, 3–9.

Koonin, E.V.; Dolja, V.V.; Krupovic, M.; Varsani, A.; Wolf, Y.I.; Yutin, N.; Zerbini, F.M.; Kuhn, J.H. Global organization and proposed megataxonomy of the virus world. *Microbiol. Mol. Biol. Rev.* **2020**, 84, e00061-19.

Koonin, E.V.; Senkevich, T.G.; Dolja, V.V. The ancient Virus World and evolution of cells. *Biol. Direct* **2006**, 1, 1-27.

Kormelink, R.; Storms, M.; Vanlent, J.; Peters, D.; Goldbach, R. Expression and subcellular location of the NSM protein of tomato spotted wilt virus (TSWV), a putative viral movement protein. *Virology* **1994**, 200, 56–65.

Kotliar, D.; Lin, A.E.; Logue, J.; Hughes, T.K.; Khoury, N.M.; Raju, S.S.; Wadsworth, M.H., 2nd; Chen, H.; Kurtz, J.R.; Dighero-Kemp, B.; et al. Single-cell profiling of Ebola virus disease in vivo reveals viral and host dynamics. *Cell* **2020**, 183, 1383–1401.

Krupovic, M.; Koonin, E.V. Multiple origins of viral capsid proteins from cellular ancestors. *Proc. Natl. Acad. Sci. USA* **2017**, 114, E2401-E2410.

Li, D.; Luo, R.; Liu, C.M.; Leung, C.M.; Ting, H.F.; Sadakane, K.; Yamashita, H.; Lam, T.W. MEGAHIT v1. 0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **2016**, 102, 3-11.

Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv* **2013**, arXiv:1303.3997.

Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**, 25, 2078–2079.

Li, M.M.; MacDonald, M.R.; Rice, C.M. To translate, or not to translate: Viral and host mRNA regulation by interferon-stimulated genes. *Trends Cell Biol.* **2015**, 25, 320–329.

Li, Z.; Wang, H.; Zhao, R.; Zhang, Z.; Xia, Z.; Zhai, J.; Huang, X. Complete genome sequence of a novel capillovirus infecting Hevea brasiliensis in China. *Arch. Virol.* **2020**, 165, 249-252.

Liu, Q.; Yang, L.; Xuan, Z.; Wu, J.; Qiu, Y.; Zhang, S.; Wu, D.; Zhou, C.; Cao, M. Complete nucleotide sequence of loquat virus A, a member of the family Betaflexiviridae with a novel genome organization. *Arch. Virol.* **2020**, 165, 223-226.

Liu, X.; Speranza, E.; Muñoz-Fontela, C.; Haldenby, S.; Rickett, N.Y.; Garcia-Dorival, I.; Fang, Y.; Hall, Y.; Zekeng, E.G.; Lüdtke, A.; et al. Transcriptomic signatures differentiate survival from fatal outcomes in humans infected with Ebola virus. *Genome Biol.* **2017**, 18, 4.

Librado, P.; Rozas, J. DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* **2009**, 25, 1451–1452.

Liu, H.; Wu, L.; Zheng, L.; Cao, M.; Li, R. Characterization of three new viruses of the family Betaflexiviridae associated with camellia ringspot disease. *Virus Res.* **2019**, 272, 197668.

Londoño, A.; Capobianco, H.; Zhang, S.; Polston, J.E. First record of Tomato chlorotic spot virus in the USA. Trop. *Plant Pathol.* **2012**, 37, 333–338.

Lopez, W.; Page, A.M.; Carlson, D.J.; Ericson, B.L.; Cserhati, M.F.; Guda, C.; Carlson, K.A. Analysis of immune-related genes during Nora virus infection of Drosophila melanogaster using next generation sequencing. *AIMS Microbiol.* **2018**, 4, 123–139.

Lovato, F.A.; Nagata, T.; de Oliveira Resende, R.; de Avila, A.C.; Inoue-Nagata, A.K. Sequence analysis of the glycoproteins of Tomato chlorotic spot virus and Groundnut ringspot virus and comparison with other tospoviruses. *Virus Genes* **2004**, 29, 321–328.

Luo, Q.; Hu, S.; Lin, Q.; Xu, F.; Peng, J.; Zheng, H.; Wu, G.; Rao, S.; Chen, J.; Lu, Y.; Guo, F.; Yan, F. Complete genome sequence of a novel foveavirus isolated from Allium sativum L. in China. *Arch. Virol.* **2021**, 166, 983-986.

Maachi, A.; Nagata, T.; Silva, J.M.F. Date palm virus A: first plant virus found in date palm trees. *Virus Genes* **2020**, 56, 792-795.

Maes, P.; Adkins, S.; Alkhovsky, S.V.; Avšič-Županc, T.; Ballinger, M.J.; Bente, D.A.; Beer, M.; Bergeron, É.; Blair, C.D.; Briese, T.; et al. Taxonomy of the order Bunyavirales: second update 2018. *Arch. Virol.* **2019**, 164, 927–941.

Marais, A.; Faure, C.; Theil, S.; Candresse, T. Characterization of the virome of shallots affected by the shallot mild yellow stripe disease in France. *PloS ONE* **2019**, 14, e0219024.

Marais, A.; Faure, C.; Candresse, T. New insights into Asian prunus viruses in the light of NGS-based full genome sequencing. *PLoS One* **2016**, 7, e0146420.

Marais, A.; Faure, C.; Mustafayev, E.; Candresse, T. Characterization of new isolates of Apricot vein clearing-associated virus and of a new Prunus-infecting virus: Evidence for recombination as a driving force in Betaflexiviridae evolution. *PLoS ONE* **2015**, 10, e0129469.

Marais, A.; Šafářová, D.; Navrátil, M.; Faure, C.; Cornaggia, D.; Brans, Y.; Suchá, J.; Candresse, T. Complete genome sequence of cherry virus T, a novel cherry-infecting tepovirus. *Arch. Virol.* **2020**, 165, 1711-1714.

Martínez, F.; Toft, C.; Hillung, J.; Giménez-Santamarina, S.; Yenush, L.; Rodrigo, G.; Elena, S.E. A comprehensive physical interaction map between the Turnip mosaic potyvirus and Arabidopsis thaliana proteomes. *Res. Sq.* **2021**, 10.21203/rs.3.rs-149993/v2.

Maree, H.J.; Blouin, A.G.; Diaz-Lara, A.; Mostert, I.; Al Rwahnih, M.; Candresse, T. Status of the current vitivirus taxonomy. *Arch. Virol.* **2020**, 165, 451-458.

Martin, D.P.; Varsani, A.; Roumagnac, P.; Botha, G.; Maslamoney, S.; Schwab, T.; Kelz, Z.; Kumar, V.; Murrell, B. RDP5: a computer program for analyzing recombination in, and removing signals of recombination from, nucleotide sequence datasets. *Virus Evol.* **2020**, 7, veaa087.

Martínez, R.T.; de Almeida, M.M.S.; Rodriguez, R.; de Oliveira, A.S.; Melo, F.L.; Resende, R.O. Identification and genome analysis of tomato chlorotic spot virus and dsRNA viruses from coinfected vegetables in the Dominican Republic by high-throughput sequencing. *Virol. J.* **2018**, 15, 24.

Melcher, U. The '30K' superfamily of viral movement proteins. *J. Gen. Virol.* **2000**, 81, 257–266.

Merkling, S.H.; Overheul, G.J.; van Mierlo, J.T.; Arends, D.; Gilissen, C.; van Rij, R.P. The heat shock response restricts virus infection in Drosophila. *Sci. Rep.* **2015**, 5, 12758.

Meyniel-Schicklin, L.; de Chassey, B.; André, P.; Lotteau, V. Viruses and interactomes in translation. *Mol. Cell Proteomics* **2012**, 11, M111.014738.

Miguel-Aliaga, I.; Jasper, H.; Lemaitre, B. Anatomy and physiology of the digestive tract of Drosophila melanogaster. *Genetics* **2018**, 210, 357–396.

van Mierlo, J.T.; Bronkhorst, A.W.; Overheul, G.J.; Sadanandan, S.A.; Ekström, J.O.; Heestermans, M.; Hultmark, D.; Antoniewski, C.; van Rij, R.P. Convergent evolution of argonaute-2 slicer antagonism in two distinct insect RNA viruses. *PLoS Pathog.* **2012**, 8, e1002872.

Morozov, S.Y.; Solovyev, A.G. Triple gene block: modular design of a multifunctional machine for plant virus movement. *J. Gen. Virol.* **2003**, 84, 1351-1366.

Muhire, B.M.; Varsani, A.; Martin, D.P. SDT: a virus classification tool based on pairwise sequence alignment and identity calculation. *PloS one* **2014**, 9, e108277.

Mumo, N.N.; Mamati, G.E.; Ateka, E.M.; Rimberia, F.K.; Asudi, G.O.; Boykin, L.M.; Stomeo, F. Metagenomic analysis of plant viruses associated with papaya ringspot disease in Carica papaya L. in Kenya. *Front. Microbiol.* **2020**, 11, 205.

Nguyen, L.T.; Schmidt, H.A.; von Haeseler, A.; Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **2015**, 32, 268-274.

Nemchinov, L.G.; Shamloul, A.M.; Zemtchik, E.Z.; Verderevskaya, T.D.; Hadidi, A. Apricot latent virus: a new species in the genus Foveavirus. *Arch. Virol.* **2000**, 145, 1801-1813.

Nicaise, V.; Candresse, T. Plum pox virus capsid protein suppresses plant pathogen□associated molecular pattern (PAMP)□triggered immunity. *Mol. Plant Pathol.* **2017**, 18, 878-886.

de Oliveira, A.S.; Melo, F.L.; Inoue-Nagata, A.K.; Nagata, T.; Kitajima, E.W.; Resende, R.O. Characterization of bean necrotic mosaic virus: a member of a novel evolutionary lineage within the Genus Tospovirus. *PLoS ONE* **2012**, 7, e38634.

Pappu, H.R.; Jones, R.A.; Jain, R.K.; Jain, R.K. Global status of tospovirus epidemics in diverse cropping systems: Successes achieved and challenges ahead. *Virus Res.* **2009**, 141, 219–236.

Paradis, E.; Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **2019**, 35, 526-528.

Park, D.; Zhang, M.; Hahn, Y. Novel Foveavirus (the family *Betaflexiviridae*) species identified in ginseng (Panax ginseng). *Acta Virol.* **2019**, 63, 155-161.

Peracchio, C.; Forgia, M.; Chiapello, M.; Vallino, M.; Turina, M.; Ciuffo, M. A complex virome including two distinct emaraviruses associated with virus-like symptoms in Camellia japonica. *Virus Res.* **2020**, 286, 197964.

Plyusnin, A.; Beaty, B.J.; Elliott, R.M.; Goldbach, R.; Kormelink, R.; Lundkvist, K.A.; Schmaljohn, C.S.; Tesh, R.B. Family – Bunyaviridae. In *Virus Taxonomy: classification and nomenclature of viruses: ninth report of the International Committee on Taxonomy of Viruses*; King, A.M.Q., Adams, M.J., Carstens, E.B., Lefkowitz, E.J., Eds.; Elsevier: San Diego, CA, USA, 2012; pp. 725–741.

Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS one* **2010**, 5, e9490.

Qi, J.; Zhou, Y.; Hua, J.; Zhang, L.; Bian, J.; Liu, B.; Zhao, Z.; Jin, S. The scRNA-seq expression profiling of the receptor ACE2 and the cellular protease TMPRSS2 reveals human organs susceptible to SARS-CoV-2 infection. *Int. J. Environ. Res. Public Health* **2021**, 18, 284.

Rambaut, A.; Lam, T.T.; Max Carvalho, L.; Pybus, O.G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2016**, 2, vew007.

Ravindra, N.G.; Alfajaro, M.M.; Gasque, V.; Huston, N.C.; Wan, H.; Szigeti-Buck, K.; Yasumoto, Y.; Greaney, A.M.; Habet, V.; Chow, R.D.; Chen, J.S.; et al. Single-cell longitudinal analysis of SARS-CoV-2 infection in human airway epithelium identifies target cells, alterations in gene expression, and cell state changes. *PLoS Biol.* **2021**, 19, e3001143.

Rehfeld, J.F. A centenary of gastrointestinal endocrinology. *Horm. Metab. Res.* **2004**, 36, 735–741.

Ren, X.; Wen, W.; Fan, X.; Hou, W.; Su, B.; Cai, P.; Li, J.; Liu, Y.; Tang, F.; Zhang, F.; Yang, Y.; et al. COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* **2021**, 184, 1895-1913.

Revell, L.J. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **2012**, 3, 217-223.

Reynard, J.S.; Brodard, J.; Remoliff, E.; Lefebvre, M.; Schumpp, O.; Candresse, T. A novel foveavirus identified in wild grapevine (Vitis vinifera subsp. sylvestris). *Arch. Virol.* **2020**, 165, 2999–3002.

Reznick, D.; Travis, J. Is evolution predictable? *Science* **2018**, 359, 738-739.

Rodrigo, G.; Carrera, J.; Ruiz-Ferrer, V.; del Toro, F.J.; Llave, C.; Voinnet, O.; Elena, S.F. A meta-analysis reveals the commonalities and differences in Arabidopsis thaliana response to different viral pathogens. *PLoS ONE* **2012**, 7, e40526.

Rotenberg, D.; Jacobson, A.L.; Schneweis, D.J.; Whitfield, A.E. Thrips transmission of tospoviruses. *Curr. Opin. Virol.* **2015**, 15, 80–89.

Russell, A.B.; Trapnell, C.; Bloom, J.D. Extreme heterogeneity of influenza virus infection in single cells. *eLife* **2018**, 7, e32303.

Sanjuán, R.; Illingworth, C.J.; Geoghegan, J.L.; Iranzo, J.; Zwart, M.P.; Ciota, A.T.; Moratorio, G.; Gago-Zachert, S.; Duffy, S.; Vijaykrishna, D. Five challenges in the field of viral diversity and evolution. *Front. Virol.* **2021**, 1, 684949.

Schulte, M.B.; Andino, R. Single-cell analysis uncovers extensive biological noise in poliovirus replication. *J. Virol.* **2014**, 88, 6205–6212.

Shi, M.; Lin, X.D.; Chen, X.; Tian, J.H.; Chen, L.J.; Li, K.; Wang, W.; Eden, J.S.; Shen, J.J.; Liu, L.; Holmes, E.C. The evolutionary history of vertebrate RNA viruses. *Nature* **2018**, 556, 197-202.

Shi, M.; Lin, X.D.; Tian, J.H.; Chen, L.J.; Chen, X.; Li, C.X.; Qin, X.C.; Li, J.; Cao, J.P.; Eden, J.S.; Buchmann, J.; et al. Redefining the invertebrate RNA virosphere. *Nature* **2016**, 540, 539-543.

Shokal, U.; Eleftherianos, I. Evolution and function of thioester-containing proteins and the complement system in the innate immune response. *Front. Immunol.* **2017**, 8, 759.

Silva, G.; Bömer, M.; Rathnayake, A.I.; Sewe, S.O.; Visendi, P.; Oyekanmi, J.O.; Quain, M.D.; Akomeah, B.; Kumar, P.L.; Seal, S.E. Molecular characterization of a new virus species identified in yam (*Dioscorea* spp.) by high-throughput sequencing. *Plants* **2019**, 8, 167.

da Silva, L.A.; Oliveira, A.S.; Melo, F.L.; Ardisson-Araújo, D.M.; Resende, F.V.; Resende, R.O.; Ribeiro, B.M. A new virus found in garlic virus complex is a member of possible novel genus of the family Betaflexiviridae (order *Tymovirales*). *PeerJ* **2019**, 7, e6285.

Singh, L.; Hallan, V.; Martin, D.P.; Ram, R.; Zaidi, A.A. Genomic sequence analysis of four new chrysanthemum virus B isolates: evidence of RNA recombination. *Arch. Virol.* **2012**, 157, 531-537.

Smith, M.D.; Wertheim, J.O.; Weaver, S.; Murrell, B.; Scheffler, K.; Kosakovsky Pond, S.L. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. *Mol. Biol. Evol.* **2015**, 32, 1342-1353.

Simmonds, P.; Adams, M.J.; Benkő, M.; Breitbart, M.; Brister, J.R.; Carstens, E.B.; Davison, A.J.; Delwart, E.; Gorbalenya, A.E.; Harrach, B.; Hull, R. Virus taxonomy in the age of metagenomics. *Nat. Rev. Microbiol.* **2017**, 15, 161-168.

Simon-Loriere, E.; Holmes, E.C. Why do RNA viruses recombine? *Nat. Rev. Microbiol.* **2011**, 9, 617-626.

Sironi, M.; Cagliani, R.; Forni, D.; Clerici, M. Evolutionary insights into host–pathogen interactions from mammalian sequence data. *Nat. Rev. Genet.* **2015**, 16, 224-236.

Steuerman, Y.; Cohen, M.; Peshes-Yaloz, N.; Valadarsky, L.; Cohn, O.; David, E.; Frishberg, A.; Mayo, L.; Bacharach, E.; Amit, I.; et al. Dissection of influenza infection in vivo by single-cell RNA sequencing. *Cell Syst.* **2018**, 6, 679–691.

Stuart, T.; Butler, A.; Hoffman, P.; Hafemeister, C.; Papalexi, E.; Mauck, W.M., 3rd; Hao, Y.; Stoeckius, M.; Smibert, P.; Satija, R. Comprehensive integration of single-cell data. *Cell* **2019**, 177, 1888–1902.

Sui, X.; McGrath, M.T.; Zhang, S.; Wu, Z.; Ling, K.S. First report of tomato chlorotic spot virus infecting tomato in New York. *Plant Dis.* **2018**, 102, 460.

Takeda, A.; Sugiyama, K.; Nagano, H.; Mori, M.; Kaido, M.; Mise, K.; Tsuda, S.; Okuno, T. Identification of a novel RNA silencing suppressor, NSs protein of Tomato spotted wilt virus. *FEBS Lett.* **2002**, 532, 75–79.

Thekke-Veetil, T.; Ho, T. Molecular characterization of a new vitivirus discovered in a blueberry plant with green mosaic symptoms. *Arch. Virol.* **2019**, 164, 2609-2611.

Thekke-Veetil, T.; McCoppin, N.K.; Hobbs, H.A.; Hartman, G.L.; Lambert, K.N.; Lim, H.S.; Domier, L. Discovery of a Novel Member of the Carlavirus Genus from Soybean (Glycine max L. Merr.). Pathogens 2021, 10, 223.

Thompson, J.R.; Dasgupta, I.; Fuchs, M.; Iwanami, T.; Karasev, A.V.; Petrzik, K.; Sanfaçon, H.; Tzanetakis, I.; van der Vlugt, R.; Wetzel, T.; Yoshikawa, N. ICTV virus taxonomy profile: Secoviridae. *J. Gen. Virol.* **2017**, 98, 529.

de la Torre-Almaráz, R.; Pallás, V.; Sánchez-Navarro, J.A. Molecular characterization of a new trichovirus from peach in Mexico. *Arch. Virol.* **2019**, 164, 2617-2620.

Tumescheit, C.; Firth, A.E.; Brown, K. CIAlign-A highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. *bioRxiv* **2020**, 10.1101/2020.09.14.291484

Turina, M.; Kormelink, R.; Resende, R.O. Resistance to Tospoviruses in Vegetable Crops: Epidemiological and Molecular Aspects. *Annu. Rev. Phytopathol.* **2016**, 54, 347–371.

Vera, M.; Pani, B.; Griffiths, L.A.; Muchardt, C.; Abbott, C.M.; Singer, R.H.; Nudler, E. The translation elongation factor eEF1A1 couples transcription to translation during heat shock response. *eLife* **2014**, 3, e03164.

Villamor, D.E.V.; Eastwell, K.C. Viruses associated with rusty mottle and twisted leaf diseases of sweet cherry are distinct species. *Phytopathology* **2013**, 103, 1287-1295.

Villamor, D.E.V.; Susaimuthu, J.; Eastwell, K.C. Genomic analyses of cherry rusty mottle group and cherry twisted leaf-associated viruses reveal a possible new genus within the family betaflexiviridae. *Phytopathology* **2015**, 105, 399-408.

Wang, L.; Lam, T.T.; Xu, S.; Dai, Z.; Zhou, L.; Feng, T.; Guo, P.; Dunn, C.W.; Jones, B.R.; Bradley, T.; Zhu, H.; Guan, Y.; Jiang, Y.; Yu, G. treeio: an R package for phylogenetic tree input and output with richly annotated and associated data. *Mol. Biol. Evol.* **2020**, 37, 599-603.

Wang, R.; Dong, J.; Wang, Z.; Zhou, T.; Li, Y.; Ding, W. Complete nucleotide sequence of a new carlavirus in chrysanthemums in China. *Arch. Virol.* **2018**, 163, 1973-1976.

Webster, C.G.; Frantz, G.; Reitz, S.R.; Funderburk, J.E.; Mellinger, H.C.; McAvoy, E.; Turechek, W.W.; Marshall, S.H.; Tantiwanich, Y.; McGrath, M.T.; et al. Emergence of Groundnut ringspot virus and Tomato chlorotic spot virus in Vegetables in Florida and the Southeastern United States. *Phytopathology* **2015**, 105, 388–398.

Webster, C.G.; Reitz, S.R.; Perry, K.L.; Adkins, S. A natural M RNA reassortant arising from two species of plant- and insect-infecting bunyaviruses and comparison of its sequence and biological properties to parental species. *Virology* **2011**, 413, 216–225.

Webster, C.G.; Rivera-Vargas, L.I.; Rodrigues, J.C.V.; Mercado, W.; Mellinger, H.C.; Adkins, S. First report of tomato chlorotic spot virus (TCSV) in tomato, pepper, and jimsonweed in Puerto Rico. *Plant Health Prog.* **2013**, 14.

Webster, C.L.; Waldron, F.M.; Robertson, S.; Crowson, D.; Ferrari, G.; Quintana, J.F.; Brouqui, J.M.; Bayne, E.H.; Longdon, B.; Buck, A.H.; et al. The discovery, distribution, and evolution of viruses associated with Drosophila melanogaster. *PLoS Biol.* **2015**, 13, e1002210.

Wesolowska-Andersen, A.; Everman, J.L.; Davidson, R.; Rios, C.; Herrin, R.; Eng, C.; Janssen, W.J.; Liu, A.H.; Oh, S.S.; Kumar, R.; et al. Dual RNA-seq reveals viral infections in asthmatic children without respiratory illness which are associated with changes in the airway transcriptome. *Genome Biol.* **2017**, 18, 12.

Wolf, Y.I.; Kazlauskas, D.; Iranzo, J.; Lucía-Sanz, A.; Kuhn, J.H.; Krupovic, M.; Dolja, V.V.; Koonin, E.V. Origins and evolution of the global RNA virome. *mBio* **2018,** 9, e02329-18.

Wu, L.; Liu, H.; Bateman, M.; Komorowska, B.; Li, R. First identification and molecular characterization of a novel cherry robigovirus. *Arch. Virol.* **2019**, 164, 3103-3106.

Xin, X.; Wang, H.; Han, L.; Wang, M.; Fang, H.; Hao, Y.; Li, J.; Zhang, H.; Zheng, C.; Shen, C. Single-cell analysis of the impact of host cell heterogeneity on infection with foot-and-mouth disease virus. *J. Virol.* **2018**, 92, e00179-18.

Xu, H.; Zhong, L.; Deng, J.; Peng, J.; Dan, H.; Zeng, X.; Li, T.; Chen, Q. High expression of ACE2 receptor of 2019-nCoV on the epithelial cells of oral mucosa. *Int. J. Oral Sci.* **2020**, 12, 8.

Yoon, J.Y.; Joa, J.H.; San Choi, K.; Do, K.S.; Lim, H.C.; Chung, B.N. Genetic diversity of a natural population of Apple stem pitting virus isolated from apple in Korea. *Plant Pathol. J.* **2014**, 30, 195.

Yu, G.; He, Q.Y. ReactomePA: An R/Bioconductor package for reactome pathway analysis and visualization. *Mol. BioSyst.* **2016**, 12, 477–479.

Yu, G.; Smith, D.; Zhu, H.; Guan, Y.; Lam, T.T. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **2017**, 8, 28-36.

Zanardo, L.G.; Silva, F.N.; Lima, A.T.M.; Milanesi, D.F.; Castilho-Urquiza, G.P.; Almeida, A.M.R.; Zerbini, F.M.; Carvalho, C.M. Molecular variability of Cowpea mild mottle virus infecting soybean in Brazil. *Arch. Virol.* **2014**, 159, 727-737.

Zanini, F.; Pu, S.Y.; Bekerman, E.; Einav, S.; Quake, S.R. Single-cell transcriptional dynamics of flavivirus infection. *eLife* **2018**, 7, e32942.

Zhang, Y.Z.; Chen, Y.M.; Wang, W.; Qin, X.C.; Holmes, E.C. Expanding the RNA virosphere by unbiased metagenomics. *Annu. Rev. Virol.* **2019**, 6, 119-139.

Zhao, L.; Cao, M.; Huang, Q.; Jing, M.; Bao, W.; Zhang, Y.; Hou, C.; Wu, Y.; Wang, Q.C. Occurrence and molecular characterization of Actinidia virus C (AcVC), a novel vitivirus infecting kiwifruit (*Actinidia* spp.) in China. *Plant Pathol.* **2020**, 6, 775-782.

Zheng, G.X.; Terry, J.M.; Belgrader, P.; Ryvkin, P.; Bent, Z.W.; Wilson, R.; Ziraldo, S.B.; Wheeler, T.B.; McDermott, G.P.; Zhu, J.; et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **2017**, 8, 14049.

Zhou, J.; Zhang, Z.; Lu, M.; Xiao, H.; Habili, N.; Li, S. Complete nucleotide sequence of a new virus, peach chlorotic leaf spot virus, isolated from flat peach in China. *Arch. Virol.* **2018**, 163, 3459-3461.

Zhu, Y.; Yongky, A.; Yin, J. Growth of an RNA virus in single cells reveals a broad fitness distribution. *Virology* **2009**, 385, 39–46.

Zou, X.; Chen, K.; Zou, J.; Han, P.; Hao, J.; Han, Z. Single-cell RNA-seq data analysis on the receptor ACE2 expression reveals the potential risk of different human organs vulnerable to 2019-nCoV infection. *Front. Med.* **2020**, 14, 185-192.